(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110475874 A
(43)申请公布日 2019.11.19

(21)申请号 201880018765.4

(22)申请日 2018.04.18

(30)优先权数据
17166836.1 2017.04.18 EP

(85)PCT国际申请进入国家阶段日
2019.09.17

(86)PCT国际申请的申请数据
PCT/EP2018/059889 2018.04.18

(87)PCT国际申请的公布数据
WO2018/192967 EN 2018.10.25

(71)申请人 安捷伦科技比利时有限公司
地址 比利时马赫伦

(72)发明人 贝努瓦·德沃热拉尔

(74)专利代理机构 上海光华专利事务所(普通合伙) 31219
代理人 余明伟 郭婧婧

(51)Int.Cl.
C12Q 1/6869(2006.01)
C12Q 1/6883(2006.01)

权利要求书1页 说明书14页

(54)发明名称

脱靶序列在DNA分析中的应用

(57)摘要

本发明涉及一种确定从怀孕女性中获得的生物样本中是否存在胎儿染色体非整倍体和/或杂合性缺失(LOH)的方法,所述方法包括:获取指示包括母体和胎儿核酸的生物样品的靶向捕获大规模平行测序的序列信息;确定从所述靶向捕获大规模平行测序获得的脱靶读取量;以及从所述脱靶读取计数中得到用于确定所述非整倍体或杂合性缺失存在与否的信息。

CN 110475874 A

1.一种确定从怀孕女性中获得的生物样本中是否存在胎儿染色体非整倍体和/或杂合性缺失(LOH)的方法,所述方法包括:

—获取指示包括母体和胎儿核酸的生物样品的靶向捕获大规模平行测序的序列信息;

—确定从所述靶向捕获大规模平行测序获得的脱靶读取量;以及

—从所述脱靶读取计数中得到用于确定所述非整倍体或杂合性缺失存在与否的信息。

2.一种确定从怀孕女性中获得的生物样本中是否存在胎儿非整倍体和/或杂合性缺失(LOH)的方法,所述样本包括母体和胎儿游离DNA,所述方法包括:

a)从所述生物样本中获得母体及胎儿DNA;

b)使所述DNA与一个或多个RNA或DNA标记探针接触,从而使所述探针与所述母体或胎儿DNA杂交;

c)捕获所述杂交DNA:探针;

d)对所述捕获的DNA进行测序,从而获得读段;

e)将所述内容映射到参考基因组;

f)分离在靶读取和脱靶读取;

g)获取脱靶读取计数;

并使用所述脱靶读取计数来确定是否存在胎儿非整倍体或LOH。

3.根据权利要求1或2所述的方法,进行深度测序。

4.根据上述权利要求所述的方法,其脱靶读取计数的最小数量是1×10^6。

5.根据上述任一权利要求所述的方法,其特征在于,所述探针指向预定靶标。

6.根据权利要求5所述的方法,其特征在于,所述探针指向所述DNA或区域中的重复区域。

7.根据权利要求5所述的方法,其特征在于,所述探针指向已知含有复发性CNV的一个或多个区域或位于所述复发性CNV侧翼的区域。

8.根据权利要求5所述的方法,其特征在于,所述探针指向序列长度在1×10^3到10×10^6碱基对之间的CNV靶标。

9.根据权利要求1到4中任一权利要求所述的方法,其特征在于,所述探针指向随机靶标。

10.根据上述任一权利要求所述的方法,其特征在于,所述在靶读取在进一步分析中被排除在外。

11.根据上述任一权利要求所述的方法,其特征在于,所获得的脱靶在参考集的基础上进行标准化。

12.根据权利要求1或2所述的方法,由在靶读取获取一个或多个参数,从而确定胎儿分数和/或检测是否存在微缺失。

13.一种检测从个体中获得的生物样本中是否存在杂合性缺失的方法,所述样本包括核酸,所述方法包括以下步骤:

—从所述样本DNA的靶向捕获大规模平行测序中获取序列信息;

—确定从所述靶向捕获大规模平行测序中获得的脱靶读取量;以及

—从所述脱靶读取计数中得出用于确定所述LOH存在与否的信息。

脱靶序列在DNA分析中的应用

技术领域

[0001] 本发明属于受试基因组分析技术领域。

背景技术

[0002] 胎儿非整倍体和其他染色体畸变影响了约千分之九的活产。历史上,诊断染色体异常的黄金标准是通过诸如绒毛膜取样和羊膜穿刺术等侵入性手术获得胎儿细胞的核型分析。

[0003] 母体循环中存在大量游离胎儿核酸的发现引发了新的无创产前基因检测的发展,这种检测可以检测染色体畸变。

[0004] 尽管近几年来临床遗传学领域取得了巨大进展,但仍然需要快速、经济和更准确的诊断方法。目前大多数可用的方法都是基于大量基因序列数据的生成,因此大部分信息是不必要的,或是在诊断前被过滤掉的。对于某些应用而言,只有有限数量的遗传材料是可用的,这表明需要一种比该领域的已知方法更准确和有效的分析方法。

[0005] 美国专利2015/066824 A1描述了一种方法,在该方法中,基因测序中产生的非必需信息与必需的基因测序数据相结合,以预测被提取样本的受试者是否存在多态性。然而,根据对孕妇的样本分析,这种方法不适合预测或监测胎儿的健康状况。

[0006] 此外,杂合性缺失(LOH)事关染色体,它会导致整个基因或等位基因的缺失,也可能导致周围染色体区域的一部分,如染色体臂或整个染色体的缺失。LOH在拷贝数减少或不减少的情况下都可发生,是许多人类癌症的一个重要特征,可以指示患者特定癌症的某些特征。因此,迫切需要更快、更灵敏、更准确的全基因组LOH筛查方法,以利用LOH信息治疗癌症患者。

[0007] Kuilman等人(2015)和Bellos等人(2014)都描述了使用基因测序过程中产生的非必要信息来检测受试者DNA拷贝数变化的方法。鉴于并非所有的杂合性缺失都会引起拷贝数的改变,这些方法不适用于对受试者的杂合性缺失进行全基因组的精确筛选。

[0008] 本发明的各种实施例中,利用了通常出于诊断目的而被视为非信息性的、无关的或被丢弃的数据。本发明所述的方法特别适用于产前诊断和肿瘤分析中的游离核酸分析,但也可用于非整倍体和遗传异常在疾病或综合征发展中起重要作用的其他领域。

发明内容

[0009] 本案提供了一个或多个受试者的生物样本的基因组或核酸序列分析的方法,利用可能位于靶向区或选定区之外的脱靶读取,所述靶向区或选定区是由例如使用大规模平行测序技术的靶向捕获方法生成的。本发明的方法允许使用在其他情况下可能被视为非信息性的或无关的基因信息的核酸测序信息。根据这些方法,这些序列信息可用于获得关于被获取序列和数据的样本状态的重要甚至关键的信息。这包括例如有关非整倍体和杂合性缺失(LOH)的信息。本发明的各种实施例中,通过将所述脱靶序列数据与在靶序列数据相结合,可以更有效地利用从样本中提取的核酸,从而减少总体的样本量和下游处理要求。这种

对现有样本处理和序列分析工作流程的改进在游离分析领域(包括胎儿染色体评估和循环肿瘤分析等应用)尤为重要。在这类应用中,通常只能获得少量或极少量的遗传物质,因此,本案的一个理想方面在于能更充分地利用样本序列数据来获得同时考虑在靶和脱靶序列信息的额外分析或诊断见解。

具体实施方式

[0010] 一、术语和缩略语:

[0011] 本发明提供的序列分析方法,可用于包括通过评估脱靶读取的相关序列数据,例如,通过靶向捕获大规模平行测序的方法进行样本分析时产生的序列数据,对受试者进行基因组分析的各种应用。这种脱靶序列读取通常被认为是非信息性的,从而被忽视或丢弃。当前技术和应用的发明者证明,通过利用脱靶序列读取数据,对检测染色体异常(如胎儿非整倍体)有用的见解和改进。脱靶读取也为其他序列分析应用,包括全基因组检测杂合性丢失(LOH),提供了一个有用的工具,这对目前技术来说可能非常困难,甚至是不可能的,特别是在浅层测序方案的背景下。

[0012] 除非另有定义,否则用于本发明的创新方面的所有术语,包括技术和科学术语,都具有本发明所属领域的普通技术人员通常所理解的含义。通过进一步的指导,术语定义也被包括其中,以便更好地理解本发明。

[0013] 这里使用的术语具有以下含义:

[0014] 除非上下文另有明确规定,“一个(a)”、“一(an)”和“这个(the)”包括复数形式。举例来说,“一个隔间(a compartment)”是指一个或多个隔间。

[0015] 本发明中所用的“大约”指可量化或可测量的值,例如参数、量、持续时间等,其含义是包括规定值的+/-20%或更小、优选+/-10%或更小、更优选+/-5%或更小、甚至更优选+/-1%或更小,以及最优选+/-0.1%或更小的变化。目前,所述变化适于在本发明中执行。然而,应当理解的是,修饰语“大约”所指的值本身也被具体披露。

[0016] 本发明中所用的“包含(comprise)”、“包含着(comprising)”、“包含了(comprises)”和“包含有(comprised of)”等同于“包括(include)”、“包括着(including)”、“包括了(includes)”或“含有(contain)”、“含有(containing)”、“含有(contains)”,为包含性或开放性术语,其规定了下述内容的存在,例如成分,并且不排除本领域已知或公开的附加、非叙述的成分、特征、元素、成员、步骤。

[0017] 按端点表述的数值范围包括纳入该范围内的所有数字和分数,以及陈述的端点。

[0018] 除非另有定义,否则在本发明的整个描述中,表达式“按重量计(%by weight)”、“重量百分比(weight percent)”、“%重量(%wt)”或“重量%(wt%)”是指基于配方的总重量的各个组分的相对重量。

[0019] 本发明所用术语“生物样本(biological sample)”指从个体(例如人类,比如一个孕妇,或其他生物有机体)获得的或与之相关的、且包含一个或多个所关注的核酸分子的任何样本。

[0020] 术语“大规模平行测序(massively parallel sequencing)”或“新一代测序(next-generation sequencing)”是指在生成的测序库的基础上,在高通量方法中用于核酸(包括DNA)测序的技术。

[0021] 术语“靶向捕获大规模平行测序(targeted-capture massively parallel sequencing)”是指待测序核酸样本可通过靶向捕获步骤富集的大规模平行测序技术,所述靶向捕获可以基于任何合适的方法进行,例如RNA或DNA探针。这种富集方法可用于减少待测序靶子或片段的总量、个数或复杂性,通过检查所选或所需的靶向基因(如染色体)区域来降低分析的总难度或成本。

[0022] 与靶向捕获技术有关的术语“组套(panel)”、“探针(probe)”或“诱饵(bait)”可包括根据特定的靶向捕获方案,用于靶向或选择所需核酸片段(例如具有特定序列、同源性或亲和力的片段或区域)的、或用于询问所选基因的分子、部分或区域。

[0023] 术语“脱靶读取(off-target reads)”应理解为通过大规模并行测序过程获得的读取,其中对所选序列的靶向捕获产生了一部分非特定序列片段,或导致了一些探针或诱饵与样本核酸的非特异性配对,由此超出了对组套、探针或诱饵的预期,例如,探针与DNA的不完全杂交。

[0024] 术语“在靶读取(on-target reads)”应理解为通过靶向捕获大规模并行测序过程获得的测序读取,并且是所用组套、探针或诱饵与样本核酸预期的或特定配对的结果,因此与捕获组套、探针或诱饵相一致。

[0025] 本发明中的术语“母体样本(maternal sample)”是指从至少一个怀孕个体(例如妇女)中获得的生物样本。

[0026] 本发明中的术语“个体(subject)”是指人类及非人类受试者,或生物有机体,例如哺乳动物、无脊椎动物、脊椎动物、真菌、酵母、细菌和病毒。尽管本发明中的示例涉及人类基因组,并且主要针对人类问题,但应当理解的是,本发明适用于任何生物有机体、植物或动物的基因组,并且可用于各种领域,包括但不限于兽医学、动物科学和研究实验室。

[0027] 术语“生物液体(biological fluid)”是指从生物来源中提取的液体,包括血液、血清、血浆、痰液、灌洗液、脑脊液、尿液、精液、汗液、眼泪、唾液、囊胚腔液等。它也指生物样本可以在其中生长的培养基,如可以培养细胞、组织或胚胎的体外培养基。如本发明所用,术语“血液(blood)”、“血浆(plasma)”和“血清(serum)”明确地包含其部分或加工部分。同样地,如果从活检、拭子、涂片等中提取样本,则术语“样本(sample)”明确地包括从活检、拭子、涂片等中提取的加工部分或部分。

[0028] 本发明中的术语“母体核酸(maternal nucleic acids)”和“胎儿核酸(fetal nucleic acids)”,分别是指孕妇受试者的核酸和孕妇腹中胎儿的核酸。如前所述,“胎儿核酸(fetal nucleic acids)”和“胎盘核酸(placental nucleic acids)”通常用于指同一类型的核酸,尽管这两种核酸之间可能存在生物学差异。

[0029] 本发明所用术语“胎儿分数(fetal fraction)”是指包含胎儿和母体核酸的样本中存在的胎儿核酸的分数表示或浓度。

[0030] 本发明中的术语“拷贝数变异(copy number variation)”或“CNV”是指与第二(合格)样本中的核酸序列的拷贝数相比,第一(测试)样本中核酸序列的拷贝数的变化,即几个碱基对(bp)或更大碱基对。“拷贝数变体(copy number variant)”是指通过将试验样本中所关注的序列与合格样本中存在的序列进行比较,而在其中发现有拷贝数差异的少数碱基对或更大核酸序列,。非限制性拷贝数变体/变异包括删除及微缺失、插入及微插入、复制和乘法。CNV可能包括染色体非整倍体和部分非整倍体。

[0031] 本发明中的术语“非整倍体(aneuploidy)”是指由于整条染色体或染色体的一部分丢失或获得而引起的遗传物质失衡。非整倍体是指染色体和亚染色体均失衡,例如但不限于缺失、微缺失、插入、微插入、拷贝数变异、重复。拷贝数变异大小可能在几个碱基对到几百万个碱基对的范围内,或者在特定情况下从一千个碱基对到几百万个碱基对的范围内。跨越数千万碱基对区域和/或相当于染色体臂重要部分的大型亚染色体异常也可称为节段性非整倍体。

[0032] 本发明中的术语“染色体非整倍体(chromosomal aneuploidy)”是指由整条染色体的丢失或获得引起的遗传物质失衡,包括种系非整倍体和镶嵌非整倍体。

[0033] 术语“杂合性缺失或LOH(loss of heterozygosity or LOH)”是指导致实质上整个基因或等位基因缺失的染色体事件,也可能会使周围染色体区域、染色体臂或整个染色体的一部分缺失。

[0034] 术语“读段(read)”是指通过实验获得的DNA序列,其组成和长度(例如约20个碱基对或更多)可用于识别更大的序列或区域,例如可比对并特定分配给染色体位置、基因组区域或基因的序列部分或片段。术语“读段(read)”、“序列读段(sequence read)”和“序列(sequences)”可在整个说明书中互换使用。

[0035] 术语“读取计数(read count)”是指与样本相关的读取次数,该样本可被映射到参考序列,例如基因组参考或所述参考基因组的一部分(读取计数可根据其相对于参考的映射位置进行组合或分组)。

[0036] 本发明所用术语“参考基因组(reference genome)”或“参考序列(reference sequence)”是指与样本(例如包含在电子核酸序列数据库中的样本)不同的预定信息或序列信息。参考基因组或序列可以是代表与选定生物有机体或物种核酸相关的核酸序列的至少一部分的序列信息的集合。参考基因组或序列可以从多个样本的核酸序列中组装而成,因此,参考基因组或序列不一定代表单一生物有机体的确切组成。在本发明的各种实施例中,所述参考可用于使测序读取从一个或多个样本映射到特定或靶向染色体或遗传序列的位置。

[0037] 本发明中的术语“测试样本(test sample)”是指包含多个核酸或含有至少一个核酸序列的核酸混合物的样本,所述核酸序列的拷贝数被怀疑经历了变异,或至少有一个核酸序列被期望确定是否存在拷贝数变化。测试样本中的核酸称为测试核酸、靶核酸、靶染色体或靶染色体片段。

[0038] 本发明中的术语“参考样本(reference sample)”是指包含多个核酸或核酸混合物的样本,所述核酸序列数据与测试样本序列数据一起用于分析或计算分数和参数,正如下文和权利要求书中所述。在本发明的各种实施例中,尽管不是必需的,但对于所关注的序列,参考样本优选正常或野生类型(例如非非整倍体)。在非整倍体分析中,参考样本可以是不包括非整倍体状态(如21三体)的序列的合格样本,并且可以用于识别试验样本中是否存在21三体等非整倍体。

[0039] 术语“参考组(reference set)”包括多个“参考样本(reference samples)”。

[0040] 基因组的术语“区间(bin)”应理解为基因组的一个片段。一个基因组可以分为几个区间,要么固定大小,要么预定大小,要么可变大小。可能的固定区间大小可以是10Kb、20Kb、30Kb、40Kb、50Kb、60Kb、70Kb等,其中Kb代表千碱基对,即对应1000个碱基对的单元。

[0041] 术语“窗口(window)”应理解为多个区间。

[0042] 术语“比对(aligned)”、“比对(alignment)”、“映射(mapped)”、“比对(aligning)”或“映射(mapping)”是指一个或多个序列,根据其核酸分子与参考基因组中已知序列的顺序确定匹配。这种比对可以手动进行,也可以通过计算机算法进行,例如作为Illumina基因组分析管道的一部分分配的核苷酸数据有效局部比对(ELAND)计算机程序。在比对中序列读段的匹配可以是100%序列匹配或小于100%(非完全匹配)。

[0043] 本发明中的术语“参数(parameter)”是指表征定量数据集和/或定量数据集之间的数字关系的数值。

[0044] 本发明所用术语“临界值(cutoff value)”或“阈值(threshold)”是指其值用于在生物样本分类的两个或多个状态(例如患病和非患病)之间进行仲裁的数值。例如,如果一个参数大于临界值,则对定量数据进行第一次分类(例如病态);或者如果该参数小于临界值,则对定量数据进行不同的分类(例如非病态)。

[0045] 本发明所用术语“不平衡(imbalance)”是指由临床相关核酸序列数量中的至少一个临界值与参考量定义的任何重大偏差。例如,参考量可以是3/5的比率,因此,如果测量比率为1:1,则会出现不平衡。

[0046] 二、脱靶序列:

[0047] 本发明的目的是提供基于在靶向捕获大规模并行测序期间获得的脱靶读取的样本遗传分析方法。这些脱靶读取对进行全面的产前诊断特别有用,但也可用于检测DNA中的畸变,如非整倍体、突变或杂合性缺失(例如在癌症组套中)。通过采用在传统方法中不考虑的脱靶读取,有限数量的可用DNA(尤其是以游离DNA为起点时)和DNA衍生测序数据得到了优化使用。脱靶读取和在靶读取可同时用于一个样本的一个或多个分析,从而限制所需处理步骤的数量,如库制备和新一代测序(NGS)和/或生物信息或计算处理步骤,否则可能集中于或只保留在靶读取,由此,本发明以最优选的方式使用有限数量的材料。

[0048] 本发明的第一方面提供了一种方法,用于确定从怀孕女性中获得的生物样本中是否存在胎儿染色体非整倍体或胎儿杂合性缺失(LOH)。所述方法具体包括以下步骤:

[0049] —获取指示包括母体和胎儿核酸的生物样品的靶向捕获大规模平行测序的序列信息;

[0050] —确定在所述靶向捕获大规模平行测序期间获得的脱靶读取量;以及

[0051] —从上述脱靶读取计数信息可以确定胎儿非整倍体或胎儿杂合性缺失是否存在。

[0052] 具体地,该方法需要从孕妇的生物样本中获取母体和胎儿的DNA。该生物样本可以是血液,但也可以是唾液或血清或来自母亲的任何其他样本,并可用于从母亲和胎儿中获得遗传数据。在测序之前,对样本中的游离DNA进行靶向富集,以获得DNA的一个子集。

[0053] 本领域已知各种靶向富集方法,包括混合捕获方法和基于PCR的扩增子捕获技术,例如,安捷伦公司(Agilent Inc.)的Sureselect®、罗氏制药公司(Roche Inc.)的Nimblegen®和亿明达(Illumina Inc.)公司的TruSEq®。靶向富集的方法通常基于标记核酸或其他分子探针的使用,这些探针能够与基因组或分离的核酸中所需或预期的区域杂交或结合。在随后的步骤中,将非杂交探针冲走,并从样本中捕获和分离杂交探针。此捕获是通过存在标签来执行的。所述标签能够与第二分子结合或连接,所述第二分子能够捕获标签区和杂交区。本领域已知的适宜标签例如生物素,可结合链霉亲和素或抗生物素蛋白。

[0054] 在随后的步骤中，对捕获的区域进行放大和测序。由此，DNA区域被分离和富集。用上述方法浓缩DNA，必然会产生脱靶和在靶读数，因为杂交是一个敏感但不完善的过程，它会在捕获预期片段时会随同捕获大量脱靶片段。

[0055] 在本发明的一个实施例中，方法中使用的探针针对预先定义的靶向区域进行了专门设计。可开发探针的合适组套或诱饵包括微缺失、CNV，例如小的复发性CNV或已知的重复区域。在一个实施例中，所述探针指向已知含有复发性CNV的一个或多个区域或位于所述复发性CNV侧翼的区域。

[0056] 在本发明的另一个实施例中，所述探针是随机设计的，且不针对特定的组套或诱饵。

[0057] 诱饵或组套的大小优选在0.1Kb到100Mb之间、更优选在1Kb到50Mb之间、1Kb到10Mb之间、10Kb到1Mb之间、甚至更优选在20Kb到0.5Mb之间。

[0058] 虽然在技术上，脱靶读取是由于探针的非特异性结合造成的，但本发明的发明人观察到了探针的非特异性结合趋势。换句话说，脱靶读取不是完全随机的，而是受所用探针序列的影响。因此，可以建立一个或多个参考样本的参考集。所述参考样本集可由用户预先定义或选择(例如，从他/她自己的参考样本中选择)。通过允许用户使用自己的参考集，用户将能够更好地捕获他/她的环境中的(例如不同的湿实验室试剂或规定、不同的NGS仪器或平台等)经常性技术变化及其变量。此外，通过高水平的自动化，技术变化(例如与人类操作有关)得以减少。在本发明的较佳实施例中，所述参考集包含预期或已知不包含(相关)非整倍体、LOH或其他基因组畸变的“健康”样本的基因组信息。

[0059] 就本发明而言，脱靶读取计数的量应至少为1×10^6，更优选为至少2×10^6、3×10^6、4×10^6、5×10^6、6×10^6、7×10^6、8×10^6、9×10^6、10×10^6读取计数。

[0060] 所述序列通过新一代测序(NGS)获得。优先采用覆盖率高的测序方法，也称为深度测序。在本发明的进一步的较佳实施例中，生产了在1×10^6和100×10^6之间的总计读数，更优选在10×10^6和50×10^6之间的总计读数，甚至更优选在15×10^6和30×10^6之间的总计读数，例如读数20×10^6。

[0061] 当前技术中可以使用双端测序和单端测序。

[0062] 优选地，单端测序新一代测序被用作降低了排序成本的单端测序。

[0063] 在获得所述靶向捕获的大规模平行测序的NGS读数后，这些读数被映射到一个参考基因组或参考基因组的一部分(区间)。所述映射是通过将读数与所述参考基因组进行比对而成的。

[0064] 随后，分离脱靶读取和在靶读取，从而隔离脱靶读取。优选地，通过自动化方式(例如，使用本领域技术人员已知的、且考虑到探针的靶向区域的适当软件)识别或隔离脱靶读取。

[0065] 确定脱靶读取的读取计数。在本发明的另一个或进一步的实施例中，同时确定了在靶和脱靶的读取计数。根据在参考基因组、区间或窗口中的位置，可以进一步细分在靶读取和/或脱靶读取的总读取量。优选地，读取计数由每个区间确定。

[0066] 在进一步的步骤中，一旦获得，读取计数可以有选择地进行规范化。在样本被设置为预定义的读取量(例如1×10^6读取或更多)的情况下，读取可以针对读取的总次数进行规范化。在本发明的另一个或进一步实施例中，标准化可在参考样本集的基础上发生，其中所

述参考样本优选(尽管不必要)整倍体或本质上的整倍体。这种参考集可以有不同的样本容量。可能的样本容量可以是100个样本,例如50个男性和50个女性样本。技术人员理解用户可以自由选择参考集。优选地,这种规范化发生在区间或窗口级别。

[0067] 优选地,对所述读取次数进行重新校准,以纠正GC含量和/或从所述样本中获得的读取总数。已知GC偏好会加重基因组装配。本领域已知各种GC校正的方法。在本发明的较佳实施例中,所述GC校正将是一个LOESS回归。在本发明的一个实施例中,本方法的用户可以选择各种可能的GC校正。

[0068] 有关GC更正的详细说明,请参见PCT/EP2016/066621,其内容已全部并入本发明中。

[0069] 随后,可以使用脱靶读取计数来获得关于胎儿非整倍体或胎儿LOH的存在与否的信息,或LOH或非整倍体的总体情况的信息(例如,在癌症组套中,详见下文)。

[0070] 可通过本领域已知的能够基于游离DNA检测胎儿非整倍体或LOH的任何算法来确定是否存在基于脱靶读取的胎儿非整倍体。此类系统包括安捷伦(Agilent)的OneSight®算法、亿明达(Illumina)的VeriSeq™或西格诺(Sequenom)的MaterniT21®Plus。一般来说,可以使用所有已知的算法,这些算法能够从获得的读取中获得一个参数,该参数表示非整倍体的存在与否。

[0071] 申请号PCT/EP2016/066621中描述了一种特别合适的方法,其内容已通过引用并入本发明中。简言之,根据比对和获得的脱靶读取计数或其衍生物(可根据所述样本的GC含量和/或读取总数进行选择性校正),计算得分,最终得出一个参数,得以确定样本中是否存在非整倍体。所得得分是从读取计数或在数学上修改读取计数得出的标准化值,由此根据用户定义的参考集进行标准化。因此,每个得分都是通过与参考集的比较来获得的。需要注意的是,当前的方法不需要数据培训或了解真实情况。根据本发明的分析可以使用参考集的本质,不需要最终用户设置任何个人选择或偏好。此外,它可以很容易地被用户实施,而无需访问专有数据库。

[0072] 术语第一得分(first score)用于指与靶向染色体或染色体片段的脱靶读取计数相关的得分。它们的集合是从一组标准化的读取次数中得出的得分集,该读取次数可能包括所述靶向染色体片段或染色体的标准化读取次数。优选地,所述第一得分表示靶向染色体或染色体片段的Z得分或标准得分。优选地,所述集合来源于从包含所述靶向染色体片段或染色体的相应染色体组或染色体片段中获得的一组z数。

[0073] 优选地,所述第一得分表示靶向染色体或染色体片段的Z数或标准数。优选地,所述集合来源于从包含所述靶向染色体片段或染色体的相应染色体组或染色体片段中获得的一组z数。

[0074] 在本发明的最佳实施例中,第一得分及得分集是基于靶向染色体或染色体片段,或所有常染色体或所有染色体(或其区域)的基因组表示来计算的,由此包括靶向染色体或染色体片段。

[0075] 这样的得分可以计算如下:

[0076] $$Zi = \frac{GRi - \mu ref,i}{\sigma ref,i}$$

[0077] 用i表示一个窗口或一个染色体或一个染色体片段,用ref表示参考集。

[0078] 对所述得分集的汇总统计可以计算为，例如，单个得分的平均值或中值，也可以计算为单个得分的标准差或中位数绝对差或平均绝对差。

[0079] 所述参数p可作为第一得分的函数和分数集的导数(例如汇总统计)计算。在本发明的较佳实施例中，所述参数将是通过得分集(或其导数)校正的第一得分与所述得分集的导数之间的比率或相关性。

[0080] 在本发明的另一实施例中，所述参数将是通过第一得分集的汇总统计更正的第一得分与不同的第二得分集的汇总统计之间的比率或相关性，其中这两组得分都包括第一得分。

[0081] 在本发明的特佳实施例中，所述参数p是通过所述得分集汇总统计矫正的第一得分和所述得分及汇总统计的比率或相关性。优选地，从平均值、中位数、标准差、中位数绝对差或平均绝对差中选择汇总统计。在本发明的一个实施例中，使用所述两种的汇总统计数据在功能上相同。在本发明的另一个更佳实施例中，所述得分集的汇总统计在分子和分母上有所不同。

[0082] 通常，本发明的适当实施例涉及以下步骤(在从生物样本上的测序过程中获得脱靶序列之后)。

[0083] —将所述获得的序列与参考基因组进行比对；

[0084] —计算一组染色体片段和/或染色体上的脱靶读取次数，从而获得读取计数；

[0085] —将所述脱靶读取计数或其导数标准化为标准化的读取次数；

[0086] —获得所述标准化读取的第一得分及其得分集，其中所述第一得分来自靶向染色体或染色体片段的标准化读取，所述得分集是来自相应一组染色体或染色体片段，包括所述靶向染色体片段或染色体的一组得分集；

[0087] —根据所述第一得分和所述得分集计算参数p，其中所述参数表示以下两者之间的比率或相关性

[0088] *通过所述得分集的汇总统计进行更正的所述第一得分，以及

[0089] *所述得分集的汇总统计。

[0090] 可能的参数p可以计算如下：

[0091] $$Z\text{of } Z_i = \frac{Z_i - \underset{j=i,a,b,\ldots}{\text{median}}(Z_j)}{\underset{j=i,a,b,\ldots}{\text{sd}}(Z_j)}$$

[0092] 其中Zi表示第一得分，Zj表示得分集，i表示靶向染色体或染色体片段，j表示染色体或染色体片段i,a,b…的合集，包括所述靶向染色体片段或染色体i。

[0093] 在本发明的另一个实施例中，所述参数p被计算为：

[0094] $$Z\text{of } Z_i = \frac{Z_i - \underset{j=i,a,b,\ldots}{\text{mean}}(Z_j)}{\underset{j=i,a,b,\ldots}{\text{mad}}(Z_j)}$$

[0095] 其中Zi表示第一得分，Zj表示得分集，i表示靶向染色体或染色体片段，j表示染色体或染色体片段i,a,b…合集，包括所述靶向染色体片段或染色体i。

[0096] 在本发明的另一个最佳实施例中，所述参数p被计算为：

[0097] $$Z\text{of}\,Z_i = \frac{Z_i - \underset{j=i,a,b,\ldots}{\text{median}}(Z_j)}{\underset{j=i,a,b,\ldots}{\text{mad}}(Z_j)}$$

[0098] 其中Zi表示第一数,Zj表示第二数数集,i表示靶向染色体或染色体片段,j表示染色体或染色体片段i,a,b…的合集,包括所述靶向染色体片段或染色体i。

[0099] 所述MAD对于数据集x_1,x_2,…,x_n可以计算为

[0100] “MAD”＝1.4826x“median”(|x_i-“median”(x)|)

[0101] 也可以使用不使用系数1.4826的MAD。

[0102] 系数1.4826用于确保在变量x,平均值μ和标准差σ呈正态分布的情况下,当n值较大时,MAD的值收敛于σ。为了确保这一点,我们可以得出常数系数应等于1/((Φˆ(-1)(3/4)),而Φˆ(-1)是标准正态分布的累积分布函数的倒数。

[0103] 根据从脱靶读取中获得的数据计算出的参数p随后可与临界值进行比较,以确定与参考量相比是否存在变化(即不平衡),例如,关于两个染色体区域(或区域集)的数量之比。临界值可通过任何数量的合适方法确定。这些方法包括Bayesian-type似然法、序贯概率比检验(SPRT)、错误发现、置信区间、接收者操作特性(ROC)。在本发明的更佳实施例中,所述临界值基于统计考虑或通过测试生物样本进行经验测定。临界值可通过试验数据或验证集进行验证,必要时可在有更多数据可用时进行修改。在本发明的一个实施例中,用户能够根据经验或以前的实验,或例如基于标准统计考虑,定义自己的临界值。如果用户希望提高测试的灵敏度,则可以降低阈值(即使其接近0)。如果用户希望增加测试的特异性,那么用户可以增加阈值(即使其远离0)。用户通常需要在敏感性和特异性之间找到平衡,这种平衡通常是实验室和应用特有的,因此,如果用户可以自己更改阈值,这会很方便。

[0104] 根据所得参数与临界值的比较,可以发现非整倍体的存在与否。

[0105] 优选地,本发明的方法尤其适用于分析与表1中给出的片段或缺失相关的非整倍体,表1中包含可通过所述方法和设备识别的染色体异常的非限制性列表。

[0106] 在本发明的另一个或进一步的实施例中,靶向染色体选自染色体X、Y、6、7、8、13、14、15、16、18、21和/或22。

[0107] 本发明的方法同样可用于评估LOH的存在与否。后者可以通过使用本领域已知的任何能够检测在脱靶读取中具有足够覆盖范围的位置集内的B等位基因频率(BAF)变化的算法来执行。本发明的方法是第一种允许全基因组筛选LOH的方法。

[0108] 这特别是因为脱靶读取的性质,这些读取不是完全随机的。

[0109] 表1

[0110]

染色体	异常	关联疾病
X	XO	特纳综合征
Y	XXY	克氏综合征
	XYY	双 Y 综合征
	XXX	三 Y 综合征
	XXXX	四 X 综合征
	Xp21 缺失	杜氏/贝克综合征，先天性肾上腺发育不良，慢性肉芽肿病
	Xp22 缺失	类固醇硫酸酯酶缺乏症
	Xp26 缺失	X-连锁淋巴增生性疾病
1	1p 单体，三体	
	1p36	1p36 缺失综合征
	1q21.1	121.1 缺失综合征；末端 1q21 缺失综合征
2	单体，三体 2q	生长迟缓，发育和精神发育迟缓，以及轻微的身体异常
	2p15 - 16.1	2pL5-16.1 缺失综合征
	2q23.1	2q23.1 缺失综合征
	2q37	2q37 缺失综合征
3	单体，三体	
	3p	3p 缺失综合征

[0111]

	3q29	3q29 缺失综合征
4	单体，三体	
	4p -	沃尔夫-赫什霍恩综合征
5	5p	猫叫综合征；乐琼综合征
	5q 单体，三体	骨髓增生异常综合征
	5q35	5q35 缺失综合征
6	单体，三体	
	6p25	6p25 缺失综合征
7	7q11.23 缺失	威廉综合征
	单体，三体	儿童单体 7 综合征；骨髓增生异常综合征
8	8q24.1 缺失	朗格-吉迪翁综合征
	8q22.1	纳布卢斯面膜样面部综合征
	单体，三体	骨髓增生异常综合征；沃坎综合征
9	单体 9p	阿尔菲综合征
	单体 9p，部分三体 9p	雷托综合征
	三体	完全三体 9 综合征；嵌合三体 9 综合征
	9p22	9p22 缺失综合征
	9q34.3	9q34.3 缺失综合征
10	单体，三体	急性淋巴细胞白血病或急性非淋巴细胞白血病
	10p14 - p13	迪格奥尔格综合征Ⅱ型
11	11p -	无虹膜；威尔姆斯瘤
	11p13	瓦格综合征
	11p11.2	Potocki Shaffer 综合征
	11p15	贝克威斯 - 韦德曼综合征
	11q -	雅各布森综合征
	单体，三体	
12	单体，三体	
13	13q -	13q -综合征；奥贝尔综合征
	13q14 缺失	

	单体，三体	帕托综合征
14	单体，三体	
15	15q11 - q13 缺失，单体	普拉德 - 威利，天使人综合征
	三体	
16	16q13.3 缺失	鲁宾斯坦 - 泰比
	单体，三体	
17	17p -	17p 综合征
	17q11.2 缺失	史密斯 - 马格尼斯
	17q13.3	米勒 - 迪克
	单体，三体	
	17p11.2 - 12 三体	腓骨肌萎缩综合征 I 型；遗传性压力敏感性周围神经病
18	18p -	18p 部分单体综合征或 Grouchy LamyThieffry 综合征
	单体，三体	爱德华兹综合征
19	单体，三体	
20	20p -	三体 20p 综合征
	20p11.2 - 12 缺失	艾欧吉勒综合征
	20q -	
	单体，三体	
21	单体，三体	唐氏综合征
22	22q11.2 缺失	迪格奥尔格综合征，腭心面综合征，锥体畸形面综合征，常染色体显性 Opitz G/BBB 综合征，Caylor 心面综合征
	单体，三体	完全三体 22 综合征

[0112]

[0113] 由于游离DNA的浓度通常较低，因此，可以对一个样本进行的不同基因测试的数量是有限的。本发明允许使用迄今为止的新数据来生成全面的遗传信息。

[0114] 同时，在靶读取也可用于进一步分析样本，从而最大限度地利用样本。脱靶读取可用于分析样本的一个或多个临床方面，在靶读取可用于分析同一样本的一个或多个第二个临床方面。

[0115] 因此，本发明还涉及一种检测胎儿非整倍体和/或杂合性缺失的方法，以及确定从一个样本中获得的遗传信息的胎儿分数和/或微缺失和/或畸变的方法，因此在上述条件下，对样本进行靶向捕获大规模平行测序，用脱靶读取计数(可选择与在靶读取计数结合)来确定是否存在胎儿非整倍体和/或杂合性缺失，并由此通过在靶读取计数测定胎儿分数和/或微缺失的存在。

[0116] 基于在靶读取的胎儿分数的测定可以通过本领域已知的任何允许基于单端读取的胎儿分数测定的算法进行，特别是PCT/EP2016/066621中所述的方法，该方法已通过引用并入本发明。简言之，胎儿分数的测定依赖于对序列的在靶读取计数的测定，最好是存在于胎儿但不在母亲体内的CNV，或在母亲体内的杂合子。对于后者，探针用于靶向捕获大规模并行测序，优选定向到一组已知的、复发的、在人群中频率相对较高的CNV。在靶读取被用于测定胎儿分数，而产生的拖把读取是确定胎儿分数和/或LOH存在的基础。

[0117] 除了测定胎儿分数外，还可以根据在靶读取的生成来检测微缺失和/或畸变。优选地，可以选择组套或诱饵覆盖一组已知与临床相关的重复性微缺失。选择性地，可以在库制备步骤中消除PCR重复。用于消除重复的合适工具包括例如使用分子条形码和/或基于位置的重复数据消除。随后，根据本领域已知的算法，所获得的在靶读取构成进一步检测是否存在微缺失的基础。

[0118] 通过当前方法分析出的适当的微缺失与综合征有关，包括但不限于迪格奥尔格综合征、普拉德-威利综合征、天使人综合征、神经纤维瘤病I型、神经纤维瘤病Ⅱ型、威廉姆斯综合征、米勒-迪克综合征、纵裂-马格尼斯综合征、鲁宾斯坦-泰比综合征、沃尔夫-赫什霍恩综合征和波托基-卢夫斯基(1p36缺失)。

[0119] 一个合适的靶向组套可以指向已知的与上述综合征相关的区域。

[0120] 综上所述，本发明允许用户生成关于非整倍体状态和来自孕妇的游离部分的DNA中存在LOH的信息。同时，也可以获得有关胎儿分数和存在微缺失的信息，所有这些都不需要从有限数量的游离DNA中进行多次库制备。这具有优势a.o.，因为它不需要分离样本来进行库制备，这将进一步减少反应混合物中存在的胎儿DNA分子的绝对量。

[0121] 本发明的方法不限于在胎儿领域和基于游离DNA检测非整倍体。目前的方法同样可以从基因组DNA、FFPE DNA或任何其他合适的DNA类型中开始使用。因此，本发明还可用于非整倍体和/或LOH的一般检测，例如在癌症检测、预防和/或风险评估领域。本发明基于产生的脱靶读取的方法允许全基因组筛选，尤其是对于杂合性缺失，这在迄今之前是不可能的。

[0122] 因此，本发明的方法同样适用于从个体获得的DNA样本中检测非整倍体和/或杂合性缺失(LOH)，所述方法包括

[0123] —对所述DNA进行靶向捕获大规模平行测序；

[0124] —将脱靶读取与在靶读取分离；

[0125] —确定在所述靶向捕获大规模平行测序期间获得的脱靶读取量；以及

[0126] —从所述脱靶读取计数中获得用于确定所述受试者中所述非整倍体或杂合性缺失存在与否的信息。

[0127] 对于一个专业人员来说，上面描述的分析母体样本的各个方面也很明显地适用于这一一般方法。

[0128] 优选地，上述方法都是由计算机实现的。为此目的，本发明同样涉及一种计算机程序产品，包括计算机可读介质，其编码有多个指令，用于控制计算机系统执行对受试者生物样本中(胎儿)非整倍体的(产前)诊断和/或筛查(胎儿)非整倍体、LOH、微缺失和/或胎儿分数的操作，其中生物样本包括核酸分子。

[0129] 此类操作包括以下步骤：

[0130] —接收生物样本中所含核酸分子的至少一部分的序列(来自患者或孕妇)

[0131] —将所述获得的序列与参考基因组进行比对;

[0132] —将在靶读取与脱靶读取分离;

[0133] —计算脱靶读取次数和(可选地)在靶读取次数;

[0134] —将所述读取计数或其导数标准化为标准化的读取次数;

[0135] —根据脱靶读数计算参数,所述参数指示(胎儿)非整倍体或LOH的存在与否。

[0136] 所述操作可由用户或专业人员在远离样本采集位置和/或湿实验室程序的环境中执行,从生物样本中提取核酸并测序。

[0137] 所述操作可以通过安装在计算机上的适配软件提供给用户,也可以存储到云中。

[0138] 在完成所需的操作后,将向专业人员或用户提供报告或得分,由此,所述报告或得分提供已分析的特征的信息。优选地,报告将包含已分析的患者或样本ID的链接。所述报告或得分可提供关于样本中非整倍体或杂合性缺失存在与否、微缺失存在与否的信息,以及当从样本是从孕妇身上获得时胎儿分数的测定,根据上述方法计算出的参数获得所述信息。报告同样可以提供非整倍体的性质(如果被检测到,例如大或小染色体畸变)和/或被分析样本的质量的信息。

[0139] 本领域技术人员应当理解,上述信息可以在一份报告中提交给专业人员。

[0140] 优选地,上述操作是数字平台的一部分,该平台可以通过各种由计算机实现的操作对样本进行分子分析。