UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

TEMPUS AI, INC.,
Petitioner,
v.

GUARDANT HEALTH, INC.,
Patent Owner.

Case IPR2025-01434

U.S. Patent No. **11,149,306**

DECLARATION OF SYLVIA D. HALL-ELLIS, Ph.D., IN SUPPORT OF PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 11,149,306

I, Sylvia D. Hall-Ellis, declare as follows:

I. INTRODUCTION

1. My name is Sylvia D. Hall-Ellis. I have been retained by Tempus AI, Inc. ("Tempus" or "Petitioner") as an expert in this proceeding before the United States Patent and Trademark Office ("PTO") Patent Trial and Appeal Board ("Board") regarding U.S. Patent No. 11,149,306 ("the '306 Patent") (the "Challenged Patent").

2. I have written this Declaration at the request of Petitioner to provide my expert opinion regarding the authenticity and public availability of several publications relevant to the concurrently filed petition for *inter partes* review of the '306 Patent. This Declaration sets forth my opinions in detail and provides the basis for my opinions.

3. I reserve the right to supplement or amend my opinions, and bases for them, in response to any additional evidence, testimony, discovery, argument, and/or other additional information that may be provided to me after the date of this declaration.

4. I am being compensated for my time spent working on this matter at my normal consulting rate, plus reimbursement for any additional reasonable expenses. My compensation is not in any way tied to the content of this declaration,

the substance of my opinions, or the outcome of this proceeding. I have no other interests in this proceeding or with any of the parties.

5. All of the materials that I considered and relied upon are discussed explicitly in this declaration.

II. QUALIFICATIONS

6. I am currently an Adjunct Professor in the School of Information at San José State University in San José, California. I obtained a Master of Library Science from the University of North Texas in 1972 and a Ph.D. in Library and Information Science from the University of Pittsburgh in 1985. Over the last sixty years, I have held various positions in the field of library and information resources. I was first employed as a librarian in 1966 and have been involved in the field of library sciences since, holding numerous positions.

7. I am a member of the American Library Association (ALA) and its Association for Library Collections & Technical Services (ALCTS) Division, and I served on the Committee on Cataloging: Resource and Description (which wrote the new cataloging rules) and as the chair of the Committee for Education and Training of Catalogers and the Competencies and Education for a Career in Cataloging Interest Group. I also served as the founding Chair of the ALCTS Division's Task Force on Competencies and Education for a Career in Cataloging. Additionally, I have served as the Chair for the ALA Office of Diversity's Committee on Diversity,

as a member of the REFORMA National Board of Directors, as a member of the Editorial Board for the ALCTS premier cataloging journal, *Library Resources and Technical Services*, and as a Co-Chair for the Library Research Round Table of the American Library Association. Currently I serve as a member of the CORE Margaret Mann Citation Committee and the CORE Awards & Scholarships Coordination Committee.

8. I have also given over one hundred presentations in the field, including several on library cataloging systems and Machine-Readable Cataloging ("MARC") standards. My current research interests include library cataloging systems, metadata, and organization of electronic resources.

9. I have been deposed thirty-nine times: (1) *Symantec Corp. v. Finjan, Inc.*, Petition for *Inter Partes Review* of U.S. Patent No. 7,613,926, IPR 2015-01892, May 26, 2016, on behalf of Symantec Corp.; (2) *Symantec Corp. v. Finjan, Inc.*, 14-cv-299-HSG (N.D. Cal.), on behalf of Symantec Corp., September 14, 2017; (3) one deposition for ten matters: *Intellectual Ventures I LLC v. AT&T Mobility LLC*; AT&T Mobility II LLC, New Cingular Wireless Services, Inc., SBC Internet Services, Inc., Wayport, Inc., and Cricket Wireless LLC, C.A. No. 12-193 (LPS); *Intellectual Ventures II LLC v. AT&T Mobility LLC*; AT&T Mobility II LLC, New Cingular Wireless Services, Inc., SBC Internet Services, Inc., Wayport, Inc., and Cricket Wireless LLC, C.A. No. 13-1631 (LPS);

Intellectual Ventures I LLC v. T-Mobile USA, Inc. and T-Mobile US, Inc., C.A. No. 13-1632 (LPS); *Intellectual Ventures II LLC v. T-Mobile USA, Inc. and T-Mobile US, Inc.*, C.A. No. 13-1633 (LPS); *Intellectual Ventures I LLC, v. Nextel Operations, Inc., Sprint Spectrum L.P., Boost Mobile, LLC and Virgin Mobile USA, L.P.*, C.A. No. 13-1634 (LPS); *Intellectual Ventures II LLC v. Nextel Operations, Inc., Sprint Spectrum L.P., Boost Mobile, LLC and Virgin Mobile USA, L.P.*, C.A. No. 13-1635 (LPS); *Intellectual Ventures I LLC, v. United States Cellular Corporation*, C.A. No. 13-1636 (LPS); *Intellectual Ventures I LLC v. United States Cellular Corporation*, C.A. No. 13-1637 (LPS); *Intellectual Ventures II LLC v. AT&T Mobility LLC, AT&T Mobility II LLC, New Cingular Wireless Services, Inc.*, C.A. No. 15-799 (LPS); *Intellectual Ventures I LLC v. T-Mobile USA, Inc. and T-Mobile US, Inc.*, C.A. No. 15-800 (LPS), on behalf of AT&T Mobility LLC; AT&T Mobility II LLC, Boost Mobile, LLC Cricket Wireless LLC, Nextel Operations, Inc., New Cingular Wireless Services, Inc., SBC Internet Services, Inc., Sprint Spectrum L.P., T-Mobile USA, Inc., T-Mobile US, Inc., United States Cellular Corporation Virgin Mobile USA, L.P., and Wayport, Inc., November 15, 2016; (4) *Hitachi Maxell, LTD., v. Top Victory Electronics (Taiwan) Co. Ltd., et al.*, 2:14-cv-1121 JRG-RSP (E.D. Texas), on behalf of Top Victory Electronics (Taiwan) Co. LTD, *et. al.*, January 20, 2016; (5) *Sprint Spectrum, L.P. v. General Access Solutions, Ltd.*, Petition for *Inter Partes Review* of U.S. Patent No. 7,173,916, IPR 2017-01885, on behalf of

Sprint Spectrum L.P., July 13, 2018; (6) *Nichia Corporation v. Vizio, Inc*., 8:16-cv-00545; on behalf of Vizio, Inc., October 12, 2018; (7) *Intellectual Ventures I LLC, v. T-Mobile USA, Inc., T-Mobile US, Inc., Ericsson Inc., and Telefonaktiebolaget LM Ericsson,* 2:17-cv-00557 (JRG), on behalf of T-Mobile USA, Inc., T-Mobile US, Inc., Ericsson Inc., and Telefonaktiebolaget LM Ericsson, October 19, 2018; (8) *Pfizer, Inc. v. Biogen, Inc.*, Petition for *Inter Partes Review* of U.S. Patent No. 8,821,873, IPR 2017-01168, on behalf of Pfizer, November 3, 2018; (9) *Finjan, Inc. v. ESET, LLC and ESET SPOL. S.R.O.*, 3:17-cv-00183-CAB-BGS, on behalf of ESET, January 15, 2019; (10) *Finjan, Inc. v. Cisco Systems, Inc.*, 5:17-cv-00072-BLF-SVK, on behalf of Cisco Systems, Inc., September 6, 2019; (11) *Facebook, Inc., Instagram, LLC and Whatsapp Inc. v. Blackberry Limited* for *Inter Partes Review* of U.S. Patent No. 9,349,120 B2, on behalf of Facebook, Inc., Instagram, LLC and Whatsapp Inc., December 20, 2019; (12) *3Shape A/S and 3Shape Inc. v. Align Technology, Inc.,* Petition for *Inter Partes Review* of U.S. Patent No. 7,156,661, IPR 2020-00222 and IPR 2020-00223, August 10, 2020, on behalf of 3Shape A/S and 3Shape Inc.; (13) *Finjan Inc. v. Rapid7, Inc. and Rapid7 LLC*, District of Delaware, 1:18-cv-01519-MN, September 15, 2020, on behalf of Rapid7 Inc. and Rapid7 LLC; (14) *VLSI Technology LLC v. Intel Corporation*, Western District of Texas, 6:19-cv-00254, 6:19-cv-00255, 6:19-cv-00256, September 23, 2020, on behalf of the Intel

Corporation; (15) *Finjan Inc. v. Sonicwall, Inc.*, Northern District of California, 5:17-cv-04467-BLF-HRL, October 27, 2020, on behalf of Sonicwall, Inc.; (16) *VLSI Technology, LLC v. Intel Corporation*, District of Delaware, 1:18-cv-00966-CFC-CJB, February 5, 2021, on behalf of the Intel Corporation; (17) *Unified Patents, LLC v. 3BCOM, LLC*, Petition for *Inter Partes Review* of U.S. Patent 7,127,210, February 11, 2021, on behalf of Unified Patents; (18) *Finjan, Inc. v. Qualsys, Inc.*, Northern District of California, 4:18-cv-07229-YGR, March 1, 2021, on behalf of Qualsys, Inc.; (19) *Qualcomm, Inc. v. Monterey Research LLC*, Petition for *Inter Partes Review* of U.S. Patent 6,534,805, May 6, 2021, on behalf of Qualcomm, Inc.; (20) *Hulu, LLC v. Sound View Innovations, LLC*, Petition for *Inter Partes Review* of U.S. Patent 5,806,062, May 14, 2021, on behalf of Hulu, LLC; and, (21) *VLSI Technology LLC v. Intel Corporation*, Western District of Texas, 6:19-cv-00254, 6:19-cv-00255 and 6:19-cv-00256, August 3, 2021, on behalf of Intel Corporation; (22) *Liquidia Technologies, Inc. v. United Therapeutics Corporation*, Petition for *Inter Partes Review* of U.S. Patent 10,716,793 B2, October 20, 2021, on behalf of Liquidia Technologies, Inc.; (23) *EcoFactor, Inc. v. Google, Inc.*, Western District of Texas, 6:20-cv-00075 (ADA), 6:20-0078-ADA, and 6:20-cv-00080 ADA, October 27, 2021, on behalf of Google, Inc.; (24) *Liquidia Technologies, Inc. v. United Therapeutics Corporation*, Petition for *Inter Partes Review* of U.S. Patent 10,716,793 B2, March

11, 2022, on behalf of Liquidia Technologies, Inc.; (25) *Juniper Networks, Inc. v. Swarm Technology LLC*, Petition for *Inter Partes Review* of U.S. Patent 9,852,004, May 5, 2022, on behalf of Juniper Networks, Inc.; and, (26) *ParkerVision, Inc. v. Intel Corporation*, Western District of Texas, 6:20-cv-108 (ADA), October 26, 2022, on behalf of the Intel Corporation; (27) *CommScope v. TQ Delta*, Petition for *Inter Partes Review* of U. S. Patents 9,094,348 and 10,833,809, February 16, 2023, on behalf of CommScope; (28) *WSOU Investments, LLC d/b/a Brazos Licensing and Development v. Salesforce, Inc.*, Western District of Texas, 6:20-cv-01164-ADA, April 18, 2023, on behalf of Salesforce, Inc.; (29) *Regents of the University of Minnesota v. AT&T Mobility LLC*, District of Minnesota, 14:cv-4999-JRT-TNL & 14-cv-4669-JRT-TNL, May 18, 2023, on behalf of AT & T Mobility LLC; (30) *VLSI Technology LLC v. Intel Corporation*, Northern District of California, 5:17-cv-0571-BLF-NC, July 11, 2023, on behalf of Intel Corporation; (31) *Trustees of Purdue University v. ST Microelectronics*, Western District of Texas, 6:21-cv-00727-ADA, July 27, 2023, on behalf of ST Microelectronics; (32) *Advanced Micro Devices, Inc. v. Realtek Semiconductor Corp.*, Petition for *Inter Partes Review* of U. S. Patents 7,936,245 & 9,590,582, August 10, 2023, on behalf of Advanced Micro Devices, Inc.; (33) *Motif Foodworks, Inc. v. Impossible Foods, Inc.*, Petition for *Inter Partes Review* of U. S. Patent 9,943,096, August 24, 2023, on behalf of Motif Foodworks, Inc.; (34) *Network System Technologies, LLC v.*

Texas Instruments, Inc., Eastern District of Texas, 2:22-cv-00482, March 20, 2024, on behalf of Texas Instruments, Inc.; (35) *Network System Technologies, LLC v. Texas Instruments, Inc.*, Eastern District of Texas, 2:22-cv-00482, May 22, 2024, on behalf of Texas Instruments, Inc.; (36) *Pardalis Technology Licensing, LLC v. International Business Machines, Corporation*, Eastern District of Texas, 2:22-cv-00452, May 23, 2024, on behalf of International Business Machines, Corporation; (37) *Meta Platforms, InstaGram, Inc., Whatsapp LLC, Facebook Technologies LLC, and Giphy, Inc. v. VL Collective IP LLC*, Petition for *Inter Partes Review* of U. S. Patent 7,436,980, July 18, 2024, on behalf of Meta Platforms, InstaGram, Inc., Whatsapp LLC, Facebook Technologies LLC, and Giphy, Inc.; (38) *LCY Biotechnology Holding, Inc. v. RadiciChimica S.p.A*, Petition for *Inter Partes Review* of U. S. Patent 11,781,148, August 6, 2024, on behalf of LCY Biotechnology Holding, Inc.; and, (39) *SAP America, Inc. v. Valtrus Innovations Ltd., Key Patent Innovations Ltd., and Patent Platform Services, LLC*, Petition for *Inter Partes Review* of U. S. Patent 7,313,575, June 18, 2025, on behalf of SAP America, Inc.

10. My full curriculum vitae is provided as Attachment A to this declaration.

III. PRELIMINARIES

A. Scope of Declaration and Legal Standards

11. I am not an attorney and will not offer opinions on the law. I am, however, rendering my expert opinion on the authenticity of the documents referenced herein and on when and how each of these documents was disseminated or otherwise made available to the extent that persons interested and ordinarily skilled in the subject matter or art, exercising ordinary diligence, could have located the documents before December 28, 2013.

12. I am informed by counsel that a printed publication qualifies as publicly accessible as of the date it was disseminated or otherwise made available such that a person interested in and ordinarily skilled in the relevant subject matter could locate it through the exercise of ordinary diligence.

13. While I understand that the determination of public accessibility under the foregoing standard rests on a case-by-case analysis of the facts particular to an individual publication, I also understand that a printed publication is rendered "publicly accessible" if it is cataloged and indexed by a library such that a person interested in the relevant subject matter could locate it (*i.e.*, I understand that cataloging and indexing by a library is sufficient, though there are other ways that a printed publication may qualify as publicly accessible). One manner of sufficient indexing is indexing according to subject matter category. I understand that the

cataloging and indexing by a single library of a single instance of a particular printed publication is sufficient, even if the single library is in a foreign country. I understand that, even if access to a library is restricted, a printed publication that has been cataloged and indexed therein is publicly accessible so long as a presumption is raised that the portion of the public concerned with the relevant subject matter would know of the printed publication. I also understand that the cataloging and indexing of information that would guide a person interested in the relevant subject matter to the printed publication, such as the cataloging and indexing of an abstract for the printed publication, is sufficient to render the printed publication publicly accessible.

14. I understand that routine business practices, such as general library cataloging and indexing practices, can be used to establish an approximate date on which a printed publication became publicly accessible.

B. Persons of ordinary skill in the art.

15. I am told by counsel that the subject matter of this proceeding relates to DNA sequencing technology.

16. I have been informed by counsel that a "person of ordinary skill in the art at the time of the invention" ("POSITA") is a hypothetical person who is presumed to be familiar with the relevant field and its literature at the time of the inventions. This hypothetical person is also a person of ordinary creativity, capable of understanding the scientific principles applicable to the pertinent field.

17. I am told by counsel that persons of ordinary skill in this subject matter or art would have a Ph.D. in bioinformatics, genetics, molecular biology or a related field, and at least five years of research in an academic or industry setting, including at least two to three years of research experience in the field of cancer genomics. I am also told by counsel that persons of ordinary skill in this subject matter or art would have had knowledge of DNA sequencing, including NGS and related sequencing methods, and related sample preparation techniques; bioinformatics methods for analyzing sequence reads and mapping sequence reads onto genomes; and methods for identifying genetic variants in a sample. I have been further informed by counsel that a person of ordinary skill in the art would have been familiar with and able to understand the information known in the art relating to these fields, including the publications discussed in this declaration.

18. It is my opinion that such a person would have been engaged in research, learning, study, and practice in the field, and possibly formal instruction so that bibliographic resources relevant to his or her research would be familiar. By not later than 2013, such a person would have had access to a vast array of long-established print resources as well as to a rich set of online resources providing indexing information, abstracts, and full text services for publications relevant to the field of this dispute.

IV. LIBRARY PROFESSIONAL PRACTICES

19. In preparing this declaration, I used authoritative databases, such as the OCLC bibliographic database, the Library of Congress Online Catalog, and the Library of Congress Subject Authorities, to confirm citation details of the various publications discussed. Unless I note otherwise below in reference to a specific serial publication, it is my expert opinion that this standard protocol was followed for the serial publication discussed below.

20. *Indexing*. A researcher may discover material relevant to his or her topic in a variety of ways. One common means of discovery is to search for relevant information in an index of periodicals and other publications. Having found relevant material, the researcher will then normally obtain it online, look for it in libraries, or purchase it from the publisher, a bookstore, a document delivery service, or other provider. Sometimes, the date of a document's public accessibility will involve both indexing and library date information. However, date information for indexing entries is often unavailable. This is especially true for online indices.

21. Indexing services use a wide variety of controlled vocabularies to provide subject access and other means of discovering the content of documents. The Library of Congress Subject Authorities includes standard forms of terms and cross references that are included in bibliographic records. Subject headings are terms that an individual seeking a document regardless of format can enter in the

search bar of the online catalog. Subjects also connect an authenticated term (one included in the Library of Congress subject headings list) with related, broader, and narrower terms. The formats in which these access terms are presented vary from service to service.

22. Online indexing services commonly provide bibliographic information, abstracts, and full-text copies of the indexed publications, along with a list of the documents cited in the indexed publication. These services also often provide lists of publications that cite a given document. A citation of a document is evidence that the document was publicly available and in use by researchers no later than the publication date of the citing document. Prominent indexing services include *Scopus,* the *IEEE Xplore* database, the *ACM Digital Library, Google Scholar,* and the *Internet Archive.*

23. A major firm known for the breadth of subjects and comprehensive treatment in the preparation of index volumes, the H. W. Wilson Company offered these reference resources since the firm was founded in 1898. The *Reader's Guide to Periodical Literature* is one of the best-known titles available from H. W. Wilson. Each volume includes a comprehensive index for 300 of the most popular and important periodicals published in the United States and Canada. Information seekers have subject access expressed in plain language terminology, author access, and cross references to find the desired results from their searches. The family of

index titles included *Science & Technology Index*, *Business Periodicals*, *Applied Science & Technology Index*, *Humanities Index*, *Biological & Agricultural Index*, and *Industrial Arts Index*. These printed indices have been superseded by digital database offerings available to information seekers through Ebsco.

24. Information seekers also used printed versions of *Chemical Abstracts* and *Index Medicus* to locate articles, scientific declarations, and research papers. *Chemical Abstracts* began publication in 1907 and by 2007 its databases "contained more than 27 million records of journal and patent literature."[1] In 2010, *Chemical Abstracts* discontinued the print index. Access is now available through two electronic databases: *CAplus* and *Registry*.[2] *Index Medicus* is a bibliographic index to medical science information, started in 1879.[3] Currently, *PubMed* includes

[1] Chemical Abstracts Service, "CAS History," *available at* https://www.cas.org/about/cas-history (last accessed August 12, 2025); American Chemical Society, "Chemical Abstracts Service," *available at* https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/cas.html (last accessed August 12, 2025).

[2] American Chemical Society, "Chemical Abstracts Service," *available at* https://www.acs.org/content/acs/en/education/whatischemistry/landmarks/cas.html (last accessed August 12, 2025).

[3] Stephen Greenberg and Patricia Gallagher, "The great contribution: *Index Medicus*, *Index-Catalogue*, and IndexCast," *Journal of Medical Library Association*, April 2009, available at

the content that had been published as *Index Medicus*.[4]

25. Established in 1836 as part of the Surgeon General of the Army's Office, the National Library of Medicine (NLM) has been instrumental in the development of access to medical books, journals, and research publications. In the 1970s, NLM introduced MEDLINE, followed by "the establishment of the National Center for Biotechnology Information in 1988, the introduction of free MEDLINE in 1997, the creation of consumer-friendly MedlinePlus in 1998, and the introduction of ClinicalTrials.gov in 2000."[5]

26. Online indexing services such as Google Scholar[6] or ScienceDirect[7] commonly provide bibliographic information, abstracts, and full-text copies of the indexed publications, along with a list of the documents cited in the indexed publication. These services also often provide lists of publications that cite a given

---

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2670211/ (last accessed August 12, 2025).

[4] U.S. National Library of Medicine, "List of All Journals Cited in PubMed®," https://www.nlm.nih.gov/bsd/serfile_addedinfo.html (last accessed August 12, 2025).

[5] U.S. National Library of Medicine, "A Brief History of NLM," available at https://www.nlm.nih.gov/about/briefhistory.html.

[6] https://scholar.google.com/.

[7] https://www.sciencedirect.com/.

document.

27. A citation of a document by another is evidence that the document was publicly available and in use no later than the publication date of the citing document.

A. MARC RECORDS AND THE ONLINE LIBRARY CATALOG

28. I am fully familiar with the library cataloging standard known as the MARC standard, which is an industry-wide standard method of storing and organizing library catalog information.[8] MARC was first developed in the 1960's by the Library of Congress. A MARC-compatible library is one that has a catalog consisting of individual MARC records for each of its items. Today, MARC is the primary communications protocol for the transfer and storage of bibliographic metadata in libraries.[9]

---

[8] The full text of the standard is available from the Library of Congress at http://www.loc.gov/marc/bibliographic/ (last visited August 12, 2025).

[9] Almost every major library in the world is MARC-compatible. *See, e.g.*, *MARC Frequently Asked Questions (FAQ)*, Library of Congress, https://www.loc.gov/marc/faq.html (last visited August 21, 2025) ("MARC is the acronym for Machine-Readable Cataloging. It defines a data format that emerged from a Library of Congress-led initiative that began nearly fifty years ago. It provides the mechanism by which computers exchange, use, and interpret bibliographic information, and its data elements make up the foundation of most

29. Since at least the early 1970s and continuing to the present day, MARC has been the primary communications protocol for the transfer and storage of bibliographic metadata in libraries.[10] As explained by the Library of Congress:

> You could devise your own method of organizing the bibliographic information, but you would be isolating your library, limiting its options, and creating much more work for yourself. Using the MARC standard prevents duplication of work and allows libraries to better share bibliographic resources. Choosing to use MARC enables libraries to acquire cataloging data that is predictable and reliable. If a library were to develop a "home-grown" system that did not use MARC records, it would not be taking advantage of an industry-wide standard whose primary purpose is to foster communication of information.
>
> Using the MARC standard also enables libraries to make use of commercially available library automation systems to manage library operations. Many systems are available for libraries of all sizes and are designed to work with the MARC format. Systems are maintained and improved by the vendor so that libraries can benefit from the latest

---

library catalogs used today."). MARC is the ANSI/NISO Z39.2-1994 (reaffirmed 2016) standard for Information Interchange Format.

[10] A complete history of the development of MARC can be found in *MARC: Its History and Implications* by Henrietta D. Avram (Washington, DC: Library of Congress, 1975) and available online from the Hathi Trust (https://babel.hathitrust.org/cgi/pt?id=mdp.39015034388556;view=1up;seq=1; last visited August 12, 2025).

> advances in computer technology. The MARC standard also allows libraries to replace one system with another with the assurance that their data will still be compatible.

Why Is a MARC Record Necessary? LIBRARY OF CONGRESS. [11]

30. Thus, almost every major library in the world is MARC-compatible. *See, e.g., MARC Frequently Asked Questions (FAQ)*, LIBRARY OF CONGRESS.[12] ("MARC is the acronym for MAchine-Readable Cataloging. It defines a data format that emerged from a Library of Congress-led initiative that began nearly fifty years ago. It provides the mechanism by which computers exchange, use, and interpret bibliographic information, and its data elements make up the foundation of most library catalogs used today."). MARC is the ANSI/NISO Z39.2-1994 standard (reaffirmed in 2016) for Information Interchange Format. The full text of the standard is available from the Library of Congress.[13]

B. MARC RECORDS

31. A MARC record comprises several fields, each of which contains specific data about the work. Each field is identified by a standardized, unique, three-digit code corresponding to the type of data that follow. For example, a work's title

[11] http://www.loc.gov/marc/umb/um01to06.html#part2.

[12] https://www.loc.gov/marc/faq.html.

[13] http://www.loc.gov/marc/bibliographic/.

is recorded in Field 245; the primary author of the work is transcribed in Field 100; an item's International Standard Book Number ("ISBN") consisting of ten or thirteen digits is transcribed in Field 020; an item's International Standard Serial Number ("ISSN") is transcribed in Field 022; the Library of Congress classification notation is recorded in Field 050; and the publication date is recorded in Field 260 under the subfield "c." If a work is a periodical, then its publication frequency is recorded in Field 310, and the publication dates (e.g., the first and last publication) are recorded in Field 362, which is also referred to as the enumeration/chronology field.[14]

32. The library that created the record is recorded in Field 040 in subfield "a" with a unique library code. When viewing the MARC record online via Online Computer Library Center's ("OCLC") bibliographic database, hovering over this code with the mouse reveals the full name of the library. I used this method of "mousing over" the library codes in the OCLC database to identify the originating library for the MARC records discussed in this declaration. Where this "mouse over" option was not available, I consulted the Directory of OCLC Libraries in order to identify the institution that created the MARC record.[15]

[14] http://www.loc.gov/marc/bibliographic/bd3xx.html.

[15] https://www.oclc.org/en/contacts/libraries.html.

33. MARC records also include several fields that include subject matter classification information. An overview of MARC record fields is available through the Library of Congress.[16] For example, 6XX fields are termed "Subject Access Fields."[17] Among these, for example, is the 650 field; this is the "Subject Added Entry – Topical Term" field.[18] The 650 field is a "[s]ubject added entry in which the entry element is a topical term." These entries "are assigned to a bibliographic record to provide access according to generally accepted thesaurus-building rules (e.g., *Library of Congress Subject Headings* (LCSH), *Medical Subject Headings* (MeSH))." *Id.* Further, MARC records include call numbers, which themselves include a classification number. For example, the 050 field is the "Library of Congress Call Number."[19] A defined portion of the Library of Congress Call (LCC) Number is the classification number, and "source of the classification number is *Library of Congress Classification* and the *LC Classification-Additions and Changes*." Thus, included in the 050 field is a subject matter classification. Further, the 082 field is the "Dewey Decimal Call Number."[20] A defined portion of the

[16] http://www.loc.gov/marc/bibliographic/.

[17] http://www.loc.gov/marc/bibliographic/bd6xx.html.

[18] http://www.loc.gov/marc/bibliographic/bd650.html.

[19] http://www.loc.gov/marc/bibliographic/bd050.html.

[20] http://www.loc.gov/marc/bibliographic/bd082.html.

Dewey Decimal Call (DDC) Number is the classification number, and "source of the classification number is the *Dewey Decimal Classification and Relative Index*." Thus, included in the 082 field is a subject matter classification. Each item in a library has a single classification number. A library selects a classification scheme (*e.g.*, the Library of Congress classification scheme just described or a similar scheme such as the Dewey Decimal classification scheme) and uses it consistently. When the Library of Congress assigns the LCC classification number, it appears as part of the 050 field. When the Library of Congress assigns the DDC classification number, it appears as part of the 082 field. If a local library assigns the classification number, it appears in a 090 field. In either scenario, the MARC record includes a classification number that represents a subject matter classification.

C. OCLC

34. The OCLC was created "to establish, maintain and operate a computerized library network and to promote the evolution of library use, of libraries themselves, and of librarianship, and to provide processes and products for the benefit of library users and libraries, including such objectives as increasing availability of library resources to individual library patrons and reducing the rate of rise of library per-unit costs, all for the fundamental public purpose of furthering ease of access to and use of the ever-expanding body of worldwide scientific, literary

and educational knowledge and information."[21] Among other services, OCLC and its members are responsible for maintaining the WorldCat database (http://www.worldcat.org/), used by independent and institutional libraries throughout the world.

35. OCLC also provides its members online access to MARC records through its OCLC bibliographic database. When an OCLC member institution acquires a work, it creates a MARC record for this work in its computer catalog system in the ordinary course of its business. MARC records created at the Library of Congress are tape-loaded into the OCLC database through a subscription to MARC Distribution Services daily or weekly. Once the MARC record is created by a cataloger at an OCLC member institution or is tape-loaded from the Library of Congress, the MARC record is then made available to any other OCLC members online, and therefore made available to the public. Accordingly, once the MARC record is created by a cataloger at an OCLC member institution or is tape-loaded from the Library of Congress or another library anywhere in the world, any publication corresponding to the MARC record has been cataloged and indexed according to its subject matter such that a person interested in that subject matter

[21] Third Article, Amended Articles of Incorporation of OCLC Online Computer Library Center, Incorporated (available at https://www.oclc.org/content/dam/oclc/membership/articles-of-incorporation.pdf).

could, with reasonable diligence, locate and access the publication through any library with access to the OCLC bibliographic database or through the Library of Congress.

36. When an OCLC member institution creates a new MARC record, OCLC automatically supplies the date of creation for that record. The date of creation for the MARC record appears in the fixed Field (008), characters 00 through 05. The MARC record creation date reflects the date on which, or shortly after which, the item was first acquired or cataloged. Initially, Field 005 of the MARC record is automatically populated with the date the MARC record was created in year, month, day format (YYYYMMDD) (some of the newer library catalog systems also include hour, minute, second (HHMMSS)). Thereafter, the library's computer system may automatically update the date in Field 005 every time the library updates the MARC record (*e.g.*, to reflect that an item has been moved to a different shelving location within the library). Field 005 is visible when viewing a MARC record via an appropriate computerized interface and shows the date upon which the most recent update occurred.

37. Once one library has cataloged and indexed a publication by creating a MARC record for that publication, other libraries that receive the publication do not create additional MARC records—the other libraries instead rely on the original MARC record. They may update or revise the MARC record to ensure accuracy, but

they do not replace or duplicate it. This practice does more than save libraries from duplicating labor. It also enhances the accuracy of MARC records. Further, it allows librarians around the world to know that a particular MARC record is authoritative (in contrast, a hypothetical system wherein duplicative records were created would result in confusion as to which record is authoritative).

38. Catalogers can create MARC records for all types of print, online, and digital resources. The date of creation of the MARC record by a cataloger at an OCLC member institution reflects when the underlying item is accessible to the public. Upwards of two-thirds to three-quarters of book sales to libraries come from a jobber or wholesaler for online and print resources. These resellers make it their business to provide books to their customers as fast as possible, often providing turnaround times of only a single day after publication. Libraries purchase a significant portion of the balance of their books directly from publishers themselves, which provide delivery on a similarly expedited schedule. In general, libraries make these purchases throughout the year as the books are published and shelve the books as soon thereafter as possible in order to make the books available to their patrons. Thus, books are generally available at libraries across the country within just a few days of publication.

D. JOURNALS

39. Catalogers can create MARC records for all types of print, online, and

digital resources. For example, MARC records cover serial publications, including both serially-published monographs and journals. OCLC hosts MARC records for more than 320 million serial publications. Serial publications are those publications that have the same collective title but are intended to be continued indefinitely with enumeration such as a volume or issue number (*e.g.*, magazines, journals, etc.). In the OCLC bibliographic database, the first issue of the serial publication is typically cataloged (*i.e.*, a corresponding MARC record is created), but the date is left open-ended with the use of a punctuation mark such as a dash. OCLC serial publication MARC records represent the entire run of the serial title. With knowledge of the first issue published, future issues can be predicted based on the information provided in the MARC record, for example in field 362. In my extensive professional experience, is it highly unusual for a library to stop collecting and shelving a serial publication prior to the end of its publication run. If a subscription to a serial publication ends its run or is cancelled before the end of its run, the library will denote that it has stopped receiving new volumes by filling in the end date in the MARC record.

40. The handling of printed journal subscriptions is shown on the cover or the first few pages of individual issues. As was the best practice among libraries, issues arrived at a central facility and were immediately received, verified as part of a subscription, checked in, and stamped or labeled with the institution's name and

date. Determining that the issue was part of the library subscription ensured that the entire set of publications for the year had been received so that they could be professionally bound and retained. This process also verified that all of the published issues arrived so that the library staff did not have to request or claim an issue that did not arrive as expected. In large public libraries with branches and multi-campus libraries within academic institutions, the journals were sorted and delivered to the subscribing unit. The new issue was placed in the public area; the older issue was stored so that it remained available.

41. The foregoing process has been standard library practice longer than I have been working in the profession. I first learned the steps in the process in the late 1970s and later supervised it. Although the checking in process has become automated and now links electronically to holdings records for the MARC record for each serial title, the manual placement of a stamp or label and placing the issue in a public area has not changed for 50 years. Unless I note otherwise below in reference to a specific serial publication, it is my expert opinion that this standard protocol was followed for the serial publications discussed below.

V. PUBLICATIONS

A. PUBLICATION 1: Exhibit 1006 ("FORSHEW")

42. Attached hereto as Exhibit 1006 is a copy of the journal article titled "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep

Sequencing of Plasma DNA" by Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana W. Y. Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M. Piskorz, Mercedes Jimenez-Linan, David Bentley, James Hadfield, Andrew P. May, Carlos Caldas, James D. Brenton, and Nitzan Rosenfeld (hereafter "Forshew"). The Forshew article was published in Volume 4, Issue 136 of the open access journal *Science Translational Medicine* and was found in the National Library of Medicine (Bethesda, Maryland). The Forshew article appears beginning on page ra68 of this issue dated May 30, 2012. Exhibit 1006 is a true and correct copy of the Forshew article (34 pages) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1006 was found within the custody of a library – a place where an authentic copy of this open access journal would likely be. Exhibit 1006 is a true and correct copy in a condition that creates no suspicion about its authenticity

43. The Forshew article is available from the publisher,[22] the American Association for the Advancement of Science, and through the National Library of

[22] https://www.science.org/doi/10.1126/scitranslmed.3003726.

Medicine.[23]

44. Attached hereto as Attachment 1A is a true and correct copy of the MARC record for the open access journal *Science Translational Medicine* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 1A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 1A.

45. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 1A, that library code is "NSD," which means that the MARC record for this serial was cataloged as part of the National Serials Data Program at the Library of Congress. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the Library of Congress created OCLC record number 299570656 on January 26, 2009, ("090126"). The "BLvl" entry in Attachment 1A is "s," which indicates that the open access journal *Science Translational Medicine* is a serial publication. Field 310 of Attachment 1A reads "Monthly." As noted in the holdings information (362 field), the National Library of Medicine has received the open access journal *Science*

[23] https://pubmed.ncbi.nlm.nih.gov/22649089/.

Translational Medicine since October 2009 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1006 corresponds to the open access journal *Science Translational Medicine* from the time the serial title began publication and includes the date of the Forshew article.

46. Attachment 1A includes an entry in field 060 ("W1 $b SC781Q")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 1A further includes a descriptor term reading "Translational Research, Biomedical $v periodical" (see Attachment 1B, National Library of Medicine subject heading D057170) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 1A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

47. As of January 26, 2009, the MARC record attached hereto as Attachment 1A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the open access journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

48. WorldCat indicates that the open access journal *Science Translational*

Medicine as cataloged at the Library of Congress is currently available from 213 libraries.[24] In view of the above, this issue of the open access journal *Science Translational Medicine* was publicly available on or shortly after May 30, 2012, because by that date it had been cataloged and indexed at the Library of Congress, made part of the OCLC bibliographic database, and received in the National Library of Medicine. For these reasons, it is my opinion that the Forshew article (Exhibit 1006) published in the open access journal *Science Translational Medicine* was publicly available on or shortly after May 30, 2012, as indicated on the publisher's website.

B. PUBLICATION 2: Exhibit 1012 ("SHIROGUCHI")

49. Attached hereto as Exhibit 1012 is a copy of the journal article titled "Digital RNA Sequencing Minimizes Sequence-Dependent Bias and Amplification Noise with Optimized Single-Molecule Barcodes" by Katsuyuki Shiroguchi, Tony Z. Jia, Peter A. Sims, and X. Sunney Xie (hereafter "Shiroguchi"). The Shiroguchi article was published in Volume 109, Number 4 of the journal *Proceedings of the National Academy of Science of the United States of America (PNAS)* and was found in the National Library of Medicine (Bethesda, Maryland). The Shiroguchi article appears beginning on page 1347 of this issue dated January 24, 2012. Exhibit 1012

[24] https://search.worldcat.org/title/299570656.

is a true and correct copy of the Shiroguchi article (pages 1347-1352) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1012 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1012 is a true and correct copy in a condition that creates no suspicion about its authenticity

50. The Shiroguchi article is available from the publisher,[25] the National Academy of Science of the United States of America, through the National Library of Medicine,[26] and the digital repository *ResearchGate.*[27]

51. Attached hereto as Attachment 2A is a true and correct copy of the MARC record for the journal *PNAS* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 2A occurred

[25] https://www.pnas.org/doi/10.1073/pnas.1118018109.

[26] https://pubmed.ncbi.nlm.nih.gov/22232676/.

[27] https://www.researchgate.net/publication/221736129_Digital_RNA_sequencing_minimizes_sequence-dependent_bias_and_amplification_noise_with_optimized_single-molecule_barcodes.

on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 2A.

52. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 2A, that library code is "DNLM," which means that the MARC record for this serial was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number 1607201 on June 26, 1975, ("750626"). The "BLvl" entry in Attachment 2A is "s," which indicates that the journal *PNAS* is a serial publication. Field 310 of Attachment 2A reads "Weekly $b 2004-." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *PNAS* since January 1915 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1012 corresponds to the journal *PNAS* from the time the serial title began publication and includes the date of the Shiroguchi article.

53. Attachment 2A includes an entry in field 060 ("W1 $b PR586A")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 2A further includes a descriptor term reading "Science $v periodical" (see Attachment 2B, National Library of Medicine subject

heading D012586) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 2A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

54. As of June 26, 1975, the MARC record attached hereto as Attachment 2A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

55. WorldCat indicates that the journal *PNAS* as cataloged at the National Library of Medicine is currently available from 1,927 libraries.[28] In view of the above, this issue of the journal *PNAS* was publicly available on or shortly after January 9, 2012, because by that date it had been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Shiroguchi article (Exhibit 1012) published in the journal *PNAS* was publicly available on or shortly after January 9, 2012, as indicated on the publisher's website.[29]

[28] https://search.worldcat.org/title/1607201.

[29] https://www.pnas.org/doi/10.1073/pnas.1118018109.

C. PUBLICATION 3: Exhibit 1014 ("SCHWARZENBACH")

56. Attached hereto as Exhibit 1014 is a copy of the journal article titled "Cell-Free Nucleic Acids as Biomarkers in Cancer Patients" by Heidi Schwarzenbach, Dave S. B. Hoon, and Klaus Pantel (hereafter "Schwarzenbach"). The Schwarzenbach article was published in Volume 11 of the journal *Nature Reviews. Cancer* and was found in the National Library of Medicine (Bethesda, Maryland). The Schwarzenbach article appears beginning on page 426 of this issue dated June 2011. Exhibit 1014 is a true and correct copy of the Schwarzenbach article (pages 426-437) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1014 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1014 is a true and correct copy in a condition that creates no suspicion about its authenticity

57. The Schwarzenbach article is available from the publisher,[30] Springer Nature Limited, and through the National Library of Medicine.[31]

[30] https://www.nature.com/articles/nrc3066.

[31] https://pubmed.ncbi.nlm.nih.gov/21562580/.

58. Attached hereto as Attachment 3A is a true and correct copy of the MARC record for the journal *Nature Reviews. Cancer* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 3A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 3A.

59. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 3A, that library code is "DNLM," which means that the MARC record for this serial was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number 48623853 on July 31, 2001, ("0107331"). The "BLvl" entry in Attachment 3A is "s," which indicates that the journal *Nature Reviews. Cancer* is a serial publication. Field 310 of Attachment 3A reads "Monthly." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *Nature Reviews. Cancer* since October 2001 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1014 corresponds to the journal *Nature Reviews. Cancer* from the time the serial title began publication and includes the date of the Schwarzenbach article.

60. Attachment 3A includes an entry in field 060 ("W1 $b NA839F")—as

described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 3A further includes a descriptor term reading "Neoplasms $v periodical" (see Attachment 3B, National Library of Medicine subject heading D009369) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 3A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

61. As of July 3, 2001, the MARC record attached hereto as Attachment 3A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

62. WorldCat indicates that the journal *Nature Reviews. Cancer* as cataloged at the National Library of Medicine is currently available from 1,652 libraries.[32] In view of the above, this issue of the journal *Nature Reviews. Cancer* was publicly available on or shortly after May 12, 2011, because by that date it had

[32] https://search.worldcat.org/title/48623853.

been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Schwarzenbach article (Exhibit 1014) published in the journal *Nature Reviews. Cancer* was publicly available on or shortly after May 12, 2011, as indicated on the publisher's website.[33]

D. PUBLICATION 4: Exhibit 1015 ("FACT SHEET")

63. Attached hereto as Exhibit 1014 is a technical report titled *Fact Sheet: Human Genomic Variation* prepared by a team at the National Human Genome Research Institute in the National Institute of Health (NIH; Bethesda, Maryland) (hereafter "Fact Sheet"). Exhibit 1015 is a true and correct copy, which I obtained from the National Human Genome Research Institute website.[34] Specifically, the text of the Fact Sheet is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document.

64. Exhibit 1015 includes logo for the NIH. Further, Exhibit 1015 was found within the custody of the issuing federal government agency (i.e., the National Human Genome Research Institute), a place where, if authentic, a copy of this

[33] https://www.nature.com/articles/nrc3066.

[34] https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation.

technical report would likely be. Exhibit 1015 is a true and correct copy in a condition that creates no suspicion about its authenticity.

65. A technical report exhibits the following characteristics: the technical report is attributed to one or more creators or authors (in this case the National Human Genome Research Institute) who affix a date of creation (in this case February 1, 2023), recognize one or more outside funding agencies (in this case the National Institutes of Health), a list of scholarly references (in this case four websites for further reading), and publishes the document through a sponsoring institution or organization (in this case the National Institutes of Health).

66. Based on finding a digital copy of Exhibit 1015 on the Internet, it is my opinion that the technical report titled *Fact Sheet: Human Genomic Variation* was available to the public by or shortly after February 1, 2023, as indicated on the last page of the document.[35] It is also my opinion that there is no reason to doubt the authenticity of Exhibit 1015 given its appearance, contents, and the web address from which it was obtained.

E. PUBLICATION 5: Exhibit 1017 ("METZKER 2010")

67. Attached hereto as Exhibit 1017 is a copy of the journal article titled

[35] https://web.archive.org/web/20230202044317/https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation.

"Sequencing Technologies -- the Next Generation" by Michael L. Metzker (hereafter "Metzker 2010"). The Metzker 2010 article was published in Volume 11, Number 1 of the journal *Nature Reviews. Genetics* and was found in the National Library of Medicine (Bethesda, Maryland). The Metzker 2010 article appears beginning on page 31 of this issue dated January 2010. Exhibit 1017 is a true and correct copy of the Metzker 2010 article (pages 31-46) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1017 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1017 is a true and correct copy in a condition that creates no suspicion about its authenticity

68. The Metzker 2010 article is available from the publisher,[36] Springer Nature Limited, and through the National Library of Medicine.[37]

69. Attached hereto as Attachment 5A is a true and correct copy of the MARC record for the journal *Nature Reviews. Genetics* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC

[36] https://www.nature.com/articles/nrg2626.

[37] https://pubmed.ncbi.nlm.nih.gov/19997069/.

record to meet current cataloging rules. The most recent enhancement to Attachment 5A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 5A.

70. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 5A, that library code is "OUN," which means that the MARC record for this serial was cataloged at the Ohio University Libraries (Athens, Ohio). As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the Ohio University Libraries created OCLC record number 45426000 on November 27, 2000, ("001127"). The "BLvl" entry in Attachment 5A is "s," which indicates that the journal *Nature Reviews. Genetics* is a serial publication. Field 310 of Attachment 5A reads "Monthly." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *Nature Reviews. Genetics* since October 2000 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1017 corresponds to the journal *Nature Reviews. Genetics* from the time the serial title began publication and includes the date of the Metzker 2010 article.

71. Attachment 5A includes an entry in field 060 ("W1 $b NA839H")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal

classification system). Attachment 5A further includes a descriptor term reading "Genetics $v periodical" (see Attachment 5B, National Library of Medicine subject heading D005823) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 5A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

72. As of November 27, 2000, the MARC record attached hereto as Attachment 5A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

73. WorldCat indicates that the journal *Nature Reviews. Genetics* as cataloged at the Ohio University Libraries is currently available from 1,645 libraries.[38] In view of the above, this issue of the journal *Nature Reviews. Cancer* was publicly available on or shortly after December 8, 2009, because by that date it had been cataloged and indexed at the Ohio University Libraries, made part of the OCLC bibliographic database, and received at the National Library of Medicine. For

[38] https://search.worldcat.org/title/45426000.

these reasons, it is my opinion that the Metzker 2010 article (Exhibit 1017) published in the journal *Nature Reviews. Genetics* was publicly available on or shortly after December 8, 2009, as indicated on the publisher's website.[39]

F. PUBLICATION 6: Exhibit 1019 ("METZKER 2006")

74. Attached hereto as Exhibit 1019 is an excerpt from the six-volume set titled *Encyclopedia of Medical Devices and Instrumentation*, 2nd edition, edited by John G. Webster. The excerpt is chapter 8, which is titled "Polymerase Chain Reaction" by Michael L. Metzger and Thomas Caskey (hereafter "Metzger"). The Metzger excerpt appears on pages 380-387 in volume 5 and was found in the National Library of Medicine (Bethesda, Maryland). The Metzger excerpt (Exhibit 1019) is a true and correct copy that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the excerpt is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1019 was found within the custody of a library – a place where an authentic copy of this six-volume set would likely be. Exhibit 1019 is a true and correct copy in a condition that creates no suspicion about its authenticity

75. Attached hereto as Attachment 6A is a true and correct copy of the

[39] https://www.nature.com/articles/nrg2626.

MARC record for the six-volume set titled *Encyclopedia of Medical Devices and Instrumentation*, 2nd edition, in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 6A occurred on June 8, 2024, as shown in field 005 ("20240608"). I personally identified and retrieved the MARC record that is Attachment 6A.

76. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 6A, that library code is "DNLM," which means that the MARC record for this six-volume set was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number 61860839 on October 4, 2005, ("051004").

77. Attachment 6A includes an entry in field 060 ("W13 $b E555 2006")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 6A further includes a descriptor term reading "Equipment and Supplies $v encyclopedia" (see Attachment 6B, National Library of Medicine subject heading D004864) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 6A was indexed according to its subject matter by virtue of the title (field

245), the classification notation (field 060), and the subject heading (field 650).

78. As of October 4, 2005, the MARC record attached hereto as Attachment 6A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that purchased the six-volume set, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

79. WorldCat indicates that the six-volume set titled *Encyclopedia of Medical Devices and Instrumentation*, 2nd edition, as cataloged at the National Library of Medicine is currently available from 450 libraries.[40] In view of the above, this six-volume set titled *Encyclopedia of Medical Devices and Instrumentation*, 2nd edition, was publicly available on or shortly after October 4, 2005, because by that date it had been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Metzger excerpt (Exhibit 1019) published in the six-volume set titled *Encyclopedia of Medical Devices and Instrumentation*, 2nd edition, was publicly available on or shortly after October 4, 2005, as indicated in field 008 of Attachment 6A.

[40] https://search.worldcat.org/title/61860839.

G. PUBLICATION 7: Exhibit 1020 ("MAMANOVA")

80. Attached hereto as Exhibit 1020 is a copy of the journal article titled "Target-Enrichment Strategies for Next Generation Sequencing" by Lira Mamanova, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner (hereafter "Mamanova"). The Mamanova article was published in Volume 7, Number 2 of the journal *Nature Methods* and was found in the National Library of Medicine (Bethesda, Maryland). The Mamanova article appears beginning on page 111 of this issue dated February 2010. Exhibit 1020 is a true and correct copy of the Mamanova article (pages 111-118 and includes supplemental figures and text in the online version (44 pages)) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1020 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1020 is a true and correct copy in a condition that creates no suspicion about its authenticity

81. The Mamanova article is available from the publisher,[41] Springer

[41] https://www.nature.com/articles/nmeth.1419.

Nature Limited, and through the National Library of Medicine.[42]

82. Attached hereto as Attachment 7A is a true and correct copy of the MARC record for the journal *Nature Methods* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 7A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 7A.

83. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 7A, that library code is "NSD," which means that the MARC record for this serial was cataloged as part of the National Serials Data Program at the Library of Congress. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the Library of Congress created OCLC record number 54447242 on February 19, 2004, ("040219"). The "BLvl" entry in Attachment 7A is "s," which indicates that the journal *Nature Methods* is a serial publication. Field 310 of Attachment 7A reads "Monthly." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *Nature Methods* since October 2004 and continues to receive the publication. Accordingly, the MARC record for Exhibit

[42] https://pubmed.ncbi.nlm.nih.gov/20111037/.

1020 corresponds to the journal *Nature Methods* from the time the serial title began publication and includes the date of the Mamanova article.

84. Attachment 7A includes an entry in field 060 ("W1 $b NA812N")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 7A further includes a descriptor term reading "Biomedical Research $x methods" (see Attachment 7B, National Library of Medicine subject heading D035843) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 7A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

85. As of February 19, 2004, the MARC record attached hereto as Attachment 7A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

86. WorldCat indicates that the journal *Nature Methods* as cataloged at the

Library of Congress is currently available from 592 libraries.[43] In view of the above, this issue of the journal *Nature Methods* was publicly available on or shortly after January 28, 2010, because by that date it had been cataloged and indexed at the Library of Congress, made part of the OCLC bibliographic database, and received at the National Library of Medicine. For these reasons, it is my opinion that the Mamanova article (Exhibit 1020) published in the journal *Nature Methods* was publicly available on or shortly after January 28, 2010, as indicated on the publisher's website.[44]

H. PUBLICATION 8: Exhibit 1029 ("LI")

87. Attached hereto as Exhibit 1029 is a copy of the journal article titled "Structure-Independent and Quantitative Ligation of Single-Stranded DNA" by Tian W. Li and Kevin M. Weeks (hereafter "Li"). The Li article was published in Volume 349, Issue 2 of the journal *Analytical Biochemistry* and was found in the National Library of Medicine (Bethesda, Maryland). The Li article appears beginning on page 242 of this issue dated February 15, 2006. Exhibit 1029 is a true and correct copy of the Li article (pages 242-246) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly

[43] https://search.worldcat.org/title/54447242.

[44] https://www.nature.com/articles/nmeth.1419.

from one page to the next; further, there are no visible alterations to the document. Exhibit 1029 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1029 is a true and correct copy in a condition that creates no suspicion about its authenticity

88. The Li article is available from the publisher,[45] Elsevier, and through the National Library of Medicine.[46]

89. Attached hereto as Attachment 8A is a true and correct copy of the MARC record for the journal *Analytical Biochemistry* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 8A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 8A.

90. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 8A, that library code is "DNLM," which means that the MARC record for this serial was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file")

---

45 https://www.sciencedirect.com/science/article/abs/pii/S0003269705007979?via%3Dihub.

46 https://pubmed.ncbi.nlm.nih.gov/16325753/.

in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number1481077 on September 25, 1973, ("730925"). The "BLvl" entry in Attachment 3A is "s," which indicates that the journal *Analytical Biochemistry* is a serial publication. Field 310 of Attachment 8A reads "Eighteen no. a year." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *Analytical Biochemistry* since June 1960 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1029 corresponds to the journal *Analytical Biochemistry* from the time the serial title began publication and includes the date of the Li article.

91. Attachment 8A includes an entry in field 060 ("W1 $b AN1915")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 8A further includes a descriptor term reading "Biochemistry $v periodical" (see Attachment 8B, National Library of Medicine subject heading D001671) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 8A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

92. As of September 25, 1973, the MARC record attached hereto as Attachment 8A was accessible through any library with access to the OCLC

bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

93. WorldCat indicates that the journal *Analytical Biochemistry* as cataloged at the National Library of Medicine is currently available from 1,518 libraries.[47] In view of the above, this issue of the journal *Analytical Biochemistry* was publicly available on or shortly after November 18, 2005, because by that date it had been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Li article (Exhibit 1029) published in the journal *Analytical Biochemistry* was publicly available on or shortly after November 18, 2005, as indicated on the publisher's website.[48]

I. PUBLICATION 9: Exhibit 1048 ("KINDE")

94. Attached hereto as Exhibit 1048 is a copy of the journal article titled "Detection and Quantification of Rare Mutations with Massively Parallel Sequencing" by Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler, and

[47] https://search.worldcat.org/title/1481077.

[48] https://www.sciencedirect.com/science/article/abs/pii/S0003269705007979?via%3Dihub.

Bert Vogelstein (hereafter "Kinde"). The Kinde article was published in Volume 108, Number 23 of the journal *Proceedings of the National Academy of Science of the United States of America (PNAS)* and was found in the National Library of Medicine (Bethesda, Maryland). The Kinde article appears beginning on page 9530 of this issue dated June 7, 2011. Exhibit 1048 is a true and correct copy of the Kinde article (pages 9530-9535) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1048 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1048 is a true and correct copy in a condition that creates no suspicion about its authenticity

95. The Kinde article is available from the publisher,[49] the National Academy of Science of the United States of America, and through the National Library of Medicine.[50]

96. Attached hereto as Attachment 9A is a true and correct copy of the MARC record for the journal *PNAS* in the National Library of Medicine online

[49] https://www.pnas.org/doi/10.1073/pnas.1105422108.

[50] https://pubmed.ncbi.nlm.nih.gov/21586637/.

catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 9A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 9A.

97. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 9A, that library code is "DNLM," which means that the MARC record for this serial was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number 1607201 on June 26, 1975, ("750626"). The "BLvl" entry in Attachment 9A is "s," which indicates that the journal *PNAS* is a serial publication. Field 310 of Attachment 9A reads "Weekly $b 2004-." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *PNAS* since January 1915 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1048 corresponds to the journal *PNAS* from the time the serial title began publication and includes the date of the Kinde article.

98. Attachment 9A includes an entry in field 060 ("W1 $b PR586A")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal

classification system). Attachment 9A further includes a descriptor term reading "Science $v periodical" (see Attachment 9B, National Library of Medicine subject heading D012586) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 9A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

99. As of June 26, 1975, the MARC record attached hereto as Attachment 9A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

100. WorldCat indicates that the journal *PNAS* as cataloged at the National Library of Medicine is currently available from 1,927 libraries.[51] In view of the above, this issue of the journal *PNAS* was publicly available on or shortly after June 17, 2011, because by that date it had been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Kinde article (Exhibit 1048) published in

[51] https://search.worldcat.org/title/1607201.

the journal *PNAS* was publicly available on or shortly after June 17, 2011, as indicated on the publisher's website.[52]

J. PUBLICATION 10: Exhibit 1055 ("BENTLEY")

101. Attached hereto as Exhibit 1055 is a copy of the journal article titled "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry" by David R. Bentley and others (hereafter "Bentley"). The Bentley article was published in Volume 456, Number 6 of the open access journal *Nature* and was found in the National Library of Medicine (Bethesda, Maryland). The Bentley article appears beginning on page 53 of this issue dated November 6, 2008. Exhibit 1055 is a true and correct copy of the Bentley article (pages 53-59 and supplemental information (55 pages) that I understand is being submitted as an exhibit in this proceeding. Specifically, the text of the article is complete; no pages are missing, and the text on each page appears to flow seamlessly from one page to the next; further, there are no visible alterations to the document. Exhibit 1055 was found within the custody of a library – a place where an authentic copy of this journal would likely be. Exhibit 1055 is a true and correct copy in a condition that creates no suspicion about its authenticity

102. The Bentley article is available from the publisher,[53] Springer Nature

[52] https://www.pnas.org/doi/10.1073/pnas.1118018109.

[53] https://www.nature.com/articles/nature07517.

Limited, and through the National Library of Medicine.[54]

103. Attached hereto as Attachment 10A is a true and correct copy of the MARC record for the open access journal *Nature* in the National Library of Medicine online catalog. The library continues to update and enhance this MARC record to meet current cataloging rules. The most recent enhancement to Attachment 10A occurred on February 27, 2025, as shown in field 005 ("20250227"). I personally identified and retrieved the MARC record that is Attachment 10A.

104. As previously noted, the library that created the record is recorded in field 040 with a unique library code. For Attachment 10A, that library code is "DNLM," which means that the MARC record for this serial was cataloged at the National Library of Medicine. As can be seen in field 008 ("Date entered on file") in the MARC record for this exhibit, a cataloger at the National Library of Medicine created OCLC record number 1586310 on April 1, 1974, ("740401"). The "BLvl" entry in Attachment 10A is "s," which indicates that the journal *Nature* is a serial publication. Field 310 of Attachment 10A reads "Weekly." As noted in the holdings information (362 field), the National Library of Medicine has received the journal *Nature* since November 1869 and continues to receive the publication. Accordingly, the MARC record for Exhibit 1055 corresponds to the journal *Nature* from the time

[54] https://pubmed.ncbi.nlm.nih.gov/18987734/.

the serial title began publication and includes the date of the Bentley article.

105. Attachment 10A includes an entry in field 060 ("W1 $b NA81")—as described above, a subject matter classification number consistent with the National Library of Medicine classification system (analogous to the Dewey Decimal classification system). Attachment 10A further includes a descriptor term reading "Science $v periodical" (see Attachment 10B, National Library of Medicine subject heading D012586) in the 650 field. Thus, as of its cataloging, the publication corresponding to the MARC record attached hereto as Attachment 10A was indexed according to its subject matter by virtue of the title (field 245), the classification notation (field 060), and the subject heading (field 650).

106. As of April 1, 1974, the MARC record attached hereto as Attachment 10A was accessible through any library with access to the OCLC bibliographic database or the online catalog at a library that subscribed to the journal, which means that the corresponding publication was publicly available on or before that same date through any library with access to the OCLC bibliographic database or through an individual library.

107. WorldCat indicates that the open access journal *Nature* as cataloged at the National Library of Medicine is currently available from 3,095 libraries.[55] In

[55] https://search.worldcat.org/title/1586310.

view of the above, this issue of the open access journal *Nature* was publicly available on or shortly after November 6, 2008, because by that date it had been received, cataloged, and indexed at the National Library of Medicine and made part of the OCLC bibliographic database. For these reasons, it is my opinion that the Bentley article (Exhibit 1055) published in the open access journal *Nature* was publicly available on or shortly after November 6, 2008, as indicated on the publisher's website.[56]

VI. SUMMARY OF OPINIONS

108. In view of the foregoing, it is my opinion that the publications described above were publicly available no later than the corresponding date listed in the table below:

Exhibit	Publication	Publicly Available On or Shortly After
1006	Forshew, Tim, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana W. Y. Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M. Piskorz, Mercedes Jimenez-Linan, David Bentley, James Hadfield, Andrew P. May, Carlos Caldas, James D. Brenton, and Nitzan Rosenfeld. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA." *Science Translational Medicine*, vol. 4, issue 136 (30 May 2012): ra68.	May 30, 2012

[56] https://www.nature.com/articles/nature07517.

Exhibit	Publication	Publicly Available On or Shortly After
1012	Shiroguchi, Katsuyuki, Tony Z. Jia, Peter A. Sims, and X. Sunney Xie. "Digital RNA Sequencing Minimizes Sequence-Dependent Bias and Amplification Noise with Optimized Single-Molecule Barcodes." *Proceedings of the National Academy of Science of the United States of America (PNAS)*, vol. 109, no. 4 (24 January 2012): 1347–1352.	January 9, 2012
1014	Schwarzenbach, Heidi, Dave S. B. Hoon, and Klaus Pantel. "Cell-Free Nucleic Acids as Biomarkers in Cancer Patients." *Nature Reviews Cancer*, vol. 11 (June 2011): 426-437.	May 12, 2011
1015	National Institute of Health. National Human Genome Research Institute. *Fact Sheet: Human Genomic Variation.* Bethesda, MD: National Institute of Health, 2023.	February 1, 2023
1017	Metzker, Michael L. "Sequencing Technologies -- the Next Generation." *Nature Reviews. Genetics*, vol. 11, no.1 (January 2010): 31-46.	December 8, 2009
1019	Metzger, Michael L., and Thomas Caskey. "Polymerase Chain Reaction." *Encyclopedia of Medical Devices and Instrumentation*, 2nd ed. (vol. 5, pp. 380-387). 6 vols. Hoboken, NJ: Wiley & Sons, 2006.	October 4, 2005
1020	Mamanova, Lira, Alison J. Coffey, Carol E. Scott, Iwanka Kozarewa, Emily H. Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J. Turner. "Target-Enrichment Strategies for Next Generation Sequencing." *Nature Methods*, vol. 7, no. 2 (February 2010): 111-118 & supplemental figures and text (44 pages)	January 28, 2010
1029	Li, Tian W., and Kevin M. Weeks. "Structure-Independent and Quantitative Ligation of Single-Stranded DNA."	November 18, 2005

Exhibit	Publication	Publicly Available On or Shortly After
	Analytical Biochemistry, vol. 349, issue 2 (15 February 2006): 242-246.	
1048	Kinde, Isaac, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler, and Bert Vogelstein. "Detection and Quantification of Rare Mutations with Massively Parallel Sequencing." *Proceedings of the National Academy of Science of the United States of America (PNAS)*, vol. 108, no. 23 (7 June 2011): 9530--9535.	June 17, 2011
1055	Bentley, David R., *et al.* "Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry." *Nature*, vol. 456, no. 6 (6 November 2008): 53-59 & supplemental material (55 pages)	November 6, 2008

109. In signing this Declaration, I recognize that the Declaration will be filed as evidence in a case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I also recognize that I may be subject to cross-examination in the case and that cross-examination will take place within the United States. If cross-examination is required of me, I will appear for cross-examination within the United States during the time allotted for cross-examination.

110. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

DATED: August 15, 2025

By: Sylvia D. Hall-Ellis
Sylvia D. Hall-Ellis, Ph.D.

Exhibit A

CURRICULUM VITAE SYLVIA D. HALL-ELLIS

EDUCATION

Ph.D., University of Pittsburgh, Pittsburgh, Pennsylvania, 1985
M.P.S., University of Denver, Denver, Colorado, 2014
Post Graduate Studies, University of Texas – San Antonio, Texas, 1975-1976
M.L.S., University of North Texas, Denton, Texas, 1972
B.A., Rockford University, Rockford, Illinois, 1971

PROFESSIONAL EXPERIENCE

1981- **Consultant for higher education, non-profit organizations, and corporations.**

2002- **Adjunct Professor, School of Information, San José State University, San José, California.** Serve as part-time faculty member teaching graduate students in technical services (cataloging, bibliographic control, classification), "core courses," and special topics.

2014-2016 **Director, Grants and Resource Development, Colorado Community College System.** Provided leadership and vision to foster the continued growth of rigorous scholarship, innovative projects, and creative work for statewide system, 13 campuses, and 50 teaching sites serving 155,000 students. Responsible for leadership and ensured efficient functioning of contract and grants in compliance with state & federal requirements and successful implementation and management. Served as a subject matter expert and liaison for college Grant Directors for all issues relating to grants and subcontracts.

2010-2014 **Senior Grant Administrator, Morgridge College of Education, University of Denver (Colorado).** Provided leadership and vision to foster continued growth of rigorous scholarship, innovative research, and creative work in the Morgridge College of Education. Ensure that contract and grants processes function effectively and efficiently for 60 faculty and researchers with a focus on the successful progression and efficient management of grants totaling $13M. Worked effectively and collegially with Department Chairs and Program Coordinators on operational grant-related management activities and with a broad range of internal and external constituencies. Supported the dissemination and promotion of faculty research and scholarship to outside constituents at conferences and through publications. Assisted Principal Investigators and grant project teams by coaching, mentoring, and financial management.

2011-2013 **Interim Director & Assistant Dean, Westminster Law Library, Sturm College of Law, University of Denver.** Planned, organized, and directed all administrative activities for the library serving students, faculty, and alumni; oversaw the employment, retention, promotion, transfer and termination of library personnel; represented the library at professional conferences and public meetings; created and promoted a climate and culture of acceptance for new programs and services, a positive high-quality image of the law library, and that reflect the organization's values, encourage excellent performance, and reward high productivity and innovation; provided leadership and set strategic direction of the organization; ensured that the library provided excellent customer service through solution-oriented staff response to patron needs and by responsiveness and continuous improvement of the organization; promoted, developed, and maintained positive working relationships with colleagues and customers including key stakeholders and groups, higher education institutions, the legal community, other regional libraries and districts statewide, and national library organizations.

2007-2014 **Associate Professor, Library & Information Science, Morgridge College of Education, University of Denver (Colorado).** Served in leadership role and worked collaboratively in program, college, campus and community environments. Advised and supervised students, taught core and specialized courses at the graduate level in an integrative, student-centered learning environment. Served on LIS, College, and University committees, and maintained working relationships with colleagues in other academic units and information professionals in the Rocky Mountain region and beyond. Served on and chair doctoral student dissertation committees. Oversaw and facilitated the College and LIS graduate student association.

2002-2007 **Assistant Professor, Library & Information Science, College of Education, University of Denver (Colorado).** Served as tenure-track faculty member teaching graduate students in "core courses," resource description and access, service learning, and independent studies. Advised graduate students, participate on LIS and College committees, and serve on doctoral student dissertation committees. Oversaw and facilitated the LIS graduate student association and alumni association.

2000-2002 **Affiliate Faculty, Library & Information Science, College of Education, University of Denver (Colorado).** Served as part-time faculty member teaching graduate students in technical services (cataloging, bibliographic control, classification), "core courses," and special topics. Oversaw and facilitated the LIS graduate student association and alumni association.

2000-2001 **Special Assistant to the Secretary's Regional Representative, U.S. Department of Education, Region VIII, Denver, Colorado.** Served as the principal advisor and representative of the U. S. Secretary of Education's Regional Representative (SRR). Ensured the implementation of major goals of the SRR and the Secretary. Provided leadership on behalf of the SRR in contacts with high-level officials in Region VIII requiring sensitive policy interpretation in communication with senior Department officials to solve problems and resolve issues raised by State and local education officials. Served as the primary contact for School-to-Work/Career, Children's Health Insurance Program, and Safe and Drug-Free Schools. Delivered technical assistance to local education agencies and institutions of higher education in technology, professional development, and school construction.

1999-2000 **Catalog Librarian, Jefferson County Public Library, Lakewood, Colorado.** Performed original, copy cataloging and classification of library materials (English and Spanish) using standard library protocols; completed original descriptive cataloging and subject analysis; enhanced brief catalog and authority records in III.

1997-1999 **Development Officer, McREL International, Aurora, Colorado.** Served as senior member of corporate management team in strategic planning, development of proposals and contracts, implementation, and evaluation of new services, products, and programs for educational agencies. Provided creative leadership to corporate committees to solicit ideas, identify goals and objectives, plan, develop, present, and evaluate professional development opportunities.

1995-1997 **Education Specialist, Education Service Center, Region One, Edinburg, Texas.** Served as member of Administrative Cabinet team in strategic planning, development of proposals and contracts, implementation, and evaluation of telecommunications capabilities, services, products, and programs for 40 school districts serving 283,000 students in 7 counties. Provided creative leadership to regional and state committees to solicit ideas, identify strategic goals and objectives, plan, develop, present, and evaluate funding opportunities and professional development for 400 librarians.

1993-1996 **Assistant Professor of Library Science, Sam Houston State University, Huntsville, Texas.** Served a faculty member teaching 400 graduate students in technical services (cataloging, bibliographic control, classification), automation, and networking. Participated in distance education program and coordinated annual conference. Conducted university and Texas Library Association-funded field research focused on library collection development and academic achievement.

1992-1993 **Head Librarian, Rocky Mountain College of Art & Design, Denver, Colorado.** Responsible for the daily operation, selection and acquisition of materials, formulation of policies for library operations, media center, and photography/slides archives. Designed and implemented library automation and delivery of electronic resources to college community.

1981-1985 **Development Officer, PRLC, Inc., Pittsburgh, Pennsylvania.** Served as senior member of corporate management team in strategic planning, development of proposals and contracts, implementation, and evaluation of new services, products, and programs for 100 institutional member organizations. Coordinated the development of proposals and contracts totaling $4,000,000 annually. Provided creative leadership to corporate committees to solicit ideas, identify goals and objectives, plan, develop, present, and evaluate professional development opportunities.

1981 **Director of Library Development, Pennsylvania Department of Education, Harrisburg, Pennsylvania.** Responsible for statewide development, technical assistance, professional development, resource sharing, children's services, institutional library services, networking, and state aid program for all libraries throughout the Commonwealth. Functioned as liaison to Governor's Advisory Council, LSCA Advisory Council, District Administrators, private colleges, universities, consortia managers, and network directors. Supervised $14,000,000 formula-based state aid program and $3,000,000 grant awards to individual libraries, consortia, and networks.

1978-1981 **Assistant Director, Southern Tier Library System, Corning, New York.** Coordinated operation of system-wide programs (technical assistance, professional development, resource sharing, technical services, outreach) to 40 public libraries in 5 counties serving 500,000 residents. Solicited ideas, identified goals, sponsored, and evaluated professional development opportunities and technical assistance sessions.

1976-1978 **Division Librarian for Technical Services, Corpus Christi Public Libraries, Corpus Christi, Texas.** Provided leadership in acquisitions, cataloging, serials control, and processing for main library and 4 branches serving 250,000 residents. Participated as senior member of library management team. Compiled and prepared technical evaluations, reports, and statistical analyses of Division operations to measure the achievement and cost of annual goals, objectives, and staff performance.

1975-1976 **System Coordinator, San Antonio Major Resource Center, San Antonio, Texas.** Served as senior member of the management team for District X Office, charged to provide technical assistance, resource sharing, media services, and professional development to librarians and staff representing 30 public library jurisdictions in 21 counties serving 1,500,000 residents. Functioned as liaison to System Director, staff, and members of governing bodies with the System Board of Directors and the Texas State Library and Historical Commission. Prepared LSCA grant applications and monitored awards totaling $1,100,000 annually.

1973-1975 **Bilingual Branch Librarian, San Antonio Public Library, San Antonio, Texas.** Worked as librarian providing reference, information, and readers' advisory services in branch serving 50,000 Spanish-speaking residents in southwest San Antonio. Participated in collection development and resource acquisition activities, specializing in children's work, Spanish language resources, and multicultural studies.

1972-1973 **Librarian, Holding Institute, Laredo, Texas.** Worked as high school librarian serving 500 boarding students in Spanish-speaking environment of private school. Provided reference, research assistance, and library instruction to students and 35 faculty members.

1966-1971 **Rockford Public Library, Rockford, Illinois.** Worked in branches as part-time as a Library Assistant, Clerk, and Page in city library serving 150,000 residents.

PUBLICATIONS

Editor-reviewed Monographs (Completed and in Progress)

Hall-Ellis, Sylvia D., and Mary Bolin. *Contemporary Cataloging in an RDA Environment: A Handbook for Students and Practitioners.* Chicago, IL: American Library Association. Under contract & In development.

RDA Testing: Lessons Learned and Challenges Revealed. Sylvia D. Hall-Ellis and Robert O. Ellett, Jr., eds. Binghamton, N.Y.: Haworth, 2012. 128 p.

Hall-Ellis, Sylvia D., Stacey L. Bowers, Christopher D. Hudson, and M. Claire Williamson. *Librarian's Handbook for Seeking, Writing, and Managing Grants.* Santa Barbara, Calif.: Libraries Unlimited, 2011. 315 p.

Hall-Ellis, Sylvia D., with Ann Jerabek, and Merrie W. Valliant. *Contemporary Cataloging: A Handbook for Practitioners and Students.* Open access text. Athens, GA: University of Georgia System Regents, 2011. 767 p.

Grealy, Deborah S. and Sylvia D. Hall-Ellis. *From Research to Practice: The Scholarship of Teaching and Learning in LIS Education.* Westport, Conn.: Libraries Unlimited, 2009. 175 p.

Hall-Ellis, Sylvia D. with J. Ann Jerabek. *Grants for School Libraries.* Westport, Conn.: Libraries Unlimited, 2003. 197 p.

Hall-Ellis, Sylvia D., Doris Meyer, Frank W. Hoffmann, with J. Ann Jerabek. *Grant Writing for Small Libraries and School Library Media Centers.* Boulder, Colo.: NetLibrary, 2001. 173 p.

Hall-Ellis, Sylvia D., Doris Meyer, Frank W. Hoffmann, with J. Ann Jerabek. *Grant Writing for Small Libraries and School Library Media Centers.* Englewood, Colo: Libraries Unlimited, 2000. 173 p.

Editor-reviewed Chapters (Completed and In Progress)

Hall-Ellis, Sylvia D. "Prepared to Lead: Talent, Skills, and Competencies." In *Telling the Technical Services Story*. Chicago, IL: American Library Association, 2020. Program Presentation for Re-Accreditation by the American Library Association, School of Information, San Jose State University, 2022.

Hall-Ellis, Sylvia D. "Grant Writing and Sponsored Research Funding for Academic Librarians." In *The New Librarianship.* Vol. 4. Bradford Lee Eden, ed. New York: Scarecrow Press, 2015. (pp. 163-174)

Hall-Ellis, Sylvia D. "Organizing Information: Technical Services." In *Information Services Today: An Introduction.* Sandra Hirsch, ed. Lantham, Md.: Rowman and Littlefield, 2015. (pp. 139-148)

Hall-Ellis, Sylvia D. "Metadata, MARC, and More." In *Rethinking Technical Services, Considering Our Profession and Ourselves: What's the Future of Our Profession?* Mary Beth Weber, ed. Lantham, Md.: Rowman and Littlefield, 2015. (pp. 29-55)

Hall-Ellis, Sylvia D., ed. "Contingent Faculty: Non-Tenure Track Faculty Series." In the *Faculty Personnel Guidelines Relating to Appointment, Promotion, and Tenure.* November 2011. Denver, Colo.: University of Denver, 2011. 42 p.

Hall-Ellis, Sylvia D., ed. "Standard VII: Information Resources." In the *Sturm College of Law Self-Study Presentation for Accreditation by the American Bar Association.* Denver, Colo.: University of Denver, Sturm College of Law, 2011. 20 p.

Hall-Ellis, Sylvia D. "Applying for Grants from Foundations, Corporations, or Government." In *The Volunteers' Guide to Fundraising: Raise Money for Your School Team, Library or Community Group.* 1st ed. Ilona M. Bray, ed. Berkeley, Calif.: Nolo, 2011. (pp. 1-38 on accompanying disc)

Hall-Ellis, Sylvia D., ed. "Standard III: The Faculty." In the *Library and Information Science Program Self-Study Document for Accreditation by the American Library Association.* Denver, Colo.: University of Denver, Morgridge College of Education, 2010. 22 p.

Hall-Ellis, Sylvia D. "Library and Information Science Programs and Education for Catalogers and Metadata Specialists: Challenges for the Twenty-first Century." In *Conversations with Catalogers in the Twenty-First Century.* Elaine R. Sanchez, ed. Santa Barbara, Calif.: ABC-Clio, 2010. (pp. 226-254)

Hall-Ellis, Sylvia D., ed. "Standard III: The Faculty." In the *Library and Information Science Program Self-Study Document for Accreditation by the American Library Association.* Denver, Colo.: University of Denver, College of Education, 2003. 15 p.

Lesesne, Teri S. and Sylvia D. Hall-Ellis. "The Selection, Evaluation, and Integration of Culturally Authentic Texts: A Case for Making the Online Catalog Reflect Parallel Cultures." In *Literacy: Traditional, Cultural, Technological.* Pittsburgh, Pa.: International Association of School Librarianship, 1995. (pp. 110-113)

Lesesne, Teri S. and Sylvia D. Hall-Ellis. *The Selection, Evaluation, and Integration of Culturally Authentic Texts: A Case for Making the Online Catalog Reflect Parallel Cultures.* In *Conference Proceedings* of the 23rd Annual International Association of School Librarianship, Pittsburgh, Pennsylvania, July 17-22, 1994. ERIC Document ED374816. 17 p.

Hall-Ellis, Sylvia D. "Curriculum Folio for School Library Media Specialist Programs." In the *National Council for Accreditation of Teacher Education Self-Study for Sam Houston State University.* Huntsville, Tex.: Sam Houston State University, College of Education and Applied Science, 1994. 25 p.

Bruntjen, Scott and Sylvia D. Hall. "Attempting to Automate: Lessons Learned Over Five Years." In *Advances in Library Administration*. Volume 4. Weston, Conn.: JAI Press, 1985. (pp. 177-192)

Peer-Reviewed Journal Articles (Completed and In Progress)

Hall-Ellis-Sylvia D. "Job Design for Cataloging and Metadata Librarians." Submitted to *Journal of Library Administration.* In progress.

Hall-Ellis, Sylvia D. "The Relationship of Situational Leadership and the Dreyfus Model of Skill Acquisition for Supervisors in Cataloging and Metadata Services." To be submitted to *Library Resources and Technical Services.* In progress.

Hall-Ellis, Sylvia D. *The Relationship of Core Competencies on Learning Outcomes and Employers' Expectations for Catalog Librarians and Metadata Specialists.* To be submitted to the *Journal of Education for Library and Information Science.* In progress.

Hall-Ellis, Sylvia D. "Building Cataloger Competencies: The Dreyfus Model as a Prototype for the Education and Professional Development of Catalog Librarians and Metadata Specialists in Bibliographic Control." To be submitted to the *Cataloging & Classification Quarterly.* In progress.

Hall-Ellis, Sylvia D. "Descriptive Impressions of Entry-Level Cataloger and Metadata Specialists Positions, 2000-2016: Reflections and Trends." To be submitted to *Library Resources and Technical Services.* In progress.

Hall-Ellis, Sylvia D. "Competencies for Metadata and Cataloging Leaders: What Employers Expect as Reflected in Position Descriptions, 2000-2016." *Cataloging & Classification Quarterly.* In progress.

Hall-Ellis, Sylvia D. "Stackable Micro-credentials – A Framework for the Future." *The Bottom Line 29*, no. 4 (April 2016), http://www.emeraldinsight.com/doi/pdfplus/10.1108/BL-02-2016-0006
Ellis, Sylvia D. "Succession Planning and Staff Development – A Winning Combination." *The Bottom Line*, 28 no. 3 (May 2015): 95-98.

Hall-Ellis, Sylvia D. "Succession Planning and Staff Development – A Winning Combination." *The Informed Librarian Online* (September 2015), http://www.informedlibrarian.com/featuredArticle.cfm?FILE=succession_1509.pdf

Hall-Ellis, Sylvia D. "Nudges and Decision Making: A Winning Combination." *The Bottom Line 28*, no. 4 (July 2015): 133-136.

Hall-Ellis, Sylvia D. "Succession Planning and Staff Development – A Winning Combination." *The Bottom Line 28*, no. 3 (May 2015): 95-98.

Hall-Ellis, Sylvia D. "Metadata Competencies for Entry-Level Positions: What Employers Expect as Reflected in Position Descriptions, 2000-2013." *Journal of Library Metadata 15*, no. 2 (June 2015): 102-134.

Hall-Ellis, Sylvia D. "Onboarding to Improve Library Retention and Productivity." *The Bottom Line 27*, no. 4 (October 2014).

Hall-Ellis, Sylvia D. "Accept, Coach, and Inspire: A Formula for Success." *The Bottom Line 27*, no. 3 (July 2014): 103-106.

Hall-Ellis, Sylvia D. *Reward Systems for Staff in Higher Education Institutions: Case Studies of 15 Higher Education Institutions in Colorado.* Capstone Thesis. Denver, Colo.: University of Denver, University College, 2014.

Hall-Ellis, Sylvia D. "Reward Systems Promote High-Performance Work Teams Achieving Library Mission." *The Bottom Line 27*, no. 2 (April 2014): 66-69.

Hall-Ellis, Sylvia D. "Investment in Staff Development for Seeking External Grant Funds." *The Bottom Line 27*, no. 1 (January 2014): 22-25.

Hall-Ellis, Sylvia D. "Staff Investments for High Performance Teams." *The Bottom Line 26*, no. 4 (October 2013): 149-152.

Hall-Ellis, Sylvia D. and Deborah S. Grealy. "Service Learning in LIS Education: A Case for Holistic Assessment." *Catholic Library World* 82, no. 3 (March 2012): 178-187.

Hall-Ellis, Sylvia D., and Deborah S. Grealy. "The Dreyfus Model of Skill Acquisition: A Career Development Framework for Succession Planning and Management in Academic Libraries." *College and Research Libraries* (June 2012). http://crl.acrl.org/content/early/2012/06/19/crl12-349.full.pdf+html

Hall-Ellis, Sylvia D. "Cataloger Competencies…What Do Employers Require?" *Cataloging & Classification Quarterly* 46, no. 3 (2008): 305-330. http://ccq.haworthpress.com/doi:10.1080/01639370802034565

Hall-Ellis, Sylvia D. "Subject Access for Readers' Advisory Services: Their Impact on Contemporary Spanish Fiction in Selected Public Library Collections." *Public Library Quarterly* 27, no.1 (2007): 1-18. http://plq.haworthpress.com/doi:10.1080/01616840802122377

Hall-Ellis, Sylvia D. "Language Proficiencies among Catalogers and Technical Services Librarians." *Technical Services Quarterly* 25, no. 2 (2007) 31-47. http://tsq.haworthpress.com/doi:10.1300/J124v25n02_03

Hall-Ellis, Sylvia D. "Cataloging Electronic Resources & Metadata: Employers' Expectations as Reflected in *American Libraries and AutoCAT,* 2000-2005." *Journal of Education for Library and Information Science* 47, no. 1 (2006): 38-51.

Hall-Ellis, Sylvia D. "Descriptive Impressions of Managerial & Supervisory Cataloger Positions as Reflected in *American Libraries, AutoCAT,* and the *Colorado State Library Jobline,* 2000-2004: A Content Analysis of Education, Competencies, and Experience." *Cataloging & Classification Quarterly* 42, no. 1 (2006): 55-88.

Hall-Ellis, Sylvia D. "Descriptive Cataloging Proficiencies among Beginning Students: A Comparison among Traditional and Virtual Class Students." *Journal of Library & Information Services in Distance Education* 2, no. 2 (2005): 13-44.

Hall-Ellis, Sylvia D. "Descriptive Impressions of Entry-Level Cataloger Positions as Reflected in *American Libraries, AutoCAT,* and the *Colorado State Library Jobline,* 2000-2003." *Cataloging & Classification Quarterly* 40, no. 2 (2005): 33-72.

Hall-Ellis, Sylvia D. and Mary Ann Berry. "The Impact of Library Resource Centers on Academic Achievement in Selected Public Schools in South Texas: Preliminary Findings." *Texas Library Journal* 71, no. 2 (Summer 1995): 94-97.

Hall-Ellis, Sylvia D. "Visual Arts as Foundation for Successful Library Automation: The Rocky Mountain College of Art & Design Experience." *Art Reference Serials Quarterly* 4 (1994): 43-54.

Hall, Sylvia D. *High School Databases: An Examination of Bibliographic Data Elements*. Dissertation. Pittsburgh, Pa.: University of Pittsburgh, School of Library and Information Science, 1985.

Technical Reports

Hall-Ellis, Sylvia D. *Preparing Catalogers and Metadata Specialists Through Competency-Based Education: The San José State University iSchool Experience.* Report for the 2021 Re-Accreditation Program Presentation to the American Library Association. San José, Calif.: San José State University iSchool, 2020.

Hall-Ellis, Sylvia D. *Colorado Community College System Grant Handbook.* Denver, Colo.: Colorado Community College System, 2015.

.

Faculty Voices: Testimonials from the Frontiers of a Development Education Redesign. Elaine DeLott Baker, Marilyn Smith, and Sylvia D. Hall-Ellis, eds. Denver, Colo.: Colorado Community College System, 2014.

Lewis, Melanie, Jamie M. Van Leeuwen, and Sylvia D. Hall-Ellis. *Project Homeless Connect 7 Event Evaluation.* Denver, Colo.: University of Colorado Denver, 2009.

Hall-Ellis, Sylvia D. *Grant Writing Tips and Resources.* Denver, Colo.: University of Denver, 2008-2014. http://portfolio.du.edu/grants

Hall-Ellis, Sylvia D. and Duan Zhang. *Project Homeless Connect 6 Event Evaluation.* Denver, Colo.: Center for Community Engagement and Service Learning, 2008.

Hall-Ellis, Sylvia D. *Strategic Plan for Denver Medical Library System: Final Report.* Denver, Colo.: HealthONE and Denver Medical Library, 2008.

Hall-Ellis, Sylvia D. and Duan Zhang. *Project Homeless Connect 5 Event Evaluation.* Denver, Colo.: Center for Community Engagement and Service Learning, 2008.

Hall-Ellis, Sylvia D. *Strategic Plan for Denver Medical Library System: Phase Two Report.* Denver, Colo.: HealthONE and Denver Medical Library, 2007.

Hall-Ellis, Sylvia D. and Duan Zhang. *Project Homeless Connect 4 Event Evaluation.* Denver, Colo.: Center for Community Engagement and Service Learning, 2007.

Hall-Ellis, Sylvia D. *Strategic Plan for Denver Medical Library System: Phase One Report.* Denver, Colo.: HealthONE and Denver Medical Library, 2007.

Hall-Ellis, Sylvia D. *Denver Public Library: Opportunities for Change.* Denver, Colo.: University of Denver, Colorado Community-Based Research Network & Metropolitan Organization of People, 2005.

Hall-Ellis, Sylvia D. *Resources in Grant Writing for Colorado Schools and Libraries.* Sponsored by the Colorado Department of Education, Educational Telecommunications Unit, 1999. Revised September 2000.

Hall-Ellis, Sylvia D. *Lewis & Clark Fellows Concept Paper for the Bicentennial Commemoration.* Vancouver, Wash.: Lewis and Clark Bicentennial Council, 1998.

Hall-Ellis, Sylvia D. and Mary Ann Berry. *The Impact of Library Resource Centers on Academic Achievement in Public Schools in Texas: The Final Report.* Prepared under the sponsorship of the Research Enhancement Fund award from Sam Houston State University, the Texas Association of School Librarians, the Children's Services Round Table, and the Young Adults' Services Round Table. Huntsville, Tex.: Sam Houston State University, College of Education and Applied Science, 1997.

Hall-Ellis, Sylvia D. *The Impact of Library Resource Centers on Academic Achievement in Selected Public Schools in Texas.* Huntsville, Tex.: Sam Houston State University, College of Education and Applied Science, 1997.

Hall-Ellis, Sylvia D. *Planning for Educational Technology: the Distance Education Implementation Phase: An Investigation of Technology and Telecommunication Needs for Region One Education Service Center, South Texas Community College, TCI Cablevision of South Texas, Inc., Tech Prep of the Rio Grande Valley, Inc., Texas State Technical College, Centro Cultural en el Valle de Texas, McAllen Independent School District, Harlingen Consolidated Independent School District, Donna Independent School District, Edinburg Consolidated Independent School District, La Joya Independent School District, Mercedes Independent School District, Pharr-San Juan-Alamo Independent School District, Sharyland Independent School District, Easter Seal Society of the Rio Grande Valley, Inc., and McAllen Memorial Library.* Harlingen, Tex.: Tech Prep of the Rio Grande Valley, 1996.

Hall-Ellis, Sylvia D. "Barcodes for the Texas Library Connection." Prepared under the sponsorship of the Office of Technology Services, Texas Education Agency. Austin, Tex.: TENET News and Conferences, March 1995.

Hall-Ellis, Sylvia D. *[et. al.] Planning for Educational Technology: A Study of Future Technology Needs for the Donna Independent School District, Region One Education Region Service Center, Tech Prep of the Rio Grande Valley, Inc. South Texas Community College, and the Department of Library Science, Sam Houston State University.* Research team: Sylvia D. Hall-Ellis, Ruth Ann Riggins, Noe Torres, Patricia G. Bubb, and Michael D. Bell. Harlingen, Tex.: Tech Prep of the Rio Grande Valley, 1994. Funded through a grant award from the Texas Education Agency.

Hall-Ellis, Sylvia D. and Mary Ann Berry. *The Impact of Library Resource Centers on Academic Achievement in Selected Public Schools in South Texas.* Huntsville, Tex.: Sam Houston State University, College of Education and Applied Science, 1993. Funded through a grant award from the Sam Houston State University Research Enhancement Fund.

Hall-Ellis, Sylvia D., Alice Krueger, and Linda Brannan. *Thesaurus Development for National Curriculum Materials Database at the Eisenhower National Clearinghouse for Mathematics and Science Education, Columbus, Ohio.* Developed under the sponsorship of Mid-Continent Regional Educational Laboratory Corporate Library, Aurora, Colo. Washington, D.C.: National Science Foundation; Columbus, Ohio: ERIC Clearinghouse for Mathematics and Science Education, 1993.

Hall-Ellis, Sylvia D. and Clarence L. Coffindaffer. *Partners in Progress: The 1992 Montana Library Federation Study Final Report*. Helena, Mont: Montana State Library and The Blue Bear Group, Inc., 1992.

Hall-Ellis, Sylvia D. and Clarence L. Coffindaffer. *Montana Federation Study Identification Paper*. Denver, Colo.: The Blue Bear Group, Inc., 1992.

Bruntjen, Scott and Sylvia D. Hall. *Recommendations on the Development of System-Wide Efforts of the Northern Waters Library Service.* Ashland, Wisc.: Northern Waters Library Service and The Blue Bear Group, Inc., 1989.

Hall, Sylvia D. *The Iowa Union List of Serials as A Part of The Iowa Locator: Planning Document*. Des Moines, Ia.: State Library of Iowa, 1988.

Hall, Sylvia D. *Iowa Computer-Assisted Network (ICAN Users' Manual. IBM Compatible Edition: Version 7.0.* Des Moines, Ia.: State Library of Iowa, 1988.

Hall, Sylvia D. *Iowa Computer-Assisted Network (ICAN Users' Manual. IBM Compatible Edition: Version 5.0.* Des Moines, Ia.: State Library of Iowa, 1986.

Hall, Sylvia D. *Iowa Computer-Assisted Network (ICAN Users' Manual. Apple MacIntosh Edition: Version 5.0.* Des Moines, Ia.: State Library of Iowa, 1986.

Bruntjen, Scott and Sylvia D. Hall. *The District of Columbia Public Library Planning for Retrospective Conversion of Bibliographic Records and Automation Issues for the 1990's: Final Report*. Central City, Colo.: The Blue Bear Group, Inc., 1986.

Bruntjen, Scott and Sylvia D. Hall. *The Iowa Locator: Final Report.* Central City, Colo.: The Blue Bear Group, Inc., 1986.

Bruntjen, Scott and Sylvia D. Hall. *Discovering Resource Sharing Treasures: Accessing Pennsylvania's High School Library Collections*. Harrisburg, Pa.: State Library of Pennsylvania, Bureau of School Library Media Resources, 1985.

Hall, Sylvia D. *Retrospective Conversion and the Completeness of Cataloging in Pennsylvania High School Libraries.* Pittsburgh, Pa.: PRLC, Inc., 1985.

Bruntjen, Scott and Sylvia D. Hall. *The Iowa Locator.* Des Moines, Ia.: State Library of Iowa, 1985- http://www.statelibraryofiowa.org/ld/silo/locator

Bruntjen, Scott and Sylvia D. Hall. *ICAN Analysis*. Des Moines, Ia.: State Library of Iowa, 1985.

Bruntjen, Scott and Sylvia D. Hall. *Zebras: Bar Codes, Standards, and Library Automation*. Central City, Colo.: The Blue Bear Group, Inc., 1985.

Hall, Sylvia D. and Scott Bruntjen. *Access Pennsylvania.* Harrisburg, Pa.: State Library of Pennsylvania, 1984- http://www.accesspa.org

Hall, Sylvia D. *Reconfiguration of the ITAC Cataloging Data Base to Machine Readable (MARC) Format*. Classified Document. Arlington, Va.: U. S. Army Intelligence and Threat Analysis Center, 1984.

Bruntjen, Scott and Sylvia D. Hall. *South Dakota Library Automation Plan: Final Report*. Pittsburgh, Pa.: PRLC, Inc., 1982. Reprinted by the South Dakota State Library, 1983.

Bruntjen, Scott and Sylvia D. Hall. *Altoona Automation Plan: Final Report prepared for the Altoona Area Public Library and the Altoona District Library Center*. Pittsburgh, Pa.: PRLC, 1983.

Bruntjen, Scott, Sylvia D. Hall and Paul J. Pugliese. *College Space Utilization Study*. Pittsburgh, Pa.: PRLC, Inc., 1983.

Hall, Sylvia D. *Pennsylvania Public Library Film Center Union Catalog Cataloging and Technical Services Procedures Manual*. Pittsburgh, Pa.: PRLC, Inc., 1982.

Hall, Sylvia D. *History of Library Development in Pennsylvania: A Review and Content Analysis of the Plans, 1955-1980*. Pittsburgh, Pa.: PRLC, Inc., 1982.

Hall, Sylvia D. *Newspapers on Microfilm: A Listing of Newspapers in Allegheny, Chemung, Schuyler, Steuben, and Yates Counties*. Ithaca, N.Y.: South Central Reference and Research Council, 1981.

Hall, Sylvia D. *OCLC/ILL Subsystem at Southern Tier Library System.* Albany, N.Y.: SUNY/OCLC Network, 1980.

Hall, Sylvia D. *Retrospective Conversion for Major Public Libraries in South Central New York State: A Technical Report*. 2 vols. Corning, N.Y.: Southern Tier Library System, 1979-1980.

Hall, Sylvia D. *The Librarian and the Mexican-American Patron in Texas.* Austin, Tex.: Texas State Library, 1978.

Other Publications

Hall-Ellis, Sylvia D., ed. *Library and Information Science Program Self-Study Document for Accreditation by the American Library Association.* St. Paul, MN: St. Catherine University, 2017. 255 p.

Hall-Ellis, Sylvia D., and Robert O. Ellett, Jr., guest editors. *RDA Testing: Lessons Learned and Challenges Revealed: A Special Issue of Cataloging & Classification Quarterly* 49, no. 7-8 (November 2011).

Hall-Ellis, Sylvia D. and Robert O. Ellett, Jr. *Fundamentals of Cataloging I: An Online ALCTS Course.* Chicago, Ill.: American Library Association, 2009.

Hall-Ellis, Sylvia D., Angela J. Kinney and Martha Ann Bace, editors. *Best Practices in Cataloging Education Identified at the 2007 ALCTS Pre-conference: A Special Issue of the Journal for Library and Information Science Education* 49, no. 2 (Spring 2008).

Hall-Ellis, Sylvia D. *Cataloging Education: A New Emphasis for the Library and Information Science Curriculum.* Presentation delivered to the ALCTS Pre-conference, 132nd Annual Conference, American Library Association, Washington, D.C., June 22, 2007. http://www.loc.gov/aba/professional/

Hall-Ellis, Sylvia D. "*Cataloger Competencies: Do the Employers Require What the Professors Teach?*" Presentation delivered to the ALCTS CCS Heads of Cataloging Discussion Group, 131st Annual Conference, American Library Association, New Orleans, Louisiana, June 26, 2006. http://dlist.sir.arizona.edu/

Hall-Ellis, Sylvia D., guest editor, with Deborah S. Grealy, Clara L. Sitter, and Tassos Petrou. *Library Students' Voices in Colorado: An Issue of Colorado Libraries* 31:1 (Spring 2005).

Hall-Ellis, Sylvia D., ed. *Library and Information Science Program Self-Study Document for Accreditation by the American Library Association.* Denver, Colo.: University of Denver, College of Education, 2003. 125 p. with CD-ROM.

Hall-Ellis, Sylvia D. and Deborah S. Grealy, guest editors. *Library Education in Colorado: An Issue of Colorado Libraries* 29, no. 2 (Summer 2003).

Hall-Ellis, Sylvia D., ed. *Library and Information Science Program Self-Study Document for Accreditation by the American Library Association.* Denver, Colo.: University of Denver, College of Education, 2003.

Hall-Ellis, Sylvia D. "Adaptive Furniture for the ADA-Qualified: An Investigative Study." 2001. Unpublished manuscript.

Hall-Ellis, Sylvia D. "Cataloging Electronic Serials: A Study in User Access and Article Delivery." 2001. Unpublished manuscript.

Hall-Ellis, Sylvia D. "E-books: Acquisition and Circulation Issues in Public Libraries." 2001. Unpublished manuscript.

Hall-Ellis, Sylvia D. "Authority Control and Its Importance in School Library Media Online Catalogs." 1999. Unpublished manuscript.

Hall, Sylvia D. "Cooperation Improving Service in San Antonio System." *Texas Libraries* XXXVII (Spring 1976): 27-29.

Hall, Sylvia D. "Summer is Sea Monstrously Fun!" *Texas Library Journal* L (December 1974): 224-226.

Hall, Sylvia D. "A Brief Annotated Bibliography of Non-book Materials on the Mexican American." *Texas Library Journal* XLVIII (October 1972): 213-232.

Book Reviews

ALA Book of Library Grant Money. Ann Kepler, ed. 8th ed. Chicago, IL: American Library Association, 2011. 978-0-8389-1058-0, $175.00 [*Colorado Libraries,* June 2012]

Developing a Compensation Plan for Your Library. Paula M. Singer and Laura L. Francisco, comps. 2nd ed. Chicago, IL: American Library Association, 2009. 978-0-8389-0985-0, $57.00 [*Colorado Libraries,* March 2012]

Frugal Librarian. Carol Smallwood, ed. Chicago, IL: American Library Association, 2011. 978-0-83891-075-7, $42.00 [*Colorado Libraries,* February 2012]

Dowlin, Ken. *Getting the Money: How to Succeed in Fundraising for Public and Nonprofit Libraries.* Westport, Conn.: Libraries Unlimited, 2008. 978-1-59158-597-X, $50.00 [*Colorado Libraries,* February 2009]

Lundahl, Mats. *Bebo de Cuba; Bebo Valdés y Su Mundo*, by Mats Lundahl. Traducción de Linda Oakeshott Dragó. Barcelona: RBA Libros, 2008. 978-849-86-7259-6, $33.99 [*Críticas*, November 2008]

Bonelli, Florencia. *Marlene*, by Florencia Bonelli. Buenos Aires: Aguilar, Altea, Taurus, Alfaguara, 2008. 978-987-04-0995-3, $24.95 [*Críticas*, October 2008]

Keret, Etgar. Traducción de Ana María Bejarano. *Pizzería Kamikaze y Otras Relatos.* México, D.F.: Editorial Sexto Piso, 2008. 968-5679-29-0, no price given [*Críticas*, September 2008]

Littauer, Marita, and Florence Littauer. *Enriquece tu Comunicación. Miami, Fla.: Editorial Unilit, 2008.* 978-0-7899-1521-4, $11.99 [*Críticas*, August 2008]

Martínez, Guillermo. *La Muerte Lenta de Luciana B.* New York: Rayo Planeta, 2008. 978-0-06-156551-9, $14.95 [*Críticas*, May 2008]

Matthews, Joseph R. *The Evaluation and Measurement of Library Services.* Westport, Conn.: Libraries Unlimited, 2007. 978-1-59158-532-9, $50.00 [*Colorado Libraries,* April 2008]

Gómez-Jurado, Juan. *A Masacre de Virginia Tech: Anatomía de una Mente Torturada.* Barcelona: Ediciones El Andén, 2007. 978-84-935789-4-7, €16.50 [*Críticas*, February, 2008]

Matthews, Joseph R. *Library Assessment in Higher Education.* Westport, Conn.: Libraries Unlimited, 2007. 978-1-59158-531-2, $45.00 [*Colorado Libraries,* February 2008]

Chacón, Inma. *La Princesa India: Cuando el Viento Azul.* México: Alfaguara, 2006. 970-770398-9, $19.95 [*Críticas*, June 2006]

Vargas Llosa, Mario. *Israel/Palestina: Paz o Guerra Santa.* fotogs. by Morgana Vargas Llosa. Spain/U.S.: Aguilar: Santillana, 2006. 84-03-09691-7, paper, $19.95 [*Críticas*, May 2006]

Findlay, Diane. *Digging into Dewey.* Fort Atkinson, Wisc.: Upstart Books, 2005. 978-1-932146-18-9, $16.95 [*Colorado Libraries,* May 2006]

Pequeñas Resistencias 4: Antología del Nuevo Cuento Norteamericano y Caribeño. Menéndez, Ronaldo, Ignacio Padilla & Enrique del Risco, eds. Madrid**:** Páginas de Espuma, 2005. 84-95642-59-X, paper, $36.95 [*Críticas*, April 2006]

Rueda, André. *Vengo a Salvar a España: Biografía de Un Franco Desconocido.* Madrid**:** Nowtilus (Investigación Abierta), 2005. photogs., bibliog. 84-9763-202-8, paper, $23.95 [*Críticas*, March 2006]

Becerra, Ángela. *El Penúltimo Sueño.* 1st ed. Barcelona: Planeta, 2005. 84-08-05795-2, $25.95. [*Críticas*, January 2006]

Aguilar Camín, Héctor. *La Conspiración de la Fortuna.* 1st ed. México: Planeta, 2005. 970-37-0368-2, $19.95 [*Críticas*, October 2005]

México en Sus Libros. Enrique Florescano with Pablo Mijangos, editors & compilers. 1st ed. México, D.F.: Taurus Aguilar, 2004. 968-19-0783-3, $16.95 [*Críticas*, September 2005]

Historia Económica de México. Coordinator, Enrique Semo. México: Universidad Nacional Autónoma de México, Editorial Oceano de México, 2004. 13 vols. 970-32-0805+3 (obra completa), $9.25 each [*Críticas*, August 2005]

Czarnowsky, Christyne A. and Michael H. Williams. *Managing Your Wired Workforce: A Practical Guide.* Denver, Colo.: Bradford Pub., 2003. 1-883726-79-4, $24.95 [*Colorado Libraries,* 2004]

Abraham Lincoln. Colección Grandes Biografías. Madrid: Edimat Libros, 2003. 84-8403-858-0, €4.95 [*Críticas*, 2004]

Alponte, Juan María. *Colón: el Hombre, el Navegante, la Leyenda.* México, D.F.: Aguilar, 2003. 968-19-1260-8, $16.95 [*Críticas*, 2004]

Olcese Salvatecci, Alfieri. *Cómo Estudiar con Èxito: Téchnicas yHhábitos Para Aprender Major.* México, D.F.: Alfaomega Grupo Editor, 2002. 970-15-0764-9 [*Críticas*, 2003]

Mujeres Como Islas: Antología de Narradoras Cubanas, Dominicanas, Puertoriqueñas. Edición: Olga Marta Pérez, Thelma Jiménez, Andrés Blanco D., eds. Santo Domingo, República Dominicana: Ediciones Ferilibro, 2002. 959-209-419-5 [*Críticas*, 2003]

Tibol, Raquel. *Los Murales de Diego Rivera: Universidad Autónoma Chapingo.* México, D.F.: Editorial RM, Universidad Autónoma Chapingo, 2002. 968-5208-08-5 [*Críticas*, 2003]

Lindsay-Poland, John. *Emperors in the Jungle: the Hidden History of the U.S. in Panama.* Durham, N.C.: Duke University Press, 2003. 0-8223-3098-9, $18.95 [*Críticas*, 2003]

Kaplan, Allison G. and Ann Marlowe Riedling. *Catalog It! A Guide to Cataloging School Library Materials.* Worthington, Ohio: Linworth Pub., 2002. 1-58683-014-7, $44.95 [*Colorado Libraries,* 2002]

Bardach, Ann Louise. *Cuba Confidential: Love and Vengeance in Miami and Havana.* New York: Random, 2002. 0-375-50489-3, $25.95 [*Library Journal*, 2002]

Swan, James. *Fundraising for Libraries: 25 Proven Ways to Get More Money for Your Library.* New York: Neal Schuman, 2002. 1-55570-433-6, $69.95 [*Colorado Libraries,* 2002]

Puíg de Lange, Victoria. *Sol Con Agua.* Nashville, Tenn.: Editorial Vistazo/Ediciones Reio Negro, 2002. 0-9724506-0-2, $17.75 [*Críticas*, 2002]

The Encyclopedia of Latin American Politics. Edited by Diana Kapiszewski; assistant editor, Alexander Kazan. Westport, Conn.: Oryx Press, 2002. 1-57356-306-4, $74.95 [*Library Journal*, 2002]

Vargas Lizano, Isabel. *Y Si Quieres Saber de Mi Pasado.* Con la colaboración de J.C. Vales. Madrid: Santillana Ediciones Generales, S.L., 2002. 84-03-09278-4, $18.95 [*Críticas*, 2002]

Ruy Sánchez, Alberto. *Los Jardines Secretos de Mogador: Voces de Tierra.* México, D.F.: Alfaguara, S.A., 2001. 968-19-0879-1, $16.95 [*Críticas*, 2002]

Chávez, Ricardo and Celso Santajuliana. *El Final de las Nubes.* Barcelona: RBA Libros, S.A., 2001. 84-7901-760-0, $11.95 [*Críticas*, 2002]

The Power of Language / El Poder de la Palabra: Selected Papers from the Second REFORMA National Conference. Edited by Lillian Castillo-Speed and the REFORMA National Conference Publications Committee. Englewood, Colo.: Libraries Unlimited, 2001. 1-563089459, $35.00 [*Colorado Libraries,* 2001]

Martínez, Rubén. *Crossing Over: A Mexican Family on the Migrant Trail.* Metropolitan: Holt, 2001. 0-8050-4908-8, $26.00 [*Library Journal*, 2001]

Guevara, Ernesto "Che." *The African Dream: the Diaries of the Revolutionary War in the Congo.* Translated from the Spanish by Patrick Camiller. With an introduction by Richard Gott and a foreword by Aleida Guevara March. New York: Grove Press, 2001. 0-8021-3834-9, $13.00 [*Library Journal*, 2001]

Cooper, Gail and Garry Cooper. *New Virtual Field Trips.* Englewood, Colo.: Libraries Unlimited, 2001. 1-56308-887-8, $27.50 [*Colorado Libraries,* 2001]

NATIONAL SERVICE & PROFESSIONAL AFFILIATIONS

American Library Association, Life Member

Association for Library Collections & Technical Services (now CORE)
Cataloging, Classification, and Metadata Section
Awards & Scholarships Coordination Committee, 2025-
Margaret Mann Citation Committee, 2023-
ALCTS Editorial Board, Member, 2015-2019
Committee on Cataloging: Description and Access, Member, 2008-2012
Test Site for RDA, Manager, 2009-2011
Nominating Committee, Member, 2007-2008
Committee for Education and Training of Catalogers, Chair, 2006-2007, Member, 2003-2007
Competencies & Education for a Career in Cataloging Interest Group, Chair, 2010-2012; Member, 2010-
President's Program Committee, 2015-2016
Education Committee, Member, 2005-2007
Task Force on Competencies and Education for a Career in Cataloging, Chair, 2007-2009
2007 Annual Meeting Pre-conference, "What They Don't Teach in Library School: Competencies, Education and Employer Expectations for a Career in Cataloging," Steering Committee Chair
Fundraising Committee, Member, 2004-2006
Office of Diversity
Committee on Diversity, Chair, 2010-2011
Diversity Research Advisory Committee, Member, 2009-2010
Spectrum Scholar Mentor, 2009
Diversity Grants Review Committee, Member, 2007, 2010, 2012
Office of Statistics and Research
Research and Statistics Committee, Member, 2005-2007; Intern, 2003-2005
Library Research Round Table
Nominating Committee, 2022-2023
Membership Committee, 2019-2021
Board of Directors, Member At-Large, 2009-2012
Office of Accreditation
Accreditation Panel Member (training completed 2004)
Accreditation Review Panel Member (training completed 2004)

REFORMA, Life Member

National Board of Directors, 2005-2013
National Fundraising Chair, 2005-2013
National Recruiting and Mentoring Committee, 2008-2010
Colorado Chapter, Secretary, 2004-2005
Colorado Chapter Liaison to National Board of Directors, 2004-2013

American Association of Law Libraries (AALL), Life Member Online Audiovisual Catalogers Association (OLAC) Colorado Association of Law Librarians (CoALL)

American Association of University Women Association for Library and Information Science Education (ALISE)

University of Denver ALISE Representative, 2003-2008; 2010-2011
Membership Advisory Committee, 2007-2010
Technical Services Education Special Interest Group, 2003-
Garfield Doctoral Dissertation Award Reviewer, 2012
Garfield Doctoral Dissertation Scholarship Reviewer, 2014

American Association of University Professors (AAUP), 2006-
American Grant Writers Association, 2014- Grant Professionals Institute, 2016- National Grants Management Association, 2014-
Grant Reviewer (National and Regional Team Leader), U.S. Department of Education, 1998- Grant Reviewer, Broadband Technologies Opportunity Program, U.S. Department of Commerce, 2009
Grant Reviewer, Institute of Museum and Library Services, 1998-2000
Grant Reviewer, Colorado Department of Education, 2014-
Grant Reviewer, Colorado State Library, 1998-
Grant Reviewer, American Association of Community Colleges *Working Connections* program, 1999
Peer Reviewer, *Journal for Library and Information Science Education,* 2005-
Peer Reviewer, *Journal of Library Metadata,* 2009
Peer Reviewer, *International Journal of Library and Information Science,* 2011
Regular Columnist, *The Bottom Line*, 2013-2016
Book Reviewer, *Library Journal*, 2001
Book Reviewer, *Críticas*, 2002-2009
Book Reviewer, *Colorado Libraries,* 2000-2012

REGIONAL SERVICE & PROFESSIONAL AFFILIATIONS

Mountain Plains Library Association
Professional Development Grants Committee, Member, 2005-2006
Professional Development Policy and Guidelines Sub-committee Chair, 2006

Colorado Association of Libraries
"Student Voices" Column Editor, *Colorado Libraries,* 2005-2006
Conference Planning Committee, Member, 2002
Technical Services and Automation Division, Chair, 2002-2003; Member, 2000-
Academic Libraries Division, Peer Review Conference Papers Committee, Member, 2007-2009
Education Committee, Member, 1988-1993
Diverse Populations Committee, Member, 2009-2013

SERVICE TO THE UNIVERSITY OF DENVER
Center for Teaching and Learning Faculty Advisory Board, 2007-2008; 2010-2012
Center for Community Engagement and Service Learning Advisory Board, 2007-2012
Faculty Senate
Executive Committee, 2008-2013
Nominations, Rules & Credentials Committee, Chair, 2008-2013; Member 2007-2013
Appointment, Promotion, and Tenure Revision Committee, 2010-2012
Grievance Policy Committee Member, 2007-2010
Law Library Director Search Committee Chair, 2012-2013
PROF Grant Review, College Representative to the University Review Team, 2006, 2008
Project Homeless Connect Evaluations, Principal Investigator, 2006-2009
University Technology Council, 2007-2009
University of Denver Hyde Interviews for Incoming Freshmen, 2003-2012

SERVICE TO THE MORGRIDGE COLLEGE OF EDUCATION
Appointment, Tenure & Review Committee
Member, voting, 2003-2005; 2007- 2009
Chair, Clinical Faculty Promotion & Tenure Policy Subcommittee, 2008-2009
Member, Community Engagement Subcommittee, 2007-2009
Member, Tenure Review Panel, 2003
Advancement and Alumni Relations Committee, Chair, 2003-2007; Member, 2002-2007

College Building Committee Member, 2004-2010
College of Education Student Association, Faculty Advisor, 2004-2007; 2009-2010
Faculty Senator, 2007-2013
Research and Scholarship Committee, Chair, 2008-2009; Member, 2002-2003, 2008-2009
Research and Grant Mentoring Committee, Chair, 2009-2010
Research Task Force Member, 2010-2012
Search Committee Member, Assistant Professor for Curriculum & Instruction, 2010-2011
Workload Task Force Member, 2010-2011

SERVICE TO THE LIBRARY & INFORMATION SCIENCE PROGRAM

Library and Information Science Student and Alumni Association, Faculty Advisor, 2002-2008
ALA Student Chapter, Faculty Advisor, 2005-2008
Beta Phi Mu Phi Chapter, Faculty Advisor, 2004-2014
Steering Committee Member, Accreditation by the American Library Association, 2001-2004
Search Committee Member, Associate Professor for LIS, 2006
Search Committee Chair, Assistant Professors for LIS, 2003, 2005, 2006, 2007
Search Committee Ex-Officio Member, Director for LIS, 2004

SERVICE TO SAM HOUSTON STATE UNIVERSITY

Rio Roundup: South Texas Literature Conference, Conference Coordinator, 1993
External Relations, Fund Raising, Grants Committee, Chair, 1993-1995
Advisory Council Committee, Chair, 1993-1995
Students, Admissions, and Advisement Committee, Chair, 1994
Institutional Effectiveness Committee, Member, 1993-1995

SERVICE TO THE COLLEGE OF EDUCATION AND APPLIED SCIENCE

Faculty Affairs Committee, Member, 1993-1995
Curriculum Committee, Member, 1993-1995
Continuing Education Committee, Member, 1993-1995

SERVICE TO THE COMMUNITY

French Teacher & Student Exchange with Laon, France, Monarch High School, Boulder Valley School District (Boulder, Colorado), 2017-
Denver/Boulder Games 2022, Board of Directors, Secretary-Treasurer, 2015-2018
United Way Campaign Committee, College of Education, 2006
Arapahoe County, Election Judge for the County Clerk and Recorder, 2000-2005
Arthritis Foundation, Rocky Mountain Chapter, Certified Educator & Trainer, 1999-2008
Bonfils Blood Center, Silver Level Donor, 2000-2016
Denver Museum of Natural History, Docent, 1992
Tech Prep of the Lower Rio Grande Valley, Inc. (Harlingen, Texas) Board of Directors, 1995-1997
Executive Committee, 1996-1997; Chair, Development Committee, 1995-1997; Chair, Fiscal Agency Committee, 1995-1996; Chair, Colleges and Universities Committee, 1996-1997
Gilpin County (Colorado) Public Library Board of Trustees, 1986-1989; Vice President, 1987-1989
City of Central (Colorado) Economic Development Committee, 1987-1989
Columbine Family Health Centers, Inc. (Nederland and Black Hawk, Colorado) Board of Directors, 1988-1989

CERTIFICATION

Permanent Public Librarian Certificate - Pennsylvania, New York, Texas
Westlaw Expert Witness, 2008-
Certified Grant Writer®, 2016-

AWARDS AND HONORS

Advanced Practitioner for Service Learning and Community Engagement, University of Denver, 2011
Platinum Star Alumnae, College of Information, Library Science & Technologies, University of North Texas, 2009
Commendation for Integration of Technologies in Teaching & Learning Environment, University of Denver, 2006
Outstanding Adjunct Faculty Member Award, College of Education, University of Denver, 2002
Beta Phi Mu, Pi, University of Pittsburgh, 1985
Alpha Lambda Sigma, University of North Texas, 1972
Albert Nelson Marquis Lifetime Achievement Award, Who's Who in America
Who's Who in American Women
Who's Who of Women Executives
Dictionary of International Biography
Who's Who in the East
Who's Who in the South and Southwest
Who's Who in the World
Who's Who of Online Professionals
Who's Who in Library and Information Science
2,000 Notable Women
Who's Who of Emerging Leaders
Who's Who in Professional and Executive Women
Who's Who in American Education
International Who's Who of Professional and Business Women
International Leaders in Achievement
International Educator of the Year
Who's Who in Finance and Industry
Who's Who in Finance and Business

Invited International and National Conference Presentations

Hall-Ellis, Sylvia D. *Presentation of the Robert O. Ellett Endowed Scholarship.* Presentation delivered to the School of Information at San Jose State University Alumni & Student Program, American Library Annal Conference, San Diego, Calif., June 29, 2024.

Driscoll, Margaret, Christy Confetti Higgins, Scott Brown, and Sylvia D. Hall-Ellis. *A View from Within: Open House Tour of Three Canvas Core Courses*. Presentation to be delivered at the School of Information, San José State University Professional Development Seminar, San José, California, October 17, 2019.

Hall-Ellis, Sylvia D. *Invest in Me -- I'm Your Future: Succession Planning for Libraries.* Keynote presentation delivered at the ALCTS President's Symposium, Boston, Mass., January 8, 2016.

Seidel, Kent E. and Sylvia D. Hall-Ellis. *Making Grants Work for You: Strategies for Doctoral Students and Early Career Scholars.* Presentation delivered at the University Council for Educational Administration, Early Career Scholars Session, Indianapolis, Ind., November 9, 2013.

Hall-Ellis, Sylvia D. *So You Want to be a Manager; Leadership Skills and Competencies for Technical Services Managers and Administrators.* Presentation delivered at the 138th Annual Conference, American Library Association, Chicago, Ill., June 29, 2013.

Seidel, Kent E., Karen S. Riley, Lyndsay Agans, Susan Korach, and Sylvia D. Hall-Ellis. *Making Grants Work for You (Instead of Just Working for Grants).* A panel discussion delivered at the University Council for Educational Administration, Early Career Scholars Session, Denver, Colo., November 17, 2012.

Hall-Ellis, Sylvia D. *After the Great TS Reorganization: The Westminster Law Library.* Presentation delivered at the 137th Annual Conference, American Library Association, Anaheim, Calif., June 23, 2012.

Hall-Ellis, Sylvia D. *Conversations with Catalogers in the 21st Century.* A panel discussion sponsored by the ALCTS Competencies for a Career in Cataloging Interest Group, delivered at the 137th Annual Conference, American Library Association, Anaheim, Calif., June 22, 2012.

Hall-Ellis, Sylvia D., moderator. *Mid-Career Leaders Program.* A panel discussion sponsored by the ALA Committee on Diversity delivered at the 136th Annual Conference, American Library Association, New Orleans, La., June 26, 2011.

Hall-Ellis, Sylvia D., moderator. *Diversity Town Hall.* A community conversation sponsored by the ALA Committee on Diversity delivered at the 136th Annual Conference, American Library Association, New Orleans, La., June 24, 2011.

LaBarre, Kathryn, Sylvia D. Hall-Ellis, Karen Anderson, Rick Hasenyager, Christopher Cronin, and Penny Baker. *Briefings from RDA Test Participants.* A panel discussion delivered at the Midwinter Conference, American Library Association, San Diego, Calif., January 7, 2011.

Miksa, Shawne, Marjorie Bloss, and Sylvia D. Hall-Ellis. *Educating the Next Generation of Catalogers: Teaching RDA.* A panel discussion delivered at the 97th Annual Conference, Association for Library and Information Science Education, San Diego, Calif., January 7, 2011.

Hall-Ellis, Sylvia D., Robert Maxwell, John Hostage, and George Prager. *RDA Panel: What Cataloging Managers Need to Know.* Presentation delivered at the 103rd Annual Conference, American Association of Law Librarians, Denver, Colo., July 12, 2010.

Hall-Ellis, Sylvia D. and Stacey L. Bowers. *Catalogers in the RDA Environment: Skill Sets, Expectations and Challenges.* Presentation delivered at the 103rd Annual Conference, American Association of Law Librarians, Denver, Colo., July 11, 2010.

Hall-Ellis, Sylvia D. *Comfortable in Your Cataloging and Metadata Specialist Skin? Or, So You Want to Hire a Cataloger.* Presentation delivered to the ALCTS Research Group at the 134th Annual Conference, American Library Association, Chicago, Ill., July 11, 2009.

Perez, Megan, Sylvia D. Hall-Ellis, and Denise Anthony. *From Novice to Expert: Collaboration for Succession Planning.* A "hot topic" presentation delivered at the 14th ACRL Conference, Seattle, Wash., March 13, 2009.

Hall-Ellis, Sylvia D. *Cataloging in the RDA Environment: Skill Sets, Expectations and Challenges.* Presentation delivered to the ALCTS Research and Publications Committee at the Midwinter Conference, American Library Association, Denver, Colo., January 24, 2009.

Hall-Ellis, Sylvia D. *LIS Cataloging Education for the 21st Century: Expectations and Challenges.* A panel discussion held at the 95th Annual Conference, Association for Library and Information Science Education, Denver, Colo., January 23, 2009.

Chu, Clara, Sylvia D. Hall-Ellis, and Mark Winston. *The Doctoral Degree & Building a Career.* A panel discussion delivered at the ALA Office of Diversity Spectrum Doctoral Fellows E.J. Josey Leadership Institute, Midwinter Conference, American Library Association, Denver, Colo., January 20, 2009.

Hall-Ellis, Sylvia D. and Robert O. Ellett, Jr. *Fundamentals of Cataloging Course: An Overview of the ALCTS Online Course.* Presentation delivered to the ALCTS Big Heads Group, 133rd Annual Conference, American Library Association, Anaheim, Calif., June 30, 2008.

Hall-Ellis, Sylvia D. *Employers' Expectations for Technical Services Librarians: What We Don't Know.* Presentation delivered to the ALCTS Research and Publications Committee Program, 133rd Annual Conference, American Library Association, Anaheim, Calif., June 28, 2008.

Hall-Ellis, Sylvia D., Virginia R. Maloney, and Mary Stansbury. *Institutional Responses to Engaged Scholarship: The Carnegie Foundation Engaged University Classification at Two Universities.* Presentation delivered to the 94th Annual Conference, Association for Library and Information Science Education, Philadelphia, Pa., January 11, 2008.

Hall-Ellis, Sylvia D. *Puzzles, Problems, and Predicaments.* Presentation delivered to the ALCTS Research Discussion Group, 132nd Annual Conference, American Library Association, Washington, D.C., June 23, 2007.

Hall-Ellis, Sylvia D. *Cataloging Education: A New Emphasis for the Library and Information Science Curriculum.* Presentation delivered to the ALCTS Pre-conference, 132nd Annual Conference, American Library Association, Washington, D.C., June 22, 2007. http://www.loc.gov/catdir/cpso/careercat.html

Ellett, Jr., Robert O. and Sylvia D. Hall-Ellis. *Copy Cataloging Done Smarter.* Presentation delivered to the International Conference on Interdisciplinary Information Sciences and Technologies (InSciT2006), October 25-28, 2006.

Hall-Ellis, Sylvia D. and Robert O. Ellett, Jr. *Cooperative Cataloging: Challenges and Opportunities for Defense Libraries.* Presentation delivered to the 1st Annual Conference of Defense Libraries, Spanish Ministry of Defense, Madrid, Spain, July 7, 2006.

Hall-Ellis, Sylvia D. *Cataloger Competencies: Do the Employers Require What the Professors Teach?* Presentation delivered to the ALCTS CCS Heads of Cataloging Discussion Group, 131st Annual Conference, American Library Association, New Orleans, La., June 26, 2006.

Grealy, Deborah S. and Sylvia D. Hall-Ellis. *From Research to Practice: The Scholarship of Teaching and Learning in LIS Education.* Presentation delivered to the at the 92nd Annual Conference, Association for Library and Information Science Education, San Antonio, Tex., January 18, 2006.

Hall-Ellis, Sylvia D. *Employers' Expectations for Entry-Level Catalogers: What Position Announcement Data Indicate.* Research paper delivered to the Technical Services Special Interests Group, 91st Annual Conference, Association for Library and Information Science Education, Boston, Mass., January 12, 2005. http://dlist.sir.arizona.edu/

Hall-Ellis, Sylvia D. *Common Errors in MARC Records Prepared by LIS Students: What Does It Mean?* Research paper delivered to the ALCTS CCS Cataloging Norms Discussion Group, Mid-Winter Conference, 90th American Library Association, San Diego, Calif., January 10, 2004.

Hall-Ellis, Sylvia D. *Visual Arts as Foundation for Successful Library Automation: The Rocky Mountain College of Art & Design Experience.* Paper delivered at the 6th Annual Conference of Higher Education, Charleston, S. C., 1993.

Hall, Sylvia D. *Design Elements for Bibliographic Databases: An Overview.* Paper delivered at the 14th Online National Conference, New York, 1983.

Invited Regional Conference Presentations

Hall-Ellis, Sylvia D., Hudson, Christopher D., Brittany Cronin, and Kathryn Michaels. "The Colorado Law Project: Meeting the Public's Need for Legal Information." Panel discussion delivered at the Mountain Plains Library Association Conference, Billings, Mont., April 9, 2011.

Grealy, Deborah S. and Sylvia D. Hall-Ellis. *Education for Information Professionals in New Mexico: Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the New Mexico Library Education Summit, Las Vegas, N.M., September 26, 2005.

Hall-Ellis, Sylvia D. *Public Library-School Library Partnerships.* Presentation delivered at the 3rd Annual Colorado Association of Libraries Conference with the Mountain Plains Library Association, Denver, Colo., October 22, 2004.

Hall-Ellis, Sylvia D. *Learn All You Can – Educational Partnership Opportunities for the Lewis and Clark Bicentennial Commemoration.* Paper delivered at the 3rd Annual Lewis and Clark Bicentennial Council National Planning Conference, Bismarck, N.D., April 26, 1998.

Invited State Conference Presentations

Grealy, Deborah S. and Sylvia D. Hall-Ellis. *Academic Library Leadership Changes: Using Succession Planning and Mentoring*. Presentation delivered at the Minnesota Library Education Conference, St. Cloud, Minn., October 10, 2013.

Hall-Ellis, Sylvia D., Merrie Valliant, and Melissa Powell. *RDA: What Is It and What Do You Need To Do With It At Your Library?* Presentation delivered at the Colorado Library Consortium Spring Conference, Ft. Morgan, Colo., April 26, 2013.

Hall-Ellis, Sylvia D. *Service Learning and the Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 4th LEADS Scholars Orientation, Denver, Colo., August 5, 2009.

Hall-Ellis, Sylvia D. *Service Learning: Enhancement to Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 3rd LEADS Scholars Orientation, Denver, Colo., June 2008.

Hall-Ellis, Sylvia D. *Law Librarianship: A Community Conversation.* Sponsored by the Colorado Association of Law Libraries. Paper delivered at the Colorado Supreme Court Library, Denver, Colo., May 14, 2008.

Hall-Ellis, Sylvia D. *Opportunities and Challenges in Law Librarianship: A Community Conversation.* Presentation delivered at the Sturm College of Law, Denver, Colo., November 7, 2007.

Ellis, Sylvia D. *Grant Writing Resources for Nursing Professionals.* Presentation delivered at the Presbyterian / St. Luke's Health One Medical Center 1st Annual Research Symposium, Denver, Colo., October 17, 2007.

Hall-Ellis, Sylvia D. *Project Homeless Connect 4 Event Evaluation – Insights and Lessons Learned.* Presentation delivered at the Homelessness Research Symposium: What is DU Doing about Homelessness in Denver, Denver, Colo., September 14, 2007.

Hall-Ellis, Sylvia D. *Service Learning and the Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 2nd LEADS Scholars Orientation, Denver, Colo., August 10, 2007.

Hall-Ellis, Sylvia D. *Education for Information Professionals in a Digital Environment: Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 15th Spring Mountains and Plains Parapros Conference, Denver, Colo., February 24, 2007.

Hall-Ellis, Sylvia D. *Public Library Service to Spanish-Speaking and Latino Residents in Denver: A Case Study.* Presentation delivered at the 4th Annual Colorado Association of Libraries Conference, Denver, Colo., November 10, 2005.

Hall-Ellis, Sylvia D. *Education for Information Professionals in a Digital Environment: Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 14th Annual Mountains and Plains Parapros Conference, Aurora, Colo., July 29, 2005.

Hall-Ellis, Sylvia D. *Educational Opportunities: Library & Information Science Graduate Education at the University of Denver.* Presentation delivered at the 13th Annual Mountains and Plains Parapros Conference, Centennial, Colo., August 6, 2004.

Hall-Ellis, Sylvia D. *Library Education & Training: Focus on the West: An LIS Faculty Member's Personal Response*. Presentation delivered at the 1st Annual Colorado Association of Libraries Conference, Keystone, Colo., October 18, 2002.

Hall-Ellis, Sylvia D. *Grant Writing for School Librarians.* Presentation delivered at the 2002 Annual Colorado Education Media Association Conference, Colorado Springs, Colo., February 15, 2002.

Hall-Ellis, Sylvia D. *Grants – Opportunities for the Future.* Paper delivered at the Southeast Regional Accountability Annual Conference, Lamar, Colo., November 12, 1998.

Hall-Ellis, Sylvia D. *The Texas Library Connection and Interlibrary Loan: An Experiment in Resource Sharing.* Paper delivered at the Texas Computer Educators' Association Annual Conference, Austin, Tex., February 6, 1997.

Hall-Ellis, Sylvia D. *Finding Grant Sources on the Internet: A Guide for Librarians.* Paper delivered at the 2nd Annual Institute for School Library Personnel, Pharr-San Juan-Alamo North High School, Pharr, Tex., July 29, 1996**.**

Hall-Ellis, Sylvia D. *Mathematical and Logical Thinking: A Critical Intelligence.* Paper delivered at the 3rd Annual Paraprofessional Conference at the University of Texas - Pan American, Edinburg, Tex., March 8, 1996.

Hall-Ellis, Sylvia D. *Cataloging Trends and Issues: Update Session.* Paper delivered at the 1st Annual Institute for School Library Personnel, South Texas Community College, McAllen, Tex., July 19, 1995.

Hall-Ellis, Sylvia D. *Grant Writing: Tips and Encouragement for School Librarians.* Paper delivered at the 1st Annual Institute for School Library Personnel, South Texas Community College, McAllen, Tex., July 18 and 19, 1995.

Hall-Ellis, Sylvia D. *How to Become an Expert Grant Writer.* Paper delivered at the 3rd Annual High School Principals' Academy, South Padre Island, Port Isabel, Tex., June 19, 1995.

Texas Library Study: Results from Regions I and II. Paper delivered at the 3rd Annual Technology Conference, Texas A&M University, College Station, Tex., November 18, 1994.

Hall-Ellis, Sylvia D. *Academic Achievement and Middle School Students.* Paper delivered at the 8th Annual Young Adult Conference, Sam Houston State University, Huntsville, Tex., November 5, 1994.

Hall-Ellis, Sylvia D. *Multimedia Resources for Library Leaders.* Paper delivered at the Institute for Librarians in A Multicultural Environment, Sam Houston State University, Huntsville, Tex., June 10, 1994.

Hall-Ellis, Sylvia D. *Finding the Resource: Empowering the User, or, the Case for Curriculum Based Subject Access to Learning Resource Center Collections.* Paper delivered at the 81st Annual Texas Library Association Conference, Corpus Christi, Tex., April 12-16, 1994.

Hall-Ellis, Sylvia D. and William H. Pichette. *Sam Houston State University Makes Use of OCLC/AMIGOS Collection Analysis CD.* Paper delivered at the 81st Annual Texas Library Association Conference, Corpus Christi, Tex., April 12-16, 1994.

Hall, Sylvia D. *Funding and Library Development in Pennsylvania: A Symbiotic Relationship.* Paper delivered at the Annual Graduate Student Colloquia, University of Pittsburgh, School of Library and Information Science, 1982.

Hall, Sylvia D. *Leadership for Public Library Trustees.* Paper delivered for the Trustees Division, Pennsylvania Library Association Annual Conference, Lancaster, Penn., 1981.

Seminar and Professional Development Presentations

Taylor, Meredith, and Sylvia D. Hall-Ellis. *Talent Management and Succession Planning.* ALCTS eForum, held March 22, 2017.

Hirsh, Sandra, Heather O'Brien, Michelle Holschuh Simmons, Michael Krasulski, and Sylvia D. Hall-Ellis. *Information Services Today: An Introduction. Part 3: Information Services: Roles in the Digital Age*. Rowan and Littlefield in partnership with Library Journal webinar, recorded February 5, 2015.

Hall-Ellis, Sylvia D. and Jennifer Sweda. *Copy Cataloging in an RDA Environment.* ALCTS eForum, held May 14 and 15, 2013.

Hall-Ellis, Sylvia D. *Law Librarianship: A Community Conversation.* Sponsored by the Colorado Association of Law Libraries, presented at the Colorado Supreme Court Library, Denver, Colo., May 14, 2008.

Hall-Ellis, Sylvia D. and Beatrice Z. Gerrish. *Reading and Libraries: Recent Research in Reading.* Presentation at the Ricks Center for Gifted Education, Denver, Colo., March 4, 2008.

Hall-Ellis, Sylvia D. *Cataloger Competencies: Do the Employers Require What the Professors Teach?* Presentation for the School of Library and Information Science, San José State University, February 12, 2008.

Hall-Ellis, Sylvia D. and Robert O. Ellett, Jr. *Cooperative Cataloging: Rules, Tools, and Conventions for Building a Multi-institutional Catalog.* Sponsored for the Spanish Ministry of Defense, Madrid, Spain, July 10, 2006.

Hall-Ellis, Sylvia D. *Cash for Kids: Grant Writing Opportunities for Youth Services Librarians.* Sponsored by the Colorado Young Adult Librarians; presented at Bemis Memorial Library, Littleton, Colo., November 13, 2002.

Hall-Ellis, Sylvia D. *MARC Records and Authority Control: Planning for Bibliographic Database Migration.* Sponsored and held at the Douglas County Public Library, Castle Rock, Colo., May 28, 2002.

Grant Writing: A Refresher for Librarians. Sponsored by Library and Information Science Program, College of Education, University of Denver; presented at University Center at Chaparral, August 12, 2000.

Hall-Ellis, Sylvia D. *Shaking the Money Tree – Basic Grant Writing for Colorado Educators.* Sponsored by the Office of Educational Telecommunications of the Colorado Department of Education.
Pikes Peak Community College, Colorado Springs, Colo., November 16, 1998.
Pueblo School District 60, Pueblo, Colo., November 12, 1998.

University of Northern Colorado, Greeley, Colo., November 10, 1998.
United Technology Educational Partnership, Grand Junction, Colo., November 9, 1998.

Hall-Ellis, Sylvia D. *Cataloging Multimedia, Kits, Globes and Map Materials in USMARC.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., April 24, 1997 and December 12, 1996.

Hall-Ellis, Sylvia D. *New Standards for School Library Media Centers.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., April 15, 1997.

Hall-Ellis, Sylvia D. *The School Library Media Specialist in the 21st Century: Visions for the Future.* Sponsored and hosted by Pharr-San Juan-Alamo Independent School District, Pharr, Tex., April 4, 1997.

Hall-Ellis, Sylvia D. *Cataloging Sound Recordings and Audio Materials in USMARC.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., March 20, 1997 and November 21, 1996.

Hall-Ellis, Sylvia D. *Evaluating and Selecting CD-ROMS for School Library Media Collections.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., March 18, 1997.

Hall-Ellis, Sylvia D. *Cataloging Audiovisual Materials in USMARC.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., February 27, 1997, November 7, 1996 and October 24, 1996.

Hall-Ellis, Sylvia D. *Introduction to Dialog: Basic Searching Strategies.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., February 14, 1997.

Hall-Ellis, Sylvia D. *Cataloging Books in USMARC.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., January 30, 1997, October 10, 1996, September 26, 1996 and August 8, 1996.

Hall-Ellis, Sylvia D. *Advanced Internet Searching Techniques for Librarians.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., January 21, 1997.

Hall-Ellis, Sylvia D. *Texas Library Connection Full-Text Searching.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., December 13, 1996 and October 25, 1996.

Hall-Ellis, Sylvia D. *Developing Evaluation Strategies for Grants and Proposals.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., December 10, 1996.

Hall-Ellis, Sylvia D. *Developing Needs Assessment for Grants and Proposals.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., November 12, 1996.

Hall-Ellis, Sylvia D. *Texas Library Connection Union Catalog Searching.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., November 8, 1996 and September 6, 1996.

Hall-Ellis, Sylvia D. *Grant Resources on the Internet.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., October 23, 1996.

Hall-Ellis, Sylvia D. *Preparing a Response to the Telecommunications Infrastructure Fund Board: Needs Assessment, Professional Development Framework, and Evaluation Strategies.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., October 16, 1996.

Hall-Ellis, Sylvia D. *Texas Library Connection: Building the Districtwide Bibliographic Database.* Sponsored and hosted by Mercedes Independent School District, Mercedes, Tex., September 13, 1996 and August 8, 1996. Sponsored and hosted by Los Fresnos Consolidated Independent School District, Los Fresnos, Tex., August 12, 1996.

Hall-Ellis, Sylvia D. *School-to-Work and Special Education: An Inclusive Partnership for Success.* Sponsored and hosted by the Office of Special Education, Region One Education Service Center, Edinburg, Tex., September 11, 1996.

Hall-Ellis, Sylvia D. *Telecommunications Infrastructure Fund Board: An Overview of Funding for Secondary Schools.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., August 31, 1996.

Hall-Ellis, Sylvia D. *Cataloging Books, Multimedia, and Realia in USMARC.* Sponsored and hosted by Edinburg Consolidated Independent School District, Edinburg, Tex., August 6, 1996.

Hall-Ellis, Sylvia D. *Internet Resources for Grant Writers.* Sponsored and hosted by Region One Education Service Center, Edinburg, Tex., April 20, 1996.

Hall-Ellis, Sylvia D. *Enhanced Grant Writing Skills for Mathematics Educators: Writing Skills for Campus Teams.* Sponsored and hosted by the Office of General Education, Region One Education Service Center, Edinburg, Tex., March 22, 1996.

Hall-Ellis, Sylvia D. *Internet Resources for Grant Writers.* Sponsored and hosted by the College of Education and Applied Science, Sam Houston State University, Huntsville, Tex., March 9, 1996.

Hall-Ellis, Sylvia D. *Grant Writing for Mathematics Educators: A Development Process for Campus Teams.* Sponsored and hosted by the Office of General Education, Region One Education Service Center, Edinburg, Tex., March 4, 1996.

Hall-Ellis, Sylvia D. *Shaking the Money Tree: Preparing Successful Technology Grant Applications.* Sponsored and hosted by the Office of Technology and Media Services, Region One Education Service Center, Edinburg, Tex., February 29, 1996 and April 29, 1996.

Hall-Ellis, Sylvia D. *District-wide Technology Planning: Technical Assistance for the Texas Education Agency Initiative.* Sponsored and hosted by the Office of Technology and Media, Region One Education Service Center, Edinburg, Tex., February 2, 1996.

Hall-Ellis, Sylvia D. *Enhanced Grant Writing Skills.* Sponsored and hosted by Pharr-San Juan-Alamo Independent School District, Pharr, Tex., January 27, 1996.

Hall-Ellis, Sylvia D. *United States Copyright Act of 1976, Video Transmissions, Computer Software and the Internet.* Sponsored and hosted by the Office of Technology and Media, Region One Education Service Center, Edinburg, Tex., January 24, 1996.

Hall-Ellis, Sylvia D. *The Grant Writing Development Process.* Sponsored and hosted by Pharr-San Juan-Alamo Independent School District, Pharr, Tex., January 20, 1996.

Hall-Ellis, Sylvia D. *MARC Cataloging of Materials for Library Media Centers.* Sponsored and hosted by the Office of Media Services, Region One Education Service Center for the Donna Independent School District, Donna, Tex., August 17, 1995.

Hall-Ellis, Sylvia D. *Classification.* Sponsored by AJ Seminars, Rockville, Maryland; presented at University Hilton Hotel, Houston, Tex., May 10, 1995.

Hall-Ellis, Sylvia D. *Library Technical Services.* Sponsored by AJ Seminars, Rockville, Maryland; presented at University Hilton Hotel, Houston, Tex., April 26, 1995.

Hall-Ellis, Sylvia D. *Grant Writing: An Introduction for Public School Administrators.* Sponsored and hosted by the Office of Administrative Services, Region One Education Service Center, Edinburg, Tex., April 19, 1995.

Hall-Ellis, Sylvia D. *The School Library Media Specialist in the 21st Century: Visions for the Future.* Sponsored and hosted by the United Independent School District, Laredo, Tex., March 31, 1995.

Hall-Ellis, Sylvia D. *Using USMARC.* Sponsored by AJ Seminars, Rockville, Maryland; presented at University Hilton Hotel, Houston, Tex., March 29, 1995.

Hall-Ellis, Sylvia D. *Library Media Center Policies and Guidelines: How to Prepare for School Board Adoption.* Sponsored and hosted by the Office of Technology and Media Services, La Joya Independent School District, La Joya, Tex., March 10, 1995.

Hall-Ellis, Sylvia D. *Developing District-wide Policies and Guidelines for Library Media Centers.* Sponsored and hosted by the Office of Technology and Media Services, La Joya Independent School District, La Joya, Tex., January 10, 1995.

Hall-Ellis, Sylvia D. *MARC Cataloging of Audiovisual Materials for Library Media Centers.* Sponsored and hosted by the Office of Library Media and Technology Services, Cypress-Fairbanks Independent School District, Houston, Tex., December 1, 1994.

Hall-Ellis, Sylvia D. *The School Library Media Specialist in the 21st Century: Visions for the Future.* Sponsored and hosted by United Independent School District, Laredo, Tex., October 14, 1994.

Hall-Ellis, Sylvia D. *The School Library Media Specialist in the 21st Century: Visions for the Future.* Sponsored and hosted by the Laredo Independent School District, Laredo, Tex., October 7, 1994.

Hall-Ellis, Sylvia D. *MARC Cataloging for Library Media Centers.* Sponsored by the Office of Library Media and Technology Services, Cypress-Fairbanks Independent School District, Houston, Tex., October 5, 1994.

Hall-Ellis, Sylvia D. *Jump to the Head of the Class: Undergraduate Library Resources Available at Sam Houston State University.* Sponsored by the Office of the Associate Vice President for Student Services, Sam Houston State University, Huntsville, Tex., October 4, 1994.

Hall-Ellis, Sylvia D. *CD-ROMs - 1994's Newest and the Best for Secondary Level Media Centers.* Sponsored and hosted by Clear Lake Independent School District, Houston, Tex., June 3, 1994.

Hall-Ellis, Sylvia D. *Introduction to Classification.* Sponsored by AJ Seminars, Rockville, Maryland; presented at Holiday Inn, Market Center, Dallas, Tex., May 18, 1994.

Hall-Ellis, Sylvia D. *Basic Descriptive Cataloging.* Sponsored by AJ Seminars, Rockville, Maryland; presented at University Hilton Hotel, Houston, Tex., May 4, 1994.

Hall-Ellis, Sylvia D. *Shaking the Money Tree - Part II: Writing Successful Grant Applications.* Sponsored by Donna Independent School District, Donna, Texas, and Region One Education Service Center Edinburg, Texas; presented at South Texas Community College Library, McAllen, Tex., April 29, 1994.

Hall-Ellis, Sylvia D. *Using MARC.* Sponsored by AJ Seminars, Rockville, Maryland; presented at Holiday Inn, Market Center, Dallas, Tex., April 6, 1994.

Hall-Ellis, Sylvia D. *Automated Authority Control.* Sponsored by AJ Seminars, Rockville, Maryland; presented at University Hilton Hotel, Houston, Tex., March 23, 1994.

Hall-Ellis, Sylvia D. *Automating the District School Library Media Centers: Choices and Opportunities.* Sponsored and hosted by the Office of Technology and Library Media Services, Fort Bend Independent School District, Sugar Land, Tex., March 4, 1994.

Hall-Ellis, Sylvia D. *Shaking the Money Tree - Part I: Preparing Successful Grant Applications.* Sponsored and hosted by the Office of Library Media Services, Donna Independent School District, Donna, Tex., February 4, 1994.

GRADUATE COURSES TAUGHT

San José State University, School of Library and Information Science
INFO 249 – Advanced Cataloging and Classification (Fall 2015, 2016; Summer 2016)
INFO 287 – Special Topics in Cataloging and Classification (Spring 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026; Fall 2020, 2021, 2022, 2023, 2024, 2025, 2026)
LIBR 248 – Beginning Cataloging and Classification (Summer 2002, 2003, 2004, 2005, 2006)
LIBR 249 – Advanced Cataloging and Classification (Summer 2003; Fall 2014; Summer 2015)

University of Denver, University College
Grant Writing Certificate Seminar for Non-Profit Organizations (Fall 2020, 2021; Spring 2021)

University of Denver, Morgridge College of Education HED
5991 – Grant Writing in Higher Education (Spring 2011)
LIS 4010 – Organization of Information (Fall 2002-2009; Winter 2005; Spring 2004-2006)
LIS 4020 – Professional Principles and Ethics (Summer 2000)
LIS 4040 – Management of Libraries and Information Centers (Fall 2010; Spring 2003, 2009 (DS); Winter 2005)
LIS 4070 – Cataloging and Classification (Winter 2008, 2009, 2010, 2011, 2012; Fall 2009 (DS)) LIS
4321 – Collection Management (Spring 2005)
LIS 4326 – LIS Research (Winter 2009 (DS); Spring 2009 (DS))
LIS 4350 – Adult Materials and Services (Summer 2006, 2009)
LIS 4379 – Social Sciences Resources (Spring 2009)
LIS 4400 – Cataloging and Classification (Spring 2000, 2001; Winter 2002-2009; Summer 2005, 2006 – course renumbered LIS 4070, September 2007)
LIS 4401 – Descriptive Cataloging (Winter 2001; Spring 2002, 2003, 2006-2009; Summer 2005)
LIS 4402 – Subject Cataloging (Spring 2001, 2009; Summer 2005, 2006, 2007, 2008)
LIS 4403 – Classification Schemes (Fall 2007, 2008, 2009 (DS)) LIS
LIS 4405 – Authority Control (Winter 2009)
LIS 4510 – Materials and Services for Children (Winter 2004; Summer 2005)
LIS 4620 – Grant Writing and Fundraising (Summer 2000, 2002, 2004, 2010; Winter 2006; Fall 2006-2009)
LIS 4700 – Seminar in Technical Services (Fall 2001)
LIS 4700 – Seminar in Public Libraries (Summer 2002)
LIS 4804 – Management of Electronic Records (Spring 2004; Fall 2005)
LIS 4902 – Capstone Projects (Winter 2003, Spring 2008, 2009, 2011)
LIS 4910 – Independent Study (every quarter Winter 2000 through Spring 2011)
LIS 4920 – Service Learning (every quarter Summer 2004 through Spring 2011)
RMS 4954 – Grant Writing (Summer 2013, 2014)
RMS 4959 – Content Analysis Methodology (Spring 2015)
QRM 4978 – Grant Writing (Summer 2011, 2012)

Rutgers University, School of Communication and Information
SC&I 522 – Cataloging and Classification (Summer 2013)

University of Arizona, College of Behavioral and Social Sciences
LIS 602 – Cataloging and Classification (Summer 1995)
LIS 612 – Advanced Online Search and Retrieval (Summer 1995)

Sam Houston State University, College of Education & Applied Science
LIS 532 - Cataloging and Classification (Fall 1993, 1994, 1995)
LIS 563 - Advanced Cataloging and Classification (Summer 1994)
LIS 567 - Research Methods (Spring 1994, Spring 1995)
LIS 591 - Educational Technology (Spring 1994, 1995)
LIS 596 - Networking and Computer Technologies in Education (Fall 1993, Summer 1994, Fall 1995)

Dissertations at the University of Denver

Bowers, Stacey L. *Library Anxiety of Law School Students: A Study Utilizing the Multidimensional Library Anxiety Scale.* Chair, May 5, 2010.

Fattor, Melissa M. *Student Engagement Differences by Ethnicity and Scale for Ninth Grade Students.* Chair, November 1, 2010.

Fulton, Roseanne. *A Case Study of Culturally Responsive Teaching in Middle School Mathematics.* Kent Seidel, Chair, June 18, 2009, Outside Chair.

Grealy, Deborah S. *Tribes and Territories in Library and Information Studies Education.* Bruce Uhrmacher, Chair, June 10, 2008, Committee Member.

McCord, J. Michael. *Developing a Standard of Care for Educational Malpractice.* Chair, April 15, 2011.

Priebe, Sarah J. *Distinguishing Effects of Domain and General Knowledge on Passage Fluency and Comprehension.* Jan Keenan, Chair. July 21, 2011, Outside Chair.

Snyder-Mondragon, Sandra M. *Institutional Factors that Impact the Retention of Graduate Students of Color in Schools of Library and Information Science: A Metaregression of Accredited Library School Statistics on Student Retention and Graduation Rates.* Kathy Green, Chair, July 24, 2009, Committee Member.

Taylor, Karen Pickles. *Effective Teaching.* Elinor Katz, Chair, July 15, 2009, Outside Chair.

Thompson, Jennifer. *Distinguishing a Western Women's College: A History of the Curriculum and Student Experience at Colorado Women's College.* Edith W. King, Chair, July 16, 2010, Committee Member.

Walker, Emelda. *Influence of Organizational Factors on Job Satisfaction of Disability Service Providers at Postsecondary Institutions.* Chair, April 29, 2010.

Dissertation Proposals at the University of Denver

Bowers, Stacey L. *Library Anxiety of Law School Students: A Study Utilizing the Multidimensional Library Anxiety Scale.* Chair, November 5, 2009.

McCord, J. Michael. *Developing a Standard of Care for Educational Malpractice.* Edith W. King, Chair. April 23, 2010.

Thompson, Jennifer. *History of Colorado Women's College.* Edith W. King, Chair, October 30, 2008. Committee Member.

Walker, Emelda. *Influence of Organizational Factors on Job Satisfaction of Disability Service Providers at Postsecondary Institutions.* Chair, July 21, 2009.

Dissertations at Other Institutions

Rodríguez-Mori, Howard. *The Information Behavior of Puerto Rican Immigrants to Central Florida, 2003-2009: Grounded Analysis of Six Case Studies Use of Social Networks during the Migration Process.* Kathleen Burnett, Chair, April 10, 2009, Florida State University, Committee Member.

Schwartz, Brian *More than a Look-up Skill: Medical Information Literacy Education in Osteopathic Medical Schools.* D. Mirah Dow, Chair, July 18, 2017. Emporia State University, Committee Member.

Snow, Karen. *A Study of the Perception of Cataloging Quality among Catalogers.* Shawne D. Miksa, Chair, August 1, 2011. University of North Texas, Committee Member.

Dissertation Proposals at Other Institutions

Schwartz, Brian *More than a Look-up Skill: Medical Information Literacy Education in Osteopathic Medical Schools.* D. Mirah Dow, Chair, December 4, 2015. Emporia State University, Committee Member.

Snow, Karen. *A Study of the Perception of Cataloging Quality among Catalogers.* Shawne D. Miksa, Chair, May 11, 2010. University of North Texas, Committee Member.

Master's Thesis at the University of Denver

Hemingson, Jeff. *Recital Paper.* Lamont School of Music, February 2010, Outside Chair.

Capstones at the University of Denver

Anthony, Alisa. *Correlation of Library and Information Science Program outcomes and Vacant Position Qualifications Listed on the Colorado State Library Jobline by Employers During the Period September 1, 2000 through August 31, 2002.* Chair, Winter Quarter, 2003.

Borden, Donna M. *Improving Emergency Communications Systems: Is a Radio Communications Network the Answer?* Fall Quarter, 2014.

Bowden, Heather L.M. *Exploring Biological Models for Long-term Data Preservation.* Chair, Spring Quarter, 2008.

Casenada, Cassandra Y. *Challenges to the Recruitment, Education, and Retention of Librarians of Color*. Chair, Spring Quarter, 2008.

Chang, Jennifer C. *Legal Research Practice and Preference: A Law Firm Perspective.* Chair, Spring Quarter 2011.

Ellis, Megan S. Fitzgerald. *Design of a Public Library Adult Volunteer Recruitment Program and Training Curriculum.* Chair, Winter Quarter, 2003.

Kircher, Kathy. *Development of a Library Pathfinder for Exobiology and Posting it on the Internet.* Chair, Winter Quarter, 2003.

Melhado, Loretta. *Design of St. John's Episcopal Church Library.* Chair, Spring Quarter, 2008.

Radcliff, Kathy. *Original Cataloging of Archival materials in the HERS Collection in Penrose Library.* Chair, Winter Quarter, 2003.

Sass, Carol Ann. *ACT Periodical Index: Access to Catholic Thinking Periodical Index: Web Index to Select Catholic Periodicals.* Spring Quarter 2000.

Stone, Sergio D. *Conducting Community Analysis for the Bemis Public Library (Littleton, Colorado) Using 2002 U.S. Census Data and the Online Outcome-Based Evaluation Toolkit.* Chair, Winter Quarter, 2003.

Tureson, Tamara. *Design and Use of an Information Audit Tool for Use in a Law Library.* Chair, Winter Quarter, 2003.

Tweed, Beth. *Evaluation of Email Reference Service in a Consumer Health Library Environment.* Chair, Winter Quarter, 2003.

GRANTS AND CONTRACTS

Student Learning in Agriculture STEM through Teacher Professional Development. Research team: Stanton Gartin (PI) and Cyndi Hofmeister, Northeast Junior College; Jeff Cash, Cheryl Sánchez, Anne-Marie Crampton, Lamar Community College; Suzanna Spears, Morgan Community College; Jack Wiley, Kerry Gabrielson, Trinidad State Junior College; Michael J. Miller, Colorado State University; Michael Womochil (Co-PI), Casey Sacks, and Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Agriculture, Agriculture and Food Research Initiative, Food, Agriculture, Natural Resources and Human Sciences Education and Literacy Initiative, $404,460 (2017-2020)

Leading and Achieving: the Colorado Agriculture Regional Consortium. Research team: Jack Wiley (PI), Kerry Gabrielson, Trinidad State Junior College; Jeff Cash (PI), Cheryl Sánchez, Anne-Marie Crampton, Lamar Community College; Suzanna Spears, Morgan Community College; Cyndi Hofmeister, Northeast Junior College; Michael Womochil (Co-PI), Casey Sacks, and Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Agriculture, Hispanic Serving Institutions Education Grants Program, $504,414 (2017-2020)

Cyber Prep Program Planning Grant. Research team: Debbie Sagen (PI), Brenda Lauer, Pikes Peak Community College; Gretchen Martin, Koiosa Insights; and, Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Commerce, Regional Alliances and Multi-stakeholder Partnerships to Stimulate (RAMPS) Cybersecurity Education and Workforce Development, $199,681 (2016-2018)

Career and Technical Education in Colorado: Pathways to Education and Employment. Research team: Heather McKay (PI), Rutgers University; Sarah Heath, Casey Sacks, Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Education, Institute of Education Sciences, $1,400,000 (2017-2021)

Pre-Alliance Planning Grant Colorado Community College Alliance. Research team: Victor Vialpondo (PI) and Janel Highfill, Community College of Aurora; Heidi Loshbaugh (Co-PI), Community College of Denver; Rick Reeves, Bill McGreevy, Liz Cox, and Kristin Aslin, Red Rocks Community College; Cathy Pellish, Front Range Community College; Samuel DeVries, Arapahoe Community College; Sylvia D. Hall-Ellis, Colorado Community College System. National Science Foundation, Louis Stokes Alliance for Minority Participation (LSAMP), $86,817 (2016-2017)

Towards Scalable Differentiated Instruction Using Technology-enabled, Competency-based, Dynamic Scaffolding. Research team: Karen Wilcox (PI) and Vijay Kumar, Center for Computational Engineering, Massachusetts Institute of Technology; Flora McMartin, Broad-based Knowledge; Quinsigamond Community College; and, Casey Sacks, and Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Education, First in the World Developmental Grant, $2,891,882 (2015-2019)

Colorado Strategic Partnerships Emergency Grant. Research team: Elise Lowe-Vaughn (PI), Amy Hodson, Celia Hardin, Barbara McBride, James Newby, Christopher Dewhurst, Nina Holland, Kate Anderson, MaryAnn Roe, Chrystalynn Chrystalynn, Elaine Edon, Mona Barnes, Tom Morgan, Rob Hanni, and Marie Valenzuela, Colorado Department of Labor and Employment; Steve Anton, Joelle Brouner, and, Katie Griego, Colorado Department of Human Services; Rebecca Holmes, Judith Martinez, and Jennifer Jirous, Colorado Department of Education; Emily Templin Lesh, Colorado Workforce Development Council; Cory Everett, Colorado Department of Regulatory Agencies; and, Casey Sacks and Sylvia D. Hall-Ellis, Colorado Community College System. U.S. Department of Labor, Strategic Partnerships Emergency Grant, $5,000,000 (2015-2017)

CAEL Jump Start Program: Competency-Based Education. Research team: Jerry Migler (PI), Casey Sacks, Debra Cohn, Thomas Hartman, and Sylvia D. Hall-Ellis, Colorado Community College System; Matt Jamison, Front Range Community College; Mike Coste, Red Rocks Community College; Amanda Corum, Pueblo Community College; Janet Colvin, Pikes Peak Community College; and, MaryAnn Matheny, Community College of Denver. Council for Adult and Experiential Learning, $15,000 (2015)

Summit on the Redesign of Developmental Education. Research team: Jerry Migler (PI), Casey Sacks, and Sylvia D. Hall-Ellis, Colorado Community College System; Chip Nava, Pueblo Community College; Kim Moultney, Arapahoe Community College; Debbie Ulibarri, Trinidad State Junior College; and, Kris Bernard, Front Range Community College. American Association of Community Colleges, $15,000 (2015-2016)

Equity in Excellence at Colorado Community Colleges. Research team: Keith Howard (PI) and Sylvia D. Hall-Ellis, Colorado Community College System; Estela Mata Bensimon, Center for Urban Education, University of Southern California; and Kerry Gabrielson, Trinidad State Junior College. Colorado Department of Higher Education, Colorado Opportunity Scholarship Program, $150,000 (2015-2016)

MBA High School of Business in Colorado. Research team: Laurie Urich (PI) and Sylvia D. Hall-Ellis, Colorado Community College System; Rudolph Sumpter and Beatrice Gerrish, Boulder Valley Schools; Keith Curry Lance, RSL Research Group. Colorado Department of Higher Education, Colorado Opportunity Scholarship Program, $501,295 (2015-2016)

Fullbridge Program in Colorado. Research team: Keith Howard (PI) and Sylvia D. Hall-Ellis, Colorado Community College System; Suzanna Spears, Fort Morgan Community College; and, Cheryl Sánchez and Anne-Marie Compton, Lamar Community College. Colorado Department of Higher Education, Colorado Opportunity Scholarship Program, $300,000 (2015-2017)

Colorado Community College System Alternative Credit Program. Research team: Jerry Migler (PI), Keith Howard, and Sylvia D. Hall-Ellis, Colorado Community College System. American Council on Education, $13,000 (2015)

Internationalization in Higher Education. Researcher: Samuel D. Museus. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $8,000 (2014-2015)

International Perspectives on Bilingual Education. Researcher: Sharolyn Pollard-Durodola. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $3,700 (2014-2015)

Cultivating Culturally Relevant and Responsive Curriculum and Pedagogy in College. Researcher: Samuel Museus. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, iRise Grant, $5,000 (2014-2015)

Project EMERGE (Educational Model for Evaluation and Replicability in Gifted Environments). Research team: Norma Hafenstein (PI) and Bruce Uhrmacher. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Lynde and Harry Bradley Foundation, $235,000 (2014-2015)

Collecting Asian American Refugee Stories. Researcher: Samuel Museus. Technical reviewer for MCE: Sylvia D. Hall-Ellis. American Educational Research Association, Research Grant, $5,000 (2014-2015)

Developing Expertise in Teaching K-5 Mathematics. Research team: Julie Sarama (PI) and Douglas H. Clements (Co-PI), in partnership with the School of Education at the University of Michigan. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, $130,344 (2013-2015)

Investigation of the Long-Term Outcomes for Special Education Students. Researcher: Antonio Olmos-Gallos. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Jefferson County School District, Office of Assessment, $40,000 (2014)

Refugee Community Collaboration. Researcher: Vicki Tomlin (PI) in partnership with the African Community Center and Jewish Family Service. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, $14,854 (2014-2015)

Access to Mathematics for All. Research team: Richard S. Kitchen (PI), Nicole M. Russell (Co-PI), and Terrence Blackman (Co-PI), Curriculum Studies and Teaching Program, Morgridge College of Education; Álvaro Árias (Co-PI, Department of Mathematics); and, James Gray (Co-PI), Department of Mathematics, Community College of Aurora. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, The Robert Noyce Scholarship Program, Capacity Building Project, $349,926 (2014-2016)

Cognitive Test Battery for Intellectual Disabilities. Research team: Karen Riley (PI), Lyndsay Agans, Jessica Lerner, and Karin Dittrick-Nathan in partnership with David Hessl (PI), The MIND Institute at the University of California – Davis, and Elizabeth Berry-Kravis, Rush University Medical Center. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Institutes of Health, Outcome Measures for Use in Treatment Trials for Individuals with Intellectual and Developmental Disabilities (R01), $2,499,996 ($588,672 at DU) (2014- 2019)

Broadening Participation in Engineering among Women and Latino/as: A Longitudinal, Multi-Site Study. Researcher: Patton Garriott (PI) in partnership with the University of North Dakota and the University of Missouri. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, HER Core Research, $677,390 ($69,992 at DU) (2014-2019)

Designing a Teacher Evaluation System to Improve Teacher Effectiveness for Culturally and Linguistically Diverse Learners. Research team: María del Carmen Sálazar (PI), Jessica Lerner, and Kathy Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Professional Research Opportunities for Faculty, $29,988 (2014-2016)

Developing a College-Going Culture in Latina/O Families: Exploring the Influence of Funds of Knowledge on Family Outreach Programs. Researcher: Judy Marquez Kiyama. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Professional Research Opportunities for Faculty, $18,720 (2014-2016)

Assessment of Quality of Life in Neutral Implantation Surgery for the Treatment of Parkinson's Disease. Researcher: Cynthia McRae. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $3,967 (2014-2015)

Pura Vida: Cloud Forest, Curriculum and Cross-Cultural Study. Research team: Norma Hafenstein (PI) and Bruce Uhrmacher (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $8,000 (2014-2015)

Online Course Development for Curriculum and Instruction. Researcher: Ruth Chao. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $3,000 (2014)

Online Course Development for Curriculum and Instruction. Researcher: María del Carmen Sálazar. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $3,000 (2014)

Online Course Development for Curriculum and Instruction. Researcher: Jessica Lerner. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $3,000 (2014)

Online Course Development for Curriculum and Instruction. Researcher: Duan Zhang. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $3,000 (2014)

The Mathematics Education of African Americans, 1866 – 1954. Researcher: Nicole M. Russell. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Academy of Education, Spencer Foundation Postdoctoral Fellowship, $55,000 (2014-2016)

Early Childhood Care and Education Study. Research team: Carrie Germeroth (PI), Melissa Mincic, and Douglas H. Clements. Technical reviewer for MCE: Sylvia D. Hall-Ellis. State of North Dakota, Department of Public Instruction, $73,500 (2013-2014)

Graduate Level Specialty in Addiction Counselor Training with Emphasis on Integration of Native American Specific Content. Research team: Ruth Chao (PI) and Michael J. Faragher (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. The Galena Foundation, $289,732 (2013-2016)

The Collecting Asian American and Pacific Islander Refugee Stories (CARS) Project. Researcher: Samuel Museus. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, $15,000 (2014)

Parents in Transition: A Multiple Case Study of Parent and Family Orientation Programs. Researcher: Judy Marquéz Kiyama. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Faculty Research Fund, $1,500 (2014-2015)

The Sistah Network: Black Women Graduate Students Supporting and Retaining Each Other. Researcher; Nicole M. Russell. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Faculty Research Fund, $2,708 (2014-2015)

Evaluation of the Northeast Denver Babies Ready for College Program. Research team: Carrie Germeroth (PI), Melissa Mincic, and Douglas H. Clements. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Mile High Montessori, $9,922 (2013-2014)

Developing Teaching Expertise in K-5 Mathematics. Research team: Julie Sarama (PI) and Douglas Clements (University of Denver), in partnership with Timothy Boerst (PI), Meghan Shaughnessy, Deborah Ball, Hyman Bass (School of Education at the University of Michigan). Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, $449,827 ($130,344 at the University of Denver (2013-2015)

Early Learning Care and Education Study Program Grant for the State of North Dakota. Research team: Carrie Germeroth (PI), Melissa Mincic, and Sheridan Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. State of North Dakota Department of Public Instruction, $73,500 (2013-2014)

Local Professional Learning Community for School Leaders. Research team: Susan Korach (PI), Kristina Hesbol, and Rebecca McClure. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Education Development Center, $10,950 (2013)

Healthy Eaters, Lifelong Movers 2: Implementing Evidence-Based School Environment, Policy, and Curricular Changes to Increase Opportunities for Healthy Eating and Physical Activity in Low Income, Rural Colorado. Research team: Elaine Berlansky, University of Colorado at Denver (PI), Nicholas Cutforth, University of Denver (Co-PI), and Allison Reeds. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Health Foundation, $3,103,108 (2013-2017)

Hughes Rare Books Library Room Renovation in the Sturm College of Law. Project team: Patti H. Marks, Sylvia D. Hall-Ellis, and Leigh Elliott. Mabel T. Hughes Charitable Trust, $34,000 (2013-2014)

The Promise Center Partnership with the Marsico Institute for Early Learning and Literacy and the City and County of Denver. Research team: Karen Riley (PI), Douglas H. Clements (Co-PI), and Sheridan Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. The Piton Foundation, $223,468 (2013-2014)

Center of Excellence for Problem Gambling. Research team: Ruth Chao (PI) and J. Mike Faragher. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Department of Behavioral Health, $68,021 (2012-2013)

United Way Implementation and Validation Review. Research team: Douglas H. Clements (PI), Amanda Moreno, and Sheridan Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Mile High United Way, $19,737 (2013)

Math/Science Partnership in Rural Districts. Research team: Kristen Bunn (PI, Eagle County Schools), Paul Michalec (Co-PI), and Alegra Reiber. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Department of Education, Colorado's Mathematics and Science Partnership Program, $750,000 (2013-2014)

Mathematics and Science Education of African Americans. Research team: Nicole Russell (PI), Sylvia D. Hall-Ellis, Steve Fisher (Penrose Library). Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Professional Research Opportunities for Faculty, $29,994 (2013-2015)

Online Course Development for Curriculum and Instruction. Researcher: Nicole Russell. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $3,000 (2013)

Discourse and Opportunity: Undocumented Students and Higher Education Policy. Researcher: Ryan Gildersleeve. Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Academy of Education, Spencer Foundation Postdoctoral Fellowship, $55,000 (2012-2014)

An Anthropological Study of the Latino Graduation Ceremony. Researcher: Ryan Gildersleeve. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Spencer Foundation, $39,900 (2012-2013)

Fragile-X and Pharmaceutical Company: Clinical Trial of AFQ056. University of California – Davis Children's Hospital MIND Institute (Sacramento), Children's Hospital Denver, and Rush Children's Hospital at Rush University Medical Center (Chicago). Karen Riley, PI. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Novartis Pharmaceuticals Corporation, $394,206 (2012-2014)

Project Engage, Phase 2, a DAPRA Grant in partnership with Total Immersion Systems, Inc., and Texas A&M University. Research team: Karen Riley (PI) and Lyndsay Agans (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. U.S. Department of Defense, $60,000 (2011-2013)

International School Psychology Practicum Exchange. Researcher: Gloria L. Miller. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, $14,951 (2012-2013)

Online Course Development for Curriculum and Instruction. Research team: Bruce Uhrmacher (PI) and Norma Hafenstein (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Teaching and Learning, $18,193 (2012-2014)

Writers in the Schools. Research team: Karen Riley (PI) and Amanda Moreno (Co-PI).Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Humanities, $4,965 (2012-2013)

Learning Ecosystem Validation Grant. Research team: Karen Riley (PI), Lyndsay Agans, Kent Seidel, and Shimelis Assefa. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Bill & Melinda Gates Foundation, $315,000 (2012-2013)

Project Words of Oral Reading and Language Development (WORLD). Research team: Jorge E. Gonzalez (PI), Texas A&M University; Laura Saenz (Co-PI), University of Texas – Pan American; and, Sharolyn Pollard-Durodola (Co-PI), University of Denver. Technical reviewer for MCE: Sylvia D. Hall-Ellis. U.S. Department of Education, Institute of Education Sciences, $53,354 (2012-2015); award $640,718, transfer from Texas A&M University

Increasing the Efficacy of an Early Mathematics Curriculum with Scaffolding Designed to Promote Self-Regulation. Research team: Douglas H. Clements (PI) and Julie Sarama (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. U.S. Department of Education, Institute of Education Sciences, $1,445,315 (2008-2014); award $4,541,975, transfer from University at Buffalo, The State University of New York

Using Rule Space and Poset-based Adaptive Testing Methodologies to Identify Ability Patterns in Early Mathematics and Create a Comprehensive Mathematics Ability Test. Research team: Douglas H. Clements (PI) and Julie Sarama (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, $323,791 (2010-2014); award $1,194,944, transfer from University at Buffalo, The State University of New York

Comprehensive Postdoctoral Training in Scientific Education Research. Researcher: Julie Sarama (PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. U.S. Department of Education, Institute of Education Sciences, $133,458 (2010-2014); award $613,353, transfer from University at Buffalo, The State University of New York

Longitudinal Study of a Successful Scaling Up Project: Extending TRIAD. Research team: Douglas H. Clements (PI), Julie Sarama (Co-PI), and Abt Associates (Carolyn Layzer, Fatih Unlu, Laurie Bozzi, Lily Fesler, Alina Martinez, Cristofer Price, James van Orden). Technical reviewer for MCE: Sylvia D. Hall-Ellis. Institute of Education Sciences, $384,940 (2011-2015; award $1,250,286, transfer from University at Buffalo, The State University of New York

Early Childhood Education in the Context of Mathematics, Science, and Literacy. Research team: Julie Sarama (PI), Douglas H. Clements (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. National Science Foundation, $990,020 (2010-2014; award $2,285,228, transfer from University at Buffalo, The State University of New York

Early Childhood Clearinghouse Information Center Redesign. Research team: Karen Riley (PI) and Amanda Moreno (Co-PI).Technical reviewer for MCE: Sylvia D. Hall-Ellis. Office of the Lt. Governor of Colorado, $12,000 (2012-2013)

Morgridge Rural Educational Leadership Initiative. Research team: Lyndsay Agans (PI), Linda Brookhart (Co-PI), Susan Korach, and Rebecca McClure. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Morgridge Family Foundation, $100,000 (2012-2014)

Center of Excellence for Problem Gambling. Research team: Patrick Sherry (PI) and J. Mike Faragher. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Department of Behavioral Health, $55,000 (2012-2013)

Intermodal Transportation Institute Research Initiatives. Researcher: Patrick Sherry. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University Transportation Centers Program, Research and Innovative Technology Administration, U.S. Department of Transportation. $600,000 (2012-2014)

Mile High United Way Social Innovation Fund Early Literacy Initiative. Gloria L. Miller, PI, Amanda Moreano, Kim Hartnett-Edwards, and Sheridan Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Mile High United Way, $89,992 (2012-2013)

International School Psychology Practicum Exchange. Researcher: Gloria L. Miller. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, $1,000 (2012-2013)

Refugee Student Art Outreach. Researcher: Karin Dittrick-Nathan. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Faculty Research Fund, $3,000 (2012-2013)

Resistance, Resilience, and Reciprocity: Centering the Voices of Black Doctoral Women with Faculty Aspirations. Researcher: Nicole M. Russell. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Faculty Research Fund, $2,965 (2012-2013)

Assessing Learning through Student Notebooks. Research team: Keith Miller, Nancy Sasaki, and Kathy Green. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Professional Research Opportunities for Faculty, $30,000 (2012-2014)

International School Psychology Practicum Exchange. Researcher: Gloria L. Miller. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $2,000 (2012-2013)

ELO in Colorado. Research team: Cynthia Hazel and Duan Zhang. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Legacy Foundation, $150,000 (2012-2013)

Creating Online LIS Courses. Research team: Mary C. Stansbury (PI), Shimelis Assefa, Denise Anthony, Xiao Hu, and Krystyna Matusiak. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Teaching and Learning, $15,000 (2011-2013)

Maritime Piracy Seminar. Researcher: Ved Nanda, The Nanda Center, Sturm College of Law. Technical reviewer for Sturm College of Law: Sylvia D. Hall-Ellis. Arsenault Family Foundation, $15,000 (2012-2013).

Early Childhood Librarianship: An Interdisciplinary, Experiential Learning MLIS. Researcher: Mary C. Stansbury (PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. Laura Bush's 21st Century Librarians Program, Institute for Museums and Library Services, $249,066 (2012-2014)

Learning Ecosystem Planning Grant. Research team: Karen Riley (PI), Lyndsay Agans, Kent Seidel, and Shimelis Assefa. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Bill & Melinda Gates Foundation, $281,217 (2011-2012)

Advanced Service Learning Practitioner Faculty Grant, Sylvia D. Hall-Ellis, University of Denver, Center for Community Engagement and Service Learning, $400 (2011)

Evaluating and Enhancing the EspeciallyMe Program. Research team: Lori D. Patton and Nicole Russell. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, Public Good Fellowship Grant, $24,780 (2012)

Choosing Excellence: Let Every Child Bloom. Research team: Shimelis Assefa (PI) and Mary C. Stansbury (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Community Engagement and Service Learning, Public Good Grant, $7,657 (2012)

K-8 STEM Content Specific Professional Development to Improve Elementary Student Achievement. Research team: Kent Seidel (PI), Nicole Russell (Co-PI), Kimberly Hartnett-Edwards, Paul Michalec, Jeff Farmer, Keith Miller, Alegra Reiber, and Nancy Sasaki. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Department of Education, Improving Teacher Quality Grant, (Title II ESEA), $307,299 (2011-2012)

Building a Better Principal Pipeline to Boost Student Achievement. Researcher: Susan Korach. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Denver Public Schools; a subcontract from the Wallace Foundation Grant, $170,000 (2011-2017)

Center of Excellence for Problem Gambling. Research team: J. Mike Faragher (Co-PI) and Bobbie Vollmer (PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Department of Behavioral Health, $79,999 (2011-2012)

An Exploration of Novice Teachers' Core Competencies: Impacts on Student Achievement, and Effectiveness of Preparation. Research team: Kent Seidel (PI), Kathy Green (Co-PI), Kimberly Hartnett-Edwards, and Duan Zhang. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Institute of Education Sciences, Effective Teachers and Effective Teaching, $990,987 (2012-2015)

Project Engage, a DAPRA Grant in partnership with Total Immersion Systems, Inc., and Texas A&M University. Research team: Karen Riley (PI) and Lyndsay Agans (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. U.S. Department of Defense, $49,964 (2011-2014)

User-centered Evaluation of Music Search Engines. Researcher: Xiao Hu. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Faculty Research Fund, $2,931 (2011-2012)

Educational Practicum in Vietnam and China to Promote the Inclusion of Young Children with Disabilities. Researcher: Gloria L. Miller. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Office of Internationalization, $5,000 (2011-2012)

Evaluation of Colorado's Enhancing Quality in Infant-Toddler (EQIT) Initiative. Research team: Virginia R. Maloney and Amanda Moreno. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Buell Foundation, $395,884 (2011-2013)

Creating Engaging Environments to Teach Pre-Algebra Mathematics to Elementary Students. Research team: Álvaro Arias, Mario López, María del Carmen Salazar, and Lyndsay Agans. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver Interdisciplinary Grant, $60,000 (2011-2012)

Morgridge Education Technology Accessible (META) Resource Project. Research team: Lyndsay Agans (PI) and Shimelis Assefa (Co-PI). Technical reviewer for MCE: Sylvia D. Hall-Ellis. Morgridge Family Foundation, $36,000 (2011)

MCE Connect: A 21st Century Framework for Faculty Development. Research team: Bruce Uhrmacher (PI), Shimelis Assefa (Co-PI), Lyndsay Agans (Co-PI), Kimberly Hartnett-Edwards, Norma Hafenstein, Xiao Hu, Paul Michalec (Co-PI), María del Carmen Salazar (Co-PI), and Sandra Snyder-Mondragon. Technical reviewer for MCE: Sylvia D. Hall-Ellis. University of Denver, Center for Teaching and Learning, $22,355 (2011-2012)

Parent Education and Parent Leadership and Advocacy. Research team: Virginia Maloney (PI) and Amanda Moreno, Marsico Institute for Early Learning and Literacy. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Health Foundation, $27,358 (2010-2011)

Intentional School Culture. Researcher: Cynthia Hazel. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Denver Public Schools, $38,040 (2010-2011)

Healthy Eaters, Lifelong Movers: Implementing Evidence-Based School Environment, Policy, and Curricular Changes to Increase Opportunities for Healthy Eating and Physical Activity in Low Income, Rural Colorado. Research team: Elaine Berlansky, University of Colorado at Denver (PI), Nicholas Cutforth, University of Denver (Co-PI), and Allison Reeds. Technical reviewer for MCE: Sylvia D. Hall-Ellis. Colorado Health Foundation, $1,683,277 (2011-2014)

Second Life Learning Community. Research team: Don McCubbrey (PI), Sylvia D. Hall-Ellis (Co-PI), Walter LaMendola, and Paul Novak. University of Denver, Center for Teaching and Learning, $13,000 (2010-2011)

Reintroducing the Value of Law Librarians to Academic and Public Librarians in Colorado through the Identification and Use of Emerging Technologies. Research team: Sylvia D. Hall-Ellis (Co-PI), Stacey L. Bowers (PI), and Christopher Hudson, in partnership with Denver Public Library, Arapahoe Library District, and the Colorado State Supreme Court Library. University of Denver, Center for Community Engagement and Service Learning, $5,773 (2010-11)

Reintroducing the Value of Law Librarians to Public Librarians through the Identification and Use of Emerging Technologies and Resources. Research team: Stacey Bowers (PI) and Sylvia D. Hall-Ellis (Co-PI). American Association of Libraries, Wolters Kluwer Law & Business Grant Program, $2,725 (2010-2011)

Faculty Service Learning Pod. Research team: María del Carmen Salazar and Nicholas Cutforth (Curriculum and Instruction Program), Frank Tuitt (Higher Education Program), Cynthia Hazel and Gloria Miller (Child, Family, and School Psychology Program), and Sylvia D. Hall-Ellis (Library and Information Science Program). University of Denver, Center for Community Engagement and Service Learning, $8,000 (2010-2011)

Lincoln Collaborative. Research team: Antonio Esquibel (Principal, Lincoln High School), María del Carmen Salazar (Curriculum and Instruction Program), and. Sylvia D. Hall-Ellis. School Improvement Grant, Denver Public Schools, $375,000 (2010-2012)

Future LEADers of America: Leaders III. Research team: Denver Public Library (Kristen Svendson, PI), the University of Denver (Sylvia D. Hall-Ellis), the Colorado Chapter of REFORMA (Orlando Archibeque), and the Colorado Association of Libraries (Martin Garnar). Laura Bush's Recruiting Librarians for the 21st Century Program, Institute for Museums and Library Services, $988,366 (2009-2012)

Teaching for Success in the Library Environment: LIS 4030 in Library 2.0. Research team: Deborah S. Grealy (PI) and Sylvia D. Hall-Ellis (Co-PI). University of Denver, Center for Teaching and Learning, $ 9,350 (2009-10)

Writing Group Faculty Grant, Sylvia D. Hall-Ellis, University of Denver, Center for Community Engagement and Service Learning, $750 (2008-2009)

Collaborative Learning Faculty Grant, Sylvia D. Hall-Ellis, University of Denver, Center for Teaching and Learning, $1,500 (2008)

Connecting Information Literacy to Learning. Research team: Lori Micho, Merrie Valliant, Amanda Samland, and Sylvia D. Hall-Ellis. Colorado State Library, LSTA Discretionary Grant Program, $19,548 (2008-2009)

Law Librarianship Fellows Program. Research team: Sylvia D. Hall-Ellis (PI), Stacey Bowers (Co-PI), and Christopher Hudson, Westminster Law Library. Laura Bush's 21st Century Librarians Program, Institute for Museums and Library Services, $999,370 (2008-2012)

Grant Writing in a Cooperative Learning Environment. Sylvia D. Hall-Ellis. University of Denver, Center for Teaching and Learning, $2,000 (2008-2009)

Project Homeless Connect 6 Event Evaluation. Research team: Sylvia D. Hall-Ellis and Duan Zhang. University of Denver Center for Community Engagement and Service Learning, $9,785 (2008)

Project Ecuador: International Learning Service Libraries – A Faculty Development Experience in Ecuador. Sylvia D. Hall-Ellis, Office of Internationalization, University of Denver, $600 (2007)

Project Ecuador: International Learning Service Libraries – A Faculty Development Experience in Ecuador. Sylvia D. Hall-Ellis, International Service Learning Office, University of Denver, $400 (2007)

Project Homeless Connect 5 Event Evaluation. Research team: Sylvia D. Hall-Ellis and Duan Zhang. University of Denver Center for Community Engagement and Service Learning, $1,000 (2007)

Project Homeless Connect 4 Event Evaluation. Research team: Sylvia D. Hall-Ellis and Duan Zhang. University of Denver Center for Community Engagement and Service Learning, $4,595 (2007)

Future LEADers of America: Leaders II. Research team: Denver Public Library (Kristen Svendson, PI), the University of Denver (Sylvia D. Hall-Ellis), and the Colorado Chapter of REFORMA (Orlando Archibeque). Laura Bush's Recruiting Librarians for the 21st Century Program, Institute for Museums and Library Services, $988,518 (2007-2010)

Strategic Planning Assistance for the Denver Medical Library. Sylvia D. Hall-Ellis. Denver Medical Library, Inc., $30,000 (2006-2007)

Destiny Software for JMAC Student Lab. Sylvia D. Hall-Ellis, gift from the Sagebrush Corporation, Minneapolis, Minnesota, $10,000 (2006)

Libraries: Tools for Education and Development Worldwide: A Faculty Development Experience in France. Sylvia D. Hall-Ellis, Office of Internationalization, University of Denver, $800 (2005)

Denver Public Library: Opportunities for Change. Sylvia D. Hall-Ellis, Colorado Community Based Research Network, $2,000 (2005)

Future LEADers of America. Research team: Denver Public Library (Letty Icolari and Steve Taylor, PIs), Emporia State University (Jim Agee), and the University of Denver (Sylvia D. Hall-Ellis). Recruiting Librarians for the 21st Century Program, Institute for Museums and Library Services, $670,315 (2005-2008)

American Association of University Professors Summer Institute Professional Development Grant. Sylvia D. Hall-Ellis, American Association of University Professors, $300 (2005)

Beta Phi Mu Alumni Tea. Sylvia D. Hall-Ellis, gift from the Office of Alumni and Parent and Relations, University of Denver, $2,000 (2005)

Spectrum Software for LIS Student Lab. Sylvia D. Hall-Ellis, gift from the Sagebrush Corporation, Minneapolis, Minnesota, $5,000 (2005)

Developing Research Capacity in Community Organizations and Residents through Training and Technical Assistance. Nicholas J. Cutforth (PI) and Sylvia D. Hall-Ellis. Piton Foundation, $25,000 (2005-2006)

Libraries: Tools for Education and Development Worldwide: A Faculty Development Experience in Argentina. Sylvia D. Hall-Ellis, Office of Internationalization, University of Denver, $500 (2004)

Developing Research Capacity in Community Organizations and Residents through Training and Technical Assistance. Research team: Nicholas J. Cutforth (PI), Gary Lichtenstein and Sylvia D. Hall-Ellis, Piton Foundation, $25,000 (2003-2004)

Increasing Spanish-Speaking and Hispanic Diversity among Library and Information Science Students at the University of Denver: Development of a Student Recruitment Model. Researcher: Sylvia D. Hall-Ellis (PI), Office of Multicultural Excellence, University of Denver, $2,750 (2003-2004)

Collection Development Enrichment to Support the Cataloging & Classification Specialization and School Librarianship within the Library & Information Science Program at the University of Denver. Research team: Sylvia D. Hall-Ellis (PI) and Deborah S. Grealy. Women's Library Association, University of Denver, $4,000 (2003)

Collection Development Enrichment to Support the Library & Information Science Program at the University of Denver. Research team: Deborah S. Grealy (PI) and Sylvia D. Hall-Ellis. Women's Library Association, University of Denver, $4,000 (2001)

Wireless LAN for Library Education. Research team: Deborah S. Grealy (PI) and Sylvia Hall-Ellis. University of Denver, Center for Teaching and Learning, $24,000 (2001-2002)

Upward Bound – A Program for South Texas Youth. Research team: Monte Churchill, Executive Director of Community Relations; Gary R. Sauceda, Outreach Coordinator; Raymond Hernandez, (PI) Associate Dean for Student Success; and Sylvia D. Hall-Ellis, South Texas Community College (McAllen, Texas). U.S. Department of Education, Office of Postsecondary Education, $2,672,003 (1999-2003)

Write Now! Improving Elementary Students' Writing Skills. Research team: Caryl G. Thomason, Assistant Superintendent (PI); Karen Tankersley, Principal; and, Hal Anderson, Director of Technology, Cheyenne Mountain School District 12 (Colorado Springs, Colorado); and Sylvia D. Hall-Ellis. Colorado Department of Education, Educational Telecommunications Unit, $189,453 (1999-2001)

Operation Quick Start -- Distance Learning in Rural Colorado. Research team: Randal Weigum, Technology Coordinator (PI); Bonnie Barns, Director of Federal Programs and Staff Development; and, Adam "Joe" Raskop, Executive Director, Southeastern Board of Cooperative Educational Services (Lamar, Colorado); and Sylvia D. Hall-Ellis. Colorado Department of Education, Educational Telecommunications Unit, $400,000 (1999-2001)

Advanced Technological Training for Information Professionals for the 21st Century. Research team: Mario Reyna, Division of Business Director (PI) and Sylvia D. Hall-Ellis, South Texas Community College (McAllen, Texas). Phi Delta Kappa and the National Science Foundation, $250,000 (1999-2001)

Working Connections – Training Information Technologies Professionals for the 21st Century. Research team: Mario Reyna, Division of Business Director (PI) and Sylvia D. Hall-Ellis, South Texas Community College (McAllen, Texas). Microsoft Corporation and the American Association of Community Colleges, $1,147,775(1999-2001)

The ROAD (Research Oriented Amplification of Development to Literacy) Program. Research team: Deborah J. Leong (PI), Metropolitan State University; Elena Bodrova, Metropolitan State University; Dmitri Semenov, Robert J. Marzano, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). Hewlett-Packard Foundation, $25,000 (1998-1999)

Press to Literacy. Research team: Deborah J. Leong (PI), Metropolitan State University; Elena Bodrova, Metropolitan State University; Dmitri Semenov, Robert J. Marzano, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). The Denver Post and the Robert S. McCormick Foundation, $49,818 (1998-1999)

International Telementor Center. Research team: David Neils, David B. Frost (PI), and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). Hewlett-Packard Philanthropy, $100,000 (1998-1999)

America Reads: Providing Tutor Training for the America Reads Challenge. Research team: Louis F. Cicchinelli (PI) and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). U.S. Department of Education, Office of Educational Research and Improvement, $306,000 (1998-1999)

Comprehensive School Reform. Research team: Louis F. Cicchinelli (PI), J. Timothy Waters, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). U.S. Department of Education, Office of Educational Research and Improvement, $285,000 (1998-1999)

Sustainable Energy Education (SEE) Program for Grades 4-8: Preparing Today's Youth for Lifelong Learning and Responsible Actions in Energy Conservation. Research team: Mary Gromko, Colorado Department of Education; Gene McCarthy, Rocky Flats Field Office, U.S. Department of Energy; Gina Kissell, National Renewable Energy Laboratory; and Barbara L. McCombs (PI), Janet L. Bishop, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). State of Colorado Governor's Office for Energy Conservation, $499,936 (1997-1999)

Review of Kansas Curriculum Standards in Mathematics and Language Arts. Research team: Robert J. Marzano (PI), John S. Kendall, David B. Frost, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). Submitted to the Kansas State Department of Education, $25,903 (1998)

International Telementor Center. Research team: David Neils, John Kuglin, Chris Rapp, David B. Frost (PI), and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). Hewlett-Packard Company, $54,752 (1998)

Genesis Mission: Education Public Outreach. Developed in partnership by the Jet Propulsion Laboratory, California Institute of Technology, Los Alamos National Laboratory, Lockheed Martin Astronautics, and the Mid-continent Regional Educational Laboratory (Aurora, Colorado). Research team: John T. Sutton (PI), Alice Krueger, Martha Henry, Greg Rawls, Shae Isaacs, Jeff Johnson, Deb Jordan, Arlene Mitchell, David B. Frost, Jana Caldwell, J. Timothy Waters, and Sylvia D. Hall-Ellis. National Aeronautics and Space Administration, $139,700,000; subcontract award, $4,750,000 (1997-2007)

North Dakota Mathematics Assessment Project. Developed in partnership by the North Dakota Department of Education and Mid-continent Regional Educational Laboratory (Mid-continent Regional Educational Laboratory). Research team: Ann Clapper, Clarence Bina, Greg Gallagher, North Dakota Department of Education; Don Burger (PI), Hillary Michaels, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). U.S. Office of Education, $1,618,214; subcontract award, $389,076 (1997-2001)

Pacific Resources for Education and Learning Distance Education: Project Evaluation. Developed in partnership by the Pacific Educational Community, the Pacific Resources for Education and Learning (PREL), and Mid-continent Regional Educational Laboratory (McREL). Research team: John W. Kofel (PI), Executive Director, PREL; J. Timothy Waters, Executive Director; Joan Buttram, Robert Keller, and Sylvia D. Hall-Ellis, Mid-continent Regional Educational Laboratory (Aurora, Colorado). U.S. Office of Education, Star Schools Program, $10,000,000; subcontract award, $500,000 (1997-2002)

Identification of Bilingual Gifted and Talented Children: A Comprehensive School Grants for Bilingual Education. Developed in partnership by Hidalgo (Texas) ISD, Los Fresnos (Texas) CISD, Progreso (Texas) ISD, University of Texas – Pan American (Edinburg, Texas), and Education Service Center, Region One (Edinburg, Texas). Research team: Hilda Medrano, Dean, College of Education, University of Texas – Pan American; Linda Phemister (PI), Janie Navarro, and Sylvia D. Hall-Ellis, Education Service Center, Region One. U.S. Department of Education, Office of Bilingual Education and Minority Languages Affairs, $1,670,633 (1997-2002)

Academics 2000: First Things First -- The Texas Goals 2000. Developed for Jim Hogg (Hebbronville, Texas) County ISD, Mirando City (Texas) ISD, and San Isidro (Texas) ISD. Research team: Hilda Medrano, Dean, College of Education, University of Texas - Pan American (Edinburg, Texas); Angie Lehmann, Amy Mares, Ellen Gonzalez (PI), and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg, Texas). Texas Education Agency, $339,987 (1997-2000)

Southwestern Bell's Learning Communities Initiative. Developed for Tech Prep of the Rio Grande Valley, Inc. (Harlingen, Texas); the Center for Professional Teacher Development, University of Texas – Brownsville; Teach for America – Rio Grande Valley (McAllen, Texas); and Education Service Center, Region One (Edinburg, Texas). Research team: Patricia G. Bubb, Executive Director (PI), Tech Prep of the Rio Grande Valley; Martin Winchester, Executive Director, Teach for America – Rio Grande Valley; Aileen Johnson, Director, School of Education, University of Texas – Brownsville; and Sylvia D. Hall-Ellis. Southwestern Bell Foundation, $100,000 (1997-1998)

Texas School to Work: Regional Implementation. Developed for Tech Prep of the Rio Grande Valley, Inc. (Harlingen), South Texas Community College (McAllen), Texas State Technical College (Harlingen), Texas Southmost College (Brownsville), Empowerment Zone of the Rio Grande Valley (Mercedes), Project VIDA (Weslaco), Youth Fair Chance (McAllen), and Education Service Center, Region One (Edinburg). Research team: Patricia G. Bubb (PI), Tech Prep of the Rio Grande Valley; Stephen Vassberg, Texas State Technical College; Ellen Trevino, Youth Fair Chance; Wanda Garza, Project VIDA; Leonardo Olivares, University of Texas – Pan American; Michael Bell, South Texas Community College; and Sylvia D. Hall-Ellis. Texas Workforce Commission, $4,250,000 (1997-2002).

Comprehensive Bilingual Education Grant for Hidalgo ISD and Roma ISD. Developed for Hidalgo (Texas) ISD, Roma (Texas) ISD, College of Education, University of Texas – Pan American (Edinburg, Texas), and Education Service Center, Region One (Edinburg, Texas). Research team: Tomas Thomas (PI), Director, Office of Bilingual Education and Sylvia D. Hall-Ellis. U.S. Department of Education, Office of Bilingual Education and Minority Languages Affairs, $1,531,361 (1997-2002)

Lopez High School, Porter High School, Rivera High School, Central Middle School, and Perkins Middle School, Brownsville (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $1,473,611 (1997-1998)

Donna High School, Todd Middle School, and Solis Middle School, Donna (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $704,052 (1997-1998)

Memorial Middle School and Nellie Schunior Middle School, La Joya Independent School District (Texas). Telecommunications Infrastructure Fund Board, $293,641 (1997-1998)

Martin High School, Christen Middle School, Cigarroa Middle School, Lamar Middle School, Laredo (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $1,200,000 (1997-1998).

Lasara Middle School, Lasara (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $246,210 (1997-1998)

Los Fresnos High School and Resaca Middle School, Los Fresnos (Texas) Consolidated Independent School District. Telecommunications Infrastructure Fund Board, $500,000 (1997-1998)

Lyford Junior High School, Lyford (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $354,997 (1997-1998)

Travis Middle School, McAllen (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $269,139 (1997-1998)

Mission High School, Mission (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $297,010 (1997-1998)

Progreso High School, Progreso (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $300,000 (1997-1998)

San Perlita High School, San Perlita (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $293,490 (1997-1998)

Myra Green Junior High School, Raymondville (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $225,000 (1997-1998)

Rio Grande City High School, Ringgold Middle School, Gruella Middle School, Rio Grande City (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $809,934 (1997-1998)

Rio Hondo Junior High School, Rio Hondo (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $297,024 (1997-1998)

Roma High School, Roma (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $333,178 (1997-1998)

San Benito High School, Miller Jordan Junior High School, San Benito (Texas) Consolidated Independent School District. Telecommunications Infrastructure Fund Board, $547,000 (1997-1998)

Santa Maria High School and Santa Maria Middle School, Santa Maria (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $400,000 (1997-1998)

Salvador High School, United Independent School District (Laredo, Texas). Telecommunications Infrastructure Fund Board, $250,879 (1997-1998)

Weslaco High School, Cabaza Middle School, Cuellar Middle School, Weslaco (Texas) Independent School District. Telecommunications Infrastructure Fund Board, $874,242 (1997-1998)

Principals' Assessment and Training Center. Developed for Education Service Center, Region One (Edinburg, Texas). Research team: Roberto Zamora, Leonel Barrera, William H. Parry, and Sylvia D. Hall-Ellis. Texas Principals' Leadership Initiative, Texas Association of Secondary School Principals, Texas Association of Elementary School Principals, and the Sid Richardson Foundation. $461,439 (1996-1998)

The Texas Library Connection -- Integrating and Sharing Resources. Developed for Hidalgo County Library System (McAllen), Cameron County Library System (Brownsville), South Texas Community College (McAllen), the University of Texas - Pan American (Edinburg), University of Texas – Brownsville, and Education Service Center, Region One (Edinburg). Research team: William R. McGee, Coordinator, Hidalgo County Library System; Joe Garcia, Director, Cameron County Library System; Michael D. Bell, Library Director, South Texas Community College; Eleanor Folger Foster, University Library, University of Texas - Pan American; Jaime Vela (PI), Director of Instructional Technology and Media Services, Ron Pontius, Instructional Technology, and Fabiola Fuentes, Media Services, Education Service Center, Region One; and Sylvia D. Hall-Ellis. Office of Library Media Services, Technology Services, Texas Education Agency. $8,000 (1996-1997)

Developing Leadership Communities for Improving Algebra I for All Students. Developed for a partnership of the Region One Statewide Systemic Initiative Team. Research team: Noel Villarreal (PI), Eduardo Cancino, and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Statewide Systemic Initiative for Reform in Mathematics, Science and Technology Education to The Charles A. Dana Center for Mathematics and Science Education, The University of Texas at Austin. $25,000 (1996-1997)

Academics 2000: First Things First -- The Texas Goals 2000. Developed for Edinburg (Texas) CISD. Research team: Hilda Medrano, Dean, College of Education, University of Texas - Pan American (Edinburg); Helen Jones, Director (PI), Gifted and Talented Education, Edinburg CISD; and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Education Agency, $750,000 (1996-2001)

Texas Teachers Empowered for Achievement in Mathematics (TEXTEAM) Institute for Algebra I. Developed for Hidalgo ISD, Jim Hogg (Hebbronville) County ISD, La Joya ISD, La Villa ISD, Lyford ISD, Mission CISD, Pharr-San Juan-Alamo ISD, Rio Hondo ISD, San Isidro ISD, San Perlita ISD, Santa Maria ISD, Santa Rosa, Sharyland ISD, Valley View ISD, Webb (Laredo) CISD, Weslaco ISD, and Zapata County (Zapata) ISD. Research team: Noel Villarreal (PI), Chuck McInteer, Ellen M. Gonzalez, Education Service Center, Region One (Edinburg); and Sylvia D. Hall-Ellis. Charles A. Dana Center for Mathematics and Science Education, University of Texas - Austin, $15,975 (1996)

Community Learning Center for La Villa, Texas. Developed for Edcouch-Elsa Independent School District. Research team: Noe Gonzalez (PI), Assistant Superintendent, Edcouch-Elsa Independent School District, and Sylvia D. Hall-Ellis. Delta Region Subzone, Rio Grande Valley (Texas) Rural Empowerment Zone, $325,000 (1996-1997)

Monte Alto (Texas) Community Learning Center. Developed for Monte Alto Independent School District. Research team: Homero A. Diaz (PI), Superintendent, Monte Alto Independent School District, and Sylvia D. Hall-Ellis. Delta Region Subzone, Rio Grande Valley (Texas) Rural Empowerment Zone, $300,000 (1996-1997)

Community Learning Center for La Villa, Texas. Developed for La Villa Independent School District. Research team: Sam Gonzalez (PI), Assistant Superintendent, La Villa Independent School District; and Sylvia D. Hall-Ellis. Delta Region Subzone, Rio Grande Valley (Texas) Rural Empowerment Zone, $325,000 (1996-1997)

Building Project for Taylor Elementary. Developed for Mercedes (Texas) Independent School District. Research team: Mrs. Denise Rivera, Librarian, and Eduardo Infante, Principal, Taylor Elementary School; Ismael S. Cantu (PI), Federal Programs Director, Mercedes Independent School District; and Sylvia D. Hall-Ellis. Delta Region Subzone, Rio Grande Valley (Texas) Rural Empowerment Zone, $375,000 (1996-1997)

Border Education Network (BEN) - Distance Education through Cable Television. Developed for Edinburg (Texas) Consolidated Independent School District. Research team: Noe Torres (PI), Library Media Specialist, Magdalena Rosas, Library Media and Technology Coordinator, Edinburg Consolidated Independent School District; and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Edinburg (Texas) Consolidated School District, $250,000 (1996-1997)

Multiservice One-Stop Open Enrollment Charter School. Developed for the Information Referral Resource Assistance, Inc. (McAllen, Texas). Research team: Pablo Perez and Aguie Pena (PI), Executive Director, Information Referral Resource Assistance, Inc.; Roberto Zamora and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Education Agency, $3,066,000 (1996-2001)

Project OK: A Community Youth Opportunities Grant for Summer, 1996. Developed for McAllen (Texas) Independent School District, St. Joseph the Worker Catholic Church, and the Office of Adult Education, Education Service Center, Region One. Research team: Father Bart Flatt, St. Joseph the Worker Catholic Church; Noe Calvillo (PI), Office of Adult Education; Maria Louisa Garcia, McAllen Independent School District; and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Protective and Regulatory Agency through the Office of the Mayor, City of McAllen, $350,000 (1996-2001).

Early Childhood: a Time of Discovery. Developed for a partnership of Lasara ISD, Rio Hondo ISD, San Perlita ISD, the School of Education, University of Texas - Brownsville, and Education Service Center, Region One (Edinburg). Research team: Hugo Rodriguez, Dean, School of Education, University of Texas - Pan American; Leonel Barrera and Jack Damron, Field Service Agents, Ellen M. Gonzalez (PI), Administrator for Student Instructional Services, Ruth Solis, Education Specialist in Special Education, and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Education Agency, $722,090 (1996-2001)

Innovative Gifted and Talented Programs for Early Childhood and Elementary Education Students. Developed for a partnership of Edinburg (Texas) CISD, the School of Education, University of Texas - Pan American (Edinburg), and Education Service Center, Region One (Edinburg). Research team: Hilda Medrano, Dean, School of Education, University of Texas - Pan American; Helen de la Garza (PI), Director of Elementary Curriculum and Instruction, Edinburg CISD; and Sylvia D. Hall-Ellis. Texas Education Agency, $750,000 (1996-2001)

Reading Recovery in Early Elementary Grades. Developed for a partnership of La Villa (Texas) ISD the School of Education, University of Texas - Pan American (Edinburg), and Education Service Center, Region One (Edinburg). Research team: Hilda Medrano, Dean, School of Education, University of Texas - Pan American; Marcario Salinas (PI), Supervisor of Elementary Curriculum and Instruction, La Villa ISD; and Sylvia D. Hall-Ellis, Education Service Center, Region One. Texas Education Agency, $750,000 (1996-2001)

Connected Mathematics Project for Middle and Junior High Students. Developed for a partnership of Michigan State University, The Charles A. Dana Center for Mathematics and Science Education, The University of Texas at Austin and Region One Statewide Systemic Initiative Team. Research team: Jack Damron, Chuck McInteer, Noel Villarreal (PI), and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). National Science Foundation through Michigan State University to the Charles A. Dana Center at the University of Texas - Austin, $539,610 (1996-1999)

Sharing Resources: Testing the Interlibrary Loan Potential of the Texas Library Connection -- Integrating Media Resources. Developed under the sponsorship of the Hidalgo County Library System (McAllen) and Education Service Center, Region One (Edinburg). Research team: William H. McGee, Hidalgo County System Coordinator; Fabiola Fuentes, Library Media; Ronald Pontius (PI), Instructional Technology, Education Service Center, Region One; and Sylvia D. Hall-Ellis. Library Media Services Program, Office of Technology Services, Texas Education Agency, $25,000 (1996-1997)

State and Federal Adult Education JOBS Program. Developed for Education Service Center, Region One (Edinburg). Research team: Noe Calvillo (PI), Director, Adult Education Program, and Sylvia D. Hall-Ellis. Adult Education Program, Texas Education Agency, $532,846 (1995-1996)

Creating Safe and Drug-Free Schools and Communities. Developed for a partnership of the Region One Consortium for Safe and Drug-Free Education Environments. Research team: Clara Contreras (PI), Health Specialist, and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Education Agency, $150,000 (1995-1996)

Developing Leadership Communities for Improving Mathematics Performance for All Students on Title I Campuses. Developed for a partnership of the Region One Statewide Systemic Initiative Team. Research team: Jack Damron, Chuck McInteer, Noel Villarreal (PI), and Sylvia D. Hall-Ellis, Education Service Center, Region One (Edinburg). Texas Statewide Systemic Initiative for Reform in Mathematics, Science and Technology Education to The Charles A. Dana Center for Mathematics and Science Education, The University of Texas at Austin, $50,669 (1995-1996)

The Impact of Library Resource Centers on Academic Achievement in Selected Public Schools in South Texas. Sylvia D. Hall-Ellis (PI). Developed under the sponsorship of the Department of Library Science, College of Education and Applied Science, Sam Houston State University (Huntsville). Texas Association of School Librarians, Children's Services Round Table, and Young Adults Round Table (Austin, Texas), $2,500 (1995)

School Library Media Specialists Fellowship Program. Sylvia D. Hall-Ellis (PI). Developed under the sponsorship of the Department of Library Science, College of Education and Applied Science, Sam Houston State University (Huntsville, Texas). U.S. Department of Education, HEA Title II-B, Library Education and Human Resource Development Program, $44,000 (1995-1996)

Planning for Educational Technology. Research team: Ruth Ann Riggins (PI), Director of Library Media and Technology Services, Donna (Texas) Independent School District; Noe Torres, Education Service Center, Region One; Patricia G. Bubb, Tech Prep of the Rio Grande Calley, Inc; Michael D. Bell, South Texas Community College; and Sylvia D. Hall-Ellis, Sam Houston State University. Developed under the sponsorship of Donna (Texas) Independent School District, Education Service Center, Region One (Edinburg), Tech Prep of the Rio Grande Valley, Inc. (Harlingen), South Texas Community College (McAllen), and the Department of Library Science, Sam Houston State University (Huntsville). Funded through Infusion of Educational Technology Planning Grant Program (H.B. 18: Models). Texas Education Agency, $18,000 (1993-1994)

The Impact of Library Resource Centers on Academic Achievement in Selected Public Schools in South Texas. Sylvia D. Hall-Ellis (PI) and Mary Ann Berry. Developed under the sponsorship of the Department of Library Science, College of Education and Applied Science, Sam Houston State University (Huntsville). Sam Houston State University Research Enhancement Fund, $7,500 (1993-1994)

Electronic Mail Resource Sharing System: Management and Operation of the Iowa Computer-Assisted Network. State Library of Iowa, $294,000 (1986-1993)

The Iowa Locator: A CD-ROM Resource Sharing Tool. State Library of Iowa, $567,000 (1986-1992)

Statewide Database Development: An OCLC Tape Analysis To Determine Feasibility. State Library of South Dakota, $15,000 (1987)

The Iowa Locator: A Feasibility Study. State Library of Iowa, $50,000 (1986)

Electronic Mail Resource Sharing System: A Feasibility Study for Libraries in the State of Iowa. State Library of Iowa, $5,000 (1985)

Access Pennsylvania: A Feasibility Study. State Library of Pennsylvania, $50,000 (1984)

Automating Library Administrative and Management Tasks: Procurement and Distribution of Microcomputer Systems for District Library Centers in the Commonwealth of Pennsylvania. State Library of Pennsylvania, $495,000 (1982)

Sharing Serial Titles Resources: Procurement and Distribution of Microfiche Readers for 550 Libraries in the Commonwealth of Pennsylvania. State Library of Pennsylvania, $500,000 (1982)

Electronic Mail System: A Pilot Project for Libraries throughout the Commonwealth of Pennsylvania. State Library of Pennsylvania, $125,000 (1982)

Literacy Program for Adults in Rural Upstate New York State: Program Outreach and Evaluation - Phase 3. Appalachian Regional Commission, $50,000 (1981)

Database Building: A Cooperative Project of the Finger Lakes Library System (Ithaca), Four County Library System (Binghamton), and the Southern Tier Library System (Corning). New York State Library, $450,000 (1980-1985)

Faces of the Southern Tier: A Professional Photographer-in-Residence. New York State National Endowment for the Humanities, $25,000 (1980)

Literacy Program for Adults in Rural Upstate New York State: Program Implementation - Phase 2. Appalachian Regional Commission, $50,000 (1980)

Small Business Resources Center: A Pilot Project for Rural Public Libraries. New York State Library, $50,000 (1980)

Media Programming: A Professional Development Program for Public Librarians in Upstate New York: A Program in Allegheny, Chemung, Schuyler, Steuben, and Yates Counties. New York State Library, $35,000 (1979)

Information and Library Resources for Inmates and Prisoners in Selected Upstate New York Facilities: A Cooperative Program in Allegheny, Chemung, Schuyler, Steuben, and Yates Counties. New York State Library, $25,000 (1979)

Library Resources for Homebound Adults in Upstate New York: An Outreach Program in Allegheny, Chemung, Schuyler, Steuben, and Yates Counties. New York State Library, $40,000 (1979)

Literacy Program for Adults in Rural Upstate New York State: Program Initiation and Establishment - Phase 1. Funded by the Appalachian Regional Commission. Awarded to the Corning (New York) Public Library, $50,000 (1979)

Information Reference Services to Homebound Adults. New York State Library, $31,000 (1978)

Books-By-Mail Services to Homebound Adults. New York State Library, $31,000 (1978)

System Headquarters Services and Programs for Public Libraries in District 10. Funded through Library Services and Construction Act Title I. Texas State Library and Historical Commission, $880,000 (1976)

County Library Development Grant for Atascosita County. Funded through Library Services and Construction Act Title I. Texas State Library and Historical Commission, $15,000 (1976)

Spanish Language Materials Collection Development Program. Texas State Library and Historical Commission, $40,000 (1975)

Establishment of System Headquarters for Alamo Regional Library System (District 10) Headquartered at the San Antonio (Texas) Public Library. Texas State Library and Historical Commission, $800,000 (1975)

SELECTED CONSULTANTSHIPS

Project Strategic Planning & Funding Proposal Development

Walden University, 2015.
Pikes Peak Library District, 2014.
Douglas County Public Libraries, 2011-2012.
Denver School for Science and Technology, 2008-2011.
Johnson & Wales University, Denver Campus, 2007-2009.
Challenges, Choices, and Images K-12 Charter School, Denver Public Schools, 2007.
Denver Medical Library, Inc., 2006-2010.
Bemis Public Library (Littleton), Ergonomic Design and Facilities Enhancement Consultant, 2003.
Curtis Park Community Center (Denver), Community Technology Center Evaluation and Proposal Development Consultant, 2003.
Colorado Community Based Research Network (Denver), Funding Research Associate, 2002.
Our Lady of the Rosary Academy (Edgewater), Learning Resource Center Development Project Consultant, 2001-2003
University of Southern Colorado (Pueblo). Technology Integration and Curriculum Enhancement into Higher Education Learning Environment. Proposal Development Consultant, 2000.
Jefferson County Library System (Lakewood), 1999-2000.
Southeastern BOCES (Lamar, CO). Distance Learning Curriculum Content Development Project. Proposal Development Consultant, 1998-2004.
Cheyenne Mountain School District 12 (Colorado Springs, CO). Technology Integration into Elementary Writing Curriculum Project. Proposal Development Consultant, 1998.
Telecommunications Infrastructure Fund Board Round #2 Application, 1997. Technical assistance to the following: La Villa (Texas) Independent School District; Mirando City (Texas) Independent School District; Monte Alto (Texas) Independent School District; San Isidro (Texas) Independent School District.
Telecommunications Infrastructure Fund Board Round #1, 1996. Technical assistance to the following: Brownsville (Texas) Independent School District; Donna (Texas) Independent School District; Edinburg (Texas) Consolidated Independent School District; Harlingen (Texas) Consolidated Independent School District; Jim Hogg County Independent School District (Hebbronville, Texas); La Feria (Texas) Independent School District; La Joya (Texas) Independent School District; La Villa (Texas) Independent School District; Los Fresnos (Texas) Consolidated Independent School District; Progreso (Texas) Independent School District; Raymondville (Texas) Independent School District; Rio Hondo (Texas) Independent School District; San Benito (Texas) Independent School District; Sharyland Independent School District (Mission, Texas); Weslaco (Texas) Independent School District; Zapata (Texas) County Independent School District.
South Texas Community College (McAllen). Strategy Development to Meet the Technology Instructional Needs for Vocational, School-To-Work and Academic Programs, 1994-2000. Development Consultant.
Donna (Texas) Independent School District. Planning for the Infusion of Technology in Middle Schools Serving High At-Risk Students, 1994-1995. Research Associate.
Pettus (Texas) Independent School District. Planning for District-wide Automation in School Library Media Centers: Preparation of Data, and CD-ROM Hardware/Software Configuration Evaluation, 1994. Principal Investigator.
Fort Bend Independent School District (Sugar Land, TX). Planning for District-wide Automation in School Library Media Centers, Preparation of Data, and CD-ROM Hardware/Software Configuration Evaluation, 1994-1995. Principal Investigator.
Northern Waters Library Services (Ashland, WI). Professional Consulting Services to Make Recommendations on the Development of System-Wide Efforts of the Northern Waters Library Service, 1988-1989. Research Associate.
State Library of Iowa (Des Moines). The Iowa Locator Compact Disc to Support Multi-Type Libraries Resource Sharing including Database Design, Building, Production, and Distribution, 1986-1991. Project Director.
State Library of Iowa (Des Moines). Iowa Computer-Assisted Network Development, Enhancement, and Operation, 1985-1992. Technical Director.
State Library of South Dakota (Pierre). Statistical Sampling and Analysis of Multi-Institutional OCLC Archive Tape for Statewide Online Database Building, 1985-1986. Research Associate.
District of Columbia Public Library (Washington, DC). Planning Document for the Retrospective Conversion of Bibliographic Records and Automation Issues in the 1990's, 1984-1985. Research Associate.

State Library of Pennsylvania (Harrisburg). Database Development for High School Libraries, 1984-1985. Research Associate.
MINITEX (Minneapolis, MN). Workshops for Retrospective Conversion, Bar Coding, Library Statistics, and OCLC Serials Format, 1985. Presenter and Research Associate.
ABC Film Consortium (Altoona, Bellefonte, Johnstown, PA). Database Building, System Design, and Implementation of Online Film Booking System, 1983-1985. Database Manager.

Library Automation -- System Design & Implementation

Douglas County Library District (Castle Rock, CO). Library Automation Technical and System Performance Specifications, 2001.
Pharr-San Juan-Alamo Independent School District (Pharr, TX). Library Automation Technical and System Performance Specifications, 1997. Project Development Consultant.
Harlingen (Texas) Independent School District. Planning for District-wide Automation in Junior High School Library Media Centers: Preparation of Data, and CD-ROM Hardware/Software Configuration Evaluation, 1996. Principal Investigator.
Sharyland (Texas) Independent School District. Design and Construction of New Elementary School Library and Technology Resources Center, 1995-1996. Principal Investigator.
South Texas Independent School District (Mercedes, TX). Design and Construction of a New High School Library and Technology Resources Center, 1995-1996. Principal Investigator.
Citizens' Library (Washington, PA). Online Community Resources Files: Design and Implementation, 1984. Research Associate.
Altoona (Pennsylvania) Hospital. Integrated Online System Upgrade Study, 1984. Research Associate.
State Library of South Dakota (Pierre). Automation Plan and State Database Development, 1982-1983. Principal Investigator.
State Library of Pennsylvania (Harrisburg). COM Production and Preparation of the Technical Specifications Document, Procurement, and Distribution of Microcomputers in the Commonwealth of Pennsylvania, 1983-1984. Research Associate.
Altoona (Pennsylvania) Area Public Library. Integrated System (Circulation/Online Catalog) Study, 1982-1983. Principal Investigator.
COPSCAULD (Council of Pennsylvania State College and University Library Directors, Edinboro, PA). Design of Online Media Catalog, 1982. Principal Investigator.
Erie (Pennsylvania) County Library System. Operations Research and Design for Automated Circulation, Hardware Upgrade and Re-Retrospective Conversion, 1982. Research Associate.
Michigan Library Consortium (Lansing, MI). Tape Management System Design and Implementation, 1982-1984. Research Associate.
State Library of Pennsylvania (Harrisburg). Preparation of the Technical Specifications Document, Procurement and Distribution of Microfiche Readers in the Commonwealth of Pennsylvania, 1982. Research Associate.
Community College of Allegheny County, Allegheny Campus (Pittsburgh, PA). Integrated Systems Study, 1981. Principal Investigator.
Community College of Allegheny County, South Campus (West Mifflin, PA). Retrospective Conversion Training Program for Handicapped Students, 1981. Technical Director.
Southern Tier Library System (Corning, NY). OCLC Acquisitions Subsystem Evaluation, 1981. Principal Investigator.
South Central Reference and Research Council (Ithaca, NY). South Central Regional Delivery System, 1980. Principal Investigator.
Southern Tier Library System (Corning, NY). Newspapers on Microfilm in the Chemung-Southern Tier Library System, 1979-1981. Principal Investigator.

Cataloging & Bibliographic Database Building

Challenges, Choices, and Images K-12 Charter School, Denver Public Schools, 2007.
Denver Medical Library, Inc., 2006-2010.
Ricks Gifted and Talented School Library, 2005-
American Humane Society (Englewood), 2004-2006.
Fisher Early Learning Center Library, 2004-

Colorado Community Based Research Network (Denver), 2004-
Boulder Valley School District (Boulder), 2002-2003.
Douglas Public Library District (Castle Rock), 2001-2002.

Bibliographic Center for Research, Inc. (Denver). Original Cataloging, 2001-2003 & 1992-1993. Professional Cataloger. Libraries include Arapahoe Library District (Centennial); Clarke College (Clarke, IA); Kaiser Permanente Center for Health Research (Portland, KS); Douglas County Library System (Carson City, NV); Grinnell College (Grinnell, IA); Kansas State University (Manhattan, KS); Kansas State Library (Topeka, KS); McPherson College (McPherson, KS); Newman University (Wichita, KS); Montana Technical University (Helena, MT); Montana State Library (Helena, MT); University of Texas Southwest Medical Center Library (Dallas, TX); Community College of Southern Nevada (Las Vegas, NV); Pikes Peak Library District (Colorado Springs); Mesa Community College (Grand Junction); Coe College (Cedar Rapids, IA); Briar Cliff College (Sioux City, IA); Tri-Care Health Systems (Aurora); The Penrose-St. Francis Healthcare System (Colorado Springs); Westminster Public Library (Westminster); Lutheran Medical Center Medical Library (Wheat Ridge); American Heritage Center, University of Wyoming (Laramie, WY); National Wildlife Research Center Library (Lakewood); University of Colorado Health Sciences Center, Denison Memorial Library (Denver); Water Resources Library, Denver District Office, Bureau of Land Management (Lakewood); Mesa State Community College Learning Resources Center (Pueblo); Front Range Community College Learning Resources Center (Westminster).

Colorado Historical Society Library (Denver). Original Cataloging of Serials for Multi-Institutional Union Listing Project, 1993. Technical Services Associate. Institutional partners: Denver Art Museum, Denver Museum of Natural History Library, and Denver Botanical Gardens Library.

Arapahoe Library District (Centennial). Original Cataloging of Major Media, 1991.

Original Cataloging Projects: Davis and Elkins College (Elkins, WV), 1989-1990; Waldorf College (Forest City, IA), 1989-1990; Kennametal Corporation (Latrobe, PA), 1984-1985; Pennsylvania Public Libraries Film Center (Harrisburg), 1982-1985.

Retrospective Conversion Projects: North Central Regional Library Service (Mason City, IA), 1987-1988; Southeastern Library Services (Davenport, IA), 1986-1987; Virginia Theological Seminary (Alexandria), 1984-1985; California University of Pennsylvania (California, PA), 1984-1985; Kennametal Corporation (Latrobe, PA), 1984-1985; West Virginia University (Morgantown), 1983-1985; Altoona Area Public Library (Altoona, PA), 1983-1984; George Washington University (Washington, DC), 1983-1984; Indiana University of Pennsylvania (Indiana, PA), 1983-1985; Health Education Center (Pittsburgh), 1983-1984; Central Pennsylvania District Library Center (Bellefonte, PA), 1982-1985; Duquesne University Law School Library (Pittsburgh), 1982-1983; Allegheny County Law Library (Pittsburgh), 1982-1984; Calgon Technical Information Center (Pittsburgh), 1982; Dow Corning Corporation, Technical Information Center (Midland, MI), 1982-1983; Tri-System Public Library Retrospective Conversion Project (Binghamton, Corning, and Ithaca, NY), 1978-1981, Technical Project Director.

Updated July 2025

CANCER GENOMICS

Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA

Tim Forshew,[1]* Muhammed Murtaza,[1,2]* Christine Parkinson,[1,2,3]* Davina Gale,[1]* Dana W. Y. Tsui,[1]* Fiona Kaper,[4†] Sarah-Jane Dawson,[1,2,3] Anna M. Piskorz,[1,2] Mercedes Jimenez-Linan,[3,5] David Bentley,[6] James Hadfield,[1] Andrew P. May,[4] Carlos Caldas,[1,2,3,7] James D. Brenton,[1,2,3,7‡] Nitzan Rosenfeld[1,2‡]

Plasma of cancer patients contains cell-free tumor DNA that carries information on tumor mutations and tumor burden. Individual mutations have been probed using allele-specific assays, but sequencing of entire genes to detect cancer mutations in circulating DNA has not been demonstrated. We developed a method for tagged-amplicon deep sequencing (TAm-Seq) and screened 5995 genomic bases for low-frequency mutations. Using this method, we identified cancer mutations present in circulating DNA at allele frequencies as low as 2%, with sensitivity and specificity of >97%. We identified mutations throughout the tumor suppressor gene *TP53* in circulating DNA from 46 plasma samples of advanced ovarian cancer patients. We demonstrated use of TAm-Seq to noninvasively identify the origin of metastatic relapse in a patient with multiple primary tumors. In another case, we identified in plasma an *EGFR* mutation not found in an initial ovarian biopsy. We further used TAm-Seq to monitor tumor dynamics, and tracked 10 concomitant mutations in plasma of a metastatic breast cancer patient over 16 months. This low-cost, high-throughput method could facilitate analysis of circulating DNA as a noninvasive "liquid biopsy" for personalized cancer genomics.

INTRODUCTION

Circulating cell-free DNA extracted from plasma or other body fluids has potentially transformative applications in cancer management (*1*–*7*). Characterization of tumor mutation profiles is required for informed choice of therapy, given that biological agents target specific pathways and effectiveness may be modulated by specific mutations (*8*–*11*). Yet, mutation profiles in different metastatic clones can differ significantly from each other or from the parent primary tumor (*12*, *13*). Evolutionary changes within the cancer can alter the mutational spectrum of the disease and its responsiveness to therapies, which may necessitate repeat biopsies (*14*–*17*). Biopsies are invasive and costly and only provide a snapshot of mutations present at a given time and location. For some applications, mutation detection in plasma DNA as a "liquid biopsy" could potentially replace invasive biopsies as a means to assess tumor genetic characteristics (*2*–*7*). Sensitive methods for detecting cancer mutations in plasma may find use in early detection screening (*1*), prognosis, monitoring tumor dynamics over time, or detection of minimal residual disease (*3*, *18*, *19*). In high-grade serous ovarian carcinomas (HGSOC), mutations in the tumor suppressor gene *TP53* have been observed in 97% of cases (*20*, *21*), but these are located throughout the gene and are difficult to assay. A cost-effective method that could detect and measure allele frequency (AF) of *TP53* mutations in plasma may be highly applicable as a biomarker for HGSOC (*22*).

Circulating DNA is fragmented to an average length of 140 to 170 base pairs (bp) and is present in only a few thousand amplifiable copies per milliliter of blood, of which only a fraction may be diagnostically relevant (*2*, *3*, *23*–*25*). Recent advances in noninvasive prenatal diagnostics highlight the clinical potential of circulating DNA (*25*–*28*), but also the challenges involved in analysis of circulating tumor DNA (ctDNA), where mutated loci and AFs may be more variable. Various methods have been optimized to detect extremely rare alleles (*1*, *2*, *6*, *7*, *29*–*31*), and can assay for predefined or hotspot mutations. These methods, however, interrogate individual or few loci and have limited ability to identify mutations in genes that lack mutation hotspots, such as the *TP53* and *PTEN* tumor suppressor genes (*32*). In patients with more advanced cancers, ctDNA can comprise as much as 1% to 10% or more of circulating DNA (*2*), presenting an opportunity for more extensive genomic analysis. Targeted resequencing has been recently used to identify mutations in selected genes at AFs as low as 5% (*33*–*35*). However, identifying mutations across sizeable genomic regions spanning entire genes at an AF as low as 2%, or in few nanograms of fragmented template from circulating DNA, has been more challenging.

In response, we describe a tool for noninvasive mutation analysis on the basis of tagged-amplicon deep sequencing (TAm-Seq), which allows amplification and deep sequencing of genomic regions spanning thousands of bases from as little as individual copies of fragmented DNA. We applied this technique for detection of both abundant and

[1]Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK. [2]Department of Oncology, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK. [3]Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK. [4]Fluidigm Corporation, 7000 Shoreline Court, Suite 100, South San Francisco, CA 94080, USA. [5]Department of Histopathology, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. [6]Illumina Cambridge, Chesterford Research Park, Little Chesterford, Cambridge CB10 1XL, UK. [7]Cambridge Experimental Cancer Medicine Centre, Cambridge CB2 0RE, UK.
*These authors contributed equally to this work.
†Present address: Illumina, Inc., 5200 Illumina Way, San Diego, CA 92122, USA.
‡To whom correspondence should be addressed. E-mail: nitzan.rosenfeld@cancer.org.uk (N.R.); james.brenton@cancer.org.uk (J.D.B.)

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

rare mutations in circulating DNA from blood plasma of ovarian and breast cancer patients. This sequencing approach allowed us to monitor changes in tumor burden by sampling only patient plasma over time. Combined with faster, more accurate sequencing technologies or rare allele amplification strategies, this approach could potentially be used for personalized medicine at point of care.

RESULTS

Targeted deep sequencing of fragmented DNA by TAm-Seq

To amplify and sequence fragmented DNA, we designed primers to generate amplicons that tile regions of interest in short segments of about 150 to 200 bases (Fig. 1A and table S1), incorporating universal

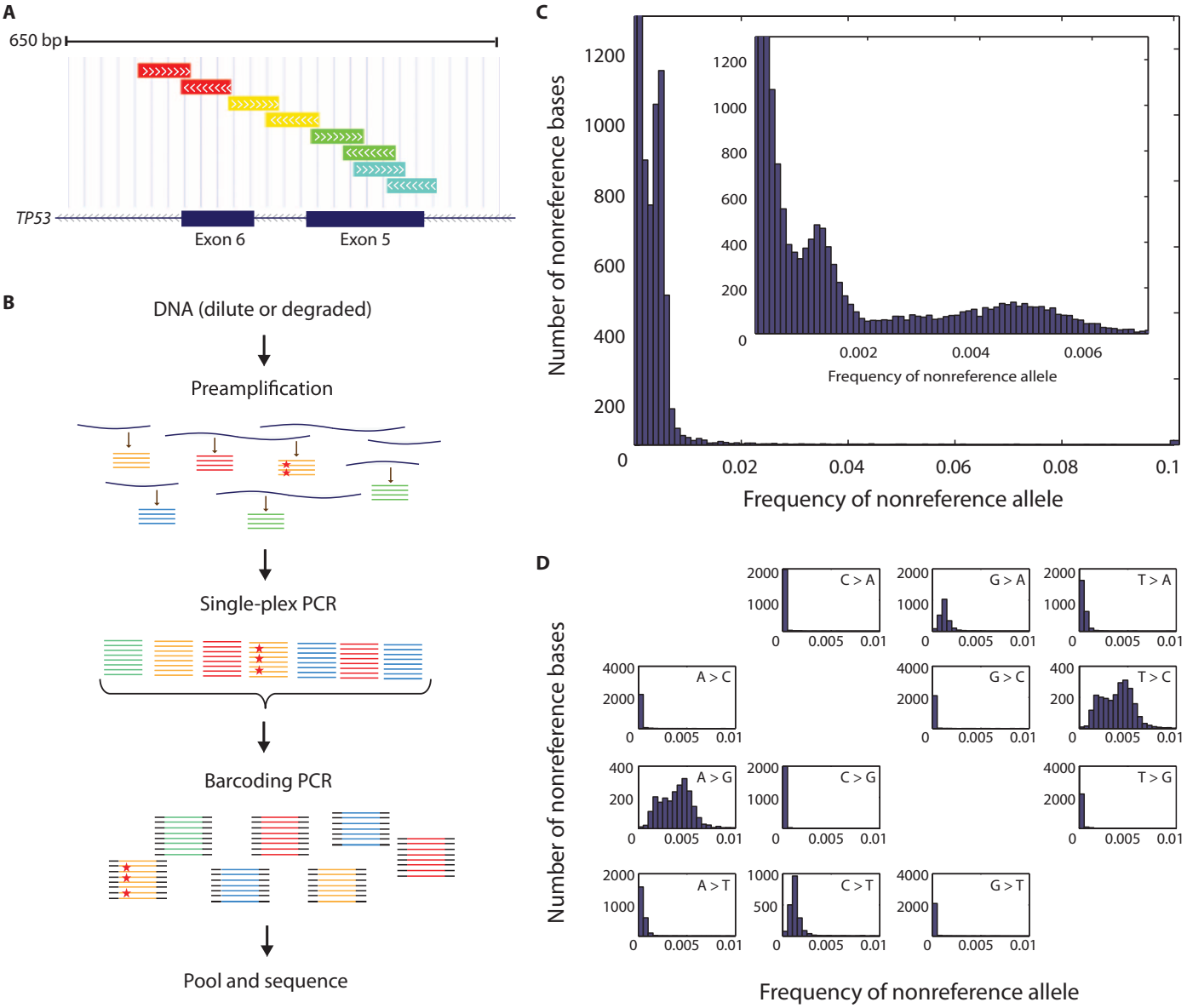


Fig. 1. Overview of tagged amplicon sequencing (TAm-Seq). (**A**) Illustration of amplicon design. Primers were designed to amplify regions of interest in overlapping short amplicons (table S1). Amplicon design is illustrated for a region covering exons 5 to 6 of *TP53*. Colored bars, segmented into forward and reverse reads, show regions covered by different amplicons (excluding primer regions). Sequencing adaptors are attached at either end, such that a single-end read generates separate sets of forward and reverse reads (fig. S1). Because amplicons are mostly shorter than 200 bp, the forward and reverse reads also partially overlap. Figure adapted from University of California, Santa Cruz, Genome Browser (http://genome.ucsc.edu/). (**B**) Workflow overview. Multiple regions were amplified in parallel. An initial preamplification step was performed for 15 cycles using a pool of the target-specific primer pairs to preserve representation of all alleles in the template material. The schematic diagram shows DNA molecules that carry mutations (red stars) being amplified alongside wild-type molecules. Regions of interest in the preamplified material were then selectively amplified in individual (single-plex) PCR, thus excluding nonspecific products. Finally, sequencing adaptors and sample-specific barcodes were attached to the harvested amplicons in a further PCR. (**C**) Distribution of observed nonreference read frequencies, averaged over 47 FFPE samples, across all loci and all nonreference bases. Inset expands the low-frequency range. (**D**) Distribution of the observed background nonreference read frequencies averaged over 47 FFPE samples for the 12 different A/C/G/T base substitutions.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

adaptors at 5′ ends (fig. S1). Performing single-plex amplification with each of these primer pairs would require dispersing the initial sample into many separate reactions, considerably increasing the probability of sampling errors and allelic loss. Multiplex amplification using a large set of primers could result in nonspecific amplification products and biased coverage. We therefore applied a two-step amplification process: a limited-cycle preamplification step where all primer sets were used together to capture the starting molecules present in the template, followed by individual amplification to purify and select for intended targets (Fig. 1B) (Supplementary Methods). The final concentration of each primer in the preamplification reaction was 50 nM, reducing the potential for interprimer interactions, and 15 cycles of long-extension (4 min) polymerase chain reaction (PCR) were used to remain in the exponential phase of amplification. We used a microfluidic system (Access Array, Fluidigm) to perform parallel single-plex amplification from multiple preamplified samples using multiple primer sets. An additional PCR step attached sequencing adaptors (fig. S1) and tagged each sample by a unique molecular identifier or "barcode" (table S2). Sequencing adaptors were separately attached at either end and the products mixed together, such that single-end sequencing generated separate sets of forward and reverse reads. We performed 100-base single-end sequencing (GAIIx sequencer, Illumina), with an additional 10 cycles using the barcode sequencing primer, generating ~30 million reads per lane. This produced an average read depth of 3250 for each of 96 barcoded samples for 48 amplicons read in two possible orientations.

Validation and sensitivity for mutation identification in ovarian tumor samples

We designed a set of 48 primer pairs to amplify 5995 bases of genomic sequence covering coding regions (exons and exon junctions) of *TP53* and *PTEN*, and selected regions in *EGFR*, *BRAF*, *KRAS*, and *PIK3CA* (table S1) by overlapping short amplicons (Fig. 1A). The sequenced regions cover mutations that account for 38% of all point mutations in the COSMIC database (v55) (*32*). We used TAm-Seq to sequence DNA extracted from 47 formalin-fixed, paraffin-embedded (FFPE) tumor specimens of ovarian cancers (table S3), which were also sequenced for *TP53* by Sanger sequencing (*36*) (Supplementary Methods). DNA extracted from FFPE samples is generally degraded and fragmented as a result of fixation and long-term ambient storage. We amplified DNA from each sample in duplicate, tagging each replicate with a different barcode. Using a single lane of sequencing, we generated 3.5 gigabases of data passing signal purity filters, producing mean read depth of 3200 above Q30 for each of the 9024 expected read groups (48 amplicons × 2 directions × 94 barcoded samples). Background frequencies of nonreference reads were ~0.1% (median, 0.03%; mean, 0.2%; in keeping with Q30 quality threshold applied), yet varied substantially between loci and base substitutions (Fig. 1C) and showed a clear bias toward purine/pyrimidine conservation (Fig. 1D). Sixty-six percent of loci had mean background rate of <0.1%, and 96% of loci had background rate of <0.6%.

The data set interrogated nearly 18,000 possible single-base substitutions for each sample, which introduces a risk of false detection. To control for sporadic PCR errors and reduce false positives, we called point mutations in a sample only if nonreference AFs were above the respective substitution-specific background distribution at a high confidence margin (0.9995 or greater), and ranked high in the list of nonreference AFs, in both replicates (Supplementary Methods). Duplicate sequencing data were obtained for 44 samples, and 43 single-base substitutions were called (table S3). These matched 100% of mutations identified by Sanger sequencing and included three additional mutations at low AFs that were below detection thresholds of Sanger sequencing (fig. S2). The upper bound of AFs that may have been missed was estimated (Supplementary Methods) at <5% for 36 of 44 FFPE samples (82%) and <10% for 42 of 44 samples (95%), with median value of 1.3% and mean value of 2.7%. Mutant AFs were highly reproducible in duplicate samples. For 42 of 43 mutations called, the difference in measured frequency between duplicates was less than 0.08, and the relative difference was 25% or less (Fig. 2A). Mutant AFs correlated significantly with tumor cellularity in the FFPE block (correlation coefficient = 0.422; $P = 0.0049$, t test) (Fig. 2B).

In a separate run, we sequenced libraries prepared from six different diluted mixtures of six FFPE samples, with a different known point mutation in *TP53* in each, to mean read depth of 5600. Of more than 100,000 possible non-SNP (single-nucleotide polymorphism) substitutions, we identified all 33 expected point mutations present at AF >1%, including 6 mutations present at AF <2%, with one false-positive called with AF = 1.9%. Using less stringent parameters (Supplementary Methods), we identified three additional mutations present at AF = 0.6% (Fig. 2C), with no additional false positives. Thus, we obtained 100% sensitivity, identifying mutations at AFs as low as 0.6%. A positive predictive value (PPV) of 100% was calculated for mutations at AF >2%, and a PPV of 90% for mutations identified at AF <2% (Fig. 2D).

Quantitative limitations of mutation detection

When applying TAm-Seq to measure a predefined mutation (as opposed to screening thousands of possible substitutions), the frequency of the mutant allele can be read out directly from the data at the desired locus. False detection is less likely, and criteria for confident mutation detection for a predefined substitution can be less stringent than those described above for de novo mutation identification (Supplementary Methods). The minimal nonreference AFs that could be detected depend on the read depth and background rates of nonreference reads, which vary per locus and substitution type. Minimal detectable frequencies increase when higher confidence margins are used (Supplementary Methods) and had a median value of 0.14% at confidence margin of 0.95 and 0.18% at confidence margin of 0.99 (fig. S3). The minimal detectable frequency would also be limited if a minimal number of reads is applied for confident mutation detection; for example, a minimum of 10 reads implies that sequencing depth of 5000 would be required to detect mutations at AF as low as 0.2%. For alleles present at ~10 or fewer copies in the starting template, reproducibility would also be limited by sampling noise, because these alleles may be over- or underrepresented in any particular reaction.

To characterize the quantitative accuracy of TAm-Seq as applied to circulating DNA, we simulated rare circulating tumor mutations by mixing plasma DNA from two healthy individuals. Using the same set of primers as used for the FFPE experiment, we identified that these two individuals differed at five known SNP loci (table S4). Total amplifiable copies in both plasma DNA samples were determined by digital PCR and mixed to obtain minor AFs ranging from 0.16% to 40% (Supplementary Methods). We sequenced diluted templates containing between 250 and <1 expected copy of the minor allele (table S5). The coefficient of variation (CV) of the observed AFs was equal on average to the inverse square root ($1/\sqrt{n}$) of the expected number of copies of the rare allele (Fig. 3A), which is the theoretical

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

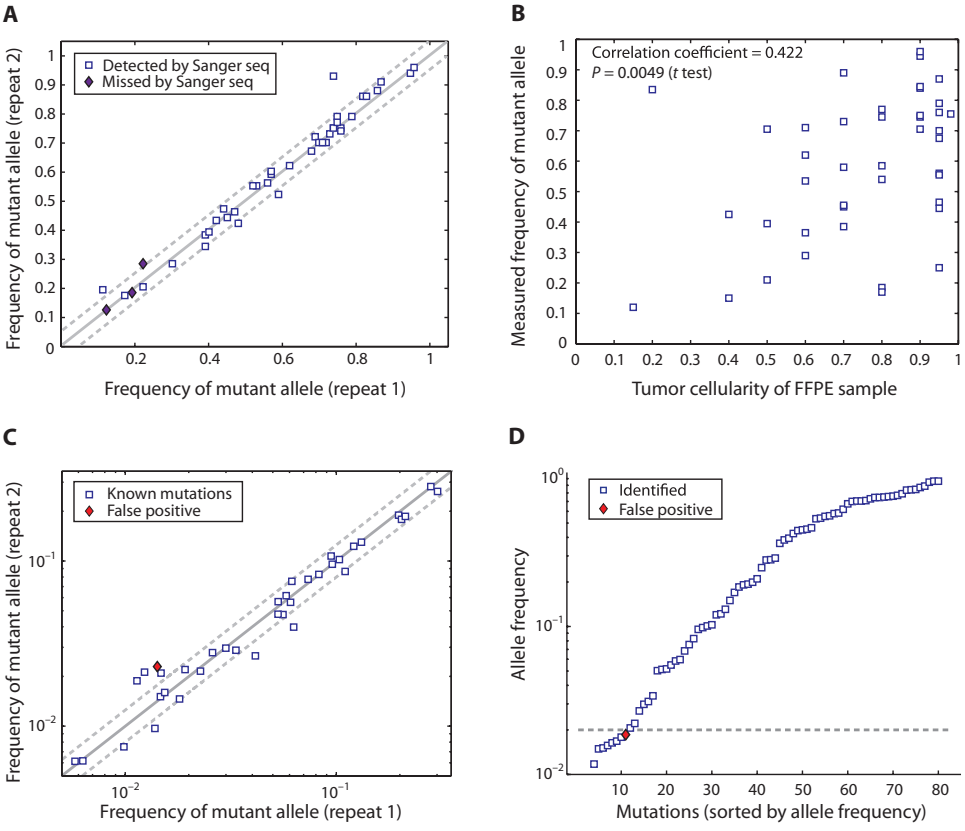


Fig. 2. Identification of mutations in ovarian cancer FFPE samples by TAm-Seq. (**A**) Concordance between duplicate measurements of AFs of mutations identified in fragmented DNA extracted from FFPE samples. The mutation frequency in each library was calculated as the fraction of reads with the mutant (nonreference) base. Solid line indicates equality. Dotted lines indicate a difference in AF of 0.05. (**B**) Correlation of AF with FFPE tumor cellularity. The measured mutant AF (average of both repeats) correlated significantly with the cellularity, estimated from histology (table S3). (**C**) Concordance between duplicate measurements of AFs of mutations identified in a mixture of DNA extracted from different FFPE samples. (**D**) Summary of mutations called in FFPE using TAm-Seq, sorted by increasing AF. Dotted line indicates AF of 2%.

limit of accuracy set by the Poisson distribution for independently segregating molecules. We compared the observed AF to the expected AF for cases where more than six copies of the minor allele were expected. Of 24 such cases, the root mean square (RMS) relative error between the expected and the observed frequency was 14%, with only 2 of 24 cases exhibiting more than 20% discrepancy. For samples with expected minor AF of 0.025, the RMS error was 23% (Fig. 3B).

Noninvasive identification of cancer mutations in plasma circulating DNA

We applied TAm-Seq to directly identify mutations in plasma of cancer patients. We studied a cohort of samples from individuals with HGSOC. These samples were first analyzed for tumor-specific mutations using digital PCR (Supplementary Methods), a method that is highly accurate (*2*, *3*, *7*, *37*) but requires design and validation of a different assay for every mutation screened and relies on previous identification of mutations in tumor samples from the same patients (*2*, *3*). We initially selected for analysis seven cases that had relatively high levels of circulating mutant *TP53* DNA in the plasma (as assessed by digital PCR). Using the equivalent amount of DNA present in 30 to 120 μl of plasma, we performed duplicate preamplification reactions for each sample. For all seven patients, *TP53* tumor mutations were identified in the circulating DNA at frequencies of 4% to 44% (Table 1). In one plasma sample collected from an ovarian cancer patient at relapse, we also identified a de novo mutation in the tyrosine kinase domain of *EGFR* (exon 21), at AF of 6% (patient 27, Table 1). We subsequently validated the presence of this mutation in plasma by performing replicate Sanger sequencing reactions of highly diluted template (Supplementary Methods), and 4 of 91 wells that were successfully Sanger-sequenced contained the *EGFR* mutation (fig. S4). We further validated the presence of this mutation by designing a sequence-specific TaqMan probe targeting this mutation and performing digital PCR (Table 1). The mutation was also identified by TAm-Seq in additional plasma collected from the same individual (sample 16, Table 2). This mutation in *EGFR* was not found in the ovarian mass removed by interval debulking surgery 15 months before the blood sample was collected, although the same sample did contain the concomitant *TP53* mutation found in the same patient's plasma, at AF of 85% (patient 27, table S3). We subsequently used TAm-Seq to sequence seven additional samples collected at the time of initial surgery including deposits in right and left ovaries and omentum. The *EGFR* mutation was detected in the two omental samples above the 0.99 confidence margin (fig. S3) at AF of 0.7%, but was not detected in the six ovarian samples (below the 0.8 confidence margin). Without previous identification in plasma, this mutation would not have been directly identified on screening those samples using high-specificity mutation identification criteria owing to its low AF. In contrast, the *TP53* mutation was identifiable in all biopsy and plasma samples (Fig. 4A). The frequency of mutant alleles in the relapsed tumor could not be directly assessed because a biopsy at relapse was not available.

We validated the TAm-Seq method on a larger panel of plasma samples in which levels of tumor-specific mutations were measured in parallel using patient-specific digital PCR assays. DNA extracted from 62 additional plasma samples collected at different time points from 37 patients with advanced HGSOC was amplified in duplicate (table S6), using DNA present in ~0.15 ml of plasma per reaction (range, 0.06 to 0.2 ml). Amplicon libraries were tagged and pooled together for sequencing with libraries prepared from 24 control samples. This generated an average sequencing depth of 650 for 62 plasma samples, sufficient to detect mutations present at AFs of 1% to 2%. Of >1.5 million possible substitutions, 42 mutations were called using the parameters previously optimized for FFPE analysis (table S6).

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

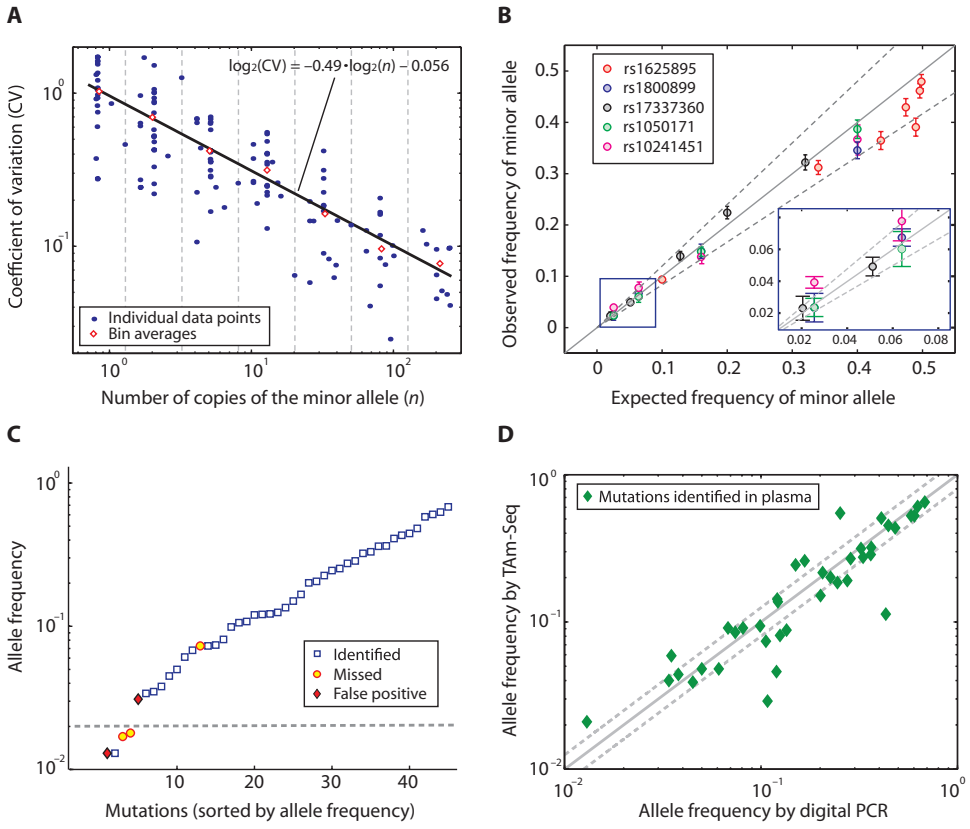


Fig. 3. Noninvasive identification and quantification of cancer mutations in plasma DNA by TAm-Seq. (**A**) Sampling noise in sequencing of sparse DNA using dilutions of plasma DNA from healthy individuals. CV of triplicate AF readings was calculated for each of the five SNPs in each of the mixes, which had varying numbers of copies of the minor allele (*n*) (blue dots). Bin averages (red diamonds) are the mean CVs calculated for each bin (bin edges denoted by the dotted vertical lines). A linear fit to the $\log_2$ of the mean CV as a function of the $\log_2$ expected copy number was calculated (black line). Two data points, with (n = 100, CV = 0.0064) and (n = 32, CV = 0.0185), were omitted from the figure for enhanced scaling. Three data points with minor allele copies of <0.8 were omitted from the analysis (n = 0.51, CV = 0.62; n = 0.41, CV = 0.86; n = 0.20, CV = 0.99). (**B**) Expected versus observed frequency of rare alleles in a dilution series of circulating DNA. Mean observed frequency was calculated for each of five SNPs for samples, where expected initial number of minor allele copies was greater than 6. Expected frequencies were calculated on the basis of quantification by digital PCR. Dotted lines represent 20% deviation from the expected frequencies. Inset highlights cases with expected minor AF <0.025. (**C**) Mutations identified in 62 plasma samples from patients with advanced HGSOC using TAm-Seq. AFs are based on digital PCR measurement for confirmed mutations (identified or missed by TAm-Seq), and on TAm-Seq for the false positives called using parameters optimized for analysis of FFPE samples. The dashed horizontal line indicates AF of 2%. Mutations detected by digital PCR at AF <1% are not shown. (**D**) AFs measured by TAm-Seq versus digital PCR for mutations identified in plasma DNA.

Table 1. Mutations identified by TAm-Seq in plasma samples from seven ovarian cancer patients. TAm-Seq was used to sequence DNA extracted from plasma of subjects with HGSOC (stage III/IV at diagnosis). Plasma was collected when patients presented with relapse disease, before initiation of chemotherapy. For patient 46, DNA from a formalin-fixed, paraffin-embedded (FFPE) sample was not included in the TAm-Seq set and the mutation was validated in FFPE by Sanger sequencing. CA125 was measured at time of plasma collection. Mean depth of coverage at the mutation locus in the TAm-Seq data was averaged over the repeats (RMS deviation = 850). AF, allele frequency; N, no; Y, yes.

Patient ID	Age at diagnosis	Time elapsed since surgery (months); number of previous lines of chemotherapy	CA125 (U/ml)	Plasma per amplification reaction (μl)	Gene	Mutation and base change (genome build hg19)		Protein change	Detected in FFPE	Mean depth (sequencing reads)	Mean AF using TAm-Seq	Mean AF using digital PCR
8	60	13; 1	2122	50	*TP53*	17:7577120	C>T	p.R273H	Y	5000	0.09	0.10
12	62	27; 3	365	50	*TP53*	17:7577579	G>T	p.Y234*	Y	5000	0.10	0.08
14	58	50; 3	260	120	*TP53*	17:7578212	G>A	p.R213*	Y	5800	0.15	0.12
25	61	9; 1	944	110	*TP53*	17:7578404	A>T	p.C176S	Y	4800	0.04	0.08
27†	68	15; 1	1051	90	*TP53*	17:7578262	C>G	p.R196P	Y	7700	0.06	0.14
					EGFR	7:55259437	G>A	p.R832H	N	5700	0.06	0.05
31	64	12; 1	313	30	*TP53*	17:7578406	C>T	p.R175H	Y	4500	0.44	0.56
46	56	30; 2	1509	30	*TP53*	17:7578406	C>T	p.R175H	Y	4200	0.23	0.30

*Indicates stop codon. †Both a *TP53* and an *EGFR* mutation were identified in this sample (Fig. 4A).

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

Table 2. Mutations identified by TAm-Seq in a set of 62 plasma samples from ovarian cancer patients. Forty mutations were identified by TAm-Seq using stringent parameters for mutation calling. Plasma samples described in this table are distinct from those in Table 1, but patients included overlap. Additional data on patients and mutations are provided in table S6.

Sample number	Plasma volume per amplification reaction (μl)	DNA amount per amplification reaction (ng)	Gene	Protein change	Mean depth (sequencing reads)	Mean AF using TAm-Seq	Mean AF using digital PCR
1	70	0.9	*TP53*	p.R273C	640	0.260	0.167
2	160	4.2	*TP53*	p.R248Q	340	0.244	0.150
3	160	5.7	*TP53*	p.R248Q	640	0.507	0.410
4	120	9.9	*TP53*	p.R213X	810	0.059	0.035
5	120	1.4	*TP53*	p.C141Y	680	0.021	0.013
6	120	2.1	*TP53*	p.C141Y	720	0.044	0.038
7	190	17.9	*TP53*	p.I195N	800	0.091	0.081
8	160	14.8	*TP53*	p.R175H	510	0.608	0.627
9	160	10.7	*TP53*	p.R175H	550	0.526	0.604
10	160	6.1	*TP53*	p.R175H	530	0.651	0.682
11	160	4.9	*TP53*	p.R175H	490	0.526	0.581
13	160	2.8	*TP53*	p.C135R	480	0.039	0.045
14	160	2.5	*TP53*	p.C135R	610	0.046	0.120
15	160	3.0	*TP53*	p.C135R	470	0.091	0.068
16†	130	3.7	*TP53*	p.R196P	1070	0.088	0.135
			EGFR	p.R832H	614	0.048	0.050
17	160	4.2	*TP53*	p.C176S	580	0.113	0.432
18	160	4.4	*TP53*	p.C176S	620	0.029	0.108
20	140	5.2	*TP53*	p.R175H	650	0.201	0.226
21	140	3.6	*TP53*	p.R175H	650	0.085	0.074
22	140	4.1	*TP53*	p.R175H	630	0.081	0.125
23	140	3.7	*TP53*	p.R175H	710	0.074	0.106
24	140	7.1	*TP53*	p.R175H	760	0.269	0.286
25	130	3.9	*TP53*	p.R273H	750	0.094	0.099
26	160	5.7	*TP53*	p.R282W	640	0.048	0.061
27	150	3.6	*TP53*	p.C141Y	480	0.321	0.364
29	150	9.5	*TP53*	p.E258K	190	0.548	0.253
31	160	3.6	*TP53*	p.C135Y	620	0.040	0.034
32	140	2.4	*TP53*	p.E56X	1480	0.137	0.122
33	160	13.2	*TP53*	p.K132N	740	0.216	0.206
34	60	5.3	*TP53*	p.K132N	570	0.151	0.201
36	160	5.8	*TP53*	p.K132N	620	0.191	0.275
37	160	9.4	*TP53*	p.K132N	530	0.287	0.362
38	160	10.1	*TP53*	p.K132N	590	0.275	0.331
39	160	16.4	*TP53*	p.K132N	700	0.315	0.323
40	160	19.7	*TP53*	p.K132N	830	0.435	0.482
41	160	15.0	*TP53*	p.K132N	730	0.452	0.445
42	160	8.5	*TP53*	p.K132N	560	0.185	0.245
43	150	3.6	*TP53*	Splicing	680	0.143	0.121
44‡	170	5.2	*TP53*	p.C238R	1543	0.071	0.073

†Both a *TP53* and an *EGFR* mutation were identified in this sample, collected from patient 27 (Table 1), 25 months after initial surgery (Fig. 4A). ‡The amplicon containing the mutation failed amplification in this sample in the initial experiment and was identified successfully in repeat analysis.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

Thirty-nine of these matched mutations detected by digital PCR in those samples (Fig. 3C). Three potential false positives were called, at AF of 3.1%, 1.3%, and 0.7% (the latter in a control sample). Using higher-stringency parameters for mutation identification (Supplementary Methods), we retained only the 39 validated mutations called, with no false positives (Table 2).

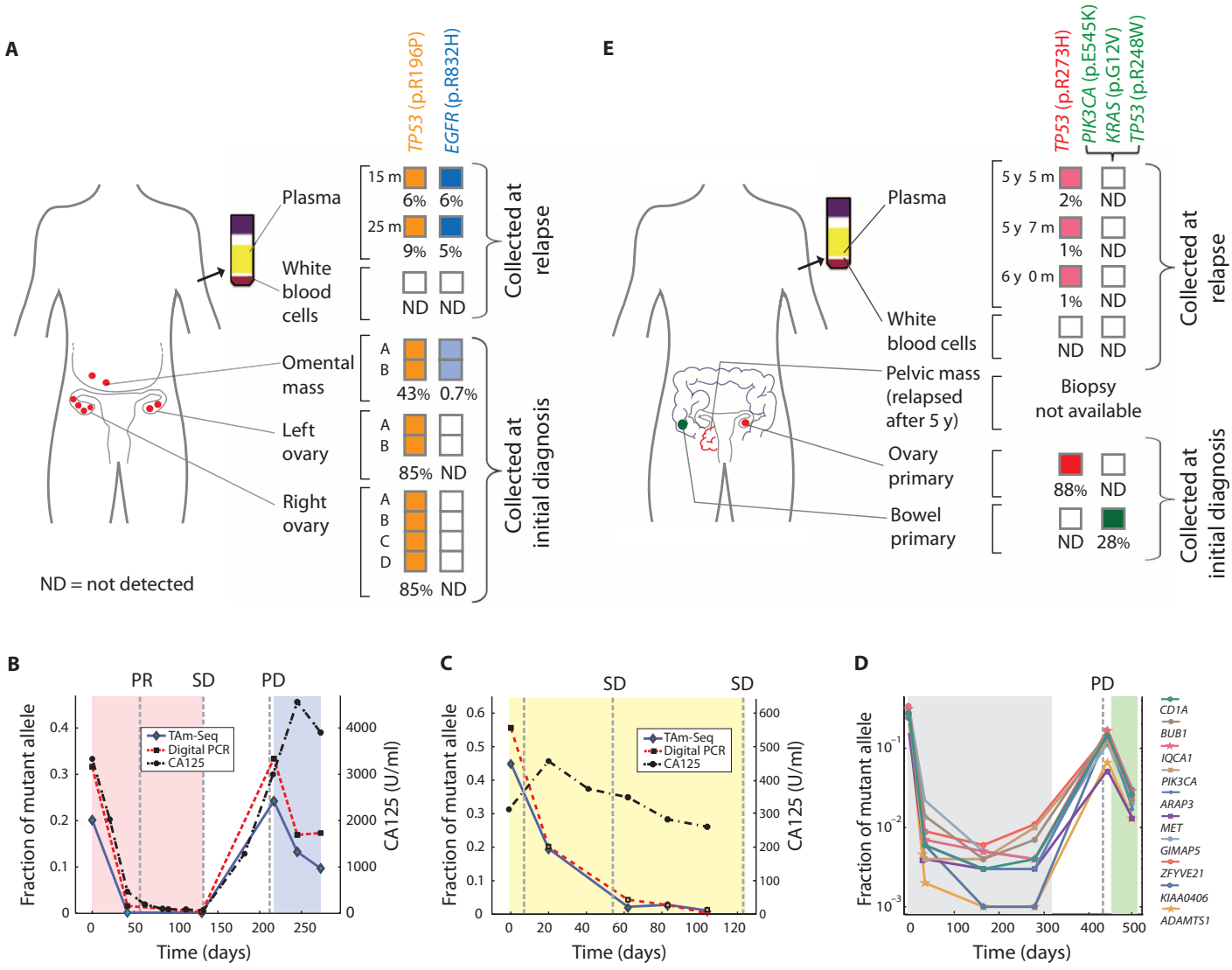


Fig. 4. Clinically relevant applications of plasma DNA sequencing using TAm-Seq. (**A**) Retrospective analysis by TAm-Seq of plasma samples collected during patient follow-up and biopsy specimens collected at initial surgery. We identified a mutation in exon 21 of *EGFR* (dark blue boxes) in two separate plasma samples, collected 15 and 25 months after initial surgery from patient 27 (Tables 1 and 2). This mutation was not directly identified in eight tumor biopsy specimens collected at the time of initial surgery (two from omental mass, two from left ovary, and four from right ovary). Having identified the mutation in the plasma samples, we examined this mutation using the lower-specificity criteria defined for mutation detection (Supplementary Methods) and detected the mutation in the two specimens that had been collected from the omentum at the time of surgery (light blue boxes) but not in the six ovarian specimens. A mutation in *TP53* was identified in all tumor and plasma samples collected from this patient (Tables 1 and 2 and table S3), but not in white blood cells (buffy coat). Percentages indicate mutant AFs. Empty boxes and "ND" indicate samples where a mutation was not identified or detected (below 0.8 confidence margin). (**B**) Monitoring frequency of mutant DNA in plasma of an ovarian cancer patient (patient 46) over time using TAm-Seq and digital PCR. TAm-Seq results are reported as the mean frequency of duplicate analyses. Parallel data are shown for digital PCR and serum CA125. Shaded regions indicate periods of chemotherapy, and vertical dashed lines indicate radiological assessment of patient responses: PR, partial response; SD, stable disease; PD, progressive disease. (**C**) Monitoring frequency of mutant DNA in plasma of an ovarian cancer patient (patient 31) over time. (**D**) Dynamics of 10 tumor-specific mutations in plasma of a breast cancer patient (not included in the other sets of samples analyzed). (**E**) Retrospective analysis of samples from synchronous primary tumors (bowel and ovarian) collected at the time of initial surgery and three plasma samples collected at relapse. In primary tumors from this patient (not included in the other sets of samples analyzed), a *TP53* mutation was identified in the ovarian cancer (red box), and mutations in *PIK3CA*, *KRAS*, and *TP53* were identified in the bowel cancer (green box). At relapse, a biopsy was not performed on the pelvic mass. The *TP53* mutation that was identified in the ovarian primary tumor (p.R273H) was detected in plasma, whereas the bowel-associated mutations were not detected.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

Of 40 point mutations detected at AF >2% by digital PCR, 38 (95%) were identified by TAm-Seq in a single experiment (Fig. 3C). One additional mutation was located in an amplicon that failed in that sample and was identified in repeated analysis; the other was likely missed by TAm-Seq owing to sampling noise, because it was found in one of the duplicate preamplified libraries but not the other (table S6). One of three mutations detected by digital PCR at 1% < AF < 2% was identified by TAm-Seq (Fig. 3C). Eleven additional point mutations detected by digital PCR at AF <1% were not detected by TAm-Seq at these settings. TAm-Seq and digital PCR measurements of AF had excellent agreement, with correlation coefficient of 0.90, increasing to 0.97 when discarding the two strongest outliers (Fig. 3D). Thus, we screened 62 samples across sizeable genomic stretches, using minute amounts of plasma DNA (median, 4 ng), and obtained 97.5% sensitivity with PPV of 100% for identifying mutations at AF >2% in plasma by TAm-Seq. Using parameters optimized for FFPE samples, one potential false positive was called at AF >2%, reducing the PPV to 97.5% (Table 3).

Monitoring levels of ctDNA

Various methods have been suggested to monitor changes in mutation load in plasma. These can have enhanced sensitivity compared to TAm-Seq for tracking individual mutations, but require design of personalized assays (*3*, *18*, *19*). None of these methods have been widely adopted. We therefore applied TAm-Seq as a generic tool to measure changes in the frequency of ctDNA over time. We studied serial plasma samples collected during follow-up and treatment of two patients with relapsed HGSOC, collected during 104 and 273 days of follow-up and treatment, respectively. Frequencies of mutant *TP53* alleles were measured by TAm-Seq and in parallel by digital PCR using a mutation-specific probe. The two methods of quantification had excellent agreement. Mutant AFs in plasma of ovarian cancer patients reflected well the clinical course of the disease compared to the serum marker CA125, showed marked decrease when systemic treatment was initiated, and increased in parallel to disease progression. In the first case (Fig. 4B), a 56-year-old woman with relapsed ovarian cancer (patient 46) was treated with fourth-line carboplatin + paclitaxel chemotherapy for six cycles (pink-shaded region). Radiology showed partial response on mid-treatment computed tomography (CT) scan. End-of-treatment CT showed stable disease. Twelve weeks from the end of her fourth-line treatment, the patient developed progressive disease. The patient then initiated fifth-line chemotherapy with liposomal doxorubicin (purple-shaded region). In the second case (Fig. 4C), a 64-year-old woman with relapsed ovarian cancer (patient 31) was treated with second-line ECX (epirubicin, cisplatin, and capecitabine) chemotherapy for six cycles. Radiology showed stable disease on mid- and end-of-treatment CT scans. The patient then remained off treatment, until she progressed 3 months later.

TAm-Seq can be flexibly adapted to sequence different genomic regions by designing primers to amplify regions of interest. We used this capability to study dynamics of multiple mutations in parallel.

Table 3. Summary of mutations identified in 69 plasma samples of ovarian cancer patients. Samples were analyzed by TAm-Seq and in parallel by digital PCR. Using parameters optimized for plasma DNA, false-positive calls were lost, whereas all confirmed calls were retained, resulting in specificity and PPV of 100%.

First set of plasma samples	
Plasma samples analyzed	7
Point mutations originally detected by digital PCR, using patient-specific assays targeting mutations identified in tumor samples	7
Point mutations identified directly in plasma by TAm-Seq	8
De novo mutations identified by TAm-Seq only, subsequently confirmed by digital PCR	1
Second set of plasma samples	
Plasma samples analyzed	62
Point mutations detected by digital PCR at AF >2%	40
Point mutations with AF >2% (by digital PCR) identified by TAm-Seq	39
Point mutations missed by TAm-Seq due to sampling error	1
Sensitivity of TAm-Seq for identifying mutations at AF >2%	97.5%
PPV of mutations called by TAm-Seq with AF >2%	97.5%*
ctDNA in ovarian cancer	
Advanced ovarian cancer patients in both sets†	38
Patients where TAm-Seq identified cancer mutations	20

*One unconfirmed substitution was called at AF >2% using parameters optimized for FFPE material. †The first set included 7 patients (Table 1), and the second set included 37 patients (table S6), 6 of whom overlap.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

Whole-genome sequencing of tumor material was used to identify tumor mutations in a patient with metastatic breast cancer undergoing two phases of chemotherapy. Ten mutations were selected, and short amplicons (<120 bp) were designed to cover the mutation loci (table S7). Serial plasma samples were collected over the course of 497 days, both before and after treatment. We performed TAm-Seq in duplicate, using DNA from 0.08 ml of plasma per amplification, and tracked dynamics of all mutations in parallel (Fig. 4D). The patient was treated with single-agent epirubicin (gray-shaded region). After 4 months off treatment, a CT scan showed progressive disease and the patient commenced further treatment with paclitaxel chemotherapy. The 10 mutations followed a common pattern of sharp decline in AF upon onset of therapy and an increase in AF upon disease progression after termination of therapy (Fig. 4D).

Finally, we used TAm-Seq to study plasma from a patient who had a history of two synchronous primary cancers, bowel and ovarian, which were resected simultaneously. After a 5-year remission, a pelvic mass of uncertain origin was detected. A biopsy was considered to guide selection of therapy but was not performed owing to risk of complications and comorbidities. The patient commenced empirically on an ovarian cancer chemotherapy regimen, to which she responded. Retrospective analysis by TAm-Seq of FFPE from the primary tumors collected at initial surgery, and three plasma samples collected serially at the time of relapse (5 years and 5 months, 5 years and 7 months, and 6 years after initial surgery), showed that the patient's plasma at relapse contained the *TP53* (p.R273H) mutation identified in the ovarian primary tumor (exceeding the 0.98, 0.93, and 0.97 confidence margins, respectively), but not the *PIK3CA* (p.E545K), *KRAS* (p.G12V), or *TP53* (p.R248W) mutations identified in the primary bowel cancer (below the 0.8 confidence margin) (Fig. 4E). Had these results been available, uncertainty and treatment delays may have been avoided, as well as the risk of prescribing chemotherapy for an inappropriate tumor site. An alternative possible outcome may have involved a finding of the *PIK3CA* or *KRAS* mutations (present in the primary bowel cancer) in the patient's plasma at the time of relapse. Such a finding, if available to clinicians at the time, may not only have led to alternate chemotherapy being offered but may have also opened the possibility of enrolment into a trial for targeted therapy with mammalian target of rapamycin (mTOR), phosphatidylinositol 3-kinase (PI3K), or mitogen-activated protein kinase kinase (MEK) inhibitors (*11*).

DISCUSSION

Detection of rare mutations in circulating DNA has long been pursued owing to its potentially transformative impact on cancer diagnosis and management. Important progress has been made using sequence-specific assays that target predefined mutations and that detect extremely rare alleles. Assays such as PCR (*6*, *7*), ligation (*5*), and primer extension/mass spectrometry (*27*) can identify specific, predefined mutations in plasma samples. Enhanced detection down to 1 mutant allele in 10,000 or more wild-type alleles can be obtained using a variety of methods, such as peptide nucleic acid and primer extension ("PPEM") (*38*), ligation followed by quantitative PCR ("LigAmp") (*39*), bead-based digital PCR in emulsions ("BEAMing") (*2*, *3*), microfluidic-based (*7*) or droplet-based digital PCR (*40*), or microinsertion/deletion/indel-activated pyrophosphorolysis ("MAP") (*29*). Nonetheless, identification of rare mutations in tumor suppressor genes such as *TP53*, which are widely mutated in cancers but lack a well-defined hotspot region, remains an elusive goal.

In patients with advanced cancers, mutant alleles can reach a sizeable fraction of DNA. For example, Dukes' D colorectal cancers have median 8% mutant AF (*2*). Screening of entire genes for mutations would therefore be useful for some applications, even if analytical selectivity is limited to a few percent. Advances in massively parallel sequencing make new approaches possible. These have largely focused on large-scale analyses, including whole-genome or whole-exome sequencing (*41*). This generates a large amount of data on genomic regions that do not, at present, inform clinical decisions. Moreover, the depth of coverage for clinically significant loci is not sufficient to detect changes that occur at low frequency (<5%). Such approaches have recently been complemented by methods for examination of individual amplicons at great depth (*30*).

The intermediate scale of sequencing is most likely to have immediate impact on clinical genomics. Targeted sequencing has been applied for tumor DNA (*34*, *35*) and cyst fluid (*33*) to detect mutations down to 5% AF, but has not been applied for analysis of circulating tumor nucleic acids. Here, we demonstrate noninvasive identification of mutant alleles in plasma, at AFs as low as 2%, by targeted deep sequencing of circulating DNA. Our TAm-Seq method uses a combination of short amplicons, two-step amplification, sample barcodes, and high-throughput PCR. Because the amplicons are short, this method effectively amplifies even small amounts of fragmented DNA such as are present in circulating DNA. The two-step amplification permits extensive primer multiplexing that enables the amplification and sequencing of sizeable genomic regions by tiling short amplicons without loss of fidelity or efficiency. Duplicate sequencing of each sample is used to avoid false positives stemming from PCR errors. Sample barcodes and high-throughput PCR reduce the per-sample costs to a range where this may be widely applicable. Preparing TAm-Seq libraries for sequencing from 48 samples takes less than 24 hours and involves only few hours of hands-on time. New platforms for massively parallel sequencing allow for fast turnaround times, which make this approach practical in a clinical setting.

The sensitivity presently achieved can provide useful diagnostic information in certain advanced cancers. We studied a cohort of subjects with advanced HGSOC in which the tumor suppressor gene *TP53* is a driver mutation (*20*). Of the 69 plasma samples collected from 38 different individuals with advanced HGSOC, we identified mutations in *TP53* in 46 samples (67%) from 20 of the cases (53%). In contrast, a previous study using a ligase detection reaction with bespoke primers found mutated *TP53* sequences in plasma for only 30% of advanced ovarian cancer patients (*5*), and a study using single-strand conformation polymorphism found no ctDNA in preoperative plasma samples from high-grade serous cancer patients (*42*).

Targeted agents, such as inhibitors of poly(adenosine diphosphate–ribose) polymerase (PARP), or tyrosine kinase inhibitors targeting epidermal growth factor receptor (EGFR), may be applicable for systemic treatment of advanced HGSOC (*8*, *10*, *22*). In a recent study of 203 HGSOC tumors, *EGFR* was found to be the most frequently mutated oncogene and was mutated in nearly 10% of cases (*10*). In one case, we identified in plasma a de novo mutation in the tyrosine kinase domain (exon 21) of *EGFR*, located 26 amino acids upstream of the L858R activating mutation widely documented for lung cancer. In a subset of tumor samples collected from the same patient 15 months earlier, this mutation was detected at AF of 0.7%, but could not have

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

been identified by analysis of those samples alone without previous knowledge of the mutation identified in plasma (Fig. 4A). In a clinical setting, identification of such a mutation could potentially guide treatment with alternative molecularly targeted therapy (*10*). Current clinical recommendations in lung adenocarcinoma suggest mutation assessment in exons 18 to 21 of *EGFR* (a region of ~560 bp) in the tumor tissue to identify patients eligible for treatment with gefitinib or erlotinib (*9*). Using a commercial PCR-based in vitro diagnostic kit (Qiagen), 28 different *EGFR* variants can be assayed (not including the mutation we identified), but the sample needs to be subdivided into seven different reactions. When sample is limited or mutant alleles are rare, this could introduce sampling errors.

Using standard amplification primers tailored to the mutation loci, we also used TAm-Seq to monitor the dynamics of 10 mutations in plasma DNA of a single patient with metastatic breast cancer, using minute amounts of input DNA. Previous studies have followed up to two mutations in any individual patient (*3*, *19*). Tracking multiple mutations can provide insight into clonal evolution and, at the same time, increases the robustness for tumor monitoring by compensating for effects of sampling noise or mutational drift. For example, if a patient has only five copies of a mutant allele per milliliter of plasma (on average), there is a 37% probability that this mutation will not be present in a 0.2-ml sample, and even a perfect assay will fail to detect residual tumor, whereas a method that measures multiple mutations in parallel can have a low likelihood of a false-negative result even if the detection rate for each mutation is less than 50%.

A current limitation of TAm-Seq is the detection limit compared to assays that target individual loci (*2*, *3*, *7*, *40*), which have been shown to detect two to three orders of magnitude lower frequencies. Our approach may be sufficient for analyzing plasma from patients with certain advanced cancers, but further improvement may be necessary before this method can be more widely used in the clinic. Higher read depth or fidelity, additional replicates, or improved algorithms could allow for enhanced mutation detection without change to protocols. An alternative strategy is through rare allele enrichment, for example, by combining TAm-Seq with protocols such as COLD-PCR (co-amplification at lower denaturation temperature PCR) (*31*).

Previously proposed methods for personalized monitoring of tumor dynamics relied on expensive custom-designed probes (*3*) or identification of rearrangements using whole-genome sequencing (*18*, *19*). These have better analytical sensitivity than currently achieved by TAm-Seq, but are difficult to implement on a routine basis. TAm-Seq strikes a balance between sensitivity and ease of use and could facilitate study and application of circulating DNA. Using TAm-Seq, we identified cancer mutations in the plasma of most advanced ovarian cancer patients and tracked dynamics of *TP53* mutations without requiring any specially designed probes. In summary, TAm-Seq is a flexible and cost-effective platform for applications in noninvasive cancer genomics and diagnostics. We have shown that this method can be used for high-throughput sequencing of plasma samples to identify and monitor levels of multiple cancer mutations in circulating DNA. This could also be applied to screen for rare mutations in a variety of heterogeneous sample types such as low-cellularity tumor specimens, cytological samples, or circulating tumor cells (*16*). With further developments, this and derivative methods may be applied in molecular screening for earlier detection or for differential diagnosis of cancer from benign masses. For genetic analysis of FFPE or small biopsy samples, TAm-Seq can be applied as is, as a cost-effective clinical aid.

MATERIALS AND METHODS

Sample collection

FFPE blocks were obtained from the pathology archives at Addenbrooke's Hospital (Cambridge, UK). Plasma samples were collected upon disease relapse, before and during chemotherapy treatment. Sample collection for this study was approved by Cambridgeshire Research Ethics Committee (REC 08/H0306/61 and 07/Q0106/63). Peripheral blood samples were collected into EDTA tubes and centrifuged at 820*g* for 10 min within 1 hour of collection to limit degradation of cell-free DNA and leukocyte lysis. Aliquots (1 ml) of plasma were centrifuged in a benchtop microfuge at 14,000 rpm for 10 min. The supernatant was transferred to sterile 1.5-ml tubes and stored at −80°C before extraction.

Extraction of DNA from FFPE and blood plasma

Paraffin blocks were cut as 8-μm sections on plain glass slides. Targeted regions for sampling were marked on adjacent hematoxylin and eosin sections by the study pathologist and recovered by scrape macrodissection. Between 3 and 20 sections were macrodissected depending on the tissue sample's size. DNA from FFPE sections was extracted with QIAamp DNA FFPE Tissue Kit (Qiagen) according to the manufacturer's instructions.

Circulating DNA was extracted from between 0.85 and 2.2 ml of plasma with the QIAamp Circulating Nucleic Acid kit (Qiagen), following the manufacturer's instructions, and with the QIAvac 24 Plus vacuum manifold. Carrier RNA was added to ACL lysis buffer to enhance binding of nucleic acids to the QIAamp membrane with the aim to enhance yields.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/4/136/136ra68/DC1
Methods
Fig. S1. PCR strategy and primer design.
Fig. S2. Sanger traces for mutations identified by tagged-amplicon sequencing.
Fig. S3. Background frequencies and detection limits for base substitutions.
Fig. S4. Replicate dilute Sanger sequencing of a mutation identified in plasma.
Table S1. Target-specific primers.
Table S2. Unique sequencing barcodes.
Table S3. Mutations identified in FFPE samples.
Table S4. SNPs identified in circulating DNA from two plasma control samples.
Table S5. Frequency of SNP alleles in dilution series of DNA from control plasma.
Table S6. Additional data for Table 2 for mutations identified in plasma samples.
Table S7. Mutations and amplicons studied in one breast cancer patient.

REFERENCES AND NOTES

1. E. Gormally, E. Caboux, P. Vineis, P. Hainaut, Circulating free DNA in plasma or serum as biomarker of carcinogenesis: Practical aspects and biological significance. *Mutat. Res.* **635**, 105–117 (2007).
2. F. Diehl, M. Li, D. Dressman, Y. He, D. Shen, S. Szabo, L. A. Diaz Jr., S. N. Goodman, K. A. David, H. Juhl, K. W. Kinzler, B. Vogelstein, Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16368–16373 (2005).
3. F. Diehl, K. Schmidt, M. A. Choti, K. Romans, S. Goodman, M. Li, K. Thornton, N. Agrawal, L. Sokoll, S. A. Szabo, K. W. Kinzler, B. Vogelstein, L. A. Diaz Jr., Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).
4. H. Schwarzenbach, D. S. Hoon, K. Pantel, Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).
5. E. M. Swisher, M. Wollan, S. M. Mahtani, J. B. Willner, R. Garcia, B. A. Goff, M. C. King, Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *Am. J. Obstet. Gynecol.* **193**, 662–667 (2005).

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

6. R. E. Board, A. M. Wardley, J. M. Dixon, A. C. Armstrong, S. Howell, L. Renshaw, E. Donald, A. Greystoke, M. Ranson, A. Hughes, C. Dive, Detection of *PIK3CA* mutations in circulating free DNA in patients with breast cancer. *Breast Cancer Res. Treat.* **120**, 461–467 (2010).
7. T. K. Yung, K. C. Chan, T. S. Mok, J. Tong, K. F. To, Y. M. Lo, Single-molecule detection of epidermal growth factor receptor mutations in plasma by microfluidics digital PCR in non–small cell lung cancer patients. *Clin. Cancer Res.* **15**, 2076–2084 (2009).
8. S. Banerjee, S. Kaye, The role of targeted therapy in ovarian cancer. *Eur. J. Cancer* **47** (Suppl. 3), S116–S130 (2011).
9. V. L. Keedy, S. Temin, M. R. Somerfield, M. B. Beasley, D. H. Johnson, L. M. McShane, D. T. Milton, J. R. Strawn, H. A. Wakelee, G. Giaccone, American Society of Clinical Oncology provisional clinical opinion: Epidermal growth factor receptor (*EGFR*) mutation testing for patients with advanced non–small-cell lung cancer considering first-line EGFR tyrosine kinase inhibitor therapy. *J. Clin. Oncol.* **29**, 2121–2127 (2011).
10. U. A. Matulonis, M. Hirsch, E. Palescandolo, E. Kim, J. Liu, P. van Hummelen, L. MacConaill, R. Drapkin, W. C. Hahn, High throughput interrogation of somatic mutations in high grade serous cancer of the ovary. *PLoS One* **6**, e24433 (2011).
11. J. A. Engelman, L. Chen, X. Tan, K. Crosby, A. R. Guimaraes, R. Upadhyay, M. Maira, K. McNamara, S. A. Perera, Y. Song, L. R. Chirieac, R. Kaur, A. Lightbown, J. Simendinger, T. Li, R. F. Padera, C. García-Echeverría, R. Weissleder, U. Mahmood, L. C. Cantley, K. K. Wong, Effective use of PI3K and MEK inhibitors to treat mutant *Kras* G12D and *PIK3CA* H1047R murine lung cancers. *Nat. Med.* **14**, 1351–1356 (2008).
12. P. J. Campbell, S. Yachida, L. J. Mudie, P. J. Stephens, E. D. Pleasance, L. A. Stebbings, L. A. Morsberger, C. Latimer, S. McLaren, M. L. Lin, D. J. McBride, I. Varela, S. A. Nik-Zainal, C. Leroy, M. Jia, A. Menzies, A. P. Butler, J. W. Teague, C. A. Griffin, J. Burton, H. Swerdlow, M. A. Quail, M. R. Stratton, C. Iacobuzio-Donahue, P. A. Futreal, The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
13. S. P. Shah, R. D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R. A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G. A. Taylor, A. E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R. L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M. A. Marra, S. Aparicio, Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
14. M. Inukai, S. Toyooka, S. Ito, H. Asano, S. Ichihara, J. Soh, H. Suehisa, M. Ouchida, K. Aoe, M. Aoe, K. Kiura, N. Shimizu, H. Date, Presence of *epidermal growth factor receptor gene* T790M mutation as a minor clone in non–small cell lung cancer. *Cancer Res.* **66**, 7854–7858 (2006).
15. S. L. Edwards, R. Brough, C. J. Lord, R. Natrajan, R. Vatcheva, D. A. Levine, J. Boyd, J. S. Reis-Filho, A. Ashworth, Resistance to therapy caused by intragenic deletion in *BRCA2*. *Nature* **451**, 1111–1115 (2008).
16. S. Maheswaran, L. V. Sequist, S. Nagrath, L. Ulkus, B. Brannigan, C. V. Collura, E. Inserra, S. Diederichs, A. J. Iafrate, D. W. Bell, S. Digumarthy, A. Muzikansky, D. Irimia, J. Settleman, R. G. Tompkins, T. J. Lynch, M. Toner, D. A. Haber, Detection of mutations in *EGFR* in circulating lung-cancer cells. *N. Engl. J. Med.* **359**, 366–377 (2008).
17. B. Norquist, K. A. Wurz, C. C. Pennil, R. Garcia, J. Gross, W. Sakai, B. Y. Karlan, T. Taniguchi, E. M. Swisher, Secondary somatic mutations restoring *BRCA1/2* predict chemotherapy resistance in hereditary ovarian carcinomas. *J. Clin. Oncol.* **29**, 3008–3015 (2011).
18. R. J. Leary, I. Kinde, F. Diehl, K. Schmidt, C. Clouser, C. Duncan, A. Antipova, C. Lee, K. McKernan, F. M. De La Vega, K. W. Kinzler, B. Vogelstein, L. A. Diaz Jr., V. E. Velculescu, Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
19. D. J. McBride, A. K. Orpana, C. Sotiriou, H. Joensuu, P. J. Stephens, L. J. Mudie, E. Hämäläinen, L. A. Stebbings, L. C. Andersson, A. M. Flanagan, V. Durbecq, M. Ignatiadis, O. Kallioniemi, C. A. Heckman, K. Alitalo, H. Edgren, P. A. Futreal, M. R. Stratton, P. J. Campbell, Use of cancer-specific genomic rearrangements to quantify disease burden in plasma from patients with solid tumors. *Genes Chromosomes Cancer* **49**, 1062–1069 (2010).
20. A. A. Ahmed, D. Etemadmoghadam, J. Temple, A. G. Lynch, M. Riad, R. Sharma, C. Stewart, S. Fereday, C. Caldas, A. Defazio, D. Bowtell, J. D. Brenton, Driver mutations in *TP53* are ubiquitous in high grade serous carcinoma of the ovary. *J. Pathol.* **221**, 49–56 (2010).
21. Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
22. R. C. Bast Jr., Molecular approaches to personalizing management of ovarian cancer. *Ann. Oncol.* **22** (Suppl. 8), viii5–viii15 (2011).
23. K. C. Chan, S. F. Leung, S. W. Yeung, A. T. Chan, Y. M. Lo, Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **14**, 4141–4145 (2008).
24. H. C. Fan, Y. J. Blumenfeld, U. Chitkara, L. Hudgins, S. R. Quake, Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. *Clin. Chem.* **56**, 1279–1286 (2010).
25. Y. M. Lo, K. C. Chan, H. Sun, E. Z. Chen, P. Jiang, F. M. Lun, Y. W. Zheng, T. Y. Leung, T. K. Lau, C. R. Cantor, R. W. Chiu, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
26. R. W. Chiu, R. Akolekar, Y. W. Zheng, T. Y. Leung, H. Sun, K. C. Chan, F. M. Lun, A. T. Go, E. T. Lau, W. W. To, W. C. Leung, R. Y. Tang, S. K. Au-Yeung, H. Lam, Y. Y. Kung, X. Zhang, J. M. van Vugt, R. Minekawa, M. H. Tang, J. Wang, C. B. Oudejans, T. K. Lau, K. H. Nicolaides, Y. M. Lo, Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: Large scale validity study. *BMJ* **342**, c7401 (2011).
27. Y. M. Lo, N. B. Tsui, R. W. Chiu, T. K. Lau, T. N. Leung, M. M. Heung, A. Gerovassili, Y. Jin, K. H. Nicolaides, C. R. Cantor, C. Ding, Plasma placental RNA allelic ratio permits non-invasive prenatal chromosomal aneuploidy detection. *Nat. Med.* **13**, 218–223 (2007).
28. Y. M. Lo, F. M. Lun, K. C. Chan, N. B. Tsui, K. C. Chong, T. K. Lau, T. Y. Leung, B. C. Zee, C. R. Cantor, R. W. Chiu, Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 13116–13121 (2007).
29. Z. Chen, J. Feng, C. H. Buzin, Q. Liu, L. Weiss, K. Kernstine, G. Somlo, S. S. Sommer, Analysis of cancer mutation signatures in blood by a novel ultra-sensitive assay: Monitoring of therapy or recurrence in non-metastatic breast cancer. *PLoS One* **4**, e7220 (2009).
30. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9530–9535 (2011).
31. J. Li, L. Wang, H. Mamon, M. H. Kulke, R. Berbeco, G. M. Makrigiorgos, Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat. Med.* **14**, 579–584 (2008).
32. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, P. A. Futreal, COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
33. J. Wu, H. Matthaei, A. Maitra, M. Dal Molin, L. D. Wood, J. R. Eshleman, M. Goggins, M. I. Canto, R. D. Schulick, B. H. Edil, C. L. Wolfgang, A. P. Klein, L. A. Diaz Jr., P. J. Allen, C. M. Schmidt, K. W. Kinzler, N. Papadopoulos, R. H. Hruban, B. Vogelstein, Recurrent *GNAS* mutations define an unexpected pathway for pancreatic cyst development. *Sci. Transl. Med.* **3**, 92ra66 (2011).
34. O. Harismendy, R. B. Schwab, L. Bao, J. Olson, S. Rozenzhak, S. K. Kotsopoulos, S. Pond, B. Crain, M. S. Chee, K. Messer, D. R. Link, K. A. Frazer, Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol.* **12**, R124 (2011).
35. N. Wagle, M. F. Berger, M. J. Davis, B. Blumenstiel, M. DeFelice, P. Pochanard, M. Ducar, P. Van Hummelen, L. E. MacConaill, W. C. Hahn, M. Meyerson, S. B. Gabriel, L. A. Garraway, High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov.* **2**, 82 (2012).
36. T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
37. B. Vogelstein, K. W. Kinzler, Digital PCR. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9236–9241 (1999).
38. X. Sun, K. Hung, L. Wu, D. Sidransky, B. Guo, Detection of tumor mutations in the presence of excess amounts of normal DNA. *Nat. Biotechnol.* **20**, 186–189 (2002).
39. C. Shi, S. H. Eshleman, D. Jones, N. Fukushima, L. Hua, A. R. Parker, C. J. Yeo, R. H. Hruban, M. G. Goggins, J. R. Eshleman, LigAmp for sensitive detection of single-nucleotide differences. *Nat. Methods* **1**, 141–147 (2004).
40. L. B. Pinheiro, V. A. Coleman, C. M. Hindson, J. Herrmann, B. J. Hindson, S. Bhat, K. R. Emslie, Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. *Anal. Chem.* **84**, 1003–1011 (2012).
41. M. Meyerson, S. Gabriel, G. Getz, Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
42. J. Otsuka, T. Okuda, A. Sekizawa, S. Amemiya, H. Saito, T. Okai, M. Kushima, Detection of p53 mutations in the plasma DNA of patients with ovarian cancer. *Int. J. Gynecol. Cancer* **14**, 459–464 (2004).

Acknowledgments: We thank H. Biggs, C. Hodgkin, S. Richardson, and L. Jones for assistance in sample collection, S. Aldridge for assistance in genomic analysis, and S. Tavaré, B. Davis, and M. Dunning for assistance in data analysis. **Funding:** We acknowledge the support of Cancer Research UK, the University of Cambridge, National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge Experimental Cancer Medicine Centre, and Hutchison Whampoa Limited. C.P. was supported in part by the Academy of Medical Sciences, Wellcome Trust, British Heart Foundation, and Arthritis Research UK. **Author contributions:** T.F., M.M., C.P., D.G., D.W.Y.T., C.C., J.D.B., and N.R. designed the study. T.F., M.M., D.W.Y.T., F.K., J.H., A.P.M.,

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

and N.R. developed methods. T.F., D.G., D.W.Y.T., A.M.P., and S.-J.D. collected the data. T.F., M.M., and N.R. analyzed TAm-Seq data. C.P., S.-J.D., C.C., and J.D.B. designed clinical studies and collected samples and clinical data. M.J.-L. performed pathological analysis. D.B. contributed sequencing data. T.F., M.M., C.P., D.G., D.W.Y.T., J.H., A.P.M., J.D.B., and N.R. interpreted the data. T.F., M.M., and N.R. wrote the manuscript with assistance from C.P., D.G., D.W.Y.T., A.P.M., J.D.B., and other authors. All authors approved the final manuscript. **Competing interests:** A.P.M. and F.K. hold equity in Fluidigm and may stand to gain by publication of these findings. D.B. and F.K. hold equity in Illumina and may stand to gain by publication of these findings.

Submitted 24 October 2011
Accepted 18 April 2012
Published 30 May 2012
10.1126/scitranslmed.3003726

Citation: T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Y. Tsui, F. Kaper, S.-J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld, Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra68 (2012).

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

Science Translational Medicine

Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA

Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana W. Y. Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M. Piskorz, Mercedes Jimenez-Linan, David Bentley, James Hadfield, Andrew P. May, Carlos Caldas, James D. Brenton and Nitzan Rosenfeld

Sci Transl Med **4**, 136ra68136ra68.
DOI: 10.1126/scitranslmed.3003726

Deep Sequencing Tumor DNA in Plasma

Five liters of circulating blood contain millions of copies of the genome, broken into short fragments; in cancer patients, a small fraction is circulating tumor DNA (ctDNA). An even smaller number harbor mutations that affect cancer outcome. Looking for diagnostic answers in circulating DNA is a challenge, but Forshew, Murtaza, and colleagues have risen to the occasion by developing a tagged-amplicon deep sequencing (TAm-Seq) method that can amplify and sequence large genomic regions from even single copies of ctDNA. By sequencing such large regions, the authors were able to identify low-level mutations in the plasma of patients with high-grade serous ovarian carcinomas.

Forshew *et al*. designed primers to amplify 5995 bases that covered select regions of cancer-related genes, including *TP53*, *EGFR*, *BRAF*, and *KRAS*. In plasma obtained from 38 patients with high levels of ctDNA, the authors were able to identify mutations in *TP53* at allelic frequencies of 2% to 65%. In plasma samples from one patient, they also identified a de novo mutation in *EGFR* that had not been detected 15 months prior in the tumor mass itself. Finally, the TAm-Seq approach was used to sequence ctDNA in plasma samples collected from two women with ovarian cancer and one woman with breast cancer at different time points, tracking as many as 10 mutations in parallel. Forshew and coauthors showed that levels of mutant alleles reflected the clinical course of the disease and its treatment—for example, stabilized disease was associated with low allelic frequency, whereas patients at relapse exhibited a rise in frequency.

Through several experiments, the authors were able to show that TAm-Seq is a viable method for sequencing large regions of ctDNA. Although this provides a new way to noninvasively identify gene mutations in our blood, TAm-Seq will need to achieve a more sensitive detection limit (<2% allele frequency) to identify mutations in the plasma of patients with less advanced cancers. Nevertheless, once optimized, this "liquid biopsy" approach will be amenable to personalized genomics, where the level and type of mutations in ctDNA would inform clinical decision-making on an individual basis.

ARTICLE TOOLS http://stm.sciencemag.org/content/4/136/136ra68

SUPPLEMENTARY MATERIALS http://stm.sciencemag.org/content/suppl/2012/05/25/4.136.136ra68.DC1

Use of this article is subject to the Terms of Service

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Translational Medicine* is a registered trademark of AAAS.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

RELATED CONTENT

http://science.sciencemag.org/content/sci/339/6119/580.full
http://stm.sciencemag.org/content/scitransmed/5/207/207ps14.full
http://stm.sciencemag.org/content/scitransmed/6/224/224ra24.full
http://stm.sciencemag.org/content/scitransmed/6/224/224ra26.full
http://stm.sciencemag.org/content/scitransmed/6/241/241ra77.full
http://stm.sciencemag.org/content/scitransmed/6/243/243fs26.full
http://stm.sciencemag.org/content/scitransmed/7/283/283ra53.full
http://stm.sciencemag.org/content/scitransmed/7/293/293ra104.full
http://stm.sciencemag.org/content/scitransmed/7/302/302ra133.full
http://science.sciencemag.org/content/sci/339/6127/1546.full
http://science.sciencemag.org/content/sci/339/6127/1559.full
http://science.sciencemag.org/content/sci/339/6127/1563.full
http://science.sciencemag.org/content/sci/339/6127/1539.full
http://science.sciencemag.org/content/sci/339/6127/1543.full
http://science.sciencemag.org/content/sci/339/6127/1567.full
http://science.sciencemag.org/content/sci/339/6127/1540.full
http://science.sciencemag.org/content/sci/339/6127/1493.full

REFERENCES

This article cites 42 articles, 16 of which you can access for free
http://stm.sciencemag.org/content/4/136/136ra68#BIBL

PERMISSIONS

http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Translational Medicine* is a registered trademark of AAAS.

Downloaded from http://stm.sciencemag.org/ by guest on December 15, 2017

www.sciencetranslationalmedicine.org/cgi/content/full/4/136/136ra68/DC1

Supplementary Materials for

Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA

Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana W. Y. Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M. Piskorz, Mercedes Jimenez-Linan, David Bentley, James Hadfield, Andrew P. May, Carlos Caldas, James D. Brenton,* Nitzan Rosenfeld*

*To whom correspondence should be addressed. E-mail: nitzan.rosenfeld@cancer.org.uk (N.R.); james.brenton@cancer.org.uk (J.D.B.)

Published 30 May 2012, *Sci. Transl. Med.* **4**, 136ra68 (2012)
DOI: 10.1126/scitranslmed.3003726

The PDF file includes:

Methods
Fig. S1. PCR strategy and primer design.
Fig. S2. Sanger traces for mutations identified by tagged-amplicon sequencing.
Fig. S3. Background frequencies and detection limits for base substitutions.
Fig. S4. Replicate dilute Sanger sequencing of a mutation identified in plasma.
Table S1. Target-specific primers.
Table S2. Unique sequencing barcodes.
Table S3. Mutations identified in FFPE samples.
Table S4. SNPs identified in circulating DNA from two plasma control samples.
Table S5. Frequency of SNP alleles in dilution series of DNA from control plasma.
Table S6. Additional data for Table 2 for mutations identified in plasma samples.
Table S7. Mutations and amplicons studied in one breast cancer patient.

SUPPLEMENTARY METHODS

Sanger sequencing

Coding sequences of the *TP53* gene (exons 2 to 11) were amplified as described previously (*36*) with the following modifications: PCR reactions were performed in 25 µl, universal primers M13 forward and M13 reverse were tagged onto both primers pairs and used to sequence in both the forward and reverse directions. To sequence exon 7, TP53-7F (CAGGTCTCCCCAAGGCGCAC) was used owing to a poly A tract downstream of the exon 7 forward primer described in (*36*): CATCCTGGCTAACGGTGAAAC. PCR products were sequenced on an ABI 3730 (Applied Biosystem). Mutational analysis was performed using Mutation Surveyor Software version 3.97 (SoftGenetics), using default settings (mutation score=5.00, mutation height=500, overlapping factor=0.20, dropping factor=0.20, SnRatio=1.00, mobility shift=imbedded algorithm). All exons were sequenced for 26 samples. For 12 samples, sequencing was performed for selected exons only.

Digital PCR

The digital PCR method and the use of the BioMark system for digital PCR analysis have been described (*7*, *37*). Samples were mixed with sequence-specific TaqMan probes, Master Mix, and 20× GE Sample Loading Reagent (Fluidigm). This mixture was then added to the 12.765 Digital Array Chip (Fluidigm) where it was partitioned into 765 separate PCR reactions. These were PCR cycled on the Fluidigm BioMark and all positive reaction chambers counted. Amplifiable copies were calculated from the number of positive wells using a Poisson correction.

Estimation of input DNA amounts

DNA amount was calculated as total amplifiable copies, including both mutant and germline alleles, as determined by digital PCR (median amplicon size 84 bp, range 58-177 bp), multiplied by the estimated amount of DNA in 1 haploid genome (3.3 pg). Amplifiable copies in control healthy plasma samples were estimated using the *FTH1* TaqMan assay Hs01694011_s1 (Life Technologies), which has an amplicon size of 180 bp.

Dilution of plasma DNA samples to determine allele quantification accuracy

After estimation of amplifiable copy numbers using digital PCR, both samples were diluted to 100 amplifiable copies per µl. A 5-µl mix was prepared containing 500 amplifiable copies, of which 400 copies were contributed by plasmaDna1 and 100 copies were contributed by plasmaDna2 (table s4). This was serially diluted 2.5× 6 times (reaching 244× dilution) in plasmaDna2. Each of the 7 mixes were then diluted 2.5× in water a further six times until we had an expected 2 total amplifiable copies of DNA. Twenty-eight of these mixes (table s5) were amplified using the described primer set (table S1) in triplicate, barcoded, and sequenced. Minor allele frequencies expected for the dilutions can be calculated using heterozygosity status for each SNP in either sample, an example that applies to 3 of the 5 SNPs (rs1800899, rs1050171 and rs10241451) we studied is given in table s5.

Replicate dilute Sanger sequencing

We performed replicate dilute Sanger Sequencing using the primers EGFR_A_F (CAGCAGGGTCTTCTCTGTTTCA) and EGFR_A_R (GGTGTTTTCACCAGTACGTTCCT). A PCR mastermix was made containing 0.05 U/μl FastStart High Fidelity Enzyme Blend (Roche), 1× FastStart High Fidelity Enzyme Buffer, 200 μM each of dNTPs, 4.5 mM $MgCl_2$, 5% DMSO, 50 nM of each primer, and 1X EvaGreen DNA binding dye (Biotium). Plasma DNA was quantified using digital PCR and diluted such that when partitioned in a 384-well plate, each reaction well would contain on average less than 1 amplifiable template. Forty-five cycles of PCR (50°C 2 min, 70°C 20 min, 95°C 10 min, 45 cycles of 95°C 15 s, 60°C 30 s, 72°C 1 min) were performed on a 7900HT Fast Real-Time PCR machine (Applied Biosystems).

TAm-Seq

Primer design for TAm-Seq

Target-specific primers were designed with universal primer sequences (termed CS1 and CS2) appended at the 5'-end (table S1). Primers were designed using Primer3 with a Tm range of 58-62°C, allowing homopolymer stretches no longer than 3 and keeping primer %GC less than 65%. Primer designs were screened against the hg19 reference genome to prevent cross-product generation, and were screened to prevent primer-dimer interactions and formation of intramolecular secondary structure. Primers were tested by amplifying from 10 ng human genomic DNA (GM17317, Coriell) in 5 μl reaction volumes containing 0.05 U/μl FastStart High Fidelity Enzyme Blend (Roche), 1× FastStart High Fidelity Enzyme Buffer, 200μM each of dNTPs, 4.5 mM MgCl2, 5% DMSO, 50 nM of each primer, and 1× Access Array sample loading solution (Fluidigm) using 35 cycles of amplification: 50°C 2 min, 70°C 20 min, 95°C 10 min, 10 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 2 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 8 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 2 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 8 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 5 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 1 cycle of 72°C for 3 min.

Barcode primers (table S2) comprised either the PE1 or PE2 sequences for Illumina cluster generation, a 10-bp barcode, followed by either CS1 or CS2 adaptor sequences. For example, one pair of barcode primers comprised 5'-PE1-CS1-3'+ 5'-PE2-BC-CS2-3', the second pair comprised 5'-PE1-CS2-3'+ 5'-PE2-BC-CS1-3'.

The sequencing reagent for read1 contained custom sequencing primers designed to anneal to both CS1 and CS2. The sequencing reagent for the index read contained custom sequencing primers designed to anneal to the reverse complements of CS1 and CS2.

Preamplification for TAm-Seq

Preamplification reactions were carried out in 10-μl reaction volumes containing 50 nM of each forward and reverse target-specific primer. Target DNA (1 to 5 μl) was added to a mastermix containing 0.5 U FastStart High Fidelity Enzyme Blend (Roche), 1× FastStart High Fidelity Enzyme Buffer, 200 μM each of dNTPs, 4.5 mM $MgCl_2$, and 5% DMSO. Reactions were subjected to 15 cycles of amplification (95°C 10 min, 15 cycles of 95°C 15 s, 60°C 4 min). Following preamplification, 4 μl Exo-SAP-it (Affymetrix) was added to each reaction and incubated for 15 min at 37°C, then for 15 min at 80°C. The preamplified samples were diluted 5-fold in PCR-grade water prior to amplification on the Access Array IFC system.

Target-specific amplification on the Access Array microfluidic system

Individual primer pairs were loaded into the primer inlets of the Access Array IFC (Fluidigm) at a final concentration of 1 µM (20× the final concentration of 50 nM required for the PCR reaction) to account for the volume ratio of sample and primer chambers (20:1) in the Access Array IFC. One µl of each preamplified sample was added to 4 µl pre-sample mix containing 0.05 U/µl FastStart High Fidelity Enzyme Blend (Roche), 1× FastStart High Fidelity Enzyme Buffer, 200 µM each of dNTPs, 4.5 mM $MgCl_2$, 5% DMSO, and 1× Access Array sample loading solution (Fluidigm). Samples and primers were loaded into the IFC using an IFC-AX controller (Fluidigm). The volume of each sample chamber within the Integrated Fluidic Circuit (IFC) was 33 nl, containing 0.7% of the input reaction volume (5 µl). The IFC was then subjected to thermal cycling using a Biomark system (Fluidigm): 35 cycles of amplification (50°C 2 min, 70°C 20 min, 95°C 10 min, 10 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 2 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 8 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 2 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 8 cycles of 95°C 15 s, 60°C 30 s, 72°C 60 s, 5 cycles of 95°C 15 s, 80°C 30 s, 60°C 30 s, 72°C 60 s, 1 cycle of 72°C for 3 min). Harvesting solution (0.05% Tween-20) was loaded onto the IFC prior to harvesting on an IFC-AX controller. One µl of harvested product was then transferred to a clean PCR plate containing 99 µl PCR-free water.

Sequencing adaptor and barcode primer addition

For each sample, 1 µl of the 100-fold diluted PCR products was added to each of two PCR plates containing 15 µl pre-sample mastermix containing 0.05 U/µl FastStart High Fidelity Enzyme Blend (Roche), 1× FastStart High Fidelity Enzyme Buffer, 200 µM each of dNTPs, 4.5 mM $MgCl_2$, and 5% DMSO. In the first plate, 4 µl of one pair of primers containing an individual 10-base barcode (BC) sequence, and sequence tags for reading in one direction (PE1-BC-CS1 + PE2-CS2) were added to each well. In the second plate, 4 µl of primers containing (PE1-BC-CS2 + PE2-CS1) were added to each well. The corresponding wells in both plates contained primers with the same barcode sequence (e.g. plate 1, well A1 = Barcode FLD0001, plate 2, well A1 = barcode FLD0001). Reaction products in plates were amplified for 15 cycles: 95°C 10 min, 15 cycles of 95°C 15 s, 60°C 30 s, 72°C 4 min, 1 cycle of 72°C for 3 min.

Quantification and clean up of DNA library

After PCR products were barcoded, they were analyzed using Agilent 2100 BioAnalyzer to ensure expected insert size (~180 bp) was obtained. They were then pooled together and purified using AMPure XP beads using a bead to amplicon ratio of 1.8:1. The library was quantified by Agilent BioAnalyzer and subjected to Illumina cluster generation. Single-end sequencing of 100 bases was performed on an Illumina GAIIx sequencer followed by a 10-base indexing (barcode) read, using custom sequencing primers targeted to the CS1 and CS2 tags for both read1 and the index read according to manufacturer's recommendations.

Analysis of sequencing data

De-multiplexing and alignment. Reads generated by the primary Illumina pipeline were demultiplexed using a known list of barcodes (table S2), allowing one base mismatch out of 10 bases. Each set of reads was aligned independently to the hg19 reference genome using bwa-short in the single-end mode. Using expected genomic positions, each set of aligned reads was separated further into its constituent amplicons. A pileup was generated for each amplicon using samtools v1.12a. Using a base quality and a mapping quality cut-off of 30, observed frequencies of non-reference alleles for every sequenced locus across all amplicons and barcodes were calculated.

Mutation identification. For each locus and base, the distribution of non-reference background allele frequencies/reads was modeled as a Normal distribution and as a Poisson distribution, fitting to a Normal distribution and using the mean number of observed non-reference reads in a set of representative barcodes as the parameter for the Poisson distribution. For each barcode, the probability of obtaining the observed frequency/number of reads (or greater) was calculated in both models, and the lower value was retained. Putative substitutions that passed a probability cut-off (confidence margin) of 0.9995 were kept for further analysis as candidate mutations.

For each barcode, all loci that passed the probability cut-off were ranked by observed frequency, corrected for the median frequency (over all samples) for each locus/base. Known SNPs obtained from the 1000 Genomes project and regions covering amplification primers were discarded, leaving 17,934 possible base calls per sample. Indels are not analyzed in the current calling algorithm. A point mutation was called in a sample if it ranked among the 6 loci with highest non-reference frequency in both duplicates, and was represented by at least 10 reads in each. If a mutation was called, it was also discarded from the pool and the process was repeated by testing the 6 highest remaining frequencies. Mutations that occurred at low frequency may have fallen outside the list of top-ranked loci, especially for more noisy samples, and may have therefore been missed. For each sample, we estimated the threshold for mutation calling by the highest background-corrected non-reference allele frequency that did not get into the top-ranked list and may have been missed by the calling algorithm. For mutations called in multiple (overlapping) amplicons, the amplicon showing best concordance between duplicates was retained.

For the sequencing data obtained from a mix of FFPE samples, we performed an additional lower-stringency analysis using a probability cut-off of 0.9995 keeping the 12 loci with highest non-reference frequency to enhance sensitivity. For the data obtained from plasma sequencing, we performed an additional higher-stringency analysis using a probability cut-off of 0.9999 keeping the 6 loci with highest non-reference frequency to enhance sensitivity.

Mutation detection. When applying TAm-Seq to measure a pre-defined mutation (as opposed to screening nearly 18,000 possible substitutions), the frequency of the mutant allele can be read out directly from the data at the desired locus. We used confidence margins of ~0.95 using a Normal distribution model for each examined substitution, and requiring a minimum of 10 representative reads, in at least one amplified library per sample. When a mutation was positively detected, its allele frequency was estimated by averaging the allele frequencies in all replicates after subtracting substitution-specific background frequencies.

SUPPLEMENTARY FIGURES

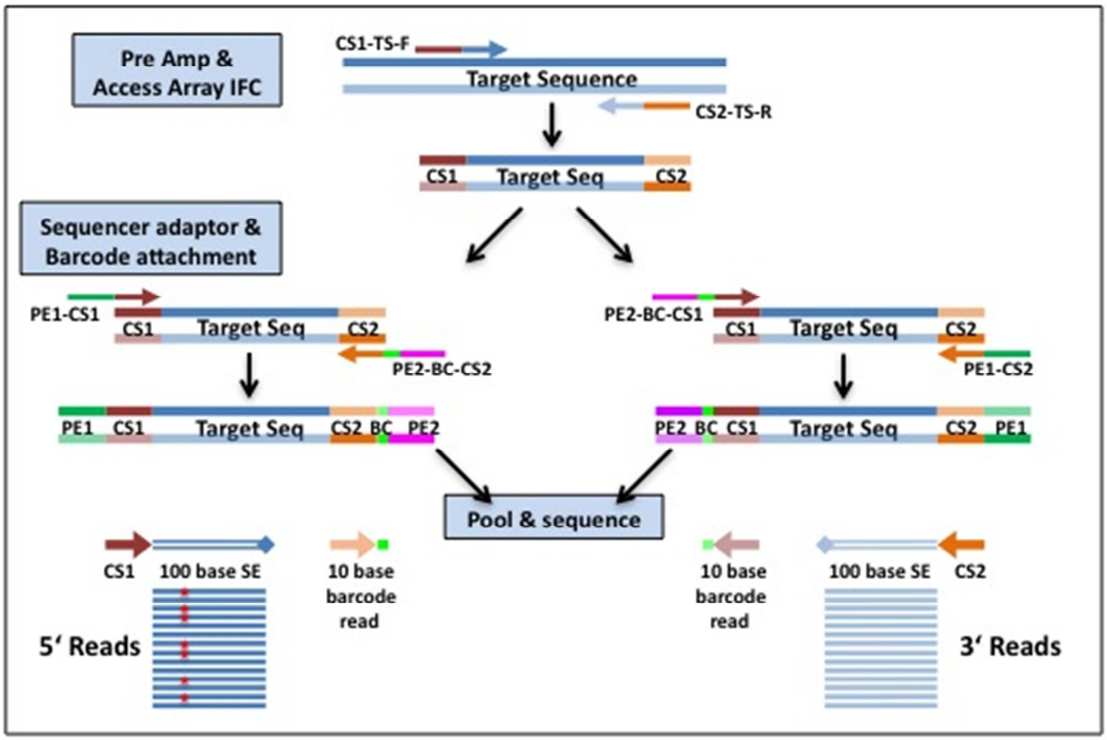


Figure S1: PCR strategy and primer design. Target-specific primers were synthesised with common adaptor sequences at their 5' ends (CS1 and CS2). These were used during the preamplification and single-plex stages to amplify the selected regions. Sequencing platform-specific adaptors and unique sample barcodes were attached during an additional round of PCR. The barcoding primers consist of Illumina sequencer adaptors (PE1 and PE2) at their 3' end, a unique 10 base barcode, and the common CS1 and CS2. Permutations of the universal tags and the sequencing adaptors enabled construction of bidirectional barcoded sequencing libraries. Using custom sequencing primers directed to the two tag sequences, sequence reads representing both strands of the amplicon pool were obtained from a single-end run. Both combinations were used in order to barcode and read amplicons in both directions. All samples were then pooled, cleaned, and then sequenced.

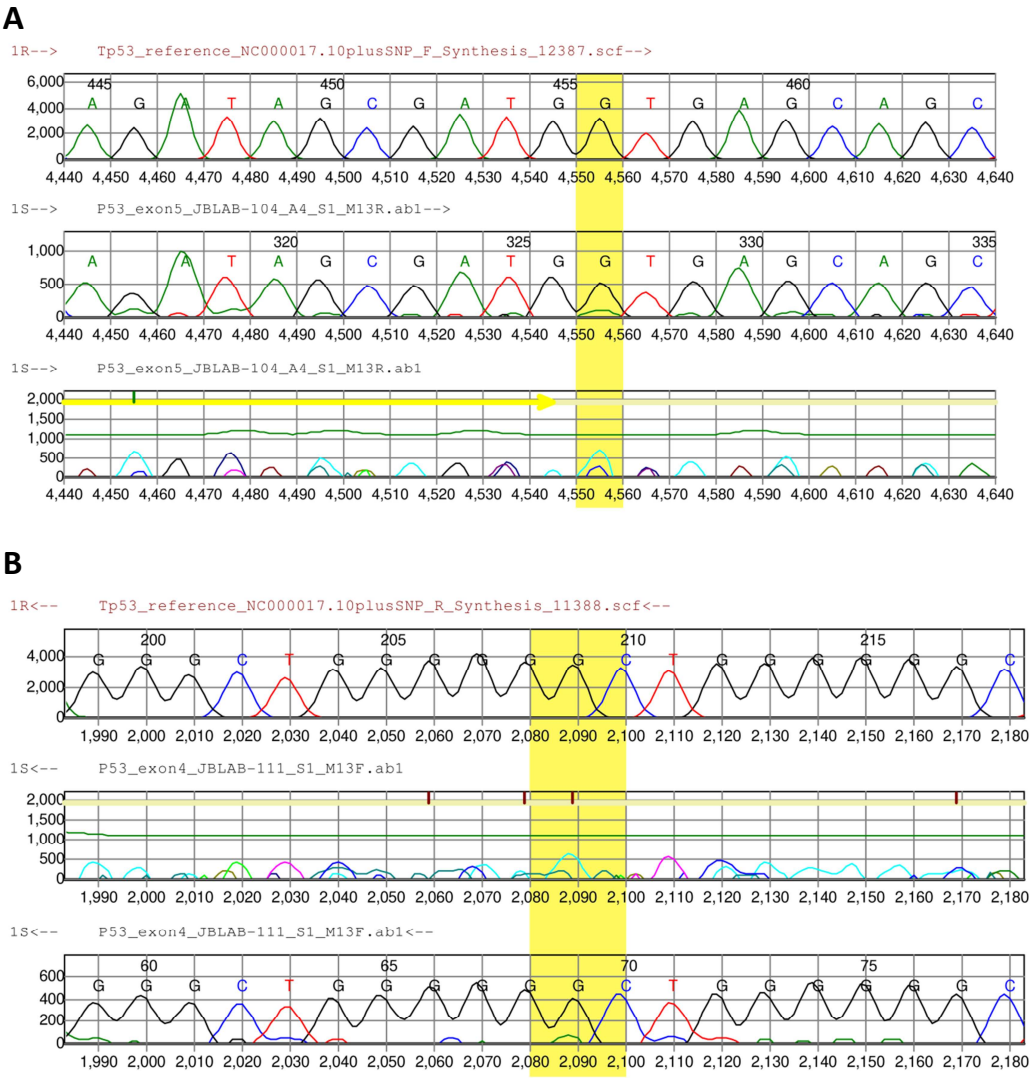


Figure S2: Sanger traces for mutations identified by tagged-amplicon sequencing. Mutational analysis of Sanger traces was performed by Mutation Surveyor Software version 3.97 (SoftGenetics) using default settings. The *TP53* gene was on the reverse strand, therefore mutations seen in the sequencing traces were the reverse complement of the mutations as indicated by the hg19 coordinates. (**A**) Sanger sequencing of sample #104. Highlighted locus shows a mutation that was identified by TAm-Seq (17:7578370C>T), but not Sanger sequencing. The Sanger sequencing trace for the sample (middle panel) shows a weak signal above the reference (top panel), but this was at background noise levels and did not pass detection thresholds (lower panel). (**B**) Sanger sequencing of sample #111. Highlighted locus shows a mutation that was identified by TAm-Seq (17:7579642C>T), but not Sanger sequencing. The Sanger sequencing trace (lower panel) showed a weak signal above the reference (top panel), but this did not pass detection thresholds (middle panel).

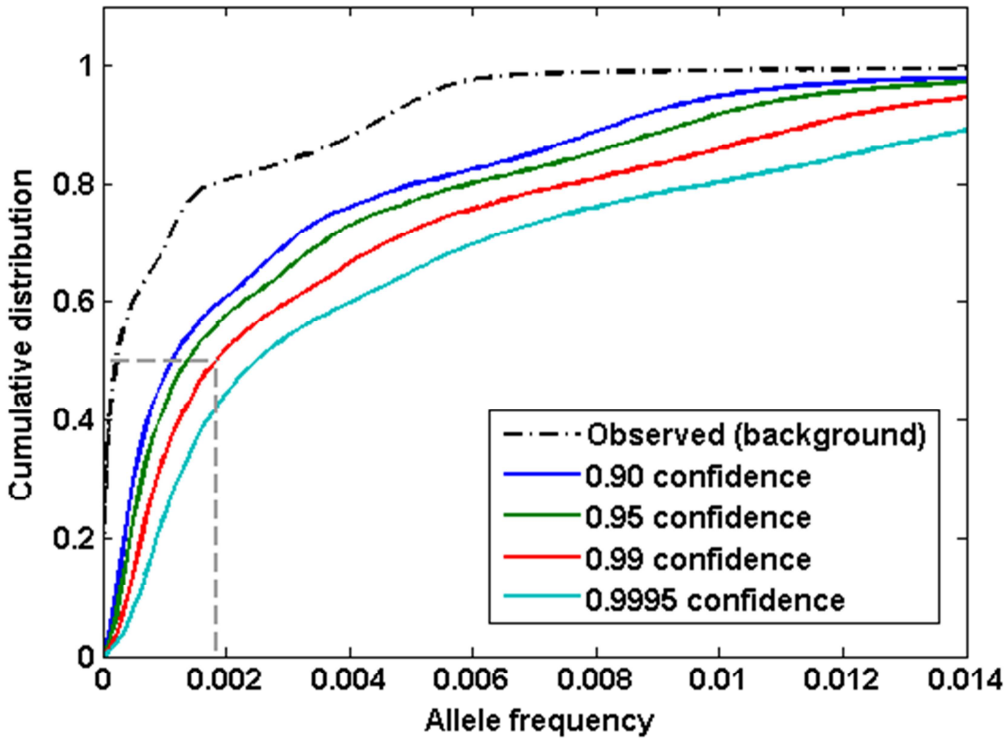


Figure S3: Background frequencies and detection limits for base substitutions. For each one of the possible base substitutions in the sequenced amplicons, we calculated the mean background read rate in the 124 barcoded libraries generated from the set of 62 plasma samples (cumulative distribution shown in black), and the standard deviation. Colored lines show the cumulative distribution of frequencies that would exceed the indicated confidence margins, using a Normal distribution model for each base substitution, given its measured mean and standard deviation. Substitutions that were covered by more than one amplicon or read direction were included multiple times, once for each amplicon/direction in which they were sequenced. Substitutions that had zero reads in all barcodes had undefined confidence limits and were omitted (n = 589). Dashed grey lines indicate the median value for 0.99 confidence margin.

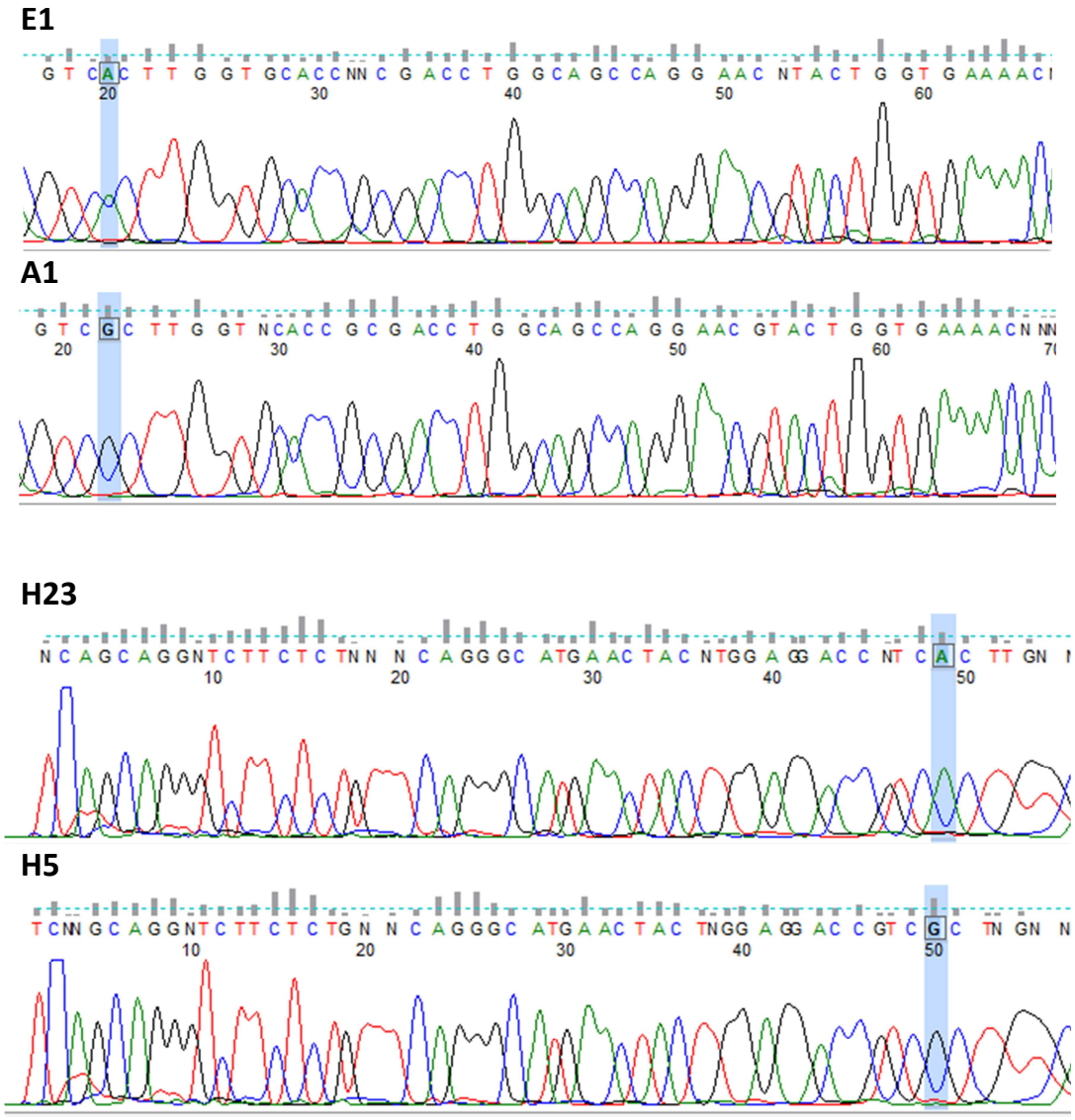


Figure S4: Replicate dilute Sanger sequencing of a mutation identified in plasma. Replicate Sanger sequencing of highly dilute template was used to validate the *EGFR* mutation that was identified *de novo* in plasma and was not found in the corresponding tumour sample from the same patient (Patient #27, Table 1). The panels show examples of mutant and wild type reads in either direction, with the mutated base (chr7:55259437G>A, genome build hg19 coordinates) highlighted in blue. E1 is an example of mutant forward read, and A1 is a matched wild-type. H23 is an example of mutant reverse read, and H5 a matched wild-type.

SUPPLEMENTARY TABLES

Table S1: Target-specific primers. Each target-specific primer consists of a universal 5' end and a target-specific 3' end as follows:

5'-ACACTGACGACATGGTTCTACA-[Target Specific-Forward]-3'.

5'-TACGGTAGCAGAGACTTGGTCT-[Target Specific-Reverse]-3'

Gene name and Amplicon ID	Target Specific Primer – Forward	Target Specific Primer- Reverse	Chr	Amplicon Start	Amplicon End
PIK3CA_E00001077674 *	CAGAGGGGAAAAATATGACAAA	AACAGAGAATCTCCATTTTAGCAC	3	178935943	178936150
PIK3CA_E00001139987	TGAGCAAGAGGCTTTGGAGT	GGTCTTTGCCTGCTGAGAGT	3	178952038	178952227
EGFR_Exon19	TCACAATTGCCAGTTAACGTCT	CCACACAGCAAAGCAGAAAC	7	55242373	55242537
EGFR_E00001601336_1	GCGTCTTCACCTGGAAGGG	CCGGACATAGTCCAGGAGG	7	55248902	55249111
EGFR_E00001601336_2	GCGTGGACAACCCCCAC	GGCTCCTTATCTCCCCTCC	7	55249005	55249213
EGFR_E00001681524_1	GGATGCAGAGCTTCTTCCCA	TTCTCTTCCGCACCCAG	7	55259352	55259542
EGFR_E00001681524_2	GGTCTTCTCTGTTTCAGGGCAT	GCTGACCTAAAGCCACCTCC	7	55259395	55259591
EGFR_E00001631695_1	GTGTCACTCGTAATTAGGTCCA	GGCCTCAGTACAAACTCATTAGC	7	55260366	55260575
EGFR_E00001779947_1	TGTTCATTCATGATCCCACTGC	CCACCAGTCACTCACACTTG	7	55266362	55266571
EGFR_E00001779947_2	TCCCTGCCAGCGAGAT	AGGGATGCAAAGGCCTCA	7	55266461	55266641
EGFR_E00001790701_1	GCCTTCTTTAAGCAATGCCATCTTTAT	CAATGGAAGCACAGACTGCAA	7	55267932	55268134
EGFR_E00001801208_1	CCCCTGCTCCTATAGCCAA	ATGAGGTACTCGTCGGCATC	7	55268806	55268987
EGFR_E00001801208_2	ACTTCTACCGTGCCCTGA	GTTCAAATGAGTAGACACAGCTT	7	55268921	55269101
EGFR_E00001773562_1	TACCCTCCATGAGGCACAC	GGAGAGCTGTAAATTCTGGCTT	7	55269336	55269516
BRAF_E00002324725	TCATAATGCTTGCTCTGATAGGA	CTGATGGGACCCACTCCAT	7	140453108	140453256
PTEN_E00001456562_1	GCAGCTTCTGCCATCTCTCT	TCCGTCTACTCCCACGTTCT	10	89624175	89624372
PTEN_E00001156351_1	TGCTGCATATTTCAGATATTTCTTTCCTTA	ATGAAAACACAACATGAATATAAACATCAAT	10	89653738	89653927
PTEN_E00001156344_1	AATCTGTCTTTTGGTTTTTCTTGATAGT	AATAGTTGTTTTAGAAGATATTTGCAAGC	10	89685172	89685367
PTEN_E00001156337_4	TATATCACTTTTAAACTTTTCTTTTAGTTGTGC	CTCGATAATCTGGATGACTCATTATTGTT	10	89690776	89690940
PTEN_E00001156330_1	TTCTTATTCTGAGGTTATCTTTTTACCAC	TCATTACACCAGTTCGTCCCT	10	89692739	89692919

PTEN_E00001156330_3	TGACCAATGGCTAAGTGAAGATGA	TCCAGGAAGAGGAAAGGAAAAACA	10	89692840	89693048
PTEN_E00001156327_1	TCTTAAATGGCTACGACCCAG	TCCAGATGATTCTTTAACAGGTAGC	10	89711775	89711942
PTEN_E00001156327_4	CAGTCAGAGGCGCTATGTGT	TCTAGATATGGTTAAGAAAACTGTTCCA	10	89711889	89712077
PTEN_E00001156321_1	TGACAGTTTGACAGTTAAAGGCAT	CACACACAGGTAACGGCTGA	10	89717547	89717726
PTEN_E00001156321_2	TGTGGTCTGCCAGCTAAAGG	TCTCCCAATGAAAGTAAAGTACAAACC	10	89717620	89717802
PTEN_E00001156321_4m	TCCACAAACAGAACAAGATGCT	GGCCTTTTCCTTCAAACAGGATT	10	89717748	89717956
PTEN_E00001156315_1m	GCAACAGATAACTCAGATTGCCTT	GTTTCCTCTGGTCCTGGTATGA	10	89720457	89720706
PTEN_E00001156315_5 *	AGGACAAAATGTTTCACTTTTGGGTAA	ACTAGATATTCCTTGTCATTATCTGCAC	10	89720649	89720799
PTEN_E00001156315_7	CCTCAGAAAAAGTAGAAAATGGAAGTC	ACAAGTCAACAACCCCCACA	10	89720706	89720915
PTEN_E00001456541_1	AGATGAGTCATATTTGTGGGTTTTCA	TCTGGATCAGAGTCAGTGGT	10	89724997	89725180
PTEN_E00001456541_2 *	GTAGAGGAGCCGTCAAATCCA	TTCATGGTGTTTTATCCCTCTTGA	10	89725068	89725264
KRAS_E00000936617	GCCTGCTGAAAATGACTGAA	AGAATGGTCCTGCACCAGTAA	12	25398163	25398329
TP53_E00001757276_1	GACCCAAAACCCAAAATGGC	TCCCTGCTTCTGTCTCCTAC	17	7572850	7573030
TP53_E00001728015_1	GGAATCCTATGGCTTTCCAACC	CCCCCTCCTCTGTTGCTG	17	7573859	7574054
TP53_00001404886_13	TCTGTATCAGGCAAAGTCATAGAA	GCCTCAAAGACAATGGCTCC	17	7576584	7576734
TP53_E00001789298_1	AGAAAACGGCATTTTGAGTGT	AAGGGTGCAGTTATGCCTCA	17	7576786	7576983
TP53_E00001789298_2	CTGGTGTTGTTGGGCAGT	ATCTCCGCAAGAAAGGGGAG	17	7576908	7577075
TP53_E00001789298_3	TGTCCTGCTTGCTTACCTCG	GCCTCTTGCTTCTCTTTTCCT	17	7577003	7577187
TP53_E00001665758_1	GGGGTCAGAGGCAAGCAG	CTTGGGCCTGTGTTATCTCC	17	7577432	7577631
TP53_E00001255919_1	GAGAAAGCCCCCCTACTGC	AGCATCTTATCCGAGTGGAAGG	17	7578091	7578274
TP53_E00001255919_3	TCCAAATACTCCACACGCAAA	GCTGCCCCCACCATGAG	17	7578229	7578406
TP53_E00001255919_5	AGCTGCTCACCATCGCTA	CCAACTGGCCAAGACCT	17	7578361	7578525
TP53_E00001255919_6	TGTGCTGTGACTGCTTGTAG	TGCCCTGACTTTCAACTCTGT	17	7578425	7578594
TP53_E00001612188_1	ATACGGCCAGGCATTGAAGT	CCTCCTGGCCCCTGTC	17	7579260	7579421
TP53_E00001612188_2	GGAAACCGTAGCTGCCCTG	AAGACCCAGGTCCAGATGAA	17	7579359	7579520
TP53_E00002359670	CAGCCTCTGGCATTCTGG	CCTGGTCCTCTGACTGCTCT	17	7579479	7579626
TP53_E00002419584	TCAAATCATCCATTGCTTGG	CCATGGGACTGACTTTCTGC	17	7579557	7579754
TP53_E00001596491_1	TTTCGCTTCCCACAGGTCTC	CAGCCAGACTGCCTTCCG	17	7579758	7579940

*Each of these 3 primer pairs potentially amplifies an additional non-target region, owing to large regions of homology and the short amplicon size. This was taken into account in data analysis.

Table S2: Unique sequencing barcodes. Platform-specific adaptors and barcodes are attached through PCR following the single-plex amplification step. The primers consisted of the PE1 and PE2 sequences for Illumina cluster generation, a 10-bp barcode, and the CS1 and CS2 adaptors, used in pairs: PE1-CS1 with PE2-BC-CS2, and PE1-CS2 with PE2-BC-CS1.

PE1: AATGATACGGCGACCACCGAGATCT.

PE2: CAAGCAGAAGACGGCATACGAGAT.

Barcode name	Barcode sequence
FLD0001	GTATCGTCGT
FLD0002	GTGTATGCGT
FLD0003	TGCTCGTAGT
FLD0004	GTCGTCGTCT
FLD0005	GTGCGTGTGT
FLD0006	GCGTCGTGTA
FLD0007	GTCGTGTACT
FLD0008	GATGTAGCGT
FLD0009	GAGTGATCGT
FLD0010	CGCTATCAGT
FLD0011	CGCTGTAGTC
FLD0012	GCTAGTGAGT
FLD0013	GAGCTAGTGA
FLD0014	CGTGCTGTCA
FLD0015	GATCGTCTCT
FLD0016	GTGCTGTCGT
FLD0017	TGAGCGTGCT
FLD0018	CATGTCGTCA
FLD0019	TCAGTGTCTC
FLD0020	GTGCTCATGT
FLD0021	CGTATCTCGA
FLD0022	GTCATGCGTC
FLD0023	CTATGCGATC
FLD0024	TGCTATGCTG
FLD0025	TGTGTGCATG
FLD0026	GAGTGTCACT
FLD0027	CTAGTCTCGT
FLD0028	GAGTGCATCT
FLD0029	TGCGTAGTCG
FLD0030	CTGTGTCGTC
FLD0031	CTGTAGTGCG
FLD0032	GTGCGCTAGT
FLD0033	TGTGCTCGCA
FLD0034	GATGCGAGCT
FLD0035	CTGTACGTGA
FLD0036	GCGATGATGA
FLD0037	TGTCGAGTCA
FLD0038	GTCTACTGTC
FLD0039	CAGTCAGAGT
FLD0040	CGCAGTCTAT
FLD0041	GTATGAGCAC
FLD0042	CGAGTGCTGT
FLD0043	TATAGCACGC
FLD0044	TCATGCGCGA
FLD0045	TATGCGCTGC
FLD0046	TCTCTGTGCA
FLD0047	CTATCGCGTG
FLD0048	TACGCTGCTG
FLD0049	CTGCATGATC
FLD0050	CGCGTATCAT
FLD0051	GTATCTCTCG
FLD0052	GCTCATATGC
FLD0053	CACTATGTCG
FLD0054	TAGCGCGTAG
FLD0055	CGTCACAGTA
FLD0056	TCGCGTGAGA
FLD0057	TACATCGCTG
FLD0058	GTGAGAGACA
FLD0059	GACTGTACGT
FLD0060	GCACGTAGCT
FLD0061	TCACGCTATG
FLD0062	CGTACTACGT
FLD0063	CAGCTGAGTA
FLD0064	GAGATCAGTC
FLD0065	TACTGAGCTG
FLD0066	TAGTAGCGCG
FLD0067	GACGTCTGCT
FLD0068	GTACTCGCGA
FLD0069	TCTGAGCGCA
FLD0070	TAGACGTGCT
FLD0071	GTGACTCGTC
FLD0072	TCGAGTAGCG
FLD0073	CGTATGATGT
FLD0074	TAGTCTGTCA
FLD0075	TGTCTCTATC
FLD0076	CTAGAGTATC
FLD0077	TATCATGTGC
FLD0078	CATGAGTGTA
FLD0079	TGTCGTCATA
FLD0080	TATCTCATGC
FLD0081	TGTGTCACTA
FLD0082	TATCGATGCT
FLD0083	TAGAGTCTGT
FLD0084	CATGCATCAT
FLD0085	TGATCAGTCA
FLD0086	CGTCTATGAT
FLD0087	GTGATACTGA
FLD0088	CTAGATCTGA
FLD0089	TATCAGTCTG
FLD0090	TCAGATGCTA
FLD0091	TATGTACGTG
FLD0092	CTATACAGTG
FLD0093	TGATACTCTG
FLD0094	TCAGCGATAT
FLD0095	CTACTGATGA
FLD0096	GTAGTACACA
FLD0097	TGCTACATCA
FLD0098	AGTGTGTCTA
FLD0099	TCATATCGCG
FLD0100	TACGTATAGC
FLD0101	CAGCTATAGC
FLD0102	TCGATGCGCT
FLD0103	GCACGCGTAT
FLD0104	GCAGTATGCG
FLD0105	TGATAGAGAG
FLD0106	GCTACTAGCG
FLD0107	TGCGAGACGT
FLD0108	CGATGACAGA
FLD0109	GACTCATGCT
FLD0110	GTCTGATACG
FLD0111	ACTAGCTGTC
FLD0112	GCGTAGACGA
FLD0113	CTCAGCAGTG
FLD0114	CAGTCTACAT
FLD0115	TACTGCAGCG
FLD0116	TACACAGTAG
FLD0117	CACATACAGT
FLD0118	CACAGTGATG
FLD0119	CGAGCTAGCA
FLD0120	GAGACTATGC
FLD0121	CAGAGCTAGT
FLD0122	CGCAGAGCAT
FLD0123	TGTACAGCGA
FLD0124	ACGTCAGTAT
FLD0125	TCACAGCATA
FLD0126	ACTGCGTGTC
FLD0127	CGATCGACTG
FLD0128	GCGAGATGTA
FLD0129	CTGATGCAGA
FLD0130	GTGACGTACG
FLD0131	CGACGCTGAT
FLD0132	CTACGATCAG
FLD0133	GCACTAGACA
FLD0134	CTAGCAGATG
FLD0135	CATGATACGC
FLD0136	GCAGCTGTCA
FLD0137	ACGTATCATC
FLD0138	AGTATCGTAC
FLD0139	GATACACTGA
FLD0140	GACTAGTCAG
FLD0141	GATGACTACG
FLD0142	CAGAGAGTCA
FLD0143	TCGATCGACA
FLD0144	ACTGATGTAG
FLD0145	ACTCGATAGT
FLD0146	GACGATCGCA
FLD0147	TCATCATGCG
FLD0148	ACATGTCTGA
FLD0149	AGTCATCGCA
FLD0150	TAGCATACAG
FLD0151	AGAGTCGCGT
FLD0152	TCTACGACAT
FLD0153	CACGAGATGA
FLD0154	ACGCACATAT
FLD0155	ACGTGCTCTG
FLD0156	ACGATCACAT
FLD0157	AGTGTACTCA
FLD0158	TGATGTATGT
FLD0159	GATATATGTC
FLD0160	TAGTACTAGA
FLD0161	TATAGAGATC
FLD0162	TCGATATCTA
FLD0163	TACATGATAG
FLD0164	TGAGATCATA
FLD0165	CTACATACTA
FLD0166	ATCAGTGTAT
FLD0167	ATCATATCTC
FLD0168	AGTAGATCAT
FLD0169	ACATAGTATC
FLD0170	ATGTATAGTC
FLD0171	ACAGTCATAT
FLD0172	ACATATACGT
FLD0173	AGCATCTATA
FLD0174	AGACTATATC
FLD0175	CAGCATCTAG
FLD0176	CGAGACGACA
FLD0177	ATCACTCATA
FLD0178	AGCTCTGTGA
FLD0179	ATGTCATGCT
FLD0180	GCTGACAGAG
FLD0181	ATACAGTCTC

FLD0182	CATAGACGTG
FLD0183	AGAGATATCA
FLD0184	ATGCTGCGCT
FLD0185	AGTCAGACGC
FLD0186	ACGATACACT
FLD0187	AGCGAGTATG
FLD0188	ATCGCTACAT
FLD0189	ATGCTAGAGA
FLD0190	AGCAGTACTC
FLD0191	ATCTAGATCA
FLD0192	ATCGCATAGA
FLD0193	TTGTTGCTGT
FLD0194	GTGTGGTTGT
FLD0195	TAGGTGGAAT
FLD0196	TGTAGGTGGA
FLD0197	TTAGTGGTGA
FLD0198	GTGAAGGTAA
FLD0199	TGTTGTGGTA
FLD0200	GTTGATGAGT
FLD0201	GGTCAGTGTA
FLD0202	GTAATGGAGT
FLD0203	CTCGTTATTC
FLD0204	GGAAGTAAGG
FLD0205	CGGTGTGTGT
FLD0206	CGTCTTCTTA
FLD0207	TGTGAATCTC
FLD0208	CTAATCGTGT
FLD0209	CTCTTAGTTC
FLD0210	GGATAGGATC
FLD0211	GGTGTCTTGT
FLD0212	GATGGTTGTA
FLD0213	CCTCGTTGTT
FLD0214	GGTTGGAGTT
FLD0215	TGGTGTCCGT
FLD0216	CGTTAGCGTA
FLD0217	TACTAGGATC
FLD0218	GTCTCAATGT
FLD0219	GATGAGGTAT
FLD0220	GGTGTTAGTG
FLD0221	CATTCTCTGA
FLD0222	CATCTGGAGT
FLD0223	GAATGGAAGA
FLD0224	GGCTGTGATC
FLD0225	TGGTGCTGGA
FLD0226	TATGGTAAGG
FLD0227	GTTCGATTGT
FLD0228	GGTAGAATGA
FLD0229	TTCTCATCGT
FLD0230	CTCAATCGTA
FLD0231	CGCTAATGTA
FLD0232	GCGTCTGAAT
FLD0233	TTCTGTTGCC
FLD0234	TTGTCCTTGC
FLD0235	CCTGTGTAGA
FLD0236	GATAAGAAGG
FLD0237	CAGGTCACAT
FLD0238	GCCATGTCAT
FLD0239	TCTGCCTATA
FLD0240	CTTAGTTCGC
FLD0241	CGTAATGAGC
FLD0242	TTGCTTAGTC
FLD0243	TCTTGTTCAC
FLD0244	GTGGCTTCGT
FLD0245	TGTTCGATAG
FLD0246	TCATTCAGTG
FLD0247	GTGGAGAGCT
FLD0248	GTAGAAGTGG
FLD0249	TGGAGCATGT
FLD0250	GAAGGAGATA
FLD0251	CGAATGTATG
FLD0252	TCGTGAATGA
FLD0253	GAATAGCTGA
FLD0254	TTGTCACATC
FLD0255	CTGGAGGCTA
FLD0256	TGTCAGCTTA
FLD0257	GTTCTTCGTA
FLD0258	TTACACGTTC
FLD0259	GTAGCCAGTA
FLD0260	TGAGAAGGTA
FLD0261	CCATATGATC
FLD0262	CGATCCTATA
FLD0263	TGACTAGCTT
FLD0264	TAACTCTGCT
FLD0265	TCGAATGTGC
FLD0266	TCGCTGAACA
FLD0267	GCGTTATTGC
FLD0268	GAACTATCAC
FLD0269	TCGAGGTACT
FLD0270	TGCGGATGGT
FLD0271	TTCGAGCTAT
FLD0272	GGTCTGGTGT
FLD0273	CTAAGTCATG
FLD0274	TTGCAGATCA
FLD0275	CTGCGAATGT
FLD0276	CTGTTCTAGC
FLD0277	CACTTGTGTG
FLD0278	TGGATGACAT
FLD0279	GATCCTGAGC
FLD0280	GTCGGTCTGA
FLD0281	TGTTACGATC
FLD0282	GTCTTGGCTC
FLD0283	GGTCGTGCAT
FLD0284	CAGGCTCAGT
FLD0285	TAGCTTCACT
FLD0286	CAGATGTCCT
FLD0287	TTACGCAGTG
FLD0288	TTCGTTCCTG
FLD0289	CACTGCTTGA
FLD0290	TCTAGCGTGG
FLD0291	GCATAATCGC
FLD0292	GTCGTAACAC
FLD0293	GAGATTGCTA
FLD0294	GGACAGATGG
FLD0295	CTTACGTTGC
FLD0296	GTGTTCGGTC
FLD0297	CTCAAGAAGC
FLD0298	TCTCGGATAG
FLD0299	CTCTGGACGA
FLD0300	CGAGCATTGT
FLD0301	CCAAGAAGAA
FLD0302	TCCTTGTTCT
FLD0303	GTAACGATGT
FLD0304	TGGACTCAGA
FLD0305	GGCATCATGC
FLD0306	GTATAACGCT
FLD0307	GCAGATAAGT
FLD0308	GTCGGCTCTA
FLD0309	TTCGATAGCA
FLD0310	GTCTAGCAGG
FLD0311	GGAACACAGG
FLD0312	TGGTTCGCTG
FLD0313	CACATTAGCG
FLD0314	GAAGCGCACT
FLD0315	GCATGCCAGT
FLD0316	GGAGACTGTA
FLD0317	TCGAACTGCA
FLD0318	GAGAGGACAT
FLD0319	GAGCACGGAA
FLD0320	GCTCTAACAT
FLD0321	TGCTGGCTTG
FLD0322	TGCATGGAGC
FLD0323	GTACTAAGAG
FLD0324	GAAGTCAAGC
FLD0325	GCGCATTATG
FLD0326	GTCCAGACAT
FLD0327	GAGACCTCTA
FLD0328	TTGCACTCAG
FLD0329	TGCGGCGATA
FLD0330	AGTTGCTAGT
FLD0331	AGGATTGAGG
FLD0332	CCAGAACAGA
FLD0333	CGTCAAGCAT
FLD0334	TTGTCGAGAC
FLD0335	GACAGGTGAC
FLD0336	CTGACAAGTG
FLD0337	CACGAAGAGC
FLD0338	CATACCTGAT
FLD0339	GACGTGCTTC
FLD0340	ATTGTGGAGT
FLD0341	TCTGGTCTCA
FLD0342	AGGTAAGAGG
FLD0343	TCCTGACAGA
FLD0344	GCACTGTTGC
FLD0345	ACCATGAGTC
FLD0346	AATGCAGTGT
FLD0347	ATATGGTGGA
FLD0348	ACTCAGTTAC
FLD0349	AAGTGCGATG
FLD0350	CCACAGAGTG
FLD0351	AGTGGTGATC
FLD0352	ACTTCTTAGC
FLD0353	GCCACATATA
FLD0354	ACGCAGGAGT
FLD0355	AATATGCTGC
FLD0356	AAGCGTAGAA
FLD0357	GACAGCAAGC
FLD0358	CTGACCGAGA
FLD0359	CGCGACTTGT
FLD0360	CATCAACATG
FLD0361	TGGCTACGCT
FLD0362	ACGCGGACTA
FLD0363	AGAGGTCGGA
FLD0364	AATCGAGCGT
FLD0365	AAGTACACTC
FLD0366	AGCTGAATGA
FLD0367	ATGCCTATCA
FLD0368	ACTGTAGGAC
FLD0369	ATAGCCGTGT
FLD0370	TCACGACGAA
FLD0371	ATCTGTCCAT
FLD0372	ACTTAGAGAG
FLD0373	AGTGGCAGGT
FLD0374	ATGAGGTCGT
FLD0375	AGGAGAAGGA
FLD0376	ACAACTGCAA
FLD0377	ATTAGCGAGT
FLD0378	ACAACGAACA
FLD0379	AGAGCGCCAA
FLD0380	AGGTAGCTCA
FLD0381	AACGCCAAGA
FLD0382	AAGGTATGAG
FLD0383	ATGGAGCACT
FLD0384	ACGGTGCTAG

Table S3: Mutations identified in FFPE samples. Diagnosis: type (FT, fallopian tube; O, ovarian; PP, primary peritoneal), histiotype (END, endometrioid, grade III; HGSOC, high-grade serous ovarian carcinoma; S/CC, mixed serous and clear cell; S/END, mixed serous papillary and endometrioid, high grade), and stage. FFPE source: sample type (B, biopsy; IDS, interval-debulking surgery; PS, primary surgery), sample tissue (FT, fallopian tube; O, ovary; OM, omentum; P, peritoneum; SC, sigmoid colon; U, uterus). Cellularity: estimated from H&E-stained slide of the FFPE block. Identified by Sanger sequencing: I, identified; FAIL, samples where one duplicate had a technical fault; Indel, indel identified (indels were not analyzed by the calling algorithm); M, missed; U, unsequenced region or sample. S-135 and S-165 were from the same patient, but collected at different timepoints. S-140 and S-106 were from the same block, cut in different batches. S-111, S-143, and S-122 had two mutations called each. S-133, S-169, and S-147 had known indels, and did not have any point mutations called. S-119 had no mutations called. S-176, S-177, and S-178 failed harvesting in one of the repeats and calling algorithm was not applied.

Patient ID	Age at diagnosis	Diagnosis: type, histiotype, and stage.			Sample ID	FFPE source		Cellularity	Gene	Mutation and base change (genome build hg19)		cDNA change	Protein change	TAm-Seq Allele frequency (repeat 1)	TAm-Seq Allele frequency (repeat 2)	Identified by Sanger sequencing
1	60	O	HGSOC	III	**S-131**	PS	O	60%	***TP53***	17:7577570	C>T	c.711G>A	p.M237I	0.72	0.70	**I**
2	63	O	HGSOC	III	**S-142**	IDS	O	40%	***TP53***	17:7578203	C>A	c.646G>T	p.V216L	0.42	0.43	**I**
3	54	O	HGSOC	III	**S-150**	IDS	O	70%	***TP53***	17:7577538	C>T	c.743G>A	p.R248Q	0.87	0.91	**I**
4	46	O	HGSOC	I	**S-167**	PS	O	70%	***TP53***	17:7578290	C>T	c.560-1G>A	p.?	0.73	0.73	**I**
5	38	O	HGSOC	III	**S-174**	PS	O	90%	***TP53***	17:7578268	A>G	c.581T>C	p.L194P	0.69	0.72	**I**
6	63	O	HGSOC	III	**S-102**	IDS	FT	60%	***TP53***	17:7578271	T>C	c.578A>G	p.H193R	0.39	0.34	**I**
7	68	O	HGSOC	III	**S-112**	IDS	O	60%	***TP53***	17:7578263	G>A	c.586C>T	p.R196*	0.30	0.28	**I**
8	60	O	HGSOC	III	**S-132**	PS	O	90%	***TP53***	17:7577120	C>T	c.818G>A	p.R273H	0.74	0.75	**I**
9	62	PP	HGSOC	IV	**S-151**	PS	SC	70%	***TP53***	17:7578479	G>A	c.451C>T	p.P151S	0.39	0.38	**I**
10	58	PP	HGSOC	IV	**S-168**	IDS	O	90%	***TP53***	17:7578508	C>T	c.422G>A	p.C141Y	0.82	0.86	**I**
11	55	O	HGSOC	III	**S-104**	IDS	O	15%	***TP53***	17:7578370	C>T	c.559+1G>A	p.?	0.12	0.12	**M**
12	62	O	HGSOC	IV	**S-115**	IDS	O	70%	***TP53***	17:7577579	G>T	c.702C>A	p.Y234*	0.57	0.59	**I**
13	45	O	HGSOC	IV	**S-123**	IDS	O	80%	***TP53***	17:7577574	T>C	c.707A>G	p.Y236C	0.53	0.55	**I**
14	58	O	HGSOC	III	**S-153**	IDS	O	95%	***TP53***	17:7578212	G>A	c.637C>T	p.R213*	0.56	0.56	**I**
15	61	O	HGSOC	IV	**S-116**	IDS	O	98%	***TP53***	17:7577121	G>A	c.817C>T	p.R273C	0.76	0.75	**I**
16	65	O	HGSOC	III	**S-127**	IDS	O	90%	***TP53***	17:7577094	G>A	c.844C>T	p.R282W	0.76	0.74	**I**
17	64	O	HGSOC	III	**S-145**	PS	FT	60%	***TP53***	17:7578176	C>T	c.672+1G>A	p.?	0.52	0.55	**I**
18	63	FT	HGSOC	III	**S-161**	PS	O	95%	***TP53***	17:7578534	C>G	c.396G>C	p.K132N	0.86	0.88	**I**
19	68	PP	HGSOC	III	**S-170**	IDS	U	50%	***TP53***	17:7579312	C>A	c.375G>T	p.T125T	0.22	0.20	**I**

20	67	PP	HGSOC	III	**S-108**	IDS	OM	20%	***TP53***	17:7578431	G>A	c.499C>T	p.Q167*	0.74	0.93	**I**
21	64	O	HGSOC	I	**S-117**	PS	O	95%	***TP53***	17:7574012	C>A	c.1015G>T	p.E339*	0.70	0.70	**I**
22	65	O	HGSOC	III	**S-128**	IDS	O	95%	***TP53***	17:7579521	C>A	c.166G>T	p.E56*	0.68	0.67	**I**
23	54	O	HGSOC	III	**S-138**	IDS	OM	95%	***TP53***	17:7577569	A>C	c.712T>G	p.C238G	0.79	0.79	**I**
24	63	FT	HGSOC	III	**S-146**	PS	O	60%	***TP53***	17:7577559	G>T	c.722C>A	p.S241Y	0.62	0.62	**I**
25	61	PP	HGSOC	III	**S-162**	IDS	OM	40%	***TP53***	17:7578404	A>T	c.526T>A	p.C176S	0.11	0.19	**I**
26	72	PP	HGSOC	III	**S-171**	IDS	OM	50%	***TP53***	17:7578265	A>T	c.584T>A	p.I195N	0.71	0.70	**I**
27	68	O	HGSOC	III	**S-179**	IDS	O	90%	***TP53***	17:7578262	C>G	c.587G>C	p.R196P	0.83	0.86	**I**
28	46	O	HGSOC	III	**S-109**	IDS	O	50%	***TP53***	17:7577548	C>T	c.733G>A	p.G245S	0.40	0.39	**I**
29	82	O	HGSOC	III	**S-118**	IDS	OM	80%	***TP53***	17:7577568	C>G	c.713G>C	p.C238S	0.74	0.75	**I**
30	69	FT	HGSOC	III	**S-130**	PS	O	80%	***TP53***	17:7579389	G>A	c.298C>T	p.Q100*	0.75	0.79	**I**
31	64	O	HGSOC	III	**S-172**	IDS	OM	70%	***TP53***	17:7578406	C>T	c.524G>A	p.R175H	0.48	0.42	**I**
32	71	O	HGSOC	III	**S-101**	IDS	P	90%	***TP53***	17:7578370	C>T	c.559+1G>A	p.?	0.95	0.94	**I**
33	73	O	HGSOC	III	**S-144**	IDS	OM	70%	***TP53***	17:7577557	A>C	c.724T>G	p.C242G	0.44	0.47	**I**
34	54	O	HGSOC	IV	**S-165**	IDS	O	95%	***TP53***	17:7578527	A>G	c.403T>C	p.C135R	0.47	0.46	**I**
					S-135	B	OM	80%	***TP53***	17:7578527	A>G	c.403T>C	p.C135R	0.57	0.60	**I**
35	55	PP	S/END	III	**S-106**	PS	O	90%	***TP53***	17:7574021	C>A	c.1006G>T	p.E336*	0.96	0.96	**I**
					S-140			90%	***TP53***	17:7574021	C>A	c.1006G>T	p.E336*	0.96	0.96	**I**
36	52	O	HGSOC	III	**S-111**	PS	O	95%	***TP53***	17:7578502	A>T	c.428T>A	p.V143E	0.45	0.44	**I**
								95%	***TP53***	17:7579642	C>T	Intron	Intron	0.22	0.28	**M**
37	45	O	END	II	**S-143**	PS	O	95%	***TP53***	17:7577538	C>T	c.743G>A	p.R248Q	0.75	0.77	**I**
								95%	***PTEN***	10:89720778	A>G	c.929A>G	p.D310G	0.59	0.52	**I**
38	55	O	END	IV	**S-122**	IDS	O	80%	***PTEN***	10:89692791	A>G	c.275A>G	p.D92G	0.19	0.18	**M**
								80%	***PTEN***	10:89685287	A>G	c.182A>G	p.H61R	0.17	0.17	**I**
39	54	O	HGSOC	III	**S-119**	PS	OM	80%								**U**
40	71	O	HGSOC	III	**S-133**	IDS	OM	80%								**Indel**
41	66	O	HGSOC	III	**S-169**	IDS	OM	50%								**Indel**
42	68	FT	S/CC	III	**S-147**	IDS	O	90%								**Indel**
43	49	FT	HGSOC	III	**S-176**	PS	O	95%							**FAIL**	
44	73	PP	HGSOC	IV	**S-177**	PS	OM	50%							**FAIL**	
45	44	O	HGSOC	III	**S-178**	IDS	OM	30%							**FAIL**	

Table S4: SNPs identified in circulating DNA from two plasma control samples.

SNP ID	Genomic location (genome build hg19) and genotypes	Normal plasma DNA 1 ("plasmaDna1")	Normal plasma DNA 2 ("plasmaDna2")
rs1625895	chr17:7578115 (T/C)	TT	TC
rs1800899	chr17:7576841 (A/G)	AG	AA
rs17337360	chr7:55260440 (C/T)	CC	TT
rs1050171	chr7:55249063 (G/A)	GA	GG
rs10241451	chr7:55248926 (T/C)	TC	TT

Table S5: Frequency of SNP alleles in dilution series of DNA from control plasma.

Total amplifiable copies \ Minor SNP percentage	200.00	80.00	32.00	12.80	5.12	2.05	0.82
2.05							40.00
5.12						40.00	16.00
12.80					40.00	16.00	6.40
32.00				40.00	16.00	6.40	2.56
80.00			40.00	16.00	6.40	2.56	1.02
200.00		40.00	16.00	6.40	2.56	1.02	0.41
500.00	40.00	16.00	6.40	2.56	1.02	0.41	0.16

Minor SNP amplifiable copies

Table S6: Additional data for Table 2 for mutations identified in plasma samples. Patient ID links to patient ID in table S3 (for patients 46 to 56, sequencing of tumour FFPE DNA was not included in table S3). Age: at diagnosis. Time: elapsed since surgery (months); chemotherapy ines: number of previous lines of chemotherapy. Vol: Plasma per amplification reaction (µL). DNA: DNA amount per amplification (in ng) was calculated from the number of total amplifiable copies (both mutant and germline allele) determined by digital PCR. Mean depth: number of sequencing reads covering this locus. AF, TAm-Seq: mean allele frequency using TAm-Seq, with the freq in each replicate given in parenthesis. Mutation call: FP, called using TAm-Seq, but not detected in FFPE DNA, and not called using more stringent calling criteria; M, mutation found in FFPE DNA and detected at allele frequency >2% in the plasma sample by digital PCR, but only found by TAm-Seq in one of the two duplicate amplicon sets; V, validated in FFPE tumour sample from the same patient (or, in the case of *EGFR* mutation in patient 27, using replicate Sanger sequencing of dilute template) and by digital PCR in plasma DNA; V*, sample failed amplification in the first run owing to a technical fault, but was reanalyzed and the mutation validated in FFPE tumour sample from the same patient and by digital PCR in plasma DNA.

Sample number	Patient ID	Age	Time (months); chemotherapy lines	Vol	DNA (ng)	Gene	Location (build hg19) and base change		Protein change	Mean depth	AF, TAm-Seq	Mean AF, digital PCR	Mutation call
1	15	61	35; 4	70	0.9	*TP53*	17:7577121	G>A	p.R273C	640	0.26 (0.227, 0.292)	0.167	V
2	3	54	44; 3	160	4.2	*TP53*	17:7577538	C>T	p.R248Q	340	0.244 (0.245, 0.242)	0.150	V
3	3	54	59; 6	160	5.7	*TP53*	17:7577538	C>T	p.R248Q	640	0.507 (0.49, 0.524)	0.410	V
4	14	58	62; 5	120	9.9	*TP53*	17:7578212	G>A	p.R213X	810	0.059 (0.056, 0.062)	0.035	V
5	10	58	48; 3	120	1.4	*TP53*	17:7578508	C>T	p.C141Y	680	0.021 (0.017, 0.024)	0.013	V
6	10	58	54; 4	120	2.1	*TP53*	17:7578508	C>T	p.C141Y	720	0.044 (0.044, 0.044)	0.038	V
7	26	72	13; 1	190	17.9	*TP53*	17:7578265	A>T	p.I195N	800	0.091 (0.081, 0.101)	0.081	V
8	31	64	21; 2	160	14.8	*TP53*	17:7578406	C>T	p.R175H	510	0.608 (0.614, 0.602)	0.627	V
9	31	64	21; 2	160	10.7	*TP53*	17:7578406	C>T	p.R175H	550	0.526 (0.529, 0.523)	0.604	V
10	31	64	21; 3	160	6.1	*TP53*	17:7578406	C>T	p.R175H	530	0.651 (0.705, 0.597)	0.682	V
11	31	64	22; 3	160	4.9	*TP53*	17:7578406	C>T	p.R175H	490	0.526 (0.548, 0.504)	0.581	V
13	47	54	12; 2	160	2.8	*TP53*	17:7578527	A>G	p.C135R	480	0.039 (0.034, 0.045)	0.045	V
14	47	54	18; 2	160	2.5	*TP53*	17:7578527	A>G	p.C135R	610	0.046 (0.047, 0.045)	0.120	V
15	47	54	20; 3	160	3	*TP53*	17:7578527	A>G	p.C135R	470	0.091 (0.133, 0.05)	0.068	V
16	27	68	25; 3	130	3.7	*TP53*	17:7578262	C>G	p.R196P	1070	0.088 (0.093, 0.084)	0.135	V
16	27	68	25; 3	130	3.7	*EGFR*	7:55259437	G>A	p.R832H	614	0.048 (0.055, 0.068, 0.008, 0.062)[†]	0.050	V

17	25	61	18; 2	160	4.2	*TP53*	17:7578404	A>T	p.C176S	580	0.113 (0.106, 0.12)	0.432	V
18	25	61	24; 3	160	4.4	*TP53*	17:7578404	A>T	p.C176S	620	0.029 (0.024, 0.033)	0.108	V
20	46	56	48; 6	140	5.2	*TP53*	17:7578406	C>T	p.R175H	650	0.201 (0.205, 0.197)	0.226	V
21	46	56	49; 6	140	3.6	*TP53*	17:7578406	C>T	p.R175H	650	0.085 (0.083, 0.087)	0.074	V
22	46	56	50; 6	140	4.1	*TP53*	17:7578406	C>T	p.R175H	630	0.081 (0.053, 0.109)	0.125	V
23	46	56	51; 6	140	3.7	*TP53*	17:7578406	C>T	p.R175H	710	0.074 (0.045, 0.103)	0.106	V
24	46	56	52; 6	140	7.1	*TP53*	17:7578406	C>T	p.R175H	760	0.269 (0.276, 0.261)	0.286	V
25	8	60	12; 1	130	3.9	*TP53*	17:7577120	C>T	p.R273H	750	0.094 (0.082, 0.106)	0.099	V
26	48	47	20; 2	160	5.7	*TP53*	17:7577094	G>A	p.R282W	640	0.048 (0.059, 0.036)	0.061	V
27	55	68	11; 2	150	3.6	*TP53*	17:7578508	C>T	p.C141Y	480	0.321 (0.342, 0.3)	0.364	V
29	50	71	0; 1	150	9.5	*TP53*	17:7577509	C>T	p.E258K	190	0.548 (0.534, 0.563)	0.253	V
31	51	62	0; 1	160	3.6	*TP53*	17:7578526	C>T	p.C135Y	620	0.04 (0.058, 0.021)	0.034	V
32	22	65	12; 1	140	2.4	*TP53*	17:7579521	C>A	p.E56X	1480	0.137 (0.147, 0.127)	0.122	V
33	18	63	48; 3	160	13.2	*TP53*	17:7578534	C>G	p.K132N	740	0.216 (0.238, 0.194)	0.206	V
34	18	63	49; 4	60	5.3	*TP53*	17:7578534	C>G	p.K132N	570	0.151 (0.173, 0.13)	0.201	V
36	18	63	56; 4	160	5.8	*TP53*	17:7578534	C>G	p.K132N	620	0.191 (0.18, 0.202)	0.275	V
37	18	63	57; 5	160	9.4	*TP53*	17:7578534	C>G	p.K132N	530	0.287 (0.245, 0.328)	0.362	V
38	18	63	58; 5	160	10.1	*TP53*	17:7578534	C>G	p.K132N	590	0.275 (0.267, 0.283)	0.331	V
39	18	63	59; 5	160	16.4	*TP53*	17:7578534	C>G	p.K132N	700	0.315 (0.302, 0.328)	0.323	V
40	18	63	60; 5	160	19.7	*TP53*	17:7578534	C>G	p.K132N	830	0.435 (0.439, 0.43)	0.482	V
41	18	63	61; 6	160	15	*TP53*	17:7578534	C>G	p.K132N	730	0.452 (0.461, 0.442)	0.445	V
42	18	63	62; 6	160	8.5	*TP53*	17:7578534	C>G	p.K132N	560	0.185 (0.2, 0.17)	0.245	V
43	56	53	15; 2	150	3.6	*TP53*	17:7578290	C>T	Splicing	680	0.143 (0.166, 0.12)	0.121	V
30	52	55	49; 4	170	5.2	*TP53*	17:7577569	A>G	p.C238R	1543	0.071(0.064, 0.077)	0.073	V*
35	18	63	50; 4	60	2.7	*TP53*	17:7578534	C>G	p.K132N	620	0.015 (0.002, 0.029)	0.073	M
12	53	51	23; 3	160	17	*PTEN*	10:89720799	T>G	Intergenic	1200	0.013 (0.014, 0.012)	-	FP
37	18	63	57; 5	160	9.4	*TP53*	17:7579790	G>A	Intronic	380	0.031 (0.031, 0.03)	-	FP

[†]The mutation locus in the *EGFR* gene was measured four separate times, in two overlapping amplicons in each of the two duplicates

Table S7: Mutations and amplicons studied in one breast cancer patient. To identify multiple mutations, genomic libraries from tumour and matched normal tissues were prepared using the standard IIlumina paired-end sample preparation kit according to the manufacturer's instructions. DNA fragments of 300 bp were sequenced using paired-end 100-bp reads on an Illumina HiSeq 2000 sequencer to 30× genomic coverage. Sequencing reads were aligned to the hg19 reference genome using ELAND (Illumina) and subsequent data analysis was performed using the CASAVA 1.8 analysis pipeline (Illumina) to identify somatic mutations. We selected 10 mutations and designed an amplicon of <120 bp to cover each mutation. We validated these mutations by TAm-Seq in the tumour material and then followed them in serial plasma samples collected from the same patient.

Gene	Genomic location of amplicon (genome build hg19)	Genomic location of mutation (genome build hg19)	Reference allele	Mutant allele
CD1A	chr1:158226710-158226810	chr1:158226762	C	T
BUB1	chr2:111430288-111430374	chr2:111430343	G	A
IQCA1	chr2:237402413-237402499	chr2:237402439	G	A
PIK3CA	chr3:178936048-178936130	chr3:178936091	G	A
ARAP3	chr5:141041694-141041773	chr5:141041739	G	A
MET	chr7:116398593-116398675	chr7:116398635	C	T
GIMAP5	chr7:150439702-150439805	chr7:150439730	C	T
ZFYVE21	chr14:104195437-104195539	chr14:104195486	G	A
KIAA0406	chr20:36641162-36641252	chr20:36641201	C	T
ADAMTS1	chr21:28214239-28214357	chr21:28214274	C	G

ATTACHMENT 1A

LDR	01903pas a2200553 a 4500
001	9915050863406676
005	20250227140630.0
007	cr|||||||||||
008	090126c20099999dcumr p | 0eng c
010	__ | **a** 2009202156
022	0_ | **a**1946-6234**l**1946-6234**9**PY**7**(Print)
022	__ | **a**1946-6242**y**1946-6234**9**EN**7**(Electronic)
035	__ | **9**101505086
040	__ | **a**NSD**c**NSD**d**NSD
041	0_ | **a**eng
042	__ | **a**pcc**a**nsdp**a**nlmcopyc
044	__ | **9**United States
060	00 | **a**W1**b**SC781Q
210	10 | **a**Sci Transl Med**2**DNLM
245	00 | **a**Science translational medicine.
246	13 | **a**STM
260	__ | **a**Washington, DC :**b**American Association for the Advancement of Science
310	__ | **a**Monthly
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	1_ | **a**Began with vol. 1, nos. 1-4 (Oct. 2009).

500	__ | **a**Description based on: Vol. 1, nos. 1-4 (Oct. 2009); title from cover.
500	__ | **a**Latest issue consulted: Vol. 1, nos. 9-12 (Dec. 2009).
500	__ | **a**Print issues ceased with v. 5 (2013).
510	1_ | **9**1**a**MEDLINE**b**v1n1, Oct 7 2009-
510	0_ | **9**1**a**PubMed**b**v1n1, Oct 7 2009-
515	__ | **a**Online version published on a weekly schedule.
530	__ | **a**Also issued online.
650	12 | **a**Translational Research, Biomedical.
655	_2 | **a**Periodical
710	2_ | **a**American Association for the Advancement of Science.
856	4_ | **u**https://www.science.org/journal/stm
911	__ | **a**Electronic only, print ceased 2013. Confirmed 20141104
914	__ | **a**Portico**b**Current to 2025 as of 20250227**c**ISSN_19466234
935	__ | **a**(DNLM)1505086-nlmdb
992	__ | **p**P2**e**EF**a**20100218
993	__ | **a**SDB**b**20100315
994	__ | **a**CDN**b**20100315
995	__ | **a**AUTH**b**20100315**c**REV**d**20250227
996	__ | **a**CONSER**b**20000101
998	__ | **a**ERMSLC 117972**b**20180329
999	__ | **a**AUTH
035	__ | **a**(OCoLC)299570656

ATTACHMENT 1B

Translational Research, Biomedical MeSH Descriptor Data 2025

MeSH Heading
Translational Research, Biomedical
Tree Number(s)
H01.770.644.145.675
Unique ID
D057170
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D057170
Annotation
TRANSLATIONAL SCIENCE, BIOMEDICAL is available for translational research as a field or the progress or status of translational research in a particular field; do not add routinely for specific studies involving "translational research"
Scope Note
The endeavour to traverse a specific step of the translation process of turning an observation in the laboratory, clinic, or community into an intervention that will improve the health of individuals and the public. (from Nat Rev Drug Discov. 2018 July ; 17(7): 455–456.)
Entry Term(s)
Translational Medical Research
Translational Research
Translational Research, Medical
Previous Indexing
Biomedical Research (2003-2009)
Research (1996-2002)
See Also
National Center for Advancing Translational Sciences (U.S.)
Public MeSH Note
2022; see TRANSLATIONAL MEDICAL RESEARCH 2012-2021; see TRANSLATIONAL RESEARCH 2010-2011
History Note
2022 (2010)
Date Established
2010/01/01
Date of Entry
2009/07/06
Revision Date
2021/07/07

Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes

Katsuyuki Shiroguchi, Tony Z. Jia, Peter A. Sims, and X. Sunney Xie[1]

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

Edited by Jonathan S. Weissman, University of California, San Francisco, CA, and approved December 14, 2011 (received for review November 1, 2011)

RNA sequencing (RNA-Seq) is a powerful tool for transcriptome profiling, but is hampered by sequence-dependent bias and inaccuracy at low copy numbers intrinsic to exponential PCR amplification. We developed a simple strategy for mitigating these complications, allowing truly digital RNA-Seq. Following reverse transcription, a large set of barcode sequences is added in excess, and nearly every cDNA molecule is uniquely labeled by random attachment of barcode sequences to both ends. After PCR, we applied paired-end deep sequencing to read the two barcodes and cDNA sequences. Rather than counting the number of reads, RNA abundance is measured based on the number of unique barcode sequences observed for a given cDNA sequence. We optimized the barcodes to be unambiguously identifiable, even in the presence of multiple sequencing errors. This method allows counting with single-copy resolution despite sequence-dependent bias and PCR-amplification noise, and is analogous to digital PCR but amendable to quantifying a whole transcriptome. We demonstrated transcriptome profiling of *Escherichia coli* with more accurate and reproducible quantification than conventional RNA-Seq.

systems biology | digital counting | next generation sequencing | gene expression | genomics

The central goal of transcriptome profiling is to accurately quantify the abundance of RNA transcripts in a sample. Although hybridization-based approaches like DNA microarrays can provide only a relative, analog measure of transcript abundance, sequencing-based approaches, such as RNA sequencing (RNA-Seq), have the advantage of removing hybridization bias among genes (1, 2) and offer the promise of true digital quantification.

The interpretation of conventional RNA-Seq is complicated by sequence-dependent bias and amplification noise from reverse transcription, adapter ligation, library amplification by PCR, solid-phase clonal amplification, and sequencing (3–5). NanoString technology mitigates these complications by eliminating enzymatic reactions and hybridizing color-coded probes directly to RNA for single-molecule detection (6), although it requires many specific probes. Other methods reduce bias in RNA-Seq by eliminating PCR and directly sequencing single molecules of RNA (7), or sequencing single molecules (8) or clonal populations (9) of cDNA. However, library amplification is desirable for sequencing small samples or single cells (10).

Conventional library amplification is based on PCR, but the exponential amplification afforded by PCR introduces noise, especially at low copy numbers (11). Digital PCR was introduced to circumvent this problem by distributing DNA molecules into many containers, each receiving zero or one molecules, which are amplified and detected by PCR (12). This technique has been successfully applied to RNA counting (13); however, it requires specific primers for each gene, which hinders high-throughput measurements.

Here we report a system-wide method for bias and noise reduction in RNA-Seq that allows the use of PCR to amplify a cDNA library before sequencing, providing accurate digital quantification of the transcriptome. In our approach, each cDNA molecule is attached to a unique barcode sequence from a large pool of barcodes before amplification (Fig. 1*A*) (14). Deep sequencing then allows quantification of the number of cDNA molecules in the original sample by counting the number of unique barcode sequences associated with a given cDNA sequence. This concept has been applied recently for studying protein-RNA interactions (15), to improve the sensitivity of DNA mutation detection (16, 17) and accuracy of DNA copy-number measurements for individual genes by threshold detection (18), and to perform karyotyping and mRNA profiling (19). However, barcode identification in these studies was not immune to errors incurred during library preparation, amplification, and sequencing, which can convert one barcode into another. Hence, a substantial fraction of reads contained misidentified barcodes (16–19), which in some cases were discarded using an artificial threshold (17–19). To avoid this complication, we designed optimized barcodes that can be ligated and amplified with minimal bias and distinguished from one another despite the accumulation of PCR mutations and sequencing errors.

Results

Barcoding Strategy for Digital RNA-Seq. Fig. 1*A* depicts the general concept of digital counting by random labeling of all target nucleic acid molecules in a sample with unique barcode sequences. To achieve unique barcoding of as many target sequences as possible, the set of barcode sequences introduced to the sample must be (*i*) much larger than the copy number of the most abundant target sequence and (*ii*) sampled randomly by the target sequences. If these two criteria are satisfied, then digital quantification of the target molecules by this method is limited only by sequencing depth and accuracy. Unlike conventional sequencing-based approaches to nucleic-acid quantification, the digital counting technique is no longer limited by intrinsic amplification noise and bias in downstream sample preparation and sequencing (Fig. 1*A*).

Implementation of the scheme in Fig. 1*A* for digital RNA-Seq requires several critical considerations. As noted above, if the barcode sequences are random, then a sequencing error at one position in a barcode will cause that barcode to be misidentified. This error-induced interconversion will occur even if the barcode sequences are nonrandom (18), unless the barcodes are carefully

Author contributions: K.S., T.Z.J., P.A.S., and X.S.X. designed research; K.S. and T.Z.J. performed research; K.S., T.Z.J., and P.A.S. analyzed the data; K.S., T.Z.J., P.A.S., and X.S.X. wrote the manuscript.

Conflict of interest statement: Harvard University has filed a provisional patent application based on this work.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE34449).

[1]To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1118018109/-/DCSupplemental.

EX1012

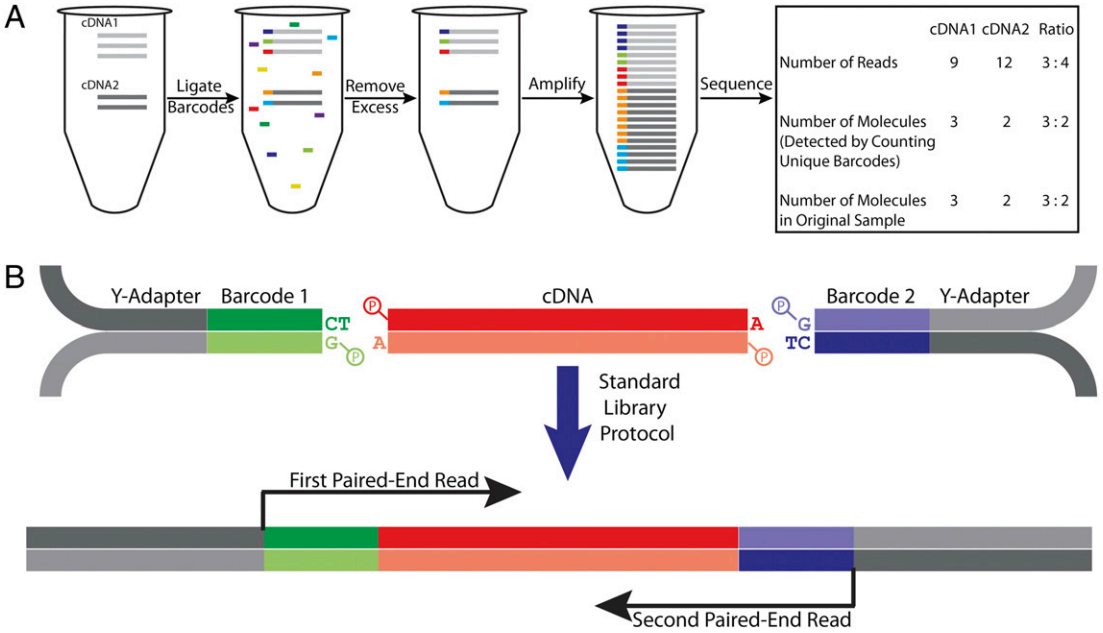


Fig. 1. Our scheme of digital RNA-Seq. (*A*) General principle of digital RNA-Seq. Assume the original sample contains two cDNA sequences, one with three copies and another with two copies. An overwhelming number of unique barcode sequences are added to the sample in excess, and five are randomly ligated to the cDNA molecules. Ideally, each cDNA molecule in the sample receives a unique barcode sequence. After removing the excess barcodes, the barcoded cDNA molecules are amplified by PCR. Because of intrinsic noise and sequence-dependent bias, the barcoded cDNA molecules are amplified unevenly. Consequently, after the amplicons are sequenced, it appears that there are three copies of cDNA1 for every four copies of cDNA2 based on the relative number of reads for each sequence. However, the ratio in the original sample was 3:2, which is accurately reflected in the relative number of unique barcodes associated with each cDNA sequence. (*B*) In our implementation of *A*, we found it advantageous to randomly ligate both ends of each phosphorylated cDNA fragment to a barcoded phosphorylated Illumina Y-shaped adapter. Note that the single T and A overhangs present on the barcodes and cDNA, respectively, are to enhance ligation efficiency. After this step, the sample is amplified by PCR and prepared for sequencing using the standard Illumina library protocol. For each amplicon, both barcode sequences and both strands of the cDNA sequence are read using paired-end deep sequencing.

designed so that multiple substitution errors and indels do not obscure their identities (20). Because DNA secondary structure can reduce amplification efficiency, the barcodes should not have significant sequence overlap or complementarity with each other, the adapter and primer sequences used in library preparation and sequencing, or the transcriptome-of-interest. Ideally, the barcode set will not contain sequence motifs that are known to be problematic for sequencing chemistries, such as long homopolymers and regions with high or low GC-content.

We used a computer program to generate a set of 145 barcode sequences 20-bp in length (Dataset S1) that satisfies the above criteria (*Materials and Methods* and *SI Materials and Methods*). The barcode sequences can sustain up to four substitution errors and remain unambiguously identifiable. In addition, a barcode that incurs up to nine substitution errors or the combination of one indel and five substitution errors will not take on the sequence of another barcode.

Instead of using a single barcode sequence to identify each target molecule in our sample (15–19), we attached a barcode sequence to both ends of each target molecule (Fig. 1*B* and *SI Materials and Methods*). If both ends of a target sequence sample all of the barcodes randomly, the target sequence will have access to $145 \times 145 = 21{,}025$ unique labels. The two barcode sequences along with the target molecule sequence were then read out by paired-end sequencing (Fig. 1*B*). This paired-end strategy dramatically reduces the number of barcodes that must be designed and synthesized, is compatible with conventional paired-end library protocols, and provides long-range sequence information that improves mapping accuracy (21, 22). In addition, attaching barcodes to both ends increases the overall randomness of barcode sampling because the two ends of a target molecule are unlikely to have a similar degree of bias. We tested and characterized this method on a set of quantified DNA spike-in sequences and a cDNA library derived from the transcriptome of *Escherichia coli*.

Quantification of Spike-in Sequences and Barcode Sampling Bias. To calibrate our digital RNA-Seq system, we measured the concentrations of five synthetic DNA spike-in sequences using the Fluidigm digital PCR platform and used them as internal standards. The spike-in samples were barcoded, added to the barcoded *E. coli* cDNA library, and quantified using the sequencing-based digital counting strategy described above. Fig. 2*A* shows that the number of digital counts (i.e., unique barcodes) observed in deep sequencing is well-correlated with the digital PCR calibration of the spike-in sequences.

To evaluate the difference between using random barcode sequences and our optimized barcode sequences, we conducted two experiments. In one experiment, we labeled the spike-in molecules with random barcode sequences (*SI Materials and Methods*), and in the second experiment we used our optimized, predetermined barcode set. We constructed the histograms of the number of reads for all barcodes observed from the most abundant spike-in sequence (Fig. 2*B*). When using random barcodes (red histogram in Fig. 2*B* and *SI Materials and Methods*), the left-most bin exhibits a large peak because a substantial fraction of barcodes are infrequently read because of sequencing errors; this causes barcodes to interconvert, generating quantification artifacts that were also evident in previous reports (16–19). In stark contrast, the left-most bin, when using optimized barcodes (green histogram in Fig. 2*B*), has no such peak because our optimized barcode sequences avoid misidentification because of sequencing errors. The effect of sequencing error on both random and optimized barcode counting is clearly shown by simulation (Fig. S1 and *SI Materials and Methods*).

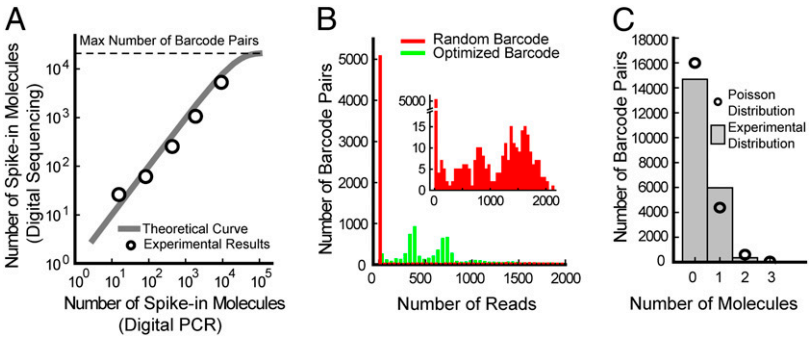


Fig. 2. Spike-in sequence quantification. (*A*) Correlation between the number of spike-in molecules for five different spike-in sequences as measured by digital PCR and digital counting of unique barcodes. The theoretical curve, which saturates because of the finite number of barcode pairs (21,025), is calculated based on the Poisson distribution (18). (*B*) Histograms of the number of reads corresponding to each observed barcode attached to the most abundant spike-in sequence for two experiments. The red histogram corresponds to a spike-in sequence labeled with random barcode sequences, and the green histogram corresponds to a spike-in sequence labeled with our optimized barcodes. Note the left-most bin in the red histogram is >10-times larger than that of the green histogram and contains a large number of unique barcodes with a low number of reads. This discrepancy is caused by various sequencing and PCR amplification errors, which generate new artifactual unique barcodes not present in the original sample and result in a large number of falsely identified unique barcodes (*SI Materials and Methods*). (*Inset*) The red histogram in greater detail. (*C*) Histogram of the number of times a barcode pair was observed with all five spike-in sequences (i.e., the number of spike-in molecules attached to a given barcode pair). Because the spike-in sequences sample the barcode pairs randomly with very little bias, the histogram follows a Poisson distribution.

We note that the green histogram in Fig. 2*B* is the distribution of the number of reads for the 5,311 uniquely barcoded molecules from a particular spike-in (*SI Materials and Methods*). Assuming each barcoded spike-in molecule is identical, the green histogram in Fig. 2*B* is essentially the probability distribution of the number of reads for a single molecule, which spans three orders-of-magnitude. This broad distribution arises primarily from intrinsic PCR amplification noise (11) in sample preparation. Given this broad single-molecule distribution, for low copy molecules in the original sample, counting the total number of reads (conventional RNA-Seq) would be catastrophic. On the other hand, this problem can be circumvented if one counts the number of different barcodes (integrated area of the histogram) using our digital RNA-Seq approach, yielding accurate quantification with single-copy resolution. The two counting schemes give the same results only when the copy number in the original sample is high, assuming there is no sequence-dependent bias.

Random sampling of the barcode sequences by each target sequence is essential for accurate digital counting. Fig. 2*C* shows that the distribution of observed molecule counts is in excellent agreement with Poisson statistics. Therefore, the five spike-in sequences sample the 21,025 barcode pairs without bias.

Digital Quantification of the *E. coli* Transcriptome. We obtained 26–32 million reads from our barcoded cDNA libraries that uniquely mapped to the *E. coli* genome (*Materials and Methods* and Dataset S2) in two replicate experiments. Fig. 3*A* shows the number of conventional and digital counts (unique barcodes) as a function of nucleotide position for the fumA transcription unit. Not surprisingly, the read density is considerably less uniform across the transcription unit than the number of digital counts, presumably because of intrinsic noise and bias in fragment amplification.

It is crucial for transcripts across the *E. coli* transcriptome to sample all barcodes evenly. Fig. 3*B* shows this distribution, which is close to Poisson but is somewhat overdispersed. Such biased sampling reduces the effective number of barcode sequences N_{eff} available. However, in our *E. coli* transcriptome sample, the copy number of the most abundant cDNA ranges from 10 to 40 copies for both counting methods. Based on Poisson statistics, even for the most abundant cDNA fragments in our sample, the required N_{eff} is ~100–400 for 95% unique labeling of all molecules (18). Because there are 21,025 barcode pairs available, on average the degree of randomness observed in Fig. 3*B* is sufficient.

The conventional method counts the number of amplicons, a quantity that is subject to bias and intrinsic amplification noise (11), rather than the number of molecules in the original sample. Conversely, in our digital counting scheme, unique barcode sequences distinguish each molecule in the sample, and so the effects of intrinsic noise are minimized. Fig. 3*C* shows how drastically different digital counting can be from conventional counting at low copy numbers, implying that digital counting of unique barcodes is advantageous, particularly for quantifying low copy fragments. We note that the correlation is stronger for high copy fragments and the same phenomenon is also observed for whole transcription units and genes (Fig. S2).

To demonstrate the superior accuracy of digital counting, we examined the uniformity of our abundance measurements within individual transcripts. Because individual transcription units were, by-and-large, intact RNA molecules following RNA synthesis, the cDNA fragments that map to one region of a given transcription unit should have the same abundance as fragments that map to a different region of the same transcription unit. We histogrammed the ratio between the variation in conventional counting ν_C and variation in digital counting ν_D for transcription units in different abundance ranges (Fig. 3*D*). A variation ratio of $\nu_C/\nu_D = 1$ indicates that both conventional and digital counting give similarly uniform abundances along the length of a transcription unit. For a transcription unit where ν_C/ν_D exceeds one, conventional counting measures abundance less consistently along the transcription unit than digital counting. The mean values of ν_C/ν_D in the two replicates are 1.4 ($s = 1.5$, where s is sample SD) and 1.2 ($s = 0.5$) for the complete set of analyzed transcription units, indicating that conventional counting is less consistent than digital counting across an average transcription unit. Furthermore, the mean value of ν_C/ν_D increases with decreasing copy number and its distribution becomes broader (Fig. 3*D*). For transcription units in the lowest abundance regime, the mean values of ν_C/ν_D are 1.9 ($s = 2.4$) and 1.3 ($s = 0.9$) for the two replicates. We conclude that, on average, digital counting outperforms conventional counting in terms of accuracy, and its performance advantage is most pronounced for low abundance transcription units.

Although Fig. 3 demonstrates that digital counting is less noisy and more accurate than conventional counting, Fig. 4 shows that digital counting is also more reproducible. We demonstrate this on the level of a single transcription unit in Fig. 4*A*, which shows

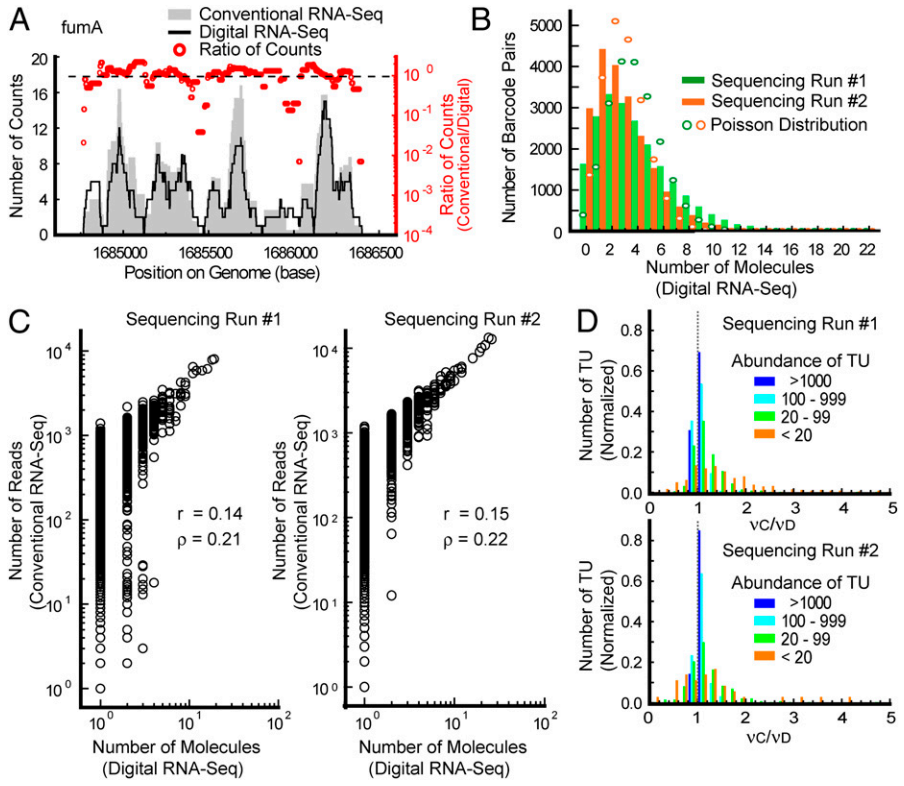


Fig. 3. Digital quantification of the *E. coli* transcriptome. (*A*) Conventional and digital counting results for the fumA transcription unit (TU) as a function of genome position. The conventional counts were calculated by using a conventional calibration curve that allows regression of the number of reads against the number of input molecules for all spike-in molecules (Fig. 2*A*). The digital counts were obtained by counting the number of unique barcodes associated with each fragment. The red dots are the ratios of these two numbers for each base. (*B*) Histograms of the number of times a barcode pair was observed with the *E. coli* cDNA sequences (i.e., the number of cDNA molecules attached to a given barcode pair) in the two replicates. Barcode sampling is more biased on average for *E. coli* cDNA fragments, but is still in reasonably good agreement with Poisson statistics. (*C*) Correlation between the number of reads (conventional counting) and the number of molecules obtained from digital counting of unique barcodes for every mapped fragment in the two replicates. For low copy molecules, the conventional counts are distributed over three orders-of-magnitude; this is because the conventional method counts amplicons, which are subject to intrinsic noise (11), rather than directly counting molecules in the original samples like the digital counting method. We note that higher copy fragments are less affected by intrinsic noise (11), as the number of molecules sequenced is greater; this effectively allows averaging over the read counts of many molecules in conventional RNA-Seq, decreasing the variance of counting in the process. (*D*) Uniformity of conventional vs. digital counting along the length of each transcription unit as a function of transcription unit abundance across the whole *E. coli* transcriptome for both replicates. We calculated the variation $\nu_D = s_D/\mu_D$ (where μ_D and s_D are the mean and sample SD of the digital counts among 99-base bins in a transcription unit, respectively) associated with digital counting and the variation $\nu_C = s_C/\mu_C$ associated with conventional counting within each transcription unit for which at least three bins contained on average at least one read. We then created the histogram of the ratio between conventional and digital counting variation (ν_C/ν_D) for transcription units in different abundance ranges for each replicate. Transcription unit abundance is the sum of all digital counts for each fragment in the transcription unit.

the ratio of counts between the two replicates for both conventional and digital counting along the fumA transcript. This ratio is consistently close to one for digital counting, but fluctuates over three orders-of-magnitude for conventional counting. We analyzed the global reproducibility of the whole transcriptome for quantification of transcription units and genes for both conventional and digital counting in Fig. 4 *B* and *C*, respectively. In both cases, the correlation between replicates is noticeably better for digital counting than conventional counting, particularly for low copy transcripts.

Discussion

Unlike previously reported methods of eliminating bias and noise from RNA-Seq (7–9), our strategy allows amplification by PCR and uses standard commercial protocols for sample preparation. However, the implementation described above also leaves considerable room for improvement. For example, one could ligate barcoded adapters directly to RNA (23, 24), reducing the bias that occurs during reverse transcription. Alternatively, a recently described protocol for processing mature mRNA from single mammalian cells could be modified to include barcoded primers for reverse transcription and second-strand synthesis before amplification (10), obviating the need for ligation.

One disadvantage of our technique is that it requires higher sequencing coverage than conventional RNA-Seq. This requirement is because both the transcriptome and the barcode set must be evenly sampled for accurate counting. However, the cost-per-base of deep sequencing continues to decrease rapidly. In our experiment, the mean number of reads per fragment was ~400. However, the spike-in sequencing reads can be randomly down-sampled 10-fold (*SI Materials and Methods*) without perturbing the correlation between abundance measured by digital PCR and digital barcode counting (Fig. S3), which implies that significantly lower coverage will suffice in many cases.

For applications where many cycles of PCR are required for sensitive detection, bias and noise reduction are crucial for accurate quantification. Although we demonstrated our technique on the *E. coli* transcriptome, we note that the maximum copy number for polyadenylated mRNA in a single mouse blastomere was found to be ~2,400 (10). With 155 optimized barcode sequences (10 more than were used in this study), one could uniquely label nearly every identical molecule in this system

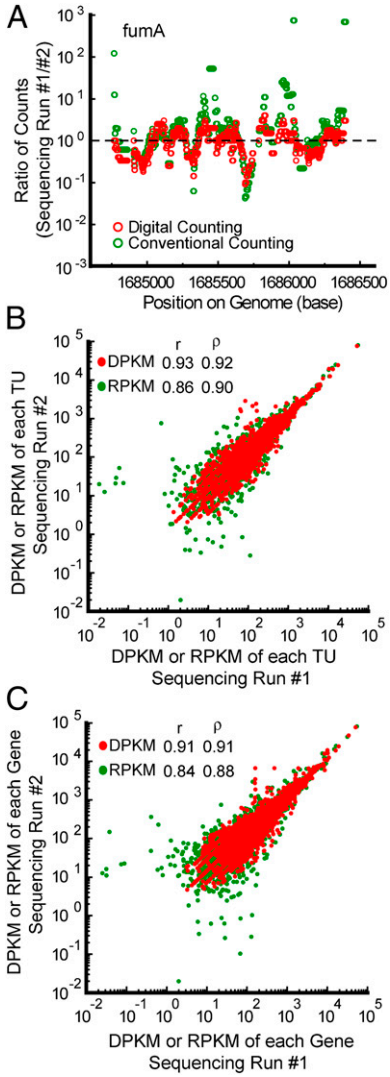


Fig. 4. Reproducibility of digital and conventional quantification of the *E. coli* transcriptome. (*A*) Ratio of counts between two replicate sequencing runs normalized by total uniquely mapped reads for digital counting plotted along with the ratio of counts between the two replicates for conventional counting of the fumA transcription unit. As expected, the ratio fluctuates over a broader range for conventional counting than digital counting along the length of the transcription unit. (*B*) Correlation between replicate sequencing runs for digital and conventional counting of transcription units. DPKM represents the uniquely mapped digital counts per kilobase per million total uniquely mapped molecules. RPKM represents the uniquely mapped reads per kilobase per million total uniquely mapped reads. (*C*) Correlation between replicate sequencing runs for digital and conventional counting of genes. Taken together, *B* and *C* demonstrate that digital counting is globally more reproducible than conventional counting.

(with 95% unique labeling for even the most abundant transcript). Hence, we expect this technique will be readily applicable to eukaryotic systems without substantial modification. In addition, we analyze the performance of digital and conventional counting in a simulation of differential expression analysis, a key application of RNA-Seq (Fig. S4). Our simulation, which accounts for experimentally measured copy number, barcode sampling bias and amplification noise distributions, shows that digital counting of unique barcodes outperforms conventional counting for differential expression analysis (*SI Materials and Methods*). Although it is always more difficult to reject the null hypothesis for low-abundance transcripts, we expect our digital counting scheme to be nonetheless more accurate than conventional counting for differential expression analysis at low copy numbers.

In addition to single-cell applications, we expect this technique to be particularly useful for nascent transcript sequencing by run-on (25) or RNA polymerase capture (26), ribosome profiling (27), and profiling of miRNA and other regulatory RNAs, which typically exist at low copy numbers. Significant recent progress has been made in minimizing bias induced by sample barcodes for multiplexed miRNA-Seq (28), and we expect that this technique could be applied to the introduction of barcodes for digital counting in any RNA-Seq experiment. In addition, one could use our approach to improve DNA sequencing experiments, such as ChIP sequencing (ChIP-Seq) (29), which is procedurally related to RNA-Seq and exposed to similar sources of bias and noise (30).

RNA-Seq holds substantial promise for basic research in biomedicine and may ultimately impact clinical diagnostics (21, 31, 32). However, challenges ranging from bias in sample preparation to limited sensitivity and remain significant. Digital RNA-Seq, along with continued improvements to sequencing technology, will lead to new applications and allow RNA-Seq to reach its full potential.

Materials and Methods

Generation and Optimization of Barcodes. We generated 2,358 random 20-base barcode candidates using a computer, such that even if a barcode accumulated nine mutations, it would not take the sequence of any other generated barcode sequences [unlike the random barcode case (*SI Materials and Methods* and Dataset S3)]. Barcode candidates containing homopolymers longer than four bases or GC-content less than 40% or greater than 60% were discarded. Barcode candidates were also discarded if each exceeded a certain degree of complementarity or sequence identity (total matches and maximum consecutive matches) with (*i*) the Illumina paired-end sequencing primers (33), (*ii*) the Illumina PCR primers PE 1.0 and 2.0, (*iii*) the 3′ end of the Illumina PCR primers PE 1.0 and 2.0, (*iv*) the whole *E. coli* genome [K-12 MG1655 strain (U00096.2)], and (*v*) all other generated barcode candidates (*SI Materials and Methods*). Any barcode candidate for which an indel mutation would place it within five point mutations of another barcode candidate was also discarded. The final population consisted of 150 barcodes, of which 145 were randomly chosen and used (Dataset S1).

***E. coli* RNA Preparation and cDNA Generation.** The cDNA library of *E. coli* [K-12 MG1655 strain (U00096.2)] was generated by a standard method (*SI Materials and Methods*).

Sample-Adapter Ligation, Sequencing Sample Preparation, and Sequencing. The cDNA library was ligated to the barcode adapter mixture, and the sequencing sample was prepared by the standard Illumina protocol with some modifications along with an internal standard (*SI Materials and Methods* and Dataset S4).

***E. coli* Transcriptome Analysis.** From the raw sequencing data, we isolated reads that contained barcode sequences that corresponded to our original list of 145 barcodes in both forward and reverse reads for each sequencing cluster that had at most one mismatch. We then aligned the first 28 bases (26 bases for the second sequencing run) of the targeted sequence of both the forward and reverse reads of each cluster to the *E. coli* genome and kept the sequences that uniquely align fewer than three mismatches and where the two reads did not map to the same sense or antisense strand of the genome. The remaining sequences were mapped to transcription units (34) and sorted by starting and ending position, as well as forward and reverse barcodes (unique tag). Mapped sequence fragments with a length of at least 1,000 bases were discarded. All sequences within the same transcription unit that had the same unique tag were analyzed further. We determined that more than one sequence with the same unique tag were identical if the distance between their center positions was less than 4 bp and if the difference in length was less than 9 bp (Figs. S5 and S6). Thus, the read counts for sequences deemed identical were summed and the sequence with more read counts was deemed as the actual correct sequence. Then, for each unique sequence, we counted the number of unique barcode tags that appeared to determine the copy number of each sequence. The genome-wide expression profile by digital counting and conventional counting are visualized by Integrated Genome Browser (*SI Materials and Methods*; see Datasets S5, S6, and S7 for specific values for datapoints shown in Fig. S2).

ACKNOWLEDGMENTS. Illumina sequencing was performed at the Harvard Faculty of Arts and Sciences Center for Systems Biology, with the help of Christian Daly, and at the Tufts University School of Medicine Genomics Core Facility, with the help of Kip Bodi and James Schiemer; digital PCR was performed by the Fluidigm Genetic Analysis Facility at the Molecular Genetics Core Facility of the Children's Hospital Boston Intellectual and Developmental Disabilities Research Center, with the help of Hal Schneider and Ta-Wei Lin; and base-calling analysis was performed at the Harvard Faculty of Arts and Sciences Research Computing Group by Jiangwen Zhang. This work was supported by National Institutes of Health (NIH) National Human Genome Research Institute Grant HG005097-01 (to X.S.X.) and NIH National Human Genome Research Institute Recovery Act Grand Opportunities Grant 1RC2HG005613-01 (to X.S.X.); and K.S. was supported by a Postdoctoral Fellowship for Research Abroad from the Japanese Society for the Promotion of Science and a fellowship from The Uehara Memorial Foundation.

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
2. Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.
3. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
4. Aird D, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.
5. Zheng W, Chung LM, Zhao H (2011) Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* 12:290.
6. Geiss GK, et al. (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* 26:317–325.
7. Ozsolak F, et al. (2009) Direct RNA sequencing. *Nature* 461:814–818.
8. Lipson D, et al. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 27:652–658.
9. Mamanova L, et al. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 7:130–132.
10. Tang F, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6:377–382.
11. Peccoud J, Jacob C (1996) Theoretical uncertainty of measurements using quantitative polymerase chain reaction. *Biophys J* 71:101–108.
12. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
13. Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314:1464–1467.
14. Hug H, Schuler R (2003) Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J Theor Biol* 221:615–624.
15. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486.
16. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res* 39:e81.
17. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
18. Fu GK, Hu J, Wang PH, Fodor SP (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA* 108:9026–9031.
19. Kivioja T, et al. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, 10.1038/nmeth.1778.
20. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5: 235–237.
21. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63.
22. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 38:4570–4578.
23. Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858–862.
24. Axtell MJ, Jan C, Rajagopalan R, Bartel DP (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell* 127:565–577.
25. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322:1845–1848.
26. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469:368–373.
27. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.
28. Alon S, et al. (2011) Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res* 21:1506–1511.
29. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316:1497–1502.
30. Nix DA, et al. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* 9:523.
31. Ozsolak F, et al. (2010) Amplification-free digital gene expression profiling from minute cell quantities. *Nat Methods* 7:619–621.
32. Shah SP, et al. (2009) Mutation of *FOXL2* in granulosa-cell tumors of the ovary. *N Engl J Med* 360:2719–2729.
33. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
34. Keseler IM, et al. (2011) EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* 39(Database issue):D583–D590.

ATTACHMENT 2A

LDR	03943cas a2200937 4500
001	994619993406676
005	20250227140822.0
007	cr|cnu||||||||
008	750626c19159999dcuwr p 0 0eng
010	__ | **a** 16010069
022	__ | **a**0027-8424**l**0027-8424**z**0744-2831**z**0273-1134**z**0273-1142**9**PY**7**(Print)
022	__ | **a**1091-6490**9**EN**7**(Electronic)
030	__ | **a**PNASA6
035	__ | **9**7505876
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
042	__ | **a**nsdp**a**lc
044	__ | **9**United States
060	00 | **a**W1**b**PR586A
210	10 | **a**Proc Natl Acad Sci U S A**2**DNLM
245	00 | **a**Proceedings of the National Academy of Sciences of the United States of America.
246	1_ | **i**Vols. for <1999-> also called:**a**PNAS
246	2_ | **0**P N A S**9**n
246	2_ | **0**PNAS USA**9**n
246	2_ | **0**PNAS**9**n
246	2_ | **0**PROC NAT ACAD SCI USA**9**n

246	2_ | **0**PROC NAT ACAD SCI**9**n
246	2_ | **0**PROC NATL ACAD SCI U S A**9**n
246	2_ | **0**PROC NATL ACAD SCI USA**9**n
246	2_ | **0**PROC NATL ACAD SCI**9**n
246	2_ | **0**PROC NATL ACAD SCI, U S A**9**n
246	2_ | **0**Proc Natl Acad Sci U**9**n
246	2_ | **0**Proc Natl Acad Sci USA**9**n
246	2_ | **0**Proc Natl Acad Sci**9**n
246	2_ | **0**Proc Natl Acad Sci, U S A**9**n
246	2_ | **0**Proc Natl Acad Sci, USA**9**n
246	2_ | **0**Proc Nattl Acad Sci USA**9**n
246	2_ | **0**Proc. Natl. Acad. Sci. USA**9**n
246	2_ | **0**Proc.Nat.Acad.Sci.Usa**9**n
246	2_ | **0**Proceedings of the National Academy of Sciences of the USA**9**n
246	2_ | **0**Proc Nat Acad Sci USA**9**om
246	2_ | **0**Proc. Nat. Acad. Sc. U. S.**9**om
260	__ | **a**Washington, DC :**b**National Academy of Sciences
300	__ | **a**v.**b**illus.
310	__ | **a**Weekly,**b**2004-
321	__ | **a**Semimonthly,**b**19??-1993
321	__ | **a**Biweekly,**b**1994-2003
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia

338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**v. 1- Jan. 15, 1915-
500	__ | **a**Not to be confused with another publication with the same title covering Mar. 1863-Oct. 1894, issued in 3 pt.: pt. 1, 1877; pt. 2, 1884; pt. 3, 1895.
500	__ | **a**Vol. 13, no. 1, pt. 2, includes Index (v. 1-10, 1915-1924) to the Proceedings; List of other publications of the Academy from 1863-1926; and List of publications of the National Research Council from 1916-1925.
500	__ | **a**Issues for 1980-1984 in two parts with consecutive page numbering: Physical sciences and Biological sciences.
510	2_ | **9**1**a**MEDLINE**b**v53n5,May 1965-
510	2_ | **9**1**a**PubMed**b**v1,1915-
510	0_ | **a**OLDMEDLINE
510	0_ | **a**PMC
510	2_ | **a**Index medicus
515	__ | **a**Some issues accompanied by supplements with distinctive titles.
530	__ | **a**Also issued online.
533	__ | **a**Microfilm.**m**v. 1 (1915)-v.15 (1929) + index v. 1 (1915)-v.10 (1924)**b**[Bethesda, Md.] :**c**National Library of Medicine,**d**1998.**e**7 microfilm reels : negative ; 35 mm.
555	__ | **a**Vols. 1-10, 1915-24, in v. 13 (Includes list of academy publications, 1863-1926, and list of publications of the National Research Council, 1916-25); Author index: Vols. 1-50, 1914-63. 1 v.
583	__ | **3**v. 16 (1930)-**a**Will not microfilm;**z**PubMed Central title.**c**20040413**5**DNLM
583	__ | **3**Vol. 84, no. 1 (JAN 1987)**c**19890818**f**Notice not published**l**Printed on acid-free paper
650	_2 | **a**Science
655	_2 | **a**Periodical

710	2_ | **a**National Academy of Sciences (U.S.)
710	2_ | **a**National Research Council (U.S.).**t**List of publications, 1916-1925.
856	4_ | **u**https://pmc.ncbi.nlm.nih.gov/journals/?term=7505876
856	4_ | **u**https://www.pnas.org/
911	__ | **a**Electronic and print, unable to acquire print format
914	__ | **a**Portico**b**Current to 2025 as of 20250227**c**ISSN_00278424
935	__ | **a**(DNLM)461999-nlmdb
995	__ | **a**AUTH**b**19750626**c**REV**d**20250227
996	__ | **a**CONSER**b**20000101
997	__ | **a**OMR
998	__ | **a**ERMSLC 170140**b**20110208
999	__ | **a**AUTH
035	__ | **a**(DNLM)P39620000(s)
035	__ | **a**(OCoLC)01607201

ATTACHMENT 2B

Science MeSH Descriptor Data 2025

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading
Science
Tree Number(s)
H01.770
Unique ID
D012586
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D012586
Annotation
use for the discipline (education, history, etc) only
Scope Note
The study of natural phenomena by observation, measurement, and experimentation.
See Also
Religion and Science
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2015/06/16

Cell-free nucleic acids as biomarkers in cancer patients

Heidi Schwarzenbach, Dave S. B. Hoon‡ and Klaus Pantel**

Abstract | DNA, mRNA and microRNA are released and circulate in the blood of cancer patients. Changes in the levels of circulating nucleic acids have been associated with tumour burden and malignant progression. In the past decade a wealth of information indicating the potential use of circulating nucleic acids for cancer screening, prognosis and monitoring of the efficacy of anticancer therapies has emerged. In this Review, we discuss these findings with a specific focus on the clinical utility of cell-free nucleic acids as blood biomarkers.

microRNAs
Small non-coding RNA molecules that modulate the activity of specific mRNA molecules by binding and inhibiting their translation into polypeptides.

**Institute of Tumour Biology, Center of Experimental Medicine, University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany.*
‡Department of Molecular Oncology, John Wayne Cancer Institute, Santa Monica, California 90404, USA.
Correspondence to K.P.
e-mail:
pantel@uke.uni-hamburg.de
doi:10.1038/nrc3066
Published online 12 May 2011

In 1948, Mandel and Métais[1] described the presence of cell-free nucleic acid (cfNA) in human blood for the first time. This attracted little attention in the scientific community and it was not until 1994 that the importance of cfNA was recognized as a result of the detection of mutated RAS gene fragments in the blood of cancer patients[2,3] (TIMELINE). In 1996, microsatellite alterations on cell-free DNA (cfDNA) were shown in cancer patients[4], and during the past decade increasing attention has been paid to cfNAs (such as DNA, mRNA and microRNAs (miRNAs)) that are present at high concentrations in the blood of cancer patients (FIG. 1). Indeed, their potential value as blood biomarkers was highlighted in a recent editorial in the journal *Science*[5].

Detecting cfNA in plasma or serum could serve as a 'liquid biopsy', which would be useful for numerous diagnostic applications and would avoid the need for tumour tissue biopsies. Use of such a liquid biopsy delivers the possibility of taking repeated blood samples, consequently allowing the changes in cfNA to be traced during the natural course of the disease or during cancer treatment. However, the levels of cfNA might also reflect physiological and pathological processes that are not tumour-specific[6]. cfNA yields are higher in patients with malignant lesions than in patients without tumours, but increased levels have also been quantified in patients with benign lesions, inflammatory diseases and tissue trauma[7]. The physiological events that lead to the increase of cfNA during cancer development and progression are still not well understood. However, analyses of circulating DNA allow the detection of tumour-related genetic and epigenetic alterations that are relevant to cancer development and progression. In addition, circulating miRNAs have recently been shown to be potential cancer biomarkers in blood.

This Review focuses on the clinical utility of cfNA, including genetic and epigenetic alterations that can be detected in cfDNA, as well as the quantification of nucleosomes and miRNAs, and discusses the relationship between cfNA and micrometastatic cells.

Biology of cfNA

The release of nucleic acids into the blood is thought to be related to the apoptosis and necrosis of cancer cells in the tumour microenvironment. Secretion has also been suggested as a potential source of cfDNA (FIG. 1). Necrotic and apoptotic cells are usually phagocytosed by macrophages or other scavenger cells[8]. Macrophages that engulf necrotic cells can release digested DNA into the tissue environment. *In vitro* cell culture experiments indicated that macrophages can be either activated or dying during the process of DNA release[8]. Fragments of cellular nucleic acids can also be actively released[9,10]. It has been estimated that for a patient with a tumour that weighs 100 g, which corresponds to 3×10^{10} tumour cells, up to 3.3% of tumour DNA may enter the blood every day[11]. On average, the size of this DNA varies between small fragments of 70 to 200 base pairs and large fragments of approximately 21 kilobases[12]. Tumour cells that circulate in the blood, and micrometastatic deposits that are present at distant sites, such as the bone marrow and liver, can also contribute to the release of cfNA[13,14].

Tumours usually represent a mixture of different cancer cell clones (which account for the genomic and epigenomic heterogeneity of tumours) and other normal cell types, such as haematopoietic and stromal cells. Thus, during tumour progression and turnover both tumour-derived and wild-type (normal) cfNA can be released into the blood. As such, the proportion of cfNA that originates from tumour cells varies owing to the state

© 2011 Macmillan Publishers Limited. All rights reserved
EX1014

At a glance

- Increased levels of circulating nucleic acids (DNA, mRNA and microRNA (miRNA)) in the blood reflect pathological processes, including malignant and benign lesions, inflammatory diseases, stroke, trauma and sepsis. During these processes nucleic acids are shed into the blood by apoptotic and necrotic cells.
- In cancer patients, circulating DNA carries tumour-related genetic and epigenetic alterations that are relevant to cancer development, progression and resistance to therapy. These alterations include loss of heterozygosity (LOH) and mutations of tumour suppressor genes (such as *TP53*) and oncogenes (such as *KRAS* and *BRAF*).
- Additional genetic alterations that are detectable on circulating DNA and used as biomarkers in cancer include the integrity of non-coding genomic DNA repeat sequences (such as *ALU* and *LINE1*). Although still in their infancy, DNA integrity assays have the potential to become a universal blood biomarker for multiple cancers.
- Epigenetic alterations in genes (such as glutathione *S*-transferase P1 (*GSTP1* and septin 9 (*SEPT9*)) and adenomatous polyposis coli (*APC*)) that are relevant to tumorigenesis and the progression of solid tumours have been detected on circulating DNA in cancer patients, and their potential clinical utility is indicated by the launch of commercial tests for cancer screening.
- The detection of circulating nucleosomes in blood indicates that cell-free DNA (cfDNA) retains at least some features of the nuclear chromatin during the process of DNA release. Initial clinical studies have indicated that monitoring the abundance of nucleosomes has potential utility for monitoring the efficacy of therapy in cancer patients.
- Dying tumour cells also discharge miRNAs, which circulate stably in the blood. The pivotal functions of miRNAs in cancer development and progression may explain the promising results of pilot studies on cancer patients using miRNA blood tests for tumour detection and prognosis.
- The cellular source of tumour-derived circulating nucleic acids is still subject to debate. After complete removal of the primary tumour the detection of cfDNA may signal the presence of micrometastatic cells in distant organs, such as the bone marrow, which pose a risk of relapse.
- Metastatic and primary tumours from the same patient can vary at the genomic, epigenomic and transcriptomic levels. Minimally invasive blood analyses of cell-free nucleic acid allow repetitive real-time monitoring of these events and will, therefore, gain clinical utility in the determination of prognosis and treatment efficacy.

and size of the tumour. The amount of cfNA is also influenced by clearance, degradation and other physiological filtering events of the blood and lymphatic circulation. Nucleic acids are cleared from the blood by the liver and kidney and they have a variable half-life in the circulation ranging from 15 minutes to several hours[7]. Assuming an exponential decay model and plotting the natural logarithm of cfDNA concentration against time, serial DNA measurements have shown that some forms of cfNA might survive longer than others. When purified DNA was injected into the blood of mice, double-stranded DNA remained in the circulation longer than single-stranded DNA[15]. Moreover, viral DNA as a closed ring may survive longer than linear DNA[15]. However, regardless of its size or configuration, cfDNA is cleared from the circulation rapidly and efficiently[16]. miRNAs seem to be highly stable, but their clearance rate from the blood has not yet been well studied in cancer patients owing to the novelty of this area of research. The nuclease activity in blood may be one of the important factors for the turnover of cfNA. However, this area of cfNA physiology remains unclear and needs further examination.

Circulating cfDNA

DNA content. In patients with tumours of different histopathological types, increased levels of total cfDNA, which consists of epigenomic and genomic, as well as mitochondrial and viral DNA, have been assessed by different fluorescence-based methods (such as, PicoGreen staining and ultraviolet (UV) spectrometry) or quantitative PCR (such as, SYBR Green and TaqMan). Although cancer patients have higher cfDNA levels than healthy control donors, the concentrations of overall cfDNA vary considerably in plasma or serum samples in both groups[17–19]. A range of between 0 and >1,000 ng per ml of blood, with an average of 180 ng per ml cfDNA, has been measured[20–23]. By comparison, healthy subjects have concentrations between 0 and 100 ng per ml cfDNA of blood, with an average of 30 ng per ml cfDNA[7]. However, it is difficult to draw conclusions from these studies, as the size of the investigated patient cohort is often small and the techniques used to quantify cfDNA vary. A large prospective study assessed the value of plasma DNA levels as indicators for the development of neoplastic or pulmonary disease. The concentration of plasma DNA varied considerably between the European Prospective Investigation into Cancer and Nutrition (EPIC) centres that were involved in the study. This variation was proposed to be due to the type of population recruited and/or the treatment of the samples[24]. However, the quantification of cfDNA concentrations alone does not seem to be useful in a diagnostic setting owing to the overlapping DNA concentrations that are found in healthy individuals with those in patients with benign and malignant disease. The assessment of cfDNA concentration might prove to be useful in combination with other blood tumour biomarkers. Following surgery, the levels of cfDNA in cancer patients with localized disease can decrease to levels that are observed in healthy individuals[25]. However, when the cfDNA level remains high, it might indicate the presence of residual tumour cells[17]. Further studies are needed for the repeat assessment of quantitative cfNA in large cohorts of patients with well-defined clinical parameters. Such investigations will be crucial if we are to use cfDNA as a prognostic biomarker, as will the isolation and processing of cfNA to defined standards (discussed below).

© 2011 Macmillan Publishers Limited. All rights reserved

cfDNA is composed of both genomic DNA (gDNA) and mitochondrial DNA (mtDNA). Interestingly, the levels of cell-free mtDNA and gDNA do not correlate in some tumour types[26,27], indicating the different nature of circulating mtDNA and gDNA. In contrast to two copies of gDNA, a single cell contains up to several hundred copies of mtDNA. Whereas gDNA usually circulates in a cell-free form, circulating mtDNA in plasma exists in both particle-associated and non-particle-associated forms[28]. Diverging results have been reported regarding whether cell-free mtDNA levels are increased and clinically relevant in cancer patients.

The cfDNA can also include both coding and non-coding gDNA that can be used to examine microsatellite instability, loss of heterozygosity (LOH), mutations, polymorphisms, methylation and integrity (size). In recent years, considerable attention has been paid to non-coding DNA, particularly repetitive sequences, such as *ALU* (which is a short interspersed nucleic element (SINE)) and as long interspersed nucleotide elements such as *LINE1* (REFS 29–31) (discussed below). *ALU* and *LINE1* are distributed throughout the genome and are known to be less methylated in cancer cells compared with normal cells[32].

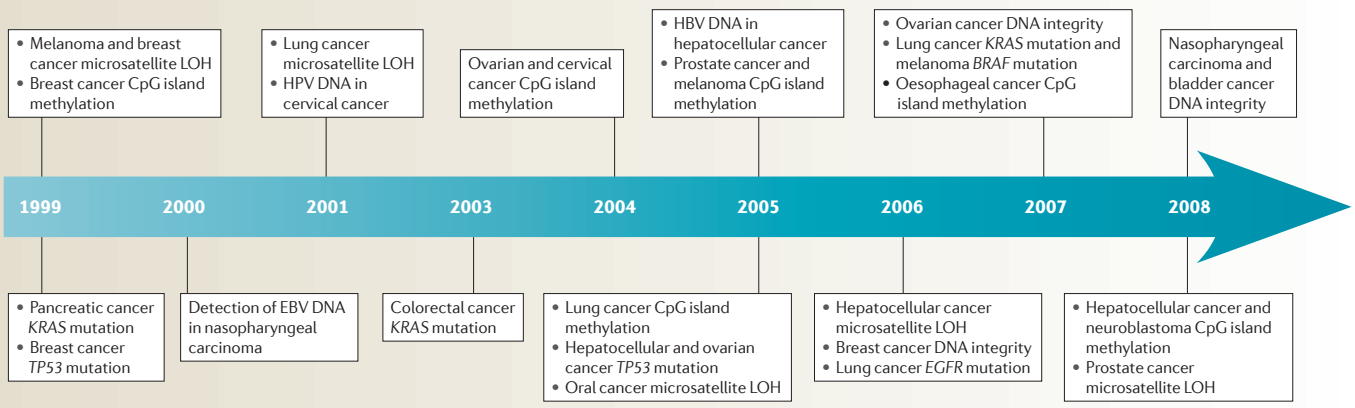
Tumour-specific LOH. Genetic alterations found in cfDNA frequently include LOH that is detected using PCR-based assays[13,18,33–38] (TABLE 1). Although similar plasma- and serum-based LOH detection methods have been used, a great variability in the detection of LOH in cfDNA has been reported. Despite the concordance between tumour-related LOH that is present in cfDNA in blood and LOH that is found in DNA isolated from matched primary tumours, discrepancies have also been found[7]. These contradictory LOH data that have been derived from blood and tumour tissue and the low incidence of LOH in cfDNA have partly been explained by technical problems and the dilution of tumour-associated cfDNA in blood by DNA released from normal cells[11,39–41]. Moreover, the abnormal proliferation of benign cells, owing to inflammation or tissue repair processes, for example, leads to an increase in apoptotic cell death, the accumulation of small, fragmented DNA in blood and the masking of LOH[42].

Alternative approaches, such as the detection of tumour-specific deletions are needed to better address the inherent problems of LOH analyses.

Tumour-specific gene mutations. The analysis of cfDNA for specific gene mutations, such as those in *KRAS* and *TP53*, is desirable because these genes have a high mutation frequency in many tumour types and contribute to tumour progression[43]. Additionally, clinically relevant mutations in *BRAF*, epidermal growth factor receptor (*EGFR*) and adenomatous polyposis coli (*APC*) have now been studied in cfDNA. Several therapeutic agents in clinical trials target the KRAS, BRAF, EGFR or p53 pathways[44,45], and require the identification of the mutation status of the patient's tumour to predict response to treatment. In this regard, cfDNA provides a unique opportunity to repeatedly monitor patients during treatment. In particular, in stage IV cancer patients, biopsies are not possible or repeat sampling of primary tumour and metastatic samples is not practical or ethical.

The major problem with this approach has been assay specificity and sensitivity. Assays targeting cfDNA mutations require that the mutation in the tumour occurs frequently at a specific genomic site. A major drawback of cfDNA assays is the low frequency of some of the mutations that occur in tumours. In general, wild-type sequences often interfere with cfDNA mutation assays. This is due to the low level of cfDNA mutations and the dilution effect of DNA fragments and wild-type DNA in circulation. In PCR-based assays technological design can significantly limit the assay sensitivity and specificity. An example is the *KRAS* mutation tissue assay that can frequently detect mutations in tumour tissues, such as the pancreas, colon and lung;

Timeline | **Detection of various forms of cfDNA in patients with different types of cancer**



The development of the detection of genetic and epigenetic alterations, as well as the measurement of DNA integrity and viral DNA, in blood from patients with different tumour types over the past decade is shown. We show only significant, prognostic findings from >40 patients with serum, plasma or bodily fluid detection of cell-free DNA (cfDNA) from individual cancers. This timeline is not meant to be comprehensive and is based on our own personal view of what have been important clinical translational events. EBV, Epstein–Barr virus; EGFR, epidermal growth factor receptor; HBV, hepatitis B virus; HPV, human papilloma virus; LOH, loss of heterozygosity.

© 2011 Macmillan Publishers Limited. All rights reserved

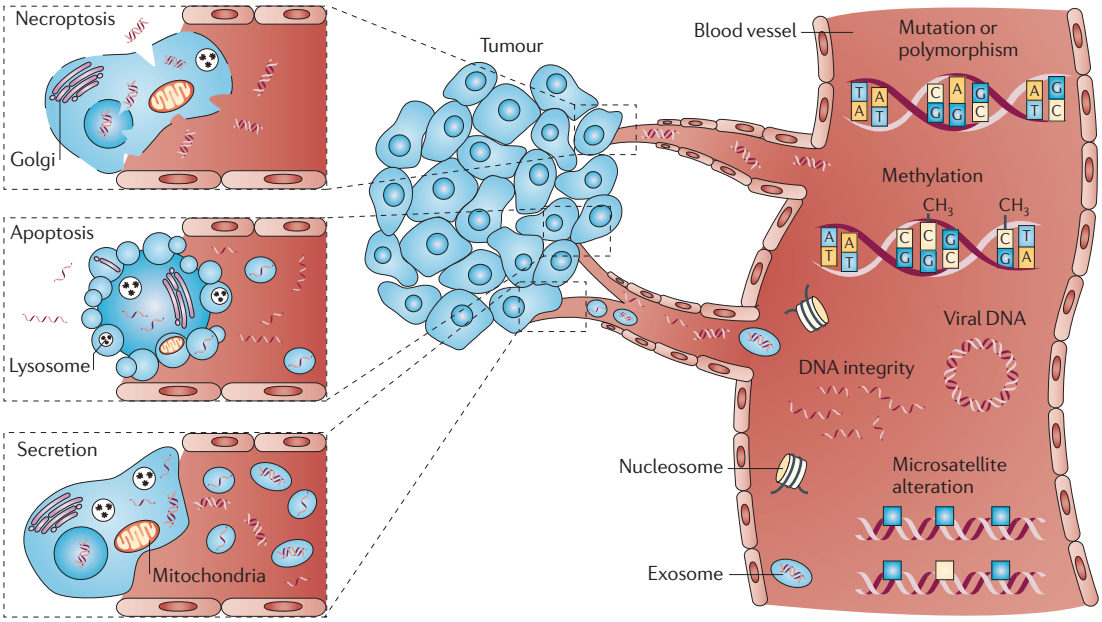


Figure 1 | **Cell-free nucleic acids in the blood.** Mutations, methylation, DNA integrity, microsatellite alterations and viral DNA can be detected in cell-free DNA (cfDNA) in blood. Tumour-related cfDNA, which circulates in the blood of cancer patients, is released by tumour cells in different forms and at different levels. DNA can be shed as both single-stranded and double-stranded DNA. The release of DNA from tumour cells can be through various cell physiological events such as apoptosis, necrosis and secretion. The physiology and rate of release is still not well understood; tumour burden and tumour cell proliferation rate may have a substantial role in these events. Individual tumour types can release more than one form of cfDNA.

however, cfDNA mutation assays using blood samples have not yet been concordantly successful[46–48]. New approaches are needed, such as cfDNA sequencing. The *BRAF* mutation V600E, which is present in >70% of metastatic melanomas, can be detected in cfDNA and has been shown to be useful in monitoring patients with melanoma who are receiving therapy[49]. This mutation has been detected in different stages of melanoma (according to the American Joint Committee on Cancer (AJCC) Cancer Staging Manual) using a quantitative real-time clamp PCR assay, with the highest levels found in the more advanced stages[49]. This is one of the first major studies to demonstrate that cfDNA mutation assays have the sensitivity to monitor patient responses before and after treatment. The utility of a cfDNA *BRAF* mutation assay has gained more importance, as new anti-BRAF drugs, such as PLX4032 (Roche)[50] and GSK2118436 (GlaxoSmithKline)[51], have shown substantial responses in patients in early clinical trials. *EGFR* mutations that occur in a specific subset of patients with lung cancers[52–54] make these tumours sensitive to EGFR-targeted therapies; however, the detection of EGFR mutations in cfDNA has not been well developed owing to issues with sensitivity and specificity. Patients whose tumours have a specific gene mutation would be strong candidates for monitoring of their cfDNA in blood for the respective specific mutation. However, sensitivity, specificity and validation need to be carried out in multicentre settings to determine true clinical utility. Alternatively, cfDNA assays might be more appropriate when used with other biomarker assays, and this might be applicable to personalized medicine, rather than diagnostic screens that can be used across a wide group of cancer patients.

DNA integrity. Another assay that is applicable to cfDNA that has gained interest in recent years is the integrity of non-coding gDNA, such as the repeat sequences of *ALU* and *LINE1*. The *ALU* and *LINE1* sequences have been referred to as 'junk DNA'; however, in recent years evidence has indicated their importance in various physiological events, such as DNA repair, transcription, epigenetics and transposon-based activity[55,56]. Approximately 17–18% of the human genome consists of *LINE1*. In normal cells *LINE1* sequences are heavily methylated, restricting the activities of these retrotransposon elements and thus preventing genomic instability. *LINE1* sequences are moderately CpG-rich, and most methylated CpGs are located in the 5′ region of the sequence that can function as an internal promoter[23]. These forms of DNA can be detected as cfDNA of different sizes, but also as methylated and unmethylated DNA. Studies on these types of cfDNA are still in their infancy; however, recent studies have shown potential prognostic and diagnostic utility[23,29–31]. The assays are based on the observation that common DNA repeat sequences are preferentially released by tumour cells that are undergoing non-apoptotic or necrotic cell death, and these fragments can be between 200 bp and 400 bp in size. The *ALU* and *LINE1* sequences are well interspersed throughout the genome on all chromosomes, so although specificity

Quantitative real-time clamp PCR assay
A technique that uses a peptide nucleic acid clamp and locked nucleic acid probes, which are DNA synthetic analogues that hybridize to complementary DNA and are highly sensitive and specific for recognizing single base pair mismatches.

© 2011 Macmillan Publishers Limited. All rights reserved

for an individual cancer type is lost in these assays, sensitivity is enhanced. Using a PCR assay, the integrity of cfDNA *ALU* sequences in blood has been shown to be sensitive for the assessment of the early stages of breast cancer progression, including micrometastasis[30]. DNA integrity cfDNA assays have also been used in testicular, prostate, nasopharyngeal and ovarian cancer[31,57–59]. These assays are still in their infancy and address an important challenge of whether a 'universal' blood biomarker for multiple cancers can be of clinical utility. Further validation of these assays will also determine their clinical utility in specific cancers.

Table 1 | **Detection of cfDNA and its alterations in patients with different tumour types***

Cancer	cfDNA	Diagnostic	Prognostic	Refs
Bladder	DNA integrity	✓	✓	123
	Methylation	✓		124
	Microsatellite alterations	✓		125
Breast	Methylation	✓	✓	126–130
	Microsatellite alterations	✓	✓	33–35
	DNA integrity		✓	30,131
	Mutation		✓	34
	Mitochondrial	✓		132
Cervical	Methylation	✓	✓	133,134
	Viral DNA	✓		135
Colorectal	Mutation	✓	✓	47,136–139
	DNA integrity	✓		31
	Methylation	✓	✓	136,140–143
Hepatocellular carcinoma	Methylation	✓	✓	144–146
	Microsatellite alterations		✓	147
	Mutation	✓	✓	148,149
	DNA integrity	✓	✓	29
	Viral DNA	✓		150
Lung	Mutation		✓	48,53,151,152
	Methylation	✓	✓	153–157
	Microsatellite alterations	✓	✓	36,37
Non-Hodgkin's lymphoma	Mutation		✓	158
	Viral DNA	✓	✓	159–161
	Methylation	✓		162
	DNA integrity	✓		162
Melanoma	Mutation	✓	✓	49,163,164
	Methylation		✓	111,115
	Microsatellite alterations	✓	✓	165–168
Ovarian	Methylation	✓	✓	169,170
	DNA integrity	✓		59
	Mutation		✓	171
	Mitochondrial	✓		172
Pancreatic	Methylation	✓		173,174
	DNA integrity	✓		31
	Mutation	✓	✓	46
Prostate	Methylation	✓	✓	38,175–179
	Microsatellite alterations	✓		13,38
	DNA integrity	✓	✓	180
	Mitochondrial		✓	26,181

*This table represents different forms of cell-free nucleic acid (cfNA) that have been detected in patients with the most prevalent cancers in both males and females[182]. This table is not meant to be comprehensive and is based on our own view of studies that offer substantial clinical insight. cfDNA, cell-free DNA.

© 2011 Macmillan Publishers Limited. All rights reserved

Epigenetic alterations. Epigenetic alterations can have a substantial effect on tumorigenesis and progression (BOX 1). Several studies have revealed the presence of methylated DNA in the serum or plasma of patients with various types of malignancy (TABLE 1). The detection of methylated cfDNA represents one of the most promising approaches for risk assessment in cancer patients.

Assays for the detection of promoter hypermethylation may have a higher sensitivity than microsatellite analyses, and can have advantages over mutation analyses. In general, aberrant DNA methylation, which seems to be common in cancer, occurs at specific CpG dinucleotides[60]. The acquired hypermethylation of a specific gene can be detected by sodium bisulphite treatment of DNA, which converts unmethylated (but not methylated) cytosines to uracil. The modified DNA is analysed using either methylation-specific PCR, with primers that are specific for methylated and unmethylated DNA, or DNA sequencing[61]. Nevertheless, to improve the assay conditions and the clinical relevance, the selection of appropriate tumour-related genes from a long list of candidate genes that are known to be methylated in neoplasia is essential. Although epigenetic alterations are not unique for a single tumour entity, there are particular tumour suppressor genes that are frequently methylated and downregulated in certain tumours[62,63]. For example, important epigenetic events in carcinogenesis include the hypermethylation of the promoter region of the genes pi-class glutathione *S*-transferase P1 (*GSTP1*) and *APC*, which are the most common somatic genome abnormalities in prostate and colorectal cancer, respectively[62,63]. Other important methylated genes that have shown prognostic utility using cfDNA assays in significant numbers of patients include RAS association domain family 1A (*RASSF1A*), retinoic acid receptor-β (*RARB*), septin 9 (*SEPT9*), oestrogen receptor-α (*ESR1*) and cyclin-dependent kinase inhibitor 2A (*CDKN2A*) (TABLE 1). The first commercial real-time PCR plasma test for the detection of early colorectal cancer (developed by Epigenomics AG and Abbott Molecular) is for the detection of *SEPT9*. This biomarker is still undergoing validation, but it demonstrates the potential diagnostic screening utility of methylated tumour-related cfDNA to differentiate cancer patients from healthy individuals and to identify the tumour type.

It is also possible to detect tumour-related alterations in histone modifications in the blood. By monitoring changes in the circulating histones and DNA methylation pattern, the antitumour effects of histone deacetylase and histone methyltransferase inhibitors may be evaluated and consequently allow a better screening of cancer patients[64,65].

Circulating nucleosomes. Circulating gDNA that is derived from tumours seems to predominantly exist as mononucleosomes and oligonucleosomes, or it is bound to the surface of blood cells by proteins that harbour specific nucleic acid-binding properties[66]. A nucleosome consists of a histone octamer core wrapped twice by a 200 bp-long DNA strand. Under physiological conditions these complexes are packed in apoptotic particles and engulfed by macrophages[67]. However, an excess of apoptotic cell death, as occurs in large and rapidly proliferating tumours or after chemotherapy treatment, can lead to a saturation of apoptotic cell engulfment and thus increased nucleosome levels in the blood[68]. The detection of circulating nucleosomes that are associated with cfDNA suggests that DNA in blood retains at least some features of the nuclear chromatin during the process of release.

Enzyme-linked immunosorbent assays (ELISAs) have been developed to quantify circulating nucleosomes. As increased concentrations are found in both benign and malignant tumours, high nucleosome levels in blood are not indicators of malignant disease[69]. However, the observed changes in apoptosis-related deregulation of proteolytic activities along with the increased serum levels of nucleosomes have been linked to breast cancer progression[70]. As typical cell-death products, the quantification of circulating nucleosomes seems to be valuable for monitoring the efficacy of cytotoxic cancer therapies[71]. For example, platinum-based chemotherapy induces caspase-dependent apoptosis of tumour cells and an increase in circulating nucleosomes in the blood of patients with ovarian cancer[17]. Moreover, the outcome of therapy can be indicated by nucleosome levels during the first week of chemotherapy and radiotherapy in patients with lung, pancreatic and colorectal cancer, as well as in patients with haematological malignancies[71].

Viral DNA. Viral cfDNA can also be detected in some tumour types. Viruses, such as human papillomavirus (HPV), hepatitis B virus (HBV) and Epstein–Barr virus (EBV), are aetiological factors in various malignancies, such as nasopharyngeal, cervical, head and neck, and hepatocellular cancer and lymphoma[72–75]. Their specific DNA may have the potential to be used as molecular biomarkers for neoplastic disease. Associations between circulating viral DNA and disease have been reported for EBV with Hodgkin's disease, Burkitt's lymphoma and nasopharyngeal carcinoma; for HBV with some forms of hepatic cell carcinoma; and for HPV with head and neck, cervical and hepatocellular cancers (TABLE 1). The clinical utility of EBV cfDNA in diagnosis and prognosis of nasopharyngeal carcinoma has been demonstrated in multiple studies with large cohorts of patients[76–80], and the use of this cfDNA has became one of the leading cfDNA blood tests for the assessment of nasopharyngeal carcinoma progression in Hong Kong, Taiwan and China, where this cancer is highly prevalent[77,78,81]. The limitation

Box 1 | **Epigenetics**

Epigenetic changes can include the methylation of gene promoter regions and histone modifications. In chromosomal regions where tumour-associated genes reside, epigenetic modifications may affect important regulatory mechanisms that normally limit malignant transformation[60]. Inactivation of tumour suppressor genes by promoter hypermethylation is thought to have a crucial role in this process[117]. DNA methylation of the cytosine base in CpG dinucleotides, which are found as isolated or clustered CpG islands, induces gene repression by inhibiting the access of transcription factors to their binding sites, and by recruiting methyl-CpG-binding proteins (MBDs) to methylated DNA together with histone-modifying enzymes[118]. Epigenetic modifications also alter the packing of nucleosomes that are implicated in transcriptional regulation[119,120].

© 2011 Macmillan Publishers Limited. All rights reserved

of most viral cfNA assays is that benign viral infections that are caused by the same viruses can complicate the interpretation of results, particularly in diagnostic screening. Establishing clinically meaningful cut-off levels is important to move these screens into the clinic.

Genometastasis. The genometastasis hypothesis describes the horizontal transfer of cell-free tumour DNA to other cells that results in transformation. If true, metastases could develop in distant organs as a result of a transfection-like uptake of dominant oncogenes that are released from the primary tumour[82]. García-Olmo *et al.*[83] showed that plasma isolated from patients with colon cancer is able to transform NIH-3T3 cells and that these cells can form carcinomas when injected into non-obese diabetic-severe combined immunodeficient mice[83]. Whether this biological function of circulating DNA has relevance in human blood is an aspect to be considered in the future.

cfDNA assay issues

One of the problems in evaluating cfNA is the standardization of assays, such as isolation technologies, standards, assay conditions, and specificity and sensitivity[7]. It remains controversial whether plasma or serum is the optimal sampling specimen. The diversity of protocols and reagents currently in use impedes the comparison of data from different laboratories.

The pre-analytical phases of cfDNA analysis such as blood collection, processing (plasma and serum), storage, baseline of patients, diurnal variations and accurate clinical conditions need to be better defined before comparisons and clinical utility can be validated[84]. A major technical issue that hampers consistency in all the cfDNA assays is the efficacy of the extraction procedures, with only small amounts of DNA obtained from plasma and serum. Another key issue is quantification before assessment on specific assay platforms. Improvement is needed in these aspects for cfDNA analysis to be more robust, consistent, comparative and informative. Extraction of cfDNA can be carried out in accordance with many methods; for example, commercial kits, company in-house procedures or individual laboratory protocols. To date no approach has been truly developed that is consistent, robust, reproducible, accurate, and validated on a large-scale patient and normal donor population. If these issues were solved a better universal standardization for the comparison of results would be provided and the clinical utility of the assays could be addressed. The development of a direct DNA assay without extraction would override many of these problems[30]. As new approaches in the assessment of cfDNA, such as next-generation sequencing, are being developed, the issue of extraction of DNA will continue to complicate cfDNA biomarker assay development and regulatory group approval.

The other major issue for cfDNA assessment is that after DNA extraction, different platform assays are used for analysis. This can vary owing to the type of cfDNA being analysed, assay sensitivity and specificity, and analytical approach. These variables are important and need to be standardized for consensus analysis and reporting. The development of PCR-based assay standardization is needed in order to report clinical and prognostic biomarker results that are similar to those outlined in the recent minimal information for the publication of quantitative real-time PCR guidelines[85]. However, this may take time to reach an international consensus, as has been apparent with the standardization of other cancer blood biomarkers. Unfortunately, the rate of approval of new cancer blood biomarkers over the past decade has been very slow. New regulatory guidelines, such as those listed for tumour biomarkers in clinical practice by the National Academy of Clinical Biochemistry (NACB USA)[84], should help to resolve some of these issues. The NACB website provides up-to-date informative detailed guidelines with references of pre-analytical and post-analytical phases, assay validation, internal quality controls, proficiency and requirements for minimizing the risk of method-related errors for biomarkers. Nevertheless, as with other types of biomarkers, new regulatory guidelines mean that developing cfNA biomarkers will be more time-consuming and costly.

Circulating cfRNA

mRNA content. Besides the quantification of cfDNA, circulating gene transcripts are also detectable in the serum and plasma of cancer patients. It is known that RNA released into the circulation is surprisingly stable in spite of the fact that increased amounts of RNases circulate in the blood of cancer patients. This implies that RNA may be protected from degradation by its packaging into exosomes[86], such as microparticles, microvesicles or multivesicles, which are shed from cellular surfaces into the bloodstream[87]. The detection and identification of RNA can be carried out using microarray technologies or reverse transcription quantitative real-time PCR[88].

Serum thyroglobulin levels are a specific and sensitive tumour marker for the detection of persistent or recurrent thyroid cancer. Levels of thyroglobulin change during thyroid hormone-suppressive therapy, as well as after stimulation with thyroid-stimulating hormone, and the levels correlate well with disease progression. The measurement of mRNA levels of thyroid-specific transcripts might be useful in the early detection of tumour relapse[89]. However, another study has shown that the detection of circulating thyroglobulin mRNA one year after thyroidectomy might be of no use in the prediction of early and midterm local and distant recurrences of this disease[90].

In patients with breast cancer, levels of *CCND1* mRNA (encoding cyclin D1) identified patients with poor overall survival in good-prognosis groups and patients who were non-responsive to tamoxifen[91]. Nasopharyngeal carcinoma has been associated with disturbances in the integrity of cell-free circulating RNA, suggesting that the measurement of plasma RNA integrity may be a useful biomarker for the diagnosis and monitoring of malignant diseases[92]. Several groups have tried to detect human telomerase reverse transcriptase (*TERT*) mRNA in plasma, and have not found any association between the presence of this mRNA and clinicopathological parameters[7].

© 2011 Macmillan Publishers Limited. All rights reserved

By contrast, Miura *et al.*[93] measured *TERT* mRNA together with *EGFR* mRNA levels in serum from patients with lung cancer and showed that *TERT* concentration correlated with tumour size, the presence of metastasis, disease recurrence and smoking. An increase in the concentration of *EGFR* mRNA correlated with advanced clinical stages, and decreased levels of *EGFR* and *TERT* were evident after surgery[93]. These findings show that although the use of mRNA has to be further assessed in large clinical trials, these data seem promising.

miRNA content. Currently, expression microarrays that cover >900 mature human miRNA sequences listed in the miRNA database (miRBase) allow the screening of deregulated transcript levels of miRNAs in different tumour tissues (BOX 2). Subsequently, the aberrant expression levels of miRNAs deduced from the array data can be examined by quantitative real-time PCR in single miRNA assays. The application of these techniques has shown conflicting quantitative data on the upregulation of circulating miRNAs from the same tumour type in different studies[94].

These discrepancies might mainly be due to the lack of an established endogenous miRNA control to normalize miRNA amounts. Indeed, a recent study[95] indicated the need for such a control. *mir-16* or the small nucleolar RNA *RNU62* are frequently used as reference genes[96], but others have argued that all tested miRNAs should have established mean expression levels to reduce the technical variation in the miRNA isolation and to more accurately assess the biological changes[97]. However, this approach is only applicable if the control miRNAs are well studied in relevant defined populations. The expression profile of blood miRNAs may change with respect to the established risk factors of the cancer patients and whether the blood samples were drawn before or after treatment, surgery or chemotherapy[94]. Therefore, for each study, the candidate reference miRNAs should be rigorously validated, as even frequently used reference miRNAs are variable under different physiological conditions or patient and donor demographics. This area is in need of universal standards to allow better comparisons and validations of specific blood miRNAs. This will continue to be a problem if the extraction of miRNAs from blood is variable from one sample to another. However, it has recently been shown that a direct miRNA assay of serum without extraction may simplify the problem and improve overall assay comparison[98]. The crucial problem is the extraction of small amounts of miRNA from plasma or serum, which is highly variable among different published papers. Because of the small size of the miRNAs and their attachment to lipids and proteins, efficient and reproducible extraction remains an inherent problem.

Nevertheless, on the basis of their biological role and involvement in transforming cells, circulating miRNAs may have potential as diagnostic, prognostic and predictive biomarkers and may also be considered as potential future therapeutic targets[94]. Although the analysis of circulating miRNAs has just begun, there are indications that circulating miRNAs may become promising biomarkers, particularly because of the strong link between their deregulation and cancer development and progression.

In 2008, the presence of miRNAs in serum was first described for patients with diffuse large B cell lymphoma[99]. miRNA expression profiles have since been shown to have signatures that are related to tumour classification, diagnosis and disease progression[95,100–102] — patients with breast cancer with advanced stage disease had significantly more miR-34a in their blood than patients at early tumour stages, and changes in miR-10b, miR-34a and miR-155 serum levels correlated with the presence of metastases[101]. Recently, an assay for circulating miR-21 was shown to be useful in the detection of early stage breast cancer progression[103], and in non-small-cell lung cancer (NSCLC) serum miRNA levels were found to be altered more than fivefold between patients with longer and shorter survival[104].

The developmental lineage and differentiation state of various tumour types might be reflected by the miRNA signature[105]. For example, detection of miR-92 in plasma could differentiate patients with colorectal cancer from patients with gastric cancer[102]. During liver development the expression of particular miRNAs has been reported to change dynamically, and one of these miRNAs, miR-500, is an oncofetal miRNA that is relevant to the diagnosis of human hepatocellular carcinoma[106].

In summary, the findings discussed above highlight the potential clinical utility of circulating miRNA profiling in cancer diagnosis. Considering the clinical relevance of miRNAs in cancer tissues, this field will inevitably grow.

cfDNA and micrometastatic cells

Besides numerous studies on the clinical utility of circulating tumour cells (CTCs) in blood, disseminated tumour cells (DTCs) in the bone marrow and cfNA in the blood of cancer patients[107–109], the investigation into combined analyses of circulating nucleic acids with CTCs or DTCs has just begun. The observed correlations of cfDNA with CTCs and DTCs suggest that cfDNA may be derived not only from the primary tumour but also from micrometastatic cells[13,14,110–114].

In a study of primary head and neck squamous cell carcinoma, microsatellite alterations in serum DNA have been reported to predict distant metastasis. In this report it was also shown that CTCs may contribute to the presence of cfDNA that is detected by microsatellite analysis[110]. In prostate cancer, the presence of CTCs significantly correlated with an increase in the detection of LOH of dematin, *CDKN2A* and *BRCA1* in cfDNA[13].

Box 2 | **Characteristics of miRNAs**

MicroRNAS (miRNAs) are a class of naturally occurring small non-coding RNA molecules. Mature miRNAs consist of 19 to 25 nucleotides and are derived from hairpin precursor molecules of 70–100 nucleotides. As 50% of human miRNAs are localized in fragile chromosomal regions, which may exhibit DNA amplifications, deletions or translocations during tumour development, their expression is frequently deregulated in cancer[121]. Therefore, miRNAs have important roles in the regulation of gene expression in cancer[122]. To date, studies on solid cancers (ovarian, lung, breast and colorectal cancer, for example) report that miRNAs are involved in the regulation of different cellular processes, such as apoptosis, cell proliferation, epithelial to mesenchymal transition and metastasis. In blood, miRNAs seem to be highly stable, because most of them are included in apoptotic bodies, microvesicles or exosomes and can withstand known mRNA degradation factors[94,103].

© 2011 Macmillan Publishers Limited. All rights reserved

The biological relevance of LOH in these regions might, therefore, contribute to a better understanding of the early steps of the metastatic cascade in this carcinoma type. Interestingly, in patients with prostate cancer who had tumour cells in their bone marrow, the frequency of LOH that was detected using cfDNA from the bone marrow plasma increased compared with patients who did not have DTCs in their bone marrow. These data suggest that tumour-specific cfDNA is also present in the bone marrow and indicate a possible relationship with bone marrow micrometastasis[14]. In breast cancer, there was no significant correlation between the presence of DTCs in the bone marrow and LOH of *CDKN2A*, *PTEN*, *BRCA1*, *BRCA2* and E-cadherin (*CDH1*) in cfDNA from blood samples[33]. Presumably, this lack of concordance is caused by the restricted set of microsatellite markers used. However, patients with DTC-positive bone marrow had higher DNA yields in their blood than patients with DTC-negative bone marrow[33], indicating that tumour cfDNA may at least partly stem from DTCs.

An association between CTCs and serum tumour-related methylated DNA has also been observed. Patients with melanoma who had CTCs and methylated *RASSF1A* and *RARB* in their blood showed a significantly poorer response to chemotherapy and a shorter time to progression, as well as poorer overall survival[111]. These findings indicate that a combined assessment of methylated cfDNA and CTCs in blood may be a useful determinant of disease status and the efficacy of systemic therapy in patients with melanoma. In patients with breast cancer, the detection of large amounts of methylated cfDNA correlated with the presence of CTCs in the blood[114]. Based on an association of cell-free, methylated *APC*, *RASSF1A* and *ESR1* molecules with CTCs, it has been suggested that CTCs are a potential source of circulating tumour-specific DNA, and that high numbers of CTCs and methylated cfDNA molecules are both a phenotypic feature of more aggressive breast tumour biology[114]. In this regard, the association of cell-free, methylated *APC* and *GSTP1* molecules with CTCs in the blood of patients with breast cancer correlated with a more aggressive tumour phenotype and an advanced disease stage[112].

Although the findings discussed above are still preliminary, they emphasize that cfDNA may also stem from CTCs that have undergone cell death when in the circulatory system.

Conclusion and perspectives

Carcinogenesis and tumour progression are complex and progressive processes that are associated with numerous genetic and epigenetic alterations, some of which can also be detected as cfNA in plasma and serum. Although there are cancer protein blood biomarkers that have been approved by the American Society of Clinical Oncology, their number and clinical use are limited. More studies are needed in large cohorts of cancer patients with well-defined clinical staging and outcomes. The cfNAs might be excellent blood cancer biomarkers, as they may be more informative, specific and accurate than protein biomarkers. Currently, efficient management of cancer patients relies on early diagnosis, precise tumour staging and monitoring of treatment. Histological evaluation of tumour tissues obtained from biopsies, as well as blood samples, are the 'gold standard' of diagnosis, but most studies usually carry out these evaluations once only. We now know that metastatic and primary tumours from the same patient can vary at the genomic, epigenomic and transcriptomic levels, thus assays that allow the repetitive monitoring of these events using blood samples would be more efficient in assessing cancer progression in patients from whom tumour tissue is not available[11,111,115,116]. Minimally invasive blood analyses of cfNA may have the potential to complement or replace the existing cancer tissue and blood biomarkers in the future.

One crucial factor in the continued development of cfNA biomarkers is addressing technical issues such as cfNA extraction (as described above) and rigorously following the guidelines of the NACB USA. This will be a major task that will require cooperation among the leading groups in the world in this field to obtain a consensus on assays and reporting results. Our recommendation is to develop a task force with expertise in cancer cfNA. As many of the new approved targeted therapies are focused on DNA aberrations, such as mutations in the *KRAS*, *BRAF* and *EGFR* genes, the investment by pharmaceutical and biotechnology companies into specific cfNA assays is likely to be highly important, particularly in monitoring drug responses. As the individual genomic profiles of a patient's tumour become more readily available, the use of cfNA assays can be better exploited for personalized medicine and for monitoring treatment efficacy.

1. Mandel, P. & Métais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C. R. Acad. Sci. Paris* **142**, 241–243 (1948).
2. Sorenson, G. D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol. Biomarkers Prev.* **3**, 67–71 (1994).
3. Vasioukhin, V. *et al.* Point mutations of the N-ras gene in the blood plasma DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.* **86**, 774–779 (1994).
4. Nawroz, H., Koch, W., Anker, P., Stroun, M. & Sidransky, D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nature Med.* **2**, 1035–1037 (1996).
5. Kaiser, J. Medicine. Keeping tabs on tumor DNA. *Science* **327**, 1074 (2010).
6. Swaminathan, R. & Butt, A. N. Circulating nucleic acids in plasma and serum: recent developments. *Ann. N. Y Acad. Sci.* **1075**, 1–9 (2006).
This review discusses the origin and biological importance of circulating nucleic acids in fetal medicine, oncology and other human-related diseases.
7. Fleischhacker, M. & Schmidt, B. Circulating nucleic acids (CNAs) and cancer-a survey. *Biochim. Biophys. Acta* **1775**, 181–232 (2007).
8. Choi, J. J., Reich, C. F., 3rd & Pisetsky, D. S. The role of macrophages in the *in vitro* generation of extracellular DNA from apoptotic and necrotic cells. *Immunology* **115**, 55–62 (2005).
9. Stroun, M. *et al.* The origin and mechanism of circulating DNA. *Ann. N. Y Acad. Sci.* **906**, 161–168 (2000).
10. Gahan, P. B. & Swaminathan, R. Circulating nucleic acids in plasma and serum. Recent developments. *Ann. N. Y Acad. Sci.* **1137**, 1–6 (2008).
11. Diehl, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc. Natl Acad. Sci. USA* **102**, 16368–16373 (2005).
12. Jahr, S. *et al.* DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.* **61**, 1659–1665 (2001).
13. Schwarzenbach, H. *et al.* Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. *Clin. Cancer Res.* **15**, 1032–1038 (2009).
A relationship between the occurrence of CTCs and circulating tumour-associated DNA in blood is described for the first time in patients with prostate cancer.

© 2011 Macmillan Publishers Limited. All rights reserved

14. Schwarzenbach, H. *et al.* Detection of tumor-specific DNA in blood and bone marrow plasma from patients with prostate cancer. *Int. J. Cancer* **120**, 1465–1471 (2007).
15. Bendich, A., Wilczok, T. & Borenfreund, E. Circulating DNA as a possible factor in oncogenesis. *Science* **148**, 374–376 (1965).
16. Emlen, W. & Mannik, M. Effect of DNA size and strandedness on the *in vivo* clearance and organ localization of DNA. *Clin. Exp. Immunol.* **56**, 185–192 (1984).
17. Wimberger, P. *et al.* Impact of platinum-based chemotherapy on circulating nucleic acid levels, protease activities in blood and disseminated tumor cells in bone marrow of ovarian cancer patients. *Int. J. Cancer* **128**, 2572–2580 (2010).
This is the first study indicating the potential value of circulating nucleosomes in monitoring the effects of chemotherapy in ovarian cancer.
18. Boddy, J. L., Gal, S., Malone, P. R., Harris, A. L. & Wainscoat, J. S. Prospective study of quantitation of plasma DNA levels in the diagnosis of malignant versus benign prostate disease. *Clin. Cancer Res.* **11**, 1394–1399 (2005).
19. Kamat, A. A. *et al.* Plasma cell-free DNA in ovarian cancer: an independent prognostic biomarker. *Cancer* **116**, 1918–1925 (2010).
20. Allen, D. *et al.* Role of cell-free plasma DNA as a diagnostic marker for prostate cancer. *Ann. N. Y Acad. Sci.* **1022**, 76–80 (2004).
21. Schwarzenbach, H., Stoehlmacher, J., Pantel, K. & Goekkurt, E. Detection and monitoring of cell-free DNA in blood of patients with colorectal cancer. *Ann. N. Y Acad. Sci.* **1137**, 190–196 (2008).
22. Chun, F. K. *et al.* Circulating tumour-associated plasma DNA represents an independent and informative predictor of prostate cancer. *BJU Int.* **98**, 544–548 (2006).
23. Sunami, E., Vu, A. T., Nguyen, S. L., Giuliano, A. E. & Hoon, D. S. Quantification of LINE1 in circulating DNA as a molecular biomarker of breast cancer. *Ann. N. Y Acad. Sci.* **1137**, 171–174 (2008).
24. Gormally, E. *et al.* Amount of DNA in plasma and cancer risk: a prospective study. *Int. J. Cancer* **111**, 746–749 (2004).
25. Catarino, R. *et al.* Quantification of free circulating tumor DNA as a diagnostic marker for breast cancer. *DNA Cell Biol.* **27**, 415–421 (2008).
26. Mehra, N. *et al.* Circulating mitochondrial nucleic acids have prognostic value for survival in patients with advanced prostate cancer. *Clin. Cancer Res.* **13**, 421–426 (2007).
27. Ellinger, J., Albers, P., Muller, S. C., von Ruecker, A. & Bastian, P. J. Circulating mitochondrial DNA in the serum of patients with testicular germ cell cancer as a novel noninvasive diagnostic biomarker. *BJU Int.* **104**, 48–52 (2009).
28. Chiu, R. W. *et al.* Quantitative analysis of circulating mitochondrial DNA in plasma. *Clin. Chem.* **49**, 719–726 (2003).
This is one of the first studies describing the technical approach and detection of mitochondria circulating DNA in plasma.
29. Tangkijvanich, P. *et al.* Serum LINE-1 hypomethylation as a potential prognostic marker for hepatocellular carcinoma. *Clin. Chim. Acta* **379**, 127–133 (2007).
30. Umetani, N. *et al.* Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J. Clin. Oncol.* **24**, 4270–4276 (2006).
First major study demonstrating a direct PCR assay for detecting *ALU* cfNA in patients with breast cancer. The study demonstrates that an *ALU* DNA integrity assay can be sensitive to detect early stage metastasis to regional tumour-draining lymph nodes.
31. Umetani, N. *et al.* Increased integrity of free circulating DNA in sera of patients with colorectal or periampullary cancer: direct quantitative PCR for ALU repeats. *Clin. Chem.* **52**, 1062–1069 (2006).
32. Schulz, W. A., Steinhoff, C. & Florl, A. R. Methylation of endogenous human retroelements in health and disease. *Curr. Top. Microbiol. Immunol.* **310**, 211–250 (2006).
33. Schwarzenbach, H. *et al.* Comparative evaluation of cell-free tumor DNA in blood and disseminated tumor cells in bone marrow of patients with primary breast cancer. *Breast Cancer Res.* **11**, R71 (2009).
First study indicating that tumour cfDNA may stem at least partly from DTCs in bone marrow.
34. Silva, J. M. *et al.* Tumor DNA in plasma at diagnosis of breast cancer patients is a valuable predictor of disease-free survival. *Clin. Cancer Res.* **8**, 3761–3766 (2002).
35. Taback, B. *et al.* Detection of tumor-specific genetic alterations in bone marrow from early-stage breast cancer patients. *Cancer Res.* **63**, 1884–1887 (2003).
36. Bruhn, N. *et al.* Detection of microsatellite alterations in the DNA isolated from tumor cells and from plasma DNA of patients with lung cancer. *Ann. N. Y Acad. Sci.* **906**, 72–82 (2000).
37. Sozzi, G. *et al.* Analysis of circulating tumor DNA in plasma at diagnosis and during follow-up of lung cancer patients. *Cancer Res.* **61**, 4675–4678 (2001).
38. Sunami, E. *et al.* Multimarker circulating DNA assay for assessing blood of prostate cancer patients. *Clin. Chem.* **55**, 559–567 (2009).
39. Coulet, F. *et al.* Detection of plasma tumor DNA in head and neck squamous cell carcinoma by microsatellite typing and p53 mutation analysis. *Cancer Res.* **60**, 707–711 (2000).
40. Hibi, K. *et al.* Molecular detection of genetic alterations in the serum of colorectal cancer patients. *Cancer Res.* **58**, 1405–1407 (1998).
41. Kopreski, M. S. *et al.* Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. *Br. J. Cancer* **76**, 1293–1299 (1997).
42. Schulte-Hermann, R. *et al.* Role of active cell death (apoptosis) in multi-stage carcinogenesis. *Toxicol. Lett.* **82–83**, 143–148 (1995).
43. De Roock, W., Biesmans, B., De Schutter, J. & Tejpar, S. Clinical biomarkers in oncology: focus on colorectal cancer. *Mol. Diagn. Ther.* **13**, 103–114 (2009).
44. Downward, J. Targeting RAS signalling pathways in cancer therapy. *Nature Rev. Cancer* **3**, 11–22 (2003).
45. Levine, A. J. & Oren, M. The first 30 years of p53: growing ever more complex. *Nature Rev. Cancer* **9**, 749–758 (2009).
46. Castells, A. *et al.* K-ras mutations in DNA extracted from the plasma of patients with pancreatic carcinoma: diagnostic utility and prognostic significance. *J. Clin. Oncol.* **17**, 578–584 (1999).
47. Ryan, B. M. *et al.* A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up. *Gut* **52**, 101–108 (2003).
48. Wang, S. *et al.* Potential clinical significance of a plasma-based KRAS mutation analysis in patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* **16**, 1324–1330 (2010).
49. Shinozaki, M. *et al.* Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. *Clin. Cancer Res.* **13**, 2068–2074 (2007).
This is the first major study to demonstrate circulating *BRAF* DNA mutation in patients with different stages of melanoma, and that cfDNA mutation detection has clinical utility for monitoring patient responses before and after therapy.
50. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363**, 809–819 (2010).
51. Kefford, R. *et al.* Phase I/II study of GSK2118436, a selective inhibitor of oncogenic mutant BRAF kinase, in patients with metastatic melanoma and other solid tumors. *J. Clin. Oncol.* **28**, 8503 (2010).
52. Ciardiello, F. & Tortora, G. EGFR antagonists in cancer treatment. *N. Engl. J. Med.* **358**, 1160–1174 (2008).
53. Kimura, H. *et al.* EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. *Br. J. Cancer* **95**, 1390–1395 (2006).
54. Kobayashi, S. *et al.* EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **352**, 786–792 (2005).
55. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18**, 1875–1883 (2008).
56. Wolff, E. M. *et al.* Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet.* **6**, e1000917 (2010).
57. Chan, K. C., Leung, S. F., Yeung, S. W., Chan, A. T. & Lo, Y. M. Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **14**, 4141–4145 (2008).
58. Ellinger, J. *et al.* Cell-free circulating DNA: diagnostic value in patients with testicular germ cell cancer. *J. Urol.* **181**, 363–371 (2009).
59. Salani, R. *et al.* Measurement of cyclin E genomic copy number and strand length in cell-free DNA distinguish malignant versus benign effusions. *Clin. Cancer Res.* **13**, 5805–5809 (2007).
60. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).
61. Kristensen, L. S. & Hansen, L. L. PCR-based methods for detecting single-locus DNA methylation biomarkers in cancer diagnostics, prognostics, and response to treatment. *Clin. Chem.* **55**, 1471–1483 (2009).
62. Ellinger, J. *et al.* CpG island hypermethylation in cell-free serum DNA identifies patients with localized prostate cancer. *Prostate* **68**, 42–49 (2008).
63. Taback, B., Saha, S. & Hoon, D. S. Comparative analysis of mesenteric and peripheral blood circulating tumor DNA in colorectal cancer patients. *Ann. N. Y Acad. Sci.* **1075**, 197–203 (2006).
64. Lane, A. A. & Chabner, B. A. Histone deacetylase inhibitors in cancer therapy. *J. Clin. Oncol.* **27**, 5459–5468 (2009).
65. Cameron, E. E., Bachman, K. E., Myohanen, S., Herman, J. G. & Baylin, S. B. Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. *Nature Genet.* **21**, 103–107 (1999).
66. Laktionov, P. P. *et al.* Cell-surface-bound nucleic acids: free and cell-surface-bound nucleic acids in blood of healthy donors and breast cancer patients. *Ann. N. Y Acad. Sci.* **1022**, 221–227 (2004).
67. Stollar, B. D. & Stephenson, F. Apoptosis and nucleosomes. *Lupus* **11**, 787–789 (2002).
68. Ward, T. H. *et al.* Biomarkers of apoptosis. *Br. J. Cancer* **99**, 841–846 (2008).
69. Holdenrieder, S. *et al.* Nucleosomes in serum of patients with benign and malignant diseases. *Int. J. Cancer* **95**, 114–120 (2001).
70. Roth, C. *et al.* Apoptosis-related deregulation of proteolytic activities and high serum levels of circulating nucleosomes and DNA in blood correlate with breast cancer progression. *BMC Cancer* **11**, 4 (2010).
71. Holdenrieder, S. *et al.* Clinical relevance of circulating nucleosomes in cancer. *Ann. N. Y Acad. Sci.* **1137**, 180–189 (2008).
72. Lo, Y. M. *et al.* Molecular prognostication of nasopharyngeal carcinoma by quantitative analysis of circulating Epstein-Barr virus DNA. *Cancer Res.* **60**, 6878–6881 (2000).
73. Kim, B. K. *et al.* Persistent hepatitis B viral replication affects recurrence of hepatocellular carcinoma after curative resection. *Liver Int.* **28**, 393–401 (2008).
74. Illades-Aguiar, B. *et al.* Prevalence and distribution of human papillomavirus types in cervical cancer, squamous intraepithelial lesions, and with no intraepithelial lesions in women from Southern Mexico. *Gynecol. Oncol.* **117**, 291–296 (2010).
75. Yu, K. H. *et al.* Quantitative analysis of cell-free Epstein-Barr virus DNA in plasma of patients with nonnasopharyngeal head and neck carcinomas. *Clin. Cancer Res.* **10**, 1726–1732 (2004).
76. Chan, A. T. *et al.* Plasma Epstein-Barr virus DNA and residual disease after radiotherapy for undifferentiated nasopharyngeal carcinoma. *J. Natl Cancer Inst.* **94**, 1614–1619 (2002).
77. Leung, S. F. *et al.* Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. *J. Clin. Oncol.* **24**, 5414–5418 (2006).
This study describes the use of plasma *EBV* in nasopharyngeal carcinoma (NPC) prognostication and monitoring during therapy. Pretherapy circulating *EBV* DNA level was shown to be an independent prognostic factor in NPC.
78. Lin, J. C. *et al.* Quantification of plasma Epstein-Barr virus DNA in patients with advanced nasopharyngeal carcinoma. *N. Engl. J. Med.* **350**, 2461–2470 (2004).
79. Lo, Y. M. *et al.* Quantitative analysis of cell-free Epstein-Barr virus DNA in plasma of patients with nasopharyngeal carcinoma. *Cancer Res.* **59**, 1188–1191 (1999).
This study describes detection of circulating *EBV* DNA in patients with NPC and provides evidence that this approach can be used for the monitoring and early detection of NPC.
80. Lo, Y. M. *et al.* Plasma cell-free Epstein-Barr virus DNA quantitation in patients with nasopharyngeal carcinoma. Correlation with clinical staging. *Ann. N. Y Acad. Sci.* **906**, 99–101 (2000).
81. Ji, M. F. *et al.* Detection of Stage I nasopharyngeal carcinoma by serologic screening and clinical examination. *Chin. J. Cancer* **30**, 120–123 (2011).

© 2011 Macmillan Publishers Limited. All rights reserved

82. Garcia-Olmo, D. C., Ruiz-Piqueras, R. & Garcia-Olmo, D. Circulating nucleic acids in plasma and serum (CNAPS) and its relation to stem cells and cancer metastasis: state of the issue. *Histol. Histopathol.* **19**, 575–583 (2004).
83. Garcia-Olmo, D. C. *et al.* Cell-free nucleic acids circulating in the plasma of colorectal cancer patients induce the oncogenic transformation of susceptible cultured cells. *Cancer Res.* **70**, 560–567 (2010).
84. Sturgeon, C. M. & Diamandis, E. P. Use of tumor markers in clinical practice: quality requirements. *Clin. Physiol. Biochem.* 1–37 (2008).
85. Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622 (2009).
86. Orozco, A. F. & Lewis, D. E. Flow cytometric analysis of circulating microparticles in plasma. *Cytometry A* **77**, 502–514 (2010).
87. Cocucci, E., Racchetti, G. & Meldolesi, J. Shedding microvesicles: artefacts no more. *Trends Cell Biol.* **19**, 43–51 (2009).
88. O'Driscoll, L. *et al.* Feasibility and relevance of global expression profiling of gene transcripts in serum from breast cancer patients using whole genome microarrays and quantitative RT-PCR. *Cancer Genomics Proteomics* **5**, 94–104 (2008).
89. Barzon, L., Boscaro, M., Pacenti, M., Taccaliti, A. & Palu, G. Evaluation of circulating thyroid-specific transcripts as markers of thyroid cancer relapse. *Int. J. Cancer* **110**, 914–920 (2004).
90. Lombardi, C. P. *et al.* Circulating thyroglobulin mRNA does not predict early and midterm recurrences in patients undergoing thyroidectomy for cancer. *Am. J. Surg.* **196**, 326–332 (2008).
91. Garcia, V. *et al.* Free circulating mRNA in plasma from breast cancer patients and clinical outcome. *Cancer Lett.* **263**, 312–320 (2008).
92. Wong, B. C. *et al.* Reduced plasma RNA integrity in nasopharyngeal carcinoma patients. *Clin. Cancer Res.* **12**, 2512–2516 (2006).
93. Miura, N. *et al.* Clinical usefulness of serum telomerase reverse transcriptase (hTERT) mRNA and epidermal growth factor receptor (EGFR) mRNA as a novel tumor marker for lung cancer. *Cancer Sci.* **97**, 1366–1373 (2006).
94. Kosaka, N., Iguchi, H. & Ochiya, T. Circulating microRNA in body fluid: a new potential biomarker for cancer diagnosis and prognosis. *Cancer Sci.* **101**, 2087–2092 (2010).
95. Mitchell, P. S. *et al.* Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl Acad. Sci. USA* **105**, 10513–10518 (2008).
This is the first major study describing the detection of circulating miRNAs in both plasma and serum, and their use in assessing patients with prostate cancer.
96. Schaefer, A. *et al.* Suitable reference genes for relative quantification of miRNA expression in prostate cancer. *Exp. Mol. Med.* **42**, 749–758 (2010).
97. Mestdagh, P. *et al.* A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.* **10**, R64 (2009).
98. Heneghan, H. M., Miller, N., Lowery, A. J., Sweeney, K. J. & Kerin, M. J. MicroRNAs as novel biomarkers for breast cancer. *J. Oncol.* **2009**, 950201 (2009).
99. Lawrie, C. H. *et al.* Detection of elevated levels of tumour-associated microRNAs in serum of patients with diffuse large B-cell lymphoma. *Br. J. Haematol.* **141**, 672–675 (2008).
The presence of miRNAs in serum was first described for cancer patients in this paper.
100. Roth, C., Kasimir-Bauer, S., Heubner, M., Pantel, K. & Schwarzenbach, H. *Circulating Nucleic Acids in Plasma and Serum*. 63–71 (Springer, 2011).
101. Roth, C. *et al.* Circulating microRNAs as blood-based markers for patients with primary and metastatic breast cancer. *Breast Cancer Res.* **12**, R90 (2010).
102. Ng, E. K. *et al.* Differential expression of microRNAs in plasma of patients with colorectal cancer: a potential marker for colorectal cancer screening. *Gut* **58**, 1375–1381 (2009).
103. Asaga, S. *et al.* Direct serum assay for microRNA-21 concentrations in early and advanced breast cancer. *Clin. Chem.* **57**, 84–91 (2011).
This study demonstrates the development of a direct blood PCR assay for detection of circulating miR-21 and its ability to assess early stage breast cancer in serum.
104. Hu, Z. *et al.* Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. *J. Clin. Oncol.* **28**, 1721–1726 (2010).
This article describes the first major study on the use of a panel of circulating miRNAs in serum to predict overall survival outcome in NSCLC.
105. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
106. Yamamoto, Y. *et al.* MicroRNA-500 as a potential diagnostic marker for hepatocellular carcinoma. *Biomarkers* **14**, 529–538 (2009).
107. Pantel, K. & Alix-Panabieres, C. Circulating tumour cells in cancer patients: challenges and perspectives. *Trends Mol. Med.* **16**, 398–406 (2010).
108. Pantel, K., Brakenhoff, R. H. & Brandt, B. Detection, clinical relevance and specific biological properties of disseminating tumour cells. *Nature Rev. Cancer* **8**, 329–340 (2008).
109. Schwarzenbach, H. & Pantel, K. *Methods of Cancer Diagnosis, Therapy and Prognosis*. 481–497 (Springer, USA, 2008).
110. Nawroz-Danish, H. *et al.* Microsatellite analysis of serum DNA in patients with head and neck cancer. *Int. J. Cancer* **111**, 96–100 (2004).
111. Koyanagi, K. *et al.* Association of circulating tumor cells with serum tumor-related methylated DNA in peripheral blood of melanoma patients. *Cancer Res.* **66**, 6111–6117 (2006).
This study indicates that a combined assessment of methylated cfDNA and CTCs in blood may be a useful determinant of disease status and efficacy of systemic therapy of metastatic melanoma.
112. Matuschek, C. *et al.* Methylated APC and GSTP1 genes in serum DNA correlate with the presence of circulating blood tumor cells and are associated with a more aggressive and advanced breast cancer disease. *Eur. J. Med. Res.* **15**, 277–286 (2010).
113. Pinzani, P. *et al.* Tyrosinase mRNA levels in the blood of uveal melanoma patients: correlation with the number of circulating tumor cells and tumor progression. *Melanoma Res.* **20**, 303–310 (2010).
114. Van der Auwera, I. *et al.* The presence of circulating total DNA and methylated genes is associated with circulating tumour cells in blood from breast cancer patients. *Br. J. Cancer* **100**, 1277–1286 (2009).
115. Mori, T. *et al.* Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. *J. Clin. Oncol.* **23**, 9351–9358 (2005).
This was the first study to demonstrate that hypermethylation of cfDNA in patients with melanoma relates to tumour progression and response to therapy, and that a panel of methylated tumour-related cfDNA can be of prognostic value before therapy.
116. Diehl, F. *et al.* Circulating mutant DNA to assess tumor dynamics. *Nature Med.* **14**, 985–990 (2008).
This article describes a new highly sensitive approach to quantify cfDNA which was applied to monitor chemotherapy.
117. Esteller, M. & Herman, J. G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J. Pathol.* **196**, 1–7 (2002).
118. Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277 (2003).
119. Zheng, Y. G., Wu, J., Chen, Z. & Goodman, M. Chemical regulation of epigenetic modifications: opportunities for new cancer therapy. *Med. Res. Rev.* **28**, 645–687 (2008).
120. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10**, 295–304 (2009).
121. Croce, C. M. Causes and consequences of microRNA dysregulation in cancer. *Nature Rev. Genet.* **10**, 704–714 (2009).
122. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
123. Ellinger, J. *et al.* Apoptotic DNA fragments in serum of patients with muscle invasive bladder cancer: a prognostic entity. *Cancer Lett.* **264**, 274–280 (2008).
124. Valenzuela, M. T. *et al.* Assessing the use of p16^{INK4a} promoter gene methylation in serum for detection of bladder cancer. *Eur. Urol.* **42**, 622–628; discussion 628–30 (2002).
125. Utting, M., Werner, W., Dahse, R., Schubert, J. & Junker, K. Microsatellite analysis of free tumor DNA in urine, serum, and plasma of patients: a minimally invasive method for the detection of bladder cancer. *Clin. Cancer Res.* **8**, 35–40 (2002).
126. Fiegl, H. *et al.* Circulating tumor-specific DNA: a marker for monitoring efficacy of adjuvant therapy in cancer patients. *Cancer Res.* **65**, 1141–1145 (2005).
127. Rykova, E. Y. *et al.* Extracellular DNA in breast cancer: cell-surface-bound, tumor-derived extracellular DNA in blood of patients with breast cancer and nonmalignant tumors. *Ann. N. Y Acad. Sci.* **1022**, 217–220 (2004).
128. Sharma, G. *et al.* CpG hypomethylation of MDR1 gene in tumor and serum of invasive ductal breast carcinoma patients. *Clin. Biochem.* **43**, 373–379 (2010).
129. Sharma, G. *et al.* Clinical significance of promoter hypermethylation of DNA repair genes in tumor and serum DNA in invasive ductal breast carcinoma patients. *Life Sci.* **87**, 83–91 (2010).
130. Skvortsova, T. E. *et al.* Cell-free and cell-bound circulating DNA in breast tumours: DNA quantification and analysis of tumour-related gene methylation. *Br. J. Cancer* **94**, 1492–1495 (2006).
131. Deligezer, U. *et al.* Effect of adjuvant chemotherapy on integrity of free serum DNA in patients with breast cancer. *Ann. N. Y Acad. Sci.* **1137**, 175–179 (2008).
132. Kohler, C. *et al.* Levels of plasma circulating cell free nuclear and mitochondrial DNA as potential biomarkers for breast tumors. *Mol. Cancer* **8**, 105 (2009).
133. Ren, C. C. *et al.* Methylation status of the fragile histidine triad and E-cadherin genes in plasma of cervical cancer patients. *Int. J. Gynecol. Cancer* **16**, 1862–1867 (2006).
134. Widschwendter, A. *et al.* DNA methylation in serum and tumors of cervical cancer patients. *Clin. Cancer Res.* **10**, 565–571 (2004).
135. Pornthanakasem, W. *et al.* Human papillomavirus DNA in plasma of patients with cervical cancer. *BMC Cancer* **1**, 2 (2001).
136. Lecomte, T. *et al.* Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. *Int. J. Cancer* **100**, 542–548 (2002).
137. Lefebure, B. *et al.* Prognostic value of circulating mutant DNA in unresectable metastatic colorectal cancer. *Ann. Surg.* **251**, 275–280 (2010).
138. Trevisiol, C. *et al.* Prognostic value of circulating KRAS2 gene mutations in colorectal cancer with distant metastases. *Int. J. Biol. Markers* **21**, 223–228 (2006).
139. Wang, J. Y. *et al.* Molecular detection of APC, K- ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. *World J. Surg.* **28**, 721–726 (2004).
140. deVos, T. *et al.* Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. *Clin. Chem.* **55**, 1337–1346 (2009).
141. Grutzmann, R. *et al.* Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS ONE* **3**, e3759 (2008).
This was the first study that demonstarated the clinical use of methylated cfDNA as a specific plasma biomarker for screening colorectal cancer.
142. He, Q. *et al.* Development of a multiplex MethyLight assay for the detection of multigene methylation in human colorectal cancer. *Cancer Genet. Cytogenet.* **202**, 1–10 (2010).
143. Tanzer, M. *et al.* Performance of epigenetic markers SEPT9 and ALX4 in plasma for detection of colorectal precancerous lesions. *PLoS ONE* **5**, e9061 (2010).
144. Chan, K. C. *et al.* Quantitative analysis of circulating methylated DNA as a biomarker for hepatocellular carcinoma. *Clin. Chem.* **54**, 1528–1536 (2008).
145. Wang, J., Qin, Y., Li, B., Sun, Z. & Yang, B. Detection of aberrant promoter methylation of GSTP1 in the tumor and serum of Chinese human primary hepatocellular carcinoma patients. *Clin. Biochem.* **39**, 344–348 (2006).
146. Wong, I. H., Lo, Y. M., Yeo, W., Lau, W. Y. & Johnson, P. J. Frequent p15 promoter methylation in tumor and peripheral blood from hepatocellular carcinoma patients. *Clin. Cancer Res.* **6**, 3516–3521 (2000).
147. Ren, N. *et al.* The prognostic value of circulating plasma DNA level and its allelic imbalance on chromosome 8p in patients with hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.* **132**, 399–407 (2006).
148. Szymanska, K. *et al.* Ser-249TP53 mutation in tumour and plasma DNA of hepatocellular carcinoma patients from a high incidence area in the Gambia, West Africa. *Int. J. Cancer* **110**, 374–379 (2004).
149. Kirk, G. D. *et al.* 249^{ser} TP53 mutation in plasma DNA, hepatitis B viral infection, and risk of hepatocellular carcinoma. *Oncogene* **24**, 5858–5867 (2005).

© 2011 Macmillan Publishers Limited. All rights reserved

150. Su, Y. W., Huang, Y. W., Chen, S. H. & Tzen, C. Y. Quantitative analysis of plasma HBV DNA for early evaluation of the response to transcatheter arterial embolization for HBV-related hepatocellular carcinoma. *World J. Gastroenterol.* **11**, 6193–6196 (2005).
151. Gautschi, O. *et al.* Origin and prognostic value of circulating KRAS mutations in lung cancer patients. *Cancer Lett.* **254**, 265–273 (2007).
152. Jian, G. *et al.* Prediction of epidermal growth factor receptor mutations in the plasma/pleural effusion to efficacy of gefitinib treatment in advanced non-small cell lung cancer. *J. Cancer Res. Clin. Oncol.* **136**, 1341–1347 (2010).
153. An, Q. *et al.* Detection of p16 hypermethylation in circulating plasma DNA of non-small cell lung cancer patients. *Cancer Lett.* **188**, 109–114 (2002).
154. Bearzatto, A. *et al.* $p16^{INK4A}$ Hypermethylation detected by fluorescent methylation-specific PCR in plasmas from non-small cell lung cancer. *Clin. Cancer Res.* **8**, 3782–3787 (2002).
155. Liu, Y. *et al.* Hypermethylation of $p16^{INK4a}$ in Chinese lung cancer patients: biological and clinical implications. *Carcinogenesis* **24**, 1897–1901 (2003).
156. Ng, C. S. *et al.* Tumor p16M is a possible marker of advanced stage in non-small cell lung cancer. *J. Surg. Oncol.* **79**, 101–106 (2002).
157. Ramirez, J. L. *et al.* 14-3-3sigma methylation in pretreatment serum circulating DNA of cisplatin-plus-gemcitabine-treated advanced non-small-cell lung cancer patients predicts survival: The Spanish Lung Cancer Group. *J. Clin. Oncol.* **23**, 9105–9112 (2005).
158. Hosny, G., Farahat, N. & Hainaut, P. TP53 mutations in circulating free DNA from Egyptian patients with non-Hodgkin's lymphoma. *Cancer Lett.* **275**, 234–239 (2009).
159. Au, W. Y., Pang, A., Choy, C., Chim., C. S. & Kwong, Y. L. Quantification of circulating Epstein-Barr virus (EBV) DNA in the diagnosis and monitoring of natural killer cell and EBV-positive lymphomas in immunocompetent patients. *Blood* **104**, 243–249 (2004).
160. Lei, K. I., Chan, L. Y., Chan, W. Y., Johnson, P. J. & Lo, Y. M. Diagnostic and prognostic implications of circulating cell-free Epstein-Barr virus DNA in natural killer/T-cell lymphoma. *Clin. Cancer Res.* **8**, 29–34 (2002).
161. Machado, A. S. *et al.* Circulating cell-free and Epstein-Barr virus DNA in pediatric B-non-Hodgkin lymphomas. *Leuk. Lymphoma* **51**, 1020–1027 (2010).
162. Deligezer, U., Yaman, F., Erten, N. & Dalay, N. Frequent copresence of methylated DNA and fragmented nucleosomal DNA in plasma of lymphoma patients. *Clin. Chim. Acta* **335**, 89–94 (2003).
163. Board, R. E. *et al.* Detection of BRAF mutations in the tumour and serum of patients enrolled in the AZD6244 (ARRY-142886) advanced melanoma phase II study. *Br. J. Cancer* **101**, 1724–1730 (2009).
164. Pinzani, P. *et al.* Allele specific Taqman-based real-time PCR assay to quantify circulating BRAFV600E mutated DNA in plasma of melanoma patients. *Clin. Chim. Acta* **411**, 1319–1324 (2010).
165. Fujimoto, A., O'Day, S. J., Taback, B., Elashoff, D. & Hoon, D. S. Allelic imbalance on 12q22–23 in serum circulating DNA of melanoma patients predicts disease outcome. *Cancer Res.* **64**, 4085–4088 (2004).
166. Fujiwara, Y. *et al.* Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients. *Cancer Res.* **59**, 1567–1571 (1999).
167. Taback, B. *et al.* Prognostic significance of circulating microsatellite markers in the plasma of melanoma patients. *Cancer Res.* **61**, 5723–5726 (2001).
168. Taback, B. *et al.* Circulating DNA microsatellites: molecular determinants of response to biochemotherapy in patients with metastatic melanoma. *J. Natl Cancer Inst.* **96**, 152–156 (2004).
169. Melnikov, A., Scholtens, D., Godwin, A. & Levenson, V. Differential methylation profile of ovarian cancer in tissues and plasma. *J. Mol. Diagn.* **11**, 60–65 (2009).
170. Muller, H. M. *et al.* Analysis of methylated genes in peritoneal fluids of ovarian cancer patients: a new prognostic tool. *Clin. Chem.* **50**, 2171–2173 (2004).
171. Swisher, E. M. *et al.* Tumor-specific p53 sequences in blood and peritoneal fluid of women with epithelial ovarian cancer. *Am. J. Obstet. Gynecol.* **193**, 662–667 (2005).
172. Zachariah, R. R. *et al.* Levels of circulating cell-free nuclear and mitochondrial DNA in benign and malignant ovarian tumors. *Obstet. Gynecol.* **112**, 843–850 (2008).
173. Liggett, T. *et al.* Differential methylation of cell-free circulating DNA among patients with pancreatic cancer versus chronic pancreatitis. *Cancer* **116**, 1674–1680 (2010).
174. Melnikov, A. A., Scholtens, D., Talamonti, M. S., Bentrem, D. J. & Levenson, V. V. Methylation profile of circulating plasma DNA in patients with pancreatic cancer. *J. Surg. Oncol.* **99**, 119–122 (2009).
175. Bastian, P. J. *et al.* Preoperative serum DNA GSTP1 CpG island hypermethylation and the risk of early prostate-specific antigen recurrence following radical prostatectomy. *Clin. Cancer Res.* **11**, 4037–4043 (2005).
176. Bryzgunova, O. E., Morozkin, E. S., Yarmoschuk, S. V., Vlassov, V. V. & Laktionov, P. P. Methylation-specific sequencing of GSTP1 gene promoter in circulating/extracellular DNA from blood and urine of healthy donors and prostate cancer patients. *Ann. N. Y Acad. Sci.* **1137**, 222–225 (2008).
177. Goessl, C., Muller, M., Heicappell, R., Krause, H. & Miller, K. DNA-based detection of prostate cancer in blood, urine, and ejaculates. *Ann. N. Y Acad. Sci.* **945**, 51–58 (2001).
178. Jeronimo, C. *et al.* Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. *Urology* **60**, 1131–1135 (2002).
179. Roupret, M. *et al.* Promoter hypermethylation in circulating blood cells identifies prostate cancer progression. *Int. J. Cancer* **122**, 952–956 (2008).
180. Ellinger, J. *et al.* Noncancerous PTGS2 DNA fragments of apoptotic origin in sera of prostate cancer patients qualify as diagnostic and prognostic indicators. *Int. J. Cancer* **122**, 138–143 (2008).
181. Ellinger, J., Muller, S. C., Wernert, N., von Ruecker, A. & Bastian, P. J. Mitochondrial DNA in serum of patients with prostate cancer: a predictor of biochemical recurrence after prostatectomy. *BJU Int.* **102**, 628–632 (2008).
182. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).

Acknowledgements
Funding was provided by Deutsche Forschungsgemeinschaft, Deutsche Krebshilfe, BMBF, Erich und Gertrud Roggenbuck-Stiftung, and the Sheldon and Miriam Adelson Foundation (to D.S.B.H.), the Melanoma Research Alliance (to D.S.B.H.), KOMEN BCTR0707528, CBCRP 16IB-0076 (to D.S.B.H.), and the US National Institutes of Health NCI R33, PO1 CA29605 and CA12582 (to D.S.B.H.).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Klaus Pantel's homepage:
www.uke.uni-hamburg.de/institute/tumorbiologie
Circulating Nucleic Acids in Plasma and Serum:
http://www.springer.com/biomed/cancer/book/978-90-481-9381-3
DISMAL: http://www.dismal-project.eu/
John Wayne Cancer Institute:
http://www.jwci.org/research-team.aspx?id=384
NACB:
http://www.aacc.org/members/nacb/pages/default.aspx

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2011 Macmillan Publishers Limited. All rights reserved
View publication stats

ATTACHMENT 3A

LDR	01550cas a22005414a 4500
001	9911241683406676
005	20250227135110.0
007	cr|cnu||||||||
008	010731c20019999enkmr p 0 a0eng
010	__ | **a** 2001243419
022	__ | **a**1474-175X**l**1474-175X**9**PY**7**(Print)
022	__ | **a**1474-1768**9**EN**7**(Electronic)
035	__ | **9**101124168
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
042	__ | **a**lcd
044	__ | **9**England
060	00 | **a**W1**b**NA839F
210	10 | **a**Nat Rev Cancer**2**DNLM
245	00 | **a**Nature reviews.**p**Cancer.
246	30 | **a**Cancer
260	__ | **a**London, UK :**b**Nature Pub. Group,**c**[c2001-
300	__ | **a**v.
310	__ | **a**Monthly
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**Vol. 1, no. 1 (Oct. 2001)-
500	__ | **a**Title from cover.
510	1_ | **9**1**a**MEDLINE**b**v1n1, Oct. 2001-
510	1_ | **9**1**a**PubMed**b**v1n1, Oct. 2001-
510	1_ | **a**Index medicus
530	__ | **a**Also available online to subscribers.
650	12 | **a**Neoplasms
650	22 | **a**Research
655	_2 | **a**Periodical
856	4_ | **u**http://www.nature.com/nrc
911	__ | **a**Electronic and print, print canceled 2017
914	__ | **a**Portico**b**Current to 2024 as of 20250227**c**ISSN_1474175X
935	__ | **a**(DNLM)1124168-nlmdb
992	__ | **p**P2**e**E4**a**20011026
993	__ | **a**SRW**b**20011030
994	__ | **a**CDN**b**20011030
995	__ | **a**AUTH**b**20011101**c**REV**d**20250227
996	__ | **a**CONSER**b**20011212

998	__ | **a**ERMSLC 108646**b**20061129
999	__ | **a**AUTH
035	__ | **a**(OCoLC)48623853

ATTACHMENT 3B

Neoplasms MeSH Descriptor Data 2025

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading
Neoplasms
Tree Number(s)
C04
Unique ID
D009369
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D009369
Annotation
general; prefer specifics; familial: consider also NEOPLASTIC SYNDROMES, HEREDITARY; metastatic cancer of unknown origin: index NEOPLASM METASTASIS
Scope Note
New abnormal growth of tissue. Malignant neoplasms show a greater degree of anaplasia and have the properties of invasion and metastasis, compared to benign neoplasms.
Entry Version
NEOPL
Entry Term(s)
Benign Neoplasm
Benign Neoplasms
Cancer
Malignancy
Malignant Neoplasm
Malignant Neoplasms
Neoplasia
Neoplasm
Neoplasms, Benign
Tumor
Tumors
Consider Also
consider also terms at CANCER, CARCINO-, ONCO-, and TUMOR
Public MeSH Note
/diagnosis was NEOPLASM DIAGNOSIS 1964-65; /etiology was NEOPLASM ETIOLOGY 1964-65; /immunology was NEOPLASM IMMUNOLOGY 1964-65; /radiotherapy was NEOPLASM RADIOTHERAPY 1964-65; /therapy was NEOPLASM THERAPY 1964-65; NEOPLASM STATISTICS was heading 1964-65; CARCINOGENESIS was heading 1977
History Note
/diagnosis was NEOPLASM DIAGNOSIS 1964-65; /etiology was NEOPLASM ETIOLOGY 1964-65; /immunology was NEOPLASM IMMUNOLOGY 1964-65; /radiotherapy was NEOPLASM RADIOTHERAPY 1964-65; /therapy was NEOPLASM THERAPY 1964-65; NEOPLASM STATISTICS was heading 1964-65; CARCINOGENESIS was heading 1977
Entry Combination
secondary:Neoplasm Metastasis
Date Established

1966/01/01

Date of Entry

1999/01/01

Revision Date

2021/06/30

page delivered in 0.142s

Connect with NLM

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Web Policies
FOIA
HHS Vulnerability Disclosure

NLM Support Center
Accessibility
Careers

NLM | NIH | HHS | USA.gov

Scan to visit

Fact Sheet

Human Genomic Variation

The vast majority of the DNA letters in peoples' genomes is identical, but a small fraction of those letters varies. This genomic variation accounts for some of the differences among people, including important aspects of their health and susceptibility to diseases.

The Big Picture

Genomic variation reflects the differences in a person's DNA compared to other peoples' DNA.

There are multiple types of variants in human genomes, ranging from small differences to large differences.

A very small subset of genomic variants contributes to human health and disease.

Researchers create reference human genome sequences to help detect genomic variants in each sequenced human genome.

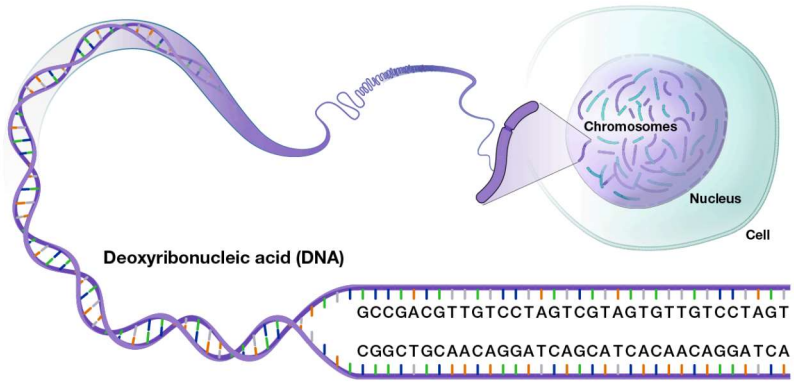
On average, a person's genome sequence is ~99.6% identical to a reference human genome sequence; that person's set of genomic variants accounts for the ~0.4% difference.

EX1015

The human pangenome is a more comprehensive framework that aims to account for genomic variation across human populations, thereby reducing biases that can come with the use of a single reference human genome sequence.

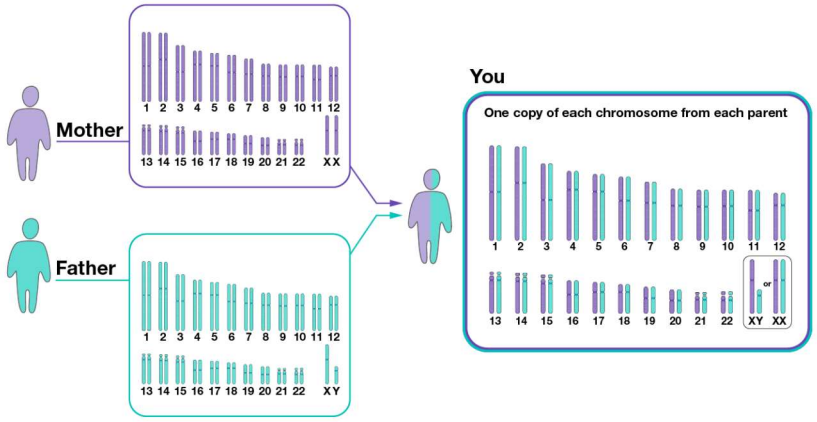
What is the human genome?

A **genome** is the complete set of DNA instructions found in every cell. **DNA** is made of four different chemicals (called nucleotides or bases), each represented by a different letter: adenine (A), thymine (T), cytosine (C) and guanine (G). The order of these letters (i.e., the DNA sequence) encodes the information that instructs each cell what to do and when to do it. The genome's DNA is packed into structures called **chromosomes**.



A genome consists of all the DNA in a cell, which is packed into chromosomes that reside in the nucleus of cells.

One copy of the human genome contains about 3 billion nucleotides, which are distributed among 23 chromosomes. Most human cells have two copies of the human genome, with one copy inherited from each parent. Cells containing two copies of each chromosome are called "diploid." Most mammals are **diploid**, but some organisms have either one set or more than two sets of each chromosome.



A person's genome consists of 23 pairs of chromosomes. One copy of each chromosome is inherited from each parent. For example, chromosomes from the depicted mother are labeled in purple, and chromosomes from the depicted father are labeled in teal. The X and Y chromosomes are known as sex chromosomes. Although there are exceptions, biological females have two X chromosomes (denoted XX) and biological males have one X chromosome and one Y chromosome (denoted XY).

How do peoples' genomes vary?

Peoples' genomes are far more similar to each other than they are different. It is frequently stated that any two peoples' genomes are ~99.9% identical to one another. This percentage is based on the finding that, on average, a single-nucleotide difference exists between two peoples' genomes once every 1,300 nucleotides or so.

However, this is an over-simplification because it only accounts for single-nucleotide differences. In reality, any two peoples' genomes are, on average, ~99.6% identical and ~0.4% different. The latter percentage reflects both single-nucleotide differences and differences that involve multiple nucleotides (discussed in detail below).

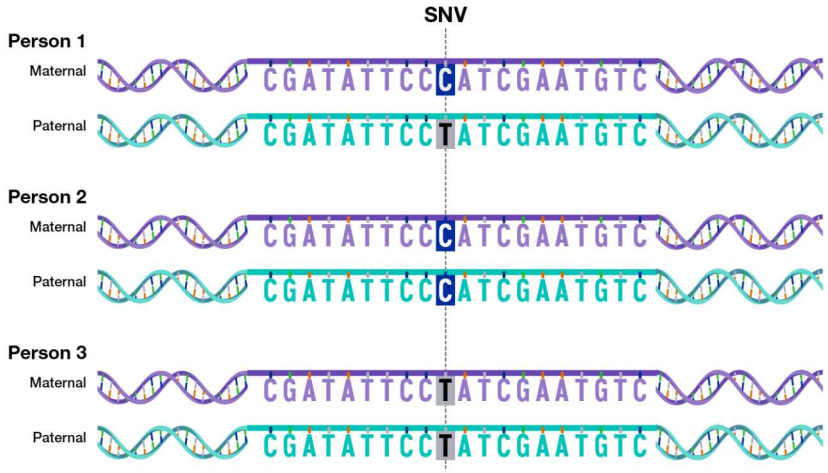
The differences among human genomes are called genomic variants. A person's set of genomic variants is part of what makes them unique. Other factors (such as diet, environment, lifestyle and social context) also contribute to a person's uniqueness. Most genomic variants have no influence on the functioning of a person's genome, but a small subset of variants do have an impact. For example, some genomic

variants influence physical characteristics, like eye color and height; others influence health conditions or how a person responds to certain medications.

What are the different types of genomic variants?

There are multiple types of genomic variants.

The smallest genomic variants are single-nucleotide variants (SNVs). Each SNV reflects a difference in a single nucleotide (or letter). For a given SNV, the DNA letter at that genomic position might be a C in one person but a T in another person. SNVs are the most common type of genomic variation. A subtype of SNVs is called a single-nucleotide polymorphism (SNP; pronounced “snip”). To be considered a SNP, a SNV must be present in at least 1% of the human population. As such, SNV is a more general term that includes both relatively common (such as SNPs) and rare single-nucleotide differences. For simplicity, we refer to all single-nucleotide differences as SNVs, regardless of their relative frequency.

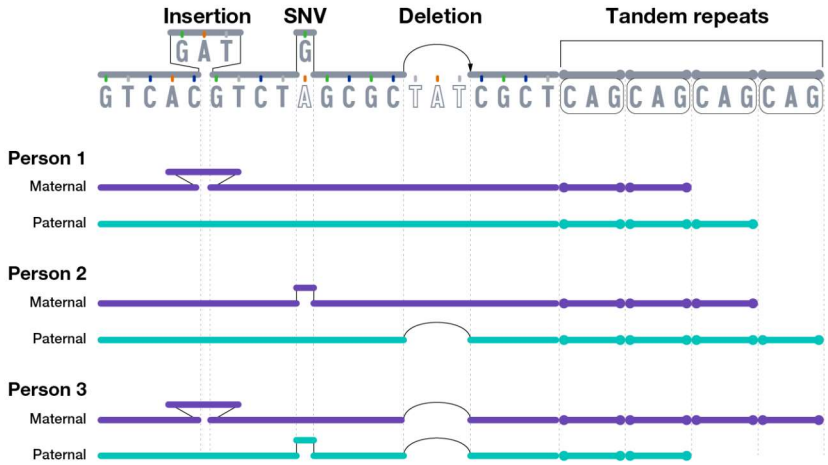


Single-nucleotide variants (SNVs) are differences of one nucleotide at a specific location in the genome. An individual may have different nucleotides at such a location on each chromosome (getting a different one from each parent), such as with Person 1. An individual may also have the same nucleotide at such a location on both chromosomes, such as with Person 2 and Person 3.

Another group of small genomic variants are insertions and deletions (often referred to as “indels”). Insertion/deletion variants reflect extra or missing DNA

nucleotides in the genome, respectively, and typically involve fewer than 50 nucleotides. Insertion/deletion variants are less frequent than SNVs but can sometimes have a larger impact on health and disease (e.g., by disrupting the function of a gene that encodes an important protein).

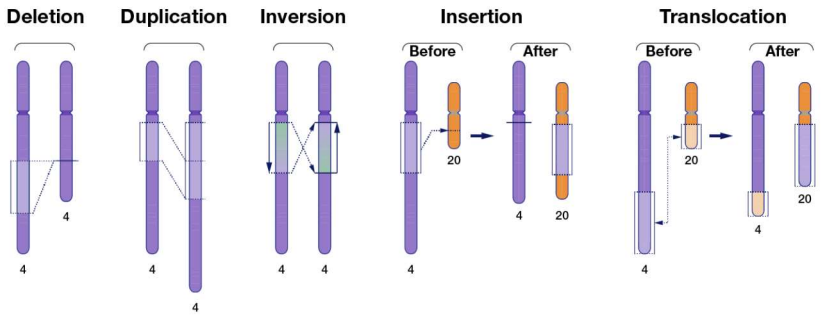
One of the most common types of insertion/deletion variants are tandem repeats (also known as microsatellites). **Tandem repeats** are short stretches of nucleotides that are repeated multiple times and are highly variable among people. Different chromosomes can vary in the number of times such short nucleotide stretches are repeated, ranging from a few times to hundreds of times. Historically, tandem repeats have been used for building maps of the human genome and DNA profiling in forensics applications.



Each person has a collection of different genomic variants. For example, Person 1 has an insertion variant; Person 2 has a SNV and deletion variant; and Person 3 has an insertion, SNV, and deletion variant. All three people have different tandem repeats. Different variants can be inherited from different parents, either the maternal genome (purple) or paternal genome (teal).

Genomic variation also extends beyond small stretches of nucleotides to larger chromosomal regions. These large-scale genomic differences are called structural variants and involve at least 50 nucleotides and as many as thousands of nucleotides that have been inserted, deleted, inverted or moved from one part of the genome to another. Tandem repeats that contain more than 50 nucleotides are considered structural variants; in fact, such large tandem repeats account for nearly

half of the structural variants present in human genomes. When a structural variant reflects differences in the total number of nucleotides involved, it is called a **copy-number variant** (CNV). Note that CNVs are distinguished from other structural variants, such as inversions and translocations, because the latter types often do not involve a difference in the total number of nucleotides.



Larger structural variants are present among human genomes. For example, human chromosomes can have missing segments (deletion variants), duplicated segments (duplication variants), inverted segments (inversion variants), added segments (insertion variants) and segments transferred from other chromosomes (translocation variants).

How are genomic variants detected?

Researchers use different approaches to detect genomic variants.

Some methods are designed to only detect known genomic variants. For example, it might be important to know if a person has inherited a particular genomic variant that is relevant to their health or healthcare; for that, researchers can perform a specific DNA test to determine whether a person has that genomic variant. With other methods, researchers can analyze a person's DNA for the presence of a large number of known genomic variants; for example, a type of test called a microarray can detect hundreds of thousands of SNVs at once.

A more comprehensive way to detect genomic variants, including those that might not yet be known, is to perform genome sequencing. Multiple methods are now available for sequencing human genomes, each only requiring a small sample of

blood, hair or cheek cells from which DNA is isolated. Today, sequencing a human genome usually costs less than $1,000, which is over a million-fold less expensive than a couple of decades ago.

When a person's genome is sequenced, both copies of each pair of chromosomes (one from each parent) are sequenced at the same time. In most routine situations (e.g., when a genome is sequenced for a medical purpose), it is not readily possible to determine the parent of origin for each detected genomic variant. This is because a human genome is not sequenced one chromosome at a time; rather, the process involves breaking up the entire genome and then piecing back together small stretches of the sequenced DNA, like a jigsaw puzzle. Without additional analyses, that process does not provide information about the parental origin of each genomic variant. A "reference" human genome sequence is also critical to sequencing a human genome and detecting genomic variants.

What is a reference genome sequence?

Detecting variants in a person's genome sequence usually requires an existing sequence for comparison — a "reference." A reference human genome sequence is an established, high-quality and well-accepted sequence of a human genome (i.e., one sequence for each of the 23 human chromosomes). A **reference human genome sequence** is not an actual sequence from one individual but pieced together from multiple people. The important feature of a reference human genome sequence is that it depicts one assigned nucleotide for every position across the human genome. In this regard, a reference human genome sequence only represents the sequence of one copy of each chromosome, whereas a person's genome sequence contains the sequences of both copies of each chromosome (i.e., people are diploid).



A reference genome sequence (green) is used for comparing individual human genome sequences (purple) to find genomic variants. Specifically, the side-by-side comparison of a newly generated human genome sequence with a reference human genome sequence allows for the detection of genomic variants in the former. Although only one generated human genome sequence is shown for the purposes of simplification, in reality, each person actually has two genome sequences (one inherited from each parent), both of which would be compared to the reference genome sequence.

Of course, the human population contains an immense amount of variation among genomes. A given reference human genome sequence is not intended to capture or represent that diversity; rather, it serves as an important data-analysis tool. Specifically, each newly generated genome sequence from a person is directly compared to a given reference human genome sequence, allowing for the detection of all differences (variants) between the person's genome and the reference genome. In short, comparison of a person's genome sequence to a reference genome sequence allows for the cataloguing of all genomic variants in that person's genome.

What is the inventory of genomic variants in a typical human genome sequence?

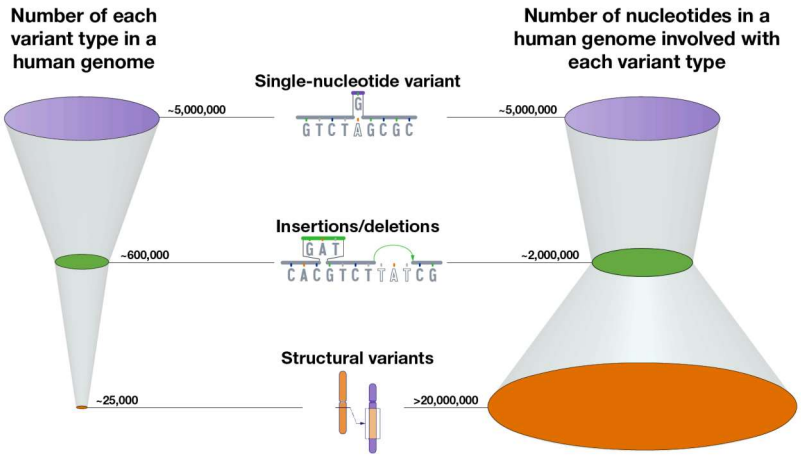
What does a typical human genome sequence look like with respect to the variants that it contains? On average, compared to a reference human genome, a person's ~6 billion-nucleotide genome sequence will have:

~5,000,000 SNVs that involve ~5,000,000 nucleotides

~600,000 insertion/deletion variants that involve ~2,000,000 nucleotides

~25,000 structural variants that involve >20,000,000 nucleotides.

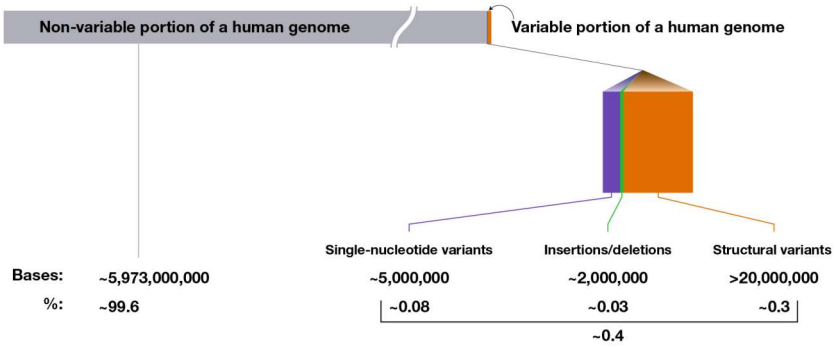
These numbers represent current estimated averages and will likely change as more human genomes are sequenced to completion. Furthermore, multiple approaches can be used to calculate how the different types of genomic variants account for the differences among genomes. In fact, calculating the total number of nucleotides involved in structural variants in each human genome is quite complicated and remains an active area of research.



On average, a person's genome sequence contains the total numbers of each of the three major types of genomic variants indicated on the left. Together, those genomic variants involve the total number of nucleotides indicated on the right. Note that each SNV only involves a single nucleotide, whereas each insertion/deletion and structural variant involves more than a single nucleotide. As a result, each type of genomic variant differs with respect to its overall contribution to genomic variation in terms of the number of nucleotides involved (e.g., a given structural variant contributes substantially more than a given SNV). Overall, insertions/deletions (in green) are less common than SNVs (purple) and involve fewer nucleotides, whereas structural variants (in orange) are less common than either SNVs or insertion/deletions but involve the greatest number of nucleotides.

That means that, on average, the complete set of genomic variants in each person's genome involves ~27,000,000 nucleotides (among the ~6,000,000,000 nucleotides in their genome). Those ~27,000,000 nucleotides reflect some type of difference at those positions in the DNA, together accounting for ~0.4% of the person's complete

genome. In other words, when accounting for the full inventory of genomic variants, a typical person's genome sequence is ~99.6 identical to (or ~0.4% different from) a reference human genome sequence (or even another person's genome sequence).



On average, a human genome sequence (consisting of ~6,000,000,000 nucleotides) is ~99.6% identical — or non-variable — compared to a reference human genome sequence. The variable portion of a human genome sequence totals ~0.4%, which consists of the indicated contributions from each of the three major types of genomic variants. These same general numbers are also found when directly comparing two peoples' genome sequences (as opposed to comparing to a reference human genome sequence).

What is a pangenome reference sequence and how is it used?

Reference human genome sequences are invaluable for detecting genomic variants in each newly sequenced human genome. But the currently available reference human genome sequences do not accurately represent the genomic diversity of the human population, and that lack of diversity can introduce biases when analyzing some peoples' genome sequences.

To address this deficiency, researchers are working to generate a more complete set of reference human genome sequences that better reflect all of humanity. A "pangenome" is the collective genome sequences of multiple individuals that better represents the genomic diversity of the species. The human pangenome reference sequence will provide a better tool for comparing genome sequences from people all over the world in an effort to detect and characterize genomic variants more completely, including those with important roles in health and disease.

Why does genome variation matter?

Genomic variation drives evolution and serves to expand biodiversity. This is true of humans as well as for plants, animals and other organisms. Such diversity keeps populations healthy and is fundamental to natural selection, an evolutionary process by which organisms adapt to changing environments.

Human genomic variation is also highly relevant in the field of medicine. Only a small fraction of genomic variants affects human health. In some cases, genomic variants directly cause diseases (such as in cystic fibrosis and sickle cell disease). In other cases, the effects of genomic variants are more subtle (such as in hypertension and diabetes, where a genomic variant might contribute to the overall risk that a person might have of the condition). Healthcare professionals are increasingly learning how to use information about patients' genomic variants to manage their medical care — something known as **genomic medicine**.

Looking forward

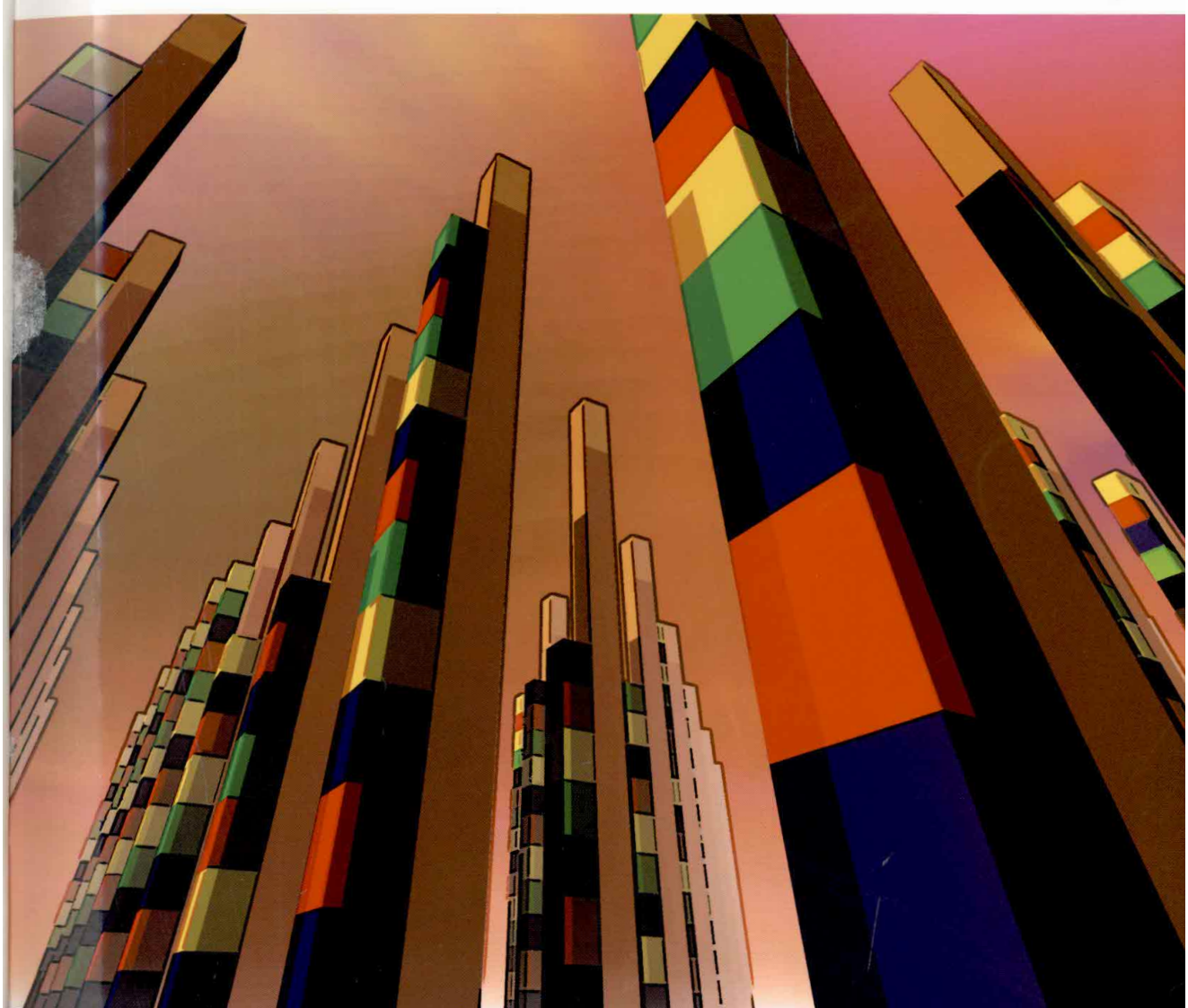
Genomics is constantly evolving and advancing, including the regular development of new experimental technologies and data-analysis methods. Such advances are readily applicable to ongoing efforts to develop new and more inclusive sets of reference human genome sequences and improve the ability to detect all variants in each newly sequenced human genome. The long-term goal is to have sufficient knowledge about genomic variation in all human populations, bringing equity in the benefits of genomic medicine.

Nature reviews. Genetics.
v. 11, no. 1 (Jan. 2010)
General Collection
W1 NA839H
2010-02-02 08:39:27

REVIEWS

january 2010 volume 11 no. 1
www.nature.com/reviews

GENETICS



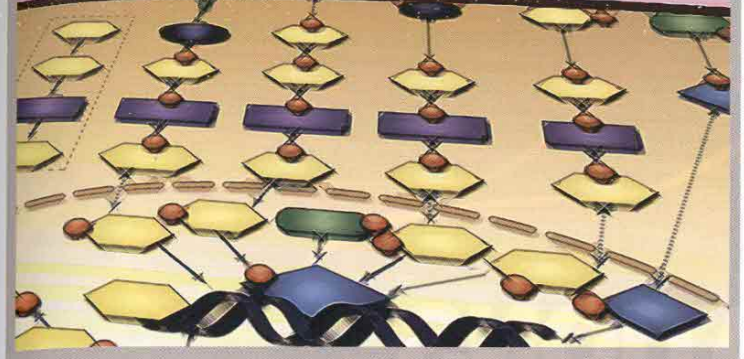
BUILDING THE SEQUENCE

A guide to the features of next-generation sequencing technologies

Evolving immune systems

Selection in human populations and origins of adaptive immunity

EX1017



Kinase function and disease p60

CONTENTS

January 2010
volume 11 no. 1

REVIEWS

17 **From evolutionary genetics to human immunology: how selection shapes host defence genes**

Luis B. Barreiro and Lluís Quintana-Murci

The recent increase in available human population genetic data is revealing signatures of selection at immunity-related genes. This Review discusses how studies of genome evolution are contributing to our understanding of how pathogens have shaped the human genome and immune response.

31 FEATURED ARTICLE **Sequencing technologies — the next generation**

Michael L. Metzker

There is an increasing demand for next-generation sequencing technologies that rapidly deliver high volumes of accurate genome information at a low cost. This Review provides a guide to the features of the different platforms, and describes the recent advances in this fast-moving area.

47 **Origin and evolution of the adaptive immune system: genetic events and selective pressures**

Martin F. Flajnik and Masanori Kasahara

How did the intricate adaptive immune system of mammals arise? New clues have recently emerged from studies of the immune systems of non-mammalian vertebrates. Here, these findings are integrated with current knowledge of macroevolutionary events and selective pressures.

60 **Kinase mutations in human disease: interpreting genotype–phenotype relationships**

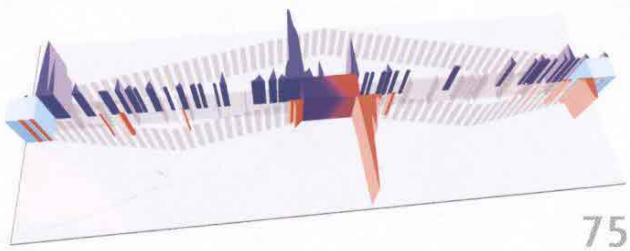
Piya Lahiry, Ali Torkamani, Nicholas J. Schork and Robert A. Hegele

Protein kinases are one of the largest gene families in humans. Through a comprehensive survey of kinase mutations the authors highlight the structure–function relationships that control the activities of these proteins and their contribution to a varied set of human diseases.

75 **RNA processing and its regulation: global insights into biological networks**

Donny D. Licatalosi and Robert B. Darnell

mRNA repertoires can be diversified by many mechanisms, including alternative splicing and alternative polyadenylation. Technological advances are now allowing genome-wide insights into the extent of RNA processing, the actions of RNA-binding proteins and how regulation at the RNA level helps to control biological systems.



75

On the web www.nature.com/reviews/genetics

Advance online publication
We operate an advance online publication (AOP) service for authors and readers to view the latest articles published online ahead of print.

Forthcoming articles
Genetics of mammalian meiosis: regulation, sex and fertility
Mary Ann Handel and John Schimenti

Personalized genomic information: preparing for the future of genetic medicine
Alan E. Guttmacher, Amy McGuire, Bruce Ponder and Kári Stefánsson

The evolution of gene duplications: classifying and distinguishing between models
Hideki Innan and Fyodor Kondrashov

Links to further information
The full text of articles includes:
- Author biographies
- Links to glossary terms
- Links to genes and proteins in databases such as Entrez, FlyBase, OMIM and UniProtKB.

At a glance provides a bullet-pointed summary of the main topics covered in each article.

More articles like this for further reading.
Each Review concludes with a 'More articles like this' box containing links to related content on NPG.

Featured article can be browsed online by registered members of nature.com.

PowerPoint figures
Figures can be downloaded as a PowerPoint file for use in presentations and for teaching.

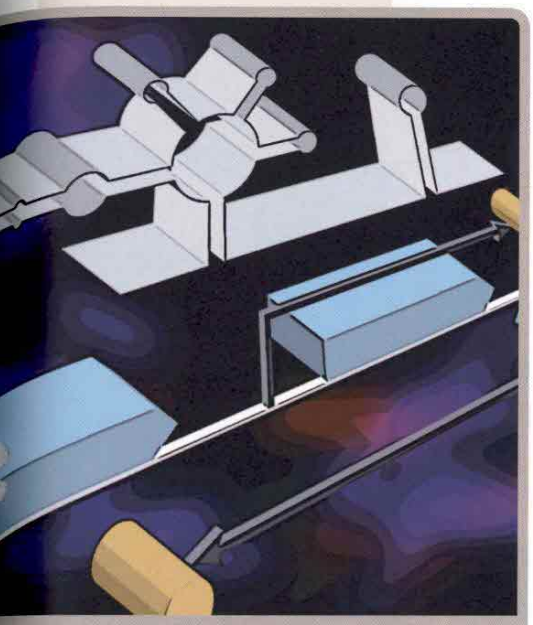
E-alert table of contents
Get monthly e-mail alerts to the content of this journal — sent FREE to your inbox — by registering online. Or sign up to receive the latest content as an RSS newsfeed by visiting www.nature.com/reviews/genetics

Online correspondence
Comments from readers about our articles.
All correspondence will be highlighted on the table of contents and on e-mail alerts.

SERIES

APPLICATIONS OF NEXT-GENERATION SEQUENCING

http://www.nature.com/nrg/series/nextgeneration



Bacterial transcriptional control p9

FROM THE EDITORS

1 An editorial commentary and introduction to this month's issue

RESEARCH HIGHLIGHTS

3 Selected highlights from the recent research literature

PROGRESS

9 **Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity**

Rotem Sorek and Pascale Cossart

Until recently, large-scale transcriptome studies in bacteria and archaea were limited by technical challenges, and there was a perception that microbial transcription was relatively simple compared with eukaryotic transcription. Now, prokaryotic transcriptomics is revealing unexpected aspects of transcriptional control, genome organization and non-coding RNAs.



Image courtesy of the Kuroda laboratory. 5

nature REVIEWS GENETICS

DISCOUNT SUBSCRIPTIONS

STUDENTS! Did you know that an online-only subscription to *Nature Reviews Genetics* costs just £55 (US$99/€85)? Combined print and online subscriptions are also available for personal and student subscribers — to subscribe visit www.nature.com/reviews/genetics

SUBSCRIPTION INFORMATION

Nature Reviews Genetics (ISSN 1471-0056) is published monthly by Nature Publishing Group.
Editorial office: The Macmillan Building, 4–6 Crinan Street, London N1 9XW, UK.
Tel: +44 (0)20 7843 3601; Fax: +44 (0)20 7843 3629
E-mail: NatureReviews@nature.com
New subscriptions/renewals/back issues and all other customer service questions should be addressed to:
Americas: *Nature Reviews Genetics*, Nature Publishing Group, Customer Services, 75 Varick Street, 9th floor, New York, NY 10013-1917, USA.
Tel: +1 212 726 9200; Fax: +1 212 696 9006
E-mail: subscriptions@natureny.com
UK/Euro/ROW (excluding Japan): *Nature Reviews Genetics*, Subscriptions Dept, Brunel Road, Basingstoke, Hants RG21 6XS, UK.
Tel: +44 (0)1256 302 827
Fax: +44 (0)1256 812 358
E-mail: subscriptions@nature.com
Japan: Japan Publications Trading Co. Ltd, 2-1 Sarugaku-cho 1 chome, Chiyoda-ku, Tokyo 101, Japan. Tel: +81 (03) 292 3755
Annual subscription rates: US: $3,520 (institutional/corporate), $265 (personal one year), $450 (personal two year), $199 (student/postdoc); Canada: add 6% GST #104911595RT001; Europe: €2,795 (institutional/corporate), €248 (personal one year), €422 (personal two year), €169 (student/postdoc); UK/ROW (excluding China, Japan, Korea): £1,806 (institutional/corporate), £160 (personal one year), £272 (personal two year), £109 (student/postdoc); Japan: contact Japan Publications Trading Co. Ltd, 2-1 Sarugaku-cho 1 chome, Chiyoda-ku, Tokyo 101, Japan. Tel: +81 (03) 292 3755. All prices are subject to local VAT/Tax where appropriate. To check if VAT/Tax applies go to www.nature.com/tax
CUSTOMER SERVICES: Feedback@nature.com
POSTMASTER: Periodicals Postage is paid at Rahway New Jersey and at additional mailing offices. Send address changes to Subscription Department, PO Box 5161, Brentwood, TN 37024-5161, USA.

SITE LICENCES

Site licence contacts:
Americas: Tel: +1 800 221 2123
E-mail: institutions@natureny.com
Asia Pacific: Tel: +81 3 3267 8769
E-mail: institutions@natureasia.com
Australia and New Zealand:
Tel: +61 3 9825 1160
E-mail: nature@macmillan.com.au
India: Tel: +91 12 4288 1054
E-mail: npgindia@nature.com
UK/Euro/ROW: Tel: +44 (0)20 7843 4759
E-mail: institutions@nature.com

DISPLAY ADVERTISING

display@natureny.com (USA) or display@nature.com
Global Head of Display Advertising:
Andrew Douglas
Tel: +44 (0) 20 7843 4960;
Fax: +44 (0) 20 7843 4749

DISPLAY ACCOUNT MANAGERS

New England:
Sheila Reardon
Tel: +1 617 399 4098; Fax +1 617 426 3717
New York, mid-Atlantic, South East:
Jim Breault
Tel: +1 212 726 9334; Fax: +1 212 726 9481
Midwest: Mike Rossi
Tel: +1 212 726 9255; Fax: +1 212 696 9481
West Coast:
George Lui
Tel: +1 415 781 3804; Fax: +1 415 781 3805
Germany, Switzerland, Austria:
Sabine Hugi-Fürst
Tel: +41 52761 3386; Fax: +41 52761 3419
UK, Ireland, Scandinavia, Spain, Portugal:
Evelina Rubio Hakansson
Tel: +44 (0) 20 7014 4079; Fax: + 44 (0) 20 7843 4749
France, The Netherlands, Belgium, Luxembourg, Italy, other Europe: Nicola Wright
Tel: +44 (0) 20 7843 4959; Fax: +44 (0) 20 7843 4749
UK, Germany, Switzerland, Austria:
Nancy Luksch
Tel: +44 (0) 20 7843 4968; Fax: +44 (0) 20 7843 4749
Japan: nature@naturejpn.com
Kate Yoneyama, Ken Mikami, NPG Nature Asia–Pacific
Tel: +81 3 3267 8765; Fax: +81 3 3267 8746
Greater China, Singapore:
Gloria To
Tel: +852 2811 7191; Fax: +852 2811 0743

This material may be protected by Copyright law (Title 17 U.S. Code)

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

Michael L. Metzker[*‡]

Abstract | Demand has never been greater for revolutionary technologies that deliver fast, inexpensive and accurate genome information. This challenge has catalysed the development of next-generation sequencing (NGS) technologies. The inexpensive production of large volumes of sequence data is the primary advantage over conventional methods. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly approaches, and recent advances in current and near-term commercially available NGS instruments. I also outline the broad range of applications for NGS technologies, in addition to providing guidelines for platform selection to address biological questions of interest.

Automated Sanger sequencing
This process involves a mixture of techniques: bacterial cloning or PCR; template purification; labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase; capillary electrophoresis; and fluorescence detection that provides four-colour plots to reveal the DNA sequence.

*Human Genome Sequencing Center and Department of Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, N1409, Houston, Texas 77030, USA.
‡LaserGen, Inc., 8052 El Rio Street, Houston, Texas 77054, USA.
e-mail: mmetzker@bcm.edu
doi:10.1038/nrg2626
Published online
8 December 2009

Over the past four years, there has been a fundamental shift away from the application of automated Sanger sequencing for genome analysis. Prior to this departure, the automated Sanger method had dominated the industry for almost two decades and led to a number of monumental accomplishments, including the completion of the only finished-grade human genome sequence[1]. Despite many technical improvements during this era, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Recent efforts have been directed towards the development of new methods, leaving Sanger sequencing with fewer reported advances. As such, automated Sanger sequencing is not covered here, and interested readers are directed to previous articles[2,3].

The automated Sanger method is considered as a 'first-generation' technology, and newer methods are referred to as next-generation sequencing (NGS). These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. The arrival of NGS technologies in the marketplace has changed the way we think about scientific approaches in basic, applied and clinical research. In some respects, the potential of NGS is akin to the early days of PCR, with one's imagination being the primary limitation to its use. The major advance offered by NGS is the ability to produce an enormous volume of data cheaply — in some cases in excess of one billion short reads per instrument run. This feature expands the realm of experimentation beyond just determining the order of bases. For example, in gene-expression studies microarrays are now being replaced by seq-based methods, which can identify and quantify rare transcripts without prior knowledge of a particular gene and can provide information regarding alternative splicing and sequence variation in identified genes[4,5]. The ability to sequence the whole genome of many related organisms has allowed large-scale comparative and evolutionary studies to be performed that were unimaginable just a few years ago. The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease. The variety of NGS features makes it likely that multiple platforms will coexist in the marketplace, with some having clear advantages for particular applications over others.

This Review focuses on commercially available technologies from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market in 2010. Nanopore sequencing is not covered, although interested readers are directed to an article by Branton and colleagues[6], who describe the advances and remaining challenges for this technology. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly, and current NGS platform performance to provide guidance on how these technologies work and how they may be applied to important biological questions. I highlight the applications of human genome resequencing using targeted and whole-genome approaches, and discuss the progress

and limitations of these methods, as well as upcoming advances and the impact they are expected to have over the next few years.

Next-generation sequencing technologies

Sequencing technologies include a number of methods that are grouped broadly as template preparation, sequencing and imaging, and data analysis. The unique combination of specific protocols distinguishes one technology from another and determines the type of data produced from each platform. These differences in data output present challenges when comparing platforms based on data quality and cost. Although quality scores and accuracy estimates are provided by each manufacturer, there is no consensus that a 'quality base' from one platform is equivalent to that from another platform. Various sequencing metrics are discussed later in the article.

In the following sections, stages of template preparation and sequencing and imaging are discussed as they apply to existing and near-term commercial platforms. There are two methods used in preparing templates for NGS reactions: clonally amplified templates originating from single DNA molecules, and single DNA-molecule templates. The term sequencing by synthesis, which is used to describe numerous DNA polymerase-dependent methods in the literature, is not used in this article because it fails to delineate the different mechanisms involved in sequencing[2,7]. Instead, these methods are classified as cyclic reversible termination (CRT), single-nucleotide addition (SNA) and real-time sequencing. Sequencing by ligation (SBL), an approach in which DNA polymerase is replaced by DNA ligase, is also described. Imaging methods coupled with these sequencing strategies range from measuring bioluminescent signals to four-colour imaging of single molecular events. The voluminous data produced by these NGS platforms place substantial demands on information technology in terms of data storage, tracking and quality control (see REF. 8 for details).

Template preparation

The need for robust methods that produce a representative, non-biased source of nucleic acid material from the genome under investigation cannot be overemphasized. Current methods generally involve randomly breaking genomic DNA into smaller sizes from which either fragment templates or mate-pair templates are created. A common theme among NGS technologies is that the template is attached or immobilized to a solid surface or support. The immobilization of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously.

Clonally amplified templates. Most imaging systems have not been designed to detect single fluorescent events, so amplified templates are required. The two most common methods are emulsion PCR (emPCR)[9] and solid-phase amplification[10]. EmPCR is used to prepare sequencing templates in a cell-free system, which has the advantage of avoiding the arbitrary loss of genomic sequences — a problem that is inherent in bacterial cloning methods. A library of fragment or mate-pair targets is created, and adaptors containing universal priming sites are ligated to the target ends, allowing complex genomes to be amplified with common PCR primers. After ligation, the DNA is separated into single strands and captured onto beads under conditions that favour one DNA molecule per bead (FIG. 1a). After the successful amplification and enrichment of emPCR beads, millions can be immobilized in a polyacrylamide gel on a standard microscope slide (Polonator)[11], chemically crosslinked to an amino-coated glass surface (Life/APG; Polonator)[12] or deposited into individual PicoTiterPlate (PTP) wells (Roche/454)[13] in which the NGS chemistry can be performed.

Solid-phase amplification can also be used to produce randomly distributed, clonally amplified clusters from fragment or mate-pair templates on a glass slide (FIG. 1b). High-density forward and reverse primers are covalently attached to the slide, and the ratio of the primers to the template on the support defines the surface density of the amplified clusters. Solid-phase amplification can produce 100–200 million spatially separated template clusters (Illumina/Solexa), providing free ends to which a universal sequencing primer can be hybridized to initiate the NGS reaction.

Single-molecule templates. Although clonally amplified methods offer certain advantages over bacterial cloning, some of the protocols are cumbersome to implement and require a large amount of genomic DNA material (3–20 μg). The preparation of single-molecule templates is more straightforward and requires less starting material (<1 μg). More importantly, these methods do not require PCR, which creates mutations in clonally amplified templates that masquerade as sequence variants. AT-rich and GC-rich target sequences may also show amplification bias in product yield, which results in their underrepresentation in genome alignments and assemblies. Quantitative applications, such as RNA–seq[5], perform more effectively with non-amplified template sources, which do not alter the representational abundance of mRNA molecules.

Before the NGS reaction is carried out, single-molecule templates are usually immobilized on solid supports using one of at least three different approaches. In the first approach, spatially distributed individual primer molecules are covalently attached to the solid support[14]. The template, which is prepared by randomly fragmenting the starting material into small sizes (for example, ~200–250 bp) and adding common adaptors to the fragment ends, is then hybridized to the immobilized primer (FIG. 1c). In the second approach, spatially distributed single-molecule templates are covalently attached to the solid support[14] by priming and extending single-stranded, single-molecule templates from immobilized primers (FIG. 1c). A common primer is then hybridized to the template (FIG. 1d). In either approach, DNA polymerase can bind to the immobilized primed template configuration to initiate the NGS reaction. Both of the above approaches are used by Helicos BioSciences. In a third approach, spatially distributed single polymerase molecules

Finished grade
A quality measure for a sequenced genome. A finished-grade genome, commonly referred to as a 'finished genome', is of higher quality than a draft-grade genome, with more base coverage and fewer errors and gaps (for example, the human genome reference contains 2.85 Gb, covers 99% of the genome with 341 gaps, and has an error rate of 1 in every 100,000 bp).

Template
This recombinant DNA molecule is made up of a known region, usually a vector or adaptor sequence to which a universal primer can bind, and the target sequence, which is typically an unknown portion to be sequenced.

Seq-based methods
Assays that use next-generation sequencing technologies. They include methods for determining the sequence content and abundance of mRNAs, non-coding RNAs and small RNAs (collectively called RNA–seq) and methods for measuring genome-wide profiles of immunoprecipitated DNA–protein complexes (ChIP–seq), methylation sites (methyl–seq) and DNase I hypersensitivity sites (DNase–seq).

Polonator
This Review mostly describes technology platforms that are associated with a respective company, but the Polonator G.007 instrument, which is manufactured and distributed by Danaher Motions (a Dover Company), is an open source platform with freely available software and protocols. Users manufacture their own reagents based on published reports or by collaborating with George Church and colleagues or other technology developers.

Fragment templates
A fragment library is prepared by randomly shearing genomic DNA into small sizes of <1 kb, and requires less DNA than would be needed for a mate-pair library.

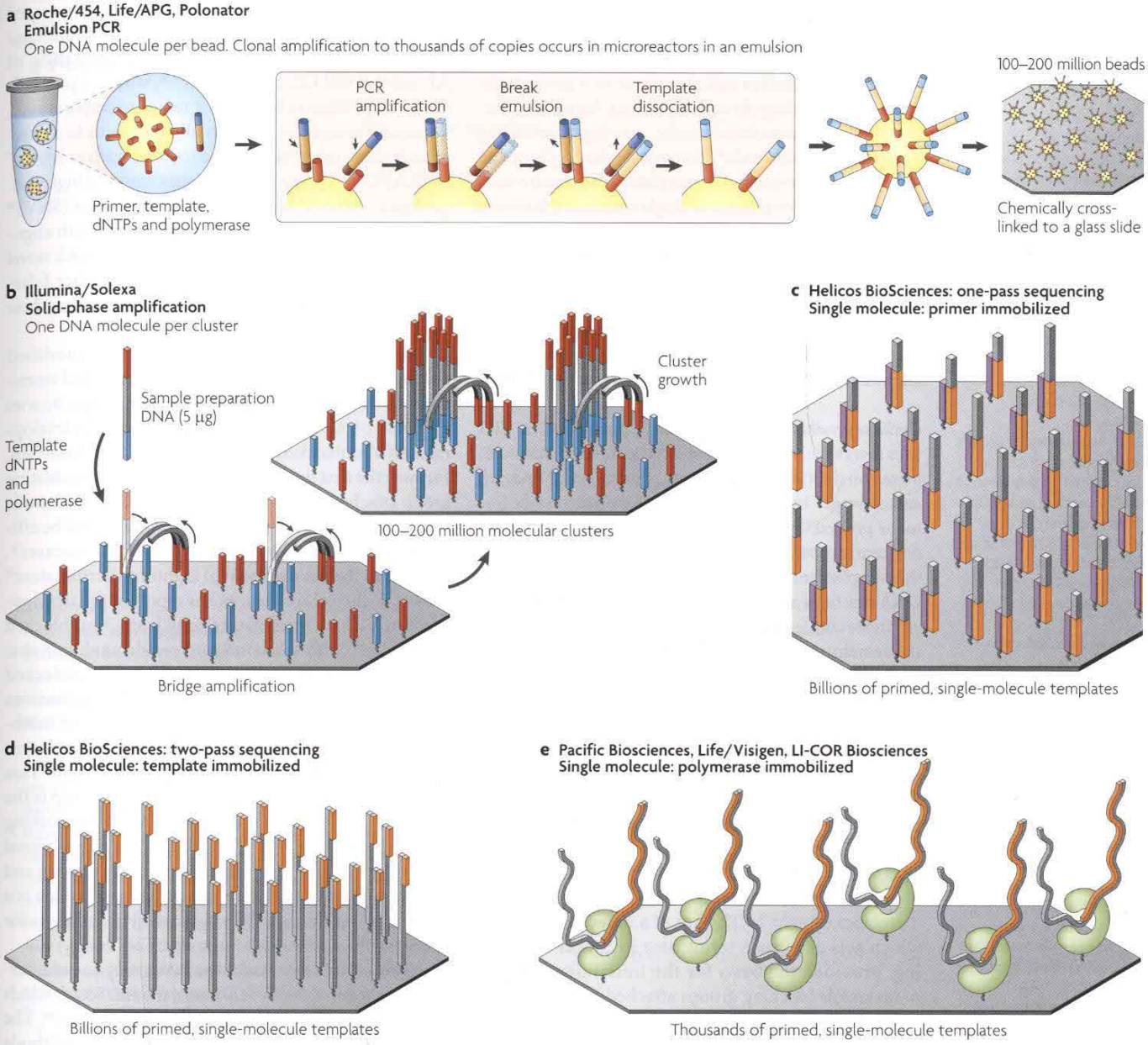


Figure 1 | **Template immobilization strategies.** In emulsion PCR (emPCR) (**a**), a reaction mixture consisting of an oil–aqueous emulsion is created to encapsulate bead–DNA complexes into single aqueous droplets. PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. EmPCR beads can be chemically attached to a glass slide or deposited into PicoTiterPlate wells (FIG. 3c). Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer (**c**); immobilization by a template (**d**); and immobilization of a polymerase (**e**). dNTP, 2′-deoxyribonucleoside triphosphate.

Mate-pair templates
A genomic library is prepared by circularizing sheared DNA that has been selected for a given size, such as 2 kb, therefore bringing the ends that were previously distant from one another into close proximity. Cutting these circles into linear DNA fragments creates mate-pair templates.

are attached to the solid support[15], to which a primed template molecule is bound (FIG. 1e). This approach is used by Pacific Biosciences[15] and is described in patents from Life/VisiGen[16] and LI-COR Biosciences[17]. Larger DNA molecules (up to tens of thousands of base pairs) can be used with this technique and, unlike the first two approaches, the third approach can be used with real-time methods, resulting in potentially longer read lengths.

Sequencing and imaging

There are fundamental differences in sequencing clonally amplified and single-molecule templates. Clonal amplification results in a population of identical templates, each of which has undergone the sequencing reaction. Upon imaging, the observed signal is a consensus of the nucleotides or probes added to the identical templates for a given cycle. This places a greater

demand on the efficiency of the addition process, and incomplete extension of the template ensemble results in lagging-strand dephasing. The addition of multiple nucleotides or probes can also occur in a given cycle, resulting in leading-strand dephasing. Signal dephasing increases fluorescence noise, causing base-calling errors and shorter reads[18]. Because dephasing is not an issue with single-molecule templates, the requirement for cycle efficiency is relaxed. Single molecules, however, are susceptible to multiple nucleotide or probe additions in any given cycle. Here, deletion errors will occur owing to quenching effects between adjacent dye molecules or no signal will be detected because of the incorporation of dark nucleotides or probes. In the following sections, sequencing and imaging strategies that use both clonally amplified and single-molecule templates are discussed.

Cyclic reversible termination. As the name implies, CRT uses reversible terminators in a cyclic method that comprises nucleotide incorporation, fluorescence imaging and cleavage[2]. In the first step, a DNA polymerase, bound to the primed template, adds or incorporates just one fluorescently modified nucleotide (BOX 1), which represents the complement of the template base. The termination of DNA synthesis after the addition of a single nucleotide is an important feature of CRT. Following incorporation, the remaining unincorporated nucleotides are washed away. Imaging is then performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye. Additional washing is performed before starting the next incorporation step. FIG. 2a depicts a four-colour CRT cycle used by Illumina/Solexa, and FIG. 2c illustrates a one-colour CRT cycle used by Helicos BioSciences.

The key to the CRT method is the reversible terminator, of which there are two types: 3′ blocked and 3′ unblocked (BOX 1). The use of a dideoxynucleotide, which acts as a chain terminator in Sanger sequencing, provided the basis for the initial development of reversible blocking groups attached to the 3′ end of nucleotides[19,20]. Blocking groups, such as 3′-*O*-allyl-2′-deoxyribonucleoside triphosphates (dNTPs)[21] and 3′-*O*-azidomethyl-dNTPs[22], have been successfully used in CRT. 3′-blocked terminators require the cleavage of two chemical bonds to remove the fluorophore from the nucleobase and restore the 3′-OH group.

Currently, the Illumina/Solexa Genome Analyzer (GA)[23] dominates the NGS market. It uses the clonally amplified template method illustrated in FIG. 1b, coupled with the four-colour CRT method illustrated in FIG. 2a. The four colours are detected by total internal reflection fluorescence (TIRF) imaging using two lasers, the output of which is depicted in FIG. 2b. The slide is partitioned into eight channels, which allows independent samples to be run simultaneously. TABLE 1 shows the current sequencing statistics of the Illumina/Solexa GA_{II} platform operating at the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC; D. Muzny, personal communication). Substitutions are the most common error type, with a higher portion of errors occurring when the previous incorporated nucleotide is a 'G' base[24]. Genome analysis of Illumina/Solexa data has revealed an underrepresentation of AT-rich[24–26] and GC-rich regions[25,26], which is probably due to amplification bias during template preparation[25]. Sequence variants are called by aligning reads to a reference genome using bioinformatics tools such as MAQ[27] or ELAND[23]. Bentley and colleagues reported high concordance (>99.5%) of single-nucleotide variant (SNV)[28] calls with standard genotyping arrays using both alignment tools, and a false-positive rate of 2.5% with novel SNVs[23]. Other reports have described a higher false-positive rate associated with novel SNV detection using these alignment tools[29,30].

The difficulty involved in identifying a modified enzyme that efficiently incorporates 3′-blocked terminators — a process that entails screening large libraries of mutant DNA polymerases — has spurred the development of 3′-unblocked reversible terminators. LaserGen, Inc. was the first group to show that a small terminating group attached to the base of a 3′-unblocked nucleotide can act as an effective reversible terminator and be efficiently incorporated by wild-type DNA polymerases[31]. This led to the development of Lightning Terminators[32] (BOX 1). Helicos BioSciences has reported the development of Virtual Terminators, which are 3′-unblocked terminators with a second nucleoside analogue that acts as an inhibitor[33]. The challenge for 3′-unblocked terminators is creating the appropriate modifications to the terminating (Lightning Terminators)[32] or inhibiting (Virtual Terminators)[33] groups so that DNA synthesis is terminated after a single base addition. This is important because an unblocked 3′-OH group is the natural substrate for incorporating the next incoming nucleotide. Cleavage of only a single bond is required to remove both the terminating or inhibiting group and the fluorophore group from the nucleobase, which is a more efficient strategy than 3′-blocked terminators for restoring the nucleotide for the next CRT cycle.

Helicos BioSciences was the first group to commercialize a single-molecule sequencer, the HeliScope, which was based on the work of Quake and colleagues[34]. The HeliScope uses the single-molecule template methods shown in FIG. 1c and FIG. 1d coupled with the one-colour (Cy5 dye) CRT method shown in FIG. 2c. Incorporation of a nucleotide results in a fluorescent signal. The HeliScope also uses TIRF to image the Cy5 dye[34], the imaging output of which is shown in FIG. 2d. Harris and colleagues[14] used Cy5-12ss-dNTPs, which are earlier versions of their Virtual Terminators that lack the inhibiting group, and reported that deletion errors in homopolymeric repeat regions were the most common error type (~5% frequency) when using the primer-immobilized strategy shown in FIG. 1c. This is likely to be related to the incorporation of two or more Cy5-12ss-dNTPs in a given cycle. These errors can be greatly reduced with two-pass sequencing, which provides ~25-base consensus reads using the template-immobilized strategy shown in FIG. 1d. At the 2009 Advances in Genome Biology and Technology (AGBT) meeting, the Helicos group reported their recent progress in sequencing the

Dephasing
This occurs with step-wise addition methods when growing primers move out of synchronicity for any given cycle. Lagging strands (for example, $n - 1$ from the expected cycle) result from incomplete extension, and leading strands (for example, $n + 1$) result from the addition of multiple nucleotides or probes in a population of identical templates.

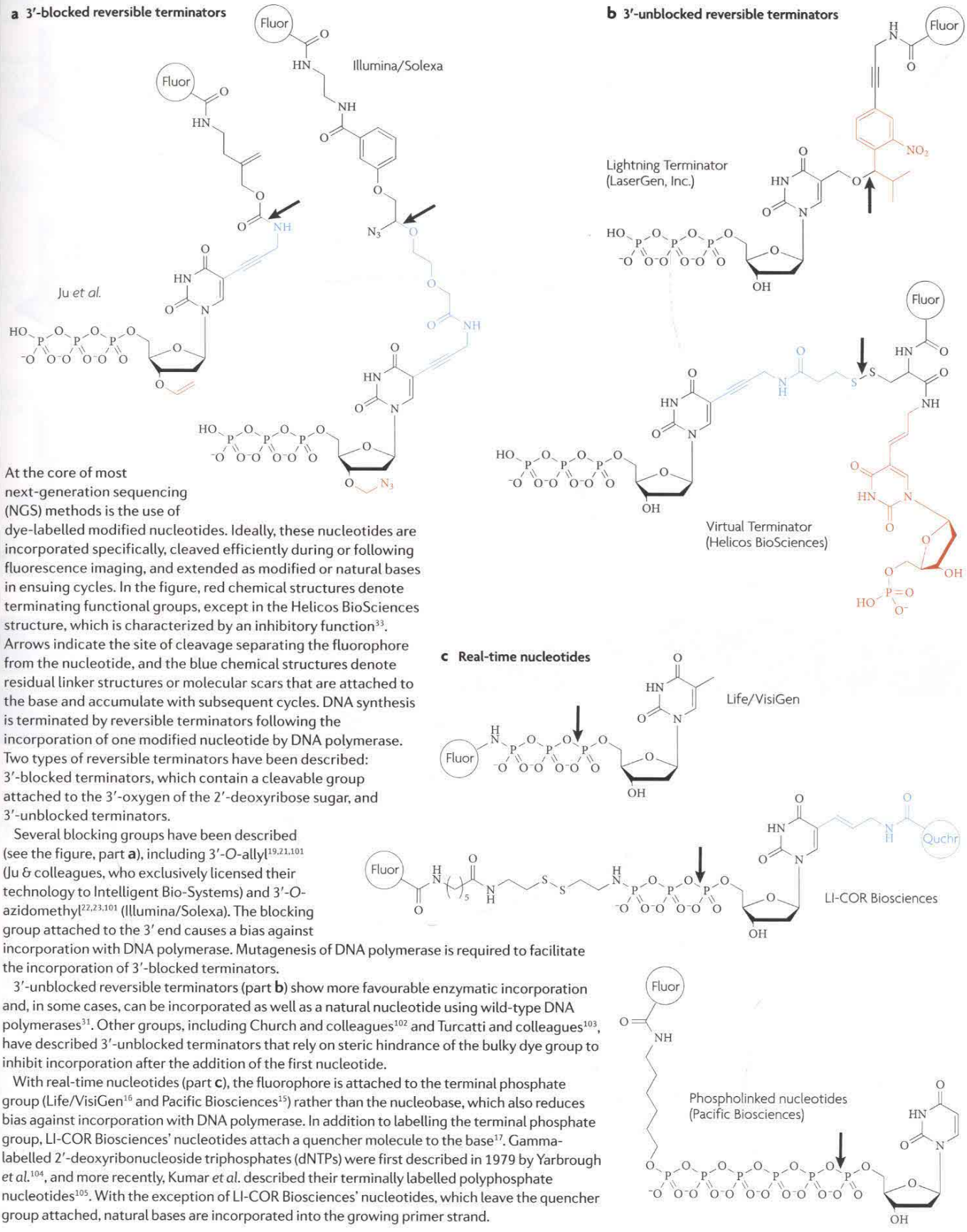
Dark nucleotides or probes
A nucleotide or probe that does not contain a fluorescent label. It can be generated from its cleavage and carry-over from the previous cycle or be hydrolysed *in situ* from its dye-labelled counterpart in the current cycle.

Total internal reflection fluorescence
A total internal reflection fluorescence imaging device produces an evanescent wave — that is, a near-field stationary excitation wave with an intensity that decreases exponentially away from the surface. This wave propagates across a boundary surface, such as a glass slide, resulting in the excitation of fluorescent molecules near (< 200 nm) or at the surface and the subsequent collection of their emission signals by a detector.

Libraries of mutant DNA polymerases
Large numbers of genetically engineered DNA polymerases can be created by either site-directed or random mutagenesis, which leads to one or more amino acid substitutions, insertions and/or deletions in the polymerase. The goal of this approach is to incorporate modified nucleotides more efficiently during the sequencing reaction.

Consensus reads
These are only useful for single-molecule techniques and are produced by sequencing the same template molecule more than once. The data are then aligned to produce a 'consensus read', reducing stochastic errors that may occur in a given sequence read.

Box 1 | Modified nucleotides used in next-generation sequencing methods



At the core of most next-generation sequencing (NGS) methods is the use of dye-labelled modified nucleotides. Ideally, these nucleotides are incorporated specifically, cleaved efficiently during or following fluorescence imaging, and extended as modified or natural bases in ensuing cycles. In the figure, red chemical structures denote terminating functional groups, except in the Helicos BioSciences structure, which is characterized by an inhibitory function[33]. Arrows indicate the site of cleavage separating the fluorophore from the nucleotide, and the blue chemical structures denote residual linker structures or molecular scars that are attached to the base and accumulate with subsequent cycles. DNA synthesis is terminated by reversible terminators following the incorporation of one modified nucleotide by DNA polymerase. Two types of reversible terminators have been described: 3′-blocked terminators, which contain a cleavable group attached to the 3′-oxygen of the 2′-deoxyribose sugar, and 3′-unblocked terminators.

Several blocking groups have been described (see the figure, part **a**), including 3′-*O*-allyl[19,21,101] (Ju & colleagues, who exclusively licensed their technology to Intelligent Bio-Systems) and 3′-*O*-azidomethyl[22,23,101] (Illumina/Solexa). The blocking group attached to the 3′ end causes a bias against incorporation with DNA polymerase. Mutagenesis of DNA polymerase is required to facilitate the incorporation of 3′-blocked terminators.

3′-unblocked reversible terminators (part **b**) show more favourable enzymatic incorporation and, in some cases, can be incorporated as well as a natural nucleotide using wild-type DNA polymerases[31]. Other groups, including Church and colleagues[102] and Turcatti and colleagues[103], have described 3′-unblocked terminators that rely on steric hindrance of the bulky dye group to inhibit incorporation after the addition of the first nucleotide.

With real-time nucleotides (part **c**), the fluorophore is attached to the terminal phosphate group (Life/VisiGen[16] and Pacific Biosciences[15]) rather than the nucleobase, which also reduces bias against incorporation with DNA polymerase. In addition to labelling the terminal phosphate group, LI-COR Biosciences' nucleotides attach a quencher molecule to the base[17]. Gamma-labelled 2′-deoxyribonucleoside triphosphates (dNTPs) were first described in 1979 by Yarbrough *et al.*[104], and more recently, Kumar *et al.* described their terminally labelled polyphosphate nucleotides[105]. With the exception of LI-COR Biosciences' nucleotides, which leave the quencher group attached, natural bases are incorporated into the growing primer strand.

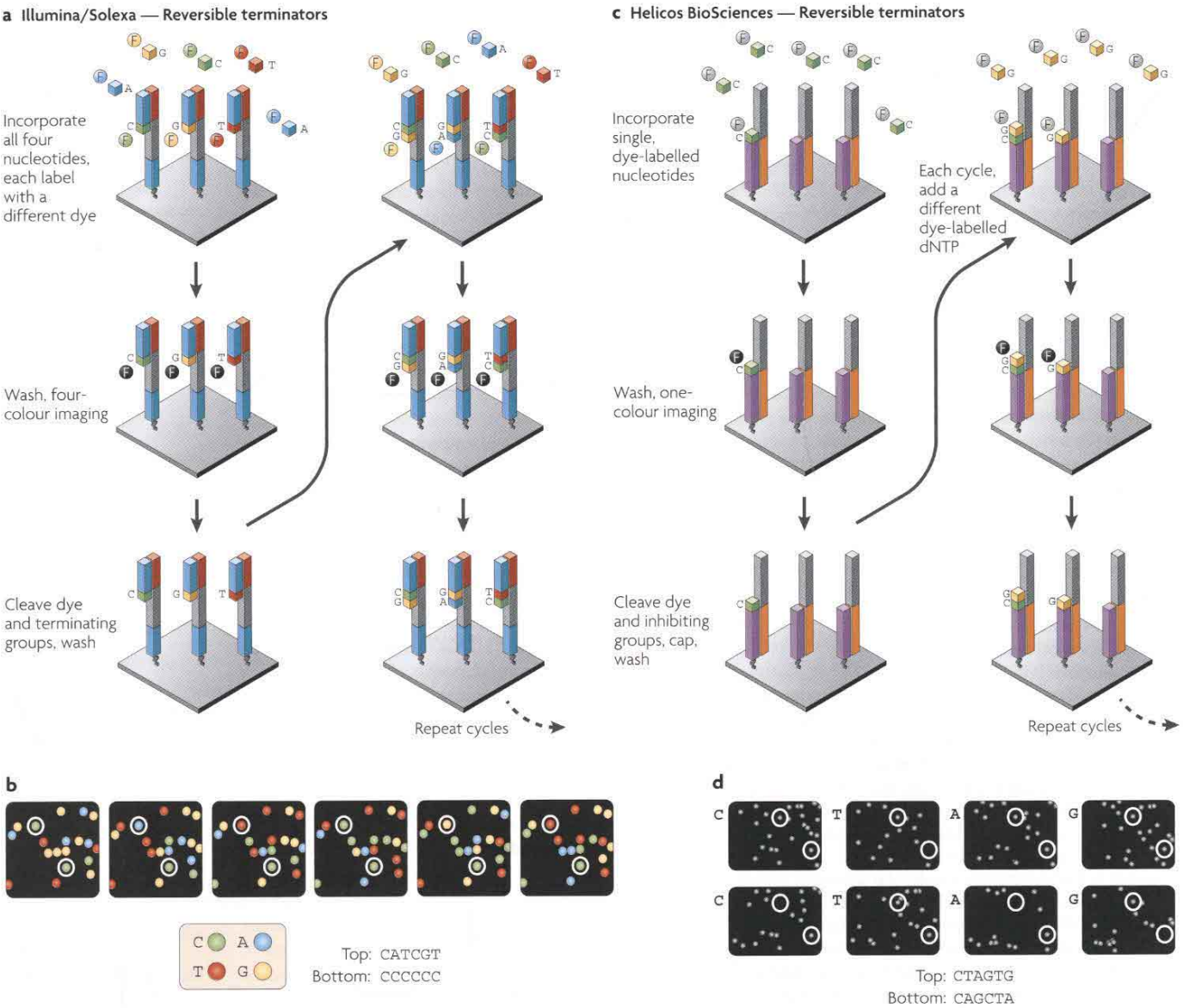


Figure 2 | **Four-colour and one-colour cyclic reversible termination methods. a** | The four-colour cyclic reversible termination (CRT) method uses Illumina/Solexa's 3′-*O*-azidomethyl reversible terminator chemistry[23,101] (BOX 1) using solid-phase-amplified template clusters (FIG. 1b, shown as single templates for illustrative purposes). Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3′-OH group using the reducing agent tris(2-carboxyethyl)phosphine (TCEP)[23]. **b** | The four-colour images highlight the sequencing data from two clonally amplified templates. **c** | Unlike Illumina/Solexa's terminators, the Helicos Virtual Terminators[33] are labelled with the same dye and dispensed individually in a predetermined order, analogous to a single-nucleotide addition method. Following total internal reflection fluorescence imaging, a cleavage step removes the fluorescent dye and inhibitory groups using TCEP to permit the addition of the next Cy5-2′-deoxyribonucleoside triphosphate (dNTP) analogue. The free sulphhydryl groups are then capped with iodoacetamide before the next nucleotide addition[33] (step not shown). **d** | The one-colour images highlight the sequencing data from two single-molecule templates.

One-base-encoded probe
An oligonucleotide sequence in which one interrogation base is associated with a particular dye (for example, A in the first position corresponds to a green dye). An example of a one-base degenerate probe set is '1-probes', which indicates that the first nucleotide is the interrogation base. The remaining bases consist of either degenerate (four possible bases) or universal bases.

Caenorhabditis elegans genome. From a single HeliScope run using only 7 of the instrument's 50 channels, approximately 2.8 Gb of high-quality data were generated in 8 days from >25-base consensus reads with 0, 1 or 2 errors. Greater than 99% coverage of the genome was reported, and for regions that showed >5-fold coverage, the consensus accuracy was 99.999% (J. W. Efcavitch, personal communication).

Sequencing by ligation. SBL is another cyclic method that differs from CRT in its use of DNA ligase[35] and either one-base-encoded probes or two-base-encoded probes. In its simplest form, a fluorescently labelled probe hybridizes to its complementary sequence adjacent to the primed template. DNA ligase is then added to join the dye-labelled probe to the primer. Non-ligated probes are washed away, followed by fluorescence

Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

imaging to determine the identity of the ligated probe[36]. The cycle can be repeated either by using cleavable probes to remove the fluorescent dye and regenerate a 5′-PO_4 group for subsequent ligation cycles (FIG. 3a) or by removing and hybridizing a new primer to the template (not shown in the figure).

Shendure and colleagues[11] used the SBL method to sequence the *Escherichia coli* MG1655 genome. Mate-pair templates were prepared with four priming sites (named A1 to A4) and were amplified by emPCR (FIG. 1a). A number of one-base-encoded probes, 1-probes to 7-probes, were used. In the first SBL cycle, the A1 primer was annealed to the template, followed by the hybridization and ligation of 1-probes, four-colour imaging and removal of the entire primer–probe strand from the solid-phase-bound template. The SBL cycle was then repeated with the A1 primer but using 2-probes, and then with the A1 primer but using 3-probes, and so on. The other three primers, A2, A3 and A4, were then cycled in an analogous manner to yield six (A2 and A4) and seven (A1 and A3) base reads (albeit discontiguous) from each genomic end, making a total of 26 base reads per mate-pair template. From two instrument runs, the authors reported the production of approximately 48 million high-quality bases, which mapped to about 70% of the *E. coli* genome[11]. This SBL method is being used on the Polonator instrument.

Life/APG has commercialized their SBL platform called support oligonucleotide ligation detection (SOLiD)[37]. The method uses two-base-encoded probes, which has the primary advantage of improved accuracy in colour calling and SNV calling, the latter of which requires an adjacent valid colour change. Colour space is a unique feature of the SOLiD system. A primer is

Two-base-encoded probe
An oligonucleotide sequence in which two interrogation bases are associated with a particular dye (for example, AA, CC, GG and TT are coded with a blue dye). '1,2-probes' indicates that the first and second nucleotides are the interrogation bases. The remaining bases consist of either degenerate or universal bases.

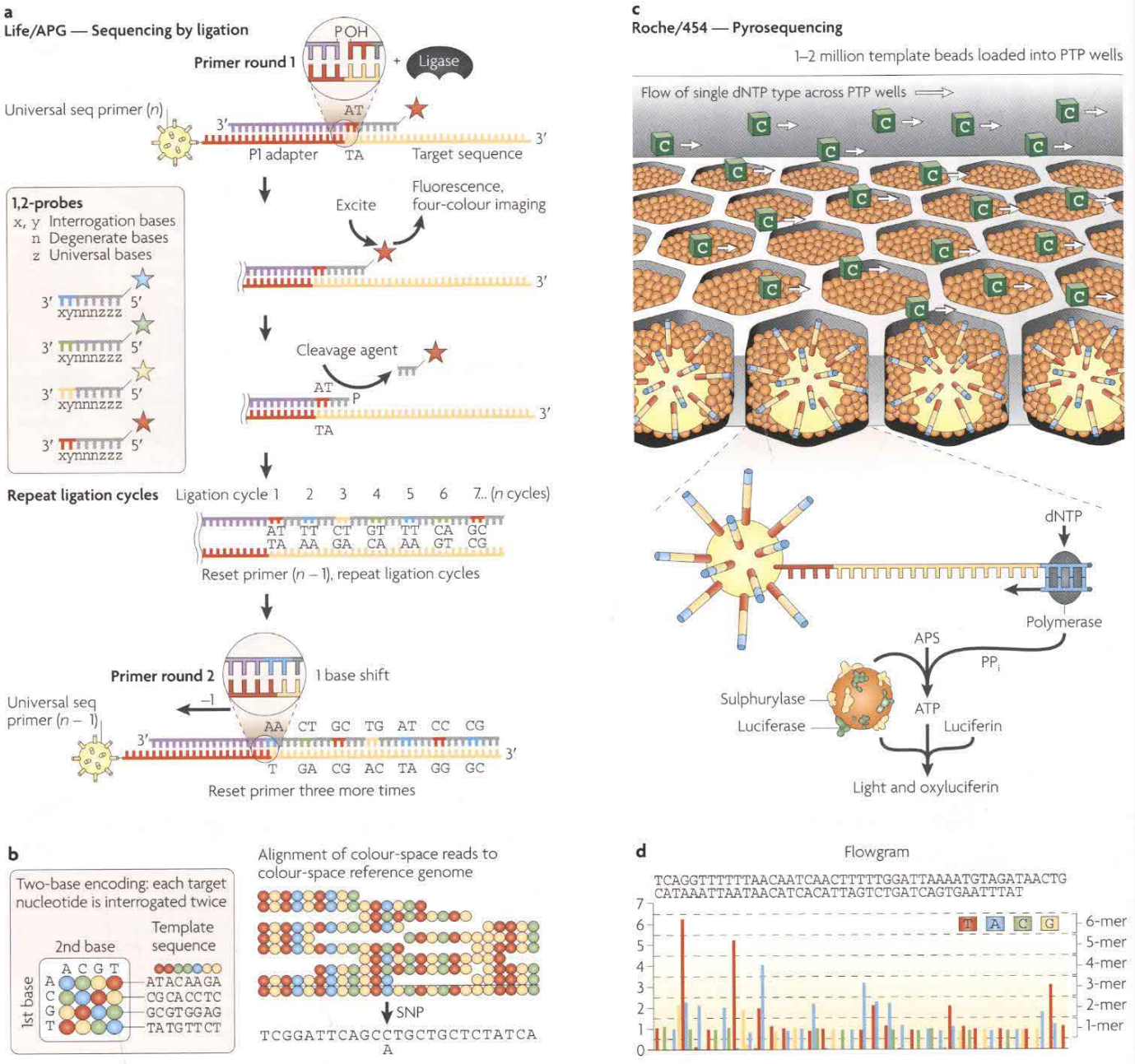


Figure 3 | **Next-generation sequencing technologies that use emulsion PCR. a** | A four-colour sequencing by ligation method using Life/APG's support oligonucleotide ligation detection (SOLiD) platform is shown. Upon the annealing of a universal primer, a library of 1,2-probes is added. Unlike polymerization, the ligation of a probe to the primer can be performed bi-directionally from either its 5′-PO_4 or 3′-OH end. Appropriate conditions enable the selective hybridization and ligation of probes to complementary positions. Following four-colour imaging, the ligated 1,2-probes are chemically cleaved with silver ions to generate a 5′-PO_4 group. The SOLiD cycle is repeated nine more times. The extended primer is then stripped and four more ligation rounds are performed, each with ten ligation cycles. The 1,2-probes are designed to interrogate the first (x) and second (y) positions adjacent to the hybridized primer, such that the 16 dinucleotides are encoded by four dyes (coloured stars). The probes also contain inosine bases (z) to reduce the complexity of the 1,2-probe library and a phosphorothiolate linkage between the fifth and six nucleotides of the probe sequence, which is cleaved with silver ions[106]. Other cleavable probe designs include RNA nucleotides[107,108] and internucleosidic phosphoramidates[107], which are cleaved by ribonucleases and acid, respectively. **b** | A two-base encoding scheme in which four dinucleotide sequences are associated with one colour (for example, AA, CC, GG and TT are coded with a blue dye). Each template base is interrogated twice and compiled into a string of colour-space data bits. The colour-space reads are aligned to a colour-space reference sequence to decode the DNA sequence. **c** | Pyrosequencing using Roche/454's Titanium platform. Following loading of the DNA-amplified beads into individual PicoTiterPlate (PTP) wells, additional beads, coupled with sulphurylase and luciferase, are added. In this example, a single type of 2′-deoxyribonucleoside triphosphate (dNTP) — cytosine — is shown flowing across the PTP wells. The fibre-optic slide is mounted in a flow chamber, enabling the delivery of sequencing reagents to the bead-packed wells. The underneath of the fibre-optic slide is directly attached to a high-resolution charge-coupled device (CCD) camera, which allows detection of the light generated from each PTP well undergoing the pyrosequencing reaction. **d** | The light generated by the enzymatic cascade is recorded as a series of peaks called a flowgram. PP_i, inorganic pyrophosphate.

hybridized to templates amplified by emPCR (FIG. 1a). The SOLiD cycle of 1,2-probe hybridization and ligation, imaging, and probe cleavage is repeated ten times to yield ten colour calls spaced in five-base intervals (FIG. 3a). The extended primer is then stripped from the solid-phase-bound templates. A second ligation round is performed with an 'n – 1' primer, which resets the interrogation bases and the corresponding ten colour calls one position to the left. Ten ligation cycles ensue, followed by three more rounds of ligation cycles. Colour calls from the five ligation rounds are then ordered into a linear sequence (that is, the colour space) and aligned to a reference genome to decode the DNA sequence (FIG. 3b). SOLiD uses two slides per run; each can be partitioned into four or eight regions called spots. TABLE 1 depicts the current sequencing statistics of the Life/APG platform operating at the BCM-HGSC (D. Muzny, personal communication). Substitutions are the most common error type. Similar to the genome analysis of Illumina/Solexa reads, SOLiD data have also revealed an underrepresentation of AT-rich and GC-rich regions[26]. Shen and colleagues[38] recently showed that MAQ alignment of SOLiD data may be undercalling true variants.

Single-nucleotide addition: pyrosequencing. Pyrosequencing is a non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions[39,40] (FIG. 3c). Unlike other sequencing approaches that use modified nucleotides to terminate DNA synthesis, the pyrosequencing method manipulates DNA polymerase by the single addition of a dNTP in limiting amounts. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses. DNA synthesis is reinitiated following the addition of the next complementary dNTP in the dispensing cycle. The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA sequence (FIG. 3d).

Margulies and colleagues[41] described the first NGS platform to integrate pyrosequencing using their PTP device. Commercialized by Roche/454, the instrument uses DNA templates prepared by emPCR (FIG. 1a), with 1–2 million beads deposited into PTP wells. Roche/454 recently released a titanium-coated PTP design, which substantially increases read length and improves data quality by reducing crosstalk between adjacent wells containing single clonally amplified beads (T. Harkins, personal communication). Smaller beads, which have sulphurylase and luciferase attached to them to facilitate light production, are loaded into wells surrounding the template beads. Individual dNTPs are then streamed across the wells and dispensed in a predetermined sequential order. The bioluminescence is imaged with a charge-coupled device (CCD) camera (FIG. 3c). TABLE 1 shows the current sequencing statistics of the Roche/454 platform operating at the BCM-HGSC (D. Muzny, personal communication). Unlike platforms that produce shorter read lengths, the Roche/454 platform does not require the run time to be doubled for the sequencing of mate-pair templates. For homopolymeric repeats of up to six nucleotides, the number of dNTPs added is directly proportional to the light signal. Insertions are the most common error type, followed by deletions.

Real-time sequencing. The next technology method to hit the commercial sector is likely to be real-time sequencing, and Pacific Biosciences is currently leading this effort[15]. Unlike reversible terminators, real-time nucleotides do not halt the process of DNA synthesis. Simply put, the method of real-time sequencing involves imaging the continuous incorporation of dye-labelled nucleotides during DNA synthesis[42]. With the Pacific Biosciences platform, single DNA polymerase molecules are attached to the bottom surface of individual zero-mode waveguide detectors (ZMW detectors)[43] (FIG. 4a) that can obtain sequence information while phospholinked nucleotides (BOX 1) are being incorporated into the growing primer strand[15] (FIG. 4b).

Other approaches have been proposed to enhance signal-to-noise measurements in real-time sequencing using more conventional detection schemes. For example, Life/VisiGen has engineered DNA polymerases with an attached fluorescent dye that, upon incorporation of their γ-labelled nucleotides, produce an enhanced signal by fluorescence resonance energy transfer[16]. LI-COR Biosciences, whose technology was acquired by Pacific Biosciences, has been developing dye-quencher nucleotides (BOX 1), which in their native state produce low signals owing to the presence of a quencher group attached to the base. The release and diffusion of the dye-labelled pyrophosphate analogue away from the immobilized DNA polymerase produces a fluorescent signal[17].

Pacific Biosciences used the highly processive, strand-displacing φ29 DNA polymerase because it efficiently incorporates phospholinked nucleotides and enables the resequencing of closed circular templates[15]. To assess the accuracy of this method, a four-colour sequencing experiment was conducted using a known 150 bp linear template. Base calls from the real-time reads were determined from their corresponding fluorescence pulses (FIG. 4b). When the reads were compared to a known sequence, 27 errors consisting of deletions, insertions and mismatches were identified, corresponding to a read accuracy of approximately 83% (131/158). Factors that led to sequencing errors included extremely short interphase intervals between two incorporation events, and the binding and release of nucleotides in the active site before incorporation into the primer strand. Given that most errors appear as stochastic events, the authors showed that repeated sequencing of the same template molecule 15 times or more could improve the consensus read accuracy to >99%[15]. At the 2009 AGBT meeting, Pacific Biosciences reported improvements to their platform; when it was used to sequence the *E. coli* genome at 38-fold base coverage, 99.3% genome coverage was obtained. The consensus accuracy reached was >99.999% for the entire genome, with read lengths averaging 964 bases (S. Turner, personal communication).

Adjacent valid colour
A nucleotide substitution will have two colour calls, one from the 5′ position and one from the 3′ position of the dinucleotide sequence. When compared with a reference genome, base substitution in the target sequence is encoded by two specific, adjacent colours. In Figure 3b, the sequence 'CCT' is encoded as blue-yellow ('CC' = blue; 'CT' = yellow), but substituting the middle 'C' for 'A' would result in two colour changes to green-red. Any other colour sequence can be discarded as an error.

Colour space
With two-base-encoded probes, the fluorescent signal or colour obtained during imaging is associated with four dinucleotide sequences having a 5′- and 3′-base. Colour space is the sequence of overlapping dinucleotides that codes four simultaneous nucleotide sequences. Alignment with a reference genome is the most accurate method for translating colour space into a single nucleotide sequence.

Zero-mode waveguide detectors
This nanostructure device is 100 nm in diameter, which is smaller than the 532 nm and 643 nm laser wavelengths used in the Pacific Biosciences platform. Light cannot propagate through these small waveguides, hence the term zero-mode. These aluminium-clad waveguides are designed to produce an evanescent wave (see the 'total internal reflection fluorescence' glossary term) that substantially reduces the observation volume at the surface of the polymerase reaction down to the zeptolitre range (10^{-21} l). This provides an advantage for the polymerization reaction, which can be performed at higher dye-labelled nucleotide concentrations.

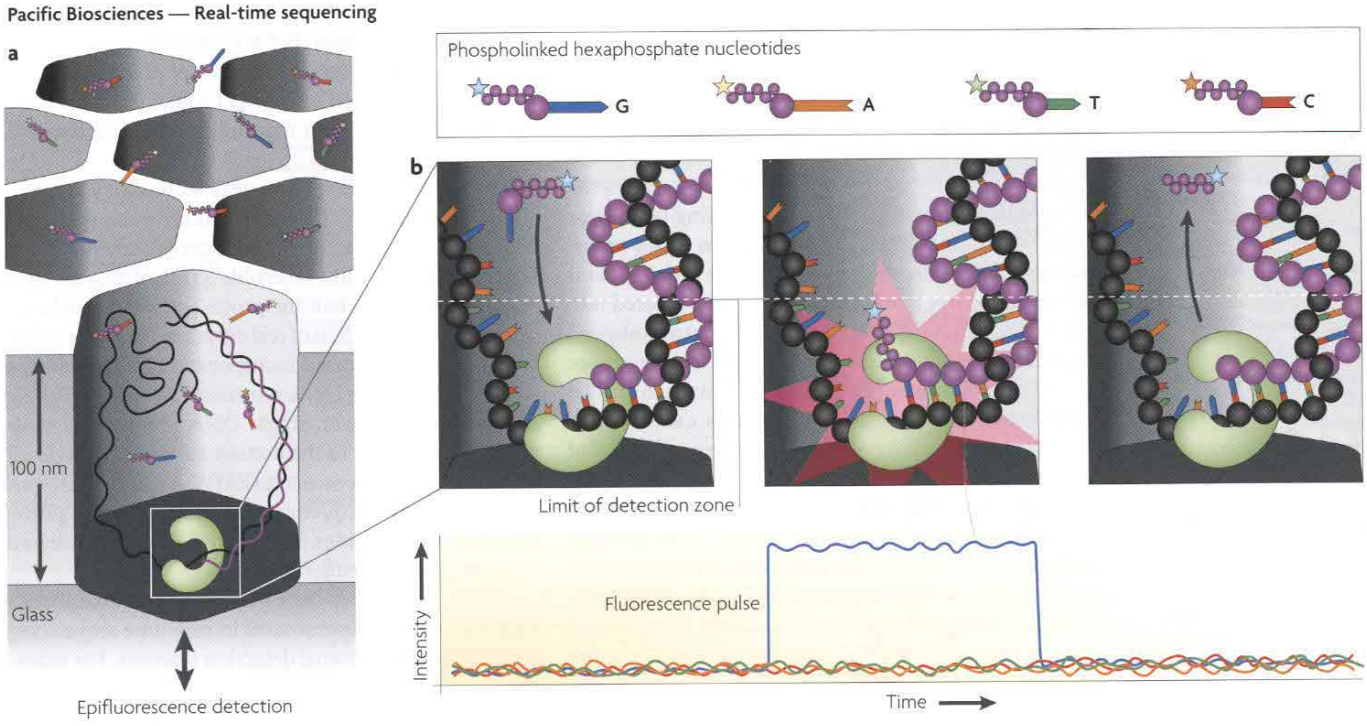


Figure 4 | **Real-time sequencing.** Pacific Biosciences' four-colour real-time sequencing method is shown. **a** | The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter the detection layer for a given period. These ZMW detectors address the dilemma that DNA polymerases perform optimally when fluorescently labelled nucleotides are present in the micromolar concentration range, whereas most single-molecule detection methods perform optimally when fluorescent species are in the pico- to nanomolar concentration range[42]. **b** | The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis and is usually on the millisecond scale. This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone on this timescale. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide.

Genome alignment and assembly

After NGS reads have been generated, they are aligned to a known reference sequence or assembled *de novo*[8,44,45]. The decision to use either strategy is based on the intended biological application as well as cost, effort and time considerations. For example, identifying and cataloguing genetic variation in multiple strains of highly related genomes, such as those found in specific species of bacteria[46–51], *C. elegans*[25,30,38] and *Arabidopsis thaliana*[52], can be accomplished by aligning NGS reads to their reference genomes. This approach is substantially cheaper and faster than Sanger sequencing. SNVs can be readily identified, although in many cases, validation of these findings is required.

There are limitations to the alignment approach, such as placing reads within repetitive regions in the reference genome or in corresponding regions that may not exist in the reference genome[28]; the latter situation may result from gaps in the reference genome or the presence of structural variants (SVs) in the genome being analysed. Mate-pair reads can resolve the correct genome assignment for some repetitive regions as long as one read in the pair is unique to the genome. A study by Egholm, Snyder and colleagues[53] showed that Roche/454 read data (averaging 258 bases), derived from 3-kb sized mate-pair genomic libraries, could capture a larger fraction of SVs in the human genome, although this approach still identified fewer SVs than traditional fosmid-end sequencing approaches[54].

De novo assemblies have been reported for bacterial genomes and mammalian bacterial artificial chromosomes[55–59], but substantial challenges exist for their application to human genomes. A reasonable strategy for improving the quality of the alignment or assembly has been to increase the read coverage. An article by Frazer and colleagues[26] challenges this approach by reporting systematic variability in local sequence coverage that was specific to different human genomic regions for the Roche/454, Illumina/Solexa and Life/APG platforms (other commercially available platforms were not evaluated). Because each NGS platform examined produced a uniquely reproducible pattern of variable sequence coverage, the mixing of different NGS read types in the alignment or assembly may remedy this shortcoming. Recently it was reported that mixing Roche/454 and Illumina/Solexa read data resulted in improved *de novo* assemblies of microbial genomes compared with assemblies based on data from either platform alone[60,61].

Fluorescence resonance energy transfer
This is generally a system that consists of two fluorescent dyes, one being a donor dye (a bluer fluorophore) and the other an acceptor dye (a redder fluorophore). When the two dye molecules are brought into close proximity (usually ≤30 nm), the energy from the excited donor dye is transferred to the acceptor dye, increasing its emission intensity signal.

Structural variants
All sequence variants other than single-nucleotide variants, including block substitutions, insertions or deletions, inversions, segmental duplications and copy-number differences.

Genome enrichment

Despite the substantial cost reductions associated with NGS technologies in comparison with the automated Sanger method[62], whole-genome sequencing is still an expensive endeavour (see below). An interim solution to this problem may be to use NGS platforms to target specific regions of interest. This strategy can be used to examine all of the exons in the genome, specific gene families that constitute known drug targets or megabase-size regions that are implicated in disease[63] or pharmacogenetics effects[64] through genome-wide association studies. The concept of targeting specific regions of the genome is well established, with PCR being the most widely used method, albeit on a small scale. PCR coupled with Sanger sequencing is suitably matched for analysing a handful of candidate genes[65], but coupling PCR with high-throughput NGS platforms for targeting strategies is not practical, as sample preparation would require handling tens of thousands of primers individually or in large multiplex groups to meet the needs of a single instrument run. A recent article by Frazer and colleagues in collaboration with RainDance Technologies reported the simultaneous amplification of 3,976 products using microdroplet PCR technology[66]. Here, a microfluidic device creates aqueous picolitre-volume droplets of forward- and reverse-targeting primers in an oil solution. The primer droplets that are targeted to different regions of the genome merge with separate picolitre droplets that contain fragmented genomic DNA and associated PCR reagents, and these mixed droplets are thermal cycled in a single tube (FIG. 5a). The authors reported an 84% capture efficiency with 90% of the targeted bases showing uniform coverage using the microdroplet PCR method sequenced with either the Roche/454 or Illumina/Solexa platform[66].

Custom-designed oligonucleotide microarrays and solution-based hybridization strategies have also been used for targeting regions of interest. For example, Roche/NimbleGen[67] oligonucleotide microarrays have been designed to provide complementary sequences for solid-phase hybridization to enrich for exons[68,69] or contiguous stretches of genomic regions[68,70] (FIG. 5b). The BCM-HGSC[68] and Cold Spring Harbor Laboratory[69] groups reported capture efficiencies of 65–77% (Roche/454 platform) and 53% (Illumina/Solexa platform), respectively, with targeted exons being covered by at least one NGS read. Recently, the BCM-HGSC group reported >90% capture efficiency with at least 10× base coverage (Roche/454 and Life/APG platforms) by chip optimization that decreases the number of probes in higher coverage regions while increasing those in low coverage regions (R. A. Gibbs, personal communication).

Other groups have captured specific genomic regions by using solution-based hybridization methods, such as molecular inversion probes (MIPs)[71–73] and biotinylated RNA capture sequences[74] (FIG. 5c). Shendure, Church and colleagues[71] were the first group to use MIPs in exon targeting by designing specific oligonucleotide ends that flank the exons of interest. However, in duplicate experiments only 20% of the targets were captured and even fewer exon regions (11%) were found in both data sets (Illumina/Solexa platform)[71]. Mindrinos, Davis and colleagues[72] recently described several technical improvements that increased the efficiency of MIP capture to >90%, with approximately 70% of targets falling within a 10-fold range in 485 exons. Shendure and colleagues[73] have also described improvements that enabled them to capture 91% of their initial 55,000 targets (Illumina/Solexa platform). An alternative method has been reported by Nusbaum and colleagues[74], who created biotinylated RNA capture sequences that are hybridized to genomic targets and subsequently enriched with streptavidin-coated magnetic beads (Illumina/Solexa platform) (FIG. 5c). This group estimated a capture efficiency of 60% for exons and 80% for genomic regions. The 1000 Genomes Project and The Exome Project have adopted microarray-based and solution-based strategies that use Roche/454, Illumina/Solexa and Life/APG platforms to target regions of interest.

Although any of the NGS platforms can perform the sequencing phase of targeted capture, some may be better suited than others (TABLE 1). Initial reports raised concerns as to the readiness of targeted capture for routine use[75,76] owing to the high degree of variability in coverage and representation, the capturing of off-target fragments, the requirement for microgram quantities of starting material and the high cost for microarrays, but improvements are being made to address these limitations[76].

Next-generation sequencing applications

The production of large numbers of low-cost reads makes the NGS platforms described above useful for many applications. These include variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assemblies of bacterial and lower eukaryotic genomes, cataloguing the transcriptomes of cells, tissues and organisms (RNA–seq)[5], genome-wide profiling of epigenetic marks and chromatin structure using other seq-based methods (ChIP–seq, methyl–seq and DNase–seq)[4], and species classification and/or gene discovery by metagenomics studies[77]. With this many applications, which platform is best suited for a given biological experiment? For example, the Illumina/Solexa and Life/APG platforms are well suited for variant discovery by resequencing human genomes because gigantic volumes of high-quality bases are produced per run (TABLE 1). Furthermore, the Helicos BioSciences platform is well suited for applications that demand quantitative information in RNA–seq[78] or direct RNA sequencing, as it sequences RNA templates directly without the need to convert them into cDNAs[79]. TABLE 1 provides an overview of NGS technologies, instrument performance and cost, pros and cons, and recommendations for biological applications; however, the rapid pace of technological advances in the field could change this information in the near future. Readers are directed to several excellent reviews on RNA–seq[5], ChIP–seq[80] and metagenomics[77].

In the following section, I highlight recent advances in sequencing personal genomes using NGS, as demand in this area is driving rapid technological developments and competitive pricing.

1000 Genomes Project
A project aimed at discovering rare sequence variants with minor allele frequencies of 1% in normal genomes derived from HapMap samples.

The Exome Project
A project aimed at developing and validating cost-effective, high-throughput technologies for resequencing all of the protein-coding regions of the human genome.

Metagenomics
The study of communities of mixed microbial genomes that reside in animals, plants and environmental niches. Samples are collected and analysed without the need to culture isolated microbes in the laboratory. The Human Microbiome Project aims to characterize a reference set of microbial genomes from different habitats within the human body, including nasal, oral, skin, gastrointestinal and urogenital regions, and to determine how changes in the human microbiome affect health and disease.

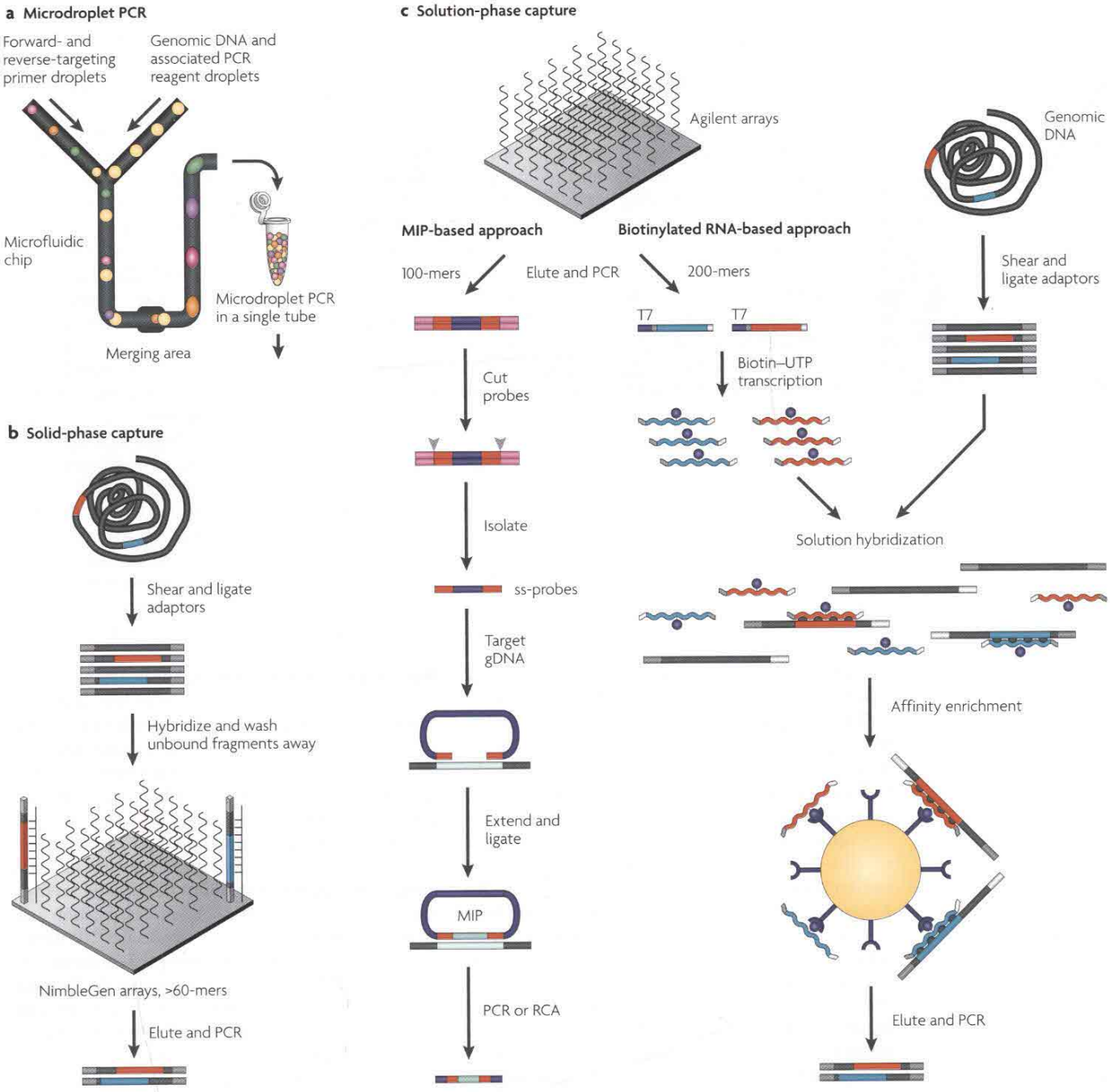


Figure 5 | **Targeted capture scheme. a** | A microfluidic nozzle creates aqueous droplets (8 pl) of forward- and reverse-targeting primers in oil (not shown), which are dispensed into a channel in a microfluidic chip. Fragmented genomic DNA and associated PCR reagents are separately dispensed as aqueous microdroplets (14 pl) and paired together with the primer droplets at a 1/1 ratio. Paired droplets pass through a merging area for coalescence into PCR-competent microdroplets, which are collected into a single tube for thermal cycling. **b** | Solid-phase capture methods ligate adaptor sequences to fragmented genomic DNA before microarray hybridization. Following stringent washing to remove non-selected genomic fragments, the enriched targets are eluted, PCR amplified and sequenced. Adaptor sequences can also be designed specifically for a given next-generation sequencing platform, which eliminates the PCR step. **c** | Solution-phase methods use custom-designed microarrays from Agilent Technologies to synthesize long, target-specific oligonucleotides, which are cleaved as a pool of probes from the support and eluted into a single tube. In the molecular inversion probe (MIP)-based approach, single-stranded probes (ss-probes) are hybridized to their corresponding templates. DNA polymerase extends the 3′-end of the probes across the exon region, and a ligation step creates closed circular DNAs. The biotinylated RNA-based approach introduces a T7 promoter sequence, which allows the incorporation of biotin-labelled uridine-5′-triphosphate (UTP) into the probe sequence by *in vitro* transcription. Following solution hybridization with genomic targets, the RNA–DNA hybrids are enriched by their binding to streptavidin-coated magnetic beads. RCA, rolling circle amplification.

Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ‖‖Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost, E. Mardis, personal communication. ¶¶R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

Personal genomes. Human genome studies aim to catalogue SNVs and SVs and their association to phenotypic differences, with the eventual goal of personalized genomics for medical purposes. In 2004, the International Human Genome Sequencing Consortium published the first, and still only, finished-grade human reference genome (currently National Center for Biotechnology Information (NCBI) build 36)[1]. Its cost was estimated at US$300 million. In October 2007, Venter and colleagues[81] described the genome sequence of J. Craig Venter using a whole-genome shotgun approach coupled with automated Sanger sequencing (TABLE 2). When the Venter genome was compared with the reference genome, 3.2 million SNVs were identified. In addition, there were over 900,000 SVs, which altogether accounted for more variant bases than the SNVs.

The first group to apply NGS to whole human genomes was Roche/454 in collaboration with Gibbs and colleagues at the BCM-HGSC, who reported the diploid genome of James D. Watson[82] (TABLE 2). When the Watson genome was compared with the reference genome, roughly 3.3 million SNVs were identified, but far fewer SVs were found than in the study by Venter and colleagues. This is important because undiscovered SVs could account for a substantial fraction of the total number of sequence variants, many of which could be potentially causative in disease[83–86]. Within the past year, five additional human genomes have been described[23,87–91], one of which was sequenced on two different NGS platforms (TABLE 2). As with the Watson genome, far fewer SVs were reported than in the study by Venter and colleagues.

The Cancer Genome Atlas
A project aimed at discovering single-nucleotide variants and structural variants that are associated with major cancers, such as brain cancer (glioblastoma multiforme), lung cancer (squamous carcinoma) and ovarian cancer (serous cystadenocarcinoma).

Personal Genome Project
A project aimed at providing open access to human genome sequences from volunteers and to develop tools for interpreting this information and correlating it with related personal medical information.

Personal genomics is also being applied to the study of disease. For example, Mardis and colleagues[29,92] have reported the sequencing of two acute myeloid leukaemia cancer genomes using the Illumina/Solexa platform, and both studies identified somatic mutations that may be associated with the disease. Gibbs and colleagues have recently described the elucidation of both allelic variants in a family with a recessive form of Charcot–Marie–Tooth disease using the Life/APG platform[93] (TABLE 2).

Several projects aimed at sequencing more individuals, including The Cancer Genome Atlas[94] and the 1000 Genomes Project, are also using the Illumina/Solexa and Life/454 platforms to sequence whole genomes. Complete Genomics (see below) recently described the first genome sequence from a Caucasian male (PGP1) enrolled in the Personal Genome Project. These projects should lead to a substantial increase in the number of personal genomes sequenced in the near future.

In comparison with automated Sanger sequencing, NGS platforms have dramatically increased throughput and substantially lowered expenditure, with several groups reporting reagent costs of below US$100,000 (TABLE 2). However, there is variability among and within NGS platforms in terms of template size and construct, read-length, throughput, and base and genome coverage (TABLE 2), and such variability makes it difficult to assess the quality (that is, the base accuracy, genome coverage and genome continuity) of genomes based on cost considerations. In June 2009, Illumina announced a personal genome sequencing service that provides 30-fold base coverage for the price of US$48,000. Complete Genomics offers a similar service with 40-fold coverage priced at US$5,000, based on a business model that is reliant on huge customer volume. Recently, Drmanac, Ballinger and colleagues sequenced the whole genomes of three individuals, including PGP1, using a modified non-cleavable probe SBL method known as combinatorial probe anchor ligation (cPAL). The number of SNVs identified was consistent with those reported in TABLE 1 for other genomes, and the group reported an amortized cost of US$4,400 in reagents[95]. Although impressive, it is unclear whether this price will be sustainable as a retail service[96] or whether the Illumina or Complete Genomics business models will see long-term profits. Achieving these cost advantages may also come with a substantial trade-off, namely the production of lower quality genomes that inadequately capture the extent of SVs[83–86]. Closing the gap between $10,000 and $1,000 will be the greatest challenge for current technology developers, and the $1,000 genome might result from as-yet-undeveloped innovations. A timetable for the $1,000 draft genome is difficult to predict, and even more uncertain is the delivery of a high-quality, finished-grade personal genome.

Closing remarks

Since 2004, the National Human Genome Research Institute has awarded more than $100 million for the development of NGS technologies[62], and these awards have facilitated much of the progress in NGS to date, as well as several commercial developments. Many companies, including IBM, Oxford Nanopore Technologies, Intelligent Bio-Systems, LaserGen, Inc. and NABsys, have NGS technologies at various stages of development and commercialization.

The production of billions of NGS reads has also challenged the infrastructure of existing information technology systems in terms of data transfer, storage and quality control, computational analysis to align or assembly read data, and laboratory information management systems for sample tracking and process management. Advances in bioinformatics are ongoing, and improvements are needed if these systems are to keep pace with the continuing developments in NGS technologies. It is possible that the costs associated with downstream data handling and analysis could match or surpass the data-production costs described in this Review.

NGS technologies have an impressive range of applications, and more are being developed. In addition to the applications described above, NGS technologies are being used to characterize the evolutionary relationships of ancient genomes[97,98] and to elucidate the role of non-coding RNAs in health and disease[99,100]. In the not too distant future, it is foreseeable that NGS technologies could be used to obtain high-quality sequence data from a genome isolated from a single cell, which would be a substantial breakthrough, particularly for cancer genomics. For this to occur, advances will be needed in techniques for efficiently isolating intact long DNA molecules and in NGS methods for accurately reading the sequence content of these molecules. The field of NGS development and applications is a fast-moving area of research, which makes this an exciting time for genomic studies.

1. International Human Genome Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
3. Hutchison, C. A. III. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.* **35**, 6227–6237 (2007).
4. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19–21 (2008).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
The Review provides a comprehensive overview of recent advances and challenges in techniques that are used in transcriptome profiling methods that use NGS technologies (RNA–seq).
6. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotech.* **26**, 1146–1153 (2008).
An excellent review of the current state of nanopore sequencing that highlights recent accomplishments and remaining challenges in the field.
7. Fan, J.-B., Chee, M. S. & Gunderson, K. L. Highly parallel genomic assays. *Nature Rev. Genet.* **7**, 632–644 (2006).
8. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149 (2008).
9. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 8817–8822 (2003).
10. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34**, e22 (2006).
11. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
This paper describes the development of the non-cleavable SBL method and shows its feasibility by sequencing the *E. coli* genome. The prototype described led to the development of the Polonator instrument.
12. Kim, J. B. *et al.* Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481–1484 (2007).
13. Leamon, J. H. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769–3777 (2003).

14. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
Developers from Helicos BioSciences and colleagues describe the development of the first single-molecule sequencing method using reversible terminators and demonstrate the technology by sequencing the M13 genome.
15. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
The authors describe the development of a real-time sequencing method using their ZMW detection system and demonstrate its feasibility by sequencing synthetic templates.
16. Hardin, S., Gao, X., Briggs, J., Willson, R. & Tu, S.-C. Methods for real-time single molecule sequence determination. US Patent 7,329,492 (2000).
17. Williams, J. G. K. System and methods for nucleic acid sequencing of single molecules by polymerase synthesis. US Patent 6,255,083 (1998).
18. Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods* **5**, 679–682 (2008).
19. Metzker, M. L. *et al.* Termination of DNA synthesis by novel 3′-modified deoxyribonucleoside triphosphates. *Nucleic Acids Res.* **22**, 4259–4267 (1994).
20. Canard, B. & Sarfati, R. DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**, 1–6 (1994).
21. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. USA* **103**, 19635–19640 (2006).
22. Guo, J. *et al.* Four-color DNA sequencing with 3′-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl Acad. Sci. USA* **105**, 9145–9150 (2008).
23. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
Developers from Illumina/Solexa and colleagues report details on their reversible terminator platform and demonstrate the technology by sequencing a flow-sorted X chromosome and the genome from a Yoruban male.
24. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
25. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**, 183–188 (2008).
26. Harismendy, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.* **10**, R32 (2009).
27. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
28. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Rev. Genet.* **10**, 241–251 (2009).
29. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
30. Sarin, S., Prabhu, S., O'Meara, M. M., Pe'er, I. & Hobert, O. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nature Methods* **5**, 865–867 (2008).
31. Wu, W. *et al.* Termination of DNA synthesis by N^6-alkylated, not 3′-*O*-alkylated, photocleavable 2′-deoxyadenosine triphosphates. *Nucleic Acid Res.* **35**, 6339–6349 (2007).
32. Wu, W., Litosh, V. A., Stupi, B. P. & Metzker, M. L. Photocleavable labeled nucleotides and nucleosides and methods for their use in DNA sequencing. US Patent Application 11/567,189 (2009).
33. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593–595 (2009).
34. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
35. Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. & Ellenberger, T. DNA ligases: structure, reaction mechanism, and function. *Chem. Rev.* **106**, 687–699 (2006).
36. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077–1080 (1988).
37. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
This paper describes Life/APG's SBL method, which uses cleavable two-base-encoded probes on the SOLiD platform. The authors demonstrate the technology through the application of genome-wide nucleosome mapping in *C. elegans*.
38. Shen, Y., Sarin, S., Liu, Y., Hobert, O. & Pe'er, I. Comparing platforms for *C. elegans* mutant identification using high-throughput whole-genome sequencing. *PLoS ONE* **3**, e4012 (2008).
39. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281**, 363–365 (1998).
40. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).
41. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
The authors describe the development of the first NGS technology using the pyrosequencing method and demonstrate its feasibility through the sequencing and *de novo* assembly of the *Mycoplasma genitalium* genome.
42. Metzker, M. L. Sequencing in real time. *Nature Biotech.* **27**, 150–151 (2009).
43. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
44. Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).
45. Chaisson, M. J., Brinza, D. & Pevzner, P. A. *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.* **19**, 336–346 (2009).
46. Hofreuter, D. *et al.* Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect. Immun.* **74**, 4694–4707 (2006).
47. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genet.* **40**, 987–993 (2008).
48. Srivatsan, A. *et al.* High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.* **4**, e1000139 (2008).
49. Suchland, R. J. *et al.* Identification of concomitant infection with *Chlamydia trachomatis* IncA-negative mutant and wild-type strains by genomic, transcriptional, and biological characterizations. *Infect. Immun.* **76**, 5438–5446 (2008).
50. Nusbaum, C. *et al.* Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. *Nature Methods* **6**, 67–69 (2009).
51. Moran, N. A., McLaughlin, H. J. & Sorek, R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. *Science* **323**, 379–382 (2009).
52. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).
53. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
54. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
55. Warren, R. L., Sutton, G. G., Jones, S. J. M. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–501 (2007).
56. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
57. Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
58. Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
59. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
60. Aury, J.-M. *et al.* High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
61. Reinhardt, J. A. *et al.* *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res.* **19**, 294–305 (2009).
62. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nature Biotech.* **26**, 1113–1115 (2008).
63. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
64. Wang, L. & Weinshilboum, R. M. Pharmacogenomics: candidate gene identification, functional validation and mechanisms. *Hum. Mol. Genet.* **17**, R174–R179 (2008).
65. Haaland, W. C. *et al.* A–β– subtype of ketosis-prone diabetes is not predominantly a monogenic diabetic syndrome. *Diabetes Care* **32**, 873–877 (2009).
66. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotech.* **27**, 1025–1031 (2009).
67. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotech.* **17**, 974–978 (1999).
68. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
69. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
70. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**, 907–909 (2007).
71. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
72. Krishnakumar, S. *et al.* A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl Acad. Sci. USA* **105**, 9296–9301 (2008).
73. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
74. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech.* **27**, 182–189 (2009).
75. Olson, M. Enrichment of super-sized resequencing targets from the human genome. *Nature Methods* **4**, 891–892 (2007).
76. Garber, K. Fixing the front end. *Nature Biotech.* **26**, 1101–1104 (2008).
77. Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A. & Versalovic, J. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* **55**, 856–866 (2009).
These authors describe the current state of metagenomics research and highlight the use of the Roche/454 platform for microbial identification through 16S ribosomal DNA phylogenetic analysis; other NGS platforms may be better suited for gene discovery efforts (see Table 2).
78. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nature Biotech.* **27**, 652–658 (2009).
79. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
80. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
The article provides a comprehensive review of recent technological advances and challenges in genome-wide profiling of DNA-binding proteins, histone modifications and nucleosomes using NGS technologies (ChIP–seq).
81. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
82. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
83. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
84. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
85. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
86. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).

87. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
88. Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
89. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
90. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
91. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nature Biotech.* **27**, 847–850 (2009).
92. Mardis, E. R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
93. Lupski, J. R. *et al.* Complete genome sequencing identifies recessive alleles in SH3TC2 causing a CMT1 neuropathy. *N. Engl. J. Med.* (in the press).
94. Collins, F. S. & Barker, A. D. Mapping the cancer genome. *Sci. Am.* **296**, 50–57 (2007).
95. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 5 Nov 2009 (doi:10.1126/science.1181498).
96. Sanderson, K. Personal genomes: standard and pores. *Nature* **456**, 23–25 (2008).
97. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).
98. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009).
99. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **136**, 629–641 (2009).
100. Carthew, R. W. & Sontheimer, E. J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
101. Barnes, C., Balasubramanian, S., Liu, X., Swerdlow, H. & Milton, J. Labelled nucleotides. US Patent 7,057,026 (2002).
102. Mitra, R. D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik & Church, G. M. Fluorescent *in situ* sequencing on polymerase colonies. *Anal. Biochem.* **320**, 55–65 (2003).
103. Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res.* **36**, e25 (2008).
104. Yarbrough, L. R., Schlageck, J. G. & Baughman, M. Synthesis and properties of fluorescent nucleotide substrates for DNA-dependent RNA polymerases. *J. Biol. Chem.* **254**, 12069–12073 (1979).
105. Kumar, S. *et al.* Terminal phosphate labeled nucleotides: synthesis, applications, and linker effect on incorporation by DNA polymerases. *Nucleosides Nucleotides Nucleic Acids* **24**, 401–408 (2005).
106. McKernan, K., Blanchard, A., Kotler, L. & Costa, G. Reagents, methods, and libraries for bead-based sequencing. US Patent Application 11/345,979 (2005).
107. Macevicz, S. C. DNA sequencing by parallel oligonucleotide extensions. US Patent 5,969,119 (1995).
108. Mir, K. U., Qi, H., Salata, O. & Scozzafava, G. Sequencing by cyclic ligation and cleavage (CycLiC) directly on a microarray captured template. *Nucleic Acids Res.* **37**, e5 (2009).

Acknowledgements

I am extremely grateful to S.-M. Ahn, J. Edwards, J. W. Efcavitch, R. A. Gibbs, T. Harkin, E. Mardis, K. McKernan, D. Muzny, S. Turner and D. Wheeler for providing current performance data for the NGS platforms, and to the National Human Genome Research Institute for their support from grants R01 HG003573, R41 HG003072, R41 HG003265 and R21 HG002443.

Competing interests statement

The author declares competing financial interests: see Web version for details.

FURTHER INFORMATION

Michael L. Metzker's homepage: http://www.bcm.edu/genetics/?pmid=10947
1000 Genomes Project: http://www.1000genomes.org
Advances in Genome Biology and Technology meeting: http://www.agbt.org
The Cancer Genome Atlas: http://cancergenome.nih.gov
The Exome Project: http://www.nhlbi.nih.gov/resources/exome.htm
Human Microbiome Project: http://nihroadmap.nih.gov/hmp
National Human Genome Research Institute: http://www.genome.gov
National Human Genome Research Institute — Fruitfly Genome Sequencing: http://www.genome.gov/11008080
***Nature Reviews Genetics* poster 'Sequencing technologies — the next generation':** http://www.nature.com/nrg/posters/sequencing/index.html
Personal Genome Project: http://www.personalgenomes.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

ATTACHMENT 5A

LDR	01822cas a2200589 a 4500
001	999627793406676
005	20250227135114.0
007	cr|cn#||||||||
008	001127c20009999enkmr p 0 a0eng d
010	__ | **a** 2001229989
022	__ | **a**1471-0056**l**1471-0056**9**PY**7**(Print)
022	__ | **a**1471-0064**9**EN**7**(Electronic)
030	__ | **a**NRGAAM
035	__ | **9**100962779
040	__ | **a**OUN**c**OUN**d**NSD**d**CAS**d**FUH**d**WAU**d**FUH
041	0_ | **a**eng
042	__ | **a**lcd**a**nlmcopyc
044	__ | **9**England
050	14 | **a**QH426**b**.N38
060	00 | **a**W1**b**NA839H
210	10 | **a**Nat Rev Genet**2**DNLM
245	00 | **a**Nature reviews.**p**Genetics.
246	30 | **a**Genetics
246	20 | **a**Nature Rev Genet**9**n
260	__ | **a**London, UK :**b**Nature Pub. Group,**c**[2000-
300	__ | **a**v. :**b**ill. ;**c**28 cm.
310	__ | **a**Monthly
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**Vol. 1, no. 1 (Oct. 2000)-
500	__ | **a**Title from cover.
510	1_ | **9**1**a**MEDLINE**b**v1n1, Oct. 2000-
510	1_ | **9**1**a**PubMed**b**v1n1, Oct. 2000-
510	1_ | **a**Index medicus
530	__ | **a**Available also by subscription via the World Wide Web.
583	__ | **3**v. 1, no. 1 (Oct. 2000)**c**20010321**f**Notice published**l**Printed on acid-free paper
650	12 | **a**Genetics
655	_2 | **a**Periodical
856	4_ | **u**http://www.nature.com/nrg/
911	__ | **a**Electronic and print, print canceled 2017
914	__ | **a**Portico**b**Current to 2024 as of 20250227**c**ISSN_14710056
935	__ | **a**(DNLM)962779-nlmdb
936	__ | **a**Vol. 2, no. 2 (Feb. 2001) LIC
992	__ | **p**P2**e**EF**a**20010202

993	__ | **a**HPN**b**20010209
994	__ | **a**CDN**b**20010209
995	__ | **a**AUTH**b**20010209**c**REV**d**20250227
996	__ | **a**CONSER**b**20000101
998	__ | **a**ERMSLC 108646**b**20061129
999	__ | **a**AUTH
035	__ | **a**(OCoLC)45426000

ATTACHMENT 5B

Genetics MeSH Descriptor Data 2025

MeSH Heading
Genetics
Tree Number(s)
H01.158.273.343
Unique ID
D005823
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D005823
Annotation
the discipline (education, history, etc) only; use GENETIC PHENOMENA or specifics treed under it for genetic processes and properties of organisms, or the qualifier / genet with specific anatomical terms, organisms, diseases, etc.
Scope Note
The branch of science concerned with the means and consequences of transmission and generation of the components of biological inheritance. (Stedman, 26th ed)
Entry Version
GENET
See Also
Genetic Structures
Histocompatibility Testing
Entry Combination
methods:Genetic Techniques
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2016/06/17

VOLUME 5

ENCYCLOPEDIA OF MEDICAL DEVICES AND INSTRUMENTATION

SECOND EDITION

JOHN G. WEBSTER

EX1019

ENCYCLOPEDIA OF

MEDICAL DEVICES AND INSTRUMENTATION

Second Edition
Volume 5

Nanoparticles – Radiotherapy Accessories

Edited by

John G. Webster
University of Wisconsin–Madison

The *Encyclopedia of Medical Devices and Instrumentation* is available online at http://www.mrw.interscience.wiley.com/emdi

WILEY-INTERSCIENCE

A John Wiley & Sons, Inc., Publication

Copyright © 2006 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222, Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warrnaties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Encylopedia of medical devices & instrumentation/by John G. Webster,

editor in chief. – 2nd ed.
p. ; cm.
Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
Includes bibliographical references and index.
ISBN-13 978-0-471-26358-6 (set : cloth)
ISBN-10 0-471-26358-3 (set : cloth)
ISBN-13 978-0-4070-04070-6 (v. 5 : cloth)
ISBN-10 0-471-04070-X (v. 5 : cloth)

1. Medical instruments and apparatus–Encyclopedias. 2. Biomedical engineering–Encyclopedias. 3. Medical physics–Encyclopedias. 4. Medicine–Data processing–Encyclopedias. I. Webster, John G., 1932- . II. Title: Encyclopedia of medical devices and instrumentation.

[DNLM: 1. Equipment and Supplies–Encyclopedias–English. W 13

E555 2006]
R856.A3E53 2006
610.2803–dc22
2005028946

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

POLYMERASE CHAIN REACTION

MICHAEL L. METZKER
Baylor College of Medicine
Houston, Taxas

THOMAS C. CASKEY
Cogene Biotech Ventures
Houston, Taxas

INTRODUCTION

Few techniques rival the impact that the polymerase chain reaction (PCR) has made in the age of molecular biology. Cloning and deoxyribonucleic acid (DNA) sequencing are other such techniques that have become embedded into everyday life on the molecular biologist's bench. Over 60 books alone (not to mention the tens of thousands of research articles) have been devoted to the strategies, methods and applications of PCR for the identification, detection and diagnosis of genetic and infectious diseases. Rightfully so, the inventor of PCR, Kary B. Mullis, was awarded the Nobel Prize in Chemistry for his discovery of the technique in 1993. However, PCR has not been without controversy. In 1989, DuPont challenged the validity of the Cetus PCR patents in federal court and with the Office of Patents and Trade Marks, and by 1991 the Cetus patents were unanimously upheld and later sold to Hoffman La Roche for $300 million. More recently, in 1993, Promega has challenged the validity of the Hoffmann La Roche *Taq* DNA polymerase patent that is currently pending. In this article, we attempt to provide a comprehensive overview for the molecular biologist when applying PCR to his/her application of interest.

DNA POLYMERASE REACTION

The DNA replication is an inherent process for the generation and evolution of future progeny in all living organisms. At the heart of this process is the DNA polymerase that primarily synthesizes new strands of DNA in a 5′→3′ direction from a single-stranded template. Most native DNA polymerases, however, are polyfunctional and show 5′-exonuclease and/or 3′-exonuclease activities that are important for cellular DNA repair and proofreading functions. Numerous molecular biology applications have harnessed these activities, such as labelling DNA by nick translation and TaqMan assays (see below), and endrepair of sheared DNA fragments and improving DNA synthesis fidelities, respectively. The PCR is an elegant, but simple, technique for the *In vitro* amplification of target DNA utilizing DNA polymerase and two specific oligonucleotide or primer sequences flanking the region of interest. PCR is a cyclic process of double-strand separation of DNA by heat denaturation, specific hybridization or annealing of short oligonucleotide primers to singlestranded DNA, and synthesis by DNA polymerase (1,2). Each cycle doubles the region marked by the primer sequences. By sequential iteration of the process, PCR exponentially generates up to a billion of copies of the target within just a few hours (Fig. 1).

The specificity of PCR is highly dependent on the careful design of unique primers with respect to the genome under investigation and the nucleotide composition of the primer sequences. Theoretically, a 16-mer (4^{16}) is of sufficient length to represent all unique primer sequences from a completely random genome size of 3 billion base pairs. In the real world, however, all genomes are not random and contain varying degrees of repetitive elements. For the human genome, Alus, LINEs (long interspersed DNA elements) and low complexity repeats are frequently observed and should be avoided in primer design when possible. There are a few simple rules for designing primer sequences that work well in PCR. In practice, PCR primers should be between 18 and 25 nucleotides long, have roughly an equal number of the four nucleotides, and show a G + C composition of 50–60%. Commercially available oligonucleotide synthesizers that show phosphamidite coupling efficiencies > 98% mean that primers of this size can usually be used in PCR without purification. A variety of computer programs are available for selecting primer sequences from a target region. Many of these programs will reveal internal hairpin structures and self-annealing primer sequences, but manual inspection of the oligonucleotide is still necessary to maximize successful PCR amplifications.

The concentrations of the PCR cocktail ingredients are also important for product specificity, fidelity and yield. In addition to *Taq* DNA polymerase and primers, the PCR mixture contains the cofactor magnesium ion (Mg^{2+}), the four 2′-deoxyribonucleoside-5′-triphosphates (dNTPs) and the buffer. In general, PCR reagent concentrations that are too high from standard conditions result in nonspecific products with high misincorporation errors, and those that are too low result in insufficient product. A typical 50 μL PCR cocktail that contains 0.4 $\mu mol \cdot L^{-1}$ of each primer, 200 $\mu mol \cdot L^{-1}$ of each dNTP, 1.5 $mmol \cdot L^{-1}$ $MgCl_2$, and 1.25 units *Taq* DNA polymerase in 10 $mmol \cdot L^{-1}$ tris-HCl, pH 8.3, 50 $mmol \cdot L^{-1}$ KCl buffer works well for most PCR applications. The optimal Mg^{2+} concentration, however, may need to be determined empirically for difficult target templates. The performance and fidelity of *Taq* DNA polymerase are sensitive to the free Mg^{2+} concentration (3), which ionically interacts with not only the dNTPs but also with the primers, the template DNA, ethylenediaminetetraacetic acid (EDTA), and other chelating agents. In most cases, the Mg^{2+} concentration will range from 1.0 to 4.0 $mmol \cdot L^{-1}$.

The number of cycles and the cycle temperature–length of time for template denaturation and primer annealing and extension are important parameters for high quality PCR results. The optimal number of cycles is dependent on the starting concentration or copy number of the target DNA and typically ranges from 25 to 35 cycles. Too many cycles will significantly increase the amount of nonspecific PCR products. For low copy number targets, such as the integrated provirus of *Human immunodeficiency virus type 1* (HIV-1) from human genomic DNA, two rounds of PCR are employed first using an outer primer pair set followed by an internal (nested) primer pair set flanking the region of interest to yield positive and specific PCR products. Brief, but effective denaturation conditions, that

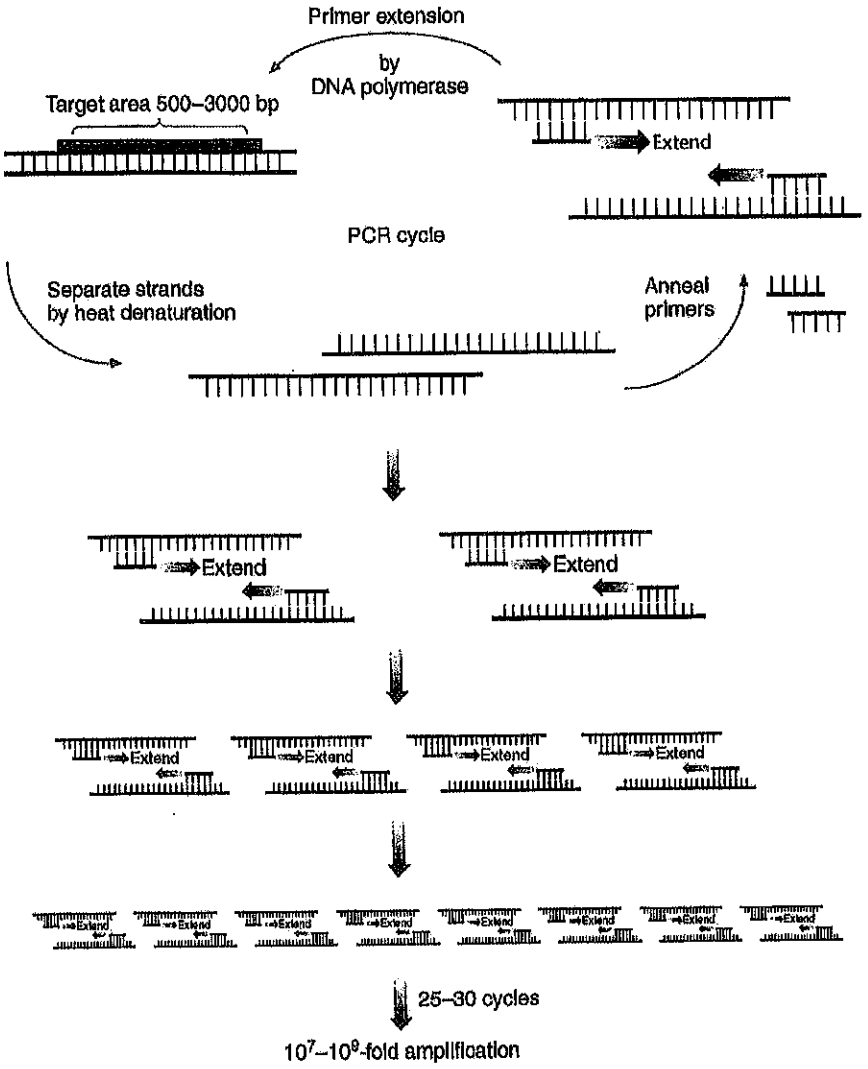


Figure 1. The PCR amplification cycle.

is 94–97 °C for 15–30 s, are necessary as *Taq* DNA polymerase has a half-life of only 40 min at 95 °C. Annealing conditions, on the other hand, are dependent on the concentration, base composition and the length of the oligonucleotide and typically range between 55 and 68 °C for 30–60 s. The length of the amplified target is directly proportional to the primer extension length of time. Primer extension is performed between 68 and 72 °C and, as a rule of thumb, is ~ 60 s for every 1 kb.

Crude extracts from blood, cerebral spinal fluid, urine, buccal smears, bacterial colonies, yeast spores, and so on are routinely used as sources of DNA for PCR templates. Due to the high sensitivity of PCR, rapid isolation protocols, such as heat and detergent disruptions, and enzymatic digestion of biological samples have been frequently used. Caution should be invoked when using crude extracts as starting materials for PCR amplifications because a number of impurities are known to inhibit *Taq* DNA polymerase. These include red blood cell components, sodium dodecyl sulfate (SDS), high salts, EDTA and too much DNA. Since only a few hundred target molecules are needed for successful PCRs, in most cases, these impurities can be effectively removed by simply diluting the starting material. Each sample should then be tested with control primers that specifically amplify a known target to determine the integrity of the crude extract. Alternatively, the isolation of the desired organism, such as HIV-1, human *Hepatitis A virus*, influenza virus, cytomegalovirus, and so on, or the isolation of specific cell fractions, such as peripheral blood mononuclear cells, can significantly increase the sensitivity and specificity of the PCR amplifications.

SENSITIVITY AND CONTAMINATION OF POLYMERASE CHAIN REACTION

Contamination is the dark side of the PCR force. The exquisite sensitivity of PCR can result in contamination from even a single molecule of foreign or exogenous DNA (4,5). To minimize false positives, standard operating procedures have been described, including the physical isolation of PCR reagents from the preparation of DNA templates and PCR products, using autoclaved solutions, premixing and aliquoting reagents, the use of disposable

gloves, avoiding splashes, the use of positive displacement pipettes, adding DNA last, and carefully choosing positive and negative controls (6). Contamination is likely to surface for DNA samples that are difficult to amplify because of sequence content, or due to poor primer design and chemical impurities in DNA extractions. This is especially true for low copy number targets or degraded samples, as greater numbers of amplification cycles are generally required to achieve the desired product. In these cases, residual amounts of exogenous DNAs can compete and override the amplification process, resulting in spurious data. The best approach to challenge dubious results is to repeat the experiment with scrupulous care to details and controls. Biological samples collected at a single time point should be divided into multiple aliquots such that independent DNA extractions and PCR experiments can be performed to verify and validate initial results. Data should be discarded if inconsistent positive and negative PCR results occur upon repetition of the experiment. While negative controls can rule out reagent contamination, sporadic contamination can go unchecked. The probability of repeating spurious contamination in a consistent manner is extremely low.

There are three sources of contaminating DNA: (*1*) carryover contamination from previously amplified PCR products; (*2*) cross-contamination between multiple source materials; and (*3*) plasmid contamination from a recombinant clone that contains the target sequence. Of the three, carryover contamination is considered to be the major source of contamination because of the relative abundance of amplified target sequences. The substitution of dUTP for dTTP in the PCR cocktail has been routinely used as a method of preventing carryover contamination. Pretreatment of subsequent PCR mixtures prior to thermal cycling with uracil DNA glycosylase results in the removal of dU from any carryover PCR product, but does not affect the template DNA or dUTP. The dU removal creates an abasic site that is heat labile and degrades during thermal cycling, thus preventing carryover amplification. Moreover, ultraviolet (UV) light can reduce work surface and reagent contamination. Cross-contamination between samples is more difficult to diagnose, and suspicious results should be repeated from independent DNA extracts and PCR experiments for samples in question. Plasmid contamination, on the other hand, can be identified by sequence analysis and comparison to all laboratory plasmid sequences.

POLYMERASE CHAIN REACTION INTRODUCES MUTATIONS

The power and ease of PCR, however, were not fully appreciated until the introduction of the thermostable DNA polymerase isolated from *Thermus aquaticus* (*Taq*) (7) and automated instrumentation in 1988. It was here that PCR could be run in fully closed and automated systems. Fresh Klenow DNA polymerase did not have to be added at each cycle and PCR could be performed at higher annealing and extension temperatures, which increased the specificity and yields of the reactions while minimizing the risks of contamination. A hot start PCR further enhances specificity by preventing the formation of nonspecific products that arise during the initial steps of thermal cycling in PCR.

Taq DNA polymerase has been shown to incorporate nucleotides incorrectly at a frequency of 1 in 9000 bases by a mutation reversion assay (8). From sequence analysis of cloned PCR products, a slightly higher error frequency was determined (1 in 4000–5000 bp) for *Taq* DNA polymerase (9). The fidelity of DNA synthesis for *Taq* DNA polymerase, however, can vary significantly with changes in free Mg^{2+} concentration, changes in the pH of the buffer, or an imbalance in the four dNTP concentrations. Polymerase misincorporation errors are minimized when the four dNTPs are equimolar and between 50 and 200 $\mu mol \cdot L^{-1}$ (9). Since *Taq* DNA polymerase lacks a 3′-exonuclease activity, misincorporated bases typically cause chain termination of DNA synthesis that are not propagated in subsequent PCR cycles. In a worst-case scenario, a mutation occurring during the first round of PCR from a single target molecule and propagated thereafter would exist at a frequency of 25% in the final PCR product. Since hundreds of target copies are routinely used as starting DNA in PCR and most misincorporations terminate DNA synthesis, the observed error frequency is $\ll 25\%$.

Cloning of full-length genes from PCR products, however, has been problematic because PCR-induced mutations can cause amino acid substitutions in the wildtype sequence. Thus, significant effort must be employed in the complete sequencing of multiple PCR clones to identify mutation-free clones or ones that contain synonymous substitutions that do not change the protein coding sequence. Accordingly, thermostable DNA polymerases that contain a 3′-exonuclease (3′-exo) activity for proofreading of misincorporated bases have been recently introduced and include DNA polymerases isolated from *Pyrococcus furiosus (Pfu), Thermococcus litoralis* (Vent), *Pyrococcus* species GB-D (Deep Vent) and *Pyrococcus woesei* (*Pwo*). The error frequencies of these DNA polymerases are two- and sixfold < *Taq* DNA polymerase (10), but these polymerases are difficult for routine use, as the 3′-exonuclease activity can easily degrade the single-stranded PCR primers. 3′-Exo DNA polymerases, however, have been successfully used in long PCR in combination with *Taq* DNA polymerase and show an approximately twofold lower error frequency than *Taq* DNA polymerase alone (10).

POLYMERASE CHAIN REACTION LENGTH LIMITATIONS

For most applications, standard PCR conditions can reliably amplify target sizes up to 3–4 kb from a variety of source materials. Target sizes > 5 kb, however, have been described in the literature using standard PCR conditions, but generally yield low quantities of PCR product. The PCR size limitation can be attributed to the misincorporation of nucleotides that occurred 1 in 4000–5000 bp that ultimately reduced the efficiency of amplifying longer target regions. A breakthrough in long PCR came through the combined use of two thermostable DNA

polymerases, one of which contains a 3′-exonuclease activity (11,12). The principle for long PCR is that the *Taq* DNA polymerase performs the high fidelity DNA synthesis part of PCR, coupled with the proofreading activity of *Pfu*, Vent or *Pwo* DNA polymerases. Once the nucleotide error is corrected, *Taq* DNA polymerase can then complete the synthesis of long PCR templates. From empirical studies, only a trace amount of the 3′-exo DNA polymerase, roughly 1% to that of *Taq* DNA polymerase or another DNA polymerase isolated from *Thermus thermophilis (Tth)*, is needed to perform long PCRs > 20 kb. Other important factors for long PCR are the isolation of high quality, high molecular weight DNA and protection against template damage, such as depurination during thermal cycling. The use of the cosolvents glycerol and dimethyl sulfoxide (DMSO) have been shown to protect against DNA damage by efficiently lowering the denaturation temperature by several degrees centigrade. The rule of thumb for primer extensions still applies for long PCRs (60 $s \cdot kb^{-1}$), although for targets > 20 kb, times extension should not exceed 22 min $cycle^{-1}$. The complexity and size of the genome under investigation can also affect the size of long PCR products. For example, PCR product lengths of 42 kb have been described for the amplification of λ bacteriophage DNA (11,12), compared with a 22 kb PCR product obtained from the human β-globin gene cluster (12).

CREATION OF NOVEL RECOMBINANT MOLECULES BY POLYMERASE CHAIN REACTION

Polymerase chain reaction can amplify both single- and double-stranded DNA templates as well as complementary DNA (cDNA) from the reverse transcription of messenger ribonucleic acid (mRNA) templates. Because of the flexibility of automated DNA synthesis, *In vitro* mutagenesis experiments can easily be performed by PCR. Recombinant PCR products can be created via the primer sequences by tolerated mismatches between the primer and the template DNA or by 5′-add-on sequences. Primer mediated mutagenesis can accommodate any nucleotide substitutions and small insertions or deletions of genetic material. The desired genetic alteration can be moved to any position within the target region by use of two overlapping PCR products with similar mutagenized ends (Fig. 2, left). This is accomplished by denaturing and reannealing the two overlapping PCR products to form heteroduplexes that have 3′-recessed ends. Following the extension of the 3′-recessed ends by *Taq* DNA polymerase, the full-length recombinant product is reamplified with the outer primers only to enrich selectively the full-length recombinant PCR product. 5′-Add-on adapters can also be used to join two unrelated DNA sequences, such as the splicing of an exogenous promoter sequence with a gene of interest (Fig. 2, right). The promoter–gene sequences are joined at the

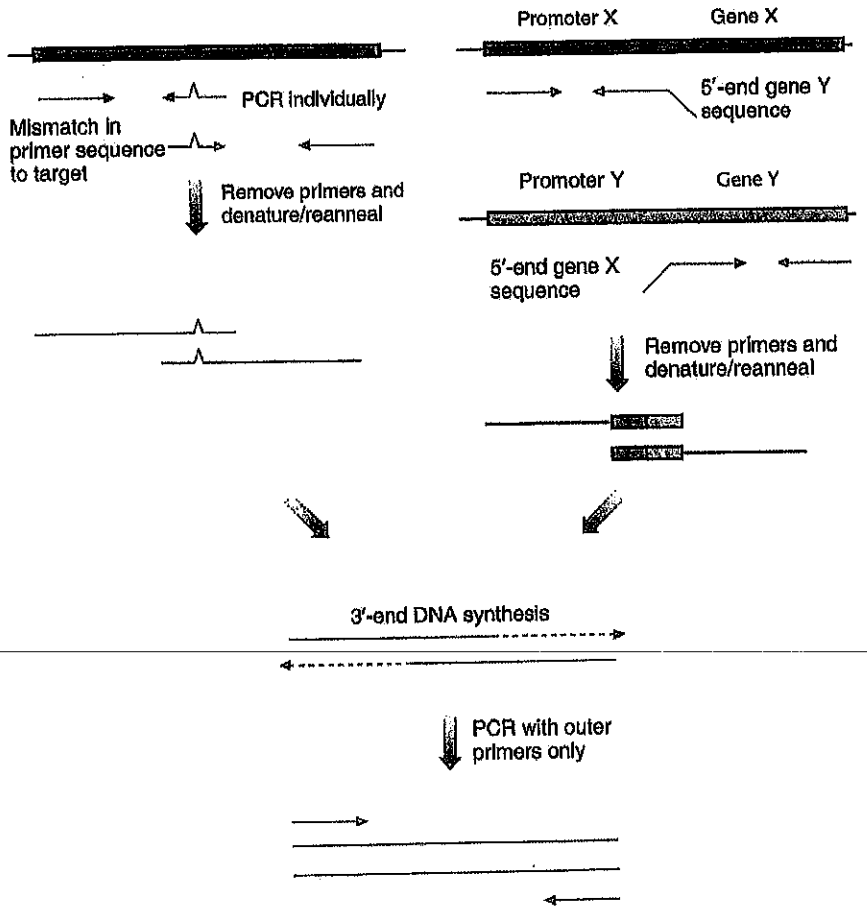


Figure 2. Creation of mutagenized or recombinant PCR products via primer mismatches (left) or 5′-add-on sequences (right).

desired junction by 5′-add-on gene specific and 5′-add-on promoter-specific adapters that can PCR amplify the promoter and the gene targets, respectively. Heteroduplexes can then be formed, as described above, from the two overlapping PCR products, which are then selectively amplified with outer primers to generate the desired full-length recombinant PCR product.

POLYMERASE CHAIN REACTION AS A DETECTION SYSTEM

Polymerase chain reaction is a powerful tool for the detection of human polymorphic variation that has been associated with hereditary diseases. Many PCR techniques have been described that can discriminate between wild-type and mutant alleles, but in this section only a few of the most frequently used techniques are discussed. Of these, DNA sequencing of PCR products is the most widely used and most sensitive method for the detection of both novel and known polymorphic differences between individuals. Complementing scanning technologies, however, have been developed for the rapid detection of allelic differences because of the high costs associated with DNA sequencing and the ability to process large numbers of samples. Singlestrand conformation polymorphism (SSCP) has been commonly used as a technique for the identification of genetic polymorphisms. Following PCR, the product is heat denatured and subjected to native or nondenaturing gel electrophoresis. Allelic differences between samples are detected as mobility band shifts by radioactive and nonradioactive labeling procedures. PCR–SSCP, however, is limited in fragment size to $\sim$ 200 bp because the accuracy in discriminating between different alleles diminishes significantly with an increase in the fragment length.

Multiplex PCR allows for the simultaneous amplification of multiple target regions and has been particularly useful for the detection of exon deletion(s) in X-linked disorders, such as Duchenne muscular dystrophy (13) and Lesch–Nyhan syndrome (14). The multiplex PCR products are resolved by gel electrophoresis and are visualized by ethidium bromide staining. The absence of specific PCR product(s) is diagnostic of exon deletion(s) in affected males, and half-dosage PCR products are diagnostic of carrier mothers (15). Moreover, up to 46 primer pairs have been simultaneously amplified by multiplex PCR with excellent success (90%) for the large-scale identification of human single nucleotide polymorphisms (SNPs) by hybridization to high density DNA chip arrays.

Genetic polymorphisms can also be identified by immobilizing the PCR product on to a nylon membrane in a dot blot format and probing by hybridization with an allele-specific oligonucleotide (ASO) that contains a 5′-biotin group. The ASO hybridization is detected by adding streptavidin-horseradish peroxidase, which binds to the biotinylated probe, followed by a colorimeter assay. The colorimeter ASO assay has been applied to the genotyping of human leucocyte antigen (HLA)–DQA alleles and the detection of β-thalassaemia mutations. More recently, multiplex PCR and colorimeter ASO methodologies have been combined in a reverse fashion, in which ASOs are immobilized on to nylon membrane strips and probed against biotinylated gene-specific multiplex PCR reactions. Allele-specific PCR products are detected by hybridization and conversion of a colourless substrate to a blue precipitate for the simultaneous genotyping of HLA–DQA 1, low density lipoprotein receptor, glycophorin A, hemoglobin G gammaglobin, D7S8, and groupspecific component.

Lastly, *In situ* PCR enables the amplification of target sequences from sections of formalin-fixed, paraffin-embedded tissue specimens to determine the levels of gene expression in specific cell types that otherwise could not be detected by conventional *In situ* hybridization. The PCR is performed directly on glass slides by overlaying the PCR mixture on to the specimen, sealing the slides to prevent evaporation, and temperature cycling using a thermal sensor or modified thermal cycler that holds glass slides.

DEGENERATE POLYMERASE CHAIN REACTION

Degenerate PCR is a powerful strategy for obtaining novel full-length cDNA sequences from limited amino acid sequence information (16). The PCR primer sequences are derived from the reverse translation of 6–9 amino acid codons, which will result in varying levels of degeneracy except for methionine and tryptophan residues. Careful attention should be exercised in the design of degenerate primers because increasing the primer complexity (i.e., using codons that show more than twofold degeneracy) will typically result in an increase in nonspecific PCR products. One approach in reducing the complexity of the degenerate primer is the use of codon bias for the particular organism from which the gene will be cloned. Alternatively, the alignment of orthologous gene sequences from other species can greatly improve the specificity of cloning the gene of interest by revealing evolutionarily conserved domains. Once the optimal primer sequence is determined, the mixture of oligonucleotides can be simultaneously synthesized and will represent all possible amino acid combinations of the degenerate sequence. The specificity of PCR should then selectively amplify the correct primer sequences to generate a gene or gene family-specific probe from which the full-length cDNA can be obtained. Degenerate PCR has been successfully used in the screening of novel gene family members such as G-protein-coupled receptors, nuclear steroid receptors and protein tyrosine kinases.

ANCIENT DNA

Phylogenetics is the study of evolutionary relationships between specimens that are inferred from contemporaneous sequences. The ability to obtain DNA sequences from specimens or even fossils that are millions of years old could equip the phylogeneticist with a powerful means of directly testing an a priori hypothesis. Following death of the tissue or organism, however, DNA is rapidly degraded by, presumably, nuclease activities and hydrolytic processes, resulting in short fragment sizes that are generally no longer than 100–150 bp. Moreover, this old DNA is largely modified by oxidative processes and by intermolecular crosslinks that render it unsuitable for cloning by standard molecular biology procedures. Short PCRs,

however, have been successfully performed from DNA samples isolated from archival and ancient specimens (17).

Museums hold vast collections of archived hospital files of patient specimens and of different species that have been collected over the last century. In a recent study, phylogenetic analyses of DNA sequences were performed from reverse transcriptase PCR (RT-PCR) of formalin-fixed, paraffin-embedded tissue specimens obtained from U.S. servicemen killed in the 1918 Spanish influenza pandemic. Viral sequences from three different gene regions were consistent with a novel H1N1 *Influenza A virus* that was most closely related to influenza strains that infect humans and swine, but not wild waterfowl, considered to be the natural reservoir for the influenza virus (18). Moreover, PCR and DNA sequencing have been performed on DNA extractions of archaeological findings, such as amplifying mitochondrial DNA sequences from a 7000 year old human brain, amplifying both mitochondrial and nuclear DNA sequences from bone specimens from a 14,000 year old saber-toothed cat, and amplifying chloroplast DNA sequences from fossil leaf samples from a 17 million-year-old Miocene *Magnolia* species.

QUANTITATIVE POLYMERASE CHAIN REACTION

Quantitative PCR (QPCR) has been widely used for detecting and diagnosing genetic deletions, for studying gene expression and for estimating the viral load of HIV-1. While DNA quantitation by multiplex PCR has been previously described (15), the quantitation of RNA has been wide reaching for the latter two areas. For many applications, estimating the relative amount of PCR product is sufficient to describe a biological observation. The absolute quantitation of RNA molecules, however, has been more difficult than for DNA because of the difficulty of generating accurate controls. Internal standards derived from synthetic RNA or cRNA have been designed to contain the same primer sequences as the target but yield a different-sized PCR product that can be easily separated by gel electrophoresis. cRNAs are not only coamplified with target sequences, but also coreverse transcribed to account for the variable efficiencies of cDNA syntheses. Moreover, QPCR is typically performed in the exponential or log phase of the amplification process (typically 14–22 cycles) to obtain accurate quantitative results. The absolute amount of target mRNA can be quantitated by serial dilutions of the target/internal control mixture and by extrapolating against the standard curve.

Both the variable range of initial target amounts and the presence of various inhibitors can, however, adversely affect the kinetics and efficiencies of PCR. Alternatively, a strategy based on a quantitative competitive (QC) approach has been used to minimize the effects of these variables. Known quantities of the competitor template, which contains the same primer sequences as the target but differs in size, are introduced into replicate PCRs containing identical quantities of the target. The point at which the intensities of the PCR products derived from the target sequence and the competitor template are equivalent is used to estimate the amount of target sequence in the original sample (19).

Recently, real-time QPCR and QCPCR (20) using a 5′-nuclease fluorogenic or TaqMan assay (21) has been developed to measure accurately the starting amounts of target sequences. Unlike gel electrophoresis, real-time QPCR has the unique advantage of being a closed-tube system, which can significantly reduce carryover contamination. Using this technique, one can easily monitor and quantitate the accumulation of PCR products during log phase amplification. The TaqMan assay utilizes dual reporter and quencher fluorescent dyes that are attached to a nonextendible probe sequence. During the extension phase of PCR, the 5′-nuclease activity of *Taq* DNA polymerase cleaves the hybridized fluorogenic probe, which releases the reporter signal and is measured during each cycle. In addition to real-time QPCR, TaqMan assays have broad utility for the identification of SNPs.

RELATED NUCLEIC ACID AMPLIFICATION PROCEDURES

Other *in vitro* systems can amplify nucleic acid targets such as the transcription-based amplification system (TAS) (6), its more recent version called the self-sustained sequence replication (3SR) (22) and the ligation-dependent Qβ-replication assay (23). These methods are best suited for the detection and semiquantitation of RNA target sequences. The strategy for TAS and 3SR is a continuous series of reverse transcription and transcription reactions that mimic retroviral replication by amplifying specific RNA sequences via cDNA intermediates. The primers contain 5′- add-on sequences for T7, T3, or SP6 promoters that are incorporated into the cDNA intermediates. The rapid kinetics of transcription-based amplifications is an attractive feature of these systems, which can amplify up to 10^7 molecules in 60 min. Short amplify products, however, which are due to incomplete transcription of the target region and incomplete RNase H digestion of the RNA–DNA hybrids, can be problematic in the TAS and 3SR assays.

Unlike PCR, TAS, or 3SR assays, the ligation-dependent Qβ-replication assay results in the amplification of probe, not target, sequences. This assay utilizes a single hybridization to the target sequence, which is embedded within, and divided between, a pair of adjacently positioned midvariant (MDV-1) RNA probes. MDV-1 RNA is the naturally occurring template for the bacteriophage Qβ RNA replicase. Following the isolation of the probe–target hybrids, ligation of the binary probes creates a full-length amplifiable MDV-1 RNA reporter. When Qβ replicase is added, newly synthesized MDV-1 RNA molecules are amplified from ligated binary probes that originally hybridized to the target sequence (23). Similar to TAS and 3SR, the Qβ-replication assay shows rapid kinetics, generating up to 10^9 molecules in 30 min, and all three methods have been successfully used for the detection and quantitation of HIV-1 RNA molecules.

LIGATION CHAIN REACTION

The ligase chain reaction (LCR) can also amplify short DNA regions of interest by iterative cycles of denaturation

and annealing/ligation steps (24). The LCR utilizes four primers: two adjacent ones that specifically hybridize to one strand of target DNA and a complementary set of adjacent primers that hybridize to the opposite strand. LCR primers must contain a 5′-end phosphate group, such that thermostable ligase (24) can join the 3′-end hydroxyl group of the upstream primer to the 5′-end phosphate group of the downstream primer. Successful ligations of adjacent primers can subsequently act as the LCR template, resulting in an exponential amplification of the target region. The LCR is well suited for the detection of SNPs because a single-nucleotide mismatch at the 3′ end of the upstream primer will not ligate and amplify, thus discriminating it from the correct base. Although LCR is generally not quantitative, linear amplifications using one set of adjacent primers, called the ligase detection reaction, can be quantitative. Coupled to PCR, linear ligation assays can also be used as a mutation detection system for the identification of SNPs using both wild-type-specific and mutant-specific primers in separate reactions. The oligonucleotide ligase assay was first reported to detect SNPs from both cloned and clinical materials using a 5′-end biotin group attached to the upstream primer and a nonisotopic label attached to the downstream primer (25). Allele-specific hybridizations and ligations can be separated by immobilization to a streptavidin-coated solid support and directly imaged under appropriate conditions without the need for gel electrophoretic analysis.

SUMMARY

Some of the general concepts and practices of PCR have been reviewed here. Not only has PCR made a major and significant impact on basic and clinical research, but it has also been well accepted and utilized in forensic science. For any scientific methodology to be accepted in the courts as evidence, it must satisfy four criteria: that the method (*1*) be subject to empirical testing, (*2*) be subject to peer review and publication, (*3*) has a known error rate, and (*4*) is generally accepted in the scientific community. The application of PCR has been admitted in the U.S. courts as evidence in criminal cases for the analysis of human DNA sequences, and in January 1997 as evidence for the phylogenetic analysis of HIV DNA sequences (26). Clearly, the scope of applications for PCR seems endless and it is truly a remarkable technique that has been widely used in molecular biology.

BIBLIOGRAPHY

1. Saiki RK, et al. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. Science 1985;230:1350–1354.
2. Mullis KB, Faloona FA. Specific synthesis of DNA *In vitro* via a polymerase-catalysed chain reaction. Methods Enzymol 1987;155:335–351.
3. Eckert KA, Kunkel TA. High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase. Nucleic Acids Res 1990;18:3739–3744.
4. Gibbs RA, Chamberlain JS. The polymerase chain reaction: a meeting report. Genes Dev 1989;3:1095–1098.
5. Kwoh DY, et al. Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. Proc Nat Acad Sci U.S.A. 1989;86:1173–1177.
6. Kwok S, Higuchi R. Avoiding false positives with PCR. Nature (London) 1989;339:237–238.
7. Saiki RK, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 1988;239:487–491.
8. Tindall KR, Kunkel TA. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. Biochemistry 1988;27:6008–6013.
9. Innis MA, Myambo KB, Gelfand DH, Brow MAD. DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified. Proc Nat Acad Sci U.S.A. 1988;85:9436–9440.
10. Cline J, Braman JC, Hogrefe HH. PCR fidelity of *Pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Res 1996;24:3546–3551.
11. Barnes WM. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. Proc Nat Acad Sci U.S.A. 1994;91:2216–2220.
12. Cheng S, Fockler C, Barnes WM, Higuchi R. Effective amplification of long targets from cloned inserts and human genomic DNA. Proc Nat Acad Sci U.S.A. 1994;91:5695–5699.
13. Chamberlain JS, et al. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acids Res 1988;23:11141–11156.
14. Gibbs RA et al. Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch–Nyhan families. Genomics 1990;7:235–244.
15. Metzker ML, Allain KM, Gibbs RA. Accurate determination of DNA in agarose gels using the novel algorithm GelScann(1.0). Computer App Biosci 1995;11:187–195.
16. Lee CC, et al. Generation of cDNA probes directed by amino acid sequence: cloning of urate oxidase. Science 1988;239:1288–1291.
17. Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Nat Acad Sci U.S.A. 1989;86:1939–1943.
18. Taubenberger JK, et al. Initial genetic characterization of the 1918 'Spanish' influenza virus. Science 1997; 275:1793–1796.
19. Gilliland G, Perrin S, Blanchard K, Bunn HF. Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction. Proc Nat Acad Sci U.S.A. 1990;87:2725–2729.
20. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res 1996;6:986–994.
21. Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5′→3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Nat Acad Sci U.S.A. 1991;88:7276–7280.
22. Guatelli JC, et al. Isothermal, *In vitro* amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. Proc Nat Acad Sci U.S.A. 1990;87:1874–1878.
23. Tyagi S, Landegren U, Tazi M, Lizardi PM, Kramer FR. Extremely sensitive, background-free gene detection using binary probes and Qβ-replicase. Proc Nat Acad Sci U.S.A. 1996;93:5395–5400.
24. Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Nat Acad Sci U.S.A. 1991;88:189–193.
25. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science 1988;241:1077–1080.
26. State of Louisiana v. Richard J Schmidt. Reasons for ruling of Louisiana State 15th Judicial District Court Judge Durwood

Conque. 15th Judicial District Court, Lafayette Parish, Louisiana. Criminal Docket 73313. 1997.

Further Reading

Erlich HA, Gelfand D, Sninsky JJ. Recent advances in the polymerase chain reaction. Science 1991; 252: 1643–1650.

Innis MA, Gelfand DH, Sninsky JJ, White TJ, editors. PCR Protocols: A Guide to Methods and Applications. Academic; San Diego: 1990.

Mullis KB, Ferré F, Gibbs RA, editors. The Polymerase Chain Reaction. Birkhäuser; Boston: 1994.

See also ANALYTICAL METHODS, AUTOMATED; DNA SEQUENCE; MICROARRAYS.

POLYMERIC MATERIALS

XIAOHUA LIU
The University of Michigan
Ann Arbor, Michigan

SHOBANA SHANMUGASUNDARAM
TREENA LIVINGSTON ARINZEH
New Jersey Institute of Technology
Newark, New Jersey

INTRODUCTION

This article aims to provide basic and contemporary information on polymeric materials used in medical devices and instrumentation. The fundamental concepts and features of polymeric materials are introduced in the first section. In the second section, the major commodity polymers used in medicine are reviewed in terms of their basic chemical and physical properties. The main part of this article, however, is devoted to polymers in biomedical engineering applications, including tissue engineering and drug delivery systems.

Polymers are a very important class of materials. A polymer can be defined as a long-chain molecule that is composed of a large number of repeating units of identical structure. Some polymers, (e.g., proteins, cellulose, and starch) are found in Nature, while many others, including polyethylene, polystyrene, and polycarbonate, are produced only by synthetic routes. Hundreds of thousands of polymers have been synthesized since the birth of polymer science. Today, polymeric materials are used in nearly all areas of daily life.

Polymers can simply be divided into two distinct groups based on their thermal processing behavior: thermoplastics and thermosets. Thermoplastics are linear or branched polymers, and they soften or melt when heated, so that they can be molded and remolded by heating. This property allows for easy processing and recycling. In comparison, thermosets are three-dimensional (3D) network polymers, and cannot be remelted. Once these polymers are formed, reheating will cause the material to scorch.

In addition to classification based on processing characteristics, polymers may also be grouped based on the chemical structure of their backbone. Polymers with one identical repeating unit in their chains are called homopolymers. The term copolymer is often used to describe a polymer with two or more repeating units. The sequence of repeating units along the polymer chain can form different structures, and copolymers can be further classified as random copolymers, alternating copolymers, block copolymers, and graft copolymers. In random copolymers, the sequence distribution of the repeating units is random, while in alternating copolymers the repeating unit are arranged alternately along the polymer chain. A block copolymer is one in which identical repeating units are clustered in blocks along the chain. In graft copolymers, the blocks of one type of repeating unit are attached as side chains to the backbone chains.

Unlike simple pure compounds, most polymers are not composed of identical molecules. A typical synthetic polymer sample contains chains with a wide distribution of chain lengths. Therefore, polymer molecular weights are usually given as averages. The number average molecular weight (M_n), which is calculated from the mole fraction distribution of different sized molecules in a sample, and the weight average molecular weight (M_w), which is calculated from the weight fraction distribution of different sized molecules, are two commonly used values. The statistical nature of polymerization reaction makes it impossible to characterize a polymer by a single molecular weight. A measure of the breadth of the molecular weight distribution is given by the ratios of molecular weight averages. The most commonly used ratio is M_w/M_n. As the weight dispersion of molecules in a sample narrows, M_w approaches M_n, and in the unlikely case that all the polymer molecules have identical weights, the ratio M_w/M_n becomes unity. Most commercial polymers have the molecular weight distribution of 1.5–10. In general, increasing molecular weight corresponds to increasing physical properties and decreasing polymer processability.

In many cases, individual polymer chains are randomly coiled and interviewed with no molecular order or structure. Such a physical state is termed amorphous. Amorphous polymers exhibit two distinctly different types of mechanical behavior. Some, like poly(methyl methacrylate, PMMA) and polystyrene are hard, rigid, glassy plastics at room temperature, while others, like polybutadiene and poly(ethyl acrylate), are soft, flexible, rubbery materials at room temperature. There is a temperature, or range of temperatures, below which an amorphous polymer is in a glassy state, and above which it is rubbery. This temperature is called the glass transition temperature (T_g). The value of T_g for a specific polymer will depend on the structure of the polymer. Side groups attached to the polymer chain will generally hinder rotation in the polymer backbone, necessitating higher temperatures to give enough energy to enable rotation to occur.

For most polymers, the T_g constitutes their most important mechanical properties. At low temperatures ($< T_g$), an amorphous polymer is glass-like, with a value of Young's modulus in the range of 10^9–10^{10} Pa, and it will break or yield at strains greater than a few percent. When the temperature is $> T_g$, the polymer becomes rubber-like, with a modulus in the range of 10^6–10^8 Pa, and it may withstand large extensions with no permanent deformation. At even

ATTACHMENT 6A

LDR	01964pam a22006014a 4500
001	9912587523406676
005	20240608231539.0
008	051004s2006 njua be 001 0 eng
010	__ | **a** 2005028946
020	__ | **a**9780471263586 (set : cloth)
020	__ | **a**0471263583 (set : cloth)
020	__ | **a**9780470040669 (v. 1 : cloth)
020	__ | **a**0470040661 (v. 1 : cloth)
020	__ | **a**9780470040676 (v. 2 : cloth)
020	__ | **a**047004067X (v. 2 : cloth)
020	__ | **z**047004067x (v. 2 : cloth)
020	__ | **a**9780470040683 (v. 3 : cloth)
020	__ | **a**0470040688 (v. 3 : cloth)
020	__ | **a**9780470040690 (v. 4 : cloth)
020	__ | **a**0470040696 (v. 4 : cloth)
020	__ | **a**9780470040706 (v. 5 : cloth)
020	__ | **z**97804070040706 (v. 5 : cloth)
020	__ | **a**047004070X (v. 5 : cloth)
020	__ | **z**047104070X (v. 5 : cloth)
020	__ | **a**9780470040713 (v. 6: cloth)
020	__ | **a**0470040718 (v. 6 : cloth)
035	__ | **9**101258752
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
042	__ | **a**pcc
044	__ | **9**United States
060	00 | **a**W 13**b**E555 2006
245	00 | **a**Encyclopedia of medical devices and instrumentation /**c**edited by John G. Webster.
246	31 | **a**Medical devices and instrumentation
250	__ | **a**2nd ed.
260	__ | **a**Hoboken, NJ :**b**Wiley-Interscience,**c**c2006.
300	__ | **a**6 v. :**b**ill.
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
500	__ | **a**Rev. ed. of: Encyclopedia of medical devices and instrumentation. 1988.
504	__ | **a**Includes bibliographical references and index.
650	12 | **a**Equipment and Supplies
655	_2 | **a**Encyclopedia
700	1_ | **a**Webster, John G.,**d**1932-2023.

935	__ | **a**(DNLM)1258752-nlmdb
992	__ | **p**P1**e**E4**a**20060324
993	__ | **a**AXE**b**20060328
994	__ | **a**CDN**b**20060328
995	__ | **a**AUTH**b**20051005**c**REV**d**20181116
996	__ | **a**rev. CIP**b**20060328
999	__ | **a**AUTH
035	__ | **a**(OCoLC)61860839

ATTACHMENT 6B

Equipment and Supplies MeSH Descriptor Data 2025

MeSH Heading
Equipment and Supplies
Tree Number(s)
E07
Unique ID
D004864
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D004864
Annotation
general or unspecified; prefer /instrum
Scope Note
Expendable and nonexpendable equipment, supplies, apparatus, and instruments that are used in diagnostic, surgical, therapeutic, scientific, and experimental procedures.
Entry Version
EQUIPMENT SUPPLIES
Entry Term(s)
Apparatus and Instruments
Device, Medical
Devices
Devices, Medical
Equipment
Inventories
Medical Device
Medical Devices
Supplies
See Also
Catalogs, Commercial as Topic
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2018/06/29

Connect with NLM
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
Web Policies
FOIA
HHS Vulnerability Disclosure

Target-enrichment strategies for next-generation sequencing

Lira Mamanova[1], Alison J Coffey[1], Carol E Scott[1], Iwanka Kozarewa[1], Emily H Turner[2], Akash Kumar[2], Eleanor Howard[1], Jay Shendure[2] & Daniel J Turner[1]

We have not yet reached a point at which routine sequencing of large numbers of whole eukaryotic genomes is feasible, and so it is often necessary to select genomic regions of interest and to enrich these regions before sequencing. There are several enrichment approaches, each with unique advantages and disadvantages. Here we describe our experiences with the leading target-enrichment technologies, the optimizations that we have performed and typical results that can be obtained using each. We also provide detailed protocols for each technology so that end users can find the best compromise between sensitivity, specificity and uniformity for their particular project.

© 2010 Nature America, Inc. All rights reserved.

The ability to read the sequence of bases that comprise a polynucleotide has had an impact on biological research that is difficult to overstate. For the majority of the past 30 years, dideoxy DNA 'Sanger' sequencing[1] has been used as the standard sequencing technology in many laboratories, and its acme was the completion of the human genome sequence[2]. However, because Sanger sequencing is performed on single amplicons, its throughput is limited, and large-scale sequencing projects are expensive and laborious: the human genome sequence took hundreds of sequencing machines several years and cost several hundred million dollars.

The paradigm of DNA sequencing changed with the advent of 'next-generation' sequencing technologies (reviewed in refs. 3,4), which process hundreds of thousands to millions of DNA templates in parallel, resulting in a low cost per base of generated sequence and a throughput on the gigabase (Gb) scale. As a consequence, we can now start to define the characteristics of entire genomes and delineate differences between them. Ultimately, whole-genome sequencing of complex organisms will become routine, allowing us to gain a deeper understanding of the full spectrum of genetic variation and to define its role in phenotypic variation and the pathogenesis of complex traits.

Nevertheless, it is not yet feasible to sequence large numbers of complex genomes in their entirety because the cost and time taken are still too great. To obtain 30-fold coverage of a human genome (90 Gb in total), would currently require several sequencing runs and would cost tens of thousands of dollars. In addition to the demands such a project would place on laboratory time and funding, the primary analysis during which the captured image files are processed, as well as storage of the sequences, would place a substantial burden on a research center's informatics infrastructure.

Consequently, considerable effort has been devoted to develop 'target-enrichment' methods, in which genomic regions are selectively captured from a DNA sample before sequencing. Resequencing the genomic regions that are retained is necessarily more time- and cost-effective, and the resulting data are considerably less cumbersome to analyze. Several approaches to target enrichment have been developed (**Fig. 1**), and there are several parameters by which the performance of each can be measured, which vary from one approach to another: (i) sensitivity, or the percentage of the target bases that are represented by one or more sequence reads; (ii) specificity, or the percentage of sequences that map to the intended targets; (iii) uniformity, or the variability in sequence coverage across target regions; (iv) reproducibility, or how closely results obtained from replicate experiments correlate; (v) cost; (vi) ease of use; and (vii) amount of DNA required per experiment, or per megabase of target.

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.S. (shendure@u.washington.edu) or D.J.T. (djt@sanger.ac.uk).

PUBLISHED ONLINE 28 JANUARY 2009; DOI:10.1038/NMETH.1419

EX1020

A technology that typically has a high specificity and uniformity will require less sequencing to generate adequate coverage of sequence data for the downstream analysis, making the sequencing more economical. In addition to these factors, when assessing which target-enrichment technology is the most appropriate for a particular project, thought must be given to how well matched each method is to the total size of intended target region, the number of samples (**Fig. 2**) and whether or not sample multiplexing is required to most efficiently use sequencer throughput.

Here we describe the most widely used approaches to target enrichment, our experiences with each and the optimizations that we have performed. We also provide detailed protocols, which we have developed with the aim of finding the best compromise between the parameters described above.

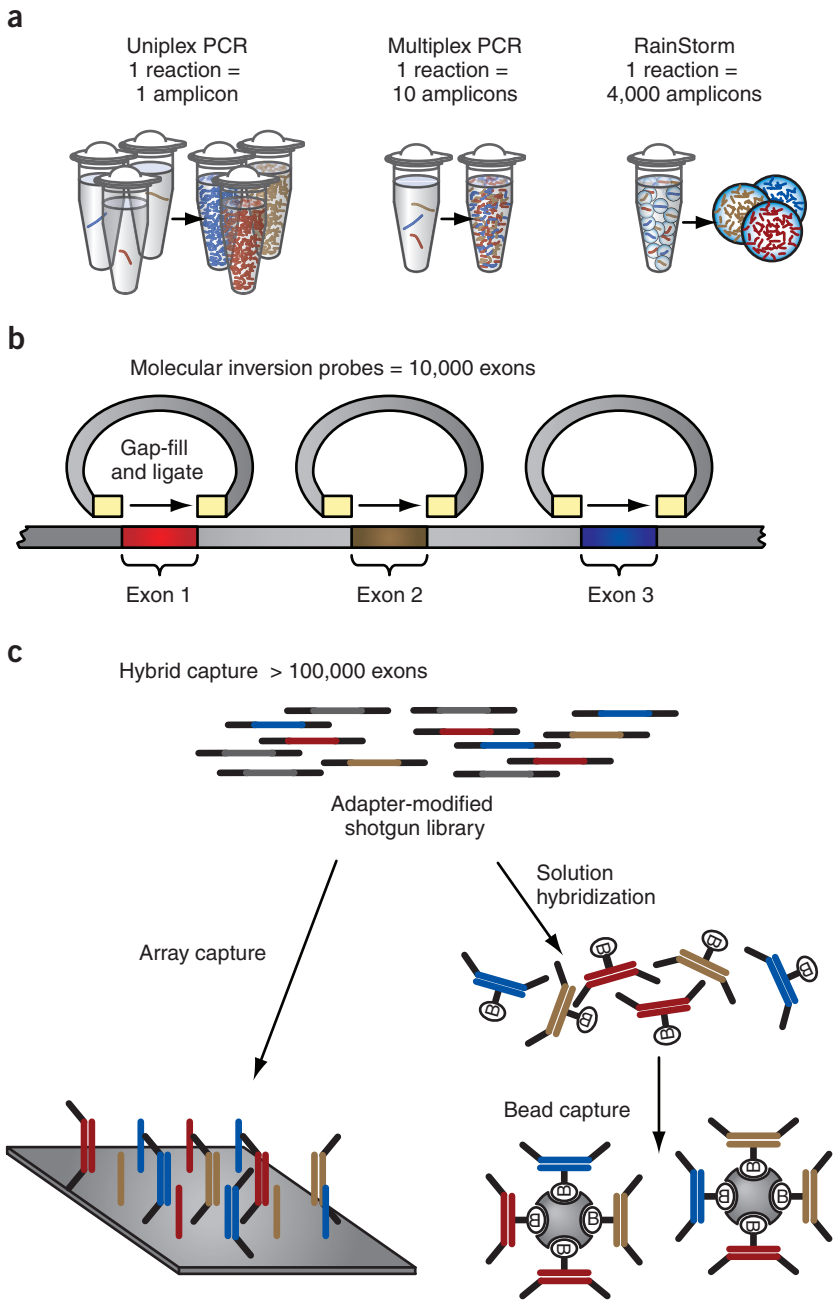


Figure 1 | Approaches to target enrichment. (**a**) In the uniplex PCR–based approach, single amplicons are generated in each reaction. In multiplexed PCR, several primer pairs are used in a single reaction, generating multiple amplicons. On the RainStorm platform, up to 4,000 primer pairs are used simultaneously in a single reaction. (**b**) In the MIP-based approach, probes consisting of a universal spacer region flanked by target-specific sequences are designed for each amplicon. These probes anneal at either side of the target region, and the gap is filled by a DNA polymerase and a ligase. Genomic DNA is digested, and the target DNA is PCR-amplified and sequenced. (**c**) In the hybrid capture–based approach, adaptor-modified genomic DNA libraries are hybridized to target-specific probes either on a microarray surface or in solution. Background DNA is washed away, and the target DNA is eluted and sequenced.

PCR

PCR has been the most widely used presequencing sample preparation technique for over 20 years[5], and it is particularly well suited to a Sanger sequencing–based approach, in which a single PCR can be used to generate a single DNA sequence and in which the sequence read length is comparable to that of a typical PCR amplicon. PCR is also potentially compatible with any next-generation sequencing platform, though to make full use of the high throughput, a large number of amplicons must be sequenced together. However, PCR is difficult to multiplex to any useful degree: the simultaneous use of many primer pairs can generate a high level of nonspecific amplification, caused by interaction between the primers, and moreover amplicons can fail to amplify[6,7]. Clever derivatives of multiplex PCR have been developed[8–10], but in practice, it is often more straightforward to perform PCRs in uniplex. Additionally, there is an upper limit to the length of amplicon that can be generated by long PCR[11]: in our experience very long PCRs tend to lack robustness, and for PCR amplification of contiguous regions, we prefer to design overlapping PCRs that are no more than 10 kilobases (kb) long. Each individual PCR must be validated and, ideally, optimized to make amplification as efficient as possible to minimize the total mass of DNA required.

After amplification, the concentration of products must be normalized before pooling to avoid sequencing one dominant PCR product above all others. There are several ways to approach normalization at this stage, but the most reliable way is to visually inspect the intensity of bands on an agarose gel, alongside a quantitative ladder. Consequently, there is an upper limit to the size of genomic target that can realistically be selected by PCR because of the workload involved. We recommend using long PCR to target regions that are up to several hundred kilobases long, as this is feasible both from the perspectives of workload and the quantity of DNA required.

By current standards, a single lane of a paired-end, 76-base sequencing run

© 2010 Nature America, Inc. All rights reserved.

would generate an average coverage of about 30,000-fold for a 100-kb target, clearly a massive excess. For the sequencing to be economical, it is necessary to barcode and pool many samples and to sequence these pools in a single lane. Several approaches to sample barcoding have been reported[12–14], but we have found ligation of barcodes to fragmented PCR amplicons to give uneven sequence coverage of different samples.

We developed a protocol for barcoding 96 samples, in which the library is prepared in 96-well plates and the barcode is included in the central region of the reverse PCR primer (**Supplementary Protocol 1**). We validated this strategy by analyzing a 25-kb region in DNA from several human populations worldwide. We sequenced 96 libraries per flowcell lane and generated 50-base paired-end sequence reads, with an additional 8 bases of sequence to generate the tag sequences. Sequence data from this study have been deposited in the European Short Read Archive. The average coverage obtained from these sequences was high: median >225-fold per lane for native DNA, and 175-fold for whole genome–amplified samples. Coverage and uniformity was poorer for whole genome–amplified samples than for genomic DNA, especially for the longest amplicon in the pool, suggesting that biases were introduced during whole-genome amplification, as has been noted previously[15,16]. However, the barcoding approach was successful, with 80% of sequenced bases covered within a twofold range of the median for the genomic samples. We called single-nucleotide polymorphisms (SNPs) at >99% of sites in approximately 98% of samples and detected 63 high-confidence SNPs; 27 of them were new and 23 were rare.

Improvements for PCR

Although this PCR-based approach was highly effective, there are several areas in which it could be improved. First, a reduction in the cost of library preparation reagents would have a major impact on the overall cost because a separate sequencing library is required for each DNA sample, making library preparation very expensive, for even a small number of lanes of sequencing. Second, improvement in the accuracy of pooling the tiled amplicons, which impacts sequence uniformity, is needed because quantifying tiled amplicons by quantitative PCR is still difficult to achieve for tens to hundreds of amplicons per sample. Third, the use of 5′-blocked primers would achieve greater sequence uniformity across amplicons[14]. Fourth, the use of a greater depth of tiling in the PCRs is another area for improvement. The failure of long PCRs has a major impact on coverage uniformity, but if every base in the target locus is covered by at least two overlapping PCRs, failure of one of these PCRs will not result in the 'loss' of that base. Finally, the use of error-correcting barcodes would allow a greater proportion of pooled sequences to be deconvoluted[17]. Using Hamming codes for tag design[18], it is possible to make tagsets in which single nucleotide-sequencing errors can be corrected, and in which two errors and single insertion-deletions can be detected unambiguously (**Supplementary Table 1**).

It is possible to design long PCR primers for close to 100% of desired targets, but in practice, not all reactions will yield a product after amplification. This can be problematic for samples in which the integrity of the DNA is low, such as clinical specimens. Similarly, when there are SNPs in the primer annealing regions, one allele may be amplified preferentially[19]. Such difficulties can usually be overcome by optimization, primer redesign, greater tiling of amplicons or using a combination of long and short PCR.

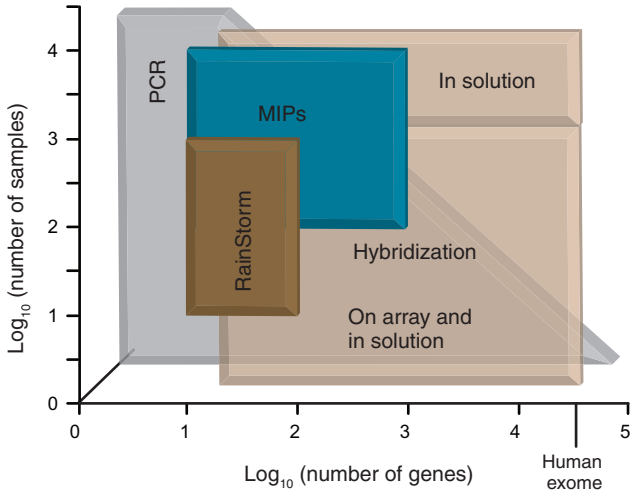


Figure 2 | Suitability of different target-enrichment strategies to different combinations of target size and sample number. Suitability was estimated from the perspective of the feasibility with which each method could be applied to the various combinations of target size and sample number, rather than the cost.

The RainStorm platform, developed by RainDance Technologies, is a convenient solution to many of the problems encountered in a standard PCR–based approach (http://www.raindancetechnologies.com/applications/next-generation-sequencing-technology.asp/). The technology uses microdroplets, similarly to emulsion PCR[20,21]. Each droplet supports an independent PCR and can be made to contain a single primer pair along with genomic DNA and other reagents. The entire population of droplets represents hundreds to thousands of distinct primer pairs and is subjected to thermal cycling, after which this emulsion is broken and products are recovered. The mixture of DNA amplicons can then be subjected to shotgun library construction and massively parallel sequencing. During the microdroplet PCR, different primer pairs cannot interact with each other, which removes one of the primary constraints on conventional multiplex PCR. The microdroplet approach also prevents direct competition of multiplex PCRs for the same reagent pool, which should improve uniformity relative to conventional multiplex PCR. The current maximum number of primer pairs that can be used is 4,000, though it is expected that the number will reach 20,000 by mid 2010 (J. Lambert, personal communication).

The proof of concept for this approach has been published recently[22]. In one experiment, the authors targeted 457 amplicons of variable size (119–956 bp) and G+C content (24–78%), totalling to 172 kb. In six samples, 84% of uniquely mapping reads aligned to targeted amplicons, and 90% of targeted bases were represented within a 25-fold abundance range. In a second experiment, they targeted 3,976 amplicons representing an aggregate target of 1.35 megabases (Mb) and observed that 79% of uniquely mapping reads aligned to targets and 97% of targeted bases were covered within a 25-fold abundance range. The specificity and uniformity of the approach compare well with those of the alternatives, and base calling demonstrated good concordance with expected HapMap genotypes. One limitation is that the approach currently has relatively high input requirements (7.5 μg per sample), but this may be reduced with optimization. In terms of the flexibility of targeting, it is reasonable to expect that this approach will have advantages and disadvantages analogous to those of conventional PCR primer design.

© 2010 Nature America, Inc. All rights reserved.

Table 1 | Performance of target-enrichment methods

	PCR	MIP	On-array hybrid capture	In-solution hybrid capture
Cost	High	<10 samples, high; >100 samples, low	Medium	<10 samples, medium; >10 samples, low
Ease of use	Low	High	Medium	High
Mass DNA	~8 µg for 1 Mb of 2× tiled, 5 kb amplicons	As little as 200 ng	10-15 µg per array for up to 30 Mb target	3 µg for up to 30 Mb target
Sensitivity	>99.5%	>98%, with stringent design constraints	98.6% of CTR[a]	>99.5% of CTR[a]
Specificity	93% for HapMap DNA samples, 72% for whole genome–amplified samples	>98%	Up to 70% mapping to CTR[a] for exons; higher for contiguous regions	Up to 80% mapping to CTR[a] for exons; higher for contiguous regions
Uniformity	80% of bases within twofold range of median	58% of CTR within tenfold coverage range; 88% within 100-fold coverage range	60% of CTR[a] within 0.5–1.5-fold of mean coverage (mapping quality[b] 30)	61% of CTR[a] within 0.5–1.5-fold of mean coverage (mapping quality[b] 30)
Reproducibility	Up to 100%	0.92 rank-order correlation[c]	For 10^7 paired-end sequences, >95% reproducibility at tenfold between two samples	For 10^7 paired end-sequences, >96% reproducibility at tenfold between two samples

[a]CTR, capture target region, that is, the regions of the desired target region to which probes could be designed after repeat masking. [b]Mapping qualities were calculated by the mapping software, MAQ, and indicate the probability that the mapping location is correct. A score of 30 or greater indicates that the quality of a read was good, and that it mapped unambiguously to that location with few mismatches. [c]Rank-order correlation in capture efficiency distributions between independent samples.

© 2010 Nature America, Inc. All rights reserved.

npg

Even with an efficient, automated PCR pipeline, it is not feasible to use conventional PCR to target genomic regions that are several megabases in size because of the high cost of primers and reagents and the DNA input requirements, particularly in large sample sets (**Fig. 2**). Similarly, there is a limit to the maximum target size that can be selected using the RainStorm platform (2–3 Mb), and its sample throughput is limited to approximately 8 per workday (**Fig. 2**). Consequently, for very large target regions such as the approximately 30 Mb human exome, or to select moderately sized regions in very large numbers of samples, other approaches to target enrichment should be used.

MOLECULAR INVERSION PROBES

Various enzymatic methods for targeted amplification are compatible with extensive multiplexing based on target circularization[23–25]. One approach in the latter category relies on the use of molecular inversion probes (MIPs), which initially had been developed for multiplex target detection and SNP genotyping[26–30]. Single-stranded oligonucleotides, consisting of a common linker flanked by target-specific sequences[31,32], anneal to their target sequence and become circularized by a ligase. Uncircularized species are digested by exonucleases to reduce background, and circularized species are PCR amplified via primers directed at the common linker. To adapt this method to perform exon capture in combination with next-generation sequencing, a DNA polymerase can be used to 'gap-fill' between target-specific MIP sequences designed to flank a full or partial exon, before ligase-driven circularization, thereby capturing a copy of the intervening sequence[24]. The assay initially demonstrated low uniformity, largely owing to inefficiencies in the capture reaction itself, but more recently an optimized, simplified protocol for MIP-based exon capture has been reported[33]. This revised protocol (**Supplementary Protocol 2**) retains the high specificity of MIP capture, with >98% of mapped reads aligning to a targeted exon but additionally, uniformity is markedly improved, with 58% of targeted bases in 13,000 targets captured to within a tenfold range and 88% to within a 100-fold range (**Fig. 3a** and **Table 1**).

The improved capture uniformity resolves the issue of stochastic allelic bias that plagued the initial proof of concept, showing that accurate genotypes can be derived from massively parallel sequencing of MIP capture products. Furthermore, MIP amplification products can be directly sequenced on a next-generation sequencing platform to interrogate variation in targeted sequences, thereby bypassing the need for shotgun library construction.

Our current view is that the approach of MIP-based capture followed by direct sequencing may be most relevant for projects involving relatively small numbers of targets but large numbers of samples (**Fig. 2**). This is based on the following characteristics. (i) Gap-fill reactions and PCRs take place in aqueous solution, in small volumes, so they are easy to scale to large numbers of samples on 96-well plates; no mechanical shearing, gel-based size purification, ligation or A-tailing is required. (ii) Sample-identifying barcodes can be nested in one of the primers used in post-capture amplification, allowing products from multiple samples to be pooled and sequenced in a single lane. (iii) As with PCR, capture is performed directly on genomic DNA rather than after conversion to a shotgun library, reducing input requirements to as low as 200 ng[34].

The main disadvantages of using MIPs for target enrichment are, first, that capture uniformity, though markedly improved, compares poorly with the most recent reports on capture by hybridization and is the foremost challenge for the approach. To help circumvent this, MIPs can potentially be grouped into sets based on similar capture efficiencies because biases tend to be systematically reproducible[34]. Also, modeling of the causes of nonuniformity can be fed back to MIP design algorithms. Second, MIP oligonucleotides can be costly and difficult to obtain in large numbers to cover large target sets. To mitigate the high cost of column-based oligonucleotide synthesis, thousands of oligos can be obtained by synthesis and release from programmable microarrays (Agilent[24]; LC Sciences). Provided that these are designed in an amplifiable format, they can potentially be used to generate MIP probes to support thousands of samples. Alternatively, one can undertake column-based synthesis of individual MIPs followed by pooling. Although the initial cost for this can be high, sufficient material is obtained to support an extraordinarily large number of capture reactions[24]. The availability of individual probes would also facilitate empirical repooling to improve capture uniformity (J.S.; unpublished data). Finally, it is worth noting that MIPs offer flexibility to address

a range of related applications, for example, DNA methylation, RNA editing and allelic imbalance in expression[34–36].

HYBRID CAPTURE

On-array capture

The principle of direct selection is well-established[37,38]: a shotgun fragment library is hybridized to an immobilized probe, nonspecific hybrids are removed by washing and targeted DNA is eluted. Roche NimbleGen and their collaborators were the first to adapt the technology to be compatible with next-generation sequencing[15,16,39]. In the original format, library DNA is hybridized to a single microarray containing 385,000 isothermal probes (the HD1 NimbleGen array), ranging from 60 to 90 bases in length, and with a total capture size of around 4–5 Mb. More recently, the HD2 array has been made available, with 2.1 million probes per array and the ability to capture up to 34 Mb on a single array (**Fig. 2**). The technology was originally designed to be used with the Roche 454 sequencer, but many groups, including ours, expended a considerable amount of effort to modify and optimize protocols for use with the Illumina Genome Analyzer. Agilent's Capture Arrays and comparative genomic hybridization (CGH) arrays are perhaps the most direct competitor to NimbleGen's HD1 arrays, though Agilent's Capture Arrays contain only 244,000 probes on the surface (10^6 for CGH arrays). We found that the performance of both NimbleGen's and Agilent's arrays is similar (**Table 1**).

There are clear advantages to on-array target enrichment of large regions over PCR-based approaches: it is far quicker and less laborious than PCR. But there are also drawbacks: working with microarray slides requires expensive hardware, such as a hybridization station. Additionally, the limit to the number of arrays that a single person can realistically perform each day is approximately 24. As arrays that are hybridized at the same time must also be eluted together, studies with very large numbers of samples are unfeasible. Finally, to have enough DNA library for a target-enrichment experiment, it is necessary to start library preparation with a relatively large amount of DNA, around 10–15 μg, though this is irrespective of whether the capture experiment is for 100 kb or an entire exome.

In-solution capture

To overcome many of these disadvantages, both Agilent and NimbleGen have also developed solution-based target-enrichment protocols. The general principle is similar to array capture, in that there are specific probes designed to target regions of interest from a sequencing library, but whereas an on-array target enrichment uses a vast excess of DNA library over probes, solution capture has an excess of probes over template, which drives the hybridization reaction further to completion using a smaller quantity of sequencing library[40]. In our experiments to test the performance of array versus solution capture, we observed that for smaller target sizes (~3.5 Mb), the uniformity and specificity of sequences obtained from a solution capture experiment tend to be slightly higher than that of array capture (**Fig. 3b,c**). Thus in the 3.5-Mb range, solution capture yields superior sequence coverage of the target regions from a similar yield of sequences. However, for whole-exome captures, both solution and array appear to perform equivalently (**Fig. 4**).

In-solution target enrichment can be performed in 96-well plates, using a thermal cycler, so it is more readily scalable than on-array enrichment and does not require specialized equipment (**Fig. 2**). The principal difference between the Agilent and NimbleGen solution capture products is the nature of the capture probes: the NimbleGen product uses 60–90-mer DNA capture probes, whereas the Agilent one uses 150-mer RNA capture probes. We have not noticed any appreciable difference between the performance of each product.

Library preparation for hybrid capture

Our aim has been to establish a robust production pipeline that can support both on-array and in-solution target enrichment. The manufacturers' workflows for these approaches are very similar and several general principles apply, which allowed us to produce a standard library preparation protocol for both approaches (**Supplementary Protocol 3**).

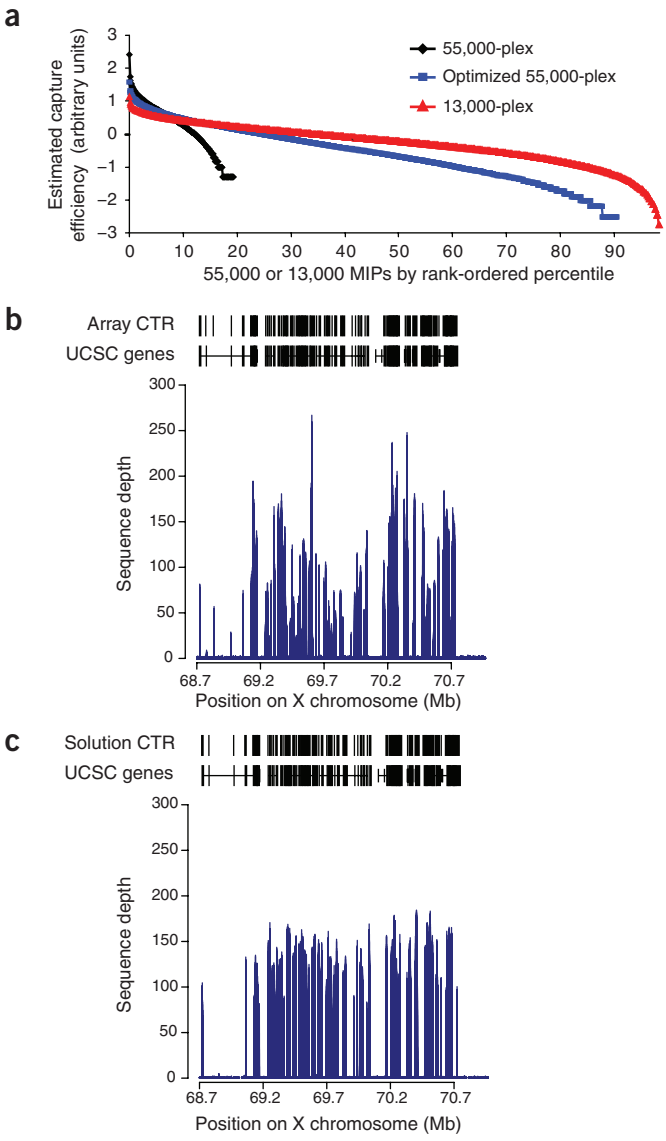


Figure 3 | Uniformity of approaches to target enrichment. (**a**) Capture efficiency obtained with the MIP-based approach, showing improvements in uniformity for optimized protocols and reduced target size. Image was adapted from ref. 33. (**b,c**) A region of human chromosome X, detailing the regions to which capture probes could be designed: the capture target region (CTR) for array capture (**b**) and for solution capture (**c**). Below the CTR are the UCSC genes, taken from the UCSC genome browser. Below this sequence depth obtained from a single lane of Illumina sequences, for a 3.5 Mb capture experiment, is shown.

© 2010 Nature America, Inc. All rights reserved.

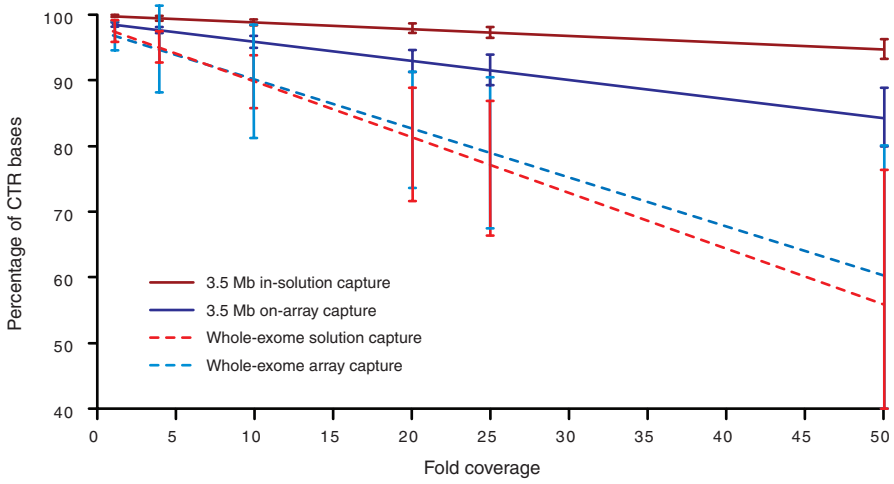


Figure 4 | Coverage plot for array and solution hybrid capture, for 3.5 Mb of exonic target and whole human exome. Values were taken from five independent array and solution experiments, using the same CTR, with each capture using a different DNA sample, and each yielding roughly 10^7 mappable sequences per lane. One lane of sequencing was used for 3.5 Mb captures, whereas two or three lanes were used for the whole exome. Error bars, s.d. ($n = 5$).

Fragment size. Fragment size, obtained by shearing or other fragmentation approaches, has a large influence over the outcome of a target-enrichment experiment, with shorter fragments invariably being captured with higher specificity than longer ones[40,41]. This is not necessarily surprising, given that a longer fragment will contain a higher proportion of off-target sequence, and the effect is especially apparent for exons, whose mean length is relatively short: 164 bp[40] (for example, a 100-bp exon that is part of a 200-bp fragment will be 50% off target just because the captured fragment is larger). However, in our experiments comparing hybrid-capture protocols, the decrease in specificity with increasing template size that we observed was more pronounced than could be accounted for by just the inclusion of off-target portions of longer template sequences and presumably reflects the increase in potential for cross-hybridization between longer fragments themselves.

We assume that there is also a lower size limit to fragments for efficient capture, but in practice the minimum fragment size is determined by the length one would wish to sequence. Longer reads would be expected to map to the reference sequence with lower ambiguity than shorter reads and can help to reduce overrepresentation toward the end of capture probes[40]. For target enrichment of human DNA, we typically generate 76-base paired-end reads, and consequently, it is useful to generate fragments that are around 200 bp to avoid overlap between reads 1 and 2 (**Supplementary Protocol 3**).

Target enrichment sample preparation protocols include a size-selection step to generate a narrow fragment size range, as this is assumed to assist with read mapping. However, this step is not compatible with a high-throughput workflow because it is too labor-intensive, and, in any case, many read-mapping software packages first align each read and then pair the reads[42,43], requiring only a maximum allowed insert size. A score or mapping quality is then assigned to the reads to indicate the probability that the reads are assigned to the incorrect location. Therefore, we investigated the effect of omitting this gel-based size-selection step by performing sequence-capture experiments on libraries prepared with and without this step. Using acoustic shearing, we could generate a sufficiently narrow fragment-size distribution that the size-selection step can be omitted (**Fig. 5a**), and when mapped using MAQ[43], we found little difference between libraries made with or without a size-selection step. In a single experiment, the percentage of mapped reads with score ≥30 (indicating that the base quality of the reads is good and that the read maps unambiguously to the selected location with few mismatches) was just over 1% lower for a library made without size selection compared to the same library after size selection (89.9 and 88.7%, respectively).

PCR optimization. The use of acoustic shearing and removal of the size-selection step resulted in a greater mass of DNA being available for target enrichment than when standard approaches are used, and this allowed us to investigate the effect of performing PCR amplification at different stages of the target-enrichment process.

We noted a negative influence of PCR amplification on the uniformity of enrichment in both on-array and in-solution methods: performing 18 cycles of PCR amplification of libraries both before and after hybridization can introduce severe bias toward neutral G+C content in the resulting sequences (**Fig. 5b**). Avoiding the PCR step altogether before hybridization greatly improved the situation (**Fig. 5c**), so it is desirable to keep PCR amplification to a minimum and only perform it after hybridization.

However, an amplification-free library preparation tends to lack robustness, especially with samples of lower integrity, such as clinical specimens, compared to intact DNA. In these cases, we recommend around six cycles of amplification before hybridization and the use of blocking adapters[41] to avoid a reduction in specificity caused by random concatenation of libraries, so-called 'daisy-chaining'. If no PCR is performed before hybridization, there is no need to use blocking adapters if sequencing is performed on Illumina's Genome Analyzer because the pre-PCR adapters are partially noncomplementary[44] and are thus not problematic in this way. We recommend that hybrid capture be performed following the manufacturers' standard protocols (**Supplementary Protocols 4**,**5**) and that 14–18 cycles of PCR be performed on the samples eluted after hybridization (**Supplementary Protocol 6**).

Prehybridization cleanup. Addition of commercial preparations of C_0t1 DNA to the hybridization reaction is reported to increase specificity[40,41]. C_0t1 DNA comprises short fragments (50–300 bp) of human placental DNA that is enriched for repetitive sequences. Thus it is capable of hybridizing to repetitive sequences in the library DNA, rendering them inert during target enrichment. C_0t1 DNA is generally added in a 5–20-fold excess over the input library. We have observed little difference in performance within this range but typically use a fivefold excess for on-array target enrichment and a 20-fold excess in solution.

Salt concentration is an important factor in determining the specificity and efficiency of hybridization. Any salts in the C_0t1 and library DNA buffers will contribute to the overall salt

© 2010 Nature America, Inc. All rights reserved.

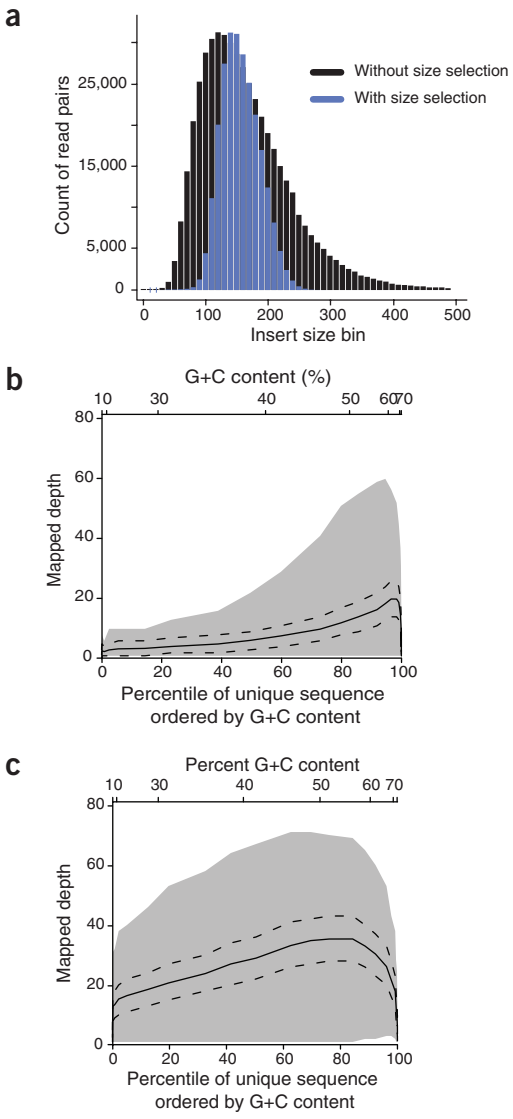


Figure 5 | Library prep optimizations for hybrid capture. (**a**) Distribution of insert sizes derived from mapped sequence data for solution capture performed with and without agarose gel–based size selection. (**b**,**c**) G+C content plot showing mapped sequence data for a 3.5 Mb array capture, in which PCR was performed before and after hybridization (**b**), or in which PCR was performed only after hybridization (**c**). The solid black line indicates the mean value, the dotted lines at either side indicate the s.d., and the shaded area shows the distribution of reads with the indicated G+C content.

concentration in the hybridization buffer, and so we prefer to desalt both the C_0t1 and library DNA before hybridization. A convenient way to achieve this is using solid-phase reversible immobilization (SPRI) beads. These are paramagnetic beads to which nucleic acids can bind reversibly, and captured DNA can be eluted in water[45] (**Supplementary Protocol 3**).

Improvements for hybrid capture

Using **Supplementary Protocols 3** and **6**, we have been able to obtain robust, reproducible target-enrichment results, both on array and in solution, allowing us to transfer target enrichment into a production environment. The turnaround time for synthesis of custom capture arrays (1 or 2 weeks) is typically shorter than for solution probes (~4 weeks), though this is likely to improve as solution probes become more established as a commercial product. We also found in-solution target enrichment to have an equivalent or slightly better performance than on-array enrichment (**Figs. 3b**,**c**, **4** and **Table 1**), and that the former was the more attractive option for high-throughput target enrichment.

Array and solution hybridization are sensitive to sample base composition, and sequences at the extremes of high A+T or G+C content can be lost through poor annealing and secondary structure, respectively. Although not a major issue for human exonic DNA, this sensitivity could be more problematic for other genomes. Another consideration is that it is seldom possible to capture all of a desired target region in a hybrid capture experiment: targets are generally subjected to repeat masking before probe design to avoid capture of homologous repetitive elements. For exonic targets, <5–15% of the primary target region can be lost in this way, leaving a region to which probes could be designed after repeat masking, or target capture region, that constitutes 85% to >95% of the primary target region. For contiguous regions, the percentage of primary target region that is represented in the capture target region is generally lower (~50–65%), but this is highly variable between regions.

CONCLUSIONS

Inevitably, there is always the temptation to quantitatively compare approaches to target enrichment. The specificity of PCR will almost certainly always exceed that of hybrid capture, and its uniformity may never be matched by either hybrid capture or MIPs. But specificity and uniformity are not everything: the chief advantage of these alternative methods is their ability to capture large target regions in a single experiment, more rapidly and conveniently than PCR. To capture the entire 30 Mb human exome, for example, would require at least 6,000 separate PCRs, each of which would need to be optimized, the products would need to be normalized, and a total of around 120 μg of genomic DNA would be required for the experiment. The same could be performed in a single hybrid capture experiment, taking a single day for the library preparation and about two additional days for the hybridization and elution, and requiring as little as 3 μg of DNA.

Target enrichment can be a highly effective way of reducing sequencing costs and saving sequencing time, and has the power to bring the field of genomics into smaller laboratories, as well as being an invaluable tool for the detection of disease-causing variants. Conversely, target enrichment increases sample preparation cost and time. Assuming that the throughput of next-generation runs and our ability to analyze large numbers of whole-genome datasets both continue to increase, and the cost per base of sequence continues to decrease, there will come a point at which it is no longer economical to perform target enrichment of single samples, compared to whole-genome sequencing. The cost of performing target enrichment by hybridization can be reduced by pooling samples before hybridization, though in our experience results from capturing pooled samples on arrays have been poorer than in solution. This is presumably a reflection of the difference in the probe:sample ratio for the two capture methods.

The logical extension of sample pooling is to perform multiplexed target enrichments in which samples are barcoded before capture. We expect this to have many applications in the future, and the technical details of this are currently being worked out.

© 2010 Nature America, Inc. All rights reserved.

We will provide updates of protocols at ftp://ftp.sanger.ac.uk/pub/pulldown/.

Accession codes. European Short Read Archive: ERA000184.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank D. MacArthur, Q. Ayub and C. Tyler-Smith for their work on long PCR and subsequent analyses, P. Akan, A. Palotie, P. Tarpey, H. Arbury and M. Humphries for their work on hybrid capture and E. Sheridan for critical reading of the standard operating procedures. This work was supported by the Wellcome Trust grant WT079643 and by US National Institutes of Health National Human Genome Research Institute grants 5R21HG004749 and 5R01HL094976.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

© 2010 Nature America, Inc. All rights reserved.

1. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
3. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
5. Saiki, R.K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).
This paper was the first description of PCR, which, coupled to electrophoretic sequencing, is the primary conventional method for targeted variation analysis.
6. Cho, R.J. *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* **23**, 203–207 (1999).
7. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
8. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
9. Meuzelaar, L.S., Lancaster, O., Pasche, J.P., Kopal, G. & Brookes, A.J. MegaPlex PCR: a strategy for multiplex amplification. *Nat. Methods* **4**, 835–837 (2007).
10. Varley, K.E. & Mitra, R.D. Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.* **18**, 1844–1850 (2008).
11. Barnes, W.M. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91**, 2216–2220 (1994).
12. Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
13. Cronn, R. *et al.* Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**, e122 (2008).
14. Harismendy, O. & Frazer, K. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. *Biotechniques* **46**, 229–231 (2009).
15. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
16. Okou, D.T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* **4**, 907–909 (2007).
17. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* **5**, 235–237 (2008).
18. Hamming, R. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–161 (1950).
19. Ikegawa, S., Mabuchi, A., Ogawa, M. & Ikeda, T. Allele-specific PCR amplification due to sequence identity between a PCR primer and an amplicon: is direct sequencing so reliable? *Hum. Genet.* **110**, 606–608 (2002).
20. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. USA* **100**, 8817–8822 (2003).
21. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16**, 652–656 (1998).
22. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol.* **27**, 1025–1031 (2009).
This paper describes the performance of the RainDance technology, which facilitates multiplex PCR by compartmentalizing primer pairs in distinct microdroplet populations that are then mixed and thermocycled in aggregate.
23. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33**, e71 (2005).
24. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
25. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* **104**, 9387–9392 (2007).
26. Faruqi, A.F. *et al.* High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. *BMC Genomics* **2**, 4 (2001).
27. Antson, D.O., Isaksson, A., Landegren, U. & Nilsson, M. PCR-generated padlock probes detect single nucleotide variation in genomic DNA. *Nucleic Acids Res.* **28**, E58 (2000).
28. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21**, 673–678 (2003).
29. Lizardi, P.M. *et al.* Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* **19**, 225–232 (1998).
30. Hardenbol, P. *et al.* Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* **15**, 269–275 (2005).
31. Nilsson, M. *et al.* Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265**, 2085–2088 (1994).
32. Landegren, U. *et al.* Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. *J. Mol. Recognit.* **17**, 194–197 (2004).
33. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).
This paper demonstrates a substantially optimized protocol for using molecular inversion probes for exon capture that also enables library-free integration of multiplex capture and next-generation sequencing.
34. Deng, J. *et al.* Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol.* **27**, 353–360 (2009).
35. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
36. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
37. Lovett, M., Kere, J. & Hinton, L.M. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88**, 9628–9632 (1991).
38. Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D. & Weissman, S.M. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88**, 9623–9627 (1991).
39. Albert, T.J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4**, 903–905 (2007).
This was one of three papers that described solid-phase, hybridization-based enrichment of targeted sequences in shotgun DNA libraries using programmable microarrays.
40. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
This paper was the first description of the method, now commercialized by Agilent, for solution-phase hybridization-based capture using complex libraries of RNA 'bait' to capture from a shotgun DNA 'pond' library.
41. Hodges, E. *et al.* Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protocols* **4**, 960–974 (2009).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
44. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
45. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).

View publication stats

nature | **methods**

Target-enrichment strategies for next-generation sequencing

Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure & Daniel J Turner

Supplementary figures and text:

Supplementary Table 1 Error-correcting barcode tags and PCR primers.

Supplementary Protocol 1 PCR and 96-well library prep standard operating procedure.

Supplementary Protocol 2 MIP standard operating procedure.

Supplementary Protocol 3 Hybrid capture library prep standard operating procedure.

Supplementary Protocol 4 Array capture standard operating procedure.

Supplementary Protocol 5 Solution capture standard operating procedure.

Supplementary Protocol 6 Hybrid capture eluate PCR standard operating procedure.

Supplementary Table 1. Error-correcting barcode tags and PCR primers.
Column B contains octamer barcodes, and Column D contains PCR primers that contain these barcodes. Column C contains the reverse complement of the barcodes given in Column B. This is the sequence that is obtained from the barcodes in the sequencing reaction, and which are robust to errors.

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
1	ACAAGCTA	TAGCTTGT	CAAGCAGAAGACGGCATACGAGATACAAGCTAGAGATCGGTCTCGGCATTC
2	AAACATCG	CGATGTTT	CAAGCAGAAGACGGCATACGAGATAAACATCGGAGATCGGTCTCGGCATTC
3	ACATTGGC	GCCAATGT	CAAGCAGAAGACGGCATACGAGATACATTGGCGAGATCGGTCTCGGCATTC
4	ACCACTGT	ACAGTGGT	CAAGCAGAAGACGGCATACGAGATACCACTGTGAGATCGGTCTCGGCATTC
5	AACGTGAT	ATCACGTT	CAAGCAGAAGACGGCATACGAGATAACGTGATGAGATCGGTCTCGGCATTC
6	CGCTGATC	GATCAGCG	CAAGCAGAAGACGGCATACGAGATCGCTGATCGAGATCGGTCTCGGCATTC
7	CAGATCTG	CAGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCTGGAGATCGGTCTCGGCATTC
8	ATGCCTAA	TTAGGCAT	CAAGCAGAAGACGGCATACGAGATATGCCTAAGAGATCGGTCTCGGCATTC
9	CTGTAGCC	GGCTACAG	CAAGCAGAAGACGGCATACGAGATCTGTAGCCGAGATCGGTCTCGGCATTC
10	AGTACAAG	CTTGTACT	CAAGCAGAAGACGGCATACGAGATAGTACAAGGAGATCGGTCTCGGCATTC
11	CATCAAGT	ACTTGATG	CAAGCAGAAGACGGCATACGAGATCATCAAGTGAGATCGGTCTCGGCATTC
12	AGTGGTCA	TGACCACT	CAAGCAGAAGACGGCATACGAGATAGTGGTCAGAGATCGGTCTCGGCATTC
13	AACAACCA	TGGTTGTT	CAAGCAGAAGACGGCATACGAGATAACAACCAGAGATCGGTCTCGGCATTC
14	AACCGAGA	TCTCGGTT	CAAGCAGAAGACGGCATACGAGATAACCGAGAGAGATCGGTCTCGGCATTC
15	AACGCTTA	TAAGCGTT	CAAGCAGAAGACGGCATACGAGATAACGCTTAGAGATCGGTCTCGGCATTC
16	AAGACGGA	TCCGTCTT	CAAGCAGAAGACGGCATACGAGATAAGACGGAGAGATCGGTCTCGGCATTC
17	AAGGTACA	TGTACCTT	CAAGCAGAAGACGGCATACGAGATAAGGTACAGAGATCGGTCTCGGCATTC
18	ACACAGAA	TTCTGTGT	CAAGCAGAAGACGGCATACGAGATACACAGAAGAGATCGGTCTCGGCATTC
19	ACAGCAGA	TCTGCTGT	CAAGCAGAAGACGGCATACGAGATACAGCAGAGAGATCGGTCTCGGCATTC
20	ACCTCCAA	TTGGAGGT	CAAGCAGAAGACGGCATACGAGATACCTCCAAGAGATCGGTCTCGGCATTC
21	ACGCTCGA	TCGAGCGT	CAAGCAGAAGACGGCATACGAGATACGCTCGAGAGATCGGTCTCGGCATTC
22	ACGTATCA	TGATACGT	CAAGCAGAAGACGGCATACGAGATACGTATCAGAGATCGGTCTCGGCATTC
23	ACTATGCA	TGCATAGT	CAAGCAGAAGACGGCATACGAGATACTATGCAGAGATCGGTCTCGGCATTC
24	AGAGTCAA	TTGACTCT	CAAGCAGAAGACGGCATACGAGATAGAGTCAAGAGATCGGTCTCGGCATTC
25	AGATCGCA	TGCGATCT	CAAGCAGAAGACGGCATACGAGATAGATCGCAGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
26	AGCAGGAA	TTCCTGCT	CAAGCAGAAGACGGCATACGAGATAGCAGGAAGAGATCGGTCTCGGCATTC
27	AGTCACTA	TAGTGACT	CAAGCAGAAGACGGCATACGAGATAGTCACTAGAGATCGGTCTCGGCATTC
28	ATCCTGTA	TACAGGAT	CAAGCAGAAGACGGCATACGAGATATCCTGTAGAGATCGGTCTCGGCATTC
29	ATTGAGGA	TCCTCAAT	CAAGCAGAAGACGGCATACGAGATATTGAGGAGAGATCGGTCTCGGCATTC
30	CAACCACA	TGTGGTTG	CAAGCAGAAGACGGCATACGAGATCAACCACAGAGATCGGTCTCGGCATTC
31	CAAGACTA	TAGTCTTG	CAAGCAGAAGACGGCATACGAGATCAAGACTAGAGATCGGTCTCGGCATTC
32	CAATGGAA	TTCCATTG	CAAGCAGAAGACGGCATACGAGATCAATGGAAGAGATCGGTCTCGGCATTC
33	CACTTCGA	TCGAAGTG	CAAGCAGAAGACGGCATACGAGATCACTTCGAGAGATCGGTCTCGGCATTC
34	CAGCGTTA	TAACGCTG	CAAGCAGAAGACGGCATACGAGATCAGCGTTAGAGATCGGTCTCGGCATTC
35	CATACCAA	TTGGTATG	CAAGCAGAAGACGGCATACGAGATCATACCAAGAGATCGGTCTCGGCATTC
36	CCAGTTCA	TGAACTGG	CAAGCAGAAGACGGCATACGAGATCCAGTTCAGAGATCGGTCTCGGCATTC
37	CCGAAGTA	TACTTCGG	CAAGCAGAAGACGGCATACGAGATCCGAAGTAGAGATCGGTCTCGGCATTC
38	CCGTGAGA	TCTCACGG	CAAGCAGAAGACGGCATACGAGATCCGTGAGAGAGATCGGTCTCGGCATTC
39	CCTCCTGA	TCAGGAGG	CAAGCAGAAGACGGCATACGAGATCCTCCTGAGAGATCGGTCTCGGCATTC
40	CGAACTTA	TAAGTTCG	CAAGCAGAAGACGGCATACGAGATCGAACTTAGAGATCGGTCTCGGCATTC
41	CGACTGGA	TCCAGTCG	CAAGCAGAAGACGGCATACGAGATCGACTGGAGAGATCGGTCTCGGCATTC
42	CGCATACA	TGTATGCG	CAAGCAGAAGACGGCATACGAGATCGCATACAGAGATCGGTCTCGGCATTC
43	CTCAATGA	TCATTGAG	CAAGCAGAAGACGGCATACGAGATCTCAATGAGAGATCGGTCTCGGCATTC
44	CTGAGCCA	TGGCTCAG	CAAGCAGAAGACGGCATACGAGATCTGAGCCAGAGATCGGTCTCGGCATTC
45	CTGGCATA	TATGCCAG	CAAGCAGAAGACGGCATACGAGATCTGGCATAGAGATCGGTCTCGGCATTC
46	GAATCTGA	TCAGATTC	CAAGCAGAAGACGGCATACGAGATGAATCTGAGAGATCGGTCTCGGCATTC
47	GACTAGTA	TACTAGTC	CAAGCAGAAGACGGCATACGAGATGACTAGTAGAGATCGGTCTCGGCATTC
48	GAGCTGAA	TTCAGCTC	CAAGCAGAAGACGGCATACGAGATGAGCTGAAGAGATCGGTCTCGGCATTC
49	GATAGACA	TGTCTATC	CAAGCAGAAGACGGCATACGAGATGATAGACAGAGATCGGTCTCGGCATTC
50	GCCACATA	TATGTGGC	CAAGCAGAAGACGGCATACGAGATGCCACATAGAGATCGGTCTCGGCATTC
51	GCGAGTAA	TTACTCGC	CAAGCAGAAGACGGCATACGAGATGCGAGTAAGAGATCGGTCTCGGCATTC
52	GCTAACGA	TCGTTAGC	CAAGCAGAAGACGGCATACGAGATGCTAACGAGAGATCGGTCTCGGCATTC
53	GCTCGGTA	TACCGAGC	CAAGCAGAAGACGGCATACGAGATGCTCGGTAGAGATCGGTCTCGGCATTC
54	GGAGAACA	TGTTCTCC	CAAGCAGAAGACGGCATACGAGATGGAGAACAGAGATCGGTCTCGGCATTC
55	GGTGCGAA	TTCGCACC	CAAGCAGAAGACGGCATACGAGATGGTGCGAAGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
56	GTACGCAA	TTGCGTAC	CAAGCAGAAGACGGCATACGAGATGTACGCAAGAGATCGGTCTCGGCATTC
57	GTCGTAGA	TCTACGAC	CAAGCAGAAGACGGCATACGAGATGTCGTAGAGAGATCGGTCTCGGCATTC
58	GTCTGTCA	TGACAGAC	CAAGCAGAAGACGGCATACGAGATGTCTGTCAGAGATCGGTCTCGGCATTC
59	GTGTTCTA	TAGAACAC	CAAGCAGAAGACGGCATACGAGATGTGTTCTAGAGATCGGTCTCGGCATTC
60	TAGGATGA	TCATCCTA	CAAGCAGAAGACGGCATACGAGATTAGGATGAGAGATCGGTCTCGGCATTC
61	TATCAGCA	TGCTGATA	CAAGCAGAAGACGGCATACGAGATTATCAGCAGAGATCGGTCTCGGCATTC
62	TCCGTCTA	TAGACGGA	CAAGCAGAAGACGGCATACGAGATTCCGTCTAGAGATCGGTCTCGGCATTC
63	TCTTCACA	TGTGAAGA	CAAGCAGAAGACGGCATACGAGATTCTTCACAGAGATCGGTCTCGGCATTC
64	TGAAGAGA	TCTCTTCA	CAAGCAGAAGACGGCATACGAGATTGAAGAGAGAGATCGGTCTCGGCATTC
65	TGGAACAA	TTGTTCCA	CAAGCAGAAGACGGCATACGAGATTGGAACAAGAGATCGGTCTCGGCATTC
66	TGGCTTCA	TGAAGCCA	CAAGCAGAAGACGGCATACGAGATTGGCTTCAGAGATCGGTCTCGGCATTC
67	TGGTGGTA	TACCACCA	CAAGCAGAAGACGGCATACGAGATTGGTGGTAGAGATCGGTCTCGGCATTC
68	TTCACGCA	TGCGTGAA	CAAGCAGAAGACGGCATACGAGATTTCACGCAGAGATCGGTCTCGGCATTC
69	AACTCACC	GGTGAGTT	CAAGCAGAAGACGGCATACGAGATAACTCACCGAGATCGGTCTCGGCATTC
70	AAGAGATC	GATCTCTT	CAAGCAGAAGACGGCATACGAGATAAGAGATCGAGATCGGTCTCGGCATTC
71	AAGGACAC	GTGTCCTT	CAAGCAGAAGACGGCATACGAGATAAGGACACGAGATCGGTCTCGGCATTC
72	AATCCGTC	GACGGATT	CAAGCAGAAGACGGCATACGAGATAATCCGTCGAGATCGGTCTCGGCATTC
73	AATGTTGC	GCAACATT	CAAGCAGAAGACGGCATACGAGATAATGTTGCGAGATCGGTCTCGGCATTC
74	ACACGACC	GGTCGTGT	CAAGCAGAAGACGGCATACGAGATACACGACCGAGATCGGTCTCGGCATTC
75	ACAGATTC	GAATCTGT	CAAGCAGAAGACGGCATACGAGATACAGATTCGAGATCGGTCTCGGCATTC
76	AGATGTAC	GTACATCT	CAAGCAGAAGACGGCATACGAGATAGATGTACGAGATCGGTCTCGGCATTC
77	AGCACCTC	GAGGTGCT	CAAGCAGAAGACGGCATACGAGATAGCACCTCGAGATCGGTCTCGGCATTC
78	AGCCATGC	GCATGGCT	CAAGCAGAAGACGGCATACGAGATAGCCATGCGAGATCGGTCTCGGCATTC
79	AGGCTAAC	GTTAGCCT	CAAGCAGAAGACGGCATACGAGATAGGCTAACGAGATCGGTCTCGGCATTC
80	ATAGCGAC	GTCGCTAT	CAAGCAGAAGACGGCATACGAGATATAGCGACGAGATCGGTCTCGGCATTC
81	ATCATTCC	GGAATGAT	CAAGCAGAAGACGGCATACGAGATATCATTCCGAGATCGGTCTCGGCATTC
82	ATTGGCTC	GAGCCAAT	CAAGCAGAAGACGGCATACGAGATATTGGCTCGAGATCGGTCTCGGCATTC
83	CAAGGAGC	GCTCCTTG	CAAGCAGAAGACGGCATACGAGATCAAGGAGCGAGATCGGTCTCGGCATTC
84	CACCTTAC	GTAAGGTG	CAAGCAGAAGACGGCATACGAGATCACCTTACGAGATCGGTCTCGGCATTC
85	CCATCCTC	GAGGATGG	CAAGCAGAAGACGGCATACGAGATCCATCCTCGAGATCGGTCTCGGCATTC

Tag number	Barcode sequence	Single correcting, double & shift detecting octamers	PCR primers
86	CCGACAAC	GTTGTCGG	CAAGCAGAAGACGGCATACGAGATCCGACAACGAGATCGGTCTCGGCATTC
87	CCTAATCC	GGATTAGG	CAAGCAGAAGACGGCATACGAGATCCTAATCCGAGATCGGTCTCGGCATTC
88	CCTCTATC	GATAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTATCGAGATCGGTCTCGGCATTC
89	CGACACAC	GTGTGTCG	CAAGCAGAAGACGGCATACGAGATCGACACACGAGATCGGTCTCGGCATTC
90	CGGATTGC	GCAATCCG	CAAGCAGAAGACGGCATACGAGATCGGATTGCGAGATCGGTCTCGGCATTC
91	CTAAGGTC	GACCTTAG	CAAGCAGAAGACGGCATACGAGATCTAAGGTCGAGATCGGTCTCGGCATTC
92	GAACAGGC	GCCTGTTC	CAAGCAGAAGACGGCATACGAGATGAACAGGCGAGATCGGTCTCGGCATTC
93	GACAGTGC	GCACTGTC	CAAGCAGAAGACGGCATACGAGATGACAGTGCGAGATCGGTCTCGGCATTC
94	GAGTTAGC	GCTAACTC	CAAGCAGAAGACGGCATACGAGATGAGTTAGCGAGATCGGTCTCGGCATTC
95	GATGAATC	GATTCATC	CAAGCAGAAGACGGCATACGAGATGATGAATCGAGATCGGTCTCGGCATTC
96	GCCAAGAC	GTCTTGGC	CAAGCAGAAGACGGCATACGAGATGCCAAGACGAGATCGGTCTCGGCATTC

Supplementary Protocol 1

PCR and 96-well library prep standard operating procedure

Adapter preparation

Custom adapters are required. These should be HPLC purified. Adapters are phosphorylated and annealed together, as described [1].

Ind_t 5' ACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3'
Ind_b 5' GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC 3'

* indicates phosphorothioate

Sample shearing

1. Take 1 µg of pooled long PCR products and make volume up to 75 µl with water.

2. Shear to approximately 200 bp by acoustic shearing: transfer to a 6 mm × 16 mm AFA fiber vial (Covaris cat. no. 520031).

3. Seal the tube with an 8 mm crimp seal cap (Covaris cat no. 520028) and crimping tool.

4. Shear with a Covaris, using the settings:

Duty cycle	20 %
Intensity	5
Cycle/burst	200
Time	150 sec

5. Transfer samples to 96-well PCR plate, gently spin and keep frozen until used.

End-repair

This protocol converts the overhangs resulting from fragmentation into blunt ends, using T4 DNA polymerase and E. coli DNA polymerase I Klenow fragment. The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs.

1. Mix the reagents below (all NEB) in a 15ml Falcon Tube, and decant into a reagent reservoir.

Nature Methods: doi: 10.1038/nmeth.1419

	1X	104X
10x T4 DNA ligase buffer with 10 mM ATP	10 µl	1040 µl
10 µM dNTP mix	4 µl	416 µl
3 U / µl T4 DNA polymerase	5 µl	520 µl
5 U / µl Klenow DNA polymerase	1 µl	104 µl
10 U / µl T4 PNK	5 µl	520 µl

2. Add 25 µl of master mix to each sample using an electronic pipette. Cover the plate with transparent cover, vortex briefly and gently spin down.

3. Incubate plate for 30 min at 20 °C in a thermocycler.

4. While incubating, prepare SPRI beads for the reaction cleanup: allow SPRI beads to come to room temperature for at least 30 minutes. Mix well, and ensure that the beads appear homogeneous and consistent in colour.

5. Add 180 µl of SPRI beads per 100 µl of end-repaired DNA in a 1.5 ml Lo-Bind tube.

6. Vortex and leave at room temperature for 5 minutes.

7. Place tubes in a magnetic rack.

8. Leave for 5 minutes or until sample is clear.

9. Carefully remove the clear solution from the tubes and discard.

10. Dispense 700 µl of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.

11. Repeat the ethanol wash once again (total of two washes).

12. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 to 10 minutes or until the residual ethanol has evaporated.

13. Add 32 µl of molecular biology grade water, vortex and incubate at room temperature for 2 minutes.

14. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.

15. Carefully remove the water and retain in a new 1.5 ml Lo-Bind tube.

16. Centrifuge the eluates at 13,000 rpm in a bench top centrifuge for 10 minutes.

17. Transfer eluates to a 96-well plate leaving behind any precipitated beads.

Nature Methods: doi: 10.1038/nmeth.1419

18. Proceed immediately with A-tailing.

A-tailing

This protocol adds an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). This prepares the DNA fragments for ligation to the adapters, which have a single 'T' base overhang at their 3' end.

1. Mix the reagents below (all NEB) in a 15ml Falcon Tube, and decant into a reagent reservoir.

	1X	104X
10x NEB buffer 2	5 µl	520 µl
1 mM dATP	10 µl	1040 µl
5 U / µl Klenow fragment (3' to 5' exo$^-$)	3 µl	312 µl

2. Add 18µl to each sample using an electronic pipette. Cover the plate with transparent cover, vortex and gently spin.

3. Incubate plate for 30 min at 37 °C in a thermocycler. Clean using SPRI beads, in the 96-well reaction plate (see End-Repair protocol above. Use 90 µl beads for each 50µl reaction, and elute in a mixture of 10 µl EB + 8.5 µl water).

4. Proceed immediately with ligation.

Ligation

This protocol ligates adapters to the ends of the DNA fragments.

1. Mix the reagents below in a 15ml Falcon Tube, and decant into a reagent reservoir.

	1X	104X
2x Quick DNA ligase buffer (NEB)	25 µl	2.6 ml
Adapter oligo mix (see above)	1.5 µl	156 µl

2. Add 26.5 µl to each sample using an electronic pipette. Cover the plate with transparent cover, vortex and gently spin.

3. Pipette 5 µl Quick Ligase (NEB) into each well using a manual pipette. Mix by pipetting, do not cover, and after pipetting in the last column leave for 15 min at room temperature (20 °C).

Nature Methods: doi: 10.1038/nmeth.1419

4. Clean using AMPure SPRI beads as described above, eluting in 20 µl EB.

5. Run an DNA 1000 chip on an Agilent BioAnalyzer 2100 on a small selection of samples. You should detect smear between 300-1000 bp.

Indexing Enrichment PCR

Indexes can be added to the central region of the reverse PCR primer. The general sequence is:

5' CAAGCAGAAGACGGCATACGAGAT-INDEX-GAGATCGGTCTCGGCATTC 3'

where INDEX represents an oligonucleotide sequence that is used to identify the sample. Indexes should be selected so that they are maximally different from one another. We typically use 8-base indexes. Indexed primers must be PAGE purified. For an example of a 96-plex set of error-correcting barcodes, and primer sequences, see **Supplementary Table 1**.

The common forward primer (HPLC purified) has the sequence:

5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'

1. Dilute indexed primers to 10 µM with water.

2. Pipette 2 µl of each library into a well of a 96-well plate.

3. Add 6 µl of 10 µM indexed primer to each well.

4. Prepare a master mix of all other reagents:

	1X	52X
10X PFx buffer (Invitrogen)	10µl	520µl
2.5 mM dNTPs	20µl	1040µl
50 mM $MgSO_4$	4µl	208µl
10 µM Forward primer	6µl	312µl
Platinum PFx Polymerase (Invitrogen)	1µl	52µl
Water	51µl	2652µl

5. Mix, and add 92 µl master mix to each well. The PCR is performed using the following program :

   94 °C for 2 min

   94 °C for 15 sec

Nature Methods: doi: 10.1038/nmeth.1419

68 °C for 45 sec x 12 cycles

4°C indefinitely

6. Clean PCR reactions SPRI beads, as described above, eluting in 18 µl EB.

7. Run an DNA 1000 chip on an Agilent BioAnalyzer 2100 on a small selection of samples. You should detect smear between 300-1000 bp.

Normalization

Normalize PCR products by qPCR before pooling. The primers used for the qPCR are locus-specific, designed to amplify one of the long PCR products in the original pool. In this way, adapter dimers are not problematic. As a concentration standard, the most concentrated library from the 6 random libraries quantified by Agilent Bioanalyzer 2100 in the preceding step, is used.

1. Dilute 2µl of the chosen concentration standard 10x, 100x and 1,000x with EB buffer.

2. Dilute all other libraries 80x in EB.

3. Perform qPCRs in duplicate, using a 2x SybrGreen master mix.

4. Pool products in equimolar ratios.

Size selection

1. Load sample pools on a 2% agarose gel in 1X TBE, and are electrophoresed against a Low Molecular Weight ladder (NEB), at 5V cm^{-1} for approximately 1 hour.

2. From each lane, cut a 250-450 bp gel slice is cut and extract the DNA.

3. Quantify pooled libraries by SYBRGreen qPCR [2].

4. Sequence using standard primers for reads 1 and 2. Indexes are sequenced after read 1, using a custom primer.

Ind_seq 5’ AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC 3’

Nature Methods: doi: 10.1038/nmeth.1419

References

1. Kozarewa I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* **6**, 291-295 (2009).

2. Quail M.A., Swerdlow H. & Turner D.J. Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18**, Unit 18 12 (2009).

Nature Methods: doi: 10.1038/nmeth.1419

Supplementary Protocol 2

Molecular Inversion Probe standard operating procedure

Generate Probes

Amplify off-array oligonucleotides (MIP precursors) using PCR: (2.5 hrs)

1. Dissolve array-derived MIP precursor oligonucleotides (mixture of 100-mers obtained from Agilent) to a final concentration of 100 nM in Tris-EDTA buffer with a pH of 8 and 0.1 % Tween.

2. Prepare the following 400 µl PCR mix in a 1.5 ml centrifuge tube. Mix and spin down:

Reagent	Volume (µl)	Final Concentration
2x iProof HF PCR master mix (Biorad)	200	1x
Oligo_Fwd_Amp Primer (100 µM)	2	500 nM
Oligo_Rev_Amp Primer (100 µM)	2	500 nM
SYBRGreen I 100 x (Invitrogen)**	1	0.2X
Template (100 nM in 0.1 % Tween)	1	250 pM
Water	194	

Split into 8 x 50 µl reactions in 0.2 ml PCR tubes. One PCR preparation can be expected to yield around 1.5 µg of amplified DNA. **Use SYBR Green when using a real-time thermocycling instrument.

3. Use the following PCR cycling program, ideally on a real-time thermocycling instrument such as the Biorad MJ Mini.

98 °C for 30 seconds

98 °C for 10 seconds
60 °C for 30 seconds x 25 cycles
72 °C for 30 seconds (read plate)

4 °C indefinitely

**We typically stop our reactions after 20 rounds of PCR. When using RT-PCR, stop the reaction slightly before the fluorescence curve plateaus to avoid over-amplification.

Nature Methods: doi: 10.1038/nmeth.1419

4. Combine and clean up PCR reactions on one column using the QIAquick PCR purification kit following the manufacturer's instructions. Elute with 90 µl elution buffer.

5. Use a Qubit High Sensitivity dsDNA Assay Kit to quantify 1 µl of the amplified DNA.

6. Analyze 1 µl amplified DNA on a 6 % TBE PAGE gel (Invitrogen) to verify amplification. Product should appear as a single band at 110 bp, as the primers add an additional 10 bp.

Digest PCR product with nicking restriction endonucleases to generate 70-mer MIPs (7.5 hrs):

7. Add 10 µl of NEB-2 (10x) and 5 µl of Nt.AlwI (10 U / µl; NEB) to 85 µl of PCR product (total volume of 100 µl)

8. Mix and split to two tubes of 50 µl each. Incubate at 37 °C for 3 hours, followed by 80 °C for 20 minutes in a thermocycler

9. Let the temperature drop to 65 °C for at least 1 minute. Add 2.5 µl of Nb.BsrDI (2 U / µl; NEB) to each of the 50 µl reactions

10. Leave at 65 °C for 3 hours, followed by 80°C for 20 minutes

11. Purify two 50 µl digestion reactions on one column using reagents from the QIAquick Nucleotide Removal Kit. Elute each column in 30 µl elution buffer. We have observed yields of 80-90 % for this step.

Quantify usable probe using a denaturing gel (2 hrs):

12. Accurate quantification of usable MIP inside the digested probe mix is important as it determines how much probe mix to add to the capture reaction.

13. Prepare two-fold dilutions of a NEB 100 bp DNA ladder (we used dilutions from 500 ng to 62 ng).

14. Mix 2x TBE-Urea sample buffer (Invitrogen) with 1 µl digested probe and the dilutions made above.

15. Denature DNA by heating to 95 °C for 5 minutes and immediately transferring to ice.

16. Run samples on a precast 6 % TBE-urea denaturing PAGE gel (Invitrogen) for 1 hr at 160 V.

Nature Methods: doi: 10.1038/nmeth.1419

17. Quantify the amount of usable MIP in the digested mixture by comparing the intensity of ladder dilutions with the intensity of the 70 bp band. Use this MIP concentration when determining the volume of probe mix to add to a capture reaction.

Capture Reaction

Note: We have found that it is no longer necessary to gel-purify the single stranded 70 bp MIP from the digested probe mix in a capture reaction. Instead, we use a blocking oligo to limit hybridization of the undigested strand of the MIP precursor (still a 100-mer) to the active 70-mer MIP in the capture reaction, which could potentially interfere with MIP hybridization to genomic DNA targets.

Hybridize probes to genomic DNA (37 hrs):

1. For each sample to capture, add the following reagents in a 0.2 ml PCR tube. The final capture reaction volume is 25 µl. Because there is no size selection of the 70 bp MIP, the volume of probe mix to add is based on the concentration of usable MIP.

Reagent	Volume (µl) per sample	Final Concentration in reaction
750 ng genomic DNA*	3	30 ng / µl*
10 x Ampligase buffer (Epicentre)	2.5	1x
40 ng (2 pmol) of MIP**	3	1.6 ng / µl**
Blocking Oligonucleotide (100 µM)	0.1	0.4 µM
Water	16.4	

* Additionally, prepare a blank capture reaction containing MIP probe but no gDNA to detect cross contamination.
** For a reaction targeting 55,000 regions in the genome. We currently aim for a ratio of MIPs to genomic DNA of 100:1 (i.e. 100 copies of each MIP in the mix for each genomic equivalent). The concentration of MIPs can be adjusted accordingly depending on complexity of the targeting reaction.

2. Denature at 95 °C for 10 minutes.

3. Incubate at 60 °C for at least 36 hours to hybridize MIPs to gDNA.

Circularize captured exons: (1 day)

4. Prepare a mix of ligase and polymerase enzymes to add to each capture reaction:

Reagent	Volume (µl)	Final

Nature Methods: doi: 10.1038/nmeth.1419

	per sample	Concentration in capture reaction
10 x Ampligase buffer (Epicentre)	0.45	1x
10 U / µl Stoffel** (Applied Biosystems)	2	0.8 U / µl
100 U / µl Ampligase** (Epicentre)	1	4 U / µl
0.25 mM dNTP**	1.25	12 µM

Prepare this mix on ice, and keep cold before adding 4.7 µl into the capture reaction.

5. Incubate at 60 °C for an additional 24 hours to allow for gap-fill and ligation to circularize captured regions.

Exonuclease select for circularized product: (1hr)

6. Prepare a mix of exonucleases to add to each capture reaction in order to remove uncaptured gDNA, excess probe and blocking oligonucleotide:

Reagent	Volume (µl) per sample	Final Concentration in reaction
Exo I 20 U / µl	2	1.7 U / µl
Exo III 100 U / µl	2	8.3 U / µl

7. Reduce the temperature of the capture reaction to 37 °C and allow it to incubate for at least one minute before adding 4 µl of exonuclease mix.

8. Incubate for 15 minutes at 37 °C.

9. Inactivate exonuclease enzymes by heating reaction at 95 °C for 2 minutes.

10. Use 5 µl of the reaction product as the template for PCR. There is no need to purify the reaction product before PCR.

Amplify and Verify Captured Product:

1. Prepare the following PCR mix in a 1.5 ml centrifuge tube. Mix and spin down.

Reagent	Volume (µl) per sample	Final Concentration in capture reaction
iProof PCR master mix (2x)	25	1x
SLXA_Paired_End_CP2_Fwd (100 µM)	0.25	500 nM
SLXA_Paired_End_CP2_Rev (100 µM)	0.25	500 nM

Nature Methods: doi: 10.1038/nmeth.1419

SYBR Green I 100X (Invitrogen)	0.25	0.5X
Water	19.25	

2. Add 45 µl of master mix to 5 µl of each sample to obtain a total reaction volume of 50 µl. Mix gently and spin down.

3. Amplify on a RT-PCR machine using the following conditions:

98 °C for 30 seconds

98 °C for 10 seconds
60 °C for 30 seconds x 25 cycles
72 °C for 60 seconds (read plate)

**When using RT-PCR, stop the reaction slightly before the fluorescence curve plateaus to avoid overamplification.

4. Purify each sample reaction on one column QIAquick PCR Purification column following the manufacturer's instructions. Elute each column in 30 µl EB buffer.

5. Use a Qubit HS dsDNA Assay Kit to quantify 1 µl of the amplified DNA.

6. Analyze 2 µl of the amplified DNA on a PAGE gel as described in step 6 to validate that the amplified product is of the expected size range. Currently, we aim for a uniform gap-fill size of 112 bp for all targets. Assuming 20 bp targeting arms, paired-end 76 bp reads enable full coverage of this gap-fill size. At this step, we consequently expect a tight band centered at 245 bp (49 bp primer (SLXA_Paired_End_CP2_Fwd) + 20 bp targeting arm + 112 bp gap-fill + 20 bp targeting arm + 44 bp primer (SLXA_Paired_End_CP2_Rev) = 245 bp).

7. Samples are now ready for analysis using the Illumina Genome Analyzer. Use PE_Capture_Sequencing and PE_Rev_Capture_Sequencing primers for sequencing

Timeline: (4 days)

Generate Probes (Steps 1-17): 12 hrs
Amplify off-array oligonucleotides using PCR: (2.5 hrs)
Digest oligonucleotides: (7.5 hrs)
Quantify usable probe using a denaturing gel: (2 hrs)

Capture Exons (Steps 18-28): 3 days

Nature Methods: doi: 10.1038/nmeth.1419

Hybridize probes to genomic DNA: (1.5 days)
Circularize captured exons: (1 day)
Exonuclease select for circularized product: (1 hr)

Amplify and Verify Captured Product (Steps 29-35): 4 hrs

Oligonucleotide Sequences:

Oligonucleotide Name	Sequence (5'→3')
General Format of MIP precursors (100-mers); x's and y's indicate variable targeting arm sequence	AGGACCGGATCAACTxxxxxxxxxxxxxxxxxxxCTTCAGCTTCCCGATA TCCGACGGTAGTGTyyyyyyyyyyyyyyyyyyyCATTGCGTGAACCGA
Oligo_Fwd_Amp	TGCCTAGGACCGGATCAACT
Oligo_Rev_Amp	GAGCTTCGGTTCACGCAATG
SLXA_Paired_End_CP2_Fwd	AATGATACGGCGACCACCGAGATCTACACGCACGATCCGACGGTA GTGT
SLXA_Paired_End_CP2_Rev	CAAGCAGAAGACGGCATACGAGATCCGTAATCGGGAAGCTGAAG
PE_Capture_Sequencing Primer	ACACGCACGATCCGACGGTAGTGT
PE_Rev_Capture_Sequencing Primer	CATACGAGATCCGTAATCGGGAAGCTGAAG
Blocking Oligonucleotide	CTTCAGCTTCCCGATATCCGACGGTAGTGT

Nature Methods: doi: 10.1038/nmeth.1419

Supplementary Protocol 3

Hybrid capture library prep standard operating procedure (array OR solution capture)

Shearing samples to 100-300bp

For array capture, dilute 20 µg genomic DNA to a total volume of 100 µl with water.
For solution capture, dilute 3 µg genomic DNA to a total volume of 100 µl with water.

1. Mix and transfer to a 6mmx16mm AFA fibre vial (Covaris cat. no. 520031).

2. Seal the tube using an 8mm metal crimp seal cap (Covaris, cat no. 520028) and crimping tool.

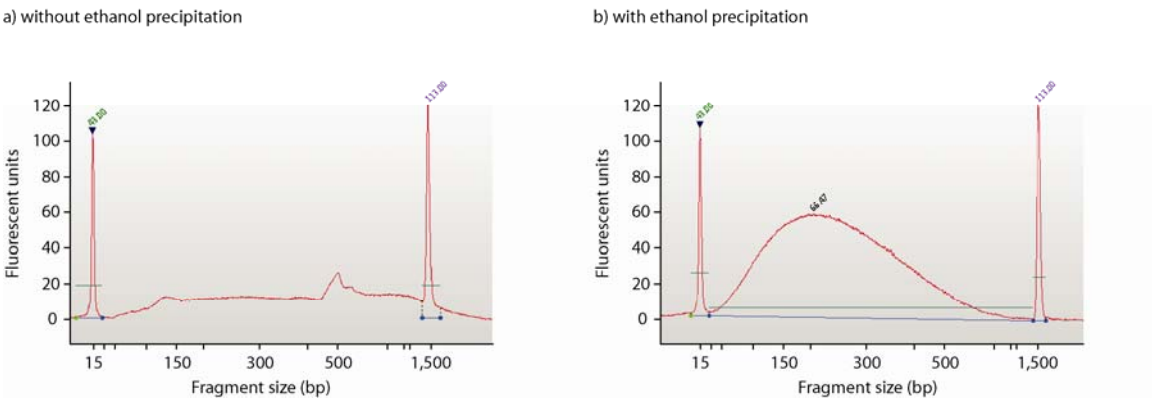
3. Shear with a Covaris, using the settings:

Duty cycle	20 %
Intensity	5
Cycle/burst	200
Time	120 sec

4. Remove the sample from the machine. Open the vial and transfer the sample into a fresh 1.5 ml Eppendorf lo-bind tube. Keep samples on ice.

5. Run 1 µl on an Agilent Bioanalyzer 2100 chip to check the quantity of fragmented DNA and to confirm the success of the fragmentation.

Impure DNA may shear badly. If there is any doubt about the purity of the sample, perform an ethanol precipitation before shearing (Figure 1).



Nature Methods: doi: 10.1038/nmeth.1419

Figure 1: Impure DNA sample sheared using identical Covaris settings a) before and b) after ethanol precipitation.

Purification after Fragmentation

Use one column for each 10 µg of DNA. For 10-20 µg DNA, divide equally between two columns. Elute with 45 µl EB per 5 µg bound DNA.

1. Add 5 x volume of Buffer PB to the fragmented DNA (for 100 µl of fragmented DNA add 500 µl buffer). Vortex.

2. Pipette 750 µl of mix per column. Pipette slowly and make sure that all the liquid comes out of the pipette tip. Centrifuge the samples for 1 minute at 13,000 rpm in a benchtop centrifuge.

3. Discard the flow-through from the collection tube. Repeat step 2 if you have any buffer mix left.

4. To wash, add 750 µl of Buffer PE to each of the columns. Centrifuge for 1 minute as above. Discard the flow-through and centrifuge again for a further minute.

5. Leave the tubes in a rack with the lids open to dry for 2 minutes.

6. Transfer the column to a clean labelled 2 ml lo-bind Eppendorf tube. For each 5 µg of bound DNA, add 47 µl of EB buffer to the centre of the column and leave for 1 minute. Centrifuge for 1 minute as above.

7. Combine eluates, if applicable. You should have a volume of 45 µl for every 5 µg in the original sample (approximately 2 µl of buffer is retained by the column).

End Repair

This protocol converts the overhangs resulting from fragmentation into blunt ends, using T4 DNA polymerase and E. coli DNA polymerase I Klenow fragment. The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs.

1. Prepare a master mix containing the following reaction mix per 5 µg sample, plus a 10 % excess (round the mass of sample UP to the nearest 5 µg):

Water	30 µl

Nature Methods: doi: 10.1038/nmeth.1419

10x T4 DNA ligase buffer with 10mM ATP	10 µl
10mM dNTP mix	4 µl
3U/ µl T4 DNA polymerase	5 µl
5U/ µl Klenow DNA polymerase	1 µl
10U/ µl T4 PNK	5 µl

2. Mix, and aliquot 55 µl of master mix into each sample tube containing the 45 µl of eluate from the previous step. Mix well and spin down.

3. Incubate for 30 minutes at room temperature (20-25 °C).

4. Clean up using one QIAquick PCR column for up to 10 µg of DNA, as described above. Elute in 34 µl EB per 5 µg DNA. This gives ~32 µl of eluate, because approximately 2 µl of buffer is retained by the column.

Addition of 'A' Bases to the 3' End of the DNA Fragments

This protocol adds an 'A' base to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). This prepares the DNA fragments for ligation to the adapters, which have a single 'T' base overhang at their 3' end.

1. Prepare a master mix containing the following reaction mix per 5 µg sample, plus a 10 % excess (round the mass of sample UP to the nearest 5 µg):

10x Klenow buffer	5 µl
1 mM dATP	10 µl

2. Mix, and aliquot 15 µl of master mix into each sample tube containing the 32 µl of end-repaired sample. Mix well and spin down.

3. Add 3 µl 5 U/ µl Klenow exo (3' to 5' exo minus). Mix and spin down.

4. Incubate for 30 minutes at 37 °C in a hot block.

5. Clean up using one QIAquick MinElute columns per 5 µg of DNA, eluting in 12 µl of EB buffer per column, in a 1.5 ml lo-bind Eppendorf tube. This gives ~10 µl of eluate, because approximately 2 µl of buffer is retained by the column.

Nature Methods: doi: 10.1038/nmeth.1419

Ligation of Adapters to DNA Fragments

This protocol ligates adapters to the ends of the DNA fragments. The procedure uses a 10:1 molar ratio of adapter to DNA insert, based on a starting quantity of 5 µg of DNA before fragmentation. The quantities below are given per 5 µg of DNA. Adjust as appropriate.

1. Prepare a master mix containing the following reaction mix per 5 µg sample, plus a 10 % excess (round the mass of sample UP to the nearest 5 µg):

2 x DNA ligase buffer	25 µl
Illumina PE Adapter oligo mix	10 µl

2. Mix and aliquot 30 µl master mix into each sample tube containing the 10 µl of A-tailed sample. Mix and spin down.

3. Add 10 µl 2,000 U/ µl T4 DNA ligase. Mix and spin down.

4. Incubate for 15 minutes at room temperature (20-25 °C).

5. Clean ligated samples and C_0t1 DNA with SPRI beads, eluting in 50 µl water. Use a 5 x excess of C0t1 for array capture and a 20 x excess for solution capture.

Alternatively, if doing a pre-hyb PCR (see below), clean up using a QIAquick PCR column as described above, eluting in 50 µl EB.

SPRI bead cleanup

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Add 90 µl of SPRI beads per 50 µl of adapter ligated sample in a 1.5 ml Lo-bind Eppendorf tube.

2. Vortex and leave at room temperature for 5 minutes.

3. Place tubes in a magnetic rack.

4. Leave for 5 minutes or until sample is clear.

5. Carefully remove the clear solution from the tubes and discard.

Nature Methods: doi: 10.1038/nmeth.1419

6. Dispense 700 µl of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.

7. Repeat the ethanol wash once again (total of two washes).

8. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 to 10 minutes or until the residual ethanol has evaporated.

9. Add 50 µl of molecular biology grade water, vortex and incubate at room temperature for 2 minutes.

10. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.

11. Carefully remove the water and retain in a new 1.5 ml lo-bind Eppendorf tube.

12. Repeat step 9-12 once more, retaining the water in the same 1.5 ml lo-bind tube. Total volume of elute should be 100 µl.

13. Centrifuge the eluate at 13,000 rpm in a bench top centrifuge for 10 minutes

14. Transfer the sample to a new 1.5 ml lo-bind Eppendorf tube leaving behind any precipitated beads.

15. Quantify 1 µl of the library using an Agilent DNA 1000 chip on a Bioanalyzer 2100 and proceed to hyb, following the manufacturer's recommended protocols.

Optional step: pre-hyb PCR

Performing a small number of PCR cycles before hybridisation can improve robustness, particularly for clinical samples, and will simplify sample indexing. Amplify each 50 µl adapter-ligated library by dividing between 4 PCR reactions.

1. Prepare a master mix containing the following reaction mix per sample, plus a 10 % excess:

10 µM PE2.1	10 µl
10 µM PEV2.2	10 µl
2 x Phusion HF master mix	100 µl
water	30 µl

Nature Methods: doi: 10.1038/nmeth.1419

2. Mix and aliquot 150 µl of master mix into each 50 µl adapter-ligated library. Mix and spin down.

3. Aliquot into four 200 µl PCR tubes (50 µl each), and perform the following temperature cycling:

98 °C	2 minutes	
98 °C	20 seconds	
65 °C	30 seconds	x 6 cycles
72 °C	30 seconds	
72 °C	5 minutes	
4 °C	indefinitely	

4. Combine all 4 reactions and clean up with SPRI beads (see above), adding 360 µl beads to the 200 µl PCR reactions, and eluting in 50 µl water.

Quantify 1 µl of library using an Agilent DNA 1000 chip on a Bioanalyzer 2100 and proceed to hybridization (Supplementary Protocol 4 or 5).

Nature Methods: doi: 10.1038/nmeth.1419

Supplementary Protocol 4

Array capture standard operating procedure

This protocol is for hybridization of adapter-ligated or PCR-amplified library DNA, so must be performed after Hybrid Capture Protocol 1.

Sample preparation

Following SPRI bead cleanup or pre-hyb PCR (see Hybrid Capture Protocol 1), lyophilize libraries using a SpeedVac.

1. Add 5.4 µl of molecular biology grade water to each sample to rehydrate.
2. Vortex and centrifuge.
3. Place sample in 70 °C heat block for 10 minutes.
4. Vortex and spin down.
5. To each sample, add:

2 x Hybridization Buffer (Nimblegen)	9.0 µl
Hybridization Component A	3.6 µl

6. Vortex and spin down.
7. Place each sample in a 95 °C heat block for 10 minutes.
8. Spin down.
9. Store at 42 °C (in hyb station tube slots) until ready for hybridization.

Prepare slides and mixers

1. Remove the mixer from the packaging (must be used within 30 minutes of opening). Blow compressed gas across the mixer and slide.

Nature Methods: doi: 10.1038/nmeth.1419

2. Position the precision mixer alignment tool (PMAT) so that the hinge is on the left and then open it.

3. Snap the mixer onto the two alignment pins on the lid of the PMAT, with the tab end of the mixer towards the inside hinge and the mixer's adhesive gasket exposed.

4. While pushing back the plastic spring with a thumb, place the slide in the base of the PMAT so that the barcode is on the right and the corner of the slide sits against the plastic spring. Remove your thumb and make sure the spring is engaging the corner of the slide and the entire slide is registered to the edge of the PMAT closest to you. Gently blow compressed gas across the mixer and slide.

5. Using forceps remove the backing from the adhesive gasket and close the lid of the PMAT so that the gasket makes contact with the slide. Lift the lid while applying pressure through the hole in the lid of the PMAT to free the mixer-slide assembly from the pins.

6. Remove the mixer-slide assembly from the PMAT. Place the mixer-slide assembly on a smooth, dark flat surface.

NOTE: *You can place the mixer-slide assembly on the back of a 42 °C heating block to facilitate adhesion of the mixer to the slid.*

7. Rub the Mixer Brayer over the mixer just hard enough to adhere the adhesive gasket and remove any bubbles. Start in the centre of the array and rub outwards. The adhesive gasket will become clear when fully adhered to both surfaces.

NOTE: *Mark the sample area on the other side of the working surface on the glass slide!*

8. Place the mixer-slide assembly in the slide bay of the Hybridization System. Make sure the assembly is seated completely within the bay.

Load and hybridize samples

NOTE: *Before loading samples onto the hyb station ensure that the array area is marked on the back of the slide.*

1. Using a Gilson Microman M100 pipette, slowly dispense the sample into the fill port. Dispense slowly to avoid introducing bubbles.

2. Tap the slide to remove any bubbles, dry any sample leaking from the ports with a clean tissue.

Nature Methods: doi: 10.1038/nmeth.1419

3. Adhere a seal tab over each fill and vent port on the mixer and rub the seal gently with the blunt end of the forceps to ensure a tight seal.

4. Close the bay clamp.

5. Turn the MAUI hyb station on/off with the mixing on/off switch. The 2 green mix mode lights on the front panel will flash for 5 seconds. Press the B mix mode switch to select the B mix mode while the light is flashing. The hyb system will recognise the slide in each occupied bay (the red indicator light will change to green if occupied).

6. Hybridize the sample to the array for 72 hours at 42 °C.

Washing arrays after hybridisation

Preparing the Wash Solutions

1. Dilute the concentrated wash solutions from the wash buffer kit. **NB:** Shown below are the quantities needed for 1 array. Volumes should be adjusted for the number of arrays to be processed. Wash buffers I, II & III come in 10X concentrated form and the stringent buffer in 2X concentrated form.

Wash Solution	Volume required (of diluted)
Wash buffer I	32 ml
Wash buffer II	164 ml
Wash buffer III	32 ml
Stringent buffer	64 ml

Preparation of the Elution and Clean-up Solutions

Method note: The 125 mM sodium hydroxide (NaOH) and 20 % acetic acid should both be prepared freshly prior to using them in the elution and clean up. Molecular biology grade water (H_2O) should be used for preparing the solutions.

2. Prepare 125 mM sodium hydroxide as detailed below. **NB:** The amounts below are for processing 1 array. The volumes must be adjusted to reflect the number of arrays being processed.

Solution	Volume H_2O	Volume 10 M NaOH

Nature Methods: doi: 10.1038/nmeth.1419

125 mM NaOH	987.5 µl	12.5 µl

3. Prepare 20 % acetic acid as detailed below. **NB:** The amounts below are for processing 1 array. The volumes must be adjusted to reflect the number of arrays being processed.

Solution	Volume H_2O	Volume glacial acetic acid
20 % acetic acid	400.0 µl	100.0 µl

4. Open a QIAquickPCR purification kit and take out the buffers to prepare them for use, see below:

Buffer	Prepare by
PE	Adding 24 ml of 100 % molecular biology grade ethanol
PB	Adding 120 µl pH indicator I to the 30 ml bottle of PB buffer
EB	Requires no further preparation

Washing the Arrays

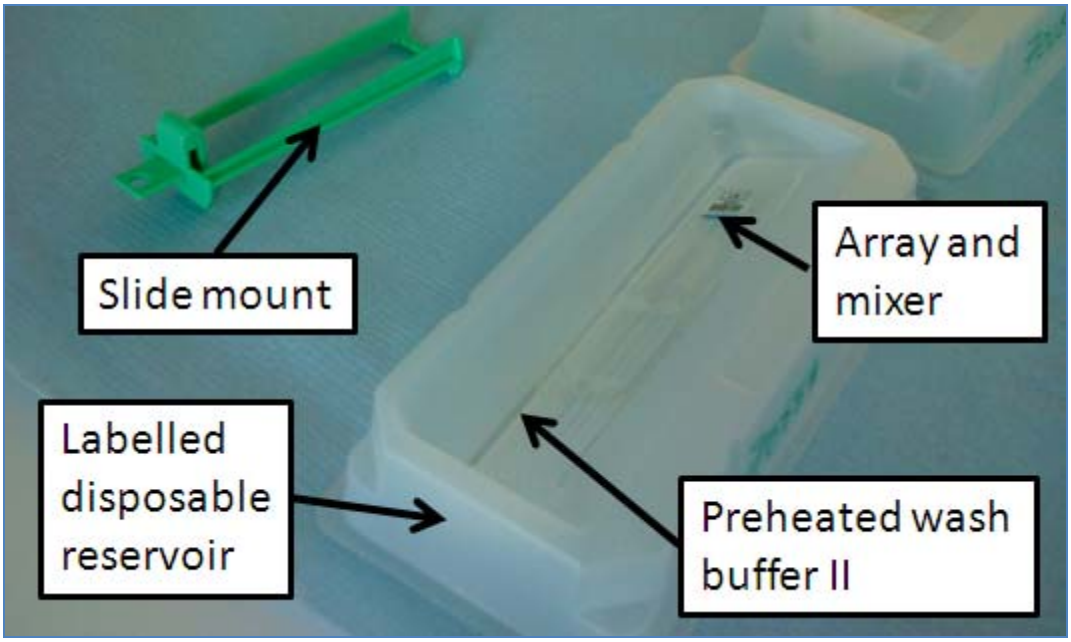
1. Ensure that there are the wash solutions as detailed below are prepared, aliquoted into falcon tubes where necessary and preheated if needed before starting the washing procedure.

NB: These quantities are for one array and must be adjusted to reflect the number of arrays to be processed.

Number of aliquots	Aliquot volume (ml)	Wash solution	Container	Temperature (°C)
1	100	Wash buffer II	Ready for reservoir	42
2	32	Wash buffer II	Labeled 50ml falcon tube	20
2	32	Stringent buffer	Labeled 50ml falcon tube	47.5
1	32	Wash buffer I	Labeled 50ml falcon tube	20
1	32	Wash buffer III	Labeled 50ml falcon tube	20

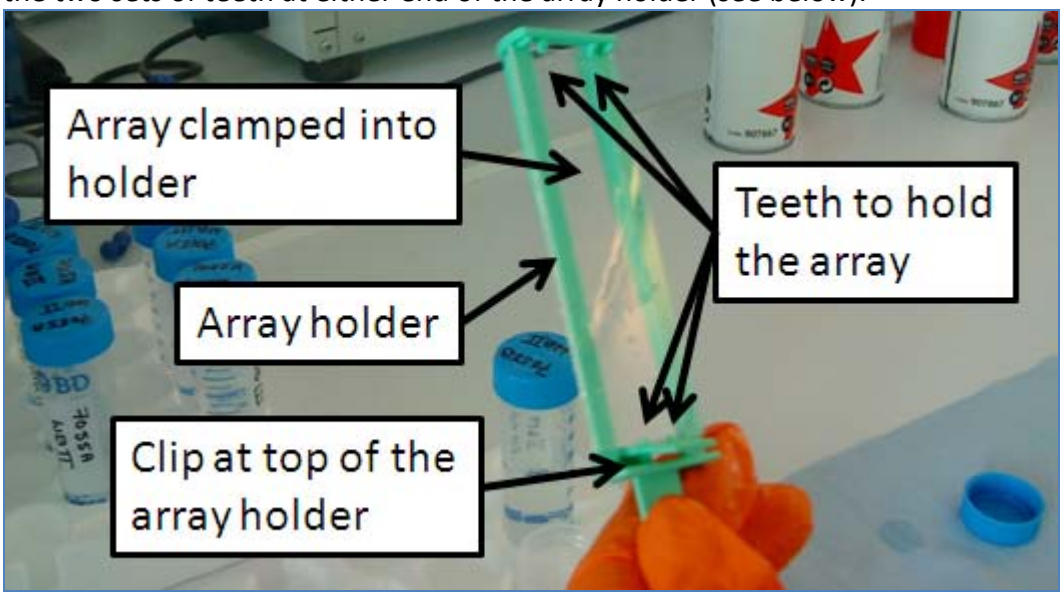
2. Label a new disposable reservoir with the sample number for each of the arrays to be washed. Remove the array from the hybridisation station and place into the reservoir. Cover the array with the 100 ml of preheated wash buffer II (see below).

Nature Methods: doi: 10.1038/nmeth.1419



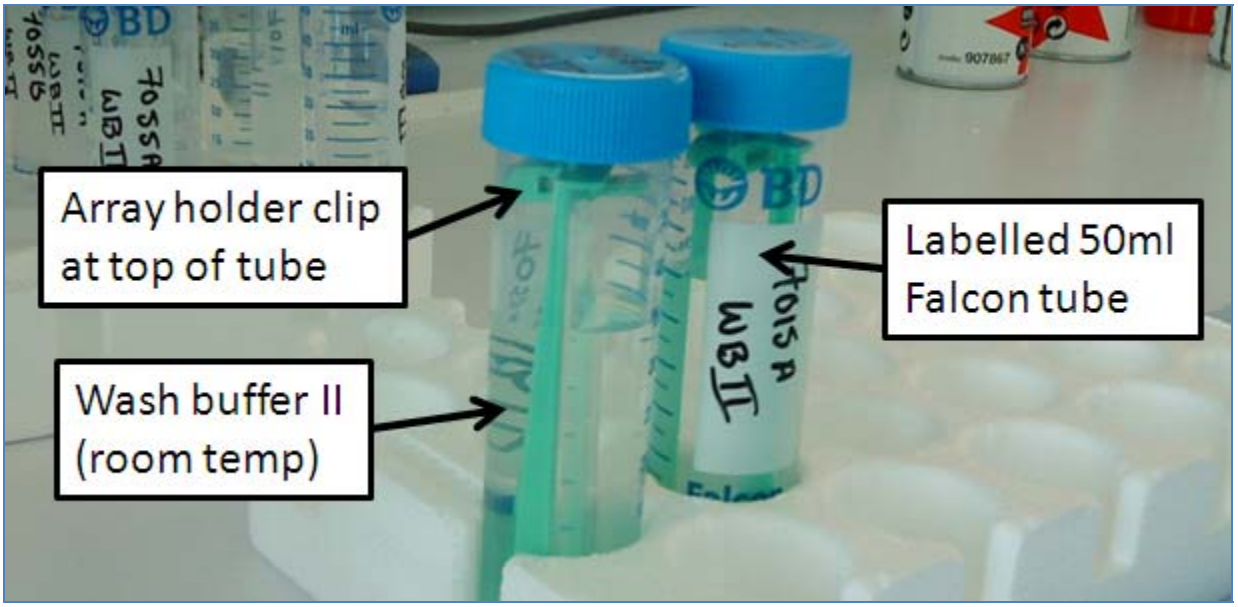
3. Carefully peel off the mixer from the array taking care not to touch the area of the array under the mixer. The mixer may now be discarded.

4. Holding the array sides only, remove the array from the reservoir and clip it into an array holder by pressing down on the top clip of the holder. The array must go under the two sets of teeth at either end of the array holder (see below).



5. Place the array-holder assembly into a labelled 50 ml Falcon tube containing room temperature wash buffer II (32.0 ml). Ensure that the array holder is orientated with the clip at the top of the tube (see below). Replace the lid of the Falcon tube.

Nature Methods: doi: 10.1038/nmeth.1419



6. Invert the Falcon tube containing the array-holder assembly ten times pausing for 1 second after each inversion before inverting again.

7. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated stringent buffer (32.0 ml at 47.5 ^{o}C). Replace the lid of the Falcon tube.

8. Invert the Falcon tube containing the array-holder assembly 10 times pausing for 1 second after each inversion before inverting again. When 10 inversions have been completed place the falcon tube with the array-holder assembly into a water bath at 42.0 ^{o}C for 5 minutes.

9. Following the incubation in the water bath invert the Falcon tube containing the array-holder assembly a further ten times pausing for 1 second after each inversion before inverting again.

10. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing a second aliquot of preheated stringent buffer (32.0 ml at 47.5 ^{o}C). Replace the lid of the Falcon tube.

11. Invert the Falcon tube containing the array-holder assembly 10 times pausing for 1 second after each inversion before inverting again. When 10 inversions have been completed place the falcon tube with the array-holder assembly into a water bath at 42.0 ^{o}C for 5 minutes.

Nature Methods: doi: 10.1038/nmeth.1419

12. Following the incubation in the water bath invert the Falcon tube containing the array-holder assembly a further ten times pausing for 1 second after each inversion before inverting again.

13. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer I (32.0 ml). Replace the lid of the Falcon tube.

14. Invert the Falcon tube containing the array-holder assembly for 2 minutes pausing for 1 second after each inversion before inverting again.

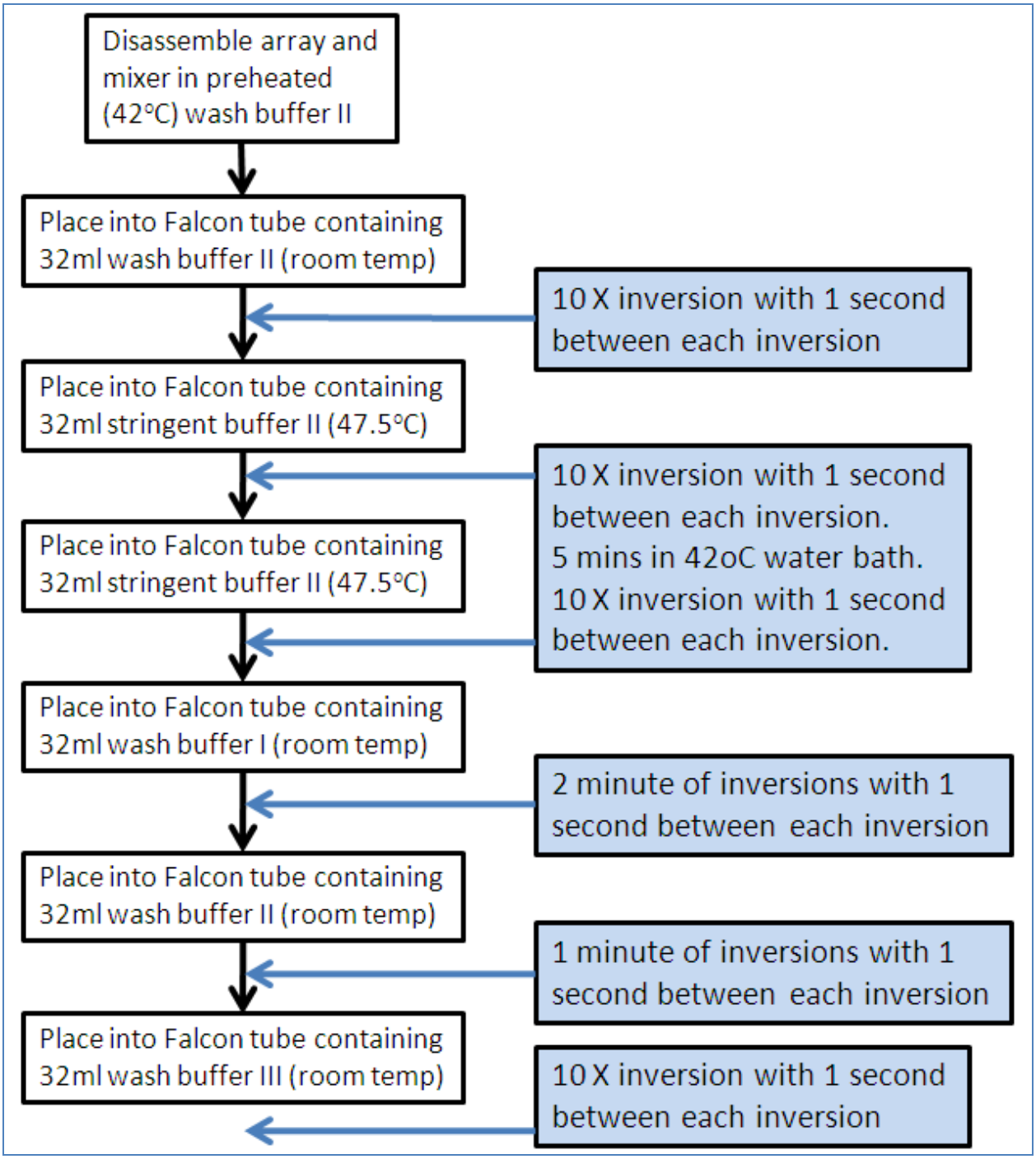
15. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer II (32.0 ml). Replace the lid of the Falcon tube.

16. Invert the Falcon tube containing the array-holder assembly for 1 minute pausing for 1 second after each inversion before inverting again.

17. Remove the array-holder assembly from the Falcon tube and place into a new labelled 50 ml Falcon tube containing preheated wash buffer III (32.0 ml). Replace the lid of the Falcon tube.

18. Invert the Falcon tube containing the array-holder assembly ten times pausing for 1 second after each inversion before inverting again (see below for flow diagram of washes).

Nature Methods: doi: 10.1038/nmeth.1419

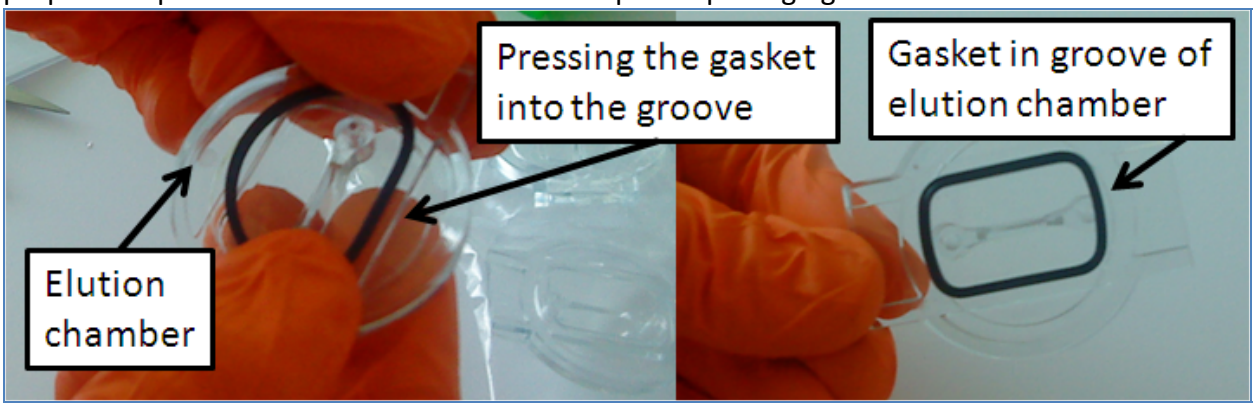


19. The array is now ready for the elution process.

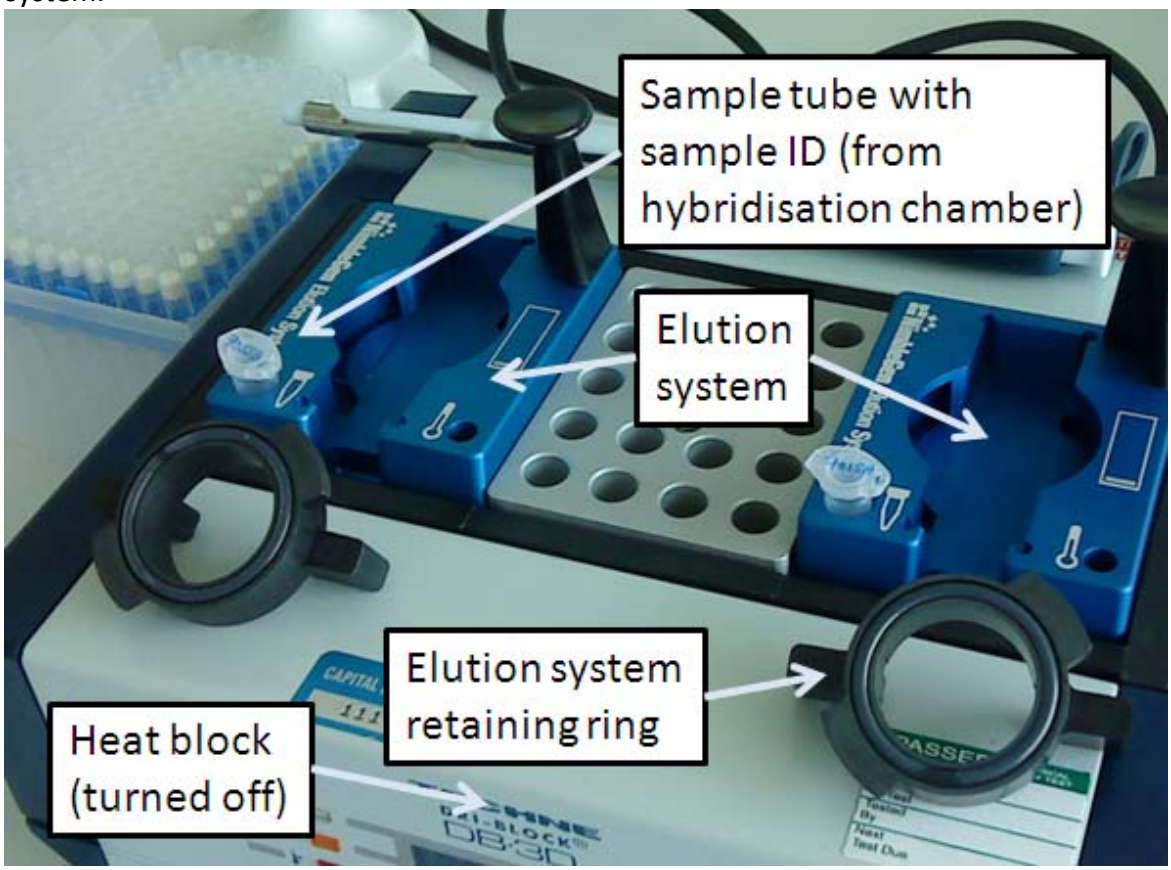
Nature Methods: doi: 10.1038/nmeth.1419

Eluting the DNA Using Sodium Hydroxide (NaOH)

1. Unwrap a new, clean EL1 elution chamber and rubber gasket. Prepare the elution chamber by pressing the gasket into the grove around the edge (see below). Once prepared replace the elution chamber into the plastic packaging until needed.

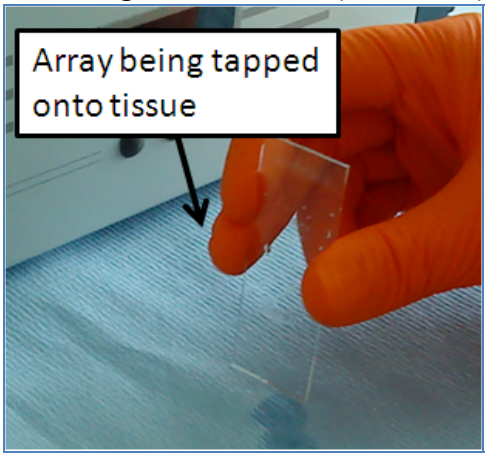


2. Place the elution system blocks into a heat block that is turned off and at room temperature for stability (see below) or place them onto the bench top. Place an elution system retaining ring ready to lock down the elution chamber near to the elution system.

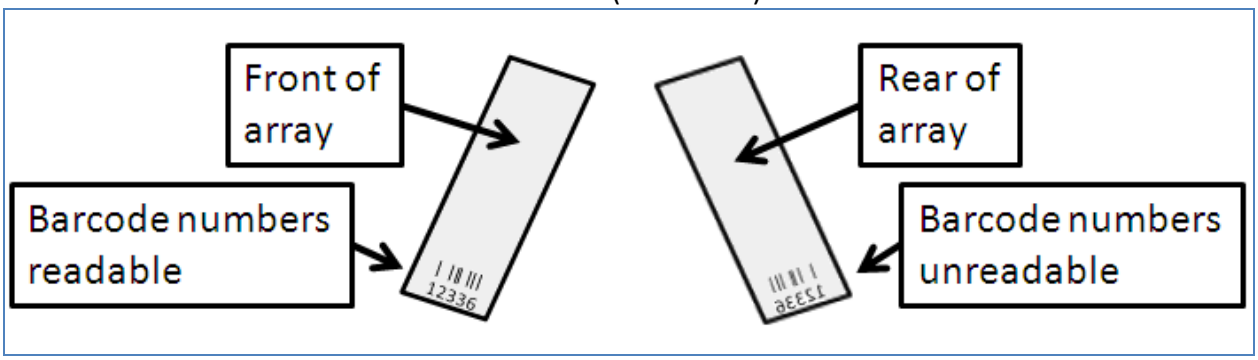


Nature Methods: doi: 10.1038/nmeth.1419

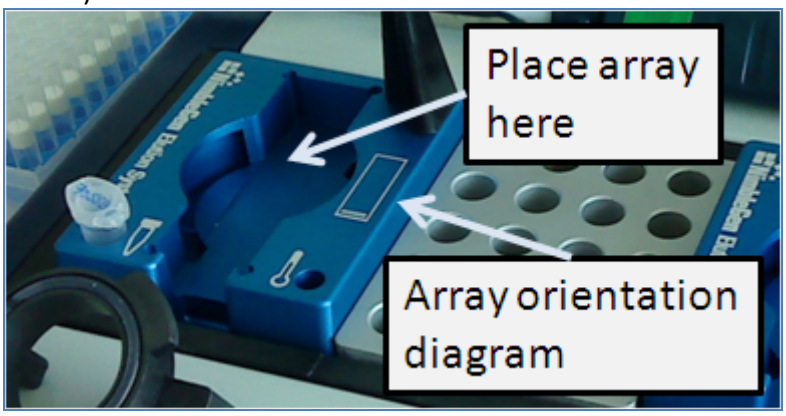
3. Remove the slide from the wash buffer III (final wash above). Unclip the array from the array holder by bending back the clip at the top of the holder. Whilst only holding the sides of the array, gently tap the bottom of the array onto some tissue to remove any remaining wash buffer III (see below)



4. Look at the array barcode (which is etched into the glass) to determine the front and back faces of the array. The front face is the side through which the numbers of the barcode can be read in the correct orientation (see below).

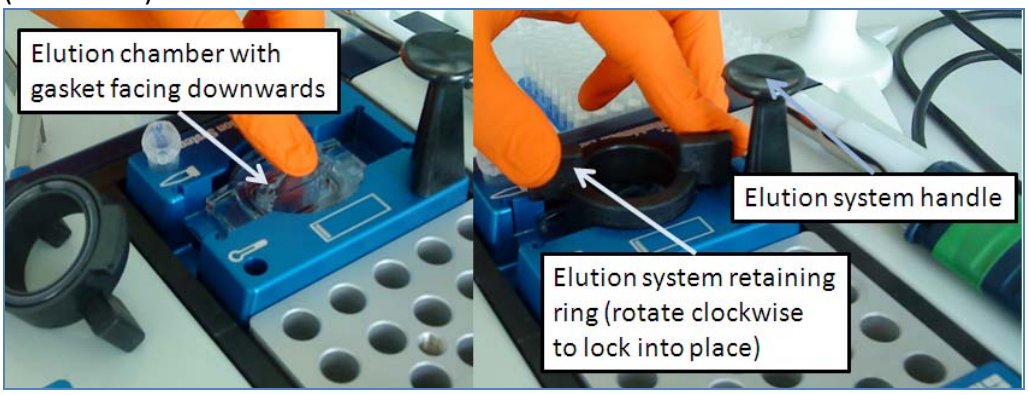


5. Wipe the rear face of the array with a low-lint tissue. Insert the array into the elution system with the barcode orientated as shown on the side of the elution system (see below).



Nature Methods: doi: 10.1038/nmeth.1419

6. Place the elution chamber (prepared as above – Step 1) into the round section of the elution system orientated so that the gasket forms a seal on the array. Place the elution system retaining ring over the elution chamber and twist clockwise to lock into position (see below).

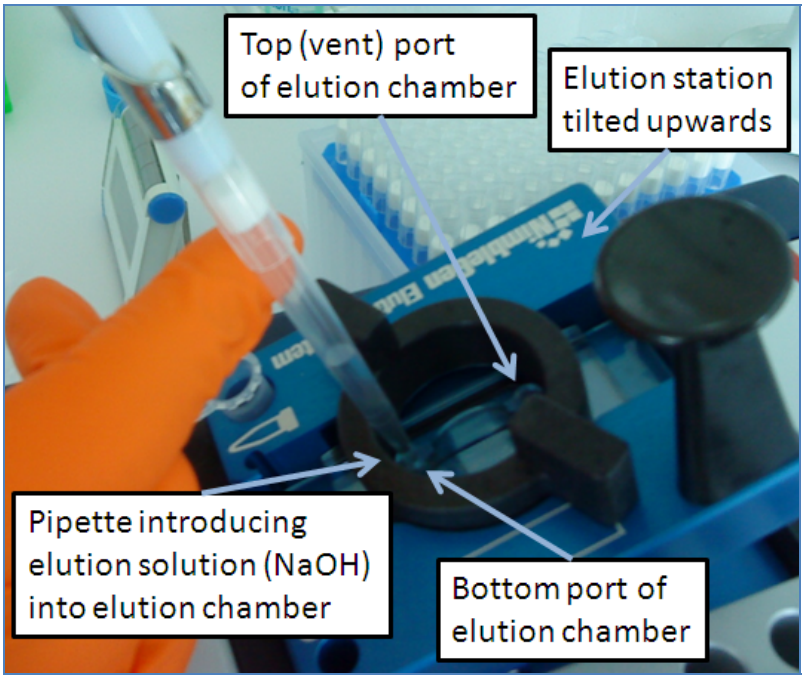


7. Tilt the elution system upwards by pulling on the elution system handle until it is locked in place.

8. Label a clean Eppendorf LoBind 1.5 ml tube with the sample name. Pipette 900 µl of the freshly prepared 125 mM NaOH into the tube.

9. Using a pipette introduce the freshly prepared sodium hydroxide (NaOH) solution (section 'Preparation of the Elution and Clean-up Solutions' above) into the elution chamber by placing the pipette tip into the bottom port of the elution chamber (see below). Introduce the liquid slowly and keep adding liquid until liquid can be seen rising up the top (vent) port of the elution chamber. **NB:** For 2.1 M arrays about 900 µl of NaOH is needed.

Nature Methods: doi: 10.1038/nmeth.1419



10. Leave the pipette tip in the bottom vent of the elution chamber to prevent any elution liquid from leaking out of the elution chamber. Return the elution system to the horizontal position and remove the pipette tip. Return any unused NaOH to the labelled 1.5 ml Eppendorf LoBind tube that it came from.

11. Incubate the array in the elution system at room temperature for 10 minutes.

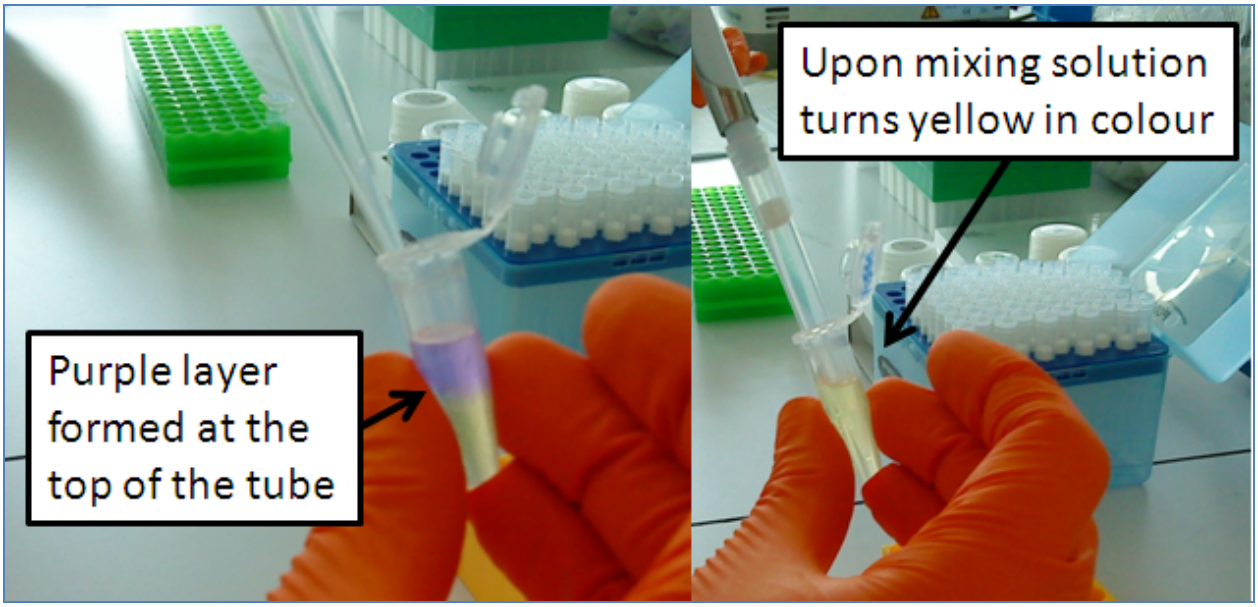
12. Following the incubation pipette out eluted DNA and NaOH from the elution chamber and return to the unused, excess NaOH in the labelled Eppendorf LoBind 1.5 ml tube.

13. Twist the elution system retaining ring anti-clockwise to release the elution chamber which can now be discarded. If necessary the elution system can be tilted again and any remaining NaOH be pipetted out and added to the rest of the eluted DNA and NaOH. The array may now be disposed of also. Clean the elution systems and retaining ring with Azo-wipes before storage.

14. Label two 1.5 ml Eppendorf LoBind tubes with the sample details and add 16 µl of 20 % acetic acid and 500 µl of PB buffer to each. Pipette up and down 10 times to mix thoroughly.

15. Split the eluate (~900 µl) and pipette about 450 µl into each of the labelled 1.5 ml Eppendorf LoBind tubes containing the acetic acid-PB buffer mixes. When first added the eluates sit on the top of the acid/buffer mixes and purple layers form at the top of the tubes (see below). Mix the solutions by pipetting up and down 10 times. The solutions should now be yellow, if it is purple the add 20 % acetic acid, 1 µl at a time mixing after each addition until it turns yellow.

Nature Methods: doi: 10.1038/nmeth.1419



Nature Methods: doi: 10.1038/nmeth.1419

16. Remove a QIAquickMinElute column (already seated in a collection tube) from the kit stored in the fridge and place into a tube rack on the bench. Label the lid of the column with the sample identifying information that appears on the eluted sample tube lids.

17. Carefully pipette 750 µl of the first DNA-PB buffer-acetic acid mixes into the appropriately labelled column ensuring that all of the liquid is dispensed from the pipette tip and that the tip does not contact the membrane of the column. **NB:** At this point there will be a column with 750 µl of DNA-PB buffer-acetic acid mix loaded onto it and a 1.5 ml Eppendorf containing the remaining 216 µl of DNA-PB buffer-acetic acid mix and a second 1.5 ml Eppendorf containing 966 µl of DNA-PB buffer-acetic acid mix. Close the lid of the first 1.5 ml Eppendorf LoBind tube and keep on the bench until step 23.

18. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.

19. Remove the column from the centrifuge and discard the liquid (eluate) that has been spun into the collection tube in which the column was seated. The column may be re-seated into the same collection tube that it was spun in after discarding the eluate.

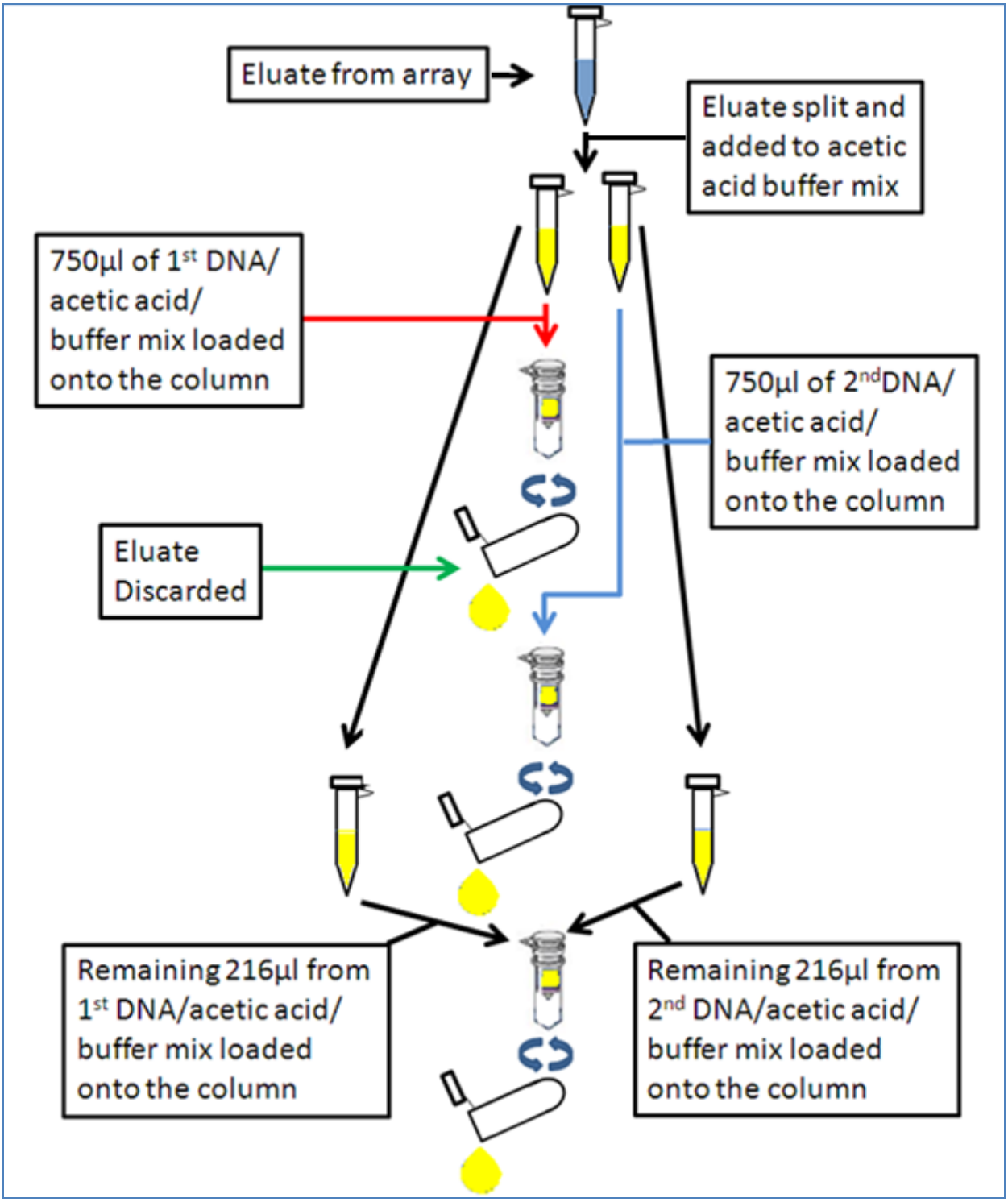
20. Carefully pipette 750 µl of the second DNA-PB buffer-acetic acid mixes from the second 1.5 ml Eppendorf LoBind tube into the appropriately labelled column. **NB:** At this point there will be a column with a second aliquot (750 µl) of DNA-PB buffer-acetic acid mix loaded onto it and two 1.5 ml Eppendorfs containing 216 µl of DNA-PB buffer-acetic acid mix each. Close the lid of the second 1.5 ml Eppendorf LoBind tube and keep on the bench until step 23.

21. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.

22. Remove the column from the centrifuge and discard the liquid (eluate) that has been spun into the collection tube in which the column was seated. The column may be re-seated into the same collection tube that it was spun in after discarding the eluate.

23. Carefully pipette the two remaining 216 µl aliquots of the DNA-PB buffer-acetic acid mixes from both of the 1.5 ml Eppendorf tubes into the appropriately labelled column.

Nature Methods: doi: 10.1038/nmeth.1419



24. Close the lids of the column and centrifuge it at 13,000 rpm for 1 minute.

Nature Methods: doi: 10.1038/nmeth.1419

25. Remove the column from the centrifuge and discard the eluate that has been spun into the collection tube in which the column was seated. The column may again be re-seated into the same collection tube that it was spun in after discarding the eluate.

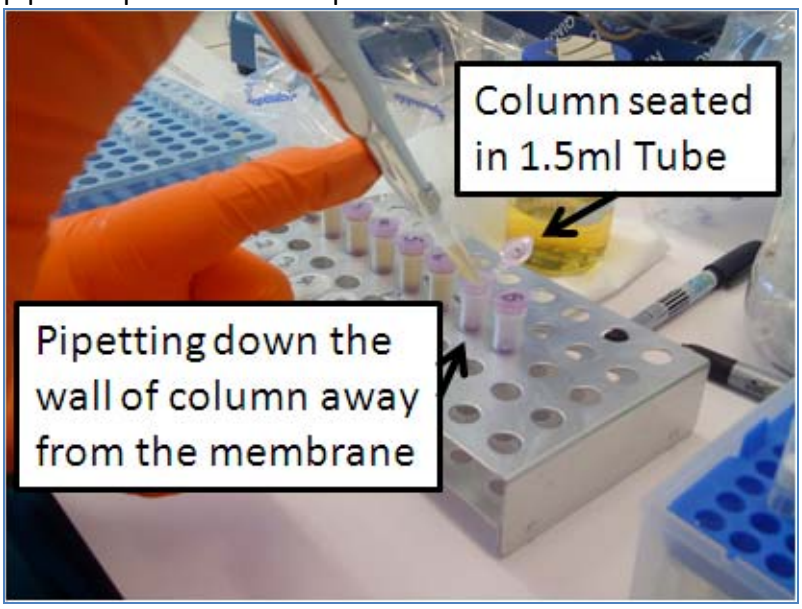
26. Using a clean pipette tip, pipette 750 µl of PE buffer (from the QIAquickPCR purification kit) into the column ensuring that all of the liquid is dispensed from the pipette tip and that the tip doesn't contact the membrane of the column.

27. Close the lid of the column and centrifuge it at 13,000 rpm for 1 minute.

28. Remove the column from the centrifuge and discard the eluate that has been spun into the collection tube in which the column was seated. This collection tube may now be discarded. Place the column into a clean 1.5 ml Eppendorf LoBind tube and place into a rack on the bench.

29. Leaving the column lid open, incubate the column and tube in the rack on the bench for 2 minutes to dry. While the columns are drying a clean 1.5 ml Eppendorf LoBind tube with the sample identifier information written on the column lid and leave in a rack on the bench.

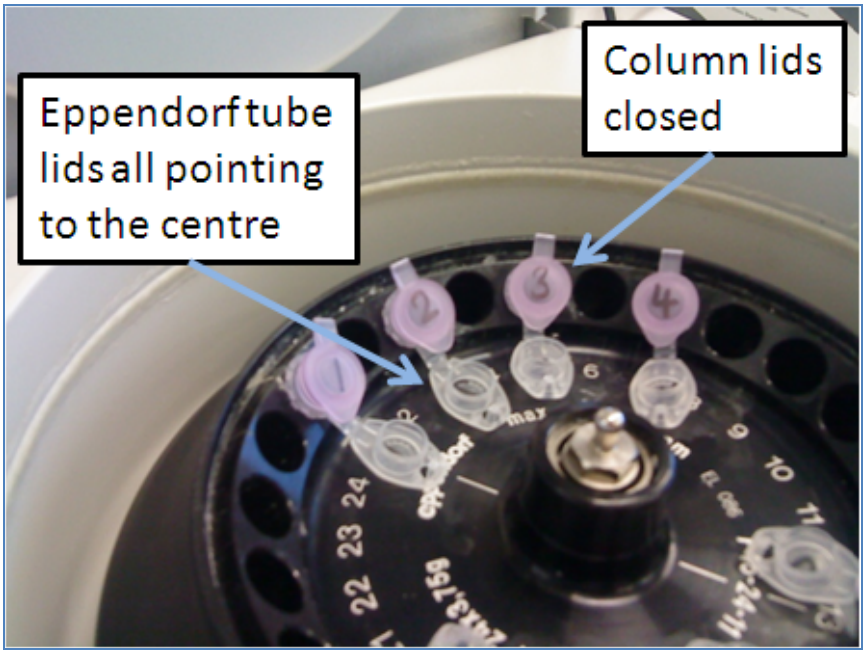
30. Transfer the column into the appropriately labelled clean 1.5 ml Eppendorf LoBind tube from step 29. Using a clean pipette tip add 34 µl of EB buffer (from the QIAquickPCR purification kit) into the column ensuring that all of the liquid is dispensed from the pipette tip and that the tip doesn't contact the membrane of the column (see below).



31. Close the lids of the columns and leave the columns on the bench for 1 minute.

Nature Methods: doi: 10.1038/nmeth.1419

32. Arrange the tube in the centrifuge rotor so that all of the Eppendorf LoBind tube lids all point towards the centre (see below), this will help prevent the Eppendorf LoBind tube lid from breaking off during the centrifugation. Centrifuge the column at 13,000 rpm for 1 minute.



33. Remove the columns and 1.5 ml Eppendorf LoBind tubes from the centrifuge and **DO NOT** discard the eluate that has been spun into the collection tube. The QIAquickMinElute column can now be discarded.

Proceed to the Post-hyb PCR (Supplementary Protocol 6).

Nature Methods: doi: 10.1038/nmeth.1419

Supplementary Protocol 5

Solution capture standard operating procedure (Agilent SureSelect protocol)

This protocol is for hybridization of adapter-ligated or PCR-amplified library DNA, so must be performed after Hybrid Capture Protocol 1.

Sample preparation

Following SPRI bead cleanup (see Hybrid Capture Protocol 1), lyophilize 500 ng library + 7.5 µg C_0t1 DNA using a SpeedVac..

1. Add 3.4 µl of molecular biology grade water to each sample to rehydrate.

2. Vortex and spin down.

3. To each sample, add:

SureSelect Block #1	2.5 µl
SureSelect Block #2	2.5 µl
SureSelect Block #3	0.6 µl

4. Prepare Hybridization Buffer as follows. Volume for 1 capture:

SureSelect Hyb #1	25 µl
SureSelect Hyb #2	1 µl
SureSelect hyb #3	10 µl
SureSelect Hyb #4	13 µl

   Note: Do NOT keep on ice.

5. Incubate 40 µl of Hybridization Buffer and library (from step 3.) at 95 °C for 5 minutes and 65 °C for at least 5 minutes. Keep at 65 °C until RNA baits are prepared (see below).

6. Dilute RNAse Block in 1:1 ratio with nuclease-free water and add 1 µl to 5 µl (500 ng) of RNA baits. Incubate for 2 min at 65 °C.

Hybridization

1. Mix 13 µl of hybridization buffer with RNA baits.

Nature Methods: doi: 10.1038/nmeth.1419

2. Add the 9 µl DNA library to the hyb buffer - RNA bait mixture.

3. Seal plate with the Greiner film and incubate for 24 hours at 65 °C with a heated lid at 105 °C.

Washing and elution

Selection with magnetic beads

1. Prepare magnetic beads by washing with SureSelect Binding buffer 3 times and resuspend in 200 µl of SureSelect Binding buffer. Incubate the hybrid-capture-bead solution on a Nutator for 30 minutes at room temperature (20 °C).

2. Mix hyb. mixture with beads and separate the beads by removing the supernatant.

3. Resuspend beads in SureSelect Wash Buffer #1.

4. Incubate the samples for 15 minutes at room temperature. Separate the beads and buffer on a Dynal magnetic separator and remove the supernatant.

5. Mix the beads in pre-warmed (65 °C) 500 µL SureSelect Wash Buffer #2. Incubate the samples for 10 minutes at 65 °C. Remove wash buffer and repeat these steps for 3 times.

6. Mix the beads in 50 µL SureSelect Elution Buffer. Incubate the samples for 10 minutes at room temperature. Separate the beads and buffer on a Dynal magnetic separator. Add 50 µL of SureSelect Neutralization Buffer.

7. Desalt the capture solution with a Qiagen MinElute PCR purification column, eluting in 34 µL buffer EB.

Proceed to the Post-hyb PCR (Supplementary Protocol 6).

Nature Methods: doi: 10.1038/nmeth.1419

Supplementary Protocol 6

Hybrid capture eluate PCR standard operating procedure (for array and solution capture eluates)

This protocol is for post-elution amplification of captured DNA from an array or solution capture (i.e. following on from Supplementary Protocol 4 or 5).

PCR primers:
PE.1 = 5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T 3′
PE.2 = 5′ CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T 3′

* indicates phosphorothioate. Both primers are HPLC purified.

1. Prepare the PCR master mix for 1 sample:

10 x PCR Buffer	5 µl
25 mM $MgCl_2$	4 µl
2.5 mM dNTPs	4 µl
10 µM PE.1	2.5 µl
10 µM PE.2	2.5 µl
Platinum®*Pfx* DNA Polymerase (Invitrogen)	0.4 µl
Sample (from capture)	31.6 µl

The total volume of the final reaction should be 50 µl.

2. Vortex briefly and spin down.

3. Run on thermocycler with the following program:

94 °C for 5 min

94 °C for 15 sec
58 °C for 30 sec x 18 cycles
72 °C for 30 sec

72 °C for 5 min
4 °C indefinitely

4. Transfer PCR product into a 1.5ml Lo-Bind tube. Run the sample on an Agilent DNA 1000 chip on a Bioanalyzer 2100. If PCR was successful, proceed to the SPRI clean up:

Nature Methods: doi: 10.1038/nmeth.1419

SPRI bead cleanup

Allow SPRI beads to come to room temperature for at least 30 minutes. Reagents need to be mixed well prior to use and should appear homogeneous and consistent in colour.

1. Take 90 µl of SPRI beads and add them to the 50 µl of PCR sample in a 1.5 ml Lo-Bind tube.

2. Vortex and hold at room temperature for 5 minutes.

3. Place tube in the magnetic rack and leave for 5 minutes or until sample is clear.

4. Carefully remove the clear solution from the tubes and discard.

5. Dispense 500 µl of 70 % ethanol into each tube while in the magnetic rack taking care not to disturb the magnetic beads. Aspirate and discard ethanol.

6. Repeat the ethanol wash once again. Total of two washes.

7. Dry the samples on a heat block (keep the lid of the tube open) at 37 °C for 5 – 10 minutes or until the residual ethanol has evaporated.

8. Add 50 µl of molecular biology grade water, vortex and incubate at room temperature (20 °C) for 2 minutes.

9. Place tubes into the magnetic rack and leave for 2-3 minutes or until sample is clear.

10. Carefully remove the water and retain in a new 1.5 ml Lo-Bind tube.

11. Repeat step 9 -12 once more, retaining the water in the same 1.5 ml Lo-Bind tube. Total volume of elute should be 100 µl.

12. Put the tube into the magnetic tool for 10 min.

13. Transfer the sample to a new 1.5 ml Lo-Bind tube leaving behind any precipitated beads.

After SPRI clean up, quantify by qPCR [1] and sequence.

References

1. Quail M.A., Swerdlow H. & Turner D.J. Improved protocols for the Illumina genome analyzer sequencing system. *Curr Protoc Hum Genet* **Chapter 18**, Unit 18 12 (2009).

Nature Methods: doi: 10.1038/nmeth.1419

ATTACHMENT 7A

LDR	01559pas a22005294a 4500
001	9912156043406676
005	20250227135121.0
007	cr|cn ||||||||
008	040219c20049999nyumr p 0 0eng c
010	__ | **a** 2004214152
022	0_ | **a**1548-7091**l**1548-7091**9**PY**7**(Print)
022	0_ | **a**1548-7105**9**EN**7**(Electronic)
035	__ | **9**101215604
040	__ | **a**NSD**c**NSD
041	0_ | **a**eng
042	__ | **a**lcd**a**nlmcopyc
044	__ | **9**United States
060	00 | **a**W1**b**NA812N
210	10 | **a**Nat Methods**2**DNLM
245	00 | **a**Nature methods.
260	__ | **a**New York, NY :**b**Nature Pub. Group,**c**c2004-
300	__ | **a**v. :**b**ill.
310	__ | **a**Monthly
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**Vol. 1, no. 1 (Oct. 2004)-
500	__ | **a**Title from cover.
510	1_ | **9**1**a**MEDLINE**b**v1n1, Oct. 2004-
510	1_ | **9**1**a**PubMed**b**v1n1, Oct. 2004-
510	1_ | **a**Index medicus
530	__ | **a**Also issued online.
650	12 | **a**Biomedical Research**x**methods
650	22 | **a**Research Design
655	_2 | **a**Periodical
856	4_ | **u**http://www.nature.com/nmeth/index.html
911	__ | **a**Electronic and print, print canceled 2017
914	__ | **a**Portico**b**Current to 2024 as of 20250227**c**ISSN_15487091
935	__ | **a**(DNLM)1215604-nlmdb
992	__ | **p**P2**e**E4**a**20050122
993	__ | **a**YMG**b**20041124
994	__ | **a**CDN**b**20041124
995	__ | **a**AUTH**b**20041206**c**REV**d**20250227
996	__ | **a**CONSER**b**20000101
998	__ | **a**ERMSLC 108646**b**20090131

999	__ | **a**AUTH
035	__ | **a**(OCoLC)54447242

ATTACHMENT 7B

Biomedical Research MeSH Descriptor Data 2025

MeSH Heading
Biomedical Research
Tree Number(s)
H01.770.644.145
Unique ID
D035843
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D035843
Annotation
biomedical research as a field or the progress or status of biomedical research; do not add routinely for specific studies in biomedicine
Scope Note
Research that involves the application of the natural sciences, especially biology and physiology, to medicine.
Entry Version
BIOMED RESEARCH
Entry Term(s)
Experimental Medicine
Investigational Medicine
Investigative Medicine
Medical Research
Research, Biomedical
Research, Medical
Previous Indexing
Research (1965-2002)
See Also
Animals, Laboratory
Public MeSH Note
2003
History Note
2003
Date Established
2003/01/01
Date of Entry
2002/07/02
Revision Date
2015/11/23

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Analytical Biochemistry 349 (2006) 242–246
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

Structure-independent and quantitative ligation of single-stranded DNA

Tian W. Li, Kevin M. Weeks *

Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599, USA

Received 13 September 2005
Available online 18 November 2005

Abstract

Ligation of an adapter oligonucleotide to a single-stranded cDNA is central to many molecular biology techniques. Current single-stranded ligation approaches suffer from low efficiencies and are strongly inhibited by preexisting DNA secondary structure. We develop an approach for ligating low concentrations of single-stranded DNAs to a DNA adapter with near-quantitative efficiency, unaffected by secondary structure in the target DNA. This efficient DNA ligation reaction will facilitate development of robust procedures for quantifying small amounts of highly structured cDNAs and their RNA templates.
© 2005 Elsevier Inc. All rights reserved.

Keywords: cDNA; Single-stranded DNA ligation

Many molecular biology manipulations, focused especially on analyzing RNA sequences and structures in cells, require that an adapter oligonucleotide be ligated onto either the 5′ or 3′ end of a target cDNA derived from an RNA of interest. Important methods in this class include mapping the 5′ or 3′ terminus of an RNA [1,2] and performing in vivo analysis of RNA structure [3,4]. A serious shortcoming of current adapter ligation reactions is that the efficiency for direct ligation of an oligonucleotide to a single-stranded cDNA is variable and is often quite low. Ligation efficiencies also decrease significantly with increasing target DNA length [5]. Thus, useful but complex approaches, involving nucleotide tailing steps prior to ligation with a double-stranded DNA adapter, have been developed [4,6]. The ligation and tailing steps are dependent on the accessibility and lack of base-paired structure at the end of the target cDNA. The recovery of a ligated product is then governed by both the abundance of a particular target cDNA and the net efficiency of adapter ligation and any nucleotide tailing steps.

We sought to ligate an adapter oligonucleotide to single-stranded cDNA-type targets at high efficiencies, independent of the secondary structure in the DNA. Using a thermostable enzyme with DNA ligase activity [7], we developed an approach for ligating single-stranded DNA targets with complex secondary structures at nearly quantitative efficiencies.

Materials and methods

Oligonucleotides

Oligonucleotides (synthesized by Midland Certified) for cDNA ligation were as follows: DNA adapter, 5′-p-ATCGA CAACA ACTCT CCTCC TCCGT GCG-(C3)-3′ (where p is a phosphate and C3 is 1,3-propanediol); DNA target (119 nt), 5′-GGTAA GCTGG CCAGC AACTT ATCTG TGTTT TTCCG ATTGT CTAGT GTCTA TGTTT GATGT TATGC GCCTG CGTCT GTACG TTCTG AACAC CCGGC CGCAA CCCTG GGAGA CGTCC CAGG-3′. PCR primers: forward, 5′-GGTAA GCTGG CCAGC AACTT ATCTG TGTTT TTCCG AT-3′; reverse, 5′-CGCAC GGAGG AGGAG AGTTG TTGTC GAT-3′. Reverse transcriptase primer for structured cDNA, 5′-GAACC GGACC GAAGC CCG-3′. Oligonucleotides for 3′ end labeling by single nucleotide extension: complement to DNA target, 5′-AAACC CTGGG ACGTC TCCCA GGGTT GCGG-3′;

* Corresponding author. Fax: +1 919 962 2388.
E-mail address: weeks@unc.edu (K.M. Weeks).

0003-2697/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2005.11.002

complement to reverse transcriptase primer, 5′-GGGTC GGGCT TCGGT CCGGT TC-3′.

3′ ^{32}P end labeling of DNA oligonucleotides

The DNA target and cDNA reverse transcriptase primer were 3′ end labeled using Klenow(-exo) polymerase and annealing to a complementary strand that allows incorporation of a single nucleotide [8]. Reactions (20 µl, 37 °C, 40 min) contained 4 µM DNA to be labeled, 4–8 µM complement, 5 U Klenow(-exo) DNA polymerase (New England Biolabs), and 2.5 µM of either α-[^{32}P]dGTP or α-[^{32}P]dATP. Reactions were quenched by dilution with formamide and purified by denaturing gel electrophoresis (10–20% [w/v] polyacrylamide, 90 mM Tris–borate, 2 mM ethylenediaminetetraacetic acid [EDTA],[1] 7 M urea). DNAs were recovered from the gel by passive elution at 37 °C for 2 h and ethanol precipitation. Target DNA and cDNA reverse transcriptase primer were resuspended in water at final concentrations of 0.1 and 0.3 µM, respectively.

Single-stranded DNA ligation

Ligation reactions (10 µl, 68 °C, 1 h) contained 50 mM Mops (pH 7.5), 1 mM dithiothreitol (DTT), 10 mM KCl, 5 mM $MgCl_2$, 2.5 mM $MnCl_2$, 20% (w/v) polyethylene glycol 6000 (PEG 6000), 25 µM ATP, ^{32}P-labeled target DNA, 1 µM adaptor DNA, and 10 U Thermophage DNA ligase (Prokaria). Ligation products were resolved by denaturing electrophoresis in 10% polyacrylamide gels and were quantified using a Phosphorimager (Molecular Dynamics).

Ligation-mediated PCR

Aliquots (8 µl) from each ligation reaction were precipitated with ethanol in the presence of 20 µg glycogen. Pellets were resuspended in 74 µl water, and one-half of this mixture was used in 100-µl PCR reactions (containing 20 mM Tris–HCl [pH 8.4], 50 mM KCl, 0.2 mM each dNTP, 2.5 mM $MgCl_2$, 0.75 µM forward primer, 0.84 µM reverse primer, 2.5 U *Taq* polymerase). After 32 PCR cycles (94 °C, 30 s; 60 °C, 30 s; 72 °C, 1 min), products were analyzed on 1.4% (w/v) agarose gels.

Construction and analysis of 3′ nested set of cDNAs

tRNA transcripts were prepared as described previously [9]. Random 2′-*O*-adducts were introduced by heating 12 µl of 0.5 µM RNA (in 5 mM Tris [pH 8.0], 0.5 mM EDTA) to 95 °C for 1 min, adding 6 µl of 333 mM Hepes (pH 8.0), and treating with 2 µl of 65 mM *N*-methylisatoic anhydride (NMIA) for 1 min at 80 °C. Radiolabeled primer was added (6 µl at 0.3 µM) and annealed at 65 °C for 5 min and then at 35 °C for 20 min, and primer extension reactions were performed (in 40 µl) exactly as described previously [9]. RNA was subsequently degraded by alkaline hydrolysis (0.2 M NaOH, 95 °C, 5 min). The 3′ nested set of cDNAs, produced in the primer extension reaction, was initially resolved on a denaturing 20% polyacrylamide sequencing gel. Extension products were identified by comparison with an adjacent sequencing marker, generated by dideoxythymidine nucleotide incorporation during primer extension. cDNAs were then excised from the gel in subsets containing approximately 20 nt each (20–40, 40–60, 60–80, 80–100, 100–120, and 120–132 nt in length). Each size-fractionated subset was individually eluted from the gel (500 mM sodium acetate [pH 6.0], 1 mM EDTA, 37 °C, 2 h), concentrated by ethanol precipitation, and resuspended in 10 µl water. Ligation reactions were performed using 1 µl of each cDNA fraction. Ligation products were resolved on 14% denaturing polyacrylamide gels and visualized by phosphorimaging.

Results and discussion

Strategy for optimizing single-stranded DNA ligation

A model system for ligating a DNA adapter to a single-stranded target DNA is shown in Fig. 1. Ligase enzymes in

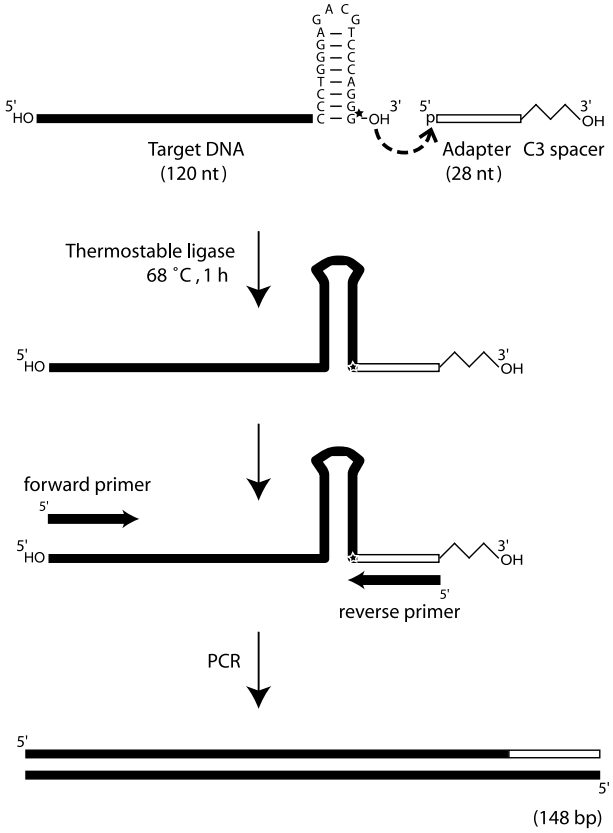


Fig. 1. End-selective single-stranded DNA ligation to create PCR templates. Dashed reaction arrow indicates ligation between 3′-OH and 5′-phosphate (p) groups. Star indicates site of 3′ radiolabel used to visualize ligation products (in Fig. 2A).

[1] *Abbreviations used:* EDTA, ethylenediaminetetraacetic acid; DTT, dithiothreitol; PEG 6000, polyethylene glycol 6000; NMIA, *N*-methylisatoic anhydride; SHAPE, selective 2′-hydroxyl acylation analyzed by primer extension.

the RNA ligase 1 family catalyze the ATP-dependent formation of a phosphodiester bond between a nucleotide 3′-hydroxyl nucleophile and a free 5′-phosphate group (Fig. 1, dashed reaction arrow) [10]. The adapter oligonucleotide was synthesized with a 5′-phosphate and a nonligatable 3′-C3 (1,3-propanediol) group to enforce unique ligation between the 3′-OH of the target DNA and the 5′-phosphate of the adapter (Fig. 1). The DNA target was radiolabeled at its 3′ end and was intentionally designed to form significant secondary structure at its 3′ terminus so that, from the onset, the ligation reaction would be optimized against an interfering secondary structure. The adapter oligonucleotide was designed such that it lacks significant secondary structure and simultaneously forms an accessible PCR primer binding site.

Efficiency of single-stranded DNA ligation

Our strategy for efficient single-stranded DNA ligation involves four elements. Reactions were performed (i) at an elevated temperature (68 °C) to reduce secondary structure in the DNA target, (ii) in 20% (w/v) PEG 6000 to increase macromolecular interactions between DNA substrates and the ligase enzyme, (iii) at saturating adapter DNA and ligase concentrations so that ligation efficiency is concentration independent for low amounts of target DNA, and (iv) with the use of an adapter DNA lacking significant self-structure.

Under these optimized conditions, the target DNA was ligated to the adapter oligonucleotide at efficiencies of at least 97% at DNA concentrations of 0.01–0.1 nM (See ligated product in Fig. 2A). The absence of a low-mobility product in reactions omitting the adapter DNA demonstrates that only linear ligated product, and not DNA circles, form in these reactions (Fig. 2A, see minus adapter lanes). At the higher target DNA concentrations of 1 and 10 nM, ligation efficiencies are 80 and 56%, respectively, and can be increased to at least 90% by extending the reaction time from 1 to 4 h (data not shown). Higher efficiencies at lower target DNA concentrations support the design principle that reactions should be performed at saturating adapter and ligase concentrations.

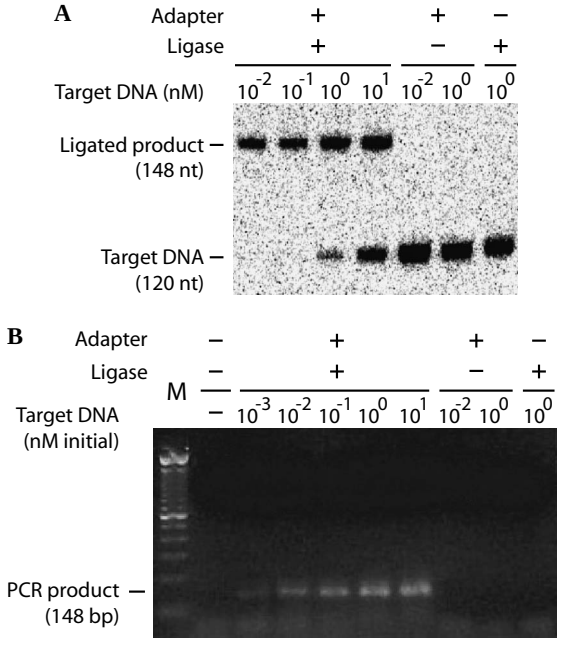


Fig. 2. Efficient single-stranded DNA ligation and selective PCR amplification. (A) Ligation reactions as a function of target DNA concentration (resolved by electrophoresis after adjusting all samples to equal ^{32}P disintegrations). (B) PCR amplification of target DNA adapter ligation products. Marker (M) is a double-stranded DNA ladder in 100-bp increments.

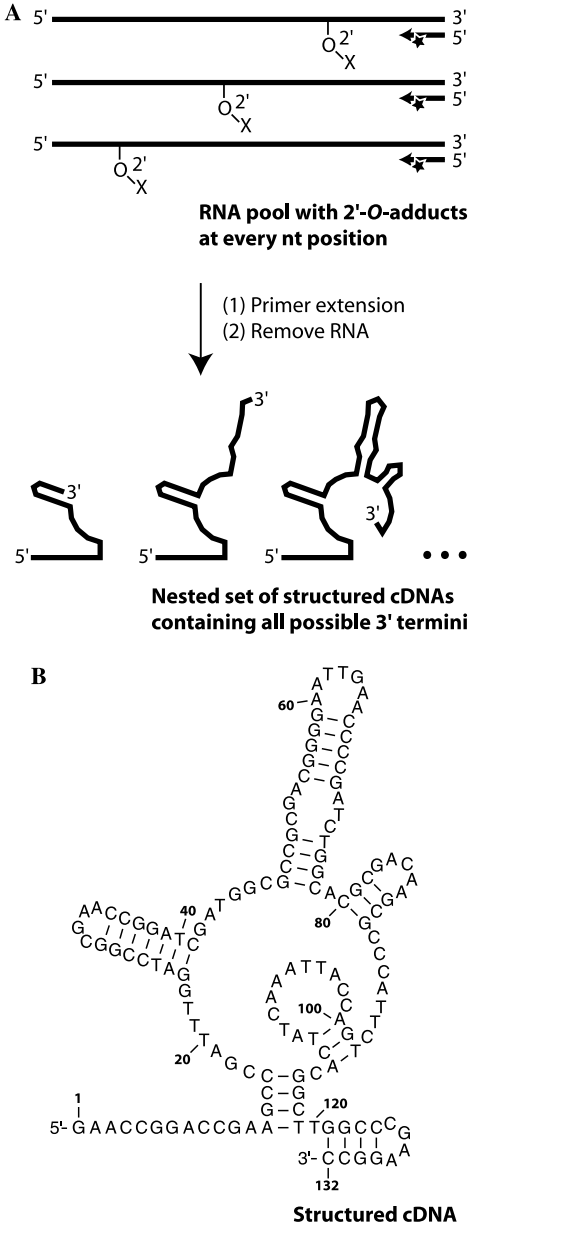


Fig. 3. Construction and secondary structure for cDNA target. (A) Scheme for synthesis of 3′ nested set of cDNA fragments. Reverse transcriptase-mediated primer extension was performed using RNA fragments carrying 2′-*O*-adducts generated by RNA SHAPE chemistry [9]. Site of ^{32}P radiolabel in reverse transcription primer is indicated with a star. (B) Calculated secondary structure for full-length cDNA target [12].

PCR amplification of ligated single-stranded DNAs

The products of the single-stranded DNA ligation can be used directly for PCR (Fig. 1). Ligation reactions (Fig. 2A) were recovered by ethanol precipitation and used without further purification in PCR (Fig. 2B). After 32 cycles, a single product is clearly observable by PCR for the ligated product formed from 10^{-2} nM starting target DNA (corresponding to 0.04 fmol DNA). PCR product synthesis, as expected, requires both the DNA adapter and ligase enzyme (Fig. 2B, control lanes). Thus, using an optimized approach for ligation of single-stranded DNAs, we form target DNA adapter products that are efficiently and selectively amplified by PCR.

Single-stranded DNA ligation independent of target DNA structure

We then evaluated the sensitivity of the single-stranded DNA ligation reaction to the secondary structure present at the 3′ end of a target DNA. We generated a set of DNAs with a common 5′ end as a complete set of nested 3′ extension products by performing a primer extension reaction using a previously described RNA template [9] (Fig. 3A). The RNA template was subjected to RNA selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry [9,11] under denaturing conditions to yield 2′-*O*-adducts at every internal RNA residue. These 2′-*O*-adducts prevent reverse transcriptase-mediated primer extension beyond the site of the adduct to yield a complete set of cDNA fragments containing every possible DNA from the 19-nt long primer to the 132-nt full-length product (Fig. 3A). Although there is some variation in individual bands, all expected extension products are observed as bands of significant intensity in a sequencing gel (Fig. 4, complete lane).

There are 113 different DNA products in the complete nested set, and the 3′ ends of these products lie in both highly structured stem loops and unpaired regions (Fig. 3B). For analysis, the complete nested set was divided into subsets containing approximately 20 DNA fragments each by excising and purifying cDNA fragments from a preparative gel (Fig. 4). Reactions were performed in the presence (+) and absence (−) of the thermostable ligase and were resolved on sequencing gels. For each size-fractionated cDNA subset, ligation to the adapter oligonucleotide is detected as the disappearance of a band in the (−) ligase lane and a shift to a product 28 nt longer in length in the (+) ligase lane. For clarity, products longer than 60 nt were subjected to extended electrophoresis to achieve improved resolution (Fig. 4, right panel).

In nearly every case, target DNAs are ligated at greater than 95% efficiency. For example, in the reactions of DNAs 40–60, 60–80, 80–100, and 100–120 nt in length, all bands corresponding to the starting DNAs are quantitatively

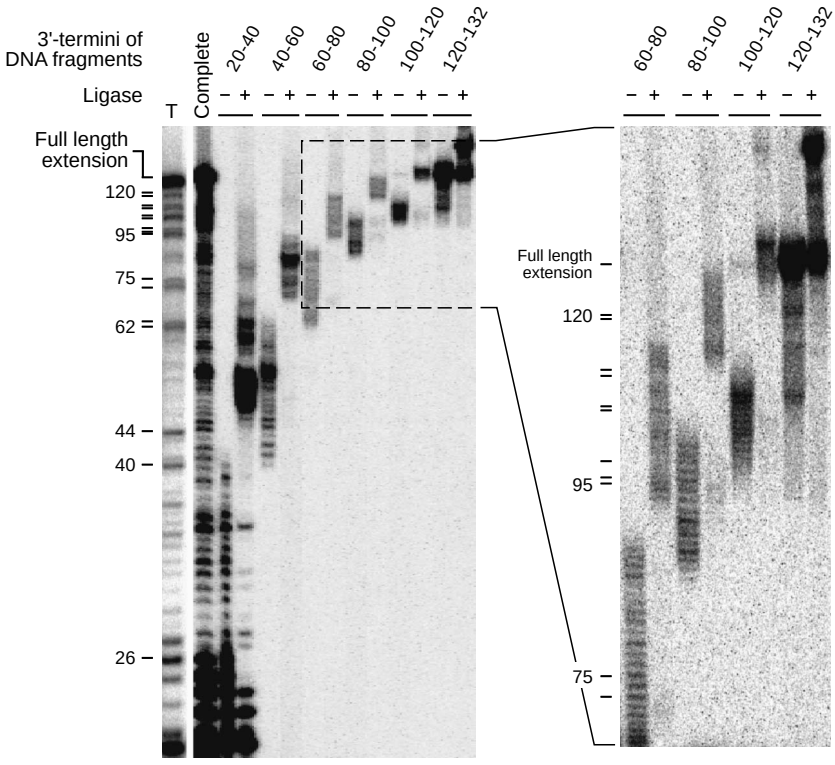


Fig. 4. Structure-independent ligation of nested structured cDNAs. Complete lane shows initial 3′ nested set of cDNA fragments; size-fractionated sets shown in remaining lanes were created from this parent ensemble. All cDNAs have identical 5′ ends but differ in length, as shown. cDNA fractions were incubated with the 28-nt adapter oligonucleotide in the presence (+) or absence (−) of thermostable ligase. Sequencing marker (T) was generated by incorporation of dideoxythymidine during primer extension from the RNA template; positions of thymidine residues are indicated at the left of each panel. Right-hand panel shows extended electrophoresis of selected reactions from main panel.

converted to product DNAs (compare − and + ligase lanes in Fig. 4). The main exception to quantitative ligation is the experiment involving DNAs 120–132 nt in length, where the full-length extension product was ligated to approximately 50%; most shorter cDNAs in this group were ligated at more than 80% efficiency.

Conclusion

Using a combination of a thermostable ligase, elevated temperature, PEG-induced molecular exclusion, and the forcing of product formation with saturating ligase and adapter DNA concentrations, we achieved efficient ligation of single-stranded DNAs independent of secondary structure in the DNA target. This approach is directly compatible with, and should yield immediate improvements to, many current strategies for analyzing RNA structure via cDNA intermediates. Efficient structure-independent cDNA ligation also represents an important technology for our ongoing program directed toward quantitative analysis of RNA structure in vivo.

Acknowledgments

This work was supported by grants from the University of North Carolina at Chapel Hill Office of Undergraduate Research and the Frances C. and William P. Smallwood Foundation (to Tian Li) and from the National Science Foundation (MCB-0416941 to Kevin Weeks). We are indebted to Kevin Wilkinson for assistance with RNA SHAPE chemistry.

References

[1] X. Liu, M.A. Gorovsky, Mapping the 5′ and 3′ ends of *Tetrahymena thermophila* mRNAs using RNA ligase mediated amplification of cDNA ends (RLM–RACE), Nucleic Acids Res. 21 (1993) 4954–4960.

[2] J.B. Edwards, J. Delort, J. Mallet, Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5′ ends of mRNAs and for constructing cDNA libraries by in vitro amplification, Nucleic Acids Res. 19 (1991) 5227–5232.

[3] E. Bertrand, M. Fromont-Racine, R. Pictet, T. Grange, Visualization of the interaction of a regulatory protein with RNA in vivo, Proc. Natl. Acad. Sci. USA 90 (1993) 3496–3500.

[4] H.H. Chen, D. Castanotto, J.M. LeBon, J.J. Rossi, A.D. Riggs, In vivo, high-resolution analysis of yeast and mammalian RNA–protein interactions, RNA structure, RNA splicing, and ribozyme cleavage by use of terminal transferase-dependent PCR, Nucleic Acids Res. 28 (2000) 1656–1664.

[5] D.C. Tessier, R. Brousseau, T. Vernet, Ligation of single-stranded oligodeoxynucleotides by T4 RNA ligase, Anal. Biochem. 158 (1986) 171–178.

[6] W.M. Schmidt, M.W. Mueller, Controlled ribonucleotide tailing of cDNA ends (CRTC) by terminal deoxynucleotidyl transferase: a new approach in PCR-mediated analysis of mRNA sequences, Nucleic Acids Res. 24 (1996) 1789–1791.

[7] T. Blondal, A. Thorisdottir, U. Unnsteinsdottir, S. Hjorleifsdottir, A. Aevarsson, S. Ernstsson, O.H. Fridjonsson, S. Skirnisdottir, J.O. Wheat, A.G. Hermannsdottir, S.T. Sigurdsson, G.O. Hreggvidsson, A.V. Smith, J.K. Kristjansson, Isolation and characterization of a thermostable RNA ligase 1 from a *Thermus scotoductus* bacteriophage TS2126 with good single-stranded DNA ligation properties, Nucleic Acids Res. 33 (2005) 135–142.

[8] Z. Huang, J.W. Szostak, A simple method for 3′-labelling of RNA, Nucleic Acids Res. 24 (1996) 4360–4361.

[9] E.J. Merino, K.A. Wilkinson, J.L. Coughlan, K.M. Weeks, RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE), J. Am. Chem. Soc. 127 (2005) 4223–4231.

[10] O.C. Uhlenbeck, R.J. Gumport, RNA ligase, Enzymes 15 (1982) 31–58.

[11] K.A. Wilkinson, E.J. Merino, K.M. Weeks, RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in $tRNA^{Asp}$ transcripts, J. Am. Chem. Soc. 127 (2005) 4659–4667.

[12] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, D.H. Turner, Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, Proc. Natl. Acad. Sci. USA 101 (2004) 7287–7292.

ATTACHMENT 8A

LDR	01857cas a2200565 4500
001	991179463406676
005	20250227140528.0
007	cr|cn ||||||||
008	730925c19609999cau x p 0 0eng
022	__ | **a**0003-2697**l**0003-2697**9**PY**7**(Print)
022	__ | **a**1096-0309**9**EN**7**(Electronic)
030	__ | **a**ANBCA2
035	__ | **9**0370535
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
044	__ | **9**United States
060	00 | **a**W1**b**AN1915
210	10 | **a**Anal Biochem**2**DNLM
245	00 | **a**Analytical biochemistry.
246	2_ | **0**ANAL BIOCHEM**9**n
246	2_ | **0**ANALYT BIOCHEM**9**n
260	__ | **a**Orlando Fl :**b**Academic Press
264	31 | **3**<2000- > :**a**San Diego, CA :**b**Elsevier
300	__ | **a**v.**b**illus.
310	__ | **a**Eighteen no. a year
321	__ | **a**Sixteen no. a year**b**19??-1994
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**v. 1- June 1960-
510	1_ | **9**1**a**MEDLINE**b**v11n2,May 1965-
510	0_ | **9**1**a**PubMed**b**v11n2,May 1965-
510	0_ | **a**OLDMEDLINE
510	1_ | **a**Index medicus
530	__ | **a**Also issued online.
533	__ | **a**Microfilm.**m**v.22(1968)-25(1968).**b**[Bethesda, Md.] :**c**National Library of Medicine,**d**1971.**e**2 microfilm reels : negative ; 35 mm.
583	__ | **3**Vol. 184, no. 1 (JAN 1990)**c**19901002**f**Notice published**l**Printed on acid-free paper
650	_2 | **a**Biochemistry
655	_2 | **a**Periodical
856	4_ | **u**http://www.sciencedirect.com/science/journal/00032697
911	__ | **a**Electronic and print, print canceled 2018
914	__ | **a**Portico**b**Current to 2025 as of 20250227**c**ISSN_00032697
935	__ | **a**(DNLM)117946-nlmdb
995	__ | **a**AUTH**b**19730925**c**REV**d**20250227

996	__ | **a**NOTCNSR**b**20000101
997	__ | **a**OMR
998	__ | **a**ERMSLC 108649**b**20061129
999	__ | **a**AUTH
035	__ | **a**(DNLM)A29850000(s)
035	__ | **a**(OCoLC)01481077

ATTACHMENT 8B

Biochemistry MeSH Descriptor Data 2025

MeSH Heading
Biochemistry
Tree Number(s)
H01.158.201
H01.181.122
Unique ID
D001671
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D001671
Annotation
the discipline (education, history, etc) only; use BIOCHEMICAL PHENOMENA or specifics treed under it for chemical processes and reactions of biological systems
Scope Note
The study of the composition, chemical structures, and chemical reactions of living things.
Entry Version
BIOCHEM
NLM Classification #
QU 21
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2009/07/06

Detection and quantification of rare mutations with massively parallel sequencing

Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W. Kinzler[1], and Bert Vogelstein[1]

The Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute, Johns Hopkins Kimmel Cancer Center, Baltimore, MD 21231

Contributed by Bert Vogelstein, April 19, 2011 (sent for review March 21, 2011)

The identification of mutations that are present in a small fraction of DNA templates is essential for progress in several areas of biomedical research. Although massively parallel sequencing instruments are in principle well suited to this task, the error rates in such instruments are generally too high to allow confident identification of rare variants. We here describe an approach that can substantially increase the sensitivity of massively parallel sequencing instruments for this purpose. The keys to this approach, called the Safe-Sequencing System ("Safe-SeqS"), are (*i*) assignment of a unique identifier (UID) to each template molecule, (*ii*) amplification of each uniquely tagged template molecule to create UID families, and (*iii*) redundant sequencing of the amplification products. PCR fragments with the same UID are considered mutant ("supermutants") only if ≥95% of them contain the identical mutation. We illustrate the utility of this approach for determining the fidelity of a polymerase, the accuracy of oligonucleotides synthesized in vitro, and the prevalence of mutations in the nuclear and mitochondrial genomes of normal cells.

diagnostics | early diagnosis | biomarkers | genetics | cancer

Genetic mutations underlie many aspects of life and death—through evolution and disease, respectively. Accordingly, their measurement is critical to several fields of research. Luria and Delbrück's classic fluctuation analysis is a prototypic example of the insights into biological processes that can be gained simply by counting the number of mutations in carefully controlled experiments (1). Counting de novo mutations in humans, not present in their parents, has similarly led to new insights into the rate at which our species can evolve (2, 3). Similarly, counting genetic or epigenetic changes in tumors can inform fundamental issues in cancer biology (4). Mutations lie at the core of current problems in managing patients with viral diseases such as AIDS and hepatitis by virtue of the drug resistance they can cause (5, 6). Detection of such mutations, particularly at a stage before their becoming dominant in the population, will likely be essential to optimize therapy. Detection of donor DNA in the blood of organ transplant patients is an important indicator of graft rejection and detection of fetal DNA in maternal plasma can be used for prenatal diagnosis in a noninvasive fashion (7, 8). In neoplastic diseases, which are all driven by somatic mutations, the applications of rare mutant detection are manifold; they can be used to help identify residual disease at surgical margins or in lymph nodes, to follow the course of therapy when assessed in plasma, and to identify patients with early, surgically curable disease when evaluated in stool, sputum, plasma, and other bodily fluids (9–11).

These examples highlight the importance of identifying rare mutations for both basic and clinical research. Accordingly, innovative ways to assess them have been devised over the years. The first methods involved biologic assays based on prototrophy, resistance to viral infection or drugs, or biochemical assays (1, 12–18). Molecular cloning and sequencing provided a new dimension to the field, as they allowed the type of mutation, rather than simply its presence, to be identified (19–24). Some of the most powerful of these newer methods are based on digital PCR, in which individual molecules are assessed one by one (25). Digital PCR is conceptually identical to the analysis of individual clones of bacteria, cells, or virus, but is performed entirely in vitro with defined, inanimate reagents. Several implementations of digital PCR have been described, including the analysis of molecules arrayed in multiwell plates, in polonies, in microfluidic devices, and in water-in-oil emulsions (25–30). In each of these technologies, mutant templates are identified through their binding to oligonucleotides specific for the potentially mutant base.

Massively parallel sequencing represents a particularly powerful form of digital PCR in that hundreds of millions of template molecules can be analyzed one by one. It has the advantage over conventional digital PCR methods in that multiple bases can be queried sequentially and easily in an automated fashion. However, massively parallel sequencing cannot generally be used to detect rare variants because of the high error rate associated with the sequencing process. For example, with the commonly used Illumina sequencing instruments, this error rate varies from ~1% (31, 32) to ~0.05% (33, 34), depending on factors such as the read length (35), use of improved base-calling algorithms (36–38), and the type of variants detected (39). Some of these errors presumably result from mutations introduced during template preparation, during the preamplification steps required for library preparation, and during further solid-phase amplification on the instrument itself. Other errors are due to base misincorporation during sequencing and base-calling errors. Advances in base calling can enhance confidence (e.g., refs. 36–39), but instrument-based errors are still limiting, particularly in clinical samples wherein the mutation prevalence can be ≤0.01% (11). In the work described herein, we show how templates can be prepared and the sequencing data obtained from them more reliably interpreted, so that relatively rare mutations can be identified with commercially available instruments.

Results

Overview. Our approach, called the Safe-Sequencing System ("Safe-SeqS"), involves two basic steps (Fig. 1). The first is the assignment of a unique identifier (UID) to each DNA template molecule to be analyzed. The second is the amplification of each uniquely tagged template, so that many daughter molecules with the identical sequence are generated (defined as a UID family). If a mutation preexisted in the template molecule used for amplification, that mutation should be present in every daughter molecule containing that UID (barring any subsequent replication or sequencing errors). A UID family in which at least 95% of family members have the identical mutation is called a "supermutant". Mutations not occurring in the original templates, such as those occurring during the amplification steps or through errors in base calling, should not give rise to supermutants. Conceptual and

Author contributions: I.K., N.P., K.W.K., and B.V. designed research; I.K., J.W., N.P., and B.V. performed research; I.K., J.W., N.P., K.W.K., and B.V. contributed new reagents/analytic tools; I.K., N.P., K.W.K., and B.V. analyzed data; and I.K. and B.V. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed: E-mail: kinzlke@jhmi.edu or bertvog@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1105422108/-/DCSupplemental.

EX1048

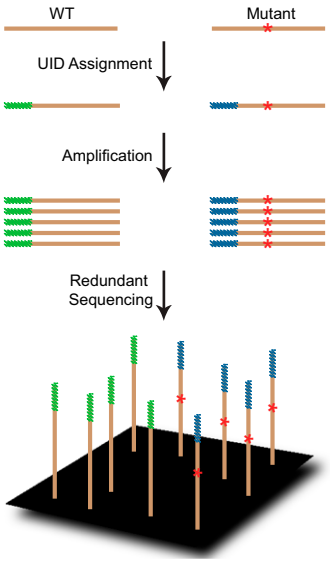


Fig. 1. Essential elements of Safe-SeqS. In the first step, each fragment to be analyzed is assigned a unique identification (UID) DNA sequence (green or blue bars). In the second step, the uniquely tagged fragments are amplified, producing UID families, each member of which has the same UID. A supermutant is defined as a UID family in which ≥95% of family members have the same mutation.

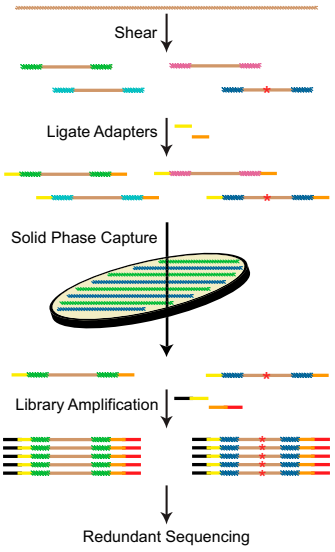


Fig. 2. Safe-SeqS with endogenous UIDs plus capture. The sequences of the ends of each fragment produced by random shearing (variously colored bars) serve as the unique identifiers (UIDs). These fragments are ligated to adapters (yellow and orange bars) so they can subsequently be amplified by PCR. One uniquely identifiable fragment is produced from each strand of the double-stranded template; only one strand is shown. Fragments of interest are captured on a solid phase containing oligonucleotides complementary to the sequences of interest. Following PCR amplification to produce UID families with primers containing 5′ "grafting" sequences (black and red bars), sequencing is performed and supermutants are defined as in Fig. 1.

practical issues related to UID assignment and supermutants are discussed in detail in *SI Materials and Methods*.

Endogenous UIDs. UIDs, sometimes called barcodes or indexes, can be assigned to nucleic acid fragments using a variety of methods. These methods include the introduction of exogenous sequences through PCR (40, 41) or ligation (42, 43). Even more simply, randomly sheared genomic DNA inherently contains UIDs consisting of the sequences of the two ends of each sheared fragment (Fig. 2 and Fig. S1). Paired-end sequencing of these fragments yields UID families that can be analyzed as described above. To use such endogenous UIDs in Safe-SeqS, we used two separate approaches: one designed to evaluate many genes simultaneously and the other designed to evaluate a single gene fragment in depth (Fig. 2 and Fig. S1, respectively).

For the evaluation of multiple genes, we ligated standard Illumina sequencing adapters to the ends of sheared DNA fragments to produce a standard sequencing library and then captured genes of interest on a solid phase (44). In this experiment, a library made from the DNA of ~15,000 normal cells was used, and 2,594 bp from six genes were targeted for capture. After excluding known single-nucleotide polymorphisms, 25,563 apparent mutations, corresponding to 2.4×10^{-4} mutations/bp, were also identified (Table 1). On the basis of previous analyses of mutation rates in human cells, at least 90% of these apparent mutations were likely to represent mutations introduced during template and library preparation or base-calling errors. Note that the error rate determined here (2.4×10^{-4} mutations/bp) is considerably lower than usually reported in experiments using the Illumina instrument because we used very stringent criteria for base calling (*SI Materials and Methods*).

With Safe-SeqS analysis of the same data, we determined that 69,505 original template molecules were assessed in this experiment (i.e., 69,505 UID families, with an average of 40 members per family, were identified) (Table 1). All of the polymorphic variants identified by conventional analysis were also identified by Safe-SeqS. However, only eight supermutants were observed among these families, corresponding to 3.5×10^{-6} mutations/bp. Thus, Safe-SeqS decreased the presumptive sequencing errors by at least 70-fold.

A strategy using endogenous UIDs was also used to reduce false-positive mutations upon deep sequencing of a single region of interest. In this case, a library prepared as described above from ~1,750 normal cells was used as template for inverse PCR using primers complementary to a gene of interest, so the PCR products could be directly used for sequencing (Fig. S1). With conventional analysis, an average of 2.3×10^{-4} mutations/bp were observed, similar to that observed in the capture experiment (Table 1). Given that only 1,057 independent molecules from normal cells were assessed in this experiment, as determined through Safe-SeqS analysis, all mutations observed with conventional analysis likely represented false positives (Table 1). With Safe-SeqS analysis of the same data, no supermutants were identified at any position.

Table 1. Safe-SeqS with endogenous UIDs

	Capture	Inverse PCR
Conventional analysis		
High-quality base pairs	106,958,863	1,041,346,645
Mean high-quality base pairs read depth	38,620×	2,085,600×
Mutations identified	25,563	234,352
Mutations/bp	2.4*E*-04	2.3*E*-04
Safe-SeqS analysis		
High-quality base pairs	106,958,863	1,041,346,645
Mean high-quality base pairs read depth	38,620×	2,085,600×
UID families	69,505	1,057
Average no. of members/UID family	40	21,688
Median no. of members/UID family	19	4
Supermutants identified	8	0
Supermutants/bp	3.5*E*-06	0.0

Exogenous UIDs. Although the results described above show that Safe-SeqS can increase the reliability of massively parallel sequencing, the number of different molecules that can be examined using endogenous UIDs is limited. For fragments sheared to an average size of 150 bp (range 125–175), 36-base paired-end sequencing can evaluate a maximum of ~7,200 different molecules containing a specific mutation (2 reads × 2 orientations × 36 bases/read × 50-base variation on either end of the fragment). In practice, the actual number of UIDs is smaller because the shearing process is not entirely random.

To make more efficient use of the original templates, we developed a Safe-SeqS strategy that used a minimum number of enzymatic steps. This strategy also permitted the use of degraded or damaged DNA, such as found in clinical specimens or after bisulfite treatment for the examination of cytosine methylation (45). As depicted in Fig. 3, this strategy employs two sets of PCR primers. The first set is synthesized with standard phosphoramidite precursors and contained sequences complementary to the gene of interest on the 3′ end and different tails at the 5′ ends of both the forward and reverse primers. The different tails allowed universal amplification in the next step. Finally, there was a stretch of 12–14 random nucleotides between the tail and the sequence-specific nucleotides in the forward primer (40). The random nucleotides form the UIDs. An equivalent way to assign UIDs to fragments, not used in this study, would employ 10,000 forward primers and 10,000 reverse primers synthesized on a microarray. Each of these 20,000 primers would have gene-specific primers at their 3′ ends and one of 10,000 specific, predetermined, nonoverlapping UID sequences at their 5′ ends, allowing for 10^8 [i.e., $(10^4)^2$] possible UID combinations. In either case, two cycles of PCR are performed with the primers and a high-fidelity polymerase, producing a uniquely tagged, double-stranded DNA fragment from each of the two strands of each original template molecule (Fig. 3). The residual, unused UID assignment primers are removed by digestion with a single strand-specific exonuclease, without further purification, and two new primers are added. The new primers, complementary to the tails introduced in the UID assignment cycles, contain grafting sequences at their 5′ ends, permitting solid-phase amplification on the Illumina instrument, and phosphorothioate residues at their 3′ ends to make them resistant to any remaining exonuclease. Following 25 additional cycles of PCR, the products are loaded on the Illumina instrument. As shown below, this strategy allowed us to evaluate the majority of input fragments and was used for several illustrative experiments.

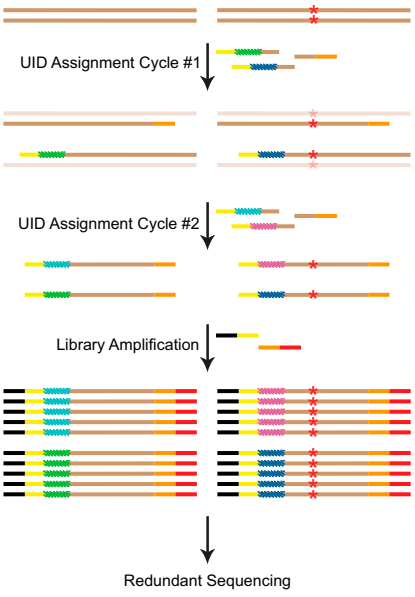


Fig. 3. Safe-SeqS with exogenous UIDs. DNA (sheared or unsheared) is amplified with a set of gene-specific primers. One of the primers has a random DNA sequence (e.g., a set of 14 Ns) that forms the unique identifier (UID) (variously colored bars), located 5′ to its gene-specific sequence, and both have sequences that permit universal amplification in the next step (yellow and orange bars). Two UID assignment cycles produce two fragments—each with a different UID—from each double-stranded template molecule, as shown. Subsequent PCR with universal primers, which also contain "grafting" sequences (black and red bars), produces UID families that are directly sequenced. Supermutants are defined as in the legend to Fig. 1.

Analysis of DNA Polymerase Fidelity. Measurement of the error rates of DNA polymerases is essential for their characterization and dictates the situations in which these enzymes can be used. We chose to measure the error rate of Phusion polymerase, as this polymerase has one of the lowest reported error frequencies of any commercially available enzyme and therefore poses a particular challenge for an in vitro-based approach. We first amplified a single human DNA template molecule, comprising a segment of an arbitrarily chosen human gene, through 19 rounds of PCR. The PCR products from these amplifications, in their entirety, were used as templates for Safe-SeqS as described in Fig. 3. In seven independent experiments of this type, the number of UID families identified by sequencing was 624,678 ± 421,274, which is consistent with an amplification efficiency of 92 ± 9.6% per round of PCR.

The error rate of Phusion polymerase, estimated through cloning of PCR products encoding β-galactosidase in plasmid vectors and transformation into bacteria, is reported by the manufacturer to be 4.4×10^{-7}errors/bp/PCR cycle. Even with very high-stringency base calling, conventional analysis of the Illumina sequencing data revealed an apparent error rate of 9.1×10^{-6} errors/bp/PCR cycle, more than an order of magnitude higher than the reported Phusion polymerase error rate (Table 2, polymerase fidelity). In contrast, Safe-SeqS of the same data revealed an error rate of 4.5×10^{-7}errors/bp/PCR cycle, nearly identical to that measured for Phusion polymerase in biological assays (Table 2, polymerase fidelity). The vast majority (>99%) of these errors were single-base substitutions (Table S1, polymerase fidelity), consistent with previous data on the mutation spectra created by other prokaryotic DNA polymerases (15, 46, 47).

Safe-SeqS also allowed a determination of the total number of distinct mutational events and an estimation of PCR cycle in which the mutation occurred. There were 19 cycles of PCR performed in wells containing a single template molecule in these experiments. If a polymerase error occurred in cycle 19, there would be only one supermutant produced (from the strand containing the mutation). If the error occurred in cycle 18, there should be two supermutants (derived from the mutant strands produced in cycle 19), etc. Accordingly, the cycle in which the error occurred is related to the number of supermutants containing that error. The data from seven independent experiments demonstrate a relatively consistent number of observed total polymerase errors ($2.2 \pm 1.1 \times 10^{-6}$ distinct mutations/bp), in reasonable agreement with the number expected from simulations ($1.5 \pm 0.21 \times 10^{-6}$ distinct mutations/bp, detailed in *SI Materials and Methods*). The data also show a highly variable timing of occurrence of polymerase errors among experiments (Table S2), as predicted from classic fluctuation analysis (1). This kind of information is difficult to derive using conventional analysis of the same next-generation sequencing data, in part because of the prohibitively high apparent mutation rate noted above.

Analysis of Oligonucleotide Composition. A small number of mistakes during the synthesis of oligonucleotides from phoshoramidite precursors are tolerable for most applications, such as routine PCR or cloning. However, for synthetic biology, wherein many oligonucleotides must be joined together, such mistakes present a major

Table 2. Safe-SeqS with exogenous UIDs

	Mean	SD
Polymerase fidelity		
Conventional analysis of seven replicates		
High-quality base pairs	996,855,791	64,030,757
Total mutations identified	198,638	22,515
Mutations/bp	2.0*E*-04	1.7*E*-05
Calculated Phusion error rate (errors/bp/cycle)	9.1*E*-06	7.7*E*-07
Safe-SeqS analysis of seven replicates		
High-quality base pairs	996,855,791	64,030,757
UID families	624,678	421,274
Members/UID family	107	122
Total supermutants identified	197	143
Supermutants/bp	9.9*E*-06	2.3*E*-06
Calculated Phusion error rate (errors/bp/cycle)	4.5*E*-07	1.0*E*-07
CTNNB1 mutations in DNA from normal human cells		
Conventional analysis of three individuals		
High-quality base pairs	559,334,774	66,600,749
Total mutations identified	118,488	11,357
Mutations/bp	2.1*E*-04	1.6*E*-05
Safe-SeqS analysis of three individuals		
High-quality base pairs	559,334,774	66,600,749
UID families	374,553	263,105
Members/UID family	68	38
Total supermutants identified	99	78
Supermutants/bp	9.0*E*-06	3.1*E*-06
Mitochondrial mutations in DNA from normal human cells		
Conventional analysis of seven individuals		
High-quality base pairs	147,673,456	54,308,546
Total mutations identified	30,599	12,970
Mutations/bp	2.1*E*-04	9.4*E*-05
Safe-SeqS analysis of seven individuals		
High-quality base pairs	147,673,456	54,308,546
UID families	515,600	89,985
Members/UID family	15	6
Total supermutants identified	135	61
Supermutants/bp	1.4*E*-05	6.8*E*-06

obstacle to success. Clever strategies for making the gene construction process more efficient have been devised (48, 49), but all such strategies would benefit from more accurate synthesis of the oligonucleotides themselves. Determining the number of errors in synthesized oligonucleotides is difficult because the fraction of oligonucleotides containing errors can be lower than the sensitivity of conventional next-generation sequencing analyses.

To determine whether Safe-SeqS could be used for this determination, we used standard phosphoramidite chemistry to synthesize an oligonucleotide containing 31 bases that were designed to be identical to that analyzed in the polymerase fidelity experiment described above. In the synthetic oligonucleotide, the 31 bases were surrounded by sequences complementary to primers that could be used for the UID assignment steps of Safe-SeqS (Fig. 3). By performing Safe-SeqS on ~300,000 oligonucleotide templates, we found that there were $8.9 \pm 0.28 \times 10^{-4}$ supermutants/bp and that these errors occurred throughout the sequence of the oligonucleotides (Fig. S2*A*). The oligonucleotides contained a large number of insertion and deletion errors, representing $8.2 \pm 0.63\%$ and $25 \pm 1.5\%$ of the total supermutants, respectively. Importantly, both the position and the nature of the errors were highly reproducible among seven independent replicates of this experiment performed on the same batch of oligonucleotides (Fig. S2*A*). This nature and distribution of errors had little in common with that of the errors produced by Phusion polymerase (Fig. S2*B* and Table S3), which were distributed in the expected stochastic pattern among replicate experiments. The number of errors in the oligonucleotides synthesized with phosphoramidites was ~60 times higher than that in the equivalent products synthesized by Phusion polymerase. These data, in toto, indicate that the vast majority of errors in the former were generated during their synthesis rather than during the Safe-SeqS procedure.

Does Safe-SeqS preserve the ratio of mutant:normal sequences in the original templates? To address this question, we synthesized two 31-base oligonucleotides of identical sequence with the exception of nucleotide 15 (50:50 C/G instead of T) and mixed them at nominal mutant/normal fractions of 3.3% and 0.33%. Through Safe-SeqS analysis of the oligonucleotide mixtures, we found that the ratios were 2.8% and 0.27%, respectively. We conclude that the UID assignment and amplification procedures used in Safe-SeqS do not greatly alter the proportion of variant sequences and thereby provide a reliable estimate of that proportion when unknown. This conclusion is also supported by the reproducibility of variant fractions when analyzed in independent Safe-SeqS experiments (Fig. S2*A*).

Analysis of DNA Sequences from Normal Human Cells. The exogenous UID strategy (Fig. 3) was then used to determine the prevalence of rare mutations in a small region of the *CTNNB1* gene isolated from ~100,000 normal human cells from three unrelated individuals. Through comparison with the number of UID families obtained in the Safe-SeqS experiments (Table 2, *CTNNB1* mutations in DNA from normal human cells), we calculated that the majority ($78 \pm 9.8\%$) of the input fragments were converted into UID families. There was an average of 68 members/UID family, easily fulfilling the required redundancy for Safe-SeqS (Fig. S3). Conventional analysis of the Illumina sequencing data revealed an average of $118{,}488 \pm 11{,}357$ mutations among the ~560 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.16 \times 10^{-4}$ mutations/bp (Table 2, *CTNNB1* mutations in DNA from normal human cells). Only an average of 99 ± 78 supermutants were observed in the Safe-SeqS analysis. The vast majority (>99%) of supermutants were single-base substitutions and the calculated mutation rate was $9.0 \pm 3.1 \times 10^{-6}$ mutations/bp (Table S1, *CTNNB1* mutations in DNA from normal human cells). Safe-SeqS thereby reduced the apparent frequency of mutations in genomic DNA by at least 24-fold (Fig. 4).

We applied the identical strategy to a short segment of mitochondrial DNA isolated from ~1,000 cells from each of seven unrelated individuals. Conventional analysis of the Illumina sequencing libraries produced with the Safe-SeqS procedure (Fig. 3) revealed an average of $30{,}599 \pm 12{,}970$ mutations among the ~150 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.94 \times 10^{-4}$ mutations/bp (Table 2, mitochondrial mutations in DNA from normal human cells). Only 135 ± 61 supermutants were observed in the Safe-SeqS analysis. As with the *CTNNB1* gene, the vast majority of mutations were single-base substitutions, although occasional single-base deletions were also observed (Table S1, mitochondrial mutations in DNA from normal human cells). The calculated mutation rate in the analyzed segment of mtDNA was $1.4 \pm 0.68 \times 10^{-5}$ mutations/bp (Table 2, mitochondrial mutations in DNA from normal human cells). Thus, Safe-SeqS thereby reduced the apparent frequency of mutations in mitochondrial DNA by at least 15-fold.

Discussion

The results described above demonstrate that the Safe-SeqS approach can substantially improve the accuracy of massively parallel sequencing (Tables 1 and 2). It can be implemented through either endogenous or exogenously introduced UIDs and can be applied to virtually any sample preparation workflow or sequencing platform. As demonstrated here, the approach can easily be used to identify rare mutants in a population of DNA templates, to measure poly-

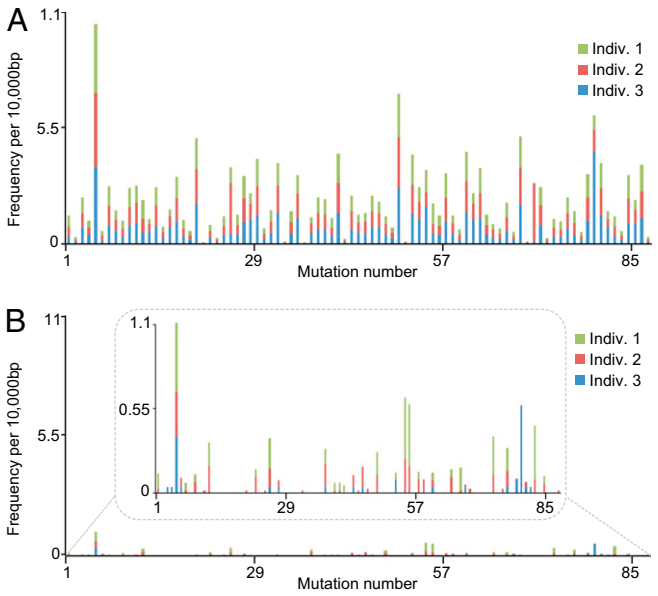


Fig. 4. Single-base substitutions identified by conventional and Safe-SeqS analysis. The exogenous UID strategy depicted in Fig. 3 was used to produce PCR fragments from the *CTNNB1* gene of three normal, unrelated individuals. Mutation numbers represent one of 87 possible single-base substitutions (3 possible substitutions/base × 29 bases analyzed). These fragments were sequenced on an Illumina GA IIx instrument and analyzed in the conventional manner (*A*) or with Safe-SeqS (*B*). Safe-SeqS results are displayed on the same scale as conventional analysis for direct comparison; the *Inset* is a magnified view. Note that most of the variants identified by conventional analysis are likely to represent sequencing errors, as indicated by their high frequency relative to Safe-SeqS and their consistency among unrelated samples.

merase error rates, and to judge the reliability of oligonucleotide syntheses. One of the advantages of the strategy is that it yields the number of templates analyzed as well as the fraction of templates containing variant bases. Previously described in vitro methods for the detection of small numbers of template molecules (e.g., refs. 29 and 50) allow the fraction of mutant templates to be determined but cannot determine the number of mutant and normal templates in the original sample.

It is of interest to compare Safe-SeqS to other approaches for reducing errors in next-generation sequencing. As mentioned in the Introduction, sophisticated algorithms to increase the accuracy of base calling have been developed (e.g., refs. 36–39). These improved base calling algorithms can certainly reduce false-positive calls, but their effectiveness is still limited by artifactual mutations occurring during the PCR steps required for library preparation as well as by any residual base-calling errors. For example, the algorithm used in the current study used very stringent criteria for base calling and was applied to short read lengths, but was still unable to reduce the error rate to less than an average of 2.0×10^{-4} errors/bp. This error frequency is at least as low as those reported with other algorithms. To improve sensitivity further, these base-calling improvements can be used together with Safe-SeqS. Travers et al. describe another powerful strategy for reducing errors (51). With this technology, both strands of each template molecule are sequenced redundantly after a number of preparative enzymatic steps. However, this approach can be performed only on a specific instrument. Moreover, for many clinical applications, there are relatively few template molecules in the initial sample and evaluation of nearly all of them is required to obtain the requisite sensitivity. The approach described here with exogenously introduced UIDs (Fig. 3) fulfills this requirement by coupling the UID assignment step with a subsequent amplification in which few molecules are lost. Our endogenous UID approaches (Fig. 2 and Fig. S1) and the one described by Travers et al. are not ideally suited for this purpose because of the inevitable losses of template molecules during the ligation and other preparative steps.

How do we know that the mutations identified by conventional analyses in the current study represent artifacts rather than true mutations in the original templates? Strong evidence supporting this is provided by the observation that the mutation prevalence in all but one experiment was similar: 2.0×10^{-4}–2.4×10^{-4} mutations/bp (Tables 1 and 2). The exception was the experiment with oligonucleotides synthesized from phosphoramidites, in which the error of the synthetic process was apparently higher than the error rate of conventional Illumina analysis when used with stringent base-calling criteria. In contrast, the mutation prevalence of Safe-SeqS varied much more, from 0.0 to 1.4×10^{-5} mutations/bp, depending on the template and experiment. Moreover, the mutation prevalence measured by Safe-SeqS in the most controlled experiment, in which polymerase fidelity was measured (Table 2, polymerase fidelity), was almost identical to that predicted from previous experiments in which polymerase fidelity was measured by biological assays. Our measurements of mutation prevalence in the DNA from normal cells are consistent with some previous experimental data. However, estimates of these prevalences vary widely and may depend on cell type and sequence analyzed (*SI Materials and Methods*). We therefore cannot be certain that the relatively low number of mutations revealed by Safe-SeqS represented errors occurring during the sequencing process rather than true mutations present in the original DNA templates. Potential sources of error in the Safe-SeqS process are described in *SI Materials and Methods*.

Like all techniques, Safe-SeqS has limitations. For example, we have demonstrated that the exogenous UIDs strategy can be used to analyze a single amplicon in depth. This technology may not be applicable to situations wherein multiple amplicons must be analyzed from a sample containing a limited number of templates. Multiplexing in the UID assignment cycles (Fig. 3) may provide a solution to this challenge. A second limitation is that the efficiency of amplification in the UID assignment cycles is critical for the success of the method. Clinical samples can contain inhibitors that reduce the efficiency of this step. This problem can presumably be overcome by performing more than two cycles in the UID assignment PCR step (Fig. 3), although this would complicate the determination of the number of templates analyzed. The specificity of Safe-SeqS is currently limited by the fidelity of the polymerase used in the UID assignment PCR step, i.e., 8.8×10^{-7} mutations/bp in its current implementation with two cycles. Increasing the number of cycles in the UID assignment PCR step to five would decrease the overall specificity to $\sim 2 \times 10^{-6}$ mutations/bp. However, this specificity can be increased by requiring more than one supermutant for mutation identification—the probability of introducing the same artifactual mutation twice or three times would be exceedingly low [$(2 \times 10^{-6})^2$ or $(2 \times 10^{-6})^3$, respectively]. In sum, there are several simple ways to vary the Safe-SeqS procedure and analysis to realize the needs of specific experiments.

Luria and Delbrück, in their classic paper in 1943, wrote that their "prediction cannot be verified directly, because what we observe, when we count the number of resistant bacteria in a culture, is not the number of mutations which have occurred but the number of resistant bacteria which have arisen by multiplication of those which mutated, the amount of multiplication depending on how far back the mutation occurred" (ref. 1, p. 495). The Safe-SeqS procedure described here can verify such predictions because the number as well as the time of occurrence of each mutation can be estimated from the data, as noted in the experiments on polymerase fidelity. In addition to templates generated by polymerases in vitro, the same approach can be applied to DNA from bacteria, viruses, and mammalian cells. We therefore expect that this

strategy will provide definitive answers to a variety of important biomedical questions.

Materials and Methods

Endogenous UIDs. To expose endogenous UIDs, DNA was fragmented to an average size of ~200 bp by acoustic shearing (Covaris) and then end-repaired, A-tailed, and ligated to Y-shaped adapters according to standard Illumina protocols. DNA was captured (44) with a filter containing 2,594 nt corresponding to six cancer genes. For the inverse PCR experiments (Fig. S1), we ligated custom adapters (IDT) (Table S4) instead of standard Y-shaped Illumina adapters to sheared cellular DNA. Inverse PCR was performed using *KRAS* forward and reverse primers (Table S4) that both contained grafting sequences for hybridization to the Illumina GA IIx flow cell (Table S4). Further details are provided in *SI Materials and Methods*.

Exogenous UIDs. Each strand of each template molecule was encoded with a 12- or 14-base UID using two cycles of amplicon-specific PCR, as described in the text and Fig. 3. The amplicon-specific primers both contained universal tag sequences at their 5′ ends for a later amplification step. The UIDs constituted 12 or 14 random nucleotide sequences appended to the 5′ end of the forward amplicon-specific primers (Table S4). Following two cycles of PCR for UID assignment, the products were digested with a single-strand DNA-specific nuclease. Primers complementary to the introduced universal tags and containing 3′-terminal phosphorothioates (Table S4) were added and 25 additional cycles of PCR were performed. Further details are provided in *SI Materials and Methods*.

Sequencing. Sequencing of all of the libraries described above was performed using an Illumina GA IIx instrument as specified by the manufacturer. High-quality reads were grouped into UID families on the basis of their endogenous or exogenous UIDs. Only UID families with two or more members were considered, as described in detail in *SI Materials and Methods*.

ACKNOWLEDGMENTS. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research; The Virginia and D. K. Ludwig Fund for Cancer Research; The Sol Goldman Center for Pancreatic Cancer Research; The Joseph L. Rabinowitz Fund for Pancreatic Cancer Research; National Cancer Institute Division of Cancer Prevention Contract N01-CN-43302; and National Institutes of Health Grants CA62924, CA43460, and CA57345. I.K. is also supported by a United Negro College Fund–Merck Graduate Fellowship.

1. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.
2. Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 328:636–639.
3. Durbin RM, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
4. Shibata D (2011) Mutation and epigenetic molecular clocks in cancer. *Carcinogenesis* 32:123–128.
5. McMahon MA, et al. (2007) The HBV drug entecavir - effects on HIV-1 replication and resistance. *N Engl J Med* 356:2614–2621.
6. Eastman PS, et al. (1998) Maternal viral genotypic zidovudine resistance and infrequent failure of zidovudine therapy to prevent perinatal transmission of human immunodeficiency virus type 1 in pediatric AIDS Clinical Trials Group Protocol 076. *J Infect Dis* 177:557–564.
7. Chiu RW, et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci USA* 105:20458–20463.
8. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 105:16266–16271.
9. Hoque MO, et al. (2003) High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. *Cancer Res* 63:5723–5726.
10. Thunnissen FB (2003) Sputum examination for early detection of lung cancer. *J Clin Pathol* 56:805–810.
11. Diehl F, et al. (2008) Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. *Gastroenterology* 135:489–498.
12. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112:29–35.
13. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.
14. Campbell F, Appleton MA, Shields CJ, Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right- and left-sided sporadic colorectal carcinomas. *J Pathol* 186:31–35.
15. Tindall KR, Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry* 27:6008–6013.
16. Kunkel TA (1985) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* 260:5787–5796.
17. van Dongen JJ, Wolvers-Tettero IL (1991) Analysis of immunoglobulin and T cell receptor genes. Part II: Possibilities and limitations in the diagnosis and management of lymphoproliferative diseases and related disorders. *Clin Chim Acta* 198:93–174.
18. Grist SA, McCarron M, Kutlaca A, Turner DR, Morley AA (1992) In vivo human somatic mutation: Frequency and spectrum with age. *Mutat Res* 266:189–196.
19. Liu Q, Sommer SS (2004) Detection of extremely rare alleles by bidirectional pyrophosphorolysis-activated polymerization allele-specific amplification (Bi-PAP-A): Measurement of mutation load in mammalian tissues. *Biotechniques* 36:156–166.
20. Monnat RJ, Jr., Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. *Proc Natl Acad Sci USA* 82:2895–2899.
21. Shi C, et al. (2004) LigAmp for sensitive detection of single-nucleotide differences. *Nat Methods* 1:141–147.
22. Keohavong P, Thilly WG (1989) Fidelity of DNA polymerases in DNA amplification. *Proc Natl Acad Sci USA* 86:9253–9257.
23. Sidransky D, et al. (1991) Identification of p53 gene mutations in bladder cancers and urine samples. *Science* 252:706–709.
24. Bielas JH, Loeb LA (2005) Quantification of random genomic mutations. *Nat Methods* 2:285–290.
25. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
26. Mitra RD, et al. (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100:5926–5931.
27. Chetverina HV, Samatov TR, Ugarov VI, Chetverin AB (2002) Molecular colony diagnostics: Detection and quantitation of viral nucleic acids by in-gel PCR. *Biotechniques* 33:150–152, 154, 156.
28. Zimmermann BG, et al. (2008) Digital PCR: A powerful new tool for noninvasive prenatal diagnosis? *Prenat Diagn* 28:1087–1093.
29. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 100:8817–8822.
30. Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314:1464–1467.
31. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.
32. Nazarian R, et al. (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468:973–977.
33. He Y, et al. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464:610–614.
34. Gore A, et al. (2011) Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63–67.
35. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
36. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: A self-optimizing base caller for next-generation sequencing. *Nat Methods* 5:679–682.
37. Rougemont J, et al. (2008) Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9:431.
38. Druley TE, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 6:263–265.
39. Vallania FL, et al. (2010) High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res* 20:1711–1718.
40. McCloskey ML, Stöger R, Hansen RS, Laird CD (2007) Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* 45:761–767.
41. Parameswaran P, et al. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35:e130.
42. Craig DW, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* 5:887–893.
43. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* 32:e135.
44. Herman DS, et al. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6:507–510.
45. Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128:683–692.
46. de Boer JG, Ripley LS (1988) An in vitro assay for frameshift mutations: Hotspots for deletions of 1 bp by Klenow-fragment polymerase share a consensus DNA sequence. *Genetics* 118:181–191.
47. Eckert KA, Kunkel TA (1990) High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Res* 18:3739–3744.
48. Kosuri S, et al. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28:1295–1299.
49. Matzas M, et al. (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28:1291–1294.
50. Li J, et al. (2008) Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat Med* 14:579–584.
51. Eid J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

Supporting Information

Kinde et al. 10.1073/pnas.1105422108

SI Materials and Methods

Endogenous UIDs. Genomic DNA from human pancreas or cultured lymphoblastoid cells was prepared using Qiagen kits. The pancreas DNA was used for the capture experiment and the lymphoblastoid cells were used for the inverse PCR experiment. DNA was quantified by optical absorbance and with qPCR. DNA was fragmented to an average size of ~200 bp by acoustic shearing (Covaris) and then end-repaired, A-tailed, and ligated to Y-shaped adapters according to standard Illumina protocols. The ends of each template molecule provide endogenous UIDs corresponding to their chromosomal positions. After PCR-mediated amplification of the libraries with primer sequences within the adapters, DNA was captured (1) with a filter containing 2,594 nt corresponding to six cancer genes. After capture, 18 cycles of PCR were performed to ensure sufficient amounts of template for sequencing on an Illumina GA IIx instrument.

For the inverse PCR experiments (Fig. S1), we ligated custom adapters (IDT) (Table S4) instead of standard Y-shaped Illumina adapters to sheared cellular DNA. These adapters retained the region complementary to the universal sequencing primer but lacked the grafting sequences required for hybridization to the Illumina GA IIx flow cell. The ligated DNA was diluted into 96 wells and the DNA in each column of 8 wells was amplified with a unique forward primer containing one of 12 index sequences at its 5′ end plus a standard reverse primer (Table S4). Amplifications were performed using Phusion HotStart I (NEB) in 50-μL reactions containing 1× Phusion HF buffer, 0.5 mM dNTPs, 0.5 μM each forward and reverse primer (both 5′-phosphorylated), and 1 unit of Phusion polymerase. The following cycling conditions were used: 1 cycle of 98 °C for 30 s and 16 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s. All 96 reactions were pooled and then purified using a Qiagen MinElute PCR Purification kit (no. 28004) and a QIAquick Gel Extraction kit (no. 28704). To prepare the circular templates necessary for inverse PCR, DNA was diluted to ~1 ng/uL and ligated with T4 DNA Ligase (Enzymatics) for 30 min at room temperature in a 600-uL reaction containing 1× T4 DNA ligation buffer and 18,000 units of T4 DNA Ligase. The ligation reaction was purified using a Qiagen MinElute kit. Inverse PCR was performed using Phusion Hot Start I on 90 ng of circular template distributed in twelve 50-μL reactions, each containing 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 uM each of *KRAS* forward and reverse primers (Table S4), and 1 unit of Phusion polymerase. The *KRAS*-specific primers both contained grafting sequences for hybridization to the Illumina GA IIx flow cell (Table S4). The following cycling conditions were used: 1 cycle of 98 °C for 2 min and then 37 cycles of 98 °C for 10 s, 61 °C for 15 s, and 72 °C for 10 s. The final purification was performed with a NucleoSpin Extract II kit (Macherey-Nagel) and eluted in 20 μL NE buffer. The resulting DNA fragments contained UIDs composed of three sequences: 2 endogenous ones, represented by the two ends of the original sheared fragments, plus the exogenous sequence introduced during the indexing amplification. As 12 exogenous sequences were used, this increased the number of distinct UIDs by 12-fold over that obtained without exogenous UIDs. This number could easily be increased by using a greater number of indexing primers.

Exogenous UIDs. Genomic DNA from normal human colonic mucosae or blood lymphocytes was prepared using Qiagen kits. The DNA from colonic mucosae was used for the experiments on *CTNNB1* and mitochondrial DNA, whereas the lymphocyte DNA was used for the experiments on *CTNNB1* and on polymerase fidelity. DNA was quantified with digital PCR (2) using primers that amplified single-copy genes from human cells (analysis of polymerase fidelity in *Analysis of DNA Polymerase Fidelity* and *CTNNB1* in *Analysis of DNA Sequences from Normal Human Cells*), with qPCR (mitochondrial DNA), or by optical absorbance (oligonucleotides in *Analysis of Oligonucleotide Composition*). Each strand of each template molecule was encoded with a 12- or 14-base UID using two cycles of amplicon-specific PCR, as described in the text and Fig. 3. The amplicon-specific primers both contained universal tag sequences at their 5′ ends for a later amplification step. The UIDs constituted 12 or 14 random nucleotide sequences appended to the 5′ end of the forward amplicon-specific primers (Table S4). These primers can generate 16.8 and 268 million distinct UIDs, respectively. It is important that the number of distinct UIDs greatly exceeds the number of original template molecules to minimize the probability that two different original templates acquire the same UID. The UID assignment PCR cycles included Phusion Hot Start II (NEB) in a 45-μL reaction containing 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 μM each forward (containing 12–14 Ns) and reverse primers, and 2 units of Phusion polymerase. To keep the final template concentrations <1.5 ng/uL, multiple wells were used to create some libraries. The following cycling conditions were used: one cycle of 98 °C for 30 s and then two cycles of 98 °C for 10 s, 61 °C for 120 s, and 72 °C for 10 s. To ensure complete removal of the first-round primers, each well was digested with 60 units of a single-strand DNA-specific nuclease (Exonuclease-I; Enzymatics) at 37 °C for 1 h. After a 5-min heat inactivation at 98 °C, primers complementary to the introduced universal tags (Table S4) were added to a final concentration of 0.5 uM each. These primers contained two terminal phosphorothioates to make them resistant to any residual Exonuclease-I activity. They also contained 5′-grafting sequences necessary for hybridization to the Illumina GA IIx flow cell. Finally, they contained an index sequence between the grafting sequence and the universal tag sequence. This index sequence enables the PCR products from multiple different individuals to be simultaneously analyzed in the same flow cell compartment of the sequencer. The following cycling conditions were used for the subsequent 25 cycles of PCR: 98 °C for 10 s and 72 °C for 15 s. No intermediate purification steps were performed in an effort to reduce the losses of template molecules.

After the second round of amplification, wells were consolidated and purified using a Qiagen QIAquick PCR Purification kit (no. 28104) and eluted in 50 μL EB buffer (Qiagen). Fragments of the expected size were purified after agarose (mtDNA libraries) or polyacrylamide (all other libraries) gel electrophoresis. For agarose gel purification, the eight 6-μL aliquots were loaded into wells of a 2% Size Select gel (Invitrogen) and bands of the expected size were collected in EB buffer as specified by the manufacturer. For polyacrylamide gel purification, ten 5-μL aliquots were loaded into wells of a 10% TBE Polyacrylamide gel (Invitrogen). Gel slices containing the fragments of interest were excised, crushed, and eluted essentially as described (3).

Analysis of Phusion Polymerase Fidelity. Amplification of a fragment of human genomic DNA within the BMX (RefSeq accession NM_203281.2) gene was first performed using the PCR conditions described above. The template was diluted so that an average of one template molecule was present in every three wells of a 96-well PCR plate. Fifty-microliter PCR reactions were then performed in 1× Phusion HF buffer, 0.25 mM dNTPs, 0.5 μM each forward and reverse primers (Table S4), and 2 units of

Phusion polymerase. The cycling conditions were 1 cycle of 98 °C for 30 s and 19 cycles of 98 °C for 10 s, 61 °C for 120 s, and 72 °C for 10 s. The primers were removed by digestion with 60 units of Exonuclease-I at 37 °C for 1 h followed by a 5-min heat inactivation at 98 °C. No purification of the PCR product was performed, either before or after Exonuclease-I digestion. The entire contents of each well were then used as templates for the exogenous UIDs strategy described above.

Sequencing. Sequencing of all of the libraries described above was performed using an Illumina GA IIx instrument as specified by the manufacturer. The total length of the reads used for each experiment varied from 36 to 73 bases. Base calling and sequence alignment were performed with the Eland pipeline (Illumina). Only high-quality reads meeting the following criteria were used for subsequent analysis: (*i*) the first 25 bases passed the standard Illumina chastity filter; (*ii*) every base in the read had a quality score ≥20; and (*iii*) ≤3 mismatches to expected sequences. For the exogenous UID libraries, we additionally required the UIDs to have a quality score ≥30. We noticed a relatively high frequency of errors at the ends of the reads in the endogenous UID libraries prepared with the standard Illumina protocol, presumably introduced during shearing or end repair, so the first and last 3 bases of these tags were excluded from analysis.

Safe-SeqS Analysis. High-quality reads were grouped into UID families on the basis of their endogenous or exogenous UIDs. Only UID families with two or more members were considered. Such UID families included the vast majority (≥99%) of the sequencing reads. To ensure that the same data were used for both conventional and Safe-SeqS analysis, we also excluded UID families containing only one member from conventional analysis. Furthermore, we identified a base as mutant in conventional sequencing analysis only if the same variant was identified in at least two members of at least one UID family (i.e., two mutations) when comparing conventional analysis to that of Safe-SeqS with exogenous UIDs. For comparison with Safe-SeqS with endogenous UIDs, we required at least two members of each of two UID families (i.e., four mutations) to identify a position as mutant in conventional analysis. With either endogenous or exogenous UIDs, a supermutant was defined as a UID family in which ≥95% of members shared the identical mutation. Thus, UID families with <20 members had to be 100% identical at the mutant position, whereas a 5% combined replication and sequencing error rate was permitted in UID families with more members. To determine polymerase fidelity using Safe-SeqS, and to compare the results with previous analyses of Phusion polymerase fidelity, it is important to realize that the previous analyses would detect only mutations present in both strands of the PCR products (4). This would be equivalent to analyzing PCR products generated with one less cycle with Safe-SeqS, and the appropriate correction was made in Table 2 (polymerase fidelity). Unless otherwise specified, all values listed in the text and tables represent means and SDs.

Error-generating processes. Apparent mutations, defined as any base call that varies from the expected base at a defined position, can result from a variety of processes:

i) Mutations present in the template DNA. For templates derived from normal human cells, these include mutations that were present in the zygote, occurred later during embryonic and adult development, or were present in a contaminant inadvertently introduced into the sample. These mutations are expected to be present in both strands of the relevant templates. If the mutation occurred only in the last cell cycle of a cell whose DNA was used as template, the mutation would be present in only one strand of the template.

ii) Chemically modified bases present in the templates. It has been estimated that there are many thousands of oxidized bases present in every human cell (5). When such DNA is amplified by Phusion polymerase, an apparent mutation in one strand may result.

iii) Errors introduced during the shearing process required to generate small fragments for sequencing. Acoustic shearing generates short-lived, high temperatures that can damage DNA.

iv) Errors introduced during end repair of the sheared fragments. The source of these errors can be polymerase infidelity or through incorporation of chemically modified bases in the dNTPs used for polymerization.

v) Errors introduced by other enzymatic steps, particularly if the enzymes are impure and contaminated with nucleases, polymerases, or ligases.

vi) Errors introduced during PCR amplification to prepare the libraries for capturing or for inverse PCR.

vii) Errors during PCR after capturing or during inverse PCR amplification.

viii) Errors introduced into the UID assignment cycles of Safe-SeqS (Fig. 3).

ix) Errors introduced into the library amplification cycles of Safe-SeqS performed with exogenous UIDs. Note that if UID assignment primers from process *viii* are not completely removed, they could potentially amplify DNA fragments containing errors introduced during these cycles, creating a new supermutant.

x) Errors introduced into the first bridge-PCR cycle on the Illumina flow cell. If amplification is inefficient, an error introduced into the second bridge-PCR cycle could also result in a cluster containing a mutation in most of its component molecules.

xi) Errors in base calling.

Achieving accuracy with Safe-SeqS. With conventional sequencing-by-synthesis approaches, all of the error-producing processes described above are relevant, resulting in a relatively high number of false-positive mutation calls (Tables 1 and 2). Safe-SeqS minimizes the number of false-positive mutation calls in several ways. Safe-SeqS with exogenous UIDs results in the fewest false-positive mutation calls because it requires the fewest enzymatic steps. With exogenous UIDs, error-generating processes *iii*–*vii* are completely eliminated because these steps are not performed. Safe-SeqS with exogenous UIDs also drastically reduces errors resulting from error-generating processes *x* and *xi* because of the way the data are analyzed.

After Safe-SeqS with exogenous UIDs, the only false-positive errors remaining should be those introduced during the UID assignment PCR cycles (error-generating process *viii*) or residual UID-containing primers during the library amplification cycles (error-generating process *ix*). The errors from error-generating process *viii* can be reduced by requiring at least two supermutants to identify a position as mutant. This requirement is reasonable because every preexisting mutation in a double-stranded DNA template should give rise to two supermutants, one from each strand. Furthermore, this requirement would reduce error-generating process *ii* (damaged bases in the original templates) because such bases, when copied, should give rise to only one supermutant. Finally, errors generated during the library amplification cycles (process *ix*) will not be amplified by residual UID-containing primers if those primers are completely removed, such as attempted in our study with excess Exonuclease-I.

With endogenous UIDs, the mistakes introduced by processes *x* and *xi* are drastically reduced because of the way in which the data are analyzed (as with exogenous UIDs). Errors introduced in processes *ii*–*vii* can be minimized by requiring that a mutation be observed in at least two UID families, for the reasons stated in the paragraph above. With this requirement, few false-positive mutations, in theory, should be identified.

In practice, the situation is complicated by the fact that the various amplifications are not perfect, so every strand of every original template molecule is not recovered as a UID family. This efficiency can vary from sample to sample, depending in part on the concentration of inhibitors present in clinical samples. Moreover, with exogenous UIDs, a polymerase error during the library amplification step can create a new UID family that was not represented in the UID assignment step. If this error occurred in a mutant template, an additional, artificial supermutant would be created.

These factors can be managed by incorporating various additional criteria into the analyses. For example, one might require UID families to contain >2, 5, or 10 members. Another requirement could be that the exogenous UIDs of supermutants not be related to any other UID in the library by a one-base difference. Discarding closely related UIDs would eliminate artificial supermutants generated during the library amplification steps (noted in the above paragraph). We routinely discarded such UIDs in our Safe-SeqS analyses, but it made little difference (<1%) in the number of supermutants identified. Specificity for mutations can be further increased by requiring more than one supermutant to identify a position as mutant, as described above for endogenous UIDs. When requiring multiple supermutants, the specificity can be further increased by requiring that each strand of the original double-stranded template contain the mutation or, when libraries are amplified using multiple wells, that rare mutations share an introduced sequence that identifies the well in which the two mutations (i.e., one from each strand) were amplified. Such decisions involve the usual trade-off between specificity and sensitivity. In our experiments with exogenous UIDs (Table 2), we required only one supermutant to identify a position as mutant and included all UID families with more than one member. As the sample preparatory steps required for Safe-SeqS with endogenous UIDs were associated with more error-generating processes than exogenous UIDs, we required two supermutants to identify a position as mutant in the experiments reported in Table 1 and also included all UID families with >1 member.

Mutation prevalences in normal human tissues. The experiments reported in Tables 1 and 2, in which >10,000 templates were assessed, show that mutations are present in the nuclear DNA of normal human cells at a frequency of 3.5×10^{-6}–9.0×10^{-6} mutants/bp depending on the region analyzed. It is difficult to determine whether this low level represents genuine mutations present in the original templates or the sum of genuine mutations plus artifactual mutations from the error-generating processes described above. Mutation prevalences in human cells have not been widely investigated, in part because they are so infrequent. However, several clever techniques to identify rare mutants have been devised and can in principle be used for comparison. Unfortunately, estimates of human mutation prevalences vary widely, ranging from as many as 10^{-5} mutants/bp to as many as 10^{-8} mutants/bp (6–12). In several of these studies, the estimates are complicated by the lack of data on the nature of the actual mutations—they could in some cases be caused by losses of whole chromosomes, in others by missense mutations, and in others mainly by nonsense mutations or small insertions or deletions. Additionally, these studies used various sources of normal cells and examined different genes, making direct comparisons difficult. Estimates of the prevalences and rates of mitochondrial DNA mutations similarly vary (13–19). It will be of interest in future work to analyze the same DNA templates and genes with various technologies to determine the basis for these different estimates.

However, let us assume that all of the mutations identified with Safe-SeqS represent genuine mutations present in the original DNA templates from normal cells. What does this tell us about the number of generations through which these cells have proceeded since the organism was conceived? There is a simple relationship between mutation rate and mutation prevalence: The mutation prevalence equals the product of the mutation rate and the number of generations that the cell has gone through since conception. The somatic mutation rate has been determined in previous studies to be $\sim 10^{-9}$ mutants/bp per generation, although this estimate also varies from study to study for reasons related to those mentioned above with respect to mutation prevalence. Combining this literature-derived estimate of mutation rate with our estimates of mutation prevalence suggests that the normal cells analyzed (lymphocytes, lymphoblastoid cell lines, or colonic mucosae) had proceeded through 3,500–8,900 generations, representing cells dividing every 3–7 d for the individuals examined in this study (average age, 65 y).

Computer simulation of polymerase-introduced errors. The timing of mutations introduced by polymerases greatly alters the final number of mutations observed (20). For example, two mutations would differ in prevalence by ~64-fold if introduced 6 cycles apart (2^6). Because polymerases introduce mutations in a stochastic manner, a simple Monte Carlo method was used for the simulations. In these simulations, we used the manufacturer's estimate of the Phusion polymerase error rate with an appropriate adjustment for ability of Safe-SeqS to detect mutations in only one strand (4). Note that errors introduced in cycle 19, as well as in the 2 UID assignment cycles, would result in changes in only one strand of the duplex—i.e., result in one supermutant rather than two. In each experiment, we assumed that there was a constant efficiency of amplification given by the total number of templates obtained at the end of the experiment (i.e., if the number of UID families was N, then we assumed that the number of templates increased by a factor of $N/2^{19}$ in each cycle). One thousand simulations were performed for each of seven experiments, and the results are reported in Table S2.

1. Herman DS, et al. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6:507–510.
2. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
3. Chory J, Pollard JD Jr. (2001) Separation of small DNA fragments by conventional gel electrophoresis. *Curr Protoc Mol Biol* Chap 2:Unit2.7.
4. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112:29–35.
5. Collins AR (1999) Oxidative DNA damage, antioxidants, and cancer. *BioEssays* 21:238–246.
6. Morley AA, Cox S, Holliday R (1982) Human lymphocytes resistant to 6-thioguanine increase with age. *Mech Ageing Dev* 19:21–26.
7. Trainor KJ, et al. (1984) Mutation frequency in human lymphocytes increases with age. *Mech Ageing Dev* 27:83–86.
8. Grist SA, McCarron M, Kutlaca A, Turner DR, Morley AA (1992) In vivo human somatic mutation: Frequency and spectrum with age. *Mutat Res* 266:189–196.
9. Williams GT, Geraghty JM, Campbell F, Appleton MA, Williams ED (1995) Normal colonic mucosa in hereditary non-polyposis colorectal cancer shows no generalised increase in somatic mutation. *Br J Cancer* 71:1077–1080.
10. Campbell F, Appleton MA, Shields CJ, Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right- and left-sided sporadic colorectal carcinomas. *J Pathol* 186:31–35.
11. Araten DJ, Nafa K, Pakdeesuwan K, Luzzatto L (1999) Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. *Proc Natl Acad Sci USA* 96: 5209–5214.
12. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.
13. Monnat RJ, Jr., Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. *Proc Natl Acad Sci USA* 82:2895–2899.
14. Bodenteich A, Mitchell LG, Merril CR (1991) A lifetime of retinal light exposure does not appear to increase mitochondrial mutations. *Gene* 108:305–309.
15. Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet* 59:501–509.
16. Khrapko K, et al. (1997) Mitochondrial mutational spectra in human cells and tissues. *Proc Natl Acad Sci USA* 94:13798–13803.
17. Heyer E, et al. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* 69:1113–1126.
18. Howell N, et al. (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: There is a difference between phylogenetic and pedigree rates. *Am J Hum Genet* 72:659–670.

19. Taylor RW, et al. (2003) Mitochondrial DNA mutations in human colonic crypt stem cells. *J Clin Invest* 112:1351–1360.

20. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.

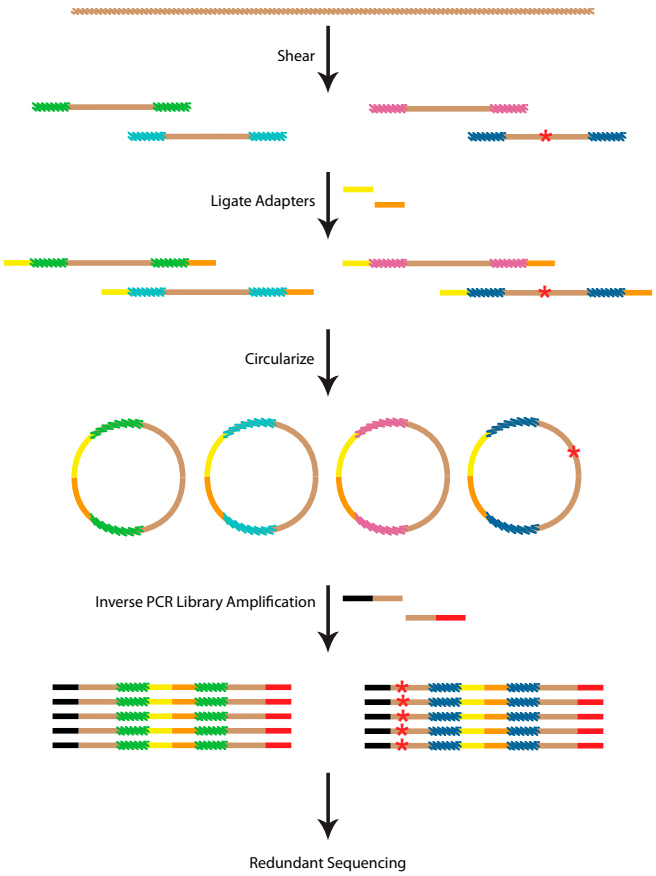


Fig. S1. Safe-SeqS with endogenous UIDs plus inverse PCR. The sequences of the ends of fragments produced by random shearing serve as unique identifiers (UIDs) (variously colored bars). These fragments are ligated to adapters (yellow and orange bars) as in a standard Illumina library preparation. One uniquely tagged fragment is produced from each strand of the double-stranded template; only one strand is shown. Following circularization with a ligase, inverse PCR is performed with gene-specific primers that also contain 5′ "grafting" sequences (black and red bars). This PCR produces UID families that are directly sequenced. Supermutants are defined as in Fig. 1.

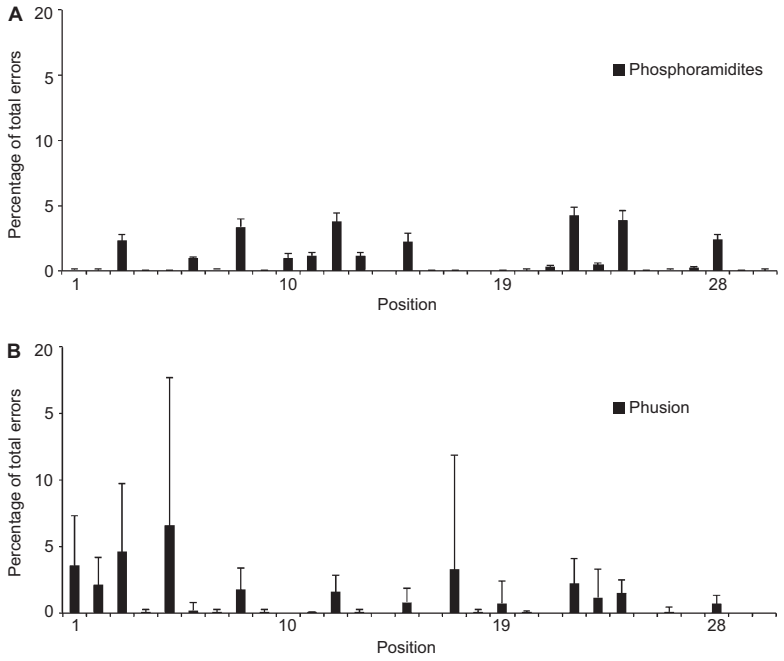


Fig. S2. Single-base substitutions position vs. error frequency in oligonucleotides synthesized with phosphoramidites and Phusion. A representative portion of the same 31-base DNA fragment synthesized with phosphoramidites (*A*) or Phusion polymerase (*B*) was analyzed by Safe-SeqS. The means and SDs for seven independent experiments of each type are plotted. Averages of 1,721 ± 383 and 196 ± 143 single-base substitution supermutants were identified in the phosphoramidite-synthesized and Phusion-generated fragments, respectively. The *y* axis indicates the fraction of the total errors at the indicated position. Note that the errors in the phosphoramidite-synthesized DNA fragment were consistent among the seven replicates, as would be expected if the errors were systematically introduced during the synthesis itself. In contrast, the errors in the Phusion-generated fragments appeared to be heterogeneous among samples, as expected from a stochastic process (20).

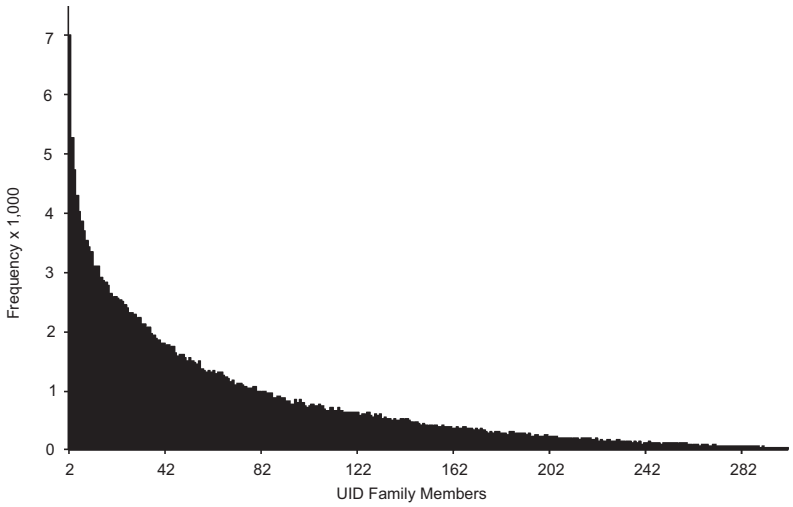


Fig. S3. UID family member distribution. The exogenous UID strategy depicted in Fig. 3 was used to produce PCR fragments from a region of *CTNNB1* from three normal, unrelated individuals (Table 2, *CTNNB1* mutations in DNA from normal human cells); a representative example of the UID families with ≤300 members (99% of total UID families) generated from one individual is shown. The *y* axis indicates the number of different UID families that contained the number of family members shown on the *x* axis.

Table S1. Fraction of single-base substitutions, insertions, and deletions with exogenous UIDs

	Mean	SD
Polymerase fidelity		
Conventional analysis of seven replicates		
Total mutations identified	198,638	22,515
% mutations represented by single-base substitutions	98.90	0.16
% mutations represented by deletions	1.05	0.16
% mutations represented by insertions	0.05	0.01
Safe-SeqS analysis of seven replicates		
Total supermutants identified	197	143
% supermutants represented by single-base substitutions	99.10	2.06
% supermutants represented by deletions	0.90	2.06
% supermutants represented by insertions	0.00	0.00
CTNNB1 mutations in DNA from normal human cells		
Conventional analysis of three individuals		
Total mutations identified	118,488	11,357
% mutations represented by single-base substitutions	97.38	0.30
% mutations represented by deletions	2.62	0.30
% mutations represented by insertions	0.00	0.00
Safe-SeqS analysis of three individuals		
Total supermutants identified	99	78
% supermutants represented by single-base substitutions	99.52	0.84
% supermutants represented by deletions	0.48	0.84
% supermutants represented by insertions	0.00	0.00
Mitochondrial mutations in DNA from normal human cells		
Conventional analysis of seven individuals		
Total mutations identified	30,599	12,970
% mutations represented by single-base substitutions	98.42	0.64
% mutations represented by deletions	1.58	0.64
% mutations represented by insertions	0.00	0.00
Safe-SeqS analysis of seven individuals		
Total supermutants identified	135	61
% supermutants represented by single-base substitutions	98.66	1.39
% supermutants represented by deletions	1.34	1.39
% supermutants represented by insertions	0.00	0.00

Table S2. Observed and expected number of errors generated by Phusion polymerase

	Observed	Expected (mean ± SD)*
Experiment 1		
Mutations represented by 1 supermutant	10	19 ± 3.7
Mutations represented by 2 supermutants	8	5.8 ± 2.3
Mutations represented by 3 supermutants	4	1.3 ± 1.1
Mutations represented by 4 supermutants	4	1.8 ± 1.3
Mutations represented by 5 supermutants	2	0.61 ± 0.75
Mutations represented by 6 supermutants	2	0.22 ± 0.44
Mutations represented by 7 supermutants	0	0.01 ± 0.10
Mutations represented by 8 supermutants	0	0.87 ± 0.86
Mutations represented by 9 supermutants	2	0.28 ± 0.51
Mutations represented by 10 supermutants	0	0.14 ± 0.38
Mutations represented by >10 supermutants	3	1.5 ± 2.7
Distinct mutations	35	32 ± 4.2
Experiment 2		
Mutations represented by 1 supermutant	19	23 ± 4.1
Mutations represented by 2 supermutants	5	9.5 ± 2.8
Mutations represented by 3 supermutants	4	2.7 ± 1.6
Mutations represented by 4 supermutants	7	2.7 ± 1.7
Mutations represented by 5 supermutants	2	0.88 ± 0.94
Mutations represented by 6 supermutants	1	0.40 ± 0.60
Mutations represented by 7 supermutants	3	0.16 ± 0.42
Mutations represented by 8 supermutants	1	0.99 ± 1.0
Mutations represented by 9 supermutants	1	0.39 ± 0.68
Mutations represented by 10 supermutants	0	0.17 ± 0.43
Mutations represented by >10 supermutants	9	1.8 ± 3.4
Distinct mutations	52	43 ± 5.1
Experiment 3		
Mutations represented by 1 supermutant	7	17 ± 3.4
Mutations represented by 2 supermutants	9	5.4 ± 2.0
Mutations represented by 3 supermutants	4	1.2 ± 1.1
Mutations represented by 4 supermutants	4	1.7 ± 1.4
Mutations represented by 5 supermutants	2	0.50 ± 0.70
Mutations represented by 6 supermutants	0	0.17 ± 0.45
Mutations represented by 7 supermutants	1	0.03 ± 0.17
Mutations represented by 8 supermutants	0	0.59 ± 0.74
Mutations represented by 9 supermutants	0	0.24 ± 0.50
Mutations represented by 10 supermutants	1	0.07 ± 0.29
Mutations represented by >10 supermutants	5	1.5 ± 2.6
Distinct mutations	33	28 ± 3.7
Experiment 4		
Mutations represented by 1 supermutant	7	15 ± 3.7
Mutations represented by 2 supermutants	8	4.1 ± 1.7
Mutations represented by 3 supermutants	2	0.70 ± 0.74
Mutations represented by 4 supermutants	1	1.5 ± 1.3
Mutations represented by 5 supermutants	3	0.21 ± 0.52
Mutations represented by 6 supermutants	2	0.08 ± 0.27
Mutations represented by 7 supermutants	1	0.0 ± 0.0
Mutations represented by 8 supermutants	2	0.65 ± 0.77
Mutations represented by 9 supermutants	2	0.17 ± 0.43
Mutations represented by 10 supermutants	0	0.05 ± 0.22
Mutations represented by >10 supermutants	1	0.92 ± 2.1
Distinct mutations	29	23 ± 3.2
Experiment 5		
Mutations represented by 1 supermutant	9	23 ± 4.1
Mutations represented by 2 supermutants	6	9.5 ± 2.8
Mutations represented by 3 supermutants	5	2.7 ± 1.6
Mutations represented by 4 supermutants	3	2.7 ± 1.7
Mutations represented by 5 supermutants	6	0.88 ± 0.94
Mutations represented by 6 supermutants	2	0.40 ± 0.60
Mutations represented by 7 supermutants	1	0.16 ± 0.42
Mutations represented by 8 supermutants	2	0.99 ± 1.0
Mutations represented by 9 supermutants	2	0.39 ± 0.68
Mutations represented by 10 supermutants	3	0.17 ± 0.43

Table S2. Cont.

	Observed	Expected (mean ± SD)*
Mutations represented by >10 supermutants	7	1.8 ± 3.4
Distinct mutations	46	43 ± 5.1
Experiment 6		
Mutations represented by 1 supermutant	4	6.7 ± 2.8
Mutations represented by 2 supermutants	7	1.5 ± 1.2
Mutations represented by 3 supermutants	1	0.10 ± 0.33
Mutations represented by 4 supermutants	2	0.60 ± 0.82
Mutations represented by 5 supermutants	0	0.07 ± 0.26
Mutations represented by 6 supermutants	0	0.01 ± 0.10
Mutations represented by 7 supermutants	1	0.0 ± 0.0
Mutations represented by 8 supermutants	1	0.39 ± 0.60
Mutations represented by 9 supermutants	0	0.01 ± 0.10
Mutations represented by 10 supermutants	0	0.0 ± 0.0
Mutations represented by >10 supermutants	2	0.50 ± 1.1
Distinct mutations	18	9.9 ± 1.4
Experiment 7		
Mutations represented by 1 supermutant	8	2.9 ± 1.6
Mutations represented by 2 supermutants	2	0.61 ± 0.79
Mutations represented by 3 supermutants	0	0.04 ± 0.24
Mutations represented by 4 supermutants	0	0.41 ± 0.59
Mutations represented by 5 supermutants	1	0.01 ± 0.10
Mutations represented by 6 supermutants	0	0.0 ± 0.0
Mutations represented by 7 supermutants	0	0.0 ± 0.0
Mutations represented by 8 supermutants	0	0.14 ± 0.35
Mutations represented by 9 supermutants	0	0.01 ± 0.10
Mutations represented by 10 supermutants	0	0.0 ± 0.0
Mutations represented by >10 supermutants	0	0.32 ± 0.93
Distinct mutations	11	4.5 ± 0.62

*See *SI Materials and Methods* for details of the simulations.

Table S3. Phosphoramidite- vs. Phusion-synthesized DNA: transitions vs. transversions comparison

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Average	SD
Phosphoramidites									
Transition supermutants	496	509	471	396	323	273	470	420	92
Transversion supermutants	1,494	1,499	1,521	1,154	944	907	1,626	1,306	298
P value*	3.4*E*-05								
Phusion									
Transition supermutants	63	275	127	5	87	182	103	120	87
Transversion supermutants	14	124	77	12	57	191	63	77	63
P value*	0.08								

**P* values were calculated using a two-tailed paired *t* test. Exp., experiment.

Table S4. Oligonucleotides used in this study

	Sequence
Endogenous UIDs	
Capture	
Adapter, strand 1	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adapter, strand 2	ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for	AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, rev	CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T
Postcapture amplification, for	AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Postcapture amplification, rev	CAAGCAGAAGACGGCATACGAGATCTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T
Sequencing primer, read 1 (Illumina)	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Sequencing primer, read 2 (Illumina)	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Inverse PCR	
Adapter, strand 1	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Adapter, strand 2	ACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-1	/5Phos/CGTGATACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-2	/5Phos/ACATCGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-3	/5Phos/GCCTAAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-4	/5Phos/TGGTCAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-5	/5Phos/CACTGTACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-6	/5Phos/ATTGGCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-7	/5Phos/GATCTGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-8	/5Phos/TCAAGTACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-9	/5Phos/CTGATCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-10	/5Phos/AAGCTAACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-11	/5Phos/GTAGCCACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, for-12	/5Phos/TACAAGACACTCTTTCCCTACACGACGCTCTTCCGAT*C*T
Whole-genome amplification, rev	/5Phos/CTCGGCATTCCTGCTGAACCGCTCTTCCGAT*C*T
Inverse PCR, antisense	AATGATACGGCGACCACCGAGATCTACACCAGCAGGCCTTATAATAAAAATAATGA
Inverse PCR, for	CAAGCAGAAGACGGCATACGAGATTGACTGAATATAAACTTGTGGTAGTTG
Sequencing primer 1 (to read internal sequences)	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Sequencing primer 2 (to read internal sequences)	CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
Index primer 1 (to read experiment indexes)	CGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Index primer 2 (to read experiment indexes)	CGGAAGAGCGGTTCAGCAGGAATGCCGAG
Exogenous UIDs	
Polymerase fidelity	
Digital PCR amplification, for	GGTTACAGGCTCATGATGTAACC
Digital PCR amplification, rev	GATACCAGCTTGGTAATGGCA
UID assignment amplification, for	CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNGGTTACAGGCTCATGATGTAACC
UID assignment amplification, rev	CACACAGGAAACAGCTATGACCATGGATACCAGCTTGGTAATGGCA
Library amplification, for-1	AATGATACGGCGACCACCGAGATCTACACCGTGATCGACGTAAAACGACGGCCA*G*T
Library amplification, for-2	AATGATACGGCGACCACCGAGATCTACACACATCGCGACGTAAAACGACGGCCA*G*T
Library amplification, for-3	AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCCA*G*T
Library amplification, for-4	AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCCA*G*T
Library amplification, for-5	AATGATACGGCGACCACCGAGATCTACACCACTGTCGACGTAAAACGACGGCCA*G*T
Library amplification, for-6	AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCCA*G*T
Library amplification, for-7	AATGATACGGCGACCACCGAGATCTACACGATCTGCGACGTAAAACGACGGCCA*G*T
Library amplification, for-8	AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCCA*G*T
Library amplification, for-9	AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCCA*G*T
Library amplification, for-10	AATGATACGGCGACCACCGAGATCTACACAAGCTACGACGTAAAACGACGGCCA*G*T
Library amplification, rev	CAAGCAGAAGACGGCATACGAGATCACACAGGAAACAGCTATGACCA*T*G
Sequencing primer (to read UID and internal sequences)	CGACGTAAAACGACGGCCAGT
Index primer (to read experiment indexes)	ACTGGCCGTCGTTTTACGTCG

Table S4. Cont.

	Sequence
CTNNB1 mutations in DNA from normal human cells	
UID assignment amplification, for	CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNNNGCAGCAACAGTCTTACCTGGACT
UID assignment amplification, rev	CACACAGGAAACAGCTATGACCATGTCCACATCCTCTTCCTCAGGATT
Library amplification, for	AATGATACGGCGACCACCGAGATCTACACCGACGTAAAACGACGGCCA*G*T
Library amplification, rev-1	CAAGCAGAAGACGGCATACGAGATATCACGCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-2	CAAGCAGAAGACGGCATACGAGATCGATGTCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-3	CAAGCAGAAGACGGCATACGAGATTGACCACACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-4	CAAGCAGAAGACGGCATACGAGATGCCAATCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-5	CAAGCAGAAGACGGCATACGAGATCAGATCCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-6	CAAGCAGAAGACGGCATACGAGATACTTGACACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-7	CAAGCAGAAGACGGCATACGAGATGATCAGCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-8	CAAGCAGAAGACGGCATACGAGATTAGCTTCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-9	CAAGCAGAAGACGGCATACGAGATGGCTACCACACAGGAAACAGCTATGACCA*T*G
Library amplification, rev-10	CAAGCAGAAGACGGCATACGAGATCTTGTACACACAGGAAACAGCTATGACCA*T*G
Sequencing primer (to read UID and internal sequences)	CGACGTAAAACGACGGCCAGT
Index primer (to read experiment indexes)	CATGGTCATAGCTGTTTCCTGTGTG
Mitochondrial mutations in DNA from normal human cells	
UID assignment amplification, for	CGACGTAAAACGACGGCCAGTNNNNNNNNNNNNNNNTTACCGAGAAAGCTCACAAGAA
UID assignment amplification, rev	CACACAGGAAACAGCTATGACCATGATGCTAAGGCGAGGATGAAA
Library amplification, for-1	AATGATACGGCGACCACCGAGATCTACACACATCGCGACGTAAAACGACGGCCA*G*T
Library amplification, for-2	AATGATACGGCGACCACCGAGATCTACACGCCTAACGACGTAAAACGACGGCCA*G*T
Library amplification, for-3	AATGATACGGCGACCACCGAGATCTACACTGGTCACGACGTAAAACGACGGCCA*G*T
Library amplification, for-4	AATGATACGGCGACCACCGAGATCTACACATTGGCCGACGTAAAACGACGGCCA*G*T
Library amplification, for-5	AATGATACGGCGACCACCGAGATCTACACGATCTGCGACGTAAAACGACGGCCA*G*T
Library amplification, for-6	AATGATACGGCGACCACCGAGATCTACACTCAAGTCGACGTAAAACGACGGCCA*G*T
Library amplification, for-7	AATGATACGGCGACCACCGAGATCTACACCTGATCCGACGTAAAACGACGGCCA*G*T
Library amplification, rev	CAAGCAGAAGACGGCATACGAGATCACACAGGAAACAGCTATGACCA*T*G
Sequencing primer 1 (to read UIDs)	CGACGTAAAACGACGGCCAGT
Sequencing primer 2 (to read internal sequences)	CCTAATTCCCCCCATCCTTAC
Index primer (to read experiment indexes)	ACTGGCCGTCGTTTTACGTCG
Analysis of phosphoramidite oligonucleotide composition	
Synthesized template, WT	GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTGTAACTTTAAAACATATTTTTGCCATTACCAAGCTGGTATC
Synthesized template, mut (*S* = 50/50 mix of C and G)	GGTTACAGGCTCATGATGTAACCTCTGTGTCTTGGTGSAACTTTAAAACATATTTTTGCCATTACCAAGCTGGTATC
UID assignment amplification, for	ACACTCTTTCCCTACACGACGCTCNNNNNNNNNNNNGGTGAGTCTGTGCAGGCAT
UID assignment amplification, rev	CTCGAGCACTGTCCTGACTGAGACGATACCAGCTTGGTAATGGCA
Library amplification, for	AATGATACGGCGACCACCGAGATCTACACCGTGATACACTCTTTCCCTACACGACGC*T*C
Library amplification, rev	CAAGCAGAAGACGGCATACGAGATCTCGAGCACTGTCCTGACTGAG*A*C
Sequencing primer (to read UID and internal sequences)	ACACTCTTTCCCTACACGACGCTC

Colors: blue, region complementary to templates; green, template-specific UID sequence; black, universal sequence; purple, experiment-specific index sequence; red, Illumina grafting primers (for hybridization to flow cell). /5Phos/, 5′ phosphate; *, phosphorothioate linkage; for, forward; rev, reverse.

ATTACHMENT 9A

LDR	03943cas a2200937 4500
001	994619993406676
005	20250227140822.0
007	cr|cnu||||||||
008	750626c19159999dcuwr p 0 0eng
010	__ | **a** 16010069
022	__ | **a**0027-8424**l**0027-8424**z**0744-2831**z**0273-1134**z**0273-1142**9**PY**7**(Print)
022	__ | **a**1091-6490**9**EN**7**(Electronic)
030	__ | **a**PNASA6
035	__ | **9**7505876
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
042	__ | **a**nsdp**a**lc
044	__ | **9**United States
060	00 | **a**W1**b**PR586A
210	10 | **a**Proc Natl Acad Sci U S A**2**DNLM
245	00 | **a**Proceedings of the National Academy of Sciences of the United States of America.
246	1_ | **i**Vols. for <1999-> also called:**a**PNAS
246	2_ | **0**P N A S**9**n
246	2_ | **0**PNAS USA**9**n
246	2_ | **0**PNAS**9**n
246	2_ | **0**PROC NAT ACAD SCI USA**9**n

246	2_ | **0**PROC NAT ACAD SCI**9**n
246	2_ | **0**PROC NATL ACAD SCI U S A**9**n
246	2_ | **0**PROC NATL ACAD SCI USA**9**n
246	2_ | **0**PROC NATL ACAD SCI**9**n
246	2_ | **0**PROC NATL ACAD SCI, U S A**9**n
246	2_ | **0**Proc Natl Acad Sci U**9**n
246	2_ | **0**Proc Natl Acad Sci USA**9**n
246	2_ | **0**Proc Natl Acad Sci**9**n
246	2_ | **0**Proc Natl Acad Sci, U S A**9**n
246	2_ | **0**Proc Natl Acad Sci, USA**9**n
246	2_ | **0**Proc Nattl Acad Sci USA**9**n
246	2_ | **0**Proc. Natl. Acad. Sci. USA**9**n
246	2_ | **0**Proc.Nat.Acad.Sci.Usa**9**n
246	2_ | **0**Proceedings of the National Academy of Sciences of the USA**9**n
246	2_ | **0**Proc Nat Acad Sci USA**9**om
246	2_ | **0**Proc. Nat. Acad. Sc. U. S.**9**om
260	__ | **a**Washington, DC :**b**National Academy of Sciences
300	__ | **a**v.**b**illus.
310	__ | **a**Weekly,**b**2004-
321	__ | **a**Semimonthly,**b**19??-1993
321	__ | **a**Biweekly,**b**1994-2003
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia

338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**v. 1- Jan. 15, 1915-
500	__ | **a**Not to be confused with another publication with the same title covering Mar. 1863-Oct. 1894, issued in 3 pt.: pt. 1, 1877; pt. 2, 1884; pt. 3, 1895.
500	__ | **a**Vol. 13, no. 1, pt. 2, includes Index (v. 1-10, 1915-1924) to the Proceedings; List of other publications of the Academy from 1863-1926; and List of publications of the National Research Council from 1916-1925.
500	__ | **a**Issues for 1980-1984 in two parts with consecutive page numbering: Physical sciences and Biological sciences.
510	2_ | **9**1**a**MEDLINE**b**v53n5,May 1965-
510	2_ | **9**1**a**PubMed**b**v1,1915-
510	0_ | **a**OLDMEDLINE
510	0_ | **a**PMC
510	2_ | **a**Index medicus
515	__ | **a**Some issues accompanied by supplements with distinctive titles.
530	__ | **a**Also issued online.
533	__ | **a**Microfilm.**m**v. 1 (1915)-v.15 (1929) + index v. 1 (1915)-v.10 (1924)**b**[Bethesda, Md.] :**c**National Library of Medicine,**d**1998.**e**7 microfilm reels : negative ; 35 mm.
555	__ | **a**Vols. 1-10, 1915-24, in v. 13 (Includes list of academy publications, 1863-1926, and list of publications of the National Research Council, 1916-25); Author index: Vols. 1-50, 1914-63. 1 v.
583	__ | **3**v. 16 (1930)-**a**Will not microfilm;**z**PubMed Central title.**c**20040413**5**DNLM
583	__ | **3**Vol. 84, no. 1 (JAN 1987)**c**19890818**f**Notice not published**l**Printed on acid-free paper
650	_2 | **a**Science
655	_2 | **a**Periodical

710	2_ | **a**National Academy of Sciences (U.S.)
710	2_ | **a**National Research Council (U.S.).**t**List of publications, 1916-1925.
856	4_ | **u**https://pmc.ncbi.nlm.nih.gov/journals/?term=7505876
856	4_ | **u**https://www.pnas.org/
911	__ | **a**Electronic and print, unable to acquire print format
914	__ | **a**Portico**b**Current to 2025 as of 20250227**c**ISSN_00278424
935	__ | **a**(DNLM)461999-nlmdb
995	__ | **a**AUTH**b**19750626**c**REV**d**20250227
996	__ | **a**CONSER**b**20000101
997	__ | **a**OMR
998	__ | **a**ERMSLC 170140**b**20110208
999	__ | **a**AUTH
035	__ | **a**(DNLM)P39620000(s)
035	__ | **a**(OCoLC)01607201

ATTACHMENT 9B

Science MeSH Descriptor Data 2025

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading
Science
Tree Number(s)
H01.770
Unique ID
D012586
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D012586
Annotation
use for the discipline (education, history, etc) only
Scope Note
The study of natural phenomena by observation, measurement, and experimentation.
See Also
Religion and Science
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2015/06/16

Vol 456|6 November 2008|doi:10.1038/nature07517

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally used long (400–800 base pair) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. Here we report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high-quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30× average depth of paired 35-base reads. We characterize four million single-nucleotide polymorphisms and four hundred thousand structural variants, many of which were previously unknown. Our approach is effective for accurate, rapid and economical whole-genome re-sequencing and many other biomedical applications.

DNA sequencing yields an unrivalled resource of genetic information. We can characterize individual genomes, transcriptional states and genetic variation in populations and disease. Until recently, the scope of sequencing projects was limited by the cost and throughput of Sanger sequencing. The raw data for the three billion base (3 gigabase (Gb)) human genome sequence, completed in 2004 (ref. 1), was generated over several years for ~$300 million using several hundred capillary sequencers. More recently an individual human genome sequence has been determined for ~$10 million by capillary sequencing[2]. Several new approaches at varying stages of development aim to increase sequencing throughput and reduce cost[3–6]. They increase parallelization markedly by imaging many DNA molecules simultaneously. One instrument run produces typically thousands or millions of sequences that are shorter than capillary reads. Another human genome sequence was recently determined using one of these approaches[7]. However, much bigger improvements are necessary to enable routine whole human genome sequencing in genetic research.

We describe a massively parallel synthetic sequencing approach that transforms our ability to use DNA and RNA sequence information in biological systems. We demonstrate utility by re-sequencing an individual human genome to high accuracy. Our approach delivers data at very high throughput and low cost, and enables extraction of genetic information of high biological value, including single-nucleotide polymorphisms (SNPs) and structural variants.

DNA sequencing using reversible terminators

We generated high-density single-molecule arrays of genomic DNA fragments attached to the surface of the reaction chamber (the flow cell) and used isothermal 'bridging' amplification to form DNA 'clusters' from each fragment. We made the DNA in each cluster single-stranded and added a universal primer for sequencing. For paired read sequencing, we then converted the templates to double-stranded DNA and removed the original strands, leaving the complementary strand as template for the second sequencing reaction (Fig. 1a–c). To obtain paired reads separated by larger distances, we circularized DNA fragments of the required length (for example, 2 ± 0.2 kb) and obtained short junction fragments for paired end sequencing (Fig. 1d).

We sequenced DNA templates by repeated cycles of polymerase-directed single base extension. To ensure base-by-base nucleotide incorporation in a stepwise manner, we used a set of four reversible terminators, 3′-*O*-azidomethyl 2′-deoxynucleoside triphosphates (A, C, G and T), each labelled with a different removable fluorophore (Supplementary Fig. 1a)[8]. The use of 3′-modified nucleotides allowed the incorporation to be driven essentially to completion without risk of over-incorporation. It also enabled addition of all four nucleotides simultaneously rather than sequentially, minimizing risk of misincorporation. We engineered the active site of 9°N DNA polymerase to improve the efficiency of incorporation of these unnatural nucleotides[9]. After each cycle of incorporation, we determined the identity of the inserted base by laser-induced excitation of the fluorophores and imaging. We added tris(2-carboxyethyl)phosphine (TCEP) to remove the fluorescent dye and side arm from a linker attached to the base and simultaneously regenerate a 3′ hydroxyl group ready for the next cycle of nucleotide addition (Supplementary Fig. 1b). The Genome Analyzer (GA1) was designed to perform multiple cycles of sequencing chemistry and imaging to collect the sequence data automatically from each cluster on the surface of each lane of an eight-lane flow cell (Supplementary Fig. 2).

To determine the sequence from each cluster, we quantified the fluorescent signal from each cycle and applied a base-calling algorithm. We defined a quality (Q) value for each base call (scaled as by the phred algorithm[10]) that represents the likelihood of each call being correct (Supplementary Fig. 3). We used the Q-values in subsequent analyses to weight the contribution of each base to sequence alignment and detection of sequence variants (for example, SNP

©2008 Macmillan Publishers Limited. All rights reserved

EX1055

calling). We discarded all reads from mixed clusters and used the remaining 'purity filtered' reads for analysis. Typically we generated 1–2 Gb of high-quality purity filtered sequence per flow cell from ~30–60-million single 35-base reads, or 2–4 Gb in a paired read experiment (Supplementary Table 1).

To demonstrate accurate sequencing of human DNA, we sequenced a human bacterial artificial chromosome (BAC) clone (bCX98J21) that contained 162,752 bp of the major histocompatibility complex on human chromosome 6 (accession AL662825.4, previously determined using capillary sequencing by the Wellcome Trust Sanger Institute). We developed a fast global alignment algorithm ELAND that aligns a read to the reference only if the read can be assigned a unique position with 0, 1 or 2 differences. We collected 0.17 Gb of aligned data for the BAC from one lane of a flow cell. Approximately 90% of the 35-base reads matched perfectly to the reference, demonstrating high raw read accuracy (Supplementary Fig. 4). To examine consensus coverage and accuracy, we used 5 Mb of 35-base purity filtered reads (30-fold average input depth of the BAC) and obtained 99.96% coverage of the reference. There was one consensus miscall, at a position of very low coverage (just above our cutoff threshold), yielding an overall consensus accuracy of >99.999%.

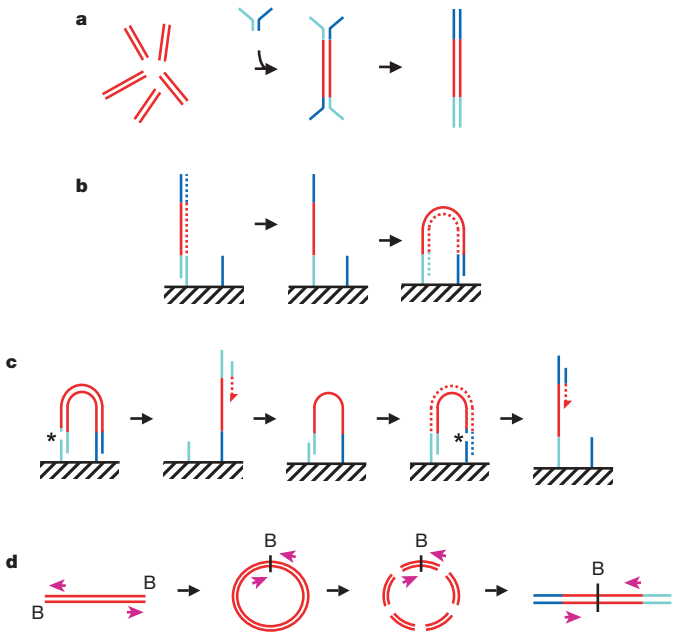


Figure 1 | Preparation of samples. **a**, DNA fragments are generated, for example, by random shearing and joined to a pair of oligonucleotides in a forked adaptor configuration. The ligated products are amplified using two oligonucleotide primers, resulting in double-stranded blunt-ended material with a different adaptor sequence on either end. **b**, Formation of clonal single-molecule array. DNA fragments prepared as in **a** are denatured and single strands are annealed to complementary oligonucleotides on the flow-cell surface (hatched). A new strand (dotted) is copied from the original strand in an extension reaction that is primed from the 3′ end of the surface-bound oligonucleotide; the original strand is then removed by denaturation. The adaptor sequence at the 3′ end of each copied strand is annealed to a new surface-bound complementary oligonucleotide, forming a bridge and generating a new site for synthesis of a second strand (dotted). Multiple cycles of annealing, extension and denaturation in isothermal conditions result in growth of clusters, each ~1 μm in physical diameter. This follows the basic method outlined in ref. 33. **c**, The DNA in each cluster is linearized by cleavage within one adaptor sequence (gap marked by an asterisk) and denatured, generating single-stranded template for sequencing by synthesis to obtain a sequence read (read 1; the sequencing product is dotted). To perform paired-read sequencing, the products of read 1 are removed by denaturation, the template is used to generate a bridge, the second strand is re-synthesized (shown dotted), and the opposite strand is then cleaved (gap marked by an asterisk) to provide the template for the second read (read 2). **d**, Long-range paired-end sample preparation. To sequence the ends of a long (for example, >1 kb) DNA fragment, the ends of each fragment are tagged by incorporation of biotinylated (B) nucleotide and then circularized, forming a junction between the two ends. Circularized DNA is randomly fragmented and the biotinylated junction fragments are recovered and used as starting material in the standard sample preparation procedure illustrated in **a**. The orientation of the sequence reads relative to the DNA fragment is shown (magenta arrows). When aligned to the reference sequence, these reads are oriented with their 5′ ends towards each other (in contrast to the short insert paired reads produced as shown in **a**–**c**). See Supplementary Fig. 17a for examples of both. Turquoise and blue lines represent oligonucleotides and red lines represent genomic DNA. All surface-bound oligonucleotides are attached to the flow cell by their 5′ ends. Dotted lines indicate newly synthesized strands during cluster formation or sequencing. (See Supplementary Methods for details.)

Detecting genetic variation of the human X chromosome

For an initial study of genetic variation, we sequenced flow-sorted X chromosomes of a Caucasian female (sample NA07340 originating from the Centre d'Etude du Polymorphisme Humain (CEPH)). We generated 278-million paired 30–35-bp purity filtered reads and aligned them to the human genome reference sequence. We carried out separate analyses of the data using two alignment algorithms: ELAND (see above) or MAQ (Mapping and Assembly with Qualities)[11]. Both algorithms place each read pair where it best matches the reference and assign a confidence score to the alignment. In cases where a read has two or more equally likely positions (that is, in an exact repeat), MAQ randomly assigns the read pair to one position and assigns a zero alignment quality score (these reads are excluded from SNP analysis). ELAND rejects all non-unique alignments, which are mostly in recently inserted retrotransposons (see Supplementary Fig. 5). MAQ therefore provides an opportunity to assess the properties of a data set aligned to the entire reference, whereas ELAND effectively excludes ambiguities from the short read alignment before further analysis.

We obtained comprehensive coverage of the X chromosome from both analyses. With MAQ, 204 million reads aligned to 99.94% of the X chromosome at an average depth of 43×. With ELAND, 192 million reads covered 91% of the reference sequence, showing what can be covered by unique best alignments. These results were obtained after excluding reads aligning to non-X sequence (impurities of flow sorting) and apparently duplicated read pairs (Supplementary Table 2). We reasoned that these duplicates (~10% of the total) arose during initial sample amplification.

The sampling of sequence fragments from the X chromosome is close to random. This is evident from the distribution of mapped read depth in the MAQ alignment in regions where the reference is unique (Fig. 2a): the variance of this distribution is only 2.26 times that of a Poisson distribution (the theoretical minimum). Half of this excess variance can be accounted for by a dependence on G+C content. However, the average mapped read depth only falls below 10× in regions with G+C content less than 4% or greater than 76%, comprising in total just 1% of unique chromosome sequence and 3% of coding sequence (Fig. 2b).

We identified 92,485 candidate SNPs in the X chromosome using ELAND (Supplementary Fig. 6). Most calls (85%) match previous entries in the public database dbSNP. Heterozygosity (π) in this data set is 4.3×10^{-4} (that is, one substitution per 2.3 kb), close to a previously published X chromosome estimate (4.7×10^{-4})[12]. Using MAQ we obtained 104,567 SNPs, most of which were common to the results of the ELAND analysis. The differences between the two sets of SNP calls are largely the consequence of different properties of the alignments as described earlier. For example, most of the SNPs found only by the MAQ-based analysis were at positions of low or zero sequence depth in the ELAND alignment (Supplementary Fig. 6c).

We assessed accuracy and completeness of SNP calling by comparison to genotypes obtained for this individual using the Illumina HumanHap550 BeadChip (HM550). The sequence data covered >99.8% of the 13,604 genotyped positions and we found excellent

©2008 Macmillan Publishers Limited. All rights reserved

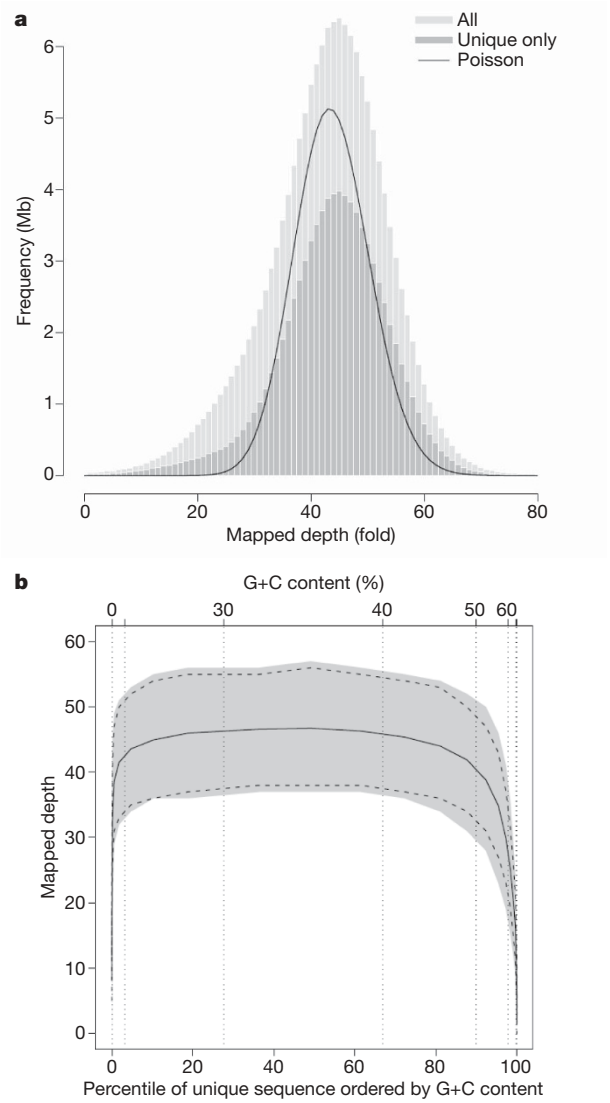


Figure 2 | X chromosome data. a, Distribution of mapped read depth in the X chromosome data set (NA07340), sampled at every 50th position along the chromosome and displayed as a histogram ('All'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('Unique only'). The solid line represents a Poisson distribution with the same mean. **b**, Distribution of X chromosome uniquely mapped reads as a function of G+C content. Note that the *x* axis is per cent G+C content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 90th centiles); the dashed lines are 10th and 90th centiles of a Poisson distribution with the same mean as the data.

agreement between sequence-based SNP calls and genotyping data (99.52% or 99.99% using ELAND or MAQ, respectively; Supplementary Table 3). There was complete concordance of all homozygous calls and a low level of 'under-calling' from the sequence data (denoted as 'GT>Seq' in Table 1) at a small number of the heterozygous sites, caused by inadequate sampling of one of the two alleles. The depth of input sequence influences the coverage and accuracy of SNP calling. We found that reducing the read depth to 15× still gives 97% coverage of genotype positions and only 1.27% of the heterozygous sites are under-called. We observed no other types of disagreement at any input depth (Supplementary Fig. 7).

We detected structural variants (defined as any variant other than a single base substitution) as follows. We found 9,747 short insertions/deletions ('short indels'; defined here as less than the length of the read) by performing a gapped alignment of individual reads (Supplementary Fig. 8). We identified larger indels based on read depth and/or anomalous read pair spacing, similar to previous approaches[13–15]. We detected 115 indels in total, 77 of which were visible from anomalous read-pair spacing (see Supplementary Tables 4 and 5). We developed Resembl, an extension to the Ensembl browser[16], to view all variants (Supplementary Fig 9). Inversions can be detected when the orientation of one read in a pair is reversed (for example, see Supplementary Fig. 10). In general, inversions occur as the result of non-allelic homologous recombination, and are therefore flanked by repetitive sequence that can compromise alignments. We found partial evidence for other inversion events, but characterization of inversions from short read data is complex because of the repeats and requires further development.

Sequencing and analysis of a whole human genome

Our X chromosome study enabled us to develop an integrated set of methods for rapid sequencing and analysis of whole human genomes. We sequenced the genome of a male Yoruba from Ibadan, Nigeria (YRI, sample NA18507). This sample was originally collected for the HapMap project[17,18] through a process of community engagement and informed consent[19] and has also been studied in other projects[20,21]. We were therefore able to compare our results with publicly available data from the same sample. We constructed two libraries: one of short inserts (~200 bp) with similar properties to the previous X chromosome library and one from long fragments (~2 kb) to provide longer-range read-pair information (see Supplementary Fig. 11 for size distributions). We generated 135 Gb of sequence (~4 billion paired 35-base reads; see Supplementary Table 6) over a period of 8 weeks (December 2007 to January 2008) on six GA1 instruments averaging 3.3 Gb per production run (see Supplementary Table 1 for example). The approximate consumables cost (based on full list price of reagents) was $250,000. We aligned 97% of the reads using MAQ and found that 99.9% of the human reference (NCBI build 36.1) was covered with one or more reads at an average of 40.6-fold depth. Using ELAND, we aligned 91% of the reads over 93% of the reference sequence at sufficient depth to call a strong consensus (>three Q30 bases). The distribution of mapped read depth was close to random, with slight over-dispersion as seen for the X chromosome data. We observed comprehensive representation across a wide range of G+C content, dropping only at the very extreme ends, but with a different pattern of distribution compared to the X chromosome (see Supplementary Fig. 12).

We identified ~4 million SNPs, with 74% matching previous entries in dbSNP (Fig. 3). We found excellent agreement of our SNP calls with genotyping results: sequence-based SNP calls covered almost all of the 552,710 loci of HM550, with >99.5% concordance of sequencing versus genotyping calls (Table 1 and Supplementary Table 7a). The few disagreements were mostly under-calls of heterozygous positions (GT>Seq) in areas of low sequence depth, providing us with a false-negative rate of <0.35% from the ELAND analysis (see Table 1). The other disagreements (0.09% of all genotypes) included errors in genotyping plus apparent tri-allelic SNPs (Supplementary Table 7a). The main cause of genotype error (0.05% of all genotypes) is the existence of a second 'hidden' SNP close to the assayed locus that disrupts the genotyping assay, leading to loss of one allele and an erroneous homozygous genotype (Supplementary Figs 13 and 14).

To examine the accuracy of SNP calling in more detail, we compared our sequence-based SNP calls with 3.7 million genotypes (HM-All) generated for this sample during the HapMap project (Table 1 and Supplementary Table 7b)[18] and found excellent concordance between the data sets. Disagreements included sequence-based under-calls of heterozygous positions in regions of low read depth. The slightly higher level of other disagreements (0.76%) seen in this analysis compared to that of the HM550 data (0.09%) is in line with the higher level of underlying genotype error rate of 0.7% for the HapMap data[18]. To refine this analysis further, we generated a set of 530,750 very high confidence reference genotypes comprising

©2008 Macmillan Publishers Limited. All rights reserved

Table 1 | Comparison of SNP calls made from sequence versus genotype data for the human genome (NA18507) and X chromosome (NA07340)

	ELAND			MAQ				
	X	Human	Human	X	Human	Human	Human	
	HM550 (13,604 SNPs) (%)	HM550 (552,710 SNPs) (%)	HM-All (3,699,592 SNPs) (%)	HM550 (13,604 SNPs) (%)	HM550 (552,710 SNPs) (%)	HM-All (3,699,592 SNPs) (%)	Combined (530,750 SNPs) (%)	(*n*)
Covered by sequence	99.77	99.60	99.24	99.91	99.74	99.29	99.78	529,589
Concordant calls	99.52	99.57	98.80	99.99	99.90	99.12	99.94	529,285
All disagreements	0.48	0.43	1.20	0.01	0.1	0.88	0.06	304
GT>Seq	0.48	0.35	0.46	0.01	0.03	0.15	0.02	130
Seq>GT	0	0.05	0.52	0	0.05	0.54	0.02	130
Other discordances	0	0.03	0.22	0	0.02	0.2	0.01	44

SNP panels referred to are HM550 (Illumina Infinium HumanHap550 BeadChip) and HM-All (complete data from phase 1 and phase 2 of the International HapMap Project). 'Combined' is a set of concordant genotypes from both sets (HM550 and HM-All; see text). GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse (that is, a heterozygous sequencing SNP call where there is a homozygous genotyping call). Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

concordant calls in both the HM550 and HM-All genotype data sets. Comparing the results of the MAQ analysis to this high confidence set (see Table 1), we found 130 heterozygote under-calls GT>Seq (that is, a false-negative rate of 0.025%). There were also 130 heterozygote over-calls Seq>GT, but most of these are probably genotype errors as 82 have a nearby 'hidden' SNP and 3 have a nearby indel. A further 41 are tri-allelic loci, leaving at most 4 potential wrong calls by sequencing (that is, false-positive rate of 4 per 529,589 positions). Finally we selected a subset of novel SNP calls from the sequence data and tested them by genotyping. We found 96.1% agreement between sequence and genotype calls (Supplementary Table 8). However, the 47 disagreements included 10 correct sequencing calls (genotyping under-calls owing to hidden SNPs) and 7 sequencing under-calls. On this basis, therefore, the false-positive discovery rate for the one million novel SNPs is 2.5% (30 out of 1,206). For the entire data set of four million SNPs detected in this analysis, the false-positive and -negative rates both average <1%.

This genome from a Yoruba individual contains significantly more polymorphism than a genome of European descent. The autosomal heterozygosity (π) of NA18507 is 9.94×10^{-4} (1 SNP per 1,006 bp), higher than previous values for Caucasians (7.6×10^{-4}, ref. 12). Heterozygosity in the pseudoautosomal region 1 (PAR1) is substantially higher (1.92×10^{-3}) than the autosomal value. PAR1 (2.7 Mb) at the tip of the short arm of chromosomes X and Y undergoes obligatory recombination in male meiosis, which is equivalent to 20× the autosome average. This illustrates a clear correlation between recombination and nucleotide diversity. By contrast, the 0.33-Mb PAR2 region has a much lower recombination rate than PAR1; we observed that heterozygosity in PAR2 is identical to that of the autosomes in NA18507. Heterozygosity in coding regions is lower (0.54×10^{-3}) than the total autosome average, consistent with the model that some coding changes are deleterious and are lost as the result of natural selection[22]. Nevertheless, the 26,140 coding SNPs (Supplementary Fig. 15) include 5,361 non-conservative amino acid substitutions plus 153 premature termination codons (Supplementary Table 9), many of which are expected to affect protein function.

a

Call	ELAND SNPs (*n*)	ELAND In dbSNP (%)	MAQ SNPs (*n*)	MAQ In dbSNP (%)
Homozygote	1,417,320	90.1	1,503,420	90.8
Heterozygote	2,411,022	63.9	2,635,776	63.8
All	3,828,342	73.6	4,139,196	73.6

b

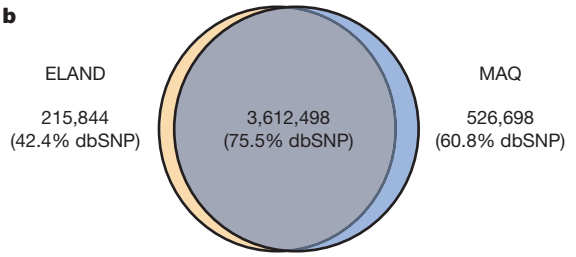


Figure 3 | SNPs identified in the human genome sequence of NA18507. **a**, Number of SNPs detected by class and percentage in dbSNP (release 128). Results from ELAND and MAQ alignments are reported separately. **b**, Analysis of SNPs detected in each analysis reveals extensive overlap. The percentage of NA18507 SNP calls that match previous entries in dbSNP is lower than that of our X chromosome study (see Supplementary Fig. 6). We expect this because individual NA07340 (from the X chromosome study) was also previously used for discovery and submission of SNPs to dbSNP during the HapMap project, in contrast to NA18507.

We performed a genome-wide survey of structural variation in this individual and found excellent correlation with variants that had been reported in previous studies, as well as detecting many new variants. We found 0.4 million short indels (1–16 bp; Supplementary Fig. 16), most of which are length polymorphisms in homopolymeric tracts of A or T. Half of these events are corroborated by entries in dbSNP, and 95 of 100 examined were present in amplicons sequenced from this individual in ENCODE regions, confirming the high specificity of this method of short indel detection. For larger structural variants (detected by anomalously spaced paired ends) we found that some were detected by both long and short insert data sets (Supplementary Fig. 17a), but most were unique to one or other data set. We observed two reasons for this: first, small events (<400 bp) are within the normal size variance of the long insert data; second, nearby repetitive structures can prevent unique alignment of read pairs (see Supplementary Fig. 17b, c). In some cases, the high resolution of the short insert data permits detection of additional complexity in a structural rearrangement that is not revealed by the long insert data. For example, where the long insert data indicate a 1.3-kb deletion in NA18507 relative to the reference, the short insert data reveal an inversion accompanied by deletions at both breakpoints (Fig. 4). We carried out *de novo* assembly of reads in this region and constructed a single contig that defines the exact structure of the rearrangement (data not shown).

We discovered 5,704 structural variants ranging from 50 bp to >35 kb where there is sequence absent from the genome of NA18507 compared to the reference genome. We observed a steadily decreasing number of events of this type with increasing size, except for two peaks (Supplementary Fig. 18). Most of the events represented by the large peak at 300–350 bp contain a sequence of the AluY family. This is consistent with insertion of short interspersed nuclear elements (SINEs) that are present in the reference genome but missing from the genome of NA18507. Similarly, the second, smaller peak at 6–7 kb is the consequence of insertion of the long interspersed nuclear element (LINE) L1 *Homo sapiens* (L1Hs) in

©2008 Macmillan Publishers Limited. All rights reserved

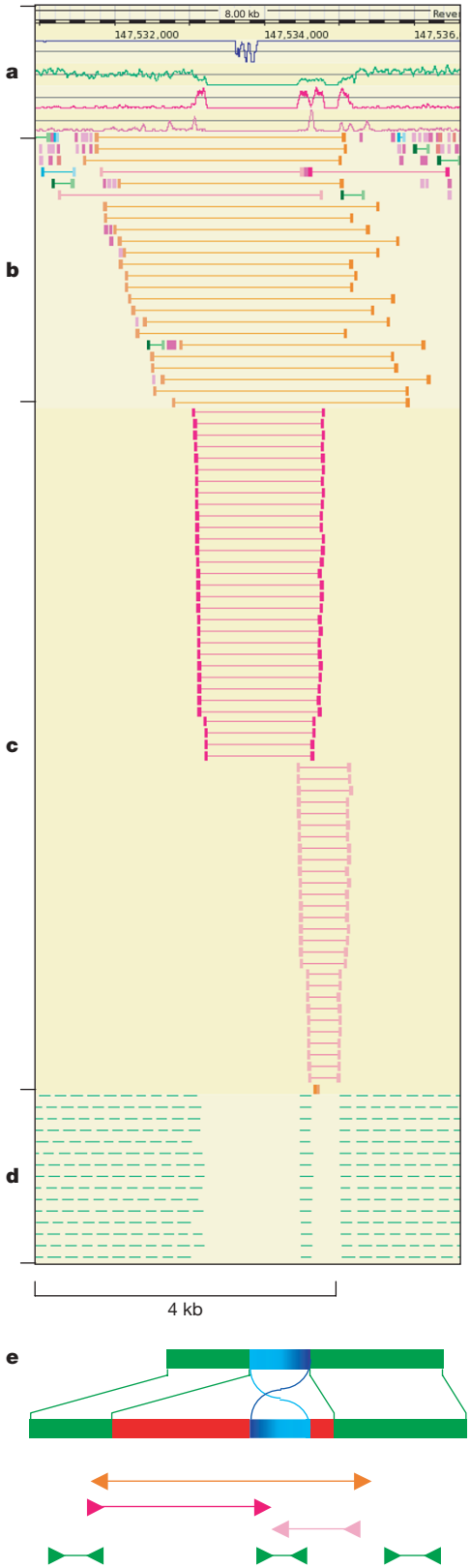


Figure 4 | Homozygous complex rearrangement detected by anomalous paired reads. The rearrangement involves an inversion of 369 bp (blue–turquoise bar in the schematic diagram) flanked by deletions (red bars) of 1,206 and 164 bp, respectively, at the left- and right-hand breakpoints. **a**, Summary tracks in the Resembl browser, denoting scale, simulated alignability of reads to reference (blue plot), actual aligned depth of coverage by NA18507 reads (green plot), density of anomalous reads indicating structural variants (red plot; peaks denote 'hotspots') and density of singleton reads (pink plot). **b**, Anomalous long-insert read pairs (orange lines denote DNA fragment; blocks at either end denote each read); the data indicate loss of ~1.3 kb in NA18507 relative to the reference. **c**, Anomalous short-insert pairs of two types (red and pink) indicate an inverted sequence flanked by two deletions. **d**, Normal short-insert read-pair alignments (each green line denotes the extent of the reference that is covered by the short fragment, including the two reads). **e**, The schematic diagram depicts the arrangement of normal and anomalous read pairs relative to the rearrangement. Top line, structure of NA18507; second line, structure of reference sequence. Green bars denote sequence that is collinear in the reference and NA18507 genomes. The turquoise–blue bar illustrates the inverted segment. Red bars indicate the sequences present in the reference but absent in NA18507. Arrows denote orientation of reads when aligned to the reference. The display in **a**–**d** is a composite of screen shots of the same window, overlapped for display purposes.

many cases. We found good correspondence between our results and the data of ref. 23, which reported 148 deletions of <100 kb in this individual on the basis of abnormal fosmid paired-end spacing. We found supporting evidence for 111 of these events. We detected a further 2,345 indels in the range 60–160 bp which are sequences present in the genome of NA18507 and absent from the reference genome (Supplementary Fig. 19). One example is shown in Supplementary Fig. 20. The 'singleton' reads on either side of the event, which have partners that do not align to the reference, form part of a *de novo* assembly that precisely defines the novel sequence and breakpoint (Supplementary Fig. 21).

Effect of sequence depth on coverage and accuracy

We investigated the impact of varying input read depth (and hence cost) on SNP calling using chromosome 2 as a model. SNP discovery increases with increasing depth: essentially all homozygous positions are detected at 15×, whereas heterozygous positions accumulate more gradually to 33× (Fig. 5a). This effect is influenced by the stringency of the SNP caller. To call each allele in this analysis we required the equivalent of two high-quality Q30 bases (as opposed to three used in full depth analyses). Homozygotes could be detected at read depth of 2× or higher, whereas heterozygote detection required at least double this depth for sampling of both alleles. Missing calls (not covered by sequence) and discordances between sequence-based SNP calls and genotype loci (mostly under-calls of heterozygotes due to low depth) progressively reduced with increasing depth (Fig. 5b). We observed very few other types of discordance at any depth; many of these are genotyping errors as described above.

Concluding remarks

Reversible terminator chemistry is a defining feature of this sequencing approach, enabling each cycle to be driven to completion while minimizing misincorporation. The result is a system that generates accurate data at very high throughput and low cost. We determined an accurate whole human genome sequence in 8 weeks to an average depth of ~40×. We built a consensus sequence, optimized methods for analysis, assessed accuracy and characterized the genetic variation of this individual in detail.

We assessed accuracy relative to genotype data over the entire fraction of the human sequence where SNP calling was possible (>90%). We established very low false-positive and -negative rates for the ~four million SNPs detected (<1% over-calls and under-calls). This compares favourably with previous individual genome analyses which reported a 24% under-calling of heterozygous positions[2,7].

Paired reads were very powerful in all areas of the analysis. They provided very accurate read alignment and thus improved the accuracy and coverage of consensus sequence and SNP calling. They were essential for developing our short indel caller, and for detecting larger structural variants. Our short-insert paired-read data set introduced a new level of resolution in structural variation detection, revealing thousands of variants in a size range not characterized previously. In

©2008 Macmillan Publishers Limited. All rights reserved

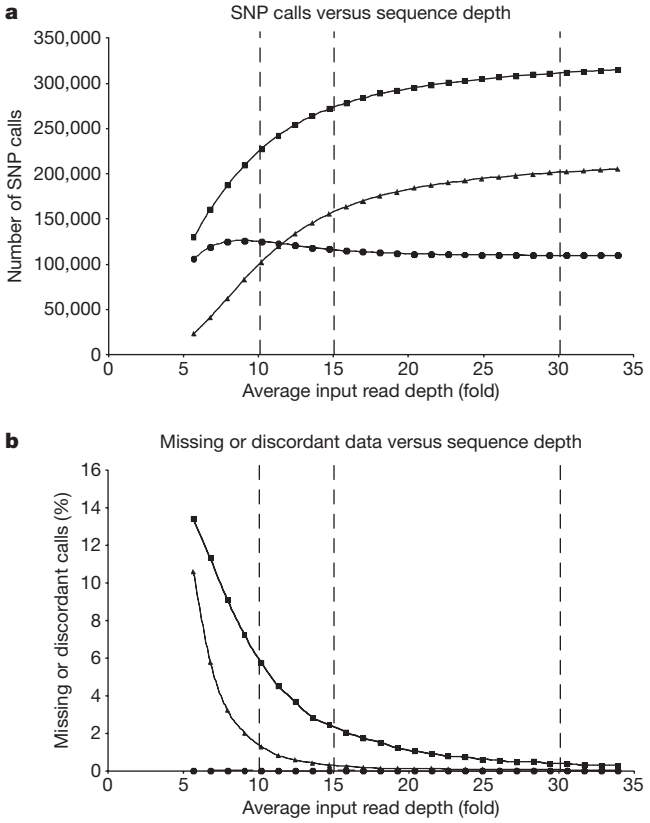


Figure 5 | Effect of sequence depth on coverage and accuracy of human genome sequencing. ELAND alignments were used for this analysis. **a**, Accumulation of sequence-based SNP calls, including all SNPs (squares), heterozygous SNPs (triangles) and homozygous SNPs (circles) with increasing input read depth. **b**, Decrease in genotype positions not covered by sequence (squares), heterozygote under-calls in sequence data relative to genotype data (triangles) and discordant SNP calls compared to genotypes (circles) with increasing input read depth. Vertical dotted lines indicate various input read depths (10×, 15×, 30× haploid genome).

some cases we determined the exact sequence of structural variants by *de novo* assembly from the same paired-read data set. Interpreting events that are embedded in repetitive sequence tracts will require further work.

Massively parallel sequencing technology makes it feasible to consider whole human genome sequencing as a clinical tool in the near future. Characterizing multiple individual genomes will enable us to unravel the complexities of human variation in cancer and other diseases and will pave the way for the use of personal genome sequences in medicine and healthcare. Accuracy of personal genetic information from sequence will be critical for life-changing decisions.

In addition to the large-scale genomic projects exemplified by the present study and others[15,24–26], the system described here is being used to explore biological phenomena in unprecedented detail, including transcriptional activity, mechanisms of gene regulation and epigenetic modification of DNA and chromatin[27–32]. In the future, DNA sequencing will be the central tool for unravelling how genetic information is used in living processes.

METHODS SUMMARY

DNA and sequencing. DNA samples (NA07340 and NA18507) and cell line (GM07340) were obtained from Coriell Repositories. DNA samples were genotyped on the HM550 array and the results compared to publicly available data to confirm their identity before use. Methods for DNA manipulation, including sample preparation, formation of single-molecule arrays, cluster growth and sequencing were all developed during this study and formed the basis for the standard protocols now available from Illumina, Inc. All sequencing was performed on Illumina GA1s equipped with a one-megapixel camera. All purity filtered read data are available for download from the Short Read Archive at NCBI or from the European Short Read Archive (ERA) at the EBI.

Analysis software. Image analysis software and the ELAND aligner are provided as part of the Genome Analyzer analysis software. SNP and structural variant detectors will be available as future upgrades of the analysis pipeline. The Resembl extension to Ensembl is available on request. The MAQ (Mapping and Assembly with Qualities) aligner is freely available for download from http://maq.sourceforge.net.

Data access. Sequence data for NA18507 are freely available from the NCBI short read archive, accession SRA000271 (ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000271). X chromosome data are freely available from ERA, accession ERA000035. Links to Resembl displays for chromosome X and human data, plus information on other available data, are provided at http://www.illumina.com/HumanGenome.

See Supplementary Methods for a detailed Methods section.

Received 24 June; accepted 2 October 2008.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
2. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
3. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
4. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309,** 1728–1732 (2005).
5. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
6. Lundquist, P. M. *et al.* Parallel confocal detection of single molecules in real time. *Opt. Lett.* **33,** 1026–1028 (2008).
7. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
8. Milton, J. *et al.* Modified nucleotides. World Intellectual Property Organization WO/2004/018497 (2004).
9. Smith, G. P. *et al.* Modified polymerases for improved incorporation of nucleotide analogues. World Intellectual Property Organization WO/2005/024010 (2005).
10. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
11. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*. doi:10.1101/gr.078212.108 (25 September 2008).
12. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
13. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37,** 727–732 (2005).
14. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
15. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genet.* **40,** 722–729 (2008).
16. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
17. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
18. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
19. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
20. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
21. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
22. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
23. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
24. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5,** 183–188 (2008).
25. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39,** 1522–1527 (2007).
26. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4,** 931–936 (2007).
27. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
28. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
29. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).

©2008 Macmillan Publishers Limited. All rights reserved

30. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
31. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133,** 523–536 (2008).
32. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 585–587 (2008).
33. Fedurco, M., Romieu, A., Williams, S., Lawrence, I. & Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.* **34,** e22 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors acknowledge the advice of A. Williamson, T. Rink, S. Benkovic, J. Berriman, J. Todd, R. Waterston, S. Eletr, W. Jack, M. Cooper, T. Brown, C. Reece and R. Cook during this work; E. Margulies for assistance with data analysis; M. Shumway for assistance with data submission; and the contributions of the administrative and support staff at all the institutions. This research was supported in part by The Wellcome Trust (to H.L., A.Sc., K.W., N.P.C, B.N.L., J.R., M.E.H. and R.D.), the Biotechnology and Biological Sciences Research Council (BBSRC) (to S.B. and D.K.), the BBSRC Applied Genomics LINK Programme (to A.Sp. and C.L.B.) and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to N.F.H. and J.C.M.). S. Balasubramanian and D. Klenerman are inventors and founders of Solexa Ltd.

Author Information Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R.B. (dbentley@illumina.com).

David R. Bentley[1], Shankar Balasubramanian[2], Harold P. Swerdlow[1]†, Geoffrey P. Smith[1], John Milton[1]†, Clive G. Brown[1]†, Kevin P. Hall[1], Dirk J. Evers[1], Colin L. Barnes[1,2], Helen R. Bignell[1], Jonathan M. Boutell[1], Jason Bryant[1], Richard J. Carter[1], R. Keira Cheetham[1], Anthony J. Cox[1], Darren J. Ellis[1], Michael R. Flatbush[3], Niall A. Gormley[1], Sean J. Humphray[1], Leslie J. Irving[1], Mirian S. Karbelashvili[3], Scott M. Kirk[3], Heng Li[4], Xiaohai Liu[1,2], Klaus S. Maisinger[1], Lisa J. Murray[1], Bojan Obradovic[1], Tobias Ost[1], Michael L. Parkinson[1], Mark R. Pratt[3], Isabelle M. J. Rasolonjatovo[1], Mark T. Reed[3], Roberto Rigatti[1], Chiara Rodighiero[1], Mark T. Ross[1], Andrea Sabot[1], Subramanian V. Sankar[3], Aylwyn Scally[4], Gary P. Schroth[3], Mark E. Smith[1], Vincent P. Smith[1], Anastassia Spiridou[1], Peta E. Torrance[1], Svilen S. Tzonev[3], Eric H. Vermaas[3], Klaudia Walter[4], Xiaolin Wu[1], Lu Zhang[3], Mohammed D. Alam[3], Carole Anastasi[1], Ify C. Aniebo[1], David M. D. Bailey[1], Iain R. Bancarz[1], Saibal Banerjee[3], Selena G. Barbour[1], Primo A. Baybayan[3], Vincent A. Benoit[1], Kevin F. Benson[1], Claire Bevis[1], Phillip J. Black[1], Asha Boodhun[1], Joe S. Brennan[1], John A. Bridgham[3], Rob C. Brown[1], Andrew A. Brown[1], Dale H. Buermann[3], Abass A. Bundu[1], James C. Burrows[3], Nigel P. Carter[4], Nestor Castillo[3], Maria Chiara E. Catenazzi[1], Simon Chang[3], R. Neil Cooley[1], Natasha R. Crake[1], Olubunmi O. Dada[1], Konstantinos D. Diakoumakos[1], Belen Dominguez-Fernandez[1], David J. Earnshaw[1,2], Ugonna C. Egbujor[1], David W. Elmore[3], Sergey S. Etchin[3], Mark R. Ewan[3], Milan Fedurco[5], Louise J. Fraser[1], Karin V. Fuentes Fajardo[1], W. Scott Furey[2], David George[3], Kimberley J. Gietzen[6], Colin P. Goddard[1], George S. Golda[3], Philip A. Granieri[3], David E. Green[1], David L. Gustafson[3], Nancy F. Hansen[7], Kevin Harnish[1], Christian D. Haudenschild[3], Narinder I. Heyer[1], Matthew M. Hims[1], Johnny T. Ho[3], Adrian M. Horgan[1], Katya Hoschler[1], Steve Hurwitz[3], Denis V. Ivanov[3], Maria Q. Johnson[3], Terena James[1], T. A. Huw Jones[1], Gyoung-Dong Kang[1], Tzvetana H. Kerelska[3], Alan D. Kersey[1], Irina Khrebtukova[3], Alex P. Kindwall[3], Zoya Kingsbury[1], Paula I. Kokko-Gonzales[1], Anil Kumar[1], Marc A. Laurent[6], Cynthia T. Lawley[6], Sarah E. Lee[1], Xavier Lee[3], Arnold K. Liao[3], Jennifer A. Loch[1], Mitch Lok[3], Shujun Luo[3], Radhika M. Mammen[1], John W. Martin[3], Patrick G. McCauley[1], Paul McNitt[3], Parul Mehta[1], Keith W. Moon[3], Joe W. Mullens[3], Taksina Newington[1], Zemin Ning[4], Bee Ling Ng[4], Sonia M. Novo[1], Michael J. O'Neill[3], Mark A. Osborne[1,2], Andrew Osnowski[1], Omead Ostadan[3,6], Lambros L. Paraschos[3], Lea Pickering[1], Andrew C. Pike[1], Alger C. Pike[3], D. Chris Pinkard[3], Daniel P. Pliskin[3], Joe Podhasky[3], Victor J. Quijano[3], Come Raczy[1], Vicki H. Rae[1], Stephen R. Rawlings[1], Ana Chiva Rodriguez[1], Phyllida M. Roe[1], John Rogers[1], Maria C. Rogert Bacigalupo[1], Nikolai Romanov[1], Anthony Romieu[5], Rithy K. Roth[3], Natalie J. Rourke[1], Silke T. Ruediger[1], Eli Rusman[3], Raquel M. Sanches-Kuiper[1], Martin R. Schenker[1], Josefina M. Seoane[3], Richard J. Shaw[1], Mitch K. Shiver[3], Steven W. Short[3], Ning L. Sizto[3], Johannes P. Sluis[3], Melanie A. Smith[1], Jean Ernest Sohna Sohna[1], Eric J. Spence[3], Kim Stevens[1], Neil Sutton[1], Lukasz Szajkowski[1], Carolyn L. Tregidgo[1], Gerardo Turcatti[5], Stephanie vandeVondele[1], Yuli Verhovsky[3], Selene M. Virk[3], Suzanne Wakelin[3], Gregory C. Walcott[3], Jingwen Wang[1], Graham J. Worsley[1], Juying Yan[3], Ling Yau[3], Mike Zuerlein[3], Jane Rogers[4]†, James C. Mullikin[7], Matthew E. Hurles[4], Nick J. McCooke[1]†, John S. West[3], Frank L. Oaks[3], Peter L. Lundberg[3], David Klenerman[2], Richard Durbin[4] & Anthony J. Smith[1]

[1]Illumina Cambridge Ltd. (Formerly Solexa Ltd), Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK. [2]Department of Chemistry, University of Cambridge, The University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, UK. [3]Illumina Hayward (Formerly Solexa Inc.), 23851 Industrial Boulevard, Hayward, California 94343, USA. [4]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [5]Manteia Predictive Medicine S.A. Zone Industrielle, Coinsins, CH-1267, Switzerland. [6]Illumina Inc., Corporate Headquarters, 9883 Towne Centre Drive, San Diego, California 92121, USA. [7]National Human Genome Research Institute, National Institutes of Health, 41 Center Drive, MSC 2132, 9000 Rockville Pike, Bethesda, Maryland 20892-2132, USA. †Present addresses: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK (H.P.S.); Oxford Nanopore Technologies, Begbroke Science Park, Sandy Lane, Kidlington OX5 1PF, UK (J.M., C.G.B.); BBSRC Genome Analysis Centre, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK (J.R.); Pronota, NV, VIB Bio-Incubator, Technologiepark 4, B-9052 Zwijnaarde/Ghent, Belgium (N.J.M.).

©2008 Macmillan Publishers Limited. All rights reserved

Experimental methods

Preparation of DNA

DNA for a bacterial artificial chromosome (bCX98J21) clone was prepared from a single colony by standard alkaline extraction method and purified by equilibrium density gradient fractionation using standard procedures. X chromosome DNA was prepared by flow-sorting X chromosomes from a culture of human lymphoblastoid cell line GM07340. Cells were arrested in metaphase using demecolcine, harvested, treated with hypotonic solution, transferred to polyamine isolation buffer for lysis. Chromosomes were harvested, stained overnight with Hoechst and Chromomycin A3 (both Sigma) and analysed on a flow cytometer (MoFlo®, DAKO) equipped with two Innova 300 series lasers (Coherent). Approximately 1.3 million chromosomes were collected and treated with proteinase K and sodium lauroyl sarcosine followed by phenylmethylsulphonyl fluoride, before recovering DNA by precipitation[34].

Samples of human genomic DNA NA07340 and NA18507 were obtained from Coriell Repositories, Camden NJ, USA.

Construction of short insert single- and paired-end libraries

Purified DNA (50 ng – 5 ug) was fragmented (6 minutes at 32 psi) through a disposable nebulizer (Invitrogen) and purified on a single Qiaquick column (Qiagen). Recovered DNA was repaired using a cocktail of T4 DNA polymerase, DNA polymerase (large fragment) and T4 polynucleotide kinase (all enzymes from New England Biolabs) and then treated with Klenow fragment (3' to 5' exo$^-$) in the presence of 0.2 mM dATP to add a dAMP to the 3' ends of the fragments. A single adaptor was ligated to the ends of the DNA. Adapters for the preparation of single read libraries comprised the oligonucleotides: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG. The adapters for the preparation of paired read libraries comprised an alternative set of two oligonucleotides: 5'ACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond) and 5'-phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG. Ligation products in the

desired size range were gel-purified and amplified with 6-18 cycles of PCR using Phusion High-Fidelity system (New England Biolabs) and the appropriate primers: for single read libraries: 5'-CAAGCAGAAGACGGCATACGAGCTCTTCCGATC$_x$T and 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T; and for paired-read libraries: 5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC$_x$T and 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC$_x$T (x = phosphorothioate bond resistant to excision by 3'-5' exonucleases). These primers include the sequences that anneal to the complementary oligonucleotides bound to the flowcell surface, plus the sequencing primer sites. Samples were diluted to a concentration of 10 nM in 10 mM Tris pH 8.5 and 0.1% Tween 20 prior to cluster formation (see below).

Construction of long range paired end libraries

Purified DNA (10 ug) was fragmented (30 seconds at 7.5 psi) through a disposable nebulizer (Invitrogen). The recovered DNA was end repaired and the 3' ends labelled using a mixture of biotin-dNTPs (PerkinElmer), dNTPs and end repair enzymes (as described above). 2 kb DNA fragments were gel purified and circularised using T3 DNA ligase (Enzymatics). The linear DNA was removed using Plasmid Safe (Epicentre) and the circular DNA fragmented by nebulization (6 minutes at 32 psi). The recovered DNA was end repaired, dAMP tailed and the ends of the DNA ligated to the paired-read adapters (see above). Biotinylated fragments were purified on M280 streptavidin beads (Dynal). The DNA was amplified from the beads using Phusion High-Fidelity system (New England Biolabs) and the appropriate primers for paired-read libraries (as described above) (see fig 1d). Library fragments in the desired size range (400-600 bp) were gel purified and diluted for cluster formation.

Preparation of flowcells

Glass 8-channel flow cells (Silex Microsystems, Sweden) were thoroughly washed and then coated for 90 min at 20°C with 2% acrylamide containing ~3.9 mg/ml *N*-(5-bromoacetamidylpentyl) acrylamide, 0.85 mg/ml tetramethylethylenediamine (TEMED) and 0.48 mg/ml potassium persulfate ($K_2S_2O_8$). Flow cell channels were

rinsed thoroughly before further use. The coated surface was then functionalised by reaction for 1 hour at 50°C with a mixture containing 0.5 uM each of two priming oligonucleotides in 10 mM potassium phosphate buffer pH 7; either A and B for single read experiments or C and D for paired read experiments (see below for sequences). Grafted flow cells were stored in 5xSSC until required.

Cluster creation for single read experiments

Cluster creation was carried out using an Illumina Cluster Station. To obtain single stranded templates, adapted DNA was first denatured in NaOH (to a final concentration of 0.1M) and subsequently diluted in cold (4 °C) hybridisation buffer (5x SSC + 0.05 % Tween 20) to working concentrations of 2 – 4 pM, depending on the desired cluster density / tile. 85 ul of each sample were primed through each lane of a flowcell at 96°C (60 ul / min). The temperature was then slowly decreased to 40 °C at a rate of 0.05 °C/sec to enable annealing to complementary adapter oligonucleotides immobilised on the flowcell surface (oligo 'A': 5´-PS-TTTTTTTTTT-(diol)3-AATGATACGGCGACCACCGA-3´; oligo 'B': 5´-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGA-3´). Hybridised template strands were extended using Taq polymerase to generate their surface-bound complement. The samples were then denatured using formamide to remove the initial seeded template. The remaining single stranded copy was the starting point for cluster creation. Clusters were amplified under isothermal conditions at 60 °C for 35 cycles using Bst polymerase for extension and formamide for denaturation during each cycle (see fig 1b). Clusters were washed with storage buffer (5x SSC) and either stored at 4 °C or used directly.

Cluster creation for paired read experiments

Paired read flowcells contained the two oligonucleotides
oligo 'C': 5'-PS-TTTTTTTTTTAATGATACGGCGACCACCGAGAUCTACAC-3' (U = 2-deoxyuridine) and
oligo 'D': 5'-PS-TTTTTTTTTTCAAGCAGAAGACGGCATACGAGoxoAT-3', (Goxo = 8-oxoguanine) immobilised on the surface in a ratio C:D = 1:1. Other than the use of a paired-end specific library, cluster creation was the same as described above.

Processing of clusters for single read experiments

Linearisation of surface-immobilised complementary oligonucleotide 'A' was achieved by incubation with linearization mix (100 mM sodium periodate, 10 mM 3-aminopropan-1-ol, 20 mM Tris pH 8.0, 50 % v/v formamide) for 20 minutes at 20 °C followed by a water wash. All exposed 3'-OH termini of DNA, either from the extended template or unextended surface oligonucleotides were blocked by dideoxy chain termination using a terminal transferase and ddNTPs. Linearised and blocked clusters were denatured with 0.1M NaOH prior to hybridisation of the sequencing primer (see fig 1c). Processed flowcells were transferred to the Illumina Genome Analyser for sequencing.

Processing of clusters for paired read experiments

For read 1, linearisation of surface immobilised oligonucleotide 'C' to retain strand 1 of each cluster was achieved by incubation with USER enzyme (as shown above). After blocking, clusters were denatured with 0.1M NaOH prior to hybridisation of the read 1 specific sequencing primer (5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT -3'). Processed flowcells were transferred to the Illumina Genome Analyser for sequencing.

Following the successful completion of sequencing of read 1 on the Genome Analyser, flowcells remained mounted and were automatically prepared for read 2 *in situ* using the Illumina Paired End module (according to operating manual). Clusters were denatured with 0.1 M NaOH to remove the products of read 1. Clusters were 3'-dephosphorylated using T4 polynucleotide kinase, and the strand that had been linearised as part of the read 1 preparation was re-synthesized isothermally as previously described for cluster creation. Linearisation to remove strand 1 of the re-synthesised clusters was achieved by the excision of 8-oxoguanine from oligo 'D' using Fpg (formamidopyrimidine DNA glycosylase, New England Biolabs). Linearised and blocked clusters were denatured with 0.1M NaOH prior to hybridisation of the read 2 specific sequencing primer (5'-CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT -3')(see fig 1c).

Sequencing on the Genome Analyser.

All sequencing runs were performed as described in the Illumina Genome Analyser operating manual. Flowcells were sequenced using standard recipes (see User Guide) in order to generate 25 and 35 base single and paired reads. Typically a single read run producing 1-2 Gb of PF data required 72 hours; paired read runs required approximately 150 hours including the time taken for automated preparation of the template for the second read.

Data Analysis

Image analysis

The image analysis program (Firecrest) first identifies the position of the DNA clusters on the images taken from the first sequencing cycle. Each initial image was band-pass filtered to remove background fluorescence and large-scale structure on the image, as well as enhance the signal-to-noise. Cluster positions were identified from a search for local maxima on the filtered image. Because of the finite accuracy of the movements of the motion stage, images taken at different sequencing cycles have random translational offsets with respect to each other. Furthermore, images taken in different frequency channels have different optical paths and wavelengths and experience further, albeit smaller, translations and scale transformations. In order to correct for the image shifts and scalings, the cluster positions that were extracted from the four images taken in the first cycle were super-imposed to construct a 'reference image' containing all detected clusters. Transformations of the image coordinates to later cycles were then obtained from a cross-correlation of the images in later cycles to the reference image. In this way, we obtained a set of four intensity measurements for each cluster and sequencing cycle. These series of intensities for each cluster are analogous to the intensity traces from Sanger sequencing. In addition to estimates of the intensities, the image analysis also extracts an estimate of the local noise or image background dispersion around the cluster for each image.

Base calling

The signals detected for the four different dye-labelled dNTPs are not independent, as the emission spectra of their dyes and the transmission and detection frequency windows may overlap between nucleotides. The relative intensities and cross-talk are described by a frequency cross-talk matrix, which characterises the intensity response

of the system to each nucleotide. We developed a method to auto-calibrate this matrix from the intensity traces and apply a correction to the extracted intensities. Because the rates of phasing and prephasing were small and consistent, the resulting intensity dissipation into different frequency cycles was also small and the accumulated loss roughly linear. Therefore we estimated phasing and prephasing rates by measuring the build-up of correlated signal between different cycles over early sequencing cycles. From these rates we derived the expected correlation of signals for each cycle and de-correlated them. The end result of these computations was a set of matrix-corrected, phasing-corrected intensity values for each cluster, from which we took whichever of the four channels gave the highest value at a given cycle to be the base call for that cycle.

Purity filtering

In order to discriminate between good reads without errors and reads derived from mixed clusters that overlap their nearest neighbours, we defined a measure of signal purity at a given cycle by taking the brightest of the four corrected intensities at that cycle as a fraction of the sum of the brightest and next brightest intensities. We discarded all reads whose corrected brightest intensity in any of the first 12 sequencing cycles was less than 60% of the sum of brightest intensity and the next brightest. We find that this criterion provides reasonable discrimination between good and bad data. Depending on the loading density, we typically kept between 50% and 70% of the raw reads. All figures quoted for accuracy and yield per flowcell refer only to this purity-filtered subset of the raw data.

Quality scoring of bases

The base caller (Bustard) provides a first estimate of the uncertainty of the base call. This is computed by propagating the noise estimates from the image analysis and integrating the resulting likelihood functions to obtain probability estimates for each of the four possible base calls. The probability estimates are transformed to scores by first converting to a log-odds scale via the formula $Q=10\log_{10}(p(X)/(1-p(X))$, where $p(X)$ for A,C,G,T is the estimated probability of the base call being X, and then rounding to integers. This scoring scheme can be thought of as a generalization of the scoring scheme made popular by the Phred base caller[10] ($Q=-10\log10(p_{error})$, where p_{error} is the probability of an incorrect base), in that the highest of the four scores - that of the called base - is

asymptotic to the Phred score, but the scheme also enables meaningful integer scores to be assigned to the three non-called bases.

This initial base quality estimate is refined by an implementation of the Phred algorithm, taking as predictors the initial confidence score for the called base together with the sequencing cycle, the purity of the called base and the minimum purity over the first 12 bases of the read. The sequence data for each flowcell lane in an experiment can be used as a training set for the same lane, with the alignment to the reference being used to determine whether or not a base is correct. Since a lane may produce several hundred million base pairs, each lane contains enough data to serve as a reasonable training set. After this procedure, the observed error rate for each base closely matches the error rate implied by the adjusted quality score (fig S3).

This "auto-calibration" procedure was used to estimate the base quality for the X chromosome and Yoruba genome analyses. However, we note that this method will underestimate the true quality of the bases since genuine differences between the sample DNA and the reference are treated as errors by the algorithm. One way of ameliorating this effect would be to exclude from the auto-calibration all reads that overlap positions known to be highly variable in the genome being sequenced – in human DNA, dbSNP positions would be obvious candidates. An alternative method is to allow one lane of the flowcell to serve as a control by using it to sequence DNA from a sample whose reference is exactly known (we have used DNA from the bacteriophage PhiX 174RF1 for this purpose). The Phred calibration table obtained from this data is then used to recalibrate the quality scores of the bases coming from the unknown samples in the other lanes. By breaking the dependency on alignments, this method has the advantage of removing the need for any knowledge at all of the reference sequence of the sample(s) in the non-control lanes, albeit at a cost of a lane's worth of data.

PhageAlign alignment

PhageAlign is a program for exhaustive alignment of reads of length k against a known reference. Both strands of the reference sequence are split into overlapping k-mers which are then sorted lexicographically. The reads are also sorted and then compared to each genomic k-mer in turn allowing any number of substitution errors.

Commonality of prefixes between lexicographically adjacent prefixes is exploited to minimise the number of base to base comparisons required, however its exhaustive nature renders it too slow for high-throughput alignment of datasets where either the number of reads or the size of the reference genome is large.

The principal use of PhageAlign is to enable an accurate measure of raw read error by allowing even noisy reads to participate in the error rate calculation, provided the aligner is able to find a unique best match for them in the reference.

ELAND alignment

In order to remap the sequence reads to a large reference genome, we developed a fast short-read alignment program called ELAND. The first few bases (32 by default, or the entire read length for reads shorter than this) are aligned to the genome allowing up to two substitution differences to obtain a set of candidate match positions for each read in the reference. Sequence from each of these positions is then used to extend the candidate alignments along the full length of each read. Finally base quality values are used to choose, where possible, the most probable of the candidate alignments.

For paired reads, a set of candidate alignments is obtained for each of the two reads as described above. Read pairs having a unique alignment of each read are first used to determine the nominal strand orientation and insert size distribution of the sample then, on a second pass, this information is used to resolve repeats and determine the anomalously paired reads that are possible indicators of structural variation.

ELAND SNP calling

For allele calling, we used only read pairs having alignments that were correctly oriented and that indicated a template insert size of within 3 standard deviations of the sample median. We further required a paired alignment score ≥6 (indicating the quality of the paired mapping). The basecalls and their associated quality values were sent to a Bayesian allele caller, which produced one or two allele calls and scores for each position in the genome. At each position, the allele caller computes log10 p(observed bases | no “A”s are present) and similarly for C,G and T. The highest two scores are then normalized by subtracting the third highest, thus obtaining log-odds scores for the two most probable alleles for which an increment in score of 3

approximately corresponds to an increase in coverage of a single base of Phred quality 30. SNPs were called where a non-reference base allele was observed, the allele call score was ≥10, and the depth at this position was no greater than three times the chromosomal mean. For heterozygous calls, we additionally required both alleles to have an allele-call score ≥10 and the ratio of their scores to be ≤3. For the reduced depth analysis (fig 5) we required an allele-call score ≥6. We excluded SNP calls that were within 15 bp of an apparent small indel (detected as described below).

ELAND structural variant detection

Hierarchical clustering of anomalous readpairs was used to identify groupings of five or more readpairs that had a similar size and position. Read pairs were defined as anomalous if they had high-confidence alignments of each individual read that nevertheless were either incorrectly oriented or implied an insert size of at least 3 standard deviations outside the sample median. These groupings were combined with other information such as depth changes, alignability and gaps in expected coverage, and a ranking system was applied. Higher (positive) ranks were assigned where the event supporting evidence was seen; negative ranks where used for regions where it would be difficult to call variants, such as the centromere, or where contradictory evidence was seen.

Some structural variants were characterised using local *de novo* assembly. We took high quality (23 of the first 25 bases had ≥Q20) anomalous pairs, singletons and their non-aligning partner in the region of interest and attempted an assembly using Velvet[35]. Contigs were aligned back to the chromosome using BLAST, in order to look for discontinuities in the alignment, indicating breakpoints.

MAQ alignment

MAQ first searches for the ungapped match with lowest mismatch score, defined as the sum of qualities at mismatching bases. To speed up the alignment, MAQ only considers positions that have 2 or fewer mismatches in the first 28bp.
Sequences that fail to reach a mismatch score threshold but whose read pair is mapped are searched with a gapped alignment algorithm in the regions defined by the read pair. To evaluate the reliability of alignments, MAQ assigns each alignment a Phred-scaled quality score which measures the probability that the true alignment is not the

one found by MAQ. MAQ always reports a single alignment, and if a read can be aligned equally well to multiple positions, MAQ will randomly pick one position and give it a mapping quality zero.

MAQ fully utilizes the read-pair information of paired reads. It is able to use this information to correct wrong alignments, to add confidence to correct alignments, and to accurately map a read to repetitive sequences if its mate is confidently aligned. With paired-end reads, MAQ also finds short insertions/deletions (indels) from the gapped alignment described above.

Calculation of Mapped read depth and distribution

For NA07340-X: After aligning the data to the reference sequence for chromosome X, the depth of mapped reads was sampled at every 50th position. The distribution of this depth is shown by the 'all' histogram in fig 2a. Then all the positions from this sample were discarded where the reference is not unique on the scale of the read length (as determined by mapping the reference to itself). The distribution of these uniquely mapped reads is shown by the 'unique' histogram in fig 2a. Comparison with the Poisson distribution having the same mean shows that there is some extra variance or overdispersion relative to the theoretical minimum.

At each of the unique positions the GC content of the reference in a surrounding window of length twice the read length was calculated. This gives an estimate of the GC content of all the reads that could have overlapped that position. Then we binned the positions by GC content, and within each bin calculated the mean depth and the 10th and 90th centiles of both the depth and a Poisson distribution with the same mean. The resulting depth-GC variation is shown in fig 2b, where the lower x-axis represents the proportion of unique reference sequence corresponding to the GC content values on the upper axis. In this plot, overdispersion at a given GC content is indicated by the Poisson 10th and 90th centile lines lying inside the shaded area at that GC value.

MAQ SNP calling

MAQ produces a consensus genotype sequence from the alignment. The consensus sequence is inferred from a Bayesian statistical model and each consensus genotype is

associated with a Phred quality which measures the probability that the consensus genotype is incorrect. Potential SNPs are detected by comparing the consensus sequence to the reference and are further filtered by a set of predefined rules. These rules are:

i) discard SNPs within the 3bp flanking region around a potential indel;
ii) discard SNPs covered by three or fewer reads;
iii) discard SNPs covered by no read with a mapping quality higher than 60;
iv) in any 10bp window, if there are 3 or more SNPs, discard them all;
v) discard SNPs with consensus quality smaller than 20; and
vi) discard a SNP if a base with consensus quality lower than 20 occurs within 3bp on either side of the target SNP.

MAQ small indel detection

MAQ regards an indel is reliable, if at least three reads contain the exact indel (identical position and indel size). MAQ only keeps one most evident indel in any 10bp window because close indels may indicate alignment artefacts.

Genome-wide de novo assembly of unaligned reads

De novo assembly was performed using Velvet[35] and unaligned read pairs. The read pairs were duplicate filtered and then further quality filtered to obtain the best 11,067,318 reads, to allow for computer memory limitations. Version 0.5.05 of velvet was used with a hash length of 23, a coverage cutoff of 5 and a maximum insert length of 350. The contigs were then filtered to ensure a minimum length of 100 bases.

MAQ structural variant detection

Anomalous read pairs with a mapping quality of at least 20 were ordered first by start position, then two anomalous pairs were allocated to the same cluster if they were overlapping and if their end positions were not further apart than a given threshold (mean of the insert lengths plus three times their standard deviation). This procedure was followed by merging neighbouring overlapping clusters. To obtain a final set of putative deletions, these candidate clusters were filtered for the number of read pairs per cluster (at least 5), for distance between leftmost and rightmost forward reads (distance is less than a given threshold, i.e. mean of the insert lengths plus three times their standard deviation), similarly for distance between leftmost and rightmost

reverse reads and for deletion size (greater than a given threshold, i.e. mean of the insert lengths plus three times their standard deviation); also read depth and repeat structures were considered. Putative deletions greater than 100kb were removed.

Mapped read depth was used to infer copy-number variants between the sample (NA07340) and reference sequences. At any given sequence position the depth of reads aligned by Maq is expected to be Poisson distributed, with mean determined by the copy numbers of both the sample and the reference at that position, the mean overall depth of coverage, and any GC or other biases. We constructed a Hidden Markov model with hidden states representing copy number differences between sample and reference and an emission or observed variable representing the mapped depth accounting for GC content. (We use a negative binomial distribution rather than a Poisson for the latter, to account for the overdispersion evident in fig 2b.) Standard HMM methods (e.g. ref 36) were then applied to infer the most probable sequence of copy-number states in the sample given the depth data.

Resembl

Resembl is an extended version of Ensembl developed at Illumina as part of a collaborative project with EBI, Sanger and Imperial College under a UK government funded LINK framework grant. It allows storage, query and viewing of Illumina re-sequencing data in a genomic context. The human X chromosome (NA07340) and the whole human genome (NA18507) re-sequencing datasets from ELAND-based alignment data were loaded into Resembl databases and Resembl websites were used for interactive data mining and QC within Illumina.

The Resembl back-end database was designed to allow storage and retrieval of the large amounts of re-sequencing data in an efficient way. It supports paired-end alignments at high coverage, as well as per-base and summary-type data on coverage and alignability. Parsing scripts were written to pre-process sort.txt files from the build process and adapted to run on a Linux clustered environment. Import scripts were written for very large-scale data loading. A system of clustered indexes was developed to allow fast retrieval of data from multi billion record tables.

The website extensions to the Ensembl browser allow visualization of re-sequencing data in a genomic context and support easy navigation back to the raw data. Extensions were made to Ensembl's Karyotype, Map and Contig View, and a new view (called ReadView) was added for closer examination of reads. Resembl websites were set-up for browsing the X and whole human genome paired-end alignments, SNPs and structural variations, as well as coverage and alignability graphical plots. . Paired alignments are categorised into Regular and Anomalous (i.e. anomalous-gapped, misoriented, chimeras, singletons) and displayed in tracks, accordingly. Suitable colour coding is used to distinguish the different types of variation. The displayed paired alignments are also filtered according to the rules used during the pipeline analysis. Candidate structural variants are highlighted by peaks in a graphical plot within ContigView. SNPs produced from the pipeline are loaded in Resembl as a user track. All the extensions were seamlessly integrated within Ensembl, and were flexibly implemented as a plugin.

34 Ng, B.L. and Carter, N.P., Factors affecting flow karyotype resolution *Cytometry A* **Vol** (9), 1028 (2006).

35 Zerbino D.R. and Birney, E. , Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs." *Genome Res* **18**(5): 821-9. (2008).

36 Rabiner L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2): 257-286. (1989).

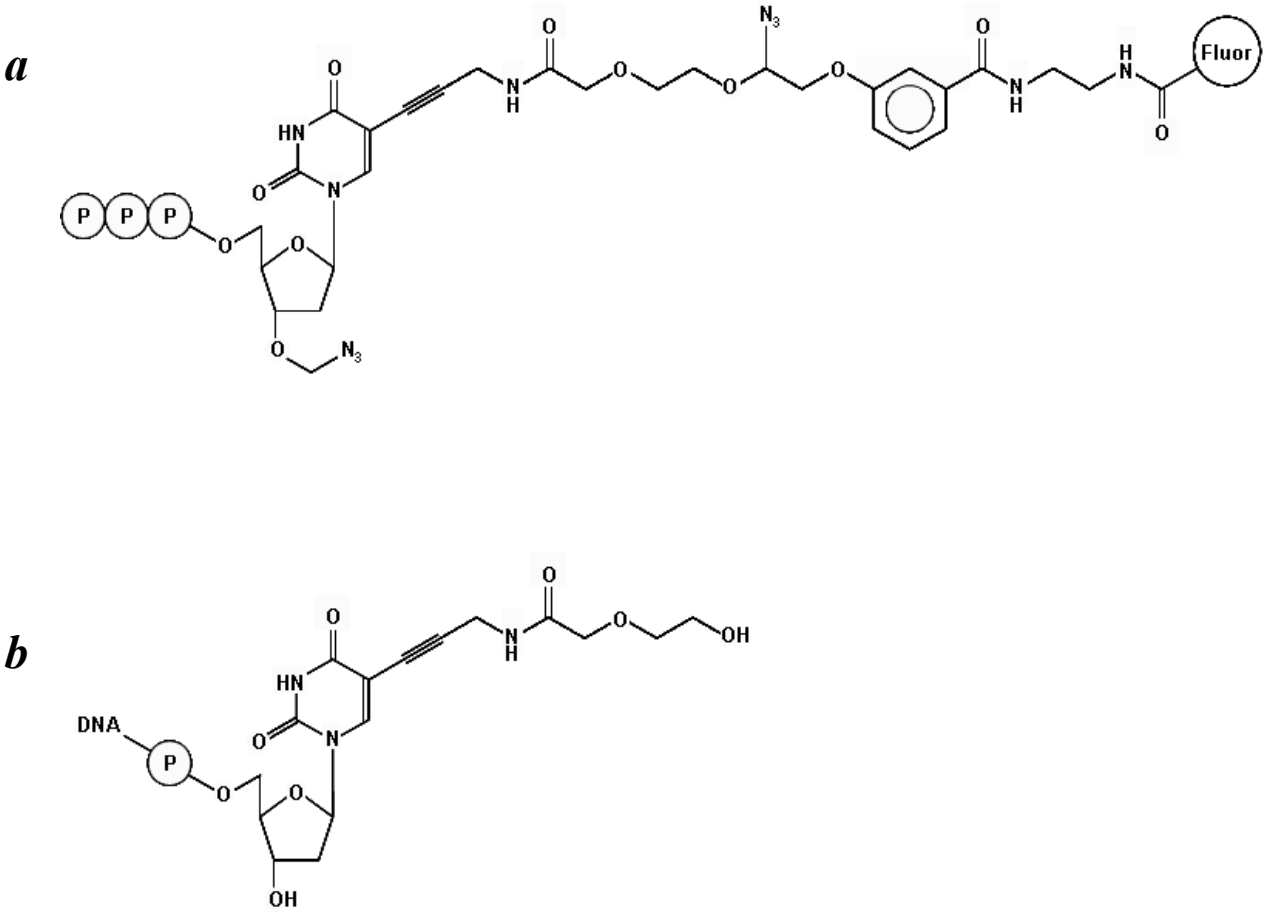


Figure S1. *a*. Structure of the reversible terminator 3'-*O*-azidomethyl 2'-deoxythymine triphosphate (T) labelled with a removable fluorophore. *b*. Structure of the incorporated nucleotide after removal of the fluorophore and terminator group. Each of the four nucleotides have an equivalent structure to the one shown here, except for the different base and a corresponding base-specific fluor.

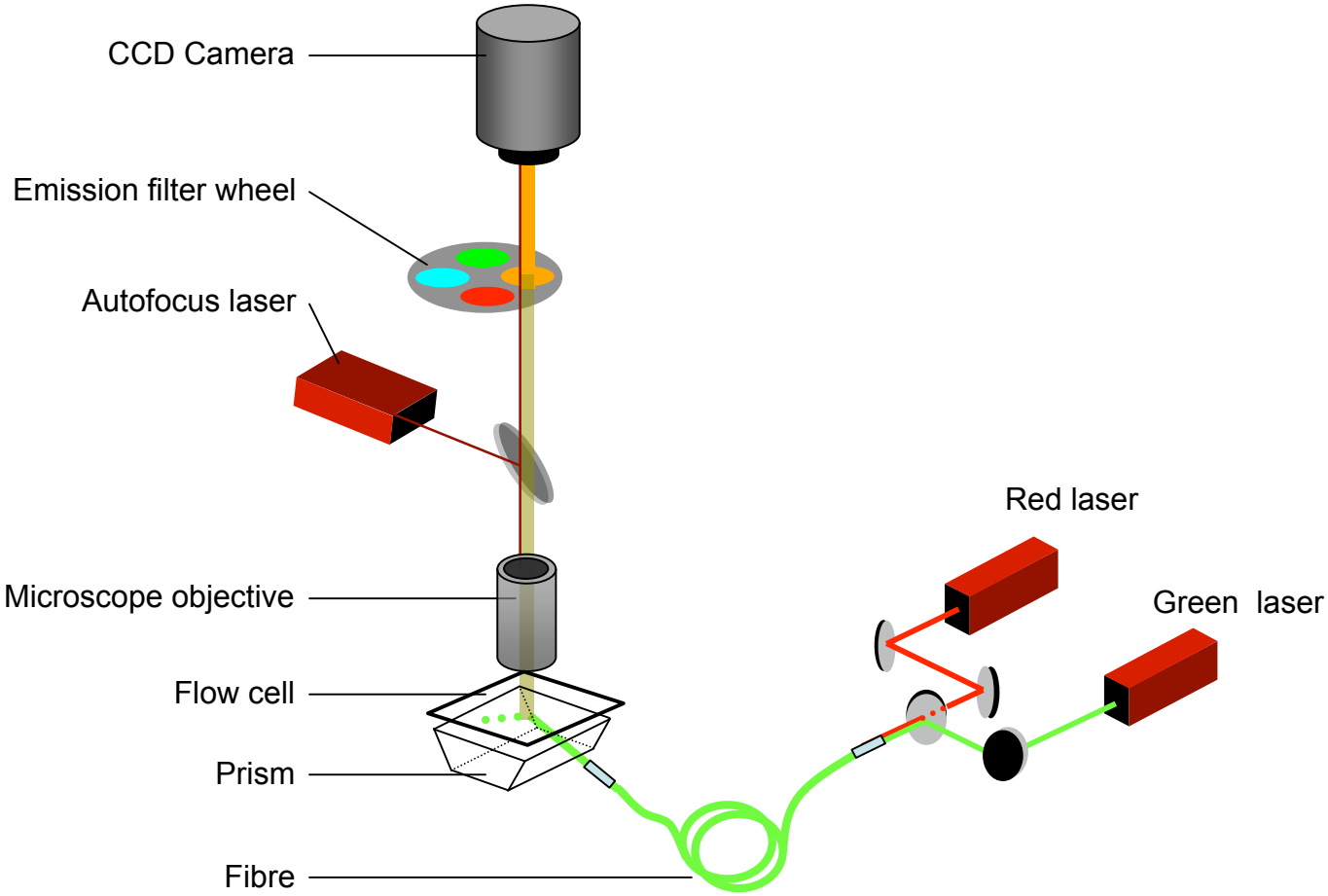


Figure S2*a*. Optical path of the Genome Analyser (GA1). Red (660 nm) and green (532 nm) lasers provide excitation beams that are directed along an optical fibre and through a prism which is in contact with the flow cell. Excitation of fluorescent nucleotides incorporated into DNA clusters on the inner surface of the flow cell (see fig S2b) leads to a base-specific emission that passes through an objective and a filter wheel and the signal is collected by a CCD camera. Autofocus utilises a third laser (635 nm) that is projected through the objective onto the flowcell (see fig S2c for details). Images are collected sequentially for each square 0.33 mm x 0.33 mm 'tile' on the flowcell surface (see fig S2d).

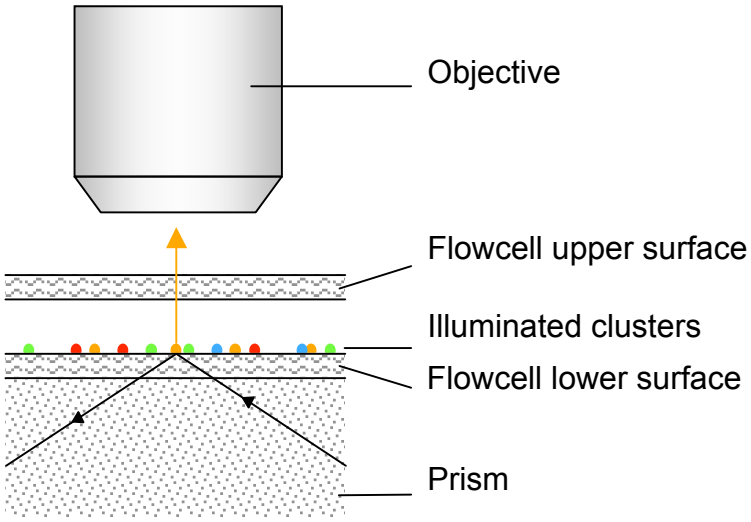


Figure S2b. Total internal reflection of the incident excitation beam at the glass-buffer interface generates an evanescent wave that excites the clusters on the surface. The fluorescence emission is captured by a custom made microscope objective, passed through a filter and is then projected onto a CCD. The evanescent wave excitation technique maximises the sensitivity of signal detection while minimising background noise.

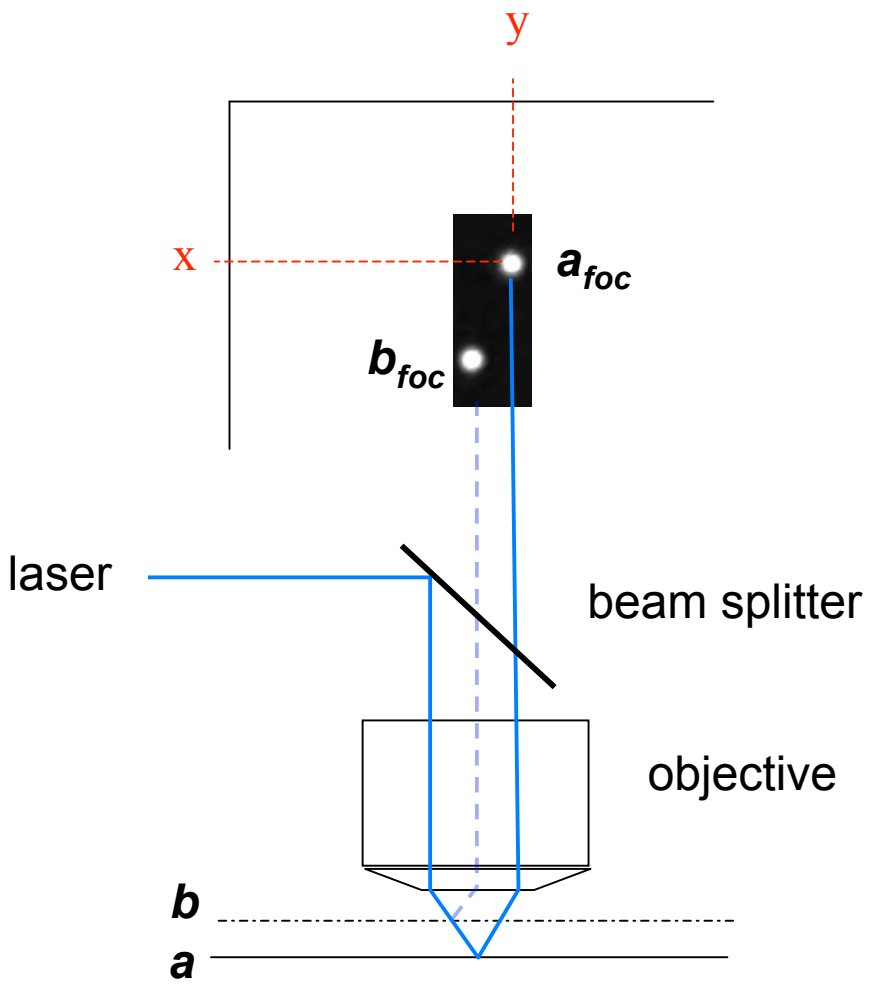


Figure S2c. Autofocus principle of operation. A reference laser beam (solid blue line) is deflected by a beam splitter along a path parallel to the optic axis of the microscope objective and is reflected from the surface of interest. Reflection from a focus plane ***a*** generates an autofocus spot $\boldsymbol{a}_{foc}$. Focus is checked manually at the beginning of the run, after which the autofocus set-up registers the correct (x,y) co-ordinates for the autofocus spot relative to the boundaries of the tile. If the surface of interest is out of focus in a subsequent cycle (as in ***b***), the laser beam follows a different path (blue dashed line) and the autofocus spot is displaced at $\boldsymbol{b}_{foc}$ Focus is automatically adjusted (z movement) as required to return the autofocus spot to the pre-registered correct position before image capture.

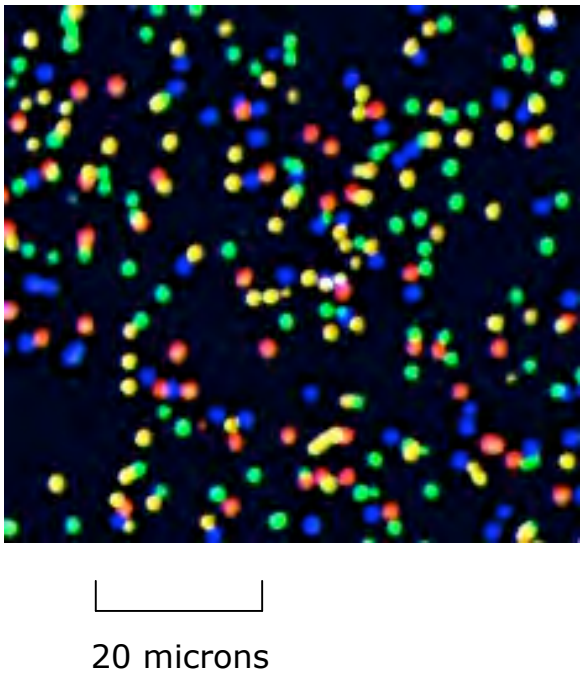


Figure S2d. Imaging clusters during the sequencing process. Part of the image of a tile with a low density of clusters is shown. Four images, one for each emission wavelength, have been artificially coloured and superimposed to show the four-colour detection system that provides the raw signal intensity for calling the individual bases.

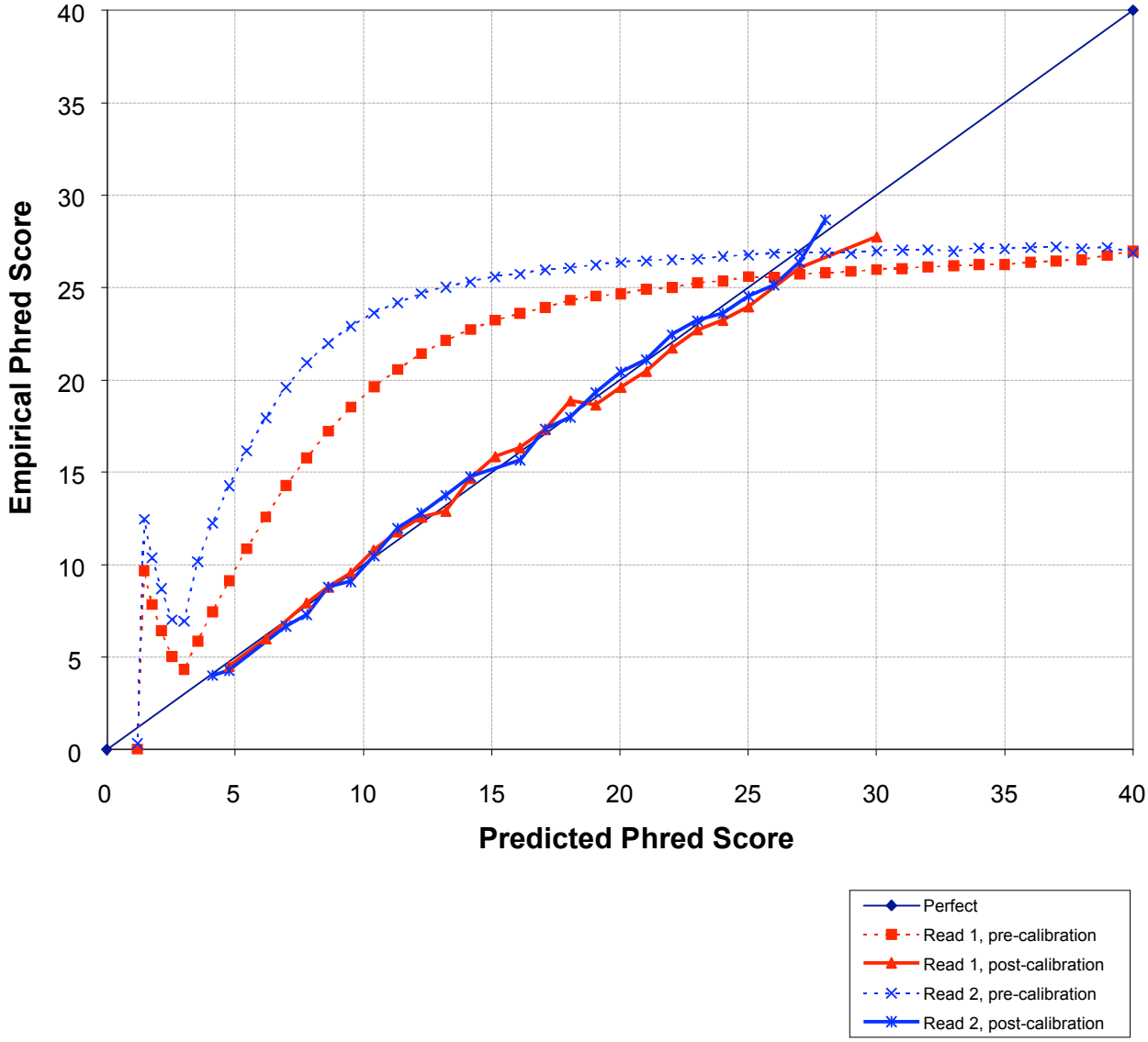


Figure S3. Effect of auto-calibration procedure on base quality scores (see supplementary methods for details). The data set is one lane of paired-read data from the NA18507 individual comprising 8,960,643 purity-filtered clusters, two 35-base reads being obtained from each cluster. Dotted lines show the original quality estimates obtained from the base caller for each read, the solid lines show the adjusted quality scores generated by the auto-calibration procedure. The probability estimates have been re-mapped from the log-odds scoring scheme in which they were originally expressed to the Phred scoring scheme of Q=-10log10(perror), where perror is the probability of an incorrect base. No attempt was made to exclude genuine differences between the sample DNA and the human reference sequence from the calibration procedure, so the calibrated quality scores likely represent a lower bound on the true quality of the data set.

a 080531_EAS192_0038_FC20GG5 – lane 5

	Lane Info		Tile Mean +/- SD for Lane							
Read	Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF clusters	& Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	5	297306	38689 +/- 3121	25740 +/- 1646	4211 +/- 867	87.22 +/- 11.11	66.67 +/- 2.40	91.29 +/- 0.20	53.86 +/- 1.44	0.57 +/- 0.07
2	5	297306	38689 +/- 3121	25740 +/- 1646	2274 +/- 504	79.47 +/- 9.69	66.67 +/- 2.40	89.80 +/- 0.40	49.65 +/- 1.78	0.69 +/- 0.09

b

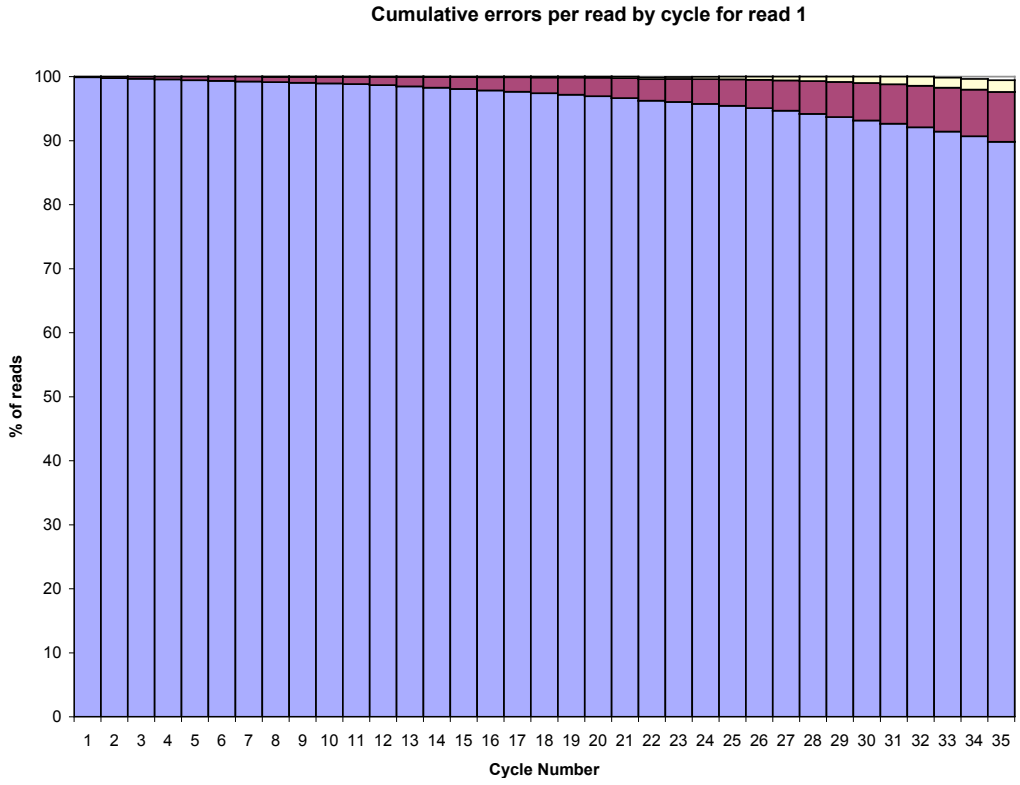


Figure S4. Sequencing summary data for the human BAC from one lane of a flow cell. ***a***, summary table for lane 5 (reads 1 and 2). Lane yield = all PF clusters per lane x read length (0.035 kbases). First cycle intensity (Int) is averaged over all PF reads and given in arbitrary units. % intensity after 20 cycles provides a measure of signal loss during the run. The % PF clusters indicates the fraction of raw clusters that provide PF sequence data after purity filtering (described in supplementary methods). In this example 91.29% of reads (read 1) aligned to the BAC reference. Note that the BAC sample also contains *E. coli* DNA; we found that a further 7% of reads aligned to the *E coli* reference. ***b***. Plot of cumulative errors per read by cycle for read 1 of the experiment summarised in a. The plot shows the % reads with 0 (blue), 1 (burgundy) or 2 (yellow) differences from the reference.

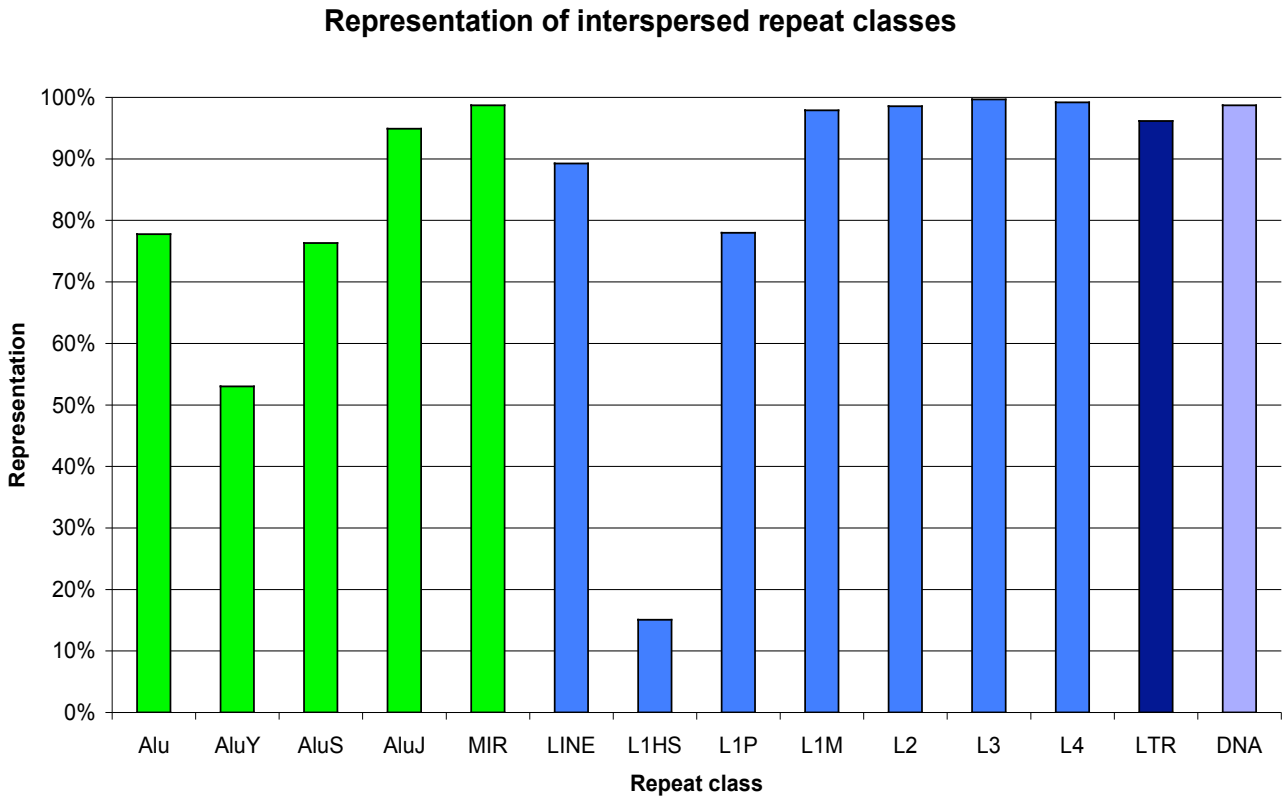


Figure S5. Representation of interspersed repeat classes in the ELAND alignment of the X chromosome dataset. Two-thirds of the low or zero sequence depth in the ELAND alignment is at interspersed repeats. The near-identical repeats arising from recently inserted retroposon elements L1Hs and members of the AluY subfamily of long and short interspersed nuclear element (SINE and LINE) classes are under-represented, while the rest are well covered with reads. Key: Alu, all Alu repeat sequences combined. Different subfamilies are as labelled: AluY - AluJ. LINE, all elements; subsets are defined by bars L1HS – L4. LTR, long terminal repeat retroposons. DNA, DNA transposons.

a

Call	ELAND SNPs n	ELAND In dbSNP %	MAQ SNPs n	MAQ In dbSNP %
Homozygote	36,402	91.7%	40,953	90.3%
Heterozygote	56,003	80.3%	63,614	77.6%
All	92,485	84.9%	104,567	82.6%

b

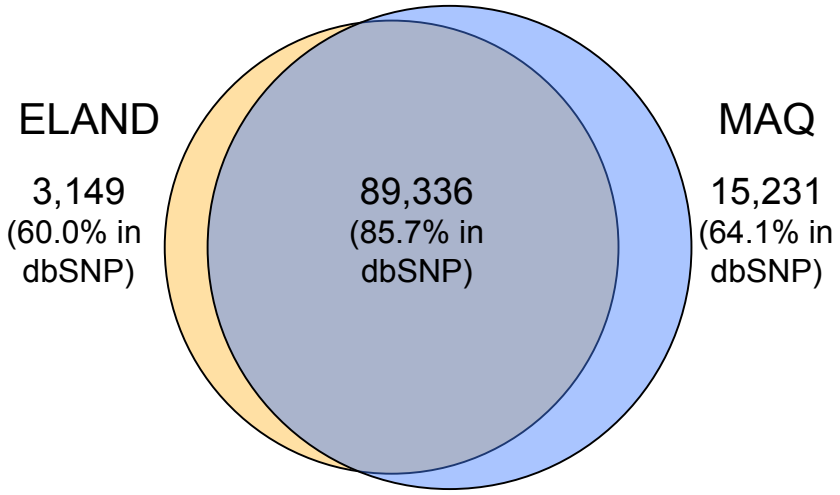


Figure S6. SNPs identified in the X chromosome sequence of NA07340. ***a***. Number of SNPs detected by class and % in dbSNP (release 128). Results from ELAND and MAQ alignments are reported separately. ***b.*** Overlap of SNPs detected in each analysis. See figure S6c (next page) for a more detailed breakdown of SNPs called by MAQ but not ELAND.

c

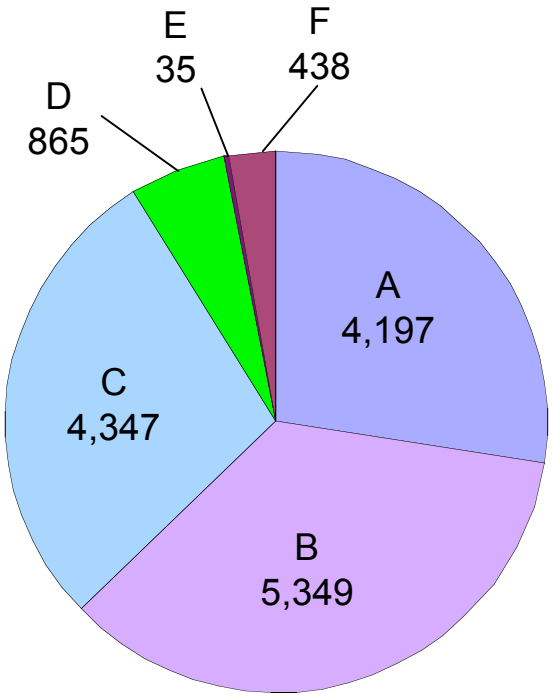


Figure S6c. Breakdown of NA07340 X-chromosome SNP calls present in MAQ but not the ELAND analysis. A: No call possible in ELAND analysis; B: MAQ calls a heterozygous position - both alleles are seen in the ELAND alignment but read depth is too low for one allele to reach the required score threshold; C: MAQ calls a heterozygous position - only the reference allele is seen in the ELAND alignment; D: Different SNP calls are made by ELAND and MAQ; E: MAQ calls a homozygous difference from the reference where ELAND calls a heterozygous position; F. MAQ calls a homozygous difference from the reference where ELAND calls the reference allele.

a

Input read depth x-fold	GT posns not covered % (of 13604)	hets undercalled % (of 4717 hets)	Discordant calls %
9	12.12	4.92	0
11	9.16	3.43	0
15	2.90	1.27	0
22	1.18	0.38	0
30	0.29	0.17	0
45	0.09	0.02	0

b

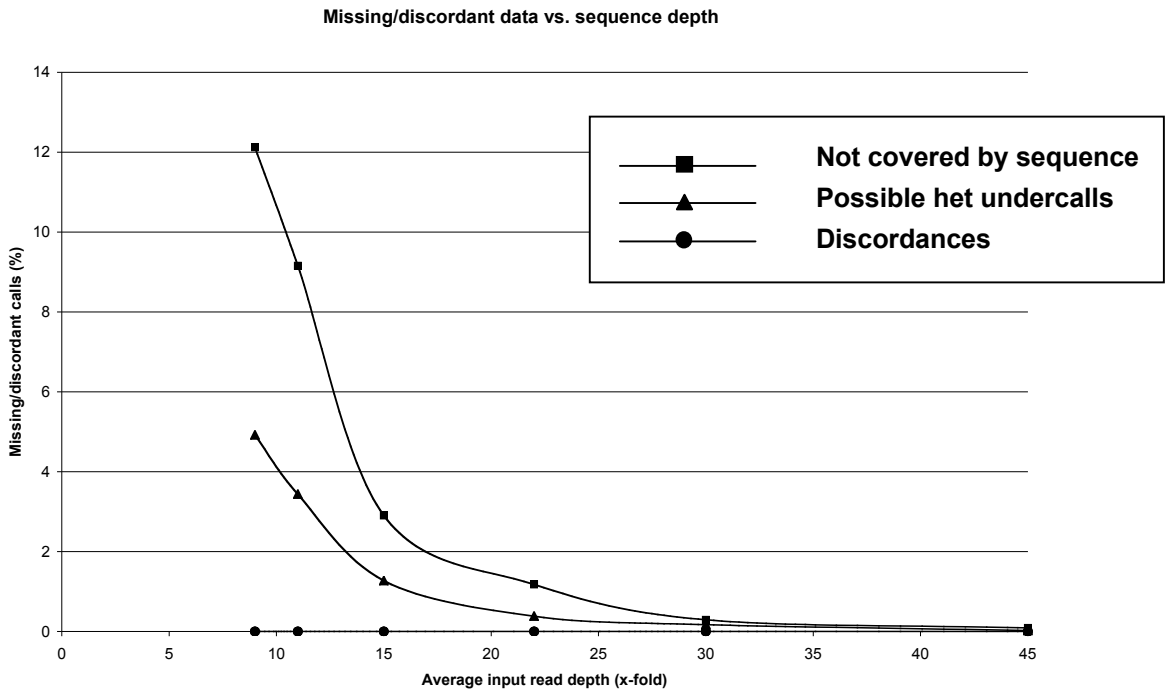


Figure S7. Analysis of SNP calls in X chromosome data for NA07340 using MAQ at different input read depths, compared to genotype data. ***a.*** Tabulated values and ***b.*** plot of percent coverage of 13604 genotype positions, percent of heterozygotes undercalled compared to genotype results, and discordant calls (all 0) as a function of Input read depth (from 9x to 45x).

a

	N	%dbSNP
Heterozygotes	5606	39.04%
Homozygotes	4142	53.96%
All indels	9747	45.38%

b

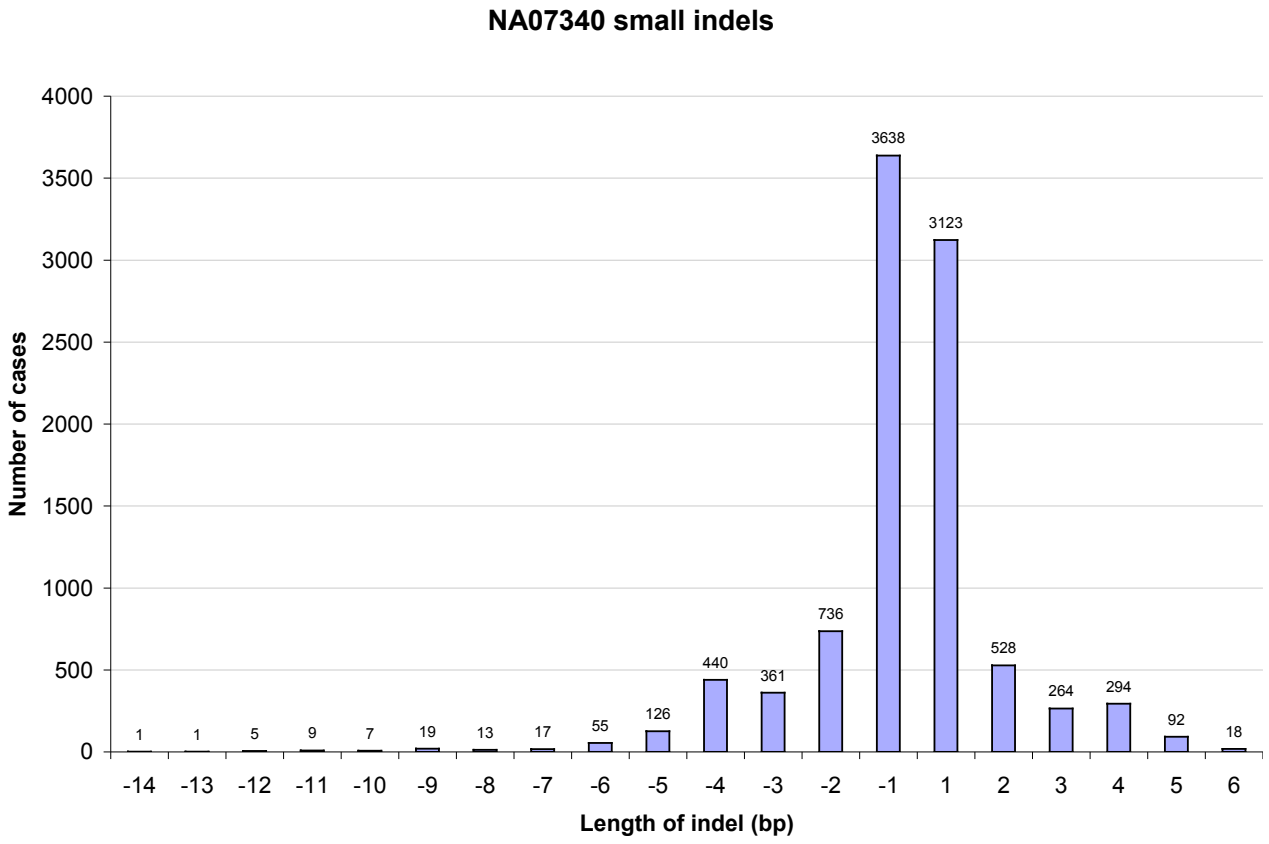


Figure S8. Analysis of short indel calls in X chromosome data for NA07340. ***a.*** total number of calls and fraction that match previous entries in dbSNP. ***b.*** Distribution of size in the 9747 indels. + and – values on the x axis correspond to presence or absence of bases in NA07340 relative to the reference sequence.

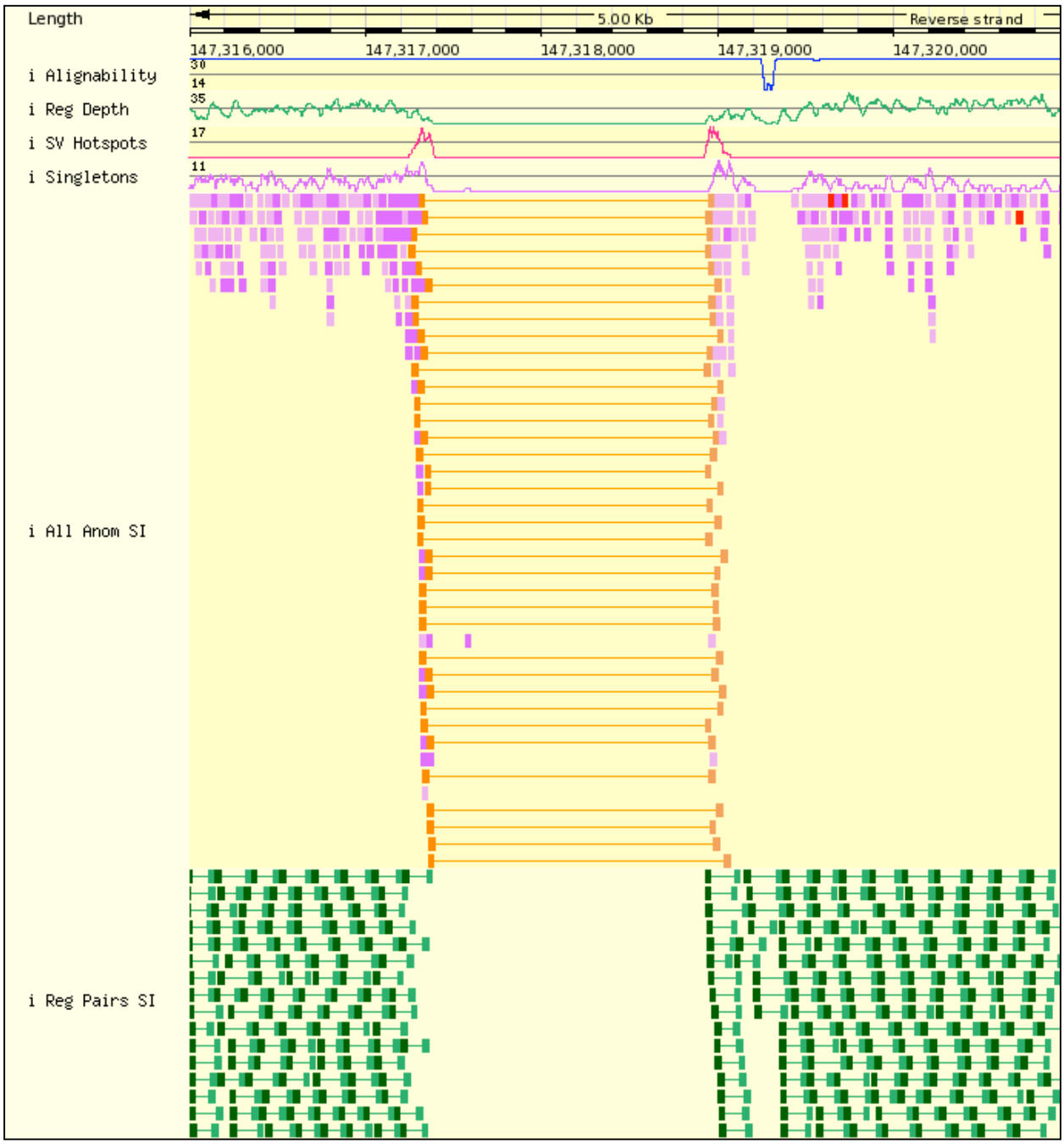


Figure S9a. homozygous 1.6 kb deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). Purple reads are singletons, i.e. reads where the other member of the pair does not align to the reference. Note also the alignability score (blue line) is high across the region. Read depth (green line) falls to zero within the deletion. The red line depicts SV hotspots, and provides a quick look-up for regions where there is a concentration of anomalous events detected on the basis of the paired read alignments. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

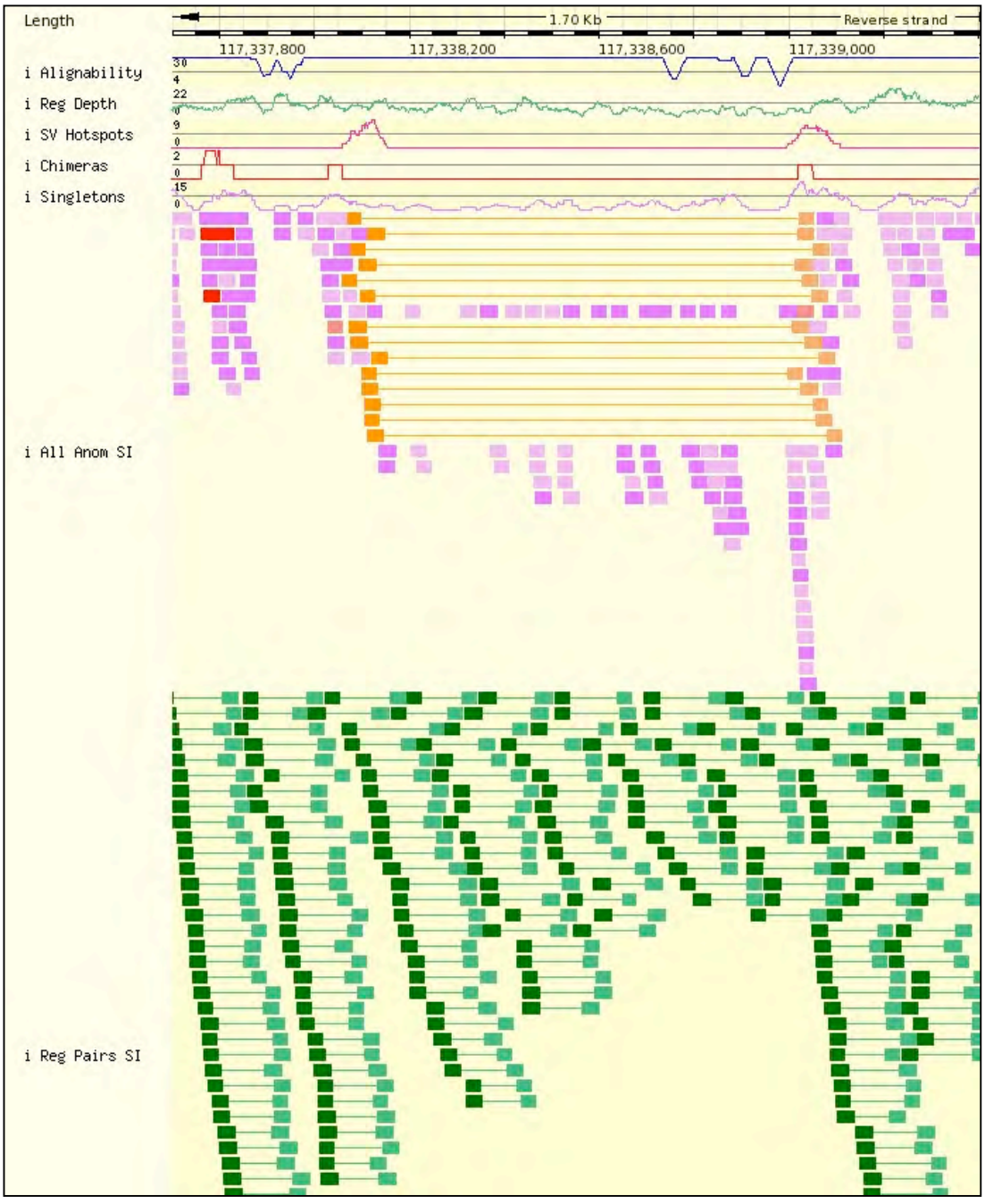


Figure S9b. Heterozygous 844 bp deletion in NA07340 relative to the reference, detected by anomalously spaced read pairs (orange) compared to regularly spaced read pairs (green). The presence of regularly spaced pairs across the entire region is indicative of heterozygosity at this locus. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

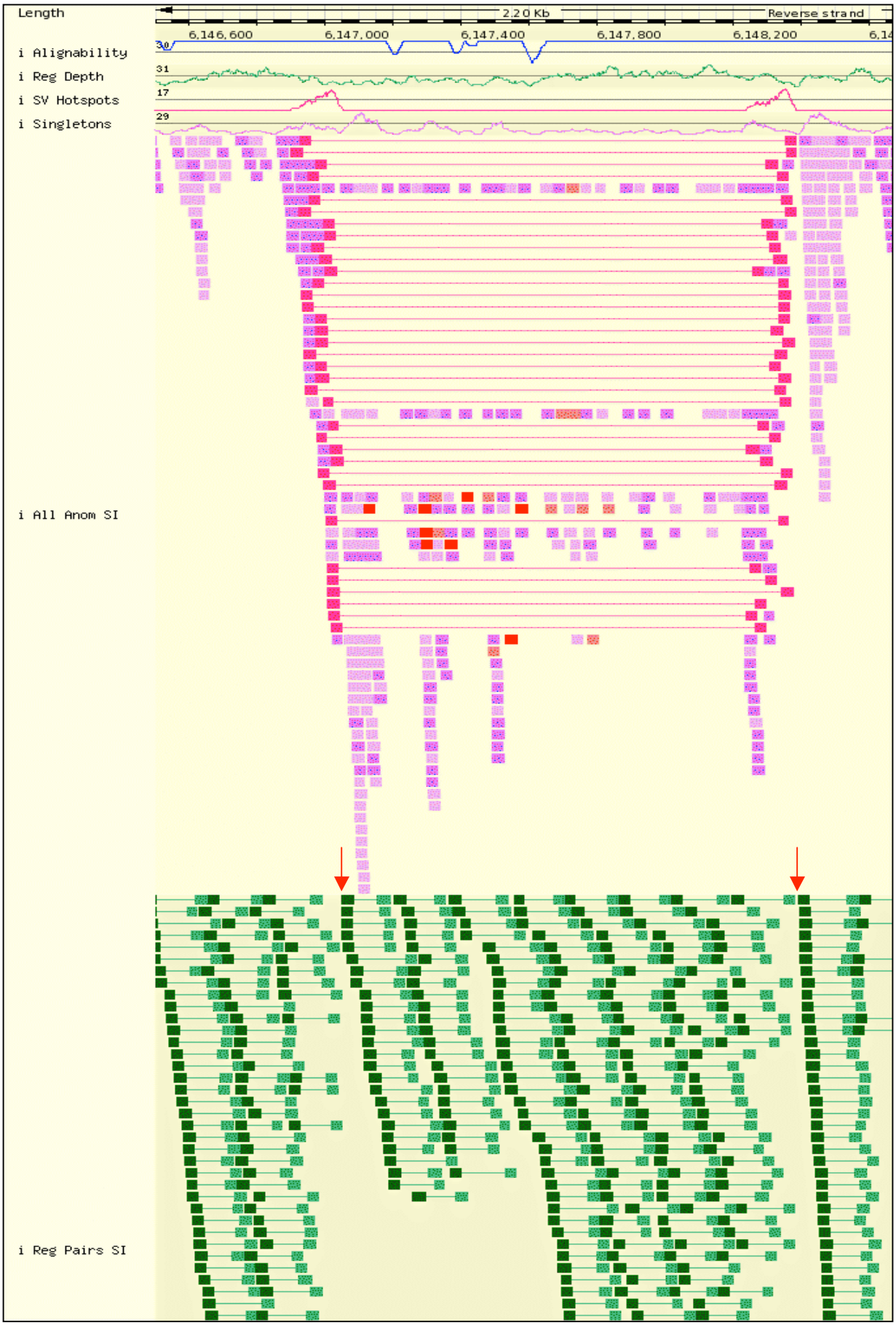


Figure S10. (See next page for legend).

Figure S10 (see previous page for figure). Homozygous 1.5 kb inversion in NA07340 relative to the reference, detected by anomalously spaced and inverted read pairs (red) compared to regularly spaced read pairs (green). Note the gaps in continuity of coverage in the normal short inserts, indicated by the red arrows, indicating that this variant is homozygous. No read pair spans either breakpoint. The reader is asked to refer to the website (link below) and to zoom in to visualise the gap running through the complete dataset. Note also the alignability score (blue line) is high across most of the region. Read depth (green line) remains high across the inversion. Purple reads are singletons, which accumulate as the other member of the read pair does not match to the reference. Singletons at each inversion breakpoint result from the disruption in alignment of the other member of the pair at or near the breaks. Red arrows indicate homozygous breaks in continuity with normal paired end mapping. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

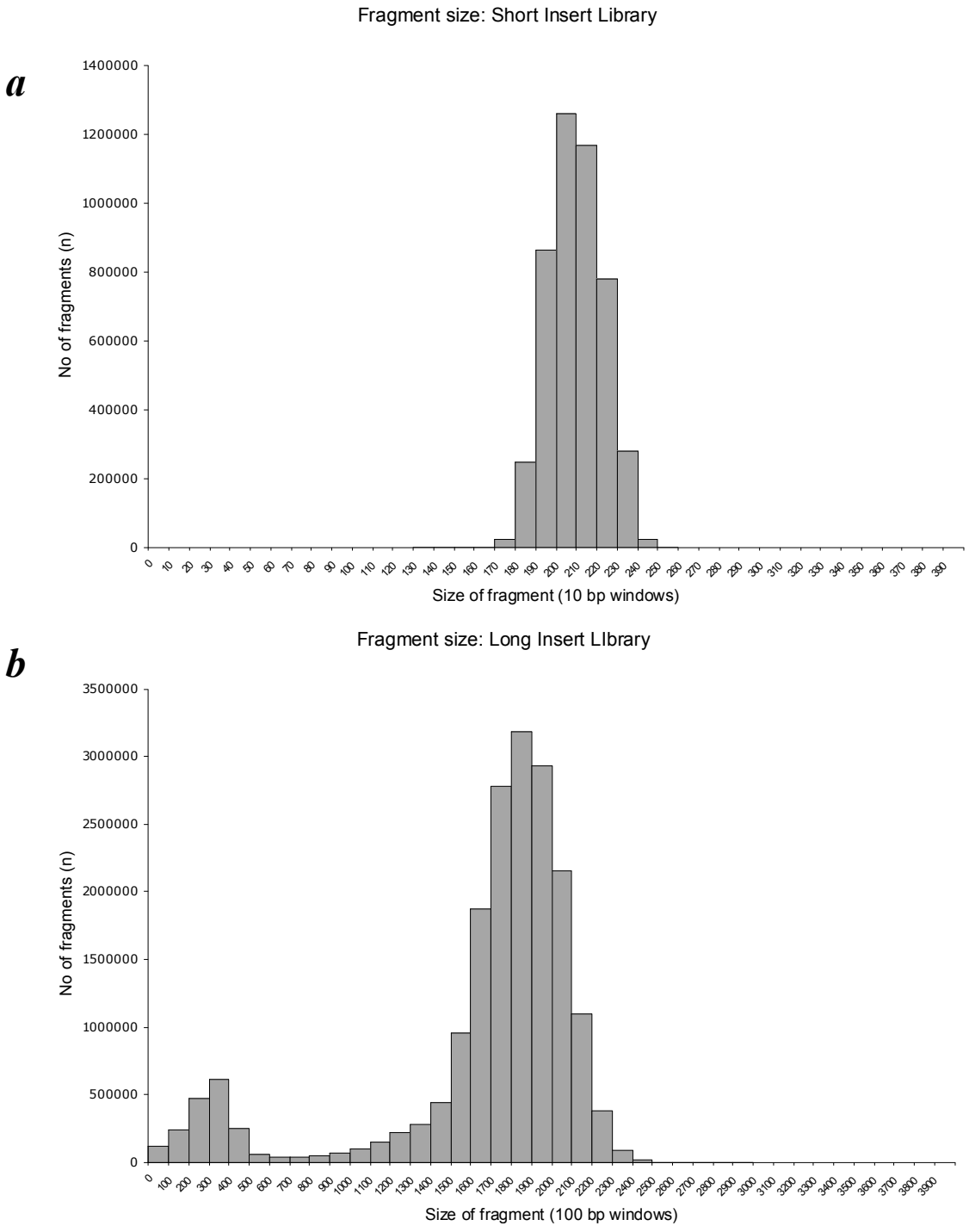


Figure S11 Size distribution for DNA fragment libraries of NA18507. Fragment lengths were determined from the separation between read pairs aligned to the reference. **a**. Short insert library, data from one lane of a flowcell were analysed and fragment lengths binned in 10 bp windows. **b.** Long insert library, data from six lanes of a flowcell with lengths binned in 100 bp windows. Note that the smaller peak on the left reflects the presence of a small fraction of short, non-junction fragments of genomic DNA that are not biotinylated. These are readily eliminated during analysis as the orientation of the paired reads is opposite to that of the real junction fragments (see fig 1 for more details).

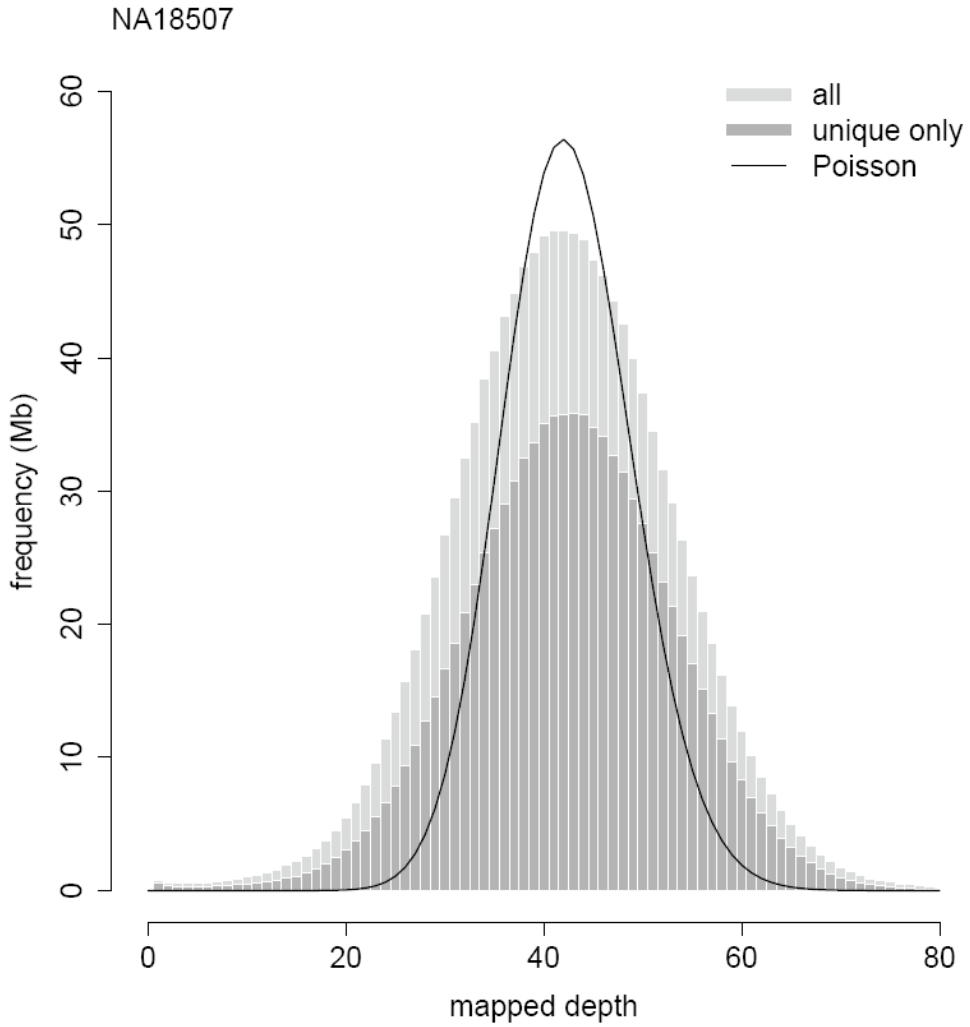


Figure S12a. Distribution of mapped read depth in the human genome dataset, sampled at every 500th position along the genome and displayed as a histogram ('all'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('unique only'). The solid line represents a Poisson distribution for unique human genome sequence

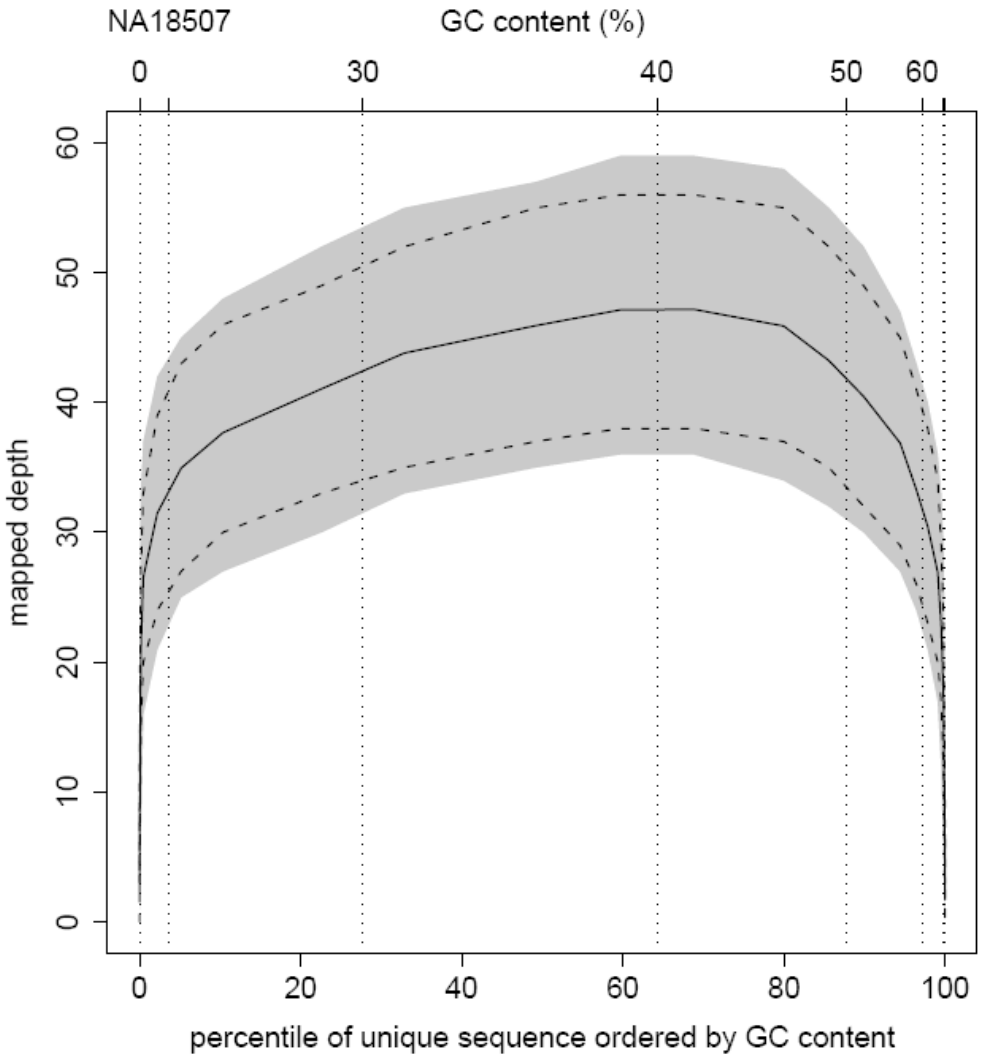


Figure S12b. Distribution of human genome NA18507 uniquely mapped reads as a function of GC content. Note that the x axis is % GC content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 90th centiles); the dashed lines are 10th and 90th centiles of a Poisson distribution with the same mean as the data. GC biases may be introduced during sample pre-amplification, cluster growth and/or sequencing. Possible contributors include reduced efficiency of GC-rich fragment amplification or loss of very short AT-rich fragments due to denaturation during sample preparation, both of which might vary from library to library and result in the difference in distribution seen here compared to the X-chromosome library (see fig. 2b).

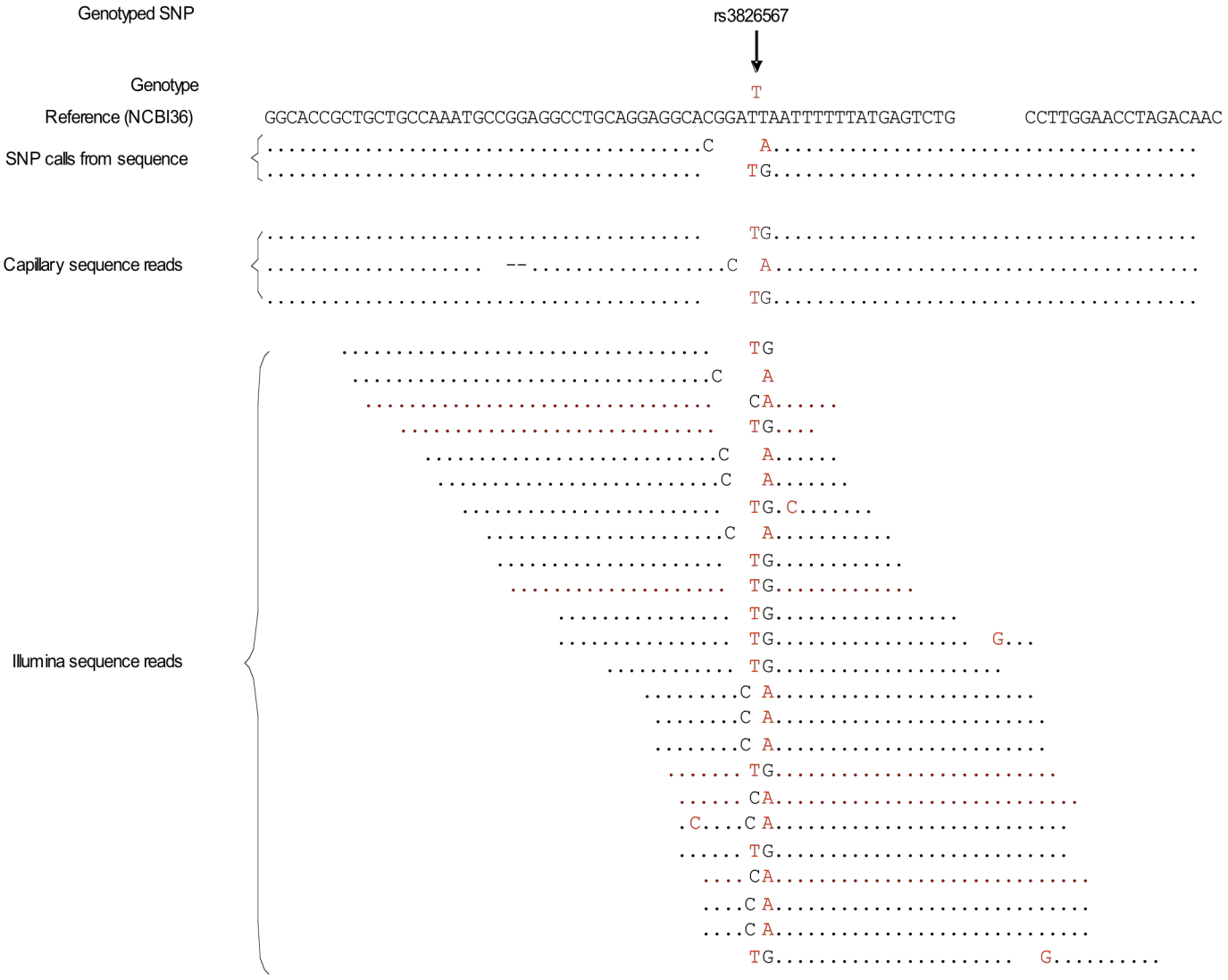


Figure S13. Sequence data reveals 'hidden' SNPs that cause genotyping errors. *a*. Example of a 'hidden' SNP detected in the sequence data. The HM550 genotyping records a 'T' only for NA18507 at SNP rs3826567 (arrowed), indicating that this individual is homozygous 'TT' at this position. Illumina sequence reads detect two alleles at this position, T and C, indicating a heterozygous position 'TC'. Capillary sequence reads obtained previously for this individual were available in the trace archive and confirm the existence of both T and C alleles. The sequence data identify the existence of a second SNP (the 'hidden SNP') directly adjacent to rs3826567 and indicate that NA18507 is heterozygous 'AG' at this position. As the genotyping assay utilises the reference sequence (in this case 'G') for design of the oligonucleotides used in the assay, the genotyping assay does not accommodate the existence of the non-reference 'A' allele and only interacts with DNA containing the reference 'G' allele, resulting in a 'T' call at rs3826567.

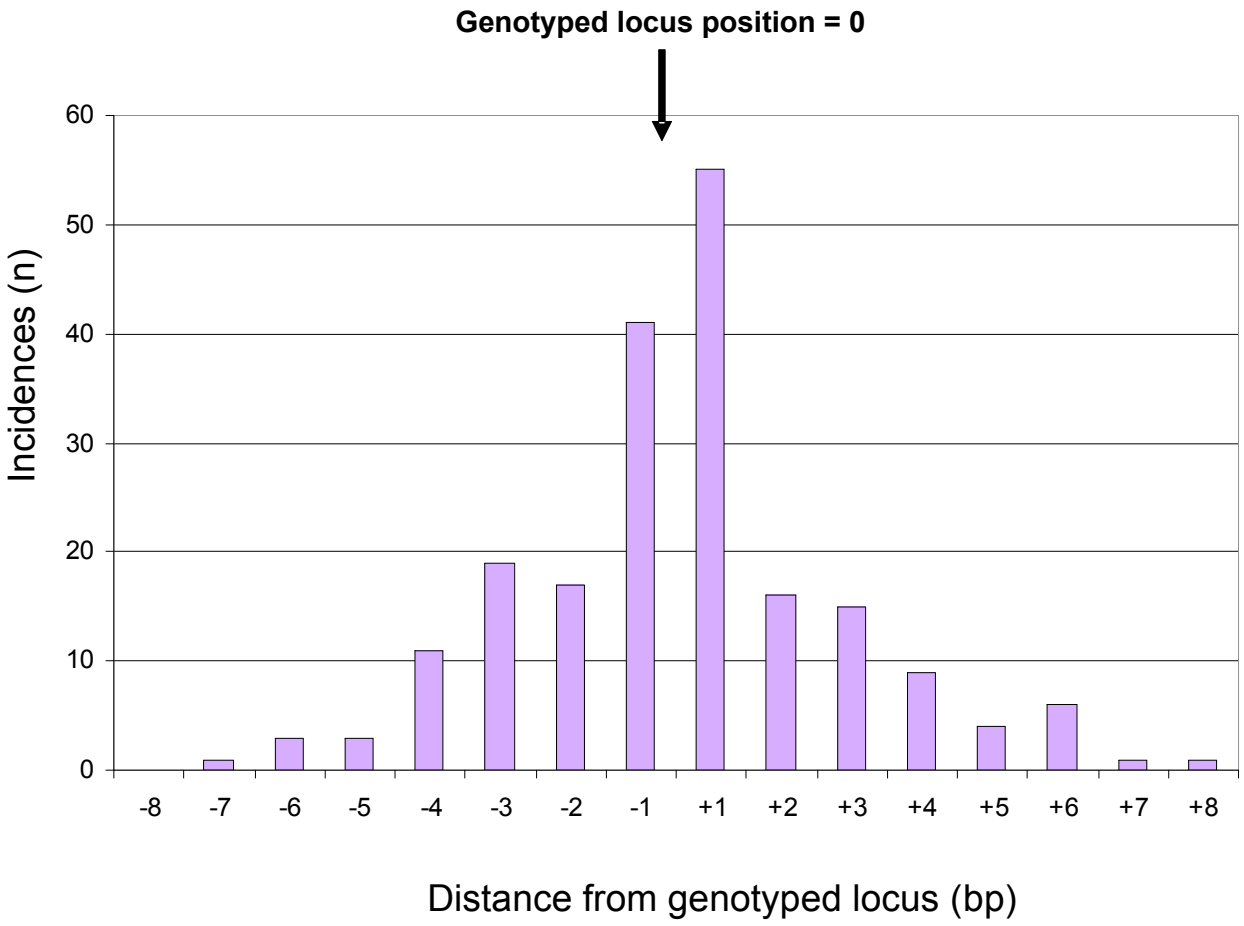
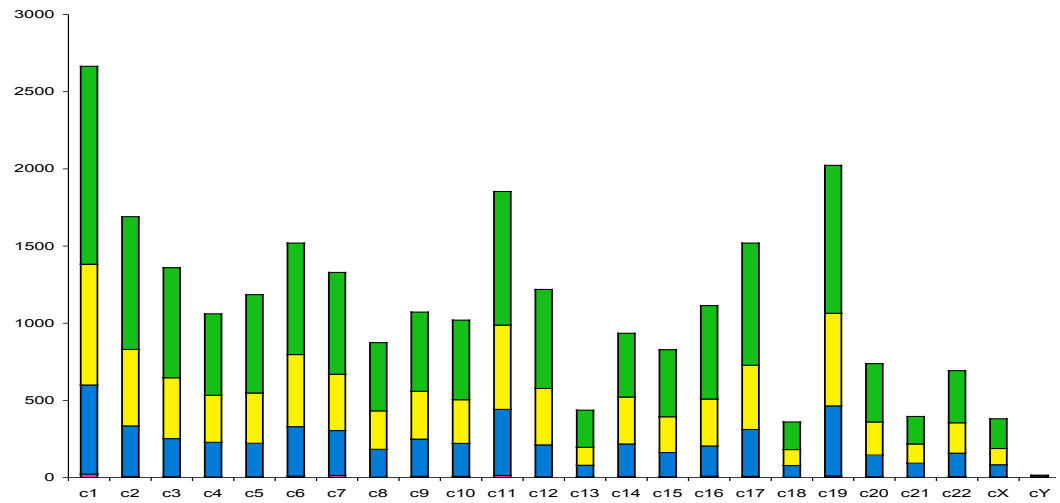
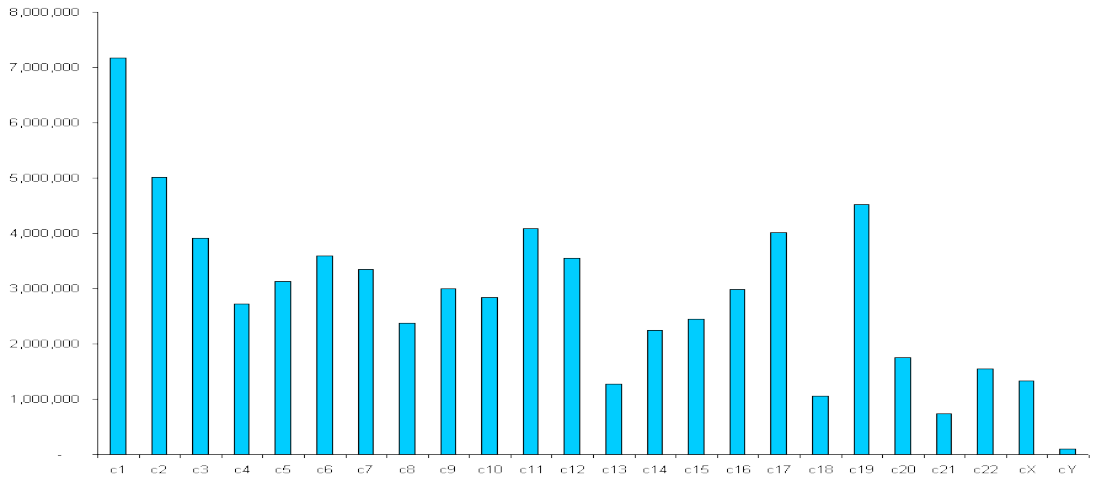


Figure S14. Distribution of 'hidden' SNPs in 202 discordant calls (see main text for discription of 'hidden' SNPs). Note that 'hidden' SNPs will affect any genotyping platform. We therefore expect a low, but non-zero, level of genotype undercalling (Seq>GT) that is highly correlated between the platforms that use these flanking sequences in assay design.

a



b



c

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12
Heterozygosity (no. of hets/covered coding bases x 10-3)	0.54	0.46	0.46	0.54	0.48	0.57	0.57	0.55	0.51	0.51	0.66	0.45
Chromosome	13	14	15	16	17	18	19	20	21	22	All coding	All genome
Heterozygosity (no. of hets/covered coding bases x 10-3)	0.49	0.58	0.45	0.57	0.53	0.42	0.63	0.62	0.65	0.61	0.54	0.99

Figure S15. Coding SNPs in NA18507, based on the gene annotation in Ensembl (release 49). ***a***. distribution of coding SNPs and their consequences by chromosome. Green: synonymous, yellow: non-synonymous, conservative, blue: non-conservative, red: premature stop codons. ***b***. distribution of coding bases by chromosome. Note close correlation of coding SNP and coding sequence distribution by chromosome (as opposed to overall chromosome length). ***c.*** Correlation of heterozygosity with coding length between chromosomes. Average autosomal coding region heterozygosity ('All coding') is approximately half that of genome average for autosomes ('All genome').

a

	n	%dbSNP
All indels	404,416	49.82%
Heterozygotes	260,620	41.44%
Homozygotes	143,796	64.99%

b

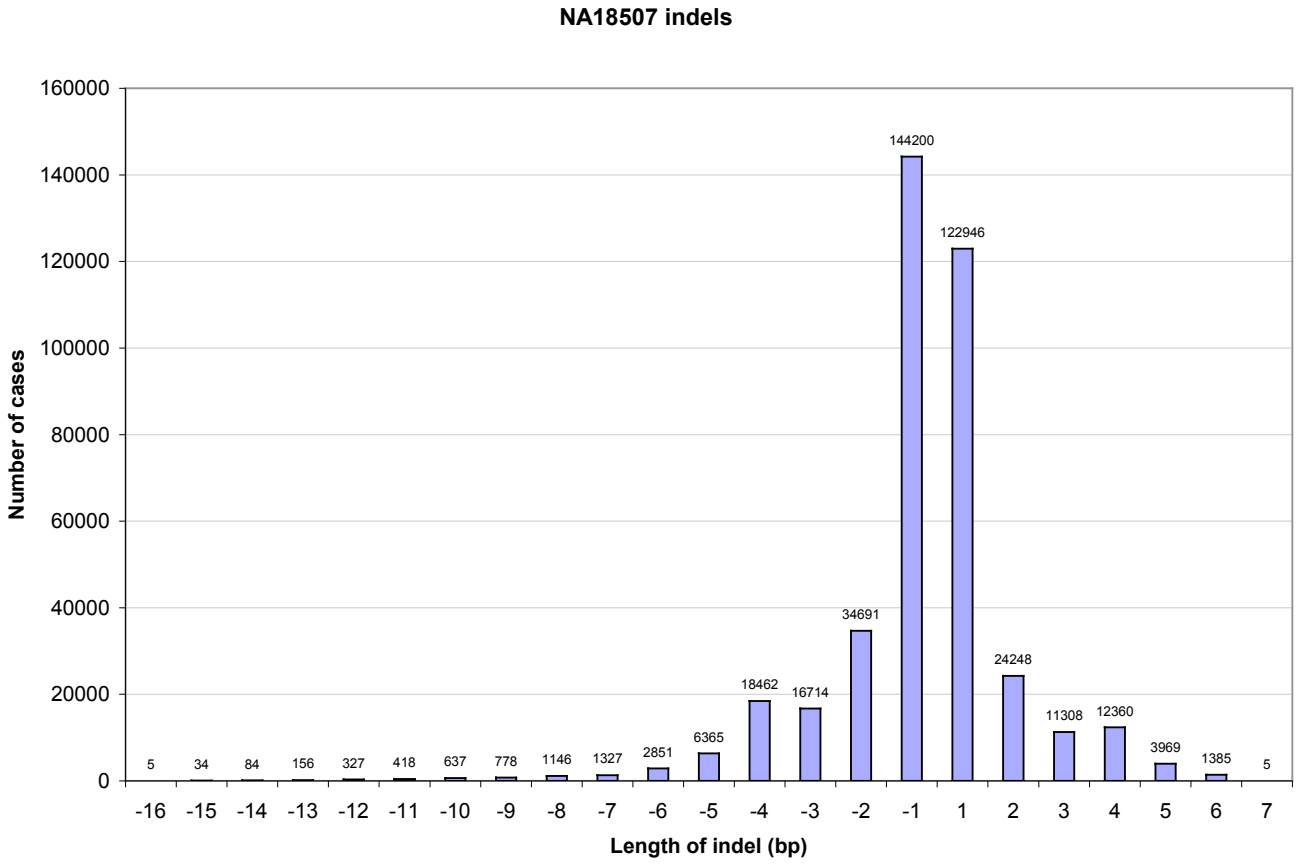


Figure S16. Analysis of short indel calls in human genome data for N18507. ***a.*** total number of calls and fraction that match previous entries in dbSNP. ***b.*** Distribution of size in the 404,416 indels. + and – values on the x axis correspond to presence or absence of bases in NA18507 relative to the reference sequence.

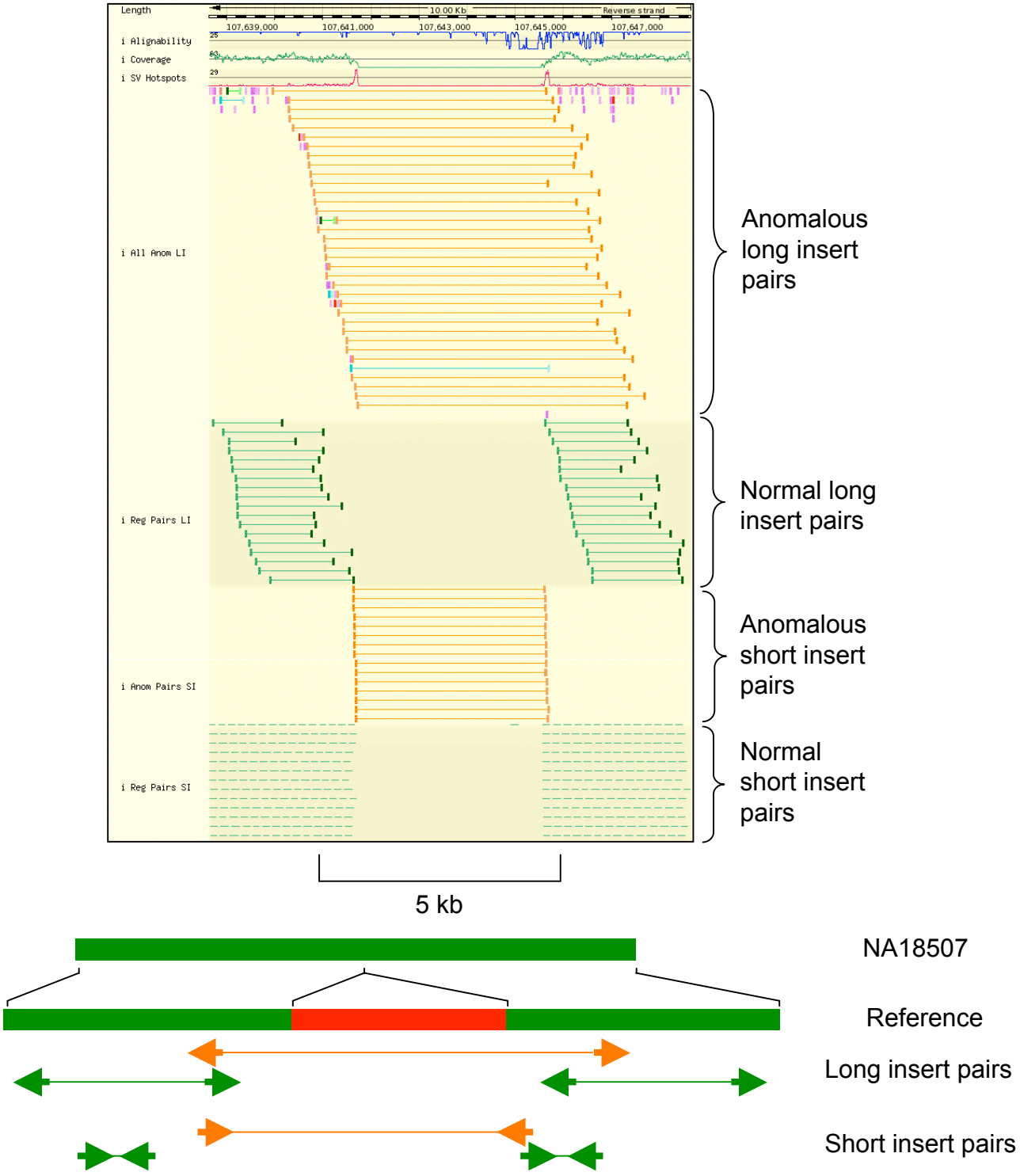


Figure S17*a*. Homozygous 3.6 kb deletion in NA18507 relative to the reference, supported by anomalously spaced long and short insert read pairs (orange) compared to regularly spaced read pairs (green). See also the schematic below. Note that the correct read orientation for long insert read pairs is different from that of the short insert read pairs and is denoted by the arrows which point in a 5' to 3' direction. See fig 1 for explanation. Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

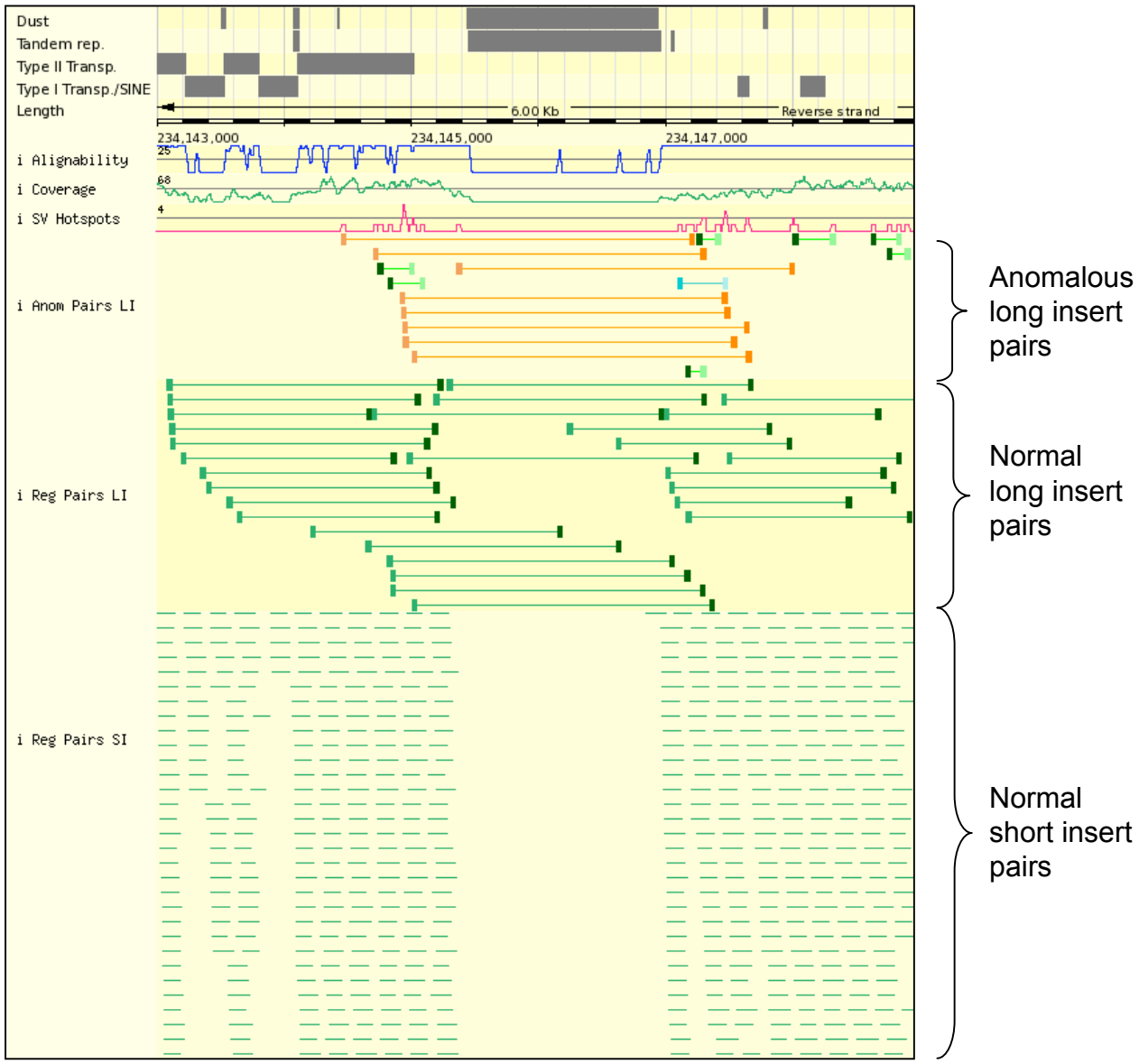


Figure S17b. Deletion in NA18507 relative to the reference, seen with anomalous long insert pairs but not short insert data. Tandem repeats (second row of display) cover ~1500 bp on the right hand side of the deletion, preventing unique alignment of reads. Short insert pairs are too short to span the tandem repeat region. Long insert pairs are longer range and right-hand reads map in unique sequence further away from the deletion (see orange blocks representing reads on right hand side of each of the long insert anomalous pairs). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

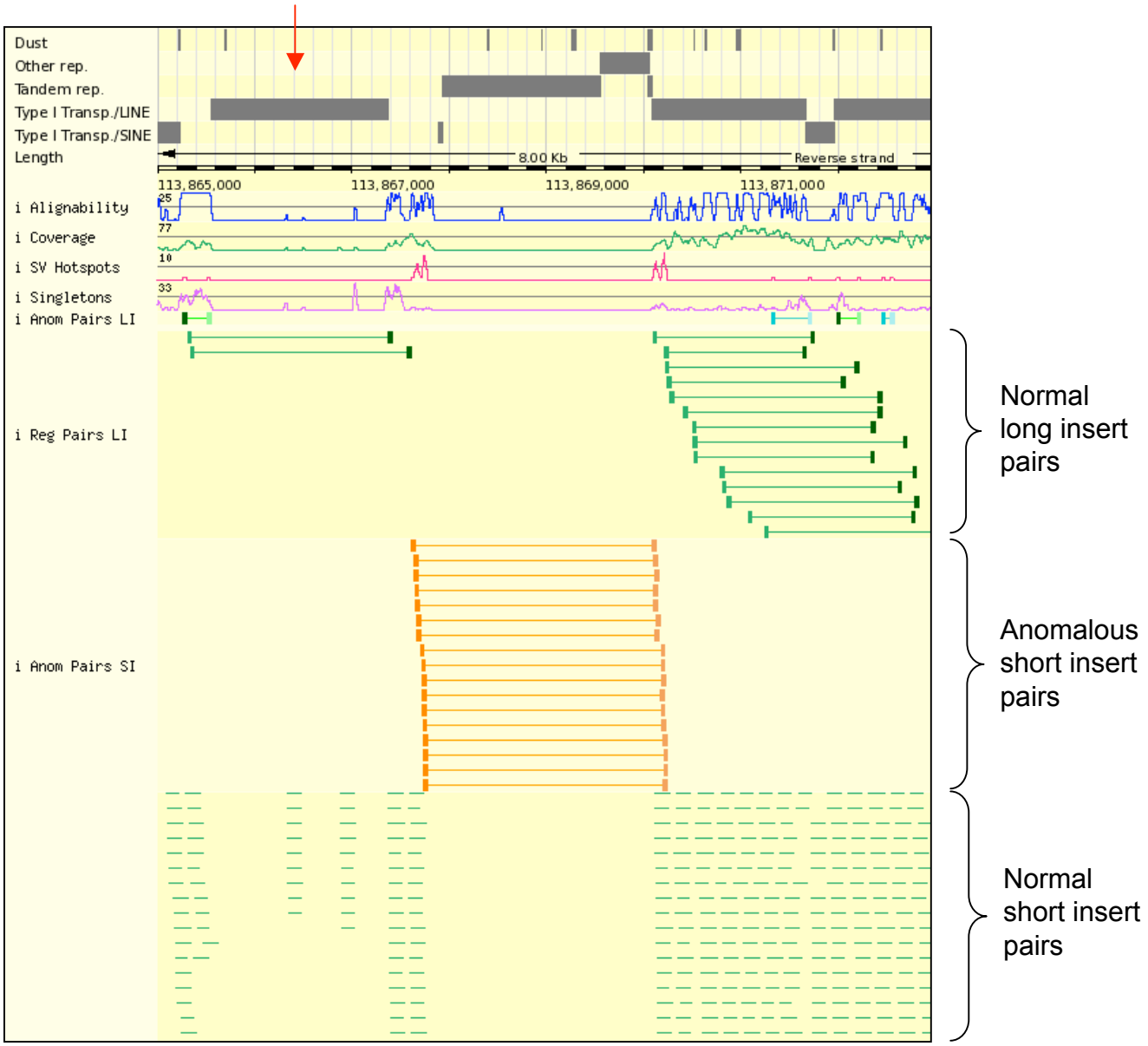


Figure S17c. Deletion in NA18507 relative to the reference, seen with anomalous short insert pairs but not long insert data. The LINE repeat on the left (see red arrow) covers 1.8 kb on the left-hand side of the deletion, preventing unique alignment of long insert reads. The higher density of short insert pairs includes a number of examples that map directly to either side of the deletion, with left ends mapping between the deletion breakpoint and the LINE repeat. Note that the LINE repeat on the right of the deletion does not appear to significantly compromise alignability (c.f. the LINE repeat on the left). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

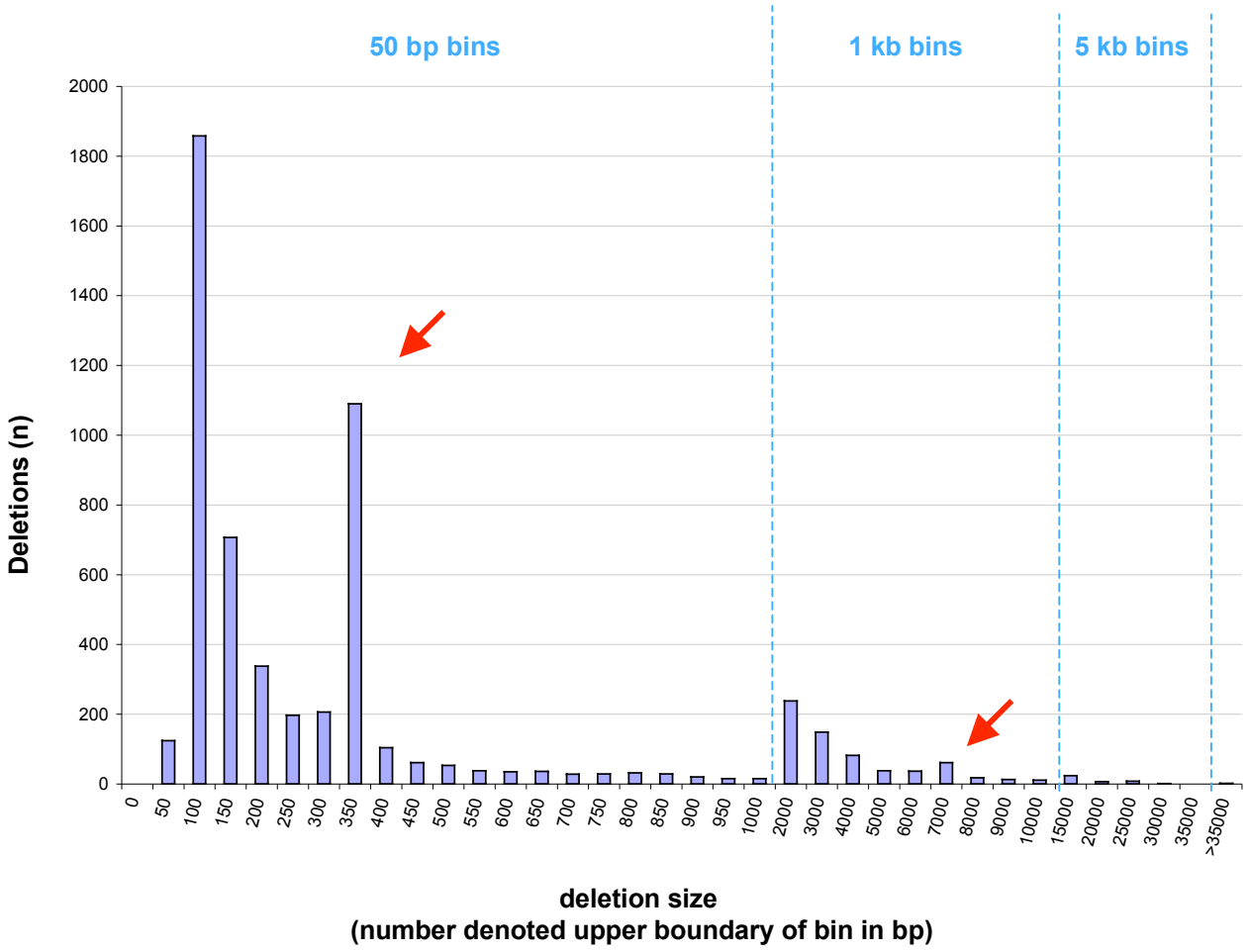


Figure S18. Size distribution of deletions in NA18507 relative to the reference. Two peaks in the distribution (arrowed) contain a substantial fraction of sequences corresponding to recently active retroposon insertions. The AluY family of SINEs are found in 90% (987/1092 of the 300-350bp fraction, and also contribute slightly to the 250-300 bp fraction. The L1HS LINE element accounts for 67% (42/61) of the 6-7 kb fraction. Note that other peaks are artefacts in the plot caused by the changes in size range of the bins.

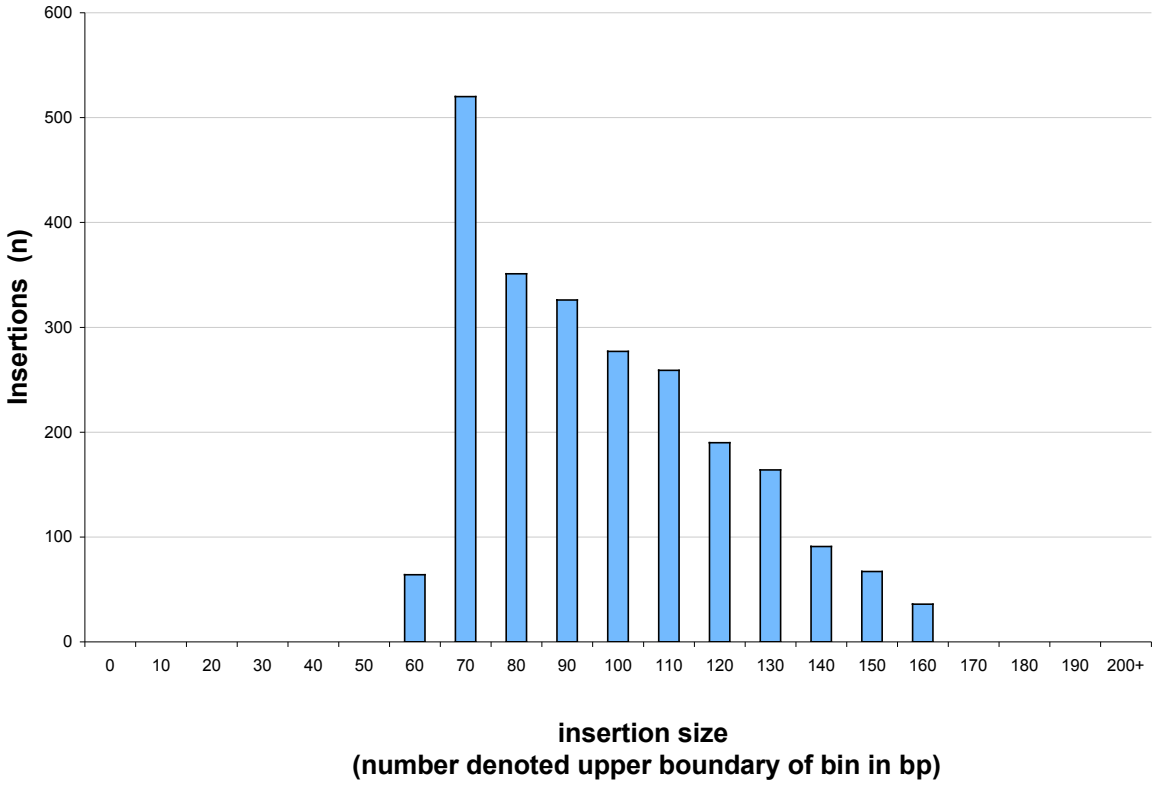


Figure S19. Size distribution of insertions in NA18507 relative to the reference, detected in the short insert dataset. Size distribution is limited by the resolution of analysis of the short insert fragment size range.

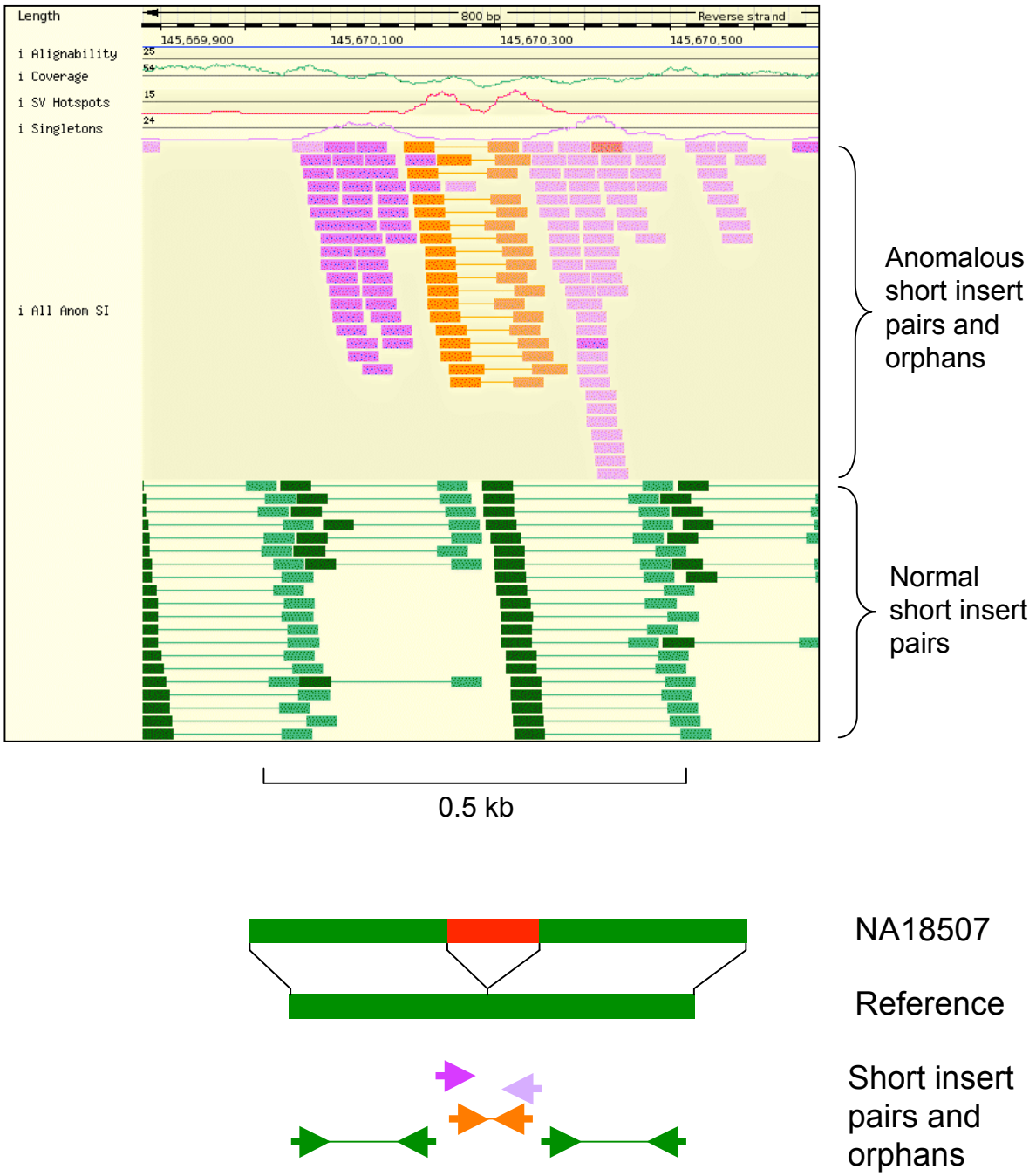


Figure S20. Homozygous 81 bp insertion detected by short insert data only. Orange bars show the anomalously short read pair spacing. Note the accumulation of singleton reads on either side, where each read has an unaligned pair. Dark and light purple denote rightwards and leftwards orientation (see schematic). Thus all singleton reads (dark purple) have unaligned pairs to the right of them; and vice versa for pale purple singletons. These unaligned pairs form part of the *de novo* assembly that results in a single contig across the sequence inserted in NA18507 relative to the reference (see fig S21). Note that this display is a composite of screen shots of the same window, overlapped for display purposes in this figure.

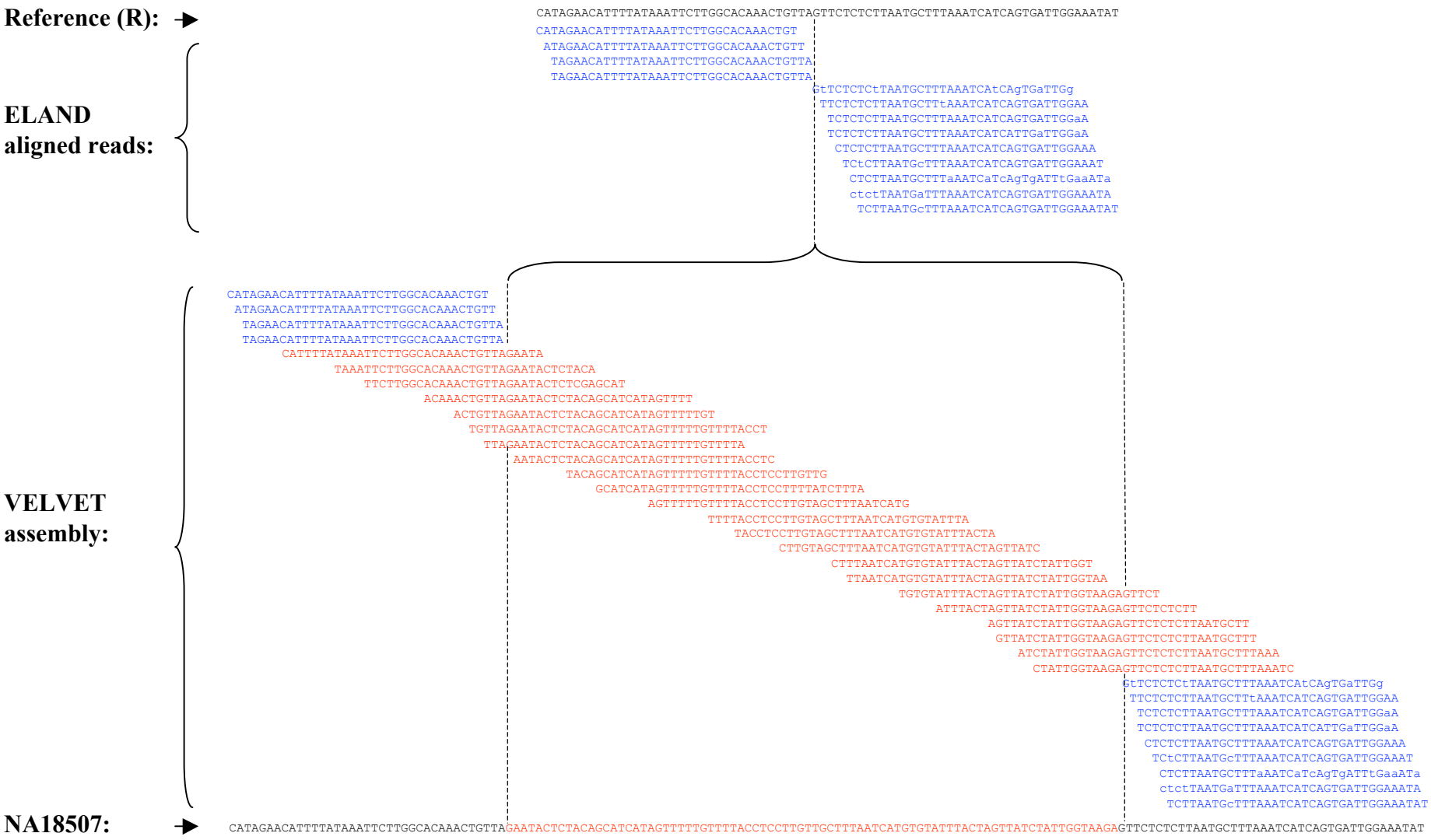


Figure S21. *De novo* assembly defines an inserted sequence in NA18507 absent from the reference. Bases present in the reference are in black, bases in NA18507 reads that align to the reference using ELAND are in blue, and bases in additional NA18507 reads that are assembled by VELVET are in red. The extra sequence in the NA18507 consensus (bottom line) is also shown in red.

a 071126_EAS56_0057_FC200433 – lanes 1-8 read 1

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	266021	31482 +/- 1612	23032 +/- 1162	3646 +/- 318	99.39 +/- 3.25	73.17 +/- 1.17	86.05 +/- 0.35	49.50 +/- 1.73	0.50 +/- 0.04
2	258725	30395 +/- 1596	22400 +/- 1396	3803 +/- 385	92.23 +/- 2.98	73.68 +/- 1.84	86.03 +/- 0.39	48.68 +/- 2.21	0.53 +/- 0.07
3	262340	30436 +/- 1382	22713 +/- 1128	3610 +/- 370	98.04 +/- 3.36	74.62 +/- 1.07	86.05 +/- 0.30	49.82 +/- 1.61	0.49 +/- 0.04
4	264384	30633 +/- 1697	22890 +/- 1321	4111 +/- 372	94.06 +/- 3.44	74.77 +/- 2.69	86.23 +/- 0.29	50.01 +/- 1.57	0.48 +/- 0.06
5	266557	30793 +/- 1492	23078 +/- 1039	4278 +/- 388	94.15 +/- 3.64	74.97 +/- 0.94	86.09 +/- 0.27	50.36 +/- 1.01	0.48 +/- 0.02
6	261586	30493 +/- 1656	22648 +/- 1194	4294 +/- 403	94.10 +/- 3.25	74.30 +/- 1.68	86.14 +/- 0.44	48.53 +/- 3.00	0.50 +/- 0.15
7	261867	30502 +/- 1835	22672 +/- 1180	4056 +/- 469	96.64 +/- 3.85	74.39 +/- 1.76	86.12 +/- 0.29	50.04 +/- 1.52	0.49 +/- 0.03
8	67955	30168 +/- 1697	21815 +/- 1276	3629 +/- 950	96.75 +/- 4.11	72.36 +/- 2.69	86.11 +/- 0.30	49.81 +/- 1.89	0.50 +/- 0.04

b 071126_EAS56_0057_FC200433 – lanes 1-8 read 2

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	266021	31482 +/- 1612	23032 +/- 1162	1954 +/- 174	92.80 +/- 6.94	73.17 +/- 1.17	84.17 +/- 0.88	42.70 +/- 2.70	0.75 +/- 0.11
2	258725	30395 +/- 1596	22400 +/- 1396	2109 +/- 206	89.61 +/- 6.56	73.68 +/- 1.84	84.45 +/- 1.06	42.62 +/- 3.25	0.74 +/- 0.14
3	262340	30436 +/- 1382	22713 +/- 1128	2045 +/- 190	95.16 +/- 5.93	74.62 +/- 1.07	84.65 +/- 0.48	44.50 +/- 2.08	0.67 +/- 0.06
4	264384	30633 +/- 1697	22890 +/- 1321	2338 +/- 220	90.64 +/- 5.90	74.77 +/- 2.69	84.97 +/- 0.39	45.20 +/- 2.11	0.64 +/- 0.05
5	266557	30793 +/- 1492	23078 +/- 1039	2260 +/- 220	92.23 +/- 5.66	74.97 +/- 0.94	84.59 +/- 0.55	44.61 +/- 2.19	0.68 +/- 0.10
6	261586	30493 +/- 1656	22648 +/- 1194	2345 +/- 216	90.92 +/- 5.23	74.30 +/- 1.68	84.81 +/- 0.74	44.47 +/- 2.63	0.67 +/- 0.08
7	261867	30502 +/- 1835	22672 +/- 1180	2209 +/- 244	93.00 +/- 5.65	74.39 +/- 1.76	84.56 +/- 0.50	44.30 +/- 2.51	0.70 +/- 0.07
8	67955	30168 +/- 1697	21815 +/- 1276	1945 +/- 524	95.72 +/- 6.79	72.36 +/- 2.69	84.41 +/- 0.78	43.70 +/- 2.72	0.72 +/- 0.08

Table S1. Summary tables for ***a*** read 1 and ***b*** read 2 of a production sequencing run of human DNA sample NA18507. For each of lanes 1-7, 330 tiles of data were collected and each tile contained approximately 22,000-23,000 clusters providing PF sequence. Lane yield = all PF clusters per lane (n) x read length (0.035 kbases). First cycle intensity (Int) is averaged over all PF reads and given in arbitrary units. % intensity after 20 cycles provides a measure of signal loss during the run. The % PF clusters indicates the fraction of raw clusters that provide PF sequence data after purity filtering (described in supplementary methods). The % align indicates what fraction of the PF reads align to the reference (in this case human genome ncbi36) using ELAND. This value will vary for different reference sequences. The % Error Rate is the average % differences for all bases in the PF dataset when compared to the reference, and includes all differences due true polymorphisms as well as alignment errors and sequence errors. The yield for this run was 3.8 Gb of PF sequence data.

	ELAND			MAQ		
	N	%		N	%	
All PF reads	278.6	100.0%		278.6	100.0%	
PF reads aligning to genome	239.3	85.9%		267.4	96.0%	
Non duplicate PF reads aligning to genome	216.8	77.8%	(90.6%)	236.7	85.0%	(88.5%)
Non duplicate PF reads aligning to X chromosome	191.9	68.9%	(88.5%)	203.9	73.2%	(86.2%)

Table S2. Distribution of reads in flow-sorted X chromosome analysis Number of reads N is given in millions. Percent values not in brackets are fraction of the total number of reads (top line value for N). Percent values in (brackets) express the proportion of reads as a fraction of the category on the previous line. Thus using ELAND, non duplicate PF reads aligning to genome are 90.6% of all PF reads aligning to genome. They represent 77.8% of all PF reads.

		Genotyped	ELAND		MAQ	
			Loci		Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	8,887	8,865	99.75%	8,885	99.98%
	Heterozygote	4,717	4,708	99.81%	4,707	99.79%
	All	13,604	13,573	99.77%	13,592	99.91%
Concordant calls	Homozygote		8,865	100.00%	8,885	100.00%
	Heterozygote		4,643	98.62%	4,706	99.98%
	All		13,508	99.52%	13,591	99.99%
All disagreements			65	0.48%	1	0.01%
GT>Seq			65	0.48%	1	0.01%
Seq>GT			0	0.00%	0	0.00%
Other discordances			0	0.00%	0	0.00%

Table S3. Comparison of X chromosome SNP calls made from sequence vs. genotype data. Genotyping of the NA07340 sample was carried out using the Illumina HumanHap550 BeadChip. GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

Left end	Right end	Length	MAQ APs	ELAND APs	Type	Repeat	Depth	Ext. evi
402924	403049	130	9	5	hom		1	3,4
435713	436369	630	11	3	hom		0	4
823102	823947	888	7	8	het		1	
827894	828076	167	9	10	hom		0	
830122	830892	735	5	3	het		1	
1130746	1131006	348	17	9	het		0	
1574560	1574825	226	10	7	hom		0	
1871088	1871680	628	4	9	het		1	4
2204889	2204968	173	5	7	het		1	4
2618720	2618890	201	12	6	hom		1	4
3044385	3044590	194	9	3	hom		0	4
4536875	4537115	248	13	11	hom		1	3,4
4914723	4915023	328	29	24	hom	AluYa5	0	2,3
5065353	5067496	2164	27	24	het		1	4
5580355	5580671	356	58	50	hom	AluYa5	1	1,2,3,4
6730947	6731159	182	5	2	hom		0	3,4
8746525	8748362	2349	6	12	het		1	4
9720010	9720194	158	8	7	hom		1	3
11635293	11641318	6069	36	35	hom	L1HS	1	1,3,4
11863127	11869352	6258	57	42	hom	L1HS	1	1,4
16337623	16338127	550	47	37	hom	L1HS	1	1,3
17430556	17430853	336	61	47	hom	AluY	1	1,3
22582505	22582796	334	34	25	hom	AluY	1	
32380501	32380550	172	6	3	hom		0	
33118510	33118820	332	31	29	hom	AluYb8	1	2,3
38281972	38282205	222	3	6	het		1	
38469315	38473453	4170	3	6	het		1	3,4
39520762	39521076	295	5	5	hom		1	1,2,3
45472524	45472804	320	58	54	hom	AluY	1	2,3
52904522	52906807	2288	11	8	het		0	
58406935	58407152	200	1	6	hom		0	3
66020510	66020630	154	18	15	het		0	3
66250543	66250693	197	51	45	hom		1	1,3,4
71039962	71040256	342	72	51	hom	AluYa5	1	2,3,4
75946512	75946687	200	20	14	het		0	
77583524	77584097	614	63	46	hom	L1HS	1	3
78833108	78835525	2408	16	25	hom		1	4
80929136	80929745	610	7	8	het		1	1,3,4
80983316	80989344	6067	13	11	hom	L1HS	0	4
84229771	84229872	134	11	6	hom		1	
86165708	86166809	1116	23	24	hom		1	1
86173414	86173872	478	7	7	het		1	
88239424	88239563	166	13	27	het		1	
89794851	89795060	159	5	5	het		0	4

Table S4 (part 1 of 2). See part 2 for legend.

Table S4 (part 2 of 2). Analysis of structural variants in X chromosome data for NA07340. Maq or Eland APs refer to the number of anomalous pairs (APs) that support the variant. A variant was called if the number of APs was 5 or more in one analysis. Thereafter, supporting data in the other analysis is recorded if there are 1 or more APs in the other analysis. Depth indicates if there was evidence for the variant on the basis of read depth (1) or not (0) in addition to the AP data. Overlap with independent datasets is indicated in the final column by 1: Evidence for a variant obtained from split fosmid reads from the same DNA detected by alignment to the reference using Ssaha; 2: entry in dbRIP; 3, entry in DGVindel (see refs 15, 23) or 4: Database of Genomic Variants. We found 39 variants in the present dataset that were supported by evidence from independent analyses of category 1, 2 or 3. Note in the case of 4 that the overlap can be across a very large region and provides corroboration but does not necessarily constitute strong supporting evidence for detection of the same event.

start[1]	length[1]	Reference copy number in region at read-length (32 bp) resolution	Copy number difference in NA07340 (diploid)[2]	Corroborated by external data[3]
4000	12000	Mostly 2	loss (-2)	1
16375	18350	Variable between 20 and 120	gain (+++)	
80000	15000	Mostly 2	loss (-2)	
97000	9000	Variable between 20 and 120	gain (+++)	
408025	1200	Approximately 40	gain (+++)	2
930225	750	Variable between 10 and 120	gain (+++)	
950000	180000	At least 2	loss (-2)	
1715775	4500	Approximately 25	loss (---)	
3750000	120000	At least 2	loss (-2)	2
9332000	11000	Mixture of 5 and 10	gain (+++)	1,2
19375625	4000	Mostly unique	loss (-1)	1
48771725	3900	Mostly 2	loss (-2)	2
49050000	210000	Tandem repeat of 16 units	gain (+++)	
49708725	300	Approximately 15	gain (+++)	1
52100000	500000	Mixture of 2 and 5	loss (--)	
52844000	6000	Variable between 2 and 10	loss (--)	
55220000	50000	Variable between 2 and 10	gain (+++)	
56810000	14000	Variable between 3 and 5	gain (++)	1,2
62250000	180000	Mixture of 2 and 10	gain(+2)	1,2
69378925	8450	Mostly unique	gain (+1)	1
73025525	200	At least 10	gain (+++)	
77402000	10000	Mostly unique	loss (-1)	1
81868975	5950	At least 5	loss (---)	1
88344000	6000	Variable up to 45	gain (+++)	1
96493275	2050	Mostly unique	loss (-2)	1
101253375	150	At least 10	gain (+++)	
114860000	60000	Tandem repeat of 20 units	gain (+++)	1,2
118758025	3500	Mostly unique	loss (-2)	1
119893875	54650	Tandem repeat of 12 units	loss (---)	2
125433375	1550	Variable between 2 and 6	gain (+++)	2
130832975	3300	Mostly unique	loss (-2)	1
134670000	130000	Variable between 2 and 12	gain (++)	2
139633075	400	Variable between 2 and 10	gain (+++)	
146168000	9000	Mostly 2	gain (++)	2
150831000	12000	Variable up to 10	loss (-1)	
154430975	8650	Mostly unique	loss (-1)	1,2
154866825	4050	At least 2	gain (++)	
154913225	350	At least 90	gain (+++)	

Table S5. Analysis of structural variants in X chromosome data for NA07340 based on read depth only.

[1] There is some imprecision in the definition of the breakpoints as genomic windows are assessed.

[2] ++/-- = 3 to 8 additional/fewer copies; +++/--- = more than 8 additional/fewer copies.

[3] 1 = CGH, 2 = Database of Genomic Variants

a

	ELAND			MAQ		
	n (billions)	%		n (billions)	%	
All PF reads	3.77	100.0%		3.77	100.0%	
PF reads aligning to genome	3.41	90.5%		3.66	97.1%	
Non duplicate PF reads aligning to genome	3.14	83.3%	(92.0%)	3.33	88.3%	(91.0%)

b

	ELAND		
	n (millions)	%	
All PF reads	296.1	100.0%	
PF reads aligning to genome	270.7	91.4%	
Non duplicate PF reads aligning to genome	228.6	77.2%	(84.4%)

Table S6. Summary of reads in human genome sequencing: ***a*** short insert library (208+/- 13 bp); ***b*.** long insert library (1,840+/- 200bp). The numbers in parentheses shows % retained from the category on the preceding line.

a

HM550 COMPARISON		Genotyped	ELAND Loci		MAQ Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	390,585	389,139	99.63%	389,909	99.83%
	Heterozygote	162,125	161,357	99.53%	161,359	99.53%
	All	552,710	550,496	99.60%	551,268	99.74%
Concordant calls	Homozygote		388,798	99.91%	389,609	99.92%
	Heterozygote		159,317	98.74%	161,115	99.85%
	All		548,115	99.57%	550,724	99.90%
All disagreements			2,381	0.43%	544	0.10%
GT>Seq			1,940	0.35%	170	0.03%
Seq>GT			258	0.05%	290	0.05%
Other discordances			183	0.03%	84	0.02%

b

HM-All COMPARISON		Genotyped	ELAND Loci		MAQ Loci	
		n	n	%	n	%
Covered by sequence	Homozygote	2,901,392	2,876,278	99.13%	2,882,021	99.33%
	Heterozygote	798,200	795,041	99.60%	791,434	99.15%
	All	3,699,592	3,671,319	99.24%	3,673,455	99.29%
Concordant calls	Homozygote		2,851,483	99.14%	2,855,192	99.07%
	Heterozygote		775,708	97.57%	785,757	99.28%
	All		3,627,191	98.80%	3,640,949	99.12%
All disagreements			44,136	1.20%	32,506	0.88%
GT>Seq			16,821	0.46%	5,602	0.15%
Seq>GT			19,268	0.52%	19,679	0.54%
Other discordances			8,047	0.22%	7,225	0.20%

Table S7. Detailed breakdown of comparison of human genome SNP calls made from sequence vs. genotype data. Genotype data were generated using the Infinium HumanHap550 BeadChip (HM550) (***a***); or obtained from the International HapMap Project (HM-All) (***b***). GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.

		SNP calls n	%
Genotyped loci	Homozygote	309	
	Heterozygote	897	
	All	1206	
Concordant calls	Homozygote	296	95.8%
	Heterozygote	863	96.2%
	All	1159	96.1%
All disagreements		47	3.9%
GT>Seq		7	0.6%
Seq>GT		22	1.8%
Other discordances		18	1.5%

Table S8. Validation of novel SNP calls by genotyping. Results for 1,206 randomly chosen novel SNP calls. There is 96.1% (1159/1206) concordance overall between sequencing and genotype calls. In all there were 47 discrepancies. GT>Seq denotes a heterozygous call by genotyping, where one of the alleles is detected by sequencing. These seven undercalls are considered to be false negatives. Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. On manual inspection we found that there were two classes of Seq>GT: Ten have nearby hidden SNPs and are therefore considered correct sequence calls. The other twelve lie in small indels that were missed in by the indel caller. These are therefore counted as false positive SNP calls in the present analysis. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call and are also counted as false positives. The total false positive discovery rate among the novel SNPs in this analysis is therefore 2.4% (30/1206).

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c1	16895539	A	C	exon:ENST00000270691	ESPNP	E638STOP	HOM	Known
c1	20374169	GA	G	exon:ENST00000247992	Q5R387	R27STOP	HET	Known
c1	32034879	GA	G	exon:ENST00000360482	SPOCD1	Q724STOP	HET	Novel
c1	46853266	AG	G	exon:ENST00000271139	MOBKL2C	R24STOP	HET	Known
c1	54954888	TA	T	exon:ENST00000395691	C1orf175	C1290STOP	HET	Known
c1	55023902	GA	G	exon:ENST00000371274	TTC22	R342STOP	HET	Novel
c1	86873963	G	C	exon:ENST00000284054	CLCA3	Y84STOP	HOM	Known
c1	89221456	AT	A	exon:ENST00000321792	NP62556.2	Y214STOP	HET	Novel
c1	100454229	CA	C	exon:ENST00000370132	DBT	E224STOP	HET	Novel
c1	110268078	AG	G	exon:ENST00000369803	CSF1	W491STOP	HET	Known
c1	111769644	TC	C	exon:ENST00000369728	OVGP1	W131STOP	HET	Known
c1	116028154	CA	C	exon:ENST00000355485	VANGL1	S338STOP	HET	Novel
c1	120138308	CT	C	exon:ENST00000369402	REG4	W114STOP	HET	Known
c1	143563747	CT	C	exon:ENST00000369356	PDE4DIP	W2351STOP	HET	Known
c1	143626918	GA	G	exon:ENST00000369356	PDE4DIP	R622STOP	HET	Known
c1	156816116	T	C	exon:ENST00000368150	OR10X1	W66STOP	HOM	Known
c1	159359477	GA	G	exon:ENST00000392188	DEDD	Q198STOP	HET	Novel
c1	169379114	T	C	exon:ENST00000236166	FMO6P	Q105STOP	HOM	Known
c1	230165470	CT	C	exon:ENST00000359647	Q6ZRE3	R116STOP	HET	Known
c1	234772923	AT	T	exon:ENST00000352231	LGALS8	L212STOP	HET	Known
c1	246179649	AT	T	exon:ENST00000357191	OR2L8	Y289STOP	HET	Known
c1	246789346	A	G	exon:ENST00000328570	OR2T29	Q24STOP	HOM	Novel
c2	31252918	CG	G	exon:ENST00000398824	N/A	Y708STOP	HET	Known
c2	89400921	GA	G	exon:ENST00000390263	IGKV1-39	R18STOP	HET	Known
c2	95883315	AG	G	exon:ENST00000357042	Q53S06	Q32STOP	HET	Known
c2	159224979	AG	G	exon:ENST00000342892	Q6ZTQ2	Q103STOP	HET	Novel
c2	207528443	AG	G	exon:ENST00000272852	CPO	W59STOP	HET	Novel
c3	15468185	GT	G	exon:ENST00000383782	COLQ	S285STOP	HET	Novel
c3	16243990	AC	C	exon:ENST00000388817	GALNTL2	S162STOP	HET	Known
c3	139505118	AG	G	exon:ENST00000383180	TXNDC6	Q171STOP	HET	Novel
c3	150980779	AG	G	exon:ENST00000383050	N/A	R130STOP	HET	Novel
c4	12948066	GA	G	exon:ENST00000382444	HSP90AB2P	W252STOP	HET	Novel
c4	70933511	AT	T	exon:ENST00000381057	HTN3	Y47STOP	HET	Known
c4	77879853	CA	C	exon:ENST00000296043	SHROOM3	Y501STOP	HET	Novel
c4	102301286	T	C	exon:ENST00000399145	N/A	Q115STOP	HOM	Known
c5	1089962	AC	C	exon:ENST00000382730	NKD2	S23STOP	HET	Known
c5	1293757	CG	C	exon:ENST00000324642	SLC6A18	Y319STOP	HET	Known
c5	32185000	AG	G	exon:ENST00000332294	Q9UI72	W17STOP	HET	Known
c5	75000878	A	G	exon:ENST00000344149	N/A	W419STOP	HOM	Known
c5	134810349	A	T	exon:ENST00000333591	C5orf20	R117STOP	HOM	Known
c6	33156444	T	G	exon:ENST00000395267	HLA-DPB1	G17STOP	HOM	Known
c6	41210331	T	C	exon:ENST00000338759	Q6ZRD8	R20STOP	HOM	Known
c6	43032225	A	G	exon:ENST00000326586	Q5T8W0	R74STOP	HOM	Known
c6	57506229	CT	C	exon:ENST00000389488	PRIM2	Q325STOP	HET	Known
c6	74076059	AG	G	exon:ENST00000370384	C6orf148	Q34STOP	HET	Known
c6	139618237	A	G	exon:ENST00000367652	TXLNB	Q123STOP	HOM	Known
c6	150253727	GA	G	exon:ENST00000367363	RAET1E	R2STOP	HET	Novel
c6	154402262	TC	C	exon:ENST00000330432	OPRM1	Q26STOP	HET	Known
c6	160823771	C	G	exon:ENST00000297289	LPAL2	Y69STOP	HOM	Novel
c7	21549488	TG	G	exon:ENST00000328843	DNAH11	E34STOP	HET	Known
c7	39615800	AG	G	exon:ENST00000357751	Q8N8G3	W43STOP	HET	Known
c7	55508201	GA	G	exon:ENST00000285279	NM30796.3	R83STOP	HET	Known
c7	63806451	A	T	exon:ENST00000344930	ZNF107	C778STOP	HOM	Novel
c7	64076102	AG	G	exon:ENST00000398695	ZNF117	R428STOP	HET	Known
c7	80138385	G	T	exon:ENST00000309881	CD36	Y325STOP	HOM	Known
c7	99072835	A	T	exon:ENST00000379712	Q6ZVR2	C18STOP	HOM	Known
c7	138727857	A	T	exon:ENST00000397584	Q8WYX0	L80STOP	HOM	Known
c7	141589238	A	G	exon:ENST00000389113	N/A	Q390STOP	HOM	Known
c7	141918182	T	C	exon:ENST00000390382	TRBV7-9	R55STOP	HOM	Novel
c7	142459539	TC	C	exon:ENST00000391508	OR6V1	Q80STOP	HET	Known
c7	143995851	TG	G	exon:ENST00000378099	TPK1	S24STOP	HET	Novel
c7	149123706	CT	C	exon:ENST00000262089	Q76B61	R1193STOP	HET	Known
c7	149153299	G	T	exon:ENST00000378016	NM198455.2	L331STOP	HOM	Known

Table S9 (part 1 of 3)

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c8	222801	CT	C	exon:ENST00000382855	N/A	Q133STOP	HET	Novel
c8	12480423	TA	T	exon:ENST00000359701	Q8NAJ9	L30STOP	HET	Known
c8	101791075	CA	C	exon:ENST00000318607	PABPC1	E345STOP	HET	Novel
c9	19347166	CT	C	exon:ENST00000380424	DENND4C	Q246STOP	HET	Novel
c9	44114866	GC	G	exon:ENST00000377553	Q8NGA9	Y224STOP	HET	Novel
c9	45060265	CA	C	exon:ENST00000377537	Q6ZVI3	C148STOP	HET	Novel
c9	66297827	TG	T	exon:ENST00000359897	Q6ZS19	Y41STOP	HET	Known
c9	125012928	CA	C	exon:ENST00000398660	N/A	S87STOP	HET	Novel
c9	134946938	GT	G	exon:ENST00000314220	Q8N7A6	S57STOP	HET	Known
c9	138754316	AG	G	exon:ENST00000341040	LCN10	Q148STOP	HET	Known
c9	139057620	A	G	exon:ENST00000371600	NPDC1	R25STOP	HOM	Known
c10	38225046	CA	C	exon:ENST00000374537	N/A	S190STOP	HET	Novel
c10	38934622	CT	C	exon:ENST00000399687	N/A	R26STOP	HET	Novel
c10	46825773	AT	T	exon:ENST00000342900	N/A	L340STOP	HET	Novel
c10	47239426	T	C	exon:ENST00000335083	N/A	R81STOP	HOM	Novel
c10	73634249	CG	G	exon:ENST00000394915	ASCC1	S78STOP	HET	Known
c10	96437552	AT	T	exon:ENST00000285979	CYP2C18	Y68STOP	HET	Known
c10	126668082	GA	G	exon:ENST00000337195	CTBP2	Q445STOP	HET	Known
c11	1662913	C	G	exon:ENST00000382165	FAM99A	S25STOP	HOM	Novel
c11	2278405	TC	C	exon:ENST00000381153	C11orf21	W23STOP	HET	Novel
c11	5733060	AT	A	exon:ENST00000317254	OR52N4	R172STOP	HET	Known
c11	6107204	GA	G	exon:ENST00000316517	Q8NH77	W263STOP	HET	Novel
c11	45029583	GA	G	exon:ENST00000378793	Q6ZS37	R107STOP	HET	Novel
c11	48242807	AT	T	exon:ENST00000320048	OR4X1	Y273STOP	HET	Known
c11	48303590	GA	G	exon:ENST00000395239	OR4C3	W14STOP	HET	Known
c11	48344426	GT	G	exon:ENST00000319813	OR4C5	Y56STOP	HET	Known
c11	60021578	CT	C	exon:ENST00000016913	MS4A12	Q71STOP	HET	Known
c11	92781120	TC	C	exon:ENST00000298050	CCDC67	R460STOP	HET	Novel
c11	104268327	A	G	exon:ENST00000375726	CASP12	R125STOP	HOM	Novel
c11	123561942	TG	T	exon:ENST00000318666	Q8NH80	Y252STOP	HET	Known
c12	3591206	T	A	exon:ENST00000382608	Q6UY24	L71STOP	HOM	Known
c12	50490305	TA	T	exon:ENST00000313893	Q8NBA4	L27STOP	HET	Novel
c12	54231266	TC	C	exon:ENST00000394256	OR6C4	R5STOP	HET	Known
c12	120699406	A	G	exon:ENST00000333310	Q9H9T1	W5STOP	HOM	Novel
c13	17941446	CT	C	exon:ENST00000342944	N/A	W241STOP	HET	Known
c13	18897913	A	G	exon:ENST00000341326	TPTE2	R324STOP	HOM	Known
c13	52020061	TC	C	exon:ENST00000258589	Q5JV89	R138STOP	HET	Novel
c13	111076351	AG	G	exon:ENST00000375713	Q5T400	W3STOP	HET	Novel
c14	22668816	GA	G	exon:ENST00000334354	SLC7A8	R179STOP	HET	Known
c14	24173254	GA	G	exon:ENST00000382542	GZMB	Q19STOP	HET	Known
c14	49872477	C	A	exon:ENST00000356146	CDKL1	L989STOP	HOM	Known
c14	74228760	GA	G	exon:ENST00000338772	Q86TS6	R95STOP	HET	Known
c14	105161972	T	G	exon:ENST00000390543	IGHG4	S290STOP	HOM	Known
c15	18726530	T	C	exon:ENST00000400226	LOC440368	W124STOP	HOM	Known
c15	19000266	CA	C	exon:ENST00000400222	NM1001413.2	E487STOP	HET	Novel
c15	53510174	AC	C	exon:ENST00000321149	DYX1C1	E417STOP	HET	Novel
c15	67114791	GA	G	exon:ENST00000388866	NOX5	W300STOP	HET	Known
c15	95102199	TC	C	exon:ENST00000398303	N/A	R167STOP	HET	Known
c16	764474	A	G	exon:ENST00000293892	NP1020361.1	Q590STOP	HOM	Novel
c16	977150	CT	C	exon:ENST00000305973	Q96S05	Q39STOP	HET	Novel
c16	1373566	AG	G	exon:ENST00000397462	UNKL	Q744STOP	HET	Novel
c16	33537201	AG	G	exon:ENST00000354689	N/A	W7STOP	HET	Known
c16	33537474	AG	G	exon:ENST00000354689	N/A	W66STOP	HET	Known
c16	33868594	GA	G	exon:ENST00000380143	Q6PQ33	R450STOP	HET	Known
c16	88172869	CG	C	exon:ENST00000268720	CPNE7	S154STOP	HET	Novel
c16	88655877	AG	G	exon:ENST00000325921	Q9NQW5-2	R73STOP	HET	Novel

Table S9 (part 2 of 3)

chromosome	position	call	ref	exon info (ensembl id)	gene	change	hom/het	known/novel
c17	3142353	TA	T	exon:ENST00000323404	OR3A1	K92STOP	HET	Known
c17	4744490	A	G	exon:ENST00000381365	Q6ZR85	W152STOP	HOM	Known
c17	20710491	GT	G	exon:ENST00000327925	NP1004306.1	S178STOP	HET	Known
c17	21144803	CT	C	exon:ENST00000342679	MAP2K3	Q102STOP	HET	Known
c17	36559089	GA	G	exon:ENST00000343246	KRTAP4-5	R153STOP	HET	Known
c17	37237375	C	G	exon:ENST00000393910	NT5C3L	S137STOP	HOM	Known
c17	64723626	TG	A	exon:ENST00000269081	ABCA10	Y261STOP	HET	other
c17	70100401	AC	C	exon:ENST00000328023	C17orf77	C207STOP	HET	Known
c18	50112154	AC	A	exon:ENST00000307844	STARD6	L114STOP	HET	Novel
c19	776155	TC	C	exon:ENST00000334630	AZU1	Q20STOP	HET	Novel
c19	9098263	GA	G	exon:ENST00000305444	OR7G3	R122STOP	HET	Known
c19	40410860	T	C	exon:ENST00000324675	TMEM162	W188STOP	HOM	Known
c19	48463827	GA	G	exon:ENST00000270077	PSG9	R127STOP	HET	Novel
c19	55246050	GA	G	exon:ENST00000377006	Q6ZNZ3	W93STOP	HET	Known
c19	56822743	GA	G	exon:ENST00000222107	SIGLEC5	R356STOP	HET	Known
c19	62334594	A	C	exon:ENST00000391707	USP29	Y913STOP	HOM	Known
c19	63065900	AT	A	exon:ENST00000396146	Q9NWJ2	R58STOP	HET	Novel
c19	63785276	TC	C	exon:ENST00000312426	NP76428.2	Q26STOP	HET	Known
c20	25703913	CA	C	exon:ENST00000376403	Q5T319	E15STOP	HET	Novel
c20	26032175	GA	G	exon:ENST00000356439	C20orf191	R81STOP	HET	Known
c20	55497442	AT	A	exon:ENST00000243919	HMG1L1	Y16STOP	HET	Known
c21	27137937	T	C	exon:ENST00000382884	ADAMTS1	W4STOP	HOM	Known
c21	45588725	GA	G	exon:ENST00000215202	C21orf111	Q188STOP	HET	Known
c22	15849049	AC	C	exon:ENST00000400588	GAB4	G163STOP	HET	Known
c22	21037728	CT	C	exon:ENST00000390293	IGLV5-48	Q106STOP	HET	Novel
c22	25398153	TC	C	exon:ENST00000382641	Q6ZVA3	W143STOP	HET	Known
c22	34886714	AG	G	exon:ENST00000349314	APOL3	Q58STOP	HET	Known
c22	37900350	CA	C	exon:ENST00000330414	YV004	G9STOP	HET	Known
c22	41000237	GT	G	exon:ENST00000332965	Q8IXR4	G223STOP	HET	Known
cX	30999550	A	C	exon:ENST00000359202	FTHL17	E148STOP	HOM	Known
cX	118489029	GT	G	exon:ENST00000317881	SLC25A5	E293STOP	HET	Known
cX	138492281	C	G	exon:ENST00000370573	MCF2	S857STOP	HOM	Novel
cX	142795134	G	T	exon:ENST00000370494	UBE2NL	L89STOP	HOM	Known
cMT	9378	A	G	exon:ENST00000391562	Q14Y83	W57STOP	HOM	Known

Table S9 (part 3 of 3). Coding SNPs predicted to cause premature termination of translation in NA18507. Note that this list excludes positions where the reference sequence also contains a termination codon within annotated coding sequence.

ATTACHMENT 10A

LDR	03177cas a2200733 4500
001	991253473406676
005	20250227135100.0
007	cr|cn#||||||||
008	740401c18699999enkwr p 0 0eng
022	__ | **a**0028-0836**l**0028-0836**9**PY**z**0302-2889**7**(Print)
022	__ | **a**1476-4687**9**EN**7**(Electronic)
030	__ | **a**NATUAS
035	__ | **9**0410462
040	__ | **a**DNLM**c**DNLM
041	0_ | **a**eng
044	__ | **9**England
060	00 | **a**W1**b**NA81
210	10 | **a**Nature**2**DNLM
222	_0 | **a**Nature**b**(London)
245	00 | **a**Nature.
246	2_ | **0**NATURE (LOND**9**n
246	2_ | **0**NATURE**9**n
246	2_ | **0**Nature (London)**9**n
246	2_ | **0**Nature (London)**9**om
246	2_ | **0**Nature (Lond)**9**om
246	2_ | **0**Natyre, Lond.**9**om
260	__ | **a**London,**b**Macmillan Journals ltd.
260	3_ | **a**Basingstoke :**b**Nature Publishing Group
300	__ | **a**v.**b**illus., ports.
310	__ | **a**Weekly
321	__ | **a**Weekly,**b**-1980
336	__ | **a**text**b**txt**2**rdacontent
337	__ | **a**unmediated**b**n**2**rdamedia
338	__ | **a**volume**b**nc**2**rdacarrier
362	0_ | **a**v. 1- (no. 1- ; Nov. 4, 1869-
500	__ | **a**Part of title on microfilm.
500	__ | **a**Issued 1971-73 in three parts: Nature, Nature: New biology, and Nature: Physical science. The latter two are cataloged separately in this library.
510	2_ | **9**1**a**MEDLINE**b**v205n976, Mar. 13, 1965-
510	2_ | **9**1**a**PubMed**b**v205n976, Mar. 13, 1965-
510	0_ | **a**OLDMEDLINE
510	4_ | **a**NLM Hidden treasure**c**p. 218 (April 1953 issue)
510	2_ | **a**Index medicus
525	__ | **a**Supplements accompany many numbers.
530	__ | **a**Also issued online.

533	__ | **a**Microfilm.**m**v. 153 (1944 Jan-Jun), v. 218, no. 5136-5148 (1968), v. 220, no. 5162-5173 (1968).**b**[Bethesda, Md.] :**c**National Library of Medicine,**d**1994-**e**5 microfilm reels : negative ; 35 mm.
555	__ | **a**Vols. 229-234, 1971; Vols. 235-240, 1972.
580	__ | **a**Vols. 229-246, Jan. 1971-Dec. 1973, issued in three parts. The two new parts, Nature: new biology, and Nature: physical sciences, continued the volume numbering of Nature in addition to carrying their own issue numbering. With v. 246, Dec. 1973, Nature: New biology, and Nature: Physical science, ceased publication and were absorbed by Nature.
580	__ | **a**Beginning in 1994 complemented by Nature structural biology, which consists of articles on structural biology for a more specialist audience.
650	_2 | **a**Science
655	_2 | **a**Periodical
787	1_ | **t**Nature: New biology**x**0090-0028**w**(OCoLC)1759490
787	1_ | **t**Nature: Physical science**x**0300-8746**w**(OCoLC)1759491
787	1_ | **t**Nature structural biology**g**1994-2003**x**1072-8368**w**(DLC) 95660668**w**(OCoLC)29209845
787	1_ | **t**Nature cell biology**g**May 1999-**x**1465-7392**w**(DLC)sn 99031593**w**(OCoLC)41423161
856	4_ | **u**http://www.nature.com/nature
911	__ | **a**Electronic and print, print canceled 2017
914	__ | **a**Portico**b**Current to 2024 as of 20250227**c**ISSN_00280836
935	__ | **a**(DNLM)125347-nlmdb
995	__ | **a**AUTH**b**19740401**c**REV**d**20250227
996	__ | **a**NOTCNSR**b**20000101
998	__ | **a**ERMSLC 108646**b**20061129
999	__ | **a**AUTH
035	__ | **a**(OCoLC)01586310
035	__ | **a**(DNLM)N07600000(s)**z**(DNLM)SR0086301(s)

ATTACHMENT 10B

Science MeSH Descriptor Data 2025

Details | Qualifiers | MeSH Tree Structures | Concepts

MeSH Heading
Science
Tree Number(s)
H01.770
Unique ID
D012586
RDF Unique Identifier
http://id.nlm.nih.gov/mesh/D012586
Annotation
use for the discipline (education, history, etc) only
Scope Note
The study of natural phenomena by observation, measurement, and experimentation.
See Also
Religion and Science
Date Established
1966/01/01
Date of Entry
1999/01/01
Revision Date
2015/06/16