

基于单体型的胎儿基因型检测方法

技术领域

本发明涉及生物技术领域，更具体而言，本发明提供了一种基于单体型的胎儿基因型检测方法。

5 背景技术

在无创产前检测单基因病方面，自 2000 年起，采用 PCR 及相关技术对孕妇外周血浆 cfDNA 的研究和临床应用的目标是对某些单基因病(Saito, Sekizawa et al. 2000)如肌强直性营养不良(Amicucci, Gennarelli et al. 2000)等进行明确诊断。这种检测方法主要针对的是单个基因或者变异位点，通过判断母亲的外周血中是否存在与自身基因型不同的外来等位基因，检测父源遗传突变和胎儿新发突变（de novo mutation）。

对于母亲杂合位点，由于孕妇血浆中混合了胎儿和母亲的 DNA 片段，存在大量母源 DNA 信息的干扰，导致胎儿是否携带母源突变难以判断，无法通过定性分析判断是否存在与母体自身不同的等位基因，因此需要对每种类型等位基因进行定量分析。数字 PCR 能够检测微小等位基因不平衡，是第一种推测母源遗传突变的方法。采用相对突变剂量（relative mutation dosage, RMD）分析，针对母亲杂合位点判断两种等位基因的比例变化及其与 FF（Fetal Fraction，胎儿来源的 cfDNA 的浓度）的比例关系，在不同的疾病上得到了验证(Tsui, Kadir et al. 2011, Barrett, McDonnell et al. 2012)。

前文提及的方法均是针对特定的变异或少部分基因区域的检测，随着下一代测序技术的发展，基于孕妇血浆中存在胎儿 cfDNA，通过对母体血浆中 cfDNA 进行全外显子测序（whole-exome sequencing, WES）或全基因组测序（whole genome sequencing, WGS）测序，在全基因组范围内推测胎儿的基因组及单基因变异、新发突变成为可能。

Lo 等人(Lo, Chan et al. 2010)率先使用 cfDNA 高深度测序及对母体的单体型进行分析，开发基于单体型剂量（relative haplotype dosage, RHDO）的方法揭示整个胎儿基因组，为该领域的研究奠定了基础。该方法利用孕妇血浆 cfDNA 测序数据中杂合位点两条等位基因的读段覆盖数的不平衡性，开发相对单体型剂量（relative haplotype dosage, RHDO）方法来判断其中胎儿基因型中，母亲杂合父亲纯合的位点。

在 Lo 等人的研究成果发表后的几年里，人们致力于改进基于单体型的方法，这些研究着眼于使用各种技术直接对母亲进行单体型分型，以避免对先证者等其他家庭成员单体型的依赖(Fan, Gu et al. 2012, Kitzman, Snyder et al. 2012)。(Fan et al., 2012)的方法利用孕妇血浆 cfDNA 数据及单细胞测序数据进行胎儿基因组推断。该方法首先将孕妇外周血中淋巴细胞提取出来完成单细胞测序，并用直接确定相位（direct deterministic phasing）的方法得到母亲的两条染色体层面的高准确性单体型。再将孕妇血浆中 cfDNA 完成全基

基因组（或外显子组）测序。通过计数孕妇血浆中孕妇杂合位点的读段数量来确定胎儿遗传自母亲的单体型。再通过计算父亲特有杂合位点在孕妇血浆中的读段数量来确定父亲杂合位点中遗传给胎儿的等位基因，之后利用这些父亲杂合位点信息完成父亲单体型的重建，从而推断出胎儿遗传自父亲的单体型。结合上述两条单体型结果，即得到胎儿基因组。

上述利用孕妇血浆 cfDNA 数据及父母单体型测序数据进行胎儿基因组推断的方法共有的局限性是依赖于单体型信息：这些方法要么需要依赖家族先证者的单体型信息来重建母体的单体型，要么依赖于应用并不广泛的复杂的实验技术获得母体单体型，需要大量的专业知识而且很耗时；这些方法由于多种原因，无法对胎儿的整个基因组进行分型，仅能覆盖部分的基因组位点，对于单体型无法覆盖的位点，无法完成胎儿基因型推断；利用单体型信息需要处理重组事件和分型的错误问题；部分方法无法判断父母双杂类型变异；这些方法无法检测新发突变；无法对 InDel 类型的胎儿基因型给出较为可靠的推断。基于单体型的方法有诸多限制，在妊娠早期由于 FF 过低和测序方法的限制等问题的存在，推测胎儿全部位点基因型仍不可行。

随着高通量测序的成本降低，研究者尝试采用更加可靠的测序方法进行高深度测序，如对孕妇血浆 cfDNA 采用免 PCR 的 WGS 测序技术。Chan 等人(Chan, Jiang et al. 2016)，采用 PCR-free 的实验方案减少测序错误及超高深度的 cfDNA WGS 测序，开发基因组范围的等位基因剂量（genome-wide relative allelic dosage, GRAD）分析方法，引领了单基因疾病全基因组 NIPD 的两个主要前沿：母体杂合位点推断和新发突变检测。该方法利用孕妇血浆 cfDNA 超高深度测序（~270×）及父母全基因组测序(whole-genome sequencing, WGS)数据进行胎儿基因组推断。该方法首先将母亲血浆 cfDNA 数据及父母 WGS 数据完成比对和变异检测，采用 PCR-free 的实验方案减少测序错误及超高深度的 cfDNA WGS 测序，开发基因组范围的等位基因剂量（genome-wide relative allelic dosage, GRAD）分析方法，从而得到母亲血浆中各个位点读段数量，之后结合孕妇血浆读段数进而得到胎儿基因组结果。优点是不需要父母单体型，可以推断新发突变。整个方法的局限性在于：没有对父母双杂的位点进行推断；需要较高的 FF；覆盖的母体杂合位点数远小于理论上的位点数。接下来，Rabinowitz 等人(Rabinowitz, Polsky et al. 2019)采用高通量测序技术对孕妇血浆 cfDNA 进行高深度测序，不需要父母单体型信息，采用贝叶斯模型结合机器学习方法，可以检测 InDel 和父母双杂位点，然而整体的检测精度相对其他算法较低。

利用孕妇血浆 cfDNA 超高深度测序数据进行胎儿基因组基因型推断，虽然可以做到胎儿基因组位点的全覆盖，但是由于对每个位点仅使用固定阈值来判断胎儿基因型，因而需要极高测序深度（300×以上），当测序深度较低（100×）或胎儿浓度较低时无法准确推断出胎儿基因型。

参考文献：

Amicucci, P., M. Gennarelli, G. Novelli and B. Dallapiccola (2000). "Prenatal Diagnosis

of Myotonic Dystrophy Using Fetal DNA Obtained from Maternal Plasma." *Clinical Chemistry* 46(2): 301-302.

Barrett, A. N., T. C. McDonnell, K. C. Chan and L. S. Chitty (2012). "Digital PCR analysis of maternal plasma for noninvasive detection of sickle cell anemia." *Clin Chem* 58(6): 1026-1032.

Chan, K. C., P. Jiang, K. Sun, Y. K. Cheng, Y. K. Tong, S. H. Cheng, A. I. Wong, I. Hudecova, T. Y. Leung, R. W. Chiu and Y. M. Lo (2016). "Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends." *Proc Natl Acad Sci U S A* 113(50): E8159-E8168.

Fan, H. C., W. Gu, J. Wang, Y. J. Blumenfeld, Y. Y. El-Sayed and S. R. Quake (2012). "Non-invasive prenatal measurement of the fetal genome." *Nature* 487(7407): 320-324.

Kitzman, J. O., M. W. Snyder, M. Ventura, A. P. Lewis, R. Qiu, L. E. Simmons, H. S. Gammill, C. E. Rubens, D. A. Santillan, J. C. Murray, H. K. Tabor, M. J. Bamshad, E. E. Eichler and J. Shendure (2012). "Noninvasive whole-genome sequencing of a human fetus." *Sci Transl Med* 4(137): 137ra176.

Lo, Y. M., K. C. Chan, H. Sun, E. Z. Chen, P. Jiang, F. M. Lun, Y. W. Zheng, T. Y. Leung, T. K. Lau, C. R. Cantor and R. W. Chiu (2010). "Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus." *Sci Transl Med* 2(61): 61ra91.

Rabinowitz, T., A. Polsky, D. Golan, A. Danilevsky, G. Shapira, C. Raff, L. Basel-Salmon, R. T. Matar and N. Shomron (2019). "Bayesian-based noninvasive prenatal diagnosis of single-gene disorders." *Genome Res* 29(3): 428-438.

Saito, H., A. Sekizawa, T. Morimoto, M. Suzuki and T. Yanaihara (2000). "Prenatal DNA diagnosis of a single-gene disorder from maternal plasma." *Lancet* 356(9236): 1170.

Tsui, N. B. Y., R. A. Kadir, K. C. A. Chan, C. Chi and Y. M. D. Lo (2011). "Noninvasive prenatal diagnosis of hemophilia by microfluidics digital PCR analysis of maternal plasma DNA." *Blood* 117(13): 3684-3691.

发明内容

针对现有技术中的至少部分问题，本发明的目的在于提供一种胎儿基因型检测方法，进行全基因组范围单位点变异检测，提供准确性并有效降低测序深度。

因此，本发明提供了一种基于单体型的胎儿基因型检测方法，所述方法包括：

1) 对于母亲和父亲的单体型分型已知的待测孕妇血浆样本，进行 cfDNA 测序，获得测序读段；

2) 将测序读段比对到参考基因组上，根据比对结果估计样本的深度分布和突变；

3) 对于四种类型的等位基因位点，根据次等位基因频率计算出胎儿样本的浓度，所

述四种类型的等位基因位点包括母亲纯合胎儿纯合、母亲纯合胎儿杂合、母亲杂合胎儿纯合和母亲杂合胎儿杂合；

4) 对于母亲父亲的基因型组合 AAAB、ABAA 和 ABAB，按如下方式推断胎儿 SNP 和 InDel 基因型，

5 a) 对于母亲父亲的基因型组合 AAAB，采用贝叶斯算法推断胎儿 SNP 基因型，并基于所述 SNP 父亲遗传的单体型类型，采用 CIVA 方法，推断胎儿遗传自父亲的 InDel 等位基因；

b) 对于母亲父亲的基因型组合 ABAA，先采用 SPRT 算法对其中的单体型可以覆盖的位点进行推断，剩余的位点采用 CIVA 方法进行推断；

10 c) 对于母亲父亲的基因型组合 ABAB，首先采用 CIVA 方法先判断胎儿遗传自父亲的位点类型；然后采用 SPRT 算法，推测胎儿遗传自母亲的位点类型；对于 SPRT 算法无法推断的位点采用 CIVA 方法推测母亲遗传给胎儿的等位基因类型；合并父源等位基因、母源等位基因得到该位点胎儿基因型；

15 d) 对于上述三种基因型组合中单体型数据无法覆盖的位点，采用贝叶斯算法确定其基因型；

e) 对于贝叶斯算法的胎儿位点与父亲母亲不同的情况，确定胎儿新发突变。

在一个实施方案中，在 4) 中，

所述贝叶斯算法包括：在得到胎儿浓度后，以此为基础建立贝叶斯模型，推测胎儿基因型；

20 所述 SPRT 算法包括：利用孕妇血浆 cfDNA 测序数据中杂合位点两条等位基因的读段覆盖数的不平衡性来判断其中胎儿基因型中，针对母亲杂合父亲纯合的位点；

所述 CIVA 方法包括：对于需要预测的胎儿位点，分别推测出该位点遗传自父亲及母亲的两种等位基因所在的两条单体型，提取两条单体型中实际在该位点的等位基因，推断胎儿基因型。

25 在一个实施方案中，在 4) 中，所述贝叶斯算法包括：对于每个位点，当已知孕妇血浆中不同等位基因的读段的深度以及胎儿浓度时，得到孕妇及胎儿基因型组合 i 的概率计算为：

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

其中， $P(A_i)$ 为基于胎儿浓度所得 n 种孕妇及胎儿基因型组合中出现 i 组合的先验概率； $P(B|A_i)$ 为基于该位点上读段深度及先验概率计算所得孕妇及胎儿基因型组合为 i 的累积概率；孕妇及胎儿基因型组合 $n = (10 \text{ 种孕妇基因型} \times 10 \text{ 种胎儿基因型})$ ；当已知父亲及母亲基因型时，则针对上述贝叶斯模型，对先验概率的组合进行修改，去除不符合孟德尔遗传规律的孕妇及胎儿基因型组合情况，重新计算每种组合的概率值；最终挑选出概率最高的组合，即得到胎儿基因型。

在一个实施方案中，在 SPRT 算法中，将母亲基因组中所有 SNP 分为 α 型和 β 型两种，同时将母亲的两条单体型编号为 Hap I 和 Hap II； α 型 SNP 中父亲的等位基因与母亲的 Hap I 中的等位基因相同，而 β 型 SNP 中父亲的等位基因则与母亲的 Hap II 中的等位基因相同；对于 α 型 SNP，当胎儿遗传到母亲 Hap I 时，则在该位点上 Hap I 的等位基因会存在读段数的增高；而当胎儿遗传到母亲 Hap II 时，则在该位点上两个等位基因的读段数趋近平衡；对于 β 型 SNP，当胎儿遗传到母亲 Hap I 时，该位点上两个等位基因的读段数趋近平衡；而当胎儿遗传到母亲 Hap II 时，则在该位点上 Hap II 对于的等位基因会存在读段数增高的情况。

在一个实施方案中，在 SPRT 算法中，通过如下两个阈值来判断每个位点两条等位基因是否平衡，

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

和

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

此处，

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

和

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

此处， q_0 是当胎儿遗传到母亲 Hap II 时位点上 Hap I 的读段总数， q_1 是当胎儿遗传到母亲 Hap I 时位点上 Hap I 的读段总数，N 为总读段数量；

对于 α 型 SNP， $q_0=0.5$ ， $q_1=ff/2+0.5$ ；对于 β 型 SNP， $q_1=0.5$ ， $q_0=0.5-ff/2$ 。

在一个实施方案中，在 4 的 e) 中，采用如下方式进一步筛选胎儿新发突变：

- i) dbSNP147、gnomAD (Genome Aggregation Database) 数据库中东亚人群次等位基因频率 (MAF) >1% 及内部数据库中记录的所有变异；
- ii) 该位点父或母或胎儿的基因型质量值 (Genotype quality, GQ) <30；
- iii) 该位点父或母或胎儿的测序深度 <20；
- iv) 该位点父或母或胎儿的基因型似然值 (genotype likelihood, PL) <30；
- v) 该位点父或母或胎儿的次等位基因频率 <0.3 或 >0.7；
- vi) 该位点父母胎儿合并后的 VCF (变异检测结果文件，由 GATK 生成) 结果中如下质控指标：QualByDepth (QD) < 2.0, RMSMappingQuality (MQ) < 50.0, FisherStrand (FS) > 60.0, StrandOddsRatio (SOR) > 3.0, MappingQualityRankSumTest (MQRankSum) < -12.5, ReadPosRankSumTest (ReadPosRankSum) < -8.0。

在一个实施方案中，在 1) 中，母亲和父亲的单体型分型分别通过对母血细胞和父亲血细胞进行 stLFR (single tube Long Fragment Read) 测序获得。

在一个实施方案中，在 1) 中，cfDNA 测序的深度为 70-200 倍，例如 100 倍。

- 5 本发明的方案将利用孕妇血浆游离 DNA 全基因组测序数据进行的胎儿基因组推断范围扩大至全基因组范围单位点变异，为胎儿遗传病筛查范围的扩大提供支持；通过将单体型算法与贝叶斯模型结合的方法，将仅利用孕妇血浆完成胎儿基因型推断的准确性进一步提高；在胎儿 cfDNA 浓度 4%，孕妇血浆游离 DNA 测序深度 100×时仍有较高的精度；将胎儿基因型推断的范围拓展到 InDel 类型。

附图说明

- 10 图 1 示出了胎儿基因型推断方法示意图。

具体实施方式

- 15 本发明的方案可以通过对孕妇外周血中游离 DNA 进行高深度全基因组测序、父母血细胞进行 stLFR 测序准确推断出胎儿基因型，从而对胎儿基因组存在的异常情况进行筛查。本发明中可以利用 stLFR 测序获得父母单体型信息，父源及母源单体型的推测可以通过引入隐马尔可夫模型、或基于片段组装等方法进一步提高准确性，从而提升整体的准确性。在本发明中，可以采用其他的单体型测序方法替代 stLFR 测序方法。本发明对胎儿可能携带的单基因病、多基因病、染色体疾病等多种遗传病的致病位点的检测有重要意义。同时，本发明的方案可应用于孕期疾病预测、胎儿表型推断等领域。

- 20 本发明的方案不依赖于先证者的单体型信息；将胎儿基因组分析的区域扩展至全基因组，将检测精度缩小至单碱基突变，将胎儿遗传性疾病检测的应用范围扩大，包括母亲纯合父亲杂合、母亲杂合父亲纯和、父母双杂的位点；检测的变异类型包括 SNP、InDel、胎儿新发突变；不依靠与胎儿浓度的高低，在胎儿浓度较低时（4%），结果仍然可以达到较高的精度；在测序深度相对较低（~100×）时依然达到较高准确性。

- 25 以下结合一个示例性实施方案，对本发明的方案进行说明。在该实施方案中，所述基于单体型的胎儿基因型检测方法包括：

1) 针对每个被分析家系，完成孕妇血浆 cfDNA 高深度测序（例如 100 倍）及母亲与胎儿父亲血细胞 stLFR 测序；

- 30 2) 所有下机数据（fq 格式）完成质控后使用 BWA（mem 模式）或其他比对软件将数据比对至人类参考基因组上；使用 Picard 等方法去除比对结果中的重复读段并计算重复率，使用 GATK 等变异检测算法中碱基质量值纠正 BQSR 功能完成比对结果的局部校正，使用 GATK 等变异检测算法中覆盖深度 Depth of Coverage 功能计算每个样本深度分布；使用 GATK 等变异检测算法中群体变异检测模式完成单位点（SNP）及小片段插入缺失（Indel）的检测；

3) 针对父本和母本的血细胞长片段 (single-tube Long fragment reads, stLFR) 测序数据, 使用 Hapcut2 或 Longhap 等工具基于 barcode 信息完成长片段的组装及样本单体型分析;

4) 胎儿样品浓度计算: 计算次要等位基因频率 (Minor allele frequency, MAF) 分布情况, 预计该分布存在四个峰值, 分别对应母亲纯合胎儿纯合、母亲纯合胎儿杂合、母亲杂合胎儿纯合、母亲杂合胎儿杂合四种类型位点; 已知 MAF 分布是由孕妇和胎儿来源的读段共同组成, 对于每个位点, 当胎儿浓度为 f 时, 母纯胎杂位点的峰值对应的 MAF 为 $f/2$; 因此, 对于 $MAF \in [0, 0.25]$, 将其分布模拟为两个正态分布的混合模型, 则可推测出胎儿来源读段分布的峰值对应的 MAF 为 $f/2$, 即得到该样本的胎儿浓度;

5) 贝叶斯算法: 得到胎儿浓度后, 以此为基础建立贝叶斯模型, 推测胎儿基因型; 对于每个位点, 当已知孕妇血浆中不同等位基因的读段的深度以及胎儿浓度时, 得到孕妇及胎儿基因型组合 i 的概率计算为:

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

其中, A_i 是 n 种组合中第 i 种组合的胎儿基因型, B 是数据观测的等位基因的碱基支持数; $P(A_i)$ 为基于胎儿浓度所得 n 种孕妇及胎儿基因型组合中出现第 i 组合的先验概率; $P(B|A_i)$ 为基于该位点上碱基测序深度及先验概率计算所得孕妇及胎儿基因型组合为 i 的累积概率; 孕妇及胎儿基因型组合 $n = (10 \text{ 种孕妇基因型} \times 10 \text{ 种胎儿基因型})$; 当已知父亲及母亲基因型时, 则针对上述贝叶斯模型, 对先验概率的组合进行修改, 去除不符合孟德尔遗传规律的孕妇及胎儿基因型组合情况, 重新计算每种组合的概率值; 最终挑选出概率最高的组合, 即得到胎儿基因型;

6) 单体型方法 a (SPRT (sequential probability ratio test) 算法 (Lo, Chan et al. 2010)): 本方法利用孕妇血浆 cfDNA 测序数据中杂合位点两条等位基因的读段覆盖数的不平衡性来判断其中胎儿基因型中, 针对母亲杂合父亲纯合的位点。

具体来说, 将母亲基因组中所有单核苷酸突变位点 (single-nucleoside polymorphism, SNP) 分为 α 型和 β 型两种, 同时将母亲的两条单体型编号为 Hap I 和 Hap II; α 型 SNP 中父亲的等位基因与母亲的 Hap I 中的等位基因相同, 而 β 型 SNP 中父亲的等位基因则与母亲的 Hap II 中的等位基因相同; 对于 α 型 SNP, 当胎儿遗传到母亲 Hap I 时, 则在该位点上 Hap I 的等位基因会存在读段数的增高; 而当胎儿遗传到母亲 Hap II 时, 则在该位点上两个等位基因的读段数趋近平衡; 对于 β 型 SNP, 当胎儿遗传到母亲 Hap I 时, 该位点上两个等位基因的读段数趋近平衡; 而当胎儿遗传到母亲 Hap II 时, 则在该位点上 Hap II 对应的等位基因会存在读段数增高的情况。

基于以上理论, 引入两个阈值来判断每个位点两条等位基因是否平衡, 这两个阈值的计算方法分别为:

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

和

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

此处，

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

和

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

此处， q_0 是当胎儿遗传到母亲 Hap II 时位点上 Hap I 的读段总数， q_1 是当胎儿遗传到母亲 Hap I 时位点上 Hap I 的读段总数， N 为总读段数量；

对于 α 型 SNP， $q_0=0.5$ ， $q_1=ff/2+0.5$ ；对于 β 型 SNP， $q_1=0.5$ ， $q_0=0.5-ff/2$ 。

通过比较每个位点上实际遗传自母亲 Hap I 及 Hap II 对应等位基因的读段累计比例与上述两个阈值的关系，即可推断出胎儿遗传自父亲和母亲的等位基因，从而确定胎儿基因型。母亲的 Hap I 上等位基因的累积深度占两条单体型总累积测序深度的比例，小于 lower boundary，胎儿遗传自母亲 Hap II，大于 upper boundary，胎儿遗传自母亲 Hap I。

7) 单体型方法 b (最近变异位点算法，closest informative variant algorithm, CIVA)：

对于需要预测的胎儿位点，分别推测出该位点遗传自父亲及母亲的两种等位基因所在的两条单体型，提取两条单体型中实际在该位点的等位基因，推断胎儿基因型；

具体来讲，根据连锁不平衡性，利用前一步贝叶斯算法或 SPRT 算法推测出来的胎儿基因型信息，利用被分析位点上下游一定区间内的距离最近的有价值的变异位点的胎儿遗传自父亲或母亲的单体型信息，推测该位点的单体型遗传情况。一定的区域长度可以选择 50 kb、100 kb、200 kb 等。

同时可以增加更加严格的判断价值的变异位点是否可信的方法：当被分析位点附近一定区域内上下游的两个价值的变异位点指向的胎儿遗传自父亲或母亲的单体型类型不一致，认定为不可信，该种情况下无法使用 CIVA。

有价值的变异位点包括下面两种：

推测父源单体型时，选取父杂母纯位点：如当父亲为 AB，母亲为 AA，胎儿为 AB 时，则判定父亲 B 等位基因所在单体型遗传至胎儿；如当父亲为 AB，母亲为 AA，胎儿为 AA 时，则判定父亲 A 等位基因所在单体型遗传至胎儿；

同理，可以推测母源单体型，利用母杂父纯位点：当母亲为 AB，父亲为 BB，胎儿为 BB 时，则判定母亲 B 等位基因所在单体型遗传至胎儿；当母亲为 AB，父亲为 BB，胎儿为 AB 时，则判定母亲 A 等位基因所在单体型遗传至胎儿；

8) 单体型与贝叶斯算法合并方法 (The combined method): 将 SNP 或 InDel 分为四类, 母亲父亲的基因型组合为 AAAB、ABAA、ABAB 和 AAAA, 合并上述贝叶斯、基于单体型信息的 SPRT、最近变异位点算法对 SNP、InDel 进行推断胎儿基因型 (图 1):

对于 SNP 类型的变异:

- 5 a) AAAB 类型的 SNP, 采用贝叶斯模型;
- b) ABAA 类型的 SNP, 先采用 SPRT 算法对其中的单体型可以覆盖的位点进行推断; 剩余的 SPRT 无法给出判断 (在 SPRT 上述方法提及的两个阈值之间) 采用 CIVA 方法;
- c) ABAB 类型的 SNP, 首先采用 CIVA 方法先判断胎儿遗传自父亲的位点类型; 然后采用 SPRT 算法, 推测胎儿遗传自母亲的位点类型; 对于 SPRT 无法推断的位点采用
- 10 CIVA 推测母亲遗传给胎儿的等位基因类型; 合并父源等位基因、母源等位基因得到该位点胎儿基因型;

d) 对于上述三种类型 SNP 中, 单体型数据无法覆盖的位点, 采用贝叶斯算法;

对于 InDel 类型的变异:

- a) AAAB 类型的 InDel, 基于前一步骤推断出的 AAAC 类型 SNP 父亲遗传的单体
- 15 型类型, 采用 CIVA 方法, 推断胎儿遗传自父亲的等位基因;
- b) ABAA 类型的 InDel, 先采用 SPRT 算法对其中的单体型可以覆盖的位点进行推断; 剩余的 SPRT 无法给出判断 (在 SPRT 上述方法提及的两个阈值之间) 采用 CIVA 方法;
- c) ABAB 类型的 InDel, 首先采用 CIVA 方法先判断胎儿遗传自父亲的位点类型;
- 20 然后采用 SPRT 算法, 推测胎儿遗传自母亲的位点类型; 对于 SPRT 无法推断的位点采用 CIVA 推测母亲遗传给胎儿的等位基因类型; 合并父源等位基因、母源等位基因得到该位点胎儿基因型;

d) 对于上述三种类型 InDel 中, 单体型数据无法覆盖的位点, 采用贝叶斯算法。

- 9) 胎儿新发 SNP 类型突变: 对于贝叶斯预测的胎儿的变异位点与父亲母亲不同,
- 25 并采用如下的过滤流程得到胎儿新发突变:

i) dbSNP147、gnomAD (Genome Aggregation Database) 数据库中东亚人群 MAF>1% 及内部数据库中记录的所有变异;

ii) 该位点父或母或胎儿的基因型质量值 (Genotype quality, GQ) <30;

iii) 该位点父或母或胎儿的测序深度<20;

- 30 iv) 该位点父或母或胎儿的基因型似然值 (genotype likelihood, PL)<30;

v) 该位点父或母或胎儿的次等位基因频率<0.3 或>0.7;

- vi) 该位点父母胎儿合并后的 VCF 结果中如下质控指标: QualByDepth (QD) < 2.0, RMSMappingQuality (MQ) < 50.0, FisherStrand (FS) > 60.0, StrandOddsRatio (SOR) > 3.0, MappingQualityRankSumTest (MQRankSum) < -12.5, ReadPosRankSumTest
- 35 (ReadPosRankSum) < -8.0;

QD: QualByDepth, 即经过深度校正的质量值。

MQ: RMSMappingQuality, 所有比对读段比对质量的均方根, 用来判断一个位点平均的比对质量。

FS: FisherStrand, 即经过 Fisher 精确检验判断当前变异是否存在链特异的倾向。

5 MQRankSum: MappingQualityRankSumTest, 对杂合位点进行的不同碱基之间比对质量的曼惠特尼秩和检验结果, 通过 Ref 和 Alt 碱基的比对质量差异来评估位点的可信度。

SOR: StrandOddsRatio, 使用 OR 值来判断变异位点链特异的程度, 值越大链特异程度越高。

10 ReadPosRankSum: ReadPosRankSumTest, 仍然是对杂合位点进行的秩和检验, 看不同的碱基是否倾向于出现在读段上的特定位置 (例如接近读段的起始或者终止)。

本发明的方法利用两种单体型算法与贝叶斯模型结合, 基于孕妇血浆深度信息及父母单体型信息进行胎儿单位点推断, 实现基因组位点全覆盖。在本发明中, 贝叶斯算法还可以结合母体与胎儿 cfDNA 的片段差异, 如片段长度、片段末端的基序 (Motif) 类型、在基因组上的比对位置的差异, 增加胎儿来源片段的权重, 提升贝叶斯算法的准确性。

15 对于 ABAB 类型位点, 首先根据贝叶斯算法判断基因组上 AAAB 位点父亲遗传给胎儿的单体型类型, 然后采用 CIVA 算法确定 ABAB 位点父亲遗传给胎儿的单体型, 采用 SPRT 算法预测 ABAB 位点母亲预测给胎儿的单体型, 联合推断出 ABAB 类型的胎儿基因型。

实施例

20 本发明人已于 2021 年内完成 50 个健康家系及 9 个疾病家系样本采集、测序及数据分析, 并基于此完成胎儿基因组推断, 以测试本胎儿基因组推断方法的可行性和准确性, 并进一步明确应用范围。

本发明在 5 个家系的母亲血浆 cfDNA 测序数据 (WGS 双端 100 bp 测序), 父亲及母亲 stLFR 数据 (WGS 双端 100 bp 测序) 上完成测试, 并使用胎儿脐血 WGS 数据 (双端 100 bp) 作为真集计算胎儿基因型推断的准确性 (表 1、2、3)。

25 5 个家系 SNP 类型位点中 AAAB 类型变异的预测精度为 $96.2\% \pm 5.8\%$ (均值 $\pm$ 标准差), ABAA 为 $96.2\% \pm 1.4\%$, ABAB 为 $87.2\% \pm 4.7\%$ 。

5 个家系 InDel 类型位点 AAAB 类型变异的预测精度为 (均值 $\pm$ 标准差) $94.6\% \pm 1.9\%$, ABAA 为 $80.2\% \pm 4.3\%$, ABAB 为 $79.3\% \pm 3.3\%$ 。

30 当胎儿浓度为 4% 时, cfDNA 测序深度 $128\times$ 时 (JK-53), SNP 的预测精度分别为, 95.6% (AAAB)、92% (ABAA)、74.7% (ABAB)

当胎儿浓度为 4%, cfDNA 测序深度 $270\times$ 时 (JK-53), 胎儿新发突变预测的敏感性为 91.3%, 阳性预测值 (PPV) 为 53.9%。

表 1. 测试家系的测序数据统计

家庭 ID	测序方法	原始读段数量	比对读段数量	比对率	平均深度	覆盖度	重复率	N50	分相 SNP 的分数
JK-7	脐带血细胞 WGS	1233682194	1231279450	99.81%	41.85	99.16%	3.75%		
JK-7	母亲 cfDNA WGS	4930998263	4899353438	99.36%	167.1	99.27%	32.58%		
JK-7	母亲血细胞 sLFR	2246895792	2213288894	98.50%	68.39	99.32%	53.12%	38692373	99.8%
JK-7	父亲血细胞 sLFR	1847735151	1825890887	98.82%	57.29	99.87%	46.48%	13595319	99.88%
JK-16	脐带血细胞 WGS	1547702979	1544031563	99.76%	52.34	99.16%	3.15%		
JK-16	母亲 cfDNA WGS	4470433381	4427313912	99.04%	150.99	99.27%	31.17%		
JK-16	母亲血细胞 sLFR	1126578266	1093285410	97.04%	34.05	99.20%	37.45%	43167036	99.78%
JK-16	父亲血细胞 sLFR	1071372133	1036265674	96.72%	32.43	99.81%	36.28%	43812809	99.84%
JK-18	脐带血细胞 WGS	1393244829	1391262033	99.86%	47.27	99.14%	3.49%		
JK-18	母亲 cfDNA WGS	3410469123	3388616846	99.36%	115.68	99.26%	22.39%		
JK-18	母亲血细胞 sLFR	1437373627	1404317757	97.70%	42.84	99.27%	36.91%	7009682	99.76%
JK-18	父亲血细胞 sLFR	1419559729	1363089772	96.02%	40.42	99.88%	43.98%	4303673	99.72%
JK-28	脐带血细胞 WGS	1388613164	1386761047	99.87%	47.38	99.13%	4.18%		
JK-28	母亲 cfDNA WGS	3311517554	3286161812	99.23%	112.03	99.27%	21.11%		
JK-28	母亲血细胞 sLFR	1444610729	1412560944	97.78%	42.08	99.27%	38.33%	1178766	99.05%
JK-28	父亲血细胞 sLFR	1284525331	1246303246	97.02%	37.82	99.85%	42.77%	1424125	99.26%
JK-53	脐带血细胞 WGS	1480341904	1477844272	99.83%	50.33	99.13%	4.23%		
JK-53	母亲血浆 cfDNA	8909756646	8870299941	99.56%	256.12	99.29%	40.84%		
JK-53	母亲血细胞 sLFR	933259760	919002827	98.47%	28.77	99.12%	17.24%	3552388	99.66%
JK-53	父亲血细胞 sLFR	1055522782	1035088382	98.06%	31.53	99.76%	13.31%	4708438	99.64%

表 2. 不同方案胎儿基因组推断在 SNP 类型位点上的准确性汇总

家庭	FF	杂合	贝叶斯模型				基于单体型的方法				融合方法	
			准确度		真预测的数目/位点数目		准确度		真预测的数目/位点数目		准确度	
			准确度	真预测的数目/位点数目	准确度	真预测的数目/位点数目	准确度	真预测的数目/位点数目	准确度	真预测的数目/位点数目	准确度	真预测的数目/位点数目
JK-7	0.09	AAAB	96.8%	631788/652391								
		ABAA	72.6%	909990/1253908			98.8%	1171962/1186610			96.1%	1209126/1253908
		ABAB	61.4%	429479/698962			96.4%	572296/593618			89.8%	627424/698962
JK-16	0.27	AAAB	97.7%	576306/590073								
		ABAA	91.8%	1014224/1104405			99.2%	1061980/1070070			98%	1082868/1104405
		ABAB	85.9%	540677/629324			97.5%	562304/576574			94.9%	597082/629324
JK-18	0.14	AAAB	96.4%	463476/480708								
		ABAA	75.7%	901292/1190790			99%	1112107/1123516			96.7%	1151107/1190790
		ABAB	65.9%	439900/667062			92.7%	488305/526841			86.5%	576938/667062
JK-28	0.12	AAAB	95.2%	313500/329369								
		ABAA	69.5%	440760/633774			98.5%	581675/590448			96%	608483/633774
		ABAB	57.3%	222910/388712			93.1%	268695/288587			83.2%	323556/388712
JK-53 (128X)	0.04	AAAB	95.6%	540046/564870								
		ABAA	59.6%	667583/1120831			97.3%	926919/952660			92.0%	998262/1085313
		ABAB	49.4%	325416/658784			89.1%	363047/407606			74.7%	481240/644645
JK-53 (256.12X)	0.04	AAAB	95%	590368/621392								
		ABAA	63.5%	706502/1112899			97.7%	979699/1002260			93.6%	1041875/1112899
		ABAB	51%	341350/668854			93.9%	462865/492889			81.8%	547115/668854

表 3. 不同方案胎儿基因组推断在 InDel 类型位点上的准确性汇总

JK-7	真预测的数目/AAAB 位点总数目		AAAB Accuracy		真预测的数目/ABAA 位点总数目		ABAA Accuracy		真预测的数目/ABAB 位点总数目		ABAB Accuracy		FF		InDel 总数目		有 SPRT 结果的 InDel		有 SPRT 结果的 InDel 比例	
	95446/123935	77%	77.3%	77%	190093/199209	95.4%	97.3%	92.7%	95069/123443	77%	80.2%	79.8%	0.09	0.27	460212	446587	335561	380533	97%	93.6%
JK-16	73714/95314	77.3%	77.3%	77.3%	142989/147016	97.3%	92.7%	95.5%	74754/93231	80.2%	79.8%	84.6%	0.27	0.14	358513	335561	335561	380533	93.6%	94.5%
JK-18	82721/104098	79.5%	79.5%	79.5%	158053/170435	92.7%	95.5%	92.3%	84550/106000	79.8%	84.6%	84.6%	0.14	0.12	402516	380533	380533	166233	94.5%	95%
JK-28	37904/42794	88.6%	88.6%	88.6%	71967/75357	95.5%	92.3%	92.3%	40697/48082	84.6%	84.6%	84.6%	0.12	0.04	174955	166233	166233	325815	95%	95%
JK-53	70971/90470	78.4%	78.4%	78.4%	131633/142619	92.3%	92.3%	92.3%	69269/92726	74.7%	74.7%	74.7%	0.04	0.04	342825	325815	325815		95%	95%

权 利 要 求 书

1. 一种基于单体型的胎儿基因型检测方法，所述方法包括：

1) 对于母亲和父亲的单体型分型已知的待测孕妇血浆样本，进行 cfDNA 测序，获得测序读段；

5 2) 将测序读段比对到参考基因组上，根据比对结果估计样本的深度分布和突变；

3) 对于四种类型的等位基因位点，根据次等位基因频率计算出胎儿样本的浓度，所述四种类型的等位基因位点包括母亲纯合胎儿纯合、母亲纯合胎儿杂合、母亲杂合胎儿纯合和母亲杂合胎儿杂合；

10 4) 对于母亲父亲的基因型组合 AAAB、ABAA 和 ABAB，按如下方式推断胎儿 SNP 和 InDel 基因型，

a) 对于母亲父亲的基因型组合 AAAB，采用贝叶斯算法推断胎儿 SNP 基因型，并基于所述 SNP 父亲遗传的单体型类型，采用 CIVA 方法，推断胎儿遗传自父亲的 InDel 等位基因；

15 b) 对于母亲父亲的基因型组合 ABAA，先采用 SPRT 算法对其中的单体型可以覆盖的位点进行推断，剩余的位点采用 CIVA 方法进行推断；

c) 对于母亲父亲的基因型组合 ABAB，首先采用 CIVA 方法先判断胎儿遗传自父亲的位点类型；然后采用 SPRT 算法，推测胎儿遗传自母亲的位点类型；对于 SPRT 算法无法推断的位点采用 CIVA 方法推测母亲遗传给胎儿的等位基因类型；合并父源等位基因、母源等位基因得到该位点胎儿基因型；

20 d) 对于上述三种基因型组合中单体型数据无法覆盖的位点，采用贝叶斯算法确定其基因型；

e) 对于贝叶斯算法的胎儿位点与父亲母亲不同的情况，确定胎儿新发突变。

2. 根据权利要求 1 所述的方法，在 4) 中，

25 所述贝叶斯算法包括：在得到胎儿浓度后，以此为基础建立贝叶斯模型，推测胎儿基因型；

所述 SPRT 算法包括：利用孕妇血浆 cfDNA 测序数据中杂合位点两条等位基因的读段覆盖数的不平衡性来判断其中胎儿基因型中，针对母亲杂合父亲纯合的位点；

30 所述 CIVA 方法包括：对于需要预测的胎儿位点，分别推测出该位点遗传自父亲及母亲的两种等位基因所在的两条单体型，提取两条单体型中实际在该位点的等位基因，推断胎儿基因型。

3. 根据权利要求 1 或 2 所述的方法，在 4) 中，所述贝叶斯算法包括：对于每个位点，当已知孕妇血浆中不同等位基因的读段的深度以及胎儿浓度时，得到孕妇及胎儿基因型组合 i 的概率计算为：

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

其中, $P(A_i)$ 为基于胎儿浓度所得 n 种孕妇及胎儿基因型组合中出现 i 组合的先验概率; $P(B|A_i)$ 为基于该位点上读段深度及先验概率计算所得孕妇及胎儿基因型组合为 i 的累积概率; 孕妇及胎儿基因型组合 $n = (10 \text{ 种孕妇基因型} \times 10 \text{ 种胎儿基因型})$; 当已知父亲及母亲基因型时, 则针对上述贝叶斯模型, 对先验概率的组合进行修改, 去除不符合孟德尔遗传规律的孕妇及胎儿基因型组合情况, 重新计算每种组合的概率值; 最终挑选出概率最高的组合, 即得到胎儿基因型。

4. 根据权利要求 1 或 2 所述的方法, 在 SPRT 算法中, 将母亲基因组中所有 SNP 分为 α 型和 β 型两种, 同时将母亲的两条单体型编号为 Hap I 和 Hap II; α 型 SNP 中父亲的等位基因与母亲的 Hap I 中的等位基因相同, 而 β 型 SNP 中父亲的等位基因则与母亲的 Hap II 中的等位基因相同; 对于 α 型 SNP, 当胎儿遗传到母亲 Hap I 时, 则在该位点上 Hap I 的等位基因会存在读段数的增高; 而当胎儿遗传到母亲 Hap II 时, 则在该位点上两个等位基因的读段数趋近平衡; 对于 β 型 SNP, 当胎儿遗传到母亲 Hap I 时, 该位点上两个等位基因的读段数趋近平衡; 而当胎儿遗传到母亲 Hap II 时, 则在该位点上 Hap II 对于的等位基因会存在读段数增高的情况。

5. 根据权利要求 4 所述的方法, 在 SPRT 算法中, 通过如下两个阈值来判断每个位点两条等位基因是否平衡,

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

和

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

此处,

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

和

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

20 此处, q_0 是当胎儿遗传到母亲 Hap II 时位点上 Hap I 的读段总数, q_1 是当胎儿遗传到母亲 Hap I 时位点上 Hap I 的读段总数, N 为总读段数量;

对于 α 型 SNP, $q_0=0.5$, $q_1=ff/2+0.5$; 对于 β 型 SNP, $q_1=0.5$, $q_0=0.5-ff/2$ 。

6. 根据权利要求 1 或 2 所述的方法, 在 4 的 e) 中, 采用如下方式进一步筛选胎儿新发突变:

25 i) dbSNP147、gnomAD (Genome Aggregation Database) 数据库中东亚人群次等位基因频率 (MAF) >1% 及内部数据库中记录的所有变异;

ii) 该位点父或母或胎儿的基因型质量值 (Genotype quality, GQ) <30;

- iii) 该位点父或母或胎儿的测序深度 <20 ;
 - iv) 该位点父或母或胎儿的基因型似然值 (genotype likelihood , PL) <30 ;
 - v) 该位点父或母或胎儿的次等位基因频率 <0.3 或 >0.7 ;
 - vi) 该位点父母胎儿合并后的 VCF (变异检测结果文件, 由 GATK 生成) 结果中如下质控指标: QualByDepth (QD) <2.0 , RMSMappingQuality (MQ) <50.0 , FisherStrand (FS) >60.0 , StrandOddsRatio (SOR) >3.0 , MappingQualityRankSumTest (MQRankSum) <-12.5 , ReadPosRankSumTest (ReadPosRankSum) <-8.0 。
- 5 7. 根据权利要求 1 或 2 所述的方法, 在 1) 中, 母亲和父亲的单体型分型分别通过对母血细胞和父亲血细胞进行 stLFR (single tube Long Fragment Read) 测序获得。
- 10 8. 根据权利要求 1 或 2 所述的方法, 在 1) 中, cfDNA 测序的深度为 70-200 倍。
9. 根据权利要求 8 所述的方法, cfDNA 测序的深度为 100 倍。