TITLE OF THE INVENTION
HIGHLY MULTIPLEX PCR METHODS AND COMPOSITIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation of U.S. Utility Application Serial No. 13/683,604, filed November 21, 2012, which is a continuation-in-part of U.S. Utility Application Serial No. 13/300,235, filed Nov. 18, 2011, and is a continuation-in-part of U.S. Utility Application Serial No. 13/110,685, filed May 18, 2011. U.S. Utility Application Serial No. 13/110,685 claims the benefit of U.S. Provisional Application Serial No. 61/395,850, filed May 18, 2010; U.S. Provisional Application Serial No. 61/398,159, filed June 21, 2010; U.S. Provisional Application Serial No. 61/462,972, filed February 9, 2011; U.S. Provisional Application Serial No. 61/448,547, filed March 2, 2011; and U.S. Provisional Application Serial No. 61/516,996, filed April 12, 2011. U.S. Utility Application Serial No. 13/300,235 claims the benefit of U.S. Provisional Application Serial No. 61/571,248, filed June 23, 2011. U.S. Utility Application Serial No. 13/683,604 also claims the benefit of U.S. Provisional Application Serial No. 61/675,020, filed July 24, 2012.

This application is also a continuation of U.S. Utility Application Serial No. 13/780,022, filed February 28, 2013, which is a continuation-in-part of U.S. Utility Application Serial No. 13/683,604, filed November 21, 2012; PCT Application No. PCT/12/58578, filed Oct. 3, 2012; U.S. Utility Application Serial No. 13/335,043, filed Dec. 22, 2011; U.S. Utility Application Serial No. 13/300,235, filed November 18, 2011; and U.S. Utility Application Serial No. 13/110,685, filed May 18, 2011, and also claims the benefit of U.S. Provisional Application Serial No. 61/634,431, filed February 29, 2012. U.S. Utility Application Serial No. 13/683,604 claims the benefit of U.S. Provisional Application Serial No. 61/675,020, filed July 24, 2012. U.S. Utility Application Serial No. 13/335,043 claims the benefit of U.S. Provisional Application Serial No. 61/426,208, filed December 22, 2010. U.S. Utility Application Serial No. 13/110,685 claims the benefit of U.S. Provisional Application Serial No. 61/395,850, filed May 18, 2010; U.S. Provisional Application Serial No. 61/398,159, filed June 21, 2010; U.S. Provisional Application Serial No. 61/462,972, filed February 9, 2011; U.S. Provisional Application Serial No. 61/448,547, filed March 2, 2011; and U.S. Provisional Application Serial No. 61/516,996, filed April 12, 2011. U.S. Utility Application Serial No. 13/300,235 claims the benefit of U.S. Provisional Application Serial

No. 61/571,248, filed June 23, 2011 and U.S. Provisional Application Serial No. 61/542,508, filed October 3, 2011.

This application is also a continuation of U.S. Utility Application Serial number 14/225,356, filed March 25, 2014, which is a continuation of PCT Application No. PCT/US2012/058578, filed October 3, 2012, which claims the benefit of U.S. Provisional Application Serial No. 61/542,508, filed October 3, 2011, and U.S. Provisional Application Serial No. 61/683,331, filed August 15, 2012. PCT Application No. PCT/US2012/058578 is also a continuation-in-part of U.S. Patent Application Serial No. 13/300,235, filed November 18, 2011. U.S. Patent Application Serial No. 13/300,235 claims the benefit of U.S. Provisional Application Serial No. 61/571,248, filed June 23, 2011, and is a continuation-in-part of U.S. Patent Application Serial No. 13/110,685, filed May 18, 2011, which claims the benefit of U.S. Provisional Application Serial No. 61/395,850, filed May 18, 2010; U.S. Provisional Application Serial No. 61/398,159, filed June 21, 2010; U.S. Provisional Application Serial No. 61/462,972, filed February 9, 2011; U.S. Provisional Application Serial No. 61/448,547, filed March 2, 2011; and U.S. Provisional Application Serial No. 61/516,996, filed April 12, 2011. The entireties of all these applications are hereby incorporated by reference for the teachings therein.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

This work was supported by Grant No. 5R44HD60423-3, awarded by the National Institutes of Health. The U.S. government may have rights in any patent issuing on this application.

FIELD OF THE INVENTION

The present invention generally relates to methods and compositions for simultaneously amplifying multiple nucleic acid regions of interest in one reaction volume.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 2

BACKGROUND OF THE INVENTION

To increase assay throughput and allow more efficient use of nucleic acid samples, simultaneous amplification of many target nucleic acids in a sample of interest can be carried out by combining many oligonucleotide primers with the sample and then subjecting the sample to polymerase chain reaction (PCR) conditions in a process known in the art as multiplex PCR. Use of multiplex PCR can significantly simplify experimental procedures and shorten the time required for nucleic acid analysis and detection. However, when multiple pairs are added to the same PCR reaction, non-target amplification products may be generated, such as amplified primer dimers. The risk of generating such products increases as the number of primers increases. These non-target amplicons significantly limit the use of the amplified products for further analysis and/or assays. Thus, improved methods are needed to reduce the formation of non-target amplicons during multiplex PCR.

Improved multiplex PCR methods would be useful for a variety of application, such as Non-Invasive Prenatal Genetic Diagnosis (NPD). In particular, current methods of prenatal diagnosis can alert physicians and parents to abnormalities in growing fetuses. Without prenatal diagnosis, one in 50 babies is born with serious physical or mental handicap, and as many as one in 30 will have some form of congenital malformation. Unfortunately, standard methods have either poor accuracy, or involve an invasive procedure that carries a risk of miscarriage. Methods based on maternal blood hormone levels or ultrasound measurements are non-invasive, however, they also have low accuracies. Methods such as amniocentesis, chorion villus biopsy and fetal blood sampling have high accuracy, but are invasive and carry significant risks. Amniocentesis was performed in approximately 3% of all pregnancies in the US, though its frequency of use has been decreasing over the past decade and a half.

Normal humans have two sets of 23 chromosomes in every healthy, diploid cell, with one copy coming from each parent. Aneuploidy, a condition in a nuclear cell where the cell contains too many and/or too few chromosomes is believed to be responsible for a large percentage of failed implantations, miscarriages, and genetic diseases. Detection of chromosomal abnormalities can identify individuals or embryos with conditions such as Down syndrome, Klinefelter's syndrome, and Turner syndrome, among others, in addition to increasing the chances of a successful pregnancy. Testing for chromosomal abnormalities is especially important as the mother's age: between the ages of 35 and 40 it is estimated that at

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 3

least 40% of the embryos are abnormal, and above the age of 40, more than half of the embryos are abnormal.

It has recently been discovered that cell-free fetal DNA and intact fetal cells can enter maternal blood circulation. Consequently, analysis of this genetic material can allow early NPD. Improved methods are desired to improve the sensitivity and specificity and reduce the time and cost required for NPD.

SUMMARY OF THE INVENTION

In one aspect, the invention features methods of amplifying target loci in a nucleic acid sample. In some embodiments, the method involves (i) contacting the nucleic acid sample with a library of test primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture; and (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that include target amplicons. In some embodiments, the method also includes determining the presence or absence of at least one target amplicon (such as at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the target amplicons). In some embodiments, the method also includes determining the sequence of at least one target amplicon (such as at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the target amplicons).

In various embodiments of any of the aspects of the invention, at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified. In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the amplified products are target amplicons. In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the targeted loci are amplified. In various embodiments, less than 60, 50, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0.5, 0.25, 0.1, or 0.05% of the amplified products are primer dimers. In some embodiments, the library of test primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 test primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer that hybridize to the same target locus. In some embodiments, the library of test primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 individual test primers that hybridize to different target loci, wherein the individual primers are not part of primer pairs.

In various embodiments of any of the aspects of the invention, the concentration of each test primer is less than 100, 75, 50, 25, 10, 5, 2, or 1 nM. In various embodiments, the GC content of the test primers is between 30 to 80%, such as between 40 to 70% or 50 to 60%, inclusive. In some embodiments, the range of GC content (*e.g.,* the maximum GC content minus minimum GC content, such as 80% - 60% = a range of 20%) of the test primers is less than 30, 20, 10, or 5%. In some embodiments, the melting temperature (T_m) of the test primers is between 40 to 80 °C, such as 50 to 70 °C, 55 to 65 °C, or 57 to 60.5 °C, inclusive. In some embodiments, the range of melting temperatures of the test primers is less than 20, 15, 10, 5, 3, or 1 °C. In some embodiments, the length of the test primers is between 15 to 100 nucleotides, such as between 15 to 75 nucleotides, 15 to 40 nucleotides, 17 to 35 nucleotides, 18 to 30 nucleotides, 20 to 65 nucleotides, inclusive. In some embodiments, the test primers include a tag that is not target specific, such as a tag that forms an internal loop structure. In some embodiments, the tag is between two DNA binding regions. In various embodiments, the test primers include a 5' region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In various embodiments, the length of the 3' region is at least 7 nucleotides. In some embodiments, the length of the 3' region is between 7 and 20 nucleotides, such as between 7 to 15 nucleotides, or 7 to 10 nucleotides, inclusive. In various embodiments, the test primers include a 5' region that is not specific for a target locus (such as a tag or a universal primer binding site) followed by a region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In some embodiments, the range of the length of the test primers is less than 50, 40, 30, 20, 10, or 5 nucleotides. In some embodiments, the length of the target amplicons is between 50 and 100 nucleotides, such as between 60 and 80 nucleotides, or 60 to 75 nucleotides, inclusive. In some embodiments, the range of the length of the target amplicons is less than 50, 25, 15, 10, or 5 nucleotides.

In various embodiments of any of the aspects of the invention, the primer extension reaction conditions are polymerase chain reaction conditions (PCR). In various embodiments, the length of the annealing step is greater than 3, 5, 8, 10, or 15 minutes. In various embodiments, the length of the extension step is greater than 3, 5, 8, 10, or 15 minutes.

In various embodiments of any of the aspects of the invention, the test primers are used to simultaneously amplify at least 1,000 different target loci in a sample that includes maternal

DNA from the pregnant mother of a fetus and fetal DNA to determine the presence or absence of a fetal chromosome abnormality. In various embodiments, the method includes ligating a universal primer binding site to the DNA molecules in the sample; amplifying the ligated DNA molecules using at least 1,000 specific primers and a universal primer to produce a first set of amplified products; and amplifying the first set of amplified products using at least 1,000 pairs of specific primers to produce a second set of amplified products. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primer pairs are used.

In various embodiments of any of the aspects of the invention, the test primers are used to simultaneously amplify at least 1,000 different target loci in sample includes DNA from an alleged father of a fetus and to simultaneously amplify the target loci in a sample that includes maternal DNA from the pregnant mother of the fetus and fetal DNA to establish whether the alleged father is the biological father of the fetus.

In various embodiments of any of the aspects of the invention, the test primers are used to simultaneously amplify at least 1,000 different target loci in one cell or multiple cells from an embryo to determine the presence or absence of a chromosome abnormality. In various embodiments, cells from a set of two or more embryos are analyzed, and one embryo is selected for *in vitro* fertilization.

In various embodiments of any of the aspects of the invention, the test primers are used to simultaneously amplify at least 1,000 different target loci in a forensic nucleic acid sample. In various embodiments, the length of the annealing step is greater than 3, 5, 8, 10, or 15 minutes.

In various embodiments of any of the aspects of the invention, the method involves using the test primers to simultaneously amplify at least 1,000 different target loci in a control nucleic acid sample to produce a first set of target amplicons and to simultaneously amplify the target loci in a test nucleic acid sample to produce a second set of target amplicons; and comparing the first and second sets of target amplicons to determine whether a target locus is present in one sample but absent in the other, or whether a target locus is present at different levels in the control sample and the test sample. In various embodiments, the test sample is from an individual suspected of having a disease or phenotype of interest (such as cancer), or an increased risk for a disease or phenotype of interest; and wherein one or more of the target loci include a sequence (*e.g.*, a polymorphism or other mutation) associated with an increased

risk for the disease or phenotype of interest, or associated with the disease or phenotype of interest. In various embodiments, the method involves using the test primers to simultaneously amplify 1,000; different target loci in a control sample that includes RNA to produce a first set of target amplicons and to simultaneously amplify the target loci in a test sample that includes RNA to produce a second set of target amplicons; and comparing the first and second sets of target amplicons to determine the presence or absence of a difference in the RNA expression levels between the control sample and the test sample. In various embodiments, the RNA is mRNA. In various embodiments, the test sample is from an individual suspected of having a disease or phenotype of interest (such as cancer) or an increased risk for the disease or phenotype of interest (such as cancer); and wherein one or more of the target loci includes a sequence (*e.g.*, a polymorphism or other mutation) associated with an increased risk for the disease or phenotype of interest or associated with the disease or phenotype of interest. In some embodiments, the test sample is from an individual diagnosed with a disease or phenotype of interest (such as cancer); and wherein a difference in the RNA expression level between the control sample and test sample indicates a target locus includes a sequence (*e.g.*, a polymorphism or other mutation) associated with an increased or decreased risk for the disease or phenotype of interest.

In some embodiments of any of the aspects of the invention, the test primers are selected from a library of candidate primers based on one or more parameters, such as the selection of primers using any of the methods of the invention. In some embodiments, the test primers are selected from a library of candidate primers based at least in part on the ability of the candidate primers to form primer dimers.

In one aspect, the invention features methods of selecting test primers from a library of candidate primers. In various embodiments, the selection involves (i) calculating on a computer an undesirability score for most or all of the possible combinations of two candidate primers from the library, wherein each undesirability score is based at least in part on the likelihood of dimer formation between the two candidate primers; (ii) removing the candidate primer with the highest undesirability score from the library of candidate primers; and (iii) if the candidate primer removed in step (ii) is a member of a primer pair, then removing the other member of the primer pair from the library of candidate primers; and (iv) optionally repeating steps (ii) and (iii), thereby selecting a library of test primers. In some embodiments, the selection method is performed until the undesirability scores for the candidate primer

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 7

combinations remaining in the library are all equal to or below a minimum threshold. In some embodiments, the selection method is performed until the number of candidate primers remaining in the library is reduced to a desired number. In various embodiments, an undesirability score is calculated for at least 80, 90, 95, 98, 99, or 99.5% of the possible combinations of candidate primers in the library. In various embodiments, the candidate primers remaining in the library are capable of simultaneously amplifying at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci. In various embodiments, the method also includes (v) contacting a nucleic acid sample that includes target loci with the candidate primers remaining in the library to produce a reaction mixture; and (vi) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that includes target amplicons.

In one aspect, the invention features methods of selecting test primers from a library of candidate primers. In various embodiments, the selection of test primers are selected from a library of candidate primers involves (i) calculating on a computer an undesirability score for most or all of the possible combinations of two candidate primers from the library, wherein each undesirability score is based at least in part on the likelihood of dimer formation between the two candidate primers; (ii) removing from the library of candidate primers the candidate primer that is part of the greatest number of combinations of two candidate primers with an undesirability score above a first minimum threshold; (iii) if the candidate primer removed in step (ii) is a member of a primer pair, then removing the other member of the primer pair from the library of candidate primers; and (iv) optionally repeating steps (ii) and (iii), thereby selecting a library of test primers. In some embodiments, the selection method is performed until the undesirability scores for the candidate primer combinations remaining in the library are all equal to or below the first minimum threshold. In some embodiments, the selection method is performed until the number of candidate primers remaining in the library is reduced to a desired number. In various embodiments, an undesirability score is calculated for at least 80, 90, 95, 98, 99, or 99.5% of the possible combinations of candidate primers in the library. In various embodiments, the candidate primers remaining in the library are capable of simultaneously amplifying at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci. In various embodiments, the method also includes (v) contacting a nucleic acid sample that includes target loci with the candidate primers remaining in the library to produce a reaction mixture; and (vi) subjecting the reaction

mixture to primer extension reaction conditions to produce amplified products that includes target amplicons.

In various embodiments of any of the aspects of the invention, the selection method involves further reducing the number of candidate primers remaining in the library by decreasing the first minimum threshold used in step (ii) to a lower second minimum threshold and optionally repeating steps (ii) and (iii). In some embodiments, the selection method involves increasing the first minimum threshold used in step (ii) to a higher second minimum threshold and optionally repeating steps (ii) and (iii). In some embodiments, the selection method is performed until the undesirability scores for the candidate primer combinations remaining in the library are all equal to or below the second minimum threshold, or until the number of candidate primers remaining in the library is reduced to a desired number.

In various embodiments of any of the aspects of the invention, the method involves, prior to step (i), identifying or selecting primers that hybridize to the target loci. In some embodiments, multiple primers (or primer pairs) hybridize to the same target locus, and the selection method is used to select a one primer (or one primer pair) for this target locus based on one or more parameters. In various embodiments, the method involves, prior to step (ii), removing a primer pair from the library that produces a target amplicon that overlaps with a target amplicon produced by another primer pair. In various embodiments, a candidate primer is selected out of a group of two or more candidate primers with equal undesirability scores for removal from the library of candidate primers based on one or more other parameters. In some embodiments, the candidate primers remaining in the library are used as a library of test primers in any of the methods of the invention. In some embodiments, the resulting library of test primers includes any of the primer libraries of the invention.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, specificity of the candidate primer for the target locus, size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 9

heterozygosity rate of the target locus, specificity of the candidate primer for the target locus; size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the test primers are used to simultaneously amplify at least 1,000 different target loci in a sample that includes maternal DNA from the pregnant mother of a fetus and fetal DNA to determine the presence or absence of a fetal chromosome abnormality. In various embodiments, the method includes ligating a universal primer binding site to the DNA molecules in the sample; amplifying the ligated DNA molecules using at least 1,000 specific primers and a universal primer to produce a first set of amplified products; and amplifying the first set of amplified products using at least 1,000 pairs of specific primers to produce a second set of amplified products. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primer pairs are used. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, specificity of the candidate primer for the target locus; size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the test primers are used to simultaneously amplify at least 1,000 different target loci in sample includes DNA from an alleged father of a fetus and to simultaneously amplify the target loci in a sample that includes maternal DNA from the pregnant mother of a fetus and fetal DNA to establish whether the alleged father is the biological father of the fetus. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, specificity of the candidate primer for the target locus; size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the test primers are used to simultaneously amplify at least 1,000 different target loci in one cell or multiple cells from an embryo to determine the presence or absence of a

chromosome abnormality. In various embodiments, cells from a set of two or more embryos are analyzed, and one embryo is selected for *in vitro* fertilization. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, specificity of the candidate primer for the target locus; size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the test primers are used to simultaneously amplify at least 1,000 different target loci in a forensic nucleic acid sample. In various embodiments, the length of the annealing step is greater than 3, 5, 8, 10, or 15 minutes. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, specificity of the candidate primer for the target locus, size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the method involves using the test primers to simultaneously amplify at least 1,000 different target loci in a control nucleic acid sample to produce a first set of target amplicons and to simultaneously amplify the target loci in a test nucleic acid sample to produce a second set of target amplicons; and comparing the first and second sets of target amplicons to determine whether a target locus is present in one sample but absent in the other, or whether a target locus is present at different levels in the control sample and the test sample. In various embodiments, the test sample is from an individual suspected of having a disease or phenotype of interest, or an increased risk for a disease or phenotype of interest; and wherein one or more of the target loci include a sequence (*e.g.*, a polymorphism) at the target locus associated with an increased risk for the disease or phenotype of interest, or associated with the disease or

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 11

phenotype of interest. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In various embodiments of any of the aspects of the invention, the undesirability scores are based at least in part on one or more parameters selected from the group consisting of heterozygosity rate of the target locus, disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, specificity of the candidate primer for the target locus, size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the target amplicon; and the method involves using the test primers to simultaneously amplify 1,000; different target loci in a control sample that includes RNA to produce a first set of target amplicons and to simultaneously amplify the target loci in a test sample that includes RNA to produce a second set of target amplicons; and comparing the first and second sets of target amplicons to determine the presence or absence of a difference in the RNA expression levels between the control sample and the test sample. In various embodiments, the RNA is mRNA. In various embodiments, the test sample is from an individual suspected of having a disease or phenotype of interest (such as cancer) or an increased risk for the disease or phenotype of interest (such as cancer); and wherein one or more of the target loci includes a sequence (*e.g.*, a polymorphism or other mutation) associated with an increased risk for the disease or phenotype of interest or associated with the disease or phenotype of interest. In some embodiments, the test sample is from an individual diagnosed with a disease or phenotype of interest (such as cancer); and wherein a difference in the RNA expression level between the control sample and test sample indicates a target locus includes a sequence (*e.g.*, a polymorphism or other mutation) associated with an increased or decreased risk for the disease or phenotype of interest. In various embodiments, at least 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified.

In one aspect, the invention features libraries of primers. In some embodiments, the primers are selected from a library of candidate primers using any of the methods of the invention. In some embodiments, the library includes primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci. In some embodiments, the library includes primers that simultaneously amplify at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000;

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 12

40,000; 50,000; 75,000; or 100,000 different target loci. In some embodiments, the library includes primers that simultaneously amplify at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci such that less than 60, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0.5, 0.25, 0.1, or 0.05% of the amplified products are primer dimers. In some embodiments, the library includes primers that simultaneously amplify 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the amplified products are target amplicons. In some embodiments, the library includes primers that simultaneously amplify target loci such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the targeted loci out of 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci are amplified. In some embodiments, the library of primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer where each pair of test primers hybridize to a target locus. In some embodiments, the library of primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs.

In various embodiments of any of the aspects of the invention, the concentration of each primer is less than 100, 75, 50, 25, 10, 5, 2, or 1 nM. In various embodiments, the GC content of the primers is between 30 to 80%, such as between 40 to 70% or 50 to 60%, inclusive. In some embodiments, the range of GC content of the primers is less than 30, 20, 10, or 5%. In some embodiments, the melting temperature of the primers is between 40 to 80 °C, such as 50 to 70 °C, 55 to 65 °C, or 57 to 60.5 °C, inclusive. In some embodiments, the range of melting temperature of the primers is less than 15, 10, 5, 3, or 1 °C. In some embodiments, the length of the primers is between 15 to 100 nucleotides, such as between 15 to 75 nucleotides, 15 to 40 nucleotides, 17 to 35 nucleotides, 18 to 30 nucleotides, or 20 to 65 nucleotides, inclusive. In some embodiments, the primers include a tag that is not target specific, such as a tag that forms an internal loop structure. In some embodiments, the tag is between two DNA binding regions. In various embodiments, the primers include a 5' region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In various

embodiments, the length of the 3' region is at least 7 nucleotides. In some embodiments, the length of the 3' region is between 7 and 20 nucleotides, such as between 7 to 15 nucleotides, or 7 to 10 nucleotides, inclusive. In various embodiments, the primers include a 5' region that is not specific for a target locus (such as another tag or a universal primer binding site) followed by a region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In some embodiments, the range of the length of the primers is less than 50, 40, 30, 20, 10, or 5 nucleotides. In some embodiments, the length of the target amplicons is between 50 and 100 nucleotides, such as between 60 and 80 nucleotides, or 60 to 75 nucleotides, inclusive. In some embodiments, the range of the length of the target amplicons is less than 50, 25, 15, 10, or 5 nucleotides.

In one aspect, the invention provides kits that include any of the primer libraries of the invention for amplifying target loci in a nucleic acid sample. In some embodiments, the kit includes instructions for using the library to amplify the target loci.

In one aspect, the invention features methods for determining a ploidy status of chromosome in a gestating fetus. In some embodiments, the method involves contacting a nucleic acid sample with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci to produce a reaction mixture; wherein the nucleic acid sample includes maternal DNA from the mother of the fetus and fetal DNA from the fetus. In some embodiments, the reaction mixture is subjected to primer extension reaction conditions to produce amplified products; the amplified products are measured with a high throughput sequencer to produce sequencing data; allele counts at the polymorphic loci are calculated on a computer based on the sequencing data; a plurality of ploidy hypotheses each pertaining to a different possible ploidy state of the chromosome are created on a computer; a joint distribution model for the expected allele counts at the polymorphic loci on the chromosome is built on a computer for each ploidy hypothesis; a relative probability of each of the ploidy hypotheses is determined on a computer using the joint distribution model and the allele counts; and the ploidy state of the fetus is called by selecting the ploidy state corresponding to the hypothesis with the greatest probability.

In one aspect, the invention features methods for determining a ploidy status of a chromosome in a gestating fetus. In an embodiment a method for determining a ploidy status

of a chromosome in a gestating fetus includes obtaining a first sample of DNA that comprises maternal DNA from the mother of the fetus and fetal DNA from the fetus, preparing the first sample by isolating the DNA so as to obtain a prepared sample, measuring the DNA in the prepared sample at a plurality of polymorphic loci on the chromosome, calculating, on a computer, allele counts at the plurality of polymorphic loci from the DNA measurements made on the prepared sample, creating, on a computer, a plurality of ploidy hypotheses each pertaining to a different possible ploidy state of the chromosome, building, on a computer, a joint distribution model for the expected allele counts at the plurality of polymorphic loci on the chromosome for each ploidy hypothesis, determining, on a computer, a relative probability of each of the ploidy hypotheses using the joint distribution model and the allele counts measured on the prepared sample, and calling the ploidy state of the fetus by selecting the ploidy state corresponding to the hypothesis with the greatest probability.

In one aspect, the invention features methods of testing for an abnormal distribution of a chromosome in a sample that includes a mixture of maternal and fetal DNA. In some embodiments, the method involves (i) contacting the sample with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture; wherein the target loci are from a plurality of different chromosomes; and wherein the plurality of different chromosomes include at least one first chromosome suspected of having an abnormal distribution in the sample and at least one second chromosome presumed to be normally distributed in the sample; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products; (iii) sequencing the amplified products to obtain a plurality of sequence tags aligning to the target loci; wherein the sequence tags are of sufficient length to be assigned to a specific target locus; (iv) assigning on a computer the plurality of sequence tags to their corresponding target loci; (v) determining on a computer a number of sequence tags aligning to the target loci of the first chromosome and a number of sequence tags aligning to the target loci of the second chromosome; and (vi) comparing on a computer the numbers from step (v) to determine the presence or absence of an abnormal distribution of the first chromosome.

In one aspect, the invention provides methods for detecting the presence or absence of a fetal aneuploidy. In some embodiments, the method involves (i) contacting a sample that includes a mixture of maternal and fetal DNA with a library of primers that simultaneously

hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different non-polymorphic target loci to produce a reaction mixture; wherein the target loci are from a plurality of different chromosomes; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that includes target amplicons; (iii) quantifying on a computer a relative frequency of the target amplicons from the first and second chromosomes of interest; (iv) comparing on a computer the relative frequency of the target amplicons from the first and second chromosomes of interest; and (v) identifying the presence or absence of an aneuploidy based on the compared relative frequencies of the first and second chromosome of interest. In some embodiments, the first chromosome is a chromosome suspected of being euploid. In some embodiments, the second chromosome is a chromosome suspected of being aneuploidy.

In one aspect, a method is disclosed for determining presence or absence of fetal aneuploidy in a maternal tissue sample comprising fetal and maternal genomic DNA, the method including (a) obtaining a mixture of fetal and maternal genomic DNA from said maternal tissue sample, (b) conducting massively parallel DNA sequencing of DNA fragments randomly selected from the mixture of fetal and maternal genomic DNA of step (a) to determine the sequence of said DNA fragments, (c) identifying chromosomes to which the sequences obtained in step (b) belong, (d) using the data of step (c) to determine an amount of at least one first chromosome in said mixture of maternal and fetal genomic DNA, wherein said at least one first chromosome is presumed to be euploid in the fetus, (e) using the data of step (c) to determine an amount of a second chromosome in said mixture of maternal and fetal genomic DNA, wherein said second chromosome is suspected to be aneuploid in the fetus, (f) calculating the fraction of fetal DNA in the mixture of fetal and maternal DNA, (g) calculating an expected distribution of the amount of the second target chromosome if the second target chromosome is euploid, using the number in step (d), (h) calculating an expected distribution of the amount of the second target chromosome if the second target chromosome is aneuploid, using the first number is step (d) and the calculated fraction of fetal DNA in the mixture of fetal and maternal DNA in step (f), and (i) using a maximum likelihood or maximum a posteriori approach to determine whether the amount of the second chromosome as determined in step (e) is more likely to be part of the distribution calculated in step (g) or the distribution calculated in step (h); thereby indicating the presence or absence of a fetal aneuploidy.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 16

In various embodiments of any of the aspects of the invention, the method also includes obtaining genotypic data from one or both parents of the fetus. In some embodiments, obtaining genotypic data from one or both parents of the fetus includes preparing the DNA from the parents where the preparing comprises preferentially enriching the DNA at the plurality of polymorphic loci to give prepared parental DNA, optionally amplifying the prepared parental DNA, and measuring the parental DNA in the prepared sample at the plurality of polymorphic loci.

In various embodiments of any of the aspects of the invention, building a joint distribution model for the expected allele count probabilities of the plurality of polymorphic loci on the chromosome is done using the obtained genetic data from the one or both parents. In some embodiments, the sample (*e.g.*, the first sample) has been isolated from maternal plasma and where the obtaining genotypic data from the mother is done by estimating the maternal genotypic data from the DNA measurements made on the prepared sample.

In one aspect, a diagnostic box is disclosed for helping to determine a ploidy status of a chromosome in a gestating fetus where the diagnostic box is capable of executing the preparing and measuring steps of any of the methods of the invention.

In various embodiments of any of the aspects of the invention, the allele counts are probabilistic rather than binary. In some embodiments, measurements of the DNA in the prepared sample at the plurality of polymorphic loci are also used to determine whether or not the fetus has inherited one or a plurality of disease linked haplotypes.

In various embodiments of any of the aspects of the invention, building a joint distribution model for allele count probabilities is done by using data about the probability of chromosomes crossing over at different locations in a chromosome to model dependence between polymorphic alleles on the chromosome. In some embodiments, building a joint distribution model for allele counts and the step of determining the relative probability of each hypothesis are done using a method that does not require the use of a reference chromosome.

In various embodiments of any of the aspects of the invention, determining the relative probability of each hypothesis makes use of an estimated fraction of fetal DNA in the prepared sample. In some embodiments, the DNA measurements from the prepared sample used in calculating allele count probabilities and determining the relative probability of each hypothesis comprise primary genetic data. In some embodiments, selecting the ploidy state

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 17

corresponding to the hypothesis with the greatest probability is carried out using maximum likelihood estimates or maximum a posteriori estimates.

In various embodiments of any of the aspects of the invention, calling the ploidy state of the fetus also includes combining the relative probabilities of each of the ploidy hypotheses determined using the joint distribution model and the allele count probabilities with relative probabilities of each of the ploidy hypotheses that are calculated using statistical techniques taken from a group consisting of a read count analysis, comparing heterozygosity rates, a statistic that is only available when parental genetic information is used, the probability of normalized genotype signals for certain parent contexts, a statistic that is calculated using an estimated fetal fraction of the sample (*e.g.*, the first sample) or the prepared sample, and combinations thereof.

In various embodiments of any of the aspects of the invention, a confidence estimate is calculated for the called ploidy state. In some embodiments, the method also includes taking a clinical action based on the called ploidy state of the fetus, wherein the clinical action is selected from one of terminating the pregnancy or maintaining the pregnancy.

In various embodiments of any of the aspects of the invention, the method may be performed for fetuses at between 4 and 5 weeks gestation; between 5 and 6 weeks gestation; between 6 and 7 weeks gestation; between 7 and 8 weeks gestation; between 8 and 9 weeks gestation; between 9 and 10 weeks gestation; between 10 and 12 weeks gestation; between 12 and 14 weeks gestation; between 14 and 20 weeks gestation; between 20 and 40 weeks gestation; in the first trimester; in the second trimester; in the third trimester; or combinations thereof.

In various embodiments of any of the aspects of the invention, a report displaying a determined ploidy status of a chromosome in a gestating fetus generated using the method. In some embodiments, a kit is disclosed for determining a ploidy status of a target chromosome in a gestating fetus designed to be used with any of the methods of the invention, the kit including a plurality of inner forward primers and optionally the plurality of inner reverse primers, where each of the primers is designed to hybridize to the region of DNA immediately upstream and/or downstream from one of the polymorphic sites on the target chromosome, and optionally additional chromosomes, where the region of hybridization is separated from the polymorphic site by a small number of bases, where the small number is selected from the

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 18

group consisting of 1, 2, 3, 4, 5, 6 to 10, 11 to 15, 16 to 20, 21 to 25, 26 to 30, 31 to 60, and combinations thereof.

In one aspect, the invention features methods for establishing whether an alleged father is the biological father of a fetus that is gestating in a pregnant mother. In some embodiments the method involves, (i) simultaneously amplifying a plurality of polymorphic loci that includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci on genetic material from the alleged father to produce a first set of amplified products; (ii) simultaneously amplifying the corresponding plurality of polymorphic loci on a mixed sample of DNA originating from a blood sample from the pregnant mother to produce a second set of amplified products; wherein the mixed sample of DNA includes fetal DNA and maternal DNA; (iii) determining on a computer the probability that the alleged father is the biological father of the fetus using genotypic measurements based on the first and second sets of amplified products; and (iv) establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus. In various embodiments, the method further includes simultaneously amplifying the corresponding plurality of polymorphic loci on genetic material from the mother to produce a third set of amplified products; wherein the probability that the alleged father is the biological father of the fetus is determined using genotypic measurements based on the first, second, and third sets of amplified products.

In one aspect, the invention provides methods of estimating relative likelihoods that each embryo from a set of embryos will develop as desired. In some embodiments, the method involves contacting a sample from each embryo with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture for each embryo, wherein the samples are each derived from one or more cells from an embryo. In some embodiments, each reaction mixture is subjected to primer extension reaction conditions to produce amplified products. In some embodiments, the method includes determining on a computer one or more characteristics of at least one cell from each embryo based on the amplified products; and estimating on a computer the relative likelihoods that each embryo will develop as desired, based on the one or more characteristics of the at least one cell for each embryo.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 19

In one aspect, the invention features methods of measuring the amount of two or more target loci in a nucleic acid sample. In some embodiments, the method involves (i) using PCR to amplify a nucleic acid sample that includes a first standard locus, a second standard locus, a first target locus, and a second target locus to form amplified products; wherein the first standard locus and the first target locus have the same number of nucleotides but have a sequence that differs at one or more nucleotides; and wherein the second standard locus and the second target locus have the same number of nucleotides but have a sequence that differs at one or more nucleotides; (ii) sequencing the amplified products to determine a standard ratio that compares the relative amount of the amplified first standard locus compared to the amplified second standard locus; wherein the standard ratio indicates the difference in PCR efficiency for the amplification of the first standard locus and the second standard locus; (iii) determining a target ratio that compares the relative amount of the amplified first target locus compared to the amplified second target locus; and (iv) adjusting the target ratio from step (iii) based on the standard ratio from step (ii) to determine the relative amount of the first target locus and the second target locus in the sample. In various embodiments, the method involves determining the absolute amount of the first target locus and the second target locus in the sample. In various embodiments, the method further includes determining the presence or absence of a target locus (*e.g*, at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci) in the sample. In various embodiments, the method involves using any of the primer libraries of the invention. In various embodiments, the method involves simultaneously amplifying 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci.

In one aspect, the invention features methods of quantitatively measuring a plurality of genetic targets in a sample for analysis. In some embodiments, the method includes (i) mixing genetic material derived from the sample for analysis with a plurality of target specific amplification reagents, and a plurality of standard sequences corresponding to the target specific amplification reagent targets; (ii) amplifying target regions of the genetic material and the standard sequences to produce target amplicons and standard sequence amplicons; and (iii) measuring the quantity of target amplicons and standard sequence amplicons produced. In some embodiments, the genetic material is present in a genetic library. In some embodiments, the genetic targets are polymorphic loci (such as SNPs). In some embodiments, the measuring

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 20

of quantity is achieved by counting sequences. In some embodiments, the method further includes determining the estimated copy number of at least one chromosome in a sample from which the genetic library was derived, wherein the determination involves comparing the number of sequence reads of a target amplicon with the number of sequence reads of a standard amplicon. In some embodiments, the standard sequences and the genetic library include universal priming sites cable of being primed by the same primer. In some embodiments, the mixing step includes at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target specific amplification reagents and at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 standard sequences. In various embodiments, the method involves using any of the primer libraries of the invention. In various embodiments, the method involves simultaneously amplifying 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target regions. In some embodiments, the relative amounts of each of the standard sequences are known. In some embodiments, the relative amounts of each of the sequences is has been calibrated with respect to a reference genome. In some embodiments, the sample for analysis includes a mixture of fetal and maternal genomes. In some embodiments, the sample for analysis is derived from the blood of a pregnant woman or derived from blood plasma. In some embodiments, the reference genome has at least one aneuploidy, such as an aneuploidy at chromosome 13, 18, 21, X, or Y. In some embodiments, the reference genome is diploid.

In one aspect, the invention features a mixture that includes a plurality of genetic standard sequences, wherein the relative amount of each genetic standard sequence in the mixture has been determined by calibration to a reference genome. In various embodiments, the mixture includes at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 genetic standard sequences. In various embodiments, the genetic standard sequences include a first universal priming site, a second universal priming site, a first target specific priming site, a second target specific priming site, and a marker sequence located between the first and second target specific priming sites, wherein the first target specific site and the second target specific priming site are located between the first and second universal priming sites. In various embodiments, the calibration involves using any of the primer libraries of the invention. In various embodiments, the calibration involves simultaneously amplifying 1,000; 2,000; 5,000; 7,500; 10,000; 20,000;

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 21

25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target regions. In some embodiments, the reference genome has at least one aneuploidy, such as an aneuploidy at chromosome 13, 18, 21, X, or Y. In some embodiments, the reference genome is diploid.

In one aspect, the invention features methods of producing a set of calibrated genetic standard sequences. In some embodiments, the method includes (i) forming an amplification reaction mixture that includes a genetic library prepared from a reference genome, a plurality of target-specific amplification primer reagent sets, and a plurality of genetic standard sequences corresponding to the target specific amplification reagent sets, (ii) amplifying the genetic library and the genetic standard sequences to produce amplicons from the target sequences and amplicons from the genetic standard sequences, (iii) measuring the quantity of the amplicons from the target sequences and amplicons from the genetic standard sequences, and (iv) determining the relative amount of each of genetic standard sequences with respect to each other, whereby the plurality of genetic standard sequences is calibrated. In various embodiments, at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 genetic standard sequences are used. In various embodiments, the method involves using any of the primer libraries of the invention. In various embodiments, the method involves simultaneously amplifying 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different sequences. In some embodiments, the reference genome has at least one aneuploidy, such as an aneuploidy at chromosome 13, 18, 21, X, or Y. In some embodiments, the reference genome is diploid.

In one aspect, the invention provides a set of genetic standard sequences that have been calibrated according to any of the methods of the invention. In one aspect, the invention provides a set of genetic standard sequences that may be calibrated either before, during or after the method is performed.

In one aspect, the invention features methods of measuring the number of copies of a gene of interest having at least one allele that has a deletion. In some embodiments, the method includes (i) mixing genetic material derived from a sample for analysis with an amplification reagent specific for the gene of interest and not capable of significantly amplifying the deletion comprising allele of the gene of interest, a standard sequence corresponding to gene of interest, an amplification reagent specific for a reference sequence, and a standard sequence corresponding to the reference sequence; (ii) amplifying the gene

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 22

sequence of interest, the standard sequence corresponding to the gene of interest, the reference sequence, and the standard sequence corresponding to the reference sequence to produce gene of interest amplicons, reference sequence amplicons, and standard sequence amplicons; and (iii) measuring the quantity of target amplicons and standard sequence amplicons produced. In some embodiments, the measuring of quantity is achieved by counting sequence reads. In some embodiments, the method further includes determining the estimated copy number of at least one chromosome in the sample from which the genetic library was derived, wherein the determination involves comparing the number of sequences of target amplicons with the number of sequences of a standard amplicons. In some embodiments, the standard sequences and the genetic library include universal priming sites capable of being primed by the same primer. In some embodiments, the relative amounts of each of the sequences have been calibrated with respect to reference genome. In various embodiments, at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 genetic standard sequences are used. In various embodiments, the method involves using any of the primer libraries of the invention. In various embodiments, the method involves simultaneously amplifying 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target regions. In some embodiments, the reference genome is diploid. In some embodiments, the sample for analysis is derived from blood.

In some embodiments of any of the aspects of the invention, preferentially enriching the DNA in the sample (*e.g.*, the first sample) at the target loci (*e.g.*, the plurality of polymorphic loci) includes obtaining a plurality of pre-circularized probes where each probe targets one of the loci (*e.g.*, polymorphic loci), where the 3' and 5' end of the probes are preferably designed to hybridize to a region of DNA that is separated from the polymorphic site of the locus by a small number of bases, where the small number is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 to 25, 26 to 30, 31 to 60, or a combination thereof, hybridizing the pre-circularized probes to DNA from the sample (*e.g.*, the first sample), filling the gap between the hybridized probe ends using DNA polymerase, circularizing the pre-circularized probe, and amplifying the circularized probe.

In some embodiments of any of the aspects of the invention, the preferentially enriching the DNA at the target loci (*e.g.*, the plurality of polymorphic loci) includes obtaining a plurality of ligation-mediated PCR probes where each PCR probe targets one of the target

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 23

loci (*e.g.*, the polymorphic loci), and where the upstream and downstream PCR probes are designed to hybridize to a region of DNA on one strand of DNA that is preferably separated from the polymorphic site of the locus by a small number of bases, where the small number is 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 to 25, 26 to 30, 31 to 60, or a combination thereof, hybridizing the ligation-mediated PCR probes to the DNA from the sample (*e.g.*, the first sample), filling the gap between the ligation-mediated PCR probe ends using DNA polymerase, ligating the ligation-mediated PCR probes, and amplifying the ligated ligation-mediated PCR probes.

In some embodiments of various aspects of the invention, preferentially enriching the DNA at the target loci (*e.g.*, plurality of polymorphic loci) includes obtaining a plurality of hybrid capture probes that target the loci (*e.g.*, the polymorphic loci), hybridizing the hybrid capture probes to the DNA in the sample (*e.g.*, the first sample) and physically removing some or all of the unhybridized DNA from the sample (*e.g.*, the first sample) of DNA.

In some embodiments of any of the aspects of the invention, the hybrid capture probes are designed to hybridize to a region that is flanking but not overlapping the polymorphic site. In some embodiments, the hybrid capture probes are designed to hybridize to a region that is flanking but not overlapping the polymorphic site, and where the length of the flanking capture probe may be selected from the group consisting of less than about 120 bases, less than about 110 bases, less than about 100 bases, less than about 90 bases, less than about 80 bases, less than about 70 bases, less than about 60 bases, less than about 50 bases, less than about 40 bases, less than about 30 bases, and less than about 25 bases. In some embodiments, the hybrid capture probes are designed to hybridize to a region that overlaps the polymorphic site, and where the plurality of hybrid capture probes comprise at least two hybrid capture probes for each polymorphic loci, and where each hybrid capture probe is designed to be complementary to a different allele at that polymorphic locus.

In some embodiments of any of the aspects of the invention, preferentially enriching the DNA a plurality of polymorphic loci includes obtaining a plurality of inner forward primers where each primer targets one of the polymorphic loci, and where the 3' end of the inner forward primers are designed to hybridize to a region of DNA upstream from the polymorphic site, and separated from the polymorphic site by a small number of bases, where the small number is selected from the group consisting of 1, 2, 3, 4, 5, 6 to 10, 11 to 15, 16 to 20, 21 to 25, 26 to 30, or 31 to 60 base pairs, optionally obtaining a plurality of inner reverse

primers where each primer targets one of the polymorphic loci, and where the 3' end of the inner reverse primers are designed to hybridize to a region of DNA upstream from the polymorphic site, and separated from the polymorphic site by a small number of bases, where the small number is selected from the group consisting of 1, 2, 3, 4, 5, 6 to 10, 11 to 15, 16 to 20, 21 to 25, 26 to 30, or 31 to 60 base pairs, hybridizing the inner primers to the DNA, and amplifying the DNA using the polymerase chain reaction to form amplicons.

In some embodiments of any of the aspects of the invention, the method also includes obtaining a plurality of outer forward primers where each primer targets one of the target (*e.g.*, polymorphic loci), and where the outer forward primers are designed to hybridize to the region of DNA upstream from the inner forward primer, optionally obtaining a plurality of outer reverse primers where each primer targets one of the target loci (*e.g.*, polymorphic loci), and where the outer reverse primers are designed to hybridize to the region of DNA immediately downstream from the inner reverse primer, hybridizing the first primers to the DNA, and amplifying the DNA using the polymerase chain reaction.

In some embodiments of any of the aspects of the invention, the method also includes obtaining a plurality of outer reverse primers where each primer targets one of the polymorphic loci, and where the outer reverse primers are designed to hybridize to the region of DNA immediately downstream from the inner reverse primer, optionally obtaining a plurality of outer forward primers where each primer targets one of the target loci (*e.g.*, the polymorphic loci), and where the outer forward primers are designed to hybridize to the region of DNA upstream from the inner forward primer, hybridizing the first primers to the DNA, and amplifying the DNA using the polymerase chain reaction.

In some embodiments of any of the aspects of the invention, preparing the sample (*e.g.*, the first sample) further includes appending universal adapters to the DNA in the sample (*e.g.*, the first sample) and amplifying the DNA in the sample (*e.g.*, the first sample) using the polymerase chain reaction. In some embodiments, at least a fraction of the amplicons that are amplified are less than 100 bp, less than 90 bp, less than 80 bp, less than 70 bp, less than 65 bp, less than 60 bp, less than 55 bp, less than 50 bp, or less than 45 bp, and where the fraction is 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 99%.

In some embodiments of any of the aspects of the invention, amplifying the DNA is done in one or a plurality of individual reaction volumes, and where each individual reaction volume contains more than 100 different forward and reverse primer pairs, more than 200

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 25

different forward and reverse primer pairs, more than 500 different forward and reverse primer pairs, more than 1,000 different forward and reverse primer pairs, more than 2,000 different forward and reverse primer pairs, more than 5,000 different forward and reverse primer pairs, more than 10,000 different forward and reverse primer pairs, more than 20,000 different forward and reverse primer pairs, more than 50,000 different forward and reverse primer pairs, or more than 100,000 different forward and reverse primer pairs.

In some embodiments of any of the aspects of the invention, preparing the sample (*e.g.*, the first sample) further comprises dividing the sample (*e.g.*, the first sample) into a plurality of portions, and where the DNA in each portion is preferentially enriched at a subset of the target loci (*e.g.*, plurality of polymorphic loci). In some embodiments, the inner primers are selected by identifying primer pairs likely to form undesired primer duplexes and removing from the plurality of primers at least one of the pair of primers identified as being likely to form undesired primer duplexes. In some embodiments, the inner primers contain a region that is designed to hybridize either upstream or downstream of the targeted locus (*e.g.*, the polymorphic locus), and optionally contain a universal priming sequence designed to allow PCR amplification. In some embodiments, at least some of the primers additionally contain a random region that differs for each individual primer molecule. In some embodiments, at least some of the primers additionally contain a molecular barcode.

In some embodiments of any of the aspects of the invention, preferential enrichment results in average degree of allelic bias between the prepared sample and the sample (*e.g.*, the first sample) of a factor selected from the group consisting of no more than a factor of 2, no more than a factor of 1.5, no more than a factor of 1.2, no more than a factor of 1.1, no more than a factor of 1.05, no more than a factor of 1.02, no more than a factor of 1.01, no more than a factor of 1.005, no more than a factor of 1.002, no more than a factor of 1.001 and no more than a factor of 1.0001. In some embodiments, the plurality of polymorphic loci are SNPs. In some embodiments, measuring the DNA in the prepared sample is done by sequencing.

In some embodiments of any of the aspects of the invention, the target loci are present on the same nucleic acid of interest (*e.g*, the same chromosome or the same region of a chromosome). In some embodiments, at least some of the target loci are present on different nucleic acids of interest (*e.g*, different chromosomes). In some embodiments, the nucleic acid sample includes fragmented or digested nucleic acids. In some embodiments, the nucleic acid sample includes genomic DNA, cDNA, or mRNA. In some embodiments, the nucleic acid

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 26

sample includes DNA from a single cell. In some embodiments, the nucleic acid sample is a blood or plasma sample that is substantially free of cells. In some embodiments, the nucleic acid sample includes or is derived from blood, plasma, saliva, semen, sperm, cell culture supernatant, mucus secretion, dental plaque, gastrointestinal tract tissue, stool, urine, hair, bone, body fluids, , tears, tissue, skin, fingernails, blastomeres, embryos, amniotic fluid, chorionic villus samples, bile, lymph, cervical mucus, or a forensic sample. In some embodiments, the target loci are segments of human nucleic acids. In some embodiments, the target loci comprise or consist of single nucleotide polymorphisms (SNPs). In some embodiments, the primers are DNA molecules.

In some embodiments of any of the aspects of the invention, the DNA in the sample (*e.g.*, the first sample) originates from maternal plasma. In some embodiments, preparing the sample (*e.g.*, the first sample) further comprises amplifying the DNA. In some embodiments, preparing the sample (*e.g.*, the first sample) further comprises preferentially enriching the DNA in the sample (*e.g.*, the first sample) at the target loci (*e.g.*, a plurality of polymorphic loci).

In various embodiments, the primer extension reaction or the polymerase chain reaction includes the addition of one or more nucleotides by a polymerase. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include ligation-mediated PCR. In various embodiments, the primer extension reaction or the polymerase chain reaction does not include the joining of two primers by a ligase. In various embodiments, the primers do not include Linked Inverted Probes (LIPs), which can also be called pre-circularized probes, pre-circularizing probes, circularizing probes, Padlock Probes, or Molecular Inversion Probes (MIPs).

It is understood that aspects and embodiments of the invention described herein include "comprising," "consisting," and "consisting essentially of" aspects and embodiments.

Definitions

Single Nucleotide Polymorphism (SNP) refers to a single nucleotide that may differ between the genomes of two members of the same species. The usage of the term should not imply any limit on the frequency with which each variant occurs.

Sequence refers to a DNA sequence or a genetic sequence. It may refer to the primary, physical structure of the DNA molecule or strand in an individual. It may refer to the

sequence of nucleotides found in that DNA molecule, or the complementary strand to the DNA molecule. It may refer to the information contained in the DNA molecule as its representation *in silico*.

Locus refers to a particular region of interest on the DNA of an individual, which may refer to a SNP, the site of a possible insertion or deletion, or the site of some other relevant genetic variation. Disease-linked SNPs may also refer to disease-linked loci.

Polymorphic Allele, also "Polymorphic Locus," refers to an allele or locus where the genotype varies between individuals within a given species. Some examples of polymorphic alleles include single nucleotide polymorphisms, short tandem repeats, deletions, duplications, and inversions.

Polymorphic Site refers to the specific nucleotides found in a polymorphic region that vary between individuals.

Allele refers to the genes that occupy a particular locus.

Genetic Data also "*Genotypic Data"* refers to the data describing aspects of the genome of one or more individuals. It may refer to one or a set of loci, partial or entire sequences, partial or entire chromosomes, or the entire genome. It may refer to the identity of one or a plurality of nucleotides; it may refer to a set of sequential nucleotides, or nucleotides from different locations in the genome, or a combination thereof. Genotypic data is typically *in silico*, however, it is also possible to consider physical nucleotides in a sequence as chemically encoded genetic data. Genotypic Data may be said to be "on," "of," "at," "from" or "on" the individual(s). Genotypic Data may refer to output measurements from a genotyping platform where those measurements are made on genetic material.

Genetic Material also "*Genetic Sample"* refers to physical matter, such as tissue or blood, from one or more individuals comprising DNA or RNA

Noisy Genetic Data refers to genetic data with any of the following: allele dropouts, uncertain base pair measurements, incorrect base pair measurements, missing base pair measurements, uncertain measurements of insertions or deletions, uncertain measurements of chromosome segment copy numbers, spurious signals, missing measurements, other errors, or combinations thereof.

Confidence refers to the statistical likelihood that the called SNP, allele, set of alleles, ploidy call, or determined number of chromosome segment copies correctly represents the real genetic state of the individual.

Ploidy Calling, also "Chromosome Copy Number Calling," or "Copy Number Calling" (CNC), may refer to the act of determining the quantity and/or chromosomal identity of one or more chromosomes present in a cell.

Aneuploidy refers to the state where the wrong number of chromosomes (*e.g.,* the wrong number of full chromosomes or the wrong number of chromosome segments, such as the presence of deletions or duplications of a chromosome segment) is present in a cell. In the case of a somatic human cell it may refer to the case where a cell does not contain 22 pairs of autosomal chromosomes and one pair of sex chromosomes. In the case of a human gamete, it may refer to the case where a cell does not contain one of each of the 23 chromosomes. In the case of a single chromosome type, it may refer to the case where more or less than two homologous but non-identical chromosome copies are present, or where there are two chromosome copies present that originate from the same parent. In some embodiments, the deletion of a chromosome segment is a microdeletion.

Ploidy State refers to the quantity and/or chromosomal identity of one or more chromosomes types in a cell.

Chromosome may refer to a single chromosome copy, meaning a single molecule of DNA of which there are 46 in a normal somatic cell; an example is 'the maternally derived chromosome 18'. Chromosome may also refer to a chromosome type, of which there are 23 in a normal human somatic cell; an example is 'chromosome 18'.

Chromosomal Identity may refer to the referent chromosome number, *i.e.* the chromosome type. Normal humans have 22 types of numbered autosomal chromosome types, and two types of sex chromosomes. It may also refer to the parental origin of the chromosome. It may also refer to a specific chromosome inherited from the parent. It may also refer to other identifying features of a chromosome.

The State of the Genetic Material or simply "Genetic State" may refer to the identity of a set of SNPs on the DNA, to the phased haplotypes of the genetic material, and to the sequence of the DNA, including insertions, deletions, repeats and mutations. It may

also refer to the ploidy state of one or more chromosomes, chromosomal segments, or set of chromosomal segments.

Allelic Data refers to a set of genotypic data concerning a set of one or more alleles. It may refer to the phased, haplotypic data. It may refer to SNP identities, and it may refer to the sequence data of the DNA, including insertions, deletions, repeats and mutations. It may include the parental origin of each allele.

Allelic State refers to the actual state of the genes in a set of one or more alleles. It may refer to the actual state of the genes described by the allelic data.

Allelic Ratio or allele ratio, refers to the ratio between the amount of each allele at a locus that is present in a sample or in an individual. When the sample was measured by sequencing, the allelic ratio may refer to the ratio of sequence reads that map to each allele at the locus. When the sample was measured by an intensity based measurement method, the allele ratio may refer to the ratio of the amounts of each allele present at that locus as estimated by the measurement method.

Allele Count refers to the number of sequences that map to a particular locus, and if that locus is polymorphic, it refers to the number of sequences that map to each of the alleles. If each allele is counted in a binary fashion, then the allele count will be whole number. If the alleles are counted probabilistically, then the allele count can be a fractional number.

Allele Count Probability refers to the number of sequences that are likely to map to a particular locus or a set of alleles at a polymorphic locus, combined with the probability of the mapping. Note that allele counts are equivalent to allele count probabilities where the probability of the mapping for each counted sequence is binary (zero or one). In some embodiments, the allele count probabilities may be binary. In some embodiments, the allele count probabilities may be set to be equal to the DNA measurements.

Allelic Distribution, or 'allele count distribution' refers to the relative amount of each allele that is present for each locus in a set of loci. An allelic distribution can refer to an individual, to a sample, or to a set of measurements made on a sample. In the context of sequencing, the allelic distribution refers to the number or probable number of reads that map to a particular allele for each allele in a set of polymorphic loci. The allele measurements may be treated probabilistically, that is, the likelihood that a given allele is present for a give sequence read is a fraction between 0 and 1, or they may be treated

in a binary fashion, that is, any given read is considered to be exactly zero or one copies of a particular allele.

Allelic Distribution Pattern refers to a set of different allele distributions for different parental contexts. Certain allelic distribution patterns may be indicative of certain ploidy states.

Allelic Bias refers to the degree to which the measured ratio of alleles at a heterozygous locus is different to the ratio that was present in the original sample of DNA. The degree of allelic bias at a particular locus is equal to the observed allelelic ratio at that locus, as measured, divided by the ratio of alleles in the original DNA sample at that locus. Allelic bias may be defined to be greater than one, such that if the calculation of the degree of allelic bias returns a value, x, that is less than 1, then the degree of allelic bias may be restated as 1/x. Allelic bias maybe due to amplification bias, purification bias, or some other phenomenon that affects different alleles differently.

Primer, also "PCR probe" refers to a single DNA molecule (a DNA oligomer) or a collection of DNA molecules (DNA oligomers) where the DNA molecules are identical, or nearly so, and where the primer contains a region that is designed to hybridize to a targeted locus (*e.g,* a targeted polymorphic locus or a nonpolymorphic locus), and may contain a priming sequence designed to allow PCR amplification. A primer may also contain a molecular barcode. A primer may contain a random region that differs for each individual molecule. The terms "test primer" and "candidate primer" are not meant to be limiting and may refer to any of the primers disclosed herein.

Library of primers refers to a population of two or more primers. In various embodiments, the library includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primers. In various embodiments, the library includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer where each pair of test primers hybridize to a target locus. In some embodiments, the library of primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs. In some embodiments, the library has both (i) primer pairs and (ii) individual primers (such as universal primers) that are not part of primer pairs.

Hybrid Capture Probe refers to any nucleic acid sequence, possibly modified, that is generated by various methods such as PCR or direct synthesis and intended to be complementary to one strand of a specific target DNA sequence in a sample. The exogenous hybrid capture probes may be added to a prepared sample and hybridized through a deanture-reannealing process to form duplexes of exogenous-endogenous fragments. These duplexes may then be physically separated from the sample by various means.

Sequence Read refers to data representing a sequence of nucleotide bases that were measured using a clonal sequencing method. Clonal sequencing may produce sequence data representing single, or clones, or clusters of one original DNA molecule. A sequence read may also have associated quality score at each base position of the sequence indicating the probability that nucleotide has been called correctly.

Mapping a sequence read is the process of determining a sequence read's location of origin in the genome sequence of a particular organism. The location of origin of sequence reads is based on similarity of nucleotide sequence of the read and the genome sequence.

Matched Copy Error, also "Matching Chromosome Aneuploidy" (MCA), refers to a state of aneuploidy where one cell contains two identical or nearly identical chromosomes. This type of aneuploidy may arise during the formation of the gametes in meiosis, and may be referred to as a meiotic non-disjunction error. This type of error may arise in mitosis. Matching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are identical.

Unmatched Copy Error, also "Unique Chromosome Aneuploidy" (UCA), refers to a state of aneuploidy where one cell contains two chromosomes that are from the same parent, and that may be homologous but not identical. This type of aneuploidy may arise during meiosis, and may be referred to as a meiotic error. Unmatching trisomy may refer to the case where three copies of a given chromosome are present in an individual and two of the copies are from the same parent, and are homologous, but are not identical. Note that unmatching trisomy may refer to the case where two homologous chromosomes from one parent are present, and where some segments of the chromosomes are identical while other segments are merely homologous.

Homologous Chromosomes refers to chromosome copies that contain the same set of genes that normally pair up during meiosis.

Identical Chromosomes refers to chromosome copies that contain the same set of genes, and for each gene they have the same set of alleles that are identical, or nearly identical.

Allele Drop Out (ADO) refers to the situation where at least one of the base pairs in a set of base pairs from homologous chromosomes at a given allele is not detected.

Locus Drop Out (LDO) refers to the situation where both base pairs in a set of base pairs from homologous chromosomes at a given allele are not detected.

Homozygous refers to having similar alleles as corresponding chromosomal loci.

Heterozygous refers to having dissimilar alleles as corresponding chromosomal loci.

Heterozygosity Rate refers to the rate of individuals in the population having heterozygous alleles at a given locus. The heterozygosity rate may also refer to the expected or measured ratio of alleles, at a given locus in an individual, or a sample of DNA.

Highly Informative Single Nucleotide Polymorphism (HISNP) refers to a SNP where the fetus has an allele that is not present in the mother's genotype.

Chromosomal Region refers to a segment of a chromosome, or a full chromosome.

Segment of a Chromosome refers to a section of a chromosome that can range in size from one base pair to the entire chromosome.

Chromosome refers to either a full chromosome, or a segment or section of a chromosome.

Copies refers to the number of copies of a chromosome segment. It may refer to identical copies, or to non-identical, homologous copies of a chromosome segment wherein the different copies of the chromosome segment contain a substantially similar set of loci, and where one or more of the alleles are different. Note that in some cases of aneuploidy, such as the M2 copy error, it is possible to have some copies of the given chromosome segment that are identical as well as some copies of the same chromosome segment that are not identical.

Haplotype refers to a combination of alleles at multiple loci that are typically inherited together on the same chromosome. Haplotype may refer to as few as two loci or to an entire chromosome depending on the number of recombination events that have occurred between a given set of loci. Haplotype can also refer to a set of single nucleotide polymorphisms (SNPs) on a single chromatid that are statistically associated.

Haplotypic Data, also "Phased Data" or "Ordered Genetic Data," refers to data from a single chromosome in a diploid or polyploid genome, *i.e.*, either the segregated maternal or paternal copy of a chromosome in a diploid genome.

Phasing refers to the act of determining the haplotypic genetic data of an individual given unordered, diploid (or polyploidy) genetic data. It may refer to the act of determining which of two genes at an allele, for a set of alleles found on one chromosome, are associated with each of the two homologous chromosomes in an individual.

Phased Data refers to genetic data where one or more haplotypes have been determined.

Hypothesis refers to a possible ploidy state at a given set of chromosomes, or a set of possible allelic states at a given set of loci. The set of possibilities may comprise one or more elements.

Copy Number Hypothesis, also "Ploidy State Hypothesis," refers to a hypothesis concerning the number of copies of a chromosome in an individual. It may also refer to a hypothesis concerning the identity of each of the chromosomes, including the parent of origin of each chromosome, and which of the parent's two chromosomes are present in the individual. It may also refer to a hypothesis concerning which chromosomes, or chromosome segments, if any, from a related individual correspond genetically to a given chromosome from an individual.

Target Individual refers to the individual whose genetic state is being determined. In some embodiments, only a limited amount of DNA is available from the target individual. In some embodiments, the target individual is a fetus. In some embodiments, there may be more than one target individual. In some embodiments, each fetus that originated from a pair of parents may be considered to be target individuals. In some embodiments, the genetic data that is being determined is one or a set of allele calls. In some embodiments, the genetic data that is being determined is a ploidy call.

Related Individual refers to any individual who is genetically related to, and thus shares haplotype blocks with, the target individual. In one context, the related individual may be a genetic parent of the target individual, or any genetic material derived from a parent, such as a sperm, a polar body, an embryo, a fetus, or a child. It may also refer to a sibling, parent or a grandparent.

Sibling refers to any individual whose genetic parents are the same as the individual in question. In some embodiments, it may refer to a born child, an embryo, or a fetus, or one or more cells originating from a born child, an embryo, or a fetus. A sibling may also refer to a haploid individual that originates from one of the parents, such as a

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 34

sperm, a polar body, or any other set of haplotypic genetic matter. An individual may be considered to be a sibling of itself.

Fetal refers to "of the fetus," or "of the region of the placenta that is genetically similar to the fetus". In a pregnant woman, some portion of the placenta is genetically similar to the fetus, and the free floating fetal DNA found in maternal blood may have originated from the portion of the placenta with a genotype that matches the fetus. Note that the genetic information in half of the chromosomes in a fetus is inherited from the mother of the fetus. In some embodiments, the DNA from these maternally inherited chromosomes that came from a fetal cell is considered to be "of fetal origin," not "of maternal origin."

DNA of Fetal Origin refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the fetus.

DNA of Maternal Origin refers to DNA that was originally part of a cell whose genotype was essentially equivalent to that of the mother.

Child may refer to an embryo, a blastomere, or a fetus. Note that in the presently disclosed embodiments, the concepts described apply equally well to individuals who are a born child, a fetus, an embryo or a set of cells therefrom. The use of the term child may simply be meant to connote that the individual referred to as the child is the genetic offspring of the parents.

Parent refers to the genetic mother or father of an individual. An individual typically has two parents, a mother and a father, though this may not necessarily be the case such as in genetic or chromosomal chimerism. A parent may be considered to be an individual.

Parental Context refers to the genetic state of a given SNP, on each of the two relevant chromosomes for one or both of the two parents of the target.

Develop As Desired, also "Develop Normally," refers to a viable embryo implanting in a uterus and resulting in a pregnancy, and/or to a pregnancy continuing and resulting in a live birth, and/or to a born child being free of chromosomal abnormalities, and/or to a born child being free of other undesired genetic conditions such as disease-linked genes. The term "develop as desired" is meant to encompass anything that may be desired by parents or healthcare facilitators. In some cases, "develop as desired" may refer to an unviable or viable embryo that is useful for medical research or other purposes.

Insertion into a Uterus refers to the process of transferring an embryo into the uterine cavity in the context of *in vitro* fertilization.

Maternal Plasma refers to the plasma portion of the blood from a female who is pregnant.

Clinical Decision refers to any decision to take or not take an action that has an outcome that affects the health or survival of an individual. In the context of prenatal diagnosis, a clinical decision may refer to a decision to abort or not abort a fetus. A clinical decision may also refer to a decision to conduct further testing, to take actions to mitigate an undesirable phenotype, or to take actions to prepare for the birth of a child with abnormalities.

Diagnostic Box refers to one or a combination of machines designed to perform one or a plurality of aspects of the methods disclosed herein. In an embodiment, the diagnostic box may be placed at a point of patient care. In an embodiment, the diagnostic box may perform targeted amplification followed by sequencing. In an embodiment the diagnostic box may function alone or with the help of a technician.

Informatics Based Method refers to a method that relies heavily on statistics to make sense of a large amount of data. In the context of prenatal diagnosis, it refers to a method designed to determine the ploidy state at one or more chromosomes or the allelic state at one or more alleles by statistically inferring the most likely state, rather than by directly physically measuring the state, given a large amount of genetic data, for example from a molecular array or sequencing. In an embodiment of the present disclosure, the informatics based technique may be one disclosed in this patent. In an embodiment of the present disclosure it may be PARENTAL SUPPORT™.

Primary Genetic Data refers to the analog intensity signals that are output by a genotyping platform. In the context of SNP arrays, primary genetic data refers to the intensity signals before any genotype calling has been done. In the context of sequencing, primary genetic data refers to the analog measurements, analogous to the chromatogram, that comes off the sequencer before the identity of any base pairs have been determined, and before the sequence has been mapped to the genome.

Secondary Genetic Data refers to processed genetic data that are output by a genotyping platform. In the context of a SNP array, the secondary genetic data refers to the allele calls made by software associated with the SNP array reader, wherein the software has made a call whether a given allele is present or not present in the sample. In the context

of sequencing, the secondary genetic data refers to the base pair identities of the sequences have been determined, and possibly also where the sequences have been mapped to the genome.

Non-Invasive Prenatal Diagnosis (NPD), or also "Non-Invasive Prenatal Screening" (NPS), refers to a method of determining the genetic state of a fetus that is gestating in a mother using genetic material found in the mother's blood, where the genetic material is obtained by drawing the mother's intravenous blood.

Preferential Enrichment of DNA that corresponds to a locus, or preferential enrichment of DNA at a locus, refers to any method that results in the percentage of molecules of DNA in a post-enrichment DNA mixture that correspond to the locus being higher than the percentage of molecules of DNA in the pre-enrichment DNA mixture that correspond to the locus. The method may involve selective amplification of DNA molecules that correspond to a locus. The method may involve removing DNA molecules that do not correspond to the locus. The method may involve a combination of methods. The degree of enrichment is defined as the percentage of molecules of DNA in the post-enrichment mixture that correspond to the locus divided by the percentage of molecules of DNA in the pre-enrichment mixture that correspond to the locus. Preferential enrichment may be carried out at a plurality of loci. In some embodiments of the present disclosure, the degree of enrichment is greater than 20. In some embodiments of the present disclosure, the degree of enrichment is greater than 200. In some embodiments of the present disclosure, the degree of enrichment is greater than 2,000. When preferential enrichment is carried out at a plurality of loci, the degree of enrichment may refer to the average degree of enrichment of all of the loci in the set of loci.

Amplification refers to a method that increases the number of copies of a molecule of DNA.

Selective Amplification may refer to a method that increases the number of copies of a particular molecule of DNA, or molecules of DNA that correspond to a particular region of DNA. It may also refer to a method that increases the number of copies of a particular targeted molecule of DNA, or targeted region of DNA more than it increases non-targeted molecules or regions of DNA. Selective amplification may be a method of preferential enrichment.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 37

Universal Priming Sequence refers to a DNA sequence that may be appended to a population of target DNA molecules, for example by ligation, PCR, or ligation mediated PCR. Once added to the population of target molecules, primers specific to the universal priming sequences can be used to amplify the target population using a single pair of amplification primers. Universal priming sequences are typically not related to the target sequences.

Universal Adapters, or 'ligation adaptors' or 'library tags' are DNA molecules containing a universal priming sequence that can be covalently linked to the 5-prime and 3-prime end of a population of target double stranded DNA molecules. The addition of the adapters provides universal priming sequences to the 5-prime and 3-prime end of the target population from which PCR amplification can take place, amplifying all molecules from the target population, using a single pair of amplification primers.

Targeting refers to a method used to selectively amplify or otherwise preferentially enrich those molecules of DNA that correspond to a set of loci, in a mixture of DNA.

Joint Distribution Model refers to a model that defines the probability of events defined in terms of multiple random variables, given a plurality of random variables defined on the same probability space, where the probabilities of the variable are linked. In some embodiments, the degenerate case where the probabilities of the variables are not linked may be used.

BRIEF DESCRIPTION OF THE DRAWINGS

The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

The presently disclosed embodiments will be further explained with reference to the attached drawings, wherein like structures are referred to by like numerals throughout the several views. The drawings shown are not necessarily to scale, with emphasis instead generally being placed upon illustrating the principles of the presently disclosed embodiments.

Figure 1: Graphical representation of direct multiplexed mini-PCR method.

Figure 2: Graphical representation of semi-nested mini-PCR method.

Figure 3: Graphical representation of fully nested mini-PCR method.

Figure 4: Graphical representation of hemi-nested mini-PCR method.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 38

Figure 5: Graphical representation of triply hemi-nested mini-PCR method.

Figure 6: Graphical representation of one-sided nested mini-PCR method.

Figure 7: Graphical representation of one-sided mini-PCR method.

Figure 8: Graphical representation of reverse semi-nested mini-PCR method.

Figure 9: Some possible workflows for semi-nested methods.

Figure 10: Graphical representation of looped ligation adaptors.

Figure 11: Graphical representation of internally tagged primers.

Figure 12: An example of some primers with internal tags.

Figure 13: Graphical representation of a method using primers with a ligation adaptor binding region.

Figure 14: Simulated ploidy call accuracies for counting method with two different analysis techniques.

Figure 15: Ratio of two alleles for a plurality of SNPs in a cell line in Experiment 4.

Figure 16: Ratio of two alleles for a plurality of SNPs in a cell line in Experiment 4 sorted by chromosome.

Figures 17A-D: Ratio of two alleles for a plurality of SNPs in four pregnant women plasma samples, sorted by chromosome.

Figure 18: Fraction of data that can be explained by binomial variance before and after data correction.

Figure 19: Graph showing relative enrichment of fetal DNA in samples following a short library preparation protocol.

Figure 20: Depth of read graph comparing direct PCR and semi-nested methods.

Figure 21: Comparison of depth of read for direct PCR of three genomic samples.

Figure 22: Comparison of depth of read for semi-nested mini-PCR of three samples.

Figure 23: Comparison of depth of read for 1,200-plex and 9,600-plex reactions.

Figure 24: Read count ratios for six cells at three chromosomes.

Figure 25: Allele ratios for two three-cell reactions and a third reaction run on 1 ng of genomic DNA at three chromosomes.

Figure 26: Allele ratios for two single-cell reactions at three chromosomes

Figure 27: Comparison of two primer libraries showing the number of loci with a particular minor allele frequency that are targeted by each primer library.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 39

Figure 28A: Graph of the electrophoresis of PCR products. Figures 28B-28M are electropherograms of lanes 1-12, respectively, in Figure 28A.

Figures 29A-29E: Cartoon depiction of a method of the invention for the determination of a fetal aneuploidy (Figure 29A). Maternal and paternal genotype data (from blood or buccal swabs) and crossover frequency data from the HapMap database are utilized to generate (Figure 29B) multiple independent hypotheses for each potential fetal ploidy state *in silico* (Figure 29C). Each of these hypotheses is expanded to include sub-hypotheses with take into consideration the different possible crossover points. The data model predicts what the sequencing data would look like (the expected allele distributions) given each hypothetical fetal genotype and at different fetal cfDNA fractions, and is compared to the actual sequencing data; the likelihood for each hypothesis is determined using Bayesian statistics. In this hypothetical example, the hypotheses with the highest likelihoods (euploidy) are indicated in red (Figure 29D). The individual likelihoods from Figure 29C are summed for each copy number hypothesis family (monosomy, disomy, or triploidy). The hypothesis with the maximum likelihood is called as the ploidy state, reveals the fetal fraction, and represents the sample-specific calculated accuracy (Figure 29E).

Figures 30A-30H: Typical graphical representations of euploidy (Figures 30A-30C), monosomy (Figure 30D), and trisomy (Figures 30E-30H). For all plots, the x-axis represents the linear position of the individual polymorphic loci along each chromosome (as indicated below the plots), and the y-axis represents the number of A allele reads as a fraction of the total (A+B) allele reads. Maternal and fetal genotypes, as well as the position on the y-axis around which the bands are centered, are indicated to the right of the plots. The plots are color-coded according to maternal genotype, such that red indicates a maternal genotype of AA, blue indicates a maternal genotype of BB, and green indicates a maternal genotype of AB. Maternal allele contributions are indicated in color in the "Fetal Genotype" column. Allele contributions are indicated as maternal|fetal, such that alleles for which the mother is AA and the fetus is AB are indicated as AA|AB. Figure 30A: Generated plots when two chromosomes are present and the fetal cfDNA fraction is 0%. This plot is from a non-pregnant woman, and thus represents the pattern when the genotype is entirely maternal. Allele clusters are thus centered around 1 (AA alleles), 0.5 (AB alleles), and 0 (BB alleles). Figure 30B**:** Generated plot when two chromosomes are present and the fetal fraction is 12%. The contribution of fetal alleles to the fraction of A allele reads shifts the position of some allele spots up or down along

the y-axis, such that the bands are centered around 1 (AA|AA alleles), 0.94 (AA|AB alleles), 0.56 (AB|AA alleles), 0.50 (AB|AB alleles), 0.44 (AB|BB alleles), 0.06 (BB|AB alleles), and 0 (BB|BB alleles). Figure 30C. Generated plot when two chromosomes are present and the fetal fraction is 26%. The pattern, including two red and two blue peripheral bands and a trio of central green bands, is readily apparent. Bands are centered around 1 (AA|AA alleles), 0.87 (AA|AB alleles), 0.63 (AB|AA alleles), 0.50 (AB|AB alleles), 0.37 (AB|BB alleles), 0.13 (BB|AB alleles), and 0 (BB|BB alleles). Figure 30D: Generated plot when one chromosome is present and the fetal fraction is 26%. The hallmark pattern of one external red and one external blue peripheral band as well as two central green bands indicated maternally-inherited monosomy. Because the fetus only contributes a single allele (A or B) to the allele reads, the internal peripheral red and blue bands are not present, and the center trio of bands condenses into two bands. Bands that are centered around 1 (AA|A alleles), 0.57 (AB|A alleles), 0.43 (AB|B alleles), and 0 (BB|B alleles). Figure 30E: Generated plot when three chromosomes are present and the fetal fraction is 27%. This pattern of two red and two blue peripheral bands as well as two central green bands indicates maternally-inherited meiotic trisomy. Bands are centered around 1 (AA|AAA alleles), 0.88 (AA|AAB alleles), 0.56 (AB|AAB alleles), 0.44 (AB|ABB alleles), 0.12 (BB|ABB alleles), and 0 (BB|BBB alleles). Figure 30F: Generated plot when three chromosomes are present and the fetal fraction is 14%. This pattern of three red and three blue peripheral bands, as well as two central green bands, indicates paternally-inherited meiotic trisomy. Bands are centered around 1 (AA|AAA alleles), 0.93 (AA|AAB alleles), 0.87 (AA|ABB alleles), 0.60 (AB|AAA alleles), 0.53 (AB|AAB alleles), 0.47 (AB|ABB alleles), 0.40 (AB|BBB alleles), 0.13 (BB|AAB alleles), 0.07 (BB|ABB alleles), and 0 (BB|BBB alleles). Figure 30G: Generated plot when three chromosomes are present and the fetal fraction is 35%. This pattern of two red and two blue peripheral bands and four central green bands indicates maternally-inherited mitotic trisomy. Bands are centered around 1 (AA|AAA alleles), 0.85 (AA|AAB alleles), 0.72 (AB|AAA alleles), 0.57 (AB|AAB alleles), 0.43 (AB|ABB alleles), 0.28 (AB|BBB alleles), 0.15 (BB|ABB alleles), and 0 (BB|BBB alleles). Figure 30H: Generated plot when three chromosomes are present and the fetal fraction is 25%. This pattern of two red and two blue peripheral bands as well as four central green bands indicates paternally-inherited mitotic trisomy. This pattern can be distinguished from that of maternally-inherited mitotic trisomy (as in Figure 30G) by the position of the internal peripheral bands. Specifically, bands are centered around 1 (AA|AAA alleles), 0.78 (AA|ABB

alleles), 0.67 (AB|AAA alleles), 0.56 (AB|AAB alleles), 0.44 (AB|ABB alleles), 0.33 (AB|BBB alleles), 0.22 (BB|AAB alleles), and 0 (BB|BBB alleles).

Figure 31: Graphical representations of (Figure 31A) euploid, (Figure 31B) T13, (Figure 31C) T18, (Figure 31D) T21, (Figure 31E) 45,X, and (Figure 31F) 47,XXY test samples as indicated. Each chromosome is indicated at the top of the plot, fetal and maternal genotypes are indicated to the right of the plots, the x-axis represents the linear position of the SNPs along each chromosome, and the y-axis indicates the number of A allele reads as a fraction of the total reads. Note the altered cluster positioning based on fetal fraction, as described herein. Each spot represents a single SNP locus. Fetal and maternal genotypes are indicated to the right of the plot, and chromosome identities are indicated at the top of the plots.

Figure 32: The combined at-birth prevalence of sex chromosome aneuploidies is greater than that of autosomal aneuploidies.

While the above-identified drawings set forth presently disclosed embodiments, other embodiments are also contemplated, as noted in the discussion. This disclosure presents illustrative embodiments by way of representation and not limitation. Numerous other modifications and embodiments can be devised by those skilled in the art which fall within the scope and spirit of the principles of the presently disclosed embodiments.

DETAILED DESCRIPTION

The present invention is based in part on the surprising discovery that often only a relatively small number of primers in a library of primers are responsible for a substantial amount of the amplified primer dimers that form during multiplex PCR reactions. Methods have been developed to select the most undesirable primers for removal from a library of candidate primers. By reducing the amount of primer dimers to a negligible amount (~0.1% of the PCR products), these methods allow the resulting primer libraries to simultaneously amplify a large number of target loci in a single multiplex PCR reaction. Because the primers hybridize to the target loci and amplify them rather than hybridizing to other primers and forming amplified primer dimers, the number of different target loci that can be amplified is increased. It was also discovered that using lower primer concentrations and much longer

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 42

annealing times than normal increases the likelihood that the primers hybridize to the target loci instead of hybridizing to each other and forming primer dimers.

During the PCR amplification and sequencing of 19,488 target loci in a genomic sample, 99.4-99.7% of the sequencing reads mapped to the genome, of those, 99.99% of the mapped to targeted loci. For plasma samples with 10 million sequencing reads, typically at least 19,350 of the 19,488 targeted loci (99.3 %) were amplified and sequenced. Being able to simultaneously amplify such a large number of target loci at once greatly decreases the amount of time and the amount of DNA required to analyze thousands of target loci. For example, DNA from a single cell is sufficient to simultaneously analyze thousands of target loci, which is important for applications in which the amount of DNA is low, such as genetic testing of a single cell from an embryo prior to *in vitro* fertilization or genetic testing of a forensic sample with little DNA. In addition, being able to analyze the target loci in one reaction volume (such as in one chamber or well) rather than splitting the sample into multiple different reactions reduces variability that can occur between reactions. In addition, methods have been developed to use reference standards to correct for amplification bias that may occur between different target loci. For example, differences in amplification efficiency between target loci due to factors such as GC content may cause differing amounts of PCR products to be produced for target loci that are actually present in the same amount. The use of reference standards similar to the target loci allows the detection of such amplification bias so that it can be corrected for during the quantitation of the target loci.

During sequencing of PCR products, artifacts such as primer dimers are detected and thus inhibit the detection of target amplicons. Because of this limitation, microarrays with hybridization probes are often used for detection since microarrays are less sensitive to interference from primer dimers. The high level of multiplexing with minimal non-target amplicons that has now been achieved allows PCR followed by sequencing to be used as an alternative to microarrays.

The multiplex-PCR methods of the invention can be in a variety of applications, such as genotyping, detection of chromosomal abnormalities (such as a fetal chromosome aneuploidy), gene mutation and polymorphism (such as single nucleotide polymorphisms, SNPs) analysis, gene deletion analysis, determination of paternity, analysis of genetic differences among populations, forensic analysis, measuring predisposition to disease, quantitative analysis of mRNA, and detection and identification of infectious agents (such as

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 43

bacteria, parasite, and viruses). The multiplex PCR methods can also be used for non-invasive prenatal testing, such as paternity testing or the detection of fetal chromosome abnormalities.

Exemplary Primer Design Methods

Highly multiplexed PCR can often result in the production of a very high proportion of product DNA that results from unproductive side reactions such as primer dimer formation. In an embodiment, the particular primers that are most likely to cause unproductive side reactions may be removed from the primer library to give a primer library that will result in a greater proportion of amplified DNA that maps to the genome. The step of removing problematic primers, that is, those primers that are particularly likely to firm dimers has unexpectedly enabled extremely high PCR multiplexing levels for subsequent analysis by sequencing. In systems such as sequencing, where performance significantly degrades by primer dimers and/or other mischief products, greater than 10, greater than 50, and greater than 100 times higher multiplexing than other described multiplexing has been achieved. Note this is opposed to probe based detection methods, *e.g.* microarrays, TAQMAN, PCR etc. where an excess of primer dimers will not affect the outcome appreciably. Also note that the general belief in the art is that multiplexing PCR for sequencing is limited to about 100 assays in the same well. Fluidigm and Rain Dance offer platforms to perform 48 or 1000s of PCR assays *in parallel reactions* for one sample.

There are a number of ways to choose primers for a library where the amount of non-mapping primer dimer or other primer mischief products are minimized. Empirical data indicate that a small number of 'bad' primers are responsible for a large amount of non-mapping primer dimer side reactions. Removing these 'bad' primers can increase the percent of sequence reads that map to targeted loci. One way to identify the 'bad' primers is to look at the sequencing data of DNA that was amplified by targeted amplification; those primer dimers that are seen with greatest frequency can be removed to give a primer library that is significantly less likely to result in side product DNA that does not map to the genome. There are also publicly available programs that can calculate the binding energy of various primer combinations, and removing those with the highest binding energy will also give a primer library that is significantly less likely to result in side product DNA that does not map to the genome.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 44

In some embodiments for selecting primers, an initial library of candidate primers is created by designing one or more primers or primer pairs to candidate target loci. A set of candidate target loci (such as SNPs) can selected based on publically available information about desired parameters for the target loci, such as frequency of the SNPs within a target population or the heterozygosity rate of the SNPs. In one embodiment, the PCR primers may be designed using the Primer3 program (the worldwide web at primer3.sourceforge.net; libprimer3 release 2.2.3, which is hereby incorporated by reference in its entirety). If desired, the primers can be designed to anneal within a particular annealing temperature range, have a particular range of GC contents, have a particular size range, produce target amplicons in a particular size range, and/or have other parameter characteristics. Starting with multiple primers or primer pairs per candidate target locus increases the likelihood that a primer or prime pair will remain in the library for most or all of the target loci. In one embodiment, the selection criteria may require that at least one primer pair per target locus remains in the library. That way, most or all of the target loci will be amplified when using the final primer library. This is desirable for applications such as screening for deletions or duplications at a large number of locations in the genome or screening for a large number of sequences (such as polymorphisms or other mutations) associated with a disease or an increased risk for a disease. If a primer pair from the library would produces a target amplicon that overlaps with a target amplicon produced by another primer pair, one of the primer pairs may be removed from the library to prevent interference.

In some embodiments, an "undesirability scores" (higher score representing least desirability) is calculated (such as calculation on a computer) for most or all of the possible combinations of two primers from a library of candidate primers. In various embodiments, an undesirability score is calculated for at least 80, 90, 95, 98, 99, or 99.5% of the possible combinations of candidate primers in the library. Each undesirability score is based at least in part on the likelihood of dimer formation between the two candidate primers. If desired, the undesirability score may also be based on one or more other parameters selected from the group consisting of heterozygosity rate of the target locus, disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, specificity of the candidate primer for the target locus, size of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, amplification efficiency of the target amplicon, and size of the

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 45

target amplicon. If multiple factors are considered, the undesirability score may be calculated based on a weighted average of the various parameters. The parameters may be assigned different weights based on their importance for the particular application that the primers will be used for. In some embodiments, the primer with the highest undesirability score is removed from the library. If the removed primer is a member of a primer pair that hybridizes to one target locus, then the other member of the primer pair may be removed from the library. The process of removing primers may be repeated as desired. In some embodiments, the selection method is performed until the undesirability scores for the candidate primer combinations remaining in the library are all equal to or below a minimum threshold. In some embodiments, the selection method is performed until the number of candidate primers remaining in the library is reduced to a desired number.

In various embodiments, after the undesirability scores are calculated, the candidate primer that is part of the greatest number of combinations of two candidate primers with an undesirability score above a first minimum threshold is removed from the library. This step ignores interactions equal to or below the first minimum threshold since these interactions are less significant. If the removed primer is a member of a primer pair that hybridizes to one target locus, then the other member of the primer pair may be removed from the library. The process of removing primers may be repeated as desired. In some embodiments, the selection method is performed until the undesirability scores for the candidate primer combinations remaining in the library are all equal to or below the first minimum threshold. If the number of candidate primers remaining in the library is higher than desired, the number of primers may be reduced by decreasing the first minimum threshold to a lower second minimum threshold and repeating the process of removing primers. If the number of candidate primers remaining in the library is lower than desired, the method can be continued by increasing the first minimum threshold to a higher second minimum threshold and repeating the process of removing primers using the original candidate primer library, thereby allowing more of the candidate primers to remain in the library. In some embodiments, the selection method is performed until the undesirability scores for the candidate primer combinations remaining in the library are all equal to or below the second minimum threshold, or until the number of candidate primers remaining in the library is reduced to a desired number.

If desired, primer pairs that produce a target amplicon that overlaps with a target amplicon produced by another primer pair can be divided into separate amplification reactions.

Multiple PCR amplification reactions may be desirable for applications in which it is desirable to analyze all of the candidate target loci (instead of omitting candidate target loci from the analysis due to overlapping target amplicons).

These selection methods minimize the number of candidate primers that have to be removed from the library to achieve the desired reduction in primer dimers. By removing a smaller number of candidate primers from the library, more (or all) of the target loci can be amplified using the resulting primer library.

Multiplexing large numbers of primers imposes considerable constraint on the assays that can be included. Assays that unintentionally interact result in spurious amplification products. The size constraints of miniPCR may result in further constraints. In an embodiment, it is possible to begin with a very large number of potential SNP targets (between about 500 to greater than 1 million) and attempt to design primers to amplify each SNP. Where primers can be designed it is possible to attempt to identify primer pairs likely to form spurious products by evaluating the likelihood of spurious primer duplex formation between all possible pairs of primers using published thermodynamic parameters for DNA duplex formation. Primer interactions may be ranked by a scoring function related to the interaction and primers with the worst interaction scores are eliminated until the number of primers desired is met. In cases where SNPs likely to be heterozygous are most useful, it is possible to also rank the list of assays and select the most heterozygous compatible assays. Experiments have validated that primers with high interaction scores are most likely to form primer dimers. At high multiplexing it is not possible to eliminate all spurious interactions, but it is essential to remove the primers or pairs of primers with the highest interaction scores *in silico* as they can dominate an entire reaction, greatly limiting amplification from intended targets. We have performed this procedure to create multiplex primer sets of up to and in some cases more than 10,000 primers. The improvement due to this procedure is substantial, enabling amplification of more than 80%, more than 90%, more than 95%, more than 98%, and even more than 99% on target products as determined by sequencing of all PCR products, as compared to 10% from a reaction in which the worst primers were not removed. When combined with a partial semi-nested approach as previously described, more than 90%, and even more than 95% of amplicons may map to the targeted sequences.

Note that there are other methods for determining which PCR probes are likely to form dimers. In an embodiment, analysis of a pool of DNA that has been amplified using a non-

optimized set of primers may be sufficient to determine problematic primers. For example, analysis may be done using sequencing, and those dimers which are present in the greatest number are determined to be those most likely to form dimers, and may be removed.

This method has a number of potential application, for example to SNP genotyping, heterozygosity rate determination, copy number measurement, and other targeted sequencing applications. In an embodiment, the method of primer design may be used in combination with the mini-PCR method described elsewhere in this document. In some embodiments, the primer design method may be used as part of a massive multiplexed PCR method.

The use of tags on the primers may reduce amplification and sequencing of primer dimer products. In some embodiments, the primer contains an internal region that forms a loop structure with a tag. In particular embodiments, the primers include a 5' region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In some embodiments, the loop region may lie between two binding regions where the two binding regions are designed to bind to contiguous or neighboring regions of template DNA. In various embodiments, the length of the 3' region is at least 7 nucleotides. In some embodiments, the length of the 3' region is between 7 and 20 nucleotides, such as between 7 to 15 nucleotides, or 7 to 10 nucleotides, inclusive. In various embodiments, the primers include a 5' region that is not specific for a target locus (such as a tag or a universal primer binding site) followed by a region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. Tag-primers can be used to shorten necessary target-specific sequences to below 20, below 15, below 12, and even below 10 base pairs. This can be serendipitous with standard primer design when the target sequence is fragmented within the primer binding site or, or it can be designed into the primer design. Advantages of this method include: it increases the number of assays that can be designed for a certain maximal amplicon length, and it shortens the "non-informative" sequencing of primer sequence. It may also be used in combination with internal tagging (see elsewhere in this document).

In an embodiment, the relative amount of nonproductive products in the multiplexed targeted PCR amplification can be reduced by raising the annealing temperature. In cases where one is amplifying libraries with the same tag as the target specific primers, the annealing temperature can be increased in comparison to the genomic DNA as the tags will contribute to

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 48

the primer binding. In some embodiments we are using considerably lower primer concentrations than previously reported along with using longer annealing times than reported elsewhere. In some embodiments the annealing times may be longer than 3 minutes, longer than 5 minutes, longer than 8 minutes, longer than 10 minutes, longer than 15 minutes, longer than 20 minutes, longer than 30 minutes, longer than 60 minutes, longer than 120 minutes, longer than 240 minutes, longer than 480 minutes, and even longer than 960 minutes. In an embodiment, longer annealing times are used than in previous reports, allowing lower primer concentrations. In various embodiments, longer than normal extension times are used, such as greater than 3, 5, 8, 10, or 15 minutes. In some embodiments, the primer concentrations are as low as 50 nM, 20 nM, 10 nM, 5 nM, 1 nM, and lower than 1 uM. This surprisingly results in robust performance for highly multiplexed reactions, for example 1,000-plex reactions, 2,000-plex reactions, 5,000-plex reactions, 10,000-plex reactions, 20,000-plex reactions, 50,000-plex reactions, and even 100,000-plex reactions. In an embodiment, the amplification uses one, two, three, four or five cycles run with long annealing times, followed by PCR cycles with more usual annealing times with tagged primers.

To select target locations, one may start with a pool of candidate primer pair designs and create a thermodynamic model of potentially adverse interactions between primer pairs, and then use the model to eliminate designs that are incompatible with other the designs in the pool.

After the selection process, the primers remaining in the library may be used in any of the methods of the invention.

Exemplary Primer Libraries

In one aspect, the invention features libraries of primers, such as primers selected from a library of candidate primers using any of the methods of the invention. In some embodiments, the library includes primers that simultaneously hybridize (or are capable of simultaneously hybridizing) to or that simultaneously amplify (or are capable of simultaneously amplifying) at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci in one reaction volume. In various embodiments, the library includes primers that simultaneously amplify (or are capable of simultaneously amplifying) between 1,000 to 2,000; 2,000 to 5,000; 5,000 to 7,500; 7,500 to 10,000; 10,000 to 20,000; 20,000 to 25,000; 25,000 to 30,000; 30,000 to 40,000; 40,000 to

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 49

50,000; 50,000 to 75,000; or 75,000 to 100,000 different target loci in one reaction volume, inclusive. In various embodiments, the library includes primers that simultaneously amplify (or are capable of simultaneously amplifying) between 1,000 to 100,000 different target loci in one reaction volume, such as between 1,000 to 50,000; 1,000 to 30,000; 1,000 to 20,000; 1,000 to 10,000; 2,000 to 30,000; 2,000 to 20,000; 2,000 to 10,000; 5,000 to 30,000; 5,000 to 20,000; or 5,000 to 10,000 different target loci, inclusive. In some embodiments, the library includes primers that simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that less than 60, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0.5, 0.25, 0.1, or 0.5% of the amplified products are primer dimers. The various embodiments, the amount of amplified products that are primer dimers is between 0.5 to 60%, such as between 0.1 to 40%, 0.1 to 20%, 0.25 to 20%, 0.25 to 10%, 0.5 to 20%, 0.5 to 10%, 1 to 20%, or 1 to 10%, inclusive. In some embodiments, the primers simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the amplified products are target amplicons. In various embodiments, the amount of amplified products that are target amplicons is between 50 to 99.5%, such as between 60 to 99%, 70 to 98%, 80 to 98%, 90 to 99.5%, or 95 to 99.5%, inclusive. In some embodiments, the primers simultaneously amplify (or are capable of simultaneously amplifying) the target loci in one reaction volume such that at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the targeted loci are amplified. In various embodiments, the amount target loci that are amplified is between 50 to 99.5%, such as between 60 to 99%, 70 to 98%, 80 to 99%, 90 to 99.5%, 95 to 99.9%, or 98 to 99.99% inclusive. In some embodiments, the library of primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 primer pairs, wherein each pair of primers includes a forward test primer and a reverse test primer where each pair of test primers hybridize to a target locus. In some embodiments, the library of primers includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 individual primers that each hybridize to a different target locus, wherein the individual primers are not part of primer pairs.

In various embodiments, the concentration of each primer is less than 100, 75, 50, 25, 20, 10, 5, 2, or 1 nM, or less than 500, 100, 10, or 1 uM. In various embodiments, the concentration of each primer is between 1 uM to 100 nM, such as between 1 uM to 1 nM, 1 to 75 nM, 2 to 50 nM or 5 to 50 nM, inclusive. In various embodiments, the GC content of the

primers is between 30 to 80%, such as between 40 to 70%, or 50 to 60%, inclusive. In some embodiments, the range of GC content of the primers is less than 30, 20, 10, or 5%. In some embodiments, the range of GC content of the primers is between 5 to 30%, such as 5 to 20% or 5 to 10%, inclusive. In some embodiments, the melting temperature (T_m) of the test primers is between 40 to 80 °C, such as 50 to 70 °C, 55 to 65 °C, or 57 to 60.5 °C, inclusive. In some embodiments, the T_m is calculated using the Primer3 program (libprimer3 release 2.2.3) using the built-in SantaLucia parameters (the world wide web at primer3.sourceforge.net). In some embodiments, the range of melting temperature of the primers is less than 15, 10, 5, 3, or 1 °C. In some embodiments, the range of melting temperature of the primers is between 1 to 15 °C, such as between 1 to 10 °C, 1 to 5 °C, or 1 to 3 °C, inclusive. In some embodiments, the length of the primers is between 15 to 100 nucleotides, such as between 15 to 75 nucleotides, 15 to 40 nucleotides, 17 to 35 nucleotides, 18 to 30 nucleotides, 20 to 65 nucleotides, inclusive. In some embodiments, the range of the length of the primers is less than 50, 40, 30, 20, 10, or 5 nucleotides. In some embodiments, the range of the length of the primers is between 5 to 50 nucleotides, such as 5 to 40 nucleotides, 5 to 20 nucleotides, or 5 to 10 nucleotides, inclusive. In some embodiments, the length of the target amplicons is between 50 and 100 nucleotides, such as between 60 and 80 nucleotides, or 60 to 75 nucleotides, inclusive. In some embodiments, the range of the length of the target amplicons is less than 50, 25, 15, 10, or 5 nucleotides. In some embodiments, the range of the length of the target amplicons is between 5 to 50 nucleotides, such as 5 to 25 nucleotides, 5 to 15 nucleotides, or 5 to 10 nucleotides, inclusive.

These primer libraries can be used in any of the methods of the invention.

Exemplary Primer Kits

In one aspect, the invention features a kit (such as kits for amplifying target loci in a nucleic acid sample) the includes any of the primer libraries of the invention. In some embodiments, a kit may be formulated that comprises a plurality of primers designed to achieve the methods described in this disclosure. The primers may be outer forward and reverse primers, inner forward and reverse primers as disclosed herein, they could be primers that have been designed to have low binding affinity to other primers in the kit as disclosed in the section on primer design, they could be hybrid capture probes or pre-circularized probes as described in the relevant sections, or some combination thereof. In an embodiment, a kit may

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 51

be formulated for determining a ploidy status of a target chromosome in a gestating fetus designed to be used with the methods disclosed herein, the kit comprising a plurality of inner forward primers and optionally the plurality of inner reverse primers, and optionally outer forward primers and outer reverse primers, where each of the primers is designed to hybridize to the region of DNA immediately upstream and/or downstream from one of the target sites (*e.g.*, polymorphic sites) on the target chromosome, and optionally additional chromosomes. In an embodiment, the primer kit may be used in combination with the diagnostic box described elsewhere in this document. In some embodiments, the kit includes instructions for using the library to amplify the target loci.

Exemplary Multiplex PCR Methods

In one aspect, the invention features methods of amplifying target loci in a nucleic acid sample that involve (i) contacting the nucleic acid sample with a library of primers that simultaneously hybridize to least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture; and (ii) subjecting the reaction mixture to primer extension reaction conditions (such as PCR conditions) to produce amplified products that include target amplicons. In some embodiments, the method also includes determining the presence or absence of at least one target amplicon (such as at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the target amplicons). In some embodiments, the method also includes determining the sequence of at least one target amplicon (such as at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the target amplicons). In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 97, 98, 99, or 99.5% of the targeted loci are amplified. In various embodiments, less than 60, 50, 40, 30, 20, 10, 5, 4, 3, 2, 1, 0.5, 0.25, 0.1, or 0.05% of the amplified products are primer dimers.

In an embodiment, a method disclosed herein uses highly efficient highly multiplexed targeted PCR to amplify DNA followed by high throughput sequencing to determine the allele frequencies at each target locus. The ability to multiplex more than about 50 or 100 PCR primers in one reaction volume in a way that most of the resulting sequence reads map to targeted loci is novel and non-obvious. One technique that allows highly multiplexed targeted PCR to perform in a highly efficient manner involves designing primers that are unlikely to hybridize with one another. The PCR probes, typically referred to as primers, are selected by creating a thermodynamic model of potentially adverse interactions between at least 500; at

least 1,000; at least 2,000; at least 5,000; at least 7,500; at least 10,000; at least 20,000; at least 25,000; at least 30,000; at least 40,000; at least 50,000; at least 75,000; or at least 100,000 potential primer pairs, or unintended interactions between primers and sample DNA, and then using the model to eliminate designs that are incompatible with other the designs in the pool. Another technique that allows highly multiplexed targeted PCR to perform in a highly efficient manner is using a partial or full nesting approach to the targeted PCR. Using one or a combination of these approaches allows multiplexing of at least 300, at least 800, at least 1,200, at least 4,000 or at least 10,000 primers in a single pool with the resulting amplified DNA comprising a majority of DNA molecules that, when sequenced, will map to targeted loci. Using one or a combination of these approaches allows multiplexing of a large number of primers in a single pool with the resulting amplified DNA comprising greater than 50%, greater than 60%, greater than 67%, greater than 80%, greater than 90%, greater than 95%, greater than 96%, greater than 97%, greater than 98%, greater than 99%, or greater than 99.5% DNA molecules that map to targeted loci.

In some embodiments the detection of the target genetic material may be done in a multiplexed fashion. The number of genetic target sequences that may be run in parallel can range from one to ten, ten to one hundred, one hundred to one thousand, one thousand to ten thousand, ten thousand to one hundred thousand, one hundred thousand to one million, or one million to ten million. Prior attempts to multiplex more than 100 primers per pool have resulted in significant problems with unwanted side reactions such as primer-dimer formation.

Targeted PCR

In some embodiments, PCR can be used to target specific locations of the genome. In plasma samples, the original DNA is highly fragmented (typically less than 500 bp, with an average length less than 200 bp). In PCR, both forward and reverse primers anneal to the same fragment to enable amplification. Therefore, if the fragments are short, the PCR assays must amplify relatively short regions as well. Like MIPS, if the polymorphic positions are too close the polymerase binding site, it could result in biases in the amplification from different alleles. Currently, PCR primers that target polymorphic regions, such as those containing SNPs, are typically designed such that the 3' end of the primer will hybridize to the base immediately adjacent to the polymorphic base or bases. In an embodiment of the present disclosure, the 3' ends of both the forward and reverse PCR primers are designed to hybridize to bases that are

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 53

one or a few positions away from the variant positions (polymorphic sites) of the targeted allele. The number of bases between the polymorphic site (SNP or otherwise) and the base to which the 3' end of the primer is designed to hybridize may be one base, it may be two bases, it may be three bases, it may be four bases, it may be five bases, it may be six bases, it may be seven to ten bases, it may be eleven to fifteen bases, or it may be sixteen to twenty bases. The forward and reverse primers may be designed to hybridize a different number of bases away from the polymorphic site.

PCR assay can be generated in large numbers, however, the interactions between different PCR assays makes it difficult to multiplex them beyond about one hundred assays. Various complex molecular approaches can be used to increase the level of multiplexing, but it may still be limited to fewer than 100, perhaps 200, or possibly 500 assays per reaction. Samples with large quantities of DNA can be split among multiple sub-reactions and then recombined before sequencing. For samples where either the overall sample or some subpopulation of DNA molecules is limited, splitting the sample would introduce statistical noise. In an embodiment, a small or limited quantity of DNA may refer to an amount below 10 pg, between 10 and 100 pg, between 100 pg and 1 ng, between 1 and 10 ng, or between 10 and 100 ng. Note that while this method is particularly useful on small amounts of DNA where other methods that involve splitting into multiple pools can cause significant problems related to introduced stochastic noise, this method still provides the benefit of minimizing bias when it is run on samples of any quantity of DNA. In these situations a universal pre-amplification step may be used to increase the overall sample quantity. Ideally, this pre-amplification step should not appreciably alter the allelic distributions.

In an embodiment, a method of the present disclosure can generate PCR products that are specific to a large number of targeted loci, specifically 1,000 to 5,000 loci, 5,000 to 10,000 loci or more than 10,000 loci, for genotyping by sequencing or some other genotyping method, from limited samples such as single cells or DNA from body fluids. Currently, performing multiplex PCR reactions of more than 5 to 10 targets presents a major challenge and is often hindered by primer side products, such as primer dimers, and other artifacts. When detecting target sequences using microarrays with hybridization probes, primer dimers and other artifacts may be ignored, as these are not detected. However, when using sequencing as a method of detection, the vast majority of the sequencing reads would sequence such artifacts and not the desired target sequences in a sample. Methods described in the prior art used to multiplex more

than 50 or 100 reactions in one reaction volume followed by sequencing will typically result in more than 20%, and often more than 50%, in many cases more than 80% and in some cases more than 90% off-target sequence reads.

In general, to perform targeted sequencing of multiple (n) targets of a sample (greater than 50, greater than 100, greater than 500, or greater than 1,000), one can split the sample into a number of parallel reactions that amplify one individual target. This has been performed in PCR multiwell plates or can be done in commercial platforms such as the FLUIDIGM ACCESS ARRAY (48 reactions per sample in microfluidic chips) or DROPLET PCR by RAIN DANCE TECHNOLOGY (100s to a few thousands of targets). Unfortunately, these split-and-pool methods are problematic for samples with a limited amount of DNA, as there is often not enough copies of the genome to ensure that there is one copy of each region of the genome in each well. This is an especially severe problem when polymorphic loci are targeted, and the relative proportions of the alleles at the polymorphic loci are needed, as the stochastic noise introduced by the splitting and pooling will cause very poorly accurate measurements of the proportions of the alleles that were present in the original sample of DNA. Described here is a method to effectively and efficiently amplify many PCR reactions that is applicable to cases where only a limited amount of DNA is available. In an embodiment, the method may be applied for analysis of single cells, body fluids, mixtures of DNA such as the free floating DNA found in maternal plasma, biopsies, environmental and/or forensic samples.

In an embodiment, the targeted sequencing may involve one, a plurality, or all of the following steps. a) Generate and amplify a library with adaptor sequences on both ends of DNA fragments. b) Divide into multiple reactions after library amplification. c) Generate and optionally amplify a library with adaptor sequences on both ends of DNA fragments. d) Perform 1000- to 10,000-plex amplification of selected targets using one target specific "Forward" primer per target and one tag specific primer. e) Perform a second amplification from this product using "Reverse" target specific primers and one (or more) primer specific to a universal tag that was introduced as part of the target specific forward primers in the first round. f) Perform a 1000-plex preamplification of selected target for a limited number of cycles. g) Divide the product into multiple aliquots and amplify subpools of targets in individual reactions (for example, 50 to 500-plex, though this can be used all the way down to singleplex. h) Pool products of parallel subpools reactions. i) During these amplifications

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 55

primers may carry sequencing compatible tags (partial or full length) such that the products can be sequenced.

Highly Multiplexed PCR

Disclosed herein are methods that permit the targeted amplification of over a hundred to tens of thousands of target sequences (*e.g.,* SNP loci) from a nucleic acid sample such as genomic DNA obtained from plasma. The amplified sample may be relatively free of primer dimer products and have low allelic bias at target loci. If during or after amplification the products are appended with sequencing compatible adaptors, analysis of these products can be performed by sequencing.

Performing a highly multiplexed PCR amplification using methods known in the art results in the generation of primer dimer products that are in excess of the desired amplification products and not suitable for sequencing. These can be reduced empirically by eliminating primers that form these products, or by performing *in silico* selection of primers. However, the larger the number of assays, the more difficult this problem becomes.

One solution is to split the 5000-plex reaction into several lower-plexed amplifications, *e.g.* one hundred 50-plex or fifty 100-plex reactions, or to use microfluidics or even to split the sample into individual PCR reactions. However, if the sample DNA is limited, such as in non-invasive prenatal diagnostics from pregnancy plasma, dividing the sample between multiple reactions should be avoided as this will result in bottlenecking.

Described herein are methods to first globally amplify the plasma DNA of a sample and then divide the sample up into multiple multiplexed target enrichment reactions with more moderate numbers of target sequences per reaction. In an embodiment, a method of the present disclosure can be used for preferentially enriching a DNA mixture at a plurality of loci, the method comprising one or more of the following steps: generating and amplifying a library from a mixture of DNA where the molecules in the library have adaptor sequences ligated on both ends of the DNA fragments, dividing the amplified library into multiple reactions, performing a first round of multiplex amplification of selected targets using one target specific "forward" primer per target and one or a plurality of adaptor specific universal "reverse" primers. In an embodiment, a method of the present disclosure further includes performing a second amplification using "reverse" target specific primers and one or a plurality of primers specific to a universal tag that was introduced as part of the target specific forward primers in

the first round. In an embodiment, the method may involve a fully nested, hemi-nested, semi-nested, one sided fully nested, one sided hemi-nested, or one sided semi-nested PCR approach. In an embodiment, a method of the present disclosure is used for preferentially enriching a DNA mixture at a plurality of loci, the method comprising performing a multiplex preamplification of selected targets for a limited number of cycles, dividing the product into multiple aliquots and amplifying subpools of targets in individual reactions, and pooling products of parallel subpools reactions. Note that this approach could be used to perform targeted amplification in a manner that would result in low levels of allelic bias for 50-500 loci, for 500 to 5,000 loci, for 5,000 to 50,000 loci, or even for 50,000 to 500,000 loci. In an embodiment, the primers carry partial or full length sequencing compatible tags.

The workflow may entail (1) extracting DNA such as plasma DNA, (2) preparing fragment library with universal adaptors on both ends of fragments, (3) amplifying the library using universal primers specific to the adaptors, (4) dividing the amplified sample "library" into multiple aliquots, (5) performing multiplex (*e.g.* about 100-plex, 1,000, or 10,000-plex with one target specific primer per target and a tag-specific primer) amplifications on aliquots, (6) pooling aliquots of one sample, (7) barcoding the sample, (8) mixing the samples and adjusting the concentration, (9) sequencing the sample. The workflow may comprise multiple sub-steps that contain one of the listed steps (*e.g.* step (2) of preparing the library step could entail three enzymatic steps (blunt ending, dA tailing and adaptor ligation) and three purification steps). Steps of the workflow may be combined, divided up or performed in different order (*e.g.* bar coding and pooling of samples).

It is important to note that the amplification of a library can be performed in such a way that it is biased to amplify short fragments more efficiently. In this manner it is possible to preferentially amplify shorter sequences, *e.g.* mono-nucleosomal DNA fragments as the cell free fetal DNA (of placental origin) found in the circulation of pregnant women. Note that PCR assays can have the tags, for example sequencing tags, (usually a truncated form of 15-25 bases). After multiplexing, PCR multiplexes of a sample are pooled and then the tags are completed (including bar coding) by a tag-specific PCR (could also be done by ligation). Also, the full sequencing tags can be added in the same reaction as the multiplexing. In the first cycles targets may be amplified with the target specific primers, subsequently the tag-specific primers take over to complete the SQ-adaptor sequence. The PCR primers may carry no tags. The sequencing tags may be appended to the amplification products by ligation.

In an embodiment, highly multiplex PCR followed by evaluation of amplified material by clonal sequencing may be used for various applications such as the detection of fetal aneuploidy. Whereas traditional multiplex PCRs evaluate up to fifty loci simultaneously, the approach described herein may be used to enable simultaneous evaluation of more than 50 loci simultaneously, more than 100 loci simultaneously, more than 500 loci simultaneously, more than 1,000 loci simultaneously, more than 5,000 loci simultaneously, more than 10,000 loci simultaneously, more than 50,000 loci simultaneously, and more than 100,000 loci simultaneously. Experiments have shown that up to, including and more than 10,000 distinct loci can be evaluated simultaneously, in a single reaction, with sufficiently good efficiency and specificity to make non-invasive prenatal aneuploidy diagnoses and/or copy number calls with high accuracy. Assays may be combined in a single reaction with the entirety of a sample such as a cfDNA sample isolated from maternal plasma, a fraction thereof, or a further processed derivative of the cfDNA sample. The sample (*e.g.*, cfDNA or derivative) may also be split into multiple parallel multiplex reactions. The optimum sample splitting and multiplex is determined by trading off various performance specifications. Due to the limited amount of material, splitting the sample into multiple fractions can introduce sampling noise, handling time, and increase the possibility of error. Conversely, higher multiplexing can result in greater amounts of spurious amplification and greater inequalities in amplification both of which can reduce test performance.

Two crucial related considerations in the application of the methods described herein are the limited amount of original sample (*e.g.*, plasma) and the number of original molecules in that material from which allele frequency or other measurements are obtained. If the number of original molecules falls below a certain level, random sampling noise becomes significant, and can affect the accuracy of the test. Typically, data of sufficient quality for making non-invasive prenatal aneuploidy diagnoses can be obtained if measurements are made on a sample comprising the equivalent of 500-1000 original molecules per target locus. There are a number of ways of increasing the number of distinct measurements, for example increasing the sample volume. Each manipulation applied to the sample also potentially results in losses of material. It is essential to characterize losses incurred by various manipulations and avoid, or as necessary improve yield of certain manipulations to avoid losses that could degrade performance of the test.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 58

In an embodiment, it is possible to mitigate potential losses in subsequent steps by amplifying all or a fraction of the original sample (*e.g.*, cfDNA sample). Various methods are available to amplify all of the genetic material in a sample, increasing the amount available for downstream procedures. In an embodiment, ligation mediated PCR (LM-PCR) DNA fragments are amplified by PCR after ligation of either one distinct adaptors, two distinct adapters, or many distinct adaptors. In an embodiment, multiple displacement amplification (MDA) phi-29 polymerase is used to amplify all DNA isothermally. In DOP-PCR and variations, random priming is used to amplify the original material DNA. Each method has certain characteristics such as uniformity of amplification across all represented regions of the genome, efficiency of capture and amplification of original DNA, and amplification performance as a function of the length of the fragment.

In an embodiment LM-PCR may be used with a single heteroduplexed adaptor having a 3-prime tyrosine. The heteroduplexed adaptor enables the use of a single adaptor molecule that may be converted to two distinct sequences on 5-prime and 3-prime ends of the original DNA fragment during the first round of PCR. In an embodiment, it is possible to fractionate the amplified library by size separations, or products such as AMPURE, TASS or other similar methods. Prior to ligation, sample DNA may be blunt ended, and then a single adenosine base is added to the 3-prime end. Prior to ligation the DNA may be cleaved using a restriction enzyme or some other cleavage method. During ligation the 3-prime adenosine of the sample fragments and the complementary 3-prime tyrosine overhang of adaptor can enhance ligation efficiency. The extension step of the PCR amplification may be limited from a time standpoint to reduce amplification from fragments longer than about 200 bp, about 300 bp, about 400 bp, about 500 bp or about 1,000 bp. Since longer DNA found in the maternal plasma is nearly exclusively maternal, this may result in the enrichment of fetal DNA by 10-50% and improvement of test performance. A number of reactions were run using conditions as specified by commercially available kits; the resulted in successful ligation of fewer than 10% of sample DNA molecules. A series of optimizations of the reaction conditions for this improved ligation to approximately 70%.

Mini-PCR

The following Mini-PCR method is desirable for samples containing short nucleic acids, digested nucleic acids, or fragmented nucleic acids, such as cfDNA. Traditional PCR

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 59

assay design results in significant losses of distinct fetal molecules, but losses can be greatly reduced by designing very short PCR assays, termed mini-PCR assays. Fetal cfDNA in maternal serum is highly fragmented and the fragment sizes are distributed in approximately a Gaussian fashion with a mean of 160 bp, a standard deviation of 15 bp, a minimum size of about 100 bp, and a maximum size of about 220 bp. The distribution of fragment start and end positions with respect to the targeted polymorphisms, while not necessarily random, vary widely among individual targets and among all targets collectively and the polymorphic site of one particular target locus may occupy any position from the start to the end among the various fragments originating from that locus. Note that the term mini-PCR may equally well refer to normal PCR with no additional restrictions or limitations.

During PCR, amplification will only occur from template DNA fragments comprising both forward and reverse primer sites. Because fetal cfDNA fragments are short, the likelihood of both primer sites being present the likelihood of a fetal fragment of length L comprising both the forward and reverse primers sites is ratio of the length of the amplicon to the length of the fragment. Under ideal conditions, assays in which the amplicon is 45, 50, 55, 60, 65, or 70 bp will successfully amplify from 72%, 69%, 66%, 63%, 59%, or 56%, respectively, of available template fragment molecules. The amplicon length is the distance between the 5-prime ends of the forward and reverse priming sites. Amplicon length that is shorter than typically used by those known in the art may result in more efficient measurements of the desired polymorphic loci by only requiring short sequence reads. In an embodiment, a substantial fraction of the amplicons should be less than 100 bp, less than 90 bp, less than 80 bp, less than 70 bp, less than 65 bp, less than 60 bp, less than 55 bp, less than 50 bp, or less than 45 bp.

Note that in methods known in the prior art, short assays such as those described herein are usually avoided because they are not required and they impose considerable constraint on primer design by limiting primer length, annealing characteristics, and the distance between the forward and reverse primer.

Also note that there is the potential for biased amplification if the 3-prime end of the either primer is within roughly 1-6 bases of the polymorphic site. This single base difference at the site of initial polymerase binding can result in preferential amplification of one allele, which can alter observed allele frequencies and degrade performance. All of these constraints make it very challenging to identify primers that will amplify a particular locus successfully

and furthermore, to design large sets of primers that are compatible in the same multiplex reaction. In an embodiment, the 3' end of the inner forward and reverse primers are designed to hybridize to a region of DNA upstream from the polymorphic site, and separated from the polymorphic site by a small number of bases. Ideally, the number of bases may be between 6 and 10 bases, but may equally well be between 4 and 15 bases, between three and 20 bases, between two and 30 bases, or between 1 and 60 bases, and achieve substantially the same end.

Multiplex PCR may involve a single round of PCR in which all targets are amplified or it may involve one round of PCR followed by one or more rounds of nested PCR or some variant of nested PCR. Nested PCR consists of a subsequent round or rounds of PCR amplification using one or more new primers that bind internally, by at least one base pair, to the primers used in a previous round. Nested PCR reduces the number of spurious amplification targets by amplifying, in subsequent reactions, only those amplification products from the previous one that have the correct internal sequence. Reducing spurious amplification targets improves the number of useful measurements that can be obtained, especially in sequencing. Nested PCR typically entails designing primers completely internal to the previous primer binding sites, necessarily increasing the minimum DNA segment size required for amplification. For samples such as maternal plasma cfDNA, in which the DNA is highly fragmented, the larger assay size reduces the number of distinct cfDNA molecules from which a measurement can be obtained. In an embodiment, to offset this effect, one may use a partial nesting approach where one or both of the second round primers overlap the first binding sites extending internally some number of bases to achieve additional specificity while minimally increasing in the total assay size.

In an embodiment, a multiplex pool of PCR assays are designed to amplify potentially heterozygous SNP or other polymorphic or non-polymorphic loci on one or more chromosomes and these assays are used in a single reaction to amplify DNA. The number of PCR assays may be between 50 and 200 PCR assays, between 200 and 1,000 PCR assays, between 1,000 and 5,000 PCR assays, or between 5,000 and 20,000 PCR assays (50 to 200-plex, 200 to 1,000-plex, 1,000 to 5,000-plex, 5,000 to 20,000-plex, more than 20,000-plex respectively). In an embodiment, a multiplex pool of about 10,000 PCR assays (10,000-plex) are designed to amplify potentially heterozygous SNP loci on chromosomes X, Y, 13, 18, and 21 and 1 or 2 and these assays are used in a single reaction to amplify cfDNA obtained from a material plasma sample, chorion villus samples, amniocentesis samples, single or a small

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 61

number of cells, other bodily fluids or tissues, cancers, or other genetic matter. The SNP frequencies of each locus may be determined by clonal or some other method of sequencing of the amplicons. Statistical analysis of the allele frequency distributions or ratios of all assays may be used to determine if the sample contains a trisomy of one or more of the chromosomes included in the test. In another embodiment the original cfDNA samples is split into two samples and parallel 5,000-plex assays are performed. In another embodiment the original cfDNA samples is split into n samples and parallel (~10,000/n)-plex assays are performed where n is between 2 and 12, or between 12 and 24, or between 24 and 48, or between 48 and 96. Data is collected and analyzed in a similar manner to that already described. Note that this method is equally well applicable to detecting translocations, deletions, duplications, and other chromosomal abnormalities.

In an embodiment, tails with no homology to the target genome may also be added to the 3-prime or 5-prime end of any of the primers. These tails facilitate subsequent manipulations, procedures, or measurements. In an embodiment, the tail sequence can be the same for the forward and reverse target specific primers. In an embodiment, different tails may be used for the forward and reverse target specific primers. In an embodiment, a plurality of different tails may be used for different loci or sets of loci. Certain tails may be shared among all loci or among subsets of loci. For example, using forward and reverse tails corresponding to forward and reverse sequences required by any of the current sequencing platforms can enable direct sequencing following amplification. In an embodiment, the tails can be used as common priming sites among all amplified targets that can be used to add other useful sequences. In some embodiments, the inner primers may contain a region that is designed to hybridize either upstream or downstream of the targeted locus (*e.g*, a polymorphic locus). In some embodiments, the primers may contain a molecular barcode. In some embodiments, the primer may contain a universal priming sequence designed to allow PCR amplification.

In an embodiment, a 10,000-plex PCR assay pool is created such that forward and reverse primers have tails corresponding to the required forward and reverse sequences required by a high throughput sequencing instrument such as the HISEQ, GAIIX, or MYSEQ available from ILLUMINA. In addition, included 5-prime to the sequencing tails is an additional sequence that can be used as a priming site in a subsequent PCR to add nucleotide barcode sequences to the amplicons, enabling multiplex sequencing of multiple samples in a single lane of the high throughput sequencing instrument.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 62

In an embodiment, a 10,000-plex PCR assay pool is created such that reverse primers have tails corresponding to the required reverse sequences required by a high throughput sequencing instrument. After amplification with the first 10,000-plex assay, a subsequent PCR amplification may be performed using a another 10,000-plex pool having partly nested forward primers (*e.g.* 6-bases nested) for all targets and a reverse primer corresponding to the reverse sequencing tail included in the first round. This subsequent round of partly nested amplification with just one target specific primer and a universal primer limits the required size of the assay, reducing sampling noise, but greatly reduces the number of spurious amplicons. The sequencing tags can be added to appended ligation adaptors and/or as part of PCR probes, such that the tag is part of the final amplicon.

Fetal fraction affects performance of the test. There are a number of ways to enrich the fetal fraction of the DNA found in maternal plasma. Fetal fraction can be increased by the previously described LM-PCR method already discussed as well as by a targeted removal of long maternal fragments. In an embodiment, prior to multiplex PCR amplification of the target loci, an additional multiplex PCR reaction may be carried out to selectively remove long and largely maternal fragments corresponding to the loci targeted in the subsequent multiplex PCR. Additional primers are designed to anneal a site a greater distance from the polymorphism than is expected to be present among cell free fetal DNA fragments. These primers may be used in a one cycle multiplex PCR reaction prior to multiplex PCR of the target polymorphic loci. These distal primers are tagged with a molecule or moiety that can allow selective recognition of the tagged pieces of DNA. In an embodiment, these molecules of DNA may be covalently modified with a biotin molecule that allows removal of newly formed double stranded DNA comprising these primers after one cycle of PCR. Double stranded DNA formed during that first round is likely maternal in origin. Removal of the hybrid material may be accomplish by the used of magnetic streptavidin beads. There are other methods of tagging that may work equally well. In an embodiment, size selection methods may be used to enrich the sample for shorter strands of DNA; for example those less than about 800 bp, less than about 500 bp, or less than about 300 bp. Amplification of short fragments can then proceed as usual.

The mini-PCR method described in this disclosure enables highly multiplexed amplification and analysis of hundreds to thousands or even millions of loci in a single reaction, from a single sample. At the same, the detection of the amplified DNA can be multiplexed; tens to hundreds of samples can be multiplexed in one sequencing lane by using

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 63

barcoding PCR. This multiplexed detection has been successfully tested up to 49-plex, and a much higher degree of multiplexing is possible. In effect, this allows hundreds of samples to be genotyped at thousands of SNPs in a single sequencing run. For these samples, the method allows determination of genotype and heterozygosity rate and simultaneously determination of copy number, both of which may be used for the purpose of aneuploidy detection. This method is particularly useful in detecting aneuploidy of a gestating fetus from the free floating DNA found in maternal plasma. This method may be used as part of a method for sexing a fetus, and/or predicting the paternity of the fetus. It may be used as part of a method for mutation dosage. This method may be used for any amount of DNA or RNA, and the targeted regions may be SNPs, other polymorphic regions, non-polymorphic regions, and combinations thereof.

In some embodiments, ligation mediated universal-PCR amplification of fragmented DNA may be used. The ligation mediated universal-PCR amplification can be used to amplify plasma DNA, which can then be divided into multiple parallel reactions. It may also be used to preferentially amplify short fragments, thereby enriching fetal fraction. In some embodiments the addition of tags to the fragments by ligation can enable detection of shorter fragments, use of shorter target sequence specific portions of the primers and/or annealing at higher temperatures which reduces unspecific reactions.

The methods described herein may be used for a number of purposes where there is a target set of DNA that is mixed with an amount of contaminating DNA. In some embodiments, the target DNA and the contaminating DNA may be from individuals who are genetically related. For example, genetic abnormalities in a fetus (target) may be detected from maternal plasma which contains fetal (target) DNA and also maternal (contaminating) DNA; the abnormalities include whole chromosome abnormalities (*e.g.* aneuploidy) partial chromosome abnormalities (*e.g.* deletions, duplications, inversions, translocations), polynucleotide polymorphisms (*e.g.* STRs), single nucleotide polymorphisms, and/or other genetic abnormalities or differences. In some embodiments, the target and contaminating DNA may be from the same individual, but where the target and contaminating DNA are different by one or more mutations, for example in the case of cancer. (see *e.g.* H. Mamon et al. *Preferential Amplification of Apoptotic DNA from Plasma: Potential for Enhancing Detection of Minor DNA Alterations in Circulating DNA*. Clinical Chemistry 54:9 (2008). In some embodiments, the DNA may be found in cell culture (apoptotic) supernatant. In some embodiments, it is possible to induce apoptosis in biological samples (*e.g.,* blood) for subsequent library

preparation, amplification and/or sequencing. A number of enabling workflows and protocols to achieve this end are presented elsewhere in this disclosure.

In some embodiments, the target DNA may originate from single cells, from samples of DNA consisting of less than one copy of the target genome, from low amounts of DNA, from DNA from mixed origin (*e.g.* pregnancy plasma: placental and maternal DNA; cancer patient plasma and tumors: mix between healthy and cancer DNA, transplantation *etc*), from other body fluids, from cell cultures, from culture supernatants, from forensic samples of DNA, from ancient samples of DNA (*e.g.* insects trapped in amber), from other samples of DNA, and combinations thereof.

In some embodiments, a short amplicon size may be used. Short amplicon sizes are especially suited for fragmented DNA (see *e.g.* A. Sikora, et sl. Detection of increased amounts of cell-free fetal DNA with short PCR amplicons. Clin Chem. 2010 Jan;56(1):136-8.)

The use of short amplicon sizes may result in some significant benefits. Short amplicon sizes may result in optimized amplification efficiency. Short amplicon sizes typically produce shorter products, therefore there is less chance for nonspecific priming. Shorter products can be clustered more densely on sequencing flow cell, as the clusters will be smaller. Note that the methods described herein may work equally well for longer PCR amplicons. Amplicon length may be increased if necessary, for example, when sequencing larger sequence stretches. Experiments with 146-plex targeted amplification with assays of 100 bp to 200 bp length as first step in a nested-PCR protocol were run on single cells and on genomic DNA with positive results.

In some embodiments, the methods described herein may be used to amplify and/or detect SNPs, copy number, nucleotide methylation, mRNA levels, other types of RNA expression levels, other genetic and/or epigenetic features. The mini-PCR methods described herein may be used along with next-generation sequencing; it may be used with other downstream methods such as microarrays, counting by digital PCR, real-time PCR, Mass-spectrometry analysis etc.

In some embodiment, the mini-PCR amplification methods described herein may be used as part of a method for accurate quantification of minority populations. It may be used for absolute quantification using spike calibrators. It may be used for mutation / minor allele quantification through very deep sequencing, and may be run in a highly multiplexed fashion. It may be used for standard paternity and identity testing of relatives or ancestors, in human,

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 65

animals, plants or other creatures. It may be used for forensic testing. It may be used for rapid genotyping and copy number analysis (CN), on any kind of material, *e.g.* amniotic fluid and CVS, sperm, product of conception (POC). It may be used for single cell analysis, such as genotyping on samples biopsied from embryos. It may be used for rapid embryo analysis (within less than one, one, or two days of biopsy) by targeted sequencing using min-PCR.

In some embodiments, it may be used for tumor analysis: tumor biopsies are often a mixture of health and tumor cells. Targeted PCR allows deep sequencing of SNPs and loci with close to no background sequences. It may be used for copy number and loss of heterozygosity analysis on tumor DNA. Said tumor DNA may be present in many different body fluids or tissues of tumor patients. It may be used for detection of tumor recurrence, and/or tumor screening. It may be used for quality control testing of seeds. It may be used for breeding, or fishing purposes. Note that any of these methods could equally well be used targeting non-polymorphic loci for the purpose of ploidy calling.

Some literature describing some of the fundamental methods that underlie the methods disclosed herein include: (1) Wang HY, Luo M, Tereshchenko IV, Frikker DM, Cui X, Li JY, Hu G, Chu Y, Azaro MA, Lin Y, Shen L, Yang Q, Kambouris ME, Gao R, Shih W, Li H. Genome Res. 2005 Feb;15(2):276-83. Department of Molecular Genetics, Microbiology and Immunology/The Cancer Institute of New Jersey, Robert Wood Johnson Medical School, New Brunswick, New Jersey 08903, USA. (2) High-throughput genotyping of single nucleotide polymorphisms with high sensitivity. Li H, Wang HY, Cui X, Luo M, Hu G, Greenawalt DM, Tereshchenko IV, Li JY, Chu Y, Gao R. Methods Mol Biol. 2007;396 - PubMed PMID: 18025699. (3) A method comprising multiplexing of an average of 9 assays for sequencing is described in: Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Varley KE, Mitra RD. Genome Res. 2008 Nov;18(11):1844-50. Epub 2008 Oct 10. Note that the methods disclosed herein allow multiplexing of orders of magnitude more than in the above references.

Targeted PCR Variants - Nesting

There are many workflows that are possible when conducting PCR; some workflows typical to the methods disclosed herein are described. The steps outlined herein are not meant to exclude other possible steps nor does it imply that any of the steps described herein are required for the method to work properly. A large number of parameter variations or other

modifications are known in the literature, and may be made without affecting the essence of the invention. One particular generalized workflow is given below followed by a number of possible variants. The variants typically refer to possible secondary PCR reactions, for example different types of nesting that may be done (step 3). It is important to note that variants may be done at different times, or in different orders than explicitly described herein. Examples that use polymorphic loci for illustration can be readily adapted for the amplification of nonpolymorphic loci if desired.

1. The DNA in the sample may have ligation adapters, often referred to as library tags or ligation adaptor tags (LTs), appended, where the ligation adapters contain a universal priming sequence, followed by a universal amplification. In an embodiment, this may be done using a standard protocol designed to create sequencing libraries after fragmentation. In an embodiment, the DNA sample can be blunt ended, and then an A can be added at the 3' end. A Y-adaptor with a T-overhang can be added and ligated. In some embodiments, other sticky ends can be used other than an A or T overhang. In some embodiments, other adaptors can be added, for example looped ligation adaptors. In some embodiments, the adaptors may have tag designed for PCR amplification.

2. Specific Target Amplification (STA): Pre-amplification of hundreds to thousands to tens of thousands and even hundreds of thousands of targets may be multiplexed in one reaction volume. STA is typically run from 10 to 30 cycles, though it may be run from 5 to 40 cycles, from 2 to 50 cycles, and even from 1 to 100 cycles. Primers may be tailed, for example for a simpler workflow or to avoid sequencing of a large proportion of dimers. Note that typically, dimers of both primers carrying the same tag will not be amplified or sequenced efficiently. In some embodiments, between 1 and 10 cycles of PCR may be carried out; in some embodiments between 10 and 20 cycles of PCR may be carried out; in some embodiments between 20 and 30 cycles of PCR may be carried out; in some embodiments between 30 and 40 cycles of PCR may be carried out; in some embodiments more than 40 cycles of PCR may be carried out. The amplification may be a linear amplification. The number of PCR cycles may be optimized to result in an optimal depth of read (DOR) profile. Different DOR profiles may be desirable for different purposes. In some embodiments, a more even distribution of reads between all assays is desirable; if the DOR is too small for some assays, the stochastic noise can be too high for the data to be too useful, while if the depth of read is too high, the marginal usefulness of each additional read is relatively small.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 67

Primer tails may improve the detection of fragmented DNA from universally tagged libraries. If the library tag and the primer-tails contain a homologous sequence, hybridization can be improved (for example, melting temperature (T_M) is lowered) and primers can be extended if only a portion of the primer target sequence is in the sample DNA fragment. In some embodiments, 13 or more target specific base pairs may be used. In some embodiments, 10 to 12 target specific base pairs may be used. In some embodiments, 8 to 9 target specific base pairs may be used. In some embodiments, 6 to 7 target specific base pairs may be used. In some embodiments, STA may be performed on pre-amplified DNA, *e.g.* MDA, RCA, other whole genome amplifications, or adaptor-mediated universal PCR. In some embodiments, STA may be performed on samples that are enriched or depleted of certain sequences and populations, *e.g.* by size selection, target capture, directed degradation.

3. In some embodiments, it is possible to perform secondary multiplex PCRs or primer extension reactions to increase specificity and reduce undesirable products. For example, full nesting, semi-nesting, hemi-nesting, and/or subdividing into parallel reactions of smaller assay pools are all techniques that may be used to increase specificity. Experiments have shown that splitting a sample into three 400-plex reactions resulted in product DNA with greater specificity than one 1,200-plex reaction with exactly the same primers. Similarly, experiments have shown that splitting a sample into four 2,400-plex reactions resulted in product DNA with greater specificity than one 9,600-plex reaction with exactly the same primers. In an embodiment, it is possible to use target-specific and tag specific primers of the same and opposing directionality.

4. In some embodiments, it is possible to amplify a DNA sample (dilution, purified or otherwise) produced by an STA reaction using tag-specific primers and "universal amplification", *i.e.* to amplify many or all pre-amplified and tagged targets. Primers may contain additional functional sequences, *e.g.* barcodes, or a full adaptor sequence necessary for sequencing on a high throughput sequencing platform.

These methods may be used for analysis of any sample of DNA, and are especially useful when the sample of DNA is particularly small, or when it is a sample of DNA where the DNA originates from more than one individual, such as in the case of maternal plasma. These methods may be used on DNA samples such as a single or small number of cells, genomic DNA, plasma DNA, amplified plasma libraries, amplified apoptotic supernatant libraries, or other samples of mixed DNA. In an embodiment, these methods may be used in the case where

cells of different genetic constitution may be present in a single individual, such as with cancer or transplants.

Protocol variants (variants and/or additions to the workflow above)

Direct multiplexed mini-PCR: Specific target amplification (STA) of a plurality of target sequences with tagged primers is shown in **Figure 1**. 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with PCR primers hybridized. 104 denotes the final PCR product. In some embodiments, STA may be done on more than 100, more than 200, more than 500, more than 1,000, more than 2,000, more than 5,000, more than 10,000, more than 20,000, more than 50,000, more than 100,000 or more than 200,000 targets. In a subsequent reaction, tag-specific primers amplify all target sequences and lengthen the tags to include all necessary sequences for sequencing, including sample indexes. In an embodiment, primers may not be tagged or only certain primers may be tagged. Sequencing adaptors may be added by conventional adaptor ligation. In an embodiment, the initial primers may carry the tags.

In an embodiment, primers are designed so that the length of DNA amplified is unexpectedly short. Prior art demonstrates that ordinary people skilled in the art typically design 100+ bp amplicons. In an embodiment, the amplicons may be designed to be less than 80 bp. In an embodiment, the amplicons may be designed to be less than 70 bp. In an embodiment, the amplicons may be designed to be less than 60 bp. In an embodiment, the amplicons may be designed to be less than 50 bp. In an embodiment, the amplicons may be designed to be less than 45 bp. In an embodiment, the amplicons may be designed to be less than 40 bp. In an embodiment, the amplicons may be designed to be less than 35 bp. In an embodiment, the amplicons may be designed to be between 40 and 65 bp.

An experiment was performed using this protocol using 1200-plex amplification. Both genomic DNA and pregnancy plasma were used; about 70% of sequence reads mapped to targeted sequences. Details are given elsewhere in this document. Sequencing of a 1042-plex without design and selection of assays resulted in >99% of sequences being primer dimer products.

Sequential PCR: After STA1 multiple aliquots of the product may be amplified in parallel with pools of reduced complexity with the same primers. The first amplification can

give enough material to split. This method is especially good for small samples, for example those that are about 6-100 pg, about 100 pg to 1 ng, about 1 ng to 10 ng, or about 10 ng to 100 ng. The protocol was performed with 1200-plex into three 400-plexes. Mapping of sequencing reads increased from around 60 to 70 % in the 1200-plex alone to over 95%.

Semi-nested mini-PCR: (see **Figure 2**) After STA 1 a second STA is performed comprising a multiplex set of internal nested Forward primers (103 B, 105 b) and one (or few) tag-specific Reverse primers (103 A). 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with Forward primer B and Reverse Primer A hybridized. 104 denotes the PCR product from 103. 105 denotes the product from 104 with nested Forward primer b hybridized, and Reverse tag A already part of the molecule from the PCR that occurred between 103 and 104. 106 denotes the final PCR product. With this workflow usually greater than 95% of sequences map to the intended targets. The nested primer may overlap with the outer Forward primer sequence but introduces additional 3'-end bases. In some embodiments it is possible to use between one and 20 extra 3' bases. Experiments have shown that using 9 or more extra 3' bases in a 1200-plex designs works well.

Fully nested mini-PCR: (see **Figure 3**) After STA step 1, it is possible to perform a second multiplex PCR (or parallel m.p. PCRs of reduced complexity) with two nested primers carrying tags (A, a, B, b). 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with Forward primer B and Reverse Primer A hybridized. 104 denotes the PCR product from 103. 105 denotes the product from 104 with nested Forward primer b and nested Reverse primer a hybridized. 106 denotes the final PCR product. In some embodiments, it is possible to use two full sets of primers. Experiments using a fully nested mini-PCR protocol were used to perform 146-plex amplification on single and three cells without step 102 of appending universal ligation adaptors and amplifying.

Hemi-nested mini-PCR: (see **Figure 4**) It is possible to use target DNA that has and adaptors at the fragment ends. STA is performed comprising a multiplex set of Forward primers (B) and one (or few) tag-specific Reverse primers (A). A second STA can be performed using a universal tag-specific Forward primer and target specific Reverse primer.

101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with Reverse Primer A hybridized. 104 denotes the PCR product from 103 that was amplified using Reverse primer A and ligation adaptor tag primer LT. 105 denotes the product from 104 with Forward primer B hybridized. 106 denotes the final PCR product. In this workflow, target specific Forward and Reverse primers are used in separate reactions, thereby reducing the complexity of the reaction and preventing dimer formation of forward and reverse primers. Note that in this example, primers A and B may be considered to be first primers, and primers 'a' and 'b' may be considered to be inner primers. This method is a big improvement on direct PCR as it is as good as direct PCR, but it avoids primer dimers. After first round of hemi nested protocol one typically sees ~99% non-targeted DNA, however, after second round there is typically a big improvement.

Triply hemi-nested mini-PCR: (see **Figure 5**) It is possible to use target DNA that has and adaptor at the fragment ends. STA is performed comprising a multiplex set of Forward primers (B) and one (or few) tag-specific Reverse primers (A) and (a). A second STA can be performed using a universal tag-specific Forward primer and target specific Reverse primer. 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with Reverse Primer A hybridized. 104 denotes the PCR product from 103 that was amplified using Reverse primer A and ligation adaptor tag primer LT. 105 denotes the product from 104 with Forward primer B hybridized. 106 denotes the PCR product from 105 that was amplified using Reverse primer A and Forward primer B. 107 denotes the product from 106 with Reverse primer 'a' hybridized. 108 denotes the final PCR product. Note that in this example, primers 'a' and B may be considered to be inner primers, and A may be considered to be a first primer. Optionally, both A and B may be considered to be first primers, and 'a' may be considered to be an inner primer. The designation of reverse and forward primers may be switched. In this workflow, target specific Forward and Reverse primers are used in separate reactions, thereby reducing the complexity of the reaction and preventing dimer formation of forward and reverse primers. This method is a big improvement on direct PCR as it is as good as direct PCR, but it avoids primer dimers. After first round of hemi nested protocol one typically sees ~99% non-targeted DNA, however, after second round there is typically a big improvement.

One-sided nested mini-PCR: (see **Figure 6**) It is possible to use target DNA that has an adaptor at the fragment ends. STA may also be performed with a multiplex set of nested Forward primers and using the ligation adapter tag as the Reverse primer. A second STA may then be performed using a set of nested Forward primers and a universal Reverse primer. 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA that has been universally amplified with Forward Primer A hybridized. 104 denotes the PCR product from 103 that was amplified using Forward primer A and ligation adaptor tag Reverse primer LT. 105 denotes the product from 104 with nested Forward primer a hybridized. 106 denotes the final PCR product. This method can detect shorter target sequences than standard PCR by using overlapping primers in the first and second STAs. The method is typically performed off a sample of DNA that has already undergone STA step 1 above – appending of universal tags and amplification; the two nested primers are only on one side, other side uses the library tag. The method was performed on libraries of apoptotic supernatants and pregnancy plasma. With this workflow around 60% of sequences mapped to the intended targets. Note that reads that contained the reverse adaptor sequence were not mapped, so this number is expected to be higher if those reads that contain the reverse adaptor sequence are mapped

One-sided mini-PCR: It is possible to use target DNA that has an adaptor at the fragment ends (see **Figure 7**). STA may be performed with a multiplex set of Forward primers and one (or few) tag-specific Reverse primer. 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation adaptors added for universal amplification. 103 denotes the single stranded DNA with Forward Primer A hybridized. 104 denotes the PCR product from 103 that was amplified using Forward primer A and ligation adaptor tag Reverse primer LT, and which is the final PCR product. This method can detect shorter target sequences than standard PCR. However it may be relatively unspecific, as only one target specific primer is used. This protocol is effectively half of the one sided nested mini PCR

Reverse semi-nested mini-PCR: It is possible to use target DNA that has an adaptor at the fragment ends (see **Figure 8**). STA may be performed with a multiplex set of Forward primers and one (or few) tag-specific Reverse primer. 101 denotes double stranded DNA with a polymorphic locus of interest at X. 102 denotes the double stranded DNA with ligation

adaptors added for universal amplification. 103 denotes the single stranded DNA with Reverse Primer B hybridized. 104 denotes the PCR product from 103 that was amplified using Reverse primer B and ligation adaptor tag Forward primer LT. 105 denotes the PCR product 104 with hybridized Forward Primer A, and inner Reverse primer 'b'. 106 denotes the PCR product that has been amplified from 105 using Forward primer A and Reverse primer 'b', and which is the final PCR product. This method can detect shorter target sequences than standard PCR.

There also may be more variants that are simply iterations or combinations of the above methods such as doubly nested PCR, where three sets of primers are used. Another variant is one-and-a-half sided nested mini-PCR, where STA may also be performed with a multiplex set of nested Forward primers and one (or few) tag-specific Reverse primer.

Note that in all of these variants, the identity of the Forward primer and the Reverse primer may be interchanged. Note that in some embodiments, the nested variant can equally well be run without the initial library preparation that comprises appending the adapter tags, and a universal amplification step. Note that in some embodiments, additional rounds of PCR may be included, with additional Forward and/or Reverse primers and amplification steps; these additional steps may be particularly useful if it is desirable to further increase the percent of DNA molecules that correspond to the targeted loci.

Nesting Workflows

There are many ways to perform the amplification, with different degrees of nesting, and with different degrees of multiplexing. In **Figure 9**, a flow chart is given with some of the possible workflows. Note that the use of 10,000-plex PCR is only meant to be an example; these flow charts would work equally well for other degrees of multiplexing.

Looped ligation adaptors

When adding universal tagged adaptors for example for the purpose of making a library for sequencing, there are a number of ways to ligate adaptors. One way is to blunt end the sample DNA, perform A-tailing, and ligate with adaptors that have a T-overhang. There are a number of other ways to ligate adaptors. There are also a number of adaptors that can be ligated. For example, a Y-adaptor can be used where the adaptor consists of two strands of DNA where one strand has a double strand region, and a region specified by a forward primer region, and where the other strand specified by a double strand region that is complementary to

the double strand region on the first strand, and a region with a reverse primer. The double stranded region, when annealed, may contain a T-overhang for the purpose of ligating to double stranded DNA with an A overhang.

In an embodiment, the adaptor can be a loop of DNA where the terminal regions are complementary, and where the loop region contains a forward primer tagged region (LFT), a reverse primer tagged region (LRT), and a cleavage site between the two (See **Figure 10**). 101 refers to the double stranded, blunt ended target DNA. 102 refers to the A-tailed target DNA. 103 refers to the looped ligation adaptor with T overhang 'T' and the cleavage site 'Z'. 104 refers to the target DNA with appended looped ligation adaptors. 105 refers to the target DNA with the ligation adaptors appended cleaved at the cleavage site. LFT refers to the ligation adaptor Forward tag, and the LRT refers to the ligation adaptor Reverse tag. The complementary region may end on a T overhang, or other feature that may be used for ligation to the target DNA. The cleavage site may be a series of uracils for cleavage by UNG, or a sequence that may be recognized and cleaved by a restriction enzyme or other method of cleavage or just a basic amplification. These adaptors can be uses for any library preparation, for example, for sequencing. These adaptors can be used in combination with any of the other methods described herein, for example the mini-PCR amplification methods.

Internally Tagged Primers

When using sequencing to determine the allele present at a given polymorphic locus, the sequence read typically begins upstream of the primer binding site (a), and then to the polymorphic site (X). Tags are typically configured as shown in **Figure 11**, left. 101 refers to the single stranded target DNA with polymorphic locus of interest 'X', and primer 'a' with appended tag 'b'. In order to avoid nonspecific hybridization, the primer binding site (region of target DNA complementary to 'a') is typically 18 to 30 bp in length. Sequence tag 'b' is typically about 20 bp; in theory these can be any length longer than about 15 bp, though many people use the primer sequences that are sold by the sequencing platform company. The distance 'd' between 'a' and 'X' may be at least 2 bp so as to avoid allele bias. When performing multiplexed PCR amplification using the methods disclosed herein or other methods, where careful primer design is necessary to avoid excessive primer primer interaction, the window of allowable distance 'd' between 'a' and 'X' may vary quite a bit: from 2 bp to 10 bp, from 2 bp to 20 bp, from 2 bp to 30 bp, or even from 2 bp to more than 30

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 74

bp. Therefore, when using the primer configuration shown in **Figure 11**, left, sequence reads must be a minimum of 40 bp to obtain reads long enough to measure the polymorphic locus, and depending on the lengths of 'a' and 'd' the sequence reads may need to be up to 60 or 75 bp. Usually, the longer the sequence reads, the higher the cost and time of sequencing a given number of reads, therefore, minimizing the necessary read length can save both time and money. In addition, since, on average, bases read earlier on the read are read more accurately than those read later on the read, decreasing the necessary sequence read length can also increase the accuracy of the measurements of the polymorphic region.

In an embodiment, termed internally tagged primers, the primer binding site (a) is split in to a plurality of segments (a', a", a'"....), and the sequence tag (b) is on a segment of DNA that is in the middle of two of the primer binding sites, as shown in **Figure 11**, 103. This configuration allows the sequencer to make shorter sequence reads. In an embodiment, a' + a" should be at least about 18 bp, and can be as long as 30, 40, 50, 60, 80, 100 or more than 100 bp. In an embodiment, a" should be at least about 6 bp, and in an embodiment is between about 8 and 16 bp. All other factors being equal, using the internally tagged primers can cut the length of the sequence reads needed by at least 6 bp, as much as 8 bp, 10 bp, 12 bp, 15 bp, and even by as many as 20 or 30 bp. This can result in a significant money, time and accuracy advantage. An example of internally tagged primers is given in **Figure 12**.

Primers with ligation adaptor binding region

One issue with fragmented DNA is that since it is short in length, the chance that a polymorphism is close to the end of a DNA strand is higher than for a long strand (*e.g.* 101, Figure 10). Since PCR capture of a polymorphism requires a primer binding site of suitable length on both sides of the polymorphism, a significant number of strands of DNA with the targeted polymorphism will be missed due to insufficient overlap between the primer and the targeted binding site. In an embodiment, the target DNA 101 can have ligation adaptors appended 102, and the target primer 103 can have a region (cr) that is complementary to the ligation adaptor tag (lt) appended upstream of the designed binding region (a) (see **Figure 13**); thus in cases where the binding region (region of 101 that is complementary to a) is shorter than the 18 bp typically required for hybridization, the region (cr) on the primer than is complementary to the library tag is able to increase the binding energy to a point where the PCR can proceed. Note that any specificity that is lost due to a shorter binding region can be

made up for by other PCR primers with suitably long target binding regions. Note that this embodiment can be used in combination with direct PCR, or any of the other methods described herein, such as nested PCR, semi nested PCR, hemi nested PCR, one sided nested or semi or hemi nested PCR, or other PCR protocols.

When using the sequencing data to determine ploidy in combination with an analytical method that involves comparing the observed allele data to the expected allele distributions for various hypotheses, each additional read from alleles with a low depth of read will yield more information than a read from an allele with a high depth of read. Therefore, ideally, one would wish to see uniform depth of read (DOR) where each locus will have a similar number of representative sequence reads. Therefore, it is desirable to minimize the DOR variance. In an embodiment, it is possible to decrease the coefficient of variance of the DOR (this may be defined as the standard deviation of the DOR / the average DOR) by increasing the annealing times. In some embodiments the annealing temperatures may be longer than 2 minutes, longer than 4 minutes, longer than ten minutes, longer than 30 minutes, and longer than one hour, or even longer. Since annealing is an equilibrium process, there is no limit to the improvement of DOR variance with increasing annealing times. In an embodiment, increasing the primer concentration may decrease the DOR variance.

Exemplary Whole Genome Amplification Methods

In some embodiments, a method of the present disclosure may involve amplifying DNA, such as the use of whole genome application to amplify a nucleic acid sample before amplifying just the target loci. Amplification of the DNA, a process which transforms a small amount of genetic material to a larger amount of genetic material that comprises a similar set of genetic data, can be done by a wide variety of methods, including, but not limited to polymerase chain reaction (PCR). One method of amplifying DNA is whole genome amplification (WGA). There are a number of methods available for WGA: ligation-mediated PCR (LM-PCR), degenerate oligonucleotide primer PCR (DOP-PCR), and multiple displacement amplification (MDA). In LM-PCR, short DNA sequences called adapters are ligated to blunt ends of DNA. These adapters contain universal amplification sequences, which are used to amplify the DNA by PCR. In DOP-PCR, random primers that also contain universal amplification sequences are used in a first round of annealing and PCR. Then, a second round of PCR is used to amplify the sequences further with the universal primer

sequences. MDA uses the phi-29 polymerase, which is a highly processive and non-specific enzyme that replicates DNA and has been used for single-cell analysis. The major limitations to amplification of material from a single cell are (1) necessity of using extremely dilute DNA concentrations or extremely small volume of reaction mixture, and (2) difficulty of reliably dissociating DNA from proteins across the whole genome. Regardless, single-cell whole genome amplification has been used successfully for a variety of applications for a number of years. There are other methods of amplifying DNA from a sample of DNA. The DNA amplification transforms the initial sample of DNA into a sample of DNA that is similar in the set of sequences, but of much greater quantity. In some cases, amplification may not be required.

In some embodiments, DNA may be amplified using a universal amplification, such as WGA or MDA. In some embodiments, DNA may be amplified by targeted amplification, for example using targeted PCR, or circularizing probes. In some embodiments, the DNA may be preferentially enriched using a targeted amplification method, or a method that results in the full or partial separation of desired from undesired DNA, such as capture by hybridization approaches. In some embodiments, DNA may be amplified by using a combination of a universal amplification method and a preferential enrichment method. A fuller description of some of these methods can be found elsewhere in this document.

Exemplary Enrichment and Sequencing Methods

In an embodiment, a method disclosed herein uses selective enrichment techniques that preserve the relative allele frequencies that are present in the original sample of DNA at each target loci (*e.g.*, each polymorphic locus) from a set of target loci (*e.g.*, polymorphic loci). While enrichment is particularly advantageous for methods for analyzing polymorphic loci, these enrichment methods can be readily adapted for nonpolymorphic loci if desired. In some embodiments the amplification and/or selective enrichment technique may involve PCR such as ligation mediated PCR, fragment capture by hybridization, Molecular Inversion Probes, or other circularizing probes. In some embodiments, methods for amplification or selective enrichment may involve using probes where, upon correct hybridization to the target sequence, the 3-prime end or 5-prime end of a nucleotide probe is separated from the polymorphic site of the allele by a small number of nucleotides. This separation reduces preferential amplification of one allele, termed allele bias. This is an improvement over methods that involve using

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 77

probes where the 3-prime end or 5-prime end of a correctly hybridized probe are directly adjacent to or very near to the polymorphic site of an allele. In an embodiment, probes in which the hybridizing region may or certainly contains a polymorphic site are excluded. Polymorphic sites at the site of hybridization can cause unequal hybridization or inhibit hybridization altogether in some alleles, resulting in preferential amplification of certain alleles. These embodiments are improvements over other methods that involve targeted amplification and/or selective enrichment in that they better preserve the original allele frequencies of the sample at each polymorphic locus, whether the sample is pure genomic sample from a single individual or mixture of individuals.

The use of a technique to enrich a sample of DNA at a set of target loci followed by sequencing as part of a method for non-invasive prenatal allele calling or ploidy calling may confer a number of unexpected advantages. In some embodiments of the present disclosure, the method involves measuring genetic data for use with an informatics based method, such as PARENTAL SUPPORT™ (PS). The ultimate outcome of some of the embodiments is the actionable genetic data of an embryo or a fetus. There are many methods that may be used to measure the genetic data of the individual and/or the related individuals as part of embodied methods. In an embodiment, a method for enriching the concentration of a set of targeted alleles is disclosed herein, the method comprising one or more of the following steps: targeted amplification of genetic material, addition of loci specific oligonucleotide probes, ligation of specified DNA strands, isolation of sets of desired DNA, removal of unwanted components of a reaction, detection of certain sequences of DNA by hybridization, and detection of the sequence of one or a plurality of strands of DNA by DNA sequencing methods. In some cases the DNA strands may refer to target genetic material, in some cases they may refer to primers, in some cases they may refer to synthesized sequences, or combinations thereof. These steps may be carried out in a number of different orders.

For example, a universal amplification step of the DNA prior to targeted amplification may confer several advantages, such as removing the risk of bottlenecking and reducing allelic bias. The DNA may be mixed an oligonucleotide probe that can hybridize with two neighboring regions of the target sequence, one on either side. After hybridization, the ends of the probe may be connected by adding a polymerase, a means for ligation, and any necessary reagents to allow the circularization of the probe. After circularization, an exonuclease may be added to digest to non-circularized genetic material, followed by detection of the circularized

probe. The DNA may be mixed with PCR primers that can hybridize with two neighboring regions of the target sequence, one on either side. After hybridization, the ends of the probe may be connected by adding a polymerase, a means for ligation, and any necessary reagents to complete PCR amplification. Amplified or unamplified DNA may be targeted by hybrid capture probes that target a set of loci; after hybridization, the probe may be localized and separated from the mixture to provide a mixture of DNA that is enriched in target sequences.

The use of a method to target certain loci followed by sequencing as part of a method for allele calling or ploidy calling may confer a number of unexpected advantages. Some methods by which DNA may be targeted, or preferentially enriched, include using circularizing probes, linked inverted probes (LIPs, MIPs), capture by hybridization methods such as SURESELECT, and targeted PCR or ligation-mediated PCR amplification strategies.

In some embodiments, a method of the present disclosure involves measuring genetic data for use with an informatics based method, such as PARENTAL SUPPORT™ (PS), which is described further herein. PARENTAL SUPPORT™ is an informatics based approach to manipulating genetic data, aspects of which are described herein. The ultimate outcome of some of the embodiments is the actionable genetic data of an embryo or a fetus followed by a clinical decision based on the actionable data. The algorithms behind the PS method take the measured genetic data of the target individual, often an embryo or fetus, and the measured genetic data from related individuals, and are able to increase the accuracy with which the genetic state of the target individual is known. In an embodiment, the measured genetic data is used in the context of making ploidy determinations during prenatal genetic diagnosis. In an embodiment, the measured genetic data is used in the context of making ploidy determinations or allele calls on embryos during in vitro fertilization. There are many methods that may be used to measure the genetic data of the individual and/or the related individuals in the aforementioned contexts. The different methods comprise a number of steps, those steps often involving amplification of genetic material, addition of oligonucleotide probes, ligation of specified DNA strands, isolation of sets of desired DNA, removal of unwanted components of a reaction, detection of certain sequences of DNA by hybridization, detection of the sequence of one or a plurality of strands of DNA by DNA sequencing methods. In some cases the DNA strands may refer to target genetic material, in some cases they may refer to primers, in some cases they may refer to synthesized sequences, or combinations thereof. These steps may be carried out in a number of different orders.

Note that in theory it is possible to target any number loci in the genome, anywhere from one loci to well over one million loci. If a sample of DNA is subjected to targeting, and then sequenced, the percentage of the alleles that are read by the sequencer will be enriched with respect to their natural abundance in the sample. The degree of enrichment can be anywhere from one percent (or even less) to ten-fold, a hundred-fold, a thousand-fold or even many million-fold. In the human genome there are roughly 3 billion base pairs, and nucleotides, comprising approximately 75 million polymorphic loci. The more loci that are targeted, the smaller the degree of enrichment is possible. The fewer the number of loci that are targeted, the greater degree of enrichment is possible, and the greater depth of read may be achieved at those loci for a given number of sequence reads.

In an embodiment of the present disclosure, the targeting or preferential may focus entirely on SNPs. In an embodiment, the targeting or preferential may focus on any polymorphic site. A number of commercial targeting products are available to enrich exons. Surprisingly, targeting exclusively SNPs, or exclusively polymorphic loci, is particularly advantageous when using a method for NPD that relies on allele distributions. There are also published methods for NPD using sequencing, for example U.S. Patent 7,888,017, involving a read count analysis where the read counting focuses on counting the number of reads that map to a given chromosome, where the analyzed sequence reads do not focused on regions of the genome that are polymorphic. Those types of methodology that do not focus on polymorphic alleles would not benefit as much from targeting or preferential enrichment of a set of alleles.

In an embodiment of the present disclosure, it is possible to use a targeting method that focuses on SNPs to enrich a genetic sample in polymorphic regions of the genome. In an embodiment, it is possible to focus on a small number of SNPs, for example between 1 and 100 SNPs, or a larger number, for example, between 100 and 1,000, between 1,000 and 10,000, between 10,000 and 100,000 or more than 100,000 SNPs. In an embodiment, it is possible to focus on one or a small number of chromosomes that are correlated with live trisomic births, for example chromosomes 13, 18, 21, X and Y, or some combination thereof. In an embodiment, it is possible to enrich the targeted SNPs by a small factor, for example between 1.01 fold and 100 fold, or by a larger factor, for example between 100 fold and 1,000,000 fold, or even by more than 1,000,000 fold. In an embodiment of the present disclosure, it is possible to use a targeting method to create a sample of DNA that is preferentially enriched in polymorphic regions of the genome. In an embodiment, it is possible

to use this method to create a mixture of DNA with any of these characteristics where the mixture of DNA contains maternal DNA and also free floating fetal DNA. In an embodiment, it is possible to use this method to create a mixture of DNA that has any combination of these factors. For example, the method described herein may be used to produce a mixture of DNA that comprises maternal DNA and fetal DNA, and that is preferentially enriched in DNA that corresponds to 200 SNPs, all of which are located on either chromosome 18 or 21, and which are enriched an average of 1000 fold. In another example, it is possible to use the method to create a mixture of DNA that is preferentially enriched in 10,000 SNPs that are all or mostly located on chromosomes 13, 18, 21, X and Y, and the average enrichment per loci is greater than 500 fold. Any of the targeting methods described herein can be used to create mixtures of DNA that are preferentially enriched in certain loci.

In some embodiments, a method of the present disclosure further includes measuring the DNA in the mixed fraction using a high throughput DNA sequencer, where the DNA in the mixed fraction contains a disproportionate number of sequences from one or more chromosomes, wherein the one or more chromosomes are taken from the group comprising chromosome 13, chromosome 18, chromosome 21, chromosome X, chromosome Y and combinations thereof.

Described herein are three methods: multiplex PCR, targeted capture by hybridization, and linked inverted probes (LIPs), which may be used to obtain and analyze measurements from a sufficient number of polymorphic loci from a maternal plasma sample in order to detect fetal aneuploidy; this is not meant to exclude other methods of selective enrichment of targeted loci. Other methods may equally well be used without changing the essence of the method. In each case the polymorphism assayed may include single nucleotide polymorphisms (SNPs), small indels, or STRs. A preferred method involves the use of SNPs. Each approach produces allele frequency data; allele frequency data for each targeted locus and/or the joint allele frequency distributions from these loci may be analyzed to determine the ploidy of the fetus. Each approach has its own considerations due to the limited source material and the fact that maternal plasma consists of mixture of maternal and fetal DNA. This method may be combined with other approaches to provide a more accurate determination. In an embodiment, this method may be combined with a sequence counting approach such as that described in US Patent 7,888,017. The approaches described could also be used to detect fetal paternity noninvasively from maternal plasma samples. In addition each approach may be applied to

other mixtures of DNA or pure DNA samples to detect the presence or absence of aneuploid chromosomes, to genotype a large number of SNP from degraded DNA samples, to detect segmental copy number variations (CNVs), to detect other genotypic states of interest, or some combination thereof.

Accurately Measuring the Allelic Distributions in a Sample

Current sequencing approaches can be used to estimate the distribution of alleles in a sample. One such method involves randomly sampling sequences from a pool DNA, termed shotgun sequencing. The proportion of a particular allele in the sequencing data is typically very low and can be determined by simple statistics. The human genome contains approximately 3 billion base pairs. So, if the sequencing method used make 100 bp reads, a particular allele will be measured about once in every 30 million sequence reads.

In an embodiment, a method of the present disclosure is used to determine the presence or absence of two or more different haplotypes that contain the same set of loci in a sample of DNA from the measured allele distributions of loci from that chromosome. The different haplotypes could represent two different homologous chromosomes from one individual, three different homologous chromosomes from a trisomic individual, three different homologous haplotypes from a mother and a fetus where one of the haplotypes is shared between the mother and the fetus, three or four haplotypes from a mother and fetus where one or two of the haplotypes are shared between the mother and the fetus, or other combinations. Alleles that are polymorphic between the haplotypes tend to be more informative, however any alleles where the mother and father are not both homozygous for the same allele will yield useful information through measured allele distributions beyond the information that is available from simple read count analysis.

Shotgun sequencing of such a sample, however, is extremely inefficient as it results in many sequences for regions that are not polymorphic between the different haplotypes in the sample, or are for chromosomes that are not of interest, and therefore reveal no information about the proportion of the target haplotypes. Described herein are methods that specifically target and/or preferentially enrich segments of DNA in the sample that are more likely to be polymorphic in the genome to increase the yield of allelic information obtained by sequencing. Note that for the measured allele distributions in an enriched sample to be truly representative of the actual amounts present in the target individual, it is critical that there is little or no

preferential enrichment of one allele as compared to the other allele at a given loci in the targeted segments. Current methods known in the art to target polymorphic alleles are designed to ensure that at least some of any alleles present are detected. However, these methods were not designed for the purpose of measuring the unbiased allelic distributions of polymorphic alleles present in the original mixture. It is non-obvious that any particular method of target enrichment would be able to produce an enriched sample wherein the measured allele distributions would accurately represent the allele distributions present in the original unamplified sample better than any other method. While many enrichment methods may be expected, in theory, to accomplish such an aim, an ordinary person skilled in the art is well aware that there is a great deal of stochastic or deterministic bias in current amplification, targeting and other preferential enrichment methods. One embodiment of a method described herein allows a plurality of alleles found in a mixture of DNA that correspond to a given locus in the genome to be amplified, or preferentially enriched in a way that the degree of enrichment of each of the alleles is nearly the same. Another way to say this is that the method allows the relative quantity of the alleles present in the mixture as a whole to be increased, while the ratio between the alleles that correspond to each locus remains essentially the same as they were in the original mixture of DNA. For some reported methods, preferential enrichment of loci can result in allelic biases of more than 1%, more than 2%, more than 5% and even more than 10%. This preferential enrichment may be due to capture bias when using a capture by hybridization approach, or amplification bias which may be small for each cycle, but can become large when compounded over 20, 30 or 40 cycles. For the purposes of this disclosure, for the ratio to remain essentially the same means that the ratio of the alleles in the original mixture divided by the ratio of the alleles in the resulting mixture is between 0.95 and 1.05, between 0.98 and 1.02, between 0.99 and 1.01, between 0.995 and 1.005, between 0.998 and 1.002, between 0.999 and 1.001, or between 0.9999 and 1.0001. Note that the calculation of the allele ratios presented here may not be used in the determination of the ploidy state of the target individual, and may only a metric to be used to measure allelic bias.

In an embodiment, once a mixture has been preferentially enriched at the set of target loci, it may be sequenced using any one of the previous, current, or next generation of sequencing instruments that sequences a clonal sample (a sample generated from a single molecule; examples include ILLUMINA GAIIx, ILLUMINA HiSeq, LIFE TECHNOLOGIES SOLiD, 5500XL). The ratios can be evaluated by sequencing through the specific alleles

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 83

within the targeted region. These sequencing reads can be analyzed and counted according the allele type and the rations of different alleles determined accordingly. For variations that are one to a few bases in length, detection of the alleles will be performed by sequencing and it is essential that the sequencing read span the allele in question in order to evaluate the allelic composition of that captured molecule. The total number of captured molecules assayed for the genotype can be increased by increasing the length of the sequencing read. Full sequencing of all molecules would guarantee collection of the maximum amount of data available in the enriched pool. However, sequencing is currently expensive, and a method that can measure allele distributions using a lower number of sequence reads will have great value. In addition, there are technical limitations to the maximum possible length of read as well as accuracy limitations as read lengths increase. The alleles of greatest utility will be of one to a few bases in length, but theoretically any allele shorter than the length of the sequencing read can be used. While allele variations come in all types, the examples provided herein focus on SNPs or variants contained of just a few neighboring base pairs. Larger variants such as segmental copy number variants can be detected by aggregations of these smaller variations in many cases as whole collections of SNP internal to the segment are duplicated. Variants larger than a few bases, such as STRs require special consideration and some targeting approaches work while others will not.

There are multiple targeting approaches that can be used to specifically isolate and enrich a one or a plurality of variant positions in the genome. Typically, these rely on taking advantage of the invariant sequence flanking the variant sequence. There are reports by others related to targeting in the context of sequencing where the substrate is maternal plasma (see, *e.g.*, Liao et al., Clin. Chem. 2011; 57(1): pp. 92-101). However, these approaches use targeting probes that target exons, and do not focus on targeting polymorphic regions of the genome. In an embodiment, a method of the present disclosure involves using targeting probes that focus exclusively or almost exclusively on polymorphic regions. In an embodiment, a method of the present disclosure involves using targeting probes that focus exclusively or almost exclusively on SNPs. In some embodiments of the present disclosure, the targeted polymorphic sites consist of at least 10% SNPs, at least 20% SNPs, at least 30% SNPs, at least 40% SNPs, at least 50% SNPs, at least 60% SNPs, at least 70% SNPs, at least 80% SNPs, at least 90% SNPs, at least 95% SNPs, at least 98% SNPs, at least 99% SNPs, at least 99.9% SNPs, or exclusively SNPs.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 84

In an embodiment, a method of the present disclosure can be used to determine genotypes (base composition of the DNA at specific loci) and relative proportions of those genotypes from a mixture of DNA molecules, where those DNA molecules may have originated from one or a number of genetically distinct individuals. In an embodiment, a method of the present disclosure can be used to determine the genotypes at a set of polymorphic loci, and the relative ratios of the amount of different alleles present at those loci. In an embodiment the polymorphic loci may consist entirely of SNPs. In an embodiment, the polymorphic loci can comprise SNPs, single tandem repeats, and other polymorphisms. In an embodiment, a method of the present disclosure can be used to determine the relative distributions of alleles at a set of polymorphic loci in a mixture of DNA, where the mixture of DNA comprises DNA that originates from a mother, and DNA that originates from a fetus. In an embodiment, the joint allele distributions can be determined on a mixture of DNA isolated from blood from a pregnant woman. In an embodiment, the allele distributions at a set of loci can be used to determine the ploidy state of one or more chromosomes on a gestating fetus.

In an embodiment, the mixture of DNA molecules could be derived from DNA extracted from multiple cells of one individual. In an embodiment, the original collection of cells from which the DNA is derived may comprise a mixture of diploid or haploid cells of the same or of different genotypes, if that individual is mosaic (germline or somatic). In an embodiment, the mixture of DNA molecules could also be derived from DNA extracted from single cells. In an embodiment, the mixture of DNA molecules could also be derived from DNA extracted from mixture of two or more cells of the same individual, or of different individuals. In an embodiment, the mixture of DNA molecules could be derived from DNA isolated from biological material that has already liberated from cells such as blood plasma, which is known to contain cell free DNA. In an embodiment, the this biological material may be a mixture of DNA from one or more individuals, as is the case during pregnancy where it has been shown that fetal DNA is present in the mixture. In an embodiment, the biological material could be from a mixture of cells that were found in maternal blood, where some of the cells are fetal in origin. In an embodiment, the biological material could be cells from the blood of a pregnant which have been enriched in fetal cells.

Circularizing Probes

Some embodiments of the present disclosure involve the use of "Linked Inverted

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 85

Probes" (LIPs), which have been previously described in the literature, to amplify the target loci before or after amplification using primers that are not LIPs in the multiplex PCR methods of the invention. LIPs is a generic term meant to encompass technologies that involve the creation of a circular molecule of DNA, where the probes are designed to hybridize to targeted region of DNA on either side of a targeted allele, such that addition of appropriate polymerases and/or ligases, and the appropriate conditions, buffers and other reagents, will complete the complementary, inverted region of DNA across the targeted allele to create a circular loop of DNA that captures the information found in the targeted allele. LIPs may also be called pre-circularized probes, pre-circularizing probes, or circularizing probes. The LIPs probe may be a linear DNA molecule between 50 and 500 nucleotides in length, and in an embodiment between 70 and 100 nucleotides in length; in some embodiments, it may be longer or shorter than described herein. Others embodiments of the present disclosure involve different incarnations, of the LIPs technology, such as Padlock Probes and Molecular Inversion Probes (MIPs).

One method to target specific locations for sequencing is to synthesize probes in which the 3' and 5' ends of the probes anneal to target DNA at locations adjacent to and on either side of the targeted region, in an inverted manner, such that the addition of DNA polymerase and DNA ligase results in extension from the 3' end, adding bases to single stranded probe that are complementary to the target molecule (gap-fill), followed by ligation of the new 3' end to the 5' end of the original probe resulting in a circular DNA molecule that can be subsequently isolated from background DNA. The probe ends are designed to flank the targeted region of interest. One aspect of this approach is commonly called MIPS and has been used in conjunction with array technologies to determine the nature of the sequence filled in. One drawback to the use of MIPs in the context of measuring allele ratios is that the hybridization, circularization and amplification steps do not happed at equal rates for different alleles at the same loci. This results in measured allele ratios that are not representative of the actual allele ratios present in the original mixture.

In an embodiment, the circularizing probes are constructed such that the region of the probe that is designed to hybridize upstream of the targeted polymorphic locus and the region of the probe that is designed to hybridize downstream of the targeted polymorphic locus are covalently connected through a non-nucleic acid backbone. This backbone can be any biocompatible molecule or combination of biocompatible molecules. Some examples of

possible biocompatible molecules are poly(ethylene glycol), polycarbonates, polyurethanes, polyethylenes, polypropylenes, sulfone polymers, silicone, cellulose, fluoropolymers, acrylic compounds, styrene block copolymers, and other block copolymers.

In an embodiment of the present disclosure, this approach has been modified to be easily amenable to sequencing as a means of interrogating the filled in sequence. In order to retain the original allelic proportions of the original sample at least one key consideration must be taken into account. The variable positions among different alleles in the gap-fill region must not be too close to the probe binding sites as there can be initiation bias by the DNA polymerase resulting in differential of the variants. Another consideration is that additional variations may be present in the probe binding sites that are correlated to the variants in the gap-fill region which can result unequal amplification from different alleles. In an embodiment of the present disclosure, the 3' ends and 5' ends of the pre-circularized probe are designed to hybridize to bases that are one or a few positions away from the variant positions (polymorphic sites) of the targeted allele. The number of bases between the polymorphic site (SNP or otherwise) and the base to which the 3' end and/or 5' of the pre-circularized probe is designed to hybridize may be one base, it may be two bases, it may be three bases, it may be four bases, it may be five bases, it may be six bases, it may be seven to ten bases, it may be eleven to fifteen bases, or it may be sixteen to twenty bases, twenty to thirty bases, or thirty to sixty bases. The forward and reverse primers may be designed to hybridize a different number of bases away from the polymorphic site. Circularizing probes can be generated in large numbers with current DNA synthesis technology allowing very large numbers of probes to be generated and potentially pooled, enabling interrogation of many loci simultaneously. It has been reported to work with more than 300,000 probes. Two papers that discuss a method involving circularizing probes that can be used to measure the genomic data of the target individual include: Porreca et al., Nature Methods, 2007 4(11), pp. 931-936.; and also Turner et al., Nature Methods, 2009, 6(5), pp. 315-316. The methods described in these papers may be used in combination with other methods described herein. Certain steps of the method from these two papers may be used in combination with other steps from other methods described herein.

In some embodiments of the methods disclosed herein, the genetic material of the target individual is optionally amplified, followed by hybridization of the pre-circularized probes, performing a gap fill to fill in the bases between the two ends of the hybridized probes, ligating the two ends to form a circularized probe, and amplifying the circularized probe,

using, for example, rolling circle amplification. Once the desired target allelic genetic information is captured by circularizing appropriately designed oligonucleotide probes, such as in the LIPs system, the genetic sequence of the circularized probes may be being measured to give the desired sequence data. In an embodiment, the appropriately designed oligonucleotides probes may be circularized directly on unamplified genetic material of the target individual, and amplified afterwards. Note that a number of amplification procedures may be used to amplify the original genetic material, or the circularized LIPs, including rolling circle amplification, MDA, or other amplification protocols. Different methods may be used to measure the genetic information on the target genome, for example using high throughput sequencing, Sanger sequencing, other sequencing methods, capture-by-hybridization, capture-by-circularization, multiplex PCR, other hybridization methods, and combinations thereof.

Once the genetic material of the individual has been measured using one or a combination of the above methods, an informatics based method, such as the PARENTAL SUPPORT™ method, along with the appropriate genetic measurements, can then be used to determination the ploidy state of one or more chromosomes on the individual, and/or the genetic state of one or a set of alleles, specifically those alleles that are correlated with a disease or genetic state of interest. Note that the use of LIPs has been reported for multiplexed capture of genetic sequences, followed by genotyping with sequencing. However, the use of sequencing data resulting from a LIPs-based strategy for the amplification of the genetic material found in a single cell, a small number of cells, or extracellular DNA, has not been used for the purpose of determining the ploidy state of a target individual.

Applying an informatics based method to determine the ploidy state of an individual from genetic data as measured by hybridization arrays, such as the ILLUMINA INFINIUM array, or the AFFYMETRIX gene chip has been described in documents references elsewhere in this document. However, the method described herein shows improvements over methods described previously in the literature. For example, the LIPs based approach followed by high throughput sequencing unexpectedly provides better genotypic data due to the approach having better capacity for multiplexing, better capture specificity, better uniformity, and low allelic bias. Greater multiplexing allows more alleles to be targeted, giving more accurate results. Better uniformity results in more of the targeted alleles being measured, giving more accurate results. Lower rates of allelic bias result in lower rates of miscalls, giving more accurate

results. More accurate results result in an improvement in clinical outcomes, and better medical care.

It is important to note that LIPs may be used as a method for targeting specific loci in a sample of DNA for genotyping by methods other than sequencing. For example, LIPs may be used to target DNA for genotyping using SNP arrays or other DNA or RNA based microarrays.

Ligation-mediated PCR

Ligation-mediated PCR may be used to amplify the target loci before or after PCR amplification using primers that are not ligated. Ligation-mediated PCR is a method of PCR used to preferentially enrich a sample of DNA by amplifying one or a plurality of loci in a mixture of DNA, the method comprising: obtaining a set of primer pairs, where each primer in the pair contains a target specific sequence and a non-target sequence, where the target specific sequence is preferably designed to anneal to a target region, one upstream and one downstream from the polymorphic site, and which can be separated from the polymorphic site by 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11-20, 21-30, 31-40, 41-50, 51-100, or more than 100; polymerization of the DNA from the 3-prime end of upstream primer to the fill the single strand region between it and the 5-prime end of the downstream primer with nucleotides complementary to the target molecule; ligation of the last polymerized base of the upstream primer to the adjacent 5-prime base of the downstream primer; and amplification of only polymerized and ligated molecules using the non-target sequences contained at the 5-prime end of the upstream primer and the 3-prime end of the downstream primer. Pairs of primers to distinct targets may be mixed in the same reaction. The non-target sequences serve as universal sequences such that of all pairs of primers that have been successfully polymerized and ligated may be amplified with a single pair of amplification primers.

Capture by Hybridization

In some embodiments, a method of the present disclosure may involve using any of the following capture by hybridization methods in addition to using multiplex PCR to amplify the target loci. Preferential enrichment of a specific set of sequences in a target genome can be accomplished in a number of ways. Elsewhere in this document is a description of how LIPs can be used to target a specific set of sequences, but in all of those applications, other targeting

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 89

and/or preferential enrichment methods can be used equally well for the same ends. One example of another targeting method is the capture by hybridization approach. Some examples of commercial capture by hybridization technologies include AGILENT's SURE SELECT and ILLUMINA's TRUSEQ. In capture by hybridization, a set of oligonucleotides that is complimentary or mostly complimentary to the desired targeted sequences is allowed to hybridize to a mixture of DNA, and then physically separated from the mixture. Once the desired sequences have hybridized to the targeting oligonucleotides, the effect of physically removing the targeting oligonucleotides is to also remove the targeted sequences. Once the hybridized oligos are removed, they can be heated to above their melting temperature and they can be amplified. Some ways to physically remove the targeting oligonucleotides is by covalently bonding the targeting oligos to a solid support, for example a magnetic bead, or a chip. Another way to physically remove the targeting oligonucleotides is by covalently bonding them to a molecular moiety with a strong affinity for another molecular moiety. An example of such a molecular pair is biotin and streptavidin, such as is used in SURE SELECT. Thus that targeted sequences could be covalently attached to a biotin molecule, and after hybridization, a solid support with streptavidin affixed can be used to pull down the biotinylated oligonucleotides, to which are hybridized to the targeted sequences.

Hybrid capture involves hybridizing probes that are complementary to the targets of interest to the target molecules. Hybrid capture probes were originally developed to target and enrich large fractions of the genome with relative uniformity between targets. In that application, it was important that all targets be amplified with enough uniformity that all regions could be detected by sequencing, however, no regard was paid to retaining the proportion of alleles in original sample. Following capture, the alleles present in the sample can be determined by direct sequencing of the captured molecules. These sequencing reads can be analyzed and counted according the allele type. However, using the current technology, the measured allele distributions the captured sequences are typically not representative of the original allele distributions.

In an embodiment, detection of the alleles is performed by sequencing. In order to capture the allele identity at the polymorphic site, it is essential that the sequencing read span the allele in question in order to evaluate the allelic composition of that captured molecule. Since the capture molecules are often of variable lengths upon sequencing cannot be guaranteed to overlap the variant positions unless the entire molecule is sequenced. However,

cost considerations as well as technical limitations as to the maximum possible length and accuracy of sequencing reads make sequencing the entire molecule unfeasible. In an embodiment, the read length can be increased from about 30 to about 50 or about 70 bases can greatly increase the number of reads that overlap the variant positions within the targeted sequences.

Another way to increase the number of reads that interrogate the position of interest is to decrease the length of the probe, as long as it does not result in bias in the underlying enriched alleles. The length of the synthesized probe should be long enough such that two probes designed to hybridize to two different alleles found at one locus will hybridize with near equal affinity to the various alleles in the original sample. Currently, methods known in the art describe probes that are typically longer than 120 bases. In a current embodiment, if the allele is one or a few bases then the capture probes may be less than about 110 bases, less than about 100 bases, less than about 90 bases, less than about 80 bases, less than about 70 bases, less than about 60 bases, less than about 50 bases, less than about 40 bases, less than about 30 bases, and less than about 25 bases, and this is sufficient to ensure equal enrichment from all alleles. When the mixture of DNA that is to be enriched using the hybrid capture technology is a mixture comprising free floating DNA isolated from blood, for example maternal blood, the average length of DNA is quite short, typically less than 200 bases. The use of shorter probes results in a greater chance that the hybrid capture probes will capture desired DNA fragments. Larger variations may require longer probes. In an embodiment, the variations of interest are one (a SNP) to a few bases in length. In an embodiment, targeted regions in the genome can be preferentially enriched using hybrid capture probes wherein the hybrid capture probes are of a length below 90 bases, and can be less than 80 bases, less than 70 bases, less than 60 bases, less than 50 bases, less than 40 bases, less than 30 bases, or less than 25 bases. In an embodiment, to increase the chance that the desired allele is sequenced, the length of the probe that is designed to hybridize to the regions flanking the polymorphic allele location can be decreased from above 90 bases, to about 80 bases, or to about 70 bases, or to about 60 bases, or to about 50 bases, or to about 40 bases, or to about 30 bases, or to about 25 bases.

There is a minimum overlap between the synthesized probe and the target molecule in order to enable capture. This synthesized probe can be made as short as possible while still being larger than this minimum required overlap. The effect of using a shorter probe length to target a polymorphic region is that there will be more molecules that overlap the target allele

region. The state of fragmentation of the original DNA molecules also affects the number of reads that will overlap the targeted alleles. Some DNA samples such as plasma samples are already fragmented due to biological processes that take place in vivo. However, samples with longer fragments by benefit from fragmentation prior to sequencing library preparation and enrichment. When both probes and fragments are short (~60-80 bp) maximum specificity may be achieved relatively few sequence reads failing to overlap the critical region of interest.

In an embodiment, the hybridization conditions can be adjusted to maximize uniformity in the capture of different alleles present in the original sample. In an embodiment, hybridization temperatures are decreased to minimize differences in hybridization bias between alleles. Methods known in the art avoid using lower temperatures for hybridization because lowering the temperature has the effect of increasing hybridization of probes to unintended targets. However, when the goal is to preserve allele ratios with maximum fidelity, the approach of using lower hybridization temperatures provides optimally accurate allele ratios, despite the fact that the current art teaches away from this approach. Hybridization temperature can also be increased to require greater overlap between the target and the synthesized probe so that only targets with substantial overlap of the targeted region are captured. In some embodiments of the present disclosure, the hybridization temperature is lowered from the normal hybridization temperature to about 40°C, to about 45°C, to about 50°C, to about 55°C, to about 60°C, to about 65, or to about 70°C.

In an embodiment, the hybrid capture probes can be designed such that the region of the capture probe with DNA that is complementary to the DNA found in regions flanking the polymorphic allele is not immediately adjacent to the polymorphic site. Instead, the capture probe can be designed such that the region of the capture probe that is designed to hybridize to the DNA flanking the polymorphic site of the target is separated from the portion of the capture probe that will be in van der Waals contact with the polymorphic site by a small distance that is equivalent in length to one or a small number of bases. In an embodiment, the hybrid capture probe is designed to hybridize to a region that is flanking the polymorphic allele but does not cross it; this may be termed a flanking capture probe. The length of the flanking capture probe may be less than about 120 bases, less than about 110 bases, less than about 100 bases, less than about 90 bases, and can be less than about 80 bases, less than about 70 bases, less than about 60 bases, less than about 50 bases, less than about 40 bases, less than about 30 bases, or less than about 25 bases. The region of the genome that is targeted by the flanking

capture probe may be separated by the polymorphic locus by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11-20, or more than 20 base pairs.

Description of a targeted capture based disease screening test using targeted sequence capture. Custom targeted sequence capture, like those currently offered by AGILENT (SURE SELECT), ROCHE-NIMBLEGEN, or ILLUMINA. Capture probes could be custom designed to ensure capture of various types of mutations. For point mutations, one or more probes that overlap the point mutation should be sufficient to capture and sequence the mutation.

For small insertions or deletions, one or more probes that overlap the mutation may be sufficient to capture and sequence fragments comprising the mutation. Hybridization may be less efficient between the probe-limiting capture efficiency, typically designed to the reference genome sequence. To ensure capture of fragments comprising the mutation one could design two probes, one matching the normal allele and one matching the mutant allele. A longer probe may enhance hybridization. Multiple overlapping probes may enhance capture. Finally, placing a probe immediately adjacent to, but not overlapping, the mutation may permit relatively similar capture efficiency of the normal and mutant alleles.

For Simple Tandem Repeats (STRs), a probe overlapping these highly variable sites is unlikely to capture the fragment well. To enhance capture a probe could be placed adjacent to, but not overlapping the variable site. The fragment could then be sequenced as normal to reveal the length and composition of the STR.

For large deletions, a series of overlapping probes, a common approach currently used in exon capture systems may work. However, with this approach it may be difficult to determine whether or not an individual is heterozygous. Targeting and evaluating SNPs within the captured region could potentially reveal loss of heterozygosity across the region indicating that an individual is a carrier. In an embodiment, it is possible to place non-overlapping or singleton probes across the potentially deleted region and use the number of fragments captured as a measure of heterozygosity. In the case where an individual caries a large deletion, one-half the number of fragments are expected to be available for capture relative to a non-deleted (diploid) reference locus. Consequently, the number of reads obtained from the deleted regions should be roughly half that obtained from a normal diploid locus. Aggregating and averaging the sequencing read depth from multiple singleton probes across the potentially deleted region may enhance the signal and improve confidence of the diagnosis. The two approaches, targeting SNPs to identify loss of heterozygosity and using multiple singleton

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 93

probes to obtain a quantitative measure of the quantity of underlying fragments from that locus can also be combined. Either or both of these strategies may be combined with other strategies to better obtain the same end.

If during testing cfDNA detection of a male fetus, as indicated by the presence of the Y-chromosome fragments, captured and sequenced in the same test, and either an X-linked dominant mutation where mother and father are unaffected, or a dominant mutation where mother is not affected would indicated heighted risk to the fetus. Detection of two mutant recessive alleles within the same gene in an unaffected mother would imply the fetus had inherited a mutant allele from father and potentially a second mutant allele from mother. In all cases, follow-up testing by amniocentesis or chorionic villus sampling may be indicated.

A targeted capture based disease screening test could be combined with a targeted capture based non-invasive prenatal diagnostic test for aneuploidy.

There are a number of ways to decrease depth of read (DOR) variability: for example, one could increase primer concentrations, one could use longer targeted amplification probes, or one could run more STA cycles (such as more than 25, more than 30, more than 35, or even more than 40)

Exemplary Methods of Determining the Number of DNA Molecules in a Sample.

A method is described herein to determine the number of DNA molecules in a sample by generating a uniquely identified molecule for each original DNA molecules in the sample during the first round of DNA amplification. Described here is a procedure to accomplish the above end followed by a single molecule or clonal sequencing method.

The approach entails targeting one or more specific loci and generating a tagged copy of the original molecules such manner that most or all of the tagged molecules from each targeted locus will have a unique tag and can be distinguished from one another upon sequencing of this barcode using clonal or single molecule sequencing. Each unique sequenced barcode represents a unique molecule in the original sample. Simultaneously, sequencing data is used to ascertain the locus from which the molecule originates. Using this information one can determine the number of unique molecules in the original sample for each locus.

This method can be used for any application in which quantitative evaluation of the number of molecules in an original sample is required. Furthermore, the number of unique molecules of one or more targets can be related to the number of unique molecules to one or

more other targets to determine the relative copy number, allele distribution, or allele ratio. Alternatively, the number of copies detected from various targets can be modeled by a distribution in order to identify the mostly likely number of copies of the original targets. Applications include but are not limited to detection of insertions and deletions such as those found in carriers of Duchenne Muscular Dystrophy; quantitation of deletions or duplications segments of chromosomes such as those observed in copy number variants; chromosome copy number of samples from born individuals; chromosome copy number of samples from unborn individuals such as embryos or fetuses.

The method can be combined with simultaneous evaluation of variations contained in the targeted by sequence. This can be used to determine the number of molecules representing each allele in the original sample. This copy number method can be combined with the evaluation of SNPs or other sequence variations to determine the chromosome copy number of born and unborn individuals; the discrimination and quantification of copies from loci which have short sequence variations, but in which PCR may amplifies from multiple target regions such as in carrier detection of Spinal Muscle Atrophy; determination of copy number of different sources of molecules from samples consisting of mixtures of different individual such as in detection of fetal aneuploidy from free floating DNA obtained from maternal plasma.

In an embodiment, the method as it pertains to a single target locus may comprise one or more of the following steps: (1) Designing a standard pair of oligomers for PCR amplification of a specific locus. (2) Adding, during synthesis, a sequence of specified bases with no or minimal complimentarity to the target locus or genome to the 5' end of the one of the target specific oligomer. This sequence, termed the tail, is a known sequence, to be used for subsequent amplification, followed by a sequence of random nucleotides. These random nucleotides comprise the random region. The random region comprises a randomly generated sequence of nucleic acids that probabilistically differ between each probe molecule. Consequently, following synthesis, the tailed oligomer pool will consists of a collection of oligomers beginning with a known sequence followed by unknown sequence that differs between molecules, followed by the target specific sequence. (3) Performing one round of amplification (denaturation, annealing, extension) using only the tailed oligomer. (4) adding exonuclease to the reaction, effectively stopping the PCR reaction, and incubating the reaction at the appropriate temperature to remove forward single stranded oligos that did not anneal to temple and extend to form a double stranded product. (5) Incubating the reaction at a high

temperature to denature the exonuclease and eliminate its activity. (6) Adding to the reaction a new oligonucleotide that is complementary to tail of the oligomer used in the first reaction along with the other target specific oligomer to enable PCR amplification of the product generated in the first round of PCR. (7) Continuing amplification to generate enough product for downstream clonal sequencing. (8) Measuring the amplified PCR product by a multitude of methods, for example, clonal sequencing, to a sufficient number of bases to span the sequence.

In an embodiment, a method of the present disclosure involves targeting multiple loci in parallel or otherwise. Primers to different target loci can be generated independently and mixed to create multiplex PCR pools. In an embodiment, original samples can be divided into sub-pools and different loci can be targeted in each sub-pool before being recombined and sequenced. In an embodiment, the tagging step and a number of amplification cycles may be performed before the pool is subdivided to ensure efficient targeting of all targets before splitting, and improving subsequent amplification by continuing amplification using smaller sets of primers in subdivided pools.

One example of an application where this technology would be particularly useful is non-invasive prenatal aneuploidy diagnosis where the ratio of alleles at a given locus or a distribution of alleles at a number of loci can be used to help determine the number of copies of a chromosome present in a fetus. In this context, it is desirable to amplify the DNA present in the initial sample while maintaining the relative amounts of the various alleles. In some circumstances, especially in cases where there is a very small amount of DNA, for example, fewer than 5,000 copies of the genome, fewer than 1,000 copies of the genome, fewer than 500 copies of the genome, and fewer than 100 copies of the genome, one can encounter a phenomenon called bottlenecking. This is where there are a small number of copies of any given allele in the initial sample, and amplification biases can result in the amplified pool of DNA having significantly different ratios of those alleles than are in the initial mixture of DNA. By applying a unique or nearly unique set of barcodes to each strand of DNA before standard PCR amplification, it is possible to exclude n-1 copies of DNA from a set of n identical molecules of sequenced DNA that originated from the same original molecule.

For example, imagine a heterozygous SNP in the genome of an individual, and a mixture of DNA from the individual where ten molecules of each allele are present in the original sample of DNA. After amplification there may be 100,000 molecules of DNA corresponding to that locus. Due to stochastic processes, the ratio of DNA could be anywhere

from 1:2 to 2:1, however, since each of the original molecules was tagged with a unique tag, it would be possible to determine that the DNA in the amplified pool originated from exactly 10 molecules of DNA from each allele. This method would therefore give a more accurate measure of the relative amounts of each allele than a method not using this approach. For methods where it is desirable for the relative amount of allele bias to be minimized, this method will provide more accurate data.

Association of the sequenced fragment to the target locus can be achieved in a number of ways. In an embodiment, a sequence of sufficient length is obtained from the targeted fragment to span the molecule barcode as well a sufficient number of unique bases corresponding to the target sequence to allow unambiguous identification of the target locus. In another embodiment, the molecular bar-coding primer that contains the randomly generated molecular barcode can also contain a locus specific barcode (locus barcode) that identifies the target to which it is to be associated. This locus barcode would be identical among all molecular bar-coding primers for each individual target and hence all resulting amplicons, but different from all other targets. In an embodiment, the tagging method described herein may be combined with a one-sided nesting protocol.

In an embodiment, the design and generation of molecular barcoding primers may be reduced to practice as follows: the molecular barcoding primers may consist of a sequence that is not complementary to the target sequence followed by random molecular barcode region followed by a target specific sequence. The sequence 5' of molecular barcode may be used for subsequence PCR amplification and may comprise sequences useful in the conversion of the amplicon to a library for sequencing. The random molecular barcode sequence could be generated in a multitude of ways. The preferred method synthesize the molecule tagging primer in such a way as to include all four bases to the reaction during synthesis of the barcode region. All or various combinations of bases may be specified using the IUPAC DNA ambiguity codes. In this manner the synthesized collection of molecules will contain a random mixture of sequences in the molecular barcode region. The length of the barcode region will determine how many primers will contain unique barcodes. The number of unique sequences is related to the length of the barcode region as N^L where N is the number of bases, typically 4, and L is the length of the barcode. A barcode of five bases can yield up to 1024 unique sequences; a barcode of eight bases can yield 65536 unique barcodes. In an embodiment, the DNA can be measured by a sequencing method, where the sequence data represents the

sequence of a single molecule. This can include methods in which single molecules are sequenced directly or methods in which single molecules are amplified to form clones detectable by the sequence instrument, but that still represent single molecules, herein called clonal sequencing.

Exemplary Methods and Reagents for Quantification of Amplification Products

Quantitation of specific nucleic acid sequences of interest is typically done by quantitative real-time PCR techniques such as TAQMAN (LIFE TECHNOLOGIES), INVADER probes (THIRD WAVE TECHNOLOGIES), and the like. Such techniques suffer from numerous shortcomings such as limited ability to achieve the simultaneous analysis of multiple sequences in parallel (multiplexation) and the ability to provide accurate quantitative data for only a narrow range of possible amplification cycles (*e.g.*, when the logarithm of PCR amplification production quantity versus the number of cycles is in the linear range). DNA sequencing techniques, particularly high throughput next-generation sequencing techniques (often referred to as massively parallel sequencing techniques) such as those employed in MYSEQ (ILLUMINA), HISEQ (ILLUMINA), ION TORRENT (LIFE TECHNOLOGIES), GENOME ANALYZER ILX (ILLUMINA), GS FLEX+ (ROCHE 454) *etc.*, can be used for by quantitative measurements of the number of copies of sequence of interest present in sample, thereby providing quantitative information about the starting materials, *e.g.*, copy number or transcription levels. High throughput genetic sequencers are amenable to the use of bar coding (*i.e.*, sample tagging with distinctive nucleic acid sequences) so as to identify specific samples from individuals thereby permitting the simultaneous analysis of multiple samples in a single run of the DNA sequencer. The number of times a given region of the genome in a library preparation (or other nucleic preparation of interest) is sequenced (number of reads) will be proportional to the number of copies of that sequence in the genome of interest (or expression level in the case of cDNA containing preparations). However, the preparation and sequencing of genetic libraries (and similar genome derived preparations) can introduce numerous biases that interfere with obtaining an accurate quantitative reading for the nucleic acid sequence of interest. For example, different nucleic acid sequences can amplify with different efficiencies during nucleic amplification steps that take place during the genetic library preparation or sample preparation.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 98

The problem with differential amplification efficiencies can be mitigated by using certain embodiments of the subject invention. The subject invention includes various methods and compositions that relate to the use of standards for inclusion in amplification processes that can be used to improve the accuracy of quantitation. The invention is of use in, among other areas, the detection of aneuploidy in a fetus by analyzing free floating fetal DNA in maternal blood, as described herein and as described, among other places, U.S. Patent No. 8,008,018; U.S. Patent No. 7,332,277; PCT Published Application WO 2012/078792A2; and PCT Published Application WO 2011/146632 A1, which are each herein incorporated by reference in its entirety Embodiments of the invention are also of use in the detection of aneuploidy in an *in vitro* generated embryos. Commercially significant aneuploidies that may be detected include aneuploidy of the human chromosomes 13, 18, 21, X and Y.

Embodiments of the invention may be used with either human or non-human nucleic acids, and may be applied to both animal and plant derived nucleic acids. Embodiments of the invention may also be used to detect and/or quantitate alleles for other genetic disorders characterized by deletions or insertions. The deletion containing alleles can be detected in suspected carriers of the allele of interest.

One embodiment of the subject invention includes standards that are present in a known quantity (relative or absolute). For example, consider a genetic library made from a genetic source that is diploid for chromosome 8 (containing locus A) and triploid for chromosome 21 (containing locus B). A genetic library can be produced from this sample that will contain sequences in quantities that are a function of the number of chromosomes present in the sample, *e.g.*, 200 copies of locus A and 300 copies of locus B. However, if locus A amplifies much more efficiently than locus B, after PCR there may be 60,000 copies of the A amplicon and 30,000 copies of the B amplicon, thus obscuring the true chromosomal copy number of the initial genomic sample when analysis by high throughput DNA sequencing (or other quantitative nucleic acid detection techniques). To mitigate this problem a standard sequence for locus A is employed, wherein the standard sequence amplifies with essentially the same efficiency as locus A. Similarly, a standard sequence for locus B is created, wherein the standard sequence amplifies with the essentially the same efficiency as locus B. A standard sequence of locus A and a standard sequence for locus B are added to the mixture prior to PCR (or other amplification techniques). These standard sequences are present in known quantities, either relative quantities or absolute quantities. Thus if a 1:1 mixture of

standard sequence A and standard sequence B were added (prior to amplification) to the mixture in the previous example, 3000 copies of the standard A amplicon would be produced and 1000 copies of the standard B amplicon would be produced, showing that locus A is amplified 3 times more efficiently than locus B, under the same set of conditions.

In various embodiments one or more selected regions of a genome containing a SNP (or other polymorphism) of interest can be specifically amplified and subsequently sequenced. This target specific amplification can take place during the formation of a genetic library for sequencing. The library can contain numerous targeted regions for amplification. In some embodiments at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 regions of interest. Examples of such libraries are described herein and can be found in U.S. Patent Application No. 2012/0270212, filed November 18, 2011, which is herein incorporated by reference in its entirety.

Many high throughput DNA sequencing techniques require the modification of the genetic starting material, *e.g.*, the litigation of universal priming sites and/or barcodes, so as to form libraries to facilitate the clonal amplification of small nucleic acid fragments prior to performing subsequent sequencing reactions. In some embodiments, one or more standard sequences are added during genetic library formation or added to a precursor component of a genetic library prior to amplification of the library. The standard sequences can be selected so as to mimic (yet be distinguishable based on nucleotide base sequence) target genomic fragments to be prepared for sequencing by a high throughput genetic sequencing technique. In one embodiment, the standard sequence can be identical to the target genomic fragment excepting one, two, three, four to ten, or eleven to twenty nucleotides. In some embodiments, when the target genetic sequence contains a SNP, the standard sequence can be identical to the SNP excepting the nucleotide at the polymorphic base, which may be chosen to be one of the four nucleotides that is not observed at that location in nature. The standard sequences can be used in a highly multiplexed analysis of multiple target loci (such as polymorphic loci). Standard sequences can be added during the process of library formation (prior to amplification) in known quantities (relative or absolute) so as to provide a standard metric for greater accuracy in determining the amount of target sequence of interest in the sample of analysis. The combination of knowledge of the known quantities of the standard sequences used in conjunction with the knowledge of the ploidy level formation of library for sequencing formed from a genome of previously characterized ploidy level, *e.g.*, known to be diploid for

all autosomal chromosomes, can be used to calibrate the amplification properties of each standard sequence with respect to its corresponding target sequence and account for variations between batches of mixtures comprising multiple standard sequences. Given that it is often necessary to simultaneously analyze a large number of loci, it is useful to produce a mixture comprising a large set standard sequences. Embodiments of the invention include mixtures comprising multiple standard sequences. Ideally the amount of each standard sequence in the mixture is known with high precision. However, it is extremely difficult to achieve this ideal because as a practical matter there is a significant amount of variation in the quantity of each standard sequence in the mixture, particularly for mixtures comprising a large number of different synthetic oligonucleotides. This variation has numerous sources, *e.g.*, variations in *in vitro* oligonucleotide synthesis reaction efficiencies between batch, inaccuracies in volume measurement, variations in pipetting, Furthermore, this variation can occur between different batches of that theoretically contain the exact same set of standard sequences in the exact same amounts. Accordingly, it is of interest to calibrate each batch of standard sequences independently. Batches of standard sequences can be calibrated against reference genomes of known chromosomal composition. Batched of standard sequences can be calibrated by sequencing the batch of standard sequences with minimal or no amplifications steps included in the sequencing protocol. Embodiments of the invention include calibrated mixtures of different standard sequences. Other embodiments of the invention include methods of calibrating mixtures of different standard sequences and calibrated mixtures of different standard sequences made by the subject methods.

Various embodiments of the subject mixtures of standard sequences and methods for using them can comprise at least 10; 100, 500; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 or more standards sequences, as well as various intermediate amounts. The number of the standard sequences can be the same as the number of target sequences selected for analysis during the generation of a targeted library for DNA sequencing. However, in some embodiments, it may be advantageous to use a lower number of standard sequences than the number of targeted regions in the library being constructed. It may be advantageous to use the lower number so as avoid coming up against the limits of the sequencing capacity of the high throughput DNA sequencer being employed. The number of standard sequences can be 50% or less than the number of targeted regions, 40% or less than the number of targeted regions, be 30% or less than the number of targeted

regions, 20% or less than the number of targeted regions, be 10% or less than the number of targeted regions, 5% or less than the number of targeted regions, 1% or less than the number of targeted regions, as well as various intermediate values. For example, if a genetic library is created using 15,000 pairs of primers targeted to specific SNP containing loci, a suitable a mixture containing 1500 standard sequences corresponding to 1500 of the 15,000 targeted loci can be added prior to the amplification step of library constructions.

The amount of standard sequences added during library construction can vary considerably among different embodiments. In some embodiments, the amount of each standard sequence can be approximately the same as the predicted amount of the target sequence present in the genomic material sample used for library preparation. In other embodiments, the amount of each standard sequence can be greater or less than the predicted amount of the target sequence present in the genomic material sample used for library preparation. While the initial relative amounts of the target sequence and the standard sequence are not critical for the function of the invention, it is preferable that the amount be within the range 100 times greater to 100 times less than the amount of the target sequence present in the genomic material sample used for library preparation. Excessive amounts of standard may use too much sequencing capacity of the DNA sequencer in a given run of the instrument. Using too low an amount of standard sequences will produce insufficient data to aid in the analysis of variation in amplification efficiency.

The standard sequences may be selected to be very similar in nucleotide base sequence to the amplified regions of interest; preferably the standard sequence has the exact same primer-binding sites as the analyzed genomic region, *i.e.*, the "target sequence." The standard sequence must be distinguishable from the corresponding target sequence at a given locus. For the sake of convenience, this distinguishable region of the standard sequence will be referred to as a "marker sequence." In some embodiments, the marker sequence region of the target sequences contains the polymorphic region, *e.g.*, a SNP, and can be flanked on both sides by primer binding regions. The standard sequence may be selected to closely match the GC content of the corresponding target sequence. In some embodiments, the primer binding regions of the standard sequence are flanked by universal priming sites. These universal priming sites are selected to match universal priming sites used in a genomic library for analysis. In other embodiments, the standard sequences do not have universal priming sites and the universal priming sites are added during the creation of a library. Standard sequences

are typically provided in single stranded form. A standard sequence is defined with respect to a corresponding target sequence and the sequence specific reagents used to amplify the target sequence. In some embodiments, the target sequence contains the polymorphism of interest, *e.g.*, a SNP, a deletion, or insertion, present in the nucleic acid sample for analysis. The standard sequence is a synthetic polynucleotide that is similar in nucleotide base sequence to the target sequence, but is nonetheless distinguishable from the target sequence by virtue of at least one nucleotide base difference, thereby providing a mechanism for distinguishing amplicon sequences derived from the standard sequence form amplicon sequences derived from the target sequence. Standard sequences are selected so as to have essentially the same amplification properties as the corresponding target sequence when amplified with the same set of amplification reagents, *e.g.*, PCR primers. In some embodiments, the standard sequences can have the same primer sequence binding sites than the corresponding target sequences. In other embodiments, the standard sequences can have a different primer sequence binding sites than the corresponding target sequences. In some embodiments, the standard sequences can be selected to produce amplicons that have the same length as the length of amplicons produced from the corresponding target sequences. In other embodiments, the standard sequences can be selected to produce amplicons that have the slightly different lengths than the length of amplicons produced from the corresponding target sequences.

After the amplification reactions have been completed, the library is sequenced on a high throughput DNA sequencer where individual molecule are clonally amplified and sequenced. The number of sequence reads for each allele of the target sequence is counted, also counted are the number of sequence reads for the standard sequence corresponding to the target sequence. The process is also carried out for at least one other pair of target sequences and corresponding standard sequences. Consider for example, locus A, X_{A1}reads for allele 1 of locus A are produced; X_{A2} reads for allele 2 of locus A are produced, and X_{AC} reads for standard sequence A are produced. The ratio of (X_{A1} plus X_{A2}) to X_{AC} is determined for each locus of interest. As discussed earlier, the process can be performed on a reference genome, *e.g.*, a genome that is known to be diploid for all chromosomes. The process can be repeated many times in order to provide a large number of read values so as to determine a mean number of reads and the standard deviation in the number of reads. The process is performed with a mixture comprising a large number of different standard sequences corresponding to different loci. By assuming that (1) X_{A1} plus X_{A2} corresponds to the known number of

chromosome, *e.g.*, 2 for the normal human female genome and (2) the standard sequences have similar amplification (and detectability) properties as their corresponding natural loci, the relative amounts of the different standard sequences in the multiplex standard mixture can be determined. The calibrated multiplex standard sequence mixture can then be used to adjust for the variability in amplification efficiency between the different loci in a multiplex amplification reaction.

Other embodiments of the invention include methods and compositions for measuring the copy number of specific genes of interest, including duplications and mutant genes characterized by large deletions that would interfere with quantitation by sequencing. Sequencing would have problems detecting alleles having such deletions. Standard sequences included the amplification process can be used to reduce this problem.

In one embodiment of the invention the target sequence for analysis is a gene having a wild type (*i.e.* functional) form and a mutant form characterized by a deletion. Exemplary of such genes is SMN1, an allele having deletion being responsible for the genetic disease spinal muscular atrophy (SMA). It is of interest to detect an individual carrying the mutant form of the gene by means of high throughput genetic sequencing techniques. The application of such techniques to the detection of deletion mutations can be problematic because, among other reasons, the lack of sequences observed in sequencing (as opposed to detecting a simple point mutation or SNP). Such embodiments employ (1) a pair of amplification primers specific for the gene of interest, where in the amplification primers will amplify the gene of interest (or a portion thereof) and will not significantly amplify the mutant allele, (2) a standard sequence corresponding to the wild type allele of the gene of interest(*i.e.*, a target sequence), but differing by at least one detectable nucleotide base, (3) a pair of amplification primers specific for a second target sequence that serves as a reference sequence, and (4) a standard sequence corresponding to the reference sequence.

In one embodiment of the invention is provided a method for measuring the number of copies of the gene of interest, where in the gene of interest has one meaning allele that comprises a deletion. The method can employ amplification reagent specific for the gene of interest, *e.g.*, PCR primers, that are specific for the gene of interest by amplifying at least a portion of the gene of interest, or the entire gene of interest, or a region adjacent to the gene of interest, while not amplifying the deletion comprising allele of the gene of interest. Additionally the subject method employs a standard sequence corresponding to the gene of

interest, wherein the standard sequence differs by at least one nucleotide base from the gene of interest (so that the sequence of the standard sequence can be readily distinguished from the naturally occurring gene of interest). Typically, the standard sequence will contain the same primer binding sites as the gene of interest so as to minimize any amplification discrimination between the gene of interest and the standard sequence corresponding to the gene of interest. The reaction will also comprises amplification reagents specific for a reference sequence. The reference sequence is a sequence of known (or at least assumed to be known) copy number in the genome to be analyzed. The reaction further comprises a standard sequence corresponding to the reference sequence. Typically, the standard sequence corresponding to the reference sequence will contain the same primer binding sites as the reference sequence so as to minimize any amplification discrimination between the reference sequence and the standard sequence corresponding to the reference sequence.

Exemplary Nucleic Acid Samples

In some embodiments, the genetic sample may be prepared and/or purified. There are a number of standard procedures known in the art to accomplish such an end. In some embodiments, the sample may be centrifuged to separate various layers. In some embodiments, the DNA may be isolated using filtration. In some embodiments, the preparation of the DNA may involve amplification, separation, purification by chromatography, liquid liquid separation, isolation, preferential enrichment, preferential amplification, targeted amplification, or any of a number of other techniques either known in the art or described herein.

In some embodiments, a method disclosed herein could be used in situations where there is a very small amount of DNA present, such as in *in vitro* fertilization, or in forensic situations, where one or a few cells are available (typically less than ten cells, less than twenty cells or less than 40 cells.) In these embodiments, a method disclosed herein serves to make ploidy calls from a small amount of DNA that is not contaminated by other DNA, but where the ploidy calling very difficult the small amount of DNA. In some embodiments, a method disclosed herein could be used in situations where the target DNA is contaminated with DNA of another individual, for example in maternal blood in the context of prenatal diagnosis, paternity testing, or products of conception testing. Some other situations where these methods would be particularly advantageous would be in the case of cancer testing where only one or a small number of cells were present among a larger amount of normal cells. The genetic

measurements used as part of these methods could be made on any sample comprising DNA or RNA, for example but not limited to: blood, plasma, body fluids, urine, hair, tears, saliva, tissue, skin, fingernails, blastomeres, embryos, amniotic fluid, chorionic villus samples, feces, bile, lymph, cervical mucus, semen, or other cells or materials comprising nucleic acids. In an embodiment, a method disclosed herein could be run with nucleic acid detection methods such as sequencing, microarrays, qPCR, digital PCR, or other methods used to measure nucleic acids. If for some reason it were found to be desirable, the ratios of the allele count probabilities at a locus could be calculated, and the allele ratios could be used to determine ploidy state in combination with some of the methods described herein, provided the methods are compatible. In some embodiments, a method disclosed herein involves calculating, on a computer, allele ratios at the plurality of polymorphic loci from the DNA measurements made on the processed samples. In some embodiments, a method disclosed herein involves calculating, on a computer, allele ratios at the plurality of polymorphic loci from the DNA measurements made on the processed samples along with any combination of other improvements described in this disclosure.

In some embodiments, this method may be used to genotype a single cell, a small number of cells, two to five cells, six to ten cells, ten to twenty cells, twenty to fifty cell, fifty to one hundred cells, one hundred to one thousand cells, or a small amount of extracellular DNA, for example from one to ten picograms, from ten to one hundred pictograms, from one hundred pictograms to one nanogram, from one to ten nanograms, from ten to one hundred nanograms, or from one hundred nanograms to one microgram.

Exemplary RNA Expression Studies

The multiplex PCR methods of the invention can be used to increase the number of target loci that can be evaluated during gene expression profiling experiments. For example, the expression levels of thousands of genes can be simultaneously monitored to determine whether a person has a sequence (such as a polymorphism or other mutation) associated with a disease (such as cancer) or an increased risk of a disease. These methods can be used to identify sequences (such as polymorphisms or other mutations) associated with an increased or decreased risk for a disease such as cancer by comparing gene expression (such as the expression of particular mRNA alleles) in samples from patients with and without the disease. Additionally, the effect of particular treatments, diseases, or developmental stages on gene

expression can be determined. Similarly, these methods can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected and uninfected cells or tissues. In these methods the number of sequencing reads can be adjusted based on the frequency of the polymorphisms that are being analyzed such that sufficient reads are performed for the polymorphisms to be detected if they are present.

In some embodiments, a sample containing RNA (such as mRNA) is amplified using a reverse transcriptase (RT) and the resulting DNA (such as cDNA) is then amplified using a DNA polymerase (PCR). The RT and PCR steps may be carried out sequentially in the same reaction volume or separately. Any of the primer libraries of the invention can be used in this reverse transcription polymerase chain reaction (RT-PCR) method. In various embodiments, the reverse transcription is performed using oligo-dT, random primers, a mixture of oligo-dT and random primers, or primers specific to the target loci. To avoid amplification of contaminating genomic DNA, primers for RT-PCR can be designed so that part of one primer hybridizes to the 3' end of one exon and the other part of the primer hybridizes to the 5' end of the adjacent exon. Such primers anneal to cDNA synthesized from spliced mRNAs, but not to genomic DNA. To detect amplification of contaminating DNA, RT-PCR primer pairs may be designed to flank a region that contains at least one intron. Products amplified from cDNA (no introns) are smaller than those amplified from genomic DNA (containing introns). Size difference in products is used to detect the presence of contaminating DNA. In some embodiments when only the mRNA sequence is known, primer annealing sites are chosen that are at least 300–400 base pairs apart since it is likely that fragments of this size from eukaryotic DNA contain splice junctions. Alternatively, the sample can be treated with DNase to degrade contaminating DNA.

Exemplary Methods for Paternity Testing

The multiplex PCR methods of the invention can be used to improve the accuracy of paternity testing since so many target loci can be analyzed at once (see, *e.g*, U.S. Publication No. 2012/0122701, filed December 22, 2011, is which is hereby incorporated by reference in its entirety). For example, the multiplex PCR method can allow thousands of polymorphic loci (such as SNPs) to be analyzed for use in the PARENTAL SUPPORT algorithm described herein to determine whether an alleged father in is the biological father of a fetus. In some

embodiments the method involves (i) simultaneously amplifying a plurality of polymorphic loci that includes at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci on genetic material from the alleged father to produce a first set of amplified products; (ii) simultaneously amplifying the corresponding plurality of polymorphic loci on a mixed sample of DNA originating from a blood sample from the pregnant mother to produce a second set of amplified products; wherein the mixed sample of DNA comprises fetal DNA and maternal DNA; (iii) determining on a computer the probability that the alleged father is the biological father of the fetus using genotypic measurements based on the first and second sets of amplified products; and (iv) establishing whether the alleged father is the biological father of the fetus using the determined probability that the alleged father is the biological father of the fetus. In various embodiments, the method further includes simultaneously amplifying the corresponding plurality of polymorphic loci on genetic material from the mother to produce a third set of amplified products; wherein the probability that the alleged father is the biological father of the fetus is determined using genotypic measurements based on the first, second, and third sets of amplified products.

Exemplary Methods for Embryo Characterization and Selection

The multiplex PCR methods of the invention can be used to improve the selection of embryos for *in vitro* fertilization by allowing thousands of target loci to be analyzed at once (*see, e.g*, U.S. Pub. No. 2011/0092763, filed May 27, 2008, filed December 22, 2011, is which is hereby incorporated by reference in its entirety). For example, the multiplex PCR method can allow thousands of polymorphic loci (such as SNPs) to be analyzed for use in the PARENTAL SUPPORT algorithm described herein to select an embryo out of a set of embryos for *in vitro* fertilization

In some embodiments, the invention provides methods of estimating relative likelihoods that each embryo from a set of embryos will develop as desired. In some embodiments, the method involves contacting a sample from each embryo with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture for each embryo, wherein the samples are each derived from one or more cells from an embryo. In some embodiments, each reaction mixture is subjected to primer extension

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 108

reaction conditions to produce amplified products. In some embodiments, the method includes determining on a computer one or more characteristics of at least one cell from each embryo based on the amplified products; and estimating on a computer the relative likelihoods that each embryo will develop as desired, based on the one or more characteristics of the at least one cell for each embryo. In some embodiments, the method includes using an informatics based method to determine the at least one characteristic, such as the PARENTAL SUPPORT algorithm described herein. In some embodiments, the characteristic includes a ploidy state. In some embodiments, the characteristic is selected from the group consisting of aneuploid, euploid, mosaic, nullsomy, monosomy, uniparental disomy, trisomy, tetrasomy, a type of aneuploidy, unmatched copy error trisomy, matched copy error trisomy, maternal origin of aneuploidy, paternal origin of aneuploidy, a presence or absence of a disease-linked gene, a chromosomal identity of any aneuploid chromosome, an abnormal genetic condition, a deletion or duplication, a likelihood of a characteristic, and combinations thereof. The characteristic may be associated with a chromosome taken from the group consisting of chromosome one, chromosome two, chromosome three, chromosome four, chromosome five, chromosome six, chromosome seven, chromosome eight, chromosome nine, chromosome ten, chromosome eleven, chromosome twelve, chromosome thirteen, chromosome fourteen, chromosome fifteen, chromosome sixteen, chromosome seventeen, chromosome eighteen, chromosome nineteen, chromosome twenty, chromosome twenty-one, chromosome twenty-two, X chromosome or Y chromosome, and combinations thereof.

Exemplary Prenatal Diagnostic Methods

The multiplex PCR methods of the present invention can be used to improve prenatal diagnostic methods, such as the determination of the ploidy status of fetal chromosomes. Given that the large number of target loci that can be simultaneously amplified, more accurate determinations can be made.

In an embodiment, the present disclosure provides *ex vivo* methods for determining the ploidy status of a chromosome in a gestating fetus from genotypic data measured from a mixed sample of DNA (*i.e.*, DNA from the mother of the fetus, and DNA from the fetus) and optionally from genotypic data measured from a sample of genetic material from the mother and possibly also from the father, wherein the determining is done by using a joint distribution model to create a set of expected allele distributions for different possible fetal ploidy states

given the parental genotypic data, and comparing the expected allelic distributions to the actual allelic distributions measured in the mixed sample, and choosing the ploidy state whose expected allelic distribution pattern most closely matches the observed allelic distribution pattern. In an embodiment, the mixed sample is derived from maternal blood, or maternal serum or plasma. In an embodiment, the mixed sample of DNA may be preferentially enriched at a target loci (*e.g.*, plurality of polymorphic loci). In an embodiment, the preferential enrichment is done in a way that minimizes the allelic bias. In an embodiment, the present disclosure relates to a composition of DNA that has been preferentially enriched at a plurality of loci such that the allelic bias is low. In an embodiment, the allelic distribution(s) are measured by sequencing the DNA from the mixed sample. In an embodiment, the joint distribution model assumes that the alleles will be distributed in a binomial fashion. In an embodiment, the set of expected joint allele distributions are created for genetically linked loci while considering the extant recombination frequencies from various sources, for example, using data from the International HapMap Consortium.

In an embodiment, the present disclosure provides methods for non-invasive prenatal diagnosis (NPD), specifically, determining the aneuploidy status of a fetus by observing allele measurements at a plurality of polymorphic loci in genotypic data measured on DNA mixtures, where certain allele measurements are indicative of an aneuploid fetus, while other allele measurements are indicative of a euploid fetus. In an embodiment, the genotypic data is measured by sequencing DNA mixtures that were derived from maternal plasma. In an embodiment, the DNA sample may be preferentially enriched in molecules of DNA that correspond to the plurality of loci whose allele distributions are being calculated. In an embodiment a sample of DNA comprising only or almost only genetic material from the mother and possibly also a sample of DNA comprising only or almost only genetic material from the father are measured. In an embodiment, the genetic measurements of one or both parents along with the estimated fetal fraction are used to create a plurality of expected allele distributions corresponding to different possible underlying genetic states of the fetus; the expected allele distributions may be termed hypotheses. In an embodiment, the maternal genetic data is not determined by measuring genetic material that is exclusively or almost exclusively maternal in nature, rather, it is estimated from the genetic measurements made on maternal plasma that comprises a mixture of maternal and fetal DNA. In some embodiments the hypotheses may comprise the ploidy of the fetus at one or more chromosomes, which

segments of which chromosomes in the fetus were inherited from which parents, and combinations thereof. In some embodiments, the ploidy state of the fetus is determined by comparing the observed allele measurements to the different hypotheses where at least some of the hypotheses correspond to different ploidy states, and selecting the ploidy state that corresponds to the hypothesis that is most likely to be true given the observed allele measurements. In an embodiment, this method involves using allele measurement data from some or all measured SNPs, regardless of whether the loci are homozygous or heterozygous, and therefore does not involve using alleles at loci that are only heterozygous. This method may not be appropriate for situations where the genetic data pertains to only one polymorphic locus. This method is particularly advantageous when the genetic data comprises data for more than ten polymorphic loci for a target chromosome or more than twenty polymorphic loci. This method is especially advantageous when the genetic data comprises data for more than 50 polymorphic loci for a target chromosome, more than 100 polymorphic loci or more than 200 polymorphic loci for a target chromosome. In some embodiments, the genetic data may comprise data for more than 500 polymorphic loci for a target chromosome, more than 1,000 polymorphic loci, more than 2,000 polymorphic loci, or more than 5,000 polymorphic loci for a target chromosome.

In an embodiment, a method disclosed herein yields a quantitative measure of the number of independent observations of each allele at a polymorphic locus. This is unlike most methods such as microarrays or qualitative PCR which provide information about the ratio of two alleles but do not quantify the number of independent observations of either allele. With methods that provide quantitative information regarding the number of independent observations, only the ratio is utilized in ploidy calculations, while the quantitative information by itself is not useful. To illustrate the importance of retaining information about the number of independent observations consider the sample locus with two alleles, A and B. In a first experiment twenty A alleles and twenty B alleles are observed, in a second experiment 200 A alleles and 200 B alleles are observed. In both experiments the ratio (A/(A+B)) is equal to 0.5, however the second experiment conveys more information than the first about the certainty of the frequency of the A or B allele. Some methods by others involve averaging or summing allele ratios (channel ratios) (*i.e.* x_i/y_i) from individual allele and analyzes this ratio, either comparing it to a reference chromosome or using a rule pertaining to how this ratio is expected to behave in particular situations. No allele weighting is implied in such methods, where it is

assumed that one can ensure about the same amount of PCR product for each allele and that all the alleles should behave the same way. Such a method has a number of disadvantages, and more importantly, precludes the use a number of improvements that are described elsewhere in this disclosure.

In an embodiment, a method disclosed herein explicitly models the allele frequency distributions expected in disomy as well as a plurality of allele frequency distributions that may be expected in cases of trisomy resulting from nondisjunction during meiosis I, nondisjunction during meiosis II, and/or nondisjunction during mitosis early in fetal development. To illustrate why this is important, imagine a case where there were no crossovers: nondisjunction during meiosis I would result a trisomy in which two different homologs were inherited from one parent; in contrast, nondisjunction during meiosis II or during mitosis early in fetal development would result in two copies of the same homolog from one parent. Each scenario would result in different expected allele frequencies at each polymorphic locus and also at all loci considered jointly, due to genetic linkage. Crossovers, which result in the exchange of genetic material between homologs, make the inheritance pattern more complex; in an embodiment, the instant method accommodates for this by using recombination rate information in addition to the physical distance between loci. In an embodiment, to enable improved distinction between meiosis I nondisjunction and meiosis II or mitotic nondisjunction the instant method incorporate into the model an increasing probability of crossover as the distance from the centromere increases. Meiosis II and mitotic nondisjunction can distinguished by the fact that mitotic nondisjunction typically results in identical or nearly identical copies of one homolog while the two homologs present following a meiosis II nondisjunction event often differ due to one or more crossovers during gametogenesis.

In some embodiments, a method disclosed herein involves comparing the observed allele measurements to theoretical hypotheses corresponding to possible fetal genetic aneuploidy, and does not involve a step of quantitating a ratio of alleles at a heterozygous locus. Where the number of loci is lower than about 20, the ploidy determination made using a method comprising quantitating a ratio of alleles at a heterozygous locus and a ploidy determination made using a method comprising comparing the observed allele measurements to theoretical allele distribution hypotheses corresponding to possible fetal genetic states may give a similar result. However, where the number of loci is above 50 these two methods is

likely to give significantly different results; where the number of loci is above 400, above, 1,000 or above 2,000 these two methods are very likely to give results that are increasingly significantly different. These differences are due to the fact that a method that comprises quantitating a ratio of alleles at a heterozygous locus without measuring the magnitude of each allele independently and aggregating or averaging the ratios precludes the use of techniques including using a joint distribution model, performing a linkage analysis, using a binomial distribution model, and/or other advanced statistical techniques, whereas using a method comprising comparing the observed allele measurements to theoretical allele distribution hypotheses corresponding to possible fetal genetic states may use these techniques which can substantially increase the accuracy of the determination.

In an embodiment, a method disclosed herein involves determining whether the distribution of observed allele measurements is indicative of a euploid or an aneuploid fetus using a joint distribution model. The use of a joint distribution model is a different from and a significant improvement over methods that determine heterozygosity rates by treating polymorphic loci independently in that the resultant determinations are of significantly higher accuracy. Without being bound by any particular theory, it is believed that one reason they are of higher accuracy is that the joint distribution model takes into account the linkage between SNPs, and likelihood of crossovers having occurred during the meiosis that gave rise to the gametes that formed the embryo that grew into the fetus. The purpose of using the concept of linkage when creating the expected distribution of allele measurements for one or more hypotheses is that it allows the creation of expected allele measurements distributions that correspond to reality considerably better than when linkage is not used. For example, imagine that there are two SNPs, 1 and 2 located nearby one another, and the mother is A at SNP 1 and A at SNP 2 on one homolog, and B at SNP 1 and B at SNP 2 on homolog two. If the father is A for both SNPs on both homologs, and a B is measured for the fetus SNP 1, this indicates that homolog two has been inherited by the fetus, and therefore that there is a much higher likelihood of a B being present on the fetus at SNP 2. A model that takes into account linkage would predict this, while a model that does not take linkage into account would not. Alternately, if a mother was AB at SNP 1 and AB at nearby SNP 2, then two hypotheses corresponding to maternal trisomy at that location could be used – one involving a matching copy error (nondisjunction in meiosis II or mitosis in early fetal development), and one involving an unmatching copy error (nondisjunction in meiosis I). In the case of a matching

copy error trisomy, if the fetus inherited an AA from the mother at SNP 1, then the fetus is much more likely to inherit either an AA or BB from the mother at SNP 2, but not AB. In the case of an unmatching copy error, the fetus would inherit an AB from the mother at both SNPs. The allele distribution hypotheses made by a ploidy calling method that takes into account linkage would make these predictions, and therefore correspond to the actual allele measurements to a considerably greater extent than a ploidy calling method that did not take into account linkage. Note that a linkage approach is not possible when using a method that relies on calculating allele ratios and aggregating those allele ratios.

One reason that it is believed that ploidy determinations that use a method that comprises comparing the observed allele measurements to theoretical hypotheses corresponding to possible fetal genetic states are of higher accuracy is that when sequencing is used to measure the alleles, this method can glean more information from data from alleles where the total number of reads is low than other methods; for example, a method that relies on calculating and aggregating allele ratios would produce disproportionately weighted stochastic noise. For example, imagine a case that involved measuring the alleles using sequencing, and where there was a set of loci where only five sequence reads were detected for each locus. In an embodiment, for each of the alleles, the data may be compared to the hypothesized allele distribution, and weighted according to the number of sequence reads; therefore the data from these measurements would be appropriately weighted and incorporated into the overall determination. This is in contrast to a method that involved quantitating a ratio of alleles at a heterozygous locus, as this method could only calculate ratios of 0%, 20%, 40%, 60%, 80% or 100% as the possible allele ratios; none of these may be close to expected allele ratios. In this latter case, the calculated allele rations would either have to be discarded due to insufficient reads or else would have disproportionate weighting and introduce stochastic noise into the determination, thereby decreasing the accuracy of the determination. In an embodiment, the individual allele measurements may be treated as independent measurements, where the relationship between measurements made on alleles at the same locus is no different from the relationship between measurements made on alleles at different loci.

In an embodiment, a method disclosed herein involves determining whether the distribution of observed allele measurements is indicative of a euploid or an aneuploid fetus without comparing any metrics to observed allele measurements on a reference chromosome that is expected to be disomic (termed the RC method). This is a significant improvement over

methods, such as methods using shotgun sequencing which detect aneuploidy by evaluating the proportion of randomly sequenced fragments from a suspect chromosomes relative to one or more presumed disomic reference chromosome. This RC method yields incorrect results if the presumed disomic reference chromosome is not actually disomic. This can occur in cases where aneuploidy is more substantial than trisomy of a single chromosome or where the fetus is triploid and all autosomes are trisomic. In the case of a female triploid (69, XXX) fetus there are in fact no disomic chromosomes at all. The method described herein does not require a reference chromosome and would be able to correctly identify trisomic chromosomes in a female triploid fetus. For each chromosome, hypothesis, child fraction and noise level, a joint distribution model may be fit, without any of: reference chromosome data, an overall child fraction estimate, or a fixed reference hypothesis.

In an embodiment, a method disclosed herein demonstrates how observing allele distributions at polymorphic loci can be used to determine the ploidy state of a fetus with greater accuracy than methods in the prior art. In an embodiment, the method uses the targeted sequencing to obtain mixed maternal-fetal genotypes and optionally mother and/or father genotypes at a plurality of SNPs to first establish the various expected allele frequency distributions under the different hypotheses, and then observing the quantitative allele information obtained on the maternal-fetal mixture and evaluating which hypothesis fits the data best, where the genetic state corresponding to the hypothesis with the best fit to the data is called as the correct genetic state. In an embodiment, a method disclosed herein also uses the degree of fit to generate a confidence that the called genetic state is the correct genetic state. In an embodiment, a method disclosed herein involves using algorithms that analyze the distribution of alleles found for loci that have different parental contexts, and comparing the observed allele distributions to the expected allele distributions for different ploidy states for the different parental contexts (different parental genotypic patterns). This is different from and an improvement over methods that do not use methods that enable the estimation of the number of independent instances of each allele at each locus in a mixed maternal-fetal sample. In an embodiment, a method disclosed herein involves determining whether the distribution of observed allele measurements is indicative of a euploid or an aneuploid fetus using observed allelic distributions measured at loci where the mother is heterozygous. This is different from and an improvement over methods that do not use observed allelic distributions at loci where the mother is heterozygous because, in cases where the DNA is not preferentially enriched or

is preferentially enriched for loci that are not known to be highly informative for that particular target individual, it allows the use of about twice as much genetic measurement data from a set of sequence data in the ploidy determination, resulting in a more accurate determination.

In an embodiment, a method disclosed herein uses a joint distribution model that assumes that the allele frequencies at each locus are multinomial (and thus binomial when SNPs are biallelic) in nature. In some embodiments the joint distribution model uses beta-binomial distributions. When using a measuring technique, such as sequencing, provides a quantitative measure for each allele present at each locus, binomal model can be applied to each locus and the degree underlying allele frequencies and the confidence in that frequency can be ascertained. With methods known in the art that generate ploidy calls from allele ratios, or methods in which quantitative allele information is discarded, the certainty in the observed ratio cannot be ascertained. The instant method is different from and an improvement over methods that calculate allele ratios and aggregate those ratios to make a ploidy call, since any method that involves calculating an allele ratio at a particular locus, and then aggregating those ratios, necessarily assumes that the measured intensities or counts that are indicative of the amount of DNA from any given allele or locus will be distributed in a Gaussian fashion. The method disclosed herein does not involve calculating allele ratios. In some embodiments, a method disclosed herein may involve incorporating the number of observations of each allele at a plurality of loci into a model. In some embodiments, a method disclosed herein may involve calculating the expected distributions themselves, allowing the use of a joint binomial distribution model which may be more accurate than any model that assumes a Gaussian distribution of allele measurements. The likelihood that the binomial distribution model is significantly more accurate than the Gaussian distribution increases as the number of loci increases. For example, when fewer than 20 loci are interrogated, the likelihood that the binomial distribution model is significantly better is low. However, when more than 100, or especially more than 400, or especially more than 1,000, or especially more than 2,000 loci are used, the binomial distribution model will have a very high likelihood of being significantly more accurate than the Gaussian distribution model, thereby resulting in a more accurate ploidy determination. The likelihood that the binomial distribution model is significantly more accurate than the Gaussian distribution also increases as the number of observations at each locus increases. For example, when fewer than 10 distinct sequences are observed at each locus are observed, the likelihood that the binomial distribution model is significantly better is

low. However, when more than 50 sequence reads, or especially more than 100 sequence reads, or especially more than 200 sequence reads, or especially more than 300 sequence reads are used for each locus, the binomial distribution model will have a very high likelihood of being significantly more accurate than the Gaussian distribution model, thereby resulting in a more accurate ploidy determination.

In an embodiment, a method disclosed herein uses sequencing to measure the number of instances of each allele at each locus in a DNA sample. Each sequencing read may be mapped to a specific locus and treated as a binary sequence read; alternately, the probability of the identity of the read and/or the mapping may be incorporated as part of the sequence read, resulting in a probabilistic sequence read, that is, the probable whole or fractional number of sequence reads that map to a given loci. Using the binary counts or probability of counts it is possible to use a binomial distribution for each set of measurements, allowing a confidence interval to be calculated around the number of counts. This ability to use the binomial distribution allows for more accurate ploidy estimations and more precise confidence intervals to be calculated. This is different from and an improvement over methods that use intensities to measure the amount of an allele present, for example methods that use microarrays, or methods that make measurements using fluorescence readers to measure the intensity of fluorescently tagged DNA in electrophoretic bands.

In an embodiment, a method disclosed herein uses aspects of the present set of data to determine parameters for the estimated allele frequency distribution for that set of data. This is an improvement over methods that utilize training set of data or prior sets of data to set parameters for the present expected allele frequency distributions, or possibly expected allele ratios. This is because there are different sets of conditions involved in the collection and measurement of every genetic sample, and thus a method that uses data from the instant set of data to determine the parameters for the joint distribution model that is to be used in the ploidy determination for that sample will tend to be more accurate.

In an embodiment, a method disclosed herein involves determining whether the distribution of observed allele measurements is indicative of a euploid or an aneuploid fetus using a maximum likelihood technique. The use of a maximum likelihood technique is different from and a significant improvement over methods that use single hypothesis rejection technique in that the resultant determinations will be made with significantly higher accuracy. One reason is that single hypothesis rejection techniques set cut off thresholds based on only

one measurement distribution rather than two, meaning that the thresholds are usually not optimal. Another reason is that the maximum likelihood technique allows the optimization of the cut off threshold for each individual sample instead of determining a cut off threshold to be used for all samples regardless of the particular characteristics of each individual sample. Another reason is that the use of a maximum likelihood technique allows the calculation of a confidence for each ploidy call. The ability to make a confidence calculation for each call allows a practitioner to know which calls are accurate, and which are more likely to be wrong. In some embodiments, a wide variety of methods may be combined with a maximum likelihood estimation technique to enhance the accuracy of the ploidy calls. In an embodiment, the maximum likelihood technique may be used in combination with the method described in US Patent 7,888,017. In an embodiment, the maximum likelihood technique may be used in combination with the method of using targeted PCR amplification to amplify the DNA in the mixed sample followed by sequencing and analysis using a read counting method such as used by TANDEM DIAGNOSTICS, as presented at the International Congress of Human Genetics 2011, in Montreal in October 2011. In an embodiment, a method disclosed herein involves estimating the fetal fraction of DNA in the mixed sample and using that estimation to calculate both the ploidy call and the confidence of the ploidy call. Note that this is both different and distinct from methods that use estimated fetal fraction as a screen for sufficient fetal fraction, followed by a ploidy call made using a single hypothesis rejection technique that does not take into account the fetal fraction nor does it produce a confidence calculation for the call.

In an embodiment, a method disclosed herein takes into account the tendency for the data to be noisy and contain errors by attaching a probability to each measurement. The use of maximum likelihood techniques to choose the correct hypothesis from the set of hypotheses that were made using the measurement data with attached probabilistic estimates makes it more likely that the incorrect measurements will be discounted, and the correct measurements will be used in the calculations that lead to the ploidy call. To be more precise, this method systematically reduces the influence of data that is incorrectly measured on the ploidy determination. This is an improvement over methods where all data is assumed to be equally correct or methods where outlying data is arbitrarily excluded from calculations leading to a ploidy call. Existing methods using channel ratio measurements claim to extend the method to multiple SNPs by averaging individual SNP channel ratios. Not weighting individual SNPs by expected measurement variance based on the SNP quality and observed depth of read reduces

the accuracy of the resulting statistic, resulting in a reduction of the accuracy of the ploidy call significantly, especially in borderline cases.

In an embodiment, a method disclosed herein does not presuppose the knowledge of which SNPs or other polymorphic loci are heterozygous on the fetus. This method allows a ploidy call to be made in cases where paternal genotypic information is not available. This is an improvement over methods where the knowledge of which SNPs are heterozygous must be known ahead of time in order to appropriately select loci to target, or to interpret the genetic measurements made on the mixed fetal/maternal DNA sample.

The methods described herein are particularly advantageous when used on samples where a small amount of DNA is available, or where the percent of fetal DNA is low. This is due to the correspondingly higher allele dropout rate that occurs when only a small amount of DNA is available and/or the correspondingly higher fetal allele dropout rate when the percent of fetal DNA is low in a mixed sample of fetal and maternal DNA. A high allele dropout rate, meaning that a large percentage of the alleles were not measured for the target individual, results in poorly accurate fetal fractions calculations, and poorly accurate ploidy determinations. Since methods disclosed herein may use a joint distribution model that takes into account the linkage in inheritance patterns between SNPs, significantly more accurate ploidy determinations may be made. The methods described herein allow for an accurate ploidy determination to be made when the percent of molecules of DNA that are fetal in the mixture is less than 40%, less than 30%, less than 20%, less than 10%, less than 8%, and even less than 6%.

In an embodiment, it is possible to determine the ploidy state of an individual based on measurements when that individual's DNA is mixed with DNA of a related individual. In an embodiment, the mixture of DNA is the free floating DNA found in maternal plasma, which may include DNA from the mother, with known karyotype and known genotype, and which may be mixed with DNA of the fetus, with unknown karyotype and unknown genotype. It is possible to use the known genotypic information from one or both parents to predict a plurality of potential genetic states of the DNA in the mixed sample for different ploidy states, different chromosome contributions from each parent to the fetus, and optionally, different fetal DNA fractions in the mixture. Each potential composition may be referred to as a hypothesis. The ploidy state of the fetus can then be determined by looking at the actual measurements, and determining which potential compositions are most likely given the observed data.

Further discussion of the points above may be found elsewhere in this document.

Non-Invasive Prenatal Diagnosis (NPD)

The process of non-invasive prenatal diagnosis involves a number of steps. Some of the steps may include: (1) obtaining the genetic material from the fetus; (2) enriching the genetic material of the fetus that may be in a mixed sample, *ex vivo*; (3) amplifying the genetic material, *ex vivo*; (4) preferentially enriching specific loci in the genetic material, *ex vivo*; (5) measuring the genetic material, *ex vivo*; and (6) analyzing the genotypic data, on a computer, and *ex vivo*. Methods to reduce to practice these six and other relevant steps are described herein. At least some of the method steps are not directly applied on the body. In an embodiment, the present disclosure relates to methods of treatment and diagnosis applied to tissue and other biological materials isolated and separated from the body. At least some of the method steps are executed on a computer.

Some embodiments of the present disclosure allow a clinician to determine the genetic state of a fetus that is gestating in a mother in a non-invasive manner such that the health of the baby is not put at risk by the collection of the genetic material of the fetus, and that the mother is not required to undergo an invasive procedure. Moreover, in certain aspects, the present disclosure allows the fetal genetic state to be determined with high accuracy, significantly greater accuracy than, for example, the non-invasive maternal serum analyte based screens, such as the triple test, that are in wide use in prenatal care.

The high accuracy of the methods disclosed herein is a result of an informatics approach to analysis of the genotype data, as described herein. Modern technological advances have resulted in the ability to measure large amounts of genetic information from a genetic sample using such methods as high throughput sequencing and genotyping arrays. The methods disclosed herein allow a clinician to take greater advantage of the large amounts of data available, and make a more accurate diagnosis of the fetal genetic state. The details of a number of embodiments are given below. Different embodiments may involve different combinations of the aforementioned steps. Various combinations of the different embodiments of the different steps may be used interchangeably.

In an embodiment, a blood sample is taken from a pregnant mother, and the free floating DNA in the plasma of the mother's blood, which contains a mixture of both DNA of maternal origin, and DNA of fetal origin, is isolated and used to determine the ploidy status of

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 120

the fetus. In an embodiment, a method disclosed herein involves preferential enrichment of those DNA sequences in a mixture of DNA that correspond to polymorphic alleles in a way that the allele ratios and/or allele distributions remain mostly consistent upon enrichment. In an embodiment, a method disclosed herein involves the highly efficient targeted PCR based amplification such that a very high percentage of the resulting molecules correspond to targeted loci. In an embodiment, a method disclosed herein involves sequencing a mixture of DNA that contains both DNA of maternal origin, and DNA of fetal origin. In an embodiment, a method disclosed herein involves using measured allele distributions to determine the ploidy state of a fetus that is gestating in a mother. In an embodiment, a method disclosed herein involves reporting the determined ploidy state to a clinician. In an embodiment, a method disclosed herein involves taking a clinical action, for example, performing follow up invasive testing such as chorionic villus sampling or amniocentesis, preparing for the birth of a trisomic individual or an elective termination of a trisomic fetus.

This application makes reference to U.S. Utility Application Serial No. 11/603,406, filed November 28, 2006 (US Publication No.: 20070184467); U.S. Utility Application Serial No. 12/076,348, filed March 17, 2008 (US Publication No.: 20080243398); PCT Application Serial No. PCT/US09/52730, filed August 4, 2009 (PCT Publication No.: WO/2010/017214); PCT Application Serial No. PCT/US10/050824, filed September 30, 2010 (PCT Publication No.: WO/2011/041485), U.S. Utility Application Serial No. 13/110,685, filed May 18, 2011, and PCT Application Serial No. PCT/12/58578, filed October 3, 2012, which are each herein incorporated by reference in its entirety. Some of the vocabulary used in this filing may have its antecedents in these references. Some of the concepts described herein may be better understood in light of the concepts found in these references.

Screening Maternal Blood Comprising Free Floating Fetal DNA

The methods described herein may be used to help determine the genotype of a child, fetus, or other target individual where the genetic material of the target is found in the presence of a quantity of other genetic material. In some embodiments the genotype may refer to the ploidy state of one or a plurality of chromosomes, it may refer to one or a plurality of disease linked alleles, or some combination thereof. In this disclosure, the discussion focuses on determining the genetic state of a fetus where the fetal DNA is found in maternal blood, but this example is not meant to limit to possible contexts that this method may be applied to. In

addition, the method may be applicable in cases where the amount of target DNA is in any proportion with the non-target DNA; for example, the target DNA could make up anywhere between 0.000001 and 99.999999% of the DNA present. In addition, the non-target DNA does not necessarily need to be from one individual, or even from a related individual, as long as genetic data from some or all of the relevant non-target individual(s) is known. In an embodiment, a method disclosed herein can be used to determine genotypic data of a fetus from maternal blood that contains fetal DNA. It may also be used in a case where there are multiple fetuses in the uterus of a pregnant woman, or where other contaminating DNA may be present in the sample, for example from other already born siblings.

This technique may make use of the phenomenon of fetal blood cells gaining access to maternal circulation through the placental villi. Ordinarily, only a very small number of fetal cells enter the maternal circulation in this fashion (not enough to produce a positive Kleihauer-Betke test for fetal-maternal hemorrhage). The fetal cells can be sorted out and analyzed by a variety of techniques to look for particular DNA sequences, but without the risks that invasive procedures inherently have. This technique may also make use of the phenomenon of free floating fetal DNA gaining access to maternal circulation by DNA release following apoptosis of placental tissue where the placental tissue in question contains DNA of the same genotype as the fetus. The free floating DNA found in maternal plasma has been shown to contain fetal DNA in proportions as high as 30-40% fetal DNA.

In an embodiment, blood may be drawn from a pregnant woman. Research has shown that maternal blood may contain a small amount of free floating DNA from the fetus, in addition to free floating DNA of maternal origin. In addition, there also may be enucleated fetal blood cells comprising DNA of fetal origin, in addition to many blood cells of maternal origin, which typically do not contain nuclear DNA. There are many methods know in the art to isolate fetal DNA, or create fractions enriched in fetal DNA. For example, chromatography has been show to create certain fractions that are enriched in fetal DNA.

Once the sample of maternal blood, plasma, or other fluid, drawn in a relatively non-invasive manner, and that contains an amount of fetal DNA, either cellular or free floating, either enriched in its proportion to the maternal DNA, or in its original ratio, is in hand, one may genotype the DNA found in said sample. In some embodiments, the blood may be drawn using a needle to withdraw blood from a vein, for example, the basilica vein. The method described herein can be used to determine genotypic data of the fetus. For example, it can be

used to determine the ploidy state at one or more chromosomes, it can be used to determine the identity of one or a set of SNPs, including insertions, deletions, and translocations. It can be used to determine one or more haplotypes, including the parent of origin of one or more genotypic features.

Note that this method will work with any nucleic acids that can be used for any genotyping and/or sequencing methods, such as the ILLUMINA INFINIUM ARRAY platform, AFFYMETRIX GENECHIP, ILLUMINA GENOME ANALYZER, or LIFE TECHNOLGIES' SOLID SYSTEM. This includes extracted free-floating DNA from plasma or amplifications (*e.g.* whole genome amplification, PCR) of the same; genomic DNA from other cell types (*e.g.* human lymphocytes from whole blood) or amplifications of the same. For preparation of the DNA, any extraction or purification method that generates genomic DNA suitable for the one of these platforms will work as well. This method could work equally well with samples of RNA. In an embodiment, storage of the samples may be done in a way that will minimize degradation (*e.g.* below freezing, at about -20 C, or at a lower temperature).

Parental Support

Some embodiments may be used in combination with the PARENTAL SUPPORT™ (PS) method, embodiments of which are described in U.S. Application No. 11/603,406 (US Publication No.: 20070184467), U.S. Application No. 12/076,348 (US Publication No.: 20080243398), U.S. Application 13/110,685, PCT Application PCT/US09/52730 (PCT Publication No.: WO/2010/017214), and PCT Application No. PCT/US10/050824 (PCT Publication No.: WO/2011/041485) which are incorporated herein by reference in their entirety. PARENTAL SUPPORT™ is an informatics based approach that can be used to analyze genetic data. In some embodiments, the methods disclosed herein may be considered as part of the PARENTAL SUPPORT™ method. In some embodiments, The PARENTAL SUPPORT ™ method is a collection of methods that may be used to determine the genetic data of a target individual, with high accuracy, of one or a small number of cells from that individual, or of a mixture of DNA consisting of DNA from the target individual and DNA from one or a plurality of other individuals, specifically to determine disease-related alleles, other alleles of interest, and/or the ploidy state of one or a plurality of chromosomes in the target individual. PARENTAL SUPPORT™ may refer to any of these methods. PARENTAL SUPPORT™ is an example of an informatics based method. Exemplary embodiments of the

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 123

PARENTAL SUPPORT™ method are illustrated in **Figures 29-31G** and described in Experiment 19.

The PARENTAL SUPPORT™ method makes use of known parental genetic data, *i.e.* haplotypic and/or diploid genetic data of the mother and/or the father, together with the knowledge of the mechanism of meiosis and the imperfect measurement of the target DNA, and possibly of one or more related individuals, along with population based crossover frequencies, in order to reconstruct, *in silico*, the genotype at a plurality of alleles, and/or the ploidy state of an embryo or of any target cell(s), and the target DNA at the location of key loci with a high degree of confidence. The PARENTAL SUPPORT™ method can reconstruct not only single nucleotide polymorphisms (SNPs) that were measured poorly, but also insertions and deletions, and SNPs or whole regions of DNA that were not measured at all. Furthermore, the PARENTAL SUPPORT™ method can both measure multiple disease-linked loci as well as screen for aneuploidy, from a single cell. In some embodiments, the PARENTAL SUPPORT™ method may be used to characterize one or more cells from embryos biopsied during an IVF cycle to determine the genetic condition of the one or more cells.

The PARENTAL SUPPORT™ method allows the cleaning of noisy genetic data. This may be done by inferring the correct genetic alleles in the target genome (embryo) using the genotype of related individuals (parents) as a reference. PARENTAL SUPPORT™ may be particularly relevant where only a small quantity of genetic material is available (*e.g.* PGD) and where direct measurements of the genotypes are inherently noisy due to the limited amounts of genetic material. PARENTAL SUPPORT™ may be particularly relevant where only a small fraction of the genetic material available is from the target individual (*e.g.* NPD) and where direct measurements of the genotypes are inherently noisy due to the contaminating DNA signal from another individual. The PARENTAL SUPPORT™ method is able to reconstruct highly accurate ordered diploid allele sequences on the embryo, together with copy number of chromosomes segments, even though the conventional, unordered diploid measurements may be characterized by high rates of allele dropouts, drop-ins, variable amplification biases and other errors. The method may employ both an underlying genetic model and an underlying model of measurement error. The genetic model may determine both allele probabilities at each SNP and crossover probabilities between SNPs. Allele probabilities may be modeled at each SNP based on data obtained from the parents and model crossover probabilities between SNPs based on data obtained from the HapMap database, as developed

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 124

by the International HapMap Project. Given the proper underlying genetic model and measurement error model, *maximum a posteriori* (MAP) estimation may be used, with modifications for computationally efficiency, to estimate the correct, ordered allele values at each SNP in the embryo.

The techniques outlined above, in some cases, are able to determine the genotype of an individual given a very small amount of DNA originating from that individual. This could be the DNA from one or a small number of cells, or it could be from the small amount of fetal DNA found in maternal blood.

Hypotheses

In the context of this disclosure, a hypothesis refers to a possible genetic state. It may refer to a possible ploidy state. It may refer to a possible allelic state. A set of hypotheses may refer to a set of possible genetic states, a set of possible allelic states, a set of possible ploidy states, or combinations thereof. In some embodiments, a set of hypotheses may be designed such that one hypothesis from the set will correspond to the actual genetic state of any given individual. In some embodiments, a set of hypotheses may be designed such that every possible genetic state may be described by at least one hypothesis from the set. In some embodiments of the present disclosure, one aspect of a method is to determine which hypothesis corresponds to the actual genetic state of the individual in question.

In another embodiment of the present disclosure, one step involves creating a hypothesis. In some embodiments it may be a copy number hypothesis. In some embodiments it may involve a hypothesis concerning which segments of a chromosome from each of the related individuals correspond genetically to which segments, if any, of the other related individuals. Creating a hypothesis may refer to the act of setting the limits of the variables such that the entire set of possible genetic states that are under consideration are encompassed by those variables.

A "copy number hypothesis," also called a "ploidy hypothesis," or a "ploidy state hypothesis," may refer to a hypothesis concerning a possible ploidy state for a given chromosome copy, chromosome type, or section of a chromosome, in the target individual. It may also refer to the ploidy state at more than one of the chromosome types in the individual. A set of copy number hypotheses may refer to a set of hypotheses where each hypothesis corresponds to a different possible ploidy state in an individual. A set of hypotheses may

concern a set of possible ploidy states, a set of possible parental haplotypes contributions, a set of possible fetal DNA percentages in the mixed sample, or combinations thereof.

A normal individual contains one of each chromosome type from each parent. However, due to errors in meiosis and mitosis, it is possible for an individual to have 0, 1, 2, or more of a given chromosome type from each parent. In practice, it is rare to see more that two of a given chromosomes from a parent. In this disclosure, some embodiments only consider the possible hypotheses where 0, 1, or 2 copies of a given chromosome come from a parent; it is a trivial extension to consider more or less possible copies originating from a parent. In some embodiments, for a given chromosome, there are nine possible hypotheses: the three possible hypothesis concerning 0, 1, or 2 chromosomes of maternal origin, multiplied by the three possible hypotheses concerning 0, 1, or 2 chromosomes of paternal origin. Let (m,f) refer to the hypothesis where m is the number of a given chromosome inherited from the mother, and f is the number of a given chromosome inherited from the father. Therefore, the nine hypotheses are (0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), and (2,2). These may also be written as H_{00}, H_{01}, H_{02}, H_{10}, H_{12}, H_{20}, H_{21}, and H_{22}. The different hypotheses correspond to different ploidy states. For example, (1,1) refers to a normal disomic chromosome; (2,1) refers to a maternal trisomy, and (0,1) refers to a paternal monosomy. In some embodiments, the case where two chromosomes are inherited from one parent and one chromosome is inherited from the other parent may be further differentiated into two cases: one where the two chromosomes are identical (matched copy error), and one where the two chromosomes are homologous but not identical (unmatched copy error). In these embodiments, there are sixteen possible hypotheses. It should be understood that it is possible to use other sets of hypotheses, and a different number of hypotheses.

In some embodiments of the present disclosure, the ploidy hypothesis refers to a hypothesis concerning which chromosome from other related individuals correspond to a chromosome found in the target individual's genome. In some embodiments, a key to the method is the fact that related individuals can be expected to share haplotype blocks, and using measured genetic data from related individuals, along with a knowledge of which haplotype blocks match between the target individual and the related individual, it is possible to infer the correct genetic data for a target individual with higher confidence than using the target individual's genetic measurements alone. As such, in some embodiments, the ploidy hypothesis may concern not only the number of chromosomes, but also which chromosomes in

related individuals are identical, or nearly identical, with one or more chromosomes in the target individual.

Once the set of hypotheses have been defined, when the algorithms operate on the input genetic data, they may output a determined statistical probability for each of the hypotheses under consideration. The probabilities of the various hypotheses may be determined by mathematically calculating, for each of the various hypotheses, the value that the probability equals, as stated by one or more of the expert techniques, algorithms, and/or methods described elsewhere in this disclosure, using the relevant genetic data as input.

Once the probabilities of the different hypotheses are estimated, as determined by a plurality of techniques, they may be combined. This may entail, for each hypothesis, multiplying the probabilities as determined by each technique. The product of the probabilities of the hypotheses may be normalized. Note that one ploidy hypothesis refers to one possible ploidy state for a chromosome.

The process of "combining probabilities," also called "combining hypotheses," or combining the results of expert techniques, is a concept that should be familiar to one skilled in the art of linear algebra. One possible way to combine probabilities is as follows: When an expert technique is used to evaluate a set of hypotheses given a set of genetic data, the output of the method is a set of probabilities that are associated, in a one-to-one fashion, with each hypothesis in the set of hypotheses. When a set of probabilities that were determined by a first expert technique, each of which are associated with one of the hypotheses in the set, are combined with a set of probabilities that were determined by a second expert technique, each of which are associated with the same set of hypotheses, then the two sets of probabilities are multiplied. This means that, for each hypothesis in the set, the two probabilities that are associated with that hypothesis, as determined by the two expert methods, are multiplied together, and the corresponding product is the output probability. This process may be expanded to any number of expert techniques. If only one expert technique is used, then the output probabilities are the same as the input probabilities. If more than two expert techniques are used, then the relevant probabilities may be multiplied at the same time. The products may be normalized so that the probabilities of the hypotheses in the set of hypotheses sum to 100%.

In some embodiments, if the combined probabilities for a given hypothesis are greater than the combined probabilities for any of the other hypotheses, then it may be considered that that hypothesis is determined to be the most likely. In some embodiments, a hypothesis may be

determined to be the most likely, and the ploidy state, or other genetic state, may be called if the normalized probability is greater than a threshold. In an embodiment, this may mean that the number and identity of the chromosomes that are associated with that hypothesis may be called as the ploidy state. In an embodiment, this may mean that the identity of the alleles that are associated with that hypothesis may be called as the allelic state. In some embodiments, the threshold may be between about 50% and about 80%. In some embodiments the threshold may be between about 80% and about 90%. In some embodiments the threshold may be between about 90% and about 95%. In some embodiments the threshold may be between about 95% and about 99%. In some embodiments the threshold may be between about 99% and about 99.9%. In some embodiments the threshold may be above about 99.9%.

Parental Contexts

The parental context refers to the genetic state of a given allele, on each of the two relevant chromosomes for one or both of the two parents of the target. Note that in an embodiment, the parental context does not refer to the allelic state of the target, rather, it refers to the allelic state of the parents. The parental context for a given SNP may consist of four base pairs, two paternal and two maternal; they may be the same or different from one another. It is typically written as "$m_1m_2|f_1f_2$," where m_1 and m_2 are the genetic state of the given SNP on the two maternal chromosomes, and f_1 and f_2 are the genetic state of the given SNP on the two paternal chromosomes. In some embodiments, the parental context may be written as "$f_1f_2|m_1m_2$." Note that subscripts "1" and "2" refer to the genotype, at the given allele, of the first and second chromosome; also note that the choice of which chromosome is labeled "1" and which is labeled "2" is arbitrary.

Note that in this disclosure, A and B are often used to generically represent base pair identities; A or B could equally well represent C (cytosine), G (guanine), A (adenine) or T (thymine). For example, if, at a given SNP based allele, the mother's genotype was T at that SNP on one chromosome, and G at that SNP on the homologous chromosome, and the father's genotype at that allele is G at that SNP on both of the homologous chromosomes, one may say that the target individual's allele has the parental context of AB|BB; it could also be said that the allele has the parental context of AB|AA. Note that, in theory, any of the four possible nucleotides could occur at a given allele, and thus it is possible, for example, for the mother to have a genotype of AT, and the father to have a genotype of GC at a given allele. However,

empirical data indicate that in most cases only two of the four possible base pairs are observed at a given allele. It is possible, for example when using single tandem repeats, to have more than two parental, more than four and even more than ten contexts. In this disclosure the discussion assumes that only two possible base pairs will be observed at a given allele, although the embodiments disclosed herein could be modified to take into account the cases where this assumption does not hold.

A "parental context" may refer to a set or subset of target SNPs that have the same parental context. For example, if one were to measure 1000 alleles on a given chromosome on a target individual, then the context AA|BB could refer to the set of all alleles in the group of 1,000 alleles where the genotype of the mother of the target was homozygous, and the genotype of the father of the target is homozygous, but where the maternal genotype and the paternal genotype are dissimilar at that locus. If the parental data is not phased, and thus AB = BA, then there are nine possible parental contexts: AA|AA, AA|AB, AA|BB, AB|AA, AB|AB, AB|BB, BB|AA, BB|AB, and BB|BB. If the parental data is phased, and thus AB ≠ BA, then there are sixteen different possible parental contexts: AA|AA, AA|AB, AA|BA, AA|BB, AB|AA, AB|AB, AB|BA, AB|BB, BA|AA, BA|AB, BA|BA, BA|BB, BB|AA, BB|AB, BB|BA, and BB|BB. Every SNP allele on a chromosome, excluding some SNPs on the sex chromosomes, has one of these parental contexts. The set of SNPs wherein the parental context for one parent is heterozygous may be referred to as the heterozygous context.

Use of Parental Contexts in NPD

Non-invasive prenatal diagnosis is an important technique that can be used to determine the genetic state of a fetus from genetic material that is obtained in a non-invasive manner, for example from a blood draw on the pregnant mother. The blood could be separated and the plasma isolated, followed by isolation of the plasma DNA. Size selection could be used to isolate the DNA of the appropriate length. The DNA may be preferentially enriched at a set of loci. This DNA can then be measured by a number of means, such as by hybridizing to a genotyping array and measuring the fluorescence, or by sequencing on a high throughput sequencer.

When sequencing is used for ploidy calling of a fetus in the context of non-invasive prenatal diagnosis, there are a number of ways to use the sequence data. The most common way one could use the sequence data is to simply count the number of reads that map to a

given chromosome. For example, imagine if you are trying to determine the ploidy state of chromosome 21 on the fetus. Further imagine that the DNA in the sample is comprised of 10% DNA of fetal origin, and 90% DNA of maternal origin. In this case, you could look at the average number of reads on a chromosome which can be expected to be disomic, for example chromosome 3, and compare that to the number of read on chromosome 21, where the reads are adjusted for the number of base pairs on that chromosome that are part of a unique sequence. If the fetus were euploid, one would expect the amount of DNA per unit of genome to be about equal at all locations (subject to stochastic variations). On the other hand, if the fetus were trisomic at chromosome 21, then one would expect there to be more slightly more DNA per genetic unit from chromosome 21 than the other locations on the genome. Specifically one would expect there to be about 5% more DNA from chromosome 21 in the mixture. When sequencing is used to measure the DNA, one would expect about 5% more uniquely mappable reads from chromosome 21 per unique segment than from the other chromosomes. One could use the observation of an amount of DNA from a particular chromosome that is higher than a certain threshold, when adjusted for the number of sequences that are uniquely mappable to that chromosome, as the basis for an aneuploidy diagnosis. Another method that may be used to detect aneuploidy is similar to that above, except that parental contexts could be taken into account.

When considering which alleles to target, one may consider the likelihood that some parental contexts are likely to be more informative than others. For example, AA|BB and the symmetric context BB|AA are the most informative contexts, because the fetus is known to carry an allele that is different from the mother. For reasons of symmetry, both AA|BB and BB|AA contexts may be referred to as AA|BB. Another set of informative parental contexts are AA|AB and BB|AB, because in these cases the fetus has a 50% chance of carrying an allele that the mother does not have. For reasons of symmetry, both AA|AB and BB|AB contexts may be referred to as AA|AB. A third set of informative parental contexts are AB|AA and AB|BB, because in these cases the fetus is carrying a known paternal allele, and that allele is also present in the maternal genome. For reasons of symmetry, both AB|AA and AB|BB contexts may be referred to as AB|AA. A fourth parental context is AB|AB where the fetus has an unknown allelic state, and whatever the allelic state, it is one in which the mother has the same alleles. The fifth parental context is AA|AA, where the mother and father are heterozygous.

Different Implementations of the Presently Disclosed Embodiments

Methods are disclosed herein for determining the ploidy state of a target individual. The target individual may be a blastomere, an embryo, or a fetus. In some embodiments of the present disclosure, a method for determining the ploidy state of one or more chromosome in a target individual may include any of the steps described in this document, and combinations thereof:

In some embodiments the source of the genetic material to be used in determining the genetic state of the fetus may be fetal cells, such as nucleated fetal red blood cells, isolated from the maternal blood. The method may involve obtaining a blood sample from the pregnant mother. The method may involve isolating a fetal red blood cell using visual techniques, based on the idea that a certain combination of colors are uniquely associated with nucleated red blood cell, and a similar combination of colors is not associated with any other present cell in the maternal blood. The combination of colors associated with the nucleated red blood cells may include the red color of the hemoglobin around the nucleus, which color may be made more distinct by staining, and the color of the nuclear material which can be stained, for example, blue. By isolating the cells from maternal blood and spreading them over a slide, and then identifying those points at which one sees both red (from the Hemoglobin) and blue (from the nuclear material) one may be able to identify the location of nucleated red blood cells. One may then extract those nucleated red blood cells using a micromanipulator, use genotyping and/or sequencing techniques to measure aspects of the genotype of the genetic material in those cells.

In an embodiment, one may stain the nucleated red blood cell with a die that only fluoresces in the presence of fetal hemoglobin and not maternal hemoglobin, and so remove the ambiguity between whether a nucleated red blood cell is derived from the mother or the fetus. Some embodiments of the present disclosure may involve staining or otherwise marking nuclear material. Some embodiments of the present disclosure may involve specifically marking fetal nuclear material using fetal cell specific antibodies.

There are many other ways to isolate fetal cells from maternal blood, or fetal DNA from maternal blood, or to enrich samples of fetal genetic material in the presence of maternal genetic material. Some of these methods are listed here, but this is not intended to be an exhaustive list. Some appropriate techniques are listed here for convenience: using

fluorescently or otherwise tagged antibodies, size exclusion chromatography, magnetically or otherwise labeled affinity tags, epigenetic differences, such as differential methylation between the maternal and fetal cells at specific alleles, density gradient centrifugation succeeded by CD45/14 depletion and CD71-positive selection from CD45/14 negative-cells, single or double Percoll gradients with different osmolalities, or galactose specific lectin method.

In an embodiment of the present disclosure, the target individual is a fetus, and the different genotype measurements are made on a plurality of DNA samples from the fetus. In some embodiments of the present disclosure, the fetal DNA samples are from isolated fetal cells where the fetal cells may be mixed with maternal cells. In some embodiments of the present disclosure, the fetal DNA samples are from free floating fetal DNA, where the fetal DNA may be mixed with free floating maternal DNA. In some embodiments, the fetal DNA samples may be derived from maternal plasma or maternal blood that contains a mixture of maternal DNA and fetal DNA. In some embodiments, the fetal DNA may be mixed with maternal DNA in maternal:fetal ratios ranging from 99.9:0.1% to 99:1%; 99:1% to 90:10%; 90:10% to 80:20%; 80:20% to 70:30%; 70:30% to 50:50%; 50:50% to 10:90%; or 10:90% to 1:99%; 1:99% to 0.1:99.9%.

The genetic data of the target individual and/or of the related individual can be transformed from a molecular state to an electronic state by measuring the appropriate genetic material using tools and or techniques taken from a group including, but not limited to: genotyping microarrays, and high throughput sequencing. Some high throughput sequencing methods include Sanger DNA sequencing, pyrosequencing, the ILLUMINA SOLEXA platform, ILLUMINA's GENOME ANALYZER, or APPLIED BIOSYSTEM's 454 sequencing platform, HELICOS's TRUE SINGLE MOLECULE SEQUENCING platform, HALCYON MOLECULAR's electron microscope sequencing method, or any other sequencing method,. All of these methods physically transform the genetic data stored in a sample of DNA into a set of genetic data that is typically stored in a memory device en route to being processed.

A relevant individual's genetic data may be measured by analyzing substances taken from a group including, but not limited to: the individual's bulk diploid tissue, one or more diploid cells from the individual, one or more haploid cells from the individual, one or more blastomeres from the target individual, extra-cellular genetic material found on the individual, extra-cellular genetic material from the individual found in maternal blood, cells from the

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 132

individual found in maternal blood, one or more embryos created from (a) gamete(s) from the related individual, one or more blastomeres taken from such an embryo, extra-cellular genetic material found on the related individual, genetic material known to have originated from the related individual, and combinations thereof.

In some embodiments, a set of at least one ploidy state hypothesis may be created for each of the chromosomes types of interest of the target individual. Each of the ploidy state hypotheses may refer to one possible ploidy state of the chromosome or chromosome segment of the target individual. The set of hypotheses may include some or all of the possible ploidy states that the chromosome of the target individual may be expected to have. Some of the possible ploidy states may include nullsomy, monosomy, disomy, uniparental disomy, euploidy, trisomy, matching trisomy, unmatching trisomy, maternal trisomy, paternal trisomy, tetrasomy, balanced (2:2) tetrasomy, unbalanced (3:1) tetrasomy, pentasomy, hexasomy, other aneuploidy, and combinations thereof. Any of these aneuploidy states may be mixed or partial aneuploidy such as unbalanced translocations, balanced translocations, Robertsonian translocations, recombinations, deletions, insertions, crossovers, and combinations thereof.

In some embodiments, the knowledge of the determined ploidy state may be used to make a clinical decision. This knowledge, typically stored as a physical arrangement of matter in a memory device, may then be transformed into a report. The report may then be acted upon. For example, the clinical decision may be to terminate the pregnancy; alternately, the clinical decision may be to continue the pregnancy. In some embodiments the clinical decision may involve an intervention designed to decrease the severity of the phenotypic presentation of a genetic disorder, or a decision to take relevant steps to prepare for a special needs child.

In an embodiment of the present disclosure, any of the methods described herein may be modified to allow for multiple targets to come from same target individual, for example, multiple blood draws from the same pregnant mother. This may improve the accuracy of the model, as multiple genetic measurements may provide more data with which the target genotype may be determined. In an embodiment, one set of target genetic data served as the primary data which was reported, and the other served as data to double-check the primary target genetic data. In an embodiment, a plurality of sets of genetic data, each measured from genetic material taken from the target individual, are considered in parallel, and thus both sets of target genetic data serve to help determine which sections of parental genetic data, measured with high accuracy, composes the fetal genome.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 133

In an embodiment, the method may be used for the purpose of paternity testing. For example, given the SNP-based genotypic information from the mother, and from a man who may or may not be the genetic father, and the measured genotypic information from the mixed sample, it is possible to determine if the genotypic information of the male indeed represents that actual genetic father of the gestating fetus. A simple way to do this is to simply look at the contexts where the mother is AA, and the possible father is AB or BB. In these cases, one may expect to see the father contribution half (AA|AB) or all (AA|BB) of the time, respectively. Taking into account the expected ADO, it is straightforward to determine whether or not the fetal SNPs that are observed are correlated with those of the possible father.

One embodiment of the present disclosure could be as follows: a pregnant woman wants to know if her fetus is afflicted with Down Syndrome, and/or if it will suffer from Cystic Fibrosis, and she does not wish to bear a child that is afflicted with either of these conditions. A doctor takes her blood, and stains the hemoglobin with one marker so that it appears clearly red, and stains nuclear material with another marker so that it appears clearly blue. Knowing that maternal red blood cells are typically anuclear, while a high proportion of fetal cells contain a nucleus, the doctor is able to visually isolate a number of nucleated red blood cells by identifying those cells that show both a red and blue color. The doctor picks up these cells off the slide with a micromanipulator and sends them to a lab which amplifies and genotypes ten individual cells. By using the genetic measurements, the PARENTAL SUPPORT™ method is able to determine that six of the ten cells are maternal blood cells, and four of the ten cells are fetal cells. If a child has already been born to a pregnant mother, PARENTAL SUPPORT™ can also be used to determine that the fetal cells are distinct from the cells of the born child by making reliable allele calls on the fetal cells and showing that they are dissimilar to those of the born child. Note that this method is similar in concept to the paternal testing embodiment of the present disclosure. The genetic data measured from the fetal cells may be of very poor quality, comprising many allele drop outs, due to the difficulty of genotyping single cells. The clinician is able to use the measured fetal DNA along with the reliable DNA measurements of the parents to infer aspects of the genome of the fetus with high accuracy using PARENTAL SUPPORT™, thereby transforming the genetic data contained on genetic material from the fetus into the predicted genetic state of the fetus, stored on a computer. The clinician is able to determine both the ploidy state of the fetus, and the presence or absence of a plurality of

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 134

disease-linked genes of interest. It turns out that the fetus is euploid, and is not a carrier for cystic fibrosis, and the mother decides to continue the pregnancy.

In an embodiment of the present disclosure, a pregnant mother would like to determine if her fetus is afflicted with any whole chromosomal abnormalities. She goes to her doctor, and gives a sample of her blood, and she and her husband gives samples of their own DNA from cheek swabs. A laboratory researcher genotypes the parental DNA using the MDA protocol to amplify the parental DNA, and ILLUMINA INFINIUM arrays to measure the genetic data of the parents at a large number of SNPs. The researcher then spins down the blood, takes the plasma, and isolates a sample of free-floating DNA using size exclusion chromatography. Alternately, the researcher uses one or more fluorescent antibodies, such as one that is specific to fetal hemoglobin to isolate a nucleated fetal red blood cell. The researcher then takes the isolated or enriched fetal genetic material and amplifies it using a library of 70-mer oligonucleotides appropriately designed such that two ends of each oligonucleotide corresponded to the flanking sequences on either side of a target allele. Upon addition of a polymerase, ligase, and the appropriate reagents, the oligonucleotides underwent gap-filling circularization, capturing the desired allele. An exonuclease was added, heat-inactivated, and the products were used directly as a template for PCR amplification. The PCR products were sequenced on an ILLUMINA GENOME ANALYZER. The sequence reads were used as input for the PARENTAL SUPPORT™ method, which then predicted the ploidy state of the fetus.

In another embodiment, a couple - where the mother, who is pregnant, and is of advanced maternal age - wants to know whether the gestating fetus has Down syndrome, Turner Syndrome, Prader Willi syndrome, or some other whole chromosomal abnormality. The obstetrician takes a blood draw from the mother and father. The blood is sent to a laboratory, where a technician centrifuges the maternal sample to isolate the plasma and the buffy coat. The DNA in the buffy coat and the paternal blood sample are transformed through amplification and the genetic data encoded in the amplified genetic material is further transformed from molecularly stored genetic data into electronically stored genetic data by running the genetic material on a high throughput sequencer to measure the parental genotypes. The plasma sample is preferentially enriched at a set of loci using a 5,000-plex hemi-nested targeted PCR method. The mixture of DNA fragments is prepared into a DNA library suitable for sequencing. The DNA is then sequenced using a high throughput

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 135

sequencing method, for example, the ILLUMINA GAIIx GENOME ANALYZER. The sequencing transforms the information that is encoded molecularly in the DNA into information that is encoded electronically in computer hardware. An informatics based technique that includes the presently disclosed embodiments, such as PARENTAL SUPPORT™, may be used to determine the ploidy state of the fetus. This may involve calculating, on a computer, allele count probabilities at the plurality of polymorphic loci from the DNA measurements made on the prepared sample; creating, on a computer, a plurality of ploidy hypotheses each pertaining to a different possible ploidy state of the chromosome; building, on a computer, a joint distribution model for the expected allele counts at the plurality of polymorphic loci on the chromosome for each ploidy hypothesis; determining, on a computer, a relative probability of each of the ploidy hypotheses using the joint distribution model and the allele counts measured on the prepared sample; and calling the ploidy state of the fetus by selecting the ploidy state corresponding to the hypothesis with the greatest probability. It is determined that the fetus has Down syndrome. A report is printed out, or sent electronically to the pregnant woman's obstetrician, who transmits the diagnosis to the woman. The woman, her husband, and the doctor sit down and discuss their options. The couple decides to terminate the pregnancy based on the knowledge that the fetus is afflicted with a trisomic condition.

In an embodiment, a company may decide to offer a diagnostic technology designed to detect aneuploidy in a gestating fetus from a maternal blood draw. Their product may involve a mother presenting to her obstetrician, who may draw her blood. The obstetrician may also collect a genetic sample from the father of the fetus. A clinician may isolate the plasma from the maternal blood, and purify the DNA from the plasma. A clinician may also isolate the buffy coat layer from the maternal blood, and prepare the DNA from the buffy coat. A clinician may also prepare the DNA from the paternal genetic sample. The clinician may use molecular biology techniques described in this disclosure to append universal amplification tags to the DNA in the DNA derived from the plasma sample. The clinician may amplify the universally tagged DNA. The clinician may preferentially enrich the DNA by a number of techniques including capture by hybridization and targeted PCR. The targeted PCR may involve nesting, hemi-nesting or semi-nesting, or any other approach to result in efficient enrichment of the plasma derived DNA. The targeted PCR may be massively multiplexed, for example with 10,000 primers in one reaction volume, where the primers target SNPs on

chromosomes 13, 18, 21, X and those loci that are common to both X and Y, and optionally other chromosomes as well. The selective enrichment and/or amplification may involve tagging each individual molecule with different tags, molecular barcodes, tags for amplification, and/or tags for sequencing. The clinician may then sequence the plasma sample, and also possibly also the prepared maternal and/or paternal DNA. The molecular biology steps may be executed either wholly or partly by a diagnostic box. The sequence data may be fed into a single computer, or to another type of computing platform such as may be found in 'the cloud'. The computing platform may calculate allele counts at the targeted polymorphic loci from the measurements made by the sequencer. The computing platform may create a plurality of ploidy hypotheses pertaining to nullsomy, monosomy, disomy, matched trisomy, and unmatched trisomy for each of chromosomes 13, 18, 21, X and Y. The computing platform may build a joint distribution model for the expected allele counts at the targeted loci on the chromosome for each ploidy hypothesis for each of the five chromosomes being interrogated. The computing platform may determine a probability that each of the ploidy hypotheses is true using the joint distribution model and the allele counts measured on the preferentially enriched DNA derived from the plasma sample. The computing platform may call the ploidy state of the fetus, for each of chromosome 13, 18, 21, X and Y by selecting the ploidy state corresponding to the germane hypothesis with the greatest probability. A report may be generated comprising the called ploidy states, and it may be sent to the obstetrician electronically, displayed on an output device, or a printed hard copy of the report may be delivered to the obstetrician. The obstetrician may inform the patient and optionally the father of the fetus, and they may decide which clinical options are open to them, and which is most desirable.

In another embodiment, a pregnant woman, hereafter referred to as "the mother" may decide that she wants to know whether or not her fetus(es) are carrying any genetic abnormalities or other conditions. She may want to ensure that there are not any gross abnormalities before she is confident to continue the pregnancy. She may go to her obstetrician, who may take a sample of her blood. He may also take a genetic sample, such as a buccal swab, from her cheek. He may also take a genetic sample from the father of the fetus, such as a buccal swab, a sperm sample, or a blood sample. He may send the samples to a clinician. The clinician may enrich the fraction of free floating fetal DNA in the maternal blood sample. The clinician may enrich the fraction of enucleated fetal blood cells in the

maternal blood sample. The clinician may use various aspects of the methods described herein to determine genetic data of the fetus. That genetic data may include the ploidy state of the fetus, and/or the identity of one or a number of disease linked alleles in the fetus. A report may be generated summarizing the results of the prenatal diagnosis. The report may be transmitted or mailed to the doctor, who may tell the mother the genetic state of the fetus. The mother may decide to discontinue the pregnancy based on the fact that the fetus has one or more chromosomal, or genetic abnormalities, or undesirable conditions. She may also decide to continue the pregnancy based on the fact that the fetus does not have any gross chromosomal or genetic abnormalities, or any genetic conditions of interest.

Another example may involve a pregnant woman who has been artificially inseminated by a sperm donor, and is pregnant. She wants to minimize the risk that the fetus she is carrying has a genetic disease. She has blood drawn at a phlebotomist, and techniques described in this disclosure are used to isolate three nucleated fetal red blood cells, and a tissue sample is also collected from the mother and genetic father. The genetic material from the fetus and from the mother and father are amplified as appropriate and genotyped using the ILLUMINA INFINIUM BEADARRAY, and the methods described herein clean and phase the parental and fetal genotype with high accuracy, as well as to make ploidy calls for the fetus. The fetus is found to be euploid, and phenotypic susceptibilities are predicted from the reconstructed fetal genotype, and a report is generated and sent to the mother's physician so that they can decide what clinical decisions may be best.

In an embodiment, the raw genetic material of the mother and the father is transformed by way of amplification to an amount of DNA that is similar in sequence, but larger in quantity. Then, by way of a genotyping method, the genotypic data that is encoded by nucleic acids is transformed into genetic measurements that may be stored physically and/or electronically on a memory device, such as those described above. The relevant algorithms that makeup the PARENTAL SUPPORT™ algorithm, relevant parts of which are discussed in detail herein, are translated into a computer program, using a programming language. Then, through the execution of the computer program on the computer hardware, instead of being physically encoded bits and bytes, arranged in a pattern that represents raw measurement data, they become transformed into a pattern that represents a high confidence determination of the ploidy state of the fetus. The details of this transformation will rely on the data itself and the computer language and hardware system used to execute the method described herein. Then,

the data that is physically configured to represent a high quality ploidy determination of the fetus is transformed into a report which may be sent to a health care practitioner. This transformation may be carried out using a printer or a computer display. The report may be a printed copy, on paper or other suitable medium, or else it may be electronic. In the case of an electronic report, it may be transmitted, it may be physically stored on a memory device at a location on the computer accessible by the health care practitioner; it also may be displayed on a screen so that it may be read. In the case of a screen display, the data may be transformed to a readable format by causing the physical transformation of pixels on the display device. The transformation may be accomplished by way of physically firing electrons at a phosphorescent screen, by way of altering an electric charge that physically changes the transparency of a specific set of pixels on a screen that may lie in front of a substrate that emits or absorbs photons. This transformation may be accomplished by way of changing the nanoscale orientation of the molecules in a liquid crystal, for example, from nematic to cholesteric or smectic phase, at a specific set of pixels. This transformation may be accomplished by way of an electric current causing photons to be emitted from a specific set of pixels made from a plurality of light emitting diodes arranged in a meaningful pattern. This transformation may be accomplished by any other way used to display information, such as a computer screen, or some other output device or way of transmitting information. The health care practitioner may then act on the report, such that the data in the report is transformed into an action. The action may be to continue or discontinue the pregnancy, in which case a gestating fetus with a genetic abnormality is transformed into non-living fetus. The transformations listed herein may be aggregated, such that, for example, one may transform the genetic material of a pregnant mother and the father, through a number of steps outlined in this disclosure, into a medical decision consisting of aborting a fetus with genetic abnormalities, or consisting of continuing the pregnancy. Alternately, one may transform a set of genotypic measurements into a report that helps a physician treat his pregnant patient.

In an embodiment of the present disclosure, the method described herein can be used to determine the ploidy state of a fetus even when the host mother, *i.e.* the woman who is pregnant, is not the biological mother of the fetus she is carrying. In an embodiment of the present disclosure, the method described herein can be used to determine the ploidy state of a fetus using only the maternal blood sample, and without the need for a paternal genetic sample.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 139

Some of the math in the presently disclosed embodiments makes hypotheses concerning a limited number of states of aneuploidy. In some cases, for example, only zero, one or two chromosomes are expected to originate from each parent. In some embodiments of the present disclosure, the mathematical derivations can be expanded to take into account other forms of aneuploidy, such as quadrosomy, where three chromosomes originate from one parent, pentasomy, hexasomy etc., without changing the fundamental concepts of the present disclosure. At the same time, it is possible to focus on a smaller number of ploidy states, for example, only trisomy and disomy. Note that ploidy determinations that indicate a non-whole number of chromosomes may indicate mosaicism in a sample of genetic material.

In some embodiments, the genetic abnormality is a type of aneuploidy, such as Down syndrome (or trisomy 21), Edwards syndrome (trisomy 18), Patau syndrome (trisomy 13), Turner Syndrome (45X), Klinefelter's syndrome (a male with 2 X chromosomes), Prader-Willi syndrome, and DiGeorge syndrome (UPD 15). Congenital disorders, such as those listed in the prior sentence, are commonly undesirable, and the knowledge that a fetus is afflicted with one or more phenotypic abnormalities may provide the basis for a decision to terminate the pregnancy, to take necessary precautions to prepare for the birth of a special needs child, or to take some therapeutic approach meant to lessen the severity of a chromosomal abnormality.

In some embodiments, the methods described herein can be used at a very early gestational age, for example as early as four week, as early as five weeks, as early as six weeks, as early as seven weeks, as early as eight weeks, as early as nine weeks, as early as ten weeks, as early as eleven weeks, and as early as twelve weeks.

In some embodiments, a method disclosed herein is used in the context of pre-implantation genetic diagnosis (PGD) for embryo selection during in vitro fertilization, where the target individual is an embryo, and the parental genotypic data can be used to make ploidy determinations about the embryo from sequencing data from a single or two cell biopsy from a day 3 embryo or a trophectoderm biopsy from a day 5 or day 6 embryo. In a PGD setting, only the child DNA is measured, and only a small number of cells are tested, generally one to five but as many as ten, twenty or fifty. The total number of starting copies of the A and B alleles (at a SNP) are then trivially determined by the child genotype and the number of cells. In NPD, the number of starting copies is very high and so the allele ratio after PCR is expected to accurately reflect the starting ratio. However, the small number of starting copies in PGD means that contamination and imperfect PCR efficiency have a non-trivial effect on the allele

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 140

ratio following PCR. This effect may be more important than depth of read in predicting the variance in the allele ratio measured after sequencing. The distribution of measured allele ratio given a known child genotype may be created by Monte Carlo simulation of the PCR process based on the PCR probe efficiency and probability of contamination. Given an allele ratio distribution for each possible child genotype, the likelihoods of various hypotheses can be calculated as described for NIPD.

Maximum Likelihood Estimates

Most methods known in the art for detecting the presence or absence of biological phenomenon or medical condition involve the use of a single hypothesis rejection test, where a metric that is correlated with the condition is measured, and if the metric is on one side of a given threshold, the condition is present, while of the metric falls on the other side of the threshold, the condition is absent. A single-hypothesis rejection test only looks at the null distribution when deciding between the null and alternate hypotheses. Without taking into account the alternate distribution, one cannot estimate the likelihood of each hypothesis given the observed data and therefore cannot calculate a confidence on the call. Hence with a single-hypothesis rejection test, one gets a yes or no answer without a feeling for the confidence associated with the specific case.

In some embodiments, the method disclosed herein is able to detect the presence or absence of biological phenomenon or medical condition using a maximum likelihood method. This is a substantial improvement over a method using a single hypothesis rejection technique as the threshold for calling absence or presence of the condition can be adjusted as appropriate for each case. This is particularly relevant for diagnostic techniques that aim to determine the presence or absence of aneuploidy in a gestating fetus from genetic data available from the mixture of fetal and maternal DNA present in the free floating DNA found in maternal plasma. This is because as the fraction of fetal DNA in the plasma derived fraction changes, the optimal threshold for calling aneuploidy vs. euploidy changes. As the fetal fraction drops, the distribution of data that is associated with an aneuploidy becomes increasingly similar to the distribution of data that is associated with a euploidy.

The maximum likelihood estimation method uses the distributions associated with each hypothesis to estimate the likelihood of the data conditioned on each hypothesis. These conditional probabilities can then be converted to a hypothesis call and confidence. Similarly,

maximum a posteriori estimation method uses the same conditional probabilities as the maximum likelihood estimate, but also incorporates population priors when choosing the best hypothesis and determining confidence.

Therefore, the use of a maximum likelihood estimate (MLE) technique, or the closely related maximum a posteriori (MAP) technique give two advantages, first it increases the chance of a correct call, and it also allows a confidence to be calculated for each call. In an embodiment, selecting the ploidy state corresponding to the hypothesis with the greatest probability is carried out using maximum likelihood estimates or maximum a posteriori estimates. In an embodiment, a method is disclosed for determining the ploidy state of a gestating fetus that involves taking any method currently known in the art that uses a single hypothesis rejection technique and reformulating it such that it uses a MLE or MAP technique. Some examples of methods that can be significantly improved by applying these techniques can be found in US Pat 8,008,018, US Patent 7,888,017, or US Patent 7,332,277.

In an embodiment, a method is described for determining presence or absence of fetal aneuploidy in a maternal plasma sample comprising fetal and maternal genomic DNA, the method comprising: obtaining a maternal plasma sample; measuring the DNA fragments found in the plasma sample with a high throughput sequencer; mapping the sequences to the chromosome and determining the number of sequence reads that map to each chromosome; calculating the fraction of fetal DNA in the plasma sample; calculating an expected distribution of the amount of a target chromosome that would be expected to be present if that if the second target chromosome were euploid and one or a plurality of expected distributions that would be expected if that chromosome were aneuploid, using the fetal fraction and the number of sequence reads that map to one or a plurality of reference chromosomes expected to be euploid; and using a MLE or MAP determine which of the distributions is most likely to be correct, thereby indicating the presence or absence of a fetal aneuploidy. In an embodiment, the measuring the DNA from the plasma may involve conducting massively parallel shotgun sequencing. In an embodiment, the measuring the DNA from the plasma sample may involve sequencing DNA that has been preferentially enriched, for example through targeted amplification, at a plurality of polymorphic or non-polymorphic loci. The plurality of loci may be designed to target one or a small number of suspected aneuploid chromosomes and one or a small number of reference chromosomes. The purpose of the preferential enrichment is to increase the number of sequence reads that are informative for the ploidy determination.

Ploidy Calling Informatics Methods

Described herein is a method for determining the ploidy state of a fetus given sequence data. In some embodiments, this sequence data may be measured on a high throughput sequencer. In some embodiments, the sequence data may be measured on DNA that originated from free floating DNA isolated from maternal blood, wherein the free floating DNA comprises some DNA of maternal origin, and some DNA of fetal / placental origin. This section will describe one embodiment of the present disclosure in which the ploidy state of the fetus is determined assuming that fraction of fetal DNA in the mixture that has been analyzed is not known and will be estimated from the data. It will also describe an embodiment in which the fraction of fetal DNA ("fetal fraction") or the percentage of fetal DNA in the mixture can be measured by another method, and is assumed to be known in determining the ploidy state of the fetus. In some embodiments the fetal fraction can be calculated using only the genotyping measurements made on the maternal blood sample itself, which is a mixture of fetal and maternal DNA. In some embodiments the fraction may be calculated also using the measured or otherwise known genotype of the mother and/or the measured or otherwise known genotype of the father. In another embodiment ploidy state of the fetus can be determined solely based on the calculated fraction of fetal DNA for the chromosome in question compared to the calculated fraction of fetal DNA for the reference chromosome assumed disomic.

In the preferred embodiment, suppose that, for a particular chromosome, we observe and analyze N SNPs, for which we have:

- Set of NR free floating DNA sequence measurements $S=(s_1,\ldots,s_{NR})$. Since this method utilizes the SNP measurements, all sequence data that corresponds to non-polymorphic loci can be disregarded. In a simplified version, where we have (A,B) counts on each SNP, where A and B correspond to the two alleles present at a given locus, S can be written as $S=((a_1,b_1),\ldots,(a_N, b_N))$, where a_i is the A count on SNP i, b_i is the B count on SNP i, and $\sum_{i=1,N} (a_i + b_i) = NR$
- Parent data consisting of
  - genotypes from a SNP microarray or other intensity based genotyping platform: mother $M=(m_1,\ldots,m_N)$, father $F=(f_1, \ldots, f_N)$, where $m_i, f_i \in$ (AA,AB, BB).

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 143

- AND/OR sequence data measurements: NRM mother measurements $SM=(sm_1,\ldots,sm_{nrm})$, NRF father measurements $SF=(sf_1,\ldots,sf_{nrf})$. Similar to the above simplification, if we have (A,B) counts on each SNP $SM=((am_1,bm_1),\ldots,(am_N, bm_N))$, $SF=((af_1,bf_1),\ldots,(af_N, bf_N))$

Collectively, the mother, father child data are denoted as D = (M,F,SM,SF,S). Note that the parent data is desired and increases the accuracy of the algorithm, but is NOT necessary, especially the father data. This means that even in the absence of mother and/or father data, it is possible to get very accurate copy number results.

It is possible to derive the best copy number estimate (H^*) by maximizing the data log likelihood $LIK(D|H)$ over all hypotheses (H) considered. In particular it is possible to determine the relative probability of each of the ploidy hypotheses using the joint distribution model and the allele counts measured on the prepared sample, and using those relative probabilities to determine the hypothesis most likely to be correct as follows:

$$H^* = \underset{H}{\operatorname{argmax}}\, LIK(D|H)$$

Similarly the a posteriori hypothesis likelihood given the data may be written as:

$$H^* = \underset{H}{\operatorname{argmax}}\, LIK(D|H) \cdot priorprob(H)$$

Where priorprob(H) is the prior probability assigned to each hypothesis H, based on model design and prior knowledge.

It is also possible to use priors to find the maximum a posteriori estimate:

$$H_{MA} = \underset{H}{\operatorname{argmax}}\, LIK(D|H)$$

In an embodiment, the copy number hypotheses that may be considered are:

- Monosomy:
  - maternal H10 (one copy from mother)
  - paternal H01 (one copy from father)
- Disomy: H11 (one copy each mother and father)
- Simple trisomy, no crossovers considered:
  - Maternal: H21_matched (two identical copies from mother, one copy from father), H21_unmatched (BOTH copies from mother, one copy from father)

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 144

- Paternal: H12_matched (one copy from mother , two identical copies from father), H12_unmatched (one copy from mother , both copies from father)

- Composite trisomy, allowing for crossovers (using a joint distribution model):
  - maternal H21 (two copies from mother, one from father),
  - paternal H12 (one copy from mother, two copies from father)

In other embodiments, other ploidy states, such as nullsomy (H00), uniparental disomy (H20 and H02), and tetrasomy (H04, H13, H22, H31 and H40), may be considered.

If there are no crossovers, each trisomy, whether the origin was mitosis, meiosis I, or meiosis II, would be one of the matched or unmatched trisomies. Due to crossovers, true trisomy is usually a combination of the two. First, a method to derive hypothesis likelihoods for simple hypotheses is described. Then a method to derive hypothesis likelihoods for composite hypotheses is described, combining individual SNP likelihood with crossovers.

LIK(D|H) for a Simple Hypothesis

In an embodiment, LIK(D|H) may be determined for simple hypotheses, as follows. For simple hypotheses H, LIK(H), the log likelihood of hypothesis H on a whole chromosome, may be calculated as the sum of log likelihoods of individual SNPs, assuming known or derived child fraction cf. In an embodiment it is possible to derive cf from the data.

$$\mathrm{LIK}(D|H) = \sum_i \mathrm{LIK}(D|H, cf, i)$$

This hypothesis does not assume any linkage between SNPs, and therefore does not utilize a joint distribution model.

In some embodiments, the Log Likelihood may be determined on a per SNP basis. On a particular SNP i, assuming fetal ploidy hypothesis H and percent fetal DNA cf, log likelihood of observed data D is defined as:

$$\mathrm{LIK}(D|H, i) = \log P(D|H, cf, i) = \log\left(\sum_{m,f,c} P(D|m, f, c, H, cf, i) \cdot P(c|m, f, H) \cdot P(m|i) \cdot P(f|i)\right)$$

where m are possible true mother genotypes , f are possible true father genotypes, where m,f $\in$ {AA,AB,BB}, and c are possible child genotypes given the hypothesis H. In particular, for monosomy c $\in \{A, B\}$, for disomy c $\in \{AA, AB, BB\}$, for trisomy c $\in \{AAA, AAB, ABB, BBB\}$.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 145

Genotype prior frequency: p(m|i) is the general prior probability of mother genotype m on SNP i, based on the known population frequency at SNP I, denoted pA_i. In particular

$$p(AA|pA_i) = (pA_i)^2,\ p(AB|pA_i) = 2(pA_i)*(1 - pA_i),\ p(BB|pA_i) = (1 - pA_i)^2$$

Father genotype probability, p(f|i), may be determined in an analogous fashion.

True child probability: $p(c|m,f,H)$ is the probability of getting true child genotype = c, given parents m, f, and assuming hypothesis H, which can be easily calculated. For example, for H11, H21 matched and H21 unmatched, p(c|m,f,H) is given below.

p(c|m,f,H)												
		H11			H21 matched				H21 unmatched			
m	f	AA	AB	BB	AAA	AAB	ABB	BBB	AAA	AAB	ABB	BBB
AA	AA	1	0	0	1	0	0	0	1	0	0	0
AB	AA	0.5	0.5	0	0.5	0	0.5	0	0	1	0	0
BB	AA	0	1	0	0	0	1	0	0	0	1	0
AA	AB	0.5	0.5	0	0.5	0.5	0	0	0.5	0.5	0	0
AB	AB	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0	0.5	0.5	0
BB	AB	0	0.5	0.5	0	0	0.5	0.5	0	0	0.5	0.5
AA	BB	0	1	0	0	1	0	0	0	1	0	0
AB	BB	0	0.5	0.5	0	0.5	0	0.5	0	0	1	0
BB	BB	0	0	1	0	0	0	1	0	0	0	1

Data likelihood: $P(D|m,f,c,H,i,cf)$ is the probability of given data D on SNP i, given true mother genotype m, true father genotype f, true child genotype c, hypothesis H and child fraction cf. It can be broken down into the probability of mother, father and child data as follows:

$$P(D|m,f,c,H,cf,i) = P(SM|m,i)P(M|m,i)P(SF|f,i)P(F|f,i)P(S|m,c,H,cf,i)$$

Mother SNP array data likelihood: Probability of mother SNP array genotype data m_i at SNP i compared to true genotype m, assuming SNP array genotypes are correct, is simply

$$P(M|m,i) = \begin{cases} 1 & m_i = m \\ 0 & m_i \neq m \end{cases}$$

Mother sequence data likelihood: the probability of the mother sequence data at SNP i, in the case of counts $S_i=(am_i,bm_i)$, with no extra noise or bias involved, is the binomial probability defined as $P(SM|m,i)=P_{X|m}(am_i)$ where $X|m \sim Binom(p_m(A), am_i+bm_i)$ with $p_m(A)$ defined as

m	AA	AB	BB	A	B	nocall
p(A)	1	0.5	0	1	0	0.5

Father data likelihood: a similar equation applies for father data likelihood. Note that it is possible to determine the child genotype without the parent data, especially father data. For example if no father genotype data F is available, one may just use $P(F|f,i) = 1$. If no father sequence data SF is available, one may just use P(SF|f,i)=1.

In some embodiments, the method involves building a joint distribution model for the expected allele counts at a plurality of polymorphic loci on the chromosome for each ploidy hypothesis; one method to accomplish such an end is described here. Free fetal DNA data likelihood: $P(S|m, c, H, cf, i)$ is the probability of free fetal DNA sequence data on SNP i, given true mother genotype m, true child genotype c, child copy number hypothesis H, and assuming child fraction cf. It is in fact the probability of sequence data S on SNP I, given the true probability of A content on SNP i $\mu(m, c, cf, H)$

$$P(S|m, c, H, cf, i) = P(S|\mu(m, c, cf, H), i)$$

For counts, where $S_i=(a_i,b_i)$, with no extra noise or bias in data involved,

$$P(S|\mu(m, c, cf, H), i) = P_x(a_i)$$

where X~Binom(p(A), a_i+b_i) with p(A)= $\mu(m, c, cf, H)$. In a more complex case where the exact alignment and (A,B) counts per SNP are not known, $P(S|\mu(m, c, cf, H), i)$ is a combination of integrated binomials.

True A content probability: $\mu(m, c, cf, H)$, the true probability of A content on SNP i in this mother/child mixture, assuming that true mother genotype = m, true child genotype = c, and overall child fraction = cf, is defined as

$$\mu(m, c, cf, H) = \frac{\#A(m) * (1 - cf) + \#A(c) * cf}{n_m * (1 - cf) + n_c * cf}$$

where #A(g) = number of A's in genotype g, $n_m = 2$ is somy of mother and n_c is ploidy of the child under hypothesis H (1 for monosomy, 2 for disomy, 3 for trisomy).

Using A Joint Distribution Model: LIK(D|H) for a Composite Hypothesis

In some embodiments, the method involves building a joint distribution model for the expected allele counts at the plurality of polymorphic loci on the chromosome for each ploidy hypothesis; one method to accomplish such an end is described here. In many cases, trisomy is usually not purely matched or unmatched, due to crossovers, so in this section results for

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 147

composite hypotheses H21 (maternal trisomy) and H12 (paternal trisomy) are derived, which combine matched and unmatched trisomy, accounting for possible crossovers.

In the case of trisomy, if there were no crossovers, trisomy would be simply matched or unmatched trisomy. Matched trisomy is where child inherits two copies of the identical chromosome segment from one parent. Unmatched trisomy is where child inherits one copy of each homologous chromosome segment from the parent. Due to crossovers, some segments of a chromosome may have matched trisomy, and other parts may have unmatched trisomy. Described in this section is how to build a joint distribution model for the heterozygosity rates for a set of alleles; that is, for the expected allele counts at a number of loci for one or more hypotheses.

Suppose that on SNP i, $LIK(D|Hm,i)$ is the fit for matched hypothesis H_m, and $LIK(D|Hu,i)$ is the fit for unmatched hypothesis H_u, and pc(i) = probability of crossover between SNPs i-1 andi. One may then calculate the full likelihood as:

$$LIK(D|H) = \sum_{E} LIK(D|E,1:N)$$

where $LIK(D|E,1:N)$ is the likelihood of ending in hypothesis E, for SNPs 1:N. E = hypothesis of the last SNP, $E \in (Hm, Hu)$. Recursively, one may calculate:

$$LIK(D|E,1:i) = LIK(D|E,i) + \log\left(\exp(LIK(D|E,1:i-1)) * (1 - pc(i)) + \exp(LIK(D|\sim E,1:i-1))\right)$$

where ~E is the hypothesis other than E (not E), where hypotheses considered are H_m and H_u. In particular, one may calculate the likelihood of 1:i SNPs, based on likelihood of 1 to (i-1) SNPs with either the same hypothesis and no crossover, or the opposite hypothesis and a crossover, multiplied by the likelihood of the SNP i

For SNP 1, i=1, $LIK(D|E,1:1) = LIK(D|E,1)$.

For SNP 2, i=2,

$$LIK(D|E,1:2) = LIK(D|E,2) + \log\left(\exp(LIK(D|E,1)) * (1 - pc(2)) + \exp(LIK(D|\sim E,1)) * pc(2)\right)$$

,

and so on for i=3:N.

In some embodiments, the child fraction may be determined. The child fraction may refer to the proportion of sequences in a mixture of DNA that originate from the child. In the context of non-invasive prenatal diagnosis, the child fraction may refer to the proportion of

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 148

sequences in the maternal plasma that originate from the fetus or the portion of the placenta with fetal genotype. It may refer to the child fraction in a sample of DNA that has been prepared from the maternal plasma, and may be enriched in fetal DNA. One purpose of determining the child fraction in a sample of DNA is for use in an algorithm that can make ploidy calls on the fetus, therefore, the child fraction could refer to whatever sample of DNA was analyzed by sequencing for the purpose of non-invasive prenatal diagnosis.

Some of the algorithms presented in this disclosure that are part of a method of non-invasive prenatal aneuploidy diagnosis assume a known child fraction, which may not always the case. In an embodiment, it is possible to find the most likely child fraction by maximizing the likelihood for disomy on selected chromosomes, with or without the presence of the parental data

In particular, suppose that LIK(D| H11, cf, chr) = log likelihood as described above, for the disomy hypothesis, and for child fraction cf on chromosome chr. For selected chromosomes in Cset (usually 1:16), assumed to be euploid, the full likelihood is:

$$\mathrm{LIK}(cf) = \sum_{i}(chr \in Cset) \square \mathrm{Lik}(D|\, H11, cf, chr) \square$$

The most likely child fraction ($\square cf \square^{\dagger} *$) is derived as $cf^{*} = \underset{cf\,\mathrm{LIK}(cf)}{\mathrm{argmax}}$.

It is possible to use any set of chromosomes. It is also possible to derive child fraction without assuming euploidy on the reference chromosomes. Using this method it is possible to determine the child fraction for any of the following situations: (1) one has array data on the parents and shotgun sequencing data on the maternal plasma; (2) one has array data on the parents and targeted sequencing data on the maternal plasma; (3) one has targeted sequencing data on both the parents and maternal plasma; (4) one has targeted sequencing data on both the mother and the maternal plasma fraction; (5) one has targeted sequencing data on the maternal plasma fraction; (6) other combinations of parental and child fraction measurements.

In some embodiments the informatics method may incorporate data dropouts; this may result in ploidy determinations of higher accuracy. Elsewhere in this disclosure it has been assumed that the probability of getting an A is a direct function of the true mother genotype, the true child genotype, the fraction of the child in the mixture, and the child copy number. It is also possible that mother or child alleles can drop out, for example instead of measuring true child AB in the mixture, it may be the case that only sequences mapping to allele A are measured. One may denote the parent dropout rate for genomic illumina data d_{pg}, parent

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 149

dropout rate for sequence data d_{ps} and child dropout rate for sequence data d_{cs}. In some embodiments, the mother dropout rate may be assumed to be zero, and child dropout rates are relatively low; in this case, the results are not severely affected by dropouts. In some embodiments the possibility of allele dropouts may be sufficiently large that they result in a significant effect of the predicted ploidy call. For such a case, allele dropouts have been incorporated into the algorithm here:

Parent SNP array data dropouts: For mother genomic data M, suppose that the genotype after the dropout is m_d, then

$$P(M|m, i) = \sum_{m_d} P(M|m_d, i) P(m_d|m)$$

where $P(M|m_d, i) = \begin{cases} 1 & m_i = m_d \\ 0 & m_i \neq m_d \end{cases}$ as before, and $P(m_d|m)$ is the likelihood of genotype m_d after the possible dropout given the true genotype m, defined as below, for dropout rate d

	md					
m	AA	AB	BB	A	B	nocall
AA	(1-d)^2	0	0	2d(1-d)	0	d^2
AB	0	(1-d)^2	0	d(1-d)	d(1-d)	d^2
BB	0	0	(1-d)^2	0	2d(1-d)	d^2

A similar equation applies for father SNP array data.

Parent sequence data dropouts: For mother sequence data SM

$$P(SM|m, i) = \sum_{m_d} P(X|m_d * am_i) P(m_d|m)$$

where $P(m_d|m)$ is defined as in previous section and $P(X|m_d * am_i)$ probability from a binomial distribution is defined as before in the parent data likelihood section. A similar equation applies to the paternal sequence data.

Free floating DNA sequence data dropout:

$$P(S|m, c, H, cf, i) = \sum_{m_d, c_d} P(S|\mu(m_d, c_d, cf, H), i) P(m_d|m) P(c_d|c)$$

where $P(S|\mu(m_d, c_d, cf, H), i)$ is as defined in the section on free floating data likelihood.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 150

In an embodiment, $p(m_d|m)$ is the probability of observed mother genotype m_d, given true mother genotype m, assuming dropout rate d_{ps}, and $p(c_d|c)$ is the probability of observed child genotype c_d, given true child genotype c, assuming dropout rate d_{cs}. If nA_T = number of A alleles in true genotype c, nA_D = number of A alleles in observed genotype c_d, where $nA_T \geq nA_D$, and similarly nB_T = number of B alleles in true genotype c, nB_D = number of B alleles in observed genotype c_d, where $nB_T \geq nB_D$ and d = dropout rate, then

$$p(c_d|c) = \binom{nA_T}{nA_D} * d^{nA_T - nA_D} * (1-d)^{nA_D} * \binom{nB_T}{nB_D} * d^{nB_T - nB_D} * (1-d)^{nB_D}$$

In an embodiment, the informatics method may incorporate random and consistent bias. In an ideal word there is no per SNP consistent sampling bias or random noise (in addition to the binomial distribution variation) in the number of sequence counts. In particular, on SNP i, for mother genotype m, true child genotype c and child fraction cf, and X = the number of A's in the set of (A+B) reads on SNP i, X acts like a X~Binomial(p, A+B), where p = $p(m, c, cf, H)$ = true probability of A content.

In an embodiment, the informatics method may incorporate random bias. As is often the case, suppose that there is a bias in the measurements, so that the probability of getting an A on this SNP is equal to q, which is a bit different than p as defined above. How much different p is from q depends on the accuracy of the measurement process and number of other factors and can be quantified by standard deviations of q away from p. In an embodiment, it is possible to model q as having a beta distribution, with parameters α, β depending on the mean of that distribution being centered at p, and some specified standard deviation s. In particular, this gives $X|q \sim Binom(q, D_i)$, where $q \sim Beta(\alpha, \beta)$. If we let $E(q) = p, V(q) = s^2$, and parameters α, β can be derived as $\alpha = pN, \beta = (1-p)N$, where $N = \frac{p(1-p)}{s^2} - 1$.

This is the definition of a beta-binomial distribution, where one is sampling from a binomial distribution with variable parameter q, where q follows a beta distribution with mean p. So, in a setup with no bias, on SNP i, the parent sequence data (SM) probability assuming true mother genotype (m), given mother sequence A count on SNP i (am_i) and mother sequence B count on SNP i (bm_i) may be calculated as:

$$P(SM|m,i)=P_{X|m}(am_i) \text{ where } X|m \sim Binom(p_m(A), am_i+bm_i)$$

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 151

Now, including random bias with standard deviation s, this becomes:

$$X|m \sim BetaBinom(p_m(A), am_i+bm_i, s)$$

In the case with no bias, the maternal plasma DNA sequence data (S) probability assuming true mother genotype (m), true child genotype (c), child fraction (cf), assuming child hypothesis H, given free floating DNA sequence A count on SNP i (a_i) and free floating sequence B count on SNP i (b_i) may be calculated as

$$P(S|m, c, cf, H, i) = P_X(a_i)$$

where $X \sim Binom(p(A), a_i+b_i)$ with $p(A)= p(m, c, cf, H)$.

In an embodiment, including random bias with standard deviation s, this becomes $X \sim BetaBinom(p(A), a_i+b_i, s)$, where the amount of extra variation is specified by the deviation parameter s, or equivalently N. The smaller the value of s (or the larger the value of N) the closer this distribution is to the regular binomial distribution. It is possible to estimate the amount of bias, *i.e.* estimate N above, from unambiguous contexts AA|AA, BB|BB, AA|BB, BB|AA and use estimated $\hat{N}$ in the above probability. Depending on the behavior of the data, N may be made to be a constant irrespective of the depth of read a_i+b_i, or a function of a_i+b_i, making bias smaller for larger depths of read.

In an embodiment, the informatics method may incorporate consistent per-SNP bias. Due to artifacts of the sequencing process, some SNPs may have consistently lower or higher counts irrespective of the true amount of A content. Suppose that SNP i consistently adds a bias of w_i percent to the number of A counts. In some embodiments, this bias can be estimated from the set of training data derived under same conditions, and added back in to the parent sequence data estimate as:

$$P(SM|m,i)=P_{X|m}(am_i) \text{ where } X|m \sim BetaBinom(p_m(A)+ w_i, am_i+bm_i, s)$$

and with the free floating DNA sequence data probability estimate as:

$$P(S|m, c, cf, H, i) = P_X(a_i) \text{ where } X \sim BetaBinom(p(A)+ w_i, a_i+b_i, s),$$

In some embodiments, the method may be written to specifically take into account additional noise, differential sample quality, differential SNP quality, and random sampling bias. An example of this is given here. This method has been shown to be particularly useful in the context of data generated using the massively multiplexed mini-PCR protocol, and was used in Experiments 7 through 13. The method involves several steps that each introduce different kind of noise and/or bias to the final model:

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 152

(1) Suppose the first sample that comprises a mixture of maternal and fetal DNA contains an original amount of DNA of size=N_0 molecules, usually in the range 1,000-40,000, where p = true %refs

(2) In the amplification using the universal ligation adaptors, assume that N_1 molecules are sampled; usually $N_1 \sim N_0/2$ molecules and random sampling bias is introduced due to sampling. The amplified sample may contain a number of molecules N_2 where $N_2 >> N_1$. Let X_1 represent the amount of reference loci (on per SNP basis) out of N_1 sampled molecules, with a variation in $p_1 = X_1/N_1$ that introduces random sampling bias throughout the rest of protocol. This sampling bias is included in the model by using a Beta-Binomial (BB) distribution instead of using a simple Binomial distribution model. Parameter N of the Beta-Binomial distribution may be estimated later on per sample basis from training data after adjusting for leakage and amplification bias, on SNPs with $0<p<1$. Leakage is the tendency for a SNP to be read incorrectly.

(3) The amplification step will amplify any allelic bias, thus amplification bias introduced due to possible uneven amplification. Suppose that one allele at a locus is amplified f times another allele at that locus is amplified g times, where $f=ge^b$, where b=0 indicates no bias. The bias parameter, b, is centered at 0, and indicates how much more or less the A allele get amplified as opposed to the B allele on a particular SNP. The parameter b may differ from SNP to SNP. Bias parameter b may be estimated on per SNP basis, for example from training data.

(4) The sequencing step involves sequencing a sample of amplified molecules. In this step there may be leakage, where leakage is the situation where a SNP is read incorrectly. Leakage may result from any number of problems, and may result in a SNP being read not as the correct allele A, but as another allele B found at that locus or as an allele C or D not typically found at that locus. Suppose the sequencing measures the sequence data of a number of DNA molecules from an amplified sample of size N_3, where $N_3 < N_2$. In some embodiments, N_3 may be in the range of 20,000 to 100,000; 100,000 to 500,000; 500,000 to 4,000,000; 4,000,000 to 20,000,000; or 20,000,000 to 100,000,000. Each molecule sampled has a probability p_g of being read correctly, in which case it will show up correctly as allele A. The sample will be incorrectly read as an allele unrelated to the original molecule with probability $1-p_g$, and will look like allele A with probability p_r, allele B with probability p_m or allele C or allele D with

probability p_o, where $p_r+p_m+p_o=1$. Parameters p_g, p_r, p_m, p_o are estimated on per SNP basis from the training data.

Different protocols may involve similar steps with variations in the molecular biology steps resulting in different amounts of random sampling, different levels of amplification and different leakage bias. The following model may be equally well applied to each of these cases. The model for the amount of DNA sampled, on per SNP basis, is given by:

$$X_3 \sim \text{BetaBinomial}(L(F(p,b),p_r,p_g), N*H(p,b))$$

where p = the true amount of reference DNA, b = per SNP bias, and as described above, p_g is the probability of a correct read, p_r is the probability of read being read incorrectly but serendipitously looking like the correct allele, in case of a bad read, as described above, and:

$$F(p,b)= pe^b/(pe^b+(1-p)),\ H(p,b) = (e^bp+(1-p))^2/e^b,\ L(p,p_r,p_g)=p*p_g+p_r*(1-p_g).$$

In some embodiments, the method uses a Beta-Binomial distribution instead of a simple binomial distribution; this takes care of the random sampling bias. Parameter N of the Beta-Binomial distribution is estimated on per sample basis on an as needed basis. Using bias correction F(p,b), H(p,b), instead of just p, takes care of the amplification bias. Parameter b of the bias is estimated on per SNP basis from training data ahead of time.

In some embodiments the method uses leakage correction $L(p,p_r,p_g)$, instead of just p; this takes care of the leakage bias, *i.e.* varying SNP and sample quality. In some embodiments, parameters p_g, p_r, p_o are estimated on per SNP basis from the training data ahead of time. In some embodiments, the parameters p_g, p_r, p_o may be updated with the current sample on the go, to account for varying sample quality.

The model described herein is quite general and can account for both differential sample quality and differential SNP quality. Different samples and SNPs are treated differently, as exemplified by the fact that some embodiments use Beta-Binomial distributions whose mean and variance are a function of the original amount of DNA, as well as sample and SNP quality.

Platform modeling

Consider a single SNP where the expected allele ratio present in the plasma is r (based on the maternal and fetal genotypes). The expected allele ratio is defined as the expected fraction of A alleles in the combined maternal and fetal DNA. For maternal genotype g_m and

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 154

child genotype g_c, the expected allele ratio is given by equation 1, assuming that the genotypes are represented as allele ratios as well.

$$r = fg_c + (1 - f)g_m \qquad (1)$$

The observation at the SNP consists of the number of mapped reads with each allele present, n_a and n_b, which sum to the depth of read d. Assume that thresholds have already been applied to the mapping probabilities and phred scores such that the mappings and allele observations can be considered correct. A phred score is a numerical measure that relates to the probability that a particular measurement at a particular base is wrong. In an embodiment, where the base has been measured by sequencing, the phred score may be calculated from the ratio of the dye intensity corresponding to the called base to the dye intensity of the other bases. The simplest model for the observation likelihood is a binomial distribution which assumes that each of the d reads is drawn independently from a large pool that has allele ratio r. Equation 2 describes this model.

$$P(n_a,n_b|r) = p_{bino}(n_a; n_a + n_b, r) = \binom{n_a + n_b}{n_a} r^{n_a}(1 - r)^{n_b} \qquad (2)$$

The binomial model can be extended in a number of ways. When the maternal and fetal genotypes are either all A or all B, the expected allele ratio in plasma will be 0 or 1, and the binomial probability will not be well-defined. In practice, unexpected alleles are sometimes observed in practice. In an embodiment, it is possible to use a corrected allele ratio $\hat{r} = 1/(n_a + n_b)$ to allow a small number of the unexpected allele. In an embodiment, it is possible to use training data to model the rate of the unexpected allele appearing on each SNP, and use this model to correct the expected allele ratio. When the expected allele ratio is not 0 or 1, the observed allele ratio may not converge with a sufficiently high depth of read to the expected allele ratio due to amplification bias or other phenomena. The allele ratio can then be modeled as a beta distribution centered at the expected allele ratio, leading to a beta-binomial distribution for $P(n_a, n_b|r)$ which has higher variance than the binomial.

The platform model for the response at a single SNP will be defined as $F(a, b, g_c, g_m, f)$ (3), or the probability of observing $n_a = a$ and $n_b = b$ given the maternal and fetal genotypes, which also depends on the fetal fraction through equation 1. The functional form of F may be a binomial distribution, beta-binomial distribution, or similar functions as discussed above.

$$F(a, b, g_c, g_m, f) = P(n_a = a, n_b = b|g_c, g_m, f) = P(n_a = a, n_b = b|r(g_c, g_m, f)) \qquad (3)$$

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 155

In an embodiment, the child fraction may be determined as follows. A maximum likelihood estimate of the fetal fraction f for a prenatal test may be derived without the use of paternal information. This may be relevant where the paternal genetic data is not available, for example where the father of record is not actually the genetic father of the fetus. The fetal fraction is estimated from the set of SNPs where the maternal genotype is 0 or 1, resulting in a set of only two possible fetal genotypes. Define S_0 as the set of SNPs with maternal genotype 0 and S_1 as the set of SNPs with maternal genotype 1. The possible fetal genotypes on S_0 are 0 and 0.5, resulting in a set of possible allele ratios $R_0(f) = \{0,f/2\}$. Similarly, $R_1(f) = \{1-f/2, 1\}$. This method can be trivially extended to include SNPs where maternal genotype is 0.5, but these SNPs will be less informative due to the larger set of possible allele ratios.

Define N_{a0} and N_{b0} as the vectors formed by n_{as} and n_{bs} for SNPs s in S_0, and N_{a1} and N_{b1} similarly for S_1. The maximum likelihood estimate $\hat{f}$ of f is defined by equation 4.

$$\hat{f} = \arg\max_f P(N_{a0}, N_{b0}|f)\, P(N_{a1}, N_{b1}|f) \qquad (4)$$

Assuming that the allele counts at each SNP are independent conditioned on the SNP's plasma allele ratio, the probabilities can be expressed as products over the SNPs in each set (5).

$$P(N_{a0}, N_{b0}|f) = \prod_{s \in S_0} P(n_{as}, n_{bs}|f) \qquad (5)$$

$$P(N_{a1}, N_{b1}|f) = \prod_{s \in S_1} P(n_{as}, n_{bs}|f)$$

The dependence on f is through the sets of possible allele ratios $R_0(f)$ and $R_1(f)$. The SNP probability $P(n_{as}, n_{bs}|f)$ can be approximated by assuming the maximum likelihood genotype conditioned on f. At reasonably high fetal fraction and depth of read, the selection of the maximum likelihood genotype will be high confidence. For example, at fetal fraction of 10 percent and depth of read of 1000, consider a SNP where the mother has genotype zero. The expected allele ratios are 0 and 5 percent, which will be easily distinguishable at sufficiently high depth of read. Substitution of the estimated child genotype into equation 5 results in the complete equation (6) for the fetal fraction estimate.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 156

$$\hat{f} = \arg\max_f \left(\max_{r \in R_0(f)} \left[P(n_a, n_b | r) \right] \prod_{s \in S_1} \left[\left(\max_{r \in R_1(f)} \left[P(n_a, n_b | r) \right] \right] \right]$$

(6)

The fetal fraction must be in the range [0, 1] and so the optimization can be easily implemented by a constrained one-dimensional search.

In the presence of low depth of read or high noise level, it may be preferable not to assume the maximum likelihood genotype, which may result in artificially high confidences. Another method would be to sum over the possible genotypes at each SNP, resulting in the following expression (7) for $P(n_a, n_b|f)$ for a SNP in S_0. The prior probability P(r) could be assumed uniform over $R_0(f)$, or could be based on population frequencies. The extension to group S_1 is trivial.

$$P(n_a, n_b|f) = \sum_{r \in R_0(f)} \left[P(n_a, n_a | r) \right] P(r) \qquad (7)$$

In some embodiments the probabilities may be derived as follows. A confidence can be calculated from the data likelihoods of the two hypotheses H_t and H_f. The likelihood of each hypothesis is derived based on the response model, the estimated fetal fraction, the mother genotypes, allele population frequencies, and the plasma allele counts.

Define the following notation:

G_m, G_c	true maternal and child genotypes
G_{af}, G_{tf}	true genotypes of alleged father and of true father
$G(g_c, g_m, g_{tf}) = P(G_c = g_c \| G_m = g_m, G_{tf} = g_{tf})$	inheritance probabilities
$P(g) = P(G_{tf} = g)$	population frequency of genotype g at particular SNP

Assuming that the observation at each SNP is independent conditioned on the plasma allele ratio, the likelihood of a paternity hypothesis is the product of the likelihoods on the SNPs. The following equations derive the likelihood for a single SNP. Equation 8 is a general expression for the likelihood of any hypothesis h, which will then be broken down into the specific cases of H_t and H_f.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 157

$$P(n_a, n_b|h, G_m, G_{tf}, f) =$$

$$\sum_{g_c \in (0,0.5,1)} [P(\square]n_a, n_b|G_c = g_c, G_m, G_{tf}, h, f)P(G_c = g_c, G_m, G_{tf}, h, f)$$

$$= \sum_{g_c \in (0,0.5,1)} [P(\square]n_a, n_b|G_c = g_c, G_m, f)P(G_c = g_c|G_m, G_{tf}, h)$$

$$= \sum_{g_c \in (0,0.5,1)} [F(\square]n_a, n_b, g_c, g_m, f)P(G_c = g_c|G_m, G_{tf}, h) \quad (8)$$

In the case of H_t, the alleged father is the true father and the fetal genotypes are inherited from the maternal genotypes and alleged father genotypes according to equation 9.

$$P(n_a, n_b|H_t, G_m, G_{tf}, f) = \sum_{g_c \in (0,0.5,1)} [F(\square]n_a, n_b, g_c, g_m, f)P(G_c = g_c|G_m, G_{tf}, H_t) \quad (9)$$

$$= \sum_{g_c \in (0,0.5,1)} [F(\square]n_a, n_b, g_c, g_m, f)G(g_c, G_m, G_{tf})$$

In the case of H_f, the alleged father is not the true father. The best estimate of the true father genotypes are given by the population frequencies at each SNP. Thus, the probabilities of child genotypes are determined by the known mother genotypes and the population frequencies, as in equation 10.

$$P(n_a, n_b|H_t, G_m, G_{tf}, f) = \sum_{g_c \in (0,0.5,1)} [F(\square]n_a, n_b, g_c, g_m, f)P(G_c = g_c|G_m, G_{tf}, H_f)$$

$$= \sum_{g_c \in (0,0.5,1)} [F(\square]n_a, n_b, g_c, g_m, f)P(G_c = g_c|G_m)$$

$$= \sum_{g_c \in (0,0.5,1)} [\sum_{g_{tf} \in (0,0.5,1)} \square F(\square]n_a, n_b, g_c, g_m, f \Big) P(G_c = g_c|G_m, G_{tf} = g_{tf})P(G_{tf} = g_{tf})$$

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 158

$$= \sum_{g_c \in (0,0.5,1)} \left[\sum_{g_{tf} \in (0,0.5,1)} \square P(\square | n_a, n_b, g_c, g_m, f) \right) G(g_c, G_m, g_{tf}) P(g_{tf})$$

The confidence C_p on correct paternity is calculated from the product over SNPs of the two likelihoods using Bayes rule (11).

$$Cp = \frac{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f)}{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f) + \prod_s P(n_{as}, n_{bs} | H_f, G_{ms}, G_{tf}, f)} \quad (11)$$

Maximum Likelihood Model using Percent Fetal Fraction

Determining the ploidy status of a fetus by measuring the free floating DNA contained in maternal serum, or by measuring the genotypic material in any mixed sample, is a non-trivial exercise. There are a number of methods, for example, performing a read count analysis where the presumption is that if the fetus is trisomic at a particular chromosome, then the overall amount of DNA from that chromosome found in the maternal blood will be elevated with respect to a reference chromosome. One way to detect trisomy in such fetuses is to normalize the amount of DNA expected for each chromosome, for example, according to the number of SNPs in the analysis set that correspond to a given chromosome, or according to the number of uniquely mappable portions of the chromosome. Once the measurements have been normalized, any chromosomes for which the amount of DNA measured exceeds a certain threshold are determined to be trisomic. This approach is described in Fan, et al. PNAS, 2008; 105(42); pp. 16266–16271, and also in Chiu et al. BMJ 2011;342:c7401. In the Chiu et al. paper, the normalization was accomplished by calculating a Z score as follows:

> Z score for percentage chromosome 21 in test case = ((percentage chromosome 21 in test case) − (mean percentage chromosome 21 in reference controls)) / (standard deviation of percentage chromosome 21 in reference controls).

These methods determine the ploidy status of the fetus using a single hypothesis rejection method. However, they suffer from some significant shortcomings. Since these methods for

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 159

determining ploidy in the fetus are invariant according to the percentage of fetal DNA in the sample, they use one cut off value; the result of this is that the accuracies of the determinations are not optimal, and those cases where the percentage of fetal DNA in the mixture are relatively low will suffer the worst accuracies.

In an embodiment, a method of the present disclosure is used to determine the ploidy state of the fetus involves taking into account the fraction of fetal DNA in the sample. In another embodiment of the present disclosure, the method involves the use of maximum likelihood estimations. In an embodiment, a method of the present disclosure involves calculating the percent of DNA in a sample that is fetal or placental in origin. In an embodiment, the threshold for calling aneuploidy is adaptively adjusted based on the calculated percent fetal DNA. In some embodiments, the method for estimating the percentage of DNA that is of fetal origin in a mixture of DNA, comprises obtaining a mixed sample that comprises genetic material from the mother, and genetic material from the fetus, obtaining a genetic sample from the father of the fetus, measuring the DNA in the mixed sample, measuring the DNA in the father sample, and calculating the percentage of DNA that is of fetal origin in the mixed sample using the DNA measurements of the mixed sample, and of the father sample.

In an embodiment of the present disclosure, the fraction of fetal DNA, or the percentage of fetal DNA in the mixture can be measured. In some embodiments the fraction can be calculated using only the genotyping measurements made on the maternal plasma sample itself, which is a mixture of fetal and maternal DNA. In some embodiments the fraction may be calculated also using the measured or otherwise known genotype of the mother and/or the measured or otherwise known genotype of the father. In some embodiments the percent fetal DNA may be calculated using the measurements made on the mixture of maternal and fetal DNA along with the knowledge of the parental contexts. In an embodiment, the fraction of fetal DNA may be calculated using population frequencies to adjust the model on the probability on particular allele measurements.

In an embodiment of the present disclosure, a confidence may be calculated on the accuracy of the determination of the ploidy state of the fetus. In an embodiment, the confidence of the hypothesis of greatest likelihood (H_{major}) may be calculated as $(1- H_{major}) / \Sigma(\text{all } H)$. It is possible to determine the confidence of a hypothesis if the distributions of all of the hypotheses are known. It is possible to determine the distribution of all of the hypotheses if

the parental genotype information is known. It is possible to calculate a confidence of the ploidy determination if the knowledge of the expected distribution of data for the euploid fetus and the expected distribution of data for the aneuploid fetus are known. It is possible to calculate these expected distributions if the parental genotype data are known. In an embodiment one may use the knowledge of the distribution of a test statistic around a normal hypothesis and around an abnormal hypothesis to determine both the reliability of the call as well as refine the threshold to make a more reliable call. This is particularly useful when the amount and/or percent of fetal DNA in the mixture is low. It will help to avoid the situation where a fetus that is actually aneuploid is found to be euploid because a test statistic, such as the Z statistic does not exceed a threshold that is made based on a threshold that is optimized for the case where there is a higher percent fetal DNA.

In an embodiment, a method disclosed herein can be used to determine a fetal aneuploidy by determining the number of copies of maternal and fetal target chromosomes in a mixture of maternal and fetal genetic material. This method may entail obtaining maternal tissue comprising both maternal and fetal genetic material; in some embodiments this maternal tissue may be maternal plasma or a tissue isolated from maternal blood. This method may also entail obtaining a mixture of maternal and fetal genetic material from said maternal tissue by processing the aforementioned maternal tissue. This method may entail distributing the genetic material obtained into a plurality of reaction samples, to randomly provide individual reaction samples that comprise a target sequence from a target chromosome and individual reaction samples that do not comprise a target sequence from a target chromosome, for example, performing high throughput sequencing on the sample. This method may entail analyzing the target sequences of genetic material present or absent in said individual reaction samples to provide a first number of binary results representing presence or absence of a presumably euploid fetal chromosome in the reaction samples and a second number of binary results representing presence or absence of a possibly aneuploid fetal chromosome in the reaction samples. Either of the number of binary results may be calculated, for example, by way of an informatics technique that counts sequence reads that map to a particular chromosome, to a particular region of a chromosome, to a particular locus or set of loci. This method may involve normalizing the number of binary events based on the chromosome length, the length of the region of the chromosome, or the number of loci in the set. This method may entail calculating an expected distribution of the number of binary results for a presumably euploid

fetal chromosome in the reaction samples using the first number. This method may entail calculating an expected distribution of the number of binary results for a presumably aneuploid fetal chromosome in the reaction samples using the first number and an estimated fraction of fetal DNA found in the mixture, for example, by multiplying the expected read count distribution of the number of binary results for a presumably euploid fetal chromosome by (1 + n/2) where n is the estimated fetal fraction. In some embodiments, the sequence reads may be treated at probabilistic mappings rather than binary results; this method would yield higher accuracies, but require more computing power. The fetal fraction may be estimated by a plurality of methods, some of which are described elsewhere in this disclosure. This method may involve using a maximum likelihood approach to determine whether the second number corresponds to the possibly aneuploid fetal chromosome being euploid or being aneuploid. This method may involve calling the ploidy status of the fetus to be the ploidy state that corresponds to the hypothesis with the maximum likelihood of being correct given the measured data.

Note that the use of a maximum likelihood model may be used to increase the accuracy of any method that determines the ploidy state of a fetus. Similarly, a confidence maybe calculated for any method that determines the ploidy state of the fetus. The use of a maximum likelihood model would result in an improvement of the accuracy of any method where the ploidy determination is made using a single hypothesis rejection technique. A maximum likelihood model may be used for any method where a likelihood distribution can be calculated for both the normal and abnormal cases. The use of a maximum likelihood model implies the ability to calculate a confidence for a ploidy call.

Further Discussion of the Method

In an embodiment, a method disclosed herein utilizes a quantitative measure of the number of independent observations of each allele at a polymorphic locus, where this does not involve calculating the ratio of the alleles. This is different from methods, such as some microarray based methods, which provide information about the ratio of two alleles at a locus but do not quantify the number of independent observations of either allele. Some methods known in the art can provide quantitative information regarding the number of independent observations, but the calculations leading to the ploidy determination utilize only the allele ratios, and do not utilize the quantitative information. To illustrate the importance of retaining

information about the number of independent observations consider the sample locus with two alleles, A and B. In a first experiment twenty A alleles and twenty B alleles are observed, in a second experiment 200 A alleles and 200 B alleles are observed. In both experiments the ratio (A/(A+B)) is equal to 0.5, however the second experiment conveys more information than the first about the certainty of the frequency of the A or B allele. The instant method, rather than utilizing the allele ratios, uses the quantitative data to more accurately model the most likely allele frequencies at each polymorphic locus.

In an embodiment, the instant methods build a genetic model for aggregating the measurements from multiple polymorphic loci to better distinguish trisomy from disomy and also to determine the type of trisomy. Additionally, the instant method incorporates genetic linkage information to enhance the accuracy of the method. This is in contrast to some methods known in the art where allele ratios are averaged across all polymorphic loci on a chromosome. The method disclosed herein explicitly models the allele frequency distributions expected in disomy as well as and trisomy resulting from nondisjunction during meiosis I, nondisjunction during meiosis II, and nondisjunction during mitosis early in fetal development. To illustrate why this is important, if there were no crossovers nondisjunction during meiosis I would result a trisomy in which two different homologs were inherited from one parent; nondisjunction during meiosis II or during mitosis early in fetal development would result in two copies of the same homolog from one parent. Each scenario results in different expected allele frequencies at each polymorphic locus and also at all physically linked loci (*i.e.* loci on the same chromosome) considered jointly. Crossovers, which result in the exchange of genetic material between homologs, make the inheritance pattern more complex, but the instant method accommodates for this by using genetic linkage information, *i.e.* recombination rate information and the physical distance between loci. To better distinguish between meiosis I nondisjunction and meiosis II or mitotic nondisjunction the instant method incorporates into the model an increasing probability of crossover as the distance from the centromere increases. Meiosis II and mitotic nondisjunction can distinguished by the fact that mitotic nondisjunction typically results in identical or nearly identical copies of one homolog while the two homologs present following a meiosis II nondisjunction event often differ due to one or more crossovers during gametogenesis.

In an embodiment, a method of the present disclosure may not determine the haplotypes of the parents if disomy is assumed. In an embodiment, in case of trisomy, the

instant method can make a determination about the haplotypes of one or both parents by using the fact that plasma takes two copies from one parent, and parent phase information can be determined by noting which two copies have been inherited from the parent in question. In particular, a child can inherit either two of the same copies of the parent (matched trisomy) or both copies of the parent (unmatched trisomy). At each SNP one can calculate the likelihood of the matched trisomy and of the unmatched trisomy. A ploidy calling method that does not use the linkage model accounting for crossovers would calculate the overall likelihood of the trisomy as a simple weighted average of the matched and unmatched trisomies over all chromosomes. However, due to the biological mechanisms that result in disjunction error and crossing over, trisomy can change from matched to unmatched (and vice versa) on a chromosome only if a crossover occurs. The instant method probabilistically takes into account the likelihood of crossover, resulting in ploidy calls that are of greater accuracy than those methods that do not.

In an embodiment, a reference chromosome is used to determine the child fraction and noise level amount or probability distribution. In an embodiment, the child fraction, noise level, and/or probability distribution is determined using only the genetic information available from the chromosome whose ploidy state is being determined. The instant method works without the reference chromosome, as well as without fixing the particular child fraction or noise level. This is a significant improvement and point of differentiation from methods known in the art where genetic data from a reference chromosome is necessary to calibrate the child fraction and chromosome behavior.

In an embodiment where a reference chromosome is not needed to determine the fetal fraction, determining the hypothesis is done as follows:

$$H^* = \underset{H}{\operatorname{argmax}}\ LIK(D|H) * \text{priorprob}(H)$$

With the algorithm with reference chromosome, one typically assumes that the reference chromosome is a disomy, and then one may either (a) fix the most likely child fraction and random noise level N based on this assumption and reference chromosome data :

$$[cfr^*, N^*] = \underset{cfr, N}{\operatorname{argmax}}\ LIK(D(ref.chrom)|H11, cfr, N)$$

And then reduce

$$LIK(D|H) = LIK(D|H, cfr^*, N^*)$$

or (b) estimate the child fraction and noise level distribution based on this assumption and reference chromosome data. In particular, one would not fix just one value for cfr and N, but assign probability p(cfr, N) for the wider range of possible cfr, N values:

$$p(cfr, N) \sim LIK(D(ref.chrom)|H11, cfr, N) * priorprob(cfr, N)$$

where priorprob(cfr, N) is the prior probability of particular child fraction and noise level, determined by prior knowledge and experiments. If desired, just uniform over the range of cfr, N. One may then write:

$$LIK(D|H) = \sum_{cfr,N} LIK(D|H, cfr, N) * p(cfr, N)$$

Both methods above give good results.

Note that in some instances using a reference chromosome is not desirable, possible or feasible. In such a case, it is possible to derive the best ploidy call for each chromosome separately. In particular:

$$LIK(D|H) = \sum_{cfr,N} LIK(D|H, cfr, N) * p(cfr, N|H)$$

$p(cfr, N|H)$ may be determined as above, for each chromosome separately, assuming hypothesis H, not just for the reference chromosome assuming disomy. It is possible, using this method, to keep both noise and child fraction parameters fixed, fix either of the parameters, or keep both parameters in probabilistic form for each chromosome and each hypothesis.

Measurements of DNA are noisy and/or error prone, especially measurements where the amount of DNA is small, or where the DNA is mixed with contaminating DNA. This noise results in less accurate genotypic data, and less accurate ploidy calls. In some embodiments, platform modeling or some other method of noise modeling may be used to counter the deleterious effects of noise on the ploidy determination. The instant method uses a joint model of both channels, which accounts for the random noise due to the amount of input DNA, DNA quality, and/or protocol quality.

This is in contrast to some methods known in the art where the ploidy determinations are made using the ratio of allele intensities at a locus. This method precludes accurate SNP noise modeling. In particular, errors in the measurements typically do not specifically depend on the measured channel intensity ratio, which reduces the model to using one-dimensional

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 165

information. Accurate modeling of noise, channel quality and channel interaction requires a two-dimensional joint model, which can *not* be modeled using allele ratios.

In particular, projecting two channel information to the ratio r where f(x,y) is r = x/y, does not lend itself to accurate channel noise and bias modeling. Noise on a particular SNP is not a function of the ratio, *i.e.* noise(x,y) $\neq$ f(x,y) but is in fact a joint function of both channels. For example, in the binomial model, noise of the measured ratio has a variance of r(1-r)/(x+y) which is not a function purely of r. In such a model, where any channel bias or noise is included, suppose that on SNP i, the observed channel X value is $x=a_iX+b_i$, where X is the true channel value, b_i is the extra channel bias and random noise. Similarly, suppose that $y=c_iY+d_i$. The observed ratio r=x/y can not accurately predict the true ratio X/Y or model the leftover noise, since (aiX+bi)/(ciY+di) is not a function of X/Y.

The method disclosed herein describes an effective way to model noise and bias using joint binomial distributions of all of the measurement channels individually. Relevant equations may be found elsewhere in the document in sections which speaks of per SNP consistent bias, P(good) and P(ref|bad), P(mut|bad) which effectively adjust SNP behavior. In an embodiment, a method of the present disclosure uses a BetaBinomial distribution, which avoids the limiting practice of relying on the allele ratios only, but instead models the behavior based on both channel counts.

In an embodiment, a method disclosed herein can call the ploidy of a gestating fetus from genetic data found in maternal plasma by using all available measurements. In an embodiment, a method disclosed herein can call the ploidy of a gestating fetus from genetic data found in maternal plasma by using the measurements from only a subset of parental contexts. Some methods known in the art only use measured genetic data where the parental context is from the AA|BB context, that is, where the parents are both homozygous at a given locus, but for a different allele. One problem with this method is that a small proportion of polymorphic loci are from the AA|BB context, typically less than 10%. In an embodiment of a method disclosed herein, the method does not use genetic measurements of the maternal plasma made at loci where the parental context is AA|BB. In an embodiment, the instant method uses plasma measurements for only those polymorphic loci with the AA|AB, AB|AA, and AB|AB parental context.

Some methods known in the art involve averaging allele ratios from SNPs in the AA|BB context, where both parent genotypes are present, and claim to determine the ploidy

calls from the average allele ratio on these SNPs. This method suffers from significant inaccuracy due differential SNP behavior. Note that this method assumes that have both parent genotypes are known. In contrast, in some embodiments, the instant method uses a joint channel distribution model that does not assume the presence of either of the parents, and does not assume the uniform SNP behavior. In some embodiments, the instant method accounts for the different SNP behavior/weighing. In some embodiments, the instant method does not require the knowledge of one or both parental genotypes. An example of how the instant method may accomplish this follows:

In some embodiments, the log likelihood of a hypothesis may be determined on a per SNP basis. On a particular SNP i, assuming fetal ploidy hypothesis H and percent fetal DNA cf, the log likelihood of observed data D is defined as:

$$\mathrm{LIK}(D|H,i) = \log P(D|H,cf,i) = \log\left(\sum_{m,f,c} P(D|m,f,c,H,cf,i)P(c|m,f,H)P(m|i)P(f|i)\right)$$

where m are possible true mother genotypes, f are possible true father genotypes, where m,f $\in$ {AA,AB,BB}, and where c are possible child genotypes given the hypothesis H. In particular, for monosomy c $\{A,B\}$, for disomy c $\in \{AA,AB,BB\}$, for trisomy c $\in \{AAA,AAB,ABB,BBB\}$. Note that including parental genotypic data typically results in more accurate ploidy determinations, however, parental genotypic data is not necessary for the instant method to work well.

Some methods known in the art involve averaging allele ratios from SNPs where the mother is homozygous but a different allele is measured in the plasma (either AA|AB or AA|BB contexts), and claim to determine the ploidy calls from the average allele ratio on these SNPs. This method is intended for cases where the paternal genotype is not available. Note that it is questionable how accurately one can claim that plasma is heterozygous on a particular SNP without the presence of homozygous and opposite father BB: for cases with low child fraction, what looks like presence of B allele could be just presence of noise; additionally, what looks like no B present could be simple allele drop out of the fetal measurements. Even in a case where one can actually determine heterozygosity of the plasma, this method will not be able to distinguish paternal trisomies. In particular, for SNPs where mother is AA, and where some B is measured in the plasma, if the father is GG, the resulting child genotype is AGG, resulting in an average ratio of 33% A (for child fraction=100%). But in the case where the

father is AG, the resulting child genotype could be AGG for matched trisomy, contributing to the 33% A ratio, or AAG for unmatched trisomy, drawing the average ratio more toward 66% A. Given that many trisomies are on chromosomes with crossovers, the overall chromosome can have anywhere between no unmatched trisomy and all unmatched trisomy, this ratio can vary anywhere between 33-66%. For a plain disomy, the ratio should be around 50%. Without the use of a linkage model or an accurate error model of the average, this method would miss many cases of paternal trisomy. In contrast, the method disclosed herein assigns parental genotype probabilities for each parental genotypic candidate, based on available genotypic information and population frequency, and does not explicitly require parental genotypes. Additionally, the method disclosed herein is able to detect trisomy even in the absence or presence of parent genotypic data, and can compensate by identifying the points of possible crossovers from matched to unmatched trisomy using a linkage model.

Some methods known in the art claim a method for averaging allele ratios from SNPs where neither the maternal or paternal genotype is known, and for determining the ploidy calls from average ratio on these SNPs. However, a method to accomplish these ends is not disclosed. The method disclosed herein is able to make accurate ploidy calls in such a situation, and the reduction to practice is disclosed elsewhere in this document, using a joint probability maximum likelihood method and optionally utilizes SNP noise and bias models, as well as a linkage model.

Some methods known in the art involve averaging allele ratios and claim to determine the ploidy calls from the average allele ratio at one or a few SNPs. However, such methods do not utilize the concept of linkage. The methods disclosed herein do not suffer from these drawbacks.

Using Sequence Length as a Prior to Determine the Origin of DNA

It has been reported that the distribution of length of sequences differ for maternal and fetal DNA, with fetal generally being shorter. In an embodiment of the present disclosure, it is possible to use previous knowledge in the form of empirical data, and construct prior distribution for expected length of both mother(P(X| maternal)) and fetal DNA (P(X| fetal)). Given new unidentified DNA sequence of length x, it is possible to assign a probability that a given sequence of DNA is either maternal or fetal DNA, based on prior likelihood of x given either maternal or fetal. In particular if P(x|maternal) > P(x|fetal), then the DNA sequence can

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 168

be classified as maternal, with P(x|maternal) = P(x|maternal)/[(P(x|maternal) + P(x| fetal)], and if p(x|maternal) < p(x|fetal), then the DNA sequence can be classified as fetal, P(x| fetal) = P(x| fetal)/[(P(x|maternal) + P(x| fetal)]. In an embodiment of the present disclosure, a distributions of maternal and fetal sequence lengths can be determined that is specific for that sample by considering the sequences that can be assigned as maternal or fetal with high probability, and then that sample specific distribution can be used as the expected size distribution for that sample.

Variable Read Depth to Minimize Sequencing Cost

In many clinical trials concerning a diagnostic, for example, in Chiu et al. BMJ 2011;342:c7401, a protocol with a number of parameters is set, and then the same protocol is executed with the same parameters for each of the patients in the trial. In the case of determining the ploidy status of a fetus gestating in a mother using sequencing as a method to measure genetic material one pertinent parameter is the number of reads. The number of reads may refer to the number of actual reads, the number of intended reads, fractional lanes, full lanes, or full flow cells on a sequencer. In these studies, the number of reads is typically set at a level that will ensure that all or nearly all of the samples achieve the desired level of accuracy. Sequencing is currently an expensive technology, a cost of roughly $200 per 5 mappable million reads, and while the price is dropping, any method which allows a sequencing based diagnostic to operate at a similar level of accuracy but with fewer reads will necessarily save a considerable amount of money.

The accuracy of a ploidy determination is typically dependent on a number of factors, including the number of reads and the fraction of fetal DNA in the mixture. The accuracy is typically higher when the fraction of fetal DNA in the mixture is higher. At the same time, the accuracy is typically higher if the number of reads is greater. It is possible to have a situation with two cases where the ploidy state is determined with comparable accuracies wherein the first case has a lower fraction of fetal DNA in the mixture than the second, and more reads were sequenced in the first case than the second. It is possible to use the estimated fraction of fetal DNA in the mixture as a guide in determining the number of reads necessary to achieve a given level of accuracy.

In an embodiment of the present disclosure, a set of samples can be run where different samples in the set are sequenced to different reads depths, wherein the number of reads run on

each of the samples is chosen to achieve a given level of accuracy given the calculated fraction of fetal DNA in each mixture. In an embodiment of the present disclosure, this may entail making a measurement of the mixed sample to determine the fraction of fetal DNA in the mixture; this estimation of the fetal fraction may be done with sequencing, it may be done with TAQMAN, it may be done with qPCR, it may be done with SNP arrays, it may be done with any method that can distinguish different alleles at a given loci. The need for a fetal fraction estimate may be eliminated by including hypotheses that cover all or a selected set of fetal fractions in the set of hypotheses that are considered when comparing to the actual measured data. After the fraction fetal DNA in the mixture has been determined, the number of sequences to be read for each sample may be determined.

In an embodiment of the present disclosure, 100 pregnant women visit their respective OB's, and their blood is drawn into blood tubes with an anti-lysant and/or something to inactivate DNAase. They each take home a kit for the father of their gestating fetus who gives a saliva sample. Both sets of genetic materials for all 100 couples are sent back to the laboratory, where the mother blood is spun down and the buffy coat is isolated, as well as the plasma. The plasma comprises a mixture of maternal DNA as well as placentally derived DNA. The maternal buffy coat and the paternal blood is genotyped using a SNP array, and the DNA in the maternal plasma samples are targeted with SURESELECT hybridization probes. The DNA that was pulled down with the probes is used to generate 100 tagged libraries, one for each of the maternal samples, where each sample is tagged with a different tag. A fraction from each library is withdrawn, each of those fractions are mixed together and added to two lanes of a ILLUMINA HISEQ DNA sequencer in a multiplexed fashion, wherein each lane resulted in approximately 50 million mappable reads, resulting in approximately 100 million mappable reads on the 100 multiplexed mixtures, or approximately 1 million reads per sample. The sequence reads were used to determine the fraction of fetal DNA in each mixture. 50 of the samples had more than 15% fetal DNA in the mixture, and the 1 million reads were sufficient to determine the ploidy status of the fetuses with a 99.9% confidence.

Of the remaining mixtures, 25 had between 10 and 15% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 2 million reads for each sample. The two sets of sequence data for each of the mixture with between 10 and 15% fetal DNA were added

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 170

together, and the resulting 3 million reads per sample which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

Of the remaining mixtures, 13 had between 6 and 10% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 4 million reads for each sample. The two sets of sequence data for each of the mixture with between 6 and 10% fetal DNA were added together, and the resulting 5 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

Of the remaining mixtures, 8 had between 4 and 6% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 6 million reads for each sample. The two sets of sequence data for each of the mixture with between 4 and 6% fetal DNA were added together, and the resulting 7 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

Of the remaining four mixtures, all of them had between 2 and 4% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 12 million reads for each sample. The two sets of sequence data for each of the mixture with between 2 and 4% fetal DNA were added together, and the resulting 13 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

This method required six lanes of sequencing on a HISEQ machine to achieve 99.9% accuracy over 100 samples. If the same number of runs had been required for every sample, to ensure that every ploidy determination was made with a 99.9% accuracy, it would have taken 25 lanes of sequencing, and if a no-call rate or error rate of 4% was tolerated, it could have been achieved with 14 lanes of sequencing.

Using Raw Genotyping Data

There are a number of methods that can accomplish NPD using fetal genetic information measured on fetal DNA found in maternal blood. Some of these methods involve making measurements of the fetal DNA using SNP arrays, some methods involve untargeted sequencing, and some methods involve targeted sequencing. The targeted sequencing may target SNPs, it may target STRs, it may target other polymorphic loci, it may target non-

polymorphic loci, or some combination thereof. Some of these methods may involve using a commercial or proprietary allele caller that calls the identity of the alleles from the intensity data that comes from the sensors in the machine doing the measuring. For example, the ILLUMINA INFINIUM system or the AFFYMETRIX GENECHIP microarray system involves beads or microchips with attached DNA sequences that can hybridize to complementary segments of DNA; upon hybridization, there is a change in the fluorescent properties of the sensor molecule that can be detected. There are also sequencing methods, for example the ILLUMINA SOLEXA GENOME SEQUENCER or the ABI SOLID GENOME SEQUENCER, wherein the genetic sequence of fragments of DNA are sequenced; upon extension of the strand of DNA complementary to the strand being sequenced, the identity of the extended nucleotide is typically detected via a fluorescent or radio tag appended to the complementary nucleotide. In all of these methods the genotypic or sequencing data is typically determined on the basis of fluorescent or other signals, or the lack thereof. These systems are typically combined with low level software packages that make specific allele calls (secondary genetic data) from the analog output of the fluorescent or other detection device (primary genetic data). For example, in the case of a given allele on a SNP array, the software will make a call, for example, that a certain SNP is present or not present if the fluorescent intensity is measure above or below a certain threshold. Similarly, the output of a sequencer is a chromatogram that indicates the level of fluorescence detected for each of the dyes, and the software will make a call that a certain base pair is A or T or C or G. High throughput sequencers typically make a series of such measurements, called a read, that represents the most likely structure of the DNA sequence that was sequenced. The direct analog output of the chromatogram is defined here to be the primary genetic data, and the base pair / SNP calls made by the software are considered here to be the secondary genetic data. In an embodiment, primary data refers to the raw intensity data that is the unprocessed output of a genotyping platform, where the genotyping platform may refer to a SNP array, or to a sequencing platform. The secondary genetic data refers to the processed genetic data, where an allele call has been made, or the sequence data has been assigned base pairs, and/or the sequence reads have been mapped to the genome.

Many higher level applications take advantage of these allele calls, SNP calls and sequence reads, that is, the secondary genetic data, that the genotyping software produces. For example, DNA NEXUS, ELAND or MAQ will take the sequencing reads and map them to the

genome. For example, in the context of non-invasive prenatal diagnosis, complex informatics, such as PARENTAL SUPPORT™, may leverage a large number of SNP calls to determine the genotype of an individual. Also, in the context of preimplantation genetic diagnosis, it is possible to take a set of sequence reads that are mapped to the genome, and by taking a normalized count of the reads that are mapped to each chromosome, or section of a chromosome, it may be possible to determine the ploidy state of an individual. In the context of non-invasive prenatal diagnosis it may be possible to take a set of sequence reads that have been measured on DNA present in maternal plasma, and map them to the genome. One may then take a normalized count of the reads that are mapped to each chromosome, or section of a chromosome, and use that data to determine the ploidy state of an individual. For example, it may be possible to conclude that those chromosomes that have a disproportionately large number of reads are trisomic in the fetus that is gestating in the mother from which the blood was drawn.

However, in reality, the initial output of the measuring instruments is an analog signal. When a certain base pair is called by the software that is associated with the sequencing software, for example the software may call the base pair a T, in reality the call is the call that the software believes to be most likely. In some cases, however, the call may be of low confidence, for example, the analog signal may indicate that the particular base pair is only 90% likely to be a T, and 10% likely to be an A. In another example, the genotype calling software that is associated with a SNP array reader may call a certain allele to be G. However, in reality, the underlying analog signal may indicate that it is only 70% likely that the allele is G, and 30% likely that the allele is T. In these cases, when the higher level applications use the genotype calls and sequence calls made by the lower level software, they are losing some information. That is, the primary genetic data, as measured directly by the genotyping platform, may be messier than the secondary genetic data that is determined by the attached software packages, but it contains more information. In mapping the secondary genetic data sequences to the genome, many reads are thrown out because some bases are not read with enough clarity and or mapping is not clear. When the primary genetic data sequence reads are used, all or many of those reads that may have been thrown out when first converted to secondary genetic data sequence read can be used by treating the reads in a probabilistic manner.

In an embodiment of the present disclosure, the higher level software does not rely on the allele calls, SNP calls, or sequence reads that are determined by the lower level software. Instead, the higher level software bases its calculations on the analog signals directly measured from the genotyping platform. In an embodiment of the present disclosure, an informatics based method such as PARENTAL SUPPORT™ is modified so that its ability to reconstruct the genetic data of the embryo / fetus / child is engineered to directly use the primary genetic data as measured by the genotyping platform. In an embodiment of the present disclosure, an informatics based method such as PARENTAL SUPPORT™ is able to make allele calls, and/or chromosome copy number calls using primary genetic data, and not using the secondary genetic data. In an embodiment of the present disclosure, all genetic calls, SNPs calls, sequence reads, sequence mapping is treated in a probabilistic manner by using the raw intensity data as measured directly by the genotyping platform, rather than converting the primary genetic data to secondary genetic calls. In an embodiment, the DNA measurements from the prepared sample used in calculating allele count probabilities and determining the relative probability of each hypothesis comprise primary genetic data.

In some embodiments, the method can increase the accuracy of genetic data of a target individual which incorporates genetic data of at least one related individual, the method comprising obtaining primary genetic data specific to a target individual's genome and genetic data specific to the genome(s) of the related individual(s), creating a set of one or more hypotheses concerning possibly which segments of which chromosomes from the related individual(s) correspond to those segments in the target individual's genome, determining the probability of each of the hypotheses given the target individual's primary genetic data and the related individual(s)'s genetic data, and using the probabilities associated with each hypothesis to determine the most likely state of the actual genetic material of the target individual. In some embodiments, the method can determining the number of copies of a segment of a chromosome in the genome of a target individual, the method comprising creating a set of copy number hypotheses about how many copies of the chromosome segment are present in the genome of a target individual, incorporating primary genetic data from the target individual and genetic information from one or more related individuals into a data set, estimating the characteristics of the platform response associated with the data set, where the platform response may vary from one experiment to another, computing the conditional probabilities of each copy number hypothesis, given the data set and the platform response characteristics, and

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 174

determining the copy number of the chromosome segment based on the most probable copy number hypothesis. In an embodiment, a method of the present disclosure can determine a ploidy state of at least one chromosome in a target individual, the method comprising obtaining primary genetic data from the target individual and from one or more related individuals, creating a set of at least one ploidy state hypothesis for each of the chromosomes of the target individual, using one or more expert techniques to determine a statistical probability for each ploidy state hypothesis in the set, for each expert technique used, given the obtained genetic data, combining, for each ploidy state hypothesis, the statistical probabilities as determined by the one or more expert techniques, and determining the ploidy state for each of the chromosomes in the target individual based on the combined statistical probabilities of each of the ploidy state hypotheses. In an embodiment, a method of the present disclosure can determine an allelic state in a set of alleles, in a target individual, and from one or both parents of the target individual, and optionally from one or more related individuals, the method comprising obtaining primary genetic data from the target individual, and from the one or both parents, and from any related individuals, creating a set of at least one allelic hypothesis for the target individual, and for the one or both parents, and optionally for the one or more related individuals, where the hypotheses describe possible allelic states in the set of alleles, determining a statistical probability for each allelic hypothesis in the set of hypotheses given the obtained genetic data, and determining the allelic state for each of the alleles in the set of alleles for the target individual, and for the one or both parents, and optionally for the one or more related individuals, based on the statistical probabilities of each of the allelic hypotheses.

In some embodiments, the genetic data of the mixed sample may comprise sequence data wherein the sequence data may not uniquely map to the human genome. In some embodiments, the genetic data of the mixed sample may comprise sequence data wherein the sequence data maps to a plurality of locations in the genome, wherein each possible mapping is associated with a probability that the given mapping is correct. In some embodiments, the sequence reads are not assumed to be associated with a particular position in the genome. In some embodiments, the sequence reads are associated with a plurality of positions in the genome, and an associated probability belonging to that position.

Counting Method to Determine Chromosome Copy Number

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 175

In one aspect, the invention features methods of testing for an abnormal distribution of a fetal chromosome by comparing the number of sequence tags that align to different chromosomes (*see, e.g.*, U.S. Pat. No. 8,296,076, filed April 20, 2012, which is hereby incorporated by reference in its entirety). As is known in the art, the term "sequence tag" refers to a relatively short (*e.g.*, 15-100) nucleic acid sequence that can be used to identify a certain larger sequence, *e.g.*, be mapped to a chromosome or genomic region or gene. In some embodiments, the method involves (i) contacting a sample that includes a mixture of maternal and fetal DNA with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture; wherein the target loci are from a plurality of different chromosomes; and wherein the plurality of different chromosomes comprise at least one first chromosome suspected of having an abnormal distribution in the sample and at least one second chromosome presumed to be normally distributed in the sample; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products; (iii) sequencing the amplified products to obtain a plurality of sequence tags aligning to the target loci; wherein the sequence tags are of sufficient length to be assigned to a specific target locus; (iv) assigning on a computer the plurality of sequence tags to their corresponding target loci; (v) determining on a computer a number of sequence tags aligning to the target loci of the first chromosome and a number of sequence tags aligning to the target loci of the second chromosome; and (vi) comparing the numbers from step (v) to determine the presence or absence of an abnormal distribution of the first chromosome.

In one aspect, the invention provides methods for detecting the presence or absence of a fetal aneuploidy by comparing the relative frequency of target amplicons between chromosomes (*see, e.g.,* PCT Publ. No. WO 2012/103031, filed Jan. 23, 2012, which is hereby incorporated by reference in its entirety). In some embodiments, the method involves (i) contacting a sample that includes a mixture of maternal and fetal DNA with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different non-polymorphic target loci to produce a reaction mixture; wherein the target loci are from a plurality of different chromosomes; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that includes target amplicons; (iii) quantifying on a computer a relative frequency of the target amplicons from the first and second

chromosomes of interest; (iv) comparing on a computer the relative frequency of the target amplicons from the first and second chromosomes of interest; and (v) identifying the presence or absence of an aneuploidy based on the compared relative frequencies of the first and second chromosome of interest. In some embodiments, the first chromosome is a chromosome suspected of being euploid. In some embodiments, the second chromosome is a chromosome suspected of being aneuploidy

Combining Methods of Prenatal Diagnosis

There are many methods that may be used for prenatal diagnosis or prenatal screening of aneuploidy or other genetic defects. Described elsewhere in this document, and in U.S. Utility Application Serial No. 11/603,406, filed November 28, 2006; U.S. Utility Application Serial No. 12/076,348, filed March 17, 2008, and PCT Application Serial No. PCT/S09/52730 is one such method that uses the genetic data of related individuals to increase the accuracy with which genetic data of a target individual, such as a fetus, is known, or estimated. Other methods used for prenatal diagnosis involve measuring the levels of certain hormones in maternal blood, where those hormones are correlated with various genetic abnormalities. An example of this is called the triple test, a test wherein the levels of several (commonly two, three, four or five) different hormones are measured in maternal blood. In a case where multiple methods are used to determine the likelihood of a given outcome, where none of the methods are definitive in and of themselves, it is possible to combine the information given by those methods to make a prediction that is more accurate than any of the individual methods. In the triple test, combining the information given by the three different hormones can result in a prediction of genetic abnormalities that is more accurate than the individual hormone levels may predict.

Disclosed herein is a method for making more accurate predictions about the genetic state of a fetus, specifically the possibility of genetic abnormalities in a fetus that comprises combining predictions of genetic abnormalities in a fetus where those predictions were made using a variety of methods. A "more accurate" method may refer to a method for diagnosing an abnormality that has a lower false negative rate at a given false positive rate. In a favored embodiment of the present disclosure, one or more of the predictions are made based on the genetic data known about the fetus, where the genetic knowledge was determined using the PARENTAL SUPPORT™ method, that is, using genetic data of individual related to the fetus

to determine the genetic data of the fetus with greater accuracy. In some embodiments the genetic data may include ploidy states of the fetus. In some embodiments, the genetic data may refer to a set of allele calls on the genome of the fetus. In some embodiments some of the predictions may have been made using the triple test. In some embodiments, some of the predictions may have been made using measurements of other hormone levels in maternal blood. In some embodiments, predictions made by methods considered diagnoses may be combined with predictions made by methods considered screening. In some embodiments, the method involves measuring maternal blood levels of alpha-fetoprotein (AFP). In some embodiments, the method involves measuring maternal blood levels of unconjugated estriol (UE_3). In some embodiments, the method involves measuring maternal blood levels of beta human chorionic gonadotropin (beta-hCG). In some embodiments, the method involves measuring maternal blood levels of invasive trophoblast antigen (ITA). In some embodiments, the method involves measuring maternal blood levels of inhibin. In some embodiments, the method involves measuring maternal blood levels of pregnancy-associated plasma protein A (PAPP-A). In some embodiments, the method involves measuring maternal blood levels of other hormones or maternal serum markers. In some embodiments, some of the predictions may have been made using other methods. In some embodiments, some of the predictions may have been made using a fully integrated test such as one that combines ultrasound and blood test at around 12 weeks of pregnancy and a second blood test at around 16 weeks. In some embodiments, the method involves measuring the fetal nuchal translucency (NT). In some embodiments, the method involves using the measured levels of the aforementioned hormones for making predictions. In some embodiments the method involves a combination of the aforementioned methods.

There are many ways to combine the predictions, for example, one could convert the hormone measurements into a multiple of the median (MoM) and then into likelihood ratios (LR). Similarly, other measurements could be transformed into LRs using the mixture model of NT distributions. The LRs for NT and the biochemical markers could be multiplied by the age and gestation-related risk to derive the risk for various conditions, such as trisomy 21. Detection rates (DRs) and false-positive rates (FPRs) could be calculated by taking the proportions with risks above a given risk threshold.

In an embodiment, a method to call the ploidy state involves combining the relative probabilities of each of the ploidy hypotheses determined using the joint distribution model

and the allele count probabilities with relative probabilities of each of the ploidy hypotheses that are calculated using statistical techniques taken from other methods that determine a risk score for a fetus being trisomic, including but not limited to: a read count analysis, comparing heterozygosity rates, a statistic that is only available when parental genetic information is used, the probability of normalized genotype signals for certain parent contexts, a statistic that is calculated using an estimated fetal fraction of the first sample or the prepared sample, and combinations thereof.

Another method could involve a situation with four measured hormone levels, where the probability distribution around those hormones is known: $p(x_1, x_2, x_3, x_4|e)$ for the euploid case and $p(x_1, x_2, x_3, x_4|a)$ for the aneuploid case. Then one could measure the probability distribution for the DNA measurements, $g(y|e)$ and $g(y|a)$ for the euploid and aneuploid cases respectively. Assuming they are independent given the assumption of euploid/aneuploid, one could combine as $p(x_1, x_2, x_3, x_4|a)g(y|a)$ and $p(x_1, x_2, x_3, x_4|e)g(y|e)$ and then multiply each by the prior $p(a)$ and $p(e)$ given the maternal age. One could then choose the one that is highest.

In an embodiment, it is possible to evoke central limit theorem to assume distribution on $g(y|a \text{ or } e)$ is Gaussian, and measure mean and standard deviation by looking at multiple samples. In another embodiment, one could assume they are not independent given the outcome and collect enough samples to estimate the joint distribution $p(x_1, x_2, x_3, x_4|a \text{ or } e)$.

In an embodiment, the ploidy state for the target individual is determined to be the ploidy state that is associated with the hypothesis whose probability is the greatest. In some cases, one hypothesis will have a normalized, combined probability greater than 90%. Each hypothesis is associated with one, or a set of, ploidy states, and the ploidy state associated with the hypothesis whose normalized, combined probability is greater than 90%, or some other threshold value, such as 50%, 80%, 95%, 98%, 99%, or 99.9%, may be chosen as the threshold required for a hypothesis to be called as the determined ploidy state.

DNA from Children from Previous Pregnancies in Maternal Blood

One difficulty to non-invasive prenatal diagnosis is differentiating fetal cells from the current pregnancy from fetal cells from previous pregnancies. Some believe that genetic matter from prior pregnancies will go away after some time, but conclusive evidence has not been shown. In an embodiment of the present disclosure, it is possible to determine fetal DNA present in the maternal blood of paternal origin (that is, DNA that the fetus inherited from the

father) using the PARENTAL SUPPORT™ (PS) method, and the knowledge of the paternal genome. This method may utilize phased parental genetic information. It is possible to phase the parental genotype from unphased genotypic information using grandparental genetic data (such as measured genetic data from a sperm from the grandfather), or genetic data from other born children, or a sample of a miscarriage. One could also phase unphased genetic information by way of a HapMap-based phasing, or a haplotyping of paternal cells. Successful haplotyping has been demonstrated by arresting cells at phase of mitosis when chromosomes are tight bundles and using microfluidics to put separate chromosomes in separate wells. In another embodiment it is possible to use the phased parental haplotypic data to detect the presence of more than one homolog from the father, implying that the genetic material from more than one child is present in the blood. By focusing on chromosomes that are expected to be euploid in a fetus, one could rule out the possibility that the fetus was afflicted with a trisomy. Also, it is possible to determine if the fetal DNA is not from the current father, in which case one could use other methods such as the triple test to predict genetic abnormalities.

There may be other sources of fetal genetic material available via methods other than a blood draw. In the case of the fetal genetic material available in maternal blood, there are two main categories: (1) whole fetal cells, for example, nucleated fetal red blood cells or erythroblats, and (2) free floating fetal DNA. In the case of whole fetal cells, there is some evidence that fetal cells can persist in maternal blood for an extended period of time such that it is possible to isolate a cell from a pregnant woman that contains the DNA from a child or fetus from a prior pregnancy. There is also evidence that the free floating fetal DNA is cleared from the system in a matter of weeks. One challenge is how to determine the identity of the individual whose genetic material is contained in the cell, namely to ensure that the measured genetic material is not from a fetus from a prior pregnancy. In an embodiment of the present disclosure, the knowledge of the maternal genetic material can be used to ensure that the genetic material in question is not maternal genetic material. There are a number of methods to accomplish this end, including informatics based methods such as PARENTAL SUPPORT™, as described in this document or any of the patents referenced in this document.

In an embodiment of the present disclosure, the blood drawn from the pregnant mother may be separated into a fraction comprising free floating fetal DNA, and a fraction comprising nucleated red blood cells. The free floating DNA may optionally be enriched, and the genotypic information of the DNA may be measured. From the measured genotypic

information from the free floating DNA, the knowledge of the maternal genotype may be used to determine aspects of the fetal genotype. These aspects may refer to ploidy state, and/or a set of allele identities. Then, individual nucleated red blood cells may be genotyped using methods described elsewhere in this document, and other referent patents, especially those mentioned in the first section of this document. The knowledge of the maternal genome would allow one to determine whether or not any given single blood cell is genetically maternal. And the aspects of the fetal genotype that were determined as described above would allow one to determine if the single blood cell is genetically derived from the fetus that is currently gestating. In essence, this aspect of the present disclosure allows one to use the genetic knowledge of the mother, and possibly the genetic information from other related individuals, such as the father, along with the measured genetic information from the free floating DNA found in maternal blood to determine whether an isolated nucleated cell found in maternal blood is either (a) genetically maternal, (b) genetically from the fetus currently gestating, or (c) genetically from a fetus from a prior pregnancy.

Prenatal Sex Chromosome Aneuploidy Determination

In methods known in the art, people attempting to determine the sex of a gestating fetus from the blood of the mother have used the fact that fetal free floating DNA (fffDNA) is present in the plasma of the mother. If one is able to detect Y-specific loci in the maternal plasma, this implies that the gestating fetus is a male. However, the lack of detection of Y-specific loci in the plasma does not always guarantee that the gestating fetus is a female when using methods known in the art, as in some cases the amount of fffDNA is too low to ensure that the Y-specific loci would be detected in the case of a male fetus.

Presented here is a novel method that does not require the measurement of Y-specific nucleic acids, that is, DNA that is from loci that are exclusively paternally derived. The Parental Support method, disclosed previously, uses crossover frequency data, parental genotypic data, and informatics techniques, to determine the ploidy state of a gestating fetus. The sex of a fetus is simply the ploidy state of the fetus at the sex chromosomes. A child that is XX is female, and XY is male. The method described herein is also able to determine the ploidy state of the fetus. Note that sexing is effectively synonymous with ploidy determination of the sex chromosomes; in the case of sexing, an assumption is often made that the child is euploid, therefore there are fewer possible hypotheses.

The method disclosed herein involves looking at loci that are common to both the X and Y chromosome to create a baseline in terms of expected amount of fetal DNA present for a fetus. Then, those regions that are specific only to the X chromosome can be interrogated to determine if the fetus is female or male. In the case of a male, we expect to see less fetal DNA from loci that are specific to the X chromosome than from loci that are specific to both the X and the Y. In contrast, in female fetuses, we expect the amount of DNA for each of these groups to be the same. The DNA in question can be measured by any technique that can quantitate the amount of DNA present on a sample, for example, qPCR, SNP arrays, genotyping arrays, or sequencing. For DNA that is exclusively from an individual we would expect to see the following:

	DNA specific to X	DNA specific to X and Y	DNA specific to Y
Male (XY)	A	2A	A
Female (XX)	2A	2A	0

In the case of DNA from a fetus that is mixed with DNA from the mother, and where the fraction of fetal DNA in the mixture is F, and where the fraction of maternal DNA in the mixture is M, such that F+M = 100%, we would expect to see the following:

	DNA specific to X	DNA specific to X and Y	DNA specific to Y
Male fetus (XY)	M + ½ F	M + F	½ F
Female fetus (XX)	M + F	M + F	0

In the case where F and M are known, the expected ratios can be computed, and the observed data can be compared to the expected data. In the case where M and F are not known, a threshold can be selected based on historical data. In both cases, the measured amount of DNA at loci specific to both X and Y can be used as a baseline, and the test for the sex of the fetus can be based on the amount of DNA observed on loci specific to only the X chromosome. If that amount is lower than the baseline by an amount roughly equal to ½ F, or by an amount that causes it to fall below a predefined threshold, the fetus is determined to be male, and if

that amount is about equal to the baseline, or if is not lower by an amount that causes it to fall below a predefined threshold, the fetus is determined to be female.

In another embodiment, one can look only at those loci that are common to both the X and the Y chromosomes, often termed the Z chromosome. A subset of the loci on the Z chromosome are typically always A on the X chromosome, and B on the Y chromosome. If SNPs from the Z chromosome are found to have the B genotype, then the fetus is called a male; if the SNPs from the Z chromosome are found to only have A genotype, then the fetus is called a female. In another embodiment, one can look at the loci that are found only on the X chromosome. Contexts such as AA|B are particularly informative as the presence of a B indicates that the fetus has an X chromosome from the father. Contexts such as AB|B are also informative, as we expect to see B present only half as often in the case of a female fetus as compared to a male fetus. In another embodiment, one can look at the SNPs on the Z chromosome where both A and B alleles are present on both the X and the Y chromosome, and where the it is known which SNPs are from the paternal Y chromosome, and which are from the paternal X chromosome.

In an embodiment, it is possible to amplify single nucleotide positions known to varying between the homologous non-recombining (HNR) region shared by chromosome Y and chromosome X. The sequence within this HNR region is largely identical between the X and Y chromosomes. Within this identical region are single nucleotide positions that, while invariant among X chromosomes and among Y chromosomes in the population, are different between the X and Y chromosomes. Each PCR assay could amplify a sequence from loci that are present on both the X and Y chromosomes. Within each amplified sequence would be a single base that can be detected using sequencing or some other method.

In an embodiment, the sex of the fetus could be determined from the fetal free floating DNA found in maternal plasma, the method comprising some or all of the following steps: 1) Design PCR (either regular or mini-PCR, plus multiplexing if desired) primers amplify X/Y variant single nucleotide positions within HNR region, 2) obtain maternal plasma, 3) PCR Amplify targets from maternal plasma using HNR X/Y PCR assays, 4) sequence the amplicons, 5) Examine sequence data for presence of Y-allele within one or more of the amplified sequences. The presence of one or more would indicate a male fetus. Absence of all Y-alleles from all amplicons indicates a female fetus.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 183

In an embodiment, one could use targeted sequencing to measure the DNA in the maternal plasma and/or the parental genotypes. In an embodiment, one could ignore all sequences that clearly originate from paternally sourced DNA. For example, in the context AA|AB, one could count the number of A sequences and ignore all the B sequences. In order to determine a heterozygosity rate for the above algorithm, one could compare the number of observed A sequences to the expected number of total sequences for the given probe. There are many ways one could calculate an expected number of sequences for each probe on a per sample basis. In an embodiment, it is possible to use historical data to determine what fraction of all sequence reads belongs to each specific probe and then use this empirical fraction, combined with the total number of sequence reads, to estimate the number of sequences at each probe. Another approach could be to target some known homozygous alleles and then use historical data to relate the number of reads at each probe with the number of reads at the known homozygous alleles. For each sample, one could then measure the number of reads at the homozygous alleles and then use this measurement, along with the empirically derived relationships, to estimate the number of sequence reads at each probe.

In some embodiments, it is possible to determine the sex of the fetus by combining the predictions made by a plurality of methods. In some embodiments the plurality of methods are taken from methods described in this disclosure. In some embodiments, at least one of the plurality of methods are taken from methods described in this disclosure.

In some embodiments the method described herein can be used to determine the ploidy state of the gestating fetus. In an embodiment, the ploidy calling method uses loci that are specific to the X chromosome, or common to both the X and Y chromosome, but does not make use of any Y-specific loci. In an embodiment, the ploidy calling method uses one or more of the following: loci that are specific to the X chromosome, loci that are common to both the X and Y chromosome, and loci that are specific to the Y chromosome. In an embodiment, where the ratios of sex chromosomes are similar, for example 45,X (Turner Syndrome), 46,XX (normal female) and 47,XXX (trisomy X), the differentiation can be accomplished by comparing the allele distributions to expected allele distributions according to the various hypotheses. In another embodiment, this can be accomplished by comparing the relative number of sequence reads for the sex chromosomes to one or a plurality of reference chromosomes that are assumed to be euploid. Also note that these methods can be expanded to include aneuploid cases.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 184

Single Gene Disease Screening

In an embodiment, a method for determining the ploidy state of the fetus may be extended to enable simultaneous testing for single gene disorders. Single-gene disease diagnosis leverages the same targeted approach used for aneuploidy testing, and requires additional specific targets. In an embodiment, the single gene NPD diagnosis is through linkage analysis. In many cases, direct testing of the cfDNA sample is not reliable, as the presence of maternal DNA makes it virtually impossible to determine if the fetus has inherited the mother's mutation. Detection of a unique paternally-derived allele is less challenging, but is only fully informative if the disease is dominant and carried by the father, limiting the utility of the approach. In an embodiment, the method involves PCR or related amplification approaches.

In some embodiments, the method involves phasing the abnormal allele with surrounding very tightly linked SNPs in the parents using information from first-degree relatives. Then Parental Support may be run on the targeted sequencing data obtained from these SNPs to determine which homologs, normal or abnormal, were inherited by the fetus from both parents. As long as the SNPs are sufficiently linked, the inheritance of the genotype of the fetus can be determined very reliably. In some embodiments, the method comprises (a) adding a set of SNP loci to densely flank a specified set of common diseases to our multiplex pool for aneuploidy testing; (b) reliably phasing the alleles from these added SNPs with the normal and abnormal alleles based on genetic data from various relatives; and (c) reconstructing the fetal haplotype, or set of phased SNP alleles on the inherited maternal and paternal homologs in the region surrounding the disease locus to determine fetal genotype. In some embodiments additional probes that are closely linked to a disease linked locus are added to the set of polymorphic locus being used for aneuploidy testing.

Reconstructing fetal diplotype is challenging because the sample is a mixture of maternal and fetal DNA. In some embodiments, the method incorporates relative information to phase the SNPs and disease alleles, then take into account physical distance of the SNPs and recombination data from location specific recombination likelihoods and the data observed from the genetic measurements of the maternal plasma to obtain the most likely genotype of the fetus.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 185

In an embodiment, a number of additional probes per disease linked locus are included in the set of targeted polymorphic loci; the number of additional probes per disease linked locus may be between 4 and 10, between 11 and 20, between 21 and 40, between 41 and 60, between 61 and 80, or combinations thereof.

Phasing the diploid data from the parents can be challenging, and there are a number of ways this can be accomplished. Some are discussed in this disclosure, others are described in greater detail in other disclosures (*see*, *e.g.*, PCT Publ. No. WO2009105531, filed February 9, 2009, and PCT Publ. No. WO2010017214, filed August 4, 2009, which are each hereby incorporated by reference in its entirety). In one embodiment, a parent can be phased by inference by measuring tissue from the parent that is haploid, for example by measuring one or more sperm or eggs. In one embodiment the parent can be phased by inference using the measured genotypic data of a first degree relative such as the parent's parent(s) or siblings. In one embodiment, the parent can be phased by dilution where the DNA is diluted, in one or a plurality of wells, to the point where there is expected to be no more than approximately one copy of each haplotype in each well, and then measuring the DNA in the one or more wells. In one embodiment, the parent genotype can be phased by using computer programs that use population based haplotype frequencies to infer the most likely phase. In one embodiment, the parent can be phased if the phased haplotypic data is known for the other parent, along with the unphased genetic data of one or more genetic offspring of the parents. In some embodiments, the genetic offspring of the parents may be one or more embryos, fetuses, and/or born children. Some of these methods and other methods for phasing one or both parents are disclosed in greater detail in, *e.g.*, U.S. Publ. No. 2011/0033862, filed August 19, 2010; U.S. Publ. No. 2011/0178719, filed Feb. 3, 2011; U.S. Publ. No. 2007/0184467, filed Nov. 22, 2006; U.S. Publ. No. 2008/0243398, filed March 17, 2008, which are each hereby incorporated by reference in its entirety.

Fetal Genome Reconstruction

In one aspect, the invention features methods for determining a haplotype of a fetus. In various embodiments, this method allows one to determine which polymorphic loci (such as SNPs) were inherited by the fetus and to reconstruct which homologs (including recombination events) are present in the fetus (and thereby interpolate the sequence between the polymorphic loci). If desired, essentially the entire genome of the fetus can be reconstructed. If there is

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 186

some remaining ambiguity in the genome of the fetus (such as in intervals with a crossover), this ambiguity can be minimized if desired by analyzing additional polymorphic loci. In various embodiments, the polymorphic loci are chosen to cover one or more of the chromosomes at a density to reduce any ambiguity to a desired level. This method has important applications for the detection of polymorphisms or other mutations of interest in a fetus since it enables their detection based on linkage (such as the presence of linked polymorphic loci in the fetal genome) rather than by directing detecting the polymorphism or other mutation of interest in the fetal genome. For example, if a parent is a carrier for a mutation associated with cystic fibrosis (CF), a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus can be analyzed to determine whether the fetal DNA include the haplotype containing the CF mutation. In particular, polymorphic loci can be analyzed to determine whether the fetal DNA includes the haplotype containing the CF mutation without having to detect the CF mutation itself in the fetal DNA.

In some embodiments, the method involves determining a parental haplotype (*e.g.*, a haplotype of the mother or father of the fetus). In some embodiments, this determination is made without using data from a relative of the mother or father. In some embodiments, a parental haplotype is determined using a dilution approach followed by SNP genotyping or sequencing as described herein and elsewhere (*see*, *e.g.*, U.S. Publ. No. 2011/0033862, filed August 19, 2010, which is hereby incorporated by reference in its entirety). Because the DNA is diluted, it is unlikely that more than one haplotype is in the same fraction (or tube). Thus, there may be effectively a single molecule of DNA in the tube, which allows the haplotype on a single DNA molecule to be determined. In some embodiments, the method includes dividing a DNA sample into a plurality of fractions such that at least one of the fractions includes one chromosome or one chromosome segment from a pair of chromosomes, and genotyping (*e.g.*, determining the presence of two or more polymorphic loci) the DNA sample in at least one of the fractions, thereby determining a parental haplotype. In some embodiments, the genotyping involves sequencing (such as shotgun sequencing). In some embodiments, the genotyping involves use of a SNP array to detect polymorphic loci, such as at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci. In some embodiments, the genotyping involves the use of multiplex PCR. In some embodiments, the method involves contacting the sample in a fraction with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000;

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 187

25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture; and subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that are measured with a high throughput sequencer to produce sequencing data.

In some embodiments, a haplotype of the mother is determined by any of the methods described herein using data from a relative of the mother. In some embodiments, a haplotype of the father is determined by any of the methods described herein using data from a relative of the father. In some embodiments, a haplotype is determined for both the father and the mother. In some embodiments, a SNP array is used to determine the presence of at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci in a DNA sample from the mother (or father) and a relative of the mother (or father). In some embodiments, the method involves contacting a DNA sample from the mother (or father) and/or a relative of the mother (or father) with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture; and subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that are measured with a high throughput sequencer to produce sequencing data. The parental haplotype may be determined based on the SNP array or sequencing data. In some embodiments, parental data may be phased by methods described or referred to elsewhere in this document.

This parental haplotype data can be used to determine if the fetus inherited the parental haplotype. In some embodiments, a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus is analyzed using a SNP array to detect at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci. In some embodiments, a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus is analyzed by contacting the sample with a library of primers that simultaneously hybridize to at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture. In some embodiments, the reaction mixture is subjected to primer extension reaction conditions to produce amplified products. In some embodiments, the amplified products are measured with a high throughput sequencer to produce sequencing data. In various embodiments, the SNP

array or sequencing data is used to determine a parental haplotype by using data about the probability of chromosomes crossing over at different locations in a chromosome (such as by using recombination data such as may be found in the HapMap database to create a recombination risk score for any interval) to model dependence between polymorphic alleles on the chromosome. In some embodiments, allele counts at the polymorphic loci are calculated on a computer based on the sequencing data. In some embodiments, a plurality of ploidy hypotheses each pertaining to a different possible ploidy state of the chromosome are created on a computer; a model (such as a joint distribution model) for the expected allele counts at the polymorphic loci on the chromosome is built on a computer for each ploidy hypothesis; a relative probability of each of the ploidy hypotheses is determined on a computer using the joint distribution model and the allele counts; and the ploidy state of the fetus is called by selecting the ploidy state corresponding to the hypothesis with the greatest probability. In some embodiments, building a joint distribution model for allele counts and the step of determining the relative probability of each hypothesis are done using a method that does not require the use of a reference chromosome.

In some embodiments, a fetal haplotype is determined for one or more chromosomes taken from the group consisting of chromosomes 13, 18, 21, X, and Y. In some embodiments, a fetal haplotype is determined for all of the fetal chromosomes. In various embodiments, the method determines essentially the entire genome of the fetus. In some embodiments, the haplotype is determined for at least 30, 40, 50, 60, 70, 80, 90, or 95% of the genome of the fetus. In some embodiments, the haplotype determination of the fetus includes information about which allele is present for at least 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci.

Compositions of DNA

When performing an informatics analysis on sequencing data measured on a mixture of fetal and maternal blood to determine genomic information pertaining to the fetus, for example the ploidy state of the fetus, it may be advantageous to measure the allele distributions at a set of alleles. Unfortunately, in many cases, such as when attempting to determine the ploidy state of a fetus from the DNA mixture found in the plasma of a maternal blood sample, the amount of DNA available is not sufficient to directly measure the allele distributions with good fidelity in the mixture. In these cases, amplification of the DNA mixture will provide sufficient

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 189

numbers of DNA molecules that the desired allele distributions may be measured with good fidelity. However, current methods of amplification typically used in the amplification of DNA for sequencing are often very biased, meaning that they do not amplify both alleles at a polymorphic locus by the same amount. A biased amplification can result in allele distributions that are quite different from the allele distributions in the original mixture. For most purposes, highly accurate measurements of the relative amounts of alleles present at polymorphic loci are not needed. In contrast, in an embodiment of the present disclosure, amplification or enrichment methods that specifically enrich polymorphic alleles and preserve allelic ratios is advantageous.

A number of methods are described herein that may be used to preferentially enrich a sample of DNA at a plurality of loci in a way that minimizes allelic bias. Some examples are using circularizing probes to target a plurality of loci where the 3' ends and 5' ends of the pre-circularized probe are designed to hybridize to bases that are one or a few positions away from the polymorphic sites of the targeted allele. Another is to use PCR probes where the 3' end PCR probe is designed to hybridize to bases that are one or a few positions away from the polymorphic sites of the targeted allele. Another is to use a split and pool approach to create mixtures of DNA where the preferentially enriched loci are enriched with low allelic bias without the drawbacks of direct multiplexing. Another is to use a hybrid capture approach where the capture probes are designed such that the region of the capture probe that is designed to hybridize to the DNA flanking the polymorphic site of the target is separated from the polymorphic site by one or a small number of bases.

In the case where measured allele distributions at a set of polymorphic loci are used to determine the ploidy state of an individual, it is desirable to preserve the relative amounts of alleles in a sample of DNA as it is prepared for genetic measurements. This preparation may involve WGA amplification, targeted amplification, selective enrichment techniques, hybrid capture techniques, circularizing probes or other methods meant to amplify the amount of DNA and/or selectively enhance the presence of molecules of DNA that correspond to certain alleles.

In some embodiments of the present disclosure, there is a set of DNA probes designed to target loci where the loci have maximal minor allele frequencies. In some embodiments of the present disclosure, there is a set of probes that are designed to target where the loci have the maximum likelihood of the fetus having a highly informative SNP at those loci. In some

embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given population subgroup. In some embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given mix of population subgroups. In some embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given pair of parents which are from different population subgroups that have different minor allele frequency profiles. In some embodiments of the present disclosure, there is a circularized strand of DNA that comprises at least one base pair that annealed to a piece of DNA that is of fetal origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that comprises at least one base pair that annealed to a piece of DNA that is of placental origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that circularized while at least some of the nucleotides were annealed to DNA that was of fetal origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that circularized while at least some of the nucleotides were annealed to DNA that was of placental origin. In some embodiments of the present disclosure, there is a set of probes wherein some of the probes target single tandem repeats, and some of the probes target single nucleotide polymorphisms. In some embodiments, the loci are selected for the purpose of non-invasive prenatal diagnosis. In some embodiments, the probes are used for the purpose of non-invasive prenatal diagnosis. In some embodiments, the loci are targeted using a method that could include circularizing probes, MIPs, capture by hybridization probes, probes on a SNP array, or combinations thereof. In some embodiments, the probes are used as circularizing probes, MIPs, capture by hybridization probes, probes on a SNP array, or combinations thereof. In some embodiments, the loci are sequenced for the purpose of non-invasive prenatal diagnosis.

In the case where the relative informativeness of a sequence is greater when combined with relevant parent contexts, it follows that maximizing the number of sequence reads that contain a SNP for which the parental context is known may maximize the informativeness of the set of sequencing reads on the mixed sample. In an embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using qPCR to preferentially amplify specific sequences. In an embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using circularizing probes (for example, MIPs) to preferentially amplify specific sequences. In an

embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using a capture by hybridization method (for example SURESELECT) to preferentially amplify specific sequences. Different methods may be used to enhance the number of sequence reads that contain a SNP for which the parent contexts are known. In an embodiment, the targeting may be accomplished by extension ligation, ligation without extension, capture by hybridization, or PCR.

In a sample of fragmented genomic DNA, a fraction of the DNA sequences map uniquely to individual chromosomes; other DNA sequences may be found on different chromosomes. Note that DNA found in plasma, whether maternal or fetal in origin is typically fragmented, often at lengths under 500 bp. In a typical genomic sample, roughly 3.3% of the mappable sequences will map to chromosome 13; 2.2% of the mappable sequences will map to chromosome 18; 1.35% of the mappable sequences will map to chromosome 21; 4.5% of the mappable sequences will map to chromosome X in a female; 2.25% of the mappable sequences will map to chromosome X (in a male); and 0.73% of the mappable sequences will map to chromosome Y (in a male). These are the chromosomes that are most likely to be aneuploid in a fetus. Also, among short sequences, approximately 1 in 20 sequences will contain a SNP, using the SNPs contained on dbSNP. The proportion may well be higher given that there may be many SNPs that have not been discovered.

In an embodiment of the present disclosure, targeting methods may be used to enhance the fraction of DNA in a sample of DNA that map to a given chromosome such that the fraction significantly exceeds the percentages listed above that are typical for genomic samples. In an embodiment of the present disclosure, targeting methods may be used to enhance the fraction of DNA in a sample of DNA such that the percentage of sequences that contain a SNP are significantly greater than what may be found in typical for genomic samples. In an embodiment of the present disclosure, targeting methods may be used to target DNA from a chromosome or from a set of SNPs in a mixture of maternal and fetal DNA for the purposes of prenatal diagnosis.

Note that a method has been reported (U.S. Patent 7,888,017) for determining fetal aneuploidy by counting the number of reads that map to a suspect chromosome and comparing it to the number of reads that map to a reference chromosome, and using the assumption that an overabundance of reads on the suspect chromosome corresponds to a triploidy in the fetus

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 192

at that chromosome. Those methods for prenatal diagnosis would not make use of targeting of any sort, nor do they describe the use of targeting for prenatal diagnosis.

By making use of targeting approaches in sequencing the mixed sample, it may be possible to achieve a certain level of accuracy with fewer sequence reads. The accuracy may refer to sensitivity, it may refer to specificity, or it may refer to some combination thereof. The desired level of accuracy may be between 90% and 95%; it may be between 95% and 98%; it may be between 98% and 99%; it may be between 99% and 99.5%; it may be between 99.5% and 99.9%; it may be between 99.9% and 99.99%; it may be between 99.99% and 99.999%, it may be between 99.999% and 100%. Levels of accuracy above 95% may be referred to as high accuracy.

There are a number of published methods in the prior art that demonstrate how one may determine the ploidy state of a fetus from a mixed sample of maternal and fetal DNA, for example: G.J. W. Liao et al. Clinical Chemistry 2011; 57(1) pp. 92-101. These methods focus on thousands of locations along each chromosome. The number of locations along a chromosome that may be targeted while still resulting in a high accuracy ploidy determination on a fetus, for a given number of sequence reads, from a mixed sample of DNA is unexpectedly low. In an embodiment of the present disclosure, an accurate ploidy determination may be made by using targeted sequencing, using any method of targeting, for example qPCR, ligand mediated PCR, other PCR methods, capture by hybridization, or circularizing probes, wherein the number of loci along a chromosome that need to be targeted may be between 5,000 and 2,000 loci; it may be between 2,000 and 1,000 loci; it may be between 1,000 and 500 loci; it may be between 500 and 300 loci; it may be between 300 and 200 loci; it may be between 200 and 150 loci; it may be between 150 and 100 loci; it may be between 100 and 50 loci; it may be between 50 and 20 loci; it may be between 20 and 10 loci. Optimally, it may be between 100 and 500 loci. The high level of accuracy may be achieved by targeting a small number of loci and executing an unexpectedly small number of sequence reads. The number of reads may be between 100 million and 50 million reads; the number of reads may be between 50 million and 20 million reads; the number of reads may be between 20 million and 10 million reads; the number of reads may be between 10 million and 5 million reads; the number of reads may be between 5 million and 2 million reads; the number of reads may be between 2 million and 1 million; the number of reads may be between 1 million and 500,000; the number of reads may be between 500,000 and 200,000; the number of reads may

be between 200,000 and 100,000; the number of reads may be between 100,000 and 50,000; the number of reads may be between 50,000 and 20,000; the number of reads may be between 20,000 and 10,000; the number of reads may be below 10,000. Fewer number of read are necessary for larger amounts of input DNA.

In some embodiments, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 13 is greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 18 is greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 21 is greater than 2%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome X is greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome Y is greater than 1%, greater than 2%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%.

In some embodiments, a composition is described comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to a chromosome, and that contains at least one single nucleotide polymorphism is greater than

0.2%, greater than 0.3%, greater than 0.4%, greater than 0.5%, greater than 0.6%, greater than 0.7%, greater than 0.8%, greater than 0.9%, greater than 1%, greater than 1.2%, greater than 1.4%, greater than 1.6%, greater than 1.8%, greater than 2%, greater than 2.5%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, or greater than 20%, and where the chromosome is taken from the group 13, 18, 21, X, or Y. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to a chromosome and that contain at least one single nucleotide polymorphism from a set of single nucleotide polymorphisms is greater than 0.15%, greater than 0.2%, greater than 0.3%, greater than 0.4%, greater than 0.5%, greater than 0.6%, greater than 0.7%, greater than 0.8%, greater than 0.9%, greater than 1%, greater than 1.2%, greater than 1.4%, greater than 1.6%, greater than 1.8%, greater than 2%, greater than 2.5%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, or greater than 20%, where the chromosome is taken from the set of chromosome 13, 18, 21, X and Y, and where the number of single nucleotide polymorphisms in the set of single nucleotide polymorphisms is between 1 and 10, between 10 and 20, between 20 and 50, between 50 and 100, between 100 and 200, between 200 and 500, between 500 and 1,000, between 1,000 and 2,000, between 2,000 and 5,000, between 5,000 and 10,000, between 10,000 and 20,000, between 20,000 and 50,000, and between 50,000 and 100,000.

In theory, each cycle in the amplification doubles the amount of DNA present; however, in reality, the degree of amplification is slightly lower than two. In theory, amplification, including targeted amplification, will result in bias free amplification of a DNA mixture; in reality, however, different alleles tend to be amplified to a different extent than other alleles. When DNA is amplified, the degree of allelic bias typically increases with the number of amplification steps. In some embodiments, the methods described herein involve amplifying DNA with a low level of allelic bias. Since the allelic bias compounds with each additional cycle, one can determine the per cycle allelic bias by calculating the nth root of the overall bias where n is the base 2 logarithm of degree of enrichment. In some embodiments, there is a composition comprising a second mixture of DNA, where the second mixture of DNA has been preferentially enriched at a plurality of polymorphic loci from a first mixture of

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 195

DNA where the degree of enrichment is at least 10, at least 100, at least 1,000, at least 10,000, at least 100,000 or at least 1,000,000, and where the ratio of the alleles in the second mixture of DNA at each locus differs from the ratio of the alleles at that locus in the first mixture of DNA by a factor that is, on average, less than 1,000%, 500%, 200%, 100%, 50%, 20%, 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%, 0.05%, 0.02%, or 0.01%. In some embodiments, there is a composition comprising a second mixture of DNA, where the second mixture of DNA has been preferentially enriched at a plurality of polymorphic loci from a first mixture of DNA where the per cycle allelic bias for the plurality of polymorphic loci is, on average, less than 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%, 0.05%, or 0.02%. In some embodiments, the plurality of polymorphic loci comprises at least 10 loci, at least 20 loci, at least 50 loci, at least 100 loci, at least 200 loci, at least 500 loci, at least 1,000 loci, at least 2,000 loci, at least 5,000 loci, at least 10,000 loci, at least 20,000 loci, or at least 50,000 loci.

Some Embodiments

In some embodiments, a method is disclosed herein for generating a report disclosing the determined ploidy status of a chromosome in a gestating fetus, the method comprising: obtaining a first sample that contains DNA from the mother of the fetus and DNA from the fetus; obtaining genotypic data from one or both parents of the fetus; preparing the first sample by isolating the DNA so as to obtain a prepared sample; measuring the DNA in the prepared sample at a plurality of polymorphic loci; calculating, on a computer, allele counts or allele count probabilities at the plurality of polymorphic loci from the DNA measurements made on the prepared sample; creating, on a computer, a plurality of ploidy hypotheses concerning expected allele count probabilities at the plurality of polymorphic loci on the chromosome for different possible ploidy states of the chromosome; building, on a computer, a joint distribution model for allele count probability of each polymorphic locus on the chromosome for each ploidy hypothesis using genotypic data from the one or both parents of the fetus; determining, on a computer, a relative probability of each of the ploidy hypotheses using the joint distribution model and the allele count probabilities calculated for the prepared sample; calling the ploidy state of the fetus by selecting the ploidy state corresponding to the hypothesis with the greatest probability; and generating a report disclosing the determined ploidy status.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 196

In some embodiments, the method is used to determine the ploidy state of a plurality of gestating fetuses in a plurality of respective mothers, the method further comprising: determining the percent of DNA that is of fetal origin in each of the prepared samples; and wherein the step of measuring the DNA in the prepared sample is done by sequencing a number of DNA molecules in each of the prepared samples, where more molecules of DNA are sequenced from those prepared samples that have a smaller fraction of fetal DNA than those prepared samples that have a larger fraction of fetal DNA.

In some embodiments, the method is used to determine the ploidy state of a plurality of gestating fetuses in a plurality of respective mothers, and where the measuring the DNA in the prepared sample is done, for each of the fetuses, by sequencing a first fraction of the prepared sample of DNA to give a first set of measurements, the method further comprising: making a first relative probability determination for each of the ploidy hypotheses for each of the fetuses, given the first set of DNA measurements; resequencing a second fraction of the prepared sample from those fetuses where the first relative probability determination for each of the ploidy hypotheses indicates that a ploidy hypothesis corresponding to an aneuploid fetus has a significant but not conclusive probability, to give a second set of measurements; making a second relative probability determination for ploidy hypotheses for the fetuses using the second set of measurements and optionally also the first set of measurements; and calling the ploidy states of the fetuses whose second sample was resequenced by selecting the ploidy state corresponding to the hypothesis with the greatest probability as determined by the second relative probability determination.

In some embodiments, a composition of matter is disclosed, the composition of matter comprising: a sample of preferentially enriched DNA, wherein the sample of preferentially enriched DNA has been preferentially enriched at a plurality of polymorphic loci from a first sample of DNA, wherein the first sample of DNA consisted of a mixture of maternal DNA and fetal DNA derived from maternal plasma, where the degree of enrichment is at least a factor of 2, and wherein the allelic bias between the first sample and the preferentially enriched sample is, on average, selected from the group consisting of less than 2%, less than 1%, less than 0.5%, less than 0.2%, less than 0.1%, less than 0.05%, less than 0.02%, and less than 0.01%. In some embodiments, a method is disclosed to create a sample of such preferentially enriched DNA.

In some embodiment, a method is disclosed for determining the presence or absence of a fetal aneuploidy in a maternal tissue sample comprising fetal and maternal genomic DNA, wherein the method comprises: (a) obtaining a mixture of fetal and maternal genomic DNA from said maternal tissue sample; (b) selectively enriching the mixture of fetal and maternal DNA at a plurality of polymorphic alleles; (c) distributing selectively enriched fragments from the mixture of fetal and maternal genomic DNA of step a to provide reaction samples comprising a single genomic DNA molecule or amplification products of a single genomic DNA molecule; (d) conducting massively parallel DNA sequencing of the selectively enriched fragments of genomic DNA in the reaction samples of step c) to determine the sequence of said selectively enriched fragments; (e) identifying the chromosomes to which the sequences obtained in step d) belong; (f) analyzing the data of step d) to determine i) the number of fragments of genomic DNA from step d) that belong to at least one first target chromosome that is presumed to be diploid in both the mother and the fetus, and ii) the number of fragments of genomic DNA from step d) that belong to a second target chromosome, wherein said second chromosome is suspected to be aneuploid in the fetus; (g) calculating an expected distribution of the number of fragments of genomic DNA from step d) for the second target chromosome if the second target chromosome is euploid, using the number determined in step f) part i); (h) calculating an expected distribution of the number of fragments of genomic DNA from step d) for the second target chromosome if the second target chromosome is aneuploid, using the first number is step f) part i) and an estimated fraction of fetal DNA found in the mixture of step b); and (i) using a maximum likelihood or maximum a posteriori approach to determine whether the number of fragments of genomic DNA determined in step f) part ii) is more likely to be part of the distribution calculated in step g) or the distribution calculated in step h); thereby indicating the presence or absence of a fetal aneuploidy.

Exemplary Cancer Diagnostic Methods

Note that it has been demonstrated that DNA that originated from cancer that is living in a host can be found in the blood of the host. In the same way that genetic diagnoses can be made from the measurement of mixed DNA found in maternal blood, genetic diagnoses can equally well be made from the measurement of mixed DNA found in host blood. The genetic diagnoses may include aneuploidy states, or gene mutations. Any claim in the instant disclosure that reads on determining the ploidy state or genetic state of a fetus from the

measurements made on maternal blood can equally well read on determining the ploidy state or genetic state of a cancer from the measurements on host blood.

In some embodiments, a method of the present disclosure allows one to determine the ploidy status of a cancer, the method including obtaining a mixed sample that contains genetic material from the host, and genetic material from the cancer; measuring the DNA in the mixed sample; calculating the fraction of DNA that is of cancer origin in the mixed sample; and determining the ploidy status of the cancer using the measurements made on the mixed sample and the calculated fraction. In some embodiments, the method may further include administering a cancer therapeutic based on the determination of the ploidy state of the cancer. In some embodiments, the method may further include administering a cancer therapeutic based on the determination of the ploidy state of the cancer, wherein the cancer therapeutic is taken from the group comprising a pharmaceutical, a biologic therapeutic, and antibody based therapy and combination thereof.

Exemplary Implementation Methods

Any of the embodiments disclosed herein may be implemented in digital electronic circuitry, integrated circuitry, specially designed ASICs (application-specific integrated circuits), computer hardware, firmware, software, or in combinations thereof. Apparatus of the presently disclosed embodiments can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and method steps of the presently disclosed embodiments can be performed by a programmable processor executing a program of instructions to perform functions of the presently disclosed embodiments by operating on input data and generating output. The presently disclosed embodiments can be implemented advantageously in one or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device, and at least one output device. Each computer program can be implemented in a high-level procedural or object-oriented programming language or in assembly or machine language if desired; and in any case, the language can be a compiled or interpreted language. A computer program may be deployed in any form, including as a stand-alone program, or as a module, component, subroutine, or other unit

suitable for use in a computing environment. A computer program may be deployed to be executed or interpreted on one computer or on multiple computers at one site, or distributed across multiple sites and interconnected by a communication network.

Computer readable storage media, as used herein, refers to physical or tangible storage (as opposed to signals) and includes without limitation volatile and non-volatile, removable and non-removable media implemented in any method or technology for the tangible storage of information such as computer-readable instructions, data structures, program modules or other data. Computer readable storage media includes, but is not limited to, RAM, ROM, EPROM, EEPROM, flash memory or other solid state memory technology, CD-ROM, DVD, or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other physical or material medium which can be used to tangibly store the desired information or data or instructions and which can be accessed by a computer or processor.

Any of the methods described herein may include the output of data in a physical format, such as on a computer screen, or on a paper printout. In explanations of any embodiments elsewhere in this document, it should be understood that the described methods may be combined with the output of the actionable data in a format that can be acted upon by a physician. In addition, the described methods may be combined with the actual execution of a clinical decision that results in a clinical treatment, or the execution of a clinical decision to make no action. Some of the embodiments described in the document for determining genetic data pertaining to a target individual may be combined with the decision to select one or more embryos for transfer in the context of IVF, optionally combined with the process of transferring the embryo to the womb of the prospective mother. Some of the embodiments described in the document for determining genetic data pertaining to a target individual may be combined with the notification of a potential chromosomal abnormality, or lack thereof, with a medical professional, optionally combined with the decision to abort, or to not abort, a fetus in the context of prenatal diagnosis. Some of the embodiments described herein may be combined with the output of the actionable data, and the execution of a clinical decision that results in a clinical treatment, or the execution of a clinical decision to make no action.

Exemplary Diagnostic Boxes

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 200

In an embodiment, the present disclosure comprises a diagnostic box that is capable of partly or completely carrying out any of the methods described in this disclosure. In an embodiment, the diagnostic box may be located at a physician's office, a hospital laboratory, or any suitable location reasonably proximal to the point of patient care. The box may be able to run the entire method in a wholly automated fashion, or the box may require one or a number of steps to be completed manually by a technician. In an embodiment, the box may be able to analyze at least the genotypic data measured on the maternal plasma. In an embodiment, the box may be linked to means to transmit the genotypic data measured on the diagnostic box to an external computation facility which may then analyze the genotypic data, and possibly also generate a report. The diagnostic box may include a robotic unit that is capable of transferring aqueous or liquid samples from one container to another. It may comprise a number of reagents, both solid and liquid. It may comprise a high throughput sequencer. It may comprise a computer.

Experimental Section

The presently disclosed embodiments are described in the following Examples, which are set forth to aid in the understanding of the disclosure, and should not be construed to limit in any way the scope of the disclosure as defined in the claims which follow thereafter. The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to use the described embodiments, and are not intended to limit the scope of the disclosure nor are they intended to represent that the experiments below are all or the only experiments performed. Efforts have been made to ensure accuracy with respect to numbers used (*e.g.* amounts, temperature, etc.) but some experimental errors and deviations should be accounted for. Unless indicated otherwise, parts are parts by volume, and temperature is in degrees Centigrade. It should be understood that variations in the methods as described may be made without changing the fundamental aspects that the experiments are meant to illustrate.

Experiment 1

The objective was to show that a Bayesian maximum likelihood estimation (MLE) algorithm that uses parent genotypes to calculate fetal fraction improves accuracy of non-invasive prenatal trisomy diagnosis compared to published methods.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 201

Simulated sequencing data for maternal cfDNA was created by sampling reads obtained on trisomy-21 and respective mother cell lines. The rate of correct disomy and trisomy calls were determined from 500 simulations at various fetal fractions for a published method (Chiu et al. BMJ 2011;342:c7401) and our MLE-based algorithm. We validated the simulations by obtaining 5 million shotgun reads from four pregnant mothers and respective fathers collected under an IRB-approved protocol. Parental genotypes were obtained on a 290K SNP array. (See **Figure 14**)

In simulations, the MLE-based approach achieved 99.0% accuracy for fetal fractions as low as 9% and reported confidences that corresponded well to overall accuracy. We validated these results using four real samples wherein we obtained all correct calls with a computed confidence exceeding 99%. In contrast, our implementation of the published algorithm for Chiu et al. required 18% fetal fraction to achieve 99.0% accuracy, and achieved only 87.8% accuracy at 9% fetal DNA.

Fetal fraction determination from parental genotypes in conjunction with a MLE-based approach achieves greater accuracy than published algorithms at the fetal fractions expected during the 1st and early 2nd trimester. Furthermore, the method disclosed herein produces a confidence metric that is crucial in determining the reliability of the result, especially at low fetal fractions where ploidy detection is more difficult. Published methods use a less accurate threshold method for calling ploidy based on large sets of disomy training data, an approach that predefines a false positive rate. In addition, without a confidence metric, published methods are at risk of reporting false negative results when there is insufficient fetal cfDNA to make a call. In some embodiments, a confidence estimate is calculated for the called ploidy state.

Experiment 2

The objective was to improve non-invasive detection of fetal trisomy 18, 21, and X particularly in samples consisting of low fetal fraction by using a targeted sequencing approach combined with parent genotypes and Hapmap data in a Bayesian Maximum Likelihood Estimation (MLE) algorithm.

Maternal samples from four euploid and two trisomy-positive pregnancies and respective paternal samples were obtained under an IRB-approved protocol from patients where fetal karyotype was known. Maternal cfDNA was extracted from plasma and roughly 10

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 202

million sequence reads were obtained following preferential enrichment that targeted specific SNPs. Parent samples were similarly sequenced to obtain genotypes.

The described algorithm correctly called chromosome 18 and 21 disomy for all euploid samples and normal chromosomes of aneuploid samples. Trisomy 18 and 21 calls were correct, as were chromosome X copy numbers in male and female fetuses. The confidence produced by the algorithm was in excess of 98% in all cases.

The method described accurately reported the ploidy of all tested chromosomes from six samples, including samples comprised of less than 12% fetal DNA, which account for roughly 30% of 1st and early 2nd-trimester samples. The crucial difference between the instant MLE algorithm and published methods is that it leverages parent genotypes and Hapmap data to improve accuracy and generate a confidence metric. At low fetal fractions, all methods become less accurate; it is important to correctly identify samples without sufficient fetal cfDNA to make a reliable call. Others have used chromosome Y specific probes to estimate fetal fraction of male fetuses, but concurrent parental genotyping enables estimation of fetal fraction for both sexes. Another inherent limitation of published methods using untargeted shotgun sequencing is that accuracy of ploidy calling varies among chromosomes due to differences in factors such as GC richness. The instant targeted sequencing approach is largely independent of such chromosome-scale variations and yields more consistent performance between chromosomes.

Experiment 3

The objective was to determine if trisomy is detectable with high confidence on a triploid fetus, using novel informatics to analyze SNP loci of free floating fetal DNA in maternal plasma.

20mL of blood was drawn from a pregnant patient following abnormal ultrasound. After centrifugation, maternal DNA was extracted from the buffy coat (DNEASY, QIAGEN); cell-free DNA was extracted from plasma (QIAAMP QIAGEN). Targeted sequencing was applied to SNP loci on chromosomes 2, 21, and X in both DNA samples. Maximum-Likelihood Bayesian estimation selected the most likely hypothesis from the set of all possible ploidy states. The method determines fetal DNA fraction, ploidy state and explicit confidences in the ploidy determination. No assumptions are made about the ploidy of a reference

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 203

chromosome. The diagnostic uses a test statistic that is independent of sequence read counts, which is the recent state of the art.

The instant method accurately diagnosed trisomy of chromosomes 2 and 21. Child fraction was estimated at 11.9% [CI 11.7-12.1]. The fetus was found to have one maternal and two paternal copies of chromosomes 2 and 21 with confidence of effectively 1 (error probability$<10^{-30}$). This was achieved with 92,600 and 258,100 reads on chromosomes 2 and 21 respectively.

This is the first demonstration of non-invasive prenatal diagnosis of trisomic chromosomes from maternal blood where the fetus was triploid, as confirmed by metaphase karyotype. Extant methods of non-invasive diagnosis would not detect aneuploidy in this sample. Current methods rely on a surplus of sequence reads on a trisomic chromosome relative to disomic reference chromosomes; but a triploid fetus has no disomic reference. Furthermore, extant methods would not achieve similarly high-confidence ploidy determination with this fraction of fetal DNA and number of sequence reads. It is straightforward to extend the approach to all 24 chromosomes.

Experiment 4

The following protocol was used for 800-plex amplification of DNA isolated from maternal plasma from a euploid pregnancy and also genomic DNA from a triploidy 21 cell line using standard PCR (meaning no nesting was used). Library preparation and amplification involved single tube blunt ending followed by A-tailing. Adaptor ligation was run using the ligation kit found in the AGILENT SURESELECT kit, and PCR was run for 7 cycles. Then, 15 cycles of STA (95˚C for 30s; 72˚C for 1 min; 60˚C for 4 min; 65˚C for 1 min; 72˚C for 30s) using 800 different primer pairs targeting SNPs on chromosomes 2, 21 and X. The reaction was run with 12.5 nM primer concentration. The DNA was then sequenced with an ILLUMINA IIGAX sequencer. The sequencer output 1.9 million reads, of which 92% mapped to the genome; of those reads that mapped to the genome, more than 99% mapped to one of the regions targeted by the targeted primers. The numbers were essentially the same for both the plasma DNA and the genomic DNA. **Figure 15** shows the ratio of the two alleles for the ~780 SNPs that were detected by the sequencer in the genomic DNA that was taken from a cell line with known trisomy at chromosome 21. Note that the allele ratios are plotted here for ease of visualization, because the allele distributions are not straightforward to read visually. The

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 204

circles represent SNPs on disomic chromosomes, while the stars represent SNPs on a trisomic chromosome. **Figure 16** is another representation of the same data as in Figure X, where the Y-axis is the relative number of A and B measured for each SNP, and where the X-axis is the SNP number where the SNPs are separated by chromosome. In **Figure 16**, SNP 1 to 312 are found on chromosome 2, from SNP 313 to 605 are found on chromosome 21 which is trisomic, and from SNP 606 to 800 are on chromosome X. The data from chromosomes 2 and X show a disomic chromosome, as the relative sequence counts lie in three clusters: AA at the top of the graph, BB at the bottom of the graph, and AB in the middle of the graph. The data from chromosome 21, which is trisomic, shows four clusters: AAA at the top of the graph, AAB around the 0.65 line (2/3), ABB around the .35 line (1/3), and BBB at the bottom of the graph.

Figures 17A-D show data for the same 800-plex protocol, but measured on DNA that was amplified from four plasma samples from pregnant women. For these four samples, we expect to see seven clusters of dots: (1) along the top of the graph are those loci where both the mother and the fetus are AA, (2) slightly below the top of the graph are those loci where the mother is AA and the fetus is AB, (3) slightly above the 0.5 line are those loci where the mother is AB and the fetus is AA, (4) along the 0.5 line are those loci where the mother and the fetus are both AB, (5) slightly below the 0.5 line are those loci where the mother is AB and the fetus is BB, (6) slightly above the bottom of the graph are those loci where the mother is BB and the fetus is AB, (1) along the bottom of the graph are those loci where both the mother and the fetus are BB. The smaller the fetal fraction, the less the separation between clusters (1) and (2), between clusters (3), (4) and (5), and between clusters (6) and (7). The separation is expected to be half of the fraction of DNA that is of fetal origin. For example if the DNA is 20% fetal, and 80% maternal, we expect (1) through (7) to be centered at 1.0, 0.9, 0.6, 0.5, 0.4, 0.1 and 0.0 respectively; see for example **Figure 17D**, POOL1_BC5_ref_rate. If, instead the DNA is 8% fetal, and 92% maternal, we expect (1) through (7) to be centered at 1.00, 0.96, 0.54, 0.50, 0.46, 0.04 and 0.00 respectively; see for example **Figure 17B**, POOL1_BC2_ref_rate. If there is not fetal DNA detected, we do not expect to see (2), (3), (5), or (6); alternately we could say that the separation is zero, and therefore (1) and (2) are on top of each other, as are (3), (4) and (5), and also (6) and (7); see *e.g.* **Figure 17C**, POOL1_BC7_ref_rate. Note that the fetal fraction for **Figure 17A**, POOL1_BC1_ref_rate is about 25%.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 205

Experiment 5

Most methods of DNA amplification and measurement will produce some allele bias, wherein the two alleles that are typically found at a locus are detected with intensities or counts that are not representative of the actual amounts of alleles in the sample of DNA. For example, for a single individual, at a heterozygous locus we expect to see a 1:1 ratio of the two alleles, which is the theoretical ratio expected for a heterozygous locus; however due to allele bias, we may see 55:45, or even 60:40. Also note that in the context of sequencing, if the depth of read is low, then simple stochastic noise could result in significant allele bias. In an embodiment, it is possible to model the behavior of each SNP such that if a consistent bias is observed for particular alleles, this bias can be corrected for. **Figure 18** shows the fraction of data that can be explained by binomial variance, before and after bias correction. In **Figure 18**, the stars represent the observed allele bias on raw sequence data for the 800-plex experiment; the circles represent the allele bias after correction. Note that if there were no allele bias at all, we would expect the data to fall along the x=y line. A similar set of data that was produced by amplifying DNA using a 150-plex targeted amplification produced data that fell very closely on the 1:1 line after bias correction.

Experiment 6

Universal amplification of DNA using ligated adaptors with primers specific to the adaptor tags, where the primer annealing and extension times are limited to a few minutes has the effect of enriching the proportion of shorter DNA strands. Most library protocols designed for creating DNA libraries suitable for sequencing contain such a step, and example protocols are published and well known to those in the art. In some embodiments of the invention, adaptors with a universal tag are ligated to the plasma DNA, and amplified using primers specific to the adaptor tag. In some embodiments, the universal tag can be the same tag as used for sequencing, it can be a universal tag only for PCR amplification, or it can be a set of tags. Since the fetal DNA is typically short in nature, while the maternal DNA can be both short and long in nature, this method has the effect of enriching the proportion of fetal DNA in the mixture. The free floating DNA, thought to be DNA from apoptotic cells, and which contains both fetal and maternal DNA, is short – mostly under 200 bp. Cellular DNA released by cell lysis, a common phenomenon after phlebotomy, is typically almost exclusively maternal, and

is also quite long – mostly above 500 bp. Therefore, blood samples that have sat around for more than a few minutes will contain a mixture of short (fetal + maternal) and longer (maternal) DNA. Performing a universal amplification with relatively short extension times on maternal plasma followed by targeted amplification will tend to increase the relative proportion of fetal DNA when compared to the plasma that has been amplified using targeted amplification alone. This can be seen in **Figure 19** which shows the measured fetal percent when the input is plasma DNA (vertical axis) vs. the measured fetal percent when the input DNA is plasma DNA that has had a library prepared using the ILLUMINA GAIIx library preparation protocol. All the dots fall below the line, indicating that the library preparation step enriches the fraction of DNA that is of fetal origin. Two samples of plasma that were red, indicating hemolysis and therefore that there would be an increased amount of long maternal DNA present from cell lysis, show a particularly significant enrichment of fetal fraction when the library preparation is performed prior to targeted amplification. The method disclosed herein is particularly useful in cases where there is hemolysis or some other situation has occurred where cells comprising relatively long strands of contaminating DNA have lysed, contaminating the mixed sample of short DNA with the long DNA. Typically the relatively short annealing and extension times are between 30 seconds and 2 minutes, though they could be as short as 5 or 10 seconds or less, or as long as 5 or 10 minutes.

Experiment 7

The following protocol was used for 1,200-plex amplification of DNA isolated from maternal plasma from a euploid pregnancy and also genomic DNA from a triploidy 21 cell line using a direct PCR protocol, and also a semi-nested approach. Library preparation and amplification involved single tube blunt ending followed by A-tailing. Adaptor ligation was run using a modification of the ligation kit found in the AGILENT SURESELECT kit, and PCR was run for 7 cycles. In the targeted primer pool, there were 550 assays for SNPs from chromosome 21, and 325 assays for SNPs from each of chromosomes 1 and X. Both protocols involved 15 cycles of STA (95˚C for 30s; 72˚C for 1 min; 60˚C for 4 min; 65˚C for 30s; 72˚C for 30s) using 16 nM primer concentration. The semi-nested PCR protocol involved a second amplification of 15 cycles of STA (95˚C for 30s; 72˚C for 1 min; 60˚C for 4 min; 65˚C for 30s; 72˚C for 30s) using an inner forward tag concentration of 29 nM, and a reverse tag concentration of 1 uM or 0.1 uM. The DNA was then sequenced with an ILLUMINA IIGAX

sequencer. For the direct PCR protocol, 73% of the reads map to the genome; for the semi-nested protocol, 97.2% of the sequence reads map to the genome. Therefore, the semi-nested protocol result in approximately 30% more information, presumably mostly due to the elimination of primers that are most likely to cause primer dimers.

The depth of read variability tends to be higher when using the semi-nested protocol than when the direct PCR protocol is used (see **Figure 20**) where the diamonds refer to the depth of read for loci run with the semi-nested protocol, and the squares refer to the depth of read for loci run with no nesting. The SNPs are arranged by depth of read for the diamonds, so the diamonds all fall on a curved line, while the squares appear to be loosely correlated; the arrangements of the SNPs is arbitrary, and it is the height of the dot that denotes depth of read rather than its location left to right.

In some embodiments, the methods described herein can achieve excellent depth of read (DOR) variances. For example, in one version of this experiment (**Figure 21**) using a 1,200-plex direct PCR amplification of genomic DNA, of the 1,200 assays: 1186 assays had a DOR greater than 10; the average depth of read was 400; 1063 assays (88.6%) had a depth of read of between 200 and 800, and ideal window where the number of reads for each allele is high enough to give meaningful data, while the number of reads for each allele is not so high that the marginal use of those reads was particularly small. Only 12 alleles had higher depth of read with the highest at 1035 reads. The standard deviation of the DOR was 290, the average DOR was 453, the coefficient of variance of the DOR was 64%, there were 950,000 total reads, and 63.1% of the reads mapped to the genome. In another experiment (**Figure 22**) using a 1,200-plex semi-nested protocol, the DOR was higher. The standard deviation of the DOR was 583, the average DOR was 630, the coefficient of variance of the DOR was 93%, there were 870,000 total reads, and 96.3% of the reads mapped to the genome. Note, in both these cases, the SNPs are arranged by the depth of read for the mother, so the curved line represents the maternal depth of read. The differentiation between child and father is not significant; it is only the trend that is significant for the purpose of this explanation.

Experiment 8

In an experiment, the semi-nested 1,200-plex PCR protocol was used to amplify DNA from one cell and from three cells. This experiment is relevant to prenatal aneuploidy testing using fetal cells isolated from maternal blood, or for preimplantation genetic diagnosis using

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 208

biopsied blastomeres or trophectoderm samples. There were 3 replicates of 1 and 3 cells from 2 individuals (46 XY and 47 XX+21) per condition. Assays targeted chromosomes 1, 21 and X. Three different lysis methods were used: ARCTURUS, MPERv2 and Alkaline lysis. Sequencing was run multiplexing 48 samples in one sequencing lane. The algorithm returned correct ploidy calls for each of the three chromosomes, and for each of the replicates.

Experiment 9

In one experiment, four maternal plasma samples were prepared and amplified using a hemi-nested 9,600-plex protocol. The samples were prepared in the following way: Up to 40 mL of maternal blood were centrifuged to isolate the buffy coat and the plasma. The genomic DNA in the maternal sample was prepared from the buffy coat and paternal DNA was prepared from a blood sample or saliva sample. Cell-free DNA in the maternal plasma was isolated using the QIAGEN CIRCULATING NUCLEIC ACID kit and eluted in 45 uL TE buffer according to manufacturer's instructions. Universal ligation adapters were appended to the end of each molecule of 35 uL of purified plasma DNA and libraries were amplified for 7 cycles using adaptor specific primers. Libraries were purified with AGENCOURT AMPURE beads and eluted in 50 ul water.

3 ul of the DNA was amplified with 15 cycles of STA (95˚C for 10 min for initial polymerase activation, then 15 cycles of 95˚C for 30s; 72˚C for 10 s; 65˚C for 1 min; 60˚C for 8 min; 65˚C for 3 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using 14.5 nM primer concentration of 9600 target-specific tagged reverse primers and one library adaptor specific forward primer at 500 nM.

The hemi-nested PCR protocol involved a second amplification of a dilution of the first STAs product for 15 cycles of STA (95˚C for 10 min for initial polymerase activation, then 15 cycles of 95˚C for 30s; 65˚C for 1 min; 60˚C for 5 min; 65˚C for 5 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using reverse tag concentration of 1000 nM, and a concentration of 16.6 u nM for each of 9600 target-specific forward primers.

An aliquot of the STA products was then amplified by standard PCR for 10 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 209

In this way, 9,600 primers were used in the single-well reactions; the primers were designed to target SNPs found on chromosomes 1, 2, 13, 18, 21, X and Y. The amplicons were then sequenced using an ILLUMINA GAIIX sequencer. Per sample, approximately 3.9 million reads were generated by the sequencer, with 3.7 million reads mapping to the genome (94%), and of those, 2.9 million reads (74%) mapped to targeted SNPs with an average depth of read of 344 and a median depth of read of 255. The fetal fraction for the four samples was found to be 9.9%, 18.9%, 16.3%, and 21.2%

Relevant maternal and paternal genomic DNA samples amplified using a semi-nested 9600-plex protocol and sequenced. The semi-nested protocol is different in that it applies 9,600 outer forward primers and tagged reverse primers at 7.3 nM in the first STA. Thermocycling conditions and composition of the second STA, and the barcoding PCR were the same as for the hemi-nested protocol.

The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at six chromosomes for the fetuses whose DNA was present in the 4 maternal plasma samples. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 99.2% except for one chromosome that was called correctly, but with a confidence of 83%.

Figure 23 shows the depth of read of the 9,600-plex hemi-nesting approach along with the depth of read of the 1,200-plex semi-nested approach described in Experiment 7, though the number of SNPs with a depth of read greater than 100, greater than 200 and greater than 400 was significantly higher than in the 1,200-plex protocol. The number of reads at the 90th percentile can be divided by the number of reads at the 10th percentile to give a dimensionless metric that is indicative of the uniformity of the depth of read; the smaller the number, the more uniform (narrow) the depth of read. The average 90th percentile/10th percentile ratio is 11.5 for the method run in Experiment 9, while it is 5.6 for the method run in Experiment 7. A narrower depth of read for a given protocol plexity is better for sequencing efficiency, as fewer sequence reads are necessary to ensure that a certain percentage of reads are above a read number threshold.

Experiment 10

In one experiment, four maternal plasma samples were prepared and amplified using a semi-nested 9,600-plex protocol. Details of Experiment 10 were very similar to Experiment 9,

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 210

the exception being the nesting protocol, and including the identity of the four samples. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 99.7%. 7.6 million (97%) of reads mapped to the genome, and 6.3 million (80%) of the reads mapped to the targeted SNPs. The average depth of read was 751, and the median depth of read was 396.

Experiment 11

In one experiment, three maternal plasma samples were split into five equal portions, and each portion was amplified using either 2,400 multiplexed primers (four portions) or 1,200 multiplexed primers (one portion) and amplified using a semi-nested protocol, for a total of 10,800 primers. After amplification, the portions were pooled together for sequencing. Details of Experiment 11 were very similar to Experiment 9, the exception being the nesting protocol, and the split and pool approach. The ploidy calls for all 21 chromosomes in the set were called correctly with confidences above 99.7%, except for one missed call where the confidence was 83%. 3.4 million reads mapped to targeted SNPs, the average depth of read was 404 and the median depth of read was 258.

Experiment 12

In one experiment, four maternal plasma samples were split into four equal portions, and each portion was amplified using 2,400 multiplexed primers and amplified using a semi-nested protocol, for a total of 9,600 primers. After amplification, the portions were pooled together for sequencing. Details of Experiment 12 were very similar to Experiment 9, the exception being the nesting protocol, and the split and pool approach. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 97%, except for one missed call where the confidence was 78%. 4.5 million reads mapped to targeted SNPs, the average depth of read was 535 and the median depth of read was 412.

Experiment 13

In one experiment, four maternal plasma samples were prepared and amplified using a 9,600-plex triply hemi-nested protocol, for a total of 9,600 primers. Details of Experiment 12 were very similar to Experiment 9, the exception being the nesting protocol which involved three rounds of amplification; the three rounds involved 15, 10 and 15 STA cycles

respectively. The ploidy calls for 27 of 28 chromosomes in the set were called correctly with confidences above 99.9%, except for one that was called correctly with 94.6%, and one missed call with a confidence of 80.8%. 3.5 million reads mapped to targeted SNPs, the average depth of read was 414 and the median depth of read was 249.

Experiment 14

In one experiment 45 sets of cells were amplified using a 1,200-plex semi-nested protocol, sequenced, and ploidy determinations were made at three chromosomes. Note that this experiment is meant to simulate the conditions of performing pre-implantation genetic diagnosis on single-cell biopsies from day 3 embryos, or trophectoderm biopsies from day 5 embryos. 15 individual single cells and 30 sets of three cells were placed in 45 individual reaction tubes for a total of 45 reactions where each reaction contained cells from only one cell line, but the different reactions contained cells from different cell lines. The cells were prepared into 5 ul washing buffer and lysed the by adding 5 ul ARCTURUS PICOPURE lysis buffer (APPLIED BIOSYSTEMS) and incubating at 56˚C for 20 min, 95˚C for 10 min..

The DNA of the single/three cells was amplified with 25 cycles of STA (95˚C for 10 min for initial polymerase activation, then 25 cycles of 95˚C for 30s; 72˚C for 10 s; 65˚C for 1 min; 60˚C for 8 min; 65˚C for 3 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using 50 nM primer concentration of 1200 target-specific forward and tagged reverse primers.

The semi-nested PCR protocol involved three parallel second amplification of a dilution of the first STAs product for 20 cycles of STA (95˚C for 10 min for initial polymerase activation, then 15 cycles of 95˚C for 30s; 65˚C for 1 min; 60˚C for 5 min; 65˚C for 5 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using reverse tag specific primer concentration of 1000 nM, and a concentration of 60 nM for each of 400 target-specific nested forward primers. In the three parallel 400-plex reactions the total of 1200 targets amplified in the first STA were thus amplified.

An aliquot of the STA products was then amplified by standard PCR for 15 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 212

In this way, 1,200 primers were used in the single cell reactions; the primers were designed to target SNPs found on chromosomes 1, 21 and X. The amplicons were then sequenced using an ILLUMINA GAIIX sequencer. Per sample, approximately 3.9 million reads were generated by the sequencer, with 500,000 to 800,000 million reads mapping to the genome (74% to 94% of all reads per sample).

Relevant maternal and paternal genomic DNA samples from cell lines were analyzed using the same semi-nested 1200-plex assay pool with a similar protocol with fewer cycles and 1200-plex second STA, and sequenced.

The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at the three chromosomes for the samples.

Figure 24 shows normalized depth of read ratios (vertical axis) for six samples at three chromosomes (1 = chrom 1; 2 = chrom 21; 3 = chrom X). The ratios were set to be equal to the number of reads mapping to that chromosome, normalized, and divided by the number of reads mapping to that chromosome averaged over three wells each comprising three 46XY cells. The three sets of data points corresponding to the 46XY reactions are expected to have ratios of 1:1. The three sets of data points corresponding to the 47XX+21 cells are expected to have ratios of 1:1 for chromosome 1, 1.5:1 for chromosome 21, and 2:1 for chromosome X.

Figure 25 shows allele ratios plotted for three chromosomes (1, 21, X) for three reaction. The reaction in the lower left shows a reaction on three 46XY cells. The left region are the allele ratios for chromosome 1, the middle region are the allele ratios for chromosome 21, and the right region are the allele ratios for chromosome X. For the 46XY cells, for chromosome 1 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 46XY cells, for chromosome 21 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 46XY cells, for chromosome X we expect to see ratios of 1 and 0, corresponding to A, and B SNP genotypes. The reaction in the lower right shows a reaction on three 47XX+21 cells. The allele ratios are segregated by chromosome as in the lower left graph. For the 47XX+21 cells, for chromosome 1 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 47XX+21 cells, for chromosome 21 we expect to see ratios of 1, 0.67, 0.33 and 0, corresponding to AAA, AAB, ABB and BBB SNP genotypes. For the 47XX+21 cells, for chromosome X we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB, and BB SNP genotypes. The plot in the upper right was made on a reaction comprising 1 ng of genomic

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 213

DNA from the 47XX+21 cell line. **Figure 26** shows the same graphs as in Figure 25, but for reactions performed on only one cell. The left graph was a reaction that contained a 47XX+21 cell, and the right graph was for a reaction that contained a 46XX cell.

From the graphs shown in **Figure 25** and **Figure 26**, it is visually apparent that there are two clusters of dots for chromosomes where we expect to see ratios of 1 and 0; three clusters of dots for chromosomes where we expect to see ratios of 1, 0.5, and 0, and four clusters of dots for chromosomes where we expect to see ratios of 1, 0.67, 0.33 and 0. The parental support algorithm was able to make correct calls on all of the three chromosomes for all of the 45 reactions.

Experiment 15

In one experiment, maternal plasma samples were prepared and amplified using a hemi-nested 19,488-plex protocol. The samples were prepared in the following way: up to 20 mL of maternal blood were centrifuged to isolate the buffy coat and the plasma. The genomic DNA in the maternal sample was prepared from the buffy coat and paternal DNA was prepared from a blood sample or saliva sample. Cell-free DNA in the maternal plasma was isolated using the QIAGEN CIRCULATING NUCLEIC ACID kit and eluted in 50 uL TE buffer according to manufacturer's instructions. Universal ligation adapters were appended to the end of each molecule of 40 uL of purified plasma DNA and libraries were amplified for 9 cycles using adaptor specific primers. Libraries were purified with AGENCOURT AMPURE beads and eluted in 50 ul DNA suspension buffer.

6 ul of the DNA was amplified with 15 cycles of STAR 1 (95˚C for 10 min for initial polymerase activation, then 15 cycles of 96˚C for 30s; 65˚C for 1 min; 58˚C for 6 min; 60˚C for 8 min; 65˚C for 4 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using 7.5 nM primer concentration of 19,488 target-specific tagged reverse primers and one library adaptor specific forward primer at 500 nM.

The hemi-nested PCR protocol involved a second amplification of a dilution of the STAR 1 product for 15 cycles (STAR 2) (95˚C for 10 min for initial polymerase activation, then 15 cycles of 95˚C for 30s; 65˚C for 1 min; 60˚C for 5 min; 65˚C for 5 min and 72˚C for 30s; and a final extension at 72˚C for 2 min) using reverse tag concentration of 1000 nM, and a concentration of 20 nM for each of 19,488 target-specific forward primers.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 214

An aliquot of the STAR 2 products was then amplified by standard PCR for 12 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

In this way, 19,488 primers were used in the single-well reactions; the primers were designed to target SNPs found on chromosomes 1, 2, 13, 18, 21, X and Y. The amplicons were then sequenced using an ILLUMINA GAIIX sequencer. For plasma samples, approximately 10 million reads were generated by the sequencer, with 9.4-9.6 million reads mapping to the genome (94-96 %), and of those, 99.95 % mapped to targeted SNPs with a mean depth of read of 460 and a median depth of read of 350. For comparison, a perfectly even distribution would be: 10M reads / 19,488 targets = 513 reads/target. For primer-dimers, 30,000 reads were from sequenced primer-dimers (0.3% of the reads generated by the sequencer). For genomic samples, 99.4-99.7% of the reads mapped to the genome, of those, 99.99% of the mapped to targeted SNPs, and 0.1% of the reads generated by the sequencer were primer-dimers.

For plasma samples with 10 million sequencing reads, typically at least 19,350 of the 19,488 targeted SNPs (99.3 %) are amplified and sequenced. For DNA samples with 2M sequencing reads, typically at least 19,000 targeted SNPs (97.5%) are amplified and sequenced. The lower number may be due to sampling noise since the number of reads is lower and the sequencer misses some of the amplified products. If desired, the number of sequencing reads can be increased to increase the number of targeted SNPs that are amplified and sequenced.

Relevant maternal and paternal genomic DNA samples amplified using a semi-nested 19,488 outer forward primers and tagged reverse primers at 7.5 nM in the STAR 1. Thermocycling conditions and composition of STAR 2, and the barcoding PCR were the same as for the hemi-nested protocol.

The average fetal fraction for 407 samples was found to be 14.8%. The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at four chromosomes (13, 18, 21, Y) for the fetuses whose DNA was present in 378 of the 407 maternal plasma samples, and at chromosome X in 375 of the 407 maternal plasma samples. The ploidy calls for all 1,887 chromosomes in the set were called correctly with confidences above 90%. 1882 of the 1887 calls were above 95%; and 1,862 of the 1,887 calls were called with confidences above 99%.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 215

A similar control experiment was performed using water instead of DNA extracted from plasma in the plasma PCR protocol. Based on six such trials of an experiment, 5-6% of the sequenced reads were primer-dimers. Other sequenced reads were due to background noise. This experiment demonstrates that even in the absence of a nucleic acid sample with target loci for the primers to hybridize to (rather than hybridizing to other primers and forming amplified primer dimers) few primer dimers are formed.

Experiment 16

The following experiment illustrates an exemplary method for designing and selecting a library of primers that can be used in any of the multiplexed PCR methods of the invention. The goal is to select primers from an initial library of candidate primers that can be used to simultaneously amplify a large number of target loci (or a subset of target loci) in a single reaction. For an initial set of candidate target loci, primers did not have to be designed or selected for each target locus. Preferably, primers are designed and selected for a large portion of the most desirable target loci.

Step 1

A set of candidate target loci (such as SNPs) were selected based on publically available information about desired parameters for the target loci, such as frequency of the SNPs within a target population or heterozygosity rate of the SNPs (worldwide web at ncbi.nlm.nih.gov/projects/SNP/; Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308-11, which are each incorporated by reference in its entirety). For each candidate locus, one or more PCR primer pairs were designed using the Primer3 program (the worldwide web at primer3.sourceforge.net; libprimer3 release 2.2.3, which is hereby incorporated by reference in its entirety). If there were no feasible designs for PCR primers for a particular target locus, then that target locus was eliminated from further consideration.

If desired, a "target locus score" (higher score representing higher desirability) can be calculated for most or all of the target loci, such as a target locus score calculated based on a weighted average of various desired parameters for the target loci. The parameters may be assigned different weights based on their importance for the particular application that the primers will be used for. Exemplary parameters include the heterozygosity rate of the target

locus, the disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, the disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, the specificity of the candidate primer(s) used to amplify the target locus, the size of the candidate primer(s) used to amply the target locus, and the size of the target amplicon.

Step 2

A thermodynamic interaction score was calculated between each primer and all primers for all other target loci from Step 1 (*see, e.g.*, Allawi, H. T. & SantaLucia, J., Jr. (1998), "Thermodynamics of Internal C-T Mismatches in DNA", *Nucleic Acids Res. 26*, 2694-2701; Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J., Jr. (1999), "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches", *Biochemistry 38*, 3468-3477; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA: Sequence Dependence and pH Effects", *Biochemistry 37*, 9435-9444.; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest Neighbor Thermodynamic Parameters for Internal G-A Mismatches in DNA", *Biochemistry 37*, 2170-2179; and Allawi, H. T. & SantaLucia, J., Jr. (1997), "Thermodynamics and NMR of Internal G-T Mismatches in DNA", *Biochemistry 36*, 10581-10594; MultiPLX 2.1 (Kaplinski L, Andreson R, Puurand T, Remm M. MultiPLX: automatic grouping and evaluation of PCR primers. Bioinformatics. 2005 Apr 15;21(8):1701-2, which are each hereby incorporated by reference in its entirety). This step resulted in a 2D matrix of interaction scores. The interaction score predicted the likelihood of primer-dimers involving the two interacting primers. The score was calculated as follows:

interaction_score = max(- deltaG_2, 0.8 * (- deltaG_1))

where

deltaG_2 = Gibbs energy (energy required to break the dimer) for a dimer that is extensible by PCR on both ends, *i.e,.* the 3' end of each primer anneals to the other primer; and

deltaG_1 = Gibbs energy for a dimer that is extensible by PCR on at least one end.

Step 3:

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 217

For each target locus, if there was more than one primer-pair design, then one design was selected using the following method:

1 For each primer-pair design for the locus, find the worst-case (highest) interaction score for the two primers in that design and all primers from all designs for all other target loci.

2 Pick the design with the best (lowest) worst-case interaction score.

Step 4

A graph was built such that each node represented one locus and its associated primer-pair design (*e.g.*, a Maximal Clique problem). One edge was created between every pair of nodes. A weight was assigned to each edge equal to the worst-case (highest) interaction score between the primers associated with the two nodes connected by the edge.

Step 5

If desired, for every pair of designs for two different target loci where one of the primers from one design and one of the primers from the other design would anneal to overlapping target regions, an additional edge was added between the nodes for the two design. The weight of these edges was set equal to the highest weight assigned in Step 4. Thus, Step 5 prevents the library from having primers that would anneal to overlapping target regions, and thus interfere with each other during a multiplex PCR reaction.

Step 6

An initial interaction score threshold was calculated as follows:

weight_threshold = max(edge_weight) - 0.05 * (max(edge_weight) - min(edge_weight))

where

max(edge_weight) is the maximum edge weight in the graph; and

min(edge_weight) is the minimum edge weight in the graph.

The initial bounds for the threshold were set as follows:

max_weight_threshold = max(edge_weight)

min_weight_threshold = min(edge_weight)

Step 7

A new graph was constructed consisting of the same set of nodes as the graph from Step 5, only including edges with weights that exceed weight_threshold. Thus, step ignores interactions with scores equal to or below weight_threshold.

Step 8

Nodes (and all of the edges connected to the removed nodes) were removed from the graph of Step 7 until there were no edges left. Nodes were removed by applying the following procedure repeatedly:

1 Find the node with the highest degree (highest number of edges). If there is more than one then pick one arbitrarily.

2 Define the set of nodes consisting of the node picked above and all of the nodes connected to it, but excluding any nodes that have degree less than the node picked above.

3 Choose the node from the set that has the lowest target locus score (lower score representing lower desirability) from Step 1. Remove that node from the graph.

Step 9

If the number of nodes remaining in the graph satisfies the required number of target loci for the multiplexed PCR pool (within an acceptable tolerance), then the method was continued at Step 10.

If there were too many or too few nodes remaining in the graph, then a binary search was performed to determine what threshold values would result in the desired number of nodes remaining in the graphs. If there were too many nodes in the graph then, the weight threshold bounds were adjusted as follows:

max_weight_threshold = weight_threshold

Otherwise (if there are two few nodes in the graph), then the weight threshold bounds were adjusted as follows:

min_weight_threshold = weight_threshold

Then, the weight threshold was adjusted follows:

weight_threshold = (max_weight_threshold + min_weight_threshold) / 2

Steps 7-9 were repeated.

Step 10

The primer-pair designs associated with the nodes remaining in the graph were selected for the library of primers. This primer library can be used in any of the methods of the invention.

If desired, this method of designing and selecting primers can be performed for primer libraries in which only one primer (instead of a primer pair) is used for amplification of a target locus. In this case, a node presents one primer per target locus (rather than a primer pair).

Experiment 17

Figure 27 is a graph comparing two primer libraries designed using the methods of the invention. This graph shows the number of loci with a particular minor allele frequency that are targeted by each primer library. During the selection of the "new pool" library, more primers were retained. This library enables the amplification of more target loci, especially target loci with relatively large minor allele frequencies (which are the more informative alleles for some method of the invention, such as for detecting fetal chromosomal abnormalities).

These primer libraries were used in the following multiplex PCR method. Blood (20-40 mL) was collected from each subject into two to four CELL-FREE™ DNA tubes (Streck). Plasma (a minimum of 7 mL) was isolated from each sample via a double centrifugation protocol of 2,000 g for 20 min, followed by 3,220 g for 30 min, with supernatant transfer following the first spin. cfDNA was isolated from 7-20 mL plasma using the QIAGEN QIAamp Circulating Nucleic Acid kit and eluted in 45 uL TE buffer. Pure maternal genomic DNA was isolated from the buffy coat obtained following the first centrifugation, and pure paternal genomic DNA was prepared similarly from a blood, saliva or buccal sample.

Maternal cfDNA, maternal genomic DNA, and paternal genomic DNA samples were pre-amplified for 15 cycles using 11,000 target-specific assays and an aliquot was transferred to a second PCR reaction of 15 cycles using nested primers. Finally, samples were prepared for sequencing by adding barcoded tags in a third 12-cycle round of PCR. Thus, 11,000 targets were amplified in a single reaction; the targets included SNPs found on chromosomes 13, 18, 21, X, and Y. The amplicons were then sequenced using an ILLUMINA GAIIx or HISEQ

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 220

sequencer. Parental genotypes were sequenced at a lower read depth (~20% of cfDNA read depth) than the fetal genotypes.

Experiment 18

If desired, the size and quantity of the PCR products can be analyzed using standard methods, such as the use of the Agilent Technologies 2100 Bioanalyzer (**Figure 28A-M)**. For example, direct PCR methods described herein without nesting were used in 2,400-plex **(Figures 28B-28G)** and 19,488-plex experiments **(Figures 28H to 28M)**. The amount of primer was 10 nM for **Figures 28B-28D and 28H to 28J**. The amount of primer was 1 nM for **Figures 28E-28G and 28K to 28M**. The amount of input DNA was 24 ng for **Figures 28B, 28E, 28H, and 28K**; 80 ng for **Figures 28C, 28F, 28I, and 28L**; and 250 ng for **Figures 28D, 28G, 28J, and 28M**. More input DNA resulted in a greater proportion of the desired 180 base pair product. The peak at 140 base pairs is a primer dimer product.

Experiment 19

A proof-of-principle study demonstrated the detection of T13, T18, T21, 45,X, and 47,XXY with equally high accuracies across all chromosomes.

Patients

Pregnant couples were enrolled at specific prenatal care centers under protocols approved by an Institutional Review Board pursuant to local laws. Inclusion criteria were at least 18 years of age, gestational age of at least nine weeks, singleton pregnancies, and signed informed consent. Blood samples were drawn from pregnant mothers, and a blood or buccal sample was collected from the father. Samples from 2 pregnancies with T13 (Patau Syndrome), 2 with T18 (Edwards Syndrome), 2 with T21 (Down's Syndrome), 2 with 45,X, 2 with 47,XXY, and 90 normal pregnancies were selected prior to testing from a cohort of ~500 women to test which chromosomal abnormalities the method detects. Normal fetal karyotype was confirmed by molecular karyotyping for the samples where post-birth child tissue was available. Euploid sample were drawn prior to invasive testing from low-risk women. Aneuploid samples were drawn at least 7 days after invasive testing and aneuploidy was confirmed via cytogenetic karyotyping or fluorescence *in situ* hybridization at independent laboratories.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 221

Sample Preparation and Multiplex PCR

For the data in **Figures 30A-E, 30G, 30H, and 31A-31G**, sample preparation and 19,488-plex-PCR were performed as described in Experiment 15. For the data in **Figure 30F**, sample preparation and 11,000-plex-PCR were performed as described in Experiment 17.

Methodology and Data Analysis

The algorithm considers parental genotypes and crossover frequency data (such as data from the HapMap database) to calculate expected allele distributions for 19,488 polymorphic loci for a very large number possible fetal ploidy states, and at various fetal cfDNA fractions. (**Figures 29A-29C**). Unlike allele ratio based-methods, it also takes into account linkage disequilibrium, and uses non-Gaussian data models to describe the expected distribution of allele measurements at a SNP given observed platform characteristics and amplification biases. It then compares the various predicted allele distributions to the actual allelic distributions as measured in the cfDNA sample (**Figure 29C**), and calculates the likelihood of each hypothesis (monosomy, disomy, or trisomy, for which there are numerous hypotheses based on the various potential crossovers) based on the sequencing data. The algorithm sums the likelihoods of each individual monosomy, disomy, or trisomy hypotheses (**Figure 29D**), and calls the ploidy state with the maximum overall likelihood as the copy number and fetal fraction (**Figure 29E**). Although laboratory researchers were not blinded to sample karyotype, the algorithm called the ploidy states without human intervention and was blind to the truth.

Data Interpretation

Graphical Representations of the Generated Data

To determine the ploidy state of a chromosome of interest, the algorithm considers the distribution of sequence counts from each of two possible alleles at 3,000 to 4,000 SNPs per chromosome. It is important to note that the algorithm makes ploidy calls using an approach that does not lend itself to visualization. Thus, for the purposes of illustration, the data is displayed here in a simplified fashion as ratios of the two most likely alleles, labeled as A and B, so that the relevant trends can be more readily visualized. This simplified illustration does not take into account some of the features of the algorithm. For example, two important aspects of the algorithm that are not possible to illustrate with a method of visualization that

displays allele ratios are: 1) the ability to leverage linkage disequilibrium, *i.e.* the influence that a measurement at one SNP has on the likely identity of a neighboring SNP, and 2) the use of non-Gaussian data models that describe the expected distribution of allele measurements at a SNP given platform characteristics and amplification biases. Also note that the algorithm only considers the two most common alleles at each SNP, ignoring other possible alleles.

The graphical representations in **Figure 30A-30H** include samples for which two, one, or three fetal chromosomes are present. Generally, these indicate euploidy (**Figures 30A-30C**) monosomy (**Figure 30D**), and trisomy (**Figures 30E-30H**), respectively. In all plots, each spot represents a single SNP, where the targeted SNPs are plotted sequentially from left to right for one chromosome along the horizontal axes. The vertical axes indicate the number of reads for the A allele as a fraction of the total number of reads for both the A and B alleles for that SNP. Note that the measurements are made on total cfDNA isolated from maternal blood, and the cfDNA includes both maternal and fetal cfDNA; thus, each spot represents the combination of the fetal and maternal DNA contribution for that SNP. Therefore, increasing the proportion of maternal cfDNA from 0% to 100% will gradually shift some spots up or down within the plots, depending on the maternal and fetal genotype. This is described in more detail below with the corresponding plots.

The spots are color-coded according to maternal genotype, as maternal genotype contributes more to the localization of each spot and the majority of trisomies are maternally-inherited; this assists in visualizing ploidy states. Specifically, SNPs for which the maternal genotype is AA are indicated in red, those for which the maternal genotype is AB are indicated in green, and those for which the maternal genotype is BB are indicated in blue.

In all cases, SNPs that are homozygous for the A allele (AA) in both the mother and the fetus are found tightly associated with the upper limit of the plots, as the fraction of A allele reads is high because there should be no B alleles present. Conversely, SNPs that are homozygous for the B allele in both the mother and the fetus are found tightly associated with the lower limit of the plots, as the fraction of A allele reads is low because there should be only B alleles. The spots that are not tightly associated with the upper and lower limits of the plots represent SNPs for which the mother, the fetus, or both are heterozygous; these spots are useful for identifying fetal ploidy, but can also be informative for determining paternal versus maternal inheritance. These spots segregate based on both maternal and fetal genotypes and fetal fraction, and as such the precise position of each individual spot along the y-axis depends

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 223

on both stoichiometry and fetal fraction. For example, loci where the mother is AA and the fetus is AB are expected to have a different fraction of A allele reads, and thus different positioning along the y-axis, depending on the fetal fraction.

Two Chromosomes Present

Figures 30A-30C depict data that indicate the presence of two chromosomes when the sample is entirely maternal (no fetal cfDNA present, **Figure 30A**), contains a moderate fetal cfDNA fraction (**Figure 30B**), or contains a high fetal cfDNA fraction (**Figure 30C**).

Figure 30A shows data obtained from cfDNA isolated from the blood of a non-pregnant woman. When there is no fetal cfDNA present and the sample contains only maternal cfDNA, the plots represent purely the euploid maternal genotype; the hallmark pattern includes "clusters" of spots: a red cluster tightly associated with the top of the plot (SNPs where the maternal genotype is AA), a blue cluster tightly associated with the bottom of the plot (SNPs where the maternal genotype is BB), and a single, centered green cluster (SNPs where the maternal genotype is AB).

When fetal cfDNA is present, the location of the spots shifts such that the clusters segregate into discrete "bands". Note that for samples with a fetal fraction of 0%, the groupings of spots are referred to as "clusters" (as in **Figure 30A**), and for all samples with a fetal fraction of >0%, the groupings of spots are referred to as "bands" (as in **Figures 30B-30J**). If the fetal fraction is high enough, these discrete bands will be readily visible. Specifically, **Figures 30B and 30C** demonstrate the characteristic pattern associated with two fetal chromosomes present at moderate and high fetal fractions, respectively. This pattern includes three central green bands that correspond to SNPs that are heterozygous in the mother, and two "peripheral" bands each at both the top (red) and bottom (blue) of the plots that correspond to SNPs that are homozygous in the mother.

Figure 30B shows data obtained from cfDNA isolated from a plasma sample from a woman carrying a euploid fetus and with a 12% fetal cfDNA fraction. Here, the clusters of spots tightly associated with the top and bottom of the plot segregate into two discrete bands each: one red and one blue external peripheral band that remains tightly associated with the upper or lower limit of the plots, and one red and one blue internal peripheral band that has separated from the limits of the plots. These internal peripheral bands, centered around 0.92 and 0.08, represent SNPs for which the maternal genotype is AA and the fetal genotype is AB

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 224

(indicated in red), and SNPs for which the maternal genotype is BB and the fetal genotype is AB (indicated in blue), respectively. The center cluster of green spots broadens, but at this fetal fraction the segregation into distinct bands is not readily visible.

At a high fetal cfDNA fraction, the typical pattern that indicates the presence of two chromosomes (a trio of green bands as well as two red and two blue peripheral bands) is readily apparent. **Figure 30C** displays data obtained from a plasma sample from a woman carrying a euploid fetus at a fetal cfDNA fraction of 26%. Here, the peripheral bands have separated such that the internal band has shifted towards the center of the plot due to the altered levels of B alleles from the increased fetal cfDNA fraction. Significantly, at higher fetal fractions, the separation of the center green cluster into three distinct bands is now readily apparent. This central trio of bands, in this case clustering around 0.37, 0.50 and 0.63, corresponds to those SNPs where the maternal genotype is AB, and the fetal genotype is AA (top), AB (middle) and BB (bottom).

These hallmark patterns, namely three green bands and four peripheral bands (two red and two blue), indicate the presence of two chromosomes, as in autosomal euploidy or for the X chromosome in a female (XX) fetus.

One Chromosome Present

When the fetus only inherits a single chromosome, and thus only inherits a single allele, heterozygosity of the fetus is not possible. As such, the only possible fetal SNP identities are A or B. Thus, maternally-inherited monosomic chromosomes have a characteristic pattern of two central green bands that represent SNPs for which the mother is heterozygous, and only have single peripheral red and blue bands that represent SNPs for which the mother is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively **(Figure 30D)**. Note the absence of internal peripheral bands. This pattern indicates the presence of one chromosome, as in maternally-inherited autosomal monosomy, or for the X chromosome in a male (XY) fetus.

Three Chromosomes Present

Trisomic chromosomes have three characteristic patterns. The first pattern indicates maternally-inherited meiotic trisomy, a meiotic error where the fetus inherited two homologous, non-identical chromosomes from the mother **(Figure 30E)**; this pattern includes

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 225

two central green bands with two each of the peripheral red and blue bands. The second pattern indicates paternally-inherited meiotic trisomy, where the fetus inherited two homologous, non-identical chromosomes from the father (**Figure 30F**); this pattern includes four central green bands and three each of the peripheral red and blue bands. The third pattern indicates either maternally- (**Figure 30G**) or paternally-inherited (**Figure 30H**) mitotic trisomy, a mitotic error where the fetus inherited two identical chromosomes from either the mother or the father; this pattern includes four central green bands with two each of the peripheral red and blue bands. Maternally- and paternally-inherited mitotic trisomies can be distinguished by the placement of the flanking red and blue bands, such that the red and blue internal peripheral bands (those not associated with the limits of the plots) are closer to the center in paternally-inherited mitotic trisomy. This is due to the paternal contribution of identical chromosomes. Note that our previous results indicate that at the blastomere stage, 66.7% of maternally-inherited trisomies are meiotic, and that only 10.2% of trisomies are paternally-inherited.

For the Y chromosome, the PS method considers a different set of hypotheses: zero, one, or two chromosomes present. As there is no maternal contribution to the sequence reads at each locus and because heterozygous loci are not possible (cases of two Y chromosomes necessarily involve two identical chromosomes), the bands remain tightly associated with the top (A alleles) or the bottom (B alleles) of the plot (data not shown), and analysis is greatly simplified, relying on quantitative allele count data. Note that since the method interrogates SNPs, it uses homologous non-recombinant SNPs from the Y chromosome, thus obtaining data on both X and Y for one probe pair.

Identifying Aneuploidies

Identification of autosomal aneuploidies using this plot-based visualization method is straightforward given a sufficient fetal fraction, and requires only identifying plots for which there are an abnormal number of chromosomes present, as described above. Combining the knowledge of copy number of the X and Y chromosomes identifies whether sex chromosome aneuploidies are present. Specifically, plots representing a fetus with a 47,XXX genotype will have a typical “three-chromosome” pattern, and plots representing a fetus with a 47,XXY genotype will have the typical “two-chromosome” pattern for the X chromosome, but will also have allele reads indicating the presence of one Y chromosome. The method is similarly able to call 47,XYY, where a “one chromosome” pattern indicates the presence of a single X

chromosome, and allele reads indicate the presence of two Y chromosomes. A fetus with a 45,X genotype will have the typical "one-chromosome" pattern for the X chromosome, and data indicating zero Y chromosomes.

Effects of Fetal Fraction

As discussed above, the number of sequence reads from the fetus contributes to the precise location of each spot along the y-axis in the plots. As fetal fraction will affect the proportion of reads that originate from the fetus and the mother, it will also affect the positioning of each spot. At a high fraction of fetal cfDNA (generally above ~20%), as in **Figures 30C-30E** and **Figures 30G and 30H**, it is readily apparent that although the spots cluster based mainly on maternal genotype, the presence of fetal DNA from alleles whose genotype is distinct from the maternal genotype shift the clusters into multiple, distinct bands. However, as the fetal fraction decreases (as in **Figures 30B and 30F**), the spots regress towards the poles and center of the plot, resulting in tighter clusters. Specifically, the set of peripheral red bands, where the maternal genotype is AA, regress towards the top of the plot; the set of peripheral blue bands, where the maternal genotype is BB, regress towards the bottom; the set of central green bands, where the mother is heterozygous, condense into a single cluster at the center of the plot (compare **Figures 30B and 30C**). Although aneuploidy is not readily apparent by eye using this visualization technique for low fetal fraction cases, the algorithm is able to identify ploidy states with a very low fetal fraction, such as 3% fetal fraction. It is able to do this because the statistical technique compares the observed data to very precise data models that predict the allele distributions for a given sample parameter set (including copy number, parental genotypes, and fetal fraction, for example). Data model precision is critical in low fetal fraction cases, as the differences between the allele distributions for different ploidy states are proportional to the fetal fraction. In addition, the algorithm is able to determine when a data set does not contain enough data to make a confident fetal ploidy determination.

Results

Sequencing reads that mapped to targeted SNPs were deemed to be informative and were used by the algorithm. More than 95% of targeted loci were observed in the sequencing results. The plots for visualizing key ploidy calls are depicted in **Figure 31A-31G**. **Figure**

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 227

31A indicates a euploid sample. Here, chromosomes 13, 18, and 21 have the typical "two chromosome" pattern (as described herein). This includes a trio of center green bands, and two red and two blue peripheral bands. This, together with the two center green bands for the X chromosome and the presence of Y chromosome bands along the plots' peripheries, indicate a euploid XY genotype.

The most prevalent autosomal trisomies, T13, T18, and T21, are indicated by the plots in **Figures 31B, 31C, and 31D**, respectively. Specifically, **Figure 31B** depicts a T13 sample. Here, chromosomes 18 and 21 display the typical "two chromosome" pattern, chromosome X displays the typical "one chromosome" pattern, and there are reads from the Y chromosome. Together, this indicates disomy at chromosomes 18 and 21, and identifies a fetal XY genotype. However, chromosome 13 depicts a typical "three chromosome" pattern – specifically. Similarly, **Figure 31C** depicts a T18 sample, and **Figure 31D** depicts a T21 sample.

The method is also able to detect sex chromosome aneuploidies, including 45,X **(Figure 31E)**, 47,XXY **(Figure 31F)**, and 47,XYY **(Figure 31G)**. Note that the method is calling copy number at chromosomes 13, 18, 21, X, and Y; the overall chromosome number is reported assuming disomy at the remaining chromosomes. The X chromosome regions of the plot depicting a 45,X sample reveals the presence of a single chromosome. However, the lack of reads from the Y chromosome, coupled with the "two chromosome" pattern for chromosomes 13, 18, and 21, indicate a 45,X genotype. Conversely, the 47,XXY samples generate a plot revealing the presence of two X chromosomes. The data also revealed reads for alleles from the Y chromosome. Together with the presence of two copies of chromosomes 13, 18, and 21, this indicates a 47,XXY genotype. A 47,XYY genotype is indicated by the presence of a "one chromosome" pattern for the X chromosome, and reads indicating the presence of two Y chromosomes.

Discussion

This method detected T13, T18, T21, 45,X, 47,XXY, and 47,XYY non-invasively from maternal blood. This method interrogates cfDNA from maternal plasma by targeted multiplex PCR amplification and high-throughput sequencing of 19,488 SNPs. This, coupled with the method's sophisticated informatics analyses that take into account parental genotypic information and numerous sample parameters, including fetal fraction and DNA quality, more robustly detects the fetal signal and makes highly accurate ploidy calls at all of the five

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 228

chromosomes implicated in the seven most common types of at-birth aneuploidy (T13, T18, T21, 45,X, 47,XXX, 47,XXY, and 47,XYY). This method offers a number of clinical advantages over previous methods, including and most significantly greater clinical coverage and sample-specific calculated accuracies (analogous to a personalized risk score).

Increased Clinical Coverage

This method offers approximately a two-fold increase in aneuploidy coverage compared to clinically available NIPT methodologies, given its ability to accurately detect autosomal trisomies and sex chromosome aneuploidies. The method presented here is the only noninvasive test that calls ploidy at the sex chromosomes with high accuracy. Prior DNA mixing experiments and separate plasma samples analyzed in our experimental assays suggest that this method will detect a larger cohort of sex chromosome anomalies, including 47,XXX. The method presented here also detects aneuploidies at chromosomes 13, 18, and 21 with high sensitivities and specificities, and with appropriate primer design is expected to be able to detect copy number at the remaining chromosomes as well.

Sample-Specific Calculated Accuracies

Significantly, this method calculates a sample-specific accuracy for ploidy calls on each chromosome in each sample. Accuracies calculated by this method are expected to significantly lower the rate of incorrect calls by identifying and flagging individual samples that have poor quality DNA or low fetal fractions that are likely to result in a poor accuracy test result. By contrast, massively parallel shotgun sequencing (MPSS)-based methods produce a positive or negative call using a single-hypothesis rejection test, and their accuracy estimate is based on a published study cohort rather than on the characteristics of the individual sample, which are assumed to have the same accuracy as the cohort. However, individual accuracies for samples with parameters in the tail of the cohort distribution may differ significantly. This is exacerbated at low fetal fractions, as in early gestational age, or for samples with low DNA quality. These samples are generally not identified and flagged for follow-up, which can result in missed calls. The present method, however, takes into account many parameters, including fetal fraction and a number of DNA quality metrics, to make each chromosome copy number call, calculating a sample-specific accuracy for that call. This allows the method to identify

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 229

individual samples with low accuracy and flag them for follow-up. This is expected to nearly eliminate missed calls, especially at the early stages of pregnancy when fetal fractions are typically low. The presumption is that a no call is much preferred to a missed call, since a no call simply requires a redraw and reanalysis.

Converting Calculated Accuracies to Traditional Risk Scores

This method can offer an adjusted risk of aneuploidy for high-risk pregnant women, where the adjusted risk takes into account an *a priori* risk (Benn P, Cuckle H, Pergament E. Non-invasive prenatal diagnosis for Down syndrome: the paradigm will shift, but slowly. Ultrasound Obstet Gynecol 2012;39:127-130, which is hereby incorporated by reference in its entirety). Although the present method offers each patient a customized calculated accuracy, for clinical use these accuracies can be converted to traditional risk scores, which also denote the risk of an aneuploid pregnancy but are expressed as fractions. Traditional risk scores take into account various parameters, including maternal age-related risk and serum levels of biochemical markers, to offer a risk score above which a mother is considered high-risk and for whom follow-up invasive diagnostic procedures are recommended. This method significantly refines this risk score, thus reducing both the false positive and false negative rates, and offering a more accurate assessment of individual maternal risk. A calculated accuracy as used here is the likelihood that the ploidy call is correct, and is expressed as a percentage, but the calculated accuracies used in Experiment 19 do not include an age-related risk. Because calculation of a risk score typically includes an age-related risk, the calculated accuracies and traditional risk scores are not interchangeable; they must be combined to convert into a traditional risk score. The formula to combine the age-related risk with the calculated accuracy is:

$$\frac{R_1 R_2}{R_1 R_2 + [1 - R_1][1 - R_2]}$$

where R_1 is the risk score as calculated by the present method and R_2 is the risk score as calculated by first trimester screening.

SNP-Based Methods Negate Issues with Amplification Variation

An inherent drawback to the counting methods used by some other methods is that they determine fetal ploidy state by measuring the ratio of the number of reads mapping to the

chromosome of interest (*e.g.*, chromosome 21) to those mapping to a reference chromosome. Chromosomes with high or low GC content, including chromosomes 13, X, and Y, amplify with high variability. This can result in signal variation that is comparable in magnitude to the fetal cfDNA signal, which can confound copy number calls by altering the ratio of allele reads from the chromosome-of-interest to those from the reference chromosome. This can result in low accuracy for chromosomes 13, X, and Y. Significantly, this problem is exacerbated at low fetal cfDNA fractions, as tends to be the case at early gestational ages.

In contrast, SNP-based methods do not rely on consistent amplification levels between chromosomes, and are thus expected to provide results that are equally accurate across all chromosomes. Because the present method looks, in part, at relative counts of different alleles at polymorphic loci, which by definition differ only by a single nucleotide, it does not require the use of reference chromosomes, and this obviates the problems with chromosome-to-chromosome amplification variation that are inherent to methods that rely on quantitating read counts. Unlike quantitative methods that require reference chromosomes that are euploid, the present method is expected to be able to detect triploidy as well as copy-number neutral anomalies like uniparental disomy.

The Importance of Early Detection

Significantly, the combined at-birth prevalence of sex chromosome aneuploidies is higher than that of the most common autosomal aneuploidies **(Figure 32)**. However, there are currently no routine non-invasive screening methods that reliably detect sex chromosome abnormalities. Thus, sex chromosome anomalies are generally detected prenatally as a side-effect of routine testing for Down syndrome or other autosomal aneuploidies; a large proportion of cases are missed entirely. Early and accurate detection is crucial for many of these disorders where early therapeutic intervention improves clinical outcomes. For example, Turner syndrome is often not diagnosed until adolescence, although its overall at-birth prevalence is 1 in 2,500 females. Growth hormone therapy is known to prevent short stature that results from the disorder, but treatments are significantly more effective when initiated prior to the age of 4. Additionally, estrogen replacement therapy can stimulate secondary sexual characteristics in patients with Turner syndrome, but again therapy must be initiated in the pre-teen years, before the syndrome is usually detected. Together, this underscores the importance of early, routine, and safe detection of sex chromosome aneuploidies. This method

offers the first approach with the potential to serve as a routine screen for sex chromosome anomalies.

Additional Applications

Because this method utilizes targeted amplification, it is uniquely poised to detect sub-microscopic anomalies, such as microdeletions and microduplications. Although non-targeted methods like MPSS have been shown to detect the DiGeorge microdeletion syndrome, this required a sufficiently high level of genomic coverage so as to make the approach unfeasible. This is because non-targeted amplification will be several orders of magnitude less efficient on submicroscopic regions, as very small fraction of the sequencing reads will be informative. Additionally, the fact that the currently available methods have trouble accurately identifying ploidy state for the sex chromosomes suggests that they will also encounter variable amplification problems on smaller chromosomal segments.

Similarly, SNP based methods can detect UPD disorders, which are copy number-neutral anomalies that will not be detected by either the current noninvasive methods that rely on counting or the traditional invasive methods like amniocentesis and CVS that rely on cytogenetic karyotyping and/or fluorescence *in situ* hybridization. This is because SNP-based methods are uniquely able to distinguish individual haplotypes, whereas the clinically available MPSS-based and targeted methods amplify non-polymorphic loci and are thus unable to determine, for example, whether the chromosomes-of-interest originate from the same parent. This means that these microdeletion/microduplication and UPD syndromes, including Prader-Willi, Angelman, and Beckwith-Wiedemann syndromes, are generally not diagnosed prenatally, and are often initially misdiagnosed postnatally. This significantly delays therapeutic intervention. Additionally, because this method targets SNPs, this method will also facilitate parental haplotype reconstruction, allowing for detection of fetal inheritance of individual disease-linked loci (Kitzman JO, Snyder MW, Ventura M, *et al.* Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med 2012;4:137ra76, which is hereby incorporated by reference in its entirety).

The results presented here confirm the expanded scope of this method for identifying prenatal aneuploidy. Specifically, by amplifying and sequencing 19,488 SNPs, this method is able to determine copy number at chromosomes 13, 18, 21, X, and Y, and is uniquely expected to detect other chromosomal abnormalities, such as triploidy and UPD, that are not detected by

any other clinically available non-invasive method. The increased clinical coverage and powerful sample-specific calculated accuracies suggest that this method may offer a viable adjunct to invasive testing for detecting fetal chromosomal aneuploidies.

All patents, patent applications, and published references cited herein are hereby incorporated by reference in their entirety. While the methods of the present disclosure have been described in connection with the specific embodiments thereof, it will be understood that it is capable of further modification. Furthermore, this application is intended to cover any variations, uses, or adaptations of the methods of the present disclosure, including such departures from the present disclosure as come within known or customary practice in the art to which the methods of the present disclosure pertain, and as fall within the scope of the appended claims. For example, any of the methods disclosed herein for DNA can be readily adapted for RNA by including a reverse transcription step to convert the RNA into DNA. Examples that use polymorphic loci for illustration can be readily adapted for the amplification of nonpolymorphic loci if desired.

NATERA EX2040
NeoGenomics v. Natera, IPR2024-00812
Page 233