

METHOD AND DEVICE FOR DETERMINING FETAL NUCLEIC ACID CONCENTRATION IN MATERNAL PLASMA

FIELD

[0001] The present disclosure relates to the field of gene detection, and particularly, to a
5 method and a device for determining a fetal nucleic acid concentration in maternal blood.

BACKGROUND

[0002] Since the discovery of fetal cell-free DNA in maternal plasma in 1997, noninvasive
prenatal testing (NIPT), which can obtain fetal DNA information by extracting circulating cell-
free DNA (cfDNA) from the maternal plasma, has made great progress. It has been confirmed
10 that a concentration of fetal cfDNA in maternal plasma cfDNA increases with an increase of
the gestational age of blood collection, and also varies in different pregnant women. The
accurate estimation of the concentration of the fetal cfDNA in maternal plasma cfDNA not only
helps to enhance an accuracy of the NIPT, but also facilitates the study of its influence on
various pregnancy complications and maternal phenotypes.

[0003] Many institutions have successively proposed methods for estimating a fetal nucleic
acid concentration through different data and different methods. These methods can be
classified into six categories: 1) a Y chromosome depth based method, which calculates the
fetal nucleic acid concentration based on the principle that cfDNA sequencing reads derived
from pregnant women in plasma cannot be mapped to a non-homologous region of Y
20 chromosome of the human reference genome, and the reads that can be uniquely mapped to the
non-homologous region of the Y chromosome of the human reference genome are all derived
from the cfDNA of the male fetuses; 2) a single nucleotide polymorphism (SNP) method
assisted by capture sequencing data, which utilizes homozygous sites but different in paternal
and maternal genome and incorporate the reads depth information of maternal plasma to
25 calculate the fetal nucleic acid concentration; 3) a sequencing reads depth distribution method,
which divides a genome into windows (for example, 50kb), calculates the total number of reads

of cfDNA in maternal plasma and a proportion of short reads in each window, uses the calculated results as input data to establish a regression model, training the model by the fetal cfDNA concentration calculated by Y chromosome depth as the true set, and using data of a super-large sample size, and then detecting the fetal cfDNA concentration using this trained model; 4) a methylation data method, in which the calculation is based on the differences in the methylation of DNA derived from different individuals or from different tissues of the same individual; 5) a cfDNA fragment length method, which estimates a concentration of fetal cfDNA by calculating a proportion of cfDNA fragments in the pregnant woman's body based on the known fact that the length of fetal cfDNA fragments in the maternal plasma have more proportion distribution between 147bp and 167bp, and the length of cfDNA of pregnant women generally distribute between 167bp to 187bp; and 6) a nucleosome arrangement method, which estimates a concentration of fetal cfDNA based on the difference in fragment length distribution of pregnant woman cfDNA and fetal cfDNA caused by different degradation degrees. Among the six methods, only methods 1), 3) and 6) can calculate fetal cfDNA concentration merely based on NIPT data, method 1) is limited to the calculation specified to male fetuses, and methods 3) and 6) can only be used for calculation of NIPT data of higher depths.

[0004] Currently, no method is capable to calculate the cfDNA concentration of both male and female fetuses indiscriminately based on ultra-low depth NIPT data.

SUMMARY

[0005] The present disclosure aims to solve one of the technical problems in the related art at least to a certain extent. In this regard, an object of the present disclosure is to provide a method and a device for determining a fetal nucleic acid concentration in maternal plasma.

[0006] In the course of long-term research, the applicant has the following findings.

[0007] As for the existing methods for calculating the fetal nucleic acid concentration, when detecting a fetal nucleic acid concentration in blood of a pregnant woman, other types of data in addition to NIPT data of blood plasma of the pregnant woman are required for assistance in most cases. Among them, 1) the single nucleotide polymorphism (SNP) method assisted by capture sequencing data requires the use of parental capture sequencing data or high-depth cfDNA sequencing data as an assistance, the calculation of fetal nucleic acid concentration can

be completed only on basis of additionally obtaining the accurate genotypes of the father and mother (or at least the mother); 2) the methylation data method requires additional methylation data of the father and mother for the calculation of fetal nucleic acid concentration. These methods require different types of auxiliary data, which on the one hand, increases the difficulty of sampling (for example, requiring to obtain a father's blood sample), and on the other hand, increases the cost of analysis.

[0008] However, the methods that only require NIPT data of maternal plasma have additional requirements for the type of NIPT data and the sequencing depth. Among them, 1) the cfDNA fragment length calculation method requires accurate estimation of cfDNA fragment length, and thus only the paired-end sequencing (PE) can be adopted; and 2) the nucleosome arrangement calculation method requires the use of a depth difference of reads in the nucleosome unit and thus has certain requirements for the depth of NIPT sequencing, and thus it is impossible to only use the data of about 0.1x depth that is at present common in NIPT detection.

[0009] The remaining two methods, which require no additional data of other types as assistance and is not specific to NIPT data, both have their own limitations in application and cannot cover all NIPT samples. Among them, 1) the Y chromosome depth calculation method can only calculate the fetal nucleic acid concentration of male fetuses, but cannot calculate the fetal nucleic acid concentration of female fetuses; and 2) the sequencing reads depth distribution calculation method can only be used to estimate samples with a high fetal nucleic acid concentration, and cannot be applied to samples with a fetal nucleic acid concentrations smaller than or equal to 5%.

[0010] Therefore, it is still necessary to further improve the method for calculating a cfDNA concentration of both male and female fetuses indiscriminately by means of low-depth NIPT data. In this regard, the present disclosure provides a method and a device for determining a fetal nucleic acid concentration in maternal blood. The method or device can use only the NIPT data of maternal plasma to determine the fetal nucleic acid concentration, without requiring the assistance of other data. In addition, the method or device can be applied to the sequencing data of ultra-low depth (for example, about 0.1x), and has no limitations on the types of sequencing, and both the paired-end sequencing and the single-end sequencing can be employed; and the

method or device is not specific to the sample types, and is applicable for both samples involving male fetuses and samples involving female fetuses.

[0011] Specifically, the present disclosure provides the following technical solutions.

[0012] According to a first aspect of the present disclosure, the present disclosure provides
5 a method for determining a fetal nucleic acid concentration in maternal blood. The method includes: (1) determining first genotype information based on alignment of sequencing data with at least a part of a reference genome, the sequencing data being derived from a nucleic acid sample of the maternal blood; (2) correcting the first genotype information based on
10 reference data by using a linkage disequilibrium relationship to obtain second genotype information; and (3) determining the fetal nucleic acid concentration based on a difference between the first genotype information and the second genotype information.

[0013] The present disclosure provides a method for determining a fetal nucleic acid concentration in maternal blood. In the method, first genotype information is obtained by alignment of sequencing data with a reference genome; the sequencing data are obtained by
15 sequencing the nucleic acid sample of the maternal bloods, and the sequencing data includes fetal nucleic acid information and maternal nucleic acid information. The sequencing data also indirectly includes paternal nucleic acid information, since a part of the fetal nucleic acid information is derived from the father. The linkage disequilibrium relationship is utilized to correct the obtained first genotype information. In other words, the part of the nucleic acid
20 information derived from the paternal parent is mixed in the sequencing data, and is derived from a different individual from the maternal information, and thus is corrected to a certain extent to obtain the second genotype information after correction. Then, a part of corrected genotype information can be determined by the difference between the first genotype information and the second genotype information, and the more the part of corrected genotype
25 information, the higher the fetal cfDNA concentration in the maternal plasma. The fetal nucleic acid concentration can be determined based on the relationship between a proportion of the corrected genotype and the fetal cfDNA concentration in the maternal plasma.

[0014] The method for determining the fetal nucleic acid concentration in maternal blood provided by the present disclosure has many advantages in the following aspects: 1) throughout
30 the method, only the NIPT data of maternal plasma is used without requiring the assistance of

other data; 2) the method can be applied to the sequencing data of ultra-low depth (for example, about 0.1x), because that the linkage disequilibrium relationship is used to estimate the changes in the genome genotype, the changes in the genome genotype can reflect the fetal nucleic acid concentration as long as the data derived from two sources (pregnant woman and fetus) are mixed, even if the sequencing depth is low; 3) the method is not specific to the types of sequencing, and both the paired-end sequencing or the single-end sequencing are applicable; 4) the method is not specific to the sample types (both samples involving male fetuses and samples involving female fetuses are applicable), and has no limitation for the fetal nucleic acid concentration.

[0015] The method provided in the present disclosure breaks through the limitations of the sequencing depth, data type, and fetal gender of the estimated data has universality, requires no additional sampling and sequencing costs, and has extremely high application value in the NIPT field.

[0016] According to the embodiments of the present disclosure, the above-mentioned method for determining the fetal nucleic acid concentration in maternal blood may further include the following technical features.

[0017] In some embodiments of the present disclosure, the sequencing data are obtained by sequencing the nucleic acid sample of the maternal blood, and a depth of the sequencing may be 10X, 5X, 1X, 0.5X, 0.2X, or 0.1X. The nucleic acid sample of the maternal blood include fetal nucleic acid information and maternal nucleic acid information, and a part of the fetal nucleic acid information is derived from the paternal parent. Therefore, even the sequencing data obtained by sequencing the nucleic acid sample of the maternal blood are sequencing data with a low sequencing depth, e.g., sequencing data of up to 10X, up to 5X, up to 1X, or even 0.1X, the sequence data can be analyzed by the method provided in the present disclosure to determine the fetal nucleic acid concentration.

[0018] In some embodiments of the present disclosure, the sequencing data is obtained through the second-generation sequencing technology or the third-generation sequencing technology. The second-generation sequencing technology is also referred to as high-throughput sequencing technology, which can obtain many sequences one time. For example, the nucleic acid can be randomly broken into small fragments (about 250bp to 300bp) by

physical or chemical means, and then these small molecule fragments are enriched by constructing a library and sequenced in a sequencer. The sequencer has regions where these fragments can be attached, and each fragment has an independent attachment region, such that the information of all attached DNA sequences can be detected in one time. The second-generation sequencing technology can detect a large number of sequences in one time, but the fragments are limited to, for example, about 250bp to 300bp, and the cost is relatively high. The commonly used second-generation sequencing technology can be Roche/454's pyrosequencing method for sequencing, or Illumina's fluorescent sequencing detection, ABI/Solid's fluorescent sequencing detection, or MGI's DNB sequencing detection, etc. The third-generation sequencing technology can make the sequencing length up to about 10KB and does not rely on PCR amplification. The third-generation sequencing technology allows a sequencing length up to about 10KB and does not rely on PCR amplification, for example, using PacBio's SMRT or single-molecule nanopore DNA sequencing technology developed by Oxford Nanopore Technologies. The sequencing data of the nucleic acid sample of maternal blood, whether obtained by the second-generation sequencing technology or the third-generation sequencing technology, or obtained by the single-end sequencing or the paired-end sequencing, can be analyzed according to the method provided in the present disclosure to determine the fetal nucleic acid concentration.

[0019] In some embodiments of the present disclosure, the reference genome includes at least one strong linkage region in the human genome. The "strong linkage region" varies with the size and structure of the studied population, and is generally defined as a region in which a probability of historical recombination between any pair of variation sites is less than 5%. The reference genome may contain one strong linkage region, two strong linkage regions, three strong linkage regions, or even more. Generally speaking, regardless of cost, the more the strong linkage regions contained in the reference genome, the more accurate the fetal nucleic acid concentration in the maternal blood finally determined after alignment and calculation is.

[0020] When selecting strong linkage regions, all or part of the strong linkage regions of the population in the genome can be found, depending on the size and structure of the studied population, and then based on the ranges of these strong linkage regions, strong linkage regions with appropriate sizes can be selected as the reference genome. Generally speaking, the greater the number and proportion of variation sites of the maternal DNA and fetal DNA covered in the

selected strong linkage region, the more accurate the fetal nucleic acid concentration in the maternal blood calculated using the reference genome containing this strong linkage region is. In some embodiments of the present disclosure, a length of the strong linkage region ranges from 5 mb to 10 mb, for example, 10 mb, 9 mb, 8 mb, 7 mb, 6 mb, or 5 mb. The above-listed lengths of the strong linkage region may have a variation of 10% to 20%, for example, the length of the strong linkage region can be 10 mb, 11 mb or 12 mb, 9 mb or 8 mb, etc. In this way, the information derived from paternal sequencing data can be accurately corrected.

[0021] In some embodiments of the present disclosure, the first genotype information is determined based on a support number of sequencing reads. When the first genotype information is obtained, the first genotype information is determined based on the support number of sequencing reads. For example, for a certain site, if 100 sequencing reads support base A, 8 sequencing reads support base G and 20 sequencing reads support base T, then the base at this site is determined to be base A. In this way, the genotype information of each site can be obtained, and the required first genotype information can be determined by aligning to at least a part of the reference genome.

[0022] In some embodiments of the present disclosure, the first genotype information includes at least one of SNV or Indel. The first genotype information includes single nucleotide variation (SNV) information and/or small fragment insertion-deletion (Indel) information, and the paternal nucleic acid information can be reflected through the correction of these information, thereby realizing accurate determination of the fetal nucleic acid concentration.

[0023] In some embodiments of the present application, the reference data includes multiple pieces of variation site information and multiple pieces of variation frequency information. Using data containing the multiple of pieces of variation site information and the multiple of pieces of variation frequency information as the reference data, the first genotype information can be corrected based on the linkage disequilibrium relationship, so as to correct some of sequencing information from the father, and then the fetal nucleic acid concentration is determined based on a correlation between the corrected information and the fetal nucleic acid concentration.

[0024] In some embodiments of the present disclosure, the correction is performed through IMPUTE2. IMPUTE2, as an imputation algorithm, which is actually a genotype completion

and correction algorithm for sites with deletion data or low accuracy. In addition to IMPUTE2, other imputation methods can also be used to perform the site correction by using LD information, such as BNEAGLE, PHASE and other software.

[0025] In some embodiments of the present disclosure, the step (3) further includes: step (3-1) of determining a difference ratio between the first genotype information and the second genotype information; and step (3-2) of determining the fetal nucleic acid concentration based on the difference ratio obtained in step (3-1) and a pre-trained formula, the formula being determined based on a plurality of training samples with known fetal nucleic acid concentrations. In view of a relationship between the plurality of known fetal nucleic acid concentrations and the difference ratio of the first genotype information and the second genotype information, different formulas or models, for example, the linear regression model, or other models that effectively integrate all information, such as the random forest model or other deep learning models, etc., are used for training to correlate the fetal nucleic acid concentration with the difference ratio. After calculating the difference ratio between the first genotype information and the second genotype information of a nucleic acid sample from a given maternal blood, the fetal nucleic acid concentration in this maternal blood can be determined by the pre-trained formula. That is, in the early stage, some samples with known fetal nucleic acid concentrations can be used as a training set that is fitted with the help of the method provided by the present disclosure using different formulas or models so as to determine the formula; and in the subsequent application, the fetal cfDNA concentration of one or more samples can be predicted without requiring additional samples with known fetal nucleic acid concentrations.

[0026] In some embodiments of the present disclosure, the term “plurality of” indicates at least 100, for example, 100, 500, 1,000, 5,000, or more. When a formula or model is used for training, the greater the number of training samples, the more accurate the formula, and the more accurate the fetal nucleic acid concentration in the maternal blood measured by the formula. Of course, too many samples can also increase the calculation cost and the cost from the sample size itself. When both the cost and training accuracy are optimal, the number of these training samples can range from 5,000 to 1,0000, for example, 5,000, 6,000, 7,000, 8,000, 9,000, or 10,000, which is not limited herein.

[0027] According to a second aspect of the present disclosure, the present disclosure provides a device for determining a fetal nucleic acid concentration in maternal blood. The device can determine the fetal nucleic acid concentration in the maternal blood by using only the NIPT data of maternal plasma without the assistance of other data. In addition, the device can be applied to the sequencing data of ultra-low depth. Besides, the device has no limitations on the types of samples, and are applicable for both samples bearing male fetuses or samples bearing female fetuses. The device includes: an alignment unit configured to determine a first genotype information based on alignment between sequencing data and at least a part of a reference genome, in which the sequencing data are derived from a nucleic acid sample of the maternal blood; a correction unit connected to the alignment unit and configured to correct the first genotype information based on reference data by using a linkage disequilibrium relationship to obtain second genotype information; and a calculation unit connected to the alignment unit and the correction unit and configured to determine the fetal nucleic acid concentration based on a difference between the first genotype information and the second genotype information.

[0028] According to the embodiments of the present disclosure, the above device for determining the fetal nucleic acid concentration in the maternal blood can further have the following technical features. These technical features are mentioned or involved in the above-mentioned method for determining the fetal nucleic acid concentration in the maternal blood, and the functions of the features are similar to those in the above-mentioned method for determining the fetal nucleic acid concentration in maternal blood, and will not be described in detail herein.

[0029] In some embodiments of the present disclosure, in the device, the sequencing data are obtained by sequencing the nucleic acid sample of the maternal blood, and a depth of the sequencing may be 10X, 5X, 1X, 0.5X, 0.2X, or 0.1X. That is, the device provided by the present disclosure can use not only high-depth or relatively-high-depth sequencing data to determine the fetal nucleic acid concentration in the maternal blood, but also low-depth sequencing or ultra-low-depth sequencing data to determine the fetal nucleic acid concentration in the maternal blood.

[0030] In some embodiments of the present disclosure, in the device, the sequencing data

are obtained by means of the second-generation sequencing technology or the third-generation sequencing technology.

[0031] In some embodiments of the present disclosure, in the device, the reference genome includes at least one strong linkage region in the human genome.

5 **[0032]** In some embodiments of the present disclosure, in the device, a length of the strong linkage region ranges from 5mb to 10mb, for example, 10mb, 9mb, 8mb, 7mb, 6mb, or 5mb. The length of the strong linkage region is not specifically limited in the present disclosure.

[0033] In some embodiments of the present disclosure, in the device, the first genotype information is determined based on a support number of sequencing reads.

10 **[0034]** In some embodiments of the present disclosure, in the device, the first genotype information includes at least one of SNV or Indel.

[0035] In some embodiments of the present disclosure, in the device, the reference data includes a plurality of pieces of variation site information and a plurality of pieces of variation frequency information.

15 **[0036]** In some embodiments of the present disclosure, in the device, the correction is performed through IMPUTE2.

[0037] In some embodiments of the present disclosure, the calculation unit further includes: a difference ratio calculation unit configured to determine a difference ratio between the first genotype information and the second genotype information; and a fetal nucleic acid
20 concentration calculation unit connected to the difference ratio calculation unit. The fetal nucleic acid concentration calculation unit is configured to determine the fetal nucleic acid concentration based on the difference ratio obtained in the difference ratio calculation unit and a pre-trained formula. The training formula is determined based on a plurality of training samples with known fetal nucleic acid concentrations.

25 **[0038]** In some embodiments of the present disclosure, the term “plurality of” may indicate 5,000 to 10,000, for example, 5,000, 6,000, 7,000, 8,000, 9,000, or 10,000. The number of samples is not specifically limited herein.

[0039] According to a third aspect of the present disclosure, the present disclosure provides

a computer device, including a memory, a processor, and a computer program stored on the memory and executable on the processor. The program, when executed by the processor, implements the method according to any one of embodiments of the first aspect of the present disclosure. Therefore, only the NIPT data of maternal plasma is required, and the fetal nucleic acid concentration can be quickly determined with the aids of the linkage disequilibrium relationship, and it can be applied to low-depth sequencing data, being not specific to the fetal nucleic acid concentration and the sample type.

[0040] According to a fourth aspect of the present disclosure, the present disclosure provides a computer readable storage medium having a computer program stored thereon. The program, when executed by a processor, implements the method according to any one of the embodiments of the first aspect of the present disclosure. Therefore, only the NIPT data of maternal plasma is required, and the fetal nucleic acid concentration can be quickly determined with the aids of the linkage disequilibrium relationship, and it can be applied to low-depth sequencing data, not specific to the fetal nucleic acid concentration and the sample type.

BRIEF DESCRIPTION OF DRAWINGS

[0041] FIG. 1 is a structural schematic diagram of a device for determining a fetal nucleic acid concentration in maternal blood according to an embodiment of the present disclosure.

[0042] FIG. 2 is a structural schematic diagram of a calculation unit in a device for determining a fetal nucleic acid concentration in maternal blood according to an embodiment of the present disclosure.

[0043] FIG. 3 is a schematic diagram of a method for determining a fetal nucleic acid concentration through model prediction according to an embodiment of the present disclosure.

[0044] FIG. 4 is a graph illustrating prediction results of fetal nucleic acid concentrations of a 1000-sample test data set in accordance with a model obtained by using 10,000 samples as a training set according to an embodiment of the present disclosure.

the LD information is used to comprehensively estimate the fetal nucleic acid concentration of the whole genome (or chromosome level). When performing the imputation on a single sample, the genotype inference algorithms such as IMPUTE2 have the same premise that the analyzed sample is diploid. Therefore, when genotypes that contradict this premise (i.e., more than two
5 haplotypes) are present at some sites, these sites will be regarded as error sites and will be corrected. Since the maternal plasma actually contains information about three types of haplotype information, i.e., two maternal haplotypes and one fetal haplotype inherited from the father, there is a certain probability that the fetal haplotype derived from the father is deemed to be error site and thus corrected during the process of imputation. Such a probability of being
10 corrected is further correlated with the fetal cfDNA concentration in maternal plasma.

[0055] Using the information about haplotype and linkage disequilibrium in the genome of the pregnant woman, by means of the application of the haplotype information in imputation, the fetal nucleic acid concentration in a non-invasive prenatal genetic testing can be calculated based on a proportion of sites corrected by the imputation.

[0056] When aligning the sequencing data to at least a part of the reference genome, a genomic region with more evident signals can be selected as the alignment region. The mentioned genomic region with more evident signals may be reflected as a region having better coverage (the coverage of the sequencing data on the genome), relatively high minor allele frequency (indicating that the site has a high variation probability in the population), and
20 relatively high proportion of variation sites, for the purposes of further extracting characteristic information and reducing background noise interference, thereby improving the accuracy of fetal nucleic acid concentration estimation.

[0057] When selecting a strong linkage region, the strong linkage region can be determined by changing a calculation window size, for example, changing a window of 5mb to a window
25 of 10mb or to a window of the entire chromosome, so as to improve the accuracy by increasing the number of effective sites in each window.

[0058] In some embodiments of the present disclosure, the step (3) further includes: step (3-1) of determining a difference ratio between the first genotype information and the second genotype information; and step (3-2) of determining the fetal nucleic acid concentration based
30 on the difference ratio obtained in step (3-1) and a pre-trained formula, where the formula is

determined based on a plurality of training samples with known fetal nucleic acid concentrations. In view of a relationship between the plurality of known fetal nucleic acid concentrations and the difference ratio between the first genotype information and the second genotype information, different formulas or models, e.g., the linear regression model, or other models that effectively integrate all information, such as the random forest model or other deep learning models, etc., are used for training to correlate the fetal nucleic acid concentration with the difference ratio. When the difference ratio between the first genotype information and the second genotype information of a nucleic acid sample from certain maternal blood is determined, the fetal nucleic acid concentration in this maternal blood can be determined by means of the pre-trained formula.

[0059] When selecting or determining the model, additional and more complete phenotypic information of pregnant women can be added as covariates in the prediction model, thereby optimizing the prediction model and increasing the accuracy of estimation.

[0060] According to another aspect of the present disclosure, the present disclosure provides a device for determining a fetal nucleic acid concentration in maternal blood. The fetal nucleic acid concentration in maternal blood can be determined with such a device, which only requires to use the NIPT data of the maternal plasma without the assistance of other data. In addition, the device can be applied to the sequencing data of ultra-low depth. Besides, the device has not specified the sample type, and both samples involving male fetuses and samples involving female fetuses can be used. As illustrated in FIG. 1, the device includes: an alignment unit configured to determine first genotype information based on alignment between sequencing data and at least a part of a reference genome, where the sequencing data are derived from a nucleic acid sample of the maternal blood; a correction unit connected to the alignment unit, the correction unit being configured to correct the first genotype information based on reference data by using a linkage disequilibrium relationship to obtain second genotype information; and a calculation unit connected to the alignment unit and the correction unit, the calculation unit being configured to determine the fetal nucleic acid concentration based on a difference between the first genotype information and the second genotype information.

[0061] In at least some embodiments, the calculation unit, as illustrated in FIG. 2, further includes: a difference ratio calculation unit configured to determine a difference ratio between

the first genotype information and the second genotype information; and a fetal nucleic acid concentration calculation unit connected to the difference ratio calculation unit. The fetal nucleic acid concentration calculation unit is configured to determine the fetal nucleic acid concentration based on the difference ratio obtained in the difference ratio calculation unit and a pre-trained formula. The formula is determined based on a plurality of training samples with known fetal nucleic acid concentrations.

[0062] The solutions of the present disclosure will be explained below in conjunction with examples. Those skilled in the art will understand that the following examples are merely used to illustrate the present disclosure, and should not be regarded as limitations of the scope of the present disclosure. Where specific techniques or conditions are not indicated in the examples, the procedures shall be carried out in accordance with the techniques or conditions described in the literatures in the field or in accordance with the product instructions. The reagents or instruments used without indicating the manufacturers are all conventional products that are commercially available.

[0063] Example 1

[0064] Example 1 provides a method for calculating a fetal nucleic acid concentration in maternal plasma using cfDNA sequencing data of the maternal plasma as input data. The specific steps are as follows:

[0065] (1) Preliminary data processing

[0066] All the raw sequencing data (in fq format) of samples used for model-training and prediction are aligned to the human reference chromosome hg38 using the samse mode in BWA after quality control. Picard is used to remove duplicated reads in the mapping results and calculate a duplication rate. The mapping results are locally corrected using a base quality value correction BQSR function in the variation detection algorithm such as GATK. A depth distribution of each sample is calculated using the Depth of Coverage function in the variation detection algorithm such as GATK. A population variation detection mode in the variation detection algorithm such as GATK is used to perform the detection of single nucleotide variation (SNV) and small fragment insertion-deletion (Indel).

[0067] (2) Extraction of raw genotype information

[0068] The results of BWA alignment and Picard duplication-removing (in bam format) in

CLAIMS

What is claimed is:

1. A method for determining a fetal nucleic acid concentration in maternal blood, comprising:

step 1 of determining first genotype information based on alignment of sequencing data with at least a part of a reference genome, wherein the sequencing data are derived from a nucleic acid sample of the maternal blood;

step 2 of correcting the first genotype information based on reference data by using a linkage disequilibrium relationship, to obtain second genotype information; and

step 3 of determining the fetal nucleic acid concentration based on a difference between the first genotype information and the second genotype information.

2. The method according to claim 1, wherein the sequencing data are obtained by sequencing the nucleic acid sample of the maternal blood.

3. The method according to claim 1, wherein the reference genome comprises at least one strong linkage region in a human genome.

4. The method according to claim 3, wherein the strong linkage region has a length ranging from 5mb to 10 mb.

5. The method according to claim 1, wherein the first genotype information is determined based on a support number of sequencing reads.

6. The method according to claim 1, wherein the first genotype information comprises at least one of SNV or Indel.

7. The method according to claim 1, wherein the reference data comprise a plurality of pieces of variation site information and a plurality of pieces of variation frequency information.

8. The method according to claim 1, wherein the correcting is performed through IMPUTE2.

9. The method according to claim 1, wherein the step 3 further comprises:

step 3-1 of determining a difference ratio between the first genotype information and the second genotype information; and

step 3-2 of determining the fetal nucleic acid concentration in accordance with the difference ratio

obtained in the step 3-1 and a pre-trained formula, wherein the pre-trained formula is determined based on a plurality of training samples with known fetal nucleic acid concentrations.

10. A device for determining a fetal nucleic acid concentration in maternal blood, comprising:

an alignment unit configured to determine a first genotype information based on alignment between sequencing data and at least a part of a reference genome, wherein the sequencing data are derived from a nucleic acid sample of the maternal blood;

a correction unit connected to the alignment unit and configured to correct the first genotype information based on reference data by using a linkage disequilibrium relationship to obtain second genotype information; and

a calculation unit connected to the alignment unit and the correction unit, the calculation unit being configured to determine the fetal nucleic acid concentration based on a difference between the first genotype information and the second genotype information.

11. The device according to claim 10, wherein the sequencing data are obtained by sequencing the nucleic acid sample of the maternal blood.

12. The device according to claim 10, wherein the reference genome comprises at least one strong linkage region in a human genome.

13. The device according to claim 12, wherein the strong linkage region has a length ranging from 5mb to 10 mb.

14. The device according to claim 10, wherein the first genotype information is determined based on a support number of sequencing reads.

15. The method according to claim 10, wherein the first genotype information comprises at least one of SNV or Indel.

16. The method according to claim 10, wherein the reference data comprise a plurality of pieces of variation site information and a plurality of pieces of variation frequency information.

17. The method according to claim 10, wherein the correction is performed through IMPUTE2.

18. The method according to claim 10, wherein the calculation unit further comprises:

a difference ratio calculation unit configured to determine a difference ratio between the first genotype information and the second genotype information; and

a fetal nucleic acid concentration calculation unit connected to the difference ratio calculation unit and configured to determine the fetal nucleic acid concentration in accordance with the difference ratio obtained in the difference ratio calculation unit and a predetermined training formula, wherein the training formula is determined based on a plurality of reference samples with known fetal nucleic acid concentrations.

19. A computer device, comprising a memory, a processor, and a computer program stored on the memory and executable on the processor,

wherein the program, when executed by the processor, implements the method according to any one of claims 1 to 9.

20. A computer readable storage medium having a computer program stored thereon, wherein the program, when executed by a processor, implements the method according to any one of claims 1 to 9.