(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103797129 B
(45)授权公告日 2016.08.17

(21)申请号 201280028976.9

(22)申请日 2012.04.12

(30)优先权数据
61/474,362 2011.04.12 US

(85)PCT国际申请进入国家阶段日
2013.12.12

(86)PCT国际申请的申请数据
PCT/US2012/033391 2012.04.12

(87)PCT国际申请的公布数据
WO2012/142334 EN 2012.10.18

(73)专利权人 维里纳塔健康公司
地址 美国加利福尼亚州

(72)发明人 里查德·P·拉瓦
布莱恩·K·利思 约翰·P·伯克

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 高瑜 郑霞

(51)Int.Cl.
C12Q 1/68(2006.01)

(56)对比文件

CN 100519761 C,2009.07.29,说明书第42页倒数第2段.

US 2010216153 A1,2010.08.26,说明书第7,109段,权利要求46-52.

LUN FIONA M F et al.Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma.《PNAS》.2008,第105卷(第50期),第19920-19925页.

DHALLAN RAVINDER et al.A non-invasive test for prenatal diagnosis based on fetal DNA present in maternal blood: a preliminary study.《The Lancet》.2007,第369卷第474-481页.

审查员 李谦

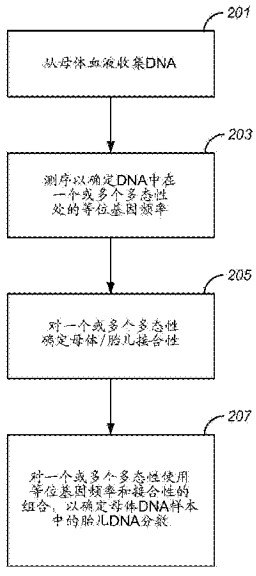
权利要求书3页 说明书44页 附图14页

(54)发明名称

使用多态计数来解析基因组分数

(57)摘要

公开一种从多态性例如小碱基变化或者插入-缺失可靠估计基因组分数(例如,胎儿分数)的方法。来自多基因组源的已测序数据用于确定对于一个或多个多态性等位基因计数。对于一个或多个多态性,接合性是分配的,而基因组分数从接合性和等位基因计数确定。某些实施例使用SNP作为相关多态性。所公开的方法能够应用作为目标在于已知多态性的故意的、预先设计的重测序研究的组成部分,或者能够用在从母体血浆产生的重叠序列中巧合发现的变化的回顾性分析中(或者其他任何存在多人DNA的混合物的设定)。



CN 103797129 B

1.一种估计在从怀孕个体的体液中获取的DNA中的胎儿DNA分数的方法，所述方法包括：

(a)接收体液样本；

(b)在提取体液中存在的母体基因组和胎儿基因组两者的DNA的条件下从样本中提取DNA；

(c)在生成包含一个或多个多态性的DNA序列读数的条件下利用核酸测序器测序已提取DNA；

(d)映射从测序体液中DNA得到的DNA序列读数到与一个或多个多态性对应的参考序列上的一个或多个多态性站，其中映射利用计算装置执行，计算装置编程成映射核酸序列到参考序列上的一个或多个多态性站；

(e)对一个或多个多态性确定所映射DNA序列读数的等位基因频率；

(f)估计已提取DNA中的胎儿DNA分数，其中估计包括利用在(e)中确定的等位基因频率对两种或更多下列接合性情况中的多态性执行一个或多个计算：

(i)怀孕个体是纯合的而胎儿是纯合的，

(ii)怀孕个体是纯合的而胎儿是杂合的，

(iii)怀孕个体是杂合的而胎儿是纯合的，以及

(iv)怀孕个体是杂合的而胎儿是杂合的，

其中(e)-(f)在一个或多个处理器上执行，而该一个或多个处理器在用于确定、分类和估计的程序指令下运行。

2.根据权利要求1所述的方法，其中(f)中的估计包括施加(e)中确定的等位基因频率到混合模型。

3.根据权利要求2所述的方法，施加等位基因频率到混合模型包括对所述一个或多个多态性的每一个，解一系列阶乘矩方程：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{d_i}$$

$$F_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)}{d_i(d_i-1)}$$

…

$$F_j = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)\cdots(a_i-j+1)}{d_i(d_i-1)(d_i-j+1)}$$

其中F_j是第j次阶乘矩，a_i是第i个多态性的少数等位基因频率，d_i是第i个多态性的覆盖，而n是多态性的数量。

4.根据权利要求3所述的方法，还包括基于接合性情况中阶乘矩与少数等位基因的可能性之间的下列关系计算胎儿DNA分数

$$F_1 \approx \sum_{i=1}^{m} a_i p_i^1$$

$$F_2 \approx \sum_{i=1}^{m} a_i p_i^2$$

…

$$F_j \approx \sum_{i=1}^{m} \alpha_i p_i^j$$

…

$$F_g \approx \sum_{i=1}^{m} \alpha_i p_i^g$$

其中，α_i是多个多态性在接合性情况i中的那部分，p_i是少数等位基因在接合性情况i中的二项式概率，m是接合性情况的数量，而g是已计算的阶乘矩的数量。

5.根据权利要求3的方法，还包括，在解一系列阶乘矩方程前，对识别为以下情况的多态性以计算性方式移除等位基因频率：

(A)在胎儿和怀孕个体中都是杂合的；

(B)在胎儿和怀孕个体中都是纯合的；或者

(C)在怀孕个体中是杂合的。

6.根据权利要求2所述的方法，其中所述混合模型对测序误差负责。

7.根据权利要求1或2所述的方法，还包括不考虑分类在组合(i)或者组合(iv)中的任意多态性。

8.根据权利要求1或2所述的方法，还包括过滤所述多个多态性，以不考虑具有大于或小于所定义阈值的少数等位基因频率的任意多态性。

9.根据权利要求1或2所述的方法，还包括施加阈值给在(e)中确定的等位基因频率。

10.根据权利要求1或2所述的方法，其中(d)中的映射包括识别多个二等位基因多态性序列。

11.根据权利要求1或2所述的方法，其中在(f)中估计DNA中的胎儿DNA分数包括将组合(iii)的数据变换成组合(ii)的数据。

12.根据权利要求11所述的方法，其中将组合(iii)的数据变换成组合(ii)的数据包括：

(a)变换数据(D,A)到(D1,A1)，其中：

A1＝0.5D-A

D1＝D

其中D是多态性的覆盖，而A是所述多态性的少数等位基因计数；

(b)三角变换；或者

(c)使用旋转矩阵。

13.根据权利要求11所述的方法，还包括施加回归计数给组合(ii)的数据和已变换的组合(iii)的数据，其中胎儿DNA分数估计为线性回归模型的回归线的斜率的两倍。

14.根据权利要求1或2所述的方法，其中：

(a)从怀孕个体的体液中获取的DNA是从怀孕个体的血浆中获取的游离DNA；或者

(b)测序在没有选择性地放大一个或多个多态性的情况下执行。

15.一种用于估计在从怀孕个体的体液中获取的DNA中的胎儿DNA分数的装置，所述装置包括：

(a)测序器，其配置成

(i)接收从包括母体基因组和胎儿基因组的DNA的体液样本提取的DNA,并且

(ii)在生成DNA序列读数条件下测序已提取的DNA;和(b)计算装置,其配置成指示一个或多个处理器以执行前述任一权利要求中的方法。

使用多态计数来解析基因组分数

[0001] 相关申请的交叉引用

[0002] 本申请要求2011年4月12日提交的美国临时申请序号61/474362的优先权，通过引用将其内容完整地结合到本文中用于所有目的。

背景技术

[0003] 母体血液中的自由浮动胎儿DNA(有时称作“游离(cell free)DNA”或“cfDNA”)发现允许从血液样本来检测染色体异常、异倍性和畸变的可能性。母体血浆中的胎儿DNA的分数(fraction)丰度不是恒定的，而且随多种因素-包括样本处理和孕龄-而变化。

[0004] 在使用DNA测序来识别染色体畸变或遗传缺陷时，重要的是了解DNA总群体中的胎儿DNA的相对丰度。例如，当胎儿分数已知时，能够通过置换方法或者线性组合的积分或经由从α到无穷大的非中心F分布的卷积来计算统计能力(识别反常情况的概率，或者灵敏度)，其中在零假设下的得分群体的重要性(错误地称作反常的最大似然)的α临界点没有畸变。

[0005] 用于检测胎儿分数的现有方法的缺点在于，它们依靠性染色体(其只能用于可靠地测量男性胚胎DNA的相对丰度)的丰度的量度或者已知为在怀孕与胚胎组织之间差异表达的基因的mRNA序列(其因孕龄或其它因素而须经表达的可变性)。

[0006] 胎儿分数的估计因若干讨厌因素而较难，包括：双亲种族差异群体遗传参数和测序误差。因此，需要在这些及其它普遍发生的混合因素存在的情况下具有健壮的方法。

发明内容

[0007] 某些所公开实施例涉及通过对母体血液样本进行测序来可靠地测量胎儿自由浮动DNA的相对丰度的计算方法。

[0008] 在具体实施例中，本发明提供从多态性(例如小碱基变化或者插入-缺失，其相对双亲种族、胚胎性别、孕龄和其它环境因素是健壮的)来可靠地估计胎儿分数的方法。本文所公开的许多示例采用SNP作为相干多态性。本发明能够作为针对已知多态性的特意预先设计的再测序研究的组成部分来应用，或者能够用于通过从母体血浆(或者其中存在来自数人的DNA的混合的任何其它设定)所生成的重叠序列的一致性所发现的变化的回顾分析中。

[0009] 本文档提供用于估计母体血液样本中的胎儿DNA的分数丰度的技术。某些所公开技术使用通过偶然发现或者在设计用于便于估计胎儿分数的预先知道SNP的面板中发现的SNP的所观测等位基因频率。

[0010] 虽然公开的许多部分涉及估计样本中的胎儿核酸的分数，但是本发明并不局限于此。本文所述的技术和装置在许多情况下能够用于从两个基因组的混合中的一个基因组-其可以或者可以不是作为父子基因组来相关的-来估计核酸的分数。

[0011] 本公开的某些方面涉及估计从怀孕个体的体液所得到的DNA中的胎儿DNA的分数。这类方法的特征可在于下列操作：(a)接收体液的样本；(b)在提取体液中存在的母体基因

组和胎儿基因组的DNA的条件下从样本中提取DNA;(c)在产生包含一个或多个多态性的DNA段序列的条件下采用核酸测序器对所提取DNA进行测序;(d)将从对液体中的DNA进行测序所得出的DNA段序列映射到参考序列上的一个或多个所指定多态性;(e)对于所指定多态性的至少一个来确定映射DNA段序列的等位基因频率;(f)基于怀孕个体的接合性和胎儿的接合性的组合来分类至少一个所指定多态性;以及(g)使用在(e)所确定的等位基因频率和来自(f)的接合性的组合来估计从怀孕个体所得到的DNA中的胎儿DNA的分数。

[0012] 映射可使用计算装置来执行,其编程为将核酸序列映射到一个或多个所指定多态性。一般来说,操作(d)-(g)的任一个可在程序指令下在一个或多个处理器上执行。

[0013] 在某些实施例中,从怀孕个体的体液所得到的DNA是从怀孕个体的血浆所得到的游离DNA。通常在没有选择性地放大一个或多个所指定多态性的任一个的情况下进行测序。

[0014] 在某些实施例中,映射从携带胎儿的个体的血液所得到的DNA段包括以计算方式将段映射到多态性的数据库。在某些实施例中,(f)中的分类将至少一个所指定多态性分为下列组合之一:(i)怀孕个体是纯合(homozygous)的,并且胎儿是纯合的,(ii)怀孕个体是纯合的,而胎儿是杂合(heterozygous)的,(iii)怀孕个体是杂合的,而胎儿是纯合的,以及(iv)怀孕个体是杂合的,并且胎儿是杂合的。

[0015] 可采用各种过滤操作。这些包括例如不予考虑组合(i)或组合(iv)中分类的任何多态性。在另一个示例中,方法还包括过滤至少一个所指定多态性,从而不予考虑具有比所定义阈值要大的少数(minor)等位基因频率的任何多态性。在又一个示例中,方法包括如下操作:过滤至少一个所指定多态性,从而不予考虑具有比所定义阈值要小的少数等位基因频率的任何多态性。

[0016] 分类操作可按照各种方式来实现。例如,它可涉及将阈值应用于(e)中确定的等位基因频率。在另一个示例中,分类操作涉及把来自(e)、对于多个多态性所得到的等位基因频率数据应用于混合模型。在一个实现中,混合模型采用阶乘矩。

[0017] 如本文所述所确定的胎儿分数可用于各种应用。在一些示例中,本文所述的方法包括如下操作:在一个或多个处理器上运行程序指令,以便将(g)中所确定的DNA的胎儿的分数自动记录在用于怀孕个体的患者病历中,其存储在计算机可读介质上。患者病历可由实验室、医生办公室、医院、保健组织、保险公司或者个人病历网站来保持。在另一个应用中,胎儿DNA的分数的估计用于规定、发起和/或改变受检者(human subject)——从其获取母体测试样本——的治疗。在另一个应用中,胎儿DNA的分数的估计用于指示和/或执行一个或多个附加测试。

[0018] 本公开的另一方面涉及用于估计从怀孕个体的体液所得到的DNA中的胎儿DNA的分数的装置。这种装置的特征可在于下列特征:(a)测序器,配置成(i)接收从包含母体基因组和胎儿基因组的体液的样本所提取的DNA,并且(ii)在产生包含一个或多个所指定多态性的DNA段序列的条件下对所提取DNA进行测序;以及(b)计算装置,配置成(例如编程为)指示一个或多个处理器执行各种操作,例如随本文所述方法操作的两个或更多所述的那些操作。在一些实施例中,计算装置配置成:(i)将核酸序列映射到参考序列上的一个或多个所指定多态性,(ii)对于所指定多态性的至少一个来确定映射DNA段序列的等位基因频率,(iii)基于怀孕个体的接合性和胎儿的接合性的组合来分类至少一个所指定多态性,以及(iv)使用等位基因频率和接合性的组合来估计从怀孕个体所得到的DNA中的胎儿DNA的分

数。

[0019] 在某些实施例中，该装置还包括用于在提取母体基因组和胎儿基因组的DNA的条件下从样本中提取DNA的工具。在一些实施方式中，该装置包括配置成提取从怀孕个体的血浆所得到的游离DNA以供测序器中测序的模块。

[0020] 在一些示例中，该装置包括多态性的数据库。该计算装置还可配置成指示一个或多个处理器通过在计算上将段映射到多态性的数据库，来映射从怀有胎儿的个体的血液所得到的DNA段。数据库中的序列是参考序列的示例。下面提供参考序列的其它示例。

[0021] 在某些实施例中，该计算装置还配置成指示一个或多个处理器将至少一个所指定多态性分类为下列组合之一：(i)怀孕个体是纯合的，并且胎儿是纯合的，(ii)怀孕个体是纯合的，而胎儿是杂合的，(iii)怀孕个体是杂合的，而胎儿是纯合的，以及(iv)怀孕个体是杂合的，并且胎儿是杂合的。在一些实施例中，计算装置还配置成指示一个或多个处理器不予考虑组合(i)或组合(iv)中所分类的任何多态性。

[0022] 在某些实施例中，该计算装置还配置成指示一个或多个处理器不予考虑具有比所定义阈值要大的少数等位基因频率的任何多态性。在一些实施例中，该计算装置还配置成指示一个或多个处理器过滤一个或多个所指定多态性，从而不予考虑具有比所定义阈值要小的少数等位基因频率的任何多态性。在某些实施例中，该计算装置还配置成指示一个或多个处理器通过将阈值应用于等位基因频率，来分类至少一个所指定多态性。

[0023] 在某些实施例中，该计算装置还配置成指示一个或多个处理器通过将对于多个多态性所得到的等位基因频率数据应用于混合模型，来分类至少一个所指定多态性。混合模型可采用阶乘矩。

[0024] 在某些实施例中，该计算装置还配置成指示一个或多个处理器将DNA的胎儿分数记录在怀孕个体的患者病因中，其存储在计算机可读介质上。患者病历可由实验室、医生办公室、医院、保健组织、保险公司或者个人病历网站来保持。

[0025] 本公开的另一方面涉及按照下列操作来估计从怀孕个体的体液所得到的DNA中的胎儿DNA的分数的方法：(a)将从怀孕个体的体液所得到的DNA段映射到多个多态性序列，其中在识别多个多态性序列的条件下对DNA测序；(b)确定多个多态性序列的每个的映射核酸的等位基因频率；以及(c)将等位基因频率应用于混合模型，以便得到从携带胎儿的个体的血液所得到的DNA中的胎儿DNA分数的估计。操作(a)-(c)的任何一个或多个可在程序指令下运行的一个或多个处理器上执行。在某些实施例中，操作(c)涉及在一个或多个处理器上运行指令，以用于求解多个多态性序列的每个的等位基因数据的阶乘矩的一系列等式。在一些实施例中，混合模型考虑测序误差。

[0026] 在某些实施例中，方法还包括在计算上去除识别为在胎儿和怀孕个体中均是杂合的多态性的等位基因频率。在一些实施方式中，在(c)之前，方法包括在计算上去除识别为在胎儿和怀孕个体中均是纯合的多态性的等位基因频率的操作。在一些实施方式中，在(c)之前，方法包括在计算上去除识别为在怀孕个体中是杂合的多态性的等位基因频率的操作。

[0027] 从怀孕个体的体液所得到的DNA可以是从怀孕个体的血浆所得到的游离DNA。从液体所得到的核酸的映射可通过将段映射到多态性的数据库来实现。

[0028] 本公开的这个方面的方法还可包括在产生包含多态性序列的DNA段序列的条件

下、采用核酸测序器对于来自怀孕个体的体液的DNA进行测序。

[0029] 在一些实施方式中，(a)中的映射包括识别多个二等位基因多态性序列。在其它实施例中，(a)中的映射包括将DNA段映射到多个预定义多态性序列。

[0030] 在一些实施例中，这个方面的方法还包括：在一个或多个处理器上运行程序指令，以便将(c)中所确定的DNA的胎儿的分数自动记录在怀孕个体的患者病历中，其存储在计算机可读介质上。患者病历可由实验室、医生办公室、医院、保健组织、保险公司或者个人病历网站来保持。

[0031] 基于胎儿DNA分数的估计，这个方面的方法还可包括规定、发起和/或改变受检者(从其获取母体测试样本)的治疗。基于胎儿DNA分数的估计，这个方面的方法还可包括指示和/或执行一个或多个附加测试。

[0032] 按照本公开的又一方面，提供方法以用于使用下列操作来估计从怀孕个体的体液所得到的DNA中的胎儿DNA的分数：(a)接收体液的样本；(b)在提取体液中存在的母体基因组和胎儿基因组的DNA的条件下从样本中提取DNA；(c)在产生DNA段序列的条件下采用核酸测序器对所提取DNA进行测序；(d)比较从液体所得出的DNA段序列，并且由比较来识别一个或多个二等位基因多态性；(e)对于所识别多态性的至少一个来确定DNA段序列的等位基因频率；(f)基于怀孕个体的接合性和胎儿的接合性的组合来分类至少一个所识别多态性；以及(g)使用在(e)中所确定的等位基因频率和来自(f)的接合性的组合来估计从怀孕个体所得到的DNA中的胎儿DNA的分数。

[0033] 映射可使用计算装置来执行，其编程为将核酸序列映射到一个或多个所指定多态性。一般来说，操作(d)-(g)的任一个可在程序指令下在一个或多个处理器上执行。

[0034] 在这个方面的某些实现中，DNA段序列的长度在大约20个碱基对与大约300个碱基对之间。

[0035] 在这个方面的某些实施例中，(f)中的分类将至少一个所识别多态性分为下列组合之一：(i)怀孕个体是纯合的，并且胎儿是纯合的，(ii)怀孕个体是纯合的，而胎儿是杂合的，(iii)怀孕个体是杂合的，而胎儿是纯合的，以及(iv)怀孕个体是杂合的，并且胎儿是杂合的。方法还可包括不予考虑组合(i)或组合(iv)中分类的任何多态性。

[0036] 按照各个实施例，这个方面的方法可包括如本文结合其它方面所述的过滤和/或分类操作。例如，这个方面的方法可包括过滤一个或多个所识别多态性，从而不予考虑具有比所定义阈值要大的少数等位基因频率的任何多态性。在一些情况下，分类至少一个所识别多态性包括将阈值应用于(e)中所确定的等位基因频率。如本文所述的混合模型的使用可用于分类所识别多态性。

[0037] 本公开的另一方面涉及用于估计胎儿DNA的分数并且包括下列元件的装置：(a)测序器，配置成(i)接收从包含母体基因组和胎儿基因组的DNA的体液样本所提取的DNA，并且(ii)对所提取DNA测序，以便产生DNA的序列段；以及(b)计算装置，配置成指示一个或多个处理器(i)将从怀孕个体的体液所得到的DNA的序列段映射到多个多态性序列，(ii)从DNA的已映射序列段确定对于多个多态性序列的每个的等位基因频率，并且(iii)将等位基因频率应用于混合模型，以便得到从携带胎儿的个体的血液所得到的DNA中的胎儿DNA分数的估计。

[0038] 用于估计胎儿DNA的分数的又一个装置包括下列元件：(a)测序器，配置成(i)接收

从包含母体基因组和胎儿基因组的DNA的体液的样本所提取的DNA，并且(ii)在产生DNA段序列的条件下对所提取DNA测序；以及(b)计算装置，配置成指示一个或多个处理器(i)比较从体液所得出的DNA段序列，并且由比较来识别一个或多个二等位基因多态性，(ii)对于所识别多态性的至少一个来确定DNA段序列的等位基因频率，(iii)基于怀孕个体的接合性和胎儿的接合性的组合来分类至少一个所识别多态性，并且(iv)使用等位基因频率和接合性的组合来估计从怀孕个体所得到的DNA中的胎儿DNA的分数。

[0039] 本文所述装置方面所采用的指令和/或硬件可提供本文所公开方法方面的计算或算法操作的任一个或多个的执行，而与以上是否明确叙述这类操作无关。

[0040] 下面将参照关联附图更详细描述所公开实施例的这些及其它特征和优点。

附图简介

[0041] 图1是示出给定基因组位置的胎儿和母体接合性状态的分类的框图。

[0042] 图2是用于实现所公开实施例的一部分的示例过程流程。

[0043] 图3提供根据测序碱基位置对于使用Eland以缺省参数与人类基因组HG18对齐的30个通道Illumina GA2数据的误差估计。

[0044] 图4是对于杂合性情况1至4的少数等位基因计数A与覆盖D(假定没有误差)的图表。

[0045] 图5示出情况3数据到情况2的变换。

[0046] 图6提供后旋转数据，其中D1选择成使得情况1和情况2、3没有重叠。E1表示情况1数据的99％上置信区间的上限。

[0047] 图7示出使用混合模型以及已知胎儿分数和估计胎儿分数的结果的比较。

[0048] 图8示出使用机器误差率作为已知参数将上偏差降低某个点。

[0049] 图9中示出使用机器误差率作为已知参数的模拟数据，增强情况1和2误差模型将上偏差极大地降低到小于低于0.2的胎儿分数的某个点。

[0050] 图10是计算机系统的示意图示，其在适当配置(例如编程)或设计时能够用作所公开实施例的分析装置。

[0051] 图11A和B示出在如一个示例中产生的染色体1(A)和染色体7的少数等位基因百分比(A/D)的变异观测(频率)的数量的直方图。

[0052] 图12A和B示出沿染色体1(A)和染色体7的等位基因频率的分布。

具体实施方式

[0053] 介绍和概述

[0054] 某些所公开实施例涉及分析从怀孕女性的血液所获取的DNA，并且使用分析来确定来自胎儿的那个DNA的分数。DNA的胎儿分数则可用于将某个置信等级归结于基于从母体血液所获取的DNA的单独分析的胎儿的另一个度量或表征。例如，从母体血液所获取的胎儿DNA样本可经过单独分析，以便检测怀孕女性所怀有的胎儿的异倍性。通过这个单独分析进行的异倍性确定可基于从母体血液所获取的DNA中存在的胎儿DNA的分数量、通过统计稳固置信等级来给出。DNA的总补体中的胎儿DNA的较低分数表明基于胎儿DNA的任何表征的低置信。

[0055] 通常但不一定，母体血液中的被分析DNA是游离DNA，但是在一些实施例中，它可以是联胞DNA。游离DNA从母体血浆中获取。从怀孕女性所获取的游离DNA含量中的胎儿DNA的量根据包括包括胎儿的孕龄在内的多种因素广泛地改变。对于典型怀孕女性，当前认为游离DNA的大约5-20％是胎儿DNA。但是，胎儿分数显著更低(例如大约1％或更低)并非不常见。在这类情况下，胎儿DNA的任何独立表征都固有地是可疑的。另一方面，一些研究人员已经报道具有高达40％或50％的胎儿DNA分数的母体游离DNA样本。

[0056] 在本文所述的某些实现中，母体DNA的胎儿分数的确定依靠在已知为持有一个或多个多态性的序列位点的多个DNA序列读数。通常但不一定，这类多态性是单核苷酸多态性(SNP)。其它类型的适当多态性包括缺失、STR(短串联重复)、插入、插入缺失标记(包括微插入缺失)等。下面提供其它示例。在某些实施例中，多态性位点存在于“参考序列”上，如下面所述。在一些实施例中，发现多态性位点，同时将序列标签相互对齐和/或与参考序列对齐。

[0057] 某些所公开方法利用如下事实：在所考虑的多态性位点的胎儿的DNA序列可能与其母体的不对应。例如，在特定SNP的位点的母体DNA可以是纯合的，而胎儿的SNP形式将是杂合的。因此，对所述的SNP所获取的序列样本的集合将是异质的，其中序列的大多数包含多数(maior)等位基因而其余分数包含少数等位基因。多数和少数等位基因的相对量通过样本中的胎儿DNA的分数来确定。

[0058] 应当指出，在纯合样本中，给定SNP或其它多态性的两种副本均包含相同等位基因，而杂合SNP或另一多态性包含多数等位基因的一个副本和少数等位基因的一个副本。因此知道，从杂合个体排他获取的DNA应当包含多数等位基因的50％和少数等位基因的50％。该知识能够用于说明胎儿DNA的分数中，如下面所述。如下面更全面所述，本文所公开的各种方法仅考虑多态性，其中在母体和胎儿DNA中共同仅存在两个等位基因。

[0059] 在一些实施方式中，多次读取从母体血液所获取的DNA，其中映射到多态性的特定位点的总读取次数被认为是多态性的“覆盖”，以及映射到那个多态性的少数等位基因的读取次数被认为是少数等位基因计数。少数等位基因计数与覆盖的比率在各种实现中是重要的。

[0060] 本文所公开的某些方法识别和表征包括来自母体和胎儿的DNA的DNA样本中的多态性的四种情况。下面，图1示出这四种情况。具体来说，在相当不感兴趣的第一情况中，母亲和胎儿在所考虑的特定多态性均是纯合的。在这种情况下，包含所述的多态性的DNA样本中的每一个序列将包含相同等位基因，并且能够不收集与来自母亲和胎儿的DNA的相对量有关的信息。但是，应当注意，在它允许研究人员或技术人员获得用于生成所考虑的序列数据的DNA测序装置的相对误差率的某个概念的意义上，这种情况可能是感兴趣的。

[0061] 分析将遇到的第二情况是一种多态性，怀孕女性对其是纯合的而胎儿对其是杂合的。在这种情况下，所检测序列的较小但是显著的分数将包含少数等位基因。具体来说，在这第二情况中，少数等位基因的频率标称地由母体血流中的胎儿DNA的分数除以二给出。

[0062] 在第三情况中，所考虑的多态性在母体DNA中是杂合的，而在胎儿DNA中是纯合的。在这种情况下，少数等位基因的频率标称地通过0.5减去DNA样本中的胎儿DNA的分数的一半来给出。

[0063] 最后，在第四情况中，所考虑的多态性在母体和胎儿中均是杂合的。在这种情况下，预计多数和少数等位基因的频率均为0.5。如同第一情况一样，第四情况对于确定DNA的

胎儿分数是比较不感兴趣的。

[0064] 如果分派有确定样本中的胎儿DNA的分数的任务的研究人员、技术人员或软件对于给定多态性知道那个多态性属于四种情况的哪一种，则可以直接估计胎儿DNA的分数(假定所考虑的多态性落入情况二或者情况三)。但是，实际上，不能先验地具有该知识。因此，要求计算装置执行本文所述的操作。

[0065] 在本文其它部分所述的某些实施例中，阈值技术用于将单个多态性分类为四种情况之一。一旦多态性经过这样分类并且被发现驻留在情况2或者3中，可以估计胎儿分数。在其它实施例中，该技术考虑分布于基因组的全部或者一部分的多个多态性。如具体示例所示，跨基因组的多个不同的SNP可用于这个目的。

[0066] 在具体实施例中，对于从母体血液样本所获取的DNA样本中的多个不同多态性来确定等位基因频率。对于这多个多态性，某个分数将对应于接合性情况1，另一个分数将对应于情况2，第三分数将对应于情况3，以及最后分数将对应于情况4。这些分数将合计为值1。混合模型或者相关技术可用于挑出这四个类别的每个中的多态性的一个或多个统计性质。具体来说，混合模型可用于确定从怀孕女性血液所获取的DNA样本中遇到的四种情况的每个的均值以及可选的方差。在具体实施例中，这是与相对于所述的多态性的计数总数(覆盖)的少数等位基因的频率关联的均值和方差。如本文其它部分所述，这四个类别的每个的平均值或者至少第二和第三类别与从母体血液所获取的DNA中的胎儿分数直接相关。

[0067] 在采用混合模型的具体实现中，对于其中考虑多态性的各位置来计算一个或多个阶乘矩。例如，阶乘矩(或者阶乘矩的集合)使用DNA序列中考虑的多个SNP位置来计算。如下式4所示，各种阶乘矩的每个是对给定位置的少数等位基因频率与覆盖的比率所考虑的所有各种SNP位置的合计。如下式5所示，这些阶乘矩还与关联上述四种接合性情况的每种的参数相关。具体来说，它们与每种情况的概率以及所考虑的多态性集合中的四种情况的每种的相对量相关。如所述，概率是母体血液的游离DNA中的胎儿DNA的分数的函数。如下面更全面说明，通过计算充分数量的这些阶乘矩(其在等式4中示出)，该方法提供充分数量的表达式来求解所有未知数。这种情况中的未知数是所考虑的多态性的群体中的四种情况的每种的相对量以及与这四种情况的每种关联的概率(并且因此还有胎儿DNA分数)。参见等式5。相似结果能够使用如下式7-12所表示的混合模型的其它形式来得到。这些特定形式仅利用落入情况1和2的多态性，其中情况3和4的多态性通过阈值技术来过滤。

[0068] 因此，阶乘矩可用作混合模型的组成部分，以便识别接合性的四种情况的任何组合的概率。以及如所述，这些概率或者对于第二和第三情况的至少那些概率与母体血液的总游离DNA中的胎儿DNA的分数直接相关。

[0069] 还应当指出，测序误差可用于降低必须求解的阶乘矩等式的系统的复杂度。在这点上，应当知道，测序误差实际上能够具有四个结果(对应于在任何给定多态性位置的四个可能碱基的每个)的任一个。

[0070] 在某些实施例中，将标签与参考染色体或基因组对齐，并且识别二等位基因多态性。这些多态性不是预先定义的或者在对齐之前识别的。它们只在对齐期间来识别，然后基于其接合性和少数等位基因计数来表征，如本文所述。这个信息用于估计基因组分数，如本文所述。

[0071] 本文所述实施例中使用的标签的长度将一般通过用于生成标签的测序方法来确

定。方法跨大范围标签长度是健壮的。在某些实现中，标签的长度在大约20与300个碱基对(或者长度在大约30至100个碱基对)之间。

[0072] 用于实现所公开实施例的一部分的示例过程流程在图2中示出。如其中所示，过程开始于201，其中从母体血液或其它体液收集DNA(游离或联胞)。由此DNA，多个序列映射到参考序列中的一个或多个多态性。这个映射为每个多态性提供等位基因频率。参见框203。

[0073] 更具体来说，在框203的过程可涉及读取多个多态性的位置的所收集DNA的序列。在一些情况下，这些可作为用于针对胎儿DNA进行的倍数性确定或其它确定的过程的组成部分来生成。因此，在一些实施例中，独立序列无需生成。将读数序列与参考序列对齐，以便使用BLAST或类似工具来使对齐为最大。

[0074] 参考序列可作为多态性的数据库来提供。在一些情况下，这是从所有多态性定义的组合扩充所产生的等位基因搜索参考集合(例如在多态性是SNP的情况下的所有SNP序列)。例如参见附录。在具体示例中，序列的长度为大约100至150个碱基对。

[0075] 回到图2，该方法对于框203的操作中考虑的多态性的一个或多个来确定母体/胎儿接合性组合。参见框205。在某些实施例中，混合模型可用于此目的。如所述，组合如下：M&F纯合，M纯合和F杂合，M杂合和F纯合，以及M&F杂合。

[0076] 最后，如框207所示，该方法使用在多态性的一个或多个的接合性情况等位基因频率的组合来估计来自母体样本的DNA中的胎儿成分的分数量。

[0077] 定义

[0078] 提供以下论述作为了解所公开实施例的某些方面和优点的辅助。

[0079] 术语“读数”指的是来自核酸样本的一部分读数序列。通常但不一定，读数表示样本中的毗连碱基对的短序列。读数可通过样本部分的碱基对序列(按照ATCG)以符号表示。它可存储在存储器装置中，并且经过适当处理，以确定它是否匹配参考序列或者满足其它标准。读数可直接从测序装置或者直接从与样本有关的所存储序列信息来得到。

[0080] 术语“标签”还指的是来自核酸样本的短序列。通常，标签包含关联信息，例如基因组中的序列的位置。为了某些目的，术语“读数”和“标签”在本文是可互换的。但是，序列读数通常与参考序列对齐，以及仅在参考基因组上的一个位点进行映射的读数称作标签。“段序列”在本文中有时与“标签”可互换地使用。

[0081] 本文中频繁的“读数”描述为长度为36个碱基对(36mers)的核酸序列。当然，所公开的实施例并不限于这种大小。在许多应用中，更少和更大的读数是适当的。对于将读数与人类基因组对齐的应用，大小30个碱基对或更大的读数一般被认为充分地将样本映射到单个染色体。大许多的标签/读数适合于某些应用。对于整个基因组测序，可使用大约1000个碱基对或更大的读数。在某些实施例中，读数可具有在大约20与10000个碱基对之间、或者大约30与1000个碱基对之间、或者大约30与50个碱基对之间的长度。

[0082] “参考序列”是生物分子序列，其频繁地是核酸、例如染色体或基因组。通常，多个读数是给定参考序列的成员。在某些实施例中，将读数或标签与参考序列进行比较，以便确定参考序列是否包含读数序列。这个过程有时称作对齐。

[0083] 在多种实施例中，参考序列比与其对齐的读数明显要大。例如，它可以是大约至少要大100倍，或者大约至少要大1000倍，或者大约至少要大10000倍，或者大约至少要大10^5倍，或者大约至少要大10^6倍，或者大约至少要大10^7倍。

[0084] 在一个示例中，参考序列是全长度人类基因组的序列。这类序列可称作基因组参考序列。在另一个示例中，参考序列局限于特定人类染色体、例如染色体13。这类序列可称作染色体参考序列。参考序列的其它示例包括其它种类的基因组以及任何种类的染色体、子染色体区域(例如链)等。

[0085] 在各个实施例中，参考序列是共有序列或者从多个个体所得出的其它组合。但是，在某些应用中，参考序列可从特定个体来获取。

[0086] 术语“对齐”指的是将读数或标签与参考序列进行比较、并且由此确定参考序列是否包含读数序列的过程。如果参考序列包含读数，则该读数可映射到参考序列，或者在某些实施例中映射到参考序列中的特定位置。在一些情况下，对齐只告知读数是否为特定参考序列的成员(即，读数在参考序列中是否存在)。例如，读数与人类染色体13的参考序列的对齐将告知该读数在染色体13的参考序列中是否存在。提供这个信息的工具可称作集合关系测试器。在一些情况下，对齐还指明参考序列中的位置(读数或标签映射到其中)。例如，如果参考序列是整个人类基因组序列，则对齐可指明读数存在于染色体13上，并且还可指明读数处于染色体13的特定链上。

[0087] “位点”是与读数或标签对应的参考序列中的唯一位置。在某些实施例中，它指定染色体(例如染色体13)的识别码、染色体链以及染色体中的准确位置。

[0088] “多态位点”是核苷酸序列发散发生的基因座。基因座可以小至一个碱基对。说明性标记具有至少两个等位基因，各以大于1%、以及更通常大于所选群体的10%或20%的频率发生。多态位点可以小至一个碱基对。术语“多态基因座”和“多态位点”在本文中可互换地使用。

[0089] 本文中的“多态序列”指的是核酸序列、例如DNA序列，其包括一个或多个多态位点，例如一个SNP或串联SNP。按照本技术的多态序列能够用于具体区分包含胎儿和母体核酸的混合的母体样本中的母体和非母体等位基因。

[0090] 详细实施例

[0091] 通常，本文所述的过程采用参考序列，其跨越一个或多个多态性，并且与所取样的DNA关联。参考序列可以是例如人类基因组、染色体或者染色体中的区域。能够为了便于估计胎儿DNA分数而指定多态性的一个或多个。被指定以用于确定胎儿分数的多态性是预先知道的多态性。例如，参考、事实和关于预先已知STR的序列信息以及相关群体数据的综合列表在STRBase中统计，其可经由万维网在ibm4.carb.nist.gov:8800/dna/home.htm来访问。来自常用STR基因座的GenBank®(http://www2.ncbi.nlm.nih.gov/cgi-bin/genbank)的序列信息也是通过STRBase可访问的。能够访问的预先已知SNP的信息是从公共可访问数据库可得到的，包括但不限于在万维网地址wi.mit.edu的Human SNP Database、在万维网地址ncbi.nlm.nih.gov、万维网地址lifesciences.perkinelmer.com的NCBI dbSNP Home Page、在万维网地址appliedbiosystems.com的Applied Biosystems by Life Technologies™(Carlsbad,CA)、在万维网地址celera.com的Celera Human SNP数据库、在万维网地址gan.iarc.fr的SNP Database of the Genome Analysis Group(GAN)。在一个实施例中，为确定胎儿分数所指定的SNP从Pakstis等人(Pakstis等人，Hum Genet 127：315-324[2010])所述的92个个体标识SNP(IISNP)的组中选取，其示为在跨群体的频率中具有极小变化($F_{st}<0.06$)，以及在全球是高度信息性，具有平均杂合性≥0.4。由本发明的方

法包含的SNP包括连锁和无连锁SNP。为了指定适当串联SNP序列，能够搜索International HapMap Consortium数据库(International HapMap Project，Nature 426:789-796[2003])。数据库在万维网的hapmap.org可得到。

[0092] 这样采用的多态性可以是为确定胎儿DNA分数所指定的预先已知多态性的面板，或者它们可在为了其它目的、例如将样本DNA标签映射到染色体的母体DNA的分析中偶然发现。

[0093] 在某些实施例中，该方法包括使用基因组的混合来对样本、例如包括胎儿和母体游离DNA的母体样本中的DNA进行测序，以便提供多个序列标记，其映射到包括参考基因组上的预先已知多态性位点的序列，并且使用在预先已知位点所映射的标签来确定胎儿分数，如下面详细描述。备选地，接着DNA的测试，通过测序技术、例如NGS所得到的序列标签被映射到参考基因组、例如hg19，以及映射到位点(多态性在其中偶然发生、即不是预先已知)的序列标签用于确定胎儿分数。

[0094] 参考序列(序列标签对其映射到预先已知多态性位点)能够是已发布参考基因组，或者它能够是所考虑的多态性的序列的人工数据库或者其它预定义集合。数据库序列的每个将跨越与多态性关联的一个或多个核苷酸。作为一个示例，参见下面在“附录1”中提供的多态性序列的列表。

[0095] 在各个实施例中，用于估计胎儿DNA分数的多态性的数量是至少2个多态性，更具体来说是对于至少大约10个多态性的每个，以及更优选地地是对于至少大约100个多态性的每个。

[0096] 在一个示例中，通过将所生成序列与从SNP定义的组合扩充所构成的参考基因组对齐，来确定SNP覆盖和等位基因频率。扩增子数据库包含由例如侧面序列的至少大约50个碱基所围绕的二等位基因变化信息。例如，具有变化信息串“[g/c]”(表示交替等位基因“g”和“c”)的扩增子可看来像是：

[0097] atcg.....accg[g/c]ccgt....

[0098] 在一些情况下，输入扩增子数据库和所生成序列以及输出SNP/等位基因计数的过程如下。

[0099] 1.从SNP定义的组合扩充来创建等位基因搜索参考集合。对于扩增子数据库中的各序列，对于变化信息串中的各等位基因，创建具有由等位基因所替代的变化信息串的等位基因序列。

[0100] a.例如，考虑上述示例扩增子序列，会创建两个序列：1)atcg.....accgGccgt...，以及2)atcg.....accgCccgt...。

[0101] b.全等位基因搜索参考集合的一个示例能够见于等位基因搜索数据库序列列表。

[0102] 2.将序列映射到等位基因搜索参考集合，仅保持仅匹配搜索集合中的一个序列的映射。

[0103] 3.通过计算匹配其等位基因序列的序列数量，来确定等位基因计数。

[0104] 本文所公开的方法假定“正常”妊娠，即，其中母亲仅怀有一个胎儿而不是双胞胎、三胞胎等的妊娠。本领域的技术人员将会理解修改，其考虑非正常妊娠，特别是胎儿数量为已知的那些妊娠。

[0105] 如所示，当确定胎儿分数时，该方法对来自母体血液的样本中的DNA进行测序，并

且计数映射到所考虑的多态性的各序列的序列标签。对于各多态性，该方法记录(taillies)映射到它的读数的总数(覆盖)以及与各等位基因关联的序列标签的数量(等位基因计数)。在一个简单示例中，具有5的覆盖的多态性可具有等位基因B的3个读数以及等位基因A的2个读数。在这个示例中，等位基因A被认为是少数等位基因，以及等位基因B被认为是多数等位基因。

[0106] 在一些实施例中，这个操作利用非常快速的测序工具，例如整体平行DNA测序工具。下面更详细地描述这类工具的示例。在一些情况下，对于单个样本读数数千或者数百万标签序列。优选地，测序按照如下方式执行：允许将被测序DNA快速直接指配给持有所考虑的多态性的特定预定义序列。一般来说，在大小30个碱基对或者更多的标签中存在用于此目的的充分信息。这个大小的标签能够明确地映射到感兴趣序列。在一个特定实施例中，过程中采用的标签序列的长度是36个碱基对。

[0107] 将标签映射到参考基因组或者映射到等位基因序列数据库中的序列(例如，参见如前面所述的附录1)，并且确定这样映射的标签数量。这将为所考虑的各多态性提供覆盖和少数等位基因计数。在一些情况下，这可与将各标签映射到23个人类染色体之一并且确定每人染色体的所映射标签的数量同时地进行。

[0108] 如所述，覆盖是读数序列的总数，其映射到参考序列中的给定多态性。映射到这种多态性的读数序列的总数中的等位基因计数具有等位基因。所有等位基因计数的总和必须等于覆盖。具有最高计数的等位基因是多数等位基因，以及具有最低计数的等位基因是少数等位基因。在某些实施例中，估计胎儿DNA分数所需的唯一信息是对于多个多态性的每个的覆盖和少数等位基因计数。在一些实施例中，还使用DNA测序装置的碱基识别误差率。

[0109] 有用的是考虑本文所公开的某些方法的数学或符号基础材料。如所述，在各个示例中，从母体血液所生成的序列与参考基因组或者其它核酸序列对齐(重叠成使得相同碱基为最大)。给定基因组位置j以及与参考对齐的序列集合，设所对齐序列之中的四个DNA碱基(“a”、“t”、“g”和“c”，又称作“等位基因”)的每个的出现次数分别为w(j,1)、w(j,2)、w(j,3)和w(j,4)。为了便于本论述，可不失一般性地假定所有变化是二等位基因。因此，可使用下列符号：

[0110] 当$B \equiv B_j \equiv \{b_j\} \equiv w_{j,i}^{(2)} = \max_{i\in\{1,2,3,4\}}\{w_{j,i}\}$时在基因组位置j的多数等位基因，作为在位置j的计数的一阶统计(多数等位基因b是对应argmax。当考虑一个以上SNP时，使用下标。)，

[0111] 当$A \equiv A_j \equiv \{a_j\} \equiv w_{j,i}^{(2)}$时在位置j的少数等位基因计数，作为在位置j的计数的二阶统计(即，第二最高等位基因计数)，

[0112] 当$D \equiv D_j = \{d_j\} = A_j + B_j$时在位置j的覆盖，以及

[0113] 测序机器误差率表示为e。

[0114] 当上下文清楚时，为了方便起见，可互换地使用符号；例如，A、Ai或{ai}对于少数等位基因或者少数等位基因计数可互换地使用。可以使用或者可以不使用下标，这取决于是否考虑一个以上SNP。(仅为了示例而使用SNP。如本文其它部分所述，可使用其它类型的多态性。)

[0115] 图1中，示出多态性接合性的四种状态的基础。如所示，母亲在给定多态性可以是

纯合或杂合的。类似地，婴儿在相同位置可以是杂合或者纯合的。如所示，情况1和2是多态性情况，其中母亲是纯合的。如果婴儿和母亲均为纯合的，则多态性是情况1多态性。如上所述，这种情况通常不是特别感兴趣的。如果母亲是纯合的而婴儿是杂合的，则胎儿分数f标称地通过少数等位基因与覆盖的比率的两倍来给出。在母亲为杂合而婴儿为纯合的多态性情况(图1的情况3)中，胎儿分数标称地为一减去少数等位基因与覆盖的比率的两倍。最后，在母亲和胎儿均为杂合的情况下，少数等位基因分数应当始终为0.5，不包括误差。对于落入情况4中的多态性无法得出胎儿分数。

[0116] 现在将进一步说明四种情况。

[0117] 情况1:母亲和婴儿纯合

[0118] -在这种情况下，不包括测序误差或污染，应当没有观测到差异。

[0119] -E(最小等位基因频率)＝E(A)＝0。

[0120] -实际上，A～(分布为)二项式分布，其通过低np的泊松分布良好地近似计算。二项式或泊松的分布率参数与测序误差率e和覆盖D相关。图3示出与人类参考基因组对齐的所生成36mer序列的失配频率。

[0121] -这种情况没有包含与胎儿分数有关的信息。

[0122] 图3提供根据测序碱基位置对于使用Eland以缺省参数与人类基因组HG18对齐的30个通道Illumina GA2数据的误差估计。

[0123] 情况2:母亲纯合而婴儿杂合

[0124] -在这种情况下，对于小胎儿分数(f)，所观测等位基因频率将显著不同。其中多数等位基因以比少数等位基因要多若干倍的频率发生。

[0125] -不包括误差，给定单个SNP位置(D,A)，E(A)＝Df/2，并且f的未偏置估计为2A/D

[0126] -不包括误差，A～二项式(f/2,D)。均值Df/2，方差(1-f/2)Df/2。[如果D＞15，则近似为正态分布]。

[0127] 情况3:母亲杂合而婴儿纯合

[0128] -在这种情况下，多数和少数等位基因的所观测频率接近，并且A/D刚好在0.5之下。

[0129] -不包括误差，E(A)＝D(1-f)/2，以及E(1-(2A/D))＝f

[0130] -不包括误差，A～二项式((1-f)/2,D)。均值D((1-f)/2)，方差D/4(1-f^2)。

[0131] 情况4:母亲杂合并且婴儿杂合

[0132] 注意，不包括误差，对此存在两种子情况。

[0133] 情况4.1:来自父亲的等位基因与母亲的等位基因不同

[0134] 这会引入第三等位基因，其是E(A)＝Df/2的少数等位基因。这些情况不应当对f的估计具有影响，因为用于向扩增子指配序列的过程将在参考SNP为二等位基因时滤出这些情况。

[0135] 情况4.2:来自父亲的等位基因匹配母亲的等位基因之一

[0136] -在这种情况下，不包括误差，两个等位基因会以1:1比例出现，使得这种情况对于胎儿分数估计不是有用的。

[0137] -不包括误差，E(A)＝0.5，以及A～二项式(0.5,D)以0.5来截取。

[0138] 图4提供对于杂合性情况1至4的少数等位基因计数A与覆盖D(假定没有误差)的图

表。

[0139] 在各个实施例中，该方法广义地涉及分析在一个或多个SNP(或者其它多态性)的等位基因频率，以便将多态性分类为在情况2和/或情况3中。与分类结合使用等位基因频率，该方法能够估计胎儿分数。

[0140] 在一些情况下，给定少数等位基因计数A和覆盖D，换言之，对于个体SNP位置，单点(D,A)允许方法进行单点估计。例如，某些方法将具有等位基因计数(D,A)的SNP分类为单个情况，并且得出如下胎儿分数估计：

[0141] ES1.1判定情况的简单阈值

[0142] 给定个体位置(SNP)，

[0143] 1.采用例如2A/D＜e或者二项式(e,D)或泊松(De)的所定义临界值，对情况1进行判定。在本发明的范围之内还可使用备选分布)。没有胎儿分数(f)估计。

[0144] 2.如果2A/D＞(0.5-e)或者二项式(0.5,D)的某个临界值，(或者其它适当近似分布)，则对情况4进行判定。对f的估计不使用位置。

[0145] 3.否则，如果2A/D＜0.25(或者另外某个手动设置或者自动估计阈值)，则对情况2进行判定。胎儿分数f估计为2A/D

[0146] 4.否则，情况3。使用胎儿分数估计f＝(1-2A/D)。

[0147] 能够通过组合来自若干SNP的等位基因计数信息以估计胎儿分数，来获得精度。

[0148] 方法EMI：通过求平均来组合多个SNP。

[0149] 取均值、中值、其它中心测量(例如：Tukey二加权、M估计量等...)。还可使用加权平均。对于可如何定义加权的示例，参见以下EM2.4。另外，可使用中心的健壮量度。

[0150] 方法EM2通过变换的来自情况2和情况3的同时估计

[0151] 对于f小于X％的情况，情况3的点(D,A)能够变换为与情况2的点一致。由此线条，公共斜率能够经由通过原点(参见图5)的回归来计算。

[0152] 基于变换的方法的一个理论缺点是情况2和3的二项式分布将具有不同形状。在典型胎儿分数等级(＜10％)，情况2数据将具有向右偏斜的接近泊松的分布，以及情况3将具有接近正态的分布。

[0153] 图5示出情况3数据到情况2的变换。现在，单个回归能够同时从两种情况来估计f。

[0154] 用于计算EM2.3的方法：

[0155] 步骤1：扔掉情况4数据

[0156] 对于数据点(D,A)，如果A＞(0.5D-T1)，则排队(D,A)不进行进一步分析。T1(D,A)，实值函数。

[0157] 步骤2：变换情况3数据

[0158] 参见图6。对于没有划为4的各数据点(D,A)，如果A＞T2×D，则将点变换到新坐标(D1,A1)。T2(D,A)，实值函数。

[0159] $$\alpha = \frac{2A}{D}$$

[0160] A1＝0.5D-A

[0161] D1＝D

[0162] 步骤3：建立阈值DT，以降低来自情况1数据的污染

[0163] 丢弃低于T2(D,A)、即实值函数的所有数据点。

[0164] 步骤4:用于剩余变换情况2和3数据的回归估计。将通过原点的回归应用于剩余点。胎儿分数估计是回归线斜率的两倍。

[0165] 注意,存在许多变换类,其能够构造成完成情况2和3数据的相同一致。示例包括三角、变换或者旋转矩阵的使用。这些推导预计包含在本公开的范围之内。此外,能够使用许多回归类(L2,L1,...)或优化。交换优化算法是轻微变化,并且涵盖在本公开的范围下。

[0166] 图6提供后旋转数据。选择D1,使得情况1和情况2和3没有重叠。E1表示情况1数据的99%上置信区间的上限。

[0167] 方法EM3加权最小二乘

[0168] 来自EM2.3的回归方法假定所有所转化数据点均具有相等方差。更适当的是考虑不同数据源、甚至来自相同杂合性模式的点的异方差性。

[0169] 步骤1至3与EM2.3相同。

[0170] 步骤4:回归

[0171] 在来自EM2.3的回归中,来自情况2数据的点将具有方差v2(f,D)=[0.5*Df-0.25*Df^2],以及来自情况3数据的点将具有方差v3(f,D)=[0.25D(1-f^2)]。假定我们对每个点给予不同加权w,如同EM2.3中一样,我们设法使下式为最小

[0172] $$Q=\sum_{i=1}^{n} w_i(a_i - sd_i)^2$$

[0173] 等式1

[0174] 将一阶导数设置为零,并且求解s:

[0175] $$\frac{\partial Q}{\partial s}=\sum_{i=1}^{n} 2w_i(d_i - sa_i)(-a_i)=0$$

$$\sum_{i=1}^{n} sa_i^2 - \sum_{i=1}^{n} 2w_i a_i x_i = 0$$

and

$$s=\frac{\sum_{i=1}^{n} 2w_i d_i a_i}{\sum_{i=1}^{n} a_i^2}$$

[0176] 其中,d_i是SNP i的覆盖,以及a_i是SNP i的(对情况3变换的)少数等位基因计数。

[0177] 等式2

[0178] 这种方法以每个点的方差的倒数进行加权,适当地估计为v2(2A/D,D)或者v3(2A/D,D)。胎儿分数估计是2×s。

[0179] 在某些实施例中,混合模型可用于将多态性的集合分类为接合性情况的两个或更多,同时从这些情况的每个的平均等位基因频率来估计胎儿DNA分数。一般来说,混合模型假定数据的特定集合由不同类型的数据的混合来组成,其各具有自己的预计分布(例如正

态分布)。该过程尝试查找每种类型的数据的均值以及可能的其它特性。在本文所公开的实施例中,存在总共四种不同数据类型(接合性情况),其构成所考虑的多态性的少数等位基因频率数据。

[0180] 在以下小节中提供混合模型的一个实现。在这个实施例中,次要作功频率A是如等式3所示的四项之和。每项对应于四个接合性情况之一。每项是多态性分数α和少数等位基因频率的二项式分布之积。αs是落入四种情况的每种上的多态性的分数。各二项式分布具有关联概率p和覆盖d。例如,情况2的少数等位基因概率通过f/2来给出。

[0181] 所公开实施例利用所考虑的等位基因频率数据的"阶乘矩"。如众所周知,分布的均值是一次矩。它是少数等位基因频率的期望值。方差是二次矩。它从平方等位基因频率的期望值来计算。

[0182] 跨所有多态性的等位基因频率数据可用于计算阶乘矩(一次阶乘矩、二次阶乘矩等),如等式4所示。如这些等式所示,阶乘矩是项对于i、数据集中的个体多态性的合计,其中在数据集中存在n个这类多态性。求和项是少数等位基因计数α_i和覆盖d_i的函数。

[0183] 有用地,阶乘矩与α_i和p_i的值具有关系,如等式5所示。从概率p_i,能够确定胎儿分数f。例如,$p_2=f/2$,以及p_3为1-f/2。因此,负责逻辑能够求解将未知数αs和ps与跨所考虑的多个多态性的少数等位基因分数的阶乘矩表达式相关的等式系统。当然,在本发明的范围之内存在用于求解混合模型的其它技术。

[0184] 有用的是还考虑本文所公开的混合模型实施例的数学或符号基础材料。以上所述的四种杂合性情况表明点(ai,di)中的ai的分布的以下二项式混合模型:

[0185] $A=\{a_i\}\sim\alpha_1 Bin(p_1,d_i)+\alpha_2 Bin(p_2,d_i)+\alpha_3 Bin(p_3,d_i)+\alpha_4 Bin(p_4,d_i)$

[0186] 其中

[0187] $1=\alpha_1+\alpha_2+\alpha_3+\alpha_4$

[0188] $m=4$

[0189] 等式3

[0190] 下面描述用于将p_i与胎儿分数和测序误差率相关的各种模型。参数α_i涉及群体特定参数,以及设这些值"浮动"的能力针对例如双亲的种族性和后裔等因素对这些方法给予附加健壮性。

[0191] 对于各种杂合性情况,能够对胎儿分数求解上式。也许,求解胎儿分数的最简易方法是通过阶乘矩方法,其中混合参数能够根据矩(其能够易于从所观测数据来估计)来表达。

[0192] 给定n个SNP位置,阶乘矩定义如下:

[0193] $$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{d_i}$$

[0194] $$F_2=\frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)}{d_i(d_i-1)}$$

[0195] …

[0196] $$F_j=\frac{1}{n}\sum_{i=1}^{n}\frac{a_i(a_i-1)\cdots(a_i-j+1)}{d_i(d_i-1)(d_i-j+1)}$$

[0197] 等式4

[0198] 阶乘矩能够与{αi,pi}相关。

[0199] $$F_1 \approx \sum_{i=1}^{m} \alpha_i p_i^1$$

[0200] $$F_2 \approx \sum_{i=1}^{m} \alpha_i p_i^2$$

[0201] …

[0202] $$F_j \approx \sum_{i=1}^{m} \alpha_i p_i^j$$

[0203] …

[0204] $$F_g \approx \sum_{i=1}^{m} \alpha_i p_i^g$$

[0205] 等式5

[0206] 能够通过在n＞2*(待估计的参数数量)时求解从以上关系式5所得出的等式系统中的{αi,pi},来识别解。显然,对于更高的g,问题在数学上变得更为困难,因为更多{αi,pi}需要估计。

[0207] 通常不可能通过在较低胎儿分数的简单阈值来准确地区分情况1与2(或者情况3与4)数据。幸亏对于简化情况模型的使用,情况1/2数据易于通过在点(2A/D)＝T的区分与情况3/4分离。T＝0.5的使用已经发现能够令人满意地执行。

[0208] 注意,采用等式4和5的混合模型方法利用所有多态性的数据,但是没有单独考虑测序误差。将第一和第二情况的数据与第三和第四情况的数据分离的适当方法能够考虑测序误差。

[0209] 在其它示例中,提供给混合模型的数据集仅包含情况1和情况2多态性的数据。这些是母亲对其为纯合的多态性。阈值技术可用于去除情况3和4多态性。例如,在采用混合模型之前消除具有大于特定阈值的少数等位基因频率的多态性。使用如简化为等式7和8的适当过滤的数据和阶乘矩,可计算胎儿分数f,如等式9所示。注意,等式7是等式3对于混合模型的这个实现的重述。还要注意,在这个具体示例中,与机器读取关联的测序误差不是已知的。因此,等式系统必须单独求解误差e。

[0210] 图7示出使用这个混合模型以及已知胎儿分数(x轴)和估计胎儿分数的结果的比较。如果混合模型完全预测胎儿分数,则绘制结果跟随虚线。然而,所估计分数非常好,特别是考虑到在应用混合模型之前消除了许多数据。

[0211] 为了进一步阐述,若干其它方法可用于来自等式3的模型的参数估计。在一些情况下,能够通过将导数设置成卡方统计的零,来查找易处理的解。在不能通过直接微分来查找简易解的情况下,二项式PDF或者其它近似多项式的泰勒级数展开能够是有效的。最小卡方估计量众所周知是有效的。

[0212] $$\chi^2(\alpha_i, p_i) = \sum_{i=1}^{n} \frac{\left(P_i - \sum \alpha_i Binomial(p_i, di)\right)^2}{Binomial(n, p)}$$

[0213] 等式6

[0214] 其中,Pi是计数i的点数。来自Le Cam的备选方法["On the Asymptotic Theory of Estimation and Testing Hypotheses",Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,Volume 1,Berkeley CA:University of CA Press,1956,第129-156页]使用似然函数的拉尔夫-牛顿迭代。来自等式5的矩解的方法能够用作迭代的起始点。

[0215] 在另一个应用下,论述解析对近似β分布的混合进行操作的预计最大化方法的混合模型的方法。

[0216] 模型情况(1+2),测序误差未知

[0217] 考虑简化模型,其仅考虑杂合性情况1和2。在这种情况下,混合分布能够写作

[0218] $A=\{a_i\}\sim\alpha_1 Bin(e,d_i)+\alpha_2 Bin(f/2,d_i)$

[0219] 其中

[0220] $1=\alpha_1+\alpha_2$

[0221] $m=4$

[0222] 等式7

[0223] 以及系统

[0224] $F_1=\alpha_1 e+(1-\alpha_1)(f/2)$

[0225] $F_2=\alpha_1 e^2+(1-\alpha_1)(f/2)^2$

[0226] $F_3=\alpha_1 e^3+(1-\alpha_1)(f/2)^3$

[0227] 等式8

[0228] 对于e(测序误差率)、α(情况1点的比例)和f(胎儿分数)来求解。其中,Fi如上式4中所定义。胎儿分数的闭式解选择为下式的实解

[0229] $$F\approx\frac{(F1-1)F2\pm\sqrt{F2}\sqrt{4F1^3+F2-3F1(2+F1)F2+4F2^2}}{2(F1^2-F2)}$$

[0230] 等式9

[0231] 其在0与1之间。

[0232] 为了计量估计量的性能,哈迪-温伯格平衡点(ai,di)的模拟数据集以设计为{1%、3%、5%、10%、15%、20%和25%}的胎儿分数和1%的恒定测序误差率来构成。1%误差率对于正使用的测序机器和协议是当前接受的误差率,并且与以上在图3中所示的Illumina Genome分析器II数据的图表一致。等式9应用于数据,并且除了四点向上偏差置之外,查找与"已知"胎儿分数的一般协议。感兴趣的是,测序误差率e估计为刚好高于1%。

[0233] 在下一个混合模型示例中,阈值或者其它过滤技术再次用于去除落入情况3和4中的多态性的数据。但是,在这种情况下,测序误差为已知。这简化胎儿DNA分数f的所产生表达式,如等式10所示。图8示出,与随等式9所采用的方式相比,这种形式的混合模型提供改进结果。

[0234] 类似方式在等式11和12中示出。这种方式认识到,只有某些测序误差增加少数等位基因计数。在每四个测序误差中而是只有一个应当增加少数等位基因计数。图9示出使用这种技术的实际与估计胎儿分数之间的非常好的协议。

[0235] 模型情况(1+2),测序误差已知

[0236] 由于所使用的机器的测序误差率在很大程度上为已知，所以计算的偏置和复杂度能够通过消除作为待求解变量的e来降低。因此，得到等式系统

[0237] $F_1=\alpha_1 e+(1-\alpha_1)(f/2)$

[0238] $F_2=\alpha_1 e^2+(1-\alpha_1)(f/2)^2$

[0239] 等式10

[0240] 对于胎儿分数f，得到解：

[0241] $$F\approx\frac{2(eF1-F2)}{(e-F1)}$$

[0242] 图8示出使用机器误差率作为已知参数将上偏差降低某个点。

[0243] 模型情况(1+2)，测序误差已知，改进的误差模型

[0244] 为了改善模型中的偏置，将以上等式的误差模型扩大为考虑如下事实：不是每一个测序误差事件将增加杂合性情况1中的少数等位基因计数A=ai。此外，允许如下事实：测序误差事件可促进杂合性情况2计数。因此，通过求解以下阶乘矩关系系统来确定胎儿分数F：

[0245] $F_1=\alpha_1 e/4+(1-\alpha_1)(e+f/2)$

[0246] $$F_2=\alpha_1\left(\frac{e}{4}\right)^2+(1-\alpha_1)(e+f/2)^2$$

[0247] 等式11

[0248] 其产生解

[0249] $$F\approx\frac{-2(e^2-5eF1+4F2)}{(e-4F1)}$$

[0250] 等式12

[0251] 图9中示出使用机器误差率作为已知参数的模拟数据，增强情况1和2误差模型将上偏差极大地降低到小于低于0.2的胎儿分数的某个点。

[0252] 实现选项

[0253] 样本

[0254] 在本文所公开的实施例中使用的样本包括基因组DNA，其是细胞的或者游离的。细胞DNA通过从相同或者不同遗传组成的全体细胞手动或机械提取基因组DNA，从全体细胞来得出。细胞DNA能够例如从相同遗传组成(其从一个受检者得出)的全体细胞、从不同受检者的全体细胞的混合或者从遗传组成中不同的全体细胞(其从一个受检者得出)的混合来得出。用于从全体细胞提取基因组DNA的方法是本领域已知的，并且根据来源性质而有所不同。

[0255] 在一些情况下，对细胞基因组DNA分片会是有利的。分片是能够是随机的，或者它能够是特定的，例如，正如使用限制内切酶酶切所取得的那样。用于随机分片的方法是本领域众所周知的，并且包括例如限制DNA酶酶切、碱处理和物理剪切。在某些实施例中，样本核酸经过分片为大约500或更多碱基对的片断，以及能够易于对其应用下一代测序(NGS)方法。在一个实施例中，样本核酸从如cfDNA(其没有经过分片)来得到。

[0256] 游离DNA是基因组DNA，其作为通常见于受检者的生物流体、如血液中的基因组片断的混合自然出现。基因组混合物能够从细胞(其通过生物过程、例如细胞凋亡自然破裂以

翻译其基因组含量)来得出。cfDNA的样本能够包括从相同种类的不同受检者的细胞的混合、从一个受检者的细胞(其遗传组成有所不同)的混合或者从不同种类、例如受检者的细胞的混合所得出的cfDNA。

[0257] 包括游离DNA的游离核酸能够通过本领域已知的各种方法从生物样本(包括但不限于血浆、血清和尿液)来得到(Fan等人,ProcNatlAcadSci 105:16266-16271[2008];Koide等人,Prenatal Diagnosis 25:604-607[2005];Chen等人,Nature Med.2:1033-1035[1996];Lo等人,Lancet 350:485-487[1997];Botezatu等人,Clin Chem.46:1078-1084,2000;以及Su等人,J Mol.Diagn.6:101-107[2004])。为了将cfDNA与细胞分离,能够使用分馏、离心法(例如密度梯度离心法)、DNA特定沉淀或者高吞吐量细胞分类和/或分离方法。用于cfDNA的手动和自动分离市场销售套装是可得到的(Roche Diagnostics,Indianapolis,IN,Qiagen,Valencia,CA,Macherey-Nagel,Duren,DE)。

[0258] 包括核酸的混合的样本(对其应用本文所述的方法)可以是生物样本,例如组织样本、生物流体样本或者细胞样本。在一些实施例中,核酸的混合通过已知方法的任一个来净化或者与生物样本隔离。样本能够是净化或者隔离多核苷酸。作为非限制性示例,生物流体包括血液、血浆、血清、汗液、泪水、唾液、尿液、唾液、耳液、淋巴、痰、脑脊髓液、ravage、骨髓悬浮液、阴道液、宫颈灌洗液、脑液、腹水、乳剂、呼吸分泌物、肠和生殖泌尿道、羊水和leukophoresis样本。在一些实施例中,样本是通过无创过程易于得到的样本,例如血液、血浆、血清、汗液、泪水、唾液、尿液、唾液、耳液、痰或粪便。优选地,生物样本是周边血液样本或者血浆和血清分数。在其它实施例中,生物样本是药签或涂片、活检标本或者细胞培养。在另一个实施例中,样本是两个或更多生物样本的组合,例如,生物样本能够包括生物流体样本、组织样本和细胞培养样本中的两个或更多。如本文所使用的术语“血液”、“血浆”和“血清”明确包含其分数或者经处理的部分。类似地,在样本从活检、药签、涂片等获取的情况下,“样本”明确包含从活检、药签、涂片等所得出的经处理分数或部分。

[0259] 在一些实施例中,样本能够从源得到,包括但不限于来自不同个体的样本、相同或不同个体的不同发展阶段、不同疾病个体(例如具有癌症或者疑似具有遗传障碍的个体)、正常个体的样本、在个体的疾病的不同阶段所得到的样本、从经受不同的疾病治疗的个体所得到的样本、来自经受不同环境因素的个体或者具有病理诱因的个体或者暴露于传染病病原(例如HIV)的样本。

[0260] 在一个实施例中,样本是母体样本,其从怀孕女性、例如孕妇来得到。在这种情况下,样本能够使用本文所述方法来分析,以便提供胎儿的潜在染色体异常的产前检查。母体样本能够是组织样本、生物流体样本或者细胞样本。作为非限制性示例,生物流体包括血液、血浆、血清、汗液、泪水、唾液、尿液、唾液、耳液、淋巴、痰、脑脊髓液、ravage、骨髓悬浮液、阴道液、宫颈灌洗液、脑液、腹水、乳剂、呼吸分泌物、肠和生殖泌尿道和leukophoresis样本。在另一个实施例中,母体样本是两个或更多生物样本的组合,例如,生物样本能够包括生物流体样本、组织样本和细胞培养样本中的两个或更多。在一些实施例中,样本是通过无创过程易于得到的样本,例如血液、血浆、血清、汗液、泪水、唾液、尿液、唾液、耳液、痰和粪便。在一些实施例中,生物样本是周边血液样本或者血浆和血清分数。在其它实施例中,生物样本是药签或涂片、活检标本或者细胞培养。

[0261] 样本还能够从体外培养组织、细胞或者其它含多核苷酸源来得到。培养样本能够

从源来获取，包括但不限于在不同介质和条件(例如pH、压力或温度)中保持的培养(例如组织或细胞)、对于不同的长度周期所保持的培养(例如组织或细胞)、采用不同因子或试剂(例如药物候选或者调制因子)所处理的培养(例如组织或细胞)或者不同类型的组织或细胞的培养。将核酸与生物源隔离的方法是众所周知的，并且将根据来源的性质而有所不同，如上所述。

[0262] 识别基因组分数中使用的多态性

[0263] 如所述，多态性可用于评估胎儿分数。在评估中使用一个或多个多态性的等位基因分数和接合性。有用多态性的示例包括但不限于单核苷酸多态性(SNP)、串联SNP、小规模多碱基缺失或插入、称作IN-DELS(又称作缺失插入多态性或DIP)、多核苷酸多态性(MNP)、短串联重复(STR)、限制片断长度多态性(RFLP)、包括微缺失的缺失、包括微插入的插入、重复、转化、易位、繁殖、复合多位点变异、副本数量变化(CNV)以及包括染色体中的序列的任何其它变化的多态性。

[0264] 在一些实施例中，所公开方法中使用的多态性包括SNP和/或STR。SNP多态性能够是单SNR、串联SNP。单SNP包括个体SNP和标签SNP、即单倍型中存在的SNP和/或单倍型域。在一些实施例中，使用多态性的组合。例如，副本数量的差能够通过包括一个或多个SNP和一个或多个STR的多态序列的组合的比较来检测。

[0265] 一般来说，任何多态位点(其能够由本文所述的测序方法所生成的读数来包含)能够用于识别包括不同基因组的DNA的样本中的基因组分数。对于实施本发明的方法有用的多态序列是从多种公共可访问数据库(其持续扩大)可得到的。例如，有用数据库非限制性地包括在万维网地址wi.mit.edu的Human SNP Database、在万维网地址ncbi.nlm.nih.gov、万维网地址lifesciences.perkinelmer.com的NCBI dbSNP Home Page、在万维网地址celera.com的Celera Human SNP数据库、在万维网地址gan.iarc.fr的SNP Database of the Genome Analysis Group(GAN)、在万维网地址atcc.org的ATCC短串联重复(STR)数据库以及在万维网地址hapmap.org的HapMap数据库。

[0266] 能够在胎儿分数评估中使用的多态性的数量能够是至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、200、300、400、500、600、700、800、900、1000或以上。例如，估计人类基因组包括至少大约1千万个SNP。因此，在来自人类受检者的样本中能够基因分型的可用多态性的数量能够是至少大约1千万个SNP以及任何一个人类基因组中存在的许多其它类型的多态性。在一些实施例中，包括第一和第二基因组的DNA、例如cfDNA的混合的样本的第一基因组中的一个或多个多态性的识别通过使用如本文所述的NGS方法的全基因组测序来执行。在一些实施例中，全基因组测序方法是NGS方法，其通过对无性系放大核酸分子进行大规模平行测序或者通过单核酸分子的大规模平行测序、即单分子测序来识别多态序列。

[0267] 应用

[0268] 源自样本中的两个不同基因组源的每个的核酸的分数能够用于各种目的。在本文所述的各个实施例中，母体样本的游离DNA中的胎儿DNA的分数用于促进产前检查，以及帮助进行与妊娠的治疗有关的判定。在其它实施例中，所考虑的基因组不是母体和胎儿。下面提供用于确定分数基因组存在的基因组源的各个示例。

[0269] 母体血液中循环的游离胎儿DNA和RNA能够用于增加数量的条件的早期无创产前检查(NIPD),用于妊娠管理以及帮助生殖决策。少量循环胎儿DNA在妊娠期间存在于母体血流中(Lo等人,Lancet 350:485-487[1997])。被认为源自正死亡胎盘细胞,游离胎儿DNA已经被表明由长度通常少于200bp的短片断组成(Chan等人,ClinChem 50:88-92[2004]),其能够早至4周受孕被辩明(Illanes等人,Early Human Dev 83:563-566[2007]),并且已知为在分娩的数小时之内从母体循环中清除(Lo等人,Am J Hum Genet 64:218-224[1999])。除了cfDNA之外,游离胎儿RNA(cfRNA)的片断也能够在源自胎儿或胎盘中转录的基因的母体血流中辩明。这些胎儿遗传元素从母体血液样本的提取和后续分析为NIPD提供新时机。

[0270] 如所述,所公开方法确定生物样本中的第二基因组的分数。方法可选地确定包括第一和第二基因组的DNA(例如cfDNA)的混合的血液样本中的多个障碍的存在或不存在。在一些实施例中,胎儿分数的确定可包括:(a)对cfDNA的混合的至少一部分进行基因组测序,以便得到多个序列标签;(b)确定多个序列标签中的多个多态性的存在或不存在;以及(c)将多个多态性与混合中的第一和/或第二基因组关联。在优选实施例中,混合对于多个多态性未强化。通过将全基因组测序方法所得到的映射标签的序列与多个参考多态性进行比较,来执行DNA的混合中的多个多态性的识别,如本文所述。

[0271] 在以上所公开的实施例中,第一基因组是胎儿基因组,以及第二基因组是母体基因组。在另一个实施例中,第一基因组是未受影响细胞的基因组,以及第二基因组是来自受影响细胞、例如癌细胞的基因组。在一些实施例中,受影响和未受影响细胞从同一受检者得出。例如,受影响细胞能够是其基因组已被障碍改变的细胞。在一些实施例中,障碍是单基因障碍。在其它实施例中,障碍是多基因障碍。障碍能够通过单个多态性、例如标签SNP或者通过单倍型中存在的多个多态性来识别。在一些实施例中,按照本方法所识别的多个多态性存在于单倍型域中。

[0272] 能够借助于本方法来识别的障碍是遗传障碍,其是至少部分由基因或染色体中的异常所引起的疾病。样本中的胎儿分数的知识能够帮助识别产前背景中的这类障碍。通过本方法所识别的障碍包括单基因、即单个基因病以及多基因、即复合病。单个基因病包括常染色体显性、常染色体隐性、X连锁显性、X连锁隐性和Y连锁。

[0273] 在常染色体显性病中,只有基因的一个突变副本将是受障碍影响的人所需的。通常,受影响受检者具有一个受影响双亲(之一),以及存在后代将继承突变基因的50%机会。作为常染色体显性的条件有时具有降低的外显率,其意味着,虽然仅需要一个突变副本,但是并非继承那个突变的全部个体继续发展该疾病。能够通过本发明来识别的常染色体显性病的示例非限制性地包括家族性高胆固醇血症、遗传性球形红细胞性贫血、马方氏综合症、1型神经纤维瘤病、遗传性非息肉病性结直肠癌以及遗传性多发性外生骨疣和亨廷顿病。

[0274] 使用本方法所检测的常染色体隐性病包括镰状细胞性贫血、囊肿性纤维化、泰氏-萨氏病、泰氏-萨氏病、黏多醣症、糖原过多症和半乳糖血。由本方法所检测的X连锁病包括迪谢内肌营养不良和血友病。在常染色体隐性病中,基因的两个副本必须对于受检者突变,以便受常染色体隐性病影响。受影响受检者通常具有各携带突变基因的单个副本的未受影响双亲(并且称作携带者)。各携带突变基因的一个副本的两个未受影响的人对各妊娠有25%机会具有被障碍影响的孩子。能够通过本方法来识别的这种类型的障碍的示例包括囊肿性纤维化、镰状红细胞疾病、泰氏-萨氏病、尼曼-皮克病、脊髓性肌肉萎缩症和罗伯茨综

合症。某些其它显型、例如湿与干耳垢也按照常染色体隐性方式来确定。X连锁显性病通过X染色体上的基因的突变来引起。只有几个障碍具有这种遗传模式，其中最典型示例是X连锁低磷酸盐血症性佝偻病。男性和女性均在这些障碍中受到影响，其中男性通常比女性受到更严重影响。一些X连锁显性条件、例如Rett综合症、2型色素失调症和艾卡尔迪综合症通常在男性中是致命的，并且因此主要在女性中看到。除了对这个发现的例外是极罕见情况，其中具有克氏综合症(47，XXY)的男孩也继承X连锁显性条件情况，并且在疾病严重性方面呈现与女性更相似的症状。传递X连锁显性病的机会在男人与妇女之间有所不同。具有X连锁显性病的男人的儿子将全部不受影响(因为他们接收其父亲的Y染色体)，以及其女儿将全部继承该条件。具有X连锁显性病的妇女对各妊娠有50％机会具有受影响胎儿，但是应当注意，在例如色素失调等情况下，只有女性后代一般是能生存的。另外，虽然这些条件没有改变出生率本身，但是具有Rett综合症或艾卡尔迪综合症的个体很少生殖。

[0275] 本方法还能够促进与X连锁病关联的多态性的识别。X连锁隐性条件也通过X染色体上的基因的突变来引起。男性比女性更频繁地受影响，以及传递障碍的机会在男人与妇女之间有所不同。具有X连锁隐性病的男人的儿子将不受影响，以及其女儿将携带突变基因的一个副本。作为X连锁隐性病(X^RX^T)的携带者的妇女有50％机会具有受影响的儿子以及50％机会具有携带突变基因的一个副本并且因此是携带者的女儿。X连锁隐性条件非限制性地包括严重疾病血友病A、迪谢内肌营养不良和雷-纳综合症以及诸如男性斑秃和红绿色盲之类的不太严重的常见条件。X连锁隐性条件因偏斜X灭活或者单体性X(特纳综合症)而在女性中有时能够明显。

[0276] Y连锁病通过Y染色体上的突变引起。因为男性从其父亲继承Y染色体，所以受影响父亲的每一个儿子将受影响。因为女性从其父亲继承X染色体，所以受影响父亲的女性后代从不受影响。由于Y染色体较小并且包含极少基因，所以存在较少Y连锁病。综合症常常包括不育，其可借助于某些生育力治疗来防止。示例是男性不育和羽片毛症。

[0277] 如所述，用于检测样本中的基因组分数的所公开方法能够用于促进从物质样本来检测异倍性。在一些实施例中，异倍性是完整染色体三体性或单体性或者部分三体性或单体性。部分异倍性通过染色体的部分的损失或增益，并且包含产生于不平衡易位、不平衡转化、缺失和插入的染色体不平衡。到目前为止，与生命相容的最常见异倍性是三倍性21、即唐氏综合症(DS)，其通过染色体21的部分或全部的存在引起。DS很少能够通过继承或偶发缺陷引起，由此染色体21的全部或部分的额外副本变为附连到另一个染色体(通常为染色体14)，以便形成单个畸变染色体。DS与智能缺陷、严重学习困难和长期健康问题、例如心脏病所引起的过度死亡率关联。具有已知临床重要性的其它异倍性包括爱德华氏综合症(三体性18)和帕套综合症(三体性13)，其在生命的前几个月经常是致命的。与性染色体的数量关联的异常也是已知的，并且包括女性出生中的例如特纳综合症(XO)等单体性X和三X综合症(XXX)以及男性出生中的克兰费尔特综合症(XXY)和XYY综合症，其全部与包括不育和智力技能的降低的各种显型关联。单体X[45，X]是说明大约7％的自然流产的早期妊娠丢失的常见原因。基于1-2/10000的45，X(又称作特纳综合症)的活产频率，估计少于1％的45，X胚胎将存活。大约30％的特纳综合症患者具有45，X细胞系以及45，XX细胞系或者包含重新排列X染色体的细胞系的镶嵌(Hook and Warburton 1983)。考虑到高胚胎致死性，活产婴儿中的显型比较缓和，并且已经假设具有特纳综合症的可能全部活产女性携带包含两个性染

色体的细胞系。单体性X能够在女性中作为45,X或者作为45,X/46XX出现,以及在男性中作为45,X/46XY出现。一般表明人类的常染色体单体性与生命不相容;但是,存在相当数量的细胞生成报告,描述在活产儿中的一个染色体21的全单体性(Vosranoval等人,Molecular Cytogen.1:13[2008];Joosten等人,Prenatal Diagn.17:271-5[1997])。本发明的方法能够用于在产前诊断这些和其它染色体异常。

[0278] 按照一些实施例,胎儿分数在确定染色体1-22、X和Y的任一个的染色体三体性的存在或不存在中能够是有用的。按照本方法能够检测的染色体三体性的示例非限制性地包括三体性21(T21;唐氏症)、三体性18(T18;爱德华氏综合症)、三体性16(T16)、硬体性20(T20)、三体性22(T22;猫眼综合症)、三体性15(T15;PraderWilli综合症)、三体性13(T13;帕套综合症)、三体性8(T8;Warkany综合症)、三体性9以及XXY(克兰费尔特综合症)、XYY或XXX三体性。存在于非镶嵌状态中的其它常染色体的完整三体性是致死因子,但是当存在于镶嵌状态时能够与生命相容。将会理解,无论存在于镶嵌还是非镶嵌状态中,各种完整三体性以及部分三体性能够按照本发明的理论在胎儿cfDNA中确定。

[0279] 能够通过本方法来确定的部分三体性的非限制性示例包括但不限于部分三体性1q32-44、三体性9p、三体性4镶嵌性、三体性17p、部分三体性4q26-qter、部分2p三体性、部分三体性1q和/或部分三体性6p/单体性6q。

[0280] 本文所公开的方法还能够用于帮助确定染色体单体性X、染色体单体性21和部分单体性,例如单体性13、单体性15、单体性16、单体性21和单体性22,其已知为涉及在妊娠流产中。通常涉及在完整异倍性中的染色体的部分单体性也能够通过本发明的方法来确定。能够按照本方法来确定的缺失综合症的非限制性示例包括通过染色体的部分缺失所引起的综合症。能够按照本发明的方法来确定的部分缺失的示例非限制性地包括染色体1、4、5、7、11、18、15、13、17、22和10的部分缺失,下面对其进行描述。

[0281] 1q21.1缺失综合症或者1q21.1(复发)微缺失是染色体1的罕见畸变。接着缺失综合症,还存在1q21.1重复综合症。虽然存在具有特定斑点上缺失综合症的DNA丢失的一部分,但是存在具有重复综合症的相同斑点上的DNA的相似部分的两个或三个副本。文献将缺失和重复称作1q21.1副本数量变化(CNV)。1q21.1缺失能够与TAR综合症(血小板减少伴桡骨缺失)关联。

[0282] Wolf-Hirschhom综合症(WHS)(OMIN#194190)是与染色体4p16.3的半合子缺失关联的邻近基因缺失综合症。Wolf-Hirschhom综合症是生性畸形综合症,其特征在于产前和产后生长不足、可变程度的发育性残疾、颅面特征(鼻子的'希腊战士头盔'外观、高额头、突出印堂、器官距离过远、高拱眉毛、凸眼、内眦赘皮、短小人中、带下嘴角的独立嘴巴和小颌)和癫痫。

[0283] 又称作5p-或5p减并且命名为猫叫综合症(OMIN#123450)的染色体5的部分缺失通过染色体5(5p15.3-p15.2)的短臂(p臂)的缺失引起。具有这种条件的婴儿常常具有听起来像猫一样的高音调哭声。该障碍的特征在于智力残疾和延迟发育、小头尺寸(小头畸形)、低出生体重和幼年的弱肌张力(张力减退)、与众不同的面部特征以及可能的心脏缺陷。

[0284] 又称作染色体7q11.23缺失综合症(OMIN 194050)的威-贝二氏综合症是产生于多系统障碍(其由包含大致28个基因的染色体7q11.23上的1.5至1.8Mb的半合子缺失所引起)的邻近基因缺失综合症。

[0285] 又称作11q缺失障碍的雅各布森综合症是产生于包括带11q24.1的染色体11的末端区的缺失的罕见先天障碍。它能够引起智力残疾、与众不同的面部外观以及多种身体问题，包括心脏缺陷和出血障碍。

[0286] 称作单体性18p的染色体18的部分单体性是罕见染色体病，其中染色体18的短臂(p)的全部或部分缺失(单体)。障碍的特征通常在于短身材、可变程度的智能迟钝、言语迟缓、头骨和面部(颅面)区域的畸形和/或附加身体异常。关联的颅面缺陷在范围和严重性方面逐个情况极大地改变。

[0287] 染色体15的副本的结构和数量的变化所引起的条件包括安格曼综合症和Prader-Willi综合症，其涉及染色体15、15q11-q13区域的相同部分中的基因活动的丢失。将会理解，若干易位和微缺失在携带者双亲中会是无症状的，但是仍然能够引起后代的主要遗传疾病。例如，携带15q11-q13微缺失的健康母亲能够生出具有安格曼综合症、严重神经变性病的孩子。因此，本发明能够用于识别胎儿中的这种部分缺失和其它缺失。

[0288] 部分单体性13q是罕见染色体病，其在染色体13的一段长臂(q)丢失(单体)时产生。天生具有部分单体性13q的婴儿可呈现低出生体重、状况和面部(卢布面区域)的畸形、骨骼异常(特别是手和脚)和其它身体异常。智力迟钝是这种条件的特性。幼年期间的死亡率在天生具有这种障碍的个体之中很高。部分单体性13q的几乎所有情况没有明显原因(偶发)而随机发生。

[0289] 司马综合症(SMS-OMIM#182290)通过染色体17的一个副本上的遗传物质的缺失或丢失引起。这种众所周知的综合症与发育延迟、智力迟钝、诸如心脏和肝脏缺陷之类的先天性异常以及诸如严重睡眠障碍和自伤行为之类的神经行为异常关联。司马综合症(SMS)在大多数(90％)情况下通过染色体17p11.2中的3.7-Mb中间缺失引起。

[0290] 又称作迪格奥尔格综合症的22q11.2缺失综合症是通过一小段染色体的缺失引起。缺失(22q11.2)在染色体对之一的长臂上的染色体的中间附近发生。这种综合症的特征广泛地改变，甚至在同一家庭的成员之中，并且影响身体的许多部分。特性征兆和症状可包括诸如先天性心脏病之类的出生缺陷、与对于闭合的神经肌肉问题最通常相关的上颚中的缺陷(腭咽关闭不全)、学习障碍、面部特征的轻微差异和复发传染。染色体区域22q11.2中的微缺失与精神分裂症的20至30倍增加风险关联。

[0291] 染色体10的短臂上的缺失与例如显形等迪格奥尔格综合症关联。染色体10p的部分单体性是罕见的，但是在呈现迪格奥尔格综合症的特征的患者的一部分中已经观测到。

[0292] 在一个实施例中，本发明的方法用于确定部分单体性，包括但不限于染色体1、4、5、7、11、18、15、13、17、22和10的部分单体性，例如部分单体性1q21.11、部分单体性rp16.3、部分单体性5p15.3-p15.2、部分单体性7q11.23、部分单体性11q24.1、部分单体性18p、染色体15的部分单体性(15q11-q13)、部分单体性13q、部分单体性17p11.2、染色体22的部分单体性(22q11.2)以及部分单体性10p也能够使用该方法来确定。

[0293] 能够按照本发明的方法来确定的其它部分单体性包括不平衡易位t(8;11)(p23.2;p15.5)；11q23微缺失；17p11.2缺失；22q13.3缺失；Xp22.3微缺失；10p14缺失；20p微缺失，[del(22)(q11.2q11.23)]，7q11.23和7q36缺失；1p36缺失；2p微缺失；1型神经纤维瘤病(17q11.2微缺失)、Yq缺失；4p16.3微缺失；1p36.2微缺失；11q14缺失；19q13.2微缺失；鲁宾斯坦-泰比(16p13.3微缺失)；7p21微缺失；米勒-狄克氏综合症(17p13.3)；以及2q37微

缺失。部分缺失能够是染色体部分的小缺失，或者它们能够是染色体的微缺失，其中单个基因的缺失能够发生。

[0294] 已经识别通过染色体臂的部分的重复所引起的若干重复综合症(参见OMIN[Online Mendelian Inheritance in Man，在ncbi.nlm.nih.gov/omim在线查看])。在一个实施例中，本方法能够用于确定染色体1-22、X和Y的任一个的段的重复和/或繁殖的存在或不存在。能够按照本方法来确定的重复综合症的非限制性示例包括染色体8、15、12和17的部分的重复，下面对其进行描述。

[0295] 8p23.1重复综合症是通过来自人类染色体8的区域的重复所引起的罕见遗传病。这种重复综合症在64000个出生中具有1的估计患病率，并且是8p23.1缺失综合症的倒数。8p23.1重复与可变显型关联，包括言语迟缓、具有突出额头和弓形眉毛的轻微畸形以及先天性心脏病(CHD)。

[0296] 染色体15q重复综合症(Dup15q)是临床可识别综合症，其产生于染色体15q11-13.1的重复。具有Dup15q的婴儿通常具有张力减退(不良肌张力)、生长迟缓；他们可能天生具有兔唇和/或上颚或者心脏、肝脏或其它器官的畸形；他们呈现某种程度的认知迟缓/残疾(智力迟钝)、言语和语言迟缓以及感觉处理障碍。

[0297] 帕-克综合症是额外#12染色体物质的结果。通常存在细胞的混合(镶嵌)，部分具有额外#12物质，以及部分正常(没有额外#12物质的46个染色体)。具有这种综合症的婴儿具有许多问题，包括严重智力迟钝、不良肌张力、“粗糙”面部特征以及突出额头。他们趋向于具有极薄的上唇，其中具有较厚的下唇和短鼻。其它健康问题包括癫痫、不良进食、僵硬关节、成人期的白内障、听力损失以及心脏缺陷。具有帕-克的人具有缩短的寿命。

[0298] 具有指定为dup(17)(p11.2p11.2)或dup 17p的遗传条件的个体在染色体17的短臂上携带额外遗传信息(称为重复)。染色体17p11.2的重复成为Potocki-Lupski综合症(PTLS)，其是新认识的遗传条件，在医疗文献上仅报道数十例。具有这种重复的患者常常具有低肌张力、不良进食以及在幼年无法茁壮成长，并且还有运动和言语转折点的迟缓发育。具有PTLS的许多个体具有发音和语言处理的困难。另外，患者可具有与在具有自闭症或者自闭症谱系病的人中所看到的相似的行为特性。具有PTLS的个体可具有心脏缺陷和睡眠窒息。包括基因PMP22的染色体17p12中的大区域的重复已知为引起夏玛丽牙疾病。

[0299] CNV与死产关联。但是，由于常规细胞遗传学的固有限制，CNV对死产的贡献被认为未被充分表示(Harris等人，PrenatalDiagn 31:932-944[2011])。方法在帮助确定部分异倍性的存在、例如染色体段的缺失和繁殖中是有用的，并且能够用于帮助识别和确定与死产关联的CNV的存在或不存在。

[0300] 本方法还能够帮助识别与遗传病关联的多态性，其是复合、多因子或者多基因的，表示它们可能与结合生活方式和环境因素的多个基因的作用关联。多因子障碍包括例如心脏病和糖尿病。虽然复合障碍常常群集在家族中，但是它们没有遗传的鲜明模式。关于血统，多基因病趋向于“在家族中流传”，但是遗传并不是如同孟德尔疾病那么简单。强环境成分与许多复合障碍、例如血压关联。本方法能够用于识别多态性，其与多基因病关联，包括但不限于哮喘、自身免疫性疾病、例如多种硬化、癌症、纤毛、颚裂、糖尿病、心脏病、高血压、发炎性肠道疾病、智力迟钝、心境障碍、肥胖、屈光不正和不育。在一些实施例中，多态性是SNP。在其它实施例中，多态性是STR。在又一些实施例中，多态性是SNP和STR的组合。

[0301] 在一个实施例中，与障碍关联的多态序列的识别包括对于与cfDNA的混合中的第二基因组对应的细胞基因组的至少一部分进行测序。通过确定在包含实质上仅从第二基因组所得出的DNA分子的第一样本中的多个多态位点的序列，确定在包含从第一和第二基因组所得出的DNA分子的混合的第二样本中的对应多个多态位点的序列，并且比较在两种样本中确定的多态序列，由此识别包括两个基因组的混合的样本的第一基因组中的多个多态性，来执行由第一基因组所贡献的多态序列的识别。例如，通过确定在母体血沉棕黄层样本、即包含实质上仅从第二基因组所得出的DNA分子的样本中的多个多态位点的序列，确定在净化血浆样本、即包含从胎儿和母体基因组所得出的cfDNA分子的混合的第二样本中的对应多个多态位点的序列，并且比较在两种样本中确定的多态序列以识别多个胎儿多态性，来执行由胎儿基因组、即第一基因组所贡献的多态序列的识别。在一个实施例中，第一基因组是胎儿基因组，以及第二基因组是母体基因组。在另一个实施例中，第一基因组是未受影响细胞的基因组，以及第二基因组是来自受影响细胞的基因组。在一些实施例中，受影响和未受影响细胞从同一受检者得出。例如，受影响细胞能够是其基因组已被障碍改变的细胞。

[0302] 在一个实施例中，估计基因组分数的所公开方法帮助检测患者体内的癌症。在各个示例中，癌症通过包括下列步骤的方法来检测：从患者提供包括从正常、即未受影响和癌症、即受影响细胞所得出的基因组的混合的样本；以及识别与癌症关联的多个多态性。在一些实施例中，样本从血液、血浆、血清和尿液中选取。在一些实施例中，样本是血浆样本。在其它实施例中，样本是尿液样本。

[0303] 在一个实施例中，识别与癌症关联的多个多态性包括增强多态目标序列的样本中的DNA。在其它实施例中，没有执行多态目标序列的样本的增强。在一些实施例中，识别与癌症关联的多个多态性包括量化多态序列的副本的数量。

[0304] 能够按照本发明的方法来识别和/或监测的癌症包括实体瘤以及血液肿瘤和/或恶性肿瘤。待治疗的各种癌症包括肉瘤、癌瘤和腺癌，并不局限于乳癌、肺癌、结直肠癌、胰腺癌、卵巢癌、前列腺癌、肾癌、肝细胞瘤、胖子癌、黑素瘤、多发性骨髓瘤、淋巴瘤、霍奇金淋巴瘤、非霍奇金淋巴瘤、儿童淋巴瘤以及淋巴球和皮肤起源淋巴瘤、白血病、儿童白血病、毛细胞白血病、急性淋巴球白血病、急性髓细胞白血病、慢性淋巴球白血病、慢性髓细胞白血病、慢性骨髓白血病以及肥大细胞白血病、髓样肿瘤、肥大细胞肿瘤、血液肿瘤以及淋巴肿瘤，包括远离初生肿瘤位点的其它组织或器官中的新陈代谢病变。

[0305] 本发明的方法例如在诊断或确定已知为与特定单倍型关联的疾病条件中的预后(prognosis)方面是有用的，以便确定新单倍型以及检测与对药物的响应性的单倍型关联。多个多态序列与多个障碍的关联能够从多个障碍的每个的单个多态序列的识别码来确定。备选地，多个多态序列与多个障碍的关联能够从多个障碍的每个的多个多态序列的识别码来确定。

[0306] 常规基因型技术局限于识别数千碱基的短基因组区域中的多态性，以及单倍型的识别依靠使用计算算法的家族数据和统计估计。全基因组测序通过直接识别基因组上的多态性，来实现单倍型的识别。按照各个实施例的单倍型的识别并不受多态性之间的中间距离限制。在一些实施例中，方法包括对母体细胞DNA进行全基因组测序。母体细胞DNA能够从没有胎儿基因组DNA的生物样本来得到。例如，母体DNA能够从母体血液的血沉棕黄层来得

到。能够确定包括多个多态序列(其跨越整个染色体)的单倍型。在一个实施例中,将胎儿单倍型与已知障碍关联单倍型进行比较,并且基于胎儿单倍型与已知障碍关联单倍型的任一个的关联来指明胎儿具有障碍或者胎儿易受障碍影响。还能够将胎儿单倍型与关联特定多态性的治疗响应性或无响应性的单倍型进行比较。所识别胎儿单倍型与已知单倍型数据库的比较允许障碍的诊断和/或预后。包括胎儿和母体cfDNA的混合的任何生物样本能够用于确定胎儿障碍的存在或不存在。优选地,生物样本从包括血浆在内的血液或者其分数或者尿液来选取。在一个实施例中,生物样本是血液样本。在另一个实施例中,生物样本是血浆样本。在又一个实施例中,生物样本是尿液样本。

[0307] 在一个实施例中,本发明提供一种用于确定多个胎儿障碍的存在或不存在的方法,包括:(a)得到包括胎儿和母体DNA的游离混合的母体血液样本;(b)对胎儿和母体DNA的游离混合的至少一部分进行全基因组测序,由此得到多个序列标签;(c)确定序列标签中的多个胎儿多态性;以及(d)确定多个胎儿障碍的存在或不存在。能够按照本方法来识别的多个胎儿障碍的示例包括本文所述的单基因和多基因病。

[0308] 在一个实施例中,本发明提供一种用于确定多个胎儿障碍的存在或不存在的方法,其包括识别与多个障碍相关单倍型关联的多个胎儿多态性。在一些实施例中,单倍型的每个包括至少2个、至少3个、至少4个、至少5个、至少10个或者至少15不同的标签多态性。单倍型中存在的标签多态性能够属于相同类型的多态性、例如所有标签SNP多态性,或者能够是多态性的组合、例如标签SNR和标签缺失。在一个实施例中,多态性是标签SNP。在另一个实施例中,多态性是标签STR。在又一个实施例中,多态性是标签SNP和标签STR的组合。标签多态性能够在基因组的编码和/或非编码区域中。多态性的识别通过使用如本文所述的NGS技术的全基因组测序来执行。

[0309] 本发明提供一种用于将副本数量变化(CNV)识别为测试样本中的感兴趣序列的多态性的方法,其中测试样本包括从两个不同基因组所得出的核酸的混合,并且是已知的或者怀疑在一个或多个感兴趣序列的量中有所不同。通过本发明的方法所确定的副本数量变化包括整体多态性的增益或丢失、涉及显微镜下可见的极大染色体段的改变以及大小范围从千碱基(kb)至兆碱基(Mb)的DNA段的子显微副本数量变化的丰度。

[0310] 人类基因组中的CNV显著影响人的多样性以及对疾病的倾向(Redon等人,Nature 23:444-454[2006],Shaikh等人,Genome Res 19:1682-1690[2009])。CNV已知为通过不同机制来贡献遗传疾病,从而在大多数情况下引起基因剂量的不平衡或者基因中断。除了其与遗传病的直接相关性之外,CNV还已知为仲裁能够有害的单倍型变化。近来,若干研究报道了如与正常控制相比的诸如自闭症、ADHD和精神分裂症之类的复合障碍中的罕见或重新CNV的增加负担,从而突出罕见或独特CNV的潜在病原性(Sebat等人,316:445-449[2007];Walsh等人,Science 320:539-543[2008])。主要由于缺失、重复、插入和不平衡易位事件,CNV产生于基因组重新排列。

[0311] 本发明的实施例提供一种评估测试样本中的感兴趣序列、例如临床相干序列的副本数量变化的方法,其中测试样本包括从两个不同基因组所得出的核酸的混合,并且是已知的或者怀疑在一个或多个感兴趣序列的量中有所不同。核酸的混合从两种或者更多种类型的细胞来得出。在一个实施例中,核酸的混合从自经过医疗条件、例如癌症的受检者所得出的正常和癌细胞来得出。

[0312] 我们认为，许多实体瘤、例如乳癌通过若干遗传畸变的积累从开始发展到转移。[Sato等人，Cancer Res.，50:7184-7189[1990]；Jongsma等人，J ClinPathol:Mol Path 55:305-309[2002]]。这类遗传畸变在其积累时可赋予增生优点、遗传不稳定性和迅速发展药物抗性的伴随能力以及增强血管形成、蛋白质水解和转移。遗传畸变可影响隐性“抑瘤基因”或者显性作用致癌基因。通过揭示突变肿瘤抑制等位基因，引起杂合性(LOH)的丢失的缺失和重组被认为在肿瘤进展中起主要作用。

[0313] 在诊断有恶性肿瘤的患者的循环中已经发现cfDNA，其中包括但不限于肺癌(Pathak等人，ClinChem 52:1833-1842[2006])、前列腺癌(Schwartzenbach等人，Clin Cancer Res 15:1032-8[2009])以及乳癌(Schwartzenbach等人，在breast-cancer-research.eom/content/11/5/R71在线可得到，[2009])。与能够在癌症患者的循环cfDNA中确定的癌症关联的基因组不稳定性的识别是潜在诊断和预后工具。在一个实施例中，本发明的方法评估包括从疑似或者已知为具有例如恶性肿瘤、肉瘤、淋巴瘤、白血病、胚组织瘤和胚细胞瘤等癌症的受检者所得出的核酸的混合的样本中的感兴趣序列的CNV。在一个实施例中，样本是从周边血液所得出(处理)的血浆样本，并且其包括从正常和癌细胞所得出的cfDNA的混合。在另一个实施例中，确定CNV是否存在所需的生物样本从来自其它生物流体(包括血清、汗液、泪水、唾液、尿液、唾液、耳液、淋巴、痰、脑脊髓液、ravage、骨髓悬浮液、阴道液、宫颈灌洗液、脑液、腹水、乳剂、呼吸分泌物、肠和生殖泌尿道、羊水和leukophoresis样本)或者在组织活检、药签或涂片中的癌和非癌细胞的混合来得出。

[0314] 感兴趣序列是核酸序列，其已知或者疑似在癌症的发展和/或进展中起作用。感兴趣序列的示例包括核酸序列，其在癌细胞中放大或缺失，如下面所述。

[0315] 与人类实体瘤关联的显性作用基因通常通过超表达或者改变表达来发挥其效果。基因放大是引起基因表达的未调控的常见机制。来自细胞生成研究的论据表明，显著放大在人类乳癌的50％以上中发生。最值得注意的是，位于染色体17(17(17q21-金2))的原癌基因人类表皮生长因子受体2(HER2)的放大引起细胞表面上的HER2受体的超表达，从而引起乳癌和其它恶性肿瘤中的过度和异常信令(Park等人，Clinical Breast Cancer 8:392-401[2008])。已经发现多种致癌基因在其它人类恶性肿瘤中被放大。人类肿瘤中的细胞致癌基因的放大的示例包括下列的放大：原髓细胞白血病细胞系HL60中和小细胞肺癌细胞系的c-myc，原发神经母细胞瘤(III和IV期)、神经母细胞瘤细胞系、视网膜母细胞瘤细胞系和原发肿瘤、小细胞肺癌系和肿瘤中的N-myc，小细胞肺癌细胞系和肿瘤中的L-myc，急性髓细胞样白血病和结肠癌细胞系中的c-myb，扁平上皮癌细胞和原发神经胶质瘤中的c-erbb，肺、结肠、膀胱和直肠的原发恶性肿瘤中的c-K-ras-2，乳腺癌细胞系中的N-ras(Varmus H.，Ann Rev Genetics 18:553-612(1984)[Watson等人，Molecular Biology of the Gene (4th ed.;Benjamin/Cummings Publishing Co.1987)])。

[0316] 涉及抑瘤基因的染色体缺失在实体瘤的发展和进展中可起重要作用。位于染色体13q14中的视网膜母细胞癌抑瘤基因(Rb-1)是最广泛表征的抑瘤基因。Rb-1基因产物、即105kDa核磷蛋白质在细胞周期调控中显然起重要作用(Howe等人，ProcNatlAcadSci(USA) 87:5883-5887[1990])。Rb蛋白质的改变或丢失表达通过经由点突变或者染色体缺失的基因等位基因的灭活引起。发现Rb-i改变不仅存在于视网膜母细胞瘤中，而且还存在于其它恶性肿瘤、例如骨肉瘤和小细胞肺癌(Rygaard等人，Cancer Res 50:5312-5317[1990)])以

及乳癌中。限制片断长度多态性(RFLP)研究已经指明,这类肿瘤类型在13q频繁丢失杂合性,表明Rb-1基因等位基因之一因总染色体缺失而已经丢失(Bowcock等人,Am J Hum Genet,46:12[1990])。包括重复、缺失和不平衡易位、其中涉及染色体6和其它同伴染色体的染色体1异常指明,染色体1的区域、特别是1q21-1q32和1p11-13可能持有致癌基因或抑瘤基因,其以致病方式与骨髓增生性肿瘤的慢性和晚期阶段相关(Caramazza等人,Eur J Hematol84:191-200[2010])。骨髓增生性肿瘤还与染色体5的缺失关联。染色体5的完全丢失或者间质缺失是骨髓增生异常综合症(MDS)中的最常见核型异常。隔离del(5q)/5q-MDS患者与具有附加核型缺陷的患者(他们趋向于发展骨髓增生性肿瘤(MPN)和急性髓细胞样白血病)相比具有更有利的预后。不平衡染色体5缺失的频率引起如下概念:5q持有一个或多个抑瘤基因,其在造血干/祖细胞(HSC/HPC)的生长控制中具有基础作用。集中在5q31和5q32的通常缺失区域(CDR)的细胞遗传学映射识别候选抑瘤基因,包括核蛋白体亚单位RPS14、转录因子Egr1/Krox20和心肌重构蛋白质α-catenin(Eisenmann等人,Oncogene 28:3429-3441[2009])。新肿瘤和肿瘤细胞系的细胞生成和等位型研究已经表明,来自染色体3p、包括3p25、3p21-22、3p21.3、3p12-13和3p14上的若干不同区域的等位基因丢失是肺、乳、肾、头和颈、卵巢、子宫颈、结肠、胰腺、食道、膀胱和其它器官的大上皮癌的宽谱中涉及的最早和最频繁的基因组异常。若干抑瘤基因映射到染色体3p区域,并且被认为间质缺失或者启动子高甲基化先于恶性肿瘤的发展中的3p或整个染色体3的丢失(Angeloni D., Briefings Functional Genomics 6:19-39[2007])。

[0317] 具有唐氏综合症(DS)的新生儿和儿童常常有先天短暂白血病,并且具有急性髓细胞样白血病和急性淋巴细胞白血病的增加风险。持有大约300个基因的染色体21可涉及在白血病、淋巴瘤和实体瘤的许多结构畸变中,例如易位、缺失和放大。此外,已经识别位于染色体21上的基因在瘤形成中起重要作用。肉体数字以及结构染色体21畸变与白血病关联,以及包括RUNX1、TMPRSS2和TFF的特定基因(其位于21q中)在瘤形成中起作用(Fonatsch C Gene Chromosomes Cancer 49:497-508[2010])。

[0318] 在一个实施例中,该方法提供评估基因放大与瘤演进程度之间的关联的手段。放大和/或缺失与癌的阶段或等级之间的相关性在预后上是重要的,因为这种信息可有助于遗传性肿瘤等级的定义,其更好地预测带有具有最坏预后的更晚期肿瘤的疾病的未来过程。另外,与早期放大和/或缺失事件有关的信息在关联那些事件中作为后续疾病进展的预测值会是有用的。通过该方法所识别的基因放大和缺失能够与其它已知参数关联,例如肿瘤等级、组织学、Brd/Urd加标索引、激素状态、淋巴转移、肿瘤大小、存活时长以及从流行病学和生物统计研究可得到的其它肿瘤性质。例如,将要由该方法测试的肿瘤DNA可包括非典型增生、导管原位癌、I-III期癌和新陈代谢淋巴结,以便准许识别放大和缺失与阶段之间的关联。进行的关联可进行可能的有效治疗干预。例如,一致放大的区域可包含过表达基因,其产物可以能够以治疗方式来攻击(例如,生长因子受体酷氨酸激酶、$p185^{HER2}$)。

[0319] 该方法能够用于通过确定从原发癌症到转移到其它位点的细胞的核酸的副本数量变化,来识别与药物抗性关联的放大和/或缺失事件。如果基因放大和/或缺失是核型不稳定性的表现(其允许药物抗性的迅速发展),则预期来自化疗耐药患者的原发肿瘤中的更大放大和/或缺失。例如,如果特定基因的放大负责药物抗性的发展,则预期那些基因周围的区域在来自化疗耐药患者的胸腔积液的肿瘤细胞中但不在原发肿瘤中一致地放大。基因

放大和/或缺失与药物抗性的发展之间的关联的发现可允许识别将会或者将不会获益于辅助疗法的患者。

[0320] 在其它实施例中，本方法能够用于识别与三核苷酸重复障碍(其是通过三核苷酸重复扩充所引起的一组遗传病)关联的多态性。三核苷酸扩充是在整个全部基因组序列发生的不稳定微卫星重复的子集。如果重复存在于健康基因中，则动态突变可增加重复计数，并且引起缺陷基因。在一个实施例中，该方法能够用于识别与脆性X综合症关联的三核苷酸重复。如与携带者中的60至230重复以及未受影响具体中的5至54重复相比，患有脆性X综合症的患者的X染色体的长臂能够包含从230至4000CGG。产生于这个三核苷酸扩充的染色体不稳定性在临床上作为智力迟钝、与众不同面部特征和男性的大睾丸症而存在。第二相关DNA三联体重复疾病、脆性X-E综合症也在X染色体上识别，但是发现是扩大CCG重复的结果。本方法能够识别与包括第I、II和III类的其它重复扩充障碍关联的三核苷酸重复。第I类障碍包括亨廷顿疾病(HD)和脊髓小脑共济失调，其通过特定基因的蛋白质编码部分中的CAG重复扩充引起。第II类扩充趋向于与异质扩充(其幅值一般较小，但是也见于基因的外显子上)在表型上更为不同。第III类包括：脆性X综合症；强直性肌营养不良；脊髓小脑共济失调、少年肌肉阵挛性癫痫和friereich共济失调中的两个。这些疾病的特征在于通常比前两组要大许多的重复扩充，以及重复位于基因的蛋白质编码区域外部。

[0321] 在其它实施例中，本方法能够识别与已知为通过通常在原本不相关蛋白质的编码区域中的增加数量的CAG重复所引起的至少10个神经性障碍关联的CAG三核苷酸重复。在蛋白质合成期间，扩充CAG重复转化为形成称作多谷氨酰胺段(“polyQ”)的一系列未中断谷氨酰胺残基。这类多谷氨酰胺段可经过增加的聚合。这些障碍的特征在于遗传的常染色体显性模式(除了脊髓延髓肌萎缩症之外，其呈现X连锁遗传)、中年发作、渐进过程以及CAG重复数量与疾病严重性和发作年龄的相关性。致病基因在所有已知多谷氨酰胺疾病中广泛地表达。PolyQ疾病的共同症状的特征在于通常在以后的生活中影响人们的神经细胞的渐进衰退。虽然这些疾病共用相同的重复密码子(CAG)和一些症状，但是不同多谷氨酰胺疾病的重复在不同的染色体上发生。能够通过本方法来识别的polyQ障碍的示例非限制性地包括DRPLA(齿状核红核苍白球路易体萎缩)、HD(亨廷顿疾病)、SBMA(脊延髓肌萎缩症或者肯尼迪病)、SCA1(1型脊髓小脑共济失调)、SCA2(2型脊髓小脑共济失调)、SCA3(3型脊髓小脑共济失调或者Machado-Joseph疾病)、SCA6(6型脊髓小脑共济失调)、SCA7(7型脊髓小脑共济失调)、SCA17(17型脊髓小脑共济失调)。能够通过本方法来识别的非polyQ障碍的示例包括FRAXA(脆性X综合症)、FXTAS(脆性X关联震颤/共济失调综合症)、FRAXE(脆性XE智力迟钝)、FRDA(Friedreich共济失调)、DM(肌强直性营养不良)、SCA8(8型脊髓小脑共济失调)、SCA12(12型脊髓小脑共济失调)。

[0322] 除了CNV在癌症中的作用之外，CNV还与增长数量的普通复杂疾病关联，包括人体免疫缺陷病毒(HIV)、自身免疫性疾病和神经精神障碍谱。

[0323] 至今，多种研究报道了炎症和免疫响应和HIV、哮喘、克罗恩病以及其它自身免疫障碍中涉及的基因的CNV之间的关联(Fanciulli等人，Clin Genet 77:201-213[2010])。例如，CCL3L1中的CNV已经涉及在HIV/AIDS易感性(CCL3L1，17q11.2缺失)、风湿性关节炎(CCL3L1，17q11.2缺失)和川崎病(CCL3L1，17q11.2重复)中；已经报道HBD-2中的CNV倾向于结肠克罗恩氏病(HDB-2，8p23.1缺失)和扯皮癣(HDB-2，8p23.1缺失)；FCGR3B中的CNV表明

倾向于系统性红斑儿狼疮中的血管球性肾炎(FCGR3B,1q23缺失,1q23重复)、抗中性粒细胞浆抗体(ANCA)关联的振荡(vasculatis)(FCGR3B,1q23缺失),并且增加发展风湿性关节炎的风险。存在至少两种炎症或自身免疫疾病,其已经表明与不同基因座位的CNV关联。例如,克罗恩病与在HDB-2的低副本数量关联,但是还与IGRM基因上游的普通缺失多态性关联,其对p47免疫相关GTPase家族成员进行编码。除了与FCGR3B副本数量的关联之外,还已经报道SLE易感性在具有补充成分CA的较低数量副本的受检者之中显著增加。

[0324] 在GSTM1(GSTM1,1q23缺失)和GSTT1(GSTT1,22q11.2缺失)基因座的基因组缺失与变应性哮喘的增加风险之间的关联在多个单独研究中已经报道。在一些实施例中,本方法能够用于确定与炎症和/或自身免疫疾病关联的CNV的存在或不存在。例如,本方法能够用于确定疑似患有HIV、哮喘或者克罗恩病的患者中的CNV的存在。与这类疾病关联的CNV的示例非限制性地包括在17q11.2、8p23.1、1q23和22q11.2的缺失以及在17q11.2和1q23的重复。在一些实施例中,本方法能够用于确定包括但不限于CCL3L1、HBD-2、FCGR3B、GSTM、GSTT1、C4和IRGM的基因中的CNV的存在。

[0325] 重新和继承DNV与若干普通神经和精神疾病之间的关联在自闭症、精神分裂症和癫痫症以及神经变性病、例如帕金森病、肌萎缩性脊髓侧索硬化(ALS)和常染色体显性阿尔兹海默病的一些病例中已经报道(Fanciulli等人,Clin Genet 77:201-213[2010])。在具有带15q11-q13的重复的自闭症和自闭症谱系障碍(ASD)患者中已经观测到细胞生成异常。按照自闭症基因组项目联盟,154CNV包括染色体15q11-q13上或者在新基因组位置(其中包括染色体2p16、1q21)以及在区域(其关联与ASD重叠的司马综合症)中的17p12的若干复发CNV。染色体16p11.2上的复发微缺失或者微重复已经突出如下观察:重新CNV基因的座位被检测,例如SHANK3(22q13.3缺失)、轴突蛋白1(NRXN1,2p16.3缺失)和神经胶(NLGN4,Xp22.33缺失),其已知为调控突触分化以及调控谷氨酸能神经递质释放。精神分裂症还与多个重新CNV关联。与精神分裂症关联的微缺失和微重复包含属于神经发育和谷氨酸能通道的基因的超表示,表明影响这些基因的多个CNV可直接有助于精神分裂症的发病机理,例如ERBB4,2q34缺失,SLC1A3,5p13.3缺失;RAPEGF4,2q31.1缺失;CIT,12。24缺失;以及具有重新CNV的多个基因。CNV还与其它神经障碍关联,包括癫痫症(CHRNA7,15q13.3缺失)、帕金森病(SNCA 4q22重复)和ALS(SMN1,5q12.2-q13.3缺失;和SMN2缺失)。在一些实施例中,本方法能够用于确定与神经系统疾病关联的CNV的存在或不存在。例如,本方法能够用于确定疑似患有自闭症、精神分裂症、癫痫症、神经变性病、例如帕金森病、肌萎缩性脊髓侧索硬化症(ALS)或者常染色体显性阿尔兹海默病的患者中的CNV的存在。本方法能够用于确定与神经系统的疾病关联的基因的CNV,非限制性地包括自闭症谱系障碍(ASD)、精神分裂症和癫痫症中的任一个,以及与神经变性病、例如帕金森病关联的基因的CNV。与这类疾病关联的CNV的示例非限制性地包括在15q11-q13、2p16、1q21、17p12、16p11.2和4q22的重复以及在22q13.3、2p16.3、Xp22.33、2q34、5p13.3、2q31.1、12.24、15q13.3和5q12.2的缺失。在一些实施例中,本方法能够用于确定基因中的CNV的存在,包括但不限于SHANK3、NLGN4、NRXN1、ERBB4、SLC1A3、RAPGEF4、CIT、CHRNA7、SNCA、SMNI和SMN2。

[0326] 新陈代谢和心血管特质、例如家族性高胆固醇血症(FH)、动脉硬化症和冠心病与CNV之间的关联在多个研究中已经报道(Fanciulli等人,Clin Genet 77:201-213[2010])。例如,在没有携带其它LDLR突变的一些FH患者的LDLR基因(LDLR,19p13.2缺失/重复)已经

观测到种系重新排列、主要是缺失。另一个示例是LAP基因，其对阿朴脂蛋白(a)(apo(a))进行编码，其血浆浓度与冠心病、心肌梗塞(MI)和中风的风险关联。包含脂蛋白Lp(a)的Apo(a)的血浆浓度在单个之间改变超过1000倍(fold)，以及这种可变性的90％在LPA基因座在遗传上确定，其中血浆浓度和Lp(a)同种型大小与'kringle 4'重复序列的高可变数量(范围5-50)成比例。这些数据指明，至少两个基因中的CNV能够与心血管风险关联。本方法能够在大型研究中用于专门搜索与心血管病的CNV关联。在一些实施例中，本方法能够用于确定与新陈代谢和心血管病关联的CNV的存在或不存在。例如，本方法能够用于确定疑似患有家族性高胆固醇血症的患者中的CNV的存在。本方法能够用于确定与新陈代谢或心血管病、例如血胆脂醇过多关联的基因的CNV。与这类疾病关联的CNV的示例非限制性地包括LDLR基因的19p13.2缺失/重复以及LPA基因中的繁殖。

[0327] 测序

[0328] 在各个实施例中，本文所述的方法采用下一代测序技术(NGS)，其中无性放大DNA模板或者单DNA分子在流式细胞中按照大规模平行方式来测序(如Volkerding等人的ClinChem 55:641-658[2009]中所述;Metzker M Nature Rev 11:31-46[2010])。除了高吞吐量序列信息之外，NGS还提供数字定量信息，因为各序列读数是可计数"序列标签"，表示个体无性DNA模板或者单DNA分子。NGS的测序技术包括焦磷酸测序、采用可逆染料终止剂的合成测序、通过寡核苷酸探针连接的测序以及实时测序。

[0329] 在各个实施例中，能够分析样本，其没有经过放大或者仅部分放大(靶放大)。在一些情况下，能够实现确定胎儿分数的方法，而无需任何类型的靶放大。

[0330] 作为测序过程的组成部分发生的全基因组放大提供充分副本，其能够通过增加测序周期的数量以提供增加的更好覆盖来覆盖。

[0331] 在优选实施例中，在全基因组测序之前非专门地增强包括从两个不同基因组所得出的DNA分子的混合的样本，即，全基因组放大在测序之前执行。

[0332] 样本DNA的非特定增强可指样本的基因组DNA片断的全基因放大，其能够用于在通过测序识别多态性之前增加样本DNA的等级。非特定增强能够是样本中存在的两个基因组之一的选择性增强。例如，非特定增强能够是选择母体样本中的胎儿基因组，其能够通过已知方法来得到，以便增加样本中胎儿与母体DNA的相对比例。备选地，非特定增强能够是样本中存在的两种基因组的非选择性放大。例如，非特定放大能够属于包括来自胎儿和母体基因组的DNA的混合的样本中的胎儿和母体DNA。用于全基因组放大的方法是本领域已知的。简并寡核苷酸引物PCR(DOP)、引物延伸PCR技术(PEP)和多重置换扩增(MDA)是全基因组放大方法的示例。在一些实施例中，包括来自不同基因组的cfDNA的混合的样本对于混合中存在的基因组的cfDNA未增强。在其它实施例中，包括来自不同基因组的cfDNA的混合的样本对于样本中存在的基因组的任一个非专门增强。

[0333] 在其它实施例中，对样本中的cfDNA专门增强。特定增强指的是对于特定序列(其在对DNA样本测序之前选择用于放大)、例如多态靶序列的基因组样本的增强。但是，所公开实施例的优点在于，不需要靶放大。多态

[0334] 一些测序技术是商业可行的，例如Affymetrix Inc.(Sunnyvale,CA))的杂交测序平台、来自454 Life Sciences(radford,CT)、Illumina/Solexa(Hayward,CA)和Helicos Biosciences(Cambridge,MA)的合成测序平台以及来自Applied Biosystems(Foster

City,CA)的连接测序平台,如下面进行描述。除了使用Helicos Biosciences的合成测序所执行的单分子测序之外,其它单分子测序技术也由所公开方法来包含,并且包括Pacific Biosciences的SMRT™技术、Ion Torrent™技术以及例如由Oxford Nanopore Technologies所研制的纳米孔测序。

[0335] 虽然自动化Sanger方法被认为是'第一代'技术,但是包括自动化Sanger测序的Sanger测序也能够被所公开方法采用。包括发展核酸成像技术、例如原子力显微镜(AFM)或者透射电子显微镜(TEM)的使用的附加测序方法也被所公开方法包含。下面描述示范测序技术。

[0336] 在一个实施例中,所公开方法中使用的DNA测序技术是Helicos True单分子测序(tSMS)(例如在Harris T.D.等人的Science 320:106-109[2008]所述)。在tSMS技术中,将DNA样本裂解为大致100至200个核苷酸的链,以及将polyA序列添加到每个DNA链的3'末端。每个链通过添加荧光加标腺苷核苷酸来加标。DNA链则杂交到流式细胞,其包含数百万oligo-T俘获位点,其固定到流式细胞表面。模板能够处于大约100百万模板/cm^2的密度。流式细胞则加载到仪器、例如HeliScop™测序器中,以及激光器照射流式细胞的表面,从而展示各模板的位置。CCD照相装置能够将模板的位置映射在流式细胞表面上。模板荧光标签则经过裂解并且冲洗掉。测序反应开始于引入DNA聚合酶以及荧光加标核苷酸。oligo-T核苷酸用作引物。聚合酶按照模板引导方式将加标核苷酸结合到引物。去除聚合酶和未结合核苷酸。引导了荧光加标核苷酸的结合的模板通过对流式细胞表面进行成像来辨别。在成像之后,裂解步骤去除荧光标签,并且该过程对其它荧光加标核苷酸重复进行,直到取得预期读数长度。序列信息随各核苷酸添加步骤来收集。通过单分子测序技术的全基因组测序在测序库的制备中排除基于PCR的放大,以及样本制备的直接性允许样本的直接测量,而不是那个样本的副本的测量。

[0337] 在一个实施例中,所公开方法中使用的DNA测序技术是454测序(Roche)(例如在Margulies,M.等人的Nature 437:376-380(2005)所述)。454测序涉及两个步骤。在第一步骤,将DNA剪切为大致300-800碱基对的片断,以及片断为平末端。寡核苷酸接头则连接到片断末端。接头用作片断的放大和测序的引物。片断能够使用例如接头B(其包含5'生物素标签)来附连到DNA俘获微珠、例如抗生蛋白链菌素涂敷微珠。附连到微珠的片断在油-水乳状液的微滴中经过PCR放大。结果是各微珠上的无性放大DNA片断的多个副本。在第二步骤,在陷阱(微微升大小)俘获微珠。焦磷酸测序对每个DNA片断并行地执行。添加一个或多个核苷酸生成光信号,其由CCD照相装置记录在测序仪器中。信号强度与所结合的核苷酸的数量成比例。焦磷酸测序利用焦磷酸盐(PPi),其在核苷酸添加时被释放。PPi在腺苷5'磷酰硫酸存在的情况下通过ATP硫酸化酶来转换成ATP。荧光素酶使用ATP将荧光素转换成氧化荧光素,并且这个反应生成光,其被辨别和分析。

[0338] 在一个实施例中,所公开方法中使用的DNA测序技术是SOLiD™技术(Applied Biosystems)。在SOLiD™连接测序中,将基因组DNA剪切为片断,以及将接头附连到片断的5'和3'末端,以便生成片断库。备选地,能够通过将接头连接到片断的5'和3'末端,使片断成圆形,酶切圆形化片断以生成内部接头,并且将接头附连到所产生片断的5'和3'末端以生成配对库,来引入内部接头。随后,克隆微珠在包含微珠、引物、模板和PCR成分的微反应器中制备。接着PCR,使模板变性,并且增强微珠以将微珠与延长模板分离。所选微珠上的模板

经过3'改性，其准许接合到载玻片。通过部分随机寡核苷酸与中心确定碱基(或者碱基对)——其通过特定荧光团来识别——依次杂交和连接，来确定序列。在记录颜色之后，连接寡核苷酸被裂解和去除，以及然后该过程重复进行。

[0339] 在一个实施例中，所公开方法中使用的DNA测序技术是Pacific Biosciences的单分子实时(SMRT™)测序技术。在SMRT测序中，在DNA合成期间对染料加标核苷酸的连续结合进行成像。将单DNA聚合酶分子附连到单独零模波长标识符(ZMW标识符)的底面，其在磷连锁核苷酸被结合到生长引物链的同时得到序列信息。ZMW是限制结构，其实现针对荧光核苷酸-其迅速扩入ZMW外(单位为微秒)-的背景来观测单核苷酸与DNA聚合酶的结合。它花费数毫秒将核苷酸结合到生长链中。在这个时间期间，荧光标签被激发，并且产生荧光信号，以及荧光标签裂开。染料的对应荧光的标识指示结合哪一个碱基。该过程重复进行。

[0340] 在一个实施例中，所公开方法中使用的DNA测序技术是纳米孔测序(例如，如Soni GV和Meller A.的ClinChem 53:1996-2001[2007]所述)。纳米孔测序DNA分析技术由多家公司在工业上研制，其中包括Oxford Nanopore Technologies(Oxford,United Kingdom)。纳米孔测序是单分子测序技术，由此DNA的单分子在经过纳米孔时直接被测序。纳米孔是直径大约1纳米的小孔。导电流体中的纳米孔的沉浸以及其两端的电位(电压)的施加因通过纳米孔的离子的导电而产生微波电流。流动的电流量对纳米孔的大小和形状敏感。当DNA分子经过纳米孔时，DNA分子上的各核苷酸在不同程度上阻隔纳米孔，从而以不同程度来改变通过纳米孔的电流的幅值。因此，当DNA分子经过纳米孔时的电流的这种变化表示DNA序列的读数。

[0341] 在一个实施例中，所公开方法中使用的DNA测序技术是化学敏感场效应晶体管(chemFET)阵列(例如，如2007年12月17日提交的美国专利发表No.2009/0026082中所述)。在该技术的一个示例中，DNA分子能够放入反应室中，以及模块分子能够杂交到结合到聚合酶的测序引物。在测序引物的3'末端将一个或多个三磷酸盐结合到新核酸链中能够由chemFET通过电流的变化来辨别。阵列能够具有多个chemFET传感器。在另一个示例中，单核酸能够附连到微珠，以及核酸能够在微珠上放大，并且单独微珠能够传递到chemFET阵列上的单独反应室，其中每个室具有chemFET传感器，并且能够对核酸测序。

[0342] 在一个实施例中，所公开方法中使用的DNA测序技术是Halcyon Molecular方法，其使用透射电子显微镜(TEM)。称作单独分子放置快速纳米传递(IMPRNT)的方法包括利用采用重原子标记有选择地加标的高分子重量(150kb或以上)DNA的单原子分辨率透射电子显微镜成像，以及以一致碱基-碱基间距将超薄膜上的这些分子排列在超密集(3nm链-链)平行阵列中。电子显微镜用于对膜上的分子进行成像，以便确定重原子标记的位置，以及从DNA提取碱基序列信息。该方法在PCT专利发表WO2009/046445中进一步描述。该方法允许以不到十分钟的时间对全人类基因组进行测序。

[0343] 在一个实施例中，DNA测序技术是Ion Torrent单分子测序，其将半导体技术与简单测序化学组对，以便在半导体芯片上将化学编码信息(A、C、G、T)直接转化为数字信息(0、1)。实际上，当核苷酸通过聚合酶结合到DNA链中时，氢离子作为副产品被释放。Ion Torrent使用微加工阱的高密度阵列按照大规模平行方式来执行这个生物化学过程。每个阱保持不同的DNA分子。阱的下面是离子敏感层，以及其下是离子传感器。当核苷酸、例如C添加到DNA模板并且然后结合到DNA链时，将释放氢离子。来自那个离子的电荷将改变溶液

的pH,其能够通过Ion Torrent的离子传感器来识别。测序器-本质上是全球最小的固态pH计-读出碱基,从化学信息直接信息直接转到数字信息。离子个人基因组测序仪(PGM™)测序器则使芯片依次充满核苷酸。如果充满芯片的下一个核苷酸不是匹配物,没有电压变化将被记录,并且没有碱基被读出。如果DNA链上存在两个相同碱基,则电压将翻倍,并且芯片将记录被读出的两个相同碱基。直接识别允许在数秒时间记录核苷酸结合。

[0344] 在一些实施例中,方法将PCR或者相关技术用于在映射样本核苷酸序列之前将其放大。但是,本文所公开的算法技术一般不要求放大,特别是用于估计基因组分数的多态性的靶放大。

[0345] 某些实施例通过杂化来采用数字PCR和测序。数字聚合酶链反应(数字PCR或dPCR)能够用于直接识别和量化样本中的核酸。数字PCR能够在乳状液中执行。单独核酸例如在微流体室装置中分离,并且各核酸通过PCR单独放大。核酸能够分离成使得存在大致0.5核酸/阱或者不超过1核酸/阱的平均数。不同探针能够用于区分胎儿等位基因和母体等位基因。能够枚举等位基因以确定副本数量。在通过杂化的测序中,杂化包括将多个多核苷酸序列与多个多核苷酸探针相接触,其中多个多核苷酸探针的每个能够可选地连结到基底。基底可能是包括已知核苷酸序列的阵列的平坦表面。杂化到阵列的模式能够用于确定样本中的存在的多核苷酸序列。在其它实施例中,各探针连结到微珠、例如磁珠等。杂化到微珠能够被识别,并且用于识别样本中的多个多核苷酸序列。

[0346] 在一个实施例中,该方法采用使用Illumina的合成测序和可逆终止剂测序化学的数百万DNA片断的大规模平行测序。模板DNA能够是基因组DNA、例如cfDNA。在一些实施例中,来自隔离细胞的基因组DNA用作模板,以及将它分片为数百碱基对的长度。在其它实施例中,cfDNA用作模板,并且不要求分片,因为cfDNA作为短片断而存在。例如,胎儿cfDNA作为<300bp的片断在血流中循环,以及母体cfDNA已经估计为作为大约0.5与1Kb之间的片断进行循环(Li等人,ClinChem,50:1002-1011(2004))。Illumina的测序技术依靠分片基因组DNA附连到平面透光表面,其上结合了寡核苷酸锚点。模板DNA经过末端修复以生成5'磷酸化平末端,以及Klenow片断的聚合酶活动用于将单A碱基添加到平磷酸化DNA片断的3'末端。这种添加制备DNA片断供连接到寡核苷酸接头,其在3'末端具有单T碱基的突出量,以增加连接效率。接头寡核苷酸是流式细胞锚点的互补。在限制稀释条件下,将接头改性单链模板DNA添加到流式细胞,并且通过杂化到锚点来固定。附连DNA片断被延长,并且经过桥放大以创建具有数亿个聚类——各包含相同模板的~1000个副本——的超高密度测序流式细胞。在一个实施例中,随机分片的基因组DNA、例如cfDNA在经过聚类放大之前使用PCR来放大。备选地,使用无放大基因组库制备,以及随机分片的基因组DNA、例如cfDNA单独使用聚类放大来增强(Kozarewa等人,Nature Methods 6:291-295[2009])。模板使用健壮四色DNA合成测序技术-其采用具有可去除荧光染料的可逆终止剂-来测序。高灵敏度荧光标识使用激光激发和全内反射光学器件来实现。大约20-40bp、例如36bp的短序列读数针对重复掩蔽参考基因组来对齐,以及基因差使用专门开发的数据分析管线软件来调用。在完成第一读数之后,模板能够就地再生,以便实现从片断的相对端的第二读数。因此,按照该方法使用DNA片断的单端或组对端测序。执行样本中存在的DNA片断的部分测序,以及计算包括预定长度、例如36bp的读数的序列标签,其被映射到已知参考基因组。

[0347] 序列读数的长度与特定测序技术关联。NGS方法提供序列读数,其大小从数十到数

百碱基对改变。在本文所述方法的一些实施例中，序列读数为大约20bp、大约25bp、大约30bp、大约35bp、大约40bp、大约45bp、大约50bp、大约55bp、大约60bp、大约65bp、大约70bp、大约75bp、大约80bp、大约85bp、大约90bp、大约95bp、大约100bp、大约110bp、大约120bp、大约130、大约140bp、大约150bp、大约200bp、大约250bp、大约300bp、大约350bp、大约400bp、大约450bp或者大约500bp。预计技术进步将实现大于500bp的单端读数，从而在生成组对端读数时实现大于大约1000bp的读数。在一个实施例中，序列读数为36bp。能够由所公开方法采用的其它测序方法包括单分子测序方法，其能够对＞5000bp的核酸分子进行测序。大量序列输出通过分析管线来传递，其将一次成像输出从测序器变换为碱基串。集成算法包执行核心一次数据变换步骤：图像分析、强度评分、碱基读出和对齐。

[0348] 映射

[0349] 各种计算方法可以用来映射每一个所识别的序列到容器(bin)，例如，通过识别样本中映射到特点基因、染色体、对偶基因，或者其他结构的所有序列。存在很多计算机算法以对齐序列，包括但不限于，BLAST(Altschul等，1990)，BLITZ(MPsrch)(Sturrock&Collins，1993)，FASTA(Person&Lipman，1988)，BOWTIE(Langmead等，Genome Biology 10：R25.1-R25.10[2009])，或者EL和(Illumina，Inc.，SanDiego，CA，USA)，在一些实施例中，容器的序列可以在技术领域已知的核酸数据库中被发现，包括但不限于，GenBank、dbEST、dbSTS、EMBL(欧洲分子生物实验室)，和DDBJ(日本DNA数据库)。BLAST或者类似工具可以用来相对于序列数据库搜索所识别的序列，而检索结果可以用来把所识别的序列整理到合适的容器中。

[0350] 装置

[0351] 测序数据的分析和由此而来的诊断一般利用计算机硬件执行，其中计算机硬件根据定义的算法和程序操作。由此，某些实施例使用处理步骤，其涉及存储在或者经由一个或多个计算机系统或者其他处理系统传输的数据。本发明的实施例还涉及用于执行这些操作的装置。该装置可以专门构造成用于所需目的，或者可以是通用计算机(或计算机组)，其通过存储在计算机中的计算机程序和/或数据结构选择性地激活或者重配置。在一些实施例中，一组处理器共同和/或并行地执行一些或者所有所述分析操作(例如，通过网络或者云计算)。用于执行本文所述方法的处理器或者处理器组可以是各种类型的，包括微控制器和微处理器，例如可编程装置(例如，CPLD和FPGA)和其他装置例如门阵列ASIC、数字信号处理器、和/或通用处理器。

[0352] 另外，一些实施例涉及有形的和/或非临时性的计算机可读介质或者计算机程序产品，其包括程序指令和/或数据(包括数据结构)用于执行各种计算机执行的操作。计算机可读介质的例子包括但不限于，半导体存储装置、磁介质例如硬盘、磁带、光学介质例如CD、光磁介质、和特别地配置成存储和执行程序指令的硬件，例如只读存储装置(ROM)和随机读取存储器(RAM)。计算机可读介质可以直接被终端用户控制，或者介质可以间接被终端用户控制。例如，直接控制的介质包括位于用户设备的介质和/或与其他实体共享的介质。例如，间接控制的介质包括用户经由外部网络和/或经由提供共享资源的服务(例如“云”)间接读取的介质。例如，程序指令包括例如编译器生成的机器代码，和包含高级代码可以由计算机利用解译器执行的文件。

[0353] 在一个实施例中，计算机程序产品被提供用于生成输出，指示源自定义基因组(例

如胎儿的)的核酸分数,和可选地其他信息例如测试样本中胎儿非整倍性的存在与否。计算机产品可以包含指令用于执行一个或多个上述方法中的任一个以确定来自特定生物体的核酸分数。如所解释的,计算机产品可以包括非临时性和/或有形的计算机可读介质,其具有记录在其上的计算机可执行的或者可编译的逻辑(例如,指令),用于使能处理器来确定基因组分数,并且,在一些情况下,确定非整倍性或者其他状况在基因组中存在与否。在一个例子中,计算机产品包括计算机可读介质,其具有记录在其上的计算机可执行的或者可编译的逻辑(例如,指令),用于使能处理器来确定胎儿分数和诊断胎儿非整倍性,包括:接收子程序用于从来自母体生物样本的至少一部分核酸分子接收测序数据,其中所述测序数据包括在一个或多个多态性的基因座的序列;计算机分配的逻辑,用于分析序列以确定对于一个或多个多态性的等位基因计数,和确定母体生物样本中的核酸的胎儿分数;和输出子程序,用于生成输出,指示样本中的核酸的胎儿分数。

[0354] 如上所述,来自考虑样本中的序列信息可以映射到多态性参考序列。另外,所映射序列信息可以用来生成对于多态性的等位基因计数和/或确定接合性情况。这些信息可以用来确定胎儿分数。在各种实施例中,多态性参考序列例如存储在数据库例如关系或者目标数据库中。应当认识到,在大多数情况下,对于没有辅助工具的人类,执行任何一个或所有这些计算操作是不实际的,甚至是不可能的。例如,从样本映射单30bp读数到多态性数据库参考序列在没有计算装置的帮助下很可能会消耗过分长的时间。当然,问题是复杂的,因为可靠的结果常常需要映射几千(例如,至少约10000)或者甚至几百万读数到一个或多个染色体。

[0355] 在某些实施例中,所公开的方法利用存储的列表或者其他组织形式的数据,其中数据与用于生成待分析核酸序列的生物体的参考多态性相关。如上所述,来自考虑样本中的序列可以对齐或者映射到所存储的多态性。个体多态性一般是这样的序列,其具有的长度足以明白地映射到从核酸样本所识别的序列。典型地,多态性分成组,一个对于每个对偶基因。在各种实施例中,参考多态性存储在数据库中,该数据库包含多态性以及它们的序列的特征。该关于多态性的信息收集可以存储在例如相关的或者目标数据库中。

[0356] 图10显示了典型的计算机系统,其在适当配置或者设计时,可以用作本发明的分析装置。计算机系统200包括任意数量的处理器202(也称作中央处理单元,或者CPU),其耦合到存储装置包括初级存储206(一般是随机存储器,或者叫RAM),初级存储204(一般为只读存储器,或者叫ROM)。CPU 202可以是各种类型,包括微控制器和微处理器,例如可编程装置(例如,CPLD和FPGA)和不可编程装置(例如门阵列ASIC或者通用微处理器)。如同已知的,初级存储204用来传输数据和指令给CPU,而初级存储206一般以双向方式用来传输数据和指令。两种初级存储装置都可以包括任意合适的计算机可读介质,例如上文所说的那些。大容量存储装置208也双向地耦合到CPU202并提供额外数据存储能力,并且可以包括上文所述的任意计算机可读介质。大容量存储装置208可以用来存储程序,数据等,且一般是次级存储介质例如硬盘。应当认识到,保存在大容量存储装置208中的信息在适当情况下可以集成在作为虚拟存储作为初级存储206一部分的标准模式中。具体的大容量存储装置例如CD-ROM 214可以单向地传递数据给CPU。

[0357] CPU 202还耦合到界面210,其连接到一个或多个输入/输出装置,例如视频监视器、鼠标、键盘、麦克风、触摸屏、传感卡读卡器、平板、尖笔、声音或者笔记识别器,或者其他

已知的输入装置，例如其他计算机。最终，CPU 202可选地可以耦合到外部装置，例如利用外部连接的数据库或者计算机或者通信网络(如212处所示)。利用该连接，可以认为CPU可以从网络接收信息，或者向网络输出信息(在执行此处描述的方法步骤时)。

[0358] 序列或者其他数据，可以直接或者间接地由用户输入进计算机。在一个实施例中，计算机系统200直接耦合到读取和/或分析放大的核酸序列的测序工具。来这些工具的序列或者其他信息通过系统200经由界面212提供用于分析。可选地，系统200处理的序列从序列存储源例如数据库或者其他库提供。在处理装置200中，存储器装置例如初级存储206或者大规模存储208缓冲或者存储(至少临时地)核酸序列。另外，存储器装置可以存储各种染色体或者基因的标签数量、所计算的副本计数等等。存储器还可以存储各种常规和/或程序用来分析序列或者映射数据的存在。这些程序/常规和/或可以包括用来执行统计分析的程序等等。

[0359] 在一个例子中，用户提供样本到测序装置中。数据通过连接到计算机的测序装置收集和/或分析。计算机上的软件允许数据收集和/或分析。数据可以存储，显示(经由监视器或其他类似装置)，和/或发送到其他位置。如所指出，计算机可以连接到因特网，用于传输数据给远程用户(例如，医师、科学家或者分析师)使用的手持装置。应当明白，数据可以在传输之前存储和/或分析。在一些实施例中，原始数据收集并发送给将要分析和/或存储数据的远程用户(或装置)。传输可经由因特网，但也可以经由卫星或者其他连接。可选地，数据可以存储在计算机可读介质上(例如，CD或者半导体存储装置)而介质可以运输给终端用户(例如通过邮件)。远程用户可以在相同或者不同放入地理位置，包括但不限于，大楼、城市、州、国家或大陆。

[0360] 在一些实施例中，本发明的方法还包括收集与多个多核苷酸序列有关的数据并且发送数据到计算机。例如，计算机可以连接到实验室设备，例如，样本收集装置，核苷酸放大装置，核苷酸测序装置，或者杂交装置。计算机然后可以收集由实验室仪器采集的可应用数据。数据可以在任意步骤存储在计算机上，例如，在实时收集时，在发送前、在发送时、或者在发送后。数据可以存储在能够从计算机取出的计算机可读介质上。收集的或者存储的数据可以从计算机传输到远程位置，例如，通过局域网或广域网，例如因特网。

[0361] 在一个方面，本发明还提供一种系统，其能够执行核苷酸测序的定量分析，带有精度为至少60％、65％、70％、75％、80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、或至少99％。核苷酸测序可以包括Sanger测序、大规模平行测序、杂交或者其他此处描述的技术。该系统可以包括各种元件，例如，实验室设备和计算机系统，并且可以配置成执行本发明公开的方法。

[0362] 在一些实施例中，装置和/或编程指令还可以包括指令，用于自动地记录与方法(例如胎儿DNA分数和可选的存在与否胎儿染色体的非整倍性)相关的信息在在患者(提供母体测试样本的受测者)病历中。患者病历可以通过例如实验室、医生办公室、医院、保健组织、保险公司或个人病历网站保持。另外，基于处理器实现的分析结果，方法还可以涉及规定、发起和/或改变受检者(母体检测样本从该受检者获取)的治疗。其黑可以涉及执行对取自受测者的其他样本的一个或多个附加测试或分析。

[0363] 例子

[0364] 从已测序变量中预测的胎儿分数：情况2

[0365] 为展示本方法可以用来可靠估计母体样本中的胎儿分数，制造了人造‘母体’样本，并且染色体1和7的所有基因座处的碱基变化被识别，以预测较小贡献基因组的分数。

[0366] 从怀孕女性隔离的cfDNA是母体和胎儿cfDNA的混合物，带有胎儿cfDNA水平为总cfDNA的中值到10％(Lo等，2010，″Maternal Plasma DNA seqeucning reveals the genome-wide genetic and mutational profile of the fetus″，Prenatal Diagnosis，2，1-12)。为制造人造母体样本，从一位母亲和她的儿子获得的基因组DNA(gDNA)(mother and son DNAs NA 10924and NA 10925；The Coriell Institute for Medical Research，Camden，NJ)用来制造混合基因组的样本。母亲和儿子gDNA各5微克被修剪到约200bp的分数，且各自确定浓度。包含来自儿子的10％DNA和来自母亲的90％的DNA的人造样本被制造，以模仿母体血液样本，其通常被认为根据孕龄含有2-40％胎儿cfDNA[Lun等，2008，″Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma″，Clinical Chemistry，54，1664-1672]。测序图书馆从人造样本的DNA准备，且利用IlluminaHiSeq2000在4条流动池道上进行50次测序循环。生成接近8亿49-mer序列读数。

[0367] 8亿读数利用GSNAP算法对齐到重复屏蔽的人类参考基因组(hg19版)(http://research-pub.gene.com/gmap/)，其中允许一次错配，而没有插入或者缺失。在基因组上映射到多个位置的序列被忽略。所有其他所映射的读数都计数为序列标签，而仅一个基因座(40和100序列标签都映射到它)被考虑作进一步分析，即，只有覆盖40和100标签的碱基被考虑。

[0368] 对于每个碱基基因座，对映射到四个碱基的每个的标签的数量进行计数。具有多于两个可能碱基的基因座被排除，而仅有映射到单等位基因和二等位基因基因座的标签被用来预测人造胎儿分数。在每个碱基基因座映射的标签的数量代表在该基因座处的覆盖(D)。在该模仿的母体样本中，期望母亲的多数等位基因(B)的贡献会反映90％比例的标签，而儿子的少数等位基因(A)的贡献会反映10％比例的标签。

[0369] 图11A和B显示染色体1和7上的变体碱基观察(频率)的数量的直方图，分别用于染色体1和7的少数等位基因百分比(A/D)。少数等位基因的百分比是在给定基因座处的对偶基因总数的百分比。例如，对于存在8例少数等位基因A和56例多数等位基因B的给定基因座，少数等位基因百分比是8％。数据显示对于少数等位基因的最大出现次数(频率)被观察到，当少数等位基因百分比为5％时(这代表一半的胎儿分数)。因此，该数据预测样本包含为10％的胎儿分数，这正好对应用来制造人造母体样本的数据。

[0370] 图12A和B显示等位基因频率分别沿染色体1和7的分布。两幅图都显示变体对偶基因沿染色体的最大数量是5％的少数等位基因频率和95％的多数等位基因频率。剩下的数据点的一些代表母亲基因组中存在的二等位基因基因座，而其他代表测序方法的噪声。每幅图的中间部分(其中没有代表变体对偶基因)与染色体的着丝点一致，这已知是染色体的重复富量区域，而对于其，标签映射在多余一个基因座上并因此被排除在分析之外。在其他区域，例如着丝点侧面的区域和对应端粒的区域，变体对偶基因是过代表的。这些区域的过代表可以归于测序方法，其中一些区域比其他区域在较大水平测序。

[0371] 因此，本方法可以用来预测胎儿分数。该方法是尤其有利的，因为不需要识别靶向序列，例如SNP，而任何染色体的任何位置的任何变体都可以用来预测胎儿分数百分比。

[0372] 其他实施例

[0373] 尽管上文根据具体过程和装置大致描述了本发明，本发明具有更大的应用范围。具体来说，本发明以检测取自怀孕个体DNA样本中的胎儿DNA分数的方式描述，但是不限定于此，而是此处描写的概念和方法也可以应用在其他背景中，例如检测具有源自两个或多个不同基因组的DNA的样本中的DNA类型的相对量。当然，本领域技术人员会认识到其他变化、修改和替代。

[0374] 例如，尽管此处描述的大多数例子和应用涉及估计取自怀有胎儿的个体的DNA样本的胎儿DNA分数，本发明也不限于此。更一般地，各种实施例提供对来自测试样本中两个不同基因组的核酸相对量的估计，该测试样本包含来自两个不同基因组的核酸的混合物，并且各种实施例已知或很有可能在一个或多个感兴趣序列的数量上是不同的。核酸混合物划分为两个或多个细胞类型。

[0375] 另外，尽管此处描述的大多数例子取自怀孕的人类，本发明也不限于此。例如，提供待测样本的个体可以是包括多核苷酸序列的有机体，例如植物、昆虫(比如苍蝇)或者动物。在一些实施例中，受检对象是哺乳动物，例如，老鼠、鼠、狗、猴子或者人。如所述，受检对象可以是怀孕个体。受检者可以是带有某种疾病的个体，例如癌症，或者被异物例如微生物(例如病毒)感染。样本可以包括来自受检对象的体液，例如，血液、血浆、血清、痰、唾液、尿液、排泄物、脓、淋巴、粘液等等。例如，样本可以是包含母体和胎儿游离DNA的混合物的母体血浆样本。一般来说，所公开的方法可以涉及从样本中测序；映射序列读数到多态性；基于接合性分类多态性；以及估计来自样本中的第二源的DNA分数。

[0376] 附录1：等位基因检索数据库序列列表

>rs560681.1|Chr.1|length=111|allele=A
CACATGCACA GCCAGCAACC CTGTCAGCAG GAGTTCCCAC CAGTTTCTTT
CTGAGAACAT CTGTTCAGGT TTCTCTCCAT CTCTATTTAC TCAGGTCACA
GGACCTTGGG G
>rs560681.2|Chr.1|length=111|allele=G
CACATGCACA GCCAGCAACC CTGTCAGCAG GAGTTCCCAC CAGTTTCTTT
CTGAGAACAT CTGTTCAGGT TTCTCTCCAT CTCTGTTTAC TCAGGTCACA
GGACCTTGGG G

>rs1109037.1|Chr.2|length=126|allele=A
TGAGGAAGTG AGGCTCAGAG GGTAAGAAAC TTTGTCACAG AGCTGGTGGT
GAGGGTGGAG ATTTTACACT CCCTGCCTCC CACACCAGTT TCTCCAGAGT
GGAAAGACTT TCATCTCGCA CTGGCA
>rs1109037.2|Chr.2|length=126|allele=G
TGAGGAAGTG AGGCTCAGAG GGTAAGAAAC TTTGTCACAG AGCTGGTGGT
GAGGGTGGAG ATTTTACACT CCCTGCCTCC CACACCAGTT TCTCCGGAGT
GGAAAGACTT TCATCTCGCA CTGGCA

[0377]

>rs9866013.1|Chr.3|length=121|allele=C
GTGCCTTCAG AACCTTTGAG ATCTGATTCT ATTTTTAAAG CTTCTTAGAA
GAGAGATTGC AAAGTGGGTT GTTTCTCTAG CCAGACAGGG CAGGCAAATA
GGGGTGGCTG GTGGGATGGGA
>rs9866013.2|Chr.3|length=121|allele=T
GTGCCTTCAG AACCTTTGAG ATCTGATTCT ATTTTTAAAG CTTCTTAGAA
GAGAGATTGC AAAGTGGGTT GTTTCTCTAG CCAGACAGGG CAGGTAAATA
GGGGTGGCTG GTGGGATGGGA

>rs13182883.1|Chr.5|length=111|allele=A
AGGTGTGTCT CTCTTTTGTG AGGGGAGGGG TCCCTTCTGG CCTAGTAGAG
GGCCTGGCCT GCAGTGAGCA TTCAAATCCT CAAGGAACAG GGTGGGGAGG
TGGGACAAAG G
>rs13182883.2|Chr.5|length=111|allele=G
AGGTGTGTCT CTCTTTTGTG AGGGGAGGGG TCCCTTCTGG CCTAGTAGAG
GGCCTGGCCT GCAGTGAGCA TTCAAATCCT CGAGGAACAG GGTGGGGAGG
TGGGACAAAG G

>rs13218440.1|Chr.6|length=139|allele=A
CCTCGCCTAC TGTGCTGTTT CTAACCATCA TGCTTTTCCC TGAATCTCTT
GAGTCTTTTT CTGCTGTGGA CTGAAACTTG ATCCTGAGAT TCACCTCTAG
TCCCTCTGAG CAGCCTCCTG GAATACTCAG CTGGGATGG
>rs13218440.2|Chr.6|length=139|allele=G
CCTCGCCTAC TGTGCTGTTT CTAACCATCA TGCTTTTCCC TGAATCTCTT
GAGTCTTTTT CTGCTGTGGA CTGAAACTTG ATCCTGAGAT TCACCTCTAG
TCCCTCTGGG CAGCCTCCTG GAATACTCAG CTGGGATGG

>rs4606077.1|Chr.8|length=114|allele=C
GCAACTCCCT CAACTCCAAG GCAGACACCA AAGCCCTCCC TGCCTGTGGC
TTTGTAGTTC TAGTGTGGGA TCTGACTCCC CACAGCCCAC CCAAAGCCGG
GGAACTCCTC ACTG
>rs4606077.2|Chr.8|length=114|allele=T
GCAACTCCCT CAACTCCAAG GCAGACACCA AAGCCCTCCC TGCCTGTGGC
TTTGTAGTTC TAGTGTGGGA TCTGACTCCC CACAGCCTAC CCAAAGCCGG
GGAACTCCTC ACTG

[0378] >rs7041158.1|Chr.9|length=117|allele=C
AATTGCAATG GTGAGAGGTT GATGGTAAAA TCAAACGGAA CTTGTTATTT
TGTCATTCTG ATGGACTGGA ACTGAGGATT TTCAATTTCC TCTCCAACCC
AAGACACTTC TCACTGG
>rs7041158.2|Chr.9|length=117|allele=T
AATTGCAATG GTGAGAGGTT GATGGTAAAA TCAAACGGAA CTTGTTATTT
TGTCATTCTG ATGGACTGGA ACTGAGGATT TTCAATTTCC TTTCCAACCC
AAGACACTTC TCACTGG

>rs740598.1|Chr.10|length=114|allele=A
GAAATGCCTT CTCAGGTAAT GGAAGGTTAT CCAAATATTT TTCGTAAGTA
TTTCAAATAG CAATGGCTCG TCTATGGTTA GTCTCACAGC CACATTCTCA
GAACTGCTCA AACC
>rs740598.2|Chr.10|length=114|allele=G
GAAATGCCTT CTCAGGTAAT GGAAGGTTAT CCAAATATTT TTCGTAAGTA
TTTCAAATAG CAATGGCTCG TCTATGGTTA GTCTCGCAGC CACATTCTCA
GAACTGCTCA AACC

>rs10773760.1|Chr.12|length=128|allele=A

ACCCAAAACA CTGGAGGGGC CTCTTCTCAT TTTCGGTAGA CTGCAAGTGT
TAGCCGTCGG GACCAGCTTC TGTCTGGAAG TTCGTCAAAT TGCAGTTAAG
TCCAAGTATG CCACATAGCA GATAAGGG
>rs10773760.2|Chr.12|length=128|allele=G
ACCCAAAACA CTGGAGGGGC CTCTTCTCAT TTTCGGTAGA CTGCAAGTGT
TAGCCGTCGG GACCAGCTTC TGTCTGGAAG TTCGTCAAAT TGCAGTTAGG
TCCAAGTATG CCACATAGCA GATAAGGG

>rs4530059.1|Chr.14|length=110|allele=A
GCACCAGAAT TTAAACAACG CTGACAATAA ATATGCAGTC GATGATGACT
TCCCAGAGCT CCAGAAGCAA CTCCAGCACA CAGAGAGGCG CTGATGTGCC
TGTCAGGTGC
>rs4530059.2|Chr.14|length=110|allele=G
GCACCAGAAT TTAAACAACG CTGACAATAA ATATGCAGTC GATGATGACT
TCCCAGAGCT CCAGAAGCAA CTCCAGCACA CGGAGAGGCG CTGATGTGCC
TGTCAGGTGC

>rs1821380.1|Chr.15|length=139|allele=C
[0379] GCCCAGATTA GATGGAACCT TTTCCTCTTT TCCAGTGCAA GACAAGCGAT
TGAAAGAAGT GGATGTGTTA TTGCGGGCAC AATGGAGCCA CTGAACTGCA
GTGCAAAAAT GCAGTAAGGC ATACAGATAG AAGAAGGAG
>rs1821380.2|Chr.15|length=139|allele=G
GCCCAGATTA GATGGAACCT TTTCCTCTTT TCCAGTGCAA GACAAGCGAT
TGAAAGAAGT GGATGTGTTA TTGCGGGCAC AATGGAGCCA CTGAACTGCA
GTGCAAAAAT GCAGTAAGGG ATACAGATAG AAGAAGGAG
>rs7205345.1|Chr.16|length=116|allele=C

TGACTGTATA CCCCAGGTGC ACCCTTGGGT CATCTCTATC ATAGAACTTA
TCTCACAGAG TATAAGAGCT GATTTCTGTG TCTGCCTCTC ACACTAGACT
TCCACATCCT TAGTGC
>rs7205345.2|Chr.16|length=116|allele=G
TGACTGTATA CCCCAGGTGC ACCCTTGGGT CATCTCTATC ATAGAACTTA
TCTCACAGAG TATAAGAGCT GATTTCTGTG TCTGCCTGTC ACACTAGACT
TCCACATCCT TAGTGC

>rs8078417.1|Chr.17|length=110|allele=C
TGTACGTGGT CACCAGGGGA CGCCTGGCGC TGCGAGGGAG GCCCCGAGCC
TCGTGCCCCC GTGAAGCTTC AGCTCCCCTC CCCGGCTGTC CTTGAGGCTC

TTCTCACACT

>rs8078417.2|Chr.17|length=110|allele=T

TGTACGTGGT CACCAGGGGA CGCCTGGCGC TGCGAGGGAG GCCCCGAGCC

TCGTGCCCCC GTGAAGCTTC AGCTCCCCTC CCTGGCTGTC CTTGAGGCTC

TTCTCACACT

>rs576261.1|Chr.19|length=114|allele=A

CAGTGGACCC TGCTGCACCT TTCCTCCCCT CCCATCAACC TCTTTTGTGC

CTCCCCCTCC GTGTACCACC TTCTCTGTCA CCAACCCTGG CCTCACAACT

CTCTCCTTTG CCAC

>rs576261.2|Chr.19|length=114|allele=C

CAGTGGACCC TGCTGCACCT TTCCTCCCCT CCCATCAACC TCTTTTGTGC

CTCCCCCTCC GTGTACCACC TTCTCTGTCA CCACCCCTGG CCTCACAACT

CTCTCCTTTG CCAC

[0380]

>rs2567608.1|Chr.20|length=110|allele=A

CAGTGGCATA GTAGTCCAGG GGCTCCTCCT CAGCACCTCC AGCACCTTCC

AGGAGGCAGC AGCGCAGGCA GAGAACCCGC TGGAAGAATC GGCGGAAGTT

GTCGGAGAGG

>rs2567608.2|Chr.20|length=110|allele=A

CAGTGGCATA GTAGTCCAGG GGCTCCTCCT CAGCACCTCC AGCACCTTCC

AGGAGGCAGC AGCGCAGGCA GAGAACCCGC TGGAAGGATC GGCGGAAGTT

GTCGGAGAGG

>rs2073383.1|Chr.22|length=140|allele=C

GCTGCAGAAT CCACAGAGCC AGACGCCCCC TGGGCCCCCA GCGCCCCCCT

GCACAAGTGG GGAAACTAGG TCATGGGGCC CAGGCAGTGT GGAAGGCGTT

GCAGGAGTTG CCCAGGGCGT GGGGTCCTCC AGCCTCAGTG

>rs2073383.2|Chr.22|length=140|allele=T

GCTGCAGAAT CCACAGAGCC AGACGCCCCC TGGGCCCCCA GCGCCCCCCT

GCACAAGTGG GGAAACTAGG TCATGGGGCC CAGGCAGTGT GGAAGGCGTT

GCAGGAGTTG CCCAGGGTGT GGGGTCCTCC AGCCTCAGTG

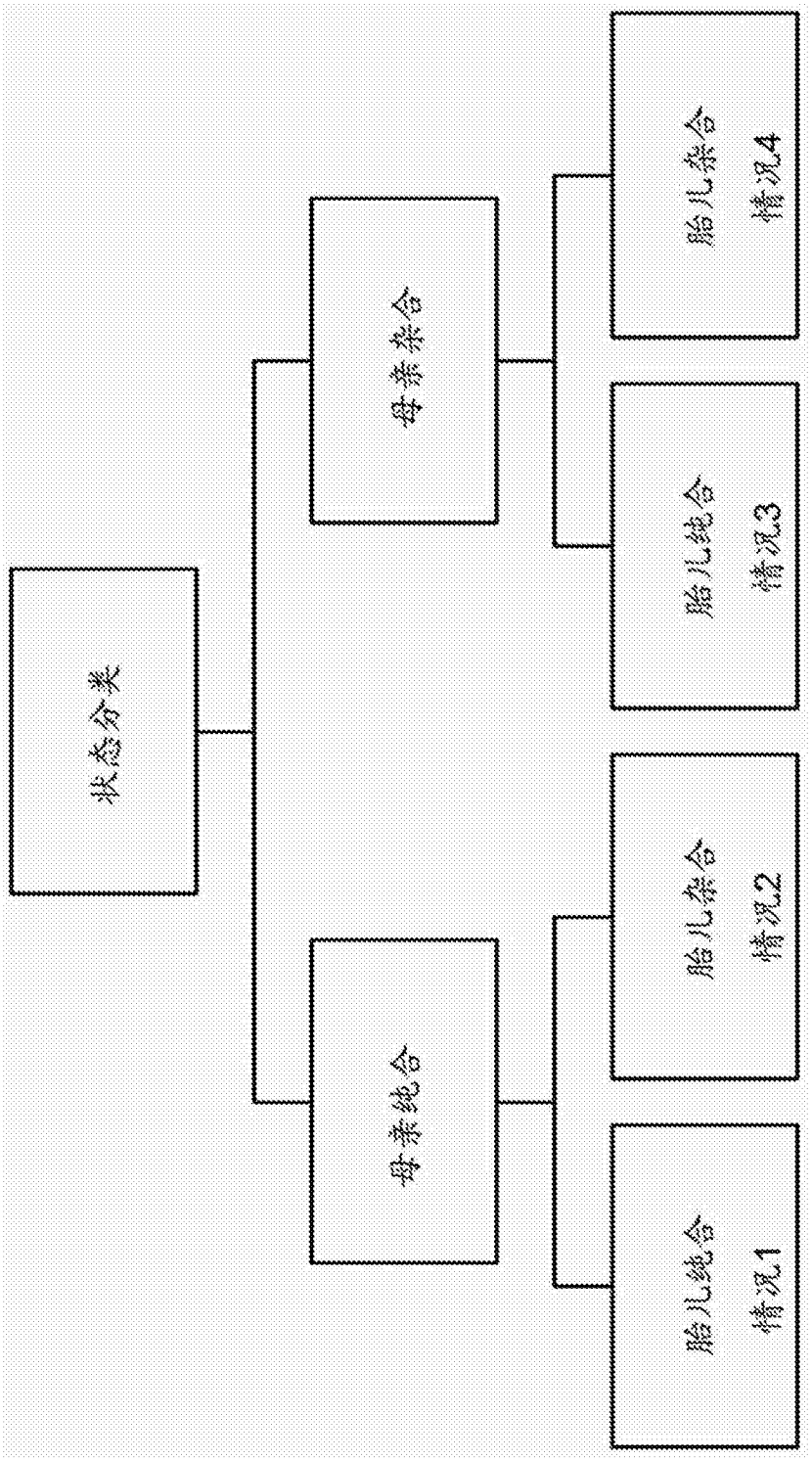


图1

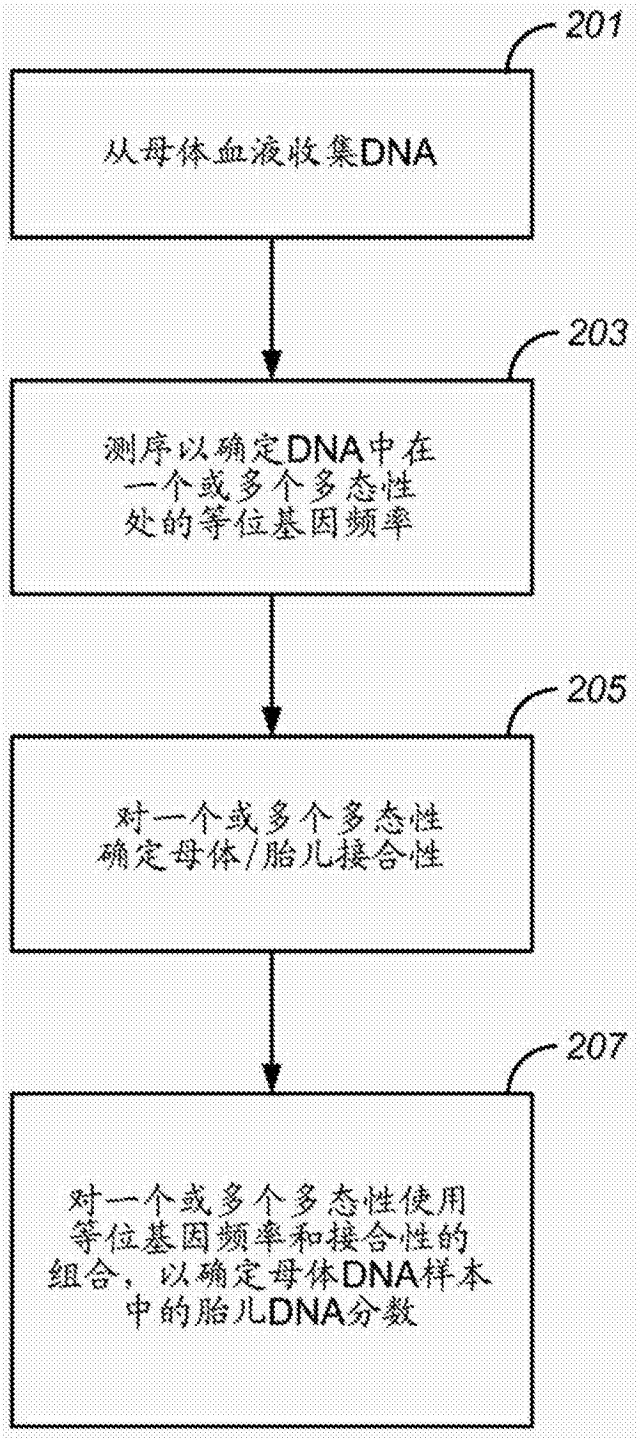

201
从母体血液收集DNA
203
测序以确定DNA中在一个或多个多态性处的等位基因频率
205
对一个或多个多态性确定母体/胎儿接合性
207
对一个或多个多态性使用等位基因频率和接合性的组合，以确定母体DNA样本中的胎儿DNA分数

图2

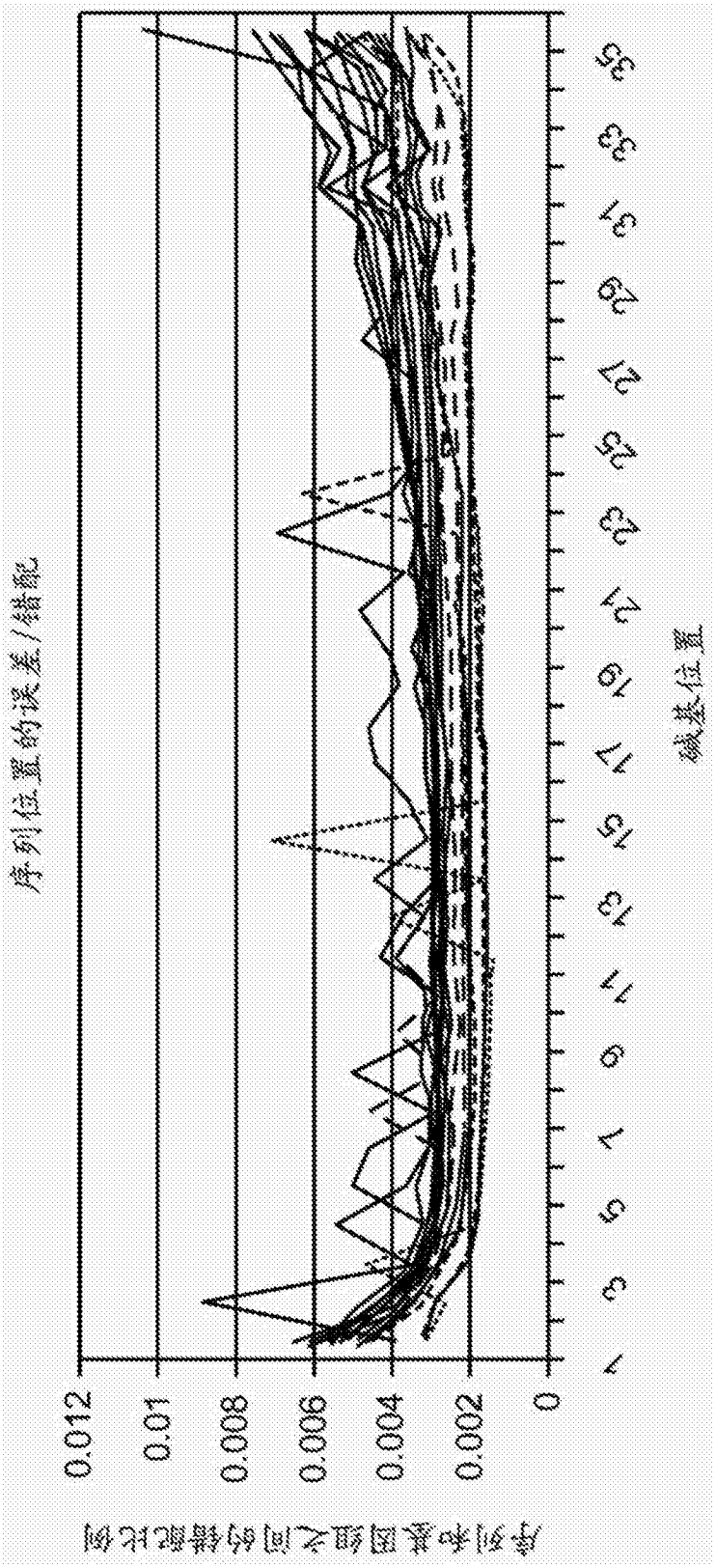

序列位置的误差/错配
0.012
0.01
0.008
0.006
0.004
0.002
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
碱基位置
序列和基因组之间的错配比例

图3

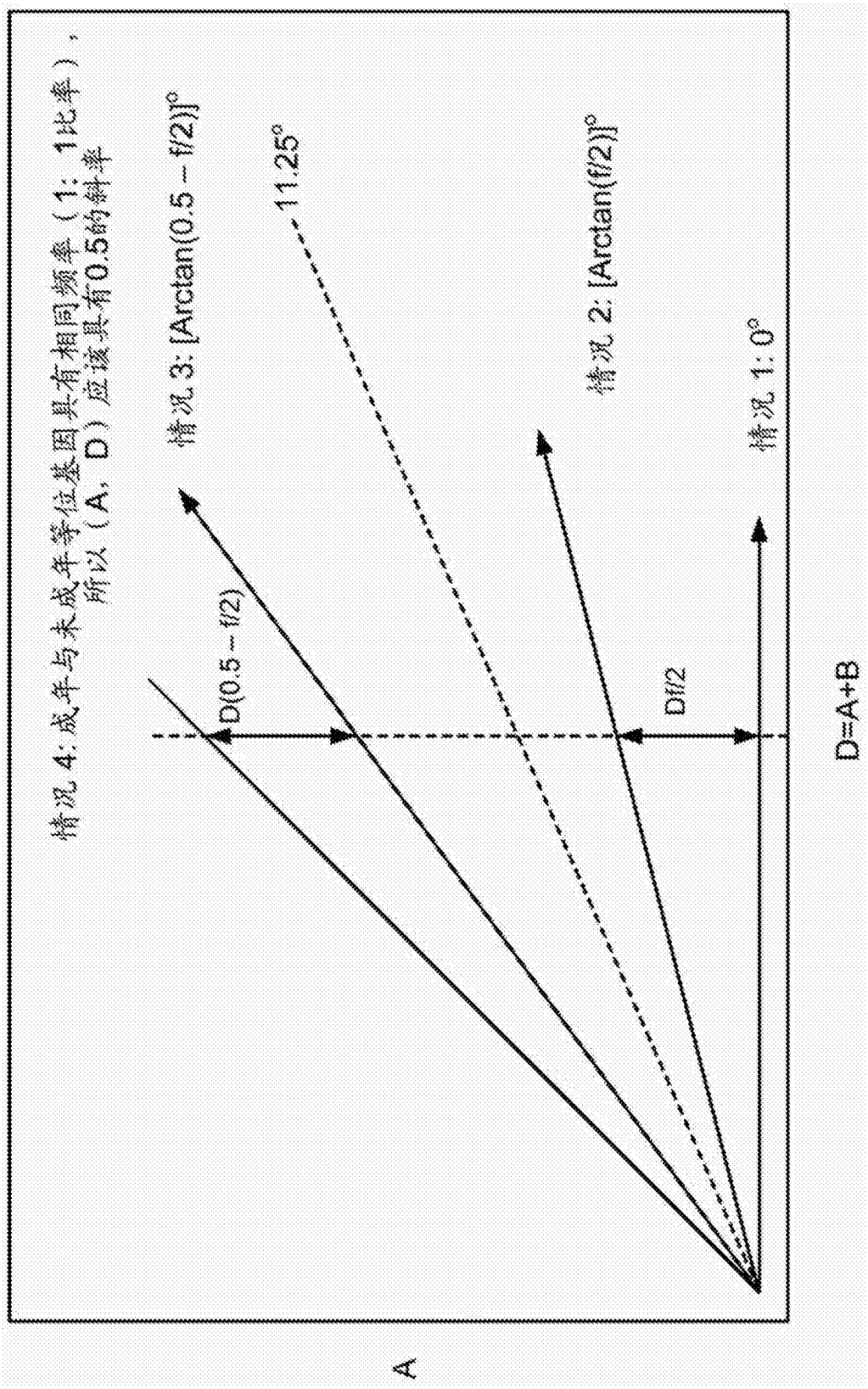

情况 4: 成年与未成年等位基因具有相同频率（1：1比率），
所以（A，D）应该具有0.5的斜率
情况 3: [Arctan(0.5 – f/2)]°
11.25°
情况 2: [Arctan(f/2)]°
情况 1: 0°
D(0.5 – f/2)
Df/2
D=A+B
A

图4

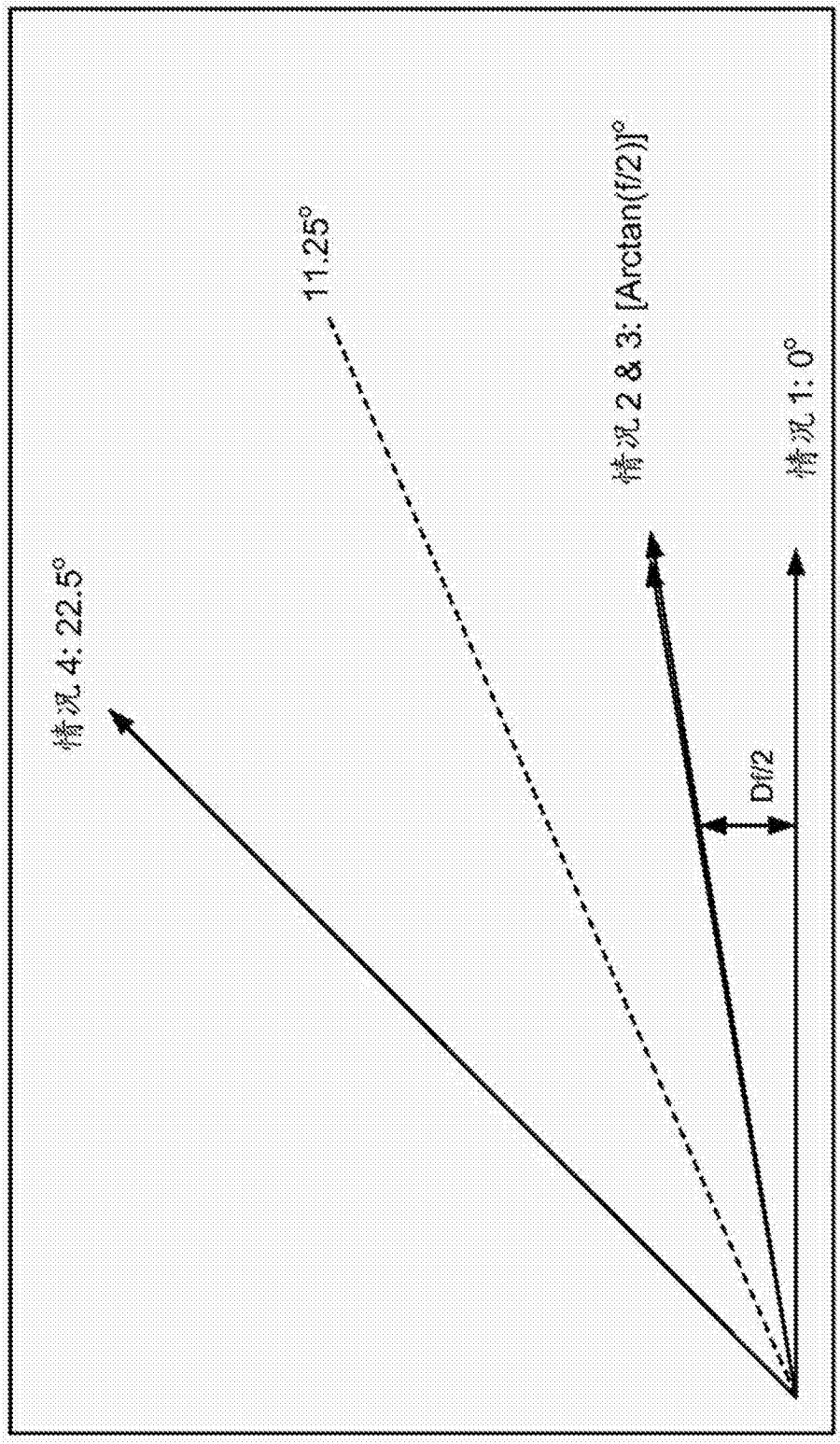


图5

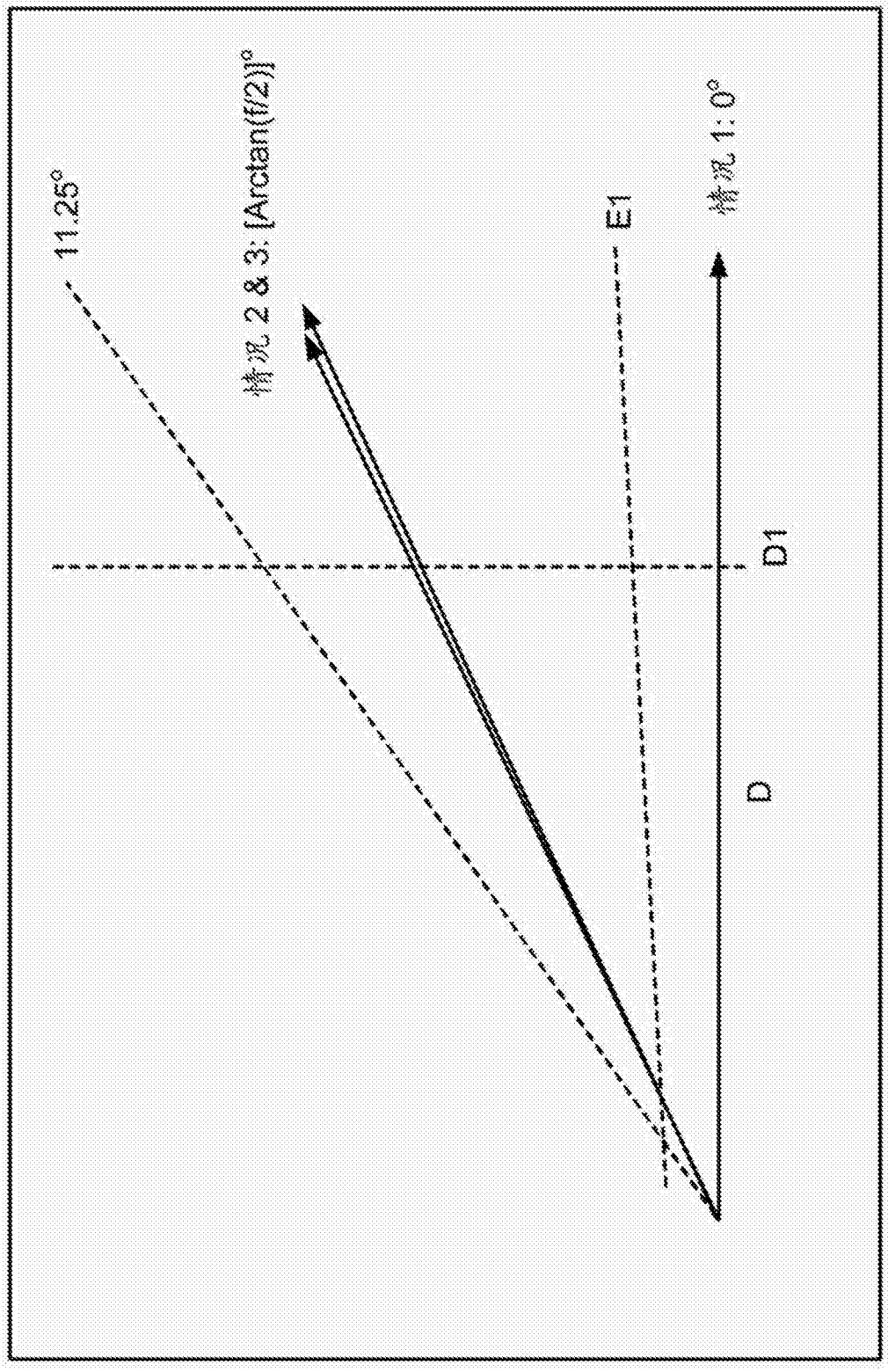


图6

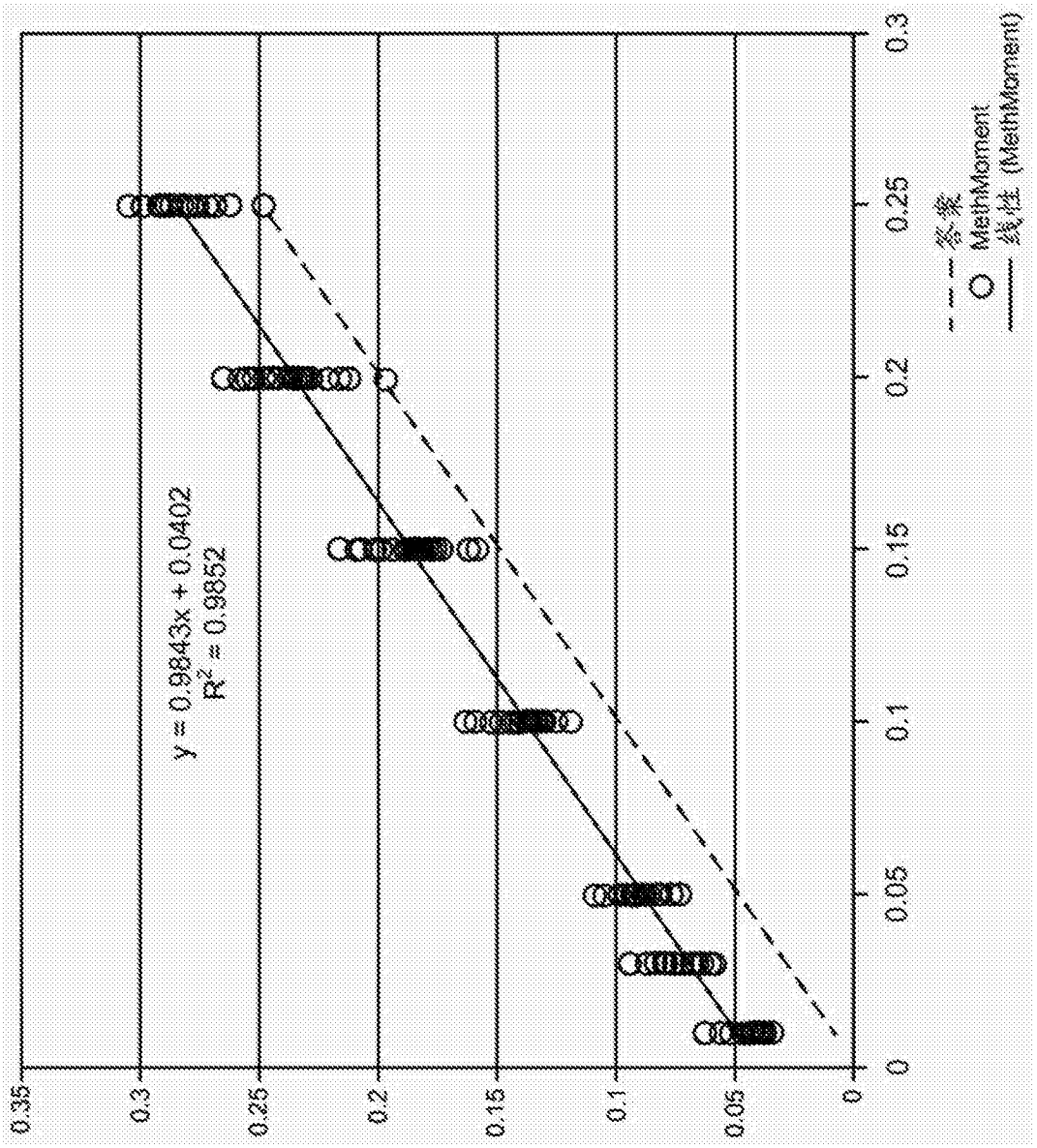

y = 0.9843x + 0.0402
R² = 0.9852
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.3
0.25
0.2
0.15
0.1
0.05
0
答案
MathMoment
线性 (MathMoment)

图7

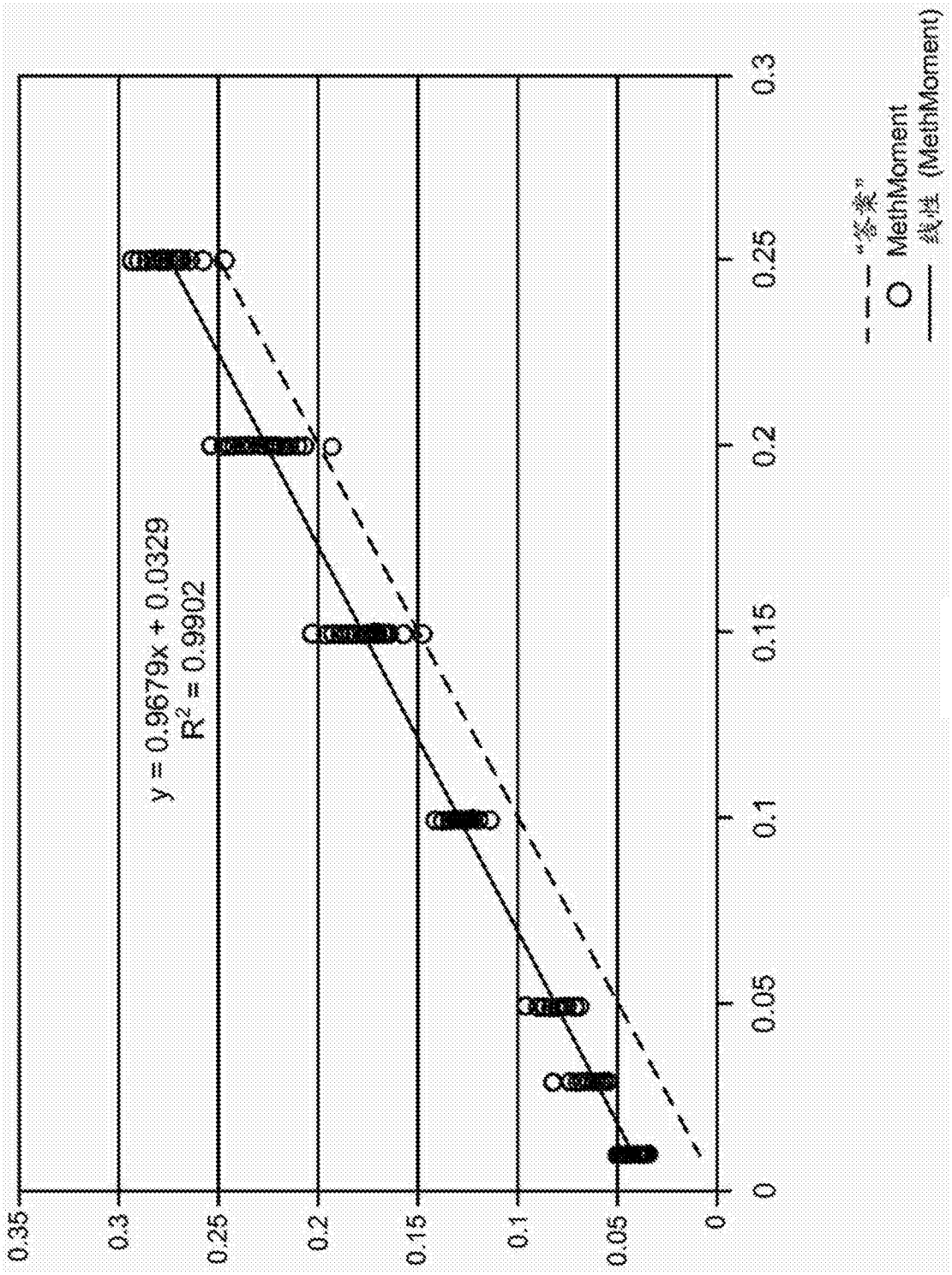

y = 0.9679x + 0.0329
R² = 0.9902
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
"答案"
MethMoment
线性 (MethMoment)

图8

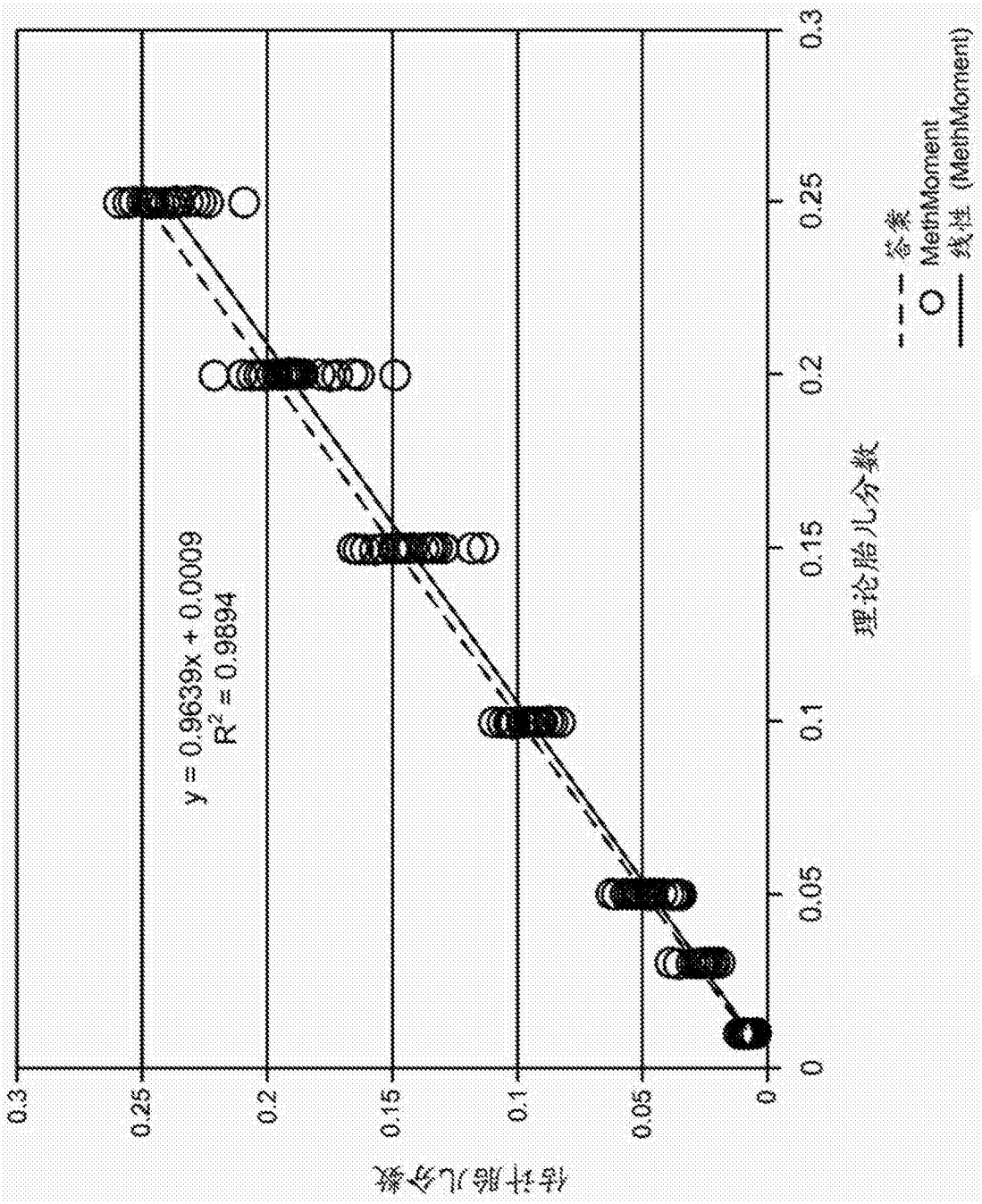

0.3
0.25
0.2
0.15
0.1
0.05
0
估计胎儿分数
理论胎儿分数
y = 0.9639x + 0.0009
R² = 0.9894
答案
MethMoment
线性 (MethMoment)

图9

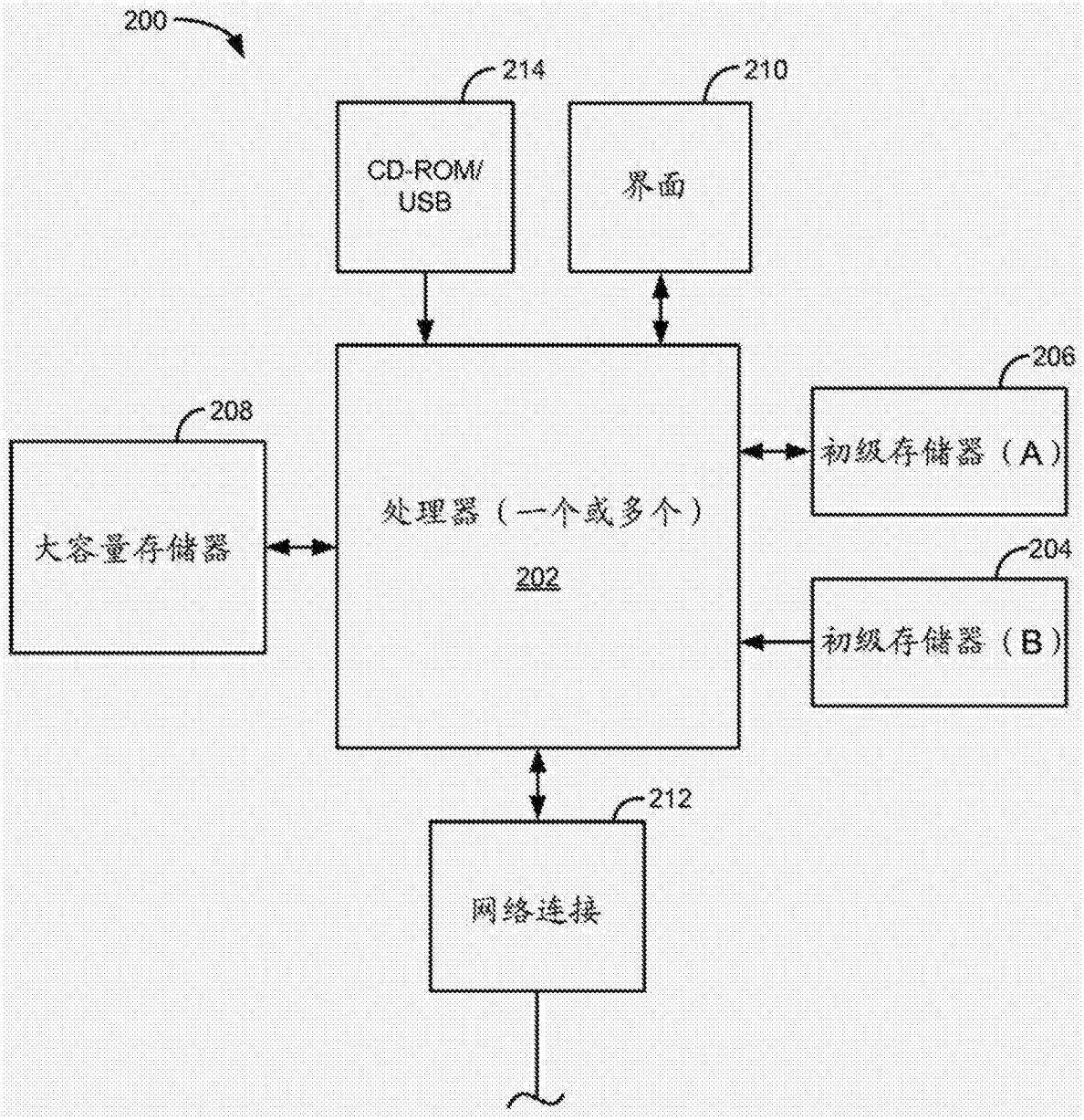

200
214
CD-ROM/
USB
210
界面
206
208
初级存储器（A）
处理器（一个或多个）
202
大容量存储器
204
初级存储器（B）
212
网络连接

图10

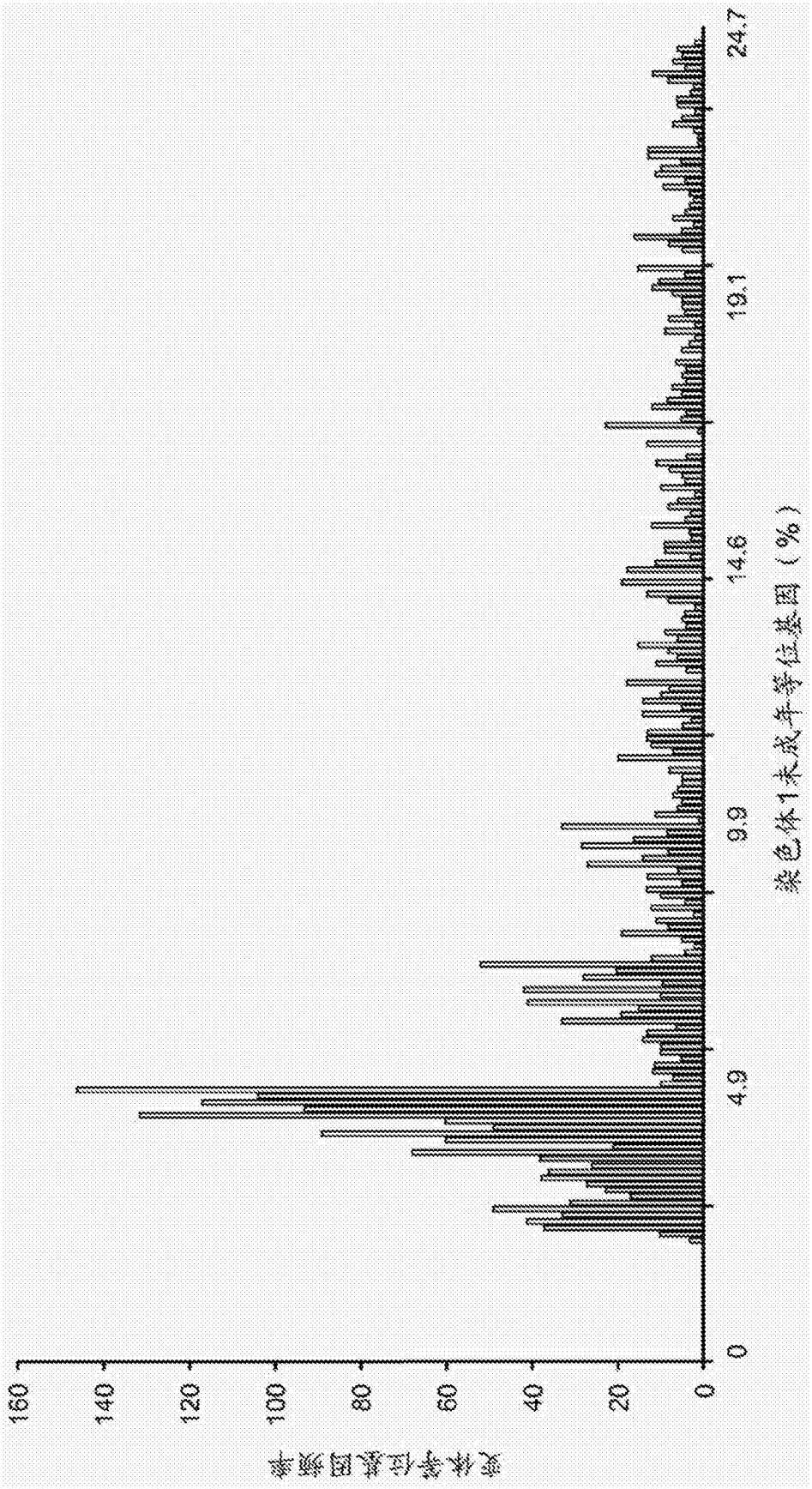


图11A

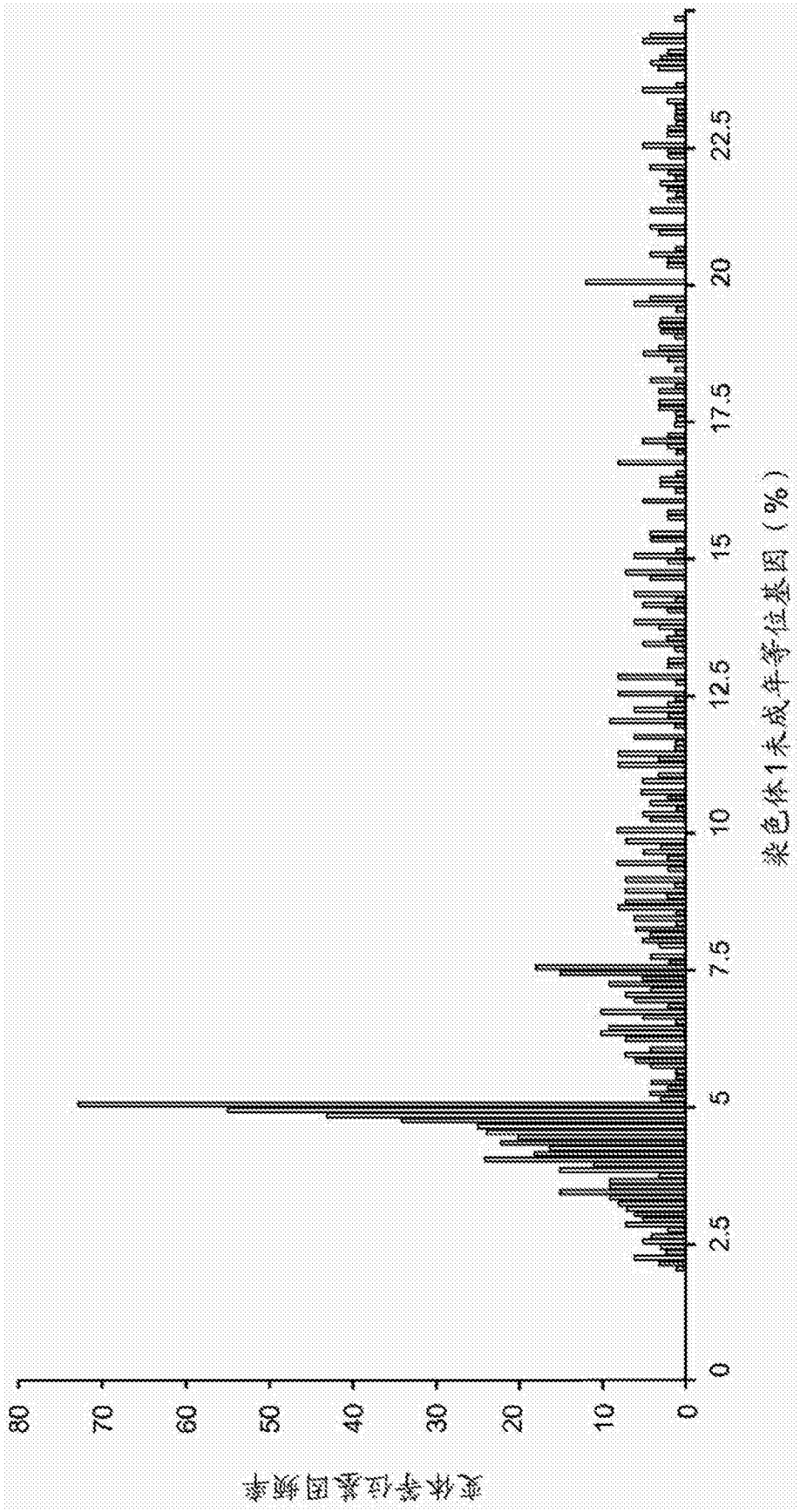

80
70
60
50
40
30
20
10
0
0
2.5
5
7.5
10
12.5
15
17.5
20
22.5
来色体1未成年等位基因（%）
来体等位基因频率

图11B

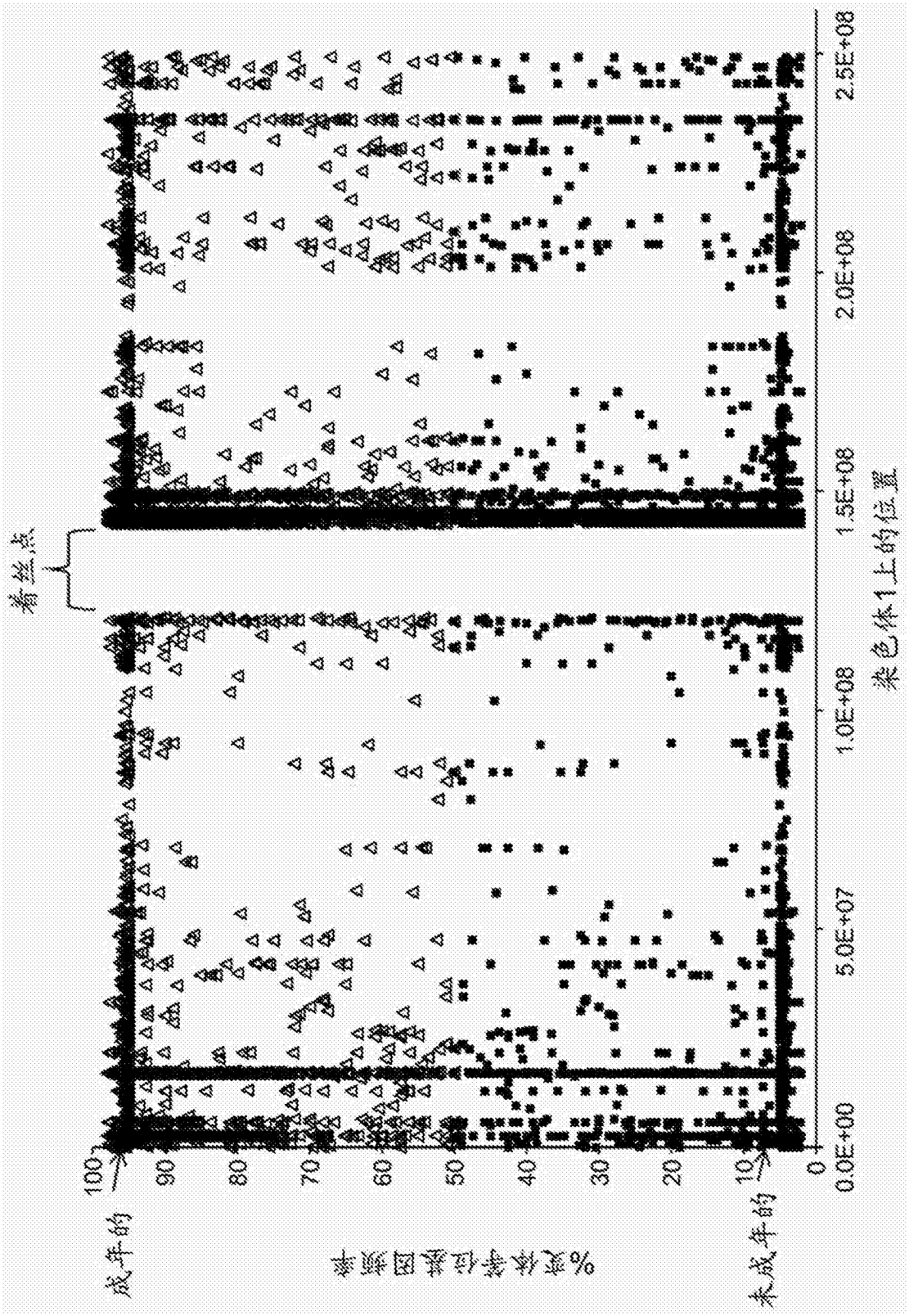

着丝点
染色体1上的位置
0.0E+00
5.0E+07
1.0E+08
1.5E+08
2.0E+08
2.5E+08
%染体条位基因频率
成年的
未成年的
100
90
80
70
60
50
40
30
20
10
0

图12A

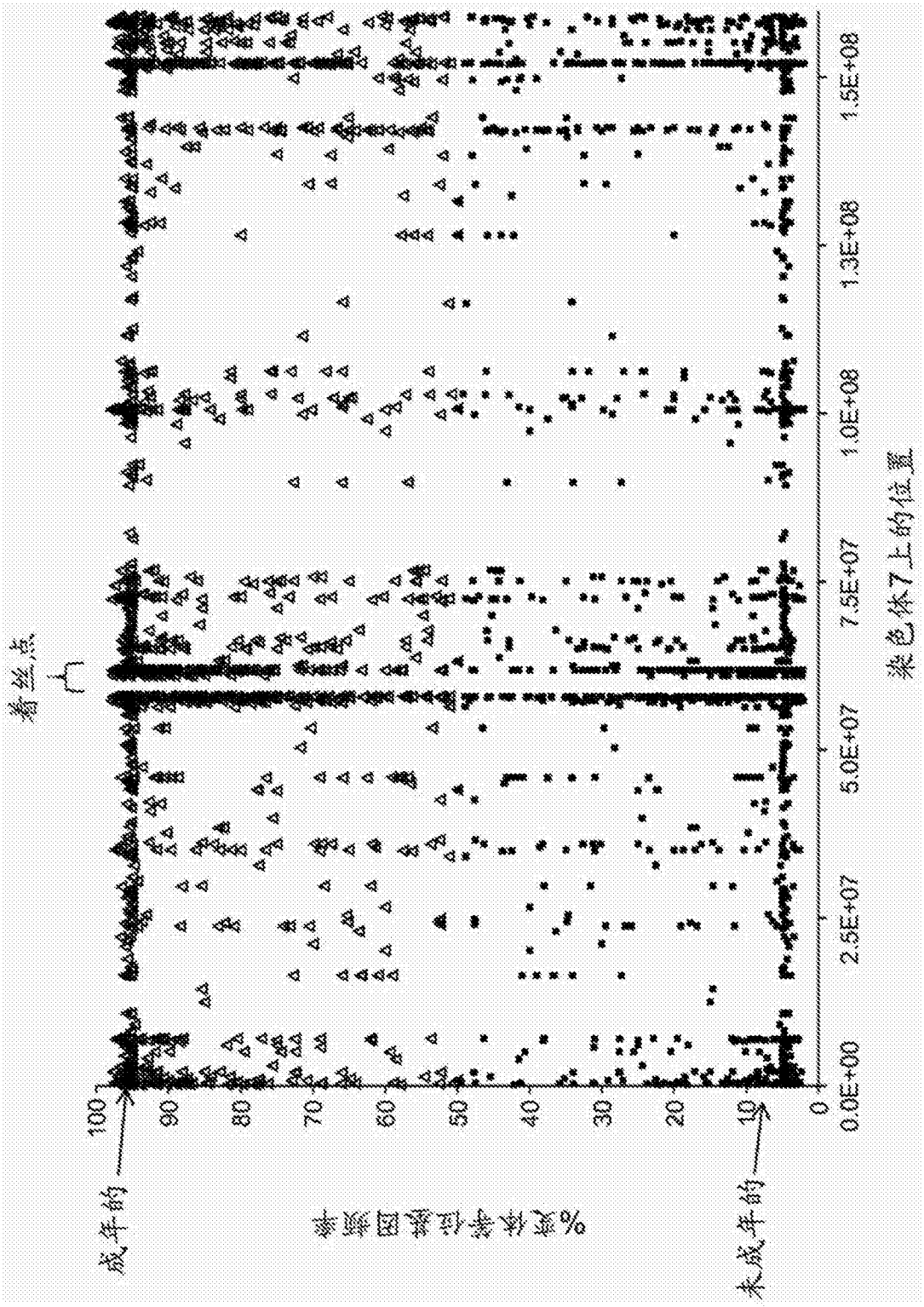

着丝点
染色体7上的位置
0.0E+00
2.5E+07
5.0E+07
7.5E+07
1.0E+08
1.3E+08
1.5E+08
100
90
80
70
60
50
40
30
20
10
0
%表体单位基因频率
成年的
未成年的

图12B