



(12)发明专利

(10)授权公告号 CN 103608818 B

(45)授权公告日 2017.12.08

(21)申请号 201180069972.0

(22)申请日 2011.11.18

(65)同一申请的已公布的文献号

申请公布号 CN 103608818 A

(43)申请公布日 2014.02.26

(30)优先权数据

61/462,972 2011.02.09 US

61/448,547 2011.03.02 US

61/516,996 2011.04.12 US

13/110,685 2011.05.18 US

61/571,248 2011.06.23 US

61/542,508 2011.10.03 US

13/300,235 2011.11.18 US

(85)PCT国际申请进入国家阶段日

2013.10.09

(86)PCT国际申请的申请数据

PCT/US2011/061506 2011.11.18

(87)PCT国际申请的公布数据

W02012/108920 EN 2012.08.16

(73)专利权人 纳特拉公司

地址 美国加利福尼亚州

(72)发明人 M·罗比诺威特茨 G·杰梅罗斯

M·班杰维齐 A·瑞安

Z·德姆科 M·希尔 B·齐默曼

J·班尼

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291

代理人 黄志华 石磊

(51)Int.Cl.

G06F 19/22(2006.01)

G01N 33/48(2006.01)

(56)对比文件

US 2007202525 A1, 2007.08.30, 全文.

US 2007184467 A1, 2007.08.09, 参见说明书第29、32-33、36、39、69-70、74-77、79、121、126-127、180、294、323段.

US 2009099041 A1, 2009.04.16, 参见说明书第9、19、41、59、65、93、108-120、124段.

WO 2010017214 A1, 2010.02.11, 全文.

审查员 张媛媛

权利要求书3页 说明书96页

序列表4页 附图18页

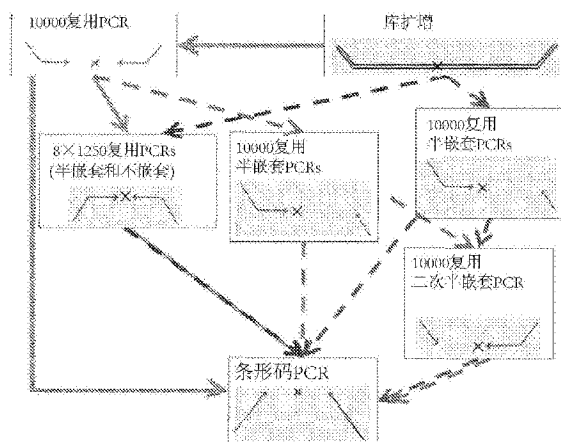
(54)发明名称

非侵入性产前倍性识别装置

(57)摘要

本发明提供了用于由从含有来自胎儿的母亲和来自胎儿的DNA的DNA的混合样本测得的基因型数据,以及任选的由母亲和父亲的基因型数据来确定妊娠中的胎儿的染色体的倍性状态的方法。所述倍性状态是通过使用联合分布模型创建不同可能的胎儿倍性状态的给定亲本基因型的数据多个预期的等位基因分布,并将预期的等位基因分布与混合样本中所测得的测量等位基因分布方式进行比较,选择预期的等位基因分布模式与所观测的等位基因分布模式最为相当的倍性状态来确定。DNA的混合样本可以以最小化等位基因偏差的方式在多个多态性位点进行优

先富集,例如使用大量多重定向PCR。



1. 一种计算机,该计算机被配置用于:

接收包括胎儿DNA和母亲DNA的DNA样本中染色体或者染色体片段上的多个多态性位点的DNA测序数据的输入;

由DNA测序数据来计算多个多态性位点的等位基因数;

创建每一个与染色体或染色体片段可能的不同倍性状态相关的多个倍性的假设;

建立用于每个倍性假设的染色体或染色体片段上的多个多态性位点的预期的等位基因数的联合分布模型;

通过使用联合分布模型和样本上测得的等位基因数来确定每个倍性假设的相对概率,以及样本中估测的胎儿DNA分数;

通过选择对应于具有最大概率的假设的倍性状态的方式来确定胎儿染色体或染色体片段的倍性状态,其中置信度估测适合于识别倍性状态确定。

2. 根据权利要求1所述的计算机,其特征在在于,样本中的DNA来源于母体血浆。

3. 根据权利要求1所述的计算机,其中所述计算机进一步被配置用于接收在胎儿父母的一方或双方的多个多态性位点上基因型数据的输入,并且其中针对每个倍性假设的染色体或染色体片段上的多个多态性位点的预期的等位基因数建立的联合分布模型是使用来自胎儿父母的一方或双方所获取基因型数据产生的。

4. 根据权利要求1所述的计算机,其中所述的样本是从母体血浆中分离出来的,其中所述计算机进一步被配置用于接收多个多态性位点的来自胎儿母亲的基因型数据的输入,并且其中来自母体的基因型数据是通过从所述DNA测序数据估测母体基因型数据来获得的。

5. 根据权利要求1所述的计算机,其中多个多态性位点为单核苷酸多态性。

6. 根据权利要求1所述的计算机,其中所述计算机被配置用于通过将有关涵盖染色体中的不同的位置的染色体概率的数据用于在染色体上的多态性等位基因之间的模型相关性建立等位基因数的联合分布模型。

7. 根据权利要求1所述的计算机,其中所述计算机被配置用于使用了样本中估测的胎儿DNA分数确定每个假设的相对概率。

8. 根据权利要求1所述的计算机,其中所述计算机被配置用于通过如下计算确定胎儿的倍性状态:

将采用联合分布模型和等位基因数确定的每个倍性假设的相对概率与每个倍性假设的相对概率结合,所述倍性假设的相对概率使用选自由读数分析,比较杂合率,一种仅当使用亲本遗传信息才能获得的统计信息,父母背景的归一化的基因型信号的概率,一种使用估测的第一样本或制备样本的胎儿DNA的分数计算的统计信息,及其组合所构成的组中的一种或多种统计技术进行计算。

9. 根据权利要求1所述的计算机,其中所述计算机进一步被配置用于计算针对确定的倍性状态进行置信度估测的计算,其中针对13号染色体,18号染色体和21号染色体的置信度超过99%。

10. 一种体外确定在妊娠中的胎儿的染色体或者染色体片段的一个或多个拷贝数的亲本起源的方法,所述方法包括:

在染色体或染色体片段的多个多态性位点上对样本中DNA测序,其中所述样本DNA包括来自胎儿母亲的母体DNA以及来自胎儿的胎儿DNA;

在计算机上由DNA测序数据来计算多个多态性位点的等位基因数；

在计算机上创建每一个与染色体或染色体片段的一个或多个拷贝数的不同亲本起源相关的多个假设；

在计算机上创建针对在妊娠中的胎儿的染色体或染色体片段的一个或多个拷贝数的亲本起源的一组至少一个假设；

用计算机建立用于每个假设的染色体或染色体片段上的多个多态性位点的预期等位基因数的联合分布模型；

用计算机通过使用联合分布模型和在样本上测得的等位基因数来确定每个假设的相对概率,以及样本中估测的胎儿DNA分数；

通过选择具有最大概率的假设来确定在妊娠中的胎儿的染色体或者染色体片段的一个或多个拷贝数的亲本起源,其中置信度估测适合于所述的亲本起源确定。

11. 根据权利要求10所述的方法,其中该方法进一步包括在测序步骤之前,在多个多态性位点优先富集样本中的DNA,并且其中在多个多态性位点优先富集的DNA包括:

获取多个连接介导的PCR探针,其中每个PCR探针定向一个多态性位点,其中相应的上游和下游的PCR探针被设定在DNA的一条链上与DNA区域杂交,其通过少数碱基与基因座上的多态性位点分离,其中少数是1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21-25,26-30,31-60,或它们的组合;

将连接介导PCR探针与来自样本的DNA杂交;

使用DNA聚合酶填补连接介导的PCR探针端之间的间隙;

连接连接介导的PCR探针;

扩增连接的连接介导的PCR探针。

12. 根据权利要求10所述的方法,其中该方法进一步包括在测序步骤之前,在多个多态性位点优先富集样本中的DNA,并且其中在多个多态性位点优先富集的DNA包括:

获取多个内部正向引物,其中每个引物定向一个多态性位点,并且内部正向引物的3'端被设计用于与来自多态性位点的DNA上游区域杂交,并通过少数碱基与多态性位点分离,并且其中少数选自1,2,3,4,5,6至10,11至15,16至20,21至25,26至30,或31至60个碱基对;

任选获取多个内部反向引物,其中每个引物定向一个多态性位点,并且内部反向引物的3'端被设计用于与来自多态性位点的DNA上游区域杂交,并通过少数碱基与多态性位点分离,并且其中少数选自1,2,3,4,5,6至10,11至15,16至20,21至25,26至30,或31至60个碱基对;

将内部正向引物引物和任选的内部反向引物与DNA杂交;以及

用聚合酶链反应扩增DNA形成扩增产物,其中所述扩增DNA是在一个或多个单独的反应室中进行,并且其中每个单独的反应室中含有超过500个不同的正向和反向引物对。

13. 根据权利要求12所述的方法,进一步包括:

获取多个外部正向引物,其中每个引物定向一个多态性位点,其中每个外部正向引物被设计用于与来自内部正向引物的DNA上游区域杂交;

任选获取多个外部反向引物,其中每个引物定向一个多态性位点,其中每个外部反向引物被设计用于与来自内部反向引物的下游的DNA区域瞬时杂交;

将外部正向引物和任选的外部反向引物与DNA杂交;

用聚合酶链反应扩增DNA。

14. 根据权利要求12所述的方法,进一步包括:

获取多个外部反向引物,其中每个引物定向一个多态性位点,其中每个外部反向引物被设计用于与来自内部反向引物的下游的DNA区域瞬时杂交;

任选获取多个外部正向引物,其中每个引物定向一个多态性位点,其中每个外部正向引物被设计用于与来自内部正向引物的上游的DNA区域杂交;

将外部反向引物和任选的外部正向引物与DNA杂交;

用聚合酶链反应扩增DNA。

15. 根据权利要求12所述的方法,进一步包括:

将通用适配子添加到样本中的DNA;以及

用聚合酶链反应扩增样本中的DNA。

16. 根据权利要求12所述的方法,其特征在于,所述内部引物通过识别可能形成不想要的引物双链的引物对,从大量引物中除去至少一对被识别为可能形成不想要的引物双链的引物来进行选择。

17. 根据权利要求10所述的方法,其中所述方法进一步包括在测序步骤之前,在多个多态性位点优先富集样本中的DNA,并且其中所述优先富集导致了在优先富集后的样本与优先富集之间的样本之间的等位基因的平均偏差程度不超过1.2倍。

18. 一种计算机,该计算机被配置用于:

a). 接收随机选自胎儿和母体的基因组DNA的混合物的DNA片段进行大规模并行DNA测序数据的输入;

b). 识别DNA片段属于的染色体或染色体片段;

c). 使用b)中的数据确定胎儿和母体的基因组DNA的混合物中至少一种第一染色体或染色体片段的量,其中,所述至少一个第一染色体或染色体片段被推定为胎儿整倍体;

d). 以b)中的数据确定所述母体和胎儿的基因组DNA的混合物中第二染色体或染色体片段的量,其中,所述第二染色体或染色体片段被推定是胎儿非整倍体;

e). 计算胎儿和母体的DNA的混合物中的胎儿DNA的分数;

f). 计算第二染色体或染色体片段的量的预期分布,针对其中所述第二染色体或染色体片段是整倍体的情况,使用c)的数值和计算的胎儿和母体的DNA的混合物中的胎儿DNA的分数;

g). 计算第二染色体或染色体片段的量的预期分布,针对第二染色体或染色体片段是非整倍体的情况,使用c)的数和在e)计算出的胎儿和母体DNA的混合物中的胎儿DNA的分数;以及

h). 使用最大似然度或最大后验概率法确定是否d)中所确定的第二染色体或染色体片段的量更有可能是f)中计算出的或在g)中计算出的部分分布其中,置信度估测适合于所述的染色体的量的确定;从而表明胎儿存在或不存在非整倍性。

非侵入性产前倍性识别装置

[0001] 相关申请

[0002] 本申请是2011年5月18日提交的美国申请No.13/110,685 的部分连续案,要求享有2010年5月18日提交的美国临时申请 No.61/395,850,2010年6月21日提交的美国临时申请 No.61/398,159,2011年2月9日提交的美国临时申请 No.61/462,972,2011年3月2日提交的美国临时申请 No.61/448,547,2011年4月12日提交的美国临时申请No.61/516,996的权利;本申请还要求享有2011年6月23日提交的美国临时申请No.61/571,248的权利,这些申请所公开的全部内容通过引用其中的教导包含于此。

发明领域

[0003] 本发明一般涉及非侵入性的产前倍体调用方法。

背景技术

[0004] 目前的产前诊断方法,可以向医生和家长提醒胎儿发育中的异常。如果不进行产前诊断,在出生时,50个婴儿中会有一个出现严重的身体或精神障碍,多达30个中有一个将有某种形式的先天性畸形。不幸的是,标准的方法要么精度差,要么涉及侵入性程序,有导致流产的风险。基于母体血液中的激素水平或超声测量的方法,虽然是非侵入性的,但是,它们也有精度低的缺点。如绒毛膜羊膜穿刺术,绒毛活检和胎儿的血液采样的方法具有精度高的优点,但是它们是侵入性的并且存在重大风险。所有在美国对所有怀孕者施行羊膜穿刺术的约占3%,但在过去十五年中,其使用频率已经下降。

[0005] 最近已发现,无细胞的胎儿DNA与完整的胎儿细胞可以进入母体血液循环。因此,对这种遗传物质进行分析可以实现早期的非侵入性产前基因诊断(NPD)。

[0006] 正常的人类在每一个健康的二倍体细胞中有两套23条染色体,分别来自父母的一个拷贝。非整倍体,一种在核细胞细胞中含有太多和/或太少的染色体的情况,被认为是导致大比例植入失败、流产和遗传性疾病的原因。染色体异常检测可以识别个体或胚胎,如唐氏综合症,克氏综合症,特纳综合症的条件下,其中包括,除了增加成功怀孕的机会。染色体异常的测试像母体的年龄一样特别重要,据估计,在35和40岁之间的有至少40%的胚胎是不正常的,而在40岁以上时,有一半以上的胚胎是不正常的。

[0007] 用于产前筛查的一些测试

[0008] 如在母体血清中第一孕期测量的,与妊娠相关的血浆蛋白A (PAPP-A) 的水平,可能与胎儿染色体异常有关,其中包括三倍体13、18和21。此外,在第一孕期的低PAPP-A水平可能预测不良的妊娠结果,包括一个小胎龄儿 (SGA) 的婴儿或死胎。孕妇经常进行第一孕期血清筛查,这通常涉及测试妇女血液中的激素PAPP-A和β人绒毛膜促性腺激素 (β-HCG)。在某些情况下,也给妇女进行超声波检查以便寻找可能的生理缺陷。特别是对胎儿颈部透明带 (NT) 测量,可以指示非整倍体的风险。在许多领域,标准的产前筛查治疗包括第一孕期结合NT测试的血清筛查。

[0009] 三重测试,也称为三重筛查,凯特林测试或巴特测试,是在第二孕期对患者进行是

否存在染色体异常和神经管缺陷高风险或低风险分类的调查。有时用“多个标记的筛选试验”术语来代替。术语“三重测试”可以涵盖的术语“双重测试”、“四合一检测”、“四重测试”和“五重测试。”

[0010] 三重测试测量血清甲胎蛋白 (AFP), 未结合雌三醇 (UE3), β 人类绒毛膜促性腺激素 (β -HCG), 浸润性滋养层抗原 (ITA) 和/或抑制体。一个阳性测试意味着一个高风险的染色体异常 (神经管缺陷), 并对这中患者再通过更敏感和特别的程序获得一个明确的诊断, 程序大多是侵入性的, 如羊膜穿刺术。三重测试, 可用于筛查的一些条件, 包括21三倍体综合征 (唐氏综合症)。除了唐氏综合症, 三重测试和四重测试还可筛查三倍体胎儿综合症18也称为爱德华氏综合症、开放的神经管缺陷, 也可以检测增加的特纳氏综合症、三倍体、16三倍体嵌合, 胎儿死亡, 综合症, 类固醇硫酸酯酶缺乏症的风险。

[0011] 发明概述

[0012] 本发明所公开是用于测定妊娠中的胎儿的染色体倍性水平的方法。根据本发明附图所示的各方面, 在一个实施方式中, 用于测定妊娠中的胎儿的染色体倍性水平的方法包括获取含有源自胎儿母体的母本DNA和来自胎儿的胎儿DNA的DNA第一样本, 通过分离DNA的方式制备第一样本以便获得制备样本, 测量染色体上的多个多态性位点的制备样本的DNA, 用计算机由制备样本测得的DNA计算多个多态性位点的等位基因数, 用计算机创建每一个与染色体可能不同倍性水平相关的多个倍性的假设, 用计算机建立用于每个倍性假设的在染色体上的多个多态性位点的预期的等位基因数的联合分布模型, 用计算机通过使用联合分布模型和制备样本上测得的等位基因计数确定每个倍性假设的相对概率, 通过选择对应于具有最大概率的假设的倍性水平的方式来识别胎儿的倍性水平。

[0013] 在一些实施方式中, 第一样本中的DNA源于母体血浆。在一些实施方式中, 所述制备第一样本的步骤还包括扩增DNA。在一些实施方式中, 所述制备第一样本的步骤还包括在多个多态性位点的第一样本中优先富集DNA。

[0014] 在一些实施方式中, 所述在多个多态性位点的第一样本中优先富集DNA包括: 获取多个预环化的探针, 其中每个探针都靶向一个多态性位点, 其中探针的3'和5'端被设计与通过少数碱基而与多态性位点区相分离的DNA区域杂交, 少数是1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 2, 21至25, 26至30, 31至60, 或它们的组合, 将预环化探针与来自第一样本的DNA进行杂交, 使用DNA聚合酶填补杂交的探针端之间的间隙, 环化预环化探针, 以及扩增环化探针。

[0015] 在一些实施方式中, 所述在多个多态性位点优先丰富DNA 包括: 获取多个连接介导的PCR探针, 其中每个PCR探针靶向一个多态性位点, 并上游和下游的PCR探针被设计在DNA的一条链上与DNA区域杂交, 其通过少数碱基与座上的多态性位点分离, 其中少数是1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21-25, 26-30, 31-60, 或它们的组合, 将连接介导PCR探针与来自第一样本的DNA杂交, 使用DNA聚合酶填补连接介导的PCR探针端之间的间隙, 连接连接介导的PCR探针, 扩增连接的连接介导的PCR探针。

[0016] 在一些实施方式中, 所述在多个多态性位点优先丰富DNA 包括: 获取多个靶向多态性位点的杂交捕获探针, 将杂交捕获探针与第一样本中的DNA杂交, 从DNA的第一样本中物理地除去部分或全部的未杂交的DNA。

[0017] 在一些实施方式中, 所述多个杂交捕获探针被设计为侧翼的但不重叠的多态性位

点区域杂交。在一些实施方式中,所述多个杂交捕获探针被设计为侧翼的但不重叠的多态性位点区域杂交,并在侧翼的捕获探针的长度,可以选自选自小于约120个碱基,小于约110个碱基,小于约100个碱基,小于约90个碱基,小于约80个碱基,小于约70个碱基,小于约60个碱基,小于约 50个碱基,小于约40个碱基、小于约30个碱基,以及少于约 25个碱基构成的集合。在一些实施方式中,混合的捕获探针杂交的区域重叠的多态性位点,其中包括至少两个为每个多态性位点的杂交捕获探针杂交捕获探针的多个,每个杂交捕获探针被设计为是互补的,以在该多态性位点的不同的等位基因。

[0018] 在一些实施方式中,所述在多个多态性位点优先丰富DNA 包括:获取多个内部正向引物,其中每个引物靶向一个多态性位点,并且内部正向引物的3'端被设计用于与来自多态性位点的 DNA上游区域杂交,并通过少数碱基与多态性位点分离,其中少数选自1,2,3,4,5,6至10,11至15,16至20,21至25, 26至30,或31至60个碱基对,任选获取多个内部反向引物,其中每个引物靶向一个多态性位点,并且内部反向引物的3'端被设计用于与来自多态性位点的DNA上游区域杂交,并通过少数碱基与多态性位点分离,其中少数选自1,2,3,4,5,6至10, 11至15,16至20,21至25,26至30,或31至60个碱基对,将内部引物与DNA杂交,用聚合酶链反应扩增DNA形成扩增产物。

[0019] 在一些实施方式中,所述方法还包括:获取多个外部正向引物,其中每个引物靶向一个多态性位点,并且外部正向引物被设计用于与来自内部正向引物的DNA上游区域杂交,任选获取多个外部反向引物,其中每个引物靶向一个多态性位点,并且外部反向引物被设计用于与来自内部反相引物的下游的DNA区域瞬时杂交,将第一引物与DNA杂交,用聚合酶链反应扩增DNA。

[0020] 所述方法还包括:获取多个外部反向引物,其中每个引物靶向一个多态性位点,并且外部反向引物被设计用于与来自内部反向引物的下游的DNA区域瞬时杂交,任选获取多个外部正向引物,其中每个引物靶向一个多态性位点,并且外部正向引物被设计用于与来自内部正相引物的上游的DNA区域杂交,将第一引物与DNA杂交,用聚合酶链反应扩增DNA。

[0021] 在一些实施方式中,所述制备第一样本还包括:将通用适配子附加到第一样本中的DNA,以及用聚合酶链反应扩增第一样本中的DNA。在一些实施方式中,至少所扩增的扩增子的一小部分是小于100bp,沸点小于90bp,小于80bp,小于70bp,小于65bp,小于60bp,小于55bp,小于50bp,或小于45bp,其所占分数是10%,20%,30%,40%,50%,60%,70%,80%,90%,或99%。

[0022] 在一些实施方式中,所述扩增DNA是在一个或多个单独的反应室中进行,其中每个单独的反应室中含有超过100个不同的正向和反向引物对,超过200个不同的正向和反向引物对,超过 500个不同的正向和反向引物对,超过1000个不同的正向和反向引物对,超过2000个不同的正向和反向引物对,超过5000个不同的正向和反向引物对,超过10000个不同的正向和反向引物对,超过20,000个不同的正向和反向引物对,超过50,000个不同的正向和反向引物对,或超过10万个不同的正向和反向引物对。

[0023] 在一些实施方式中,所述制备第一样本还包括:将第一样本分成多个部分,其中各部分中的DNA优先富集在多个多态性位点的子集中。在一些实施方式中,所述内部引物通过识别可能形成不想要的引物双链的引物对,除去至少其中一对引物被认定为可能形成不想要的引物双链的大量引物来进行选择。在一些实施方式中,所述内引部物包含设计用于杂

交靶向多态性位点的上游或下游区域,并任选地包含被设计为可以PCR扩增的通用引物序列。在一些实施方式中,至少一些引物中另外含有因各单独的引物分子各不相同的随机区域。在一些实施方式中,至少一些引物另外含有分子条形码。

[0024] 在一些实施方式中,所述方法还包括从胎儿父母的一方或双方获取基因型数据。在一些实施方式中,所述从胎儿父母的一方或双方获取基因型数据包括:从所述父母制备DNA,其中所述制备包括在多个多态性位点优先富集DNA制得制备的亲本DNA,任选地扩增制备的亲本DNA,以及测量制备样本中亲本DNA的多个多态性位点。

[0025] 在一些实施方式中,所述建立用于染色体上的多个多态性位点的预期等位基因数概率的联合分布模型是通过使用从父母一方或双方获得的遗传数据来进行的。在一些实施方式中,所述第一样本已从母体血浆中分离出来,并且从母体中获得基因型数据是通过从制备样本测量DNA估算母体基因型数据来进行的。

[0026] 在一些实施方式中,所述优先富集产生制备样本和选自不超过2倍,不超过1.5倍、不超过1.2倍,不超过1.1倍,不超过1.05倍,不超过1.02倍,不超过1.01倍,不超过1.005倍,不超过1.002倍,不超过1.001倍和不超过1.0001倍的倍数的第一样本之间的等位基因的平均偏置程度。在一些实施方式中,所述多个多态性位点是SNP。在一些实施方式中,所述测量制备样本中DNA的步骤是通过测序进行的。

[0027] 在一些实施方式中,公开了一种诊断盒,用于帮助测定妊娠中的胎儿的染色体倍体水平,其中所述诊断盒能够执行根据权利要求1所述的方法的制备和测量步骤。

[0028] 在一些实施方式中,所述的等位基因数是概率而不是二进制数。在一些实施方式中,所述测量多个多态性位点的制备样本中的DNA也可用于确定胎儿是否对遗传了疾病相关的单倍型中的一个或者多个。

[0029] 在一些实施方式中,所述建立用于等位基因计数概率的分布模型是通过将有关涵盖染色体中的不同的位置的染色体概率的数据用于建立在染色体上的多态等位基因之间的模型相关性进行的。在一些实施方式中,所述建立用于等位基因数,以及确定每个假设的相对概率的步骤是通过采用需要使用的参照染色体的方法来进行的。

[0030] 在一些实施方式中,所述确定每个假设的相对概率的步骤利用制备样本中胎儿DNA的估测分数。在一些实施方式中,所述用于计算等位基因计数概率,并确定每个假设的相对概率制备样本的DNA测量包括原始遗传数据。在一些实施方式中,所述选择对应于所述具有最大概率的假设的倍性水平是采用最大似然估计或最大后验估计进行的。

[0031] 在一些实施方式中,所述识别胎儿的倍性水平的步骤还包括:结合各倍体假设的相对概率确定使用联合分布模型和等位基因数概率,其每个倍性假设的相对概率是使用统计技术计算的,统计技术取自包括阅读计数分析,比较杂合率,统计信息,仅当亲本遗传信息用于确定的亲本的归一化的基因型信号,使用估计的胎儿小部分的所述第一样本或样本的统计信息,计算,以及它们的组合的概率。

[0032] 在一些实施方式中,置信估计适合于识别倍性水平。在一些实施方式中,采取基于所识别的胎儿的倍性水平的临床操作,其特征在于,临床操作选自终止妊娠或维持妊娠。

[0033] 在一些实施方式中,所述方法能够在4和5之间的孕周,5和6之间的孕周;6和7之间孕周;7和8之间的孕周;8和9之间孕周,9和10之间孕周,10和12之间孕周,12和14之间孕周,14和20之间孕周,20和40之间孕周进行,在第一季度,在第三季度,第四季度,或其组

合。

[0034] 在一些实施方式中,一种报告,其显示测得的孕育中的胎儿的染色体的倍性水平,采用所述方法制做。在一些实施方式中,公开了一种试剂盒,用于测定在妊娠中的胎儿染色体的倍性水平被设计以根据权利要求9所述的方法使用,该试剂盒包括:多个内部正向引物和任选的多个内部反向引物,其中每个引物被设计用于与来自靶染色体上,和任选的附加的染色体上的一个多态性位点的上游和/或下游的DNA区域杂交,其中杂交区域通过少数碱基与多态性位点分开的,其中的少数选自由1,2,3,4,5,6 至10,11至15,16至20,21至25,26至30,31至60,以及它们的组合构成的集合。

[0035] 在一些实施方式中,一种用于确定包括胎儿和母体的基因组 DNA的母体组织样本中存在或不存在胎儿非整倍性的方法,该方法包括:a)从所述母体组织样本中获取胎儿和母体的基因组 DNA的混合物,b)进行大规模并行DNA片段的DNA测序随机选自步骤a)中的胎儿和母体的基因组DNA的混合物来确定所述DNA片段的序列,c)确定步骤b)中得到的序列的染色,d)使用步骤c)中的数据确定胎儿和母体的基因组DNA的混合物中至少一种第一染色体量,其中,所述至少一个第一染色体被推定为胎儿整倍体,e)以步骤c)中的数据确定的第二染色体所述母体和胎儿的基因组DNA的混合物中第二染色体的量,其中,所述第二染色体被推定是胎儿非整倍体,f)计算母体和胎儿的基因组DNA的混合物中的胎儿DNA的分数,g)计算第二靶染色体的量的预期分布,如果第二靶染色体是整倍体,使用步骤d)的数,h)计算第二靶染色体的量的预期分布,如果第二靶染色体是非整倍体,使用步骤d)的数和在步骤f)计算出的胎儿和母体DNA的混合物中的中胎儿DNA的分数,和i)使用最大似然或最大后验概率的方法确定是否步骤e)中所确定的量的第二染色体更有可能是在步骤g)中或在步骤h)中计算出的分布计算出的分布分数;从而表明胎儿非整倍体的存在或不存在。

附图说明

[0036] 本发明所公开的实施方式将进一步参考附图来进行说明,其中,几个视图中的相同的结构均用相同的标号表示。所示附图并不一定按比例绘制,而一般是着重于表示本发明所公开的実施方式的原理。

[0037] 图1:直接多重微型PCR法的图示。

[0038] 图2:半巢式微型PCR法的图示。

[0039] 图3:全巢式微型PCR法的图示。

[0040] 图4:半巢式微型PCR法的图示。

[0041] 图5:三重半巢式微型PCR法的图示。

[0042] 图6:单边巢式微型PCR法的图示。

[0043] 图7:单边微型PCR法的图示。

[0044] 图8:反向半巢式微型PCR法的图示。

[0045] 图9:半巢式法的一些可能的工作流程。

[0046] 图10:环形连接适配子的图示。

[0047] 图11:内标引物的图示。

[0048] 图12:一些内标引物的例子。

[0049] 图13:使用具有连接适配子结合区域的引物的方法的图示。

- [0050] 图14:采用两种不同分析技术的计算方法的模拟倍性识别精度。
- [0051] 图15:在实验4中的细胞系的多个SNP位点的两种等位基因型的比率。
- [0052] 图16:在实验4中的细胞系的多个SNP位点的两种等位基因型的比率通过染色体排序。
- [0053] 图17:四个孕妇的血浆样本中的多个单核苷酸多态性的两种等位基因型的比率,排序条件为:染色体。
- [0054] 图18:数据修正前后分数数据可以通过二项式方差的方式来解释。
- [0055] 图19:图形显示样本中胎儿DNA相对富集后接着就是短库制备协议。
- [0056] 图20:深度读图比较直接PCR和半巢式法。
- [0057] 图21:深度阅读比较三个基因组样本的直接PCR。
- [0058] 图22:深度阅读比较三个样本的半巢式微型PCR。
- [0059] 图23:深度阅读比较1,200-重 (plex) 及9,600重反应。
- [0060] 图24:三个染色体上读取六细胞的计数率。
- [0061] 图25:在三个染色体上进行1ng的基因组DNA的两个三细胞反应和第三反应的等位基因比率。
- [0062] 图26:三个染色体上的两个单细胞反应的等位基因比率。
- [0063] 虽然上述确定的附图阐述本申请的实施方式,其他实施方式也是可以预期的,正如在讨论中所指出的。本申请仅以图示的方式提供了说明性实施方式,但不限于此。那些本领域技术人员在本发明公开的实施方式的原理的范围和精神之内可以衍生出许多其他的修改和实施方式。
- [0064] 发明详述
- [0065] 在一个实施方式中,本发明提供了用于从由混合DNA (即,来自胎儿母体的DNA,和来自胎儿的DNA) 样本测得的基因型数据和由来自母体也可以是父亲的基因物质样本测得的可选基因型数据确定妊娠胎儿染色体倍性状态的体外方法。其中确定是通过使用一个联合分布模型对父母基因型数据给出的不同可能的胎儿倍性状态创建一组预期的等位基因分布、对预期的等位基因分布和混合样本中测得的实际等位基因分布进行对比、并选择预期的等位基因分布模式与观察到的等位基因分布模式最相匹配的倍性状态达到的。在一个实施方式中,混合样本来自母体血液、或母体的血清或血浆。在一个实施方式中,混合DNA样本被优先在多个多态性位点富集。在一个实施方式中,优先富集可以通过等位基因偏差最小化的方式实现。在一个实施方式中,本发明涉及的DNA组成已优先在多个位点富集,因此等位基因偏差低。在一个实施方式中,等位基因分布通过对源自混合样本的DNA测序测得。在一个实施方式中,联合分布模型假定等位基因以二项式方式分布。在一个实施方式中,为基因连锁位点创建预期的联合等位基因分布组时,同时考虑现存的从各种来源的重组频率,例如,使用来自国际单体型图团体的数据。
- [0066] 在一个实施方式中,本发明提供了非侵入性产前诊断 (NPD) 的方法,具体地,通过观察等位基因的多态性位点的基因型数据上测量DNA的混合物,具体地,通过观察由DNA混合物测得的基因型数据中在多个多态性位点等位基因的测量结果确定胎儿的非整倍性状态。其中某些等位基因测量结果表示一个非整倍体的胎儿,而其他等位基因测量结果指标一个整倍体胎儿。在一个实施方式中,基因型数据的测量是通过对来自母体血浆的DNA混合

物进行测序。在一个实施方式中,可能会对DNA 样本的对应于经过等位基因分布计算的多个位点的DNA分子进行优先富集。在一个实施方式中,测定一个仅或几乎仅包含来自母体的遗传物质的DNA的样本,可能也是仅或几乎仅包含来自父亲的遗传物质的DNA的样本。在一个实施方式中,父母一方或双方的遗传测量与估计的胎儿组分一起用于创建多个预期的等位基因分布,对应于不同可能的潜在胎儿遗传状态。预期的等位基因分布可被称为假设。在一个实施方式中,在本质,母体的数据不是通过测量上完全是或几乎完全是母体的遗传物质确定的,而是根据对包含母体和胎儿DNA混合物的母体血浆进行遗传测量估计的。在一些实施方式中,假设可以包括在一个或多个染色体上的胎儿倍性,胎儿中某一染色体的某一片段继承自父母一方或他们的组合。在一些实施方式中,胎儿的倍性状态是通过比较观察到的等位基因的测量结果和不同的假设,其中至少有一些假设对应于不同倍性状态、选择对应于最有可能给出正确观察等位基因测量结果的假设的倍性状态确定的。在一个实施方式中,该方法包括使用源自部分或全部测量的SNP的等位基因测量数据,而不管位点是否是纯合或杂合的,因此,不涉及仅使用杂合位点的等位基因。这种方法可能不适合只涉及一个多态性位点的遗传数据的情况下。当遗传数据包含十多个目标染色体的多态性位点或二十多个多态性位点时,这种方法是特别有利的。当遗传数据包含超过50目标染色体的多态性位点,超过100或超过200个目标染色体的多态性位点时。这种方法是特别有利的,在一些实施方式中,遗传数据可能包括超过500个目标染色体的多态性位点,超过1000个多态性位点,超过2000个多态性位点,或5000个目标染色体的多态性位点。

[0067] 在一个实施方式中,本发明所公开的方法使用选择性富集技术,该技术在源自一组多态性位点的每个多态性位点保持出现在初始DNA样本中的相对等位基因频率在一些实施方式中,扩增和/或选择性富集技术可能涉及PCR,如连接介导PCR,片段杂交捕获,分子倒置探针,或其他圆形探针。在一些实施方式中,用于扩增或选择性富集的方法可能涉及使用探针,正确的目标序列杂交后,核苷酸探针的3-引物端或5-引物端通过少量核苷酸从等位基因的多态性位点分离。这种分离降低一个等位基因的优先扩增,被称为等位基因偏差。这是一种涉及使用探针改进的方法,其中,正确杂交探针的3-引物端或5-引物端直接与一个等位基因的多态性位点相邻或非常接近。在一个实施方式中,探针的杂交区域排除可能或肯定包含一个多态性位点。在杂交位置的多态性位点可能会导致不平等杂交,或在某些等位基因完全抑制杂交,导致某些等位基因优先扩增。这些实施方式,是对其他涉及目标扩增和或选择性富集的方法的改进,不管样本是源自一个单一个体的纯基因组样本或多人的混合样本,他们在每一个多态型位点更好地保留样本的初始等位基因频率。

[0068] 在一个实施方式中,本发明所公开的方法采用高效的高度复用的目标PCR扩增DNA,然后由高通量测序确定在每个目标基因位点的等位基因频率。这种在一个反应方式中复用超过约50 或100PCR引物,将所得序列读数目标位点地图的能力是新颖的和非显而易见的。一种允许高复用PCR、以高效方式实施的技术涉及不同于杂交的引物设计。在至少500,至少1000,至少为 5000,至少10000,至少20000,至少50000,或至少100000个潜在的引物对,或意想不到的引物和样品DNA的相互作用之间,通过创建一个具有潜在不良相互作用热力学模型选择PCR探针,然后利用该模型消除库中其它不兼容设计。另一种允许高度复用目标PCR,以高效率的方式实施的技术将部分或完整的嵌套方法用于目标PCR。使用这些方法中的一个或组合,在一个单一的池中与扩增的DNA一起,允许复用至少300,至少800,至少

1200,至少4000,或至少10000个引物。该包括大部分DNA分子,当测序,将映射到目标位点。使用这些方法中的一个或组合,在一个单一的池中与扩增的DNA一起,允许复用大量引物,该DNA 含有大于50%,大于80%,大于90%,大于95%,大于98%的或大于99%的映射到目标位点的DNA分子。

[0069] 在一个实施方式中,本发明所公开的方法对一个多态性位点的每个等位基因的多个独立观察产生定量测量。这不像大多数的方法,如微阵列或定性PCR,只提供关于两个等位基因比例的信息,但不定量提供对两种等位基因独立观察的数目。对提供定量独立观察数目信息的方法,仅比例用预倍性计算,但定量信息本身是没有用的。要说明保留有关独立观察数目的信息的重要性,考虑具有两个等位基因的样本位点,A和B。在第一个实验中,观察20个A等位基因和20个B等位基因,在第二个实验中,观察到200A等位基因和200B等位基因。在这两个实验中,两种等位基因的比值($A/(A+B)$)都等于0.5,但是第二个实验比第一试验传达了更多确定性的A或B等位基因的频率信息。

[0070] 在现有技术中已知的一些方法包括对源自单个等位基因的等位基因比率(信道比)(即 x_i/y_i)求平均或求和,并分析这个比例,不是用它与参照染色体对比就是使用一个关于这个比率在特定情况下是怎样的预期行为的规则。这些本领域中已知的方法没有包含等位基因加权,其假设大约相同数量的每个等位基因的 PCR产品可以确保与所有的等位基因的行为方式相同。这样的方法有许多缺点,并且更重要的是,排除了使用本发明在其他地方所描述的一些改进。

[0071] 在一个实施方式中,本发明所公开的一种方法在二倍体中对预期的等位基因频率分布,以及也许在减数分裂过程I中不分离、在减数分裂过程II中不分离、和/或在胎儿发育的早期有丝分裂过程中不分离造成的三倍体中预期的多个等位基因频率分布显式建模。为了说明为什么这很重要,想象一种没有交叉的情况:在减数分裂过程I中不分离会导致一个三倍体,其中两个不同的同源继承自一个父本,相反,在减数分裂过程II或在胎儿发育早期的有丝分裂过程中不分离将导致来自一个父本的同样同源的两份拷贝。每个场景都将在每个多态性位点导致不同预期的等位基因频率,并且由于遗传连锁,在共同考虑的所有位点都会出现上述情况。在同源之间,导致遗传物质交换的交叉会造成遗传模式更复杂。在一个实施方式中,除了使用基因位点之间的物理距离外,本方法通过使用重组率信息可以容纳这种情况。在一个实施方式中,为了能提高第一减数分裂不分离、第二减数分裂不分离或有丝分裂不分离之间的区分,本方法合并到一个在到着丝粒的距离的增加时提高交叉概率的模型中。第二减数分裂和有丝分裂不分离可通过有丝分裂不分离通常会导致一个同源染色体的相同或几乎相同的拷贝的事实区分,而目前的两个同系物,由于在配子发生过程中的一个或多个交叉,伴随的第二减数分裂不分离事件往往有所不同。

[0072] 在一些实施方式中,本发明所公开的方法包括对比所观察到的等位基因的测量结果和对应于可能的胎儿遗传非整倍体的理论假设,而且不包括在杂合基因位点定量等位基因比例的步骤。其中,基因位点的数目小于20,用一种包含在一个杂合基因位点定量等位基因比例的方法确定倍性和使用一种包含对比所观察的等位基因测定结果和对应于可能的胎儿遗传状态的理论等位基因分布假设的方法确定倍性可得到类似的结果。然而,其中的位点的数目达到50以上时,这两种方法可能得到显著不同的结果,位点的数目达到400以上,达到1000或2000以上时,这两种方法都极有可能给更加显著差异的结果。这些差异是由

于这样的事实,进行的方法,该方法包括在杂合基因位点定量等位基因的比例而不独立测量每个等位基因的大小和聚合或平均比例,排除了使用包括使用一个联合分布模型,执行连接分析,用一个二项式分布模型,和/或其它先进的统计方法的技术。而使用一种包括对比所观察到的等位基因测量结果和对应于可能的胎儿的遗传状态的理论等位基因分布的假设的方法,可以使用这些能大大增加测定的准确度技术。

[0073] 在一个实施方式中,本发明所公开的方法包括使用一个联合分布模型确定观察到的等位基因的分布测量结果是否表示一个整倍体或非整倍体胎儿。联合分布模型的使用是一个不同于通过独立处理多态性位点确定杂合度比率的方法并对显著更高精度的结果有显著改进。不受任何特定的理论的约束,但相信,他们具有更高的精度的原因之一是联合分布模型考虑到了SNPs之间的连接,在导致形成成长为胎儿胚胎配子的减数分裂期间,发生交叉的可能性。在为一个或多个假设创建期望的等位基因测量分布时使用连接概念的目的在于比不实用连接创建的预期等位基因测量分布能更好地与实际情况相对应。例如,设想有两个SNP 位点,1和2彼此靠近,母体是在一个同源上的SNP1上的A和 SNP2上的A,以及在一个同源2上的SNP1上的B和SNP2上的B。如果父体是在两个同源上的两个SNP为点上的A,而测得 B在胎儿的SNP1上,这显示同源而被胎儿继承,因此,在SNP2 出现胎儿的B存在很高的可能性。考虑了连接的一个模型会预测这种情况,而不考虑连接的一个模型则不能预测这种情况。另外,如果母体是在SNP1上的AB,而AB靠近SNP2,那么可以使用在该位置对应于孕妇三倍体的两个假设——一个涉及匹配的拷贝错误(在第二减数分裂不分离或胎儿发育早期有丝分裂不分离),另一个涉及匹配的拷贝错误(在第一减数分裂不分离)在匹配的拷贝错误三倍体的情况下,如果胎儿在SNP1从母体继承了AA,那么胎儿是更容易在SNP2从母体继承AA或BB,但不是AB。在一个不匹配的拷贝错误的情况下,胎儿会在两个SNP位点从母体继承AB。通过考虑了连接的倍性识别方法确定的等位基因分布假设能做出折现预测,因此,比没有考虑连接的倍性识别方法能在相当大的程度上对应于实际等位基因的测量。需要注意的是,当使用基于计算等位基因的比率和汇总那些等位基因比率的方法时,达成连接是不可能的。

[0074] 原因之一是,可以相信,当测序是用来测量等位基因时,使用包含对于观察的等位基因测量和对应于可能的胎儿遗传状态的理论假设的方法进行倍性确定具有更高的精度。这种方法比其他方法可以从读数的等位基因总数低的数据获取更多的信息。例如,一种基于计算和汇总等位基因比率的方法,会产生不成比例的加权随机噪声。例如,想象一种涉及使用测序测量等位基因的情况,有一组基因位点,其中对每个位点只检测到五个序列读数。在一个实施方式中,对每一个等位基因,数据可能与虚拟的等位基因分布比较,并根据序列读数的数目进行加权;因此,从这些测量获得的数据将适当地加权,并纳入到总体的确定结果中。这是相对于涉及在杂合位点定量等位基因比率的方法,由于这种方法只能计算0%,20%,40%,60%,80%或100%的比率作为可能的等位基因比率,这些比率没有一个能与预期的等位基因比率接近。在后一种情况下,计算出的等位基因比率,要么由于不充分的读书必须要丢弃,要么进行不成比例的加权,因而引入随机噪声到测定结果,从而降低判断的准确性。在一个实施方式中,单个等位基因的测量可能会被视为独立的测量,其中,相同位点等位基因测量之间的关系与不同位点测量之间的关系没有什么不同。

[0075] 在一个实施方式中,本发明所公开的方法包括,在不对任何度量与在一个预期为

二倍体的参考染色体上观察到的等位基因的分布的测量(称为RC方法)进行比较时,确定观察到的等位基因的测量是否表示一个整倍体或非整倍体胎儿。这是对诸如使用鸟枪法测序方法,即通过评估来自与一个或多个假定的二倍体参考染色体相关的嫌疑染色体的随机测序片段的比例检测非正北体的方法进行了显著改进的方法。如果假定二倍体参考染色体实际上不是二倍体时,这种RC方法会产生不正确的结果。这可能发生在一条染色体的非整倍体大大多于三倍体,或胎儿是三倍体以及所有的常染色体都是三倍体的情况下。在一个女性三倍体(69,XXX)胎儿的情况下,事实上是不存在二倍体染色体的。本发明所述的方法不需要参考染色体就能够在女性三倍体胎儿中正确地识别三倍体染色体。对于每个染色体、假设、胎儿分数和噪音水平,一个联合分布模型可以是合适的,不需要任何参考染色体数据、完整的胎儿分数估计、或一个确定的参考假设。

[0076] 在一个实施方式中,本发明所公开的方法演示如何在多态性位点观察等位基因分布用于确定胎儿的倍性状态时,比现有技术的方法具有更高的精度。在一个实施方式中,该方法使用靶向测序获得母体-胎儿混合基因型以及在多个SNP多态性位点任选的母体和/或父亲的基因型,根据不同的假设,先建立各种预期的等位基因频率分布,然后观察从母体-胎混合物获得的定量等位基因信息,并评估哪一种假设适合数据最好,与数据适合最好的假设对应的遗传状态被称为正确的遗传状态。在一个实施方式中,本发明所公开的方法也使用的吻合程度生成置信,所谓的遗传状态是正确的遗传状态。在一个实施方式中,本发明所公开的方法包括使用分析在具有不同父母背景的位点发现的等位基因分布的算法,并比较所观察到的等位基因分布和预期的等位基因分布以获取针对不同父母背景(不同的亲本基因型到模式)下的不同倍性状态。这不同于不使用能在一个母体-胎儿的混合样本中每个位点估计每个等位基因独立实例数目的方法,而是基于此方法的一种改进的方法。在一个实施方式中,本发明所公开的方法包括,使用在母本是杂合的基因位点所观察到等位基因分布结果,来确定观察到的等位基因分布的测量结果是否表示一个整倍体或非整倍体胎儿。这不同于不使用在杂合母本位点观察的等位基因分布的方法,而是一种改进方法,因为在对特定目标个体不知道是信息富集的位点不进行DNA优先富集或对DNA进行优先富集的情况下,在确定倍性时允许使用来自一组测序数据的两倍数量的基因测量数据,会产生更准确的确定结果。

[0077] 在一个实施方式中,本发明所公开的方法使用一个联合分布模型,假定在每个基因位点的等位基因频率的频道具有符合多项式(从而SNP多态性位点是双等位基因时符合二项式)的性质。在一些实施方式中,联合分布模型采用 β -二项式分布。当使用一种测量技术时,如测序,为出现在每个基因位点的每个等位基因提供定量测量,二项式模型可以应用到每个基因位点,相关的等位基因频率程度及其置信度可以确定。本领域已知从等位基因比率产生倍性识别的方法,或其中有定量等位基因信息被丢弃的方法,尚无法确定所观察到的比率的把握性。本方法是一种与计算等位基因比率和汇总这些比率以确定倍性不同的改进方法,因为任何涉及在特定基因位点计算等位基因比率,然后汇总这些比率的方法,必须假定所测得的给定等位基因或位点的DNA数量的强度将按照高斯模式分布的。本发明所公开的方法不涉及计算等位基因比率。在一些实施方式中,本发明所公开的方法可能涉及把在多个基因位点对每个等位基因的观察数量合并到一个模型中。在一些实施方式中,本发明所公开的方法可能涉及计算的预期分布本身,允许使用一个比任何假设等位基因测量

为高斯分布的模型更准确的联合二项式分布的模型。随着位点数的增加,二项式分布模型比高斯分布明显更准确的可能性提高。例如,当少于20个位点在接受查询时,二项分布模型显著更好的可能性是很低的。然而,当超过100,特别是超过400,尤其是超过1000,尤其是超过超过2000个位点时,二项式分布模型将具有非常高的明显比高斯分布模型更准确可能性,从而产生更准确的倍性确定。随着在每个位点观察数目的增加,二项分布模型比高斯分布更准确的可能性也随之提高。例如,当在每个被观察的基因位点观察到少于10个不同的序列时,二项式分布模型显著更好的可能性较低。然而,当每个位点超过50个序列读数,尤其是超过 100个序列读数,尤其是超过200个序列读数,特别是超过300 个序列读数时,二项式分布模型将有非常高的可能性明显比高斯分布模型更准确,从而产生更准确的倍性确定。

[0078] 在一个实施方式中,本发明所公开的方法使用测序来测定 DNA样本中每个位点上的每个等位基因的实际数量。每个测序读数可能被映射到一个特定的位点,并处理成一个二进制序列的读数,交替地,每个读数的属性和/或映射的概率可交替地合并为测序读数的一部分,形成一个概率序列读数,即映射到一个给定的位点可能的全部或部分序列读数的数目。使用二进制计数或计数的概率有可对每一组测量值使用二项式分布,围绕这些计数数量计算置信区间。使用二项分布的这种能力允许倍性估计更准确,以及计算更精确的置信区间。这不同于那些使用强度测量等位基因存在数量的方法,并进行了改进。例如使用微阵列的方法,或用电泳条带用荧光阅读器测量荧光标记的DNA强度进行测量的方法,本方法进行了改进。在一个实施方式中,本发明所公开的方法使用本方面的数据集,为该数据集确定估计等位基因分布频率的参数。这是一种改进的方法,利用训练的数据集或先前的数据集为现有预期的等位基因频率分布,或可能预期的等位基因比率设定参数。这是因为存在不同条件集涉及每个基因样本的收集和测量,因此,一个使用实时的数据集中的数据为用于倍性样本测定的联合分布模型确定参数的方法,往往是更准确的。

[0079] 在一个实施方式中,本发明所公开的方法包括使用最大似然技术,确定观察到的等位基因测量分布是否表示一个整倍体或非整倍体胎儿。最大似然技术的使用是不同于使用单个假设抑制技术方法的显著改进方法,这种方法所得到的测定结果具有显著的更高精度其中一个原因是单个假设抑制技术只是基于一个测量分布而不是两个设置切断阈值,这意味着阈值通常不是最佳的。另一个原因是,最大似然技术允对每一个样品进行切断阈值的优化,而不是对所有样品确定一个切断阈值而不考虑每一个样本的具体特征另一个原因是,最大似然技术的使用允许对每个倍性识别进行置信计算。对每个识别进行之心计算的能力允许一个医生知道哪些识别是准确的,哪些更可以是错误的。在一些实施方式中,各种各样的方法可以与最大似然估计技术结合,以提高倍性识别的准确度。在一个实施方式中,最大似然技术可能与在美国专利7888017中所描述的方法结合使用。在一个实施方式中,最大似然技术可能与以下方法结合使用,即在混合样品中用定向 PCR扩增技术扩增DNA,接着用一个读数的方法测序和分析,如串列诊断学中所使用的,见International Congress of Human Genetics2011,in Montreal in October2011.在一个实施方式中,本发明所公开的方法包括在混合样本中的估算胎儿DNA分数,并使用该估计计算倍性识别和倍性识别的置信度。注意,这是两个不同的和有区别的方法,使用估计胎儿基因分数作为有效胎儿基因分数的筛查,然后使用一个单一的不考虑到胎儿基因分数,也不对识别产生置信计算的假设

抑制技术进行倍性识别。

[0080] 在一个实施方式中,本发明所公开的方法考虑了数据噪声倾向,并包含通过对每次测量附加一个概率误差。使用从用附带概率估计的测量数据获得的一组假设中选择正确的假设的最大似然技术,更有可能降低不正确的测量,正确的测量值将用于获得倍性识别的计算。更准确地,这种方法系统地减少了不正确的倍性测定数据造成的影响。这是一个改进了的,假定所有的数据是同样正确的方法,或是从获得倍性识别的计算中任意排除外围数据的方法。用通道比测量的现有方法要求通过取单个SNP通道比平均值的办法将方法扩展到多个SNP多态性位点。不通过基于SNP多态性的质量和观察到的读数深度的预期的测量方差对个体的SNP多态性位点加权,会降低统计量的准确性,导致倍性识别的准确性显著降低,特别是在处在边界线的情况下。

[0081] 在一个实施方式中,本发明所公开的方法并不预设这样的常识,即对胎儿来说,SNP多态性或其他多态性位点是杂合的。此方法允许在父本基因型信息不可用的情况下作出倍性识别。这是一个改进了的方法,其中的常识是,为了适当地选择目标基因位点、或解释对胎儿/母体混合DNA样本做出的遗传测量,必须提前知道SNP多态性位点是杂合的。

[0082] 本文描述的方法用于只有少量DNA样本可用或胎儿DNA百分含量很低时特别有利。这是由于当只有少量的DNA可用时,基因位点的等位基因发生相应较高丢失率,和/或当在胎儿和母体DNA的混合样本中胎儿DNA百分比较低,发生相应较高的胎儿等位基因丢失率。高等位基因丢失率,是指对目标个体,有很大比例的等位基因没有测定到,导致不准确的胎儿分数计算的结果,以及不准确的倍性确定。由于本文公开的方法可使用联合分布模型,该模型考虑了SNP之间的继承模式的连接,可明显获得更准确倍性确定。本文描述的方法,当混合物中出现的胎儿DNA分子小于40%、小于30%、小于20%、小于10%、小于8%、甚至小于6%时,允许作出一个准确的倍性测定。

[0083] 在一个实施方式中,当一个个体的DNA与相关的个体DNA混合时,基于测量,有可能确定一个个体的倍性状态。在一个实施方式中,DNA的混合物是母体血浆中发现的游离的DNA,可能包括来自母体的DNA、已知的核型和基因型,也可以是与胎儿的DNA、未知的核型和基因型的混合物。为了获得不同的倍性状态、父母每一方对胎儿不同的染色体贡献、以及任选地,混合物中不同的胎儿DNA分数,使用来自父母一方或双方的已知基因型信息预测混合样品中DNA的多个潜在的遗传状态是可能的。每种潜在的组成可被称为一个假设,而胎儿的倍性状态可以由实际测量确定,然后确定哪种潜在的组成最有可能给出观测数据。

[0084] 在一些实施方式中,本发明所公开的方法可用于非常少量的DNA存在的情况,如在体外受精,或法医只有一个或几个(通常少于10个细胞,少于20个细胞或少于40个细胞)细胞可用的情况。在这些实施方式中,本发明所公开的方法可用于从少量未被其他DNA污染的DNA进行倍性识别,但用少量的DNA进行倍性识别非常困难的。

[0085] 在一些实施方式中,本发明所公开的方法可用于与目标DNA被另一个个体的DNA污染的情况,例如,产前诊断背景下的母体血液中,亲子鉴定,或产受孕测产品。其他一些情况下,如只有一个或少数细胞存在大量的正常细胞中间的癌症测试情况下,这些方法将是特别有利的。作为这些方法中的一部分使用的遗传测量,可以用于任何包含DNA或RNA的样本,例如但不限于:血液,血浆,体液,尿液,头发,眼泪,唾液,组织,皮肤,指甲,卵裂球,胚胎羊水,绒毛膜样本,粪便,胆汁,淋巴液,宫颈粘液,精液,或其他细胞或由核酸组成的材

料。在一个实施方式中,本发明所公开的方法可以与核酸的检测方法,如测序,微阵列,定量PCR,数字PCR,或用于检测核酸的其他方法一起运行。如果由于某种原因,认为是可取的,可以计算在一个位点的等位基因数的概率的比率,在这些方法是兼容的条件下,结合本发明描述的方法,等位基因比率可用于确定倍性状态,在一些实施方式中,本发明所公开的方法包括在计算机上在处理样本上测得的DNA的多个多态性位点计算等位基因比率。在一些实施方式中,本发明所公开的方法包括在计算机上在处理样本上测得的DNA的多个多态性位点计算等位基因比率以及本发明中描述的任何其他改进。

[0086] 在本文中的其他地方可能会发现对以上各点的进一步讨论。

[0087] 非侵入性产前诊断(NPD)

[0088] 非侵入性的产前诊断过程中包括一些步骤。某些步骤可包括:(1)从胎儿获得遗传物质,(2)可能在一个混合样本中,体外富集胎儿的遗传物质;(3)体外扩增遗传物质;(4)在体外优先富集遗传物质中的特定位点(5)在体外测量遗传物质;(6)在体外,用计算机分析基因型数据。本发明在此描述了降低到六个相关步骤的方法。至少有一些方法步骤不直接作用于身体。在一个实施方式中,本发明涉及脱离于身体的、适用于组织和其对他生物物质进行分离的治疗和诊断方法。至少一些方法步骤在计算机上执行。

[0089] 本发明的一些实施方式中,允许临床医生用一种非侵入性的方式确定妊娠胎儿的遗传状态,这样婴儿的健康不用置于收集胎儿的遗传物质的危险中,并且母体也不需要接受侵入性程序。此外,在某些方面,本发明允许遗传状态测定具有很好的测量精度,精确显著高于如非侵入性的母体血清分析物筛查,例如在产前护理广泛使用的三重测试。

[0090] 本发明所公开的高精度方法,如本发明所述,是信息学方法用于基因型数据分析的结果。现代技术进步产生了使用诸如高通测序和基因分型阵列方法从一个基因样本中测量大量遗传信息的能力。本文所公开的方法让临床医生更好地利用大量数据,对胎儿遗传状态做出更准确的诊断。下面给出了多个实施方式方式的细节。不同的实施方式可能涉及对上述步骤的不同组合。具有不同步骤的不同实施方式的各种组合可以互换使用。

[0091] 在一个实施方式中,从怀孕母体采取血液样本,分离自由浮动于母体血液血浆中的DNA,其中包含含有母系源的DNA和胎儿源的DNA的混合物,用于确定胎儿的倍性状态。在一个实施方式中,本发明所公开的方法包括优先富集那些混合物中的DNA序列,这些DNA序列以富集后等位基因比率和/或基因型频率分布大多保持一致的一种方式与多态型等位基因对应。在一个实施方式中,本发明所公开的方法包括基于扩增的高效定向PCR,以便产品分子有一个很高的比例与目标基因位点对应。在一个实施方式中,本发明所公开的方法包括对一种包含源自母体的DNA和源自胎儿的DNA的DNA混合物测序。在一个实施方式中,本文所公开的方法包括用测量等位基因分布来确定母体妊娠中胎儿的倍性状态。在一个实施方式中,本发明所公开的方法包括向临床医生报告所确定的倍性状态。在一个实施方式中,本发明所公开的方法包括采取的临床动作,例如,执行跟进如绒毛膜绒毛取样或羊膜穿刺术等侵入性测试,为准备一个三倍体个体的诞生,或三倍体胎儿的选择性终止。

[0092] 本申请参考了美国实用新型序列号为11/603406,申请日为2006年11月28日,(美国专利公开号:20070184467);美国实用新型,序列号为12/076348,申请日为2008年3月17日,(美国专利公开号:20080243398);PCT实用新型,序列号为PCT/US09/52730,申请日为2009年8月4日(PCT公开号:W0/2010/017214);PCT实用新型,序列号为PCT/US10/050824

号,申请日为2010年9月30日(PCT公开号:W0/2011/041485),和美国实用新型,序列号为13/110685,申请日为2011年5月18日。在本案中使用的词汇可能在这些引用有其来路。本文所描述的一些概念,可能更好地理解在这些参考文献中发现的概念。

[0093] 筛查含有自由浮动胎儿DNA的母体血液

[0094] 本发明所述的方法,也可以用以帮助确定一个孩子,胎儿,或其他目标个体的基因型,所发现的目标的遗传物质存在于在其他的遗传物质中。在一些实施方式中,基因型可能与一个或多个染色体的倍性状态有关,它可以是指一个或多个等位基因链接的疾病,或它们的某种组合。在本发明中,讨论的重点是确定胎儿的遗传状态,其中胎儿的DNA存在于孕妇血液中,但这个例子并不意味着限制可能本方法应用到的背景。此外,本方法可适用于目标DNA的量占其与非目标DNA任何比例的情况,例如,可以使目标DNA的构成可以是任何DNA0.000001到99.999999%之间的数值。外,非目标DNA并不一定需要来自一个个体,或者甚至是来自一个相关的个体,只要求来自部分或全部相关的非目标个体的遗传数据已知的。在一个实施方式中,本发明所公开的方法可以用来确定源自含有胎儿DNA的母体血液中的胎儿的基因型数据。它也可能被用于,在怀孕子宫中有多个胎儿、或样本中可能存在其他污染DNA的情况,例如,源自其他已经出生的兄弟姐妹的DNA。

[0095] 这种技术可以利用胎儿的血细胞通过胎盘绒毛进入母体血液循环的现象。通常情况下,只有极少数的胎儿细胞以这种方式进入母体血液循环(没有足够的胎儿产妇出血产生阳性 Kleihauer-Betke试验)。胎儿细胞可以被整理出来并通过各种技术寻找特定的DNA序列进行分析,但没有侵入性程序本身具有的风险。这种技术也可利用游离的胎儿DNA通过胎盘组织的细胞凋亡释放DNA的方式进入母体循环的现象,其中,胎盘组织中包含与胎儿基因型相同的DNA。在母体血浆中发现游离的DNA已被证明含有胎儿DNA的比例高达30%至40%的。

[0096] 在一个实施方式中,血液可以从孕妇获得。有研究表明,母体血液中除了含有源自母体的自由浮动DNA外可能还含有少量的游离的来自胎儿的DNA。此外,除了通常不包含核DNA母系起源许多血细胞,也有可能存在包括胎儿起源的DNA的摘除的胎儿血细胞。有在本领域中许多已知的方法分离胎儿DNA,或在胎儿DNA新建富集的胎儿DNA分数。例如,色谱方法已显示可创建在胎儿DNA中富集的一定分数的胎儿DNA。

[0097] 一旦一个母体的血液、血浆、或其它流体的样本在手,该样本是以相对非侵入性的方式抽取的,而且样本含一定量的胎儿DNA,无论胎儿DNA是存在于细胞中的或是游离的,也无论是富集于占一定比例的母体DNA中的,或者以其原比例存在,人们就可能对在上述样本中发现的DNA进行基因分型。在一些实施方式中,血液可以用针管从血液回流的静脉中抽取,例如,大殿静脉。本发明所述的方法可以用来确定胎儿的基因型数据。例如,它可以被用于确定在一个或多个染色体的倍性状态,它可以被用于确定一个或一组的SNP多态性位点的特征,包括插入,缺失和易位。它可用于确定一个或多个的单倍型,包括父源的一种或多种基因型的特点。

[0098] 注意,此方法将与可以用于任何基因分型和/或测序方法的核酸一起使用,如INFINIUM ILLUMINA阵列平台,AFFYMETRIX GENECHIP,ILLUMINA GENOME ANALYZER,或LIFE TECHNOLOGIES' SOLID SYSTEM。这包括从血浆或血浆扩增提取的游离的DNA(例如,全基因组扩增,PCR),来自其他细胞类型的基因组DNA(例如来自全血中的人类淋巴细胞)或相同细胞

类型的扩增。对于DNA制备,适于这些平台之一的产生基因组的提取或纯化方法也可以使用。这种方法可以同样出色的用于RNA样本。在一个实施方式中,样本将以最大限度地减少降解的方式存储,(如低于冰点,在约-20℃,或在较低的温度)。

[0099] 家长支持

[0100] 一些实施方式可以结合使用PARENTAL SUPPORT™方法,这些实施方式描述在以下专利中有所描述:美国专利申请号 11/603406 (美国专利公开号:20070184467),美国专利申请号 12/076348 (美国专利公开:20080243398),美国专利申请号 13/110,685,PCT申请 PCT/US09/52730 (PCT公开号: WO/2010/017214) 和PCT申请号PCT/US10/050824 (PCT公开号: WO/2011/041485),上述专利均以引用方式并入本发明。PARENTAL SUPPORT™是一种以信息为基础的方法,可用于分析遗传数据。在一些实施方式中,本文公开的方法可被视为分PARENTAL SUPPORT™方法的一部。在一些实施方式中,PARENTAL SUPPORT™方法是一个可以被用来确定一个目标个体遗传数据的方法的集合,其测量精度高,使用该个体的一个或少量细胞,或包含来自目标个体和一个或多个其他个体DNA的 DNA的混合物,具体而言,以确定与疾病相关的等位基因、其他受关注等位基因,和/或目标个体的一个或多个染色体的倍性状态。亲本间的SUPPORT™可参阅下列任一方法。PARENTAL SUPPORT™是一个以信息为基础的方法的例子。

[0101] PARENTAL SUPPORT™方法利用已知的亲本的遗传数据,即母体和/或父亲的单倍型和/或二倍体遗传数据、连同减数分裂机制的知识和对目标DNA的不完善测量、和可能的一个或多个相关个体,以人口为基础的交叉频率,使用计算机重构在多个等位基因位点的基因型、和/或一个胚胎或任何目标细胞的倍性状态,和具有高置信度的在关键位点位置的目标DNA。PARENTAL SUPPORT™方法不仅可以重建测定结果不理想的SNPs (SNPs),也可以重建插入和删除,以及SNP多态性或未经测量的全部DNA 区域。此外,PARENTAL SUPPORT™方法既可以测量多种与疾病关联的位点也可以从一个单细胞进行非整倍体的筛查。在一些实施方式中,PARENTAL SUPPORT™方法可用来表征来自IVF 周期期间的胚胎活检的一个或多个细胞以确定一个或多个细胞的遗传状态。

[0102] PARENTAL SUPPORT™方法允许噪声遗传数据的清除。这可通过使用相关个体的基因型(父母)作为参考,从而推断在目标基因组(胚胎)中的正确的基因的等位基因。在只提供少量的遗传物质(例如,PGD)和由于遗传物质数量有限而导致直接测量的基因型数据存在固有的噪声情况,使用PARENTAL SUPPORT™方法可以是特别适宜的。在只提供少部分的源自目标个体(e.g.NPD)的遗传物质和用于源自其他个体的污染基因信号导致直接测量的基因型数据存在固有的噪声情况,使用 PARENTAL SUPPORT™方法可以是特别适宜的。PARENTAL SUPPORT™方法能够对胎儿重建高度精确有序的二倍体等位基因序列,加上染色体片段的拷贝数,即使传统的,无序的二倍体测量也可以通过高比率的等位基因舍弃、接收、可变的扩增偏差和其他误差得以表征。该方法可以同时使用潜在的遗传模型和潜在的测量误差模型。遗传模型可确定每个SNP等位基因的概率和两个SNP之间的交叉概率。等位基因概率可以基于从父母获得的数据在每个SNP建模以及基于从如国际人类基因组单体型图计划开发的HapMap数据库获得数据在SNP之间对较差概率建模。给出正确的潜在遗传模型和测量误差模型,可使用最大后验(MAP)估计和计算效率修正,在胚胎中的每个SNP估计正确、有序的等位基因数值。在某些情况下,以上概述的技术,在一个个体给出非常小量的

源自该个体的DNA情况下,能够确定这个个体基因型。这可以从一个或少量细胞的获得的DNA,或者可以从母体血液中的获得的少量的胎儿DNA。

[0103] 定义

[0104] SNPs (SNP) 是指单核苷酸在同一物种两个成员的基因组之间可能会有所不同。术语使用不意味着对每个变体发生频率的任何限制。

[0105] 序列是指一个DNA序列或基因序列。也可能指主要的,DNA 分子的物理结构,或一个单独的链。它可能指在DNA分子发现的核苷酸序列,或与该DNA分子互补链的。它可能指包含在DNA 分子中的信息,如在计算机中用来代表DNA molecule的信息。

[0106] 基因位点是指在一个个体的DNA上所关注的特定区域,这可以是指一个SNP,一个可能的插入或删除位点,或其他一些相关的遗传变异。与疾病相关的SNP多态性也可以是指与疾病相关的位点。

[0107] 多态性等位基因也称“多态性基因位点”是指一个等位基因或基因位点,其中在给定的物种中基因型在不同个体之间变化。多态等位基因的一些例子包括SNPs、短串联重复序列、删除、拷贝和倒置。

[0108] 多态性位点是指不同个体之间变化的多态区域发现的特定核苷酸。

[0109] 等位基因是指占据一个特定的基因位点的基因。

[0110] 遗传数据也称“基因型数据”是指描述一个或多个个体的基因组属性的数据。它可以是指一个或一组位点,部分或整个序列,部分或整个染色体,或整个基因组。它可以是指一个或多个核苷酸的特征,它可以是指一组顺序的核苷酸或基因组中不同的位置的核苷酸,或它们的组合。基因型数据通常是基于计算机上的,然而,它也可以是指在一个序列中以化学编码的遗传数据表示的实际的核苷酸。基因型的数据可能会说成“在...上”“...的”“在...”“从...”或“在个体上”。基因型数据可能指从基因分形平台的测量输出,其中那些测量是对遗传物质进行的。

[0111] 遗传物质也称“遗传样本”是指身体物质,例如来自含有DNA 或RNA的一个或多个个体的组织或血液。

[0112] 有噪声的遗传数据是指以下任何遗传数据:舍弃的等位基因,不确定的碱基对测量,测量不正确的碱基对,缺失的碱基对测量,不确定的插入或删除测量,不确定染色体片段拷贝数的测量,杂散信号,失踪测量,其它错误,或它们的组合。

[0113] 置信度是指统计的可能性,所谓的SNP,等位基因,等位基因组,倍性识别,或正确代表个体实际遗传状态的,确定的染色体片段拷贝数。

[0114] 倍性识别也称“染色体拷贝数识别”或“拷贝数识别”(CNC),可以指确定数量和/或确定存在于细胞中的一个或多个染色体的染色体特征的行为。

[0115] 非整倍体是指错误的染色体数目存在于细胞中的状态。在对人类的体细胞的情况下,它可以是指一个细胞中不包含22对常染色体和一对性染色体的情况。在对一个人类的配子的情况下,它可以是指一个细胞不包含23条染色体中的一条的情况。在对单染色体类型的情况下,它可以是指多于或少于两个同源但没有相同的染色体拷贝存在,或其中有两个源于相同父体的染色体拷贝存在的情况。

[0116] 倍性状态是指在一个细胞中的一个或多个染色体类型的数量和/或染色体特征。

[0117] 染色体是指一个单一的染色体拷贝,意指一个单一的DNA 分子,在一个正常的体

细胞中有46个染色体,一个例子是“源于母体18号染色体”。染色体也可能指染色体的类型,其中在一个正常的人类体细胞中有23个,一个例子是“18号染色体”。

[0118] 染色体特征是指对照染色体的数目,即染色体类型。正常的人类有22种类型的已编号的常染色体和两种类型的性染色体。它也可能指亲本起源的染色体。它也可以是指一个特定的继承自父母的染色体。它也可能指染色体的其他属性特征。

[0119] 遗传材料状态或简称“遗传状态”是指DNA上的一组SNP的特征,遗传物质分阶段的单倍型,以及DNA序列,包括插入、缺失、重复和突变。它也可能指一个或多个染色体的、染色体片段的或染色体片段组的倍性状态。

[0120] 等位遗传数据是指一组涉及一个或多个等位基因的基因型数据的。它可以是指分阶段的单倍型数据。它可以是指SNP的特征,它可以是指DNA序列数据,包括插入、缺失、重复和突变。它可能包括的每个等位基因的亲本源。

[0121] 等位基因状态是指在一组中的一个或多个等位基因中的基因的实际状态。它可以是指等位遗传数据所描述的基因的实际状态。

[0122] 等位基因的比率或等位基因比例,是指,存在于一个样本或一个个体中,在一个基因位点每种等位基因数量之间的比比率。当测定样本通过测序检测时,等位基因的比例可以指映射到基因位点每个等位基因序列读数的比率。当样本通过基于强度测量的方法进行检测时,等位基因比率可以指如该测量方法估计的存在于该基因位点的每个等位基因数量的比率

[0123] 等位基因计数是指映射到一个特定的基因位点的序列的数目,并且如果该位点是多态性的,它是指映射到每个等位基因的序列的数目。如果每个等位基因数以二进制方式的表示,那么等位基因数是整数。如果等位基因以概率计数,那么等位基因计数可以是一个分数。

[0124] 等位基因计数概率是指可能映射到一个特定的基因位点或一组在一个多态性基因位点的等位基因,结合的概率映射的序列的数目。需要注意的是等位基因计数相当于等位基因计数概率,其中对每个计数序列的概率映射是二进制(0或1)的。在一些实施方式中,等位基因计数概率可以是二进制的。在一些实施方式中,等位基因计数概率可以被设置为等于DNA测量结果。

[0125] 等位基因分布,或“等位基因计数分布”是指在一组基因位点的每个基因位点存在的每个等位基因的相对量。等位基因分布可以指一个个体,一个样本,或对样本的一组测量结果。在测序的背景中,等位基因分布是指在一组多态性位点中的每个等位基因中,映射到一个特定的等位基因的读数的数目或可能数目。等位基因测量值以概率的方式进行处理,也就是说,对于给定的以便获取序列读数的等位基因存在的可能性是一个0与1之间的分数,或者它们可以一个二进制的方式处理,也就是说,任何给定读数的被认为恰好是特定等位基因的零个或一个拷贝。

[0126] 等位基因分布模式是指一组对应于不同父母背景的不同等位基因分布。某些等位基因分布模式可能表示一定的倍性状态。

[0127] 等位基因偏差是指杂合基因位点测得的等位基因的比率与出现在DNA初始样本中的比率不同的程度。在某一特定位置等位基因偏差的程度等于在该基因位点观察到的等位基因比率,如所测定的,除以初始DNA样本中该位点的等位基因的比率。等位基因偏差可以

定义为大于1,这样,如果等位基因偏差程度的计算返回一个小于1的值 x ,那么等位基因偏差程度可能重述为 $1/x$ 。等位基因偏差可能由于扩增偏差、纯化偏差、或一些影响不同等位基因的其他的现象而不同。

[0128] 引物,也称为“PCR探针”是指单个DNA分子(一个DNA寡聚物)或DNA分子的集合(多个DNA寡聚物),其中的DNA 分子是相同的,或几乎相同的,并且其中的引物中包含的区域是设计用来是杂交目标多态性位点的,且包含一个底物序列设计允许PCR扩增的。一个引物也可能包含分子条形码。一个引物可能包含一个对每个个体分子不同的随机区域。

[0129] 杂交捕获探针是指任何核酸序列,可能被修饰过,使用的各种方法,如PCR或直接合成产生的,并与一个样本中特定目标 DNA序列的一条链是互补的。外源杂交捕获探针可以被添加到一个准备好的样本并通过一个deanture-再退火过程杂合以生成外源性-内源性片段的双链。这些双链体然后可以在物理上分开的样本通过各种手段。

[0130] 序列读数是指表示使用克隆测序法测定的核苷酸碱基序列的数据。克隆测序可能会产生代表一个原始的DNA分子的单个、或克隆,或簇的序列数据。一个序列读数在序列的每个碱基位置也可能有相关的质量分数,其表示核苷酸被正确调用的概率。

[0131] 映射序列读数是在某一特定生物体的基因组序列中,确定序列读数的源的位置的过程。序列读数的源的位置是基于该序列读数的碱基序列和基因组序列的相似性确定的。

[0132] 匹配拷贝错误,也称为“匹配的染色体非整倍体”(MCA),是指一种非整倍体的状态,其中一个细胞中包含两个相同或几乎相同的染色体。这种类型的非整倍体可能在减数分裂期间配子形成的过程中出现,并且可以被称为减数分裂不分离错误。这种类型的错误可能会在有丝分裂中出现。匹配三倍体可以是指,可以是指一个给定染色体的三个拷贝出现的情况,并且其中的两个拷贝相同的。

[0133] 非匹配拷贝错误,也称“独特的染色体非整倍体”(UCA),指的是一种非整倍性状态,其中一个细胞中含有来自同一亲本的两个染色体,这可以是同源的,但不完全相同。这种类型的非整倍体可能出现在减数分裂过程中,可以被称为减数分裂错误。非匹配的三倍体可以是指一个给定染色体的三个拷贝出现在一个个体中的情况,并且其中的两个拷贝来自同一亲本,并且同源但不完全相同,请注意,不匹配的三倍体,可以是指来自一个亲本的两个同源染色体出现的情况,并且其中的某些染色体片段是完全相同的而另一些片段仅仅是同源。

[0134] 同源染色体是指含有相同组的基因的染色体拷贝,其中的基因通常在减数分裂过程中是配对的。

[0135] 相同的染色体是指含有相同的一组基因的染色体拷贝,并且对每个基因,它们含有的相同等位基因组是相同的,或几乎是相同的。

[0136] 等位基因脱失(ADO)是指一个给定等位基因的同源染色体的一组碱基对中的至少一个碱基对未检测到情况。

[0137] 基因位点脱失(LDO)是指一个给定等位基因的同源染色体的一组碱基对中的两个碱基对未检测到的情况。

[0138] 纯合子是指具有类似的等位基因作为相应的染色体位点。

[0139] 杂合子是指具有不同的等位基因作为相应的染色体位点。

[0140] 杂合率是指在在一个给定基因位点具有杂合等位基因的人群中的个体比率。杂合

性率也可以是指,在一个个体中的给定的基因位点中的,或一个DNA样本中的,期望或测得等位基因的比率。

[0141] 高信息量的SNPs (HISNP) 是指一个SNP,其中胎儿有一个不存在于母体基因型中的等位基因。

[0142] 染色体区域是指染色体片段,或一个完整的染色体。

[0143] 染色体片段是指可以以大小从一个碱基对整个染色体进行排列的染色体的一段。

[0144] 染色体是指无论一个完整的染色体或一个染色体的段或一个染色体的部分。

[0145] 拷贝是指染色体片段的拷贝数量。它可以是指染色体片段的相同拷贝,或染色体片段不相同的、同源的拷贝,其中,该染色体片不同的拷贝段含有基本上类似的一组基因位点,其中一个或多个等位基因是不同的。请注意,在非整倍体的某些情况下,如 M2拷贝错误,有可能有一些给定的染色体片段的拷贝是相同,以及一些相同染色体片段的拷贝是不相同的。

[0146] 单倍型是指通常是被一起继承到位于同一染色体上多个位点上的等位基因的组合。单倍型可以是指尽可能少的两个位点,或依赖于给定基因位点之间的已发生重组事件数量的整个染色体。单倍型也可以指在一个单一的统计学关联的染色单体上的一组SNPs (SNPs)。

[0147] 单倍型数据也称“相控数据”或“有序遗传数据”指的是来自一个二倍体或多倍体基因组的单个染色体的数据,即,在二倍体基因组中,或是隔离母系的亦或是父系的染色体的拷贝。

[0148] 调相是指确定一个个体的无序的单倍型遗传数据、二倍体 (或多倍体) 遗传数据的行为。对在一个染色体上发现的一组等位基,它可能指确定一个等位基因上的两个基因那个是与一个个体的两个同源染色体的中的每一个都相关的行为。

[0149] 阶段性数据是指一个或多个单倍型已经确定的遗传数据。

[0150] 假设指在一个给定的染色体上的可能的倍性状态,或在给定位点上的一组可能的等位基因状态。该组可能性可以包括一个或多个元素。

[0151] 拷贝数假设也称为“倍性状态假设”指一种涉及在一个个体中一个染色体拷贝数量的假设。也可能指涉及每个染色体特征的假设,包括每个染色体的源的亲本,并且其亲本的两个染色体出现在该个体中。如果有的话,它也可能指涉及来自一个相关个体的染色体或染色体片段与来自一个个体的给定的染色体基因对应的假设。

[0152] 目标个体是指遗传状态正在确定的个体。在一些实施方式中,只有有限数量来自目标个体的DNA。在一些实施方式中,目标个体是一个胎儿。在一些实施方式中,可能有一个以上的目标个体。在一些实施方式中,每个源自一对亲本的胎儿可能被认为是目标个体。在一些实施方式中,正在确定的遗传数据是一个或一组等位基因调用。在一些实施方式中,正在确定的遗传数据是一个倍性调用。

[0153] 相关个体是指任何个体,其与目标个体是基因相关的,因而共享单倍型块。在一个背景中,相关的个体可以是一个目标个体的基因亲本,或者任何源自一个亲本的遗传物质,如精子、极体、胚胎、胎儿、或儿童。它也可能指兄弟姐妹,父母或祖父母。

[0154] 兄弟姐妹是指任何个体,在所考虑的问题中,其遗传父母是相同的个体。在一些实施方式中,它可以是指一个初生的孩子,一个胚胎,或胎儿,或源自一个初生的孩子、一个胚

胎、或胎儿的一个或多个细胞。兄弟姐妹也可以是指源自一个亲本的单倍体个体,如精子、极体,或任何其他组的单倍型遗传物质。一个个体本身可被视为一个兄弟姐妹。

[0155] 胎儿指“胎儿的”或“基因与胎儿相似的胎盘区域的”。在孕妇中,胎盘的某些部分是与胎儿的基因相似的,孕妇血液中发现的自由浮动胎儿DNA可能起源于胎盘上的与胎儿基因型相匹配的这部分。需要注意的是在胎儿中染色体一半的遗传信息是继承自胎儿的母体。在一些实施方式中,从这些母系遗传的来自胎儿细胞的染色体的DNA被认为是“胎儿起源的”,而不是“母系起源的”。

[0156] 胎儿源DNA是指原是细胞一部分的DNA,该细胞的基因型基本上相当于胎儿的基因型。

[0157] 母源DNA是指原本是细胞一部分的DNA,该细胞的基因型基本上相当于母体基因型。

[0158] 孩子可能指一个胚胎、一个卵裂球、或一个胎儿。请注意,在本发明公开的实施方式中,所描述的概念也同样适用于一个个体,该个体是出生的孩子、一个胎儿、一个胚胎或其中的一组细胞。术语孩子的使用可能只是意味着个体是指孩子是父母的遗传后代。

[0159] 亲本是指一个体的父亲或母体的遗传。一个个体通常有两个亲本,一个母体和一个父亲,虽然这不一定是这种情况,如在基因或染色体嵌合体的情况下。亲本可能会被认为是单个的。

[0160] 父母背景是指在源自目标的一个或两个两个亲本的每个相关的两个染色体上,一个给定的SNP遗传状态。

[0161] 预期发育,也称“正常发育”,是指一个可行的胚胎植入子宫并导致怀孕,和/或怀孕继续和导致活产,和/或一个出生的无染色体异常的孩子,和/或一个出生的无不期望的其他遗传状况,如有与疾病相关的基因,的孩子。术语“”是指包括父母和保健促进者期望的任何情况。在某些情况下,“预期发育”可以指用于医学研究或其他目的不可行的或可行的胚胎。

[0162] 插入子宫是指在体外受精背景下胚胎转移到子宫腔的过程。

[0163] 孕妇血浆是指源自孕妇的血液的血浆部分。

[0164] 临床决定是指任何采取或不采取行动的决定,其结果个体的健康和生存。在产前诊断的背景下,临床决定可能指流产或不流产胎儿的决定。也可以指一个进一步的测试、采取行动以减轻不良型的决定,或采取行动以备伴随异常的孩子出生的临床决定。

[0165] 诊断盒是指一台机器或一个机器的组合,设计成执行本文所公开的方法的一个或多个方面。在一个实施方式中,诊断盒可放置在一个病人护理点。在一个实施方式中,诊断盒可以执行随后测序的目标扩增。在一个实施方式中,诊断盒可以单独使用或在一个技术人员的帮助下运作。

[0166] 基于信息的方法是指在很大程度上依赖于统计意义上的一个大数据量的方法。在产前诊断的背景中,是指设计的用来确定一个或多个染色体的倍性状态,或通过统计推断确定一个或多个等位基因的等位基因状态最可能的状态的一种方法,而不是通过直接物理测量状态的方法,给出大量的遗传数据,例如,从一个分子阵列或序列中。在本发明的一个实施方式中,信息为基础的技术可能在本发明中公开。在本发明的一个实施方式中,它可能 PARENTAL SUPPORT™。

[0167] 主遗传数据是指由基因分型平台输出的模拟强度信号。在 SNP阵列的背景下,主遗传数据指进行任何基因型调用之前的强度信号。在测序的背景下,主遗传数据是指在任何碱对特征被确定之前,以及在序列被映射到基因组之前,来自定序器的模拟测量结果、类似的色谱图。

[0168] 二次遗传数据是指通过基因分型平台输出的处理过的遗传数据。在一个SNP阵列的背景中,二次遗传数据是指由软件与相关的SNP阵列阅读器对等位基因的识别,其中,无论样本中是否存在给定的等位基因,该软件都已经给出了一个调用。在测序背景下,二次遗传数据,指序列的碱基对特征已被确定,也可能该序列已被映射到基因组中。

[0169] 非侵入性产前诊断(NPD),或也称“非侵入性产前筛查”(NPS),是指是使用母体血液中发现的遗传物质确定母体孕育的胎儿的遗传状态的方法,其中,遗传物质是通过抽取母体的静脉血液的方式获得的。

[0170] 对应于一个基因位点的DNA的优先富集,或在一个位点的 DNA的优先富集,是指任何方法所导致使富集后的DNA混合物中对应基因位点的DNA分子的百分比高于富集前的DNA的混合物中的对应基因位点的DNA分子中的方法。该方法可以包括对应基因位点的DNA分子的选择性扩增。该方法可以包括除去与基因位点不对应的DNA分子。该方法可以包括多种方法的结合。富集度被定义为富集后的DNA混合物中对应基因位点的DNA分子的百分比除以富集前的DNA的混合物中的对应基因位点的 DNA分子的百分比。优先富集也可以在多个基因位点进行。在本发明的一些实施方式中,富集的程度大于20。在本发明的一些实施方式中,富集的程度大于200。在本发明的一些实施方式中,富集度大于2000。当优先富集在多个基因位点进行,富集度可指在一组基因位点中的所有位点的富集度的平均富集度。

[0171] 扩增是指一种增加DNA分子的拷贝数目方法。

[0172] 选择性扩增可以指一种方法,用于增加一个特定DNA分子、或对应于一个特定DNA区域的DNA分子的拷贝的数目。它也可以指一种方法,用于增加一个特定目标DNA分子、或目标DNA 区域而不增加非目标DNA分子或DNA区域拷贝的数目。选择性扩增可以是一种优先富集的方法。

[0173] 通用底物序列是指可附加到一个目标DNA分子群体的DNA 序列,例如通过连接、PCR或连接介导的PCR。一旦附加到目标分子的群体,使用一个单一的扩增引物对,可用特别针对通用底物序列的引物扩增目标群体。通用底物序列通常与目标序是不相关的。

[0174] 通用适配子或“连接适配子”或“库标记”是DNA分子中含有通用底物序列,可以被共价连接到目标双链DNA分子的群体的 5端和3端。此外适配子向发生PCR扩增的目标群体的5端和3 端提供通用底物序列,使用一个单一的扩增引物对,扩增源自目标群体的所有分子。

[0175] 定位指在一个DNA混合物中,用于选择性地扩增或以其他方式优先富集那些与一组位点对应的DNA分子的一种方法。

[0176] 联合分布模型是指一个定义事件概率的模型,而事件定义为多个随机变量,给定的多个随机变量定义在相同的概率空间上,其中的变量的概率是相互关联的。在一些实施方式中,在变量概率不相关联的退化情况下未链接也可使用。

[0177] 假设

[0178] 在本发明的背景中,一个假设是指一个可能的遗传状态。它可以是指一个可能的

倍性状态。它可以是指一个可能的等位基因状态。一组假设可以是指一组可能的遗传状态，一组可能的等位基因的状态，一组可能的倍性状态，或它们组合的。在一些实施方式中，假设一组可以被设计成，使得该组的一个假设将对应于任何给定个体的实际遗传状态。在一些实施方式中，的一组假设可以这样设计，每一个可能的基因状态可被改组中的至少一个假设描述。在本发明的一些实施方式中，在所关注的问题中，一种方法的一个方面是，确定哪个假设对应于有关个体的实际遗传状态。

[0179] 在本发明的另一个实施方式中，一个步骤涉及到创建一个假设。在一些实施方式中，它可以是一个拷贝数假设。在一些实施方式中，它可能包含一个涉及假设，每一个个体的哪一个染色体片段一般对应于其他个体的，如果有的话，哪一个片段。创建一个假设可以是指设置变量的限制的行为，使得这些变量包含整组考虑中的可能的遗传状态。

[0180] 一个“拷贝数的假设，”也被称为“倍性假设”，或“倍性状态假设，”可以是指在目标个体中，一个给定的有关染色体拷贝、染色体类型、或染色体部分的可能倍性状态的假设。它也可能指在个体中的一个以上染色体类型的倍性状态。一组拷贝数假设可以是指一组假设，其中每个假设对应一个个体中不同可能的倍性状态。一组假设可能涉及一组可能的倍性状态、一组可能的亲本单倍型的贡献，混合样本中一组可能的胎儿DNA的百分比，或它们的组合。

[0181] 一个正常的个体，包含分别来自父母的每条染色体的类型之一。然而，由于在减数分裂和有丝分裂中的错误，一个个体可能可能有0, 1, 2, 或多个来自父母双方的一个给定的染色体类型。在实践中，很少见到多于2个来自一个亲本的给定染色体的情况。在本发明中，一些实施方式中只考虑了来自一个亲本的0, 1, 或 2份拷贝给定染色体可能的假设，考虑更多或更少源自亲本的可能拷贝只是一个简单的延伸。在一些实施方式中，对于一个给定的染色体，有9个可能的假设：三个涉及源自母体的0, 1, 或2 染色体的可能的假设乘以三个涉及源自父体的0, 1, 或2染色体的可能的假设。设 (M, F) 代表这种加深，其中m是一个给定的染色体遗传自母体，f是一个给定的染色体遗传自父亲。因此，九中假假设是 (00), (0, 1), (0, 2), (10), (1, 1), (1, 2), (20), (2, 1), (2, 2)。这些也可写为H00, H01, H02, H10, H12, H20, H21, 和H22。不同的假设对应不同的倍性状态。例如，(1, 1) 指的是一个正常的二体染色体，(2, 1) 是指一个母体的三倍体，以及 (0, 1) 是指一个父体的单倍体。在一些实施方式中，两个染色体继承自一个亲本和一个染色体继承自另一个亲本的情况下从一个家长和一个染色体继承自其他家长可能进一步分化两种情况：一个情况是两条染色体是相同的（匹配拷贝错误），一个情况是两条染色体是同源的，但不相同的（非匹配拷贝错误）。在这些实施方式中，有16个可能的假设。但应当理解，使用其他不同假设数目的其他假设也是可能的。

[0182] 在本发明的一些实施方式中，倍性是指一个假设源自其他相关的个体染色体对应于在目标个体的基因组中发现的染色体的假设。在一些实施方式中，该方法的一个关键是相关个体共享单倍型块、使用从相关个体测量的遗传数据、以及单倍型块在目标个体和相关个体之间匹配知识的事实，对一个目标个体推断的正确的遗传数据比单独使用目标个体的遗传的测量结果具有较高置信是可能的。因此，在一些实施方式中，倍性假设可能不仅涉及染色体的数目，而且相关个体的染色体与目标个体中的其他的一个或多个染色体也可以是相同、或几乎相同的。

[0183] 一旦定义一组假设，当算法在输入的遗传数据操作时，它们可能会为每个所考虑

的假设输出确定的统计概率。使用相关的遗传数据作为输入,对每一个假设,可通过数学计算,概率值、如所指出的一个或多个专业技术、算法、和/或在本发明的其他地方描述的方法确定各种假设的概率。

[0184] 一旦对不同假设的概率做出估计,如通过多种技术所确定的,它们可结合起来。对每一种假设,这可能意味着,把每种技术确定的概率相乘。对假设的概率结果可能进行归一化处理。请注意,一个倍性假设是指一个可能的染色体倍性状态。

[0185] “合并概率“,也被称为”结合假设“或结合专业技术结果的过程,应该是对本领域的技术人员熟悉的在现有技术中的线性代数概念。合并概率的一个可能的方法如下:当一个专业技术用于评估一组给定的一组遗传数据的一组假设时,该方法的输出是以一对一方式与每一个假设相关联的一组概率。当一组概率被第一个专业技术确定时,其中的每个概率与假设集中的每一个假设相关联,与第二个专业技术确定的一组概率相结合,与,其中每一个都与相同的一组假设集中的每一个假设相关联,然后将这两个集的概率相乘。这意味着,对于集合中的每一个假设,将这两个如两个专家方法所确定的与该假设相关联的概率乘得到对应的输出概率结果。该过程可以扩展到任何数量的专业技术。如果只使用一个专业技术,那么输入概率和输出概率是相同的。如果使用两个以上的专业技术,那么将相关概率在同时相乘。加过可能被规范化处理,这样假设集合中的概率总和为100%。

[0186] 在一些实施方式中,如果一个给定假设的合并概率大于其他任何假设的合并概率,那么这一假设可能被确定为最有可能的假设。在一些实施方式中,一个假设如果归一化的概率大于阈值,可被确定为是最有可能的,可以称为倍性状态或其他遗传状态。在一个实施方式中,这可能意味着,与该假设相关染色体的数目及特征可被称为倍性状态。在一个实施方式中,这可能意味着与该假设关联的等位基因的特征可以被称为等位基因状态。在一些实施方式中,阈值可以是约50%和约80%之间。在一些实施方式中,阈值可以是在大约80%和大约90%之间。在一些实施方式中,阈值可以是约90%和约95%之间。在一些实施方式中,阈值可以是大约在95%和大约99%。在一些实施方式中,阈值可以是大约在99%和大约在99.9%之间。在一些实施方式中,阈值可以在大约99.9%以上。

[0187] 亲本背景

[0188] 亲本背景是指一个给定的等位基因的遗传状态,相关的两个染色体每一个是对目标的一个或两个两个本的。请注意,在一个实施方式中,亲本背景是指指等位基因状态目标的,更确切地说,它是指亲本的等位基因状态。对于一个给定的SNP的亲本背景父母背景可含有4个碱基对,两个是父体的和两个是母体的,它们可以是彼此相同的或不同的。通常写为“M1M2|F1F2”,其中m1和m2是给定SNP的两个母体染色体上的遗传状态,f1和f2是给定的SNP的两个父体的染色体上的遗传状态。在一些实施方式中,亲本背景可写为“F1F2|M1M2”。请注意,下标“1”和“2”是指在给定的等位基因上第一和第二的染色体的基因型,并注意,选择那个染色体比被标记为“1”,及那个染色体被标记为“2”是任意的。

[0189] 请注意,本发明中,A和B一般用来代表碱基对身份,A或B同样可以很好地代表C(胞嘧啶),G(鸟嘌呤),A(腺嘌呤),T(胸腺嘧啶)。例如,如果在一个基给定SNPs(SNP)的等位基因上,母亲一条染色体上针对那个SNP的基因型是T,同源染色体上是G,父亲两条同源染色体上针对那个SNP的基因型是G,可以说,目标个体等位基因的父母状况是AB|BB,也可以说,该等位基因有AB|AA的父母状况。需要注意的是,从理论上讲,任何四种可能的核苷

酸都可能出现在给定的等位基因上,因此,比如母亲有基因型AT,父亲在给定的等位基因上可能有基因型 GC。然而,经验数据表明,在大多数情况下,在一个给定的等位基因上仅观察到四种碱基对中的两种。例如,当使用单一的串联重复序列时,可能会有超过2种,超过4种,甚至超过10中父母状况。本发明中所讨论的内容,假定在一个给定的等位基因中只有两种可能的碱基对被观察到,虽然通过修改所公开的实施方式,可以将此假设不成立的情况考虑在内。

[0190] A“父母状况”可以是指一组或子集具有相同父母状况的目标SNP。例如,如果一个人测定一个给定目标个体的一个指定染色体上的1000个等位基因,那么AA|BB状况可以是指1000个等位基因组中的全套基因,在此,目标母亲的基因型是纯合的,目标父亲的基因型是纯合的,但是产妇基因型和父亲基因型在该位点上并不相同。如果父母的数据不分段,从而 $AB=BA$,则有九种可能的父母状况:AA|AA,AA|AB,AA|BB,AB|AA,AB|AB,AB|BB,BB|AA,BB|AB,BB|BB。如果父母的数据分段,因此 $AB \neq BA$,则可能有16种不同的父母背景:AA|AA,AA|AB,AA|BA,AA|BB,AB|AA,AB|AB,AB|BA,AB|BB,BA|AA,BA|AB,BA|BA,BA|BB,BB|AA,BB|AB,BB|BA和BB|BB。一方式杂合的父母状况的SNP组可能要参照杂合的状况。

[0191] 非侵入性产前诊断中使用父母状况

[0192] 非侵入性产前诊断是一个重要的技术,借助非侵入方式,通过遗传物质确定胎儿的遗传状态的是一种非侵入性的方式中得到的,例如可以对怀孕母亲抽血。血液被分离,得到单独的血浆,然后分离出血浆的DNA。通过大小选择方法得到适当长度的DNA。一组基因位点的DNA可以被优先富集。这种DNA可以通过一系列方式加以测定,如通过基因分型阵列杂交,测定荧光,或通过高通量测序仪进行测序。

[0193] 在非侵入性产前诊断的情况中,当测序用于胎儿的倍性识别时,有许多方法用来处理测序数据。人们用来处理测序数据的最常用方法是,简单地数一数读取映射到一个给定的染色体的数量。例如,设想一下,你正试图测定胎儿21号染色体的倍性水平。进一步设想,DNA样本中包括10%胎儿来源的DNA以及90%母亲来源的DNA。在这种情况下,可以看染色体上的平均读取次数(用来预测二倍体),例如3号染色体,并将之与21号染色体上的读取次数进行对比,在此,读取次数用来校正染色体上具有部分专一序列的碱基对数量。如果胎儿是整倍体,会预期每单位基因组上DNA的量在所有的位置(可随机变化)上都是相等的。另一方面,如果胎儿在21号染色体上市三倍体,就会预期在21号染色体上每个遗传单位会有比基因组其他位置上稍多一些DNA。尤其是在混合物中会多出来自21号染色体的约5%的DNA。当测序法被用来测量DNA时,可以预期会有因21号染色体有别于其他染色体的特有片段多处5%的专一读图次数。人们可以观察来自一个特定染色体的高于一般阈值的DNA数量,并与该染色体专一读图序列数量进行校正,作为异倍性诊断的基础。另一种用来检测异倍性的方法与上述方法类似,所不同的是将父母状况考虑在内。

[0194] 若考虑哪些等位基因作为目标,需要考虑那些比其他父母状况更有用的父母状况。例如,AA|BB和对称的状况BB|AA是最有用的状况,因为胎儿已知携带了有别于母体的等位基因。由于对称性的原因,AA|BB和BB|AA两种状况都可简称为AA|BB。另一组有用的父母状况是AA|AB和BB|AB,因为在这些情况下,胎儿有50%的机会携带一个母亲没有的等位基因。由于对称性的原因,无论是AA|AB,还是BB|AB,都可以简称为AA|AB。第三组有用的父母状况是AB|AA和AB|BB的,因为在这些情况下,胎儿携带已知的父亲等位基因,该等位基因在

产妇的基因组中也存在。由于对称性的原因,AB|AA和AB|BB状况可以被称为 AB|AA。第四组有用的父母状况是AB|AB,此时胎儿有一个未知的等位基因状态,无论等位基因是什么状态,该基因在母亲体内有相同的等位基因。第五个有用的父母状况是AA|AA,此时父亲和母体是杂合的。

[0195] 目前公开的实施方式的不同实现

[0196] 这里公开的方法用于确定目标个体的倍性水平。目标个体,可以是一个卵裂球,胚胎或胎儿。在本发明的一些实施方式中,一种用于确定目标个体中的一个或多个染色体的倍性水平的方法,可包含本发明档中描述的任意步骤,以及它们的组合:

[0197] 在一些实施方式中,用于确定胎儿遗传状态的遗传物质的来源可以是胎儿细胞,比如从母亲血液中分离得到的成核胎儿的血血红细胞。该方法可能涉及一个从孕妇获得的学业样本。该方法可能涉及运用可视化技术分离胎儿的血血红细胞,这种可视化技术基于以下思想:某一特定的颜色组合与成核血血红细胞唯一关联,一种类似的颜色组合与母亲血液中任何其他细胞不关联。与成核胎儿血血红细胞相关的的研的组合包括围绕核的蛋白的红色,这种颜色可能由不同的颜色造成,并且和物质的颜色可以染成,例如蓝色。通过从母体血液中分离出细胞并在平板上展开,然后确定这些同时看到红色(血红蛋白)和蓝色的点(核材料)从而能够识别有核血红细胞的位置。然后用显微操作器提取这些核血血红细胞,使用基因分型和/或测序技术来测量在那些细胞中的基因材料的基因型特征。

[0198] 在一个实施方式中,人们可用一种只对胎儿血红蛋白发出荧光而不对母体血红蛋白发出荧光的染料对核血血红细胞染色,这样,可剔除一个核血血红细胞染色源自母体或胎儿之间的不确定性。本发明的一些实施方式中可能涉及染色或以其他方式标记核材料。本发明的一些实施方式中可能涉及使用胎儿细胞的特异性抗体专门标记胎儿的核材料。

[0199] 还有许多其他的方法从母体血液中分离胎儿细胞,或从母体的血液中分离胎儿DNA,或,富集存在于母体中的遗传物质的胎儿遗传物质的样本。这里列出了其中的一些方法,但是这里并不打算给出一个详尽的列表。为方便起见,这里列出了一些适适用的技术:使用荧光或其他标记的抗体、尺寸排阻色谱法、磁或其他标记的亲亲和标签、表观遗传差异如在特定等位基因上母体和胎儿细胞之间的差异性甲基化、通过CD45/14耗损及从CD45/14阴性细胞及进行CD71阳性选择实现的密度梯度离心、单或双 Percoll梯度和渗透压、或半乳糖凝集素方法。

[0200] 在本发明的一个实施方式中,目标个体是一个胎儿,并且不同的基因型测量结果是由源自胎儿的多个DNA样本得出的。在本发明的一些实施方式中,胎儿的DNA样本是从分离的胎儿细胞的到的,其中胎儿细胞可能与母体细胞混合在一起。在本发明的一些实施方式中,胎儿的DNA样本来自游离的胎儿DNA,其中的胎儿DNA可能自由浮动母体DNA混合在一起。在一些实施方式中,胎儿的DNA可能衍生自含有附体DNA和胎儿DNA的母体血浆或母体血液。在一些实施方式中,胎儿DNA可能与母体DNA以母体:胎儿为99.9:0.1%~99.1%;99.1%~90.10%;90.10%~80.20%;80.20%~70.30%,70.30%~50.50%;50.50%~10.90%;或10.90%~1.99%;1.99%~0.1:99.9%的比例范围混合。

[0201] 在一些实施方式中,基因样本来制备过的和/或纯化过的。在本领域有很多已知的标准方法来完成这样的工作。在一些实施方式中,样本可能离心方法被分离成多个层。在一些实施方式中,该DNA可使用过滤方法分离。在一些实施方式中,DNA制备的可能涉及扩增、

分离、通过色谱法纯化、液-液分离、分离、优先富集、优先扩增、目标扩增,或任何数量的本领域已知的或本文所描述的其他技术。

[0202] 在一些实施方式中,本发明所公开的方法可能涉及扩增 DNA。DNA的扩增是一个可以通过各种各样的方法,将少量遗传物质转换成含有遗传数据的大量遗传物质的过程,包括但不限于聚合酶链反应; (PCR)。扩增DNA的方法之一是全基因组扩增(WGA)。有很多的方法可用于WGA:连接介导的PCR (LM-PCR),退化寡核苷酸引物PCR (DOP-PCR),多重置换扩增(MDA)。在LM-PCR中,称为适配子的短DNA序列连接到 DNA的两个钝端。这些适配子包含用于通过PCR扩增DNA的普遍扩增序列。在DOP-PCR法中,随机引物也包含普遍扩增序列用于第一轮退火和PCR。然后,进行第二轮PCR进一步用通用引物序列扩增该序列。MDA使用phi-29聚合酶,这是已用于单细胞分析的、一个高加工性和非特异性的用于DNA拷贝的酶。从一个单细胞扩增物质的主要限制是:(1) 必须使用极稀 DNA浓度或非常小的体积的反应混合物,和(2) 可靠地在整个基因组从蛋白质解离DNA的难度。无论如何,单细胞的全基因组扩增成功的用于各种发明已有数十年时间。还有从DNA样本中扩增DNA的其他方法。DNA扩增将初始DNA样本转换成该序列族中有相似DNA样本A,但数量更大。在某些情况下,可能并不需要扩增。

[0203] 在一些实施方式中,DNA可以使用普遍扩增方法进行扩增,如WGA或MDA扩增。在一些实施方式中,DNA可通过定向扩增法扩增,例如,使用定向PCR,或圆形探针。在一些实施方式中,DNA可能被优先富集,用靶向扩增法或导致从不希望的 DNA全部或部分分离出希望的DNA的方法,例如杂合法捕获。在一些实施方式中,DNA可以被扩增,通过使用一个通用的扩增和优先富集的组合方法进行扩增。这些方法的一些更全面的描述,可以在本文档中的其他地方发现。

[0204] 目标个体和/或相关的个体的遗传数据可以从分子状态转换为电子状态,通过适当的遗传物质使用选择一组工具和技术的方法,包括但不限于这些方法:通过测量使用的工:基因分型微阵列、高通量测序。一些高通量测序方法包括桑格DNA测序、焦磷酸测序ILLUMINA Solexa技术平台、Illumina公司的基因组分析仪,或APPLIED BIOSYSTEM's 454测序平台,HELICOS's TRUE SINGLE MOLECULE SEQUENCING平台,HALCYON MOLECULAR的分子电子显微镜测序方法,或其他任何测序方法。所有这些方法将储存在基因样本中的遗传数据完全转换成一套通常储存在一个记忆装置中,并在程序中进行处理。

[0205] 一个有关个体的遗传数据可以通过分析一组物质得以测量,其中的物质包括但不限于:个体的散装二倍体组织,个体的一个或多个二倍体细胞,个体的一个或多个单倍体细胞、目标个体的一个或多个卵裂球、从个体找到的细胞外遗传物质,从母体血液中发现找到的个体的细胞外遗传物质、从母体血液中发现的个体的细胞、从相关个体产生于配子的一个或多个胚胎、一个一个或多个取自胚胎的卵裂球、从相个体发现的细胞外的遗传物质、已知源自相关个体的遗传物质,以及它们的组合。

[0206] 在一些实施方式中,为所关注目标个体的每个染色体类型,可能会创建一组至少一个倍性状态假设。每一个倍性状态假设可以是指一个可能的目标个体的染色体或染色体片段的倍性状态。该组假设可能包括在目标个体的染色体中预期存在部分或全部可能的倍性状态。一些可能的倍性状态可能包括缺体、单倍体、二倍体、单亲二倍体、正倍体、三倍体、匹配三倍体、不匹配的三倍体,母体三倍体、父体三倍体,四被踢体,平衡四倍体(2:2)、不平衡(3:1)四倍体、六倍体、五倍体、其他非整倍体,以及它们的组合。这些非整倍体的任

何状态可以是混合的或部分非整倍体,如不平衡易位、平衡易位、罗伯逊易位、重组、删除、插入、交叉、以及它们的组合。

[0207] 在一些实施方式中,所确定的倍性状态的知识可用于作出临床决定。这方面的知识,通常以物理排列的方式被存储在一个记忆装置中,然后可被转化成一份报告。然后该报告可能会为采取行动的依据。例如,可以是终止妊娠的临床决定;或者,可以是继续妊娠的临床决定。在一些实施方式中,临床决定可能涉设计用来减少严重的一种表遗传性疾病的风险一种干预,或一个采取相应的步骤的决定,为一个特殊需要儿童做准备。

[0208] 在本发明的一个实施方式,本文所描述的任何一种方法,可能会被修改,以允许用于来自相同目标个体的多个目标,例如,从同一个孕妇抽取的多个血样。这可能会提高模型的准确度,多个基因测量可能提供更多的数据,以此来确定该目标基因型。在一个实施方式中,一组目标遗传数据用作所报告的主要数据,另一组用作核实该主要目标遗传数据。在一个实施方式中,有多组遗传数据,其中每个数据都是从取自目标个体的遗传物质所测得,这些数据被认为是平行的,因此,这两组遗传数据用于帮助确定具有高测量准确性的亲本的遗传数据中,那一部分数据构成胎儿的基因组。

[0209] 在一个实施方式中,该方法可以用于亲子鉴定的目的。例如,给定来自母体的基于SNP的基因型信息,而来自一个男人的基因型信息可以是或可能不是父亲的遗传,,如果男性的基因型信息的确代表实际的孕期胎儿父亲的遗传,有可能确定从混合的样本测得的基因型信息。做到这一点简单的方法是简单地看一下母体是AA是的情况下,可能的父亲是AB还是BB。在这种情况下,人们可能期望看到这时父亲分别贡献一半(AA|AB)还是全部(AA|BB)。考虑到预期的ADO,可直接确定所观察到胎儿 SNP多态性是否与父亲那些SNP多态性相关。

[0210] 本发明的一个实施方式可能如下:一个孕妇,想知道如果她的胎儿患有唐氏综合症,和/或如果会患囊性纤维化,她不希望承担患有这些及疾病之一一个孩子。医生取得她的血液,并用一种标志物对血红蛋白染色使其出现一个清晰的红色,用量一种标志物染色核物质使其出现明显蓝色。已知产妇的血血红细胞是典型的无核细胞,而胎儿细胞含有细胞核的比例很高,通过识别同时显示红色和蓝色的颜色,医生是凭视觉分离一些有核血血红细胞。医生用显微操作器从载玻片上拾取这些细胞,并且将它们到实验室对10个单个细胞进行扩增和基因分型。通过使用基因测量,PARENTAL SUPPORT™方法能够确定10个细胞中的6个是母体血液细胞,PARENTAL SUPPORT™方法也能够确定10个细胞中的4个是胎儿细胞。如果一个孩子已经从出生孕妇,PARENTAL SUPPORT™方法也可以用来确定胎儿细胞是不同于出生的孩子的细胞,通过对胎儿细胞的可靠的等位基因分识别,它们与已出生的孩子的细胞是不相似的。注意,此方法是在概念上类似于本发明的父系测试实施方式。由于对单个细胞的基因分型的困难,从胎儿细胞测得的遗传数据可能质量很差,包括许多等位基因丢失。临床医生可以使用测得的胎儿DNA和可靠的亲本的DNA测量结果,用PARENTAL SUPPORT™方法来准确地推断胎儿的基因组的特征,从而将来自胎儿的遗传物质上含有的遗传数据转化成预测的存储在电脑上的胎儿遗传状态。临床医生对胎儿的倍性状态,以及对存在或不存在所关注的与多个疾病相关的基因都能够确定。事实证明,胎儿是整倍体,而不是囊性纤维化携带者,而母体决定继续妊娠。

[0211] 在本发明的一个实施方式中,孕妇希望确定她的胎儿是否患有任何整个染色体异

常。她去找她的医生,并给出了她的血液样本,她和她的丈夫从面颊拭子上给出了他们自己的DNA样本。一位实验室研究员使用MDA协议扩增亲本的DNA,并对亲本 DNA基因分型,并用ILLUMINA INFINIUM阵列在大量SNP上测量亲本遗传数据。然后,研究人员振荡血样,取出血浆,使用尺寸排阻色谱法分离游离的DNA的样本。接着,研究人员使用的一个或多个荧光抗体,例如一个专门对胎儿血红蛋白的荧光抗体分离胎儿的有核红血细胞。然后研究人员的取出已分离的或经过富集的胎儿遗传物质,并使用适当地设的70-聚体寡核苷酸库对其进行扩增,使得每个寡核苷酸的两个端部对应于目标等位基因的任一侧上的侧翼序列。加入聚合酶,连接酶和适当的试剂后,寡核苷酸进行填空环化、捕获期望的等位基因。溶液中加入核酸外切酶,热灭活,产品直接作为模板进行PCR扩增。对PCR产品在Illumina基因组分析仪进行上测序。将测序读数用作 PARENTAL SUPPORT™方法的输入,然后预测胎儿的倍性状态。

[0212] 在另一个实施方式中,一对夫妇-其中怀孕的母亲是高龄孕妇-想知道孕期胎儿是否有唐氏综合症、特纳氏综合症、普拉德威利综合症、或其他一些整个染色体异常,产科医生需要从母亲和父亲抽取血样。血液被送到一个实验室,一个技术人员离心母母亲的样本以分离血浆和白膜层。

[0213] 在白膜层和父亲的血液样本中的DNA,经过扩增转换,并且将在扩增的遗传物质中经过编码的遗传数据进一步从以分子形式存储的遗传数据转换称电子形式存储遗传数据,通过将遗传物质在高通过测序器上运行,测量父母的基因型。该血浆样本在一组位点用一个5000-plex半嵌套定向PCR方法经过优先富集。该DNA片段的混合物制备成适合于测序的DNA库。然后用高通量测序方法,例如ILLUMINA GAIIx基因组分析仪,对DNA使用进行测序。测序过程将以分子形式编码的信息转换成以计算机硬件中以电子形式编码的信息。一个以信息为基础的技术包括本发明公开的实施方式,如PARENTAL SUPPORT™,可用于确定胎儿的倍性状态。这可能涉及从制备样本测量的DNA测量结果,在计算机上计算多个多态性位点的等位基因计数概率;在计算机上创建多个倍性假设,其中每个与不同可能的染色体的倍性状态有关;在计算机上,对每个倍性假设,为染色体上多个多态性位点的预期的等位基因计数建立一个联合分布模型;在计算机上,用该联合分布模型和在制备的样本测得的等位基因计数确定每个倍性假设的相对概率;并调用胎儿的倍性状态通过选择对应的具有最大概率的假设的倍性状态。即可确定胎儿有唐氏综合症。打印一份报告,或以电子方式发送给孕妇的产科医生,产科医生将诊断转送给该妇女。该名妇女、她的丈夫和医生坐下来讨论他们的选择。基于胎儿的患有三倍体疾病的知识,这对夫妻决定终止妊娠。

[0214] 在一个实施方式中,一家公司可能决定提供一种从母体抽血设计用来检测孕期胎儿非整倍体的诊断技术。他们的产品可能涉及一位母亲出现在她的产科医生面前,而产科医生可能可能抽去她的血液。产科医生也可能会收集来自胎儿的父亲的遗传样本。临床医生可能会从母体血液中分离血浆,并纯化来自血浆中的 DNA。临床医生也可能从母体血液中分离白膜层,从白膜的制备DNA。临床医生也可能从父亲的遗传样本制备DNA。临床医生可使用本发明中描述的分子生物学技术,将普遍扩增标签附加到来自血浆样本中的DNA上。临床医生被通用标记过的DNA。临床医生可能通过多个技术,包括杂交捕获和定向PCR,优先富集DNA。定向PCR可能涉及巢式、半球巢式或半嵌套、或任何其他能导致有效地富集源自血浆的DNA的方法。

[0215] 定向PCR可被大量复用的,例如,在一个反应中用10000 个引物,引物定向到13、18、21号染色体上的SNP,X和那些是常见的X和Y两个染色体上的位点,以及任选的其他染色体的基因位点。这种选择性的富集和/或扩增可能涉及用不同的标签、分子条形码、,扩增标签和/或测序标签标记每一个个体分子。然后,临床医生可能对血浆样本测序,也可能对制备的孕产妇和/或父亲的DNA测序。

[0216] 分子生物学的步骤可以通过一个诊断盒全部或部分执行,序列数据可被输入到一台计算机上,或其他类型的计算平台,如在“云”中找到的。计算平台可以在目标多态性位点从定序器的测量结果计算等位基因计数定序测量的。计算平台可以为每个13、18、21号染色体、X和Y染色体创建多个与缺体、单倍体、二倍体、匹配的三倍体、以及非匹配三倍体相关的一个倍性假设。

[0217] 计算平台可在染色体上的目标位点为所要查询的5个中的每一个染色体的每个倍性假设建立一个获取等位基因计数的联合分布模型。计算平台可确定一个概率,每个倍性假设的确是用这个模型和在优先富集的源自样本血浆的DNA上的每一个测得的等位基因计数得到的。对每个13号、18号、21号、X和Y 染色体,通过选择对应与具有最大概率值的恰当假设的倍性状态,计算平台可识别胎儿的倍性状态。可生成一份含有该确定的倍性状态的报告,它也可以以电子方式发送到产科医生的输出设备上显示,或是一份打印的硬拷贝报告送至产科医生。产科医生可能会通知患者以及可选的胎儿的父亲,他们可能会决定哪些临床选项是开放的,哪一种是最可取的。

[0218] 在另一个实施方式中,一个孕妇,以下简称为“母亲”可能会决定她想知道她的胎儿(ES)是否携带任何遗传异常或其他状况。她可能会想在她有信心继续妊娠之前确保不会有任何严重畸形。她可能会去找她的妇产科医生,妇产科医生可能采集她的血液样本。他也可能采集一种遗传样本,如用口腔拭子,从她的脸颊部。他可能还需要从胎儿的父亲采集基因样本,如用口腔拭子,从精液样,或血液样本。他可能将样本送给临床医生。临床医生可能在母体的血液样本中富集部分游离的胎儿DNA。母体血液样本中,临床医生可能在母体的血液样本中富集部分去核的胎儿血细胞。临床医生可使用本发明所述的方法各个方面来确定的胎儿的遗传数据。该遗传数据可包括胎儿的倍性状态,和/或一个或多个胎儿中与疾病相关的等位基因的特征。可能会产生一个综合产前诊断结果的报告。该报告可发送或邮寄给医生,他可能会告诉母亲胎儿的遗传状态。基于胎儿有一个或多个染色体或基因异常的事实或不希望的状况,该母亲可能决定终止妊娠。基于胎儿没有任何染色体或基因的异常事实,或任何其他所关注的遗传状况,她也可决定继续妊娠的。

[0219] 另一个例子可能涉及通过精子捐献者已人工授精的孕妇,并且已怀孕。她希望她所怀的胎儿有遗传性疾病的风险降到最低。她已经在抽血者处抽血,使用本发明中描述的技术分离三个有核的胎儿血红细胞,并且还从母亲和遗传学上的父亲的采集了组织样本。对从胎儿、母亲和父亲的遗传物质适当扩增及用 ILLUMINA INFINIUM BEADARRAY基因分型,并使用本发明描述的方法高精度的清洁和阶段处理父母和胎儿的基因型,以及为胎儿确定倍性识别。胎儿被发现是整倍体,从重构的胎儿基因型预测出表型敏感性,并生成一个报告,发送到母体的医师,使他们可以决定什么样的临床决定可以是最好的。

[0220] 在一个实施方式中,原始的母亲和父亲的遗传物质通过扩增的方法转化的一定量的与序列中相似、但数量较大的DNA。然后,通过基因分型方法,由核酸编码的基因型数据,

转化为遗传测量结果,以物理和/或电子的形式存储在一个记忆装置上,如上所述。使用一种编程语言,将构成PARENTAL SUPPORT™算法相关的算法和本发明详细讨论的相关部分翻译成计算机程序。然后,在计算机上的硬件上通过执行所述计算机程序,代替以物理形式对位和字节编码、排列模式表示原始测量数据,将其成为转换成为代表高置信的胎儿倍性状态测定结果。此变换中的细节将依赖于数据本身、计算机语言和执行本发明所述的硬件系统。然后,可将被以物理配置形式代表一个高质量的胎儿倍性测定结果的数据转换成一份报告发送给保健医生。此变换可以使用打印机或计算机显示器进行。报告可以是印刷拷贝,在纸张上或其他合适的介质上,或它可以是电子形式的。在电子报告的情况下,它可以被发送,它可以以物理方式存储的位于保健医生可及的计算机记忆装置上,它也可被显示在屏幕上以便于阅读。在屏幕显示的情况下,通过对在显示装置上像素的物理转换,数据可能会被转化为可读的格式。这种转换可以通过物理方式向荧光屏发射发射电子实现,通过改变电荷的方法,以物理方式改变在屏幕上的一组特定的像素的透明度,屏幕可能位于基板前面,发射或吸收光子。这种转换可以通过改变液晶中纳米级的分子取向的方式来完成,例如,在一组特定的像素,液晶分子从向列相到胆甾醇型相或近晶相变换。这种变换可以完成通过电流引起光子从由多个排列成有意义图案的发光二极管构成的一组特定的像素发射光子来完成。这种变换可以通过任何其他方式来完成,以用来显示信息,如在计算机屏幕,或一些其他的输出设备或传递信息的方法。保健医生可能基于报告采取行动的,于是报告中的数据转化成行动。该行动可以是继续或中止妊娠,在这种情况下具有遗传异常的妊娠中的胎儿被变换成非生物胎儿。此处列出的转换可能会被合并,例如,医生可能将孕妇和父亲的遗传物质,通过一系列本发明中概述的公开的步骤,转换成包括对遗传异常胎儿流产,或包括继续妊娠的到医疗决定。另外,可以将一组基因型测量结果转换成一份报告,帮助医生治疗其怀孕的患者。

[0221] 在本发明的一个实施方式中,本发明所述的方法可用于确定胎儿的倍性状态,即使宿主母体,即孕妇,不是她所携带胎儿的亲生母体。在本发明的一个实施方式,本发明所述的方法可用于确定胎儿的倍性状态,仅使用孕妇血液样本,而不需要父亲的遗传样本。

[0222] 在本发明公开的实施方式中,一些数学方法用来建立关于有限数目非整倍体的状态的假设。在某些情况下,例如,只有零、一个或两个染色体预计是源自每个家长的。在本发明的一些实施方式中,这种数学推导可扩展到其他形式的非整倍体,如四倍体,其中的三个染色体源自一个父母,五倍体、六倍体等,而没有改变本发明的基本概念。同时,也可以把重点放在一个更小的数目倍性状态,例如,仅三倍体和二倍体。请注意,指示非正数染色体的倍性测定结果可能表示在遗传物质样本中的嵌合。

[0223] 在一些实施方式中,基因异常是一个非整倍体类型,如唐氏综合症(或21三倍体),爱德华综合症(18三倍体),Patau 综合症(13三倍体),特纳综合症(45X),克氏综合症(携带2X染色体的男性),普拉德-威利综合症,以及DiGeorge氏综合症(UPD15)。如上所述的先天性疾患,通常是不希望的,胎儿罹患一个或多个表型异常,的知识可能为终止妊娠的决定提供依据、采取必要的预防措施以备一个有特殊需要的孩子的出生,或采取一些治疗方法的诞生意味着,以减轻染色体异常的严重程度。

[0224] 在一些实施方式中,本发明所述的方法可用于在一个非常早期的胎龄,例如早在四周,早在五周,早在六周,早在七周,早在八周,早在九周,早在10周,早在11周,或早在12

周。

[0225] 请注意,已被证明在宿主血液中可发现源自生活在宿主体内癌症的DNA。以同样的方式,可以用源自混合母体血液中发现的DNA的测量进行基因诊断,基因诊断同样可以很好地用源自宿主的血液中发现的混合的DNA的测量进行。本发明中的任何声明,进行母体血液确定倍性状态或胎儿基因状态的读数与确定源自宿主血液的基因状态或癌症的基因状态的完全读数相同。

[0226] 在一些实施方式中,本发明公开的一种方法允许人们确定癌症倍性状态,该方法包括:获得包含源自宿主遗传物质和癌症和遗传物质的混合样本,测量混合的样本中的DNA,计算混合样本中癌症起源的DNA的分数;使用对混合的样本的测量结果和计算的分数确定癌症的倍性状态。在一些实施方式中,该方法可以进一步基于包括基于癌症倍性状态测定结果实施癌症的治疗。其中,所述癌症治疗选自以下组中的治疗措施包括一种药物、生物治疗、基于治疗和它们的组合的抗体。

[0227] 在一些实施方式中,本发明所公开的方法可用于植入前遗传学诊断 (PGD) 的背景下,体外受精过程中的胚胎选择,其中目标个体是胚胎,从源自第3天胚胎的一个或两个细胞的活检,或源自第五天、第六天的胚胎滋养层的活检,亲本的基因型数据可以用来进行关于胚胎的倍性测定。在一个在PGD设置中,只测定孩子的DNA,并且只检测少量细胞,一般是一个到5个,但也可多达10个、20个或50个。A和B等位基因的初始的拷贝 (在SNP) 的总数,一般由孩子基因型和细胞的数量确定。在 NPD过程中,起始拷贝数目非常高,这样期望经过PCR后等位基因比率正确地反映起始比率。然而,在PGD过程中,少量的开始拷贝意味着污染和不完善的PCR相率对接下来的PCR等位基因比率存在重要影响。这种影响可能比在预测测序后等位基因比率测量的偏差时的读入深度更重要。基于PCR的探测效率和污染概率,通过蒙特卡罗模拟方法模拟PCR过程可以创建一个已知孩子基因型的测得等位基因比率的分布。对每个可能的孩子基因型给出一个等位基因率分布,可以计算各种假设的可能性,如在NIPD描述的。

[0228] 本文所公开的任何一个实施方式可以在数字电子电路,集成电路,专门设计的ASIC(特定用途集成电路)、计算机硬件、固件,软件、或它们的组合中实施。在本发明公开的实施方式中的装置可以清楚的体现在一个机器可读存储装置中由可编程处理器执行的以计算机程序产品的方式实施;并且本发明公开的实施方式的方法步骤可以由可编程处理器执行的指令程序实施,通过输入数据操作和产生输出的方式实现本发明公开的实施方式的功能。本发明公开的实施方式可以方便地在一个或多个计算机程序实现,计算机程序在一个可编程系统上是可执行的和/或可翻译的,可编程系统包括至少一个可以是专用的或通用的可编程处理器,从一个存储系统耦合以接收数据和指令,并传输数据和指令到一个存储系统,至少有一个输入设备,和至少有一个输出设备。各计算机程序可以在一个高层次的程序或面向目标编程语言,或在汇编或机器语言中实现,在任何情况下,语言可以是编译的或解释的语言。计算机程序可以以任何形式部署,包括作为独立的程序,或者作为模块,组件,子例程或其他适于在计算环境中使用的单元。计算机程序可以部署到执行或翻译的一台计算机上,或者在位于一个站点的多台计算机上,或者分布在由通信网络互联的多个站点上。

[0229] 计算机可读存储介质,如本文所用,是指物理或有形的存储 (而不是信号),包括

但不限于以任何用于有形信息存储的方法或技术实现的易失性和非易失性,可移动和不可移动介质,如任何方法或技术作为计算机可读指令,数据结构,程序模块或其它数据。计算机可读存储介质包括,但不限于,RAM,ROM, EPROM,EEPROM,闪存或其它固态存储器技术,CD-ROM, DVD或其它光盘存储,磁带盒,磁带,磁盘存储器或其它磁存储设备,或任何其他物理或材料的介质,可用于有形地存储所需要的信息或数据或指令,并且可以由计算机或处理器访问。

[0230] 本发明所描述的任何一种方法,可包括以物理格式输出数据,如在计算机屏幕上,或在一张纸上打印输出。在本文件的其他地方的对任何实施方式的解释中,应当理解,所描述的方法可以由医生做出的以一种格式与输出的可操作的数据进行组合。此外,所描述的方法,也可以与一个临床决定的实际执行相组合,导致临床治疗,或一个不作行动的临床决定的执行。本发明文件中所描述的用于确定有关目标个体的遗传数据的某些实施方式可以在IVF背景下与选择一个或多个胚胎转移的决定相结合,任选地,与转移胚胎到准妈妈子宫的过程相。本发明文件所描述的用于确定有关目标个体的遗传数据的一些实施方式中,在产前诊断的背景下,可以与一个潜在的染色体异常的通知相结合,或不与其结合,与医疗专业,任选地与流产胎儿,或不流产胎儿的决定相结合。本发明文件所描述的一些实施方式可与可操作的数据输出,一个临床决定的执行导致在临床治疗,或一个不作任何动作的临床决定的执行相结合。

[0231] 定向富集和测序

[0232] 为非侵入性产前等位基因识别或倍性识别,在一组目标位点富集一个DNA的样本,作为方法的一部分,接下来进行测序的一种技术的使用可赋予一些意想不到的优点。在本发明的一些实施方式中,该方法包括使用遗传数据测量和基于信息的方法,例如,PARENTAL SUPPORT™ (PS),一些实施方式的最终结果是可操作的胚胎或胎儿遗传数据。有许多方法可用于测量个体和/或相关个体的遗传数据的方法的包含于方法中的一部分。在一个实施方式中,本文所公开了富集一组目标等位基因浓度的方法,该方法包括以下步骤中的一个或多个:定向扩增遗传物质,加入位点特异性寡核苷酸探针,连接指定的DNA链,分离一组所需的DNA,去除不需要的反应组分,通过杂合检测目标DNA序列,以及通过DNA测序方法检测一个或多个的DNA链的序列。在某些情况下,DNA链可以是指目标的遗传材料,在某些情况下他们可以是指引物,在某些情况下,他们他们可以是合成的序列或它们的组合。这些步骤可以多个不同的顺序实施。对于给定的高度可变的分子生物学性质,那种方法,或那种步骤的组合将表现不佳、表现好、或在各种情况下最好通常是不明显的。

[0233] 例如,在定向扩增前,采取一个通用的扩增DNA的步骤可赋予一些优点,如消除瓶颈风险和减少等位基因偏差。该DNA 也可能与能与两个目标序列中相邻区域,一边一个,杂交的寡核苷酸探针混合。杂交后,探针的端部可通过加入一个聚合酶、一种连接的措施、和任何必要的试剂被连接,以使探针的环化。环化后,可以加入核酸外切酶以消化非环化的遗传物质,然后检测环化探针。DNA可以与能与目标序列中的两个相邻区域,一边一个,杂交的PCR引物混合。杂交后,探针的端部可通过加入一个聚合酶、一种连接的措施、和任何完成PCR扩增的试剂被连接。扩增或未经扩增的DNA可以通过定向一组基因位点的杂交捕获探针定向。杂交后,探针可以从混合物定位和分离,从而提供在目标序列中扩增的DNA混合物。

[0234] 在一些实施方式中,,目标遗传物质的检测可以用多重复用的方式进行。以并行反

式运行的遗传目标序列的数量可以是1到10个,10到100个,100到1000个,1000到10000个,10000到100000个,100000到1000000个,或1000000到10000000个不等。请注意,在现有技术的公开中,包括成功复用PCR反应的涉及高达约50或100个引物库,没有更多的。之前尝试复用每个库池多余100个引物时导致产生了副反应的重大问题,如引物二聚体的形成。

[0235] 在一些实施方式中,本方法可用于基因分型单个细胞、一个小数量的细胞、2至5个细胞、6到10个细胞、10到20个细胞、20到50细胞、50到100个细胞,100个到1000个细胞或少量的胞外DNA的,例如从1到10皮克,从10到100皮克,从100皮克到1毫微克,1到10毫微克,10到100毫微克,或从几百毫微克到1微克。

[0236] 为等位基因识别或倍性识别,定向某些基因位点,作为方法的一部分,接下来进行测序的一种方法的使用可赋予一些意想不到的优点。一些可用于定向DNA,或优先富集的方法包括使用圆形探针,连接倒置探针(LIPs,MIPS),杂交捕获的方法,如SURESELECT,以及定向PCR或连接介导的PCR扩增策略。

[0237] 在一些实施方式中,本发明公开的方法包括使用的遗传数据测量和基于信息的方法,例如,PARENTAL SUPPORT™(PS)。PARENTAL SUPPORT™是本发明发明所述的基于信息操纵遗传数据的方法。一些实施方式的最终结果是可操作的胚胎或胎儿的遗传数据,然后在这些可操作数据的基础上做出临床上决定。PS方法背后的算法利用测得的目标个体的遗传数据,目标个体往往是胚胎或胎儿,以及测得的相关个体的遗传数据,利用目标个体的遗传状态是已知的条件,能够提高算法的准确性。在一个实施方式中,测得的遗传数据用于在产前基因诊断中的倍性测定背景中。在一个实施方式中,测得的遗传数据用于在体外受精中的倍性测定或胚胎等位基因识别的背景中。

[0238] 在上述的背景下,有许多方法可用于测量个体和/或相关个体的遗传数据。这些不同的方法包括了一些步骤,这些步骤通常涉及遗传物质的扩增,以及寡核苷酸探针,连接指定的DNA链,分离所需的DNA集,去除不需要的反应组分,通过杂交检测某些DNA的序列,通过DNA测序方法检测一个或多个DNA链的序列的。在某些情况下,DNA链可能指目标基因物质,在某些情况下,他们可能指引物,在某些情况下,他们可能指合成的序列或它们的组合。这些步骤可以多个不同的顺序实施。对于给定的高度可变的分子生物学性质,哪种方法,哪那种步骤的组合将表现不佳、表现好、或在各种情况下最好通常是不明显的。

[0239] 请注意,在理论上定向任何数量基因组中位点是可能的,在任何地方,从一个位点到超过一百万位点。如果对DNA样本进行定向,然后测序,由定序器所读取的将富集的等位基因的百分比与其在样本中的天然丰度有关。在任何地方,富集程度可以(甚至更少)从百分之一到十倍,一百倍,一千倍,甚至几百万倍。在人类基因组中有大约有30亿个碱基对,和核苷酸,包括约75万个多态性位点。定向的位点越多,富集程度就越小是可能的。定向位点的数目越少,富集的程度就越大是可能的,并且,对于给定数量的序列读数,在这些位点可达到更大的读数深度。

[0240] 在本发明的一个实施方式中,定向或优先可能完全集中在SNPs。在一个实施方式中,定位或优先可能集中在任何多态性位点。一些商业定位产品可供富集外显子。令人惊讶的是,当使用用于依赖于等位基因分布的NPD方法时,专用于定向SNPs,或专用于多态性位点,是特别有利的。也有发表的采用测序进行NPD方法,例如美国专利7888017,涉及一个读数计数分析,其中读数计数集中于计数映射到一个给定的染色体的读数的数量,其中被分

析的序列读数不集中于多态性基因组的区域。这些类型的方法,不注重多态性等位基因不会惠及尽可能多的一组等位基因的定向或优先富集。

[0241] 在本发明的一个实施方式中,有可能使用一个关注于SNPs 的定向方法,在基因组的多态性区域富集一个基因样品。在一个实施方式中,有可能关注于一个小数量的SNPs,例如在1和100 之间的SNPs或数量较多,例如,在100和1000之间,在1000 和10000之间,在10000和100000之间或超过10万的SNPs。在一个实施方式中,有可能关注于一个或少数几个与活体三倍体产儿相关的染色体,例如13、18、21号染色体、X和Y染色体,或它们的某种组合。在一个实施方式中,有可能以一个小的因子富集目标SNPs,例如1.01倍和100倍,或通过一个较大的因子,例如在100倍和1000000倍之间,或甚至超过1000000倍。在本发明的一个实施方式中,有可能使用的一个定位方法,创建一个在基因组多态性的区域优先富集的DNA样本。在一个实施方式中,有可能使用这种方法创建一个具有任何这些特征的DNA 混合物,其中的DNA的混合物中包含母体的DNA和游离的胎儿的DNA。在一个实施方式中,有可能使用这种方法来创建一个具有这些因素的任何组合的DNA的混合物。例如,本发明所述的方法可被用于生成一个包括母体DNA和胎儿DNA的混合物,并是对DNA对应于200个SNPs进行优先富集的,所有这些都位于18号染色体或21号染色体,并且每个的平均富集水平是 1000倍。在另一个例子中,有可能使用该方法来创建一个优先在全部或大多数位于13、18、21号染色体、X和Y染色体上的 10000个SNPs富集的DNA混合物,每个基因位点的平均富集水平超过500倍。本发明所述的任何定向方法可用于创建优先在某些位点富集的DNA的混合物。

[0242] 在一些实施方式中,本发明的方法进一步包括:使用高通量的DNA定序器测量混合部分中的DNA,其中,混合组分中的 DNA,包含一个不相称的数目的源自一个或多个染色体的序列,其中所述的一个或多个染色体是从以下的染色体组中取得的,包括13号染色体,18号染色体,21号染色体,X染色体,Y染色体以及它们的组合。

[0243] 这里所描述的三种方法:多重PCR,杂交定向捕获和链接倒置探针(LIPs),其可以用于从足够数量的、源自母体血浆样本的多态性位点获取和分析测量结果以便检测胎儿的非整倍性;这并不意味着排除选择性富集目标位点的其它方法。其他方法也可能同样在不改变方法的本质的情况下很好地使用。在每一种情况下测定的多态性可能包括SNPs (SNPs),小插入缺失,或STR 基因位点。一种优选的方法包括SNPs的使用。每一种方法都产生等位基因频率数据;可分析每个目标基因位点的等位基因频率和/或来自这些基因位点的联合等位基因频率分布数据,以确定胎儿的倍性。由于有限的源材料和母体血浆含有孕妇和胎儿DNA 混合物的事实,每种方法都有自己的考虑。该方法与其他方法结合使用,以提供更精确地确定。在一个实施方式中,这种方法也可以与序列计数的方法组合,如在美国专利7888017中所描述的。描述的方法也可用于从母体血浆样本非侵入性地检测胎儿亲子鉴定。此外,每一种方法可以应用到其他的DNA混合物或纯的 DNA样本,以检测非整倍体染色体的存在或不存在,对大量的、源自降解的DNA样本的SNP进行基因分型,检测片段的拷贝数变异(CNVs),以检测所关注的其他基因型状态,或它们的某种组合。

[0244] 准确测量样本中的等位基因分布

[0245] 目前的测序方法可用于估计样本中的等位基因分布。一个这样的方法包括随机抽样DNA库中的序列,被称为鸟枪法测序。在测序数据一个特定等位基因的比例通常是非常低的,可通过简单的统计确定。人类基因组包含约30亿个碱基对。因此,如果使用的测序方法

对100个基点读数,在每30万序列读数中,一个特定的等位基因被测量一次。

[0246] 在一个实施方式中,本发明的方法用于,从测得的该染色体基因位点的等位基因分布,确定两个或两个以上不同的、在DNA 样本中包含同一组中的位点的单倍型存在或不存在。不同的单倍型可能代表源自一个个体的两个不同的同源染色体,源自一个三倍体个体的三种不同的同源染色体,源自母体和胎儿的三个不同的同源单倍型,其中的一个单倍型在母体和胎儿之间是共享的,源自母体和胎儿的三个或四个单倍型,其中的一个或两个单倍型在母体和胎儿之间是共享的,或其他组合。单倍型之间的多态性等位基因往往具有更多的信息,但是对任何等位基因,其中对这个相同的等位基因母体和父体不都是纯合子,在已有简单的读数分析信息之外通过测得的等位基因分布,将产生有用的信息。

[0247] 然而,这样的—个样本的鸟枪法测序,是非常低效的,因为在样本中不同的多态性之间它对区域产生很多不是多态多态性的序列,或不感兴趣的染色体,因此不揭示关于目标单型型的比例信息。本发明所描述的方法是,样本中具体的目标和/或优先富集的DNA片段在基因组中更可能是多态的,从而提高由测序得到的等位基因信息的产量。需要注意,对于在一个富集的能真正代表出现在目标个体中数量的样本中检测等位基因分布,与在目标片段上的一个给定位点的其他等位基因相比。很少或没有优先富集某个等位基因是至关重要的。在本领域中已知的目前定向多态等位基因的方法设计成确保至少有一些被检测到的任何存在的等位基因。然而,这些方法没有为无偏地测量存在于原始混合物中的多态性等位基因的等位基因分布的目而设计。任何特定的定向富集方法,将能够产生富集的样本,其中测得的等位基因分布能比其他方法的等位基因分布更好地准确代表存在与原有的未经扩增的样本的等位基因分布是废显而易见的。在理论上,虽然许多富集方法可以预料能完成这样一个目标,但一个熟知本领域的一个普通的技术人员清楚地知道,现有的扩增、定向和其他优先扩增方法存在—种很大的随机或确定性偏差。本文所描述的方法的一个实施方式,允许多个DNA混合物中发现的等位基因对应于基因组中的一个给定的位点,其中的基因组是被扩增的或以—种各等位基因的富集度几乎—样的方式优先富集的。另一种方法认为,该方法允许出现在该混合物中相对数量的等位基因以—个整体增加,而对应于每个基因位点的等位基因之间的比率与他们在原来的DNA混合物中的比率基本保持不变。用在现有技术中的方法对基因位点优先富集可以导致超过1%,超过2%,超过5%,甚至10%以上的等位基因偏压。这个优先富集可能是由于使用杂交捕获方法时的捕获偏差,或扩增偏差,偏差在每一个周期中可能很小,但当与超过20,30或40个周期复合时,可能会变得很大。本发明的目的,保持基本上相同的比率意思是原始混合物中的等位基因比率除以所得到的混合物中的等位基因的比率是在0.95到1.05之间,在0.98和1.02之间,在0.99之间和1.01,在0.995和1.005之间,在0.998和1.002之间,在0.999 和1.001之间,或在0.9999和1.0001。请注意,这里提出的等位基因比率的计算,不能用于目标个体的倍性状态的确定,可能只有—个用于测量等位基因偏差的指标。

[0248] 在一个实施方式中,一旦混合物在—组目标位点是被优先富集过的,可能使用任何—个之前的,现在的、或下一代测序隆样本的仪器(从—个单一的分子产生的样本;例子包括ILLUMINA GAIIx,ILLUMINA HISEQ,LIFE TECHNOLOGIES SOLiD, 5500XL)对其进行测序。通过目标区域内的特定等位基因测序可以对比率做出评价。根据等位基因类型和据此确定不同等位基因的量,可对这些测序读数进行进行分析何计数。对—到几个碱基长度的

变异,等位基因的检测将通过测序进行,并且为了评估该捕获分子的等位基因组合,扩展所讨论的等位基因序列对数的跨度是必不可少。用于检测基因型的捕获分子的总数可通过增加读测序的长度得到提高。对所有分子全测序将保证在所提供富集库中收集最大量的数据。然而目前测序是昂贵,并且有一种可以使用一个较小的数量的序列读数测量等位基因分布的方法将有很大的价值。此外,当读书长度增加时,对最大可能长度的读数存在技术上的限制以及精度的限制。具有最实用价值的等位基因中只有一到几个碱基长度,但在理论上可以使用任何短于测序读数长度的等位基因。虽然在所有类型的等位基因都存在变体,本文所提供的例子集中在SNPs或只有几个相邻的碱基对含有变体的情况。较大的变体如片段拷贝数目变体可以通过合并这些较小的变异,在很多情况下,如复制片段内部SNP的全部集合进行检测。变体大于几个碱基时,如STRs需要特殊考虑而且一些定向方法可以使用而其他方法则不能使用。

[0249] 有多中定向方法可用于基因组中按特性分离和富集一个或多个变异位置,通常情况下,这些依靠利用未变异序列位于变异序列两侧的特点。在测序背景下,存在与定向相关的的现有技术,其中,培养基是母体血浆(见,例如,Liao等人,临床化学,2011, 57(1):第92-101)。然而,在现有技术中的这些方法都使用定向外显子的定向探针,不专注于定向基因组多态区域。在一个实施方式中,本发明的方法包括使用专注于或几乎完全专注于多态区域的定向探针。在一个实施方式中,本发明的方法包括使用专注于或几乎完全专注于SNPs的定向探针目标,在本发明的一些实施方式中,定向的多态性位点包括至少10%的SNPs,至少20%的SNPs,至少30%的SNPs,至少40%的SNPs,至少50%的SNPs,至少60%的SNPs,至少70%的SNPs,至少80%的SNPs,至少90%的SNPs,至少95%的SNPs,至少98%的SNPs,至少99%的SNPs,至少99.9%的SNPs,或专注于SNPs。

[0250] 在一个实施方式中,本发明的方法可以被用来确定源自DNA分子混合物中的基因型(在特定的位点的DNA的碱基组成)和这些基因型的相对比例,其中,这些DNA分子可能起源于一个或多个基因不同的个体。在一个实施方式中,本发明的方法可用于确定一组多态性位点的基因型,和存在于这些位点的不同等位基因数量的相对比例。在一个实施方式中,多态性位点的可能包括全部SNPs。在一个实施方式中,多态性位点可能包括SNPs、单一的串联重复序列,和其他多态性。在一个实施方式中,本发明的方法可用于确定在DNA混合物中的一组多态性位点的等位基因的相对分布,其中包括DNA混合物包括源自母体的DNA和源于胎儿的DNA。在一个实施方式中,联合等位基因分布是基于从孕妇血液离的DNA混合物确定的。在一个实施方式中,在一组基因位点等位基因分布可用于确定妊娠中的胎儿的一个或多个染色体上的倍性状态。

[0251] 在一个实施方式中,DNA分子的混合物可以是来自于从一个个体的多个细胞中提取的。在一个实施方式中,源自原始细胞集合的ADN可能包括有相同或不同的基因型的二倍体或单倍体细胞的混合物中,如果该个体是镶嵌的(生殖细胞或体细胞)。在一个实施方式中,DNA分子的混合物也可以是来自于从单细胞中提取的。在一个实施方式中,DNA分子混合物也可以是源自相同个体或不同个体的两个或多个细胞提取的。在一个实施方式中,DNA分子的混合物也可能源自从已经脱离细胞生物物质中分离的DNA,所述的生物物质,如众所周知的,含有无细胞的DNA的血浆。。在一个实施方式中,这种生物物质可以是源自一个或多个个体的DNA的混合物,在怀孕的情况下,它已被证明胎儿DNA存在于该混合物中。在一个实

施方式中,生物物质可以是源自母体血液中发现的细胞,其中的某些细胞是胎儿起源的细胞。在一个实施方式中,生物物质可以是源自一个怀孕的血液中的细胞,其中已对胎儿细胞进行了富集的。

[0252] 圆形探针

[0253] 本发明的一些实施方式中涉及的“链接倒置探针”(LIPs)的使用,先前已在文献中描述过。LIPs是一个通用术语,指包含的技术涉及一个圆形DNA分子的创建,其中的探针设计成能与目标等位基因任一侧的DNA目标区域进行杂交,例如,加入适当的聚合酶和/或连接酶、适当条件、缓冲剂和其它试剂,将完成的互补的、倒置的DNA区域,横跨目标等位基因创建一个捕获目标等位基因中发现的信息的DNA圆环。LIPs也可以称为被预环化的探针,或预环化探针,或圆化的探针。LIPs探针可以是在 50和500个核苷酸长度之间的线性DNA分子,以及在一个实施方式中长度在70和100核苷酸之间;在某些实施方式中的,它可以是长于或短于本文所描述。本发明的其他实施方式中涉及不同的LIPs技术的变形,如的挂锁探针和分子倒置探针(MIPS)。

[0254] 用于测序,对特定位置定向的一种方法是合成一个探针,其中探针的3'和5'端的在目标区域任一侧相邻的位置,以倒置的方式退火定向目标DNA,这样,加入DNA聚合酶和DNA连接酶可导致从3'末端延伸,加入碱基到与目标分子(间隙填充)互补的单链探针是,接下来连接新的3'端到初始探针的5'端导致形成一个环状DNA分子,其后它可从背景DNA分离。该探针端部设计成位于感兴趣目标区域的侧翼。这种方法一个方面通常被称为MIPS并且与阵列技术一起已被用于填充序列的性质。在测定等位基因比率的背景中,使用的MIPS一个缺点是杂交环化和扩增步骤在相同的位点不能以相同的速率进行。这导致测量的等位基因比率不能代表存在于原始混合物中的实际等位基因比率。。

[0255] 在一个实施方式中,圆形探针的构成是这样的,目标多态性基因位点上游杂化的探针区域和目标多态性基因位点下游杂化的探针区域通过非核酸骨架以共价键连接。该骨架可以是任何生物相容性的分子或生物相容的分子的组合。可能的生物相容分子的一些例子是聚(乙二醇),聚碳酸酯,聚氨酯,聚乙烯,聚丙烯,聚砜类聚合物,硅树脂,纤维素,含氟聚合物,丙烯酸类化合物,苯乙烯嵌段共聚物,和其他的嵌段共聚物。

[0256] 在本发明的一个实施方式中,这种方法已被修改为可以很容易地适合于作为询问填充序列测序一种手段。为了保留原始样本中的原始等位基因的比例的至少一个重要的考虑因素必须考虑在内。在间隙填充区域中不同等位基因之间的可变位置必须不能太接近的探针结合位点,因为DNA聚合酶导致的变体差异会出现起始偏差。另一个考虑是,额外的变体可能存在于与间隙填充的区域的变体相关的探针的结合位点,这可能会导致来自不同等位基因的不均等扩增。在本发明的一个实施方式中,设计成与距离目标等位基因变体位置(多态性位点)一个或几个碱基的位置杂交。杂交相差的的目标等位基因的变体位置的基准。多态性位点(SNPs或以其他方式)之间的碱基和预环化探针的3'端和5'端设计用于杂交的该碱基数量可以是一个碱基、可以是两个碱基、可以是三个碱基,可以是四个碱基,可以是5个碱基,可以是6个碱基,可以是7到10个碱基,可以是11到15个碱基,或者它可以是16到20个碱基,20到30个碱基,或30至60个碱基。正向和反向引物可被设计杂交不同数目的距离多态性位点碱基。圆形探针可以用目前的DNA合成技术大量生成,允许使非常大数量探针生成的和潜在的汇集,使许多位点同时询问。据报道,有用超过300000探针同时工作的。讨论

包括圆形探针方法的两篇论文可被用于测量目标个体的基因组数据包括：波雷卡等人，自然方法，2007年4(11)，第931-936页；和Turner等人，自然的方法，2009,6(5)，第315-316页。在这些论文中所描述的方法，也可以与本文所描述的其他方法结合使用。这两篇文献中的方法的某些步骤，也可以与本发明所述的其他方法的其他步骤组合使用。

[0257] 在本文公开的方法的一些实施方式中，目标个体的遗传物质被任意扩增，随后通过预环化探针杂交，进行间隙填充以填补在杂交探针两端之间的碱基，连接两端形成一个环化探针，以及扩增环化探针，使用，例如，滚环扩增。一旦所需的目标等位基因的遗传信息被由适当设计的环化寡核苷酸探针捕获，如在LIPs 系统中，环形探针的基因序可被测量以得到所需的序列数据。在一个实施方式中，适当设计的寡核苷酸探针可直接用于未扩增的目标个体的遗传物质，之后扩增。需要注意的是，也可以使用一些扩增程序可用于扩增原始的遗传材料，或环化的LIPs，包括滚动圆扩增，丙二醛(MDA)，或其他扩增协议。也可以使用不同的方法测量目标基因组遗传信息，例如，使用高通量测序，Sanger 测序，以及其他测序方法，通过杂交捕集，通过环化捕集，多重 PCR，以及其他杂交方法，以及他们的组合。

[0258] 一旦使用一个上述方法或一种上述方法组合的方法、一个基于信息的方法如PARENTAL SUPPORT™方法，对一个个体的遗传物质进行了测量，那么，一个适当的遗传测量可用于确定高该个体的一个或多个染色体的倍性状态，和/或一个或一组的等位基因的遗传状态，特别是那些与一种疾病或感兴趣的遗传状态相关的等位基因。请注意，用LIPs进行基因序列的复用捕获，然后使用测序基进行因分型已有报道。然而，基于LIPs策略的序列的数据用于在一个单细胞中的、少数细胞中的、或者细胞外 DNA的、遗传物质的扩增，还没有被用于确定目标个体倍性状态的目的。

[0259] 由遗传数据，如通过杂交阵列，如ILLUMINA INFINIUM阵列，或AFFYMETRIX的基因芯片测量的，应用信息为基础的方法来确定一个个体的倍性状态，已经在本文其他地方提及及的文献中有所描述。然而，本发明所述的方法是对先前在文献中描述的方法的改进。例如，基于LIPs的方法，然后通过高通量测序意外提供更好的基因型数据，由于该方法具有更好的复用性能，更好的捕捉特异性，更好的均匀性，和低等位基因偏差。更大的复用允许定向更多的目标等位基因，获得更准确的结果。均匀性更好导致更多用于测量的目标等位基因，从而提供更准确的结果。较低的等位基因偏差率导致错误之比率较低，从而给出更准确的结果。更准确的结果导致改善临床结果，以及更好的医疗处理。

[0260] 重要的是要注意，LIPs可作为一种方法，定向DNA样本中特定的位点用于使用一些方法分形而不是测序。例如，LIPs 可以用于定向用于基于SNP阵列或其它DNA或RNA微阵列进行基因分型。

[0261] 连接介导PCR

[0262] 连接介导的PCR是一种PCR方法用于通过扩增DNA的混合物中的一个或多个位点优先富集DNA样本，该方法包括：获取一组引物对，其中引物对中的每个引物包含一个特定的目标序列和一个非目标序列，其中特定的目标序列设计成对源自多态性位点的上游和下游的目标区域退火，可以以0,1,2,3,4,5,6, 7,8,9,10,11-20,21-30,31-40,41-50,51-100，或超过100 个从多态性位点分离；从上游引物的3' 引物端杂交的DNA填充单链它和下游引物的带有与目标分子核苷酸互补的5'引物端之间的区域，连接上个上游引物杂交的碱基到相邻的下游引物5' 端的碱基；用包含在上游引物5' 引物端的非定向序列扩增只杂交和连接

的分子;区分目标的引物对可以混合在同一反应。非目标序列作为通用的序列,例如,已成功杂交并连接的所有引物对可能本一对扩增引物对所引物扩增。

[0263] 杂交捕获

[0264] 在目标基因组优先富集一组特定的序列可以有多种方式可以完成。本发明中的其他地方是描述了LIPs可以用来怎样定向一组特定的序列,但在所有这些应用中,其他定向和/或优先富集方法可以同样适用于相同的目的。另一种定向方法的一个例子是通过杂交方法的捕获。通过杂交捕获的技术商业化的一些例子包括AGILENT's SURE SELECT和ILLUMINA's TruSeq。在杂交捕获中,与所需的目标序列互补或几乎互补的一组寡核苷酸允许与DNA混合物杂交,然后,从混合物中物理分离。一旦所需要的序列与目标寡核苷酸杂交,物理去除的目标寡核苷酸的效果也去除目标序列。一旦除去杂交的寡核苷酸,它们可以被加热到高于其熔融温度并被扩增。某些物理移除目标寡核苷酸的方法是通过共价键结合目标的寡核苷酸到固体支持物上,例如磁性珠,或一个芯片上。另一种物理移除目标的寡核苷酸的方法是通过共价键以与另一分子基团有很强的亲和力结合到分子基团。这样的分子对的一个例子是生物素和抗蛋白链霉素,如在SURE SELECT中使用的。因此,目标序列可以以共价连接到一个生物素分子上,杂交后,用抗生物素蛋白链霉素附加的固体支持物可用于拉下生物素化的寡核苷酸,杂交到目标序列。

[0265] 杂交捕获包括与感兴趣的目标分子目标互补的探针杂交到目标分子。杂交捕获探针最初开发用于定向和富集存在于目标之间相对均匀的基因组的大部分。在该应用中,重要的是所有目标被足够地均匀性扩增及所有区域可以通过测序检测,然而,没有考虑保留在原始样本中的等位基因比例。接下来的捕获,存在于样本中的等位基因可通过对捕获的分子直接测序确定。对这些测序结果进行分析并根据等位基因类型计数。然而,使用目前的技术,所测得的捕获序列的等位基因分布通常不代表原来的等位基因分布。

[0266] 在一个实施方式中,等位基因的因检测通过测序进行。为了捕捉到多态性位点的等位基因的特征,评估捕获分子的组成,测序读数的跨度涵盖考虑之中的等位基因是必不可少的。由于捕获分子往往具有可变的长度,根据测序不保证能够覆盖的变体的位置,除非对整个分子测序。然而,考虑到成本和技术上的限制,以最大可能的长度和测序读数精度对整个分子测序是不可行的。在一个实施方式中,读数长度从30至增加到约50或约70个碱基可以大大提高读数数量以便覆盖目标序列中变体的位置。

[0267] 另一种增加读数数量以询问感兴趣的位置的方法是降低探针的长度,只要不会导致在底层的富集等位基因的偏差。合成探针的长度应足够长,使得两个所设计的与两个在1个位点发现的不同的等位基因杂交的探针,将以近乎均匀的方式与白原始样本中的不同的等位基因杂交。目前,在本领域中已知的方法描述的通常是长度超过120个碱基的探针。在当前实施方式中,如果等位基因是一个或几个碱基,则捕获探针可以是少于约110个碱基,少于约100个碱基,少于约90个碱基,少于约80个碱基,少于约70个碱基,少于约60个碱基,少于约50个碱基,少于约40个碱基,少于约30个碱基,以及少于约25个碱基,这足以确保与富集所有等位基因相等。当用杂交捕获技术富集DNA混合物,是包含从血液,例如母体血液中分离的游离的DNA的混合物, DNA的平均长度是相当短的,通常少于200个碱基。使用更短的探针可导致一个捕获探针捕获所需要DNA片段的更大机会。较大的变化,可能需要更长的探针。在一个实施方式中,感兴趣的变化是一个(一个SNP)的到几个碱基的长度。在一个实

施方式中,在基因组中的目标区域可以使用杂交捕获探针优先富集,其中杂交捕获探针可以短于90个碱基的长度,也可以是小于80个碱基,小于70个碱基,小于60个碱基,小于50个碱基,小于40个碱基,小于30个碱基,或少于25个碱基。在一个实施方式中,为了增加所需的等位基因测序的机会,设计的用于杂交位于多态等位基因侧翼区域的探针长度可以从90个碱基以上减少到约80个碱基,或到约70个碱基,或到约60个碱基,或到约50个碱基,或到约40个碱基,或到约30个碱基,或到约 25个碱基。

[0268] 为了捕获,在合成探针和目标分子之间存在一个最小重叠区域。合成探针可以尽可能地短到比所需的最小重叠区域稍大。使用一个短的探针长度定向一个多态性区域的效果是将有重叠目标等位基因区域的更多分子。原始DNA分子的碎片状态也影响重叠目标等位基因读数的数量。一些DNA样本,如因发生在体内的生物过程已经破碎的血浆样本。然而,具有较长片段的样本得益于测序库制备和富集前的破碎。当探针和片段两个很短(约60~80bp)时,可实现几乎没有不重叠兴趣的临界区域放入序列读数最大特异性。

[0269] 在一个实施方式中,可以以最大限度地调整杂交条件,在初始样本中存在的不同的等位基因捕获中提高均匀。在一个实施方式中,降低杂交温度使在等位基因之间的杂交偏差的差异最小化。在本领域中已知的方法避免使用较低温度下的进行杂交,因为降低温度有增加探针杂交意想不到的目标的效果。然而,当目标是保留等位基因比率具有最大保真度时,使用较低的杂交温度的方法可提供最佳化的准确的等位基因的比率,尽管事实上,目前的技术教导远离这种方法。也可以提高杂交温度增加需要的目标与合成探针之间的更大重叠,以便只捕获具有大量重叠的目标区域的目标。在本发明的一些实施方式中,杂交温度从正常的杂交温度降低到约40℃,到约45℃,到约50℃,到约55℃,到约 60℃,到约65,或到约70℃。在一个实施方式中,杂交捕获探针可以这样设计,捕获探针具有与在位于多态性等位基因侧翼的区域发现的DNA互补的DNA,该捕获探针区域不是紧邻该多态性位点。相反,捕获探针可以这样设计,捕获探针设计成与位于目标多态性位点侧翼的DNA杂交,该捕获探针的区域从该捕获探针部分分离,将以范德华力与该多态性位点以一个等价于一个或少量碱基的长度的很小距离接触。在一个实施方式中,杂交捕获探针设计成与一个区域杂交,该区域位于多态性位点的侧翼但不会越过它,这可被称为一个侧翼捕获探针。侧翼捕获探针的长度小于约120个碱基,小于约110个碱基,小于约100个碱基,小于约90个碱基,以及可以小于80个碱基,小于约70个碱基,小于约60个碱基,小于约50个碱基,小于约40个碱基,小于约30个碱基或少于约25个碱基。由侧翼捕获探针定向的该基因组区域可被多态性座分离成,1,2,3,4,5,6,7,8,9,10, 11-20,或20个以上的碱基对。

[0270] 一个定向捕获的表述基于使用定向序列捕获的疾病筛查试验。定制定向序列捕获,如男鞋目前提供的AGILENT(SURE SELECT),ROCHE-NIMBLEGEN,或ILLUMINA。捕获探针可以是定制设计的,以确保捕获不同类型的突变。对点突变重叠点突变的一个或多个探针,应足以捕获并对突变测序。

[0271] 对于小的插入或缺失,重叠,突变的一个或多个探针可以是有效捕获并对含有突变的片段测序。在限制探针捕获效率之间,杂交的效率可能较低,通常被设计为针对参考基因组序列。为了确保捕获突变的片段,可以设计两个探针,一个匹配正常的等位基因,一个匹配突变等位基因。较长的探针可强化杂交。多个重叠的探针可强化捕获。最后,放置一个直接相邻但不重叠的探针,该突变可允许正常和突变等位基因有相对类似的捕获效率。

[0272] 对于简单的串联重复序列 (STR), 探针重叠这些高度可变的位点不可能很好地捕捉片段。为了强化捕获, 一个探针可以放在相邻的, 但不重叠的变异位点。然后该片段可以正常测序揭示 STR 的长度和组成。

[0273] 对于大的缺失, 目前用于外显子捕获系统中的一系列的重叠探针, 一个共同的方法可能起作用。然而, 使用这种方法, 可能难以确定个体是否是杂交的。定向和评估位于捕获区域内的 SNPs 可能揭示横跨显示一个体是一个载体的区域的杂合性缺失。在一个实施方式中, 有可能将非重叠的或单体探针横跨整个可能缺失区域, 和使用捕获片段的数目来衡量杂合性。在一个个体携带一个大的缺失的情况下, 二分之一数目的片段预计将可用于捕捉相对于不可删除 (二倍体) 参考基因位点。因此, 从缺失区域获得的读数应有大约一半是从一个正常的二倍体基因位点获得的。横跨潜在的缺失区域聚合和平均从多个的单体探针得到的测序读数深度可以增强信号及提高诊断的置信度。也可以结合这两种方法, 定向 SNPs 确定杂合性丢失和使用多个单体探针获得位于该位点潜在片段的数量的定量测量结果。其中一个或两个策略也可以与其他策略结合, 以更好地取得的相同的结果。

[0274] 如果在测试过程中, cfDNA 检测男性胎儿, 如由存在 Y 染色体片段所显示的, 在相同的测试中捕获和测序, 无论是 X-连接的显性突变, 其中母体和父亲是不受影响, 或母体不影响显性突变表示对胎儿有高层次的风险。检测到在同一个母体不受影响的基因中两个突变的隐性等位基因, 就意味着胎儿从父亲继承了突变等位基因以及一个潜在的源自母体的第二个突变的等位基因。在所有的情况下, 后续的羊膜穿刺或绒毛取样测试可能表明上述情况。

[0275] 一个基于疾病筛查试验的定向捕获可以与一个基于非整倍体的非侵入性产前诊断试验的定向捕获结合。

[0276] 有很多减少读数深度 (DOR) 变异的方法: 例如, 一个可能会增加引物的浓度的方法, 一个可以使用较长的定向扩增探针的方法, 或一个可以运行更多的 STA 周期 (如 25 个以上, 30 个以上, 35 个以上, 甚至 40 个以上) 的方法。

[0277] 定向 PCR

[0278] 在一些实施方式中, PCR 可用于定向基因组中特定的位置。在血浆样本中, 初始的 DNA 是高度破碎的 (通常小于 500bp, 平均长度小于 200bp)。在 PCR 中, 正向和反向引物, 必须对相同的片段退火以便使之扩增。因此, 如果片段很短, PCR 检测也必须扩增相对较短的区域。像 MIPS 一样, 如果多态性位置太靠近聚合酶结合位点, 可能会在源自不同等位基因导致扩增偏差。目前, 定向多态性的区域 PCR 引物, 如那些含有 SNPs, 通常是这样设计成引物的 3' 端将与紧邻的多态基的碱基杂交。

[0279] 在本发明的一个实施方式中, 在正向和反向 PCR 引物的 3' 端设计成与距离目标等位基因变异位置 (多态性位点) 相差一个或几个位置碱基杂交。的的目标等位基因的变体位置的基准。多态性位点 (SNPs 或以其他的) 和所设计的引物的 3' 端杂交的碱基之间的数量可以是一个碱基, 可以是两个碱基, 可以是 3 个碱基, 可以是 4 个碱基, 可以是 5 个碱基, 可以是 6 个碱基, 可以是 7 到 10 个碱基, 可以是 11 到 15 个碱基, 或者可以是 16 到 20 个碱基。正向和反向引物可设计成与距离多态性位点相差不同数目的碱基杂交。

[0280] PCR 检测可以大量生成, 但是当超出了大约 100 个 PCR 检测时, 不同的 PCR 检测之间的相互作用使得它们很难复用。各种复杂的分子方法可用于提高复用水平, 但它可能仍然

被限制为每个反应中少于100,可能是200,或可能是500个检测。大量的DNA样本可以被分割在多个子反应中,然后在重组之前测序。对于总体的样本或某些DNA分子的亚群的样本是有限,分割样本会引入统计噪声。在一个实施方式中,一个小数量或有限数量的DNA样本可以是指数量低于10皮克,介于10和100皮克之间,100皮克和1纳克之间,介于1纳克10纳克之间,或者介于 10纳克和100纳克之间。请注意,这种方法是对少量DNA特别有用的,而其他方法包括分割成多个池子的方法,可能会导致显著与统计噪声相关的问题,当运行于任何数量的DNA样本时,这种方法仍然提供最大限度地减少偏差的好处。在这些情况下,可以使用通用的预先扩增步骤以增加整体的样本数量。理想情况下,预扩增步骤应该没有明显改变的等位基因分布。

[0281] 在一个实施方式中,本发明的方法可以生成大量特定目标位点的PCR产品,具体地,1000到5000个位点,5000至10000 个位点或超过10000个位点,从有限的样本,如单细胞或体液 DNA,通过测序或一些其他的基因分型方法进行基因分型。目前,对超过5至10个目标进行多重PCR反应提出了一个重大的挑战,并常常受阻于引物副产品,如引物二聚体,和其他人工制品。当使用微阵列杂交探针检测目标序列时,引物二聚体和其他人工制品可能会被忽略,因为这些都不能检出的。然而,当使用测序作为检测的方法时,测序绝大多数的序列读数将对这些人工制品测序,而不是所需的样本中的目标序列。现有技术中描述的方法用于在一个反应中复用50或100个以上的反应,接着测序,通常会导致超过20%,通常会导致超过50%,在许多情况下超过 80%,以及在某些情况下,超过90%的脱靶序列读数。

[0282] 在一般情况下,进行一个样本的多(n个)目标定向测序个的(大于50,大于100,大于500,或大于1000),可以将样品分割到扩增个体目标的多个平行反应中。这可以在PCR多孔板,或者在商业平台,如FLUIDIGM ACCESS ARRAY (48每个样本8反应,在微流控芯片上)或DROPLET PCR by RAIN DANCE TECHNOLOGY (100到几千个目标)上进行。不幸的是,使用有限数量的DNA样本时这些分割和池的方法是有问题的,因为往往没有足够的基因组拷贝,以确保在每个孔中都有一个基因组中每个区域的一个拷贝。这是一个特别严重的问题,当多态性位点是被定向的,在多态性位点上等位基因的相对比例是必要的,因为由分割和分池引入的随机噪声导致对出现在初始DNA样本中的等位基因的比例的测量结果不准确。这里描述的是一种能够有效且高效地扩增许多PCR反应的方法,能够适用使用于只有数量有限的DNA可提供的情况。在一个实施方式中,该方法可应用于单细胞,体液,DNA的混合物如母体血浆中发现的游离的 DNA,活性组织切片检查,环境和/或法医样本的分析。

[0283] 在一个实施方式中,定向测序可能包括一个,多个或所有以下步骤。a) 用DNA片段两端上的配子序列生成和扩增的一个库。b) 在库扩增后分成多个反应。c) 用DNA片段两端上的配子序列生成和选择性地扩增一个库。d) 对每个目标使用一个目标特异性“正向”引物和一个标记特异性引物进行1000-10000-plex 选定目标的扩增。e) 在这个产品中进行第二次扩增,使用“反向”目标特异性引物和一个(或多个)对在第一轮作为目标特定正向引物一部分引入的通用标签具有特异性的引物。f) 对选定的目标进行1000-plex的有限数量的周期的预扩增。g) 将产品以多个等份划分并在单一反应中扩增划分的子池,例如,50到500-plex的子池,虽然都可使用单一方式。h) 合并并行的子池反应产品。i) 在这些扩增的引物中进行兼容标记的测序(部分或全部长度),使得产品被测序。

[0284] 高复用PCR

[0285] 本发明所公开的方法允许定向扩增超过一百到数以万计的源自血浆中基因组DNA的目标序列(如SNP位点)。扩增的样本可以是相对不含引物二聚体的产品,并在目标位点具有低的等位基因偏差。如果在扩增期间或之后的产品附加上了与测序兼容的配子,这些产品的分析可以通过测序进行。

[0286] 使用本领域中已知的方法的执行一个高复用的PCR扩增会导致过量于所需扩增产品和不适合用于测序的引物二聚体生成,通过消除形成这些产品的引物或通过硅片上进行引物筛选。可以根据经验减小这些产品生成。然而,检测数字越大,这个问题变得更加困难。

[0287] 一种解决方案是将5000个复用的反应分割成几个低复用的扩增,如100个50个复用或50个100个复用的反应,或用微流控或甚至将样本拆分割成单个的PCR反应。但是,如果样本DNA 是有限的,例如在从怀孕血浆进行非侵入性产前诊断中,应避免在复用反应之间分割样品,因为这将导致瓶颈。

[0288] 本发明所述的方法次是全局地第一扩增对样本的血浆DNA 进行全局性扩增,然后对每个反应以较为适中数目的目标序列将样本分成多个复用目标富集反应。在一个实施方式中,本发明所公开的方法可用于在多个位点优先富集DNA混合物中,该方法包括以下步骤中的一个或多个:从DNA混合物生成并扩增一个库,其中库中的分子具有连接在DNA碎片两端的配子序列。将扩增的库分割成多个反应,执行选定的目标的第一轮多重扩增,对每个目标使用一个目标特异性“正向”引物以及一个或多个通用特异性配子配子的“反向”引物第一轮普遍的“反向”引物。

[0289] 在一个实施方式中,本发明公开的一种方法进一步包括使用“反向”目标特异性引物和一个或多个对通用标签具有特异性的引物进行第二次扩增,其中通用标签是作为目标特异性引物的一部分在第一轮引入的。在一个实施方式中,该方法可能包括完全嵌套,半嵌套,半嵌套,一侧完全嵌套,一侧半嵌套,一侧半嵌套的PCR方法。在一个实施方式中,本发明的方法的用于在多个基因位点优先富集DNA混合物,该方法包括对选定的目标进行有限数量周期的复用预扩增,将产品分割成多个等份和单个反应的目标的扩增子池,合并并行的子池反应池产品。注意,这种方法可用于以一种方式执行定向扩增,导致低水平的等位基因偏差在50至500个位点,在500至5000个位点,在5000至50000 个位点,或甚至为5万至50万个基因位点的扩增。在一个实施方式中,引物携带部分或全部长度的测序的兼容标记。

[0290] 的工作流程可能需要(1)提取血浆DNA,(2)用两端都带有通用适配子的碎片制备片段库,(3)使用特定于适配子的通用引物扩增,(4)将扩增样本“库”分成多个等分样本,(5)对等分样本进行复用扩增(例如,每个目标的一个目标特异性引物和标记特异性引物约100个复用,1000个复用,或10000个复用),(6)合并一个样本的等分样本(7)为样本编条形码,(8)混合该样本并调整浓度,(9)对样本测序。工作流程可以包括多个子步骤包含一个已列出的步骤(如步骤(2)的制备库步骤可能三种酶的步骤(钝化尾端、DNA拖尾和适配子连接)和三个纯化步骤)。工作流程的步骤可以合并,拆分或以不同的顺序(如编条形码和合并样本)执行。

[0291] 重要的是要注意,库扩增可以在这样一种方单核小体DNA 片段如在孕妇的血液循环中发现的游离的胎儿DNA细胞(源自胎盘的)。需要注意的是PCR检测可以有标签,例如序列标签,(通常为15-25个碱基的截短形式)。复用后,合并PCR复用样本,然后由一个标记特

异性PCR完成(包括条形码)标记(也可以通过连接)。在复用中时也在相同的反应可以加入全测序标记。在第一个周期中,可以用目标特异性引物扩增目标,其后的标记特异性引物接管以完成SQ-配子序列。PCR引物可以不携带标记。测序标记可以通过连接附加到扩增产品。

[0292] 在一个实施方式中,高度复用的多重PCR后跟通过克隆测序评估扩增物质可能被用来检测胎儿非整倍。然而,鉴于传统的多重PCR只能同时评估高达50个位点,本发明所描述的方法可以用来同步评估超过50个位点同时进行,超过100个位点同时进行,超过500个位点同时进行,超过1000个位点同时进行,超过5000多个位点同时进行,超过10000个位点同时进行,超过50000个位点同时进行,超过10万个位点同时进行。实验已经表明,对高达,包括并超过10000个不同的位点同时进行评估,在单个反应中,以足够好的效率和特异性,进行非侵入性的产前非整倍体的诊断和/或具有精度高的拷贝数调用。在单一反应中,检测可以与完整的或部分的分离自母体血浆的cfDNA样本,或进一步加工的源自cfDNA样本衍生物结合。cfDNA或衍生工具也可以分割成多个平行的多重反应。最佳的样本分割和多重复用决定于交易的各种性能规格。由于材料的数量有限,将样本拆分成多个部分能引入采样噪声,处理时间,并增加了出错的可能性。相反,较高的复用可以导致更大量的假扩增和更大的不均等的扩增,这两者都可以降低测试性能。

[0293] 在本发明所述的方法的应用中有两个关键相关的因素是,在获得等位基因频率或其他测量结果的材料中的有限的原始血浆数量和原始分子数量。如果原来的分子数低于某一水平,随机抽样噪声变得显著,并且可能会影响测试的准确性。通常情况下,如果测量是对包含相当于每个目标位点500-1000原始分子的一个样本进行的,可以得到用于进行非侵入性的产前非整倍体诊断的,质量足够好的数据。有很多提高不同的测量值数目方法,例如增加样本体积的方法。施加到样本的每个操作也可能潜在地导致材料的损失。表征由各种操作产生的损失和避免损失是必不可少的,或有必要提高某种操作的收率从而避免因降低测试性能而造成的损失。

[0294] 在一个实施方式中,有可能在随后的步骤,通过扩增全部或部分的原始cfDNA样本以减轻潜在损失。有很多方法可用于扩增样本中全部遗传物质,提高可供用于下游程序的量。在一个实施方式中,连接一个不同的适配子,两个不同的适配子,或许多不同的适配子后,使连接介导的PCR(LM-PCR)的DNA片段进行PCR扩增。在一个实施方式中,多重置换扩增(MDA) ϕ -29聚合酶被用于等温扩增所有的DNA。在DOP-PCR和变异中,随机引物用于扩增的原始物质的DNA。每一种方法都有一定的特性,如扩增横跨所有代表基因组区域的均匀性,捕获和扩增原始DNA的效率,以及扩增性能是片段的长度的函数。

[0295] 在一个实施方式中,LM-PCR方法可与一个单一具有3-引物酪氨酸的异源双链化适配子一起使用。异源双链化适配子允许使用单个适配子分子,在第一轮PCR,该适配子分子在原始DNA片段的5-引物段和3-引物段可被转换成两个不同序列。在一个实施方式中,有可能利用粒析、或产品AMPURE、TASS,或其他类似的方法对扩增库分级。在连接之前,样本DNA可以被钝端,然后一个单一的腺苷碱基加入到3-引物端。DNA连接前用限制性内切酶或其他一些裂解方法裂解。在连接中,样品片段3-引物端的腺苷和互补的酪氨酸悬适配子可以提高连接效率。

[0296] 从时间角度看,减少扩增片段长于超过约200个基点,约300个基点,约400bp左右,约500个基点,或约1000个基点降低扩增,PCR扩增的延伸步骤可能会受到限制。由于母体血

浆中发现的较长的DNA几乎完全是产妇的,这可能会导致胎儿DNA的富集度在10-50%,并改善测试性能。一些反应运行使用的条件如所指定的市售试剂盒,导致成功连接少于10%的样本DNA分子。一系列反应条件的优化可改进这种连接至约70%。

[0297] 微型-PCR

[0298] 传统的PCR检测方法的设计导致不同胎儿分子的大量损失,但通过设计很短的PCR检测可以大大减少损失,这种PCR检测称为微型PCR检测。产妇血清中胎儿cfDNA高度破碎,碎片尺寸近似地以高斯模式分布,平均值为160个基点,标准差为15个基点,最小尺寸为约100个基点,最大尺寸为约220个基点的片段大小分布。涉及定向目标的多态性的碎片分布的开始和结束位置,而不一定是随机分布,在个体目标和所有目标之间总体上变化很大,在源自该位点各种碎片中,一个的特定目标基因的多态性位点可能占据从开始到结束的任何位置。请注意,术语微型PCR同样也可能指正常PCR,没有额外的约束或限制。

[0299] 在PCR过程中,扩增只发生于包括正向和反向引物位点的模板DNA片段。因为胎儿cfDNA片段短,考虑到两个引物位点都存在的可能性,包括正向和反向引物位点的胎儿片段具有长度L的可能性是扩增子的长度与片段长度的比率。在理想的条件下,检测,其中的扩增子是45,50,55,60,65,或70个基点将成功地分别扩增72%,69%,66%,63%,59%,56%的可用模板片段分子。扩增子的长度为5按两端的正向和反向引发位点之间的距离。短于典型的使用在本领域中已知的那些扩增子的长度,可能会在所需的多态性位点导致更有效的只需要短序列读数的测量。在一个实施方式中,相当大部分扩增子应该是小于100基点的,小于90基点的,小于80基点的,小于70基点的,小于65基点的,小于60个基点的,小于55基点的,小于50基点的或小于45个基点的。

[0300] 请注意,在现有技术已知的方法中,通常避免使用如本发明所述的那些短的检测,因为它们不是必需的,,通过限制引物的长度,退火特性,正向和反向引物之间的距离。在引物的设计上施加了相当大约束。

[0301] 还要注意,如果任何引物的3-引物端在大约1-6个多态性位点的碱基之内,偏置扩增是有潜力的。这种在初始聚合酶结合的位点的单碱基差异,可能会导致一个等位基因优先扩增,它可以改变观察到的等位基因频率和降低性能。所有这些约束使得它对确定成功扩增一个特定的位点的引物非常具有挑战性,此外,是设计大型成套的在相同的多重反应中兼容的引物。在一个实施方式中,内正向和反向引物的3'端被设计成与一个源自多态性位点上游侧的DNA区域杂交,并通过一个小数目的碱基与多态性位点数分开。理想的情况下,可能的碱基数是6和10个碱基之间,但同样也可能在4和15个碱基之间,在3和20个碱基之间,在2个和30个碱基之间,或1和60个碱基之间,实现基本相同的目的。

[0302] 多重PCR可以只包含单一的一轮PCR,其中所有目标被扩增,或者它可以包含一轮PCR,后面跟一轮或多轮巢式PCR或巢式PCR的某些变体。巢式PCR由包括随后的一轮或多轮PCR扩增,使用一个或多个新的对内部结合的引物,至少由一个碱基结合到前一轮中使用的引物。在随后的反应中,巢式PCR减少杂散扩增目标的数量,通过扩增只有那些来自前一轮的具有正确的内部序列的扩增产品。减少杂散扩增目标能提高可获得的有用的测量值的数目,特别是在测序过程中。

[0303] 巢式PCR通常需要设计完全处于之前引物结合点内部的引物,必要地提高扩增所需的最小DNA片段的尺寸。对于样本如母体血浆cfDNA,其中的DNA是高度破碎的。较大的检

测尺寸减少了不同cfDNA分子的数量,从此可以一个测量结果。在一个实施方式中,为了抵消这种效果,人们可以使用部分的嵌套的方法,其中一个或第二个两个引物都重叠从部分碱基内部延伸的第一结合位点,在最低限度地增加在整个检测尺寸的同时,实现额外的特异性。

[0304] 在一个实施方式中,设计了多重PCR检测池用于扩增潜在的杂合SNP或其它多态性或非多态性位点的一个或多个染色体,这些检测方法用于一单个反应中扩增DNA。PCR检测的数目可以在聚合酶链反应检测的数量可以是在50和200PCR检测之间的,200和1000检测之间的,1000和5000PCR检测之间的,或在5000和20000PCR检测之间的(分别是50到200个复用,200到1000个复用,1000到5000个复用,5000至20000个复用,超过20000个复用)在一个实施方式中,设计了约10000PCR检测(10000个复用的)多重复用池用于扩增染色体X,Y,13,18,和21以及1或2,这些检测方法在一单个反应中扩增cfDNA,该DNA是从一个血浆样本、绒毛膜绒毛样本、羊膜穿刺术样本、单个或少量细胞、其他体液或组织、癌症或其它遗传物质获得的。每个位点的SNP频率可以通过克隆或其他扩增子测序的方法确定。等位基因频率分布的统计分析或所有检测比率,可用于确定包含在该试验中的样本中是否含有一个或多个染色体一个三倍体。在另一个实施方式中,原始的cfDNA样本被分成两个样本进行平行的5000个复用的检测。在另一个实施方式中,原始的cfDNA样本被分成n个样本进行平行的($\sim 10000/n$)复用的检测,其中n介于2至12之间,或介于12和24之间,或介于24和48之间,或介于48和96之间。收集数据并用类似的、已经描述的方法分析。注意,此方法同样也适用于检测易位,缺失,重复,和其他染色体异常。

[0305] 在一个实施方式中,对目标基因组没有同源性的尾部也可被添加到任何引物的3-引物端或5-引物端。这些尾部便于后续的操作、程序、或测量。在一个实施方式中,尾部序列可以是相同的正向和反向的定向特异性引物。在一个实施方式中,不同的尾部可以用于正向和反向定向异性引物。在一个实施方式中,多个不同的尾部,可以用于不同的位点或基因位点组。某些尾部可以在所有位点位点的子集之间共享。例如,使用正向和反向尾部对应的任何现有测序平台所要求的正向和反向序列能跟随扩增进行直接测序。在一个实施方式中,尾部可以在所有的扩增了的目标之间用作常见的引发位点,用于添加其他有用的序列。在一些实施方式中,内部引物可以包含一个设计用来或与定向多态性位点的上游或与下游杂交的区域。在一些实施方式中,引物可以含有一个分子的条形码。在一些实施方式中,引物可包含一个设计用来允许PCR扩增的通用的引发序列。

[0306] 在一个实施方式中,创建一个10000个复用检测池,使得正向和反向引物具有对应于一种高通量测序仪,如HISEQ,GATIX,或ILLUMINA提供的MYSEQ所需的正向和反向序列的尾部,此外,包括测序尾部的5-端是一个额外的序列,可以使用添加作为引发位点在随后的PCR中以便对扩增子增加一个核苷酸条形码序列,能使多个样品在高通量测序仪单个通道进行多重测序。

[0307] 在一个实施方式中,创建一个10000个复用PCR检测池,反向引物具有对应于高通量测序仪所需的反向序列的尾部,用第一10000复用池检测扩增后,可以进行随后的PCR扩增,使用另一个经部分嵌套的正向引物(例如6嵌套碱)10000复用池,对所有目标进行扩增,一个与反向序列尾部对应的反向引物包含在第一轮中。后续一轮的部分嵌套的扩增只用一个定向特异性引物,并且一个通用引物限制需要的测定尺寸、降低采样噪声,但可大幅度降

低杂散扩增子的数量。测序标签,可以添加到附加的连接适配子和/或作为PCR探针一部分,使得标签是最终的扩增子的一部分。

[0308] 胎儿的组分会影响测试的性能。有很多的方法可用于富集母体血浆中发现的胎儿DNA组分。胎儿组分可以由先前所描述和已经讨论过的LM-PCR法以及通过定向去除母体长片段的方法提高。在一个实施方式中,在目标位点多重PCR扩增之前,可以执行一个额外的多重PCR反应以便有选择地去除长的和大的对应于在随后的多重PCR中要定向的位点的母体碎片。设计额外的引物,对距离多态型比预期无细胞胎儿DNA片段之间存在的距离较大的位点退火。这些引物在目标的多态性位点进行多重PCR之前可用于一个周期的多重复用PCR反应。这些远端引物分子或基团标记,可以允许标记的DNA片段的选择性识别。在一个实施方式中,这些DNA分子可以用生物素分子以共价键修饰,允许经过一个周期的PCR后含有这些引物的新形成的双链DNA的除去。在第一轮形成的双链DNA可能是起源于产妇。杂化材料的去除可以由所使用的磁链霉素抗生素蛋白珠完成。还有其他的标记方法能同样出色地起作用。在一个实施方式中,尺寸选择方法可用于富集更短链的DNA样本,例如小于约800个基点,小于约500个基点,或少于约300个个基点的碱基对。对于小片段的扩增,可以继续像往常一样进行。

[0309] 本发明中描述的微型PCR法能在单一反应中高度复用扩增和分析数百至数千或甚至上百万源自一个单一样本位点。同样,扩增的DNA的检测也可以通过使用条形码PCR在一个测序通道中复用,几十到数百个样本可以被复用在同一个测序通道。此复用检测已成功测试49复用,一个更高程度的复用也是可能的。

[0310] 实际上,在一个单一的测序运行中,允许数百个样本在成千上万的SNPs上进行基因分型。对于这些样本,该方法允许基因型和杂合性率的测定,以及同时测定拷贝数,这两者都可能被用于非整倍体检测的目的。这种方法特别适用于从母体血浆中所发现的游离的DNA检测怀孕中的胎儿的非整倍性。该方法可以作为鉴定胎儿性别、和/或预测的胎儿父亲身份的方法的一部分使用。它可以做作为突变剂量的方法的一部分使用。该方法可以用于任何数量的DNA或RNA,并且定向的区域可以是SNPs,其他多态区域,非多态的区域,以及它们的组合。

[0311] 在一些实施方式中,可以使用连接介导的通用PCR扩增DNA片段。连接介导的通用PCR扩增可用与扩增血浆DNA,然后,它可以被分为多个平行的反应。也可以用于优先扩增短的片段,从而富集胎儿的组分。在一些实施方式中,通过连接为片段片段增加标签就能检测较短的片段,使用更短的引物的目标序列的特定部分和/或在较高的温度退火,从而降低非特定的反应。

[0312] 发明所描述的方法可用于若干目的,其中有一个目标DNA集混合了许多污染DNA。在一些实施方式中,目标DNA和污染DNA可以是源自基因相关的个体。例如,一个胎儿的遗传异常(目标)可以从含有胎儿(目标)DNA和也含有母体(污染性)DNA的母体血浆进行检测;该异常包括全染色体异常(如非整倍体),部分染色体异常(如缺失,重复,倒置,易位),多核苷酸多态性(如STR基因位点),单核苷酸多态性,和/或其他遗传异常或差异在一些实施方式中,目标和有污染的DNA可以是来自同一个体的,但是其中的目标和有污染的DNA因一个或多个突变而不同,例如,在癌症的情况下,(see e.g. H. Mamon et al. Preferential Amplification of Apoptotic DNA from Plasma: Potential for Enhancing Detection

of Minor DNA Alterations in Circulating DNA. *Clinical Chemistry* 54:9 (2008)). 在一些实施方式中, DNA可能会发现在细胞培养(细胞凋亡)的上清液。在一些实施方式中,也能够诱导细胞凋亡的生物样本(如血液)用于随后的库制备,扩增和/或测序。为实现这一目的一些有利的工作流程和协议在本发明中的其他地方给出。

[0313] 在一些实施方式中,与目标DNA可能源自单细胞、源自含有小于一个目标基因组的拷贝的DNA样本,源自的数量的DNA、源自来自混合源的DNA(例如怀孕血浆:胎盘和母体DNA,癌症患者的血浆和肿瘤:健康和癌症之间的DNA的混合体,移植物等)、源自其他体液、源自细胞培养、源自培养上清、源自法医的DNA样本,源自古代的DNA样本(例如昆虫被困在琥珀中),源自DNA的其他样本,以及它们的组合。

[0314] 在一些实施方式中,可以使用一个短的扩增子的尺寸。短扩增子尺寸尤其适合与DNA碎片(参见,例如, A. Sikora, et al. Detection of increased amounts of cell-free fetal DNA with short PCR amplicons. *Clin Chem*. 2010 Jan; 56 (1):136-8.)

[0315] 使用短扩增子的大小可能会导致一些显著的好处。短扩增子的大小可能会导致优化的扩增效率。短扩增子的大小通常会产生更短的产品,因此,非特异性引物的机会就更少。更短的产品可以更密集地群集上测序流动池,集群会更小。请注意,本发明所述的方法可同样适用于较长的PCR扩增子。如果必要的话,扩增子的长度可能会增加,例如,当测序序列伸展较大。用146个复用定向扩增、100到200个基点长度的检测作为第一个步骤以嵌套方案在单细胞和基因组DNA上运行的实验,得出了呈阳性的结果。

[0316] 在一些实施方式中,本发明所述的方法可用于扩增和/或检测 SNPs,拷贝数,核苷酸甲基化, mRNA水平,以及其他类型的 RNA表达水平,其他的遗传和/或后生功能。这里所描述的微型-PCR方法可用于新一代测序,它可能与其他方法如微阵列方法,数字PCR计数,实时PCR,质谱光谱分析法等方法一起使用。

[0317] 在一些实施方式中,本文所描述的微型PCR扩增方法可用于为少数群体准确定量方法的一部分。它可用于使用峰值校准器进行绝对定量。通过非常深的测序方法它可以被用于突变/次要等位基因的定量,并可以以一个高度复用的方式运行。它可用于亲戚或祖先、人类、动物、职务或其他生物之间,标准的父系关系和身份确定测试,它可用于法医检测。它可用于快速基因分型和拷贝数分析(CN),对任何一种材料,例如:羊水和CVS,精子,产品的概念验证(POC)。它可用于单细胞分析,如胚胎活检样本的基因分型。使用微型-PCR通过定向测序,它可用于快速胚胎分析(小于一天,一天,或两天范围内的活检)。

[0318] 在一些实施方式中,它可用于肿瘤的分析:肿瘤活检通常是健康细胞和肿瘤细胞的混合物。定向PCR允许SNPs和位点深度测序至接近没有背景序列。它可用于对肿瘤DNA的拷贝数和杂合性损失分析。据说肿瘤DNA可存在于许多不同的体液或肿瘤患者的组织中。它可用于检测肿瘤的复发和/或肿瘤的筛查。它可用于种子的质量控制测试。它可用于育种,或钓鱼目的。请注意,任何这些方法同样可以很好地用于定向非多态性位点的倍体识别目的。

[0319] 描述构成本发明所公开的方法的一些基本法方法的文献包括:描述的一些背后的根本方法,本文公开的方法包括:(1) Wang HY, Luo M, Tereshchenko IV, Frikker DM, Cui X, Li JY, Hu G, Chu Y, Azaro MA, Lin Y, Shen L, Yang Q, Kambouris ME, Gao R, Shih W, Li H. *Genome Res*. 2005 Feb; 15 (2):276-83. Department of Molecular Genetics,

Microbiology and Immunology/The Cancer Institute of New Jersey,Robert Wood Johnson Medical School,New Brunswick,New Jersey08903,USA. (2) High-throughput genotyping of single nucleotide polymorphisms with high sensitivity. Li H, Wang HY,Cui X,Luo M,Hu G,Greenawalt DM, Tereshchenko IV,Li JY,ChuY,Gao R.Methods Mol Biol.2007;396 -PubMed PMID:18025699. (3) A method comprising multiplexing of an average of 9 assays for sequencing is described in: Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes.Varley KE,Mitra RD.Genome Res. 2008Nov;18(11):1844-50.Epub2008Oct10.需要注意的是本发明所公开的方法允许多路复用的数量级超过上述参考文献中的数量级。

[0320] 引物设计

[0321] 高度复用PCR往往导致非常高比例的DNA产品的生成,从而导致无用的副反应,如引物二聚体形成的生成。在一个实施方式中,可以从引物库中除去最有可能造成无用副反应的特定的引物从而给出一个将导致映射到基因组的更大比例的扩增DNA的引物库。去除有问题的引物的步骤,那就是,这些尤其可能形成二聚体的引物去接下来的测序分析竟然有意外地具有极高的 PCR复用水平。如在系统如测序系统中,其中因引物二聚体和/或其他有危害产品存在而性能显著降低,比其他描述的复用的系统大于10倍,大于50倍,大于100倍的高复用已经实现。请注意,这与基于探测的检测方法相反,例如微阵列,TaqMan,PCR等,过量的引物二聚体不明显会影响结果。还要注意的,在本领域一般的概念是,在同一个孔中复用PCR测序被限制在大约 100检测的范围。例如Fluidigm和Rain Dance,提供的平台可以平行反应的方式对一个样本执行48或1000PCR检测。

[0322] 有很多方法为库选择引物,使库中的非映射引物二聚体或其它引物有害产品的量最小化。经验数据表明,少量的“坏”的引物对大量的非映射引物二聚体的副反应负有责任。去除这些“坏”引物可以增加映射到目标位点的序列读数百分比。一种方法识别‘坏’的引物的方法是观察用定向扩增方法扩增的DNA测序数据,可以去除那些看到有最大频率的引物二聚体,以便给出一个明显不太可能导致不映射到基因组的副产品DNA的引物库。也有公开提供的程序,可以计算各种引物组合的结合能,并去除那些结合能最高的组合,也会给出一个明显不太可能导致副产品不映射到基因组的DNA的引物库。

[0323] 引物复用的大量对包含的检测施加相当大的约束。意外地相互作用的导致杂散的扩增产品。微型PCR的尺寸限制,可能会导致进一步的限制。在一个实施方式中,有可能开始于一个具有非常大的数量的潜在SNP目标(大于约500至100万之间),并试图设计设计扩增每个SNP的引物。其中,可以设计引物,用已公开的DNA双链形成的热力学参数,通过评估在所有可能的引物对之间杂散引物双链形成的可能性,有可能试图确定可能形成杂散的产品引物对。引物的相互作用可以通过与这种相互作用相关的得分函数分类,具有最差相互作用得分的引物被消除直到达到所需要的引物的数量。SNPs是杂合的情况是最有用的,有可能对检测列表分级,并选得杂合最兼容的检测。实验已经证实,具有相互作用得分高的引物最有可能引物二聚体。在高复用下,不可能的消除所有杂散相互作用,但是在硅片上去除具有最高相互作用得分的引物或引物对是必不可少的,因为他们可以主导整个反应,极大地限制了预定目标的扩增。我们已经完成了此过程,可创建高达10000个引物的复用引物组。由于这个程序,改进是巨大的,相比与从一个没有去除最差引物反应的10%相比,能对目标

产品进行80%以上,90%以上,95%以上,98%以上,甚至99%以上的扩增,如通过对所有PCR产品测序所确定的。部分半嵌套的方法结合使用时,如上所述,90%以上,甚至超过95%的扩增子可映射到目标序列。

[0324] 请注意,存在确定那一个PCR探针可能形成二聚体的方法,在一个实施方式中,分析已经使用一组非优化的引物扩增的一池DNA可以是足以确定有问题的引物。例如,分析可使用测序进行,以最大数目存在的二聚体被确定为是那些最有可能形成二聚体的,可以去除。

[0325] 此方法有许多潜在的应用,例如用于SNP基因分型,杂合率的测定,拷贝数测量,和其他定向测序应用。在一个实施方式中,该引物设计的方法可以与本文其他地方所描述的微型PCR方法组合使用。在一些实施方式中,引物设计的方法,可以作为一个庞大的多重复用PCR方法的一部分使用。

[0326] 在引物上使用标签可以降低引物二聚体产品的扩增和测序。标记的引物可以用来缩短必要的定向特异性序列到低于20个以下,15个以下,12个以下,甚至10个以下的碱基对。当目标序列是在引物结合位点内的碎片时,或者它可设计成引物的设计,这可能是偶然的标准引物设计。这种方法的优点包括:增加检测数量,可设计为某种最大扩增子长度的检测,以及缩短“无信息”引物的序列的测序。它也可以结合内部标记使用(见本文件的其他地方)。

[0327] 在一个实施方式中,多重复用定向PCR扩增过程中非生产性产品的相对数量可以通过提高退火温度减少。在用相同的标记作为目标特异性引物扩增库的情况下,退火温度可以提高到与基因组DNA相比,标签将继续与引物结合。在一些实施方式中,我们使用比以前报道的非常低的引物浓度,使用比其他地方报道的更长的退火时间。在一些实施方式中,退火的时间可能会超过10分钟,超过20分钟,超过30分钟,超过60分钟,超过120分钟,超过240分钟,超过480分钟,甚至更长的时间超过960分钟。在一个实施方式中,使用比以往的报道的较长的退火时间,允许较低的引物浓度。在一些实施方式中,引物的浓度低至50 nM, 20nM, 10nM, 5nM, 1nM, 以及低于1 μ M。这令人惊讶导致高度复用反应的鲁棒性能,例如1000复用的反应,2000复用的反应,5000复用的反应,10000复用的反应,20000复用的反应,50000复用的反应,甚至100000复用的反应。在一个实施方式中,扩增使用一个,两个,三个,四个或五个周期运行和较长的退火时间,随后的PCR循环和标记的引物使用更常见的退火时间。

[0328] 选择目标位置,可以从一池候选引物对的设计开始,创建一个引物对之间的潜在不良相互作用的热力学模型,然后使用该模型来消除与池中的其他设计不相容的设计。

[0329] 定向PCR变体-嵌套

[0330] 进行PCR时,有许多是可能的工作流程;在此描述了本文所公开方法的一些典型的工作流程。本文概述的步骤并不意味着排除其他可能的步骤,也不意味着任何此处所描述的步骤是该方法正常工作所需的步骤。大量参数的变化或其他修改在文献中已知的,并且可以不影响本发明的本质。下面给出一个特定的广义工作流程跟随多个可能的变体。该变体通常是指可能的第二次PCR反应,例如,不同类型的嵌套,可以用(步骤3)。重要的是要注意,变体可以在不同的时间或以不同的顺序工作而不是显式地在此描述。

[0331] 1. 样本中的DNA可以具有连接适配子,通常被称为库标记或附加的连接适配子标签(LTs),其中连接适配子包含一个普遍的引发序列,跟随一个普遍扩增。在一个实施方式

中,这可以通过使用旨在破碎后创建测序库的标准协议实现。在一个实施方式中,DNA样本可以是钝端的,然后一个A可以附加在3'末端。一个Y-适配子和T-外悬可以添加和连接。在一些实施方式中,可以使用除了A或T-外悬以外的其他的粘性末端。在一些实施方式中,可以添加其他适配子,例如环连接适配子。在一些实施方式中,适配子可具有设计用于PCR扩增的标记。

[0332] 2. 特定的目标扩增 (STA): 扩增前的数百到数千到数万甚至数十万个目标的预扩增可以被复用在反应中。STA通常是在从10到30个周期中运行,虽然也可在5到40个周期运行,在2到50个周期,甚至是从1到100个周期。引物可以是添加尾的,例如,一个简单的工作流,或避免大比例的二聚体测序。请注意,通常情况下,两个携带相同的标签引物的二聚体,将不能被扩增或有效地测序。在一些实施方式中,可以进行在1和10个循环的之间的PCR,在一些实施方式中,也可以进行10至20个循环的PCR,在一些实施方式中,也可以进行20至30个循环的PCR,在一些实施方式之间也可以进行30至40个循环的PCR,在一些实施方式中,也可以进行超过40个循环的PCR。扩增可以是线性扩增。PCR循环数进行优化,以导致一个最佳的读数深度(DOR)轮廓。为不同的目的不同DOR轮廓可能是需要的。在一些实施方式中,在所有检测之间,更均匀的读数分布是可取的;如果DOR某些检测的太小,随机噪声过高的数据是没有用的,而如果读数的深度过高每个额外的读数边际效用相对较小。

[0333] 引物的尾部可从通用标记库中提高DNA片段的检测。如果该库标记和引物尾部含有一个同源的序列,则杂交可以得到改善(例如,降低熔化温度(T_m)),如果只有一部分引物目标序列在DNA样品片段中,引物可以延长。在一些实施方式中,可以使用13个或更多的目标特异性碱基对。在一些实施方式中,可以使用10至12目标特异性碱基对。在一些实施方式中,可以使用8至9的目标特异性碱基对。在一些实施方式中,可以使用6至7的目标特异性碱基对。在一些实施方式中,STA也可以对预扩增DNA进行,例如MDA,RCA,其他的全基因组扩增,或适配子介导的通用PCR。在一些实施方式中,STA可以对富集或贫化一定序列和群体的样本进行,例如大小选择,定向捕获,定向降解。

[0334] 3. 在一些实施方式中,有可能进行第二次多重复用的PCR或引物延伸反应以提高特异性,并减少不良产品。例如,全嵌套,半嵌套,半嵌套,和/或细分成更小测定池的平行反应,是提高特异性的可以使用的所有技术。实验表明,样本分割成三个400复用反应比一个1200复用的具有完全相同引物的反应可导致产品DNA具有更大特异性。同样,实验表明,样本分割成4个2400复用的反应一个9600复用的具有完全相同引物的反应可导致产品DNA具有更大特异性。在一个实施方式中,有可能用相同和相反的方向性的目标特异性和标记的特异性引物。

[0335] 4. 在一些实施方式中,有可能扩增一个DNA样品(稀释,纯化的或以其他方式),该样本是通过使用标记的特异性引物和“普遍扩增”,例如,扩增许多或所有的预扩增的和标记的目标,由STA反应生成的。引物可能包含额外的功能序列,例如条形码,或一个完整的适配子序列,该适配子序列是在高通量测序平台上测序所必需的。

[0336] 这些方法可用于分析任何的DNA样本,当DNA样本特别小,或当它是一个DNA源自多于一个个体,如在母体血浆中的情况下放入样本时,这些方法是特别有用。这些方法也可以用于DNA样本,如一个单一的或少量的细胞、基因组DNA、血浆DNA、扩增的血浆库、扩增的细胞凋亡上清液库、或其他混合DNA样本。在一个实施方式中,这些方法可用于不同遗传体

质的细胞可能存在在一个单一个体的情况下,如癌症或移植。

[0337] 协议变体(变体和/或增加上述工作流程)

[0338] 直接复用的微型-PCR:在图1中所示的多个标签的引物与目标序列的特异性靶扩增(STA)。101表示在X点102还具有多态性位点的双链DNA表示连接添加的适配子普遍扩增的双链DNA。103表示已普遍扩增PCR引物杂交的单链DNA。图104表示在最终的PCR产品。在一些实施方式中,STA可以做100多个,200多名,500多名,1000多名,2000多名,5000元以上,1万多,2万多,超过50000个,超过10万或更多超过20万的目标。在随后的反应,标签特异性引物扩增所有目标序列和加长的标签,包括所有必要的序列进行测序,包括样本的指标。在一个实施方式中,引物可能不被标签,或只有特定的引物可被标记。传统的适配子连接测序适配子可能会添加。在一个实施方式中,最初的引物上可以带有标记。

[0339] 在一个实施方式中,引物设计的扩增DNA的长度出乎意料的短。先前的技术表明人员的艺术证明本领域熟练的普通的人的典型设计是100基点的扩增子。在一个实施方式中,扩增子可被设计为成小于80基点的。在一个实施方式中,扩增子可被设计为小于70基点的。在一个实施方式中,扩增子可被设计为小于60基点的。在一个实施方式中,扩增子可被设计为小于50个基点的。在一个实施方式中,扩增子可被设计为小于45基点的。在一个实施方式中,扩增子可被设计为小于40基点的。在一个实施方式中,扩增子可被设计为小于35基点的。在一个实施方式中,扩增子可被设计为介于40和65基点的。

[0340] 进行一个实验,使用该协议使用1200个复用扩增。基因组DNA和怀孕血浆两个两者都使用约70%的映射到目标序列的序列读数。详细细节在本文档中其他地方给出。没有设计和及检测选择的1042个复用的测序导致在>99%的序列中出现引物二聚体产品。

[0341] 顺序PCR:在STA之后,对该产品的多个等份用相同的引物和降低复杂度的扩增池以并行的方式进行扩增,第一轮扩增可以给予足够的材料来拆分。这种方法特别适合小样本,例如那些约6到100皮克,约100皮克到1纳克,约1毫微克至10毫微克,或约10毫微克至100毫微克。把1200复用分成三个400个复用进行协议。测序读数映射从单独的1200复用的约60%至70%增加至95%以上。

[0342] 半嵌套的微型PCR:(参见图2)在STA1之后,执行一个第二次STA1,包括一个内部嵌套的正向引物(103B,105B)和一个(或几个)有特定标签的反向引物的多重集(103A)。101表示在具有多态性位点的感兴趣的双链DNA在X点。102表示双链DNA具有为普遍扩增而添加的连接适配子。103表示已被用正向引物B和杂交的反向引物A普遍扩增的单链DNA。104表示103的PCR产品。105表示104的产品,具有杂交的嵌套正向引物B和已经是103和104之间发生的PCR的部分分子的反向标签A。106表示最终的PCR产品。使用此工作流程通常有大于95%的序列映射到预定的目标。嵌套引物可重叠外部正向引物序列,但引入了额外的3'-端碱基。在一些实施方式中,有可能使用1和20之间的额外的3'碱基。实验已经表明,在1200复用中使用9个或更多的额外的3'碱基,设计工作得很好。

[0343] 完全嵌套的微型-PCR(见图3)在STA步骤1之后,有可能执行一个第二次的多重复用PCR(或平行复用,降低复杂度的PCR反应),使用两个再有标签(A,A,B,b)的嵌套引物。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示双链DNA,具有为普遍扩增而添加的连接适配子。103表示已被用正向引物B和杂交的反向引物A普遍扩增的单链DNA。104表示103的PCR产品。105表示104的产品,具有杂交的嵌套正向引物B和嵌套反向A。106表示

最终的PCR产品。在一些实施方式中,有可能使用两套完整的引物集。使用完全嵌套微型PCR协议的实验执行146复用扩增单个或三个细胞而不使用添加普遍连接适配子和扩增的102步骤。

[0344] 半嵌套的微型PCR:(参见图4)有可能使用在片段端部有一个适配子的目标DNA。进行STA包括一个正向引物(B)和一个(或几个)特定标记的反向引物(A)复用集。第二个STA可以使用通用标签特定的正向引物和目标特定的反向引物执行。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示双链DNA,具有为普遍扩增而添加的连接适配子。103表示已被用正向引物B和杂交的反向引物A普遍扩增的单链DNA。104表示已被用反向引物A和连接适配子标签引物LT普遍扩增的103的PCR产品。105表示源于104与正向引物B杂交的产品。106表示最终的PCR产品。在本工作流程中,针目标特定的正向和反向引物用于不同的反应,从而降低了反应的复杂性,并防止正向和反向引物二聚体的形成。请注意,在这个例子中,引物A和B可以被认为是第一个引物,引物'a'和'b'可被认为是内引物。这种方法对直接PCR是一个很大的改进,因为它和直接PCR一样好,但它避免了引物二聚体。经过第一轮半嵌套协议之后,通常看到~99%的非目标DNA,但是,经过第二轮之后,通常有一个很大的改进。

[0345] 三重半嵌套的微型-PCR:(参见图5),有可能使用在片段端部有一个适配子的目标DNA。进行STA包括:正向引物(B)和一个(或几个)标签特定的反向引物(A)和(a)的复用集。使用通用标签特定的正向引物和目标特定的反向引物进行第二个STA。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示双链DNA,具有为普遍扩增而添加的连接适配子。103表示已被用杂交的反向引物A普遍扩增的单链DNA。104表示已被用反向引物A和连接适配子标签引物LT普遍扩增的源自103的PCR产品。105表示源自104产品,具有杂交的正向引物B。106表示使用反向引物和正向引物B扩增的源自105的PCR产品。107表示源自106的产品,具有杂交的反向引物'a'。图108表示最终的PCR产品。请注意,在这个例子中,引物'a'和B可以被认为是内引物,并且引物A可被认为是第一引物。任选地,A和B都可以被认为是第一个引物,以及'a'可以被认为是内引物。反向和正向引物的指定是可以切换的。在本工作流程中,目标特定的正向和反向引物用于不同的反应,从而降低了反应的复杂性,并防止正向和反向引物二聚体的形成。这种方法对直接PCR是一个很大的改进,因为它和直接PCR一样好,但它避免了引物二聚体。经过第一轮半嵌套协议之后,通常看到~99%的非目标DNA,但是,经过第二轮之后,通常有一个很大的改进。

[0346] 单边嵌套的微型PCR:(参见图6),有可能使用在片段端部有一个适配子的目标DNA。也用一个嵌套正向引物和连接适配子标签作为反向引物的复用集进行STA。然后,进行第二个STA,使用一个嵌套正向引物和一个通用的反向引物集。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示双链DNA,具有为普遍扩增而添加的连接适配子。103表示已被用杂交的反向引物A普遍扩增的单链DNA。103表示已普遍扩增正向引物A杂交的单链DNA。104表示已被用反向引物A和连接适配子标签引物LT普遍扩增的源自103的PCR产品。104是103的PCR产品扩增正向引物和连接适配子标签反向引物LT。105表示源自104的产品,具有杂交的正向引物B。106表示最终的PCR产品。这种方法可以在第一和第二STA,通过使用重叠的引物,检测比标准PCR更短的目标序列。该方法典型地是对以经历了上面的步骤STA1-添加通用标签和扩增的DNA的样本进行;两个嵌套引物只在一侧,另一侧使用该的

标签。该方法对凋亡上清和妊娠血浆库执行。使用此工作流程大约有60 %的序列映射到预期的目标。请注意,包含反向适配子序列的读数不被映射,所以,如果那些包含反向适配子序列的读数不被映射,这个数字预计是更高的。

[0347] 单边微型PCR:有可能使用在片段端部有一个适配子的目标 DNA(见图7)。进行STA,使用一个正向引物和一个(或几个)标记特定的反向引物复用集。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示具有为普遍扩增而添加的连接适配子的双链DNA。103表示具有杂交的正向引物A的单链DNA。104表示已被用正向引物A和连接适配子标签引物LT普遍扩增的源自103的PCR产品,也是最终的PCR产品。这种方法可以比标准PCR检测更短的目标序列。然而,它可以是相对非特异性的,因为只使用一个目标特定的引物。该协议实际上是半单边嵌套的微型PCR。

[0348] 反向半嵌套微型PCR:有可能使用在片段端部有一个适配子的目标DNA(参见图8)。进行STA,使用一个正向引物和一个(或几个)标记特定的反向引物复用集。101表示具有多态性位点的感兴趣的双链DNA在X点。102表示具有为普遍扩增而添加连接适配子的双链DNA。103表示具有杂交的反向引物B的单链DNA。104表示已被用反向引物B和连接适配子标签引物LT普遍扩增的源自103的PCR产品。105表示具有杂交的正向引物A和内反向引物'b'的源自104的PCR产品。106表示使用正向引物A和反向引物'b'扩增的源自105的PCR产品,也是最后的PCR产品。这种方法可以比标准的PCR检测更短的目标序列。

[0349] 也还可以有更多的变体,简单的迭代或组合上述方法,如双嵌套PCR,其中使用了三套引物。另一种变体是一边是嵌套另一边是半嵌套的微型-PCR,其中STA也可以进行,用一组复用嵌套的正向引物和一个(或几个)标记特异性的反向引物。

[0350] 需要注意的是,在所有这些变体中,正向引物和反向引物的特征是可以互换的。请注意,在一些实施方式中,嵌套的变体可以同样很好地运行,而不需包括附加的适配子标签和一个通用的扩增步骤的初始库制备。请注意,在一些实施方式中,可能包括额外附加的PCR轮数,额外的正向和/或反向引物和扩增步骤,如果进一步增加对应于目标位点的DNA分子的百分比是可取的,这些额外的步骤可能特别有用。

[0351] 嵌套工作流程

[0352] 有许多方法执行扩增,用不同程度的嵌套和不同程度的复用。在图9中,给出了一个具有一些可能的工作流程图。请注意,使用1000个复用的PCR只意味着是一个例子,这些流程图同样适用于其他程度的多重复用。

[0353] 环形连接适配子

[0354] 当添加通用标签适配子,例如,处于创建一个测序为目的库,也有一些方法连接适配子。一种方法是钝端DNA样本,进行A-接尾,连接有一个T型外悬的适配子。也有很多其他的方法连接适配子。也有一些可以连接的适配子。例如,可使用一个Y-适配子,该适配子包含两条DNA链,其中一条链有一个双链区域,一个区域由一个正向引物指定,其中的另一个由一个在第一个链链上与该双链区域互补的双链区域指定,一个区域具有反向引物,该双链区域,当退火时,可以含有一个为了达到连接到有A外悬的双链DNA的目的的T-外悬。

[0355] 在一个实施方式中,适配子可以是一个环形的DNA其中末端区域是互补的,并且环型区域中包含一个的正向引物标记的区域(LFT),一个反向引物标记的区域(LRT),和一个两者之间的分裂位点(参见图10)。101是指有双链、钝端目标DNA。102是指有A-接尾的目标

DNA。103是指有T外悬‘T’和分裂点‘Z’的环形连接适配子。104是指添加环形连接适配子的目标DNA。105是指目标DNA,带有添加的可在在分裂点的裂解的连接适配子。LFT是指连接适配子正向标签,及LRT指连接适配子反向标签。互补区域可能以一个最终可能在T外悬结束,或可用于连接到目标DNA其他功能。分裂点可以是一系列由3UNG切割的尿嘧啶,或一个由限制性内切酶或其他方法、或只是一个基本的扩增识别和切割的序列。这些适配子可以用于任何库制备,例如,用于测序。这些适配子可以与本文所描述的任何其他方法结合使用,例如微型PCR扩增方法。

[0356] 内部标记的引物

[0357] 当使用测序确定一个存在于给定多态性位点的等位基因时,序列读数通常从引物结合位点(a)的上游开始,然后到多态性位点(X)。标签通常被配置成如图11中所示,左边。101指的是单链目标DNA带有感兴趣的‘X’上的多态性位点和添加标签‘b’的引物‘a’。为了避免非特异性杂交,引物结合位点(目标DNA区域与‘a’互补)的长度通常是18至30基点。序列标签‘b’通常是约20个基点,在理论上这些可以是任何长度超过约15个基点的序列,尽管很多人使用测序平台公司出售的引物序列。‘a’和‘X’之间的距离‘d’可以至少是2基点,以避免等位基因偏差。当用本发明公开的方法或其他方法执行多重PCR扩增时,仔细的引物设计对避免过多的引物与引物之间的相互作用是必要的,以,允许‘d’,‘a’和‘X’之间可距离的窗口‘d’可能有很大不同:从2个基点至10个基点,从2个基点至20个基点,从2个基点至30个基点,甚至从2个基点至30个基点以上。因此,当使用图11所示引物配置时,左边,测序读数必须是一个至少为40基点的,以或的足够长的读数时间来测量多态性位点,根据‘a’和‘d’长度,序列读数可能需要达到60或75基点。一般,序列读数越长,的时间越长,一个给定数量的序列读数花费的成本和时间就越高,因此,最大限度地减少所需的读出长度,可节省时间和金钱。此外,平均来说,在该读数中前期读取的碱基比后期读取的更准确,降低必要的序列读数长度也能提高对多态性区域的测量的准确度。

[0358] 在一个实施方式中,称为内部标记的引物,引物的结合位点(a)的分割成多个片段(a’,a”,a’”....),序列标签(b)DNA片段上,DNA片段位于两个引物结合点上,如在图11中的103所示。这种配置允许的测序器进行较短序列的读数。在一个实施方式中,一个a’+a”应该是至少约18个基点,并且可以是长达30,40,50,60,80,100或超过100基点。在一个实施方式中,a”应该是至少约6个碱基,在一个实施方式中为约是8到16基点的。所有其他因素相同的情况下,使用内部标记引物序列的长度可以减少读数所需的长度至少6个基点,多达8个基点,10个基点,12个基点,15个基点,甚至多达20个或30bp基点。这可能会导致在一个显著的金钱、时间和准确性上的优势。一个内部标记的引物的例子在图12中给出。

[0359] 带有连接适配子结合区域的引物

[0360] 片段化DNA的存在的一个问题是,因为它长度是短的,多态性,靠近一条DNA链的端部的机会高于长链的(例如,101,图10)。由于一个多态性PCR的捕获需要一个在多态性两端长度合适的引物结合位点。大量的DNA链与目标多态性将由于引物和目标结合位点之间不充足的重叠而被错过。在一个实施方式中,与目标DNA101可以有附加的连接适配子102,目标引物103,就可以有一个区域(cr),该区域是与附加到设计的连接区域(a)的上游侧的连接适配子标签(1t)互补的,(见图13);这样,在连接区域(101区域,与a互补)小于通常杂交所需的18基点的情况下,在引物上与库标签互补的该区域(cr)能够提高结合能到一个PCR

可以继续进行的点。请注意,任何由于较短的结合区造成的特异性的丢失,可以由其他有足够长目标结合区的 PCR引物得到弥补。请注意,本实施方式可以与在直接PCR或任何本发明所描述的其他方法结合使用,如嵌套的PCR、半嵌套的PCR、半嵌套PCR、单边嵌套的套的PCR、或者半(semi)或半(hemi)嵌套的PCR,或其他PCR协议。当使用的测序数据确定倍性用结合的分析方法,包括比较观察的等位基因数据和各种假设预期预期的额等位基因分布,每个额外的具有低读数深度的等位基因读数将比具有高读数深度的等位基因读数产生更多的信息。因此,理想的情况下,人会希望看到深度均匀(DOR)的读数,其中每个位点将有类似数量有代表性的序列读数。因此,尽量减少DOR方差是可取的。在一个实施方式中,有可能通过增加退火时间,以降低DOR方差的系数(这可以被定义为DOR的标准偏差/平均的DOR)。在一些实施方式中,退火温度可以长于2分钟,长于4分钟,长于10分钟,长于30分钟,以及长于一小时,甚至更长时间。由于退火是一个平衡的过程中,随着退火时间的增加,改善DOR偏差的改善是限制的。在一个实施方式中,增加引物的浓度可能会降低DOR方差。

[0361] 诊断盒

[0362] 在一个实施方式中,本发明公开的,包括能够部分地或完全地执行本发明中描述的任何一种方法的诊断盒。在一个实施方式中,诊断盒可设在医生的办公室、医院实验室、或任何适当的有利于病人护理的位置。该诊断盒可以以全自动的方式运行所有的方法,或者该诊断盒可能需要一个或多个步骤由技师手动完成。在一个实施方式中,该诊断盒可以是能够至少对母体血浆中测得的基因型数据进行分析。在一个实施方式中,该诊断盒可以与一些反法连接,发送诊断盒上测量的基因型数据到外部计算设备,然后可以分析这些基因型数据,并且还可能生成一个报告。该诊断盒可以包括一个机器人单元,能够从一个容器转移水溶液或液体样本到另一个容器。它可包括多种固体和液体试剂。它可包括一个高通测序器。也可包括一台计算机。

[0363] 引物试剂盒

[0364] 在一些实施方式中,可以配制一个试剂盒,其包括多个设计的引物以实现本发明中描述的方法。引物可以是外正向和反向引物,内正向和反向引物如本发明所公开的,它们可能是被设计为对试剂盒中其他引物具有低结合亲和力引物,如本发明在引物设计一节中所公开的,他们可以杂交捕获探针或预环化的探针如在有关章节中所描述的,或它们的某种组合。在一个实施方式中,可以配制一种试剂盒,设计与本发明所公开的方法一起使用的,用于确定一个在妊娠中胎儿的目标染色体的倍性状态,该试剂盒包括多个内正向引物和任意的内反向引物,任意的内正向引物和外反向引物外,其中,每个引物设计成与DNA的紧接一个目标染色体上的和任意的附加的染色体上的多态性位点的上游和/或下游区域杂交。在一个实施方式中,引物试剂盒可以与在本文其他地方描述的与诊断盒结合使用。

[0365] DNA的组成

[0366] 当对胎儿和母体血液的混合物中测得的测序数据进行信息学分析以确定对胎儿的基因组有关的信息时,例如胎儿的倍性状态,在一组位点上测量等位基因分布是有利的。不幸的是,在许多情况下,例如,当从母体血样的血浆中发现的DNA混合物试图确定胎儿的倍性状态时,可提供的DNA的数量直接在混合物中用于测量混合物中具有良好精度的等位基因分布是不充分的。

[0367] 在这些情况下,DNA混合物的扩增将提供足够数量的DNA分子,因此可以测得所需

的具有良好精度的等位基因分布。然而,目前通常用于在DNA进行测序的扩增的扩增方法常常很偏,这意味着,他们不同时以相同的量扩增多态性位点的等位基因。一个很偏的扩增可以导致等位基因分布与在原混合物等位基因分布是相当不同。对于大多数用途,不需要对存在于等位基因多态性位点的相对量进行高度精确的测量。与此相反,在一个本发明的实施方式中,特别针对富集多态性等位基因并保留等位基因比率的扩增或富集方法是有利的。

[0368] 本文中描述的许多方法可用于在DNA样本的多个位点,以一种最小化偏差的方式优先富集等位基因。某些例子使用圆形探针定向多个位点,其中预环化探针的3'末端和5'末端设计成与距离目标等位基因的多态性位点一个或几个位置的碱基杂交的。另一种方法是使用PCR探针,其中PCR探针的3'末端设计成与距离目标等位基因的多态性位点一个或几个位置的碱基杂交的。另一种方法是使用一个分裂池的方法生成DNA混合物,其中。优先富集的基因位点用低等位偏差富集而没有直接多重复用的弊端。另一种方法是使用一种杂交捕捉的方法,其中捕获探针是这样设计的,捕获探针的区域设计成与位于目标多态性位点两侧的基因杂交,并由一个或少量碱基从多态性位点分开。

[0369] 在从一组的多态性位点测量等位基因分布来确定个体倍性状态的情况下。在为遗传测量制备样本时,在DNA样本中保留等位基因的相对数量是可取的。这种制备包括WGA扩增,定向扩增,选择性富集技术,杂交捕获技术,圆化的探针或其它方法意味着,扩增DNA量的和/或选择性地增强对应的一定等位基因的DNA分子的存在。

[0370] 在本发明的一些实施方式中,有一组DNA探针设计成定向位点,其中的位点有最大的次要等位基因频率。在本发明的一些实施方式中,有一组探针设计成定向位点,其中的位点有胎儿的最大似然率,其中的胎儿在那些位点有一个高度提供信息的 SNP。在本发明的一些实施方式中,有一组探针设计成定向位点,其中的探针是对一个给定群体亚群进行了优化。在本发明的一些实施方式中,在本发明的一些实施方式中,有一组探针设计成定向位点,其中的探针对给定的群体亚群的混合物进行了优化。在本发明的一些实施方式中,在本发明的一些实施方式中,有一组探针设计成定向位点,其中的探针是对给定一对父母进行了优化,其中父母是来自不同的具有不同的次要等位基因频率轮廓的群体的亚群。在本发明的一些实施方式中,有一个环化的DNA链,其中包括至少一个已退火成一个胎儿来源DNA片断的碱基对。在本发明的一些实施方式中,有一个环化的DNA链,其中包括至少一个已退火成一个胎盘来源DNA片断的碱基对。在本发明的一些实施方式中,有一个环化的DNA链,在环化的同时至少一些核苷酸退火为胎儿来源的DNA。在本发明的一些实施方式中,有一个环化的DNA链,在环化的同时至少一些核苷酸退火为胎盘来源的DNA。在本发明的一些实施方式中,有一组探针,其中的一些探针定向单一的串联重复序列,以及一些探针定向单一的核苷酸多态性。在一些实施方式中,被选择的位点是为了非侵入性的产前诊断的目的。在一些实施方式中,探针用于非侵入性的产前诊断的目的。在一些实施方式中,使用一种方法定向位点,该方法包括圆形探针、MIPs、杂交探针捕获、在SNP阵列上的探针、或它们的组合。为了非侵入性的产前诊断的目的对位点测序。

[0371] 在序列的相对信息量较大,且与父背景相关的情况下,遵循使包含一个父背景是已知的SNP的序列读数的数量最大化可以使在混合样本上的这组测序读数的信息量最大化。在一个实施方式中,包含一个父背景是已知的SNP的测序读数的数量可以通过使用定

量PCR优先扩增特定序列增强的。在一个实施方式中,包含一个父背景是已知的SNP的测序读数的数量可以通过使用圆形探针(例如,MIP)优先扩增特定序列增强的。在一个实施方式中,包含一个父背景是已知的SNP的测序读数的数量可以通过使用杂交捕获方法(例如,SureSelect)优先扩增特定序列增强的。可以使用不同的方法增强包含一个父背景是已知的SNP序列读数的数量。在一个实施方式中,定位可通过延长连接,连接但不延长,杂交捕获或PCR完成。

[0372] 在一个碎片基因DNA样本中,一小部分DNA序列唯一地映射到个体染色体;其他的DNA序列可以在不同的染色体上发现。注意,血浆中发现的DNA,无论是母体或胎儿起源的通常是成碎片的,其长度通常长500个基点。在一个典型的基因组样本中,大约有3.3%的可映射序列将映射到13号染色体,2.2%的可映射序列将映射到18号染色体,1.35%的可映射序列将映射到21号染色体,4.5%的可映射序列将映射在女性的X染色体;2.25%的可映射序列将映射到X染色体(男性)和0.73%的可映射序列将映射到Y染色体(男)。这些染色体最有可能是胎儿中的非整倍体。另外,在短序列中,大约20个序列中有一个将含有一个SNP,用包含在dbSNP上的SNPs,这个比例可以是给定的较高的,因此有可能有许多未被发现的SNPs。

[0373] 在本发明的一个实施方式中,定向方法可用于在一个映射到给定染色体的DNA样本中提高DNA分数,使得该分数显著高于上面列出的对于典型的基因组样本的百分比。在本发明的一个实施方式中,定位方法可以用于提高DNA样本中的一部分DNA,使得包含一个SNPs的序列的百分比显著大于可能会在通常的基因组样本发现发现的。在本发明的一个实施方式中,在一个用于产前诊断的母体和胎儿的DNA混合物,定向方法可用于从一个染色体或从一组SNPs对DNA定向。

[0374] 注意到,一种用于通过计数映射到一个怀疑的染色体读数数量和将其与映射到一个参考染色体读数数量进行比较,以及使用一个假设,即在怀疑染色体上的大量读数对应于在该染色体上的胎儿的一个三倍体,来确定胎儿非整倍性的方法已被报道(美国专利7888017)。那些用于产前诊断的反法不能针对任何形式的应用,也不能描述针对产前诊断的应用。

[0375] 在对混合样本测序中通过利用定向方法有可能用较少的序列读数实现一定程度精度。该精读可能指灵敏度,它可以是指特异性,或者它可以是指它们的某种组合。理想水平的精度可以在90%和95%之间,它可以是在95%和98%之间,它可以是在98%和99%之间,它可以是99%和99.5%之间,它可以是99.5%之间和99.9%,它可以是99.9%和99.99%之间,它可以是99.99%和99.999%之间,它可以是99.999%和100%之间。准确度在95%以上的水平可以被称为高准确度。

[0376] 在现有技术中有多种公开的方法,演示如何从母体和胎儿的DNA混合样本确定胎儿的倍性,例如:GJ.W.Liao et al.Clinical Chemistry2011;57(1)pp.92-101。这些方法集中于沿着每条染色体的数千个位置。沿着一条可被定向的染色体的位置的数量,同时仍然对胎儿具有一个高准确度的倍性测定,对于从一个DNA的混合样本中给定的序列读数的数目是出乎意料的低。在本发明的一个实施方式中,一个准确的倍性确定可以通过使用定向测序,使用任何定向方法实现,例如,定量PCR,配体介导的PCR,其他的PCR方法,杂交捕获,或圆化探针,其中,定向所需的沿染色体位点的数目可以在5000和2000位点之间,可

以在 2000和1000位点之间,可以是在1000和500位点之间,可以是在500和300位点之间,可以是在300和200个位点,可以是在 200和150位点之间,可以是在150和100位点之间,可以是在 100和50个位点,可以是在50和20位点之间,可以是在20和 10位点之间。理想情况下,它可以是在100和500位点之间。通过定向较少的位点数目和执行出乎意料低少量的序列读数可实现高水平的准确性。读数的数量可以介于100百万和50百万读数之间,读数的数量可以介于50百万和20百万读数之间,读数的数量可以介于20百万和10百万读数之间,读数的数量可以介于10百万和500万读数之间,读数的数量可以介于500万和200 万读数之间,读数的数量可以介于200万和100万读数之间,读数的数量可以介于100万和50万读数之间,读数的数量可以介于50万和20万读数之间,读数的数量可以介于20万和10万读数之间,读数的数量可以介于10万和50000读数之间,读数的数量可以介于50000和20000读数之间,读数的数量可以介于 20000和10000读数之间,读数的数量可以低于10000读数。较少数目的读数需要输入大量的DNA输入。

[0377] 在一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到13号染色体上的序列的百分比是大于4%,大于5%,大于6%,大于7%,大于8%,大于9%,大于10%,大于12%,大于15%,大于20%,大于 25%,或大于30%。在一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到18 号染色体上的序列的百分比是大于3%,大于4%,大于5%,大于6%,大于7%,大于8%,大于9%,大于10%,大于12%,大于15%,大于20%,大于25%,或大于30%。在一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到21号染色体上的序列的百分比是大于2%,3%以上,更大的比4%,大于5%,大于6%,大于7%,大于8%,大于9%,大于10%,大于12%,大于15%,大于20%时,大于25%,或大于30%。在一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到X号染色体上的序列的百分比是大于6%,大于7 %,大于8%,大于9%,大于10%,大于12%,大于15%,大于20%,大于25%,或大于30%。在一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到Y号染色体上的序列的百分比是大于1%,大于2%,大于3%,大于4%,大于5%,大于6%,大于7%,大于8%,大于9%,大于10%,大于12%,大于15%时,大于20%,大于25%,或大于30%。

[0378] 在一些实施方式中,所述一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到包含至少一个单核苷酸多态性的染色体的序列的百分比是大于0.2%,大于0.3%,大于0.4%,大于0.5%,大于0.6%,大于0.7%,大于0.8%,大于0.9%,大于1%,大于1.2%,大于1.4%时,大于1.6%,大于1.8%,大于2%,大于2.5%,大于3%,大于4%,大于5 %,大于6%,大于7%,大于8%,大于9%,大于10%,大于 12%,大于15%或大于20%,并且,其中的染色体取自染色体组13,18,21,X,或Y。在本发明的一些实施方式中,有一种组合物含有胎儿来源和孕产妇来源的DNA的混合物,其中,唯一地映射到包含源自一组单核苷酸多态性中的至少一个单核苷酸多态性的染色体的序列的百分比是大于0.15%,大于0.2%,大于0.3%,大于0.4%,大于0.5%,大于0.6%,大于0.7%,大于0.8%,大于0.9%,大于1%的,大于1.2%,大于1.4%,大于1.6%,大于1.8%,大于2%,大于2.5%,大于3%,大于 4%,大于5%,大于6%,大于大于8%,7%,9%以上,大于 10%,大于12%,大于15%,或大于20%,并且,其中的染色体取自染色体组13,18,21,X and Y,以及在一组单

核苷酸多态性中的单个单核苷酸多态性数目是在1和10之间,在10和20之间,在20和50之间,在50和100之间,在100和200之间,在200和500之间,在500和1000之间,在1000和2000之间,在2000和5000之间,在5000和10000之间,在10000和20000之间,在20000和50000之间,在50000和100000之间。

[0379] 在理论上,在扩增的每个循环中,存在的DNA量增加一倍;但是,事实上,扩增的程度略低于两倍。在理论上,扩增,包括定向扩增,将导致DNA混合物的无偏差的扩增,但是,事实上,不同的等位基因与其他等位基因先比往往被扩增到不同的程度。当DNA被扩增时,等位基因偏差程度通常会随扩增步骤数目的增加而增加。在一些实施方式中,本发明所述的方法包括以一个较低水平的等位基因偏差扩增DNA。由于等位基因偏差与每个增加的周期合成,可以通过计算整体偏差的 n 次方根确定每个周期的等位基因偏差,其中 n 是富集程度以2为底的对数。在一些实施方式中,有一种组合物,包含一个第二DNA混合物,其中第二DNA混合物已被在源自第一DNA混合物的多个多态性位点优先富集,其中的富集程度至少为10,至少为100,至少为1000,至少为10000,至少为100000或至少为1000000,其中在第二DNA混合物中每个位点的等位基因比率不同于在第一DNA混合物中那个位点的等位基因比率,他们之间的因数是,平均来说,小于1000%,500%,200%,100%,50%,20%,10%,5%,2%,1%,0.5%,0.2%,0.1%,0.05%,0.02%,或0.01%。在一些实施方式中,有一种组合物,包含一个第二DNA混合物,其中第二DNA混合物已被在源自第一DNA混合物的多个多态性位点优先富集,其中对多个多态性位点的每个周期的等位基因偏差是,平均来说,小于10%,5%,2%,1%,0.5%,0.2%,0.1%,0.05%或0.02%。在一些实施方式中,多个多态性位点包括至少10个位点,至少20个位点,至少50个位点,至少100个位点,至少200个位点,至少500个位点,至少1000个位点,至少2000个位点,至少5000个位点,至少10000个位点,至少20000个位点,或至少50000个位点。

[0380] 最大似然估计

[0381] 本领域已知的,用于检测生物现象或医疗状况的存在或不存在的大多数方法包括使用一个单一的假设抑制试验,测量条件与度量,如果度量的一侧,给定的阈值时,条件是存在的,而在另一侧的阈值的度量值跌倒,条件是不存在的。单假设拒绝测试看起来只有空和替代假设之间时,空分布决定。在不考虑备用分布,不能每个假设观测数据估计的可能性,因此不能计算在电话会议上的信心。因此,用一个单一的假设抑制试验,得到回答是或否的具体情况相关联的信心没有感觉。

[0382] 在一些实施方式中,本发明所公开的方法使用最大似然法能够检测生物现象或医疗状况存在或不存在。可以调整,适合于每一种情况下,这是一个实质性的改进的方法,使用一个单一的假设作为调用存在或不存在的条件的阈值的抑制技术。这是特别相关的诊断技术,旨在确定遗传数据可从胎儿和母体的DNA存在于自由浮动发现母体血浆中的DNA的混合物在一个孕期胎儿非整倍体的存在或不存在。这是因为,在血浆中的胎儿DNA的小部分来自分数的变化,最佳阈值,用于调用非整倍体与整倍体变化。由于胎儿的分数下降,与非整倍体相关联的数据的分布变得越来越相似的分布相关联的数据带整倍体的。

[0383] 最大似然估计法使用的分布相关的每一种假设,估计每个假设条件的数据的可能性。在这些条件概率可以被转换为一个假设呼叫和信心。同样,最大后验估计方法使用相同的条件概率最大似然估计,也包括人口先验时,选择最好的假设,并确定信心。

[0384] 因此,使用的最大似然估计 (MLE) 技术,或密切相关的最大后验概率 (MAP) 的技术,给两个优点,第一,它增加一个正确的呼叫的机会,并且还允许为每个计算的置信打电话。在一个实施方式中,选择对应于具有最大概率的假设的倍性状态进行用最大似然估计或最大后验概率估计。在一个实施方式中,揭示一种方法,用于确定涉及目前在本领域中已知的任何方法,它用一个假设抑制技术和重新在于它采用了最大似然估计或MAP技术的妊娠中的胎儿的倍性状态。通过应用这些技术,可以显著改善的方法的一些例子可以在美国专利8008018,美国专利7888017,或美国专利7332277中找到。

[0385] 在一个实施方式中,描述了一种方法,用于确定胎儿非整倍体的存在或不存在的母体血浆样本包括胎儿和母体的基因组DNA,该方法包括:获得母体血浆样本,测量发现血浆样本中的DNA片段,具有高通量测序的序列映射到染色体上,并确定序列的数目读数映射到每个染色体,计算血浆样本中的胎儿DNA的小部分,计算目标染色体是将期望的量的预期分布目前,如果,如果所述第二目标染色体整倍体和一个或多个预期如果该染色体非整倍体,将期望的分布,使用胎儿组分和序列的数目读数映射到一个或多个参考染色体预期整倍体,使用最大似然估计或 MAP确定的分布中的哪一个是最可以是正确的,从而表明胎儿非整倍体的存在或不存在的。在一个实施方式中,测量从血浆中可能涉及的DNA进行大规模并行鸟枪法测序。在一个实施方式中,测量从血浆样本中的DNA可能涉及已优先富集,例如通过有针对性的扩增,在多个多晶型或非多态性位点的DNA测序。的多个位点可被设计为针对一个或少数的怀疑非整倍体染色体和一个或少数的参考染色体。的优先富集的目的是要增加的测序读数的数目,倍性测定信息。

[0386] 倍性识别信息法

[0387] 本发明所描述的是一种用于确定给定的序列数据的胎儿的倍性水平方法。在一些实施方式中,这个序列数据可在一个高通量测序器上测量。在一些实施方式中,序列数据在源于游离从母体的血液中分离的DNA上测量,其中,所述游离DNA包括一些母系起源的DNA,以及一些胎儿/胎盘来源的DNA。本节将描述本发明的一个实施方式,其中胎儿的倍性水平是假设被分析为未知的并将从数据中估测的混合物中的胎儿DNA的分数来测定。这部分还将描述一个实施方式,其中混合物中的胎儿DNA的分数(“胎儿分数”)或胎儿DNA的百分比可以通过另一种方法来测量,并且被假定为在确定胎儿的倍性水平中是已知的。在一些实施方式中,只使用母体血液样本本身的基因分型的测量值来计算计算胎儿分数,这是胎儿和母体DNA的混合物中的。在一些实施方式中,该分数也可以用母体的测得的或已知的基因型和/或父亲的测得的或已知的基因型计算。在另一个实施方式中,胎儿的倍性水平可以完全是基于与假设为二倍体的对照染色体胎儿DNA的计算分数相比的待测染色体的的计算出的胎儿DNA的分数来测量。

[0388] 在优选的实施方式中,假设,对于一个特定的染色体,我们 $\tilde{N}$ SNP位点,我们观察和分析N SNPs,对于其我们有:

[0389] • 设置NR游离DNA序列测量值 $S = (s_1, \dots, s_{NR})$ 。由于这种方法利用SNP测量,对应于非多态性位点的所有序列数据均可以忽略不计。在一个简化的版本中,在那里每个SNP我们有(A,B)计数,其中A和B对应于存在于一个给定的基因座上的两个等位基因,S可以被写为: $S = ((a_1, b_1), \dots, (a_N, b_N))$,其中 a_i 是在SNP i计数, b_i 是SNP i计数,以及 $\sum_{i=1:N} (a_i + b_i) = NR$

[0390] • 父亲数据组成的

[0391] 源自SNP微列阵或其他基于基因分型平台的强度的基因型：

[0392] 母亲 $M = (m_1, \dots, m_N)$ ，父亲 $F = (f_1, \dots, f_N)$ ，其中 $m_i, f_i \in (AA, AB, BB)$ 。

[0393] 和/或序列数据测量：NRM母体测量值 $SM = (sm_1, \dots, sm_{nrm})$ ，NRF 父亲测量值 $SF = (sf_1, \dots, sf_{nrf})$ 。类似上述的简化，如果我们对每个 SNP $SM = ((am_1, bm_1), \dots, (am_N, bm_N))$ ， $SF = ((af_1, bf_1), \dots, (af_N, bf_N))$ 有 (A,B) 计数

[0394] $SM = ((am_1, bm_1), \dots, (am_N, bm_N))$ ， $SF = ((af_1, bf_1), \dots, (af_N, bf_N))$

[0395] 总的来说，母体，父亲的子数据表示为 $D = (M, F, SM, SF, S)$ 。注意需要亲本数据，并增加了该算法的准确性，但不是必要的，尤其是父亲的数据。这意味着，即使在母体和/或父亲的数据的情况下，有可能获得非常精确的拷贝数结果。

[0396] 这是可能获得最好的拷贝数估计 (H^*) 通过将似然所考虑的所有假设 (H) 似然数据最大化 $LIK(D|H)$ 。尤其是，有可能通过使用联合分布模型和制备样本上测得的等位基因计数来确定每个倍性假设的相对概率，并利用这些相对概率来确定最有可能是正确的假设，如下所示：

[0397] $H^* = \underset{H}{\operatorname{argmax}} LIK(D|H)$

[0398] 同样的，后验假设似然给定数据的可写为：

[0399] $H^* = \underset{H}{\operatorname{argmax}} LIK(D|H) * \text{先验概率}(H)$

[0400] 其中先验概率 (H) 是分配给每个基于模型的设计和先验知识的假设 H 的先验概率。

[0401] 这也可能是使用先验找到最大后验估计：

[0402] $H_{MA} = \underset{H}{\operatorname{argmax}} LIK(D|H)$

[0403] 在一个实施方式中，可能被考虑的拷贝数的假说是：

[0404] • 单倍体：

[0405] 母体 $H10$ (一个拷贝源自母体)

[0406] 父体 $H01$ (一个拷贝源自父体)

[0407] • 二倍体： $H11$ (母体和父体各一份拷贝)

[0408] • 单纯三体综合征，被认为是无交叉：

[0409] 母体： $H21_{\text{配对的}}$ (两个完全相同的拷贝源自母体，一个拷贝源自父体)， $H21_{\text{不配对的}}$ (两个拷贝均源自母体，一个拷贝源自父体)

[0410] 父亲： $H12_{\text{配对的}}$ (一个拷贝源自母体，两个完全相同的拷贝来自父亲)， $H12_{\text{不配对的}}$ (一个拷贝源自母体，两个拷贝源自父亲)

[0411] 复合三体综合征，允许交叉 (使用联合分布模型)：

[0412] $\emptyset$ 母体 $H21$ (两个拷贝来自母体，一个拷贝来自父亲)，

[0413] $\emptyset$ 父亲 $H12$ (一个拷贝来自母体，两个拷贝来自父亲)

[0414] 在其它实施方式中，其他倍体水平，如零倍体 ($H00$)，单亲二倍体 ($H20$ 和 $H02$)，四倍性 ($H04, H13, H22, H31$ 和 $H40$) 均可以考虑。

[0415] 如果有没有交叉，每个三体综合征，无论起源是减数分裂，减数分裂I，还是减数分裂II，都将是配对或不配对的三体之一。由于交叉，真正的三体综合征是通常是两者的结

合。首先,说明用于简单假设的假设似然法。然后说明用于复合假设的假设似然法,以交叉的方式将个体的SNP相结合。

[0416] 用于简单假设的LIK(D|H)

[0417] 在一个实施方式中,LIK(D|H)可以被确定为简单假设,如下所示。LIK(H),对于简单的假设H,LIK(H),在整条染色体假设H的对数似然可以以个体SNP的对数似然的总和来计算,假设已知的或派生的子本分数比照。在一个实施方式中,有可能从数据得出比照。

$$[0418] \quad \text{LIK}(D|H) = \sum_i \text{LIK}(D|H, cf, i)$$

[0419] 这个假设不假设任何SNPs之间的联系,因此不采用联合分布模型。

[0420] 在一些实施方式中,对数似然可以以每个SNP为基准被确定。对于一个特定的SNP i,假设胎儿倍性假设H%以及胎儿DNA 比照,D定义为观察到的数据的对数似然

[0421]

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

[0422] 其中m可能是真母体基因型,F可能是真父体基因型,其中 $m, f \in \{AA, AB, BB\}$ 和c可能是子基因型给定假设H.特别是,单倍体 $\in \{A, B\}$,二倍体 $\in \{AA, AB, BB\}$,三体 $\in \{AAA, AAB, ABB, BBB\}$ 。

[0423] 基因型先验频率: $p(m|i)$ 是母体基因型m的SNPI一般先验概率,根据SNP I的已知一般频率,记为 pA_i 。特别是

$$[0424] \quad p(AA|pA_i) = (pA_i)^2, p(AB|pA_i) = 2(pA_i) * (1-pA_i), p(BB|pA_i) = (1-pA_i)^2$$

[0425] 父亲的基因型的概率 $p(f|i)$,可以以相似的方式来确定。

[0426] 真子概率: $p(c|m, f, H)$ 是得到真子本基因型=C的可能性,给定亲体M,F,假定假设H,它可以很容易地计算出。例如,对于H11,H21配对和H21不配对,下面给出 $p(c|m, f, H)$ 。

[0427]

		p(c m,f,H)											
m	f	H11			H21 配对的				H21 不配对的				
		AA	AB	BB	AAA	AAB	ABB	BBB	AAA	AAB	ABB	BBB	
AA	AA	1	0	0	1	0	0	0	1	0	0	0	
AB	AA	0.5	0.5	0	0.5	0	0.5	0	0	1	0	0	
BB	AA	0	1	0	0	0	1	0	0	0	1	0	
AA	AB	0.5	0.5	0	0.5	0.5	0	0	0.5	0.5	0	0	
AB	AB	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0	0.5	0.5	0	
BB	AB	0	0.5	0.5	0	0	0.5	0.5	0	0	0.5	0.5	
AA	BB	0	1	0	0	1	0	0	0	1	0	0	
AB	BB	0	0.5	0.5	0	0.5	0	0.5	0	0	1	0	
BB	BB	0	0	1	0	0	0	1	0	0	0	1	

[0428] 数据的似然性: $P(D|m, f, c, H, i, cf)$ 是给定的数据D基于SNP i 的概率,真母体基因型m,真父体基因型F,真子本基因型C,假设H以及子本分数比照。它可以细分为母体,父体和子本的数据的概率如下:

[0429] $P(D|m, f, c, H, cf, i) = P(SM|m, i) P(M|m, i) P(SF|f, i) P(F|f, i) P(S|m, c, H, cf, i)$

[0430] 母体SNP阵列数据似然性:将SNP i 的母体SNP阵列基因型数据与真实基因型 m 比较,假设SNP阵列基因型是正确的,只是

$$[0431] \quad P(M|m, i) = \begin{cases} 1 & m_i = m \\ 0 & m_i \neq m \end{cases}$$

[0432] 母体序列数据的似然性:SNP i 母体序列数据的概率,在计数 $S_i = (a_{mi}, b_{mi})$ 的情况下,不含多余的噪音或偏差,二项式概率定义为 $P(SM|m, i) = P_{X|m}(a_{mi})$,其中 $X|m \sim \text{Binom}(p_m(A), a_{mi} + b_{mi})$ 定义为

[0433]

m	AA	AB	BB	A	B	未识别
$p(A)$	1	0.5	0	1	0	0.5

[0434] 父亲数据的可能性:类似的公式适用于父亲的数据的可能性。注意,如果没有亲本数据,尤其是父亲的数据,就不可能确定孩子的基因型。例如,如果没有父亲的基因型数据 F ,可能只是使用 $p(F|f, i) = 1$ 。如果没有父亲的序列数据 SF ,可能只是使用 $P(SF|f, i) = 1$ 。

[0435] 在一些实施方式中,所述方法包括建立预期的等位基因计数的每个倍性假设在染色体上的多个多态性位点的联合分布模型;这里描述的是这样的端部。游离胎儿DNA数据的可能性: $P(S|m, c, H, cf, i)$ 是关于SNP i 的游离胎儿DNA序列数据的概率是我真正的孩子,给定真母体基因型 m ,孩子的基因型 C ,子本拷贝数假设 H ,假设子本分数比照。事实上,实际上关于SNP i 序列数据 S 的概率,给定SNP i $\mu(m, c, cf, H)$ 的A含量的真实概率。

[0436] $P(S|m, c, H, cf, i) = P(S|\mu(m, c, cf, H), i)$

[0437] 计数,其中 $S_i = S_i = (a_i, b_i)$,所涉及的数据没有额外的噪音或偏差,

[0438] $P(S|\mu(m, c, cf, H), i) = P_x(a_i)$

[0439] 其中 $X \sim (p(A), a_i + b_i)$, $P(A) = \mu(M, C, \text{比照} H)$ 。在更复杂的不知道精确定位和每SNP (A,B)计数的情况下,是综合二项式的结合 $P(S|\mu(M, C, H, CF), i)$ 。

[0440] 真实A满意率: $\mu(M, C, H, CF)$,该母体/孩子混合物中的SNP i 的真实A满意率,假设真实母体基因型= m ,真实子本基因型= C ,总子本分数= cf ,被定义为

$$[0441] \quad \mu(m, c, cf, H) = \frac{\#A(m) * (1 - cf) + \#A(c) * cf}{n_m * (1 - cf) + n_c * cf}$$

[0442] 其中 $\#A(g) = A$ 的基因型数 g ,是母体的体性,假设 H 下的孩子的倍性(1是单体性,2单二体性,3是三体性)。

[0443] 使用用于复合假说的联合分布模型:LIK(D|H)

[0444] 在一些实施方式中,所述方法包括建立预期的等位基因计数的每个倍性假设在染色体上的多个多态性位点的联合分布模型;这里描述的是这样的端部。在许多情况下,三体性通常不是纯粹的配对或不配对,由于交叉,因此在本节的结果为复合假设H21(母体三体综合征)和H12(父亲三体综合征)的推导,结合配对和不配对的三体综合征,说明可能有交叉。

[0445] 在三倍体症情况下,如果没有交叉,三倍体症将只是配对或不配对的三倍体症。配对的三倍体症是孩子从一个亲本那里继承了两个相同的染色体片段的拷贝。而不配对的三

倍体症是孩子从亲本那里继承了每个同源染色体片段的一个拷贝。由于交叉,染色体的某些片段可能具有配对的三体性,而其他部分可能具有不配对的三体性。本节中描述的是如何建立一组的等位基因的杂合率的联合分布模型,也就是说,在一个或多个假设的多个位点的等位基因的预期数。

[0446] 假设SNP I , $LIK(D|H_m, i)$ 为适合配对假说 H_m , $LIK(D|H_u, i)$ 是合适的为无与伦比的假说 H_u , 和 $pc(i) = \text{SNPs之间交叉概率} i-1$ 安迪。然后计算出一个完整的可能性:

[0447] $LIK(D|H) = \sum_E LIK(D|E, 1:N)$

[0448] 其中 $LIK(D|E, 1:N)$ 以假设 E 为似然度结果, 对于 SNPs $1:N$ 。 E = 最近SNP的假设, $E \in (H_m, H_u)$ 。递归的, 可以计算:

[0449] $LIK(D|E, 1:i) = LIK(D|E, i) + \log(\exp(LIK(D|E, 1:i-1)) * (1 - pc(i)) + \exp(LIK(D|\sim E, 1:i-1)) * pc(i))$

[0450] 其中 $\sim E$ 是假设而不是 E (非 E), 其中所考虑的假设是 H_m 和 H_u 。特别是, 可以基于 1 到 $(i-1)$ SNPs 的似然度, 采用同样的假设和没有交叉, 或相反的假设和一个交叉, 乘以 SNP i 的似然度来计算 $1:i$ SNPs 的似然度

[0451] 对于 SNP1, $i=1$, $LIK(D|E, 1:1) = LIK(D|E, 1) \dots$

[0452] 对于

[0453] SNP2, $i=2$, $LIK(D|E, 1:2) = LIK(D|E, 2) + \log(\exp(LIK(D|E, 1)) * (1 - pc(2)) + \exp(LIK(D|\sim E, 1)) * pc(2))$,

[0454] 对 $i=3:N$ 等,

[0455] 在一些实施方式中, 可以确定子分数。子分数可以参考源于子本的DNA混合物中的序列的比例。上下文中的非侵入性的产前诊断, 子本分数可以指母体血浆中的序列的比例来源于胎儿或部分含胎儿的基因型的胎盘。可以指由母体血浆制备的, 可以富集胎儿DNA的DNA样本中的子本分数。因此, 确定DNA样本中的子本分数的一个目的是用于使用一种可以对胎儿进行倍性识别的算法, 子本分数可以指为非侵入性的分析产前诊断进行测序分析的任何DNA样本。

[0456] 本发明中所提出的算法是假定一种已知的子分数情况下进行非侵入性的产前非整倍体诊断法一部分, 这可能并非总是如此。在一个实施方式中, 可能通过对选定的染色体的二倍性概率最大化的方式发现最有可能的子分数, 在有或没有亲本数据存在条件下。

[0457] 特别是, 假设 $LIK(D|H_{11}, cf, chr) =$ 如上所述的对数似然, 二倍性假设, 号染色体 chr 上的和子本分数 cf 。对于选定的染色体 $Cset$ 中 (通常为 $1:16$), 假设是整倍体, 完全似然值是:

[0458] $LIK(cf) = \sum_{chr \in Cset} Lik(D|H_{11}, cf, chr)$

[0459] 最大可能的子本分数是 (cf^*) 被导出如 $cf^* = \arg\max_{cf} LIK(cf)$ 。

[0460] 有可能使用任何染色体集, 也可以推导子本分数而无需假设对照染色体上的整倍性。使用这种方法, 有可能确定在下列情况下的任何的子本分数: (1) 具有亲本阵列数据, 和母体血浆鸟枪法测序数据, (2) 具有亲本阵列数据, 和母本血浆定向测序数据, (3) 基于亲本和母体血浆的定向测序数据; (4) 基于母亲和母本血浆分数的定向测序数据; (5) 基于母体血浆分数的定向测序数据; (6) 其他亲本和子本分数测量值组合。

[0461] 在一些实施方式中, 该信息学方法可合并数据缺失, 这可能会导致在倍性更高精

度倍性的测定。在本发明的其他部分,已假定获得一个A的概率是一个真母体基因型、真子本基因型、在该混合物中的子本分数和子本拷贝数的直接函数。也有可能母体或孩子的等位基因是缺失的,父本对基因组Illumina公司数据的缺失率可表示为 d_{pg} ,父本对序列数据的缺失率可表示为 d_{ps} 以及子本对序列数据的缺失率可表示为 d_{cs} 。在一些实施方案中,母体的缺失率可能被认为是零,而子本的缺失率相对较低,在这种情况下,结果都不会受到缺失的严重影响。在一些实施方式中,等位基因缺失概率可能是足够大的以至于对预测倍性识别产生显著影响。对于这样情况,等位基因的缺失已被纳入到这里的算法:

[0462] 家长SNP阵列数据缺失:母体的基因组数据M,假设缺失后的基因型是 m_d ,那么

$$[0463] \quad P(M|m, i) = \sum_{m_d} P(M|m_d, i) P(m_d|m)$$

[0464] 其中, $P(M|m_d, i) = \begin{cases} 1 & m_i = m_d \\ 0 & m_i \neq m_d \end{cases}$ 同前, $P(m_d|m)$ 是给出真正的基因型m在可能得缺失后的基因型可能性 m_d ,定义如下,对于缺失率d

m	m _d					
	AA	AB	BB	A	B	未识别
[0465] AA	(1-d) ²	0	0	0 2d(1-d)	0	d ²
AB	0	(1-d) ²	0	d(1-d)	d(1-d)	d ²
BB	0	0	(1-d) ²	0	2d(1-d)	d ²

[0466] 类似的公式适用于父体SNP阵列数据。

[0467] 亲本序列数据丢失:对于母体的序列数据SM

$$[0468] \quad P(SM|m, i) = \sum_{m_d} P_{X|m_d}(am_i) P(m_d|m)$$

[0469] 其中 $P(m_d|m)$ 定义为如在游离数据可能性章节中所定义的。 $P_{X|m_d}(am_i)$ 来自异构二项式分布的概率定义为如之前父本数据可能性一节中所定义的。一个类似的公式用于母本序列数据。

[0470] 游离DNA序列数据丢失:

$$[0471] \quad P(S|m, c, H, cf, i) = \sum_{m_d, c_d} P(S|\mu(m_d, c_d, cf, H), i) P(m_d|m) P(c_d|c)$$

[0472] 其中 $P(S|\mu(m_d, c_d, cf, H), i)$ 定义为如在游离数据可能性章节中所定义的。。

[0473] 在一个实施方式中, $p(m_d|m)$ 是所观察到的母亲基因型 m_d 的概率,给定的真母亲基因型为m,假定缺失率为 d_{ps} ,而 $p(c_d|c)$ 是观察到的胎儿基因型 c_d 的概率,给定的真子本基因型为c,假定缺失率为 d_{cs} . 如果 nA_T =在真基因型c中等位基因A的数目, nA_D =在观察到的基因型 c_d 中等位基因A的数目,其中 $nA_T \geq nA_D$, 同样地 nB_T =在真基因型c中等位基因B的数目, nB_D = 在观察到的基因型 c_d 中等位基因B的数目,其中 $nB_T \geq nB_D$ 已及 d =缺失率,那么

[0474]

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

[0475] 在一个实施方式中该基于信息学的方法可以合并随机的和一致性性的偏差,以理想的话说,在序列计数的数量中没有一个 SNP一致性采样偏差或随机噪声(处理二项式分布偏差外),特别是,在SNP i ,对母亲的基因型 m ,真子本基因型 c 和子本分数 cf ,在SNP i 上(A+B)的读数集中 $X=A$ 的数量, X 就像 $X \sim \text{Binomial}(p, A+B)$,其中 $p = \mu(m, c, cf, H) = A$ 内容的真概率。

[0476] 在一个实施方式中,信息学方法可合并随机偏差。这是常有的情况下,假设在测量中有一个偏置,以便得到一个A在此SNP 的概率等于到 q ,这是有一个点与 p 在如上所定义的不同。 P 与 q 有多大不同取决于测量过程的精度和一些其他因素,其数值可以通过从 p 远离 q 的标准偏差进行量化。在一个实施方式中,有可能以具有beta分布对 q 模型,参数 α, β 取决于以 p 为中心的平均分布,和一些指定的标准差 s 。特别是,这给出了 $X|q \sim \text{Bin}(q, D_i)$,其中 $q \sim \text{Beta}(\alpha, \beta)$ 。如果我们令 $E(q) = p, V(q) = s^2$,可以导出参数 α, β ,形如 $\alpha = pN, \beta = (1-p)N$,其中

$$N = \frac{p(1-p)}{s^2} - 1。$$

[0477] 这是定义的一个 β -二项分布,其中一个是从带有变参数 q 的二项式分布抽样,其中 q 服从一个平均值为 p 的 β -二项式分布分布。所以,在SNP上的一个步骤无偏差,假设真母体基因型(m),给定在SNP i 上的母体序列A的计数(am_i)以及在SNP i 母体序列B的计数(bm_i),则父本序列数据(SM)的概率可以计算如下:

[0478] $P(SM|m, i) = P_{X|m}(am_i)$ 其中 $X|m \sim \text{Binom}(p_m(A), am_i + bm_i)$

[0479] 现在,包括随机偏差和标准差 s ,上式变为:

[0480] $X|m \sim \text{BetaBinom}(p_m(A), am_i + bm_i, s)$

[0481] 在没有偏差的情况下,假设真母体基因型(m),真子本基因型(c),子本分数(cf),假定子本假设为 H ,给定游离DNA序列A在SNP i 上的计数(a_i)和游离DNA序列B在SNP i 上的计数(b_i),母体血浆DNA序列数据(S)的概率可以计算如下:

[0482] $P(S|m, c, cf, H, i) = P_x(a_i)$

[0483] 其中 $X \sim \text{Binom}(p(A), a_i + b_i)$, $p(A) = \mu(m, c, cf, H)$ 。

[0484] 在一个实施方式中,包括随机偏差与标准差为 s ,这成为 $X \sim \text{BetaBinom}(p(A), a_i + b_i, s)$,其中额外变化量由的偏差参数 s 指定,或等效于 N 。 s 值越小(或 N 的值越大)分布就越接近正规的二项式分布。从明确的AA|AA, BB|BB, AA|BB, BB|AA背景,有可能故居偏差量,即估计上述 N ,并用于上述概率估计 N 。根据数据的特性, N 可以是一个常数,不考虑读数 $a_i + b_i$ 深度, $a_i + b_i$ 的函数,或对大深度读数使偏差较小的 $a_i + b_i$ 的函数。

[0485] 在一个实施方式中,信息学方法可合并每个SNP偏差。由于人工执行测序过程,不论A的实际容量如何,可能有一些SNPs 可能始终具有较低或较高的计数,假设在该SNP i 始终加一个 w_i %的偏差到A的数量计数。在一些实施方式中,这个偏差可以从在相同的条件下获得的训练数据集估计,并添加回估计的父本序列数据中,为:

[0486] $P(SM|m, i) = P_{X|m}(am_i)$

[0487] 其中 $X|m \sim \text{BetaBinom}(p_m(A) + w_i, am_i + bm_i, s)$

[0488] 以及游离DNA序列数据的概率估计为:

[0489] $P(S|m, c, cf, H, i) = P_x(a_i)$ 其中 $X \sim \text{BetaBinom}(p(A) + w_i, a_i + b_i, s)$,

[0490] 在一些实施方式中,该方法可能会被写入具体考虑额外的噪声,差分样本的质量,

差分SNP的质量,和随机取样偏差。这里给出了一个这样的例子。在使用大规模复用的微型PCR协议产生的数据的背景下,以及用于实验7至13时,该方法已被证明是特别有用的。该方法包括几个步骤,每一个步骤都会对最后终的模型引入不同类型的造成和或偏差:

[0491] (1) 假设第一含有母体和胎儿DNA的混合物样本中包含的原始DNA量= N_0 个分子,通常会在1000-40000范围内,其中 p = 实际的%参考的。

[0492] (2) 在使用通用连接适配子扩增时,假定采样 N_1 个分子;通常 $N_1 \sim N_0/2$ 个分子,从而由于采用而引入了随机采样偏差。扩增的样本可能包含 N_2 个分子, $N_2 \gg N_1$ 。设 X_1 代表参考位点(在每个SNP碱基上)的数量,出自 N_1 采样的分子,带有一个偏差 $P_1 = X_1/N_1$,引入了随机抽样偏差贯穿整个协议的其余部分。这个抽样偏差通过使用 β -二项式(BB)分布而不是一个简单的二项式分布包含在模型中。后面,经过调整泄漏和扩增偏差,在SNPs, $0 < P < 1$,之后根据来自训练数据的每个样本的偏差,可估计 β -二项分布的参数 N 。泄漏是不正确地读出一个SNP的倾向。

[0493] (3) 扩增步骤将放大等位基因的偏差,由于可能的不均匀扩增从而扩增偏差被引入。假设在一个位点的一个等位基因被扩增 f 倍,另一个在该位点的等位基因扩增 g 倍,其中 $f = g e^b$,其中 $b=0$ 表示没有偏差。该偏置参数 b ,以0为中心,表示A等位基因有多多多或少地得到扩增对照B等位基因在一个特定SNP的扩增。参数 b 在不同的SNP之间可能会有所不同。偏差参数 b 可以估集每个SNP的偏差,例如从训练数据。

[0494] (4) 测序步骤包括一个测序一个扩增的分子的样本。在这个步骤中可能存在泄漏,泄漏是一个SNP被读错的情况。泄漏可能导致的任何数量的问题,可能会导致在被读数一个SNP不是正确的等位基因A,但如另外一个等位基因B在该位点被发现,或如一个等位基因C或D通常不会在该位点发现。假设测序测量源自一个扩增的大小为 N_3 样品多个DNA分子, $N_3 < N_2$ 。在一些实施方案中, N_3 的可能范围在20,000至100,000;100,000至500,000;500,000至4,000,000;4,000,000至20,000,000;或20,000,000至100,000,000。每个采样分子被正确读取的概率是 p_g ,在这种情况下,它会显示像等位基因A一样正确。该样本将被不正确地读取,如一个与原始分子不相关的等位基因,其概率为 $1-p_g$,并且看起来像等位基因A,概率为 p_r ,像等位基因B,概率为 p_m 或像等位基因C或等位基因D,概率 p_o ,其中 $p_r + p_m + p_o = 1$ 。参数 p_g, p_r, p_m, p_o 从训练数据在每个SNP碱基上进行估计。

[0495] 不同的协议可能包含相似且有所变化的步骤,这些在该分子生物学步骤中的变化,导致不同数量的随机抽样,不同水平的扩增和不同泄漏偏差。下面的模型也可以同样地适用于每一种情况。用于采样DNA数量的模型,在每个SNP上的偏差由下式给出:

[0496] $X_3 \sim \text{BetaBinomial}(L(F(p, b), p_r, p_g), N * H(p, b))$

[0497] 其中 p =对照DNA数量的真值, B =每个SNP上的偏差,如上所述, p_g 是一个正确读取的概率, p_r 是不正确读取的读数概率,但看起来像正确的等位基因,如果是一个坏读取,如上所述,则:

[0498] $F(p, b) = p e^b / (p e^b + (1-p))$, $H(p, b) = (e^b p + (1-p))^2 / e^b$, $L(P, P_r, P_g) = p * P_g + P_r * (1 - p_g)$

[0499] 在一些实施方案中,该方法使用一个 β -二项式分布,而不是一个简单的二项式分布,这需要照顾到随机抽样偏差。 β -二项分布在样本的每个碱上以一个基于需求的偏差进行估计。使用偏差校正 $F(F(p, b), H(p, b))$,而不是仅仅 P ,照顾到扩增偏差。偏差参数 b 在每个

SNP的碱基上从提前训练的数据进行估计。

[0500] 在一些实施方式中,该方法使用泄漏校正 $L(P, P_r, P_g)$,而不是只有 P ;泄漏偏差,即改变SNP与样本的质量。在一些实施方案中,参数 p_g, P_i, P_o 在每一SNP偏差上从提前训练的数据进行估计。在一些实施方案中,参数 p_g, P_i, P_o 在使用中可能随当前的样本被更新,以说明变化的样本的质量。

[0501] 本发明描述的模型是相当普遍的而且能说明两个差分采样质量和差分SNP的质量。不同的样本和SNP的处理方式不同,所举出的事实是一些实施方式中使用的 β -二项式分布,其均值和方差的是原始DNA量,以及样本和SNP质量的函数。

[0502] 平台建模

[0503] 考虑单个SNP,其中存在于血浆中预期的等位基因比率是 r (基于对母体和胎儿的基因型)。预期的等位基因比率被定义为在母体和胎儿的DNA的结合中预期的A等位基因的分数。对于母体基因型 g_m 和子本基因型 g_c ,预期的等位基因比率由式(1)给定,假设基因型也被表示为等位基因比率。

$$[0504] \quad r = fg_c + (1-f)g_m \quad (1)$$

[0505] 在SNP的观察包括对每个出现的等位基因映射的读数的数目, n_a 和 n_b ,其加和为读数深度 d 。假设阈值已经被应用到映射的概率及PHRED分数,这样映射和等位基因的观察可以被认为是正确的。一个PHRED的得分是一个数值度量,涉及一个特定的测量在一个特定碱基的特定测量是错误的概率。在一个实施方式中,其中的碱基已通过测序测量,PHRED的得分可以从对应于所谓的碱基的染色强度与碱基的染色强度的比值计算。对于观察可能性最简单的模型是一个二项式分布,其中假定每个 d 读数是独立从具有等位基因比率为 r 的一个大池中独立得到的。公式2描述了这一模型。

$$[0506] \quad P(n_a, n_b | r) = p_{\text{bino}}(n_a; n_a + n_b, r) = \binom{n_a + n_b}{n_a} r^{n_a} (1-r)^{n_b} \quad (2)$$

[0507] 二项式模型可以多种方式扩展。当母体和胎儿的基因型或全是A或全是B时,预血浆中期等位基因比率将是0或1,因而二项式概率将不能很好地定义。在实践中,有时可观察到意想不到的等位基因。在一个实施方式中,有可能使用一个校正后的等位基因比率 $r = 1/(n_a + n_b)$ 容许小数目的意想不到的等位基因数。在一个实施方式中,有可能使用训练数据模拟出现在每个SNP的意外等位基因的比率,并使用该模型来校正预期等位基因的比率。当预期的等位基因比率是0或1,由于扩增偏差或其它现象,对预期等位基因比率具有足够高的读数深度时,所观察到的等位基因比率都可能不收敛。那么等位基因比率可以以预期的等位基因比率为中心的beta分布建模,得到一个比二项式具有较高方差的 β -二项式分布 $P(n_a, n_b | r)$ 。

[0508] 该平台在一个单个SNP的响应模型将定义为 $F(a, b, g_c, g_m, f)$ (3),或观察概率 $n_a = a$ 和 $n_b = b$ 给出母本和胎儿基因型,其也通过公式(1)取决于胎儿分数。 F 的函数形式可以是一种二项分布、 β -二项式分布、或者上述讨论的相似的公式。

$$[0509] \quad F(a, b, g_c, g_m, f) = P(n_a = a, n_b = b | g_c, g_m, f) = P(n_a = a, n_b = b | r(g_c, g_m, f)) \quad (3)$$

[0510] 在一个实施方式中,子本分数可以确定如下。产前测试胎儿分数 f 的最大似然估计,可能来自没有父本信息的使用。这可能是与父系遗传数据不可用有关,例如记录的父亲不是实际上的胎儿的遗传父亲。从该SNPs集估计胎儿分数,其中母本的基因型是0或1,产生

一组只有两种可能的胎儿基因型。定义 S_0 为具有母本基因型是0的SNPs集合, S_1 为具有母本基因型是1的SNPs 集。可能的胎儿基因型对 S_0 是0和0.5,从而产生在一组可能的等位基因比率 $R_0(f) = \{0, f/2\}$,同理, $R_1(f) = \{1-f/2, 1\}$ 。这种方法可以很简单的被扩展到包括其中母本基因型为0.5的SNPs,但由于可能的等位基因比率集合较大,这些SNPs提供的信息量较小。

[0511] 定义 N_{a0} 和 N_{b0} 是由 n_{as} 和 n_{bs} 对SNPs s 在 S_0 中形成的向量,以及 N_{a1} 和 N_{b1} 类似如对 S_1 。则 f 的最大似然性估计 $\hat{f}$ 由公式(4)定义。

$$[0512] \quad \hat{f} = \arg \max_f P(N_{a0}, N_{b0}|f) P(N_{a1}, N_{b1}|f) \quad (4)$$

[0513] 假设在每个SNP的等位基因计数独立于取决于在该SNP的血浆等位基因比率的条件,概率可表示为在每个集合上SNPs的产品(5)。

$$[0514] \quad P(N_{a0}, N_{b0}|f) = \prod_{s \in S_0} P(n_{as}, n_{bs}|f) \quad (5)$$

$$[0515] \quad P(N_{a1}, N_{b1}|f) = \prod_{s \in S_1} P(n_{as}, n_{bs}|f)$$

[0516] f 上的依赖关系是通过可能的等位基因比率集合 $R_0(f)$ and $R_1(f)$ 。SNP的概率 $P(n_{as}, n_{bs}|f)$ 可以近似为取决于 f 假设的最大似然基因型。在相当高的胎儿分数和读数深度,最大似然基因型的选择将是高置信的。例如,在胎儿分数为10%,深度为1000的读数,可考虑在其母亲有0基因型的SNP。预期等位基因比例为0%和5%,在足够高的读数深度这将是很容易分辨的。把估计子本基因型带入公式(5)得到用于胎儿分数估计的完全公式(6)。

$$[0517] \quad \hat{f} = \arg \max_f \left[\prod_{s \in S_0} \left(\max_{r_s \in R_0(f)} P(n_{as}, n_{bs}|r_s) \right) \prod_{s \in S_1} \left(\max_{r_s \in R_1(f)} P(n_{as}, n_{bs}|r_s) \right) \right] \quad (6)$$

[0518] 胎儿分数必须在范围 $[0, 1]$,所以通过约束的一维搜索可以方便地实现优化。

[0519] 在低读数深度和高噪声水平出现时,不假设最大似然的基因型可能是可取的,这可能会导致人为的高置信度。另一种方法是在每个SNP对可能的基因型加和,对于一个集合 S_0 中的SNP,对 $P(n_a, n_b|f)$ 得到下面的表达式(7)。可以假设先验概率 $P(r)$ 在 $R_0(f)$ 是均匀的,或者可以基于群体的频率。 S_1 组的扩展在是微不足道的。

$$[0520] \quad P(n_a, n_b|f) = \sum_{r \in R_0(f)} P(n_a, n_b|r) P(r) \quad (7)$$

[0521] 在一些实施例中的概率可以推导如下。可以从两个假设 H_t 和 H_f 的数据似然性计算出一个置信度。基于响应模型、估计胎儿组分和母亲的基因型、等位基因群体频率和血浆等位基因计数推导出每个假设的似然性。

[0522] 定义如下符号:

[0523] G_m, G_c 真母本和子本基因型s

[0524] G_{af}, G_{tf} 所谓的父亲和真父亲的真基因型

[0525] $G(g_c, g_m, g_{tf}) = P(G_c = g_c | G_m = g_m, G_{tf} = g_{tf})$ 继承概率

[0526] $P(g) = P(G_{tf} = g)$ 在特定SNP的基因型群体频率

[0527] 假设在每个SNP的观察是独立于取决于血浆等位基因比率的,一个父亲身份加深的似然性是在该SNP上的似然性的结果。下列公式导出了对于一个单个SNP的似然性。公式8是一个对于任何假设 h 的似然性的通式,那么可以将其拆分成针对 H_t 和 H_f 的特定情况。

$$\begin{aligned}
P(n_a, n_b | H_t, G_m, G_{tf}, f) &= \sum_{g_c \in \{0,0.5,1\}} P(n_a, n_b | G_c = g_c, G_m, G_{tf}, h, f) P(G_c = g_c | G_m, G_{tf}, h, f) \\
[0528] \quad &= \sum_{g_c \in \{0,0.5,1\}} P(n_a, n_b | G_c = g_c, G_m, f) P(G_c = g_c | G_m, G_{tf}, h) \\
&= \sum_{g_c \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, h) \quad (8)
\end{aligned}$$

[0529] 在 H_t 的情况下,所谓的父亲是真父亲和胎儿的基因型根据式(9)继承自母本的基因型和所谓的父亲的基因型。

$$\begin{aligned}
P(n_a, n_b | H_t, G_m, G_{tf}, f) &= \sum_{g_c \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_t) \quad (9) \\
[0530] \quad &= \sum_{g_c \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, G_{tf})
\end{aligned}$$

[0531] 在 H_f 的情况下,所谓的父亲不是真父亲。真正父亲的基因型的最佳估计由在每个SNP的群体频率给出。因此,子本基因型的概率由已知的母亲的基因型和群体频率确定,如式(10)。

$$\begin{aligned}
P(n_a, n_b | H_f, G_m, G_{tf}, f) &= \sum_{g_c \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_f) \\
[0532] \quad &= \sum_{g_c \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m) \\
&= \sum_{g_c \in \{0,0.5,1\}} \sum_{g_{tf} \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf} = g_{tf}) P(G_{tf} = g_{tf}) \\
&= \sum_{g_c \in \{0,0.5,1\}} \sum_{g_{tf} \in \{0,0.5,1\}} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, g_{tf}) P(g_{tf})
\end{aligned}$$

[0533] 在正确的父亲身份上的置信度 C_p 用贝叶斯法则(11)从两个似然性假设的SNPs上的结果计算。

$$[0534] \quad C_p = \frac{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f)}{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f) + \prod_s P(n_{as}, n_{bs} | H_f, G_{ms}, G_{tf}, f)} \quad (11)$$

[0535] 采用胎儿百分比分数的最大似法模型

[0536] 通过测量母体血清中所含的游离DNA,或通过测量任何混合样本中的基因型材料来确定胎儿的倍性状态,是一种很有意义的工作。有很多方法,例如,进行读数计算析,假定,如果胎儿的特定染色体为三体型的,则来自母体血液中发现的染色体的DNA的总量将会相对于对照染色体升高。一种检测这种胎儿的三体综合症的方法是归一化每种染色体预期的DNA的量,例如,SNP位点的数目对应于一个给定的染色体,根据对应于给定染色体的分析组中的SNP数,或根据染色体唯一可映射部分的数目。一旦测量已被归一化,任何所测得的DNA的量超过某阈值的染色体被确定为三体型的。如Fan等人所述的方法,PNAS, 2008; 105(42); pp.16266-16271,以及Chiu等人的论文,BMJ 2011;342:e7401。通过计算Z评分归一化如下:

[0537] 试验示例中21号染色体百分比Z分数=((试验示例中21号染色体百分比)-(对照例中21号染色体平均百分比))/(对照例中21号染色体百分比的标准偏差)。

[0538] 这些方法使用单假设拒绝法确定胎儿的倍体状态。然而,他们有着一些明显的缺点。因为这些用于根据样本中的胎儿DNA样本中的百分比来确定在胎儿倍性的方法是不变的,他们使用一个切断值;这样做的结果是,测定的精度不是最优的,在混合物中胎儿DNA的百分比相对较低这种情况下,将出现最差的精度。

[0539] 在一个实施方式中,本申请的用于确定胎儿的倍性状态的方法包括考虑到样本中的胎儿DNA的分数。在本申请的另一个实施方式中,该方法包括使用极大似然估计。在一个实施方式中,本发明公开的方法包括计算起源于胎儿或胎盘的样本中DNA的百分数。在一个实施方式中,用于识别非整倍性的所述阈值根据所计算出的胎儿DNA的百分比进行自适应

调整。在一些实施方式中,用于估计DNA混合物中的胎儿来源的DNA的百分比的方法,包括获取包括来自母体的遗传物质和来自胎儿的遗传物质的混合样本,获取来自胎儿父亲的遗传样本,测量混合样本中的DNA,测量父亲样本中的DNA,并使用混合样本的和父亲样本的DNA测量值来计算DNA混合物中的胎儿来源的DNA的百分比。

[0540] 在本申请的一个实施方式中,胎儿DNA的分数,或混合物中胎儿DNA的百分比的可以被测量。在一些实施方式中,所述分数可以只使用基于母体血浆样本本身(这是胎儿和母体DNA的混合物)进行的基因分型测量来计算。在一些实施方式中,还可以用母体的测得的或其他已知的基因型和/或父亲的测得的或已知的基因型来计算该分数。在一些实施方式中,可以使用基于母体和胎儿DNA的混合物以及亲本的背景信息的测量值来胎儿DNA的百分数。在一个实施方式中,胎儿DNA分可以使用群体频率计算以调整基于有关特定等位基因测量值的概率模型。

[0541] 在本申请的一个实施方式中,可以基于胎儿倍性状态的测定的准确性来计算置信度。在一个实施方式中,具有最大似然度(Hmajor)的假设置信度可以被计算为 $(1-H_{major})/\Sigma$ (所有的H)。如果所有的假设的分布是已知的,就可以确定一个假设的置信度。如果亲本的基因型信息是已知的,就可以确定所有的假设的分布。如果整倍体胎儿的数据的预期分布和非整倍体胎儿的数据的预期分布信息是已知的,就可以计算倍性判定的置信度。如果亲本的基因型信息是已知的,就可以计算这些预期的分布。在一个实施方式中,可以使用正常假设附近的和反常假设附近的检验统计量的分布信息来确定识别的和精确的阈值这二者的可靠性,以及进行更可靠识别。当混合物中的胎儿DNA的量和/或百分比很低时,这是特别有用的。这将有助于避免的情况被发现是整倍体的胎儿实际上是非整倍体的情况,因为检验统计量,例如Z统计量不会不超过一个基于针对较高百分比的胎儿DNA的情况被优化的阈值而获得的阈值。

[0542] 在一个实施方式中,本发明所公开的方法可以用于通过确定在母体和胎儿的遗传物质的混合物中的母体和胎儿的目标染色体的数目拷贝来确定胎儿非整倍体型。该方法可能需要获得包含母体和胎儿的遗传物质的母体组织;在一些实施方式中,该母体组织可能是从母体血液中分离出来的母体血浆或组织。该方法可能还需要通过处理上述母体组织从上述母体组织获得母体和胎儿的遗传物质的混合物。该方法可能需要将所获得的遗传物质分配成多个反应样本,随机提供包括来自目标染色体的目标序列的个体反应样本,不包括来自目标染色体的目标序列个体反应样本,例如对样本进行高通量测序。该方法可能需要分析个体反应样本中存在还是不存在遗传物质的靶序列,以提供表示反应样本中的推测的整倍体胎儿染色体存在还是不存在的二进制结果的第一个数,以及表示反应样本中的推测的非整倍体胎儿染色体存在还是不存在的二进制结果的第二个数。任何一个二进制结果的数都可以计算,例如,通过计算映射到一个特定的染色体的序列读数的信息学技术的方式,染色体的特定区域,特定的位点或位点集。该方法可以包括基于染色体长度,染色体区域的长度归一化的二进制事件数,或该集中的位点数。该方法可能需要使用第一个数来计算用于推测反应样本中的整倍体胎儿染色体的二进制结果数的预期分布。该方法可能需要使用第一个数以及估计的混合物中发现的胎儿DNA的分数来计算用于推测反应样本中的非整倍体胎儿染色体的二进制结果数的预期分布,例如,用于推测整倍体型胎儿染色体的二进制结果的预期读数计算分布乘以 $(1+n/2)$,其中n是估计胎儿分数。

[0543] 在一些实施方式中,序列读数可以以概率映射的方式处理而不是二进制的结果;这种方法会产生更高的精确度,但需要更大的计算能力。可以采用多种方法来估算胎儿分数,其中一些在本申请的其他部分有介绍的。该方法可以包括使用最大似然法来确定是否第二数对应于为整倍体或为非整倍体的可能的胎儿染色体非整倍体。该方法可以包括识别胎儿的倍性状态,以使该倍性状态与具有最大似然度的正确给定的测量数据的假设相对应。

[0544] 需要注意的是,也可以使用可以用于提高确定胎儿的倍性状态的任何方法的准确度的最大似然度模型。同样的,可以计算确定胎儿的倍性状态的任何方法的置信度。最大似然度模型可以使得采用单假设拒绝技术倍性测量的任何方法的准确性得以改善。最大似然度模型可以用于可以针对正常和不正常情况下计算似然度分布的任何方法。最大似然度模型的使用意味着计算倍体识别置信度的能力。

[0545] 该方法的进一步讨论

[0546] 在一个实施方式中,本发明所公开的方法利用多态性位点的每个等位基因的独立观测数的定量测量,该方法不包括计算等位基因比。这与诸如一些基于微阵列法的方法不同,其提供了在一个位点的两个等位基因的比例的信息,但不能量化两种等位基因的的独立观测数目。一些在本领域中已知的方法,可以提供有关独立观测数目的定量信息,但进行倍性测定的计算只利用等位基因比率,不使用的量化信息。要说明保留有关独立观测数的信息的重要性,要考虑具有两个等位基因A和B的样本。在第一个实验中考察20个A等位基因和20个B等位基因,在第二个实验中考察200个A等位基因和200个B等位基因。在这两个实验中的(A/(A+B))的比=0.5,然而,第二个实验比第一个实验传达了更多关于A或B等位基因频率的确定性信息。本方法,使用定量数据而不是利用等位基因比率,来更准确地模拟最有可能在每个多态性位点的等位基因频率。

[0547] 在一个实施方式中,即时法建立了汇集来自多个多态性位点的测量值的遗传模型,以更好地区别二倍体和三倍体症并确定三倍体类型。此外,本方法合并遗传连锁信息,以提高该方法的准确性。与本领域已知的方法相比,这是一条染色体上的所有多态性位点的等位基因比的平均值。本发明公开的方法明确模拟由减数分裂过程I不分离、减数分裂过程II不分离、在胎儿形成早期的减数分裂过程中的不分离产生的二倍体和三倍体中预期的等位基因频率分布。为了说明为什么这一点很重要,如果减数分裂过程I中没有交叉不分离,将产生其中的两个不同的同源体继承自父母中的一个的三倍体;而减数分裂过程II不分离、在胎儿形成早期的减数分裂过程中的不分离将产生来自父母中的一个的相同同源体的两个拷贝。每个推测都在每个多态性位点,还有在共同考虑的所有物理连接位点(即相同染色体上的基因位点)产生不同的预期等位基因频率。交叉,导致同源体之间的遗传物质的交换,使继承模式更为复杂,但即时法可允许使用遗传连锁信息,即位点之间的重组率信息和物理距离。为了更好地区别减数分裂过程I不分离、减数分裂过程II不分离,为着丝点的距离增加即时法结合模型增加交叉概率。在两个同源体随后提供一个减数分裂不分离事件通常由于在配子发生过程中的一个或多个交叉而不同时,减数分裂过程II不分离可以通过减数分裂不分离通常会产生一个同源体的两个相同的或相近的拷贝的事实来辨别。

[0548] 在一个实施方式中,如果假定二倍体,本申请方法可能无法确定单倍体的父母。在一个实施方式中,在三体综合征的情况下,即时法可以通过使用血浆带有来自一个亲本的

两个拷贝的事实来确定关于父母一方或双方的单倍型,母本阶段信息可以通过两个拷贝都继承自待测亲本的算法来确定。特别是,一个孩子可以继承亲本(配对三倍体)的两个相同的拷贝的任意一个,或者是亲本的两个拷贝(不配对三倍体)。每个SNP,都可以计算出配对三倍体和不配对三倍体的似然性。由于交叉而不使用联动模式的倍性识别法将以所有染色体的配对和不配对的三倍体症的简单的加权平均的方式计算三倍体症的总似然度。然而,由于导致脱节错误和跨越的生物学机制,只有发生交叉,一条染色体上三倍体才可以从配对变成不配对(反之亦然)。即时法概率考虑到交叉的可能性,从而使得倍性识别比那些不这样做的方法具有更高的精度。

[0549] 在一个实施方式中,对照染色体用来确定子本分数和噪声水平值或概率分布。在一个实施方式中,子本分数,噪声水平,和/或概率分布仅使用从其倍性状态要被测定的染色体获得的遗传信息来确定。即时法没有对照染色体,以及没有固定的特定的子本分数或噪音水平就可以进行。这是本领域的已知方法一个显着的改善和不同点,其中来自对照染色体的遗传数据是校正子本分数和染色体行为的必须的。

[0550] 在一个实施方式中,确定胎儿的分数不需要对照对照染色体,确定假设如下:

[0551] $H^* = \operatorname{argmax}_H \text{LIK}(D|H) * \text{先验概率}(H)$

[0552] 采用利用对照染色体的算法,通常假定对照染色体是一个二倍体,然后

[0553] (a) 基于这个假设,并对照染色体数据可以定出最有可能的子分数和随机噪声水平N:

[0554] 然后减小

[0555] $[cfr^*, N^*] = \operatorname{argmax}_{cfr, N} \text{LIK}(D(\text{ref. chrom})|H11, cfr, N)$

[0556] 及然后简化为

[0557] $\text{LIK}(D|H) = \text{LIK}(D|H, cfr^*, N^*)$

[0558] 或(b)基于这个假设并对照染色体数据估计子本分数和噪音水平分布。特别是,人们不会修复只是一个值cfr和N,但分配可能性p(cfr, N)范围更广, N值的概率为p(cfr, N):

[0559] $p(cfr, N) \sim \text{LIK}(D(\text{ref. chrom})|H11, cfr, N) * \text{priorprob}(cfr, N)$

[0560] 其中先验性概率(cfr, N),特别是子本分数和噪声水平的先验概率,在了解和实验之前测定。如果需要的话,只是在均匀cfr 范围内,然后写出:

[0561] $\text{LIK}(D|H) = \sum_{cfr, N} \text{LIK}(D|H, cfr, N) * p(cfr, N)$

[0562] 上述两种方法都得到了良好的效果。

[0563] 请注意,在某些情况下使用对照染色体是不需要的,可能或可行的。在这种情况下,可能分别对每个染色体进行最好的倍性识别。特别是:

[0564] $\text{LIK}(D|H) = \sum_{cfr, N} \text{LIK}(D|H, cfr, N) * p(cfr, N|H)$

[0565] 可能如上述所确定的,为p(cfr, N|H)以上,分别对每个染色体,假定假设H,不只是为假定二倍体的对照染色体。可以使用这种方法,保持噪声和子本分数的参数固定,或固定的参数中的任何一个,或保持每个染色体,每个假设的概率的形式中的这两个参数中。

[0566] DNA测量容易有噪声和/或出错,特别是测量的DNA的量很小,或者其中的DNA混合污染了DNA。这种噪声将导致基因型数据不准确,并倍性识别不太准确。在一些实施方式中,噪声建模的平台建模或一些其他的方法也可以用于计算噪声对倍性测定的有害影响。即时方法使用两个通道的联合模型,其造成随机噪声的原因是由于输入DNA的量,DNA质量和/或协议质量。

[0567] 与本领域中已知的一些方法相比,使用在一个位点的等位基因强度的比率进行倍性的测定。这种方法阻碍了准确的SNP噪声建模。特别是,测量误差通常不具体取决于所测量的信道的强度比,其减少了使用一维信息模型。噪声,信道质量和信道的相互作用的精确建模需要两维的联合模型,这是不能用等位基因比率进行建模。

[0568] 特别是,将两个信道信息用于比率 r ,其中, $f(x,y)$ 是 $r=x/y$,并不有助于信道噪声和偏置建模。一个特定SNP的噪声不是其比值的函数,即噪声 $(x,y) \neq f(x,y)$,但实际上是在两个通道上的联合函数。例如,在二项式模型中,噪声测量比具有 $r(1-r)/(x+y)$ 的方差,这完全是 r 的函数。在这种模型中,包含任何通道的偏置或噪音,假定在SNP i ,观测到的通道 X 值为 $x=a_iX+b_i$,其中 X 是真通道值, b_i 是额外的通道偏置和随机噪声。同样,假设 $y=c_iY+d_i$ 。所观察到的比率 $r=x/y$ 不能准确预测的真实比例 X/Y 或模拟剩余噪声,因为 $(a_iX+b_i)/(c_iY+d_i)$ 不是 X/Y 的函数。

[0569] 本发明所披露的方法描述一种采用所有各测量通道联合二项式分布模拟噪声和偏差的有效的方法。相关方程可以在本申请其他部分找到,其显示每个SNP的恒定偏差, $P(\text{好})$ 和 $P(\text{ref}|\text{差})$, $P(\text{mut}|\text{差})$,其有效地调整SNP行为。在一个实施方式中,本申请的方法使用了BetaBinomial分布,从而避免了仅仅依靠对等位基因比率的限制的做法,而是在两个通道数的基础上的行为模型。

[0570] 在一个实施方式中,本发明披露的方法可以通过使用所有可用的测量值由母体血浆中发现的遗传数据来识别妊娠中的胎儿的倍性。在一个实施方式中,本发明披露的方法可以通过使用仅来自亲本的背景的子集的测量值由母体血浆中发现的遗传数据来识别妊娠中的胎儿的倍性。本领域中一些已知的方法只使用测得的遗传数据,其中父母背景来自AA|BB背景,也就是说,如果不是不同的等位基因,父母两人在一个给定的位点均是纯合型的。采用这种方法的一个问题是,一小部分多态性位点是来自于AA|BB背景,通常小于10%。在本发明方法所公开的实施方式中,该方法不使用亲本背景是AA|BB的位点制得的母体血浆遗传测量值。在一个实施方式中,该即时法只将血浆测量值用于这些具有AA|AB,AB|AA,AB|AB亲本背景的多态性位点。

[0571] 本领域中已知的方法包括计算来自其中存在父母双方的基因型的AA|BB背景下的SNP的等位基因比的平均值,并要求由基于这些SNP的平均等位基因比来确定倍性识别。该方法由于SNP的行为差异有着明显的不准确性。注意,该方法假定两个亲本因型是已知的。相反,在一些实施方式中,即时法采用了联合通道分布模式,不假定父母任意一方的存在,并且不假定均匀的SNP行为。在一些实施方式中,即时法导致不同的单核苷酸多态性的行为/权重。在一些实施方式中,即时法不需要对一个或两个亲本基因型的信息。即时法如何做到这一点的例子如下:

[0572] 在一些实施方式中,假设的对数似然度可以在每个SNP基础上确定。对于一个特定的SNP i ,假定胎儿倍性假设 H 和胎儿DNA百分比 cf ,所观测的数据 D 的对数似然度倍定义

为:

[0573]

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

[0574] 其中m是可能的真实母体基因型,f是可能的真实父亲的基因型,其中m,f ∈ {AA, AB, BB}, 其中c是可能的给定的假设H 时的子本基因型,特别是。单倍体c ∈ {A,B}, 二倍体c ∈ {AA,AB,BB}, 三倍体c ∈ {AAA,AAB,ABB,BBB}。请注意,包括亲本的基因型数据通常产生更准确的倍性测定结果,然而,亲本的基因型数据对于操作良好的即时法不是必须的。

[0575] 本领域中一些已知的方法包括计算来自其母体是纯合型,而血浆(AA|AB或AA|BB背景中任意一个)中测出不同等位基因的SNP的等位基因比的平均值,以及由基于这些SNP的平均等位基因比来确定倍性识别的要求。该方法适用于父亲的基因型无法获得的情况。请注意,在一个特定的SNP上血浆是杂合的,在与父亲BB相对的位置而不存在纯合的情况下宣称有多准确是值得商榷的,因为在这种情况下,子本分数低,看上去是存在B 的等位基因可能只是存在的噪声。另外,看上去没有B存在可能只是胎儿测量结果缺失等位基因。即使在实际可以确定血浆的杂合性的情况下,该方法也无法辨别父亲的三体性。特别地,对于母亲是AA的SNP,并在血浆中测出一些B的情况,如果父亲是 GG,得到的孩子基因型AGG,结果是平均比例为33%的A(孩子分数=100%)。但是在父亲是AG的情况下,得到的孩子基因型可能会是与三体综合症配对的AGG,结果比率是33%的A,而对于与三体综合症不配对的AAG,结果平均比率更倾向于66 %的A。鉴于许多三倍体基于染色体交叉,整个染色体介于不含不配对三倍体和含有全部不配对三倍体之间的任何部分,该比率在33-66%之间变化。对于普通二倍体,比例应该是50%左右。不使用连接模型或平均数准确度误差模型,这种方法会错过许多父源三体症情况。与此相反,本发明所公开的方法,基于能够获得的基因型信息和群体频率,为每个亲本基因型候选者指定亲本基因型概率,并不明确要求亲本的基因型。此外,即使在父母因型数据的存在或不存在的的情况下,本发明所公开的方法也能够检测到三体综合征,并且可以采用连接模型通过识别配对到不配对的三体综合征可能的交叉点来补偿。

[0576] 本领域中用于计算来自母本或父本的基因型均不知道的 SNP的等位基因比的平均值,并由基于这些SNP的平均比确定倍性识别的方法。然而,实现这些目的的方法没有披露。本发明所公开的方法能够使倍性识别在这样的情况下更准确,并且在本发明的其他部分所公开的具体实施方式,使用联合概率最大似然法和选择性地利用SNP噪声和偏差模型,以及连接模型。

[0577] 本领域中一些已知的方法包括平均等位基因比率,如权利要求由一个或几个SNP的平均等位基因比来确定的倍性识别。然而,这样的方法不使用联动的概念。本发明公开的方法不具有这些缺点。

[0578] 使用序列长度为先验确定DNA的起源

[0579] 据报道母体和胎儿的DNA序列长度的分布是不同的,胎儿的通常较短。在本申请的一个实施方式中,可以使用经验数据形式的以前的信息,构造母体(P(X|母体))和胎儿DNA(P(X|胎儿))的预期长度的先验分布。由于给定的新的身份不明的DNA 序列具有长度x,根据x给定的是母体还是胎儿的先验概率,可以指定给定的DNA序列是母体或胎儿DNA的概率。

特别是,如果 $P(X|\text{母体}) > P(X|\text{胎儿})$,则该DNA序列可以被归类为母体,用 $P(X|\text{母体}) = P(|\text{母体}) / [P(X|\text{母体}) + P(X|\text{胎儿})]$,并且如果 $p(x|\text{母体}) < p(x|\text{胎儿})$,则该DNA序列可以被归类为胎儿, $P(x|\text{胎儿}) = P(x|\text{胎儿}) / [P(x|\text{母体}) + P(x|\text{胎儿})]$ 。在本申请的一个实施方式中,能够确定母体和胎儿的序列长度分布对于考虑所述概率可以被指定为具有高概率母体或胎儿的样本很有效,则该样本的具体分布可以作为该样本的预期大小分布。

[0580] 可变读数深度最小化序列成本

[0581] 在许多与诊断有关的临床试验中,例如,在Chiu等人,BMJ 2011;342:c7401,许多参数被设定的协议,然后对试验中的每个患者使用相同的参数执行相同的协议。在使用测序作为测量遗传物质的方法来确定母本的妊娠中的胎儿的倍性状态的情况下,相关的参数是读数的数量。读数的数量可以指实际读数的数量,预计读数、部分泳道、全部泳道、或测序器上的全流细胞的数量。在这些研究中,读数的数量通常设置在确保所有或几乎所有的样本达到所需的精度水平的水平。测序是目前一种昂贵的技术,每 5个可映射100万的读数的成本大约200美元,当价格下降时,任何可以使得基于测序的诊断以较少的读数在类似的准确性水平下实施的方法,将比如节省相当经费。

[0582] 倍性测定的准确度通常取决于若干因素,包括读数的数量和混合物中胎儿DNA的分数。当混合物中胎儿DNA的分数较高时,准确度通常较高。同时,如果读数的数量较多,准确度通常较高。可能出现两种情况,其中倍性状态以相当的准确度测定,在第一种情况下其混合物中的胎儿DNA具有小于第二种情况的分数,在第一种情况下需要测序的读数大于第二种情况的。可以使用估计的混合物中胎儿DNA的分数作为达到给定的精度水平所必须测定的读数的数量一个指标。

[0583] 在本申请的一个实施方式中,可以跑一组样本,其中该组中的不同的样本对不同的读数深度测序,其中,每个样本的读数的数量以每个混合物中胎儿DNA计算分数所能达到的给定精度水平进行选择。在本申请的一个实施方式中,这可能需要进行混合样本的测定以确定该混合物中的胎儿DNA分数;可以采用测序的方式估测胎儿分数,可以用TaqMan来完成,可以用qPCR来完成,可以用SNP阵列来完成,可以用可以区分给定位点的不同等位基因的任何方法来完成。当与实际测得的数据比较时,通过包含一组所考虑的假设中涵盖全部的或一组选定的胎儿分数的假设,可以淘汰对于胎儿分数估计的需要。该混合物中的胎儿DNA分数确定后,对每个样本要读取的序列数可以被确定。在本申请的一个实施方式中,100名孕妇访问其各自的OB,他们的血被抽入含有抗lysan和/或灭活DNA酶的血管。他们每个人都带回家一个试剂盒,让其孕育的胎儿的父亲给出唾液样本。所有100对配偶的这两套遗传物质被送回实验室,母体血液旋转降速分离白细胞层,以及血浆。所述血浆含有母体DNA以及胎盘来源的DNA的混合物。使用SNP阵列对母体白细胞层和父亲的血液进行基因分型,而母体血浆样本中的DNA SURESELECT 杂交探针定位。用探针展开的DNA用于生成100标记库,每个用于一个母体样本,其中每个样本都以不同的标记来标记。从每个库中抽取一部分,各部分混合在一起,并加入多路复用形式的 ILLUMINA HISEQ DNA测序仪的两个泳道上,其中每个泳道产生约50,00万可映射的读数,基于100复合混合物产生约100万可映射的读数,或每个样本约100万读数。序列读数被用来确定每个混合物中胎儿DNA的分数。50个样本的混合物中具有超过 15%以上的胎儿DNA,在100万读数足以以99.9%的置信度确定胎儿的倍性状态。

[0584] 剩余的混合物中,有25例胎儿DNA在10和15%之间;由这些混合物形成的相关库的每部分复合起来和并在HISEQ的一条泳道上跑,每个样本产生额外的200万读数。胎儿DNA在10 和15%的混合物两组序列数据加到一起,每个样本获得300万读数足以以99.9%的置信度确定这些胎儿的倍性状态。

[0585] 剩余的混合物中,有13例胎儿DNA在6和10%之间;由这些混合物形成的相关库的每部分复合起来和并在HISEQ的一条泳道上跑,每个样本产生额外的400万读数。胎儿DNA在6和 10%的混合物两组序列数据加到一起,每个样本获得500万读数足以以99.9%的置信度确定这些胎儿的倍性状态。

[0586] 剩余的混合物中,有8例胎儿DNA在4和6%之间;由这些混合物形成的相关库的每部分复合起来和并在HISEQ的一条泳道上跑,每个样本产生额外的600万读数。胎儿DNA在4和6 %的混合物两组序列数据加到一起,每个样本获得700万读数足以以99.9%的置信度确定这些胎儿的倍性状态。

[0587] 剩余的四个混合物中,所有的胎儿DNA都在6和10%之间;由这些混合物形成的相关库的每部分复合起来和并在HISEQ的一条泳道上跑,每个样本产生额外的1200万读数。胎儿DNA在 2和4%的混合物两组序列数据加到一起,每个样本获得1300万读数足以以99.9%的置信度确定这些胎儿的倍性状态。

[0588] 这种方法需要HISEQ仪上有六条测序泳道以使超过100个的样本达到99.9%的准确率。如果每个样本要求跑相同的数量,以确保每次倍性测定的准确率为99.9%,可以采取25条测序泳道,如果可以允许没有识别率或4%的错误率,则14条测序泳道就可以实现。

[0589] 使用原始的基因分型数据

[0590] 有许多方法可以使用从孕妇血液中发现的胎儿DNA上测得胎儿遗传信息完成NPD。这些方法包括利用SNP阵列进行胎儿 DNA的测量,一些方法包括非定向测序,而一些方法包括定向测序。定向测序可以定向SNP位点,其可以定向STR,其可以定向其他多态性位点,其可以定向非多态性位点,或它们的一些组合。这些方法的一些可以包括使用商业或专有的等位基因识别仪,其可以通过来自进行测量的装置中的传感器的强度数据来识别等位基因的身份。例如,ILLUMINA的INFINIUM系统或 AFFYMETRIX GENECHIP微阵列系统包括有孔玻璃珠或附有可以与DNA的互补片段杂交的DNA序列的微晶片,可以检测到的传感器分子荧光特性的变化。还有测序方法,例如ILLUMINA SOLEXA GENOME SEQUENCER或ABI SOLID GENOME SEQUENCER,其中,DNA片段的所述基因序列被测序;DNA 互补链的延伸后的链被测序,扩展核苷酸的身份通常是通过荧光或互补核苷酸上所附的无线标记进行检测。在所有这些方法中基因型或测序数据通常是基于荧光或其他信号,或其缺失测得的。这些系统通常配有低级别的软件包。由荧光或其他检测装置(原始遗传数据)的模拟输出识别特异性等位基因(二次遗传数据)。例如,在的给定基于SNP阵列的等位基因的情况下,该软件可以进行识别,例如,如果荧光强度高于或低于某一阈值时,某SNP 存在还是不存在。同样,测序仪的输出是显示所检测到的每种染料的荧光水平的色谱图,该软件将识别某碱基对是A或T或C 或G。高通量测序仪通常进行一系列这样的测量,识别表示被测序的DNA序列的最有可能的结构的读数。这里所定义的色谱的直接模拟输出值是原始遗传数据,通过软件进行碱基对/SNP识别在本发明中被认为是二次遗传数据。在一个实施方式中,原始数据是指原始强度的数据是基因型分型平台的未经处理的输出,其中基因型分型平台是指SNP阵列,或指测

序平台。该二次遗传数据是指处理过的遗传数据,其中其进行等位基因识别,或序列数据已指定碱基对,和/或序列读取已经被映射到基因组。

[0591] 许多更高级的应用软件利用这些等位基因识别,SNP识别和序列读数,也就是说,二次遗传数据,基因分型软件包。例如, DNA NEXUS,ELAND或MAQ将提取测序读数,并将它们映射到基因组。例如,本发明中,非侵入性的产前诊断,复杂的信息系统,如PARENTAL SUPPORTTM,可以利用大量的SNP识别来确定个体的基因型。此外,本发明中,胚胎植入前遗传学诊断,可以提取被映射到基因组中的一组序列读数,通过将被映射到每个染色体或部分的染色体的序列读数的计数归一化处理,就可以确定个体的倍性状态。本发明中,在非侵入性产前诊断时,可以提取一组由存在于孕妇血浆中的DNA测得的序列读数,并将它们映射到基因组。然后可以对被映射到每个染色体或部分的染色体的的序列读数进行归一化处理,并使用该数据来确定个体的倍性状态。例如,可能会得出这样的结论,具有大得不成比例的读数的量的染色体是来自被抽取血液的母本的妊娠胎儿的三倍体。

[0592] 然而,实际上,测量仪器的最初输出的是模拟信号。当某个碱基对是由与测序软件相关的软件(例如可以识别碱基对的一个 T的软件),则实际上,识别是软件认为是最有可能的识别。然而,在某些情况下,该识别可能置信度很低,例如,模拟信号可能表明特定的碱基对只有90%可能是T,10%可能是A。在另一个例子中,与SNP阵列读数相关的基因型识别软件可以识别某等位基因是G。然而,实际上,基本模拟信号可能表示等位基因是G的可能只有70%,和等位基因是T的可能为30%。在这些情况下,更高级的应用软件时使用基因型识别和由较低级的软件进行的序列识别,他们将丢失一些信息。也就是说,原始遗传数据,由基因分型平台直接测得,可能比由所附的软件包确定的二次基因混乱,但包含更多的信息。二次遗传数据序列映射到基因组中时,许多读数被剔除,因为有些碱基读取没有足够的清晰度,或映射不清晰。当使用原始遗传数据序列读数时,所有的或许多的在原始读数被转化成二次基因读数序列时那些可能已被剔除的读数,可以通过处理的概率法的读数的方式使用。

[0593] 在本申请的一个实施方式中,较高级的软件并不依赖于等位基因识别,SNP识别或由较低级的软件确定的序列读数。相反,较高级的软件将计算基础建立在由基因分型平台直接测得的模拟信号。在本发明的一个实施方式中,基于信息系统的方法,例如修改PARENTAL SUPPORTTM,以便能够重建胚胎/胎儿/子本的遗传数据被设计为直接使用由基因分型平台所测得的原始遗传数据。在本发明的一个实施方式中,基于信息系统的方法,例如PARENTAL SUPPORTTM可以使用的原始遗传数据而不是使用二次遗传数据进行等位基因识别和/或染色体拷贝数识别。在本发明的一个实施方式中,所有遗传识别、SN识别、序列读数,序列映射都通过使用由基因分型平台所测得的原始强度数据并以概率的方式来进行处理,而不是将原始遗传数据转化成二次遗传识别。在一个实施方式中,来自制备样本的DNA的测量值用于计算等位基因数的概率,并确定包含原始遗传数据的每个假设的相对概率。

[0594] 在一些实施方式中,该方法可以提高含有至少一个相关的个体的遗传数据的目标个体的遗传数据的准确性,该方法包括获得目标个体的基因组的特异性原始遗传数据和相关个体的基因组的特异性遗传数据,建立一个或多个可能关于对应于目标个体的基因组中的染色体片段的来自相关个体的染色体片段假设集,确定每个给定的目标个体的原始遗传数据的假设的概率,使用与每个假设有关的概率来确定目标个体的实际遗传物质的最可能

状态。在一些实施方式中,所述方法可确定的目标个体的基因组中染色体的一个片段的拷贝的数目,该方法包括创建关于有多少染色体片段的拷贝存在于目标个体的基因组中拷贝数假设集,将来自目标个体的原始遗传数据和来自一个或多个相关个体的遗传信息并入数据集,估测与数据集相关的平台响应的特征,其中平台响应可以该从一个实验变化到另一个,计算每个拷贝数假设的条件概率,给定的数据集和平台的响应特征,并基于最有可能的拷贝数假设确定染色体片段的拷贝数。在一个实施方式中,本申请的方法可确定目标个体中至少一个染色体的倍性状态,该方法包括:获取来自目标个体和来自一个或多个相关个体的原始遗传数据,创建目标个体的每个染色体的至少一个倍性状态假设集,采用一种或多种专业技术来确定该集中每个倍性状态假设的统计概率,对于所采用的每种专业技术,给定所获得的遗传数据,结合,对于每个倍性状态的假设,由一个或多个专业技术所确定的统计概率,基于所结合的每个倍性状态假设的统计概率确定目标个体中的每个染色体倍性状态的统计概率。在一个实施方式中,本申请的方法可确定目标个体,和来自目标个体的父母一方或双方,和任选的来自一个或多个相关的个体的等位基因集中的等位基因水平,该方法包括:获得来自目标个体,和来自父母一方或双方,和来自任何相关的个体的原始遗传数据,创建目标个体的,和父母一方或双方,和任选的任何相关的个体的至少一种等位基因的假设集,确定给定的所获得的遗传数据的假设集中的每个等位基因的统计概率,确定目标个体的,和父母一方或双方,和任选的任何相关的个体的等位基因集中每个等位基因的等位基因状态,基于各等位基因的假设的统计概率。

[0595] 在一些实施方式中,混合样本的遗传数据可以包括序列数据,其中所述序列数据可能不能唯一地映射到人类基因组。在一些实施方式中,混合样本的遗传数据可以包括序列数据,其中序列数据映射到基因组中的多个位置,其中每个可能的映射都与在给定的映射是正确的概率相关。在一些实施方式中,序列读数是假定与基因组中的特定位置相关联。在一些实施方式中,所述序列读数与基因组中的多个位置相关,以及并与属于该位置的概率相关。

[0596] 产前诊断结合方法

[0597] 有许多方法可用于非整倍性或其他遗传缺陷的产前诊断或产前筛查。本发明档中的其他部分所述的,以及美国申请号 11/603,406,2006年11月28日申请;美国申请号12/076,348, 2008年3月17日申请,以及美国PCT申请号No. PCT/S09/52730 是这样的一种方法,该方法使用相关个体的遗传数据,来提高目标个体的遗传数据的准确度,如胎儿,是已知的,或估计的。其他用于产前诊断的方法涉及测量母体血液中某些激素的水平,这些激素与各种遗传异常有关。这方面的一个例子,被称为三联筛查试验,该试验中,测量母体血液中的几种(通常为两种,三种,四种或五种)不同的激素水平。在多重法被用来确定一个给定的结果的似然性,其中没有一种方法就其本身而言是确定的,可以将这些方法所给定的信息结合起来使得预测比单独的方法更准确。在三联筛查试验中,由三种不同的激素给出的信息相结合,可使得遗传性异常的预测比个体激素水平可以预测到的更准确。

[0598] 本发明公开了一种用于使得关于胎儿的遗传状态的预测的更精确的方法,特别是胎儿的遗传异常的可能性,其包括将使用多种方法预测的胎儿的遗传异常的预测结合相结合。“更准确的”的方法可以指在给定的假阳性率条件下,用于诊断具有较低的假阴性率的异常。在本申请的一个有利的实施方式中,基于已知是关于胎儿的遗传学数据进行一种或

多种预测,其中使用 PARENTAL SUPPORTTM法确定遗传数据,也就是说,使用与胎儿相关的个体的遗传数据来确定胎儿的遗传数据具有更高的准确性。在一些实施方式中,所述遗传数据可以包括胎儿的倍性状态。在一些实施方式中,遗传数据可以指胎儿的基因组中一组等位基因的识别。在一些实施方式中,可以使用三联筛查试验进行一些预测。在一些实施方式中,可以使用母体血液中其他的激素水平的测量来进行一些预测。在一些实施方式中,通过所考虑的诊断方法进行的预测,可以与所考虑的筛选的方法作出的预测相结合。在一些实施方式中,所述方法包括测量母体血液中的 α -胎甲球蛋白(AFP)。在一些实施方式中,所述方法包括测量母体血液中未结合的雌三醇(UE3)水平。在一些实施方式中,所述方法包括母体血液中的 β ,人绒毛膜促性腺激素(β -hCG)水平的测量。在一些实施方式中,所述方法包括测量母体血液中侵入性滋养层细胞膜抗原(ITA)的水平。在一些实施方式中,所述方法包括测量母体血液中的抑制素的水平。在一些实施方案中,所述方法包括测量母体血液中妊娠相关血浆蛋白A(PAPP-A)的水平。在一些实施方案中,所述方法包括测量母体血液中其他激素或孕母体血清标记的水平。在一些实施方式中,以使用其他方法进行一些预测。在一些实施方式中,可以使用完全综合试验来进行一些预测,例如,将超声波与妊娠12周左右的验血,以及与16周左右第二次验血相结合。在一些实施方式中,所述方法包括测量颈部透明组织厚度(NT)。在一些实施方式中,所述方法包括使用所测得的上述激素水平进行预测。在一些实施方式中,所述方法包括上述方法的组合。

[0599] 有许多与预测相结合的方法,例如,可将激素测量转换成中位数的倍数(MoM),然后转换成似然比(LR)。同样的,其他的测量值可以转化成使用NT分布的混合模型的LRS。NT的LR和生化标记的景观资源可以乘以年龄和妊娠相对危险率以获得各种条件,如21三体综合征风险度。检测率(DRS)和假阳性率(FPRS)可以通过采取高于给定的风险阈值的风险比计算。

[0600] 在一个实施方式中,识别倍性状态的方法包括将使用联合分布模型确定的各倍体假设的相对概率与使用源自确定胎儿三倍综合症风险度的其他方法的统计技术计算的倍性假设的相对概率相结合。包括但不限于:读计数分析,比较杂合率,仅亲本遗传信息才能获得的统计信息,某亲本背景的归一化基因型信号的概率统计信息,采用所估测的第一样本或制备样本以及它们的组合的胎儿分数来计算统计量。

[0601] 另一种方法涉及一种4个测得的激素水平的情况,围绕这些激素的概率分布是已知的:整倍体的情况下 $p(x_1, x_2, x_3, x_4 | e)$ 和非整倍体的情况下 $p(x_1, x_2, x_3, x_4 | a)$ 。可以测量DNA测量的概率分布,整倍体和非整倍体的情况下为 $g(y | e)$ 和 $g(y | a)$ 。假设他们是独立的整倍体/非整倍体的的给定假设,可以结合为 $p(x_1, x_2, x_3, x_4 | a) g(y | a)$ 和 $p(x_1, x_2, x_3, x_4 | e) g(y | e)$,然后每个乘以给定母体年龄的事先 $p(a)$ 和 $p(e)$ 。然后,可以选择一个最高的。

[0602] 在一个实施方式中,可以依中心极限定理以假设 $g(y | a \text{ or } e)$ 上的分布是高斯分布,并测量通过观察多个样本测量平均和标准偏差。在另一个实施方式中,可以假设它们不是独立的给定结果,并收集足够的样本来估计联合分布 $p(x_1, x_2, x_3, x_4 | a \text{或} e)$ 。

[0603] 在一个实施方式中,目标个体的倍性状态被确定为涉及最大概率的假设的倍性状态。在某些情况下,一个假设的归一化的,组合的概率大于90%。每个假设都涉及一个或一组,倍性状态,和涉及归一化的,组合概率大于90%的假设的倍性状态,或一些其他的阈值,如50%,80%,95%,98%,99%,或99.9%,可以被选择为被称为所确定的倍性状态的假设

所需的阈值。

[0604] 来先前怀孕中母体血浆的子本DNA

[0605] 非侵入性产前诊断的困难之一是将来自正在孕期的胎儿细胞与先前怀孕中的胎儿细胞区别开。有些人认为,来自怀孕之前遗传物质一段时间后会消失,但还没有确凿的证据显示。在本申请所公开的一个实施方式中,用PARENTAL SUPPORT™(PS)法和父本基因组信息能够确定父系起源的母体血液中的胎儿DNA的存在(即,胎儿DNA遗传自父亲)。这种方法可以利用阶段性的亲本遗传信息。使用祖父母的遗传数据(如从祖父精液测得的遗传数据),或来自其他出生的孩子或流产样本的遗传数据可以对来自未分阶段的遗传型数据的亲本基因型阶段化。人们还可以通过基于HapMap分阶段的方式来将为分阶段的遗传信息分阶段,或父本细胞的遗传信息的单体型。成功的单体型已通过染色体是紧密结合的减数分裂阶段捕获细胞,并采用微流控技术把不同的染色体置于不同的孔中被证明。在另一个实施方式中,可以使用分阶段父母的单倍型数据来检测到来自父亲的一个以上的同系物的存在,这意味着血液中存在来自一个以上的子本的遗传物质。通过浓缩预期将是胎儿整倍体的染色体,人们可以排除胎儿受三体综合症困扰的可能性。此外,确定胎儿DNA是否来自当前的父亲是可能的,在这种情况下,人们可以使用其他方法,例如三倍性检验来预测的遗传异常。

[0606] 有可能是其他来源的胎儿遗传物质可通过其他方法而不是抽血获得。在母体血液中可获得胎儿遗传物质的情况下,有两个主要的类别:(1)完整的胎儿细胞,例如,有核胎儿红细胞或有核红细胞,和(2)游离胎儿DNA。在完整胎儿细胞的情况下,有一些证据表明,胎儿细胞可以存留于母体血液中相当长的一段时间使得可以从含有来自以前怀孕的孩子和胎儿DNA的孕妇体内分离细胞。也有证据表明,游离胎儿DNA在大约几个星期从该系统消失。难题之一是如何确定其包含在细胞中的个体的身份,即,确保测得的遗传物质不是以前怀孕的胎儿。在本申请的一个实施方式中,母体遗传物质信息可以用来确保待测遗传物质不是母体遗传物质。有很多方法做到这一点,这包括基于信息学的方法,诸如PARENTAL SUPPORT™,如本发明件或任何在本发明件中引用的专利中所描述的。

[0607] 在本申请的一个实施方式中,来自怀孕的母体抽取的血液被分离成含有游离胎儿DNA的一小部分,以及含有有核红细胞的一小部分。所述游离DNA可任选地被富集,并且DNA的基因型信息可被测量。从所测得的来自游离DNA的基因型信息,母体的基因型信息可用于确定胎儿基因型的情况。在这些方面,可参考倍性状态,和/或一组等位基因特征。然后,可以使用在本发明的其他部分所述的以及其他相关专利,特别是那些对本发明件第一部分中提到的方法来对个体有核红细胞进行基因分型。

[0608] 母体基因组信息使得人们可以确定任何给定的单个血细胞是否遗传于母体。而且上述测得的胎儿的基因型情况使得人们可以确定单个血细胞是否遗传自目前妊娠的胎儿。本质上讲,本申请所公开的内容使得人们可以使用母体的遗传信息,和可能是来自其他相关的个体,如父亲的遗传信息,以及从来自母体血液的游离DNA所测得的遗传信息来确定从母体血液中分离出的有核细胞是否是(a)遗传自母体,(b)遗传自目前孕育的胎儿的,或(c)遗传自从前怀孕的胎儿。

[0609] 产前性别染色体非整倍体测定

[0610] 在本领域已知的方法中,人们试图从母体的血液确定妊娠中的胎儿的性别是利用

胎儿游离DNA (ffDNA) 是存在于母体血浆中的事实。如果人们能够在母体血浆中检测到Y染色体的特异性基因座,这就意味着妊娠中的胎儿是男性。然而,在使用在现有技术中已知的方法时,在血浆中没有检测到Y染色体的特异性基因座并不总是保证妊娠中的胎儿是女性,在某些情况下,作为的ffDNA量太低不足以确保Y染色体的特异性基因座在男性胎儿情况下被检测到。

[0611] 这里介绍的是一种新的方法,该方法不需要测量Y染色体特异性核酸,即,DNA单纯来自于父源的基因座。先前所披露的,亲本支持法,使用交叉频率数据,亲本基因型数据,和信息技术,来确定妊娠中的胎儿的倍性状态。胎儿性别的仅仅是胎儿性染色体倍性状态。一个孩子XX为女性,XY是男性。本发明所述的方法也能够确定胎儿的倍性状态。需要注意的是性别鉴定是性染色体倍性测定的有效的代名词;性别鉴定的情况下,假设往往是,孩子是整倍体,因此很少有可能的假设。

[0612] 本发明公开的方法包括寻找X和Y染色体共同的基因座以产生一个适合于胎儿的有关胎儿DNA预期量的基准。然后,只对 X染色体的特定区域可以询问,以确定是否胎儿是女性还是男性。在男性的情况下,我们希望看到来自X染色体的特异性基因座的胎儿DNA比来自X和Y染色体的特异性基因座的少。相反的,在女性胎儿的情况下,我们预计,这些组中每个的DNA的量都是相同的。待测DNA通过任何可以测量样本中存在的DNA 的量的技术测量,例如,qPCR,SNP阵列,基因分型阵列,或测序。对于完全是来自于一个个体的DNA,我们希望看到:

[0613]

	X 特异性 DNA	X 和 Y 特异性 DNA	Y 特异性 DNA
男性 (XY)	A	2A	A
女性 (XX)	2A	2A	0

[0614] 在来自胎儿的DNA与来自母体的DNA混合的情况下,混合物中的胎儿DNA的分数为F,而该混合物中的母体DNA的分数是M,这样 $F+M=100\%$,我们希望看到以下:

[0615]

	X 特异性 DNA	X 和 Y 特异性 DNA	Y 特异性 DNA
男性胎儿 (XY)	$M + \frac{1}{2} F$	$M + F$	$\frac{1}{2} F$
女性胎儿 (XX)	$M + F$	$M + F$	0

[0616] 在F和M是已知的情况下,可以计算预期的比例,所观察到的数据可以与预期的数

据进行比较。其中M和F都不知道的情况下,可以基于历史数据选择阈值。在这两种情况下,所测得的在X和Y两个特异性基因座的DNA的量可以用作基准,性别测试可以基于所测得的单纯X染色体的特异性基因座上的DNA的量进行。如果该量低于数值大致等于 $1/2F$ 的基准值,或由一个使它低于预定义的阈值的量,确定胎儿是男性,如果该量约等于基准值,或者如果不低一个使它低于一个预定义的阈值的量,胎儿被确定为女性。

[0617] 在另一个实施方式中,人们可以只寻找X和Y染色体共同的那些基因座,通常被称为Z染色体。Z染色体上的基因座的一个子集,通常总是一个X染色体上的A,而在Y染色体上的B。如果发现来自Z染色体的SNP有B基因型,则胎儿被识别为男性,如果发现来自Z染色体上的SNP只有A基因型,则胎儿被识别为女性。在另一个实施方式中,人们可以寻找仅在X染色体发现的基因座。如在B存在的的情况下,资料例如AA|B是特别翔实,则表明对胎儿有一条X染色体来自父亲。资料如AB|B也很翔实,则我们希望看到,与男性胎儿相比,B的存在往往只有女性胎儿的水平的一半。在另一个实施方式中,人们可以寻找其等位基因A和B均存在于X和Y染色体上的Z染色体上的SNP,其中SNP是来自父亲的Y染色体的,并且是来自父亲的X染色体是已知的。

[0618] 在一个实施方式中,可以将已知的单核苷酸位置扩增到Y染色体和X染色体共用的同源非重组(HNR)区域内不同。该HNR区域内的序列在X和Y染色体之间在很大程度上是相同的。这个相同的区域内有单核苷酸位置,当在人群中的X染色体,Y染色体之间不变时,其在X和Y染色体之间是不同的。每次PCR检测可扩增来自存在于X和Y两个染色体上的基因座的序列。在每个扩增的序列内都将有使用测序或其他方法可以检测到单个碱基。

[0619] 在一个实施方式中,胎儿性别可以由母体血浆中发现的胎儿游离DNA确定,该方法包括以下步骤中的部分或全部:1)设计PCR引物扩增X/Y变异体HNR区域内的单核苷酸位置(常规或微型PCR,如果需要附加倍增)2)获得母体血浆,3)用HNR X/Y PCR检测PCR扩增来自母体血浆的目标,4)测序扩增子,5)检查一个或多个扩增序列内所存在的Y型等位基因的序列数据。一个或多个的存在表示一个男性胎儿。所有扩增子中缺失所有Y等位基因表示是女性胎儿。

[0620] 在一个实施方式中,人们可以使用定向测序来测量母体血浆的DNA和/或亲本基因型。在一个实施方式中,人们可以忽略所有的明确源自父系来源的DNA的序列。例如,本发明中的AA|AB,可以对A序列数进行计数,并忽略所有的B序列。为了确定适于上述算法杂合率,人们可以将所观察的A序列的数与给定探针的总序列的预期的数目进行比较。人们可以有許多方法计算每个样本基础上的每个探针的预期的序列数。另一种方法是针对一些已知的纯合子等位基因,然后用历史数据使每个探头的读取数与已知的纯合子等位基因的读取数相符。对于每个样本,人们之后测量纯合等位基因的读取数目,然后使用该测量值,以及经验导出的关系,来估计在每个探针的序列读取数目。

[0621] 在一些实施方式中,通过结合由多种方法得出的预测结果可以确定胎儿的性别。在一些实施方式中使用本申请所述的多个方法。在一些实施方式中,所述多种方法中的至少一个是来自本申请中描述的方法。

[0622] 在一些实施方式中,本发明所述的方法可以被用来确定妊娠中胎儿的倍性状态。在一个实施方式中,倍性识别方法使用X染色体的特异性基因座,或X和Y染色体的共同的基因座,但不使用任何Y染色体特异性基因座。在一个实施方式中的倍性的识别方法使用一个

或多个以下操作：X染色体特异性基因座，X和 Y两种染色体共同的基因座，以及Y染色体的特异性基因座。在一个实施方式中，其中性染色体的比率是相似的，例如，45,X（Turner综合征），46,XX（正常女性）和47,XXX（三体综合征X），所述分化可以通过将等位基因分布与根据各种假说预期的等位基因分布进行比较实现。在另一个实施方式中，这可以通过将性染色体的序列的相关读取数与一个或多个被假定为整倍体的对照染色体相比较实现。还要注意的，这些方法可以扩大到包括非整倍体的情况下。

[0623] 单基因病诊断

[0624] 在一个实施方式中，用于确定胎儿的倍性状态的方法可能会延伸使单基因疾病同时检测能够进行。单基因病诊断将相同的定向法变为用于非整倍体检测，并需要额外的具体目标。在一个实施方式中，单基因的NPD诊断是通过连锁分析的方式进行。在许多情况下，直接测试的cfDNA样本是不可靠的，因为母体DNA的存在使得确定胎儿是否继承母体的突变几乎不可能进行。独特的父系来源的等位基因的检测是具有挑战性的，但如果疾病是显性的，并由父亲携带，仅仅完全是提供信息，由此限制了该方法的利用。在一个实施方式中，所述方法包括PCR或相关的扩增方法

[0625] 在一些实施方式中，该方法包括利用来自一级亲属中的信息通过周围的联系非常紧密的亲本中的SNP来对异常等位基因分阶段。然后，亲本支持可能基于来自这些SNP的定向测序数据进行，以判断同源性，正常或不正常，由胎儿从父母双方那儿继承。只要SNP有足够的联系，胎儿的基因型的继承可以非常可靠地确定。在一些实施方式中，该方法包括（a）将一组SNP位点加入到稠密侧指定的一组常见的疾病到用于非整倍体检验的多重池；（b）用来基于自不同亲属的遗传数据的正常和不正常的等位基因对于来自添加的SNP的等位基因可靠的划分阶段，及（c）重建胎儿二倍性，或分阶段SNP等位基因继承母体和父亲的同系物在周边地区的疾病基因，以确定胎儿的基因型。在一些实施方案中，添加与疾病密切相关的另外的探针，是紧密相连的一种疾病挂钩轨迹的一套被用于非整倍体检测的多态性位点。

[0626] 重组胎儿双倍型的是具有挑战性的，因为样本是孕母体和胎儿DNA的混合物。在一些实施方式中，该方法加入阶段性的单核苷酸多态性与疾病等位基因相关信息，然后考虑到来自位置特异性重组的似然度和从母体血浆遗传测量中观察到的数据，以取得最可能的胎儿基因型的单核苷酸多态性与重组数据的物理距离。

[0627] 在一个实施方式中，一组定向多态性位点包含了每种疾病连接位点一些额外的探针，每种疾病连接位点的额外的探针的数量可能在4和10之间，在11和20之间，在21和40之间，在41和60之间，在61和80，或是它们的组合。

[0628] 样本中的DNA分子数的确定

[0629] 本发明中描述了一种方法，通过产生一种对DNA扩增的第一轮过程中样本中的每个原始DNA分子的唯一识别分子来确定样本中的DNA分子的数量。这里描述一个程序，实现上述的最后部分，然后是单分子或克隆测序法。

[0630] 该方法需要针对一个或多个特异性位点，并产生原始分子的一个标记的拷贝，以这种方式，来自每个目标位点的大多数或所有的标记分子将有一个唯一的标记，并且其序列彼此不相同，该条形码使用克隆或单分子测序。每一个单一序列的条形码序列代表原样本中的一个独特的分子。同时，测序数据被用来确定分子起源的位点。使用此信息，可以决定每个基因座在原始样本中的独特分子的数量。

[0631] 该方法可用于定量评价的原始样本中分子的数目所必需的任何应用中。此外，一个或多个目标的独特分子的数目可以与一个或多个其他目标的独特的分子的数量相关联以确定相对拷贝数，等位基因分布，或等位基因的比例。另外，从不同的目标检测到的拷贝的数目可以由分布建模，以便确定最有可能的原目标的拷贝数量。应用领域包括但不限于检测的插入和删除，如发现杜氏肌营养不良症携带者；量化染色体的缺失或重复片段，如观察到的拷贝数突变体；来自出生的个体样本的染色体拷贝数；来自未出生的个体样本，如胚胎或胎儿的染色体拷贝数。

[0632] 该方法可同时与评估包含在针对序列的突变体结合。这可以用来确定代表原样本中的每个等位基因的分子的数量。此拷贝数法，可以与评价SNP位点或其它序列变异体结合，以确定出生和未出生的个体的染色体拷贝数；来自具有短的序列变异体的基因位点的拷贝的鉴别和定量，但其中PCR可由多个目标区域扩增，如脊肌萎缩的载波检测；来自含有不同个体的混合物的样本的不同来源的分子的拷贝数的测定，如源自母体血浆的游离DNA的胎儿非整倍体检测。

[0633] 在一个实施方式中，所述涉及到单一目标位点的方法可以包括一个或多个以下步骤：(1) 设计用于特异性位点的PCR扩增的标准的低聚物对。(2) 添加，在合成过程中，将特异性碱基非或最小的互补性加入到目标位点或将基因组加到一个特定目标的低聚物的5'端。这个序列中，被称为末端，是一个已知的序列，以用于后续扩增，然后是一个随机核苷酸序列。这些随机核苷酸包含随机区域。随机区域包含每个探针分子之间的概率不同的随机生成的核酸序列。因此，下述合成，末端低聚物池由以已知的序列开头的，接着是分子间不同的未知的序列，然后是特异性目标序列的低聚物的集合组成。(3) 仅使用末端低聚物进行一轮扩增(变性，退火，延伸)。(4) 核酸外切酶加入反应，有效地终止PCR反应，该反应在适当的温度下温育，以除去没有退火至室温和延伸形成的双链产物的单链寡核苷酸。(5) 高温孵化反应以使核酸外切酶变性，并消除其活性。(6) 将新的寡核苷酸添加到该反应以与和其他定向特异性低聚物一起用于用于第一反应的低聚物的尾部互补，以使由第一轮PCR产生的产物进行PCR 扩增。(7) 继续扩增以产生足够的产品用于下游克隆测序。(8) 通过多种方法测量扩增的PCR产物，例如，克隆测序，有足够数目的碱基覆盖序列。

[0634] 在一个实施方式中，本申请方法的涉及并行地或以其他方式针对多个位点。不同靶位点的引物可以独立产生，并混合来创建多重PCR池。在一个实施方式中，原始样本可以被分为子池，而不同位点可以在重组和测序之前定向每个子池。在一个实施方式中，标记步骤和许多扩增循环可以在该池被细分前进行以确保在分裂前的所有目标都能有效的定位，使用细分池中更小的组的引物通过连续扩增来改善后续扩增。

[0635] 一种应用的例子，这项技术尤其特别有用的是非侵入性产前非整倍性诊断，其中给定的基因座的等位基因的比例或多个基因座的等位基因分布可以用来帮助确定存在于胎儿中的染色体的拷贝的数量。在这种情况下，需要扩增存在于初始样本中的DNA，同时保持各等位基因的相对量。在某些情况下，特别是在DNA 是一个非常小的量，例如，少于5,000个拷贝的基因组，少于1,000个拷贝的基因组，少于500个拷贝的基因组中，且少于100个拷贝的基因组，人们可以遇到被称为瓶颈这种现象。这是初始样本中的任何给定的等位基因的拷贝中的一小部分，扩增偏差可以导致其那些等位基因的比值比在初始混合物中的DNA具有显著的不同的DNA的扩增池。在标准的PCR扩增前通过将一种独特的或几乎唯一的条形

码用于DNA的每一条链,它可能从源于相同的原始分子的一组n个相同序列的DNA分子排除n-1个DNA拷贝。

[0636] 例如,假设在个体基因组中的有一个杂合SNP,并且来自每个等位基因分子的个体的DNA的混合物存在于DNA原始样本中。扩增后有可能对应于该位点有100,000个DNA分子。由于随机过程,DNA的比例可能是为1:2至2:1的任何数,但是,由于每个原分子都采用一个独特的标记来标记,这将有确定扩增池中的DNA源于每个等位基因整整10DNA分子。因此,该方法将比不用这种方法得到更准确的每个等位基因的相对量的测量值。对于需要将等位基因偏差的相对量最小化的方法,该方法将提供更准确的数据。

[0637] 测序片段与目标位点的结合可以多种方式来实现。在一个实施方式中,有足够的长度的序列可以从目标片段获得以覆盖分子条形码,以及对应于靶序列足够数量的独特的碱基对,以实现目标位点的明确识别。在另一个实施方式中,包含随机生成的分子条形码的分子的条形码编码引物也可以包含识别将要关联的目标的条形码的特异性条形码(位点条形码)。这种位点条形码在针对个体目标的所有分子条形码引物之间都相同,因此所有由此产生的所有扩增子,但不同于所有其他目标。在一个实施方式中,本发明所描述的标记方法也可以与单侧嵌套协议结合。

[0638] 在一个实施方式中,分子条形码编码引物的设计和生成可能被化为如下实践中:分子条形码编码引物可以由不与继之以随机分子的条形码区域,继之以目标特异性序列的目标序列互补。分子条形码的该序列5'可以用于子序列的PCR扩增,并且可以包含用于测序的库的扩增子变换的序列中。可以多种方式生成的随机分子条形码序列。优选的方法以这样一种方式合成标记引物分子,以包含条形码区域的合成过程中的所有四种碱基。所有或各个碱基的组合可以使用IUPAC DNA歧义代码指定。以这种方式合成的分子集合将包含一种在其分子的条形码区域的序列的随机的混合物。条形码区域的长度将决定多少引物将包含独特的条形码。独特的序列的数目与条形码区域的长度有关,其中N是碱基数,通常为4,L是条形码的长度。含有5个碱基的条形码可达1024 独特序列;含有8个碱基的条形码可以产生65536个独特的条形码。在一个实施方式中,该DNA可以用测序法测量,其中,序列数据表示单分子序列。这可以包括单分子直接测序的方法或单分子通过测序仪扩增形成克隆,但仍然代表单分子,在此称为克隆测序的方法。

[0639] 实施方式

[0640] 在一些实施方式中,本发明公开了一种用于形成提供所确定的妊娠中的胎儿的染色体倍性状态的报告的方法,该方法包括:获取含有源自胎儿母体的母本DNA和来自胎儿的胎儿DNA的 DNA第一样本;从胎儿父母的一方或双方获取基因型数据;通过分离DNA的方式制备第一样本以便获得制备样本;测量多个多态性位点的制备样本中的DNA;用计算机计算由制备样本测得的DNA计算多个多态性位点的等位基因数或等位基因数概率;用计算机创建关于与染色体不同的可能倍性状态相应的染色体上多个多态性位点的期望等位基因数概率的假设;采用联合分布模型和适合于所述制备样本的等位基因数概率用计算机建立每个倍性假设相对性概率;通过选择对应于具有最大概率的假设的倍性状态的方式来识别胎儿的倍性状态;以及形成提供所确定的倍性状态的报告。

[0641] 在一些实施方式中,该方法是用来确定多个各自的母体中的多个妊娠中胎儿的倍性状态的方法,该方法还包括:确定源于制备样本的胎儿的DNA的百分数;其中,测量制备样

本中的DNA的步骤是通过对每个制备样本中的多个DNA分子的测序进行的,其中经测序较多DNA分子来自具有较小的分数的胎儿DNA的制备样本而不是具有较大分数的胎儿DNA的制备样本。

[0642] 在一些实施方式中,该方法是用来确定多个各自的母体中的多个妊娠中胎儿的倍性状态的方法,其中,测量制备样本中的DNA,对于每个胎儿而言,是通过对第一部分制备样本的DNA进行测序得到的第一组测量值,所述方法还包括:测量每个胎儿的每个倍性假设的第一相对概率,给定的第一组DNA测量值;对来自这些胎儿的第二部分制备样本进行再次测序,其中每个倍性假设的第一个确定相对概率表明对应于非整倍体胎儿倍性假设具有显著的但不是决定性的概率,以给出第二组测量值;使用第二组测量值和任选的第一组测量值进行第二相对概率测定来获得胎儿的倍性假设;通过选择对应于依据第二相对概率测定值具有最大概率的假设的倍性状态来识别再次测序的第二样本的倍性状态。

[0643] 在一些实施方式中,本发明公开了一种成分的物质,所述成分的物质包括:具有优先富集的DNA的样本,其中所述优先富集的DNA样本从源自第一样本的DNA的多个多态性位点优先富集,其中所述第一样本的DNA由来自母体血浆的母体DNA和胎儿DNA的混合物组成,其富集程度的至少为2倍,并且其中的第一样本与优先富集的样本之间的等位基因偏置,平均,选自由小于2%,小于1%,小于0.5%,小于0.2%,小于0.1%,小于0.05%,小于0.02%,小于0.01%构成的集合中。在一些实施方式中,公开了一种产生具有这种优先富集的DNA的样本方法。

[0644] 在某些实施方式中,公开了一种用于确定包括胎儿和母体的基因组DNA的母体组织样本中存在或不存在胎儿非整倍性的方法,该方法包括:(a)从所述母体组织样本中获取胎儿和母体的基因组DNA的混合物;(b)选择性地富集胎儿和母体DNA的混合物的多个多态性等位基因;(c)分配选择性富集的源自胎儿和母体的基因组DNA的混合物的片段,具有提供含有单一的基因组DNA分子或单一的基因组DNA分子的扩增产物反应样本的步骤(d)对在步骤c的反应样本中选择性富集的基因组DNA片段进行大规模并行的DNA测序以确定所述选择性富集的片段的序列;(e)识别步骤d中得到的染色体的序列;(f)分析步骤d的数据来确定i)来自步骤d的属于至少一个被推定为母体和胎儿的第一目标染色体的基因组DNA片段的数目,及ii)来自步骤d的属于第二目标染色体的基因组DNA片段的数目,其中所述第二染色体被怀疑是在胎儿体内的非整倍体;(g)如果所述第二目标染色体整倍体,使用在步骤f的i)部分确定的数来计算来自步骤d的第二目标染色体的基因组DNA片段的数目的预期的分布;(h)如果所述第二目标染色体是非整倍体,使用步骤f的i)部分的第一个数和步骤b)的混合物中发现的胎儿DNA的估计分数来计算来自步骤d的所述第二目标染色体的基因组DNA片段数的预期的分布;以及(i)使用最大似法或最大后验法来确定是否在步骤f)的部分ii)确定的基因组DNA片段的数量更可能是部分在步骤g)中计算出的部分分布,或在步骤h)中计算出的分布,从而表明胎儿非整倍体的存在或不存在。

实施例

[0645] 本发明公开的实施方式在下面的实施方式中进行描述,其被详细解释说明以帮助理解本发明,其不应该被解释为以任何方式限制由其后的权利要求所限定的本申请的范围。提出下列实施方式以向本领域中的普通技术人员提供如何使用所描述的实施方式的完

整的披露和说明,并且既非旨在限制本申请的范围,也非旨在代表的下面的实验都是全部的或唯一的实验。已作出努力以确保所使用数的准确性(如金额,温度等),但仍存在一些实验性的错误和偏差。除非另有说明,份是(体积)份,温度为摄氏度。应该理解的是,所描述的方法可以在不改变该实验的想要说明的基本方面有所变化。

[0646] 试验1

[0647] 本实验的目的是表明贝叶斯最大似然估计(MLE)算法,与公布的方法相比,该算法使用父母的基因型来计算胎儿分数提高了非侵入性产前三体性诊断的准确性。

[0648] 母体cfDNA的模拟测序数据可以通过基于三倍性-21和各自的母体细胞系获得的取样读数产生。正确的二体性和三体性识别比率以适于公布的方法(Chiu等人BMJ2011; 342:c7401)和基于MLE的算法不同胎儿分数由500个模拟来确定。我们通过基于IRB批准协议从四个孕妇和各自的父亲获得的500万鸟枪法读数验证了模拟。亲本的基因型通过290K的SNP阵列获得。(见图14)

[0649] 在模拟中,胎儿分数低至9%时,基于MLE的方法达到了99.0%的准确率,所报道的置信度与总精度对应良好。我们使用四个实际样本验证了这些结果,这四个实际样本中我们得到所计算的置信度超过99%的完全正确的识别。相比之下,我们实施Chiu等人公布的算法,需要18%的胎儿成分,达到99.0%的准确率,而9%的胎儿DNA只能达到87.8%的准确率。

[0650] 源自父母的基因型的胎儿分数测定值与基于MLE的方法结合对于第一和第二孕期早期所预期的胎儿分数可以比公布的算法达到更高的精度。此外,本发明所公开的方法产生置信度量,对于测定结果的可靠性是至关重要的,尤其是对于倍体检测比较困难的低胎儿成分。所发布方法使用准确度较低的阈值法基于多组二倍体培训数据和预定义了假阳性率的一种方法来识别倍性。此外,如果没有一个置信度量,当没有足够的cfDNA来进行识别时,所公开的方法报告胎儿的假阴性结果的是有风险的。在一些实施方式中,置信估计适合于所识别的胎儿的倍性状态。

[0651] 试验2

[0652] 本实验的目的是通过使用定向测序法与与父母的基因型和贝叶斯最大似然估计(MLE)算法的HapMap数据相结合的方式改善胎儿三倍性18,21,和X特别包括低分数胎儿的样本的非侵入性检测。

[0653] 来自四个整倍体和两个三倍性阳性的孕妇的母体样本和各自父亲的样本是在IRB批准的协议下从胎儿染色体组型已知的父母获得的。母本cfDNA是从血浆中提取的,并获得大约1000万序列读取,接着优先富集定向特异性SNP。同样,对父母样本进行相似测序以获得基因型。

[0654] 上述算法正确地识别所有整倍体样本的18和21号染色体二倍性和非整倍体样本的正常染色体。18和21三体性倍正确识别,正如在男性和女性胎儿的X染色体拷贝数。在所有情况下,通过算法产生的置信度都超过98%。

[0655] 所述方法准确地报道了源自6个样本的所有测试的染色体倍性,其包括含有小于12%的胎儿DNA的样本,其占了大约30%的第一孕期和第二个孕期初的样本。瞬时MLE算法和公布的方法之间的关键的区别在于它使得父基因型和HapMap数据精度提高,并产生一个置信度量。在低胎儿成分条件下,所有方法的精度均变差,重要的是在胎儿cfDNA不足以进

行可靠识别时要正确地识别样本。其他研究人员已经使用了Y染色体特异性探针来估测男性胎儿的胎儿分数,而同时父母基因型使男女双方的胎儿分数估测成为可能。公布的方法,利用不相关的鸟枪法测序的另一个固有限制是倍性识别精度因不同因素,如GC丰度,而在染色体间变化。瞬时定向测序法在很大程度上是独立于这种染色体规模变化并获得较一致的染色体之间的性能。

[0656] 试验3

[0657] 本实验的目的是要确定是否可以利用新型信息学分析母体血浆中游离胎儿DNA的SNP位点以高置信度检测出三倍体胎儿的三倍性。

[0658] 一名怀孕的患者采取20ml的血液接着进行异常超声。离心后,从白细胞层提取母本DNA(DNEASY,QIAGEN);从血浆中提取细胞游离DNA(QIAAMP QIAGEN)。在两个DNA样本中的染色体2,21,X上应用定向测序分析SNP位点。最大似然贝叶斯估计选择从所有可能的倍性状态组最有可能的假设。该方法可以确定胎儿DNA分数,倍性测定中的倍性状态和明确的置信度。关于对照染色体倍性不作任何假设。所述诊断使用与序列读取计数无关的检验统计量,这是本领域近期最先进的水平。

[0659] 即时法准确诊断第2和第21号染色体的三倍体症。子本分数估计为11.9% [CI11.7-12.1]。胎儿被发现含有第2和第21号染色体的一个母亲和两个父亲的拷贝有效置信度为1(概率误差 <10⁻³⁰)。这是分别是由第2和21号染色体上的92,600和258,100 读数实现的。

[0660] 这是来自其胎儿是三倍体症的母体血液的三倍体症染色体的非侵入性产前诊断第一证据,已通过中期染色体组型确认。现存的非侵入性诊断方法不会在这个样本中不会检测到非整倍体。目前的方法依赖于相对于二体对照染色体更多剩余的三倍体染色体的序列读数。此外,现存的方法对于这种分数的胎儿DNA 和这种数量的序列读数将无法达到同样高置信度。将该方法扩展到所有24个染色体很简单。

[0661] 试验4

[0662] 采用以下协议使用标准的PCR对来自整倍体妊娠的母体血浆分离出的DNA还有来自三倍体症第21号细胞系的基因组 DNA进行800-重扩增(意味着不使用嵌套)。库的制备和扩增涉及具有A尾化的单管钝端化。采用AGILENT SURESELECT试剂盒中找到的连接试剂盒进行适配子连接,PCR反应进行7个周期。然后,在第2、21和X染色体上使用800个不同的定向单核苷酸多态性位点的引物进行15个周期的STA(95℃,30秒;72℃,1分钟;60℃,4分钟;65℃,1分钟;72℃,30秒)。以12.5nM 的引物浓度进行反应。随后用ILLUMINA IIGAX测序器对DNA 测序。该测序器输出190读数,其中92%读取到基因组中;这些读取到到基因组中的读数,99%以上读取到定向的引物的一个定向的区域中。血浆DNA和基因组DNA的数目基本上都是相同的。图15示出通过测序器检测出的基因组DNA中的~780个SNP位点中的两个等位基因的比率,所述基因组DNA取自的具有已知 21号染色体的三体综合症的细胞系。注意这里示出的等位基因比是为方便可视化,因为等位基因分布不是简单的视觉阅读的。圆圈代表二体染色体上的SNP,而星星代表三倍体症染色体上的 SNP。图16是与图X中相同的数据的另一种表示,其中Y轴为每个SNP测定的A和B的相对数量,而X轴是由染色体分离的SNP的SNP数。在图16中,2号染色体上发现1至312个SNP,21号染色体上发现313至605个SNP,这是三倍体症,在X染色体上有606到800个SNP。来自染色体2和X的数据表明二体染色体,以三组

的相对序列数显示:AA在图的顶部,BB在图的底部,而AB在图的中间。源自21号染色体的数据,这是三倍体症,显示了四组:AAA在图的顶部,AAB在线0.65周围(2/3),ABB在线0.35的周围(1/3),BBB在图的底部。

[0663] 图17示出相同的800-重协议的但来自孕妇的四个血浆样本的DNA上测得的数据。对于这四个样本,我们希望看到七组点:(1)沿图的顶部的这些位点中母体和胎儿都是AA,(2)在稍微低于图的顶端那些位点中母体是AA和胎儿是AB,(3)略高于线0.5的那些位点中母体是AB和胎儿是AA,(4)沿0.5线的那些位点对母体和胎儿都AB,(5)稍低于0.5的线那些位点中母体是AB和胎儿是BB,(6)略高于图的底部这些位点中母体是BB和胎儿是AB,(1)沿图的底部的那些位点中母体和胎儿都是BB。较小的胎儿分数,则集群(1)和(2)之间,群集(3),(4)及(5)之间,以及集群(6)和(7)之间的更少分离。

[0664] 分离预期为胎儿来源的DNA分数的一半。例如,如果DNA有20%胎儿的,有80%是母体的,我们预期(1)至(7)分别集中于1.0,0.9,0.6,0.5,0.4,0.1和0.0,例如见图17,POOL1_BC5_ref_rate。如果DNA改为,8%是胎儿,92%是母体的,我们预期(1)至(7)被集中在1.00,0.96,0.54,0.50,0.46,0.04和0.00;例如,参见图17,POOL1_BC2_ref_rate。如果没有检测到胎儿DNA,我们不期望看到(2),(3),(5)或(6);取而代之的,我们可以说分离度是零,因此,(1)和(2)在彼此的顶部,(3),(4)及(5),以及(6)和(7)同样,例如见图17,POOL1_BC7_ref_rate。请注意,在图17中,胎儿分数POOL1_BC1_ref_rate为25%左右。

[0665] 实验5

[0666] 大多数DNA扩增和测量方法都会产生一些等位基因偏差,其中的两个等位基因,通常会发现在一个位点进行检测,其强度或计数并不代表DNA样本的等位基因的实际数量。例如,对于一个单一的个体,在杂合位点,我们期望看到两个等位基因的比例为1:1,这是杂合基因座预期的理论比;但由于等位基因偏差,我们可以看到55:45,甚至至60:40。另外请注意,在测序的背景下,如果读取的深度低,那么简单的随机噪声可能会导致显著的等位基因偏差。在一个实施方式中,这可能是每个SNP的行为模型,这样如果观察到特定等位基因的一致的偏差,该偏置可以被校正。图18显示了偏差校正之前和之后可以用二项式方差进行解释部分数据。在图18中,星号表示800-重试验的原始序列数据所观察到的等位基因偏差;圆圈代表校正后的等位基因偏差。需要注意的是,如果根本没有等位基因偏差,我们希望数据沿着x=y线下降。使用150重定向扩增产生的数据扩增DNA产生一组类似的数据与偏差校正后的1:1线非常接近。

[0667] 试验6

[0668] 使用连接适配子,以适配子标记的特异性引物,对DNA进行全扩增,其中引物退火和延伸时间限制到几分钟对富集部分更短的DNA链有影响。大多数库协议被设计用于创建适合于包含这样的步骤的测序的DNA库,示例协议是公开的和本领域公知的。在本发明的一些实施方式中,含有通用标记适配子与血浆DNA相连,并使用适配子标记特异性引物扩增。在一些实施方式中,通用标记可与测序中所用的标记相同,它可以是仅用于PCR扩增的通用标记,或者它可以是一组标记。由于胎儿DNA一般本来就短,而母体的DNA本质上可以可长和可短,因此,这种方法对于富集混合物中胎儿DNA的比例有影响。游离DNA,被认为是源自凋亡细胞的DNA,其中包含胎儿和母体的DNA,是短的-大多在200基点。细胞DNA由细胞裂解释放,放血后的一个普遍现象,通常是几乎完全是母体,也是相当长的-大多是在500基点

以上。因此,周围的超过几分钟的血样将包含短(胎儿+母体)和更长的(母体)的DNA。以相对较短的延伸时间对母体血浆进行全扩增,然后再进行定向扩增,与单纯使用定向扩增的血浆相比,往往会增加胎儿DNA的相对比例。这可以从图19看出,当输入是血浆DNA(垂直轴)时,测得的胎儿分数,而当输入DNA是具有使用ILLUMINA GATIIx库制备协议制备的库的血浆DNA(垂直轴)时,测得的胎儿百分比。所有的点落在下面的行,表示该库准备工序丰富的分数胎儿来源的DNA。两个血浆样本均为红色表示溶血,因此会有从细胞裂解长母体DNA目前量增加,显示胎儿分数库时准备进行有针对性的扩增前的一个特别显著的富集。本发明所公开的方法是特别有用的情况下,有溶血或其他一些情况时有发生,其中包括相对长链的DNA污染的细胞裂解,与长的DNA短的DNA混合样本污染。通常情况下,相对短的退火和延伸时间是30秒和2分钟之间,尽管它们可能是尽可能短,5或10秒或更少,或只要是5或10分钟。

[0669] 试验7

[0670] 1,200复杂的DNA扩增分离从母体血浆从整倍体妊娠的基因组DNA从细胞株三倍体21使用直接PCR协议,也是一个半嵌套的方法用于以下协议。库准备和扩增涉及单管A尾钝结局。使用连接试剂盒,发现在安捷伦SureSelect试剂盒的变形例的运行适配子连接,PCR反应进行7个周期。有针对性的引物池,共有550从21号染色体上的SNP检测,325从每1和X染色体的SNP检测这两个协议涉及15个周期的STA(95°C 30秒,72°C,1分钟;60°C的4分钟65°C 30秒,72°C 30秒),采用16 纳米引物浓度。半嵌套PCR协议涉及15个周期的STA(95°C 30秒,72°C,1分钟,60°C,4分;65°C的30秒72°C 30秒)第二扩增使用内前向标记浓度为29nM,和反向标记1uM 的0.1uM的浓度。随后的DNA测序与Illumina的IIGAX序器。对于直接PCR协议,73%的读取映射到基因组中的半套式协议,97.2%的序列读取映射到基因组中的。因此,半嵌套协议,结果在约30%以上的信息,想必这主要是由于引物是最有可能导致引物二聚体消除。

[0671] 读变异的深度趋向较大,比直接PCR的协议是用来钻石参阅位点的读出与半嵌套协议运行的深度(参见图20),当使用半嵌套协议和正方形是指读位点的运行没有嵌套深度。深度阅读SNP 位点被布置为钻石,这样的钻石都属于曲线上,而出现的平方松散corelated的单核苷酸多态性的安排是任意的,并且是深度的点表示的高度读取而不是它的左到右的位置。

[0672] 在一些实施方式中,本发明所述的方法可以实现优异的深度读(DOR)的方差。例如,在这个实验中的一个版本(图21)使用1200复杂的直接PCR扩增的基因组DNA,1200检测:1186 测定有DOR大于10读出的平均深度为400;1063测定(88.6%)的深度200和800之间的读取,读取每个等位基因的数量是足够高,以提供有意义的的数据,而读取的每个等位基因的数量是没有那么高的理想的窗口的边缘使用那些读特别小。只有12个等位基因有较高的深度阅读1035读取最高。在DOR的标准偏差为290,平均DOR为453,本DOR的方差系数为64%,有950000 总读取,读取映射到基因组中的63.1%。在另一个实验中(图22),使用1200复杂的半套式协议,在DOR较高。的DOR的标准偏差为583,平均DOR为630,的DOR方差系数为93%,87万总读取,读取映射到基因组中的96.3%。请注意,在这两种情况下,单核苷酸多态性的深度阅读为母,使弯曲线代表读母体深度的安排。父亲和孩子之间的分化并不显著,它是唯一的趋势是显着的这个解释的目的。

[0673] 试验8

[0674] 在一个实验中,半嵌套1200重PCR协议用于从一个细胞以及从三个细胞扩增DNA。这个实验是有关使用从母体血液分离出的胎儿细胞的产前非整倍体试验或用于用活检卵裂球或滋养外胚层样本进行胚胎植入前遗传学诊断。每个条件有3个重复,每2个体重复1和3细胞(46XY和47XX+21)。分析定向染色体1,21和X。一个测序泳道中跑48重样本进行测序:ARCTURUS,MPERv2和碱裂解。测序复运行在一个测序车道。每个重复,该算法对每次重复的三个染色体都要返回正确的倍性识别。

[0675] 试验9

[0676] 在一项实验中,制备四个母体血浆样本,并采用半嵌套9,600 复用协议扩增。此样本的制备按以下方式:对最多40毫升母体血液倍离心并分离为白细胞层和血浆。母体的基因组DNA制备由白细胞层制备,而父体的DNA由血液样本或唾液样本制备。母体血浆中的脱细胞DNA使用QIAGEN循环核酸试剂盒分离,并根据制造商的说明,在45微升TE缓冲液中洗脱。通用连接适配子被追加到35微升纯化血浆DNA的每个分子端和使用适配子的特异性引物进行扩增库7个周期。库用AGENCOURT AMPURE 珠纯化并用50微升水洗脱。

[0677] 对3微升的DNA用14.5nM的引物浓度的9600定向特异性标记的反向引物和一个库适配子特异性正向引物(500nM)扩增 15个STA循环(聚合酶初始激活95℃,10分钟;然后15个周期,95℃,30秒;72℃,10秒;65℃,1分钟;60℃,8分钟,65℃,3分钟和72℃,30秒;最后在72℃下延长2分钟)。

[0678] 半嵌套PCR协议包括15个STA周期的第一STA产品的稀释产物的第二次扩增(聚合酶初始激活95℃,10分钟,然后15 个循环,95℃,30秒;65℃,1分钟;60℃,5分钟;65℃,5 分钟和72℃,30秒;72℃下最后延长2分钟)使用1000nM的反向标记浓度,每个9600定向特异性正向引物的浓度为16.6u nM。

[0679] 等分的STA产物之后通过标准PCR以1uM的标记特异性正向和条形码反向引物扩增10个循环以产生条形码测序库。等分的每个库与具有不同的条形码的,使用核酸纯化柱纯化的库混合。

[0680] 以这种方式,9600条引物被用于在单井反应;所述引物倍设计用于染色体1,2,13,18,21,X和Y上发现的目标SNP。用ILLUMINA GAIIX测序器对扩增产物进行测序。每个样本由测序器产生约3.9百万读数,3700000读数映射到基因组(94%),并且,290万读数(74%)映射到目标SNP,平均读取深度为344,读取中间深度为255。四个样本的胎儿分数被发现是9.9%,18.9 %,16.3%,和21.2%。

[0681] 有关母体和父亲的基因组DNA样本扩增9600半嵌套重协议和测序。半嵌套的协议是不同的,它适用于9600外的正向引物和标记的反向引物在第一STA7.3纳米。热循环条件和组合物,所述第二STA,和条形编码PCR为半嵌套协议相同。

[0682] 用本发明所公开的信息学的方法进行测序数据分析,被称为倍性状态在6条染色体的胎儿的DNA中存在4母体血浆样本。集合中的所有28条染色体被称为套数识别正确的置信度在99.2 %以上,除一条染色体被称为正确的,但与83%的置信度。

[0683] 图23示出了1200复用半嵌套方法的读取深度和实验7中所述的半嵌套的的9600复用半嵌套方法的读取的深度,但读取的SNP位点的数量与深度大于100,大于200大于400,显着高于在1200复用协议中的。上面的第90个百分位的读取次数,可划分的数在第10百分读

取,得到读取的深度的均匀性的指标的一个无量纲的度量,数字越小,越均匀(狭窄)的深度读。平均90 percentile/10th的百分位的比率为11.5实验9中运行的方法,虽然它是5.6实验7中运行的方法。较窄的阅读深度,对于一个给定的协议的复杂性是更好的测序效率,较少的序列读出是必要的,以确保目标读取比例以上的读数阈值。

[0684] 试验10

[0685] 在一项实验中,制备四个母体血浆样本,并用半嵌套9600 重协议扩增。实验10的详细信息与试验9非常相似,不相似之处是嵌套协议,还包括四个样本的特征。对所述组中所有28条染色体进行倍性识别,其正确识别的置信度大于99.7%。7.6百万(97%) 读数被读取到基因组中,而6.3百万(80%)的读数被读取到定向SNP中。读取的平均深度为751,而读取深度的中间数为396。

[0686] 试验11

[0687] 在一项实验中,三个母体血浆样本被分成5等分,各等分使用2400多重引物(4份)或1200多重引物(1份) 扩增,并用半嵌套协议扩增,共10,800引物。扩增后,各部分合并测序。实验11的详细信息与实验9非常相似,不相似之处是嵌套协议,还有分裂和合并的方法。对所述组中所有21条染色体进行倍性识别,其正确识别的置信度大于99.7%,除了一条未识别的置信度为83%外。340万读数读取读取到定向SNP,读取的深度为 404和读取深度的中间数为258。

[0688] 试验12

[0689] 在一项实验中,四个母体血浆样本分割成四个相等的部分,并使用2400复用的引物扩增的各部分,使用半嵌套协议,以总的9600引物扩增。扩增后,合并的部分测序。有关实验12实验细节与试验9非常相似,异常之处是嵌套协议,分裂和合并的方法。对集合中的所有28条染色体的倍性识别被正确地识别,置信度在97%以上,除了一个丢失的识别的置信度为78%外。450 万读取映射到有目标SNP,平均读取深度为535和读取深度的中间数412。

[0690] 试验13

[0691] 在一项实验中,制备四个母体血浆样本,并使用9600重三重半嵌套协议进行扩增,共9600个引物。有关实验12的详细信息与实验9非常相似,唯一不同的是嵌套协议,该协议涉及三轮扩增;三轮分别涉及15,10和15个STA周期。对所述组中所有 27和28条染色体进行倍性识别,其正确识别的置信度大于99.9 %,除了一条正确识别的置信度为94.6%,一条未识别的置信度为80.8%外。350万读数读取读取到定向SNP,读取的深度为414 和读取深度的中间数为249。

[0692] 试验14

[0693] 在一项实验中,45套细胞扩增1,200复杂的半嵌套协议,测序,并在三个染色体套数确定。请注意,这个实验是为了模拟的条件进行单细胞活组织切片检查,从第3天的胚胎植入前基因诊断,或滋养外胚层活检,从第5天的胚胎。15个单独的单细胞和30套3细胞放置在45个体的反应管中,总共45个反应,每个反应包含只从一个细胞系的细胞,但在不同的反应包含来自不同的细胞系的细胞。星u1洗涤液制备成细胞裂解加入星u1 ARCTURUS PICOPURE裂解液(Applied Biosystems)上,和孵化在56℃,20分钟,95℃为10分钟。

[0694] 单相/三相细胞的DNA进行扩增25个循环STA(95℃,10 分钟的初始聚合酶激活,然后25个循环的95℃30秒72℃10秒 65℃,1分钟60℃为8分65℃,3分钟和72℃30秒和最后一

次延长为2分钟,在72℃)采用50纳米的1200目标特定的正向和反向引物标记引物浓度。

[0695] 半嵌套PCR协议涉及三个平行第二扩增稀释第一STA的产品为20个周期的STA (95℃为10分钟的初始聚合酶激活,然后 15个循环的95℃30秒65℃1分钟,60℃为5分钟;65℃,5分钟和72℃30秒;最后延伸为2分钟,在72℃)使用反向标记为1000nm 的特异性引物浓度,浓度为每400个具体目标嵌套正向引物60 纳米。在三个平行的400重反应的共1200个目标,扩增在第一 STA这样扩增。

[0696] 的STA的产品的等分样本,然后通过标准的PCR扩增15个循环,1uM的标记特定的正向和条形码的反向引物,以产生条形码的测序库。等分不同的条形码库,每个库混合使用离心柱纯化。

[0697] 以这种方式,1200的引物中使用的单细胞反应的引物的设计目标SNP位点发现1,21和X染色体上的扩增产物进行测序使用一个ILLUMINA的GAIIIX测序仪。每个样本,约3.9万读取所产生的序,有500,000至800,000万映射到基因组(每个样本的74%至94%的读取)读取。

[0698] 有关母体和父亲从细胞系的基因组DNA样本进行分析,使用相同的半巢1200复杂的检测池具有类似的协议,用更少的周期和1200复杂的第二个STA和测序。

[0699] 测序数据的分析,用信息学的方法所披露者外,倍体的状态被称为在三个染色体的样本。

[0700] 图24示出了6个样本在三个染色体(1=铬12=铬213=铬倍) 的读出率(垂直轴)的归一化深度。的比率被设定为映射到该染色体,归一化,读取的数目等于读取映射到该染色体的数目除以平均三口井,每一个都包括三个46XY细胞。的三组数据点对应于的46XY反应,预期的比率为1:1。三组数据点对应于47XX21 细胞预计将有21号染色体和X染色体为2:1的比率为1:1的1 号染色体,1.5:1

[0701] 图25显示了等位基因比率3染色体作图,(1,21,X) 的三个反应。该反应在左下方显示三个46XY细胞的反应。左区域1 号染色体上的等位基因比率,在中间区域是21号染色体和右区域的等位基因比率对于46XY细胞X染色体等位基因比率,我们希望看到比为1的1号染色体,0.5和0,对应于AA,AB和 BB的SNP的基因型。对于的46XY细胞,染色体21,我们期望看到的比率为1,0.5和0,对应AA,AB和BB SNP基因型。X染色体为46XY细胞,我们希望看到比1和0,对应到A和B 的SNP基因型。该反应在右下示出了三个47XX21细胞反应。被分隔开的左下方图形中染色体等位基因比率。47XX21细胞,1 号染色体,我们期望看到的比率为1,0.5和0,对应AA,AB 和BB SNP基因型。47XX21细胞,21号染色体上,我们希望看到比1,0.67,0.33和0,对应AAA,AAB,ABB,BBB SNP 基因型。47XX21细胞,X染色体,我们期望看到的比率为1,0.5和0,对应AA,AB,BB SNP基因型。在右上方的情节作出的反应,包括从47XX21细胞系的基因组DNA1纳克。图26显示了如图25所示的曲线图相同,但只有一个细胞上进行反应。左边的图是包含一个47XX21细胞反应,右图所示为包含46XX 细胞反应。

[0702] 从图25和图26所示的图形,它是明显可见的是,有两簇点染色体,我们期望看到比1和0;三组染色体点,我们希望看到比 1,点0.5和0,四组染色体我们希望看到比1,0.67,0.33和0。父母的支持算法能够做出正确的呼吁所有三个染色体为所有45 个反应。

[0703] 本申请的所引用的所有专利,专利申请,出版的参考文献均通过引用的方式全文并入本发明。虽然本发明所公开的方法在其相关的具体实施方式进行了描述,但应该了解,

其还可以进行进一步的修改。此外,本申请意在涵盖本发明所公开的方法的任何变化,应用或调整,包括偏离本发明所公开的而属于在本技术领域属于公知或习惯做法,均落在所附的权利要求书的范围内。

序列表

<110> 吉恩安全网络公司

<120> 非侵入性产前倍性识别方法

<130> 118103.011702/PCT

<140> 没分配

<141> 2011-11-18

<150> 13/110,685

<151> 2011-05-18

<150> 61/395,850

<151> 2010-05-18

<150> 61/398,159

<151> 2010-06-21

[0001] <150> 61/462,972

<151> 2011-02-09

<150> 61/448,547

<151> 2011-03-02

<150> 61/571,248

<151> 2011-06-23

<150> 61/516,996

<151> 2011-04-12

<160> 12

<170> PatentIn version 3.5

<210> 1

<211> 42

<212> DNA

<213> 人工序列

<220>

<223> 合成结构

	<400> 1 aactcacata gcacacgacg ctcttccgat ctgcaagca ca	42
	<210> 2 <211> 39 <212> DNA <213> 人工序列	
	<220> <223> 合成结构	
	<400> 2 tcctctgtga cacgacgctc ttccgatctc cctgctctt	39
	<210> 3 <211> 40 <212> DNA <213> 人工序列	
[0002]	<220> <223> 合成结构	
	<400> 3 tcctctctct acacgacgct ctccgatctc cgggctgtca	40
	<210> 4 <211> 42 <212> DNA <213> 人工序列	
	<220> <223> 合成结构	
	<400> 4 tacatccttg agacacgaag ctcttccgat ctgctgtgca gt	42
	<210> 5 <211> 42 <212> DNA <213> 人工序列	

	<220>	
	<223> 合成结构	
	<400> 5	
	tttgettggag ctacacgacg ctcttccgat ctcgaggatt tc	42
	<210> 6	
	<211> 42	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 合成结构	
	<400> 6	
	gtcttatggt ggacacgacg ctcttccgat ctcaaagcca gt	42
	<210> 7	
	<211> 50	
[0003]	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 合成结构	
	<400> 7	
	aactcacata gctgatcggt acacgacgct ctccgatct tgcaagcaca	50
	<210> 8	
	<211> 47	
	<212> DNA	
	<213> 人工合成	
	<220>	
	<223> 合成结构	
	<400> 8	
	tcctctgtgt gatcggtaca cgacgtctt ccgatctccc tgetctt	47
	<210> 9	
	<211> 48	

	<212>	DNA	
	<213>	人工序列	
	<220>		
	<223>	合成结构	
	<400>	9	
		tcctctctct tgatecgtac acgacgcctt ccgatctcg ggctgtca	48
	<210>	10	
	<211>	50	
	<212>	DNA	
	<213>	人工序列	
	<220>		
	<223>	合成结构	
	<400>	10	
		tacatccttg agtgatecgt acacgacgct ctccgatct gctgtgcagt	50
[0004]	<210>	11	
	<211>	50	
	<212>	DNA	
	<213>	人工序列	
	<220>		
	<223>	合成结构	
	<400>	11	
		tttgcttgag ctgatecgt acacgacgct ctccgatct egggagtttc	50
	<210>	12	
	<211>	50	
	<212>	DNA	
	<213>	人工序列	
	<220>		
	<223>	合成结构	
	<400>	12	
		gtcttatggt ggtgatecgt acacgacgct ctccgatct caaagccagt	50

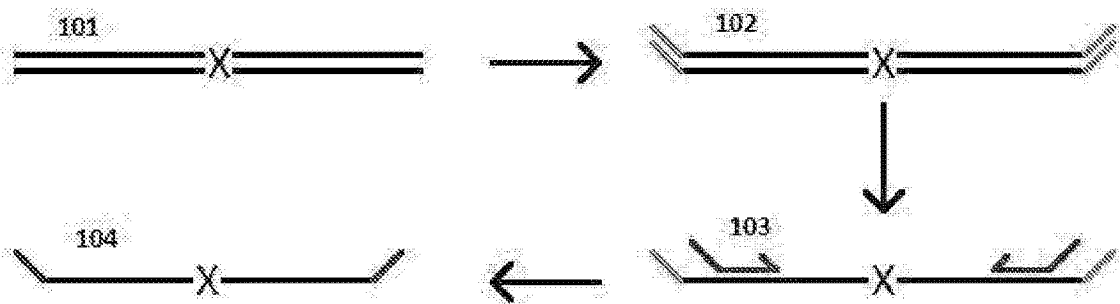


图1

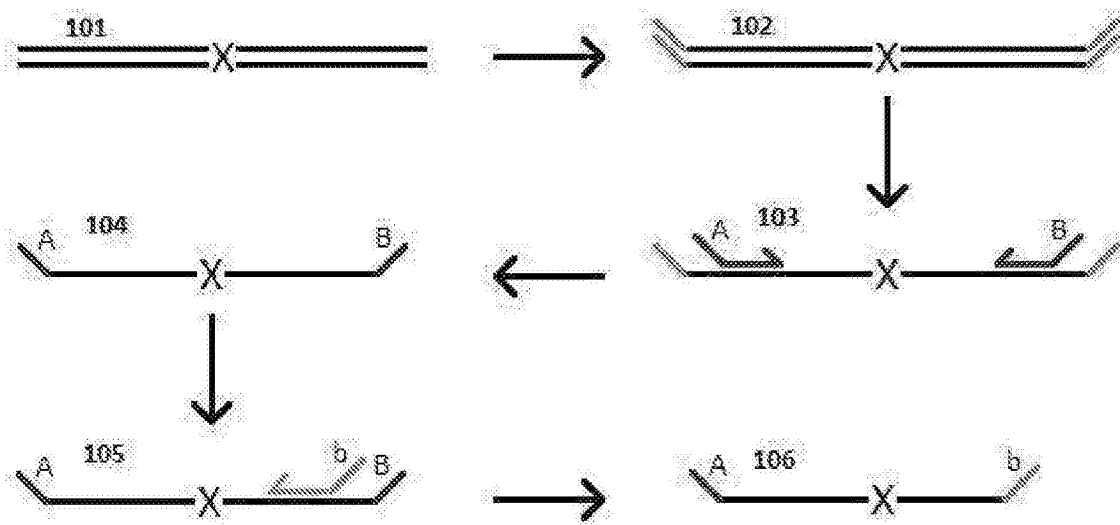


图2

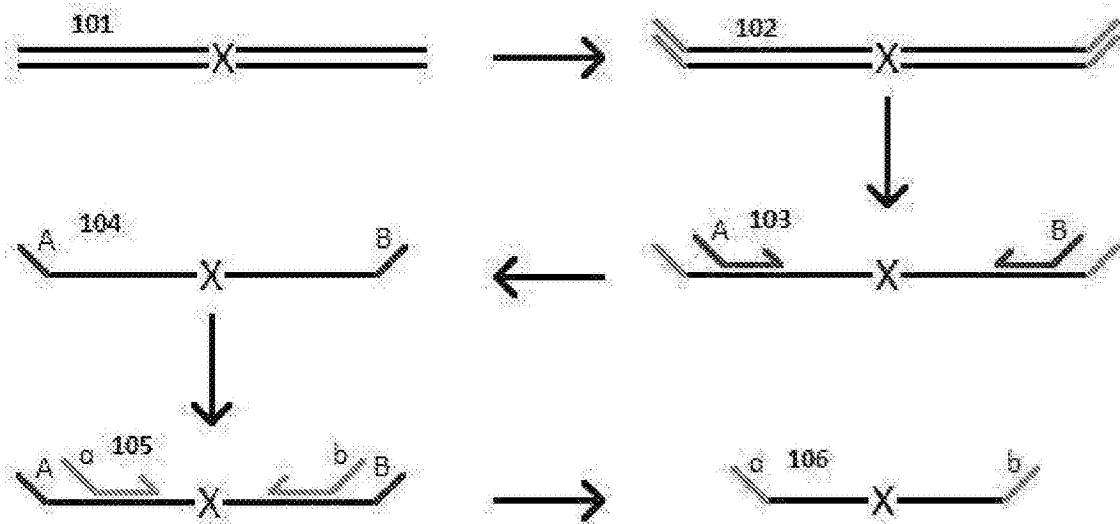


图3

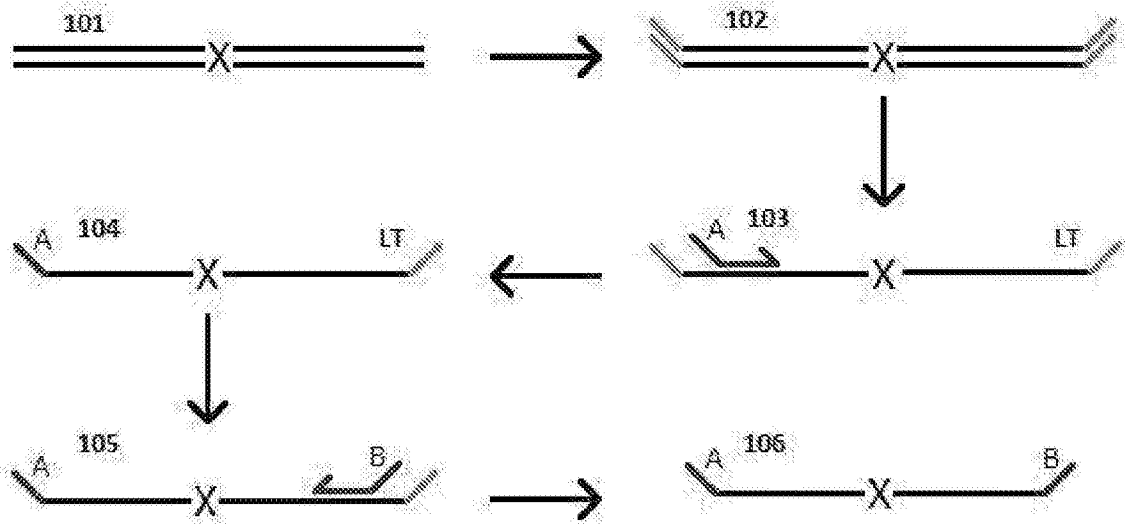


图4

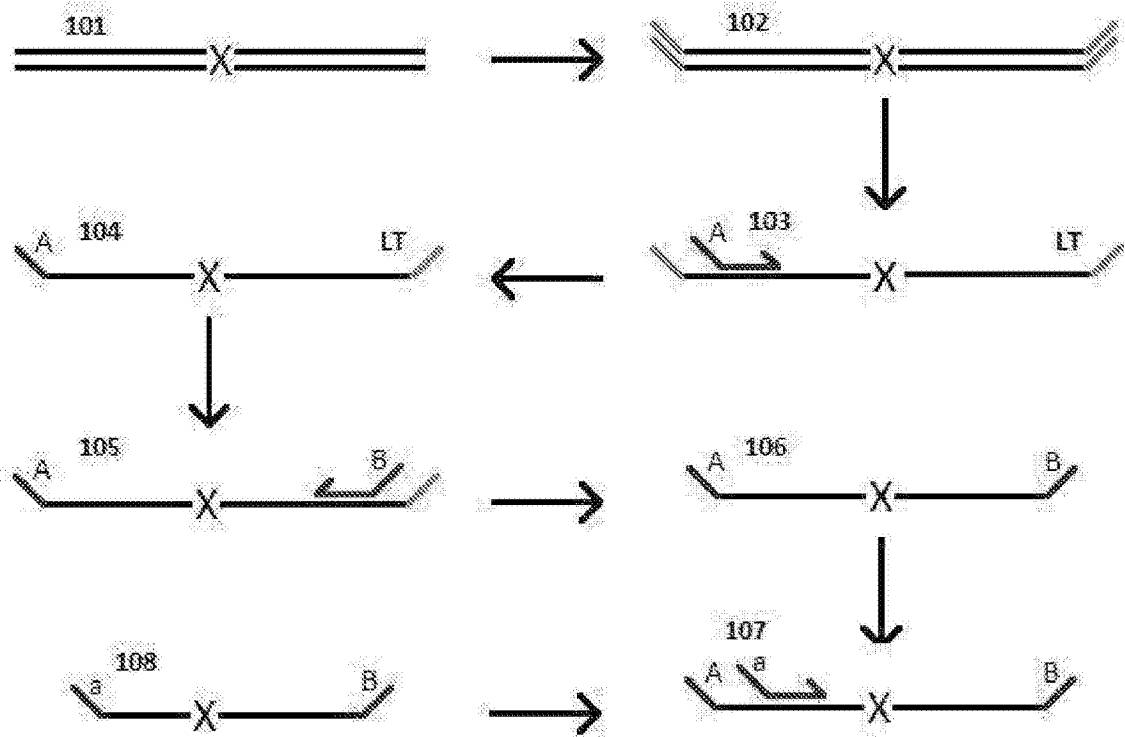


图5

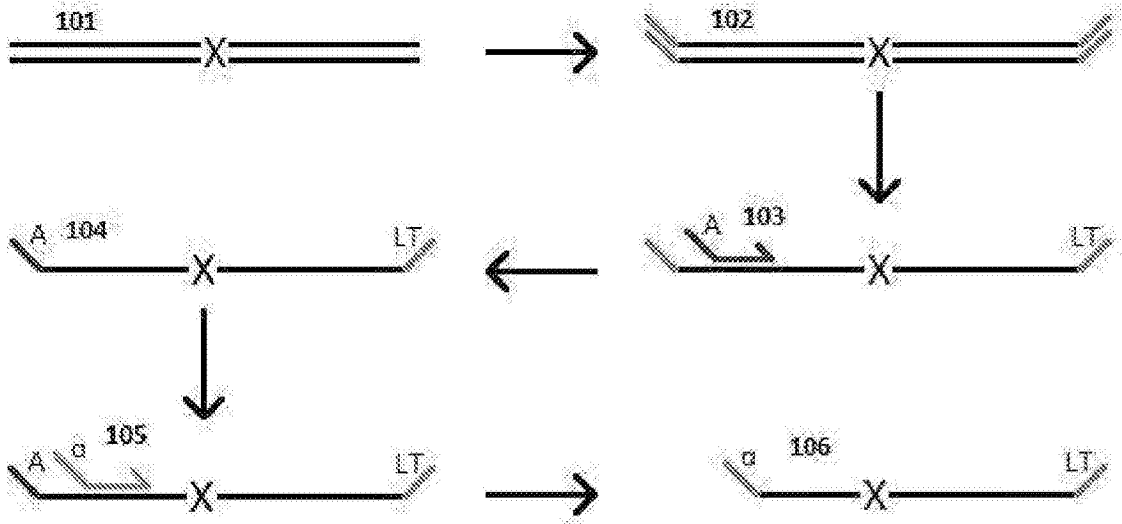


图6

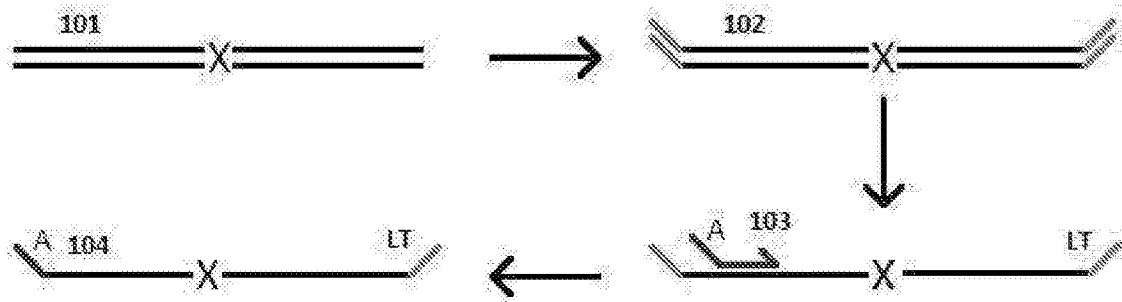


图7

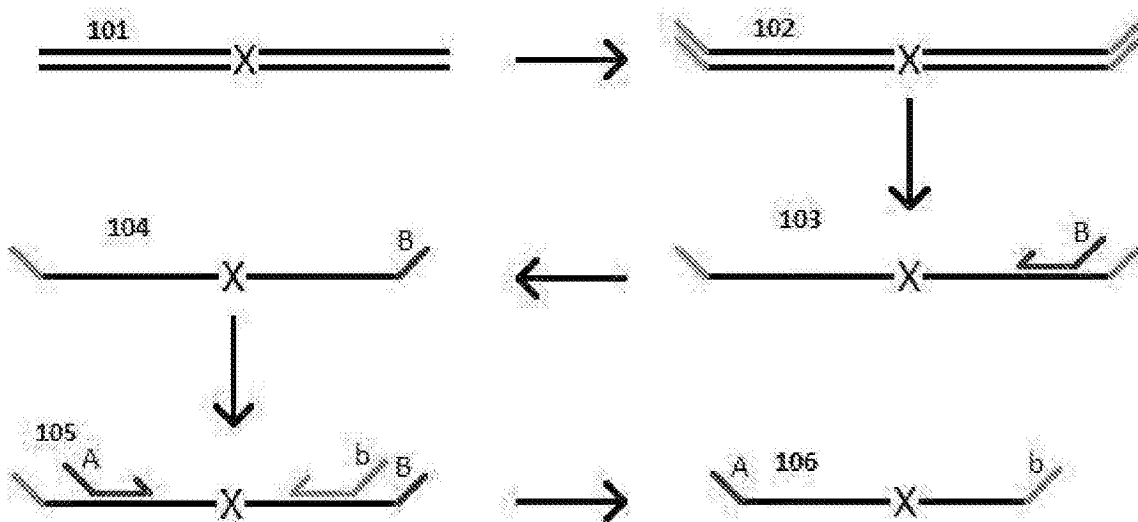


图8

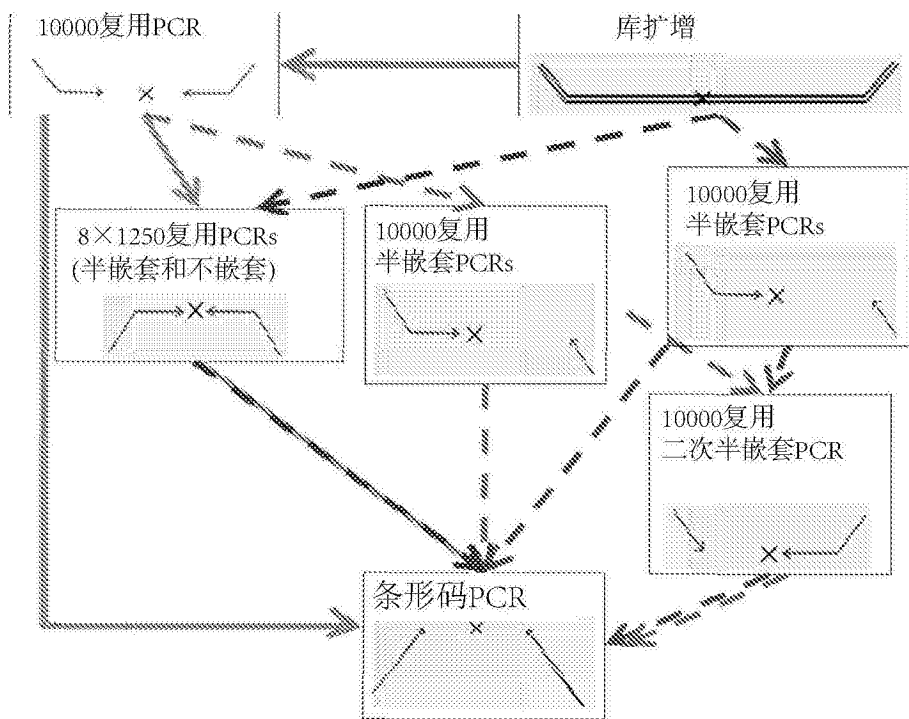


图9

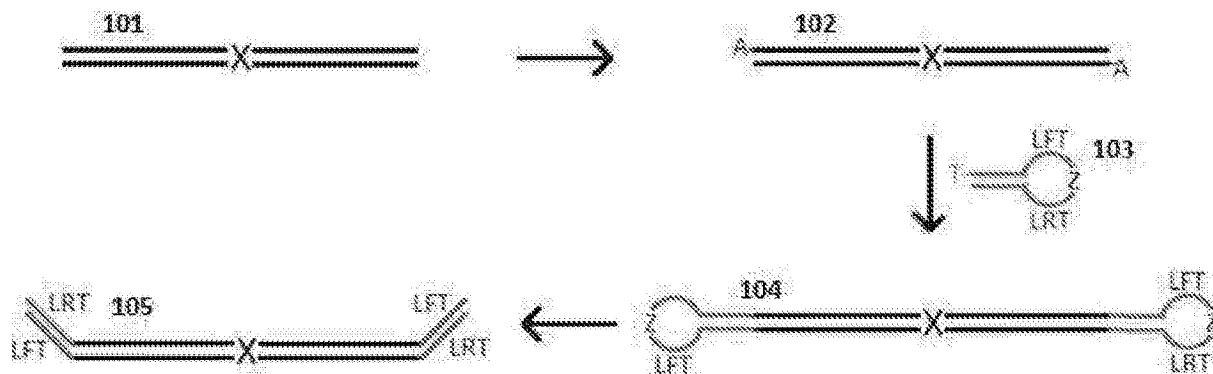


图10

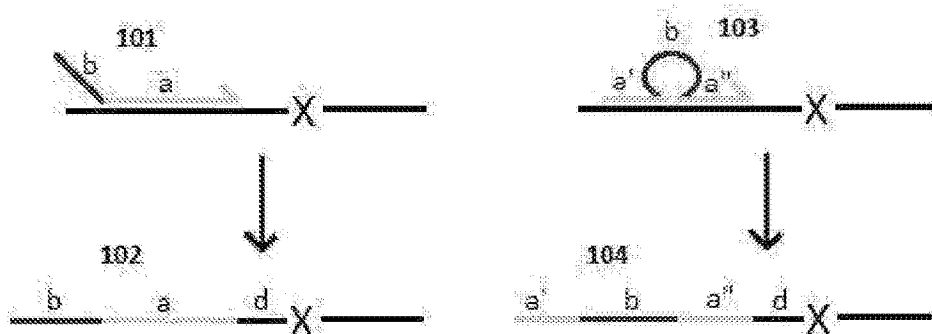


图11

测序适配子序列位于引物序列内部，目标特异性序列位于其两侧，在每个引物的3'端有10个碱基是目标特异性的。引物用实时PCR成功测试。

rs8130564	int-tag 1.10	AACTCACATAGCAGACGACGCTCTTCGATCTTGCAAGCACA	seqid1
rs2832093	int-tag 2.10	TCCTCGTGACACGACGCTCTTCGATCTCCCTGCTCTT	seqid2
rs12011281	int-tag 3.10	tcctctctctACACGACGCTCTTCGATCTGGGCTGTCA	seqid3
rs6719561	int-tag 4.10	TACATCCTTGAGACACGACGCTCTTCGATCTGCTGTGCAGT	seqid4
rs10187018	int-tag 5.10	tttgcttgagctACACGACGCTCTTCGATCTcgggagttc	seqid5
rs10460481	int-tag 6.10	gctttatggttgACACGACGCTCTTCGATCTcaagccagt	seqid6

测序适配子序列位于引物序列内部，目标特异性序列位于其两侧，内部标签形成发夹结构，这带来具有目标特异性的引物两端接近而阻碍到该“内部标签”的非特异性结合，在每个引物的3'端有10个碱基是目标特异性的。引物用实时PCR成功测试。

rs8130564	loop-int-tag 1.10	AACTCACATAGCtgatrggtACACGACGCTCTTCGATCTTGCAAGCACA	seqid7
rs2832093	loop-int-tag 2.10	TCCTCGTGtgatrggtACACGACGCTCTTCGATCTCCCTGCTCTT	seqid8
rs12011281	loop-int-tag 3.10	tcctctctctgaggggtACACGACGCTCTTCGATCTGGGCTGTCA	seqid9
rs6719561	loop-int-tag 4.10	TACATCCTTGAGtgaggggtACACGACGCTCTTCGATCTGCTGTGCAGT	seqid10
rs10187018	loop-int-tag 5.10	tttgcttgagctgctgggtACACGACGCTCTTCGATCTcgggagttc	seqid11
rs10460481	loop-int-tag 6.10	gctttatggttgctgggtACACGACGCTCTTCGATCTcaagccagt	seqid12

图12

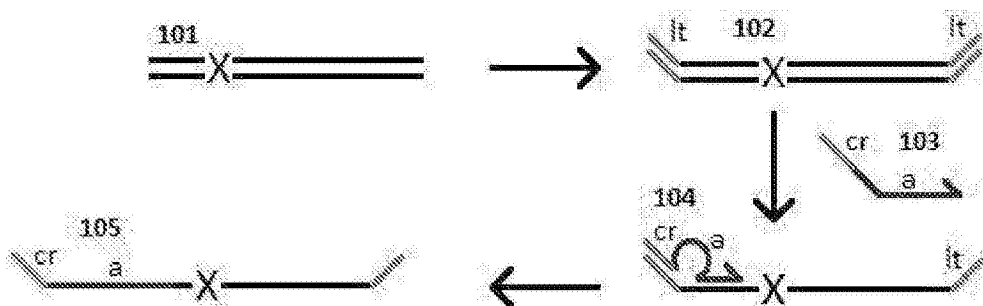


图13

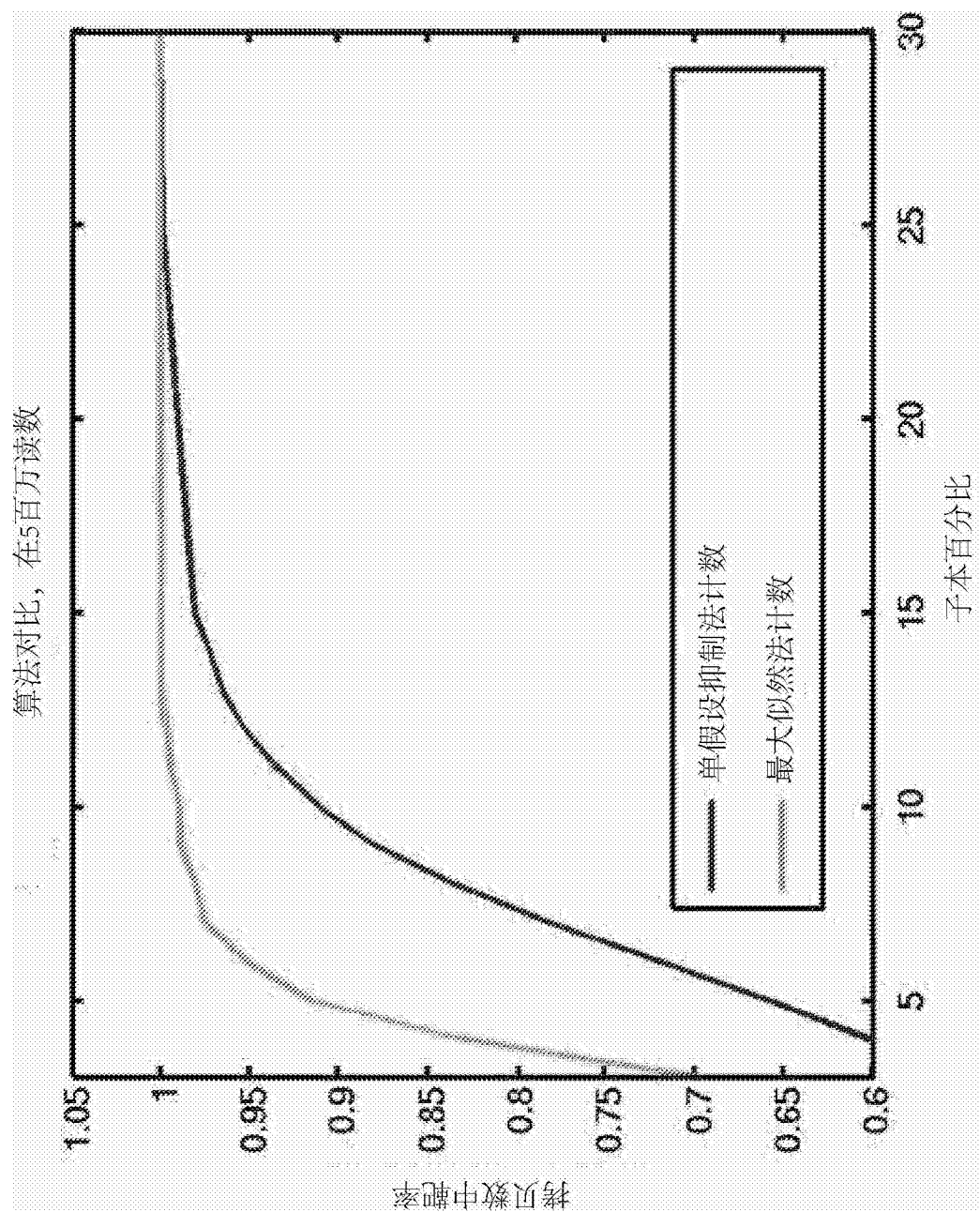


图14

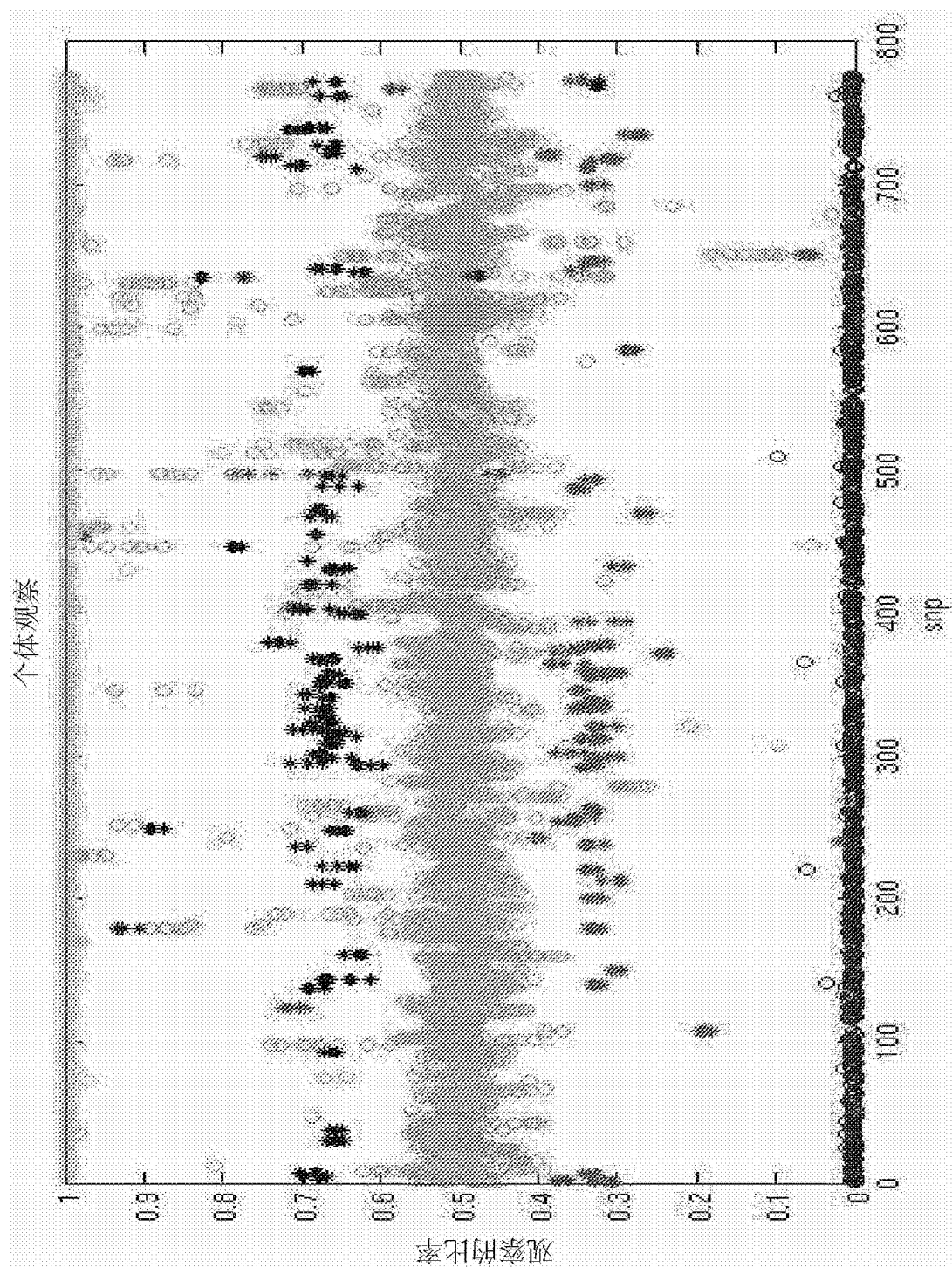


图15

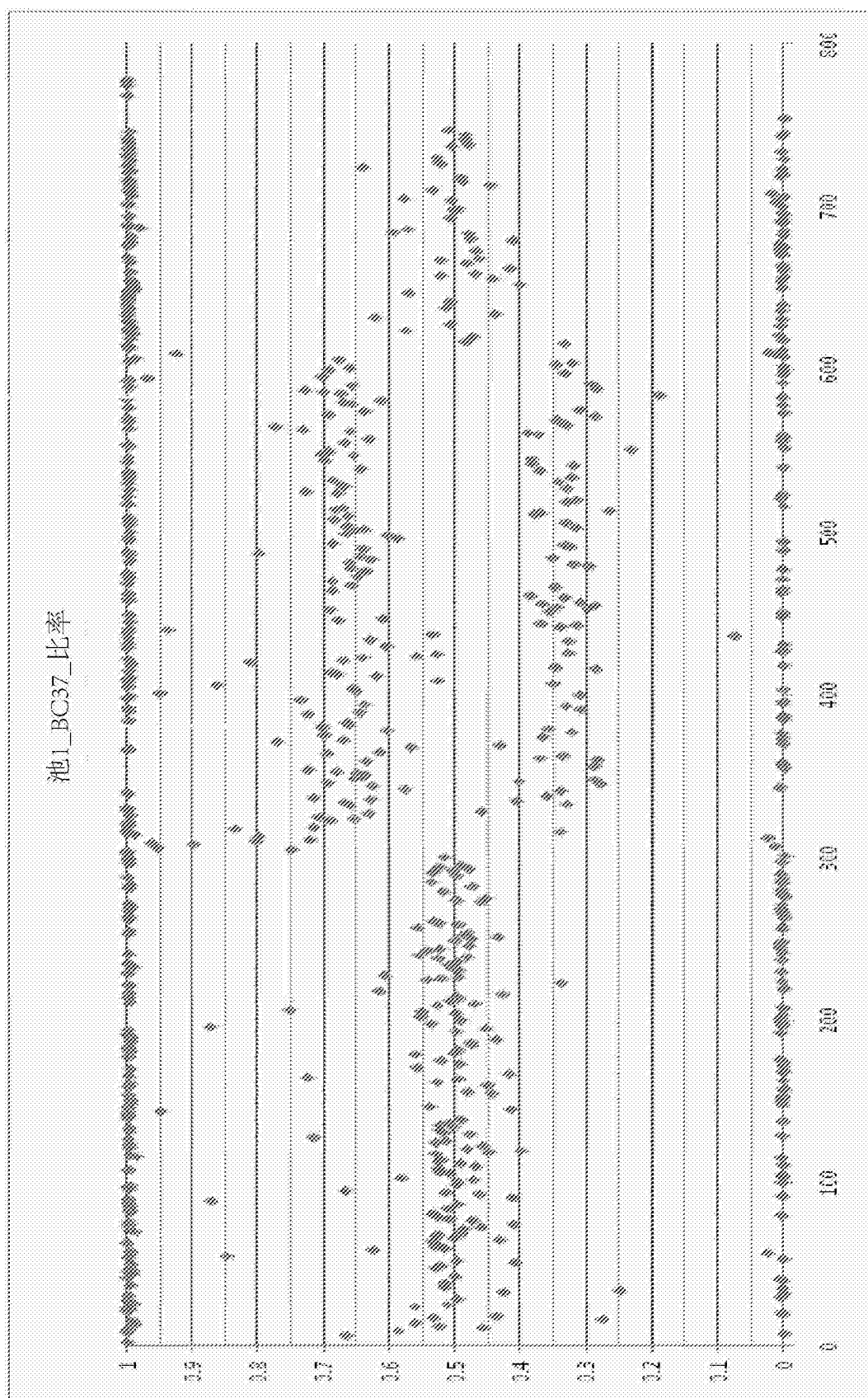


图16



图17

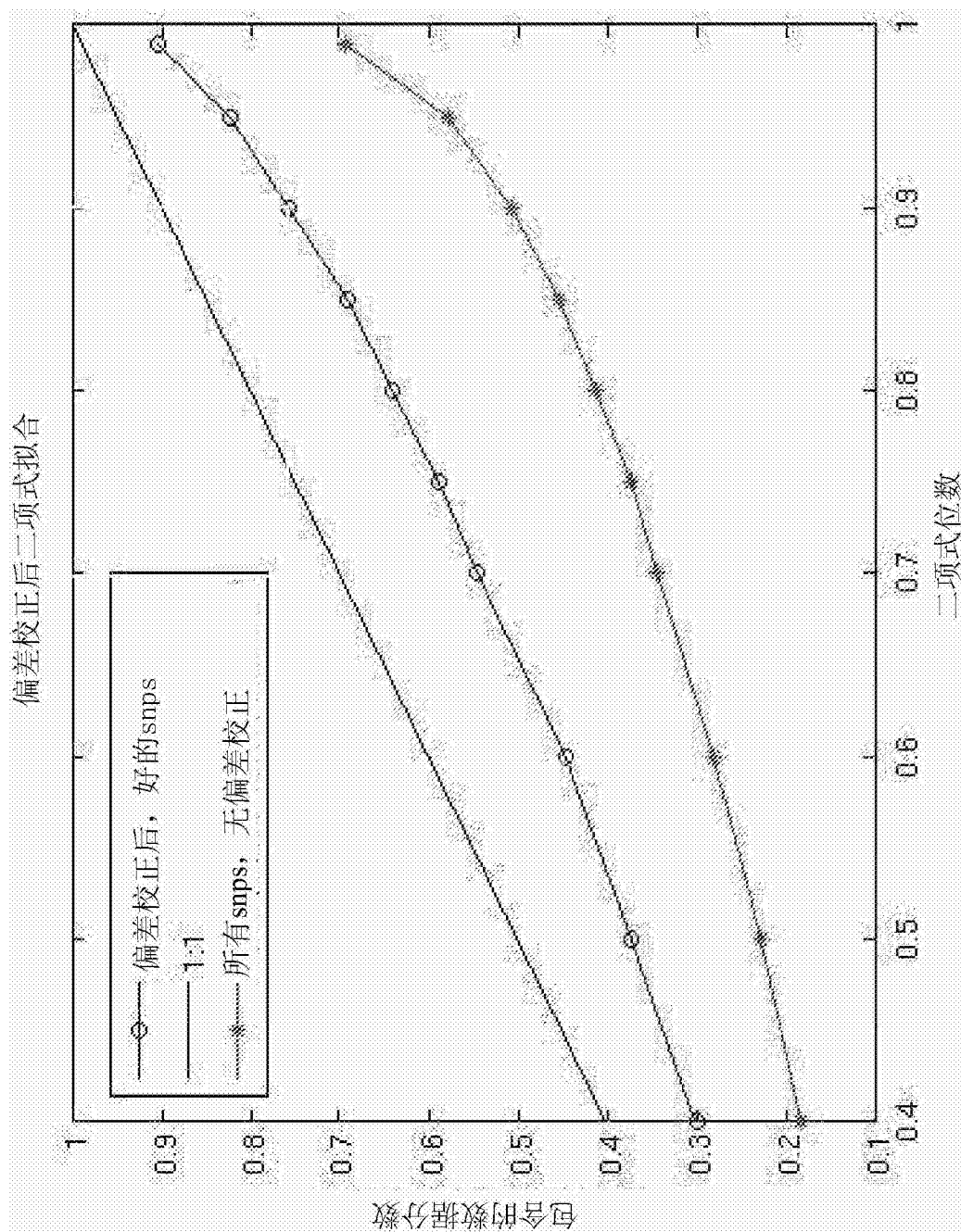


图18

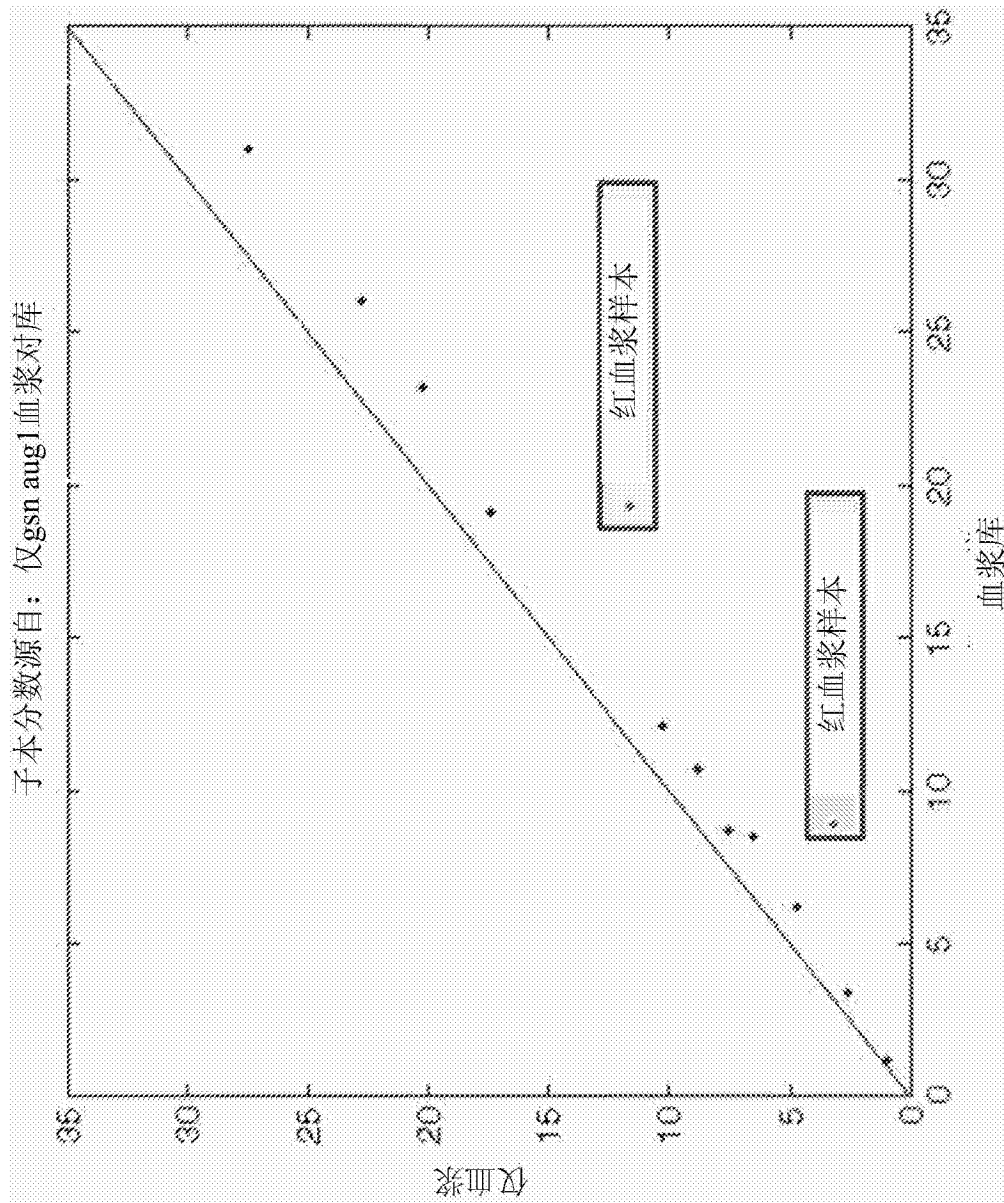


图19

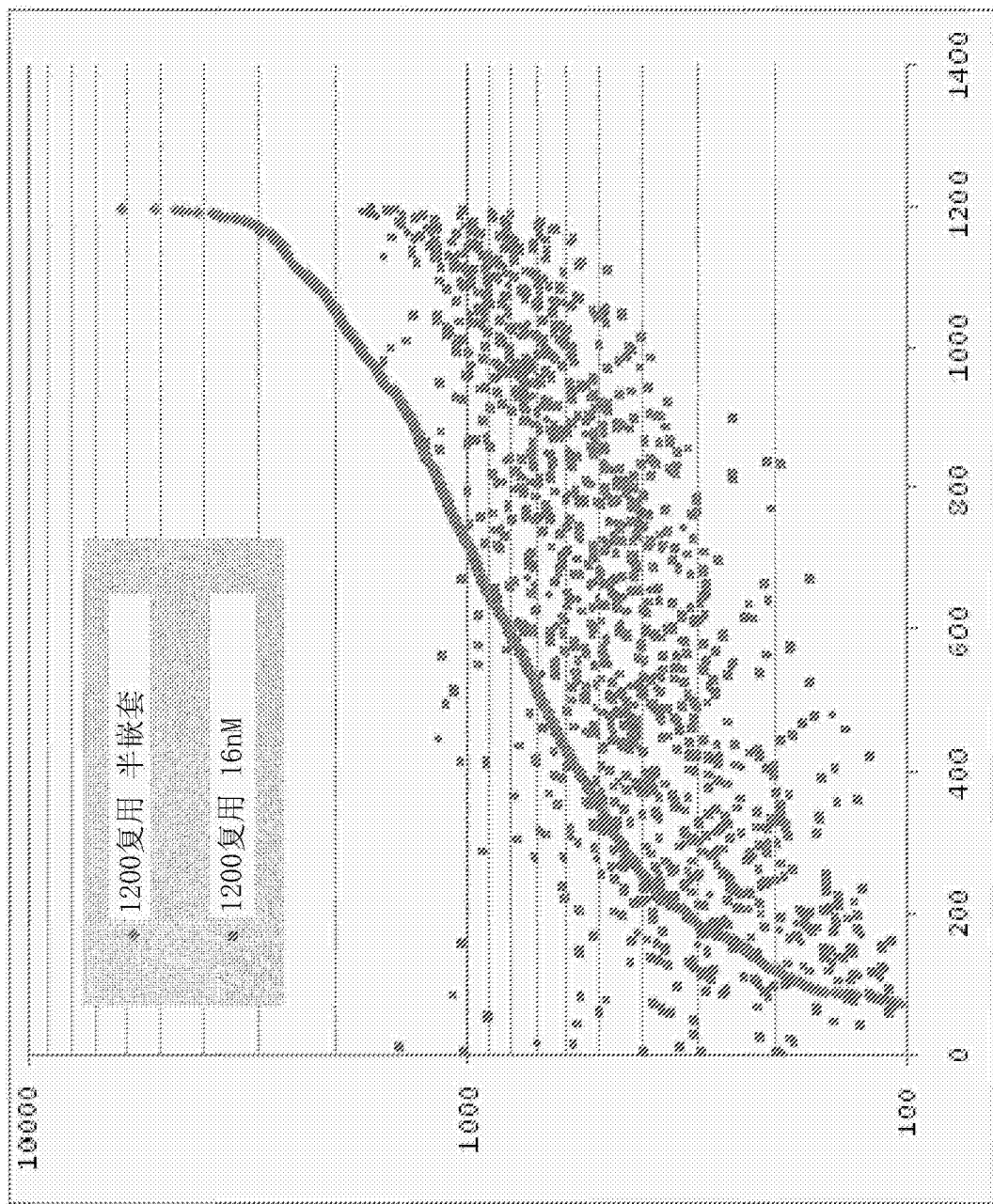


图20

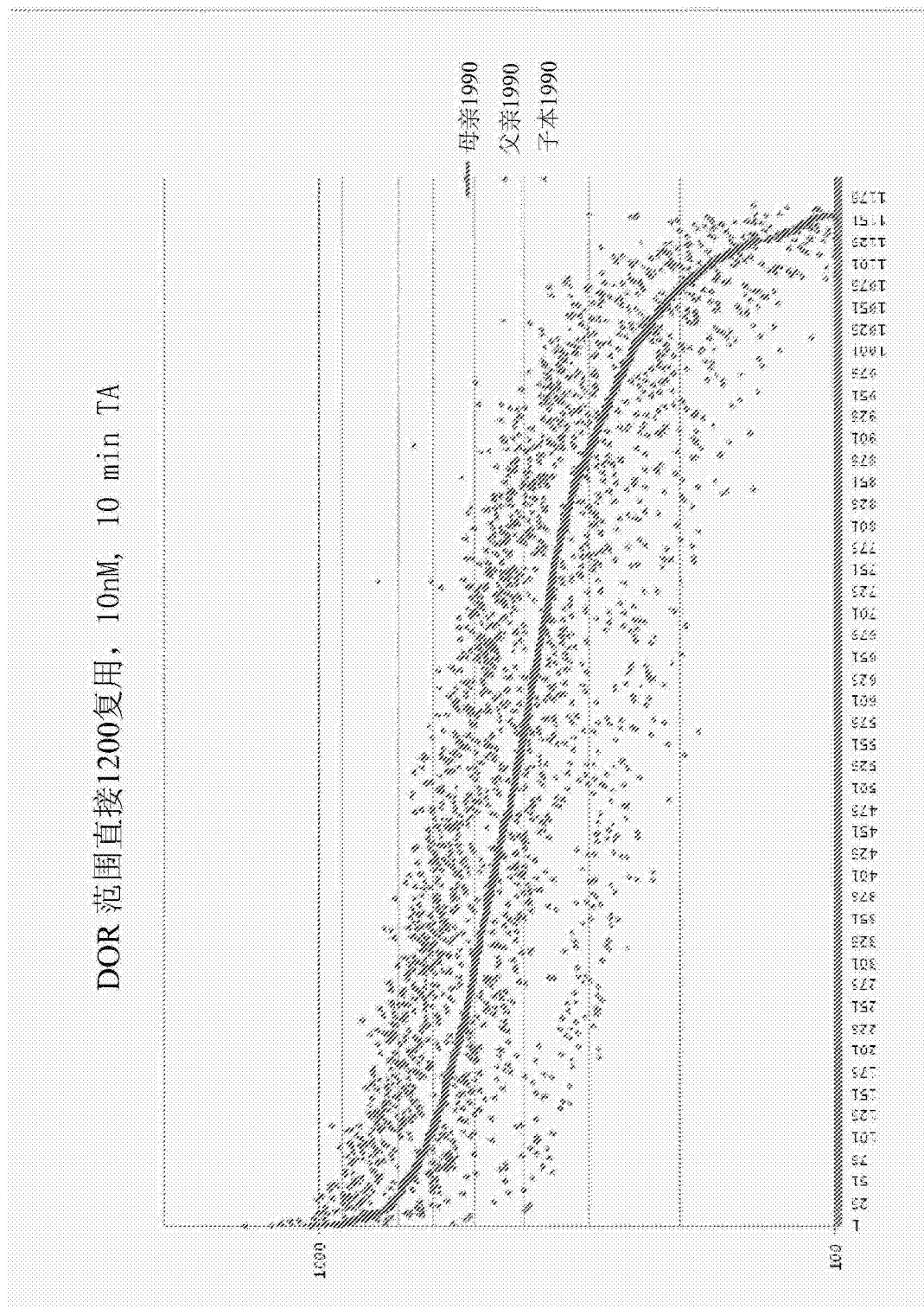


图21

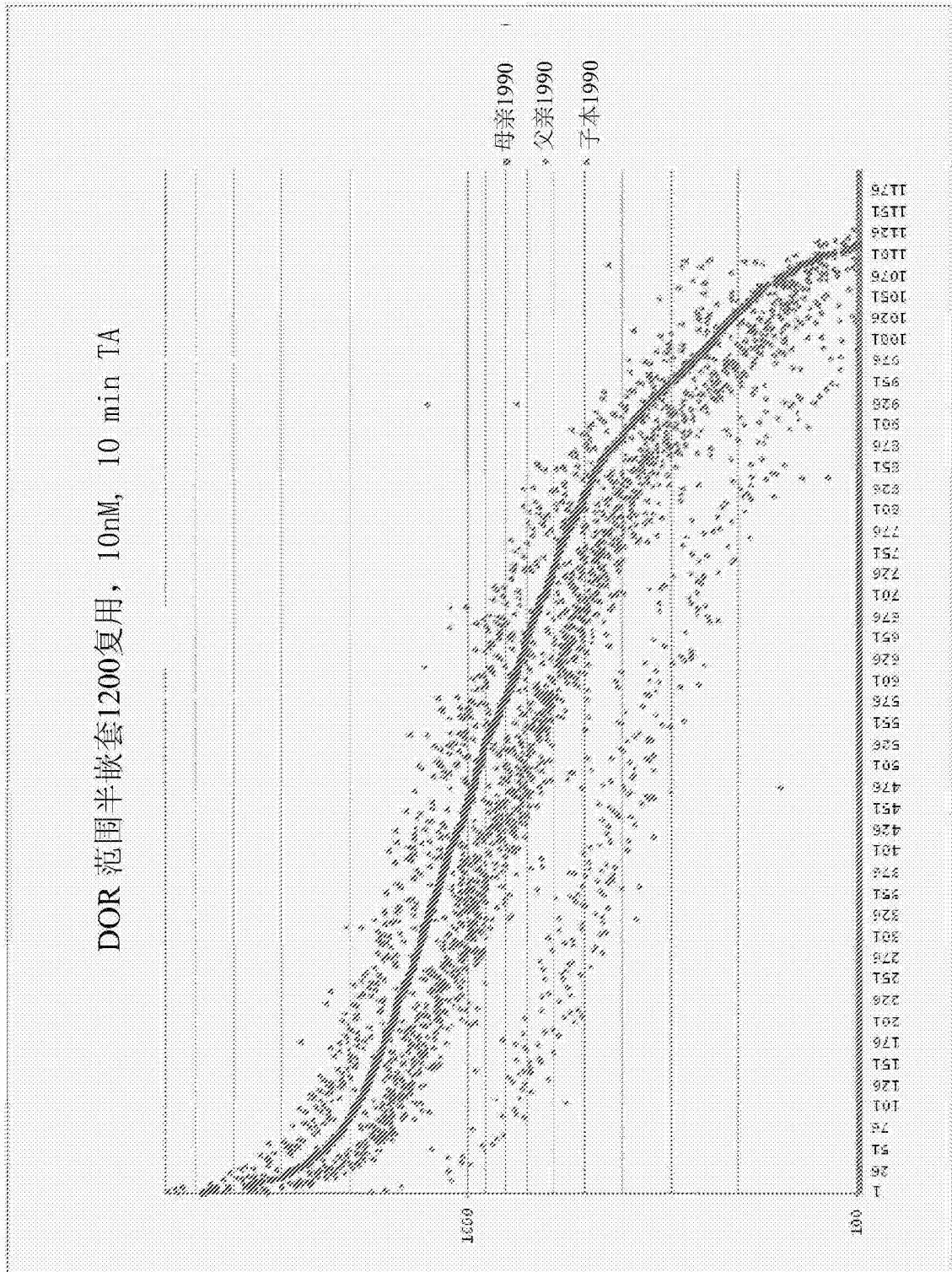


图22

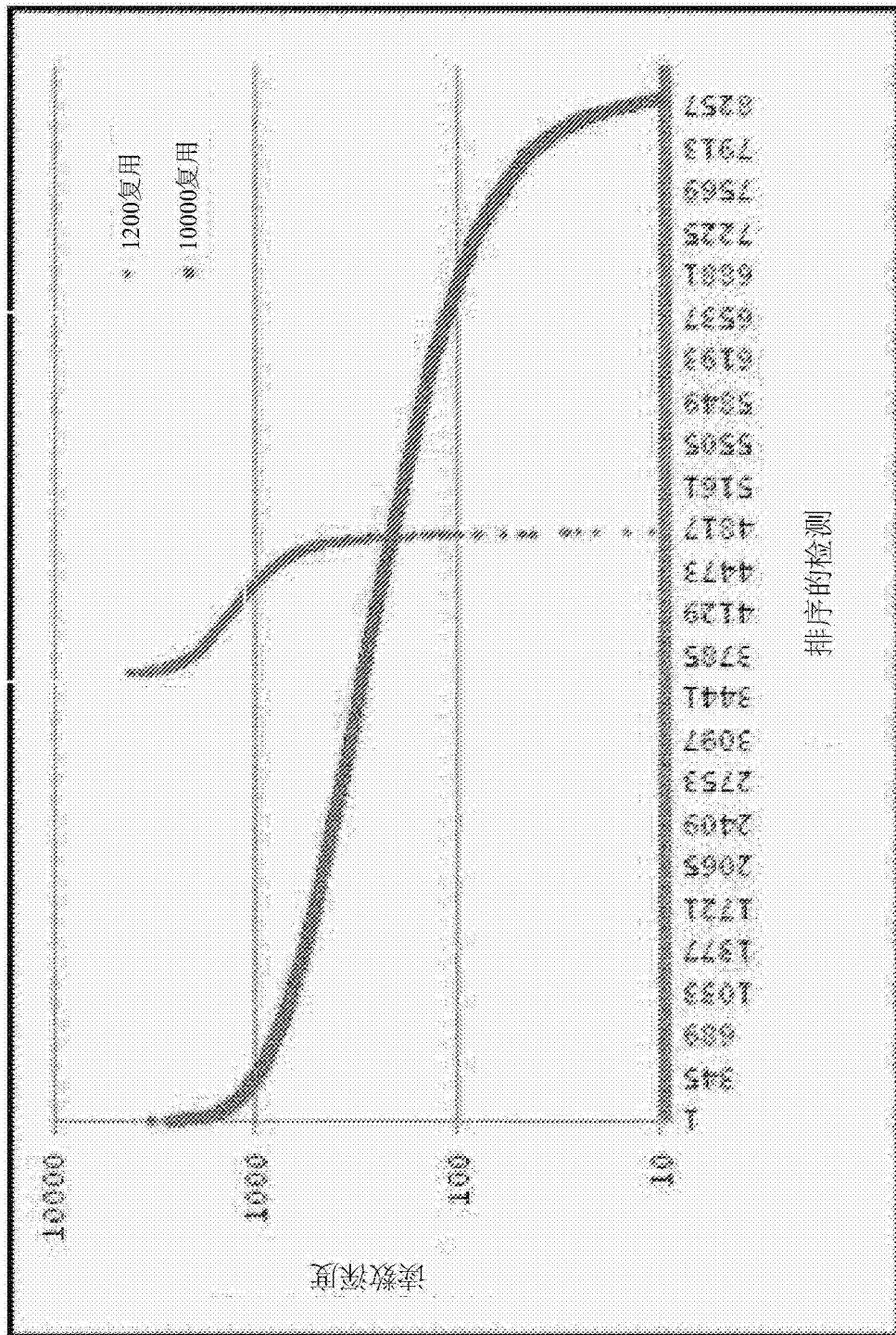


图23

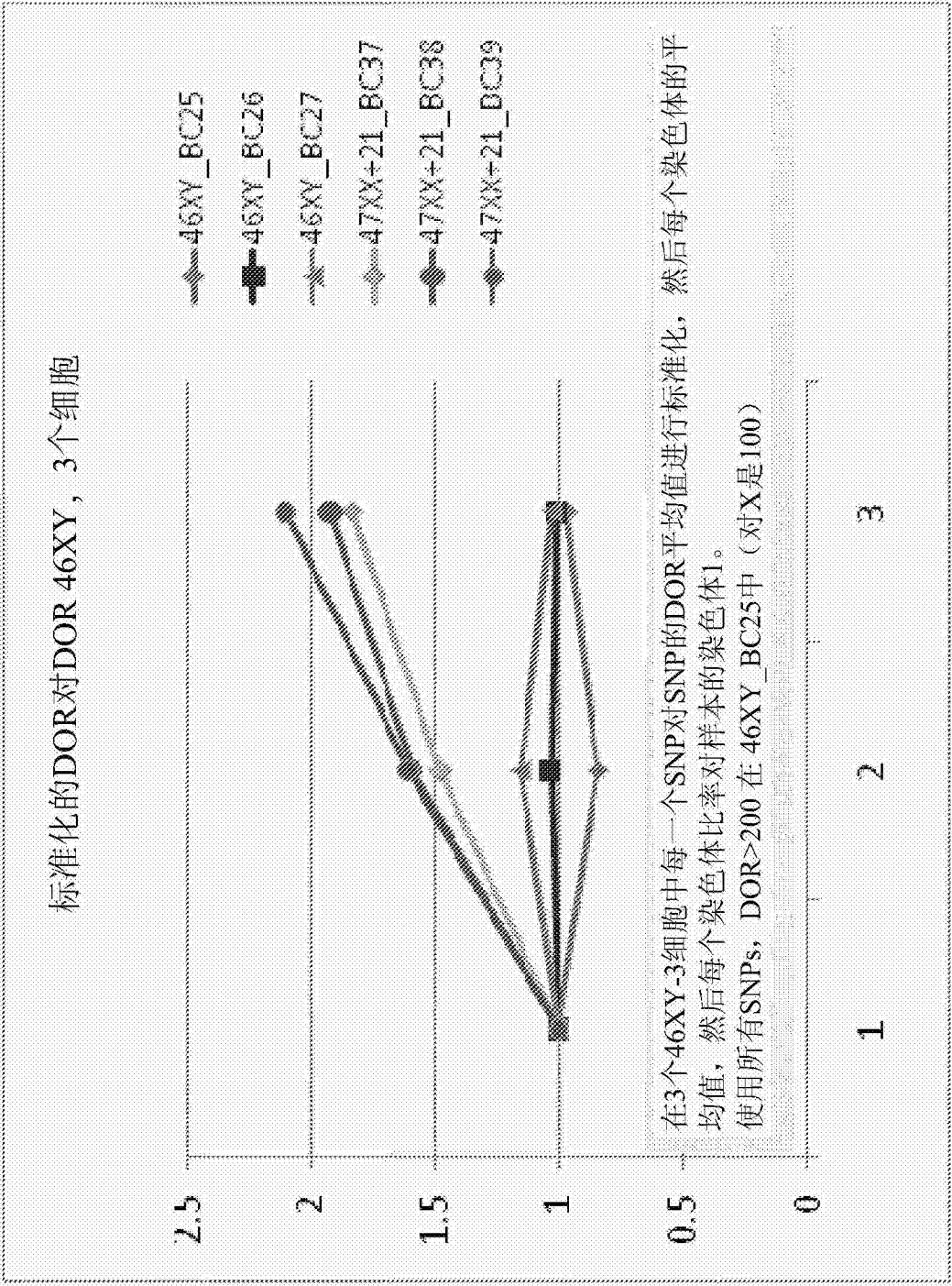


图24

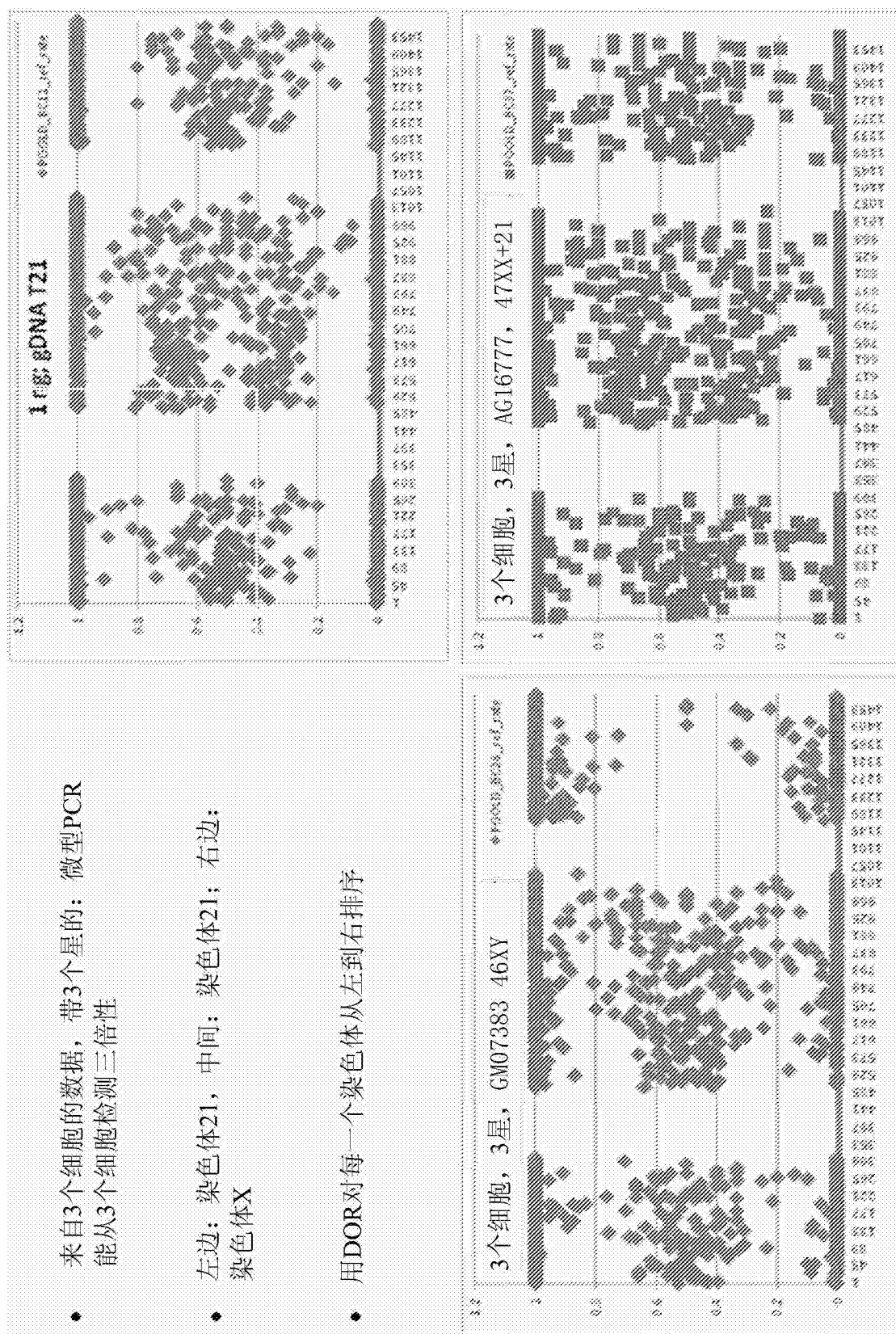


图25

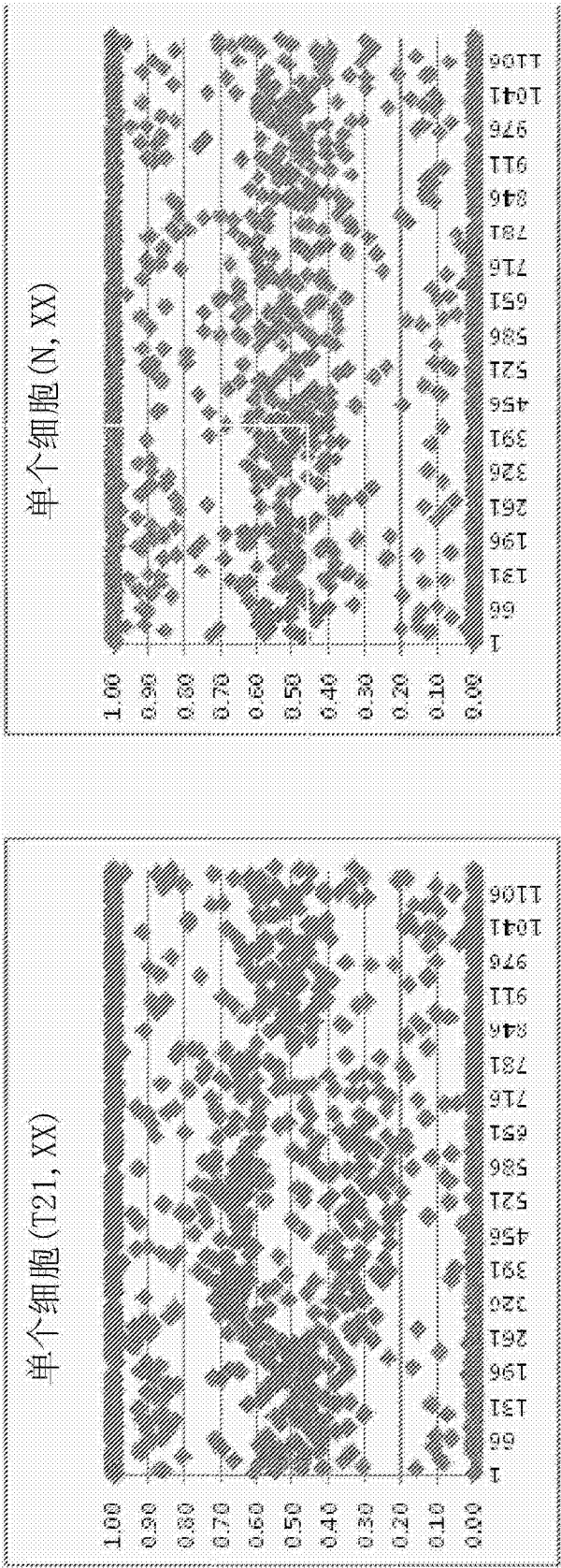


图26