

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. _____

VIA EFS _____

INVENTOR(S)		
Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Jin Billy	Li	Medford, Massachusetts 02155
Gregory	Porreca	Ocean City, New Jersey 08226
Uri	Laserson	Boston, Massachusetts 02215
Additional inventors are being named on the _____ separately numbered sheets attached hereto.		
TITLE OF THE INVENTION (500 characters max):		
EVALUATION OF TARGET LENGTH IN NUCLEIC ACID CAPTURE-BASED ASSAYS		
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>2</u> ☐ Other (specify) _____		
☒ Specification (e.g. description of the invention) Number of Pages <u>25</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 Director of the United States Patent and Trademark Office is enclosed to cover the filing fee and application size fee (if applicable).		
☒ Payment by credit card. Form PTO-2038 is attached.		
☒ 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

1649742.1

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

Pat R H Waller

Date

May 28, 2009

TYPED or PRINTED NAME

Patrick R.H. Waller

REGISTRATION NO.
(if appropriate)

41,418

TELEPHONE

617.646.8000

Docket Number:

G0814.70001US00

EVALUATION OF TARGET LENGTH IN NUCLEIC ACID CAPTURE-BASED ASSAYS

Field of Invention

5 The invention relates to methods for detecting nucleic acid deletions or insertions in regions containing nucleic acid sequence repeats.

Background of the Invention

10 Genomic regions that contain nucleic acid sequence repeats are often the site of genetic instability due to the amplification or contraction of the number of sequence repeats (e.g., the insertion or deletion of one or more units of the repeated sequence). of Instability in the length of genomic regions that contain high numbers of repeat sequences has been associated with a number of hereditary and non hereditary diseases and conditions.

15 For example, “Fragile X syndrome, or Martin-Bell syndrome, is a genetic syndrome which results in a spectrum of characteristic physical, intellectual, emotional and behavioral features which range from severe to mild in manifestation. The syndrome is associated with the expansion of a single trinucleotide gene sequence (CGG) on the X chromosome, and results in a failure to express the FMR-1 protein which is required for
20 normal neural development. There are four generally accepted forms of Fragile X syndrome which relate to the length of the repeated CGG sequence; Normal (29-31 CGG repeats), Premutation (55-200 CGG repeats), Full Mutation (more than 200 CGG repeats), and Intermediate or Gray Zone Alleles (40 - 60 repeats).”

25 Other examples include cancer, which has been associated with microsatellite instability (MSI) involving an increase or decrease in the genomic copy number of nucleic acid repeats at one or more microsatellite loci (e.g., BAT-25 and/or BAT-26).

 The are currently many sequencing-based assays for determining the number of nucleic acid sequence repeats at a particular locus and identifying the presence of nucleic acid insertions or deletions.

30

Summary of the Invention

 Aspects of the invention provide methods and compositions for identifying nucleic acid insertions or deletions in genomic regions of interest without determining

the nucleotide sequences of these regions. Aspects of the invention are particularly useful for detecting nucleic acid insertions or deletions in genomic regions containing nucleic acid sequence repeats (e.g., di- or tri-nucleotide repeats). However, the invention is not limited to analyzing nucleic acid repeats and may be used to detect insertions or deletions in any target nucleic acid of interest. Aspects of the invention are particularly useful for analyzing multiple loci in a multiplex assay.

In some embodiments, aspects of the invention relate to determining whether an amount of target nucleic acid that is captured in a genomic capture assay is higher or lower than expected. In some embodiments, a statistically significant deviation from an expected amount (e.g., higher or lower) is indicative of the presence of a nucleic acid insertion or deletion in the genomic region of interest. In some embodiments, the amount is a number of nucleic acid molecules that are captured. In some embodiments, the amount is a number of independently captured nucleic acid molecules in a sample. It should be appreciated that the captured nucleic acids may be literally captured from a sample, or their sequences may be captured without actually capturing the original nucleic acids in the sample. For example, nucleic acid sequences may be captured in an assay that involves a template-based extension of nucleic acids having the region of interest, in the sample.

Aspects of the invention are based on the recognition that the efficiency of certain capture techniques is affected by the length of the nucleic acid being captured. Accordingly, an increase or decrease in the length of a target nucleic acid (e.g., due to an insertion or deletion of a repeated sequence) can alter the capture efficiency of that nucleic acid. In some embodiments, a difference in the capture efficiency (e.g., a statistically significant difference in the capture efficiency) of a target nucleic acid is indicative of an insertion or deletion in the target nucleic acid. It should be appreciated that the capture efficiency for a target nucleic acid may be evaluated based on an amount of captured nucleic acid (e.g., number of captured nucleic acid molecules) relative to a control amount (e.g., based on an amount of control nucleic acid that is captured). However, the invention is not limited in this respect and other techniques for evaluating capture efficiency also may be used.

According to aspects of the invention, evaluating the capture efficiency as opposed to determining the sequence of the entire repeat region reduces errors associated

with sequencing through repeat regions. Repeat sequences often give rise to stutters or skips in sequencing reactions that make it very difficult to accurately determine the number of repeats in a target region without running multiple sequencing reactions under different conditions and carefully analyzing the results. Such procedures are
5 cumbersome and not readily scalable in a manner that is consistent with high throughput analyses of target nucleic acids. Accordingly, aspects of the invention are useful to increase the sensitivity of detecting insertions or deletions in target regions, particularly target regions containing repeated sequences.

In some embodiments, aspects of the invention relate to capturing genomic
10 nucleic acid sequences using a molecular inversion probe (e.g., MIP or Padlock probe) technique, and determining whether the amount (e.g., number) of captured sequences is higher or lower than expected. In some embodiments, the amount (e.g., number) of captured sequences is compared to an amount (e.g., number) of sequences captured in a control assay. The control assay may involve analyzing a control sample that contains a
15 nucleic acid from the same genetic locus having a known sequence length (e.g., a known number of nucleic acid repeats). However, a control may involve analyzing a second (e.g., different) genetic locus that is not expected to contain any insertions or deletions. The second genetic locus may be analyzed in the same sample as the locus being interrogated. The second genetic locus may be a locus that is not characterized by the
20 presence of nucleic acid repeats (and thus not expected to contain insertions or deletions of the repeat sequence).

In some embodiments, a target nucleic acid region that is being evaluated may be determined by the identify of the targeting arms of a probe that is designed to capture the target region (or sequence thereof). For example, the targeting arms of a MIP probe may
25 be designed to be complementary (e.g., sufficiently complementary for selective hybridization) to genomic regions flanking a target region suspected of containing an insertion or deletion. It should be appreciated that two targeting arms may be designed to be complementary (e.g., sufficiently complementary for selective hybridization) to the two flanking regions that are immediately adjacent (e.g., immediately 5' and 3',
30 respectively) to a region of a sequence repeat on one strand of a genomic nucleic acid. However, one or both targeting arms may be designed to hybridize several bases (e.g., 1-5, 5-10, 10-25, 25-50, or more) upstream or downstream from the repeat region in such a

way that the captured sequence includes a region of unique genomic sequence that on one or both sides of the repeat region.

In some embodiments, two or more (e.g., 2, 3, 4, 5, 6, 7, 8, 9, 10 or more) different loci may be interrogated in parallel in a single assay (e.g., in a multiplex assay).

5 In some embodiments, the ratio of captured nucleic acids for each locus may be used to determine whether a nucleic acid insertion or deletion is present in one locus relative to the other. For example, the ratio may be compared to a control ratio that is representative of the two loci when neither one has an insertion or deletion relative to control sequences (e.g., sequences that are normal or known to be associated with
10 healthy phenotypes for those loci). However, the amount of captured nucleic acids may be compared to any suitable control as discussed herein.

The locus of a captured sequence may be identified by determining a portion of unique sequence 5' and/or 3' to the repeat region in the target nucleic acid suspected of containing a deletion or insertion. This does not require sequencing the captured repeat
15 region itself. However, some or all of the repeat region also could be sequenced as aspects of the invention are not limited in this respect.

Aspects of the invention may be combined with one or more sequence-based assays (e.g., SNP detection assays), for example in a multiplex format, to determine the genotype of one or more regions of a subject.

20 According to some aspects of the invention, methods of detecting a polymorphism in a nucleic acid in a biological sample are provided. In some embodiments, the methods comprise evaluating the efficiency of capture at one or more loci and determining whether one or both alleles at that locus contain an insertion or deletion relative to a control locus (e.g., a locus indicative of a length of repeat sequence
25 that is associated with a healthy phenotype).

Accordingly, aspects of the invention relate to methods for determining whether a target nucleic acid has an abnormal length by evaluating the capture efficiency of a target nucleic acid in a biological sample from a subject, wherein a capture efficiency that is different from a reference capture efficiency is indicative of the presence, in the
30 biological sample, of a target nucleic acid having an abnormal length. It should be appreciated that the term "abnormal" is a relative term based on a comparison to a "normal" length. In some embodiments, a normal length is a length that is associated

with a normal (e.g., healthy or non-carrier phenotype). Accordingly, an abnormal length is a length that is either shorter or longer than the normal length. In some embodiments, the presence of an abnormal length is indicative of an increased risk that the locus is associated with a disease or a disease carrier phenotype. In some embodiments, the abnormal length is indicative that the subject is either has a disease or condition or is a carrier of a disease or condition (e.g., associated with the locus). However, it should be appreciated that the description of embodiments relating to detecting the presence of an abnormal length also support detecting the presence of a length that is different from an expected or control length.

10 In some embodiments, aspects of the invention relate to estimating the length of a target nucleic acid (e.g., of a sub-target region within a target nucleic acid). In some embodiments, aspects of the invention relate to methods for estimating the length of a target nucleic acid by contacting the target nucleic acid with a plurality of detection probes under conditions that permit hybridization of the detection probes to the target
15 nucleic acid, wherein each detection probe is a polynucleotide that comprises a first arm that hybridizes to a first region of the target nucleic acid and a second arm that hybridizes to a second region of the target nucleic acid, wherein the first and second regions are on a common strand of the target nucleic acid, and wherein the nucleotide sequence between the 5' end of the first region and the 3' end of the second region is the nucleotide
20 sequence of a sub-target nucleic acid; and capturing a plurality of sub-target nucleic acids that are hybridized with the plurality of detection probes; and measuring the frequency of occurrence of a sub-target nucleic acid in the plurality of sub-target nucleic acids, wherein the frequency of occurrence of the sub-target nucleic acid in the plurality of sub-target nucleic acids is indicative of the length of the sub-target nucleic acid. It
25 should be appreciated that methods for estimating a nucleic acid length may involve comparing a capture efficiency for a target nucleic acid region to two or more reference efficiencies for known nucleic acid lengths in order to determine whether the target nucleic acid region is smaller, intermediate, or larger in size than the known control lengths. In some embodiments, a series of nucleic acids of known different lengths may
30 be used to provide a calibration curve for evaluating the length of a target nucleic acid region of interest.

In some embodiments, the capture efficiency of a target region suspected of having a deletion or insertion is determined and comparing the capture efficiency to a reference indicative of a normal capture efficiency. In some embodiments, the capture efficiency is lower than the reference capture efficiency. In some embodiments, the
5 subject is identified as having an insertion in the target region. In some embodiments, the capture efficiency is higher than the reference capture efficiency. In some embodiments, the subject is identified as having a deletion in the target region. In some embodiments, the subject is identified as being heterozygous for the insertion. In some embodiments, the subject is identified as being heterozygous for the deletion.

10 In some embodiments, aspects of the invention relate to capturing a sub-target nucleic acid (or a sequence of a sub-target nucleic acid). In some embodiments, a molecular inversion probe technique is used. In some embodiments, a molecular inversion probe is a single linear strand of nucleic acid that comprises a first targeting arm at its 5' end and a second targeting arm at its 3' end, wherein the first targeting arm
15 is capable of specifically hybridizing to a first region flanking one end of the sub-target nucleic acid, and wherein the second targeting arm is capable of specifically hybridizing to a second region flanking the other end of the sub-target nucleic acid on the same strand of the target nucleic acid. In some embodiments, the first and second targeting arms are between about 10 and about 100 nucleotides long. In some embodiments, the
20 first and second targeting arms are about 10-20, 20-30, 30-40, or 40-50 nucleotides long. In some embodiments, the first and second targeting arms are about 20 nucleotides long. In some embodiments, the first and second targeting arms have the same length. In some embodiments, the hybridization Tms of the first and second targeting arms are similar. In some embodiments, the hybridization Tms of the first and second targeting arms are
25 within 2-5 C of each other. In some embodiments, the hybridization Tms of the first and second targeting arms are identical. In some embodiments, the sub-target nucleic acid contains a nucleic acid repeat. In some embodiments, the nucleic acid repeat is a dinucleotide or trinucleotide repeat. In some embodiments, the sub-target nucleic acid contains 10-100 copies of the nucleic acid repeat in the absence of an abnormal increase
30 or decrease in nucleic acid repeats. In some embodiments, the sub-target nucleic acid is a region of the Fragile-X locus that contains a nucleic acid repeat. In some embodiments, one or both targeting arms hybridize to a region on the target nucleic acid

that is immediately adjacent to a region of nucleic acid repeats. In some embodiments, one or both targeting arms hybridize to a region on the target nucleic acid that is separated from a region of nucleic acid repeats by a region that does not contain any nucleic acid repeats. In some embodiments, the molecular inversion probe further
5 comprises a primer-binding region that can be used to sequence the captured sub-target nucleic acid and optionally the first and/or second targeting arm.

In some embodiments, aspects of the invention relate to evaluating the length of a plurality of different target nucleic acids in a biological sample. In some embodiments, the plurality of target nucleic acids are analyzed using a plurality of different molecular
10 inversion probes. In some embodiments, each different molecular inversion probe comprises a different pair of first and second targeting arms at each of the 3' and 5' ends. In some embodiments, each different molecular inversion probe comprises the same primer-binding sequence.

In some embodiments, aspects of the invention relate to analyzing nucleic acid
15 from a biological sample obtained from a subject. In some embodiments, the biological sample is a blood sample. In some embodiments, the biological sample is a tissue sample. In some embodiments, nucleic acids are analyzed in biological samples obtained from a plurality of different subjects. In some embodiments, nucleic acids from a biological sample are analyzed in multiplex reactions. It should be appreciated that a
20 biological sample contains a plurality of copies of a genome derived from a plurality of cells in the sample. Accordingly, a sample contains a plurality of independent copies of a target nucleic acid region of interest, the capture efficiency of which can be used to evaluate its size as described herein.

In some embodiments, methods are provided for diagnosing multiple heritable
25 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 or 2) 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
30 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, aspects of the invention relate to evaluating a nucleic acid capture efficiency by determining an amount of target nucleic acid that is captured (e.g., an amount of sub-target nucleic acid sequences that are captured). In some embodiments, the amount of target nucleic acid that is captured is determined by determining a number of independently captured target nucleic acid molecules (e.g., the amount of independently captured molecules that have the sequence of the sub-target region). In some embodiments, the amount of target nucleic acid that is captured is compared to a reference amount of captured nucleic acid. In some embodiments, the reference amount is determined by determining a number of independently captured molecules of a reference nucleic acid. In some embodiments, the reference nucleic acid is a nucleic acid of a different locus in the biological sample that is not suspected of containing a deletion or insertion. In some embodiments, the reference nucleic acid is a nucleic acid of known size and amount that is added to the capture reaction. As described herein, a number of independently captured nucleic acid sequences can be determined by contacting a nucleic acid sample with a preparation of a probe (e.g., a MIP probe as described herein). It should be appreciated that the preparation comprises a plurality of copies of the same probe and accordingly a plurality of independent copies of the target region may be captured by different probe molecules. The number of probe molecules that actually capture a sequence can be evaluated by determining an amount or number of captured molecules using any suitable technique. This number is a reflection of both the number of target molecules in the sample and the efficiency of capture of those target molecules, which in turn is related to the size of the target molecules as described herein. Accordingly, the capture efficiency can be evaluated by controlling for the abundance of the target nucleic acid, for example by comparing the number or amount of captured target molecules to an appropriate control (e.g., a known size and amount of control nucleic acid, or a different locus that should be present in the same amount in the biological sample and is not expected to contain any insertions or

deletions). It should be appreciated that other factors may affect the capture efficiency of a particular target nucleic acid region (e.g., the sequence of the region, the GC content, the presence of secondary structures, etc.). However, these factors also can be accounted for by using appropriate controls (e.g., known sequences having similar properties, the same sequences, other genomic sequences expected to be present in the biological sample at the same frequency, etc., or any combination thereof).

In some embodiments, aspects of the invention relate to identifying a subject as having an insertion or deletion in one or more alleles of a genetic locus if the capture efficiency for that genetic locus is statistically significantly different than a reference capture efficiency.

It should be appreciated that hybridization conditions used for any of the capture techniques described herein (e.g., MIP capture techniques) can be based on known hybridization buffers and conditions.

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 the detection of deletions or insertions is important.

Brief Description of the Drawings

Figure 1 is reproduced from Deng et al., Nature Biotech. 27:353-60, (see Supplemental Fig. 1G) and illustrates: An “analysis of the effects of probe characteristics on capturing efficiency. The statistical package R and its *effects* module were used for this analysis. A linear model was used, and each individual factor was assumed to be independent. The dashed lines represent a 95% confidence interval.

Shorter target sequences were captured with higher efficiency than long target sequences ($p < 2 \times 10^{-16}$).”

Figure 2 illustrates a non-limiting scheme of padlock capture of a region that includes both repetitive regions (thick wavy line) and the adjacent unique sequence (thick strait line). The regions of the probe are indicated with the targeting arms shown as regions “1” and “3.” An intervening region that may be, or include, a sequencing primer binding site is shown as “2.” After the padlock is circularized and amplified, it can be end-sequenced to obtain the sequence of the unique sequence, which represents the repetitive region of interest.

Figure 3 illustrates a non-limiting hypothetical relationship between target gap size and the relative number of reads of the repetitive region, which is measured by the ratio of the repeat “representative” reads vs. the “control” region reads. The unit of y-axis is arbitrary.

Description of the Invention

Aspects of the invention relate to detecting the presence of an insertion or deletion at a genomic locus without requiring the locus to be sequenced. Aspects of the invention are particularly useful for detecting an insertion or deletion in a nucleic acid region that contains high levels of sequence repeats. The presence of sequence repeats at a genetic locus is often associated with relatively high levels of polymorphism in a population due to insertions or deletions of one or more of the sequence repeats at the locus. The polymorphisms can be associated with diseases or predisposition to diseases (e.g., certain polymorphic alleles are recessive alleles associated with a disease or condition). However, the presence of sequence repeats often complicates the analysis of a genetic locus and increases the risk of errors when using sequencing techniques to determine the precise sequence and number of repeats at that locus.

Aspects of the invention relate to determining the size of a genetic locus by evaluating the capture frequency of a portion of that locus suspected of containing an insertion or deletion (e.g., due to the presence of sequence repeats) using a nucleic acid capture technique (e.g., a nucleic acid sequence capture technique based on molecular inversion probe technology). According to aspects of the invention, a statistically significant difference in capture efficiency for a genetic locus of interest in different

biological samples (e.g., from different subjects) is indicative of different relative lengths in those samples. It should be appreciated that the length differences may be at one or both alleles of the genetic locus. Accordingly, aspects of the invention may be used to identify polymorphisms regardless of whether biological samples being interrogated at

5 heterozygous or homozygous for the polymorphisms. According to aspects of the invention, subjects that contain one or more loci with an insertion or deletion can be identified by analyzing capture efficiencies for nucleic acids obtained from one or more biological samples using appropriate controls (e.g., capture efficiencies for known nucleic acid sizes, capture efficiencies for other regions that are not suspected of

10 containing an insertion or deletion in the biological sample(s), or predetermined reference capture efficiencies, or any combination thereof. However, it should be appreciated that aspects of the invention are not limited by the nature or presence of the control. In some embodiments, if a statistically significant variation in capture efficiency is detected, a subject may be identified as being at risk for a disease or condition

15 associated with insertions or deletions at that genetic locus. In some embodiments, the subject may be analyzed in greater detail in order to determine the precise nature of the insertion or deletion and whether the subject is heterozygous or homozygous for one or more insertions or deletions. For example, gel electrophoresis of an amplification (e.g., PCR) product of the locus, or Southern blotting, or any combination thereof can be used

20 as an orthogonal approach to verify the length of the locus. In some embodiments, a more exhaustive and detailed sequence analysis of the locus can be performed to identify the number and types of insertions and deletions. However, other techniques may be used to further analyze a locus identified as having an abnormal length according to aspects of the invention.

25 Accordingly, aspects of the invention relate to detecting abnormal nucleic acid lengths in genomic regions of interest. In some embodiments, the invention aims to estimate the size of genomic regions that are hard to be accessed, such as repetitive elements. However, it should be appreciated that methods of the invention do not require that the precise length be estimated. In some embodiments, it is sufficient to determine

30 that one or more alleles with abnormal lengths are present at a locus of interest (e.g., based on the detection of abnormal capture efficiencies).

In a non-limiting example, fragile X can be used to illustrate aspects of the invention where the size of trinucleotide repeats (genotype) is linked to a symptom (phenotype). However, it should be appreciated that fragile X is a non-limiting example and similar analyses may be performed for other genetic loci (e.g., independently or
5 simultaneously in multiplex analyses).

According to aspects of the invention, repetitive regions are surrounded by non-repetitive unique sequences, which can be used to amplify the repeat-containing regions using, for example, PCR or padlock (MIP)-based method.

In addition to the repetitive regions, a probe (e.g., a MIP or padlock probe) can be
10 designed to include at least a sequence that is sufficient to be uniquely identified in the genome (or target pool). After the probe is circularized and amplified, the amplicon can be end-sequenced so that the unique sequence can be identified and served as the “representative” of the repetitive region (Figure 2). Although capturing efficiency is overall negatively correlated with target length, different probe sequences may have
15 unique features. Therefore, multiple probes could be designed and tested so that an optimal one is chosen to be sensitive enough to differentiate repetitive sizes of roughly 0-150bp, 150-600bp, and beyond, which represent normal, premutation and full mutation of fragile X syndrome, respectively. However, it should be appreciated that other probe sizes and sequences can be designed, and optionally optimized, to distinguish a range of
20 repeat region size differences (e.g., length differences of about 3-30 bases, about 30-60 bases, about 60-90 bases, about 90-120 bases, about 120-150 bases, about 150-300 bases, about 300-600 bases, about 600-900 bases, or any intermediate or longer length difference). It should be appreciated that a length difference may be an increase in size or a decrease in size.

25 In some embodiments, an initial determination of an unexpected capture frequency is indicative of the presence of size difference. In some embodiments, an increase in capture frequency is indicative of a deletion. In some embodiments, a decrease in capture frequency is indicative of an insertion. However, it should be appreciated that depending on specific sequence parameters and the relative sizes of the
30 capture probes, the target region, and the deletions or insertions, a change in capture frequency can be associated with either an increase or decrease in target region length.

In some embodiments, the precise nature of the change can be determined using one or more additional techniques as described herein.

Use of molecular inversion probes has been demonstrated for detection of single nucleotide polymorphisms (Hardenbol et al. 2005 Genome Res 15:269-75) and for
5 preparative amplification of large sets of exons (Porreca et al. 2007 Nat Methods 4:931-6, Krishnakumar et al. 2008 Proc Natl Acad Sci USA 105:9296-301). In both cases, oligonucleotide probes are designed which have ends ('targeting arms') that hybridize up-stream and down-stream of the locus that is to be amplified. Following hybridization, polymerase fill-in and ligation reactions are performed to convert the hybridized probe to
10 a covalently-closed, circular molecule containing the desired target. PCR or rolling circle amplification plus exonuclease digestion of non-circularized material is performed to isolate and amplify the circular targets from the starting nucleic acid pool. Since one of the main benefits of the method is the potential for a high degree of multiplexing, generally thousands of targets are captured in a single reaction containing thousands of
15 probes.

Accordingly, in some aspects a MIP probe includes a linear nucleic acid strand that contains two hybridization sequences or targeting arms, one at each end of the linear probe, wherein each of the hybridization sequences is complementary to a separate
20 sequence on a the same strand of a target nucleic acid, and wherein these sequences on the target nucleic acid flank the two ends of the target nucleic acid sequence of interest. It should be appreciated that upon hybridization, the two ends of the probe are inverted with respect to each other in the sense that both 5' and 3' ends of the probe hybridize to the same strand to separate regions flanking the target region (as illustrated in Figure 2 for example).

25 In some embodiments, the hybridization sequences are between about 10-100 nucleotides long, for example between about 10-30, about 30-60, about 60-90, or about 20, about 30, about 40, or about 50 nucleotides long. However, other lengths may be used depending on the application. In some embodiments, the hybridization Tms of both targeting arms of a probe are designed or selected to be similar. In some embodiments,
30 the hybridization Tms of the targeting arms of a plurality of probes designed to capture different target regions are selected or designed to be similar so that they can be used together in a multiplex reaction. Accordingly, a typical size of a MIP probe prior to fill-

in is about 60-80 nucleotides long. However, other sizes can be used depending on the sizes of the targeting arms and any other sequences (e.g., primer binding or tag sequences) that are present in the MIP probe. In some embodiments, MIP probes are designed to avoid sequence-dependent secondary structures. In some embodiments, MIP probes are designed such that the targeting arms do not overlap with known polymorphic regions. In some embodiments, targeting arms that can be used for capturing the repeat region of the Fragile X locus can have the following sequences or complementary to these sequences depending on the strand that is captured.

left: CTCCGTTTCGGTTTCACTTC

right: ATCTTCTCTTCAGCCCTGCT

The typical captured size using these targeting arms is about 100 nucleotides in length (e.g., about 30 repeats of a tri-nucleotide repeat).

In some embodiments, the number of reads obtained for the “representative” of the repetitive region is not informative to estimate the target length because it is dependent on the total number of reads obtained. To overcome this, it is useful to include one or more probes that target other “control” regions where no or minimal polymorphism exists among populations. Because of the systematic consistency of capturing efficiency (Figure 2), the ratio of reads obtained for the repetitive “representative” to reads obtained for the control region(s) will be tuned using DNA with defined numbers of repeats. Ultimately, the ratio can serve as a measure of the repeat length (Figure 3).

In some embodiments, to better tell targets with similar size range apart, the whole repetitive region can be sequenced by making a shotgun library (e.g., by making a shotgun library from a captured sequence, for example a sequence captured using a MIP probe). The longer the repeat is, the more short reads of repeats will be obtained. Therefore, the target length will contribute twice to the relative number of “repetitive” reads, which will gain better resolution of differentiating targets. In some embodiments, the expectation is that the number of reads from any given repeat will be a direct function of the number of repeats present. However, in some embodiments, a Poisson sampling-induced spread may need to be considered and in some embodiments may be sufficiently large to limit the resolution.

Other non-limiting examples of loci at which insertions or deletions or repeat sequences may be associated with a disease or condition are provided in Tables 1 and 2. It should be appreciated that the presence of an abnormal length at any one or more of these loci may be evaluated according to aspects of the invention. In some embodiments, two or more of these loci or other loci may be evaluated in a single multiplex reaction using different probes designed to hybridize under the same reaction conditions to different target nucleic acid in a biological sample.

Table 1.

Polyglutamine (PolyQ) Diseases

Type	Gene	Normal/wildtype	Pathogenic
DRPLA (Dentatorubropallidoluysian atrophy)	ATN1 or DRPLA	6 - 35	49 - 88
HD (Huntington's disease)	HTT (Huntingtin)	10 - 35	35+
SBMA (Spinobulbar muscular atrophy or Kennedy disease)	Androgen receptor on the X chromosome.	9 - 36	38 - 62
SCA1 (Spinocerebellar ataxia Type 1)	ATXN1	6 - 35	49 - 88
SCA2 (Spinocerebellar ataxia Type 2)	ATXN2	14 - 32	33 - 77
SCA3 (Spinocerebellar ataxia Type 3 or Machado-Joseph disease)	ATXN3	12 - 40	55 - 86
SCA6 (Spinocerebellar ataxia Type 6)	CACNA1A	4 - 18	21 - 30
SCA7 (Spinocerebellar ataxia Type 7)	ATXN7	7 - 17	38 - 120
SCA17 (Spinocerebellar ataxia Type 17)	TBP	25 - 42	47 - 63

Table 2.

Non-Polyglutamine Diseases

Type	Gene	Codon	Normal/wildtype	Pathogenic
FRAXA (Fragile X syndrome)	<i>FMRI</i> , on the X-chromosome	CGG	6 - 53	230+
FXTAS (Fragile X-associated tremor/ ataxia syndrome)	<i>FMRI</i> , on the X-chromosome	CGG	6 - 53	55-200
FRAXE (Fragile XE mental retardation)	AFF2 or FMR2, on the X-chromosome	GCC	6 - 35	200+
FRDA (Friedreich's ataxia)	FXN or X25, (frataxin)	GAA	7 - 34	100+
DM (Myotonic dystrophy)	DMPK	CTG	5 - 37	50+
SCA8 (Spinocerebellar ataxia Type 8)	OSCA or SCA8	CTG	16 - 37	110 – 250
SCA12 (Spinocerebellar ataxia Type 12)	PPP2R2B or SCA12	CAG On 5' end	7 - 28	66 - 78

The following examples illustrate aspects and embodiments of the invention and are not intended to be limiting or restrictive. Many variations of the invention will become apparent to those skilled in the art upon review of this specification. The full scope of the invention should be determined by reference to the claims, along with their full scope of equivalents, and the specification, along with such variations.

Examples

MIP capture of 'target' locus and 'control' loci

In some embodiments, to accurately quantify the efficiency of target locus capture, at least three sets of control loci are captured in parallel that have *a priori* been shown to serve as proxies for various lengths of target locus. For example, if the target

locus is expected to have a length between 50 and 1000 bp, then sets of control loci having lengths of 50, 250, and 1000 bp could be captured (e.g. 20 loci per set should provide adequate protection from outliers), and their abundance digitally measured by sequencing. These loci should be chosen such that minimal variation in efficiency
5 between samples and on multiple runs of the same sample is observed (and are therefore 'efficiency invariant'). These will serve as 'reference' points that define the shape of the curve of abundance-vs-length. Determining the length of the target is then simply a matter of 'reading' the length from the appropriate point on the calibration curve.

In some embodiments, the statistical confidence one has in the estimate of target
10 length from this method is driven largely by three factors: 1) reproducibility / variation of the abundance data used to generate the calibration curve; 2) goodness of fit of the regression to the 'control' datapoints; 3) reproducibility of abundance data for the target locus being measured. Statistical bounds on 1) and 2) will be known in advance, having been measured during development of the assay. Additionally, statistical bounds on 3)
15 will be known in general in advance, since assay development should include adequate population sampling and measure of technical reproducibility. Standard statistical methods should be used to combine these three measures into a single *P* value for any given experimental measure of target abundance.

In some embodiments, given the set of calibration observations, and a linear
20 regression fit to that data, the regression can be used to predict the length value for *n* observations of the target locus whose length is unknown. First, choose an acceptable range for the confidence interval of the length estimate. For example, in the case of distinguishing "normal" (87-93bp) from "premutation" (165-600bp) potential cases of Fragile X, the goal is to measure length to sufficient precision to distinguish 93 bp from
25 165 bp. The predicted response value, computed when *n* observations is substituted into the equation for the regressed line, will have arbitrary precision. However, if for example a 95% confidence level is desired, that 95% confidence interval must be sufficiently short that it does not overlap both the "normal" and "premutation" length ranges. Continuing the example, if one calculates a length of 190 from *n*=400 MIP
30 observations, and based on the regression from calibration data, the 95% confidence interval is 190 +/- 20 bp, one can conclude the sample represents a "premutation" length with 95% certainty. Conversely, if the calibration data were less robust, error estimates

of the regression would be higher, leading to larger confidence intervals on the predicted response value. In some embodiments, if the 95% CI were calculated as 190 +/- 100 bp from n=400, one could not determine whether the predicted response value corresponds to a "normal" or "premutation" length.

- 5 In some embodiments, the confidence interval for a predicted response is calculated as:

The estimate for the response $\hat{y}$ is identical to the estimate for the mean of the response: $\hat{y} = b_0 + b_1x^*$. The confidence interval for the predicted value is given by $\hat{y} \pm t^* s_{\hat{y}}$, where $\hat{y}$ is the fitted value corresponding to x^* . The value t^* is the
10 *upper (1 - C)/2 critical value for the t(n - 2) distribution.*

In some embodiments, a technique for analyzing a locus of interest can involve the following steps.

- 15 ***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)
- 20 • 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.

- 25 ***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)
- 30 • 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.

5 ***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-Ilmn (platform-specific amplification sequence plus 'circle constant region'-specific sequence)
- 10 • 1.5 ul 10 uM CP-2-RA-Ilmn (platform-specific amplification sequence plus 'circle constant region'-specific sequence)
- 0.4 ul 25 mM dNTP mix (Invitrogen)
- 2.0 ul heat-inactivated exonuclease reaction mix
- 1.5 ul 10X SybrGreen (Invitrogen)
- 15 • 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
- 20 2. 95C for 30 sec
3. 58C for 60 sec
4. 72C for 60 sec
5. goto 2, N more times

25 ***Perform sequencing (e.g., next-generation sequencing) on sample for digital quantitation according to manufacturer's instructions (e.g., Illumina, ABI)***

It should be appreciated that the preceding examples are non-limiting and aspects of the invention may be implemented as described herein using alternative techniques and/or protocols that are available to one of ordinary skill in the art.

30 What is claimed is:

Claims

1. A method for determining whether a target nucleic acid has an abnormal length, the method comprising evaluating the capture efficiency of a target nucleic acid in a biological sample from a subject, wherein a capture efficiency that is different from
5 a reference capture efficiency is indicative of the presence, in the biological sample, of a target nucleic acid having an abnormal length.

2. The method of claim 1, comprising determining the capture efficiency of a target region suspected of having a deletion or insertion and comparing the capture
10 efficiency to a reference indicative of a normal capture efficiency.

3. The method of claim 1 or 2, wherein the capture efficiency is lower than the reference capture efficiency.

15 4. The method of claim 3, wherein the subject is identified as having an insertion in the target region.

5. The method of claim 1 or 2, wherein the capture efficiency is higher than the reference capture efficiency.
20

6. The method of claim 5, wherein the subject is identified as having a deletion in the target region.

7. The method of claim 4, wherein the subject is identified as being
25 heterozygous for the insertion.

8. The method of claim 6, wherein the subject is identified as being heterozygous for the deletion.

30 9. The method of any prior claim, wherein a sub-target nucleic acid is captured from the target nucleic acid using a molecular inversion probe.

10. The method of claim 9, wherein the molecular inversion probe comprises a first targeting arm at its 5' end and a second targeting arm at its 3' end, wherein the first targeting arm is capable of specifically hybridizing to a first region flanking one end of the sub-target nucleic acid, and wherein the second targeting arm is capable of
5 specifically hybridizing to a second region flanking the other end of the sub-target nucleic acid on the same strand of the target nucleic acid.

11. The method of claim 10, wherein the first and second targeting arms are between about 10 and about 100 nucleotides long.

10

12. The method of claim 11, wherein the first and second targeting arms are about 10-20, 20-30, 30-40, or 40-50 nucleotides long.

13. The method of claim 12, wherein the first and second targeting arms are
15 about 20 nucleotides long.

14. The method of any one of claims 10-13, wherein the first and second targeting arms have the same length.

20 15. The method of any one of claims 10-14, wherein the hybridization Tms of the first and second targeting arms are similar.

16. The method of claim 15, wherein the hybridization Tms of the first and second targeting arms are within 2-5 C of each other.

25

17. The method of claim 15, wherein the hybridization Tms of the first and second targeting arms are identical.

18. The method of any one of claims 9-17, wherein the sub-target nucleic acid
30 contains a nucleic acid repeat.

19. The method of claim 18, wherein the nucleic acid repeat is a dinucleotide or trinucleotide repeat.

20. The method of claim 18, wherein the sub-target nucleic acid contains 10-
5 100 copies of the nucleic acid repeat in the absence of an abnormal increase or decrease in nucleic acid repeats.

21. The method of claim 18, wherein the sub-target nucleic acid is a region of the Fragile-X locus that contains a nucleic acid repeat.

10

22. The method of any one of claims 10-17, wherein one or both targeting arms hybridize to a region on the target nucleic acid that is immediately adjacent to a region of nucleic acid repeats.

15 23. The method of any one of claims 10-17, wherein one or both targeting arms hybridize to a region on the target nucleic acid that is separated from a region of nucleic acid repeats by a region that does not contain any nucleic acid repeats.

24. The method of any one of claims 10-17, wherein the molecular inversion
20 probe further comprises a primer-binding region that can be used to sequence the captured sub-target nucleic acid and optionally the first and/or second targeting arm.

25. The method of any of the prior claims, wherein a plurality of different target nucleic acids are analyzed in a biological sample.

25

26. The method of claim 25, wherein the plurality of target nucleic acids are analyzed using a plurality of different molecular inversion probes.

27. The method of claim 26, wherein each different molecular inversion
30 probe comprises a different pair of first and second targeting arms at each of the 3' and 5' ends.

28. The method of claim 27, wherein each different molecular inversion probe comprises the same primer-binding sequence.

29. The method of any prior claim, wherein the biological sample is a blood
5 sample.

30. The method of any prior claim, wherein the biological sample is a tissue sample.

10 31. The method of any prior claim, wherein the capture efficiency is evaluated by determining an amount of target nucleic acid that is captured.

32. The method of claim 31, wherein the amount of target nucleic acid that is captured is determined by determining a number of independently captured target nucleic
15 acid molecules.

33. The method of claim 31 or 32, wherein the amount of target nucleic acid that is captured is compared to a reference amount of captured nucleic acid.

20 34. The method of claim 33, wherein the reference amount is determined by determining a number of independently captured molecules of a reference nucleic acid.

35. The method of claim 34, wherein the reference nucleic acid is a nucleic acid of a different locus in the biological sample that is not suspected of containing a
25 deletion or insertion.

36. The method of claim 34, wherein the reference nucleic acid is a nucleic acid of known size and amount that is added to the capture reaction.

30 37. The method of any prior claim, wherein a subject is identified as having an insertion or deletion in one or more alleles of the genetic locus if the capture efficiency is statistically significantly different that the reference capture efficiency.

38. A method for estimating the length of a target nucleic acid, the method comprising:

- 5 (i) contacting the target nucleic acid with a plurality of detection probes under conditions that permit hybridization of the detection probes to the target nucleic acid, wherein each detection probe is a polynucleotide that comprises a first arm that hybridizes to a first region of the target nucleic acid and a second arm that hybridizes to a second region of the target nucleic acid, wherein the first and second regions are on a common strand of the target
10 nucleic acid, and wherein the nucleotide sequence between the 5' end of the first region and the 3' end of the second region is the nucleotide sequence of a sub-target nucleic acid; and
- (ii) capturing a plurality of sub-target nucleic acids that are hybridized with
15 the plurality of detection probes; and
- (iii) measuring the frequency of occurrence of a sub-target nucleic acid in the plurality of sub-target nucleic acids, wherein the frequency of occurrence of the sub-target nucleic acid in the plurality of sub-target nucleic acids is indicative of the length of the sub-target nucleic acid.

Abstract

Aspects of the invention relate to methods for detecting abnormal nucleic acid lengths in a biological sample by using assays capable of detecting abnormal sequence capture efficiencies.

Figure 1

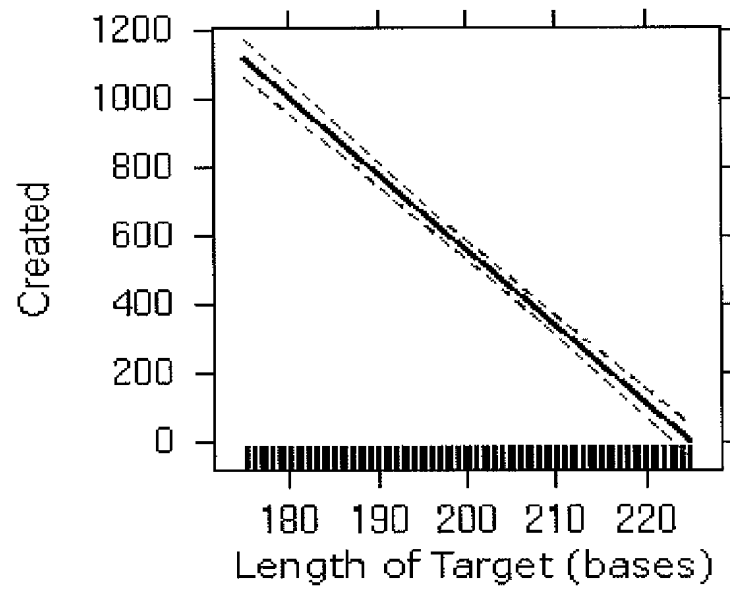


Figure 2

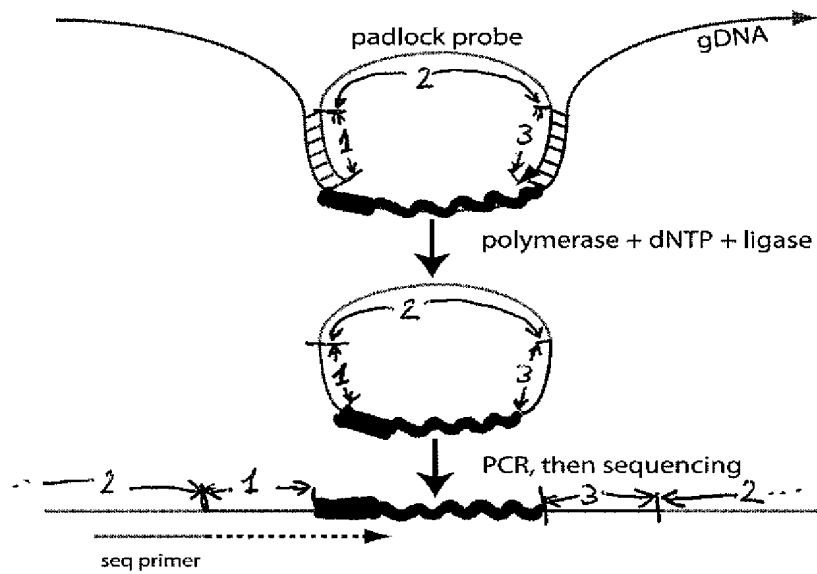


Figure 3

