(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 107312844 B
(45)授权公告日 2021.01.22

(21)申请号 201710549501.9

(22)申请日 2010.11.05

(65)同一申请的已公布的文献号
申请公布号 CN 107312844 A

(43)申请公布日 2017.11.03

(30)优先权数据
61/259,076 2009.11.06 US
61/360,399 2010.06.30 US

(62)分案原申请数据
201080059269.7 2010.11.05

(73)专利权人 香港中文大学
地址 中国香港新界

(72)发明人 卢煜明 陈君赐 郑文莉 赵慧君

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 洪欣

(51)Int.Cl.
C12Q 1/6809(2018.01)
C12Q 1/6883(2018.01)
G16B 20/00(2019.01)

(56)对比文件
CN 1452665 A,2003.10.29
CN 1498276 A,2004.05.19
CN 1469932 A,2004.01.21
CN 101137760 A,2008.03.05
WO 2008070862 A3,2008.11.20
WO 2007028155 A3,2007.06.21
EP 1524321 A1,2005.04.20
CN 1798974 A,2006.07.05
PaigeB.Larrabee等.Microarray Analysis of Cell-Free Fetal DNA in Amniotic Fluid: A Prenatal Molecular Karyotyne.《Am.J.Hum.Genet》.2004,第485-491页.

审查员 童欣

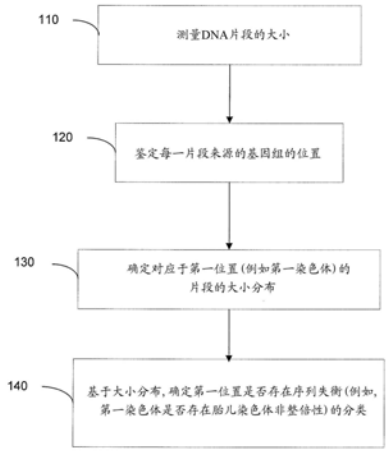
权利要求书3页 说明书27页 附图36页

(54)发明名称

基于大小的基因组分析

(57)摘要

提供了进行序列失衡的产前诊断的系统、方法和装置。移位(例如向较小大小分布的移位)可以表示某些情形下的失衡。例如,来自高危染色体的核酸片段的大小分布可以用于确定胎儿染色体非整倍性。不同染色体的大小秩次可以用于确定高危染色体秩次从预期秩次的变化。而且,可以将一条染色体的统计大小值间的差异与另一条染色体的统计大小值进行比较以鉴定大小的显著移位。还可以利用大小分布确定相对于母亲的基因型或单体型,母体样品中是否存在序列失衡,从而提供胎儿的基因型或单体型来确定胎儿的基因型和单体型。



CN 107312844 B

1.分析获自怀有胎儿的孕妇个体的生物样品的计算机系统,其中所述生物样品包括为核酸序列一部分的核酸分子,所述生物样品包括来自所述胎儿和所述孕妇个体的核酸分子,所述计算机系统包括:

对于所述生物样品中多个核酸分子的每一个:

测量所述核酸分子的大小的装置;和

鉴定所述核酸分子来源于哪一条核酸序列的装置;

从来自第一序列的核酸分子的大小计算第一统计值的装置;

鉴定具有与所述第一序列的GC含量类似的GC含量的一条或多条参照序列的装置;

从来自所述一条或多条参照序列的核酸分子的大小计算第二统计值的装置;

利用所述第一统计值和所述第二统计值确定参数的装置,所述参数是所述第一统计值与所述第二统计值的差值或比值;以及

基于所述参数与截止值的比较确定所述第一序列是否存在序列失衡分类的装置。

2.如权利要求1所述的计算机系统,其中用于测量所述多个核酸分子的每一个的大小的装置包括:

接收所述生物样品中包含的多个核酸分子的至少一部分的序列数据的装置,其中每个核酸分子的所述部分包含各自核酸分子的两个末端。

3.如权利要求1所述的计算机系统,其中所述第一统计值包括来自所述第一序列的核酸分子测量的大小的中值或平均大小。

4.如权利要求1所述的计算机系统,其中所述核酸分子的大小是长度或与长度有关的测量的参数。

5.如权利要求1所述的计算机系统,其中所述核酸分子的大小是分子量。

6.如权利要求1所述的计算机系统,其中所述核酸分子的大小是对应于长度的荧光强度。

7.如权利要求1所述的计算机系统,其中所述第一序列是染色体,以及所述序列失衡是胎儿染色体非整倍性。

8.如权利要求4所述的计算机系统,其中所述所述一条或多条参照序列是一条或多条参照染色体。

9.如权利要求4所述的计算机系统,其中所述一条或多条参照序列包含一条染色体。

10.如权利要求4所述的计算机系统,其中所述一条或多条参照序列是多条染色体。

11.如权利要求1所述的计算机系统,其还包含:

计算所述一条或多条参照序列的GC含量的装置;以及

计算所述第一序列的GC含量的装置。

12.如权利要求1所述的计算机系统,其中鉴定所述核酸分子来源于哪一条核酸序列的装置包括:

从所述核酸分子的至少一部分的序列数据获得序列的装置;以及

将所述序列与人基因组对齐的装置。

13.如权利要求8所述的计算机系统,其中获得所述一条或多条参照序列的GC含量以及所述第一序列的GC含量用于具体的测序平台。

14.如权利要求1所述的计算机系统,其中所述生物样品包括血液、血浆、血清、获自母

体血液的胎儿细胞、尿液、唾液或子宫灌洗液。

15.如权利要求1所述的计算机系统，其中所述生物样品包括含胎儿细胞的母体血液。

16.计算机可读介质，其存储用于控制处理器来分析获自怀有胎儿的孕妇个体的生物样品的多个指令，其中所述生物样品包括为核酸序列一部分的核酸分子，所述生物样品包括来自所述胎儿和所述孕妇个体的核酸分子，所述多个指令包括：

对于所述生物样品中多个核酸分子的每一个：

测量所述核酸分子的大小；和

鉴定所述核酸分子来源于哪一条核酸序列；

从来自第一序列的核酸分子的大小计算第一统计值；

鉴定具有与所述第一序列的GC含量类似的GC含量的一条或多条参照序列；

从来自所述一条或多条参照序列的核酸分子的大小计算第二统计值；

利用所述第一统计值和所述第二统计值确定参数，所述参数是所述第一统计值与所述第二统计值的差值或比值；以及

基于所述参数与截止值的比较确定所述第一序列是否存在序列失衡分类。

17.如权利要求16所述的计算机可读介质，其中测量所述多个核酸分子的每一个的大小包括：

接收所述生物样品中包含的多个核酸分子的至少一部分的序列数据，其中每个核酸分子的所述部分包含各自核酸分子的两个末端。

18.如权利要求16所述的计算机可读介质，其中所述第一统计值包括来自所述第一序列的核酸分子测量的大小的中值或平均大小。

19.如权利要求16所述的计算机可读介质，其中所述核酸分子的大小是长度或与长度有关的测量的参数。

20.如权利要求16所述的计算机可读介质，其中所述核酸分子的大小是分子量。

21.如权利要求16所述的计算机可读介质，其中所述核酸分子的大小是对应于长度的荧光强度。

22.如权利要求16所述的计算机可读介质，其中所述第一序列是染色体，以及所述序列失衡是胎儿染色体非整倍性。

23.如权利要求20所述的计算机可读介质，其中所述所述一条或多条参照序列是一条或多条参照染色体。

24.如权利要求20所述的计算机可读介质，其中所述一条或多条参照序列包含一条染色体。

25.如权利要求20所述的计算机可读介质，其中所述一条或多条参照序列是多条染色体。

26.如权利要求16所述的计算机可读介质，所述多个指令还包含：

计算所述一条或多条参照序列的GC含量；以及

计算所述第一序列的GC含量。

27.如权利要求16所述的计算机可读介质，其中鉴定所述核酸分子来源于哪一条核酸序列包括：

从所述核酸分子的至少一部分的序列数据获得序列；以及

将所述序列与人基因组对齐。

28.如权利要求24所述的计算机可读介质，其中获得所述一条或多条参照序列的GC含量以及所述第一序列的GC含量用于具体的测序平台。

29.如权利要求16所述的计算机可读介质，其中所述生物样品包括血液、血浆、血清、获自母体血液的胎儿细胞、尿液、唾液或子宫灌洗液。

30.如权利要求16所述的计算机可读介质，其中所述生物样品包括含胎儿细胞的母体血液。

基于大小的基因组分析

[0001] 相关申请的交叉引用

[0002] 本申请要求以下美国临时申请的优先权并且是其非临时申请:于2009年11月6日提交、题目为“染色体失衡的检测(Detection of Chromosomal Imbalance)”的美国临时申请61/259076以及于2010年6月30日提交、题目为“基于大小的基因组分析(Size-Based Genomic Analysis)”的美国临时申请61/360399,其全部内容出于所有目的通过引用并入本文。

[0003] 本申请还涉及于2008年7月23日提交、题目为“利用大规模平行基因组测序诊断胎儿染色体非整倍性(Diagnosing Fetal Chromosomal Aneuploidy Using Massively Parallel Genomic Sequencing)”的美国申请12/178,181(代理案卷号:016285-005220US)、题目为“利用富集基因组测序诊断胎儿染色体非整倍性(Diagnosing Fetal Chromosomal Aneuploidy Using Genomic Sequencing With Enrichment)”的美国申请12/614350(代理案卷号:016285-005221US)、以及同时提交的题目为“来自母体生物样品的胎儿基因组分析(Fetal Genomic Analysis From A Maternal Biological Sample)”的美国申请(代理案卷号:016285-006710US),这些申请的全部内容出于所有目的通过引用并入本文。

[0004] 背景

[0005] 1997年对母体血浆中胎儿DNA的发现已经开启了新的无创性产前诊断的可能性(Lo et al Lancet 1997;350:485-487)。这一技术已经被快速转化为临床应用,检测胎儿来源的、父代遗传的基因或序列,例如用于胎儿性别确定和用于胎儿RhD状态确定。然而,涉及既存在于母体又存在于胎儿基因组的基因组靶标例如染色体21的产前诊断应用更具挑战性。

[0006] 最近,已表明具有较高定量精确度的单分子计数技术,可能是这一问题的有前景的解决方案(Lo et al Proc Natl Acad Sci USA 2007;104:13116-13121;Fan et al Anal Chem 2007;79:7576-7579;美国专利申请11/701,686;Chiu et al Trends Genet 2009;25:324-331;Chiu et al Proc Natl Acad Sci USA 2008;105:20458-20463;Fan et al Proc Natl Acad Sci USA 2008;105:16266-16271)。这种方法通过观测来自所选基因组位置的分子数目在疾病和健康间的定量差异实现诊断目标。例如,对于胎儿唐氏综合症的诊断,当胎儿罹患三体21时,染色体21的分子数会增加(Chiu et al Proc Natl Acad Sci USA 2008;105:20458-20463;Fan et al Proc Natl Acad Sci USA 2008;105:16266-16271)。

[0007] 然而,这种计数技术可能受有限数目的数据点或其他缺点的影响。因此,需要提供具有优于现存技术的某些优势的进行产前诊断的新方法、系统和装置。

[0008] 概述

[0009] 本发明的某些实施方案能够提供可以利用基于大小的分析在获自孕妇个体的生物样品中进行序列失衡(例如,胎儿染色体非整倍性)的产前诊断的系统、方法和装置。例如,高危染色体的核酸分子片段的大小分布可以用于确定胎儿染色体的非整倍性。一些实

施方案也可以检测其它序列失衡，诸如生物样品(包含母亲和胎儿DNA)中的序列失衡，其中所述失衡是相对于母亲的基因型、突变状态或单体型。如果样品完全来自母亲，而不是来自胎儿和母亲，可以相对于预期大小分布通过对应于特定序列的片段(核酸分子)的大小分布来确定这种失衡。移位(例如到较小的大小分布)可以表示在某些情形下的失衡。

[0010] 在一个实施方案中，来自各个染色体的片段相对于彼此的大小分布(例如，代表大小分布的统计值)的秩次(ranking)用于确定失衡。例如，测试样品中高危染色体的片段大小的秩次可以与获自参照生物样品的高危染色体的秩次相比较。可以基于比较进行诊断。例如，如果秩次变化了(例如，指示核酸片段大小的降低)指定量，那么可以作出高危染色体中存在胎儿染色体非整倍性的诊断。在利用这种秩次分析的多个实施方案中，可以利用所有22个常染色体和性染色体，或者可以利用染色体的子集。

[0011] 在另一实施方案中，利用高危染色体片段的大小(例如，代表大小分布的统计值)与参照染色体片段的大小之间的差值。例如，如果大小的差值大于或小于截止值(也称为阈值)，则可以作出高危染色体中存在胎儿染色体非整倍性的诊断。

[0012] 根据一个实施例方案，提供了在获自孕妇个体的生物样品中进行序列失衡的产前诊断的方法。生物样品包括核酸分子，其是核酸序列的一部分。对于生物样品中多个核酸分子中的每个，测量核酸分子的大小，以及鉴定核酸分子来源于哪一核酸序列。计算机系统确定了对应于第一序列的核酸分子的大小分布。基于确定的大小分布，确定第一序列是否存在序列失衡的分类。

[0013] 根据另一实施例方案，提供了在获自孕妇个体的生物样品中进行胎儿染色体非整倍性产前诊断的方法。对于生物样品中多个核酸分子中的每个，测量核酸分子的大小，以及鉴定核酸分子来源于哪条染色体。计算机系统从对应于染色体的核酸分子的大小计算统计值。计算多条染色体的每条的统计值。基于统计值确定染色体的秩次。将确定的第一染色体的秩次与获自参照生物样品的第一染色体的另一秩次比较。基于比较，确定第一染色体是否存在胎儿染色体非整倍性的分类。根据另一实施例方案，提供了在获自孕妇个体的生物样品中进行胎儿染色体非整倍性产前诊断的方法。对于生物样品中多个核酸分子中的每个，测量核酸分子的大小，并鉴定核酸分子来源于哪条染色体。计算机系统从对应于第一染色体的核酸分子的大小计算第一统计值。计算机系统从对应于一条或多条第二染色体的核酸分子的大小计算第二统计值。确定第一统计值和第二统计值间的间隔值(separation value)。将间隔值与一个或多个截止值相比较。基于比较，确定第一染色体是否存在胎儿染色体非整倍性的分类。

[0014] 本发明的其他实施方案涉及与本文描述的方法有关的系统和计算机可读介质。在一个实施方案中，计算机可读介质包含接受数据和分析数据的指令，而不是指导机器产生数据(例如，测序核酸分子)的指令。在另一实施方案中，计算机可读介质的确包含指导机器产生数据的指令。在一个实施方案中，计算机程序产品包含储存多个指令的计算机可读介质，所述指令用于控制处理器以执行本文描述的方法的操作。实施方案还涉及被配置为执行本文描述的任一方法的步骤的计算机系统，其可能具有执行单独的步骤或单独的步骤组的不同部件。

[0015] 参考包括附图和权利要求书的说明书的其他部分会认识到本发明的实施方案的其他特征和优势。本发明的其他特征和优势，以及多种实施方案的结构和操作在下文结合

附图详细描述。在附图中，相同的参考号可以表示相同的或功能类似的元件。

[0016] 附图简述

[0017] 图1是流程图，说明了按照本发明的实施方案在获自孕妇个体的生物样品中进行序列失衡产前诊断的方法100。

[0018] 图2是显示了按照本发明的实施方案，在与染色体对齐的序列大小方面作出的染色体的中值秩次(当使用第1版的Illumina基因簇生成试剂盒(Cluster Generation Reagent Kit)时)的图。

[0019] 图3是证明了按照本发明的实施方案来自母体血浆中不同染色体序列的大小分析可以用于胎儿染色体21非整倍性无创性产前检测的图。

[0020] 图4是流程图，说明了按照本发明的实施方案，利用大小统计值的秩次在获自孕妇个体的生物样品中进行胎儿染色体非整倍性的产前诊断的方法。

[0021] 图5是表，说明了按照本发明的实施方案，染色体21与染色体7和染色体14在与它们对齐的序列大小方面的比较。

[0022] 图6是流程图，说明了按照本发明的实施方案，利用基因组位置的片段大小的统计值比较，在获自孕妇个体的生物样品中进行序列失衡的产前诊断的方法。

[0023] 图7显示了按照本发明的实施方案，由短片段贡献的总长度分数(F)相对于截止大小(w)的图700。

[0024] 图8显示了按照本发明的实施方案，对于整倍体孕妇，染色体21(实线)和参照染色体(除染色体13、18和21外的所有常染色体)(点划线)的短片段贡献的总长度分数(F)相对于截止大小的图800。

[0025] 图9显示了按照本发明的实施方案，对于21三体孕妇，染色体21(实线)和参照染色体(除染色体13、18和21外的所有常染色体)(点划线)的F值相对于截止大小的图900。

[0026] 图10显示了按照本发明的实施方案，对于整倍体和21三体孕妇，染色体21与参照染色体(除染色体13、18和21外的所有常染色体)间的$\Delta F_{(21-Ref)}$相对于大小截止的图1000。

[0027] 图11显示了按照本发明的实施方案，120个孕妇的性别和分类的表1100。

[0028] 图12说明了按照本发明的实施方案，不同疾病或无疾病状态的四种样品的不同染色体的秩次。

[0029] 图13显示了按照本发明的实施方案，120例整倍体、13三体、18三体和21三体的染色体13的秩次。

[0030] 图14显示了按照本发明的实施方案，120例整倍体、13三体、18三体和21三体的染色体18的秩次。.

[0031] 图15显示了按照本发明的实施方案，120例整倍体、13三体、18三体和21三体的染色体21的秩次。

[0032] 图16是表，说明了按照本发明的实施方案，染色体13与染色体5和染色体6在与它们对齐的序列大小方面的比较。染色体5和染色体6与染色体13的比较用于检测在13三体孕妇中染色体21序列大小的变化。包括来自整倍体和18三体孕妇的结果用于比较。

[0033] 图17是表，说明了按照本发明的实施方案，染色体18与染色体12和染色体14在与它们对齐的序列大小方面的比较。按照本发明的实施方案，包括来自整倍体和13三体孕妇的结果用于比较。

[0034] 图18显示了按照本发明的实施方案，在150bp处染色体18和参照染色体间由短片段贡献的总长度分数的差值（$\Delta F_{(18-Ref)}$）。

[0035] 图19显示了按照本发明的实施方案，在150bp处染色体21和参照染色体间由短片段贡献的总长度分数的差值（$\Delta F_{(21-Ref)}$）。

[0036] 图20显示了按照本发明的实施方案，不同染色体(NCBI build 36,第48版)的GC含量列表。

[0037] 图21显示了按照本发明的实施方案，在150bp处染色体13和参照染色体间由短片段贡献的总长度分数的差值（$\Delta F_{(18-Ref)}$）。

[0038] 图22是显示了按照本发明的实施方案，与染色体21对齐的序列的中值大小和与Y染色体对齐的序列百分比间的相互关系图。

[0039] 图23A-23C是显示了按照本发明的实施方案，分别与染色体18、13和21对齐的序列的中值大小和与Y染色体对齐的序列百分比间的相互关系图。

[0040] 图24显示了利用母体血浆DNA分析，对本发明的实施方案与无创性检测胎儿非整倍性(13三体和18三体)另一方法的的精确度的比较。.

[0041] 图25A-25C显示了按照本发明的实施方案，孕妇和胎儿基因型的不同情形的图解。

[0042] 图26显示了按照本发明实施方案的一个实例，其中母亲是杂合的，父亲是纯合的。

[0043] 图27显示了按照本发明实施方案的一个实例，其中当亲代单体型如图26所显示时，胎儿从母亲遗传了Hap I。

[0044] 图28显示了表，说明了按照本发明的实施方案，染色体22上的α型单核苷酸多态性(SNP)的大小分析。

[0045] 图29显示了表，说明了按照本发明的实施方案，染色体22上的β型SNP的大小分析。

[0046] 图30显示了按照本发明的实施方案，染色体22上的α型和β型SNP的$\Delta F_{(Hap\ I-Hap\ II)}$的图。

[0047] 图31A是提供了按照本发明的实施方案，没有靶标富集的血浆DNA的大小分析的表。

[0048] 图31B是提供了按照本发明的实施方案，有靶标富集的血浆DNA的大小分析的表。

[0049] 图32是按照本发明的实施方案，有和无靶标富集的T21和整倍体样品的ΔF的图。

[0050] 图33显示了可用于按照本发明的实施方案的系统和方法的示例性计算机装置的方框图。

[0051] 定义

[0052] 本文使用的术语“生物样品”指取自个体(例如，人类，诸如孕妇)并且包含一种或多种感兴趣的核酸分子的任何样品。

[0053] 术语“核酸”或“多核苷酸”指脱氧核糖核酸(DNA)或核糖核酸(RNA)及其单链或双链形式的聚合物。除非特别限定，该术语包括含天然核苷酸已知类似物的核酸，所述类似物具有与参照核酸类似的结合特性并以与天然存在的核苷酸类似的方式代谢。除非另外指明，特定核酸序列还暗含其保守性修饰的变体(例如，简并密码子取代物)、等位基因、直系同源物、SNP、拷贝数变体和互补序列以及明确指明的序列。特别地，简并密码子取代可以通过产生这样的序列来实现：在该序列中一个或多个选择的(或全部)密码子的第三位被混合碱基和/或脱氧肌苷残基取代(Batzer et al.,Nucleic Acid Res.19:5081(1991)；

Ohtsuka et al.,J.Biol.Chem.260:2605-2608(1985)；和Rossolini et al.,Mol.Cell.Probes 8:91-98(1994))。术语核酸可与基因、cDNA、mRNA、小非编码RNA、微小RNA(miRNA)、与Piwi-相互作用的RNA、以及由基因或基因座编码的短发夹RNA(shRNA)交互使用。

[0054] 术语“基因”意指参与产生多肽链或转录的RNA产物的DNA节段。它可以包括编码区之前和之后的区域(前导序列和尾随序列)以及个别编码节段(外显子)间的间隔序列(内含子)。

[0055] 如本文所用的术语“临床相关的核酸序列”(也被称为靶标序列或染色体)可以指对应于其潜在失衡被测试的较大基因组序列节段或对应于较大基因组序列自身的多核苷酸序列。一个实例是染色体21的序列。其它实例包括染色体18、13、X和Y。其它实例包括，胎儿从其父母之一或两者遗传的突变的基因序列或基因多态性或拷贝数变异，或者作为胎儿中重头合成的突变。在一些实施方案中，多个临床相关的核酸序列或临床相关核酸序列等同的多个标志物可以用于提供检测失衡的数据。例如，来自21号染色体的5个不连续序列的数据，能够以累加的方式(additivefashion)用于确定可能的21号染色体失衡，从而将所需的样品体积有效地降低至1/5。

[0056] 本文使用的术语“参照核酸序列”指其大小分布用于针对靶标序列比较的核酸序列。参照核酸序列的实例包括染色体、染色体的一部分、特定等位基因(例如母亲的等位基因)、特定单体型、基因组或者人工合成的核酸序列。这种参照核酸序列可以内在地存在于样品中或在样品处理或分析期间外源性地添加。在一些实施方案中，参照核酸序列展示了代表无疾病的健康状态的大小图谱。

[0057] 本文所用的术语“基于”意思是“至少部分基于”，并指在确定另一值中使用的一值(或结果)，诸如发生在方法输入和该方法输出关系中。本文所用术语“导出”也指方法输入和该方法输出的关系，诸如发生在当推导是公式的计算时。如本文使用的术语“参数”意指表征定量数据集和/或定量数据集间的数值关系的数值。例如，第一核酸序列的第一量和第二核酸序列的第二量间的比值(或比值的函数)是参数。

[0058] 如本文所用的，术语“基因座(locus)”或其复数形式“基因座(loci)”是具有变异的任意长度的核苷酸(或碱基对)在基因组的位置或地址。

[0059] 如本文所用的术语“序列失衡”指由至少一个截止值限定的临床相关核酸序列的量与参照量任一显著的偏差。序列失衡可以包括染色体剂量失衡、等位基因失衡、突变剂量失衡、拷贝数失衡、单体型剂量失衡以及其他类似的失衡。作为例子，等位基因或突变剂量失衡可以发生在当胎儿具有不同于母亲的基因型时，由此在样品的特定基因座产生失衡。

[0060] 如本文所用的术语“染色体非整倍性”指染色体的定量数量与二倍体基因组的染色体数量的变化。这种变化可以是增加或丢失。它可以涉及一条染色体的全部或染色体区域。

[0061] 本文所用的术语“单体型”指在相同染色体或染色体区域上同时传递的多个基因座上的等位基因的组合。单体型可以指少至一对基因座或指染色体区，或指整个染色体。术语“等位基因”指在同一物理基因组座的可选DNA序列，其可以导致或可以不导致不同的表型特征。在任一特定的具有每一染色体的两个拷贝(除男性人个体的性染色体外)的二倍体生物体中，每一基因的基因型包含存在于该基因座的等位基因对，其在纯合子中是相同的，

在杂合子中是不同的。生物体种群或种类一般包括在多个个体间的每一基因座上的多个等位基因。其中在种群中发现超过一个等位基因的基因组基因座被命名为多态位点。基因座的等位基因变异可测量为种群中存在的等位基因数(即多态程度)或杂合子的比例(即杂合率)。如本文所用的,术语“多态性”指人基因组中任一个体间的变异,而不管其频数。这种变异的实例包括但不限于单核苷酸多态性、简单串联重复序列多态性、插入-缺失多态性、突变(其可以是疾病导致的)以及拷贝数变异。

[0062] 详述

[0063] 已经发现,存在于孕妇血浆中的胎儿DNA分子一般比母体来源的分子更短(Chan et al Clin Chem 2004;50:88-92;Li et al Clin Chem 2004;50:1002-1011;美国专利申请20050164241)。本发明的某些实施方案可以从母体血浆DNA中确定胎儿是否通过改变了来源于基因组特定部分的DNA分子的大小而过度表现(over-representation)或低表现(under-representation)了基因组的该部分。因为胎儿DNA占母体血浆中少部分的DNA,所以母体血浆中大小变化的总程度可能不明显,因此可能难以检测。在一些实施方案中,测量很多分子的大小以达到疾病和健康间的统计学上显著性差异。

[0064] I.确定片段的大小

[0065] 可以测量很多DNA分子的大小的一个方法是通过大规模平行基因组测序。这可以通过例如Illumina基因组分析仪平台(利用合成测序)(Bentley DR et al Nature 2008;456:53-59)、ABI SOLiD(利用连接测序)(McKernan et al Genome Res 2009;19:1527-1541)、Roche 454平台(Marguelis et al Nature 2005;437:376-380)以及Helicos单分子测序平台(Harris et al Science 2008;320:106-109)来实施。还预期也可以利用其它的大规模平行测序平台,例如Pacific Biosciences(单分子、实时(SMRT™)技术)(Eid et al Science 2009;323:133-138)、纳米孔测序(Clarke J et al.Nat Nanotechnol 2009;4:465-470)、半导体测序(例如通过Ion Torrent(www.iontorrent.com)等。

[0066] 从这种测序获得DNA大小信息的一种方式是实施双末端(PE)测序,其中DNA分子的两末端被测序。然后,对应于分子两末端的序列可以映射(map)回参照基因组(例如,参照人基因组或参照马基因组,或者感兴趣的任何动物的基因组)。在一个实施方案中,单独地对于每一末端而言,两末端的每一末端测序的长度长到足够被映射回参照人基因组(例如,约10-24个碱基或25-36个碱基)。在另一实施方案中,仅一部分序列可以没有错配地映射回人基因组的非重复序列区域。在一方面,如果两条序列一起用于映射中,映射可以是明确的。在这种情形下,即便每一末端过短而不能可信地映射回,利用两条序列也可以提供明确的映射。可以通过两条序列末端的基因组坐标的相减计算出分子大小。

[0067] 在另一实施方案中,可以通过完整DNA分子的完全测序或接近完全的测序,而不是仅两末端测序来获得分子的大小。这可以通过具有相对长读长的测序平台有效实现,诸如Roche 454平台、Pacific Biosciences单分子、实时(SMRT™)技术以及Ion Torrent技术(www.iontorrent.com)。

[0068] 上述测序方法的通量可以随着使用索引或条形编码而增加(Cronn et al.Nucleic Acids Res 2008;36:e122)。因此,样品或患者特异性索引或条形码可以添加至特定核酸测序文库中的核酸片段。然后,将大量的这种文库(每种都具有样品或患者特异性索引或条形码)混合在一起,并一起测序。测序反应后,可以基于条形码或索引收集来自

每一样品或患者的测序数据。该策略可以增加通量并由此增加本发明的成本效益。

[0069] 在另一实施方案中，可以在进行大小分析前选择或分离生物样品中的核酸分子。在一个变型中，将核酸分子用会优先结合来自基因组(例如，染色体21、18、13或X中的一个)所选基因座的核酸分子的装置(例如，含探针的微阵列或溶液)处理，然后在结合的核酸分子子集上进行大小分析。在这种实施方案中，可以使用Nimblegen序列捕获系统(www.nimblegen.com/products/seqcap/index.html)或Agilent SureSelect靶标富集系统(Target Enrichment System(www.opengenomics.com/SureSelect_Target_Enrichment_System))、或者类似的平台。在另一实施方案中，未结合的核酸子集可被差别移除或降解或消化。

[0070] 至少一些实施方案可以用任一单分子分析平台来进行，其中染色体来源和分子的长度可以被分析，例如电泳、光学方法(例如光学作图法和其变型，en.wikipedia.org/wiki/Optical_mapping#cite_note-Nanocoding-3，以及Jo et al.Proc Natl Acad Sci USA 2007;104:2673-2678)、基于荧光的方法、基于探针的方法、数字PCR(基于微流体或基于乳剂，例如BEAMing(Dressman et al.Proc Natl Acad Sci USA 2003;100:8817-8822)、RainDance(www.raindancetech.com/technology/pcr-genomics-research.asp))、滚环扩增、质谱分析法、熔解分析法(或熔解曲线分析)、分子筛技术等。作为质谱分析法的例子，更长的分子会具有更大的质量(大小值的例子)。

[0071] 在一个实例中，利用双末端测序方案，通过Illumina基因组分析仪系统对血浆DNA分子进行随机测序。在该实验中，利用第1版的Illumina双末端(PE)簇生成试剂盒。每一末端测序36bp。利用Illumina提供的GAPipeline-1.0软件包中的eland_成对程序，将每一序列的两个末端与重复序列掩蔽的人基因组(NCBI Build 36，第48版)对齐。每一末端的36bp中仅有32bp被用于对齐目的。

[0072] 在一些实施方案中，满足以下标准的PE读码可以用于随后的分析：(1)每一建议对的单独成员都在测序流通池的同一基因簇位置进行测序，并且能够以人参照基因组预期的正确方向被对齐到同一染色体；(2)该对的两个成员测序的读码能够被对齐到重复序列掩蔽的人参照基因组，而无任何核苷酸错配；(3)该对的每一成员的测序读码具有大于4的单值性分数(uniqueness score)；以及(4)表现出小于600bp的插入物大小的对。然后按照两末端每个的位置计算每一对齐的序列的大小。

[0073] II.利用大小分布以确定非整倍性状态

[0074] 图1是流程图，说明了按照本发明的实施方案，在获自孕妇个体的生物样品中进行序列失衡的产前诊断的方法100。尽管方法100主要结合分析胎儿染色体非整倍性被描述，但方法100的其他实施方案和本文的其他方法可以应用于其他序列失衡(例如，基因型或突变的鉴定)。方法100和本文提及的其他方法可以完全或部分由包括一个或多个处理器的计算机系统来实施。

[0075] 方法100和本文描述的任一方法可以完全或部分用可以被配置为执行步骤的包括处理器的计算机系统来实施。因此，实施方案涉及计算机系统，其被配置为执行本文描述的任一方法的步骤，可能具有执行单独步骤或单独的步骤组的不同组件。尽管表现为编号的步骤，但本文方法的步骤可以同时执行或以不同顺序执行。另外地，这些步骤的一部分可以与来自其他方法的其他步骤的一部分一起使用。并且，步骤的全部或一部分可以是任选的。

另外地，任一方法的任一步骤都可以用模块、电路、或执行这些步骤的其他装置来执行。

[0076] 在步骤110中，测量生物样品中的至少一些核酸分子(例如DNA或RNA)的大小。核酸分子还可以被称为片段，因为它们是完整基因组的片段。可以通过任何适合的方法来测量大小，例如，上文描述的方法。

[0077] 步骤120鉴定每一核酸分子来源的基因组的位置。该位置可以是基因组的任一部分，所述基因组在提供的实例中是人基因组，但也可以是其它基因组。例如，该位置可以是片段来源的染色体编号、可以通过基因组坐标来限定(例如，特异性坐标或坐标范围)的染色体部分，以及甚至可以是片段来源的(起源的)的两个染色体(假定整倍体)中的一个。

[0078] 在一个实施方案中，可以通过测序以及将序列信息与参照人基因组序列比较来实施这一鉴定。在另一实施方案中，该鉴定可以通过与具有已知染色体起源的探针组杂交来实施。探针可以用一种或多种荧光标记物来标记，以微阵列形式或在溶液中。在另一实施方案中，核酸分子可以通过在溶液中或在固体表面上的探针组来捕获，然后对捕获的(或者剩余的未被捕获的)核酸分子进行测序。在一些实施方案中，如果要鉴定染色体非整倍性外的序列失衡，鉴定片段起源于哪一染色体的步骤可以是任选的。

[0079] 在步骤130中，确定对应于第一位置(例如第一染色体)的核酸分子的大小分布。多种实施方案可以利用多种大小分布。在一些实施方案中，大小分布涉及一条染色体的片段相对于其它染色体片段的大小的秩次(例如，平均数(average)、中值或均值(mean))。在其它实施方案中，大小分布可以涉及染色体片段实际大小的统计值。在一项实施中，统计值可以包括染色体片段的任何平均数、均值或中值大小。在另一实施中，统计值可以包括低于截止值的片段的总长度，其可以除以所有片段或者至少低于较大截止值的片段的总长度。

[0080] 在步骤140中，基于确定的大小分布，确定第一位置是否存在序列失衡(例如，胎儿染色体非整倍性)的分类。在一个实施方案中，将染色体的秩次与参照秩次(例如整倍体样品的秩次)比较。如果变化是明显的(例如超过阈值)，然后样品可以被分类为非整倍体。在另一实施方案中，比较两个染色体间或染色体组间的实际大小的统计值。例如，在单独的统计值间获取差值，并将差值与截止值比较。

[0081] II.大小分布(秩次)

[0082] 实施方案可以利用样品核酸片段的大小值以确定是否存在染色体失衡。片段在被进行测序后也称为序列。在一个实施方案中，确定多个染色体片段的大小分布，以及基于分布的统计值(例如平均数、均值或中值)将染色体求秩。方便起见，术语“大小”在本文可以与大小的统计值同义使用。应该清楚何时术语“大小”指特定片段的大小和一组片段大小的统计测量。

[0083] A.秩次

[0084] 图2是图200，其说明了来源自母体血浆中不同染色体的DNA片段的测量的大小分布。因为测量的大小不仅能够反映体内DNA片段大小，而且能够反映分析步骤的贡献，所以预期大小分布可以随平台的不同而不同(例如，对于Illumina基因组分析仪和对于ABI SOLiD平台)，并且甚至对于特定平台当使用不同版本的试剂也会变化。然而，只要利用相同平台或试剂类型分析参照样品和测试样品，可以以平台/试剂非依赖方式利用实施方案。例如，如果可以确定和校正任何误差，或者如果可以表明平台和/或试剂类型具有严密匹配的分析性能，一些实施方案还可以利用不同的平台和/或试剂类型。

[0085] 在图2中，从与22个常染色体和染色体X对齐的片段序列的大小方面对这些染色体进行了比较。在Y轴上，秩次是序列大小的降序排列，即具有最长序列的染色体秩次为1，具有最短序列的染色体秩次为23。在一个实施方案中，利用SigmaStat(SPSS)软件进行非参量比较(秩次的Kruskal-Wallis单因素方差分析，随后是Bonferroni-校正的成对比较)。比较可以是代表大小的任何统计值的比较，包括单独大小的每个的秩次以及每一染色体序列的单独秩次的统计分析。在一项实施中，允许在秩次编号中有节点(ties)和跳跃。

[0086] 在一些实施方案中，将每一序列映射至特定染色体。然后，对于每一染色体，确定映射至该染色体的序列的一个或多个统计值。对于每一染色体的大小，可以计算不同类型的统计值(例如，均值和中值)。然后将对应的统计值求秩。例如，可以将每一染色体的平均大小彼此比较。如果使用多于一个的统计值，那么可以将统计值合并(例如，按照某一公式，诸如加权平均数)，然后可以将这种合并的统计值求秩。在一个实施方案中，可以合并特定染色体的统计值的秩次(例如，如上述提及的统计值)，然后将合并的秩次彼此比较。

[0087] 在其它实施方案中，所有的序列都按照大小单独求秩。也就是说，如果有1,000,000个序列，秩次从1到一百万，在秩次编号中有可能的节点和跳跃。然后可以将映射至相同染色体的所有序列的秩次加在一起。秩次总和可以除以对齐到特定染色体的序列数以得到该染色体的平均序列秩次。具有最高平均序列秩次的染色体可以被标记为最长的(秩次为1，在Y轴上最高)，以及具有最低平均序列秩次的染色体是最短的(秩次为23，在Y轴上最低)。在另一实施方案中，可以确定中值秩次。具有整倍体男性胎儿、整倍体女性胎儿和21三体男性胎儿的病例的中值秩次显示在图2中。

[0088] 在X-轴上，染色体的排列为来源自每一染色体的DNA片段的大小分布的降序排列(X染色体除外(见后文))。在一个实施方案中，对于这种秩次，仅适用整倍体病例。具有最长的测量大小(即长度)的染色体被排列在左侧。X染色体放在图的右侧是因为它的秩次由胎儿性别控制。

[0089] 如上提及的，测量的大小可以随平台的不同而不同(例如，随着从Illumina系统到另一系统而变化)。因此，在一方面，大小可以指“测量的”大小，而不是实际大小。大小甚至可以在从一个版本的Illumina试剂盒变换到另一个时而变化，例如当版本1的双末端基因簇生成试剂盒变化到版本2时。在一个实施方案中，使用者可以进行他们特定系统的秩次评定。

[0090] 从图2中可以看出，当与怀男性胎儿的孕妇相比时，X染色体的平均秩次在怀女性胎儿的孕妇血浆中更低(即变得更短)。这一观测结果的解释是由胎儿释放的DNA片段短于来自母亲的DNA片段。因此，通过释放双倍剂量的X染色体，女性胎儿会降低来自母体血浆的X染色体的片段的总测量的大小。相反地，男性胎儿仅能释放单剂量的X染色体。

[0091] 还可以从图2中看出，与怀整倍体胎儿的孕妇相比，怀21三体胎儿的孕妇血浆中染色体21的秩次下降(即变得更短)。该观测结果的解释可以再次追溯到来自母体血浆染色体21的片段的测量的大小。相反地，整倍体胎儿的每一胎儿细胞仅能释放双剂量的染色体21。

[0092] 图3是图300，证明了按照本发明的实施方案，来自母体血浆不同染色体的序列的大小分析可以用于胎儿染色体21非整倍性无创性产前检测。在该实例中，染色体大小由与其他染色体相比时的大小秩次来表示。因此，较大的大小秩数指示染色体在母体血浆中具有较短的DNA片段大小。

[0093] 图300证明了当与胎儿是整倍体的孕妇(秩次为9-18)相比时,怀21三体胎儿的孕妇(秩次为18-21)的母体血浆中染色体21的大小秩数更大(由此指示更短的DNA片段)。该观测结果的解释为胎儿DNA短于母体DNA,以及来自21三体胎儿的染色体21的额外剂量会导致母体血浆中染色体21序列统计值的总体缩短。

[0094] B.利用秩次的方法

[0095] 图4是流程图,说明了按照本发明的实施方案,利用大小统计值的秩次在获自孕妇个体的生物样品中进行胎儿染色体非整倍性产前诊断的方法。

[0096] 在步骤410中,测量获自生物样品的多个核酸分子(片段)中每一个的大小。要注意的是,该多个核酸分子可以是获得的所有核酸分子的子集。当测序被作为大小测量的一部分进行时,多个核酸分子甚至可以是被测序的所有核酸分子的子集。

[0097] 在步骤420中,鉴定多个核酸分子的每个来源于哪一染色体。在多个实施方案中,步骤410和420的顺序可以颠倒或同时发生。例如,在双末端测序条件下,序列的基因组比对可以提供其染色体位置及其长度(通过起始和末端基因组坐标的相减)。在一个实施方案中,染色体可以如在步骤120中那样被鉴定。

[0098] 在步骤430中,对于多条染色体的每一条,从对应于染色体的核酸分子的大小计算统计值。可以本文描述的任一方式计算统计值。例如,统计值可以包括初始求秩阶段的结果,如上文所述的。在一个实施方案中,可以仅利用对应于任一特定染色体的核酸分子的一部分。

[0099] 在步骤440中,基于统计值确定染色体的秩次。在一个实施方案中,可以利用基础排序算法来确定秩次。在另一实施方案中,可以实施更为复杂的比较,诸如秩次的Kruskal-Wallis单因素方差分析以及随后的Bonferroni-校正的成对比较,或者其它合适的方法。在不同实施方案中,秩次可以是整数、分数、实数(例如在一范围内)、或基于规则的字母数字秩次(例如,A-X)。

[0100] 在步骤450中,将确定的第一染色体秩次与获自参照生物样品的第一染色体的另一秩次进行比较。在一个实施方案中,比较是确定的秩次针对截止阈值(例如单值或范围)的比较,该截止阈值从一个或多个参照生物样品的秩次来确定。如果秩次是18(或19)或更高,这种截止值可以是如从图3中所确定的。在另一实施方案中,可以确定两种样品间第一染色体的秩次的差值,并且可以将该差值与截止值比较。在一项实施中,分析参照生物样品以确定它不含有感兴趣的疾病,以及甚至可以确定样品不含有可能导致秩次问题的任何相关疾病。

[0101] 在步骤460中,基于比较,确定第一染色体是否存在胎儿染色体非整倍性的分类。在一个实施方案中,分类可以是疾病或没有疾病的二元分类。在另一实施方案中,分类可以是三元的,因为可以利用不确定的分类。在另一实施方案中,分类可以包括特定分类的概率,由此有效地具有不仅三个分类。

[0102] III.大小分布(大小的统计值)

[0103] 在其它实施方案中,大小分布可以包括片段大小的统计值(例如,特定基因组位置的实际或绝对值的统计值),而不是秩次。在一个实施方案中,可以将第一染色体的实际大小与同一测试样品的一条或多条参照染色体的实际大小进行比较。例如,第一染色体和一条或多条参照染色体的这些实际值间的间隔值(例如,差值或比值)可以与截止值进行比

较。在一项实施中，截止值可以从参照样品中确定。在另一实施方案中，测试样品和参照生物样品间的染色体片段实际大小间的间隔值可以连同截止值一起被使用。在另一实施方案中，可以将染色体片段的实际大小针对截止值进行比较以获得可信的分类。

[0104] A.绝对大小

[0105] 一些实例表明，技术人员可以通过将来源于染色体21的片段的绝对大小与来源于一条或多条参照染色体的片段的绝对大小进行比较，从而实现21三体的无创性产前检测。在一个实施方案中，染色体7和14被选为参照染色体，因为它们在母体血浆中的大小值(例如，绝对大小或大小秩次)相对接近于染色体21。在实践中，参照染色体可以是任一染色体，其具有的片段大小在其与整倍体样品染色体21(或其他感兴趣的染色体)的关系上(例如在特异性分析平台和/或试剂类型方面)是一致的。

[0106] 图5是表500，说明了按照本发明的实施方案，在与各自染色体对齐的序列大小方面对染色体21与染色体7和染色体14的比较。表500的数据获自16个测试样品。对于每一样品，显示了染色体7、14和21的每一条的片段的平均大小。还提供了平均值间的差值。p值表明了每一差值会发生在健康样品中的可能性。

[0107] 如可以从图5的表500看到的，对于所有21三体的孕妇而言，与染色体21对齐的序列明显短于(例如通过平均大小)与染色体7和染色体14对齐的序列(Mann-Whitney秩和检验，p-值<0.001)。这种程度的统计学显著性缩短并没有在整倍体孕妇中观察到。因此，表500表明对于所有21三体孕妇而言，染色体21和染色体7间的平均片段大小的差值大于1bp，而没有整倍体病例显示大于1bp的差值。因此，1bp可以提供确定分类的精确截止值。类似地，对于所有21三体孕妇而言，染色体14的平均片段大小一致地大于染色体21的平均片段大小。实际上，如果0.5bp的截止值用于当与来自染色体21的片段比较时观测到染色体14片段的“延长”，可以将所有21三体病例与非21三体案列区分开。因此，在一个实施方案中，截止值可以从一个或多个参照样品确定。

[0108] B.利用绝对大小的方法

[0109] 图6是流程图，说明了按照本发明的实施方案，利用基因组位置的片段大小的统计值的比较，在获自孕妇个体的生物样品中进行序列失衡的产前诊断的方法。在一方面，方法600可以涉及基于第一染色体的片段大小和一条或多条参照染色体的片段大小的间隔值(例如差值或比值)确定序列失衡的分类。

[0110] 在步骤610中，测量获自生物样品的多个核酸分子的大小。注意的是，多个核酸分子可以被获得并且包括如步骤410提及的相似片段。

[0111] 在步骤620中，鉴定每一核酸分子来源的基因组位置。位置可以是基因组的任一部分，如步骤120和其它地方所描述的。例如，鉴定多个核酸分子的每一个来源于哪一染色体。可以通过映射至参照基因组进行这一确定，如本文所述的。

[0112] 在步骤630中，从对应于第一基因组位置(例如，第一染色体)的核酸分子的大小计算第一统计值。在一个实施方案中，第一统计值可以是对应于第一染色体的片段的平均、均值或中值大小。在另一实施方案中，第一统计值可以包括低于第一大小的片段的长度总和，所述第一大小可以是一种类型的截止值。例如，小于200bp的每一片段可以将其长度求和。总和可以除以另一数值，诸如对应于第一染色体的所有片段长度之和或者大于第二大小截止值(其可以与第一大小相同)的片段长度之和。例如，第一统计值可以是低于第一大小截

止值的片段总长度相对于片段总长度的比值，或者小片段总长度相对于大片段总长度的比值。

[0113] 在步骤640中，从对应于第二基因组位置(例如，第二染色体)的核酸分子的大小计算第二统计值。第二染色体可以视为参照染色体。在一个实施方案中，可以计算多个参照染色体的统计值。在一项实施中，可以合并统计值，以使统计值可以是来自一条或多条第二染色体的统计值。在另一实施方案中，多个参照染色体的统计值可以单独比较，如上文提及的。

[0114] 在步骤650中，比较第一统计值和第二统计值以获得间隔值。在一个实施方案中，间隔值可以是确定的第一统计值和第二统计值间的差值。在另一实施方案中，间隔值可以是两个统计值的比值。在另一实施方案中，可以确定多个间隔值，例如，对于每一个第二统计值的间隔值，其可以相对于每一参照染色体来计算。

[0115] 在步骤660中，可以将间隔值与一个或多个截止值比较。在一个实施方案中，可以进行多个间隔值的每一个的比较。例如，如上文提及的，可以确定第一统计值与每一个第二间隔值间的不同间隔值。在不同实施中，每一间隔值可以与相同或不同的截止值比较。在另一实施方案中，将间隔值与两个截止值比较以确定间隔值是否在特定范围内。该范围可以包括一个截止值以确定是否出现非正常数据点(例如非整倍性)，以及第二截止值可以用于确定数据点是否可能由测量或分析中的误差导致(例如，如果间隔值大于曾经预期的，甚至对于有疾病的样品)。

[0116] 在步骤670中，基于比较确定第一基因组位置是否存在序列失衡(例如，胎儿染色体非整倍性)的分类。在一个实施方案中，多个截止值(例如N个截止值)可以用于单个间隔值。在这种实施方案中，可以确定N+1个分类。例如，两个截止值可以用于确定整倍体(正常或健康的)、不确定的以及非整倍体(患病的或不健康的)的分类。在其中进行多个比较(例如，对于每一间隔值的比较)的另一实施方案中，分类可以基于每一比较。例如，基于规则的方法可以考虑由每一比较产生的分类。在一项实施中，只有当所有分类一致时，才提供明确的分类。在另一实施中，使用多数分类。在另一实施中，可以基于每一间隔值与各自截止值的接近程度利用更为复杂的公式，以及可以分析这些接近值(closeness value)来确定分类。例如，接近值可以求和(与其他因子一起，诸如标准化)，并且可以将结果与另一截止值比较。

[0117] 在其它实施方案中，方法600的变型还可以应用于第一染色体的统计值与截止值的直接比较，所述截止值可以从参照样本得到。一些实施方案还可以用于分析来自非怀孕个体的生物样品。这种分析可以仅考虑样品的所有片段的大小的统计值，并将统计值或间隔值与截止值比较以确定是否存在序列失衡。如果被分类为存在失衡，可以进行失衡位置的进一步分析，例如通过分析特定基因组位置(例如，染色体)的统计大小值和/或间隔值。

[0118] C.利用短片段的总长度

[0119] 如上文提及的，在一些实施方案中，还可以通过由短DNA片段贡献的总DNA长度的分数来反映血浆DNA的大小分布。例如，大小分布可以包括低于截止值的片段的总长度，其可以除以所有片段或者至少低于较大截止值的片段的总长度。相反地，还可以通过由长DNA片段贡献的总DNA长度的分数来反映血浆DNA的大小分布。例如，大小分布可以包括高于截至值的片段的总长度，其可以除以所有片段或至少低于较小截止值的片段的总长度。作为

另一个例子，还可以利用小与大的比值。一个实施方案利用150bp作为来限定短血浆DNA片段的截止值。然而，任何截止值，例如130bp、140bp、160bp和170bp也可以作为限定短DNA片段的截止值。注意的是，如本文所用的，在提及单链片段长度时，碱基对还可以与多个核苷酸(nt)同义。

[0120] 在一个实施方案中，由短DNA片段贡献的DNA长度分数的计算可以如下：F＝由短DNA片段贡献的DNA长度的分数；S＝所有短DNA片段长度之和(等于或低于截止值)；以及T＝样品中所有DNA片段的总长度，与它们的长度无关。由此该分数可以通过F＝S/T提供，其是大小统计值的一个实例。F的计算可以用于来自样品的所有片段，或者用于基因组的特定位置，例如用于特定染色体。

[0121] 在一项实施中，可以确定样品中所有DNA片段的总长度。然后可以选择截止大小(w)，低于该截止大小的DNA片段被定义为“短片段”。截止大小可以变化，以及可以选择截止大小以适合不同诊断目的。可以通过计算等于或短于截止大小的所有DNA片段的长度之和来计算短DNA片段的总长度。由短DNA片段贡献的总长度分数可以如下计算：

[0122] $F＝\sum^{w}$长度/$\sum^{600}$长度，其中

[0123] $\sum^{w}$长度表示长度等于或低于截止值w(bp)的DNA片段长度之和；以及

[0124] $\sum^{600}$长度表示等于或低于600bp的DNA片段的长度之和。

[0125] 统计值F还可以用于使用秩次的实施方案中。例如，可以从一组单独基因组位置(例如染色体)中每一个的片段计算F。

[0126] 出于示例目的，在下文实例中通过短于600bp的片段之和计算总长度。然而，其它大小限制，例如400bp、500bp和700bp可以用于计算“总长度”。在该实例中，基于600bp或更小的DNA片段计算总长度，因为Illumina基因组分析仪(Solexa)系统在扩增和测序长于600bp的DNA片段时不是有效的。此外，将分析限制在短于600bp的DNA片段还可以避免由基因组的结构变异引起的偏差。在存在结构变异，例如重排(KiddJM et al,Nature 2008;453:56-64)的情况下，当通过将DNA片段末端映射至参照基因组在生物信息学上估计大小时，可能高估DNA片段的大小。此外，所有成功测序并映射至参照基因组的DNA片段中大于99.9％的片段小于600bp，并因此包括等于或短于600bp的所有片段，会提供样品中DNA片段的大小分布的代表性估计。

[0127] 如上文讨论的，对于胎儿具有额外拷贝的染色体，可以观测到DNA片段向更短大小分布的移位。在一个方面，在具有非整倍性风险的染色体(靶染色体)和无非整倍性风险的染色体(参照染色体)间由短DNA片段贡献的总长度分数的差值测量可以是定量测量以确定来源于这些染色体的DNA片段的大小分布是否不同。

[0128] 在一个实施方案中，我们将$F_{(Tar)}$和$F_{(Ref)}$分别定义为具有非整倍性风险的染色体和参照染色体的短DNA片段贡献的总长度分数。靶染色体和参照染色体间由短DNA片段贡献的总长度分数的差值($\Delta F_{(Tar-Ref)}$)可以计算为$\Delta F_{(Tar-Ref)}＝F_{(Tar)}-F_{(Ref)}$。例如，$\Delta F_{(21-1)}$是染色体21和染色体1的短DNA片段贡献的总长度分数的差值。在胎儿染色体非整倍性产前诊断中应用$\Delta F_{(Tar-Ref)}$将在下述章节中讨论。在另一实施方案中，$F_{(Tar)}/F_{(Ref)}$的比值还可以以与使用$\Delta F_{(Tar-Ref)}$类似方式用作间隔值。

[0129] 可以将靶染色体和参照染色体的F值间的差值作为统计值以确定胎儿的靶染色体是否是三体的。当胎儿的靶染色体是三体的时，胎儿的三体染色体会为母体血浆贡献额外

剂量的短胎儿DNA，因此导致靶染色体序列大小分布明显缩短。这种靶染色体序列的大小分布的缩短会导致由靶序列($F_{靶标}$)的短DNA片段贡献的序列长度分数的增加。因此，在$F_{靶标}$和$F_{ref.}$间的差值ΔF会增加。

[0130] 图7显示了对于怀男性胎儿的母体血浆样品，由短片段(F)贡献的总长度分数相对截止大小(w)的图700。将与常染色体和Y染色体对齐的DNA片段的F值相对于用于定义“短DNA片段”的截止大小标绘于纵轴。在怀男孩的孕妇中，与Y染色体对齐的DNA分子代表从男性胎儿释放的DNA。因为母体血浆中的大多数循环DNA分子来源于母亲，所以与常染色体对齐的DNA片段应该主要代表母体DNA片段。F值随着截止大小增加，并且当样品中的所有DNA片段短于或等于截止大小时，F值接近1.0的值。两类DNA分子间的大小分布差异可以通过它们的F值的差值反映。较高的F值表明由短片段贡献的总长度分数更高，并由此表明DNA片段的大小分布更短。

[0131] 如图700所示的，来自Y染色体的DNA分子的大小分布短于来自常染色体的DNA的大小分布。特别地，Y染色体的F值上升早于常染色体的F值，从而导致在80bp至350bp间F_{Y}高于$F_{常染色体}$。将Y染色体和常染色体间的F值的差值($\Delta F_{(Y-常染色体)}$)进一步相对截止大小作图，并且由虚线表示，其在80bp至350bp间是正值。$\Delta F_{(Y-常染色体)}$的最大值是出现在约150bp的0.23。如下文的实例所说明的，高危染色体和参照染色体间由短片段贡献的总片段长度分数差值(ΔF)是量化它们大小分布差异的有用的间隔值。进一步举例来说，技术人员可以利用任何大小截止值，例如130bp-170bp的大小截止值来确定ΔF值。

[0132] 图8显示了对于整倍体孕妇，由染色体21(实线)和参照染色体(除染色体13、18和21外的所有常染色体)(点划线)的短片段贡献的总长度分数(F)相对于截止大小的图800。两个F值的差值($\Delta F_{(21-Ref)}=F_{(chr21)}-F_{(Ref)}$)由虚线表示。因为染色体21和参照染色体的DNA片段的大小分布在整倍体孕妇中是相似的，所以对任一截止大小而言，$\Delta F_{(21-Ref)}$的值接近零。

[0133] 图9显示了对于21三体孕妇，染色体21(实线)和参照染色体(除染色体13、18和21外的所有常染色体)(点划线)的F值相对于截止大小的图900。$\Delta F_{(21-Ref)}$由虚线表示。由于来自胎儿的染色体21的额外剂量，母体血浆中染色体21的DNA片段的大小分布短于参照染色体的分布。DNA片段的大小分布的差异反映为$\Delta F_{(21-Ref)}$的正值，其在约150bp处达到0.016的最大值。

[0134] 图10显示了对于整倍体和21三体孕妇，染色体21和参照染色体(除染色体13、18和21外的所有常染色体)间的$\Delta F_{(21-Ref)}$相对于大小截止值的图1000。在21三体病例中观测到增加的$\Delta F_{(21-Ref)}$，但对于整倍体病例而言，$\Delta F_{(21-Ref)}$在每一大小截止值处均接近零。因为在约150bp处观测到最大$\Delta F_{(21-Ref)}$，150bp处的差值可以作为用于确定是否存在染色体21序列大小分布的任何明显缩短的间隔值。然而，可以利用在整倍体和三体病例间出现显著差异的任何大小，例如，但不限于140bp、145bp、155bp和160bp。在该实例中，对于21三体和整倍体孕妇，在150bp处观测到的染色体21和参照染色体总长度分数的差值分别为0.016和-0.002。

[0135] 由短DNA片段贡献的总DNA长度分数形式中的这种差异可以用于区分整倍体和非整倍体胎儿。可以以多种方式检测差异。在一个实施方案中，可以将特定大小截止值处的ΔF值(间隔值的一个例子)与截止值比较以确定样品的类别(分类)。在另一实施方案中，可以

存在ΔF峰值，并且可以将该值与一个或多个类别截止值比较。在多个实施方案中，峰值可以是最大或最小值、接近max/min值的平均值、或者与max/min值有关或来源于max/min值的其它值)。还可以利用间隔值(例如ΔF)的其它统计值，诸如峰宽、或对应于峰的特定大小截止值的位置。

[0136] 在一个实施方案中，获自特定基因组位置或整个基因组的多个片段的F值(或本文描述的其它统计值，诸如小片段长度除以大片段长度)可以用于确定是否存在异常状态。例如，如果统计值超过截止值，可以鉴定存在异常状态，因为小片段的量在正常范围之外。除孕妇外，这还可以用于其它患者以鉴定除了胎儿疾病之外的疾病。

[0137] 在一些实施方案中，可以在片段大小分析之前进行物理大小分级。在一个实施方案中，核酸分子可以被分为两种大小的部分(例如，大于200bp的部分，以及小于或等于200bp的部分)，然后可以比较这些大小部分中每个上的所选择的染色体(例如，染色体21)的大小分布。在存在胎儿三体(例如，21三体)的条件下，分子大小较小的大小部分与分子大小较大的大小部分相比相对丰度增加。

[0138] 在其它实施方案中，可以利用低于长度截止值而不是大小分布的多个片段。例如，可以相对于一条或多条参照染色体，比较靶染色体(例如，染色体21)的低于长度截止值的片段数(例如，差值或比值)。在一个实施方案中，低于长度截止值的片段数除以片段总数以获得百分比，并且可以在靶染色体和一条或多条参照染色体间比较该百分比以提供参数。得到的参数(例如，差值或比值)可以与截止值(例如1％)比较。在一方面，可以在上述百分比最高的长度处选择长度截止值。

[0139] V.利用秩次的实例

[0140] 除了21三体外，母体血浆中的片段大小分析也可以用作其它胎儿染色体非整倍性的无创性产前检测，诸如13三体、18三体和性染色体非整倍性(诸如特纳(Turner)综合征、克氏(Klinefelter)综合征以及XYY等)。当染色体异常仅涉及特定染色体的一部分(例如，由染色体易位导致的21三体)时，也可以利用实施方案。在这种情形下，会观测到来自受影响的染色体区的DNA片段的片段大小异常。

[0141] 图11显示了母体血浆DNA文库的表1100，该文库利用多重样品制备试剂盒(Illumina)按照生产商的说明书构建。将具有可区别的条形码的每两个样品导入一个通道，随后在Illumina基因组分析仪II上进行标准的多重双末端测序。可以基于条形码区分样品。分析120个孕妇的血浆样品。胎儿的性别和染色体非整倍性状态显示在表1100中。

[0142] 图12说明了按照本发明的实施方案，不同疾病或无疾病状态的四种样品的不同染色体的秩次。如前所述，将22条常染色体以及染色体X按照它们的片段大小求秩。正如可以看到的，基于版本2的Illumina基因簇生成试剂试剂盒的染色体的相对秩次不同于使用版本1的试剂盒时的相对秩次(图2)。如下文所示，实施方案(例如方法400)允许非整倍体病例与整倍体病例区分开来。

[0143] A.13三体

[0144] 在该实例中，我们证实了实施方案用于13三体的产前诊断。图13显示了120个孕妇的染色体13秩次的图，例如其可以源于方法400。在该图中，T13、T18、T21和Eu分别指13三体、18三体、21三体和整倍体孕妇。在23个13三体孕妇的18个(78.3％)中，染色体13的秩次为22或以下，而在97个非13三体孕妇中仅有2个(2.1％)的染色体13的秩次为22或以下。因

此，利用22的截止秩次，用于胎儿13三体产前诊断的染色体片段大小的秩次分析的灵敏度和特异度分别为78.3％和97.9％。

[0145] 在图13中，可以看出，整倍体以及18和21三体孕妇的染色体13的秩次高于(即代表秩次的数较小)13三体孕妇。换句话说，当与其它染色体上的序列相比时，13三体孕妇的染色体13序列看起来短于非13三体孕妇的染色体13序列。13三体孕妇中染色体序列的明显缩短是由于胎儿DNA对染色体13的贡献增加，这归因于胎儿的额外的染色体13。

[0146] B.18三体

[0147] 图14显示了120个病例的染色体18的秩次。在30个18三体病例的26个(86.7％)中，染色体18的秩次低于(即表示秩次的数大)13，而90个非18三体病例中没有一个具有低于13的秩次。因此，利用秩次13作为截止值，用于胎儿18三体产前诊断的染色体片段大小的秩次分析的灵敏度和特异性分别为86.7％和100％。

[0148] 在该分析中，我们在18三体、13三体、21三体和整倍体孕妇中比较了染色体18的大小秩次。就染色体18而言，后三组可以被视为“正常”对照，因为它们没有染色体18的剂量异常。如可以从图14中看到的，整倍体和13三体病例的染色体18的秩次集中在1-3附近。另一方面，18三体病例的染色体18的秩次为13至22，表明染色体18片段的大小短于整倍体、21三体和13三体病例中的染色体18片段的大小。同样，额外的染色体18解释了这些观测结果。

[0149] C.21三体

[0150] 图15显示了120个病例的染色体21的秩次。在9个21三体病例的8个(88.9％)中，染色体21秩次为22或以下，而在111个非21三体病例中没有一个的染色体21具有22或以下的秩次。因此，利用秩次22作为截止值，用于胎儿21三体产前诊断的染色体片段大小秩次分析的灵敏度和特异性分别为88.9％和100％。

[0151] V.利用大小差值的实例

[0152] A.13三体

[0153] 在接下来的这些实例中，我们证实了技术人员可以通过比较来源于染色体13的片段的绝对大小和来源于一条或多条参照染色体的片段的绝对大小而实现13三体的无创性产前检测，如方法600所描述的。该实例利用与前述实例中的数据集相同的数据集。作为示例，对于13三体检测，我们选择了染色体5和染色体6作为参照染色体。

[0154] 如从图16的表1600可以看出的，对于所有13三体孕妇的同一样品内，与染色体13对齐的序列显著短于与染色体5和染色体6对齐的序列(Mann-Whitney秩和检验，p-值≤0.001)。作为对照，包括了整倍体和18三体孕妇。在整倍体和18三体孕妇中，染色体13的剂量是正常的。如从表1600可以看出的，来源于染色体13的序列的这种统计显著的大小异常没有在整倍体和18三体孕妇中发现。

[0155] 而且，对于所有13三体孕妇，在同一样品内，染色体13和染色体5间的平均片段大小的差值大于0.4bp，而非13三体病例中没有一个显示大于0.4bp的差值。类似地，对于所有13三体孕妇，染色体13和染色体6间的平均片段大小的差值大于0.5bp，而没有一个非13三体病例显示大于0.5bp的差值。

[0156] B.18三体

[0157] 作为示例，对于18三体检测，选择染色体14作为参照染色体。对于非18三体孕妇，在图17的表1700中可以看到，来源于染色体18的序列在统计学上明显长于来源于染色体14

的序列(Mann-Whitney秩和检验,p-值≤0.005)。然而,对于18三体病例,染色体18序列并不显著长于与染色体14对齐的序列。利用截止值0bp,基于来自染色体18和14的DNA片段平均大小间的差值,可以区分整倍体和18三体病例。这些观测结果可以通过以下事实来解释:当胎儿是18三体时,短于母体染色体18序列的胎儿来源的染色体18序列的额外剂量会降低这种序列的总大小。这会导致染色体18的总大小分布更接近染色体14的分布。

[0158] V.利用总长度贡献的实例

[0159] 在下面的实例中,参照染色体由除染色体13、18和21外的常染色体组成。

[0160] 图18显示了在150bp处染色体18和参照染色体间由短片段贡献的总长度分数的差值($\Delta F_{(18-Ref)}$)。利用$\Delta F_{(18-Ref)}$为0.0003的诊断截止值,检测的18三体孕妇的灵敏度为93.3%,以及特异性为100%。

[0161] 图19显示了在150bp处染色体21和参照染色体间由短片段贡献的总长度分数的差值($\Delta F_{(21-Ref)}$)。利用$\Delta F_{(21-Ref)}$为0.007的诊断截止值,检测的21三体孕妇的灵敏度为100%,以及特异性为100%。

[0162] VII.参照染色体的选择

[0163] 通过母体血浆DNA的大小分析,例如利用方法600,可以以多种方式选择一条或多条参照染色体用于胎儿染色体非整倍性的无创性产前诊断。在多个实施方案中,可以选择不同的参照染色体。

[0164] 第一类参照染色体是如下的那些:其中来源于它们的DNA片段在特定的分析平台(或者具有严密匹配的分析性能的分析平台)上展现出与母体血浆中来源于可能涉及非整倍性染色体(例如染色体21、18或13)的那些片段相似的大小分布。在一个实施方案中,在该类分析中,如果高危染色体的大小秩次或绝对大小显示出从参照染色体的统计学显著的下降,则检测到了胎儿染色体非整倍性。在其它实施方案中,技术人员可以测量来源于高危染色体的片段和来自参照染色体的片段间的平均或中值大小差值。

[0165] 第二类参照染色体是如下的那些:其中来源于它们的DNA片段在统计学上短于胎儿是整倍体时母体血浆中来源于可能涉及非整倍性染色体(例如染色体21、18或13)的片段。当高危染色体利用特定平台测量时是最长的一个的时候,会出现这种情形。例如在图12中,当在母体血浆中测量时,染色体18-来源的DNA片段是常染色体中最长的。因此,技术人员可以选择其片段在统计学上显著短于染色体18的参照染色体。在一个实施方案中,在这类分析中,如果不能发现高危染色体的大小秩次或绝对大小与参照染色体在统计学上有显著差异,则在病例中检测出了胎儿染色体非整倍性。例如,这种策略已经用于图17的分析,如上文所描述的。

[0166] 第三类参照染色体是如下的那些:其中来源于它们的DNA片段在统计学上显著长于胎儿是整倍体时母体血浆中来源于可能涉及非整倍性染色体(例如染色体21、18或13)的片段。当高危染色体利用特定平台测量时是最短的一个的时候,会出现这种情形。因此,技术人员可以选择其片段在统计学上显著长于高危染色体的参照染色体。在一个实施方案中,在这类分析中,如果参照染色体和高危染色体间的大小秩次、秩次或绝对大小的差值增加,则在病例中检测出了胎儿染色体非整倍性。

[0167] 第四类参照染色体是具有相似GC含量的那些。染色体的GC含量可以影响测序反应中的定量读取。使由染色体间的GC含量差异引起的可能偏差最小化的一种方法是选择具有

相似GC含量的合适参照染色体。图20显示了不同染色体(NCBI build 36,第48版)的GC含量的列表。以GC含量递增的顺序列出染色体。现在提供了利用具有相似GC含量的参照染色体的实例。

[0168] 图21显示了在150bp处染色体13和参照染色体间由短片段贡献的总长度分数的差值($\Delta F_{(13\text{-}Ref)}$)。这里通过示例,我们利用染色体3、4、5和6作为参照染色体用于$\Delta F_{(13\text{-}Ref)}$分析。如图20所显示的,染色体3-6具有36.53%-38.79%的GC含量,其与染色体13的39.07%的GC含量相似。利用0.0038的诊断截止值,ΔF分析检测的13三体病例的灵敏度为95.7%,特异性为99.0%。

[0169] 预期GC偏差问题可能在不同程度上影响不同的测序平台。例如,利用不需要预扩增的平台,诸如Helicos平台(Harris TD,et al.Single-molecule DNA sequencing of a viral genome.Science 2008;320:106-9)、纳米孔(Lund J,Parviz BA.Scanning probe and nanopore DNA sequencing:core techniques and possibilities.Methods Mol Biol 2009;578:113-22)、或者来自Pacific Biosciences的单分子实时系统(Eid J,et al.Real-time DNA sequencing from single polymerase molecules.Science 2009;323:133-8),可能允许对染色体参照组有更广的选择。

[0170] IV.胎儿DNA浓度的利用

[0171] 在21三体(或其它三体)孕妇中,胎儿会将额外剂量的短于来自母体细胞片段的染色体21片段释放入母体血浆中。技术人员会预期到这些较短片段的浓度与母体血浆中胎儿DNA的浓度相关。换句话说,Y染色体来源的序列的分数浓度(fractional concentration)越高,21三体孕妇中染色体21来源的序列的测量大小会越短。

[0172] 图22显示了来自怀男性21三体胎儿的多个孕妇的结果。如从图22可以看出的,在染色体21序列的中值大小和与Y染色体对齐的序列百分比间的确存在负相关(r=-0.942,Pearson相关)。如果利用染色体21的大小秩次,也可以预期到类似的趋势,即当胎儿DNA分数浓度增加时,染色体21的秩数会增加,表明较短的片段。因为存在相关性,实施方案可以利用胎儿DNA浓度的测量作为本文描述的任一方法的参数。

[0173] 在这类分析中,可以通过本领域技术人员已知的任何方法,测量母体血浆中胎儿DNA的分数浓度和绝对浓度。如果胎儿是男性,可以通过母体血浆中来源于Y染色体的序列的分数浓度来测量胎儿DNA的浓度。另一实例是利用父本遗传的基因标志物诸如单核苷酸多态性或简单串联重复序列多态性或插入-缺失多态性。另一实例是利用表观遗传学标志物,诸如胎儿和母体DNA间差异甲基化的区域(Poon et al.Clin Chem 2002;48:35-41;Chiu et al.Am J Pathol 2007;170:941-950;Chan et al.Clin Chem 2006:52:2211-2218;美国专利6,927,028)。上述标志物可以利用本领域技术人员已知的方法来分析,这些方法包括聚合酶链式反应(PCR)、数字PCR、测序、大规模平行测序和定向的大规模平行测序。

[0174] 在一个实施方案中,技术人员可以改变诊断阈值(例如,本文描述的任一方法的截止值)用于检测与母体血浆中测量的胎儿DNA浓度有关的染色体非整倍性。因此,对于具有相对较高胎儿DNA的母体血浆样品,预期来源于可能涉及非整倍性染色体的血浆DNA分子的缩短程度比具有相对较低胎儿DNA浓度的母体血浆样品更为显著。

[0175] 因为ΔF与大小分布有关,ΔF还显示了与胎儿DNA浓度的相关性。图23A显示了Δ

$F_{(18-Ref)}$(染色体18对参照染色体)和胎儿DNA浓度间的相关性。30个T18病例中的10个怀有男性胎儿,由此可以通过这些样品中Y染色体序列的分数浓度来估计胎儿DNA的分数浓度。在$\Delta F_{(18-Ref)}$和Y染色体序列的分数浓度间存在显著的相关性(r=0.879,Spearman相关)。这些结果提示,母体血浆中染色体18序列的大小分布的缩短程度与18三体孕妇的母体血浆中胎儿DNA的分数浓度相关。

[0176] 低于诊断截止值的病例由空心圆表示,而具有超过截止值的差值的病例由实心圆表示。ΔF值小于诊断截止值(0.0003)(由空心圆表示)的两个病例当与ΔF值大于0.0003(由实心圆表示)的病例相比时,具有相对较低的胎儿DNA分数浓度。低胎儿DNA分数浓度可能是为什么这两个病例在图18中进行的分析中缺失的原因。因此,在一个实施方案中,如果胎儿浓度在样品中是低的,分类可能被忽略或重做。

[0177] 图23B显示了$\Delta F_{(21-Ref)}$(染色体21对参照染色体)和胎儿DNA浓度间的相关性。9个T21病例中的5个怀有男性胎儿。在$\Delta F_{(21-Ref)}$值与Y染色体序列的分数浓度间存在显著相关性(r=0.9,Spearman相关)。

[0178] 图23C显示了$\Delta F_{(13-Ref)}$(染色体13对染色体3、4、5和6)和胎儿DNA浓度间的相关性。23个13三体病例中的14个怀有男性胎儿。可以通过样品中Y染色体序列的分数浓度估计胎儿DNA的分数浓度。在$\Delta F_{(13-Ref)}$和Y染色体序列的分数浓度间存在正相关(r=0.644,Spearman相关)。$\Delta F_{(13-Ref)}$低于诊断截止值(0.0038)的病例用空心圆表示。低胎儿DNA分数浓度可能是为什么该病例在图21中进行的分析中缺失的原因。

[0179] IX大小分析和分子计数方法的比较

[0180] 图24显示了利用母体血浆DNA分析对本发明的实施方案与无创性检测胎儿非整倍性(13三体和18三体)的另一方法的精确度的比较。该实例说明了利用大小的实施方案与基于分子计数的方法(美国专利申请11/701,686;Chiu et al Trends Genet 2009;25:324-331;Chiu et al Proc Natl Acad Sci USA 2008;105:20458-20463;Fan et al Proc Natl Acad Sci USA 2008;105:16266-16271;美国专利公开2009/0029377)的比较。利用Chiu等描述的方法(Proc Natl Acad Sci USA 2008;105:20458-20463)分析了8个母体血浆样品(两个整倍体,两个18三体以及四个13三体)。对于每一病例,将之前Chiu等报道的利用Z-分值的分子计数方法利用分子计数与基于大小分析的实施方案的结果相比较。

[0181] 对于z-分值的计算,首先计算每一病例的感兴趣的染色体的表现百分比(percentage representation)。然后,计算参照病例的染色体表现的均值和标准偏差。在该数据集中,病例1、2、5、6、7和8用作参照组用于计算染色体18表现的均值和标准偏差。病例1、2、3和4用作参照组用于计算染色体13表现的均值和标准偏差。z-分值定义为偏离参照组均值的标准偏差数。染色体的显著过表现(overrepresentation)定义为z-分值大于3。病例3和4怀18三体胎儿,因此染色体18片段在她们的血浆中是过表现的。病例5、6、7和8怀13三体胎儿,但仅病例5和7在她们的血浆中显示了染色体13的过表现。尽管病例6和8怀13三体胎儿,但在母体血浆中没有显示染色体13的明显过表现。

[0182] 通过将与特定染色体对齐的所有片段的大小和与参照染色体对齐的片段大小进行比较,检测到了染色体的DNA片段大小的明显缩短。Mann-Whitney检验用于比较,并且将P-值小于<0.0001定义为存在显著差异。对于染色体13片段的大小分析,参照染色体是染色体5。对于怀整倍体或18三体胎儿的所有病例,染色体13的片段大小与染色体5的片段大小

没有显著差异。对于怀13三体胎儿的四个病例，染色体13片段显著短于染色体5片段，这暗示与非13三体病例相比，染色体13的片段大小缩短。因此与通过z-分值方法鉴定了四个13三体病例中的两个相比，通过本发明正确地鉴定所有四个13三体病例。

[0183] 对于染色体18片段的大小分析，参照染色体是染色体12。对于所有怀整倍体或13三体胎儿的病例，染色体18的片段显著长于染色体12的片段。对于怀18三体胎儿的两个病例，染色体18片段的大小与染色体12片段的大小没有显著差异，这暗示与怀非18三体胎儿的病例相比，染色体18的片段大小是缩短的。换句话说，两个18三体病例被正确地分类。

[0184] X.多态性检测和遗传病症的诊断

[0185] 母体血浆DNA的大小分析还可以用于胎儿基因型的无创性检测。胎儿基因型可以用于确定胎儿是否遗传了突变基因，具有特定等位基因的失衡或者其它序列失衡或用于其它目的。在这个实施方案中，一个等位基因可以是参照基因组位置(序列)，另一等位基因可以是受测试的基因组位置。因此，利用参照序列的任一方法还可以应用于确定基因型或其它序列失衡。

[0186] 在一个实施方案中，可以通过大小差异(失衡)是否存在于母体样品中的等位基因间(例如当母亲在该等位基因处是杂合时)来确定序列失衡(以及由此可能的基因型)。例如，如果在样品中的等位基因间没有大小属性的差异，则可以确定胎儿具有与母亲相同的基因型。作为另一个实例，如果在样品中的等位基因间存在大小属性的差异，则可以确定胎儿具有与母亲不同的基因型。

[0187] 在以下的实例中，对于特定的基因座，母亲是杂合的(即具有一个拷贝的N等位基因和一个拷贝的M等位基因，表示为NM)。字母N和M在名称上分别代表野生型(N代表正常)和突变型(M代表突变)等位基因。然而，N和M等位基因可以对应于任何两个不同的等位基因，并不必然对应于野生型和/或突变型。在一个实施方案中，M可以被视为高危基因组序列，N被视为参照序列。在这种情况下，技术人员可以理解利用参照序列的上述任一方法可以用于确定基因型。

[0188] 在非怀孕妇女中，携带两个等位基因的分子的平均大小会是相同的。然而，在孕妇的血浆中，存在来自母亲和胎儿的DNA分子的混合物。母体来源的DNA分子长于胎儿来源的DNA分子。如果母亲和胎儿都具有两个等位基因(即都具有N和M等位基因)，这两个等位基因会具有长DNA分子和短DNA分子的相等贡献。因此，产生的N和M等位基因的大小分布会是相同的。相反地，如果母亲和胎儿的基因型不同，例如如果母亲是NM，胎儿是MM，则N和M等位基因的大小分布会是不同的。换句话说，血浆中两个等位基因的大小分布受胎儿基因型的影响。图25A-25C显示了按照本发明的实施方案，孕妇和胎儿的基因型在不同情形下的图解。

[0189] 在图25A中，胎儿具有NN基因型，而母亲的基因型是NM。条状物2510的长度分别表示来自母亲和胎儿的两个等位基因中的一个的片段的平均(均值)大小。如上所述，胎儿比母亲具有更小的平均大小。因此，长条状物代表母亲DNA，而短条状物代表胎儿DNA。

[0190] 因为母亲和胎儿都贡献了等位基因N，而只有母亲贡献了等位基因M，所以具有等位基因N的分子大小分布短于具有等位基因M的分子的大小分布。换句话说，在具有NM基因型的孕妇中，与等位基因M相比，等位基因N的较短的大小分布暗示胎儿基因型是NN。因此，当N的大小分布(即平均大小)小于M某一截止值(例如，百分比或绝对值)时，可以鉴定胎儿具有两个野生型(N)等位基因。

[0191] 在图25B中，胎儿具有NM的基因型。母亲和胎儿都贡献等位基因N和M。因此，具有等位基因N和M的分子的大小分布是相同的。在具有NM基因型的孕妇中，等位基因M和N的相同的大小分布将指示胎儿基因型为NM。因此，当N的大小分布(例如平均大小)在某一截止值(例如，百分比或绝对值)内约等于M时，可以鉴定胎儿具有一个野生型(N)等位基因和一个突变型(M)等位基因。

[0192] 在图25C中，胎儿具有MM基因型。因为母亲和胎儿都贡献了等位基因M，而只有母亲贡献了等位基因N，所以具有等位基因M的分子的大小分布会短于具有等位基因N的分子的大小分布。在具有NM基因型的孕妇中，相比等位基因N，等位基因M更短的大小分布指示胎儿基因型是MM。因此，当M的大小分布(即平均大小)小于N某一截止值(例如，百分比或绝对值)时，可以鉴定胎儿具有两个突变型(M)等位基因。

[0193] 这一方法还可以用于分析其中母亲是纯合(如NN或MM)的情况。如果胎儿具有不同的基因型，母体样品的大小分布也会改变，由此可以确定胎儿的基因型。同样，如果大小分布没有改变，则可以确定基因型与母亲的基因型相同，如上文所述的其中母亲是杂合的病例。

[0194] 在一些实施方案中，确定相对于母亲的基因型是否存在失衡(例如，一个M一个N时暗示没有失衡)或其它情况可以利用截止值来进行。例如，如果与母亲基因型存在足够大的偏差(例如，通过百分比)，则胎儿可以被确定为遗传了具有较小大小分布的等位基因。在一个实施方案中，截止值可以取决于母体样品中胎儿核酸的百分比。如果存在较高百分比的胎儿核酸，则可以预期较大的偏差，由此可以使用较大的截止值(例如，对于一个等位基因相对于另一等位基因的大小分布的差值)。如果存在较低百分比的胎儿核酸，则预期有较小的偏差，由此可以使用较小的截止值。

[0195] 在一个实施方案中，父亲的基因型可以用于确定母亲的哪个等位基因的大小分布由于胎儿核酸发生了变化。在其中父亲的基因型是纯合的情况下，这可以允许将可能的胎儿基因型缩小至仅判断从母亲遗传了什么等位基因，因为来自父亲的那个等位基因是已知的。经过这种缩小后，胎儿基因组的确定可能更为精确，因为只有两种可能性需要被检验。在一项实施中，使用相同的截止值，而不管从父亲遗传了哪个基因型。在另一项实施中，可以使用不同的截止值。

[0196] 在多个实施方案中，本文提及的任一大小分布都可以用于这种序列失衡确定中。在一些实施方案中，还可以提供精确水平。例如，除了失衡和平衡分类外，还可以使用“未确定的”分类。以这种方式，可以高可信度地确定一些测定，而中间区域的值可能需要进一步的数据点。

[0197] XI.胎儿单体型遗传的大小分析

[0198] 大小分析的应用还可以进一步延伸至确定哪个母体单体型传递给了胎儿。单体型可以指多个基因座处的等位基因。术语“单体型”的定义可以在本申请的定义章节中找到。胎儿单体型可以用于确定胎儿是否遗传了突变基因，是否具有特定等位基因的失衡或用于其他目的。因此，可以以与基因型相似的方式利用单体型，但因为存在更多的基因座，较小体积的血液样品可以用于实现确定胎儿单体型的相同或甚至更好的统计学可信度。在一方面，可以相对于母亲单体型确定序列失衡。在本文，单体型可以表现为一系列多态性(例如SNP)，每一多态性位于序列变异已知的基因组中的特定位置处。

[0199] 在一个实施方案中，提供了通过分析母体血浆中的序列失衡确定胎儿单体型的方法。在一方面，单体型间大小属性(size profile)间隔值(例如差值)用于确定序列失衡。在一个实施方案中，母亲的单体型(例如Hap I和Hap II)通过母体样品(例如，不含有胎儿核酸的样品)的分析(例如测序(He D et al.Bioinformatics 2010;26:i183-i190)或单分子单体分型(Ding C et al.Proc Natl Acad Sci USA 2003;100:7449-7453以及Xiao M et al.Nat Methods 2009;6:199-201)来确定。在另一实施方案中，母亲的单体型可以利用对母亲的父母、兄弟姐妹、之前的孩子或其他亲属的分析来确定。在另一实施方案中，对于强连锁不平衡处的多态性，了解母亲在一基因座处的基因型可以暗示在其它基因座处的基因型，例如当等位基因正常地出现在同一序列(即单体型)中时。因此，母亲的单体型可以从基因型的一个测量隐含地确定。还可以确定超过一个基因座处的基因型，其中每一个确定的基因型可以暗示在其它基因座处的基因型，从而推导出单体型。

[0200] 在一个实施方案中，还可以确定父亲的基因型。这一信息可以用于确定父亲在特定SNP处是纯合的还是杂合的。可以直接确定父亲在每一基因座的等位基因。因此，技术人员可以确定是否每一个父本等位基因与母亲的Hap I或Hap II上的等位基因相同(称为α型或β型SNP)。

[0201] 在另一实施方案中，还确定了父亲的单体型。父亲的单体型可以通过对父本样品的分析(例如测序(He D et al.Bioinformatics 2010;26:i183-i190)或单分子单体分型(Ding C et al.Proc Natl Acad Sci USA 2003;100:7449-7453以及Xiao M et al.Nat Methods 2009;6:199-201)，或者通过对父亲的父母、兄弟姐妹、之前的孩子或其他亲属的分析来确定。可选择地，对于强连锁不平衡处的多态性，了解父亲在一个基因座处的基因型可以暗示在其它基因座处的基因型，例如当等位基因正常地出现在同一序列(即单体型)中时。因此，父亲的单体型可以从基因型的一个测量隐含地确定。还可以确定超过一个基因座处的基因型，其中每一个确定的基因型可以暗示在其它基因座处的基因型，从而推导出单体型

[0202] 这种实施方案的一个应用可以是通过分析其中父亲和母亲都是杂合的SNP来确定含胎儿核酸的母体样品是否存在序列失衡(例如，通过检测大小属性的差异)。由此从是否存在序列失衡推导出胎儿基因型或单体型。

[0203] 图26显示了一个实例，其中在待分析的SNP基因座处母亲是杂合的，父亲是纯合的。为了进行大小分析，集中在母亲是杂合的且父亲是纯合的SNP的子集上。母亲的两个同源染色体分别被称为Hap I和Hap II。对于这些SNP中的每一个，我们可以确定两个母体等位基因的哪个位于Hap I上，哪个位于Hap II上。如果父本等位基因与Hap I上的母体等位基因相同时，SNP被定义为α型；以及如果父本等位基因与Hap II上的母体等位基因相同时，SNP被定义为β型。确定胎儿基因组的进一步描述可以见于上文提及的申请“Fetal Genomic Analysis From A Maternal Biological Sample(来自母体生物样品的胎儿基因组分析)”中。

[0204] 一旦母亲的单体型和父亲的基因型或单体型已知，可以通过鉴定在SNP子集中是否存在序列失衡来分析与每一SNP相关的片段的大小分布以确定胎儿单体型。在可选的实施方案中，并不确切地知道父本基因型或单体型，但基于例如受测试群体的已知基因型或单体型的频数，技术人员可以利用如统计程序推导出可能的父本基因型或单体型。然后确

定每一SNP的序列失衡(如通过大小失衡确定的),例如在X节中讨论的。

[0205] 例如,如果父亲对于Hap I上的等位基因是纯合的(α型),则要么不存在大小失衡(胎儿从母亲遗传了Hap II,因此正如母亲一样,是Hap I和Hap II杂合的),要么存在大小失衡,其中Hap I具有较小的大小分布(胎儿从母亲遗传了Hap I,因此是Hap I纯合的)。如果父亲是Hap II纯合的(如下文描述的β型),则要么不存在失衡(胎儿从母亲遗传了Hap I,因此是Hap I和Hap II杂合的),要么存在失衡,其中Hap II具有较小的大小分布(胎儿从母亲遗传了Hap II,因此是Hap II纯合的)。还可以利用不确定的分类,或者在平衡和失衡间变化水平的确定性。一般地,类似的截止值用于任一类型。在多个实施方案中,可以利用本文提及的任一大小分布。

[0206] 在一个实施方案中,可以分析多个SNP位置中每一个的片段的大小分布以确定胎儿的两个单体型。例如,可以将一个单体型上SNP等位基因片段的大小分布间的差异与另一单体型上的SNP等位基因的大小分布进行比较。可以以多种方式进行统计学分析,例如通过确定每一SNP,然后将大多数(平衡、失衡以及或许包括不确定的)看作单体型。作为另一个实例,可以合计整个SNP的大小(例如以便获得与截止值比较的平均或中值大小分布)。或者,可以利用两者的结合。另一实施可以利用数据点的极值,例如特定SNP的最小差值。

[0207] 图27显示了其中当父本单体型如图26所示时,胎儿从母亲遗传了Hap I的实例。对于α型SNP(由非阴影框围绕),胎儿已经从父亲遗传了与位于母体Hap I上的等位基因相同的等位基因。因此,胎儿对于Hap I上的等位基因是纯合的。因此,母体血浆中Hap I上的等位基因的大小分布短于Hap II上的等位基因。对于β型SNP(由阴影框围绕),胎儿已从父亲遗传了与母体Hap II上的等位基因相同的等位基因。因此,胎儿会是杂合的。因此,母体血浆中Hap I和Hap II上等位基因的大小分布具有相同的大小分布。

[0208] 在一个实施方案中,同一类型的SNP(α型或β型)可以一起分析。对于α型SNP,携带Hap I上等位基因的分子的大小分布会短于携带Hap II上等位基因的分子的大小分布。对于β型SNP,携带Hap I和Hap II上等位基因的分子的大小分布是相同的。换句话说,如果携带Hap I的分子的大小分布短于携带Hap II的分子的大小分布,胎儿是Hap I纯合的。如果携带Hap I和Hap II的分子的大小分布相同,则胎儿是杂合的。

实施例

[0209] 下面的实验用于测试胎儿单体型遗传性分析的精确度。招募了来产科门诊进行β-地中海贫血产前诊断的一对夫妇。采集了父亲和母亲的血液样品。对于母亲,血液样品在12周妊娠的绒毛膜绒毛取样(CVS)前采集。CVS后,储存一部分用于实验。

[0210] 从父亲和母亲的血沉棕黄层(buffy coat)和CVS样品中提取DNA。通过Affymetrix人类全基因组SNP阵列6.0(Affymetrix Genome-Wide Human SNP Array 6.0)系统对这些DNA样品进行分析以确定父亲、母亲和胎儿的基因型。在该实验中,CVS数据用于推导母体单体型。然而,在测试的临床实施中,可以通过上文描述的其他方法推导母体单体型。CVS数据还可以用于确认利用本文不需要CVS的方法进行确定的精确度。

[0211] 在目前的示例中,我们集中在其中母亲是杂合的,且父亲是纯合的信息性SNP子集上。在SNP的该子集中,夫妇和胎儿的基因型用于构建母亲的单体型。我们将单体型I(Hap I)定义为母亲传递给胎儿的一系列等位基因,而将单体型II(Hap II)定义为胎儿没有从母

亲接收的一系列等位基因。

[0212] 然后,我们将信息性SNP分为两个亚型,即α型和β型。对于α型SNP,父本等位基因与Hap I上母体等位基因相同。对于这些SNP,胎儿从父母遗传了相同的等位基因(Hap I上的等位基因),因此对于Hap I上的SNP,胎儿是纯合的。对于β型SNP,父本等位基因与Hap II上母体等位基因相同。对于这些SNP,胎儿从母亲遗传了Hap I上的等位基因,以及从父亲遗传了不同的等位基因(与Hap II上的等位基因相同),因此胎儿是杂合的。

[0213] 利用Illumina基因组分析仪平台,将从母亲血浆中提取的DNA进行大规模平行测序。进行血浆DNA分子的双末端测序。每一分子在每末端测序50bp,因此每一分子总共测序100bp。利用位于深圳的北京基因组研究所的SOAP2程序(soap.genomics.org.cn/)(Li R et al.Bioinformatics 2009,25(15):1966-7),将每一序列的两末端与非重复序列掩蔽的人基因组(non-repeat-masked human genome)(从UCSC genome.ucsc.edu下载的Hg18NCBI.36)对齐。

[0214] 在一个实施方案中,由单体型短片段贡献的总长度分数的统计值用于确定哪个母体单体型传递给胎儿。作为实例,染色体22用于说明大小分析如何用于推导哪个母体单体型传递给了胎儿。首先,我们将染色体22分为7个节段,每一节段包含50个其中母亲是杂合的,且父亲是纯合的信息性SNP(染色体22和节段都是序列的例子)。对于每一节段,覆盖这些信息性SNP的DNA片段(为序列一部分的分子的例子)被分为两组,即Hap I和Hap II,按照这些片段对应于这两个母体单体型的哪一个而分。对于每一节段,确定映射至母体Hap I和母体Hap II的所有片段的总长度之和。然后,类似地,确定每一节段中映射至母体Hap I和Hap II的短片段的总长度之和。出于示例的目的,在该实施例中,将150bp或更短的片段定义为用于计算短片段的总长度的短片段。从这些长度中,对于映射至每一节段内Hap I和Hap II的DNA片段,可以计算由短片段贡献的总长度的分数。

[0215] 图28显示的表说明了按照本发明的实施方案,对染色体22上α型SNP的大小分析。$\Delta F_{(Hap\ I-Hap\ II)}$表示Hap I和Hap II间由短片段贡献的总长度分数的差值。当排除了仅由28个SNP构成的最后一个节段时,$\Delta F_{(Hap\ I-Hap\ II)}$的范围为0.0288至0.0701。$\Delta F_{(Hap\ I-Hap\ II)}$为正值表示Hap I的F值总是大于Hap II的F值。因为F值被定义为由短片段贡献的总长度的分数,较高的F值表明由短片段贡献的总长度的分数较高。换句话说,这些结果表明对于被分析的由50个SNP组成的每个区域而言,携带Hap I上等位基因的DNA片段短于携带Hap II上等位基因的DNA片段。这表明Hap I的大小分布短于Hap II的大小分布。因此,我们可以推导出胎儿对于Hap I上的等位基因是纯合的。如上提到的,在一个实施方案中,可以基于Hap I和Hap II中的两个基因组位置间的间隔值(例如,ΔF),对每一节段进行分类,然后基于各个节段的分类进行总分类。在另一实施方案中,可以从每一节段的间隔值确定总间隔值(例如,平均间隔值),以及该总的统计值可以用于确定分类。

[0216] 在一个实施方案中,约0.02的截止值可以用于确定是否存在失衡。如果使用ΔF的中值或平均数,可以利用更大的截止值并且仍然是精确的。截止值还可以用于不确定的结果,例如在0.015至0.025间的区域可以是不确定的,需要进一步的分析。

[0217] 图29显示的表说明了按照本发明的实施方案,对染色体22上的β型SNP的大小分析。ΔF(Hap I-Hap II)表示由Hap I和Hap II的短片段贡献的总长度分数的差值。ΔF(Hap I-Hap II)值的范围为-0.0203至0.0207,中值为0.0003。小ΔF(Hap I-Hap II)值与映射至

Hap I和Hap II的片段的相同大小分布相配。因此,我们可以推导出对于Hap I和Hap II,胎儿是杂合的。因为β型SNP被定义为父本等位基因与Hap II上母体等位基因相同的SNP,该结果暗示胎儿从母亲遗传了Hap I。

[0218] 图30显示了按照本发明的实施方案,染色体22上的α型和β型SNP的ΔF(Hap I-Hap II)的图。对于α型SNP,Hap I片段大小分布短于Hap II片段的大小分布,这导致ΔF(Hap I-Hap II)值大于零。对于β型SNP,在Hap I和Hap II片段大小分布间没有差异,因此ΔF(Hap I-Hap II)值聚集在0附近。利用0.025的截止值,对于所有13个α型节段和21个β型节段,ΔF(Hap I-Hap II)分析可以正确地推导胎儿Hap I的遗传性。

[0219] XII.利用靶向测序的实施例

[0220] 下文的实施例证实了本发明的实施方案基于大小的诊断方法可以用于靶向测序模式。在这种模式中,具有诊断兴趣的基因组区域被特异性靶向以测序。与其中一些测序力用于与诊断应用不直接相关的区域的涉及随机测序的情形相比,这一模式具有集中在感兴趣区域进行测序的优势。因此,预期靶向测序模式能够增加通量并降低系统成本。可以利用本领域技术人员已知的任何模式来实施靶向测序,包括液相捕获系统(例如,Agilent SureSelect系统)、固相捕获系统(例如,Roche NimbleGen系统)或者通过靶标特异性扩增(例如,RainDance系统)。

[0221] 从8个孕妇的前三个月期间收集血液样品。通过DSP DNA血液微型试剂盒(DSP DNA Blood Mini Kit)(Qiagen)从每一病例的3.2mL血浆中提取DNA。在绒毛膜绒毛样品(母体血液样品采集后收集的)上进行的染色体核型分析指示四个胎儿具有T21(UK229、UK510、UK807、PW421),而其他四个是整倍体男性(PW226、PW263、PW316、PW370)。

[0222] 按照生产商关于染色质免疫沉淀测序(ChromatinImmunoprecipitationSequencing)样品制备的操作步骤,通过双末端样品制备试剂盒(Illumina),将每一病例5-30ng血浆DNA用于DNA文库构建。利用QIAquick PCR纯化试剂盒(Qiagen)中提供的旋转柱直接纯化适配子连接的DNA,而没有进一步的大小选择。然后将适配子连接的DNA利用标准引物的15个循环的PCR进行扩增。引物是Illumina的PCR引物PE 1.0和2.0。DNA文库通过利用NanoDrop ND-1000分光光度计(NanoDrop Technologies)进行定量,并利用DNA 1000试剂盒(Agilent)在2100Bioanalyzer上泳动,从而核对大小分布。每一样品以约290bp的平均大小产生0.6-1μg扩增的血浆DNA文库。

[0223] SureSelect人全外显子(SureSelect Human All Exon)捕获文库获自Agilent,并覆盖了37.8Mb的人类外显子(目录号:5190-2310)。对于该研究中的所有8个病例,按照生产商的说明,将每一病例500ng扩增的血浆DNA文库与捕获探针在65℃下孵育24小时。杂交后,通过利用链霉亲和素包被的磁珠(Dynal DynaMag-2Invitrogen)拉下生物素化的探针/靶标杂交物来选择捕获的靶标,并用MinElute PCR纯化试剂盒(Qiagen)纯化。最后,通过利用来自Agilent的SureSelect GA PE引物的12个循环的PCR扩增,富集靶向的DNA文库。通过QIAquick PCR纯和试剂盒(Qiagen)纯化PCR产物。

[0224] 将有或没有靶标富集的8对文库加载于两个流通池的16个通道上,然后通过利用36-bp×2双末端模式的基因组分析仪IIx(Illumina)进行测序。利用短寡核苷酸比对程序2(Short Oligonucleotide Alignment Program 2)(http://soap.genomics.org.cn/),将所有36-bp测序的读数与未掩蔽的人参照基因组(Hg18)(http://genome.ucsc.edu)对齐。

双末端读数的片段大小被限定为从40bp到600bp。从两末端最远的核苷酸的坐标推断出每一测序的DNA片段的大小。

[0225] 用于胎儿21三体产前检测的大小分析

[0226] 在该实施例中，计算染色体21和参照染色体由短片段贡献的总长度的分数，分别表示为F_{21}和F_{ref}。参照染色体由除染色体13、18和21外的所有常染色体组成。通过600bp或更小的所有DNA片段的长度之和计算总长度。染色体21和参照染色体间由短片段贡献的长度分数的差值(ΔF)计算为$F_{21}-F_{ref}$。

[0227] 图31A的表提供了按照本发明的实施方案，没有靶标富集的血浆DNA的大小分析。不同的列提供了染色体21小于或等于150bp的片段的总长度、染色体21小于或等于600bp的片段总长度、以及两者比值F_{21}。其它列提供了参照染色体小于或等于150bp的片段的总长度、参照染色体小于或等于600bp的片段总长度、以及两者比值F_{ref}。最后一列是两个分数间的差值ΔF。

[0228] 图31B的表提供了按照本发明的实施方案，有靶标富集的血浆DNA的大小分析。图31B的列具有与图31B的表相同的数据格式。

[0229] 图32是有或没有靶标富集的T21和整倍体样品的ΔF的图。对于没有靶标富集的样品，利用ΔF的0.005截止值，可以区分来自T21和整倍体孕妇的血浆样品，精确度为100％。对于有靶标富集的样品，利用ΔF的0.004的截止值，可以区分来自T21和整倍体孕妇的血浆样品，精确度为100％。该实施例证实了利用靶向测序可以进行基于大小的分析。对于T21的检测，利用染色体21和参照染色体的靶向测序是有利的，这样使得50％的测序针对前者，而其余的针对后者。这种设计可以降低测序力在与T21检测不直接相关的区域的浪费。这种设计可以允许利用多重测序术(例如，通过利用带索引或条形码的测序术)对来自多个患者的样品进行测序。

[0230] 利用任何合适的计算机语言，如Java、C++或使用例如常规或面向对象技术的Perl，本申请所述的任何软件组件或函数可以作为由处理器运行的软件代码来执行。软件代码可在用于存储和/或传输的计算机可读介质上存储为一系列指令或命令，合适的介质包括随机存取存储器(RAM)、只读存储器(ROM)、诸如硬盘或软盘的磁性介质或诸如光盘(CD)或DVD(多功能数码光盘)的光学介质、闪存等。计算机可读介质可以是此类存储或传输装置的任何组合。

[0231] 此类程序也可以利用适合通过有线、光学和/无线网络传播的载波信号来编码和传输，该网络符合包括国际互联网在内的各种协议。因此，本发明实施方案的计算机可读介质，可以利用此类程序编码的数据信号产生。用程序代码编码的计算机可读介质可以与兼容的装置组装，或由其它装置(如经由互联网下载)独立地提供。任何此类计算机可读介质可以位于一个计算机程序产品上或在该产品内(例如，硬盘或整个计算机系统)，并且可以存在于系统或网络内不同计算机程序产品上或在该产品内内。计算机系统可以包括显示屏、打印机或向用户提供本文所提到的任何结果的其他合适的显示器。

[0232] 计算机系统的实例显示在图33中。图33显示的子系统通过系统总线3375相互连接。显示了其它子系统，诸如打印机3374、键盘3378、硬盘3379、与显示适配器3382相连的显示屏3376以及其它。与I/O控制器3371相连的外围装置和输入/输出(I/O)装置，可以通过本领域已知的多种元件诸如串行端口3377与计算机系统连接。例如，串行端口3377或外部界

面3381可以用于将计算机装置与诸如因特网的广域网、鼠标输入装置或扫描仪连接。通过系统总线的相互连接允许中央处理器3373与每一子系统通讯，并且控制来自系统内存3372或硬盘3379的指令的执行，以及子系统间信息的交换。系统内存3372和/或硬盘3379可以体现为计算机可读介质。本文提及的任何值可以从一个组件输出到另一组件，并且可以输出至用户。

[0233] 计算机系统可以包括多个相同的组件或子系统，例如通过外部界面3381或者通过内部界面连接在一起的组件或子系统。在一些实施方案中，计算机系统、子系统或装置可以通过网络进行通信。在这种情形下，一台计算机可以被认为是客户端，而另一台计算机是服务器，其中每一台都可以是同一计算机系统的一部分。客户端和服务器都可以包括多个系统、子系统或组件。

[0234] 可以以任何合适的方式合并特定实施方案的具体细节，而不会偏离本发明实施方案的精神和范畴。然而，本发明的其它实施方案可以涉及与每一单独方面有关的具体实施方案，或这些单独方面的具体合并。

[0235] 出于示例和描述的目的，上文呈现了本发明示例性实施方案的描述。不意图是全面的或将本发明限制为所述的准确形式，并且根据上文的教导，可以做出许多修饰和变化。为了最好地解释本发明的原理及其实践应用而选择和描述了实施方案，由此使本领域技术人员在各种实施方案中，并且通过适于所考虑的具体用途的各种修饰来最佳地利用本发明。

[0236] 出于所有目的，本文所引用的所有出版物、专利和专利申请都通过引用的方式以其整体并入本文。

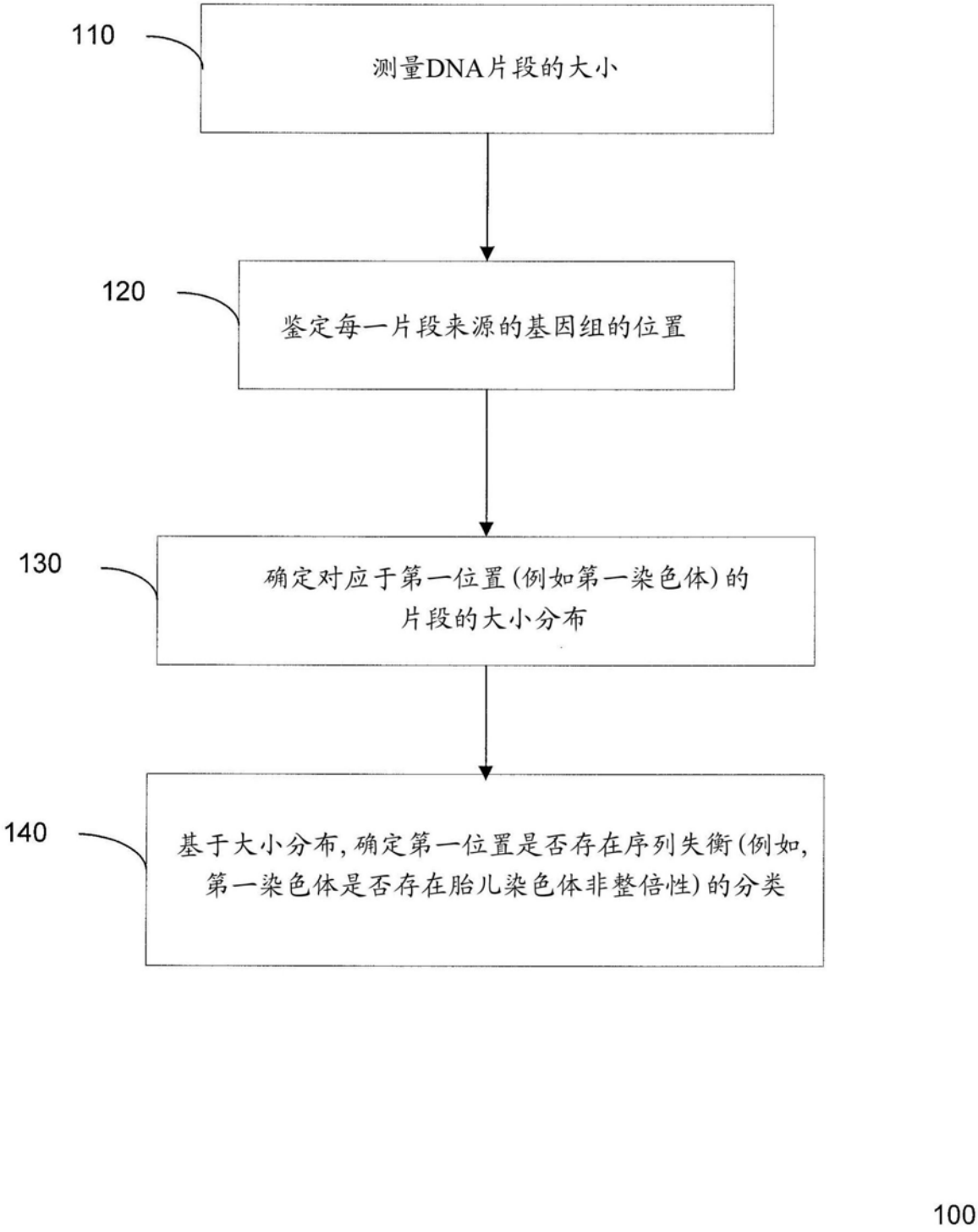

110
测量DNA片段的大小
120
鉴定每一片段来源的基因组的位置
130
确定对应于第一位置(例如第一染色体)的片段的大小分布
140
基于大小分布,确定第一位置是否存在序列失衡(例如,第一染色体是否存在胎儿染色体非整倍性)的分类
100

图1

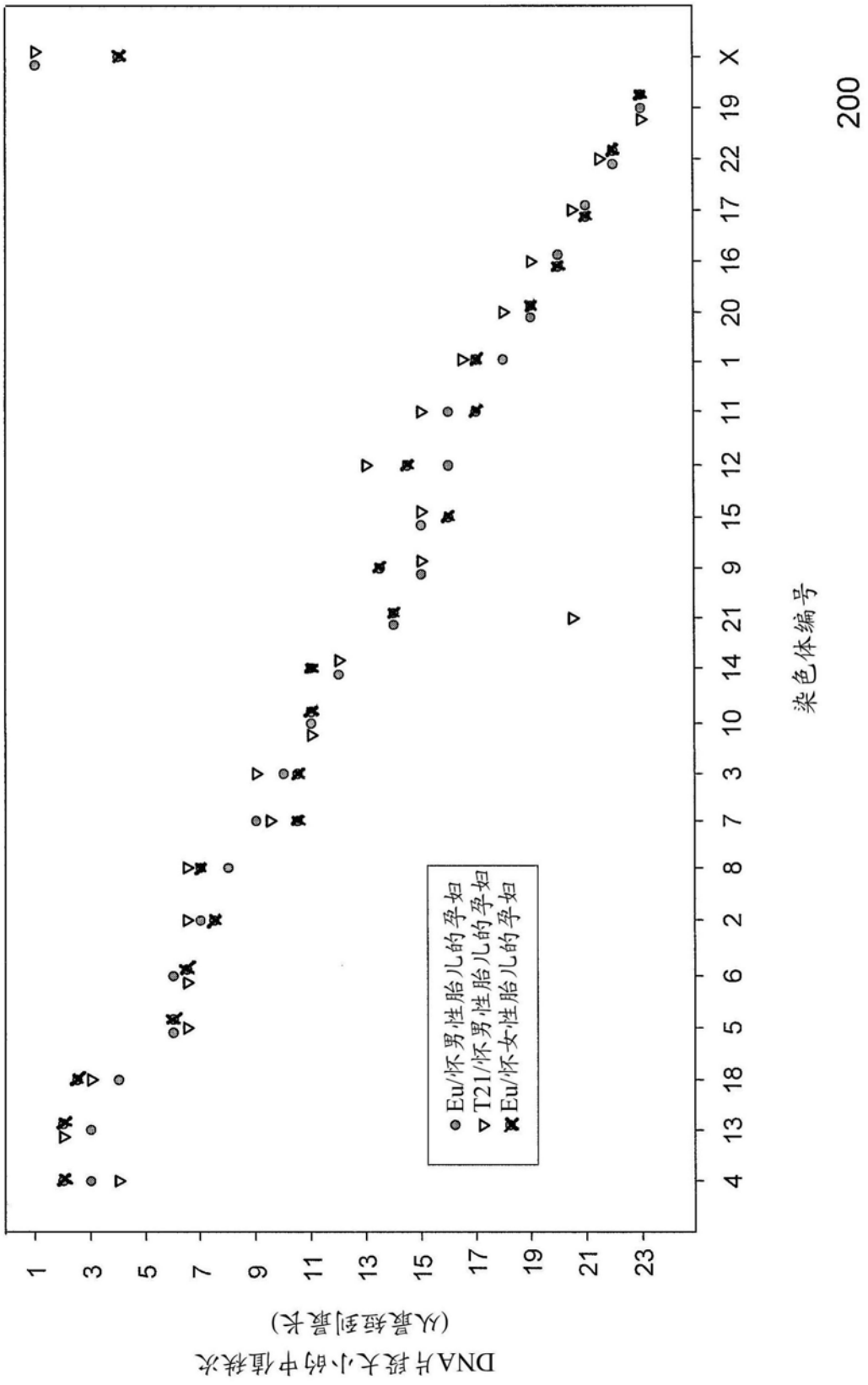

DNA片段大小的中值秩次
(从最短到最长)
染色体编号
200
Eu/怀男性胎儿的孕妇
T21/怀男性胎儿的孕妇
Eu/怀女性胎儿的孕妇

图2

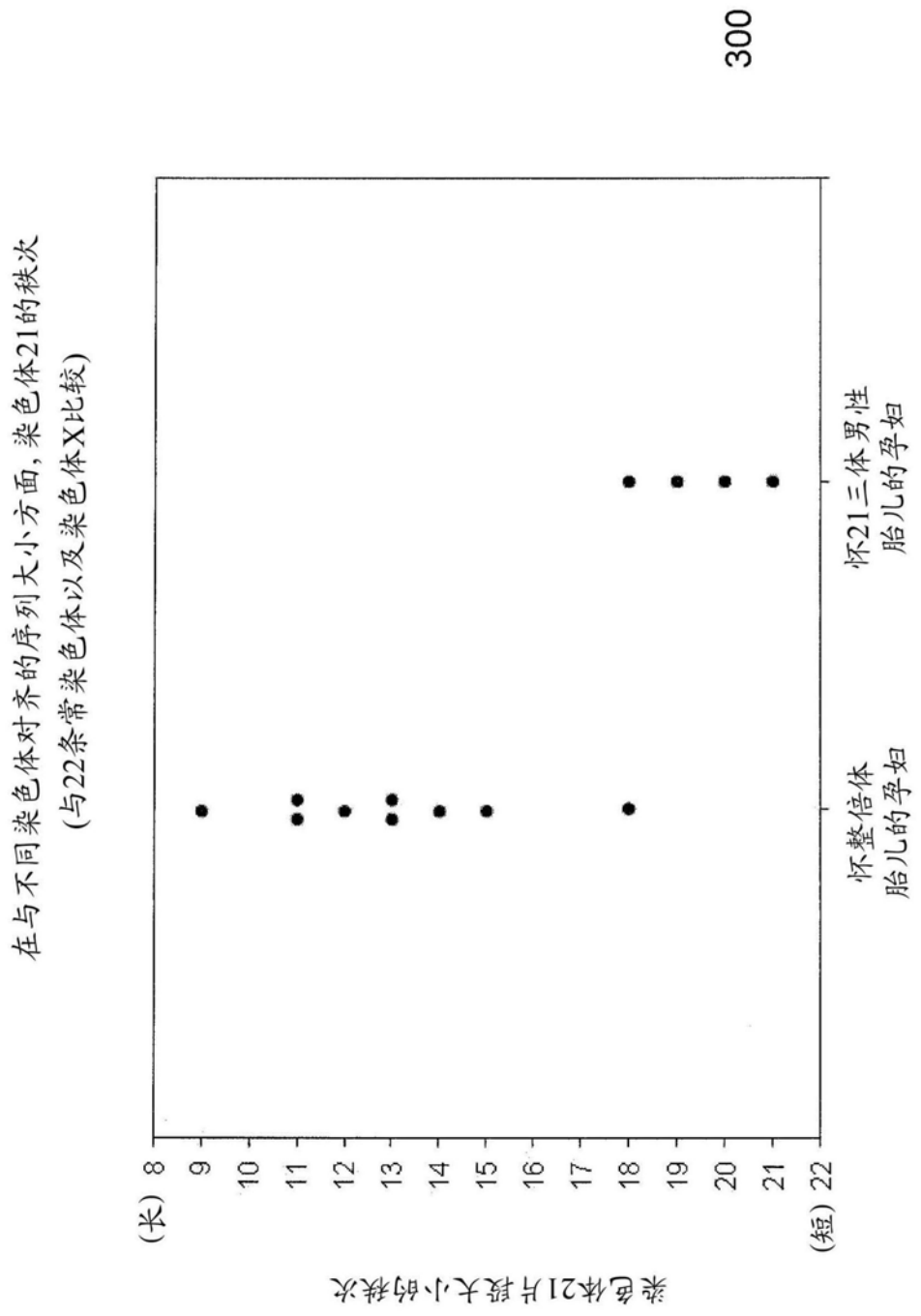

300
在与不同染色体对齐的序列大小方面，染色体21的秩次
(与22条常染色体以及染色体X比较)
(长) 8
9
10
11
12
13
14
15
16
17
18
19
20
21
(短) 22
染色体21片段大小的秩次
怀整倍体
胎儿的孕妇
怀21三体男性
胎儿的孕妇

图3

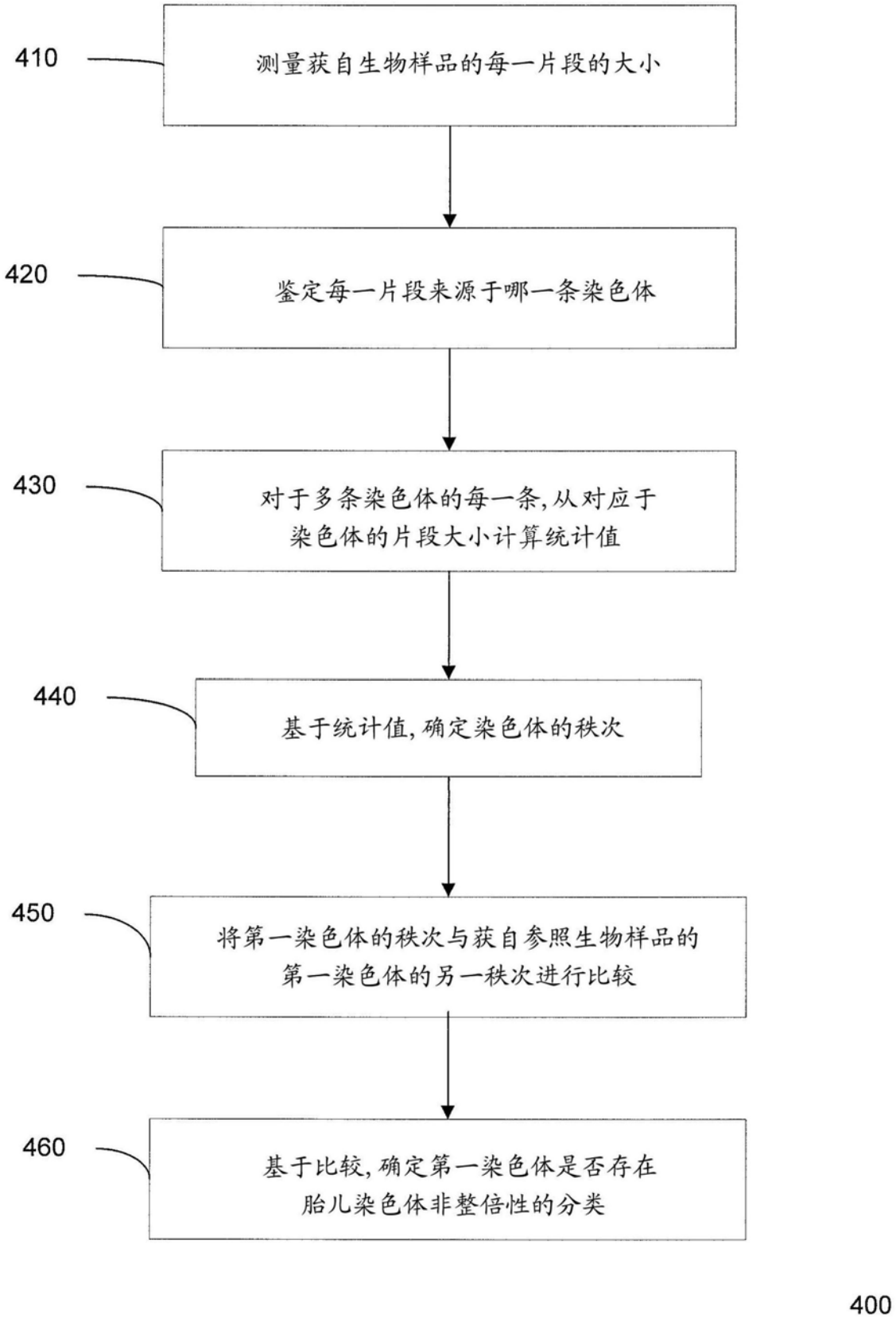

410
测量获自生物样品的每一片段的大小
420
鉴定每一片段来源于哪一条染色体
430
对于多条染色体的每一条,从对应于染色体的片段大小计算统计值
440
基于统计值,确定染色体的秩次
450
将第一染色体的秩次与获自参照生物样品的第一染色体的另一秩次进行比较
460
基于比较,确定第一染色体是否存在胎儿染色体非整倍性的分类
400

图4

样品名称	三个月	胎儿核型	染色体21片段的平均大小	染色体7片段的平均大小	染色体7和染色体21的均值间的差值	P值*(染色体7对染色体21)	染色体14片段的平均大小	染色体14和染色体21的均值间的差值	P值*(染色体14对染色体21)
M4800-男性	3rd	整倍体	156.2	157.0	0.8	0.08	156.1	-0.1	0.822
M4801-男性	3rd	整倍体	151.2	151.7	0.5	0.085	151.5	0.3	0.564
M4814-女性	3rd	整倍体	157.8	157.3	-0.5	0.106	156.8	-0.9	0.022
PW006-Eu-男性	1st	整倍体	146.0	146.1	0.0	0.527	145.9	-0.1	0.988
PW007-Eu-男性	1st	整倍体	152.1	152.7	0.6	0.033	152.3	0.2	0.679
PW008-Eu-男性	1st	整倍体	147.2	147.3	0.1	0.538	147.0	-0.2	0.688
PW012-Eu-男性	1st	整倍体	146.7	146.8	0.1	0.333	146.7	0.1	0.68
PW020-Eu-男性	1st	整倍体	139.7	140.0	0.3	0.174	139.7	0.0	0.741
PW009-Eu-女性	1st	整倍体	151.9	152.2	0.3	0.068	152.1	0.2	0.405
PW010-Eu-女性	1st	整倍体	148.3	148.5	0.2	0.101	148.4	0.1	0.286
PW016-Eu-女性	1st	整倍体	141.5	141.1	-0.4	0.782	141.2	-0.3	0.842
PW022-Eu-女性	1st	整倍体	140.8	140.7	-0.1	0.362	140.5	-0.4	0.94
M2849-T21-男性	1st	T21	145.9	147.0	1.1	<0.001	146.8	0.8	<0.001
M4386-T21-男性	1st	T21	144.3	146.3	2.0	<0.001	145.9	1.6	<0.001
M4467-T21-男性	1st	T21	141.0	142.7	1.7	<0.001	142.3	1.3	<0.001
M4620-T21-男性	1st	T21	149.7	151.3	1.6	<0.001	150.9	1.2	<0.001

500

图5

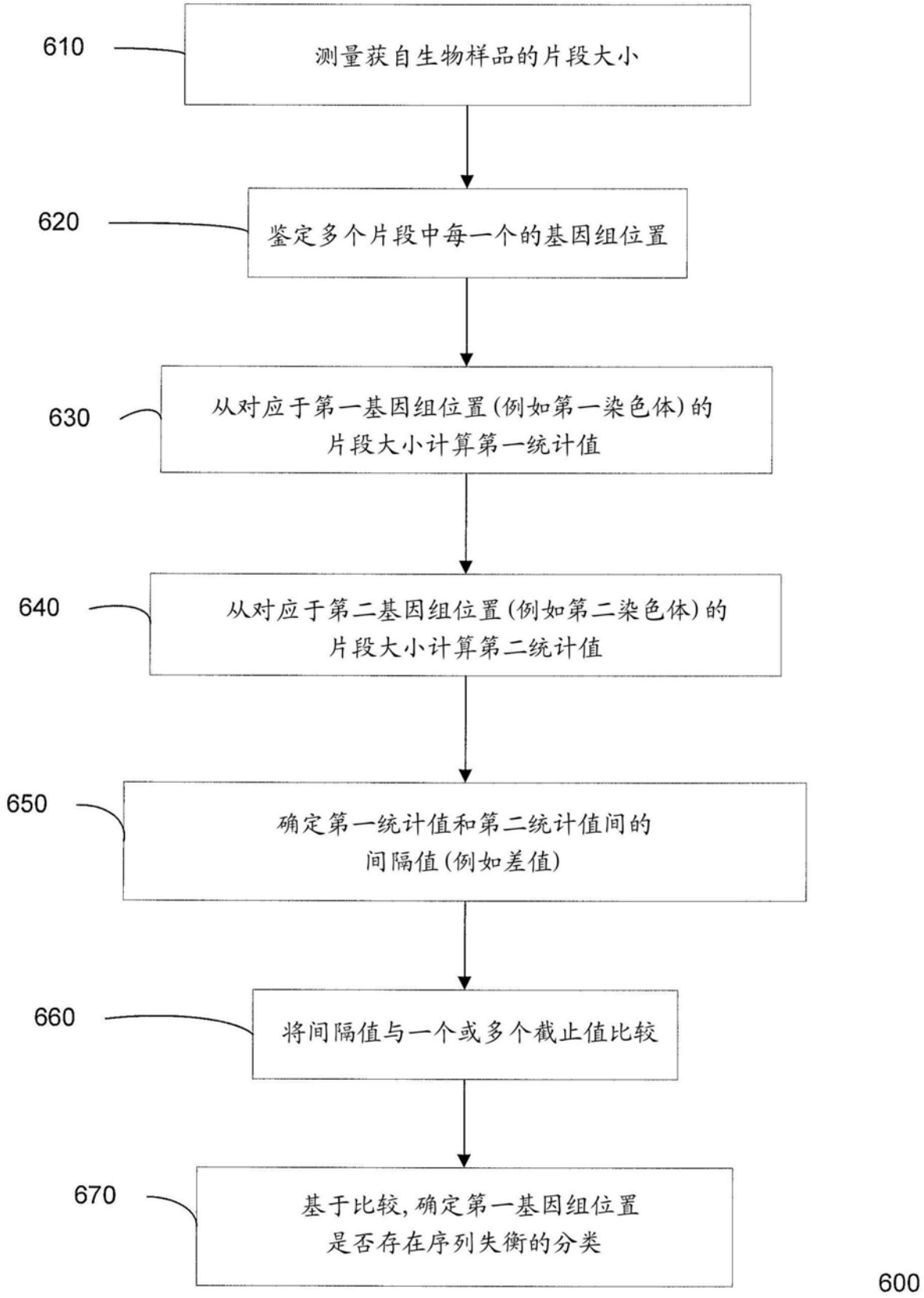

610
测量获自生物样品的片段大小
620
鉴定多个片段中每一个的基因组位置
630
从对应于第一基因组位置(例如第一染色体)的片段大小计算第一统计值
640
从对应于第二基因组位置(例如第二染色体)的片段大小计算第二统计值
650
确定第一统计值和第二统计值间的间隔值(例如差值)
660
将间隔值与一个或多个截止值比较
670
基于比较,确定第一基因组位置是否存在序列失衡的分类
600

图6

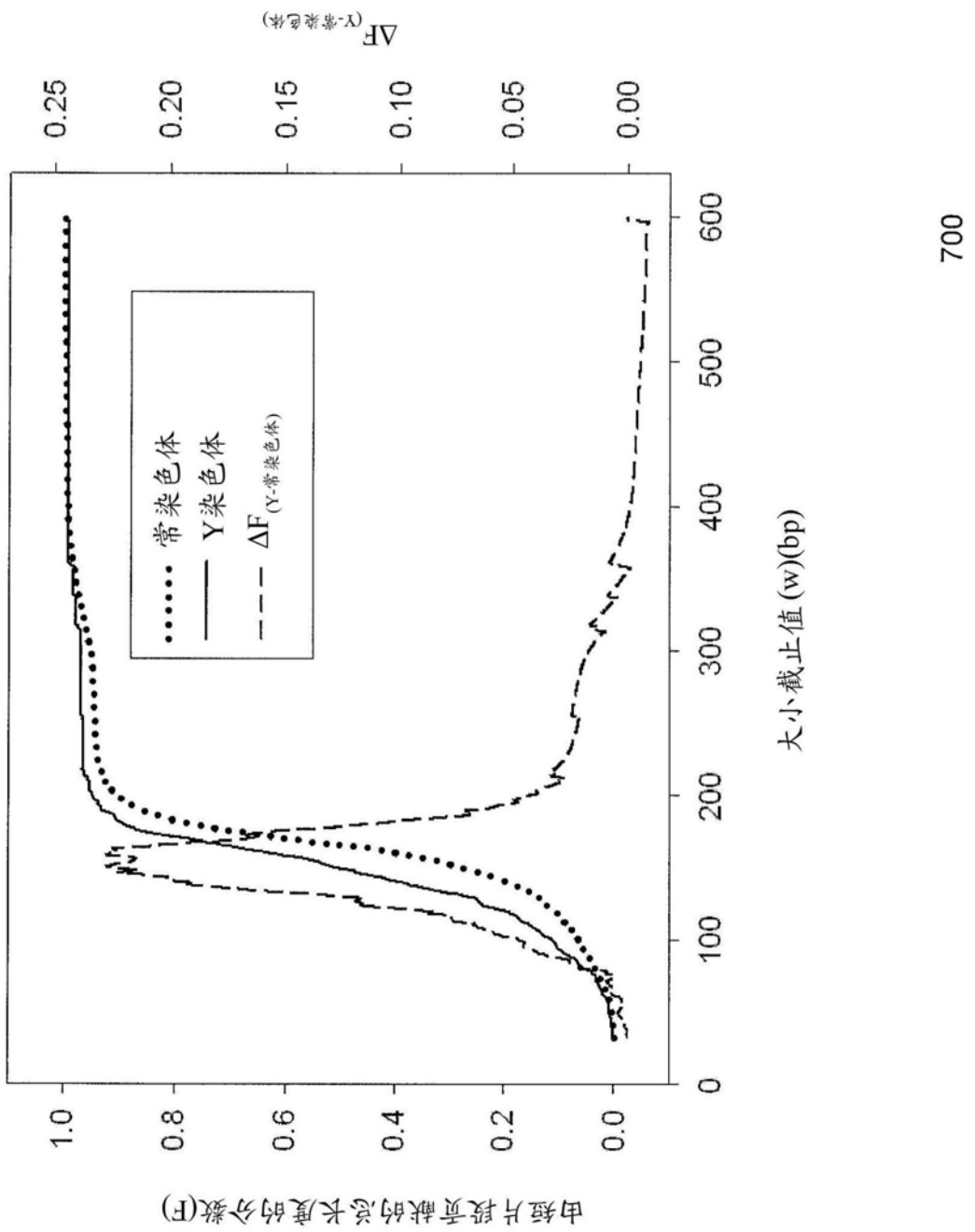

由短片段贡献的总长度的分数(F)
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700
大小截止值(w)(bp)
常染色体
Y染色体
ΔF(Y-常染色体)
ΔF(Y-常染色体)
0.25
0.20
0.15
0.10
0.05
0.00

图7

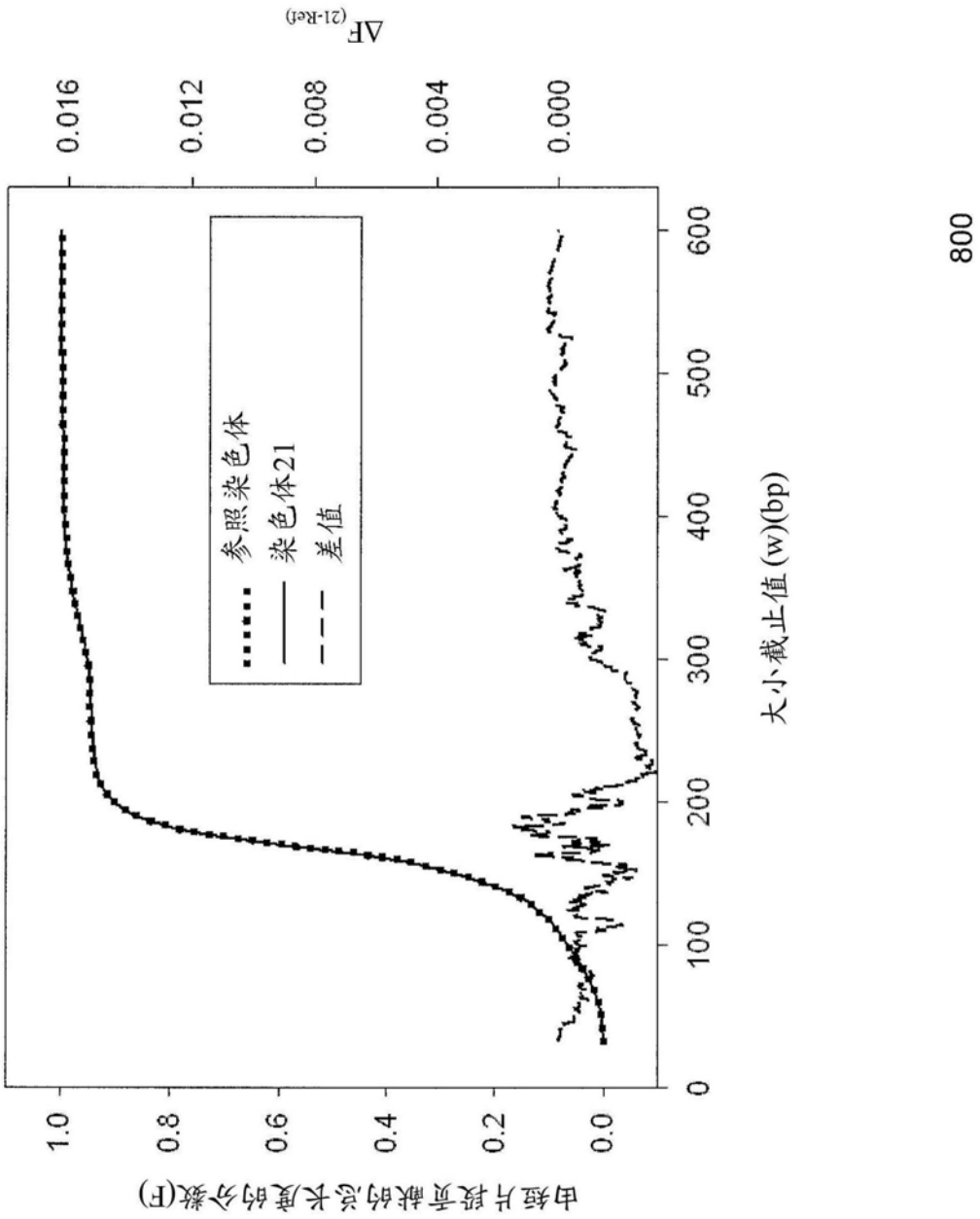

ΔF(21-Ref)
0.016
0.012
0.008
0.004
0.000
参照染色体
染色体21
差值
由短片段贡献的总长度的分数(F)
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
800
大小截止值(w)(bp)

图8

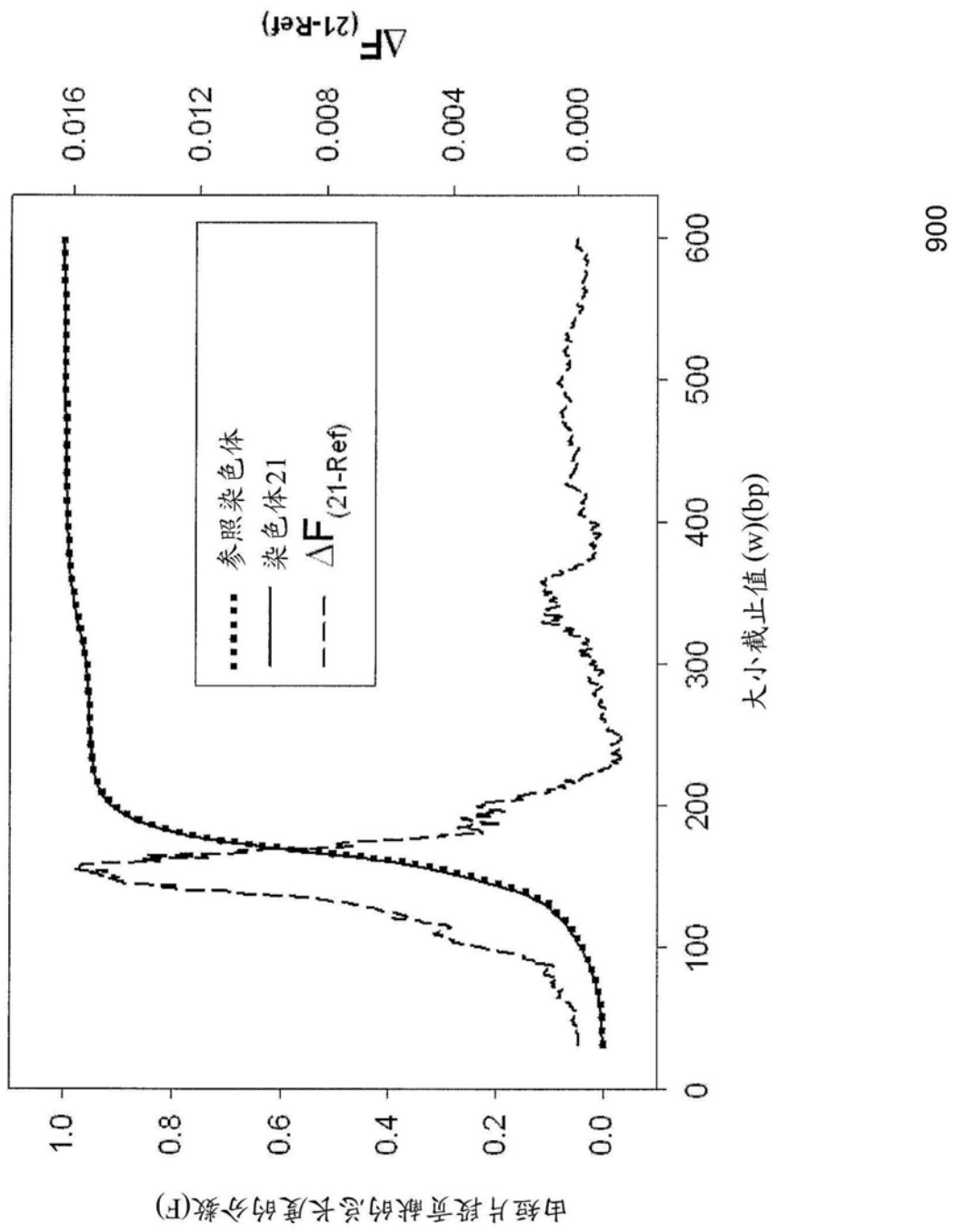

ΔF(21-Ref)
0.016
0.012
0.008
0.004
0.000
参照染色体
染色体21
ΔF(21-Ref)
由短片段贡献的总长度的分数(F)
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
900
大小截止值(w)(bp)

图9

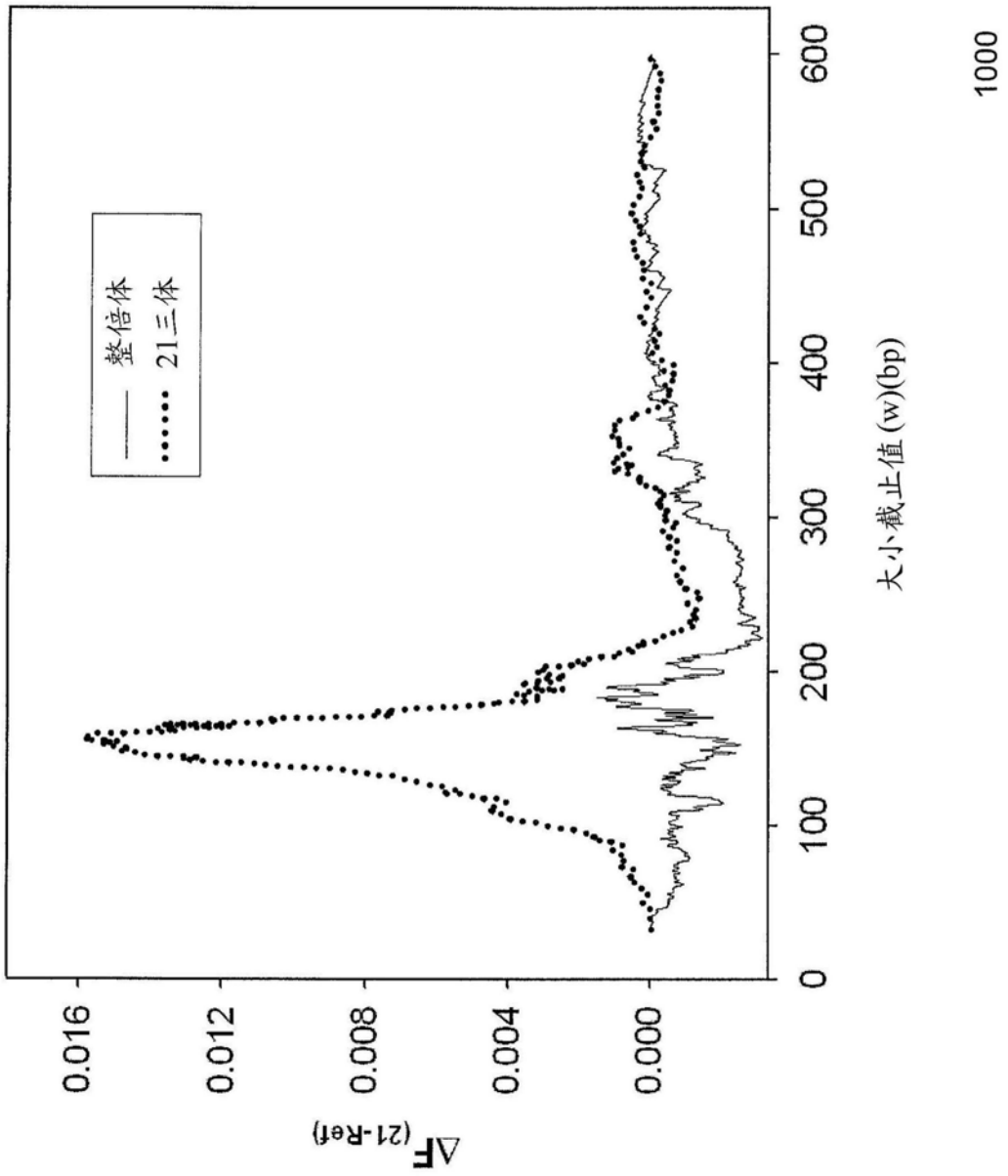

整倍体
21三体
ΔF(21-Ref)
0.016
0.012
0.008
0.004
0.000
0
100
200
300
400
500
600
1000
大小截止值(w)(bp)

图10

	病例数		
	女性	男性	总计
13三体 (T13)	9	14	23
18三体 (T18)	20	10	30
21三体 (T21)	4	5	9
整倍体	23	35	58
总计	56	64	120

1100

图11

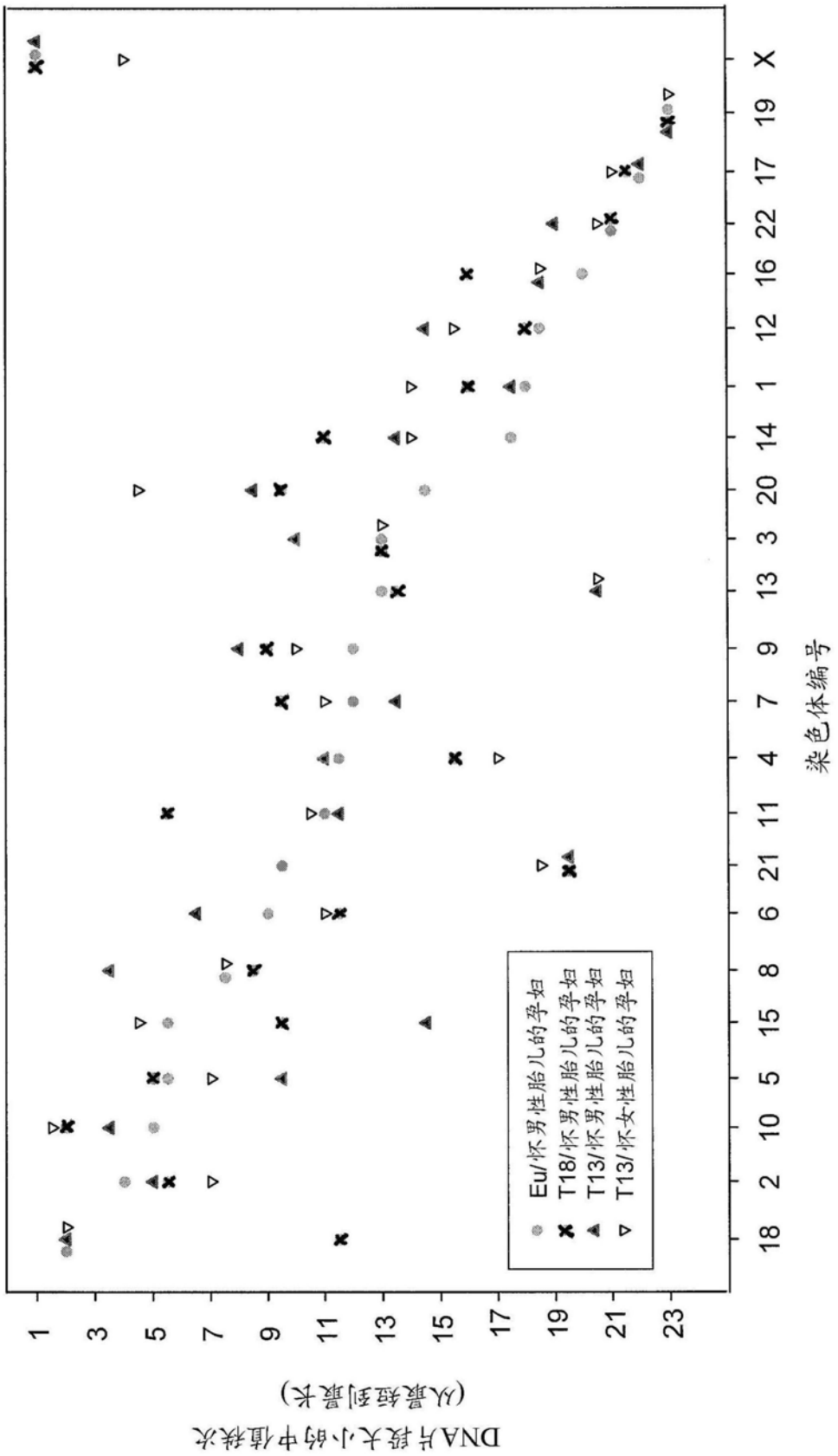

DNA片段大小的中值秩次
(从最短到最长)
染色体编号
Eu/怀男性胎儿的孕妇
T18/怀男性胎儿的孕妇
T13/怀男性胎儿的孕妇
T13/怀女性胎儿的孕妇

图12

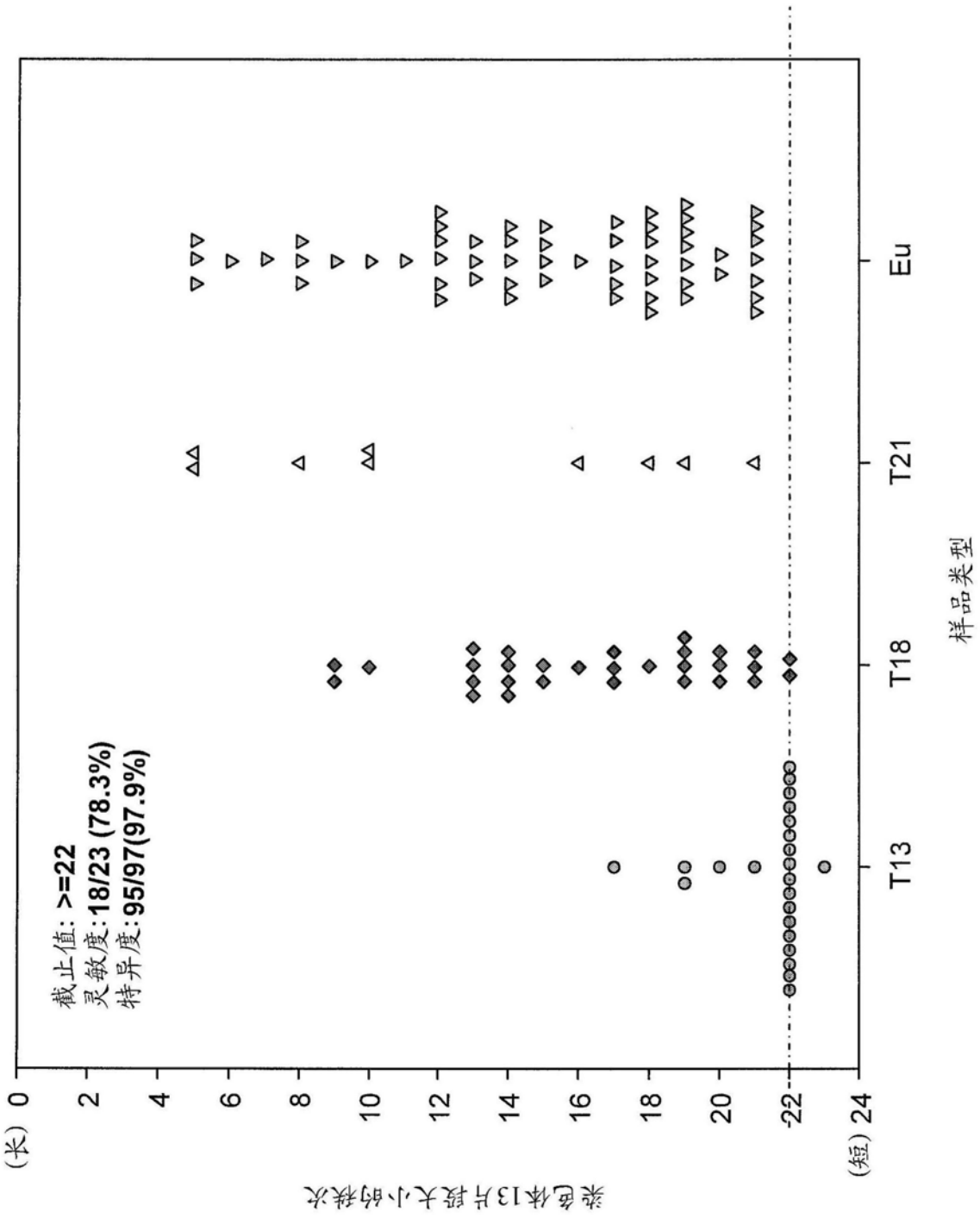

截止值: >=22
灵敏度:18/23 (78.3%)
特异度:95/97(97.9%)
(长) 0
2
4
6
8
10
12
14
16
18
20
22
(短) 24
染色体13片段大小的秩次
T13
T18
T21
Eu
样品类型

图13

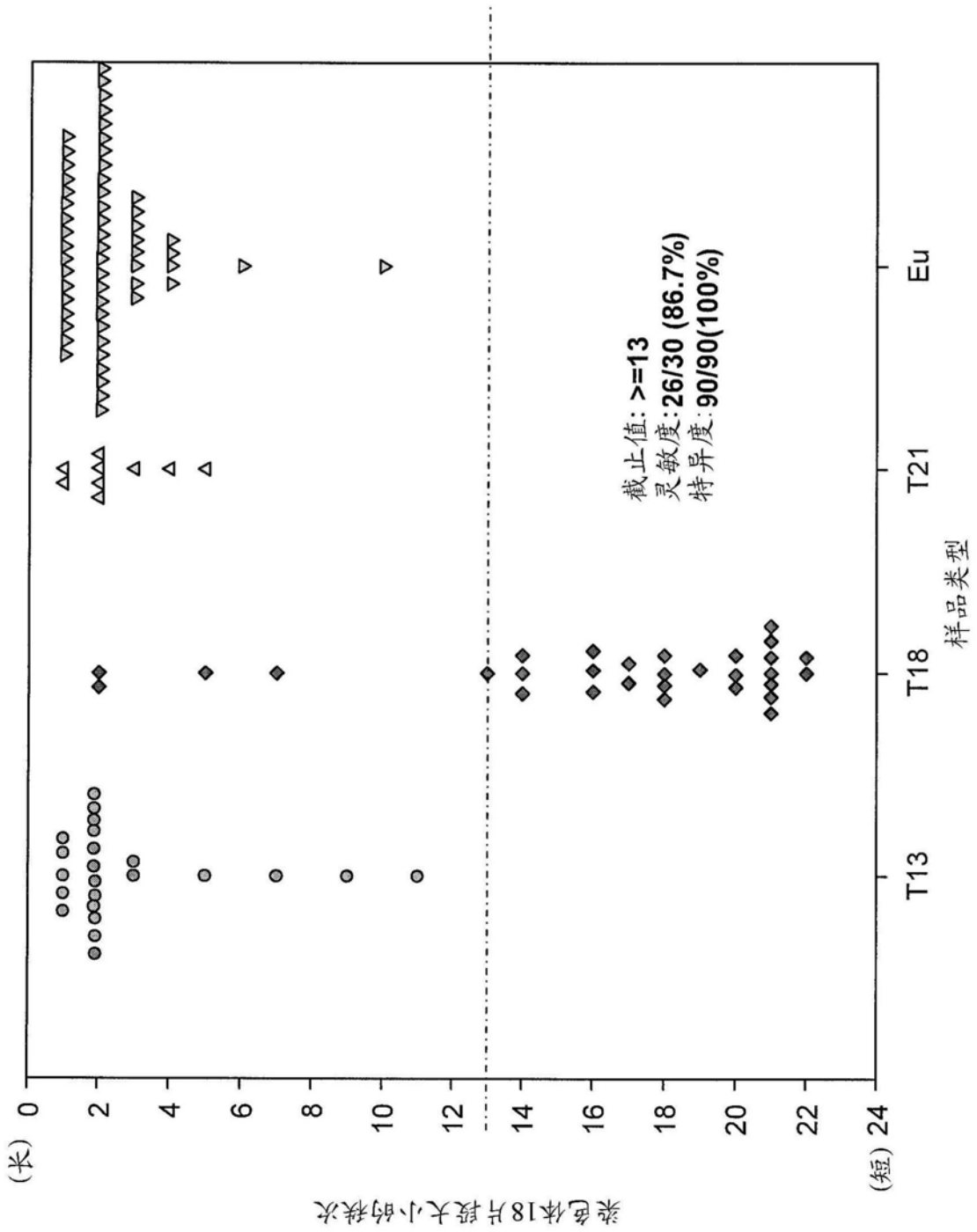

染色体18片段大小的秩次
(长)
(短)
0
2
4
6
8
10
12
14
16
18
20
22
24
T13
T18
T21
Eu
样品类型
截止值: >=13
灵敏度:26/30 (86.7%)
特异度:90/90(100%)

图14

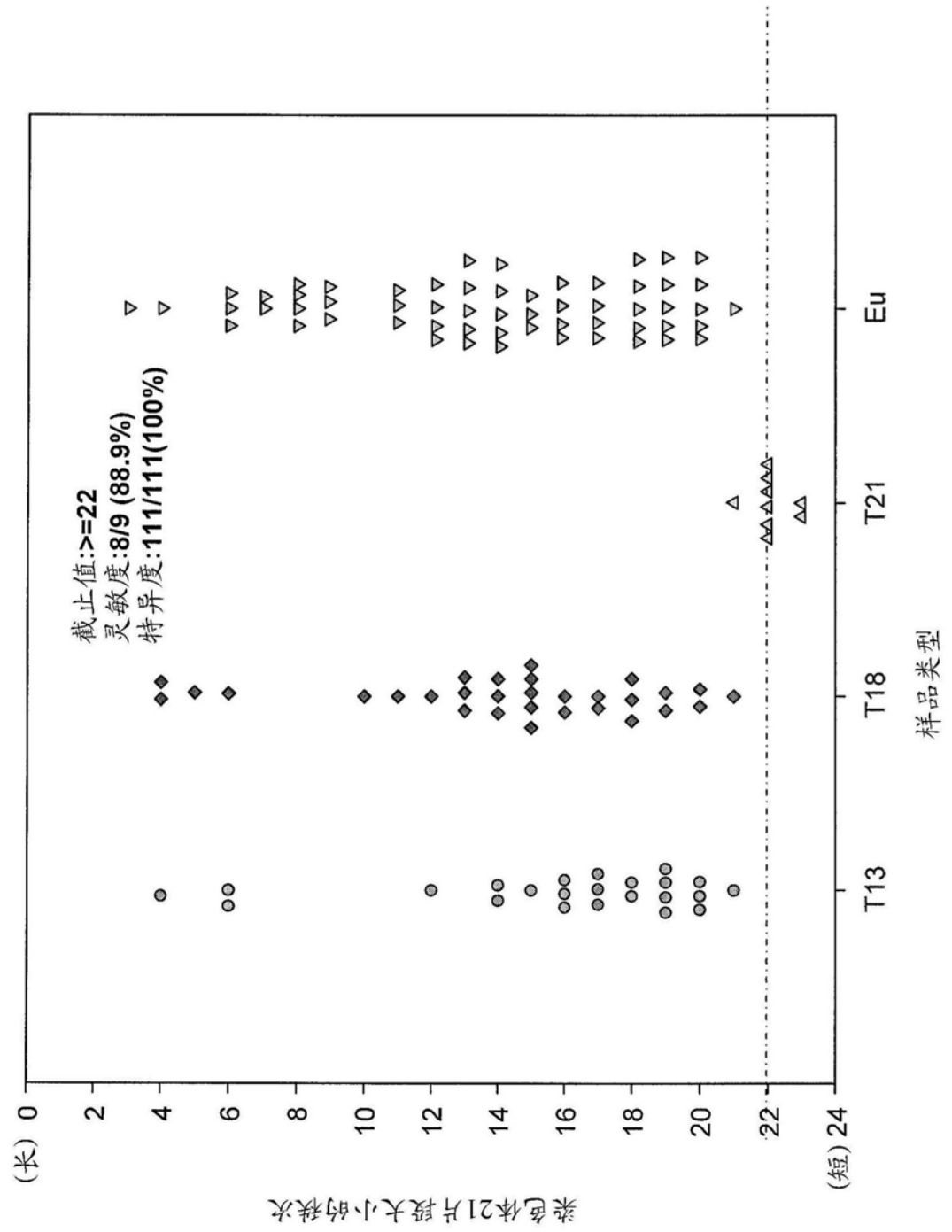

截止值:>=22
灵敏度:8/9 (88.9%)
特异度:111/111(100%)
染色体21片段大小的秩次
(长) 0
2
4
6
8
10
12
14
16
18
20
22
(短) 24
T13
T18
T21
Eu
样品类型

图15

样品名称	核型	染色体13片段的平均大小	染色体5片段的平均大小	染色体5和染色体13均值间的差值	P值*(染色体5对染色体13)	染色体6片段的平均大小	染色体6和染色体13的均值间的差值	P值*(染色体6对染色体13)
87164 - Eu - 男性	整倍体	154.969	155.198	0.229	0.159	155.037	0.068	0.732
87213 - Eu - 男性	整倍体	154.965	155.294	0.329	0.09	155.339	0.374	0.256
94355 - T18 - 男性	18三体	163.908	164.11	0.202	0.035	164.143	0.235	0.276
96233 - T18 - 男性	18三体	163.751	164.074	0.323	0.356	163.623	-0.128	0.371
92394 - T13 - 男性	13三体	165.15	165.754	0.604	<0.001	165.901	0.751	<0.001
61175 - T13 - 男性	13三体	156.165	156.581	0.416	<0.001	156.696	0.531	<0.001
96342 - T13 - 女性	13三体	157.064	158.127	1.063	<0.001	158.301	1.237	<0.001
PW093 - T13 - 女性	13三体	159.038	159.846	0.808	<0.001	159.679	0.641	0.001

1600

图16

样品名称	核型	染色体18片段的平均大小	染色体14片段的平均大小	染色体18和染色体14均值间的差值	P值(染色体18对染色体14)
87164- Eu-男性	整倍体	155.595	154.614	0.981	<0.001
87213- Eu-男性	整倍体	155.977	155.142	0.835	<0.001
94355- T18-男性	18三体	163.637	163.914	-0.277	0.541
96233- T18-男性	18三体	163.897	164.378	-0.481	0.393
92394- T13-男性	13三体	166.163	165.715	0.448	<0.001
61175- T13-男性	13三体	157.186	156.606	0.58	<0.001
96342- T13-女性	13三体	158.326	158.013	0.313	0.003
PW093- T13-女性	13三体	160.469	159.503	0.966	<0.001

1700

图17

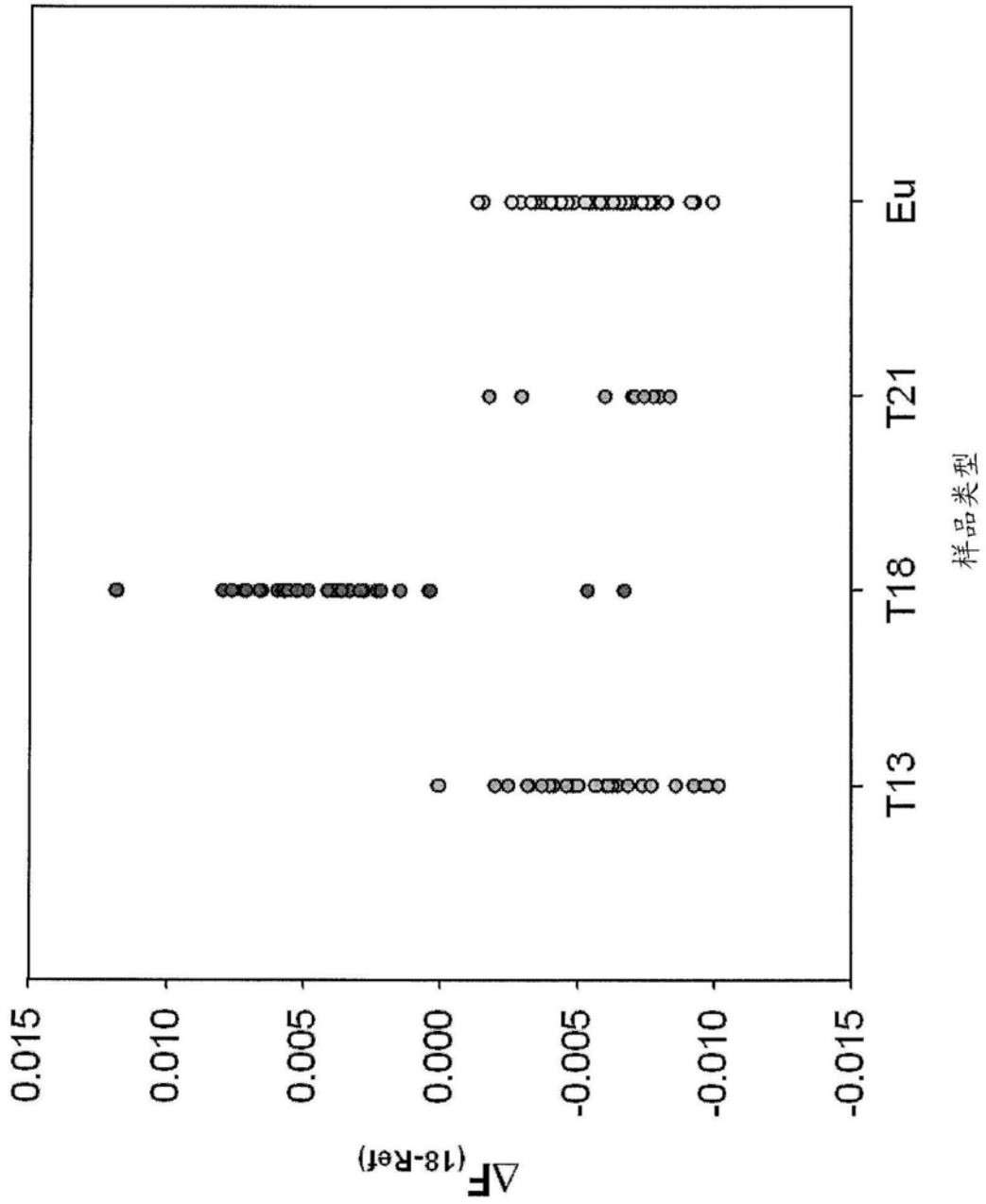

0.015
0.010
0.005
0.000
-0.005
-0.010
-0.015
ΔF(18-Ref)
T13
T18
T21
Eu
样品类型

图18

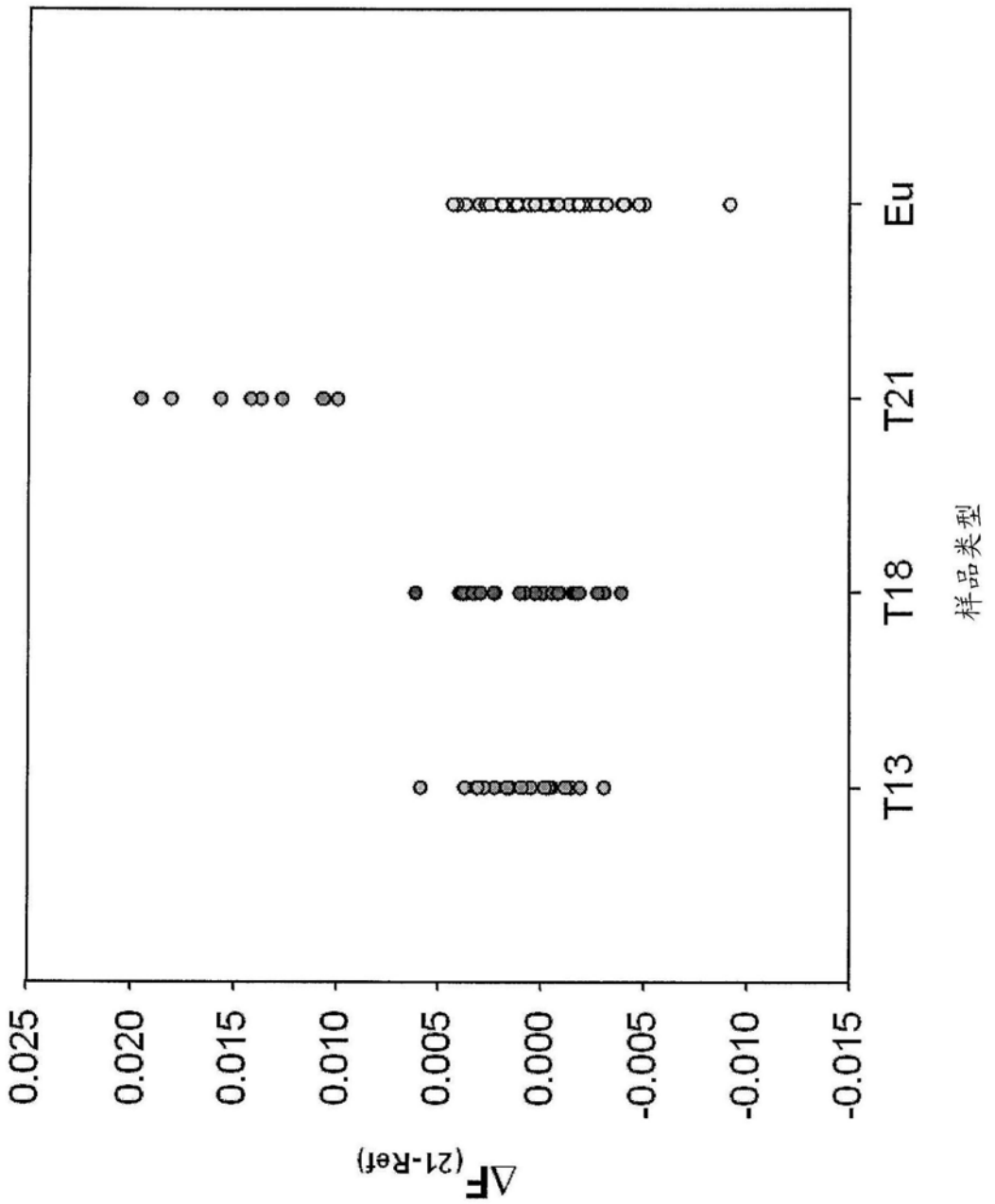

0.025
0.020
0.015
0.010
0.005
0.000
-0.005
-0.010
-0.015
ΔF (21-Ref)
T13
T18
T21
Eu
样品类型

图19

Chr.	GC含量(%)
4	36.53
13	37.10
6	38.60
5	38.61
3	38.79
X	38.96
18	39.07
8	39.45
2	39.67
12	39.80
7	39.94
21	40.00
14	40.33
9	41.03
10	41.41
1	41.52
11	41.89
15	42.35
20	44.65
16	45.22
17	46.05
22	49.89
19	50.65

图20

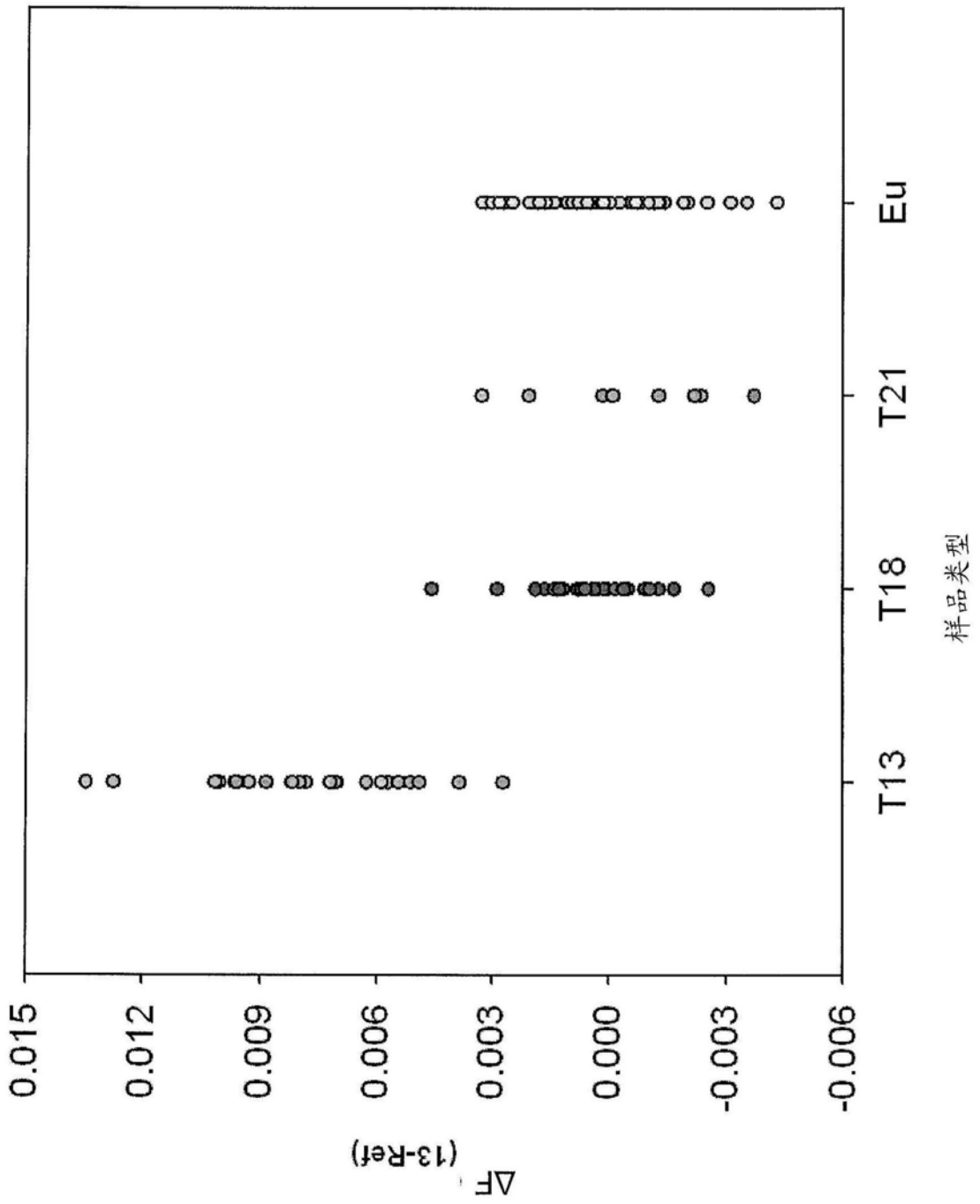

0.015
0.012
0.009
0.006
0.003
0.000
-0.003
-0.006
ΔF (13-Ref)
T13
T18
T21
Eu
样品类型

图21

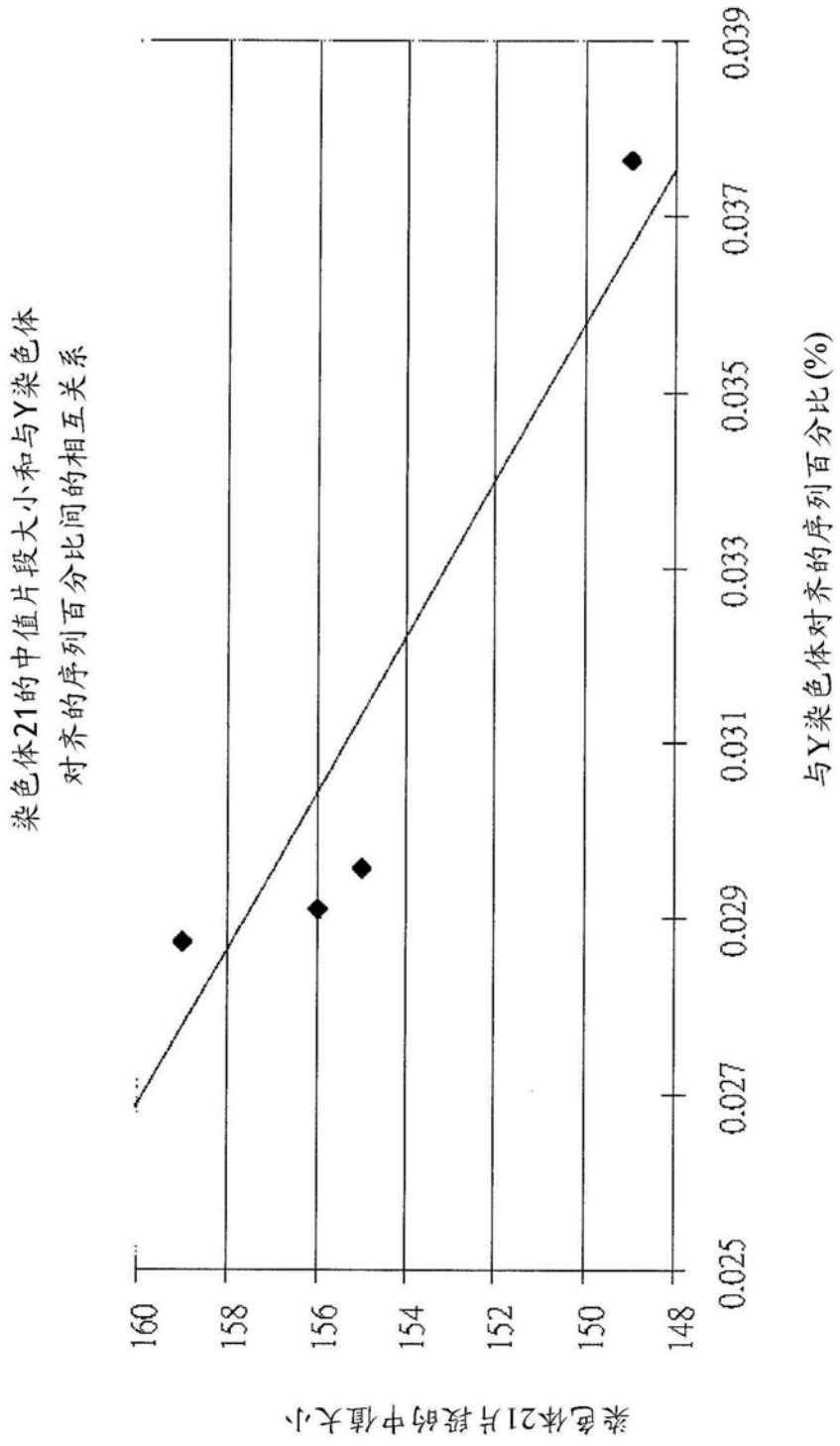

染色体21的中值片段大小和与Y染色体
对齐的序列百分比间的相互关系
染色体21片段的中值大小
160
158
156
154
152
150
148
0.025
0.027
0.029
0.031
0.033
0.035
0.037
0.039
与Y染色体对齐的序列百分比(%)

图22

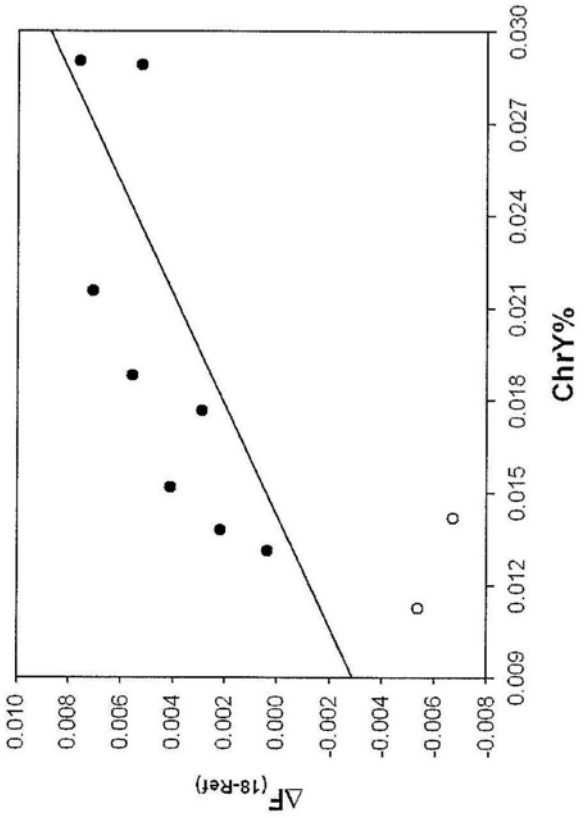

ΔF(18-Ref)
0.010
0.008
0.006
0.004
0.002
0.000
-0.002
-0.004
-0.006
-0.008
0.009
0.012
0.015
0.018
0.021
0.024
0.027
0.030
ChrY%

图23A

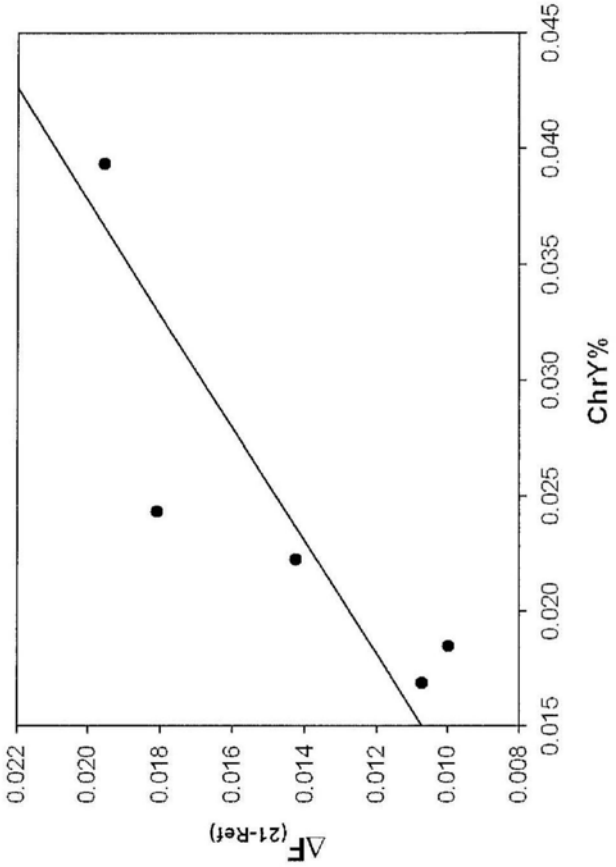

ChrY%
0.045
0.040
0.035
0.030
0.025
0.020
0.015
ΔF (21-Ref)
0.022
0.020
0.018
0.016
0.014
0.012
0.010
0.008

图23B

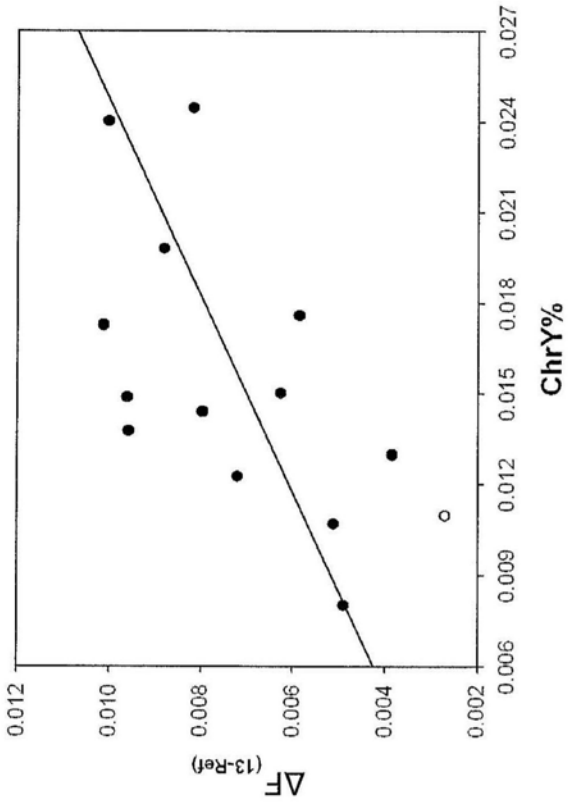


图23C

样品ID	核型	染色体18的Z-分值	染色体13的Z-分值	检测的染色体18 DNA片段的显著缩短	检测的染色体13 DNA片段的显著缩短
1	整倍体	1.07	0.84	无	无
2	整倍体	-0.64	0.08	无	无
3	18 三体	3.62	-1.43	有	无
4	18 三体	6.07	0.51	有	无
5	13 三体	0.61	3.55	无	有
6	13 三体	-1.04	2.06	无	有
7	13 三体	2.67	5.72	无	有
8	13 三体	-1.39	1.55	无	有

图24

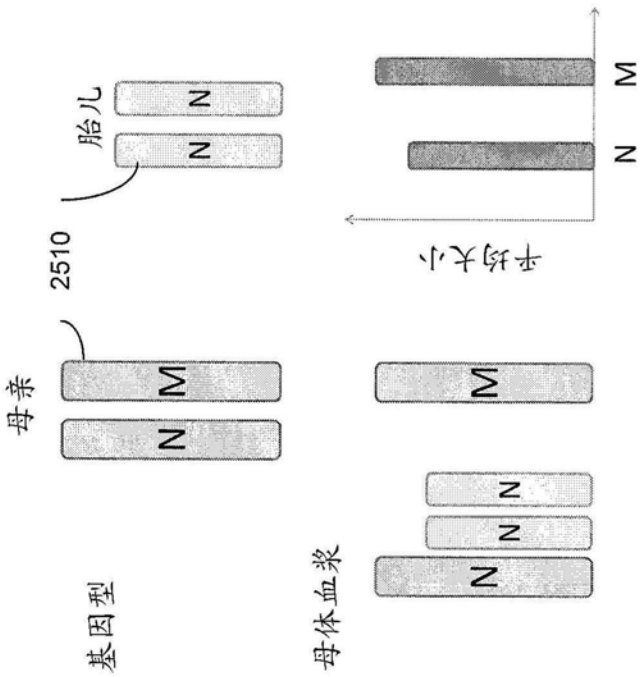

胎儿
N
N
2510
母亲
M
N
基因型
母体血浆
N
N
N
M
N
平均大小
M
N

图25A

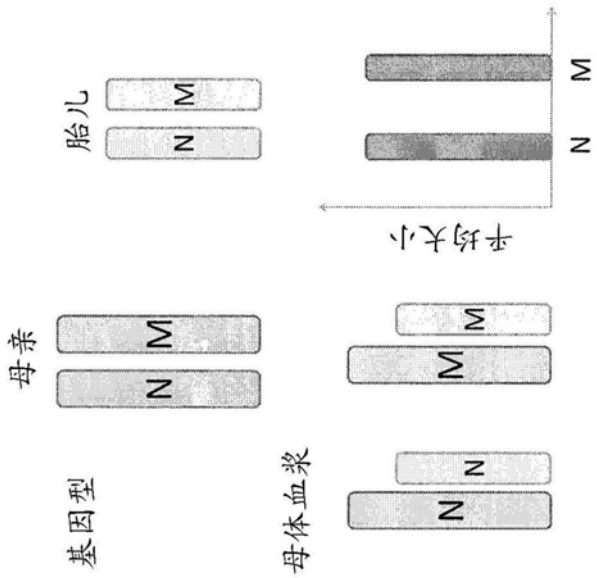

胎儿
M
N
母亲
M
N
基因型
母体血浆
M
M
N
N
平均大小
M
N

图25B

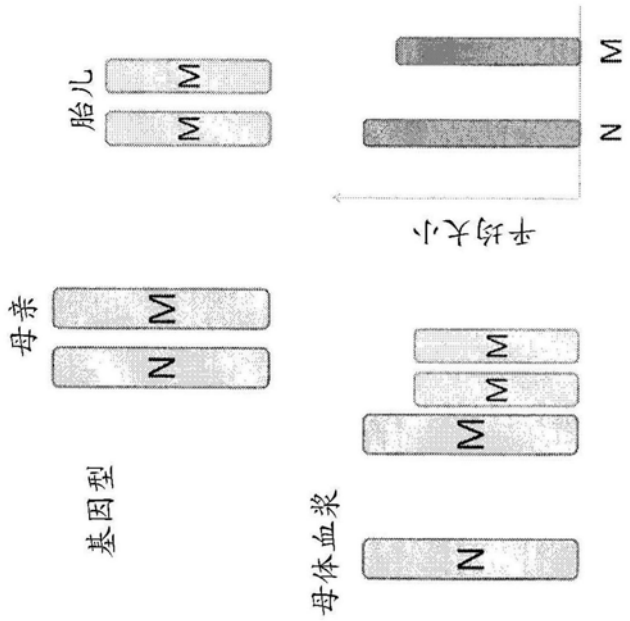


图25C

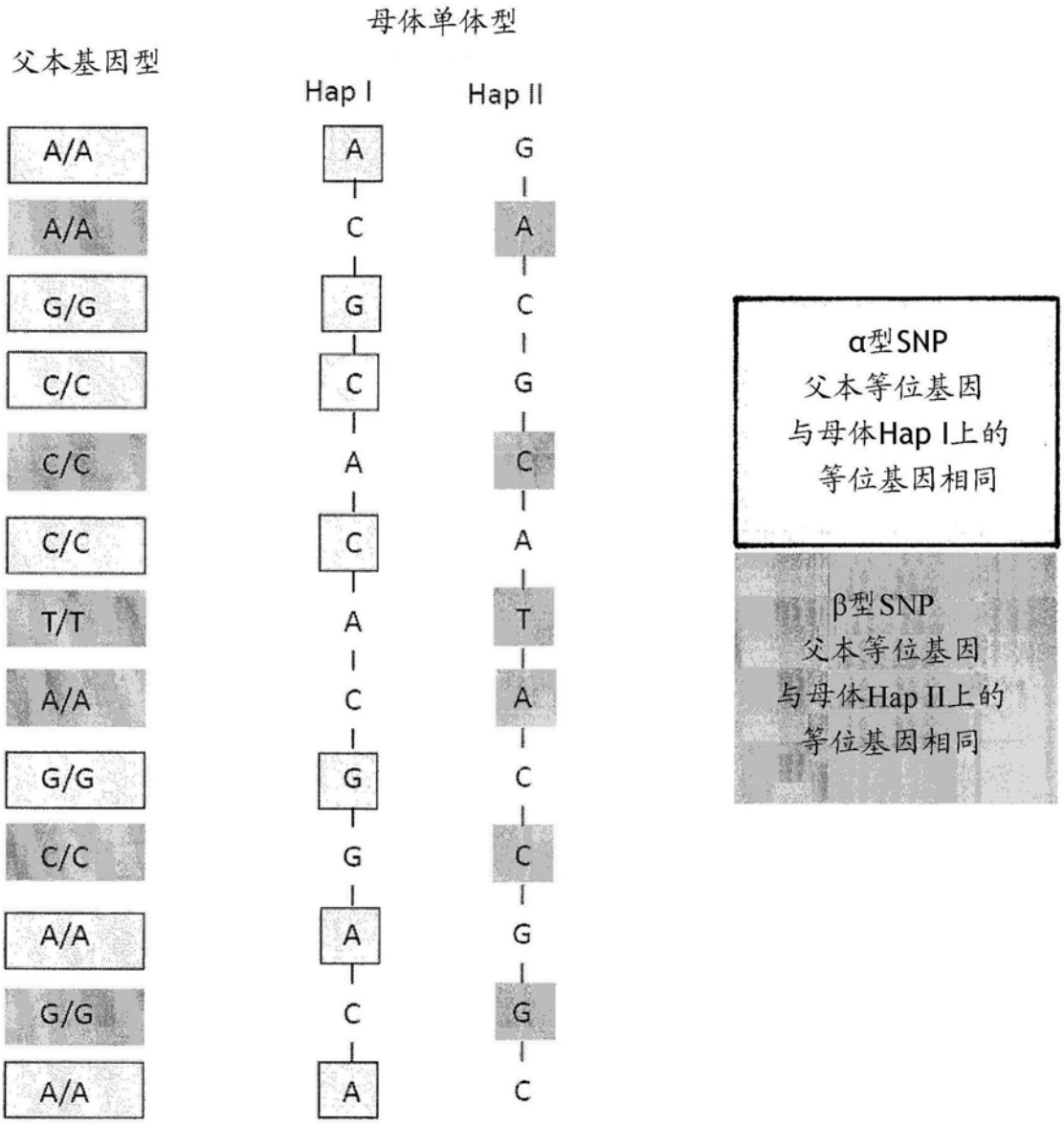

母体单体型
父本基因型
Hap I
Hap II
A/A
A/A
G/G
C/C
C/C
C/C
T/T
A/A
G/G
C/C
A/A
G/G
A/A
A
C
G
C
A
C
A
C
G
G
A
C
A
C
G
C
A
T
A
C
C
G
G
C
α型SNP
父本等位基因
与母体Hap I上的
等位基因相同
β型SNP
父本等位基因
与母体Hap II上的
等位基因相同

图26

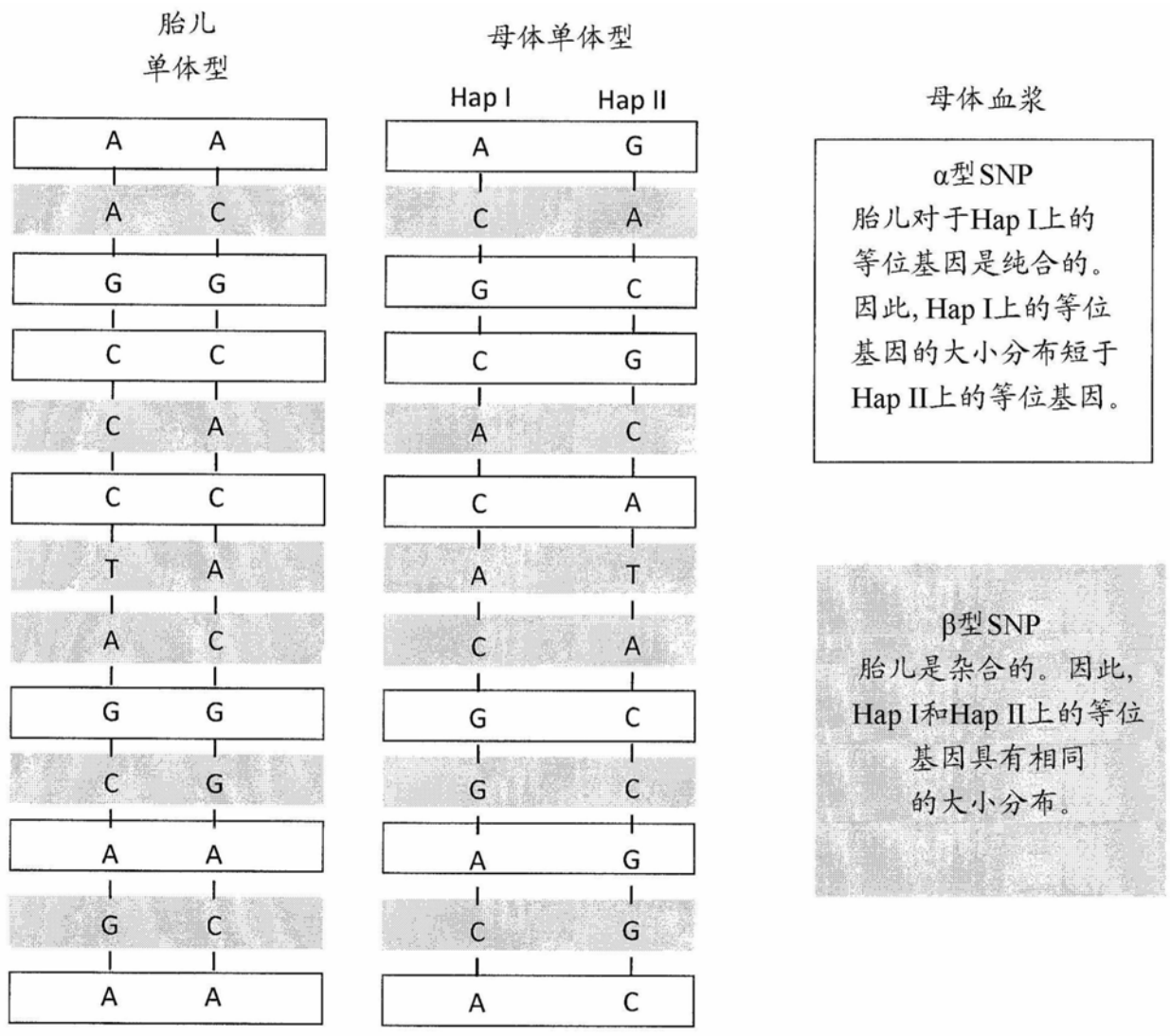

胎儿
单体型
母体单体型
Hap I
Hap II
母体血浆
A A
A C
G G
C C
C A
C C
T A
A C
G G
C G
A A
G C
A A
A G
C A
G C
C G
A C
C A
A T
C A
G C
G C
A G
C G
A C
α型SNP
胎儿对于Hap I上的
等位基因是纯合的。
因此, Hap I上的等位
基因的大小分布短于
Hap II上的等位基因。
β型SNP
胎儿是杂合的。因此,
Hap I和Hap II上的等位
基因具有相同
的大小分布。

图27

α型SNP					
节段中SNP的数目	起始SNP	终止SNP	由短片段贡献的总长度的分数		$\Delta F_{(Hap\ I\ -\ Hap\ II)}$
			Hap I	Hap II	
50	rs2027649	rs5993883	0.2595	0.2096	0.0499
50	rs2239395	rs1002286	0.2648	0.2219	0.0429
50	rs4822458	rs3747134	0.2905	0.2204	0.0701
50	rs5761557	rs4410	0.2659	0.2037	0.0623
50	rs134784	rs4822998	0.2569	0.2221	0.0348
50	rs5762936	rs5998473	0.2882	0.2302	0.0580
50	rs5754086	rs8140669	0.2599	0.2311	0.0288
50	rs5999854	rs5756540	0.2659	0.2249	0.0411
50	rs229535	rs2413637	0.2953	0.2425	0.0528
50	rs4820431	rs11090087	0.2793	0.2199	0.0594
50	rs1023469	rs1972489	0.2884	0.2438	0.0445
50	rs5764858	rs5769218	0.2644	0.2001	0.0643
50	rs6009043	rs5769452	0.2481	0.2118	0.0363
28	rs17825762	rs131815	0.2648	0.2407	0.0241

图28

β型SNP					
节段中SNP的数目	起始SNP	终止SNP	由短片段贡献的总长度的分数		$\Delta F_{(Hap\ I\ -\ Hap\ II)}$
			Hap I	Hap II	
50	rs2159071	rs873387	0.2565	0.2414	0.0151
50	rs11917	rs78908	0.2548	0.2577	-0.0029
50	rs589089	rs7288450	0.2495	0.2589	-0.0093
50	rs5759868	rs2301492	0.2431	0.2477	-0.0046
50	rs2301497	rs713911	0.2411	0.2204	0.0207
50	rs4820740	rs5752850	0.2181	0.2195	-0.0014
50	rs5752851	rs5752964	0.2344	0.2372	-0.0028
50	rs1894473	rs135472	0.2603	0.2510	0.0093
50	rs135475	rs5754558	0.2324	0.2457	-0.0133
50	rs4821148	rs11705488	0.2381	0.2380	0.0001
50	rs390647	rs362246	0.2270	0.2224	0.0046
50	rs362214	rs6000130	0.2404	0.2401	0.0003
50	rs739206	rs12628179	0.2341	0.2193	0.0148
50	rs17298479	rs2143921	0.2597	0.2643	-0.0047
50	rs5758913	rs926542	0.2655	0.2582	0.0073
50	rs7291629	rs133761	0.2156	0.2360	-0.0203
50	rs133755	rs6007919	0.2359	0.2332	0.0027
50	rs135570	rs136646	0.2496	0.2466	0.0030
50	rs80454	rs8136986	0.2354	0.2372	-0.0018
50	rs5768117	rs133662	0.2339	0.2330	0.0009
49	rs6007851	rs739365	0.2376	0.2372	0.0004

图29

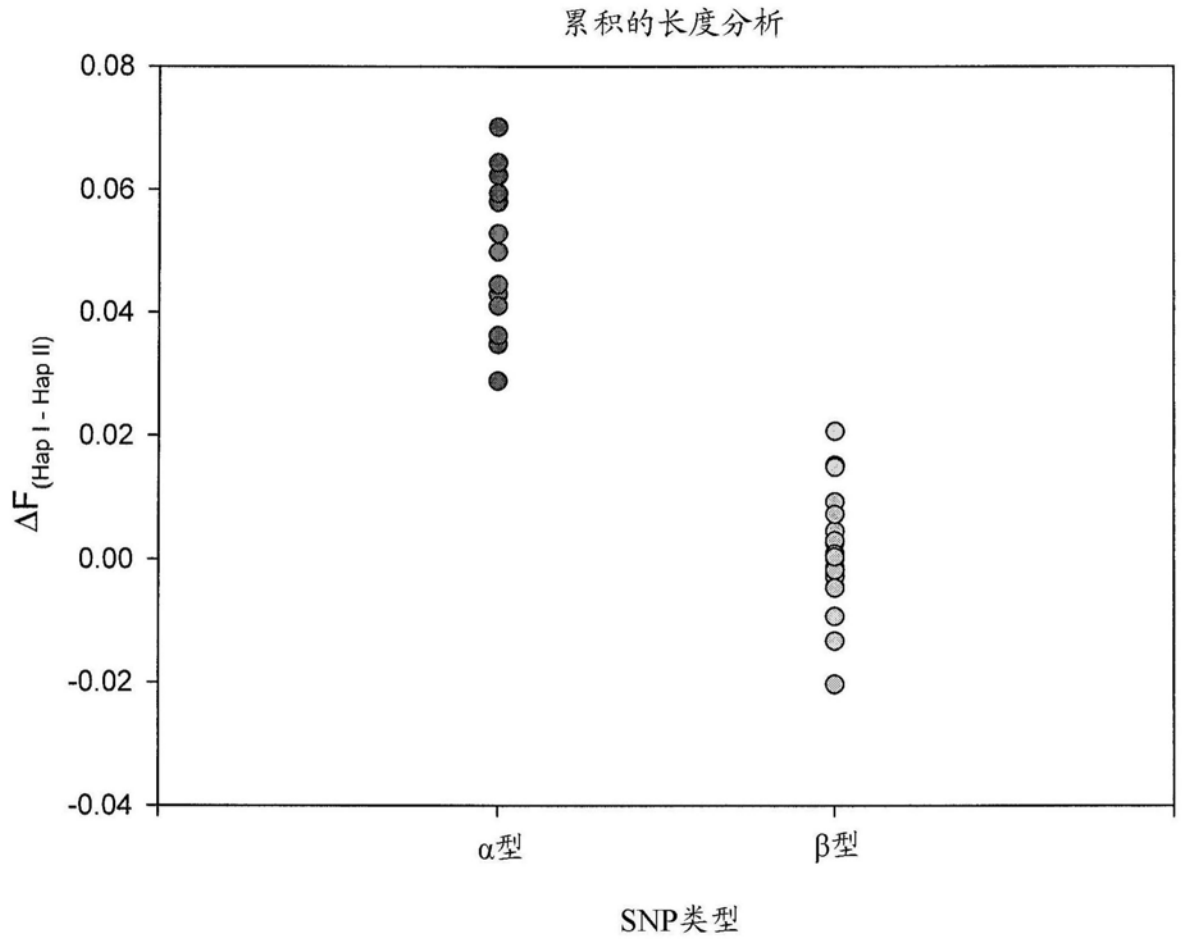


图30

样品代码	靶标富集	胎儿状态	由染色体21小于或等于150 bp的DNA片段贡献的总长度(10^6 bp)	由染色体21小于或等于600 bp的DNA片段贡献的总长度(10^6 bp)	F_{21}	由参照染色体小于或等于150 bp的DNA片段贡献的总长度(10^6 bp)	由参照染色体小于或等于600 bp的DNA片段贡献的总长度(10^6 bp)	F_{ref}	ΔF
uk 99229	无	T21	10.2	36.0	0.283	653.1	2422.2	0.270	0.013
pw 421	无	T21	10.6	35.6	0.299	651.2	2315.5	0.281	0.018
uk 99510	无	T21	9.8	32.0	0.307	607.1	2082.2	0.292	0.016
uk 99807	无	T21	12.6	30.1	0.417	745.2	1905.8	0.391	0.026
pw 226	无	整倍体	10.4	33.7	0.310	734.7	2368.4	0.310	0.000
pw 316	无	整倍体	10.9	38.4	0.284	765.2	2691.8	0.284	0.000
pw 263	无	整倍体	7.7	25.9	0.297	537.5	1810.8	0.297	0.001
pw 370	无	整倍体	5.9	23.9	0.247	411.8	1671.0	0.246	0.000

图31A

样品代码	靶标富集	胎儿状态	由染色体21小于或等于150 bp的DNA片段贡献的总长度(10^6 bp)	由染色体21小于或等于600 bp的DNA片段贡献的总长度(10^6 bp)	F_{21}	由参照染色体小于或等于150 bp的DNA片段贡献的总长度(10^6 bp)	由参照染色体小于或等于600 bp的DNA片段贡献的总长度(10^6 bp)	F_{ref}	ΔF
uk 99229	有	T21	5.7	24.9	0.231	434.1	1938.3	0.224	0.007
pw 421	有	T21	5.1	22.4	0.226	362.4	1685.0	0.215	0.011
uk 99510	有	T21	4.4	19.0	0.229	314.4	1435.1	0.219	0.010
uk 99807	有	T21	6.6	21.3	0.312	455.1	1553.2	0.293	0.019
pw 226	有	整倍体	5.1	21.4	0.237	413.2	1720.2	0.240	-0.004
pw 316	有	整倍体	4.9	22.6	0.215	394.9	1811.5	0.218	-0.003
pw 263	有	整倍体	4.1	19.3	0.211	327.0	1526.6	0.214	-0.004
pw 370	有	整倍体	3.7	20.4	0.180	296.9	1621.0	0.183	-0.003

图31B

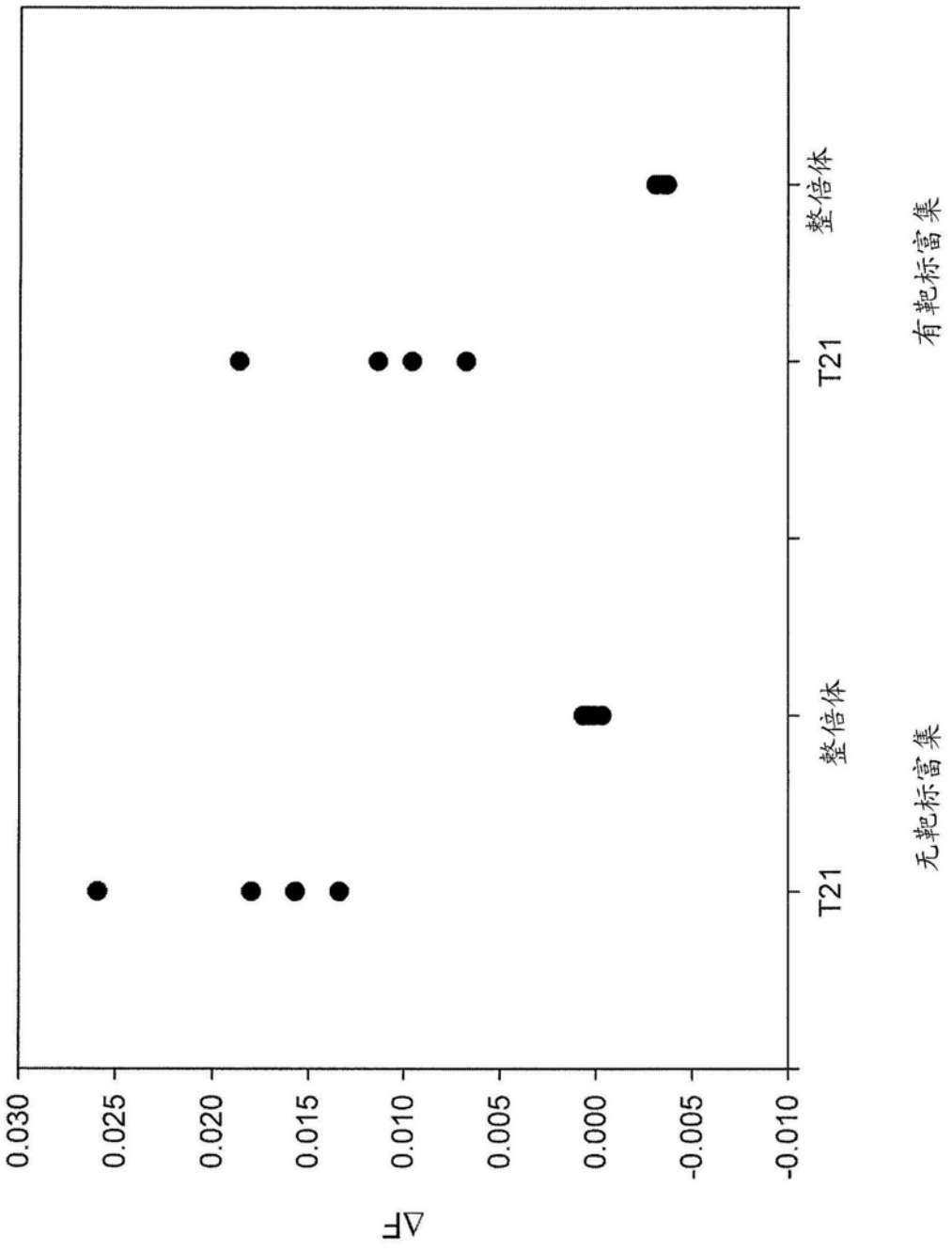

0.030
0.025
0.020
0.015
0.010
0.005
0.000
-0.005
-0.010
ΔΓ
T21
整倍体
T21
整倍体
无靶标富集
有靶标富集

图32

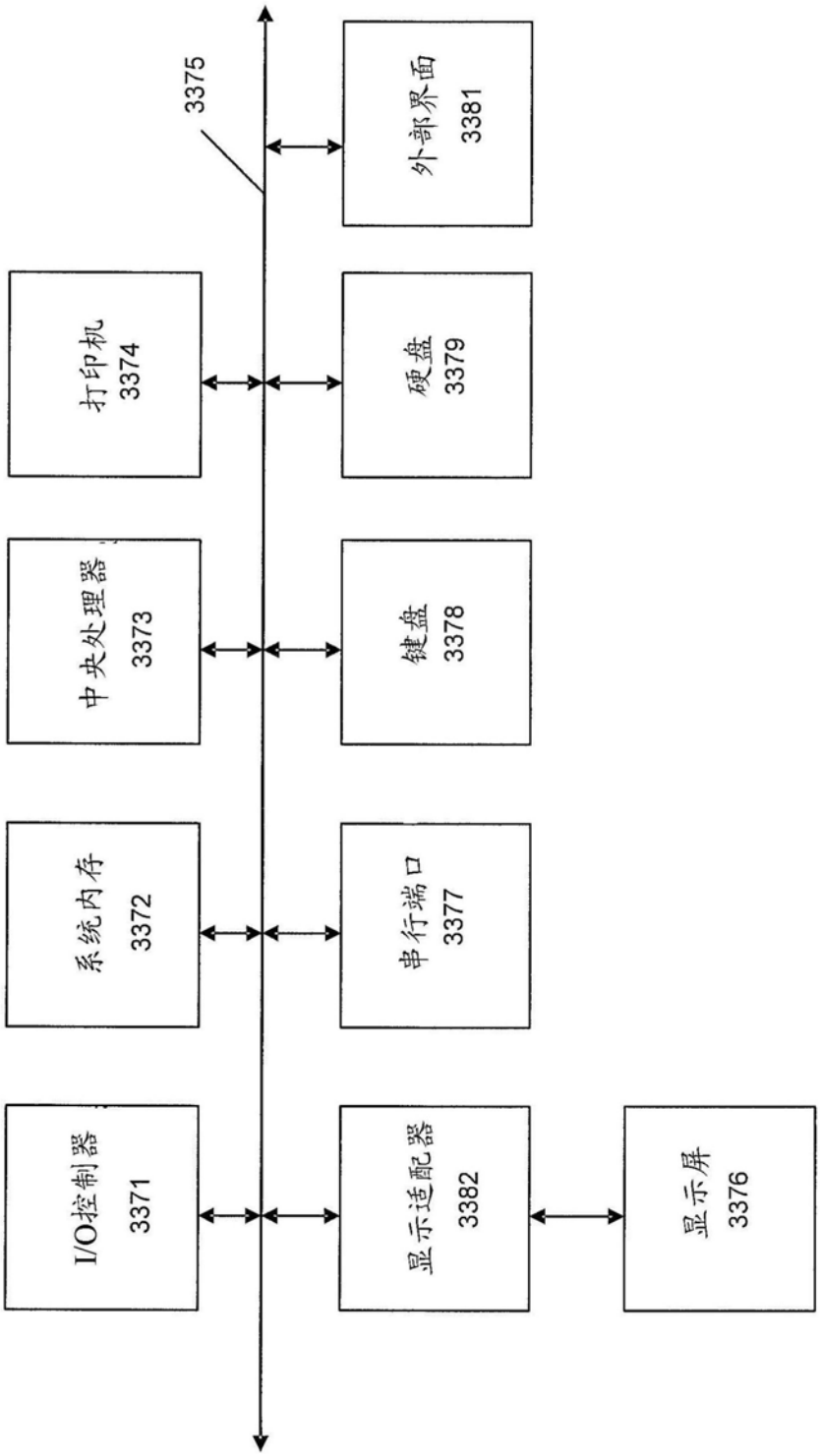

3375
外部界面
3381
打印机
3374
硬盘
3379
中央处理器
3373
键盘
3378
系统内存
3372
串行端口
3377
I/O控制器
3371
显示适配器
3382
显示屏
3376

图33