

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PROVISIONAL APPLICATION FOR PATENT COVER SHEET – Page 1 of 2

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

Express Mail Label No. _____

INVENTOR(S)		
Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Uri	Laserson	Boston, Massachusetts 02215
Greg	Porreca	Ocean City, New Jersey 08226
Additional inventors are being named on the _____ separately numbered sheets attached hereto.		
TITLE OF THE INVENTION (500 characters max):		
METHODS AND COMPOSITIONS FOR REDUCING DIAGNOSTIC ERRORS ASSOCIATED WITH BIAS IN PREPARATIONS OF NUCLEIC ACIDS		
Direct all correspondence to: CORRESPONDENCE ADDRESS		
☒ The address corresponding to Customer Number: 23628		
OR		
☐ Firm or Individual Name		
Address		
City	State	Zip
Country	Telephone	Email
ENCLOSED APPLICATION PARTS (check all that apply)		
☒ Application Data Sheet. See 37 CFR 1.76		
☐ CD(s), Number of CDs _____		
☒ Drawing(s) Number of Sheets <u>4</u>		
☐ Other (specify) _____		
☒ Specification (e.g. description of the invention) Number of Pages <u>36</u>		
Fees Due: Filing Fee of \$220 (\$110 for small entity). If the specification and drawings exceed 100 sheets of paper, an application size fee is also due, which is \$270 (\$135 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).		
METHOD OF PAYMENT OF THE FILING FEE AND APPLICATION SIZE FEE FOR THIS PROVISIONAL APPLICATION FOR PATENT		
☒ Applicant claims small entity status. See 37 CFR 1.27.		
☐ A check or money order made payable to the <i>Director of the United States Patent and Trademark Office</i> is enclosed to cover the filing fee and application size fee (if applicable).		110.00
☒ Payment by credit card. ~~Form PTO-2038 is attached.~~		TOTAL FEE AMOUNT (\$)
☒ The Director is hereby authorized to charge the filing fee and application size fee (if applicable) or credit any overpayment to Deposit		
Account Number: <u>23/2825</u>		

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

PROVISIONAL APPLICATION COVER SHEET
Page 2 of 2

PTO/SB/16 (12-08)

Approved for use through 06/30/2010. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.	
☒	No.
☐	Yes, the name of the U.S. Government agency and the Government contract number are: _____

WARNING:

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

SIGNATURE _____ /Patrick R.H. Waller/	Date _____ May 18, 2009
TYPED or PRINTED NAME _____ Patrick R.H. Waller	REGISTRATION NO. _____ 41,418 (if appropriate)
TELEPHONE _____ 617.646.8000	Docket Number: _____ G0814.70002US00

METHODS AND COMPOSITIONS FOR REDUCING DIAGNOSTIC ERRORS ASSOCIATED WITH BIAS IN PREPARATIONS OF NUCLEIC ACIDS

Field of the Invention

The invention relates to preparative steps in DNA sequencing-related technologies that reduce bias and increase the reliability and accuracy of downstream quantitative applications.

Background of the Invention

There are currently many genomics assays that utilize next-generation (polony-based) sequencing to generate data, including genome resequencing, RNA-seq for gene expression, bisulphite sequencing for methylation, and Immune-seq, among others. In order to make quantitative measurements (including genotype calling), these methods utilize the counts of sequencing reads of a given genomic locus as a proxy for the representation of that sequence in the original sample of nucleic acids. The majority of these techniques require a preparative step to construct a high-complexity library of DNA molecules that is representative of a sample of interest. This may include chemical or biochemical treatment of the DNA (e.g., bisulphite treatment), capture of a specific subset of the genome (e.g., padlock probe capture, solution hybridization), and a variety of amplification techniques (e.g., polymerase chain reaction, whole genome amplification, rolling circle amplification).

Systematic and random errors are common problems associated with genome amplification and sequencing library construction techniques. For example, genomic sequencing library may contain an over- or under-representation of particular sequences from a source genome as a result of errors (bias) in the library construction process. Such bias can be particularly problematic when it results in target sequences from a genome being absent or undetectable in the sequencing libraries. For example, an under-representation of particular allelic sequences (e.g., heterozygotic alleles) from a genome in a sequencing library can result in an apparent homozygous representation in a sequencing library. As most downstream sequencing library quantification techniques depend on stochastic counting processes, these problems have typically been addressed by sampling enough (over-sampling) to obtain a minimum number of observations necessary to make statistically significant decisions. However, the strategy of oversampling is generally limited to

elimination of low-count Poisson noise, and the approach wastes resources and increases the expense required to perform such experiments. Moreover, oversampling can result in a reduced statistical confidence in certain conclusions (e.g., diagnostic calls) based on the data. Accordingly, new approaches are needed for overcoming bias in sequencing library preparatory methods.

Brief Summary of the Invention

Aspects of the disclosure are based, in part, on the discovery of methods for overcoming problems associated with systematic and random errors (bias) in genome capture, amplification and sequencing methods, namely high variability in the capture and amplification of nucleic acids and disproportionate representation of heterozygous alleles in sequencing libraries. Accordingly, in some embodiments, the disclosure provides methods that reduce errors associated with the variability in the capture and amplification of nucleic acids. In other embodiments, the methods improve allelic representation in sequencing libraries and, thus, improve variant detection outcomes. In certain embodiments, the disclosure provides preparative methods for capturing target nucleic acids (e.g., genetic loci) that involve the use of differentiator tag sequences to uniquely tag individual nucleic acid molecules. In some embodiments, the differentiator tag sequence permit the detection of bias based on the occurrence of combinations of differentiator tag and target sequences observed in a sequencing reaction. In other embodiments, the methods reduce errors caused by bias, or the risk of bias, associated with the capture, amplification and sequencing of genetic loci, e.g., for diagnostic purposes.

Aspects of the invention relate to providing sequence tags (referred to as differentiator tags) that are useful to determine whether target nucleic acid sequences identified in an assay are from independently isolated target nucleic acids or from multiple copies of the same target nucleic acid molecule (e.g., due to bias in a preparative step, for example, amplification). This information can be used to help analyze a threshold number of independently isolated target nucleic acids from a biological sample in order to obtain sequence information that is reliable and can be used to make a genotype conclusion (e.g., call) with a desired degree of confidence. This information also can be used to detect bias in one or more nucleic acid preparative steps.

In some embodiments, a diagnosis, prognosis, or disease risk assessment is provided to a subject based on a genotype determined for that subject at one or more genetic loci (e.g., based on the analysis of a biological sample obtained from that subject). In some

embodiments, an assessment is provided to a couple, based on their respective genotypes at one or more genetic loci, of the risk of their having one or more children having a genotype associated with a disease or condition (e.g., a homozygous or heterozygous genotype associated with a disease or condition). In some embodiments, a subject or a couple may seek genetic or reproductive counseling in connection with a genotype determined according to embodiments of the invention. The methods disclosed herein are useful for any application where reduction of bias, e.g., associated with genomic isolation, amplification, sequencing, is important. For example, detection of cancer mutations in a heterogeneous tissue sample, detection of mutations in maternally-circulating fetal DNA, and detection of mutations in cells isolated during a preimplantation genetic diagnostic procedure.

Accordingly, in some aspects, methods of genotyping a subject are provided. In some embodiments, the methods comprise determining the sequence of at least a threshold number of independently isolated nucleic acids, wherein the sequence of each isolated nucleic acid comprises a target nucleic acid sequence and a differentiator tag sequence, wherein the threshold number is a number of unique combinations of target nucleic acid and differentiator tag sequences, wherein the isolated nucleic acids are identified as independently isolated if they comprise unique combinations of target nucleic acid and differentiator tag sequences, and wherein the target nucleic acid sequence is the sequence of a genomic locus of a subject.

In some embodiments, the isolated nucleic acids are products of a circularization selection-based preparative method, e.g., molecular inversion probe capture products. In other embodiments, the isolated nucleic acids are products of an amplification-based preparative methods. In other embodiments, the isolated nucleic acids are products of hybridization-based preparative methods.

Circularization selection-based preparative methods selectively convert regions of interest (target nucleic acids) into a covalently-closed circular molecule which is then isolated typically by removal (usually enzymatic, e.g. with exonuclease) of any non-circularized linear nucleic acid. Oligonucleotide probes (e.g., molecular inversion probes) are designed which have ends that flank the region of interest (target nucleic acid) and, optionally, primer sites, e.g., sequencing primer sites. The probes are allowed to hybridize to the genomic target, and enzymes are used to first (optionally) fill in any gap between probe ends and second ligate the probe closed. Following circularization, any remaining (non-target) linear nucleic acid is typically removed, resulting in isolation (capture) of target nucleic acid. Circularization selection-based preparative methods include molecular inversion probe

capture reactions and ‘selector’ capture reactions. In some embodiments, molecular inversion probe capture of a target nucleic acid is indicative of the presence of a polymorphism in the target nucleic acid.

In amplification-based (e.g., PCR-based or LCR-based, etc.) preparative methods, genomic loci (target nucleic acids) are isolated directly by means of a polymerase chain reaction or ligase chain reaction (or other amplification method) that selectively amplifies each locus using a pair of oligonucleotide primers. It is to be understood that primers will be sufficiently complementary to the target sequence to hybridize with and prime amplification of the target nucleic acid. Any one of a variety of art known methods may be utilized for primer design and synthesis. One or both of the primers may be perfectly complementary to the target sequence. Degenerate primers may also be used. Primers may also include additional nucleic acids that are not complementary to target sequences but that facilitate downstream applications, including for example restriction sites and differentiator tag sequences. Amplification-based methods include amplification of a single target nucleic acid and multiplex amplification (amplification of multiple target nucleic acids in parallel).

Hybridization-based preparative methods involve selectively immobilizing target nucleic acids for further manipulation. It is to be understood that one or more oligonucleotides (immobilization oligonucleotides), which comprise differentiator tag sequences, and which may be from 15 to 170 nucleotides in length, are used which hybridize along the length of a target region of a genetic locus to immobilize it. In some embodiments, immobilization oligonucleotides, are either immobilized before hybridization is performed (e.g., Roche/Nimblegen ‘sequence capture’), or are prepared such that they include a moiety (e.g. biotin) which can be used to selectively immobilize the target nucleic acid after hybridization by binding to e.g., streptavidin-coated microbeads (e.g. Agilent ‘SureSelect’).

In some embodiments, the methods disclosed herein comprise determining the sequence of molecular inversion probe capture products, each comprising a molecular inversion probe and a target nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence, and wherein the target nucleic acid is a captured genomic locus of a subject, and genotyping the subject at the captured genomic locus based on the sequence of at least a threshold number of unique combinations of target nucleic acid and differentiator tag sequences of molecular inversion probe capture products.

In some embodiments, the methods disclosed herein comprise obtaining molecular inversion probe capture products, each comprising a molecular inversion probe and a target

nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence, wherein the target nucleic acid is a captured genomic locus of the subject, amplifying the molecular inversion probe capture products, and genotyping the subject by determining, for each target nucleic acid, the sequence of at least a threshold number of unique combinations of target nucleic acid and differentiator tag sequence of molecular inversion probe capture products. In certain embodiments, obtaining comprises capturing target nucleic acids from a genomic sample of the subject with molecular inversion probes, each comprising a unique differentiator tag sequence. In specific embodiments, capturing is performed under conditions wherein the likelihood of obtaining two or more molecular inversion probe capture products with identical combinations of target and differentiator tag sequences is equal to or less than a predetermined value, optionally wherein the predetermined value is about 0.05.

In one embodiment, the threshold number for a specific target nucleic acid sequence is selected based on a desired statistical confidence for the genotype. In some embodiments, the methods further comprising determining a statistical confidence for the genotype based on the number of unique combinations of target nucleic acid and differentiator tag sequences.

According to some aspects, methods of analyzing a plurality of genetic loci are provided. In some embodiments, the methods comprise obtaining a plurality of molecular inversion probe capture products each comprising a molecular inversion probe and a target nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence, amplifying the plurality of molecular inversion probe capture products, determining numbers of occurrence of combinations of target nucleic acid and differentiator tag sequence of molecular inversion probe capture products in the amplified plurality, and if the number of occurrence of a specific combination of target nucleic acid sequence and differentiator tag sequence exceeds a predetermined value, detecting bias in the amplification of the molecular inversion probe comprising the specific combination. In some embodiments, the methods further comprise genotyping target sequences in the plurality, wherein the genotyping comprises correcting for bias, if detected.

In some embodiments, the target nucleic acid is a gene (or portion thereof) selected from Table 1. In some embodiments, the genotyping comprises determining the sequence of a target nucleic acid (e.g., a polymorphic sequence) at one or more (both) alleles of a genome (a diploid genome) of a subject. In certain embodiments, the genotyping comprises

determining the sequence of a target nucleic acid at both alleles of a diploid genome of a subject, wherein the target nucleic acid comprises, or consists of, a sequence of Table 2.

Brief Description of the Drawings

Figure 1a, b, and c depict various methods for combining differentiator tag sequence and target sequences (NNNN depicts a differentiator tag sequence).

Figure 2 depicts a method for genotyping based on target and differentiator tag sequences. Sequencing reads of target and differentiator tag sequences are collapsed to make diploid genotype calls.

Figure 3 depicts results of a simulation of a MIP capture reaction in which MIP probes, each having a differentiator tag sequence of 15 nucleotides, are combined with 10000 target sequence copies (e.g., genome equivalents). In this simulated reaction, the probability of capturing one or more copies of a target sequence having the same differentiator tag sequence is 0.05. The Y axis reflects the number of observations. The X axis reflects the number of independent occurrences of target:differentiator tag combinations.

Figure 4 depicts a graph of sequencing coverage, which is needed to ensure that alleles are sampled to sufficient depth (either 10x or 20x minimum sampling per allele, assuming 1000 targets). X axis is total per-target coverage required, and Y axis is the probability that a given total coverage will result in at least 10x or 20x coverage for each allele.

Detailed Description of the Invention

Aspects of the disclosure are based, in part, on the discovery of methods for overcoming problems associated with systematic and random errors (bias) in genome capture, amplification and sequencing methods, namely high variability in the capture and amplification of nucleic acids and disproportionate representation of heterozygous alleles in sequencing libraries. Accordingly, in some embodiments, the disclosure provides methods that reduce variability in the capture and amplification of nucleic acids. In other embodiments, the methods improve allelic representation in sequencing libraries and, thus, improve variant detection outcomes. In certain embodiments, the disclosure provides preparative methods for capturing target nucleic acids (e.g., genetic loci) that involve the use of differentiator tag sequences to uniquely tag individual nucleic acid molecules. In some

embodiments, the differentiator tag sequence permits the detection of bias based on the frequency with which pairs of differentiator tag and target sequences are observed in a sequencing reaction. In other embodiments, the methods reduce errors caused by bias, or the risk of bias, associated with the capture, amplification and sequencing of genetic loci, e.g., for diagnostic purposes.

Aspects of the invention relate to associating unique sequence tags (referred to as differentiator tag sequences) with individual target molecules that are independently captured and/or analyzed (e.g., prior to amplification or other process that may introduce bias). These tags are useful to distinguish independent target molecules from each other thereby allowing an analysis to be based on a known number of individual target molecules. For example, if each of a plurality of target molecule sequences obtained in an assay is associated with a different differentiator tag, then the target sequences can be considered to be independent of each other and a genotype likelihood can be determined based on this information. In contrast, if each of the plurality of target molecule sequences obtained in the assay is associated with the same differentiator tag, then they probably all originated from the same target molecule due to over-representation (e.g., due to biased amplification) of this target molecule in the assay. This provides less information than the situation where each nucleic acid was associated with a different differentiator tag. In some embodiments, a threshold number of independently isolated molecules (e.g., unique combinations of differentiator tag and target sequences) is analyzed to determine the genotype of a subject.

According to some aspects, the disclosure relates to multiplex diagnostic methods. In some embodiments, the multiplex diagnostic methods comprise capturing a plurality of target nucleic acids (genetic loci) in parallel (e.g., a genetic locus of Table 1) using capture probes having unique differentiator tag sequences. In some embodiments, the genetic loci possess one or more polymorphisms (e.g., a polymorphism of Table 2) the genotypes of which correspond to disease causing alleles. Accordingly, in some embodiments, the disclosure provides methods for assessing multiple heritable disorders in parallel. In some embodiments, methods are provided for diagnosing multiple heritable disorders in parallel at a pre-implantation, prenatal, perinatal, or postnatal stage. In some embodiments, the disclosure provides methods for analyzing multiple genetic loci (e.g., a plurality of target nucleic acids selected from Table 1) from a patient sample, such as a blood, pre-implantation embryo, chorionic villus or amniotic fluid sample. A patient or subject may be a human. However, aspects of the invention are not limited to humans

and may be applied to other species (e.g., mammals, birds, reptiles, other vertebrates or invertebrates) as aspects of the invention are not limited in this respect. A subject or patient may be male or female. In some embodiments, in connection with reproductive genetic counseling, samples from a male and female member of a couple may be analyzed. In some embodiments, for example, in connection with an animal breeding program, samples from a plurality of male and female subjects may be analyzed to determine compatible or optimal breeding partners or strategies for particular traits or to avoid one or more diseases or conditions.

In some embodiments, the invention relates to compositions comprising pools (libraries) of preparative nucleic acids that each comprise “differentiator tag sequences” for detecting and reducing the effects of bias, and for genotyping target nucleic acid sequences. As used herein, a “differentiator tag sequence” is a sequence of a nucleic acid (a preparative nucleic acid), which in the context of a plurality different isolated nucleic acids, identifies a unique, independently isolated nucleic acid. Typically, differentiator tag sequences are used to identify the origin of a target nucleic acid at one or more stages of a nucleic acid preparative method. For example, in the context of a multiplex nucleic acid capture reaction, differentiator tag sequences provide a basis for differentiating between multiple independent, target nucleic acid capture events. Also, in the context of a multiplex nucleic acid amplification reaction, differentiator tag sequences provide a basis for differentiating between multiple independent, primary amplicons of a target nucleic acid, for example. Thus, combinations of target nucleic acid and differentiator tag sequence (target:diffentiator tag sequences) of an isolated nucleic acid of a preparative method provide a basis for identifying unique, independently isolated target nucleic acids. Differentiator tags may be synthesized by random nucleotide addition, for example. Differentiator tag sequences are typically of a predefined length, which is selected to control the likelihood of producing unique target:diffentiator tag sequences in a preparative reaction (e.g., amplification-based reaction, a circularization selection-based reaction, e.g., a MIP reaction). Differentiator tag sequences may be, up to 5, up to 6, up to 7 up to 8, up to 9, up to 10, up to 11, up to 12, up to 13, up to 14, up to 15, up to 16, up to 17, up to 18, up to 19, up to 20, up to 21, up to 22, up to 23, up to 24, up to 25, or more nucleotides in length. For purposes of genotyping, isolated nucleic acids are identified as independently isolated if they comprise unique combinations of target nucleic acid and differentiator tag sequences, and observance of threshold numbers of unique

combinations of target nucleic acid and differentiator tag sequences provide a certain statistical confidence in the genotype.

During a library preparation process, each nucleic acid molecule may be tagged with a unique differentiator tag sequence in a configuration that permits the differentiator tag sequence to be sequenced along with the target nucleic acid sequence of interest (the nucleic acid sequence for which the library is being prepared, e.g., a polymorphic sequence). The incorporation of the nucleic acid comprising a differentiator tag sequence at a particular step allows the detection and correction of biases in subsequent steps of the protocol.

A large library of unique differentiator tag sequences may be created by using degenerate, random-sequence polynucleotides of defined length. The differentiator tag sequences of the polynucleotides may be read at the final stage of the sequencing. The observations of the differentiator tag sequences may be used to detect and correct biases in the final sequencing read-out of the library. For example, the total possible number of differentiator tag sequences, which may be produced, e.g., randomly, is 4^N , where N is the length of the differentiator tag sequence. Thus, it is to be understood that the length of the differentiator tag sequence may be adjusted such that the size of the population of MIPs having unique differentiator tag sequences is sufficient to produce a library of MIP capture products in which identical independent combinations of target nucleic acid and differentiator tag sequence are rare. As used herein combinations of target nucleic acid and differentiator tag sequences, may also be referred to as “target:differentiator tag sequences”.

In the final readout of a sequencing process, each read may have an additional unique differentiator tag sequence. In some embodiments, when differentiator tag sequences are distributed randomly in a library, all the unique differentiator tag sequences will be observed about an equal number of times. Accordingly, the number of occurrences of a differentiator tag sequence may follow a Poisson distribution.

In some embodiments, overrepresentation of target:differentiator tag sequences in a pool of preparative nucleic acids (e.g., amplified MIP capture products) is indicative of bias in the preparative process (e.g., bias in the amplification process). For example, target:differentiator tag sequence combinations that are statistically overrepresented are indicative of bias in the protocol at one or more steps between the incorporation of the differentiator tag sequences into MIPs and the actual sequencing of the MIP capture products.

The number of reads of a given target:differentiator tag sequence may be indicative (may serve as a proxy) of the amount of that target sequence present in the originating sample. In some embodiments, the numbers of occurrence of sequences in the originating

sample is the quantity of interest. For example, using the methods disclosed herein, the occurrence of differentiator tag sequences in a pool of MIPs may be predetermined (e.g., may be the same for all differentiator tag sequences). Accordingly, changes in the occurrence of differentiator tag sequences after amplification and sequencing may be indicative of bias in the protocol. Bias may be corrected to provide an accurate representation of the composition of the original MIP pool, e.g., for diagnostic purposes.

According to some aspects, a library of preparative nucleic acid molecules (e.g., MIPs, each nucleic acid in the library having a unique differentiator tag sequence, may be constructed such that the number of nucleic acid molecules in the library is significantly larger than the number prospective target nucleic acid molecules to be captured using the library. This ensures that products of the preparative methods include only unique target:differentiator tag sequence; e.g., in a MIP reaction the capture step would undersample the total population of unique differentiator tag sequences in the MIP library. For example, an experiment utilizing 1 µg of genomic DNA will contain about ~150,000 copies of a diploid genome. For a MIP library, each MIP in the library comprising a randomly produced 12-mer differentiator tag sequence (~1.6 million possible unique differentiator tag sequences), there would be more than 100 unique differentiator tag sequences per genomic copy. For a MIP library, each MIP in the library comprising a randomly produced 15-mer differentiator tag sequence (~1 billion possible unique differentiator tag sequences), there would be more than 7000 unique differentiator tag sequences per genomic copy. Therefore, the probability of the same differentiator tag sequence being incorporated multiple times is incredibly small. Thus, it is to be appreciated that the length of the differentiator tag sequence is to be selected based on the amount of target sequence in a MIP capture reaction and the desired probability for having multiple, independent occurrences of target:differentiator tag sequence combinations.

The skilled artisan will appreciate that as part of a MIP library preparation process, adapters may be ligated onto the ends of the molecules of interest. Adapters often contain PCR primer sites (for amplification or emulsion PCR) and/or sequencing primer sites. In addition, barcodes may be included, for example, to uniquely identify individual samples (e.g., patient samples) that may be mixed together. (See, e.g., USPTO Publication Number US 2007/0020640 A1 (McCloskey et al.)

The actual incorporation of the random differentiator tag sequences can be performed through various methods known in the art. For example, nucleic acids comprising

differentiator tag sequences may be incorporated by ligation. This is a flexible method, because molecules having differentiator tag sequence can be ligated to any blunt-ended nucleic acids. The sequencing primers must be incorporated subsequently such that they sequence both the differentiator tag sequence and the target sequence. Alternatively, the sequencing adaptors can be synthesized with the random differentiator tag sequences at their 3' end (as degenerate bases), so that only one ligation must be performed. Another method is to incorporate the differentiator tag sequence into a PCR primer, such that the primer structure is arranged with the common adaptor sequence followed by the random differentiator tag sequence followed by the PCR priming sequence (in 5' to 3' order). A differentiator tag sequence and adaptor sequence (which may contain the sequencing primer site) are incorporated as tags. Another method to incorporate the differentiator tag sequences is to synthesize them into a padlock probe prior to performing a gene capture reaction. The differentiator tag sequence is incorporated 3' to the targeting arm but 5' to the amplification primer that will be used downstream in the protocol. Another method to incorporate the differentiator tag sequences is as a tag on a gene-specific or poly-dT reverse-transcription primer. This allows the differentiator tag sequence to be incorporated directly at the cDNA level.

In some embodiments, at the incorporation step, the distribution of differentiator tag sequences can be assumed to be uniform. In this case, bias in any part of the protocol would change the uniformity of this distribution, which can be observed after sequencing. This allows the differentiator tag sequence to be used in any preparative process where the ultimate output is sequencing of many molecules in parallel.

Differentiator tag sequences may be incorporated into probes (e.g., MIPs) of a plurality when they are synthesized on-chip in parallel, such that degeneracy of the incorporated nucleotides is sufficient to ensure near-uniform distribution in the plurality of probes. It is to be appreciated that amplification of a pool of unique differentiator tag sequences may itself introduce bias in the initial pool. However, in most practical cases, the scale of synthesis (e.g., by column synthesis, chip based synthesis, etc.) is large enough that amplification of an initial pool of differentiator tag sequences is not necessary. By avoiding amplification or selection steps on the pool of unique differentiator tag sequences, potential bias may be minimized.

One example of the use of the differentiator tag sequences is in genome re-sequencing. Considering that the raw accuracy of most next-generation sequencing instruments is relatively low, it is crucial to oversample the genomic loci of interest.

Furthermore, since there are two alleles at every locus, it is important to sample enough to ensure that both alleles have been observed a sufficient number of times to determine with a sufficient degree of statistical confidence whether the sample is homozygous or heterozygous. Indeed, the sequencing is performed to sample the composition of molecules in the originating sample. However, after multiple reads have been collected for a given locus, it is possible that due to bias (e.g., caused by PCR amplification steps), a large fraction of the reads are derived from a single originating molecule. This would skew the population of target sequences observed, and would affect the outcome of the genotype call. For example, it is possible that a locus that is heterozygous is called as homozygous, because there are only a few observations of the second allele out of many observations of that locus. However, if information is available on differentiator tag sequences, this situation could be averted, because the over-represented allele would be seen to also have an over-represented differentiator tag sequence (i.e., the sequences with the overrepresented differentiator tag sequence all originated from the same single molecule). Therefore, the sequences and corresponding distribution of differentiator tag sequences can be used as an additional input to the genotype-calling algorithm to significantly improve the accuracy and confidence of the genotype calls.

In some aspects, the disclosure provides methods for analyzing a plurality of target sequences which are genetic loci or portions of genetic loci (e.g., a genetic locus of Table 1). The genetic loci may be analyzed by sequencing to obtain a genotype at one or more polymorphisms (e.g., SNPs). Exemplary polymorphisms are disclosed in Table 2. The skilled artisan will appreciate that other polymorphisms are known in the art and may be identified, for example, by querying the Entrez Single Nucleotide Polymorphism database, for example, by searching with a GeneID from Table 1.

The mutations listed in Table 2 are documented polymorphisms in several disease-associated genes (CFTR is mutated in cystic fibrosis, GBA is mutated in Gaucher disease, ASPA is mutated in Canavan disease, HEXA is mutated in Tay Sachs disease). The polymorphisms are of several types: insertion/deletion polymorphisms which will cause frameshifts (and thus generally interrupt protein function) unless the insertion/deletion length is a multiple of 3bp, and substitutions which can alter the amino acid sequence of the protein and in some cases cause complete inactivation by introduction of a stop codon.

The methods disclosed here are also useful for diagnosis of one or more heritable disorders. In some embodiments, a heritable disorder that may be diagnosed with the methods disclosed herein is a genetic disorder that is prevalent in the Ashkenazi Jewish

population. In some embodiments, the heritable disorders are selected from: 21-Hydroxylase-Deficient Congenital Adrenal Hyperplasia; ABCC8-Related Hyperinsulinism; Alpha-Thalassemia, includes Constant Spring, & MR associated; Arylsulfatase A Deficiency-Metyachromatic Leukodystrophy; Biotinidase Deficiency-Holocarboxylase Synthetase Deficiency; Bloom's Syndrome; Canavan Disease; CFTR-Related Disorders-cystic fibrosis; Citrullinemia Type I; Combined MMA & Homocystinuria-dblC; Dystrophinopathies (DMD & BMD); Familial Dysautonomia; Fanconi Anemia-FANCC; Galactosemia-Classical: Galactokinase Deficiency & Galactose Epimerase Deficiency; Gaucher Disease; GJB2-Related DFNB 1 Nonsyndromic Hearing Loss and Deafness; Glutaric acidemia Type 1; Hemoglobinopathies beta-chain disorders; Glycogen Storage Disease Type 1A; Maple Syrup Urine Disease; Types 1A, 1B, 2, 3; Medium Chain Acyl-Coenzyme A; Dehydrogenase Deficiency-MCADD; Methylmalonic Acidemia; Mucopolidosis IV; Nemaline Myopathy; Nieman-Pick Type A-Acid Sphingomyelinase Deficiency; Non-Ketotic Hyperglycinemia-Glycine Encephalopathy; Ornithine Transcarbamylase Deficiency; PKU Phenylalanine Hydroxylase Deficiency; Propionic Acidemia; Short Chain Acyl-CoA Dehydrogenase Deficiency-SCADD; Smith-Lemli-Opitz Syndrome; Spinal Muscular Atrophy (SMN1)-SMA; Tay Sachs-HexA Deficiency; Usher Syndrome-Type I (Type IB, Type IC, Type ID, Type IF, Type IG) ; X-Linked Mental Retardation ARX-Related Disorders; X-Linked Mental Retardation with Cerebellar Cypoplasia and distinctive Facial Appearance; X-Linked Mental Retardation; includes 9, 21, 30, 46, 58, 63, 88, 89; X-linked mental retardation: FM1-Related Disorders-FRXA, Fragile X MR; X-linked SMR: Renpenning Syndrome 1; Zellweger Spectrum disorders - Peroxisomal Bifunctional Enzyme Deficiencies including Zellweger, NALD, infantile Refsums.

Table 1: Target Nucleic Acids

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
CYP21A2	1589	cytochrome P450, family 21, subfamily A, polypeptide 2	CAH1; CPS1; CA21H; CYP21; CYP21B; P450c21B; MGC150536; MGC150537; CYP21A2	201910	NG_008337.1	6p21.3
ABCC8	6833	ATP-binding cassette, sub-family C (CFTR/MRP), member	HI; SUR; HHHF1; MRP8; PHHI; SUR1; ABC36;	600509	NG_008867.1	11p15.1

Gene name	Gene ID	Description	Gene aliases	OMI M	RefSeqGene	Chromosome map position
		8	HRINS; TNDM2; ABCC8			
ATRX	546	alpha thalassemia/mental retardation syndrome X-linked (RAD54 homolog, <i>S. cerevisiae</i>)	SHS; XH2; XNP; ATR2; SFM1; RAD54; MRXHF1; RAD54L; ZNF-HX; MGC2094; ATRX	300032	NG_008838.1	Xq13.1-q21.1
ARSA	410	arylsulfatase A	MLD; ARSA	607574	NG_009260.1	22q13.31-qter; 22q13.33
PSAP	5660	prosaposin	GLBA; SAPI; FLJ00245; MGC110993; PSAP	176801	NG_008835.1	10q21-q22
BTD	686	biotinidase	BTD	609019	NG_008019.1	3p25
HLCS	3141	holocarboxylase synthetase (biotin-(propionyl-Coenzyme A-carboxylase (ATP-hydrolysing)) ligase)	HCS; HLCS	609018	NC_000021.7	21q22.1; 21q22.13
BLM	641	Bloom syndrome, RecQ helicase-like	BS; RECQ2; RECQL2; RECQL3; MGC126616; MGC131618; MGC131620; BLM	604610	NG_007272.1	15q26.1
ASPA	443	aspartoacylase (Canavan disease)	ASP; ACY2; ASPA	608034	NG_008399.1	17pter-p13
CFTR	1080	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)	CF; MRP7; ABC35; ABCC7; CFTR/MRP; TNR-CFTR; dJ760C5.1; CFTR	602421	NC_000007.12	7q31.2
ASS1	445	argininosuccinate synthetase 1	ASS; CTLN1; ASS1	603470	NG_011542.1	9q34.1
MMACHC	25974	methylmalonic aciduria (cobalamin deficiency) cblC type, with homocystinuria	cblC; FLJ25671; DKFZp564I122; RP11-291L19.3; MMACHC	609831	NC_000001.9	1p34.1
IKBKA P	8518	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein	FD; DYS; ELP1; IKAP; IKI3; TOT1; FLJ12497; DKFZp781H1425; IKBKAP	603722	NG_008788.1	9q31
FANCC	2176	Fanconi anemia, complementation group	FA3; FAC; FACC;	227645	NG_011707.1	9q22.3

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
		C	FLJ14675; FANCC			
GALK1	2584	galactokinase 1	GK1; GALK; GALK1	604313	NG_008079.1	17q24
GALT	2592	galactose-1-phosphate uridylyltransferase	GALT	606999	NC_000009.10	9p13
GALE	2582	UDP-galactose-4-epimerase	SDR1E1; FLJ95174; FLJ97302; GALE	606953	NG_007068.1	1p36-p35
GBA	2629	glucosidase, beta; acid (includes glucosylceramidase)	GCB; GBA1; GLUC; GBA	606463	NG_009783.1	1q21
GJB2	2706	gap junction protein, beta 2, 26kDa	HID; KID; PPK; CX26; DFNA3; DFNB1; NSRD1; DFNA3A; DFNB1A; GJB2	121011	NG_008358.1	13q11-q12
GCDH	2639	glutaryl-Coenzyme A dehydrogenase	GCD; ACAD5; GCDH	608801	NG_009292.1	19p13.2
G6PC	2538	glucose-6-phosphatase, catalytic subunit	G6PT; GSD1; GSD1a; MGC163350; G6PC	232200	NG_011808.1	17q21
HBB	3043	hemoglobin, beta	CD113t-C; beta-globin; HBB	141900	NG_000007.3	11p15.5
BCKDHA	593	branched chain keto acid dehydrogenase E1, alpha polypeptide	MSU; MSUD1; OVD1A; BCKDE1A; FLJ45695; BCKDHA	608348	NC_000019.8	19q13.1-q13.2
BCKDHB	594	branched chain keto acid dehydrogenase E1, beta polypeptide	E1B; FLJ17880; dJ279A18.1; BCKDHB	248611	NG_009775.1	6q13-q15
DBT	1629	dihydrolipoamide branched chain transacylase E2	E2; E2B; BCATE2; MGC9061; DBT	248610	NG_011852.1	1p31
DLD	1738	dihydrolipoamide dehydrogenase	E3; LAD; DLDH; GCSL; PHE3; DLD	238331	NG_008045.1	7q31-q32
ACADM	34	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	MCAD; ACAD1; MCADH; FLJ18227; FLJ93013; FLJ99884; ACADM	607008	NG_007045.1	1p31
MMAA	166785	methylmalonic aciduria (cobalamin deficiency) cblA type	cblA; MGC120010; MGC120011; MGC120012; MGC120013; MMAA	607481	NG_007536.1	4q31.22

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
MMAB	326625	methylmalonic aciduria (cobalamin deficiency) cblB type	ATR; cblB; MGC20496; MMAB	607568	NG_007096.1	12q24
MUT	4594	methylmalonyl Coenzyme A mutase	MCM; MUT	609058	NG_007100.1	6p12.3
MCOLN1	57192	mucolipin 1	ML4; MLIV; MST080; TRPML1; MSTP080; TRP-ML1; TRPM-L1; MCOLN1	605248	NC_000019.8	19p13.3-p13.2
ACTA1	58	actin, alpha 1, skeletal muscle	ACTA; ASMA; CFTD; MPFD; NEM1; NEM2; NEM3; CFTD1; CFTDM; ACTA1	102610	NG_006672.1	1q42.13
TPM3	7170	tropomyosin 3	TM3; TRK; NEM1; TM-5; TM30; TM30nm; TPMsk3; hscp30; MGC3261; FLJ41118; MGC14582; MGC72094; OK/SW-cl.5; TPM3	191030	NG_008621.1	1q21.2
TNNT1	7138	troponin T type 1 (skeletal, slow)	ANM; TNT; STNT; TNTS; FLJ98147; MGC104241; TNNT1	191041	NG_011829.1	19q13.4
NEB	4703	nebulin	NEM2; NEB177D; FLJ11505; FLJ36536; FLJ39568; FLJ39584; DKFZp686C1456; NEB	161650	NG_009382.1	2q22
SMPD1	6609	sphingomyelin phosphodiesterase 1, acid lysosomal	ASM; NPD; SMPD1	607608	NG_011780.1	11p15.4-p15.1
GLDC	2731	glycine dehydrogenase (decarboxylating)	GCE; NKH; GCSP; HYGN1; MGC138198; MGC138200; GLDC	238300	NC_000009.10	9p22
GCSH	2653	glycine cleavage system protein H (aminomethyl carrier)	GCE; NKH; GCSH	238330	NC_000016.8	16q23.2
AMT	275	aminomethyltransferase	GCE; NKH; GCST; AMT	238310	NC_000003.10	3p21.2-p21.1
OTC	5009	ornithine carbamoyltransferase	OCTD; MGC129967; MGC129968; MGC138856; OTC	300461	NG_008471.1	Xp21.1
PAH	5053	phenylalanine hydroxylase	PH; PKU; PKU1; PAH	612349	NG_008690.1	12q22-q24.2

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
DHPR	5860	quinoid dihydropteridine reductase	DHPR; PKU2; SDR33C1; FLJ42391; QDPR	612676	NG_008763.1	4p15.31
PTS	5805	6-pyruvoyltetrahydropterin synthase	PTPS; FLJ97081; PTS	261640	NG_008743.1	11q22.3-q23.3
PCCA	5095	propionyl Coenzyme A carboxylase, alpha polypeptide	PCCA	232000	NG_008768.1	13q32
PCCB	5096	propionyl Coenzyme A carboxylase, beta polypeptide	DKFZp451E113; PCCB	232050	NG_008939.1	3q21-q22
ACADS	35	acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain	SCAD; ACAD3; ACADS	606885	NG_007991.1	12q22-qter
DHCR7	1717	7-dehydrocholesterol reductase	SLOS; DHCR7	602858	NC_000011.8	11q13.2-q13.5
SMN1	6606	survival of motor neuron 1, telomeric	SMA; SMN; SMA1; SMA2; SMA3; SMA4; SMA@; SMNT; BCD541; T-BCD541; SMN1	600354	NG_008691.1	5q13
HEXA	3073	hexosaminidase A (alpha polypeptide)	TSD; MGC99608; HEXA	606869	NG_009017.1	15q23-q24
MYO7A	4647	myosin VIIA	DFNB2; MYU7A; NSRD2; USH1B; DFNA11; MYOVIIA; MYO7A	276903	NG_009086.1	11q13.5
USH1C	10083	Usher syndrome 1C (autosomal recessive, severe)	PDZ73; AIE-75; DFNB18; PDZ-45; PDZ-73; NY-CO-37; NY-CO-38; ush1cpst; PDZ-73/NY-CO-38; USH1C	605242	NC_000011.8	11p15.1-p14
CDH23	64072	cadherin-like 23	USH1D; DFNB12; FLJ00233; FLJ36499; KIAA1774; KIAA1812; MGC102761; DKFZp434P2350; CDH23	605516	NG_008835.1	10q21-q22
PCDH15	65217	protocadherin 15	USH1F; DFNB23; DKFZp667A1711; PCDH15	605514	NG_009191.1	10q21.1
SANS	124590	Usher syndrome 1G (autosomal recessive)	SANS; ANKS4A; FLJ33924; USH1G	607696	NG_007882.1	17q25.1

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
ARX	170302	aristaless related homeobox	ISSX; PRTS; MRX29; MRX32; MRX33; MRX36; MRX38; MRX43; MRX54; MRX76; MRX87; MRXS1; ARX	300382	NG_008281.1	Xp21
OPHN1	4983	oligophrenin 1	OPN1; MRX60; OPHN1	300127	NG_008960.1	Xq12
JARID1C	8242	lysine (K)-specific demethylase 5C	MRXJ; SMCX; MRXSJ; XE169; JARID1C; DXS1272E; KDM5C	314690	NG_008085.1	Xp11.22-p11.21
FTSJ1	24140	FtsJ homolog 1 (E. coli)	JM23; MRX9; SPB1; TRM7; CDLIV; MRX44; FTSJ1	300499	NG_008879.1	Xp11.23
SLC6A8	6535	solute carrier family 6 (neurotransmitter transporter, creatine), member 8	CRT; CT1; CRTR; MGC87396; SLC6A8	300036	NC_000023.9	Xq28
DLG3	1741	discs, large homolog 3 (Drosophila)	MRX; MRX90; NEDLG; NE-Dlg; SAP102; SAP102; KIAA1232; DLG3	300189	NC_000023.9	Xq13.1
TM4SF2	7102	tetraspanin 7	A15; MXS1; CD231; MRX58; CCG-B7; TM4SF2; TALLA-1; TM4SF2b; DXS1692E; TSPAN7	300096	NG_009160.1	Xp11.4
ZNF41	7592	zinc finger protein 41	MRX89; MGC8941; ZNF41	314995	NG_008238.1	Xp11.23
FACL4	2182	acyl-CoA synthetase long-chain family member 4	ACS4; FACL4; LACS4; MRX63; MRX68; ACSL4	300157	NG_008053.1	Xq22.3-q23
PQBP1	10084	polyglutamine binding protein 1	SHS; MRX55; MRXS3; MRXS8; NPW38; RENS1; PQBP1	300463	NC_000023.9	Xp11.23
PEX1	5189	peroxisomal biogenesis factor 1	ZWS1; PEX1	602136	NG_008341.1	7q21.2
PXMP3	5828	peroxisomal membrane protein 3, 35kDa	PAF1; PEX2; PMP3; PAF-1; PMP35; RNF72; PXMP3	170993	NG_008371.1	8q21.1
PEX6	5190	peroxisomal biogenesis factor 6	PAF2; PAF-2; PXAAA1; PEX6	601498	NG_008370.1	6p21.1
PEX10	5192	peroxisomal biogenesis	NALD; RNF69;	602859	NG_008342.1	1p36.32

Gene name	Gene ID	Description	Gene aliases	OMIM	RefSeqGene	Chromosome map position
		factor 10	MGC1998; PEX10			
PEX12	5193	peroxisomal biogenesis factor 12	PAF-3; PEX12	601758	NG_008447.1	17q12
PEX5	5830	peroxisomal biogenesis factor 5	PXR1; PTS1R; PTS1-BP; FLJ50634; FLJ50721; FLJ51948; PEX5	600414	NG_008448.1	12p13.31
PEX26	55670	peroxisomal biogenesis factor 26	FLJ20695; PEX26M1T; Pex26pM1T; PEX26	608666	NG_008339.1	22q11.21

Table 2: Polymorphisms

Gene name	Gene ID	SNP ID	Mutation
CFTR	1080	rs63500661	TCACATCACCAAGTTAAAAAAAAAAAAA[A/G]GGGGCGGGGGGCAGAATGAAAATT
CFTR	1080	rs63107760	AAACAAGGATGAATTAAGTTTTTTTT[-/T]AAAAAAGAAACATTTGGTAAGGGGA
CFTR	1080	rs62469443	ATCACCAAGTTAAAAAAAAAAAAAGGG[A/G]CGGGGGGGCAGAATGAAAATTGCAT
CFTR	1080	rs62469442	CTATTGAACCAGAACCAAACAGGAAT[A/G]CCATAGCATTTTGTAAACTAAACTG
CFTR	1080	rs62469441	CAGGAGTTCAAGACCAGCCTACTAAA[A/C]CACACACACACACACACACACACAC
CFTR	1080	rs62469439	GATTAAATAATAGTGTTTATGTACCC[C/G]GCTTATAGGAGAAGAGGGTGTGTGT
CFTR	1080	rs62469438	ATTGTTATCTTTTCATATAAGGTAAC[A/T]GAGGCCAGAGAGATTAAATAACAT
CFTR	1080	rs62469437	TAATTTTAATTAAGTAAATTTAATTG[A/G]TAGATAAATAAGTAGATAAAAATA
CFTR	1080	rs62469436	GTATAAAAAAAAAAAAAAAAAAAGTT[A/T]GAATGTTTTCTTGCATTCA GAGCCT
CFTR	1080	rs62469435	ATACTAAAAATTTAAAGTTCTCTTGC[A/G]ATATATTTTCTTAATATCTTACATC
CFTR	1080	rs62469434	TGCTGGGATTACAGGCGTGAGCCACC[A/G]CGCCTGGCCTGATGGGACATATTTT
CFTR	1080	rs62469433	CTACAATATAAGTATAGTATTGCAAA[A/C]CCATCAGGAAGGGTGTTAACTATTT

Gene name	Gene ID	SNP ID	Mutation
CFTR	1080	rs61763210	GTTGTCTCCAAACTTTTTTCAGGTG[-/AGA]AGGTGGCCAACCGAGCTTCGGAAAG
CFTR	1080	rs61720488	TTTTTTCATAAAAGATTATATAAAGG[A/C]TATTGCTTTTGAATCACAAACACTA
CFTR	1080	rs61481156	ATCTAGTGAGCAGTCAGGAAAGAGAA[C/T]TTCCAGATCCTGGAAATCAGGGTTA
CFTR	1080	rs61443875	TAGAGTATAAAAAAAAAAAAAAAAAAAAA[-/A]GTTTGAATGTTTTCTTGCATTGAGA
CFTR	1080	rs61312222	TGCAAATGCCAACTATCAAAGATATT[C/G]GAGTATACTGTCAATAAACTTCATA
CFTR	1080	rs61159372	TCCTCAACAGTTAGAAACAATATTTT[C/G]AGTGATTTCCTATGCCAACTTTACT
CFTR	1080	rs61094145	TTTTTGTTATTGTTGTTAAATAAGTG[A/G]GAATTCAATACAGTATAATGTCTGT
CFTR	1080	rs61086387	CTTGAAATCGGATATATATATATATA[-/TGTATATATATATATATATATATATATACATATATATATATA]GTATTATCCCTGTTTTACAGTTTT
CFTR	1080	rs60996744	AGAGGGGCTGTGAAGGACACCAAGGA[A/G]GAGACTAAGAGCCAGGAGGGAAAC
CFTR	1080	rs60960860	TAGAGTTTATTAGCTTTTACTACTCT[A/G]CTTAGTTACTTTGTGTTACAGATA
CFTR	1080	rs60923902	ACTAGTGATGATGAGCTTCTTTTCAT[-/AT]GTTTGTGGCTGCATAAATGCTTC
CFTR	1080	rs60912824	GCAGAGAAAAGAGGGGCTGTGAAGGA[C/G]ACCAAGGAGGAGACTAAGAGCCAGG
CFTR	1080	rs60887846	TTCAGAGGTCTACCACTGGTGCATAC[G/T]CTAATCACAGTGTCGAAAATTTTAC
CFTR	1080	rs60793174	AAGAAAGAGCAAAAGAGGGCAAACCT[C/T]TCATACATTTTTGATGTGCGAAACCA
CFTR	1080	rs60788575	CCTAAAGTTTAAAAAGAAAAAAAAAAAA[-/A]GGAAGAAGGAATTAATAATCCAAAG
CFTR	1080	rs60760741	GTGTGTGTGTGTATATATATATATAT[A/T]TATATATTTTTTTTTTCTGAGCCA
CFTR	1080	rs60456599	AAACTGTTGATGTTTTCATTTATTTA[C/G]ATCATTGGAAAACCTTAGATTCTAG
CFTR	1080	rs60363249	TTTATCCATTCTTAACCAGAACAGAC[A/G]TTTTTTCAGAGCTGGTCCAGGAAAA
CFTR	1080	rs60355115	TTGAAATCGGATATATATATATATAT[A/G]TATATATATATATATATATATATAT

Gene name	Gene ID	SNP ID	Mutation
GBA	2629	rs12034326	AAGCAGCCCTGGGGAGTCGGGGCGGG[A/G]CCTGGATTGAAAAGAGACGGTCAC
GBA	2629	rs11558184	CTCCAAGTTCTGGGAGCAGAGTGTGC[A/G]GCTAGGCTCCTGGGATCGAGGGATG
GBA	2629	rs11430678	GTTCCCTCCAGTAAtttttttttt[-/G/T]gttttgagacagagtcttgcctgt
GBA	2629	rs11264345	CTAGTACCTTACTTCCCTCAAGTTCA[A/T]TCATCTCACAGATATTTCCTGAGCA
GBA	2629	rs10908459	aattagccgtgcgtggtggcgggtgc[C/T]tgtaatcccacgtactgggaggct
GBA	2629	rs10796940	CCATGGCCAGCCGGGGAGGGGACGGG[A/C]ACACACAGACCCACACAGAGACTCA
GBA	2629	rs10668496	agcgagagagagagagagagagagag[-/AG]gagCCGGCGCGAGAACTACGCATGC
GBA	2629	rs7416991	CGTAGCAGTTAGCAGATGATAGGCGG[C/G/T]GAAATCTTATTTACAGGGCATTAA
GBA	2629	rs4024049	CTGGCCCTGGTGACAGTGGGGCTGTG[C/T]GTGGGGCCAGAGCCTTCTCAGAGGT
GBA	2629	rs4024048	CAGATACTGGCCCTGGTGACAGTGGG[A/G]CTGTGCGTGGGGCCAGAGCCTTCTC
GBA	2629	rs4024047	GACAGATACTGGCCCTGGTGACAGTG[G/T]GGCTGTGCGTGGGGCCAGAGCCTTC
GBA	2629	rs3841430	GGCTCctctctctctctctctct[-/TC]gctcgtctctcgtctctcgtctct
GBA	2629	rs3754485	GTTTCAGACCAGCCTGGCCAACATAG[C/T]GAAACCCCATCTCTACTAAAATAA
GBA	2629	rs3205619	AGTGGGCGATTGGATGGAGCTGAGTA[C/T]GGGGCCCATCCAGGCTAATCACACC
GBA	2629	rs2990227	CCGGGCTCCGTGAATGTTTGTACAT[C/G]TCTGAAGAACGTATGAATTACATAA
GBA	2629	rs2990226	GAATCCCAACCCCGACGCTCGTCGCC[C/G]GGCTCCGTGAATGTTTGTACATGT
GBA	2629	rs2990225	GCGAATCCCAACCCCGACGCTCGTCG[C/T]CGGGCTCCGTGAATGTTTGTACACAT
GBA	2629	rs2990224	TGGGCAGAAGTCAGGGTCCAAAGAAA[G/T]GGCAAAGAAAAGTGTcagtggctca
ASPA	443	rs63751297	TAAGAAAGACGTTTTTGATTTTTTTC[A/G]GACTTCTCTGGCTCCACTACCCTGC
ASPA	443	rs62071301	CTGATTCCCTGGCCAGGAGCGGTGGCT[C/T]ACGCCTGTAATCCCAGCGCTTTGGG
ASPA	443	rs62071300	TAAAAATGCTGATTCCTGGCCAGGAG[C/T]GGTGGCTCACGCCTGTAATCCAGC

Gene name	Gene ID	SNP ID	Mutation
ASPA	443	rs62071299	TTTAAAAATGCTGATTCTTGGCCAGG[A/C]GCGGTGGCTCACGCCTGTAA TCCCA
ASPA	443	rs62071297	CAAGACCTGTCAAAGATCTGAGAAAT[A/T]TTACCCGACTTACAAGCTA ACCATT
ASPA	443	rs61697033	ACTGTAATAAGTGCTGTAAAAGAAAT[A/G]CACAAAATAATATAGCAGA GGGTAT
ASPA	443	rs60743592	CTTGAGGTCAGGAGTTCAAGACCAGT[C/T]TGGGCAACATGGGGAAAAAC CTTGTC
ASPA	443	rs60666840	AGGTTGCAGTGAGCCGAGATCATGCC[A/G]TTGCACTCCAGCCGGGGCA AAAAA
ASPA	443	rs60147514	ACAAGTGTCTTGAAATTATCTGTGAT[C/T]TGCTATAGAGCAATACTTTT GTAAA
ASPA	443	rs59930743	GTGGGTATATGCAGCTCTATGCACTA[C/T]CTGCTCATTTATTTGGTAAAT CTAA
ASPA	443	rs59690349	TGTGTGTGTGTGCGTGTGTGTGTGTG[- /TGTGTGTG]ATCATAAGAGTGGCTGCAGCAAAC
ASPA	443	rs59676360	AGTCTGGAGTGCAATGGTGCAATCTC[A/G]GCTCACTGCAGCCTCCACCT CCGGG
ASPA	443	rs59335404	CTCCTAATGGATATTTCTAAATTTT[G/T]CTGAACAGAATTTAACTTGA GCTGG
ASPA	443	rs58879097	ATTTAAAAATGGATTTCTAGAAAAAC[A/G]ATCACATACTTGAATATTTT AGCAA
ASPA	443	rs58686774	CTATAAATGGGTAGCATGAGGGATTC[A/G]AGGAGGTGGCTGAAAGAAG CACGTA
ASPA	443	rs57511162	AAGAAACCAAGCATAGTAGAGTGTTA[A/G]AAAACCAAAGCAACTAAA CAACTGT
ASPA	443	rs55859596	CGGGGCTCAGAACTTGTAACAGAAAA[A/T]TAAAATATACTCCACTCAA GGGAAT
ASPA	443	rs55742972	TACTACACTTCACGGATACTGTACTT[- /GTACTT]TTTTTCCAAATTGAAGGTTTTTGGC
ASPA	443	rs55640436	TGTFTTTTGTFTTTGTFTTTGTFTTT[- /GTTTTGTFTTT]TGAGATGGAGTCTCGCTCTGTCGCC
ASPA	443	rs36225687	TTTGCCTTACTACACTTCACGGATAC[- /TGACT]TGTACTTTTTTTCCAAATTGAAGGT
ASPA	443	rs36051310	GAGGTGGCTGAAAGAAGCACGTATCC[- /C]TGATGGCATGGTTGCGGGTTATATG
ASPA	443	rs36034906	GAGAAAAGCAGTTCCTGGAACACCCC[- /C]ACCCCTTAACCCCTTATCTCTGCTT
ASPA	443	rs36033666	TTACATATGTATACATGTGCCATGTT[-

Gene name	Gene ID	SNP ID	Mutation
			/T]GGTGTGCCGCACCCATTAAC TCGTC
ASPA	443	rs35730123	CTTTTCCAGATTTTTTTTTTTTTT[- /T]GAGACAGAGTTTCACTCTTGTG
ASPA	443	rs35629100	TTTGGAATCTTAAGCTTTTATTTGG[- /G]TGTCACAGAGAAACAGGATCTGTAT
ASPA	443	rs35614631	TACTTTAAGTTTTAGGGTACATGTGC[- /A]CCATGTGCAGGTTTGTACATATGT
ASPA	443	rs35225782	ATTCATGACCAGCCACATAAATGCAC[- /A]GTATTACTTCGCAAGCATGCCAATG
ASPA	443	rs35178659	GTGCACTAGAATTAGCTAAAGTGGGG[- /G]AAAAAAGATGCATTTGATGGTCTA
ASPA	443	rs35095578	AACCTCCACCTCCCAGGTTCAAGAGA[- /A]TTCTCCTGCCTCAGCCTCCCAAGTA
ASPA	443	rs35002210	CCTCCCTGTGATCCGAAGTAGCAGAC[A/G]TACTTAACTTCCATGGTGGATTGTT
ASPA	443	rs34744839	AAAACATTATTATATCTAGAAAAAA[- /A]TGTATCTTAACCATTGTGGGAAGTG
ASPA	443	rs34680506	TTGAAGGTAAAATCATAGGGAGTTGG[- /G]AGCTGTCCTCTTGCCTGAATCAGT
ASPA	443	rs34365618	ACTTGTGGCCTTTTTGGAGAGGTTAG[- /CA]ACTCTGAAACTCTGTCCCTGGACC
ASPA	443	rs34275920	GAAGGAGAAAAAGAGAGGAAATAAGT[- /T]AAAATAATAAACACAATTAATAAAG
ASPA	443	rs34109510	TGTATACATGTGCCATGTTGGTGTGC[C/T]GCACCCATTAAC TCGTCATTTAGCA
ASPA	443	rs34054576	TCACCTGTCACCTCCTATAGAACTTT[- /C]CCCTGACCCTCCTCTATAGCATTA
ASPA	443	rs34015272	ATAAATGATCATCATTCACAGTAGGG[- /G]TTTTGTTTTGTTTTTTTCTGGAAA
ASPA	443	rs34002091	ACAGACATATCTACAAACACACTTTT[- /T]CACATATTTGTGTAAGTCATTTATG
ASPA	443	rs28940574	AAAGACAATAAATAACGCTCAATG[A/C]AAAAAGTATTCGCTGCTGTTACAT
ASPA	443	rs28940279	TACCGTGTACCCCGTGTGTTGTGAATG[A/C]GGCCGCATATTACGAAAAGAAAGAA
ASPA	443	rs17850703	CAGGGCTGGAGGTAAAACCATTTATT[A/G]CTAACCCAGAGCAGTGAA GAAGTG
ASPA	443	rs17222495	TTCTTCATTGCCTATTGAAGAGAGAG[C/T]GGAATGCTTTGGTTGCCAGATATGG

Gene name	Gene ID	SNP ID	Mutation
ASPA	443	rs17175228	CACAAGATCTCATTACTCAGGAGCTG[C/T]CCAAGTGTCTAATGTACTTAGTTAA
ASPA	443	rs16953074	TTCTGTGTAACATTTTCATTTAAGCAA[A/G]GGATTTCGGCAAATCAAAAATGTCA
ASPA	443	rs16953070	TAAAACGTATTGAAGGTATTATTGAC[G/T]CTGTTGAAGCAAAGAGAACAAAACA
HEXA	3073	rs62022858	ATCTGCTCTTCCAGTTGGATGACAAG[C/T]CTTGCTGTCTAACACCTGCTGCAGA
HEXA	3073	rs62022857	CCATTTTTTGTGTGATTTTTTTTTTTC[C/T]TGAATACTTTTTATCGCAGTTGGTT
HEXA	3073	rs62017872	CCCTGTCTCTAAAAGAAAAAAAAAAAA[A/G]AAAAAAAAAAGAAAACAAACCCAA
HEXA	3073	rs62017871	AGTGGCTCCAAAAAGGTCATGGAACC[C/T]CTTGAGGATGATGCAAATTGACTCT
HEXA	3073	rs61662730	TAAAGTTACTTTTCTTTTATTGACTT[C/T]CCCTTATTTTTTAACCTTATGCTTT
HEXA	3073	rs61329913	CAGAGTTAAAAAAAAAAAAAAAAAAAA[-/A]GGAAGTAGCAGCAACAGCTTGGA
HEXA	3073	rs60920713	GTTGCCCAGGGTTGAGTGCAGAGGCA[C/T]ATCTTGGCTCACAGCAACCTCTGCC
HEXA	3073	rs60783213	AAGGCTTTTTTTTTTTTTTTTTTTT[-/TTTT]GAGACAGAGTCTTGCTGTGTCACCC
HEXA	3073	rs60644867	GCCTACATTCTGCAAAGAGGAGGGAA[C/G]ATTCACAGCTCCATACTTGAACCCCT
HEXA	3073	rs60288568	CCAAAGGAGAATAGCTCTAGGGGAGG[C/G]AGGTGGATGAGTATGCATGGGGAG
HEXA	3073	rs59888548	GACTCCATCTCAAAAAAAAAAAAAAAAAA[-/A]TGCAGTCTAATGGCAGAATTAGACT
HEXA	3073	rs59733856	TTATTTATTTATTTATTTATTTTGA[A/G]ACAGGGTCTCTGTTGTCCAGGCTGG
HEXA	3073	rs59427837	TTTTGAGGCAGGGTCTCACTCTGTTG[C/T]CCAGGGTTGAGTGCAGAGGCACATC
HEXA	3073	rs59171976	CGCCTTGCGAAGGCCCCACAGCTTGC[C/T]TGTGACAAACGTTCATAGGCAATG
HEXA	3073	rs58706602	GGAGGTCTGTACAAAGCACCACTAC[C/T]TCATGGGTCAGTTTCCACAGCAGAA
HEXA	3073	rs58696963	GAATCTTATAATTCACTGTGTACCTC[-/CCTC]TGTTTCATTTTTTCGCAATTGA
HEXA	3073	rs58610850	AACATAGTATCTAATATAGCTTTACA[C/T]CCAAAGCCAAAATATGAATA

Gene name	Gene ID	SNP ID	Mutation
			CACTG
HEXA	3073	rs58016062	TTGTTTTGTTTTGTTTGGGGGGGGGG[-/G]TTGTTTTCTGAGAGGGAGTCTTGC
HEXA	3073	rs57733983	CATACCAAAGGGCAGCTGGAGGGATA[C/T]AGACGGAAGTCATGTGGAGAGTGAA
HEXA	3073	rs57476645	CAGGTGTGAGCCACCACGACCACCAA[A/T]TTAGCTCTTTTACTCCTTCCTTC
HEXA	3073	rs56870003	AGTGGTAGCTGATTTTGCTTCTGGAT[A/C]CTTGCCACCTTCCCCTCTTTAAT
HEXA	3073	rs56338339	AAAGACCTGTTTCTTAAAAAAAAAAAA[-/AGAAAAAAAAAAAA]GAAAGAAAAGAAAAGAAAAAACAG
HEXA	3073	rs55995352	TAAAAAATCTTTCAATGAGGAGATGT[C/T]CCCAGAGCAAGACAGCTGTAGGATG
HEXA	3073	rs55860138	AAAAGAAAAAAAAAAAAAAAAAAAAA[-/A]GAAAACAAAACCCAAACCCATAAAG
HEXA	3073	rs55743646	CCTGTCTCTAAAAGAAAAAAAAAAAAA[A/G]AAAAAAAAAGAAAACAAAACCCAAA
HEXA	3073	rs55665666	GTTATCATAGAAAAATATCACACTCT[-/GT]CTGTATCCCCACTTCCAGAACTGT
HEXA	3073	rs36106892	CAGGAGCTCATAGAATTACATACAAT[-/C]TTTTTTTTTTTTTTTGAGACAGCG
HEXA	3073	rs36091525	TTGAGAATCTTATAATTCAGTGTGA[-/CCTC]CCTCTGTTTCATATTTTCGCAATTG
HEXA	3073	rs35949555	CCACTACCACAGTGCCTAGAGAACAA[C/T]ATGTGTTTAATAATATTTAAATAAT
HEXA	3073	rs35827424	CCCTGTCTCTAAAAGAAAAAAAAAAAA[-/A]AAAAAAAAAGAAAACAAAACCCAA
HEXA	3073	rs35729578	CCATTATATCATTCATTTCCCCTCA[-/T]TTTCTTCATTCCAACCAAGATATAT
HEXA	3073	rs35649102	TCCGTCTCAAAAAAAAAAAAAAAAAAG[-/A]GAAAGGAATTATCTCATGTATACA
HEXA	3073	rs35118677	CTGGGGCAGTTAAAAAGAAAAACAAA[-/C]CCCTGGTCCCTGCCCTTGAGGAGAT
HEXA	3073	rs35005352	CTCCAGGGTCCCATTCCAGGACCACA[-/C]GCCTGCTACCTCTGCAGCTCACTCA
HEXA	3073	rs34736306	GGATTGACATATACCAGTTAGACGGA[-/T]TTTTTTTTTCCATAAACCAGGCTCA
HEXA	3073	rs34607939	ACAAATAATTACTACATATCTACAAC[A/G]TTCCAGATACAGAAGAAATGGCCAA

Gene name	Gene ID	SNP ID	Mutation
HEXA	3073	rs34496117	TAAACACACTTGAAACATCATATAAA[-/ATG]ATATTACTACAAGACTTAACCGTAA
HEXA	3073	rs34300017	ACACAGGTAATCCATGTTTATTATAG[-/A]AAAATGCCACATTACTCTTTATTGA
HEXA	3073	rs34206496	AGTTATCATAGAAAAATATCACACTC[-/TG]TCTGTATCCCCACTTCCAGAACTG
HEXA	3073	rs34110830	AATGAACTTACAGGAAGGTAATATAT[-/G]GGAAATAAACATCTTATTGAATTTA
HEXA	3073	rs34093438	GGACCCCTGAAAGGCACAAGACACCC[-/T]TTCAGGTTACACTTCCTGAAAAGCT
HEXA	3073	rs34085965	CCACCAATCACCAGAGCCTTCTGCTC[A/G]GGGGTACCTGAGGGAAAAC AAGCAA
HEXA	3073	rs34004907	AAAGACTGAAAAAACATTTCATAACTA[-/T]TTTTCTTGTTATCCTCGGAAATGTC
HEXA	3073	rs28942072	TATCTTCATCTTGGAGGAGATGAGGT[C/T]GATTTACCTGCTGGAAGTC CAACC
HEXA	3073	rs28942071	TTCCTATGAACGTTTGTACACTTC[C/T]GCTGTGAGTTGCTGAGGCGA GGTGT
HEXA	3073	rs28941771	GCTTGCTGTTGGATACATCTCGCCAT[C/T]ACCTGCCACTCTCTAGCATCC TGGA
HEXA	3073	rs28941770	CCGGGGCTTGCTGTTGGATACATCTC[G/T]CCATTACCTGCCACTCTCTA GCATC

Examples

Example 1: Use of differentiator tag sequences to detect and correct bias in a MIP-capture reaction of a set of exon targets

The first step in performing the detection/correction is to determine how many differentiator tag sequences are necessary for the given sample. In this example, 1000 genomic targets corresponding to 1000 exons were captured. Since the differentiator tag sequence is part of the probe, it will measure/report biases that occur from the earliest protocol steps. Also, being located in the backbone, the differentiator tag sequence can easily be sequenced from a separate priming site, and therefore not impact the total achievable read-length for the target sequence. MIP probes are synthesized using standard column-based oligonucleotide synthesis by any number of vendors (e.g. IDT), and differentiator tag sequences are introduced as ‘degenerate’ positions in the backbone. Each degenerate

position increases the total number of differentiator tag sequences synthesized by a factor of 4, so a 10 nt degenerate region implies a differentiator tag sequence complexity of ~1e6 species.

Hybridize capture probes to human genomic sample

Assemble hybridization reaction:

- 1.0 ul capture probe mix (~2.5 pmol)
- 2.0 ul 10X Ampligase buffer (Epicentre)
- 6.0 ul 500 ng/ul human genomic DNA (~16.7 fmol)
- 11 ul dH₂O

In a thermal cycler, heat reaction to 95C for 5 min to denature genomic DNA, then cool to 60C. Allow to incubate at 60C for 40 hours.

Convert hybridized probes into covalently-closed circular products containing subtargets

Prepare fill-in / ligation reaction mixture:

- 0.25 ul 2 mM dNTP mix (Invitrogen)
- 2.5 ul 10X Ampligase buffer (Epicentre)
- 5.0 ul 5 U/ul Taq Stoffel fragment (Applied Biosystems)
- 12.5 ul 5 U/ul Ampligase (Epicentre)
- 4.75 ul dH₂O

Add 1.0 ul of this mix to the hybridized probe reaction, and incubate at 60C for 10 hours.

Purify circularized probe / subtarget products from un-reacted probes and genomic DNA

Prepare exonuclease reaction mixture:

- 21 ul fill-in / ligation reaction product
- 2.0 ul 10X exonuclease I buffer (New England Biolabs)
- 2.0 ul 20 U/ul exonuclease I (New England Biolabs)

- 2.0 ul 100 U/ul exonuclease III (New England Biolabs)

Incubate at 37C for 60 min, then heat-inactivate by incubating at 80C for 15 min.

Immediately cool to 4C for storage.

Amplify circular material by PCR using primers specific to the 'constant region' of the probes

Prepare PCR mixture:

- 5.0 ul 10X Accuprime reaction buffer (Invitrogen)
- 1.5 ul 10 uM CP-2-FA (5'-GCACGATCCGACGGTAGTGT-3')
- 1.5 ul 10 uM CP-2-RA (5'-CCGTAATCGGGAAGCTGAAG-3')
- 0.4 ul 25 mM dNTP mix (Invitrogen)
- 2.0 ul heat-inactivated exonuclease reaction mix
- 1.5 ul 10X SybrGreen (Invitrogen)
- 0.4 ul 2.5 U/ul Accuprime Pfx polymerase (Invitrogen)
- 37.7 ul dH₂O

Thermal cycle in real-time thermal cycler according to the following protocol, but stop cycling before amplification yield plateaus (generally 8-12 cycles):

1. 95C for 5 min
2. 95C for 30 sec
3. 58C for 60 sec
4. 72C for 60 sec
5. goto 2, N more times

Prepare a shotgun next-generation sequencing library for analysis

- Purify desired amplicon population from non-specific amplification products by gel extraction.
- Concatemerize amplicons into high-molecular weight products suitable for shearing
- Mechanically shear, using either a nebulizer, BioRuptor, Hydroshear, Covaris, or similar instrument. DNA should be sheared into fragments several hundred basepairs in length.

- Ligate adapters required for amplification by the sequencing platform used. If necessary, purify ligated product from unligated product and adapters.

Perform sequencing of library according to manufacturer's directions (e.g. Illumina, ABI, etc), reading both the target sequence and the differentiator tag sequence.

Analyze data by correcting for any biases detected by quantitation of differentiator tag sequence abundance

Construct a table of target:differentiator tag abundances from the read data, e.g.:

Target ID	Differentiator tag sequence ID	Count
1	3547	1
2	4762	1
1	9637	1
1	1078	5
3	4762	1
1	2984	1

All 'count' entries should be '1', since any particular target:differentiator tag mapping will not occur more than once by chance, and therefore will only be observed if bias was present somewhere in the sample preparation process. For any target:differentiator tag combination observed more than once, all such reads are 'collapsed' into a single read before consensus basecalls are determined. This will cancel the effect of bias on consensus basecall accuracy. Figure 2 depicts a method for making diploid genotype calls in which repeat target:differentiator tag combination are collapsed.

Example 2: Differentiator tag sequence design for MIP capture reactions

For a set of targets, the number of differentiator tag sequences necessary to be confident (within some statistical bounds) that a certain differentiator tag sequence will not be observed more than once by chance in combination with a certain target sequence was

determined. The total number of unique differentiator tag sequences for a certain differentiator tag sequence length is determined as $4^{(\text{Length in nucleotides of the differentiator tag sequence})}$. For a molecular inversion probe capture reaction that uses MIP probes having differentiator tag sequences, the probability of performing the capture reaction and capturing one or more copies of a target sequence having the same differentiator tag sequence is calculated as: $p = 1 - [N! / (N-M)!] / [N^M]$, wherein N is the total number of possible unique differentiator tag sequences and M is the number of target sequence copies in the capture reaction. Thus, by varying the differentiator tag sequence length it is possible to perform a MIP capture reaction in which the probability of capturing one or more copies of a target sequence having the same differentiator tag sequence is set at a predetermined probability value.

For example, for a differentiator tag sequence of 15 nucleotides in length, there are 1,073,741,824 possible differentiator tag sequences. A MIP capture reaction in which MIP probes, each having a differentiator tag sequence of 15 nucleotides, are combined with 10000 target sequence copies (e.g., genome equivalents), the probability of capturing one or more copies of a target sequence having the same differentiator tag sequence is 0.05. In this example, the MIP reaction will produce very few (usually 0, but occasionally 1 or more) targets where multiple copies are tagged with the same differentiator tag sequence. Figure 3 depicts results of a simulation for 100000 capture reactions having 15 nucleotide differentiator tag sequences and 10000 target sequences.

Example 3: Assessment of the probability for obtaining enough sequencing reads to make accurate base-calls at multiple independent loci, as a function of sequencing coverage.

Monte Carlo simulations were performed to determine sequencing coverage requirements. The simulations assume 10000 genomic copies of a given locus (target) half mom alleles and half dad alleles. The simulations further assume 1% efficiency of capture for the MIP reaction. The simulation samples from a capture mix 100 times without replacement to create a set of 100 capture products. The simulation then samples from the set of 100 capture products with replacement (assuming unbiased amplification) to generate 'reads' from either mom or dad. The number of reads sampled depends on the coverage. The number of independent reads from both mom and dad necessary to make a high-quality base-call (assumed to be 10 or 20 reads) were then determined. The process was repeated 1000 times for each coverage level, and the fraction of times that enough reads from both parents were

successfully obtained was determined. This fraction was raised to the power 1000, assuming we have 1000 independent loci that must obtain successful base-calls, plotted (See Figure 4). Result show that roughly 50x coverage is required to capture each allele $\geq 10x$ with > 0.95 probability.

It will be clear that the methods may be practiced other than as particularly described in the foregoing description and examples. Numerous modifications and variations of the present disclosure are possible in light of the above teachings and, therefore, are within the scope of the claims. Preferred features of each aspect of the disclosure are as for each of the other aspects *mutatis mutandis*. The documents including patents, patent applications, journal articles, or other disclosures mentioned herein are hereby incorporated by reference in their entirety. In the event of conflict, the disclosure of present application controls, other than in the event of clear error.

What is claimed is:

Claims

1. A method of genotyping a subject, the method comprising:
determining the sequence of at least a threshold number of independently isolated nucleic acids, wherein the sequence of each isolated nucleic acid comprises a target nucleic acid sequence and a differentiator tag sequence,
wherein the threshold number is a number of unique combinations of target nucleic acid and differentiator tag sequences, wherein the isolated nucleic acids are identified as independently isolated if they comprise unique combinations of target nucleic acid and differentiator tag sequences, and wherein the target nucleic acid sequence is the sequence of a genomic locus of a subject.
2. The method of claim 1, wherein the isolated nucleic acids are molecular inversion probe capture products.
3. A method of genotyping a subject, the method comprising:
determining the sequence of molecular inversion probe capture products, each comprising a molecular inversion probe and a target nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence, and wherein the target nucleic acid is a captured genomic locus of a subject, and
genotyping the subject at the captured genomic locus based on the sequence of at least a threshold number of unique combinations of target nucleic acid and differentiator tag sequences of molecular inversion probe capture products.
4. A method of genotyping a subject, the method comprising:
obtaining molecular inversion probe capture products, each comprising a molecular inversion probe and a target nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence, wherein the target nucleic acid is a captured genomic locus of the subject,
amplifying the molecular inversion probe capture products, and
genotyping the subject by determining, for each target nucleic acid, the sequence of at least a threshold number of unique combinations of target nucleic acid and differentiator tag sequence of molecular inversion probe capture products.

5. The method of claim 4, wherein the obtaining comprises capturing target nucleic acids from a genomic sample of the subject with molecular inversion probes, each comprising a unique differentiator tag sequence.

6. The method of claim 4 or 5, wherein the capturing is performed under conditions wherein the likelihood of obtaining two or more molecular inversion probe capture products with identical combinations of target and differentiator tag sequences is equal to or less than a predetermined value, optionally wherein the predetermined value is about 0.05.

7. The method of any one of claims 4-6, wherein the threshold number for a specific target nucleic acid sequence is selected based on a desired statistical confidence for the genotype.

8. The method of any one of claims 4-6, further comprising determining a statistical confidence for the genotype based on the number of unique combinations of target nucleic acid and differentiator tag sequences.

9. A method of analyzing a plurality of genetic loci, the method comprising:
obtaining a plurality of molecular inversion probe capture products each comprising a molecular inversion probe and a target nucleic acid, wherein the sequence of the molecular inversion probe comprises a differentiator tag sequence and, optionally, a primer sequence,
amplifying the plurality of molecular inversion probe capture products,
determining numbers of occurrence of combinations of target nucleic acid and differentiator tag sequence of molecular inversion probe capture products in the amplified plurality, and

if the number of occurrence of a specific combination of target nucleic acid sequence and differentiator tag sequence exceeds a predetermined value, detecting bias in the amplification of the molecular inversion probe comprising the specific combination.

10. The method of claim 9, further comprising genotyping target sequences in the plurality, wherein the genotyping comprises correcting for bias, if detected.

11. The method of any of claims 1-10, wherein the target nucleic acid sequence is a sequence of a gene selected from Table 1.

Abstract

The invention relates to preparative steps in DNA sequencing-related technologies that reduce the impact of bias in a genetic analysis and increase the reliability and accuracy of downstream quantitative evaluations or applications.

FIGURE 1

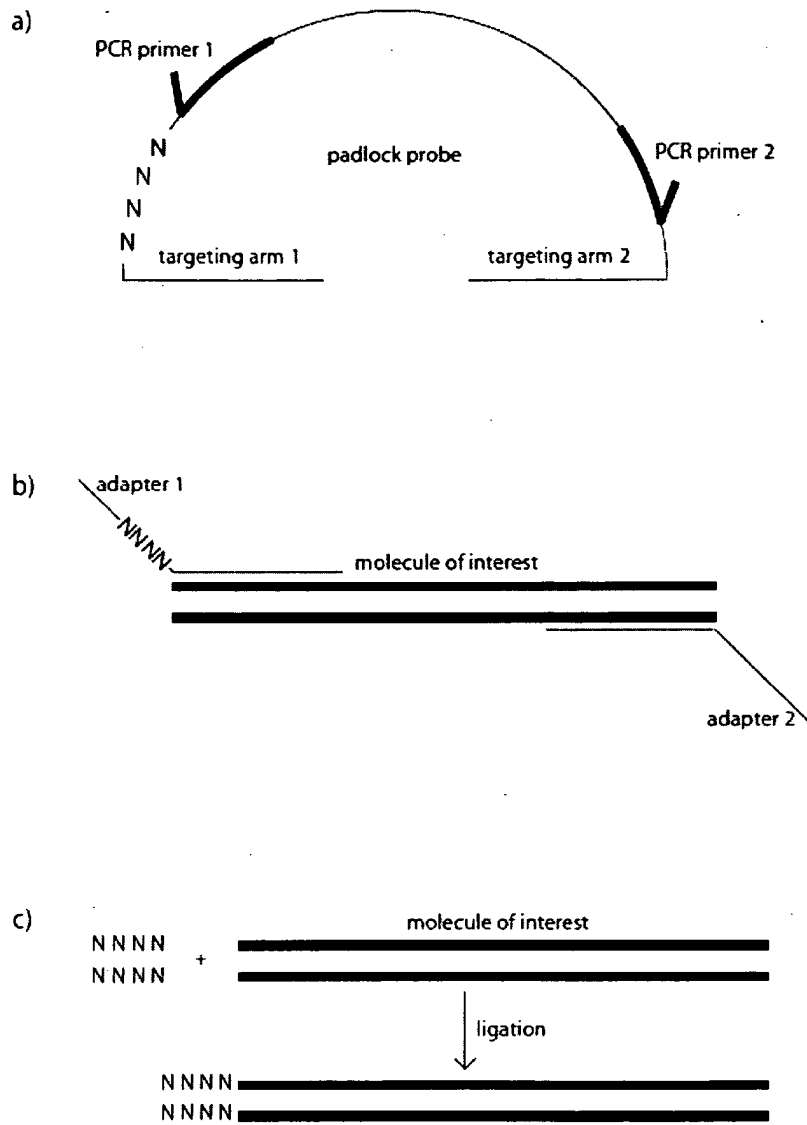


FIGURE 2

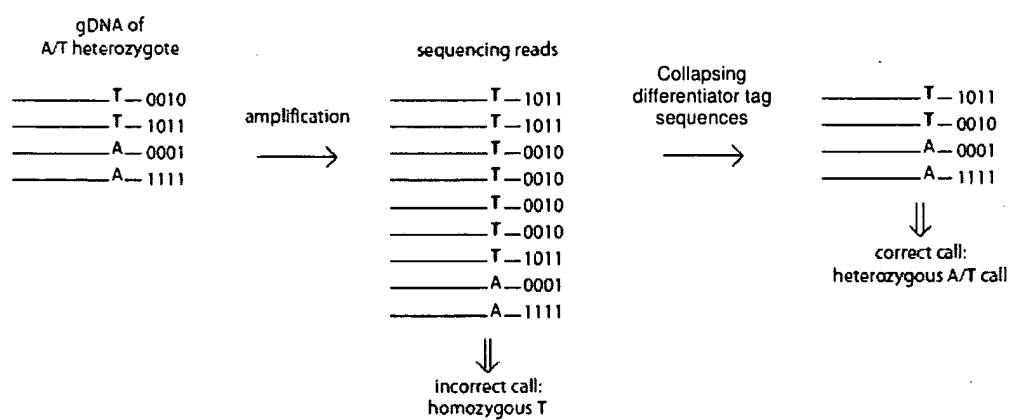


FIGURE 3

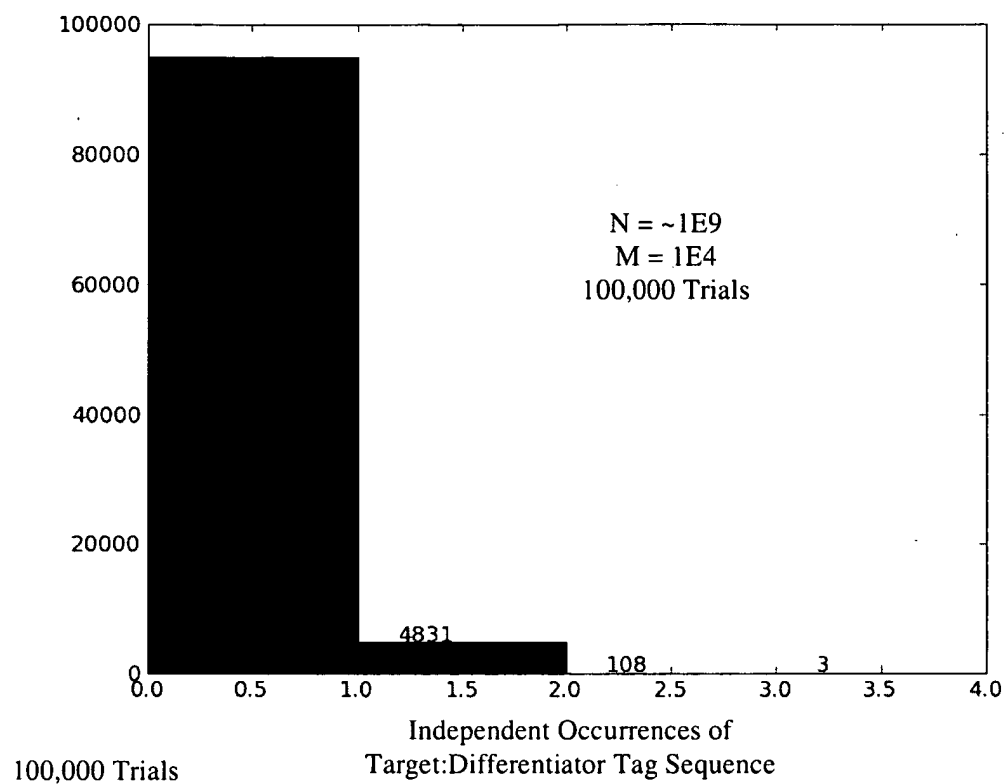


FIGURE 4

