

1. A method for detecting gene mutations, wherein the method comprises the steps of:

step A, performing high-throughput sequencing of cell-free DNAs in maternal peripheral blood to obtain sequencing data;

step B, aligning the sequencing data with those of a reference genome to obtain SNP sites;

step C, performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; and

step D, identifying mutations leading to fetal gene mutation from the fetal genotype in the target mixed genotype;

wherein the mixed genotype refers to the pseudo-tetraploid genotype, which is composed of genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBA, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBA, BBAB, and BBBB, where A represents the reference allele of each SNP sites, and B represents the mutant allele of each SNP sites. The seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

2. The method according to claim 1, wherein when the initial fetal concentration is not the true fetal concentration, the step of obtaining the target mixed genotype comprises:

step C1, performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the

SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites;

step C2, selecting the initial mixed genotype suitable for calculating a second fetal concentration as a second mixed genotype;

step C3, calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data;

step C4, comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ;

step C5, assessing the relationship between the difference value Δf and a pre-defined value; and

step C6, when Δf is greater than the pre-defined value, repeating steps C1 to C5 with the f' as f ; and when the Δf is less than or equal to the pre-defined value, taking the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype.

3. The method according to claim 1 or 2, wherein the step of performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises:

obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1,

$$\sum P(G_j | S) = 1 \quad (1)$$

wherein, G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, $P(G_j|S)$ represents the probability of the mixed genotype G_j at a SNP site under the S condition;

obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from the formula (2) by selecting any one mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number of occurrence of the mutant allele at the SNP sites, the number of occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ;

then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with maximum occurrence probability as the mixed genotype with maximum probability at the i -th SNP site.

4. The method according to claim 3, wherein $P(G_{ij})$ in the formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the

following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

5. The method according to claim 3, wherein $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number of occurrence of the mutant allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele when the mixed genotype of the i -th SNP site is G_{ij} .

6. The method according to claim 5, wherein depending on the mixed genotype G_{ij} , the theoretical probability $f(b)$ of the occurrence of a mutant allele is respectively calculated as follows, when a mixed genotype of the i -th SNP site is G_{ij} :

when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0;
when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; and
when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$;
when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5;
when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$;
when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and
when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1;
wherein the f represents the initial fetal concentration.

7. The method according to claim 2, wherein the initial fetal

concentration is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%; and preferably the pre-defined value is ≤ 0.001 .

8. The method according to claim 2, wherein the second mixed genotype is selected from any one or two or more of the following four mixed genotypes: AAAB, ABAA, ABBB, and BBAB.

9. The method according to claim 1, wherein the steps of identifying mutations leading to fetal gene mutation from the fetal genotype in the target mixed genotype comprises:

filtering the polymorphic sites with a high incidence in the human population in a fetal genotype in the target mixed genotype of each of the SNP sites to obtain preliminary candidate mutation sites;

filtering SNP sites of synonymous mutations and nonsense mutations and mutations occurring in non-conserved regions, from the preliminary candidate mutation sites to obtain candidate mutation sites; and

performing literature review and clinical data review on the candidate mutation sites to obtain the mutations leading to the fetal gene mutation.

10. A device for detecting gene mutations, wherein the device comprises:

a detection module for performing high-throughput sequencing of cell-free DNA existed in maternal peripheral blood to obtain sequencing data;

an alignment module for aligning the sequencing data with a reference genomic sequence to obtain SNP sites;

a target mixed genotype determination module for performing mixed genotyping at each SNP site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP site, and taking the mixed

genotype with the maximum probability as the target mixed genotype of each of the SNP sites; and

a mutation site screening module for identifying mutations that lead to fetal gene mutation according to the fetal genotype in the target mixed genotype of each of the SNP sites;

wherein the mixed genotype refers to the pseudo-tetraploid genotypes, which is composed of genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

11. The device according to claim 10, wherein when the initial fetal concentration is not the true fetal concentration, the target mixed genotype determination module comprises:

a pre-estimation module for performing mixed genotyping for each SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype;

a selection module for selecting the initial mixed genotype suitable for calculating a second fetal concentration as a second mixed genotype;

a calculation module for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data;

a comparison module for comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ;

an assessment module for assessing a relationship between the difference

value Δf and a pre-defined value;

an iteration module for repeatedly executing the pre-estimation module, the selection module, the calculation module, the comparison module and the assessment module with the f' as f , when the Δf is greater than the pre-defined value; and

a labelling module for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype when the Δf is less than or equal to the pre-defined value.

12. The device according to claim 10 or 11, wherein the step of performing mixed genotyping at each SNP site by the target mixed genotype determination module using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP sites comprises:

obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes,

$$\sum P(G_j | S) = 1 \quad (1)$$

S represents one of the SNP sites, and $P(G_j | S)$ represents the probability of the mixed genotype G_j at a SNP site under the S condition;

obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from the formula (2) by selecting any mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number of occurrence of the mutant allele at each SNP site, the number occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ;

then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with maximum occurrence probability as the initial mixed genotype with maximum probability at the i -th SNP site.

13. The device according to claim 12, wherein $P(G_{ij})$ in the formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

14. The device according to claim 12, wherein $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number of occurrence of the mutant allele at the i -th SNP site, k represents the number of occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele in the fetus when the mixed genotype of the i -th SNP site is G_{ij} .

15. The device according to claim 14, wherein depending on the mixed genotype G_{ij} , the theoretical probability $f(b)$ of the occurrence of a mutant allele in the fetus is respectively calculated as follows, when a mixed genotype of the i -th SNP site is G_{ij} :

when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0;
when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$;
when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$;
when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5;
when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$;
when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and
when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1;
wherein the f represents the initial fetal concentration.

16. The device according to claim 11, wherein the initial fetal concentration in the pre-estimation module is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%, and more preferably, the pre-defined value in the assessment module is ≤ 0.001 .

17. The device according to claim 11, wherein the second mixed genotype in the calculation module is selected from any one or more of the following four mixed genotypes: AAAB, ABAA, ABBB, and BBAB.

18. The device according to claim 11, wherein the mutation site screening module comprises:

a high-incidence polymorphic site filtration sub-module for filtering out polymorphic sites with high incidence in the human population in a fetal genotype in the target mixed genotype of each of the SNP sites to obtain preliminary candidate mutation sites;

a gene mutation site screening sub-module for filtering SNP sites of synonymous mutations, nonsense mutations and mutations occurring in non-conserved regions, from the preliminary candidate mutation sites to obtain candidate mutation sites; and

a literature and clinical data review sub-module for performing literature review and clinical data review on the candidate mutation sites to obtain the mutations leading to the fetal gene mutation.

19. A kit for genotyping of a pregnant woman and a fetus, wherein the kit comprises:

reagents and apparatuses for enriching cell-free DNA from a maternal peripheral blood plasma and performing high-throughput sequencing;

an apparatus for aligning the sequencing data obtained by the high-throughput sequencing with those of a reference genomic sequence to obtain SNP sites; and

an apparatus for performing mixed genotyping at each SNP site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP site,

and taking the mixed genotype with the maximum probability as the target mixed genotype of each of the SNP sites;

wherein the mixed genotype refers to pseudo-tetraploid genotypes composed of genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

20. The kit according to claim 19, wherein when the initial fetal concentration is not a true fetal concentration, the apparatus for obtaining the target mixed genotype of each of the SNP sites comprises:

a first calculation element for performing mixed genotyping at each SNP site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among 7 mixed genotypes of each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites;

a selection element for selecting the initial mixed genotype suitable for calculating a second fetal concentration, and recording it a second mixed genotype;

a second calculation element for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data;

a comparison element for comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ;

an assessment element for assessing whether the Δf is greater than a pre-defined value;

an iteration element for repeatedly operating the first calculation element,

the selection element, the second calculation element, the comparison element and the assessment element with the f' as f , when the Δf is greater than the pre-defined value; and

a labelling element for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype when the Δf is less than or equal to the pre-defined value.

21. The kit according to claim 19 or 20, wherein in the apparatus for obtaining the target mixed genotype, the step of performing mixed genotyping at each SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each of the SNP sites comprises:

obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents

$$\sum P(G_j | S) = 1 \quad (1)$$

any one of the seven mixed genotypes, S represents one of the SNP sites, and $P(G_j | S)$ represents the probability of the mixed genotype G_j at an SNP site under the S condition;

obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from the formula (2) by selecting any mixed genotype G_{j^*} from G_j as the reference mixed genotype:

$$P(G_{ij^*} | S_i) = \frac{P(S_i | G_{ij^*})P(G_{ij^*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number of occurrence of the mutant allele at the SNP sites, the number of occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ;

then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with maximum occurrence probability as the mixed genotype with maximum probability at the i -th SNP site.

22. The kit according to claim 21, wherein $P(G_{ij})$ in the formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

23. The kit according to claim 20, wherein $P(S_i|G_{ij})$ in the formula (4) is

calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number of occurrence of mutant allele at the i -th SNP site, k represents the number of occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence a mutant allele when the mixed genotype of the i -th SNP site is G_{ij} .

24. The kit according to claim 23, wherein $f(b)$ in the formula (7) is calculated according to the following formulas respectively, depending on the mixed genotype G_{ij} :

when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0;

when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$;

when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$;

when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5;

when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$;

when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and

when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1;

wherein the f represents the initial fetal concentration.

25. The kit according to claim 20, wherein the initial fetal concentration in the pre-estimation element is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%, and preferably the pre-defined value in the assessment element is ≤ 0.001 .

26. The kit according to claim 20, wherein the second mixed genotype in the second calculation element is selected from any one or more of the following four mixed genotypes: AAAB, ABAA, ABBB, and BBAB.

Method, Device and Kit for Detecting Fetal Genetic Mutation

Technical Field

The present invention relates to the field of biological information, and in particular to a method, device and kit for detecting a fetal gene mutation.

Background Art

Prenatal diagnosis, also known as intrauterine diagnosis, refers to the assessment of congenital diseases (comprising malformations and hereditary diseases) using various methods before the birth of the fetus. It provides the scientific basis for the termination of the pregnancy. Among them, prenatal diagnosis of hereditary diseases mainly targets chromosomal diseases and Mendelian inheritant disease. Mendelian inheritant disease refers to a disease transmitted according to Mendel' law, which is usually caused by a single gene mutation controlled by a pair of alleles, involving changes in a single nucleotide to the entire gene, therefore this type of disease is also called the single-gene defect. As of June 25, 2013, the OMIM (online mendelian inheritance in man) database has encompassed 4,912 single-gene defects with clear molecular mechanisms, involving 2,992 pathogenic genes.

In terms of current strategies of prenatal testing, genetic diagnosis can be divided into two main categories: direct genetic diagnosis and indirect genetic diagnosis. Direct genetic diagnosis means direct detection of the pathogenic gene itself, and such method is mainly applicable to families with clear gene mutation sites, types, and pathogenicity of probands. In terms of current tests of prenatal diagnosis for single-gene defects, genetic diagnosis can be divided into two types: pre-implantation genetic diagnosis (PGD) and prenatal

diagnosis during pregnancy, depending on time periods of the diagnosis. The development of high-throughput sequencing technologies has greatly accelerated the innovation of clinical detection technologies.

Prenatal diagnosis in pregnancy comprises invasive prenatal diagnosis and non-invasive prenatal diagnosis. Non-invasive prenatal diagnosis (NIPD) is also known as a prenatal diagnosis technology that is not invasive. With the discovery of cell-free fetal DNAs (cffDNAs) in maternal plasma, non-invasive prenatal diagnosis is becoming increasingly popular due to its low risk. However, due to the very slight difference between the maternal DNA and the cell-free fetal DNA, a large amount of maternal DNA background undoubtedly increases the difficulty of detecting the cell-free fetal DNA, especially in the detection of point mutations.

Recently, Liao et al. and Lo et al. performed sequencing analysis on the plasma cell-free DNAs of pregnant women with a human genome coverage up to 65x. They detected over 95% of specific paternal SNPs carried by the fetus, derived genetic maps of genomes of the fetus and pregnant woman according to the sequencing results, and successfully detected a fetus with an inherited 4-bp known mutation in a thalassemia gene from the father.

Although the above-mentioned method can be used to derive the genetic map of the genome of a fetus by sequencing analysis of the plasma cell-free DNAs from a pregnant woman, it needs to combine the genetic information derived from the father. Sequencing multiple samples would undoubtedly increase the cost of sequencing significantly, and the dependence on genetic information derived from the father may also be limited. In addition, the above-mentioned method has problems of requiring whole-genome sequencing, with high sequencing depths, and only assessing mutations associated with the paternal source. Therefore, there is still a need to improve the existing

detection method.

Summary of the Invention

The present invention aims to provide a method and a device for detecting gene mutations and a kit for performing genotyping for a pregnant woman and her fetus, so that all SNPs of the fetus within the range of sequencing data are detected while the cost of detection is reduced.

In order to achieve the above object of the present invention, according to an aspect of the present invention, a method for detecting gene mutations is provided, the method comprising the steps of: performing high-throughput sequencing of cell-free DNA in maternal peripheral blood to obtain sequencing data; aligning the sequencing data with those of a reference genome to obtain SNP sites; performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; and identifying mutant alleles that lead to fetal gene mutation according to the fetal genotype in the target mixed genotype; wherein the mixed genotype refers to pseudo-tetraploid genotype, which is composed of genotypes of the pregnant woman and her fetus, the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABAB, ABAB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABAB, ABAB, BBAB, and BBBB, where A represents the reference allele of each SNP sites, and B represents the mutant allele of each SNP sites. The seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

Further, when the initial fetal concentration is not the true fetal concentration, the step of obtaining the target mixed genotype comprises: step

C1, performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites; step C2, selecting the initial mixed genotype suitable for calculating a second fetal concentration as a second mixed genotype; step C3, calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; step C4, comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ; step C5, assessing the relationship between the difference value Δf and a pre-defined value; and step C6, when Δf is greater than the pre-defined value, repeating steps C1 to C5 with the f' as f ; and when the Δf is less than or equal to the pre-defined value, taking the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype.

Further, the step of performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1,

$$\sum P(G_j | S) = 1 \quad (1)$$

wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, $P(G_j | S)$ represents the probability of the mixed genotype G_j at a SNP site under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from formula (2) by selecting any one mixed genotype G_{j^*} from G_j as the reference mixed genotype:

$$P(G_{ij^*} | S_i) = \frac{P(S_i | G_{ij^*})P(G_{ij^*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij^*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij^*})P(G_{ij^*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j^*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number of occurrence of the mutant allele at the SNP sites, the number of occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ; then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with maximum occurrence probability as the mixed genotype with maximum probability at the i -th SNP site.

Further, $P(G_{ij})$ in the formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

Further, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number of occurrence of the mutant allele at the i -th SNP site, k represents the number of occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele when the mixed genotype of the i -th SNP site is G_{ij} .

Further, depending on the mixed genotype G_{ij} , the theoretical probability $f(b)$ of the occurrence of a mutant allele is respectively calculated as follows, when the mixed genotype of the i -th SNP site is G_{ij} : when the mixed genotype of the i -th SNP site is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the initial fetal concentration.

Further, the initial fetal concentration is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%; and more preferably the pre-defined value is ≤ 0.001 .

Further, the second mixed genotype is selected from any one or two or

more of the following four mixed genotypes: AAAB, ABAA, ABBB, and BBAB.

Further, the step of identifying mutations leading to fetal gene mutation from the fetal genotype in the mixed genotype comprises: filtering the polymorphic sites with a high incidence in the human population in a fetal genotype in the target mixed genotype of each of the SNP sites to obtain preliminary candidate mutation sites; filtering SNP sites of synonymous mutations and nonsense mutations and mutations occurring in a non-conserved regions, from the preliminary candidate mutation sites to obtain candidate mutation sites; and performing literature review and clinical data review on the candidate mutation sites to obtain the mutations leading to the fetal gene mutation.

According to another aspect of the present invention, a device for detecting gene mutations is provided, the device comprising: a detection module for performing high-throughput sequencing of cell-free DNA existed in peripheral blood of a pregnant woman to obtain sequencing data; an alignment module for aligning the sequencing data with a reference genomic sequence to obtain SNP sites; a target mixed genotype determination module for performing mixed genotyping at each SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP sites, and taking the mixed genotype with the maximum probability as the target mixed genotype of each of the SNP sites; and a mutation site screening module for identifying mutation sites that lead to fetal gene mutations according to the fetal genotype in the target mixed genotype of each of the SNP sites; wherein the mixed genotype refers to the pseudo-tetraploid genotypes, which is composed of genotypes of the pregnant woman and her fetus, the mixed

genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

Further, when the initial fetal concentration is not the true fetal concentration, the target mixed genotype determination module comprises: a pre-estimation module for calculating with a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype; a selection module for selecting the initial mixed genotype suitable for calculating a second fetal concentration as a second mixed genotype; a calculation module for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; a comparison module for comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ; a assessment module for assessing a relationship between the difference value Δf and a pre-defined value; a iteration module for repeatedly executing the pre-estimation module, the selection module, the calculation module, the comparison module and the assessment module with the f' as f , when the Δf is greater than the pre-defined value; and a labelling module for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype when the Δf is not greater than the pre-defined value.

Further, the step of performing mixed genotyping at each SNP sites by the target mixed genotype determination module using a Bayesian model and an

initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, and $P(G_j | S)$

$$\sum P(G_j | S) = 1 \quad (1)$$

represents the probability of the mixed genotype of the SNP site being G_j at an SNP site under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from formula (2) by selecting any mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i | G_{ij})$ is obtained by a binomial distribution formula using

the number occurrence of the mutant allele at each SNP sites, the number occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ; then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with the maximum occurrence probability as the mixed genotype with the maximum probability at the i -th SNP site.

Further, $P(G_{ij})$ in the above-mentioned formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

Further, $P(S_i | G_{ij})$ in the above-mentioned formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number occurrence of the mutant allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele in the fetus when the mixed genotype of the i -th SNP site is G_{ij} .

Further, depending on the mixed genotype G_{ij} , the theoretical probability $f(b)$ of the occurrence of a mutant allele in the fetus is calculated as follows respectively, when a mixed genotype of the i -th SNP site is G_{ij} : when the

mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the initial fetal concentration.

Further, the initial fetal concentration in the pre-estimation module is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%, and more preferably, the pre-defined value in the assessment module is ≤ 0.001 .

Further, the second mixed genotype in the calculation module is selected from any one or more of the following four mixed genotypes: AAAB, ABAA, ABBB, and BBAB.

Further, the mutation site screening module comprises: a high-incidence polymorphic site filtration sub-module for filtering out polymorphic sites with high incidence in the human population in a fetal genotype in the target mixed genotype of each of the SNP sites to obtain preliminary candidate mutation sites; a gene mutation screening sub-module for filtering SNP sites of synonymous mutations, nonsense mutations and mutations occurring in non-conserved regions, from the preliminary candidate mutation sites to obtain candidate mutation sites; and a literature and clinical data review sub-module for performing literature review and clinical data review on the candidate mutation sites to obtain the mutations site leading to the fetal gene mutation.

According to still another aspect of the present invention, a kit for genotyping of a pregnant woman and her fetus is provided, the kit comprising: reagents and apparatuses for enriching cell-free DNA from peripheral blood

plasma of the pregnant woman and performing high-throughput sequencing; an apparatus for aligning the sequencing data obtained by the high-throughput sequencing with those of a reference genomic sequence to obtain SNP sites; and an apparatus for obtaining a mixed genotype with the maximum probability among seven mixed genotypes of each SNP sites using the Bayesian model and an initial fetal concentration f , and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; wherein the mixed genotype refers to pseudo-tetraploid genotypes composed of genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

Further, when the initial fetal concentration is not the true fetal concentration, the apparatus for obtaining the target mixed genotype of each of the SNP sites comprises: a first calculation element for performing mixed genotyping at each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among 7 mixed genotypes of each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites; a selection element for selecting the initial mixed genotype suitable for calculating a second fetal concentration, and recording it the second mixed genotype; a second calculation element for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; a comparison element for comparing the second fetal

concentration f' with the initial fetal concentration f to obtain a difference value Δf ; an assessment element for assessing whether the Δf is greater than a pre-defined value; an iteration element for repeatedly operating the first calculation element, the selection element, the second calculation element, the comparison element and the assessment element with the f' as f , when the Δf is greater than the pre-defined value; and a labelling element for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype when the Δf is not greater than the pre-defined value.

Further, in the apparatus for obtaining the target mixed genotype, the step of performing mixed genotyping at each SNP sites using the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, and

$$\sum P(G_j | S) = 1 \quad (1)$$

$P(G_j | S)$ represents the probability of the mixed genotype G_j at a SNP site under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from formula (2) by selecting any one

mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number occurrence of the mutant allele at each SNP sites, the number occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ; then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with the maximum occurrence probability as the initial mixed genotype at the i -th SNP site.

Further, $P(G_{ij})$ in the formula (4) is obtained by multiplying the probability of genotype G' of the pregnant woman and the probability of genotype G' of the fetus, which are calculated using the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is the population mutation frequency of the i -th SNP site.

Further, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number occurrence of the mutant allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele in the fetus when the mixed genotype of the i -th SNP site is G_{ij} .

Further, the $f(b)$ in the formula (7) is calculated according to the following formulas respectively, depending on the mixed genotype G_{ij} : when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the initial fetal concentration.

Further, the initial fetal concentration in the pre-estimation element is a pre-estimated fetal concentration, preferably the pre-estimated fetal concentration is 10%, and the pre-defined value in the assessment element is ≤ 0.001 .

Further, the second mixed genotype in the second calculation element is selected from any one or more of the following four mixed genotypes: AAAB, ABAA, ABBA, and BBAB.

Applying the technical solutions of the present invention, SNP sites having mixed maternal and fetal genomic informations can be obtained by high-throughput sequencing and alignment with a reference genomic sequence,

and genotypes of cell-free fetal DNA and the mother's own DNA in the peripheral blood of the pregnant woman can be typed using the pseudo-tetraploid genotyping model proposed by the present invention, thereby enabling the detection of all possible gene mutations in the fetus only using peripheral blood of the pregnant woman. The method of the present invention reduces separate sequencing for samples derived from the father and/or mother; and has not special requirement of the sequencing technology, wherein the target region sequencing can be used to obtain sequencing data, thereby reducing the cost of sequencing. Furthermore, the present method can detect fetal gene mutations at all SNP sites within the range of sequencing data, providing convenient and diversified services for prenatal diagnosis.

Brief Description of the Drawings

The accompanying drawings of the descriptions constitute a part of the present application, and are used for providing further understanding of the present invention, the illustrations embodiments of the present invention and thereof are intended to explain the present invention and are not intended to limit the invention. In the drawings:

Figure 1 shows a flow chart of a method for detecting gene mutations in a preferred embodiment of the present application;

Figure 2 shows an operation flowchart in Example 1 of the present application; and

Figure 3 shows a graphical results of verification using the existing mutation detection method on the gene mutation detected by the method of the present application in Example 2.

Detailed Description of Embodiments

It should be noted that the embodiments and the features in the embodiments in the present application may be combined with each other without conflict. The invention will be described in detail below through the drawings in conjunction with the embodiments.

Glossaries:

The population mutation frequency refers to the proportion of mutation of a gene in a particular population, for example the mutation frequency per thousand Asian people.

The pre-defined value reflects the level of detection resolution, and can be reasonably set according to the actual situation of sequencing. For example, when the sequencing depth is $\geq 1000\times$, the preferred pre-defined value is ≤ 0.001 .

Fetal concentration is the ratio of cell-free fetal DNAs in plasma of a pregnant woman to total cell-free DNAs in plasma. The fetal concentration f can be obtained by experimental methods well known to those skilled in the art, or can be preliminarily pre-estimated according to common knowledge in the art, for example, 5% to 20%.

In the present invention, the high-throughput sequencing of cell-free DNAs in the peripheral blood of pregnant women can be either whole genome sequencing (WGS) or target region capture sequencing of the genes of interest.

In the present invention, the mixed genotype refers to a pseudo-tetraploid genotype composed of genotypes of a pregnant woman and a fetus, and both A and B are haplotypes. The first two haplotypes of the mixed genotype represent the diploid genotype of the mother, and the latter two haplotypes represent the diploid genotype of the fetus. A represents a reference allele of each of the SNP site, and B represents a mutant allele of each of the SNP site. For a site of the sequencing data, if it is consistent with the base of the corresponding site in the

reference genome, it is a reference genotype, and otherwise it is a mutant genotype. The mixed genotype, for example, may be AAAB, which means that the diploid genotype of the mother is AA and is a homozygous reference type, and the diploid genotype of the fetus is AB and is a mutation carrying type.

In the present invention, the population mutation frequency refers to the proportion of the number of cells or individuals in which a mutation occurs within a specific population, for example the mutation frequency per thousand Asians.

As mentioned in the background art section, the method for detecting fetal gene mutations using high-throughput sequencing in the prior art typically requires additional paternal and maternal sample information and can only detect a Y chromosome-linked monogenic disease. In order to reduce the detection cost and provide diversified prenatal testing services, a method for detecting a gene mutation is provided in a typical embodiment of the present invention, as shown in Figure 1, the method comprising the steps of: performing high-throughput sequencing of cell-free DNAs in peripheral blood of a pregnant woman to obtain sequencing data; aligning the sequencing data with a reference genome to obtain SNP sites; performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; and identifying out mutation sites leading to fetal gene mutations according to genotypes of the fetus in the target mixed genotype; wherein the mixed genotype refers to pseudo-tetraploid genotypes formed by genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types consisting of AAAA, AAAB, ABAA, ABAB, ABBA, BBAB, and BBBB, and

the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites.

Based on the fact that cell-free DNA in the peripheral blood of the pregnant woman comprises both the maternal DNA and the fetal DNA, and current technical means are difficult to completely separate the DNA from two sources, the concept of pseudo-tetraploid is proposed by the inventor, the tetraploid obtained by mixing genotypes of the pregnant woman and the fetus is called pseudo-tetraploid, and at each site of the genome, the genotype of the site obtained by mixing the genotype of the pregnant woman and the genotype of the fetus is called a mixed genotype. In order to assess the probability of occurrence of a mutant genotype at each site, A represents the normal reference allele at that site; B represents the mutant allele at the site.

By placing the diploid genotype at each site of the pregnant woman in front, and placing the diploid genotype at the corresponding site of the fetus behind to indicate a mixed genotype at a site of the pseudo-tetraploid, seven possible mixed genotypes of AAAA, AAAB, ABAA, ABAB, ABBB, BBAB and BBBB can be obtained, a mixed genotype with the maximum probability at each of the SNP sites can be deduced from sequencing data to obtain a target mixed genotype at the site, thereby obtaining the genotype of the fetus from the target mixed genotype.

Further, the inventors proposed the idea of mixed genotyping of the above-mentioned pseudo-tetraploid using conditional probability and the Bayesian model.

By the above-mentioned method of the present invention, SNP sites having mixed maternal and fetal genomic informations can be obtained only

by performing high-throughput sequencing and sequence alignment of cell-free DNA in peripheral blood of the mother; the fetal and the maternal genotype at each of the SNP sites of cell-free DNA in the peripheral blood of the pregnant woman can be determined based on the concept of mixed genotyping proposed by the present invention, thereby achieving detection of all possible gene mutations in the fetus using only peripheral blood of the pregnant woman. On one hand, the present method reduces the sequencing of the paternal and maternal samples, and reduces the cost of sequencing; and on the other hand, it also facilitates the detection of fetal gene mutations under certain special conditions, such as the case where the paternal sample is not available, and thus provides diversified services for prenatal diagnosis.

In the above-mentioned method of the present invention, based on the concepts of pseudo-tetraploid and mixed genotyping of pseudo-tetraploid proposed by the present invention, those skilled in the art can perform genotyping for a mixed genotype of pseudo-tetraploid using the conditional probability and a Bayesian model, so as to obtain the genotype of the fetus at the SNP site, which lays a foundation for screening mutation sites that cause fetal gene mutations. According to the sources of peripheral blood samples of pregnant women, an initial fetal concentration is divided into two cases: known or unknown. When the fetal concentration is known, the initial fetal concentration f is a true fetal concentration. A mixed genotype with the maximum probability for each of the SNP sites can be calculated using the initial fetal concentration f and the Bayesian model. When the fetal concentration is unknown, a derivation process for the true fetal concentration is required.

In a preferred embodiment of the present invention, when the initial fetal concentration is not the true fetal concentration, the step of obtaining a target

mixed genotype comprises: step C1, performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites; step C2, selecting the initial mixed genotype suitable for calculating a second fetal concentration as a second mixed genotype; step C3, calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; step C4, comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference Δf ; step C5, determining a relationship between the difference Δf and the pre-defined value; step C6, when Δf is greater than the pre-defined value, repeating steps C1 to C5 with the f' as f ; when the Δf is not greater than the pre-defined value, taking the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype.

In the above-mentioned process for mixed genotyping of pseudotetraploid, since the fetal concentration is unknown, the occurrence probability of seven possible mixed genotypes is calculated at a fetal concentration that is preliminarily pre-estimated according to common sense, for example, 5% to 15%, as the initial fetal concentration, thereby obtaining the mixed genotype with the maximum probability at each of the SNP sites. In combination with the actual sequencing data, the actual fetal concentration is calculated by using the mutations from the mother or the fetus, and then the calculated fetal concentration is then compared with the pre-estimated initial fetal concentration. If the difference is less than a pre-defined value, then the calculated fetal concentration needs to be taken as the initial concentration in step C1, and the steps C1 to C5 are repeated until the difference between the

calculated fetal concentration at some point and the initial concentration in step C1 of the cycle is less than the pre-defined value, After the termination of the cycle, the mixed genotype with the maximum probability at each of the SNP sites at the initial concentration in the step C1 of the cycle is recorded as the target mixed genotype of each of the SNP sites. The above-mentioned pre-defined value reflects the level of the detection resolution and can be reasonably set according to the actual situation. For example, when the sequencing depth is $\geq 1000\times$, a preferred pre-defined value is ≤ 0.001 .

In the above-mentioned preferred embodiment, the selection principle of the mixed genotype for facilitating calculation of the fetal concentration can be rationally selected according to the calculation method. In a preferred embodiment of the invention, the above-mentioned mixed genotype for calculation of the fetal concentration includes, but not limited to, any one or more of AAAB, ABAA, ABBB, and BBAB. The mutant alleles or reference alleles in these mixed genotypes are only from the pregnant woman or the fetus, and the concentration of one of them can be calculated based on the number of times the mutant alleles and the reference alleles are detected in the sequencing data, so that it is very easy to obtain the concentration in the fetus.

In a preferred embodiment of the present invention, the above-mentioned step of performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites,

$$\sum P(G_j | S) = 1 \quad (1)$$

and $P(G_j|S)$ represents the probability of the mixed genotype of the SNP site being G_j at an SNP site under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the occurrence probability of G_j genotype at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively; and obtaining the following formula (3) from the formula (2) by selecting any one mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number occurrence of a mutant allele at each of the SNP sites, the number occurrence of a reference allele corresponding to the mutant allele, and the initial fetal concentration f ; then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with

maximum occurrence probability as the mixed genotype with maximum probability at the i -th SNP site.

In the above-mentioned preferred embodiment of the present invention, the method for calculation of the mixed genotype with the maximum probability at a SNP site is converted into a calculation of a ratio between an occurrence probability of the above-mentioned seven mixed genotypes at the SNP site and an occurrence probability of one of the mixed genotypes, so that the mixed genotype with the maximum occurrence probability at the SNP site is indirectly obtained, and thus is inferred to be an initial mixed genotype of the site.

In the above-mentioned method of the present invention, in the case that a mutant genotype is known to occur at a site, those skilled in the art can calculate $P(G_{ij})$ in the formula (4) using the population mutation frequency in the thousand human genome database. In a preferred embodiment of the present invention, $P(G_{ij})$ is calculated by the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6).$$

In the above-mentioned formula (6), G' represents the above-mentioned three separate possible genotypes occurring at a site in the pregnant woman or the fetus, and then an occurrence probability of a mixed genotype at the specific site is the product of the occurrence probability of genotype G' at the site of the pregnant woman and the occurrence probability of genotype G' at the site of the fetus, wherein θ is the population mutation frequency of the i -th SNP site. The parameter θ is obtained from the population mutation frequency in the thousand human genome database.

In the above-mentioned method of the present invention, the numerical

value of $P(S_i|G_{ij})$ is obtained depending on the difference between the number occurrence of mutant alleles and the number occurrence of reference alleles in actual sequencing data, and the difference in the initial fetal concentration when a site is a specific mixed genotype G_{ij} , using a binomial distribution formula. In a specific embodiment of the present invention, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k + r - 1}{r - 1} \times f(b)^r \times (1 - f(b))^k$$

wherein r represents the number occurrence of the mutant allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of the mutant allele of the fetus when the mixed genotype of the i -th SNP site is G_{ij} .

In the above-mentioned embodiment, $f(b)$ is related to the concentration of cell-free fetal DNA in peripheral blood of the pregnant woman, and can be calculated using a conventional fetal concentration calculation method such as Fetal Quant (see Lench N, Barrett A, Fielding S, *et al.* The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made [J]. Prenatal diagnosis, 2013, 33(6): 555-562.).

After obtaining the fetal concentration, when a site is one of the mixed genotypes, a possible paternal genotype is derived for the mixed genotype, thereby deducing the theoretical occurrence probability of a mutant allele when a specific mixed genotype occurs.

As described above, in the present invention, when the difference value between the final second fetal concentration calculated from the initial fetal concentration by iteration using the expectation maximization algorithm and the initial fetal concentration is not significantly different from the pre-defined value, the initial fetal concentration or the second fetal concentration in this

case is the true concentration of the fetus in the sample. Assuming that the initial fetal concentration (pre-estimated fetal concentration) f is 10%, the mixed genotypes of all SNP sites when $f=10\%$ are calculated; the fetal concentration f' (second fetal concentration f') is calculated according to frequencies of the reference allele and the mutant allele actually detected for a mixed genotype; and if the difference value between f' and f is less than the pre-defined value, then the iteration ends, and the corresponding f' at the end of the iteration is the true fetal concentration. More preferably, the above-mentioned pre-defined value is less than or equal to 0.001.

Specifically, the following algorithm is used to calculate the true fetal concentration:

step 0: pre-estimating that fetal concentration f is 10%;

iteration:

step 1: inferring the fetal genotype according to the mixed genotyping, and calculating the fetal concentration f' according to the $f(b)$ in the genotype;

step 2: calculating Δf , wherein $\Delta f = D(f - f')$;

step 3: if $\Delta f < \varepsilon$, the iteration ends, wherein ε represents any small positive number; and

step 4: obtaining the fetal concentration $f = f(b)$;

wherein the function $D()$ represents a distance function, which measures a difference between two variables.

The process of calculation of the true fetal concentration herein is illustrated below by examples.

For example, 3 SNP sites are selected for calculation, all the genotypes are assumed as AAAB type, f is the initial pre-estimated fetal concentration, and f' is the second fetal concentration deduced from the detected frequencies of A and B.

For the first SNP, A and B are detected 19 times and once respectively, assuming $f=10\%$, the probability values of the 7 genotypes are calculated, and it is derived that the mixed genotype of the SNP is AAAA after comparison, which does not meet the hypothesis of AAAB, and the first SNP should be removed;

for the second SNP, A and B are detected 16 times and 4 times respectively, assuming $f=10\%$, the probability values of the 7 genotypes are calculated, and it is derived that the mixed genotype of the SNP is AAAB after comparison, which meets the hypothesis of AAAB, and in this case, $f'=40\%$; and

for the third SNP, A and B are detected 18 times and twice respectively, assuming $f=10\%$, the probability values of the 7 genotypes are calculated, and it is derived that the mixed genotype of the SNP is AAAB after comparison, which meets the hypothesis of AAAB, and in this case, $f'=20\%$.

Combining cases of the above three SNPs, the second and the third cases are accepted, the first one is excluded, then the average value of f' is 30%, and the difference value between the assumed f and the detected and deduced f' is greater than 0.001; therefore, iterative calculation needs to continue until the difference value there between is less than 0.001.

The above-mentioned method for calculating the fetal concentration f by the iterative algorithm of the present invention has advantages of high accuracy and no limitation by the gender, as compared with the method for calculating the fetal concentration f using X chromosome in a male fetus and a methylation method in a female fetus in the prior art.

After obtaining the fetal concentration using the above-mentioned method, depending on the mixed genotype G_{ij} the theoretical probability $f(b)$ of the occurrence of the mutant allele can be respectively calculated according to the

following formulas, when a mixed genotype of the i -th SNP site is G_{ij} : when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the initial fetal concentration.

$f(b)$ refers to the probability of the mutant allele in the mixed genotype of pseudo-tetraploid, and thus only the occurrence of B in the mixed genotype of pseudo-tetraploid needs to be calculated, as shown in the Table 1 below: (assuming that the probability of mixed genotypes of pseudo-tetraploid is 1)

Table 1:

Mixed genotype	Mother		Fetus		Occurrence probability $f(b)$ of B genotype
	Genotype	Concentration	Genotype	Concentration	
1-AAAA	AA	$1-f$	AA	f	B does not occur, so it is 0
2-AAAB	AA	$1-f$	AB	f	$f/2$
3-ABAA	AB	$1-f$	AA	f	$(1-f)/2=0.5-f/2$
4-ABAB	AB	$1-f$	AB	f	$1/2$
5-ABBB	AB	$1-f$	BB	f	$(1-f)/2+f=0.5+f/2$
6-BBAB	BB	$1-f$	AB	f	$1-f/2$
7-BBBB	BB	$1-f$	BB	f	1

In the method of the present invention, the above-described mixed genotyping enables the deducing of the target mixed genotype of each SNP sites, thereby obtaining the genotype of the fetus. After obtaining the genotype of the fetus, the pathogenic mutations leading to a gene mutation can be found. In a particular embodiment of the present invention, the step of identifying the mutation from SNP sites according to a difference in the fetal genotype in the target mixed genotype of each of the SNP sites comprises: filtering the

polymorphic sites with a high incidence in the human population in various SNP sites for which fetal genotypes are deduced, to obtain preliminary candidate mutation sites; filtering SNP sites of synonymous mutations, nonsense mutations and mutations occurring in non-conserved regions, from the preliminary candidate sites to obtain candidate mutation sites; and performing literature review and clinical data review on the candidate mutation sites to obtain the mutations leading to the fetal gene mutation.

In the above-mentioned embodiment, in the process of analyzing each SNP site of the fetus to find the pathogenic mutations, the high-frequency SNP sites which cause differences between different individuals in the human population are deleted, because these sites are obviously not the pathogenic mutations. In the present invention, the high-incidence polymorphic sites in the human population are removed using the dnSNP135 public database and the Freq_1000g2012feb (thousand human genome) database which have been collated by the medical community. After removing the SNP sites caused by individual differences, preliminary candidate mutations are obtained, and then mutation prediction softwares commonly used in the field is used to filter harmful mutations, for example, ANNOVAR software can screen whether the mutations cause amino acid change, that is, whether the mutations cause asense mutations, and can also filter whether the mutations occur in conserved sequence regions.

After the above-mentioned software filtering, it is also necessary to perform artificial interpretation of possible pathogenic mutations that have been identified. The so-called “artificial interpretation” means to find the site information associated with a monogenic disease from possible pathogenic mutations by the review of existing databases and literatures, and perform corresponding interpretation. Furthermore, the method of the present invention

is not limited to detecting the presence or absence of hot spot mutations leading to known monogenic diseases, it can also detecting non-hot spot mutations of known monogenic diseases and unreported potential pathogenic genes and their mutations. Therefore, the method can provide diversified services to customers according to their different needs.

In the above-mentioned methods of the present invention, the step of performing high-throughput sequencing of cell-free DNA in peripheral blood of a pregnant woman to obtain sequencing data comprises performing sample DNA library construction at first using the high-throughput sequencing method commonly used in the art, and then sequencing using existing high-throughput sequencing platforms. In a preferred embodiment of the present invention, the step of performing high-throughput sequencing of cell-free DNA in peripheral blood of a pregnant woman to obtain sequencing data comprise: extracting plasma DNA from the peripheral blood of the pregnant woman, and enriching cell-free DNA in the plasma DNA to obtain enriched DNA; performing library construction for the enriched DNA to obtain a sequencing library; and performing high-throughput sequencing of the sequencing library to obtain the sequencing data.

In the above-mentioned preferred embodiment, since the amount of the cell-free DNA in the peripheral blood of the pregnant woman is relatively low in the maternal plasma, which is substantially not more than 10%, the step of extraction and enrichment of the cell-free DNA is required before the high-throughput sequencing. For the extraction and enrichment step, suitable extraction and enrichment methods can be selected by those skilled in the art depending on the diversity of samples and the requirement of data respectively. For example, QIAmp DNA Blood Mini Kit from Qiagen, Germany, or commercially available similar reagents from other companies, or self-made

relevant reagents for extraction and enrichment of peripheral blood of a pregnant woman can be used for the extraction and enrichment.

After the step of performing library construction for the above-mentioned enriched DNA to obtain a sequencing library, different target regions are selected for sequencing, depending on detection purposes of different samples. During the actual operation of the present invention, the step of performing target region sequencing on the library to obtain a sequencing library containing the target regions is also included before performing the high-throughput sequencing. In a more preferred embodiment of the invention, the step of performing exon capture hybridization on the sequencing library is added, so that the subsequent high throughput sequencing is performed only for exons. Since introns are usually cleaved off during the transcription process of a gene, and exons are final regions encoding a protein, only performing exon sequencing can increase an effective data volume and improve efficiency of sequencing. After obtaining a sequencing library for exons, it is also possible to detect mutations of known monogenic diseases within specific target regions depending on detection purposes and/or detection samples, or to detect all mutations in the sequencing data as entirety.

In the above-mentioned preferred embodiment of the invention, different methods or reagents may be selected for capture depending on the target regions to be captured. For example, a capture kit from Roche NimbleGen, US, or a self-made kit or a commercially available kit from other companies with a similar function can be used to perform target region sequencing.

In another typical implementation of the present invention, a device for detecting gene mutations is provided, the device comprising: a detection module for performing high-throughput sequencing of cell-free DNA exsited in peripheral blood of a pregnant woman to obtain sequencing data; an

alignment module for aligning the sequencing data with a reference genomic sequence to obtain SNP sites; a target mixed genotype determination module for performing mixed genotyping at each SNP site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP site, and taking the mixed genotype with the maximum probability as the target mixed genotype of each of the SNP sites; and a mutation site screening module for identifying mutations that lead to fetal gene mutation according to the fetal genotype in the target mixed genotype of each of the SNP sites; wherein the mixed genotype refers to the pseudo-tetraploid genotypes, which is composed of genotypes of the pregnant woman and the fetus, and the mixed genotype is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7.

In the above-mentioned device of the present invention, SNP sites in the maternal and fetal genomic information different from those of a reference genome are obtained by the detection module and the alignment module, and a mixed genotype of the pseudo-tetraploid genotypes composed of genotypes of the pregnant woman and the fetus is typed by the target mixed genotype determination module to obtain genotypes of the mother and the fetus at each of the SNP sites, thereby achieving detection of all possible gene mutations in the fetus using only the peripheral blood sample of the pregnant woman. The device of the present invention not only reduces the separate sequencing of the paternal and maternal samples (cellular genome samples from the peripheral blood), and reduces the cost of sequencing, but also facilitates the detection of

fetal gene mutations under certain special conditions, such as the case where the paternal sample is not available, and thus provides diversified services for prenatal diagnosis.

In the above-mentioned target mixed genotype determination module of the present invention, when the initial fetal concentration is the true fetal concentration, those skilled in the art would obtain a target mixed genotype of each of the SNP sites by modifying the conditional probability and the Bayesian model, based on the concepts of pseudo-tetraploid and the mixed genotype of pseudo-tetraploid proposed by the present invention. In a preferred embodiment of the present invention, when the initial fetal concentration is not the true fetal concentration, the target mixed genotype determination module comprises: a pre-estimation module for calculating with a Bayesian model and an initial fetal concentration f to obtain the mixed genotype with the maximum probability among 7 mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype; a selection module for selecting the initial mixed genotype suitable for calculating a second fetal concentration, which is recorded as the second mixed genotype; a calculation module for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; a comparison module for comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference value Δf ; a assessment module for assessing whether the Δf is greater than a pre-defined value; a iteration module for repeatedly executing the pre-estimation module, the selection module, the calculation module, the comparison module and the assessment module with the f' as f , when the Δf is greater than the pre-defined value; and a labelling module for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype

when the Δf is not greater than the pre-defined value.

The above-mentioned specific calculation method for calculating a mixed genotype with the maximum probability among 7 mixed genotypes for each of the SNP sites using the initial fetal concentration f can be obtained by modifying the conditional probability and the Bayesian model in many ways. Preferably, when calculating the probabilities of the seven mixed genotypes, probability calculation formulas of the seven mixed genotypes are divided by a probability calculation formula of one specific mixed genotype respectively, thereby obtaining a ratio between the probability of each mixed genotype and the probability of the specific mixed genotype, a mixed genotype with the largest ratio is the mixed genotype with the maximum probability at the SNP site, i.e. the initial mixed genotype of the SNP site. The above-mentioned specific mixed genotype may be any one of seven mixed genotypes, and may be reasonably selected according to convenience of calculation.

In the calculation module in the above-mentioned preferred embodiment, the selection principle of the mixed genotype for calculation of the fetal concentration can be rationally selected according to the calculation method. In a preferred embodiment of the invention, the above-mentioned mixed genotype for calculation of the fetal concentration includes, but is not limited to, any one or more of AAAB, ABAA, ABBB, and BBAB. The mutant alleles or reference alleles in these mixed genotypes are only from the pregnant woman or the fetus, and the concentration of one of them can be calculated based on the number occurrence of the mutant alleles and the reference alleles are detected in the sequencing data, so that it is very easy to obtain the concentration of the fetus.

In another preferred embodiment of the present invention, the step of performing mixed genotyping at each SNP site by the target mixed genotype

determination module using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites,

$$\sum P(G_j | S) = 1 \quad (1)$$

and $P(G_j | S)$ represents the probability of the mixed genotype G_j at a SNP site under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from formula (2) by selecting any mixed genotype G_{j*} from G_j as the reference mixed genotype::

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population

mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number occurrence of the mutant allele at each the SNP site, the number occurrence of the reference allele corresponding to the mutant allele, and initial fetal concentration f ;

then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding the mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with the maximum occurrence probability as the initial mixed genotype with the maximum probability at the i -th SNP site.

In the above-mentioned preferred embodiment of the present invention, the method for calculation of the mixed genotype at a SNP site using the conditional probability and the Bayesian model is converted into a calculation of a ratio between an occurrence probability of the above-mentioned seven mixed genotypes at the SNP site and an occurrence probability of one of the mixed genotypes, so that the mixed genotype with the maximum occurrence probability at each of the SNP sites is indirectly obtained, and thus is inferred to be the mixed genotype with the maximum probability at each of the SNP sites.

In the above-mentioned device of the present invention, in the case that a mutant genotype is known to occur at a site, those skilled in the art can calculate $P(G_{ij})$ in the formula (4) using the population mutation frequency in the thousand human genome database. In a preferred embodiment of the present invention, in the above-mentioned fetal genotype determination module (a target mixed genotype determination module), $P(G_{ij})$ in the formula (4) is calculated by the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6).$$

In the above-mentioned formula (6), G' represents the above-mentioned three separate possible genotypes occurring at each site in the pregnant woman or the fetus, and then an occurrence probability of a mixed genotype at the specific site is the product of the probability of genotype G' at the site of the pregnant woman and the probability of genotype G' at the site of the fetus, wherein θ is the population mutation frequency of the i -th SNP site. The parameter θ is obtained from the population mutation frequency in the thousand human genome database.

In the above-mentioned devices of the present invention, in the fetal genotype determination module, the numerical value of $P(S_i|G_{ij})$ in the formula (4) is obtained depending on the difference between the number occurrence of mutant alleles and the number occurrence of reference alleles in actual sequencing data, and the difference in the initial fetal concentration when a site is a specific mixed genotype G_{ij} , using a binomial distribution formula. In a specific embodiment of the present invention, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k + r - 1}{r - 1} \times f(b)^r \times (1 - f(b))^k \quad (7)$$

wherein r represents the number occurrence of the allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of the occurrence of a mutant allele in the fetus when the mixed genotype of the i -th SNP site is G_{ij} .

In the above-mentioned embodiment, $f(b)$ is related to the concentration of cell-free fetal DNA in peripheral blood of the pregnant woman, and can be

calculated using a conventional fetal concentration calculation method such as Fetal Quant (see Lench N, Barrett A, Fielding S, *et al.* The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made [J]. Prenatal diagnosis, 2013, 33(6): 555-562.).

After obtaining the fetal concentration, when a site is one of the mixed genotypes, a possible paternal genotype is derived for the mixed genotype, thereby deducing the theoretical occurrence probability of a mutant allele in the fetus when a specific mixed genotype occurs.

In another preferred embodiment of the present invention, the initial fetal concentration f is calculated by iteration using the expectation maximization algorithm. Assuming that the initial pre-estimated fetal concentration f is 10%, the mixed genotypes of all SNP sites when $f=10\%$ are calculated; the actual fetal concentration f' is calculated according to frequencies of the reference genotypes and the mutant alleles actually detected for some mixed genotypes; and if the difference value between f' and f is less than the pre-defined value, then the iteration ends, and the corresponding f' at the end of the iteration is the fetal concentration f . More preferably, when the above-mentioned pre-defined value is less than or equal to 0.001, the iteration ends. The specific algorithm used is the same as that in the foregoing detection method, and details are not described herein again. Similarly, when the above-mentioned mixed genotype G_{ij} of the i -th SNP site is any one of the 7 types, the theoretical probability $f(b)$ of the occurrence of the mutant allele is the same as in Table 1.

In the above-mentioned device of the present invention, in the above-mentioned target mixed genotype determination module, a target mixed genotype can be deduced by the pseudo-tetraploid genotyping for each of the SNP sites, thereby a fetal genotype at each of the SNP sites can be obtained from the target mixed genotype, and pathogenic mutations can thus be found

after finding the fetal genotype. In a typical embodiment of the present invention, the mutation site screening module in the above-mentioned device comprises: a high-incidence polymorphic site filtration sub-module for filtering out polymorphic sites with high incidence in the human population in a fetal genotype in the target mixed genotype of each of the SNP sites to obtain preliminary mutation sites; a gene mutation site screening sub-module for filtering SNP sites of synonymous mutations, nonsense mutations and mutations occurring in non-conserved regions, from the preliminary mutation sites to obtain candidate mutation sites; and a literature and clinical data review sub-module for performing screening on the candidate mutation sites to obtain the mutations leading to pathogenic gene mutations which has been recorded in literatures and clinical data.

In the above-mentioned embodiment, in the process of analyzing each SNP site of the fetus to find the pathogenic mutations, the high-frequency SNP sites which cause individual differences in the human population are deleted by the high-incidence polymorphic site filtration sub-module, because these sites are obviously not the pathogenic mutations. In the present invention, the high-incidence polymorphic sites in the human population are removed by the above-mentioned high-incidence polymorphic site filtration sub-module using the dnSNP135 public database and the Freq_1000g2012feb (thousand human genome) database which have been collated by the medical community. After removing the SNP sites caused by individual differences, fetus-specific SNP sites are obtained, and then the sites which actually cause a gene mutation is screened by the gene mutation site screening sub-module. The module can use a mutation prediction module commonly used in the art for filtering harmful mutations. ANNOVAR module can screen whether the mutations cause amino acid change, that is, whether the mutations cause sense mutations, and can also

filter whether the mutations occur in conserved sequence regions.

After the above-mentioned gene mutation site screening sub-module is subjected to the above-mentioned mutation prediction module screening, an artificial interpretation sub-module for artificial interpretation of possible pathogenic mutations that have been filtered is further included. The so-called “artificial interpretation sub-module” means to perform alignment between SNP sites which have been filtered by the mutation prediction module and pathogenic sites which are reviewed from existing databases and literatures, so as to find the site information associated with monogenic diseases and perform corresponding interpretation. The above-mentioned device of the present invention is not limited to detecting the presence or absence of hot spot mutation site leading to a known monogenic disease, it can also detect non-hot spot mutations of known monogenic diseases and unreported potential pathogenic genes and their mutations. Therefore, the method can provide diversified services to customers according to their different needs.

In the above-mentioned devices of the present invention, the detection module is a process for preparing a sequencing library from cell-free DNA enriched from peripheral blood plasma of a pregnant woman and performing high-throughput sequencing to obtain sequencing data. The step of high-throughput sequencing of cell-free DNA in peripheral blood of a pregnant woman to obtain sequencing data comprises performing sample DNA library construction at first by the high-throughput sequencing method commonly used in the art, and then sequencing using existing high-throughput sequencing platforms. In a preferred embodiment of the present invention, the above-mentioned detection device comprises a process of extracting plasma DNA from the peripheral blood of the pregnant woman, and enriching cell-free DNA in the plasma DNA to obtain enriched DNA; performing library

construction for the enriched DNA to obtain a sequencing library; and performing high-throughput sequencing of the sequencing library to obtain the sequencing data.

In the above-mentioned preferred embodiment, since the amount of the cell-free DNA in the peripheral blood of the pregnant woman is relatively low in the maternal plasma, which is substantially not more than 10%, the above-mentioned detection device further comprises the step of extraction and enrichment of the cell-free DNA before the high-throughput sequencing. Suitable extraction and enrichment methods can be selected depending on diversity samples and requirement data respectively. For example, QIAmp DNA Blood Mini Kit from Qiagen, Germany, or commercially available similar reagents from other companies, or self-made relevant reagents for extraction and enrichment of peripheral blood of a pregnant woman can be used for the extraction and enrichment.

In the above-mentioned detection device, after the step of performing library construction for the above-mentioned enriched DNA to obtain a sequencing library, different target regions can further be selected for sequencing, depending on detection purposes of different samples. During the actual operation of the present invention, in the above-mentioned detection device, the step of performing target region sequencing on the sequencing library to obtain a sequencing library containing the target regions is also included after obtaining the sequencing library and before performing the high-throughput sequencing. In a more preferred embodiment of the invention, the step of performing exon capture hybridization on the sequencing library is added, so that the subsequent high throughput sequencing is performed only for exons. Since introns are usually cleaved off during the transcription process of a gene, and exons are regions encoding a protein, only performing exon

sequencing can increase an effective data volume and improve efficiency of sequencing. After obtaining a sequencing library for exons, it is also possible to detect mutations of known monogenic diseases within specific target regions depending on detection purposes and/or detection samples, or to detect all mutations in the sequencing data as entirety.

In the above-mentioned preferred embodiment of the invention, different methods or reagents may be selected for capture depending on the target regions to be captured. For example, a capture kit from Roche NimbleGen, US, or a self-made kit or a commercially available kit from other companies with a similar function can be used to perform target region sequencing.

In another typical implementation of the present invention, a kit for genotyping of a pregnant woman and her fetus is provided, the kit comprising: reagents and apparatuses for enriching cell-free DNA from a peripheral blood plasma sample of the pregnant woman and performing high-throughput sequencing; an apparatus for aligning the sequencing data obtained by the high-throughput sequencing with those of a reference genomic sequence to obtain SNP sites; and an apparatus for performing mixed genotyping at each site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes of each SNP site, and taking the mixed genotype with the maximum probability as the target mixed genotype of each of the SNP sites; wherein the mixed genotype refers to pseudo-tetraploid genotypes formed of genotypes of the pregnant woman and the fetus, and is any one of seven types, AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites, the seven types are sequentially numbered as type 1, type 2, type 3,

type 4, type 5, type 6 and type 7.

In the above kit of the present invention, the mixed genotype at each of the SNP sites is deduced using a mixed genotype of pseudo-tetraploid formed of genotypes of the pregnant woman and the fetus and a conditional probability and a Bayesian model, thereby obtaining genotypes of the mother and the fetus, thereby achieving detection of all possible genotypes in the fetus using only a peripheral blood sample of the pregnant woman. The kit of the present invention not only reduces the sequencing of the paternal and/or maternal samples, and reduces the cost of sequencing, but also provides convenience and diversified services for the detection of fetal genotypes under certain special conditions, such as the case when the paternal sample is not available.

In the above-mentioned kit of the present invention, when the initial fetal concentration is the true fetal concentration, those skilled in the art would obtain a mixed genotype of the present invention by modifying the conditional probability and the Bayesian model, based on the pseudo-tetraploid and mixed genotyping of pseudo-tetraploid proposed by the present invention. In a preferred embodiment of the present invention, when the initial fetal concentration is not the true fetal concentration, the apparatus for obtaining the target mixed genotype in the above-mentioned kit comprises: a first calculation element for performing mixed genotyping at each SNP site using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among 7 mixed genotypes of each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites; a selection element for selecting the initial mixed genotype suitable for calculating a second fetal concentration, and recording it the mixed genotype for calculation of the fetal concentration (the second mixed genotype); a second calculation element for calculating a

calculated fetal concentration f' according to the mixed genotype for calculation of the fetal concentration and sequencing data; a comparison element for comparing calculated fetal concentration f' and the initial fetal concentration f to obtain a difference value Δf ; an assessment element for determining whether the Δf is greater than a pre-defined value; an iteration element for repeatedly operating the first calculation element, the selection element, the second calculation element, the comparison element and the assessment element with the f' as f , when the Δf is greater than the pre-defined value; and a labelling element for labelling the initial mixed genotype corresponding to the initial fetal concentration as the target mixed genotype when the Δf is not greater than the pre-defined value.

An apparatus for performing mixed genotyping for each SNP site to obtain a target mixed genotype of each of the SNP sites calculates probabilities of 7 mixed genotypes at each of the SNP sites using a pre-estimated initial fetal concentration to obtain a mixed genotype with the maximum probability at each of the SNP sites; takes the mixed genotype with the maximum probability as an initial mixed genotype; then selects an initial mixed genotype suitable for calculation of the fetal concentration as a second mixed genotype by a selection element; then calculates a second fetal concentration f' according to the second mixed genotype and sequencing data by a second calculation element; then assesses the difference between an initial fetal concentration and a calculated fetal concentration according to a difference value Δf which is obtained by comparing the initial fetal concentration with the calculated fetal concentration by a comparison element and an assessment element; when the difference value Δf is greater than a pre-defined value, records the second fetal concentration f' as an initial fetal concentration f by a iteration element for cyclic execution of the above-mentioned pre-estimation element, the selection

element, the calculation element, the comparison element and the assessment element, until when the difference value Δf is less than the pre-defined value, it is considered that the initial fetal concentration is not significantly different from the calculated fetal concentration, i.e. the initial fetal concentration in this case is the true concentration in the fetus; thereby labelling the mixed genotype with the maximum probability calculated from the initial fetal concentration as the target mixed genotype by a labelling element.

The selection principle of the second mixed genotype by the selected element in the above-mentioned preferred embodiment can be reasonably selected according to the calculation method. In a preferred embodiment of the invention, the above-mentioned second mixed genotype includes, but is not limited to, any one or more of AAAB, ABAA, ABBB, and BBAB. The mutant alleles or reference alleles in these mixed genotypes are only from the pregnant woman or the fetus, and the concentration of one of them can be calculated based on the number occurrence of the mutant alleles and the reference alleles are detected in the sequencing data, so that it is suitable to obtain the concentration in the fetus.

In another preferred embodiment of the present invention, in the above-mentioned apparatus for obtaining a target mixed genotype, the step of performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises: obtaining the following formula (1) based on the sum of the conditional probability of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S

$$\sum P(G_j | S) = 1 \quad (1)$$

represents one of the SNP sites, and $P(G_j|S)$ represents the probability of the

mixed genotype G_j at an SNP under the S condition; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents the probability of occurrence of G_j at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively;

obtaining the following formula (3) from formula (2) by selecting any one mixed genotype G_{j*} from G_j as the reference mixed genotype:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the corresponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype G_j at the i -th SNP site to a probability of the mixed genotype G_{j*} at the i -th SNP site under the S_i condition; $P(G_{ij})$ is calculated from the population mutation frequency, and $P(S_i|G_{ij})$ is obtained by a binomial distribution formula using the number occurrence of the mutant allele at each of the SNP sites, the number occurrence of the reference allele corresponding to the mutant allele, and the initial fetal concentration f ; then by the following formula (5), finding the mixed genotype with maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with maximum occurrence probability as the initial mixed genotype at the i -th SNP site

$$G = \arg \max(\varphi_j) \quad (5).$$

In the above-mentioned preferred embodiment of the present invention, in the kit, the method for calculation of the mixed genotype at a SNP site using the conditional probability and the Bayesian model is converted by the above-mentioned first calculation element into a calculation of a ratio between the probability of the above-mentioned seven mixed genotypes at the SNP site and the probability of one of the mixed genotypes, so that the mixed genotype with the maximum probability at the SNP site is indirectly obtained, and thus is recorded as the mixed genotype at the site.

In the above-mentioned kit of the present invention, in the case that a mutant genotype is known to occur at a site, those skilled in the art can calculate $P(G_{ij})$ in the formula (4) using the population mutation frequency in the thousand human genome database. In a preferred embodiment of the present invention, in the above-mentioned apparatus for deducing a fetal genotype of each of the SNP sites, $P(G_{ij})$ in the formula (4) is calculated by the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6).$$

In the above-mentioned formula (6), G' represents the above-mentioned three separate possible genotypes occurring at a site in the pregnant woman or the fetus, and then an probability of a mixed genotype at the specific site is the product of the probability of genotype G' at the site of the pregnant woman and the probability of genotype G' at the site of the fetus, wherein θ is the population mutation frequency of the i -th SNP site. The parameter θ is obtained from the population mutation frequency in the thousand human genome database.

In the above-mentioned kits of the present invention, in the apparatus for deducing a fetal genotype of each of the SNP sites, the numerical value of $P(S_i|G_{ij})$ in the formula (4) is obtained depending on the difference between the number occurrence of mutant alleles and the number occurrence of reference alleles in actual sequencing data, and the difference in the fetal concentration when a site is a specific mixed genotype G_{ij} , using a binomial distribution formula. In a preferred embodiment of the present invention, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number occurrence the allele at the i -th SNP site, k represents the number occurrence of the reference allele at the i -th SNP site, and $f(b)$ represents the theoretical probability of occurrence of a mutant allele when the mixed genotype of the i -th SNP site is G_{ij} .

In the above-mentioned embodiment, $f(b)$ is related to the concentration of cell-free fetal DNAs in peripheral blood of the pregnant woman, and can be calculated using a conventional fetal concentration calculation method such as Fetal Quant (see Lench N, Barrett A, Fielding S, *et al.* The clinical implementation of non-invasive prenatal diagnosis for single-gene disorders: challenges and progress made [J]. Prenatal diagnosis, 2013, 33(6): 555-562.).

After obtaining the fetal concentration, when a site is one of the mixed genotypes, a possible paternal genotype is derived for the mixed genotype, thereby deducing the theoretical occurrence probability of a mutant allele in the fetus when a specific mixed genotype occurs.

In another preferred embodiment of the present invention, the initial fetal concentration f is calculated by iteration using the expectation maximization algorithm. Assuming that the initial fetal concentration f is 10%, the mixed

genotypes of all SNP sites when $f=10\%$ are calculated; the actual fetal concentration f' (i.e. the second fetal concentration f') is calculated according to frequencies of the reference genotypes and the mutant alleles actually detected for mixed genotypes; and if the difference value between f' and f is less than the pre-defined value, then the iteration ends, and the corresponding f' at the end of the iteration is the fetal concentration f . More preferably, when the above-mentioned pre-defined value is 0.001, the iteration ends. The specific algorithm used is the same as that in the foregoing detection method, and details are not described herein again.

After obtaining the fetal concentration using the above-mentioned method, the theoretical probability $f(b)$ of the occurrence of the mutant allele when a mixed genotype of the i -th SNP site is G_{ij} can be calculated according to the following formulas respectively, depending on the mixed genotype G_{ij} : when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the fetal concentration, and the specific calculation method is the same as shown in Table 1, which is not detailed herein again.

Beneficial effects of the present invention will be further described below in conjunction with examples.

It should be noted that Example 1 is carried out in accordance with the flowchart shown in Figure 2. All reagents used in the following examples are from NEB unless otherwise specified; and the methods used can be carried out by conventional methods in the art unless otherwise specified.

Example 1. A method for detecting gene mutations

Experiment 1: Sample preparation and cell-free DNA extraction

(1) Peripheral blood extracted from a pregnant woman was placed in a centrifuge for centrifugation at a speed of 1600 g for 10 min, and then the plasma was collected.

(2) After obtaining the peripheral blood plasma extracted from the pregnant woman, cell-free DNA in the plasma was extracted using the QIAamp DNA Blood Mini Kit (Qiagen, Germany, catlog # 51106) by following the method written in the user's manual.

Experiment 2: Capture enrichment and library construction

2.1 End-repair of cell-free DNA in plasma of the pregnant woman

Experimental objective: since the cell-free DNA extracted from the plasma of the pregnant woman are double-stranded DNA fragments which are either blunt-ended or contain 3' or 5' overhangs. In this step, the overhangs were phosphorylated to blunt ends by T4 DNA polymerase, a large fragment of *E. coli* DNA polymerase I (Klenow fragment) and polynucleotide kinase T4. The 3' to 5' exonuclease activity of the large fragment of DNA polymerase I removes the 3' overhangs and the T4 DNA polymerase activity fills the 5' overhangs. Eventually the cell-free DNAs have blunt ends.

Experimental materials, reagents and instruments: cell-free DNA of Experiment 1; a mixture of dNTPs (10 mM); T4 DNA polymerase (3 units/ μ L); Klenow fragment (5 units/ μ L); T4 PNK (T4 polynucleotide kinase, 10 units/ μ L) and PNK buffer; magnetic beads for DNA purification; and a PCR instrument.

Experimental procedure:

A. Formulating the following reaction system:

Plasma DNA of a pregnant woman	40 μ L
--------------------------------	------------

A mixture of dNTPs (10 mM)	2 μ L
T4 DNA polymerase	1 μ L
Klenow fragment	1 μ L
T4 PNK (T4 polynucleotide kinase)	1 μ L
PNK buffer	5 μ L
Total volume	160 μ L

B. Incubating in a PCR instrument at 20°C for 30 minutes;

C. Purifying the reaction product with magnetic beads, and eluting with 19.5 μ L elution buffer (EB).

2.2 Adding base “A” at a 3' end of cell-free DNA fragments

Implementation objective: since subsequent adapter sequences contain a single “T” base at the 3' end needs to be ligated with the end-repaired cell-free DNA fragments, it is therefore necessary to first add a complementary “A” bases at the 3' end of the end-repaired fragments. This step is accomplished by using the Klenow fragment absent of 3' to 5' exonuclease activity.

Experimental materials, reagents and instruments: end-repaired cell-free DNA; Klenow buffer (10x); dATP (1 mM); Klenow fragment (absent of 3' to 5' exonuclease activity, 5 U/ μ L); and a PCR instrument.

Experimental procedure:

A. Formulating the following reaction system:

End-repaired cfDNA	19.5 μ L
Klenow buffer (10 $\times$)	2.5 μ L
dATP (1 mM)	2.5 μ L
Klenow fragment	0.5 μ L
Total volume	25 μ L

B. Incubating in a PCR instrument at 37°C for 30 minutes.

2.3 Ligating adapters to both ends of the DNA fragments

Experimental objective: in order to enable the DNA fragment with added “A” to be specifically amplified in the subsequent PCR steps, it is necessary to

use a DNA ligase to ligate specific adapters (i.e. an annealing product of each of adapter sequence 1 and adapter sequence 2) at both ends of the DNA.

Experimental materials, reagents and instruments: DNA with added “A” base; DNA ligase buffer (2×); DNA ligase (1 U/μL); adapter sequence 1 and adapter sequence 2; a PCR instrument; and magnetic beads for DNA purification.

The sequence of the adapter sequence 1 is

SEQ ID NO: 1: 5' P-GATCGGAAGAGCACACGTCT-3';

The sequence of the adaptor sequence 2 is SEQ ID NO: 2:

5' P-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3' (Illumina).

Experimental procedure:

A. Formulating the following reaction system:

DNAs obtained in step 2.2	23 μL
DNA ligase buffer (2×)	25 μL
DNA ligase (1 unit/μL)	1 μL
Adapter (20 pmol/μL)	1 μL
Total volume	50 μL

B. Incubating in a PCR instrument at 20°C for 15 minutes.

C. Purifying the reaction product with magnetic beads, and eluting with 38.2 μL elution buffer (EB).

2.4 PCR amplification of the DNA fragments with adapters ligated to both ends.

Experimental objective: PCR amplification was carried out on the DNA fragment modified by ligating adapters to both ends. On the one hand, the complementary sequence of sequencing primer attached to both ends of the treated cell-free DNA fragments could be filled during the process of PCR, and on the other hand, sufficient amount of the DNA fragments could be obtained to continue the subsequent sequencing steps.

Experimental materials, reagents and instruments: DNA fragments with adapters at each of the two ends; 10×Pfx DNA polymerase amplification buffer; a mixture of dNTPs (10 mM); MgSO₄ (50 mM); PCR primer 1 (10 pmol/μL); PCR primer 2 (10 pmol/μL); and Pfx DNA polymerase (2.5 U/μL).

The sequence of PCR primer 1 is SEQ ID NO: 3:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3';

The sequence of PCR primer 2 is SEQ ID NO: 4:

5'-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'.

Experimental procedure:

A. Formulating the following reaction system:

DNAs obtained in step 2.3	38.2 μL
10×PFX DNA polymerase amplification buffer	5 μL
A mixture of dNTPs (10 mM)	2 μL
MgSO ₄ (50 mM)	2 μL
PCR primer 1 (10 pmol/μL)	1 μL
PCR primer 2 (10 pmol/μL)	1 μL
PFX DNA polymerase	0.8 μL
Total volume	50 μL

B. Performing amplification according to the following PCR procedures:

step one: incubating at 94°C for 2 minutes; step two: denaturation at 94°C for 15 seconds; annealing at 62°C for 30 seconds; extension at 72°C for 30 seconds, and repeat step two for 15 cycles; step three: incubating at 72°C for 10 minutes; and step four, finishing the reaction, and preserving at 4°C.

C. Purifying the reaction product with magnetic beads and eluting with ddH₂O.

D. Completing library preparation, and measuring the concentration by Agilent's DNA detector 2100 and the concentration was measured as 21.34 ng/ μ L.

Experiment 3: Target region capture

3.1 Library hybridization

After the library was quantified, exon capture hybridization was performed using the capture kit SeqCap EZ Human Exome+UTR Kit (Cat# 06740308001) from Roche NimbleGen, USA.

Experimental materials, reagents: DNA library; SeqCap.EZ Exome+UTR.Library; Cot DNA; SeqCap EZ Hyb and Wash Kit; HE oligo sequence and TS-INV-HE index oligo sequence;

wherein, the HE oligo sequence is SEQ ID NO: 5:

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'.

The TS-INV-HE index oligo sequence is SEQ ID NO: 6:

5'-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'.

Experimental procedure:

A. Formulating the following reaction system:

Reagent	Amount
DNA library obtained in step 2.4	1 μ g
Cot DNA	5 μ g
HE oligo	1000 pmol (1 μ L of 1000 μ M)
TS-INV-HE index oligo	1000 pmol (1 μ L of 1000 μ M)

B. Drying at 56°C with a vacuum concentrator after finishing the above procedure.

After evaporating samples to dry, adding 7.5 μ L 2 $\times$ Hybridization Buffer

and 3 μ L Hybridization Component A, mixing uniformly, and performing denaturation at 95°C for 10 minutes.

C. After finishing the denaturation, transferring the above-mentioned mixture to 0.2 mL PCR tube with a convex cap, and adding 4.5 μ L SeqCap EZ Exome+UTR Library. Vortexing for 3 seconds, mixing thoroughly, and centrifuging at the maximum speed for 10 seconds.

D. Placing a sample mixture to be hybridized at 47°C for heating for 64 to 72 hours for hybridization.

3.2 Elution and recovery of captured DNA

Experimental reagents: streptavidin Dynabeads; 1 $\times$ Stringent Wash Buffer; 1 $\times$ Wash Buffer I; 1 $\times$ Wash Buffer II; 1 $\times$ Wash Buffer III and ddH₂O.

Experimental procedure:

A. Transferring the sample library to a 0.2 mL PCR tube containing streptavidin Dynabeads, pipetting up and down 10 times and mixing well; then placing the 0.2 mL PCR tube in a heating module and incubate at 47°C for 45 minutes. During the incubation, vortexing the tube every 15 minutes to mix the reagents, so that the DNA fragments could bind to the magnetic beads.

B. After incubating for 45 min, adding 100 μ L 1 $\times$ Wash Buffer I at 47°C to a 15 μ L captured DNA sample. Vortexing and mixing uniformly for 10 sec. Transferring all components in the 0.2 mL PCR tube to a 1.5 mL centrifuge tube. Placing the centrifuge tube on a magnetic stand to gather the magnetic beads, and discarding the supernatant.

Then removing the 1.5 mL centrifuge tube from the magnetic stand, and adding 200 μ L 1 $\times$ Stringent Wash Buffer preheated at 47°C. Pipetting up and down 10 times to mix uniformly. After mixing, placing the sample on a heating module at 47°C for 5 minutes, and washing twice with a 1 $\times$ Stringent Wash

Buffer at 47°C. Placing the 1.5 mL centrifuge tube on a magnetic stand again, and discarding a supernatant.

Adding 200 µL 1×Wash Buffer I at room temperature to the above-mentioned 1.5 mL centrifuge tube, and vortexing and mixing uniformly for 2 min. Placing the centrifuge tube on a magnetic stand and discarding the supernatant.

Adding 200 µL 1×Wash Buffer II at room temperature to the above-mentioned 1.5 mL centrifuge tube, and vortexing and mixing uniformly for 1 min. Placing the centrifuge tube on a magnetic stand and discarding a supernatant.

Adding 200 µL 1×Wash Buffer III at room temperature to the above-mentioned 1.5 mL centrifuge tube, and vortexing and mixing uniformly for 30 sec. Placing the centrifuge tube on a magnetic stand and discarding a supernatant.

Removing the 1.5 mL centrifuge tube from the magnetic stand, adding 50 µL ddH₂O to elute the magnetic bead-captured sample. Storing the magnetic bead-sample mixture at -20°C.

3.3 PCR amplification of captured DNA

Experimental objective: due to the very low concentration of DNA samples captured, PCR amplification is needed to meet the requirements of subsequent experiments.

Experimental materials and reagents: captured DNA by hybridization; 10×Pfx DNA polymerase amplification buffer; a dNTP mixture (10 mM); MgSO₄ (50 mM); PCR primer 3 (10 pmol/µL) (Invitrogen); PCR primer 4 (10 pmol/µL); and Pfx DNA polymerase (2.5 U/µL).

The sequence of the PCR primer 3 is SEQ ID NO: 7:

5'-AATGATACGGCGACCAACCGAGATCTACACTCTTTCCCTACACG

ACGCTCTTCCGATCT-3';

The sequence of the PCR primer 4 is SEQ ID NO: 8:

5'-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTT
CAGACGTGTGCTCTTCCGATCT-3'.

Experimental procedure:

A. Formulating the following reaction system:

Captured DNA	38.2 μ L
10 $\times$ PFX DNA polymerase	5 μ L
amplification buffer	
A mixture of dNTPs (10 mM)	2 μ L
MgSO ₄ (50 mM)	2 μ L
PCR primer 1 (10 pmol/ μ L)	1 μ L
PCR primer 2 (10 pmol/ μ L)	1 μ L
PFX DNA polymerase	0.8 μ L
Total volume	50 μ L

B. Performing amplification according to the following PCR procedures:

step one: incubating at 94°C for 2 minutes; step two: denaturation at 94°C for 15 seconds; annealing at 62°C for 30 seconds; extension at 72°C for 30 seconds, repeat step two for 13 cycles; step three: incubating at 72°C for 10 minutes; and step four, finishing the reaction, and preserving at 4°C.

C. Purifying the reaction product with magnetic beads and eluting with 30 μ L elution buffer ddH₂O.

D. After preparing the library, and measuring the concentration using Agilent's DNA detector 2100 and the concentration was measured as 13.60 ng/ μ L.

Experiment 4: Loading and sequencing

The DNA molecules in the sequencing library were made into DNA clusters using cBot instrument from Illumina, and the resulting DNA clusters

were subjected to 100 cycles of double-end sequencing on an Illumina HiSeq 2000 (or Illumina HiSeq 2500) sequencer.

The raw image data files obtained by high-throughput sequencing (Illumina HiSeqTM2000) were converted by the CASAVA Base Calling raw sequenced sequence (Sequenced Reads) which were also known as Raw Data or Raw Reads. The results were saved as a FASTQ (abbreviated as fq) file format containing the sequence information of the sequenced sequences (reads) and the corresponding sequencing quality information.

Experiment 5: Processing of raw sequencing data

5.1 Filtering out unqualified reads

Raw reads obtained by sequencing contain reads with adapters and reads low sequencing quality (over 50% nucleotide bases have a sequencing quality score of $Q \leq 5$ in a read). In order to ensure the quality of the analysis result, the raw reads should be filtered to obtain reads with qualified sequencing quality and with adapters removed (also known as clean reads), and subsequent analysis is based on the filtered reads. The following sequences are filtered out: (1) reads containing N at a ratio of greater than 5%; (2) low-quality reads (nucleotide bases with a quality value of $Q \leq 5$ accounts for 50% or more of the entire read length); and (3) reads contaminated with adapters.

The raw data statistics for the samples are shown in Table 2. The modified Q30 bases rate (%) indicates the proportion of bases with a quality value greater than 30 (an error rate of less than 0.1%) in the total sequence after filtration. The larger the value, the better the sequencing quality. Generally, if the index is greater than 85%, the sequencing quality is considered qualified. If it is less than 85%, then re-sequencing is required.

Table 2:

Raw reads	Clean reads	Effect rate (%)	Q30 (%)
-----------	-------------	-----------------	---------

140,070,440	136,958,780	97.78	86.76
-------------	-------------	-------	-------

Effect rate (%): The percentage of reads obtained by dividing the clean reads to the raw reads. The clean reads were obtained by removing the following ones in the raw reads: 1. Low-quality reads, which nucleotide bases with a quality value of $Q \leq 5$ accounts for 50% or more of the entire read length; 2. reads containing N at a ratio of greater than 5%; and 3. reads with adapter contamination.

5.2 Mapping quality control

By mapping filtered clean reads to a reference genome (HG19, NCBI built 37) using the mapping software BWA (bwa-0.7.5a), the mapping results are subjected to mapping quality control to obtain a mapping file.

The quality control points comprise the data mapping rate, capture specificity, target region sequencing depth, target region sequencing depth distribution, PCR duplication rate and the like. The results of the mapping quality control are shown in Table 3.

Table 3:

Duplication rate (%)	Mapping rate (%)	Capture specificity	Target region sequencing depth (X) (Target Average depth)
5.54	97.44	72.38	139.78

In the above-mentioned Table 3, the capture specificity means that reads are completely mapped to a target region, and reads are partially in the target region and are partially outside of the region; the Target Average depth (X) refers to the depth of a target region; the duplication rate (%) involves reads that are duplicated due to PCR amplification; the mapping rate (%) refers to a ratio of the reads mapped to hg19 reference genome in the raw data using BWA, and generally 90% or more can be considered as normal results.

Then, based on the above-mentioned mapping file, useless data (such as duplication reads, etc.) are removed, and a set of sites with the nucleotide

sequence different from the reference genome are obtained;

Finally, statistic analysis of the sequencing results of the above-mentioned differential base sites are performed, for example, 40 As and 60 Ts being detected at a site, and some other information such as the location of a site.

Experiment 6: Mixed genotyping of pseudo-tetraploid

Target areas of the pregnant woman and fetus is directly genotyped only by genetic information derived from peripheral blood of the pregnant woman as mentioned above. According to the pseudo-tetraploid genotyping model of the present invention, a mixed genotype of a pseudo-tetraploid composed of a genotype of the pregnant woman and a genotype of the fetus is deduced to obtain genotypes of the pregnant woman and the fetus at each corresponding site. The following content takes the specific situation of a site as an example to illustrate the process of deducing the genotypes of the pregnant woman and the fetus.

If at the site, the reference allele is detected 50 times, the mutant allele is detected 8 times, the fetal concentration is 8%, the population mutation frequency is 0.03, and G_{j*} in the formula (4) is G_4 , then the formula (4') is obtained

$$\varphi_j = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{i4})P(G_{i4})} \quad (4')$$

In combination with the formula (4') to the formula (7) and Table 1, a probability ratio φ of the 7 mixed genotypes and the fourth mixed genotype is calculated, thereby obtaining a corresponding mixed genotype with the highest probability ratio φ . The specific calculation process is as follows:

when the mixed genotype is AAAA,

$$\varphi_1 = \frac{P(S_i | G_{i1})P(G_{i1})}{P(S_i | G_{i4})P(G_{i4})} = \frac{\binom{50+8-1}{8-1} 0^8 (1-0)^{50} \times (1-0.03)^2 (1-0.03)^2}{\binom{50+8-1}{8-1} 0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

when the mixed genotype is AAAB,

$$\varphi_2 = \frac{P(S_i | G_{i2})P(G_{i2})}{P(S_i | G_{i4})P(G_{i4})} = \frac{0.04^8 (1-0.04)^{50} \times (1-0.03)^2 \times 2 \times 0.03 \times (1-0.03)}{0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

when the mixed genotype is ABAA,

$$\varphi_3 = \frac{P(S_i | G_{i3})P(G_{i3})}{P(S_i | G_{i4})P(G_{i4})} = \frac{0.46^8 (1-0.46)^{50} \times (1-0.03)^2 \times 2 \times 0.03 \times (1-0.03)}{0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

when the mixed genotype is ABAB,

$$\varphi_4 = \frac{P(S_i | G_{i4})P(G_{i4})}{P(S_i | G_{i4})P(G_{i4})} = 1 ;$$

when the mixed genotype is ABBB,

$$\varphi_5 = \frac{P(S_i | G_{i5})P(G_{i5})}{P(S_i | G_{i4})P(G_{i4})} = \frac{0.54^8 (1-0.54)^{50} \times (0.03)^2 \times 2 \times 0.03 \times (1-0.03)}{0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

when the mixed genotype is BBAB,

$$\varphi_6 = \frac{P(S_i | G_{i6})P(G_{i6})}{P(S_i | G_{i4})P(G_{i4})} = \frac{0.96^8 (1-0.96)^{50} \times (0.03)^2 \times 2 \times 0.03 \times (1-0.03)}{0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

and when the mixed genotype is BBBB,

$$\varphi_7 = \frac{P(S_i | G_{i7})P(G_{i7})}{P(S_i | G_{i4})P(G_{i4})} = \frac{1^8 (1-1)^{50} \times (0.03)^2 (0.03)^2}{0.5^8 (1-0.5)^{50} \times (2 \times 0.03 \times (1-0.03))^2} ;$$

All of the φ values are compared, obtaining $\varphi_2 > \varphi_3 > \varphi_4 > \varphi_5 > \varphi_6 > \varphi_7 = \varphi_1$, and therefore the most likely mixed genotype of this site is type AAAB. Further, it is deduced that the genotype of the pregnant woman is AA, and the genotype of the fetus is AB. According to the above-mentioned principle, a mixed genotype of each of all the variant sites (SNP sites) in the mapping result file is obtained, thereby the genotype of the fetus is obtained.

Experiment 7: Determination of pathogenic mutation sites

7.1 Filtering out non-pathogenic mutation sites

A fetal genotype at each variant site obtained from the results of pseudo-tetraploid typing is compared with the following databases, respectively, and the variant sites fulfilling the following criteria in the databases are filtered out: (1) high-frequency mutations in the dbSNP135 public database; and (2) polymorphic sites in the Freq 1000g2012feb (thousand human genome) database. (3) The mutation sites of synonymous mutations, nonsense mutations and non-conserved regions are filtered out according to the mutation prediction software. The sites that appeared in all of the above-mentioned three screening conditions are excluded to obtain fetal specific variant sites, as shown in Table 4.

Table 4:

Total number of SNPs	Non-synonymous mutation site	ASN_freq<0.05	dbSNP	Differential base site
111407	100046	8622	100501	1049

7.2 Based on the published literatures and clinical data, pathogenic mutations are determined from the above-mentioned filtered mutation sites.

Example 2. An example of non-invasive single-gene defect diagnosis of an osteogenesis imperfecta for a fetus

Sample information: a pregnant woman, 28 years old, gravida 3 para 0, with regular menstrual of 5-6 days, and a menstrual cycle of 29 days. The last menstrual period was March 27, 2012, and the expected date of birth was January 4, 2013. She conceived naturally, had no history of fever, rash or the like during early pregnancy, and had no history of exposure to radiation or poisons. At gestational week 13, *Toxoplasma gondii*, rubella, cytomegalovirus, and herpes simplex virus were all test negative; at gestational week 14+, width

of nuchal translucency (NT) of the fetus detected by B-mode ultrasound is 0.14 cm; and at gestational week 17+, 21-trisomy risk probability was $< 1:50000$ and 18-trisomy risk probability was $< 1:50000$ indicated by the serological screening. At gestational week 26+, B-mode ultrasound in a healthcare hospital suggested dysplasia of fetal femurs and tibias; and at gestational week 26+, reexamination through B-mode ultrasound suggested that fetal skull was thin with reduced echo, the length of bilateral femurs was 3.3 cm and bent into an angle, and the tibias and fibulas were also bent into an angle. It was considered that the fetal femurs, and the tibias and fibulas formed angles. At gestational week 30+, the pregnant woman carried out genetic mutation analysis using the method of the present invention, and the fetus was diagnosed to have osteogenesis imperfecta and the pregnant woman was recommended to terminate the pregnancy.

10 ml of peripheral blood was drawn from the pregnant woman, and according to the method of Example 1, cell-free DNA was extracted, captured and enriched, and the enriched DNA was subjected to high-depth sequencing through the HiSeq platform.

Pseudo-tetraploid typing: the sequencing result was subjected to quality control, low-quality data were filtered out, the remaining data were mapped to the genome, and according to the mapping result, the fetal genotyping information was deduced through the pseudo-tetraploid typing model of the present invention, and screened for whether it is related to osteogenesis.

Mutation site screening: 111407 raw mutations were filtered according to the steps in Example 1 for mutation sites, and finally 7 mutations were obtained. Literature review and clinical data review were performed on the screened 7 mutations, and one mutation was finally determined (COL1A1:NM_000088:c.G2596A:p.G866S) as a pathogenic mutation leading

to osteogenesis imperfecta.

Verification of results: the fetal umbilical cord blood sample and peripheral blood samples of the pregnant woman and the fetal father were used to verify the pathogenic mutation obtained for the sample of the present example, and the results are shown in Figure 3 (in Figure 3, MF, FF or C-F represents the sense strand of gene of the mother, the father, and the fetus respectively). As shown in Figure 2, the fetus truly contains the pathogenic mutation at this site.

Example 3. Verification of the pseudo-tetraploid typing model of the present invention.

The above-mentioned peripherals blood sample of the pregnant woman is detected and analyzed by the pseudo-tetraploid typing model of the present invention, meanwhile, the somatic cell detection method of “a fetal cord blood sample + a peripheral blood sample of the pregnant woman” is used to verify and evaluate the validity of the pseudo-tetraploid typing model of the present invention.

In combination with the somatic cell sequencing result of the pregnant woman and the somatic cell sequencing result of the cord blood sample, comparing with the result of pseudo-tetraploid genotyping was compared and the accuracy of the method of the present invention are shown in Table 5 below; and the detection rate obtained by comparing with the somatic cell sequencing result of a pregnant woman and the somatic cell sequencing result of a cord blood sample in the prior art is shown in Table 6.

Table 5: Mixed genotyping of pseudo-tetraploid and detection rates

Mixed genotype	Number of matched sites	Number of unmatched sites ^A	Total number of sites ^A	Accuracy /%
AAAA	89078219	2056	89080275	99. 99
AAAB	10157	5654	15811	64. 24

ABAA	9968	6051	16019	62. 23
ABAB	16203	11711	27914	58. 05
ABBB	4936	4077	9013	54. 77
BBAB	7274	1646	8920	81. 54
BBBB	28284	821	29105	97. 18
Total				
(without AAAA)	76822	29960	106782	71. 94

Notes: since AAAA does not contain mutant allele B, the number of sites for this genotype is not counted in the total number of sites. The total number of sites in the above-mentioned table is the sum of the number of mapped base sites in the mother + the number of mapped base sites in the fetus by somatic cell sequencing for the mother and cord blood respectively (each of the repeatitive sites in mother and her fetus was only counted once for calculation).

The total number of sites ^A represents the total number of positive sites detected using the pseudo-tetraploid genotyping model of the present invention.

Number of matched sites is the number of true positive sites, representing the number of sites in the above-mentioned total number of sites ^A which are determined by the pseudo-tetraploid typing model of the present invention with true mutations. (The somatic cell sequencing performed using mother's blood and cord blood was consider as the gold standard for maternal and fetal genotype detection, and the method of determining maternal and fetal genotypes by the pseudo-tetraploid model is a subject method, and by comparing the subject method of the present invention with the gold standard, the site with consistent result is counted as a true negative or true positive site, and the inconsistent site is recorded as a false positive or false negative site.)

Number of unmatched sites is the number of false negative sites, representing the number of sites that are not determined by the pseudo-tetraploid typing of the present invention in the total number of sites ^A.

Positive detection accuracy is measured as an index for evaluating the subject detection method, pseudo-tetraploid in this case. It is calculated as true positive/(true positive + false negative) $\times$ 100%, i.e. a ratio of the number of true positive sites detected by pseudo-tetraploid method to the total number of sites ^A.

Table 6: Somatic cell detection results and the detection rate

Mixed genotype	Number of matched sites	Number of unmatched sites ^B	Total number of sites ^B	Detection rate/%
AAAA	89078219	10897	89089116	99. 99
AAAB	10157	1808	11965	84. 89
ABAA	9968	5820	15788	63. 14
ABAB	16203	8872	25075	64. 62
ABBB	4936	2897	7833	63. 02
BBAB	7274	575	7849	92. 67
BBBB	28284	1384	29668	95. 34
Total				
(without AAAA)	76822	21356	98178	78. 25

The detection accuracy is the ratio of the number of positive sites detected by the subject method, pseudo-tetraploid in this case, to the number of true positive sites detected by the gold standard. In Table 6 above, the total number of sites ^B is the number of sites that have true positive mutations as determined by a somatic cell sequencing result of a pregnant woman and a somatic cell sequencing result of a cord blood sample. Number of matched sites is the number of true positive sites detected using the pseudo-tetraploid typing model. Number of unmatched sites is the number of true positive sites (i.e. the number

of false negative sites) that were undetected by the pseudo-tetraploid typing model. Thus, the true positive/(false negative + true positive) $\times$ 100% in Table 6 represents the detection rate.

Example 4

A device for detecting gene mutations comprises:

a detection module for performing high-throughput sequencing of cell-free DNAs in peripheral blood of a pregnant woman to obtain sequencing data. It comprises instruments for sequencing the cell-free DNAs in the peripheral blood of the pregnant woman which include cBot instrument from Illumina and Genome AnalXzer from Illumina, HiSeq2000 sequencer or HiSeq2500 sequencer or SOLiD series of sequencers from ABI.

Preferably, the detection module further comprising a region capture sub-module for performing target region capture on the sequencing library constructed from the enriched cell-free DNAs to obtain a sequencing library for high-throughput sequencing.

An alignment module for aligning the sequencing data with a reference genomic sequence to obtain SNP sites;

a target mixed genotype determination module for performing mixed genotyping for each of the SNP sites using a Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; wherein the mixed genotype refers to pseudo-tetraploid genotypes formed by genotypes of the pregnant woman and the fetus, the mixed genotype is any one of seven types consisting of AAAA, AAAB, ABAA, ABAB, ABAB, ABAB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABAB, ABAB, BBAB, and BBBB, and are sequentially numbered as type 1, type 2,

type 3, type 4, type 5, type 6 and type 7, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites; a target mixed genotype of fetus at each of the SNP sites is deduced by performing mixed genotyping according to conditional probability and the Bayesian model.

Preferably, when the initial fetal concentration is not the true fetal concentration, the target mixed genotype determination module comprises: a pre-estimation module for calculating probabilities of 7 mixed genotypes of each of the SNP sites respectively with the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability for each of the SNP sites, and taking the mixed genotype with the maximum probability as an initial mixed genotype; a selection module for selecting the initial mixed genotype, if suitable, as a second mixed genotype to calculate the second fetal concentration, which is recorded as a second mixed genotype; a calculation module for calculating a second fetal concentration f' according to the second mixed genotype and the sequencing data; a comparison module for comparing the second fetal concentration f' with the initial fetal concentration f to obtain a difference Δf ; a determination module for determining whether the Δf is greater than a pre-defined value; an iteration module for repeatedly executing the pre-estimation module, the selection module, the calculation module, the comparison module and the determination module with the f' as f , when the Δf is greater than the pre-defined value; and a labelling module for labelling the initial mixed genotype of each of the SNP sites as the target mixed genotype when the Δf is not greater than the pre-defined value.

Preferably, the initial fetal concentration is 10%; more preferably, the pre-defined value is 0.001; and further preferably, the mixed genotype for calculation of the fetal concentration is selected from any one or more of the

following four mixed genotypes: AAAB, ABAA, ABBA, and BBAB.

Preferably, the step block of performing mixed genotyping for each of the SNP sites by the target mixed genotype determination module using the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises: obtaining the following formula (1) on the basis that the conditional probability sum of the seven mixed genotypes is 1,

$$\sum P(G_j | S) = 1 \quad (1)$$

wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, and $P(G_j | S)$ represents a probability of the mixed genotype of an SNP site being G_j when the SNP site is S ; obtaining the following formula (2) from the Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents a probability of occurrence of G_j genotype at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively; obtaining the following formula (3) from formula (2) by selecting any one mixed genotype G_{j*} from G_j as a reference:

$$P(G_{ij*} | S_i) = \frac{P(S_i | G_{ij*})P(G_{ij*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the cooreponding side of formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij*})P(G_{ij*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype

of the i -th SNP site G_j to the probability of the reference mixed genotype of the i -th SNP site G_{j*} under the S_i condition; $P(G_{ij})$ is calculated from a population mutation frequency, and $P(S_i|G_{ij})$ is obtained by the binomial distribution using the number occurrence of a mutant allele at each of the SNP sites, the number of occurrence of a reference allele corresponding to the mutant allele, and the initial fetal concentration; then by the following formula (5)

$$G = \arg \max (\varphi_j) \quad (5)$$

finding a mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with the maximum occurrence probability as the mixed genotype with the maximum probability at the i -th SNP site.

Preferably, $P(G_{ij})$ in the above-mentioned formula (4) is obtained by multiplying a probability of genotype G' of the pregnant woman by a probability of genotype G' of the fetus in the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is a population mutation frequency of the i -th SNP site.

Preferably, in the fetal genotype determination module, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number occurrence of a mutant allele at the i -th SNP site, k represents the number occurrence of a reference allele at the i -th SNP site, and $f(b)$ represents a theoretical occurrence probability of a mutant allele in the fetus when a mixed genotype of the i -th SNP site is G_{ij} .

Preferably, the theoretical occurrence probability $f(b)$ of a mutant allele in

the fetus when a mixed genotype of the i -th SNP site is G_{ij} is calculated as follows respectively, depending on the mixed genotype G_{ij} : when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the fetal concentration, the fetal concentration is calculated by iteration using the expectation maximization algorithm. Assuming that the initial f is 10%, the mixed genotypes of all SNP sites when $f = 10\%$ are calculated; the fetal concentration f' is calculated according to actual frequencies of the reference alleles and the mutant alleles detected for mixed genotypes; and if the difference between f' and f is less than the pre-defined value, then the iteration ends, and the corresponding f' at the end of the iteration is the fetal concentration f . More preferably, when the fetal concentration is iteratively calculated using the expectation maximization algorithm, and the pre-defined value is ≤ 0.001 , the iteration ends.

A mutation site screening module is used to screen a mutation site from various SNP sites according to the genotype of each of the SNP sites of the fetus.

Preferably, the mutation site screening module comprises: a high-frequency polymorphic site filtration sub-module for filtering out polymorphic sites with a high occurrence frequency in the human population in each of the SNP sites of a fetal genotype to obtain preliminary candidate mutation sites; for example, the high-frequency polymorphic sites in the human population are removed using the dnSNP135 public database and the

Freq_1000g2012feb (thousand human genome) database which have been currently collated by the medical community, and the specific SNP sites of the fetus are obtained, and then the sites which could actually cause gene mutations are screened by the gene mutation site screening sub-module.

A gene mutation site screening sub-module is used for filtering out SNP sites, which result in synonymous mutations and nonsense mutations and occur in a non-conserved region, from the preliminary candidate mutation sites to obtain candidate mutation sites. The module can use a mutation prediction module commonly used in the art for performing harmful mutation screening. For example, ANNOVAR module can screen whether the mutation causes an amino acid change, that is, whether the mutation is nonsynonymous, and can also screen whether the mutation occurs in a conserved sequence region.

A Literature and clinical data screening sub-module is used for performing screening on the candidate mutation sites to obtain the pathogenic mutation site that has been recorded in the literature and clinical data. SNP sites which have been screened by the mutation prediction module and pathogenic sites which are retrieved from existing databases and literatures are aligned, so as to find the site information associated with a monogenic disease and perform corresponding interpretation. The presence or absence of a hot spot mutation site leading to a known monogenic disease can be detected, and non-hot spot mutation sites of a known monogenic disease and an unreported potential pathogenic gene and its mutation sites can also be detected.

Example 5

A kit for genotyping of pregnant women and fetuses comprises:

reagents and apparatuses for the enrichment of cell-free DNA from peripheral blood plasma of the pregnant woman and high-throughput sequencing, wherein the detection reagents can comprise various reagents or

chemicals used in steps such as cell-free DNA extraction, separation, enrichment, detection, and library construction, and the detection apparatuses can comprise 1.5 ml EP tubes, PCR tubes, pipettes, 96-well plates, high-throughput sequencers and the like;

an apparatus for aligning the sequencing data produced by high-throughput sequencing with the reference genome to obtain SNP sites, wherein the apparatus comprises various hardware modules, which are stored with specific storage media, for performing the above-mentioned alignment function using a computer terminal or a mobile terminal; and

an apparatus for performing mixed genotyping for each of the SNP sites using the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites, and taking the mixed genotype with the maximum probability as a target mixed genotype of each of the SNP sites; wherein pseudo-tetraploid refers to pseudo-tetraploid genotypes composed of genotypes of the pregnant woman and the fetus, the genotype of the pseudo-tetraploid is recorded as a mixed genotype which is any one of seven types consisting of AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB, and the AAAA, AAAB, ABAA, ABAB, ABBB, BBAB, and BBBB are sequentially numbered as type 1, type 2, type 3, type 4, type 5, type 6 and type 7, where A represents a reference allele of each of the SNP sites, and B represents a mutant allele of each of the SNP sites.

When the initial fetal concentration is not the true fetal concentration, the apparatus for obtaining the target mixed genotype of each of the SNP sites comprises: a first calculation element for calculating probabilities of 7 mixed genotypes of each of the SNP sites respectively with the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum

probability, and taking the mixed genotype with the maximum probability as an initial mixed genotype of each of the SNP sites; a selection element for selecting the initial mixed genotype as a mixed genotype, if suitable, to calculate a second fetal concentration, and recording it as a mixed genotype for calculation of the fetal concentration; a second calculation element for calculating a second fetal concentration f' according to the mixed genotype for calculation of the fetal concentration and sequencing data; a comparison element for comparing the second fetal concentration f' and the initial fetal concentration f to obtain a difference Δf ; a determination element for determining whether the Δf is greater than a pre-defined value; a circulation element for repeatedly operating the first calculation element, the selection element, the second calculation element, the comparison element and the determination element with the f' as f when the Δf is greater than the pre-defined value; and a labelling element for labelling the initial mixed genotype corresponding to the initial fetal concentration f as the target mixed genotype when the Δf is not greater than the pre-defined value.

Preferably, in the apparatus for obtaining the target mixed genotype, the step of performing mixed genotyping for each of the SNP sites using the Bayesian model and an initial fetal concentration f to obtain a mixed genotype with the maximum probability among seven mixed genotypes for each of the SNP sites comprises: obtaining the following formula (1) on the basis that a conditional probability sum of the seven mixed genotypes is 1, wherein G_j represents any one of the seven mixed genotypes, S represents one of the SNP sites, and

$$\sum P(G_j | S) = 1 \quad (1)$$

$P(G_j | S)$ represents a probability of the mixed genotype of the SNP site being G_j when an SNP site is the S ; obtaining the following formula (2) from the

Bayesian model

$$P(G_{ij} | S_i) = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i)} \quad (2)$$

wherein in the formula (2), $P(G_{ij})$ represents a probability of occurrence of G_j genotype at the i -th SNP site, and j value corresponds to the sequentially numbered mixed genotype, which are 1, 2, 3, 4, 5, 6 or 7 respectively; obtaining the following formula (3) from formula (2) by selecting any one mixed genotype G_{j^*} from G_j as a reference:

$$P(G_{ij^*} | S_i) = \frac{P(S_i | G_{ij^*})P(G_{ij^*})}{P(S_i)} \quad (3)$$

dividing each side of the formula (2) with the cooreponding side of the formula (3) to obtain the following formula (4)

$$\varphi_j = \frac{P(G_{ij} | S_i)}{P(G_{ij^*} | S_i)} = \frac{P(S_i | G_{ij})P(G_{ij})}{P(S_i | G_{ij^*})P(G_{ij^*})} \quad (4)$$

wherein, φ_j represents the ratio of the probability of the mixed genotype of the i -th SNP site being G_j to a probability of the mixed genotype of the i -th SNP site being G_{j^*} under the S_i condition; $P(G_{ij})$ is calculated from a population mutation frequency, and $P(S_i|G_{ij})$ is obtained by the binomial distribution formula using the number occurrence of a mutant allele at each of the SNP sites, the number occurrence of a reference allele corresponding to the mutant allele, and the initial fetal concentration; then by the following formula (5)

$$G = \arg \max(\varphi_j) \quad (5)$$

finding a mixed genotype with the maximum occurrence probability among the seven mixed genotypes, and recording the mixed genotype with the maximum occurrence probability as the mixed genotype with the maximum probability at

the i -th SNP site.

Preferably, $P(G_{ij})$ in the formula (4) is obtained by multiplying a probability of genotype G' of the pregnant woman by a probability of genotype G' of the fetus in the following formula (6)

$$\begin{cases} P(G' = AA) = (1 - \theta)^2 \\ P(G' = AB) = 2\theta(1 - \theta) \\ P(G' = BB) = \theta^2 \end{cases} \quad (6)$$

wherein θ is a population mutation frequency of the i -th SNP site.

Preferably, $P(S_i|G_{ij})$ in the formula (4) is calculated by the following formula (7):

$$P(S_i | G_{ij}) = \binom{k+r-1}{r-1} \times f(b)^r \times (1-f(b))^k \quad (7)$$

wherein r represents the number occurrence of a mutant allele at the i -th SNP site, k represents the number occurrence of a reference allele at the i -th SNP site, and $f(b)$ represents a theoretical occurrence probability of a mutant allele in the fetus when a mixed genotype of the i -th SNP site is G_{ij} .

Preferably, the theoretical occurrence probability $f(b)$ of a mutant allele in the fetus when a mixed genotype of the i -th SNP site is G_{ij} is calculated as follows respectively, depending on the mixed genotype G_{ij} : when the mixed genotype G_{ij} is G_{i1} , the value of the $f(b)$ is 0; when the mixed genotype G_{ij} is G_{i2} , the value of the $f(b)$ is $f/2$; when the mixed genotype G_{ij} is G_{i3} , the value of the $f(b)$ is $0.5-f/2$; when the mixed genotype G_{ij} is G_{i4} , the value of the $f(b)$ is 0.5 ; when the mixed genotype G_{ij} is G_{i5} , the value of the $f(b)$ is $0.5+f/2$; when the mixed genotype G_{ij} is G_{i6} , the value of the $f(b)$ is $1-f/2$; and when the mixed genotype G_{ij} is G_{i7} , the value of the $f(b)$ is 1; wherein the f represents the fetal concentration, the fetal concentration is calculated by iteration using the expectation maximization algorithm. Assuming that the initial f is 10%, the

mixed genotypes of all SNP sites when $f = 10\%$ are calculated; the fetal concentration f' is calculated according to the actual frequencies of the reference allele and the mutant alleles actually detected for mixed genotypes; and if the difference between f' and f is less than the pre-defined value, then the iteration ends, and the corresponding f' at the end of the iteration is the fetal concentration f . More preferably, when the fetal concentration is iteratively calculated using the expectation maximization algorithm, and the pre-defined value is ≤ 0.001 , the iteration ends.

The above-mentioned apparatus for deducing the fetal genotype of each of the SNP sites by mixed genotyping using pseudo-tetraploid comprises various hardware modules, which are stored with specific storage media, for performing the above-mentioned calculation, determination or confirming function using a computer terminal or a mobile terminal. The above-mentioned various calculation means, as parts of the apparatus, can separately perform or can be assembled into an apparatus to perform the above-mentioned calculation function, and thus components that load or store the above-mentioned calculation means are also constituents of the apparatus.

From the above description, it can be seen that the method, device and kit for the non-invasive prenatal gene mutation diagnosis of the present invention can infer the fetal genotype and determine whether the genotype would cause a corresponding disease, only by the genetic information of cell-free DNA in the peripheral blood of the pregnant woman. The genetic information of the fetus's father or mother is not required. It simplifies the process of non-invasive single-gene defect detection and reduces the cost of the test. In addition, the present invention can not only detect a specific single-gene defect, but also can detect multiple single-gene defects simultaneously.

It will be apparent to those skilled in the art that some of the modules,

elements or steps of the present application described above may be implemented by a general-purpose computing device, and they may be integrated on a single computing device or distributed across a net composed of multiple computing devices. Alternatively, they may be implemented by program codes executable by the computing device, and accordingly they may be stored in a storage device for execution by the computing device; or they may be separately fabricated into individual integrated circuit modules, or multiple modules or steps may be implemented by fabricating them as a single integrated circuit module. As such, the present application is not limited to a combination of any particular hardware and software.

Only the preferred embodiments of the present invention are described above, and are not intended to limit the present invention, and various modifications and changes can be made to the present invention for those skilled in the art. Any modification, equivalent substitution, improvement and the like made within the spirit and principle of the present invention are intended to be included within the scope of the present invention.