

Petitioner QIAGEN's Presentation

QIAGEN Sciences, LLC v. Tecan Genomics, Inc.

IPR2025-00028; IPR2025-00029

Patents 10,036,012 and 10,876,108



The '012 and '108 Claims



The Independent Claim Of The '012 Patent



US010036012B2

(12) **United States Patent**
Amorese et al.

(10) **Patent No.:** US 10,036,012 B
(45) **Date of Patent:** *Jul. 31, 2017

(54) **COMPOSITIONS AND METHODS FOR TARGETED NUCLEIC ACID SEQUENCE ENRICHMENT AND HIGH EFFICIENCY LIBRARY GENERATION**
(71) Applicant: NuGen Technologies, Inc., San Carlos, CA (US)
(72) Inventors: Doug Amorese, Los Altos, CA (US); Chris Armour, Kirkland, WA (US); Nurith Kurn, Palo Alto, CA (US)
(73) Assignee: NUGEN TECHNOLOGIES, INC., San Carlos, CA (US)
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.
This patent is subject to a terminal disclaimer.

(21) Appl. No.: 15/471,785
(22) Filed: Mar. 28, 2017

(65) **Prior Publication Data**
US 2017/0298345 A1 Oct. 19, 2017

Related U.S. Application Data
(63) Continuation of application No. 13/750,768, filed on Jan. 25, 2013, now Pat. No. 9,650,628.
(60) Provisional application No. 61/591,241, filed on Jan. 26, 2012.

(51) **Int. Cl.**
C12N 15/10 (2006.01)
C12N 15/66 (2006.01)
(52) **U.S. Cl.**
CPC C12N 15/1068 (2013.01); C12N 15/1072 (2013.01); C12N 15/66 (2013.01); C12N 2525/191 (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**
U.S. PATENT DOCUMENTS

4,362,867 A 12/1982 Paddock
4,455,066 A 7/1984 Cornblith et al.
4,469,863 A 9/1984 Tr'o et al.
4,582,877 A 4/1986 Fairchild et al.
4,875,187 A 10/1989 Duck et al.
4,935,357 A 6/1990 Szybalaki
4,942,124 A 7/1990 Church
4,988,617 A 1/1991 Landegren et al.
4,996,143 A 2/1991 Heller et al.
5,011,769 A 4/1991 Duck et al.
5,034,506 A 7/1991 Summerton et al.
5,035,998 A 7/1991 Hartley
5,043,272 A 8/1991 Hartley
5,082,830 A 1/1992 Brakel et al.
5,090,591 A 2/1992 Long
5,130,238 A 7/1992 Malek et al.
5,169,766 A 12/1992 Schuster et al.

FOREIGN PATENT DOCUMENTS
CA 2449256 A1 11/2002
EP 0305627 B1 12/1993
(Continued)

OTHER PUBLICATIONS
Amorese, et al. Improved pathogen sequencing of host-pathogen RNA-sequence. AGBT 2016. Advances in Genome Biology Technology Conference, Feb. 10, 2016, Orlando, FL. NuGen Technologies. Poster Presentation. 2 pages.
Amorese, et al. U.S. Appl. No. 14/211,261, filed Mar. 14, 2014.
Amorese, et al. U.S. Appl. No. 14/778,564, filed Sep. 16, 2015.
Amorese et al., U.S. Appl. No. 14/836,936, filed Aug. 26, 2015.
Amorese, et al. U.S. Appl. No. 14/995,339, filed Jan. 7, 2016.
Armour, Chris, U.S. Appl. No. 13/643,056, filed Oct. 23, 2012.
Armour, Chris, U.S. Appl. No. 14/920,254, filed Oct. 22, 2015.
Armour, Chris, U.S. Appl. No. 14/995,382, filed Jan. 14, 2016.
Armour, et al., U.S. Appl. No. 14/390,012, filed Oct. 1, 2014.
Huang et al., The behavior of 5-hydroxymethylcytosine in brain sequencing. PLOS one, Jan. 2015, 5(1):e8888. 9 Pages.
Hodges, et al., U.S. Appl. No. 15/047,448, filed Feb. 18, 2016.
Jing, et al., Spectra of BRCA1 and BRCA2 mutations in Korean patients with breast cancer: the importance of whole-genome sequencing. Journal of Human Genetics. vol. 57, Pages 212-215, Jan. 2012.

ABSTRACT
The present invention provides methods, compositions and kits for targeted nucleic acid sequence enrichment in nucleic acid sample and for high efficiency nucleic acid library generation for next generation sequencing (NGS). Specifically, the methods, compositions and kits provided herein are useful for the production and capture of amplification-ready, target-specific and strand-specific regions of interest from nucleic acid samples containing complex DNA.

22 Claims, 4 Drawing Sheets

1. A method for sequencing an enriched nucleic acid sequence of interest, the method comprising:

a) obtaining a nucleic acid fragment ligated to a partial duplex adaptor, wherein the nucleic acid fragment comprises a nucleic acid sequence of interest, wherein the partial duplex adaptor comprises a first adaptor sequence, and wherein the partial duplex adaptor comprises a first strand and a second strand, wherein the first strand is longer than the second strand;

b) annealing one or more oligonucleotides in solution to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the partial duplex adaptor, wherein the one or more oligonucleotides comprise a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;

c) extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment ligated to the partial duplex adaptor with a polymerase, thereby generating one or more oligonucleotide extension products comprising sequence complementary to the first adaptor sequence at a first end, sequence complementary to the nucleic acid sequence of interest, and the second adaptor sequence at a second end;

d) amplifying the one or more oligonucleotide extension products using a first primer that anneals to a complement of the first adaptor sequence and a second primer that anneals at its 3' end to a complement of the second adaptor sequence to enrich for the nucleic acid sequence of interest, wherein products of the amplifying comprise a 3' end with sequence complementary to a sequence on a surface;

e) annealing a strand of the products of the amplifying to the sequence on the surface using the 3' end with sequence complementary to the sequence on the surface; and

f) sequencing the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.

The Independent Claim Of The '108 Patent

 US 10,876,108 B2	
(12) United States Patent Amorese et al.	(10) Patent No.: US 10,876,108 B2 (45) Date of Patent: *Dec. 29, 2020
(54) COMPOSITIONS AND METHODS FOR TARGETED NUCLEIC ACID SEQUENCE ENRICHMENT AND HIGH EFFICIENCY LIBRARY GENERATION	5,034,506 A 7/1991 Sumner et al. 5,035,996 A 7/1991 Hartley 5,045,272 A 8/1991 Hartley 5,082,830 A 1/1992 Benkel et al. 5,090,591 A 2/1992 Long 5,130,338 A 7/1992 Malek et al. 5,169,766 A 12/1992 Schuster et al. 5,171,534 A 12/1992 Smith et al. 5,194,370 A 3/1993 Berninger et al. 5,216,141 A 6/1993 Boomer 5,234,809 A 8/1993 Sumner et al. 5,235,033 A 8/1993 Sumner et al. 5,242,794 A 9/1993 Whalley et al. 5,386,023 A 1/1995 Sanghvi et al. 5,390,491 A 1/1995 Kacian et al. 5,409,818 A 4/1995 Dreyer et al. 5,418,149 A 5/1995 Gelfand et al. 5,427,929 A 6/1995 Richards et al. 5,480,784 A 1/1996 Kacian et al. 5,494,810 A 2/1996 Bittany et al. 5,508,169 A 4/1996 Dugan et al. 5,508,176 A 4/1996 Rose et al. 5,510,270 A 4/1996 Fodor et al. 5,525,471 A 6/1996 Zeng 5,545,522 A 8/1996 Van et al. 5,554,516 A 9/1996 Kacian et al. 5,554,517 A 9/1996 Dreyer et al. 5,556,752 A 9/1996 Lockhart et al. (Continued)
(71) Applicant: Nugen Technologies, Inc. , San Carlos, CA (US)	
(72) Inventors: Doug Amorese , Los Altos, CA (US); Chris Amour , Kirkland, WA (US); Nurith Kurn , Palo Alto, CA (US)	
(73) Assignee: NUGEN TECHNOLOGIES, INC. , Redwood City, CA (US)	
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 203 days. This patent is subject to a terminal disclaimer.	
(21) Appl. No.: 16/017,340	
(22) Filed: Jun. 25, 2018	
(65) Prior Publication Data US 2019/0078082 A1 Mar. 14, 2019	FOREIGN PATENT DOCUMENTS CA 2444926 A1 11/2002 EP 0365627 B1 12/1993 (Continued)
Related U.S. Application Data	OTHER PUBLICATIONS Rothberg et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011; 475: 348-352 + Supplementary Information. (Year: 2011). Huang et al. The behavior of 5-hydroxymethylcytosine in Bisulfite sequencing. PLoS one. Jan. 2010; 5(1):e8008. 9 Pages. International search report and written opinion dated Feb. 5, 2016 for PCT/US2015/047053. Jang, et al. Spectra of BRCA1 and BRCA2 mutations in Korean patients with breast cancer: the importance of whole-genome sequencing. Journal of Human Genetics. vol. 57, pp. 212-215, Jan. 5, 2012. (Continued)
(63) Continuation of application No. 15/471,785, filed on Mar. 28, 2017, now Pat. No. 10,036,012, which is a continuation of application No. 13/750,768, filed on Jan. 25, 2013, now Pat. No. 9,650,628.	
(60) Provisional application No. 61/591,241, filed on Jan. 26, 2012.	
(51) Int. Cl. C12N 15/10 (2006.01) C12N 15/66 (2006.01)	
(52) U.S. Cl. C12N 15/1068 (2013.01); C12N 15/1072 (2013.01); C12N 15/66 (2013.01); C12Q 25/35 (2013.01); C12Q 1/6844 (2013.01); C12Q 1/6855 (2013.01); C12Q 1/6844 (2013.01); C12Q 1/6855 (2013.01); C40B 30/04 (2006.01)	
(58) Field of Classification Search CPC C12N 15/1068; C12N 15/66; C12N 15/1072; C12Q 25/35; C12Q 1/6844; C12Q 1/6855; C40B 30/04; C40B 50/06 See application file for complete search history.	Primary Examiner —Angelo M. Bertagna (74) Attorney, Agent, or Firm —Brown Rudnick LLP; Thomas C. Meyers
(56) References Cited U.S. PATENT DOCUMENTS 4,362,867 A 12/1982 Paddock 4,458,066 A 7/1984 Cantor et al. 4,469,863 A 9/1984 Ts'o et al. 4,582,877 A 4/1986 Fairclark et al. 4,876,187 A 10/1989 Duck et al. 4,935,157 A 6/1990 Szabadai 4,942,124 A 7/1990 Church et al. 4,988,617 A 1/1991 Landegren et al. 4,996,141 A 2/1991 Heller et al. 5,011,769 A 4/1991 Duck et al.	ABSTRACT The present invention provides methods, compositions and kits for targeted nucleic acid sequence enrichment in a nucleic acid sample and for high efficiency nucleic acid library generation for next generation sequencing (NGS). Specifically, the methods, compositions and kits provided herein are useful for the production and capture of amplification-ready, target-specific and strand-specific regions of interest from nucleic acid samples containing complex DNA.
19 Claims, 4 Drawing Sheets	

1. A method for sequencing an enriched nucleic acid sequence of interest, the method comprising:

a) annealing one or more oligonucleotides in solution in a reaction mixture to the nucleic acid sequence of interest in a nucleic acid fragment, wherein the reaction mixture comprises a plurality of nucleic acid fragments, wherein the nucleic acid fragment comprising the nucleic acid sequence of interest comprises a first adaptor sequence, wherein the one or more oligonucleotides comprise a 3' portion with at least 10 bases designed to be complementary to the nucleic acid sequence of interest and a 5' tail portion comprising a second adaptor sequence that is non-complementary to the nucleic acid sequence of interest;

b) extending the one or more oligonucleotides annealed to the nucleic acid sequence of interest in the nucleic acid fragment comprising the first adaptor sequence with a polymerase, in the reaction mixture, thereby generating one or more oligonucleotide extension products comprising sequence complementary to the first adaptor sequence at a first end, sequence complementary to the nucleic acid sequence of interest, and the second adaptor sequence at a second end;

c) amplifying the one or more oligonucleotide extension products, in the reaction mixture, using a first primer that anneals to the complement of the first adaptor sequence and a second primer that anneals at its 3' end to a complement of the second adaptor sequence, thereby enriching the nucleic acid sequence of interest, by generating amplified products comprising the enriched nucleic acid sequence of interest; and
d) sequencing the amplified products comprising the enriched nucleic acid sequence of interest on a massively parallel sequencing platform.

Motivation to Combine



PO Statements About Motivation To Combine

Tecan's Sur-Reply

for the reasons explained be [Petitioner] assumes there is a motivation to combine where none exists. Petitioner cannot demonstrate that a POSA would have been

Tecan's Sur-Reply

EX1011.0016, [0080]. Not once does Shapero suggest that MARA can or should be used with any other instrument than its arrays. Accordingly, Petitioner's list of out-

What Is Shapero's (Ex. 1006) MARA Method?

MARA: a novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays

Michael H. Shapero*, Jane Zhang, Ann Loraine, Weiwei Liu, Xiaojun Di, Guoying Liu and Keith W. Jones

Affymetrix, Inc., Genotyping Research, 3380 Central Expressway, Santa Clara, CA 95051, USA

Received October 8, 2004; Revised November 10, 2004; Accepted November 24, 2004

ABSTRACT

We have developed a locus-specific DNA target preparation method for highly multiplexed single nucleotide polymorphism (SNP) genotyping called MARA (Multiplexed Anchored Runoff Amplification). The approach uses a single primer per SNP in conjunction with restriction enzyme digested, adapter-ligated human genomic DNA. Each primer is composed of common sequence at the 5' end followed by locus-specific sequence at the 3' end. Following a primary reaction in which locus-specific products are generated, a secondary universal amplification is carried out using a generic primer pair corresponding to the oligonucleotide and genomic DNA adapter sequences. Allele discrimination is achieved by hybridization to high-density DNA oligonucleotide arrays. Initial multiplex reactions containing either 250 primers or 750 primers across nine DNA samples demonstrated an average sample call rate of ~95% for 250- and 750-plex MARA. We have also evaluated >1000- and 4000-primer plex MARA to genotype SNPs from human chromosome 21. We have identified a subset of SNPs corresponding to a primer conversion rate of ~75%, which show an average call rate over 95% and concordance >99% across seven DNA samples. Thus, MARA may potentially improve the throughput of SNP genotyping when coupled with allele discrimination on high-density arrays by allowing levels of multiplexing during target generation that far exceed the capacity of traditional multiplex PCR.

INTRODUCTION

Long before the completion of the human genome sequence, it was clear that sites of genetic variation could be used as markers to identify disease segregation patterns among families (1). This approach led to the identification of a number of genes involved in rare, monogenic disorders. With the complete sequence now available (2,3), attention has turned to the challenge of identifying common disease genes. Although

multiple so single nucleotide marker of abundance the ongoing will identify expected t studies des complex d drug respon there is sti cost-effect types in la There a SNPs that approaches e.g. the us cleotide pr as DNA p approaches and labeli rently use there is a biding the significant be genoty of the target sequences and improving the ability to differentiate correct and incorrect genotype calls. Traditionally, standard multiplex PCR is difficult to design and enable at levels beyond several dozen loci per reaction (8). Even with careful optimization of input reagents (9) and efforts to minimize non-specific primer dimers (10), multiplex PCR is not a routine laboratory technique. A number of alternative approaches to standard multiplex PCR have been described over the years. These include the use of universal primer sequences added to the ends of locus-specific sequences (11), the use of PCR suppression in which one target-specific primer is used in conjunction with a primer common to adapters present on all DNA fragments (12,13) and the use of acrylamide spheres with covalently attached primers for solid-phase PCR amplification (14). While these methods do improve multiplexing levels as compared with standard approaches, none of the described methods has demonstrated reactions >100-plex and thus are still of moderate utility for a high-throughput

ABSTRACT

We have developed a locus-specific DNA target preparation method for highly multiplexed single nucleotide polymorphism (SNP) genotyping called MARA (Multiplexed Anchored Runoff Amplification). The approach uses a single primer per SNP in

*To whom correspondence should be addressed. Tel: +1 408 731 5171; Fax: +1 408 481 0422; Email: michael_shapero@affymetrix.com

MARA: a novel approach for locus-specific SNP genotyping using high-density DNA oligonucleotide arrays

Michael H. Shapero*, Jane Zhang, Ann Loraine, Guoying Liu and Keith W. Jones

Altamira, Inc., Genotyping Research, 3380 Central Expressway

Received October 8, 2004; Revised November 10, 2004; Accepted November 10, 2004

ABSTRACT

We have developed a locus-specific DNA target preparation method for highly multiplexed single nucleotide polymorphism (SNP) genotyping called MARA (Multiplexed Anchored Runoff Amplification). The approach uses a single primer per SNP in conjunction with restriction enzyme digested, adapter-ligated human genomic DNA. Each primer is composed of common sequence at the 5' end followed by locus-specific sequence at the 3' end. Following a primary reaction in which locus-specific products are generated, a secondary universal amplification is carried out using a generic primer pair corresponding to the oligonucleotide and genomic DNA adapter sequences. Allele discrimination is achieved by hybridization to high-density DNA oligonucleotide arrays. Initial multiplex reactions containing either 250 primers or 750 primers across nine DNA samples demonstrated an average sample call rate of ~95% for 250- and 750-plex MARA. We have also evaluated >1000- and 4000-primer plex MARA to genotype SNPs from human chromosome 21. We have identified a subset of SNPs corresponding to a primer conversion rate of ~75%, which show an average call rate over 95% and concordance >99% across seven DNA samples. Thus, MARA may potentially improve the throughput of SNP genotyping when coupled with allele discrimination on high-density arrays by allowing levels of multiplexing during target generation that far exceed the capacity of traditional multiplex PCR.

INTRODUCTION

Long before the completion of the human genome sequence, it was clear that sites of genetic variation could be used as markers to identify disease segregation patterns among families (1). This approach led to the identification of a number of genes involved in rare, monogenic disorders. With the complete sequence now available (2,3), attention has turned to the challenge of identifying common disease genes. Although

*To whom correspondence should be addressed. Tel: +1 408 731 5171; Fax: +1 408 731 5172; Email: shapero@altamira.com
Nucleic Acids Research, Vol. 32 No. 22 © Oxford University Press

There are a multitude of genotyping methodologies for SNPs that have been described over the years (5–7). These approaches can vary in terms of how alleles are discriminated, e.g. the use of nucleic acid hybridization with short oligonucleotide probes alone or in combination with enzymes such as DNA polymerases or DNA ligases. Additionally, these approaches can vary in the nature of both the assay formats and labeling and detection strategies. However, many currently used approaches share a common feature in that there is a requirement for an amplification of the region harboring the polymorphism prior to actual genotyping. PCR significantly increases the copy number of the sequence to be genotyped, thereby effectively reducing the complexity of the target sequences and improving the ability to differentiate correct and incorrect genotype calls. Traditionally, standard multiplex PCR is difficult to design and enable at levels beyond several dozen loci per reaction (8). Even with careful optimization of input reagents (9) and efforts to minimize non-specific primer dimers (10), multiplex PCR is not a routine laboratory technique. A number of alternative approaches to standard multiplex PCR have been described over the years. These include the use of universal primer sequences added to the ends of locus-specific sequences (11), the use of PCR suppression in which one target-specific primer is used in conjunction with a primer common to adapters present on all DNA fragments (12,13) and the use of acrylamide spheres

(16), it is limited in practice by the requirement for large quantities (25 µg) of input genomic DNA and, therefore, is not well suited for high-throughput analysis. Thus, one of the continuing challenges in SNP genotyping is the development of molecular methods capable of amplifying locus-specific SNPs in a highly multiplexed manner.

evenly spaced SNPs are used (4,17,18). In order to leverage the power of SNPs in genetic studies, novel target preparation methods need to be developed that, when combined with methods for allele discrimination, allow rapid and accurate genotypic information at a fraction of the cost of current approaches. High-density arrays provide a method for allele discrimination and have been used successfully for SNP dis-

EX1006 at 2 (left. col.)

EX1006 at 1 (right col.)

Shapero's Conclusion

MARA: a novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays

Michael H. Shapero*, Jane Zhang, Ann Loraine, Weiwei Liu, Xiaojun Di, Guoying Liu and Keith W. Jones

Affymetrix, Inc., Genotyping Research, 3380 Central Expressway, Santa Clara, CA 95051, USA

Received October 8, 2004; Revised November 10, 2004; Accepted November 24, 2004

ABSTRACT

We have developed a locus-specific DNA preparation method for highly multiplexed nucleotide polymorphism (SNP) genotyping. MARA (Multiplexed Anchored Runoff Amplification) uses a single primer per locus in conjunction with restriction enzyme digestion of adapter-ligated human genomic DNA. Each locus is composed of common sequence at the 5' end followed by locus-specific sequence at the 3' end. Following a primary reaction in which locus-specific products are generated, a secondary universal amplification is carried out using a generic primer corresponding to the oligonucleotide and DNA adapter sequences. Allele discrimination is achieved by hybridization to high-density DNA oligonucleotide arrays. Initial multiplex reactions using either 250 primers or 750 primers across samples demonstrated an average sample coverage of ~95% for 250- and 750-plex MARA. We have evaluated >1000- and 4000-primer plex MARA to SNPs from human chromosome 21. We have identified a subset of SNPs corresponding to a polymorphism rate of ~75%, which show an average coverage over 95% and concordance >99% across samples. Thus, MARA may potentially improve the throughput of SNP genotyping when coupled with allele discrimination on high-density arrays, allowing levels of multiplexing during target genotyping that far exceed the capacity of traditional multiplexing.

INTRODUCTION

Long before the completion of the human genome sequence, it was clear that sites of genetic variation could be used as markers to identify disease segregation patterns among families (1). This approach led to the identification of a number of genes involved in rare, monogenic disorders. With the complete sequence now available (2,3), attention has turned to the challenge of identifying common disease genes. Although

DNA samples. These results suggest that MARA may serve as a target preparation method for highly multiplexed locus-specific SNP genotyping that far exceeds the capacity of standard traditional multiplex PCR. When coupled with a method for allelic discrimination such as high-density array hybridization, this method provides access to large numbers of locus-specific SNPs in a single reaction and thus may facilitate the use of SNPs in genome-wide association studies.

These include the use of universal primer sequences added to the ends of locus-specific sequences (11), the use of PCR suppression in which one target-specific primer is used in conjunction with a primer common to adapters present on all DNA fragments (12,13) and the use of acrylamide spheres with covalently attached primers for solid-phase PCR amplification (14). While these methods do improve multiplexing levels as compared with standard approaches, none of the described methods has demonstrated reactions >100-plex and thus are still of moderate utility for a high-throughput

*To whom correspondence should be addressed. Tel: +1 408 731 5171; Fax: +1 408 481 0422; Email: michael_shapero@affymetrix.com

Jones (Ex. 1011), MARA Description



US 20050142577A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2005/0142577 A1**
Jones et al. (43) **Pub. Date: Jun. 30, 2005**

(54) **METHODS FOR GENOTYPING SELECTED
POLYMORPHISM**

(60) Provisional application No. 60/493,085, filed on Aug.
5, 2003.

(75) Inventors: **Keith W. Jones**, Sunnyvale, CA (US);
Michael H. Shapero, Redwood City,
CA (US); **Stephen P.A. Foder**, Palo
Alto, CA (US)

Publication Classification

(51) **Int. Cl.⁷** **C12Q 1/68; C12P 19/34**
(52) **U.S. Cl.** **435/6; 435/91.2**

Correspondence Address:
AFFYMETRIX, INC
ATTN: CHIEF IP COUNSEL, LEGAL DEPT.
3380 CENTRAL EXPRESSWAY
SANTA CLARA, CA 95051 (US)

(57) **ABSTRACT**

Methods for genotyping polymorphisms using a locus specific primer that is complementary to a region near a selected polymorphism are described. Methods for synthesizing pools of locus specific primers that incorporate some degenerate positions are also disclosed. A plurality of different sequence capture probes are synthesized simultaneously using degenerate oligonucleotide synthesis. The sequence of the locus specific regions of the capture probes are related in that they have some bases that are identical in each sequence in the plurality of sequences and positions that vary from one locus specific region to another. The sequences are selected based on proximity to a polymorphism of interest and because they conform to a similar sequence pattern.

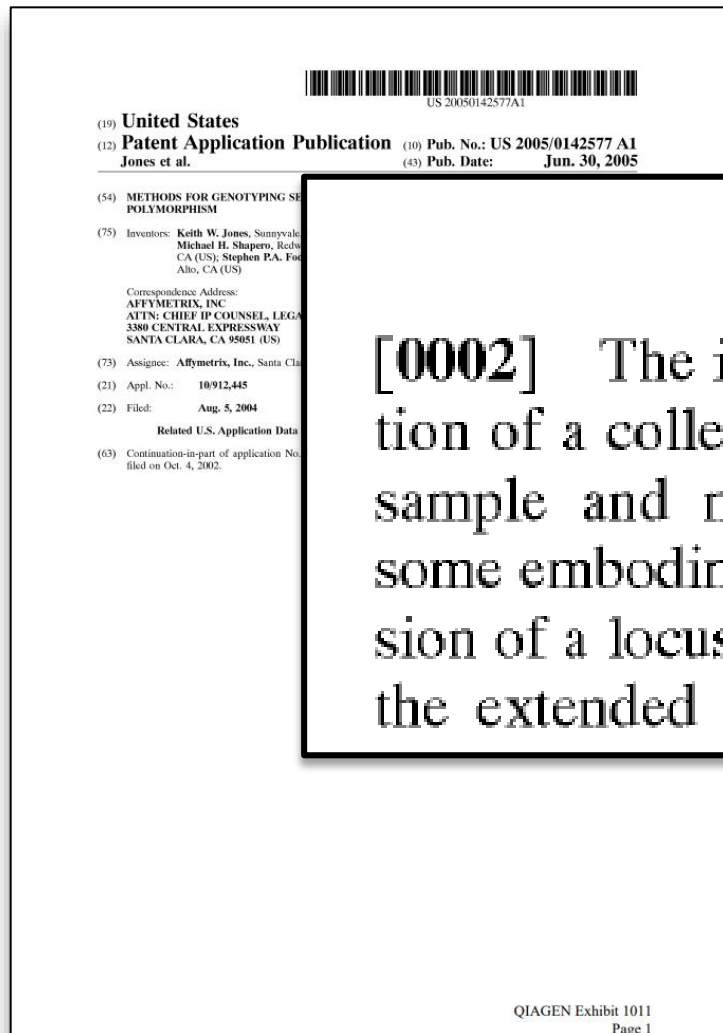
(73) Assignee: **Affymetrix, Inc.**, Santa Clara, CA

(21) Appl. No.: **10/912,445**

(22) Filed: **Aug. 5, 2004**

Related U.S. Application Data

(63) Continuation-in-part of application No. 10/264,945,
filed on Oct. 4, 2002.



FIELD OF THE INVENTION

[0002] The invention relates to enrichment and amplification of a collection of target sequences from a nucleic acid sample and methods of analyzing amplified product. In some embodiments target sequences are amplified by extension of a locus-specific primer followed by amplification of the extended locus-specific primer with a generic pair of

Jones (Ex. 1011) Continued

(19) United States
(12) Patent Application
Jones et al.

(54) METHODS FOR GENOTYPING
POLYMORPHISM

(75) Inventors: Keith W. Jones
Michael H. Smith
CA (US); Stephen
Alto, CA (US)

Correspondence Address:
AFFYMETRIX, INC.
ATTN: CHIEF IP COUNSEL
3380 CENTRAL EXPRESSWAY
SANTA CLARA, CA 95051

(73) Assignee: Affymetrix, Inc.

(21) Appl. No.: 10/912,444

(22) Filed: Aug. 5, 2002

Related U.S. App.

(63) Continuation-in-part of application
filed on Oct. 4, 2002.

[0075] (C) Multiplexed Locus Specific Genotyping

[0076] Generally, the invention provides methods for highly multiplexed locus specific amplification of nucleic acids and methods for analysis of the amplified products. In preferred embodiments the amplified targets prepared using Multiplexed Anchored Runoff Amplification (MARA) are analyzed to determine the genotype of SNPs. In some embodiments the invention combines the use of capture probes that comprise a common sequence and a locus-specific region with adaptor-modified sample nucleic acid; the adaptor comprises a second common sequence. The capture probes are extended to produce copies of the sample DNA that contain common priming sequences flanking the target sequence. The copies are amplified with a generic set of primers that recognize the common sequences. The amplified product may be analyzed by hybridization to an array of probes.

QIAGEN Exhibit 1011
Page 1

[0080] A second embodiment of the methods is illustrated in FIG. 2. Capture probes are designed with a locus specific region (LS1 or LS2) and a common sequence (A1) as in FIG. 1. In this embodiment the capture probes further comprise a tag sequence that is unique for each species of capture probe designed. (For a description of tags and tag probes, see, U.S. Ser. No. 08/626,285.) The capture probes are attached to the array through hybridization of the tag sequence to a substantially complementary tag probe sequence that is attached to the array. The tag probes may be attached to the array in discrete locations. Different species of tag probes are present at different discrete, spatially addressable locations. Adaptor-ligated genomic DNA is hybridized to the array so that the capture probes hybridize to target sequences in the sample. The capture probes are extended as in FIG. 1 to incorporate the target sequence and common sequence A2. The extended capture probes are released and amplified using primers A1 and A2. The amplified product may then be analyzed by, for example, hybridization to an array. Information about the region of interest can be determined by analysis of the hybridization pattern. The amplified sample may be analyzed by any method known in the art, for example, MALDI-TOF mass spec, capillary electrophoresis, OLA, dynamic allele specific hybridization (DASH) or TaqMan® (Applied Biosystems, Foster City, Calif.). For other methods of genotyping analyses see Syvanen, *Nature Rev. Gen.* 2:930-942 (2001) which is herein incorporated by reference in its entirety.

Goals of Shapero & Patents: To Prepare Enriched Libraries For Analysis



Shapero (Ex. 1006) *cf* '012 Patent (Ex. 1001)

Shapero

and labeling and detection strategies. However, many currently used approaches share a common feature in that there is a requirement for an amplification of the region harboring the polymorphism prior to actual genotyping. PCR significantly increases the copy number of the sequence to be genotyped, thereby effectively reducing the complexity of the target sequences and improving the ability to differentiate correct and incorrect genotype calls. Traditionally, standard multiplex PCR is difficult to design and enable at levels beyond several dozen loci per reaction (8). Even with careful optimization of input reagents (9) and efforts to minimize non-specific primer dimers (10), multiplex PCR is not a routine laboratory technique. A number of alternative approaches to

Ex. 1006.6-7

'012 Patent

Although the cost of whole genome sequencing is decreasing and the throughput of the NGS platforms is increasing, it is nonetheless often more practical and cost-effective to select genomic regions of interest for sequencing and analysis. Target enrichment is a commonly employed strategy in genomic DNA sequencing in which genomic regions of interest are selectively captured from a DNA sample before sequencing. Focused target enrichment is an important tool especially in the fields of study where sequencing of a large number of samples is necessary (e.g. population-based studies of disease markers or SNPs), making whole genome sequencing cost-prohibitive. Similarly, improvements have been made that enable DNA libraries to be made from nucleic acid from fewer number of cells, but these are bound by the limitations of the efficiency of ligation reactions.

Several approaches to target enrichment have been developed which vary from one another in terms of sensitivity, specificity, reproducibility, uniformity, cost and ease of use. The target enrichment methods commonly employed today can be divided into three major categories, each with its distinct advantages and disadvantages: 1) PCR-based methods; 2) capture-by-hybridization, i.e. on-array or in-solution hybrid capture; and 3) capture-by-circularization, i.e. molecular inversion probe-based methods.

The PCR-based methods employ highly parallel PCR amplification, where each target sequence in the sample has a corresponding pair of unique, sequence-specific primers. The simultaneous use of numerous primer pairs makes multiplex PCR impractical due to high level of non-specific amplification and primer-primer interactions. Recently developed microdroplet PCR technology (Tewhey et al., 2009) in which each amplification reaction is physically

Ex. 1001 Col.1:23-Col.2:30

Shapero (Ex. 1006) *cf* '012 Patent (Ex. 1001)

Shapero

(16), it is limited in practice by the requirement for large quantities (25 µg) of input genomic DNA and, therefore, is not well suited for high-throughput analysis. Thus, one of the continuing challenges in SNP genotyping is the development of molecular methods capable of amplifying locus-specific SNPs in a highly multiplexed manner.

Ex. 1006 at 2 of 9 Left Col.

'012 Patent

There is a need for improved methods for selective target enrichment that allow for low-cost, high throughput capture of genomic regions of interest without specialized instrumentation. Additionally, there is also a need for high efficiency nucleic acid library generation. The methods of the invention described herein fulfill these needs.

Ex.1001, Col.2:25-30

'012 Patent

General

The methods of the invention can be used for the selective enrichment of a plurality of defined target sequences from complex DNA with a set of common primers and adaptors, thus circumventing the need for multiplex PCR and multiple primer pairs. A multiplicity of target regions of interest are envisioned: for example, the regions of interest can represent all known coding regions, the entire exome, selected regions

Ex.1001, Col.2:25-30

Types Of Analysis Patents Identify



The '012 Patent (Ex. 1001) - Enrichment Uses

(12) **United States Patent**
Amorese et al.

(54) **COMPOSITIONS AND METHODS FOR TARGETED NUCLEIC ACID SEQUENCE ENRICHMENT AND HIGH EFFICIENCY LIBRARY GENERATION**

(71) Applicant: **NuGEN Technologies, Inc.**, San Carlos, CA (US)

(72) Inventors: **Doug Amorese**, Los Altos, CA (US);
Chris Armour, Kirkland, WA (US);
Nurith Kurn, Palo Alto, CA (US)

(73) Assignee: **NUGEN TECHNOLOGIES, INC.**, San Carlos, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.
This patent is subject to a terminal disclaimer.

(21) Appl. No.: **15/471,785**

(22) Filed: **Mar. 28, 2017**

(65) **Prior Publication Data**
US 2017/0298345 A1 Oct. 19, 2017

Related U.S. Application Data
(63) Continuation of application No. 13/750,768, filed on Jan. 25, 2013, now Pat. No. 9,650,628.
(60) Provisional application No. 61/591,241, filed on Jan. 26, 2012.

(51) **Int. Cl.**
C12N 15/10 (2006.01)
C12N 15/66 (2006.01)

(52) **U.S. Cl.**
CPC **C12N 15/1068** (2013.01); **C12N 15/1072** (2013.01); **C12N 15/66** (2013.01); **C12Q 2525/191** (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**
U.S. PATENT DOCUMENTS

4,362,867 A	12/1982	Paddock
4,458,666 A	7/1984	Caruthers et al.
4,469,863 A	9/1984	Tr'o et al.
4,582,877 A	4/1986	Fairchuk et al.
4,875,187 A	10/1989	Duck et al.
4,935,357 A	6/1990	Szybalaki
4,942,124 A	7/1990	Church
4,988,617 A	1/1991	Landegren et al.
4,996,143 A	2/1991	Heller et al.
5,011,769 A	4/1991	Duck et al.
5,034,506 A	7/1991	Summerton et al.
5,035,996 A	7/1991	Hartley
5,043,272 A	8/1991	Hartley
5,082,830 A	1/1992	Brakel et al.
5,090,591 A	2/1992	Long
5,130,238 A	7/1992	Malek et al.
5,169,766 A	12/1992	Schuster et al.

Methods of Amplification

The methods, compositions and kits described herein can be useful to generate amplification-ready products for downstream applications such as massively parallel sequencing (i.e. next generation sequencing methods), generation of libraries with enriched population of sequence regions of interest, or hybridization platforms. Methods of amplification are well known in the art. Suitable amplification reactions can be exponential or isothermal and can include any DNA amplification reaction, including but not limited to

interest from nucleic acid samples containing complex DNA.

22 Claims, 4 Drawing Sheets

The '012 Patent (Ex. 1001)



(12) United States Patent Amorese et al.		(10) Patent No.: (45) Date of Patent:
(54) COMPOSITIONS AND METHODS FOR TARGETED NUCLEIC ACID SEQUENCE ENRICHMENT AND HIGH EFFICIENCY LIBRARY GENERATION	5,171,534 A 5,194,570 A 5,216,141 A 5,234,809 A 5,235,033 A 5,242,794 A 5,380,023 A 5,399,491 A 5,409,818 A 5,418,149 A 5,427,929 A 5,480,784 A 5,494,810 A 5,508,169 A 5,508,178 A 5,510,270 A 5,525,471 A 5,545,522 A 5,554,516 A 5,554,517 A 5,556,752 A 5,565,340 A 5,573,913 A 5,578,832 A 5,589,339 A	12/1992 S 3/1993 B 6/1993 B 8/1993 B 8/1993 S 9/1993 W 1/1995 S 3/1995 K 4/1995 D 5/1995 C 6/1995 R 1/1996 R 2/1996 B 4/1996 D 4/1996 R 4/1996 F 6/1996 Z 8/1996 V 9/1996 K 9/1996 D 9/1996 L 10/1996 C 11/1996 R 12/1996 B
(71) Applicant: NuGen Technologies, Inc., San Carlos, CA (US)		
(72) Inventors: Doug Amorese, Los Altos, CA (US); Chris Armour, Kirkland, WA (US); Nurith Kurn, Palo Alto, CA (US)		
(73) Assignee: NUGEN TECHNOLOGIES, INC., San Carlos, CA (US)		
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days. This patent is subject to a terminal disclaimer.		
(21) Appl. No.: 15/471,785		(Continued)
(22) Filed: Mar. 28, 2017		FOREIGN PATENT
(65) Prior Publication Data US 2017/0298345 A1 Oct. 19, 2017	CA EP	2444926 A1 0305627 B1 (Continued)
Related U.S. Application Data		OTHER PUBLICATIONS
(63) Continuation of application No. 13/750,768, filed on Jan. 25, 2013, now Pat. No. 9,650,628.		Amorese, et al. Improved pathologic RNA-sequence. AGBT 2016. Adv. Biology Conference, Feb. 10, 2016. gies. Poster Presentation. 2 pages.
(60) Provisional application No. 61/591,241, filed on Jan. 26, 2012.		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
(51) Int. Cl. C12N 15/10 (2006.01) C12N 15/66 (2006.01)		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
(52) U.S. CL. C12N 15/1068 (2013.01); C12N 15/1072 (2013.01); C12N 15/66 (2013.01); C12N 2525/191 (2013.01)		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
(58) Field of Classification Search None See application file for complete search history.		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
(56) References Cited		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
U.S. PATENT DOCUMENTS		Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,362,867 A	12/1982 Paddock	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,455,066 A	7/1984 Cornblith et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,469,863 A	9/1984 Ts'o et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,582,877 A	4/1986 Fairchild et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,875,187 A	10/1989 Duck et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,935,357 A	6/1990 Szybalaki	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,942,121 A	7/1990 Church	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,988,617 A	1/1991 Landegren et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
4,996,143 A	2/1991 Heller et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,011,769 A	4/1991 Duck et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,034,506 A	7/1991 Summerton et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,035,996 A	7/1991 Hartley	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,043,272 A	8/1991 Hartley	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,082,830 A	1/1992 Binkal et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,090,591 A	2/1992 Long	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,130,238 A	7/1992 Malek et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.
5,169,766 A	12/1992 Schuster et al.	Amorese, et al. U.S. Appl. No. 14/000,000, filed Jan. 26, 2013.

In some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal.

Downstream Applications

An important aspect of the invention is that the methods and compositions disclosed herein can be efficiently and cost-effectively utilized for downstream analyses, such as next generation sequencing or hybridization platforms, with minimal loss of biological material of interest. The methods of the present invention can also be used in the analysis of genetic information of selective genomic regions of interest (e.g., analysis of SNPs or other disease markers) as well as genomic regions which may interact with the selective Sequencing

The '012 Patent (Ex. 1001)



(12) United States Patent Amorese et al.		(10) Patent No.: U.S. Pat. No. 8,110,709
		(45) Date of Patent: 12/19/2012
(54) COMPOSITIONS AND METHODS FOR TARGETED NUCLEIC ACID SEQUENCE ENRICHMENT AND HIGH EFFICIENCY LIBRARY GENERATION	5,171,534 A	12/19/2012
(71) Applicant: NuGEN Technologies, Inc., San Carlos, CA (US)	5,194,370 A	3/19/2013
(72) Inventors: Doug Amorese, Los Altos, CA (US); Chris Armour, Kirkland, WA (US); Norith Kurn, Palo Alto, CA (US)	5,216,141 A	6/19/2013
(73) Assignee: NUGEN TECHNOLOGIES, INC., San Carlos, CA (US)	5,234,809 A	8/19/2013
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.	5,235,033 A	8/19/2013
This patent is subject to a terminal disclaimer.	5,242,794 A	9/19/2013
(21) Appl. No.: 15/471,785	5,360,023 A	1/19/2015
(22) Filed: Mar. 28, 2017	5,399,491 A	3/19/2015
(65) Prior Publication Data	5,409,818 A	4/19/2015
US 2017/029845 A1 Oct. 19, 2017	5,418,149 A	5/19/2015
Related U.S. Application Data	5,427,929 A	6/19/2015
(63) Continuation of application No. 13/750,768, filed on Jan. 25, 2013, now Pat. No. 9,650,628.	5,494,810 A	2/19/2016
(60) Provisional application No. 61/591,241, filed on Jan. 26, 2012.	5,508,169 A	4/19/2016
(51) Int. Cl.	5,508,178 A	4/19/2016
C12N 15/10 (2006.01)	5,510,270 A	4/19/2016
C12N 15/66 (2006.01)	5,525,471 A	6/19/2016
(52) U.S. CL.	5,545,522 A	8/19/2016
CPC C12N 15/1068 (2013.01); C12N 15/1072 (2013.01); C12N 15/66 (2013.01); C12O 2525/191 (2013.01)	5,554,516 A	9/19/2016
(58) Field of Classification Search	5,554,517 A	9/19/2016
None	5,556,732 A	9/19/2016
See application file for complete search history.	5,565,340 A	10/19/2016
(56) References Cited	5,573,913 A	11/19/2016
U.S. PATENT DOCUMENTS	5,578,832 A	11/19/2016
4,362,867 A	5,580,339 A	12/19/2016
4,455,066 A		
4,469,863 A		
4,582,877 A		
4,875,187 A		
4,935,357 A		
4,942,124 A		
4,988,617 A		
4,996,143 A		
5,011,769 A		
5,034,506 A		
5,035,996 A		
5,043,272 A		
5,082,830 A		
5,090,591 A		
5,130,238 A		
5,169,766 A		

Genetic Analysis

The methods of the present invention can be used in the analysis of genetic information of selective genomic regions of interest as well as genomic regions which may interact with the selective region of interest. Amplification methods as disclosed herein can be used in the devices, kits, and methods known to the art for genetic analysis, such as, but not limited to those found in U.S. Pat. Nos. 6,449,562, 6,287,766, 7,361,468, 7,414,117, 6,225,109, and 6,110,709. In some cases, amplification methods of the present invention can be used to amplify target nucleic acid of interest for DNA hybridization studies to determine the presence or absence of polymorphisms. The polymorphisms, or alleles, can be associated with diseases or conditions such as genetic disease. In other cases the polymorphisms can be associated with susceptibility to diseases or conditions, for example,

Methods Of Analysis Identified For Shapero's MARA

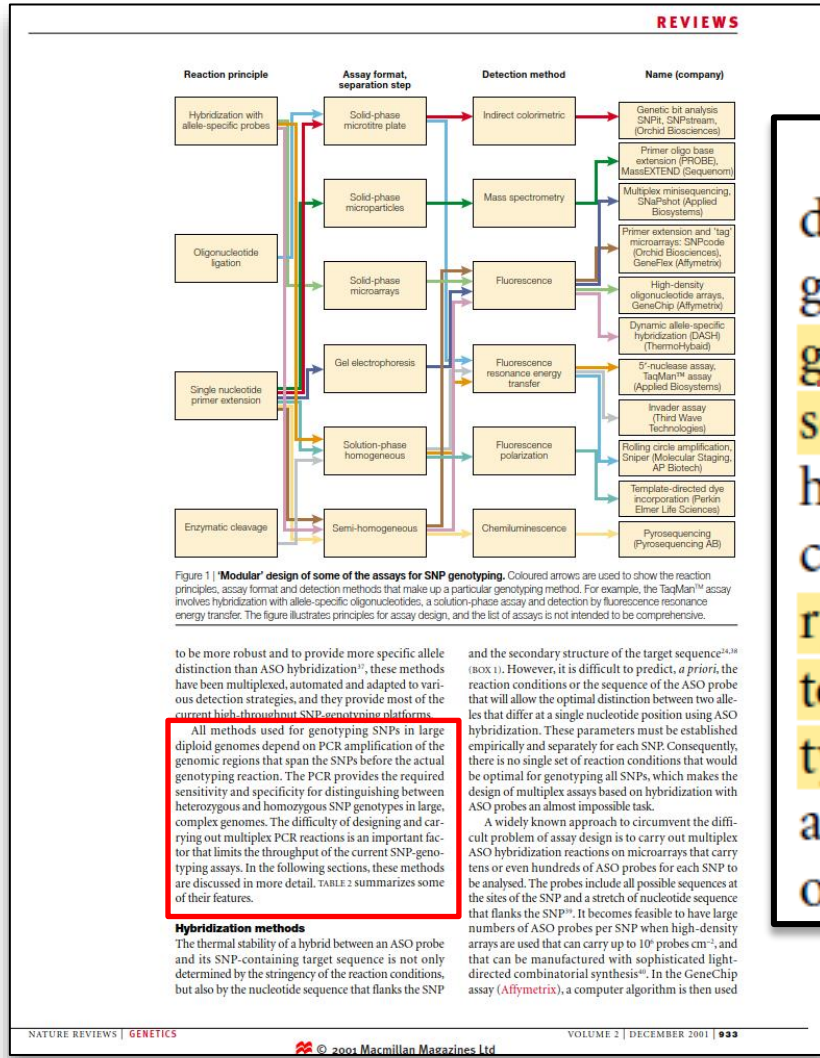


ACCESSING GENETIC VARIATION: GENOTYPING SINGLE NUCLEOTIDE POLYMORPHISMS

Ann-Christine Syvänen

Understanding the relationship between genetic variation and biological function on a genomic scale is expected to provide fundamental new insights into the biology, evolution and pathophysiology of humans and other species. The hope that single nucleotide polymorphisms (SNPs) will allow genes that underlie complex disease to be identified, together with progress in identifying large sets of SNPs, are the driving forces behind intense efforts to establish the technology for large-scale analysis of SNPs. New genotyping methods that are high throughput, accurate and cheap are urgently needed for gaining full access to the

abundant genetic variation of organisms.



All methods used for genotyping SNPs in large diploid genomes depend on PCR amplification of the genomic regions that span the SNPs before the actual genotyping reaction. The PCR provides the required sensitivity and specificity for distinguishing between heterozygous and homozygous SNP genotypes in large, complex genomes. The difficulty of designing and carrying out multiplex PCR reactions is an important factor that limits the throughput of the current SNP-genotyping assays. In the following sections, these methods are discussed in more detail. TABLE 2 summarizes some of their features.

Syvanen (Ex. 1012)

Table 2 Features of SNP-genotyping methods			
Method	Most significant advantage	Disadvantage	Special feature
<i>Hybridization methods</i>	Widely used	Limited genotype discrimination	
Reverse dot blot ²⁸	Multiplexing possible	Prone to non-specific background	Precursor to microarrays
GeneChip microarrays ³⁹	Very high probe density	High failure rates ^{41,42} , expensive	High-tech manufacturing process ⁴⁰
DASH ⁴⁴	Inexpensive labelling method	Complex design rules	Dedicated instrument
PNA ⁴⁶ and LNA ⁴⁸ probes	Specific as ASO probes	Not widely available	Innovative chemical design
TaqMan ^{49,50}	Simplicity of assay	Expensive probes	Quantitative real-time PCR
Molecular Beacons ^{51,52}	Simplicity of assay	Expensive probes	Versatile stem-loop structure
<i>Allele-specific PCR</i>	Real-time PCR assay	Requires optimization	
Intercalating dye ⁶⁰	Inexpensive labelling method	Non-specific products detected	Accurate quantification ¹³⁸
FRET primers ^{61,62}	Simplicity of assay	Expensive primers	Universal FRET primers
AlphaScreen ⁶⁴	Detects specific product	Expensive probes	Applicable to ASO hybridization
<i>Primer extension</i>	Accurate genotyping	Similar reagents as in PCR	
SNPstream, GBA ⁶⁸	Inexpensive and robust	Multiple detection steps	Automated high-throughput process
Multiplex minisequencing, SNaPshot ^{69,71}	Multiplexing capacity	Size separation step	Compatible with capillary DNA sequencers
Pyrosequencing ^{72,74}	Sequencing of up to 50 bases	Expensive, difficult to multiplex	Dedicated instrument

REVIEWS

ACCESSING GENETIC VARIATION GENOTYPING SINGLE NUCLEOTIDE POLYMORPHISMS

Ann-Christine Syvänen

Understanding the relationship between genetic variation and biological function on a large scale is expected to provide fundamental new insights into the biology, evolution and pathophysiology of humans and other species. The hope that single nucleotide polymorphisms (SNPs) will allow genes that underlie complex disease to be identified with progress in identifying large sets of SNPs, are the driving forces behind intense efforts to establish the technology for large-scale analysis of SNPs. New genotyping methods with high throughput, accurate and cheap are urgently needed for gaining full access to the abundant genetic variation of organisms.

**LINKAGE DISEQUILIBRIUM
MAPPING**
Analysing single nucleotide polymorphism alleles in population-based studies to identify loci that are associated with a particular disease or phenotype.

Department of
Medical Sciences—
Molecular Medicine,
Uppsala University
Hospital,
75183 Uppsala, Sweden.
e-mail: Ann-Christine.
Syvanen@medsci.uu.se

Comparison of genomic DNA sequences in different individuals reveals some positions at which two, or in some cases more than two, bases can occur. These single nucleotide polymorphisms (SNPs) are highly abundant, and are estimated to occur at 1 out of every 1,000 bases in the human genome^{1,2}. Depending on where a SNP occurs, it might have different consequences at the phenotypic level. SNPs in the coding regions of genes that alter the function or structure of the encoded proteins are a necessary and sufficient cause of most of the known recessively or dominantly inherited monogenic disorders. These SNPs are routinely analysed for diagnostic purposes. Another important group of SNPs are those that alter the primary structure of a protein involved in drug metabolism. These SNPs are targets for pharmacogenetic analyses³. Missense SNPs in the coding regions of genes, such as the two SNPs in the *apolipoprotein E* gene⁴ and the *factor V Leiden* mutation⁵, can also contribute to common disease. This type of SNP can be analysed to assess the risk of an individual for a particular disease. In addition, it is likely that SNPs in the regulatory regions of genes might influence the risk of common disease. However, most SNPs are located in non-coding regions of the genome, and have no direct known impact on the phenotype of an individual.

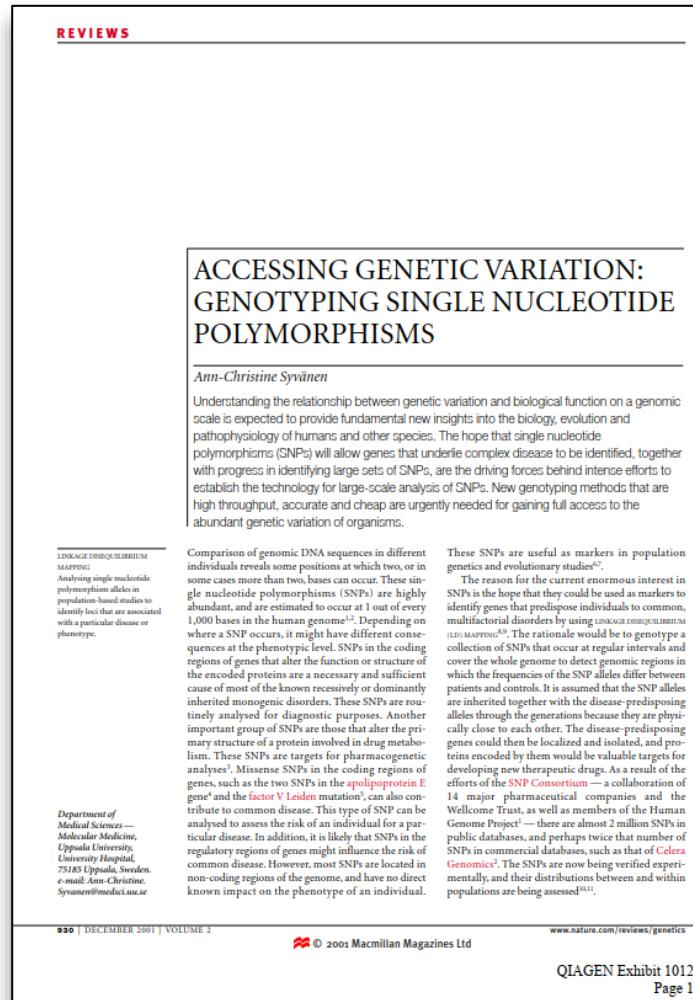
These SNPs are useful as markers in genetics and evolutionary studies^{6,7}.

The reason for the current enormous interest in SNPs is the hope that they could be used to identify genes that predispose individuals to multifactorial disorders by using *LINKAGE DISEQUILIBRIUM MAPPING*^{8,9}. The rationale would be to collect a large number of SNPs that occur at regular intervals across the whole genome to detect genes in which the frequencies of the SNP alleles differ between patients and controls. It is assumed that genes that are inherited together with the disease alleles through the generations because they are located close to each other. The disease-causing genes could then be localized and isolated. The proteins encoded by them would be valuable targets for developing new therapeutic drugs. As part of the SNP Consortium—a collaboration of 14 major pharmaceutical companies and the Wellcome Trust, as well as members of the Genome Project¹⁰—there are almost 2 million SNPs in public databases, and perhaps twice that number in commercial databases, such as *Genomics*¹¹. The SNPs are now being used to identify genes, and their distributions between populations are being assessed^{12,13}.

© 2001 Macmillan Magazines Ltd

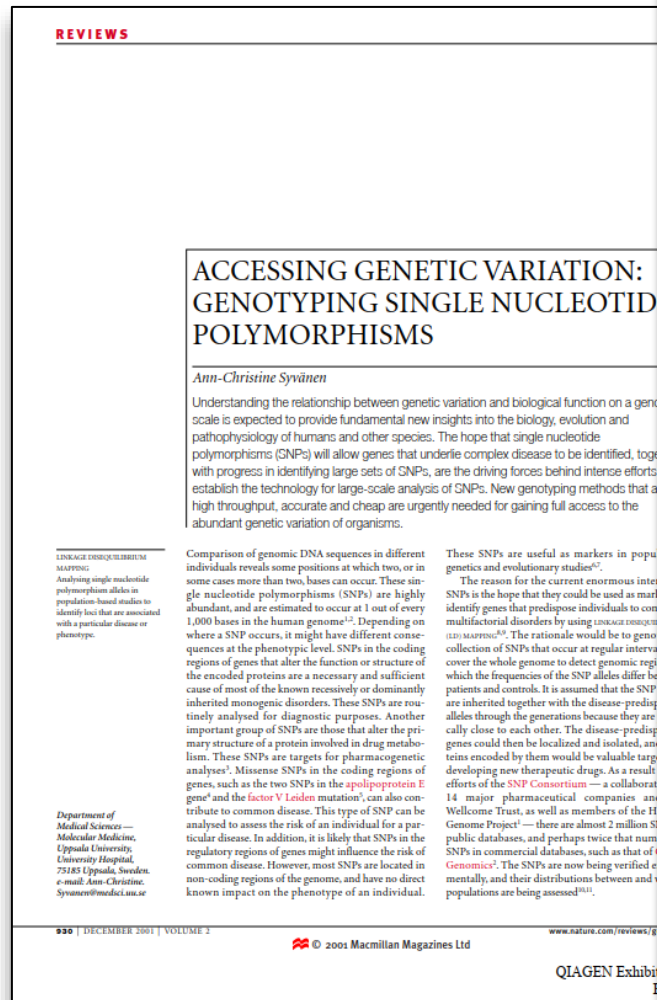
QIAGEN Exhibit 1012
Page 1

In the PYROSEQUENCING (Pyrosequencing AB) method, primer extension is monitored by enzyme-mediated luminometric detection of pyrophosphate, which is released on incorporation of deoxynucleotide triphosphates^{72,73}. The genotype of a SNP is deduced by sequential addition and degradation of the nucleotides using apyrase in a dedicated instrument that operates in a 96-well or 384-well microtitre plate format⁷⁴. Using pyrosequencing, short 30–50-bp sequences of DNA that flank a SNP can be determined. A limitation of the method is that the sequential identification of bases prevents genotyping of several SNPs per reaction in diploid genomes.



Future trends

In practice, the requirement of a PCR amplification step to achieve sensitive and specific SNP genotyping is the principal factor that limits the throughput of assays today. Multiplex PCR amplification of more than ten DNA fragments is difficult to carry out reproducibly owing to the generation of spurious amplification products^{39,41,42,80}. Recently, a PCR strategy based on a single universal PCR primer for amplification of a reduced representation of the human genome that avoids the problem of carrying out multiplex PCR was applied to SNP discovery and validation¹¹⁷. Similar generic amplification strategies could also be devised for genotyping panels of known SNPs to circumvent multiplex PCRs with locus-specific primers. Another suggested approach for avoiding non-specific primer interactions during multiplex PCR is to carry out the amplifications with primers immobilized at physically distinct locations in 'colonies' (PCR colonies)¹¹⁸, gel pads¹¹⁹ or using microelectronic arrays¹²⁰. SNP genotyping by mini-sequencing⁶⁷ or pyrosequencing⁷³ would be compatible with these formats.



Despite the numerous technical advances in detection and multiplexing strategies, no technique clearly represents a breakthrough. It is difficult to predict which, if any, existing SNP-genotyping technology will facilitate the required 100- or 1,000-fold increase in throughput required for whole-genome SNP analysis in sample collections of relevant size. The next generation technology for SNP analysis should avoid the PCR amplification step, and the technology could be based on the analysis of single DNA molecules, which would allow direct determination of the haplotypes formed by the SNPs. Such futuristic candidate technologies might derive from a further development of the system for haplotyping by carbon nanotube atomic force microscopy probes¹²⁹ or a further refinement of the microscopic technology that has recently allowed direct optical mapping of the whole *Escherichia coli* genome¹³⁰. Hopefully, future SNP-genotyping technology will be more elegant than the incremental and brute-force expansion of existing technologies in the way that the Sanger sequencing method has been exploited for genome sequencing. One thing is certain — once

Delseny (Ex. 1007) “High Throughput DNA Sequencing”



4. Applications of classical sequencing and NGSTs to solve biological problems

The availability of several plant genomes has already allowed for addressing a few important questions: How are the plant genomes organized? How do they evolve? And which genes are specific to each species?

NGSTs provide a number of new opportunities such as sequencing several individual genomes within a species to answer questions about genetic diversity at a previously unimaginable scale. The technologies open the way to using sequencing as a genotyping tool. NGSTs also provide transcriptome data with considerably increased depth. Changes in genome structure and sequence modifications in different cells and tissues during development or in response to environmental cues can be investigated. Some of these questions have been recently addressed [6,7,71,72]. In the next sections we focus on what we have learned from analysis of the few available plant genomes and recent developments induced by the widespread use of NGSTs in plant biology laboratories.

Delseny (Ex. 1007) – Advantages over Array



4.3. Population structure, genetic diversity and genotyping

morphisms (SNPs). SNPs are the most abundant variants. Their discovery initially relied on classical sequencing of PCR fragments and re-sequencing by hybridization. A core collection of

* * * *

Although important, these observations give an incomplete view of genetic variation because many SNPs are missed and because in Deis and repeats cannot be investigated by hybridization techniques. They will soon be overwhelmed by new data resulting from NGTs. Re-sequencing of three accessions of *Arabidopsis* [the reference Col-0 accession and two highly divergent accessions, Bur-0 and Tsu-1) by use of Illumina technology has been reported [113]. In fact, this re-sequencing is part of a much larger project,

* * * *

So far, genotyping has involved mostly microsatellites, a technology with limited resolution and requiring much time and human resources. The ultimate resolution is at the base-pair level and NGTs now open the possibility to genotype a line by sequencing it, provided a reference genome is available. This strategy has recently

been illustrated in rice [125]: in total 150 RILs derived from a cross between 93-11 x Nipponbare, (the two indica and japonica rice reference genomes) were re-sequenced with use of the Illumina machine and the barcode technology. With this multiplexing, the sequencing of the 150 lines could thus be completed in two runs. In all, 1,500,000 SNPs were detected and the authors could map >5000 recombination points (33.8 per line, on average). This resolution is -35 x higher than that available with use of classical markers. The time of acquisition of data is 20 x faster than with previous methods. Such an approach can substitute for the genotyping microarray strategy because of its higher resolution.

Delseny (Ex. 1007) – Targeted Genotyping



4.4. Targeted sequencing

For many applications, it is sufficient to re-sequence DNA fragments. So far, many targeted sequencing projects have addressed comparative sequencing of a locus in different, more or less closely related species. With the intensification of sequencing, more samples can be compared. As an example, the organisation of the *Adh1* locus across nine cultivated and wild rice genomes has been reported [126]. These species have differentiated during the last 15-20 My. A total of 46 genes and 4 pseudogenes were identified, with 38 belonging to 8 multigene families. Analysis of each of these families revealed lineage-specific gain and loss of gene members. Transposable elements had a clear role in mediating replacement of intergenic space, gene disruption and gene movement. Similar conclusions were drawn from a study of the agronomically important *MONOCULM1* locus, which controls tiller number [127].

Because many genomes contain large amounts of repetitive sequences and transposable elements, selecting relevant portions might be of interest. This selection is particularly important in medicine, in which sequencing the same locus in a large range of patients is an important target. This finding has led to the development of methods to capture the desired sequences [128]. These approaches will certainly be applied in plants to assess the genetic diversity associated with a number of genes contributing to important agronomic traits, one of the main goals of the international Generation Challenge Program, <http://www.generationcp.org/subprogramme2.php> [129].

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan M. Friedman^{1,3}

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will fall and throughput improve even more in the next few years. Clinical use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the genomic signatures of novel infectious agents. In addition to providing these entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical applications where they are currently being used.

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets of interest, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data transfer capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will require training to use the sequence data effectively, and appropriate methods will need to be developed to deal with the incidental discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

Introduction

DNA sequencing was first described by Maxim and Gilbert¹ and Sanger et al. in 1977.² Subsequent improvements to the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At each step, more sophisticated DNA sequencing instruments, programs, and bioinformatics have provided more automation and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

costs of sequencing to date, but at hand,⁷ it seems only a matter of time before massively parallel sequencing becomes routinely available in clinical laboratories.

Massively parallel sequencing will allow simultaneous screening for mutations in hundreds of loci in genetically heterogeneous disorders, whole-genome screening for novel mutations, and sequence-based detection of novel pathogens that cause human disease. In addition, massively parallel sequencing will permit clinical application of our expanding knowledge of pharmacogenetics, cancer genetics, epigenetics, and complex traits. As with any new clinical test, analysis of clinical utility will have to be undertaken and clear standards and guidelines will need to be put in place before massively parallel sequencing can be routinely offered.

This review describes currently available massively parallel sequencing platforms and their potential impact on clinical testing in medical genetics, with consideration of technical issues that pertain to clinical laboratories and ethical issues that need to be addressed before massively parallel sequencing can be incorporated into routine clinical care.

Sanger Sequencing
Clinical DNA sequencing is usually performed using the Sanger method, which involves the use of DNA polymerase, a primer, and four deoxynucleoside triphosphates (dNTPs) to synthesize a complementary DNA strand. The resulting DNA fragments are then separated by size using gel electrophoresis, and the sequence is determined by reading the bands.

Massively Parallel Sequencing
This section describes the various massively parallel sequencing technologies that are currently available. These include sequencing-by-synthesis (SBS), sequencing-on-chip (SOC), and sequencing-by-ligation (SBL). Each of these methods has its own strengths and weaknesses, and the choice of method will depend on the specific application.

Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan M. Friedman^{1,3}

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will fall and throughput improve even more in the next few years. **Clinical use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the genomic signatures of novel infectious agents. In addition to providing these entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical applications where they are currently being used.**

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets of interest, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data transfer capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will require training to use the sequence data effectively, and appropriate methods will need to be developed to deal with the incidental discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

costs of data storage and analysis. Massively parallel sequencing has had little impact on clinical diagnostics to date, but with the promise of the \$1000 genome close at hand,⁷ it seems only a matter of time before massively parallel sequencing becomes routinely available in clinical laboratories.

Massively parallel sequencing will allow simultaneous screening for mutations in hundreds of loci in genetically heterogeneous disorders, whole-genome screening for novel mutations, and sequence-based detection of novel pathogens that cause human disease. In addition, massively parallel sequencing will permit clinical application of our expanding knowledge of pharmacogenetics, cancer genetics, epigenetics, and complex traits. As with any new clinical test, analysis of clinical utility will have to be undertaken and clear standards and guidelines will need to be put in place before massively parallel sequencing can be routinely offered.

This review describes currently available massively parallel sequencing platforms and their potential impact on clinical testing in medical genetics, with consideration of technical issues that pertain to clinical laboratories and ethical issues that need to be addressed before massively parallel sequencing can be incorporated into routine clinical care.

¹Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada; ²Child & Family Research Institute, Vancouver, British Columbia V5Z 4S6, Canada; ³Child & Family Research Institute, Vancouver, British Columbia V5Z 4S6, Canada; *Correspondence: tucker@interchange.ubc.ca
DOI: 10.1016/j.ajhg.2009.06.022. ©2009 by The American Society of Human Genetics. All rights reserved.

Tucker (Ex. 1013)

Whole-genome sequencing is not yet practical in the clinical setting, and it is likely that massively parallel

* * * * *

sequencing will initially be used to sequence selected genomic regions. Potential applications, which are discussed in more detail below, include testing many different loci for mutations simultaneously in a patient with a genetically heterogeneous disease or screening a large number of samples for mutations in a set of candidate genes. The latter

is made possible by molecular “barcoding,” which involves adding a short DNA sequence tag that is unique to a particular patient to every DNA fragment made from that patient’s sample. Several patient samples can then be pooled and sequenced together, and the sequences obtained from each patient can be separated bioinformatically.²⁷

Tucker (Ex. 1013) - Targeted

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan M. Friedman^{1,3}

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will continue to decrease. Massively parallel sequencing has had little impact on clinical diagnostics, but it is likely that costs of data storage and analysis. Massively parallel sequencing has had little impact on clinical diagnostics, but it is likely that costs of data storage and analysis. Massively parallel sequencing has had little impact on clinical diagnostics, but it is likely that costs of data storage and analysis.

Routine clinical use of massively parallel sequencing requires higher accuracy, better ways to select genomic subsets, and improvements in the functionality, speed, and data analysis software. In addition, substantial clinical laboratory computer infrastructure, data storage, and capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will need training to use the sequence data effectively, and methods will need to be developed to deal with the discovery of pathogenic mutations and variants of clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and to increase the responsibility of geneticists to ensure that the vast amount of personal genomic data produced is used in a medically and socially appropriate manner.

Introduction

DNA sequencing was first described by Maxam and Sanger et al. in 1977.¹ Subsequent improvements to the Sanger method have increased the efficiency of DNA sequencing more than three orders of magnitude. At the same time, more sophisticated DNA sequencing instruments, and bioinformatics have provided more efficient and higher throughput. Several massively parallel sequencing methods have become available in the past couple of years, allowing larger-scale projects to be completed, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.²⁻⁶

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

Targeted sequencing requires substantial up-front preparation to select the DNA segments of interest. PCR with modifications to permit higher multiplexing is one useful way to do this. Examples of some available methods are shown in Figures 6A–6C. Such methods have been used to amplify hundreds of selected exons from a DNA sample.^{27–29}

Alternatively, targeted regions can be obtained by direct hybridization to oligonucleotide arrays containing probes complementary to the regions of interest. The array is

then denatured, and the fragments obtained can be amplified or directly sequenced, depending on the depth of hybridization (Figure 6D). This method has been successful on a large number of target regions with both the 454 FLX Sequencer³⁰ and the Illumina Genome Analyzer.^{31–34} Several other novel methods have recently been developed to enrich segments of interest.^{35–37} However, any enrichment strategy is unlikely to be effective for all genomic applications, highlighting the importance of further development of enrichment techniques.

Massively Parallel Sequencing Platforms

This section provides a brief overview of commercially available massively parallel sequencing platforms. For more detailed discussion of massively parallel platforms, readers are referred to recent reviews.^{9,10}

The Illumina Genome Analyzer, which uses "sequencing by synthesis"^{11,12} to produce single reads of 75+ basepairs (bp) (Figure 2), can currently generate about 17 gigabases

¹Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6H 3N1, Canada; ²BC Cancer Agency Genome Sciences Centre, Vancouver, British Columbia V5Z 4S6, Canada; ³Child & Family Research Institute, Vancouver, British Columbia V6H 3N1, Canada
*Correspondence: ttucker@interchange.ubc.ca
DOI 10.1016/j.ajhg.2009.06.022. ©2009 by The American Society of Human Genetics. All rights reserved.

Tucker (Ex. 1013) - Genotyping

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan M. Friedman^{1,3}

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will fall and throughput improve even more in the next few years. Clinical use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the genomic signatures of novel infectious agents. In addition to providing these entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical applications where they are currently being used.

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets of interest, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data transfer capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will require training to use the sequence data effectively, and appropriate methods will need to be developed to deal with the incidental discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

Introduction

DNA sequencing was first described by Maxim and Gilbert¹ and Sanger et al. in 1977.² Subsequent improvements to the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At each step, more sophisticated DNA sequencing instruments, programs, and bioinformatics have provided more automation and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

costs of data sequencing to date, but at hand,⁷ it parallel seq laboratories

Massively screening for heterogeneous novel mutations pathogenic, massively pre tion of our cancer gene any new cl to be unde need to be ing can be

This rev parallel seq on clinical of technical ethical issu parallel seq ical care.

Sanger Seq Clinical DN lary-based, usually pre interest. Th that involv primer and proach can cies at a co

Massively 2 This section available n more detail readers are The Illum by synthe (bp) (Figure

Interpretation of Data

Six billion base pairs of DNA per patient are a lot of data to interpret. Computers and software can help—in fact, they are essential—but clinical interpretation of genome sequence data will always require a well-trained and experienced genetics professional. As a practical matter, only a subset of a person's genome can actually be examined for variants that cause or predispose to disease. This subset may either be selected before sequencing is done with a technique like those described in the previous section or, if whole-genome sequencing is performed, bioinfor-

* * * * *

The subset examined for a particular patient may be different under different circumstances. If the goal is to obtain genotypes of a comprehensive set of single-nucleotide polymorphisms (SNPs), copy-number variants (CNVs), and other structural variants (SVs) for disease prediction, it would make most sense to focus on these polymorphic regions. If the goal is to identify a sequence mutation in an unidentified locus, sequencing all exons and adjacent promoter regions may be most informative. If the goal is to survey a more limited subset of loci for rare sequence mutations and copy number changes that are known to produce a genetically heterogeneous condition such as autism, a more targeted approach may be optimal. In other circumstances, such as

Applications in Clinical Genetics Labs: Improvements in Current Diagnostic Capabilities

Sanger sequencing has been used clinically in conjunction with PCR for more than a decade to identify sequence mutations and other variants of selected Mendelian disease genes, but wider application of sequencing to clinical testing has been limited by cost and throughput. More recently, many clinical genetics laboratories have adopted array genomic hybridization as a means of detecting copy number changes that cause intellectual disability and other birth defects. Here we consider how massively parallel sequencing could be used to perform these and other forms of genetic testing that are currently available clinically more cost effectively and with higher throughput.

¹Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia Centre, Vancouver, British Columbia V5Z 4S6, Canada; ²Child & Family Research Institute, ³Correspondence: tucker@interchange.ubc.ca
DOI: 10.1016/j.ajhg.2009.06.022. ©2009 by The American Society of Human Genetics. All rights reserved.

Tucker (Ex. 1013) - Advantages over Array

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genomics

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and J. Michael

Massively parallel sequencing has reduced the cost of genomic sequencing by more than three orders of magnitude, and it seems likely that cost of sequencing will continue to improve even more in the next few years. The use of massively parallel sequencing will provide a means of rapid, high-throughput screening of thousands of loci for pathogenic variants and by sequencing biological specimens for the detection of novel infectious agents. In addition to entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in many situations where they are currently being used.

Routine clinical use of massively parallel sequencing requires higher accuracy, better ways to select genomic regions of interest, and improvements in the functionality, speed, and data analysis software. In addition, substantial expansion of laboratory computer infrastructure, data storage, and capacity will be needed to handle the extremely large volumes of data produced. Clinicians and laboratory personnel will need training to use the sequence data effectively, and methods will need to be developed to deal with the discovery of pathogenic mutations and variants of clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially appropriate manner.

Introduction

DNA sequencing was first described by Maxam and Sanger et al. in 1977.² Subsequent improvements in the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At each step, more sophisticated DNA sequencing instruments, programs, and bioinformatics have provided more automation and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

Recognizing Disease Predisposing and Protective Factors

Massively parallel sequencing has distinct advantages as a means of recognizing variants that may predispose to, or protect against, the development of common complex diseases. SNP arrays, even those with millions of features, provide genotypes of only a small fraction of the variants present in an individual. In contrast, massively parallel sequencing could provide complete information on all SNPs and other disease-associated variants that are present. This may be especially important in people whose origin is not from a population for which the tag-SNPs on most genotyping arrays provide optimal coverage. In addition, massively parallel sequencing could detect rare variants as well as the common ones for which genotyping chips are designed, and rare variants may be especially important in recognizing people with predispositions to developing certain complex diseases.^{43,44}

interest. The DNA is then sequenced by "cycle sequencing" that involves several rounds of template denaturation, primer annealing, and extension (Figure 1).⁸ This approach can achieve read lengths of ~1 Kb and high accuracies at a cost of about \$500 per megabase (Mb).⁹

Massively Parallel Sequencing Platforms

This section provides a brief overview of commercially available massively parallel sequencing platforms. For more detailed discussion of massively parallel platforms, readers are referred to recent reviews.^{9,10}

The Illumina Genome Analyzer, which uses "sequencing by synthesis"^{11,12} to produce single reads of 75+ basepairs (bp) (Figure 2), can currently generate about 17 gigabases

Cancer genomes are very heterogeneous because of the genetic instability and clonal evolution that characterize tumor development. Moreover, most tumors are composed of several different types of cells, some of which may not be part of the malignant clone. Consequently, mutations involved in tumor development or progression may be present in only a small fraction of the cells and may be missed with array studies and PCR-based assays of whole-tumor DNA. Massively parallel sequencing could accurately measure cancer-associated genetic alterations that occur in only a small fraction of the cells tested because the independent amplification and sequencing of millions of different DNA fragments from each specimen permits the accurate detection of rare sequences if the depth of coverage is sufficient.

* * * *

Identification of Structural Variants

Structural variants (SVs) include copy number variants (CNVs) as well as inversions and other chromosomal rearrangements that do not involve a change in copy number. Most CNVs occur as benign polymorphisms, but pathogenic CNVs have recently been found to be the most frequent recognizable cause of intellectual disability and some other birth defects.^{63,64} Array genomic hybridization is now used clinically to detect such pathogenic CNVs, but this technology does not detect balanced SVs. Other CNVs are involved in modulation of complex traits⁶⁵ and disease susceptibility.^{66,67} Massively parallel sequencing can reliably identify both balanced and unbalanced SVs⁶⁸ and provides much higher resolution than is possible with array genomic hybridization, permitting better genotype-phenotype correlation.

Tucker (Ex. 1013) - Targeted MPS

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genomics

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan

Massively parallel sequencing has reduced the cost of the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will continue to improve even more in the next few years. The use of massively parallel sequencing will provide a window into the cause of many diseases of unknown etiology through the simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the signatures of novel infectious agents. In addition to providing entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical settings where they are currently being used.

Routine clinical use of massively parallel sequencing requires higher accuracy, better ways to select genomic subregions, and improvements in the functionality, speed, and data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and capacity will be needed to handle the extremely large amounts of data produced. Clinicians and laboratory personnel will need training to use the sequence data effectively, and methods will need to be developed to deal with the discovery of pathogenic mutations and variants of uncertain significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially appropriate manner.

Introduction

DNA sequencing was first described by Maxam and Sanger et al. in 1977.¹ Subsequent improvements in the Sanger method have increased the efficiency more than three orders of magnitude. More sophisticated DNA sequencing instruments, and bioinformatics have provided more information and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale projects for whole-genome sequencing, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing becomes feasible for smaller laboratories to adopt, although doing so involves substantial costs. These include not just the massive sequencing machines themselves, but also the

Simultaneous Screening for Mutations at Multiple Loci

Some conditions that are seen frequently by clinical geneticists may be caused by Mendelian mutations of many dozens or hundreds of different genetic loci. Examples include intel-

* * * *

Massively parallel sequencing could provide the opportunity to test hundreds or even thousands of candidate loci simultaneously. This could be done by whole-genome sequencing and selective bioinformatic analysis or by sequencing candidate regions selected by array capture or one of the other methods of targeted amplification described above (Figure 6).

¹Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V6H 3N1, Canada; ²BC Cancer Agency Genome Sciences Centre, Vancouver, British Columbia V5Z 4S6, Canada; ³Child & Family Research Institute, Vancouver, British Columbia V6H 3N1, Canada
*Correspondence: ttucker@interchange.ubc.ca
DOI: 10.1016/j.ajhg.2009.06.022. ©2009 by The American Society of Human Genetics. All rights reserved.

Tucker (Ex. 1013) - Pool/Barcode For Capacity

REVIEW

Massively Parallel Sequencing The Next Big Thing in Genetics

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan S. Tucker¹

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will continue to decrease and throughput improve even more in the next few years. The use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through genome-wide screening of thousands of loci for pathogenic variants, and by sequencing biological specimens for the detection of novel infectious agents. In addition to providing entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical settings where they are currently being used.

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data processing capacity will be needed to handle the extremely large amounts of data produced. Clinicians and laboratory personnel will need training to use the sequence data effectively, and additional methods will need to be developed to deal with the discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

Introduction

DNA sequencing was first described by Maxam and Gilbert¹ and Sanger et al. in 1977.² Subsequent improvements in the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At the same time, more sophisticated DNA sequencing instruments, such as microarrays and bioinformatics have provided more information and higher throughput. Several massively parallel sequencing methods have become available in the past couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing becomes feasible for smaller laboratories to adopt the technology, although doing so involves substantial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

Mutation Detection in Mendelian Disease

Massively parallel sequencing could be used to sequence very large as well as smaller Mendelian disease genes fully, covering all exons as well as the associated 5', 3', and intronic sequences. This improved coverage could increase the sensitivity of mutation detection over current methods, which often employ a screening technique such as SSCP prior to sequencing and limit the segments sequenced to the exons that are most frequently affected by mutations or, at best, just to the exons and the immediately surrounding bases.^{30,32,42} However, the capacity of current massively parallel sequencing platforms is too great for efficient sequencing of most single genes. For example, a single lane on an Illumina flow cell would provide more than 68× coverage of the entire genomic segment containing the DMD gene (MIM #300377) in a male. A solution to this problem would be to pool DNA samples from several family members or from unrelated patients that have been prepared by locus-specific PCR and barcoded so that the DNA from each individual can be distinguished.²⁷

Dr. Bustamante (Ex. 1120) - Genotyping by NGS

PO Sur-Reply

I. PETITIONER FAILS TO DEMONSTRATE A MOTIVATION TO COMBINE

Petitioner's arguments are almost entirely based on conclusory statements.

See Paper 22, 2 (“[B]y 2010, scientists were genotyping using NGS.”). None of Petitioner's cited references (*i.e.*, Shapero, Delseny, and Syvänen) include such broad statements, and Petitioner fails to identify evidence in the record to support its allegations that a POSA would combine the prior art to arrive at the claimed methods.

Ex. 1120 at 2.

Dr. Bustamante Testimony

Q. You're not disputing, are you, that by 2010, whether for actual clinical diagnostic applications or for research applications, groups of people, scientists were genotyping using next generation sequencing, are you?

A. I am not disputing that.

Ex. 1120 at 367.

Metzker 2009 (Ex. 1025)

This Review focuses on commercially available technologies from Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences, the Polonator instrument and the near-term technology of Pacific Biosciences, who aim to bring their sequencing device to the market in 2010. Nanopore sequencing is not covered, although interested readers are directed to an article by Branton and colleagues⁶, who describe the advances and remaining challenges for this technology. Here, I present a technical review of template preparation, sequencing and imaging, genome alignment and assembly, and current NGS platform performance to provide guidance on how these technologies work and how they may be applied to important biological questions. I highlight the applications of human genome resequencing using targeted and whole-genome approaches, and discuss the progress

VOLUME 11 | JANUARY 2010 | 31

Alleged Metzker Teaching Away



Genome enrichment

Despite the substantial cost reductions associated with NGS technologies in comparison with the automated Sanger method⁶², whole-genome sequencing is still an expensive endeavour (see below). An interim solution to this problem may be to use NGS platforms to target specific regions of interest. This strategy can be used to examine all of the exons in the genome, specific gene families that constitute known drug targets or megabase-size regions that are implicated in disease⁶³ or pharmacogenetics effects⁶⁴ through genome-wide association studies. The concept of targeting specific regions of the genome is well established, with PCR being the most widely used method, albeit on a small scale. PCR coupled with Sanger sequencing is suitably matched for analysing a handful of candidate genes⁶⁵, but coupling PCR with high-throughput NGS platforms for targeting strategies is not practical, as sample preparation would require handling tens of thousands of primers individually or in large multiplex groups to meet the needs of a single instrument run. A recent article by Frazer and colleagues in collaboration with RainDance Technologies reported the simultaneous amplification of 3,976 products using microdroplet PCR technology⁶⁶. Here, a microfluidic device creates aqueous picolitre-volume droplets of forward- and reverse-targeting primers in an oil solution. The primer droplets that are targeted to different regions of the genome merge with separate picolitre droplets that contain fragmented genomic DNA and associated PCR reagents, and these mixed droplets are thermal cycled in a single tube (FIG. 5a). The authors reported an 84% capture efficiency with 90% of the targeted bases showing uniform coverage using the microdroplet PCR method sequenced with either the Roche/454 or Illumina/Solexa platform⁶⁶.

Custom-designed oligonucleotide microarrays and solution-based hybridization strategies have also been used for targeting regions of interest. For example, Roche/NimbleGen⁶⁷ oligonucleotide microarrays have been designed to provide complementary sequences for solid-phase hybridization to enrich for exons^{68,69} or contiguous stretches of genomic regions^{68,70} (FIG. 5b). The BCM-HGSC⁶⁸ and Cold Spring Harbor Laboratory⁶⁹ groups reported capture efficiencies of 65–77% (Roche/454 platform) and 53% (Illumina/Solexa platform), respectively, with targeted exons being covered by at least one NGS read. Recently, the BCM-HGSC group reported >90% capture efficiency with at least 10× base coverage (Roche/454 and Life/APG platforms) by chip optimization that decreases the number of probes in higher coverage regions while increasing those in low coverage regions (R. A. Gibbs, personal communication).

Other groups have captured specific genomic regions by using solution-based hybridization methods, such as molecular inversion probes (MIPs)^{71–73} and biotinylated RNA capture sequences⁷⁴ (FIG. 5c). Shendure, Church and colleagues⁷¹ were the first group to use MIPs in exon targeting by designing specific oligonucleotide ends that flank the exons of interest. However, in duplicate experiments only 20% of the targets were captured and even fewer exon regions (11%) were found in both data

Next-generation sequencing applications

The production of large numbers of low-cost reads makes the NGS platforms described above useful for many applications. These include variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assemblies of bacterial and lower eukaryotic genomes, cataloguing the transcriptomes of cells, tissues and organisms (RNA-seq)⁵, genome-wide profiling of epigenetic marks and chromatin structure using other seq-based methods (ChIP-seq, methyl-seq and DNase-seq)⁴, and species classification and/or gene discovery by metagenomics studies⁷⁷. With this many applications, which platform is best suited for a given biological experiment? For example, the Illumina/Solexa and Life/APG platforms are well suited for variant discovery by resequencing human genomes because gigantic volumes of high-quality bases are produced per run (TABLE 1). Furthermore, the Helicos BioSciences platform is well suited for applications that demand quantitative information in RNA-seq⁷⁸ or direct RNA sequencing, as it sequences RNA templates directly without the need to convert them into cDNAs⁷⁹. TABLE 1 provides an overview of NGS technologies, instrument performance and cost, pros and cons, and recommendations for biological applications; however, the rapid pace of technological advances in the field could change this information in the near future. Readers are directed to several excellent reviews on RNA-seq⁵, ChIP-seq⁸⁰ and metagenomics⁷⁷.

Tucker (Ex. 1013) - Pool/Barcode For Capacity

REVIEW

Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine

Tracy Tucker,^{1,*} Marco Marra,^{1,2} and Jan M. Friedman^{1,3}

Massively parallel sequencing has reduced the cost and increased the throughput of genomic sequencing by more than three orders of magnitude, and it seems likely that costs will fall and throughput improve even more in the next few years. Clinical use of massively parallel sequencing will provide a way to identify the cause of many diseases of unknown etiology through simultaneous screening of thousands of loci for pathogenic mutations and by sequencing biological specimens for the genomic signatures of novel infectious agents. In addition to providing these entirely new diagnostic capabilities, massively parallel sequencing may also replace arrays and Sanger sequencing in clinical applications where they are currently being used.

Routine clinical use of massively parallel sequencing will require higher accuracy, better ways to select genomic subsets of interest, and improvements in the functionality, speed, and ease of use of data analysis software. In addition, substantial enhancements in laboratory computer infrastructure, data storage, and data transfer capacity will be needed to handle the extremely large data sets produced. Clinicians and laboratory personnel will require training to use the sequence data effectively, and appropriate methods will need to be developed to deal with the incidental discovery of pathogenic mutations and variants of uncertain clinical significance. Massively parallel sequencing has the potential to transform the practice of medical genetics and related fields, but the vast amount of personal genomic data produced will increase the responsibility of geneticists to ensure that the information obtained is used in a medically and socially responsible manner.

Introduction

DNA sequencing was first described by Maxim and Gilbert¹ and Sanger et al. in 1977.² Subsequent improvements to the Sanger method have increased the efficiency and accuracy more than three orders of magnitude. At each step, more sophisticated DNA sequencing instruments, programs, and bioinformatics have provided more automation and higher throughput. Several massively parallel sequencing methods have become available in the last couple of years, allowing larger-scale production of genomic sequence, and the number of human genomes sequenced with such instrumentation is now increasing rapidly.³⁻⁶

As the cost of massively parallel sequencing falls, it becomes feasible for smaller laboratories to adopt this technology, although doing so involves substantial initial costs. These include not just the massively parallel sequencing machines themselves, but also the associated

Sanger
Clinical
lary-b
usual
inter
that
prime
proac
cies a

Mass
This
avali
more
reade
The
by sy
(bp) (

¹Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada; ²Centre, Vancouver, British Columbia V5Z 4S6, Canada; ³Child & Family Research Institute, Edmonton, Alberta T6G 2G4, Canada
*Correspondence: ttucker@interchange.ubc.ca
DOI 10.1016/j.ajhg.2009.06.022. ©2009 by The American Society of Human Genetics

Mutation Detection in Mendelian Disease

Massively parallel sequencing could be used to sequence very large as well as smaller Mendelian disease genes fully, covering all exons as well as the associated 5', 3', and intronic sequences. This improved coverage could increase the sensitivity of mutation detection over current methods, which often employ a screening technique such as SSCP prior to sequencing and limit the segments sequenced to the exons that are most frequently affected by mutations or, at best, just to the exons and the immediately surrounding bases.^{30,32,42} However, the capacity of current massively parallel sequencing platforms is too great for efficient sequencing of most single genes. For example, a single lane on an Illumina flow cell would provide more than 68× coverage of the entire genomic segment containing the *DMD* gene (MIM #300377) in a male. A solution to this problem would be to pool DNA samples from several family members or from unrelated patients that have been prepared by locus-specific PCR and barcoded so that the DNA from each individual can be distinguished.²⁷

Other Alleged Teaching Away Factors



1. Enrichment Level Requirements

What level of enrichment does the patent require that the library preparation methodology achieve?

A. There's no --

MS. CIPRIANO: Objection to form.

A. There's no specific level of enrichment that's outlined.

Dr. Bustamante (Ex. 1120) - Testimony Re: Enrichment

Q. We have a situation where we created a library using a single primer.

A. Okay.

Q. And only 25 percent of the fragments that get created that get read on the Illumina sequencer are really what you want.

A. Okay.

Q. And 75 percent are not what you want.

A. Okay.

Ex. 1120: 71:22-72:10

As a biostatistician or somebody who worked with raw data from a Illumina sequencer from prior to 2012, would you say that that data, which was only 25 percent of what you want, can't be used?

A. That's not what I said.

You can, certainly, use it and you might be able to interpret it and there might be situations where the sample is so precious that you are willing to have only a couple percent, right.

Ex. 1120: 73:9-20.

Dr. Bustamante Testimony: 25% Enrichment and MPS

Q. We have a situation where we created a library using a single primer.

A. Okay. 09:42:10

Q. And only 25 percent of the fragments that get created that get read on the Illumina sequencer are really what you want.

A. Okay. 09:42:21

Q. And 75 percent are not what you want.

A. Okay. 09:42:24

Q. Would you say that that would be characteristic of "a relatively high rate" of off-target hybridization, as Dr. Brooks was talking about in Paragraph 14?

A. And, again, just to clarify, 25 percent are on target, 75 percent are off target?

Q. Correct.
A. That's the readout from the experiment?

Q. Correct.
A. I would say that's a high degree of off target.

Q. Okay. "That's a high degree of off target period"?

A. Yes.

As a biostatistician or somebody who worked with raw data from a Illumina sequencer from prior to 2012, would you say that that data, which was only 25 percent of what you want, can't be used?

A. That's not what I said. You can, certainly, use it and you might be able to interpret it and there might be situations where the sample is so precious that you are willing to have only a couple percent, right.

We did a lot of work in ancient DNA sequencing where that was the case. We've done work in other situations where, you know, the minority of the sample is what you want.

Dr. Bustamante (Ex. 1120) Neandertal Enrichment

11 Q. But these authors pre-2010 11:48:42
 12 didn't run from the hills and say, ah, I 11:48:45
 13 can't use why NGS, I have to use an 11:48:48
 14 Illumina sequencer to sequence this, even 11:48:49
 15 though some of the stuff I'm putting on 11:48:52
 16 the machine might only be 4 percent of 11:48:53
 17 what I want -- 96 percent is garbage and 11:48:56
 18 4 percent is what I want? They didn't 11:49:00
 19 run for the hills, did they? 11:49:01
 20 A. As I've stated in my 11:49:02
 21 testimony, there are some biological 11:49:04
 22 questions that are so interesting that 11:49:06
 23 you are absolutely willing to throw many 11:49:07
 24 resources at the problem. This is the 11:49:10
 25 team that won the Nobel Prize for this 11:49:13
 Page 169

Ex. 1120 at 169.

A Draft Sequence of the Neandertal Genome

RESEARCH ARTICLE

collagen to allow a direct date. The third bone, V33.26, comes from layer G (sublayer unknown) and has not been previously used for large-scale DNA sequencing. It was directly dated to $44,450 \pm 550$ years B.P. (OxA-V-2291-18, uncalibrated).

Sequencing library construction. A total of nine DNA extracts were prepared from the three bones (table S4) using procedures to minimize laboratory contamination that we have developed over the past two decades (22, 41). Samples of each extract were used to construct Roche/454 sequencing libraries that carry the project-specific tag sequence 5'-TGAC-3' in their 3'-ends. Each library was amplified with the primers used in the 454 sequencing emulsion PCR process. To estimate the percentage of endogenous Neandertal DNA in the extracts, we carried out sequencing runs using the 454 Life Sciences GS FLX platform and mapped the reads against the human, chimpanzee, rhesus, and mouse genomes as well as all nucleotide sequences in GenBank. DNA sequences with a significantly better match to the primate genomes than to any of the other sources of sequences were further analyzed. Mitochondrial DNA contamination from modern humans was estimated by primer extension capture (46) using six biotinylated primers that target informative differences between human and Neandertal mtDNA (45), followed by sequencing on the GS FLX platform. Extracts that contained more than 1.5% hominin DNA relative to other DNA were used to construct further libraries. These were similarly analyzed to assess the percentage of hominin DNA and, if found suitable, were used for production sequencing on the 454 Life Sciences GS FLX/Titanium and Illumina GAII platforms.

Enrichment of Neandertal DNA. Depending on the extract, between 95 and 99% of the DNA sequenced in the libraries was derived from nonprimate organisms, which are presumably derived from microbes that colonized the bone after the death of the Neandertals. To improve the ratio of Neandertal to microbial DNA, we identified restriction enzymes that preferentially cut bacterial DNA sequences in the libraries and treated the libraries with these to increase the relative

proportion of Neandertal DNA in the libraries (SOM Text 1). Such enzymes, which have recognition sites rich in the dinucleotide CpG, allowed a 4- to 6-fold increase in the proportion of Neandertal DNA in the libraries sequenced. This is expected to bias the sequencing against GC-rich regions of the genome and is therefore not suitable for arriving at a complete Neandertal genome sequence. However, for producing an overview of the genome at about one-fold coverage, it drastically increases the efficiency of data production without unduly biasing coverage, especially in view of the fact that GC-rich sequences are over-represented in ancient DNA sequencing libraries (23, 45) so that the restriction enzyme treatment may help to counteract this bias.

Sequencing platforms and alignments. In the initial phase of the project, we optimized DNA extraction technology and library construction [e.g., (47)]. In a second phase, we carried out production sequencing on the 454 Life Sciences GS FLX platform from the bones V33.16 and V33.26 (0.5 Gb and 0.8 Gb of Neandertal sequence, respectively). In the third phase, we carried out production sequencing on the Illumina/Solexa GAII platform from the bones V33.16, V33.25, and V33.26 (1.2 Gb, 1.3 Gb, and 1.5 Gb, respectively) (table S4). Each molecule was sequenced from both ends (SOM Text 2), and bases were called with the machine learning algorithm Ibis (48). All reads were required to carry correct clean-room tags, and previous data where these tags were not used (40, 42) were not included in this study. Except when explicitly stated, the analyses below are based on the largest data sets, generated on the Illumina platform. In total, we generated 5.3 Gb of Neandertal DNA sequence from about 400 mg of bone powder. Thus, methods for extracting and sequencing DNA from ancient bones are now efficient enough to allow genome-wide DNA sequence coverage with relatively minor damage to well-preserved paleontological specimens.

The dominant type of nucleotide misincorporation when ancient DNA is amplified and sequenced is due to deamination of cytosine residues (25). This causes C to T transitions in the DNA sequences,

Downloaded from https://www.science.org on September 16, 2025

Ex. 1102 at 2.

Bustamante (Ex. 1120) Re Enrichment Levels

Q. So these people sequenced on 15:30:37
a massively parallel platform, as 15:30:40
reported in the literature, even though 15:30:42
there was only 65 -- at the low end 15:30:44
65 percent of on target, right? 15:30:47

A. Yes. 15:30:53

Q. Okay. And then if we go down, 15:30:53
the sentence continues. And I'll just -- 15:30:56
it continues after the last paren. So 15:31:00
"(Roche/454 platform) and 53% 15:31:03
(Illumina/Solexa platform), respectively, 15:31:09
with targeted exons being covered by at 15:31:12
least one NGS read." 15:31:15

Do you see that? 15:31:17

A. I do. 15:31:17

Q. So, when somebody enriched 15:31:18
for an Illumina platform, they, actually, 15:31:23
reported data, even useful data, even 15:31:26
though the percent of on-target reads was 15:31:30
53 percent, right? 15:31:35

A. With, one single read, right. 15:31:39
Whether you can draw any inference 15:31:40
from a single NGS read is an open 15:31:42

Page 324

- Testimony About Results Presented in Metzker Ex. 1025 at 11

Dr. Bustamante (Ex. 1120) Re: Dr. Brooks (Ex.2026.37-38)

Q. Alright. Let's go back to Paragraph 16 of the Brooks Declaration.

Why in 2012 with single primer extension would you have expected only 50 percent of the reads to be target sequenced that you were interested in?

A. I can only speculate as to why Dr. Brooks expected 50 percent.

Q. Do you agree with him?

A. I'm not sure I would agree to 50 percent.

Ex. 1120 at 279

Q. So is there a number in 2011 level of reads that you get that are on target that are -- that you would do and that you would not do in a massively parallel system?

A. It really depends on the biological question. So I would say that for some problems -- and if you're going to get plenty of depth of coverage, you can tolerate 50 percent of target. For others, 50 percent off target might not be sufficient to draw good conclusions until -- you'd have to do more sequencing in order to understand how well it's working.

Ex. 1120 at 281

Dr. Bustamante: Predicting On Target Level

1 CARLOS D. BUSTAMANTE, Ph.D.
 2 there? 09:35:06
 3 MS. CIPRIANO: Objection, 09:35:07
 4 compound. 09:35:08
 5 A. Could you restate the 09:35:09
 6 question? 09:35:10
 7 Q. Sure. 09:35:10
 8 Do you have any idea quantitatively 09:35:11
 9 what he means by "relatively high off 09:35:13
 10 target"? 09:35:17
 11 A. I can't read Dr. Brooks' mind. 09:35:18
 12 I can't tell you what he had in mind when 09:35:21
 13 he wrote that, in terms of quantitatively. 09:35:23
 14 Q. Alright. Putting yourself 09:35:25
 15 back in 2012, what is your estimation for 09:35:26
 16 how much off target you're going to get? 09:35:31
 17 A. As I said, it depends on the 09:35:33
 18 genome. It depends on the gene region. 09:35:35
 19 It depends on a myriad of factors. It 09:35:39
 20 could be high. 09:35:44
 21 Q. So greater than 50 percent or 09:35:45
 22 less than 50 percent? 09:35:48
 23 A. It depends on the region of 09:35:49
 24 interest. But, you know, it depends on 09:35:52
 25 how many copies we're thinking about, 09:35:56
 Page 64

Ex. 1120 at 64.

1 CARLOS D. BUSTAMANTE, Ph.D.
 2 so... 09:36:00
 3 Q. So, again, if we were taking 09:36:00
 4 1 loci, do you think that somebody -- 09:36:02
 5 would having a reference genome and using 09:36:03
 6 a computer program to design their 09:36:06
 7 primers, single primer, you think it's a 09:36:10
 8 reasonable statement to believe that in 09:36:13
 9 that situation, the primer designed by a 09:36:16
 10 computer program with a reference genome 09:36:18
 11 would bind off target to the fragment 09:36:20
 12 50 percent of the time? 09:36:24
 13 MS. CIPRIANO: Objection, 09:36:25
 14 asked and answered. 09:36:26
 15 A. I believe I've answered that 09:36:29
 16 question. 09:36:30
 17 Q. Answer it again. 09:36:30
 18 A. It depends on the conditions. 09:36:31
 19 Q. So sometimes it might bind 09:36:34
 20 off target more than 50 percent of the 09:36:35
 21 time, sometimes it might bind 90 percent 09:36:37
 22 of the time? 09:36:40
 23 MS. CIPRIANO: Objection, 09:36:41
 24 form. 09:36:41
 25 A. Could you simplify the 09:36:44
 Page 65

Ex. 1120 at 65

1 CARLOS D. BUSTAMANTE, Ph.D.
 2 question for me? 09:36:45
 3 Q. Well, you said, "it depends." 09:36:45
 4 Sometimes it would bind on target 09:36:47
 5 90 percent of the time, sometimes it 09:36:49
 6 would bind 50 percent of the time; is 09:36:51
 7 that what you're saying? 09:36:53
 8 A. That's not what I said. I 09:36:54
 9 said, "It depends on the conditions." 09:36:55
 10 It depends on the genome. 09:36:56
 11 It depends on the gene region. 09:36:58
 12 It depends on how many actual copies 09:37:00
 13 there are in the genome, not how many 09:37:03
 14 copies are annotated in the reference 09:37:04
 15 genome. 09:37:07
 16 There are some regions that are 09:37:07
 17 very hard to sequence. There are other 09:37:08
 18 regions that aren't so hard to sequence. 09:37:12
 19 Some of the time you'll get perfectly 09:37:13
 20 good sequence from a single primer. 09:37:15
 21 Sometimes you won't. It depends on the 09:37:19
 22 conditions. 09:37:21
 23 And, again, today looking back at 09:37:22
 24 that, we know the technology works well 09:37:24
 25 and you can program it well. So putting 09:37:26
 Page 66

Ex. 1120 at 66.

So is solution-phase better than solid? Not necessarily. Given the low cost of sequencing, tight specificity may not be that important. As sequencing prices drop, “if half your reads map someplace else, or aren’t directly mappable to your region of interest, maybe you don’t care—you just throw them away,” says Zwick, who adds the microarray probes might even help coverage by capturing targeted sequences that are much longer than the probes themselves.

2. Shapero On-Target/Enrichment Levels

Bustamante on: Possible Shapero Off Target Causes

Q. Well other than primer dimers 16:14:01
and off-target hybridization of the 16:14:03
single specific target primer, what else 16:14:08
would be contributing to a lack of 16:14:11
specificity in the library generation, as 16:14:13
described by Shapero Exhibit 1006? 16:14:18

A. In general, those are the 16:14:31
features that generally characterize 16:14:33
artifacts that would be concerning for 16:14:40
next generation sequencing. There could 16:14:44
be additional artifacts generated through 16:14:47
more nonspecific binding and sort of more 16:14:52
random priming. Although, again, there's 16:14:58
no necessary evidence of that in this 16:15:00
paper given the high affinity in the 16:15:02
readout through the array hybridization 16:15:05

Bustamante (Ex. 1120) Re: Shapero Primer/Dimers

19 Q. So given, you know, the length 13:22:08
20 of the primers that are being used in the 13:22:10
21 250 flex, what are the possible size 13:22:14
22 ranges of primer dimers that could 13:22:16
23 possibly occur in the 250 plex? 13:22:18
24 A. We can turn back to the 13:22:21
25 beginning of the paper where they give us 13:22:22
Page 228

1 CARLOS D. BUSTAMANTE, Ph.D.
2 the sizes. 13:22:24
3 So, I believe, we're looking at 23 13:22:42
4 plus 25, so 48 base pairs. That would be 13:22:47
5 one primer. 13:22:52

6 Q. So, if primer dimers were 13:22:54
7 formed between the primers, given the 13:22:57
8 lengths that they have, what are the 13:23:00
9 possible size ranges? 13:23:02
10 A. We would expect them to 13:23:03
11 migrate at about a hundred base pairs, 13:23:05
12 right. They're below a hundred base 13:23:07
13 pairs. 13:23:08
14 Q. And in Figure 2(A), which is 13:23:09
15 a gel of run of all the amplicons, do you 13:23:11
16 see a shred of evidence that there are 13:23:14
17 any primer dimers in this? 13:23:16
18 A. I do not -- 13:23:18
19 MS. CIPRIANO: Objection to 13:23:19
20 form 13:23:20
21 A. I do not see a band. As I 13:23:20
22 said, I don't know where exactly they cut 13:23:22
23 off. It could be it's runoff. I don't 13:23:24
24 have reason to believe that. 13:23:30
25 But, again, I believe, the focus 13:23:31
Page 229

Ex. 1120 at 228-229.

9 Q. Well, there's a hundred base 13:23:50
10 pair marker in there, right? 13:23:52
11 A. And there's no band at a 13:23:52
12 hundred base pairs. 13:23:53
13 Q. Well, there is a hundred base 13:23:53
14 pair marker annotated in Figure 2(A), 13:23:56
15 right? 13:23:59
16 A. And it's not very bright. So 13:23:59
17 I don't know if they've run the gel off. 13:24:01
18 There is no hundred base pair band that I 13:24:04
19 see on the gel. 13:24:07
20 Q. So you think annotating a 13:24:11
21 hundred base pairs was just a phantom 13:24:13
22 thing that they just threw in there for 13:24:15
23 fun? 13:24:18
24 A. No, they could have -- again, 13:24:18
25 you're asking me about whether there is 13:24:19
Page 230

1 CARLOS D. BUSTAMANTE, Ph.D.
2 evidence in the gel for primer dimers. I 13:24:20
3 don't see evidence. But I also don't see 13:24:23
4 a band right at a hundred base pair and I 13:24:25
5 also believe in many ways it's not 13:24:28
6 relevant because the hybridization does 13:24:30
7 the work. 13:24:31

Ex. 1120 at 230-231.

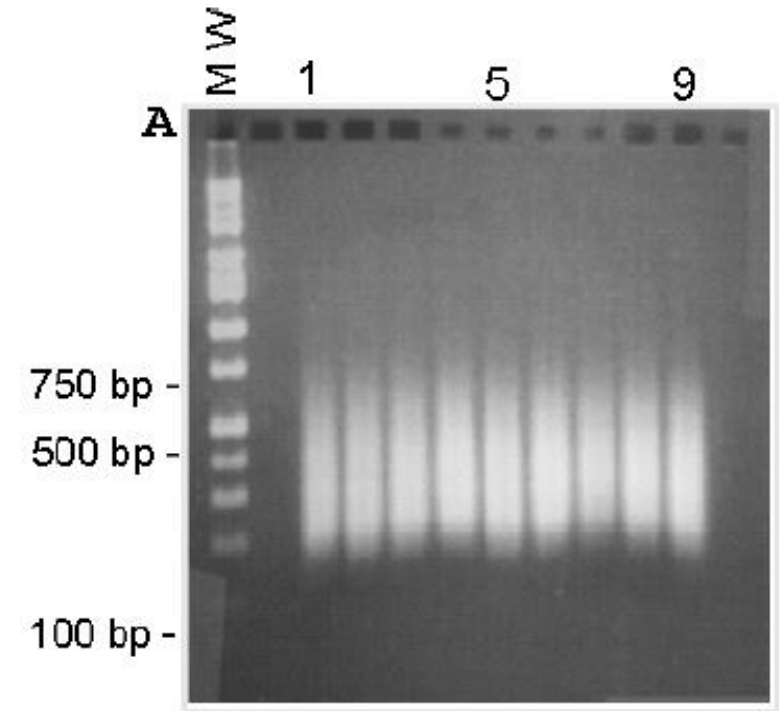


Figure 2. (A) 250-plex 2° PCR products separated on a 2%, 1×TBE agarose gel. (B) Negative control (no locus-specific primers) 2° PCR products separated on a 2%, 1×TBE agarose gel. (C) p502 chip image using 250 primer-plex 2° PCR products as target derived from input genomic DNA NA10830 (D) p502 chip image using negative control 2° PCR products as target derived from input genomic DNA NA10830.

Ex. 1006; see *a/so* Metzker Decl. Ex. 1121 ¶¶ 14-16

PO Sur-Reply Argument 1) Re: Primer/Dimer

Second, to the extent that any primer dimers remained in the solution, the data as presented did not capture them. EX1120.0228:7-15. As Dr. Bustamante explained, the 100 bp ladder band (*i.e.*, the control) was entirely missing from the gel. EX1120.0230:5-19. While the author's annotation indicates where the band was expected to be, the gel does not show the 100 bp control. The missing control suggests that there was an issue with the portion of the gel where primer dimers would have been expected, and Dr. Metzker concedes this possibility. EX1120.0229:21-230:8; EX2213.0045:12-20.

PO Sur-Reply at 12

PO's Cite To Dr. Metzker Testimony Re "Issue"

9 Q. Yeah, so the data that you cite to in that
10 paragraph is the gel read in Figure 2A of the
11 publication. Correct?

12 A. Well, paragraph 14 is really responding to
13 the gel in Figure 2A in particular reference
14 to the lack of primer dimers in the gel, and
15 in refuting Dr. Bustamante's suggestion that
16 maybe the reason why he couldn't see primer
17 dimers in Figure 2A is because maybe they
18 just ran off the gel, which is kind of -- I
19 mean, it's possible, but not the way that the
20 data are presented.

21 Because the data are presented where
22 there are molecular weight markers, and I

would have expected -- and I say this in
paragraph 14 -- that if there were primer
dimers, they would be slightly under 100 base
pairs because each primer is almost 48 base
pairs.

And I've run lots of PCRs and lots
of gels. Primer dimers are very prominent
when they're in a library, and what I'm
speaking to here is that Figure 2A does not
show any evidence whatsoever of primer dimers
being created from the 250-plex MARA
reaction.

Q. Thank you, Dr. Metzker.

So just to confirm the answer to my
question, the gel read is the only data that
you cite in that paragraph. Right?

A. With respect to the absence of primer dimers.
That's -- that's what I'm talking -- there's
other information in Figure 2A that also
speaks to on-target specificity. For
example, the size range of the expected
extension products are exactly what you --

PO Sur-Reply Argument 2) Re: Primer/Dimers

First, before the reads were run on the gel, Shapero used an exonuclease digestion step to remove products other than the full-length extension products. EX1006.0004. As a result, primer dimers and other partial sequences would have been removed from the mixture, reducing the chance that any small fragments would be seen on a gel.

Dr. Bustamante Re: Off Target Hybridization

10 Q. Well other than primer dimers 16:14:01
 11 and off-target hybridization of the 16:14:03
 12 single specific target primer, what else 16:14:08
 13 would be contributing to a lack of 16:14:11
 14 specificity in the library generation, as 16:14:13
 15 described by Shapero Exhibit 1006? 16:14:18
 16 A. In general, those are the 16:14:31
 17 features that generally characterize 16:14:33
 18 artifacts that would be concerning for 16:14:40
 19 next generation sequencing. There could 16:14:44
 20 be additional artifacts generated through 16:14:47
 21 more nonspecific binding and sort of more 16:14:52
 22 random priming. Although, again, there's 16:14:58
 23 no necessary evidence of that in this 16:15:00
 24 paper given the high affinity in the 16:15:02
 25 readout through the array hybridization 16:15:05

Ex. 1120 at 350.

are shown in Figure 2A. Each of the nine genomic DNA samples used in MARA show a reproducible smear from 150 to 550 bp when the 250 locus-specific primers are present. This result is consistent with the expected size distribution of products which ranges from 100 to 500 bp, with a mean size of 287 bp. As shown in Figure 2B, the yield of secondary amp-

Ex. 1006 at pg. 4 of 9.

Cont'd Dr. Bustamante Ex. 1120 Re Off Target

Q. Well, they actually did an analysis of what they should expect size-wise and this gel that they ran shows that their PCR -- the PCR results that they got were, actually, consistent with the amplicon sizes that they would expect, which is also consistent with their primers priming on target, correct?

A. It's absolutely consistent with their primers priming on target and could also be consistent with primers priming off target and giving a size distribution in the same range of 150 to 550 base pairs, which would be expected, given the restriction digest that's being done. And if there were homology, you simply replicate those regions.

Ex. 1120 at 214.

13 No. 2, what I said is what they're
14 reporting here is a size distribution
15 consistent with what they expected giving
16 a wide range of product lengths between
17 150 to 500 base pairs consistent with
18 on-target priming.
19 I'm happy to agree with the experts
20 that this is strong evidence of on-target
21 priming. If you had off-target priming,
22 you couldn't rule that out by looking at
23 the size distribution, because those
24 off-target products could well be within
25 this range that you're going to hybridize

Ex. 1120 at 215.

3. POSA's Alleged Disregard of SPE

Siebert Ex. 1044/1101.2 & Ex. 1121.342-343

We then mapped the PCR products by digestion with *Bam*HI and *Pvu*II. The restriction enzyme pattern obtained is shown in Figure 3. The release of common DNA fragments at 1.6 kb with *Bam*HI and 1.4 kb with *Pvu*II from the larger of the two PCR products was predicted from the known sequence. These data thus verified the gene-specificity of the walk for all five adaptor-ligated DNA libraries. The results also show that provisional restriction maps of the unknown sequence can be obtained without cloning.

We have found that the high specificity of the method described is primarily due to the suppression effect of the adaptor/adaptor primer combination, although in some cases additionally having an amine group on the lower strand of the adaptor (vectorette blocking) yielded better results.

13 Q. And do you disagree with 16:04:56
14 Siebert's conclusion, which the editors 16:04:57
15 of Nucleic Acid Research allowed to be 16:05:01
16 published, which says that the method and 16:05:05
17 the data that is supplied here shows that 16:05:08
18 the methodology had high specificity in 16:05:11
19 enriching or walking? 16:05:16
20 A. Remember, this is a paper that 16:05:18
21 has to do with genome walking where you're 16:05:21
22 trying to extend into an unknown region 16:05:23
23 from a known region. It's an exploratory 16:05:25
24 technique and there's really one product 16:05:29
25 that is being generated versus, as we've 16:05:34
2 discussed, application in more of the 16:05:40
3 multiplex setting. 16:05:45
4 So, in the context as described 16:05:46
5 here for genome walking, this was 16:05:47
6 accurately described in the paper. 16:05:53
7 Q. So, when it says, "high 16:05:54
8 specificity," it's walking in the region 16:05:57
9 they want it to walk in and not walking 16:05:59
10 at high rates in other regions that they 16:06:02
11 don't want it to walk in, correct? 16:06:05
12 A. For the example that they 16:06:07
13 describe and the particular gene that 16:06:09
14 they used to demonstrate the utility. 16:06:11



Miura Ex. 2047 & Dr. Bustamante Ex. 1120

ABSTRACT

Motivation: In contrast with conventional PCR using a pair of specific primers, some applications utilize a single unique primer in combination with a common primer, thereby relying solely on the former for specificity. These applications include rapid amplification of cDNA ends (RACE), adaptor-tagged competitive PCR (ATAC-PCR), PCR-mediated genome walking and so forth. Since the primers designed by conventional methods often fail to work in these applications, an improved strategy is required, particularly, for a large-scale analysis.

Results: Based on the structure of 'off-target' products in the ATAC-PCR, we reasoned that the practical determinant of the specificity of primers may not be the uniqueness of entire sequence but that of the shortest 3'-end subsequence that exceeds a threshold of duplex stability. We termed such a subsequence as a 'specificity-determining subsequence' (SDSS) and developed a simple algorithm to predict the performance of the primer: the algorithm identifies the SDSS of each primer and examines its uniqueness in the target genome. The primers designed using this algorithm worked much better than those designed using a conventional method in both ATAC-PCR and 5'-RACE experiments. Thus, the algorithm will be generally useful for improving various PCR-based applications.

Availability: The source code of the program is available upon request from the authors or can be obtained from <http://itolab.cb.k.u-tokyo.ac.jp/GATC/>

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: ito@k.u-tokyo.ac.jp

Ex. 2047 at 1.

These two sets of primers were used for the 5'-RACE, and the products were resolved by gel electrophoresis (Fig. 4A). Since it is difficult to judge whether the 5'-RACE has worked solely from the electrophoretic patterns, we cloned the products, randomly picked 12 clones and sequenced them for identification. As summarized in Figure 4B, the primers designed using the SDSS algorithm worked strikingly better than those designed using the 3'-end 12mer specificity algorithm. More than 90% of the products obtained using the SDSS primers were derived from the bona fide targets (Fig. 4B). Even in the cases where no distinct bands were observed on the gel, bona fide targets with different ends were detected. These were presumably due to truncated reverse transcription. In contrast, the purity of the products obtained by using the other primer set was rather poor (~10%) (Fig. 4B). These results clearly demonstrate the efficiency of the SDSS algorithm in designing primers for RACE, particularly for less abundant transcripts such as the 10 ORFs examined above.

Ex. 2047 at 6.

Q. And Miura, actually, reports	16:22:42
that using the particular algorithm that	16:22:47
he has developed -- and this is 2005.	16:22:50
So, just a year after Shapero and,	16:22:54
at least, six years before the '012 and	16:22:57
'108 purported inventors filed their	16:23:01
application, he reports on designing	16:23:04
primers for single primer extension,	16:23:09
which are 90 percent specific, correct?	16:23:12
A. In certain scenarios.	16:23:16
Q. So, perusing the	16:23:18

Ex. 1120 at 357.

Disclosure of Invention

Technical Problem

[10] Therefore, the present inventors have developed a method capable of notably improving upon the above problems. A low amplification efficiency is caused by insufficient ligation between the restricted genomic DNA and cassette, and non-specific amplification is attributed to PCR reaction by only cassette primers in stead of a combination of the known region-specific primers and cassette primers. Thus, to amplify the desired region only, amplification by only cassette primers should be inhibited, and amplification by a combination of the known region-specific primers and cassette primers should only be performed. By the present inventors, the restricted genomic DNA fragment was treated with ddNTP, and ligated with a complementary cassette to increase the efficiency of ligation between the genome and cassette and to prevent non-specific amplification, thereby completing the present invention.

[63] Step 2) is a step of modifying the 3'-OH end of the restricted DNA with a nucleotide analogue. In the present invention, the 3'-OH end of genomic DNA is modified with ddNTP for the prevention of further polymerization, non-specific amplification can be remarkably reduced. For example, when the genomic DNA fragment having



'012 Identification of and Solutions For SPE Alleged "Problems"

4. References Allegedly Citing Shapero Negatively

Howell (2006) Ex. 2130

Instead of amplifying the entire genome, directed amplification of selected portions of the genome may be more appropriate for some applications. PCR is commonly used for amplification of a single target sequence or a few sequences in parallel. However, for high levels of parallelism, a large number of separate amplification reactions must be performed.

As briefly noted previously, the so-called adaptor ligation method uses restriction endonucleases to digest a genomic sample into fragments of a known distribution. Common adaptor oligonucleotides are then ligated to the ends of the generated fragments. A subsequent PCR amplification with a single pair of primers can be adjusted to preferentially amplify fragments within a certain range of sizes. This way, a subset of the genome is amplified and a reduction of sequence complexity is achieved.^{109,133} A variant of the method uses adaptors, one common primer, and a single specific primer per target sequence to achieve amplification of specific genomic sequences.^{134,135}

If restriction digestion is carried out with type II S restriction enzymes, different end sequences are produced for different fragments. This way, adaptor sequences can be selected to circularize a certain population of fragments. These fragments can then be amplified and the procedure repeated, allowing the selective amplification of a single fragment from a genomic DNA sample.¹³⁶ This method may also be multiplexed by using several adaptor sequences, but incompatibilities of end sequences limit the extent to which this can be performed.

None of the methods above achieves the ideal goal of amplifying a large set of arbitrary sequences without the production of undesired amplification products. In the recently described selector amplification method, restriction fragments are made single-stranded by heat denaturation, circularized by hybridization to specific selector probes, and ligated to a common oligonucleotide, thus incorporating a common sequence motif in the single-stranded circles formed from the selected fragments. The selector probe end sequences are 10 to 20 nt in length to achieve specificity. Furthermore, structure-specific cleavage, using the same mechanism as in the Invader assay, can be utilized to remove a 5' portion of the fragment prior to ligation, allowing the creation of circles of similar lengths. All circles can then be amplified by PCR using a single primer pair for multiplex amplification for hundreds to thousands of selected fragments of interest.¹³⁷

KNOWN VERSUS UNKNOWN VARIATIONS

The vast majority of techniques described here are applied to SNPs where sequence and mapping information are known. Other methods are more adept at looking for sequence variants where exact information is unavailable. For example, if a particular gene has been implicated in a given disease, scanning the exons and promoters for variation could be desirable from a diagnostic point of view.

A number of methods can speed the search for such unknown sequence variants. Single-strand conformation polymorphism (SSCP),¹³⁸ denaturing gradient gel electrophoresis (DGGE),¹³⁹ and denaturing high performance liquid chromatography (dHPLC)¹⁴⁰ are examples of such methods. Although these three techniques can identify DNA fragments bearing sequence differences, the position or nature of genetic change is difficult to characterize based on results of these methods alone. Therefore, these methods are combined with or in fact replaced by sequencing techniques to derive exact sequence information.

One further technological development that could drastically affect the way we approach analysis of the genome is the emergence of powerful and economical sequencing techniques. Rather than looking at single base positions, the strategy would involve decoding the entire genome or extensive selected regions thereof. Complete information of the genetic sequence of an individual would thus become available, obviating the need for genotyping specific single base positions. Research activities aimed at achieving "\$1000" genome sequencing¹⁵² are under way, making cataloguing entire genome sequence information in populations a foreseeable possibility. In the meantime, many of the technologies described here will certainly assist in harvesting the fruits of the human genome initiative and play key roles in biomedical research for years to come.

Methodology

Open Access

Simple allele-discriminating PCR for cost-effective and rapid genotyping and mapping

Minh Bui^{1,2} and Zhongchi Liu^{*1}

Abstract

Background: Single nucleotide polymorphisms (SNPs) are widely observed between individuals, ecotypes, and species, serving as an invaluable molecular marker for genetic, genomic, ecological and evolutionary studies. Although, a large number of SNP-discriminating methods are currently available, few are suited for low-throughput and low-cost applications. Here, we describe a genotyping method named Simple Allele-discriminating PCR (SAP), which is ideally suited for the small-scale genotyping and gene mapping routinely performed in small to medium research or teaching laboratories.

Currently, a large variety of techniques for high-throughput SNP genotyping are available [7,8]. They can be grouped into four main classes: allele-specific hybridization, allele-specific nucleotide incorporation, allele-specific oligonucleotide ligation, and allele-specific invasive cleavage. For example, the TaqMan genotyping method [9] and the Amplifluor SNP HT genotyping System [10] are PCR-based, suitable for large scale high-throughput applications. Both methods, however, require expensive instrumentation and reagents such as synthetic oligonucleotides labeled with different fluorescent dyes. Various genome re-sequencing methods are also extremely powerful for large scale SNP-discrimination [11,12], yet are impractical for assaying a selected set of SNPs in specific genomic regions, and are usually beyond the reach of small to medium sized laboratories with limited resources.

Morin (Ex. 2135)

per completed genotype. As genotyping methods are constantly evolving, and different laboratories will choose different methods based on their throughput needs and available equipment, we expect the cost of genotyping to be lower in the future, but variable among laboratories. The method we have developed serves several purposes, minimizing costs while allowing for reasonably high-throughput genotyping. Additional modifications will be needed to further decrease costs, increase throughput to allow for efficient amplification of a larger number of loci in the multiplex PCR (e.g. Shapero *et al.* 2004), and facilitate higher throughput genotyping (e.g. using a multiplex SNP assay format).

5. Alleged Lack of Accuracy of NGS

Moorthie (Ex. 2169)

Sanger sequencing has a low (but non-zero) error rate of around 10^{-4} to 10^{-5} for single calls (one error per 1,000–10,000 bases), but the accuracy for detecting heterozygous variants is much more difficult to assess and is almost certainly context dependent to some extent. When it comes to detection of a low level variant, for example mosaic or somatic mutation, the limit of detection in terms of minor allele representation is only around 20% for Sanger sequencing. Current NGS platforms have a somewhat higher raw error rate of around 10^{-2} to 10^{-3} (depending on read length), but this is easily offset by increasing read depth (i.e. consensus accuracy). In fact the desired accuracy can essentially be determined by altering the read depth appropriately. In contrast to Sanger sequencing detection of low level variants is basically limited by the raw error rate and would be substantially below 0.1%.

Read depth, genome coverage and uniformity

The read depth or depth of coverage refers simply to the number of times a base is sequenced in a single run of the machine. The required read depth varies depending upon

Currently, the error rates of raw sequence data produced by all of the massively parallel sequencing platforms are higher than with Sanger sequencing, but the overall error rate is reduced because of the high degree of sequencing depth, typically 40-fold for a diploid genome, that is necessary to achieve complete coverage with massively parallel

6. Alleged Teachings To Avoid PCR

Bronner and Quail Ex. 2155.15,18, 19

BASIC PROTOCOL 4

PCR AMPLIFICATION OF THE LIBRARY

After the libraries have been shown to be adapter ligated we typically amplify the DNA with 8 to 18 cycles of PCR to increase the quantity of the library, enrich for fully ligated fragments, and tail fragments with the nucleotide sequences necessary for cluster amplification. Excessive PCR cycling should be avoided as it can lead to bias in the library and can result in unusual size profiles when libraries are analyzed, for example, on an Agilent Bioanalyzer chip. Over-amplification depletes the primer pool, resulting in an accumulation of non-extended single stranded template, which after denaturation anneals at

Ex. 2155 at 15.

ALTERNATE PROTOCOL 6

SEQUENCING WITHOUT PCR

PCR amplification during library prep allows lower amounts of starting material to be used, but introduces bias (Quail et al., 2012b). To eliminate this, we developed a PCR-free library method (Kozarewa et al., 2009). Here, full-length Illumina Y adapters are ligated directly to A-tailed fragments. Since these adapters contain the full complement of sequencing primer

Ex. 2155 at 19.

Note: Unnecessary PCR amplification steps will exacerbate amplification biases, and so they should be avoided wherever possible.

Ex. 2155 at 18.

Bronner and Quail Ex. 2155 – 454 Pyro MPS

This dominance is, however, quite a recent thing. In the last 5 years, we have seen the rise and fall of alternative sequencing technologies, some of which are still quite frequently used today. For instance in 2005, the first of the next-generation DNA sequencers, 454's GS20 (now Roche 454), became commercially available (Margulies et al., 2005). This method revolutionized the paradigm of DNA sequencing. Instead of a single sequencing reaction generating a single sequence, the 454 introduced massively parallel sequencing. The Roche 454 technology uses emulsion PCR to generate beads coated in amplicons that are derived from single template molecules. About a million of these beads are then sequenced in parallel by pyrosequencing (Ronaghi et al., 1998). Images of the beads are analyzed to generate high quality sequences. In this way, throughput is increased, cost reduced and

Arid Ex. 2042 – Avoiding Not Practical

Obviously, the best way to avoid bias during PCR is to avoid library amplification by PCR altogether [6,21]. On the other hand, PCR-free libraries require relatively large amounts of input DNA and are thus impractical for many sample types. Furthermore, there is no enrichment of sequenceable fragments carrying adapters on both ends, and the yield of such fragments is very sensitive to variations in DNA quality and purity, which in turn can affect the efficiency of end repair and adapter ligation. We also note that preparing PCR-free libraries alone

does not necessarily guarantee unbiased sequencing data as significant bias is introduced elsewhere in the process.

Importantly, PCR-free protocols are not readily amenable to automation and are therefore not the best choice as the default protocol in high-throughput sequencing facilities such as the Broad Institute Genome Sequencing Platform, where currently about 1,000 Illumina sequencing libraries are made per week to support a diverse portfolio of sequencing projects. Improved PCR conditions like the one described here will likely satisfy the vast majority of projects. Enhancing the coverage of high-GC loci is critical for human genome and exome sequencing in cancer and medical genetics, the major sequencing applications in terms of bases generated. Solving the loss of AT-rich loci remains a challenge, but has less of an impact on human genome sequencing and on the sequencing field as a whole. We expect that PCR-free methods, which are invaluable and superior for extreme base compositions at both ends of the %GC spectrum, will be reserved for projects that are most sensitive to base-composition bias and can supply input DNA of sufficient quality and quantity.

PO's New References

QIAGEN's EPO Response (Ex. 2211)

2.3 Documents D2-D4 all require a forward and a reverse primer to each target. None of the documents D1-D4 relates to single primer extension, and therefore none of them are actually closest prior art. The closest prior art would be a document relating to single primer extension, which is different from regular PCR requiring a target-specific primer pair, and to scaling single primer extension up to a multiplex reaction. There is no indication in any of the prior art cited that the addition of Mg^{2+} and MutS would be beneficial to single primer extension reactions, let alone to upscaling these to a multiplex reaction. Thus the technical problem to be solved over the prior art must be reformulated as providing an improved multiplex polynucleotide amplification method that uses a single target-specific primer per target region (*i.e.*, a highly multiplex single primer extension reaction).

Secondary Considerations



No Teaching Away or Skepticism

Brooks Declaration

14. Based on my experience in the field, I was skeptical that the claimed methods would provide efficient sequencing data. Because the claimed methods involve annealing and extending a target-specific primer to enrich a sequence of interest (rather than, for example, annealing a forward target-specific and reverse target-specific primer pair at a sequence of interest and performing a polymerase chain reaction with the forward target-specific and reverse target-specific primer), I thought the technology would provide poor target specificity that is not consistent with enrichment worthy of high quality sequencing data. I thought that the target-specific primers would bind to off-target sites at a relatively high rate, leading to non-specific enriched nucleic acid sequences that would flood the sequencing system and confound sequence analysis. I thought that in order to generate enriched sequences that would provide highly specific sequencing results, one would want to amplify a sequence of interest with, for example, a primer pair including a forward target-specific primer and reverse target-specific primer in a polymerase chain reaction using the forward target-specific primer and reverse target-specific primer for each sequence of interest.

No Teaching Away or Skepticism

Bustamante Transcript

What level of enrichment does the patent require that the library preparation methodology achieve?

A. There's no --

MS. CIPRIANO: Objection to form.

A. There's no specific level of enrichment that's outlined.

EX1120.160:9-16.

Bustamante Transcript

Q. In other words, even knowing going in that 95 percent of what you might put onto the Illumina sequencer was not going to be what you wanted, researchers did not run away from sequencing that on an Illumina sequencer, correct?

A. As a first attempt at a Neanderthal draft genome, that was the case.

EX1120.163:19-164:3.

Bustamante Transcript

Q. So is there a number in 2011 level of reads that you get that are on target that are -- that you would do and that you would not do in a massively parallel system?

A. It really depends on the biological question. So I would say that for some problems -- and if you're going to get plenty of depth of coverage, you can tolerate 50 percent of target. For others, 50 percent off target might not be sufficient to draw good conclusions until -- you'd have to do more sequencing in order to understand how well it's working.

EX1120 at 281:8-22.

Bustamante Transcript

Q. But the preparation of the libraries is no different than the preparation of the libraries as described in Shapero, right?

Prior to putting it on the array, prior to putting it on the Illumina sequencer, those steps are identical.

A. I would -- I would say that the MARA method, as described, is one example of the broader SPET, as described in this patent.

EX1120 at 265:7-17.

Nike v. Adidas

"Substantial evidence supports the Board's finding that other methods of minimizing waste—a problem undisputedly recognized by Nishida—had existed before the date of the invention."

955 F.3d 45, 55 (Fed. Cir. 2020).

In re Ethicon, Inc.

"[T]he Board's finding that the evidence relied on by Ethicon to support its claims of commercial success, industry praise, and unexpected results did not establish that any success, praise or unexpected results were due to the [claimed] 85:15 VDF:HFP coating, rather than to an unclaimed feature"

844 F.3d 1344, 1352 (Fed. Cir. 2017).

Industry Praise Directed to Library Preparation Technique and Multiplexing

As explained by Barchi, the claimed technology's "reliance on single primers, rather than on primer pairs, greatly simplifies panel design, and allows higher levels of multiplexing than PCR-based genotyping." *Id.*

POR at 60.

In a study performed by Italian and British research institutions, the researchers explained that "[u]nlike other methods, [the claimed] technique allows for the execution of cost-effective genotyping experiments while providing full control on target sites," which it attributed in part to the "high efficiency enrichment system" and "the scalability to tens of thousands of probes in a single reaction." EX2107.¶295; EX2068.0548. It reported that the study "results suggest that [the claimed] SPET genotyping is a very effective approach for genotyping a large number of loci," and "suggest[ed] that it might be a valuable tool for large-scale genotyping studies in virtually all species." EX2068.0550. .

POR at 61.

Industry Praise – No Multiplexing Requirement

Bustamante Transcript

Q. In fact, the claims of the '012 and '108 patent don't require multiplexing at all, do they?

A. There is -- none of the limitations, specifically, require multiplexing.

Ex.1120 at 297:25-298:3.

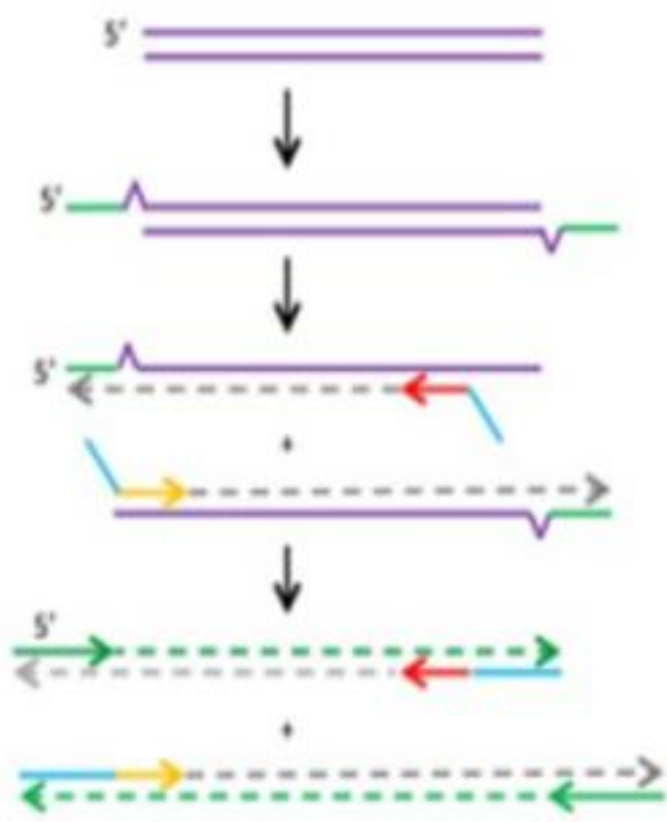
No Evidence of Copying

70. On information and belief, QIAGEN has been aware of Tecan Genomics' proprietary enrichment technology and families of the Asserted Patents since 2013. In December 2013, NuGEN and QIAGEN entered into discussions related to QIAGEN's potential acquisition of NuGEN. As part of these discussions, QIAGEN performed extensive due diligence. As a result of that formal due diligence, QIAGEN spoke to many senior-level employees at NuGEN and reviewed various documents. As part of this due diligence, QIAGEN received unfettered access to confidential information concerning NuGEN's research and development of technologies, its entire IP portfolio, and any other trade secrets that NuGEN possessed.

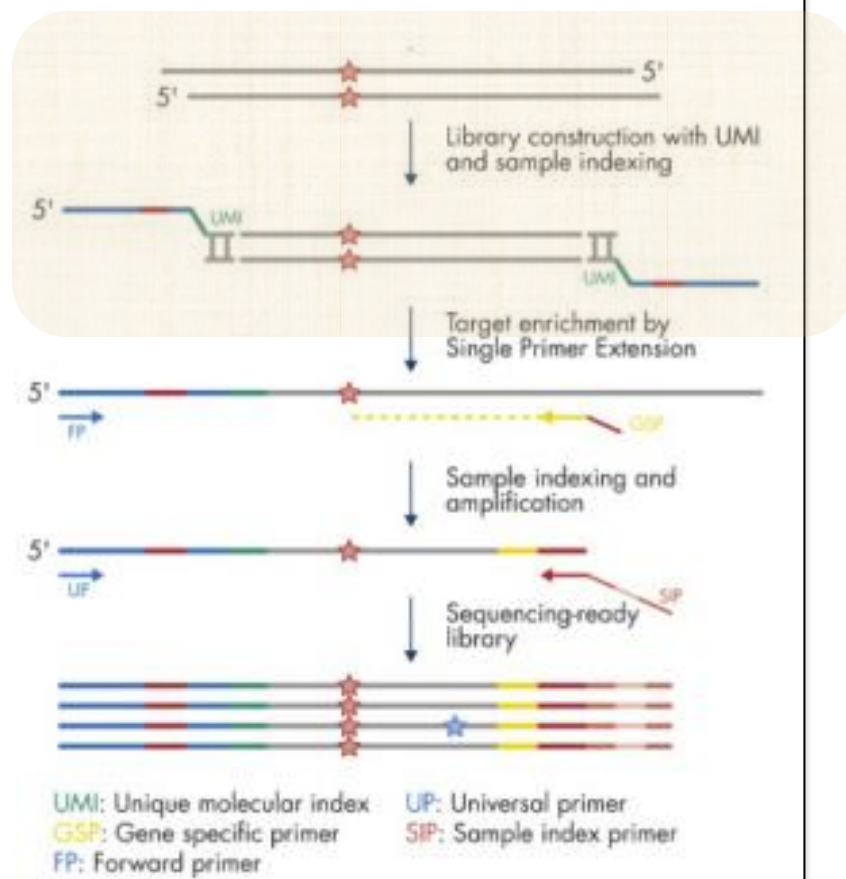
RESPONSE: QIAGEN admits that NuGEN and QIAGEN entered into discussions related to QIAGEN's potential acquisition of NuGEN. QIAGEN denies that the discussions between NuGEN related to QIAGEN's potential acquisition of NuGEN began in December 2013. QIAGEN admits that it performed some level of due diligence regarding acquisition of NuGEN. QIAGEN admits that QIAGEN's due diligence involved discussions with QIAGEN employees and reviewing various documents. QIAGEN denies that it received unfettered access to confidential information concerning NuGEN's research and development of technologies, its entire IP portfolio, and any other trade secrets that NuGEN possessed." The remaining allegations and characterizations are too vague and undefined in for example the claim of "families of the Asserted Patents" and "Tecan enrichment technology" for QIAGEN to understand what is being alleged and on that basis denies the allegations.

No Evidence of Copying

Ovation Target Enrichment



QIAseq Targeted DNA Panel



Bustamante Transcript,
Ex. 1120 at 390:25-402:9



QIAGEN's 2018 Marketing Statements on QIAseq

Application Note

QIAseq™ SPE technology for Illumina®: Redefining amplicon sequencing

Ex.2148 at 1.

Efficient coverage of both hotspots and entire CDS region

With SPE, you do not need to settle for hotspots. Instead, up to 20,000 primers can be pooled in an individual pool of primers, resulting in maximum coverage of around 500 genes, regardless of whether you are targeting hotspots, CDS, intronic regions, promoters, etc.

Ex.2148 at 4.

Uniform coverage by reducing the number of required primers and pools

SPE utilizes a reduced number of required primers resulting in highly uniform enrichment and sequencing due to a reduction in primer dropouts (Figure 5).

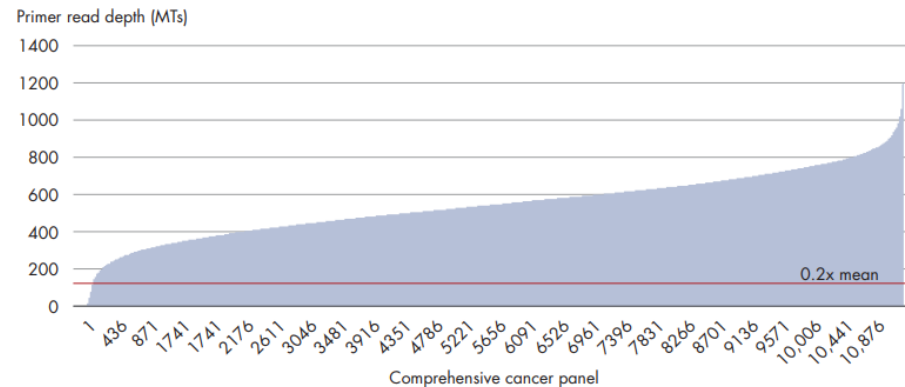


Figure 5. Highly uniform enrichment and sequencing. In this experiment, 6000 SNPs were enriched from 20 ng of NA12878 DNA. Library was constructed for sequencing on a MiSeq®, with 4,000,000 reads generated. The panel achieved a uniformity of 99.5% at 0.2x of mean coverage, and 96% at 0.5x of mean coverage.

Ex.2148 at 4.

Single primer extension (SPE) is a revolutionary approach that redefines amplicon-based enrichment and sequencing. With SPE, each genomic region is targeted by only one region-specific primer plus a universal adapter primer that binds to sequences introduced through library adapters (Figure 4).

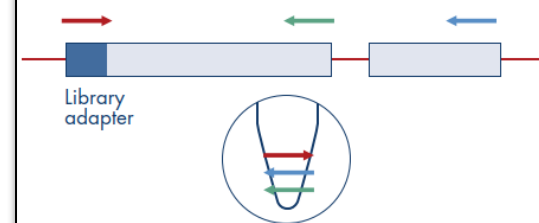


Figure 4. Primer design using single primer extension (SPE). Amplicon-based enrichment is performed in a single reaction regardless of the size of targeted regions.

Ex.2148 at 2.

QIAGEN 2015 EP Patent Specification (Ex. 2210)

[0031] the present inventors have identified a solution to increasing primer annealing extension efficiency while improving primer extension specificity via a combination of Mg^{2+} at a concentration higher than in a typical PCR reaction and a MutS protein. This combination is especially beneficial when a large number of very low concentrations of primers are used in single primer extension process to enrich a large genomics region for sequencing analysis.

Ex.2210 at 5.

SUMMARY

In one aspect, the present disclosure provides a method for performing a highly multiplex single primer extension reaction, comprising:

a. extending at least about 1,000 different primers in a single primer extension reaction using a plurality of target nucleic acids as templates in the presence of (1) a MutS protein, and (2) Mg^{2+} at a minimum concentration of about 6 mM to generate extension products.

Ex.2210 at 6.

Brooks Declaration

14. Based on my experience in the field, I was skeptical that the claimed methods would provide efficient sequencing data. Because the claimed methods involve annealing and extending a target-specific primer to enrich a sequence of interest (rather than, for example, annealing a forward target-specific and reverse target-specific primer pair at a sequence of interest and performing a polymerase chain reaction with the forward target-specific and reverse target-specific primer), I thought the technology would provide poor target specificity that is not consistent with enrichment worthy of high quality sequencing data. I thought that the target-specific primers would bind to off-target sites at a relatively high rate, leading to non-specific enriched nucleic acid sequences that would flood the sequencing system and confound sequence analysis. I thought that in order to generate enriched sequences that would provide highly specific sequencing results, one would want to amplify a sequence of interest with, for example, a primer pair including a forward target-specific primer and reverse target-specific primer in a polymerase chain reaction using the forward target-specific primer and reverse target-specific primer for each sequence of interest.

EX2026 at 37, ¶14.

Bustamante Transcript

Q. Alright. Let's go back to Paragraph 16 of the Brooks Declaration.

Why in 2012 with single primer extension would you have expected only 50 percent of the reads to be target sequenced that you were interested in?

A. I can only speculate as to why Dr. Brooks expected 50 percent.

Q. Do you agree with him?

A. I'm not sure I would agree to 50 percent.

EX1120 at 279:6-16.

End

