

HAPLOTYPE-BASED METHOD FOR DETECTING FETAL GENOTYPE

DESCRIPTION

TECHNICAL FIELD

[0001] The present disclosure belongs to the field of biotechnology. More specifically, the present disclosure provides a haplotype-based method for detecting fetal genotype.

BACKGROUND OF THE INVENTION

[0002] Since 2000, in the field of non-invasive prenatal detection for monogenic diseases, researches and clinical applications of cfDNA in the peripheral plasma of pregnant women by PCR and related technologies were aimed at providing a clear diagnosis of the monogenic disease (Saito, Sekizawa *et al.* 2000), such as myotonic dystrophy (Amicucci, Gennarelli *et al.* 2000). Such detection methods mainly target a single gene or variant, and detect paternally inherited mutations and fetal *de novo* mutations by determining the presence of foreign alleles in maternal peripheral blood that differ from the genotypes of mother.

[0003] For maternal heterozygous loci, due to the mixing of fetal and maternal DNA fragments in plasma of pregnant women, the interference of a large amount of maternal DNA information makes it difficult to determine whether a fetus carries maternal mutations and impossible to determine whether there is an allele different from the mother by qualitative analysis. Therefore, quantitative analysis is necessary for each type of allele. Digital PCR, the first method to infer maternally inherited mutations, is capable of detecting minor allele imbalances. The use of relative mutation dosage (RMD) analysis in determining the change of the proportion of the two alleles and its relationship with FF (Fetal Fraction, the concentration of cfDNA from fetus) on maternal heterozygous loci has been validated in different diseases (Tsui, Kadir *et al.* 2011, Barrett, McDonnell *et al.* 2012).

[0004] The above mentioned methods all focus on specific variants or a small gene region. With the development of next generation sequencing technology, it becomes possible to perform whole-exome sequencing (WES) or whole genome sequencing (WGS) of cfDNA in maternal plasma to infer fetal genome, monogenic variant and *do novo* mutation in whole genome, based on the presence of fetal cfDNA in plasma of pregnant women.

[0005] Lo *et al.* used high-depth sequencing of cfDNA to analyze the maternal haplotype, and

with known maternal and paternal haplotypes to obtain sequencing reads;

2) aligning the sequencing reads to a reference genome, and determining the depth distribution and mutations of the sample according to the alignment results;

3) calculating the fetal cfDNA concentration according to the minor allele frequency for four types of allelic loci, wherein the four types of allelic loci comprise: an allelic locus where the mother is homozygous and the fetus is homozygous, an allelic locus where the mother is homozygous and the fetus is heterozygous, an allelic locus where the mother is heterozygous and the fetus is homozygous, and an allelic locus where the mother is heterozygous and the fetus is heterozygous; and

4) for the parental genotype combinations AAAB, ABAA and ABAB, inferring fetal SNP and InDel genotypes as follows:

a) for the parental genotype combination AAAB, using Bayesian algorithm to infer the fetal SNP genotype, and using CIVA method to infer the fetal InDel alleles inherited from the father based on the SNP haplotype inherited from the father;

b) for the parental genotype combination ABAA, using SPRT algorithm to infer a locus that can be covered by the haplotype and using CIVA method to infer the remaining loci;

c) for the parental genotype combination ABAB, firstly using CIVA method to determine the fetal locus type inherited from the father; then, using SPRT algorithm to infer the fetal locus type inherited from the mother; for the locus that cannot be inferred by SPRT algorithm, using CIVA method to infer the fetal allele type inherited from the mother; combining the paternal allele and maternal allele, so as to obtain the fetal genotype of the locus;

d) for a locus that cannot be covered by the haplotype data in the above three genotype combinations, using Bayesian algorithm to determine the genotypes of the locus; and

e) in the case where the fetal locus inferred by Bayesian algorithm is different from the parental locus, a fetal *de novo* mutation is identified.

[0013] In one embodiment, in 4),

the Bayesian algorithm comprises: constructing a Bayesian model based on the fetal concentration to infer the fetal genotype;

the SPRT algorithm comprises: identifying the fetal genotype at a locus where the mother is heterozygous and the father is homozygous by using the imbalance of read coverage of two

loci where mother is heterozygous and father is homozygous, and the loci where both mother and father are heterozygous; detect variant types comprising SNP, InDel and fetal *de novo* mutations; do not depend on the fetal concentration, with high precision when the fetal concentration is low (4%); and can achieve high accuracy when the sequencing depth is relatively low ($\sim 100\times$).

[0024] The embodiments of the present disclosure are explained in accompanying with an exemplary embodiment as below. In this embodiment, the haplotype-based method for detecting fetal genotype comprises:

[0025] 1) performing high-depth sequencing (e.g., $100\times$) of cfDNA in the plasma of a pregnant woman and stLFR sequencing of maternal and fetal paternal blood cells for each family to be analyzed;

[0026] 2) aligning all raw data (fq format) after quality control to the human reference genome by using alignment software such as BWA (mem mode); using methods such as Picard to remove the duplicate reads in the alignment and calculating the duplication rate; using the Base Quality Score Recalibration (BQSR) function in variant calling algorithms such as GATK to conduct the local correction in the alignment; using the Depth of Coverage function in variant calling algorithms such as GATK to calculate the depth distribution of each sample; using the population variant calling mode in variant calling algorithms such as GATK to detect single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels) of small fragments;

[0027] 3) using tools such as Hapcut2 or Longhap to assemble long fragments and analyze the sample haplotype based on barcode information for the sequencing data of single-tube long fragment reads (stLFRs) of paternal and maternal blood cells;

[0028] 4) calculating fetal cfDNA concentration: calculating the distribution of minor allele frequency (MAF), which is expected to have four peaks corresponding to four types of loci: a locus where the mother is homozygous and the fetus is homozygous, a locus where the mother is homozygous and the fetus is heterozygous, a locus where the mother is heterozygous and the fetus is homozygous, and a locus where the mother is heterozygous and the fetus is heterozygous; for each locus, when the fetal concentration is f , MAF corresponding to the peak of the locus where the mother is homozygous and the fetus is heterozygous is $f/2$ given that MAF distribution is obtained from reads of pregnant women and fetuses; therefore, for $MAF \in [0, 0.25]$, if the distribution is simulated as a mixed model of two normal distributions, it can be inferred that MAF corresponding to the peak in the read distribution obtained from the fetus is $f/2$, and the fetal concentration in the sample can be obtained;

WHAT IS CLAIMED IS:

1. A haplotype-based method for detecting fetal genotype, the method comprising:
 - 1) performing cfDNA sequencing on a plasma sample of a pregnant woman to be detected with known maternal and paternal haplotypes to obtain sequencing reads;
 - 2) aligning the sequencing reads to a reference genome, and determining the depth distribution and mutations of the sample according to the alignment results;
 - 3) calculating the fetal cfDNA concentration according to the minor allele frequency for four types of allelic loci, wherein the four types of allelic loci comprise: an allelic locus where the mother is homozygous and the fetus is homozygous, an allelic locus where the mother is homozygous and the fetus is heterozygous, an allelic locus where the mother is heterozygous and the fetus is homozygous, and an allelic locus where the mother is heterozygous and the fetus is heterozygous; and
 - 4) for the parental genotype combinations AAAB, ABAA and ABAB, inferring fetal SNP and InDel genotypes as follows:
 - a) for the parental genotype combination AAAB, using Bayesian algorithm to infer the fetal SNP genotype, and using CIVA method to infer the fetal InDel alleles inherited from the father based on the SNP haplotype inherited from the father;
 - b) for the parental genotype combination ABAA, using SPRT algorithm to infer a locus that can be covered by the haplotype and using CIVA method to infer the remaining loci;
 - c) for the parental genotype combination ABAB, firstly using CIVA method to determine the fetal locus type inherited from the father; then, using SPRT algorithm to infer the fetal locus type inherited from the mother; for the locus that cannot be inferred by SPRT algorithm, using CIVA method to infer the fetal allele type inherited from the mother; combining the paternal allele and maternal allele, so as to obtain the fetal genotype at the locus;
 - d) for a locus that cannot be covered by the haplotype data in the above three genotype combinations, using Bayesian algorithm to determine the genotypes of the locus; and
 - e) in the case where the fetal locus inferred by Bayesian algorithm is different from the parental locus, a fetal *de novo* mutation is identified.
2. The method of claim 1, in 4),
 - the Bayesian algorithm comprises: constructing a Bayesian model based on the fetal

concentration to infer the fetal genotype;

the SPRT algorithm comprises: identifying the fetal genotype at a locus where the mother is heterozygous and the father is homozygous by using the imbalance of read coverage of two alleles at heterozygous loci in the plasma cfDNA sequencing data of the pregnant woman; and

the CIVA method comprises: inferring two haplotypes of two alleles inherited from the father and the mother on the fetal locus to be predicted, respectively, and extracting the alleles that are actually at the locus in the two haplotypes, so as to infer the fetal genotype.

3. The method of claim 1 or 2, in 4), the Bayesian algorithm comprises: for each locus, when the read depth of different alleles and fetal concentration in the plasma of the pregnant woman are known, the probability of obtaining a combination i of maternal and fetal genotypes is calculated as follows:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

wherein, $P(A_i)$ is a prior probability of the combination i in n combinations of maternal and fetal genotypes obtained based on the fetal concentration; $P(B|A_i)$ is the cumulative probability of the combination i of maternal and fetal genotypes calculated based on the read depth at the locus and the prior probability; n combinations of maternal and fetal genotypes = (10 maternal genotypes $\times$ 10 fetal genotypes); when the paternal and maternal genotypes are known, the combination in the calculation of the prior probability is modified for the above Bayesian algorithm, to remove the combinations of maternal and fetal genotypes that do not conform to Mendel's laws of heredity, and the probability value of each combination is recalculated; and finally, the combination with highest probability is selected, so as to obtain the fetal genotype.

4. The method of claim 1 or 2, in the SPRT algorithm, all SNPs in the maternal genome are classified as α type and β type, and two haplotypes of the mother are numbered as Hap I and Hap II; the α type SNP is defined as the allele of the father in the SNP being the same as that of the mother in Hap I, and the β type SNP is defined as the allele of the father in the SNP being the same as that of the mother in Hap II; for the α type SNP, the number of allele reads at the locus in Hap I increases when the fetus inherits Hap I from the mother; however, the numbers of reads of two alleles at the locus are close to balance when the fetus inherits Hap II from the mother; for the β type SNP, the numbers of reads of two alleles at the locus are close to balance when the

fetus inherits Hap I from the mother; however, the number of allele reads at the locus in Hap II increases when the fetus inherits Hap II from the mother.

5. The method of claim 4, in the SPRT algorithm, the following two thresholds are used to determine whether two alleles at each locus are balanced,

$$\text{upper boundary} = \frac{(\ln 1200)/N - \ln d}{\ln g}$$

and

$$\text{lower boundary} = \frac{(\ln 1/1200)/N - \ln d}{\ln g}$$

wherein,

$$d = \frac{1 - q_1}{1 - q_0}$$

and

$$g = \frac{q_1(1 - q_0)}{q_0(1 - q_1)}$$

wherein, q_0 is the number of reads of Hap I at the locus when the fetus inherits Hap II from mother, q_1 is the number of reads of Hap I at the locus when the fetus inherits Hap I from mother, and N is the total number of reads;

for the α type SNP, $q_0=0.5$, $q_1=ff/2+0.5$; for the β type SNP, $q_1=0.5$, $q_0=0.5-ff/2$.

6. The method of claim 1 or 2, in e) of 4), a *de novo* mutation in the fetus is further screened by the following criteria:

i) all variants recorded in dbSNP147, Genome Aggregation Database (gnomAD) with minor allele frequency (MAF) > 1% in East Asian population, and internal database;

ii) a locus with the paternal or maternal or fetal genotype quality (GQ) < 30;

iii) a locus with the paternal or maternal or fetal sequencing depth < 20;

iv) a locus with the paternal or maternal or fetal genotype likelihood (PL) < 30;

v) a locus with the paternal or maternal or fetal minor allele frequency < 0.3 or > 0.7; and

vi) a locus with the following quality control criteria in the combined VCF (Variant Call Format, generated by GATK) of parent and fetus: QualByDepth (QD) < 2.0, RMSMappingQuality (MQ) < 50.0, FisherStrand (FS) > 60.0, StrandOddsRatio (SOR) > 3.0,

MappingQualityRankSumTest(MQRankSum) <-12.5, ReadPosRankSumTest (ReadPosRankSum) < -8.0.

7. The method of claim 1 or 2, in 1), the maternal and paternal haplotypes are obtained by stLFR (single tube Long Fragment Read) sequencing of the maternal and paternal blood cells, respectively.
8. The method of claim 1 or 2, in 1), the depth of cfDNA sequencing is 70-200×.
9. The method of claim 8, the depth of cfDNA sequencing is 100×.