1. 一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；以及

基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息，所述预定区域碱基信息包括胎儿的单倍型，其中包括，

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

2. 根据权利要求 1 所述的方法，其特征在于，所述胎儿基因组 DNA 样本是从孕妇外周血中提取的。

3. 根据权利要求 1 所述的方法，其特征在于，所述测序是利用选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种对所述测序文库进行的。

4. 根据权利要求 1 所述的方法，其特征在于，进一步包括将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

5. 根据权利要求 4 所述的方法，其特征在于，所述参照序列为人类参考基因组。

6. 根据权利要求 1 所述的方法，其特征在于，所述胎儿遗传相关个体是所述胎儿的父母。

7. 根据权利要求 1 所述的方法，其特征在于，在所述惠特比算法中，采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ~ 30，N 为所述预定区域的长度，

采用

$$a_{ij}=\Pr(q_t=i|q_{t-1}=j)=\begin{cases}(1-p_r)^2 & x_t=x_{t-1},y_t=y_{t-1}\\ (1-p_r)\cdot p_r & x_t=x_{t-1},y_t\neq y_{t-1}\ \ or\ \ x_t\neq x_{t-1},y_t=y_{t-1}\\ p_r^2 & x_t\neq x_{t-1},y_t\neq y_{t-1}\end{cases}$$

作为重组转移矩阵，其中，$p_r = re/N$。

8. 根据权利要求 7 所述的方法，其特征在于，re = 25。

9. 根据权利要求 4 所述的方法，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{base}=\sum_{k=\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{\bar{k}})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{\bar{k}})$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

PACIFIC BIOSCIENCES OF CALIFORNIA, INC. EXHIBIT - 1005 - Part 3
Pg.781

10. 根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

11. 根据权利要求 10 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

12. 一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；

分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息，所述预定区域碱基信息包括胎儿的单倍型，其中包括，

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

13. 根据权利要求 12 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。

14. 根据权利要求 12 所述的系统，其特征在于所述测序装置为选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种。

15. 根据权利要求 12 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

16. 根据权利要求 12 所述的系统，其特征在于，所述惠特比算法采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ～ 30，N 为所述预定区域的长度，

采用

$$a_{ik}=\Pr(q_t=k|q_{t-1}=j)=\begin{cases}(1-p_r)^2 & x_t=x_{t-1},y_t=y_{t-1}\\ (1-p_r)p_r & x_t=x_{t-1},y_t\neq y_{t-1}\ \ or\ \ x_t\neq x_{t-1},y_t=y_{t-1}\\ p_r^2 & x_t\neq x_{t-1},y_t\neq y_{t-1}\end{cases}$$

作为重组转移矩阵，其中，$p_r = re/N$。

17. 根据权利要求 16 所述的系统，其特征在于，re = 25。

18. 根据权利要求 15 所述的系统，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{t,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{\bar{k}})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_k)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

[0001] 优先权信息

[0002] 无

技术领域

[0003] 本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

[0004] 遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为3个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第21号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。

[0005] 然而，目前的相关检测手段仍有待改进。

发明内容

[0006] 本发明旨在至少解决现有技术中存在的技术问题之一。

[0007] 在本发明的一个方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：针对胎儿基因组DNA样本，构建测序文库；对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态（hidden states），可以将血浆的测序数据当做观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities），观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），我们则可以通过诸如惠特比算法（Viterbi algorithm）根据隐马尔可夫模型推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0008] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测，从而可以有效地对胎儿基因组的遗传信息进行产前确定。

[0009] 在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，该计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，基于胎儿的测序结果，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0010] 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0011] 本发明的上述和 / 或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

[0012] 图 1 为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

[0013] 图 2 为根据本发明的一个实施例的用于确定胎儿基因组中预定区域核酸序列的系统的结构示意图。

具体实施方式

[0014] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

[0015] 需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

[0016] 确定胎儿基因组中预定区域碱基信息的方法

[0017] 在本发明的第一方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

Pg.785

[0018] 首先，针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，胎儿基因组 DNA 样本的来源并不受特别限制。根据本发明的一些实施例，可以采用任何含有胎儿核酸的孕妇样本。例如，根据本发明的实施例，可以采用的孕妇样本为孕妇乳汁、尿液和外周血。其中，优选孕妇外周血。采用孕妇外周血作为胎儿基因组 DNA 样本的来源，可以有效地实现无创取样方式获得胎儿基因组 DNA，从而可以在不影响胎儿正常发育的前提下，对胎儿的基因组进行有效监测。关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide (Part#1005361 ;Feb2010) 或 Paired-End SamplePrep Guide(Part#1005063 ; Feb2010)，通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

[0019] 在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台 (MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan ;11(1):31-46) 包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等)、ABI-Solid 和 Roche-454(焦磷酸测序) 测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术 (True Single Molecule DNA sequencing)，Pacific Biosciences 公司单分子实时测序 (single molecule real-time(SMRT™))，以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等 (Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4):244-245)。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

[0020] 任选地，在得到测序结果之后，可以将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据。在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参照序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI36.3,HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0021] 接下来，基于与预定区域对应的测序数据，确定预定区域中的部分核酸序列；以及基于所确定的预定区域中的部分核酸序列，按照惠特比算法，确定预定区域的其他核酸序列，以便获得预定区域的核酸序列。根据本发明的实施例，可以通过基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以利用惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息。由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0022] 下面参考图 1，对利用惠特比算法借助隐马尔可夫模型进行分析的原理进行详细描述：

[0023] 在本文中所使用的术语“胎儿遗传相关个体”指的是在遗传意义上，与胎儿之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胎儿遗传相关个体”为胎儿的亲代例如父母。由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态 (hidden states)，可以将血浆的测序数据当做观察序列 (observations)，借助先验数据推算出状态转移概率 (transition probabilities)、观察序列概率分布 (observation symbol probabilities) 和初始状态概率分布 (initial state distribution)，我们则可以通过惠特比算法 (Viterbi algorithm) 推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。

[0024] 详细分析步骤如下：

[0025] 记号：

[0026] I. 需要检测的位点数为 N。

[0027] II. 父母的单倍型分别记为 $FH = \{fh_0, fh_1\}$ 和 $MH = \{mh_0, mh_1\}$，

[0028] 其中，

[0029] $mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$，$fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$，

[0030] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$，

[0031] $k \in \{0, 1\}$，$i = 1, 2, 3, \ldots, N$。

[0032] III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

[0033] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$，

[0034] 其中，$x_i \in \{0, 1\}$，$y_i \in \{0, 1\}$，

[0035] 下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

[0036] 而所有可能出现的隐藏状态组成了集合 Q。

[0037] IV. 测序数据记为，$S = \{s_1, \ldots, s_i, \ldots, s_N\}$，

[0038] 其中，$s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,T}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0039] V. 平均胎儿浓度和平均测序错误率分别记为 θ 和 e。

[0040] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

[0041] I. 初始状态概率分布记为 $\pi = \{\pi_j\}$ $(j \in Q)$。

Pg.787

[0042] 根据本发明的实施例，在没有参考数据的情况下，可以设$\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0043] II. 根据本发明的实施例，记单倍体重组概率为$p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0044] III. 根据本发明的实施例，记单倍体重组转移矩阵记为$A = \{a_{jk}\}\ (j, k \in Q)$，其中$a_{jk}$为隐藏状态转移的概率，即

[0045] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \ \ or \ \ x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

[0046] 胎儿单倍型$h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$的下标 x_i和 y_i组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态。举个例子，$x_i = 0$ 代表，“母源性染色体上，对应基因座上等位基因型为$m_{i,0}$”。

[0047] 第二步，构建观察序列概率矩阵；

[0048] 根据本发明的实施例，记观察序列概率矩阵为$B = \{b_{i,j}(s_i)\}\ (i = 1, 2, 3, \ldots, N,\ j \in Q)$，其中$b_{i,j}(s_i)$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j = \{x_i, y_i\}$）时，观测到这种测序信息的可能性”，即

[0049] $$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_{i0}, m_{i1}\}) = \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}!\, n_{i,C}!\, n_{i,G}!\, n_{i,T}!} (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

[0050] 其中$P_{i,base}$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j = \{x_i, y_i\}$）时，该碱基出现的可能性”，即

[0051] $$P_{i,base} = \Pr(base | q_i = j, \{m_{i0}, m_{i1}\}) = \sum_{k \in \{0,1\}} \frac{1}{2}(1-e)\Delta(base, m_{ik}) + \frac{1}{2}e \cdot \Delta(base, m_{i\bar{k}}) + \frac{1}{2}e \cdot \Delta(base, f_{ik}),$$

[0052] 其中指示函数

[0053] $$\Delta(x, y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

[0054] 此步是进行 HMM 参数，每个位点的观察序列概率分布$b_{i,j}(s_i)$计算，即计算每个位点上不同胎儿单倍型（隐藏状态）下，血浆出现当前测序数据（观察序列）的可能性。

[0055] 第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）；

[0056] 定义局部概率$\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_{i-1}(q_{i-1}) \cdot a_{q_{i-1}, q_i}\right) \cdot b_{q_i}(s_i)$

Pg.788

[0057] 定义逆向指针 $\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$。

[0058] 这里所使用的术语局部概率 $\delta_i(q_i)$ 和逆向指针 $\Psi_i(q_i)$ 都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 LawrenceR. Rabiner, PROCEEDINGS OF THE IEEE, Vol. 77, No. 2, 1989 年 2 月，通过参照将其全文并入本文。

[0059] 第四步，确定最终状态，并回溯最优路径：

[0060] 确定最终状态，$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$

[0061] 按照逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^*=\Psi_i(q_i)$ $(i=1,2,3,\ldots,N-1)$。

[0062] 第五步，输出结果。

[0063] 由此，能够有效地对胎儿基因组的序列进行分析。相比其他已有的产前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上：

[0064] 1）根据本发明实施例，检测的位点不仅针对父源性位点，对母源性位点，即母亲杂合位点，同样可以很好地检测出来胎儿是否遗传了母亲的致病位点，检测准确率可高达 95％以上，且可以检测多种变异类型，扩大了疾病检测的范围。

[0065] 2）根据本发明实施例，不但可以通过一次测序获得多个位点、多种疾病的信息，对于一些在孕妇血浆中覆盖程度较低，单纯提高测序深度依然无法准确判定的基因序列，可以通过本方法推断得到，结果准确可靠。

[0066] 3）根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

[0067] 另外，根据本发明实施例，本发明的确定胎儿基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。除了可进行产前无创检测胎儿基因组信息，达到疾病检测的目的，还可以进行无创产前亲子鉴定，在孩子出生前判定孩子父亲身份，为一些涉及抚养责任和义务、财产纠纷、性侵案等协助侦破。

[0068] 用于确定胎儿基因组中预定区域碱基信息的系统

[0069] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括：文库构建装置 100、测序装置 200 以及分析装置 400。

[0070] 根据本发明的实施例，文库构建装置 100 适于针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所构建的测序文库进行测序，以便获得测序结果，所得到的测序结果由多个测序数据构成。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置，该 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。由此，该系统可以适用于进行无创产前检测。

[0071] 根据本发明的实施例，任选地，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参照序列进行比对，以便确定与预定区域对应的测序数据。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，提高后续对测

Pg.789

序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan;11(1):31-46）包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time(SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4):244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCB136.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0072] 根据本发明的实施例，分析装置 400 适于基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

[0073] 根据本发明的实施例，惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re = 25 ~ 30，优选 25，N 为所述预定区域的长度。

[0074] 采用

[0075] $$a_{ij}=\Pr(q_t=k|q_{t-1}=j)=\begin{cases}(1-p_r)^2 & x_t=x_{t-1},y_t=y_{t-1}\\ (1-p_r)\cdot p_r & x_t=x_{t-1},y_t\neq y_{t-1}\ \ or\ \ x_t\neq x_{t-1},y_t=y_{t-1}\\ p_r^2 & x_t\neq x_{t-1},y_t\neq y_{t-1}\end{cases}$$

[0076] 作为重组转移矩阵，$p_r = re/N$。

[0077] 根据本发明的实施例，将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0078] $$P_{base}=\sum_{base\in\{A,C,G,T\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_x)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_y)+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{xy})$$

[0079] 其中，

[0080] $$\Delta(x,y)=\begin{cases}1-\varepsilon & x=y\\ \varepsilon/3 & x\neq y\end{cases}$$

[0081] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0082] 由此，利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定

胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0083] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

[0084] 在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

[0085] 需要说明的是，本领域技术人员能够理解，在前面所描述的确定胎儿基因组中预定区域核酸序列的方法的特征和优点也适合于确定胎儿基因组中预定区域核酸序列的系统，为描述方便，不再详述。

[0086] 计算机可读介质

[0087] 在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0088] 根据本发明的实施例，指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re = 25 ～ 30，优选 25，N 为所述预定区域的长度，

[0089] 采用

[0090] $$a_{jk}=\Pr(q_t=k\mid q_{t-1}=j)=\begin{cases}(1-p_r)^2 & x_t=x_{t-1},y_t=y_{t-1}\\ (1-p_r)\cdot p_r & x_t=x_{t-1},y_t\neq y_{t-1}\ \ or\ \ x_t\neq x_{t-1},y_t=y_{t-1}\\ p_r^2 & x_t\neq x_{t-1},y_t\neq y_{t-1}\end{cases}$$

[0091] 作为重组转移矩阵，其中，$p_r = re/N$。

[0092] 根据本发明的实施例，所述指令将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0093] $$P_{t,base}=\sum_{k\in\{ACGT\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_x)+\frac{1}{2}\varepsilon\cdot\Delta(base,f_x)$$

[0094] 其中，

[0095] $$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

[0096] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0097] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

[0098] 就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0099] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

[0100] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0101] 此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0102] 下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自Illumina公司。

[0103] 一般方法

[0104] 本发明实施例的主要步骤包括：

[0105] 1）无创采取含有胎儿遗传物质的孕妇样品，提取其中含有的遗传物质。

[0106] 2）胎儿家庭成员如父母和外祖父母等基因组DNA提取和纯化。

[0107] 3）各遗传物质根据不同测序平台测序要求进行文库构建。

[0108] 4）测序获得的数据进行过滤，过滤条件根据质量值，接头污染等来设定。

[0109] 5）获得的高质量序列根据需要进行组装处理，组装结果与人类基因组参考序列进行比对。获得唯一比对的序列，带入模型进行分析。

[0110] 分析模型：

[0111] 记号：

[0112] L.需要检测的位点数为N。

Pg.792

[0113] II. 父母的单倍型分别记为 FH = {fh_0, fh_1} 和 MH = {mh_0, mh_1}，

[0114] 其中，

[0115] mh_k = {$m_{1,k}$, ..., $m_{i,k}$, ..., $m_{N,k}$}，fh_k = {$f_{1,k}$, ..., $f_{i,k}$, ..., $f_{N,k}$}，

[0116] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$。

[0117] $k \in \{0, 1\}$，$i = 1, 2, 3, \ldots, N$。

[0118] III. 将未知的胎儿单倍型记为 H = {h_0, h_1}，特别地，h_0和 h_1分别遗传自母亲和父亲。

[0119] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$。

[0120] 其中，$x_i \in \{0, 1\}$，$y_i \in \{0, 1\}$，

[0121] 下标 x_i和 y_i组成的序列对，q_i = {x_i, y_i} 组成了我们需要解码的隐藏状态，

[0122] 而所有可能出现的隐藏状态组成了集合 Q。

[0123] IV. 测序数据记为，S = {s_1, ..., s_i, ..., s_N}

[0124] 其中，s_i = {$n_{i,A}$, $n_{i,C}$, $n_{i,G}$, $n_{i,T}$}，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0125] V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

[0126] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵；

[0127] I. 初始状态概率分布记为 π = {π_j} ($j \in Q$)，

[0128] 根据本发明的实施例，在没有参考数据的情况下，可以设$\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0129] II. 根据本发明的实施例，记单倍体重组概率为 p_r = re/N，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0130] III. 根据本发明的实施例，记单倍体重组转移矩阵记为 A = {a_{jk}} ($j, k \in Q$)，其中 a_{jk}为隐藏状态转移的概率，即

[0131] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r) \cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

[0132] 第二步，构建观察序列概率矩阵；

[0133] 根据本发明的实施例，记观察序列概率矩阵为 B = {$b_{i,j}(s_i)$} ($i = 1, 2, 3, \ldots, N$, $j \in Q$)，其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，观测到这种测序信息的可能性”，即

[0134] $$b_{i,j}(s_i) = \Pr\left(s_i | q_i = j, \{m_0, m_1\}\right) = \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}! n_{i,C}! n_{i,G}! n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}}$$

[0135] 其中 $P_{i,base}$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，该碱基出

Pg.793

现的可能性”，即

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right)$$

[0136] $$= \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_0\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_0\right),$$

[0137] 其中指示函数

[0138] $$\Delta(x,y) = \begin{cases} 1-\varepsilon & x = y \\ \varepsilon/3 & x \neq y \end{cases}$$

[0139] 第二步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）：

[0140] 定义局部概率 $\delta_i(q_i) = \left(\max_{q_{i-1}\in Q} \delta_{i-1}(q_{i-1})\cdot a_{q_{i-1}q_i}\right)\cdot b_{q_i}(s_i)$

[0141] 定义逆向指针 $\Psi_i(q_i) = \arg\max_{q_{i-1}\in Q} \delta_{i-1}(q_{i-1})\cdot a_{q_{i-1}q_i}$

[0142] 第四步，确定最终状态状态，并回溯最优路径：

[0143] 确定最终状态，$q_N^* = \arg\max_{q_N\in Q} \delta_N(q_N)$，

[0144] 根据逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i(q_{i+1})$ $(i = 1,2,3,\ldots,N-1)$。

[0145] 第五步，输出结果。

[0146] 实施例 1

[0147] 样品收集及处理：

[0148] (1) 所取样品包括一个家庭中父亲和母亲孕期的外周血，胎儿出生后取脐带血，以 EDTA 抗凝管收集，Oragene® DNA 唾液采集 /DNA 纯化试剂盒 OG-250 采集祖父母和外祖父母唾液。

[0149] (2) 祖父母和外祖父母唾液 DNA 提取后用 Infinium® HD Human610-Quad BeadChip 基因芯片进行分型。

[0150] (3) 取母亲孕期外周血，1600g，4℃离心 10 分钟，将血细胞和血浆分开，血浆再以 16000g，4℃离心 10 分钟，进一步去除残留的白细胞。孕妇血浆用 TIANamp Micro DNA Kit(TIANGEN) 提取 DNA，得到母亲和胎儿基因组 DNA 混合物，并且从全血分离血浆后剩余的白细胞中提取母亲基因组 DNA。将所得到的血浆 DNA 根据 Illumia® 公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0151] (4) 父亲外周血、母亲外周血白细胞和胎儿脐带血则直接用 TIANamp Micro DNA Kit(TIANGEN) 提取试剂盒提取基因组 DNA。

Pg.794

[0152] 除血浆DNA样品外，将所获得的所有DNA样品，需用Covaris™打断仪打断至500bp大小的片段。将获得的DNA片段以及血浆DNA样品根据Illumia®公司HiSeq2000™测序仪的上机要求进行建库，具体步骤如下：

[0153] 末端修复反应体系：

[0154]

10×T4 多核苷酸激酶缓冲液	10 μl
dNTPs(10mM)	4 μl
T4 DNA 聚合酶	5 μl
Klenow 片段	1 μl
T4 多核苷酸激酶	5 μl
DNA 片段	30μl
ddH_2O 补齐至 100 μl	

[0155] 20℃反应30分钟后，使用PCR Purification Kit(QIAGEN)回收末端修复产物。将所得到的产物最后溶于34μl的EB缓冲液中。

[0156] 末端添加碱基A反应体系：

[0157]

10×Klenow 缓冲液	5μl
dATP(1mM)	10μl
Klenow (3'-5' exo⁻)	3μl
DNA	32μl

[0158] 37℃温育30分钟后，经MinElute® PCR Purification Kit(QIAGEN)纯化并溶于12μl的EB中。

[0159] 接头连接反应体系：

[0160]

2x Rapid DNA 连接缓冲液	25μl
PEI Adapter oligomix(20uM)	10μl
T4 DNA 连接酶	5μl
添加碱基A的产物	10μl

[0161] 20℃反应15分钟后，使用PCR Purification Kit(QIAGEN)回收连接产物。将所得到的产物最后溶于32μl的EB缓冲液中。

[0162]

PCR 反应体系：	
接头连接反应产物	10 μl
Phusion DNA Polymerase Mix	25 μl
PCR 引物 (10 pmol/μl)	1 μl
Index N(10 pmol/μl)	1 μl
超纯水	13 μl

[0163] 反应程序如下：

[0164]

98℃	30 s	
98℃	10 s	10 个循环
65℃	30 s	
72℃	30 s	
72℃	5 min	
4℃	Hold	

[0165] 使用 PCR Purification Kit(QIAGEN) 回收 PCR 产物。样品最后溶于 50 μ l 的 EB 缓冲液中。

[0166] 构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后，用 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0167] (5) 父母基因组测序分型：

[0168] a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI36.3,HG18）。

[0169] b. 使用 SOAPsnp（南方汉族 (CHS) 家系数据使用的是千人计划数据）对数据进行一致序列 (consensus sequence, CNS) 构建。

[0170] c. 提取出标记位点的基因型。

[0171] (6) 父母单倍体型推断：

[0172] a. 构建含祖辈与父母基因型的群体基因型矩阵，即提取父母、祖辈和南方汉族家系在标记位点的基因型。

[0173] b. 使用 BEAGLE 对父母的单倍型进行推断。

[0174] (7) 胎儿单倍体型推断：

[0175] a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI36.3,HG18）。

[0176] b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

[0177] 构建初始状态概率向量：采取无参考数据模式，即各个初始状态概率相等，均为 0.25。

[0178] 单倍体重组转移矩阵，保守地，我们取 re = 25（其余按一般方法所述）。

[0179] c. 统计每个位点的测序信息，并构建观察序列概率矩阵（其余按一般方法所述）。

[0180] d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

[0181] e. 确定最终状态状态，并回溯最优路径。

[0182] f. 输出。

[0183] 根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下：

[0184]

			预测						合计		
			纯合			杂合					
			总点数	准确数	准确率	总点数	准确数	准确率	总点数	准确数	准确率
实际检测	父源	纯合	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		杂合	85,409	84,735	99.87%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		284,961	284,287	99.76%	108,087	103,912	96.14%	373,048	368,199	98.70%
父源总体			4,881	4,881	100.00%	1,716	1,478	86.13%	6,559	6,359	96.95%

[0185] 工业实用性

[0186] 本发明的确定胎儿基因组中预定区域碱基信息的方法、用于确定胎儿基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胎儿基因组中预定区域的核酸序列进行分析。

[0187] 尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

[0188] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

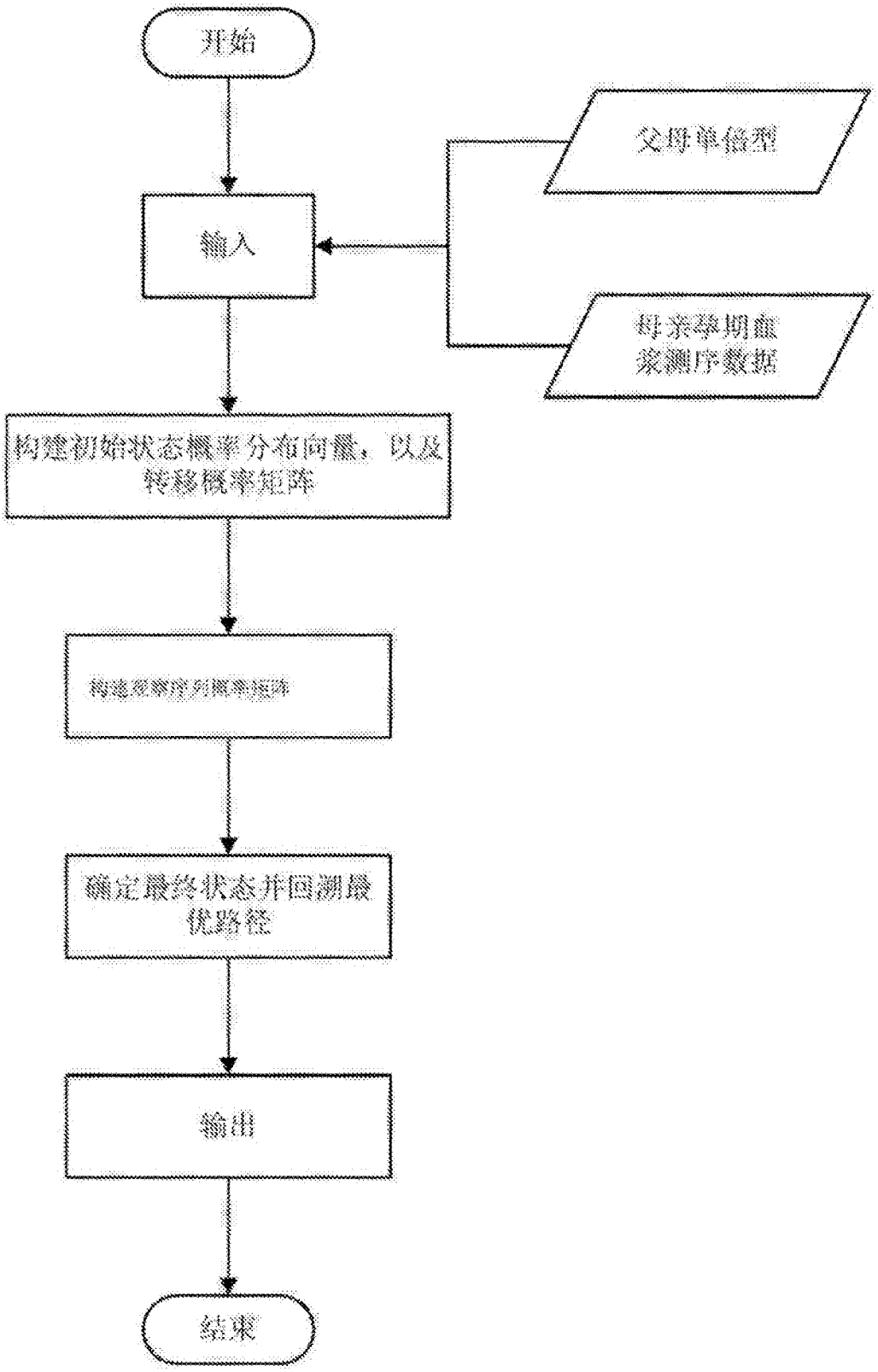

开始
父母单倍型
输入
母亲孕期血
浆测序数据
构建初始状态概率分布向量，以及
转移概率矩阵
确定最终状态并回溯最
优路径
输出
结束

图 1

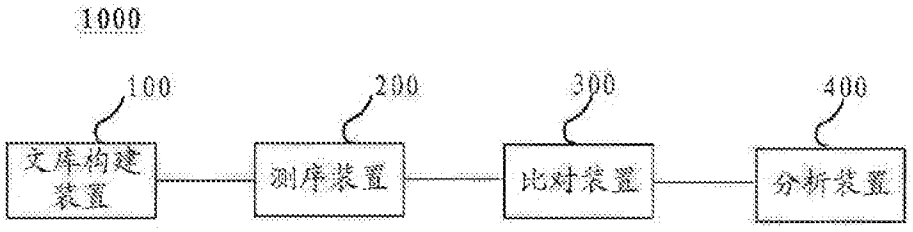


图 2

(12) STANDARD PATENT (11) Application No. **AU 2012380221 B2**
(19) AUSTRALIAN PATENT OFFICE

(54) Title
Method, system and computer readable medium for determining base information in predetermined area of fetus genome

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2012380221** (22) Date of Filing: **2012.05.14**

(87) WIPO No: **WO13/170429**

(43) Publication Date: **2013.11.21**
(44) Accepted Journal Date: **2016.09.29**

(71) Applicant(s)
BGI Diagnosis Co., Ltd.

(72) Inventor(s)
Chen, Shengpei;Ge, Huijuan;Li, Xuchao;Yi, Shang;Wang, Jian;Wang, Jun;Yang, Huanming;Zhang, Xiuqing

(74) Agent / Attorney
Shelston IP Pty Ltd., Level 21, 60 Margaret Street, Sydney, NSW, 2000

(56) Related Art
WO 2007/092473
US 2012/095697

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2013 年 11 月 21 日 (21.11.2013)

WIPO | PCT

(10) 国际公布号
WO 2013/170429 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2012/075478

(22) 国际申请日: 2012 年 5 月 14 日 (14.05.2012)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **深圳华大基因健康科技有限公司 (BGI HEALTH SERVICE CO., LTD.)** [CN/CN]; 中国广东省深圳市盐田区北山道 146 号北山工业区 11 栋 2、3 楼, Guangdong 518083 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **陈盛培 (CHEN, Shengpei)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **葛会娟 (GE, Huijuan)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李旭超 (LI, Xuchao)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **易赏 (YI, Shang)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **汪建 (WANG, Jian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **杨焕明 (YANG, Huanming)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

张秀清 (ZHANG, Xiuqing) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

(54) 发明名称 : 确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) Abstract: Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(57) 摘要: 提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胎儿基因组中预定区域碱基信息的方法，包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对测序文库进行测序，以便获得胎儿的测序结果，该胎儿的测序结果由多个测序数据构成；基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。

WO 2013/170429 A1

2012380221 01 Sep 2016

METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

TECHNICAL FIELD

Embodiments of the present disclosure generally relate to a method of determining base information of a predetermined region in a fetal genome, and a system and a computer readable medium thereof.

BACKGROUND

Any discussion of the prior art throughout the specification should in no way be considered as an admission that such prior art is widely known or forms part of common general knowledge in the field.

Genetic diseases are one kind of diseases caused by changes of genetic materials, having characteristics of being congenital, familial, permanent and hereditary. The genetic diseases may be categorized into 3 classes: monogenetic disease, polygenetic disorder and chromosome abnormality. In which the monogenetic disease is mostly because of genetic function abnormality caused by dominant or recessive inheritance of a single disease-causing gene; while the polygenetic disorder is a kind of disease caused by a plurality of gene changes, which may be influenced by external environment to some extent; and the chromosome abnormality includes number abnormality and structure abnormality, with a most common example being as a Down's Syndrome resulting from Trisomy 21, of which a child patient presenting congenital traits such as mongolism and abnormal body shape, etc. Since there are no effective therapeutic treatments for genetic diseases so far, it can only pertinently perform supportive treatments or drug remission with expensive cost, which may bring heavy burdens both in economy and spirit for societies and families. Thus, it is extremely necessary to do some preventive work by detecting pathological status with a fetus before birth, to achieve a purpose of good prenatal and postnatal care.

However, related detection method still needs to be improved.

2012380221 01 Sep 2016

Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise", "comprising", and the like are to be construed in an inclusive sense as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".

SUMMARY

According to a first aspect, the invention provides a method of determining base information of a predetermined region in a fetal genome, comprising following steps:

constructing a sequencing library based on a genomic DNA sample of a fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

According to a second aspect, the invention provides a system for determining base information of a predetermined region in a fetal genome, comprising:

a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus;

a sequencing apparatus, connected to the library constructing apparatus, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the

sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

According to a third aspect, the invention provides a computer readable medium comprising a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model, which causes the apparatus to perform:

constructing a sequencing library based on the genomic DNA sample of the fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using the hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

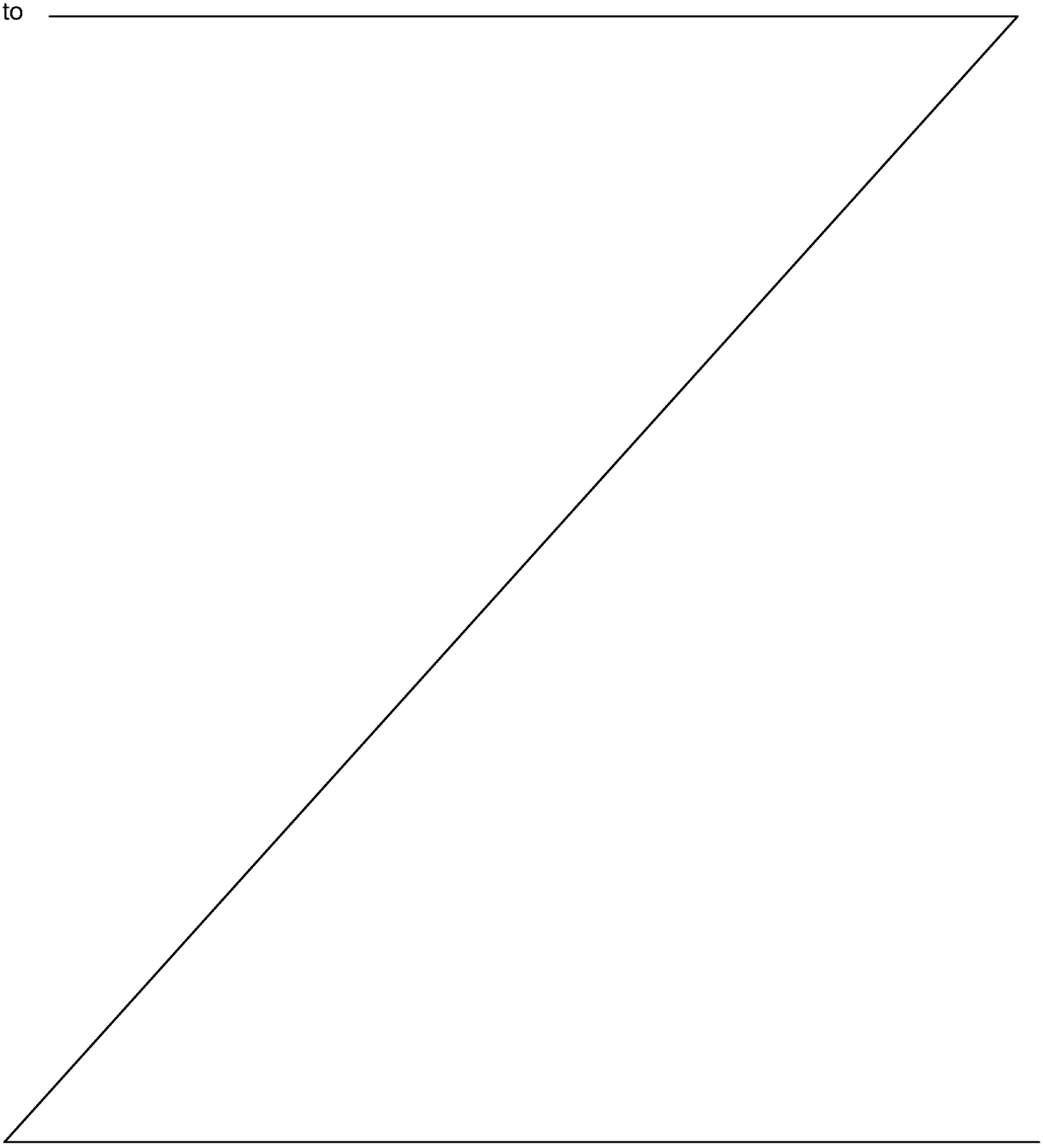
constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

Embodiments of the present disclosure seek to solve at least one of the problems existing in the related art to at least some extent, or to provide a useful alternative.

Embodiments of a first broad aspect of the present disclosure provide a method of determining base information of a predetermined region in a fetal genome. According to



1c

embodiments of the disclosure, the method may comprise: constructing a sequencing library based on a genomic DNA sample of a fetus; subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. A formation of offspring genome equals to a random recombination with parental generation's genome (i.e., an interchange of haplotype recombination, and a random combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth. Thus, according to embodiments of the present disclosure, in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, nucleic acid sequence of a predetermined region in a fetal genome may be determined, by which a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Embodiments of a second broad aspect of the present disclosure provide a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the system may comprise: a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus; a sequencing apparatus, connected to the library constructing apparatus, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. Using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be

determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Embodiments of a third broad aspect of the present disclosure provide a computer readable medium. According to embodiments of the present disclosure, the computer readable medium including a plurality of instructions is adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model. Using the computer readable medium of the present disclosure may effectively execute the plurality of instructions by a processor, to determine a nucleic acid sequence of the predetermined region in the fetal genome in virtue of the hidden Markov Model, for example using the Viterbi algorithm based on the sequencing data of the fetus combining with genetic information of a related individual, by which prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Additional aspects and advantages of embodiments of present disclosure will be given in part in the following descriptions, become apparent in part from the following descriptions, or be learned from the practice of the embodiments of the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects and advantages of embodiments of the present disclosure will become apparent and more readily appreciated from the following descriptions made with reference the accompanying drawings, in which:

Fig.1 is a flow chart showing an analyzing process using a hidden Markov Model according to an embodiment of the present disclosure; and

Fig.2 is a schematic diagram showing a system for determining base information of a predetermined region in a fetal genome according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

Reference will be made in detail to embodiments of the present disclosure. The same or similar elements and the elements having same or similar functions are denoted

by like reference numerals throughout the descriptions. The embodiments described herein with reference to drawings are explanatory, illustrative, and used to generally understand the present disclosure. The embodiments shall not be construed to limit the present disclosure.

It should note that terms such as "first" and "second" are used herein for purposes of description and are not intended to indicate or imply relative importance or significance. Thus, features defined with “first”, “second” may explicitly or implicitly include one or more the features. Furthermore, in the description of the present disclosure, unless otherwise stated, “a/the plurality of” means two or more.

Method of determining base information of a predetermined region in a fetal genome

In a first aspect of the present disclosure, there is provided a method of determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the method may comprise:

firstly, constructing a sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, source of the genomic DNA sample of the fetus is not subjected to any special restrictions. According to some embodiments of the present disclosure, any pregnant samples containing a nucleic acid of a fetal may be used. For example, according to embodiments of the present disclosure, the pregnant sample may be breast milk, urine and peripheral blood from a pregnant woman. In which, the pregnant peripheral blood is preferred. Using the pregnant peripheral blood as the source of the genomic DNA sample of the fetus may effectively realize obtaining the genomic DNA sample of the fetus by noninvasive sampling, by which the fetal genome may be effectively monitored in the premise of having no influence on normal development of fetal growth. As for methods and processes of constructing a sequencing library for the nucleic acid sample, a person skilled in the art may appropriately select depending on different sequencing technology. Detailed process may refer to procedure provided by sequencer manufacturer, such as Illumina Company, for example, refer to Multiplexing Sample Preparation Guide (Part#1005361; Feb 2010) or Paired-End SamplePrep Guide (Part#1005063; Feb 2010) from Illumina Company, which are incorporated herein for reference. According to embodiments of the present disclosure, methods and devices for extracting a nucleic acid from a biological sample are not

subjected to any special restrictions, which may be performed using a commercial nucleic acid extracting kit.

After being constructed, obtained sequencing library is applied to a sequencer, to obtain a corresponding sequencing result consisting of a plurality of sequencing data. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus.

Optionally, after being obtained, the sequencing result may be aligned to a reference sequence, to determine sequencing data corresponding to the predetermined region. Term of "predetermined region" used herein should be broadly understood, referring to any region of a nucleic acid molecule containing a possible predetermined event. For SNP analysis, it may be a region containing SNP site. For analyzing chromosome aneuploidy, the predetermined region refers to entire or part of the chromosome to be analyzed, i.e., selecting sequencing data deriving from the chromosome. Methods of

selecting sequencing data deriving from a corresponding region in the sequencing result are not subjected to any special restrictions. According to embodiments of the present disclosure, all obtained sequencing data may be aligned to a reference sequence with a known nucleic acid, to obtain the sequencing data deriving from the predetermined region. In addition, according to embodiments of the present disclosure, the predetermined region may also be a plurality of dispersal points which are not discontinuous in a genome. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

Then, determining a part of a nucleic acid sequence of the predetermined region based on sequencing data corresponding to the predetermined region; and determining other parts of the nucleic acid sequence based on determined part of the nucleic acid sequence of the predetermined region using Viterbi algorithm, to obtain the nucleic acid sequence of the predetermined region. According to embodiments of the present disclosure, the base information of the predetermined region is determined based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. According to embodiments of the present disclosure, the base information of the predetermined region is determined using the hidden Markov Model is performed based on Viterbi algorithm. Thus, a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Referring to Fig.1, a principal for analysis using Viterbi algorithm in virtue of a hidden Markov Model is descripted in details below:

In the genetic sense, term of “a related individual” refers to individuals having a genetic relationship with a fetus. For example, according to embodiments of the present disclosure, “a related individual” may be a rental generation of a fetus, such as parents. Thus, a formation of offspring genome equals to a random recombination with parental generation’s genome (i.e., an interchange of haplotype recombination, and a random

combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth.

Steps of analyzing are show below in details:

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

$mh_k = \{m_{1,k},...,m_{i,k},...,m_{N,k}\}$, $fh_k = \{f_{1,k},...,f_{i,k},...,f_{N,k}\}$,

$\forall fh_{i,k}, mh_{i,k} \in \{A,C,G,T\}$,

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

$h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$, $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$,

in which $x_i \in \{0,1\}$, $y_i \in \{0,1\}$,

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1,...,s_i,...,s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively

recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ ($j \in Q$).

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ ($j,k \in Q$), in which a_{jk} represents a probability of hidden states transition, i.e.,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ ($i = 1,2,3,...,N$, $j \in Q$), in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$b_{i,j}(s_i) = \Pr\left(s_i \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot \left(P_{i,A}\right)^{n_{i,A}} \cdot \left(P_{i,C}\right)^{n_{i,C}} \cdot \left(P_{i,G}\right)^{n_{i,G}} \cdot \left(P_{i,T}\right)^{n_{i,T}},$$

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, f_{y_i}\right),$$

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

in which, an indicator function is

Such step is to perform HMM parameter, calculating a probability distribution of observation in each site $b_{i,j}(s_i)$, i.e., calculating a possibility presenting current sequencing data (observations) in the pregnant plasma, assuming different fetal haplotypes in each site.

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

Definition: partial probability $$\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i),$$

Definition: reversal cursor $$\Psi_i(q_i) = \arg\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}.$$

Terms of "partial probability $\delta_i(q_i)$" and "reversal cursor $\Psi_i(q_i)$" used herein both follow classic definitions of Viterbi algorithm. Detailed descriptions for the definition of the parameter may refer to Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77, No.2, February 1989, which is incorporated herein by reference.

Step 4, determining a final state, and tracing back an optional path

Determination of the final state, $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$.

The most possible fetal haplotype $q_i^* = \Psi_i(q_i)$ ($i = 1,2,3,...,N-1$) is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

Thus, the sequence of the fetal genome may be effectively analyzed. Comparing to other existing method of antenatal detection, the method of the present disclosure may have following technical advantages, mainly embodying in accuracy and amount of genetic information obtainable:

1) According to embodiments of the present disclosure, a site to be detected is not limited to a parental site, for a maternal site, i.e., a maternal heterozygous site, whether a fetus inherits a maternal pathopoiesia site may also be detected excellently, with an accuracy up to 95% or more; and a plurality of abnormality types can be detected, which enlarges a range of disease detection.

2) According to embodiments of the present disclosure, information of a plurality of site and diseases may be obtained by one time of sequencing; while those gene sequence, having a low coverage in the pregnant plasma which is not able to be accurately determined only by enhancing sequencing depth, may be obtained by the method of the present disclosure, with an accurate and liable result.

3. According to embodiments of the present disclosure, a plotting with a genetic disease may be performed, some related diseases may be directly deduced with information of other sites, with a large amount of information obtained for one time, which has a more instructive meaning for clinical detection.

In addition, according to embodiments of the present disclosure, the method of

determining base information of a predetermined region in a fetal genome, not limited to a certain genetic polymorphic sites such as SNP or STR, is adapted for all genetic polymorphic sites, which may be parallel used for a plurality of sites, to verify each other. Besides applying to antenatal noninvasive detect genomic information of a fetus, achieving a purpose of disease detection, the method of the present disclosure may also be used in noninvasive antenatal paternity identification, i.e., determining an identity of a fetus' father prior birth, providing assistance for disputes involving rearing responsibilities and obligations, property and sexual assault cases, etc.

System for determining base information of a predetermined region in a fetal genome

In another aspect of the present disclosure, there is provided a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, referring to Fig.2, the system 1000 may comprises: a library constructing apparatus 100, a sequencing apparatus 200 and an analyzing apparatus 400.

According to embodiments of the present disclosure, the library constructing apparatus 100 is adapted for constructing sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, the sequencing apparatus 200 is connected to the library constructing apparatus 100, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data. According to embodiments of the present disclosure, the system 1000 may also comprise a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood. Thus, the system may be adapted for noninvasive antenatal detection.

According to embodiments of the present disclosure, optionally, the system may also comprise an aligning apparatus 300. According to embodiments of the present disclosure, the aligning apparatus 300 is connected to the sequencing apparatus 200, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination

Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

According to embodiments of the present disclosure, the analyzing apparatus 400 is connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model.

According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination

probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

Analysis with sequencing data, which is detailed descripted above, is also adapted to the system for determining base information of a predetermined region in a fetal genome, which is omitted for brevity.

Thus, using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

Terms of “connected” should be broadly understood, which may refer to a direct connection or indirect connection, as long as achieving the above functional connection.

It should note that a person skilled in the art may understand that features and advantages of the method of determining base information of a predetermined region in a fetal genome described above may also adapted to the system for determining base information of a predetermined region in a fetal genome, which are omitted for brevity.

Computer readable medium

In a further aspect of the present disclosure, there is provided a computer readable medium. According to embodiments of the present disclosure, the computer readable medium includes a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model. Thus, using the computer readable medium may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

According to embodiments of the present disclosure, the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm. According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

wherein

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$$

.

Analysis with sequencing data, which is detailed descripted above, is also adapted to the computer readable medium, which is omitted for brevity.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

As to the specification, "computer readable medium" may be any device adaptive for including, storing, communicating, propagating or transferring programs to be used by or in combination with the instruction execution system, device or equipment. More specific examples of the computer readable medium comprise but are not limited to: an electronic connection (an electronic device) with one or more wires, a portable computer enclosure (a magnetic device), a random access memory (RAM), a read only memory (ROM), an erasable programmable read-only memory (EPROM or a flash memory), an optical fiber device and a portable compact disk read-only memory (CDROM). In addition, the computer readable medium may even be a paper or other appropriate medium capable of printing programs thereon, this is because, for example, the paper or other appropriate medium may be optically scanned and then edited, decrypted or processed with other appropriate methods when necessary to obtain the programs in an electric manner, and then the programs may be stored in the computer memories.

It should be understood that each part of the present disclosure may be realized by the hardware, software, firmware or their combination. In the above embodiments, a plurality of steps or methods may be realized by the software or firmware stored in the memory and executed by the appropriate instruction execution system. For example, if it is realized by the hardware, likewise in another embodiment, the steps or methods may be realized by one or a combination of the following techniques known in the art: a discrete logic circuit having a logic gate circuit for realizing a logic function of a data signal, an application-specific integrated circuit having an appropriate combination logic gate circuit, a programmable gate array (PGA), a field programmable gate array (FPGA), etc.

Those skilled in the art shall understand that all or parts of the steps in the above exemplifying method of the present disclosure may be achieved by commanding the related hardware with programs. The programs may be stored in a computer readable storage medium, and the programs comprise one or a combination of the steps in the method embodiments of the present disclosure when run on a computer.

In addition, each function cell of the embodiments of the present disclosure may be

integrated in a processing module, or these cells may be separate physical existence, or two or more cells are integrated in a processing module. The integrated module may be realized in a form of hardware or in a form of software function modules. When the integrated module is realized in a form of software function module and is sold or used as a standalone product, the integrated module may be stored in a computer readable storage medium.

Reference will be made in detail to examples of the present disclosure. It would be appreciated by those skilled in the art that the following examples are explanatory, and cannot be construed to limit the scope of the present disclosure. If the specific technology or conditions are not specified in the examples, a step will be performed in accordance with the techniques or conditions described in the literature in the art (for example, referring to J. Sambrook, *et al.* (translated by Huang PT), Molecular Cloning: A Laboratory Manual, 3rd Ed., Science Press) or in accordance with the product instructions. If the manufacturers of reagents or instruments are not specified, the reagents or instruments may be commercially available, for example, from Illumina company.

General method

The method according to embodiments of the present disclosure mainly comprises following steps:

1) noninvasive sampling a pregnant sample containing fetal genetic materials, extracting genomic DNA therefrom;

2) extracting and purifying genomic DNA sample from family members of the fetus, such as parents or grandparents thereof;

3) constructing a sequencing library with every genetic material in accordance with an requirement for different sequencing platform;

4) filtering obtained sequencing data, with filtering criteria based on quality value, adaptor contamination and etc;

5) assembling obtained high-quality sequences as required, aligning an assembled result to a human genome reference sequence, to obtain uniquely-mapped sequences for analyzing using the model.

Analysis model

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

$mh_k = \{m_{1,k}, ..., m_{i,k}, ..., m_{N,k}\}$, $fh_k = \{f_{1,k}, ..., f_{i,k}, ..., f_{N,k}\}$,

$\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$,

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

$h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$, $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$,

in which $x_i \in \{0,1\}$, $y_i \in \{0,1\}$,

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1, ..., s_i, ..., s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ ($j \in Q$).

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of

each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ ($j,k \in Q$), in which a_{jk} represents a probability of hidden states transition, i.e.,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ ($i = 1,2,3,...,N$, $j \in Q$), in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$P_{i,base} = \Pr(base | q_i = j, \{m_0, m_1\}) = \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon \cdot \Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon \cdot \Delta(base, f_{y_i}),$$

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

in which, an indicator function is

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

$$\delta_i(q_i)=\left(\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}(s_i)$$

Definition: partial probability

$$\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$$

Definition: reversal cursor .

Step 4, determining a final state, and tracing back an optional path

Determination of the final state, $q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$.

The most possible fetal haplotype $q_i^*=\Psi_i(q_i)$ ($i=1,2,3,...,N-1$) is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

EXAMPLE 1

Sample collection and treatment

(1) collected sample included: peripheral blood extracted from a father and a pregnant mother within a family, and fetal umbilical cord blood after birth, all of which were collected in a tube containing EDTA for anticoagulation; saliva were collected from four grandparents using a Oragene® DNA saliva collection/DNA purification kit OG-250.

(2) extracted saliva DNA of the four grandparents were subjected to genotyping using Infinium® HD Human610-Quad BeadChip gene chip.

(3) the peripheral blood collected from the pregnant mother was centrifuged with 1600 g at 4°C for 10 min, to separate blood cells and plasma. Then obtained plasma was centrifuged with 16000 g at 4°C for 10 min, to further remove residual leukocytes, to obtain final plasma of the pregnant mother. Then genomic DNA was extracted from the final plasma of the pregnant mother using TIANamp Micro DNA Kit (TIANGEN), to obtain a genomic DNA mixture of mother and fetus thereof. Then maternal genomic DNA was extracted from removed residual leukocytes. Obtained plasma DNA were subjected to library construction based on requirement for HiSeq2000™ sequencer of Illumia® sequencer. Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected

to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(4) parental peripheral blood, leukocytes extracted from maternal peripheral blood and fetal umbilical cord blood were extracted with their respective genomic DNA using TIANamp Micro DNA Kit (TIANGEN).

Except for plasma DNA sample, all obtained DNA sample needed to be fragmented using Covaris™ to have a length of 500 bp. Obtained DNA fragments and plasma DNA sample were subjected to library construction based on the requirement for HiSeq2000™ sequencer of Illumia® sequencer, with a detailed procedure:

End-reparing reacting system:

10× T4 Polynucleotide kinase buffer	10 μL
dNTPs (10 mM)	4 μL
T4 DNA polymerase	5 μL
Klenow fragments	1 μL
T4 Polynucleotide kinase	5 μL
DNA fragments	30 μL
ddH_2O	up to 100 μL

After reacting at 20°C for 30 min, PCR Purification Kit (QIAGEN) was used in recycling end-repaired products. Then the recycled end-repaired products were finally dissolved in 34μL of EB buffer.

A reacting system for adding base A at end:

10× Klenow buffer	5 μL
dATP (1 mM)	10 μL
Klenow (3'-5' exo^-)	3 μL
DNA	32 μL

After incubating at 37°C for 30 min, obtained products were purified by MinElute® PCR Purification Kit (QIAGEN) and dissolved in 12 μL of EB buffer, to obtain DNA samples added with base A at end.

Ligating adaptor reacting system:

2× Rapid DNA ligating buffer	25 μL
PEI Adapter oligo-mix (20 μM)	10 μL

T4 DNA ligase	5 μL
DNA sample added with base A at end	10 μL

After reacting at 20°C for 15 min, PCR Purification Kit (QIAGEN) was used in recycling ligated products. The ligated products were finally dissolved in 32 μL of EB buffer.

PCR reacting system:

Ligated product	10 μL
Phusion DNA Polymerase Mix	25 μL
PCR primer (10 pmol/μL)	1 μL
Index N (10 pmol/μL)	1 μL
UltraPureTM Water	13 μL

Reacting procedure was shown as below:

98°C	30 s	
98°C	10 s	10 cycles
65°C	30 s	
72°C	30 s	
72°C	5 min	
4°C	Hold	

PCR Purification Kit (QIAGEN) was used in recycling PCR products, which were finally dissolved in 50 μL of EB buffer.

Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(5) parental and maternal genomes sequencing genotyping

a. the sequencing data were aligned to a human reference genome (Hg19, NCBI 36.3) using SOAP2.

b. obtained data were subjected to consensus sequence (CNS) construction using SOAPsnp (thousands of planning data were used for Southern Han (CHS) pedigree data).

c. genotypes of a maker site were extracted.

(6) determination of parents' haplotypes

a. constructing a group genotype matrix containing ancestors' and parents' genotypes, i.e., extracting genotypes in the marker site of parents, ancestors and Southern Han pedigree.

b. deducing parents' haplotypes using BEAGLE.

(7) determination of fetal haplotype

a. aligning plasma sequencing data to a human reference genome ((Hg19, NCBI 36.3) using SOAP2;

b. constructing a probability vector of initial states, and a transition matrix of haplotypes recombination,

constructing the probability vector of initial states: taking a model of non-reference data, i.e., probabilities of every initial states were equal, being 0.25.

constructing the transition matrix of haplotypes recombination: conservatively, $re = 25$ (others were same as descriptions in "general method");

c. calculating sequencing information of each site, and constructing a probability matrix of observations (others were same as descriptions in "general method");

d. constructing a partial probability matrix, and a reversal curse (others were same as descriptions in "general method");

e. determining a final state, and tracing back an optional path; and

f. outputting.

According to genotyping results, the accuracy thereof were shown below:

			mother						total		
			homozygosis			heterozygosis					
			site number	accurate number	accuracy	site number	accurate number	accuracy	site number	accurate number	accuracy
autosome	father	homozygosis	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		heterozygosis	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
			264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
chromosome X			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

Industrial Applicability

The method of determining base information of a predetermined region in a fetal genome, the system for determining base information of a predetermined region in a fetal genome and a computer readable medium according to embodiments of the present disclosure may be effectively applied in analyzing the nucleic acid sequence of the predetermined region in the fetal genome.

Although explanatory embodiments have been shown and described, it would be appreciated by those skilled in the art that the above embodiments cannot be construed to limit the present disclosure, and changes, alternatives, and modifications can be made in the embodiments without departing from spirit, principles and scope of the present disclosure.

Reference throughout this specification to "an embodiment," "some embodiments", "one embodiment", "another example", "an example", "a specific examples", or "some examples", means that a particular feature, structure, material, or characteristic described in connection with the embodiment or example is included in at least one embodiment or example of the present disclosure. Thus, the appearances of the phrases such as "in some embodiments," "in one embodiment", "in an embodiment", "in another example, "in an example," "in a specific examples," or "in some examples," in various places throughout this specification are not necessarily referring to the same embodiment or example of the present disclosure. Furthermore, the particular features, structures, materials, or characteristics may be combined in any suitable manner in one or more embodiments or examples.

201238022 1 01 Sep 2016

WHAT IS CLAIMED IS:

1. A method of determining base information of a predetermined region in a fetal genome, comprising following steps:

constructing a sequencing library based on a genomic DNA sample of a fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

2. The method of claim 1, wherein the genomic DNA sample of the fetus is extracted from pregnant peripheral blood.

3. The method of claim 1 or claim 2, further comprising a step of aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the reference sequence is a human reference genome,

wherein the step of aligning the sequencing result of the fetal genome to the reference sequence to determine sequencing result deriving from the predetermined region further comprises:

determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}$.

4. The method of any one of claims 1 to 3, wherein the related individual is a parent of the fetus.

5. The method of any one of claims 1 to 4, wherein the step of determining the base information of the predetermined region using the hidden Markov Model is performed based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

6. The method of any one of claims 1 to 5, wherein the predetermined region is a site previously determined as having a genetic polymorphism,

wherein optionally the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

7. A system for determining base information of a predetermined region in a fetal genome, comprising:

a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus;

a sequencing apparatus, connected to the library constructing apparatus, and

adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

8. The system of claim 7, further comprising a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood.

9. The system of claim 2 or claim 8, further comprising an aligning apparatus, connected to the sequencing apparatus, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x=y \\ e/3 & x\neq y \end{cases}$.

10. The system of any one of claims 7 to 9, wherein the analyzing apparatus is adapted for determining the base information of the predetermined region using a hidden Markov Model based on Viterbi algorithm, wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

11. A computer readable medium comprising a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model, which causes the apparatus to perform:

constructing a sequencing library based on the genomic DNA sample of the fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using the hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

12. The computer readable medium of claim 11, wherein the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr\left(q_i = k | q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

13. The computer readable medium of claim 11 or claim 12, wherein the plurality of instructions are adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

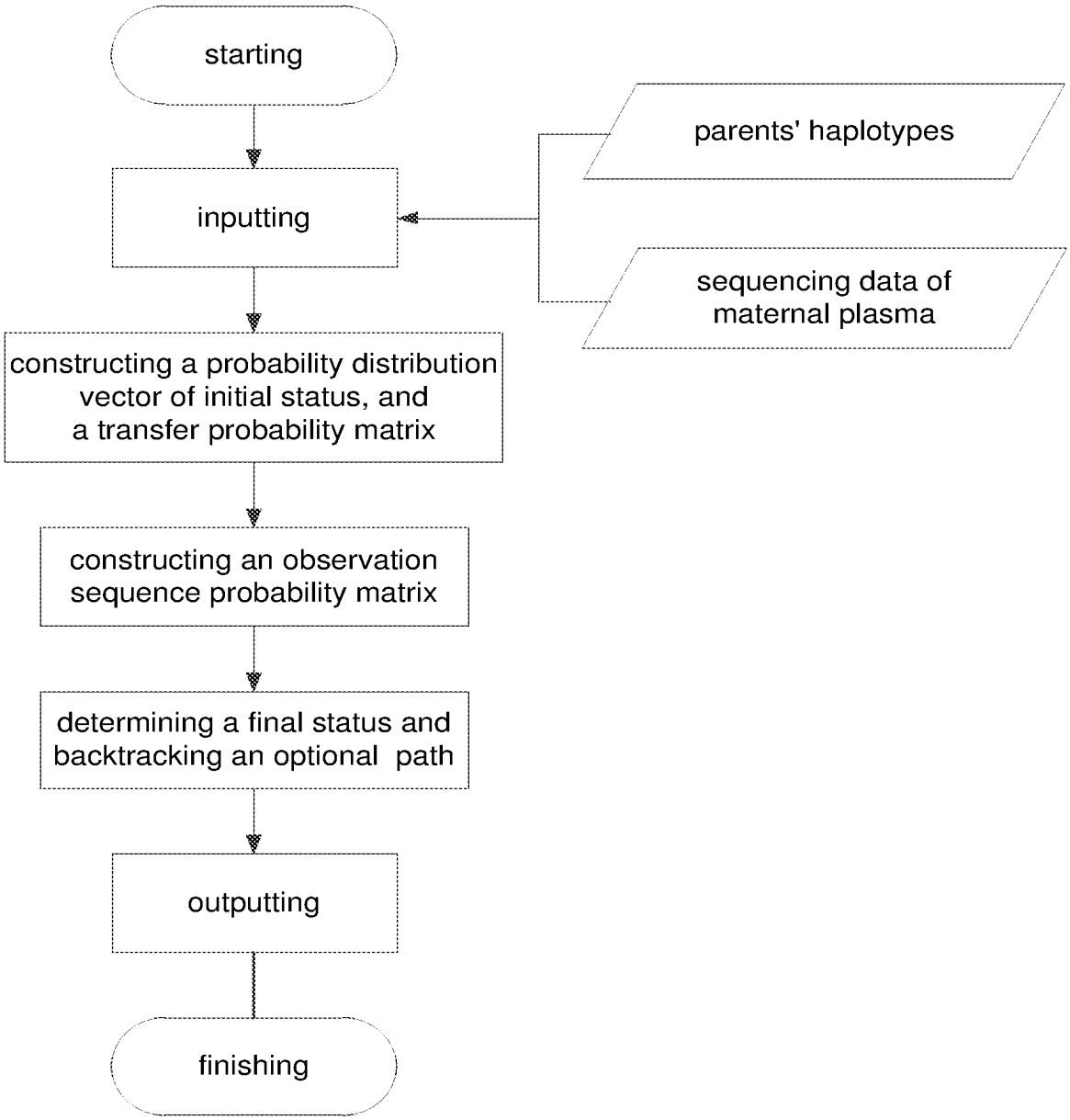

starting
inputting
parents' haplotypes
sequencing data of maternal plasma
constructing a probability distribution vector of initial status, and a transfer probability matrix
constructing an observation sequence probability matrix
determining a final status and backtracking an optional path
outputting
finishing

Fig.1

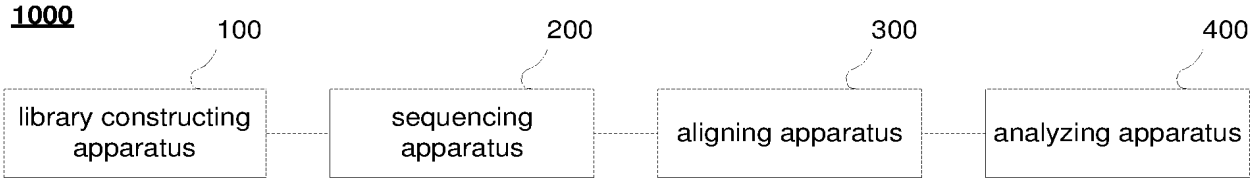

1000
100
200
300
400
library constructing apparatus
sequencing apparatus
aligning apparatus
analyzing apparatus

Fig.2

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2017 (26.01.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/012544 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/CN2016/090642

(22) International Filing Date:
20 July 2016 (20.07.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/194,702 20 July 2015 (20.07.2015) US

(71) Applicant: **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Office of Research and Knowledge Transfer Services, Room 301, Pi Ch'iu Building, Shatin, New Territories, Hong Kong 999077 (CN).

(72) Inventors: **LO, Yuk-Ming Dennis**; 4th Floor, 7 King Tak Street, Homantim, Kowloon, Hong Kong 999077 (CN). **CHAN, Kwan Chee**; Flat C, 15th Floor, Block 5, Peak One, Mei Tin Road, Shatin, New Territories, Hong Kong 999077 (CN). **CHIU, Rossa Wai Kwun**; House 31, Double Haven, 52 Ma Lok Path, Shatin, New Territories, Hong Kong 999077 (CN). **JIANG, Peiyong**; Flat 7 on 1st Floor of Block B of Kwong Lam Court, Nos. 62-66 (No.62) Siu Lek Yuen Road, Shatin, New Territories, Hong Kong 999077 (CN). **SUN, Kun**; Rm 711, Block A, Staff Quarters, Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, New Territories, Hong Kong (CN).

(74) Agent: **INSIGHT INTELLECTUAL PROPERTY LIMITED**; 19 A, Tower A, InDo Building, No. 48A Zhichun Road, Haidian District, Beijing 100098 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

WO 2017/012544 A1

(54) Title: METHYLATION PATTERN ANALYSIS OF HAPLOTYPES IN TISSUES IN DNA MIXTURE

(57) Abstract: Systems, apparatuses, and method are provided for determining the contributions of different tissues to a biological sample that includes a mixture of cell-free DNA molecules from various tissues types, e.g., as occurs in plasma or serum and other body fluids. Embodiments can analyze the methylation patterns of the DNA mixture (e.g., methylation levels at particular loci) for a particular haplotype and determine fractional contributions of various tissue types to the DNA mixture, e.g., of fetal tissue types or tissue types of specific organs that might have a tumor. Such fractional contributions determined for a haplotype can be used in a variety of ways.

METHYLATION PATTERN ANALYSIS OF HAPLOTYPES IN TISSUES IN DNA MIXTURE

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] The present application claims priority to U.S. Provisional Application No. 62/194,702, entitled "Methylation Pattern Analysis Of Haplotypes In Tissues In A DNA Mixture" filed July 20, 2015, the entire contents of which is herein incorporated by reference for all purposes.

BACKGROUND

[0002] It has previously been demonstrated that through the analysis of plasma DNA of a pregnant woman carrying a fetus, the maternal haplotypes inherited by the fetus can be deduced using the process of relative haplotype dosage analysis (RHDO) (Lo et al. Sci Transl Med 2010; 2: 61ra91 and US Patent 8,467,976). The haplotype information for the pregnant woman can be used. The haplotype information can be obtained using family analysis or a method for the direct analysis of the haplotype (e.g. Fan et al. Nat Biotechnol 2011; 29: 51-57; Snyder et al. Nat Rev Genet 2015; 16: 344-358). SNPs that are heterozygous in the mother but homozygous in the father can be used for the RHDO analysis.

[0003] Such a use of specific SNPs can limit the loci that can be used, and therefore limit the amount of data and accuracy. Such use of specific SNPs may also limit the clinical utility of the method as DNA samples from additional family members may not be available, and methods for the direct analysis of haplotypes would add costs to the analysis.

BRIEF SUMMARY

[0004] Embodiments are described for determining the contributions of different tissues to a biological sample that includes a mixture of cell-free DNA molecules from various tissues types, e.g., as occurs in plasma or serum and other body fluids. Embodiments can analyze the methylation patterns of the DNA mixture (e.g., methylation levels at particular loci) for a particular haplotype and determine fractional contributions of various tissue types to the DNA

mixture, e.g., of fetal tissue types or tissue types of specific organs that might have a tumor. Such fractional contributions determined for a haplotype can be used in a variety of ways.

[0005] In some embodiments, two fractional contributions of a tissue type can be determined using methylation levels of two sets of cell-free DNA molecules from a maternal sample, each set being for a different one of two parental haplotypes of a parent of a fetus, for a chromosomal region being analyzed. In various implementations, the maternal sample can be the plasma or serum sample from a woman pregnant with one or more fetuses. The two fractional contributions can be used to identify a portion of the fetal genome. For example, a separation value between the two fractional contributions of fetal tissue can indicate the fetal genotype at a locus and can indicate which of the two parental haplotypes is inherited by the fetus. For instance, the higher fractional contribution can indicate the inherited haplotype, and both can be inherited if the separation value is less than a threshold; both haplotypes could be inherited when both parents share the haplotype (or allele for a genotype) for the region being analyzed.

[0006] In some embodiments, only one fractional contribution of fetal tissue is determined for one haplotype. When the one fractional contribution exceeds a reference value (e.g., as determined from other samples), the fetus can be determined to have inherited the one haplotype for the region being analyzed.

[0007] In some embodiments, two methylation levels can be determined for two sets of cell-free DNA molecules from a maternal sample, each set being for a different one of two parental haplotypes of a parent of a fetus, as part of identifying a portion of the fetal genome. The two methylation levels can be compared to each other to identify which haplotype is inherited by the fetus, e.g., by which methylation level is lower. For example, a fetus contributes cell-free DNA molecules that are hypomethylated, and a measurement of a lower methylation level of one haplotype indicates that the one haplotype is inherited by the fetus.

[0008] In some embodiments, a sequence imbalance can be detected for a target chromosomal region of a fetus using a mixture of cell-free DNA molecules from a plurality of tissues types. Target heterozygous loci can be identified for the target chromosomal region having a first target haplotype and a second target haplotype having different alleles. A first target fractional contribution of the fetal tissue type in the mixture can be determined using methylation levels at the target heterozygous loci, where the methylation levels are determined using a target set of

cell-free DNA molecules located at (i.e., covering) the loci of the first haplotype. Similarly, a first reference fractional contribution of the fetal tissue type can be determined. A separation value of the first target fractional contribution and the first reference fractional contribution can be compared to a threshold value to determine whether the fetus has a sequence imbalance. If the two fractional contributions are significantly different, then a sequence imbalance can be determined. The specific threshold being used can depend on the specific sequence imbalance (e.g., an amplification or a deletion) being tested.

[0009] In some embodiments, a fractional contribution of a first haplotype in a first tissue type can be used to determine whether the first tissue type has a disease state. The first haplotype can have a signature specific to healthy cells or to abnormal cells. Thus, the first haplotype can be not present in healthy cells of the organism, or present in healthy cells of the organism and not present in abnormal cells that may be in the mixture. A separation value between the first fractional contribution and a reference fractional contribution can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state.

[0010] In some embodiments, the tissue origin of a copy number aberration can be determined using methylation deconvolution. A first chromosomal region can be identified as exhibiting a copy number aberration. The haplotypes in the first chromosomal region can be determined. For each of M tissue types, a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution can be determined. The tissue type with the highest separation value can be identified as the tissue of origin.

[0011] Other embodiments are directed to systems and computer readable media associated with methods described herein.

[0012] A better understanding of the nature and advantages of embodiments of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] FIG. 1 is a flowchart illustrating a method of analyzing a DNA mixture of cell-free DNA molecules to determine fractional contributions from various tissue types from methylation levels according to embodiments of the present invention.

[0014] FIG. 2 shows a schematic diagram showing several potential applications of DNA methylation deconvolution (e.g., using plasma) and its applications according to embodiments of the present invention.

[0015] FIG. 3A shows a graph of percentage contributions of different organs to plasma DNA for 15 pregnant women according to embodiments of the present invention. FIG. 3B shows a plot 350 of a correlation between the plasma DNA fractions contributed by the placenta deduced from plasma DNA methylation deconvolution and the fetal DNA fractions deduced using fetal-specific SNP alleles according to embodiments of the present invention.

[0016] FIG. 4 shows a table of percentage contributions determined from a plasma DNA tissue mapping analysis among pregnant women according to embodiments of the present invention.

[0017] FIG. 5 shows plots of percentage contributions of organs other than the placenta by plasma DNA tissue mapping and fetal DNA fractions based on fetal-specific SNP alleles according to embodiments of the present invention.

[0018] FIG. 6 shows a table of percentage contributions from plasma DNA tissue mapping analysis among the non-pregnant healthy control subjects according to embodiments of the present invention.

[0019] FIG. 7 shows a table of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the first set of markers (with high organ specificity) according to embodiments of the present invention.

[0020] FIG. 8 shows a table of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the second set of markers (with low organ specificity) according to embodiments of the present invention.

[0021] FIG. 9A is a plot showing the correlation between the estimated fetal DNA fraction (contribution from the placenta) and the fetal DNA fraction determined by counting the fetal-specific alleles in the maternal plasma samples.

[0022] FIG. 9B is a plot showing absolute difference between the estimation from methylation markers and fetal DNA fraction determined by fetal-specific alleles counting.

[0023] FIG. 10A is a graph showing placental contribution to plasma DNA deduced using markers with different selection criteria according to embodiments of the present invention. FIG. 10B is a graph showing the accuracy of plasma DNA deconvolution using markers with low variability (category i) and high variability (category ii) in the same type of tissue.

[0024] FIG. 11A shows a first scenario where the fetus has inherited the M allele from the mother and has a genotype of MN at a particular locus according to embodiments of the present invention. FIG. 11B shows a second scenario where the fetus has inherited the N allele from the mother and has the genotype of NN at a particular locus according to embodiments of the present invention.

[0025] FIG. 12A shows a determination of a maternal haplotype inherited by a fetus using methylation deconvolution according to embodiments of the present invention. FIG. 12B shows an illustration of the paternal haplotype methylation analysis according to embodiments of the present invention.

[0026] FIG. 13 is a flowchart illustrating a method 1300 of determining a portion of a fetal genome from a maternal sample using methylation deconvolution according to embodiments of the present invention.

[0027] FIG. 14 is a flowchart illustrating a method 1400 of determining a portion of a fetal genome from a maternal sample using methylation levels according to embodiments of the present invention.

[0028] FIG. 15 shows chromosomal aneuploidy detection based on haplotype deconvolution for maternal haplotypes according to embodiments of the present invention.

[0029] FIG. 16 shows chromosomal aneuploidy detection based on haplotype deconvolution for paternal haplotypes according to embodiments of the present invention.

[0030] FIG. 17 is a flowchart of a method 1700 for detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female according to embodiments of the present invention.

[0031] FIG. 18 shows an illustration of haplotype deconvolution for organ transplant monitoring according to embodiments of the present invention.

[0032] FIG. 19 is a flowchart illustrating a method of analyzing a biological sample of an organism to detect whether a first tissue type has a disease state associated with a first haplotype according to embodiments of the present invention.

[0033] FIG. 20 shows a plot of copy number aberrations detected in the plasma of an HCC patient according to embodiments of the present invention.

[0034] FIG. 21 is a flowchart illustrating a method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration according to embodiments of the present invention.

[0035] FIG. 22 shows a block diagram of an example computer system 10 usable with system and methods according to embodiments of the present invention.

TERMS

[0036] A "methylome" provides a measure of an amount of DNA methylation at a plurality of sites or loci in a genome. The methylome may correspond to all of the genome, a substantial part of the genome, or relatively small portion(s) of the genome. Examples of methylomes of interest are the methylomes of organs (e.g. methylomes of brain cells, bones, the lungs, the heart, the muscles and the kidneys, etc.) that can contribute DNA into a bodily fluid (e.g. plasma, serum, sweat, saliva, urine, genital secretions, semen, stools fluid, diarrheal fluid, cerebrospinal fluid, secretions of the gastrointestinal tract, ascitic fluid, pleural fluid, intraocular fluid, fluid from a hydrocele (e.g. of the testis), fluid from a cyst, pancreatic secretions, intestinal secretions, sputum, tears, aspiration fluids from breast and thyroid, etc.). The organs may be transplanted organs. The methylome of a fetus is another example.

[0037] A "plasma methylome" is a methylome determined from the plasma or serum of an animal (e.g., a human). The plasma methylome is an example of a cell-free methylome since

plasma and serum include cell-free DNA. The plasma methylome is also an example of a mixed methylome since it is a mixture of fetal/maternal methylome or tumor/patient methylome or DNA derived from different tissues or organs or donor/recipient methylome in the context or organ transplantation.

[0038] A "site" (also called a "genomic site") corresponds to a single site, which may be a single base position or a group of correlated base positions, e.g., a CpG site or larger group of correlated base positions. A "locus" may correspond to a region that includes multiple sites. A locus can include just one site, which would make the locus equivalent to a site in that context.

[0039] The "methylation index" for each genomic site (e.g., a CpG site) can refer to the proportion of DNA fragments (e.g., as determined from sequence reads or probes) showing methylation at the site over the total number of reads covering that site. A "read" can correspond to information (e.g., methylation status at a site) obtained from a DNA fragment. A read can be obtained using reagents (e.g. primers or probes) that preferentially hybridize to DNA fragments of a particular methylation status. Typically, such reagents are applied after treatment with a process that differentially modifies or differentially recognizes DNA molecules depending of their methylation status, e.g. bisulfite conversion, or methylation-sensitive restriction enzyme, or methylation binding proteins, or anti-methylcytosine antibodies, or single molecule sequencing techniques that recognize methylcytosines and hydroxymethylcytosines.

[0040] The "methylation density" of a region can refer to the number of reads at sites within the region showing methylation divided by the total number of reads covering the sites in the region. The sites may have specific characteristics, e.g., being CpG sites. Thus, the "CpG methylation density" of a region can refer to the number of reads showing CpG methylation divided by the total number of reads covering CpG sites in the region (e.g., a particular CpG site, CpG sites within a CpG island, or a larger region). For example, the methylation density for each 100-kb bin in the human genome can be determined from the total number of cytosines not converted after bisulfite treatment (which corresponds to methylated cytosine) at CpG sites as a proportion of all CpG sites covered by sequence reads mapped to the 100-kb region. This analysis can also be performed for other bin sizes, e.g. 500 bp, 5 kb, 10 kb, 50-kb or 1-Mb, etc. A region could be the entire genome or a chromosome or part of a chromosome (e.g. a chromosomal arm). The methylation index of a CpG site is the same as the methylation density for a region when the

region only includes that CpG site. The "proportion of methylated cytosines" can refer the number of cytosine sites, "C's", that are shown to be methylated (for example unconverted after bisulfite conversion) over the total number of analyzed cytosine residues, i.e. including cytosines outside of the CpG context, in the region. The methylation index, methylation density and proportion of methylated cytosines are examples of "methylation levels." Apart from bisulfite conversion, other processes known to those skilled in the art can be used to interrogate the methylation status of DNA molecules, including, but not limited to enzymes sensitive to the methylation status (e.g. methylation-sensitive restriction enzymes), methylation binding proteins, single molecule sequencing using a platform sensitive to the methylation status (e.g. nanopore sequencing (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) and by the Pacific Biosciences single molecule real time analysis (Flusberg et al. Nat Methods 2010; 7: 461-465)).

[0041] A "methylation profile" (also called methylation status) includes information related to DNA methylation for a region. Information related to DNA methylation can include, but not limited to, a methylation index of a CpG site, a methylation density of CpG sites in a region, a distribution of CpG sites over a contiguous region, a pattern or level of methylation for each individual CpG site within a region that contains more than one CpG site, and non-CpG methylation. A methylation profile of a substantial part of the genome can be considered equivalent to the methylome. "DNA methylation" in mammalian genomes typically refers to the addition of a methyl group to the 5' carbon of cytosine residues (i.e. 5-methylcytosines) among CpG dinucleotides. DNA methylation may occur in cytosines in other contexts, for example CHG and CHH, where H is adenine, cytosine or thymine. Cytosine methylation may also be in the form of 5-hydroxymethylcytosine. Non-cytosine methylation, such as N^6-methyladenine, has also been reported.

[0042] "Methylation-aware sequencing" refers to any sequencing method that allows one to ascertain the methylation status of a DNA molecule during a sequencing process, including, but not limited to bisulfite sequencing, or sequencing preceded by methylation-sensitive restriction enzyme digestion, immunoprecipitation using anti-methylcytosine antibody or methylation binding protein, or single molecule sequencing that allows elucidation of the methylation status.

[0043] A "tissue" corresponds to a group of cells that group together as a functional unit. More than one type of cells can be found in a single tissue. Different types of tissue may consist of

different types of cells (e.g., hepatocytes, alveolar cells or blood cells), but also may correspond to tissue from different organisms (mother vs. fetus) or to healthy cells vs. tumor cells. "Reference tissues" correspond to tissues used to determine tissue-specific methylation levels. Multiple samples of a same tissue type from different individuals may be used to determine a tissue-specific methylation level for that tissue type.

[0044] A "biological sample" refers to any sample that is taken from a subject (*e.g.*, a human, such as a pregnant woman, a person with cancer, or a person suspected of having cancer, an organ transplant recipient or a subject suspected of having a disease process involving an organ (e.g., the heart in myocardial infarction, or the brain in stroke, or the hematopoietic system in anemia) and contains one or more nucleic acid molecule(s) of interest. The biological sample can be a bodily fluid, such as blood, plasma, serum, urine, vaginal fluid, fluid from a hydrocele (e.g. of the testis), or vaginal flushing fluids, pleural fluid, ascitic fluid, cerebrospinal fluid, saliva, sweat, tears, sputum, bronchoalveolar lavage fluid, etc. Stool samples can also be used. In various embodiments, the majority of DNA in a biological sample that has been enriched for cell-free DNA (e.g., a plasma sample obtained via a centrifugation protocol) can be cell-free (as opposed to cells), e.g., greater than 50%, 60%, 70%, 80%, 90%, 95%, or 99%. The centrifugation protocol can include 3,000 g x 10 minutes, obtaining the fluid part, and recentrifuging at 30,000 g for another 10 minutes to remove residual cells.

[0045] The term "level of cancer" can refer to whether cancer exists (i.e., presence or absence), a stage of a cancer, a size of tumor, whether there is metastasis, the total tumor burden of the body, and/or other measure of a severity of a cancer (e.g. recurrence of cancer). The level of cancer could be a number or other indicia, such as symbols, alphabet letters, and colors. The level could be zero. The level of cancer also includes premalignant or precancerous conditions (states) associated with mutations or a number of mutations. The level of cancer can be used in various ways. For example, screening can check if cancer is present in someone who is not known previously to have cancer. Assessment can investigate someone who has been diagnosed with cancer to monitor the progress of cancer over time, study the effectiveness of therapies or to determine the prognosis. In one embodiment, the prognosis can be expressed as the chance of a patient dying of cancer, or the chance of the cancer progressing after a specific duration or time, or

the chance of cancer metastasizing. Detection can mean 'screening' or can mean checking if someone, with suggestive features of cancer (e.g. symptoms or other positive tests), has cancer.

[0046] The term "sequence imbalance" of a chromosomal region can refer to any significant deviation in an amount of cell-free DNA molecules from the chromosomal region relative to an expected value, if the organism was healthy. For example, a chromosomal region may exhibit an amplification or a deletion in a certain tissue, thereby resulting in a sequence imbalance for the chromosomal region in a DNA mixture containing DNA from the tissue, mixed with DNA from other tissues. As examples, the expected value can be obtained from another sample or from another chromosomal region that is assumed to be normal (e.g., an amount representative of two copies for a diploid organism). A chromosomal region can be composed of multiple disjoint subregions.

[0047] A "type" for a genomic locus (marker) corresponds to specific attributes for a locus across tissue types. The description primarily refers to type I loci and type II loci, whose properties are provided in detail below. A locus of a given type can have specific statistical variation in methylation levels across tissue types. A "category" for a genomic locus (marker) corresponds to specific variation in methylation levels for a locus across different individuals for a same tissue type. A set of genomic loci (markers) can be composed of any number of loci of various types and/or categories. Thus, a set of loci corresponds to loci selected for a particular measurement and does not connote any particular properties of the loci in the set.

[0048] A "separation value" corresponds to a difference or a ratio involving two values, e.g., two fractional contributions or two methylation levels. The separation value could be a simple difference or ratio. The separation value can include other factors, e.g., multiplicative factors. As other examples, a difference or ratio of functions of the values can be used, e.g., a difference or ratio of the natural logarithms (ln) of the two values. A separation value can include a difference and a ratio.

[0049] The term "*classification*" as used herein refers to any number(s) or other characters(s) that are associated with a particular property of a sample. For example, a "+" symbol (or the word "positive") could signify that a sample is classified as having deletions or amplifications. The classification can be binary (e.g., positive or negative) or have more levels of classification (e.g., a scale from 1 to 10 or 0 to 1). The term "*cutoff*" and "*threshold*" refer to a predetermined number

used in an operation. A threshold value may be a value above or below which a particular classification applies. Either of these terms can be used in either of these contexts.

DETAILED DESCRIPTION

[0050] Methylation differences among tissue types (e.g., fetal tissues, liver etc.) in a DNA mixture (e.g., plasma) may be used to differentiate properties of haplotypes in a specific tissue type. For example, the methylation levels of two maternal haplotypes in the plasma of a pregnant woman can be used to determine which haplotype is inherited from the mother to the fetus. As another example, the methylation levels of two haplotypes in fetal tissue can be used to detect a sequence imbalance (e.g., an aneuploidy) in the fetus. Other tissue types may also be analyzed, e.g., to detect a disease state in a specific tissue type. The tissue type from which a copy number aberration originates may also be determined.

[0051] Some embodiments can determine percentages of cell-free DNA in plasma (or other DNA mixture) from various tissue types using known methylation levels at certain genomic sites for the specific tissue types. For example, methylation levels at the genomic sites can be measured for a liver sample, and these tissue-specific methylation levels can be used to determine how much cell-free DNA in the mixture is from the liver. Methylation levels can be measured for tissue types that provide substantial contributions to the DNA mixture, so that a predominance (e.g., more than 90%, 95%, or 99%) of the cell-free DNA mixture can be accounted for. Such other samples can include, but not limited to, some or all of the following: lung, colon, small intestines, pancreas, adrenal glands, esophagus, adipose tissues, heart, and brain.

[0052] A deconvolution process can be used to determine fractional contributions (e.g., percentage) for each of the tissue types for which tissue-specific methylation levels are known. In some embodiments, a linear system of equations can be created from the known tissue-specific methylation levels and the mixture methylation levels at the specified genomic sites, and the fractional contributions that best approximate the measured mixture methylation levels can be determined (e.g., using least squares).

[0053] Once the fractional contributions are determined, the fractional contributions can be used for various purposes. For example, differences in fractional contributions of fetal tissue can be used to determine which haplotype is inherited from a parent. The alleles at one or more

heterozygous loci can be determined for each of two parental haplotypes. Cell-free DNA at the one or more heterozygous loci can be used to determine two fractional contributions: one for each haplotype. For example, cell-free DNA molecules having alleles of a first haplotype can be used to determine a first fractional contribution, and cell-free DNA molecules having alleles of a second haplotype can be used to determine a second fractional contribution. The inherited haplotype will correspond to the higher fractional contribution for fetal tissue.

[0054] Further, an inherited haplotype will have lower methylation levels due to the general hypomethylation of fetal cell-free DNA. The methylation levels for the two haplotypes can be compared, and the haplotype with lower methylation level can be identified as being the inherited haplotype.

[0055] As another example, a sequence imbalance can be detected in a target chromosomal region of a fetus. A target fractional contribution of the fetal tissue type in the mixture can be determined for a first haplotype in the target chromosomal region. Similarly, a reference fractional contribution of the fetal tissue type can be determined for a reference chromosomal region. A separation value between the two contributions can be compared to a threshold value to determine whether the fetus has a sequence imbalance (e.g., an aneuploidy).

[0056] As another example, a first haplotype can have a signature specific to healthy cells or to abnormal cells. A separation value between the fractional contribution determined for the first haplotype and a reference fractional contribution can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state. As examples, the first haplotype can be of a transplanted organ or a tumor, or only be in healthy cells and not in a transplanted organ or a tumor. The disease state may be whether the transplanted organ is being rejected, or whether a tumor is increasing in size or metastasized (e.g., after a surgery did not remove all of the tumor).

[0057] As another example, the tissue origin of a copy number aberration can be determined using methylation deconvolution. A first chromosomal region can be identified as exhibiting a copy number aberration. For each of M tissue types, a corresponding separation value between the fractional contributions of the two haplotypes in the first chromosomal region can be determined. The tissue type with the highest separation value can be identified as the tissue of origin.

[0058] Methylation deconvolution is first described, and then the selection of methylation markers and the accuracy of the methylation deconvolution are described. The use of the fractional contributions to determine part of a fetal genome is then described.

I. COMPOSITION OF DNA MIXTURE BY METHYLATION DECONVOLUTION

[0059] Different tissue types can have different levels of methylation for a genomic site. These differences can be used to determine the fractional contributions of DNA from the various tissue types in a mixture. Thus, the composition of a DNA mixture can be determined by a tissue-specific methylation pattern analysis. The examples below discuss methylation densities, but other methylation levels can be used.

A. Single Genomic Site

[0060] The principle of methylation deconvolution can be illustrated using a single methylation genomic site (methylation marker) to determine a composition of a DNA mixture from an organism. Assume that tissue A is completely methylated for the genomic site, i.e. methylation density (MD) of 100% and tissue B is completely unmethylated, i.e. MD of 0%. In this example, methylation density refers to the percentage of cytosine residues with the context of CpG dinucleotides being methylated in the region of interest.

[0061] If the DNA mixture C is composed of tissue A and tissue B and the overall methylation density of the DNA mixture C is 60%, we can deduce the proportional contribution of tissues A and B to the DNA mixture C according to the following formula:

$$MD_C = MD_A \times a + MD_B \times b,$$

where MD_A, MD_B, MD_C represent the MD of tissues A, tissue B and the DNA mixture C, respectively; and a and b are the proportional contributions of tissues A and B to the DNA mixture C. In this particular example, it is assumed that tissues A and B are the only two constituents of the DNA mixture. Therefore, a + b = 100%. Thus, it is calculated that tissues A and B contribute 60% and 40%, respectively, to the DNA mixture.

[0062] The methylation densities in tissue A and tissue B can be obtained from samples of the organism or from samples from other organisms of the same type (e.g., other humans, potentially

of a same subpopulation). If samples from other organisms are used, a statistical analysis (e.g., average, median, geometric mean) of the methylation densities of the samples of tissue A can be used to obtain the methylation density MD_A, and similarly for MD_B.

[0063] Genomic site can be chosen to have minimal inter-individual variation, for example, less than a specific absolute amount of variation or being within a lowest portion of genomic sites tested. For instance, for the lowest portion, embodiments can select only genomic sites having the lowest 10% of variation among a group of genomic sites tested. The other organisms can be taken from healthy persons, as well as those with particular physiologic conditions (e.g. pregnant women, or people with different ages or people of a particular sex), which may correspond to a particular subpopulation that includes the current organism being tested.

[0064] The other organisms of a subpopulation may also have other pathologic conditions (e.g. patients with hepatitis or diabetes, etc.). Such a subpopulation may have altered tissue-specific methylation patterns for various tissues. The methylation pattern of the tissue under such disease condition can be used for the deconvolution analysis in addition to using the methylation pattern of the normal tissue. This deconvolution analysis may be more accurate when testing an organism from such a subpopulation with those conditions. For example, a cirrhotic liver or a fibrotic kidney may have a different methylation pattern compared with a normal liver and normal kidney, respectively. Thus, if a patient with liver cirrhosis was screened for other diseases, it can be more accurate to include a cirrhotic liver as one of the candidates contributing DNA to the plasma DNA, together with the healthy tissues of other tissue types.

B. Multiple Genomic Sites

[0065] More genomic sites (e.g., 10 or more) may be used to determine the constitution of the DNA mixture when there are more potential candidate tissues. The accuracy of the estimation of the proportional composition of the DNA mixture is dependent on a number of factors including the number of genomic sites, the specificity of the genomic sites (also called "sites") to the specific tissues, and the variability of the sites across different candidate tissues and across different individuals used to determine the reference tissue-specific levels. The specificity of a site to a tissue refers to the difference in the methylation density of the genomic sites between the particular tissue and other tissue types.

[0066] The larger the difference between their methylation densities, the more specific the site to the particular tissue would be. For example, if a site is completely methylated in the liver (methylation density = 100%) and is completely unmethylated in all other tissues (methylation density = 0%), this site would be highly specific for the liver. Whereas, the variability of a site across different tissues can be reflected by, for example, but not limited to, the range or standard deviation of methylation densities of the site in different types of tissue. A larger range or higher standard deviation would allow a more precise and accurate determination of the relative contributions of the different organs to the DNA mixture mathematically. The effects of these factors on the accuracy of estimating the proportional contribution of the candidate tissues to the DNA mixture are illustrated in the later sections of this application.

[0067] Here, we use mathematical equations to illustrate the deduction of the proportional contribution of different organs to the DNA mixture. The mathematical relationship between the methylation densities of the different sites in the DNA mixture and the methylation densities of the corresponding sites in different tissues can be expressed as:

$$\overline{MD_i} = \sum_k (p_k \times MD_{ik}),$$

where $\overline{MD_i}$ represents the methylation density of the site i in the DNA mixture; p_k represents the proportional contribution of tissue k to the DNA mixture; MD_{ik} represents the methylation density of the site i in the tissue k. When the number of sites is the same or larger than the number of organs, the values of individual p_k can be determined. The tissue-specific methylation densities can be obtained from other individuals, and the sites can be chosen to have minimal inter-individual variation, as mentioned above.

[0068] Additional criteria can be included in the algorithm to improve the accuracy. For example, the aggregated contribution of all tissues can be constrained to be 100%, i.e.

$$\sum_k p_k = 100\%.$$

Furthermore, all the organs' contributions can be required to be non-negative:

$$p_k \geq 0, \forall k$$

[0069] Due to biological variations, the observed overall methylation pattern may not be completely identical to the methylation pattern deduced from the methylation of the tissues. In

such a circumstance, mathematical analysis would be required to determine the most likely proportional contribution of the individual tissues. In this regard, the difference between the observed methylation pattern in the DNA and the deduced methylation pattern from the tissues is denoted by W.

$$W = O - \sum_k (p_k \times M_k)$$

where O is the observed methylation pattern for the DNA mixture and M_k is the methylation pattern of the individual tissue k. p_k is the proportional contribution of tissue k to the DNA mixture. The most likely value of each p_k can be determined by minimizing W, which is the difference between the observed and deduced methylation patterns. This equation can be resolved using mathematical algorithms, for example by, but not limited to, using quadratic programming, linear/non-linear regression, expectation-maximization (EM) algorithm, maximum likelihood algorithm, maximum a posteriori estimation, and the least squares method.

C. Method of Methylation Deconvolution

[0070] As described above, a biological sample including a mixture of cell-free DNA molecules from an organism can be analyzed to determine the composition of the mixture, specifically the contributions from different tissue types. For example, the percentage contribution of the cell-free DNA molecules from the liver can be determined. These measurements of the percentage contributions in the biological sample can be used to make other measurements of the biological sample, e.g., identifications of where a tumor is located, as is described in later sections.

[0071] FIG. 1 is a flowchart illustrating a method 100 of analyzing a DNA mixture of cell-free DNA molecules to determine fractional contributions from various tissue types from methylation levels according to embodiments of the present invention. A biological sample includes a mixture of cell-free DNA molecules from M tissues types. The biological sample can be any one of various examples, e.g., as mentioned herein. The number M of tissue types is greater than two. In various embodiments, M can be 3, 7, 10, 20, or more, or any number in between. Method 100 can be performed at least partially using a computer system, as can other methods described herein.

[0072] At block 110, N genomic sites are identified for analysis. The N genomic sites can have various attributes, e.g., as described in more detail in section II, which describes type I and type II

genomic sites. As examples, the N genomic sites can include type I or type II sites only, or a combination of both. The genomic sites can be identified based on analyses of one or more other samples, e.g., based on data obtained from databases about methylation levels measured in various individuals.

[0073] Specific genomic sites can be selected to provide a desired level of accuracy. For example, loci that have at least a threshold variability can be used, as opposed to just using loci that are specific to one tissue type. A first set (e.g., 10) of the genomic sites can be selected such that each have a coefficient of variation of methylation levels of at least 0.15 across M tissue types and such that each has a difference between a maximum and a minimum methylation level for the M tissue types that exceeds 0.1 for one or more other samples. This first set of genomic sites may not have a specific methylation signature for a specific tissue type, e.g., only or predominantly methylated in the specific tissue type. Such a first set is referred to as type II sites. These genomic sites can be used in combination with genomic sites that do have a specific signature, which are referred to as type I sites.

[0074] Using the type II sites can ensure that the full space of methylation levels across the tissue types is spanned by the genomic sites, thereby providing increased accuracy over the type I sites. Just using more type I sites provides redundant basis vectors for the methylation space (i.e., more genomic sites that have the same pattern as other sites), while adding other genomic sites whose methylation levels have various values across different tissues adds new basis vectors for discriminating fractional contributions via the linear system of equations.

[0075] In some embodiments, at least 10 of the N genomic sites each have a coefficient of variation of methylation levels of at least 0.15 across the M tissue types. The at least 10 genomic sites can also each have a difference between a maximum and a minimum methylation level for the M tissue types that exceeds 0.1. These methylation properties of the genomic loci can be measured for one sample or a set of samples. The set of samples may be for a subpopulation of organisms that includes the instant organism being tested, e.g., a subpopulation having a particular trait that is shared with the instant organism. These other samples can be referred to as reference tissues, and different reference tissues may be used from different samples.

[0076] At block 120, N tissue-specific methylation levels are obtained at the N genomic sites for each of M tissue types. N is greater than or equal to M, so that the tissue-specific methylation

levels can be used in the deconvolution to determine the fractional percentages. The tissue-specific methylation levels can form a matrix A of dimensions N by M. Each column of the matrix A can correspond to a methylation pattern for a particular tissue type, where the pattern is of methylation levels at the N genomic sites.

[0077] In various embodiments, the tissue-specific methylation patterns can be retrieved from public database(s) or previous studies. In examples herein, the methylation data for neutrophils and B cells were downloaded from the Gene Expression Omnibus (Hodges et al. Mol Cell 2011;44:17-28). Methylation patterns for other tissues (hippocampus, liver, lung, pancreas, atrium, colon (including its various parts, e.g. sigmoid colon, transverse colon, ascending colon, descending colon), adrenal gland, esophagus, small intestines and CD4 T cell) were downloaded from the RoadMap Epigenomics project (Ziller et al. Nature 2013;500:477-81). The methylation patterns for the buffy coat, placenta, tumor and plasma data were from published reports (Lun et al. Clin Chem. 2013;59:1583-94; Chan et al. Proc Natl Acad Sci U S A. 2013;110:18761-8). These tissue-specific methylation patterns can be used to identify the N genomic sites to be used in the deconvolution analysis.

[0078] At block 130, the biological sample including a mixture of cell-free DNA molecules from the M tissues types is received. The biological sample may be obtained from the patient organism in a variety of ways. The manner of obtaining such samples may be non-invasive or invasive. Examples of non-invasively obtained samples include certain types of fluids (e.g. plasma or serum or urine) or stools. For instance, plasma includes cell-free DNA molecules from many organ tissues, and is thus useful for analyzing many organs via one sample.

[0079] At block 140, cell-free DNA molecules from the biological sample are analyzed to identify their locations in a reference genome corresponding to the organism. For example, the cell-free DNA molecules can be sequenced to obtain sequence reads, and the sequence reads can be mapped (aligned) to the reference genome. If the organism was a human, then the reference genome would be a reference human genome, potentially from a particular subpopulation. As another example, the cell-free DNA molecules can be analyzed with different probes (e.g., following PCR or other amplification), where each probe corresponds to a genomic location, which may cover a heterozygous and one or more CpG sites, as is described below.

[0080] A statistically significant number of cell-free DNA molecules can be analyzed so as to provide an accurate deconvolution for determining the fractional contributions from the M tissue types. In some embodiments, at least 1,000 cell-free DNA molecules are analyzed. In other embodiments, at least 10,000 or 50,000 or 100,000 or 500,000 or 1,000,000 or 5,000,000 cell-free DNA molecules or more can be analyzed. The total number of molecules to analyze can depend on M and N, and the desired precision (accuracy). In various examples, the total number of cell-free DNA analyzes can be less than 500,000, one million, two million, five million, ten millsion, 20 million, or 50 million.

[0081] At block 150, N mixture methylation levels are measured at the N genomic sites using cell-free DNA molecules that are each located at any one of N genomic sites of the reference genome. A DNA molecule can be identified as located at a genomic site or a locus by one or more bases of the DNA molecule corresponding to one or more base positions of the genomic site or locus. Thus, the sequence of the DNA molecule would cover one or more base positions of the genomic site or locus. This information can be determined based on the locations determined in block 140. Such an identification of a DNA molecule located at a site of a locus can be used for any similar block of methods described herein.

[0082] The N mixture methylation levels refer to methylation levels in the mixture of the biological sample. As an example, if a cell-free DNA molecule from the mixture is located at one of the N genomic sites, then a methylation index for that molecule at the site can be included in an overall methylation density for that site. The N mixture methylation levels can form a methylation vector b of length N, where b corresponds to observed values from which the fractional contribution of each corresponding tissue type can be determined.

[0083] In one embodiment, the methylation levels for the genomic sites in the DNA mixture can be determined using whole genome bisulfite sequencing. In other embodiments, the methylation levels for the CpG sites can be determined using methylation microarray analysis, such as the Illumina HumanMethylation450 system, or by using methylation immunoprecipitation (e.g. using an anti-methylcytosine antibody) or treatment with a methylation-binding protein followed by microarray analysis or DNA sequencing, or by using methylation-sensitive restriction enzyme treatment followed by microarray or DNA sequencing, or by using methylation aware sequencing e.g. using a single molecule sequencing method (e.g. by a nanopore sequencing (Schreiber et al.

Proc Natl Acad Sci 2013; 110: 18910-18915) or by the Pacific Biosciences single molecule real time analysis (Flusberg et al. Nat Methods 2010; 7: 461-465)). Tissue-specific methylation levels can be measured in a same way. In yet other embodiments, other methods, for example, but not limited to targeted bisulfite sequencing, methylation-specific PCR, non-bisulfite based methylation-aware sequencing (e.g. by single molecule sequencing platforms (Powers et al. Efficient and accurate whole genome assembly and methylome profiling of E. coli. BMC Genomics. 2013;14:675)) can be used for the analysis of the methylation level of the plasma DNA for plasma DNA methylation deconvolution analysis.

[0084] At block 160, M values of a composition vector are determined. Each M value corresponds to a fractional contribution of a particular tissue type of the M tissue types to the DNA mixture. The M values of the composition vector can be solved to provide the N mixture methylation levels (e.g., methylation vector b) given the matrix A composed of NxM tissue-specific methylation levels (i.e. N tissue-specific methylation levels for each of the M tissue types). The M fractional contributions can correspond to a vector x that is determined by solving Ax=b. When N is greater than M, the solution can involve a minimization of errors, e.g., using least-squares.

[0085] At block 170, the composition vector is used determine an amount of each of the M tissue types in the mixture. The M values of the composition vector may be taken directly as the fractional contributions of the M tissue types. In some implementations, the M values can be converted to percentages. Error terms can be used to shift the M values to higher or lower values.

D. Applications

[0086] As mentioned above, the fractional contributions can be used in further measurements of the biological sample and other determinations, e.g., whether a particular chromosomal region has a sequence imbalance, whether a particular tissue type is diseased, and to determine which haplotype of two parental haplotypes is inherited by the fetus of a pregnant female from which the sample was obtained.

[0087] FIG. 2 shows a schematic diagram showing several potential applications of DNA methylation deconvolution (e.g., using plasma) according to embodiments of the present invention. In FIG. 2, a biological sample 205 is subjected to genome-wide bisulfite sequencing at 210. At

230, plasma DNA tissue mapping uses tissue-specific methylation profiles 220 to determine tissue contribution percentages. Example tissue-specific methylation profiles are shown as liver, blood cells, adipose tissues, lungs, small intestines, and colon. The contribution percentages can be determined as described above and elsewhere, e.g., solving Ax=b. Examples of applications include prenatal testing 241, cancer detection and monitoring 242, organ transplant monitoring, and organ damage assessment 244.

[0088] A list of methylation markers (genomic sites) that are useful for determining the contributions of different organs to the plasma DNA can be identified by comparing the methylation profiles (FIG. 2) of different tissues, including the liver, lungs, esophagus, heart, pancreas, sigmoid colon, small intestines, adipose tissues, adrenal glands, colon, T cells, B cells, neutrophils, brain and placenta. In various examples, whole genome bisulfite sequencing data for the liver, lungs, esophagus, heart, pancreas, colon, small intestines, adipose tissues, adrenal glands, brain and T cells were retrieved from the Human Epigenome Atlas from the Baylor College of Medicine (www.genboree.org/epigenomeatlas/index.rhtml). The bisulfite sequencing data for B cells and neutrophils were from the publication by Hodges et al. (Hodges et al; Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol Cell 2011; 44: 17-28). The bisulfite sequencing data for the placenta were from Lun et al (Lun et al. Clin Chem 2013; 59:1583-94). In other embodiments, markers can be identified from datasets generated using microarray analyses, e.g. using the Illumina Infinium HumanMethylation450 BeadChip Array.

II. SELECTION OF METHYLATION MARKERS

[0089] Above, we have described the principle of using methylation analysis to determine the composition of a DNA mixture. In particular, the percentage contribution of different organs (or tissues) to the plasma DNA can be determined using methylation analysis. In this section, we further describe the method for the selection of methylation markers and clinical applications of this technology.

[0090] The results of determining the composition of the DNA mixture by methylation analysis are affected by the methylation markers used for the deconvolution of the composition of the DNA

mixture. Thus, the selection of appropriate genomic methylation markers can be important for the accurate determination of the constitution of the DNA mixture.

A. Criteria for a Methylation Marker for Deconvolution

[0091] For marker selection, the following three attributes may be considered. (i) It is desirable for a methylation marker to have a low variability in the methylation level measured in the same tissue type across different individuals. As the determination of the composition of the DNA mixture is dependent on the recognition of the tissue-specific methylation patterns, the low variability in methylation level in the same tissue type across different individuals would be useful for accurate identification of the tissue-specific patterns in the DNA mixture. In embodiments where the tissue-specific methylation levels are obtained from samples of other organisms (e.g., from a database), the low variability means that the methylation levels from the other samples are similar to the tissue-specific methylation levels for the current organism being tested.

[0092] (ii) It is desirable for a methylation marker to have a high variability in methylation levels across different tissues. For a particular marker, a higher difference in the methylation levels across different tissues can provide a more precise determination of the contribution of different tissues to the DNA mixture. In particular, an improvement in precision can be obtained by using one set of markers having attribute (ii) and another set of markers having attribute (iii).

[0093] (iii) It is desirable for a methylation marker to have a particularly different methylation level in a particular tissue when compared with those from most or all of the other tissues. In contrast to point (ii) above, a marker can have low variability in the methylation level of most tissues but its methylation level in one particular tissue is different from most of the other tissues. This marker would be particularly useful for the determination of the contribution of the tissue that has a different methylation level from other tissues.

B. Example

[0094] A principle of marker selection is illustrated in the following hypothetical examples in table 1.

	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5	Marker 6
Liver 1	20%	69%	9%	9%	10%	90%

Liver 2	50%	70%	10%	10%	10%	90%
Liver 3	90%	71%	11%	11%	10%	90%
Heart	20%	20%	30%	13%	12%	12%
Lung	30%	30%	60%	17%	14%	84%
Colon	40%	40%	90%	20%	80%	80%

Table 1. Methylation densities in different tissues for 6 hypothetical methylation markers.

[0095] In this hypothetical example, marker 2 has lower variability in methylation density in the liver from three individuals when compared with marker 1. Therefore, marker 2 is superior to marker 1 as a signature for determining the contribution of the liver in a DNA mixture.

[0096] Compared with marker 4, marker 3 has a higher variability in methylation density across different tissue types. The same level of change in the estimated contribution from the different tissues would provide a bigger change in the deduced methylation density of the DNA mixture for marker 3 than for marker 4 according to the mathematical relationship discussed above. Therefore, the estimation of the contribution of each tissue can be more precise with marker 3.

[0097] Marker 5 has a low variability in methylation density across the liver, heart and lung. Their methylation densities vary from 10% to 14%. However, the methylation density of colon is 80%. This marker would be particularly useful for determining the contribution of the colon in the DNA mixture. Similarly, the heart is hypomethylated compared with the other tissues for marker 6. Therefore, the contribution of the heart can be accurately determined by marker 6. Thus, the combination of markers 5 and 6 would be able to accurately determine the contributions of the colon and the heart. The addition of markers 2 and 3 would then be sufficient to deduce the contribution of each of the four organs, including the liver, heart, lung and colon.

C. Different Types of Markers

[0098] A methylation marker may not necessarily need to have all of the above three attributes. A type I methylation marker would typically have attribute (iii) above. A number of such markers may also have attribute (i). On the other hand, a type II methylation marker would typically have attribute (ii) above. A number of such markers may also have attribute (i). It is also possible that a particular marker may have all three attributes.

[0099] In some embodiments, markers are broadly divided into two types (type I and type II). Type I markers have tissue specificity. The methylation level of these markers for a particular group of one or more tissues is different from most of the other tissues. For example, a particular tissue can have a significant methylation level compared with the methylation level of all the other tissues. In another example, two tissues (e.g., tissue A and tissue B) have similar methylation levels, but the methylation levels of tissues A and B are significantly different from those of the remaining tissues.

[0100] Type II markers have a high inter-tissue methylation variability. The methylation levels of these markers are highly variable across different tissues. A single marker in this category may not be sufficient to determine the contribution of a particular tissue to the DNA mixture. However, a combination of type II markers, or in combination with one or more type I markers can be used collectively to deduce the contribution of individual tissues. Under the above definition, a particular marker can be a type I marker only, a type II marker only, or be simultaneously both a type I and type II marker.

1. Type I Markers

[0101] In one embodiment, a type I marker can be identified by comparing the methylation density of the marker with the mean and standard deviation (SD) of methylation densities of this particular marker for all candidate tissues. In one implementation, a marker is identified if its methylation density in one tissue is different from the mean of all the tissues by 3 standard deviations (SD).

[0102] The methylation profiles of 14 tissues obtained from sources mentioned above were studied to select markers. In one analysis, a total of 1,013 type I markers were identified (markers labeled type I in Table S1 of Appendix A of U.S. Provisional Application No. 62/158,466) using the above criteria. In other embodiments, other cutoffs between the particular tissues and the mean methylation densities can be used, for example, but not limited to 1.5 SD, 2 SD, 2.5 SD, 3.5 SD and 4 SD. In yet another embodiment, a type I marker can identified through the comparison of the methylation density of the particular tissue to the median methylation density of all tissues.

[0103] In other embodiments, the type I markers can be obtained when more than one tissue (for example, but not limited to two, three, four or five tissues) show significantly different

methylation densities than the mean methylation density of all the candidate tissues. In one implementation, a cutoff methylation density can be calculated from the mean and SD of the methylation densities of all the candidate tissues. For illustration purpose, the cutoff can be defined as 3 SD higher or lower than the mean methylation densities. A marker is selected when the methylation densities of more than one (for example, but not limited to two, three, four, five, or more than five) tissues are more than 3 SD higher than the mean methylation density or more than 3 SD lower than the mean methylation density of the tissues.

2. Type II Markers

[0104] For identification of type II markers, the mean and SD of methylation densities across all 14 candidate tissues were calculated and the ratio of SD to the mean was denoted as the coefficient of variation (CV). In this illustrative example, we used a cutoff of >0.25 for the CV to identify the qualified type II markers, as well as the difference between the maximum and minimum methylation densities for the group of tissues exceeding 0.2. Using these criteria, 5820 type II markers were identified (markers labeled type II in Table S1 of Appendix A). In other embodiments, other cutoffs for the CV, for example but not limited to 0.15, 0.2, 0.3 and 0.4, can be used. In yet other embodiments, other cutoffs for the difference between the maximum and minimum methylation densities, for example, but not limited to 0.1, 0.15, 0.25, 0.3, 0.35, 0.4, 0.45 and 0.5, can be used.

[0105] In other embodiments, the average values across multiple samples of the same tissue type can be used to measure a variation of the methylation levels across different tissues. For example, 10 methylation levels of a same genomic site from 10 samples can be averaged to obtain a single methylation level for the genomic site. A similar process can be performed to determine average methylation levels for other tissue types for the genomic site. The average values across tissue types can then be used for determining whether the genomic site has significant variation across tissue types. Other statistical values can be used besides an average, e.g., a median or a geometric mean. Such statistical values can be used to identify type I and/or type II markers.

[0106] The different samples of a same tissue type (e.g., from different individuals) can be used to determine a variation of methylation levels across the different samples. Thus, if there are multiple samples of the same tissue type, embodiments can further measure the variation of a

particular marker amongst such samples of the same tissue type. A marker with a low variation across samples would be a more reliable marker than one with a high variation. Further details of markers and deconvolution can be found in commonly-owned U.S. Patent Publication 2016/0017419, entitled "Methylation Pattern Analysis Of Tissues In A DNA Mixture," by Chiu et al., and PCT Publication WO2014/043763 entitled "Non-Invasive Determination Of Methylome Of Fetus Or Tumor From Plasma."

D. Different Categories of Markers

[0107] A "category" for a genomic locus (methylation marker) corresponds to specific variation in methylation levels for a locus across different individuals for a same tissue type. Different categories can have different ranges of variation among a particular tissue type across individuals. A first category of methylation markers might have a difference of 10% in the methylation levels or lower among the individuals tested. A second category of methylation markers might have a difference of more than 10% in the methylation levels among the individuals tested. The use of methylation markers with low inter-individual variations (first category markers) would potentially improve the accuracy of determining the contribution of the particular organ in the DNA mixture.

E. Identification of Potential Methylation Markers

[0108] In some embodiments, potential methylation markers were identified in the following manner. Such potential methylation markers can then be subjected to the above criteria to identify type I and type II markers. In other embodiments, an identification of type I or type II is not needed. And, other embodiments may use other techniques to identify potential methylation markers.

[0109] In some embodiments, all CpG islands (CGIs) and CpG shores on autosomes were considered for potential methylation markers. CGIs and CpG shores on sex chromosomes were not used so as to minimize variation in methylation levels related to the sex-associated chromosome dosage difference in the source data. CGIs were downloaded from the University of California, Santa Cruz (UCSC) database (genome.ucsc.edu/, 27,048 CpG islands for the human genome) (Kent et al., The human genome browser at UCSC, Genome Res. 2002;12(6):996-1006) and CpG shores were defined as 2 kb flanking windows of the CpG islands (Irizarry et al. The

human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009; 41(2):178 - 186). Then, the CpG islands and shores were subdivided into non-overlapping 500 bp units and each unit was considered as a potential methylation marker.

[0110] The methylation densities (i.e., the percentage of CpGs being methylated within a 500 bp unit) of all the potential loci were compared between the 14 tissue types. As previously reported (Lun et al. Clin Chem. 2013; 59: 1583-94), the placenta was found to be globally hypomethylated when compared with the remaining tissues. Thus, the methylation profile of the placenta was not included at the marker identification phase. Using the methylation profiles of the remaining 13 tissue types, the two types of methylation markers were identified. For example, type I markers can refer to any genomic sites with methylation densities that are 3 SD below or above in one tissue when compared with the mean level of the 13 tissue types. Type II markers can considered highly variable when (A) the methylation density of the most hypermethylated tissue is at least 20% higher than that of the most hypomethylated one; and (B) the SD of the methylation densities across the 13 tissue types when divided by the mean methylation density (i.e. the coefficient of variation) of the group is at least 0.25. Lastly, in order to reduce the number of potentially redundant markers, only one marker may be selected in one contiguous block of two CpG shores flanking one CpG island.

F. Selection based on Application

[0111] The set of methylation markers chosen for particular applications can be varied depending on the parameters of the desired applications. For example, for applications focusing on haplotype or allele analysis, useful markers would be those located on the same cell-free DNA molecules as one of the heterozygous alleles. As cell-free DNA molecules (e.g., plasma DNA) are usually less than 200 bp, useful markers can be CpG sites within 200 bp of a heterozygous locus (e.g., a SNP). As another example, for applications in which the release of DNA from a particular tissue into plasma is of special significance, one can select a preferentially larger number of methylation markers that are differentially methylated in this tissue type (e.g. type I marker) when compared with the others in the marker set.

[0112] The number and choice of methylation markers in the deconvolution analysis can be varied according to the intended use. If the fractional contribution of the liver is of particular interest, e.g. in a patient who has received a liver transplant, more type I liver specific markers can be used in the deconvolution analysis to increase the precision of the quantification of the contribution of the transplanted liver to the plasma DNA.

III. COMPOSITION ACCURACY

[0113] As described above, embodiments can identify the tissue contributors of plasma DNA. In various examples, genome-wide bisulfite sequencing of plasma DNA was performed and analyzed with reference to methylation profiles of different tissues. Using quadratic programming as an example, the plasma DNA sequencing data were deconvoluted into proportional contributions from different tissues. Embodiments were tested for pregnant women, patients with hepatocellular, lung and colorectal carcinoma, and subjects following bone marrow and liver transplantation.

[0114] In most subjects, white blood cells were the predominant contributors to the circulating DNA pool. The placental contributions in pregnant women correlated with the proportional contributions as revealed by fetal-specific genetic markers. The graft-derived contributions to the plasma in the transplant recipients correlated with those determined using donor-specific genetic markers. Patients with hepatocellular, lung or colorectal cancer showed elevated plasma DNA contributions from the organ with the tumor. The liver contributions in hepatocellular carcinoma patients also correlated with measurements made using tumor-associated copy number aberrations.

[0115] In cancer patients and in pregnant women exhibiting copy number aberrations in plasma, methylation deconvolution pinpointed the tissue type responsible for the aberrations. In a pregnant woman diagnosed as having follicular lymphoma during pregnancy, methylation deconvolution indicated a grossly elevated contribution from B-cells into the plasma DNA pool and localized B-cells (instead of the placenta) as the origin of the copy number aberrations observed in plasma. Accordingly, embodiments may serve as a powerful tool for assessing a wide range of physiological and pathological conditions based on the identification of perturbed proportional contributions of different tissues into plasma.

A. Contribution of Different Types of Blood Cells

[0116] As an example of the methylation deconvolution, we determined the contribution of different tissues and cell types to the circulating DNA. Two blood samples were collected from two patients suffering from systemic lupus erythematosus (SLE). After collection, the venous blood samples were centrifuged at 1,500 g for 10 minutes. After centrifugation, the blood cells and the plasma were separated. DNA was then extracted from the blood cells. The DNA was bisulfite converted and sequenced using one lane of a flow cell in a HiSeq2000 sequencer. Two blood cell samples were analyzed using the cell-type-specific methylation pattern analysis. The methylation patterns of neutrophils, lymphocytes, the esophagus, colon, pancreas, liver, lung, heart, adrenal glands and hippocampus were included as potential candidates of the blood cell DNA. 609 methylation markers were selected for the analysis. The whole blood samples of the two subjects were also sent for cell counting to determine the fractional composition of the neutrophils and lymphocytes of the blood cells.

	Blood sample 1		Blood sample 2	
	Cell type-specific methylation pattern analysis	Blood cell counting	Cell type-specific methylation pattern analysis	Blood cell counting
Neutrophils	90.5%	93.6%	89.4%	89.9%
Lymphocytes	9.5%	6.4%	10.6%	10.1%
Esophagus	0%	-	0%	-
Colon	0%	-	2%	-
Pancreas	0%	-	0%	-
Liver	0%	-	1%	-
Lung	1%	-	1%	-
Heart	0%	-	3%	-
Adrenal gland	0%	-	0%	-
Hippocampus	0%	-	0%	-

Table 2. Blood tissue contributions by deconvolution pattern analysis and cell counting

[0117] For methylation pattern analysis, neutrophils and lymphocytes were determined as the major components constituting the blood cell DNA. The relative proportion of the contribution of neutrophils and lymphocytes resemble their relative abundance in the blood samples according to the cell counting analysis.

B. Pregnant Women

[0118] The contributions of different tissues, including the liver, lung, pancreas, colon, hippocampus, small intestines, blood cells, heart, adrenal gland, esophagus and placenta, were analyzed using methylation analysis of the plasma DNA of pregnant women. As the placental genotype is in general identical to the fetus's genotype but different from the pregnant woman's genotype, the precise contribution of the placenta to the maternal plasma can be accurately determined by counting the number of fetal specific-alleles in the sample.

1. Composition and Correlation to Fetal DNA Percentage

[0119] Genome-wide bisulfite sequencing of plasma DNA was performed for 15 pregnant women, five from each of first, second and third trimesters. Methylation deconvolution was performed and the percentage contributions from different tissues were deduced. The contributions of different organs were determined based on the methylation levels (such as methylation densities) of all the type I and type II markers in table S1 using quadratic programming analysis.

[0120] FIG. 3A shows a graph 300 of percentage contributions of different organs to plasma DNA for 15 pregnant women according to embodiments of the present invention. Each bar corresponds to the results of one sample. The different colors represent the contributions of different organs into plasma. These results show that the white blood cells (i.e. neutrophils and lymphocytes) are the most important contributors to the plasma DNA pool. This observation is consistent with those previously obtained following bone marrow transplantation (Lui YY et al. Clin Chem 2002; 48: 421-7).

[0121] FIG. 4 shows a table 400 of percentage contributions determined from a plasma DNA tissue mapping analysis among pregnant women according to embodiments of the present invention. These results also show that the placenta is another key contributor of the plasma DNA in pregnant women, with fractional concentrations from 9.9% to 38.4%.

[0122] We also measured the placental contributions using paternally-inherited fetal single nucleotide polymorphism (SNP) alleles that were not possessed by the pregnant women as previously described (31). To analyze the fetal-specific SNP alleles, the genotypes of the fetuses were determined by analyzing the chorionic villus samples or the placenta. The genotypes of the

pregnant women were determined by analyzing the blood cells. The SNP-based results show the independent validation of the methylation deconvolution results.

[0123] FIG. 3B shows a plot 350 of a correlation between the plasma DNA fractions contributed by the placenta deduced from plasma DNA methylation deconvolution and the fetal DNA fractions deduced using fetal-specific SNP alleles according to embodiments of the present invention. Plot 350 shows that the placental contributions determined by methylation deconvolution have a strong correlation with the fetal DNA fractions measured using SNPs ($r = 0.99$, $p < 0.001$, Pearson correlation). Accordingly, a good positive correlation is observed between the values of the two parameters, suggesting that the plasma DNA methylation deconvolution accurately determines the contribution of the placenta to the maternal plasma samples.

[0124] FIG. 5 shows plots of percentage contributions of organs other than the placenta by plasma DNA tissue mapping and fetal DNA fractions based on fetal-specific SNP alleles according to embodiments of the present invention. The X-axis represents the fetal DNA fractions estimated by SNP-based analysis and the Y-axis represents the percentage contribution deduced by plasma tissue DNA mapping analysis. Plasma DNA contributions of the neutrophils show a reverse correlation. This is likely due to the fact that neutrophils are a major contributor to the plasma DNA pool and hence, as the placental contribution increases, the relative contribution from the neutrophils would by necessity decrease. The methylation deconvolution results of the remaining tissues show no correlation with the fetal DNA fraction.

[0125] FIG. 6 shows a table 600 of percentage contributions from plasma DNA tissue mapping analysis among the non-pregnant healthy control subjects according to embodiments of the present invention. When the process was applied to plasma of non-pregnant healthy controls, placental contribution was absent in most samples (median: 0%; interquartile range: 0% to 0.3%).

2. Comparison of Selected Markers vs. Random Markers

[0126] The accuracy of the percentage contributions was tested with select markers relative to random markers. Different composition calculations were done for different sets of markers. One set was chosen based on criteria mention above, and the other was a random set. The results show

that it is important to judicially choose the methylation markers (genomic loci) use, in order to obtain accurate results.

[0127] Eleven pregnant women and four healthy non-pregnant subjects were recruited for this analysis. Their plasma DNA was bisulfite converted and sequenced using the Illumina HiSeq2000 sequencer. Each plasma sample was sequenced with one lane of a sequencing flow cell. The sequence reads were then analyzed using a bioinformatic program, Methy-Pipe (Jiang P. PLoS One 2014; 9: e100360). This program can align the bisulfite-converted sequence reads to the reference genome and determine the methylation status of each CpG site on each sequenced fragment.

[0128] The first set of markers have high specificity for identifying the different tissues in the plasma DNA. For each tissue type, markers that have the biggest difference in methylation density compared with the other tissues were selected. The markers were determined from genomic regions containing at least one CpG dinucleotide. In this example, CpG islands (CGIs) were used as potential markers, having a high frequency of CpG sites in a particular stretch of DNA. CGIs in this particular example are downloaded from the University of California, Santa Cruz (UCSC) database: (genome.ucsc.edu). In total, we obtained 27,048 CpG islands from the human genome. The median size of a CpG island is 565 bp (range: 200 bp to 45 kb). 90% of the islands are less than 1.5 kb.

[0129] For each methylation marker, the difference in methylation density between the tissue-of-interest and the other tissues was determined. The difference is then expressed as the number of standard deviations (SDs) across the other tissues. For the tissue-of-interest, all the markers were ranked according to this difference in methylation density. The 20 markers with the biggest difference above (10 markers) and below (10 markers) the mean methylation densities of the other tissues were selected. The number of markers can vary, for example, but not limited to 5, 15, 20, 30, 40, 50, 100 and 200.

[0130] In addition, markers with a high variability across all the different tissues were also selected. In this example, markers with >50% difference between the tissues with the highest and lowest methylation densities were selected. In other applications, other values, for example, but not limited to 20%, 30%, 40%, 60%, 70% and 80%, can be used. Furthermore, the variability of methylation densities across different tissues was also calculated based on the mean and SD. In

this example, a marker was also selected if the value of SD is more than two times the mean. In other applications, other cutoff values, for example, but not limited to 1, 1.5, 2.5 and 3, can also be used. Based on these selection criteria, 344 methylation markers were selected for the first set.

[0131] For the second set, 341 markers were randomly selected from the 27,048 CGIs discussed above. All the CGIs were first numbered from 1 to 27,048. Then a random number (between 1 and 27,048) was generated by a computer for marker selection. This process was then repeated until a total of 341 markers were selected. If a random number generated had been used, another one would be generated. This set of markers is expected to have a much lower specificity in identifying the tissue-specific methylation patterns. Thus, the accuracy of determining the composition of the plasma DNA is expected to be reduced.

[0132] FIG. 7 shows a table 700 of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the first set of markers (with high organ specificity) according to embodiments of the present invention. The fetal DNA fractions were determined by counting fetal-specific alleles and are shown in the bottom row. In each of the four non-pregnant control subjects, the contribution of the placenta to the plasma was determined to be close to 0%. This indicates the specificity of this approach.

[0133] FIG. 8 shows a table 800 of the estimated contributions of different organs to the plasma DNA for 11 pregnant women and 4 non-pregnant healthy subjects using the second set of markers (with low organ specificity) according to embodiments of the present invention. The fetal DNA fractions determined by counting fetal-specific alleles are shown in the bottom row. Using these less specific markers, a relatively non-concordant percentage of contribution from the placenta was observed, and considerable contributions from the placenta were observed in the four non-pregnant control subjects. This indicates that the tissue specificity of the markers is important in this approach.

[0134] FIG. 9A is a plot 900 showing the correlation between the estimated fetal DNA fraction (contribution from the placenta) and the fetal DNA fraction determined by counting the fetal-specific alleles in the maternal plasma samples. The results from the two techniques have good correlation using the first set of methylation markers. However, using the second set of methylation markers, the estimation by using the methylation analysis showed significant deviation from the true values determined using fetal-specific alleles counting.

[0135] FIG. 9B is a plot 950 showing absolute difference between the estimation from methylation markers and fetal DNA fraction determined by fetal-specific alleles counting. The median error of the estimation using methylation analysis were 4% and 8% using the first set of markers and the second set of markers, respectively.

C. Effect of Different Criteria

[0136] As described above, various criteria can be used to identify markers of different types. For example, a type I marker can be identified by a methylation level in a particular tissue that is different from the mean methylation level for all tissues, e.g., at least by a specific threshold, such as 3 SD. And, for type II markers, criteria of a certain variation and maximum difference are used. Sections below show accuracy of different criteria for identifying markers.

1. Performance of Markers with Less Stringent Criteria

[0137] We compared the performance of methylation deconvolution analysis using markers with different variability across different tissues. The placental contributions to plasma DNA were determined for 15 pregnant women based on two sets of markers with different selection criteria. Both sets of markers include all the type I markers as described in previous sections. However, the selection criteria of type II markers are different for the two sets of markers.

[0138] Set I markers include all the 5,820 type II markers fulfilling the criteria of having methylation density CV >0.25 and the difference between the maximum and minimum methylation densities for the groups of tissues exceeding 0.2. For Set II markers, the CV requirement was >0.15 and the difference between the maximum and minimum methylation densities for the groups of tissues exceeded 0.1. There were 8,511 type II markers in this set of markers.

[0139] FIG. 10A is a graph 1000 showing placental contribution to plasma DNA deduced using markers with different selection criteria according to embodiments of the present invention. The vertical axis corresponds to placental contribution deduced using the set II markers. The horizontal axis corresponds to placental contribution deduced using the set I markers. There was a good correlation between the placental contribution results based on the two sets of markers with different selection criteria (r=0.99, Pearson correlation). Accordingly, good accuracy can be

obtained using the requirements of CV >0.15 and of the difference between the maximum and minimum methylation densities for the groups of tissues exceeding 0.1.

2. Effect of Methylation Level Variation within Same Type of Tissue

[0140] To investigate if the variation in methylation level of markers between the same type of tissues (e.g. from different individuals) would affect the performance of deconvolution analysis, we analyzed placental tissues from two pregnant cases. Two categories of methylation markers were identified. Specifically, the two categories were identified based on their similarity in methylation levels in two placental tissues. Markers of category i have a methylation density of 10% or lower. Markers of category ii have high variability between the two placental tissues (difference in methylation density of more than 10%).

[0141] FIG. 10B is a graph 1050 showing the accuracy of plasma DNA deconvolution using markers with low variability (category i) and high variability (category ii) in the same type of tissue. Plasma DNA deconvolution was performed to determine the placental contribution to the plasma DNA for 15 pregnant women. For each marker, the mean of the methylation densities of the two placental tissues were used to represent the methylation level of the placenta in the analysis. For each of the deconvolution analysis using the category i and category ii markers, a total of 1,024 markers were used.

[0142] The amount of placentally-derived DNA in plasma was further determined based on the proportion of the fetal-specific SNP alleles. The percentage contribution deduced by the methylation deconvolution analysis based on category i and category ii markers were then compared with the results based on fetal-specific SNP alleles. The median deviation of the derived placental contribution from the value estimated based on fetal-specific alleles was 2.7% and 7.1% using category i and category ii markers, respectively. Thus, the use of category i markers which had lower inter-individual variation in the tissue methylation level gave better accuracy in the methylation deconvolution analysis.

[0143] Significantly higher difference between the results from methylation deconvolution and fetal-specific allele analysis was observed when markers with high variability within the same type of tissue (category ii) were used ($P<0.0001$, Wilcoxon sign-rank test). In other words, the use of markers with low variability within the same type of tissue would increase the accuracy of

methylation deconvolution analysis. Accordingly, markers can be selected based on the variability within the same type of tissues, for example, but not limited to the value of CV and the difference between the maximum and minimum methylation density for the same type of tissues.

IV. DECONVOLUTION OF FETAL SIGNATURES

[0144] If a genomic signature (e.g., a particular SNP allele) is known, embodiments can determine which tissue is the origin of such signatures. Thus, if a particular signature is representative of a fetus (e.g., a paternal allele at a particular locus), then the fractional contribution for the signature would be substantial for the placental tissue.

[0145] To illustrate that single nucleotide alteration can also be used to determine the tissue of origin that the alteration is derived from, we analyzed the plasma DNA of a pregnant woman. The placenta and the maternal buffy coat were genotyped to identify the SNPs that the mother was homozygous and the fetus was heterozygous. We denote the allele shared by the fetus and the mother as A and the fetal-specific allele as B. Therefore, the mother had a genotype of AA and the fetus had a genotype of AB at each of these SNPs.

[0146] After bisulfite sequencing of the maternal plasma DNA, all DNA fragments carrying the fetal-specific allele (B allele) and at least one CpG site were selected and used for downstream analysis. A total of 1.31 billion fragments were sequenced and 677,140 fragments carrying the fetal-specific allele (B allele) were used for the deconvolution analysis. All CpG sites that were covered by at least 10 DNA fragments were used for deconvolution analysis. Other numbers of DNA fragments covering a site can be used, such as 5, 15, 20, 25, or 30. As the B allele was fetal specific, these DNA fragments were expected to be derived from the placenta.

Tissue	Contribution (%)
Liver	0.9
Lung	0.0
Colon	0.0
Small intestines	0.0
Pancreas	0.5
Adrenal glands	0.0
Esophagus	3.1
Adipose tissues	0.0
Heart	0.0
Brain	0.3

T cells	0.0
B cells	0.0
Neutrophil	0.0
Placenta	95.2

Table 3. Methylation deconvolution analysis using fetal-specific allele.

[0147] In Table 3, from the methylation deconvolution analysis, it was shown that the placenta was deduced to be the major contributor for these DNA fragments carrying fetal-specific SNP alleles. These results suggest that the methylation deconvolution analysis accurately identified the tissue origin of these DNA fragments carrying fetal-specific alleles.

[0148] This shows that a particular allele can be attributed to a fetus. Such a technique is described in more detail below for determining genotypes and haplotypes of a fetus using methylation deconvolution analysis.

V. DETERMINATION OF FETAL GENOME (MUTATIONAL ANALYSIS)

[0149] For noninvasive prenatal testing, the analysis of the inheritance of a maternal mutation using maternal plasma DNA is a challenging task. For example, if a pregnant woman is heterozygous for a mutation, the analysis on the mutational status of the fetus using maternal plasma DNA analysis would be technically difficult because both the mutant and the normal alleles would be present in her plasma, regardless of the mutational status of her fetus. Previously, a number of different approaches have been developed to address this problem (Lun et al. Proc Natl Acad Sci U S A. 2008;105:19920-5; Lo et al. Sci Transl Med. 2010;2:61ra91; Lam et al. Clin Chem. 2012;58:1467-75). The principle of these previous approaches involves the comparison between the relative amounts of the mutation and the normal allele in maternal plasma. To enhance the statistical power of the comparisons, some of these approaches further involve the comparisons of the relative amounts of SNP alleles linked to the mutation and those linked to the normal allele. As an alternative or in addition, some embodiments of the present invention can deduce the mutational status of the fetus by methylation deconvolution analysis.

A. Contribution for Alleles using Methylation Deconvolution

[0150] In this example, a genotype of the fetus is determined. Assume the genotypes of the father and mother are NN and MN at a particular locus, respectively. M and N denote the mutant

and normal alleles, respectively. In this scenario, the fetus can inherit either the M allele or the N allele from the mother. Therefore, there are two possible genotypes for the fetus, namely MN and NN. In maternal plasma, DNA carrying the fetal genotype is actually derived from the placenta. Thus, these DNA fragments would exhibit the placental methylation profile.

[0151] FIG. 11A shows a first scenario where the fetus has inherited the M allele from the mother and has a genotype of MN at a particular locus according to embodiments of the present invention. In the top part of FIG. 11A (labeled Genotypes), the father is shown as having genotype NN, the mother is shown as having genotype MN, and the fetus is shown as having genotype MN. The DNA fragments that exhibit the placental methylation profile are marked with a P, where are shown on the fetal genotype. For example, the placental methylation profile can correspond to certain methylation levels at genomic sites near the particular locus. DNA fragments that align to the particular locus can also include genomic sites near the locus (e.g., within 200 bp of the locus), and thus can be used to measure methylation levels for the methylation deconvolution analysis. Considering the genotypes of the parents, the M allele is specific for the mother and the N allele is shared between the father and the mother.

[0152] In the bottom part of FIG. 11A (labeled Maternal plasma), instances of the two alleles M and N are shown, with each instance representing a different DNA molecule in the plasma at the locus of interest. Only a small number of DNA molecules are shown for illustration purposes. In this example, the fetal DNA percentage is assumed to be 25%, as shown by 25% of the DNA molecules being marked with a P.

[0153] In the maternal plasma sample, we selectively analyzed the DNA fragments carrying the M allele and performed the methylation deconvolution analysis. Because the fetus has a genotype of MN, the placenta would contribute both M and N alleles to the maternal plasma DNA. Therefore, some of the DNA fragments carrying the M allele would also carry the placenta-specific methylation profile at genomic sites near the locus. The methylation deconvolution analysis would indicate that some of the DNA fragments carrying the M allele would be derived from the placenta, and thus the fetal genotype does include the M allele.

[0154] FIG. 11B shows a second scenario where the fetus has inherited the N allele from the mother and has the genotype of NN at a particular locus according to embodiments of the present invention. In this situation, only the DNA fragments carrying the N allele would exhibit the

placental methylation profile in maternal plasma. Thus, the selective analysis of the DNA fragments carrying the M allele with methylation deconvolution would indicate that these DNA fragments do not have significant contribution from the placenta. Accordingly, it can be determined that the fetus does not have M, and thus has a genotype of NN.

[0155] In some embodiments, the placental contribution for the M and N alleles can be compared. Here, we assume that the fetal DNA accounts for approximately 10% of the total maternal plasma DNA. The selective deconvolution of the M and N alleles would be useful to indicate which allele the fetus has inherited from the mother. The expected results are shown in Table 4 below:

	Fetal genotype	
	MN	NN
Placental contribution for plasma DNA carrying the M allele	Approximately 10%	Insignificant (close to zero)
Placental contribution for plasma DNA carrying the N allele	Approximately 10%	Approximately 20%
Ratio of placental contribution for M and N alleles (M:N)	1:1	0:2

Table 4. Placental contributions for M and N alleles for NN paternal genotype.

[0156] In Table 4, the percentage placental contribution of the M and N alleles can be compared. An approximately equal placental contribution for the two alleles (e.g., within a threshold of each other) suggests that the fetal genotype is MN. On the other hand, a significantly higher placental contributions for the N allele compared with the M allele would indicate a fetal genotype of NN.

[0157] In another embodiment, the paternal genotype does not need to be taken into account. In this situation, the possible genotypes of the fetus include MM, MN, and NN.

	Fetal genotype		
	MN	NN	MM
Placental contribution for plasma DNA carrying the M allele	Approximately 10%	Insignificant (close to zero)	Approximately 20%
Placental contribution for plasma DNA carrying the N allele	Approximately 10%	Approximately 20%	Insignificant (close to zero)

Ratio of placental contribution for M and N alleles (M:N)	1:1	0:2	2:0

Table 5. Placental contributions for M and N alleles for unknown paternal genotype.

[0158] In table 5, the placental contribution for the DNA fragments carrying the M and N alleles for different fetal genotypes are shown. When the fetus has a genotype of MM, the placental contribution for the M allele would be significantly higher than that for the N allele. When the fetus has a genotype of NN, the placental contribution for the N allele would be significantly higher than that for the N allele. When the fetus has a genotype of NM, the placental contribution for the M allele would be approximately equal to the placental contribution for the N allele.

[0159] Accordingly, where the paternal genotype is not known, fractional contributions can be determined for both alleles. That is, a first fractional contribution can be determined using a first set of cell-free DNA molecules that align to the locus and include N. Methylation levels of the first set of cell-free DNA molecules can be measured at K genomic sites near the locus. And, a second fractional contribution can be determined using a second set of cell-free DNA molecules that align to the locus and include M. Methylation levels of the second set of cell-free DNA molecules can be measured at the K genomic sites near the locus. For the first scenario of the fetal genotype being MN, the fractional contributions determined for either allele would be about the same, as can be tested to by determining whether the fractional contributions are within a threshold value of each other.

[0160] To illustrate the feasibility of this approach, we analyzed the plasma DNA of a pregnant woman. The plasma DNA was bisulfite converted and analyzed using massively parallel sequencing. In addition, the placenta and blood cells were analyzed to determine the genotype of the fetus and the mother. For illustration purposes, a SNP located within the *KLF2* gene was analyzed. For this SNP, the genotypes of the mother and the fetus were CG and CC, respectively. With this genotype combination, the placenta would contribute the C allele to the maternal plasma, but all the G alleles in the maternal plasma would be derived from the maternal tissues.

[0161] In the sequencing data, there were 24 fragments carrying the G allele and 55 fragments carrying the C allele. The CpG sites within these DNA fragments were used for methylation deconvolution. In this analysis, an objective is to determine the placental contribution of the two

alleles. To illustrate the principle, only the placenta and the blood cells were considered as candidate tissues for the methylation deconvolution analysis. In another embodiment, three or more types of tissues can be used as candidates. In yet another embodiment, tissues expected to have significant contribution, for example blood cells, liver, lung, intestines and placenta, can be used as candidates.

	C allele	G allele	C/G ratio
Placenta	62.6%	1.8%	34
Blood cells	37.4%	98.2%	

Table 6. Placental contributions for C and G alleles for unknown paternal genotype.

[0162] In Table 6, the contribution from the placenta was deduced to be 62.6% and 1.8% for the C allele and the G allele, respectively. The ratio of placental contribution for C/G is 34. These results suggest that the genotype of the fetus would be CC. This is consistent with the genotyping result of the placental tissue.

[0163] This embodiment is different from and potentially has more utility than a previous method for noninvasive prenatal testing based on the analysis of allelic ratio for DNA with a specific methylation pattern (Tong et al. Clin Chem 2006; 52: 2194-202). In this previous method, tissue-specific DNA is first identified from a DNA mixture (e.g. plasma DNA) based on methylation pattern. For example, a particular gene is completely unmethylated in the blood cells and methylated in the placenta. The identification is performed using an enzyme that leaves the methylated placental DNA intact.

[0164] Thus, all the methylated DNA molecules remaining in the plasma would be derived from the placenta rather than from the blood cells. Then, the allelic ratio for a SNP located on the placenta-derived DNA molecules can be determined by measuring amounts of the different alleles at the locus using the intact placental DNA. When the fetus is heterozygous for the SNP, the ratio of the two alleles in the placenta-specific DNA would be approximately 1. However, if the fetus is affected by an aneuploid chromosome and has three copies of the chromosome carrying this particular SNP, the ratio of the two alleles would be either 1:2 or 2:1.

[0165] In this previous method, the tissue-specific DNA molecules need to be first identified based on a methylation status that is unique to the tissue of interest. The methylated DNA molecules are unique for the placenta because the blood cells are completely unmethylated for the

targeted region. However, in this present embodiment, the uniqueness of a certain methylation state is not required. The candidate tissues only need to be different in their methylation profiles, accordingly more loci can be used, thereby enabling haplotype deconvolution. Thus, the tissue contributions can be determined for the different alleles based on their methylation profiles. Further, the previous method may be more susceptible to statistical variations as the numbers of fetal reads with each allele are compared directly to each other. Whereas, when the placental contributions are compared to each other, the numbers of fetal reads are not compared directly to each other. Instead, the placental contribution is determined from all of the reads (methylated or not), and thus the placental contributions can be the same, even when the number of fetal reads differs. Thus, a coverage bias to one haplotype can be accounted for.

B. Determination of Inherited Haplotype using Deconvolution

[0166] It has previously been demonstrated that through the analysis of plasma DNA (or other cell-free DNA) of a pregnant woman carrying a fetus, the maternal haplotypes inherited by the fetus can be deduced using the process of relative haplotype dosage analysis (RHDO) (Lo et al. Sci Transl Med 2010; 2: 61ra91 and U.S. Patent 8,467,976). In this method, one uses the haplotype information for the pregnant woman. This latter information can be obtained using family analysis or a method for the direct analysis of the haplotype (e.g. Fan et al. Nat Biotechnol 2011; 29: 51-57; Snyder et al. Nat Rev Genet 2015; 16: 344-358). SNPs that are heterozygous in the mother but homozygous in the father can be used for the RHDO analysis. Such a use of specific SNPs can limit the loci that can be used, and therefore limit the amount of data and accuracy. Embodiments may not be so restricted to such specific SNPs. Further, embodiments can be used in combination with the above references to provide increased accuracy.

[0167] Embodiments can use methylation deconvolution to determine placental contributions using the cell-free DNA molecules for two haplotypes. The placental contributions can be compared to determine which haplotype is inherited by the fetus. Embodiments can start with deduced maternal or paternal haplotypes, and then measure the methylation levels of plasma DNA molecules containing SNP alleles in each of those deduced haplotypes. One can then perform methylation deconvolution. The fetal haplotype can be identified as the one with the highest placental contribution from the methylation deconvolution analysis. In all of the above embodiments, the paternal or maternal haplotypes can, instead of being a deduced one, also be

determined by family analysis (i.e. by analyzing the DNA of other family members) or by a direct method (e.g. the method described by Fan et al Nat Biotechnol 2012).

1. Maternal Haplotypes

[0168] In this example, we demonstrate plasma DNA methylation deconvolution analysis can be used for deducing the maternal haplotypes inherited by an unborn fetus. A source of genomic DNA from the pregnant woman, e.g. the buffy coat DNA, can be subjected to genotyping, e.g. using a microarray. Then, the maternal genotyping results are entered into a haplotype deduction program (e.g., IMPUTE2, Howie et al. PLoS Genet. 2009;7:e1000529) to deduce the likely first maternal haplotype and the second maternal haplotype. Population-specific genotype and haplotype information can be taken into consideration for improving the accuracy of deduction. In other embodiments, the parental haplotypes can be worked out by single molecule analysis, for example but not limited to the methods described by Fan et al (Nat Biotechnol. 2011;29:51-7) , Kaper et al (Proc Natl Acad Sci U S A. 2013;110:5552-7), Lan et al. (Nat Commun. 2016;7:11784) and Selvaraj et al (Nat Biotech 2013;31:1111–1118). Then, maternal plasma DNA can be subjected to genome-wide bisulfite sequencing and alignment to reference genomic sequences. Methylation deconvolution can then be performed for each of the predicted haplotypes. As fetal DNA in maternal plasma is predominantly of placental origin, the maternal haplotype inherited by the fetus is the one that shows the highest placental contribution.

[0169] The maternal haplotype information can be used to link the SNP alleles and the CpG sites on the same homologous chromosome together. Then, DNA fragments from the same chromosome copy (haplotype) can be identified using the SNP alleles. The CpG sites (or other sites) on this particular chromosome copy (haplotype) can be used for the methylation deconvolution. As the number of CpG sites that can be used for deconvolution would be proportional to the number of SNPs on the homologous chromosome and much bigger than the number of CpG sites linked to a single SNP in the haplotype-based deconvolution analysis, this method would be more precise than the deconvolution analysis using CpG site(s) that are linked to a single SNP. The principle is illustrated in FIG. 12A.

[0170] FIG. 12A shows a determination of a maternal haplotype inherited by a fetus using methylation deconvolution according to embodiments of the present invention. In the top part of

FIG. 12A, the two haplotypes of the mother and the fetus are shown at three loci that the mother is heterozygous. The two maternal haplotypes are labeled as Hap I and Hap II. In this example, the fetus has inherited Hap I from the mother. For illustration purpose, only the SNP loci that the mother is heterozygous are shown. For illustration purposes, the father is homozygous for each of these loci in this example. However, the same principle extends to scenarios that the father is heterozygous without any change.

[0171] In the bottom part of FIG. 12A (labeled Maternal plasma), instances of the two alleles at each locus are shown, with each instance representing a different DNA molecule in the plasma at the locus of interest. Only a small number of DNA molecules are shown for illustration purposes. In this example, the fetal DNA percentage is assumed to be 20%, as shown by 20% of the DNA molecules being marked with a P.

[0172] In maternal plasma, the DNA molecules carrying the fetal genotype are derived from the placenta and therefore carrying the placental specific methylation patterns. The circles labeled with "P" represent the CpG sites exhibiting the placental methylation pattern near the heterozygous locus. A read including a heterozygous locus and a neighboring site can be used for measuring a methylation level for detecting the placental methylation pattern. In this example, an objective is to determine if the fetus has inherited Hap I or Hap II from the mother. To achieve this, plasma DNA fragments that carry alleles on Hap I and cover at least one CpG site are selected for methylation deconvolution. As the fetus has inherited Hap I from the mother, the placenta would contribute a significant proportion to this pool of plasma DNA molecules. On the other hand, when the fragments carrying alleles on Hap II are analyzed with methylation deconvolution, a very low contribution from the placenta would be observed.

[0173] To illustrate this, we analyzed the maternal plasma sample stated above for Table 6. We focused on a 5-Mb region on chromosome 1. SNPs where the mother was heterozygous and the fetus was homozygous were selected for the analysis. For each of these SNP loci, the alleles that were shared between the mother and the fetus formed one haplotype (denoted as Hap I) and the alleles that were present only on the maternal genome formed another haplotype (denoted as Hap II). Thus, in this example, there are two maternal haplotypes (Hap I and Hap II) and the fetus has inherited Hap I from the mother. In the maternal plasma, the DNA fragments carrying the alleles on Hap I and those carrying the alleles on Hap II were analyzed separately using methylation

deconvolution. All the CpG sites on the same plasma DNA molecule of a heterozygous SNP were used for the deconvolution analysis. In this example, none of these CpG sites overlapped with type I or type II markers.

	Hap I	Hap II
Liver	0%	0%
Lung	0%	6.7%
Colon	3.4%	6.2%
Small intestines	0%	10.6%
Pancreas	4.1%	0%
Adrenal glands	0%	4.6%
Esophagus	0%	0%
Adipose tissues	3.7%	3.6%
Heart	0%	0%
Brain	6.8%	10.6%
T cells	6.8%	21%
B cells	8.9%	11.7%
Neutrophil	12.7%	25%
Placenta	53.5%	0%

Table 7. Methylation deconvolution for Hap I and Hap II.

[0174] Table 7 shows deconvolution of plasma DNA fragments carrying the alleles on the two maternal haplotypes, namely Hap I and Hap II. The fetus had inherited the maternal Hap I. From this deconvolution analysis, the placenta was deduced to contribute 53.5% of the plasma DNA fragments carrying the alleles on Hap I. On the other hand, there was no contribution from the placenta to the plasma DNA fragments carrying the alleles on Hap II. Therefore, the methylation deconvolution analysis had accurately predicted that the fetus had inherited Hap I from the mother. Greater accuracy may be achieved using CpG sites that overlap with type I and/or type II markers.

[0175] As a further example, to demonstrate the practical utility of this approach, another pregnant woman was recruited. Maternal peripheral blood was taken. The blood sample was fractionated into plasma and the cellular components. The maternal buffy coat was analyzed using an Illumina HumanOmni2.5-8 BeadChip array. We used IMPUTE2 (Howie et al. PLoS Genet. 2009;7:e1000529) to deduce the phase of 851 heterozygous SNPs on a 5 Mb region on the

telomeric end of chromosome 1p. The haplotype phasing was based on reference haplotypes of 1000 genomes (mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.tgz).

[0176] After the phased haplotypes were obtained, the CpG sites linked to the two haplotypes were used to perform methylation deconvolution. All the CpG sites on the same plasma DNA molecule of a heterozygous SNP were used for the deconvolution analysis. In this example, none of these CpG sites overlapped with type I or type II markers. Among the 851 SNPs used for the deconvolution, 820 (96.2%) were on intron and intergenic regions. None of them overlapped with CpG islands or shores.

	Hap I	Hap II
Liver	0	0
Lung	0	5.4
Colon	0	6.2
Small intestine	0	0
Pancreas	0	25
Adrenal glands	0	0
Esophagus	0	0
Adipose tissues	0	17.8
Heart	0	0
Brain	0	0
T cells	11	7.9
B cells	0	0
Neutrophils	20.2	28.4
Placenta	68.9	9.3

Table 8. Methylation deconvolution for Hap I and Hap II.

[0177] Table 8 shows a deconvolution of plasma DNA fragments carrying the alleles on the two maternal haplotypes deduced from a panel of reference haplotypes. The two haplotypes are named Hap I and Hap II. The deduced Hap I has significantly higher amount of placental contribution than Hap II, namely, 68.9% versus 9.3%. Thus, the maternal Hap I was deduced to have been inherited by the fetus. The maternal inheritance relied on haplotype deduction was consistent with results from maternal and fetal genotypes.

[0178] An advantage of this method is that one is not restricted to SNPs for which the father of the fetus is homozygous and the mother of the fetus is heterozygous. Indeed, in the above example,

we had performed the analysis without knowing or deducing the paternal genotype or haplotype. This is an advantage over previously described methods ((Lo et al. Sci Transl Med 2010; 2: 61ra91, US Patent 8,467,976, Fan et al. Nature 2012; 487: 320- 324, Kitzman et al. Sci Transl Med 2012; 4: 137ra76).

[0179] In some embodiments, a first fractional contribution for a first haplotype can be compared to a reference value derived based on the fetal DNA fraction to determine whether the haplotype has been inherited by the fetus. The cutoffs can be calculated as for example but not limited to 1 time, 1.2 times, 1.4 times, 1.6 times, 1.8 times, 2 times, 2.2 times, 2.4 times, 2.6 times or 2.8 times of the fetal DNA fraction. In this manner, the second fractional contribution for a second haplotype does not need to be determined, if the first fractional contribution is sufficiently large.

[0180] In some embodiments, the inherited haplotype may have a deconvoluted fractional concentration double that of the fetal fraction and the non-inherited one has an insignificant contribution. The contribution of the non-inherited haplotype may not have a zero contribution as the paternal haplotypes may give noise to this analysis because some paternal alleles may be the same as the maternal alleles. If the level of noise is high, the fractional contribution of the second haplotype can be determined, and the one with a higher deconvoluted fraction can be deduced to be inherited by the fetus.

[0181] Some implementations could test both haplotypes using the reference value, to confirm that only one is inherited. If both appear to be inherited, then the two fractional contributions can be compared to each other. Additionally, if both appear to be inherited, the paternal genome can be checked, as the fetus could have inherited a paternal haplotype that matches the non-inherited maternal haplotype.

[0182] In other embodiments, the second fractional contribution can be used to determine the reference value, e.g., the second fractional contribution plus a threshold value. Thus, the reference value can be a sum of the second fractional contribution and a threshold value.

2. Paternal haplotype

[0183] In another embodiment, methylation deconvolution analysis can be applied for the analysis of the paternal haplotype inheritance.

[0184] FIG. 12B shows an illustration of the paternal haplotype methylation analysis according to embodiments of the present invention. Methylation deconvolution can be performed on the maternal plasma DNA fragments carrying the alleles on paternal Hap III and Hap IV. As Hap III has been inherited by the fetus, the placental contribution would be higher for Hap III compared with Hap IV. Thus, the paternal inheritance of the fetus can be deduced.

[0185] This embodiment has advantages over previous methods based on the analysis of paternal-specific alleles. For example, for SNP at position 1, the A allele is present in the father, but not in the mother. Therefore, the detection of the paternal-specific A allele in maternal plasma indicates the inheritance of the Hap III by the fetus. However, for SNP at position 2, both the C and T alleles are not fetal-specific. In this situation, paternal-specific allele analysis cannot be used. However, the methylation deconvolution analysis does not require the presence of paternal-specific allele. Thus, SNPs that are heterozygous in both the father and the mother can be used for the methylation deconvolution analysis of the two paternal haplotypes.

[0186] Accordingly, a similar process, as used for maternal haplotypes, can be used to determine which paternal haplotype is inherited. In FIG. 12B, the placental contribution for Hap III would be higher than the placental contribution from Hap IV. The paternal haplotypes can be determined in a same or similar manner as the maternal haplotypes can be determined.

3. Method using Deconvolution

[0187] FIG. 13 is a flowchart illustrating a method 1300 of determining a portion of a fetal genome from a maternal sample using methylation deconvolution according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type. The fetus has a father and a mother being the pregnant female. The portion of the fetal genome can be an entire chromosome copy or just part of the chromosome copy. The determined portions of the fetal genome can be combined to provide information on different portions of the fetal genome, up to the entire fetal genome.

[0188] At block 1310, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1310 can be performed using techniques described in block 140 of method 100 of FIG. 1. For example, at least 1,000 cell-free DNA molecules can be analyzed to determine where

the cell-free DNA molecules are located, and methylation levels can be measured as described below. Further, the cell-free DNA molecules are analyzed to determine a respective allele of the cell-free DNA molecule. For example, an allele of a DNA molecule can be determined from a sequence read obtained from sequencing or from a particular probe that hybridizes to the DNA molecule, where both techniques can provide a sequence read (e.g., the probe can be treated as the sequence read when there is hybridization).

[0189] At block 1320, a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus are determined. One skilled in the art will be aware of various techniques to determine haplotypes of a parent. The haplotypes may be determined from the same sample as used to determine methylation levels below or from a different sample. In some implementations, the haplotypes can be determined from cellular samples, e.g., the buffy coat of a blood sample or the tissue of another organ. Examples of determining haplotypes are provided in U.S. Patent No. 8,467,976, which is incorporated by reference in its entirety. The first parent can be the mother or the father. Other examples of methods for detecting the parental haplotypes include, but not limited to the methods described by Fan et al (Nat Biotechnol 2011; 29: 51-57), Snyder et al (Nat Rev Genet 2015; 16: 344-358), the GemCode technology from 10X Genomics (www.10xgenomics.com/), and Targeted Locus Amplification (TLA) technology from Cergentis (www.cergentis.com/).

[0190] At block 1330, one or more heterozygous loci are identified from the first and second haplotypes. Each heterozygous locus has a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype. The one or more heterozygous loci may be a first plurality of heterozygous loci, where a second plurality of heterozygous loci can correspond to a different chromosomal region.

[0191] At block 1340, a first set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the heterozygous loci from block 1330 and includes a corresponding first allele, so that the cell-free DNA molecule can be identified as corresponding to the first haplotype. It is possible for a cell-free DNA molecule to be located at more than one of the heterozygous loci, but typically a read would only include one heterozygous locus. Each of the first set of cell-free DNA molecules also includes at least one of N genomic sites, where the genomic sites are used to measure the methylation levels. N is an integer,

e.g., greater than or equal to 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000, or 5,000. Thus, a read of a cell-free DNA molecule can indicate coverage of 1 site, 2 sites, etc.

[0192] At block 1350, N first mixture methylation levels are measured at the N genomic sites (e.g., CpG sites) using the first set of the plurality of cell-free DNA molecules. One first mixture methylation level can be measured for each of the N genomic sites. Block 1350 can be performed in a similar manner as block 150 of method 100 of FIG. 1. In some embodiments, the measurement of the methylation level of a DNA molecule can use methylation-aware sequencing results, which may also be used to determine the location and respective allele of the DNA molecule. One skilled in the art will be aware of the various techniques that can be used to determine a methylation status of a site on a DNA molecule.

[0193] At block 1360, a first fractional contribution of the fetal tissue type in the mixture is determined using the N first methylation levels. In some embodiments, block 1360 can be performed via blocks 160 and 170 of method 100 of FIG. 1. Thus, a fractional contribution can be determined simultaneously for a panel of M tissue types. Block 1360 may use N issue-specific methylation levels at N genomic sites, determined for each of M tissue types, e.g., as in block 120 of method 100 of FIG. 1.

[0194] At block 1370, a second set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the heterozygous loci from block 1330 and includes a corresponding second allele, so that the cell-free DNA molecule can be identified as corresponding to the second haplotype. Each of the second set of cell-free DNA molecules also includes at least one of the N genomic sites, where the genomic sites are used to measure the methylation levels.

[0195] At block 1380, N second mixture methylation levels at the N genomic sites are measured using the second set of the plurality of cell-free DNA molecules. Block 1380 may be performed in a similar manner as block 1350.

[0196] At block 1385, a second fractional contribution of the fetal tissue type in the mixture is determined using the N second methylation levels. Block 1385 may be performed in a similar manner as block 1360.

[0197] At block 1390, a first separation value is computed between the first fractional contribution and the second fractional contribution. Examples of separation values are described herein, e.g., including a difference or a ratio.

[0198] At block 1395, the portion of fetal genome is determined at the one or more heterozygous loci based on the first separation value. Thus, an inherited haplotype of the first parent can be determined. For example, the first separation value can be a ratio of the first fractional contribution and the second fractional contribution. The portion of the fetal genome can be determined to have one or more copies of the first haplotype and no copies of the second haplotype when the ratio is greater than a threshold value. Examples of threshold values include but not limited to 1.3, 1.4, 1.5, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8 and 3.0. The portion of the fetal genome can be determined to have one or more copies of the second haplotype and no copies of the first haplotype when the ratio is less than a threshold value. Examples of threshold values include but not limited to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8. The portion of the fetal genome can be determined to have the first haplotype and the second haplotype when the ratio is equal to one within a cutoff value. Examples of cutoff values include but not limited to 0.85, 0.9, 0.95, 1.0, 1.05, 1.1 and 1.15. Both haplotypes might be inherited when both parents have a same haplotype in the region being analyzed.

[0199] As another example, the first separation value is a difference of the first fractional contribution and the second fractional contribution. The portion of the fetal genome can be determined to have one or more copies of the first haplotype and no copies of the second haplotype when the difference is greater than a threshold value. Examples of threshold values include but not limited to 1%, 1.5%, 2%, 2.5%, 3%, 4%, 5%, 6%, 7%, 8%, 10%, 12%, 14%, 16%, 18% and 20%. The portion of the fetal genome can be determined to have one or more copies of the second haplotype and no copies of the first haplotype when the difference is less than a threshold value, e.g., where the threshold value is a negative number.

[0200] The inherited haplotype of the other parent can also be determined. For example, a second plurality of heterozygous loci of the first chromosomal region can be identified in the genome of the other parent. Fractional contributions can be determined for each of the other parent's haplotypes, and a separation value can be used to determine the inherited haplotype of the other parent.

[0201] For example, the first plurality of heterozygous loci and the second plurality of heterozygous loci can be the same loci or be different. Each of the second plurality of heterozygous loci can include a corresponding third allele in a first haplotype of the other parent (e.g., a first paternal haplotype) and a corresponding fourth allele in a second haplotype of the other parent (e.g., a second paternal haplotype). The third and fourth alleles can be the same as the first and second alleles. In addition to the first and second set of cell-free DNA molecules for the first parent, a third set of the plurality of cell-free DNA molecules can each be located at any one of the second plurality of heterozygous loci, include the corresponding third allele of the heterozygous locus, and include at least one of K genomic sites. The K genomic sites may be the same or different than the N genomic sites used for the first parent. In a similar manner as with the first parent, K third mixture methylation levels can be measured at the K genomic sites using the third set of the second plurality of cell-free DNA molecules, and a third fractional contribution of the fetal tissue type in the mixture can be determined using the K third methylation levels. The third fractional contribution corresponds to the first haplotype of the other parent (e.g., the first paternal haplotype).

[0202] A fourth set of the plurality of cell-free DNA molecules can each be located at any one of the second plurality of heterozygous loci, include the corresponding fourth allele of the heterozygous locus, and include at least one of the K genomic sites. Thus, the fourth set of DNA can be used to test the second haplotype of the other parent. K fourth mixture methylation levels at the K genomic sites can be measured using the fourth set of the second plurality of cell-free DNA molecules, and a fourth fractional contribution of the fetal tissue type in the mixture can be determined using the K fourth methylation levels. A second separation value can be computed between the third fractional contribution and the fourth fractional contribution, and the portion of fetal genome at the second plurality of heterozygous loci can be determined based on the second separation value. The inherited haplotype from the other parent can be determined in a similar manner as for the first parent. The fourth fractional contribution corresponds to the second haplotype of the other parent (e.g., the second paternal haplotype).

[0203] In some embodiments, the second fractional contribution does not need to be determined. Instead, a haplotype can be determined to be inherited if the corresponding fractional contribution is sufficiently high. For example, the first fractional contribution can be compared to a reference

value to determine whether the fetus inherited the first haplotype at the first chromosomal region. The fetus can be determined to have inherited the first haplotype at the first chromosomal region when the first fractional contribution exceeds the reference value.

[0204] In other embodiments, the reference value may be determined from the second fractional contribution. For example, the reference value can be a sum of the second fractional contribution and a threshold value. The sum with the threshold value can ensure that the first fractional contribution is sufficiently greater than the second fractional contribution.

[0205] A separate determination of inheritance can be made for the second haplotype by comparing the second fractional contribution to the reference value to determine whether the fetus inherited the second haplotype at the first chromosomal region. The fetus can be determined to have inherited the second haplotype at the first chromosomal region when the second fractional contribution exceeds the reference value. If both the fractional contributions are determined to exceed the reference value, the two fractional contributions can be compared to each other to determine if one is significantly greater than the other (e.g., using a threshold). The haplotypes of the other parent can be determined to identify whether one of these haplotypes is the same as the haplotypes of the first parent, thereby explaining that both haplotypes of the first parent could have been inherited.

C. Determination of Inherited Haplotype using Methylation Levels

[0206] Other embodiments can use the general hypomethylation of cell-free fetal DNA to identify the inherited haplotype as the one with the lower overall methylation level. Embodiments can start with deduced maternal or paternal haplotypes, and then measure the methylation levels of plasma DNA molecules containing SNP alleles in each of those deduced haplotypes. In one implementation of analyzing the maternal haplotypes, the methylation levels of the two deduced maternal haplotypes can be compared, and the one with the lower methylation level would be predicted to be the haplotype inherited by the fetus. In another implementation of analyzing the paternal haplotypes, the methylation levels of the two deduced paternal haplotypes can be compared, and the one with the lower methylation level would be predicted to be the haplotype inherited by the fetus.

1. Example

[0207] As an example, a methylation level of each of the two maternal haplotypes can be determined. As the placental tissue is relatively hypomethylated compared with other tissues, it is expected that the maternal haplotype inherited by the fetus would be more hypomethylated than the one that is not inherited by the fetus. The methylation densities were tested in the maternal plasma using the actual haplotypes of the mother, which were deduced using the maternal, paternal, and fetal genotypes.

	Hap I	Hap II
Overall methylation densities	65%	87%

Table 9. Methylation densities for actual Hap I and Hap II.

[0208] Table 9 shows methylated densities of the two maternal haplotypes in the maternal plasma. As Hap I was the actual haplotype inherited by the fetus by genotyping, the results of the methylation analysis of the haplotype correctly identified the inheritance.

[0209] In other embodiments, the maternal haplotypes can be deduced based on the genotypes of the mother alone, or reference haplotypes of the population from haplotype database can also be used for this analysis. The maternal haplotypes used in this example were phased using the IMPUTE2 program. Thus, deduced maternal haplotypes can also be used in this analysis.

	Hap I	Hap II
Overall methylation densities	68%	76%

Table 10. Methylation densities for deduced Hap I and Hap II.

[0210] Table 10 shows methylated densities of the two deduced maternal haplotypes in the maternal plasma. The deduced maternal haplotype that was inherited by the fetus had lower methylation densities. An example of a statistical procedure that one can use to determine whether one haplotype has a sufficiently lower methylation density includes the chi-square test. A separation between the two methylation levels can be required to be sufficiently large (e.g., greater than a threshold) to make the determination. If the separation is not sufficient, then an indeterminate classification can be made. In some embodiments, a determination of an inheritance of both haplotypes can be determined if the separation is not sufficiently large and if both methylation levels are below a threshold level, which may be characterized by inclusion of fetal

DNA. For example, tables 9 and 10 indicate that a methylation density below 70% may indicate that the fetus has inherited that haplotype. Both haplotypes may be inherited when the parents share a haplotype for the region being analyzed.

[0211] In another embodiment, the overall methylation densities of the maternal plasma DNA carrying the paternal Hap III and Hap IV can be compared. Similar to the maternal haplotype analysis, the fetus would be deduced as having inherited the paternal haplotype that has lower overall methylation densities.

2. Method using Methylation Levels

[0212] FIG. 14 is a flowchart illustrating a method 1400 of determining a portion of a fetal genome from a maternal sample using methylation levels according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type. The fetus has a father and a mother being the pregnant female. The portion of the fetal genome can be an entire chromosome copy or just part of the chromosome copy. The determined portions of the fetal genome can be combined to provide the entire fetal genome, as with other methods described herein.

[0213] At block 1410, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1410 can be performed in a similar manner as block 1310 of method 1300 of FIG. 13.

[0214] At block 1420, a first haplotype and a second haplotype of a first chromosomal region are determined of a first parental genome of a first parent of the fetus. Block 1420 can be performed in a similar manner as block 1320 of FIG. 13. In some embodiments, genotypes of the first parent's genome can be determined at the plurality of heterozygous loci using a sample from the first parent, e.g., a blood sample or other tissue that may or may not include fetal DNA. A plurality of reference haplotypes can be obtained, e.g., from databases of reference genomes. The first haplotype and the second haplotype can be deduced using the genotypes and the plurality of reference haplotypes. For example, the alleles of each genotype can be compared against the reference haplotypes, and any haplotypes that do not include the alleles at the corresponding locus can be discarded. Once two reference haplotypes remain, those haplotypes can be identified as being the first haplotype and the second haplotype.

[0215] At block 1430, a plurality of heterozygous loci is identified from the first and second haplotypes. Each heterozygous locus has a first allele in the first haplotype and a second allele in the second haplotype.

[0216] At block 1440, a first set of the plurality of cell-free DNA molecules is identified. Block 1440 can be performed in a similar manner as block 1340 of FIG. 13.

[0217] At block 1450, a first mixture methylation level is measured using the first set of the plurality of cell-free DNA molecules. For example, the first mixture methylation level may be a methylation density for the cell-free DNA molecules of the first set. The methylation density can be computed as a total methylation density for all of the cell-free DNA molecules of the first set. In another example, separate methylation densities can be computed for each locus, and the separate methylation densities can be combined to obtain the first mixture methylation level, e.g., an average of the separate methylation densities.

[0218] At block 1460, a second set of the plurality of cell-free DNA molecules is identified. 1460 can be performed in a similar manner as block 1370 of FIG. 13.

[0219] At block 1470, a second mixture methylation level is measured using the second set of the plurality of cell-free DNA molecules. For example, the second mixture methylation level may be a methylation density for the cell-free DNA molecules of the second set.

[0220] At block 1480, it is determined which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower. As part of block 1480, a separation value can be determined between the first mixture methylation level and the second mixture methylation level, and compared to a threshold value. The threshold value can ensure that the lower level is sufficiently lower. The threshold value can be determined using the chi-square test. For example, measurements can be taken of samples where the inherited haplotype is known, and a distribution of the separation values can be determined, and a threshold value can be selected that accurately determines the inherited haplotype in the training data obtained from the samples. Methods 1300 and 1400 can also be combined, with each being performed as a check, and the inherited haplotype determined if both methods are consistent with each other.

D. Selection of loci

[0221] Various embodiments can be used for the comparison of the methylation levels or the fraction contributions of the two deduced maternal haplotypes in the maternal plasma. In one embodiment, the number of SNP loci to be analyzed can be determined before the analysis. For example, the number of SNP loci used in the haplotype deconvolution analysis can be determined according to a number of factors, for example, but not limited to the desired statistical power, the mean difference in methylation levels in the placenta and blood cells in the region of interest, and the number of molecules being analyzed for each SNP.

[0222] The size of the region of interest can be fixed, and all SNPs within the region of interest can be used in the analysis. The size of the region of interest can be determined taking into account a number of factors, for example, but not limited to the desired statistical power, the mean difference in methylation level in the placenta and blood cells in the region of interest, the number of molecules being analyzed for each SNP, and the chance of meiotic recombination with the region of interest.

[0223] In other embodiments, the number of SNPs and the size of the region to be analyzed are not determined before the analysis. For example, the number of SNPs can be sequentially increased until the data are sufficient to arrive at a statistically significant conclusion concerning which maternal haplotype is statistically significantly less methylated than the other one. For instance, the SNPs on the region of interest can be arranged in an ascending order of their genomic coordinates. Then statistical testing can be carried out with the data of the SNP with the lowest number of genomic coordinate. If this is sufficient to make a conclusion regarding which haplotype is less methylated statistically, then a conclusion is made. Similarly, the SNPs can be arranged in a descending order, with a highest number of genomic coordinate that is sufficient being used.

[0224] If the statistical accuracy is not sufficient, another statistical comparison can be performed starting from the next SNP with a higher number of genomic coordinates. On the other hand, if the data of the first SNP are not sufficient for one to conclude that one haplotype is less methylated than the other one (or that the separation value between fractional contributions is not sufficiently large), the data of another SNP can be added and another round of statistical testing is

carried out. This procedure can be continued until the accumulated data are sufficient to make a statistically significant conclusion. A number of statistical tests can be performed to compare the methylation levels of the two haplotypes, for example, but not limited to Student's t-test, Mann-Whitney rank-sum test and Chi-square test. The level of statistical significance can be determined based on the desired confidence of the conclusion, for example, but not limited to adopting a P-value of 0.05, 0.01, 0.001, 0.0001, or 0.00001.

E. Combinations with RHDO

[0225] In some embodiments, results generated by the RHDO analysis of U.S. Patent 8,467,976 can be combined with present methylation embodiments to arrive at a more accurate procedure for diagnosis or to reduce an amount of sequencing required. For example, fetal haplotypes can be determined using present embodiments and using results of the RHDO analysis of U.S. Patent 8,467,976, and the determined fetal haplotypes from both of the techniques can be compared. For example, the results from the two analyses would be accepted only if they are concordant. Further analysis can be performed if the two analyses show different conclusions, e.g., measurements can be repeated at higher depth of coverage on the genome.

[0226] For such a combined approach to be most cost-effective, it is preferred to have one type of sequencing that can yield data for both methods. In one embodiment, this can be done by a single molecule method that would generate sequencing as well as methylation information, e.g. using the Single Molecule Real Time sequencing technology from Pacific Biosciences, or nanopore sequencing (e.g. from Oxford Nanopore Technologies). These are two examples of methylation-aware sequencing. In another embodiment, the RHDO analysis can be performed on bisulfite sequencing results. For such an embodiment, any maternal and paternal genetic information can be determined also using bisulfite sequencing. Bisulfite sequencing is thus another example of methylation-aware sequencing. Furthermore, other methylation aware sequencing technologies can be used, such as oxidative bisulfite sequencing (Booth et al. Science 2012; 336: 934-937) or Tet-assisted bisulfite sequencing (Yu et al. Cell 2012; 149: 1368-1380). The latter examples would allow one to analyze the 5-methylcytosine distribution of the analyzed DNA molecules.

F. Uses of knowledge of fetal genome

[0227] The noninvasive prenatal analysis of the fetal genome can be used to determine if a fetus has inherited a disease from the parents. This is particularly useful for the detection of monogenic diseases, for example congenital adrenal hyperplasia (New et al. J Clin Endocrinol Metab 2014;99:E1022-30), beta-thalassemia (Lam et al. Clin Chem. 2012;58:1467-75) and hereditary muscular dystrophies (Genet Med 2015;17:889-96). If a monogenic disease is detected, various treatments can be performed, e.g., the pregnancy can be terminated, treatment provided before pregnancy, or after birth. For example, steroid treatment can be given prenatally to a pregnant woman confirmed of having a fetus affected by congenital adrenal hyperplasia to avoid abnormal sexual development.

VI. HAPLOTYPE DECONVOLUTION ANALYSIS FOR ANEUPLOIDY DETECTION

[0228] The haplotype deconvolution can also be used to detect a sequence imbalance of a chromosomal region of a fetus, such as aneuploidies, microdeletions, or microamplifications (e.g., microduplications). For example, a fractional contribution of a haplotype in one region can be compared to a fractional contribution of another haplotype in another region.

A. Mother

[0229] FIG. 15 shows a chromosomal aneuploidy detection based on haplotype deconvolution for maternal haplotypes according to embodiments of the present invention. In this illustration, the mother has two maternal haplotypes, namely Hap I and Hap II. For illustration purposes, we assume that 80% of her plasma DNA was derived from her own cells and 20% was derived from the placenta, which are example percentages in commonly measured ranges. This method can be generally applied for pregnancies with different fetal DNA percentages. Knowledge of the fetal DNA percentage is not required, but is simply provided for illustration, although a measurement of the fetal DNA percentage may be performed in various ways, e.g., using fetal-specific alleles or fetal-specific methylation markers.

[0230] The fetus has inherited Hap I and another haplotype from the father, namely Hap III. The placentally-derived DNA would exhibit the fetal genotypes, and a sequence imbalance can be detected by analyzing the fractional contribution resulting from the placentally-derived DNA.

[0231] As illustrated above, the fetal inheritance of the maternal haplotype can be determined through the deconvolution of the two maternal haplotypes. Analysis for the placental contribution to the maternal DNA can be performed for each of the two maternal haplotypes. The maternal haplotype inherited by the fetus (Hap I in this example) would have a much higher placental contribution compared with the maternal haplotype that is not inherited by the fetus (Hap II). The placental contribution for Hap I would be positively correlated with the fetal DNA fraction in the maternal plasma.

[0232] After determining which maternal haplotype has been inherited by the fetus, the dosage of the chromosome the fetus has inherited from the mother can be further determined through maternal haplotype deconvolution. In this illustration, two chromosomal regions are analyzed using maternal haplotype deconvolution. In one embodiment, the reference chromosome (RefChr) is a chromosome or a chromosomal region which is unlikely to be affected by a chromosomal aneuploidy. The reference chromosomal region is shown on the left side of FIG. 15. The target chromosome (TargetChr) is a chromosome or a chromosomal region which is potentially affected by a chromosomal aneuploidy. The target chromosomal region is shown on the right side of FIG. 15. The two regions can be for different regions of a same chromosome or for regions of two different chromosomes.

[0233] In the example shown, the fetus has been deduced to have inherited Hap I from the mother for both the reference chromosome and target chromosome through the methylation deconvolution of the Hap I and the Hap II at each region. Then, the placental contribution to maternal plasma DNA for Hap I can be compared between the reference chromosome and the target chromosome. If the placental contribution of Hap I for the target chromosomal region is significantly different than the placental contribution of Hap I for the reference chromosomal region (e.g., higher for amplification or lower for deletion), then a sequence imbalance can be identified.

[0234] For illustration purpose, we use the detection of trisomy as an example. However, other types of chromosome aneuploidies, including monosomy, amplification of a subchromosomal region or deletion of a subchromosomal region can also be detected using this method. For trisomy, the extra copy of the affected chromosome can be inherited from the father (denoted as Trisomy (F)) or the mother (denoted as Trisomy (M)). In over 90% of trisomy 21 cases, the extra

copy of chromosome 21 is derived from the mother (Driscoll et al. N Engl J Med 2009;360: 2556–2562). In the scenario of Trisomy (M), the placental contribution of Hap I for the target chromosome would be higher than that for the reference chromosome. In FIG. 15, Trisomy (M) is shown with two instances of Hap I, which would provide a higher placental contribution for the target region than the one instance of Hap I for the reference region.

[0235] Whether the placental contribution of Hap I for the target chromosome is higher than that for the reference chromosome can be determined by comparing a separation value between the two placental contributions and a threshold, which may be based on a separate measurement of the fetal DNA percentage. A higher fetal DNA percentage would result in a higher expected separation value between the two placental contributions, and thus the threshold can be set higher. For example, with the fetal DNA percentage being 20%, the placental contribution of Hap I for the reference region would be about 20% and the placental contribution of Hap I for the target region would be about 36.4%.

[0236] For instance, assume that 10 DNA molecules exist at the reference chromosome, then two of them are fetal and eight of them are maternal. For the two fetal DNA molecules, one is derived from Hap I and one is derived from Hap III. For the eight maternal DNA molecules, four are Hap I and four are Hap II. For the target region, there would be an extra DNA molecule of Hap I from the fetus. Thus there would be two fetal Hap I DNA molecules and 4 maternal Hap I DNA molecules total, providing 2/6=33.3%. The threshold value for the difference (e.g., 13.3%) can be placed between 0 and 13.3% to provide optimal specificity and sensitivity. A distribution of separation values can be determined from a reference group of samples. In the scenario of Euploid, the placental contributions would be approximately equal, e.g., the separation value would be less than the threshold. One skilled in the art will know how to select a suitable threshold based on the description herein and in U.S. Patent 8,467,976, and other references cited herein.

[0237] In one embodiment, the ratio (or other separation value) of placental contribution of Hap I between the target and reference chromosomes for a group of pregnant women, each known to be carrying a euploid fetus can be used as a reference interval. The ratio in the tested case can be compared with this reference group to determine if a significant elevation of placental contribution of Hap I is present for the target region relative to the reference region. In the example of 20% fetal DNA, the ratio would be 33.3/20 = 1.67. The ratio can be generalized to 2/(1+f), where f

represents the fetal DNA fraction. In another embodiment, the difference in the placental contribution of Hap I between the target and the reference chromosomes can be determined. This difference is then compared with a reference group.

B. Father

[0238] In another embodiment, haplotype deconvolution of the paternal haplotypes (Hap III and Hap IV) can be performed in the maternal plasma. The analysis of the paternal haplotypes can be performed in a similar manner as for the maternal haplotypes.

[0239] FIG. 16 shows a chromosomal aneuploidy detection based on haplotype deconvolution for paternal haplotypes according to embodiments of the present invention. In this illustration, the father has two paternal haplotypes, namely Hap III and Hap IV. As in FIG. 15, the fetus has inherited Hap I from the mother and Hap III from the father.

[0240] In the scenario where the extra copy of chromosome is derived from the father (Trisomy (F)), the placental contribution of Hap III would be higher for the target chromosome than for the reference chromosome. This is shown for the Trisomy (F) example, where two copies of Hap III are shown. As described above for the maternal haplotypes, a separation value between the placental contributions of Hap III for the target and reference regions can be compared to a threshold to determine whether an extra copy of Hap III exists for the target region. In various embodiments, a ratio or difference of the two placental contributions of the tested case can be compared with a reference group of pregnant women, each known to be carrying a euploid fetus so as to determine if the fetus has a chromosomal trisomy for the target chromosome, or amplification or deletion of a target chromosomal region. The threshold can be based on the separation values for the reference group of euploid fetuses, a reference group of aneuploidy fetuses, or both. A separate measure of fetal DNA percentage can also be used, as is described herein.

C. Method of detecting a sequence imbalance

[0241] FIG. 17 is a flowchart of a method 1700 for detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female according to embodiments of the present invention.

[0242] At block 1710, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1710 can be performed in a similar manner as block 1310 of method 1300 of FIG. 13.

[0243] At block 1720, a first target haplotype of a target chromosomal region of a first parental genome of a first parent of the fetus is determined, and a first reference haplotype of a reference chromosomal region of the first parental genome is determined. Block 1720 can be performed in a similar manner as block 1320 of FIG. 13. The target chromosomal region and the reference chromosomal region can be an entire chromosome or just part of a chromosome. Thus, the target chromosomal region can be a first chromosome and the reference chromosomal region can be a second chromosome different from the first chromosome. The first parent can be the mother or the father of the fetus.

[0244] The target chromosomal region can be selected based on various criteria. For example, a plurality of target regions can be selected, as may occur to test many non-overlapping regions of a specified size, such as 1 Mb, 5 Mb, 10 Mb, 20 Mb, 50 Mb, etc. As another example, the target chromosomal region can be selected based on a copy number analysis that identifies the region as having more DNA molecules than expected, e.g., as described in U.S. Patent Publications 2009/0029377 and 2011/0276277.

[0245] In some embodiments, it can be determined that the fetus has inherited the first target haplotype from the first parent and that the fetus has inherited the first reference haplotype from the first parent. The determination can include embodiments of FIG. 13 or FIG. 14. For example, determining that the fetus has inherited the first target haplotype from the first parent can include determining a second target fractional contribution of the fetal tissue type in the mixture corresponding to the second target haplotype, computing a second separation value between the first target fractional contribution and the second target fractional contribution, and determining that the fetus has inherited the first target haplotype from the first parent based on the second separation value.

[0246] At block 1730, a plurality of target heterozygous loci are identified of the target chromosomal region of the first parental genome. Each target heterozygous locus includes a corresponding first target allele in the first target haplotype and a corresponding second target allele in a second target haplotype of the first chromosomal region of the first parental genome.

Referring back to the example of FIG. 15, the target heterozygous loci have corresponding first target alleles of {G,T,A} on Hap I and have corresponding second target alleles of {A,G,C} on Hap II.

[0247] At block 1740, a target set of the plurality of cell-free DNA molecules is identified. Each cell-free DNA molecule of the target set is located at any one of the target heterozygous loci, includes a corresponding first target allele, and includes at least one of N genomic sites in the target chromosomal region. Block 1740 can be performed in a similar manner as described herein. For example, sequence reads can be mapped to a reference genome, where the target set of plurality of cell-free DNA molecules aligns to any one of the target heterozygous loci.

[0248] At block 1750, N first mixture methylation levels are measured at the N genomic sites using the target set of the plurality of cell-free DNA molecules. Block 1750 can be performed in a similar manner as block 1350 of FIG. 13.

[0249] At block 1760, a first fractional contribution of the fetal tissue type in the mixture is determined using the N first methylation levels. Block 1760 can be performed in a similar manner as block 1360 of FIG. 13.

[0250] At block 1770, a plurality of reference heterozygous loci are identified for the reference chromosomal region of the first parental genome. Each reference heterozygous locus includes a corresponding first reference allele in the first reference haplotype and a corresponding second reference allele in a second reference haplotype of the reference chromosomal region of the first parental genome. Referring back to the example of FIG. 15, the reference heterozygous loci have corresponding first target alleles of {A,T,C} on Hap I and have corresponding second target alleles of {T,C,A} on Hap II.

[0251] At block 1775, a reference set of the plurality of cell-free DNA molecules is identified. Each cell-free DNA molecule of the reference set is located at any one of the reference heterozygous loci, includes a corresponding first reference allele, and includes at least one of K genomic sites in the reference chromosomal region.

[0252] At block 1780, K reference mixture methylation levels are measured at the K genomic sites using the reference set of the plurality of cell-free DNA molecules.

[0253] At block 1785, a first reference fractional contribution of the fetal tissue type is determined in the mixture using the K reference methylation levels.

[0254] At block 1790, a first separation value is computed between the first target fractional contribution and the first reference fractional contribution.

[0255] At block 1795, the first separation value is compared to a threshold value to determine a classification of whether the fetus has a sequence imbalance for the target chromosomal region. If the first separation value exceeds the threshold value then a sequence imbalance can be identified. The threshold value can be determined as described above, e.g., based on separation values seen in a reference group of samples not having a sequence imbalance and/or a reference group of samples having the sequence imbalance. As examples, the classification can be positive, negative, or indeterminate for the sequence imbalance being tested.

[0256] Different threshold values can be used, depending on the type of sequence imbalance. For example, if the sequence imbalance is a deletion, then the first separation value would be expected to be a negative value. In such a case, the threshold value can be a negative number, and the comparison can determine that the first threshold value exceeds the threshold value by being a larger negative number. If the sequence imbalance being tested is an amplification, then it can be tested whether the separation value is greater than the threshold value. Thus, the threshold value used can be dependent on the type of sequence imbalance being tested.

VII. DECONVOLUTION OF SIGNATURES TO IDENTIFY DISEASED TISSUE

[0257] If a genomic signature (e.g., a particular SNP allele) is known, embodiments can determine which tissue is the origin of such signatures. As the cell-free DNA molecules exhibiting the signatures are from the tissue of origin, the tissue of origin can be identified from the fractional contributions determined using cell-free DNA molecules exhibiting the signatures. Thus, cell-free DNA molecules with a signature of a transplanted organ (e.g., a signature of a haplotype of the transplanted organ) can be used to monitor changes in amounts of cell-free DNA molecules from the transplanted organ with high sensitivity, e.g., given that a high fractional contribution of the DNA in the mixture would be from the transplanted organ. Examples are provided for transplants to show that the technique is accurate. In another example, a signature of a tumor can be used to identify tissues within which the tumor resides.

A. Organ Transplantation

[0258] As an example for organ transplantation, we analyzed the plasma of a patient who had received liver transplantation and a patient who had received bone marrow transplantation. For each case, the donor-specific SNP alleles were identified through the genotyping of the tissues from the patients and the donors. For the liver transplant recipient, a biopsy of the donor liver and the blood cells of the recipient were sequenced. For the bone marrow transplant case, the buccal swab (recipient genotype) and the blood cells (donor genotype) were sequenced. The plasma DNA samples were sequenced after bisulfite conversion. Sequenced DNA fragments carrying a donor-specific SNP allele and at least one CpG site were used for downstream methylation deconvolution analysis. A total of 72 million and 121 million reads were sequenced for the patients who had received liver and bone marrow transplantation, respectively. For the two cases, 38 and 5355 fragments were used for deconvolution analysis, respectively.

Tissue type	Liver transplant recipient	Bone marrow transplant recipient
Liver	45.4	4.4
Lung	0.0	1.5
Colon	29.3	6.3
Small intestines	0.0	1.8
Pancreas	0.0	0.0
Adrenal glands	0.0	0.0
Esophagus	0.0	0.0
Adipose tissues	0.0	14.8
Heart	0.0	0.0
Brain	14.5	9.6
T-cells	0.0	12.3
B-cells	5.9	16.6
Neutrophils	4.9	32.8

Table 6. Fractional contributions for different organs to plasma DNA fragments carrying donor-specific alleles in the two transplant recipients.

[0259] Table 6 shows methylation deconvolution analysis on plasma DNA fragments carrying donor-specific alleles in a liver transplant recipient and a bone marrow transplant recipient. The

numbers represent the percentage contribution of different tissues to the donor-specific plasma DNA fragments. For the liver transplant case, the liver was shown to be the most important contributor to these DNA fragments. For the bone marrow transplant case, the hematopoietic system (including the T-cells, B-cells and neutrophils) was the major contribution of the donor-specific DNA fragments. These results indicate that methylation deconvolution can accurately indicate the tissue origin of DNA fragments having single nucleotide alterations. A small amount of sequenced fragments were attributed to other tissues probably because of measurement imprecision as relatively small number of donor-specific fragments were used for the deconvolution analysis.

[0260] The fractional contributions for tissue associated with the transplanted organ can be determined in the above manner and monitored. With the baseline fractional contribution (an example of a reference fractional contribution) relatively high as a result of using only cell-free DNA molecules exhibiting the donor signature, small changes in total amount of donor DNA in plasma can be detected. Accordingly, methylation deconvolution analysis can be applied for the monitoring of organ transplantation.

[0261] As can be seen above for the liver transplant, methylation deconvolution is not absolutely specific. In this analysis, plasma DNA fragments carrying the donor-specific alleles were used for methylation deconvolution analysis. These fragments are specific for the donor and should be derived only from the liver in this liver transplant recipient. Therefore, the theoretical contribution of the liver should be 100%. Another possibility is that certain cell types are present in different types of tissues making the liver methylation profile overlapping with other tissues. For example, the connective tissue cells in the liver may also be present in other organs. But, relative percentages from other patients or other samples (e.g., at other times) of the instant patient can identify whether more cell-free DNA molecules are being released.

[0262] In various embodiments, the donor signature can correspond to a particular haplotype of the donor genome or both haplotypes in a chromosomal region. Methylation deconvolution can be performed using cell-free DNA molecules located on the particular donor haplotypes, and increases in the fractional contribution of the particular haplotype can be monitored. If a significant increase occurs (e.g., as measured by a percentage or absolute threshold), then a rejection of the transplanted organ can be identified.

[0263] FIG. 18 shows an illustration of haplotype deconvolution for organ transplant monitoring according to embodiments of the present invention. The donor has haplotypes labeled Hap I and Hap II, and the recipient has haplotypes labeled Hap III and Hap IV. The donor has a signature at locus 1 and locus 3, as the alleles are not found on the recipient haplotypes. Locus 2 and locus 4 do not have a donor signature. Thus, embodiments may use DNA molecules that are located at locus 1 and locus 3 as part of a deconvolution process.

[0264] The plasma DNA deconvolution can be used to determine whether the determined fractional contribution from the transplanted organ is at a baseline or increased relative to the baseline. In some embodiments, the fractional contributions can be determined for each of Hap I and Hap II separately, if different signatures exist; such different signatures can exist at different loci. In other embodiments, a single fractional contribution can be determined for both haplotypes, e.g., when they share a signature. In the example shown in FIG. 18, Hap I and Hap II do share a signature at locus 1 and locus 3.

[0265] Accordingly, the contribution of the transplanted organ can be determined using haplotype deconvolution. The increase in the contribution of the haplotype to the transplanted organ would be useful to indicate the increased contribution of the organ to the plasma DNA. In various embodiments, the baseline level can be determined from a cohort of transplant recipients not having rejection or from a cohort of transplant recipients having rejection. When using recipients having rejection, the baseline level can be determined as below those from a cohort of transplant recipients having rejection.

[0266] As mentioned above, the donor may have two identical haplotypes or the recipient can also have two identical haplotypes. Furthermore, the donor and recipient may share a haplotype. As long as the donor or the recipient has a unique haplotype, a change in a percentage of cell-free DNA molecules from the donor tissue can be determined. In the former, a rejection will be detected when one sees an increase in the contribution of the donor-unique haplotype in plasma (or other sample). In the latter, a rejection will be detected when one sees a decrease in the contribution of the recipient-unique haplotype in plasma.

[0267] Accordingly, some embodiments can use a first haplotype that is present in normal cells of the organism and not being present in abnormal cells that may be in the mixture. This would correspond to the latter example above, when the recipient has a unique haplotype. Another

example is when a patient has a unique haplotype in healthy cells compared to a tumor (e.g., previously found in the organism). In this embodiment, the first tissue type can be determined to have the disease state when the first separation value is less than the threshold value.

[0268] In some embodiments, if the transplanted organ is detected as being rejected, treatment can be provided. For example, a change is dosage of anti-rejection medication can be provided. As another example, a new organ can be obtained, and surgery can be performed to remove the old transplanted organ and put in the new transplanted organ.

B. Hepatocellular carcinoma (HCC)

[0269] As an example for determining a tissue of origin for a cancer signature or aberration (or monitoring for a tumor that was known to exist or have existed), we analyzed the plasma of an HCC patient. The tumor and the blood cells of the patients were sequenced to identify the cancer-specific single nucleotide mutations. Sequenced DNA fragments carrying a cancer-specific mutation and at least one CpG site were used for downstream methylation deconvolution analysis. A total of 11,968 fragments were used for deconvolution analysis. In addition to the methylation profiles from the normal tissue organs, we have also included the methylation profile of HCC tissues as candidate tissue of origin.

[0270] In another embodiment, more types of tumor tissues can be considered as candidate tissues for the mutations. In one embodiment, the methylation profiles of the common cancers, for example but not limited to colorectal cancer, lung cancer, breast cancer, pancreatic cancer, prostate cancer, bladder cancer, cervical cancer and ovarian cancer can be included as candidate tissues. In yet another embodiment, only the most possible cancers specific to the patient can be included in the analysis. For example, in female patients, the breast cancer, ovarian cancer, colorectal cancer and cervical cancer are considered. In yet another embodiment, the ethnic origin and the age are considered in the selection of the candidate tissues.

[0271] Table 7 shows methylation deconvolution of the plasma DNA fragments carrying cancer-associated mutations. The deconvolution analysis accurately determined that the DNA fragments carrying cancer-associated mutations are predominantly derived from the liver cancer tissues.

Tissue	Contribution (%)

Liver	0.0
Lung	0.0
Colon	0.0
Small intestines	0.0
Pancreas	0.0
Adrenal glands	0.0
Esophagus	0.0
Adipose tissues	0.0
Heart	0.0
Brain	0.0
T-cells	0.0
B-cells	0.0
Neutrophil	4.6
Liver cancer	95.4
Placenta	0.0

Table 7. Fractional contributions for HCC patient using cancer mutation.

[0272] In some embodiments, the tumor can initially be identified by detecting a copy number aberration, e.g., as described in U.S. Patent Nos. 8,741,811 and 9,121,069. The particular tissue of origin can be determined, e.g., as described in U.S. Patent Application 14/994,053 based on patterns of copy number aberrations previously identified in various tumors. Once the tumor has been identified, treatment can be performed, e.g., by surgery, radiotherapy, or chemotherapy. Either way, a biopsy can be obtained after the tissue of origin is determined. A cancer-specific point mutation can be determined from the biopsy or from DNA fragments in plasma (e.g., as described in U.S. patent Publication 2014/0100121, or other mixtures that are associated with the copy number aberration.

[0273] After treatment, a key change would be the disappearance of the genomic aberrations, including the copy number aberration and point mutation. When these aberrations are gone, the analysis of the genomic signature of the point mutation in the affected regions would give a change in the tissue contribution via the methylation deconvolution analysis. If the tumor comes back in the future, the cancer-associated changes in tissue composition (as determined using methylation deconvolution analysis) would be seen again. For example, the fractional contribution can be compared to a reference fractional contribution, and if a change is detected, then new courses of treatment can be provided.

[0274] In various embodiments, the cancer-specific mutation can be on only one haplotype or on both haplotypes, e.g., in a manner similar to the donor example above. Thus, as with the donor, the fractional contributions can be determined for each of Hap I and Hap II separately, if different signatures exist; such different signatures can exist at different loci. In other embodiments, a single fractional contribution can be determined for both haplotypes, e.g., when they share a signature.

C. Imprinting

[0275] In another embodiment, the haplotype deconvolution analysis can be applied for the analysis of the genomic regions showing tissue-specific imprinting. It has been shown that the differential methylation of the paternally and maternally inherited alleles in different tissue organs is a common phenomenon (Baran et al. Genome Res 2015;25:927-36). Haplotype deconvolution would be useful for the monitoring of the contribution of the organ exhibiting tissue-specific imprinting. For example, when the paternally and maternally inherited haplotypes have different methylation status in the liver but not in other tissues, methylation deconvolution can be performed on both the paternally and maternally inherited haplotypes. In one embodiment, both the paternal and maternal methylation patterns can be included as candidate tissues in the analysis.

D. Method using Genomic Signature

[0276] FIG. 19 is a flowchart illustrating a method 1900 of analyzing a biological sample of an organism to detect whether a first tissue type has a disease state associated with a first haplotype according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type. Method 1900 is performed at least partially using a computer system.

[0277] At block 1910, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 1910 can be performed using techniques described in block 140 of method 100 of FIG. 1. For example, at least 1,000 cell-free DNA molecules can be analyzed to determine where the cell-free DNA molecules are located, and methylation levels can be measured as described below. Further, the cell-free DNA molecules are analyzed to determine a respective allele of the cell-free DNA molecule. For example, an allele of a DNA molecule can be determined from a sequence read or from a particular probe that hybridizes to the DNA molecule.

[0278] At block 1920, one or more loci are identified. Each locus has a first allele on a first haplotype of a first chromosomal region. The first haplotype has a property of either: (1) not being present in healthy cells of the organism, but instead may be from a tumor or transplanted tissue, as examples; or (2) being present in normal cells of the organism and not being present in abnormal cells that may be in the mixture. Thus, the first haplotype has a genomic signature. In this manner, there is a difference between the healthy (normal) cells and the abnormal cells, thereby allowing embodiments to track a fractional contribution of one or the other, or both, so as to track an extent (e.g., fractional contribution) of the abnormal cells. With property (1), the first haplotype is associated with a disease state, e.g., cancer or a rejection of transplanted tissue. Thus, a particular cancer can have the first haplotype in a cancer genome of the particular cancer.

[0279] The one or more first alleles may be identified at the one or more loci on the first haplotype by obtaining a tissue sample (e.g., of the tumor or transplanted tissue) and analyzing DNA molecules of the tissue sample to determine the first haplotype. Such a tissue sample may be obtained from a biopsy, and method 1900 can be used to test if the cancer has metastasized to other tissues, or has recurred after surgery. Each one of the loci may be a heterozygous locus or a homozygous locus in the abnormal cells. For example, in FIG. 18, locus 1 and locus 3 are homozygous in the donor organ. But, ultimately, more than one allele would be observed in the plasma for all the loci, as each locus would have a signature for healthy cells or for abnormal cells. Thus, two haplotypes would exist across the tissue types, but a single tissue type might have only one haplotype in a region being analyzed.

[0280] At block 1930, a first set of the plurality of cell-free DNA molecules is identified. Each of the plurality of cell-free DNA molecules is located at any one of the loci from block 1920 and includes a corresponding first allele at the one locus, so that the cell-free DNA molecule can be identified as corresponding to the first haplotype. Each of the first set of cell-free DNA molecules also includes at least one of N genomic sites, where the genomic sites are used to measure the methylation levels. N is an integer, e.g., greater than or equal to 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000, or 5,000.

[0281] At block 1940, N first mixture methylation levels are measured at the N genomic sites using the first set of the plurality of cell-free DNA molecules. One first mixture methylation level can be measured for each of the N genomic sites. Block 1940 can be performed in a similar

manner as block 150 of method 100 of FIG. 1. In some embodiments, the measurement of the methylation level of a DNA molecule can use methylation-aware sequencing results, which may also be used to determine the location and respective allele of the DNA molecule.

[0282] At block 1950, a first fractional contribution of the first tissue type in the mixture is determined using the N first methylation levels. In some embodiments, block 1950 can be performed via blocks 160 and 170 of method 100 of FIG. 1. Thus, a fractional contribution can be determined simultaneously for a panel of M tissue types. Block 1950 may use N issue-specific methylation levels at N genomic sites, determined for each of M tissue types, e.g., as in block 120 of method 100 of FIG. 1.

[0283] At block 1960, a separation value between the first fractional contribution and a reference fractional contribution is computed. Examples of separation values are described herein. The reference fractional contribution can be determined using samples from organisms that are healthy for the first tissue type. For a transplant example, the reference fractional contribution can be determined from one or more measurements of biological samples of organisms whose transplanted first tissue is not being rejected.

[0284] At block 1970, the separation value can be compared to a threshold value to determine a classification of whether the first tissue type has a disease state. For example, if the first haplotype is associated with cancer, then an appreciable first fractional contribution indicates that the first tissue type has cancer, as can be measured by the separation value exceeding the threshold value (e.g., when the reference fractional contribution is zero). The amount that the first fractional contribution exceeds the threshold value can indicate a certain level of cancer. As another example, the first haplotype can be specific to transplanted tissue, and a high contribution relative to the reference can indicate the organism is rejecting the transplanted tissue.

[0285] In an embodiment where the first haplotype is present in normal cells of the organism and not present in abnormal cells that may be in the mixture, the first tissue type can be determined to have the disease state when the first separation value is less than the threshold value. An example of a disease state is preeclampsia which can be associated with a spectrum of pathologic changes in a fetal tissue such as the placenta. As an example, in such a situation, if the first haplotype is specific to the fetus, e.g. paternally inherited haplotype, it may be increased in the maternal plasma in a pregnancy complicated with preeclampsia.

[0286] In some embodiments, a second haplotype for the diseased tissue, e.g., the transplanted tissue or a tumor may also be used. Thus, a second fractional contribution can be computed and compared to the reference fractional contribution. Accordingly, a second set of the plurality of cell-free DNA molecules can each be located at any one of the one or more loci, include a corresponding second allele on a second haplotype of the first chromosomal region, and include at least one of the N genomic sites. The second haplotype would have the same property of being only from healthy cells or abnormal cells.

[0287] A plurality of tissue types can be tested (e.g., using method 100 of FIG. 1), so as to determine a tissue of origin of the first haplotype, e.g., when it is associated with cancer. Accordingly, fractional contributions of other tissue types in the mixture can be determined using the N first methylation levels, and corresponding separation values between the corresponding fractional contributions and respective reference fractional contribution can be compared to the threshold value to determine a classification of whether each of the other tissue types has the particular cancer. Different tissue could have different reference fractional contributions.

VIII. IDENTIFYING TISSUE OF ORIGIN OF CNA OF CANCER

[0288] In some embodiments, an origin of a tumor may not be known. Thus, it can be difficult to identify point mutations in a tumor, as may be used for method 1900 of FIG. 19 or other methods described herein. Additionally, a tumor may not have a significant number of point mutations, but may have chromosomal regions exhibiting amplifications and deletions (examples of copy number aberrations).

[0289] To address this problem, embodiments can use a copy number analysis to identify regions that exhibit a copy number aberration (CNA). Typically, a CNA occurs on only one haplotype of a region. As only one haplotype has an amplification or a deletion, there will be a relatively large difference between the fractional contributions of the tissue type within which the tumor resides.

[0290] The CNA analysis may be performed in a variety of ways, e.g., as described in U.S. Patent Nos. 8,741,811 and 9,121,069. For example, the human genome (or genome for other type of organism) can be partitioned into approximately 3,000 non-overlapping 1-Mb bins. The number of reads mapping to each 1-Mb bin can be determined. After correcting for GC bias (Chen EZ, et

al. (2011) *PLoS One* 6(7):e21791), the sequence read density of each bin can be calculated. For each bin, the sequence read density of the test case can be compared to the values of the reference control subjects. Copy number gains and losses may be defined as 3 standard deviations above and below, respectively, the mean of the controls. Accordingly, identifying a first chromosomal region as exhibiting a copy number aberration can be based on a first amount of cell-free DNA molecules that are located in the first chromosomal region.

[0291] To determine the tissue origin of copy number aberrations in plasma, plasma DNA tissue mapping can be performed using the methylation markers located within the genomic regions exhibiting such aberrations in plasma. In the examples below for the cancer patients, mapping of plasma DNA copy number aberrations was performed only in cases with aberrations affecting a contiguous chromosome region of at least 30 Mb so that a sufficient number of methylation markers could be used for mapping.

A. Identifying regions with copy number aberration (CNA)

[0292] A 62-year-old male patient with HCC was recruited from the Department of Surgery, Prince of Wales Hospital, Hong Kong with informed consent. Ten milliliters of venous blood were collected in EDTA tubes at diagnosis and 3 months after the resection of the tumor. The blood samples were centrifuged at 3000 g for 10 minutes to separate the blood cells from the plasma. The plasma was recentrifuged at 30000 g for 10 minutes to remove the remaining cells.

[0293] DNA extracted from the blood cells was used for phasing the SNPs to construct the haplotypes of the patient using the 10x genomics platform following the manufacturer's instruction. High molecular weight DNA was extracted from blood or tissue samples using MagAttract HMW DNA kit (QIagen, Germany). The quality of DNA was verified by Genomic DNA Analysis ScreenTape on a 4200 TapeStation system (Agilent, Germany). DNA was quantified by dsDNA HS Assay kit on a Qubit 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA). Sample indexing and library preparation were performed using the GemCode system and its associated reagents (10X Genomics, Pleasanton, CA) (Zheng et al. Nat Biotechnol. 2016 Mar;34:303-11). In brief, 1ng of DNA was inputted for GEM reactions in which individual DNA molecules were partitioned to introduce specific barcodes and extend the DNA. After GEM reactions, sequencing libraries were prepared according to the manufacturer's recommendations.

The libraries were quantified by qPCR using KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA). The normalized libraries were sequenced on a HiSeq 2500 sequencer (Illumina, San Diego, CA), with paired-end sequencing of 98-bp, 14-bp I5 and 8-bp I7 index reads. Sequencing results were analyzed using Long Ranger software suite (10X Genomics) so that all heterozygous SNPs were phased and the two haplotypes of the patient was determined.

[0294] The plasma samples were sequenced using the Illumina to a depth of 17x. Copy number aberrations were detected in the plasma of the HCC patient according to the method as previously described (Chan et al. Clin Chem. 2013;59:211-24).

[0295] FIG. 20 shows a plot of copy number aberrations detected in the plasma of an HCC patient according to embodiments of the present invention. The inner circle represents the result of the plasma sample collected at diagnosis (pre-operation) and the outer circle represents the result of the plasma sample collected at 3 months after the resection of the tumor (post-operation). Each dot represents a 1-Mb region. The green, red, and grey dots represent regions with copy number gain, copy number loss and no copy number change, respectively. Copy number aberrations were detected in the plasma sample at diagnosis and these changes disappeared after the tumor was removed.

[0296] In FIG. 20, two regions are highlighted for having a CNA. Region 2010 has a copy number gain, and region 2020 has a copy number loss. The haplotypes of these regions can be determined using any tissue sample of the subject, and not just a tumor sample. The difference in copy number is what is driving the difference in the fractional contributions, and that difference should be greatest in the tissue type with the tumor.

B. Determining the tissue origin of the copy number aberrations

[0297] We performed methylation deconvolution analysis for the two haplotypes independently. For illustration purposes, the two haplotypes are named Hap I and Hap II. The plasma DNA molecules covering heterozygous SNPs and at least one CpG site were used for this analysis. Plasma DNA molecules carrying the SNP alleles on Hap I were analyzed independently from those carrying alleles on Hap II. The methylation status of the CpG sites were used for methylation deconvolution for molecules mapped to Hap I and Hap II independently. As a result, the tissue contribution to Hap I and Hap II in plasma DNA could be determined.

[0298] First we focused on the regions with amplification. For illustration purposes, we analyzed the amplified region on chromosome 1q as an example.

	At diagnosis	After tumor resection
Hap I	34,119	11,131
Hap II	26,582	11,176

[0299] Table 11 shows the number of sequence reads from the two haplotypes. At diagnosis, the number of reads mapped to Hap I was increased compared with the number of reads mapped to Hap II. This indicated that Hap I is amplified relative to Hap II. This observation is compatible with the fact that a particular chromosome is duplicated in cancer rather than both homologous chromosomes being amplified to the same extent, which was in line with the fact that the copy number aberrations occurs preferentially on one haplotype (Adey A. et al, Nature. 2013;500:207-11;LaFramboise T. et al, PLoS Comput Biol. 2005;1(6):e65). The difference in the dosage of the two haplotypes disappeared after the tumor was resected. The difference in the absolute sequence reads number between the plasma samples taken at diagnosis and after tumor resection was due to the difference in the total number of sequence reads generated for the two plasma samples.

	At diagnosis			After tumor resection		
	Hap I	Hap II	Difference	Hap I	Hap II	Difference
liver	19.7	8.0	11.7	21.3	21.9	-0.6
Lung	5.4	0	5.4	0	0	0
Colon	0	0	0	0	0	0
Brain	0	0	0	9.0	9.0	0
Heart	0	17.0	-17	3.0	2.5	-0.5
Blood cell	74.9	75.0	0	66.7	66.6	0.1
Total	100	100	0	100	100	0

[0300] Table 12 shows a percentage contribution of different tissues to plasma DNA for the two haplotypes at diagnosis and after tumor resection. At diagnosis, the contributions of liver to plasma DNA were 19.7% and 8.0% for Hap I and Hap II, respectively. A difference of 11.7% was the highest among different types of tissue. This indicated that the dosage difference between Hap I and Hap II in the plasma was most likely contributed from the contribution of the liver. This further indicated that the likely origin of the chromosomal aberration was from the liver because the copy number changes was most likely due to the duplication of Hap I in the sequence read

counts analysis. In another embodiment, the difference in the contribution for Hap I and Hap II can be ranked to indicate the relative likelihood of different tissues being the source of the copy number aberrations.

[0301] The value for the heart is -17, which is in the opposite direction of the copy number aberration identified by Table 11. Thus, although the absolute value for the heart is larger than the absolute value for the liver, the opposite sign would discount the heart as being a viable candidate for the tissue type of the origin of the tumor. As the total contribution of all organs is 100%, the positive difference in the contributions of liver results in other tissues having negative values.

[0302] Similarly, this haplotype-specific methylation deconvolution can also be performed on regions with copy number loss. For illustration purpose, we performed this analysis on a region on chromosome 1p that exhibited copy number loss.

	At diagnosis	After tumor resection
Hap I	19,973	8,323
Hap II	12,383	7,724

[0303] Table 13 shows a number of sequence reads from the two haplotypes. At diagnosis, the number of reads mapped to Hap II was decreased compared with the number of reads mapped to Hap I. In tumor tissues, most of the regions with chromosome copy number loss would only involve the deletion of one of the two chromosomes. Thus the relative reduction in the dosage of Hap II was compatible with the deletion of Hap II. The difference in the dosage of the two haplotypes disappeared after the tumor was resected indicated that the amount of tumor-derived DNA had decreased or disappeared from the plasma.

	At diagnosis			After tumor resection		
	Hap I	Hap II	Difference (Hap I – Hap II)	Hap I	Hap II	Difference (Hap I – Hap II)
liver	13.3	5.5	7.8	10.2	13.2	-3
Lung	0	0	0	4.1	0.5	3.6
Colon	3.8	0	3.8	8.6	17.5	-8.9
Brain	0	0	0	0	0	0
Heart	3.7	0	3.7	25.5	19.4	6.1
Blood cell	79.2	94.5	-15.3	51.6	49.4	2.2
Total	100	100	0	100	100	0

[0304] Table 14 shows a percentage contribution of different tissues to plasma DNA for the two haplotypes at diagnosis and after tumor resection. At diagnosis, the contributions of liver to plasma DNA were 13.3% and 5.5% for Hap I and Hap II, respectively. A difference of 7.8% was the highest among different types of tissue. This indicated that the dosage difference between Hap I and Hap II in the plasma was most likely contributed from the contribution of the liver. This further indicated that the likely origin of the chromosomal aberration was from the liver because the copy number changes was most likely due to the deletion of Hap II in the sequence read counts analysis. In another embodiment, the difference in the contribution for Hap I and Hap II can be ranked to indicate the relative likelihood of different tissues being the source of the copy number aberrations.

C. Method of determining tissue origin of tumor

[0305] FIG. 21 is a flowchart illustrating a method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration according to embodiments of the present invention. The biological sample includes a mixture of cell-free DNA molecules from a plurality of tissues types that include a first tissue type.

[0306] At block 2110, a plurality of cell-free DNA molecules from the biological sample are analyzed. Block 2110 can be performed using techniques described in block 1910 of FIG. 1 and block 140 of method 100 of FIG. 1, as well as other blocks describing similar features.

[0307] At block 2115, a first chromosomal region is identified as exhibiting a copy number aberration in the organism based on a first amount of cell-free DNA molecules that are located in the first chromosomal region. As an example, plasma DNA analysis is performed to identify regions that exhibit copy number aberrations. The aberration can correspond to over or under-representation. In some embodiments, the genome can be separated into bins (e.g., 1-Mb bins), and the amount of cell-free DNA molecules from a particular bin can be determined (e.g., by mapping sequence reads to that part of a reference genome). The amount for a particular bin can be normalized (e.g., with respect to an average amount for a bin) can an over or under-representation can be identified.

[0308] Other techniques than counting DNA molecules mapping to a particular region can be used. For example, a distribution of sizes of the DNA molecules aligning to the first chromosomal

region can be used to detect the CNA. For example, cell-free tumor DNA is smaller than cell-free DNA from normal cells. This difference in size can be used to detect differences in a size distribution (e.g., average size or ratio of number of DNA molecules at different sizes) between two haplotypes for the region, or between the region and another region.

[0309] At block 2120, a first haplotype and a second haplotype of the organism in the first chromosomal region are determined. The two haplotypes may have been determined as part of block 2115. The two haplotypes can be determined using the same cell-free mixture or from a different sample, e.g., a cellular sample.

[0310] At block 2130, one or more heterozygous loci of the first chromosomal region are identified. Each heterozygous locus includes a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype. Block 2130 may be performed in a similar manner as other similar blocks of methods described herein.

[0311] At block 2140, a first set of the plurality of cell-free DNA molecules are identified. Each DNA molecule of the first set is located at any one of the one or more heterozygous loci, includes the corresponding first allele of the heterozygous locus, and includes at least one of N genomic sites. N is an integer greater than or equal to 2. Block 2140 may be performed in a similar manner as other similar blocks of methods described herein.

[0312] At block 2150, N first mixture methylation levels at the N genomic sites are measured using the first set of the plurality of cell-free DNA molecules. Block 2150 may be performed in a similar manner as other similar blocks of methods described herein.

[0313] At block 2160, a second set of the plurality of cell-free DNA molecules is identified. Each DNA molecule of the second set is located at any one of the one or more heterozygous loci, includes the corresponding second allele of the heterozygous locus, and includes at least one of the N genomic sites. Block 2160 may be performed in a similar manner as other similar blocks of methods described herein.

[0314] In some embodiments, a first number of cell-free DNA molecules in the first set of the plurality of cell-free DNA molecules can be determined, and a second number of cell-free DNA molecules in the second set of the plurality of cell-free DNA molecules can be determined, e.g., as shown in table 11. It can be determined which number is higher, thereby providing information

about the expected separation value for the tissue of origin, e.g., which haplotype should have a higher fractional contribution.

[0315] The first set of the plurality of cell-free DNA molecules can have a first size distribution, and the second set of the plurality of cell-free DNA molecules can have a second size distribution. A statistical value of a size distribution of the DNA molecules can be determined for each haplotype, thereby providing a first statistical value and a second statistical value. The haplotype with the smaller size distribution would be expected to have a higher copy number than the other haplotype, as tumor cell-free DNA is known to be smaller, as described in U.S. Patent No. 8,741,811. Examples of a statistical value of a size distribution are a ratio of a number of DNA molecules at different sizes, an average size, or a percentage of the DNA molecules at a specific size (e.g., below a size cutoff).

[0316] At block 2170, N second mixture methylation levels at the N genomic sites are measured using the second set of the plurality of cell-free DNA molecules. Block 2170 may be performed in a similar manner as other similar blocks of methods described herein.

[0317] Blocks 2180 and 2190 may be performed for each of a plurality of M tissue types. The M tissue types can include a default list of tissue types that are screened and for which reference methylation levels may be known. The default list can include tissues that cancer is most predominantly seen. M is an integer greater than one.

[0318] At block 2180, a computer system determines a corresponding first fractional contribution of the tissue type in the mixture using the N first methylation levels. The computer system determines a corresponding second fractional contribution of the tissue type in the mixture using the N second methylation levels. Block 2180 may be performed in a similar manner as other similar blocks of methods described herein.

[0319] At block 2190, a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution is computed. Various separation values can be used, e.g., as are described herein.

[0320] At block 2195, the first tissue type is identified as being an origin of the copy number aberration based on a first separation value of the first tissue type having a maximal value among the corresponding separate values. The determination can require that the highest separation value

be sufficiently higher than the second highest separation value. For example, the difference can be required to be at least a threshold value, e.g., 1%, 2%, 35, 4%, 5%, 6%, or 7%. In one implementation, a difference between the first separation value and a next highest separation value can be compared to a threshold to determine a classification of how likely the first tissue type is the origin of the copy number aberration. Thus, even if the difference is not above the threshold, a probability or other classification can be provided. For instance, a linear relationship can be used from 0 to the threshold, where the probability is 100% once the difference is equal to the threshold.

[0321] Depending on how the separation value is determined, the maximal value can be a maximum negative number or a maximum positive number. For example, the difference values in table 14 could be determined using Hap II – Hap I. Whether the maximum should be a positive or a negative value can be determined using the analysis of DNA molecules on each haplotype, e.g., a count as in Table 13 or a size analysis as described above. In some implementations, the separation value can always be determined such that a maximum positive value is expected, e.g., by subtracting the fractional contribution of the haplotype with a lower copy number from the fractional contribution of the haplotype with the higher copy number.

[0322] After the origin is identified, an investigation using imaging modalities, e.g. computed tomography (CT) scan or magnetic resonance imaging (MRI), of the subject (entire subject or specifically of the candidate organ) can be performed to confirm or rule out the presence of a tumor in the organ. If presence of a tumor is confirmed, treatment can be performed, e.g., surgery (by a knife or by radiation) or chemotherapy.

IX. COMPUTER SYSTEM

[0323] Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in FIG. 22 in computer apparatus 10. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a computer system can include multiple computer apparatuses, each being a subsystem, with internal components. A computer system can include desktop and laptop computers, tablets, mobile phones and other mobile devices.

[0324] The subsystems shown in FIG. 22 are interconnected via a system bus 75. Additional subsystems such as a printer 74, keyboard 78, storage device(s) 79, monitor 76, which is coupled to display adapter 82, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 71, can be connected to the computer system by any number of means known in the art such as input/output (I/O) port 77 (e.g., USB, FireWire®). For example, I/O port 77 or external interface 81 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer system 10 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 75 allows the central processor 73 to communicate with each subsystem and to control the execution of a plurality of instructions from system memory 72 or the storage device(s) 79 (e.g., a fixed disk, such as a hard drive, or optical disk), as well as the exchange of information between subsystems. The system memory 72 and/or the storage device(s) 79 may embody a computer readable medium. Another subsystem is a data collection device 85, such as a camera, microphone, accelerometer, and the like. Any of the data mentioned herein can be output from one component to another component and can be output to the user.

[0325] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 81 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server, where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0326] Aspects of embodiments can be implemented in the form of control logic using hardware (e.g. an application specific integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As used herein, a processor includes a single-core processor, multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

[0327] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language

such as, for example, Java, C, C++, C#, Objective-C, Swift, or scripting language such as Perl or Python using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission. A suitable non-transitory computer readable medium can include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0328] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different computer products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0329] Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, units, circuits, or other means for performing these steps.

[0330] The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention. However, other

embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

[0331] The above description of example embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above.

[0332] A recitation of "a", "an" or "the" is intended to mean "one or more" unless specifically indicated to the contrary. The use of "or" is intended to mean an "inclusive or," and not an "exclusive or" unless specifically indicated to the contrary. Reference to a "first" component does not necessarily require that a second component be provided. Moreover reference to a "first" or a "second" component does not limit the referenced component to a particular location unless expressly stated.

[0333] All patents, patent applications, publications, and descriptions mentioned herein are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

WHAT IS CLAIMED IS:

1. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying one or more heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining, by the computer system, a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

computing a first separation value between the first fractional contribution and the second fractional contribution;

determining the portion of fetal genome at the one or more heterozygous loci based on the first separation value.

2. The method of claim 1, wherein the one or more heterozygous loci are a first plurality of heterozygous loci.

3. The method of claim 2, wherein the first haplotype is a first maternal haplotype, and wherein the second haplotype is a second maternal haplotype.

4. The method of claim 3, further comprising:

identifying a second plurality of heterozygous loci of the first chromosomal region in a paternal genome, each of the second plurality of heterozygous loci including a corresponding third allele in a first paternal haplotype and a corresponding fourth allele in a second paternal haplotype, wherein the paternal genome corresponds to the father of the fetus;

identifying a third set of the plurality of cell-free DNA molecules that each:

is located at any one of the second plurality of heterozygous loci,

includes the corresponding third allele of the heterozygous locus, and

includes at least one of K genomic sites;

measuring K third mixture methylation levels at the K genomic sites using the third set of the second plurality of cell-free DNA molecules;

determining a third fractional contribution of the fetal tissue type in the mixture using the K third methylation levels;

identifying a fourth set of the plurality of cell-free DNA molecules that each:

is located at any one of the second plurality of heterozygous loci,

includes the corresponding fourth allele of the heterozygous locus, and

includes at least one of the K genomic sites;

measuring K fourth mixture methylation levels at the K genomic sites using the fourth set of the second plurality of cell-free DNA molecules;

determining a fourth fractional contribution of the fetal tissue type in the mixture using the K fourth methylation levels;

computing a second separation value between the third fractional contribution and the fourth fractional contribution;

determining the portion of fetal genome at the second plurality of heterozygous loci based on the second separation value.

5. The method of claim 2, wherein the first separation value is a ratio of the first fractional contribution and the second fractional contribution, and wherein the portion of the fetal genome is determined to have the first haplotype and the second haplotype when the ratio is equal to one within a cutoff value.

6. The method of claim 2, wherein the first separation value is a difference of the first fractional contribution and the second fractional contribution.

7. The method of claim 2, wherein the portion of the fetal genome is determined to have one or more copies of the first haplotype and no copies of the second haplotype when the first separation value is greater than a threshold value.

8. The method of claim 2, wherein the portion of the fetal genome is determined to have one or more copies of the second haplotype and no copies of the first haplotype when the first separation value is less than a threshold value.

9. The method of claim 1, wherein N is 10 or more.

10. The method of claim 1, wherein the N mixture methylation levels form a methylation vector b, and wherein determining the first fractional contribution of the fetal tissue type includes:

for each of M tissue types:

obtaining N tissue-specific methylation levels at the N genomic sites, N being greater than or equal to M, wherein the tissue-specific methylation levels form a matrix A of dimensions N by M, the M tissue types including the fetal tissue type;

solving for a composition vector x that provides the methylation vector b for the matrix A; and

for each of one or more components of the composition vector x:

using the component to determine a corresponding fractional contribution of a corresponding tissue type of the M tissue types in the mixture.

11. The method of claim 10, wherein the M tissue types include liver, lung, neutrophils, lymphocytes, erythroblasts, heart, large intestines, small intestines, and placenta, wherein the fetal tissue type is the placenta.

12. The method of claim 1, wherein measuring N first mixture methylation levels at the N genomic sites includes analyzing methylation-aware sequencing results, and wherein the locations of the plurality of cell-free DNA molecules are determined using the methylation-aware sequencing results.

13. The method of claim 1, wherein the first separation value includes a ratio of the first fractional contribution and the second fractional contribution.

14. The method of claim 1, wherein the first separation value includes a difference of the first fractional contribution and the second fractional contribution.

15. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying one or more heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in a second haplotype of the first chromosomal region of the first parental genome;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels; and

comparing the first fractional contribution to a reference value to determine whether the fetus inherited the first haplotype at the first chromosomal region, wherein the fetus inherited the first haplotype at the first chromosomal region when the first fractional contribution exceeds the reference value.

16. The method of claim 15, wherein the reference value is determined by:

determining the second haplotype of the first chromosomal region of the first parental genome;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

measuring a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

determining the reference value as a sum of the second fractional contribution and a threshold value.

17. The method of claim 15, further comprising:

determining the second haplotype of the first chromosomal region of the first parental genome;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels;

comparing the second fractional contribution to the reference value to determine whether the fetus inherited the second haplotype at the first chromosomal region, wherein the fetus inherited the second haplotype at the first chromosomal region when the second fractional contribution exceeds the reference value.

18. A method of determining a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first haplotype and a second haplotype of a first chromosomal region of a first parental genome of a first parent of the fetus;

identifying a plurality of heterozygous loci of the first chromosomal region of the first parental genome, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the plurality of heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites;

measuring a first mixture methylation level using the first set of the plurality of cell-free DNA molecules;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the plurality of heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring a second mixture methylation level using the second set of the plurality of cell-free DNA molecules; and

determining which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower.

19. The method of claim 18, wherein the first mixture methylation level is a methylation density of the first set of the plurality of cell-free DNA molecules, and wherein the second mixture methylation level is the methylation density of the second set of the plurality of cell-free DNA molecules.

20. The method of claim 18, wherein determining which of the first haplotype and the second haplotype is inherited includes:

determining a separation value between the first mixture methylation level and the second mixture methylation level; and

comparing the separation value to a threshold value.

21 . The method of claim 20, wherein the threshold value is determined using Student's t-test, Mann-Whitney rank-sum test, or Chi-square test.

22. The method of claim 18, wherein determining the first haplotype and the second haplotype of the first chromosomal region of the first parental genome includes:

determining genotypes of the first parent genome at the plurality of heterozygous loci using a sample from the first parent;

obtaining a plurality of reference haplotypes; and

deducing the first haplotype and the second haplotype using the genotypes and the plurality of reference haplotypes.

23. A method of detecting a sequence imbalance in a portion of a fetal genome of an unborn fetus of a pregnant female using a biological sample from the pregnant female, wherein the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including maternal tissue types and a fetal tissue type, the fetus having a father and a mother being the pregnant female, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference human genome; and

determining a respective allele of the cell-free DNA molecule;

determining a first target haplotype of a target chromosomal region of a first parental genome of a first parent of the fetus;

identifying a plurality of target heterozygous loci of the target chromosomal region of the first parental genome, each target heterozygous locus including a corresponding first target allele in the first target haplotype and a corresponding second target allele in a second target haplotype of the target chromosomal region of the first parental genome;

identifying a target set of the plurality of cell-free DNA molecules that each:

is located at any one of the target heterozygous loci,

includes the corresponding first target allele of the target heterozygous locus, and

includes at least one of N genomic sites in the target chromosomal region, N being an integer greater than or equal to 2;

measuring N target mixture methylation levels at the N genomic sites using the target set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first target fractional contribution of the fetal tissue type in the mixture using the N target methylation levels; and

determining a first reference haplotype of a reference chromosomal region of the first parental genome, the reference chromosomal region being different than the target chromosomal region;

identifying a plurality of reference heterozygous loci of the reference chromosomal region of the first parental genome, each reference heterozygous locus including a corresponding first reference allele in the first reference haplotype and a corresponding second reference allele in a second reference haplotype of the reference chromosomal region of the first parental genome;

identifying a reference set of the plurality of cell-free DNA molecules that each:

is located at any one of the reference heterozygous loci,

includes the corresponding first reference allele of the reference heterozygous locus, and

includes at least one of K genomic sites in the reference chromosomal region;

measuring K reference mixture methylation levels at the K genomic sites using the reference set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first reference fractional contribution of the fetal tissue type in the mixture using the K reference methylation levels;

computing a first separation value between the first target fractional contribution and the first reference fractional contribution; and

comparing the first separation value to a threshold value to determine a classification of whether the fetus has the sequence imbalance for the target chromosomal region.

24. The method of claim 23, wherein the sequence imbalance is an amplification of the target chromosomal region.

25. The method of claim 23, wherein the threshold value used is dependent on a type of the sequence imbalance being tested.

26. The method of claim 23, further comprising:

measuring a fetal DNA percentage in the biological sample; and

using the fetal DNA percentage to determine the threshold value.

27. The method of claim 23, further comprising:

determining the threshold value using a reference group of samples, all of which have the sequence imbalance or do not have the sequence imbalance.

28. The method of claim 23, further comprising:

determining that the fetus has inherited the first target haplotype from the first parent; and

determining that the fetus has inherited the first reference haplotype from the first parent.

29. The method of claim 28, wherein determining that the fetus has inherited the first target haplotype from the first parent includes:

determining a second target fractional contribution of the fetal tissue type in the mixture corresponding to the second target haplotype;

computing a second separation value between the first target fractional contribution and the second target fractional contribution;

determining that the fetus has inherited the first target haplotype from the first parent based on the second separation value.

30. The method of claim 23, wherein the target chromosomal region is a first chromosome and the reference chromosomal region is a second chromosome different from the first chromosome.

31. The method of claim 23, wherein the first parent is the mother of the fetus.

32. A method of analyzing a biological sample of an organism, the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference genome corresponding to the organism;

determining a respective allele of the cell-free DNA molecule;

identifying one or more loci, each having a corresponding first allele on a first haplotype of a first chromosomal region, wherein the first haplotype is:

not present in healthy cells of the organism, or

present in the healthy cells of the organism and not present in abnormal cells that may be in the mixture;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one locus of the one or more loci,

includes the corresponding first allele of the one locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

determining, by the computer system, a first fractional contribution of the first tissue type in the mixture using the N first methylation levels;

determining a first separation value between the first fractional contribution and a reference fractional contribution; and

comparing the first separation value to a threshold value to determine a classification of whether the first tissue type has a disease state.

33. The method of claim 32, wherein the first haplotype is on a first chromosome, and wherein the first tissue type has the first haplotype on both copies of the first chromosome.

34. The method of claim 32, wherein the first haplotype is not present in the healthy cells of the organism, and wherein the first haplotype is associated with the disease state.

35. The method of claim 34, wherein the first tissue type is determined to have the disease state when the first separation value is greater than the threshold value.

36. The method of claim 34, further comprising:

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more loci;

includes a corresponding second allele on a second haplotype of the first chromosomal region, the second haplotype not being present in the healthy cells of the organism, and the second haplotype associated with the disease state; and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

determining, by the computer system, a second fractional contribution of the first tissue type in the mixture using the N second methylation levels;

determining a second separation value between the second fractional contribution and the reference fractional contribution; and

comparing the second separation value to the threshold value as part of determining the classification of whether the first tissue type has the disease state.

37. The method of claim 34, wherein the disease state is cancer.

38. The method of claim 37, wherein a particular cancer has the first haplotype in a cancer genome of the particular cancer.

39. The method of claim 38, wherein the particular cancer is selected from a group consisting of liver cancer, lung cancer, pancreas cancer, atrium cancer, colon cancer, sigmoid colon cancer, transverse colon cancer, ascending colon cancer, descending colon cancer, adrenal gland cancer, esophagus cancer, small intestines cancer, and CD4 T cell cancer.

40. The method of claim 37, wherein the classification of whether the first tissue type has cancer comprises classification of the presence or absence of a cancer, classification of a stage of cancer, classification of a size of a tumor, and/or classification of metastasis.

41. The method of claim 37, further comprising performing, by the computer system:

determining a plurality of corresponding fractional contributions of other tissue types in the mixture using the N first mixture methylation levels;

determining corresponding separation values between the plurality of corresponding fractional contributions and corresponding reference fractional contributions; and

comparing the corresponding separation values to the threshold value to determine classifications of whether each of the other tissue types has the particular cancer.

42. The method of claim 34, wherein the corresponding first allele at each of the one or more loci is a cancer-specific mutation.

43. The method of claim 32, further comprising:

determining the corresponding first allele at each of the one or more loci on the first haplotype by:

analyzing DNA molecules of a tissue sample having the first haplotype to determine the first haplotype.

44. The method of claim 32, wherein the disease state is a rejection of a transplant of the first tissue type in the organism.

45. The method of claim 44, wherein the reference fractional contribution is determined from one or more measurements of biological samples of organisms with the first tissue type being healthy.

46. The method of claim 44, wherein the reference fractional contribution is determined from one or more measurements of biological samples of organisms whose transplanted first tissue is not being rejected.

47. The method of claim 32, wherein the reference fractional contribution is zero.

48. The method of claim 32, wherein the N first mixture methylation levels form a methylation vector b, and wherein determining the first fractional contribution of the first tissue type includes:

for each of M tissue types:

obtaining N tissue-specific methylation levels at the N genomic sites, N being greater than or equal to M, wherein the tissue-specific methylation levels form a matrix A of dimensions N by M, the M tissue types including the first tissue type, M being an integer greater than one;

solving for a composition vector x that provides the methylation vector b for the matrix A

for each component of one or more components of the composition vector x:

using the component to determine a corresponding fractional contribution of a corresponding tissue type of the M tissue types in the mixture.

49. The method of claim 32, wherein the first haplotype is present in the healthy cells of the organism and not present in the abnormal cells that may be in the mixture.

50. The method of claim 32, wherein the first tissue type is determined to have the disease state when the first separation value is less than the threshold value.

51. The method of claim 49, wherein the abnormal cells are from a tumor.

52. The method of claim 49, wherein the abnormal cells are from donor tissue.

53. A method of analyzing a biological sample of an organism to identify an origin of a chromosomal aberration, the biological sample including a mixture of cell-free DNA molecules from a plurality of tissues types, including a first tissue type, the method comprising:

analyzing, by a computer system, a plurality of cell-free DNA molecules from the biological sample, the plurality of cell-free DNA molecules being at least 1,000 cell-free DNA molecules, wherein analyzing a cell-free DNA molecule includes:

identifying a location of the cell-free DNA molecule in a reference genome corresponding to the organism;

determining a respective allele of the cell-free DNA molecule;

identifying a first chromosomal region as exhibiting a copy number aberration in the organism based on a first amount of cell-free DNA molecules that are located in the first chromosomal region;

determining a first haplotype and a second haplotype of the organism in the first chromosomal region;

identifying one or more heterozygous loci of the first chromosomal region, each heterozygous locus including a corresponding first allele in the first haplotype and a corresponding second allele in the second haplotype;

identifying a first set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding first allele of the heterozygous locus, and

includes at least one of N genomic sites, N being an integer greater than or equal to 2;

measuring N first mixture methylation levels at the N genomic sites using the first set of the plurality of cell-free DNA molecules;

identifying a second set of the plurality of cell-free DNA molecules that each:

is located at any one of the one or more heterozygous loci,

includes the corresponding second allele of the heterozygous locus, and

includes at least one of the N genomic sites;

measuring N second mixture methylation levels at the N genomic sites using the second set of the plurality of cell-free DNA molecules;

for each tissue type of M tissue types:

determining, by the computer system, a corresponding first fractional contribution of the tissue type in the mixture using the N first methylation levels, M being an integer greater than one;

determining, by the computer system, a corresponding second fractional contribution of the tissue type in the mixture using the N second methylation levels;

computing a corresponding separation value between the corresponding first fractional contribution and the corresponding second fractional contribution;

identifying the first tissue type as being the origin of the copy number aberration based on a first separation value of the first tissue type having a maximal value among the corresponding separate values.

54. The method of claim 53, further comprising:

determining a difference between the first separation value and a next highest separation value; and

comparing the difference to a threshold to determine a classification of how likely the first tissue type is the origin of the copy number aberration.

55. The method of claim 53, wherein identifying the first chromosomal region as exhibiting the copy number aberration includes:

comparing the first amount of cell-free DNA molecules that are located in the first chromosomal region to a reference value determined from amounts of cell-free DNA molecules that are located in other chromosomal regions.

56. The method of claim 53, wherein the copy number aberration is a copy number gain.

57. The method of claim 53, further comprising:

determining a first number of cell-free DNA molecules in the first set of the plurality of cell-free DNA molecules;

determining a second number of cell-free DNA molecules in the second set of the plurality of cell-free DNA molecules; and

using the first number and the second number to determine whether the maximal value is to be positive or negative.

58. The method of claim 53, wherein analyzing the cell-free DNA molecule includes: determining a size of the cell-free DNA molecule, wherein the first set of the plurality of cell-free DNA molecules has a first size distribution, and wherein the second set of the plurality of cell-free DNA molecules has a second size distribution, the method further comprising:

determining a first statistical value of the first size distribution;

determining a second statistical value of the second size distribution; and

using the first statistical value and the second statistical value to determine whether the maximal value is to be positive or negative.

59. The method of claim 58, further comprising:

determining which of the first haplotype and the second haplotype has a higher copy number using the first statistical value and the second statistical value; and

computing the corresponding separation values such that the first separation value is positive.

60. The method of any one of claims 1-59, wherein analyzing the cell-free DNA molecules comprises bisulfite sequencing, sequencing preceded by methylation-sensitive restriction enzyme digestion, immunoprecipitation using anti-methylcytosine antibody or methylation binding protein, or single molecule sequencing that allows elucidation of the methylation status.

61. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform an operation of any method of claims 1-59.

62. A system comprising:

the computer product of claim 61; and

one or more processors for executing instructions stored on the computer readable medium.

63. A system comprising means for performing any method of claims 1-59 .

64. A system configured to perform any method of claims 1-59.

65. A system comprising modules that respectively perform the steps of any method of claims 1-59.

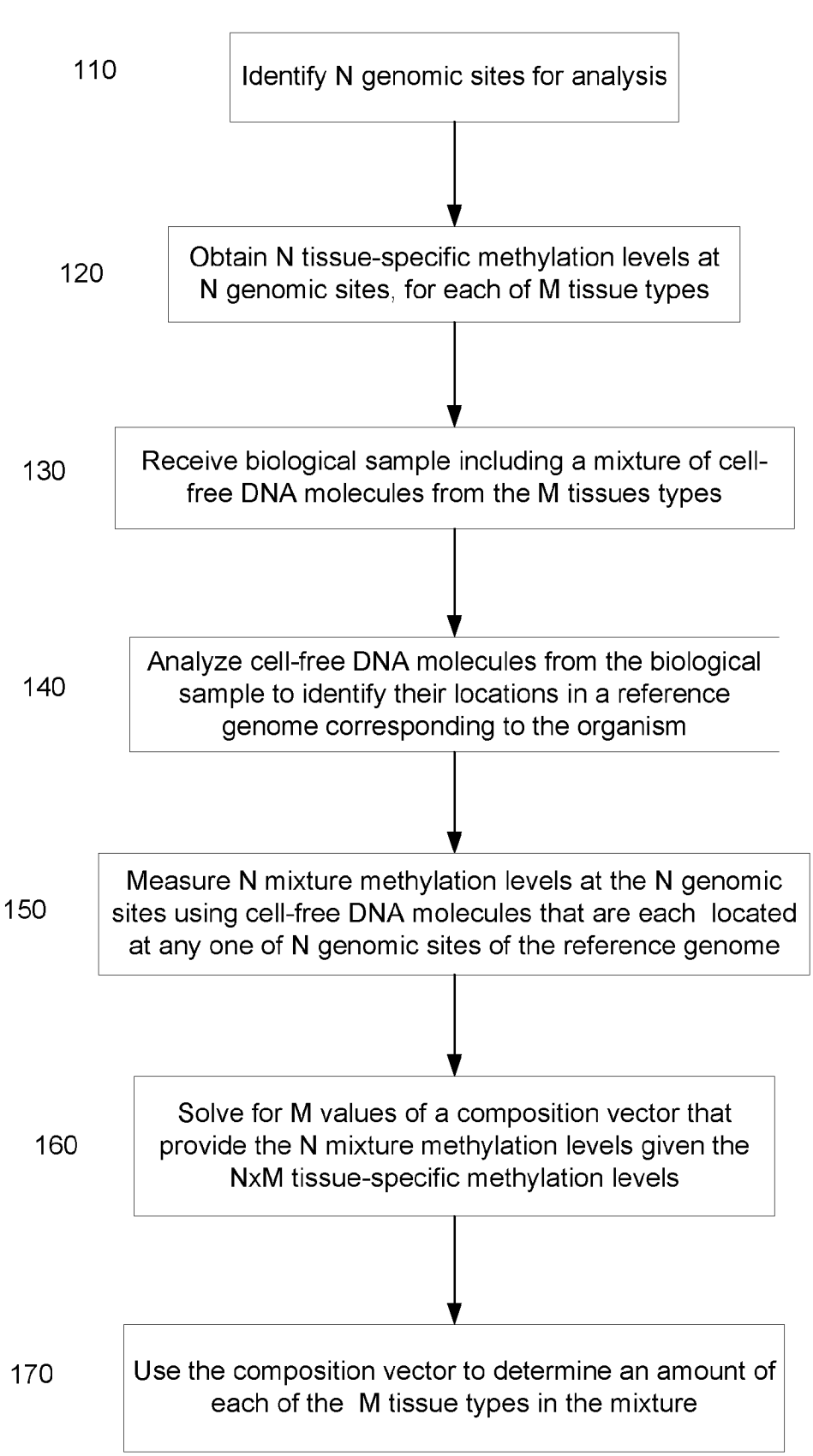

100
110
Identify N genomic sites for analysis
120
Obtain N tissue-specific methylation levels at N genomic sites, for each of M tissue types
130
Receive biological sample including a mixture of cell-free DNA molecules from the M tissues types
140
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome corresponding to the organism
150
Measure N mixture methylation levels at the N genomic sites using cell-free DNA molecules that are each located at any one of N genomic sites of the reference genome
160
Solve for M values of a composition vector that provide the N mixture methylation levels given the NxM tissue-specific methylation levels
170
Use the composition vector to determine an amount of each of the M tissue types in the mixture

FIG. 1

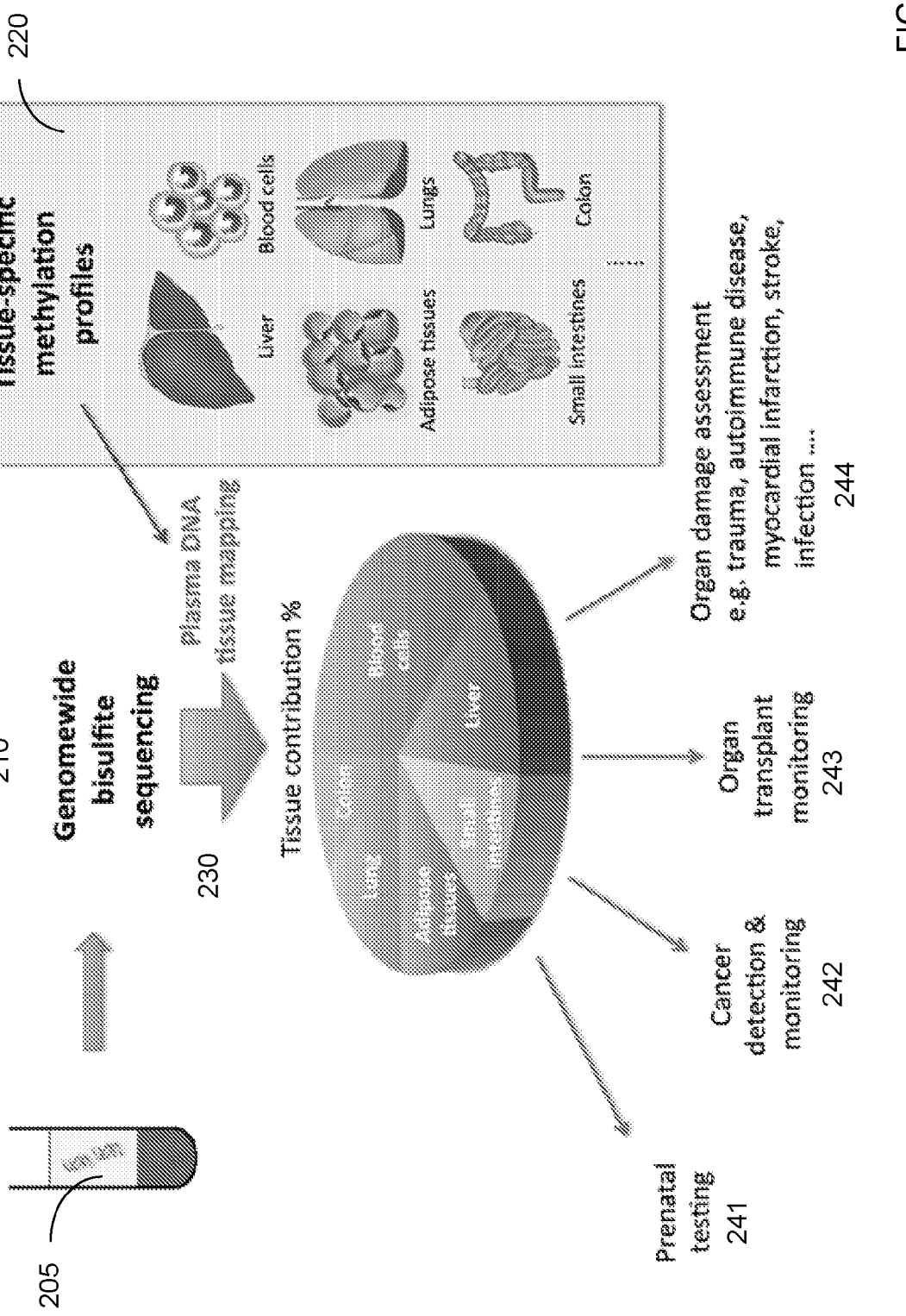

205
210
Genomewide bisulfite sequencing
230
Plasma DNA tissue mapping
220
Tissue-specific methylation profiles
Liver
Blood cells
Adipose tissues
Lungs
Small intestines
Colon
Tissue contribution %
Prenatal testing 241
Cancer detection & monitoring 242
Organ transplant monitoring 243
Organ damage assessment e.g. trauma, autoimmune disease, myocardial infarction, stroke, infection 244

FIG. 2

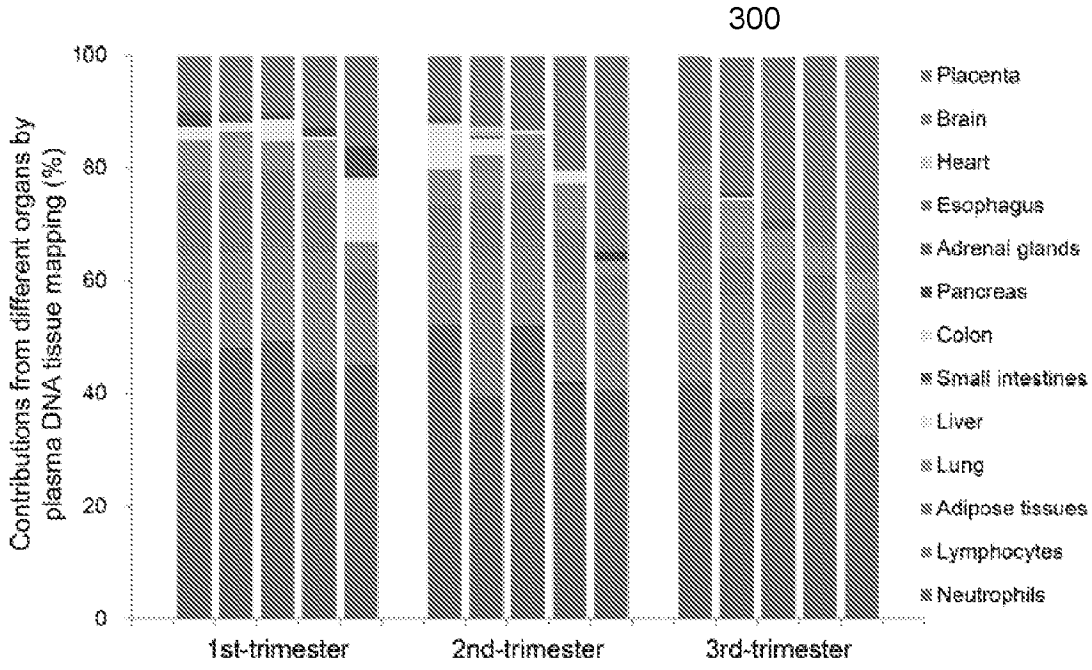

300
Contributions from different organs by plasma DNA tissue mapping (%)
100
80
60
40
20
0
1st-trimester
2nd-trimester
3rd-trimester
Placenta
Brain
Heart
Esophagus
Adrenal glands
Pancreas
Colon
Small intestines
Liver
Lung
Adipose tissues
Lymphocytes
Neutrophils

FIG. 3A

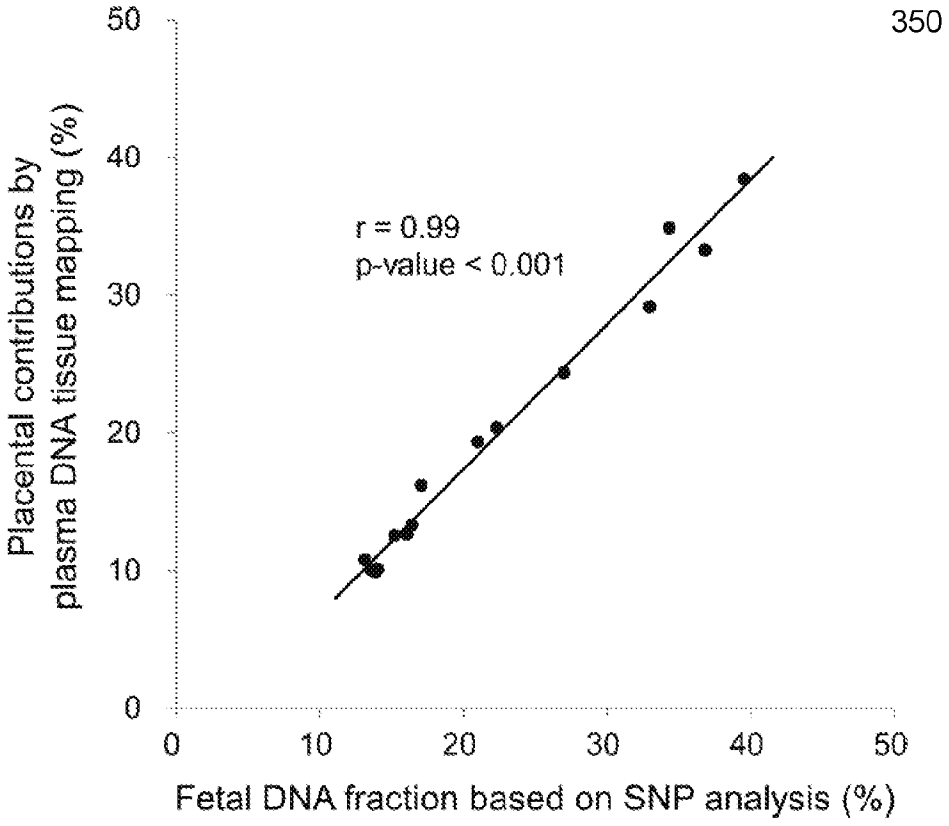

350
Placental contributions by plasma DNA tissue mapping (%)
50
40
30
20
10
0
r = 0.99
p-value < 0.001
0
10
20
30
40
50
Fetal DNA fraction based on SNP analysis (%)

FIG. 3B

400

Case no.	Percentage contributions of plasma DNA (%)													SNP-based fetal DNA fraction (%)	Gestational age (weeks)
	Liver	Lungs	Colon	Small intestines	Pancreas	Adrenal glands	Esophagus	Adipose tissues	Heart	Brain	Neutrophils	Lymphocytes	Placenta		
1	1.5	9.3	0.0	0.0	0.0	1.9	0.0	12.1	0.0	0.0	48.1	17.0	10.0	13.6	13 1/7
2	2.5	7.6	0.0	2.8	0.0	0.0	0.0	12.6	0.0	0.0	46.1	18.6	9.9	13.9	13
3	3.8	5.8	0.0	0.7	0.0	0.6	0.0	12.4	0.0	0.0	49.2	17.4	10.1	14.0	12 6/7
4	0.6	9.2	0.0	1.8	0.0	0.0	0.0	10.6	0.0	0.0	44.1	21.1	12.7	16.1	13
5	11.1	5.2	0.0	5.7	0.0	0.0	0.0	5.1	0.0	0.0	45.0	11.7	16.2	17.1	12 6/7
6	8.1	6.4	0.0	0.0	0.0	1.3	0.0	4.5	0.0	0.0	51.9	17.0	10.8	13.1	21 3/7
7	3.0	8.2	0.0	0.0	0.0	0.5	0.0	10.1	1.7	0.0	39.5	24.5	12.5	15.2	20 6/7
8	0.8	8.5	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	51.7	15.5	13.3	16.4	22 2/7
9	2.6	8.0	0.0	0.0	0.0	0.0	0.0	9.2	0.0	0.0	41.7	18.1	20.4	22.3	22 1/7
10	0.0	2.6	0.0	0.0	1.6	0.0	0.0	4.1	0.0	0.0	40.5	16.3	34.9	34.3	21 /7
11	0.0	7.4	0.0	0.0	0.0	0.0	0.0	9.8	0.0	0.0	41.5	22.0	19.3	21.0	38 1/7
12	0.6	10.1	0.0	0.8	0.0	0.0	0.0	6.6	0.0	0.0	39.0	18.7	24.3	27.0	38 3/7
13	0.0	7.4	0.0	0.0	0.0	1.7	0.0	6.3	0.0	0.0	36.7	18.7	29.2	33.0	38 1/7
14	0.0	5.9	0.0	0.0	0.0	0.0	0.0	8.3	0.0	0.0	39.6	12.9	33.3	36.8	38 2/7
15	0.0	7.3	0.0	0.0	0.0	0.0	0.0	7.2	0.0	0.0	32.5	14.6	38.4	39.5	38 3/7

FIG. 4

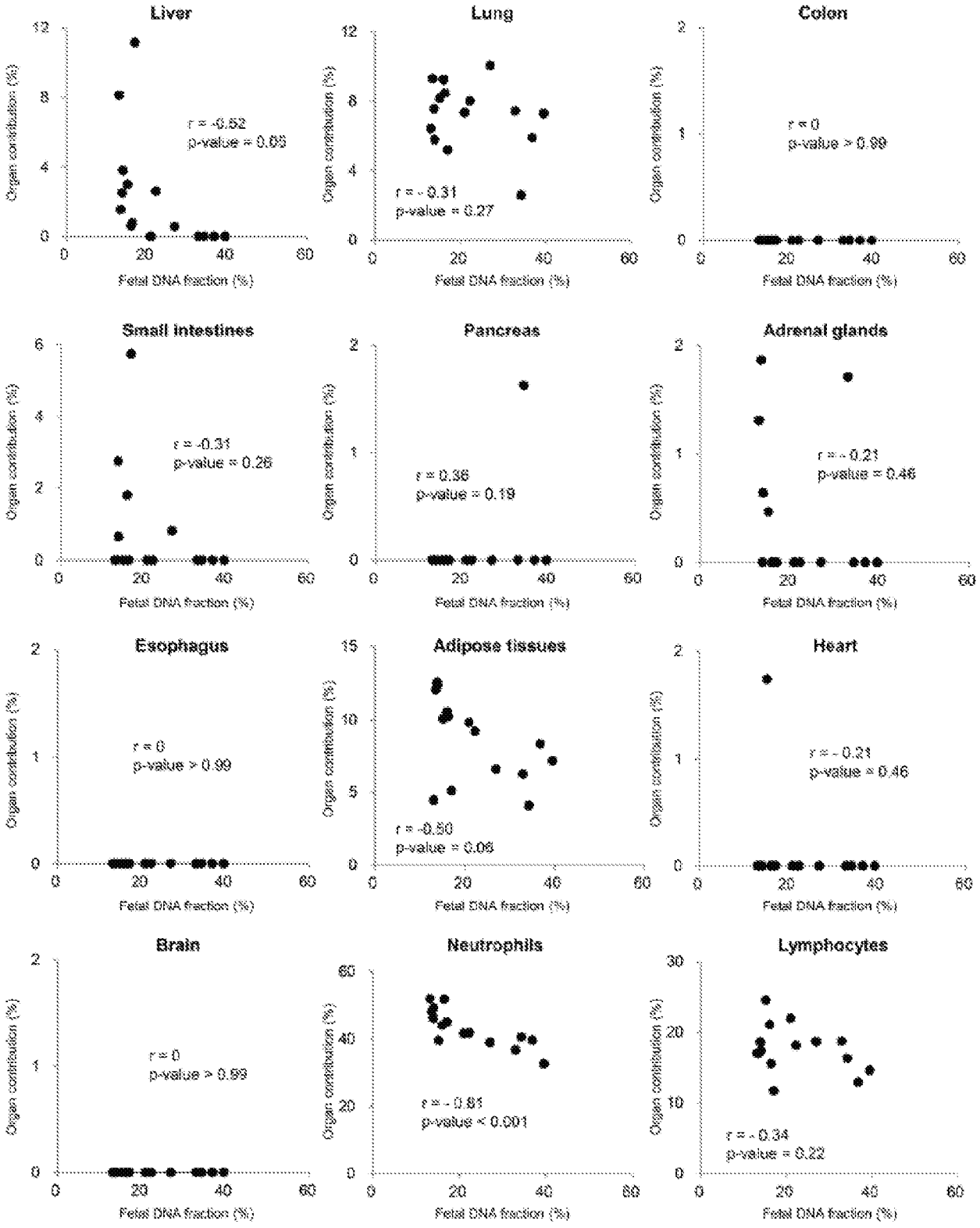

Liver
r = -0.52
p-value = 0.05
Lung
r = - 0.31
p-value = 0.27
Colon
r = 0
p-value > 0.99
Small intestines
r = -0.31
p-value = 0.26
Pancreas
r = 0.38
p-value = 0.19
Adrenal glands
r = - 0.21
p-value = 0.46
Esophagus
r = 0
p-value > 0.99
Adipose tissues
r = -0.50
p-value = 0.06
Heart
r = - 0.21
p-value = 0.46
Brain
r = 0
p-value > 0.99
Neutrophils
r = - 0.91
p-value < 0.001
Lymphocytes
r = - 0.34
p-value = 0.22
Organ contribution (%)
Fetal DNA fraction (%)

FIG. 5

600

Case no.	Percentage contributions of plasma DNA (%)												
	Liver	Lungs	Colon	Small intestines	Pancreas	Adrenal glands	Esophagus	Adipose tissues	Heart	Brain	Neutrophils	Lymphocytes	Placenta
1	5.8	6.8	0.0	1.3	0.0	0.0	0.0	6.2	2.9	0.0	58.2	18.8	0.0
2	5.1	5.2	0.0	7.3	0.0	0.0	0.0	4.5	0.0	0.0	61.8	16.1	0.0
3	5.6	7.8	0.0	2.1	0.0	0.0	0.0	6.5	0.6	0.0	58.8	18.6	0.0
4	4.1	6.4	0.0	4.5	0.0	0.0	0.0	5.9	1.3	0.4	59.0	18.4	0.0
5	4.4	2.5	0.7	6.4	0.0	0.0	0.0	5.5	2.7	0.6	59.6	17.1	0.3
6	5.8	4.1	0.0	5.3	0.0	0.0	0.0	5.2	0.5	0.8	55.2	22.7	0.5
7	4.9	4.1	0.0	4.5	0.0	0.0	0.0	6.3	0.0	0.0	60.2	20.0	0.0
8	4.4	7.4	0.0	5.3	0.0	0.0	0.0	3.1	0.3	0.3	61.7	17.5	0.0
9	4.5	9.4	0.0	4.0	0.0	0.0	0.0	6.2	0.1	1.2	55.9	18.6	0.0
10	4.6	0.1	0.2	1.3	0.0	0.0	0.0	7.5	0.0	2.4	60.0	23.9	0.0
11	4.7	4.5	0.0	5.5	0.0	0.0	0.0	5.6	0.9	0.5	62.4	16.0	0.0
12	7.3	4.9	0.0	4.4	0.0	0.0	0.0	6.0	1.8	0.9	54.0	20.7	0.0
13	2.4	10.9	0.0	1.7	0.0	0.0	0.0	5.7	0.0	0.6	53.9	24.5	0.3
14	3.8	3.5	0.0	0.0	0.0	0.0	0.0	8.9	0.0	0.0	61.2	22.7	0.0
15	5.3	7.2	0.0	4.2	0.0	0.2	0.0	5.1	0.0	0.1	56.3	21.6	0.0
16	4.9	6.5	0.0	7.9	0.0	0.0	0.0	4.3	0.5	1.2	55.4	18.9	0.4
17	18.6	12.8	0.0	0.0	0.0	0.0	0.0	4.4	0.0	0.0	47.7	16.5	0.0
18	8.3	5.0	0.0	2.1	0.0	0.0	0.0	7.5	0.0	0.2	56.2	19.4	1.3
19	6.0	0.0	1.0	2.3	0.0	0.0	0.0	9.3	0.0	0.0	54.5	26.3	0.6
20	2.0	4.8	0.0	0.0	0.0	0.5	0.0	10.5	0.0	0.0	58.5	22.5	1.1
21	8.5	0.0	0.0	4.5	0.0	0.0	0.0	9.4	0.0	0.0	53.9	23.4	0.2
22	6.4	11.1	0.0	4.0	0.0	0.0	0.0	6.7	0.0	0.0	53.7	18.1	0.0
23	2.0	2.8	0.0	8.7	0.0	0.0	0.0	6.5	1.1	0.0	62.5	15.0	1.3
24	5.4	2.2	0.0	2.7	0.0	0.0	0.0	6.9	0.0	1.6	65.2	15.2	0.9
25	21.0	0.0	0.0	7.1	0.0	0.0	0.0	3.8	0.0	0.0	51.6	16.4	0.0
26	7.0	6.4	0.0	6.1	0.0	0.0	0.0	4.1	1.6	0.0	55.9	19.0	0.0
27	7.9	10.4	0.0	0.3	0.0	0.0	0.0	4.8	0.0	0.0	59.1	17.6	0.0
28	5.4	9.7	0.0	4.2	0.0	0.0	0.0	5.1	0.0	0.0	55.7	20.0	0.0
29	6.2	6.6	0.0	0.0	0.0	0.0	0.0	6.3	0.0	0.0	60.7	20.2	0.0
30	6.9	5.1	0.0	5.2	0.0	0.0	0.0	6.4	1.4	0.0	58.8	16.2	0.0
31	14.0	6.4	0.0	4.5	0.0	0.0	0.0	4.9	0.7	0.0	53.2	16.2	0.0
32	9.9	5.5	0.0	1.9	0.0	0.0	0.0	6.3	0.6	0.0	60.1	15.6	0.0
Median	5.5	5.3	0.0	4.2	0.0	0.0	0.0	6.1	0.0	0.0	58.3	18.7	0.0
Lower quantile	4.6	3.9	0.0	1.9	0.0	0.0	0.0	5.0	0.0	0.0	55.0	16.5	0.0
Upper quantile	7.1	7.3	0.0	5.3	0.0	0.0	0.0	6.6	0.8	0.5	60.1	21.0	0.3

FIG. 6

700

	Pregnant women											Non-pregnant subjects			
	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4
Buffy coat	58%	64%	65%	53%	83%	52%	75%	71%	52%	66%	69%	87%	83%	96%	80%
Esophagus	0%	0%	0%	2%	0%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%
Small intestines	0%	0%	0%	4%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Colon	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	0%	0%
Pancreas	0%	1%	0%	1%	0%	0%	0%	0%	1%	3%	2%	1%	0%	0%	0%
Liver	3%	8%	10%	1%	0%	11%	6%	7%	4%	9%	8%	7%	7%	3%	6%
Lung	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Heart	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	10%	0%	0%	0%	2%
Adrenal gland	0%	1%	0%	1%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	5%
Hippocampus	1%	3%	0%	1%	2%	4%	1%	4%	3%	3%	4%	2%	5%	0%	5%
Placenta	38%	24%	25%	36%	15%	34%	18%	15%	40%	19%	7%	2%	0%	1%	2%
Fetal DNA fraction determined by counting fetal specific alleles	32%	21%	26%	28%	11%	25%	17%	13%	13%	32%	3%	0%	0%	0%	0%

FIG. 7

800

	Pregnant women											Non-pregnant subjects			
	1	2	3	4	5	6	7	8	9	10	11	1	2	3	4
Buffy coat	43%	44%	47%	40%	47%	41%	38%	41%	43%	39%	45%	10%	11%	1%	13%
Esophagus	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Small intestines	0%	4%	1%	2%	3%	0%	1%	5%	0%	0%	0%	0%	0%	0%	0%
Colon	25%	18%	10%	23%	14%	23%	24%	20%	27%	25%	16%	17%	7%	22%	27%
Pancreas	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Liver	4%	8%	11%	4%	8%	11%	9%	0%	0%	0%	0%	0%	0%	0%	0%
Lung	0%	0%	0%	1%	0%	0%	0%	3%	0%	0%	11%	0%	0%	0%	0%
Heart	0%	0%	1%	3%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%
Adrenal gland	0%	3%	6%	0%	0%	0%	7%	4%	3%	7%	7%	24%	23%	20%	4%
Hippocampus	7%	10%	7%	11%	12%	10%	8%	15%	9%	11%	16%	44%	42%	41%	50%
Placenta	21%	13%	18%	16%	15%	16%	11%	12%	18%	18%	6%	4%	16%	15%	7%
Fetal DNA fraction determined by counting fetal specific alleles	32%	21%	26%	28%	11%	25%	17%	13%	13%	32%	3%	0%	0%	0%	0%

FIG. 8

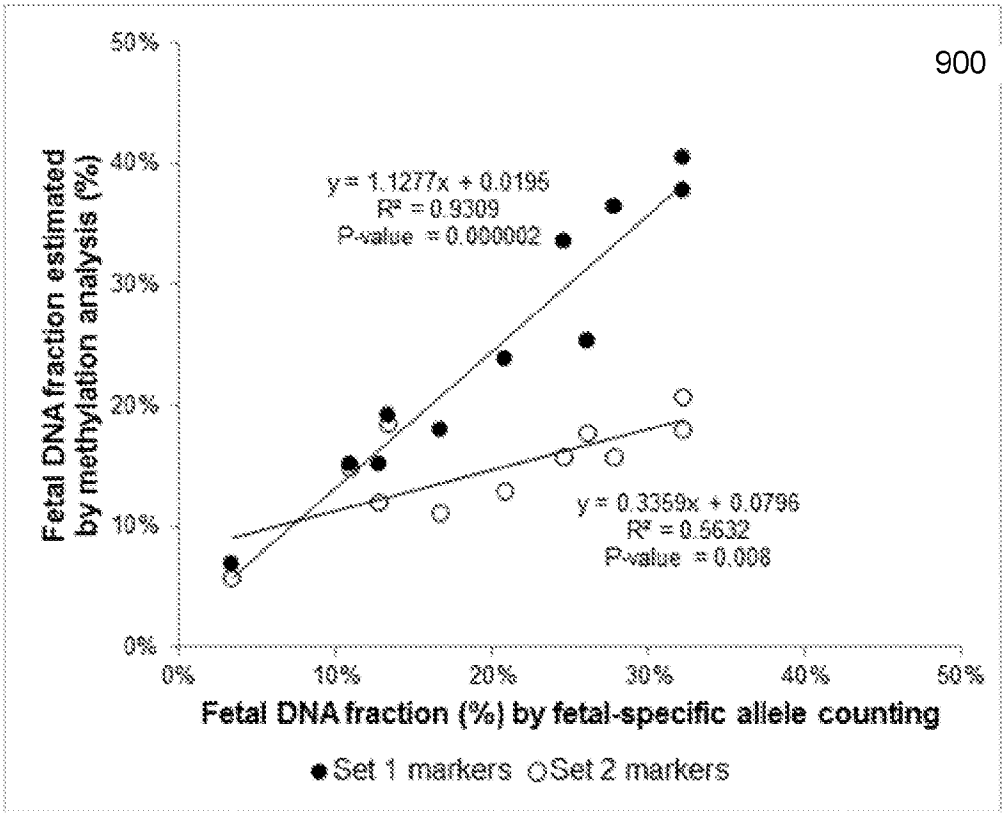

900
50%
40%
30%
20%
10%
0%
Fetal DNA fraction estimated by methylation analysis (%)
y = 1.1277x + 0.0195
R² = 0.9309
P-value = 0.000002
y = 0.3359x + 0.0796
R² = 0.5632
P-value = 0.008
0%
10%
20%
30%
40%
50%
Fetal DNA fraction (%) by fetal-specific allele counting
Set 1 markers
Set 2 markers

FIG. 9A

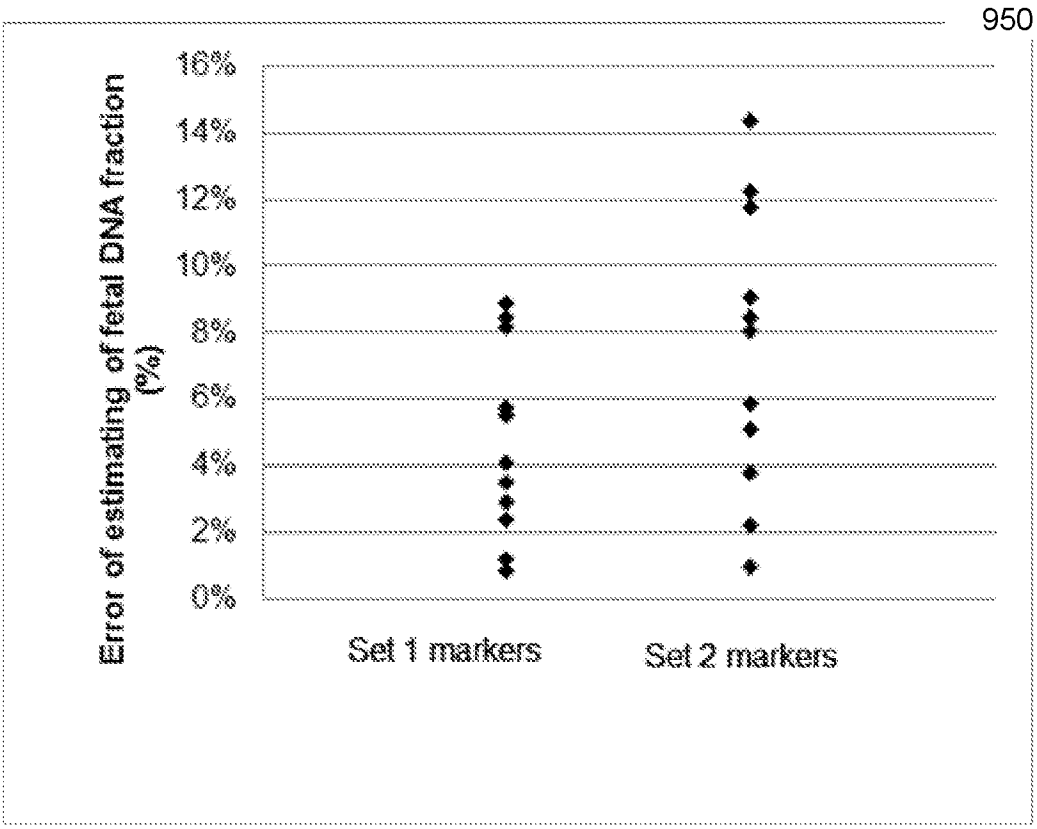

950
16%
14%
12%
10%
8%
6%
4%
2%
0%
Error of estimating of fetal DNA fraction (%)
Set 1 markers
Set 2 markers

FIG. 9B

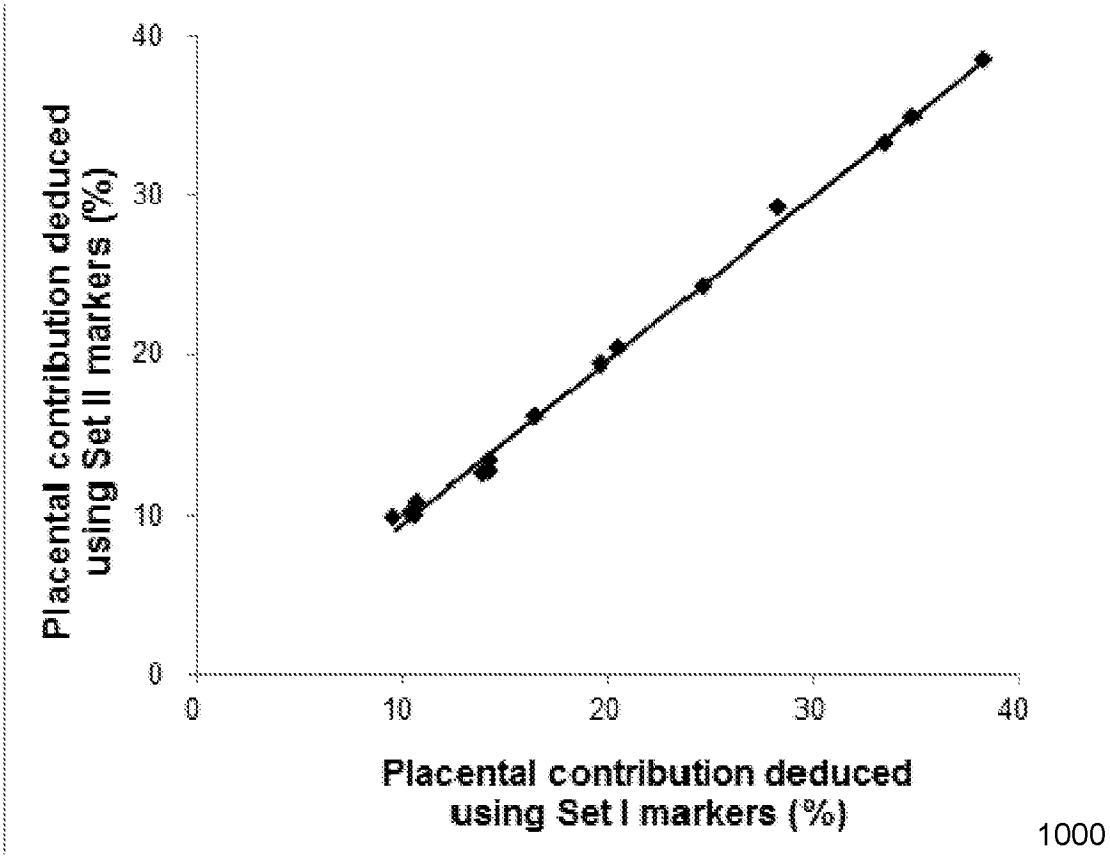


FIG. 10A

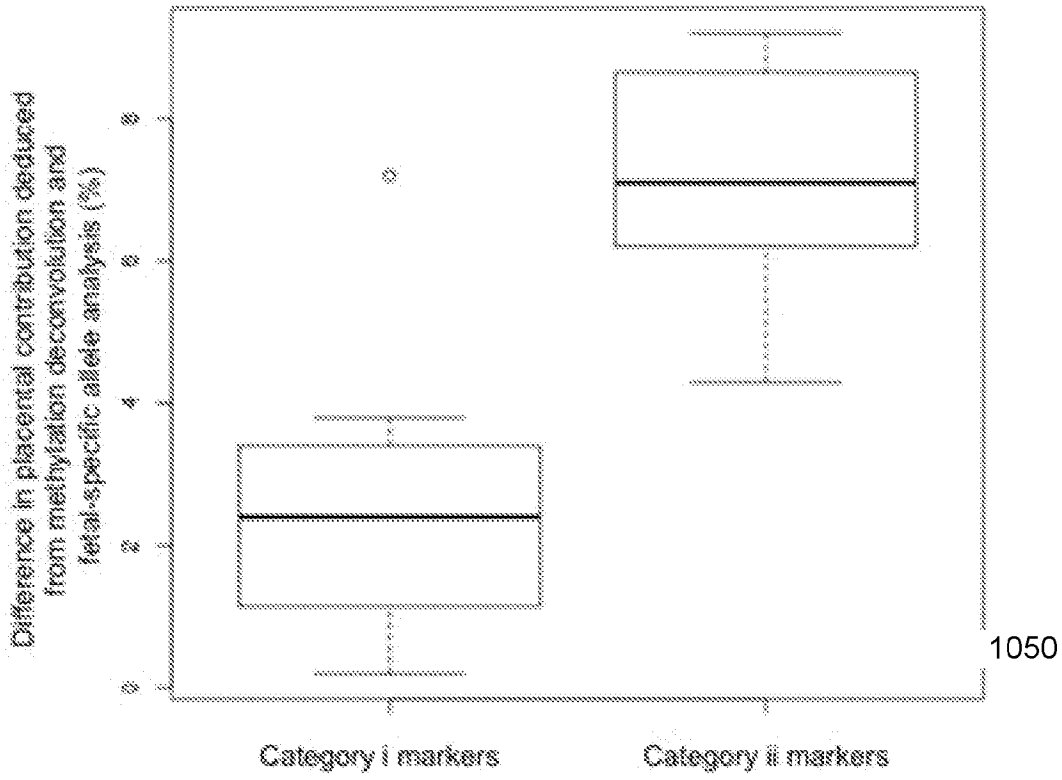


FIG. 10B

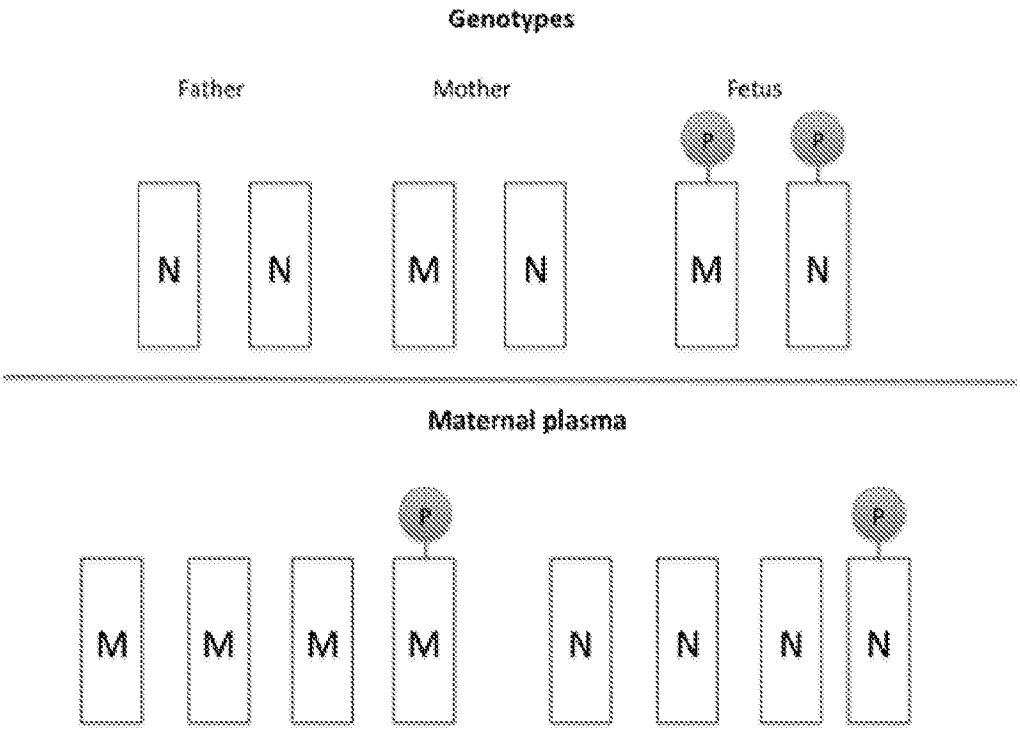


FIG. 11A

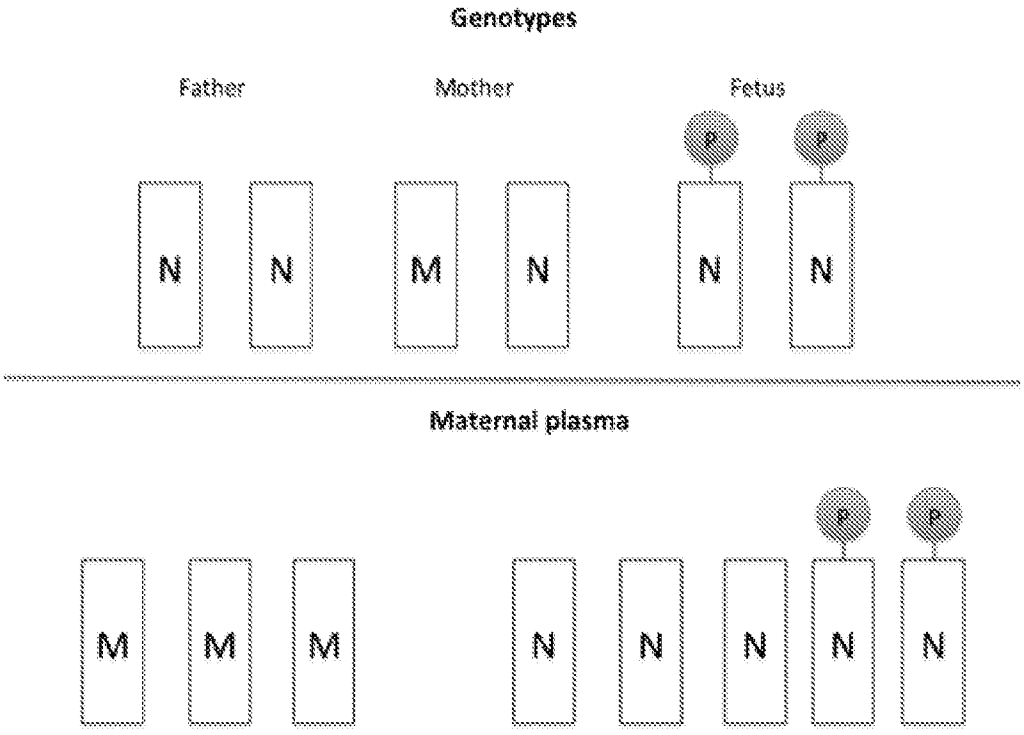


FIG. 11B

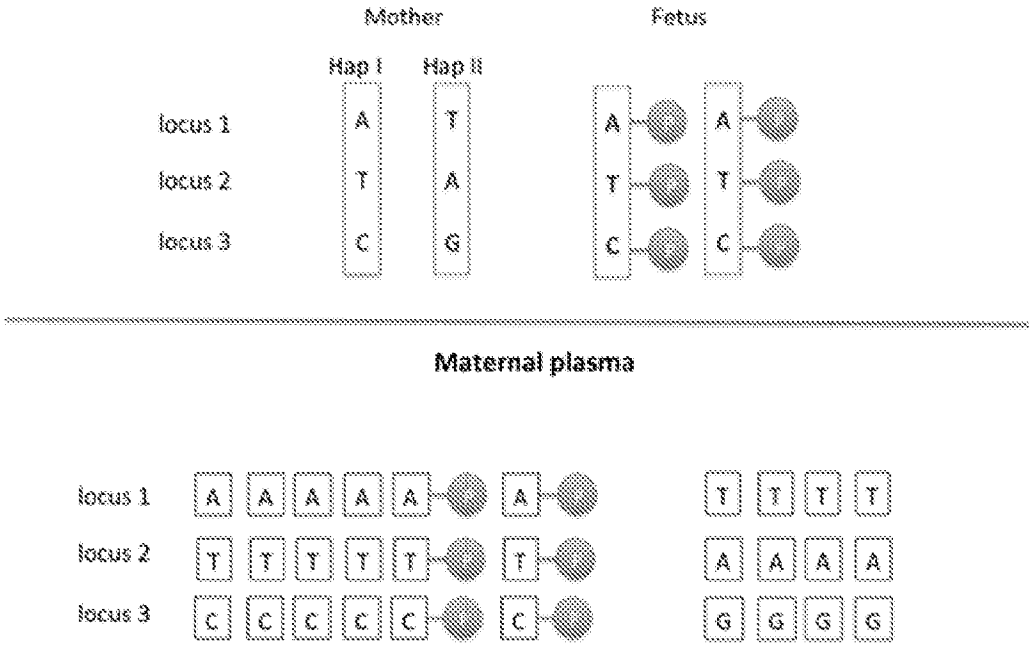


FIG. 12A

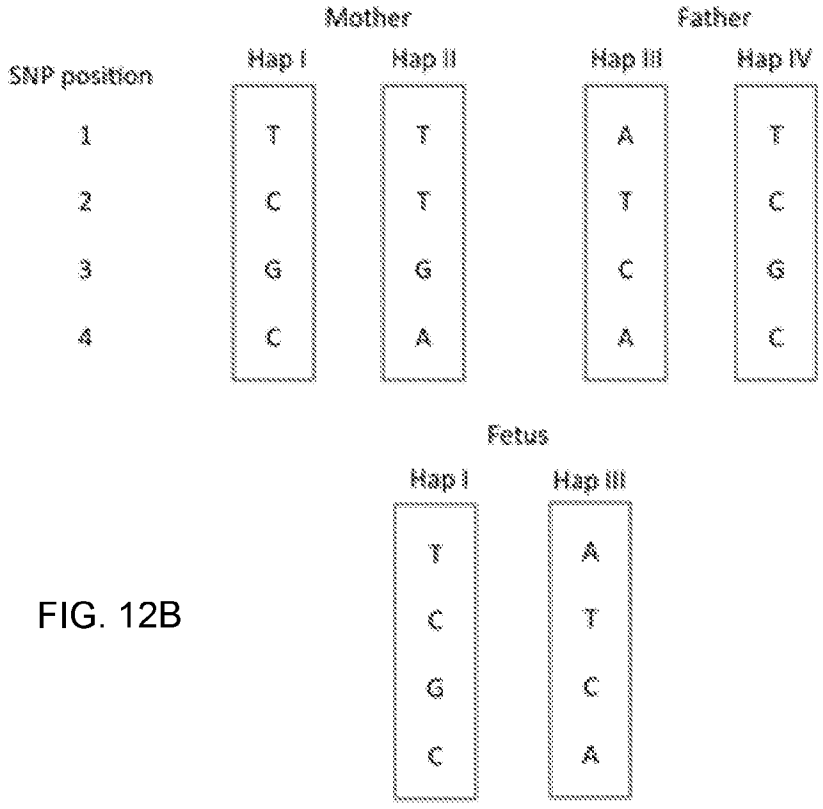


FIG. 12B

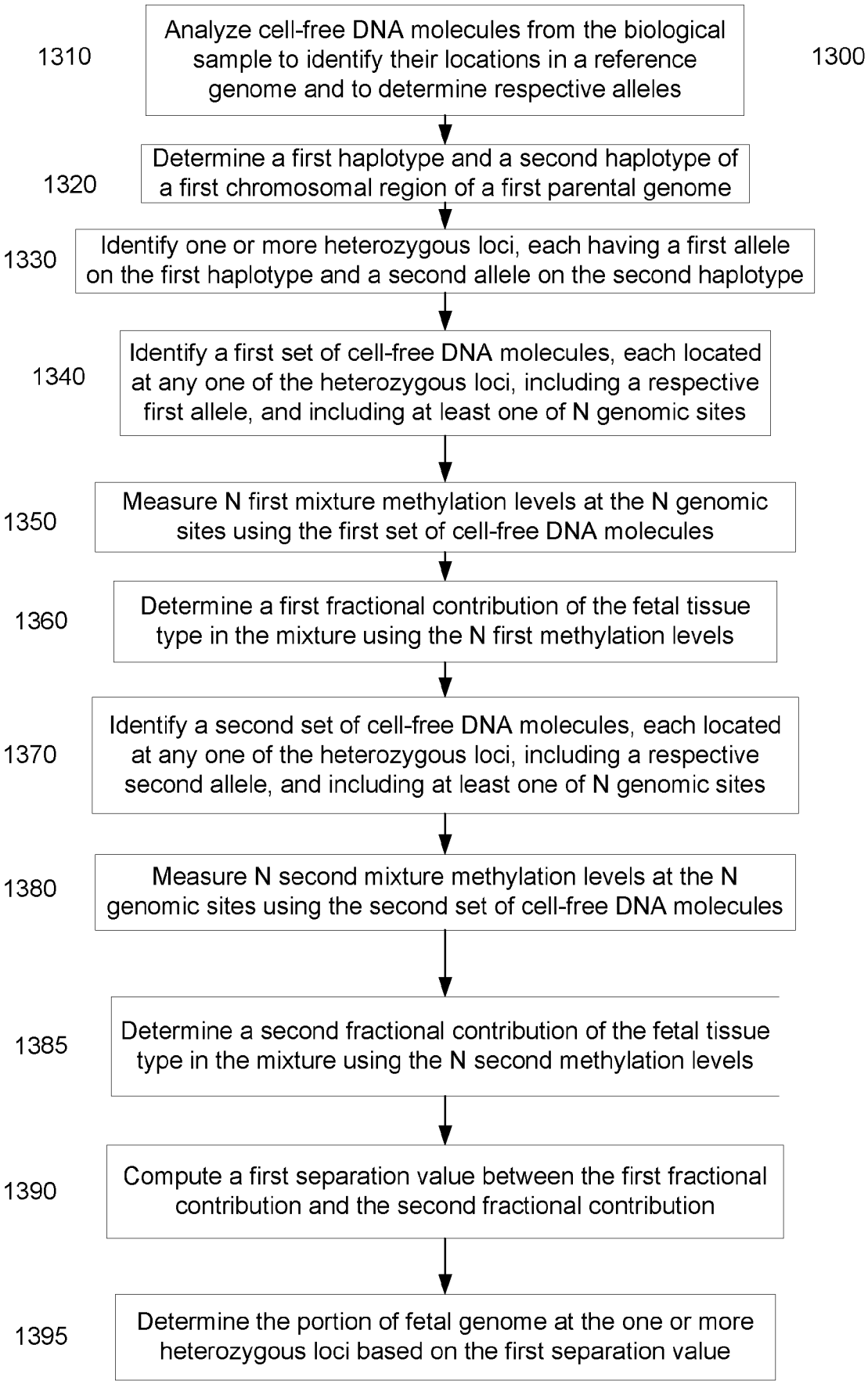

1300
1310
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1320
Determine a first haplotype and a second haplotype of a first chromosomal region of a first parental genome
1330
Identify one or more heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
1340
Identify a first set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective first allele, and including at least one of N genomic sites
1350
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
1360
Determine a first fractional contribution of the fetal tissue type in the mixture using the N first methylation levels
1370
Identify a second set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective second allele, and including at least one of N genomic sites
1380
Measure N second mixture methylation levels at the N genomic sites using the second set of cell-free DNA molecules
1385
Determine a second fractional contribution of the fetal tissue type in the mixture using the N second methylation levels
1390
Compute a first separation value between the first fractional contribution and the second fractional contribution
1395
Determine the portion of fetal genome at the one or more heterozygous loci based on the first separation value

FIG. 13

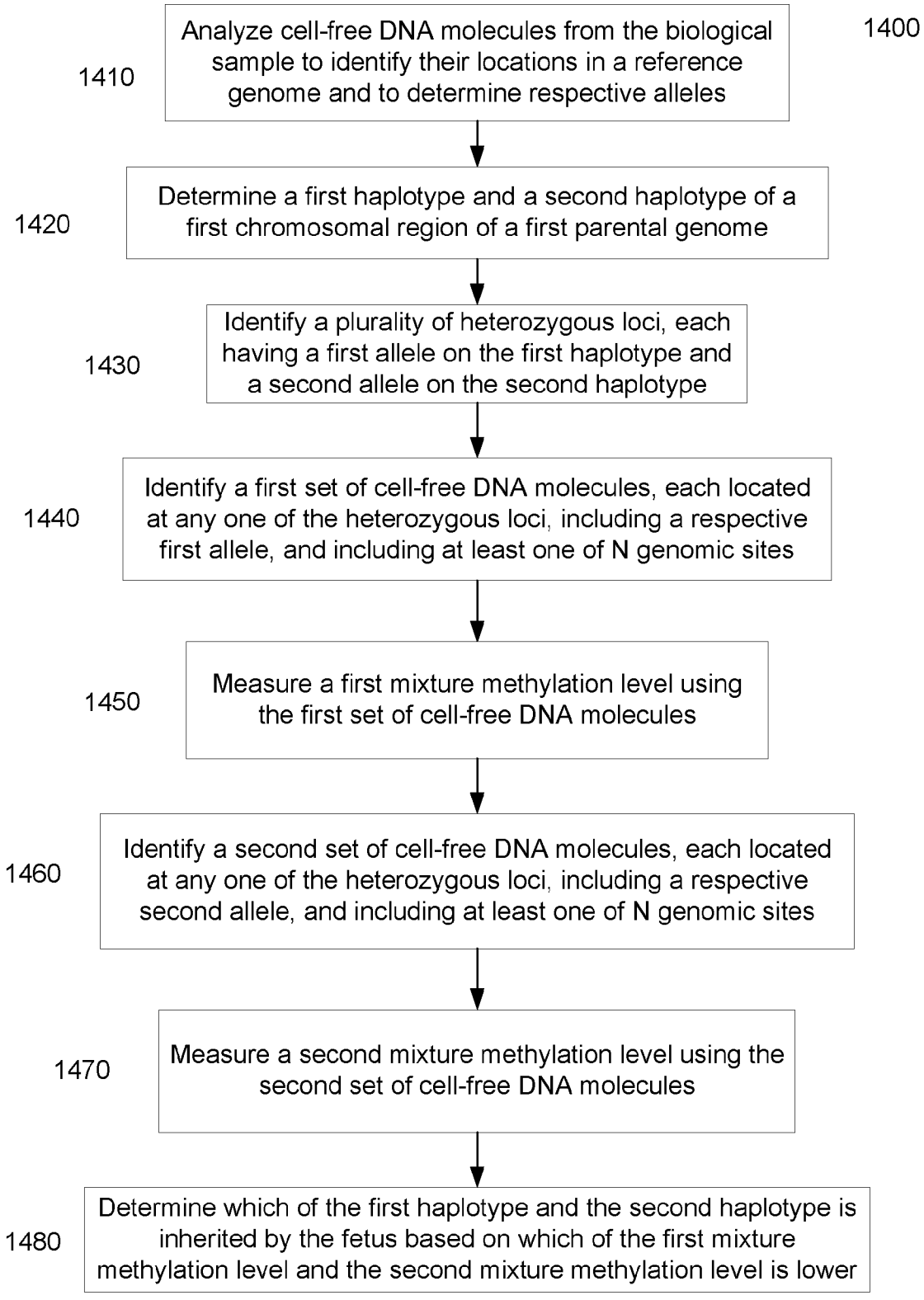

1400
1410
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1420
Determine a first haplotype and a second haplotype of a first chromosomal region of a first parental genome
1430
Identify a plurality of heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
1440
Identify a first set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective first allele, and including at least one of N genomic sites
1450
Measure a first mixture methylation level using the first set of cell-free DNA molecules
1460
Identify a second set of cell-free DNA molecules, each located at any one of the heterozygous loci, including a respective second allele, and including at least one of N genomic sites
1470
Measure a second mixture methylation level using the second set of cell-free DNA molecules
1480
Determine which of the first haplotype and the second haplotype is inherited by the fetus based on which of the first mixture methylation level and the second mixture methylation level is lower

FIG. 14

FIG. 15

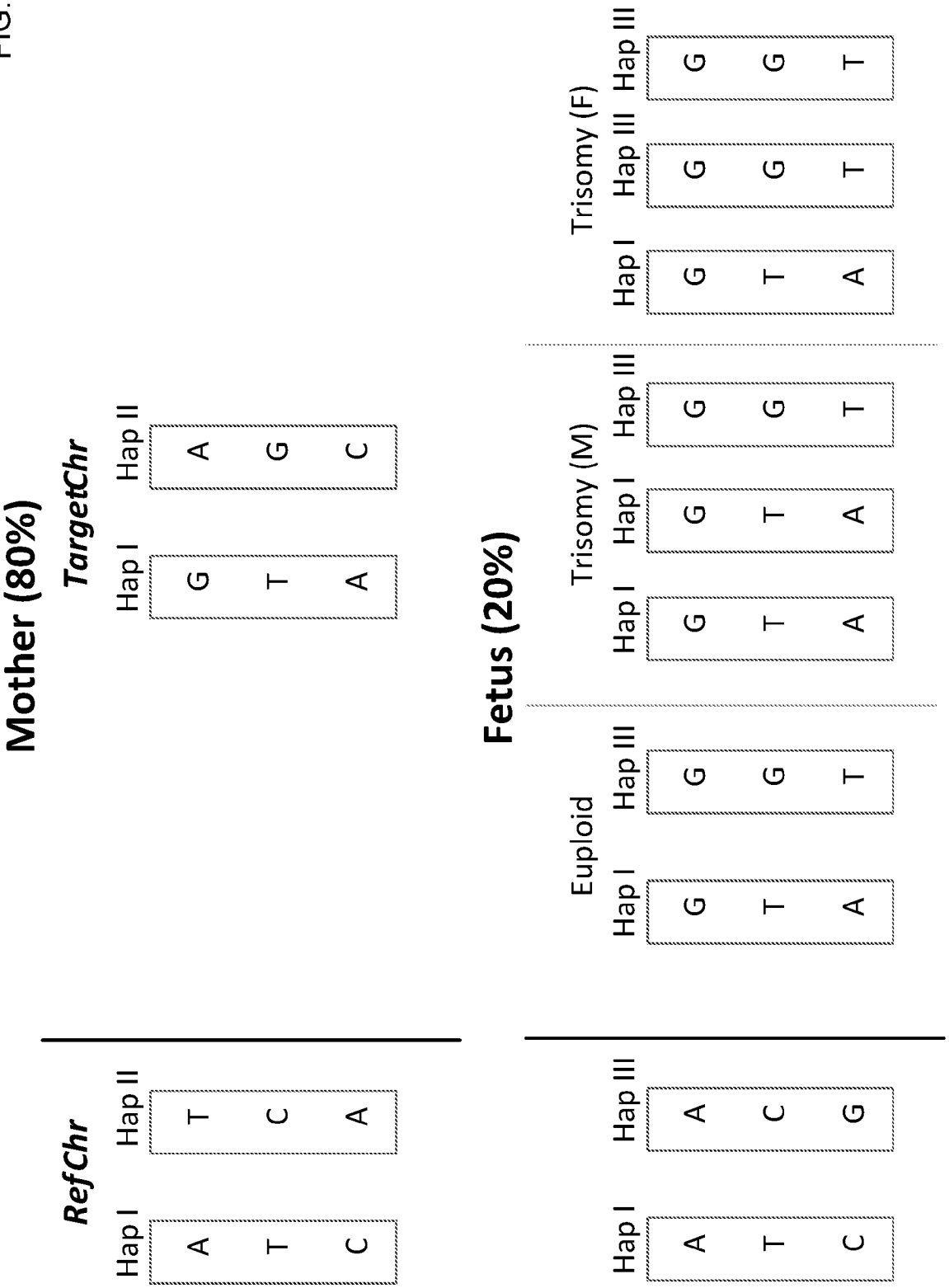
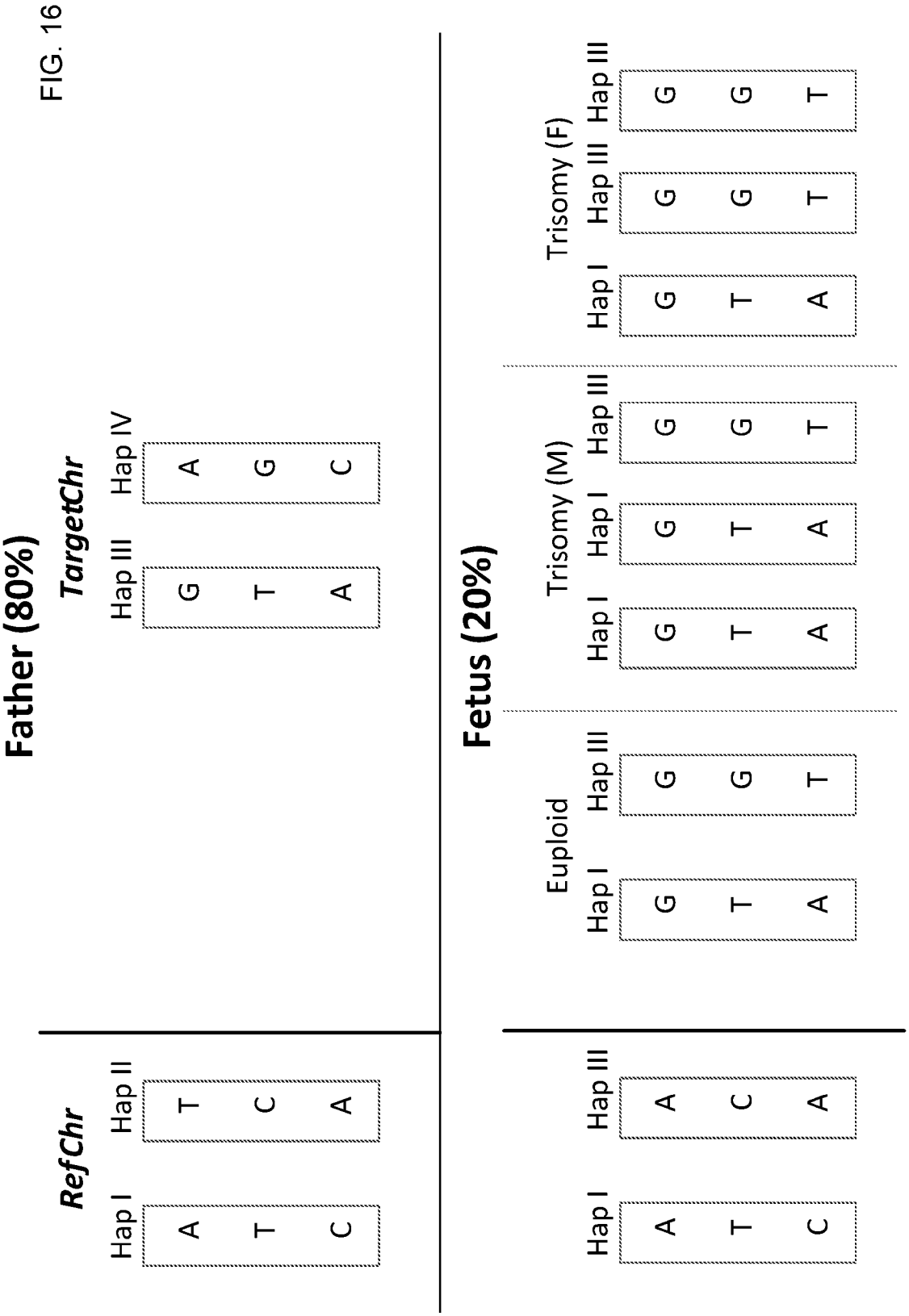

Mother (80%)
RefChr
Hap I
A
T
C
Hap II
T
C
A
TargetChr
Hap I
G
T
A
Hap II
A
G
C
Fetus (20%)
Euploid
Hap I
G
T
A
Hap III
G
G
T
Hap III
A
C
G
Trisomy (M)
Hap I
G
T
A
Hap I
G
T
A
Hap III
G
G
T
Trisomy (F)
Hap I
G
T
A
Hap III
G
G
T
Hap III
G
G
T

FIG. 16


Father (80%)
RefChr
Hap I
A
T
C
Hap II
T
C
A
TargetChr
Hap III
G
T
A
Hap IV
A
G
C
Fetus (20%)
Euploid
Hap I
G
T
A
Hap III
G
G
T
Trisomy (M)
Hap I
G
T
A
G
T
A
Hap III
G
G
T
Trisomy (F)
Hap I
G
T
A
Hap III
G
G
T
Hap III
A
C
A
Hap I
A
T
C

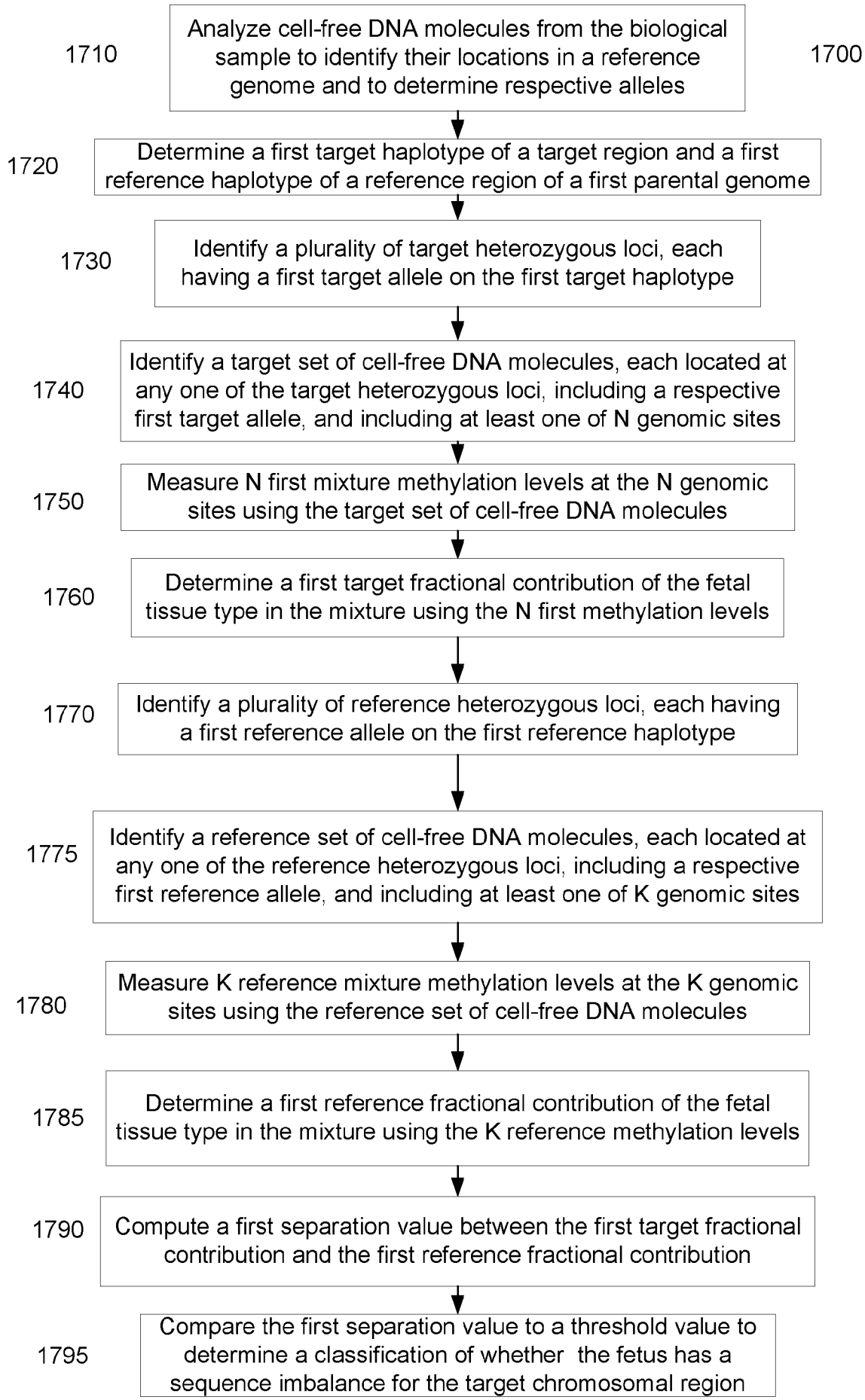

1700
1710
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1720
Determine a first target haplotype of a target region and a first reference haplotype of a reference region of a first parental genome
1730
Identify a plurality of target heterozygous loci, each having a first target allele on the first target haplotype
1740
Identify a target set of cell-free DNA molecules, each located at any one of the target heterozygous loci, including a respective first target allele, and including at least one of N genomic sites
1750
Measure N first mixture methylation levels at the N genomic sites using the target set of cell-free DNA molecules
1760
Determine a first target fractional contribution of the fetal tissue type in the mixture using the N first methylation levels
1770
Identify a plurality of reference heterozygous loci, each having a first reference allele on the first reference haplotype
1775
Identify a reference set of cell-free DNA molecules, each located at any one of the reference heterozygous loci, including a respective first reference allele, and including at least one of K genomic sites
1780
Measure K reference mixture methylation levels at the K genomic sites using the reference set of cell-free DNA molecules
1785
Determine a first reference fractional contribution of the fetal tissue type in the mixture using the K reference methylation levels
1790
Compute a first separation value between the first target fractional contribution and the first reference fractional contribution
1795
Compare the first separation value to a threshold value to determine a classification of whether the fetus has a sequence imbalance for the target chromosomal region

FIG. 17

Donor
Hap I
Hap II
Recipient
Hap III
Hap IV
Locus 1
T
T
A
A
Locus 2
C
T
T
C
Locus 3
G
G
C
C
Locus 4
C
A
A
C
Plasma DNA deconvolution
Stable
Rejection
Tissues
Hap I
Hap II
Hap I
Hap II
Transplanted organ
Baseline level

FIG. 18

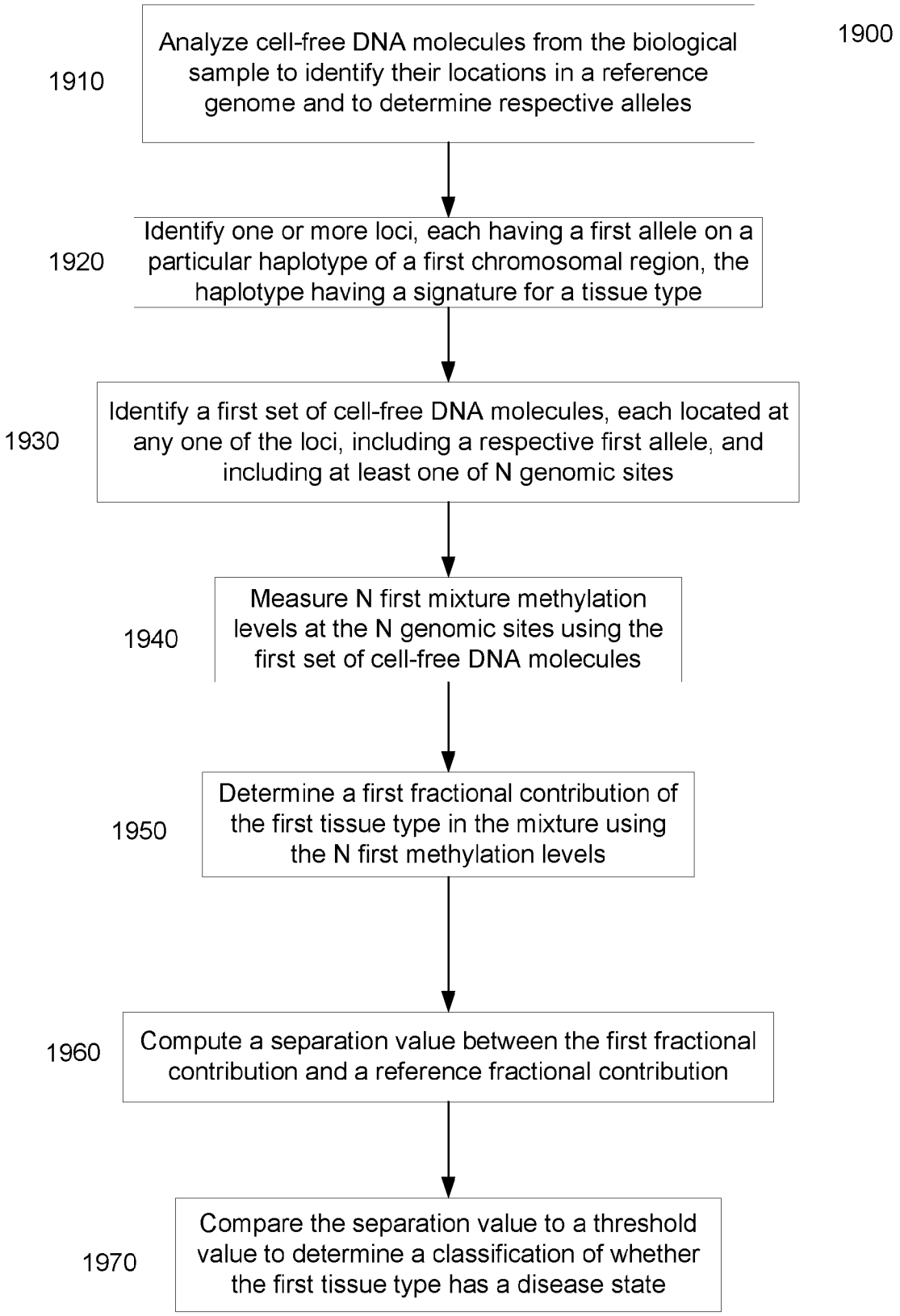

1900
1910
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
1920
Identify one or more loci, each having a first allele on a particular haplotype of a first chromosomal region, the haplotype having a signature for a tissue type
1930
Identify a first set of cell-free DNA molecules, each located at any one of the loci, including a respective first allele, and including at least one of N genomic sites
1940
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
1950
Determine a first fractional contribution of the first tissue type in the mixture using the N first methylation levels
1960
Compute a separation value between the first fractional contribution and a reference fractional contribution
1970
Compare the separation value to a threshold value to determine a classification of whether the first tissue type has a disease state

FIG. 19

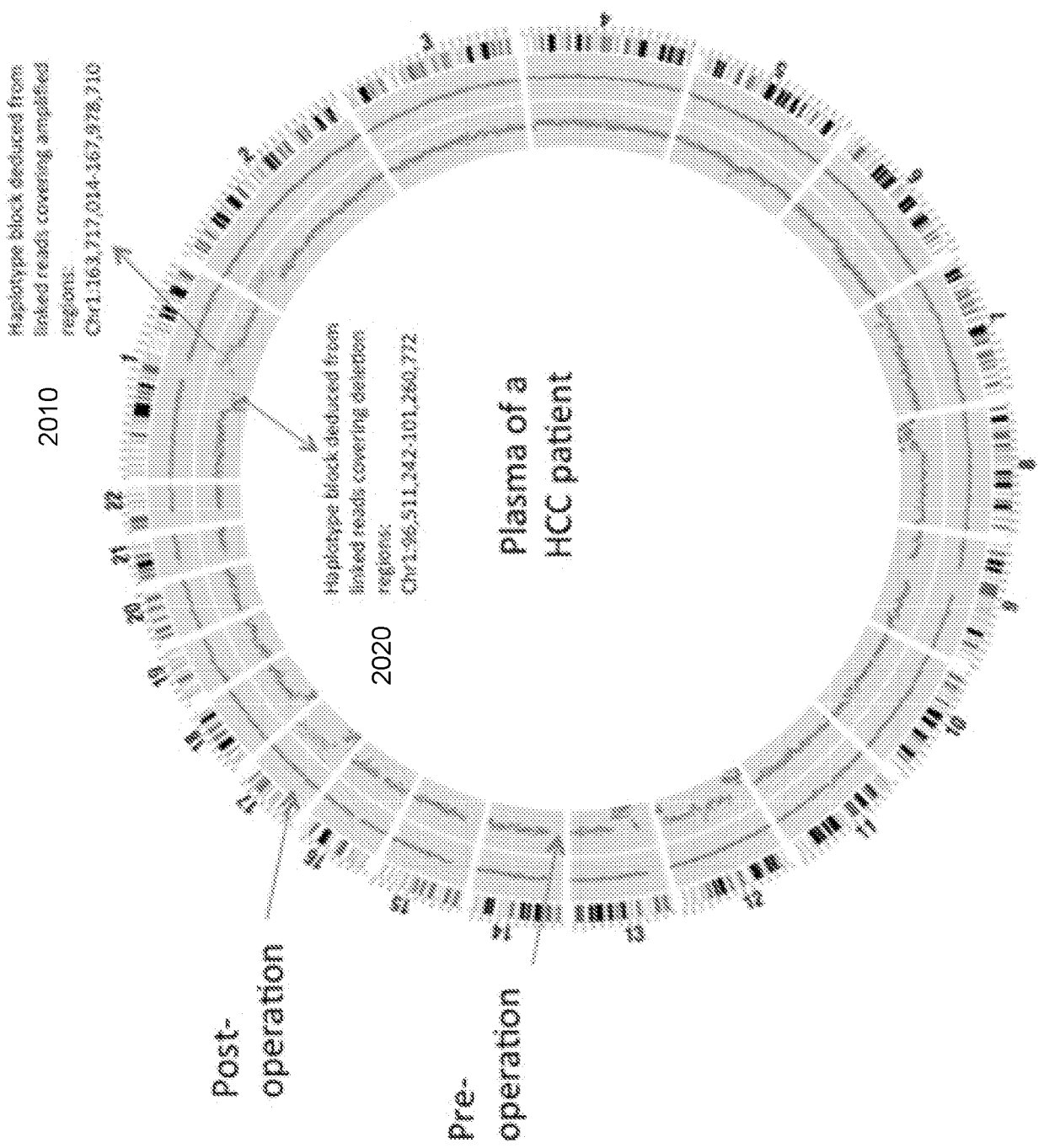

Haplotype block deduced from linked reads covering amplified regions:
Chr1:363,717,034-367,978,710
2010
Haplotype block deduced from linked reads covering deletion regions:
Chr1:96,311,242-101,260,772
2020
Plasma of a HCC patient
Post-operation
Pre-operation

FIG. 20

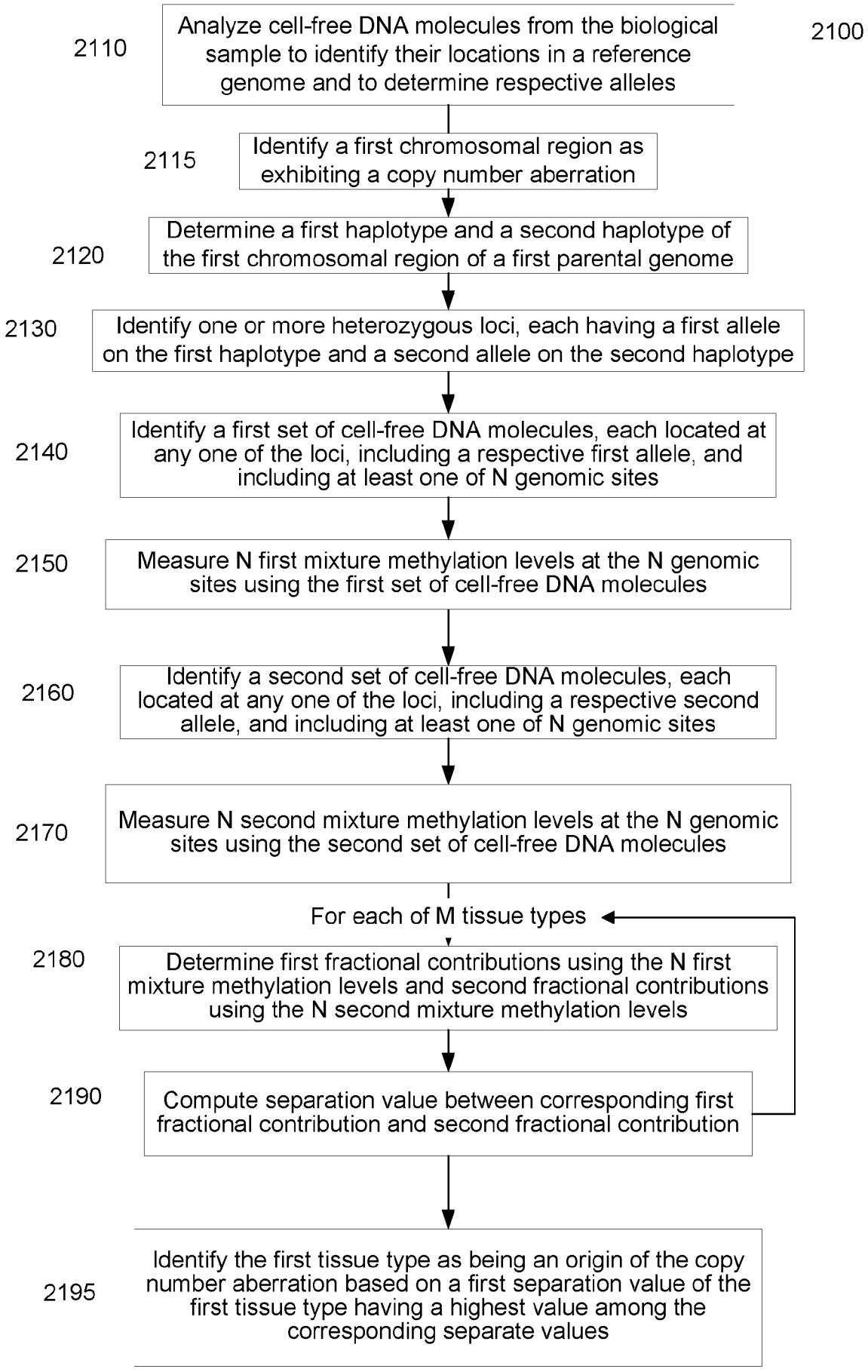

2100
2110
Analyze cell-free DNA molecules from the biological sample to identify their locations in a reference genome and to determine respective alleles
2115
Identify a first chromosomal region as exhibiting a copy number aberration
2120
Determine a first haplotype and a second haplotype of the first chromosomal region of a first parental genome
2130
Identify one or more heterozygous loci, each having a first allele on the first haplotype and a second allele on the second haplotype
2140
Identify a first set of cell-free DNA molecules, each located at any one of the loci, including a respective first allele, and including at least one of N genomic sites
2150
Measure N first mixture methylation levels at the N genomic sites using the first set of cell-free DNA molecules
2160
Identify a second set of cell-free DNA molecules, each located at any one of the loci, including a respective second allele, and including at least one of N genomic sites
2170
Measure N second mixture methylation levels at the N genomic sites using the second set of cell-free DNA molecules
For each of M tissue types
2180
Determine first fractional contributions using the N first mixture methylation levels and second fractional contributions using the N second mixture methylation levels
2190
Compute separation value between corresponding first fractional contribution and second fractional contribution
2195
Identify the first tissue type as being an origin of the copy number aberration based on a first separation value of the first tissue type having a highest value among the corresponding separate values

FIG. 21

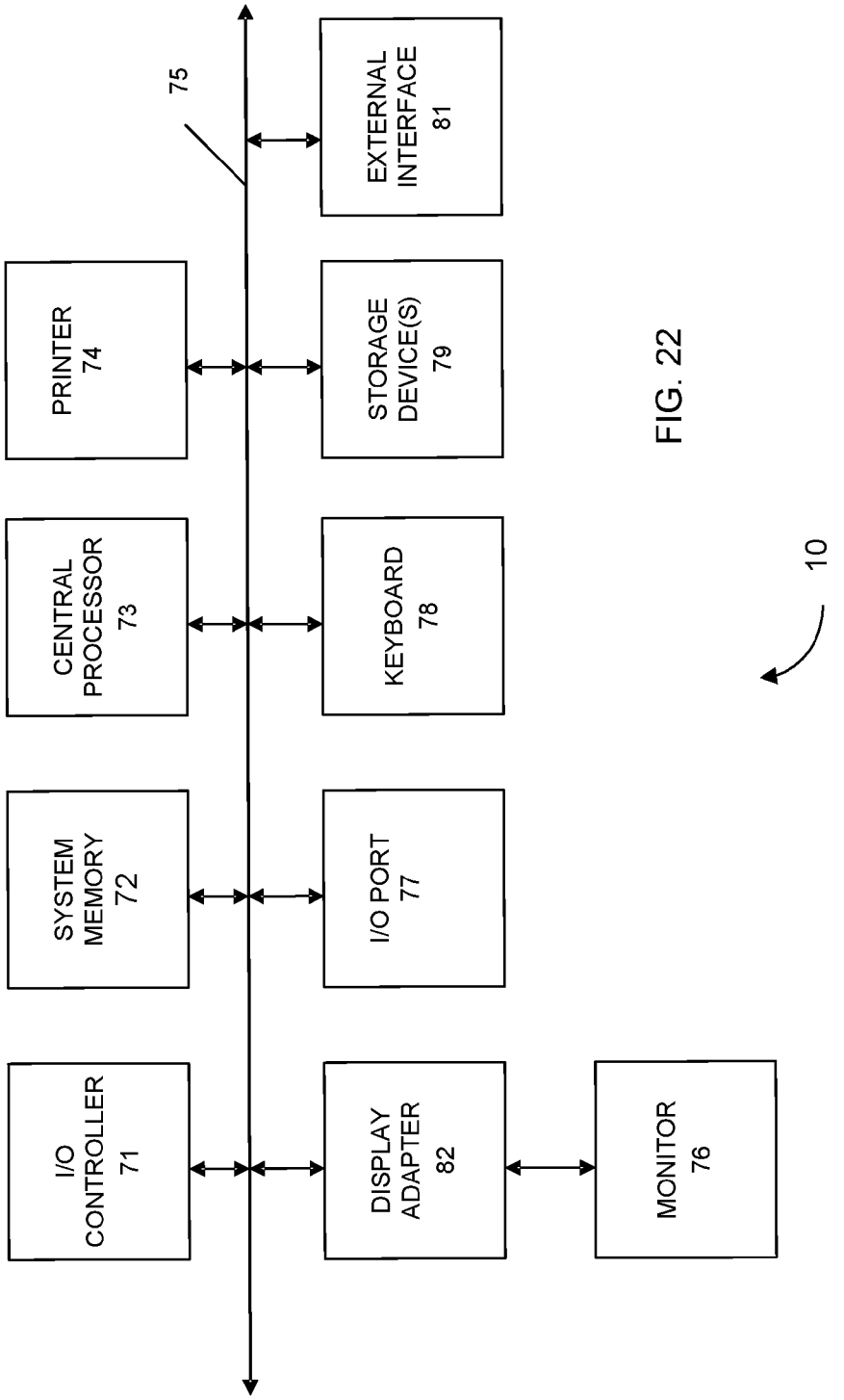

75
EXTERNAL INTERFACE 81
PRINTER 74
STORAGE DEVICE(S) 79
CENTRAL PROCESSOR 73
KEYBOARD 78
SYSTEM MEMORY 72
I/O PORT 77
I/O CONTROLLER 71
DISPLAY ADAPTER 82
MONITOR 76
10

FIG. 22

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/090642

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI,SIPOABS,CNABS,HKABS,CNKI,Pubmed,Google scholar:fetal, fetus, allele, haplotype, chromosome, parental, methylation, cell-free, non-invasive, THE CHINESE UNIVERSITY OF HONG KONG

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2014043763 A1 (UNIV HONG KONG CHINESEET AL.) 27 March 2014 (2014-03-27) see the whole document	1-65
A	WO 2011057094 A1 (UNIV HONG KONG CHINESEET AL.) 12 May 2011 (2011-05-12) see the whole document	1-65

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 October 2016	**19 October 2016**

Name and mailing address of the ISA/CN	Authorized officer
STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**LI,Kangqi**
Facsimile No. **(86-10)62019451**	Telephone No. **(86-10)62411034**

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/090642

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2014043763 A1	27 March 2014	JP 2015536639 A	24 December 2015
		AU 2013317708 A1	09 April 2015
		HK 1207124 A1	22 January 2016
		KR 20150082228 A	15 July 2015
		MX 2015003571 A	22 June 2015
		CA 2884066 A1	27 March 2014
		EP 2898100 A1	29 July 2015
		TW 201418474 A	16 May 2014
		EA 201500327 A1	30 July 2015
		SG 11201501927V A	29 April 2015
		HK 1206797 A1	15 January 2016
		EP 2898100 A4	11 May 2016
		IL 237495 A	30 April 2015
		CN 104781422 A	15 July 2015
		PH 12015500547 A1	04 May 2015
		IL 237495 D0	30 April 2015
WO 2011057094 A1	12 May 2011	IL 219521 A	26 February 2015
		CN 105779280 A	20 July 2016
		IL 219521 D0	28 June 2012
		EP 2496717 A1	12 September 2012
		AU 2010315037 B2	18 September 2014
		US 2013253844 A1	26 September 2013
		US 8467976 B2	18 June 2013
		AU 2010315037 A8	12 July 2012
		JP 5540105 B2	02 July 2014
		CA 2779695 A1	12 May 2011
		CA 2779695 C	24 May 2016
		TW I458976 B	01 November 2014
		MX 2012005214 A	21 September 2012
		AU 2010315037 B9	23 April 2015
		TW 201122470 A	01 July 2011
		US 2013323731 A1	05 December 2013
		JP 2014193165 A	09 October 2014
		EA 201200690 A1	30 May 2013
		CN 102770558 B	06 April 2016
		AU 2010315037 A1	07 June 2012
		CN 102770558 A	07 November 2012
		US 2011105353 A1	05 May 2011
		JP 2013509884 A	21 March 2013

Espacenet

Bibliographic data: KR101896167 (B1) — 2018-09-07

DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

Inventor(s): 로, 육-밍 데니스, ; 치우, 로사 와이 큔, ; 찬, 콴 치, ; 지, 베니 충 잉, ; 총, 카 춘

Applicant(s): 더 차이니즈 유니버시티 오브 홍콩

Classification: - international: ***C12Q1/68; G06F19/18; G06F19/22***
- cooperative: **C12Q1/68 (EA)**; **C12Q1/6809 (CN)**; **C12Q1/6827 (EP, CN, KR, US)**; **C12Q1/6869 (CN)**; **C12Q1/6883 (EP, CN, KR, US)**; **C12Q1/6888 (KR, US)**; **G01N33/48 (EA)**; G01N33/50 (EA); **G16B20/00 (CN, KR, US)**; **G16B20/10 (EP)**; **G16B30/00 (CN, KR)**; C12Q2600/112 (KR, US); C12Q2600/154 (CN, KR, US); C12Q2600/156 (KR, US); G01N2800/387 (KR, US); G16B30/00 (EP, US); Y02A90/10 (EP, US) more

Application number: KR20167021213 20080723 Global Dossier

Priority number(s): US20070951438P 20070723 ; WO2008GB02524 20080723

Also published as: AU2008278839 (A1) AU2008278839 (B2) AU2008278843 (A1) BRPI0814670 (A2) BRPI0814670 (B1) CA2693081 (A1) CA2693081 (C) CA2694007 (A1) CA2694007 (C) CA2900927 (A1) CA2900927 (C) CA3009992 (A1) CA3029497 (A1) CA3076142 (A1) CA3076159 (A1) CN101849236 (A) CN101971178 (A) CN101971178 (B) CN103849684 (A) CN103853916 (A) CN103853916 (B) CN103902809 (A) CN103902809 (B) CN106676188 (A) CN106834474 (A) CN106834474 (B) CN106834481 (A) CN106886688 (A) CN106886688 (B) CN107083424 (A) CN107083425 (A) CY1114773 (T1) CY1117525 (T1) DK2183693 (T3) DK2183693 (T4) DK2183693 (T5) DK2514842 (T3) DK2527471 (T3) DK2557519 (T3) EA017966 (B1) EA028642 (B1) EA035451 (B1) EA035451 (B9) EA201000231 (A1) EA201201551 (A1) EA201300072 (A1) EA201600280 (A1) EA201791612 (A2) EA201791612 (A3) EP2183692 (A1) EP2183692 (B1) EP2183693 (A1) EP2183693 (B1) EP2183693 (B2) EP2514842 (A2) EP2514842 (A3) EP2514842 (B1) EP2527471 (A2) EP2527471 (A3) EP2527471 (B1) EP2557517 (A2) EP2557517 (A3) EP2557518 (A2) EP2557518 (A3) EP2557518 (B1) EP2557519 (A2) EP2557519 (A3)

EP2557519 (B1) EP2557520 (A2) EP2557520 (A3)
EP3067807 (A1) EP3540739 (A1) EP3656870 (A1)
EP3745405 (A1) EP3770275 (A1) ES2441807 (T3)
ES2441807 (T5) ES2571738 (T3) ES2792802 (T3)
HK1144024 (A1) HK1177768 (A1) HK1182195 (A1)
HK1199067 (A1) HK1224033 (A1) HRP20140009 (T1)
HRP20140009 (T4) HRP20160493 (T1) HUE030510 (T2)
IL203311 (A) IL233261 (A) JP2010534068 (A) JP2010534069 (A)
JP2014073134 (A) JP2015142588 (A) JP2016185162 (A)
JP2017148073 (A) JP2019000113 (A) JP2019013245 (A)
JP2020031663 (A) JP2020031664 (A) JP2020115887 (A)
JP5519500 (B2) JP5736170 (B2) JP6151739 (B2)
JP6383837 (B2) JP6522554 (B2) JP6629940 (B2)
JP6695392 (B2) KR101646978 (B1) KR101829564 (B1)
KR101829565 (B1) KR101916456 (B1) KR101966262 (B1)
KR101972994 (B1) KR102060911 (B1) KR102076438 (B1)
KR102112438 (B1) KR102128960 (B1) KR102147626 (B1)
KR102197512 (B1) KR20100058503 (A) KR20100075826 (A)
KR20160030404 (A) KR20160113145 (A) KR20160113146 (A)
KR20160113147 (A) KR20160113148 (A) KR20170127073 (A)
KR20180100713 (A) KR20180114251 (A) KR20180121695 (A)
KR20190114039 (A) KR20190114041 (A) KR20190143494 (A)
KR20200055151 (A) KR20200100860 (A) KR20210006468 (A)
MX2010000846 (A) MX341573 (B) MX346069 (B) NZ582702 (A)
NZ600407 (A) PL2183693 (T3) PL2183693 (T4) PL2183693 (T5)
PL2514842 (T3) PT2183693 (E) SG183062 (A1) SI2183693 (T1)
SI2183693 (T2) SI2514842 (T1) US10208348 (B2)
US2009029377 (A1) US2009087847 (A1) US2014045181 (A1)
US2014256559 (A1) US2014256560 (A1) US2014329695 (A1)
US2014329696 (A1) US2019136323 (A1) US2020056242 (A1)
US8706422 (B2) US8972202 (B2) US9051616 (B2)
WO2009013492 (A1) WO2009013496 (A1) ZA201000524 (B) less

Abstract not available for KR101896167 (B1)
Abstract of corresponding document: WO2009013492 (A1)

Methods, systems, and apparatus are provided for determining whether a nucleic acid sequence imbalance exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, the ratio of the two sequences (or sets of sequences) are chosen. The cutoff value may be determined based at least in part on the percentage of fetal DNA in a sample, such as maternal plasma, containing a background of maternal nucleic acid sequences. The cutoff value may also be determined based on an average concentration of a sequence per reaction. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as sequential probability ratio testing (SPRT).

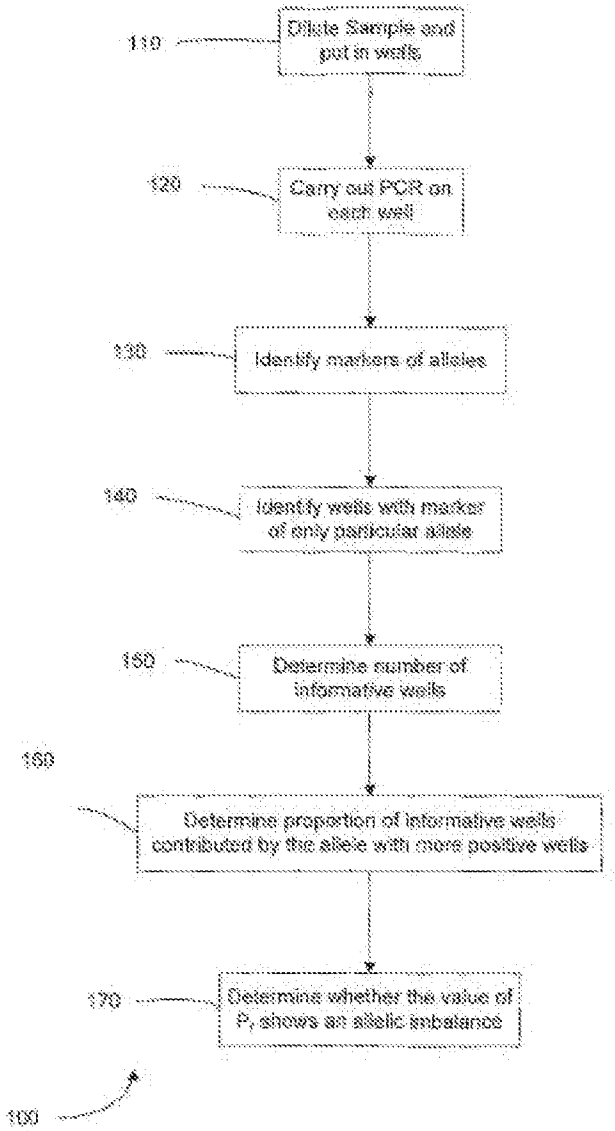

110
Dilute Sample and put in wells
120
Carry out PCR on each well
130
Identify markers of alleles
140
Identify wells with marker of only particular allele
150
Determine number of informative wells
160
Determine proportion of informative wells contributed by the allele with more positive wells
170
Determine whether the value of P_1 shows an allelic imbalance
100

FIG. 1

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년09월07일
(11) 등록번호 10-1896167
(24) 등록일자 2018년08월31일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) *C12Q 1/68* (2018.01)
G06F 19/22 (2017.01)
(52) CPC특허분류
G06F 19/18 (2013.01)
C12Q 1/6827 (2018.05)
(21) 출원번호 10-2016-7021213(분할)
(22) 출원일자(국제) 2008년07월23일
심사청구일자 2016년09월01일
(85) 번역문제출일자 2016년08월02일
(65) 공개번호 10-2016-0113147
(43) 공개일자 2016년09월28일
(62) 원출원 특허 10-2010-7003906
원출원일자(국제) 2008년07월23일
심사청구일자 2013년07월23일
(86) 국제출원번호 PCT/GB2008/002524
(87) 국제공개번호 WO 2009/013492
국제공개일자 2009년01월29일
(30) 우선권주장
60/951,438 2007년07월23일 미국(US)
(56) 선행기술조사문헌
JP2007515947 A*
KR1020020064298 A*
WO2007075836 A2*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
더 차이니즈 유니버시티 오브 홍콩
중국 홍콩 엔. 티. 새틴 더 차이니즈 유니버시티 오브 홍콩
(72) 발명자
로, 육-밍 데니스
홍콩, 구룡, 호만틴, 7 킹 탁 스트리트, 4 플로어
치우, 로사 와이 쿤
홍콩, 뉴 테리토리스, 타이 포, 1 훙 람 드라이브, 콘스텔레이션 코브, 블록 1, 플랫 1에이
(뒷면에 계속)
(74) 대리인
손민

전체 청구항 수 : 총 20 항 심사관 : 박진아

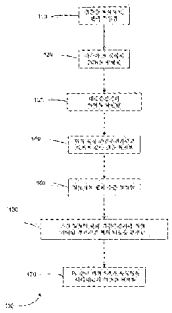
(54) 발명의 명칭 **핵산 서열 불균형의 결정**

(57) 요 약

본 발명에서는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하는 방법, 시스템 및 장치가 제공된다. 예를 들어, 2개의 서열 (또는 서열 세트)의 비의 불균형을 측정하기 위한 하나 이상의 분별값이 선택된다. 분별값은 적어도 부분적으로 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 중 태아 DNA의 백

(뒷면에 계속)

대 표 도 - 도1



분율을 기준으로 결정될 수 있다. 분별값은 또한 반응 당 서열의 평균 농도를 기준으로 결정될 수 있다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 순차 확률비 검증 (SPRT)을 사용하여 결정될 수 있다.

(52) CPC특허분류
C12Q 1/6883 (2018.05)
C12Q 1/6888 (2018.05)
G06F 19/22 (2018.05)
C12Q 2600/112 (2013.01)
C12Q 2600/154 (2013.01)
C12Q 2600/156 (2013.01)
G01N 2800/387 (2013.01)

(72) 발명자
찬, 콴 치
홍콩, 구룡, 메이 푸 선 추엔, 브로드웨이 스트리트, 블록 34, 13/에프, 플랫 에이
지, 베니 충 잉
홍콩, 뉴 테리토리스, 마 온 산, 라 코스타, 타워 2, 플랫 18이
총, 카 춘
홍콩, 뉴 테리토리스, 타이 포, 푸 신 에스테이트, 신 킹 하우스, 29/에프, 플랫 06

명 세 서

청구범위

청구항 1

핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된 비일시적 컴퓨터 판독가능 기록 매체로서, 상기 지시는 하기를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체:

다수의 반응으로부터 데이터를 수득하는 단계로, 상기 데이터는

(1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이터; 및

(2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 단계;

두 데이터 세트로부터 파라미터를 결정하는 단계;

생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값(cutoff value)을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 생물학적 샘플 내에 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 2

제1항에 있어서, 서열 불균형이 종양에서 염색체의 획득 또는 소실과 관련되는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 3

제1항에 있어서, 제1 유형의 세포가 제1 유기체로부터의 것이고, 제2 유형의 세포가 제2 유기체로부터의 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 4

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 제1 백분율의 결정이 여성 대상체와 태아 사이의 다형성 차이의 정량화를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 5

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 태아가 남성이며, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 Y 염색체 DNA 농도를 결정하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 6

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 제1 유전자좌에서의 태아-특이적 메틸화 패턴을 나타내는 DNA 분자의 양을 제1 유전자좌에서의 DNA 분자의 총량과 비교하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 7

제6항에 있어서, 태아-유도된 DNA 분자가 과-메틸화되고, 모체-유도된 DNA 분자가 저-메틸화되는 것인, 비일시

적 컴퓨터 판독가능 기록 매체.

청구항 8

제1항에 있어서, 상기 지시가

반응 당, 과소발현된 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인, 참조 핵산 서열의 제1 평균 농도를 결정하는 단계; 및

제1 평균 농도에 제1 백분율로부터 유도된 인자를 곱하여 참조 핵산 서열이 아닌 핵산 서열의 제2 평균 농도를 얻는 단계를 추가로 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 9

제8항에 있어서, 상기 지시가 참조 핵산 서열에 대한 데이타로부터 유도된 값이 입력된 역 확률 분포를 사용하여 다수의 반응 각각에서 참조 핵산 서열의 평균 농도를 결정하는 단계를 더 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 10

태아를 임신한 여성 대상체의 생물학적 샘플 내의 태아 DNA의 분획 농도를 결정하는 방법으로서, 생물학적 샘플은 여성 대상체 및 태아로부터 핵산 분자를 포함하고, 모체는 제1 대립유전자에 대한 제1 유전자좌에서 동형접합성이고 태아는 제1 대립유전자 및 제1 대립유전자와 상이한 제2 대립유전자에 대한 제1 유전자좌에서 이형접합성이며, 상기 방법은 하기를 포함하는 것인, 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제1의 다수의 반응으로부터 제1 데이터를 수득하는 단계로, 상기 반응은 다수의 관심 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내고, 상기 데이터는

(1) 제1 대립유전자에 대해 양성인 반응의 제1 수를 나타내는 제1 세트의 정량 데이터; 및

(2) 제2 대립유전자에 대해 양성인 반응의 제2 수를 나타내는 제2 세트의 정량 데이터를 포함하는, 단계; 및

제1 수를 제2 수와 비교하여 태아 DNA의 분획 농도를 결정하는 단계.

청구항 11

제10항에 있어서, 태아 DNA의 분획 농도가 제1 수 및 제2 수를 사용하여 백분율로서 결정되는 것인 방법.

청구항 12

제10항에 있어서, 하기 단계를 더 포함하는 것인 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제2의 다수의 반응으로부터 제2 데이터를 수득하는 단계로, 상기 데이터는

(1) 제1 유전자좌를 포함하지 않는 하나 이상의 임상적으로 관련된 핵산 서열의 제3 양을 나타내는 제3 세트의 정량 데이터; 및

(2) 하나 이상의 임상적으로 관련된 핵산 서열과 상이한 하나 이상의 배경 핵산 서열의 제4 양을 나타내는 제4 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 여성 대상체 및 태아로부터의 핵산 분자로부터 유래된 것인, 단계;

제3 및 제4 데이터 세트로부터 제3 양과 제4 양 사이의 상대적 양을 제공하는 파라미터를 결정하는 단계;

태아 DNA의 분획 농도로부터 제1 분별값을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 13

제12항에 있어서, 서열 분균형이 염색체 이수성(aneuploidy)인 방법.

청구항 14

제12항에 있어서, 임상적으로 관련된 핵산 서열 중 하나가 유전자 다형성의 대립유전자이고, 배경 핵산 서열 중 하나가 유전자 다형성의 또 다른 대립유전자인 방법.

청구항 15

제12항에 있어서, 제2의 다수의 반응을 수행하기 전에 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계를 더 포함하는 방법.

청구항 16

제15항에 있어서, 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계가 혼성화로 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열을 선별하기 위해 올리고뉴클레오티드 어레이를 사용하는 단계를 포함하는 것인 방법.

청구항 17

제15항에 있어서, 생물학적 샘플을 농축시키는 단계가 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열로부터 DNA를 증폭시키는 단계를 포함하는 것인 방법.

청구항 18

제12항에 있어서, 태아 DNA의 분획 농도가 파라미터 결정 후 결정되는 것인 방법.

청구항 19

제12항에 있어서, 제1의 다수의 반응이 제2의 다수의 반응과 동일한 것인 방법.

청구항 20

제10항에 있어서, 반응이 서열분석 반응 또는 증폭 반응인 방법.

발명의 설명

기 술 분 야

[0001] 우선권 주장

[0002] 본 출원은 2007년 7월 23일 출원된, 발명의 명칭이 "핵산 서열 불균형의 결정(DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE)"인 미국 가출원 번호 제60/951438호 (사건관리 번호 제016285-005200US호)를 우선권으로 주장하며 이의 정규출원이고, 상기 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0003] 관련 출원의 상호 참조

[0004] 본 출원은 또한 동시에 출원된, 발명의 명칭이 "게놈 서열분석을 사용한 태아 염색체 이수성의 진단(DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING)"인 정규출원 (사건관리 번호 제016285-005220US호)에 관한 것이고, 이 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0005] 발명의 분야

[0006] 본 발명은 일반적으로 2개의 상이한 핵산 서열간 불균형을 결정하는 유전자형 및 질환의 진단 시험, 보다 특히 모체 혈액 샘플의 시험을 통한 태아에서의 다운 증후군, 다른 염색체 이수성(aneuploidy), 돌연변이 및 유전자형의 확인에 관한 것이다. 본 발명은 또한 암의 검출, 이식의 모니터링 및 감염성 질환의 모니터링에 관한 것이다.

배경기술

[0008] 유전 질환, 암 및 다른 상태는 종종 2개의 상응하는 염색체 또는 대립유전자 또는 다른 핵산 서열에서의 불균형으로부터 초래되거나 이러한 불균형을 생성한다. 즉, 다른 서열에 비해 한 서열의 양이 정상적인 경우보다 많거나 적다. 통상적으로, 정상 비율은 대등한 50/50 비율이다. 다운 증후군 (삼염색체 21)은 추가 염색체 21의 불균형을 갖는 질환이다.

[0009] 삼염색체 21의 통상적인 출생 전 진단 방법은 한정된 태아 소실 위험을 낳는 침습적 절차, 예컨대 양수검사 또는 융모막 융모 생검에 의해 태아 물질을 샘플링하는 것을 포함한다. 비-침습적 절차, 예컨대 초음파검사 및 생화학 마커에 의한 스크리닝은 한정적인 침습적 진단 절차 전에 임신 여성을 위험-등급화하는데 사용되었다. 그러나, 이러한 스크리닝 방법은 전형적으로 핵심 염색체 이상 대신에 삼염색체 21과 관련된 부대징후를 측정하므로, 최적에 못미치는 진단 정확성 및 다른 단점들 (예컨대 재태 기간에 의해 많은 영향을 받음)을 갖는다.

[0010] 1997년에 모체 혈장 중 순환 무-세포 태아 DNA의 발견은 비-침습적 출생 전 진단의 새로운 가능성을 제공하였다 (문헌 [Lo, YMD and Chiu, RWK 2007 Nat Rev Genet 8, 71-77]). 이 방법이 성별-연계된 유전자 장애 (문헌 [Costa, JM et al. 2002 N Engl J Med 346, 1502]) 및 특정 단일 유전자 장애 (문헌 [Lo, YMD et al. 1998 N Engl J Med 339, 1734-1738])의 출생 전 진단에 용이하게 적용되었지만, 태아 염색체 이수성의 출생 전 검출에 적용하기에는 상당한 난점이 있었다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 첫째, 태아 핵산은 모체 혈장 중에서, 종종 분석에 방해가 될 수 있는 다량의 모체 기원 배경 핵산과 공존한다 (문헌 [Lo, YMD et al. 1998 Am J Hum Genet 62, 768-775]). 둘째, 태아 핵산은 모체 혈장 중에서 우세하게는 무-세포 형태로 순환하여, 태아 게놈 내 유전자 또는 염색체의 투여 정보 유도를 어렵게 만든다.

[0011] 이들 난점을 극복하는 유의한 개발이 최근에 이루어졌다 (문헌 [Benachi, A & Costa, JM 2007 Lancet 369, 440-442]). 한 접근법은 모체 혈장 중 태아-특이적 핵산을 검출하여, 모체 배경 방해 문제를 극복하는 것이다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 염색체 21의 투여는 태반-유래 DNA/RNA 분자에서 다형성 대립유전자의 비율로부터 추론되었다. 그러나, 이러한 방법은, 샘플이 보다 적은 양의 표적 유전자를 함유하는 경우 덜 정확하며, 하나의 다형성이 사용되는 경우 단지 그 집단의 아군인, 표적 다형성에 대해 이형접합성인 태아에게만 적용될 수 있다.

[0012] 달란 등(Dhallan et al)의 문헌 [Dhallan, R, et al. 2007, supra Dhallan, R, et al. 2007 Lancet 369, 474-481]에는, 모체 혈장에 포름알데히드를 첨가함으로써 순환 태아 DNA의 비율을 증가시키는 별법의 전략이 기재되어 있다. 모체 혈장 중 태아에 의해 기여된 염색체 21 서열의 비율은, 염색체 21 상의 단일 뉴클레오티드 다형성 (SNP)에 대한 부계-유전 태아-특이적 대립유전자 대 비-태아-특이적 대립유전자의 비율을 평가함으로써 결정되었다. SNP 비율은 참조 염색체에 대해 유사하게 추정되었다. 이어서, 태아 염색체 21의 불균형은, 염색체 21에 대한 SNP 비율과 참조 염색체에 대한 SNP 비율 사이의 통계적으로 유의한 차이를 검출함으로써 추론되었으며, 상기 유의성은 0.05 이하의 고정 p-값을 사용하여 정의된다. 고 집단 적용범위를 보장하기 위해, 염색체당 500개 초과의 SNP가 표적화되었다. 그러나, 고비율로 증가시키기 위한 포름알데히드의 효과와 관련하여 논쟁이 있었으므로 (문헌 [Chung, GTY, et al. 2005 Clin Chem 51, 655-658]), 방법의 재현성은 더 평가될 필요가 있다. 또한, 각각의 태아 및 모체는 각각의 염색체에 대해 상이한 수의 SNP의 정보를 제공하므로, SNP 비율 비교를 위한 통계적 시험의 영향력은 건 마다 변할 것이다 (문헌 [Lo, YMD & Chiu, RWK. 2007 Lancet 369, 1997]). 또한, 이들 접근법은 유전자 다형성의 검출에 좌우되므로, 이들 다형성에 이형접합성인 태아에게 국한된다.

[0013] 삼염색체 21 및 정배수체 태아로부터 수득한 양수세포 배양물에서 염색체 21 유전자좌 및 참조 유전자좌의 중합효소 연쇄 반응 (PCR) 및 DNA 정량화를 사용하여, 짐메르만 등(Zimmermann et al) (문헌 [2002 Clin Chem 48, 362-363])은 삼염색체 21 태아에서 염색체 21 DNA 서열의 1.5배 증가를 기준으로 상기 2군의 태아를 구별할 수 있었다. DNA 주형 농도의 2배 차이는 단지 한 임계 사이클(threshold cycle (Ct))의 차이를 구성하기 때문에, 1.5배 차이의 구별은 통상적인 실시간 PCR의 한계였다. 보다 미세한 정도의 정량 구별을 달성하기 위해, 별법의 전략이 필요하다. 따라서, 본 발명의 일부 실시양태는 이러한 목적을 위해 디지탈 PCR (문헌 [Vogelstein, B et al. 1999 Proc Natl Acad Sci USA 96, 9236-9241])을 사용한다.

[0014] 디지탈 PCR은 핵산 샘플에서 대립유전자 비율이 벗어난 것을 검출하기 위해 개발되었다 (문헌 [Chang, HW et al. 2002 J Natl Cancer Inst 94, 1697-1703]). 임상적으로, 디지탈 PCR은 종양 DNA 샘플에서 이형접합성의 소실 (LOH)을 검출하는데 유용한 것으로 나타났다 (문헌 [Zhou, W. et al. 2002 Lancet 359, 219-225]). 디지

탈 PCR 결과의 분석을 위해, 순차 확률비 검증 (SPRT)이 이전 연구에 의해 채택되어, 샘플 중 LOH의 존재를 제시하는 것과 제시하지 않는 것으로 실험 결과를 분류하였다 (문헌 [El Karoui at al. 2006 Stat Med 25, 3124-3133]). 이전 연구에서 사용된 방법에서, LOH를 결정하기 위한 분별값 (cutoff value)은 DNA 중 2개 대립유전자의 고정 참조 비율 2/3을 사용하였다. 모체 혈장 중 태아 핵산의 양, 비율 및 농도는 변할 수 있기 때문에, 이들 방법은 모체 혈장 중 모체 핵산의 배경에서 태아 핵산을 사용하여 삼염색체 21을 검출하는데 적합하지 않다.

[0015] 순환 태아 핵산 분석을 기준으로 하는 태아 삼염색체 21 (및 다른 불균형) 검출을 위한 비-침습적 시험, 특히 유전자 다형성 및/또는 태아-특이적 마커의 사용에 독립적인 비-침습적 시험을 수행하는 것이 바람직하다. 또한, 정확한 분별값 결정 및 서열 계수를 수행하여, 정확성에 필요한 모체 혈장 핵산 분자의 양 및/또는 데이타의 웰 수를 감소시킬 수 있음으로써, 증가된 효율 및 비용-효과를 제공하는 것이 바람직하다. 또한, 비-침습적 시험이 고 감수성 및 특이성을 가져 오류 진단을 최소화하는 것이 바람직하다.

[0016] 모체 혈장 중 태아 DNA 검출에 대한 또다른 적용은 베타-지중해빈혈과 같은 단일 유전자 장애의 출생 전 진단에 대한 것이다. 그러나, 태아 DNA는 단지 모체 혈장 중 소수 분획의 DNA를 구성하므로, 이러한 접근법은 단지 태아에서 부계로부터 유전되었지만 모계로부터는 존재하지 않는 돌연변이를 검출할 수 있을 뿐이라고 생각된다. 이러한 예로는, 베타-지중해빈혈을 일으키는 베타-글로빈 유전자의 코돈 41/42에서의 4 bp 결실 (문헌 [Chiu RWK et al. 2002 Lancet, 360, 998-1000]) 및 낭포성 섬유증을 일으키는 낭포성 섬유증 막횡단 전도 조절자 유전자의 Q890X 돌연변이 (문헌 [Gonzalez-Gonzalez et al 2002 Prenat Diagn, 22, 946-8])가 포함된다. 그러나, 베타-지중해빈혈 및 낭포성 섬유증 둘 다는, 이 질환이 그 자체로 표출되기 전에 태아에게 각각의 부모로부터 돌연변이가 유전될 필요가 있는 상염색체 열성 상태이므로, 부계-유전 돌연변이만의 검출은 단지 상기 질환에 걸린 태아를 가질 위험을 25%에서 50%로 증가시킬 것이다. 진단적으로 이는 이상적이 아니다. 따라서, 현존하는 접근법의 주된 진단 적용은, 부계-유전된 태아 돌연변이가 모체 혈장에서 검출될 수 없는 경우, 이후 태아가 동형접합성 질환 상태에 걸리는 것으로부터 배제될 수 있는 경우의 시나리오를 위해서일 것이다. 그러나, 진단적으로, 이러한 접근법은 부계 돌연변이의 음성 검출을 기준으로 결론을 내린다는 단점을 갖는다. 따라서, 상기 제한 없이, 완전한 태아 유전자형 (동형접합성 정상, 동형접합성 돌연변이 또는 이형접합성)이 모체 혈장으로부터 결정되도록 하는 접근법이 매우 바람직할 것이다.

발명의 내용

해결하려는 과제

[0018] 개요

[0019] 본 발명의 실시양태는, 핵산 서열 불균형 (예를 들어, 대립유전자 불균형, 돌연변이 불균형 또는 염색체 불균형)이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어, 2개의 서열 (또는 서열 세트) 양의 비율의 불균형을 결정하기 위한 하나 이상의 분별값이 선택된다.

[0020] 한 실시양태에서, 분별값은 적어도 부분적으로 생물학적 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 또는 혈청 또는 뇨 중 태아 (임상적으로 관련된 핵산) 서열의 백분율을 기준으로 결정된다. 또다른 실시양태에서, 분별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다.

[0021] 분별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 발견, 신뢰 구간, 수신자 판단 특성 (receiver operating characteristic, ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 시험의 양을 더 최소화시켰다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다.

[0022] 한 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하는 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 다수의 반응 각각에서, 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인 참조 핵산 서열의 평균 농도로부터 제1 분별값을 유도하는 단계; 파라미터를 제1 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를

결정하기 위한 방법이 제공된다.

[0023] 또다른 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하며, 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래된 것인 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서, 제1 유형 세포로부터의 핵산 서열의 양을 측정하여 생성된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법이 제공된다.

[0024] 본 발명의 다른 실시양태는 본원에 기재된 방법과 관련된 시스템 및 컴퓨터 판독가능 매체에 관한 것이다. 예시적 실시양태에 따르면, 본 발명은 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된, 비일시적 컴퓨터 판독가능 기록 매체에 관한 것으로서, 상기 지시는 (1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이터; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이터를 포함하고, 상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 다수의 반응으로부터 데이터를 수득하는 단계; 두 데이터 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 제1 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함한다.

[0025] 하기 상세한 설명 및 수반되는 도면을 참조로 하여 본 발명의 성질 및 이점을 보다 잘 이해할 수 있다.

발명의 실시 수단

[0027] 정의

[0028] 본원에 사용된 용어 "생물학적 샘플"은 대상체 (예를 들어, 임신 여성과 같은 인간)로부터 취하며 하나 이상의 관심 핵산 분자(들)을 함유하는 임의의 샘플을 지칭한다.

[0029] 용어 "핵산" 또는 "폴리뉴클레오티드" 단일가닥 또는 이중가닥 형태의 데옥시리보핵산 (DNA) 또는 리보핵산 (RNA) 및 이들의 중합체를 지칭한다. 구체적으로 한정하지 않는 한, 상기 용어는 참조 핵산과 유사한 결합 성질을 가지며 천연 발생 뉴클레오티드와 유사한 방식으로 대사되는 천연 뉴클레오티드의 공지된 유사체를 함유하는 핵산을 포함한다. 달리 나타내지 않는 한, 특정 핵산 서열은 또한 암묵적으로, 그의 보존적으로 변형된 변이체 (예를 들어, 퇴행 코돈 치환체), 대립유전자, 오소로그(ortholog), SNP 및 상보적 서열, 및 명백하게 나타내어진 서열을 포함한다. 구체적으로, 퇴행 코돈 치환체는 하나 이상의 선택된 (또는 모든) 코돈의 제3 위치가 혼합 염기 및/또는 데옥시이노신 잔기로 치환되는 서열을 생성함으로써 달성될 수 있다 (문헌 [Batzer et al, Nucleic Acid Res. 19:5081 (1991)], [Ohtsuka et al., J. Biol. Chem. 260:2605-2608 (1985)] 및 [Rossolini et al., Mol. Cell. Probes 8:91-98 (1994)]). 용어 핵산은 유전자, cDNA, mRNA, 소형 비코딩 RNA, 마이크로 RNA (miRNA), 피위(Piwi)-상호작용 RNA, 및 유전자 또는 유전자좌에 의해 코딩되는 단형 헤어핀 RNA (shRNA)와 상호교환적으로 사용된다.

[0030] 용어 "유전자"는 폴리펩티드 쇄를 생성하는데 관여하는 DNA 절편을 의미한다. 코딩 영역 전 및 후 영역 (리더 및 트레일러) 뿐만 아니라 개별 코딩 절편 (엑손) 사이의 삽입 서열 (인트론)을 포함할 수 있다.

[0031] 본원에 사용된 용어 "반응"은 관심있는 특정 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내는 화학적, 효소적 또는 물리적 작용에 관여하는 임의의 과정을 지칭한다. "반응"의 한 예는 중합효소 연쇄 반응 (PCR)과 같은 증폭 반응이다. "반응"의 또다른 예는 합성 또는 라이게이션에 의한 서열분석 반응이다. "정보제공 반응"은 관심있는 하나 이상의 특정 폴리뉴클레오티드 서열의 존재를 나타내며 단지 하나의 관심 서열이 존재하는 한가지 경우에서의 반응이다. 본원에 사용된 용어 "웰"은 제한된 구조, 예를 들어 웰-형상 바이알, 셀, 또는 PCR 어레이 챔버 내의 예정된 위치에서의 반응을 지칭한다.

[0032] 본원에 사용된 용어 "임상적으로 관련된 핵산 서열"은 잠재적 불균형에 대해 시험될 보다 큰 게놈 서열의 절편에 상응하는 폴리뉴클레오티드 서열 또는 보다 큰 게놈 서열 그 자체를 지칭할 수 있다. 한 예는 염색체 21의 서열이다. 다른 예는 염색체 18, 13, X 및 Y를 포함한다. 또다른 예는 돌연변이 유전자 서열 또는 유전자 다

형성 또는 부모의 한쪽 또는 양쪽으로부터 태아에게 유전될 수 있는 카피 수 변형을 포함한다. 또다른 예는 악성 종양에서 돌연변이, 결실 또는 증폭된 서열, 예를 들어 이형접합성 소실 또는 유전자 복제가 일어나는 서열을 포함한다. 일부 실시양태에서, 다수의 임상적으로 관련된 핵산 서열 또는 동등하게는 임상적으로 관련된 핵산 서열의 다수의 마커는 불균형 검출을 위한 데이타를 제공하는데 사용될 수 있다. 예를 들어, 염색체 21 상의 5개의 비-연속 서열로부터의 데이타는 가능한 염색체 21 불균형의 결정을 위해 추가적인 방식으로 사용되어, 필요한 샘플 부피를 1/5로 유효하게 감소시킬 수 있다.

[0033] 본원에 사용된 용어 "배경 핵산 서열"은 임상적으로 관련된 핵산 서열에 대한 정상 비율이 공지되어 있는 (예를 들어 1 대 1 비율인) 핵산 서열을 지칭한다. 한 예로서, 배경 핵산 서열 및 임상적으로 관련된 핵산 서열은 이형접합성으로 인해 구별되는, 동일한 염색체로부터의 2개의 대립유전자이다. 또다른 예로, 배경 핵산 서열은 임상적으로 관련된 핵산 서열인 다른 대립유전자에 대해 이형접합성인 한 대립유전자이다. 또한, 각각의 배경 핵산 서열 및 임상적으로 관련된 핵산 서열 중 일부는 상이한 개체로부터 유래될 수 있다.

[0034] 본원에 사용된 용어 "참조 핵산 서열"은 반응 당 평균 농도가 공지되어 있거나 동등하게는 측정되는 핵산 서열을 지칭한다.

[0035] 본원에 사용된 용어 "과다표출된(overrepresented) 핵산 서열"은 2개의 관심 서열 (예를 들어, 임상적으로 관련된 서열 및 배경 서열) 중 생물학적 샘플에서 다른 서열보다 풍부한 핵산 서열을 지칭한다.

[0036] 본원에 사용된 용어 "~기준으로"는 "적어도 부분적으로 ~기준으로"를 의미하며, 방법의 투입과 방법의 산출의 관계에서 일어나는, 한 값 (또는 결과)이 또다른 값의 결정에 사용되는 것을 지칭한다. 본원에 사용된 용어 "유도"는 그 유도가 식의 계산인 경우에 일어나는, 방법의 투입과 방법의 산출의 관계를 지칭한다.

[0037] 본원에 사용된 용어 "정량 데이타"는 하나 이상의 수치 값을 제공하는 하나 이상의 반응으로부터 얻은 데이타를 의미한다. 예를 들어, 특정 서열에 대한 형광 마커를 나타내는 웰의 수는 정량 데이타일 것이다.

[0038] 본원에 사용된 용어 "파라미터"는 정량 데이타 세트, 및/또는 정량 데이타 세트들 사이의 수치 관계를 특성화하는 수치 값을 의미한다. 예를 들어, 제1 핵산 서열의 제1 양과 제2 핵산 서열의 제2 양 사이의 비율 (또는 비율의 함수)이 파라미터이다.

[0039] 본원에 사용된 용어 "분별값"은 생물학적 샘플에 대한 분류의 2가지 이상의 상태 (예를 들어 질환 상태 및 비질환 상태)를 중재하는데 사용되는 수치 값을 의미한다. 예를 들어, 파라미터가 분별값보다 큰 경우 정량 데이타의 제1 분류가 만들어지거나 (예를 들어 질환 상태), 또는 파라미터가 분별값보다 작은 경우 정량 데이타의 상이한 분류가 만들어진다 (예를 들어 비질환 상태).

[0040] 본원에 사용된 용어 "불균형"은 하나 이상의 분별값으로 정의되는, 참조 정량으로부터 임상적으로 관련된 핵산 서열의 정량의 임의의 유의한 편차를 의미한다. 예를 들어, 참조 정량이 3/5의 비율일 수 있고, 따라서 측정 비율이 1:1인 경우에 불균형이 일어날 것이다.

[0042] 발명의 상세한 설명

[0043] 본 발명은, 다른 비-임상적으로 관련된 서열 (예를 들어, 염색체 또는 대립유전자 불균형)에 관하여 임상적으로 관련된 핵산 서열의 참조 (예를 들어 비질환 상태) 양에 비교되는 증가 또는 감소가 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어 2개의 서열 (또는 서열 세트) 양의 비율과 관련하여 참조 양에 비교되는 변화 (즉 불균형)이 존재하는지 여부를 결정하기 위해 하나 이상의 분별값이 선택된다. 참조 양에서 검출된 변화는 다른 비-임상적 관련 서열에 대한 임상적 관련 핵산 서열에 관하여 임의의 편차 (상향 또는 하향)일 수 있다. 따라서, 참조 상태는 임의의 비율 또는 다른 양 (예를 들어 1-1 대응 이외의 것)일 수 있고, 변화를 의미하는 측정 상태는 하나 이상의 분별값에 의해 결정된 바와 같이 참조 양과 상이한 다른 양 또는 임의의 비율일 수 있다.

[0044] 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래될 수 있다. 예를 들어, 태아/태반 세포로부터 유래된 태아 핵산 서열은 생물학적 샘플, 예컨대 모체 세포로부터 유래된 모체 핵산 서열의 배경을 함유하는 모체 혈장 중에 존재한다. 따라서, 한 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플 중 제1 유형 세포의 백분율을 기준으로 결정된다. 샘플 중 태아 서열의 백분율은 임의의 태아-유래 유전자좌에 의해 결정될 수 있으며, 임상적으로 관련된 핵산 서열을 측정하는 것에 국한되지 않는다는 것을 주목한다. 또다른 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플, 예컨대 체내 비-악성 세포로부터 유래된 핵산 서열의 배경을 함유하는 혈장, 혈청, 타액 또는 뇨 중 종양 서열의 백분율을

기준으로 결정된다.

[0045] 또다른 실시양태에서, 분별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 웰의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 발견, 신뢰 구간, 수신자 판단 특성 (ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 시험의 양을 더 최소화시킨다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다. 디지탈 PCR이 제시되더라도, 다른 방법이 사용될 수 있다.

[0046] 디지탈 PCR은 대부분의 양성 증폭이 단일 주형 분자로부터의 신호를 반영하도록 매우 희석된 핵산에 대한 멀티플렉스 PCR 분석을 포함한다. 이로써 디지탈 PCR은 개별 주형 분자의 계수를 가능하게 한다. 분석된 PCR의 총 수 중 양성 증폭의 비율로 원래 또는 비-희석 샘플 중의 주형 농도를 평가할 수 있다. 이러한 기술은 다양한 유전 현상의 검출을 가능하게 하도록 제시되었으며 (상기 문헌 [Vogelstein, B et al. 1999]), 종양 샘플 중 이형접합성 소실의 검출 (상기 문헌 [Zhou, W. et al. 2002]) 및 암 환자의 혈장 중 이형접합성 소실의 검출 (상기 문헌 [Chang, HW et al. 2002])을 위해 이전부터 사용되었다. 디지탈 PCR에 의한 주형 분자 정량화는 리포터 염료와 핵산 농도 사이의 투여량-반응 관계에 의존하지 않기 때문에, 그의 분석 정확도는 실시간 PCR의 분석 정확도보다 이론적으로 월등하여야 한다. 따라서, 디지탈 PCR은 표적 유전자좌와 참조 유전자좌 사이의 보다 미세한 정도의 정량 차이를 잠재적으로 구별할 수 있게 한다.

[0047] 이를 시험하기 위해, 본 발명자들은 먼저 디지탈 PCR이 모체 혈장 중 염색체 21로부터의 태반 전사체인 PLAC4 mRNA (문헌 [Lo, YMD, et al. 2007 Nat Med 13, 218-223]) 대립유전자 비율을 결정하고, 이로써 삼염색체 21 및 정배수체 태아를 구별할 수 있는지 여부를 평가하였다. 이러한 접근법을 디지탈 RNA-SNP 방법이라 칭한다. 이어서, 본 발명자들은 디지탈 PCR의 증가된 정확도로 인해 유전자 다형성에 좌우되지 않으면서 태아 염색체 이수성을 검출할 수 있는지 여부를 평가하였다. 본 발명자들은 이를 디지탈 상대적 염색체 투여 (RCD) 분석이라 칭한다. 전자의 접근법은 다형성-의존적이지만 정량 구별에 있어서 보다 낮은 정확도를 필요로 하는 반면, 후자의 접근법은 다형성-독립적이지만 정량 구별에 있어서 보다 높은 정확도를 필요로 한다.

[0049] I. 디지탈 RNA-SNP

[0050] **A. 개관**

[0051] 디지탈 PCR은 DNA 샘플 중 두 대립유전자의 벗어난(skewing) 대립유전자 비율의 존재를 검출할 수 있다. 예를 들어, 디지탈 PCR은 종양 DNA 샘플 중 이형접합성 소실 (LOH)을 검출하는데 사용되어 왔다. DNA 샘플 중 두 대립유전자, 즉 A 및 G가 존재하고, A 대립유전자는 LOH가 있는 세포 중에 소실된다고 가정한다. LOH가 종양 샘플 중 50%의 세포에 존재하는 경우, DNA 샘플 중 G:A의 대립유전자 비율은 2:1일 것이다. 그러나, LOH가 종양 샘플에 존재하지 않는 경우, G:A의 대립유전자 비율은 1:1일 것이다.

[0052] 도 1은 디지탈 PCR 실험을 설명하는 흐름도 100이다. 단계 110에서, DNA 샘플을 희석시킨 후, 독립된 웰로 분배한다. 본 발명자들은 일부 혈장 핵산 종이 원래의 샘플에서 이미 상당히 희석되도록 결정했다는 것을 주목한다. 따라서, 필요한 농도로 이미 존재하는 경우, 일부 주형에 대한 희석은 필요하지 않다. 이전의 연구 (예를 들어, 상기 문헌 [Zhou et al 2002])에서는, 특정 "주형 DNA"의 평균 농도가 웰 당 두 주형 중 하나의 대략 0.5 분자가 될 정도로 DNA 샘플을 희석시켰다. 용어 "주형 DNA"는 A 또는 G 대립유전자를 가리키는 것으로 보이고, 이러한 특정 농도에 대한 이론적 근거는 제공되지 않았다는 것을 주목한다.

[0053] 단계 120에서, 각각의 웰에서, PCR 과정은 A 및/또는 G 대립유전자를 동시에 검출하기 위해서 실시된다. 단계 130에서, 각각의 웰 중 마커가 식별된다 (예를 들어, 형광으로) (예를 들어 A, G, A 및 G, 또는 어느 쪽도 아님). LOH의 부재 하에, DNA 샘플 중 A 및 G 대립유전자의 과다는 동일하다 (셀 당 각각 한 카피). 따라서, 웰이 A 대립유전자 및 G 대립유전자에 대해 양성일 확률은 동일할 것이다. 이것은 유사한 수의 웰이 A 또는 G 대립유전자에 대해 양성인 것에 의해 반영될 것이다. 그러나, LOH가 종양 샘플 중 50% 또는 그 보다 많은 세포에 존재하는 경우, G 및 A 대립유전자의 대립유전자 비율은 적어도 2:1일 것이다. 이전의 방법들은 단순히 샘플이 적어도 50% 양성인 것으로 추정하였다. 따라서, 웰이 G 대립유전자에 대해 양성일 확률은 A 대립유전자에 대해 양성일 확률보다 높을 것이다. 그 결과, G 대립유전자에 대해 양성인 웰의 수는 A 대립유전자에 대해 양성인 웰의 수보다 더 많을 것이다.

[0054] 단계 140에서, 디지탈 PCR 결과를 분류하기 위해, 각각의 대립유전자에 대해 양성이나 다른 대립유전자에 대해서는 양성이 아닌 웰의 수를 계수할 것이다. 상기 예에서, A 대립유전자에 대해 양성이나 G 대립유전자에 대해

음성인 웰의 수, 및 G 대립유전자에 대해 양성이나 A 대립유전자에 대해 음성인 웰의 수를 계수한다. 한 실시양태에서, 덜 양성인 웰을 나타내는 대립유전자는 참조 대립유전자로 간주된다.

[0055] 단계 150에서, 정보제공 웰의 총 수는 두 대립유전자 중 어느 하나에 대해 양성인 웰의 수의 합으로 결정된다. 단계 160에서, 더 많은 양성 웰의 대립유전자에 의해 기여되는 정보제공 웰의 비율 (P_r) (파라미터의 한 예)이 계산된다.

[0056] P_r = 더 많은 양성 웰의 대립유전자에 대해서만 양성인 웰의 수/단지 하나의 대립유전자 (A 또는 G)에 대해 양성인 웰의 총 수.

[0057] 다른 실시양태는 하나의 대립유전자를 갖는 모든 웰을 하나 이상의 대립유전자를 갖는 모든 웰로 나눈 것을 사용할 수 있다.

[0058] 단계 170에서, P_r 값이 대립유전자 불균형을 나타내는지 여부가 결정된다. 정확하고 효율적인 것이 바람직하므로, 이 작업은 간단하지 않다. 불균형을 결정하기 위한 한 방법은 베이지안(Bayesian)-유형 우도 방법, 순차 확률비 검증 (SPRT)을 사용한다. SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 방법이다. 즉, SPRT는 디지탈 PCR의 결과를 분류하여 벗어난 대립유전자의 존재 또는 부재를 시사하는 통계적 방법이다. SPRT는 분석될 웰의 수를 최소화하여 소정의 통계적 검증력 및 정확도를 획득하도록 하는 이점을 갖는다.

[0059] 예시적 SPRT 분석에서, 실험 결과는 귀무가설 (null hypothesis) 및 대립가설 (alternative hypothesis)에 대해 시험될 것이다. 대립가설은 샘플 중 벗어난 대립유전자 비율이 있는 경우 수용된다. 귀무가설은 샘플 중 벗어난 대립유전자 비율이 없는 경우 수용된다. P_r 값은 귀무가설 또는 대립가설을 수용하기 위해서 두 분별값과 비교될 것이다. 두 가설 모두 수용되지 않는 경우, 샘플은 관찰된 디지탈 PCR 결과가 바람직한 통계적 신뢰도로 샘플을 분류하기에 충분하지 않다는 것을 의미하는 미분류로 표시될 것이다.

[0060] 귀무가설 또는 대립가설을 수용하기 위한 분별값은 전형적으로, 가설에서 언급된 가정 하에 고정된 P_r 값을 기준으로 계산되어 왔다. 귀무가설에서, 샘플은 벗어난 대립유전자 비율을 나타내지 않는 것으로 가정된다. 따라서, 각 웰이 A 및 G 대립유전자에 대해 양성일 확률은 동일할 것이므로, 예상 P_r 값은 1/2일 것이다. 대립가설에서, 예상 P_r 값은 2/3 또는 0.5와 2/3의 대략 중간 정도, 예를 들어 0.585로 여겨진다. 또한, 제한된 수의 실험으로 인해, (.585+3/N)의 상한값 및 (.585-3/N)의 하한값이 선택될 수 있다.

[0062] **B. 다운 증후군의 검출**

[0063] 본 발명의 한 실시양태에서, 디지탈 SNP는 임신 여성의 혈장으로부터 태아 다운 증후군을 검출하는데 사용된다. 태아/태반 세포에 특이적인 마커를 사용하여, 염색체 21 중 대립유전자의 비율을 측정할 수 있다. 예를 들어, PLAC4 대립유전자의 관찰된 과다표출 정도가 통계적으로 유의한지를 결정하기 위해서, SPRT가 사용된다.

[0064] 하나의 예시적 실시양태에 따라, 디지탈 RNA-SNP는 염색체 21로부터 전사되고 태반에 의해 발현되는 PLAC4 mRNA에 위치한 A/G SNP의 다형성 대립유전자, rs8130833의 비율에서의 불균형을 결정한다. 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 한편 (1:1 게놈 비율); 삼염색체 21에서, 삼염색체성 염색체 21은 태아 게놈 중 하나의 SNP 대립유전자의 추가 카피와 회합되어 2:1 비율을 나타낼 것이다. 디지탈 PCR 분석의 목적은 분석된 샘플 중 두 PLAC4 대립유전자의 양이 동등한지 또는 다른지를 결정하는 것이다. 따라서, A 및 G PLAC4 대립유전자는 모두 표적 주형이다. 실시간 PCR 검정은 PLAC4 mRNA를 증폭시키도록 설계되었고, 두 SNP 대립유전자는 태크만(TaqMan) 형광 프로브에 의해 구별되었다. 분석 단계의 도식적 설명이 도 2a에 나타나 있다.

[0065] 도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 방법 200을 설명한다. 단계 210에서, 샘플을 얻는다. 단계 220에서, 핵산 서열, 예를 들어 PLAC4 mRNA를 추출된 RNA 샘플에서 정량화한다. 한 실시양태에서, 이것은 PLAC4 mRNA에 대해 실시간 PCR로 실시된다. 한 측면에서, 이 단계는 조작자에게 표적이 디지탈 PCR 분석의 '영역'에 도달하기 전에 어느 만큼의 희석이 요구되는지에 대한 아이디어를 제공한다.

[0066] 단계 230에서, 샘플을 희석시킨다. 단계 240에서, 희석시킨 샘플의 농도를 측정한다. 희석시킨 샘플의 농도가 약 1 주형/웰 (즉, 참조 또는 비-참조 서열 또는 어느 하나의 대립유전자)인지 확인할 수 있다. 일부 실시양태는 이 측정을 위한 IV절에 기술된 기술을 사용한다. 예를 들어, 본 발명자들은 실시간 PCR 분석을 위해서 희석시킨 샘플을 96 웰로 배분하여 알맞게 희석되었는지를 확인하였다. 희석 농도는 또한 모르는 상태로 남겨두어

이 단계가 제거되도록 할 수 있고, 이는 추후 설명될 것이다.

[0067] 단계 250에서, 디지탈 PCR을 어레이의 각 웰 상에서 실시하였다. 예를 들어, 동일하게 희석시킨 샘플을 실시간 PCR 분석을 위해서 384-웰에 배분하였다. PCR 결과로부터, 각 핵산 서열에 대한 마커의 양 및 정보제공 웰의 수를 확인한다. 정보제공 웰은 A 또는 G 대립유전자에 대해서만 양성이나, 둘 모두에 대해서 양성은 아닌 것으로 정의된다. 단계 260에서, 예상 P_r 값을 계산한다. 이러한 단계들은 추후 더욱 자세하게 설명될 것이다. 이 계산은 단계 250에서 측정된 값들로부터 파라미터를 결정하는 것을 포함한다. 예를 들어, 웰 당 실제 평균 주형 농도가 계산될 수 있다.

[0068] 단계 270에서, SPRT 또는 다른 우도-비율 시험은 불균형이 존재하는지 여부를 결정하기 위해서 실시될 수 있다. 정배수체 사례에 대해, 동일한 수의 A-양성 및 G-양성 웰이 예상된다. 그러나, 삼염색체 21 태아로부터의 주형 분자가 분석되는 경우, 하나의 대립유전자만을 함유하는 웰의 수는 다른 대립유전자만을 함유하는 웰의 수보다 많아야 한다. 요컨대, 삼염색체 21에 대한 대립유전자 불균형이 예상된다.

[0069] 상기 언급된 것과 같이, SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 베이지안-유형 우도 방법이다. 삼염색체 21 검출을 위한 디지탈 PCR 분석에서, 대립가설은 대립유전자 불균형이 존재하는 경우 (즉, 삼염색체 21이 검출되는 경우) 수용되고; 귀무가설은 대립유전자 불균형이 존재하지 않는 경우 (즉, 삼염색체 21이 검출되지 않는 경우) 수용된다. 더 많은 수의 계수의 대립유전자는 잠재적으로 과다표출된 대립유전자로 지칭되고, 모든 정보제공 웰 중 그의 비율 (P_r)이 계산될 것이다. SPRT는, P_r이 삼염색체 21 샘플에 대해 예상되는 충분한 정도의 대립유전자 불균형을 나타내는지를 결정하기 위해서 적용된다.

[0070] 조작 면에 있어서, SPRT는 어느 하나의 가설을 수용하거나 또는 거절하기 위한 확률 경계를 정의하기 위해서 구축된 한 쌍의 SPRT 곡선을 갖는 그래프를 사용하여 적용되고 해석될 수 있다. 도 3은 본 발명의 한 실시양태에 따라 다운 증후군을 결정하기 위한 SPRT 곡선을 갖는 그래프를 설명한다. SPRT 곡선은 신뢰성 있게 분류될 수 있는 경우 주어진 정보제공 웰의 총 수 (x-축)에 대해 잠재적으로 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 요구되는 비율 P_r (y-축)을 플롯팅한다. 도 3에 설명된 것과 같이, 상단 곡선은 대립가설을 수용하기 위한 확률 경계를 설정하고; 하단 곡선은 귀무가설을 수용하기 위한 확률 경계를 설정한다.

[0071] 실험적으로 유도된 P_r 값은 어느 하나의 가설을 수용하거나 또는 거절하기 위한 예상 P_r 값과 비교될 것이다. 귀무가설이 수용되는 경우, 샘플은 정배수체 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 대립가설이 수용되는 경우, 샘플은 삼염색체 21 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 별법으로, 주어진 정보제공 계수의 수에 대한 P_r이 질환 분류에 대해 요구되는 통계적 신뢰성 수준에 아직 도달하지 못한 경우, 어느 하나의 가설도 받아들여질 수 없다. 이러한 경우들은 더 많은 데이타가 이용가능하기 전까지는 분류불가능으로 간주된다. 질환 분류가 가능하지 않은 경우, 집계된 데이타가 SPRT에 의해 분류가능해지기 전까지 추가의 384-웰 플레이트를 실시할 수 있다.

[0072] 따라서, SPRT는 주어진 수준의 신뢰성을 위해서 다른 통계적 방법보다 더 적은 양의 시험이 요구된다는 이점을 제공한다. 실용적인 측면에서, SPRT는 요구되는 양의 데이타가 축적되어 불필요한 추가적 분석을 최소화하자마자 어느 하나의 가설의 수용 또는 거절을 가능하게 한다. 이 특징은 일반적으로 이용가능한 주형 분자의 수가 제한적인 경우 낮은 농도로 존재하는 혈장 핵산의 분석과 특히 관련이 있다. 엄격한 분류에 추가하여, 분류는 또한 백분율 정확성을 포함할 수 있다. 예를 들어, 분별값과의 비교로 인한 분류는 샘플이 특정 백분율의 핵산 서열 불균형의 우도를 나타내거나, 또는 동등하게는 측정된 불균형이 특정 백분율 또는 다른 값에 정확하다는 것을 제공할 수 있다.

[0073] 유사한 접근법을 적용하여, 모체 혈장 또는 혈청 중 태아 핵산을 사용하여 돌연변이 또는 유전자 다형성과 관련한 태아의 유전자형을 결정할 수 있다. 태아는 그의 모계로부터 그의 게놈의 절반을 물려받을 것이라는 것을 상기해야 한다. 설명을 위해서, 두 대립유전자 A 및 B가 있는 특정 유전자좌를 고려한다. 모계가 AB 유전자형의 이형접합체인 경우, 태아는 이론적으로는 AA, BB 또는 AB의 유전자형을 가질 수 있다. 태아가 AB 유전자형, 즉 모계와 같은 유전자형을 갖는 경우, 모체 혈장에는 (모계 및 태아 모두로부터의) AB 유전자형의 핵산만 존재할 것이다. 따라서, 핵산 또는 대립유전자 균형이 모체 혈장에서 보여진다. 반면, 태아가 AA 또는 BB의 유전자형을 갖는 경우, 모체 혈장에서 A 또는 B 대립유전자 각각이 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 생각은 또한 질환-유발성 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 척수 근위축증을 유발하는 돌연변이)에도 적용가능한데, 각 경우 A는 야생형 대립유전자로, B는 돌연변이 대립유전자로 고

려될 수 있다.

[0075] II. 디지탈 RCD

[0076] 디지탈 RNA-SNP의 단점은 디지탈 RNA-SNP가 분석된 SNP에 대한 이형접합체의 경우에만 적용가능하다는 것이다. 순환성 태아 핵산 분석을 기준으로 한 삼염색체 21 태아 또는 다른 태아 염색체 이수성 (예를 들어, 삼염색체 18, 13 및 성 염색체 이수성)을 검출하기 위한 비침습성 시험이 유전자 다형성의 사용과 무관한 경우 디지탈 RNA-SNP가 이상적이라는 것은 하나의 개선된 점이다. 따라서, 한 실시양태에서, 염색체 양은 참조 염색체, 즉 이 연구에서 염색체 1에 위치한 유전자좌에 상대적인 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정된다. 정배수체 태아의 게놈에서 염색체 21 대 염색체 1의 비율의 2:2로부터의 변화는 삼염색체 21의 경우와 구분된다. 삼염색체 21 검출을 위한 디지탈 PCR 분석에서, 비교할 두 가설은 염색체 불균형이 존재하지 않는 (즉, 삼염색체 21이 검출되지 않는) 귀무가설 및 염색체 불균형이 존재하는 (즉, 삼염색체 21이 검출되는) 대립가설일 것이다.

[0077] 이 접근법은 다른 염색체 이수성, 예를 들어 삼염색체 18에서의 염색체 18, 삼염색체 13에서의 염색체 13, 터너 증후군에서의 염색체 X와 연관된 다른 염색체에 일반화될 수 있다. 또한, 염색체 1과는 별개로, 관련 이수성과 연관되지 않은 다른 염색체도 또한 참조 염색체로 사용될 수 있다. 유사한 접근법이 또한 부분적으로, 암에서 통상적으로 결실되는 염색체 대 참조 염색체의 비율 변화를 분석함으로써 암 검출에 적용될 수 있다. 전자의 예로는 결장직장암에서의 염색체 5q, 폐암에서의 염색체 3p 및 비인두 암종에서의 염색체 9p가 포함된다. 도 2b는 서열 불균형으로 나타나는 일부 통상적인 암-관련 염색체 이상을 열거한다.

[0078] 도 2a는 또한 본 발명의 한 실시양태에 따른 디지탈 RCD 방법 205를 설명한다. 단계 220-230에 대한 한 실시양태에서, 추출된 DNA를, 예를 들어, 나노드롭(Nanodrop) 기술로 정량화하고, 웰 당 염색체 21 또는 표준화 염색체 (예컨대, 염색체 1)로부터의 약 하나의 표적 주형의 농도로 희석시킨다. 단계 240의 한 실시양태에서, 384-웰 플레이트 중 태크만 프로브를 모두 사용하는 디지탈 RCD 분석을 진행하기 전에 약 37% 수준의 웰이 음성인지 여부를 확인하기 위해서, 단지 96-웰 포맷에서 염색체 1 프로브를 사용하는 검정에 의해 희석시킨 DNA 샘플을 분석함으로써 확인할 수 있다. 37%의 유의성은 IV절에서 추후 기술될 것이다.

[0079] 단계 240의 시험 및 단계 250의 결과는 한 쌍의 태크만 프로프에 의해 구별되는 파라로그(paralogous) 서열 변이에 의해 식별되는 두 염색체 모두에 존재하는 파라로그 서열을 증폭시키도록 고안된 실시간 PCR 검정으로 수행될 수 있다 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915]). 이러한 맥락에서, 정보제공 웰은 염색체 21 또는 염색체 1 유전자좌에 대해 양성인, 그러나 둘 모두에 대해서 양성은 아닌 웰로서 정의된다. 정배수체 태아에 대해, 어느 하나의 유전자좌에 대해 양성인 정보제공 웰의 수는 대략 동등해야 한다. 삼염색체 21 태아에 대해, 염색체 1보다 염색체 21에 대해 양성인 웰의 과다표출이 있어야 한다. 과다표출의 정확한 비율은 하기 절에 기술되어 있다.

[0081] III. 태아 서열의 혼입률

[0082] 상기 기술된 방법 200 및 205의 실시양태의 단점은 태아 특이적 마커가 요구된다는 것이다. 따라서, 본 발명의 한 실시양태에서, 비-태아 특이적 마커가 사용된다. 비-태아 특이적 마커를 사용하기 위해서, 본 발명의 실시양태는 모체 혈장 (즉, 생물학적 샘플) 중 태아 DNA 분획 농도를 측정한다. 이러한 정보와 함께, 보다 유용한 P_r 값이 하기와 같이 계산될 수 있다.

[0083] 모체 혈장 중 태아 DNA의 낮은 분획 백분율로도, 삼염색체 21 태아는 모체 혈장으로 방출된 태아 DNA의 게놈-등가물 (GE) 당 염색체 21 서열의 추가적 양에 기여할 것이다. 예를 들어, 태아에 의해 기여된 5 GE/mL DNA와 50 GE/mL의 총 DNA (즉, 10% 태아 DNA 분획 농도)를 함유하는 정배수체 임신으로부터의 모체 혈장 샘플은 모체 혈장 (밀리리터) 당 염색체 21 서열의 총 100개의 카피 (90개의 모체 카피 + 10개의 태아 카피)를 함유해야 한다. 삼염색체 21 임신에 대해, 태아 GE는 각각 3개 카피의 염색체 21에 기여하여, 모체 혈장 중 염색체 21 서열의 총 105개의 카피/mL (90개의 모체 카피 + 15개의 태아 카피)가 되도록 한다. 따라서, 10% 태아 DNA 농도에서, 삼염색체 임신의 모체 혈장 중 염색체 21 유래 서열의 양은 정배수체 경우의 1.05배일 것이다. 따라서, 분석적 접근법이 이러한 적은 정도의 정량 차이를 결정하도록 개발될 수 있다면, 삼염색체 21 태아의 비침습성 출생 전 진단에 대한 다형성-독립적 시험이 가능할 것이다.

[0084] 따라서, 과다표출의 정도는 분석된 DNA 샘플 중 태아 DNA 분획 농도에 좌우될 것이다. 예를 들어, 태반 DNA를 분석하는 경우, 태아 게놈 중 이론적 RCD 비율은 3:2, 즉 1.5-배 차이이어야 한다. 그러나, 상기 기술된 것과 같이, 이론적 RCD 비율은 10% 태아 DNA를 함유하는 모체 혈장 샘플을 분석하는 경우 1.05로 감소할 것이다.

실험적으로 유도된 P_r은 단지 염색체 21 유전자좌에 양성인 웰의 수를 총 정보제공 웰의 수로 나누어 계산된다. 실험적으로 유도된 P_r은 계산된 P_r 및 이론적 RCD 비율과 함께 SPRT 분석에 적용된다.

[0085] 도 4는 본 발명의 실시양태에 따라 태아 핵산의 백분율을 사용하여 질환 상태를 결정하는 방법 400을 나타낸다. 단계 410에서, 태아 물질의 분획 백분율이 측정된다. 한 실시양태에서, 분획 백분율은 비-태아-특이적 마커 (즉, 모계 및 태아 모두에 존재하는 유전자 서열)와 관련하여 태아-특이적 마커 (예를 들어, Y-염색체, 유전자 다형성 마커 (예컨대, SNP), 태반 후성 기호 (placental epigenetic signatures))의 양을 측정하여 결정된다. 실시간 PCR, 디지탈 PCR, 서열분석 반응 (대규모 병렬 게놈 서열분석을 포함) 또는 임의의 다른 정량 방법에 의해 실질적 측정을 수행할 수 있다. 한 측면에서, 이 측정에 대해 잠재적으로 대립유전자 불균형이 있을 수 있는 유전자 표적을 사용하지 않는 것이 바람직하다.

[0086] 단계 420에서, 샘플을 희석시키고, 희석시킨 샘플을 웰에 넣고, 각 웰에서 반응을 측정하는 것을 비롯한 디지탈 PCR 또는 다른 측정 방법이 실시된다. 단계 430에서, PCR 결과는 상이한 참조 핵산 서열 (예컨대, 염색체 또는 대립유전자)의 마커를 식별하기 위해서 사용된다. 단계 440에서, 과다표출된 서열의 실제 비율 (P_r)이 계산된다. 단계 450에서, 질환 상태를 결정하기 위한 분별값은 샘플 중 태아 물질의 백분율을 사용하여 계산된다. 단계 460에서, 실제 P_r 및 분별값으로부터, 불균형이 존재하는지 여부가 결정된다.

[0087] 한 실시양태에서, 참조 핵산 서열의 분획 백분율은 디지탈 RNA-SNP 방법에 혼입된다. 따라서, 암 세포로 인한 LOH를 연구하는 경우, 50% 미만 암 세포를 갖는 종양 샘플로 실시할 수 있다. 이것은 또한 50% 초과 암 세포를 갖는 샘플에 대해서도 사용되어, 더욱 정확한 P_r을 수득함으로써 부정확한 진단으로 이끌 수 있는 잘못된 양성의 수를 감소시킬 수 있다. 또다른 실시양태에서, 태아 핵산 백분율은 디지탈 PCR 방법에 혼입되어, 태아가 모체 혈장 핵산 분석으로부터 부모의 유전자 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 척수 근위축증을 유발하는 돌연변이) 또는 다형성을 물려받는지 여부를 결정한다.

[0089] <u>IV. 웰 당 평균 농도의 혼입</u>

[0090] 이전 방법 (예를 들어, 상기 문헌 [Zhou, W. et al. 2002])의 또다른 단점은 웰 당 주형의 평균 농도 (m)가 웰 당 1로 요구된다는 것이다. 정확한 농도를 얻는 것이 어렵다는 것을 감안할 때, 이것은 부정확해질 수 있다. 또한, 웰 당 1 주형의 정확한 농도로도, 이전 방법은 웰 중 주형의 통계적 분포를 간과하였다. 이전 방법에서, 즉, 기존 알고리즘에서, 대립가설을 수용하기 위한 예상 P_r 값은 대립유전자 비율로 가정되어, 웰 당 주형 DNA의 평균 농도와 무관하였다.

[0091] 그러나, 희석시킨 샘플에서 주형의 자연적인 통계적 편차로 인해, 정확하게 웰 당 1 주형은 아닐 것이다. 본 발명의 실시양태는 하나 이상의 서열의 평균 농도를 측정하여, 분별값, 즉 예상 P_r을 계산하는데 사용된다. 한 측면에서, 이러한 계산은 상이한 핵산 서열을 함유하는 웰의 확률을 측정하여 예상 P_r을 결정하기 위해서 사용되는 통계적 분포를 포함한다.

[0092] 한 실시양태에서, 평균 농도는 일례로 DNA 샘플 중 더 낮은 농도의 핵산 서열인 한 참조 핵산 서열로 고려된다. 불균형이 없는 샘플의 경우, 샘플 중 두 서열의 농도는 동일할 것이고, 어느 하나는 참조 대립유전자로 간주될 수 있다. 예를 들어 LOH가 있는 샘플의 경우, 암 세포에서 결실된 대립유전자는 참조 대립유전자로 간주될 것이다. 참조 대립유전자의 평균 농도는 m_r로 표시될 것이다. 또다른 실시양태에서, 더 높은 농도의 서열이 참조 서열로 간주될 수 있다.

[0094] **A. 디지탈-SNP, SPRT 및 디지탈 PCR을 사용한 예**

[0095] 도 5는 본 발명의 한 실시양태에 따라 평균 주형 농도를 사용하여 질환 상태를 결정하는 방법 500을 보여준다. 단계 510에서, 상이한 서열의 양을 측정한다. 이것은, 예를 들어 상기 설명된 것과 같은 디지탈 PCR 실험에서 마커를 계수하여 실시할 수 있다. 그러나, 이것은 증폭 단계를 포함하지 않거나 또는 형광 마커를 사용하지 않지만 다른 성질, 예컨대 질량, 특정 광학적 성질 또는 염기-쌍 성질과 같은 물성을 사용할 수 있는 다른 방법에 의해서도 실시할 수 있다.

[0096] 단계 520에서, 과다표출된 서열의 실제 비율이 결정된다. 이것은 상기 기술된 것과 같이 단지 그 서열만 나타내는 웰의 수를 취하고, 정보제공 웰의 수로 나누어 실시될 수 있다. 단계 530에서, 하나 이상의 서열의 평균 농도가 측정된다 (참조 서열). 한 실시양태에서, 참조 서열은 과다표출된 서열이다. 또다른 실시양태에서, 참

조 서열은 과소표출된(underrepresented) 서열이다. 측정은 디지탈 PCR 실험에서 참조 서열에 대해 음성인 웰의 수를 계수하여 실시할 수 있다. 음성 웰의 비율과 평균 주형 농도 사이의 관계는 다음 소절에 기술된 것과 같은 푸아송 (Poisson) 분포에 의해 설명된다.

[0097] 단계 540에서, 상이한 서열에 대해 양성인 웰의 예상 양이, 예를 들어 푸아송 분포를 사용하여 계산된다. 예상 양은 웰 당 서열, 웰 당 평균 서열, 서열을 함유하는 웰의 수 또는 임의의 다른 적합한 양의 확률일 수 있다. 단계 550에서, 예상 P_r은 예상 양으로부터 계산된다. 단계 560에서, 분별값은, 예를 들어 SPRT를 사용하여 예상 P_r로부터 계산된다. 단계 570에서, 핵산 서열 불균형의 분류가 결정된다. 방법 500의 특정 측면이 이제 기술된다.

[0099] 1. 서열 예상 양의 결정

[0100] 웰 (반응 또는 반응 혼합물) 당 평균 농도가 단계 530으로부터 알려지면, 서열을 나타내는 웰의 예상 수가 단계 540에서 계산될 수 있다. 이 양은 %, 분수 값, 또는 정수 값으로 표현될 수 있다. 설명을 위해 특정 예를 사용하면서, 웰 당 참조 주형의 평균 농도 (m_r)는 웰 당 0.5이고, PLAC4 SNP, rs8130833에서 삼염색체 21 태아의 유전자형은 AGG인 것으로 가정한다. 따라서, 참조 주형은 A 대립유전자일 것이고, 과다표출된 주형은 G 대립유전자일 것이다.

[0101] 한 실시양태에서, 푸아송 분포는 디지탈 PCR과 같은 측정 과정의 웰의 반응 혼합물 중 A 대립유전자의 분포로 가정한다. 다른 실시양태에서는, 이항 분포와 같은 다른 분포 함수가 사용된다.

[0102] 푸아송 방정식은 $P(n) = \frac{m^n e^{-m}}{n!}$ 이고, 여기서 n은 웰 당 주형 분자의 수이고; P(n)은 특정 웰 중 n 주형 분자의 확률이고; m은 특정 디지탈 PCR 실험에서 한 웰 중 주형 분자의 평균 수이다.

[0103] 따라서, 0.5의 평균 A-대립유전자 농도에서 A 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0104] $P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} \approx 0.6065$

[0105] 따라서, 하나 이상의 A 대립유전자의 분자를 함유하는 임의의 웰의 확률은: 1 - 0.6065 = 0.3935이다. 그러므로, 약 39%의 웰이 하나 이상의 A 대립유전자의 분자를 함유할 것으로 예상될 것이다.

[0106] 비-참조 핵산 서열의 경우, 삼염색체 21 태아의 각 세포에 대해서, A 대 G의 게놈 비율은 1:2일 것이다. 추출된 RNA 또는 DNA 샘플 중 A 대 G 비율이 변화하지 않는 것으로 가정하면, 웰 당 G 대립유전자의 평균 농도는 A 대립유전자의 평균 농도의 2배, 즉 2 × 0.5 = 1일 것이다.

[0107] 따라서, 평균 G-대립유전자 농도인 1에서 G 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0108] $P(0) = \frac{1^0 e^{-1}}{0!} = e^{-1} \approx 0.3679$

[0109] 따라서, 하나 이상의 G 대립유전자 분자를 함유하는 임의의 웰의 확률은: 1 - 0.3679 = 0.6321일 것이다. 그러므로, 약 63%의 웰이 하나 이상의 G 대립유전자 분자를 함유하는 것으로 예상될 것이다.

[0110] 2. 과다표출된 서열 비율의 결정

[0111] 예상 양이 계산된 후, 과다표출된 핵산 서열의 비율이 결정될 수 있다. A 대립유전자 및 G 대립유전자로 웰을 채우는 것이 독립적이라고 가정하면, 두 대립유전자를 모두 함유하는 웰의 확률은 0.3935 × 0.6321 = 0.2487일 것이다. 그러므로, 약 25%의 웰이 두 대립유전자를 모두 함유하는 것으로 예상될 것이다.

[0112] A 대립유전자는 함유하되 G 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 하나 이상의 A 대립유전자를 함유하는 웰의 수에서 A 및 G 대립유전자를 모두 함유하는 웰의 수를 뺀 것이다: 0.3935 - 0.2487 = 0.1448. 유사하게, G 대립유전자는 함유하되 A 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 0.6321 - 0.2487 = 0.3834일 것이다. 정보제공 웰은 A 대립유전자 또는 G 대립유전자에 대해 양성이나, 둘 모두에 대해서는 양성이 아닌 웰로서 정의된다.

[0113] 따라서, 디지탈 RNA-SNP 분석에서 G 대립유전자에 상대적으로 A 대립유전자를 함유하는 웰의 예상 비율은 0.1448/0.3834이다. 즉, 단지 G 대립유전자에 대해서만 양성인 웰의 비율은 단지 A 대립유전자에 대해서만 양성인 웰의 2.65배이다. 이것은 과다표출된 대립유전자가 다른 대립유전자의 2배인 태아 게놈 비율과 대조적이다.

[0114] SPRT 분석에 대해, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_r)은 SPRT 곡선을 사용하여 계산되고 해석된다. 현 예에서, 정보제공 웰의 비율은 0.1448 + 0.3834 = 0.5282일 것이다. 따라서, m_r 0.5에서 삼염색체 21 경우의 예상 P_r은 0.3834/0.5282 = 0.73이다.

[0115] 평균 주형 농도 (m)는 푸아송 방정식의 주요 파라미터이므로, P_r은 m에 따라변할 것이다. 도 6은 본 발명의 실시양태에 따라 웰 당 평균 참조 주형 농도 (m_r)로 표현된 주형 농도의 범위에 대한 삼염색체 21 샘플의 예상 디지탈 RNA-SNP 대립유전자 비율 및 P_r을 표로 만든 표 600을 나타낸다. 표 600은 일련의 웰 당 평균 참조 주형 농도 (m_r)에 대한 예상 대립유전자 비율 및 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_r)을 나타낸다.

[0116] 예상 P_r 값은 비-선형 방식으로 웰 당 참조 대립유전자의 평균 농도 (m_r)에 따라 변한다. 표 600에 제시된 것과 같이, 대립가설을 수용하기 위한 예상 P_r 값은 m_r과 함께 증가할 것이다. 귀무가설을 수용하기 위한 예상 P_r 값이 0.5로 고정되어 있으므로, 대립유전자 불균형이 있는 샘플 및 대립유전자 불균형이 없는 샘플은 m_r이 증가할 경우, P_r 값의 측면에서 보다 광범위하게 나뉘어질 것이다. 다른 실시양태에서는 귀무가설을 수용하기 위한 값이 0.5가 아닐 수 있다는 것을 주목한다. 이것은 정상 비율이 1:1과 다른 경우, 예를 들어, 5:3인 경우 일어날 수 있으며, 따라서 비율이 5:3으로부터 벗어나는 경우 불균형이 발생할 것이다. 두 상이한 핵산 서열 양의 차이는 개별 경우에 따라 결정될 것이다.

[0117] 그러나, 이전 방법 (예를 들어, 상기 문헌 [Zhou, W. et al. 2002])은 LOH 샘플에 대해 고정된 예상 P_r 값을 사용하였으므로, LOH가 있는 샘플에 대한 P_r 값을 과소평가하였다 (대립가설 수용). 과소평가의 정도는 m_r과 함께 증가할 것이다. 즉, DNA 샘플 중 참조 대립유전자의 평균 농도가 더 높을수록, 이전 방법은 더욱 부정확해질 것이다. 대립가설을 수용하기 위한 P_r의 과소평가는 귀무가설 및 대립가설 모두를 수용하기 위한 분별값의 계산을 부정확하게 할 것이다.

[0119] 3. 예상 P_r을 기준으로 한 분별값의 계산

[0120] SPRT를 사용하는 실시양태에 대해, 문헌 [El Karoui at al. (2006)]으로부터 SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식을 사용할 수 있다. 또한, 귀무가설 또는 대립가설을 수용하기 위해서 바람직한 통계적 신뢰성의 수준은 방정식에서 임계값 우도 비율(threshold likelihood ratio)을 조정하면서 변할 수 있다. 이 연구에서, 임계값 우도 비율 8이 암 검출의 맥락에서 대립유전자 불균형이 있는 샘플과 대립유전자 불균형이 없는 샘플을 구별하는 만족할만한 성능을 제공하는 것으로 나타났기 때문에, 상기 값 8이 사용된다. 따라서, 한 실시양태에서, SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식은 다음과 같다:

[0121] 상한 = $[(\ln 8)/N - \ln \delta]/\ln \gamma$

[0122] 하한 = $[(\ln 1/8)/N - \ln \delta]/\ln \gamma$

[0123] 여기서, $\delta = (1 - \Theta_1)/(1 - \Theta_0)$,

[0124] $\gamma = -(\Theta_1(1 - \Theta_0)/\Theta_0(1 - \Theta_1)$,

[0125] Θ_0 = 귀무가설이 사실인 경우 비-참조 대립유전자를 함유하는 정보제공 웰의 비율 = 0.5 (하기 참조),

[0126] Θ_1 = 대립가설이 사실인 경우 비-참조 (즉, 과다표출된) 대립유전자를 함유하는 정보제공 웰의 비율,

[0127] N = 정보제공 웰의 수 = 어느 하나의 대립유전자에만 양성인 웰의 수.

[0128] (ln은 자연 로그, 즉 $\log_e$를 나타내는 수학 기호임)

[0129] 귀무가설을 수용하기 위한 Θ_0의 결정을 위해서, 샘플은 정배수체 태아를 갖는 임신 여성으로부터 얻어지는 것으로 가정된다. 이러한 가정 하에, 어느 하나의 주형에 양성인 웰의 예상 수는 1:1일 것이므로, 비-참조 대립유전자를 함유하는 정보제공 웰의 예상 비율은 0.5일 것이다.

[0130] 대립가설을 수용하기 위한 Θ_1의 결정을 위해서, 샘플은 삼염색체 21 태아를 갖는 임신 여성으로부터 얻어지는 것으로 가정된다. 디지탈 RNA-SNP 분석을 위한 삼염색체 21 경우의 예상 P_r의 계산은 표 600에 상술되어 있다. 따라서, 디지탈 RNA-SNP 분석을 위한 Θ_1은 표 600의 마지막 열에 제시된 데이타를 참조한다.

[0132] 4. 평균 농도의 측정

[0133] m_r의 측정은 당업자에게 공지되거나 또는 공지될 다양한 메커니즘에 의해 실시될 것이다. 한 실시양태에서, m_r 값은 디지탈 PCR 분석의 실험 과정 중 결정된다. m_r 값과 참조 대립유전자에 대해 양성인 웰의 총 수 사이의 관계는 분포 (예를 들어, 푸아송 분포)에 의해 좌우될 수 있으므로, m_r은 다음 식을 사용하여 참조 대립유전자에 대해 양성인 웰의 수로부터 계산될 수 있다:

[0134] m_r = - ln (1 - 참조 대립유전자에 대해 양성인 웰의 비율).

[0135] ln은 자연 로그, 즉 $\log_e$라는 것에 주목한다. 이 접근법은 디지탈 PCR 실험을 위해 사용되는 DNA 샘플 중 m_r의 직접적이고 정확한 평가를 제공한다.

[0136] 이 방법은 바람직한 농도를 얻기 위해서 사용될 수 있다. 예를 들어, 샘플의 추출된 핵산은 방법 200의 단계 240에서 실시한 것과 같이 반응 웰 당 하나의 주형 분자와 같은 특정 농도로 희석될 수 있다. 푸아송 분포를 사용하는 한 실시양태에서, 주형이 없는 웰의 예상 비율은 e^{-m}으로 계산될 수 있고, 여기서 m은 웰 당 주형 분자의 평균 농도이다. 예를 들어, 웰 당 하나의 주형 분자의 평균 농도에서, 주형 분자가 없는 웰의 예상 비율은 e^{-1}, 즉 0.37 (37%)로 주어진다. 나머지 63%의 웰은 하나 이상의 주형 분자를 함유할 것이다. 전형적으로, 디지탈 PCR 실행에서 양성 웰 및 정보제공 웰의 수가 계수될 것이다. 정보제공 웰의 정의 및 디지탈 PCR 데이타가 해석되는 방식은 용도에 좌우된다.

[0137] 다른 실시양태에서, 웰 당 평균 농도 m_r은 또다른 정량화 방법, 예를 들어 질량 분광법을 사용하는 정량 실시간 PCR, 반-정량 경쟁적 PCR, 실-경쟁적 PCR 등에 의해 측정된다.

[0139] **B. 디지탈 RCD**

[0140] 평균 농도를 사용하는 디지탈 RCD를 상기 기재된 디지탈 SNP 방법과 유사한 방식으로 수행할 수 있다. 참조 염색체 (비-염색체 21) 마커, 염색체 21 마커, 및 두 마커 모두에 대해 양성인 웰의 수를 디지탈 PCR에 의해 측정할 수 있다. 디지탈 SNP 분석에 대한 웰 당 참조 마커의 평균 농도 (m_r)의 계산에서와 같은 푸아송 확률 함수에 따라, 염색체 21 마커의 양성에 상관없이, 웰 당 참조 마커의 평균 농도 (m_r)를 참조 마커에 대해 음성인 웰의 총 수로부터 계산할 수 있다.

[0141] 이후, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성으로부터 얻어진 것과 같은 혈장 샘플을 분류하기 위해 SPRT 분석을 사용할 수 있다. 태아가 정배수체인 경우 귀무가설이 수용될 것이다. 이 시나리오에서, 참조 마커 및 염색체 21 마커에 대해 양성인 웰의 예상 비율은 1:1일 것이며, 이에 따라, 염색체 21 마커에 대해 양성 신호를 갖는 정보제공 웰의 예상 비율은 0.5일 것이다. 태아가 염색체 21에 대해 삼염색체인 경우 대립가설이 수용될 것이다. 이 시나리오에서, 샘플 DNA가 전적으로 태아로부터 유래된 경우, 각 웰에서의 염색체 21 마커의 평균 농도는 참조 마커의 평균 농도 (m_r)의 3/2배일 것이다.

[0142] 디지탈 RCD를 사용하여 태아-특이적 마커, 예를 들어 태반 후성 기호의 검출을 통해 염색체 양을 측정할 수 있지만 (문헌 [Chim, SSC. et al. 2005 Proc Natl Acad Sci USA 102, 14753-14758]), 디지탈 RCD 분석의 실시양태는 비-태아-특이적 마커를 사용한다. 따라서, 비-태아 특이적 마커를 사용하는 경우 태아 물질의 백분율을 측정하는 추가의 단계가 나타날 것이다. 따라서, 웰 당 염색체 21 마커의 평균 농도는 샘플 중 태아 DNA의 비율에 의존적일 것이며, m_r [(200% + 태아 DNA 백분율)/200%]을 사용하여 계산할 수 있다.

[0143] 예시용 특정 실시예를 다시 사용하여, 웰 당 참조 주형인 염색체 1의 평균 농도 (m_r)는 0.5인 것으로 가정하며,

50%의 DNA는 태아로부터 유래된 것으로 가정하고, 샘플 중 50%의 DNA는 모체로부터 유래된다.

[0144] 따라서, 푸아송 분포를 사용하여, 평균 농도가 웰 당 0.5인 경우 염색체 1 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0) = \frac{0.5^{0} e^{-0.5}}{0!} = e^{-0.5} \approx 0.6065$$

[0145]

[0146] 따라서, 염색체 1 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.6065 = 0.3935일 것이다. 따라서, 대략 39%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다.

[0147] 이러한 삼염색체 21 태아의 각각의 세포에 대하여, 염색체 21 대 염색체 1의 게놈 비율은 3:2일 것이다. DNA 샘플에서 염색체 21과 염색체 1 사이의 비율은 태아 DNA 분획 농도 (태아 DNA %)에 의존적일 것이며, 3×태아 DNA % + 2(1 - 태아 DNA %) : 2×태아 DNA % + 2×(1 - 태아 DNA %)일 것이다. 따라서, 이 경우에서, 태아 DNA 분획 농도가 50%인 경우, 비는 (3×50% + 2×50%) / (2×50% + 2×50%) = 1.25일 것이다. 디지탈 SNP 방법이 태아 특이적 마커를 사용하지 않은 경우, 이러한 계산을 사용하여 비-참조 서열의 평균 농도를 계산할 수도 있다.

[0148] 따라서, 웰 당 염색체 1 유전자좌의 평균 농도가 0.5인 경우, 웰 당 염색체 21 유전자좌의 평균 농도는 1.25×0.5 = 0.625이다. 따라서, 평균 농도가 웰 당 0.625인 경우 염색체 21 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0) = \frac{0.625^{0} e^{-0.625}}{0!} = e^{-0.625} \approx 0.5353$$

[0149]

[0150] 따라서, 염색체 21 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.5353 = 0.4647일 것이다. 따라서, 대략 46%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다. 유전자좌 중 하나로의 웰의 충전이 독립적이라고 가정하면, 유전자좌 모두를 함유하는 웰의 확률은 0.3935×0.4647 = 0.1829일 것이다. 따라서, 대략 18%의 웰이 두 유전자좌 모두를 함유하는 것으로 예상될 것이다.

[0151] 염색체 1 유전자좌를 함유하되 염색체 21 유전자좌는 함유하지 않는 것으로 예상되는 웰의 비율은, 하나 이상의 염색체 1 유전자좌를 함유하는 웰의 수에서 유전자좌 모두를 함유하는 웰의 수를 차감한 것, 즉 0.3935 - 0.1829 = 0.2106일 것이다. 유사하게, 염색체 21 유전자좌를 함유하되 유전자좌 모두는 함유하지 않는 것으로 예상되는 웰의 비율은 0.4647 - 0.1829 = 0.2818일 것이다. 정보제공 웰은 염색체 1 유전자좌 또는 염색체 21 유전자좌 중 하나에 대해 양성이되 둘 모두에 대해서는 아닌 웰로서 정의된다.

[0152] 따라서, 디지탈 RCD 분석에서의 염색체 21 대 염색체 1 예상 비율은 0.2818/0.2106 = 1.34이다. 바꾸어 말하면, 염색체 21 유전자좌에 대해서만 양성인 웰의 비율은 염색체 1 유전자좌에 대해서만 양성인 웰의 비율의 1.34배이다. 이는 DNA 샘플에서의 1.25의 비율과는 대조적이다.

[0153] SPRT 분석에 대하여, 염색체 21 유전자좌에 대해 양성인 정보제공 웰의 비율 (P_r)을 SPRT 곡선을 사용하여 계산 및 해석하는 것이 필요할 것이다. 현재 예에서, 정보제공 웰의 비율은 0.2106 + 0.2818 = 0.4924일 것이다. 따라서, 0.5의 m_r에서 50% 태아 DNA를 갖는 삼염색체 21 경우의 예상 P_r은 0.2818/0.4924 = 0.57이다.

[0154] 평균 주형 농도 (m)가 푸아송 방정식에서의 중요 파라미터이기 때문에, P_r은 m에 의해 변할 것이다. 도 7은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도의 범위에서 삼염색체 21 샘플 중 10%, 25%, 50% 및 100%의 태아 DNA 분획 농도에 대한 예상 P_r을 표로 만든 표 700을 나타낸다. 디지탈 RCD 분석에 대한 삼염색체 21 경우의 예상 P_r에 대한 계산을 표 700에서 상술한다. 따라서, 표 700에서 상응하는 예상 P_r 값을 나타내는 열로부터 여러 태아 DNA 분획 농도를 갖는 샘플의 디지탈 RCD 분석에 대한 Θ_1을 얻을 수 있다.

[0156] **C. 결과**

[0157] 1. 상이한 m_r의 비교

[0158] 대립유전자 또는 염색체 불균형의 이론값 (태아 게놈에서와 같음)과 실험적으로 예상되는 정도 사이의 차이, 및 m_r 값 범위에 대한 후자의 측정을 위한 계산값에 대한 근거를 표 600 및 700에 나타낸다. 삼염색체 21 샘플의 디지탈 RNA-SNP 분석에서, m_r = 0.5인 경우, 참조 대립유전자만을 함유하는 웰에 대한 과다표출된 대립유전자만을 함유하는 웰, 즉 디지탈 RNA-SNP 비율은 2.65이다 (표 600). 100% 태아 DNA로 구성된 표본의 디지탈 RCD 분석에서, m_r = 0.5인 경우, 염색체 1 유전자좌에 대해 단독으로 양성인 웰에 대한 염색체 21 유전자좌에 대해 단독으로 양성인 웰, 즉 디지탈 RCD 비율은 1.7이다 (표 700) (P_r = 0.63, 따라서 디지탈 RCD 비율은 0.63/(1 - 0.63) = 1.7임). 태아 DNA 분획 농도가 감소하기 때문에, 디지탈 RCD 비율은 동일한 m_r에 대해 감소한다 (표 700).

[0159] 표 600 및 700에서 나타낸 것과 같이, 대립유전자 또는 염색체 과다표출 정도는 m_r에 따라 증가한다. 그러나, 정보제공 웰의 백분율은 m_r = 0.5 근처에서 그의 최대에 도달하며, m_r이 추가로 증가함에 따라 점차 감소한다. 실제로, 표본에 대한 주형 분자의 양이 제한되지 않는 경우 정보제공 웰의 비율의 감소는 분석된 웰의 총 수의 증가에 의해 상쇄될 수 있지만, 추가의 웰은 시약 비용의 증가를 요구할 것이다. 따라서, 최적 디지탈 PCR 수행은 주형 농도와 샘플 당 시험된 웰의 총 수 간에 상호 의존적이다.

[0161] 2. SPRT 곡선을 사용하는 예

[0162] 상기 논의된 것과 같이, 디지탈 PCR 실험에 대한 예상 대립유전자 또는 염색체 불균형 정도는 반응 혼합물 (예를 들어, 웰) 당 실제 주형 농도에 의존적이다. 본 발명자들은 참조 대립유전자를 기준으로 한 주형 농도, 즉 웰 당 평균 참조 주형 농도 (m_r)를 기재한다. 상기 방정식에서 나타낸 것과 같이, 예상 P_r을 사용하여 상부 및 하부 SPRT 곡선의 플롯팅을 결정할 수 있다. 예상 P_r이 또한 m_r 값에 의존적이기 때문에, SPRT 곡선의 플롯팅은 본질적으로 m_r 값에 의존적일 것이다. 따라서, 실제로, 디지탈 PCR 데이타 세트의 실제 m_r에 관련된 한 세트의 SPRT 곡선이 특정 실행으로부터 P_r의 해석을 위해 사용될 필요가 있다.

[0163] 도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에 있어서 0.1, 0.5 및 1.0의 m_r 값에 대한 SPRT 곡선에서의 차이의 정도를 예시하는 플롯 800을 나타낸다. 각 세트의 디지탈 PCR 데이타는 특정 실행의 정확한 m_r 값에 관련된 특정 곡선으로 해석되어야 한다. 디지탈 RNA-SNP 및 RCD 방법에 대한 예상 대립유전자 또는 염색체 불균형 정도가 상이하기 때문에 (전자: 2:1, 후자: 3:2), 2개 디지탈 PCR 시스템에 대하여 상이한 세트의 SPRT 곡선이 필요하다는 것을 주목한다. 실험적으로 유도된 P_r을 디지탈 PCR 실행의 상응하는 m_r에 의해 선택된 관련 SPRT 곡선을 이용하여 해석한다. 이는 고정된 세트의 곡선이 사용된 경우 디지탈 PCR에 의한 LOH의 분자 검출을 위해 앞서 보고된 SPRT를 사용한 것과는 대조적이다.

[0164] SPRT를 사용하여 디지탈 PCR 데이타를 해석하기 위한 실제적 방법을 가상 디지탈 RNA-SNP 실행을 사용하여 하기 예시한다. 각 경우의 디지탈 RNA-SNP 분석 이후, A 대립유전자에 대해서만 양성인 웰, G 대립유전자에 대해서만 양성인 웰, 또는 두 대립유전자 모두에 대해서 양성인 웰의 수를 계수한다. 참조 대립유전자는 보다 적은 수의 양성 웰을 갖는 대립유전자로서 정의된다. 푸아송 확률 밀도 함수에 따라, 다른 대립유전자가 양성인 지에 관계없이, 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여 m_r 값을 계산한다. 본 발명자의 가설 예의 데이타는 다음과 같다:

[0165] 96-웰 반응에서, 20개의 웰은 A 대립유전자에 대해서만 양성이고, 24개의 웰은 G 대립유전자에 대해서만 양성이며, 33개의 웰은 두 대립유전자 모두에 대해 양성이다. A 대립유전자가 참조 대립유전자로 간주되는데, A-양성 웰이 G-양성 웰에 비해 적기 때문이다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 20 - 33 = 43이다. 따라서, 푸아송 방정식을 사용하여 다음과 같이 m_r을 계산할 수 있다: $-\ln(43/96) = 0.80$. 이 경우의 실험적으로 결정된 P_r은 24/(20 + 24) = 0.55이다.

[0166] 표 600에 따르면, m_r = 0.8에서의 삼염색체 21 샘플의 예상 P_r은 0.76이다. 따라서, 이 경우 Θ_1은 0.76이다. Θ_1 = 0.76을 기준으로 한 SPRT 곡선을 사용하여 실험적으로 유도된 P_r (이 경우 0.55)을 해석할 것이다. P_r = 0.55를 관련된 SPRT 곡선 상에 피팅하는 경우, 데이타 지점은 하부 곡선 아래에 있다. 따라서, 본 경우는 정배수체로 분류되며, 도 3을 참조한다.

[0168] 3. 기존 방법과의 비교

[0169] 도 9a는 96-웰 디지탈 RNA-SNP 분석에서의 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교 표 900을 나타낸다. 도 9b는 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교의 표 950이다. 신규 알고리즘은 디지탈 PCR 데이타로부터 유도된 m_r에 특이적인 SPRT 곡선의 선택을 의미한다. 기존 알고리즘은 모든 디지탈 PCR 실행에 대한 고정된 세트의 SPRT 곡선의 사용을 의미한다. 분류의 정확도에 대한 분별값의 부정확한 계산의 효과는 표 900에 나타낸 시뮬레이션 분석에 의해 밝혀진다.

[0170] 기존 연구에서 고정된 세트의 SPRT 곡선의 사용과 비교하여, 분류불가능한 데이타의 비율은 표 900 및 950에서 나타난 것과 같이 본 발명의 방법에서 훨씬 적다. 예를 들어, 본 발명의 방법을 사용하여, m_r = 0.5에서, 14% 및 0%의 삼염색체 21 샘플이 각각 96-웰 및 384-웰 디지탈 RNA-SNP 분석에 대하여 분류불가능할 것이나, 고정된 곡선의 사용에 대해서는 각각 62% 및 10%가 분류불가능할 것이다 (표 900). 따라서, 본 발명의 방법은 보다 적은 수의 정보제공 웰로 질환 분류를 가능하게 한다.

[0171] 표 900에서 나타낸 것과 같이, 신규 알고리즘은 0.1 내지 2.0의 모든 m_r 값에 대한 대립유전자 비율이 벗어나거나 벗어나지 않은 샘플의 분류에서 보다 정확하다. 예를 들어, m_r이 1.0과 같고, 96-웰 디지탈 RNA-SNP 실행을 수행하는 경우, 신규 알고리즘은 각각 88% 및 92%의, 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플을 정확하게 분류하며, 반면 기존 알고리즘을 사용하여 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플의 정확한 분류의 백분율은 단지 각각 19% 및 36%이다.

[0172] 신규 알고리즘을 사용하여, 대립유전자 비율이 벗어난 샘플과 벗어나지 않은 샘플의 분리는 m_r에 따라 증가할 것이다. 결과적으로, 분류 정확도는 m_r에 따라 증가할 것이다. m_r이 2.0 초과로 증가하는 경우 분류 정확도에 대한 2개 샘플 그룹의 분리에서의 증가의 효과는 감소될 것인데, 이는 정보제공 웰의 백분율에서의 감소 때문이다. 반면, 기존 알고리즘의 사용에서, m_r이 증가하는 경우 분류 정확도는 유의하게 감소하는데, 이는 예상 P_r 값과 그의 참값의 편차 증가 때문이다.

[0173] 본 발명자의 실험 및 시뮬레이션 데이타는 디지탈 RNA-SNP가 삼염색체 21 검출에 대해 효과적이며 정확한 방법임을 나타낸다. 모체 혈장에서의 *PLAC4* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 13개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지탈 PCR 실험만이 정확한 분류에 필요하였다 (도 13b의 표 1350). 이에 따라, 상기 균질한 실시간 디지탈 PCR-기반 방법은 RNA-SNP 분석에 대한 질량 분광법-기반 방법의 또다른 형태를 제공한다 (상기 문헌 [Lo, YMD, et al. 2007 Nat Med]). 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 종을 태아 염색체 이수성의 디지탈 PCR-기반 검출에 사용할 수 있다는 것을 구상한다. 한 예는 태아 후성 마커이며 (문헌 [Chim, SSC et al. (2005) Proc Natl Acad Sci USA 102, 14753-14758]; [Chan, KCA et al. (2006) Clin Chem 52, 2211-2218]), 이는 최근에 후성 대립유전자 비율 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용되었다 (문헌 [Tong, YK et al. (2006) Clin Chem 52, 2194-2202]). 따라서, 본 발명자들은 디지탈 EAR이 가능한 분석 기법일 것으로 예상한다.

[0175] V. 증가 %, 다중 마커 및 PCR 변형법

[0176] 상기 기재된 것과 같이, 태아 DNA만이 임신 11주 내지 17주 사이에 대략 3%의 평균 분획 농도로 모체 혈장 DNA의 소수 분획을 구성하는 경우, 모체 혈장으로부터 추출된 DNA에 본 발명의 실시양태를 적용하는 것이 복잡해질 수 있다. 그러나, 본원에서 나타낸 것과 같이, 디지탈 RCD는 심지어 이수체 DNA가 소수 집단으로서 존재하는 경우에도 이수성 검출을 가능하게 한다. 임신 초기 동안 존재할 수 있는, 태아 DNA 분획 농도의 감소에 대하여, 보다 많은 수의 정보제공 계수가 디지탈 RCD에 대해 요구된다. 도 12의 표 1200에서 요약한 것과 같이, 본 작업의 의미는 본 발명자들이 한 세트의 기준 파라미터, 예를 들어 요구되는 태아 DNA 분획 및 총 주형 분자를 제공하였다는 것이며, 이들에 대해 진단 검정이 수립될 수 있다. 본 발명자의 견해에서, 25%의 태아 DNA 분획 농도에 대한 7680개의 총 수의 반응은 특히 매력적인 세트의 기준 파라미터이다. 이러한 파라미터는 표 1200에서 나타낸 것과 같이 정배수체 및 삼염색체 21 샘플이 정확하게 97%로 분류가능하게 한다.

[0177] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한된다 (문헌 [Lo, YMD. et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에서, 태아 및 모체 모두로부터의 기여에 대하여, 상염색체 유전

자좌, β-글로빈 유전자의 중앙 모체 혈장 농도는 986 카피/mL인 것으로 나타났다 (문헌 [Lo, YMD. et al. 1998 Am J Hum Genet 62, 768-7758]). 7,680개의 분자를 포획하기 위해, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 얻을 수 있는 상기 혈장 부피는 일반적인 실시의 한도이다. 그러나, 본 발명자들은 다중 세트의 chr21 및 참조 염색체 표적을 디지탈 RCD 분석을 위해 합할 수 있다는 것을 구상한다. 5쌍의 chr21 및 참조 염색체 표적에 대하여, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 멀티플렉스 단일 분자 PCR을 수행할 수 있다. 이러한 멀티플렉스 단일 분자 분석의 견고성(robustness)은 단일 분자 일배체형화(haplotyping)에 대해 이미 입증되었다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci USA 100, 7449-7453]).

[0178] 별법으로, 25%의 태아 DNA 분획 농도를 달성하기 위해, 방법은 모체 혈장에서 태아 DNA의 선택적 풍부화 (문헌 [Li, Y. et al. 2004 Clin Chem 50, 1002-1011]) 또는 모체 DNA 배경의 억제 (문헌 [Dhallan, R et al. 2004 JAMA 291, 1114-1119]) 또는 둘 모두를 가능하게 할 수 있다. 태아 DNA 풍부화 및 모체 DNA 억제를 위한 물리적 방법과는 별개로, 특정 DNA 메틸화 패턴을 나타내는 태아 DNA 분자의 표적화와 같은 분자 풍부화 전략을 사용하는 것이 또한 가능할 것이다 (문헌 [Chim, SSC et al. 2005 Proc Natl Acad Sci USA 102,14753-14758], [Chan, KCA et al. 2006 Clin Chem 52, 2211-2218]; [Chiu, RWK et al. 2007 Am J Pathol 170, 941-950.])

[0179] 추가적으로, 디지탈 PCR 수행에 대한 현행 연구에 사용되는 것과 같은 디지탈 실시간 PCR 분석의 수동 설정을 위한 다수의 또다른 방법이 현재 존재한다. 이러한 다른 방법에는 미세유체 디지탈 PCR 칩 (문헌 [Warren, L et al. 2006 Proc Natl Acad Sci USA 103, 17807-17812]; [Ottesen, EA et al. 2006 Science 314, 1464-1467]), 에멀젼 PCR (문헌 [Dressman, D et al. 2003 Proc Natl Acad Sci USA 100, 8817-8822]), 및 예를 들어 로슈(Roche) 454 플랫폼, 일루미나 솔렉사(Illumina Solexa) 플랫폼, 및 어플라이드 바이오시스템즈(Applied Biosystems)의 솔리드(SOLiD)™ 시스템을 사용하는 대규모 병렬 게놈 서열분석 (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]) 등이 있다. 후자에 대하여, 본 발명자의 방법은 또한 단일 DNA 분자에 대한 대규모 병렬 서열분석 방법에 적용가능하며, 이는 증폭 단계, 예컨대 헬리코스 트루(Helicos True) 단일 분자 DNA 서열분석 기술 (문헌 [Harris TD et al. 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시즈(Pacific Biosciences)의 단일 분자, 실시간 (SMRT™) 기술, 및 나노공극(nanopore) 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53, 1996-2001])을 요구하지 않는다. 이들 방법을 사용하여, 디지탈 RNA-SNP 및 디지탈 RCD를 다수의 샘플에 대하여 빠르게 수행함으로써, 비-침습적 출생 전 진단에 대해 본원에서 제안된 방법의 임상적 실행가능성을 향상시킬 수 있다.

도면의 간단한 설명

[0181] 도 1은 디지탈 PCR 실험을 설명하는 흐름도이다.

도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 및 RCD 방법을 예시한다.

도 2b는 암에서 자주 검출되는 염색체 이상의 예에 대한 표를 도시한다.

도 3은 본 발명의 실시양태에 따른 다운 증후군을 결정하는데 사용된 SPRT 곡선을 갖는 그래프를 예시한다.

도 4는 본 발명의 실시양태에 따른 태아 세포의 백분율을 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 5는 본 발명의 실시양태에 따른 평균 농도를 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 6은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도 범위에 대해 삼염색체 21 샘플의 예상되는 디지탈 RNA-SNP 대립유전자 비율 및 P_r을 표로 만든 표를 도시한다.

도 7은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_r)로서 표현된 주형 농도 범위에서 삼염색체 21 샘플 중 분획 태아 DNA 농도 10%, 25%, 50% 및 100%에 대해 예상되는 P_r을 표로 만든 표를 도시한다.

도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에서 m_r 값 0.1, 0.5 및 1.0에 대해 SPRT 곡선의 차이 정도를 예시하는 플롯을 도시한다.

도 9a는 본 발명의 실시양태에 따른 96-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기 위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 9b는 본 발명의 실시양태에 따른 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기

위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 10은 본 발명의 실시양태에 따른 소정의 정보제공 카운트에 대해 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아 및 분류가능하지 않은 태아의 백분율을 도시하는 표이다.

도 11은 본 발명의 실시양태에 따른 순수 (100%) 태아 DNA 샘플에 대한 디지탈 RCD 분석용 컴퓨터 시뮬레이션을 도시하는 표 1100이다.

도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플을 분류하기 위해 m_r=0.5에서 디지탈 RCD 분석의 정확성을 컴퓨터 시뮬레이션한 결과를 도시하는 표 1200이다.

도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지탈 RNA-SNP 분석에 대한 표 1300을 도시한다.

도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지탈 RNA-SNP 분석에 대한 표 1350을 도시한다.

도 14a 내지 14c는 본 발명의 실시양태에 따른 RCD 분석으로부터 생성된 분별값 곡선을 예시하는 플롯을 도시한다.

도 15a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 15b는 본 발명의 실시양태에 따른 한 모체 혈장 샘플로부터의 12개 반응 패널의 디지탈 RNA-SNP 데이타에 대한 표를 도시한다.

도 15c는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 16a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반의 디지탈 RNA-SNP 분석에 대한 표를 도시한다.

도 16b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반에 대한 디지탈 RNA-SNP 데이타의 SPRT 해석을 도시한다.

도 17은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 50% 태반/모체 혈액 세포 DNA 혼합물의 디지탈 RCD 분석에 대한 표를 도시한다.

도 18은 본 발명의 실시양태에 따른 정확한 분류에 대한 판정 경계값을 예시하는 SPRT 곡선을 도시한다.

도 19는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 양수 샘플의 디지탈 RCD 분석에 대한 표를 도시한다.

도 20은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 임신 (E = 정배수체; T18 = 삼염색체 18)으로부터의 태반 DNA 샘플의 디지탈 RCD 분석에 대한 표를 도시한다.

도 21은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 (E = 정배수체; T21 = 삼염색체 21; U = 미분류)의 50% 태반/모체 혈액 세포 DNA 혼합물의 멀티플렉스 디지탈 RCD 분석에 대한 표를 도시한다.

도 22a 및 22b는 본 발명의 실시양태에 따른 50% 정배수체 또는 삼염색체 21 태반 게놈 DNA/50% 모체 연막(buffy coat) DNA 혼합물의 멀티플렉스 디지탈 RCD 분석에 대한 표를 도시한다. 미분류(Unclass)는 분류가능하지 않음을 나타내고, T21은 삼염색체 21을 나타낸다.

도 23은 남성 및 여성 파트너 둘 다가 동일한 돌연변이를 전달하는 시나리오를 도시한다.

도 24a는 본 발명의 실시양태에 따른 여성/남성 및 남성/남성 DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 24b는 본 발명의 실시양태에 따른 25% 여성 및 75% 남성 DNA를 갖는 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 25는 본 발명의 실시양태에 따른 HbE 돌연변이에 대해 모체 혈장 샘플을 모방하는 15% 내지 50% DNA 혼합

물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 26a는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 5% 내지 50% DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 26b는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 20% DNA 혼합물의 디지탈 RMD 분석에 대한 표를 도시한다.

도 27은 본 발명의 실시양태에 따른 시스템 및 방법으로 이용가능한 예시적 컴퓨터 장치의 블록 다이아그램을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0182] **실시예**

[0183] 하기 실시예는 청구된 발명을 예시하되 제한하지 않으려고 한다.

[0185] I. 컴퓨터 시뮬레이션

[0186] SPRT 방법을 사용하여 삼염색체 21 진단의 정확도를 평가하기 위해 컴퓨터 시뮬레이션을 수행하였다. 마이크로소프트 엑셀 2003 소프트웨어 (마이크로소프트 코포레이션(Microsoft Corp.; 미국)) 및 윈도우용 SAS 9.1 소프트웨어 (에스에이에스 인스티튜트 인크.(SAS Institute Inc., 미국 노스 캐롤라이나주))를 사용하여 컴퓨터 시뮬레이션을 수행하였다. 디지탈 PCR의 수행은 참조 주형 농도 (m_r), 정보제공 계수의 수, 및 대립유전자 또는 염색체 불균형의 추정 정도 (P_r) 사이의 상호작용이다. 이들 변수 각각의 범위에 대하여 별도의 시뮬레이션을 수행하였다. 디지탈 RNA-SNP 및 디지탈 RCD에 대한 SPRT 곡선의 결정 경계가 상이하기 때문에, 2개 시스템에 대한 시뮬레이션 분석을 별도로 수행하였다.

[0187] 시뮬레이션된 각각의 디지탈 PCR 조건 (즉, m_r, 태아 DNA 분획 농도, 총 웰 수)에 대하여, 2회의 시뮬레이션을 수행하였다. 제1회는 시험된 샘플을 정배수체 태아를 갖는 임신 여성으로부터 얻은 시나리오를 시뮬레이션하였다. 제2회는 시험된 샘플을 삼염색체 21 태아를 갖는 임신 여성으로부터 얻은 경우의 시나리오를 시뮬레이션하였다. 각각의 시뮬레이션에서, 5000명의 태아의 시험을 시뮬레이션하였다.

[0189] A. RNA-SNP

[0190] 디지탈 RNA-SNP에 대하여, m_r = 0.1 내지 m_r = 2.0으로 384-웰 실험의 시뮬레이션을 수행하였다. 각각의 m_r 값에서, 본 발명자들은 5000명의 정배수체 및 5000명의 삼염색체 21 태아를 시험하는 시나리오를 시뮬레이션하였다. 주어진 m_r에 적절한 SPRT 곡선을 사용하여 10,000명의 태아를 분류하였다. 도 10은 본 발명의 실시양태에 따른, 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아, 및 주어진 정보제공 계수에 대해 분류가능하지 않은 태아의 백분율을 나타내는 표 1000이다. 정배수체 및 이수체 경우의 진단에 대한 정확도는 0.5 내지 2.0 사이의 m_r에 대해 모두 100%이다. m_r = 0.1인 경우, 57% 및 88%의 정배수체 및 삼염색체 21 태아만이 384-웰의 분석 이후 정확하게 분류될 수 있었다.

[0191] 하기 단계에 기재된 것과 같이 시뮬레이션 데이타를 생성하였다.

[0192] 단계 1에서, 각 웰에 대하여, SAS 프로그램의 무작위 (푸아송) 함수 (www.sas.com/technologies/analytics/statistics/index.html)를 사용하여 2개 난수를 생성하여 각각 A 및 G 대립유전자를 나타내었다. 무작위 (푸아송) 함수는 0으로부터 출발하는 양의 정수 (즉, 0, 1, 2, 3,....)를 생성할 것이며, 생성되는 각 정수의 확률은 웰 당 대립유전자의 평균 농도를 나타내는 주어진 평균값에 대한 푸아송 확률 밀도 함수에 따른 상기 수의 확률에 의존하였다. A 대립유전자를 나타내는 난수가 0보다 큰 경우, 즉 A 대립유전자의 1개 이상의 분자를 함유한 경우, 웰은 A 대립유전자에 대해 양성으로서 간주되었다. 유사하게, G 대립유전자를 나타내는 난수가 0보다 큰 경우, 웰은 G 대립유전자에 대해 양성으로서 간주되었다.

[0193] 정배수체 태아를 갖는 임신 여성의 시나리오를 시뮬레이션하기 위해, A 대립유전자 및 G 대립유전자에 대한 난수를 생성하는데 동일한 평균 값을 사용하였다. 예를 들어, m_r = 0.5에서의 디지탈 RNA-SNP 분석을 시뮬레이션하는 분석에서, A 대립유전자 또는 G 대립유전자 중 하나에 대한 평균 값을 웰 당 0.5 분자의 대립유전자 중 하나에 대한 평균 농도를 의미하는 0.5와 동일하게 설정하였다. 푸아송 방정식을 사용하여, 0.5의 평균

농도에서, A 또는 G 대립유전자에 대해 양성인 웰의 비율은 동일할 것이며, 0.3935였다 (표 600 참조).

[0194] m_r = 0.5에서 삼염색체 21 태아를 갖는 임신 여성의 디지탈 RNA-SNP 분석을 시뮬레이션한 경우, 웰 당 과다표출된 대립유전자의 평균 농도는 참조 대립유전자의 평균 농도의 2배, 즉 1이 되는 것으로 예상될 것이다. 이 상황에서, 과다표출된 대립유전자에 대해 양성인 웰의 확률은 0.6321이었다 (표 600 참조).

[0195] 디지탈 PCR 웰에 대한 난수의 생성 이후, 웰을 하기 상태 중 하나로서 분류할 수 있었다:

[0196] a. A 및 G 대립유전자 모두에 대해 음성

[0197] b. A 및 G 대립유전자 모두에 대해 양성

[0198] c. A 대립유전자에 대해 양성이되, G 대립유전자에 대해 음성

[0199] d. G 대립유전자에 대해 양성이되, A 대립유전자에 대해 음성.

[0200] 단계 2에서, 원하는 수의 웰 (현재 시뮬레이션에 대하여 384-웰)을 생성할 때까지 단계 1을 반복하였다. A 대립유전자에 대해서만 양성인 웰의 수 및 G 대립유전자에 대해서만 양성인 웰의 수를 계수하였다. 덜 양성인 웰을 갖는 대립유전자는 참조 대립유전자로서 간주되며, 보다 양성인 웰을 갖는 대립유전자는 잠재적으로 과다표출된 대립유전자로서 간주되었다. 정보제공 웰의 수는 대립유전자 중 하나에 대해 양성이되, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 잠재적으로 과다표출된 대립유전자를 함유하는 정보제공 웰의 비율 (P_r)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선에 대한 상한 및 하한을 본 발명의 실시양태에 따라 계산하였다.

[0201] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 각각의 2개 시나리오에 대해 5000회 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1000에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주된다. 두 군 모두에 대하여, 미리-명시된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0202] 단계 4에서, 단계 1 내지 3을 0.1 내지 2.0 범위의 m_r에 대해 0.1 증가량으로 수행하였다.

[0204] B. RCD

[0205] 도 11은 본 발명의 실시양태에 따라 0.1 내지 2.0 범위의 m_r에서 순수한 (100%) 태아 DNA 샘플에 대한 디지탈 RCD 분석용 컴퓨터 시뮬레이션을 나타내는 표 1100이다. 태아 DNA 분획 농도가 낮아졌기 때문에, 염색체 21 과다표출 정도는 감소되며, 이에 따라 정확한 질환 분류를 위해 보다 많은 수의 정보제공 웰이 요구된다. 따라서, m_r = 0.5에서 384개 내지 7680개 웰 범위의 총 웰 수에 대하여 50%, 25% 및 10%의 태아 DNA 농도에 대해 추가로 시뮬레이션을 수행하였다.

[0206] 도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플의 분류에 대하여 m_r = 0.5에서 디지탈 RCD 분석의 정확도의 컴퓨터 시뮬레이션 결과를 나타내는 표 1200이다. 디지탈 RCD의 수행은 보다 높은 태아 DNA 분획 농도를 갖는 경우에 대해 보다 양호하다. 25% 태아 DNA 농도 및 7680개 총 수의 PCR 분석에서, 97%의 정배수체 및 이수체 경우 모두가 부정확한 분류 없이 분류가능할 것이다. 나머지 3%의 경우는 분류가 달성될 수 있을 때까지 추가 분석을 요구한다.

[0207] 디지탈 RCD 분석을 시뮬레이션하기 위한 절차는 디지탈 RNA-SNP 분석에 대해 기재된 것과 유사하였다. 시뮬레이션을 위한 단계를 하기 기재한다.

[0208] 단계 1에서, 푸아송 확률 밀도 함수 하의 2개의 난수를 생성하여 참조 유전자좌, 염색체 1 및 염색체 21 유전자좌를 나타내었다. 정배수체 태아를 갖는 대상체에 대하여, 염색체 1 및 염색체 21 유전자좌 모두의 평균 농도는 동일하였다. 상기 시뮬레이션 분석에서, 웰 당 각 유전자좌에 대한 0.5의 평균 주형 농도를 사용하였다. 삼염색체 21 태아를 갖는 대상체에 대하여, 상기 시뮬레이션에서의 m_r은 0.5였으나, 웰 당 염색체 21 유전자좌의 평균 농도는 표 700에서 나타낸 것과 같이 시험된 샘플에서의 태아 DNA 분획 농도에 의존할 것이다. 웰에 대한 참조 및/또는 염색체 21 유전자좌의 분포를 웰 당 유전자좌의 적절한 평균 농도를 갖는 푸아송 확률 밀도

함수에 따라 생성된 각각의 유전자좌를 나타내는 난수에 의해 측정하였다.

[0209] 단계 2에서, 원하는 수의 웰, 예를 들어 384-웰 플레이트 실험에 대하여 384개 웰이 생성될 때까지 단계 1을 반복하였다. 염색체 1 단독 및 염색체 21 단독에 대해서만 양성인 웰의 수를 계수하였다. 정보제공 웰의 수는 염색체 중 하나에 대해 양성이되, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 염색체 21에 대해 양성인 정보제공 웰의 비율 (P_r)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선의 상한 및 하한을 SPRT 분석에 대해 앞선 섹션에서 기재된 것과 같이 계산하였다.

[0210] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 2개 시나리오 각각에 대해 5000회의 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1100에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주되었다. 두 군 모두에 대하여, 미리-명시된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0211] 단계 4에서, 단계 1 내지 3을 384개 내지 7680개 범위의 총 웰 수에서 10% 25%, 50% 및 100% 태아 DNA를 갖는 샘플에 대해 반복하였다.

[0213] II. 삼염색체 21 검출의 검증

[0214] **A. PLAC4에 대한 RNA-SNP**

[0215] 염색체 21 상의 PLAC4 유전자에 대한 rs8130833 SNP를 사용하여 디지탈 RNA-SNP의 실제적 실행가능성을 입증하였다 (문헌 [Lo, YMD et al. 2007 Nat Med 13, 218-223]). 2개 정배수체 및 2개 삼염색체 21 이형접합 태반으로부터의 태반 DNA 및 RNA 샘플을 분석하였다. 디지탈 RNA-SNP 프로토콜을 사용하되, 역전사 단계를 생략하였으며, 이에 따라 필수적으로 절차를 디지탈 DNA-SNP 분석으로 전환하여, 태반 DNA 샘플을 분석하였다. 정확한 경우 분류의 가능성과 정보제공 웰의 비율 사이에 균형을 맞추기 위해, 본 발명자들은 웰 당 임의의 유형의 하나의 대립유전자를 목적으로 하는 샘플을 희석하고, 96-웰 디지탈 PCR 분석에 의해 확인하였다. 이후, 384-웰 디지탈 RNA-SNP 실험을 하였다. P_r 및 m_r을 계산하고, 상기 m_r 값에 대한 SPRT 곡선을 데이타 해석을 위해 사용하였다.

[0216] 도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 태반 조직에서의 디지탈 RNA-SNP 분석의 표 1300을 나타낸다. 질량 분광 검정에 의해 유전자형을 측정하였다. 실험적으로 얻어진 P_r이 분류불가능한 영역 아래에 있는 경우 "정배수체"로 지정하였고, 실험적으로 얻어진 P_r이 분류불가능한 영역 위에 있는 경우 삼염색체 21을 나타내는 "T21"로 지정하였다. T21, 삼염색체 21. 1회 384-웰 실험으로 DNA 및 RNA 샘플 모두를 사용하여 각각의 상기 경우를 정확하게 분류하였다.

[0217] 추가적으로, 본 발명자들은 정배수체 태아를 갖는 9명의 여성 및 삼염색체 21 태아를 갖는 4명의 여성으로부터의 혈장 RNA 샘플을 시험하였다. 도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 여성으로부터의 모체 혈장의 디지탈 RNA-SNP 분석의 표 1350을 나타낸다. 모든 경우를 정확하게 분류하였다. 한 삼염색체 21 경우 (M2272P)에 대한 초기 결과는 1회 384-웰 실험 이후 SPRT 곡선 사이의 분류불가능한 영역 내에 있었다. 따라서, 추가의 384-웰 실행을 수행하였다. 총 768개 웰로부터의 응집된 데이타로부터 신규 m_r 및 P_r 값을 계산하였으며, 상기 m_r 값을 기준으로 선택된 신규 세트의 SPRT 곡선을 사용하여 분류를 수행하였다. 이후, 상기 경우를 이수체로서 정확하게 스코어링하였다.

[0218] 본 발명자의 실험 및 시뮬레이션 데이타는 디지탈 RNA-SNP가 삼염색체 21 검출에 대한 효과적이며 정확한 방법임을 나타낸다. 모체 혈장 중 *PLAC4* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 13개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지탈 PCR 실험만이 정확한 분류를 위해 요구되었다. 이에 따라, 상기 균질한 실시간 디지탈 PCR-기반 방법은 RNA-SNP 분석을 위한 질량 분광법-기반 방법과는 또다른 방법을 제공한다. 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 종이 태아 염색체 이수성의 디지탈 PCR-기반 검출에 대해 사용될 수 있다는 것을 구상한다. 한 예는 최근에 후성 대립유전자 비 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용된 태아 후성 마커이다 (문헌 [Tong YK et al. 2006 Clin Chem, 52, 2194-2202]). 따라서, 본 발명자들은 디지탈 EAR이 가능한 분

석 기술일 것이라고 예상한다.

[0220] B. RCD

[0221] 또한, 염색체 21 및 1에서의 파라로그 서열을 표적화하는 PCR 분석을 사용하여 삼염색체 21 검출에 대한 디지탈 RCD의 실제적 실행가능성을 연구하였다. 파라로그 유전자좌를 본원에서 예로서 사용하였다. 염색체 21 및 임의의 다른 참조 염색체에 대한 비-파라로그 서열 또한 RCD에 대해 사용할 수 있다. 2개 정배수체 및 2개 삼염색체 21 태반으로부터의 태반 DNA 샘플을 웰 당 염색체 중 하나로부터 대략 1개 표적 주형으로 희석시켰으며, 96-웰 디지탈 PCR 분석에 의해 확인하였다. 각각의 확인된 샘플을 384-웰 디지탈 RCD 실험에 의해 분석하였으며, P_r 및 m_r 값을 계산하였다. 디지탈 RCD에 대하여, 염색체 1 파라로그는 참조 주형이었다. 상기 m_r 값을 사용하여 데이타 해석을 위한 상응하는 세트의 SPRT 곡선을 선택하였다. 모든 태반 샘플을 도 14a에서 나타낸 것과 같이 정확히 분류하였다.

[0222] 모체 혈장에서의 태아 DNA의 시나리오와 같은 디지탈 RCD 방법을 사용하여 과량의 정배수체 DNA와 혼합된 삼염색체 21 DNA를 검출할 수 있는지를 입증하기 위해, 정배수체 모체 혈액 세포 DNA의 배경 중 50% 및 25%의 삼염색체 21 태반 DNA를 함유하는 혼합물을 분석하였다. 10개의 삼염색체 21 및 10개의 정배수체 경우로부터의 태반 DNA를 각각 동일한 양의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 20개의 50% DNA 혼합물을 제조하였다. 도 14b는 본 발명의 실시양태에 따른 50% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 (1440)을 나타낸다. 유사하게는, 5개의 삼염색체 21 및 5개의 정배수체 경우로부터의 태반 DNA를 각각 3배 과량의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 10개의 25% DNA 혼합물을 제조하였다. 도 14c는 25% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 1470을 나타낸다. 도 14b 및 14c에서 나타낸 것과 같이, 모든 정배수체 및 이수체 DNA 혼합물을 정확하게 분류하였다.

[0223] 각각의 샘플은 도 14b 및 14c에 표시된 것과 같이 여러 384-웰 디지탈 PCR 분석 이후 분류가능한 지점에 도달하였다. 50% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 5개의 범위가 요구된다. 25% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 7개의 범위가 요구된다. 매 384회의 디지탈 PCR 분석의 추가와 함께 정확하게 분류되는 경우의 누적 비율은 표 1200에서 제시된 컴퓨터 시뮬레이션에 의해 예상된 것과 같았다.

[0225] III. 디지탈 PCR을 이용한 방법

[0226] A. 디지탈 RNA-SNP

[0227] 모든 RNA 샘플을 먼저 써모스크립트(ThermoScript) 역전사효소 (인비트로젠(Invitrogen))를 사용하여 유전자-특이적 역전사 프라이머에 의해 역전사시켰다. 역전사 프라이머의 서열은 5'-AGTATATAGAACCATGTTTAGGCCAGA-3' (서열 1) (미국 아이오와주 코럴빌 소재의 인테그레이티드 DNA 테크놀러지스(Integrated DNA Technologies))였다. 디지탈 RNA-SNP를 위한 역전사된 RNA (즉, cDNA) 샘플 및 DNA 샘플 (예를 들어, 태반 DNA)의 후속 처리는 본질적으로 동일하였다. 디지탈 PCR 분석 전에, 먼저 DNA 및 cDNA 샘플을 프라이머 5'-CCGCTAGGGTGTCTTTTAAGC-3' (서열 2), 5'-GTGTTGCAATACAAAATGAGTTTCT-3' (서열 3), 및 형광 프로브 5'-(FAM)ATTGGAGCAAATTC(MGBNFQ)-3' (서열 4) (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) (여기서, FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처임)로 이루어진, PLAC4에 대한 실시간 PCR 검정을 사용하여 정량화하였다.

[0228] 앰플리콘이 특정된 HPLC-정제된 단일 가닥 합성 DNA 올리고뉴클레오티드 (싱가포르 소재의 프롤리고(Proligo))의 계열 희석에 의해 보정 곡선을 제작하였다. 서열은 5'-CGCCGCTAGGGTGTCTTTTAAGCTATTGGAGCAAATTCAAATTTGGCTTAAAGAA AAAGAAACTCATTTTGTATTGCAACACCAGGAGTATCCCAAGGGACTCG-3' (서열 5)였다. 반응 부피 25 μL로 2X 태크만 유니버설 PCR 마스타 믹스(TaqMan Universal PCR Master Mix) (어플라이드 바이오시스템즈)를 사용하여 반응을 설정하였다. 각각의 프라이머 400 nM 및 프로브 80 nM을 각각의 반응에 사용하였다. ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈)에서, 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 45주기를 수행하였다. 그 다음, 웰 당 대략 1개의 주형 분자에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 DNA 또는 cDNA 샘플의 계열 희석에

착수하였다. 이러한 농도에서는, 반응 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되며, 이는 96-웰 디지탈 실시간 PCR 분석을 수행함으로써 최초로 확인되었다. 이어서, 한 세트의 비-인트론 관통(spanning) 프라이머: 전방향 프라이머 5'-TTTGTATTGCAACACCATTTGG-3' (서열 6) 및 앞서 기재된 유전자-특이적 역전사 프라이머를 사용하여 384-웰 플레이트에서 디지탈 RNA-SNP 분석을 수행하였다.

[0229] PLAC4 서열 상 rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 태크만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 5'- (FAM)TCGTCGTCTAACTTG(MGBNFQ)-3' (서열 7) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 8)였다. 반응 부피 5 μL로 2X 태크만 유니버셜 PCR 마스타 믹스를 사용하여 반응을 설정하였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 107 nM 및 대립유전자-A-특이적 프로브 357 nM을 함유하였다. 반응은 ABI PRISM 7900HT 서열 검출 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 45주기를 수행하였다. 반응 동안, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 상기 소프트웨어는 기준선 및 임계값을 자동으로 계산하였다. A 또는 G 대립유전자에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다.

[0231] **B. 디지탈 RCD 분석**

[0232] 본 연구에 사용되는 모든 태반 및 모체 연막 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀러지(NanoDrop Technology))에 의해 정량화하였다. 6.6 pg/세포의 전환을 사용하여 DNA 농도를 카피/μL로 전환하였다. 웰 당 대략 1개의 주형에 상응하는 DNA의 양을, DNA 샘플을 계열 희석함으로써 결정하고 96-웰 포맷에서 실시간 PCR 검정에 의해 확인하였는데, 여기서 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되었다. 확인 플레이트에 대한 PCR 설정은 참조 염색체에 대한 프로브를 추가한 것만 제외하고는 이하에 기재된 것과 동일하였다. 디지탈 RCD 분석에서, 염색체 21 및 1 상의 파라로그 유전자좌 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915])를 먼저 전방향 프라이머 5'-GTTGTTCTGCAAAAAACCTTCGA-3' (서열 9) 및 역방향 프라이머 5'-CTTGGCCAGAAATACTTCATTACCATAT-3' (서열 10)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그를 표적으로 하도록 2개의 염색체-특이적 태크만 프로브를 설계하였는데, 그들의 서열은 각각 5'-(FAM)TACCTCCAT AATGAGTAAA(MGBNFQ)-3' (서열 11) 및 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3' (서열 12)였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스 (어플라이드 바이오시스템즈), 각각의 프라이머 450 nM 및 각각의 프로브 125 nM을 함유하였다. 총 반응 부피는 5μL/웰이었다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 50주기를 수행하였다. 모든 실시간 PCR 실험은 ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈) 상에서 수행하였고, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 디폴트 기준선 및 매뉴얼 임계값을 사용하였다. 염색체 21 또는 염색체 1에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다. SPRT에 의해 질환 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트가 분석될 것이다.

[0234] <u>IV. 미세유체-기반 디지탈 PCR의 사용</u>

[0235] **A. 디지탈 RNA-SNP**

[0236] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 PCR 분석의 성능을 증명한다. 플루이다임 바이오마크(Fluidigm BioMark)™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0237] 정배수체 또는 삼염색체 21 태아를 가진 임신 여성으로부터 태반 조직 및 모체 말초 혈액 샘플을 채취하였다. 프라이머 연장에 이은 질량 분광법에 의해 태반 DNA 샘플에서 PLAC4 유전자 상 rs8130833 SNP의 유전자형 결정을 수행하였다. 태반 및 모체 혈장 샘플로부터 RNA를 추출하였다.

[0238] 모든 RNA 샘플을 써모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 역전사 프라이머 (5'-AGTATATAGAACCATGTTTAGGCCAGA-3' (서열 13))에 의해 역전사시켰다. 태반 cDNA

샘플에 대해, 웰 당 대략 1개의 주형 분자에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 계열 희석을 수행하였다.

[0239] 12.765 디지탈 어레이 (플루이다임)에 의해 바이오마크 시스템™ (플루이다임) 상에서 디지탈 PCR을 수행하였다. 각각의 디지탈 어레이는 12개의 샘플-검정 혼합물을 수용하기 위한 12개의 패널로 이루어져 있다. 각각의 패널을 7 nL 반응물/웰을 수행하기 위한 765개 웰로 추가 분할하였다. PLAC4 유전자 상의 rs8130833 SNP 영역을 전방향 프라이머 (5'-TTTGTATTGCAACACCATTTGG-3' (서열 14)) 및 앞서 기재된 유전자-특이적 역전사 프라이머에 의해 증폭시켰다. rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 태크만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 5'-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3' (서열 15) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 16)였다. 반응 부피 10 μL로 2X 태크만 유니버셜 PCR 마스타 믹스를 사용하여 1개의 어레이 패널에 대한 반응을 설정하였다. 각각의 반응은 1X 태크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 53.5 nM, 대립유전자-A-특이적 프로브 178.5 nM 및 cDNA 샘플 3.5 μL를 함유하였다. 각각의 태반 cDNA 샘플에 대해서는 1개의 반응 패널이 사용된 반면, 각각의 모체 혈장 샘플에 대해서는 12개의 패널이 사용되었다. 샘플-검정 혼합물을 나노플렉스(NanoFlex)™ IFC 제어기 (플루이다임)에 의해 디지탈 어레이에 로딩하였다. 반응은 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0240] 1개의 정배수체 및 2개의 T21 이형접합체 태반으로부터의 태반 RNA 샘플을 765-웰 반응 패널에서 분석하였다. 각각의 샘플에 대해, A 또는 G 대립유전자 (둘 다는 아님)에 대해 양성인 것을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 과다표출된 대립유전자의 비율 (P_r)을 결정하였다. 이들 실행의 웰 당 정확한 평균 참조 주형 농도 (m_r)에 적당한 SPRT 곡선을 적용하여, 실험적으로 얻은 P_r이 정배수체를 지시하는지 T21 샘플을 지시하는지를 결정하였다. 도 15a에 나타낸 바와 같이, 상기 접근법을 사용하여 모든 RNA 샘플이 정확하게 분류되었다.

[0241] 정배수체 태아를 가진 4명의 여성 및 삼염색체 21 태아를 가진 1명의 여성으로부터 채취한 혈장 RNA 샘플을 추가로 시험하였다. 각각의 샘플을 12개의 765-웰 반응 패널, 즉 혈장 RNA 샘플 당 9180회의 반응으로 분석하였다. 도 15b는 상기 혈장 RNA 샘플의 12개 패널 각각에 대한 정보제공 웰의 수를 나타낸다. 표에 나타낸 바와 같이, 혈장 샘플 중 주형 농도는 임의의 1개의 반응 패널 중의 정보제공 웰의 수로는 SPRT 분류에 충분하지 않을 정도로 묽었다. 상기 샘플을 정배수체 샘플로 분류하기 전에, 3개의 반응 패널로부터의 정보제공 웰을 합해야 했다 (도 15c). 도 15c는 2 내지 12개의 패널로부터 종합한 데이타를 사용하여 모든 혈장 사례가 정확하게 분류될 수 있었음을 나타낸다.

[0242] 디지탈 PCR을 수행하기 위한 매뉴얼 방법과 비교해서, 상기 미세유체-기반 방법은 훨씬 더 빠르고 덜 노동 집약적이었다. 전체 과정이 2시간 반 내에 완료될 수 있었다.

[0244] **B. 삼염색체 18의 출생 전 검출을 위한 디지탈 RNA-SNP 분석**

[0245] 본 실시예에서는, 삼염색체 18 태아에 대한 다형성 대립유전자의 비율 불균형을 검출하기 위해, 염색체 18 상의 태반-발현 전사체인 세르핀 펩티다제 억제제 클레이드 B (오브알부민) 멤버 2 (SERPINB2) mRNA에 대한 디지탈 PCR-기반 대립유전자 식별 검정을 사용하였다. QIAamp DNA 미니 키트 (독일 힐덴 소재의 퀴아젠(Qiagen)) 및 TRIzol 시약 (미국 캘리포니아주 칼즈배드 소재의 인비트로젠)을 제조업자의 프로토콜에 기재된 바와 같이 각각 사용하여, 태반 조직 샘플로부터의 DNA 및 RNA 추출을 수행하였다. 추출한 태반 RNA 샘플을 DNase I (인비트로젠) 처리하여 오염된 게놈 DNA를 제거하였다. 앞서 기재된 바와 같은 매스어레이 컴팩트(MassARRAY Compact) (미국 샌디에고 소재의 시퀘놈(Sequenom))를 사용하여 호모지너스 매스익스텐드(Homogenous MassEXTEND) (hME) 검정에 의해 태반 조직 DNA 샘플에서 SERPINB2 유전자 상 rs6098 SNP의 유전자형 결정을 수행하였다.

[0246] 써모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 프라이머 5'-CGCAGACTTCTCACCAAACA-3' (서열 17) (미국 아이오와주 코럴빌 소재의 인테그레이티드 DNA 테크놀러지스)에 의해 태반 조직 RNA 샘플 상에서 SERPINB2 전사체에 대한 역전사를 수행하였다. 모든 cDNA 샘플을, 반응 웰 당 1개의 주형 분자의 평균 농도에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 하는 농도로 희석하였다. 태크만 유니버셜 PCR 마스타 믹스 (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) 및 바이오마크™ PCR 시약 (미국 샌프란시스코 소재의 플루이다임)을 사용하여 디지탈 PCR을

설정하였다. 전방향 프라이머 5'-CTCAGCTCTGCAATCAATGC-3' (서열 18) (인테그레이티드 DNA 테크놀러지스) 및 역방향 프라이머 (역전사에 사용된 유전자-특이적 프라이머와 동일한 것)는 600 nM의 농도로 사용하였다. SERPINB2 서열 상 rs6098 SNP의 A 또는 G 대립유전자를 표적으로 하는 2개의 테크만 프로브는 5'-(FAM)CCACAGGGAATTATTT(MGBNFQ)-3' (서열 19) 및 5'-(VIC)CCACAGGGGATTATTT(MGBNFQ)-3' (서열 20) (어플라이드 바이오시스템즈)였다. FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처이며, 각각 300 nM 및 500 nM의 농도로 사용하였다. 각각의 샘플-시약 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)를 사용하여 바이오마크™ 12.765 디지탈 어레이 상의 765개 반응 웰에 분할하였다. 분할 후, 어레이를 열 증폭 및 형광 검출을 위해 바이오마크™ 실시간 PCR 시스템 (플루이다임)에 넣었다. 반응을 50℃에서 2분 동안 개시하고, 95℃에서 5분 동안 지속한 후, 95℃에서 15초 및 59℃에서 1분의 45주기를 수행하였다. 증폭 후, 정보제공 웰 (A 또는 G 대립유전자 단독에 대해 양성인 것)의 수 및 대립유전자 둘 다에 대해 양성인 웰의 수를 계수하여 순차 확률비 검증 (SPRT) 분석에 제공하였다.

[0247] 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 반면 (1:1), 삼염색체 18에 대해서는, 한 대립유전자의 추가 카피가 존재함으로써 태아 게놈 내에 2:1의 비율이 주어질 것이다. 상이한 샘플의 해석에 대한 일련의 SPRT 곡선을 생성시켰다. 이들 곡선은 분류에 필요한 정보제공 웰의 주어진 총 수 (x축)에 대한, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 예상 비율 P_r (y축)을 나타낸다. 각각의 샘플에 대해, 실험적으로 유도된 P_r을 예상 P_r 값과 비교하였다. 상부 곡선 위쪽의 샘플은 삼염색체 18로 분류되고, 하부 곡선 아래쪽의 샘플은 정배수체로 분류되었다. 두 곡선 사이의 공간은 분류불가능한 영역이다.

[0248] 태아 삼염색체 18의 검출을 위한 디지탈 RNA-SNP 분석의 실행가능성을 SERPINB2 유전자 상 rs6098 SNP를 사용하여 증명하였다. 먼저 이형접합체 사례를 확인하기 위해서, 질량 분광법에 의해 정배수체 및 삼염색체 18 태아를 갖는 대상체로부터의 태반 조직 DNA 샘플의 유전자형을 결정하였다. 9개의 정배수체 및 3개의 삼염색체 18 이형접합체 태반을 발견하여 디지탈 RNA-SNP 분석에 제공하였다. 각각의 샘플에 대해 P_r 및 m_r을 계산하고, 상기 m_r값에 대한 SPRT 곡선을 질환 분류를 위해 사용하였다. 도 16a에 나타낸 바와 같이, 모든 샘플이 정확하게 분류되었다. 삼염색체 18 태반에 대한 P_r값은 분류불가능한 영역의 위쪽인 반면, 정배수체 태반에 대한 P_r값은 상기 영역의 아래쪽에 속하였다.

[0249] m_r = 0.1, 0.2 및 0.3을 기준으로 한 SPRT 곡선을 갖는 샘플을 도 16b에 나타내었다. 이들 데이타는 디지탈 RNA-SNP 방법이 삼염색체 18 임신에 대한 매우 유용한 진단 도구임을 시사한다. 2개의 곡선은 분류불가능한 영역에 대한 경계선을 나타낸다. 상부 곡선 위쪽의 데이타 지점을 갖는 샘플은 이수체로 분류되고, 하부 곡선 아래쪽의 데이타 지점을 갖는 샘플은 정배수체로 분류되었다.

[0251] **C. 디지탈 RCD 분석**

[0252] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 RCD 분석의 성능을 증명한다. 플루이다임 바이오마크™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0253] 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 태반 조직, 모체 혈액 세포 및 양수 샘플을 얻었다. 10개의 T21 및 10개의 정배수체 사례로부터의 태반 DNA를 각각 동량의 정배수체 모체 혈액 세포 DNA와 혼합함으로써 20개의 50% DNA 혼합물을 제조하였다. 혼합물 샘플 중 정확한 태아 비율을 확실히 하기 위해, 추출한 DNA를 먼저 260 nm에서 광학 밀도 (OD) 측정에 의해 정량화하였다. 그 다음, 12.765 디지탈 어레이 (플루이다임)를 사용하여 바이오마크™ 시스템 (플루이다임)에 의해 이들을 디지탈 정량화하였다. 샘플의 정량화 검정은 참조 염색체에 대한 프로브를 사용한 것만 제외하고는 이하에 기재된 것과 동일하였다.

[0254] 50% DNA 혼합물 및 양수 샘플 내 염색체 양을, 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정하였다. 먼저 염색체 21 및 1 상의 1쌍의 파라로그 유전자좌의 101-bp 앰플리콘을 전방향 프라이머 5'-GTTGTTCTGCAAAAAACCTTCGA-3' (서열 21) 및 역방향 프라이머 5'-CTTGGCCAGAAATACTTCATTACCATAT-3' (서열 22)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그 사이를 구별하도록 2개의 염색체-특이적 테크만 프로브를 설계하였는데, 그들의 서열은 각각

5'-(FAM)TACCTCCATAATGAGTAAA(MGBNFQ)-3' (서열 23) 및 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3' (서열 24)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 환언하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0255] 삼염색체 18 (T18)을 검출하기 위한 디지탈 RCD 접근법의 용도를 증명하기 위해, 염색체 21 및 18 상의 파라로그 서열을 표적으로 하는 또다른 검정을 설계하였다. 먼저 염색체 21 및 18 상의 파라로그 유전자좌의 128-bp 앰플리콘을 전방향 프라이머 5'-GTACAGAAACCACAAACTGATCGG-3' (서열 25) 및 역방향 프라이머 5'-GTCCAGGCTGTGGGCCT-3' (서열 26)에 의해 동시 증폭시켰다. 염색체 21 및 18 파라로그 사이를 구별하도록 2개의 염색체-특이적 태크만 프로브를 설계하였는데, 그들의 서열은 각각 5'-(FAM)AAGAGGCGAGGCAA(MGBNFQ)-3' (서열 27) 및 5'-(VIC)AAGAGGACAGGCAAC(MGBNFQ)-3' (서열 28)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 환언하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0256] 모든 실험은 12.765 디지탈 어레이 (플루이다임)를 사용하여 바이오마크™ 시스템 (플루이다임) 상에서 수행하였다. 반응 부피 10 μL로 2X 태크만 유니버설 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 1개의 패널에 대한 반응을 설정하였다. 각각의 반응은 1X 태크만 유니버설 PCR 마스타 믹스, 각각의 프라이머 900 nM, 각각의 프로브 125 nM 및 50% 태반/모체 혈액 세포 DNA 샘플 3.5 μL를 함유하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)에 의해 디지탈 어레이에 로딩하였다. 검출을 위해 바이오마크™ 시스템 상에서 반응을 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0257] 정배수체 및 T21 50% 태반/모체 혈액 세포 DNA 샘플을 디지탈 어레이 상에서 chr21/chr1 검정에 의해 분석하였다. 각각의 샘플에 대해, chr21 또는 chr1 마커 단독 (둘 다는 아님)에 대해 양성인 웰을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 과다표출된 마커의 비율 (P_r)을 결정하였다. 디지탈 PCR 패널 중 어느 하나에 대한 웰 당 정확한 평균 참조 주형 농도 (m_r)에 적당한 SPRT 곡선을 적용하여, 실험적으로 얻은 P_r이 정배수체 샘플을 지시하는지 T21 샘플을 지시하는지를 결정하였다. 결정이 이루어질 수 있을 때까지, 분류되지 않은 채로 남아있는 샘플에 대한 여분의 패널로부터 데이타를 종합하였다. 도 17에 나타낸 바와 같이, 모든 50% 태반/모체 혈액 세포 DNA 샘플이 상기 접근법을 사용하여 필요한 1 내지 4개 패널 범위의 데이타에 의해 정확하게 분류되었다. 도 18에 나타낸 바와 같이, 정확한 분류를 위한 결정 경계선을 나타내기 위해 SPRT 곡선을 또한 플롯팅하였다.

[0258] 정배수체 태아를 가진 23명의 여성 및 T21 태아를 가진 6명의 여성으로부터 얻은 양수 샘플에 대한 RCD 분석을 추가로 적용하였다. 각각의 샘플을 단일 765-웰 반응 패널에서 chr21/chr1 검정에 의해 분석하였다. 도 19는 SPRT 분류 개요를 나타낸다. 도 19에 나타낸 바와 같이, 모든 29개 샘플이 정확하게 분류되었다. 따라서, 디지탈 RCD 방법은 출생 전 진단에 사용되는 다양한 종류의 샘플 유형, 예컨대 양수 및 융모막 융모 생검에서 초위성체 (문헌 [Levett LJ, et al. A large-scale evaluation of amnio-PCR for the rapid prenatal diagnosis of fetal trisomy. Ultrasound Obstet Gynecol 2001; 17: 115-8]) 또는 단일 뉴클레오티드 다형성 (SNP) (문헌 [Tsui NB, et al. Detection of trisomy 21 by quantitative mass spectrometric analysis of single-nucleotide polymorphisms. Clin Chem 2005; 51: 2358-62]) 마커 또는 실시간 비-디지탈 PCR (문헌 [Zimmermann B, et al. Novel real-time quantitative PCR test for trisomy 21. Clin Chem 2002; 48: 362-3])을 사용하는 삼염색체의 검출에 대한 대안적인 접근법이다.

[0259] T18 사례를 검출하기 위한 시도로, 3개의 정배수체 및 5개의 T18 태반 DNA 샘플 상에 chr21/chr18 검정을 적용하였다. 모든 정보제공 웰 중 과다표출된 마커의 비율 (P_r)을 계산하였다. 이들 모두는 1개의 T18 사례가 정배수체로 잘못 분류된 것을 제외하고는 정확하게 분류되었다. 결과를 도 20에 요약하였다.

[0261] V. 질량 분광법 플랫폼에 대한 멀티플렉스 디지탈 RCD 검정의 이용

[0262] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한적이다 (문헌 [Lo YMD. et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에 상염색체 유전자좌, β-글로빈 유전자의 중앙 모체 혈장 농도는 태아 및 모계 모두에 기인하여 986개 카피/mL인 것으로 밝혀진 바 있다 (문헌 [Lo YMD. et al. 1998 Am J

Hum Genet 62, 768-7758]). 7,680개의 분자를 포획하기 위해서, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 수득가능한 상기 부피의 혈장은 통상적인 실행의 한계이다. 그러나, 본 발명자들은 chr21 및 참조 염색체 표적의 다중 세트를 디지탈 RCD 분석을 위해 합할 수 있는 것으로 생각한다. 5쌍의 chr21 및 참조 염색체 표적에 대해, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 멀티플렉스 단일 분자 PCR을 수행할 수 있었다. 이러한 멀티플렉스 단일 분자 분석의 견고성은 단일 분자 일배체형화에 대해 미리 입증된 바 있다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci U S A 100, 7449-7453]).

[0263] 일례에서, 태반 조직 및 모체 혈액 세포 샘플을 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 수득하였다. 5개의 정배수체 및 5개의 T21 태반 DNA 샘플을 각각 동일한 비율의 모체 혈액 세포 DNA와 혼합하여 50% 태아 DNA를 갖는 혈장 샘플을 모방한 10개의 DNA 혼합물을 생성하였다. 혼합물 샘플 중 정확한 태아 비율을 확실하게 하기 위해, 추출된 DNA를 먼저 260 nm에서의 광학 밀도 (OD) 측정으로 정량화하였다. 이어서, 이를 384-웰 포맷에서 실시간 PCR에 의해 디지탈 방식으로 정량화하였다. 샘플을 정량화하기 위한 검정은 상기 디지탈 RCD 분석의 예에 기재된 바와 동일하였다.

[0264] 50% 믹스 내 염색체 양을 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석으로 결정하였다. 상기 방법은 디지탈 상대적 염색체 양 (RCD) 분석으로 칭해진다. 염색체 21 및 1 상의 한 쌍의 파라로그 유전자좌의 121-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'-ACGTTGGATGGTTGTTCTGCAAAAAACCTTCGA-3' (서열 29) 및 역방향 프라이머 5'-ACGTTGGATGCTTGGCCAGAAATACTTCATTACCATAT-3' (서열 30)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 1 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-CTCATCCTCACTTCGTACCTC-3' (서열 31)이다.

[0265] T21 사례를 검출하는 멀티플렉싱 디지탈 PCR 검정의 유용성을 입증하기 위해서, 염색체 21 및 18 상의 파라로그 서열을 표적으로 한 또다른 디지탈 RCD 검정을 고안하였다. 염색체 21 및 18 상의 파라로그 유전자좌의 148-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'- ACGTTGGATGGTACAGAAACCACAAACTGATCGG-3' (서열 32) 및 역방향 프라이머 5'-ACGTTGGATGGTCC AGGCTGTGGGCCT-3' (서열 33)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 18 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-ACAAAAGGGGGAAGAGG-3' (서열 34)이다.

[0266] 멀티플렉스 디지탈 RCD 분석은 프라이머 연장 프로토콜을 이용하여 수행되었다. PCR 반응은 진앰프(GeneAmp) PCR 코어 시약 키트 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 5 μL로 설정하였다. 각각의 반응물은 1X 완충액 II, 2 mM $MgCl_2$, 200 μM dNTP 믹스, 0.2 U 앰플리태크 골드(AmpliTaq Gold), 각각 200 nM의 4종의 프라이머 및 50% DNA 믹스를 함유하였다. 상기 검정/샘플 혼합물을 384-웰 PCR 플레이트에 분배하고, 50℃에서 2분 동안 반응을 개시한 후, 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40회 주기를 수행하였다.

[0267] PCR 생성물을 쉬림프 알칼린 포스파타제 (SAP) 처리하여 혼입되지 않은 dNTP를 제거하였다. 상기 혼합물을 37℃에서 40분 동안, 이어서 85℃에서 5분 동안 인큐베이션하였다. 이어서, 프라이머 연장 반응을 수행하였다. 간단히 말해서, 연장 칵테일 중 chr21/chr1 검정으로부터의 771 nM의 연장 프라이머, chr21/chr18 검정으로부터의 1.54 μM의 연장 프라이머, 0.67 U 써모시퀘나제(Thermosequenase) (시퀘놈), 및 각각 64 μM의 ddCTP, ddGTP, dATP 및 dTTP를 SAP-처리된 PCR 생성물에 첨가하였다. 반응 조건은 94℃에서 2분, 이어서 94℃에서 5초, 50℃에서 5초 및 72℃에서 5초의 80회 주기였다. 16 μL의 물 및 3 mg의 클린 레진(Clean Resin) (시퀘놈)을 최종 처리(clean up)를 위해 연장 생성물에 첨가하였다. 상기 혼합물을 회전기에서 20 내지 30분 동안 혼합한 후, 361 g에서 5분 동안 원심 분리단계를 수행하였다. 15 내지 25 nL의 최종 생성물을 매스어레이 나노 분산기 S (시퀘놈)로 스펙트로칩(SpectroCHIP) (시퀘놈) 상에 분배하였다. 스펙트로칩으로부터의 데이타 획득을 매스어레이 분석기 컴팩트 질량 분광기 (시퀘놈)에서 수행하였다. 분석을 위해 질량 데이타를 매스어레이 타이퍼(MassARRAY Typer) (시퀘놈) 소프트웨어로 임포트하였다.

[0268] 5개의 정배수체 및 5개의 T21 50% 태반/모체 DNA 샘플을 이중 RCD 검정으로 분석하였다. 각 샘플에 대해, 오직 chr21 또는 chr1 또는 chr18 마커에만 양성인 웰을 포함하여 각각의 검정으로부터 정보제공 웰의 수를 계수하였다. 전체 정보제공 웰 중 chr21 마커의 비율 (P_r)을 각각의 RCD 검정에 대해 개별적으로 계산하였다. 이어서, 순차 확률비 검증 (SPRT)을 적용하여 P_r이 정배수체 샘플을 나타내는 지 또는 T21 샘플을 나타내는 지 결

정하였다. 이렇게 하여, 각각의 플레이트를 2회 계수함으로써 필요한 웰의 수를 감소시켰다.

[0269] chr21/chr1 검정을 통상 먼저 적용하였다. 샘플이 미분류 상태로 있는 경우, chr21/chr18 검정으로부터의 값을 추가 계산에 포함시킬 것이다. 여분의 플레이트를 결정이 이루어질 수 있을 때까지 미분류 상태로 있는 샘플에 사용하였다. 도 21에 나타낸 바와 같이, 모든 정배수체 50% 믹스 샘플을 단일 384-웰 플레이트를 사용하여 정확히 분류하였다. 여러 T21 사례는 정확한 분류를 위해 2개 이상의 플레이트를 필요로 하였다. 하나의 검정만을 이용한 경우, 분류 달성시 필요한 정보제공 웰의 수에 이르기 위해서 더 많은 플레이트가 필요할 것이다. 예를 들어, RCD 검정 중 어느 하나를 단독 이용한 경우에는 사례 N0230에 대한 데이타가 분류될 수 없었다. 그러나, 2개의 검정으로부터의 데이타를 합한 경우에는 정확한 분류가 이루어졌다. 이중 RCD 검정을 이용하지 않은 경우, 추가의 분석 플레이트가 필요할 것이다. 본 발명자들은 보다 높은 수준의 멀티플렉싱 검정으로 웰 수의 추가 감소를 예상할 것이다.

[0270] 또다른 예에서, 본 발명자들은 염색체 21, 및 다른 비-염색체-21 상염색체 상에 위치한 그의 상응하는 파라로그 파트너에 대한 4종의 상이한 앰플리콘을 표적으로 하는 4-플렉스 검정을 개발하였다. 이러한 4-플렉스 검정은 정배수체 및 삼염색체 21 임신 여성으로부터의 샘플의 디지탈 RCD 분석, 이어서 SPRT 분류에 사용되었다. 태반 샘플로부터의 DNA 추출은 퀴아앰프 조직 키트 (독일 힐덴 소재의 퀴아젠)를 사용하여 수행되었다.

[0271] 상기 연구에 사용된 모든 태반 및 모체 연막 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀러지)로 정량화하였다. DNA 농도를 6.6 pg/세포의 전환을 이용하여 게놈 당량 (GE)/μL로 전환하였다. 웰 당 대략 1개의 주형에 해당하는 DNA의 양을 DNA 샘플의 계열 희석으로 결정하였다. 이러한 조건하에, 본 발명자들은 대략 37%의 웰이 음성 증폭을 나타낼 것으로 예상할 것이다. 멀티플렉스 디지탈 RCD 분석에서, 4 세트의 파라로그 서열 표적을 선택하였다: 염색체 21 및 1 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTTGATGAAGTCTCATCTCTACTTCG-3' (서열 35) 및 역방향 프라이머 5'-ACGTTGGATGCAATAAGCTTGGCCAGAAATACT-3' (서열 36)를 사용하여 동시 증폭시켜 81 bp의 앰플리콘을 생성하였다. 염색체 21 및 7 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGAATTTAAGCTAAATCAGCCTGAACTG-3' (서열 37) 및 역방향 프라이머 5'-ACGTTGGATGGTTTCTCATAGTTCATCGTAGGCTTAT-3' (서열 38)를 사용하여 동시 증폭시켜 82 bp의 앰플리콘을 생성하였다. 염색체 21 및 2 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTCAGGCAGGGTTCTATGCAG-3' (서열 39) 및 역방향 프라이머 5'-ACGTTGGATGAGGCGGCTTCCTGGCTCTTG-3' (서열 40)를 사용하여 동시 증폭시켜 101 bp의 앰플리콘을 생성하였다. 염색체 21 및 6 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGCTCGTCTCAGGCTCGTAGTT-3' (서열 41) 및 역방향 프라이머 5'-ACGTTGGATGTTTCTTCGAGCCCTTCTTGG-3' (서열 42)를 사용하여 동시 증폭시켜 102 bp의 앰플리콘을 생성하였다. 각각의 반응물은 10X 완충액 II (어플라이드 바이오시스템즈), $MgCl_2$ 및 각각 100 nM의 프라이머를 함유하였다. 총 반응 부피는 5 μL/웰이었다. 95℃에서 5분 동안 반응을 개시한 후, 95℃에서 30초, 62℃에서 30초 및 72℃에서 30초의 45회 주기를 수행하고, 72℃에서 7분 동안 최종 연장을 수행하였다. 모든 통상적인 PCR 증폭은 진앰프 PCR 시스템 9700 (어플라이드 바이오시스템즈) 상에서 수행되었다. 혼입되지 않은 뉴클레오티드를 쉬림프 알칼린 포스파타제 (SAP) 처리로 비활성화시켰다. 각각의 반응물은 10X SAP 완충액 (시쿼놈) 및 SAP 효소 (시쿼놈)를 함유하였다. 2 μL의 SAP 믹스를 각각의 PCR에 첨가하였다. SAP 반응물을 37℃에서 40분 동안 및 85℃에서 5분 동안 인큐베이션하였다. SAP 처리 후, iPLEX 골드(Gold) 키트 (시쿼놈)를 사용하여 PCR 생성물 상에서 프라이머 연장 반응을 수행하였다. 염색체 21 및 1 상의 파라로그 유전자좌에 대한 파라로그 서열 미스매치 (PSM)는 연장 프라이머 5'-GTCTCATCTCTACTTCGTACCTC-3' (서열 43)에 의해 검사되었다. 염색체 21 및 7 상의 파파로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TTTTACGCTGTCCCCATTT-3' (서열 44)에 의해 검사되었다. 염색체 21 및 2 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-GGTCTATGCAGGAGCCGAC-3' (서열 45)에 의해 검사되었다. 염색체 21 및 6 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TGGGCGCGGGAGCGGACTTCGCTGG-3' (서열 46)에 의해 검사되었다. 각각의 반응물은 10X iPLEX 완충액 (시쿼놈), iPLEX 종결 믹스 (시쿼놈), iPLEX 효소 (시쿼놈), 및 1.03 μM로 사용된 염색체 21 및 6에 대한 PSM용 연장 프라이머를 제외하고는 각각 343 nM의 연장 프라이머를 함유하였다. 2 μL의 iPLEX 믹스를 5 μL의 PCR 생성물에 첨가하였다. 200-쇼트-사이클 프로그램에 따라 iPLEX 반응을 순환시켰다. 간단히 말해서, 샘플을 먼저 94℃에서 35초 동안 변성시킨 후, 52℃에서 5초 동안 어닐링하고, 80℃에서 5초 동안 연장하였다. 어닐링 및 연장 주기를 총 5회의 주기 동안 4회 이상 반복하고, 이어서 94℃, 5초의 변성 단계로 역 순환시킨 후, 다시 5회 주기의 어닐링 및 연장 루프를 수행하였다. 1회의 변성 단계와 함께 5회의 어닐링 및 연장 주기를 39회 반복하여 총 40회에 이르게 하였다. 72℃에서 3분 동안 최종 연장을 수행하였다. iPLEX 반응 생성물을 16 μL의 물로 희석하고, 각각의 PCR에 대해 6 mg의 수지로 탈염시켰다. 384-웰 플레이트를 1600 g에서 3분 동안 원심분리한

후, 스펙트로칩 (시쿼놈) 및 매트릭스-보조 레이저 탈착/이온화 비행시간형 (MALDI-TOF) 질량 분광법 MS 분석기 (시쿼놈) 상에 분배하였다.

[0272] 염색체 21에 대해서만 또는 4개의 검정 각각에 대한 참조 염색체에 대해서만 양성인 웰의 수를 독립적으로 기록하였다. 각각의 검정에 있어서, 염색체 21 및 참조 염색체에 대한 푸아송 교정된 분자 수를 계산하였다. 염색체 21에 대한 푸아송 교정된 분자 수의 합 뿐만 아니라 전체 4개의 검정으로부터의 푸아송 교정된 참조 염색체 수의 합을 계산하였고, 이는 4-플렉스 검정에 대한 정보제공 계수로 간주되었다. P_r 값은 4-플렉스 검정에 대한 염색체 21 계수를 4-플렉스 검정에 대한 염색체 21 및 참조 염색체 계수의 합으로 나눈 값이었다. 실험적으로 유도된 P_r 값을 SPRT 분석에 적용하였다. SPRT에 의해 질환 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트를 분석할 것이다. 총 2개의 50% 정배수체 태반 게놈 DNA/50% 모체 연막 DNA 믹스 및 2개의 50% 삼염색체 21 태반 게놈 DNA/50% 모체 연막 DNA 믹스를 분석하였다.

[0273] 실험적으로 유도된 P_r 값을 귀무가설 또는 대립가설 시험시의 예상 P_r 값과 비교할 것이다. 다르게는, 주어진 수의 정보제공 계수에 대한 P_r이 질환 분류에 필요한 통계적 신뢰 수준에 도달하지 못한 경우, 귀무가설 또는 대립가설은 모두 수용될 수 없을 것이다. 이러한 경우는 보다 많은 데이타가 이용가능해질 때까지 분류불가능인 것으로 간주되었다.

[0274] 각 샘플의 결과 및 SPRT 분류를 도 22a 및 22b에 표로 작성하였다. 2개의 정배수체 샘플은 SPRT 분류에 도달할 수 있기 전에 2개 및 5개의 384-웰 멀티플렉스 디지탈 RCD 분석을 필요로 하였다. 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 삼염색체 21 샘플은 둘 다 단지 하나의 384-웰 멀티플렉스 디지탈 RCD 분석으로 각각 정확히 분류되었다. 유사하게, 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 그러나, 4-플렉스 검정으로부터의 혼합 계수는 정확한 SPRT 분류를 가능하게 하였다. 이들 데이타는, 멀티플렉스 디지탈 RCD의 이용에 의해 정보제공 계수의 유효 수가 단일-플렉스 디지탈 RCD 검정의 이용과 비교하여 주어진 수의 수행된 디지탈 PCR 분석에 대해 상당히 증가하였음을 예증하였다.

[0276] VI. 디지탈 후성 상대적 염색체 양의 사용

[0277] 여기서, 본 발명자들은, 염색체 이수성에 관여하는 염색체 (즉, 삼염색체 21에서 염색체 21) 및 참조 염색체 상의 태아-특이적 DNA 메틸화 패턴 또는 다른 후성 변화를 나타내는 후성 마커를 디지탈 PCR 분석에 적용하는, 소위 디지탈 후성 상대적 염색체 양 (디지탈 ERCD)이라 불리는 접근법을 약술한다. 정상 태아를 가진 임신 여성으로부터 추출된 혈장 DNA 내 참조 염색체 후성 마커에 대해 양성인 웰 수에 대한 염색체 21 후성 마커에 대해 양성인 웰 수의 비율은 참조 범위를 제공할 것이다. 상기 비율은 태아가 삼염색체 21을 갖는 경우 증가할 것으로 예상될 것이다. 하나 초과의 염색체 21 마커 및 하나 초과의 참조 염색체 마커를 본 분석에서 사용할 수 있다는 것이 당업자에게 명백하다.

[0278] 태아 (태반)-특이적 메틸화 패턴을 나타내는 염색체 21 상의 유전자의 일례는 완전카르복실라제 합성효소 (Holocarboxylase synthetase, HLCS) 유전자이다. HLCS는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 이는 본원에 참조로 포함된 미국 특허 출원 제11/784499호에 포함된다. 태아 (태반)-특이적 메틸화 패턴을 나타내는 참조 염색체 상의 유전자의 일례는 염색체 3 [10] 상의 RASSF1A 유전자이다. RASSF1A는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 본원에 참조로 포함된 미국 특허 출원 제11/784501호를 참조한다.

[0279] 모체 혈장을 이용한 태아 내 삼염색체 21의 디지탈 PCR 검출에 대한 과-메틸화된 HLCS 및 과-메틸화된 RASSF1A의 적용에서, 모체 말초 혈액을 먼저 수집한다. 이어서, 상기 혈액을 원심분리하고, 혈장을 수집한다. 이어서, 혈장으로부터의 DNA를 당업계에 널리 공지된 기술, 예컨대 퀴아앰프 블러드(Blood) 키트 (퀴아젠)를 사용하여 추출한다. 이어서, 혈장 DNA를 1종 이상의 메틸화-감수성 제한 효소, 예컨대 HpaII 및 BstUI를 사용하여 소화시킨다. 이들 메틸화-감수성 제한 효소는 태아의 과-메틸화된 서열을 무손상 상태로 두는 반면, 이들 유전자의 모체, 비-메틸화된 버전은 절단할 것이다. 이어서, 소화된 혈장 DNA 샘플을, 제한 효소 처리되었지만 무손상인 HLCS 또는 RASSF1A 서열의 분자가 반응 웰 당 평균 대략 0.2 내지 1개 검출될 정도로 희석한다. 2개의 실시간 PCR 시스템을 사용하여 희석된 DNA를 증폭시킬 것이며; 하나는 2개의 프라이머, 및 HLCS 유전자에 특이적인 1개의 태크만 프로브 (서열이 비-메틸화된 경우 제한 효소(들)에 의해 절단될 영역 포함)를 사용하고; RASSF1A 유전자에 관한 다른 하나는 유사하게 2개의 프라이머 및 1개의 태크만 프로브를 사용한다. 후자의 RASSF1A 프라이머/프로브 세트와 관련하여, 일례가 문헌 [Chan et al 2006, Clin Chem 52, 2211-2218]에 기재

되어 있다. HLCS 및 RASSF1A 표적에 관한 태크만 프로브는 상이한 형광 리포터, 예컨대 각각 FAM 및 VIC를 가질 것이다. 이어서, 384-웰 플레이트를 사용하여 디지탈 PCR 실험을 수행한다. HLCS에 대해서만 양성으로 스코어링된 웰의 수 및 RASSF1A에 대해서만 양성으로 스코어링된 웰의 수를 계수할 것이며, 이들 계수의 비율을 구할 것이다. HLCS:RASSF1A 비율은 정상 정배수체 태아를 가진 임신 여성과 비교하여 삼염색체 21 태아를 가진 임신 여성으로부터 얻은 모체 혈장에 대해 보다 높을 것으로 예상될 것이다. 과다표출의 정도는 디지탈 PCR 수행시 웰 당 평균 참조 주형 농도에 따라 달라질 것이다.

[0280] 이들 결과를 스코어링 하는 다른 방법, 예를 들어 RASSF1A에 대해 동시에 양성인 것에 관계 없이 HLCS에 대해 양성인 웰의 수를 계수하고, 반대로 HLCS에 대해 동시에 양성인 것에 관계 없이 RASSF1A에 대해 양성인 웰의 수를 계수하는 방법이 가능할 것이다. 또한, 비율을 계산하는 대신 HLCS 및 RASSF1A 계수의 총 수 또는 이들의 차이를 이용하여 태아의 삼염색체 21 상태를 나타낼 수 있었다.

[0281] 플레이트에서 디지탈 PCR을 수행하는 것 외에도, 디지탈 PCR의 다양한 변형, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 폴로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법 등을 이용할 수 있다는 것이 당업자에게 명백할 것이다. 디지탈 PCR의 이러한 변형은 단지 예로서 제시되며, 제한적이지 않다.

[0282] 실시간 PCR 외에도, 질량 분광법과 같은 방법을 이용하여 디지탈 PCR 결과를 스코어링할 수 있다는 것이 당업자에게 명백할 것이다.

[0283] HLCS 및 RASSF1A의 태아 및 모체 버전을 식별하기 위해 메틸화-감수성 제한 효소를 사용하는 것 외에도, 메틸화 상태를 확인하기 위한 다른 방법, 예를 들어 바이술파이트 변형법, 메틸화-특이적 PCR, 메틸화된 시토신에 대한 항체를 사용하는 면역침강법, 질량 분광법 등을 적용할 수도 있음이 당업자에게 명백할 것이다.

[0284] 또한, 본 특허 출원의 상기 예 및 기타 예에서 설명된 접근법을 모체 뇨, 양수, 자궁경관 세척액, 융모막 융모, 모체 타액 등을 비롯한 태아 DNA가 발견될 수 있는 기타 체액에서도 사용할 수 있음이 당업자에게 명백할 것이다.

[0286] VII. 에멀젼 PCR 및 기타 전략을 이용한 대규모 병렬 게놈 서열분석

[0287] 여기서, 본 발명자들은, 핵산 분자의 디지탈 판독이 모체 혈장 중 태아 염색체의 이수성, 예를 들어 삼염색체 21을 검출하기 위해 사용될 수 있기 때문에 핵산 분자의 디지탈 판독의 또다른 예를 기재할 것이다. 태아 염색체 이수성은 염색체 또는 염색체 영역의 비정상적 용량(들)으로부터 기인한다. 비침습성 시험법이 감수성 및 특이성이 높기 때문에 오진을 최소화하기에 바람직하다. 그러나, 태아 DNA는 낮은 절대 농도로 존재하고, 모체 혈장 및 혈청 중의 모든 DNA 서열에 대해 소량의 부분으로 나타난다. 따라서, 디지탈 PCR 샘플링 표적화에 특이적인 유전자좌의 갯수는 동일한 생물학적 표본 내에서 무한히 증가될 수 없다. 따라서, 많은 세트의 특이적 표적 유전자좌를 분석함으로써, 디지탈 PCR 샘플링 수행 횟수를 증가시키지 않으면서도 표본으로부터 얻을 수 있는 데이타의 양을 증가시킬 수 있다.

[0288] 따라서, 본 실시양태는 모체 배경 핵산을 함유한 생물학적 샘플 중에 소량의 집단으로서 존재하는 제한된 양의 태아 핵산으로부터 추론될 수 있는 유전 정보의 양을 최대화함으로써 태아 염색체의 이수성을 비-침습적으로 검출한다. 한 측면에서, 얻어진 유전 정보의 양은 정확한 진단을 위해서는 충분하지만, 요구되는 생물학적 샘플의 투입량 및 비용을 포괄할 정도로 지나치게 과량인 것은 아니다.

[0289] 454 플랫폼 (로슈) (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]), 일루미나 게놈 분석기 (또는 솔렉사 플랫폼) 또는 솔리드 시스템 (어플라이드 바이오시스템즈) 또는 헬리코스 트루 싱글 몰레큘 DNA 서열분석 기술 (문헌 [Harris TD et al. 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술, 및 나노기공 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53: 1996-2001]) 상에서 달성가능한 것과 같은 대규모 병렬 서열분석은 고도로 다중화된 표본으로부터 단리된 많은 핵산 분자를 병렬 방식으로 서열분석한다 (문헌 [Dear Brief Funct Genomic Proteomic 2003; 1: 397-416]). 이들 각각의 플랫폼은, 클로닝에 의해 확장된 또는 비-증폭된 핵산 절편의 단일 분자를 서열분석한다.

[0290] 대다수의 서열분석 판독이, 수십만 내지 수백만 또는 더 가능하게는 수백만 또는 수백억 정도로, 각각의 진행에서 각각의 샘플로부터 생성되기 때문에, 얻어진 서열분석된 판독은 원래 표본 중 핵산 종 혼합물의 대표적인 프로파일을 형성한다. 예를 들어, 서열분석된 판독의 일배체형, 전사체 및 메틸화 프로파일은 원래 표본의 그것과 유사하다 (문헌 [Brenner et al Nat Biotech 2000; 18: 630-634]; [Taylor et al Cancer Res 2007; 67: 8511-8518]). 각각의 표본으로부터 서열을 대량 샘플링하기 때문에, 핵산 풀(pool)의 서열분석으로부터 몇배의

적용범위 또는 고 중복성으로 생성된 것과 같은 동일한 서열의 수는 또한 원래 샘플 중 특정 핵산 종의 수 또는 유전자위 수에 대한 좋은 정량적 표출이다.

[0291] 한 실시양태에서, 랜덤 서열분석은 임신 여성의 혈장 중에 존재하는 DNA 절편에 대해 수행되고, 본래 태아 또는 모계로부터 유래되는 게놈 서열을 얻는다. 랜덤 서열분석은 생물학적 샘플에 존재하는 핵산 분자의 랜덤 부분을 샘플링 (서열분석)하는 것을 포함한다. 서열분석이 랜덤하게 이루어지기 때문에, 핵산 분자 (및 따라서 게놈)의 상이한 하위세트 (분획)를 각각의 분석에서 서열분석할 수 있다. 본 실시양태는 이러한 하위세트가 샘플마다 및 심지어 동일한 샘플을 사용하여 이루어질 수 있는 분석마다 다양한 경우라 해도 작용할 것이다. 분획의 예는 게놈의 약 0.1%, 0.5% 또는 1%이다. 다른 실시양태에서, 분획은 이들 값 중 적어도 하나이다.

[0292] 이어서, 바이오인포메틱스 절차를 이용하여 이들 DNA 서열 각각을 인간 게놈에 위치시킬 수 있다. 상기 서열의 비율은 이들이 인간 게놈의 반복 영역 또는 개인간 변이, 예를 들어 카피수 변이(copy number variation)에 적용되는 영역에 존재하기 때문에 추후 분석으로부터 감산될 수 있다. 이에 따라, 관심 염색체 및 하나 이상의 다른 염색체의 양을 측정할 수 있다.

[0293] 한 실시양태에서, 염색체 이수성, 예를 들어 염색체 21 또는 염색체 18 또는 염색체 13에 잠재적으로 포함되는 염색체의 파라미터 (예를 들어, 분획 표출)는 예를 들어 바이오인포메틱스 절차의 결과로부터 계산될 수 있다. 분획 표출은 모든 서열 (예를 들어, 모든 염색체의 일부 측정치) 또는 특정 하위세트의 염색체 (예를 들어, 시험된 것과는 다른 염색체)의 양을 기준으로 얻어질 수 있다.

[0294] 이후, 상기 분획 표출은 정상 태아 (즉, 정배수체)를 포함하는 임신에서 확립된 참조 범위와 비교한다. 일부 변형된 절차에서, 참조 범위를 특정 모계 혈장 샘플 중 태아 DNA의 분획 농도 (f)에 따라 조정하는 것이 가능하다. f 값은 태아가 남성인 경우 Y 염색체에 대해 맵핑가능한 서열을 이용하여 서열분석 데이타세트로부터 측정될 수 있다. f 값은 또한 별도 분석에서, 예를 들어 태아 후성 마커를 사용하여 (문헌 [Chan KCA et al 2006 Clin Chem 52, 2211-8]) 또는 단일 뉴클레오티드 다형성 분석으로부터 측정될 수 있다.

[0295] 한 측면에서, 표본 중의 핵산 풀을 100% 미만의 게놈 적용범위에서 서열분석하고, 포획된 핵산 분자의 비율 중에서, 대부분의 각각의 핵산 종만을 일단 서열분석하는 경우에, 특정 유전자위 또는 염색체의 용량 불균형이 또한 정량적으로 측정될 수 있다. 다시 말하면, 유전자위 또는 염색체의 용량 불균형은 표본의 모든 맵핑가능한 서열분석된 태그 중 상기 유전자위의 표출 백분율로부터 추론된다.

[0296] 대규모 병렬 게놈 서열분석 접근법에 대한 한 측면에서, 모든 염색체로부터의 대표적인 데이타는 동시에 생성될 수 있다. 특정 절편의 기원은 해당 시간 전에 선택되지 않는다. 서열분석은 랜덤하게 이루어지고, 이어서 데이타베이스 조사를 수행하여 상기 특정 절편이 어디로부터 유래되었는지 알 수 있다. 이것은 염색체 21로부터의 특정 절편 및 염색체 1로부터의 또다른 절편을 증폭시키는 상황과 대비된다.

[0297] 한 예에서, 이러한 서열의 부분은 상기 설명 예에서 염색체 21과 같은 이수성에 포함되는 염색체로부터 유래할 것이다. 이러한 서열분석 실행으로부터 다른 서열이 다른 염색체로부터 유래될 것이다. 다른 염색체와 비교되는 염색체 21의 상대적 크기를 고려함으로써, 참조 범위 내에서, 상기 서열분석 실행으로부터의 염색체 21-특이적 서열의 정규화 도수를 얻을 수 있다. 태아가 삼염색체 21을 갖는 경우, 서열분석 실행으로부터의 염색체 21-유래된 서열의 정규화 도수는 증가할 것이고, 따라서 삼염색체 21이 검출된다. 정규화 도수의 변화 정도는 분석된 샘플 중 태아 핵산의 분획 농도에 따라 좌우될 것이다.

[0298] 한 실시양태에서, 본 발명자들은 인간 게놈 DNA 및 인간 혈장 DNA 샘플의 단일-말단 서열분석을 위해 일루미나 게놈 분석기를 사용하였다. 일루미나 게놈 분석기는 유동 셀로 명명된 고체 표면 상에서 포획된, 클로닝에 의해 확장된 단일 DNA 분자를 서열분석한다. 각각의 유동 셀은 8개의 개별 표본 또는 표본 풀의 서열분석을 위한 8개 레인을 포함한다. 각각의 레인은 오직 인간 게놈 서열의 30억개의 염기쌍의 분획인 약 200Mb의 서열을 생성할 수 있다. 각각의 게놈 DNA 또는 혈장 DNA 샘플은 유동 셀의 한 레인을 이용하여 서열분석하였다. 생성된 짧은 서열 태그를 인간 참조 게놈에 대해 정렬시키고, 염색체 기원을 기록하였다. 각각의 염색체에 대해 정렬된 서열분석된 개별 태그의 총 갯수를 인간 참조 게놈 또는 비질환 대표 표본으로부터 예상되는 바와 같은 각각의 염색체의 상대적인 크기와 비교하여 표로 만들었다. 이렇게 하여, 염색체 획득 또는 소실을 확인하였다.

[0299] 기재된 접근법은 현재 기재된 유전자/염색체 용량 전략의 하나의 예시일 뿐이다. 별법으로, 페어링 말단의 서열분석을 수행할 수 있다. 문헌 [Campbell et al, Nat Genet 2008; 40: 722-729]에 기재된 바와 같이 참조 게놈에서 예상되는 것으로부터 서열분석된 절편의 길이를 비교하는 대신, 정렬된 서열분석된 태그의 갯수를 카운팅하고, 염색체 위치에 따라 분류하였다. 염색체 영역 또는 전체 염색체의 획득 또는 소실은 태그 카운트를 참

조 게놈에서의 예상되는 염색체 크기 또는 비질환 대표 표본의 예상되는 염색체 크기와 비교함으로써 측정하였다.

[0300] 또다른 실시양태에서, 실행 시 서열분석되는 핵산 풀의 분획은 서열분석 전에 추가로 하위 선택된다. 예를 들어, 올리고뉴클레오티드 어레이와 같이 혼성화를 기초로 하는 기술을 이용하여, 특정 염색체, 예를 들어 잠재적 이수체 염색체 및 시험된 이수성에 포함되지 않는 다른 염색체(들)로부터의 핵산 서열을 먼저 하위 선택할 수 있다. 또다른 예는 샘플 풀로부터의 핵산 서열의 특정 하위-집단이 서열분석 이전에 하위 선택되거나 또는 풍부화된다는 것이다. 예를 들어, 모체 혈장 중 태아 DNA 분자가 모체 배경 DNA 분자보다 더 짧은 절편으로 구성되어 있는 것으로 보고되어 있다 (문헌 [Chan et al Clin Chem 2004; 50: 88-92]). 따라서, 분자 크기에 따라 샘플 내 핵산 서열을 분획화하기 위해 당업자들에게 공지된 하나 이상의 방법, 예를 들어 겔 전기영동 또는 크기 배제 컬럼, 또는 미세유체-기반 접근법에 의한 방법을 사용할 수 있다. 그러나, 별법으로, 모체 혈장 중의 무-세포 태아 DNA 분석 예에서는, 태아 핵산 부분이 모체 배경을 억제하는 방법에 의해, 예컨대 포름알데히드의 첨가에 의해 풍부화될 수 있다 (문헌 [Dhallan et al JAMA 2004; 291 : 1114-9]).

[0301] 로슈 454 플랫폼, 어플라이드 바이오시스템즈 솔리드 플랫폼, 헬리코스 트루 싱글 몰레큘 DNA 서열분석 기술, 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술 및 나노기공 서열분석에 의해서와 같은 다른 단일 분자 서열분석 전략이 본 출원에서 유사하게 사용될 수 있다.

[0302] 결과 및 (예를 들어, 서열분석 및 파라미터 계산을 위한) 추가 논의의 예는 공동 출원된 출원 "게놈 서열분석을 사용한 태아 염색체 이수성 진단" (대리인 관리 번호 016285-005220US)에서 발견될 수 있고, 이것은 본원에 참고로 포함된다. 분별값을 측정하기 위해 본원에 기재된 방법은, 예를 들어 본 섹션에 기재된 바와 같이, 반응물을 서열분석하는 경우 실행될 수 있음을 주목한다.

[0303] 모체 혈장 중 태아 DNA의 분획 농도의 측정은 또한 서열분석 진행과 별도로 수행될 수 있다. 예를 들어, Y 염색체 DNA 농도는 실시간 PCR, 미세유체 PCR 또는 질량 분석기를 사용하여 미리 측정할 수 있다. 사실, 태아 DNA 농도는 여성 태아에 적용할 수 있는 Y 염색체 이외의 유전자좌를 사용하여 측정될 수 있다. 예를 들어, 챈 등(Chan et al)은 태아-유래된 메틸화된 RASSF1A 서열이 모체 유래의 비메틸화된 RASSF1A 서열을 배경으로 하여 임신 여성의 혈장에서 검출됨을 밝혀냈다 (문헌 [Chan et al, Clin Chem 2006;52:2211-8]). 따라서, 태아 DNA의 분획 농도는 메틸화된 RASSF1A 서열의 양을 전체 RASSF1A (메틸화 및 비메틸화 RASSF1A) 서열의 양으로 나눔으로써 측정될 수 있다.

[0304] 혈액 응고 동안 모체 혈액 세포로부터 DNA가 방출되기 때문에, 본 발명을 실행하기 위해서는 모체 혈장이 모체 혈청보다 바람직할 것으로 예상된다. 따라서, 혈청을 사용하는 경우, 태아 DNA의 분획 농도가 모체 혈청에서보다 모체 혈장에서 더 낮을 것으로 예상된다. 다르게 말하면, 모체 혈청을 사용하는 경우, 동일한 임신 여성으로부터 동시에 얻어지는 혈장 샘플과 비교했을 때, 진단하고자 하는 태아 염색체 이수성에 대해 더 많은 서열이 생성될 필요가 있을 것으로 예상된다.

[0305] 그러나, 태아 DNA의 분획 농도를 측정하기 위한 또다른 대안 방법은 임신 여성 및 태아 사이의 다형성 차이의 정량화를 통하는 것일 것이다 (문헌 [Dhallan R, et al. 2007 Lancet, 369, 474-481]). 이러한 방법의 예는 임신 여성이 동형접합성이고 태아가 이형접합성인 다형성 부위를 표적으로 할 것이다. 태아-특이적 대립유전자의 양을 공통 대립유전자의 양과 비교하여 태아 DNA의 분획 농도를 측정할 수 있다.

[0306] 비교 게놈 혼성화, 마이크로어레이 비교 게놈 혼성화, 정량적 실시간 중합효소 연쇄 반응 (하나 이상의 특이적 서열(들)을 검출하고 정량화함)을 비롯한, 염색체 이상을 검출하기 위한 기존의 기술과 달리, 대규모 병렬 서열분석은 미리 측정되거나 또는 미리 한정된 세트의 DNA 서열의 검출 또는 분석에 따라 좌우되지 않는다. 표본 풀로부터의 DNA 분자의 랜덤한 대표 분획을 서열분석한다. 다양한 염색체 영역에 대해 정렬된 상이한 서열 태그의 갯수는 종양성 DNA를 함유하거나 또는 함유하지 않은 표본 사이에서 비교한다. 염색체 이상은 표본에서 임의의 소정 염색체 영역에 대해 정렬된 서열의 갯수 (또는 백분율)의 차이로 나타낼 것이다.

[0307] 또다른 예에서, 혈장 무-세포 DNA에 대한 서열분석 기술을 이용하여 특정 암의 검출을 위해 혈장 DNA에서의 염색체 이상을 검출할 수 있다. 상이한 암은 한 세트의 전형적인 염색체 이상을 갖는다. 다중 염색체 영역의 변화 (증폭 및 결실)가 사용될 수 있다. 따라서, 증폭된 영역에 대해 정렬된 서열의 비율이 증가하고, 감소된 영역에 대해 정렬된 서열의 비율이 감소될 것이다. 염색체 당 표출 백분율은 전체 게놈과 관련하여 임의의 소정 염색체의 게놈 표출의 백분율로서 표현된 참조 게놈 중 각각의 상응하는 염색체에 대한 크기와 비교할 수 있다. 직접 비교 또는 참조 염색체에 대한 비교가 또한 사용될 수 있다.

[0309] VIII. 돌연변이 검출

[0310] 모체 혈장 중의 태아 DNA는 소량의 집단으로 존재하며, 모체 혈장 DNA 중 평균 3% 내지 6%가 태아에 의한다. 이러한 이유로 인해서, 본 분야에서 행해진 기존 연구의 대부분은 태아가 부계로부터 물려받고 모체 혈장 중의 대다수의 모체 DNA 배경과 구별되는 DNA 표적을 검출하는데 집중되었다. 이와 같이 기존에 검출된 표적의 예는 Y 염색체상의 SRY 유전자 [Lo YMD et al. 1998 Am J Hum Genet, 62, 768-775] 및 모계가 RhD-음성인 경우의 RHD 유전자 [Lo YMD et al. 1998 N Engl J Med, 339, 1734-1738]를 포함한다.

[0311] 태아 돌연변이 검출을 위해서 모체 혈장을 사용하는 기존의 전략은 부계가 보인자(carrier)인 상염색체 우성 상태로 제한되며, 부계 및 모계가 상이한 돌연변이를 갖는 경우에는 직접적인 돌연변이 검출을 수행하거나 또는 연관 분석 [Ding C. et al 2004 Proc Natl Acad Sci USA 101, 10762-10767]을 실시하여 상염색체 열성 질환은 제외한다. 이러한 기존 전략에는 유의한 한계가 있었다. 예를 들어, 남성 파트너와 여성 파트너가 동일한 돌연변이를 갖는 쌍에서는 모체 혈장 중에서의 직접적인 돌연변이 검출로는 의미있는 태아 진단을 실시할 수 없다.

[0312] 이러한 시나리오는 도 23에 예시되어 있다. 상기 시나리오에서는, 3종의 가능한 태아 유전자형, NN, NM 및 MM이 존재하는데, 여기서 N은 정상적인 대립유전자를 나타내고, M은 돌연변이 대립유전자를 나타낸다. 돌연변이 대립유전자의 예는 낭포성 섬유증, 베타-지중해빈혈, 알파-지중해빈혈, 겸상적혈구 빈혈, 척수 근위축증, 선천성 부신 과형성 등을 일으키는 것을 포함한다. 이러한 장애의 다른 예는 OMIM (Online Mendelian Inheritance in Man) [www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar]에서 찾을 수 있다. 모체 혈장에서, 대부분의 DNA는 모계로부터 유래되며 NM일 것이다. 상기한 3종의 태아 유전자형 중 어느 것에서도, 모체 혈장 중에서의 독특한 검출을 허용하는 임의의 독특한 태아 대립유전자는 존재하지 않는다. 따라서, 통상적인 전략은 여기에 적용될 수 없다.

[0313] 본원에 기재한 실시양태는 이러한 시나리오를 취급할 수 있다. 모계 및 태아가 둘다 NM인 시나리오에서는 N 대립유전자 및 M 대립유전자가 대립유전자 균형을 이루고 있을 것이다. 그러나, 모계가 NM이고 태아가 NN인 경우에는 모체 혈장 중에서 N 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 한편, 모계가 NM이고 태아가 MM인 경우에는 모체 혈장 중에서 M 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 따라서, 태아 돌연변이 검출을 위한 귀무가설에서는 태아가 NM 유전자형인 경우에 대립유전자 불균형이 존재하지 않는다고 가정한다. 대립가설에서는, 대립유전자 불균형이 존재하고 N 또는 M 대립유전자 중 어느 것이 과다표출되었는지의 여부에 따라 태아 유전자형은 NN 또는 MM일 수 있다고 가정한다.

[0314] 대립유전자 불균형의 존재 또는 부재는 본원에 기재한 실시양태를 사용한 디지탈 PCR로 결정될 수 있다. 첫번째 시나리오에서, 특정 부피의 모체 혈장은 모계로부터의 50개 세포 및 태아로부터의 50개 세포의 100개 세포로부터 방출된 DNA를 함유한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 50%이다. 모계가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 모계에 의한다. 태아가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 태아에 의한다. 따라서, N 대립유전자와 M 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 NN 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 N 대립유전자가 존재할 것이다. 따라서, 총 150개의 N 대립유전자와 50개의 M 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 N이 M에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다.

[0315] 반대의 상황에서, 태아가 MM 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 M 대립유전자가 존재할 것이다. 따라서, 150개의 M 대립유전자와 50개의 N 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 M이 N에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 대립유전자 불균형은 디지탈 PCR로 결정될 수 있다. 양성 웰의 수가 더 적은 대립유전자가 참조 주형으로 간주된다. 디지탈 RNA-SNP 및 디지탈 RCD 분석과 유사하게, 디지탈 PCR 실험에서 대립유전자의 실제 분포는 푸아송 확률 밀도 함수에 의해 좌우된다. 따라서, 본 시나리오에서는 대립유전자 불균형의 이론적인 정도가 3:1이지만, 대립유전자 불균형의 예상 정도는 디지탈 PCR 분석 동안에 웰 당 평균 주형 농도에 따라 달라질 것이다. 따라서, SPRT 분석에서와 같은, 웰 당 평균 참조 주형 농도 (m_r)에 적절한 해석 분별값이 사례 분류에 사용될 필요가 있다.

[0316] 추가로, 결정할 필요가 있는 대립유전자 불균형의 정도는 태아 DNA의 분획 농도에 따라 달라진다. 상기한 예와는 반대로, 특정 부피의 모체 혈장이 모계로부터의 90개 세포 및 태아로부터의 10개 세포의 100개 세포로부터 방출된 DNA를 함유하는 경우를 가정한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 10%이

다. 모계가 유전자형 NM인 경우에는 90개의 N 대립유전자 및 90개의 M 대립유전자가 모계에 의한다. 태아가 유전자형 NM인 경우에는 10개의 N 대립유전자 및 10개의 M 대립유전자가 태아에 의한다. 따라서, N 대립유전자와 M 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 NN 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 N 대립유전자가 존재할 것이다. 따라서, 총 110개의 N 대립유전자와 90개의 M 대립유전자가 존재할 것이다.

[0317] 즉, N과 M 사이에는 N이 M에 대하여 과다표출된 대립유전자 불균형이 존재할 것이다. 반대의 상황에서, 태아가 MM 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 M 대립유전자가 존재할 것이다. 따라서, 110개의 M 대립유전자와 90개의 N 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 M이 과다표출된 대립유전자 불균형이 존재할 것이다. 태아 DNA의 분획 농도가 10%인 경우에 대립유전자 불균형의 이론적인 정도는 110:90일 것이고, 이는 상기 예에서 나타낸 바와 같이 50% 태아 DNA가 존재하는 경우의 3:1 비율과 상이하다. 따라서, SPRT 분석에서와 같은, 태아 DNA의 분획 농도에 적절한 해석 분별값이 사례 분류에 사용될 필요가 있다.

[0318] 따라서, 혈장 DNA를 추출할 것이다. 혈장 샘플 중 모체 및 태아 DNA의 양은 예를 들어 기존에 수립된 실시간 PCR 검정 [Lo, et al. 1998 Am J Hum Genet 62, 768-775] 또는 당업자에게 공지된 다른 유형의 정량화 수단, 예를 들어 SNP 마커 [Dhallan R et al. 2007 Lancet, 369, 474-481] 및 태아 후성 마커 [Chan KCA et al. 2006 Clin Chem, 52, 2211-2218]를 이용하여 정량한다. 태아 DNA의 백분율(%)을 계산한다. 이후, 디지탈 PCR 분석 동안에 각 반응 웰이 평균 1개의 주형 분자 (N 대립유전자일 수도 있고, 또는 M 대립유전자일 수도 있음)를 함유하도록 정량된 혈장 DNA 샘플 (예를 들어 희석하거나 농축된 것)을 제조한다. 한쌍의 프라이머, 및 2개의 태크만 프로브 (1개는 N 대립유전자에 특이적이고, 다른 하나는 M 대립유전자에 특이적임)를 사용하여 디지탈 PCR 분석을 수행한다. 오직 M에만 양성인 웰의 수 및 오직 N에만 양성인 웰의 수를 계수한다. 이들 웰의 비율을 사용하여 대립유전자 불균형의 증거가 있는지 여부를 결정한다. 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다. 이러한 분석의 한 변형법에서는 M에만 양성이거나 M 및 N에 양성인 웰의 수를 계수하는 것도 가능하고, N에만 양성이거나 M 및 N에 양성인 웰의 수를 계수하는 것도 가능하며, 이들 계수치의 비율을 도출하는 것도 가능하다. 한번 더, 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다.

[0319] 디지탈 RMD (상대적 돌연변이 양(Relative Mutant Dosage))라고 불리는 태아 유전자 돌연변이의 양 결정을 여성/남성 (XX/XY) DNA 혼합물을 사용하여 검증하였다. 남성 및 여성으로부터의 혈액 세포 DNA를 남성 DNA와 각각 혼합하여, 도 24a에 나타낸 바와 같이, XY의 배경에 XX 또는 XY 유전자형을 각각 25% 및 50%의 분획 농도로 갖는 샘플을 수득한다.

[0320] 추가로, 혈액 세포 샘플을 또한 12명의 남성 및 12명의 여성 대상체로부터 수득하였다. 여성 혈액 세포 DNA (유전자형 XX)를 3배 과량의 남성 혈액 세포 DNA (유전자형 XY)와 각각 혼합하여, XY 유전자형을 갖는 DNA 75%의 배경에 XX 유전자형을 갖는 DNA 25%를 함유하는 12개의 DNA 혼합물을 생성하였고, 이 결과는 도 24b에 나타냈다.

[0321] SPRT 분석의 목적은 배경 DNA에 존재하는 소량의 유전자형을 결정하는 것이었다. 75% XY DNA의 배경에 25% XX DNA를 함유하는 DNA 혼합물에서, 소량의 대립유전자는 DNA 75%에서 유래된 Y이다. 샘플 중의 DNA 25%가 XX 유전자형이기 때문에, 샘플 중에 총 200개의 DNA 분자가 존재하는 경우에는 150개의 분자가 XY 개체에서 유래된 것이다. 따라서, Y 대립유전자의 수는 75개일 것으로 예상된다. 또한, 남성 DNA (유전자형 XY)의 비율에 의한 X 대립유전자의 수 역시 75개이다. 여성 (유전자형 XX)에 의한 X 대립유전자의 수는 50 (= 2×25)이다. 따라서, X/Y 비율은 125/75 = (1 + 25%)/(1 - 25%) = 5/3이다.

[0322] 본 연구의 두번째 부분에서, 혈액 세포 샘플을 베타-글로빈 유전자, 즉, 헤모글로빈, 베타 (HBB) 유전자에 HbE (G→A) 및 CD41/42 (CTTT/-) 돌연변이를 갖는 남성 및 여성 대상체로부터 수득하였다. 모든 가능한 유전자형 (MM, MN 또는 NN)을 갖는 남성 태아를 보유하는 이형접합 모계 (MN (여기서, M = 돌연변이, N = 야생형))로부터 수득한 모체 혈장 샘플을 모방하기 위해서, 야생형 대립유전자에 대해 동형접합 (NN)이거나 2종의 돌연변이 중 하나에 대해 이형접합 (MN)인 남성으로부터의 혈액 세포 DNA를 동일 돌연변이에 대하여 이형접합 (MN)인 여성으로부터 수집한 혈액 세포 DNA 샘플과 각각 혼합하였다. 이로써, 다양한 분획 농도의 남성/돌연변이 DNA를 갖는 DNA 혼합물이 생성되었다. 또한, CD41/42 결실에 대해 동형접합 (MM)인 여성으로부터의 혈액 세포 DNA 샘플도 사용하여 DNA 혼합물을 제조하였다. SPRT 분류를 위한 남성 비율을 정확하게 하기 위해서, 각 DNA 혼합물 중 남성 DNA의 분획 농도를 ZFY/X 검정으로 결정하였다.

[0323] 디지탈 ZFY/X 검정을 이용하여 SPRT를 검증하고 또한 DNA 혼합물 중 남성 DNA의 분획 농도를 결정하였다. 염색체 X에서의 아연 핑거(Zinc Finger) 단백질 서열 (ZFX) 및 염색체 Y에서의 아연 핑거 단백질 서열 (ZFY)의 양을 디지탈 PCR 분석으로 결정하였다. 전방향 프라이머 5'-CAAGTGCTGGACTCAGATGTAACTG-3' (서열 47) 및 역방향 프라이머 5'-TGAAGTAATGTCAGAAGCTAAAACATCA-3' (서열 48)를 사용하여 ZFX 및 ZFY 유전자좌의 87-bp 앰플리콘을 우선 동시 증폭시켰다. 염색체 X 및 Y 파라로그를 구별하는 2종의 염색체-특이적 태크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)TCTTTAGCACATTGCA(MGBNFQ)-3' (서열 49) 및 5'-(FAM)TCTTTACCACACTGCAC(MGBNFQ)-3' (서열 50)였다.

[0324] DNA 혼합물 중의 돌연변이 양을 돌연변이 대립유전자에 대한 정상 대립유전자의 디지탈 PCR 분석으로 결정하였다. HbE 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 앰플리콘을 전방향 프라이머 5'-GGGCAAGGTGAACGTGGAT-3' (서열 51) 및 역방향 프라이머 5'-CTATTGGTCTCCTTAAACCTGTCTTGTAA-3' (서열 52)로 우선 동시 증폭시켰다. 정상 (G) 및 돌연변이 (A) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 태크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)TTGGTGGTGAGGCC(MGBNFQ)-3' (서열 53) 및 5'-(FAM)TTGGTGGTAAGGCC(MGBNFQ)-3' (서열 54)였다. HbE 돌연변이에 대한 결과를 도 25에 나타냈다.

[0325] CD41/42 결실 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 및 83-bp 앰플리콘을 전방향 프라이머 5'-TTTTCCCACCCTTAGGCTGC-3' (서열 55) 및 역방향 프라이머 5'-ACAGCATCAGGAGTGGACAGATC-3' (서열 56) 각각을 사용하여 우선 동시 증폭시켰다. 정상 (결실 없음) 및 돌연변이 (결실 있음) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 태크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)CAGAGGTTCTTTGAGTCCT(MGBNFQ)-3' (서열 57) 및 5'-(FAM)AGAGGTTGAGTCCTT(MGBNFQ)-3' (서열 58)였다. HbE 돌연변이에 대한 결과를 도 26a 및 도 26b에 나타냈다.

[0326] 이러한 실험은 12.765 디지탈 어레이즈 (플루이다임)를 사용하는 바이오마크™ 시스템 (플루이다임)에서 수행하였다. 1개의 패널에 대한 반응은 2× 태크만 유니버설 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 10 ㎕로 설정하였다. CD41/42 및 ZFY/X 검정의 경우, 각 반응물은 1× 태크만 유니버설 PCR 마스타 믹스, 900 nM씩의 각 프라이머, 125 nM씩의 각 프로브 및 1 ng/㎕의 DNA 혼합물 3.5 ㎕를 함유하였다. HbE 검정의 경우, 정상 (G) 및 돌연변이 (A) 대립유전자를 표적으로 하는, 각각 250 nM 및 125 nM의 프로브를 첨가하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)로 디지탈 어레이에 로딩하였다. 반응은 신호 검출을 위해 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시한 후에 95℃에서 10분 동안 진행하고, 95℃, 15초 및 57℃ (ZFY/X 및 CD41/42의 경우) 또는 56℃ (HbE의 경우), 1분의 50회 주기를 수행하였다. 각 경우마다 적어도 1개의 반응 패널을 사용하였고, 결정이 이루어질 수 있을 때까지 미분류 상태로 있는 샘플을 위한 여분의 패널로부터 데이타를 합하였다.

[0327] 또한, 디지탈 PCR이 당업자에게 공지된 방법, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 폴로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법을 이용하여 수행될 수 있다는 것이 당업자에게 명백할 것이다.

[0329] IX. 암을 이용한 예

[0330] 한 실시양태에서, 본 발명을 수행하여 암성 종양에서 발생할 수 있는 것과 같이 대립유전자 비율이 벗어나거나 벗어나지 않은 샘플을 분류할 수 있다. 한 측면에서, 각 경우에 오직 A 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 오직 G 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 및 2종의 대립유전자 둘다에 대해서 양성 신호를 갖는 웰의 수를 디지탈 PCR로 결정하였다. 참조 대립유전자는 양성 웰의 수가 더 적은 대립유전자로 정의된다 (2종의 대립유전자가 둘다 동일한 수의 양성 웰을 갖는 시나리오가 있다고 한다면, 둘 중 어느 것이라도 참조 대립유전자로 사용될 수 있음). 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여, 다른 대립유전자가 양성인지 여부와는 관계 없이 푸아송 확률 밀도 함수에 따라 웰 당 참조 대립유전자의 평균 농도 추론치 (m_r)를 계산하였다. 본 연구진은 상기 계산을 예시하는 가정적인 예를 사용한다.

[0331] 96웰 반응에서, 20개 웰은 A 대립유전자 대해 양성이고, 24개 웰은 G 대립유전자에 대해 양성이며, 28개 웰은 이들 2개의 대립유전자 모두에 대해 양성이다. A 대립유전자에 대하여 양성인 웰의 수가 더 적기 때문에, A 대립유전자를 참조 대립유전자로 간주한다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 20 - 28 = 48개이다.

따라서, m_r은 푸아송 분포로 계산될 수 있고, -ln(48/96) = 0.693이다.

[0332] LOH 검출과 관련하여, 귀무가설에서는 1개의 대립유전자 결실이 존재하여 샘플에서 대립유전자 비율이 벗어나지 않았다고 가정한다. 이러한 가정하에서, 2종의 대립유전자에 대한 양성 웰의 수의 예상 비율은 1:1이고, 이에 따라 잠재적으로 과다표출된 대립유전자를 함유하는 정보제공 웰 (오직 1종의 대립유전자에 대해서만 양성인 웰)의 예상 비율은 0.5이다.

[0333] 299LOH 검출과 관련하여, 대립가설에서는 샘플의 세포 중 50%에서 1개의 대립유전자 결실이 존재하여 샘플에서 대립유전자 비율이 벗어나 있다고 가정한다. 과다표출된 대립유전자와 참조 대립유전자 사이의 대립유전자 비율이 2:1이기 때문에, 웰 당 과다표출된 대립유전자의 평균 농도는 참조 대립유전자의 평균 농도의 2배일 것이다. 그러나, 과다표출된 대립유전자에 양성인 웰의 수는 참조 대립유전자에 양성인 웰의 수의 단순히 2배가 아니라 푸아송 분포에 따를 것이다.

[0334] 정보제공 웰은 A 또는 G 대립유전자에 양성이지만 상기 2종의 대립유전자 둘다에는 양성이 아닌 웰로 정의된다. 대립유전자 비율이 벗어난 샘플의 경우에 과다표출된 대립유전자를 함유하는 웰의 수의 예상 비율의 계산치는 표 600에 나타낸 바와 같다. 상기 예에서, 종양 세포의 50%에서 LOH가 존재하는 경우에는 웰 당 G 대립유전자의 평균 농도가 2×0.693 = 1.386일 것이다. LOH가 종양 세포의 50% 초과에 존재하는 경우에는 웰 당 G 대립유전자의 평균 농도가 하기 식에 따를 것이다: $1/[1-(\text{LOH를 갖는 비율})] \times m_r$.

[0335] G 대립유전자에 양성인 웰의 예상 비율은 $1 - e^{-1.386} = 0.75$ (즉, 75% 또는 72개의 웰)일 것이다. A 또는 G 대립유전자에 대한 웰의 양성이 독립적이라고 가정할 때, 0.5 × 0.75 = 0.375의 웰이 A 및 G 대립유전자 둘다에 양성일 것이다. 따라서, 0.5 - 0.375 = 0.125의 웰이 오직 A 대립유전자에만 양성일 것이고, 0.75 - 0.375 = 0.375의 웰이 오직 G 대립유전자에만 양성일 것일 것이다. 따라서, 정보제공 웰의 비율은 0.125 + 0.375 = 0.5일 것이다. G 대립유전자를 갖는 정보제공 웰의 예상 비율은 0.375/0.5 = 0.75이다. 이후, P_r에 대한 이러한 예상 값을 사용하여, 샘플에 대립유전자 비율이 벗어난 경우 (즉, 여기서는 LOH)가 존재하는지의 여부를 결정하기 위한 적절한 SPRT 곡선을 구축할 수 있다.

[0336] 이어서, 디지탈 PCR 분석을 통해 실험적으로 결정된, 비-참조 대립유전자를 갖는 정보제공 웰의 실제 비율 (P_r)을 사용하여, 귀무가설 또는 대립가설 중 어느 것이 받아들여질 지, 또는 더 많은 웰을 사용한 추가의 분석이 필요할 지를 결정하였다. 귀무가설 또는 대립가설을 받아들이기 위한 P_r의 결정 한계값은 임계값 우도 비율 8을 기준으로 하여 계산되는데, 이는 상기 값이 대립유전자 불균형이 존재하는 샘플과 존재하지 않는 샘플을 암 검출과 관련하여 식별하는데 만족스러운 성능을 제공하는 것으로 나타났기 때문이다 ([Zhou, W, et al. (2001) Nat Biotechnol 19, 78-81], [Zhou et al 2002 (상기 문헌)]). 상기 예에서, 정보제공 웰의 수는 20 + 24 = 44이고, 실험적으로 수득된 P_r은 24/44 = 0.5455이다. 결정 한계값은 귀무가설을 받아들이기 위해서는 ≤ 0.5879이고, 대립가설을 받아들이기 위해서는 ≥ 0.6739이다. 따라서, 이러한 예에서의 샘플은 대립유전자 비율이 벗어난 NOT으로 분류된다.

[0337] 결론적으로, 본 연구진은 샘플 중의 서열 불균형을 검출하는 접근법을 알아냈다. 한 실시양태에서, 본 발명은 태아 염색체 이수성의 비-침습적 검출에 사용될 수 있고, 예를 들어 모체 혈장 중 태아 핵산의 분석을 통해 삼염색체 21을 검출할 수 있다. 이러한 접근법은 또한 양수, 융모막 융모 샘플, 모체 뇨, 자궁경관 샘플, 모체 타액 등을 포함하는, 태아 핵산을 함유하는 다른 생물학적 물질에도 적용될 수 있다. 첫째, 본 연구진은 삼염색체 21 태아를 보유하는 여성의 모체 혈장 중 염색체 21에서의 태반-발현된 전사체인 PLAC4 mRNA에서 SNP의 대립유전자 불균형을 결정하기 위한 본 발명의 용도를 입증하였다. 둘째, 본 연구진은 본 발명이 삼염색체 21의 비-침습적 태아 검출을 위해서 RCD (상대적 염색체 양(Relative Chromosome Dosage)) 분석을 통해 비-다형성 기재의 방법으로 사용될 수 있음을 입증하였다. 이러한 디지탈 RCD-기반의 접근법은 태아 DNA를 함유하는 샘플 중 염색체 21의 전체 카피수가 참조 염색체에 비하여 과다표출되었는지 여부를 직접 평가하는 것을 수반한다. 디지탈 RCD는 정교한 장치 없이도 25% 태아 DNA를 함유하는 샘플에서 삼염색체 21을 검출할 수 있게 한다. 본 연구진은 순차 확률비 검증 (SPRT)을 적용하여 디지탈 PCR 데이타를 해석하였다. 컴퓨터 시뮬레이션 분석을 실시하여 질환 분류 알고리즘의 높은 정확성을 확인하였다.

[0338] 추가로, 본 연구진은 상기 접근법이 염색체 이수성 이외의 다른 형태의 핵산 서열 불균형의 결정, 예를 들어 모체 혈장 중 태아 돌연변이의 검출 또는 다형성 검출 및 혈장 중 종양-유래의 핵산 분석에 의한 악성 세포 게놈의 국소적인 획득 및 소실의 검출에 적용될 수 있음을 알아냈다.

[0339] 본 명세서에 기재한 임의의 소프트웨어 성분 또는 기능은 임의의 적합한 컴퓨터 언어, 예를 들어 자바(Java), C++ 또는 펄(Perl) (예를 들어 통상의 기술 또는 객체 지향(object-oriented) 기술을 사용한 것)을 사용한 프로세서에 의해 실행되는 소프트웨어 코드로서 시행될 수 있다. 소프트웨어 코드는 일련의 지시 또는 명령어로서 저장을 위한 컴퓨터 판독가능 매체에 저장될 수 있으며, 적합한 매체는 랜덤 액세스 메모리 (RAM, random access memory), 리드 온리 메모리 (ROM, read only memory), 자성 매체, 예컨대 하드 드라이브 또는 플로피 디스크, 또는 광학 매체, 예컨대 콤팩트 디스크 (CD) 또는 DVD (디지탈 버서틀 디스크(digital versatile disk)), 플래쉬 메모리 등을 포함한다. 컴퓨터 판독가능 매체는 이러한 저장 장치의 임의의 조합일 수 있다.

[0340] 이러한 프로그램은 또한 다양한 프로토콜, 예를 들어 인터넷에 따라 유선, 광학 및/또는 무선 네트워크를 통해 전송되도록 맞춰진 캐리어 신호를 사용하여 암호화되고 전송될 수도 있다. 이와 같이, 본 발명의 실시양태에 따른 컴퓨터 판독가능 매체가 이러한 프로그램으로 암호화된 데이타 신호를 사용하여 생성될 수 있다. 프로그램 코드로 암호화된 컴퓨터 판독가능 매체는 호환가능한 장치와 함께 패키지될 수도 있고, 또는 다른 장치와 별도로 제공 (예를 들어, 인터넷 다운로드에 의함)될 수도 있다. 임의의 이러한 컴퓨터 판독가능 매체는 1개의 컴퓨터 프로그램 제품 (예를 들어 하드 드라이브 또는 전체 컴퓨터 시스템)상에 있거나 그 내부에 있을 수 있고, 시스템 또는 네트워크 내부의 상이한 컴퓨터 프로그램 제품상에 있거나 그 내부에 있을 수 있다. 컴퓨터 시스템은 모니터, 프린터 또는 본원에서 언급한 임의의 결과를 사용자에게 제공하기 위한 다른 적합한 디스플레이를 포함할 수 있다.

[0341] 컴퓨터 시스템의 예는 도 27에 나타냈다. 도 27에 나타낸 하위시스템은 시스템 버스 (2775)를 통해 상호연결된다. 디스플레이 어댑터 (2782)에 커플링된 추가의 하위시스템, 예컨대 프린터 (2774), 키보드 (2778), 고정 디스크 (2779), 모니터 (2776) 등을 나타냈다. I/O 제어기 (2771)에 커플링된 주변기기 및 입력/출력 (I/O) 장치는 직렬 포트 (2777)과 같이 당업계에 공지된 임의의 수의 수단을 통해 컴퓨터 시스템에 연결될 수 있다. 예를 들어, 직렬 포트 (2777) 또는 외부 인터페이스 (2781)을 사용하여 컴퓨터 장치를 광범위한 범위의 네트워크, 예를 들어 인터넷, 마우스 입력 장치 또는 스캐너에 연결할 수 있다. 시스템 버스를 통한 상호연결은 중앙 프로세서 (2773)가 각각의 하위시스템과 소통하고 시스템 메모리 (2772) 또는 고정 디스크 (2779)로부터의 지시의 수행 및 하위시스템들 사이의 정보 교환을 제어하도록 한다. 시스템 메모리 (2772) 및/또는 고정 디스크 (2779)는 컴퓨터 판독가능 매체를 포함할 수 있다.

[0342] 본 발명의 예시적인 실시양태에 관한 상기 기재는 예시와 설명을 위해서 제공된 것이다. 이는 본 발명을 상기한 형태로만 한정하거나 제한하려는 것이 아니며, 상기한 기술에 비추어 많은 변형과 변화가 가능하다. 본 발명의 이론 및 이것의 실제적인 적용을 최적으로 설명하고 그에 따라 당업자가 본 발명을 고려되는 특별한 용도에 맞게 다양하게 변형시켜 다양한 실시양태에서 최적으로 사용할 수 있도록 하는 실시양태를 선택하여 기재하였다.

[0343] 본원에서 언급한 모든 간행물, 특허 및 특허 출원서는 어떠한 목적으로도 그 전문이 본원에 참고로 포함된다.

SEQUENCE LISTING

<110> Lo, Yuk-Ming Dennis
Chiu, Rossa Wai Kwun
Chan, Kwan Chee
See, Benny Chung Ying
Ng, Ellen Cheng
The Chinese University of Hong Kong

<120> Determining Nucleic Acid Sequence
Imbalance

<130> 015236-005188PC

<140> PCT/CN08/001824
<141> 2008-07-23

<150> US 60/961,438
<151> 2007-07-23

<160> 38

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR and digital RNA-DNA analysis
PLAC4 gene-specific reverse transcription primer

<400> 1
agtatatgga accatgttta ggccagg 27

<210> 2
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR PLAC4 primer

<400> 2
tcgctagcgt gtctttaag a 21

<210> 3
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR PLAC4 primer

<400> 3
ctattgcaat acaaaatcac tttct 25

[0344]

<210> 5
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic real-time PCR FLACS fluorescent probe

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (13)...(13)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 4
attggagcaa attc 14

<210> 6
<211> 104
<212> DNA
<213> Artificial Sequence

<220>
<223> single-stranded synthetic DNA oligonucleotide
amplicon

<400> 5
cgccgctccg gtgtcatttc agctattggc gcaaattccc attggcattc aagaaaaagc 60
aactcatttt gtattgcaac accaggagta tcccaaggga ctcg 104

<210> 6
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital RNA-SNP analysis non-intron
spanning forward primer

<400> 6
tttgtattgc aacaccattt gg 22

<210> 7
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-B-specific TaqMan probe targeting
rs1130833 SNP of FLACS

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

[0345]

<220>
<221> modified_base
<222> (15)...(15)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 7
tcatcgtcta actta 15

<210> 8
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting rs8130833 SNP of FLAC4

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (17)...(17)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 8
attcgtdatc taactg 17

<210> 9
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification forward primer for paralogous loci on chromosome 21 and 1

<400> 9
gttgttctgc aaaaaacct cga 23

<210> 10
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification reverse primer for paralogous loci on chromosome 21 and 1

<400> 10
ctagcaaga aatatttat tacaatat 28

3

[0346]

<210> 11
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<400>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (18)...(18)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 11
ccctcccctg aggagtaaa 19

<210> 12
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (20)...(20)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 12
cgtgccctg tactgtgtaa 20

<210> 13
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR PLAC4 gene-specific reverse
transcription primer

<400> 13
agtatatagg accatgttg ggccaga 27

[0347]

<210> 14
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification forward primer (or rs8130833) SNP region of PLAC4 gene

<400> 14
ttgtattgc aatacacatt gg 22

<210> 15
<211> 15
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting PLAC4 rs8130833 SNP

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (15)...(15)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 15
tactgctata aattg 15

<210> 16
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting PLAC4 rs8130833 SNP

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (17)...(17)
<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 16
attagtcatc tcactg 16

§

[0348]

<210> 17
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification serpin peptidase inhibitor clade B (ovalbumin) member 3 (SERPINB3) gene-specific reverse transcription primer

<400> 17
cttcagcctc tcactcaaca 20

<210> 18
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR amplification serpin peptidase inhibitor clade B (ovalbumin) member 3 (SERPINB3) forward primer

<400> 18
ctcagcctcg caatcaatgc 20

<210> 19
<211> 16
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting serpin peptidase inhibitor clade B (ovalbumin) member 3 (SERPINB3) rs6095 SNP

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (16)...(16)
<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 19
cccacgcgaa ttattt 16

<210> 20
<211> 16
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting serpin peptidase inhibitor clade B (ovalbumin) member 3 (SERPINB3) rs6095 SNP

[0349]

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (18)...(18)
<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 20
cctcagtggga ttattt 16

<210> 21
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis cn-amplification
forward primer for 101-bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 21
gttgttctgc aaaaaaattt cga 23

<210> 22
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis cn-amplification
reverse primer for 101-bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 22
ctttgcctaga aatatttaat taccatat 28

<210> 23
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (19)...(19)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

[0350]

<400> 23
tgccttcata atgagtaaa 19

<210> 24
<211> 26
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (25)...(25)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 24
cgtaccctcg tcatgtgtaa 20

<210> 25
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital MCB analysis co-amplification
forward primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 25
gtacagaaac cacaaatgg tcgg 24

<210> 26
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital MCB analysis co-amplification
reverse primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 26
gtccagcctg tgggcct 17

<210> 27
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

[0351]

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by 6-carboxyfluorescein (6FAM)

<220>
<221> modified_base
<222> (14)...(14)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 87
aagaggcgag gcaa 14

<210> 88
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 18 paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (14)...(15)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 88
aagaggacag gcaac 15

<210> 89
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
forward primer for 121-bp amplicon for paralogous
loci on chromosome 21 and X

<400> 89
acgttggatg gtgattatgc aaaaaacctt aga 33

<210> 90
<211> 36
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
reverse primer for 121-bp amplicon for paralogous
loci on chromosome 21 and X

[0352] 9

<400> 30
acgttggatg ctgggcaagg aatacttcat taccatat 38

<210> 31
<211> 21
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
extension primer for 121-bp amplicon for
paralogous loci on chromosome 21 and 1

<400> 31
ctcatcaataa attagtcact c 21

<210> 32
<211> 34
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
forward primer for 148-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 32
acgttggatg gtacagaaac caccaactga tcag 34

<210> 33
<211> 27
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
reverse primer for 148-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 33
acgttggatg gtccaggcta tggaact 27

<210> 34
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
extension primer for 148-bp amplicon for
paralogous loci on chromosome 21 and 18

<400> 34
accaaaaggg gaagagg 17

<210> 35
<211> 35
<212> DNA
<213> Artificial Sequence

[0353]

<210>
<223> synthetic digital PCR analysis co-amplification
forward primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 35
acgttggatg ttgatgaagt ctcatatata attcg 35

<210> 36
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
reverse primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 36
acgttggatg caatcagcct ggtcaqaaat act 33

<210> 37
<211> 37
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
forward primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 37
acgttggatg gaatttaagc tcaatcagcc tgaactg 37

<210> 38
<211> 37
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
reverse primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 38
acgttggatg gtttctcata gttcatagta ggcttat 37

<210> 39
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification
forward primer for 101 bp amplicon for paralogous
loci on chromosome 21 and 2

<400> 39
acgttggatg tcaggtaggg ttctatgcag 30

[0354] 11

<210> 40
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital NGS analysis co-amplification
reverse primer for 182 bp amplicon for paralogous
loci on chromosome 21 and 2

<400> 40
acgttggatg agagcgcctc ctgcctcttg 30

<210> 41
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital NGS analysis co-amplification
forward primer for 182 bp amplicon for paralogous
loci on chromosome 21 and 6

<400> 41
acgttggatg gccggtcttc cgctcgtcgt t 31

<210> 42
<211> 30
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital NGS analysis co-amplification
reverse primer for 182 bp amplicon for paralogous
loci on chromosome 21 and 6

<400> 42
acgttggatg ttttttcgag ccctcctgga 30

<210> 43
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence
mismatches (PSMM) on paralogous loci on chromosomes
21 and 1

<400> 43
gtttcatctc tacttcgtac ccc 23

<210> 44
<211> 16
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence
mismatches (PSMM) on paralogous loci on chromosomes
21 and 7

[0355] 13

<400> 44
tttacgctg tcccaattt 19

<210> 45
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 2

<400> 45
agtcatgca gaagcagaa 19

<210> 46
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 6

<400> 46
tgggcgcgg agcggaattc gatgg 25

<210> 47
<211> 25
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification forward primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 47
caaatgcagg actcagatgt aactg 25

<210> 48
<211> 28
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR analysis co-amplification reverse primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 48
tgaagtaatg tcagaagata aaacatca 28

<210> 49
<211> 18
<212> DNA
<213> Artificial Sequence

[0356]

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome X paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (16)...(16)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 49
tcttttgcac attgca 16

<210> 50
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome Y paralog

<220>
<221> modified_base
<222> (1)...(1)
<223> a modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (17)...(17)
<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 50
tcattacaaa actgcac 17

<210> 51
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR co-amplification forward
primer for 87-bp amplicon of beta-globin
(hemoglobin beta (HBB)) HbE mutation forward
primer

<400> 51
gggcaaggtg aacgtggat 19

<210> 52
<211> 29
<212> DNA
<213> Artificial Sequence

[0357] 18

<220>
<223> synthetic digital PCR co-amplification forward primer for 97-bp amplicon of beta-globin (hemoglobin beta (HBB)) HbE mutation reverse primer

<400> 52
ctattggcct cctaaaaccc gtctcgtaa 29

<210> 53
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-G-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation normal allele

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (14)...(14)
<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 53
ttggtggtga ggcc 14

<210> 54
<211> 14
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-A-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation mutant allele

<220>
<221> modified_base
<222> (1)...(1)
<223> t modified by 6-carboxyfluorescein (FAM)

<220>
<221> modified_base
<222> (14)...(14)
<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 54
ttggtggtaa ggcc 14

<210> 55
<211> 20
<212> DNA
<213> Artificial Sequence

19

[0358]

<220>
<223> synthetic digital PCR co-amplification forward primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 55
tttcccaaa cctaggcgc 20

<210> 56
<211> 23
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic digital PCR co-amplification reverse primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 56
acagcatcag gatgggcag atc 23

<210> 57
<211> 19
<212> DNA
<213> Artificial Sequence

<220>
<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation normal allele (without deletion)

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC

<220>
<221> modified_base
<222> (19)...(19)
<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 57
cagaggttat ttgaggtcc 19

<210> 58
<211> 15
<212> DNA
<213> Artificial Sequence

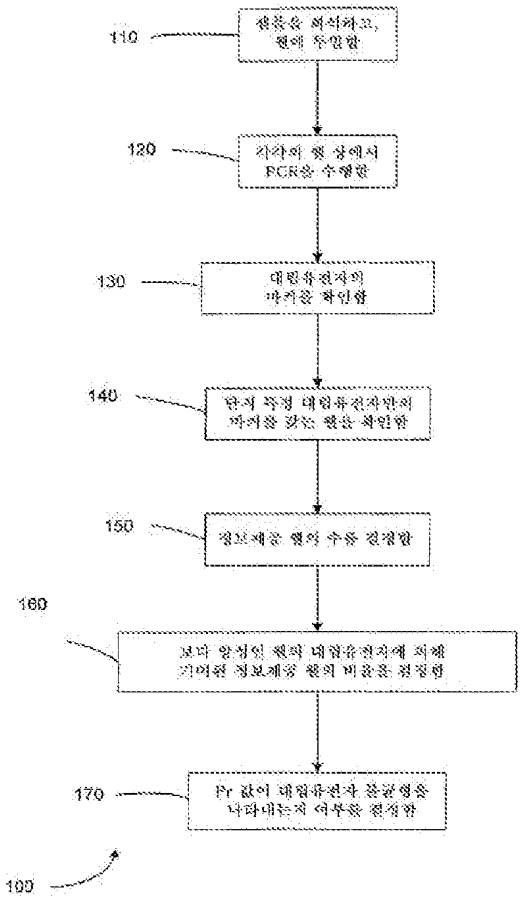
<220>
<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation mutant allele (with deletion)

<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by 6-carboxyfluorescein (FAM)

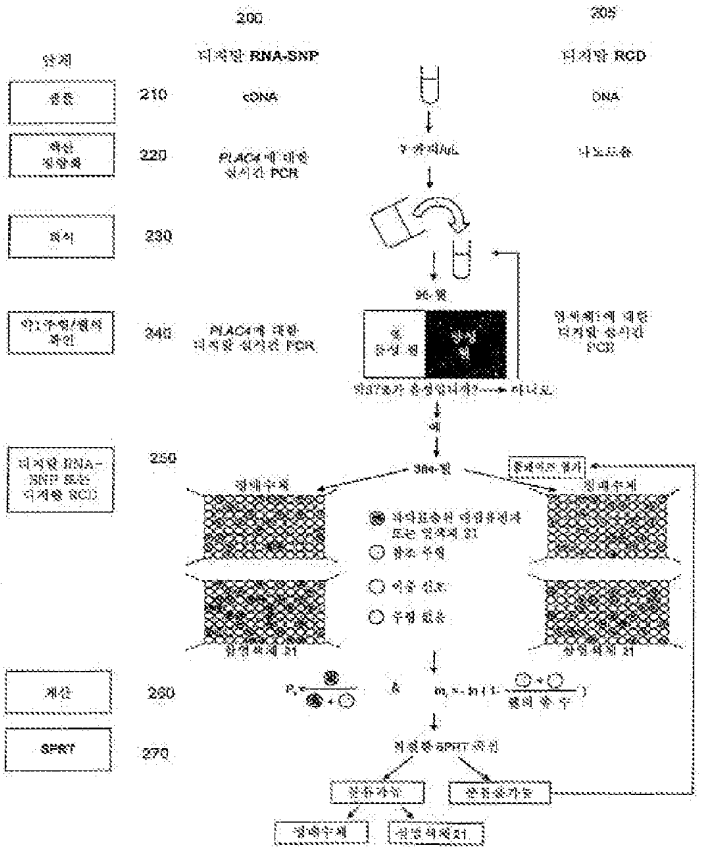
[0359]

도면

도면1


110
샘플을 분석하고, 형의 유전형
120
각각의 형 상에서 PCR을 수행함
130
대립유전자의 마커를 확인함
140
단일 특정 대립유전자만의 마커를 갖는 형을 확인함
150
정보제공 형의 수를 결정함
160
보다 앞서인 형의 대립유전자에 저해 기여된 정보제공 형의 비율을 결정함
170
Pr 값이 대립유전자 분균형을 나타내는지 여부를 결정함
100

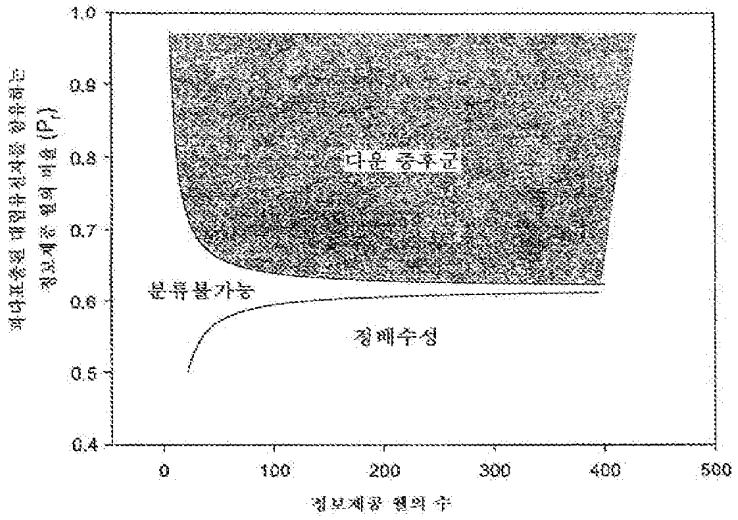
도면2a



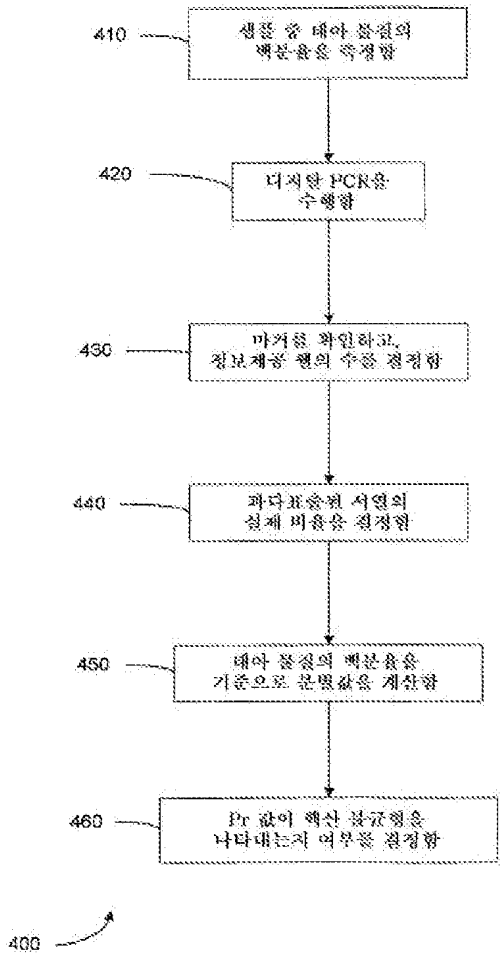
도면2b

샘플	종양지방 전여체 이상	
	획득	소실
종양1	7	5p; 11q; 17p
종양2	1q; 8q; 17q; 20q	8p; 13q; 16q; 18q
종양3	7; 8q; 13q; 20q	1p; 5q; 8p; 15q; 17p; 18q
종양4	1q; 8q; 20q	4q; 8p; 13q; 16q; 17p
종양5	1q; 3q; 5p; 8q	3p; 8p; 9p; 13q; 17p
종양6	3q; 5p; 8q; 19q	3p; 4q; 5q; 10q; 13q; 16q
종양7	1q; 3q; 8; 12; 19	1p; 3p; 9; 11q; 13q; 14q; 16q
종양8	7; 8q; X-q	8p; 13q; 16q

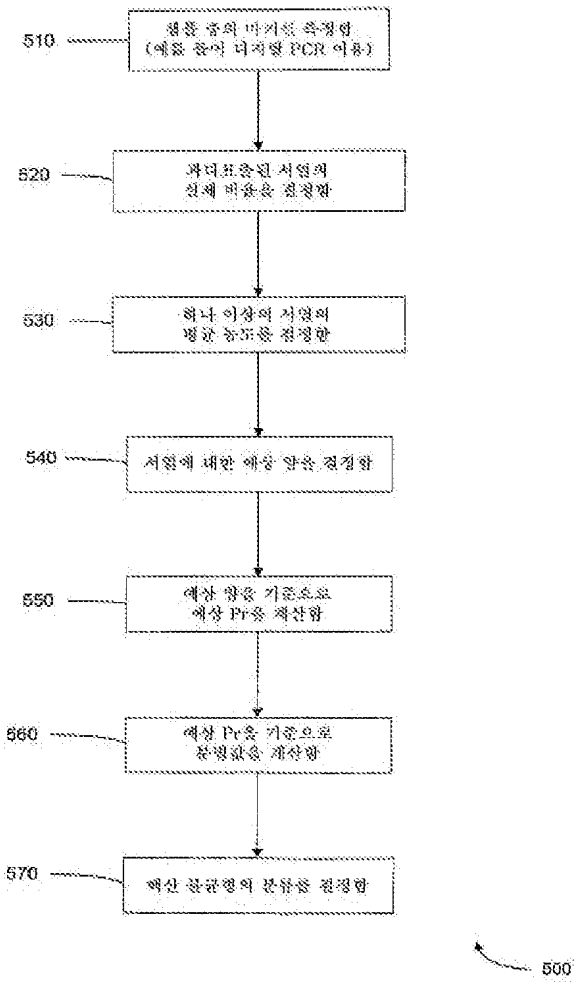
도면3


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
다운 증후군
분류불가능
정배수성
정보제공 랩의 수

도면4


410
생물 중 태아 물질의 백분율을 측정함
420
디지탈 PCR을 수행함
430
마커를 확인하고, 정보제공 랩의 수를 결정함
440
과다표현된 서열의 실제 비율을 결정함
450
태아 물질의 백분율을 기준으로 문턱값을 계산함
460
Pr 값이 핵산 불균형을 나타내는지 여부를 결정함
400

도면5



도면6

m	참조 샘플에서의 실제 양성의 평균 비율	과다표출된 서열에 대한 양성인 분포 비율	두 샘플에서의 모두에 양성인 분포 비율	참조 샘플에서의 음성인 분포 비율	과다표출된 서열에 대한 음성인 분포 비율	두 샘플 중에서 음성인 분포 비율	표준편차 값의 비율	Pr
0.1	0.0982	0.1853	0.0173	0.0779	0.1681	2.11	0.2419	0.58
0.2	0.1803	0.3297	0.0598	0.1535	0.2669	2.02	0.3914	0.69
0.3	0.2592	0.4552	0.1168	0.2422	0.3542	2.35	0.4765	0.7
0.4	0.3357	0.5507	0.1815	0.3483	0.3895	2.49	0.5173	0.71
0.5	0.4015	0.6321	0.2487	0.4448	0.3824	2.66	0.5282	0.73
0.6	0.4512	0.6988	0.3153	0.5359	0.3835	2.82	0.5194	0.74
0.7	0.5034	0.7534	0.3793	0.6241	0.3741	3.01	0.4985	0.75
0.8	0.5507	0.7981	0.4395	0.7112	0.3586	3.25	0.4698	0.76
0.9	0.5934	0.8347	0.4955	0.7981	0.3394	3.49	0.4375	0.78
1	0.6331	0.8647	0.5466	0.8851	0.3181	3.72	0.4036	0.79
1.1	0.6671	0.8892	0.5937	0.9719	0.2966	4	0.3695	0.8
1.2	0.6988	0.9089	0.6374	0.9934	0.2779	4.32	0.3373	0.81
1.3	0.7285	0.9247	0.6784	0.994	0.2623	4.57	0.3088	0.82
1.4	0.7564	0.9375	0.7166	0.9958	0.2516	4.68	0.2854	0.83
1.5	0.7829	0.9478	0.7522	0.9967	0.2453	4.68	0.2685	0.85
1.6	0.8081	0.9562	0.7856	0.9975	0.2437	4.65	0.2582	0.86
1.7	0.8323	0.9628	0.8168	0.9975	0.2455	4.67	0.2539	0.87
1.8	0.8547	0.9727	0.8459	0.9978	0.2508	4.65	0.2526	0.88
1.9	0.8764	0.9776	0.8734	0.9979	0.2582	4.69	0.2532	0.88
2	0.8967	0.9817	0.8988	0.9979	0.2659	4.69	0.2487	0.89

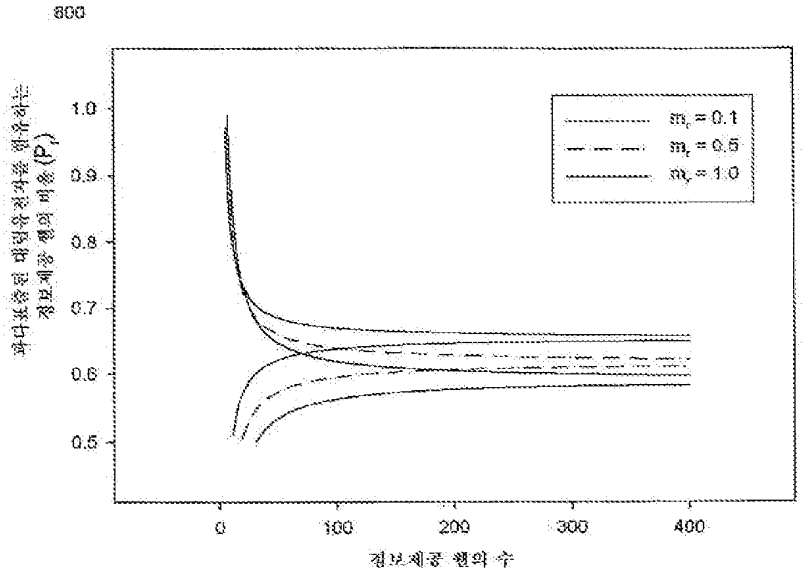
600

도면7

m_c	샘플 중 태아 DNA의 백분율							
	10%		25%		50%		100%	
	정보제공 행의 백분율	P_1	정보제공 행의 백분율	P_1	정보제공 행의 백분율	P_1	정보제공 행의 백분율	P_1
0.1	18%	0.51	18%	0.53	19%	0.56	21%	0.61
0.2	30%	0.51	31%	0.53	32%	0.56	35%	0.61
0.3	39%	0.51	40%	0.53	41%	0.57	43%	0.62
0.4	45%	0.51	45%	0.54	46%	0.57	48%	0.63
0.5	48%	0.52	49%	0.54	49%	0.57	51%	0.63
0.6	50%	0.52	50%	0.54	50%	0.58	51%	0.64
0.7	50%	0.52	50%	0.54	50%	0.58	50%	0.65
0.8	49%	0.52	49%	0.54	48%	0.58	48%	0.65
0.9	48%	0.52	47%	0.55	47%	0.59	46%	0.66
1	46%	0.52	45%	0.55	44%	0.59	43%	0.67
1.1	44%	0.52	43%	0.55	42%	0.6	40%	0.68
1.2	41%	0.52	40%	0.55	39%	0.6	37%	0.69
1.3	39%	0.52	38%	0.55	36%	0.6	34%	0.69
1.4	36%	0.52	35%	0.56	33%	0.61	31%	0.7
1.5	34%	0.52	33%	0.56	31%	0.61	28%	0.71
1.6	31%	0.52	30%	0.56	28%	0.62	26%	0.72
1.7	29%	0.53	28%	0.56	26%	0.62	23%	0.73
1.8	27%	0.53	25%	0.57	24%	0.63	21%	0.73
1.9	24%	0.53	23%	0.57	21%	0.63	19%	0.74
2	22%	0.53	21%	0.57	20%	0.64	17%	0.75

700

도면8



도면9a

99	분류	신규 방오피막층		기존 방오피막층		99	분류	신규 방오피막층		기존 방오피막층	
		정박수명	운항거리 21	정박수명	운항거리 21			정박수명	운항거리 21	정박수명	운항거리 21
0.1	정박	3%	37%	7%	35%	1.1	정박	39%	39%	14%	35%
	부착률	3%	2%	3%	1%		부착률	1%	0%	4%	1%
	이탈률	97%	61%	97%	65%		이탈률	60%	2%	82%	64%
0.2	정박	6%	38%	0%	40%	1.2	정박	97%	99%	11%	35%
	부착률	3%	2%	4%	0%		부착률	1%	0%	4%	0%
	이탈률	90%	34%	96%	60%		이탈률	2%	8%	84%	65%
0.3	정박	34%	75%	6%	39%	1.3	정박	91%	98%	7%	34%
	부착률	4%	2%	4%	0%		부착률	1%	2%	4%	0%
	이탈률	62%	24%	90%	61%		이탈률	8%	2%	89%	66%
0.4	정박	54%	81%	14%	35%	1.4	정박	92%	91%	2%	34%
	부착률	3%	1%	4%	0%		부착률	1%	2%	4%	0%
	이탈률	43%	19%	82%	61%		이탈률	8%	9%	91%	66%
0.5	정박	64%	85%	20%	38%	1.5	정박	92%	99%	2%	32%
	부착률	3%	1%	4%	0%		부착률	1%	1%	4%	0%
	이탈률	33%	14%	76%	62%		이탈률	8%	0%	93%	68%
0.6	정박	74%	89%	23%	38%	1.6	정박	92%	99%	1%	31%
	부착률	2%	1%	4%	0%		부착률	1%	1%	4%	0%
	이탈률	24%	11%	73%	62%		이탈률	8%	10%	94%	69%
0.7	정박	80%	89%	25%	37%	1.7	정박	93%	89%	1%	30%
	부착률	2%	1%	4%	0%		부착률	1%	1%	4%	0%
	이탈률	18%	10%	71%	63%		이탈률	7%	10%	96%	70%
0.8	정박	83%	91%	23%	37%	1.8	정박	92%	88%	0%	28%
	부착률	2%	1%	4%	0%		부착률	1%	1%	4%	0%
	이탈률	15%	9%	73%	63%		이탈률	7%	11%	96%	72%
0.9	정박	86%	91%	21%	36%	1.9	정박	92%	87%	0%	26%
	부착률	1%	1%	4%	0%		부착률	1%	1%	3%	0%
	이탈률	13%	8%	75%	64%		이탈률	8%	12%	97%	74%
1	정박	88%	92%	19%	36%	2	정박	92%	86%	0%	24%
	부착률	1%	0%	4%	0%		부착률	1%	1%	4%	0%
	이탈률	11%	8%	77%	64%		이탈률	9%	13%	96%	76%

930

도면9b

99	분류	신규 방오피막층		기존 방오피막층		99	분류	신규 방오피막층		기존 방오피막층	
		정박수명	운항거리 21	정박수명	운항거리 21			정박수명	운항거리 21	정박수명	운항거리 21
0.1	정박	59%	82%	46%	85%	1.1	정박	1000%	1000%	82%	98%
	부착률	3%	1%	3%	1%		부착률	0%	0%	1%	0%
	이탈률	40%	21%	69%	105%		이탈률	0%	0%	7%	105%
0.2	정박	89%	98%	79%	90%	1.2	정박	1000%	1000%	90%	98%
	부착률	1%	0%	2%	0%		부착률	0%	0%	1%	0%
	이탈률	10%	2%	19%	10%		이탈률	0%	0%	9%	10%
0.3	정박	96%	99%	88%	96%	1.3	정박	1000%	1000%	95%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	1%	0%
	이탈률	4%	1%	11%	10%		이탈률	0%	0%	18%	10%
0.4	정박	98%	100%	92%	98%	1.4	정박	1000%	1000%	97%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	1%	0%
	이탈률	1%	0%	7%	10%		이탈률	0%	0%	12%	10%
0.5	정박	100%	100%	94%	99%	1.5	정박	1000%	1000%	98%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	1%	0%
	이탈률	0%	0%	6%	10%		이탈률	0%	0%	13%	10%
0.6	정박	100%	100%	94%	98%	1.6	정박	1000%	1000%	82%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	1%	0%
	이탈률	0%	0%	5%	10%		이탈률	0%	0%	17%	10%
0.7	정박	100%	100%	95%	99%	1.7	정박	1000%	1000%	79%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	2%	0%
	이탈률	0%	0%	5%	10%		이탈률	0%	0%	19%	10%
0.8	정박	100%	100%	95%	99%	1.8	정박	1000%	1000%	74%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	2%	0%
	이탈률	0%	0%	5%	10%		이탈률	0%	0%	24%	10%
0.9	정박	100%	100%	94%	99%	1.9	정박	1000%	1000%	70%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	2%	0%
	이탈률	0%	0%	6%	10%		이탈률	0%	0%	28%	10%
1	정박	100%	100%	94%	99%	2	정박	1000%	1000%	67%	98%
	부착률	0%	0%	1%	0%		부착률	0%	0%	3%	0%
	이탈률	0%	0%	6%	10%		이탈률	0%	0%	30%	10%

950

도면10

m_c	분류 결과	정확수치	T21	m_b	분류 결과	정확수치	T21
0.1	정확	57%	88%	0.9	정확	100%	100%
	부정확	3%	1%		부정확	0%	0%
	미분류	40%	11%		미분류	0%	0%
0.2	정확	88%	98%	1	정확	100%	100%
	부정확	1%	0%		부정확	0%	0%
	미분류	10%	2%		미분류	0%	0%
0.3	정확	96%	99%	1.1	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	4%	1%		미분류	0%	0%
0.4	정확	99%	100%	1.2	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	1%	0%		미분류	0%	0%
0.5	정확	100%	100%	1.3	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.6	정확	100%	100%	1.4	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.7	정확	100%	100%	1.6	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%
0.8	정확	100%	100%	2	정확	100%	100%
	부정확	0%	0%		부정확	0%	0%
	미분류	0%	0%		미분류	0%	0%

1000

도면11

m_c	분류 결과	정확수치	실험결과 T21	m_c	분류 결과	정확수치	실험결과 T21
0.1	정확	38%	44%	0.7	정확	95%	94%
	부정확	2%	2%		부정확	0%	1%
	미분류	60%	54%		미분류	5%	5%
0.2	정확	64%	69%	0.8	정확	95%	95%
	부정확	2%	2%		부정확	1%	0%
	미분류	34%	29%		미분류	4%	5%
0.3	정확	78%	82%	0.9	정확	96%	96%
	부정확	2%	1%		부정확	0%	0%
	미분류	20%	17%		미분류	4%	4%
0.4	정확	85%	87%	1	정확	97%	96%
	부정확	1%	1%		부정확	0%	0%
	미분류	14%	12%		미분류	3%	4%
0.5	정확	89%	90%	1.5	정확	97%	96%
	부정확	1%	1%		부정확	0%	0%
	미분류	10%	9%		미분류	3%	4%
0.6	정확	92%	93%	2	정확	98%	95%
	부정확	1%	1%		부정확	0%	0%
	미분류	7%	7%		미분류	4%	5%

1100

도면12

정의 수	분포 형태	검출 율 차이 DNA 복원율							
		10%		25%		50%		100%	
		정방수행	정방수행 아님	정방수행	정방수행 아님	정방수행	정방수행 아님	정방수행	정방수행 아님
384	정확	0%	0%	6%	7%	47%	49%	89%	90%
	부정확	0%	0%	0%	1%	2%	2%	1%	1%
	미분류	100%	100%	93%	92%	51%	49%	10%	9%
768	정확	0%	0%	23%	34%	73%	76%	99%	99%
	부정확	0%	0%	2%	2%	2%	1%	0%	0%
	미분류	100%	100%	75%	74%	25%	23%	2%	1%
1152	정확	1%	1%	39%	40%	86%	87%	100%	100%
	부정확	0%	0%	2%	2%	1%	1%	0%	0%
	미분류	99%	99%	59%	58%	13%	12%	0%	0%
1536	정확	2%	2%	50%	51%	93%	93%	100%	100%
	부정확	0%	0%	2%	2%	1%	1%	0%	0%
	미분류	98%	97%	48%	47%	7%	6%	0%	0%
1920	정확	4%	4%	60%	60%	96%	96%	100%	100%
	부정확	0%	1%	2%	2%	0%	0%	0%	0%
	미분류	96%	95%	38%	37%	4%	3%	0%	0%
3840	정확	18%	19%	84%	83%	100%	100%	100%	100%
	부정확	1%	1%	1%	1%	0%	0%	0%	0%
	미분류	81%	80%	15%	15%	18%	22%	0%	0%
7680	정확	45%	45%	97%	97%	100%	100%	100%	100%
	부정확	2%	2%	0%	0%	0%	0%	0%	0%
	미분류	55%	53%	3%	3%	0%	0%	0%	0%

1200

도면13a

샘플	염기서열*	염색 수	전체 데이터셋에서 관찰 염색인 염색 수				m_i*	SPRT 결과		
			A 염색	G 염색	AG	모두 염색		P_i*	염색수행 영역*	분류*
참조 DNA										
N877	AG	384	89	83	126	86	0.79	0.61	0.63 - 0.65	정확수행
N790	AG	384	102	83	78	126	0.82	0.65	0.61 - 0.63	정확수행
N435	AGG	384	49	157	120	48	0.63	0.76	0.62 - 0.64	T21
N985	AAG	384	135	69	82	98	0.5	0.66	0.61 - 0.63	T21
샘플 RNA										
V933	AG	384	109	93	71	117	0.56	0.53	0.61 - 0.63	정확수행
V943	AG	384	89	105	74	121	0.55	0.53	0.61 - 0.63	정확수행
N855	AGG	384	52	138	95	99	0.48	0.73	0.61 - 0.63	T21
T215	AAG	384	146	56	138	42	0.71	0.72	0.62 - 0.64	T21

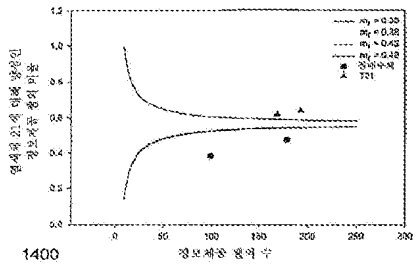
1300

도면13b

샘플	염기서열	염색 수	개별 데이터셋에서 관찰 염색인 염색 수				m	SPRT 결과		
			A 염색	G 염색	AG	모두 염색		비율	염색 수행 영역	분류
M23890P	AG	384	90	100	97	97	0.67	0.626	0.62 - 0.64	정확수행
M23891P	AG	384	97	105	85	117	0.65	0.52	0.61 - 0.63	정확수행
M2473SP	AG	384	98	92	34	162	0.3	0.582	0.58 - 0.62	정확수행
M2526P	AG	384	28	26	3	324	0.08	0.609	0.64 - 0.64	정확수행
M2528P	AG	384	112	85	44	143	0.41	0.568	0.60 - 0.62	정확수행
M2601P	AG	384	80	101	73	121	0.55	0.528	0.61 - 0.63	정확수행
M2607P	AG	384	73	91	57	163	0.47	0.555	0.60 - 0.63	정확수행
M2630P	AG	384	88	80	52	175	0.57	0.577	0.59 - 0.62	정확수행
M2639P	AG	384	71	66	17	240	0.21	0.559	0.58 - 0.62	정확수행
M2525P	AAG	384	116	53	21	200	0.21	0.675	0.58 - 0.61	T21
M2273P	AAG	768	248	127	112	283	0.37	0.66	0.60 - 0.61	T21
M2718P	AGG	384	66	114	66	138	0.42	0.633	0.60 - 0.62	T21
M1519P	AGG	384	56	130	54	142	0.34	0.691	0.59 - 0.62	T21

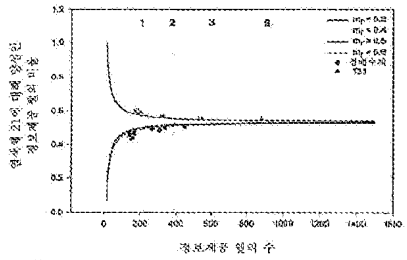
1350

도면14a



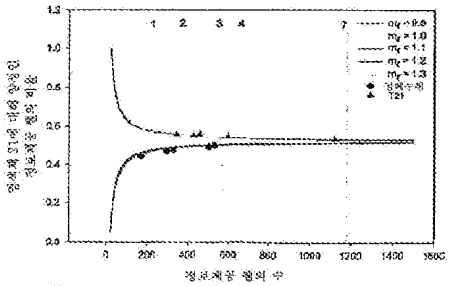
1400

도면14b



1430

도면14c



1460

도면15a

샘플	유전자형	염의 수	개별 대립유전자에 대해 양성인 염의 수				SPRT 결과			
			A 단독	G 단독	GA	모두 음성	m^a	P^b	분류불가능한 영역	분류[c]
V943	AG	765	168	209	117	271	0.47	0.554	0.609- 0.621	정확수체
N435	AGG	765	92	281	105	287	0.30	0.753	0.597- 0.610	T21
T21-5	AAG	765	287	126	141	231	0.43	0.679	0.607- 0.618	T21

[a] 염 당 알로 분포의 수

[b] P, 표준 유기 작에 의해 계산하였음:

$$\frac{\text{전단표준된 대립유전자에 대해 양성인 염의 수}}{\text{A에 대해서만 양성인 염의 수 + G에 대해서만 양성인 염의 수}}$$

[c] P=m 분류불가능한 영역 속에 있는 경우에 "정확수체"로 지정하였고; P>m 분류불가능한 영역 위에 있는 경우에 "T21"로 지정하였음. T21, 삼염색체 21.

도면15b

재현 #	염의 수	개별 대립유전자에 대해 양성인 염의 수			
		A 단독	G 단독	GA	모두 음성
S1	765	14	15	0	736
S2	765	7	10	0	748
S3	765	13	9	1	744
S4	765	7	8	0	750
S5	765	12	8	0	745
S6	765	8	14	0	743
S7	765	13	9	0	743
S8	765	9	8	0	748
S9	765	14	15	0	736
S10	765	9	12	0	744
S11	765	8	17	0	740
S12	765	11	9	0	745

도면16c

샘플	유전자형	좌위	개별 대립유전자에 대해 관찰된 횟수					SPRT 결과		
			합의 수	A 단독	G 단독	GA	오류 음성	Pi	결론불가능한 영역	분류
M2879P	AG	5	3825	82	103	5	3635	0.557	0.570- 0.602	정배수체
M2876P	AG	3	2295	35	40	0	2220	0.533	0.548- 0.625	정배수체
M3078P	AG	3	2295	33	33	1	2228	0.500	0.541- 0.631	정배수체
M2975P	AG	12	9180	56	46	0	9078	0.549	0.556- 0.615	정배수체
M2831P	AAG	2	1530	42	13	0	1476	0.764	0.631- 0.640	T21

[a] 분류에 필요한 최소의 수

도면16a

샘플	유전자형[a]	A 단독	G 단독	둘 다	π_c[b]	P_i[c]	SPRT 분석 결론불가능한 영역[d]	분류[e]
N3505	AG	157	131	48	0.27	0.545	0.59-0.61	정배수체
N3530	AG	174	132	52	0.28	0.569	0.59-0.61	정배수체
N3531	AG	146	129	28	0.23	0.531	0.58-0.60	정배수체
N3533	AG	140	135	45	0.29	0.525	0.59-0.61	정배수체
N2772	AG	159	149	79	0.35	0.516	0.60-0.61	정배수체
N3506	AG	154	145	69	0.32	0.515	0.60-0.61	정배수체
N3565	AG	94	73	7	0.11	0.563	0.58-0.61	정배수체
N3560	AG	101	109	38	0.18	0.519	0.58-0.61	정배수체
N3561	AG	91	70	17	0.12	0.565	0.58-0.61	정배수체
N828	AAG	193	89	22	0.16	0.632	0.58-0.60	T18
CA2	AGG	12	28	0	0.016	0.700	0.51-0.66	T18
CA7	AGG	99	171	56	0.23	0.638	0.59-0.61	T18

[a] 유전자형은 검정 분석법에 의해 결정하였음.

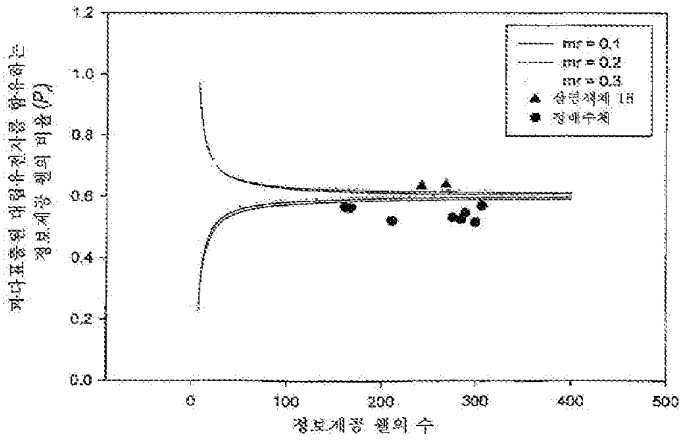
[b] π_c 값은 각 염색체 유형 분석의 평균 수용 비율임.

[c] P_i 값은 각각표준화된 대립유전자에 따른 양성인 정보 제공 횟수의 비율을 나타냄.

[d] 분류불가능한 영역은 π_c 값에 따라 변함.

[e] 유전 분류불가능한 영역 아래에 있는 경우에 "정배수체"로 지정하였으며; Pi이 분류불가능한 영역 위에 있는 경우에 "T18"로 지정하였음.

도면16b

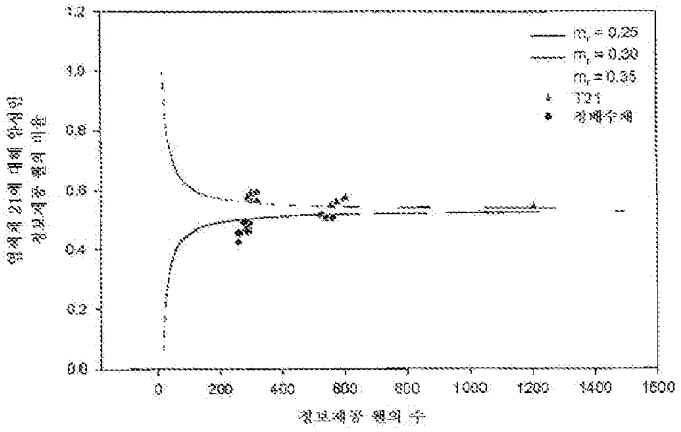


도면17

수준	SNP	서열에 따른 염기서열 돌연변이		염기서열 돌연변이 종류			m_r*	상관계수		
		수	종수	종 수	Chi2's 검정	Chi1 검정		R^2	정확한 구간추정 계산	종수*
721	N0284	1	386	88	172	191	0.382	0.558	0.557 - 0.559	721
	N0374	2	668	118	357	551	0.28	0.550	0.516 - 0.547	721
	N0392	1	288	73	168	126	0.262	0.573	0.569 - 0.582	721
	N0349	2	676	107	322	252	0.27	0.551	0.518 - 0.548	721
	N0495	1	389	97	178	129	0.233	0.569	0.569 - 0.569	721
	N0456	1	918	86	199	129	0.26	0.564	0.569 - 0.569	721
	N0607	2	813	134	365	257	0.32	0.573	0.519 - 0.568	721
	N0786	1	262	52	167	122	0.26	0.575	0.565 - 0.580	721
	N0893	4	1508	299	557	558	0.31	0.546	0.526 - 0.555	721
	N0921	1	317	105	180	137	0.38	0.568	0.519 - 0.568	721
중복	N0195	1	282	56	111	151	0.34	0.424	0.564 - 0.568	중복수체
	N0192	1	205	55	132	150	0.34	0.483	0.563 - 0.561	중복수체
	N0195	2	521	108	272	255	0.27	0.518	0.517 - 0.548	중복수체
	N0221	1	542	86	205	257	0.27	0.537	0.517 - 0.547	중복수체
	N0282	1	285	57	144	151	0.32	0.486	0.558 - 0.559	중복수체
	N0483	1	282	70	135	145	0.33	0.456	0.554 - 0.563	중복수체
	N0527	1	282	55	137	156	0.23	0.486	0.556 - 0.560	중복수체
	N0619	1	281	50	136	145	0.31	0.484	0.555 - 0.561	중복수체
	N0625	1	281	60	134	157	0.33	0.480	0.557 - 0.560	중복수체
	N0691	2	568	116	287	275	0.30	0.552	0.518 - 0.547	중복수체

* 해당 염기서열 돌연변이 수
*m_r 값은 0.01 간격 이하 생략함
$\frac{\text{서열종류에 대해 염기서열 Chi2's에서 선택 염기서열 종류 수}}{\text{Chi1's에 대해서만 선택된 염기서열 종류 수 + Chi1's에 대해서만 선택된 종류 수}}$
* m_r이 최적값 이하에 있는 경우에는 "중복수체"로 표시하였으며, 이러한 염기서열을 갖지 않은 경우에 "721"로 표시하였음.

도면18



도면19

수준	SNP	서열에 따른 염기서열 돌연변이 수	염기서열 돌연변이 종류 수				m_r*	상관계수		
			종수	Chi2's 검정	Chi1 검정			R^2	정확한 구간추정 계산	종수*
721	A0276	1	380	269	248	131	0.74	0.846	0.567 - 0.569	721
	A2008	1	345	573	332	115	1.02	0.870	0.578 - 0.586	721
	A2178	1	384	268	288	195	0.68	0.823	0.571 - 0.588	721
	A56536	1	384	509	248	120	0.39	0.870	0.551 - 0.578	721
	A59750	1	268	462	186	98	1.02	0.704	0.566 - 0.616	721
	A59407A	1	378	66	268	120	0.21	0.821	0.548 - 0.572	721
중복	A0228	1	367	275	195	182	0.98	0.564	0.576 - 0.582	중복수체
	A0297	1	282	228	184	208	0.84	0.486	0.572 - 0.583	중복수체
	A0295	1	378	222	190	183	0.76	0.510	0.568 - 0.585	중복수체
	A2383	1	284	48	157	177	0.35	0.580	0.541 - 0.578	중복수체
	A56835A	1	378	126	182	178	0.69	0.518	0.556 - 0.578	중복수체
	A54855A	1	359	94	188	188	0.38	0.525	0.550 - 0.554	중복수체
	A54884A	1	372	271	184	176	0.65	0.522	0.575 - 0.580	중복수체
	A57883A	1	415	159	200	212	0.74	0.484	0.568 - 0.564	중복수체
	A57903A	1	216	97	159	157	0.35	0.503	0.548 - 0.575	중복수체
	A56837A	1	188	9	87	111	0.17	0.458	0.523 - 0.570	중복수체
	A58214A	1	182	5	88	85	0.12	0.484	0.528 - 0.568	중복수체
	A58890A	1	384	133	191	205	0.55	0.458	0.561 - 0.568	중복수체
	A52282A	1	334	342	172	157	1.02	0.830	0.582 - 0.587	중복수체
	A55625A	1	128	8	94	84	0.12	0.529	0.518 - 0.591	중복수체
	A52439A	1	378	32	202	176	0.44	0.538	0.565 - 0.579	중복수체
	A1156	1	354	128	186	182	0.48	0.529	0.552 - 0.577	중복수체
	A1171	1	369	100	186	174	0.44	0.538	0.554 - 0.570	중복수체
	A1182	1	288	40	130	130	0.28	0.529	0.542 - 0.578	중복수체
	A1185	1	349	403	139	113	1.61	0.541	0.564 - 0.622	중복수체
	A1375	1	378	65	201	172	0.42	0.828	0.564 - 0.576	중복수체
	A1375	1	369	66	186	188	0.56	0.529	0.560 - 0.576	중복수체
	A1386	1	387	67	188	182	0.47	0.433	0.566 - 0.577	중복수체
	A1586	1	274	124	157	134	0.67	0.551	0.556 - 0.577	중복수체

* 해당 염기서열 돌연변이 수
*m_r 값은 0.01 간격 이하 생략함
$\frac{\text{서열종류에 대해 염기서열 Chi2's에서 선택 염기서열 종류 수}}{\text{Chi1's에 대해서만 선택된 염기서열 종류 수 + Chi1's에 대해서만 선택된 종류 수}}$
* m_r이 최적값 이하에 있는 경우에는 "중복수체"로 표시하였으며, 이러한 염기서열을 갖지 않은 경우에 "721"로 표시하였음.

표 20

유형	시료	사용된 SNP 수	정확도를 얻어낸 수	재수집 결과 수			R_c	SPRT 분석		
				동일 염색체	Chr18 염색체	Chr21 염색체		R_c^2	판정 불가능한 염색체	분류
정상수준	N0230	1	360	248	181	192	0.68	0.476	0.573 - 0.589	E
	N0625	1	373	225	179	181	0.78	0.484	0.559 - 0.586	E
	N0851	1	358	206	152	156	0.67	0.536	0.554 - 0.553	E
T18	N0528	1	543	259	329	214	0.55	0.605	0.573 - 0.590	T18
	N0530	1	346	152	249	106	0.52	0.694	0.574 - 0.592	T18
	N0936	1	372	286	209	182	0.88	0.663	0.573 - 0.590	E
	N1978	1	343	115	221	122	0.37	0.644	0.561 - 0.574	T18
	N1751	1	303	84	161	115	0.31	0.630	0.548 - 0.575	T18

* 총 참 음주 분석의 수
** R_c 값은 아래 식에 따라 계산하였다:
$$R_c = \frac{\text{21번염색체 염색체증가가 (chr18)인 것을 검출한 횟수}}{\text{chr21에 염색체가 증가되는 횟수 + chr18에 염색체가 증가된 횟수}}$$
*** R_c이 분류불가능한 영역 이내에 있는 정수는 '영역수준'으로 지정하였고, R_c이 분류불가능한 영역 외에 있는 정수는 'T21'로 지정하였다.

표 21

특성	시료	증폭부위의 수	결과	정보성을 얻은 염색체 수	재수집된 염색체 수			R_c	SPRT 분석		
					동일 염색체	Chr21 염색체	염색체 양쪽		R_c^2	분류불가능한 염색체	분류
정상	N0155	1	chr21/chr1	149	78	71	77	0.92	0.480	0.555 - 0.584	E
		1	chr21/chr18	152	68	78	73	0.44	0.525	0.557 - 0.584	E
			정상	300	141	149	150	0.48	0.500	0.552 - 0.568	E
	N2192	1	chr21/chr1	162	73	70	82	0.56	0.432	0.492 - 0.561	E
		1	chr21/chr18	183	68	85	90	0.53	0.508	0.556 - 0.578	E
			정상	365	141	153	122	0.55	0.572	0.577 - 0.587	E
	N0355	1	chr21/chr1	102	58	31	51	0.45	0.559	0.490 - 0.581	E
		2	chr21/chr18	165	45	91	84	0.41	0.491	0.488 - 0.550	E
			정상	327	103	162	165	0.43	0.455	0.512 - 0.558	E
	N0281	1	chr21/chr1	173	80	81	92	0.47	0.458	0.484 - 0.570	E
		1	chr21/chr18	168	91	88	81	0.42	0.512	0.489 - 0.553	E
			정상	335	104	186	173	0.45	0.480	0.514 - 0.557	E
	N0552	1	chr21/chr1	155	97	101	55	0.64	0.543	0.503 - 0.575	E
		1	chr21/chr18	176	111	91	65	0.57	0.517	0.504 - 0.557	E
			정상	322	208	192	120	0.60	0.529	0.532 - 0.562	E
		2	chr21/chr1	174	120	89	65	0.69	0.466	0.503 - 0.577	E
			정상	626	311	277	259	0.65	0.517	0.525 - 0.552	E
T21	N0334	1	chr21/chr1	159	85	106	85	0.51	0.500	0.502 - 0.575	T21
		1	chr21/chr18	172	88	80	86	0.50	0.517	0.488 - 0.570	T21
			정상	331	133	205	171	0.58	0.539	0.517 - 0.556	T21
		2	chr21/chr1	175	73	90	80	0.53	0.517	0.495 - 0.575	T21
			정상	349	206	253	237	0.61	0.532	0.534 - 0.549	T21
		2	chr21/chr18	178	82	85	83	0.47	0.534	0.495 - 0.575	T21
			정상	727	358	527	546	0.59	0.532	0.527 - 0.548	T21
		3	chr21/chr1	184	108	103	81	0.65	0.560	0.504 - 0.575	T21
			정상	911	376	496	495	0.61	0.528	0.525 - 0.548	T21
		3	chr21/chr18	188	102	120	78	0.63	0.508	0.503 - 0.572	T21
			정상	1008	478	610	499	0.56	0.550	0.531 - 0.544	T21
	N0332	1	chr21/chr1	178	108	103	70	0.63	0.566	0.508 - 0.577	T21
		1	chr21/chr18	181	65	116	61	0.55	0.578	0.501 - 0.574	T21
			정상	324	151	243	161	0.59	0.568	0.519 - 0.557	T21
	N0365	1	chr21/chr1	169	92	93	76	0.64	0.525	0.499 - 0.582	T21
		1	chr21/chr18	156	48	87	75	0.38	0.624	0.488 - 0.586	T21
			정상	324	198	170	154	0.42	0.532	0.512 - 0.568	T21
		2	chr21/chr1	149	98	85	63	0.41	0.577	0.495 - 0.565	T21
			정상	573	176	266	317	0.41	0.541	0.519 - 0.551	T21
		3	chr21/chr18	188	43	109	68	0.34	0.595	0.487 - 0.580	T21
			정상	641	217	355	289	0.40	0.558	0.523 - 0.546	T21
	N0706	1	chr21/chr1	182	82	105	76	0.53	0.558	0.498 - 0.575	T21
		1	chr21/chr18	175	53	97	81	0.58	0.545	0.492 - 0.575	T21
			정상	380	165	203	157	0.64	0.564	0.515 - 0.558	T21
	N0399	1	chr21/chr1	186	108	108	72	0.61	0.600	0.505 - 0.586	T21
		1	chr21/chr18	182	88	104	84	0.50	0.555	0.508 - 0.571	T21
			정상	368	171	212	158	0.55	0.578	0.518 - 0.556	T21

* 총 참 음주 분석의 수
** R_c 값은 아래 식에 따라 계산하였다:
$$R_c = \frac{\text{21번염색체 염색체증가가 (chr21)인 것을 검출한 횟수}}{\text{chr21에 염색체가 증가되는 횟수 + 참조에 염색체가 정상인 횟수}}$$
*** R_c이 분류불가능한 영역 이내에 있는 정수는 '영역수준'으로 지정하였고, R_c이 분류불가능한 영역 외에 있는 정수는 'T21'로 지정하였다.

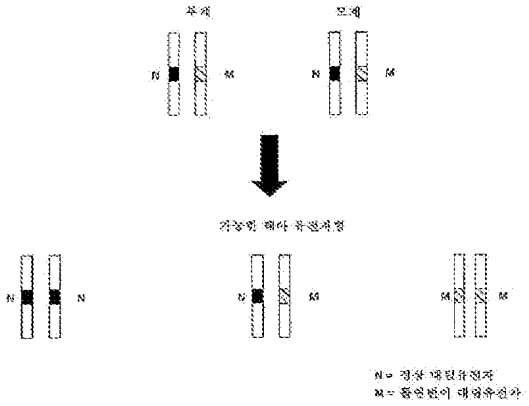
도면22a

		부여된 코딩된 숫자			
		영역체 21	참조 영역체	R	분류
정해주제 사례 1	컴정 1	37	33	0.52345751	여분류(Duplicate)
	컴정 2	31	42	0.42506279	여분류
	컴정 3	29	37	0.44166785	여분류
	컴정 4	39	26	0.6006691	여분류
	플레이트 1 소계	136	138	0.495821	여분류
	컴정 1	32	25	0.56605316	여분류
	컴정 2	27	24	0.53165424	여분류
	컴정 3	10	17	0.3651535	여분류
	컴정 4	24	18	0.5626754	여분류
	플레이트 2 소계	93	84	0.52474	여분류
	총계	229	223	0.507182	정해주제
사례 2	컴정 1	46	35	0.56321585	여분류
	컴정 2	31	45	0.41246525	여분류
	컴정 3	32	22	0.59954763	여분류
	컴정 4	38	33	0.5228176	여분류
	플레이트 1 소계	146	135	0.518153	여분류
	컴정 1	28	23	0.56272135	여분류
	컴정 2	23	22	0.51186633	여분류
	컴정 3	23	15	0.60888187	여분류
	컴정 4	29	26	0.5293546	여분류
	플레이트 2 소계	102	85	0.545204	여분류
	플레이트 1+2 소계	248	221	0.529593	여분류
	컴정 1	17	35	0.4122149	여분류
	컴정 2	25	16	0.5027299	여분류
	컴정 3	13	26	0.3382602	여분류
	컴정 4	27	24	0.5316872	여분류
	플레이트 3 소계	82	91	0.478775	여분류
	플레이트 1~3 소계	331	311	0.515069	여분류
	컴정 1	15	9	0.6269193	여분류
	컴정 2	14	15	0.4824153	여분류
	컴정 3	19	17	0.5146313	여분류
	컴정 4	21	18	0.5270050	여분류
	플레이트 4 소계	69	60	0.532264	여분류
	플레이트 1~4 소계	399	372	0.517942	여분류

도면22b

	컴정 1	10	13	0.4337957	여분류
	컴정 2	11	14	0.4390082	여분류
	컴정 3	7	16	0.3018022	여분류
	컴정 4	19	10	0.6579055	여분류
	플레이트 5 소계	48	54	0.469851	여분류
	총계	447	426	0.5123444	정해주제
T21					
사례 3	컴정 1	50	33	0.60012253	여분류
	컴정 2	55	47	0.54456055	여분류
	컴정 3	48	28	0.6196526	여분류
	컴정 4	56	36	0.6108731	여분류
	총계	208	144	0.59079	T21
사례 4	컴정 1	56	36	0.61065513	여분류
	컴정 2	38	32	0.53905912	여분류
	컴정 3	40	24	0.62788068	여분류
	컴정 4	47	27	0.634511	여분류
	총계	180	119	0.603476	T21

도면23



도면24a

사례	샘플	적중률 %	해법의 수	ZFY	ZFX	ALL POS	NET ZFY	NET ZFX	참조 대립유전자	m_c	SPRT°
1	XY + XX	50%	1	239	465	155	64	310	Y	0.37	XX
	(XX %)	25%	1	328	447	205	123	242	Y	0.56	XX
2	$XY_1 + XY_2$	50%	1	414	405	220	194	185	X	0.85	XY
	(XY_2 %)	25%	1	432	425	245	187	181	X	0.81	XY

도면24b

사례	해법의 수	ZFY	ZFX	ALL POS	NET ZFY	NET ZFX	참조 대립유전자	m_c	SPRT
1	1	81	171	18	63	153	Y	0.11	XX
2	1	103	175	23	80	152	Y	0.14	XX
3	1	110	175	28	82	147	Y	0.16	XX
4	1	111	156	24	87	132	Y	0.18	XX
5	2	236	322	50	186	272	Y	0.17	XX
6	1	116	184	29	87	155	Y	0.18	XX
7	1	94	169	19	75	150	Y	0.13	XX
8	1	91	155	24	67	131	Y	0.13	XX
9	3	333	447	74	259	373	Y	0.16	XX
10	1	110	154	24	86	130	Y	0.16	XX
11	1	83	170	26	57	144	Y	0.11	XX
12	1	85	159	16	69	143	Y	0.12	XX

표 25

사례	계통의 수	조직 유전자형[a]	적아 유전자형[a]	이상형 적아 %	A	G	ALL POS	NET A	NET G	상대 빈도	결과 해석	유전자좌수	m	SPRT[c]
1	1	AG	GG	19%	364	316	116	148	200	42.6	A	2	0.42	GG
2	1	AG	GG	19%	388	276	91	127	185	44.8	A	2	0.32	GG
3	1	AG	AG	19%	348	251	97	151	164	48.4	A	2	0.39	AG
4	1	AG	GG	19%	338	240	91	102	186	48.7	A	2	0.28	GG
5	1	AG	AG	19%	338	245	75	155	176	48.6	A	2	0.32	AG
6	1	AG	AG	19%	318	225	105	169	169	49.1	A	2	0.34	AG
7	1	AG	AG	19%	284	278	94	183	166	49.6	A	2	0.49	AG
8	1	AG	GG	19%	298	303	134	162	210	49.7	A	2	0.48	GG
9	1	AG	AG	19%	275	285	102	178	165	49.0	A	2	0.43	AG
10	1	AG	AA	19%	307	279	120	217	165	NM	G	2	0.48	AA
1	1	AG	GG	20%	248	259	85	152	185	53.0	A	2	0.33	GG
2	1	AG	GG	20%	191	289	78	112	230	59.5	A	2	0.28	GG
3	1	AG	AG	20%	248	218	75	172	143	53.2	G	2	0.30	U
4	1	AG	GG	33%	158	257	99	99	198	55.6	A	2	0.23	GG
5	1	AG	AG	20%	231	215	71	160	144	53.3	G	2	0.33	AG
6	1	AG	AG	33%	231	265	99	165	165	50.0	A	2	0.45	AG
7	1	AG	AG	20%	255	231	83	158	148	52.7	G	2	0.33	AG
8	1	AG	GG	33%	248	311	104	142	207	59.3	A	2	0.39	GG
9	1	AG	AG	20%	267	249	85	164	166	54.6	G	2	0.39	AG
10	1	AG	AA	20%	303	288	77	238	138	NM	G	2	0.31	AA
1	1	AG	GG	25%	242	304	99	146	205	55.7	A	2	0.38	GG
2	1	AG	GG	25%	289	314	88	124	226	54.8	A	2	0.32	GG
3	1	AG	AG	25%	304	212	68	148	132	58.8	A	2	0.31	AG
4	1	AG	GG	25%	185	309	78	108	214	66.8	A	2	0.28	GG
5	1	AG	AG	25%	233	248	77	156	156	54.5	A	2	0.38	AG
6	1	AG	AG	25%	282	265	110	182	165	57.3	A	2	0.48	AG
7	1	AG	AG	25%	281	275	95	166	160	58.5	A	2	0.42	AG
8	1	AG	GG	25%	236	327	86	146	224	58.7	A	2	0.37	GG
9	1	AG	AG	25%	271	265	104	167	176	59.2	A	2	0.44	AG
10	1	AG	AA	25%	243	154	46	197	108	NM	G	2	0.22	AA
1	1	AG	GG	50%	136	282	57	79	205	59.0	A	2	0.28	GG
2	1	AG	GG	50%	163	356	78	85	280	66.8	A	2	0.28	GG
3	1	AG	AG	50%	205	208	68	137	155	65.7	A	2	0.30	AG
4	1	AG	GG	50%	167	356	66	77	290	69.8	A	2	0.19	GG
5	1	AG	AG	50%	247	246	91	156	155	57.7	G	2	0.39	AG
6	1	AG	AG	50%	216	245	68	150	177	69.8	A	2	0.24	AG
7	1	AG	AG	50%	255	253	103	152	179	51.9	A	2	0.41	AG
8	1	AG	GG	50%	156	306	83	73	233	67.6	A	2	0.33	GG
9	1	AG	AG	50%	135	226	74	161	151	55.7	G	2	0.35	AG
10	1	AG	AA	50%	317	196	35	282	71	NM	G	2	0.15	AA

[a] A = 동형접합 대립유전자, G = 이형접합 대립유전자.
[b] 결과 해석 DNA 농도 의사형 ZFYY 증폭에 의해 결정됨.
[c] 대상은 여성으로, 의사형 ZFYY의 증폭은 사용되지 않음.
[d] SPRT는 사례 10을 제외하고 의사형 ZFYY 증폭으로 결정된 적아 또는 수형태임을; U, 미분류.

표 26a

사례	조직 유전자형[a]	적아 유전자형[a]	이상형 적아 %	M	N	ALL POS	NET M	NET N	1단계 적아 농도	1단계 SPRT	2단계 의사형 ZFY/ZFX 비율 적아 농도	2단계 SPRT	결과 해석	유전자좌수	m
1	MN	MM	3	849	753	273	573	480	9%	MM	7.8%	MM	N	2	0.40
	MN	MM	3	863	748	283	574	458	10%	MM	8.6%	MM	N	2	0.39
	MN	MM	1	323	219	72	244	140	25%	MM	24.1%	MM	N	2	0.34
	MN	MM	1	418	190	82	326	99	50%	MM	57.3%	MM	N	2	0.23
2	MN	MM	9	2559	2475	857	1672	1589	5%	U	2.4%	U	N	2	0.48
	MN	MM	3	895	807	300	595	507	10%	MM	9.7%	MM	N	2	0.46
	MN	MM	1	277	253	97	180	156	25%	MM	24.2%	MM	N	2	0.42
	MN	MM	1	272	295	101	172	194	50%	MM	58.6%	MM	N	2	0.44

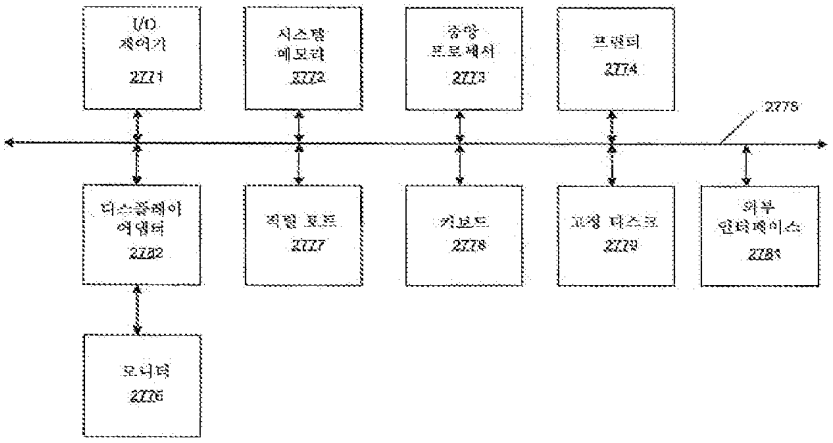
[a] M = CTTT 결실 존재, N = CTTT 결실 부재.
[b] SPRT는 여성군 별로 수행하였음.
[c] SPRT는 의사형 ZFY/X 증폭으로 결정된 적아 또는 수형태임을; U, 미분류.

표 26b

사례	여성의 수	조직 유전자형[a]	적아 유전자형[a]	여성의 적아 %	M	N	ALL POS	NET M	NET N	결과 해석 농도[b]	결과 해석	유전자좌수	m	SPRT[c]
1	1	MN	MM	20%	203	200	58	145	142	28.0	M	2	0.30	MM
2	1	MN	MN	20%	149	204	33	110	165	29.7	M	2	0.22	MM
3	1	MN	MN	20%	128	212	31	97	181	30.7	M	2	0.18	MM
4	1	MN	MM	20%	173	196	45	128	151	27.0	M	2	0.26	MM
5	1	MN	NN	20%	147	230	47	100	183	24.3	M	2	0.21	MM
6	1	MN	MN	20%	199	206	45	154	161	29.2	M	2	0.30	MM
7	1	MN	NN	20%	186	237	53	107	179	18.7	M	2	0.24	MM
8	1	MN	MN	20%	203	180	43	160	137	33.4	M	2	0.27	MM
9	1	MN	NN	20%	175	241	61	112	338	16.9	M	2	0.26	MM
10	1	MN	MN	20%	174	190	52	122	136	16.8	M	2	0.28	U
11	1	MN	NN	20%	167	242	49	118	193	17.7	M	2	0.25	MM
12	1	MN	MM	20%	215	199	47	186	152	21.6	M	2	0.30	MM

[a] M = CTTT 결실 존재, N = CTTT 결실 부재.
[b] 결과 해석 DNA 농도 의사형 ZFYY 증폭으로 결정됨.
[c] SPRT는 의사형 ZFY/X 증폭으로 결정된 적아 또는 수형태임을; U, 미분류.

도면27



서 열 목 록

SEQUENCE LISTING

<110> Lo, Yuk-Ming Dennis

Chiu, Rossa Wai Kwun

Chan, Kwan Chee

Zee, Benny Chung Ying

Chong, Ka Chun

The Chinese University of Hong Kong

<120> Determining a Nucleic Acid Sequence mbalance

<130> IPA161076-CN

<140> WO PCT/GB08/02524

<141> 2008-07-23

<150> US 60/951,438

<151> 2007-07-23

<160> 58

<170> FastSEQ for Windows Version 4.0

<210> 1

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR and digital RNA-SNP analysis

PLAC4 gene-specific reverse transcription primer

<400> 1

agtatataga accatgttta ggccaga 27

<210> 2

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 primer

<400> 2

ccgctagggt gtcttttaag c 21

<210> 3

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 primer

<400> 3

gtgttgcaat acaaaatgag tttct 25

<210> 4

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic real-time PCR PLAC4 fluorescent probe

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 4

attggagcaa attc 14

<210> 5

<211> 104

<212> DNA

<213> Artificial Sequence

<220>

<223> single-stranded synthetic DNA oligonucleotide amplicon

<400> 5

cgccgctagg gtgtctttta agctattgga gcaaattcaa atttggctta aagaaaaaga 60

aactcatttt gtattgcaac accaggagta tcccaaggga ctcg 104

<210> 6

<211> 22

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RNA-SNP analysis non-intron spanning forward primer

<400> 6

tttgtattgc aacaccattt gg 22

<210> 7

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting rs8130833 SNP of PLAC4

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (15)...(15)

<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 7

tcgtcgtcta acttg 15

<210> 8

<211> 17

<212> DNA

<

213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting rs8130833 SNP of PLAC4

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (17)...(17)

<223> g modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 8

attcgtcatc taacttg 17

<210> 9

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification forward primer for paralogous loci on chromosome 21 and 1

<400> 9

gttgttctgc aaaaaacctt cga 23

<210> 10

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for paralogous loci on chromosome
21 and 1

<400> 10

cttggccaga aatacttcat taccatat 28

<210> 11

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (19)...(19)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 11

tacctccata atgagtaaa 19

<210> 12

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog
<220>
<221> modified_base
<222> (1)...(1)
<223> c modified by fluorescent reporter VIC
<220>
<221> modified_base
<222> (20)...(20)
<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)
<400> 12
cgtacctctg taatgtgtaa 20
<210> 13
<211> 27
<212> DNA
<213> Artificial Sequence
<220>
<223> synthetic digital PCR PLAC4 gene-specific reverse

transcription primer
<400> 13
agtatataga accatgttta ggccaga 27
<210> 14
<211> 22
<212> DNA
<213> Artificial Sequence
<220>
<223> synthetic digital PCR amplification forward primer
for rs8130833 SNP region of PLAC4 gene
<400> 14
tttgtattgc aacaccattt gg 22
<210> 15
<211> 15
<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting
PLAC4 rs8130833 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (15)...(15)

<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 15

tcgtcgtcta acttg 15

<210> 16

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting
PLAC4 rs8130833 SNP

<220>

<221> modified_base

<222
> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (17)...(17)

<223> g modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 16

attcgtcatc taacttg 17

<210> 17

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) gene-specific reverse transcription

primer

<400> 17

cgcagacttc tcaccaaaca 20

<210> 18

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR amplification serpin
peptidase inhibitor clade B (ovalbumin) member 2
(SERPINB2) forward primer

<400> 18

ctcagctctg caatcaatgc 20

<210> 19

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting

serpin peptidase inhibitor clade B (ovalbumin)
member 2 (SERPINB2) rs6098 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (16)...(16)

<223> t modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 19

ccacagggaa ttattt 16

<210> 20

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting

serpin peptidase inhibitor clade B (ovalbumin)
member 2 (SERPINB2) rs6098 SNP

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (16)...(16)

<223> t modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 20

ccacagggga ttattt 16

<210> 21

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification

forward primer for 101-bp amplicon for paralogous

loci on chromosome 21 and 1

<400> 21

gttgttctgc aaaaaacctt cga 23

<210> 22

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 101-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 22

cttggccaga aatacttcat taccatat 28

<210> 23

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (19)...(19)

<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 23

tacctccata atgagtaaa 19

<210> 24

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 1 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (20)...(20)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 24

cgtacctctg taatgtgtaa 20

<210> 25

<211> 24

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 25

gtacagaaac cacaaactga tcgg 24

<210> 26

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for 128-bp amplicon for paralogous
loci on chromosome 21 and 18

<400> 26

gtccaggctg tgggcct 17

<210> 27

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 21 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> a modified by minor groove binding non-fluorescent
quencher (MGBNFQ)

<400> 27

aagaggcgag gcaa 14

<210> 28

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe
designed to target chromosome 18 paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (15)...(15)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 28

aagaggacag gcaac 15

<210> 29

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification forward primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 29

acgttggatg gttgttctgc aaaaaacctt cga 33

<210> 30

<211> 38

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 30

acgttggatg cttggccaga aatacttcat taccatat 38

<210> 31

<211> 21

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification extension primer for 121-bp amplicon for paralogous loci on chromosome 21 and 1

<400> 31

ctcatcctca cttcgtacct c 21

<210> 32

<211> 34

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification forward primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 32

acgttggatg gtacagaaac cacaaactga tcgg 34

<210> 33

<211> 27

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification reverse primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 33

acgttggatg gtccaggctg tgggcct 27

<210> 34

<211> 17

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification extension primer for 148-bp amplicon for paralogous loci on chromosome 21 and 18

<400> 34

acaaaagggg gaagagg 17

<210> 35

<211> 35

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
forward primer for 81 bp amplicon for paralogous
loci on chromosome 21 and 1

<400> 35

acgttggatg ttgatgaagt ctcatctcta cttcg 35

<210> 36

<211> 33

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
reverse primer for 81 bp amplicon for paralogous

loci on chromosome 21 and 1

<400> 36

acgttggatg caataagctt ggccagaaat act 33

<210> 37

<211> 37

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification
forward primer for 82 bp amplicon for paralogous
loci on chromosome 21 and 7

<400> 37

acgttggatg gaatttaagc taaatcagcc tgaactg 37

<210> 38

<211> 37

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

reverse primer for 82 bp amplicon for paralogous loci on chromosome 21 and 7

<400> 38

acgttggatg gtttctcata gttcatcgta ggcttat 37

<210> 39

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 101 bp amplicon for paralogous loci on chromosome 21 and 2

<400> 39

acgttggatg tcaggcaggg ttctatgcag 30

<210> 40

<211> 30

<212> DNA

<

213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

reverse primer for 101 bp amplicon for paralogous loci on chromosome 21 and 2

<400> 40

acgttggatg aggcggcttc ctggctcttg 30

<210> 41

<211> 31

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification

forward primer for 102 bp amplicon for paralogous loci on chromosome 21 and 6

<400> 41

acgttggatg gctcgtctca ggctcgtagt t 31

<210> 42

<211> 30

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital RCD analysis co-amplification reverse primer for 102 bp amplicon for paralogous loci on chromosome 21 and 6

<400> 42

acgttggatg tttcttcgag cccttcttgg 30

<210> 43

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 1

<400

> 43

gtctcatctc tacttcgtac ctc 23

<210> 44

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 7

<400> 44

ttttacgctg tccccattt 19

<210> 45

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 2

<400> 45

ggtctatgca ggagccgac 19

<210> 46

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic extension primer for paralogous sequence mismatches (PSM) on paralogous loci on chromosomes 21 and 6

<400> 46

tgggcgcggg agcggacttc gctgg 25

<210> 47

<211> 25

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification forward primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 47

caagtgctgg actcagatgt aactg 25

<210> 48

<211> 28

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR analysis co-amplification reverse primer for 87-bp amplicon for zinc finger protein sequences on chromosome X (ZFX) and Y (ZFY)

<400> 48

tgaagtaatg tcagaagcta aaacatca 28

<210> 49

<211> 16

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome X paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (16)...(16)

<223> a modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 49

tctttagcac attgca 16

<210> 50

<211> 17

<212

> DNA

<213> Artificial Sequence

<220>

<223> synthetic chromosome-specific TaqMan probe designed to target chromosome Y paralog

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (17)...(17)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 50

tctttaccac actgcac 17

<210> 51

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward primer for 87-bp amplicon of beta-globin (hemoglobin beta (HBB)) HbE mutation forward primer

<400> 51

gggcaaggtg aacgtggat 19

<210> 52

<211> 29

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward primer for 87-bp amplicon of beta-globin (hemoglobin beta (HBB)) HbE mutation reverse primer

<400> 52

ctattggtct ccttaaacct gtcttgtaa 29

<210> 53

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-G-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation normal allele

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 53

ttggtggtga ggcc 14

<210> 54

<211> 14

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-A-specific TaqMan probe targeting beta-globin (hemoglobin beta (HBB)) HbE mutation mutant allele

<220>

<221> modified_base

<222> (1)...(1)

<223> t modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (14)...(14)

<223> c modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 54

ttggtggtaa ggcc 14

<210> 55

<211> 20

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification forward primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 55

ttttcccacc cttaggctgc 20

<210> 56

<211> 23

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic digital PCR co-amplification reverse primer for 87-bp and 83-bp amplicon of CD41/42 deletion mutation normal and mutant allele

<400> 56

acagcatcag gagtggacag atc 23

<210> 57

<211> 19

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation normal allele (without deletion)

<220>

<221> modified_base

<222> (1)...(1)

<223> c modified by fluorescent reporter VIC

<220>

<221> modified_base

<222> (19)...(19)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 57

cagaggttct ttgagtcct 19

<210> 58

<211> 15

<212> DNA

<213> Artificial Sequence

<220>

<223> synthetic allele-specific TaqMan probe targeting CD41/42 deletion mutation mutant allele (with deletion)

<220>

<221> modified_base

<222> (1)...(1)

<223> a modified by 6-carboxyfluorescein (FAM)

<220>

<221> modified_base

<222> (15)...(15)

<223> t modified by minor groove binding non-fluorescent quencher (MGBNFQ)

<400> 58

agaggttgag tcctt 15

Electronic Acknowledgement Receipt

EFS ID:	44609436
Application Number:	17379544
International Application Number:	
Confirmation Number:	2092
Title of Invention:	DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS
First Named Inventor/Applicant Name:	Yuk-Ming Dennis Lo
Customer Number:	20350
Filer:	David Dah-Wei Hsu/Kelly Kearney
Filer Authorized By:	David Dah-Wei Hsu
Attorney Docket Number:	080015-028420US-1249377
Receipt Date:	27-DEC-2021
Filing Date:	19-JUL-2021
Time Stamp:	09:38:29
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Transmittal Letter	IDS_Transmittal_080015-1249377.pdf	103551 ae7a0f2e897063429cbaf9add29eb6ef144a035a	no	2

Warnings:

Information:					
2	Information Disclosure Statement (IDS) Form (SB08)	IDS_080015-1249377.pdf	208098 27a8cea6ef2f56ef1a1f08f3f27ffde888f0f341	no	5
Warnings:					
Information:					
This is not an USPTO supplied IDS fillable form					
3	Foreign Reference	WO_2009013492_A1.pdf	6216178 804f191e26f32d22327296a706e373d56286199b	no	115
Warnings:					
Information:					
4	Foreign Reference	WO_2010027484_A2.pdf	6061866 05791f73c935947fb67afe46a16ce46e353a479c	no	117
Warnings:					
Information:					
5	Foreign Reference	WO_2010068289_A2.pdf	5801936 58c828f46f3a4d864a8656c20d68013ece94d106	no	103
Warnings:					
Information:					
6	Foreign Reference	WO_2013170429_A1.pdf	4005660 f809d374c8776ff595a84afc2d40a10438226720	no	30
Warnings:					
Information:					
7	Foreign Reference	WO_2013188846_A1.pdf	9283413 8828bfcbc40899fabbe244e3922b9bfb8dc753ab	no	197
Warnings:					
Information:					
8	Foreign Reference	CN_104053789_A.pdf	778706 f594ed3ca33ee6f89908fe9dca88915e4d76ff9a	no	2
Warnings:					
Information:					

9	Foreign Reference	WO_2014153757_A1.pdf	4707471 d043c6954e4e4c6da2e418b8f43710c58b5414ff	no	32
Warnings:					
Information:					
10	Foreign Reference	CN_104053789_B.pdf	5200902 35b616a5593fcedec83a46e2f1906c1745701b4a	no	21
Warnings:					
Information:					
11	Foreign Reference	AU_2012380221_B2.pdf	1478424 28b6c0347e4585c647421268884f0179f029154	no	34
Warnings:					
Information:					
12	Foreign Reference	WO_2017012544_A1.pdf	5975388 3800f6f739e07bbe8af3e2237762925dceffa64e	no	127
Warnings:					
Information:					
13	Foreign Reference	KR_101896167_B1.pdf	5814678 568ce21f9f2ff6a335c467a0f7ba1f33abf507eb	no	98
Warnings:					
Information:					
14	Non Patent Literature	NPL_BEAULAURIER_2019.pdf	1988572 b16745bc8206a2a37bf59964e172026fde05e6ae	no	16
Warnings:					
Information:					
15	Non Patent Literature	NPL_BLOW_2016.pdf	18147066 2101c0a78cb1df971a15d1162e8211e60643c881	no	28
Warnings:					
Information:					

16	Non Patent Literature	NPL_CLARK_2012.pdf	9886876 b7955cb27d6f9c602a360c789b1142b5dab24f85	no	12
Warnings:					
Information:					
17	Non Patent Literature	NPL_CLARK_2013.pdf	882293 2b95961fa1cd0b120eae4e9eebdd8dc8f42d063b	no	10
Warnings:					
Information:					
18	Non Patent Literature	NPL_EID_2009.pdf	941854 8ec7c4030db696ecc95632baab4548b042517c06	no	6
Warnings:					
Information:					
19	Non Patent Literature	NPL_FENG_2013.pdf	663439 67a3110cb2dafda48d78beea0b1696404fef765a	no	10
Warnings:					
Information:					
20	Non Patent Literature	NPL_FENG_2014.pdf	2371275 a010c24372c794bfddba511745a8e356fff7a96f	no	12
Warnings:					
Information:					
21	Non Patent Literature	NPL_FLUSBERG_2010.pdf	6416793 47c4d531cbe1b3f741a9bf02ceb41b7085d11ddd	no	5
Warnings:					
Information:					
22	Non Patent Literature	NPL_HE_2017.pdf	847766 a7835e9e6cb505f62ad6a07854c02b201bb3aa6d	no	2
Warnings:					
Information:					

23	Non Patent Literature	NPL_KILE_2010.pdf	2328082 6b4234d04a636837018dbf04b63b0ec846ba77ea	no	9
Warnings:					
Information:					
24	Non Patent Literature	NPL_LIU_2019.pdf	17033812 1ae01136db52452663cc1e4b951c74cbf7faf9cb	no	12
Warnings:					
Information:					
25	Non Patent Literature	NPL_LIU_2020.pdf	873085 1c5276b58cd92ade7d3102a581a665b5b4799b68	no	11
Warnings:					
Information:					
26	Non Patent Literature	NPL_NI_2019.pdf	7768142 03fdd28c849c9125ff18faf8cdd085887ebdff44	no	10
Warnings:					
Information:					
27	Non Patent Literature	NPL_SCHADT_2013.pdf	13965512 d6836cde90c5ba79968fdd868371d48028c628c7	no	13
Warnings:					
Information:					
28	Non Patent Literature	NPL_SUGAI_2017.pdf	3037818 b45de86adea4ac8d64316784b25be027eec6124a	no	10
Warnings:					
Information:					
29	Non Patent Literature	NPL_SUZUKI_2016.pdf	7780927 ca343769d51fbf786c8bcd17db4272bf7e79fa09	no	9
Warnings:					
Information:					

30	Non Patent Literature	NPL_SUZUKI_2017.pdf	24272446 70e58d7dfe7bde7600ccb818d0ffb60571aaf031	no	117

Warnings:	
Information:	
Total Files Size (in bytes):	174842029

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

I hereby certify that this correspondence is being filed via EFS-Web with the United States Patent and Trademark Office on _____.

KILPATRICK TOWNSEND & STOCKTON LLP

By:_____

PATENT
Attorney Docket No.: 080015-1249377-028420US
Client Ref. No.: 19/MED/901 US Con1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	Confirmation No.: 2092
Yuk-Ming Dennis Lo et al.	Examiner: Laura B. Goddard
Application No.: 17/379,544	Technology Center/Art Unit: 1642
Filed: July 19, 2021	
For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS	INFORMATION DISCLOSURE STATEMENT
Customer No.: 20350	

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR §1.98(a)(2), copies of any cited U.S. Patents and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR §1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information. This Information Disclosure Statement is being filed before the mailing date of the first Office Action on the merits.

Applicant believes that no fee is required for submission of this statement. However, if any fees are due for the submission of this Information Disclosure Statement, please deduct those fees from Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP
dhsu@kilpatricktownsend.com
1400 Wewatta Street
Suite 600
Denver, CO, 80202
United States of America

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
17/379,544	07/19/2021	Yuk-Ming Dennis Lo	080015-028420US-1249377

CONFIRMATION NO. 2092

PUBLICATION NOTICE

20350
Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

OC00000130049877

Title:DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Publication No.US-2021-0363571-A1
Publication Date:11/25/2021

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publically available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently http://www.uspto.gov/patft/.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Public Records Division. The Public Records Division can be reached by telephone at (571) 272-3150 or (800) 972-6382, by facsimile at (571) 273-3250, by mail addressed to the United States Patent and Trademark Office, Public Records Division, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently https://portal.uspto.gov/pair/PublicPair. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

Office of Data Managment, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/379,544	07/19/2021	Yuk-Ming Dennis Lo	080015-028420US-1249377	2092

20350 7590 11/23/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
11/23/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

Kilpatrick Townsend & Stockton LLP - West Coast Mailstop: IP Docketing - 22
1100 Peachtree Street Suite 2800
Atlanta, GA 30309

In re Application of	:
Lo et al.	: **DECISION ON PETITION**
Application No. 17/379,544	: **UNDER 37 C.F.R. § 1.84(a)(2)**
Filed: 07/19/2021	:
Attorney Docket No. 080015-028420US-1249377	:

This is a decision on the Petition to Accept Color Drawings under 37 C.F.R 1.84(a)(2), received in the United States Patent and Trademark Office on July 19, 2021.

37 C.F.R. § 1.84(a)(2) in part states:
[t]he Office will accept color drawings in utility patent applications only after granting a petition filed under this paragraph explaining why the color drawings are necessary---
Any such petition must include the following:

1. Payment of the fee set forth under 37 C.F.R. 1.17(h),
2. Submission of three (3) sets of color drawings, or one (1) set of color drawings if filed by EFS-Web, and
3. The first paragraph of the brief description of the drawings in the Specification contains the following required language:

"The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee."

A review of the Patent and Trademark records indicate that the petition did not meet the following requirement(s): 1 ☐ 2 ☐ 3 ☐ Other ☑

In accordance with 37 CFR 1.84(a)(2), color drawings are only accepted on rare occasions when they are the only practical medium by which to disclose the subject matter to be patented and a petition for color drawings must explain why the color drawings are necessary . See also MPEP 608.02(VIII). Applicant's explanation does not clearly explain why color drawings are the only practical medium by which to disclose the subject matter sought to be patented. An explanation should include which figure(s) require color as the only practical medium, what features of the drawings require color, and why the features can only be depicted using color.

Unless a color drawings petition is filed and granted, Applicant will be required to either cancel the drawings or provide substitute black and white drawings. The specification will also need to be amended to remove the requisite color drawings language. See MPEP 608.02 (VIII).

Accordingly, the petition is **DISMISSED**

A renewed petition filed under 37 C.F.R. 1.84(a)(2) must be filed within TWO (2) MONTHS of this decision. No extension of time is permitted. No additional fee is required with the renewed petition.

Should there be any questions about this decision, please contact MARIANNE C SEIDEL, by letter addressed to Director, TC 1600 at the address listed above, or by telephone at (571)272-0584, or by facsimile sent to the general Office facsimile number, 571-273-8300.

/MARIANNE C SEIDEL/
Quality Assurance Specialist, Technology Center 1600

/ANNE M. GUSSOW/
Supervisory Patent Examiner, Technology Center 1600

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/379,544	07/19/2021	Yuk-Ming Dennis Lo	080015-028420US-1249377	2092

20350 7590 11/23/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
GODDARD, LAURA B

ART UNIT	PAPER NUMBER
1642	

NOTIFICATION DATE	DELIVERY MODE
11/23/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Office Action Summary	Application No. 17/379,544	Applicant(s) Lo et al.	
	Examiner LAURA B GODDARD	Art Unit 1642	AIA (FITF) Status Yes

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 2 MONTHS FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☑ Responsive to communication(s) filed on 7/19/2021.

☐ A declaration(s)/affidavit(s) under **37 CFR 1.130(b)** was/were filed on _____.

2a)☐ This action is **FINAL.** 2b) ☑ This action is non-final.

3)☐ An election was made by the applicant in response to a restriction requirement set forth during the interview on _____; the restriction requirement and election have been incorporated into this action.

4)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims*

5) ☑ Claim(s) 1-30 is/are pending in the application.

5a) Of the above claim(s) _____ is/are withdrawn from consideration.

6) ☐ Claim(s) _____ is/are allowed.

7) ☐ Claim(s) _____ is/are rejected.

8) ☐ Claim(s) _____ is/are objected to.

9) ☑ Claim(s) 1-30 are subject to restriction and/or election requirement

* If any claims have been determined allowable, you may be eligible to benefit from the **Patent Prosecution Highway** program at a participating intellectual property office for the corresponding application. For more information, please see http://www.uspto.gov/patents/init_events/pph/index.jsp or send an inquiry to **PPHfeedback@uspto.gov.**

Application Papers

10)☐ The specification is objected to by the Examiner.

11)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

Certified copies:

a)☐ All b)☐ Some** c)☐ None of the:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

** See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☐ Notice of References Cited (PTO-892)

2) ☐ Information Disclosure Statement(s) (PTO/SB/08a and/or PTO/SB/08b) Paper No(s)/Mail Date _____.

3) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date _____.

4) ☐ Other: _____.

Notice of Pre-AIA or AIA Status

The present application, filed on or after March 16, 2013, is being examined under the first inventor to file provisions of the AIA.

Election/Restriction

Restriction to one of the following inventions is required under 35 U.S.C. 121:

I. Claims 1-13, drawn to a method for determining whether a sample nucleic acid molecule is of fetal or maternal origin, wherein the sample nucleic acid molecule is cell-free and obtained from a biological sample of a female subject pregnant with a fetus, the method comprising: (a) receiving data acquired by measuring pulses in an optical signal corresponding to nucleotides sequenced in the sample nucleic acid molecule and obtaining, from the data, values for the following properties: for each nucleotide: an identity of the nucleotide, a position of the nucleotide within the sample nucleic acid molecule, a width of the pulse corresponding to the nucleotide, and an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide; (b) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties: the identity of the nucleotide, a position of the nucleotide with respect to a target position within the1 window, the width of the pulse corresponding to the nucleotide, and the

interpulse duration; (c) for each of one or more nucleotides of the nucleotides sequenced in the sample5 nucleic acid molecule, inputting the input data structure into a model, the model configured to: receive the input data structure, generated using the data acquired by measuring the pulses in the optical signal corresponding to the nucleotides, apply optimized parameters of the model to the properties of the input data structure; and output a methylation status specifying whether the nucleotide at the target position in the respective window has a methylation; and (d) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, determining, using the model, the methylation status for whether the methylation is present in the nucleotide when the nucleotide is at the target position within the window in the input data structure; and (e) determining whether the sample nucleic acid molecule is of fetal or maternal origin using the methylation statuses of the one or more nucleotides, classified in C12Q 1/6816.

II. Claims 14-17, drawn to A method of determining whether a copy number aberration is present in a region of a fetal genome, the method comprising: for each sample nucleic acid molecule of a plurality of sample nucleic acid molecules, wherein the plurality of sample nucleic acid molecules is cell-free and obtained from a biological sample of a female subject pregnant with a fetus: (a) receiving data acquired by measuring pulses in an optical signal corresponding to nucleotides sequenced in the sample nucleic acid molecule and obtaining, from the data, values for the following properties: for each nucleotide: an identity of the nucleotide, a position of the nucleotide within the sample nucleic acid molecule, a width of the pulse corresponding to the

nucleotide, and an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide; (b) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties: the identity of the nucleotide, a position of the nucleotide with respect to a target position within the window, the width of the pulse corresponding to the nucleotide, and the interpulse duration; (c) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, inputting the input data structure into a model, the model configured to: receive the input data structure, generated using the data acquired by measuring the pulses in the optical signal corresponding to the nucleotides, apply optimized parameters of the model to the properties of the input data structure; and output a methylation status specifying whether the nucleotide at the target position in the respective window has a methylation, (d) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, determining, using the model, a methylation status for whether the methylation is present in the nucleotide when the nucleotide is at the target position within the window in the input data structure; and (e) identifying the sample nucleic acid molecule as aligning to the region; determining a methylation level of the region using the plurality of methylation statuses of the one or more nucleotides for each sample nucleic acid molecule of the plurality of sample nucleic acid molecules; and determining

whether the copy number aberration is present at the region of the fetal genome using the methylation level, classified in C12Q 1/6816.

III. Claims 18-27, drawn to a method for determining whether a sample nucleic acid molecule is of fetal or maternal origin, wherein the sample nucleic acid molecule is cell-free and obtained from a biological sample of a female subject pregnant with a fetus, the method comprising: (a) sequencing the sample nucleic acid molecule by measuring pulses in an optical signal corresponding to nucleotides in the sample nucleic acid molecule; (b) receiving data acquired by sequencing the sample nucleic acid molecule, and obtaining, from the data, values for the following properties: for each nucleotide: an identity of the nucleotide, a position of the nucleotide within the sample nucleic acid molecule, a width of the pulse corresponding to the nucleotide, and an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide; (c) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the window, the properties: the identity of the nucleotide, a position of the nucleotide with respect to a target position within the window, the width of the pulse corresponding to the nucleotide, and the interpulse duration; (d) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, inputting the input data structure into a model, the model configured to: receive the input data structure, generated using the

data acquired by measuring the pulses in the optical signal corresponding to the nucleotides, apply optimized parameters of the model to the properties of the input data structure; and output a methylation status specifying whether the nucleotide at the target position in the respective window has a methylation; and (e) determining whether the sample nucleic acid molecule is of fetal or maternal origin using the methylation statuses of the one or more nucleotides, classified in C12Q 1/6816.

IV. Claims 28-30, drawn to a method of determining whether a copy number aberration is present in a region of a fetal genome, the method comprising: for each sample nucleic acid molecule of a plurality of sample nucleic acid molecules, wherein the plurality of sample nucleic acid molecules is cell-free and obtained from abiological sample of a female subject pregnant with a fetus: (a) sequencing the sample nucleic acid molecule by measuring pulses in an optical signal corresponding to nucleotides in the sample nucleic acid molecule; (b) receiving data acquired by sequencing the sample nucleic acid molecule, and obtaining, from the data, values for the following properties: for each nucleotide: an identity of the nucleotide, a position of the nucleotide within the sample nucleic acid molecule, a width of the pulse corresponding to the nucleotide, and an interpulse duration representing a time between the pulse corresponding to the nucleotide and a pulse corresponding to a neighboring nucleotide; (c) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, creating an input data structure, the input data structure comprising a window of the nucleotides sequenced in the sample nucleic acid molecule, wherein the input data structure includes, for each nucleotide within the

window, the properties: the identity of the nucleotide, a position of the nucleotide with respect to a target position within the window, the width of the pulse corresponding to the nucleotide, and the interpulse duration; (d) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, inputting the input data structure into a model, the model configured to: receive the input data structure, generated using the data acquired by measuring the pulses in the optical signal corresponding to the nucleotides, apply optimized parameters of the model to the properties of the input data structure; and output a methylation status specifying whether the nucleotide at the target position in the respective window has a methylation, (e) for each of one or more nucleotides of the nucleotides sequenced in the sample nucleic acid molecule, determining, using the model, a methylation status for whether the methylation is present in the nucleotide when the nucleotide is at the target position within the window in the input data structure; and (f) identifying the sample nucleic acid molecule as aligning to the region; determining a methylation level of the region using the plurality of methylation statuses of the one or more nucleotides for each sample nucleic acid molecule of the plurality of sample nucleic acid molecules; and determining whether the copy number aberration is present at the region of the fetal genome using the methylation level, classified in C12Q 1/6816.

The inventions are independent or distinct, each from the other because:

Inventions I-IV are directed to related methods. The related inventions are distinct if: (1) the inventions as claimed are either not capable of use together or can have a materially different design, mode of operation, function, or effect; (2) the

inventions do not overlap in scope, i.e., are mutually exclusive; and (3) the inventions as claimed are not obvious variants. See MPEP § 806.05(j). In the instant case, the inventions as claimed are drawn to materially distinct methods which differ at least in objectives, method steps, reagents used, response variables, and criteria for success. Furthermore, the inventions as claimed do not encompass overlapping subject matter and there is nothing of record to show them to be obvious variants.

Restriction for examination purposes as indicated is proper because all the inventions listed in this action are independent or distinct for the reasons given above and there would be a serious search and/or examination burden if restriction were not required because one or more of the following reasons apply: (a) the inventions have acquired a separate status in the art due to their recognized divergent subject matter; (b) the inventions require a different field of search (for example, employing different search queries for the different steps and reagents); (c) the prior art applicable to one invention would not likely be applicable to another invention; (d) the inventions are likely to raise different non-prior art issues under 35 U.S.C. 101 and/or 35 U.S.C. 112, first paragraph.

Applicant is advised that the reply to this requirement to be complete must include (i) an election of an invention to be examined even though the requirement may be traversed (37 CFR 1.143) **and (ii) identification of the claims encompassing the elected invention**.

The election of an invention may be made with or without traverse. To reserve a right to petition, the election must be made with traverse. If the reply does not distinctly and specifically point out supposed errors in the restriction requirement, the election shall be treated as an election without traverse. Traversal must be presented at the time of election in order to be considered timely. Failure to timely traverse the requirement will result in the loss of right to petition under 37 CFR 1.144. If claims are added after the election, applicant must indicate which of these claims are readable upon the elected invention.

Should applicant traverse on the ground that the inventions are not patentably distinct, applicant should submit evidence or identify such evidence now of record showing the inventions to be obvious variants or clearly admit on the record that this is the case. In either instance, if the examiner finds one of the inventions unpatentable over the prior art, the evidence or admission may be used in a rejection under 35 U.S.C. 103 or pre-AIA 35 U.S.C. 103(a) of the other invention.

Applicant is reminded that upon the cancellation of claims to a non-elected invention, the inventorship must be corrected in compliance with 37 CFR 1.48(a) if one or more of the currently named inventors is no longer an inventor of at least one claim remaining in the application. A request to correct inventorship under 37 CFR 1.48(a) must be accompanied by an application data sheet in accordance with 37 CFR 1.76 that identifies each inventor by his or her legal name and by the processing fee required under 37 CFR 1.17(i).

Any inquiry concerning this communication or earlier communications from the examiner should be directed to LAURA B GODDARD whose telephone number is (571)272-8788. The examiner can normally be reached Mon-Fri, 7am-3:30pm.

Examiner interviews are available via telephone, in-person, and video conferencing using a USPTO supplied web-based collaboration tool. To schedule an interview, applicant is encouraged to use the USPTO Automated Interview Request (AIR) at http://www.uspto.gov/interviewpractice.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Misook Yu can be reached on 571-272-0839. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of published or unpublished applications may be obtained from Patent Center. Unpublished application information in Patent Center is available to registered users. To file and manage patent submissions in Patent Center, visit: https://patentcenter.uspto.gov. Visit https://www.uspto.gov/patents/apply/patent-center for more information about Patent Center and https://www.uspto.gov/patents/docx for information about filing in DOCX format. For additional questions, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/Laura B Goddard/
Primary Examiner, Art Unit 1642

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
17/379,544	07/19/2021	Yuk-Ming Dennis Lo	080015-028420US-1249377	2092

20350 7590 10/27/2021

Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

EXAMINER
CENTRAL, DOCKET

ART UNIT	PAPER NUMBER
OPAP	

NOTIFICATION DATE	DELIVERY MODE
10/27/2021	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

KTSDocketing2@kilpatrick.foundationip.com
ipefiling@kilpatricktownsend.com

Decision Granting Request for Prioritized Examination (Track I)	**Application No.** 17/379,544	**Applicant(s)** Lo et al.	
	Examiner TERRI S JOHNSON	**Art Unit** OPET	**AIA (FITF) Status** Yes

1. THE REQUEST FILED 19 July 2021 IS **GRANTED** .

The above-identified application has met the requirements for prioritized examination

A. ☑ for an original nonprovisional application (Track I).
B. ☐ for an application undergoing continued examination (RCE).

2. **The above-identified application will undergo prioritized examination**. The application will be accorded special status throughout its entire course of prosecution until one of the following occurs:

A. filing a **petition for extension of time** to extend the time period for filing a reply;

B. filing an **amendment to amend the application to contain more than four independent claims, more than thirty total claims** , or a multiple dependent claim;

C. filing a **request for continued examination** ;

D. filing a notice of appeal;

E. filing a request for suspension of action;

F. mailing of a notice of allowance;

G. mailing of a final Office action;

H. completion of examination as defined in 37 CFR 41.102; or

I. abandonment of the application.

Telephone inquiries with regard to this decision should be directed to TERRI JOHNSON at (571)272 -2991. In his/her absence, calls may be directed to Petition Help Desk at (571) 272-3282.

/TERRI S JOHNSON/ Paralegal Specialist, OPET	

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	TOT CLAIMS	IND CLAIMS
17/379,544	07/19/2021		4980	080015-028420US-1249377	30	4

CONFIRMATION NO. 2092

FILING RECEIPT

20350
Kilpatrick Townsend & Stockton LLP - West Coast
Mailstop: IP Docketing - 22
1100 Peachtree Street
Suite 2800
Atlanta, GA 30309

OC000000127705793

Date Mailed: 08/16/2021

Receipt is acknowledged of this non-provisional utility patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF FIRST INVENTOR, and TITLE OF INVENTION. Fees transmitted by check or draft are subject to collection.

Please verify the accuracy of the data presented on this receipt. If an error is noted on this Filing Receipt, please submit a written request for a corrected Filing Receipt, including a properly marked-up ADS showing the changes with strike-through for deletions and underlining for additions. If you received a "Notice to File Missing Parts" or other Notice requiring a response for this application, please submit any request for correction to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections provided that the request is grantable.

Inventor(s)

Yuk-Ming Dennis Lo, Homantin, CHINA;
Rossa Wai Kwun Chiu, Shatin, CHINA;
Kwan Chee Chan, Mei Foo Sun Cheun, CHINA;
Peiyong Jiang, Pak Shek Kok, CHINA;
Suk Hang Cheng, Fanling, CHINA;
Wenlei Peng, Shatin, CHINA;
On Yee Tse, Fanling, CHINA;

Applicant(s)

The Chinese University of Hong Kong, Shatin, HONG KONG;

Power of Attorney: The patent practitioners associated with Customer Number 20350

Domestic Priority data as claimed by applicant

This application is a CON of 16/995,607 08/17/2020 PAT 11091794
which claims benefit of 63/051,210 07/13/2020
and claims benefit of 63/019,790 05/04/2020
and claims benefit of 62/991,891 03/19/2020
and claims benefit of 62/970,586 02/05/2020
and claims benefit of 62/887,987 08/16/2019

Foreign Applications for which priority is claimed (You may be eligible to benefit from the **Patent Prosecution Highway** program at the USPTO. Please see http://www.uspto.gov for more information.) - None.

Foreign application information must be provided in an Application Data Sheet in order to constitute a claim to foreign priority. See 37 CFR 1.55 and 1.76.

Permission to Access Application via Priority Document Exchange: Yes

Permission to Access Search Results: Yes

Applicant may provide or rescind an authorization for access using Form PTO/SB/39 or Form PTO/SB/69 as appropriate.

If Required, Foreign Filing License Granted: 08/13/2021

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 17/379,544**

Projected Publication Date: 11/25/2021

Non-Publication Request: No

Early Publication Request: No
Title

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Preliminary Class

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications: No

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign

patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4258).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

SelectUSA

The United States represents the largest, most dynamic marketplace in the world and is an unparalleled location for business investment, innovation, and commercialization of new technologies. The U.S. offers tremendous resources and advantages for those who invest and manufacture goods here. Through SelectUSA, our nation works to promote and facilitate business investment. SelectUSA provides information assistance to the international investor community; serves as an ombudsman for existing and potential investors; advocates on behalf of U.S. cities, states, and regions competing for global investment; and counsels U.S. economic development organizations on investment attraction best practices. To learn more about why the United States is the best country in the world to develop technology, manufacture products, deliver services, and grow your business, visit http://www.SelectUSA.gov or call +1-202-482-6800.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 17/379,544

APPLICATION AS FILED - PART I

	(Column 1)		(Column 2)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED		NUMBER EXTRA	RATE($)	FEE($)		RATE($)	FEE($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A		N/A	N/A			N/A	320
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A		N/A	N/A			N/A	700
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A		N/A	N/A			N/A	800
TOTAL CLAIMS (37 CFR 1.16(i))	30	minus 20 =	* 10			OR	x 100 =	1000
INDEPENDENT CLAIMS (37 CFR 1.16(h))	4	minus 3 =	* 1				x 480 =	480
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $310 ($155 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).							1680
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))								0.00
* If the difference in column 1 is less than zero, enter "0" in column 2.				TOTAL			TOTAL	4980

APPLICATION AS AMENDED - PART II

		(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT A	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

		(Column 1)		(Column 2)	(Column 3)					
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
AMENDMENT B	Total (37 CFR 1.16(i))	*	Minus	**	=	x =		OR	x =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	x =		OR	x =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest found in the appropriate box in column 1.

==========
Sequence Listing was accepted.
If you need help call the Patent Electronic Business Center at (866) 217-9197 (toll free).
Reviewer: Zhang, Yizhu (ASRC)
Timestamp: [year=2021; month=7; day=21; hr=9; min=2; sec=52; ms=545;]
==========

Validated By CRFValidator v 1.0.5

Application No: 17379544 Version No: 1.0

Input Set:

Output Set:

Started: 2021-07-19 20:52:47.761
Finished: 2021-07-19 20:52:47.882
Elapsed: 0 hr(s) 0 min(s) 0 sec(s) 121 ms
Total Warnings: 5
Total Errors: 0
No. of SeqIDs Defined: 5
Actual SeqID Count: 5

Error code	Error Description
W 213	Artificial or Unknown found in <213> in SEQ ID (1)
W 213	Artificial or Unknown found in <213> in SEQ ID (2)
W 213	Artificial or Unknown found in <213> in SEQ ID (3)
W 213	Artificial or Unknown found in <213> in SEQ ID (4)
W 213	Artificial or Unknown found in <213> in SEQ ID (5)

SEQUENCE LISTING

<110> THE CHINESE UNIVERSITY OF HONG KONG

<120> DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

<130> 080015-028420US-1249377

<140> US 17/379,544
<141> 2021-07-19

<150> 16/995,607
<151> 2020-08-17

<150> 63/051,210
<151> 2020-07-13

<150> 63/019,790
<151> 2020-05-04

<150> 62/991,891
<151> 2020-03-19

<150> 62/970,586
<151> 2020-02-05

<150> 62/887,987
<151> 2019-08-16

<160> 5

<170> PatentIn version 3.5

<210> 1
<211> 36
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 1
gccuguaauc ccagcacuuu guuuuagagc uaugcu 36

<210> 2
<211> 67
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 2
agcauagcaa guuaaaauaa ggcuaguccg uuaucaacuu gaaaaagugg caccgagucg 60

gugcuuu 67

<210> 3
<211> 36
<212> RNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 3
agggucucgc ucugucgccc guuuuagagc uaugcu 36

<210> 4
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 4
atacgtacgt 10

<210> 5
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
oligonucleotide

<400> 5
atacgtacgt 10

SCORE Placeholder Sheet for IFW Content

Application Number: 17379544 Document Date: 07/19/2021

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Drawing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

DocCode – SEQ.TXT

SCORE Placeholder Sheet for IFW Content

Application Number: 17379544 Document Date: 07/19/2021

The presence of this form in the IFW record indicates that the following document type was received in electronic format on the date identified above. This content is stored in the SCORE database.

Since this was an electronic submission, there is no physical artifact folder, no artifact folder is recorded in PALM, and no paper documents or physical media exist. The TIFF images in the IFW record were created from the original documents that are stored in SCORE.

- Sequence Listing

At the time of document entry (noted above):

- USPTO employees may access SCORE content via DAV or via the SCORE web page.
- External customers may access SCORE content via PAIR using the Supplemental Content tab.

Form Revision Date: March 1, 2019

PTO/AIA/80 (07-12)
Approved for use through 11/30/2014, OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number

POWER OF ATTORNEY TO PROSECUTE APPLICATIONS BEFORE THE USPTO

I hereby revoke all previous powers of attorney given in the application identified in the attached statement under 37 CFR 3.73(c).

I hereby appoint:

☒ Practitioners associated with Customer Number: **20350**

OR

☐ Practitioner(s) named below (if more than ten patent practitioners are to be named, then a customer number must be used):

Name	Registration Number	Name	Registration Number

As attorney(s) or agent(s) to represent the undersigned before the United States Patent and Trademark Office (USPTO) in connection with any and all patent applications assigned only to the undersigned according to the USPTO assignment records or assignment documents attached to this form in accordance with 37 CFR 3.73(c).

Please change the correspondence address for the application identified in the attached statement under 37 CFR 3.73(c) to:

☒ The address associated with Customer Number: **20350**

OR

☐ Firm or Individual Name			
Address			
City		State	Zip
Country			
Telephone		Email	

Assignee Name and Address:
The Chinese University of Hong Kong
Knowledge Transfer Office
Pi Chiu Building
Shatin, New Territories, Hong Kong SAR
People's Republic of China

A copy of this form, together with a statement under 37 CFR 3.73(c) (Form PTO/SB/96 or equivalent) is required to be filed in each application in which this form is used. The statement under 37 CFR 3.73(c) may be completed by one of the practitioners appointed in this form, and must identify the application in which this Power of Attorney is to be filed.

SIGNATURE of Assignee of Record
The individual whose signature and title is supplied below is authorized to act on behalf of the assignee

Signature		Date	26 Feb 2014
Name	Mr. TS CHAN	Telephone	(852) 3943 1578
Title	Acting Director Knowledge Transfer Office, The Chinese University of Hong Kong		

This collection of information is required by 37 CFR 1.31, 1.32 and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**
If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2. #

PATENT APPLICATION

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

Inventors:

Yuk-Ming Dennis Lo, a citizen of United Kingdom, residing at
Homantin, Kowloon, Hong Kong SAR China

Rossa Wai Kwun Chiu, a citizen of Australia, residing at
Shatin, New Territories, Hong Kong SAR China

Kwan Chee Chan, a citizen of Hong Kong, residing at
Mei Foo Sun Chuen, Kowloon, Hong Kong SAR China

Peiyong Jiang, a citizen of China, residing at
Pak Shek Kok, Tai Po, New Territories, Hong Kong SAR China

Suk Hang Cheng, a citizen of Hong Kong, residing at
Fanling, New Territories, Hong Kong SAR China

Wenlei Peng, a citizen of the China, residing at
Shatin, New Territories, Hong Kong SAR China

On Yee Tse, a citizen of Hong Kong, residing at
Fanling, New Territories, Hong Kong SAR China

Assignee: The Chinese University of Hong Kong
Office of Research and Knowledge Transfer Services
Room 301, Pi Ch'iu Building
Shatin, New Territories Hong Kong

Entity: Large

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] The present application is a continuation application of U.S. Patent Application No. 16/995,607, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on August 17, 2020, which claims the benefit of priority to U.S. Provisional Application No. 63/051,210, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on July 13, 2020; U.S. Provisional Application No. 63/019,790, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on May 4, 2020; U.S. Provisional Application No. 62/991,891, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on March 19, 2020; U.S. Provisional Application No. 62/970,586, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on February 5, 2020; and U.S. Provisional Application No. 62/887,987, entitled "DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS," filed on August 16, 2019, the entire contents of all of which are herein incorporated by reference for all purposes.

REFERENCE TO A "SEQUENCE LISTING" SUBMITTED AS ASCII TEXT FILES VIA EFS-WEB

[0002] The instant application contains a Sequence Listing which has been filed electronically in ASCII format and is hereby incorporated by reference in its entirety. Said ASCII copy, created on May 24, 2021, is named 080015-028420US-1249377_SL.txt and is 1,709 bytes in size.

BACKGROUND

[0003] The existence of base modifications in nucleic acids varies throughout different organisms including viruses, bacteria, plants, fungi, nematodes, insects, and vertebrates (e.g. humans), etc. The most common base modifications are the addition of a methyl group to different DNA bases at different positions, so-called methylation. Methylation has been found on

cytosines, adenines, thymines and guanines, such as 5mC (5-methylcytosine), 4mC (N4-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine), 5caC (5-carboxylcytosine), 1mA (N1-methyladenine), 3mA (N3-methyladenine), 7mA (N7-methyladenine), 3mC (N3-methylcytosine), 2mG (N2-methylguanine), 6mG (O6-methylguanine), 7mG (N7-methylguanine), 3mT (N3-methylthymine), and 4mT (O4-methylthymine). In vertebrate genomes, 5mC is the most common type of base methylation, followed by that for guanine (i.e. in the CpG context).

[0004] DNA methylation is essential for mammalian development and has notable roles in gene expression and silencing, embryonic development, transcription, chromatin structure, X chromosome inactivation, protection against activity of the repetitive elements, maintenance of genomic stability during mitosis, and the regulation of parent-of-origin genomic imprinting.

[0005] DNA methylation plays many important roles in the silencing of promoters and enhancers in a coordinated manner (Robertson, 2005; Smith and Meissner, 2013). Many human diseases have been found to be associated with aberrations of DNA methylation, including but not limited to the process of carcinogenesis, imprinting disorders (e.g. Beckwith–Wiedemann syndrome and Prader–Willi syndrome), repeat-instability diseases (e.g. fragile X syndrome), autoimmune disorders (e.g. systemic lupus erythematosus), metabolic disorders (e.g. type I and type II diabetes), neurological disorders, aging, etc.

[0006] The accurate measurement of methylomic modification on DNA molecules would have numerous clinical implications. One widely used method to measure DNA methylation is through the use of bisulfite sequencing (BS-seq) (Lister *et al.*, 2009; Frommer *et al.*, 1992). In this approach, DNA samples are first treated with bisulfite which converts unmethylated cytosine (i.e. C) to uracil. In contrast, the methylated cytosine remains unchanged. The bisulfite modified DNA is then analyzed by DNA sequencing. In another approach, following bisulfite conversion, the modified DNA is then subjected to polymerase chain reaction (PCR) amplification using primers that can differentiate bisulfite converted DNA of different methylation profiles (Herman *et al.*, 1996). This latter approach is called methylation-specific PCR.

[0007] One disadvantage of such bisulfite-based approaches is that the bisulfite conversion step has been reported to significantly degrade the majority of the treated DNA (Grunau, 2001). Another disadvantage is that the bisulfite conversion step would create strong CG biases (Olova *et al.*, 2018), resulting in the reduction of signal-to-noise ratios typically for DNA mixtures with heterogeneous methylation states. Furthermore, bisulfite sequencing would not be able to sequence long DNA molecules because of the degradation of DNA during bisulfite treatment. Thus, there is a need to determine the modification of bases of nucleic acids, without prior chemical (e.g. bisulfite conversion) and nucleic acid amplification (e.g. using the PCR).

BRIEF SUMMARY

[0008] We have developed a new method that, in one embodiment, allows the determination of base modifications, such as 5mC in nucleic acids without template DNA pre-treatment such as enzymatic and/or chemical conversions, or protein and/or antibody binding. While such template DNA pre-treatment is not necessary for the determination of the base modifications, in examples that are shown, certain pre-treatment (e.g. digestion with restriction enzymes) may serve to enhance aspects of the invention (e.g. allowing the enrichment of CpG sites for analysis). The embodiments present in this disclosure could be used for detecting different types of base modification, for example, including but not limited to 4mC, 5hmC, 5fC, and 5caC, 1mA, 3mA, 7mA, 3mC, 2mG, 6mG, 7mG, 3mT, and 4mT, etc. Such embodiments can make use of features derived from sequencing, such as kinetic features, that are affected by the various base modifications, as well as an identity of nucleotides in a window around a target position whose methylation status is determined.

[0009] Embodiments of the present invention can be used for, but is not limited to, single molecule sequencing. One type of single molecule sequencing is single molecule, real-time sequencing in which the progress of the sequencing of a single DNA molecule is monitored in real-time. One type of single molecule, real-time sequencing is that commercialized by Pacific Biosciences using their Single Molecule, Real-Time (SMRT) system. Methods may use the pulse width of a signal from sequencing bases, the interpulse duration (IPD) of bases, and the identity of the bases in order to detect a modification in a base or in a neighboring base. Another single

molecule system is that based on nanopore sequencing. One example of a nanopore sequencing system is that commercialized by Oxford Nanopore Technologies.

[0010] The methods we have developed can serve as tools to detect base modifications in biological samples to assess the methylation profiles in the samples for various purposes including but not limited to research and diagnostic purposes. The detected methylation profiles can be used for different analysis. The methylation profiles can be used to detect the origin of DNA (e.g., maternal or fetal, tissue, bacterial, or DNA obtained from tumor cells enriched from the blood of a cancer patient). Detection of aberrant methylation profiles in tissues aids the identification of developmental disorders in individuals, identify and prognosticate tumors or malignancies.

[0011] Embodiments of the present invention may include analyzing the relative methylation levels of haplotypes of an organism. An imbalance in the methylation levels between the two haplotypes may be used to determine a classification of a disorder. A higher imbalance may indicate the presence of a disorder or a more severe disorder. The disorder may include cancer.

[0012] Methylation patterns in a single molecule can identify chimera and hybrid DNA. Chimeric and hybrid molecules may include sequences from two different genes, chromosomes, organelles (e.g. mitochondria, nucleus, chloroplasts), organisms (mammals, bacteria, viruses, etc.), and/or species. Detecting junctions of chimeric or hybrid DNA molecules may allow for detecting gene fusions for various disorders or diseases, including cancer, prenatal, or congenital disorders.

[0013] A better understanding of the nature and advantages of embodiments of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0015] **FIG. 1** illustrates SMRT sequencing of molecules carrying base modifications according to embodiments of the present invention.

[0016] **FIG. 2** illustrates SMRT sequencing of molecules carrying methylated and unmethylated CpG sites according to embodiments of the present invention.

[0017] **FIG. 3** illustrates interpulse durations and pulse width according to embodiments of the present invention.

[0018] **FIG. 4** shows an example of a measurement window of the Watson strand of DNA for detecting a base modification according to embodiments of the present invention.

[0019] **FIG. 5** shows an example of a measurement window of the Crick strand of DNA for detecting a base modification according to embodiments of the present invention.

[0020] **FIG. 6** shows an example of a measurement window by combining data from the Watson strand of DNA and its complementary Crick strand for detecting any base modification according to embodiments of the present invention.

[0021] **FIG. 7** shows an example of a measurement window by combining data from the Watson strand of DNA and its nearby region's Crick strand for detecting any base modification according to embodiments of the present invention.

[0022] **FIG. 8** shows examples of measurement windows of the Watson strand, the Crick strand, and both strands for determining methylation states at CpG sites according to embodiments of the present invention.

[0023] **FIG. 9** shows a general procedure of constructing analytical, computational, mathematical, or statistical models for classifying base modifications according to embodiments of the present invention.

[0024] **FIG. 10** shows a general procedure of classifying base modifications according to embodiments of the present invention.

[0025] **FIG. 11** shows a general procedure of constructing the analytical, computational, mathematical, or statistical models for classifying methylation states at CpG sites using samples with known methylation states of the Watson strand according to embodiments of the present invention.

[0026] **FIG. 12** shows a general procedure of classifying methylation states of the Watson strand for an unknown sample according to embodiments of the present invention.

[0027] **FIG. 13** shows a general procedure of constructing the analytical, computational, mathematical, or statistical models for classifying methylation states at CpG sites using samples with known methylation states of the Crick strand according to embodiments of the present invention.

[0028] **FIG. 14** shows a general procedure of classifying methylation states of the Crick strand for an unknown sample according to embodiments of the present invention.

[0029] **FIG. 15** shows a general procedure of constructing the statistical models for classifying methylation states at CpG sites using samples with known methylation states from both the Watson and Crick strands according to embodiments of the present invention.

[0030] **FIG. 16** shows a general procedure of classifying methylation states of an unknown sample from the Watson and Crick strands according to embodiments of the present invention.

[0031] **FIGS. 17A** and **17B** show the performance of a training dataset and a testing dataset for determining methylation according to embodiments of the present invention.

[0032] **FIGS. 18A** and **18B** show the performance of a training dataset and a testing dataset for determining methylation according to embodiments of the present invention.

[0033] **FIGS. 19A** and **19B** show the performance of a training dataset and a testing dataset at different sequencing depths for determining methylation according to embodiments of the present invention.

[0034] **FIGS. 20A** and **20B** show the performance of a training dataset and a testing dataset for different strands for determining methylation according to embodiments of the present invention.

[0035] **FIGS. 21A** and **21B** show the performance of a training dataset and a testing dataset for different measurement windows for determining methylation according to embodiments of the present invention.

[0036] **FIGS. 22A** and **22B** show the performance of a training dataset and a testing dataset for different measurement windows using downstream bases only for determining methylation according to embodiments of the present invention.

[0037] **FIGS. 23A** and **23B** show the performance of a training dataset and a testing dataset for different measurement windows using upstream bases only for determining methylation according to embodiments of the present invention.

[0038] **FIG. 24** shows the performance of methylation analysis using kinetic patterns associated with downstream and upstream bases using asymmetric flanking sizes in the training dataset according to embodiments of the present invention.

[0039] **FIG. 25** shows the performance of methylation analysis using kinetic patterns associated with downstream and upstream bases using asymmetric flanking sizes in the testing dataset according to embodiments of the present invention.

[0040] **FIG. 26** shows the relative importance of features in regard to classification of methylation states at CpG sites according to embodiments of the present invention.

[0041] **FIG. 27** shows the performance of the motif-based IPD analysis for methylation detection without using the pulse width signal according to embodiments of the present invention.

[0042] **FIG. 28** is a graph of a principal component analysis technique using 2-nt upstream and 6-nt downstream of a cytosine that is subjected to methylation analysis according to embodiments of the present invention.

[0043] **FIG. 29** is a graph of a performance comparison between a method using principal component analysis and a method using a convolutional neural network according to embodiments of the present invention.

[0044] **FIGS. 30A** and **30B** show the performance of a training dataset and a testing dataset for different analytical, computational, mathematical, or statistical models using upstream bases only for determining methylation according to embodiments of the present invention.

[0045] **FIG. 31A** shows an example of one approach for generating molecules with unmethylated adenines by whole genome amplification according to embodiments of the present invention. FIG. discloses SEQ ID NO: 4.

[0046] **FIG. 31B** shows an example of one approach for generating molecules with methylated adenines by whole genome amplification according to embodiments of the present invention. FIG. discloses SEQ ID NO: 5.

[0047] **FIGS. 32A** and **32B** show interpulse duration (IPD) values across sequenced A bases in template DNA of the Watson strand between unmethylated and methylated datasets according to embodiments of the present invention.

[0048] **FIG. 32C** shows a receiver operating characteristic curve for determining methylation in the Watson strand according to embodiments of the present invention.

[0049] **FIGS. 33A** and **33B** show interpulse duration (IPD) values across sequenced A bases in template DNA of the Crick strand between unmethylated and methylated datasets according to embodiments of the present invention.

[0050] **FIG. 33C** shows a receiver operating characteristic curve for determining methylation in the Crick strand according to embodiments of the present invention.

[0051] **FIG. 34** illustrates 6mA determination of the Watson strand according to embodiments of the present invention.

[0052] **FIG. 35** illustrates 6mA determination of the Crick strand according to embodiments of the present invention.

[0053] **FIGS. 36A** and **36B** show the determined probability of being methylated for sequenced A bases of the Watson strand between uA and mA datasets using a measurement window based convolutional neural network model according to embodiments of the present invention.

[0054] **FIG. 37** shows an ROC curve for the detection of 6mA using a measurement window based CNN model for sequenced A bases of the Watson strand according to embodiments of the present invention.

[0055] **FIG. 38** shows a performance comparison between IPD-metric-based 6mA detection and a measurement window based 6mA detection according to embodiments of the present invention.

[0056] **FIGS. 39A** and **39B** show the determined probability of being methylated for those sequenced A bases of the Crick strand between uA and mA datasets using measurement window based CNN model according to embodiments of the present invention.

[0057] **FIG. 40** shows the performance of 6mA detection using measurement window based CNN model on sequenced A bases of the Crick strand according to embodiments of the present invention.

[0058] **FIG. 41** shows examples of methylation states across A bases in a molecule including the Watson and Crick strands according to embodiments of the present invention.

[0059] **FIG. 42** shows an example of enhanced training by selectively using A bases in mA dataset with IPD values greater than its 10^{th} percentile according to embodiments of the present invention.

[0060] **FIG. 43** is a graph of the percentages of unmethylated adenines in the mA dataset versus the number of subreads in each well according to embodiments of the present invention.

[0061] **FIG. 44** shows methyladenine patterns between the Watson and Crick strands of a double-stranded DNA molecule in a testing dataset according to embodiments of the present invention.

[0062] **FIG. 45** is a table showing the percentage of fully-unmethylated molecules, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns in training and testing datasets according to embodiments of the present invention.

[0063] **FIG. 46** illustrates representative examples for molecules with fully-unmethylated molecules regarding adenine sites, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns according to embodiments of the present invention.

[0064] **FIG. 47** shows an example of a long read (6,265 bp) harboring a CpG island (as shaded in yellow) according to embodiments of the present invention.

[0065] **FIG. 48** is a table showing that the 9 DNA molecules were sequenced by Pacific Biosciences SMRT sequencing and which overlapped with imprinted regions according to embodiments of the present invention.

[0066] **FIG. 49** shows an example of a genomic imprinting according to embodiments of the present invention.

[0067] **FIG. 50** shows an example for the determination of methylation patterns in an imprinted region according to embodiments of the present invention.

[0068] **FIG. 51** shows a comparison of methylation levels deduced by between the new approach and conventional bisulfite sequencing according to embodiments of the present invention.

[0069] **FIGS. 52A** and **52B** show the performance of detection of methylation of plasma DNA according to embodiments of the present invention. (A) The relationship between the predicted probability of methylation versus the ranges of methylation levels quantified by bisulfite sequencing. (B) The correlation between the methylation levels determined by Pacific Biosciences (PacBio) sequencing according to the embodiments present in this disclosure (y-axis) and methylation levels quantified by bisulfite sequencing (x-axis) in 10-Mb resolution.

[0070] **FIG. 53** shows a correlation of the genomic presentation (GR) of the Y chromosome between Pacific Biosciences SMRT sequencing and BS-seq according to embodiments of the present invention.

[0071] **FIG. 54** shows an example of CpG block based detection of methylation using CpG blocks each of which harbor a series of CpG sites according to embodiments of the present invention. 5mC: methylation; C: unmethylation.

[0072] **FIGS. 55A** and **55B** show training and testing of methylation calling for human DNA molecules using the CpG block-based approach according to embodiments of the present invention. (A) Performance in the training dataset. (B) Performance in an independent testing dataset.

[0073] **FIGS. 56A** and **56B** show copy number changes in tumor tissue according to embodiments of the present invention.

[0074] **FIGS. 57A** and **57B** show copy number changes in tumor tissue according to embodiments of the present invention.

[0075] **FIG. 58** shows a schematic illustration of plasma DNA tissue mapping from the plasma of a pregnant woman using the methylation levels deduced according to embodiments of the present invention.

[0076] **FIG. 59** shows a correlation between the placental contribution to maternal plasma DNA deduced and fetal DNA fraction deduced by Y chromosome reads according to embodiments of the present invention.

[0077] **FIG. 60** shows a table summarizing sequencing data from different human tissue DNA samples according to embodiments of the present invention.

[0078] **FIG. 61** shows an illustration of various ways of analyzing methylation patterns according to embodiments of the present invention.

[0079] **FIGS. 62A** and **62B** show a comparison of methylation densities at the whole-genome level quantified by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0080] **FIGS. 63A**, **63B**, and **63C** show different correlations of overall methylation levels quantified by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0081] **FIGS. 64A** and **64B** show methylation patterns at 1-Mnt resolution for a hepatocellular carcinoma (HCC) cell line and a buffy coat sample from a health control subject with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0082] **FIGS. 65A** and **65B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for an HCC cell line (HepG2) and buffy coat sample from a healthy control subject.

[0083] **FIGS. 66A** and **66B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC cell line (HepG2) and buffy coat sample from a healthy control subject.

[0084] **FIGS. 67A** and **67B** show methylation patterns at 1-Mnt resolution for an HCC tumor tissue and adjacent normal tissue with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0085] **FIGS. 68A** and **68B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0086] **FIGS. 69A** and **69B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0087] **FIGS. 70A** and **70B** show methylation patterns at 1-Mnt resolution for an HCC tumor tissue and adjacent normal tissue with methylation levels determined by bisulfite sequencing and by single molecule, real-time sequencing according to embodiments of the present invention.

[0088] **FIGS. 71A** and **71B** show scatter plots of methylation levels at a 1-Mnt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0089] **FIGS. 72A** and **72B** show scatter plots of methylation levels at a 100-knt resolution determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention for HCC tumor tissue and adjacent normal tissue.

[0090] **FIG. 73** shows an example of the aberrant pattern of methylation nearby the tumor suppressor gene *CDKN2A* according to embodiments of the present invention.

[0091] **FIGS. 74A** and **74B** show differential methylation regions detected by single molecule, real-time sequencing according to embodiments of the present invention.

[0092] **FIG. 75** shows methylation patterns of hepatitis B virus DNA between HCC tissues and adjacent non-tumoral tissues using single molecule, real-time sequencing according to embodiments of the present invention.

[0093] **FIG. 76A** shows methylation levels of hepatitis B virus DNA in liver tissues from patients with cirrhosis but without HCC using bisulfite sequencing according to embodiments of the present invention.

[0094] **FIG. 76B** shows methylation levels of hepatitis B virus DNA in HCC tissues using bisulfite sequencing according to embodiments of the present invention.

[0095] **FIG. 77** illustrates methylation haplotype analysis according to embodiments of the present invention.

[0096] **FIG. 78** shows the size distribution of the sequenced molecules determined from consensus sequences according to embodiments of the present invention.

[0097] **FIGS. 79A**, **79B**, **79C**, and **79D** show examples of allelic methylation patterns in the imprinted regions according to embodiments of the present invention.

[0098] **FIGS. 80A**, **80B**, **80C**, and **80D** show examples of allelic methylation patterns in non-imprinted regions according to embodiments of the present invention.

[0099] **FIG. 81** shows a table of methylation levels of allele-specific fragments according to embodiments of the present invention.

[0100] **FIG. 82** shows an example for determining the placental origin of plasma DNA in pregnancy using methylation profiles according to embodiments of the present invention.

[0101] **FIG. 83** illustrates fetal-specific DNA methylation analysis according to embodiments of the present invention.

[0102] **FIGS. 84A**, **84B**, and **84C** show the performance of different measurement window sizes across different reagent kits for SMRT-seq according to embodiments of the present invention.

[0103] **FIGS. 85A**, **85B**, and **85C** show the performance of different measurement window sizes across different reagent kits for SMRT-seq according to embodiments of the present invention.

[0104] **FIGS. 86A**, **86B**, and **86C** show the correlation of overall methylation levels quantified by bisulfite sequencing and SMRT-seq (Sequel II Sequencing Kit 2.0) according to embodiments of the present invention.

[0105] **FIGS. 87A** and **87B** show a comparison of overall methylation level between various tumor tissues and paired adjacent non-tumoral tissues according to embodiments of the present invention.

[0106] **FIG. 88** shows determining the methylation status using a sequence context determined from a circular consensus sequence (CCS) according to embodiments of the present invention.

[0107] **FIG. 89** shows an ROC curve for the detection of methylated CpG sites using a sequence context determined from CCS according to embodiments of the present invention.

[0108] **FIG. 90** shows an ROC curve for the detection of methylated CpG sites without CCS information and without prior alignment to a reference genome according to embodiments of the present invention.

[0109] **FIG. 91** shows an example of preparing molecules for single molecule, real-time sequencing according to embodiments of the present invention.

[0110] **FIG. 92** shows an illustration of the CRISPR/Cas9 system according to embodiments of the present invention.

[0111] **FIG. 93** shows an example of a Cas9 complex for introducing two cuts spanning an end-blocked molecule of interest according to embodiments of the present invention.

[0112] **FIG. 94** shows methylation distribution of Alu regions determined by bisulfite sequencing and single molecule, real-time sequencing according to embodiments of the present invention.

[0113] **FIG. 95** shows the distribution of methylation levels of Alu regions determined by model using results from single molecule, real-time sequencing according to embodiments of the present invention.

[0114] **FIG. 96** shows a table of tissues and the methylation levels of Alu regions in the tissues according to embodiments of the present invention.

[0115] **FIG. 97** shows clustering analysis for different cancer types using methylation signals related to Alu repeats according to embodiments of the present invention.

[0116] **FIG. 98A** and **98B** show the effect of read depth on the overall methylation level quantification in the testing datasets which were involved with the whole genome amplification and M.SsssI treatment according to embodiments of the present invention.

[0117] **FIG. 99** shows a comparison between overall methylation levels determined by SMRT-seq (Sequel II Sequencing Kit 2.0) and BS-seq with the use of different subread depth cutoffs according to embodiments of the present invention.

[0118] **FIG. 100** is a table showing the effect of the subread depth on the correlation of methylation levels between two measurements by SMRT-seq (Sequel II Sequencing Kit 2.0) and BS-seq according to embodiments of the present invention.

[0119] **FIG. 101** shows the subread depth distribution with respect to fragment sizes in data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0120] **FIG. 102** shows a method of detecting a modification of a nucleotide in a nucleic acid molecule according to embodiments of the present invention.

[0121] **FIG. 103** shows a method for detecting a modification of a nucleotide in a nucleic acid molecule according to embodiments of the present invention.

[0122] **FIG. 104** illustrates relative haplotype-based methylation imbalance analysis according to embodiments of the present invention.

[0123] **FIG. 105A** and **105B** are a table of the haplotype blocks showing differential methylation levels between Hap I and Hap II in the tumor DNA compared with the adjacent non-tumoral tissue DNA for the case TBR3033 according to embodiments of the present invention.

[0124] **FIG. 106** is a table of the haplotype blocks showing differential methylation levels between Hap I and Hap II in the tumor DNA compared with the adjacent normal tissue DNA for the case TBR3032 according to embodiments of the present invention.

[0125] **FIG. 107A** is a table summarizing the number of haplotype blocks showing methylation imbalance between two haplotypes between tumor and adjacent nontumoral tissues on the basis of data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0126] **FIG. 107B** is a table summarizing the number of haplotype blocks showing methylation imbalance between two haplotypes in tumor tissues for different tumor stages on the basis of data generated by Sequel II Sequencing Kit 2.0 according to embodiments of the present invention.

[0127] **FIG. 108** illustrates relative haplotype-based methylation imbalance analysis according to embodiments of the present invention.

[0128] **FIG. 109** shows a method of classifying a disorder in an organism having a first haplotype and a second haplotype according to embodiments of the present invention.

[0129] **FIG. 110** illustrates creating human-mouse hybrid fragments for which the human part is methylated while the mouse part is unmethylated according to embodiments of the present invention.

[0130] **FIG. 111** illustrates creating human-mouse hybrid fragments for which the human part is unmethylated while the mouse part is methylated according to embodiments of the present invention.

[0131] **FIG. 112** shows the length distribution of DNA molecules in a DNA mixture (sample MIX01) after ligation according to embodiments of the present invention.

[0132] **FIG. 113** illustrates a junction region by which a first DNA (A) and a second DNA (B) are joined together according to embodiments of the present invention.

[0133] **FIG. 114** illustrates methylation analysis for the DNA mixture according to embodiments of the present invention.

[0134] **FIG. 115** shows a boxplot of the probabilities of being methylated for CpG sites in the sample MIX01 according to embodiments of the present invention.

[0135] **FIG. 116** shows length distribution of DNA molecules in the DNA mixture after cross-ligation of sample MIX02 according to embodiments of the present invention.

[0136] **FIG. 117** shows a boxplot of the probabilities of being methylated for CpG sites in the sample MIX02 according to embodiments of the present invention.

[0137] **FIG. 118** is a table comparing methylation determined by bisulfite sequencing and Pacific Biosciences sequencing for MIX01 according to embodiments of the present invention.

[0138] **FIG. 119** is a table comparing methylation determined by bisulfite sequencing and Pacific Biosciences sequencing for MIX02 according to embodiments of the present invention.

[0139] **FIGS. 120A** and **120B** show methylation levels in 5-Mb bins for human-only and mouse-only DNA for MIX01 and MIX02 according to embodiments of the present invention.

[0140] **FIGS. 121A** and **121B** show methylation levels in 5-Mb bins for the human part and the mouse part of human-mouse hybrid DNA fragments for MIX01 and MIX02 according to embodiments of the present invention.

[0141] **FIGS. 122A** and **122B** are representative graphs showing methylation states in a single human-mouse hybrid molecule according to embodiments of the present invention.

[0142] **FIG. 123** shows a method of detecting chimeric molecules in a biological sample according to embodiments of the present invention.

[0143] **FIG. 124** illustrates a measurement system according to embodiments of the present invention.

[0144] **FIG. 125** shows a block diagram of an example computer system usable with systems and methods according to embodiments of the present invention.

[0145] **FIG. 126** shows *Msp*I-based targeted single molecule, real-time sequencing with the use of DNA end repair and A-tailing according to embodiments of the present invention.

[0146] **FIGS. 127A** and **127B** show the size distribution of *Msp*I-digested fragments according to embodiments of the present invention.

[0147] **FIG. 128** shows a table with the number of DNA molecules for certain selected size ranges according to embodiments of the present invention.

[0148] **FIG. 129** is a graph of the percentage coverage of CpG sites within CpG islands versus size of DNA fragments after restriction enzyme digestion according to embodiments of the present invention.

[0149] **FIG. 130** shows *Msp*I-based targeted single molecule, real time sequencing without the use of DNA end repair and A-tailing according to embodiments of the present invention.

[0150] **FIG. 131** shows *Msp*I-based targeted single molecule, real time sequencing with a reduced probability of adapter self-ligation according to embodiments of the present invention.

[0151] **FIG. 132** is a graph of the overall methylation levels between placenta and buffy DNA samples determined by *Msp*I-based targeted single molecule, real-time sequencing according to the embodiments of the present invention.

[0152] **FIG. 133** shows a clustering analysis of placental and buffy coat samples using their DNA methylation profiles determined by *Msp*I-based targeted single molecule, real-time sequencing according to the embodiments of the present invention.

TERMS

[0153] A "*tissue*" corresponds to a group of cells that group together as a functional unit. More than one type of cells can be found in a single tissue. Different types of tissue may consist of different types of cells (e.g., hepatocytes, alveolar cells or blood cells), but also may correspond to tissue from different organisms (mother vs. fetus; tissues in a subject who has received transplantation; tissues of an organism that are infected by a microorganism or a virus) or to healthy cells vs. tumor cells. "Reference tissues" can correspond to tissues used to determine tissue-specific methylation levels. Multiple samples of a same tissue type from different individuals may be used to determine a tissue-specific methylation level for that tissue type.

[0154] A "*biological sample*" refers to any sample that is taken from a human subject. The biological sample can be a tissue biopsy, a fine needle aspirate, or blood cells. The sample can also be for example, plasma or serum or urine from a pregnant woman. Stool samples can also be used. In various embodiments, the majority of DNA in a biological sample from a pregnant woman that has been enriched for cell-free DNA (e.g., a plasma sample obtained via a centrifugation protocol) can be cell-free, e.g., greater than 50%, 60%, 70%, 80%, 90%, 95%, or 99% of the DNA can be cell-free. The centrifugation protocol can include, for example, 3,000 g x 10 minutes, obtaining the fluid part, and re-centrifuging at for example, 30,000 g for another 10

minutes to remove residual cells. In certain embodiments, following the 3,000 g centrifugation step, one can follow up with filtration of the fluid part (e.g. using a filter of pore size of 5 μm, or smaller, in diameter).

[0155] A "*sequence read*" refers to a string of nucleotides sequenced from any part or all of a nucleic acid molecule. For example, a sequence read may be a short string of nucleotides (e.g., 20-150) sequenced from a nucleic acid fragment, a short string of nucleotides at one or both ends of a nucleic acid fragment, or the sequencing of the entire nucleic acid fragment that exists in the biological sample. A sequence read may be obtained in a variety of ways, e.g., using sequencing techniques or using probes, e.g., in hybridization arrays or capture probes, or amplification techniques, such as the polymerase chain reaction (PCR) or linear amplification using a single primer or isothermal amplification.

[0156] A "*subread*" is a sequence generated from all bases in one strand of a circularized DNA template that has been copied in one contiguous strand by a DNA polymerase. For example, a subread can correspond to one strand of circularized DNA template DNA. In such an example, after circularization, one double-stranded DNA molecule would have two subreads: one for each sequencing pass. In some embodiments, the sequence generated may include a subset of all the bases in one strand, e.g., because of the existence of sequencing errors.

[0157] A "*site*" (also called a "*genomic site*") corresponds to a single site, which may be a single base position or a group of correlated base positions, e.g., a CpG site or larger group of correlated base positions. A "locus" may correspond to a region that includes multiple sites. A locus can include just one site, which would make the locus equivalent to a site in that context.

[0158] A "*methylation status*" refers to the state of methylation at a given site. For example, a site may be either methylated, unmethylated, or in some cases, undetermined.

[0159] The "*methylation index*" for each genomic site (e.g., a CpG site) can refer to the proportion of DNA fragments (e.g., as determined from sequence reads or probes) showing methylation at the site over the total number of reads covering that site. A "read" can correspond to information (e.g., methylation status at a site) obtained from a DNA fragment. A read can be

obtained using reagents (e.g. primers or probes) that preferentially hybridize to DNA fragments of a particular methylation status at one or more sites. Typically, such reagents are applied after treatment with a process that differentially modifies or differentially recognizes DNA molecules depending on their methylation status, e.g. bisulfite conversion, or methylation-sensitive restriction enzyme, or methylation binding proteins, or anti-methylcytosine antibodies, or single molecule sequencing techniques (e.g. single molecule, real-time sequencing and nanopore sequencing (e.g. from Oxford Nanopore Technologies)) that recognize methylcytosines and hydroxymethylcytosines.

[0160] The "*methylation density*" of a region can refer to the number of reads at sites within the region showing methylation divided by the total number of reads covering the sites in the region. The sites may have specific characteristics, e.g., being CpG sites. Thus, the "CpG methylation density" of a region can refer to the number of reads showing CpG methylation divided by the total number of reads covering CpG sites in the region (e.g., a particular CpG site, CpG sites within a CpG island, or a larger region). For example, the methylation density for each 100-kb bin in the human genome can be determined from the total number of cytosines not converted after bisulfite treatment (which corresponds to methylated cytosine) at CpG sites as a proportion of all CpG sites covered by sequence reads mapped to the 100-kb region. This analysis can also be performed for other bin sizes, e.g. 500 bp, 5 kb, 10 kb, 50-kb or 1-Mb, etc. A region could be the entire genome or a chromosome or part of a chromosome (e.g. a chromosomal arm). The methylation index of a CpG site is the same as the methylation density for a region when the region only includes that CpG site. The "proportion of methylated cytosines" can refer the number of cytosine sites, "C's", that are shown to be methylated (for example unconverted after bisulfite conversion) over the total number of analyzed cytosine residues, i.e. including cytosines outside of the CpG context, in the region. The methylation index, methylation density, count of molecules methylated at one or more sites, and proportion of molecules methylated (e.g., cytosines) at one or more sites are examples of "*methylation levels*." Apart from bisulfite conversion, other processes known to those skilled in the art can be used to interrogate the methylation status of DNA molecules, including, but not limited to enzymes sensitive to the methylation status (e.g. methylation-sensitive restriction enzymes),

methylation binding proteins, single molecule sequencing using a platform sensitive to the methylation status (e.g. nanopore sequencing (Schreiber et al. Proc Natl Acad Sci 2013; 110: 18910-18915) and by single molecule, real-time sequencing (e.g. that from Pacific Biosciences) (Flusberg et al. Nat Methods 2010; 7: 461-465)).

[0161] A "*methylome*" provides a measure of an amount of DNA methylation at a plurality of sites or loci in a genome. The methylome may correspond to all of the genome, a substantial part of the genome, or relatively small portion(s) of the genome.

[0162] A "*pregnant plasma methylome*" is the methylome determined from the plasma or serum of a pregnant animal (e.g., a human). The pregnant plasma methylome is an example of a cell-free methylome since plasma and serum include cell-free DNA. The pregnant plasma methylome is also an example of a mixed methylome since it is a mixture of DNA from different organs or tissues or cells within a body. In one embodiment, such cells are the hematopoietic cells, including, but not limited to cells of the erythroid (i.e. red cell) lineage, the myeloid lineage (e.g., neutrophils and their precursors), and the megakaryocytic lineage. In pregnancy, the plasma methylome may contain methylomic information from the fetus and the mother. The "cellular methylome" corresponds to the methylome determined from cells (e.g., blood cells) of the patient. The methylome of the blood cells is called the blood cell methylome (or blood methylome).

[0163] A "*methylation profile*" includes information related to DNA or RNA methylation for multiple sites or regions. Information related to DNA methylation can include, but not limited to, a methylation index of a CpG site, a methylation density (MD for short) of CpG sites in a region, a distribution of CpG sites over a contiguous region, a pattern or level of methylation for each individual CpG site within a region that contains more than one CpG site, and non-CpG methylation. In one embodiment, the methylation profile can include the pattern of methylation or non-methylation of more than one type of base (e.g. cytosine or adenine). A methylation profile of a substantial part of the genome can be considered equivalent to the methylome. "DNA methylation" in mammalian genomes typically refers to the addition of a methyl group to the 5' carbon of cytosine residues (i.e. 5-methylcytosines) among CpG dinucleotides. DNA

methylation may occur in cytosines in other contexts, for example CHG and CHH, where H is adenine, cytosine or thymine. Cytosine methylation may also be in the form of 5-hydroxymethylcytosine. Non-cytosine methylation, such as N^6-methyladenine, has also been reported.

[0164] A "*methylation pattern*" refers to the order of methylated and non-methylated bases. For example, the methylation pattern can be the order of methylated bases on a single DNA strand, a single double-stranded DNA molecule, or another type of nucleic acid molecule. As an example, three consecutive CpG sites may have any of the following methylation patterns: UUU, MMM, UMM, UMU, UUM, MUM, MUU, or MMU, where "U" indicates an unmethylated site and "M" indicates a methylated site. When one extends this concept to base modifications that include, but not restricted to methylation, one would use the term "*modification pattern*," which refers to the order of modified and non-modified bases. For example, the modification pattern can be the order of modified bases on a single DNA strand, a single double-stranded DNA molecule, or another type of nucleic acid molecule. As an example, three consecutive potentially modifiable sites may have any of the following modification patterns: UUU, MMM, UMM, UMU, UUM, MUM, MUU, or MMU, where "U" indicates an unmodified site and "M" indicates a modified site. One example of base modification that is not based on methylation is oxidation changes, such as in 8-oxo-guanine.

[0165] The terms "*hypermethylated*" and "*hypomethylated*" may refer to the methylation density of a single DNA molecule as measured by its single molecule methylation level, e.g., the number of methylated bases or nucleotides within the molecule divided by the total number of methylatable bases or nucleotides within that molecule. A hypermethylated molecule is one in which the single molecule methylation level is at or above a threshold, which may be defined from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%. A hypomethylated molecule is one in which the single molecule methylation level is at or below a threshold, which may be defined from application to application, and which may change from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%.

[0166] The terms "hypermethylated" and "hypomethylated" may also refer to the methylation level of a population of DNA molecules as measured by the multiple molecule methylation levels of these molecules. A hypermethylated population of molecules is one in which the multiple molecule methylation level is at or above a threshold which may be defined from application to application, and which may change from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 95%. A hypomethylated population of molecules is one in which the multiple molecule methylation level is at or below a threshold which may be defined from application to application. The threshold may be 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 95%. In one embodiment, the population of molecules may be aligned to one or more selected genomic regions. In one embodiment, the selected genomic region(s) may be related to a disease such as cancer, a genetic disorder, an imprinting disorder, a metabolic disorder, or a neurological disorder. The selected genomic region(s) can have a length of 50 nucleotides (nt), 100 nt, 200 nt, 300 nt, 500 nt, 1000 nt, 2 knt, 5 knt, 10 knt, 20 knt, 30 knt, 40 knt, 50 knt, 60 knt, 70 knt, 80 knt, 90 knt, 100 knt, 200 knt, 300 knt, 400 knt, 500 knt, or 1 Mnt.

[0167] The term "*sequencing depth*" refers to the number of times a locus is covered by a sequence read aligned to the locus. The locus could be as small as a nucleotide, or as large as a chromosome arm, or as large as the entire genome. Sequencing depth can be expressed as 50x, 100x, etc., where "x" refers to the number of times a locus is covered with a sequence read. Sequencing depth can also be applied to multiple loci, or the whole genome, in which case x can refer to the mean number of times the loci or the haploid genome, or the whole genome, respectively, is sequenced. Ultra-deep sequencing can refer to at least 100x in sequencing depth.

[0168] The term "*classification*" as used herein refers to any number(s) or other characters(s) that are associated with a particular property of a sample. For example, a "+" symbol (or the word "positive") could signify that a sample is classified as having deletions or amplifications. The classification can be binary (e.g., positive or negative) or have more levels of classification (e.g., a scale from 1 to 10 or 0 to 1).

[0169] The terms "*cutoff*" and "*threshold*" refer to predetermined numbers used in an operation. For example, a cutoff size can refer to a size above which fragments are excluded. A threshold value may be a value above or below which a particular classification applies. Either of these terms can be used in either of these contexts. A cutoff or threshold may be "a reference value" or derived from a reference value that is representative of a particular classification or discriminates between two or more classifications. Such a reference value can be determined in various ways, as will be appreciated by the skilled person. For example, metrics can be determined for two different cohorts of subjects with different known classifications, and a reference value can be selected as representative of one classification (e.g., a mean) or a value that is between two clusters of the metrics (e.g., chosen to obtain a desired sensitivity and specificity). As another example, a reference value can be determined based on statistical analyses or simulations of samples.

[0170] The term "*level of cancer*" can refer to whether cancer exists (i.e., presence or absence), a stage of a cancer, a size of tumor, whether there is metastasis, the total tumor burden of the body, the cancer's response to treatment, and/or other measure of a severity of a cancer (e.g. recurrence of cancer). The level of cancer may be a number or other indicia, such as symbols, alphabet letters, and colors. The level may be zero. The level of cancer may also include premalignant or precancerous conditions (states). The level of cancer can be used in various ways. For example, screening can check if cancer is present in someone who is not previously known to have cancer. Assessment can investigate someone who has been diagnosed with cancer to monitor the progress of cancer over time, study the effectiveness of therapies or to determine the prognosis. In one embodiment, the prognosis can be expressed as the chance of a patient dying of cancer, or the chance of the cancer progressing after a specific duration or time, or the chance or extent of cancer metastasizing. Detection can mean 'screening' or can mean checking if someone, with suggestive features of cancer (e.g. symptoms or other positive tests), has cancer.

[0171] A "*level of pathology*" (or level of disorder) can refer to the amount, degree, or severity of pathology associated with an organism, where the level can be as described above for cancer. Another example of pathology is a rejection of a transplanted organ. Other example pathologies

can include gene imprinting disorders, autoimmune attack (e.g., lupus nephritis damaging the kidney or multiple sclerosis), inflammatory diseases (e.g., hepatitis), fibrotic processes (e.g. cirrhosis), fatty infiltration (e.g. fatty liver diseases), degenerative processes (e.g. Alzheimer's disease), and ischemic tissue damage (e.g., myocardial infarction or stroke). A heathy state of a subject can be considered a classification of no pathology.

[0172] A "*pregnancy-associated disorder*" include any disorder characterized by abnormal relative expression levels of genes in maternal and/or fetal tissue. These disorders include, but are not limited to, preeclampsia, intrauterine growth restriction, invasive placentation, pre-term birth, hemolytic disease of the newborn, placental insufficiency, hydrops fetalis, fetal malformation, HELLP syndrome, systemic lupus erythematosus, and other immunological diseases of the mother.

[0173] The abbreviation "*bp*" refers to base pairs. In some instances, "bp" may be used to denote a length of a DNA fragment, even though the DNA fragment may be single stranded and does not include a base pair. In the context of single-stranded DNA, "bp" may be interpreted as providing the length in nucleotides.

[0174] The abbreviation "nt" refers to nucleotides. In some instances, "nt" may be used to denote a length of a single-standed DNA in a base unit. Also, "nt" may be used to denote the relative positions such as upstream or downstream of the locus being analyzied. In some contexts concerning technological conceptualization, data presentation, processing and analysis, "nt" and "bp" may be used interchangeably.

[0175] The term "*sequence context*" can refer to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA. Such a stretch of DNA could be surrounding a base that is subjected to or the target of base modification analysis. For example, the seqeunce context can refer to bases upstream and/or downstream of a base that is subjected to base modification analysis.

[0176] The term "*kinetic features*" can refer to features derived from sequencing, including from single molecule, real-time sequencing. Such features can be used for base modification

analysis. Example kinetic features include upstream and downstream sequence context, strand information, interpulse duration, pulse widths, and pulse strength. In single molecule, real-time sequencing, one is continuously monitoring the effects of activities of a polymerase on a DNA template. Hence, measurements generated from such a sequencing can be regarded as kinetic features, e.g., nucleotide sequences.

[0177] The term "*machine learning models*" may include models based on using sample data (e.g., training data) to make predictions on test data, and thus may include supervised learning. Machine learning models often are developed using a computer or a processor. Machine learning models may include statistical models.

[0178] The term "*data analysis framework*" may include algorithms and/or models that can take data as an input and then output a predicted result. Examples of "data analysis frameworks" include statistical models, mathematical models, machine learning models, other artificial intelligence models, and combinations thereof.

[0179] The term "*real-time sequencing*" may refer to a technique that involves data collection or monitoring during progress of a reaction involved in sequencing. For example, real-time sequencing may involve optical monitoring or filming the DNA polymerase incorporatingincorporate a new base.

[0180] The term "*about*" or "*approximately*" can mean within an acceptable error range for the particular value as determined by one of ordinary skill in the art, which will depend in part on how the value is measured or determined, i.e., the limitations of the measurement system. For example, "about" can mean within 1 or more than 1 standard deviation, per the practice in the art. Alternatively, "about" can mean a range of up to 20%, up to 10%, up to 5%, or up to 1% of a given value. Alternatively, particularly with respect to biological systems or processes, the term "about" or "approximately" can mean within an order of magnitude, within 5-fold, and more preferably within 2-fold, of a value. Where particular values are described in the application and claims, unless otherwise stated the term "about" meaning within an acceptable error range for the particular value should be assumed. The term "about" can have the meaning as commonly

understood by one of ordinary skill in the art. The term "about" can refer to ±10%. The term "about" can refer to ±5%.

DETAILED DESCRIPTION

[0181] Achieving bisulfite-free determination of a base modification, include a methylated base, is the subject of different research efforts but none have been shown to be commercially viable. Recently, a bisulfite-free method for detecting 5mC and 5hmC has been published (Y. Liu *et al.*, 2019) using a mild condition for 5mC and 5hmC base conversion. This method involves multiple steps of enzymatic and chemical reactions including ten-eleven translocation (TET) oxidation, pyridine borane reduction, and PCR. The efficiency for each step of conversion reaction as well as PCR bias would adversely affect the ultimate accuracy in 5mC analysis. For example, the 5mC conversion rate has been reported to be around 96%, with a false-negative rate of around 3%. Such performance would potentially limit one's ability to detect certain subtle changes of methylation in a genome. On the other hand, the enzymatic conversion would not be able to perform equally well across the genome. For example, the conversion rate of 5hmC was 8.2% lower than that for 5mC, and the conversion rate for non-CpG was 11.4% lower than that for CpG contexts (Y. Liu *et al.*, 2019). Thus, the ideal situation is the development of approaches for measuring base modifications of a native DNA molecule without any prior conversion (chemical or enzymatic, or combinations thereof) step and even without an amplification step.

[0182] There were a number of proof-of-concept studies (Q. Liu *et al.*, 2019; Ni *et al.*, 2019) in which the electric signals produced by a long-read nanopore sequencing approach (e.g., using the system developed by Oxford Nanopore Technologies) enabled one to detect methylation states with the use of a deep learning method. In addition to Oxford Nanopore, there are other single molecule sequencing approaches that allow long reads. One example is the single molecule, real-time sequencing. One example of a single molecule, real-time sequencing is that commercialized the Pacific Biosciences SMRT system. As the principle of single molecule, real-time sequencing (e.g., the Pacific Biosciences SMRT system) is different from that of a non-optical based nanopore system (e.g. by Oxford Nanopore Technologies), approaches for base modification detection developed for such non-optical based nanopore system cannot be used for

single molecule, real-time sequencing. For example, a non-optical nanopore system is not designed for capturing the patterns of fluorescent signals produced by immobilized DNA polymerase based DNA synthesis (employed by single molecule, real-time sequencing such as by the Pacific Biosciences SMRT system). As a further example, in the Oxford Nanopore sequencing platform, each measured electric event is associated with a k-mer (e.g., 5-mer) (Q. Liu *et al.*, 2019). However, in the Pacific Biosciences SMRT sequencing platform, each fluorescent event is generally associated with a single incorporated base. Furthermore, a single DNA molecule would be sequenced multiple times in Pacific Biosciences SMRT sequencing including Watson and Crick strands. Conversely, for the Oxford Nanopore long-read sequencing approach, sequence readout is performed once for each of the Watson and Crick strands.

[0183] It has been reported that the polymerase kinetics would be affected by methylation states in the sequences of *E. coli* (Flusberg *et al.*, 2010). Previous studies showed that when compared with the detection of 6mA, 4mC, 5hmC, and 8-oxo-guanine, it is much more challenging to use the polymerase kinetics of single molecule, real-time sequencing for deducing the methylation states (5mC versus C) of a particular CpG in a single molecule. The reason is that the methyl group is small and oriented towards the major groove and is not involved in base pairing, leading to very subtle interruption in the kinetics caused by 5mC (Clark *et al.*, 2013). Hence, there is a paucity of approaches for determining the methylation states of cytosines at the single-molecule level.

[0184] Suzuki et al developed an algorithm (Suzuki *et al.*, 2016) attempting to combine the interpulse duration (IPD) ratios for neighboring CpG sites to increase the confidence in identifying the methylation states of those sites. However, this algorithm only allowed one to predict a genomic region of being completely methylated or completely unmethylated, but lacked the ability to determine intermediate methylation patterns.

[0185] Regarding single molecule, real-time sequencing, current approaches only used one or two parameters independently, achieving a very limited accuracy in detecting 5mC because of the measurement difference between 5-methylcytosine and cytosine. For example, Flusberg *et al.* demonstrated that IPD was altered in base modifications including N6-methyladenosine, 5-

methylcytosine, and 5-hydroxymethylcytosine. However, pulse width (PW) of the sequencing kinetics was not found to have a significant effect. Hence, in the method they used for predicting base modification, using the detection of N6-methyladenosine as an example, only IPD but not PW was used.

[0186] In follow-up publications by the same group (Clark *et al.*, 2012; Clark *et al.* 2013), IPD but not PW was incorporated in the algorithms for the detection of 5-methylcytosine. In Clark *et al.* 2012, the detection rate of 5-methylcytosine without converting it to 5-methylcytosine only ranged from 1.9% to 4.3%. Furthermore, in Clark *et al.* 2013, the authors had further reaffirmed the subtlety of the kinetic signature of 5-methylcytosine. To overcome the low sensitivity of detecting 5-methylcytosine, Clark *et al.* further developed a method which converted 5-methylcytosine to 5-carboxylmethylcytosine using Ten-eleven translocation (Tet) proteins so as to improve the sensitivity of 5-methylcytosine (Clark *et al.* 2013) because the alteration of IPD caused by 5-carboxylcytosine was much more than by 5-methylcytosine.

[0187] In a more recent report by Blow *et al.,* the IPD ratio-based method previously described by Flusberg *et al* was used to detect the base modifications in 217 bacterial and 13 archaeal species with 130-fold read coverage per organism (Blow *et al.*, 2016). Among all the base modifications they identified, only 5% involved 5-methylcytosine. They attributed this low detection rate of 5-methylcytosine to the low sensitivity of single-molecule real-time sequencing for detecting 5-methylcytosine. In most bacteria, a set of sequence motifs were targeted by DNA methyltransferases (MTases) for methylation (e.g. 5'-GmATC-3' by Dam or 5'-CmCWGG-3' by Dcm in *E. coli*) at nearly *all* of these motifs in the genome, with only a small fraction of these motif sites remaining non-methylated (Beaulaurier *et al.* 2019). Furthermore, the use of the IPD-based method to classify the methylation status of the second C in the 5'-CCWGG-3' motif with, or without treatment, with Tet proteins yielded detection rates of 5-methylcytosine of 95.2% and 1.9%, respectively (Clark *et al.* 2013). Taken as a whole, the IPD method without prior base conversion (e.g., using Tet proteins) missed the majority of 5-methylcytosine.

[0188] In the studies mentioned above (Clark *et al.*, 2012; Clark *et al.*, 2013; Blow *et al.*, 2016), IPD-based algorithms were used without consideration of the sequence context at where

the candidate base modification was located. Other groups have attempted to take into account the sequence context of a nucleotide for the detection of base modification. For example, Feng *et al.* used a hierarchical model to analyze IPDs for the detection of 4-methylcytosine and 6-methyladenosine in a respective sequence context (Feng *et al.* 2013). However, in their method, they only considered the IPD at the base of interest and the sequence context adjacent to that base, but did not use the IPD information of all neighboring bases adjacent to the base of interest. In addition, PW was not considered in the algorithm, and they did not present any data on the detection of 5-methylcytosine.

[0189] In another study, Schadt *et al.* developed a statistical method, called conditional random field, to analyze the IPD information of the base of interest and the neighboring bases to determine if the base of interest was a 5-methylcytosine (Schadt *et al.*, 2012). In this work, they also considered the IPD interaction between these bases by inputting them into an equation. However, they did not input the nucleotide sequence, namely A, T, G, or C, in their equation. When they applied the method to determine the methylation status of the M.Sau3AI plasmid, the area under ROC curve was close to 0.5 even at an 800-fold sequence coverage of the plasmid sequence. Moreover, in their method, they had not taken into account PW in their analysis.

[0190] In yet another study by Beckman *et al.*, they compared the IPD of all sequences that shared the same 4-nt or 6-nt motif in the genome between a target bacterial genome and a completely unmethylated genome, e.g., obtained through whole genome amplification (Beckman *et al.* 2014). The purpose of such analysis was only to identify motifs that would be more frequently affected by base modifications. In the study, they only considered the IPD of a potentially modified base but not the IPD of the neighboring base or PW. Their method was not informative about the methylation status of individual nucleotide.

[0191] In summary, these previous attempts of utilizing IPD only or with combination of sequence information in the neighboring nucleotides for grouping data were not able to determine the base modification of 5-methylcytosine with meaningful or practical accuracy. In a recent review by Gouil *et al.*, the authors concluded that because of the low signal-to-noise ratio, the detection of 5-methylcytosine in a single molecule using single-molecule real-time

sequencing is inaccurate (Gouil *et al.*, 2019). In these previous studies, it remains unknown as to whether it may be feasible to use the kinetic features for genomewide methylomic analysis, especially for complex genomes such as human genomes, cancer genomes, or fetal genomes.

[0192] In contrast to previous studies, some embodiments of methods described in this disclosure are based on measuring and utilizing IPD, PW, and sequence context for every base within the measurement window. We reasoned that if we can use a combination of multiple metrics, for example, concurrently making use of features including upstream and downstream sequence context, strand information, IPD, pulse widths as well as pulse strength, we might be able to achieve the accurate measurement of base modifications (e.g. mC detection) at single-base resolution. Sequence context refers to the base compositions (A, C, G, or T) and the base orders in a stretch of DNA. Such a stretch of DNA could be surrounding a base that is subjected to or the target of base modification analysis. In one embodiment, the stretch of DNA could be proximal to a base that is subjected to base modification analysis. In another embodiment, the stretch of DNA could be far away from a base that is subjected to base modification analysis. The stretch of DNA could be upstream and/or downstream of a base that is subjected to base modification analysis.

[0193] In one embodiment, the features of upstream and downstream sequence context, strand information, IPD, pulse widths as well as pulse strength, which are used for base modification analysis, are referred to as kinetic features.

[0194] The embodiments present in this disclosure can be used for DNA obtained from, but not limited to, cell lines, samples from an organism (e.g. solid organs, solid tissues, a sample obtained via endoscopy, blood, or plasma or serum or urine from a pregnant woman, chorionic villus biopsy, etc.), samples obtained from the environment (e.g. bacteria, cellular contaminants), food (e.g., meat). In some embodiments, the methods present in this disclosure can also be applied following a step in which a fraction of the genome is first enriched, e.g. using hybridization probes (Albert *et al.*, 2007; Okou *et al.*, 2007; Lee *et al.*, 2011), or approaches based on physical separation (e.g. based on sizes, etc) or following restriction enzyme digestion (e.g. MspI), or Cas9-based enrichment (Watson *et al.*, 2019). While the invention does not

require enzymatic or chemical conversion to work, in certain embodiments, such a conversion step can be included to further enhance the performance of the invention.

[0195] Embodiments of the present disclosure allow for improved accuracy or practicality or convenience in detecting base modifications or measuring modification levels. The modification may be detected directly. Embodiments may avoid enzymatic or chemical conversion, which may not preserve all modification information for detection. Additionally, certain enzymatic or chemical conversions may not be compatible with certain types of modifications. Embodiments of the present disclosure may also avoid amplification by PCR, which may not transfer base-modification information to the PCR products. Additionally, both strands of DNA may be sequenced together, thereby enabling the pairing of the sequence from one strand with its complementary sequence to the other strand. By contrast, PCR amplification splits the two strands of double-stranded DNA, so such pairing of sequences is difficult.

[0196] Methylation profiles, determined with or without enzymatic or chemical conversion, can be used for analyzing biological samples. In one embodiment, the methylation profiles can be used to detect the origin of cellular DNA (e.g., maternal or fetal, tissue, viral, or tumor). Detection of aberrant methylation profiles in tissues aid the identification of developmental disorders in individuals and the identification and prognostication of tumors or malignancies. Imbalances in methylation levels between hapltoypes can be used to detect disorders, including cancer. Methylation patterns in a single molecule can identify chimeric (e.g., between a virus and human) and hybrid DNA, (e.g., between two genes normally unfused in a natural genome); or between two species (e.g., through genetic or genomic manipulation).

[0197] Methylation analysis may be improved by enhanced training, which may include narrowing the data used in a training set. Specific regions may be targeted for analysis. In embodiments, such targeting can involve an enzyme that either alone, or in combination with other reagent(s), may cleave a DNA sequence or a genome based on its sequence. In some embodiments the enzyme is a restriction enzyme that recognizes and cleaves a specific DNA sequence(s). In other embodiments, more than one restriction enzymes with different recognition sequences can be used in combination. In some embodiments, the restriction enzyme may cleave

or not cleave based on the methylation status of the recognition sequences. In some embodiments, the enzyme is one within the CRISPR/Cas family. For example, genomic regions of interest can be targeted using a CRISPR/Cas9 system or other system based on guide RNA (i.e., short RNA sequences which bind to a complementary target DNA sequences and in the process guides an enzyme to act at a target genomic location). In some instances, methylation analysis may be possible without alignment to a reference genome.

I. METHYLATION DETECTION WITH SINGLE MOLECULE, REAL-TIME SEQUENCING

[0198] Embodiments of the present disclosure allow for directly detecting base modifications, without enzymatic or chemical conversion. Kinetic features (e.g., sequence context, IPD, and PW) obtained through single molecule, real-time sequencing can be analyzed with machine learning to develop a model to detect modification or the absence of a modification. Modification levels may be used to determine the origin of DNA molecules or the presence or level of the disorder.

[0199] Using Pacific Biosciences SMRT sequencing as an example of single molecule, real-time sequencing for illustration purposes, a DNA polymerase molecule is positioned at the bottom of wells that serve as zero-mode waveguides (ZMW). The ZMW is a nanophotonic device for confining light to a small observation volume, which can be a hole whose diameter is very small and does not allow the propagation of light in the wavelength range used for detection such that only emission of optical signals from dye-labeled nucleotide incorporated by the immobilized polymerase are detectable against a low and constant background signal (Eid *et al.*, 2009). The DNA polymerase catalyzes the incorporation of fluorescently labeled nucleotides into complementary nucleic acid strands.

[0200] **FIG. 1** shows an example of molecules carrying base modifications that were sequenced by single-molecule circular consensus sequencing. Molecules 102, 104, and 106 carry base modifications. DNA molecules (e.g., molecule 106) may be ligated with hairpin adapters to form ligated molecule 108. Ligated molecule 108 can then form circularized molecule 110. The

circularized molecules can bind to immobilized DNA polymerase and can initiate the DNA synthesis. Molecules not carrying base modifications can also be sequenced.

[0201] **FIG. 2** shows an example of molecules carrying methylated and/or unmethylated CpG sites that were sequenced by single molecule, real-time sequencing. DNA molecules were first ligated with hairpin adapters to form circularized molecules which would bind to immobilized DNA polymerase and to initiate the DNA synthesis. In FIG. 2, DNA molecule 202 is ligated with hairpin adapters to form ligated molecule 204. Ligated molecule 204 then forms circularized molecule 206. The molecules without CpG sites can also be sequenced. Circularized molecule 206 includes an unmethylated CpG site 208, which may still be sequenced.

[0202] Once the DNA synthesis was initialized, fluorescently dye-labeled nucleotides would be incorporated into the newly synthesized strand by the immobilized polymerase on the basis of a circular DNA template, leading to the emission of optical signals. Because the DNA templates were circularized, the entire circular DNA template would go through the polymerase multiple times (i.e. one nucleotide in a DNA template would be sequenced multiple times). A sequence generated from the process, in which all bases in the circularized DNA template entirely passed through the DNA polymerase, is called a subread. One molecule in a ZMW would generate multiple subreads because the polymerase can continue around the entire circular DNA template multiple times. In one embodiment, a subread may only contain a subset of the sequence, base modifications or other molecular information of the circular DNA template because, in one embodiment, of the existence of sequencing errors.

[0203] As illustrated in **FIG. 3**, the arrival times and durations of the resultant fluorescence pulses would allow one to measure the polymerase kinetics. The interpulse duration (IPD) is a metric for the length of a time period between two emission pulses each of which would be suggestive of an incorporated fluorescently labeled nucleotide in a nascent strand (FIG. 3). As shown in FIG. 3, the pulse width (PW) is another metric, reflecting polymerase kinetics, in association with the duration of the pulses related to a base call. PW could be the duration of the pulse at 0% of the height of signal peak (i.e., fluorescent intensity of dye-labeled nucleotide as incorporated). In one embodiment, PW could be defined by, for example, but not limited to, the

duration of the pulse at 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the height of signal peak. In some embodiments, the PW may be the area under the peak divided by the height of the signal peak.

[0204] Such polymerase kinetics such as IPDs have been shown to be affected by base modifications such as N6-methyladenine (6mA), 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) in synthetic and microbial sequences (e.g. *E. coli*) (Flusberg *et al.*, 2010). Flusberg *et al.* 2010 did not use sequence context and IPD as independent inputs to detect a modification, which resulted in a model that lacked practically meaningful accuracy for detection. Flusberg *et al.* only used sequence context to confirm 6mA occurred in GATC. Flusberg *et al.* is silent as to using sequence context in combination with IPD as inputs to detect methylation status.

[0205] The weak interruptions conferred on the new base incorporation to the 5-methylcytosine in complementary strands make methylation calling extremely challenging even for relatively simple microbial genomes when using IPD signals only, as it was reported that the detection of methylation motif $C^{m}CWGG$ was only ranged from 1.9% to 4.3% (Clark *et al.*, 2013). For example, the analytic software package (SMRT Link v6.0.0) provided by Pacific Biosciences is not able to perform 5mC analysis. Furthermore, a previous version of the SMRT Link v5.1.0 required one to use the Tet1 enzyme to convert 5mC to 5-carboxylcytosine (5caC) prior to methylation analysis, since the IPD signals associated with 5caC would be enhanced (Clark *et al.*, 2013). Thus, it is not surprising that there are no studies showing the feasibility of using single molecule, real-time sequencing to analyze native DNA in a genomewide manner for the human genome.

II. MEASUREMENT WINDOW PATTERNS AND MACHINE LEARNING MODELS

[0206] Techniques to detect modifications in bases without enzymatic or chemically converting the modification and/or the base are desired. As described herein, modifications in a target base may be detected using kinetic feature data obtained from single molecule, real-time sequencing for bases surrounding the target base. Kinetic features may include interpulse

duration, pulse width, and sequence context. These kinetic features may be obtained for a measurement window of a certain number of nucleotides upstream and downstream of the target base. These features (e.g., at particular locations in the measurement window) can be used to train a machine learning model. As an example of the sample preparation, the two strands of a DNA molecule may be connected by hairpin adapters, thereby forming a circular DNA molecule. The circular DNA molecule allows for kinetic features to be obtained for either or both of the Watson and Crick strands. A data analysis framework can be developed based on the kinetic features in the measurement windows. This data analysis framework may then be used to detect modifications, including methylation. The section describes various techniques for detecting modifications.

A. Using single strand

[0207] As shown in **FIG. 4**, as an example, we obtained the subreads of the Watson strand from Pacific Biosciences SMRT sequencing to analyze one particular base regarding the states of base modifications. In FIG. 4, the 3 bases from each side of a base that was subjected to base modification analysis would be defined as a measurement window 400. In one embodiment, sequence context, IPDs, and PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and downstream sequence and one nucleotide for base modification analysis) were compiled into a 2-dimensional (i.e. 2-D) matrix as a measurement window. In the example shown, the measurement window 400 is for one subread of the Watson strand. Other variations are described herein.

[0208] The first row 402 of the matrix indicated the sequence that was studied. In the second row 404 of the matrix, the position of 0 represented the base for base modification analysis. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, upstream of the base that was subjected to base modification analysis. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt and 3-nt, respectively, downstream of the base that was subjected to base modification analysis. Each position includes 2 columns, which contain the corresponding IPD and PW values. The following 4 rows (rows 408, 412, 416, and 420) corresponded to 4 types of nucleotides (A, C, G, and T) in the strand (e.g. Watson strand),

respectively. The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position. As shown in FIG. 4, at the relative position of 0, the IPD and PW values were shown in the row indicating 'G' in the Watson strand, suggesting that a guanine was called in the sequence result at that position. The other grids in a column that did not correspond to a sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix (FIG. 4) would be 5'-GATGACT-3' for the Watson strand.

[0209] As shown in one embodiment depicted in **FIG. 5**, the measurement window could be applied to data from the Crick strand. We obtained the subreads of the Crick strand from single molecule, real-time sequencing to analyze one particular base regarding the states of base modifications. In FIG. 5, the 3 bases from each side of a base that was subjected to base modification analysis and the base subjected to base modification analysis would be defined as a measurement window. In one embodiment, sequence context, IPDs, PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and downstream sequence and one nucleotide for base modification analysis) were compiled into a 2-dimensional (i.e. 2-D) matrix as a measurement window. The first row of the matrix indicated the sequence that was studied. In the second row of the matrix, the position of 0 represented the base for base modification analysis. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt and 3-nt, respectively, upstream of the base that was subjected to base modification analysis. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt and 3-nt, respectively, downstream of the base that was subjected to base modification analysis. Each position includes 2 columns, which contained the corresponding IPD and PW values. The following 4 rows corresponded to 4 types of nucleotides (A, C, G, and T) in this strand (e.g. the Crick strand). The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position. As shown in FIG. 5, at the relative position of 0, the IPD and PW values were shown in the row indicating 'T' in the Crick strand, suggesting that a thymine was called in the sequence result at that position. The other grids in a column that did not correspond to a sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix (FIG. 5) would be 5'-ACTTAGC-3' for the Crick strand.

B. Using both Watson and Crick strands

[0210] **FIG. 6** shows an embodiment where the measurement window could be implemented in a way that data from the Watson strand and its complementary Crick strand could be combined. As shown in FIG. 6, we obtained the subreads of the Watson and Crick strands from single molecule, real-time sequencing to analyze one particular base for modifications. In one embodiment, the measurement window from the Crick strand of the circular DNA template was complementary to the measurement window from the Watson strand, which was subjected to base modification analysis. In FIG. 6, the 3 bases from each side of the first base in the Watson strand that was subject to base modification analysis and the first base would be defined as the first measurement window. The 3 bases from each side of the second base in the Crick strand and the second base would be defined as the second measurement window. The second base was complementary to the first base. In one embodiment, the sequence context, IPDs, PWs for these 7 bases (i.e. 3-nucleotide (nt) upstream and downstream sequence and one nucleotide for base modification analysis) from the Watson and Crick strands were compiled into a 2-dimensional (i.e. 2-D) matrices. These measurement windows from the Watson and Crick strands were considered as the first and second measurement windows, respectively.

[0211] The first row of the matrix of the Watson and Crick strands indicated the sequence that was studied. In the second row of the matrix of the Watson strand, the position of 0 represented the first base for base modification analysis. The position of 0 shown in the second row of the matrix of the Crick strand represented the second base complementary to the first base. The relative positions of -1, -2, and -3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, upstream of the first and second bases. The relative positions of +1, +2, and +3 indicated the position 1-nt, 2-nt, and 3-nt, respectively, downstream of the first and second bases. Each position derived from the Watson and Crick strands would correspond to 2 columns which contained the corresponding IPD and PW values. The following 4 rows in the matrices of the Watson and Crick strands corresponded to 4 types of nucleotides (A, C, G, and T) in the specific strand (e.g., the Crick strand), respectively. The presence of IPD and PW values in the matrix depended on which corresponding nucleotide type was sequenced at a particular position.

[0212] As shown in FIG. 6, at the relative position of 0, the IPD and PW values were shown in the row indicating 'A' in the Watson strand and 'T' in the Crick strand, suggesting that an adenine and thymine were called in the sequence result at that position of the Watson and Crick strands, respectively. The other grids in a column that did not correspond to sequenced base would be coded as '0'. As an example, the sequence information corresponding to the 2-D digital matrix of the Watson strand (FIG. 6) would be 5'-ATAAGTT-3'. The sequence information corresponding to the 2-D digital matrix of the Crick strand (FIG. 6) would be 5'-AACTTAT-3'.

[0213] As shown in this example, data from the Watson and Crick strands can be combined to form a new matrix, which may also be considered as a measurement window. This new matrix can be used used as a single sample that is used to train a machine learning model. Thus, all of the values in the new matrix can be treated as separate features, although the particular placement in the 2D matrix can have an impact, e.g., when a convolutional neural network (CNN) is used. The sequence context at the various positions for the different strands can be conveyed via the non-zero entries in the matrix.

[0214] **FIG. 7** shows that the measurement window could be implemented in a way that data from the Watson and Crick strands are not the positions exactly complementary to each other. As shown in FIG. 7, the first measurement window was 5'-ATAAGTT-3'; and the second measurement window was 5'-GTAACGC-3'. In some embodiments, the Watson and Crick strands may be shifted from each other so that the positions are not complementary.

[0215] **FIG. 8** shows that a measurement window could be used to analyze the methylation states at CpG sites. The position of 0 corresponds to the cytosine of the CpG site, and thus there is a shift by one position between the two strands, so that the C is at the 0 position for both strands. Accordingly, only a portion of the sequences included in the measurement window from the Watson and Crick strands are complementary to each other. In other embodiments, all of the sequences in the measurement window from the Watson and Crick strands can be complementary to each other. In yet other embodiments, none of the sequences in the measurement window from the Watson and Crick strands are complementary to each other.

[0216] In one embodiment, for a measurement window, the length of DNA stretch surrounding a base that was subjected to base modification analysis could be asymmetrical. For example, X-nt upstream and Y-nt downstream of that base could be used for base modification analysis. X could include, but is not limited to, 0, 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000; Y could include, but was not limited to, 0, 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000.

C. Training models and detecting modifications

[0217] **FIG. 9** shows a general procedure as to how to use the measurement window to determine any base modifications. DNA samples known to be unmodified and modified were subjected to single molecule, real-time sequencing. The modified DNA (e.g., modified molecule 902) means that the base (e.g., base 904) has the modification (e.g., methylation) at the site. The unmodified DNA (e.g., unmodified molecule 906) means that the base (e.g., base 908) does not have the modification at the site. Both sets of DNA can be artificially created or processed to form the modified/unmodified DNA.

[0218] At stage 910, the samples can then undergo single molecule, real-time sequencing. As part of SMRT sequencing, circular molecules could be sequenced multiple times by passing through the immobilized DNA polymerase repeatedly. The sequence information obtained from each time would be deemed as a subread. Thereby, one circular DNA template would generate multiple subreads. The sequencing subreads can be aligned to a reference genome using, for example but not limited to, BLASR (Mark J Chaisson et al, BMC Bioinformatics. 2012; 13: 238). In various other embodiments, BLAST (Altschul SF et al, J Mol Biol. 1990;215(3):403-410), BLAT (Kent WJ, Genome Res. 2002;12(4):656-664), BWA (Li H et al, Bioinformatics. 2010;26(5):589-595), NGMLR (Sedlazeck FJ et al, Nat Methods. 2018;15(6):461-468), LAST (Kielbasa SM et al, Genome Res. 2011;21(3):487-493) and Minimap2 (Li H, Bioinformatics. 2018;34(18):3094-3100) could be used for aligning

subreads to a reference genome. The alignment can allow the data from multiple subreads to be combined (e.g., averaged) as the data in each subread for the same postion can be identified.

[0219] At stage 912, from the alignment result, IPDs, PWs, and sequence context surrounding a base that was subjected to base modification analysis were obtained. At stage 914, the IPDs, PWs, and sequence context were recorded in a certain structure, for example but not limited to, 2-D matrix as shown in FIG. 9.

[0220] At stage 916, a number of 2-D matrices containing the reference kinetic patterns derived molecules with known base modifications were used to train the analytical, computational, mathematical, or statistical model(s). At stage 918, a statistical model is developed resulting from the training. For simplicity, FIG. 9 shows only a statistical model developed by training, but any model or data analysis framework can be developed. Example data analysis frameworks include machine learning models, statistical models, and mathematical models. The statistical models could include, but was not limited to, linear regression, logistic regression, deep recurrent neural network (e.g., long short-term memory, LSTM), Bayes classifier, hidden Markov model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, and support vector machine (SVM). A DNA stretch surrounding a base that was subjected to base modification analysis could be X-nt upstream and Y-nt downstream of that base, namely "measurement window."

[0221] The data structures can be used in a training process, as the correct outputs (i.e., modification state) are known. For example, the IPDs, PWs, and sequence context corresponding to 3-nt upstream and downstream of a base from the Watson and/or Crick strand(s) can be used for constructing the 2-D matrix to be used to train the statistical model(s) for classifying base modifications. In this manner, the training can provide a model that can classify a base modification at a position of a nucleic acid with a previously known status.

[0222] **FIG. 10** shows a general procedure as to how the statistical model(s) learned from DNA samples that carried known states of base modifications can detect base modifications. A sample with unknown states of base modifications was subjected to SMRT sequencing. The

sequencing subreads were aligned to a reference genome using, for example, the techniques mentioned above. In addition or instead, the subreads can be aligned to each other. Yet other embodiments can use just one subread or analyze them independently such that alignment is not performed.

[0223] For a base that was subjected to base modification analysis, one would obtain IPDs, PWs, and sequence context from the Watson and/or Crick strand(s) in the alignment results using a comparable measurement window as used in the training step (FIG. 9), and was associated with that base. In another embodiment, the measurement windows between training and testing procedures would be different. For example, the size of the measurement windows between training and testing procedures might be different. Those IPDs, PWs, and sequence context would be transformed into a 2-D matrix. Such 2-D matrix of a testing sample would be compared with the reference kinetic features to determine base modifications. For example, the 2-D matrix of a testing sample can be compared with reference kinetic features through the statistical model(s) that were learned from the training samples, so that the base modifications at sites in nucleic acid molecules in a testing sample could be determined. The statistical models could include, but was not limited to, linear regression, logistic regression, deep recurrent neural network (e.g., long short-term memory, LSTM), Bayes classifier, hidden Markov model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, and support vector machine (SVM).

[0224] **FIG. 11** shows a general procedure as to how the method could be made for classifying methylation states at CpG sites. DNA samples known to be unmethylated and methylated at CpG sites were subjected to single molecule, real-time sequencing. The sequencing subreads were aligned to a reference genome. Watson strand data were used.

[0225] From the alignment result, IPDs, PWs, and sequence context surrounding a cytosine at a CpG site that was subjected to methylation analysis were obtained and recorded in a certain structure, for example but not limited to, 2-D matrix as shown in FIG. 11. A number of 2-D matrices containing the reference kinetic patterns derived molecules with known methylation

states were used to train the statistical model(s). A stretch of DNA surrounding a base under interrogation could be X-nt upstream and Y-nt downstream of that base, namely the "measurement window." X could include, but is not limited to, 0, 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000; Y could include, but is not limited to, 0, 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, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 100, 150, 200, 300, 400, 500, 1000, 2000, 4000, 5000, and 10000. In one embodiment, IPDs, PWs, and sequence context corresponding to 3-nt upstream and downstream of a base from the Watson strand could be used for constructing the 2-D matrix that was used to train the statistical model(s) to classify base modifications.

[0226] **FIG. 12** shows a general procedure of classifying methylation states of an unknown sample. A sample with unknown methylation states was subject to single molecule, real-time sequencing. The sequencing subreads were aligned to a reference genome.

[0227] For a cytosine of a CG site in the alignment result, one would obtain IPDs, PWs, and sequence context from the Watson strand using a comparable measurement window which was applied in the training step (FIG. 11), associated with that base whose modification was under interrogation. Those IPDs, PWs, and sequence context can be transformed into a 2-D matrix. Such 2-D matrix of a testing sample would be compared with the reference kinetic patterns illustrated in FIG. 11 to determine methylation states. X11

[0228] **FIG. 13** and **FIG. 14** show that kinetic features from the Crick strand could be used for training and testing procedures as elaborated above, similar to procedures with the Watson strand. The statistical model(s) could be the same or different models. When different models, they could be used to obtain independent classifications, which can be compared, e.g., if they are in agreement then a modification status is identified. If they are not in agreement, then an unclaossified status could be identified. When they are the same model, the data can be combined into a single data structure, e.g., the matrix in FIG. 6.

[0229] **FIG. 15** and **FIG. 16** show that kinetic features from both the Watson and Crick strands could be used for training and testing procedures as elaborated above. DNA samples known to be unmethylated and methylated at CpG sites were subjected to single molecule, real-time sequencing. The sequencing subreads were aligned to a reference genome, although alignment of the subreads to each other is possible, as can be done for other methods described herein.

[0230] For a subread in the alignment result, IPDs, PWs, and sequence context surrounding a cytosine of a CpG site which was subject to methylation analysis were obtained. Because DNA molecules were circularized through the use of two hairpin adaptors (e.g., following a SMRTBell template preparation protocol), the circular molecules could be sequenced more than once, thereby generating multiple subreads of a molecule. The subreads can be used for generating circular consensus sequencing (CCS) reads. In general for all methods described herein, one ZMW could generate multiple subreads but only correspond to one CCS read.

[0231] In some embodiments, the fully unmethylated dataset could be created by PCR on human DNA fragments. For example, the fully methylated dataset could be produced through human DNA fragments treated by CpG methyltransferase M.SssI, in which all CpG sites were assumed to be methylated. In other examples, another CpG methyltransferase could be used, such as M.MpeI. In other embodiments, synthetic sequences with known methylation states or pre-existing DNA samples with different methylation levels, or hybrid methylated states creating by restriction enzyme cutting of methylated and unmethylated DNA molecules followed by ligation (which would create a proportion of chimeric methylated/unmethylated DNA molecules) could be used for training the methylation prediction models or classifiers.

[0232] The transformation of kinetic patterns, including sequence context, IPD and pulse width (PW), can be a 2-D matrix comprising features from Watson and Crick strands for analyzing methylation sates at CG sites, as illustrated in FIG. 15. This approach allowed us to accurately capture the subtle kinetic changes caused by methylated cytosines as well as its nearby sequence context. As with any of the various methods described herein, for each CpG present in a subread, the measurement window of (e.g., 3-bases upstream and downstream of a

cytosine of a CpG site) can be used for subsequent analyses, thus leading to a total of 7 nucleotides (including cytosine of a CpG site) being analyzed together. The IPD and PW for each base among those 7 nucleotides can be calculated. To capture the sequence context attributing to kinetic changes, the IPD and PW signals can be compiled into a particular base call, relative sequencing positions, and the strand information as shown in FIG. 15. Such a data structure is named 2-D digital matrix of kinetics for simplicity.

[0233] Such a 2-D digital matrix is analogous to a "2-D digital image". For instance, the first row of the 2-D digital matrix contained the relative positions surrounding a cytosine of a CpG locus that was subjected to methylation analysis, with 3-nt upstream and downstream of that cytosine site. The position of 0 represented the cytosine site whose methylation was to be determined. The relative positions of -1 and -2 indicated the 1-nt and 2-nt upstream of the cytosine that was in question. The relative positions of +1 and +2 indicated the 1-nt and 2-nt downstream of the cytosine that would be used. Each position would correspond to 2 columns which contained the corresponding IPD and PW values. Each row corresponded to the 4 types of nucleotides (A, C, G, and T) in the Watson and Crick strands. The filling of IPD and PW values in the matrix depended on which corresponding nucleotide type was preset in the sequenced result (i.e. subread) at a particular position.

[0234] As shown in FIG. 15, at the relative position of 0, the IPD and PW values were shown in the row of 'C' in the Watson strand, suggesting that cytosine was called at that position. The other grids in a column that did not correspond to a sequenced base would be coded as '0'. As an example, the sequence information corresponded the 2-D digital matrix (FIG. 15) would be 5'-ATACGTT-3' and 5'-TAACGTA-3' for the Watson and Crick strands, respectively. In this context, upstream and downstream sequences flanking a cytosine of a CpG site in the Watson and Crick strands would be different. Since the methylation at CpG sites would be symmetrical between the Watson and Crick strands (Lister *et al.*, 2009), the kinetics in both strands were used for training the methylation prediction model in one preferred embodiment. In another embodiment, the Watson and Crick strands might be used for training the methylation prediction model separately.

[0235] Considering the high data throughput of single molecule, real-time sequencing, in one embodiment, a deep learning algorithm (e.g. convolutional neural networks (CNN)) (LeCun *et al.*, 1989) may be suited for distinguishing the methylated CpGs from unmethylated CpGs. Other algorithms could also be used in addition or instead, for example, but not limited to, linear regression, logistic regression, deep recurrent neural network (e.g., long short-term memory, LSTM), Bayes classifier, hidden Markov model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, and support vector machine (SVM), etc. The training can used the Watson and Crick strands separately or in a combined new matrix, as described in FIGS. 6-8.

[0236] Another transformation of kinetic patterns could be a N-dimensional matrix. N could, for example, be 1, 3, 4, 5, 6, and 7. For example, the 3-D matrix would be a stack of 2-D matrices stratified according to the number of tandem CG sites for a DNA stretch being analyzed, in which the 3rd dimension would be the number of tandem CG sites in that DNA stretch. The pulse strength or pulse magnitude (e.g. measured by the peak height of a pulse, or by the area under the pulse signal) might be also incorporated into a matrix in some embodiments. The pulse strength (a metric for the amplitude of the pulse peak, FIG. 3) could be either added to an extra column adjacent to columns in association with PW and IPD values on top of the original 2-D matrix, or added to a 3rd dimension to form a 3-D matrix.

[0237] As further exampples, a 2D matrix of 8(row)x21(column) can be transformed into a 1-D matrix (i.e. vector) comprising 168 elements. And we can scan this 1-D matrix, e.g., to perform CNN or other modeling. As another example, methods can split an 8x21 2-D matrix to multiple smaller matrices, e.g., two 4x21 2-D matrices. Putting these two smaller matrices together in a vertical direction provides a 3-D matrix (i.e. x=21, y=4, z=2). Methods can scan the 1st 2-D matrix and then the 2nd 2-D matrix, to form the data presentation for machine learning. The data can be split further to form a higher dimensional matrix. Additinally, secondary structure information can be added to the data structure, e.g., an extra matrix (1-D matrix) on top of 2-D matrix. Such an extra matrix can code whether each base within the measurement window is involved in a secondary structure (e.g. stem-loop structure), for example, the base involving the “stem” is coded as 0 and base involving the “loop” is coded as 1.

[0238] In one embodiment, the methylation status of a CpG site within a single DNA molecule can be expressed as a probability of being methylated based on a statistical model, rather than giving a qualitative result of "methylated" or "unmethylated." A probability of 1 indicates that, based on the statistical model, a CpG site may be deemed as methylated. A probability of 0 indicates that, based on the statistical model, a CpG site may be deemed as unmethylated. In subsequent downstream analysis, a cutoff value can be used to classify if a particular CpG site is classified as "methylated" or "unmethylated" based on the probability. The possible values of the cutoff include 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, or 95%. The predicted probability of being methylated for a CpG site greater than a predefined cutoff may be classified as "methylated," while the probability of being methylated for a CpG site not greater than a predefined cutoff may be classified as "unmethylated." A desired cutoff would be obtained from training dataset using, for example, receiver operating characteristics (ROC) curve analysis.

[0239] FIG. 16 shows a general procedure of classifying methylation states of an unknown sample from the Watson and Crick strands. sample with unknown methylation states was subjected to single molecule, real-time sequencing. The sequencing subreads can be aligned to a reference genome or to each other, as with other methods, to determine consensus values (e.g., average, median, mode, or other statistical value) for a given position. As shown, the measured values for the two strands can be combined into a single 2D matrix.

[0240] For a cytosine of a CG site in the alignment result, one would obtain IPDs, PWs, and sequence context from the Watson strand using a comparable measurement window (3-nt upstream and downstream of a cytosine of a CpG site) as applied in the training step (FIG. 16), associated with that base whose modification was under interrogation, although different sized windows can be used. Such a 2-D matrix of a testing sample can be compared with the reference kinetic patterns illustrated in FIG. 16 to determine the methylation states.

III. EXAMPLE MODEL TRAINING FOR DETECTION OF METHYLATION

[0241] To test the feasibility and validity of proposed approaches, we prepared a placenta DNA library with M.SssI treatment (methylated library) and PCR amplification (unmethylated

library) prior to single molecule, real-time sequencing. We obtained 44,799,736 and 43,580,452 subreads for methylated and unmethylated libraries, respectively, corresponding to 421,614 and 446,285 circular consensus sequences (CCSs). As a result, each molecule was sequenced with a median of 34 and 32 times in methylated and unmethylated libraries. The data set was generated from DNA prepared by the Pacific Biosciences Sequel Sequencing Kit 3.0. This kit was developed to be used for use of the original Pacific Biosciences Sequel sequencer. To differentiate the Sequel from its successor, the Sequel II, we herein refer to the original Sequel as Sequel I. Hence, the Sequel Sequencing Kit 3.0 would be referred herein as the Sequel I Sequencing Kit 3.0. Sequencing kits designed for the Sequel II sequencer include the Sequel II Sequencing Kit 1.0 and Sequel II Sequencing Kit 2.0 that are also described in this disclosure.

[0242] We used 50% of the sequenced molecules generated from methylated and unmethylated libraries to train a statistical model (and used the remaining 50% for validation), which in this case is a convolutional neural network (CNN) model. As an example, the CNN model can have one or more convolutional layers (e.g., 1D or 2D layers). A convolutional layer can use one or more different filters, with each filter using a kernel that operates on matrix values local (e.g., in neighboring or surrounding) to a particular matrix element, thereby providing a new value to the particular matrix element. One implementation used two 1D-convolution layers (each with 100 filters with a kernel size of 4). The filters can be applied separately and then combined (e.g., in a weighted average). A resulting matrix can be smaller than the input matrix.

[0243] The convolutional layers can be followed by a ReLU (rectified linear unit) layer, which can be followed by a dropout layer with a dropout rate of 0.5. The ReLU is an example of an activiation function that can operate on the individual values resulting in the new matrix (image) from the convolutional layer(s). Other activation functions (e.g., sigmoid, softmax, etc.) can also be used. One or more of such layers can be used. The dopout layer can be used on the ReLU layer or on a maximum pooling layer and act as a regularization to prevent overfitting. The dropout layer can be used during the training process to ignore different (e.g., random) values during different iterations of an optimization process (e.g., to reduce a cost/loss function) that is performed as part of training.

[0244] A maximum pooling layer (e.g., a pool size of 2) may be used after the ReLU layer. The the maximum pooling layer can act similar to the convolution layer but instead of taking a dot product between the input and the kernel, the maximum of the region from the input overlapped by the kernel can be taken. Further convolutional layer(s) can be used. For example, the data from a pooling layer can be input to another two 1D-convolution layers (e.g., each with 128 filters with a kernel size of 2 followed by a ReLU layer), further using a dropout layer with a dropout rate of 0.5. A maximum pooling layer with a pool size of 2 was used. Finally, a fully connected layer (e.g., with 10 neurons followed by a ReLU layer) can be used. An output layer with one neuron can be followed by a sigmoid layer, thereby yielding the probability of methylation. Various settings of layers, filters, and kernel sizes can be adapted. In this training dataset, we used 468,596 and 432,761 CpG sites from methylated and unmethylated libraries.

A. Results of training and testing datasets

[0245] **FIG. 17A** shows the probability of being methylated for each CpG site in each single DNA molecule in the training dataset. The probability of methylation was much higher in the methylated library than the unmethylated library. For a cutoff of 0.5 for the probability of being methylated, 94.7% of unmethylated CpG sites were correctly predicted to be unmethylated, and 84.7% of methylated CpG were correctly predicted to be methylated.

[0246] **FIG. 17B** shows the performance of the testing dataset. We used a model trained by the training dataset to predict the methylation states of 469,729 and 432,024 CpG sites in an independent testing dataset from methylated and unmethylated libraries. For a cutoff of 0.5 for probability of being methylated, 94.0% of unmethylated CpG sites were correctly predicted to be unmethylated, and 84.1% of methylated CpG were correctly predicted to be methylated. These results suggested that the use of a novel transformation of kinetics coupled with sequence context could enable the determination of methylation states in DNA (e.g. from human subjects).

[0247] We evaluated the power of each feature (sequence context, IPD, and PW) in predicting the methylation state of CpG by including a subset of the features in the model. In the training dataset, models with (i) sequence context only, (ii) IPD only, and (iii) PW only gave area-under-the-curve (AUC) values of 0.5, 0.74 and 0.86, respectively. While combining IPD and sequence

context improved the performance with an AUC of 0.86. The combined analysis of sequence context ("Seq"), IPD, and PW substantially improved the performance with an AUC of 0.94 (**FIG. 18A**). The performance of an independent testing dataset was comparable to the training dataset (**FIG. 18B**).

[0248] We defined the subread depth of a CpG site as the average number of subreads covering it and its surrounding 10 bp. As shown in **FIG. 19A** and **FIG. 19B**, the higher the subread depth of a CpG site, the higher the accuracy of detection of methylation we would achieve. For example, as shown in the testing dataset (**FIG. 19B**), if the depth of each CpG site was at least 10, the AUC of predicting methylation states would be 0.93. However, if the subread depth of each CpG site is at least 300, the AUC of predicting methylation states would be 0.98. On the other hand, even for the depth of 1, we could achieve an AUC of 0.9, suggesting that our approach could achieve methylation prediction with the use of low sequencing depth.

[0249] To test the effect of strand information on the performance of methylation analysis, the sequence context, IPD and PW originating from the Watson and Crick strands were used to train according to the embodiments present in this disclosure, respectively. **FIG. 20A** and **FIG. 20B** showed that it is feasible to use a single strand, namely either the Watson or Crick strand, for training and testing as the AUC could be achieved up to 0.91 and 0.87 in training and testing datasets. The use of both strands (e.g., as described in FIGS. 6-8) including the Watson and Crick strands would give rise to the best performance (AUC: 0.94 and 0.90 in training and testing datasets, respectively), suggesting the strand information would be important to achieve the optimal performance.

[0250] We further tested the different number of nucleotides upstream and downstream of a CpG site, to study how this parameter affected the performance of according to the embodiments present in this disclosure developed in this disclosure. **FIG. 21A** and **FIG. 21B** show that the number of nucleotides upstream and downstream of a cytosine in the context of CpG would affect the accuracy of methylation prediction. For example, as an illustration purpose, considering but not limited to 2 nucleotides (nt), 3 nt, 4 nt, 6 nt, 8 nt, 10 nt, 15 nt, and 20 nt upstream and downstream of a cytosine being analyzed, the AUC of a method using 2 nt

upstream and downstream of cytosine being interrogated would be only 0.50 in both training and testing datasets, whereas the AUC of a method using 15 nt upstream and downstream of a cytosine being interrogated would increase to 0.95 and 0.92 in the training and testing datasets. These results suggested that varying the length of upstream and downstream regions flanking cytosines being analyzed would allow to figure out the optimal performance. In one embodiment, as shown in FIG. 21B, one would use 3 nt upstream and downstream of a cytosine to determine the methylation states, which could achieve an AUC of 0.89.

[0251] In one embodiment, one could use asymmetrical sequences flanking a cytosine being interrogated to perform the analysis according to the embodiments present in this disclosure. For example, 2 nt upstream combined with 1 nt, 3 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt downstream of a cytosine could be used; 3 nt upstream combined with 1 nt, 2 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt downstream of a cytosine could be used; 4 nt upstream combined with 1 nt, 2 nt, 3 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt downstream of a cytosine could be used. As another example, 2 nt downstream combined with 1 nt, 3 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt upstream of a cytosine could be used; 3 nt downstream combined with 1 nt, 2 nt, 4 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt upstream of a cytosine could be used; 4 nt downstream combined with 1 nt, 2 nt, 3 nt, 5 nt, 6 nt, 7 nt, 8 nt, 9 nt, 10 nt, 11 nt, 12 nt, 13 nt, 14 nt, 15 nt, 16 nt, 17 nt, 18 nt, 19 nt, 20 nt, 25 nt, 30 nt, 35 nt, and 40 nt upstream of a cytosine could be used. By taking advantage of IPDs, PWs, strand information, and sequence context in association with n-nt upstream and m-nt downstream of a cytosine could provide an improved accuracy in determining the methylation states in certain embodiments. Such varying measurement windows could be applied to other types of base modification analysis, such 5hmC, 6mA, 4mC, and oxoG, or any modification disclosed herein. Such varying measurement windows could include DNA secondary structure analysis, such as

G-quadruplex and stem-loop structure. Such an example is explained above. Such secondary structure information could also be added as another column in a matrix.

[0252] **FIG. 22A** and **FIG. 22B** show that it is feasible to determine the methylation states using kinetic patterns associated with only downstream bases of at least 3 bases. According to the embodiments present in this disclosure, with the use of features associated with cytosine and its downstream 3, 4, 6, 8, and 10 bases, the AUCs of the determination of methylation states in training dataset were 0.91, 0.92, 0.94, 0.94, and 0.94, respectively, in the training dataset; the AUCs were 0.87, 0.88, 0.90, 0.90, and 0.90, respectively, in the testing dataset.

[0253] **FIG. 23A** and **FIG. 23B** show, however, if one only uses the features associated with the upstream bases, the classification power appears to diminish in one's ability in distinguishing the methylation states. The AUCs in the training dataset and the testing dataset were all 0.50 for 2 to 10 upstream bases.

[0254] **FIG. 24** and **FIG. 25** show that different combinations of upstream and downstream bases would allow one to achieve an optimal classification power in determining methylation states. For example, the features associated with 8-base upstream and 8-base downstream of a cytosine would achieve a best performance in this dataset, with an AUC of 0.94 and 0.91 in the training and testing datasets, respectively.

[0255] **FIG. 26** shows the relative importance of features in regard to classification of methylation states at CpG sites. 'W' and 'C' in the brackets indicate the strand information, 'W' for the Watson strand and 'C' for the Crick strand. The importance of each feature, including sequence context, IPD and PW, was determined using random forest. The random forest tree analysis showed that the feature importance of IPDs and PWs peaked in the downstream of a cytosine that was under interrogation, revealing that the major contributions to classification power were the IPDs and PWs downstream to a cytosine that was under interrogation.

[0256] The random forest was composed of multiple decision trees. During the construction of the decision tree, Gini impurity was used to determine which decision logic for decision nodes should be taken. Important features that have more influence on the final classification outcome

were likely in nodes closer to the root of the decision tree, while unimportant features that have less influence on the final classification outcome are likely in nodes further away from the root. So, the feature importance could be estimated by computing the average distance relative to the roots of all decision trees in the random forest.

[0257] In some embodiments, the consensus of methylation calls at CpG sites between the Watson and Crick strands could be further used for improving the specificity. For example, it could be required that both strands showing methylated would be called as a methylated state, and both strands showing unmethylated would be called as unmethylated state. Since the methylation at CpG sites was known to be typically symmetrical, the confirmation from each strand can improve the specificity.

[0258] In various embodiments, the overall kinetic features from a whole molecule might be used for the determination of methylation states. For example, the methylation in a whole molecule would affect kinetics of the whole molecule during single molecule, real-time sequencing. By modelling sequencing kinetics of the entire template DNA molecule including IPDs, PWs, fragment sizes, strand information, and sequence context, it may improve the accuracy of the classification as to whether a molecule is methylated or not. As an example, the measurement windows may be the entire template molecule. Statistical values (e.g., mean, median, mode, percentile, etc) of IPD, PW, or other kinetic features may be used for determining methylation of a whole molecule.

B. Limitations of other analysis techniques

[0259] It was reported that the detection of methylation based on IPD for a particular C in a particular sequence motif was very low, for example, a sensitivity of only 1.9% (Clark *et al.*, 2013). We also attempted to reproduce such analysis by combining different sequence motifs with IPDs without using the PW metric, and just using a cutoff for the IPD and not the data structures as descrined herein. For example, 3-nt upstream and downstream flanking a CpG being interrogated were extracted. IPDs of that CpG were stratified into different groups (4096 groups for the 6 positions) depending on the context of 6-nt flanking sequences (i.e. upstream and downstream 3 nt, respectively) that was centered on that CpG. The IPDs between methylated

and unmethylated CpGs within the same sequence motif were studied using ROC. For example, IPDs of CpG in the unmethylated "AATCGGAC" motif and methylated "AAT^{m}CGGAC" motif were compared, showing an AUC of 0.48. Thus, the use of cutoffs in a particular sequence group performed poorly relatively to embodiment that use various

[0260] **FIG . 27** shows the performance of the above motif-based IPD analysis (Beckmann et al. BMC Bioinformatics. 2014) for methylation detection without using the pulse width signal. The vertical bar plots represent the averaged AUCs across different k-mer motifs flanking CpG sites being studied (i.e. the number of bases surrounding CpG sites being interrogated). FIG. 27 showed that the averaged AUCs for IPD based discriminative powers between methylated and unmethylated cytosines across different k-mer motifs (e.g. 2-mer, 3-mer, 4-mer, 6-mer, 8-mer, 10-mer, 15-mer, 20-mer surrounding CpG sites in question) was found to be less than 60%. These results suggested that the consideration of the IPD of the candidate nucleotide in a given motif context without taking into account the IPDs of the neighboring nucleotides (Flusberg *et al.*, 2010) would be inferior to the methods disclosed herein for the determination of CpG methylation.

[0261] We also tested the method present in Flusberg *et al.* study (Flusberg *et al.*, 2010). We analyzed a total of 5,948,348 DNA segments which were 2-nt upstream and 6-nt downstream of a cytosine that was subjected to methylation analysis. There were 2,828,848 segments that were methylated and 3,119,500 segments that were unmethylated. As shown in **FIG. 28**, the signals deduced from the principal component analysis with the use of IPDs and PWs were found to be largely overlapping between fragments with methylated cytosines (mC) and unmethylated cytosines (C), suggesting that the method as described by Flusberg *et al* lacks practically meaningful accuracy. These results suggested that the principal component analysis, which linearly combined PW and IPD values at the bases and neighboring bases, as used in Flusberg *et al*'s study (Flusberg *et al.*, 2010) could not reliably or meaningfully differentiate 5-methylcytosine and unmethylated cytosines.

[0262] **FIG. 29** shows that the AUC of the method based on the principal component analysis for which two principal components were used in Flusberg *et al*'s study (Flusberg *et al.*, 2010)

involving IPDs and PWs were much less accurate (AUC: 0.55) than the approach based on a convolutional neural network involving IPD and PW as well as sequence context as shown in our disclosure (AUC: 0.94).

C. Other mathematical/statistical models

[0263] In another embodiment, other mathematical/statistical models, for example including but not limited to a random forest and logistic regression, could be trained by adapting features developed above. As for the CNN model, the training and testing datasets were constructed from the DNA with M.SssI treatment (methylated) and PCR amplification (unmethylated), which were used to train a random forest (Breiman, 2001). In this random forest analysis, we described each nucleotide with 6 features: IPD, PW, and a 4-component binary vector encoding the base identity. In such a binary vector, A, C, G, and T was coded with [1,0,0,0], [0,1,0,0], [0,0,1,0] and [0,0,0,1], respectively. For each CpG site being analyzed, we incorporated the information of its 10 nt upstream and downstream in both strands, forming a 252-dimension (252-D) vector, with each feature representing one dimension. The training dataset described above with the 252-D vectors was used to train a random forest model, as well as the logistic regression model. The trained model was used to predict the methylation states in an independent testing dataset. The random forest was comprised of 100 decision trees. During tree construction, bootstrap samples were used. While splitting the node of each decision tree, Gini impurity was employed to determine the best split, and a maximum of 15 features would be considered in each split. Also, each leaf of the decision tree was required to contain at least 60 samples.

[0264] **FIG. 30A** and **FIG. 30B** show performance of a method using a random forest and logistic regression for methylation prediction. FIG. 30A shows AUC values in training dataset for CNN, random forest, and logistic regression. FIG. 30B shows AUC values in testing dataset for CNN, random forest and logistic regression. The AUC of a method using random forest achieved 0.93 and 0.86 in the training and testing dataset, respectively.

[0265] The training dataset described with the same 252-D vectors was used to train a logistic regression model. The trained model was used to predict the methylation states in an independent testing dataset. A logistic regression model with L2 regularization (Ng and Y., 2004) was fitted

with the training dataset. As shown in FIG. 30A and FIG. 30B, the AUC of a method using logistic regression would achieve 0.87 and 0.83 in the training and testing dataset, respectively.

[0266] Therefore, these results suggested that certain models (for example, but not limited to the random forest and logistic regression) other than CNN could be used for methylation analysis using the features and analytical protocols we developed in this disclosure. These results also suggested that CNN implemented according to the embodiments in this disclosure with an AUC of 0.90 in the testing dataset (FIG. 30B) was superior to both the random forest (AUC: 0.86) and the logistic regression (AUC: 0.83).

D. Determination of 6mA modifications of nucleic acids

[0267] In addition to methylated CpG, the methods described herein can also detect other DNA base modifications. For example, methylated adenine, including in the form of 6mA, can be detected.

1. 6mA detection using kinetic features and sequencing context

[0268] To evaluate the performance and utility of the embodiments disclosed for the determination of base modifications of nucleic acids, we further analyzed N6-adenine methylation (6mA). In one embodiment, approximately 1 ng of human DNA (e.g. extracted from placental tissues) was amplified to obtain 100 ng DNA product through whole genome amplification with unmethylated adenine (uA), unmethylated cytosine (C), unmethylated guanine (G), and unmethylated thymine (T).

[0269] **FIG. 31A** shows an example of one approach for generating molecules with unmethylated adenines by whole genome amplification. In the figure, "uA" denotes an unmethylated adenine and "mA" denotes a methylated adenine. Whole genome amplification was performed using exonuclease-resistant thiophosphate-modified random hexamers as primers, which bind randomly over a genome, allowing the polymerase (e.g. Phi29 DNA polymerase) to amplify the DNA (e.g., by isothermal linear amplification). At stage 3102, a double-stranded DNA is denatured. At stage 3106, the amplification reaction is initiated when a number of random hexamers (e.g., 3110) were annealed to the denatured template DNA (i.e.

single-stranded DNA). As shown in 3114, when the hexamer-mediated DNA synthesis of strand 3118 proceeded in the 5' to 3' direction and arrived at the next hexamer-mediated DNA synthesis site, the polymerase displaced the newly-synthesized DNA strand (3122) and continued the strand extension. The displaced strands became single-stranded DNA templates for binding of random hexamers again and might initiate new DNA synthesis. Repeated hexamer annealing and strand displacement in an isothermal process would result in a high yield of amplified DNA products. This amplification described here may fall under the technique of multiple displacement amplification (MDA).

[0270] The amplified DNA products were further fragmented into, for example, but not limited to, fragments with sizes of 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, or other desired size ranges. The fragmentation process may include enzymatic digestion, nebulization, hydrodynamic shearing, and sonication, etc. As a result, the original base modifications such as 6mA may be nearly eliminated by whole genome amplification with unmethylated A (uA). FIG. 31A shows possible fragments (3126, 3130, and 3134) of the DNA products, with both strands having unmethylated A. Such whole-genome amplified DNA products without mA were subjected to single-molecule, real-time sequencing to generate uA dataset.

[0271] **FIG. 31B** shows an example of one approach for generating molecules with methylated adenines by whole genome amplification. In the figure, "uA" denotes an unmethylated adenine and "mA" denotes a methylated adenine. Approximately 1 ng human DNA was amplified to obtain 10 ng DNA product through whole genome amplification with 6mA and unmethylated C, G, and T. The methylated adenines can be produced through a series of chemical reactions (J D Engel *et al.* J Biol Chem. 1978;253:927-34). As illustrated in FIG. 31B, the whole genome amplification was performed using exonuclease-resistant thiophosphate-modified random hexamers as primers which bind randomly over a genome, allowing the polymerase (e.g. Phi29 DNA polymerase) to amplify the DNA (e.g. by isothermal linear amplification), similar to FIG. 31A. Exonuclease-resistant thiophosphate-modified random hexamers are resistant to the 3'→5' exonuclease activity of proofreading DNA polymerases. Thus, during amplification, the random hexamers will be protected from degradation.

[0272] The amplification reaction initiated when a number of random hexamers were annealed to the denatured template DNA (i.e. single-stranded DNA). When the hexamer-mediated DNA synthesis proceeded in the 5' to 3' direction and arrived at the next hexamer-mediated DNA synthesis site, the polymerase displaced the newly-synthesized DNA strand and continue the strand extension. The displaced strands became single-stranded DNA templates for binding of random hexamers again and initiating new DNA synthesis. Repeated hexamer annealing and strand displacement in an isothermal process would result in a high yield of amplified DNA products.

[0273] The amplified DNA products were further fragmented into, for example, but not limited to, fragments with sizes of 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 5 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 60 kb, 70 kb, 80 kb, 90 kb, 100 kb, or other combinations in length. As shown in FIG. 31B, the amplified DNA products would include different forms of methylation patterns across adenine sites in each strand. For example, both strands of a double-stranded molecule may be methylated with respect to adenines (Molecule I), which would be generated when two strands are derived from DNA synthesis during whole genome amplification.

[0274] As another example, one strand of a double-stranded molecule may contain interlacing methylation patterns across adenine sites (Molecule II). An interlacing methylation pattern is defined as one that includes a mixture of methylated and unmethylated bases present in a DNA strand. In the following examples, we use an interlacing adenine methylation pattern that includes a mixture of methylated and unmethylated adenines present in a DNA strand. This type of double-stranded molecule (Molecule II) would possibly be generated because an unmethylated hexamer containing unmethylated adenines was bound to a DNA strand and initiated DNA extension. Such an amplified DNA product containing the hexamer with unmethylated adenines would be sequenced. Alternatively, this type of double-stranded molecule (Molecule II) would be initiated by fragmented DNA from original template DNA containing unmethylated adenines, since such fragmented DNA could be bound to a DNA strand as a primer. Such an amplified DNA product containing part of the original DNA with unmethylated

adenines in a strand would be sequenced. As the unmethylated hexamer primers are only a small portion of the resulting DNA strands, the majority of fragments will still contain 6mA.

[0275] As another example, one strand of a double-stranded DNA molecule may be methylated across adenine sites but the other stand may be unmethylated (Molecule III). This type of double-stranded molecule may be generated when an original DNA strand without methylated adenines is provided as a template DNA molecule for producing a new strand with methylated adenines.

[0276] Both strands may be unmethylated (Molecule IV). This type of double-stranded molecule may be due to the reannealing of two original DNA strands without methylated adenines.

[0277] The fragmentation process may include enzymatic digestion, nebulization, hydrodynamic shearing, and sonication, etc. Such whole-genome amplified DNA products may be predominantly methylated in terms of A sites. This DNA with mA was subjected to single-molecule, real-time sequencing to generate mA dataset.

[0278] For the uA dataset, we sequenced 262,608 molecules with a median of 964 bp in length using single-molecule, real-time sequencing. The median subread depth was 103 x. Of the subreads, 48% could be aligned to a human reference genome using the BWA aligner (Li H *et al.* Bioinformatics. 2009;25:1754-60). As an example, one could employ the Sequel II System (Pacific Biosciences) to carry out single-molecule, real-time sequencing. The fragmented DNA molecules were subjected to single-molecule real-time (SMRT) sequencing template construction using a SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences). Sequencing primer annealing and polymerase binding conditions were calculated with the SMRT Link v8.0 software (Pacific Biosciences). Briefly, sequencing primer v2 was annealed to the sequencing template, and then a polymerase was bound to templates using a Sequel II Binding and Internal Control Kit 2.0 (Pacific Biosciences). Sequencing was performed on a Sequel II SMRT Cell 8M. Sequencing movies were collected on the Sequel II system for 30 hours with a Sequel II Sequencing Kit 2.0 (Pacific Biosciences).

[0279] For the mA dataset, we sequenced 804,469 molecules with a median of 826 bp in length using single-molecule, real-time sequencing. The median subread depth was 34 x. Of the subreads, 27% could be aligned to a human reference genome using the BWA aligner (Li H *et al.* Bioinformatics. 2009;25:1754-60).

[0280] In one embodiment, the kinetic characteristics including but not limited to IPD and PW were analyzed in a strand-specific manner. For the sequencing results derived from the Watson strand, 644,318 A sites without methylation randomly selected from the uA dataset and 718,586 A sites with methylation randomly selected from the mA dataset were used to constitute a training dataset. Such a training dataset was used to establish the classification models and/or thresholds for differentiating between methylated and unmethylated adenines. A testing data set was constituted from 639,702 A sites without methylation and 723,320 A sites with methylation. Such a testing dataset was used to validate the performance for a model/threshold deduced from a training dataset.

[0281] We analyzed sequencing results originating from the Watson strands. **FIG. 32A** shows interpulse duration (IPD) values across the training dataset of the uA and mA datasets. For the training dataset, IPD values across sequenced A sites were observed to be higher in the mA dataset (median: 1.09; range: 0 – 9.52) than the uA dataset (median: 0.20; range: 0 – 9.52) (*P* value < 0.0001; Mann Whitney U test).

[0282] **FIG. 32B** shows IPD for the testing dataset of the uA and mA datasets. When we studied the IPD values across sequenced A sites in the testing dataset, we observed that IPD values were higher in the mA dataset than the uA dataset (median 1.10 versus 0.19; *P* value < 0.0001; Mann Whitney U test).

[0283] **FIG. 32C** shows the area under the receiver operating characteristic (ROC) curve using the IPD cutoff. The true positive rate is on the y-axis, and the false positive rate is on the x-axis. The area under the receiver operating characteristic curve (AUC) in differentiating sequenced A bases in template DNA molecules with and without methylation using corresponding IPD values was 0.86 for both training and testing datasets.

[0284] In addition to results from the Watson strand, we analyzed sequencing results originating from the Crick strands. **FIG. 33A** shows IPD values across the training dataset of uA and mA datasets. For the training dataset, IPD values across sequenced A sites were observed to be higher in the mA dataset (median: 1.10 range: 0 - 9.52) than in the uA dataset (median: 0.19; range: 0 – 9.52) (*P* value < 0.0001; Mann Whitney U test).

[0285] **FIG. 34B** shows IPD values for the testing dataset of uA and mA datasets. The higher IPD values across sequenced A sites were also observed in the mA dataset for the testing dataset, compared with the uA dataset (median 1.10 versus 0.19; *P* value < 0.0001; Mann Whitney U test).

[0286] **FIG. 33C** shows the area under ROC curve. The true positive rate is on the y-axis, and the false positive rate is on the x-axis. The area under the ROC curve (AUC) value in differentiating sequenced A bases in template DNA molecules with and without methylation using corresponding IPD values was 0.86 and 0.87 for the training and testing datasets, respectively.

[0287] **FIG. 34** shows an illustration for 6mA determination of the Watson strand using a measurement window according to embodiments of the present invention. Such a measurement window may include kinetic features such as IPD and PW and nearby sequence context. Determination of 6mA may be performed similarly as determination of methylated CpG.

[0288] **FIG. 35** shows an illustration for 6mA determination of the Crick strand using a measurement window according to embodiments of the present invention. Such a measurement window may include kinetic features such as IPD and PW and nearby sequence context.

[0289] As an example, 10 bases from each side of sequenced A base in a template DNA that was being interrogated were used to construct a measurement window. The feature values including IPDs, PWs, and sequence context were used to train a model using a convolutional neural network (CNN) according to the methods disclosed herein. In other embodiments, the statistical models may include, but are not limited to, linear regression, logistic regression, deep recurrent neural network (e.g. long short term memory, LSTM), Bayes classifier, hidden Markov

model (HMM), linear discriminant analysis (LDA), k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), random forest algorithm, and support vector machine (SVM), etc.

[0290] **FIG. 36A** and **FIG. 36B** show the determined probability of being methylated for sequenced A bases of the Watson strand between uA and mA datasets using a measurement window based CNN model. FIG. 36A shows that a CNN model was learned from a training dataset. As an example, the CNN model made use of two 1D-convolution layers (each with 64 filters with a kernel size of 4 followed by a ReLU (rectified linear unit) layer), followed by a dropout layer with a dropout rate of 0.5. A maximum pooling layer with a pool size of 2 was used. Then it flowed into two 1D-convolution layers (each with 128 filters with a kernel size of 2 followed by a ReLU layer), further using a dropout layer with a dropout rate of 0.5. A maximum pooling layer with a pool size of 2 was used. Finally, a fully connected layer with 10 neurons followed by a ReLU layer, with an output layer with one neuron followed by a sigmoid layer, thereby yielded the probability of methylation. The other settings of layers, filters, kernel sizes could be adapted, e.g., as described herein for other methylation (e.g., CpG). In this training dataset regarding the sequencing results of the Watson strand, we used 644,318 and 718,586 A bases from unmethylated and methylated libraries.

[0291] Based on the CNN model, for the Watson strand related data, sequenced A bases in template DNA molecules from mA database gave rise to a much higher probability of methylation in both the training and testing datasets, in comparison with those A bases present in the uA dataset (P value < 0.0001; Mann Whitney U test). For the training dataset, the median probability of methylation on A sites in the uA dataset was 0.13 (interquartile range, IQR: 0.09 – 0.15), whereas that value in mA dataset was 1.000 (IQR: 0.998 – 1.000).

[0292] FIG. 36A shows the probability of methylation determined for the testing dataset. For the testing dataset, the median probability of methylation on A sites in the uA dataset was 0.13 (IQR: 0.10 – 0.15), whereas that value in mA dataset was 1.000 (IQR: 0.997 – 1.000). FIGS. 36A and 36B show that a measurement window based CNN model can be trained to detect methylation in a testing data set.

[0293] **FIG. 37** is an ROC curve for the detection of 6mA using a measurement window based CNN model for sequenced A bases of the Watson strand. True positive rate is on the y-axis and false positive rate is on the x-axis. The figure shows that AUC value in differentiating sequenced A sites with and without methylation using a CNN model was 0.94 and 0.93 for training and testing datasets which consisted of the Watson strand sequencing results, respectively. It suggested that it was feasible to use the disclosure herein to determine methylation states on A sites using data of the Watson strand. If we used the determined probability of methylation of 0.5 as cutoff, 99.3% specificity and 82.6% sensitivity could be achieved for 6mA detection. FIG. 37 shows that a measurement window based CNN model can be used to detect 6mA with high specificity and sensitivity. The accuracy of the model can be compared to a technique using only an IPD metric.

[0294] **FIG. 38** shows a performance comparison between IPD-metric-based 6mA detection and a measurement window based 6mA detection. Sensitivity is plotted on the y-axis and specificity is plotted on the x-axis. FIG. 38 shows that the performance using measurement window based 6mA classification according to the disclosure herein (AUC: 0.94) was superior to that conventional method using IPD metric only (AUC: 0.87) (P value < 0.0001; DeLong's test). The measurement window based CNN model outperformed the IPD-metric-based detection.

[0295] **FIG. 39A** and **39B** show the determined probability of being methylated for those sequenced A bases of the Crick strand between uA and mA datasets using measurement window based CNN model. FIG. 39A shows the training dataset and FIG. 39B shows the testing dataset. Both figures plot the probability of methylation on the y-axis. FIGS. 39A and 39B show that on the basis of CNN model, for the Crick strand related data, sequenced A bases in template DNA molecules from mA database gave rise to a much higher probability of methylation in both training and testing dataset, in comparison with those A bases present in uA database (P value < 0.0001; Mann-Whitney U test).

[0296] **FIG. 40** shows the performance of 6mA detection using measurement window based CNN model on sequenced A bases of the Crick strand. The true positive rate is on the y-axis. The false positive rate is on the x-axis. FIG. 40 shows that AUC value in differentiating

sequenced A sites with and without methylation using a CNN model was 0.95 and 0.94 for training and testing datasets which consisted of the Crick strand sequencing results, respectively. The performance using the CNN approach disclosed herein (AUC: 0.94) was also shown to be superior to that only using IPD metric (0.87) (P value < 0.0001). The results suggested that it was feasible to use the disclosure herein to determine methylation states on A sites using data of the Crick strand. If we used the determined probability of methylation of 0.5 as cutoff, 99.3% specificity and 83.0% sensitivity could be achieved for 6mA detection. FIG. 40 shows that a measurement window based CNN model can be used to detect 6mA with high specificity and sensitivity.

[0297] **FIG. 41** shows examples of methylation states across A bases in a molecule including the Watson and Crick strands. The white dots represent unmethylated adenines. The black dots represent methylated adenines. The horizontal lines with dots represent a strand of a double-stranded DNA molecule. Molecule 1 shows that both the Watson and Crick strands are determined to be unmethylated across A bases. Molecule 2 shows that the Watson strand was nearly all unmethylated while the Crick strand was nearly all methylated. Molecule 3 shows that both the Watson and Crick strands were determined to be nearly all methylated across A bases.

2. Enhanced training using selective dataset

[0298] As shown in FIGS. 36A, 36B, 39A, and 39B, there was a bimodal distribution of probability of methylation across sequenced A bases in template DNA molecules in mA dataset. In other words, there existed some molecules with uA signals in the mA dataset. This was further evidenced by the existence of fully-unmethylated molecules and hemi-methylated molecules in mA dataset (FIG. 41). One possible reason may be that molecules with uA in DNA templates would still account for a considerable portion in the mA dataset after whole genome amplification, as molecules with 6mA would lead to the decreased efficiency of amplifying DNA during the whole-genome amplification step. This explanation was supported by the fact that 1 ng genomic DNA amplified with 6mA would only lead to 10 ng DNA products, whereas 1 ng genomic DNA amplified with unmethylated A would give rise to 100 ng DNA products under the same amplification conditions. Therefore, for the mA dataset, original template DNA

molecules whose adenines are usually unmethylated (e.g. 0.051%) (Xiao CL *et al.* Mol Cell. 2018;71:306-318) would account approximately 10% of the total adenines.

[0299] In one embodiment, when one attempted to train a CNN model for differentiating between mA and uA, one would selectively use those A bases with relatively higher IPD values in the mA dataset so as to reduce the influence of the uA data on training the model for mA detection. Only A bases with IPD values above a certain cutoff value may be used. The cutoff value may correspond to a percentile. In one embodiment, one would use those A bases in mA dataset with IPD values greater than the value at the 10^{th} percentile. In some embodiments, one would use those A with IPD values greater than value at the 1^{st}, 5^{th}, 15^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th}, or 95^{th} percentile. The percentile may be based on data from all nucleic acid molecules in a reference sample or multiple reference samples.

[0300] **FIG. 42** shows the performance with enhanced training by selectively using A bases in mA dataset with IPD values greater than its 10^{th} percentile. FIG. 42 shows the true positive rate on the y-axis and the false positive rate on the x-axis. The figure shows that with the use of A bases in the mA dataset with IPD values greater than the 10^{th} percentile to train a CNN model, the AUC in differentiating between mA and uA bases would increase to 0.98, which was superior to the model (AUC: 0.94) trained by data without the selection according to IPD values prior to training. It suggested that the selection of mA sites using IPD values to create a training dataset would aid in improving the discriminative power.

[0301] To further confirm the existence of molecules with uA bases in mA dataset, we hypothesized that the percentage of uA in mA dataset would enrich in those wells with more subreads, as the 6mA present in a molecule would slow down the polymerase elongation when generating a new strand, in comparison with a molecule without 6mA.

[0302] **FIG. 43** shows a graph of the percentages of unmethylated adenines in the mA dataset versus the number of subreads in each well. The y-axis shows the percentage of uA in the mA dataset. The x-axis shows the number of subreads in each well. The testing dataset was reanalyzed using the enhanced model that was trained by using mA sites after removal of A sites whose IPD values were below the 10^{th} percentile. A gradual increase of uA (i.e. rising from 14.6

to 55.05%) was observed as the number of subreads per well increased, including from 1 to 10 subreads per sequencing well, to 10 to 20 subreads per well, to 40 to 50 subreads per well, 60 to 70 subreads per well and above 70. Thus, wells that have a high number of subreads tend to have low mA. The methylation of A may retard the progression of the sequencing reaction. Hence, it is more likely that sequencing wells with a high subread depth will be unmethylated with regard to A. This behavior can be exploited for detection of the unmethylated molecules using a cutoff value for the number of subreads associated with the molecule, e.g., greater than 70 subreads can be identified as majority unmethylated.

[0303] **FIG. 44** shows methyladenine patterns between the Watson and Crick strands of a double-stranded DNA molecule in a testing dataset. The methylation of A is asymmetric, and thus the behavior is different between the two strands. Most of the molecules were methylated due to the incorporation of mA, with some residual unmethylated A. The y-axis shows the methyladenine level of the Crick strand. The x-axis shows the methyladenine level of the Watson strand. Each dot represents a double-stranded molecule. Using the enhanced model that was trained by selected mA sites, the double-stranded molecules may be categorized into different groups according to the methylation level of each strand as follows:

(a) For a double-stranded DNA molecule, the methyladenine levels of the Watson and Crick strands were both greater than 0.8. Such a double-stranded molecule was defined as a fully-methylated molecule regarding adenine sites (FIG. 44, area A). The methyladenine level of a strand was defined as the percentage of A sites that were determined to be methylated among total A sites in that strand.

(b) For a double-stranded DNA molecule, the methyladenine level of one strand was greater than 0.8 whereas the other strand was less than 0.2. Such a molecule was defined as hemi-methylated molecule regarding adenine sites (FIG. 44, areas B1 and B2).

(c) For a double-stranded DNA molecule, the methyladenine levels of the Watson and Crick strands were both less than 0.2. Such a double-stranded molecule was defined as a fully-unmethylated molecule regarding adenine sites (FIG. 44, area C).

(d) For a double-stranded DNA molecule, the methyladenine levels of the Watson and Crick strands did not belong to groups a, b, and c. Such a double-stranded molecule was defined as

a molecule with interlacing methylation patterns regarding adenine sites (FIG. 44, area D). The interlacing methylation patterns were defined as a mixture of methylated and unmethylated adenines present in a DNA strand.

[0304] In some other embodiments, the cutoffs of methyladenine levels for defining unmethylated strand may be, but are not limited to, less than 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5. The cutoffs of methyladenine levels for defining methylated strand would be but not limited to greater than 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99.

[0305] **FIG. 45** is a table showing the percentage of fully-unmethylated molecules, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns in training and testing datasets. The molecules in the testing dataset may be classified into fully-unmethylated molecules (7.0%) with respect to adenine sites, hemi-methylated molecules (9.8%), fully-methylated molecules (79.4%), and molecules with interlacing methyladenine patterns (3.7%). These results were comparable to the results shown in the training dataset, for which there were fully-unmethylated molecules (7.0%) with respect to adenine sites, hemi-methylated molecules (10.0%), fully-methylated molecules (79.4%), and molecules with interlaced methyladenine patterns (3.6%).

[0306] **FIG. 46** illustrates representative examples for molecules with fully-unmethylated molecules regarding adenine sites, hemi-methylated molecules, fully-methylated molecules, and molecules with interlaced methyladenine patterns. The white dots represent unmethylated adenines. The black dots represent methylated adenines. The horizontal lines with dots represent a strand of a double-stranded DNA molecule.

[0307] In embodiments, one may improve the performance in differentiating between methylated and unmethylated adenines by increasing the purity of 6mA bases which were used for training a CNN model. To this end, one may increase the time duration of DNA amplification reaction so that increased newly-produced DNA products can dilute the effect of unmethylated adenines contributed from original DNA templates. In other embodiments, one may incorporate biotinylated bases during DNA amplification with 6mA. The newly-produced DNA products with 6mA can be pulled down and enriched using streptavidin coated magnetic beads.

3. Uses of 6 mA methylation profiles

[0308] DNA 6mA modification is present in the genomes of bacteria, archaea, protists, and fungi (Didier W *et al.* Nat Rev Micorbiol. 2009;4:183-192). It was also reported that 6mA existed in the human genome, accounting for 0.051% of the total adenines (Xiao CL *et al.* Mol Cell. 2018;71:306-318). Considering the low content of 6mA in a human genome, in one embodiment, one can create a training dataset by adjusting the ratio of 6mA in dNTP mix (N represents unmodified A, C, G, and T) in the step of whole genome amplification. For example, one could use the ratio of 6mA to dNTP of 1:10, 1:100, 1:1000, 1:10000, 1:100000, or 1:1000000. In another embodiment, adenine DNA methyltransferase M. EcoGII may be used to create 6mA training dataset.

[0309] The amount of 6mA was lower in gastric and liver cancer tissues, and this 6mA downregulation correlated with increased tumorigenesis (Xiao CL *et al.* Mol Cell. 2018;71:306-318). On the other hand, it was reported that higher levels of 6mA were present in glioblastoma (Xie *et al.* Cell. 2018;175:1228-1243). Thus, the approach for 6mA as disclosed herein would be useful for studying cancer genomics (Xiao CL *et al.* Mol Cell. 2018;71:306-318; Xie *et al.* Cell. 2018;175:1228-1243). In addition, 6mA was found to be more prevalent and abundant in mammalian mitochondrial DNA, showing in association with hypoxia (Hao Z *et al.* Mol Cell. 2020; doi:10.1016/j.molcel.2020.02.018). Thus, the approach for 6mA detection in this disclosure would be useful for studying the mitochondrial stress response under different clinical conditions such as pregnancy, cancer, and autoimmune diseases.

IV. RESULTS AND APPLICATIONS

A. Detecting methylation

[0310] Detecting methylation at CpG sites using methods described above was performed for different biological samples and genomic regions. As an example, methylation determination with cell-free DNA in the plasma of pregnant women using single molecule, real-time sequencing was verified against methylation determination using bisulfite sequemcing. The methylation results may be used for different applications, including determining copy number

and diagnosing disorders. The methods described below are not limited to CpG sites and may also be applied to any modification described herein.

1. Detection of methylation for long DNA molecules in placenta tissue

[0311] Single molecule, real-time sequencing could sequence DNA molecules kilobases in length (Nattestad *et al.*, 2018). The deciphering of methylation states for CpG sites using the invention described here would allow one to infer the haplotype information of the methylation states by synergistically making use the long-read information of single molecule, real-time sequencing. To demonstrate the feasibility of inferring the long-read methylation states as well as its haplotype information, we sequenced a placenta tissue DNA with 478,739 molecules which were covered by 28,913,838 subreads. There were 7 molecules greater than 5 kb in size. Each was on average covered by 3 subreads.

[0312] **FIG. 47** shows that methylation states along the long DNA molecule with 6,265 bp in size (i.e. a haplotype block), which was sequenced in a ZMW with ZMW hole number m54276_180626_162240/40763503 and mapped to the genomic location of chr1:113246546-113252811 in the human genome. '-' represented the non-CpG nucleotide; 'U' represents the unmethylated state at a CpG site; and 'M' represents the methylated state at a CpG site. The region 4710 highlighted in yellow indicated a CpG island region which was known to be unmethylated in general (FIG. 47). The majority of CpG sites in that CpG island were deduced to be unmethylated (96%). In contrast, 75% of CpG sites outside the CpG island were deduced to be unmethylated. These results suggested that the methylation level outside of the CpG island (e.g. CpG island shore/shelf) was higher than that of the CpG island. The mixture of methylated and unmethylated states in a haplotype arrangement in the regions outside that CpG island would indicate the variability of methylation patterns. Such observations were generally in line with current understanding (Zhang *et al.*, 2015; Feinberg and Irizarry, 2010). Thus, this disclosure has enabled one to call different methylation states along a long molecule including methylation and unmethylation states, implying that the haplotype information of methylation states could be phased. Haplotype information refers to the linking of the methylation states of CpG sites on a contiguous stretch of DNA.

[0313] In one embodiment, we could use this approach herein for analyzing methylation states along a haplotype to detect and analyze the imprinted regions. Imprinted regions are subjected to epigenetic regulation that causes methylation states in a parent-of-origin fashion. For example, one important imprinted region is located on human chromosome 11p15.5 and contains the imprinted genes *IGF2,* H19, and *CDKN1C* (*P57^{kip2}*) which are strong regulators of fetal growth (Brioude *et al,* Nat Rev Endocrinol. 2018;14:229-249). The genetic and epigenetic aberrations in imprinted regions would be associated with diseases. Beckwith-Wiedemann syndrome (BWS) is an overgrowth syndrome, with patients often presenting with macroglossia, abdominal wall defects, hemihyperplasia, enlarged abdominal organs and an increased risk of embryonal tumors during early childhood. BWS is considered to be caused by genetic or epigenetic defects within 11p15.5 regions (Brioude *et al,* Nat Rev Endocrinol. 2018;14:229-249). A region called ICR1 (imprinting control region 1) which is located between *H19* and *IGF2* is differentially methylated on the paternal allele. ICR1 directs parent of origin-specific expression of *IGF2.* Thus, the genetic and epigenetic aberrations in ICR1 would lead to aberrant expression of *IGF2* which is one of possible reasons causing BWS. Thus, the detection of methylation states along the imprinted regions would be of clinical significance.

[0314] We downloaded data for 92 imprinted genes from a public database which curates currently reported imprinted genes (http://www.geneimprint.org/). The regions 5-kb upstream and downstream of these imprinted genes were used for further analysis. Among these regions, 160 CpG islands are associated with these imprinted genes. We obtained 324,248 circular consensus sequences from a placenta sample. After removing the circular consensus sequences with low quality and short overlapped regions with the CpG islands (e.g. smaller than 50% of the length of that relevant CpG island), we obtained 9 circular consensus sequences overlapping with 9 CpG islands which corresponded to 8 imprinted genes.

[0315] **FIG. 48** is a table showing that the 9 DNA molecules were sequenced by single molecule, real-time sequencing and overlapping with imprinted regions, including H19, WT1-AS, WT1, DLK1, MEG3, ATP10A, LRRTM1, and MAGI2. The 6th column contained the DNA stretches overlapping with CpG islands involving the imprinted regions. 'U' represents an unmethylated cytosine at the CpG context; 'M' represents a methylated cytosine in the CpG

context. '*' represents a CpG site that was not covered in the sequencing result; '-' represents a nucleotide from non-CpG sites; the genotype is indicated in the brackets if a molecule overlaps with a single nucleotide polymorphism (SNP). The 7th column indicates the methylation states for a whole molecule. A molecule can be called as methylated if the majority of CpG sites (e.g. more than 50%) were shown to be methylated according to embodiments present in this disclosure; otherwise it would be called as unmethylated.

[0316] Among 9 DNA molecules, 5 DNA molecules (55.6%) were called as methylated, which was not significantly deviated from the expectation in which 50% of DNA molecules would be methylated. As shown in the 6th column of the table of FIG. 48, the majority of the CpG sites were shown to methylated or unmethylated in a concerted manner, i.e., as a methylation haplotype. One embodiment is that a molecule would be called as methylated if the majority of CpG sites (e.g. more than 50%) were shown to be methylated according to embodiments present in this disclosure, it would be called as unmethylated otherwise. Other cutoffs, for determining if a molecule is methylated or not, could be used, for example, but not limited to, at least 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of CpG sites in a molecule were analyzed to be deemed methylated.

[0317] In another embodiment, we could use molecules concurrently comprising at least one SNP and at least one CpG site analysis to determine if a region might be associated with an imprinted region or a known imprinted gene might be aberrant (e.g. loss of imprinting). For illustration purposes, **FIG. 49** shows that the first molecule from an imprinting region carried the allele 'A'; and the second molecule from that imprinting region carried the allele 'G'. Assuming that the imprinting region was paternally imprinted, the first molecule from the maternal haplotype was fully unmethylated; and the second molecule from the paternal haplotype was fully methylated. In one embodiment, such an assumption would provide the ground truth of methylation states, allowing for testing the performance of base modification detection according to the embodiments present in this disclosure.

[0318] FIG. 49 shows an example for the determination of methylation patterns in an imprinted region. DNA in a biologic sample was extracted and ligated with hairpin adapters to

form the circular DNA molecules. Sequence information and base modifications (e.g. methylation states at CpG sites) regarding those circular DNA molecules were not known. Those circular DNA molecules were subjected to single molecule, real-time sequencing. IPDs, PWs, and sequence context for bases in each subread originating from those circular DNA molecules were determined after the subreads were mapped to the reference genome. In addition, genotypes of those molecules were determined. IPDs, PWs, and sequence context in a measurement window associated with CG sites would be compared with the reference kinetic patterns according to the embodiments present in this disclosure to determine the methylation states for each CpG. If two molecules with different alleles showed different methylation patterns in a way that one was fully unmethylated and the other was fully methylated, the genomic region associated with these two molecules would be an imprinted region. In one embodiment, if such genomic region happened to be a known imprinted region, for example, as illustrated in FIG. 49, the methylation patterns for these two molecules were in line with the expected methylation patterns (i.e. ground truth) in a normal situation. It may suggest the accuracy of the methods for the classification of methylation states according to the embodiments present in this disclosure. In one embodiment, the derivation between measured methylation patterns according to embodiments present in this disclosure and expected methylation patterns would indicate the aberrations of imprinting, for example, the loss of imprinting.

[0319] **FIG. 50** shows an example for the determination of methylation patterns in an imprinted region. In one embodiment, the imprinting pattern could be further determined through analyzing the methylation patterns of that region across a certain pedigree tree. For example, the analysis of methylation patterns and allelic information across the paternal, maternal genomes and offspring could be performed. Such a pedigree tree could further include the paternal or maternal grandfather's, paternal or maternal grandmother's genomes or other relevant genomes. In another embodiment, such analysis could be extended into family trio (mother, father and child) datasets in a certain population, for example, obtaining methylation and genotype information for each individual according to embodiments present herein.

[0320] As shown after the classification, both genotype (allele in box) and the methylation status can be determined. For each of the molecules, a methylation pattern at each site can be

provided (e.g., all methylated or all unmethylated) so as to identify which parent the molecule is inherited from. Or, a methylation density can be determined, and one or more cutoffs can classify whether the molecule is hypermethylated (e.g., > 80% or other % and from one parent) or hypomethylated (e.g., <20% or other % and from the other parent).

2. Detection of methylation for cfDNA molecules

[0321] As another example, cell-free DNA (cfDNA) methylation has been also increasingly recognized as important molecular signals for non-invasive prenatal testing. For example, we have shown that cfDNA molecules from regions carrying tissue-specific methylation can be used for determining the proportional contributions from different tissues such as neutrophils, T cells, B cells, liver, placenta in the plasma of pregnant women (Sun *et al.*, 2015). The feasibility of using plasma DNA methylation of pregnant women to detect trisomy 21 has also been demonstrated (Lun *et al.*, 2013). cfDNA molecules in maternal plasma were fragmented with a median size of 166 bp, which is much shorter than artificially-fragmented *E. coli* DNA with approximately 500 bp in size. It has been reported that cfDNA is non-randomly fragmented, for example, end motifs of plasma DNA in association with the tissue origins such as from the placenta. Such characteristic properties of cell-free DNA give an extremely different sequence context from artificially-fragmented *E. coli* DNA. Thus, it remains unknown whether such polymerase kinetics would allow for quantitatively deducing the methylation levels, typically for cell-free DNA molecules. The disclosures in this patent application would be applicable to, but not limited to, cell-free DNA methylation analysis in the plasma of pregnant women, for example by using the methylation prediction model trained from above-said tissue DNA molecules.

[0322] Using single molecule, real-time sequencing, six plasma DNA samples of pregnant women with a male fetus were sequenced with a median of 30,738,399 subreads (range: 1,431,215-105,835,846), corresponding to a median of 111,834 CCS (range: 61,010-503,582). Each plasma DNA was sequenced with a median of 262 times (range: 173-320). The data set was generated from DNA prepared by Sequel I Sequencing Kit 3.0

[0323] To evaluate the detection of methylation for cfDNA molecules, we used bisulfite sequencing (Jiang *et al.*, 2014) to analyze the methylation of the above-said 6 plasma DNA samples of pregnant women. We obtained a median of 66 million paired-end reads (58-82 million paired-end reads). The median overall methylation was found to be 69.6% (67.1%-72.0%).

[0324] **FIG. 51** shows the comparison of methylation levels deduced by the new approach and conventional bisulfite sequencing. The y-axis is the methylation levels predicted according to the embodiments present in this patent application. The x-axis is the methylation levels deduced by bisulfite sequencing. A median of 314,675 CpG sites (range: 144,546-1,382,568) were analyzed for plasma DNA results generated from single molecule, real-time sequencing. The median proportion of CpG sites that were predicted to be methylated was 64.7% (range: 60.8-68.5%), which appeared to be comparable with results deduced from bisulfite sequencing. As shown in FIG. 51, there was a good correlation (r: 0.96, p-value=0.0023) between overall methylation levels deduced by the single molecule, real-time sequencing with the present methylation prediction approach and bisulfite sequencing.

[0325] Because of the shallow depth of bisulfite sequencing, it might not be robust for deducing the methylation levels (i.e. the fraction of sequenced CpG being methylated) for each CpG in the human genome. Instead, we calculated the methylation levels in some regions with multiple CpG sites, by aggregating read signals covering CpG sites of a genomic region in which any two consecutive CpG sites were within 50 nt and the number of CpG sites was at least 10. The percentage of sequenced cytosine among the sum of sequenced cytosines and thymines across CpG sites in a region indicated the methylation levels of that region. The regions were divided into different groups according to regional methylation levels. The probability of methylation predicted by the model learned from the previous training datasets (i.e. tissue DNA) was elevated accordingly as the methylation levels increased (**FIG. 52A**). These results further suggested the feasibility and validity of using single molecule, real-time sequencing to predict methylation states of cfDNA molecules in pregnant women. **FIG. 52B** showed that the methylation level in a 10-Mb genomic window estimated using single molecule, real-time

sequencing according to the embodiments present in this disclosure was well corrected with that by bisulfite sequencing ($r = 0.74$; p-value < 0.0001)

[0326] **FIG. 53** showed that the genomic representations (GR) of Y chromosome in maternal plasma of pregnant women that was measured by single molecule, real-time sequencing were well correlated with those measured by BS-seq ($r = 0.97$; P-value $= 0.007$). These results suggested that single molecule, real-time sequencing would also allow the accurate quantification of DNA molecules originating from non-hematopoietic tissues such as placenta, whose contributed DNA generally represented a minority. In other words, this disclosure demonstrated the feasibility for simultaneously analyzing copy number aberrations and methylation states for native molecules without any base conversions and amplifications prior to sequencing.

3. CpG block-based method

[0327] Some embodiments can perform a methylation analysis on a number of genomic regions harboring multiple CpG sites, for example but not limited to 2, 3, 4, 5, 10, 20, 30, 40, 50, 100 CpG sites, etc. The size of such a genomic region can be, for example, but not limited to, 50, 100, 200, 300, and 500 nt, etc. The distance between CpG sites in this region could be, for example but not limited to 10, 20, 30, 40, 50, 100, 200, 300 nt, etc. In one embodiment, we could merge any two consecutive CpG sites within 50 nt to form a CpG block such that the number of CpG sites in this block was more than 10. In such a block-based method, multiple regions can be combined into one window represented as a single matrix, effectively treating the regions together.

[0328] As an example, as shown in **FIG. 54**, the kinetics of all subreads associated with a CpG block were used for methylation analysis. The projected IPD profiles of the upstream and downstream 10 nt flanking at each CpG in that block were artificially aligned relative to CpG sites to calculate the average IPD profile (**FIG. 54**). The word "projected" means that we had aligned the subread kinetic signals to each corresponding CpG site in question. The average IPD profiles for a CpG block were used for training a model (e.g. using artificial neural network, ANN for short) to identify the methylation states for each block. The ANN analysis included an

input layer, two hidden layers, and an output layer. Each CpG block was characterized by a feature vector of 21 IPD values that would be input to the ANN. The first hidden layer included 10 neurons with ReLu as an activation function. The second hidden layer included 5 neurons with ReLu as an activation function. Finally, the output layer included 1 neuron with Sigmoid as an activation function which would output the probability of methylation. A CpG site showing the probability of methylation > 0.5 was deemed as methylation, otherwise deemed as unmethylation. The average IPD profile may be used for analyzing the methylation state of a whole molecule. The whole molecule may be considered methylated if a certain number of sites above a threshold (e.g., 0, 1, 2, 3, etc.) are methylated or if the molecule has a certain methylation density.

[0329] There were 9,678 and 9,020 CpG blocks in unmethylated and methylated libraries, each of which harbored at least 10 CpG sites. Those CpG blocks covered 176,048 and 162,943 CpG sites for unmethylated and methylated libraries. As shown in **FIG. 55A** and **FIG. 55B**, we could achieve greater than 90% overall accuracy in predicting methylation states in both training dataset and testing dataset. However, such embodiment relied on CpG blocks would greatly reduce the number of CpGs that were able to be assessed. By definition, the requirement of the least number of CpG sites would restrict the methylation analysis to some particular genomic regions (e.g. preferentially analyzing CpG islands).

B. Determination of origin or disorder

[0330] Methylation profiles may be used to detect tissue origin or determine the classification of a disorder. Methylation profile analysis may be used in conjunction with other clinical data, including imaging, conventional blood panels, and other medical diagnostic information. Methylation profiles may be determined using any method described herein.

1. Determination of copy number aberration

[0331] This section shows that SMRT is accurate for determining copy number, and thus the methylation profile and copy number profile can be analyzed concurrently.