



UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY. DOCKET NO	TOT CLAIMS	IND CLAIMS
61/798,925	03/15/2013		125	S11-230 PR1		

CONFIRMATION NO. 6603

86012
VLP Law Group LLP
555 Bryant Street
Suite 820
Palo Alto, CA 94301

FILING RECEIPT



Date Mailed: 05/14/2013

Receipt is acknowledged of this provisional patent application. It will not be examined for patentability and will become abandoned not later than twelve months after its filing date. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please submit a written request for a Filing Receipt Correction. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections**

Inventor(s)

Maximilian DIEHN, San Carlos, CA;
Arash A. Alizadeh, San Mateo, CA;
Aaron M. Newman, Palo Alto, CA;
Scott V. Bratman, Palo Alto, CA;

Applicant(s)

Maximilian DIEHN, San Carlos, CA;
Arash A. Alizadeh, San Mateo, CA;
Aaron M. Newman, Palo Alto, CA;
Scott V. Bratman, Palo Alto, CA;

Power of Attorney:

David Roise--47904

If Required, Foreign Filing License Granted: 05/09/2013

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 61/798,925**

Projected Publication Date: None, application is not eligible for pre-grant publication

Non-Publication Request: No

Early Publication Request: No

**** SMALL ENTITY ****

Title

IDENTIFICATION AND USE OF CIRCULATING TUMOR MARKERS

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications:

page 1 of 3

TEMPUS EX1025

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at <http://www.uspto.gov/web/offices/pac/doc/general/index.html>.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, <http://www.stopfakes.gov>. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4158).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign Assets Control, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit <http://www.SelectUSA.gov> or call +1-202-482-6800.

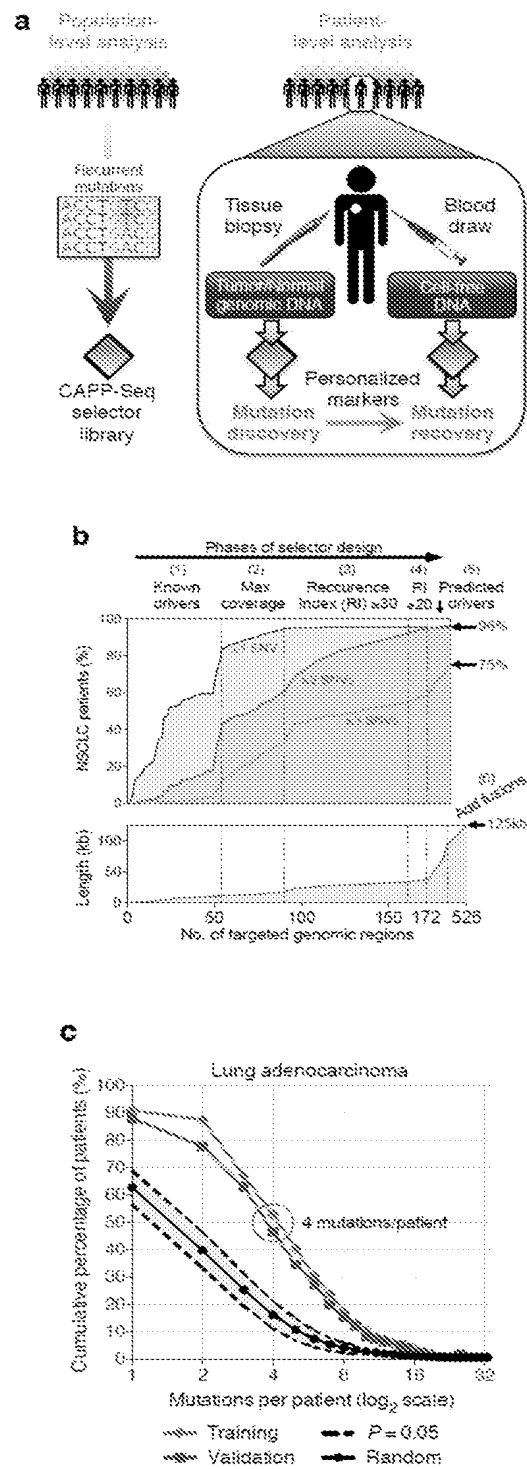


Figure 1

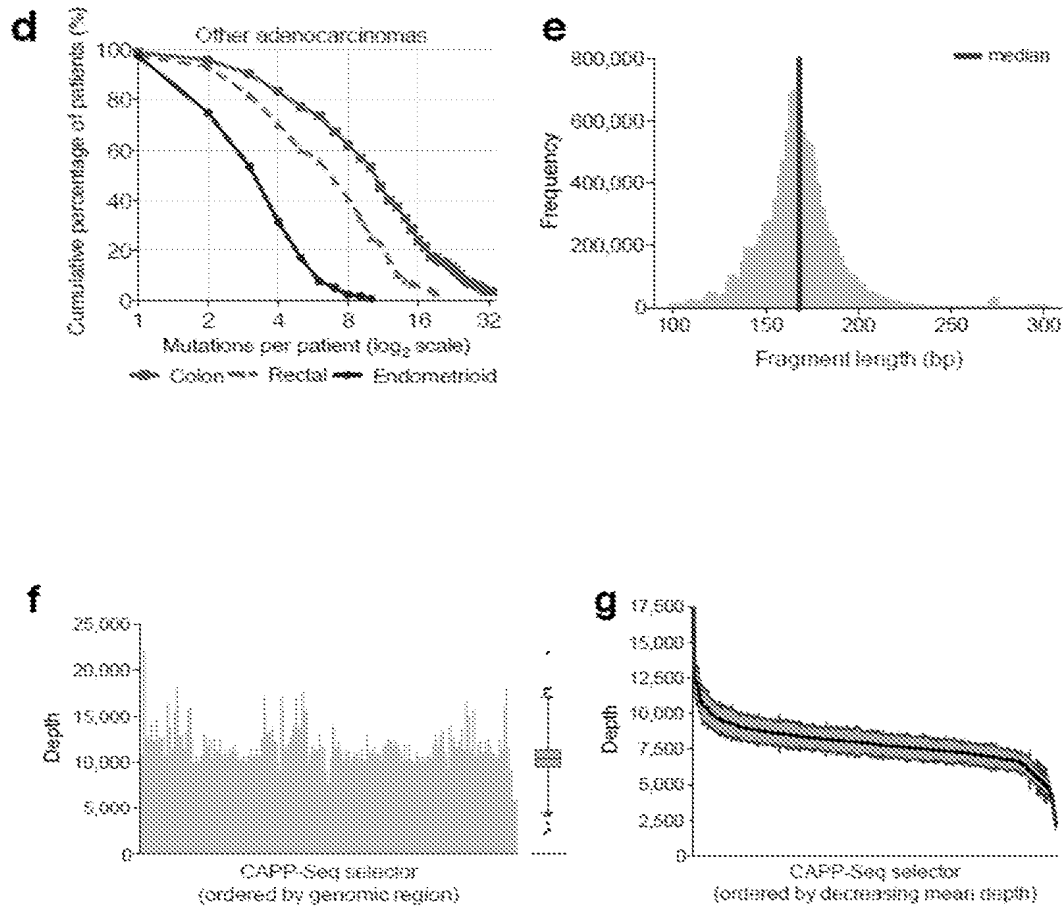


Figure 1 (cont.)

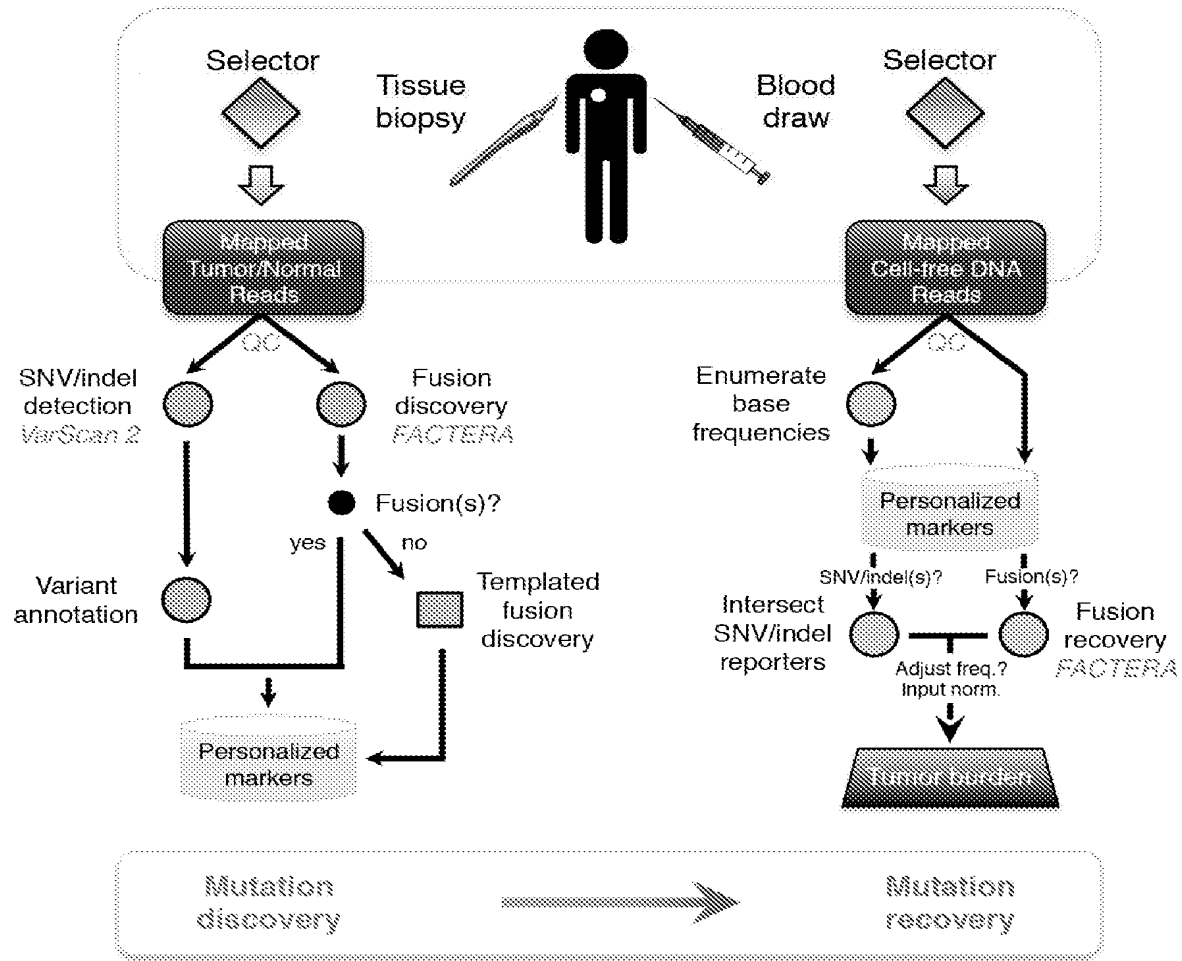


Figure 2

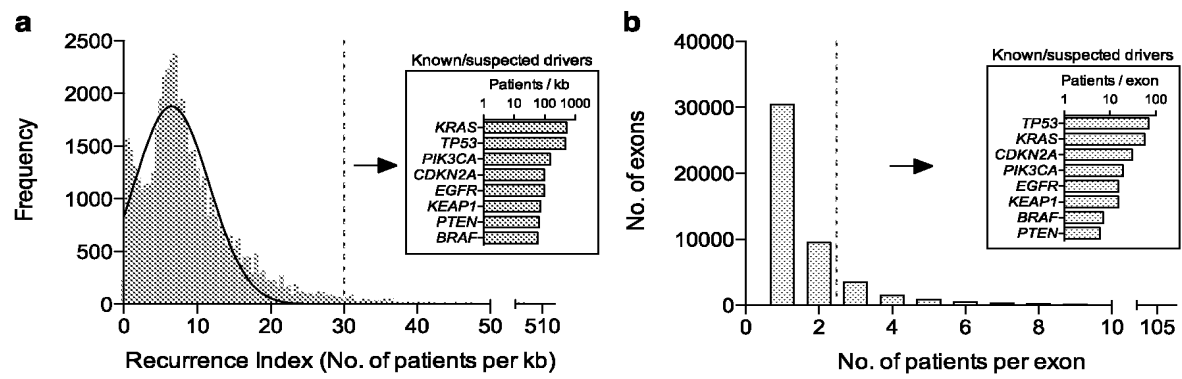


Figure 3

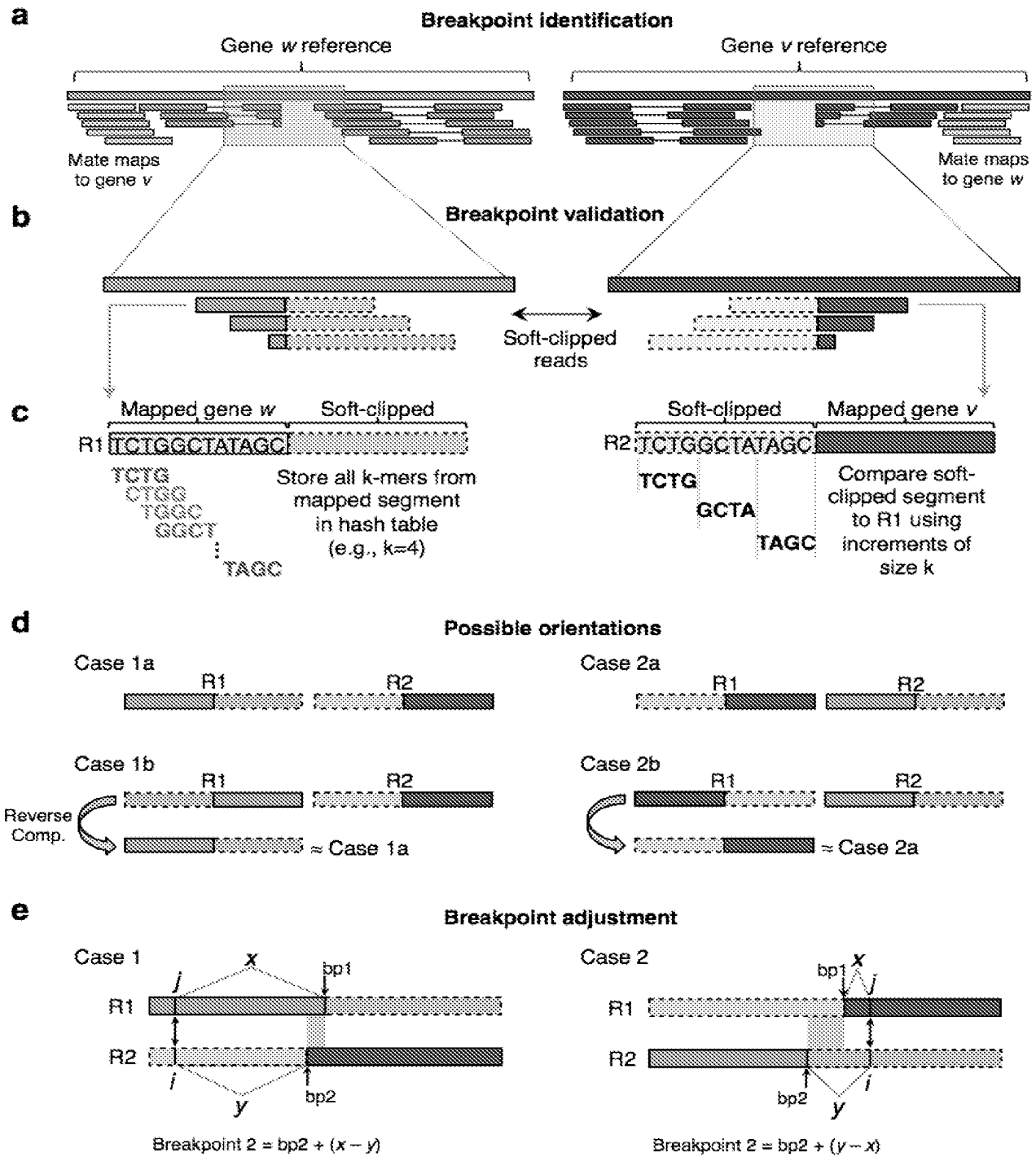


Figure 4

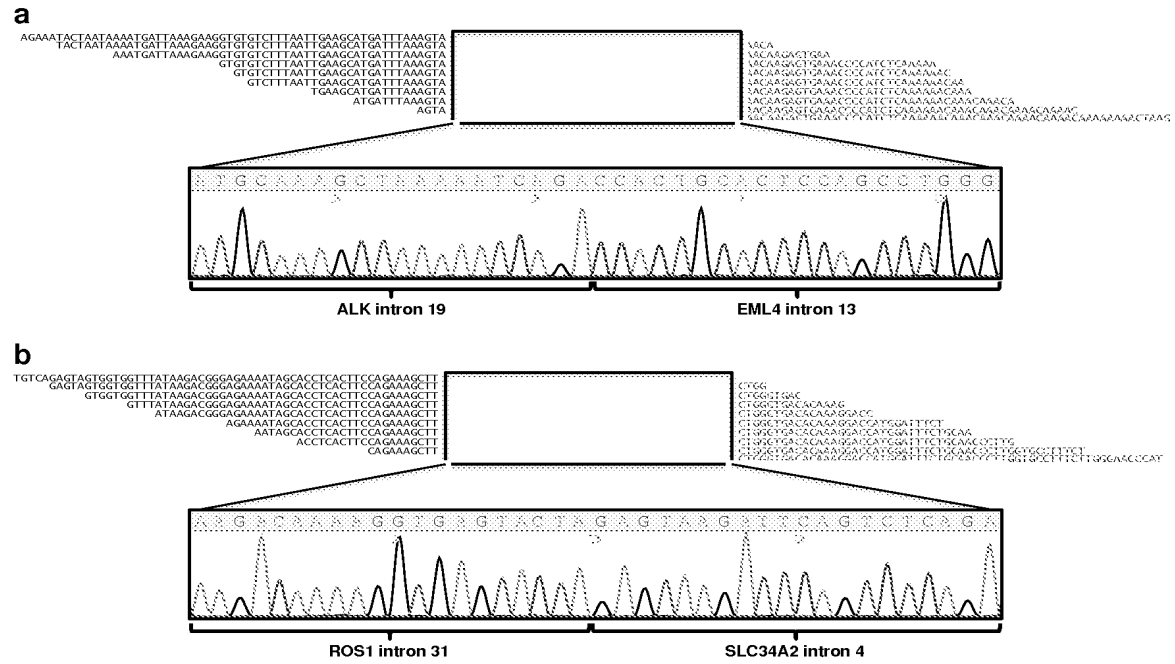


Figure 5

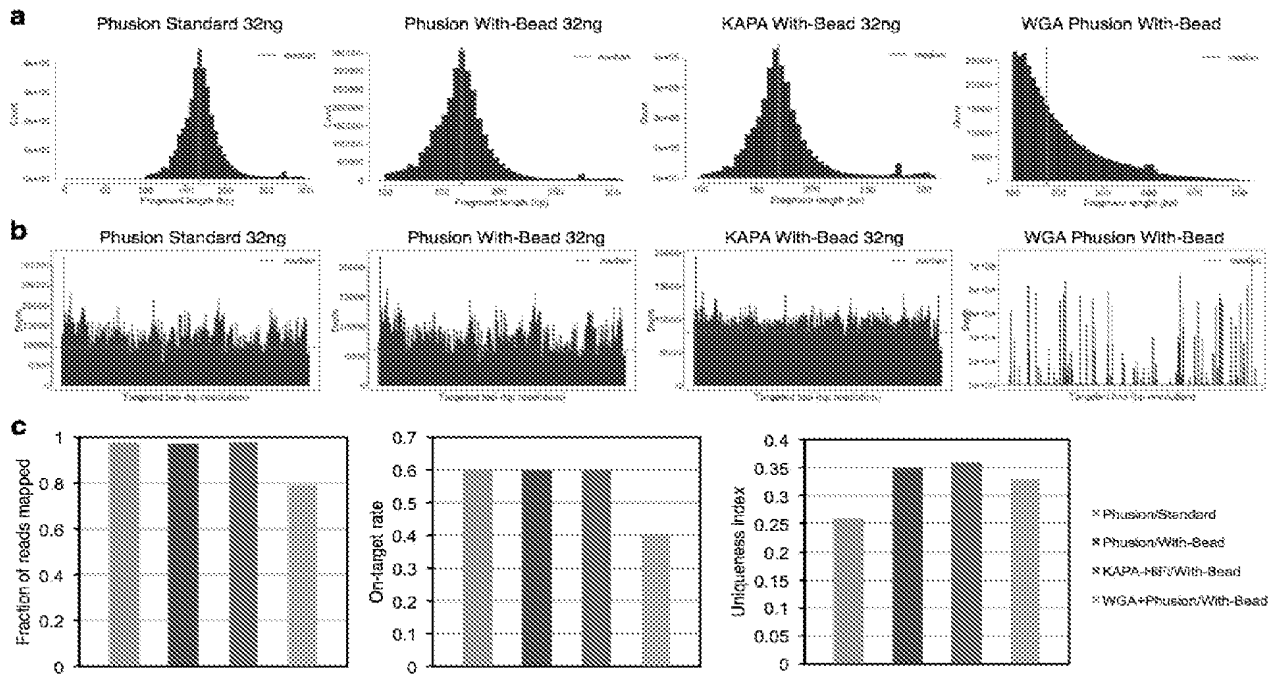


Figure 6

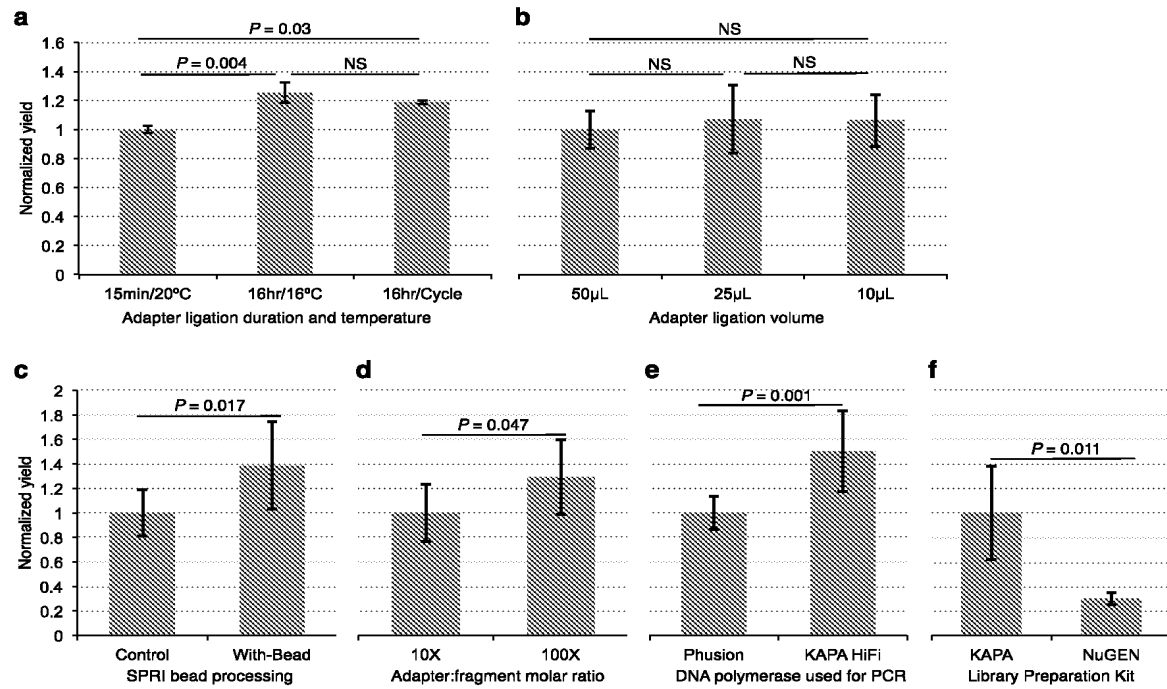


Figure 7

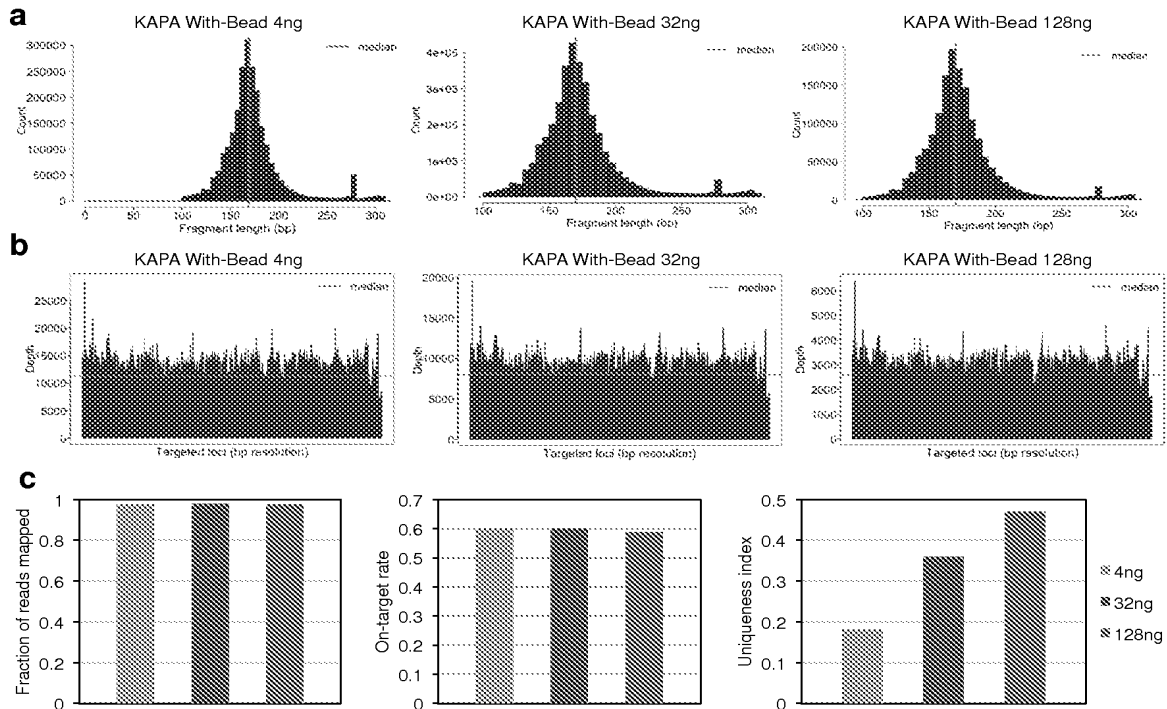


Figure 8

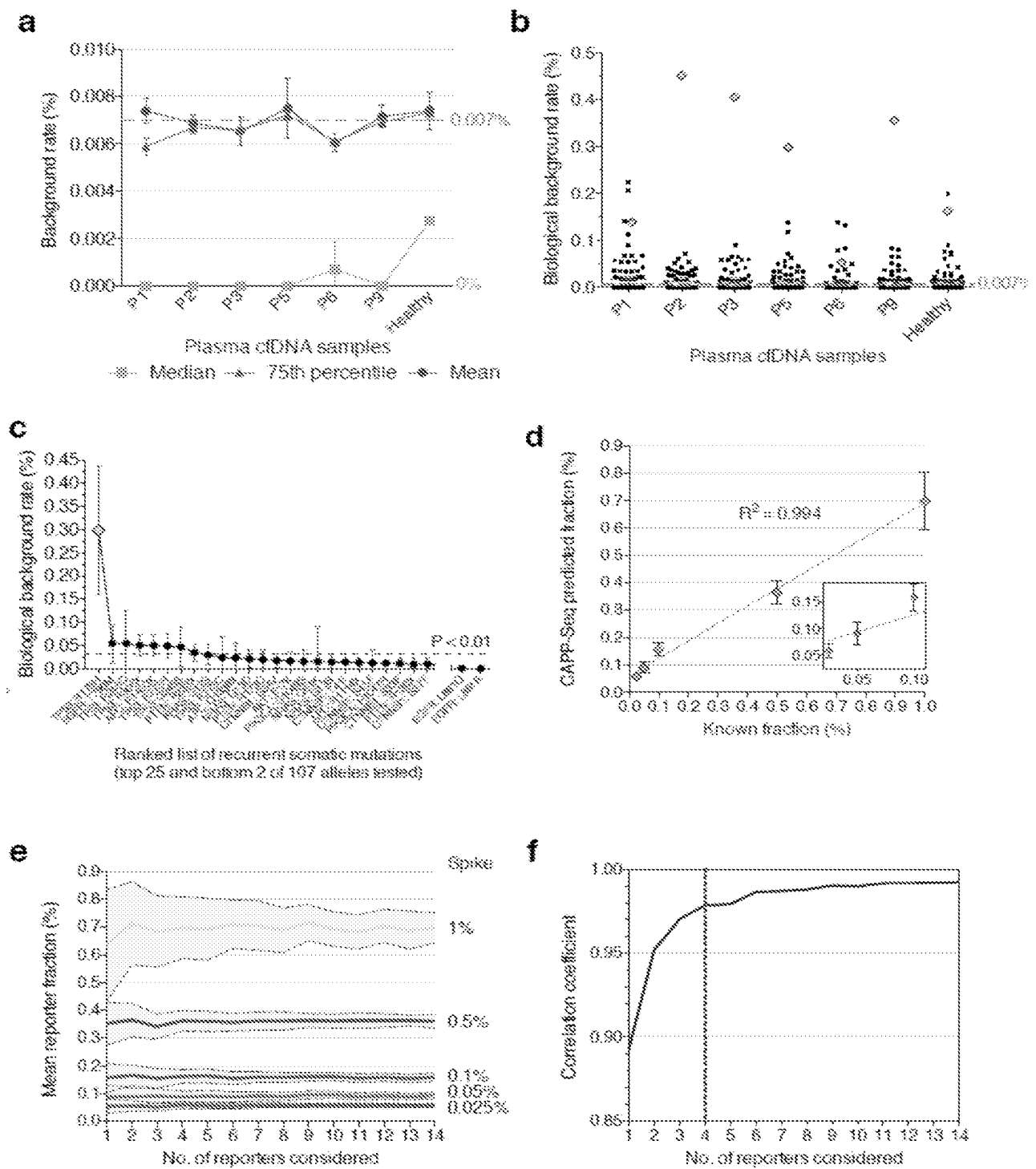


Figure 9

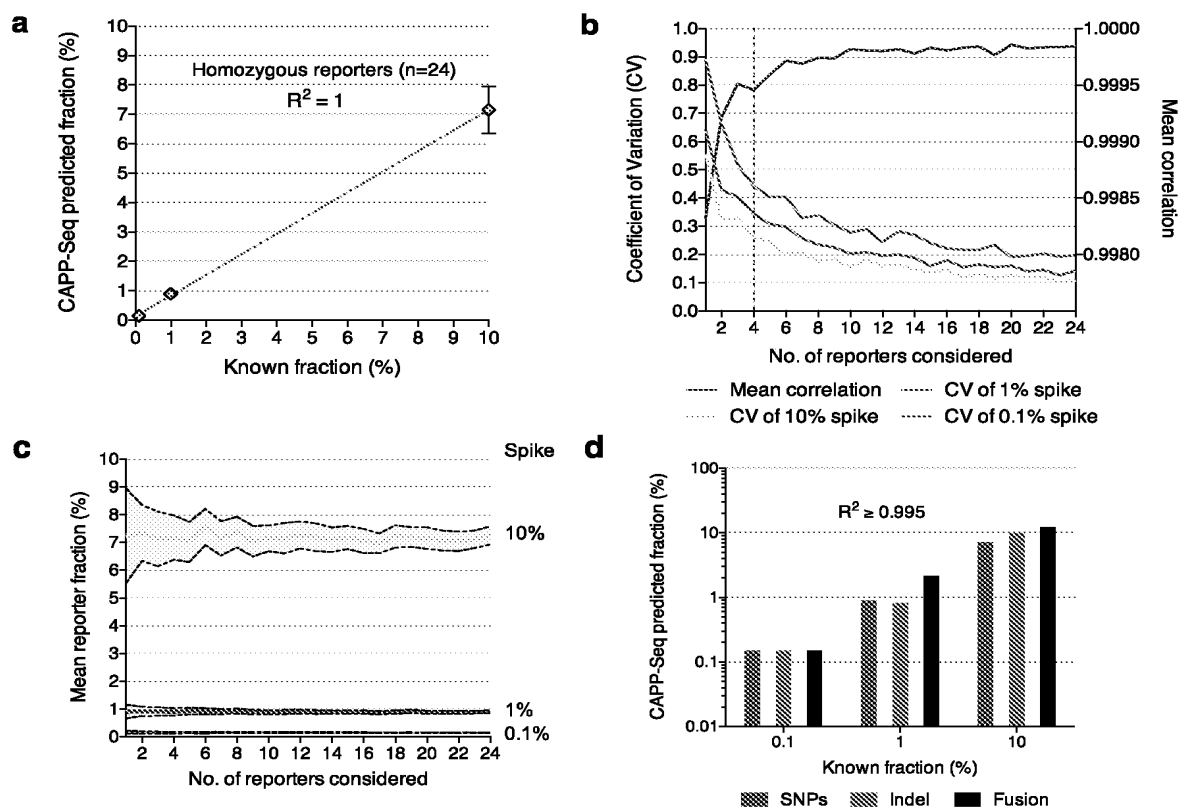


Figure 10

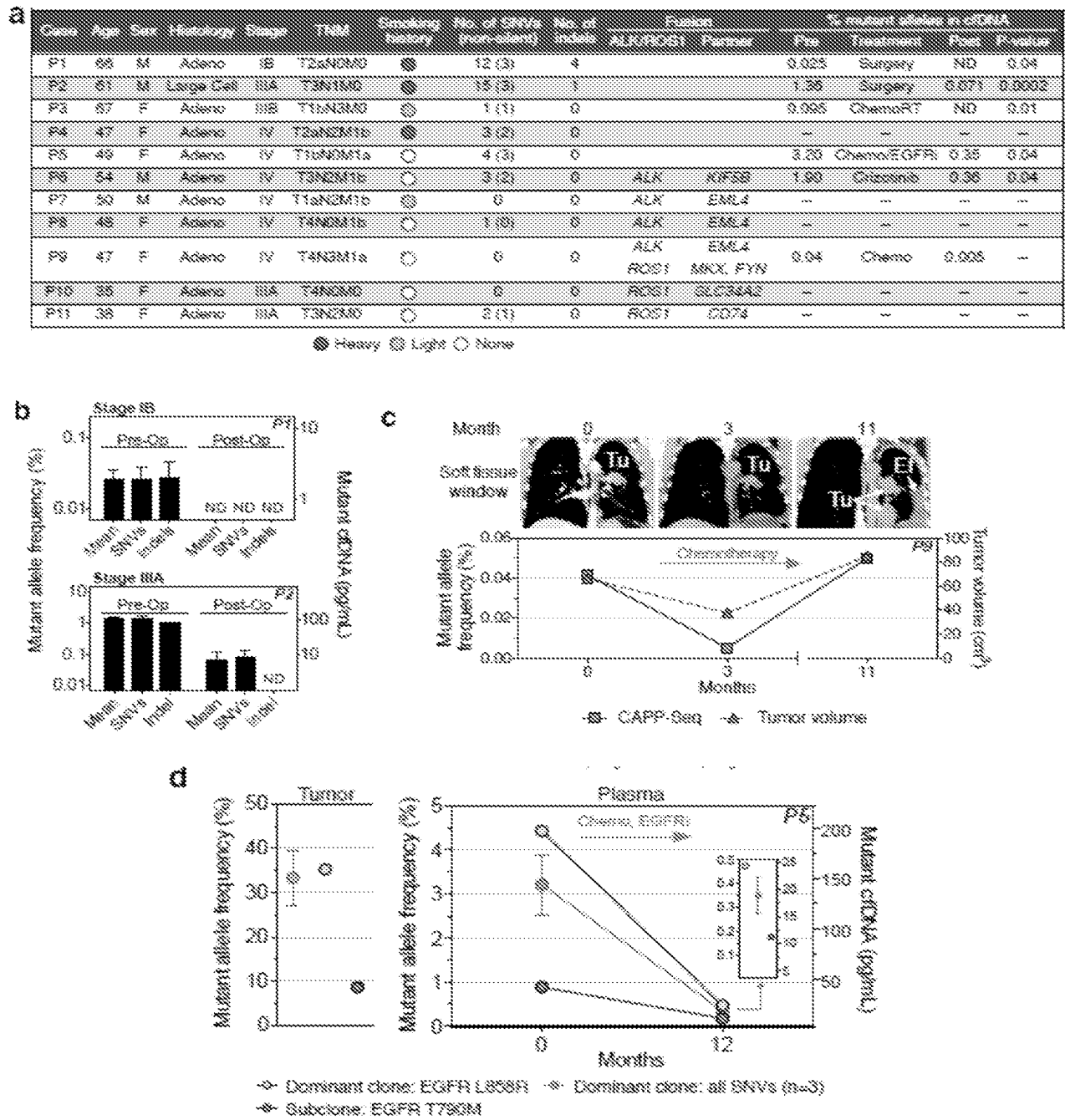


Figure 11

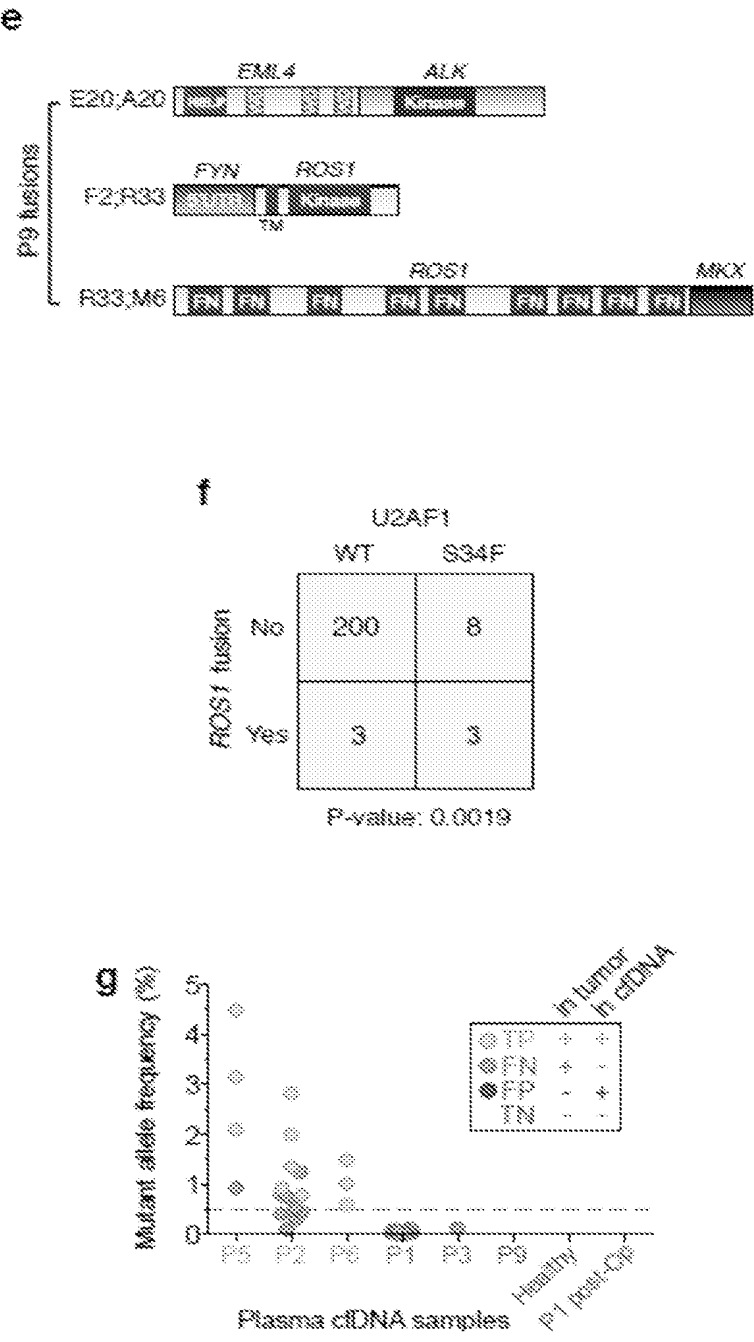


Figure 11 (cont.)

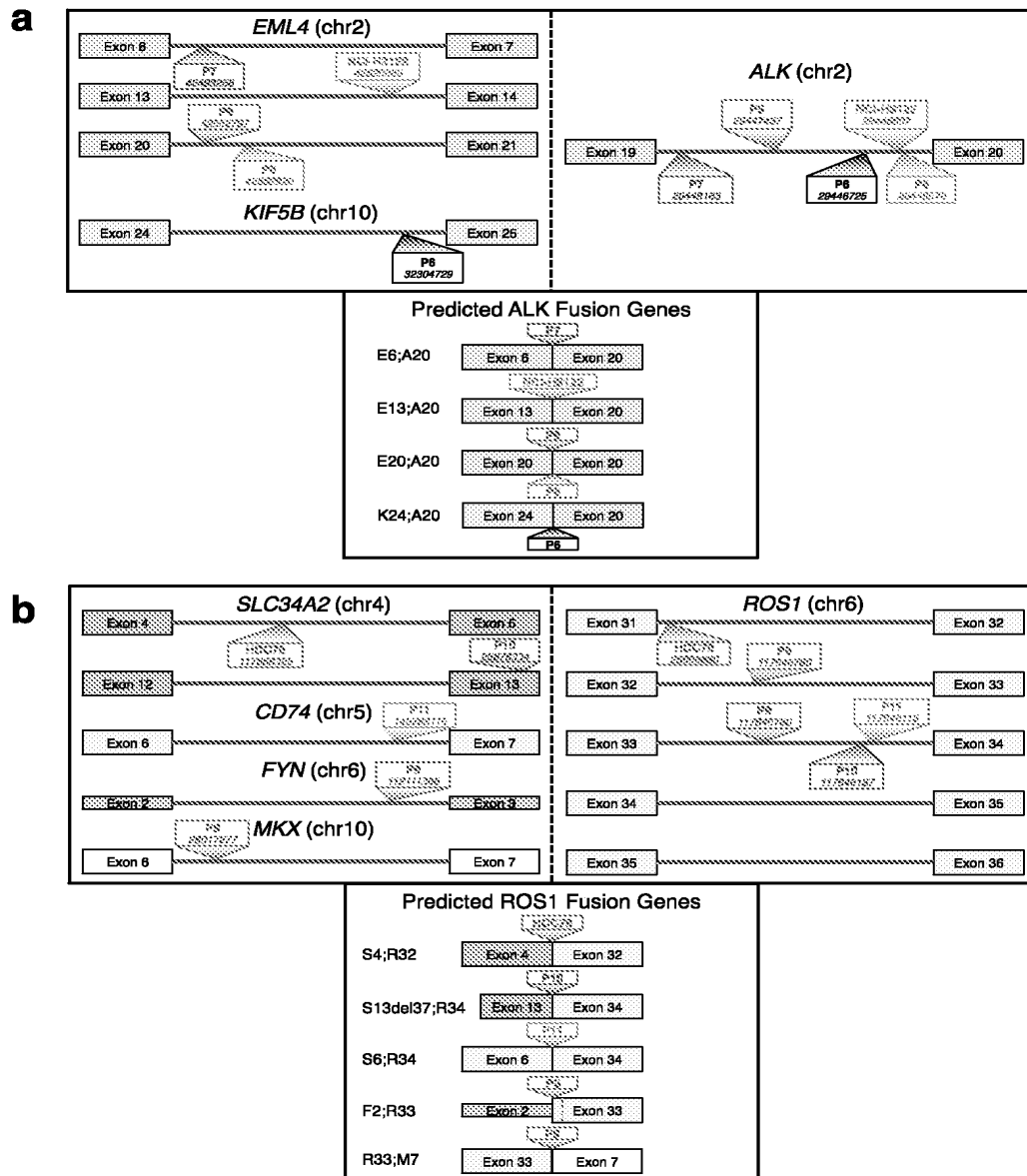
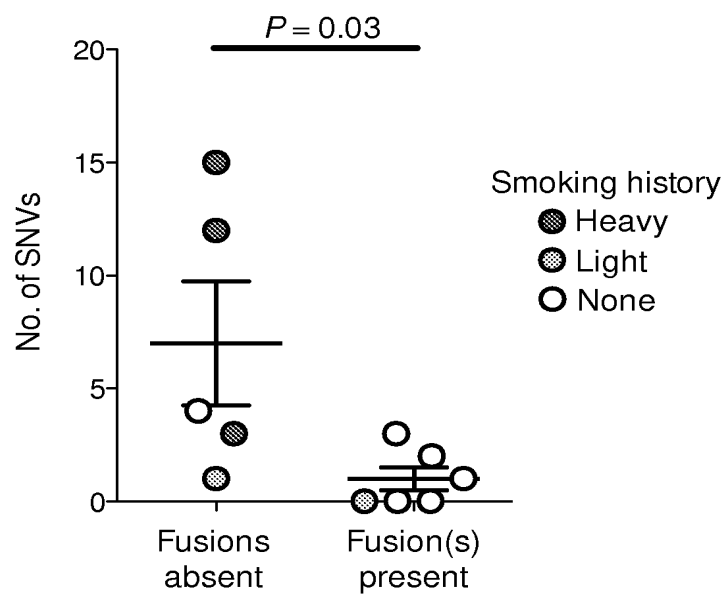


Figure 12

**Figure 13**

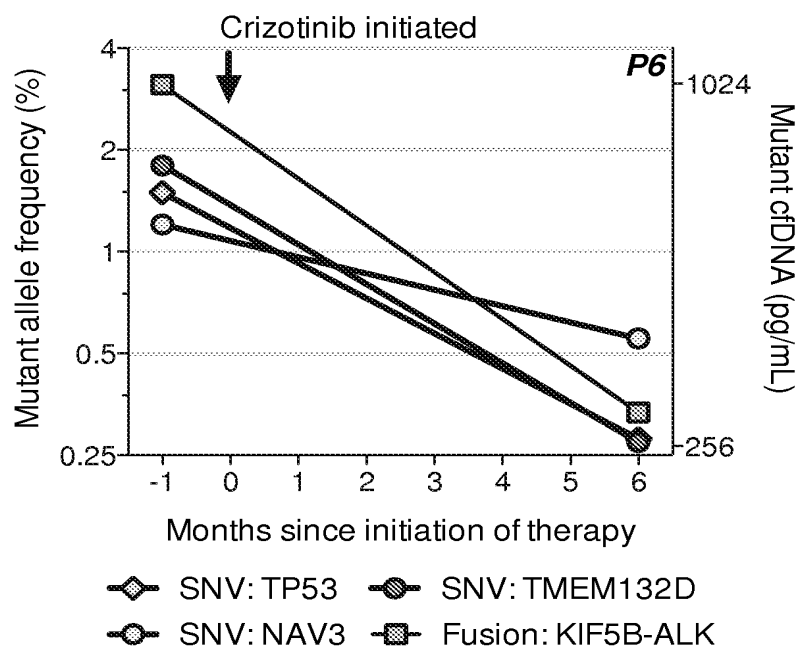


Figure 14

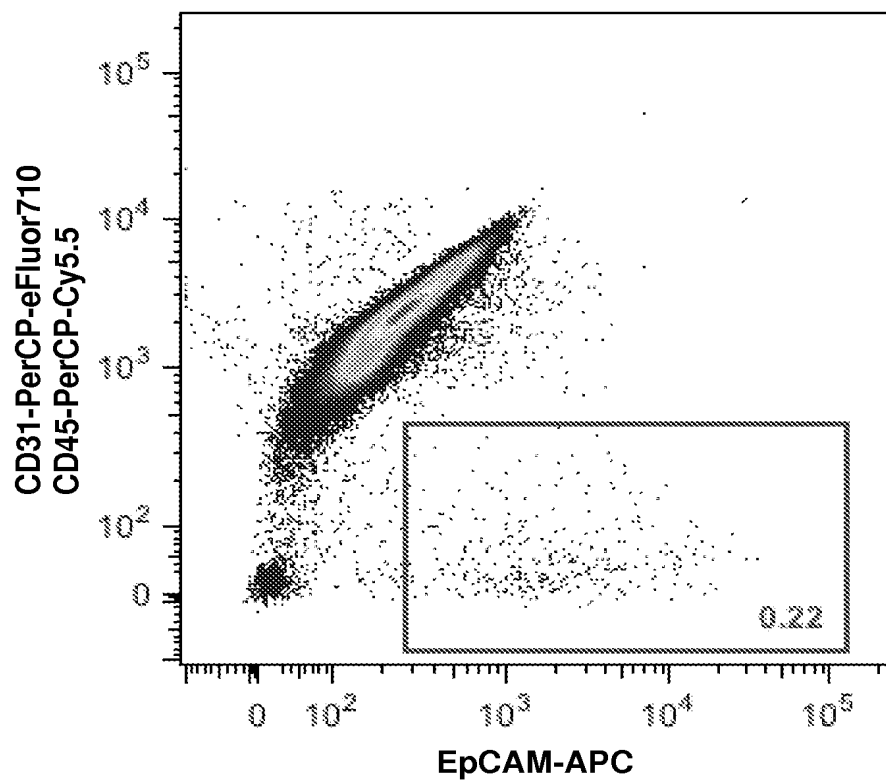


Figure 15

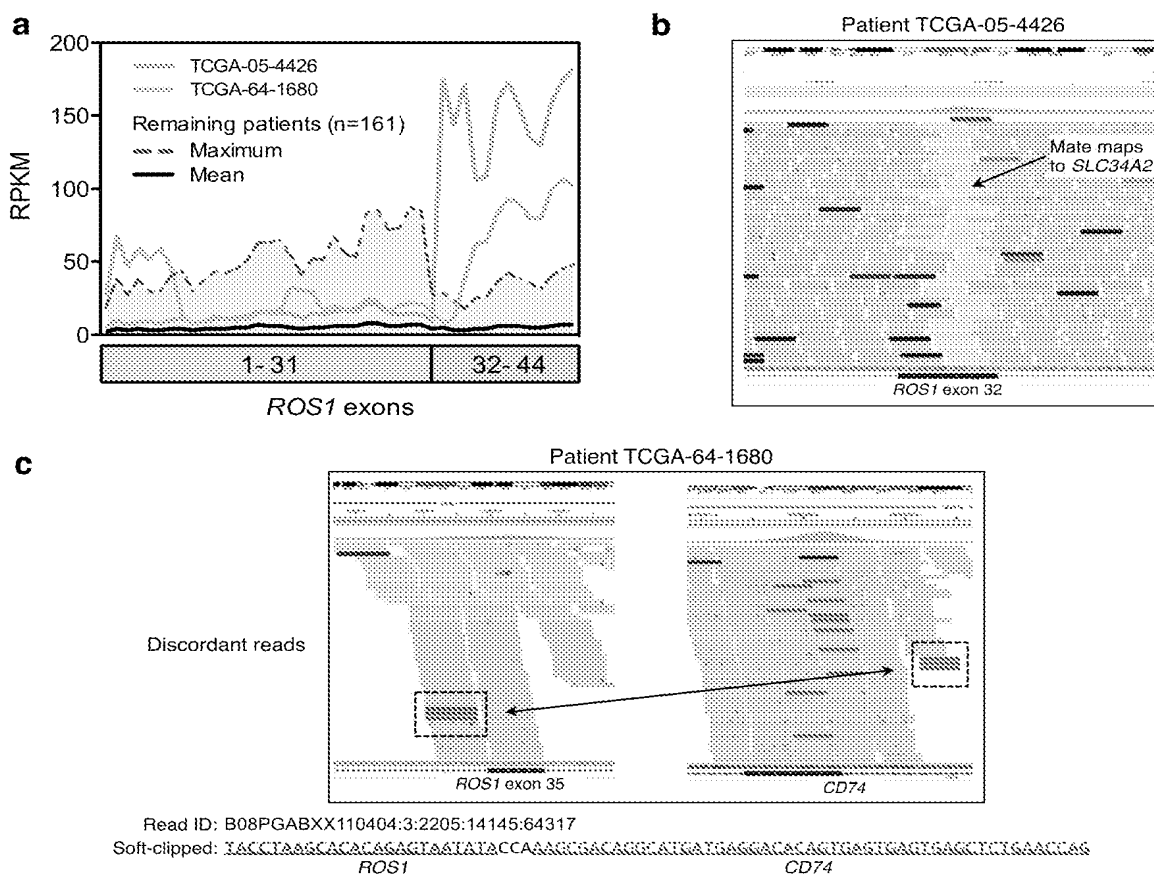


Figure 16

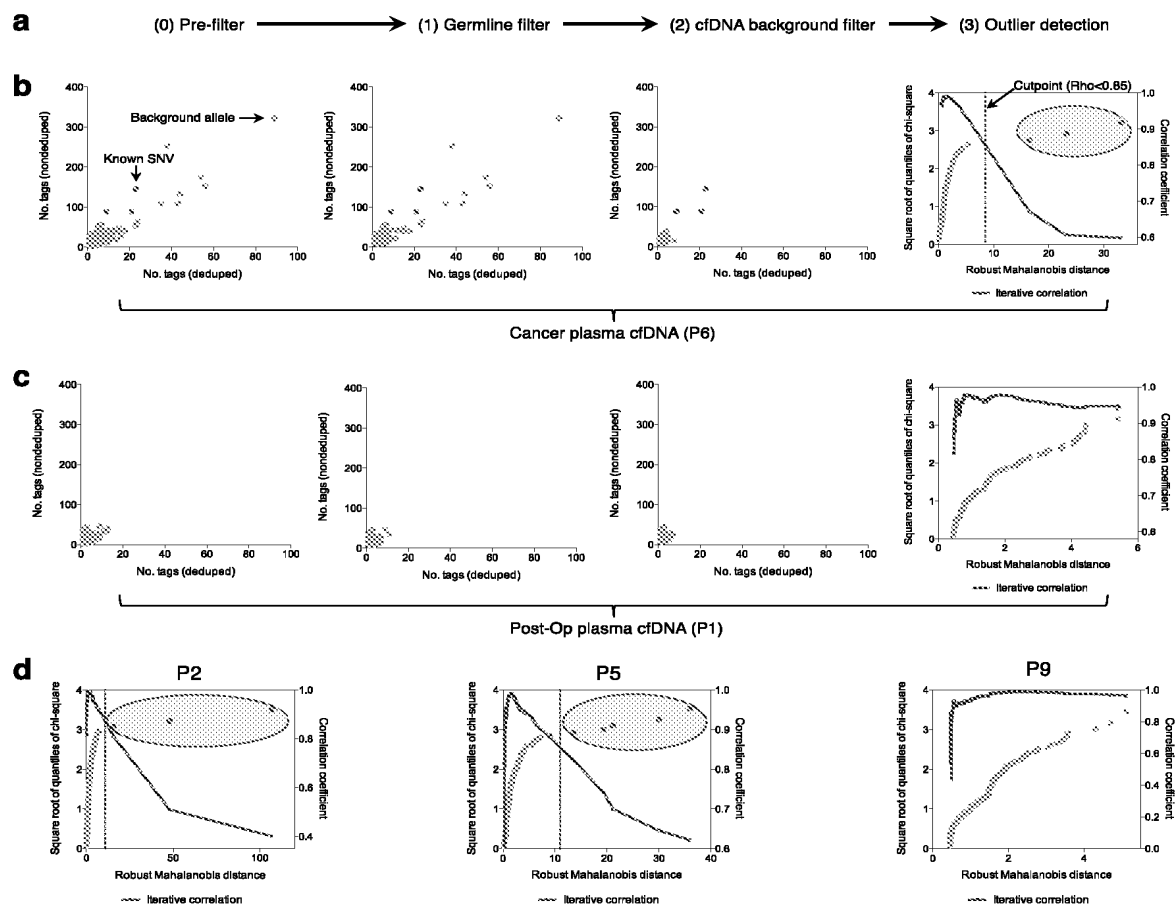


Figure 17

Provisional Application for Patent Cover Sheet

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c)

Inventor(s)

Inventor 1

Remove

Given Name

Middle Name

Family Name

City

State

Country

Maximilian

DIEHN

San Carlos

CA

US

Inventor 2

Remove

Given Name

Middle Name

Family Name

City

State

Country

Arash

A.

Alizadeh

San Mateo

CA

US

Inventor 3

Remove

Given Name

Middle Name

Family Name

City

State

Country

Aaron

M.

Newman

Palo Alto

CA

US

Inventor 4

Remove

Given Name

Middle Name

Family Name

City

State

Country

Scott

V.

Bratman

Palo Alto

CA

US

All Inventors Must Be Listed – Additional Inventor Information blocks may be generated within this form by selecting the **Add** button.

Add

Title of Invention

IDENTIFICATION AND USE OF CIRCULATING TUMOR MARKERS

Attorney Docket Number (if applicable)

S11-230 PR1

Correspondence Address

Direct all correspondence to (select one):

☒ The address corresponding to Customer Number

☐ Firm or Individual Name

Customer Number

86012

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

☒ No.

☐ Yes, the name of the U.S. Government agency and the Government contract number are:

Entity Status

Applicant claims small entity status under 37 CFR 1.27

☒ Yes, applicant qualifies for small entity status under 37 CFR 1.27

☐ No

Warning

Petitioner/applicant is cautioned to avoid submitting personal information in documents filed in a patent application that may contribute to identity theft. Personal information such as social security numbers, bank account numbers, or credit card numbers (other than a check or credit card authorization form PTO-2038 submitted for payment purposes) is never required by the USPTO to support a petition or an application. If this type of personal information is included in documents submitted to the USPTO, petitioners/applicants should consider redacting such personal information from the documents before submitting them to USPTO. Petitioner/applicant is advised that the record of a patent application is available to the public after publication of the application (unless a non-publication request in compliance with 37 CFR 1.213(a) is made in the application) or issuance of a patent. Furthermore, the record from an abandoned application may also be available to the public if the application is referenced in a published application or an issued patent (see 37 CFR 1.14). Checks and credit card authorization forms PTO-2038 submitted for payment purposes are not retained in the application file and therefore are not publicly available.

Signature

Please see 37 CFR 1.4(d) for the form of the signature.

Signature	/David A. Roise/			Date (YYYY-MM-DD)	2013-03-15
First Name	David A.	Last Name	Roise	Registration Number (If appropriate)	47904

This collection of information is required by 37 CFR 1.51. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 8 hours to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **This form can only be used when in conjunction with EFS-Web. If this form is mailed to the USPTO, it may cause delays in handling the provisional application.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that : (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Identification and Use of Circulating Tumor Markers

Statement of Governmental Support

[0001] This invention was made with government support under grant number W81XWH-12-1-0285 awarded by the Department of Defense. The government has
5 certain rights in this invention.

Background of the Invention

[0002] Analysis of cancer-derived cell-free DNA (cfDNA) has the potential to revolutionize detection and monitoring of cancer. Noninvasive access to malignant DNA is particularly attractive for solid tumors, which cannot be repeatedly sampled
10 without invasive procedures. In non-small cell lung cancer (NSCLC), PCR-based assays have been used previously to detect recurrent point mutations in genes such as *KRAS* or *EGFR* in plasma DNA (Taniguchi et al. (2011) *Clin. Cancer Res.* 17:7808-7815; Gautschi et al. (2007) *Cancer Lett.* 254:265-273; Kuang et al. (2009) *Clin. Cancer Res.* 15:2630-2636; Rosell et al. (2009) *N. Engl. J. Med.* 361:958-967), but the
15 majority of patients lack mutations in these genes. Other studies have proposed identifying patient-specific chromosomal rearrangements in tumors via whole genome sequencing (WGS), followed by breakpoint qPCR from cfDNA (Leary et al. (2010) *Sci. Transl. Med.* 2:20ra14; McBride et al. (2010) *Genes Chrom. Cancer* 49:1062-1069). While sensitive, such methods require optimization of molecular assays for
20 each patient, limiting their widespread clinical application. More recently, several groups have reported amplicon-based deep sequencing methods to detect cfDNA

mutations in up to 6 recurrently mutated genes (Forsheew et al. (2012) *Sci. Transl. Med.* 4:136ra168; Narayan et al. (2012) *Cancer Res.* 72:3492-3498; Kinde et al. (2011) *Proc. Natl Acad. Sci. USA* 108:9530-9535). While powerful, these approaches are limited by the number of mutations that can be interrogated (Rachlin et al. (2005) *BMC Genomics* 6:102) and the inability to detect genomic fusions.

[0003] PCT International Patent Publication No. 2011/103236 describes methods for identifying personalized tumor markers in a cancer patient using "mate-paired" libraries. The methods are limited to monitoring somatic chromosomal rearrangements, however, and must be personalized for each patient, thus limiting their applicability and increasing their cost.

[0004] U.S. Patent Application Publication No. 2010/0041048 A1 describes the quantitation of tumor-specific cell-free DNA in colorectal cancer patients using the "BEAMing" technique (Beads, Emulsion, Amplification, and Magnetism). While this technique provides high sensitivity and specificity, this method is for single mutations and thus any given assay can only be applied to a subset of patients and/or requires patient-specific optimization. U.S. Patent Application Publication No. 2012/0183967 A1 describes additional methods to identify and quantify genetic variations, including the analysis of minor variants in a DNA population, using the "BEAMing" technique.

[0005] U.S. Patent Application Publication No. 2012/0214678 A1 describes methods and compositions for detecting fetal nucleic acids and determining the fraction of cell-free fetal nucleic acid circulating in a maternal sample. While sensitive, these methods analyze polymorphisms occurring between maternal and fetal nucleic acids rather than polymorphisms that result from somatic mutations in tumor cells. In addition, methods that detect fetal nucleic acids in maternal circulation require much less sensitivity than methods that detect tumor nucleic acids in cancer patient circulation, because fetal nucleic acids are much more abundant than tumor nucleic acids.

[0006] U.S. Patent Application Publication Nos. 2012/0237928 A1 and 2013/0034546 describe methods for determining copy number variations of a sequence of interest in a test sample comprising a mixture of nucleic acids. While potentially applicable to the analysis of cancer, these methods are directed to measuring major structural changes in nucleic acids, such as translocations, deletions, and amplifications, rather than single nucleotide variations.

[0007] U.S. Patent Application Publication No. 2012/0264121 A1 describes methods for estimating a genomic fraction, for example, a fetal fraction, from polymorphisms such as small base variations or insertions-deletions. These methods do not, however, make use of optimized libraries of polymorphisms, such as, for example, libraries
5 containing recurrently-mutated genomic regions.

[0008] U.S. Patent Application Publication No. 2013/0024127 A1 describes computer-implemented methods for calculating a percent contribution of cell-free nucleic acids from a major source and a minor source in a mixed sample. The methods do not, however, provide any advantages in identifying or making use of optimized
10 libraries of polymorphisms in the analysis.

[0009] PCT International Publication No. WO 2010/141955 A2 describes methods of detecting cancer by analyzing panels of genes from a patient-obtained sample and determining the mutational status of the genes in the panel. The methods rely on a relatively small number of known cancer genes, however, and they do not provide any
15 ranking of the genes according to effectiveness in detection of relevant mutations. In addition, the methods were unable to detect the presence of mutations in the majority of serum samples from actual cancer patients.

[0010] There is thus a need for new and improved methods to detect and monitor tumor-related nucleic acids in cancer patients.
20

Summary of the Invention

[0011] The present invention addresses these and other problems by providing novel methods and systems relating to the characterization, diagnosis, and monitoring of
25 cancer. In particular, according to one aspect, the invention provides methods for creating a library of recurrently mutated genomic regions comprising:

identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer;

wherein the library comprises the plurality of genomic regions;

30 the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0012] In specific embodiments of these methods, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0013] In other specific method embodiments, at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0014] In still other specific method embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0015] In some embodiments, the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0016] In other embodiments, the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0017] In some embodiments, the library comprises a plurality of genomic regions encoding a plurality of driver sequences, more specifically known driver sequences or driver sequences that are recurrently mutated in the specific cancer.

[0018] In some embodiments, the library comprises a plurality of genomic regions that are recurrently rearranged in the specific cancer.

[0019] In preferred embodiments, the specific cancer is a carcinoma, and in more preferred embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0020] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0021] In another aspect, the invention provides methods for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0022] In specific embodiments of this aspect of the invention, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0023] In other specific embodiments, at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0024] In still other specific embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0025] In some embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0026] In other embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0027] In some embodiments, the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences, more specifically known driver sequences or driver sequences that are recurrently mutated in the specific cancer.

[0028] In some embodiments, the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

[0029] In preferred embodiments, the specific cancer is a carcinoma, and in more preferred embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0030] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0031] In some embodiments, the methods further comprising the steps of:
obtaining a cell-free nucleic acid sample from the subject; and
identifying the patient-specific genetic alteration in the cell-free nucleic acid sample.

[0032] In specific embodiments, the step of identifying the patient-specific genetic alteration in the cell-free nucleic acid sample comprises sequencing a genomic region comprising the patient-specific genetic alteration in the cell-free sample.

[0033] In other specific embodiments, the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample comprises the step of enriching the plurality of target regions in the tumor nucleic acid sample and the genomic nucleic acid sample, and in more specific embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

[0034] In still other specific embodiments, the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample, and in still more specific embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

[0035] In some embodiments, the methods further comprise the step of quantifying the cancer-specific genetic alteration in the cell-free sample.

[0036] In yet another aspect, the invention provides methods for screening a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a cell-free nucleic acid sample from a subject;

sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and

identifying a cancer-specific genetic alteration in the cell-free sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0037] In specific embodiments, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

[0038] In other specific embodiments, at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0039] In still other specific embodiments, at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

[0040] In particular embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

[0041] In other particular embodiments, each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

[0042] In still other particular embodiments, the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences, and, more particularly, the driver sequences are known driver sequences or are recurrently mutated in the specific cancer.

[0043] In yet still other particular embodiments, the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

[0044] In some embodiments, the specific cancer is a carcinoma, including, for example, an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

[0045] In specific embodiments, the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

[0046] In other specific embodiments, the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample, and, in some embodiments, the enriching step comprises use of a custom library of biotinylated DNA.

Brief Description of the Drawings

[0047] Figure 1. Development of Cancer Personalized Profilng by Deep Sequencing (CAPP-Seq). (a) Schematic depicting design of CAPP-Seq selectors and their application for assessing circulating tumor DNA. (b) Multi-phase design of the NSCLC CAPP-Seq selector. (c) Analysis of the number of SNVs per lung adenocarcinoma covered by the NSCLC CAPP-Seq selector in the TCGA WES cohort (*Training*; N=229) and an independent lung adenocarcinoma WES data set (*Validation*; N=183) (Imielinski et al. (2012) *Cell* 150:1107-1120). (d) Number of SNVs per patient identified by the NSCLC CAPP-Seq selector in WES data from three adenocarcinomas from TCGA, colon (COAD), rectal (READ), and endometrioid (UCEC) cancers. (e-f) Quality parameters from a representative CAPP-Seq analysis of plasma cfDNA, including length distribution of sequenced cfDNA fragments (e), and depth of sequencing coverage across all genomic regions in the selector (f). (g) Variation in sequencing depth across cfDNA samples from 4 patients.

[0048] Figure 2. CAPP-Seq computational pipeline. Major steps of the bioinformatics pipeline for mutation discovery and quantitation in plasma are schematically illustrated.

[0049] Figure 3. Statistical enrichment of recurrently mutated NSCLC exons captures known drivers.

[0050] Figure 4. Development of the FACTERA algorithm. Major steps used by FACTERA (see Detailed Methods) to precisely identify genomic breakpoints from aligned paired-end sequencing data are anecdotally illustrated using two hypothetical

genes, w and v . (a) Improperly paired, or "discordant," reads (indicated in yellow) are used to locate genes involved in a potential fusion (in this case, w and v). (b) Because truncated (i.e., soft-clipped) reads may indicate a fusion breakpoint, any such reads within genomic regions delineated by w and v are also further analyzed. (c) Consider soft-clipped reads, R1 and R2, whose non-clipped segments map to w and v , respectively. If R1 and R2 derive from a fragment encompassing a true fusion between w and v , then the mapped portion of R1 should match the soft-clipped portion of R2, and vice versa. This is assessed by FACTERA using fast k-mer indexing and comparison. (d) Four possible orientations of R1 and R2 are depicted. However, only Cases 1a and 2a can generate valid fusions (see Detailed Methods). Thus, prior to k-mer comparison (panel c), the reverse complement of R1 is taken for Cases 1b and 2b, respectively, converting them into Cases 1a and 2a. (e) In some cases, short sequences immediately flanking the breakpoint are identical, preventing unambiguous determination of the breakpoint. Let iterators i and j denote the first matching sequence positions between R1 and R2. To reconcile sequence overlap, FACTERA arbitrarily adjusts the breakpoint in R2 (i.e., bp2) to match R1 (i.e., bp1) using the sequence offset determined by differences in distance between bp2 and i , and bp1 and j . Two cases are illustrated, corresponding to sequence orientations described in (d).

[0051] Figure 5. Application of FACTERA to NSCLC cell lines NCI-H3122 and HCC78, and Sanger-validation of breakpoints. (a) Pile-up of a subset of soft-clipped reads mapping to the *EML4-ALK* fusion identified in NCI-H3122 along with the corresponding Sanger chromatogram. (b) Same as (a), but for the *SLC34A2-ROS1* translocation identified in HCC78.

[0052] Figure 6. Improvements in CAPP-Seq performance with optimized library preparation procedures.

[0053] Figure 7. Optimizing allele recovery from low input cfDNA during Illumina library preparation.

[0054] Figure 8. CAPP-Seq performance with various amounts of input cfDNA.

[0055] Figure 9. Analysis of CAPP-Seq background, allele detection threshold, and linearity. (a) Analysis of background rate for 6 NSCLC patient plasma samples and a healthy individual (Detailed Methods). (b) Analysis of biological background in (a) focusing on 107 recurrent somatic mutations from a previously reported SNaPshot

panel (Su et al. (2011) *J. Mol. Diagn.* 13:74-84). Mutations found in a given patient's tumor were excluded. The mean frequency for each patient (horizontal red line) was within confidence limits of the mean background limit of 0.007% (horizontal blue line; panel a). A single outlier mutation (*TP53* R175H) is indicated by an orange diamond.

5 (c) Individual mutations from (b) ranked by most to least recurrent, according to median frequency across the 7 samples. (d) Dilution series analysis of expected versus observed frequencies of mutant alleles using CAPP-Seq. Dilution series were generated by spiking fragmented HCC78 DNA into control cfDNA. (e) Analysis of the effect of the number of SNVs considered on the estimates of fractional abundance (95%
10 confidence intervals shown in gray). (f) Analysis of the effect of the number of SNVs considered on the mean correlation coefficient between expected and observed cancer fractions (blue dashed line) using data from panel (d). 95% confidence intervals are shown for (a)-(c). Statistical variation for (d) is shown as s.e.m.

[0056] Figure 10. Empirical spiking analysis of CAPP-Seq using two NSCLC cell
15 lines. (a) Expected and observed (by CAPP-Seq) fractions of NCI-H3122 DNA spiked into control HCC78 DNA are linear for all fractions tested (0.1%, 1%, and 10%; $R^2 = 1$). (b) Using data from (a), analysis of the effect of the number of SNVs considered on the estimates of fractional abundance (95% confidence intervals shown in gray). (c)
20 Analysis of the effect of the number of SNVs considered on the mean correlation coefficient and coefficient of variation between expected and observed cancer fractions (dashed lines) using data from panel (a). (d) Expected and observed fractions of the *EML4-ALK* fusion present in HCC78 are linear ($R^2 = 0.995$) over all spiking concentrations tested (see Figure 5(b) for breakpoint verification). The observed
25 *EML4-ALK* fractions were normalized based on the relative abundance of the fusion in 100% H3122 DNA (see Detailed Methods for details). Moreover, a single heterozygous insertion (indel) discovered within the selector space of NCI-H3122 (chr7: 107416855, +T) was concordant with defined concentrations (shown are observed fractions adjusted for zygosity).

[0057] Figure 11. Application of CAPP-Seq for noninvasive detection and
30 monitoring of circulating tumor DNA. (a) Characteristics of 11 patients included in this study (Table 3). P-values reflect a two-sided paired t-test for patients with reporter SNVs detected at both time points; other p-values were determined as described in

Methods. ND, mutant DNA was not detected above background. Dashes, plasma sample not available. Smoking history, ≥ 20 pack years (heavy), >0 pack years (light). (b-d) Disease monitoring using CAPP-Seq. Mutant allele frequencies (left y-axis) and absolute concentrations (right y-axis) are shown. The lower limit of detection (defined in Figure 2(a)-(b)) is indicated by the dashed lines. (b) Pre- and post-surgery circulating tumor DNA levels quantified by CAPP-Seq in a Stage IB and a Stage IIIA NSCLC patient. Complete resections were achieved in both cases. (c) Disease burden changes in response to chemotherapy in a Stage IV NSCLC patient with three rearrangement breakpoints identified by CAPP-Seq. Tumor volume based on CT measurements and CAPP-Seq mutant allele frequencies are shown. Tu, tumor; Ef, pleural effusion. (d) Detection and monitoring of a subclonal *EGFR* T790M resistance mutation in a patient with Stage IV NSCLC. The fractional abundance of the dominant clone and T790M-containing clone are shown in the primary tumor (left) and plasma samples (right). (e) Predicted transcripts of three fusion genes detected in case P9. (f) Statistically significant co-occurrence of *ROS1* fusions and *U2AF1* S34F mutations in NSCLC ($P = 0.0019$; two-sided Fisher's exact test). (g) Exploratory analysis of the potential application of CAPP-Seq for cancer screening. Pre-treatment plasma samples from panel (a) and a plasma sample from a healthy individual were examined for the presence of mutant allele outliers without knowledge of the primary tumor mutations (see Detailed Methods). Error bars represent s.e.m.

[0058] Figure 12. Base-pair resolution breakpoint mapping for all patients and cell lines enumerated by FACTERA. Gene fusions involving ALK (a) and ROS1 (b) are graphically depicted. Schematics in the top panels indicate the exact genomic positions (HG19 NCBI Build 37.1/GRCh37) of the breakpoints in *ALK*, *ROS1*, *EML4*, *KIF5B*, *SLC34A2*, *CD74*, *MKX*, and *FYN*. Bottom panels depict exons flanking the predicted gene fusions with notation indicating the 5' fusion partner gene and last fused exon followed by the 3' fusion partner gene and first fused exon. For example, in S13del37;R34 exons 1-13 of *SLC34A2* (excluding the 3' 37 nucleotides of exon 13) are fused to exons 34-43 of *ROS1*. Exons in *FYN* are from its 5'UTR and precede the first coding exon. The green dotted line in the predicted *FYN-ROS1* fusion indicates the first in-frame methionine in *ROS1* exon 33, which preserves an open reading frame

encoding the ROS1 kinase domain. All rearrangements were each independently confirmed by PCR and/or FISH.

[0059] Figure 13. Presence of fusions is inversely related to the number of SNVs detected by CAPP-Seq. For each patient listed in Figure 11(a) the number of identified SNVs versus the presence or absence of detected genomic fusions are plotted. The shading of the symbols is identical to Figure 11(a), and indicates smoking history. Statistical significance was determined using a two-sided Wilcoxon rank sum test, and error bars indicate s.e.m.

[0060] Figure 14. Different types of reporters are similarly useful for disease monitoring. Three SNVs and an ALK translocation identified in patient 6 are concordant at each time point, showing a comparable drop in fractional abundance after treatment with the ALK kinase inhibitor Crizotinib. Due to small differences in measured allele frequencies at each time point, linear regression was used to fit all allele frequencies to their adjusted mutant cfDNA concentrations ($R^2 = 0.93$). Thus, the scale on the right y-axis is interpolated. To accurately quantify disease burden, translocation and SNV frequencies were adjusted based on differences in zygosity and sequencing depth in the tumor sample (see Detailed Methods).

[0061] Figure 15. Flow cytometry-analysis of P9 pleural effusion. Flow cytometry of cryopreserved cells from a pleural effusion revealed only 0.22% of cells stained positive for the epithelial marker, EpCAM, and negative for the lineage markers CD31 (endothelial cells) and CD45 (immune cells). FACS was used to enrich tumor cells and analysis of tumor-enriched genomic DNA identified 3 fusions (Figure 11(e)), while unsorted low purity tumor specimen hampered *de novo* fusion discovery using FACTERA (Detailed Methods).

[0062] Figure 16. Analysis of RNA-Seq data from lung adenocarcinoma patients in TCGA identifies 2 candidate cases with *ROS1* rearrangements. (a) *ROS1* fusions are known to result in over-expression of the C-terminal kinase domain, and breakpoints typically occur downstream of exon 31 (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870; Rikova et al. (2007) *Cell* 131:1190-1203; Takeuchi et al. (2012) *Nat. Med.* 18:378-381). Exon-level RPKM values for *ROS1* are plotted for 163 LUAD patients. Two patients (TCGA-05-4426 and TCGA-64-1680) have expression patterns suggestive of *ROS1* fusions. (b,c) Pileups of RNA-Seq reads in these two patients

illustrate an abundance of reads mapping to regions surrounding *ROS1* exon boundaries. Colored reads indicate discordant pairs, consistent with *ROS1* fusions. Such pairs map to *SLC34A2* for patient TCGA-05-4426 (b) and *CD74* for patient TCGA-64-1680 (c). A single soft-clipped RNA-Seq read supports a *ROS1-CD74* fusion event in TCGA-64-1680.

[0063] Figure 17. Non-invasive cancer screening with CAPP-Seq, related to Figure 11(g). (a) Steps to identify candidate SNVs in plasma cfDNA demonstrated using a patient sample with NSCLC (P6, see Table 3). Following stepwise filtration, outlier detection is applied (Detailed Methods). (b) Same as (a), but using a plasma cfDNA sample from a patient who had their tumor surgically removed. No SNVs are identified, as expected. (c) Three additional representative samples applying retrospective screening to patients analyzed in this study. P2 and P5 samples have confirmed tumor-derived SNVs, while P9 is cancer positive but lacks tumor-derived SNVs. Red points, confirmed tumor-derived SNVs; Green points, background noise.

Detailed Description of the Invention

[0064] Tumors continually shed DNA into the circulation, where it is readily accessible. Stroun et al. (1987) *Eur J Cancer Clin Oncol* 23:707-712. Provided herein are methods for the ultrasensitive detection of circulating tumor DNA called Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq). Also provided are methods for creating libraries of recurrently mutated genomic regions used in the CAPP-Seq methods. CAPP-Seq targets hundreds of recurrently mutated genomic regions and simultaneously detects point mutations, insertions/deletions, and rearrangements. CAPP-Seq for non-small cell lung cancer has been demonstrated herein with a design that identified mutations in >95% of tumors. CAPP-Seq accurately quantified circulating tumor DNA from early and advanced stage tumors and identified mutant alleles down to 0.025% with a detection limit of <0.01%. Tumor-derived DNA levels paralleled clinical responses to diverse therapies and CAPP-Seq identified actionable mutations in plasma. Moreover, CAPP-Seq identified significant co-occurrence of *ROS1* translocations with *U2AF1* splicing factor mutations. Finally, the utility of CAPP-Seq for cancer screening is also described. CAPP-Seq can be routinely applied to noninvasively detect and monitor tumors, thus facilitating personalized cancer therapy.

Methods for Creating Libraries

[0065] According to one aspect of the invention, methods for creating a library of recurrently mutated genomic regions are provided. The methods comprise the step of identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer, wherein the library comprises the plurality of genomic regions, the plurality of genomic regions comprises at least 10 different genomic regions, and at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

[0066] It should be understood that the term "library" represents a compilation or collection of individual components. Thus, a library of recurrently mutated genomic regions is a compilation or collection of recurrently mutated genomic regions. The libraries of the instant disclosure are useful because they include a large number of potentially mutated genomic regions within a minimal length of genomic sequence. Use of these libraries to identify genetic alternations in specific patient samples is particularly advantageous because the libraries do not need to be optimized on a patient-by-patient basis.

[0067] The libraries created according to the instant methods comprise genomic regions that are recurrently mutated in a specific cancer. The identification of these recurrent mutations benefits greatly from the availability of databases such as, for example, The Cancer Genome Atlas (TCGA) and its subsets (<http://cancergenome.nih.gov/>). Such databases serve as the starting point for identifying the recurrently mutated genomic regions of the instant libraries. The databases also provide a sample of mutations occurring within a given percentage of subjects with a specific cancer.

[0068] The libraries created according to the instant methods comprise a plurality of genomic regions, wherein the plurality of genomic regions comprises at least 10 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, at least 500, or even more different genomic regions.

[0069] It should be understood that the inclusion of larger numbers of genomic regions generally increases the likelihood that a unique mutation will be identified to distinguish tumor nucleic acid in a subject from the subject's genomic nucleic acid.

Including too many genomic regions in the library is not without a cost, however, since the number of genomic regions is directly related to the length of nucleic acids that must be sequenced in the analysis. At the extreme, the entire genome of a tumor sample and a genomic sample could be sequenced, and the resulting sequences could be compared to note any differences. Such a brute force approach is not possible, however, with the vanishingly small quantities of tumor nucleic acid present in a cell-free sample.

[0070] The libraries of the instant disclosure address this problem by identifying genomic regions that are recurrently mutated in a particular cancer, and then ranking those regions to maximize the likelihood that the region will include a distinguishing genetic alteration in a particular tumor. The library of recurrently mutated genomic regions, or "selectors", can be used across an entire population for a given cancer, and does not need to be optimized for each subject.

[0071] The term "mutation", as used herein, refers to a genetic alteration in the genome of an organism, specifically to a change in the nucleotide sequence of the organism. Examples of mutations include point mutations, where a single nucleotide is changed in the genome, and larger-scale changes in the genome, such as rearrangements, insertions, deletions, and amplifications. A recurrent mutation is a mutation that has been identified in more than one individual.

[0072] The terms "patient" and "subject" are used interchangeably. These are typically individuals that suffer from the cancer of interest. While the individuals are typically human individuals, the methods and systems of the instant disclosure could also be applied to other species, in particular, to other animal species, for example, livestock animals and pets.

[0073] The libraries of recurrently mutated genomic regions disclosed herein are created for a given type of cancer using one or more of the following design phases:
Phase 1: Identify known "driver" genes, i.e., genes that are known to be mutated frequently in the particular cancer.

Phase 2: Maximize patient coverage by selecting genomic regions that contain recurrent mutations in multiple subjects with the particular cancer and ranking those selections to maximize the number of patients identified by mutations in those regions.

Phases 3 and 4: Further ranking of genomic regions containing recurrent mutations by maximizing the "recurrence index".

Phase 5: Add genomic regions from genes predicted to harbor "driver" mutations in the particular cancer.

5 Phase 6: Add genomic regions covering fusions and their flanking regions.

[0074] It should be understood, however, that the above-described phases of selector design are independent of one another and may be applied separately or in a different order within the methods of library creating and still achieve the desired result.

[0075] Application of the above approaches for recurrently mutated genomic regions
10 in non-small cell lung cancer results in the library shown in Table 1. All genomic regions included in the selector, along with their corresponding HUGO gene symbols and genomic coordinates, as well as patient statistics for NSCLC and a variety of other cancers, are shown, organized by selector design phase. The percentage of coverage of NSCLC patients as the Table 1 library was developed is shown in Figure 1(b). Also
15 shown in the bottom panel of this figure is the cumulative length of genomic regions (in kb) as the library is created according to the above phasing. The three curves in the top panel show percentage coverage of patients with at least one distinguishing mutation between tumor and genomic sequences (≥ 1 SNVs), at least two distinguishing mutations between tumor and genomic sequences (≥ 2 SNVs), and at least three
20 distinguishing mutations between tumor and genomic sequences (≥ 3 SNVs). As is apparent from these graphs, the library created according to the instant methods identifies genomic regions that are highly likely to include identifiable mutations in tumor sequences. This library includes a relatively small total number of genomic regions and thus a relatively short cumulative length of genomic regions and yet
25 provides a high overall coverage of likely mutations in a population. The library does not, therefore, need to be optimized on a patient-by-patient basis. The relatively short cumulative length of genomic regions also means that the analysis of cancer-derived cell-free DNA using these libraries is highly sensitive and allows the sequencing of this DNA to a great depth.

30 **[0076]** Accordingly, the libraries of recurrently mutated genomic regions created using the instant methods comprise a plurality of genomic regions that are recurrently mutated in a specific cancer, and the plurality of genomic regions comprises at least 10

different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 25 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 50 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 100 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 150 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 200 different genomic regions. In some embodiments, the plurality of genomic regions comprises at least 500 different genomic regions or even more.

[0077] In some embodiments, the plurality of genomic regions comprises at most 5000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 2000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 1000 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 500 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 200 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 150 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 100 different genomic regions. In some embodiments, the plurality of genomic regions comprises at most 50 different genomic regions or even fewer.

[0078] Importantly, the libraries of recurrently mutated genomic regions created according to the instant methods enable the identification of patient- and tumor-specific mutations within the genomic regions in a high percentage of subjects. Specifically, in these libraries, at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. In some embodiments, at least two mutations within the plurality of genomic regions are present in at least 60% of all subjects with the specific cancer. In specific embodiments, at least three mutations, or even more, within the plurality of genomic regions are present in at least 60% of all subjects with the specific cancer.

[0079] In some embodiments, in the libraries of recurrently mutated genomic regions created according to these methods, at least one mutation within the plurality of

genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0080] In specific embodiments, at least two mutations within the plurality of genomic regions are present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0081] In more specific embodiments, at least three mutations, or even more, within the plurality of genomic regions are present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, 99.9% or even higher percentages of all subjects with the specific cancer.

[0082] As previously noted, the cumulative length of genomic regions in the libraries of recurrently mutated genomic regions created according to the instant methods are relatively short, thus minimizing sequencing costs associated with the analytical methods relying on these libraries and maximizing their sensitivity. In some embodiments, the cumulative length of genomic regions is at most 30 megabases (Mb). In some embodiments, the cumulative length of genomic regions is at most 20 Mb, 10 Mb, 5 Mb, 2 Mb, or 1 Mb. In some embodiments, the cumulative length of genomic regions is at most 500 kilobases (kb), 200 kb, 100 kb, 50 kb, 20 kb, 10 kb, or even fewer.

[0083] In some embodiments, the library of recurrently mutated genomic regions created according to the instant methods comprises the genomic regions displayed in Table 1, or a subset of those genomic regions.

[0084] The instant methods include the step of identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer. As noted elsewhere, the libraries are particularly useful in methods for analyzing cancer-specific gene alterations in solid tumors, because those alterations can be detected in cell-free nucleic acids present in blood samples. Accordingly, the libraries created according to these methods include genomic regions that are recurrently mutated in a solid tumor. In some embodiments, the solid tumor is a carcinoma. In specific embodiments, the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma. The methods are also applicable to genomic regions that are recurrently mutated in other cancers, however. Specifically, the other cancer may be, for example, a sarcoma, a leukemia, a lymphoma, or a myeloma.

Systems

[0085] The methods for creating a library of recurrently mutated genomic regions, as disclosed herein, are typically implemented by a programmed computer system.

Therefore, according to another aspect, the instant disclosure provides computer systems for creating a library of recurrently mutated genomic regions. Such systems comprise at least one processor and a non-transitory computer-readable medium storing computer-executable instructions that, when executed by the at least one processor, cause the computer system to carry out the above-described methods for creating a library.

Methods for Analyzing Genetic Alterations

[0086] The libraries created according to the above-described methods are useful in the analysis of genetic alterations, particularly in comparing tumor and genomic sequences in a patient with cancer. As shown in Figure 2, a tissue biopsy sample from the patient may be used to discover mutations in the tumor by sequencing the genomic regions of the selector library in tumor and genomic nucleic acid samples and comparing the results. Because the selector libraries are designed to identify mutations in tumors from a large percentage of all patients, it is not necessary to optimize the library for each patient.

[0087] Accordingly, in this aspect of the invention, methods are provided for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample.

[0088] In these methods, the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer; the plurality of genomic regions comprises at least 10 different genomic regions; and at least one

mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. More specifically, the plurality of target regions may correspond to the plurality of genomic regions found in the libraries of recurrently mutated genomic regions created using the above-described methods. In other words, 5 in various embodiments, the number of different genomic regions in the plurality of genomic regions, the number of mutations within the plurality of genomic regions that are present in a specific percentage of all subjects with the specific cancer, the percentage of all subjects with the specific cancer with at least one mutation within the plurality of genomic regions, the specific composition of the plurality of genomic 10 regions, the types of cancer, and the cumulative length of the plurality of genomic regions have the values disclosed above for the methods of creating a library.

[0089] In some embodiments, the plurality of target regions used in the methods for analyzing a cancer-specific genetic alteration in a subject corresponds to the library of recurrently mutated genomic regions displayed in Table 1, or a subset of those genomic 15 regions.

[0090] It should be understood that the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may occur in a single step or in separate steps. For example, it may be possible to obtain a single tissue sample from a patient, for example from a biopsy sample, that includes both 20 tumor nucleic acids and genomic nucleic acids. It is also within the scope of this step to obtain the tumor nucleic acid sample and the genomic nucleic acid sample from the subject in separate samples, in separate tissues, or even at separate times.

[0091] The step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may also include the process of extracting 25 a biological fluid or tissue sample from the subject with the specific cancer. These particular steps are well understood by those of ordinary skill in the medical arts, particularly by those working in the medical laboratory arts.

[0092] The step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may additionally include procedures to 30 improve the yield or recovery of the nucleic acids in the sample. For example, the step may include laboratory procedures to separate the nucleic acids from other cellular components and contaminants that may be present in the biological fluid or tissue

sample. As noted, such steps may improve the yield and/or may facilitate the sequencing reactions.

[0093] It should also be understood that the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer may be performed by a commercial laboratory that does not even have direct contact with the subject. For example, the commercial laboratory may obtain the nucleic acid samples from a hospital or other clinical facility where, for example, a biopsy or other procedure is performed to obtain tissue from a subject. The commercial laboratory may thus carry out all the steps of the instantly-disclosed methods at the request of, or under the instructions of, the facility where the subject is being treated or diagnosed.

Methods for Screening

[0094] The methods of the instant invention may also be applied to the detection of cancer in a patient, where there is no prior knowledge of the presence of a tumor in the patient. Accordingly, in this aspect of the invention are provided methods for screening a cancer-specific genetic alteration in a subject comprising the steps of:

- obtaining a cell-free nucleic acid sample from a subject;
- sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and
- identifying a cancer-specific genetic alteration in the cell-free sample.

[0095] In these methods, the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer. In some embodiments, the plurality of genomic regions comprises at least 10 different genomic regions, and at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer. More specifically, the plurality of target regions may correspond to the plurality of genomic regions found in the libraries of recurrently mutated genomic regions created using the above-described methods. In other words, in various embodiments, the number of different genomic regions in the plurality of genomic regions, the number of mutations within the plurality of genomic regions that are present in a specific percentage of all subjects with the specific cancer, the percentage of all subjects with the specific cancer with at least one mutation within the plurality of genomic regions, the specific composition of the plurality of genomic

regions, the types of cancer, and the cumulative length of the plurality of genomic regions have the values disclosed above for the methods of creating a library.

[0096] In some embodiments, the plurality of target regions used in the methods for screening a cancer-specific genetic alteration in a subject corresponds to the library of recurrently mutated genomic regions displayed in Table 1, or a subset of those genomic regions.

[0097] It will be readily apparent to one of ordinary skill in the relevant arts that other suitable modifications and adaptations to the methods and applications described herein may be made without departing from the scope of the invention or any embodiment thereof. Having now described the present invention in detail, the same will be more clearly understood by reference to the following Examples, which are included herewith for purposes of illustration only and are not intended to be limiting of the invention.

EXAMPLES

Noninvasive and Ultrasensitive Quantitation of Circulating Tumor DNA by Hybrid Capture and Deep Sequencing

[0098] To overcome the limitations of prior methods, an ultrasensitive and specific strategy for analysis of cancer-derived cfDNA (Cancer Personalized Profiling by Deep Sequencing (CAPP-Seq)) that can simultaneously detect single nucleotide variants (SNVs), insertions/deletions (indels), and rearrangements, without the need for patient-specific optimization has been developed. CAPP-Seq employs an adaptable "selector" to enrich recurrently mutated regions in the cancer of interest using a custom library of biotinylated DNA oligonucleotides (Ng et al. (2010) *Nat. Genetics* 42:30-35). To use CAPP-Seq for monitoring circulating tumor DNA, this selector is typically applied first to matched tumor and normal genomic DNA to identify a patient's cancer-specific genetic aberrations and then directly to cfDNA in order to quantify these mutations (Figure 1a and Figure 2).

[0099] The design of an NSCLC CAPP-Seq selector is shown in Figure 1(b). Phase 1: Genomic regions harboring known/suspected driver mutations in NSCLC. Phases 2-4: Addition of exons containing recurrent SNVs using WES data from lung adenocarcinomas and squamous cell carcinomas from TCGA (N=407). Regions were

selected iteratively to maximize the number of mutations per tumor while minimizing selector size. Recurrence index = total unique patients with mutations covered per kb of exon. Phases 5-6: Exons of predicted NSCLC drivers (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181) and introns/exons harboring breakpoints in rearrangements involving *ALK*, *ROS1*, and *RET* were added. Bottom: increase of selector length during each design phase. Figure 1(c) shows an analysis of the number of SNVs per lung adenocarcinoma covered by the NSCLC CAPP-Seq selector in the TCGA WES cohort (*Training*; N=229) and an independent lung adenocarcinoma WES data set (*Validation*; N=183) (Imielinski et al. (2012) *Cell* 150:1107-1120). Results are compared to selectors randomly sampled from the exome ($P < 1.0 \times 10^{-6}$ for the difference between random selectors and the NSCLC CAPP-Seq selector). Figure 1(d) shows the number of SNVs per patient identified by the NSCLC CAPP-Seq selector in WES data from three adenocarcinomas from TCGA, colon (COAD), rectal (READ), and endometrioid (UCEC) cancers. Figures 1(e) and (f) show quality parameters from a representative CAPP-Seq analysis of plasma cfDNA, including length distribution of sequenced cfDNA fragments 1(e), and depth of sequencing coverage across all genomic regions in the selector 1(f). Figure 1(g) illustrates the variation in sequencing depth across cfDNA samples from 4 patients. The envelope above and below the solid line represents s.e.m. Figure 2 illustrates the CAPP-Seq computational pipeline. See Detailed Methods section for details.

[0100] For the initial implementation of CAPP-Seq we focused on NSCLC, although our approach is generalizable to any cancer for which a comprehensive list of recurrent mutations has been identified. We employed a multi-phase approach to design a NSCLC-specific selector, aiming to identify genomic regions recurrently mutated in this disease (Figure 1b, Table 1, and Methods). We began by including exons covering recurrent mutations in potential driver genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Forbes et al. (2010) *Nucleic Acids Res.* 38:D652-657) as well as other sources (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181) (e.g. *KRAS*, *EGFR*, *TP53*). Next, using whole exome sequencing (WES) data from 407 NSCLC patients profiled by The Cancer Genome Atlas (TCGA), an iterative algorithm was applied to maximize the number of mutations per patient while minimizing selector size. The approach relied on a

recurrence index that identified known driver mutations as well as uncharacterized genes that are frequently mutated and are therefore likely to be involved in NSCLC pathogenesis (Figure 3 and Table 1).

Table 1. Recurrently mutated genomic regions in NSCLC.

Design phase	Sample design			Patients cohort				Coverage: Unique L100 & SCC patients (n=607)				Coverage: All L100 & SCC samples (n=116)			
	Regions covered	Genes covered	Length (bp)	Gene	Chr	Start (bp)	End (bp)	Length (bp)	Patients covered	No. patients per exon	FI	No. samples with SNV	% patients ≥ 1 SNV	% patients ≥ 2 SNVs	% patients ≥ 3 SNVs
Known drivers	1	1	130	AKT1	chr14	103249424	103249533	130	1	1	7.7	1	0.00	0.00	0.00
Known drivers	2	280	119	BRCA1	chr17	140453174	140453183	120	9	8	66.7	9	2.21	0.00	0.00
Known drivers	3	369	119	BRCA1	chr17	140453175	140453190	116	16	7	53.9	16	3.93	0.00	0.00
Known drivers	4	3	1079	CDKN2A	chr12	213714375	213714382	108	16	20	59.4	16	3.93	0.00	0.00
Known drivers	5	3	1079	CDKN2A	chr12	213714375	213714382	352	53	7	59.4	53	12.02	0.00	0.00
Known drivers	6	4	1353	CTNNA3	chr13	412636162	412636169	124	57	4	28.2	55	13.02	0.00	0.00
Known drivers	7	5	1392	EGFR	chr7	55241813	55241788	124	59	1	24.2	54	14.95	0.00	0.00
Known drivers	8	5	1483	EGFR	chr7	55241814	55241788	100	59	1	24.2	54	14.95	0.00	0.00
Known drivers	9	5	1689	EGFR	chr7	55241814	55241817	187	89	4	50.0	84	15.97	1.23	0.00
Known drivers	10	5	1828	EGFR	chr7	55241814	55241817	187	89	4	50.0	84	15.95	1.23	0.00
Known drivers	11	8	1828	EGFR	chr7	55241814	55241817	157	81	0	0.0	74	15.90	1.72	0.25
Known drivers	12	6	2113	ERBB2	chr17	37890164	37890263	100	81	0	0.0	74	15.90	1.72	0.25
Known drivers	13	7	2293	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	14	7	2405	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	15	8	2593	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	16	8	2740	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	17	8	3477	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	18	8	4117	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	19	8	4295	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	20	9	4445	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	21	9	4577	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	22	11	4789	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	23	11	4831	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	24	12	5198	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	25	13	5417	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	26	13	5850	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	27	13	7339	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	28	13	7489	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	29	14	7689	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	30	14	7731	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	31	15	7917	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	32	15	8179	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	33	16	8259	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	34	16	8345	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	35	16	8431	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	36	16	8498	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	37	16	8678	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	38	16	8819	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	39	16	8927	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	40	16	9013	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	41	17	9039	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	42	17	9099	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	43	17	9143	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	44	17	9214	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	45	17	9252	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	46	17	10081	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	47	17	10140	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	48	17	10329	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	49	17	10524	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	50	18	10632	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	51	18	10773	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	52	18	10837	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	53	18	11187	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Known drivers	54	18	11352	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	55	19	11472	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	56	20	11527	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	57	21	11641	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	58	21	11749	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	59	22	11851	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	60	23	11850	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	61	24	12084	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	62	25	12257	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	63	26	12339	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	64	27	12505	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	65	28	12590	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	66	29	12692	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	67	30	12801	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	68	31	12894	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	69	32	12988	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	70	33	13093	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	71	34	13240	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	72	35	13400	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	73	36	13599	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	74	37	13813	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	75	38	14033	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	76	39	14250	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	77	40	14472	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	78	41	14696	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	79	42	14921	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	80	43	15147	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25
Mac coverage	81	44	15349	ERBB2	chr17	37890164	37890263	187	85	4	21.4	79	20.86	1.72	0.25

82	43	15551	LPJ1B	CH2	141685646	141685646	152	377	1	7	34.7	150	92.83	55.77	29.48	388	1	8	39.6	155	92.80	55.81	29.59
83	44	15559	GPJ2B	CH2	141724655	141724655	198	378	1	6	38.0	154	92.87	56.27	29.98	389	1	8	38.0	154	92.84	56.09	30.07
84	45	16031	CJ2L	CH1	157869564	157869564	322	379	1	12	37.3	147	93.12	57.00	30.98	390	1	12	37.3	152	93.08	58.80	31.03
85	46	16509	SPTA1	CH1	158967266	158967266	219	380	1	8	38.5	149	93.32	57.78	31.03	391	1	8	38.5	149	93.32	57.78	31.03
86	47	16832	ADAM7520	CH1	162812428	162812428	142	381	1	5	35.2	144	93.81	58.48	31.20	392	1	5	35.2	148	93.58	59.23	31.28
87	48	18835	ADAM7520	CH2	143863983	143863983	143	382	1	5	35.0	144	93.88	58.48	31.20	393	1	5	35.0	149	93.79	59.23	31.28
88	49	23588	FLJ6C	CH19	149067147	149067147	155	383	0	35	31.3	109	95.08	58.30	43.24	394	0	36	33.2	115	95.23	60.78	43.24
89	50	17109	NLRF4	CH19	156823163	156823163	172	384	2	17	34.9	140	93.86	59.72	31.84	394	1	6	34.9	150	94.03	59.23	31.86
90	51	17344	CH1F6	CH5	19473825	19473825	142	384	2	17	34.9	140	94.35	59.96	32.68	396	2	17	34.6	147	94.51	59.43	32.70
91	52	18281	MYH2	CH17	145030351	145030351	145	388	2	5	32.0	134	94.84	61.92	32.92	398	2	5	34.5	149	94.99	59.43	32.94
92	53	18317	ORAI1A	CH11	151365369	151365369	937	388	0	30	32.0	134	94.84	61.92	32.92	398	0	30	32.0	141	94.99	61.94	33.56
93	54	20245	ORAI1A	CH11	147875130	147875130	1038	388	0	32	30.9	121	94.84	63.88	37.59	388	0	34	32.8	133	94.99	63.25	37.71
94	55	21178	ORAI1A	CH11	151423242	151423242	962	388	0	26	28.0	128	94.84	65.11	38.33	388	0	28	30.2	127	94.99	64.68	38.42
95	56	22224	ORAI1A	CH11	155423572	155423572	941	387	1	27	29.0	117	95.09	68.34	39.01	389	1	28	30.1	123	95.23	68.83	39.88
96	57	23588	FLJ6C	CH19	149067147	149067147	155	383	0	35	31.3	113	95.09	68.34	39.01	389	0	33	31.5	119	95.23	68.83	42.00
97	58	23588	FLJ6C	CH19	149067147	149067147	155	383	0	35	31.3	109	95.08	58.30	43.24	394	0	36	33.2	115	95.23	60.78	43.24
98	59	23588	FLJ6C	CH19	149067147	149067147	155	383	0	35	31.3	109	95.08	58.30	43.24	394	0	36	33.2	115	95.23	60.78	43.24
99	59	24062	TNN	CH1	170816438	170816438	199	387	0	9	35.2	105	95.08	68.29	43.24	394	0	9	35.2	111	95.23	63.78	43.20
100	60	24389	GAT3A	CH10	170816438	170816438	199	387	0	9	35.2	105	95.08	68.29	43.24	394	0	9	35.2	108	95.23	63.78	43.20

170	118	36894	ACER2	chr8	19424867	19424838	143	389	0	3	210	4	9558	9459	5749	401	0	4	280	4	9570	9475	5752
171	120	37103	AMOT	chrX	112024413	112024346	234	389	0	5	214	3	9558	9484	5823	401	0	8	256	3	9570	9499	5776
172	121	37215	CBP2A	chr9	138193718	138193727	112	389	0	3	283	2	9558	9509	5923	401	0	3	298	2	9570	9523	5778
173	122	38109	INHEA	chr7	41729497	41730140	884	389	0	17	190	2	9558	9509	5937	401	0	17	190	2	9570	9523	5847
174	122	38468	INHEA	chr7	41739484	41739972	389	389	0	3	7.7	2	9558	9509	5946	401	0	3	7.7	2	9570	9523	5947
175	123	39065	EPHA5	chr4	86193631	86193637	107	389	0	3	280	2	9558	9509	5948	401	0	3	280	2	9570	9523	5948
176	123	39362	EPHA5	chr4	86187660	86187648	157	389	0	2	12.7	2	9558	9509	5948	401	0	2	12.7	2	9570	9523	5948
177	123	39367	EPHA5	chr4	86201649	86201643	195	389	0	2	10.3	2	9558	9509	5971	401	0	2	10.3	2	9570	9523	5971
178	123	39108	EPHA5	chr4	86213771	86213821	151	389	0	3	19.9	2	9558	9509	5971	401	0	3	19.9	2	9570	9523	5971
179	123	39319	EPHA5	chr4	86217106	86217316	211	389	0	4	19.0	2	9558	9509	5955	401	0	4	19.0	2	9570	9523	5943
180	123	39420	EPHA5	chr4	86218470	86218640	101	389	0	4	19.0	2	9558	9509	6044	401	0	4	19.0	2	9570	9523	5943
181	123	39807	EPHA5	chr4	86230734	86230920	167	389	0	3	18.0	2	9558	9509	6044	401	0	3	21.6	2	9570	9523	5943
182	123	39734	EPHA5	chr4	86231775	86231775	127	389	0	3	23.8	2	9558	9509	6044	401	0	3	23.8	2	9570	9523	6038
183	123	39635	EPHA5	chr4	86233953	86233158	101	389	0	2	19.8	2	9558	9509	6093	401	0	2	19.8	2	9570	9523	6038
184	123	39638	EPHA5	chr4	86242366	86242788	101	389	0	0	0.0	2	9558	9509	6093	401	0	0	0.0	2	9570	9523	6038
185	123	40040	EPHA5	chr4	86257091	862570184	164	389	0	2	14.2	2	9558	9509	6093	401	0	2	14.2	2	9570	9523	6038
186	123	40041	EPHA5	chr4	86257091	862570184	164	389	0	1	6.2	2	9558	9509	6093	401	0	1	6.2	2	9570	9523	6038
187	123	40227	EPHA5	chr4	86283815	86283801	125	389	0	0	14.8	2	9558	9509	6093	401	0	0	14.8	2	9570	9523	6038
188	123	40264	EPHA5	chr4	86350364	86350369	324	389	0	5	14.8	2	9558	9509	6093	401	0	5	14.8	2	9570	9523	6038
189	123	40821	EPHA5	chr4	86361105	86361281	157	389	0	6	9.1	2	9558	9509	6093	401	0	6	9.1	2	9570	9523	6038
190	123	41168	EPHA5	chr4	86373783	86373722	955	389	0	0	0.0	2	9558	9509	6093	401	0	0	0.0	2	9570	9523	6038
191	123	41588	EPHA5	chr4	86379482	86379481	102	389	0	1	5.5	2	9558	9509	6093	401	0	1	5.5	2	9570	9523	6038
192	123	41770	EPHA5	chr4	86385270	863853469	162	389	0	0	0.0	2	9558	9509	6093	401	0	0	0.0	2	9570	9523	6038
193	124	41871	EPHA3	chr3	86176340	86175441	102	389	0	2	19.6	2	9558	9509	6093	401	0	2	19.6	2	9570	9523	6038
194	124	41873	EPHA3	chr3	862559009	862559870	862	389	0	6	9.1	2	9558	9509	6093	401	0	6	9.1	2	9570	9523	6038
195	124	42835	EPHA3	chr3	863930085	863930221	157	389	0	4	25.5	2	9558	9509	6143	401	0	4	25.5	2	9570	9523	6082
196	124	42792	EPHA3	chr3	863930084	863931240	337	389	0	3	8.9	2	9558	9509	6143	401	0	3	8.9	2	9570	9523	6082
197	124	43429	EPHA3	chr3	86444968	86445111	128	389	0	2	15.9	2	9558	9509	6187	401	0	2	15.9	2	9570	9523	6110
198	124	43465	EPHA3	chr3	86444967	86444858	180	389	0	1	5.3	2	9558	9509	6187	401	0	1	5.3	2	9570	9523	6110
199	124	43549	EPHA3	chr3	86456418	86456521	104	389	0	0	0.0	2	9558	9509	6187	401	0	0	0.0	2	9570	9523	6134
200	124	43549	EPHA3	chr3	86457188	86457289	102	389	0	0	0.0	2	9558	9509	6187	401	0	0	0.0	2	9570	9523	6134
201	124	43561	EPHA3	chr3	86462290	86462416	127	389	0	3	23.8	2	9558	9509	6187	401	0	3	23.8	2	9570	9523	6134
202	124	43778	EPHA3	chr3	86462290	86462416	127	389	0	0	0.0	2	9558	9509	6187	401	0	0	0.0	2	9570	9523	6134
203	124	43865	EPHA3	chr3	86463854	86465540	187	389	0	1	5.3	2	9558	9509	6187	401	0	1	5.3	2	9570	9523	6134
204	124	44068	EPHA3	chr3	86476238	86477336	101	389	0	4	19.0	2	9558	9509	6187	401	0	4	19.0	2	9570	9523	6134
205	124	44277	EPHA3	chr3	86490299	86490509	211	389	0	0	0.0	2	9558	9509	6187	401	0	0	0.0	2	9570	9523	6134
206	124	44428	EPHA3	chr3	86496374	86496524	151	389	0	1	8.6	2	9558	9509	6187	401	0	1	8.6	2	9570	9523	6134
207	124	44623	EPHA3	chr3	86496374	86496524	151	389	0	2	10.3	2	9558	9509	6218	401	0	2	10.3	2	9570	9523	6134
208	124	44780	EPHA3	chr3	86521813	86521769	157	389	0	3	19.1	2	9558	9509	6265	401	0	3	19.1	2	9570	9523	6205
209	124	44887	EPHA3	chr3	86526546	86526562	107	389	0	1	9.3	2	9558	9509	6265	401	0	1	9.3	2	9570	9523	6205
210	124	44889	PTFRD	chr9	8317657	8317658	102	389	0	2	19.8	2	9558	9509	6265	401	0	2	19.8	2	9570	9523	6205
211	125	45138	PTFRD	chr9	8319680	8319686	137	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
212	125	45282	PTFRD	chr9	8301581	8331736	458	389	0	1	8.4	2	9558	9509	6265	401	0	1	8.4	2	9570	9523	6205
213	125	45408	PTFRD	chr9	8338821	8339047	127	389	0	2	15.7	2	9558	9509	6265	401	0	2	15.7	2	9570	9523	6205
214	125	45537	PTFRD	chr9	8340942	8340949	128	389	0	1	7.8	2	9558	9509	6265	401	0	1	7.8	2	9570	9523	6205
215	125	45717	PTFRD	chr9	8341089	8341268	180	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
216	125	46004	PTFRD	chr9	8341892	8341978	287	389	0	2	7.0	2	9558	9509	6265	401	0	2	7.0	2	9570	9523	6205
217	125	46160	PTFRD	chr9	8376385	8376980	166	389	0	1	6.4	2	9558	9509	6265	401	0	1	6.4	2	9570	9523	6205
218	125	46281	PTFRD	chr9	8376906	8376720	121	389	0	1	8.3	2	9558	9509	6265	401	0	1	8.3	2	9570	9523	6205
219	125	46383	PTFRD	chr9	8386231	8386407	177	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
220	125	46383	PTFRD	chr9	8404535	8404660	125	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
221	125	46884	PTFRD	chr9	8436580	8436580	101	389	0	1	9.9	2	9558	9509	6265	401	0	1	9.9	2	9570	9523	6205
222	125	46785	PTFRD	chr9	8437168	8437268	101	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
223	125	46888	PTFRD	chr9	8448724	8449537	114	389	0	3	28.3	2	9558	9509	6265	401	0	3	28.3	2	9570	9523	6205
224	125	47001	PTFRD	chr9	8454536	8454537	102	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
225	125	47001	PTFRD	chr9	8454536	8454537	102	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
226	125	47374	PTFRD	chr9	8465465	8465575	214	389	0	6	28.3	2	9558	9509	6265	401	0	7	53.2	2	9570	9523	6205
227	125	47476	PTFRD	chr9	8470989	8471390	102	389	0	1	3.6	2	9558	9509	6265	401	0	1	3.6	2	9570	9523	6205
228	125	47737	PTFRD	chr9	8484118	8484378	251	389	0	5	19.2	2	9558	9509	6265	401	0	5	19.2	2	9570	9523	6205
229	125	47839	PTFRD	chr9	8485228	8485327	102	389	0	0	0.0	2	9558	9509	6265	401	0	0	0.0	2	9570	9523	6205
230	125	48028	PTFRD	chr9	8485761	8486349	539	389	0	4	8.8	2	9558	9509	6314	401	0	5	8.5	2	9570	9523	6277
231	12																						

258	128	52790	KDR	CH4	55942509	105	389	0	1	8,7	2	95,59	95,09	85,11	401	0	1	8,7	2	95,70	95,23	84,88	
259	128	52800	KDR	CH4	55942509	106	389	0	3	28,3	0	95,59	95,09	85,11	401	0	3	28,3	0	95,70	95,23	84,88	
260	128	52810	KDR	CH4	55942509	107	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
261	128	52820	KDR	CH4	55942509	108	389	0	2	18,5	2	95,59	95,09	85,11	401	0	2	18,5	2	95,70	95,23	84,88	
262	128	52830	KDR	CH4	55942509	113	389	0	0	1	7,5	2	95,59	95,09	85,11	401	0	1	7,5	2	95,70	95,23	84,88
263	128	52840	KDR	CH4	55942509	148	389	0	2	13,5	2	95,59	95,09	85,11	401	0	2	13,5	2	95,70	95,23	84,88	
264	128	52850	KDR	CH4	55942509	153	389	0	5	14,8	2	95,59	95,09	85,11	401	0	5	14,8	2	95,70	95,23	84,88	
265	128	52860	KDR	CH4	55942509	157	389	0	2	19,2	2	95,59	95,09	85,11	401	0	2	19,2	2	95,70	95,23	84,88	
266	128	52870	KDR	CH4	55942509	159	389	0	2	12,7	2	95,59	95,09	85,11	401	0	2	12,7	2	95,70	95,23	84,88	
267	128	52880	KDR	CH4	55942509	165	389	0	1	8,1	2	95,59	95,09	85,11	401	0	1	8,1	2	95,70	95,23	84,88	
268	128	52890	KDR	CH4	55942509	168	389	0	1	8,8	2	95,59	95,09	85,11	401	0	1	8,8	2	95,70	95,23	84,88	
269	128	52900	KDR	CH4	55942509	178	389	0	2	11,2	2	95,59	95,09	85,11	401	0	2	11,2	2	95,70	95,23	84,88	
270	128	52910	KDR	CH4	55942509	182	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
271	128	52920	KDR	CH4	55942509	184	389	0	2	10,0	2	95,59	95,09	85,11	401	0	2	10,0	2	95,70	95,23	84,88	
272	128	52930	KDR	CH4	55942509	186	389	0	4	9,3	2	95,59	95,09	85,11	401	0	4	9,3	2	95,70	95,23	84,88	
273	128	52940	KDR	CH4	55942509	187	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
274	128	52950	KDR	CH4	55942509	191	389	0	1	9,9	2	95,59	95,09	85,11	401	0	1	9,9	2	95,70	95,23	84,88	
275	128	52960	KDR	CH4	55942509	192	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
276	128	52970	KDR	CH4	55942509	197	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
277	128	52980	KDR	CH4	55942509	197	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
278	127	52989	KDR	CH4	55942509	197	389	0	1	6,3	2	95,59	95,09	85,11	401	0	1	6,3	2	95,70	95,23	84,88	
279	127	52990	KDR	CH4	55942509	197	389	0	0	0	2	95,59	95,09	85,11	401	0	0	0	2	95,70	95,23	84,88	
280	127	52991	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
281	127	52992	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
282	127	52993	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
283	127	52994	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
284	127	52995	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
285	127	52996	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
286	127	52997	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
287	127	52998	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
288	127	52999	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
289	127	53000	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
290	127	53001	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
291	127	53002	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
292	127	53003	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
293	127	53004	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
294	127	53005	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
295	127	53006	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
296	127	53007	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
297	127	53008	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
298	127	53009	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
299	127	53010	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
300	127	53011	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
301	127	53012	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
302	127	53013	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
303	127	53014	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
304	127	53015	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
305	128	53020	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
306	128	53025	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
307	128	53030	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
308	128	53035	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
309	128	53040	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
310	128	53045	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
311	128	53050	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
312	128	53055	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
313	128	53060	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
314	128	53065	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
315	128	53070	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
316	128	53075	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
317	128	53080	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
318	128	53085	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
319	128	53090	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
320	128	53095	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
321	128	53100	KDR	CH4	55942509	197	389	0	1	4,1	2	95,59	95,09	85,11	401	0	1	4,1	2	95,70	95,23	84,88	
322	128	53105	KDR	CH4	55942509	197	389	0	1	4,1	2												

[illegible]

[0101] Figure 3 illustrates how the statistical enrichment of recurrently mutated NSCLC exons captures known drivers. Two metrics were employed to prioritize exons with recurrent mutations for inclusion in the CAPP-Seq NSCLC selector. The first, termed Recurrence Index (RI), is defined as the number of unique patients (i.e. tumors) with somatic mutations per kilobase of a given exon and the second metric is based on the minimum number of unique patients (i.e. tumors) with mutations in a given kb of exon. Exons containing at least one non-silent SNV genotyped by TCGA (n=47,769) in a combined cohort of 407 lung adenocarcinoma (LUAD) and squamous cell carcinoma (SCC) patients were analyzed. As shown in Figure 3(a), known/suspected NSCLC drivers are highly enriched at $RI \geq 30$ (inset), comprising 1.8% (n=861) of analyzed exons. As shown in Figure 3(b), known/suspected NSCLC drivers are highly enriched at ≥ 3 patients with mutations per exon (inset), encompassing 16% of analyzed exons.

[0102] Approximately 8% of NSCLCs contain clinically actionable rearrangements involving the receptor tyrosine kinases, *ALK*, *ROS1* and *RET* (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870; Kwak et al. (2010) *N. Engl. J. Med.* 363:1693-1703; Pao & Hutchinson (2012) *Nat. Med.* 18:349-351). To utilize the personalized nature and low false detection rate of structural rearrangements (Leary et al. (2010) *Sci. Transl. Med.* 2:20ra14; McBride et al. (2010) *Genes Chrom. Cancer* 49:1062-1069), introns and exons spanning recurrent fusion breakpoints in these genes were included in the final design phase (Figure 1b). To detect fusions in tumor and plasma DNA, a breakpoint-mapping algorithm called FACTERA was developed (Figure 4). Application of FACTERA to next generation sequencing (NGS) data from 2 NSCLC cell lines known to harbor fusions with previously uncharacterized breakpoints (Koivunen et al. (2008) *Clin. Cancer Res.* 14:4275-4283; Rikova et al. (2007) *Cell* 131:1190-1203) readily identified the breakpoints in both cases (Figure 5).

[0103] Collectively, the NSCLC CAPP-Seq selector design targets 521 exons and 13 introns from 139 recurrently mutated genes, in total covering ~125kb (Figure 1b). Within this small target (0.004% of the human genome), CAPP-Seq identifies a median of 4 point mutations and covered 96% of patients with lung adenocarcinoma or squamous cell carcinoma. To validate the number of mutations covered per tumor, we examined the selector region in WES data from an independent cohort of 183 lung

adenocarcinoma patients (Imielinski et al. (2012) *Cell* 150:1107-1120). The selector covered 88% of patients with a median of 4 SNVs per patient, thus validating our selector design algorithm ($P < 1.0 \times 10^{-6}$; Figure 1c). When compared to randomly sampling the exome, regions targeted by CAPP-Seq captured ~4-fold as many mutations per patient (at the median, Figure 1c). Due to similarities in key oncogenic machinery across cancers (Hanahan & Weinberg (2011) *Cell* 144:646-674), we hypothesized that our NSCLC selector would perform favorably on other carcinomas. Indeed, when applied to TCGA WES data, the selector successfully captured 99% of colon, 98% of rectal, and 97% of endometrioid uterine carcinomas, with a median of 12, 7, and 3 mutations per patient, respectively (Figure 1d). This demonstrates the value of targeting hundreds of recurrently mutated genomic regions and suggests that a CAPP-Seq selector could be designed to simultaneously cover mutations for a wide variety of human malignancies.

[0104] Using this CAPP-Seq selector, we profiled a total of 52 samples including NSCLC cell lines, primary tumor specimens, peripheral blood leukocytes (PBLs), and cfDNA isolated from plasma of patients with NSCLC before and after various cancer therapies (Table 2). To assess and optimize the performance of CAPP-Seq, we first applied it to cfDNA purified from healthy control plasma. Approximately 60% of reads mapped within the selector target region (Table 2). Sequenced cfDNA fragments had a median length of 169bp (Figure 1e), closely corresponding to the length of DNA contained within a chromosome (Fan et al. (2008) *Proc. Natl Acad. Sci. USA* 105:16266-16271). To optimize library preparation from small quantities of cfDNA we explored a variety of modifications to the ligation and post-ligation amplification steps including temperature, incubation time, enzyme source, and "with-bead" clean-up. The optimized protocol increased recovery efficiency by >300% and decreased bias for libraries constructed from as little as 4ng of cfDNA (Figures 6-8). Consequently, fluctuations in sequencing depth were minimal (Figure 1f-g) and unlikely to impact performance.

Table 2. Profile of samples using NSCLC CAPP-Seq selector

Sample	DNA mass used for library (ng)	Library mass used for capture (ng)	Total reads mapped	Fraction of properly paired reads	Read on-target rate	Median depth	Median fragment length
H3122 0.1% into HCC78	128	111	99.0%	96.8%	69.5%	8688	173
H3122 1% into HCC78	128	111	98.9%	96.7%	69.8%	8657	171
H3122 10% into HCC78	128	111	98.9%	96.5%	69.8%	6890	170
H3122 100%	128	111	99.0%	96.8%	68.6%	6739	174
HCC78 100%	128	111	99.0%	96.9%	69.7%	7602	172
cfDNA 100% 6cycles	32	83.3	97.5%	86.7%	60.3%	8280	168
HCC78 10% into cfDNA 4cycles	128	83.3	97.5%	83.3%	59.3%	2682	170
HCC78 10% into cfDNA 8cycles SigmaWGA	624	83.3	79.5%	72.0%	50.4%	15	158
HCC78 10% into cfDNA 6cycles	32	83.3	97.7%	87.2%	60.4%	8261	169
HCC78 10% into cfDNA 8cycles NEBNextOvernightBead	32	83.3	96.9%	91.8%	61.1%	6258	166
HCC78 10% into cfDNA 8cycles OrigNEBNext 15minLig	32	83.3	98.0%	93.1%	60.9%	9862	167
HCC78 10% into cfDNA 4ng 9cycles	4	83.3	97.6%	87.6%	60.5%	11630	169
P11 PBL	500	83.3	96.7%	93.8%	59.0%	6970	169
P11 Tumor	500	83.3	93.4%	88.3%	61.3%	7700	156
P6 PBL	500	83.3	96.7%	92.6%	67.2%	3848	152
P6 Tumor	1000	83.3	87.0%	81.8%	64.7%	2445	158
P8 PBL	500	83.3	96.9%	93.0%	65.8%	4021	154
P8 Tumor	500	83.3	91.7%	85.4%	63.6%	5331	151
P10 PBL	400	83.3	96.9%	93.6%	65.3%	4572	161
P10 Tumor	500	83.3	94.0%	89.6%	65.1%	5335	157
P7 PBL	500	83.3	97.1%	93.5%	67.1%	3552	155
P7 Tumor	500	83.3	94.1%	89.3%	64.0%	4793	162
HCC78 0.025% into cfDNA	32	83.3	98.2%	87.0%	46.3%	3913	169
HCC78 0.05% into cfDNA	32	83.3	98.1%	86.1%	44.7%	6549	169
HCC78 0.1% into cfDNA	32	83.3	98.4%	88.1%	44.9%	6897	169
HCC78 0.5% into cfDNA	32	83.3	98.8%	89.8%	46.2%	8096	169
HCC78 1% into cfDNA	32	83.3	98.5%	89.8%	46.5%	7779	171
P6-1 cfDNA	17	83.3	98.6%	91.3%	46.4%	11172	166
P6-2 cfDNA	20	83.3	98.5%	92.0%	46.6%	8455	166
P9 PBL	500	83.3	97.0%	94.4%	59.2%	5441	172
P9 Tumor	69	83.3	99.2%	97.3%	55.3%	7312	239
P3 PBL	500	83.3	99.3%	97.8%	57.0%	8838	235
P3 Tumor	500	83.3	99.3%	98.0%	66.0%	9562	204
P2 PBL	500	83.3	99.2%	97.5%	57.7%	7680	235
P2 Tumor	500	83.3	99.0%	97.1%	62.3%	7247	204
P4 PBL	500	83.3	99.1%	96.5%	56.5%	7331	227
P4 Tumor	200	83.3	97.5%	94.1%	60.0%	3968	189
P1 PBL	500	83.3	99.3%	97.1%	57.1%	7336	220
P1 Tumor	500	83.3	94.6%	90.1%	60.9%	976	192
P5 PBL	500	83.3	99.2%	97.2%	58.7%	8155	219
P5 Tumor	100	83.3	98.8%	97.0%	63.5%	6930	187
P9-1 cfDNA	12	83.3	99.1%	84.2%	65.6%	6839	172
P9-2 cfDNA	17	83.3	98.4%	83.9%	65.2%	6043	169
P9-3 cfDNA	16	83.3	99.4%	88.7%	67.6%	8141	167
P3-1 cfDNA	15	83.3	99.2%	86.0%	63.5%	7057	170
P3-2 cfDNA	16	83.3	99.3%	86.5%	63.5%	10089	171
P2-1 cfDNA	13	83.3	99.4%	86.9%	67.3%	6876	172
P2-2 cfDNA	16	83.3	99.5%	96.4%	63.6%	5248	185
P1-1 cfDNA	13	83.3	99.0%	85.0%	64.6%	5079	171
P1-2 cfDNA	7	83.3	99.4%	84.7%	64.1%	6487	172
P5-1 cfDNA	9	83.3	99.3%	87.8%	66.6%	7604	169
P5-2 cfDNA	15	83.3	99.4%	88.0%	67.5%	10451	170

[0105] Figure 6 illustrates the improvements in CAPP-Seq performance achieved

5 with optimized library preparation procedures. Using 32ng of input cfDNA from plasma, standard versus "with bead" (Fisher et al. (2011) *Genome biology* 12:R1) library preparation methods were compared, as well as two commercially available DNA polymerases (Phusion and KAPA HiFi). Template pre-amplification by Whole Genome Amplification (WGA) using Degenerate Oligonucleotide PCR (DOP) were
10 also compared. Indices considered for these comparisons included (a) length of the

captured cfDNA fragments sequenced, (b) depth and uniformity of sequencing coverage across all genomic regions in the selector, and (c) sequence mapping and capture statistics, including uniqueness. Collectively, these comparisons identified KAPA HiFi polymerase and a "with bead" protocol as having most robust and uniform performance.

[0106] Figure 7 illustrates the optimization of allele recovery from low input cfDNA during Illumina library preparation. Bars reflect the relative yield of CAPP-Seq libraries constructed from 4 ng cfDNA, calculated by averaging quantitative PCR measurements of 4 pre-selected reporters within CAPP-Seq with pre-defined amplification efficiencies. (a) Sixteen hour ligation at 16°C increases ligation efficiency and reporter recovery. (b) Adapter ligation volume did not have a significant effect on ligation efficiency and reporter recovery. (c) Performing enzymatic reactions "with-bead" to minimize tube transfer steps increases reporter recovery. (d) Increasing adapter concentration during ligation increases ligation efficiency and reporter recovery. Reporter recovery is also higher when using KAPA HiFi DNA polymerase compared to Phusion DNA polymerase (e) and when using the KAPA Library Preparation Kit with the modifications in a - d compared to the NuGEN SP Ovation Ultralow Library System with automation on a Mondrian SP Workstation (f). Relative reporter abundance was determined by qPCR using the $2^{-\Delta Ct}$ method. All values are mean $\pm$ s.d. N.S., not significant. Based on these results, it was estimated that combining the methodological modifications in a and c – e improves yield in NGS libraries by 3.3-fold.

[0107] Figure 7 illustrates the performance of CAPP-Seq with various amounts of input cfDNA. (a) Length of the captured cfDNA fragments sequenced. (b) Depth of sequencing coverage across all genomic regions in the selector. (c) Sequence mapping and capture statistics. As expected, more input cfDNA mass correlates with more unique fragments sequenced.

[0108] The detection limit of CAPP-Seq is affected by the absolute number of available cfDNA molecules in a given volume of peripheral blood, as well as PCR and sequencing errors (i.e. "technical" background). The latter primarily affects substitutions/SNVs as opposed to other CAPP-Seq reporters (i.e., indels (Minoche et al. (2011) *Genome Biol.* 12:R112) and rearrangements). Separately, mutant cfDNA could

be present in the absence of cancer due to contributions from pre-neoplastic cells from diverse tissues (i.e., "biological" background). The combined background from these sources was measured by assessing the error rate at each nucleotide position across the selector in plasma cfDNA from 6 patients and a healthy individual, excluding tumor-derived mutations. Mean and median background rates of ~0.007% and ~0% (not detected, N.D.) were found, respectively (Figure 9 (a)). Next, we hypothesized that if significant biological background is present, it should be highest for recurrently mutated positions in cancer driver genes. We therefore analyzed mutation rates of 107 recurrent cancer-associated SNVs (Su et al. (2011) *J. Mol. Diagn.* 13:74-84) in the same 7 plasma samples, again excluding those SNVs found in corresponding tumors. Though the median fractional abundance was comparable (~0%, N.D.), the mean was marginally higher at 0.012% (Figure 9 (b)). However, only one cancer-associated mutation (*TP53* R175H) was detectable in plasma at levels significantly above global background ($P<0.01$). Since this allele was detected at a median frequency of ~0.3% across all samples (Figure 9(c)), we hypothesize that it reflects true biological background and thus excluded it as a potential CAPP-Seq reporter. Collectively, this analysis suggests that biological background is not a significant factor for disease monitoring at the current detection limits of CAPP-Seq.

[0109] Next, the allele frequency detection limit and linearity of CAPP-Seq was benchmarked by spiking defined concentrations of fragmented genomic DNA from a NSCLC cell line into cfDNA from a healthy individual (Figure 9(d)) or into genomic DNA from a second NSCLC line (Figure 10(a)). CAPP-Seq accurately detected variants at fractional abundances between 0.025% and 10% with high linearity ($R^2\geq 0.994$). Analyses of the influence of the number of SNV reporters on error metrics showed only marginal improvements above a threshold of 4 reporters per tumor (Figures 9(e)-(f), 10 (b)-(c)), equivalent to the median number of SNVs per NSCLC identified by the NSCLC selector. Finally, whether fusion breakpoints and indels could also serve as linear reporters was tested. It was found that the fractional abundance of these mutations correlated highly with expected concentrations ($R^2\geq 0.995$; Figure 10(d)).

[0110] Having designed, optimized, and benchmarked CAPP-Seq, it was applied to the discovery of somatic mutations in tumors collected from a diverse group of NSCLC

patients (n=11; Figure 11(a) and Table 3). To test the breakpoint enumeration capability of CAPP-Seq, 6 patients with clinically confirmed fusions were included. These translocations served as positive controls, along with SNVs in other tumors previously identified by clinical assays (N=9; Table 3). Tumor samples included formalin fixed surgical or biopsy specimens and pleural fluid. At a mean sequencing depth of ~6,000x in tumor and paired germline samples, CAPP-Seq confirmed all previously identified SNVs and fusions (3 and 8, respectively) and discovered many additional somatic variants (Figure 11(a) and Table 3). Moreover, CAPP-Seq characterized breakpoints and partner genes at base pair resolution for each of the 8 rearrangements (Figure 12). Tumors containing fusions were almost exclusively from never smokers and, as expected (Govindan et al. (2012) *Cell* 150:1121-1134), contained fewer SNVs than those lacking fusions (Figure 13). Excluding patients with fusions (<10% of the TCGA design cohort), CAPP-Seq identified a median of 4 SNVs per patient as we had predicted (Figure 1(b)-(c)).

Table 3. Characteristics of patients used for noninvasive detection and monitoring of circulating tumor DNA by CAPP-Seq.

Case	Age	Sex	Histology	Grade and Other Histological Features	TNM Stage	Stage Group	Smoker	Pack-Years	Tumor Source	Germline Source	SNVs by Clinical Assays	Fusions Detected by FISH
P1	66	M	Adenocarcinoma	Papillary type	T2aN0M0	IB	Yes	20	FFPE cores	Frozen PBL	KRAS G13D EGFR L858R; EGFR T790M	ALK ALK ALK ROS1 ROS1
P2	61	M	Large Cell	NOS	T3N1M0	IIIA	Yes	80	FFPE cores	Frozen PBL		
P3	67	F	Adenocarcinoma	Acinar type	T1bN3M0	IIIB	Yes	15	FFPE cores	Frozen PBL		
P4	47	F	Adenocarcinoma	Micropapillary and papillary type	T2aN2M1b	IV	Yes	45	FFPE cores	Frozen PBL		
P5	49	F	Adenocarcinoma	Well differentiated	T1bN0M1a	IV	No	0	FFPE cores	Frozen PBL		
P6	54	M	Adenocarcinoma	NOS	T3N2M1b	IV	No	0	Fresh	Frozen PBL		
P7	50	M	Adenocarcinoma	Poorly differentiated	T1aN2M1b	IV	Yes	4	FFPE cores	Frozen PBL		
P8	48	F	Adenocarcinoma	Mucinous type	T4N0M1b	IV	No	0	FFPE cores	Frozen PBL		
P9	49	M	Adenocarcinoma	Not otherwise specified (NOS)	T4N3M1a	IV	No	0	Fresh	Frozen PBL		
P10	35	F	Adenocarcinoma	NOS	T4N0M0	IIIA	No	0	FFPE cores	Frozen PBL		
P11	38	F	Adenocarcinoma	Well-to-moderately differentiated	T3N2M0	IIIA	No	0	FFPE cores	Frozen PBL		

Related to Figures 11 (a) and 14, regarding smoking history, ≥20 pack years was considered heavy and >0 pack years was considered light.

[0111] To explore the potential clinical utility of CAPP-Seq for disease monitoring and minimal residual disease detection, we next applied CAPP-Seq to serial plasma samples collected from a subset of these same 11 patients (N=6), all of whom had pre- and post-treatment samples available (Figure 11; Table 4). Starting from ~15 ng of plasma cfDNA (~3mL of peripheral blood) and sequenced to a mean depth of nearly 8,000x (Table 3), CAPP-Seq detected cancer-derived cfDNA in both early and advanced stage patients (Table 4). Among patients with SNV or indel reporters, all showed a significant reduction in cancer cfDNA burden following treatment, consistent with radiographic response assessment by computed tomography (CT) (Figure 11(a)). These included two patients—one with stage IB adenocarcinoma (P1) and another with

stage IIIA large cell carcinoma (P2)—who underwent surgery with complete tumor resection (Figure 11(b)). Post-treatment cancer-derived cfDNA was undetectable in the Stage I patient but was above background for the Stage IIIA patient suggesting that residual cancer cells remained after surgery even though a complete resection was thought to have been achieved. In a third case (P6), CAPP-Seq detected 3 SNVs and a *KIF5B-ALK* fusion, and both mutation types reported similar fractional abundances of mutant cfDNA (Figure 14). Next, we analyzed a patient with 3 fusions and no detectable SNVs/indels (P9), but from whom 3 serial cfDNA samples were collected. Abundance of fusion product in the plasma was highly correlated with tumor burden and correctly indicated initial response to therapy followed by relapse ($R^2=0.97$; Figure 11(c)). Finally, in a fifth patient (P5), CAPP-Seq identified a sub-clonal population harboring the T790M *EGFR* gatekeeper mutation (Kobayashi et al. (2005) *N. Engl. J. Med.* 352:786-792) (Figure 11(d)). The ratio between clones was identical in the tumor and pre-treatment plasma cfDNA but changed after treatment with cytotoxic chemotherapy followed by a 3rd generation *EGFR* inhibitor (Figure 11(d), inset), suggesting that CAPP-Seq can detect clinically relevant subclones and monitor clonal dynamics during therapy. Taken together, these data demonstrate the potential utility of CAPP-Seq as a noninvasive clinical assay for measuring tumor burden in early and advanced stage NSCLC and for monitoring tumor-derived cfDNA during therapy.

20

Table 4. Monitoring of cfDNA in patients using CAPP-Seq.

Case	Mutant allele	Ref. allele	Chr	Position	Time point 1				Time point 2				Time point 3			
					Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %	Mutant allele depth	Total depth	Mutant allele %	Final %
P1	A	G	chr1	156786560	0	4572	0.000	0.000	3	6202	0.048	0.048	-	-	-	-
P1	T	G	chr1	157805043	0	1838	0.000	0.000	0	2266	0.000	0.000	-	-	-	-
P1	G	C	chr1	248525206	0	2828	0.000	0.000	0	4529	0.000	0.000	-	-	-	-
P1	C	T	chr2	33500291	1	943	0.106	0.106	0	943	0.000	0.000	-	-	-	-
P1	A	C	chr4	55940307	0	5856	0.000	0.000	0	8817	0.000	0.000	-	-	-	-
P1	G	A	chr4	55963940	0	5742	0.000	0.000	0	7335	0.000	0.000	-	-	-	-
P1	A	C	chr4	55969672	0	5856	0.000	0.000	0	7431	0.000	0.000	-	-	-	-
P1	C	T	chr6	117642148	0	5266	0.000	0.000	4	6849	0.058	0.058	-	-	-	-
P1	T	G	chr9	8375700	3	5535	0.054	0.054	0	7322	0.000	0.000	-	-	-	-
P1	T	C	chr9	8725025	1	927	0.121	0.121	0	1398	0.000	0.000	-	-	-	-
P1	T	G	chr10	43611663	0	3722	0.000	0.000	0	4555	0.000	0.000	-	-	-	-
P1	T	G	chr15	88522525	1	4919	0.020	0.020	4	6736	0.059	0.059	-	-	-	-
P1	+G	-	chr17	7578474	0	1752	0.000	0.000	0	2373	0.000	0.000	-	-	-	-
P1	A	G	chr17	28552444	1	4484	0.022	0.022	0	6485	0.000	0.000	-	-	-	-
P1	+T	-	chr17	28553494	0	3557	0.000	0.000	0	4713	0.000	0.000	-	-	-	-
P1	-T	-	chr17	28552185	3	3594	0.081	0.081	0	3247	0.000	0.000	-	-	-	-
P2	A	C	chr2	50453926	49	6724	0.729	0.957	0	4981	0.000	0.000	-	-	-	-
P2	G	A	chr3	89457148	40	4938	0.827	0.957	0	4211	0.000	0.000	-	-	-	-
P2	T	G	chr3	89469296	5	4667	0.107	0.214	2	3625	0.055	0.110	-	-	-	-
P2	T	A	chr3	89480240	15	5073	0.286	0.591	0	4321	0.000	0.000	-	-	-	-
P2	T	A	chr4	66189659	4	950	0.421	0.842	5	1436	0.348	0.696	-	-	-	-
P2	T	C	chr4	66242668	16	2107	0.759	0.759	0	1555	0.000	0.000	-	-	-	-
P2	A	C	chr5	179522747	46	2220	2.072	0.935	0	1377	0.000	0.000	-	-	-	-
P2	C	T	chr6	117648229	70	7819	0.895	0.985	0	5985	0.000	0.000	-	-	-	-
P2	A	C	chr12	78400637	35	7907	0.443	0.885	1	6326	0.016	0.016	-	-	-	-
P2	T	G	chr12	78400910	106	8211	1.291	0.985	1	6289	0.016	0.016	-	-	-	-
P2	T	C	chr17	7577551	112	5529	1.990	0.990	2	3914	0.052	0.104	-	-	-	-
P2	T	G	chr19	1207247	15	1124	1.335	0.935	0	747	0.000	0.000	-	-	-	-
P2	+A	-	chr2	79314100	15	3280	0.468	0.35	0	2390	0.000	0.000	-	-	-	-
P3	A	C	chr17	7578253	6	5345	0.095	0.095	0	8583	0.000	0.000	-	-	-	-
P4	T	C	chr7	55249071	42	4735	0.897	0.957	10	5537	0.179	0.179	-	-	-	-
P5	G	T	chr7	55259515	503	11349	4.432	0.957	58	12222	0.475	0.475	-	-	-	-
P5	A	G	chr11	55135338	86	4063	2.117	0.957	10	4798	0.208	0.208	-	-	-	-
P5	T	C	chr17	7577097	227	7429	3.058	0.957	35	9723	0.370	0.370	-	-	-	-
P6	A	G	chr12	78400791	84	13970	0.601	0.957	28	10128	0.276	0.276	-	-	-	-
P6	T	G	chr12	129822187	78	8680	0.899	0.957	9	6004	0.136	0.136	-	-	-	-
P6	A	G	chr17	7578275	140	9376	1.493	0.957	22	7897	0.279	0.279	-	-	-	-
P6	KIF5B-ALK	-	chr10/chr2	-	28	15006	0.187	0.187	2	9983	0.020	0.020	-	-	-	-
P9	EML4-ALK	-	chr2/chr2	-	0	10588	0.000	0.000	0	13547	0.000	0.000	0	13521	0.000	0.000
P9	FYN-ROS1	-	chr6/chr6	-	0	3261	0.000	0.000	0	6826	0.000	0.000	2	10693	0.019	0.019
P9	ROS1-MKX	-	chr6/chr10	-	10	8029	0.125	0.125	1	6485	0.015	0.015	13	9843	0.131	0.131

Bolded reporters indicate potential homozygous alleles (see Table 3 and Detailed Methods).

Note that mutant cfDNA percentages for P5 were calculated from the 3 SNVs representing the dominant clone (see Figures 11 (a) and 11 (d)); EGFR T790M (chr7:55249071 C->T) was not included.

Final allelic percentages reflect any adjustments made based on estimated zygosity (using inferred homozygous reporters) and/or sequencing coverage. See Detailed Methods for details.

[0112] In addition to its potential clinical utility, CAPP-Seq analysis promises to yield novel biological insights. For example, in one patient's tumor (P9), we identified both a classic *EML4-ALK* fusion and two previously unreported fusions involving *ROS1: FYN-ROS1* and *ROS1-MKX* (Figure 11(e), Figure 15). While the potential function of these novel *ROS1* fusions is unknown, to the best of our knowledge this is the first observation of *ROS1* and *ALK* fusions in the same NSCLC patient. All fusions were confirmed by qPCR amplification of genomic DNA, and were independently recovered in plasma samples (Table 4). Separately, among cases with a *ROS1* rearrangement, we found an unexpected enrichment for S34F missense mutations in *U2AF1*, the 35 kD subunit of the U2 spliceosomal complex auxiliary factor. This SNV was initially described as a recurrent heterozygous mutation in myelodysplastic syndrome (Graubert et al. (2012) *Nat. Genet.* 44:53-57; Yoshida et al. (2011) *Nature* 478:64-69). While *U2AF1* mutations (Imielinski et al. (2012) *Cell* 150:1107-1120) and *ROS1* translocations (Bergeth et al. (2012) *J. Clin. Oncol.* 30:863-870) were recently reported to occur individually in ~3% and ~1.7% of lung adenocarcinomas, respectively, combining the samples we profiled with publicly available data (Detailed Methods), we observed a significant enrichment for *U2AF1* S34F mutations tumors

harboring *ROS1* fusions (in 3 of 6; P=0.0019; Figure 11(f), Figure 16 and Detailed Methods).

[0113] Finally, we explored whether CAPP-Seq analysis of cfDNA could potentially be used for cancer screening. As proof-of-principle, we blinded ourselves to the mutations present in each patient's tumor and developed a statistical method to test for the presence of cancer DNA in each pre-treatment plasma sample in our cohort (Figure 17). This method identified mutant DNA in all plasma samples containing tumor-derived mutant alleles above fractional abundances of 0.5%. Mutant DNA below this level could not be detected by our algorithm, but no mutations were falsely called, indicating the high specificity of this approach (Figure 11(g) and Detailed Methods). Since ~95% of nodules identified in patients at high risk for NSCLC by low-dose CT are false positives (Aberle et al. (2011) *N. Engl. J. Med.* 365:395-409), CAPP-Seq could potentially serve as a complementary noninvasive screening test. However, methodological improvements to further lower the detection threshold will be required to detect early stage tumors.

[0114] In conclusion, we have developed a flexible method for ultrasensitive and specific assessment of circulating tumor DNA. CAPP-Seq overcomes limitations of previously proposed methods for cfDNA analysis by simultaneously measuring multiple types of mutations without patient-specific optimization and by covering mutations in the majority of patients. Moreover, due to multiplexing, CAPP-Seq is highly economical, and per sample costs for plasma cfDNA are expected to drop further as NGS costs continue to fall. Our method has the potential to accelerate the personalized detection, therapy, and monitoring of cancer patients. We anticipate that CAPP-Seq will prove valuable in a variety of clinical settings, including the assessment of cancer DNA in alternative biological fluids and specimens with low cancer cell content.

Methods

Patient Selection

[0115] Between April 2010 and June 2012, patients undergoing treatment for newly diagnosed or recurrent NSCLC were enrolled in a study approved by the Stanford

University Institutional Review Board. Enrolled patients had not received blood transfusions within 3 months of blood collection. Patient characteristics are in Table 3.

Sample Collection and Processing

5 [0116] Peripheral blood from consented patients was collected in EDTA Vacutainer tubes (BD). Blood samples were processed within 3 hours of collection. Plasma was separated by centrifugation at 2,500 x g for 10 min, transferred to microcentrifuge tubes, and centrifuged at 16,000 x g for 10 min to remove cell debris. The cell pellet from the initial spin was used for isolation of germline genomic DNA from PBLs
10 (peripheral blood leukocytes) with the DNeasy Blood & Tissue Kit (Qiagen) . Matched tumor DNA was isolated from FFPE specimens or from the cell pellet of pleural effusions. Genomic DNA was quantified by Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen).

15 *Cell-free DNA Purification and Quantification*

[0117] Cell-free DNA (cfDNA) was isolated from 1-5 mL plasma with the QIAamp Circulating Nucleic Acid Kit (Qiagen). Absolute quantification of purified cfDNA was determined by quantitative PCR (qPCR) using an 81 bp amplicon on chromosome 1 (Fan et al. (2008) *Proc. Natl Acad. Sci. USA* 105:16266-16271) and a dilution series of
20 intact male human genomic DNA (Promega) as a standard curve. Power SyberGreen was used for qPCR on a HT7900 Real Time PCR machine (Applied Biosystems). Standard PCR thermal cycling parameters were used.

Illumina NGS Library Construction

25 [0118] Indexed Illumina NGS libraries were prepared from cfDNA and shorn tumor, germline, and cell line genomic DNA. For patient cfDNA, 7-32ng DNA was used for library construction without additional shearing or fragmentation. For tumor, germline, and cell line genomic DNA, 69-1000ng DNA was sheared prior to library construction with a Covaris S2 instrument using the recommended settings for 200bp fragments. See
30 Table 2 for details.

[0119] The NGS libraries were constructed using the KAPA Library Preparation Kit (Kapa Biosystems) employing a DNA Polymerase possessing strong 3'-5' exonuclease

(or proofreading) activity and displaying the lowest published error rate (i.e. highest fidelity) of all commercially available B-family DNA polymerases (Quail et al. (2012) *Nat. Methods* 9:10-11; Oyola et al. (2012) *BMC Genomics* 13:1). The manufacturer's protocol was modified to incorporate with-bead enzymatic and cleanup steps (Fisher et al. (2011) *Genome Biol.* 12:R1). Briefly, following the end repair reaction, Agencourt AMPure XP beads (Beckman-Coulter) were added to bind and wash the DNA fragments. The DNA was then eluted directly into 50µL 1X A-tailing buffer containing the A-tailing enzyme. Following the A-tailing reaction, the DNA fragments were forced to bind to the same AMPure XP beads by adding 90 µL (1.8X) of PEG buffer (20% PEG-8000 in 2.5M NaCl). After washing, the DNA was eluted into 50µL 1X ligation buffer with ligase and 100-fold molar excess of indexed Illumina TruSeq adapters. Ligation was performed for 16 hours at 16°C. Single-step size selection was performed by adding 40µL (0.8X) of PEG buffer to enrich for ligated DNA fragments. The ligated fragments were then amplified using 500 nM Illumina backbone oligonucleotides and a variable number of PCR cycles (between 4 and 9) depending on input DNA mass. In order to minimize bias and maximize recovery of GC-rich templates, all PCR reactions were carried out in a BioRad DNA Engine Thermal Cycler with a ramp rate of 2.2°C/sec or an Eppendorf Vapo Protect Mastercycler with the Safe ramp rate setting.

[0120] Library purity and concentration was assessed by spectrophotometer (NanoDrop 2000) and qPCR (KAPA Biosystems), respectively. Fragment length was determined on a 2100 Bioanalyzer using the DNA 1000 Kit (Agilent).

Design of Library for Hybrid Selection

[0121] Custom hybrid selection was performed with the SeqCap EZ Choice Library, v2.0 (Roche NimbleGen). The custom SeqCap library was designed through the NimbleDesign portal (v1.2.R1) using genome build HG19 NCBI Build 37.1/GRCh37 and with Maximum Close Matches set to 1. Input genomic regions were selected according to the most frequently mutated genes and exons in NSCLC. These regions were identified from the COSMIC database, TCGA, and other published sources as described in the Detailed Materials. Final selector coordinates are provided in Table 1.

Hybrid Selection and High Throughput Sequencing

[0122] NimbleGen SeqCap EZ Choice was used according to the manufacturer's protocol with modifications. Between 9 and 12 indexed Illumina libraries were included in a single capture reaction. Prior to hybrid selection, the libraries were
5 quantified with a NanoDrop 2000 spectrophotometer, and 83-111 ng of each library was added (1 µg total DNA per capture reaction). Following hybrid selection, the captured DNA fragments were amplified with 12-to-14 cycles of PCR using 1X KAPA HiFi Hot Start Ready Mix and 2 µM Illumina backbone oligonucleotides in 4-to-6 separate 50 µL reactions. The reactions were then pooled and processed with the QIAquick PCR
10 Purification Kit (Qiagen). Multiplexed libraries were sequenced using 2 x 100bp paired-end runs on an Illumina HiSeq 2000.

Mapping and Quality Control of NGS Data

[0123] Paired-end reads were mapped to the hg19 reference genome with BWA 0.6.2
15 (default parameters) (Li & Durbin (2009) *Bioinformatics* 25:1754-1760), and sorted/indexed with SAMtools (Li et al. (2009) *Bioinformatics* 25:2078-2079). QC was assessed using a custom Perl script to collect a variety of statistics, including mapping characteristics, read quality, and selector on-target rate (i.e., number of unique reads that intersect the selector space divided by all aligned reads), generated respectively by
20 SAMtools flagstat, FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), and BEDTools coverageBed (Quinlan & Hall (2010) *Bioinformatics* 26:841-842). Importantly, we used a custom version of coverageBed modified to count each read at most once. Plots of fragment length distribution and sequence depth/coverage were automatically
25 generated for visual QC assessment. To mitigate the impact of sequencing errors, analyses not involving fusions were restricted to properly paired reads, and high-quality bases with a Phred quality score of at least 30 ($\leq 0.1\%$ probability of a sequencing error) were further analyzed.

Analysis of Detection Thresholds by CAPP-Seq

[0124] Two dilution series were performed to assess the linearity and accuracy of CAPP-Seq for quantitating tumor-derived cfDNA. In one experiment, shorn genomic

DNA from a NSCLC cell line (HCC78) was spiked into cfDNA from a healthy individual, while in a second experiment, shorn genomic DNA from one NSCLC cell line (NCI-H3122) was spiked into shorn genomic DNA from a second NSCLC line (HCC78). A total of 32 ng DNA was used for library construction. Following mapping and quality control, homozygous reporters were identified as alleles unique to each sample with at least 20x sequencing depth at an allelic fraction >80%. Fourteen such reporters were identified between HCC78 genomic DNA and plasma cfDNA (Figure 9 (d), (e)), whereas 24 reporters were found between NCI-H3122 and HCC78 genomic DNA (Figure 10).

CAPP-Seq Bioinformatics Pipeline

[0125] Details of bioinformatics methods are supplied in the Detailed Methods, and a graphical schematic is provided in Figure 2. Briefly, for detection of SNVs and indels, we employed VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) with strict post-processing filters to improve variant call confidence, and for fusion identification and breakpoint characterization we used a novel algorithm, termed FACTERA (Detailed Methods). To quantify tumor burden in plasma cfDNA, allele frequencies of reporter SNVs/indels were assessed using the output of SAMtools mpileup (Li et al. (2009) *Bioinformatics* 25:2078-2079), and fusions, if detected, were enumerated with FACTERA.

Statistical Analysis

[0126] The NSCLC selector was validated *in silico* using an independent cohort of lung adenocarcinomas (Imielinski et al. (2012) *Cell* 150:1107-1120) (Figure 1(c)). To assess statistical significance, we analyzed the same cohort using 10,000 random selectors sampled from the exome, each with an identical size distribution to the CAPP-Seq NSCLC selector. The performance of random selectors had a Gaussian distribution, and p-values were calculated accordingly. Note that all identified somatic lesions were considered in this analysis.

[0127] We used Monte Carlo sampling to estimate the distribution of background alleles across the NSCLC selector (Figure 9 (a), (c); Detailed Methods). For each plasma sample, *background* alleles were defined as alleles remaining after exclusion of

germline and/or somatic variant calls made by VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) (somatic p-value=0.01; otherwise, default parameters), and with a Phred quality score ≥ 30 . To evaluate the impact of reporter number on tumor burden estimates, we also performed Monte Carlo sampling (1,000x), varying the number of reporters available $\{1, 2, \dots, \max n\}$ in two spiking experiments (Figure 9 (d)-(f); Figure 10 (b)-(d)).

[0128] To assess the significance of tumor burden estimates in plasma cfDNA, we compared patient-specific SNV frequencies against the null distribution of background SNVs across the selector. Briefly, patient-specific background was quantified using the method described for Figure 9 (a) (Detailed Methods), but using the number of SNVs identified in the patient's tumor. For patients with at least 1 SNV, but no other reporter types, tumor-derived cfDNA was considered *not detectable* if mean SNV fractions fell below the 95th percentile of background alleles (i.e., $P \geq 0.05$) (Figure 11 (a)). (Due to the ultra-low false detection rate for indels (Minoche et al. (2011) *Genome Biol.* 12:R112) and fusion breakpoints, these mutation types were considered detected when present with >0 read support.) For patients with detectable disease in only 1 time point, the corresponding empirical p-value is shown in Figure 11 (a). To assess normality, we analyzed the patient with the most reporter alleles (i.e., P2; Figure 11 (a)), and found that fractional abundance measurements fit a normal distribution (D'Agostino and Pearson omnibus normality test). Thus, for patients with detectable tumor-derived cfDNA in two time points and with at least 3 cfDNA SNVs/indels, the change in tumor burden was statistically assessed using a two-sided paired t-test. For P9, who lacked reporter SNVs/indels, statistical significance was estimated by correlation of CAPP-Seq measurements with known tumor volume (as measured by CT scans).

[0129] Additional details on cell lines, tumor cell sorting, optimizations of library preparation, mutation/translocation validation, CAPP-Seq design and analytical pipelines including FACTERA translocation detection tool, and additional statistical methods are presented in the Detailed Methods.

Detailed Methods

A. Molecular Biology Methods

A1. Cell Lines

[0130] The lung adenocarcinoma cell lines NCI-H3122 and HCC78 were obtained from ATCC and DSMZ, respectively, and grown in RPMI 1640 with L-glutamine (Gibco) supplemented with 10% fetal bovine serum (Gembio) and 1% penicillin/streptomycin cocktail. Cells were maintained in mid-log-phase growth in a 37°C incubator with 5% CO₂. Genomic DNA was purified from freshly harvested cells with the DNeasy Blood & Tissue Kit (Qiagen).

A2. Pleural Fluid Processing and Flow Cytometry, and Cell Sorting

[0131] Cells from pleural fluid from patients P9 and P6 were harvested by centrifugation at 300 x g for 5 min at 4°C and washed in FACS staining buffer (HBSS + 2% heat-inactivated calf serum [HICS]). Red blood cells were lysed with ACK Lysing Buffer (Invitrogen), and clumps were removed by passing through a 100 µm nylon filter. Filtered cells were spun down and resuspended in staining buffer. While on ice, the cell suspension was blocked for 20 min with 10 µg/mL rat IgG and then stained for 20 min with APC-conjugated mouse anti-human EpCAM (BioLegend, clone 9C4), PerCP-Cy5.5-conjugated mouse anti-human CD45 (eBioscience, clone 2D1), and PerCP-eFluor710-conjugated mouse anti-human CD31 (eBioscience, clone WM59). After staining, cells were washed and resuspended with staining buffer containing 1 µg/mL DAPI, analyzed, and sorted with a FACS Aria II cell sorter (BD Biosciences). Cell doublets and DAPI-positive cells were excluded from analysis and sorting. CD31⁻CD45⁻EpCAM⁺ cells were sorted into staining buffer, spun down, and flash frozen in liquid nitrogen. DNA was isolated with the QIAamp DNA Micro Kit (Qiagen).

A3. Optimization of NGS Library Preparation from Low Input cfDNA

[0132] Any method for detecting mutant cfDNA relies on its ability to interrogate each cfDNA molecule in the circulation in order to maximize sensitivity. For this reason, we used the QIAamp Circulating Nucleic Acid kit (Qiagen) with carrier RNA as per the manufacturer's protocol to isolate cfDNA. We also took specific steps to improve the Illumina library preparation workflow.

[0133] Protocols for Illumina library construction were compared in a step-wise manner with the goal of (1) optimizing adapter ligation efficiency, (2) reducing the necessary number of PCR cycles following adapter ligation, (3) preserving the naturally occurring size distribution of cfDNA fragments, and (4) minimizing variability in depth of sequencing coverage across all captured genomic regions. Initial optimization was done with NEBNext DNA Library Prep Reagent Set for Illumina (New England BioLabs), which includes reagents for end-repair of the cfDNA fragments, A-tailing, adapter ligation, and amplification of ligated fragments with Phusion High-Fidelity PCR Master Mix. Input was 4 ng cfDNA (obtained from plasma of the same healthy volunteer) for all conditions. Relative allelic abundance in the constructed libraries was assessed by qPCR of 4 genomic loci (Roche NimbleGen: NSC-0237, NSC-0247, NSC-0268, and NSC-0272) and compared by the $2^{-\Delta C_t}$ method.

[0134] Ligations were performed at 20°C for 15 min (as per the manufacturer's protocol), at 16°C for 16 hours, or with temperature cycling for 16 hours as previously described (Lund et al. (1996) *Nucl. Acids Res.* 24:800-801). Ligation volumes were varied from the standard (50 μ L) down to 10 μ L while maintaining a constant concentration of DNA ligase, cfDNA fragments, and Illumina adapters. Subsequent optimizations incorporated ligation at 16°C for 16 hours in 50 μ L reaction volumes.

[0135] Next, we compared standard SPRI bead processing procedures, in which new AMPure XP beads are added after each enzymatic reaction and DNA is eluted from the beads for the next reaction, to with-bead protocol modifications as previously described (Fisher, S. et al. (2011) *Genome Biol.* 12:R1). We compared 2 concentrations of Illumina adapters in the ligation reaction: 12 nM (10-fold molar excess to cfDNA fragments) and 120 nM (100-fold molar excess).

[0136] Using the optimized library preparation procedures, we next compared the NEBNext DNA Library Prep Reagent Set (with Phusion DNA Polymerase) to the KAPA Library Preparation Kit (with KAPA HiFi DNA Polymerase). The KAPA Library Preparation Kit with our modifications was also compared to the NuGEN SP Ovation Ultralow Library System with automation on Mondrian SP Workstation.

30

A4. Evaluation of Library Preparation Modifications on CAPP-Seq Performance

[0137] We performed CAPP-Seq on 32 ng cfDNA using standard library preparation procedures with the NEBNext kit, or with optimized procedures using either the NEBNext kit or the KAPA Library Preparation Kit. In parallel we performed CAPP-Seq on 4 ng and 128 ng cfDNA using the KAPA kit with our optimized procedures. Indexed libraries were constructed, and hybrid selection was performed in multiplex. The post-capture multiplexed libraries were amplified with Illumina backbone primers for 14 cycles of PCR and then sequenced on a paired-end 100 bp lane of an Illumina HiSeq 2000.

[0138] We also evaluated CAPP-Seq on ultralow input following whole genome amplification (WGA). For WGA we chose not to use multiple displacement amplification with Φ 29 DNA polymerase due given the small size of cfDNA fragments in plasma (Figure 1(e)), and due to concern for chimera formation, which would confound analysis of recurrent gene fusions in NSCLC by CAPP-Seq. Instead we used SeqPlex DNA Amplification Kit (Sigma-Aldrich), which employs degenerate oligonucleotide primer PCR. We used the upper limit of input into this kit (1 ng) and performed whole genome amplification according to the manufacturer's protocol. Briefly, 1 ng cfDNA was amplified with real-time monitoring with SYBR Green I (Sigma-Aldrich) on a HT7900 Real Time PCR machine (Applied Biosystems). The amplification was terminated after 17 cycles yielding 2.8 μ g DNA. The primer removal step yielded ~600 ng DNA, and this entire amount was used for library preparation using the NEBNext kit with optimized procedures as described above.

A5. Validation of Variants Detected by CAPP-Seq

[0139] All structural rearrangements and a subset of tumoral SNVs detected by CAPP-Seq were independently confirmed by qPCR and/or Sanger sequencing of amplified fragments. For HCC78, a 120 bp fragment containing the *SLC34A2-ROS1* breakpoint was amplified from genomic DNA using the primers: 5'-AGACGGGAGAAAATAGCACC-3' and 5'-ACCAAGGGTTGCAGAAATCC-3'. A 141 bp fragment containing exon 2 of *U2AF1* was amplified using the primers: 5'-CATGTGTTTGATATCTTCCCAGC-3' and 5'-CTGGCTAAACGTCGGTTTATTG-3'. For NCI-H3122, a 143 bp fragment containing the *EML4-ALK* breakpoint was

amplified using the primers: 5'-GAGATGGAGTTTCACTCTTGTTGC-3' and 5'-GAACCTTTCCATCATACTTAGAAATAC-3'. 5 ng genomic DNA was used as template with 250 nM oligos and 1X Phusion PCR Master Mix (NEB) in 50 μ L reactions. Products were resolved on 2.5% agarose gel and bands of the expected size were removed. The amplified DNA fragments were purified using the Qiaquick Gel Extraction Kit (Qiagen) and submitted for Sanger sequencing (Elim Biopharm). For P9, genomic DNA breakpoints were confirmed by qPCR using the primers: 5'-TCCATGGAAGCCAGAAC-3' and 5'-ATGCTAAGATGTGTCTGTCA-3' for *EML4-ALK*; 5'-CCTTAACACAGATGGCTCTTGATGC-3' and 5'-TCCTCTTTCCACCTTGGCTTTCC-3' for *ROS1-MKX*; and 5'-GGTTCAGAACTACCAATAACAAG-3' and 5'-ACCTGATGTGTGACCTGATTGATG-3' for *FYN-ROS1*. For qPCR, 10 ng of pre-amplified genomic DNA was used as template with 250 nM oligos and 1X Power SyberGreen Master Mix in 10 μ L reactions performed in triplicate on a HT7900 Real Time PCR machine (Applied Biosystems). Standard PCR thermal cycling parameters were used. Amplification of amplicons spanning all 3 breakpoints detected in P9 were confirmed in tumor genomic DNA as well as plasma cfDNA, and PBL genomic DNA was used as a negative control. Separately, at least 88% of SNVs and indels detected were bona fide somatic mutations in tumors, as 38 of 46 of them were independently observed above 0.025% allele frequency in plasma cfDNA and/or were independently confirmed by SNaPshot clinical assays.

B. Bioinformatics and Statistical Methods

B1. Analysis of CAPP-Seq Background

[0140] The CAPP-Seq background rate was estimated by Monte Carlo sampling of allelic frequencies across the NSCLC selector (Figure 9 (a)). Plasma cfDNA samples were pre-filtered to remove all variant calls and dominant alleles. Specifically, for each patient, we excluded germline, loss of heterozygosity (LOH), and/or somatic variant calls made by VarScan 2 (Koboldt et al. (2012) *Genome Res.* 22:568-576) (somatic p-value=0.01; otherwise, default parameters). We sampled 4 random background alleles across this subset of the selector (equal to the median number of SNVs per NSCLC patient detected by CAPP-Seq) and calculated their mean allelic frequency, only

considering bases discordant with the prevailing genotype of the plasma sample at those 4 positions. This process was iterated 10,000 times, and mean, median, and 75th percentile statistics were collected. The entire procedure was then repeated for 5 total simulations, shown in Figure 2a.

5 [0141] We likewise applied Monte Carlo simulation to estimate the probability of finding a background allele in plasma cfDNA at a given fractional abundance (Figure 9 (c)). For consistency with the ranking of alleles in Figure 9 (c), we populated a vector containing the mean background allele frequency for each genomic position across 7 plasma cfDNA samples, each filtered to remove dominant alleles as described above.
10 Alleles were randomly sampled from this vector 10,000 times to identify the allele frequency with an empirical p-value of 0.01.

B2. ROS1 and U2AF1 Co-association Analysis

B2.1 Assembly of ROS1 and U2AF1 mutant NSCLC

15 [0142] We included only cases in which the status of both *ROS1* fusion status and *U2AF1* S34 mutation was known. There were 163 such cases from TCGA (genotyped for *U2AF1* by whole exome sequencing and for *ROS1* fusions by RNA-Seq as detailed below), 23 cases from Imielinski et al. (2012) *Cell* 150:1107-1120, 17 cases from Govindan et al. (2012) *Cell* 150:1121-1134, and 13 cases from the present study (11
20 patients and 2 NSCLC cell lines). *U2AF1* S34F mutations were detected in 11 cases (5 from TCGA, 3 from Imielinski et al., 1 from Govindan et al., and 2 from the present study), and *ROS1* fusions were detected in 6 cases (2 from TCGA, described below, and 4 from the present study). Significance testing was performed using the Fisher's exact test, and a two-tailed P-value is reported.

25

B2.2. Analysis of Whole Transcriptome Sequencing Data from TCGA for ROS1 fusions

[0143] We identified two TCGA lung adenocarcinoma patients, TCGA-05-4426 and TCGA-64-1680, harboring candidate *ROS1* fusions (Figure 16 (a)). Importantly, the latter patient also has the *U2AF1* S34F missense mutation reported in this study and in
30 prior literature (see above). To further analyze both patients' putative rearrangements, whole transcriptome RNA-Seq data (.bam files) were obtained using the UCSC GeneTorrent system (<https://cghub.ucsc.edu/downloads.html>) and realigned to hg19

using BWA 0.6.2 using default parameters (Li & Durbin (2009) *Bioinformatics* 25:1754-1760). Importantly, mapped RNA-Seq reads extended significantly past coding regions, allowing for improved assessment of fusion events (Figure 16 (b), (c)). From a manual inspection of associated RPKM expression data across *ROS1* exons (Figure 16 (a)), we suspected that breakpoint sites for these fusions may lie directly upstream of *ROS1* exons 32 and 35, respectively. Using the Integrated Genome Viewer (IGV) (Robinson et al. (2011) *Nat. Biotechnol.* 29:24-26), we found improperly paired (or discordant) reads near these exons that link *ROS1* to its well-described partners, *SLC34A2* and *CD74*, respectively (Figure 16 (b), (c)). Indeed, by applying FACTERA's templated fusion discovery (detailed below) to patient TCGA-64-1680, we recovered a single read near *ROS1* exon 35 that also maps to *CD74* (Figure 16 (c)). Collectively, these data strongly support the existence of expressed *ROS1* fusions in these two TCGA patients.

B3. CAPP-Seq Selector Design

[0144] Most human cancers are relatively heterogeneous for somatic mutations in individual genes. Specifically, in most human tumors, recurrent somatic alterations of single genes account for a minority of patients, and only a minority of tumor types can be defined using a small number of recurrent mutations (<5-10) at predefined positions. Therefore, the design of the selector is vital to the CAPP-Seq method because (1) it dictates which mutations can be detected in with high probability for a patient with a given cancer, and (2) the selector size (in kb) directly impacts the cost and depth of sequence coverage. For example, the hybrid selection libraries available in current whole exome capture kits range from 51-71 Mb, providing ~40-60 fold maximum theoretical enrichment versus whole genome sequencing. The degree of potential enrichment is inversely proportional to the selector size such that for a ~100kb selector, >10,000 fold enrichment should be achievable.

[0145] We employed a six-phase design strategy to identify and prioritize genomic regions for the CAPP-Seq NSCLC selector as detailed below. Three phases were used to incorporate known and suspected NSCLC driver genes, as well as genomic regions known to participate in clinically actionable fusions (phases 1, 5, 6), while another three phases employed an algorithmic approach to maximize both the number of

patients covered and SNVs per patient (phases 2-4). The latter relied upon a metric that we termed "Recurrence Index" (RI), defined as the number of NSCLC patients with SNVs that occur within a given kilobase of exonic sequence (i.e., No. of patients with mutations / exon length in kb). RI thus serves to measure patient-level recurrence frequency at the exon level, while simultaneously normalizing for gene/exon size. As a source of somatic mutation data uniformly genotyped across a large cohort of patients, in phases 2-4, we analyzed non-silent SNVs identified in TCGA whole exome sequencing data from 178 patients in the Lung Squamous Cell Carcinoma dataset (SCC) (Hammerman et al. (2012) *Nature* 489:519-525) and from 229 patients in the Lung Adenocarcinoma (LUAD) datasets (TCGA query date was March 13, 2012). Thresholds for each metric (i.e. RI and patients per exon) were selected to statistically enrich for known/suspected drivers in SCC and LUAD data (Figure 9). RefSeq exon coordinates (hg19) were obtained via the UCSC Table Browser (query date was April 11, 2012).

[0146] The following algorithm was used to design the CAPP-Seq selector (parenthetical descriptions match design phases noted in Figure 1 (b)).

- Phase 1 (*Known drivers*)

[0147] Initial seed genes were chosen based on their frequency of mutation in NSCLCs. Analysis of COSMIC (v57) (Forbes et al. (2010) *Nucl. Acids Res.* 38:D652-657) identified known driver genes that are recurrently mutated in $\geq 9\%$ of NSCLC (denominator ≥ 500 cases). Specific exons from these genes were selected based on the pattern of SNVs previously identified in NSCLC. The seed list also included single exons from genes with recurrent mutations that occurred at low frequency but had strong evidence for being driver mutations, such as BRAF exon 15, which harbors V600E mutations in $<2\%$ of NSCLC (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181; Okuda et al. (2008) *Cancer Sci.* 99:2280-2285; Su et al. (2011) *J. Mol. Diagn.* 13:74-84; Tsao et al. (2007) *J. Clin. Oncol.* 25:5240-5247; Chaft et al. (2012) *Mol. Cancer Ther.* 11:485-491; Paik et al. (2011) *J. Clin. Oncol.* 29:2046-2051; Stephens et al. (2004) *Nature* 431:525-526; Jin et al. (2010) *Lung Cancer* 69:279-283; Malanga et al. (2008) *Cell Cycle* 7:665-669).

- Phase 2 (*Max. coverage*)

[0148] For each exon with SNVs covering ≥ 5 patients in LUAD and SCC, we selected the exon with highest RI that identified at least 1 new patient when compared to the prior phase. Among exons with equally high RI, we added the exon with minimum overlap among patients already captured by the selector. This was repeated until no further exons met these criteria.

- Phase 3 ($RI \geq 30$)

[0149] For each remaining exon with an $RI \geq 30$ and with SNVs covering ≥ 3 patients in LUAD and SCC, we identified the exon that would result in the largest reduction in patients with only 1 SNV. To break ties among equally best exons, the exon with highest RI was chosen. This was repeated until no additional exons satisfied these criteria.

- Phase 4 ($RI \geq 20$)

[0150] Same procedure as phase 3, but using $RI \geq 20$.

- Phase 5 (*Predicted drivers*)

[0151] We included all exons from additional genes previously predicted to harbor driver mutations in NSCLC (Ding et al. (2008) *Nature* 455:1069-1075; Youn & Simon (2011) *Bioinformatics* 27:175-181).

- Phase 6 (*Add fusions*)

[0152] For recurrent rearrangements in NSCLC involving the receptor tyrosine kinases *ALK*, *ROS1*, and *RET*, the introns most frequently implicated in the fusion event and the flanking exons were included.

[0153] All exons included in the selector, along with their corresponding HUGO gene symbols and genomic coordinates, as well as patient statistics for NSCLC and a variety of other cancers, are provided in Table 1, organized by selector design phase.

C. CAPP-Seq Computational Pipeline

C1. Mutation Discovery: SNVs/indels

[0154] For detection of somatic SNV and insertion/deletion events, we employed VarScan 2 (Koboldt et al. (2012) *Genome Res* 22:568-576) (somatic p-value=0.01, minimum variant frequency=5%, and otherwise default parameters). Somatic variant calls (SNV or indel) present at less than 0.5% mutant allelic frequency in the paired normal sample (PBLs), but in a position with at least 1000x overall depth in PBLs and 100x depth in the tumor, and with at least 1x read depth on each strand, were retained (Table 3). While the selector was designed to predominantly capture exons, in practice, it also captures limited sequence content flanking each targeted region. For instance, this phenomenon is the basis for the (thus far) uniformly successful recovery by CAPP-Seq of fusion partners (which are not included within the selector) for kinase genes such as *ALK* and *ROS1* recurrently rearranged in NSCLC. As such, we also considered variant calls detected within 500bps of defined selector coordinates. These calls were eliminated if present in non-coding *repeat* regions, since repeats may confound mapping accuracy. Repeat sequence coordinates were obtained using the RepeatMasker track in the UCSC table browser (hg19). Variant annotation was performed using the SeattleSeq Annotation 137 web server (<http://snp.gs.washington.edu/SeattleSeqAnnotation137/>). Complete details for all identified SNVs and indels are provided in Table 2.

[0155] By manual inspection, two patients (P2 and P6) had SNVs with frequencies consistent with potential heterozygous and homozygous alleles. We labeled these alleles accordingly (Table 3), and based on our assumption of zygosity in these two patients, we adjusted measured fractions of heterozygous reporters in plasma cfDNA to better estimate tumor burden (Table 4).

C2. Mutation Discovery: Fusions

[0156] For practical and robust de novo enumeration of genomic fusion events and breakpoints from paired-end next-generation sequencing data, we developed a novel heuristic approach, termed FACTERA (FACile Translocation Enumeration and Recovery Algorithm). FACTERA has minimal external dependencies, works directly on a preexisting .bam alignment file, and produces easily interpretable output. Major

steps of the algorithm are summarized below, and are complemented by a graphical schematic to illustrate key elements of the breakpoint identification process (Figure 4).

[0157] As input, FACTERA requires a .bam alignment file of paired-end reads produced by BWA (Li & Durbin (2009) *Bioinformatics* 25:1754-1760), exon coordinates in .bed format (e.g., hg19 RefSeq coordinates), and a .2bit reference genome to enable fast sequence retrieval (e.g., hg19). In addition, the analysis can be optionally restricted to reads that overlap particular genomic regions (.bed file), such as the CAPP-Seq selector used in this work.

[0158] FACTERA processes the input in three sequential phases: identification of discordant reads, detection of breakpoints at base pair-resolution, and *in silico* validation of candidate fusions. Each phase is described in detail below.

C2.1. Identification of discordant reads

[0159] To iteratively reduce the sequence space for gene fusion identification, FACTERA, like other algorithms (e.g. BreakDancer (Chen et al. (2009) *Nat. Methods* 6:677-681)), identifies and classifies discordant read pairs. Such reads indicate a nearby fusion event since they either map to different chromosomes or are separated by an unexpectedly large insert size (i.e. total fragment length), as determined by the BWA mapping algorithm. The bitwise flag accompanying each aligned read encodes a variety of mapping characteristics (e.g., improperly paired, unmapped, wrong orientation, etc.) and is leveraged to rapidly filter the input for discordant pairs. The closest exon of each discordant read is subsequently identified, and used to cluster discordant pairs into distinct gene-gene groups, yielding a list of genomic regions R adjacent to candidate fusion sites. For each member gene of a discordant gene pair, the genomic region R_i is defined by taking the minimum of all 3' exon/read coordinates in the cluster, and the maximum of all 5' exon/read coordinates in the cluster. These regions are used to prioritize the search for breakpoints in the next phase (Figure 4 (a)).

C2.2 Detection of breakpoints at base pair-resolution

[0160] Discordant read pairs may be introduced by NGS library preparation and/or sequencing artifacts (e.g., jumping PCR). However, they are also likely to flank the breakpoints of *bona fide* fusion events. As such, all discordant gene pairs identified in

the preceding phase are ranked in decreasing order of discordant read depth (duplicate fragments are eliminated to correct for possible PCR bias), and genomic regions with a depth of at least 2x (by default) are further evaluated for potential breakpoints. Within each region, FACTERA analyzes all properly paired reads in which one of the two

5 reads is "soft-clipped," or truncated (e.g., see Figure 4 (a)). Soft-clipped reads allow for precise breakpoint determination, and are easily identified by parsing the CIGAR string associated with each mapped read, which compactly specifies the alignment operation used on each base (e.g. My = y contiguous bases were mapped, Sx = x bases were skipped). To simplify this step, only soft-clipped reads with the following two

10 patterns are considered, SxMy and MySx, and the number of skipped bases x is required to be at least 16 (≤ 1 in 4.3B by random chance) to reduce the impact of non-specific sequence alignments.

[0161] To validate potential genomic breakpoints, defined as the edges of soft-clipped reads, FACTERA executes the following routine, depicted in Figure 4. For

15 each discordant gene pair (e.g. genes w and v in Figure 4 (a)), all candidate breakpoints are tabulated, and the support (i.e. read frequency) for each is determined. Breakpoints supported by less than 2 reads (by default) are excluded from further analysis. Starting with the two breakpoints with highest support, FACTERA selects a representative soft-clipped read for each breakpoint, such that the length of the clipped sequence is closest

20 to half of the read length (Figure 4(b)). If the mapped region of one read matches the soft-clipped region of the other, FACTERA records a putative fusion event. To assess inter-read concordance (e.g. see reads 1 and 2 in Figure 4 (c)), FACTERA employs the following algorithm. The mapped region of read 1 is parsed into all possible subsequences of length k (i.e., k-mers) using a sliding window (k=10, by default). Each

25 k-mer, along with its lowest sequence index in read 1, is stored in a hash table data structure, allowing k-mer membership to be assessed in constant time (Figure 4 (c), left panel). Subsequently, the soft clipped sequence of read 2 is parsed into non-overlapping subsequences of length k, and the hash table is interrogated for matching k-mers (Figure 4 (c), right panel). If a minimum matching threshold is achieved ($=0.5 \times$

30 the minimum length of the two compared subsequences), then the two reads are considered concordant. FACTERA will process at most 1000 (by default) putative breakpoint pairs for each discordant gene pair. Moreover, for each gene pair,

FACTERA will only compare reads whose orientations are compatible with valid fusions. Such reads have soft-clipped sequences facing opposite directions (Figure 4 (d), top panel). When this condition is not satisfied, FACTERA uses the reverse complement of read 1 for k-mer analysis (Figure 4 (d), bottom panel).

5 [0162] In some instances, genomic subsequences flanking the true breakpoint may be nearly or completely identical, causing the aligned portions of soft-clipped reads to overlap. Unfortunately, this prevents an unambiguous determination of the breakpoint. As such, FACTERA incorporates a simple algorithm to arbitrarily adjust the breakpoint in one read (i.e., read 2) to match the other (i.e., read 1). Depending upon read
10 orientation, there are two ways this can occur, both of which are illustrated in Figure 4 (e). For each read, FACTERA calculates the distance between the breakpoint and the read coordinate corresponding to the first k-mer match between reads. For example, as anecdotally illustrated in Figure 4 (e), x is defined as the distance between the breakpoint coordinate of read 1 and the index of the first matching k-mer, j , whereas y
15 denotes the corresponding distance for read 2. The offset is estimated as the difference in distances (x, y) between the two reads (see Figure 4 (e)).

C2.3. In silico validation of candidate fusions

[0163] To confirm each candidate breakpoint *in silico*, FACTERA performs a local
20 realignment of reads against a template fusion sequence (± 500 bp around the putative breakpoint) extracted from the .2bit reference genome. BLAST is currently employed for this purpose, although BLAT or other fast aligners could be substituted. A BLAST database is constructed by collecting all reads that map to each candidate fusion sequence, including discordant reads and soft-clipped reads, as well as all unmapped
25 reads in the original input .bam file. All reads that map to a given fusion candidate with at least 95% identity and a minimum length of 90% of the input read length (by default) are retained, and reads that span or flank the breakpoint are counted. As a final step, output redundancies are minimized by removing fusion sequences within a 20bp interval of any fusion sequence with greater read support and with the same sequence
30 orientation (to avoid removing reciprocal fusions).

[0164] FACTERA produces a simple output text file, which includes for each fusion sequence, the gene pair, the chromosomal sequence coordinates of the breakpoint, the

fusion orientation (e.g., forward-forward or forward-reverse), the genomic sequences within 50bp of the breakpoint, and depth statistics for reads spanning and flanking the breakpoint. Fusions identified in patients analyzed in this work are provided in Table 3.

5 C2.4. Experimental validation of FACTERA

[0165] To experimentally evaluate the performance of FACTERA, we generated NGS data from two NSCLC cell lines, HCC78 (21.5M x 100bp paired-end reads) and NCI-H3122 (19.4M x 100bp paired-end reads), each of which has a known rearrangement (*ROS1* and *ALK*, respectively) (Bergethon et al. (2012) *J. Clin. Oncol.* 30:863-870; McDermott et al. (2008) *Cancer Res.* 68:3389-3395) with a breakpoint that has, to the best of our knowledge, not been previously published. FACTERA readily revealed evidence for a reciprocal *SLC34A2-ROS1* translocation in the former and an *EML4-ALK* fusion in the latter. Precise breakpoints predicted by FACTERA were experimentally validated by PCR amplification and Sanger sequencing (Figure 5; see also *Validation of Variants Detected by CAPP-Seq*). Importantly, FACTERA completed each run in practical time (~90 sec), using only a single thread on a hexa-core 3.4 GHz Intel Xeon E5690 chip. These initial results illustrate the utility of FACTERA as part of the CAPP-Seq analysis pipeline.

20 C2.5. Templated fusion discovery

[0166] We implemented a user-directed option to "hunt" for fusions within expected candidate genes. A fusion could be missed by FACTERA if the fusion detection criteria employed by FACTERA are incompletely satisfied—such as if discordant reads, but not soft-clipped reads, are identified—and will most likely occur when fusion allele frequency in the tumor is extremely low. As input, the method is supplied with candidate fusion gene sequences as "baits". All unmapped and soft-clipped reads in the input .bam file are subsequently aligned to these templates (using blastn) to identify reads that have sufficient similarity to both (for each read, 95% identity, e-value < 1.0e-5, and at least 30% of the read length must map to the template, by default). Such reads are output as a list to the user for manual analysis.

[0167] We tested this simple approach on a low purity tumor sample found to harbor an *ALK* fusion by FISH, but not FACTERA (i.e., case P9). Using templates for *ALK*

and its common fusion partner, *ELM4*, we identified 4 reads that mapped to both, in a region with an overall depth of ~1900x. The estimated allele frequency of 0.21% is strikingly similar to the 0.22% tumor purity measured by FACS (Figure 15), confirming the utility of the templated fusion discovery method. We subsequently FACS-depleted CD45+ immune populations and re-sequenced this patient's tumor. In the enriched tumor sample, FACTERA identified the *EML4-ALK* fusion, along with two novel ROS1 fusions (Figure 4 (e), Table 3).

C3. Mutation Recovery: SNVs/indels

[0168] Using a custom Perl script, previously identified reporter alleles were intersected with a SAMtools mpileup file generated for each plasma cfDNA sample, and the number and frequency of supporting reads was calculated for each reporter allele. Only reporters in properly paired reads at positions with at least 500x overall depth were considered.

C4. Mutation Recovery: Fusions

[0169] For enumeration of fusion frequency in sequenced plasma DNA, FACTERA executes the last step of the discovery phase (i.e., *in silico validation of candidate fusions*, above) using the set of previously identified fusion templates. The fusion allele frequency is calculated as α / β , where α is the number of breakpoint-spanning reads, and β is the mean overall depth within a genomic region ± 5 bps around the breakpoint. Regarding the NSCLC selector described in this work, the latter calculation was always performed on the single gene contained in the NSCLC selector library. If both fusion genes are targeted within a selector library, overall depth is estimated by taking the mean depth calculated for both genes.

[0170] Notably, in some cases we observed lower fusion allele frequencies than would be expected for heterozygous alleles (e.g., see cell line fusions in Table 3). This was seen in cell lines, in an empirical spiking experiment, and in one patient's tumor and plasma samples (i.e., P6), and could potentially result from inefficient "pull-down" of fusions whose partners are not represented in the selector. Regardless, fusions are useful reporters—they possess virtually no background signal and show linear behavior over defined concentrations in a spiking experiment (Figure 10 (d)). Moreover, allelic

frequencies in plasma are easily adjusted for such inefficiencies by dividing the measured frequency in plasma by the corresponding frequency in the tumor. In cases where sequenced tumor tissue is impure, tumor content can be estimated using the frequencies of SNVs (or indels) as a reference frame, allowing the fusion fraction to be normalized accordingly (Table 4). As for SNVs/indels, only fusions present in at least one plasma sample were included in calculations of tumor burden.

C5. Screening Plasma cfDNA without Knowledge of Tumor DNA

[0171] We devised the following statistical algorithm as a first step toward non-invasive cancer screening with plasma cfDNA. The method identifies candidate SNVs using iterative models of (i) background noise in paired germline DNA (in this work, PBLs), (ii) base-pair resolution background frequencies in plasma cfDNA across the selector, and (iii) sequencing error in cfDNA. Anecdotal examples are provided in Figure 17. The algorithm works in four main steps, detailed below.

[0172] As input, the algorithm takes allele frequencies from a single plasma cfDNA sample and analyzes high quality *background alleles*, defined in a first step for each genomic position as the non-dominant base with highest fractional abundance. Only alleles with depth of at least 500x and strand bias <90% (conservative, by default) are analyzed. For consistency with variant calling, we allowed the screening approach to interrogate selector regions within 500bp of defined coordinates, expanding the effective sequence space from ~125kb to ~600kb.

[0173] Second, the binomial distribution is used to test whether a given input cfDNA allele is significantly different from the corresponding paired germline allele (Figure 17 (a)-(b)). Here the probability of success is taken to be the frequency of the background allele in PBLs, and the number of trials is the allele's corresponding depth in plasma cfDNA. To avoid contributions from alleles in rare circulating tumor cells that might contaminate PBLs, input alleles with a fractional abundance greater than 0.5% in paired PBLs (by default) or a Bonferroni-adjusted binomial probability greater than 2.08×10^{-8} are not further considered (alpha of $0.05 / [\sim 600\text{kb} * 4 \text{ alleles per position}]$).

[0174] Third, a database of cfDNA background allele frequencies is assembled. Here, we used samples analyzed in the present study (i.e., pre-treatment NSCLC samples and

1 sample from a healthy volunteer), except the input sample is left out to avoid bias. Based on the assumption that all background allele fractions follow a normal distribution, a Z-test is employed to test whether a given input allele differs significantly from typical cfDNA background at the same position (Figure 17 (a)-(b)).

5 All alleles within the selector are evaluated, and those with an average background frequency of 5% or greater (by default) or a Bonferroni-adjusted single-tailed Z-score < 5.6 are not further considered (alpha of 0.05, adjusted as above).

[0175] Finally, candidate alleles are tested for remaining possible sequencing errors. This step leverages the observation that non-tumor variants (i.e., "errors") in plasma
10 cfDNA tend to have a higher duplication rate than *bona fide* variants detectable in the patient's tumor (data not shown). As such, the number of supporting reads is compared for each input allele between nondeduped (all fragments meeting QC criteria; see Methods) and deduped data (only unique fragments meeting QC criteria). An outlier analysis is then used to distinguish candidate tumor-derived SNVs from remaining
15 background noise (Figure 17 (a)-(c)). Specifically, to reveal outlier tendency in the data, the square root of the robust distance Rd (Mahalanobis distance) is compared against the square root of the quantiles of a chi-squared distribution Cs . This transformation reveals natural separation between true SNVs and false positives in cancer patients (Figure 17 (a), (c)), and notably, reveals an absence of outlier structure
20 in patient samples lacking tumor-derived SNVs (Figure 17 (b), (c)). To automatically call SNVs without prior knowledge, the screening approach iterates through data points by decreasing Rb and recalculating the Pearson's correlation coefficient Rho between Rd and Cs for points 1 to i , where Rd_i is the current maximum Rd . The algorithm iteratively reports outliers (i.e., candidate SNVs) until it terminates when $Rho \geq 0.85$.

25 [0176] Importantly, this approach positively identified 60% of the cancer samples with tumor-derived SNVs analyzed in this study with no false positive calls (Figure 11 (g)). When corresponding germline DNA from PBLs are unavailable, one can skip the 2nd step in this screening routine. After removal of germline SNVs with an allelic fraction $> 20\%$, this modified approach identified no SNVs when applied to a healthy
30 volunteer.

[0177] All patents, patent publications, and other published references mentioned herein are hereby incorporated by reference in their entireties as if each had been individually and specifically incorporated by reference herein.

[0178] While specific examples have been provided, the above description is

5 illustrative and not restrictive. Any one or more of the features of the previously described embodiments can be combined in any manner with one or more features of any other embodiments in the present invention. Furthermore, many variations of the invention will become apparent to those skilled in the art upon review of the specification. The scope of the invention should, therefore, be determined by reference
10 to the appended claims, along with their full scope of equivalents.

What is Claimed is:

1. A method for creating a library of recurrently mutated genomic regions comprising:
 - identifying a plurality of genomic regions from a group of genomic regions that are recurrently mutated in a specific cancer;
 - wherein the library comprises the plurality of genomic regions;
 - the plurality of genomic regions comprises at least 10 different genomic regions; and
 - at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.
2. The method of claim 1, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.
3. The method of claim 1, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.
4. The method of claim 1, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.
5. The method of claim 1, wherein the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.
6. The method of claim 1, wherein the identifying step comprises for each genomic region in the plurality of genomic regions, ranking the genomic region to maximize the ratio between the number of all subjects with the

specific cancer having at least one mutation within the genomic region and the length of the genomic region.

7. The method of claim 1, wherein the library comprises a plurality of genomic regions encoding a plurality of driver sequences.

8. The method of claim 7, wherein the driver sequences are known driver sequences.

9. The method of claim 7, wherein the driver sequences are recurrently mutated in the specific cancer.

10. The method of claim 1, wherein the library comprises a plurality of genomic regions that are recurrently rearranged in the specific cancer.

11. The method of claim 1, wherein the specific cancer is a carcinoma.

12. The method of claim 11, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

13. The method of claim 1, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

14. A method for analyzing a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a tumor nucleic acid sample and a genomic nucleic acid sample from a subject with a specific cancer;

sequencing a plurality of target regions in the tumor nucleic acid sample and in the genomic nucleic acid sample to obtain a plurality of tumor nucleic acid sequences and a plurality of genomic nucleic acid sequences; and

comparing the plurality of tumor nucleic acid sequences to the plurality of genomic nucleic acid sequences to identify a patient-specific genetic alteration in the tumor nucleic acid sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

15. The method of claim 14, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

16. The method of claim 14, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

17. The method of claim 14, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

18. The method of claim 14, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

19. The method of claim 14, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer

having at least one mutation within the genomic region and the length of the genomic region.

20. The method of claim 14, wherein the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences.

21. The method of claim 20, wherein the driver sequences are known driver sequences.

22. The method of claim 20, wherein the driver sequences are recurrently mutated in the specific cancer.

23. The method of claim 14, wherein the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

24. The method of claim 14, wherein the specific cancer is a carcinoma.

25. The method of claim 24, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

26. The method of claim 14, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

27. The method of any one of claims 14-26, further comprising the steps of:

obtaining a cell-free nucleic acid sample from the subject; and
identifying the patient-specific genetic alteration in the cell-free nucleic acid sample.

28. The method of claim 27, wherein the step of identifying the patient-specific genetic alteration in the cell-free nucleic acid sample comprises sequencing a genomic region comprising the patient-specific genetic alteration in the cell-free sample.

29. The method of claim 27, wherein the step of obtaining a tumor nucleic acid sample and a genomic nucleic acid sample comprises the step of enriching the plurality of target regions in the tumor nucleic acid sample and the genomic nucleic acid sample.

30. The method of claim 29, wherein the enriching step comprises use of a custom library of biotinylated DNA.

31. The method of claim 27, wherein the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample.

32. The method of claim 27, further comprising the step of quantifying the cancer-specific genetic alteration in the cell-free sample.

33. A method for screening a cancer-specific genetic alteration in a subject comprising the steps of:

obtaining a cell-free nucleic acid sample from a subject;

sequencing a plurality of target regions in the cell-free sample to obtain a plurality of cell-free nucleic acid sequences; and

identifying a cancer-specific genetic alteration in the cell-free sample;

wherein the plurality of target regions are selected from a plurality of genomic regions that are recurrently mutated in the specific cancer;

the plurality of genomic regions comprises at least 10 different genomic regions; and

at least one mutation within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

34. The method of claim 33, wherein the plurality of genomic regions comprises at least 25, at least 50, at least 100, at least 150, at least 200, or at least 500 different genomic regions.

35. The method of claim 33, wherein at least two mutations within the plurality of genomic regions or at least three mutations within the plurality of genomic regions is present in at least 60% of all subjects with the specific cancer.

36. The method of claim 33, wherein at least one mutation within the plurality of genomic regions is present in at least 60%, 70%, 80%, 90%, 95%, 98%, 99%, or 99.9% of all subjects with the specific cancer.

37. The method of claim 33, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the number of all subjects with the specific cancer having at least one mutation within the genomic region.

38. The method of claim 33, wherein each genomic region in the plurality of genomic regions is identified by ranking the genomic region to maximize the ratio between the number of all subjects with the specific cancer having at least one mutation within the genomic region and the length of the genomic region.

39. The method of claim 33, wherein the plurality of genomic regions comprises genomic regions encoding a plurality of driver sequences.

40. The method of claim 39, wherein the driver sequences are known driver sequences.

41. The method of claim 39, wherein the driver sequences are recurrently mutated in the specific cancer.

42. The method of claim 33, wherein the plurality of genomic regions comprises genomic regions that are recurrently rearranged in the specific cancer.

43. The method of claim 33, wherein the specific cancer is a carcinoma.

44. The method of claim 43, wherein the carcinoma is an adenocarcinoma, a non-small cell lung cancer, or a squamous cell carcinoma.

45. The method of claim 33, wherein the cumulative length of the plurality of genomic regions is at most 30 Mb, 20 Mb, 10 Mb, 5 Mb, 2 Mb, 1 Mb, 500 kb, 200 kb, 100 kb, 50 kb, 20 kb, or 10 kb.

46. The method of claim 33, wherein the step of obtaining a cell-free nucleic acid sample comprises the step of enriching the plurality of target regions in the cell-free nucleic acid sample.

47. The method of claim 46, wherein the enriching step comprises use of a custom library of biotinylated DNA.

Abstract

Methods for creating a library of recurrently mutated genomic regions and for using the library to analyze cancer-specific and patient-specific genetic alterations in a patient are provided. The methods can be used to measure tumor-derived
5 nucleic acids in patient blood and thus to monitor the progression of disease. The methods can also be used for cancer screening.

Electronic Patent Application Fee Transmittal

Application Number:

Filing Date:

Title of Invention:

IDENTIFICATION AND USE OF CIRCULATING TUMOR MARKERS

First Named Inventor/Applicant Name:

Maximilian Diehn

Filer:

David Allen Roise

Attorney Docket Number:

S11-230 PR1

Filed as Small Entity

Provisional Filing Fees

Description

Fee Code

Quantity

Amount

**Sub-Total in
USD(\$)**

Basic Filing:

Provisional Application Filing Fee

2005

1

125

125

Pages:

Claims:

Miscellaneous-Filing:

Petition:

Patent-Appeals-and-Interference:

Post-Allowance-and-Post-Issuance:

Extension-of-Time:

Description	Fee Code	Quantity	Amount	Sub-Total in USD(\$)
Miscellaneous:				
Total in USD (\$)				125

Electronic Acknowledgement Receipt

EFS ID:	15266671
Application Number:	61798925
International Application Number:	
Confirmation Number:	6603
Title of Invention:	IDENITIFICATION AND USE OF CIRCULATING TUMOR MARKERS
First Named Inventor/Applicant Name:	Maximilian Diehn
Customer Number:	86012
Filer:	David Allen Roise
Filer Authorized By:	
Attorney Docket Number:	S11-230 PR1
Receipt Date:	15-MAR-2013
Filing Date:	
Time Stamp:	19:25:41
Application Type:	Provisional

Payment information:

Submitted with Payment	yes
Payment Type	Credit Card
Payment was successfully received in RAM	\$ 125
RAM confirmation Number	24511
Deposit Account	504929
Authorized User	ROISE, DAVID

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

Charge any Additional Fees required under 37 C.F.R. Section 1.16 (National application filing, search, and examination fees)

File Listing:					
Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Drawings-only black and white line drawings	20130315_1417234_TPO_Drawings_as_filed.pdf	4748729	no	19
			e853497f40e1bb019d42ff4bbf12881cd88373c4		
Warnings:					
Information:					
2	Provisional Cover Sheet (SB16)	20130315_1417234_TPO_Provisional_cover_sheet_executed.pdf	2071850	no	3
			002fe4454895bc03ee2507bc4ff9b4241bb9fa08		
Warnings:					
Information:					
3		20130315_1417234_TPO_Specification_Claims_Abstract_as_filed.pdf	8103124	yes	70
			c7f3c5f40ee2183576c47ed73df031a8f2d14045		
	Multipart Description/PDF files in .zip description				
	Document Description		Start	End	
	Specification		1	62	
	Claims		63	69	
	Abstract		70	70	
Warnings:					
Information:					
4	Fee Worksheet (SB06)	fee-info.pdf	29260	no	2
			e6fc963f9c1511f02b5faf64cb0f30fa67b2d72e		
Warnings:					
Information:					
Total Files Size (in bytes):			14952963		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

SCORE Placeholder Sheet for IFW Content

Application Number: 61798925

Document Date: 03/15/2013

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

- Drawings – Other than Black and White Line Drawings

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

To access the documents in the SCORE database, refer to instructions developed by SIRA.

At the time of document entry (noted above):

- Examiners may access SCORE content via the eDAN interface.
- Other USPTO employees can bookmark the current SCORE URL (<http://es/ScoreAccessWeb/>).
- External customers may access SCORE content via the Public and Private PAIR interfaces.