(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

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

优先权信息

无

技术领域

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

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。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。

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。

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。

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。

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

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：

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

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。

发明详细描述

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。

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。

确定胎儿基因组中预定区域碱基信息的方法

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：

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; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），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。

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（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;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 Methods 6 (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。

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 NCBI 36.3，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。

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。

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:

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。

详细分析步骤如下:

记号:

I.需要检测的位点数为 N。

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

其中，

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

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

$k\in\{0,1\}$， $i=1,2,3,...,N$。

III. 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。

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

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

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，

而所有可能出现的隐藏状态组成了集合 Q。

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

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

V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

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:

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

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。

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。

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

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

胎儿单倍型 $h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$， $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$ 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 $m_{i,0}$ ”。

第二步，构建观察序列概率矩阵:

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$（ $i = 1,2,3,...,N$， $j \in Q$ ），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

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

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

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

其中指示函数

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

。

此步是进行 HMM 参数，每个位点的观察序列概率分布 $b_{i,j}\left(s_i\right)$ 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。

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）:

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

定义 逆向指针 $\Psi_i\left(q_i\right) = \arg\max_{q_{i-1}\in Q}\delta_i\left(q_i\right)\cdot a_{q_{i-1}q_i}$

这里所使用的术语局部概率 $\delta_i\left(q_i\right)$ 和逆向指针 $\Psi_i\left(q_i\right)$ 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 Lawrence R. 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。

第四步，确定最终状态，并回溯最优路径:

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

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

第五步，输出结果。

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:

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。

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。

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。

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。

用于确定胎儿基因组中预定区域碱基信息的系统

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。

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。

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（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;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 Methods 6 (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 NCBI 36.3，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。

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。

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，

采用

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

作为重组转移矩阵， p_r=re/N。

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:

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

其中，

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

。

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。

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。

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。

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。

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。

计算机可读介质

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。

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，

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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:

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

其中，

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

。

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。

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。

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。

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。

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。

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。

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。

一般方法

本发明实施例的主要步骤包括:

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。
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。
3) 各遗传物质根据不同测序平台测序要求进行文库构建。
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。
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。

分析模型:

记号:

I.需要检测的位点数为 N。

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

其中，

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

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

$k \in \{0,1\}$， $i = 1,2,3,...,N$。

III. 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。

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

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

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，

而所有可能出现的隐藏状态组成了集合 Q。

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

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

V. 平均胎儿浓度和平均测序错误率分别记为 ε和 e。

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:

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

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。

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。

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

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

第二步，构建观察序列概率矩阵:

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$（ $i = 1,2,3,...,N$， $j \in Q$），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

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

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，该碱基出现的可能性”，即

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

其中指示函数

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

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）:

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

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

第四步，确定最终状态状态，并回溯最优路径:

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

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

第五步，输出结果。

实施例 1

样品收集及处理:

（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。

（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。

（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®Bioanalyzer 2100 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）。

（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。

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：

末端修复反应体系：

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	

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。

末端添加碱基 A 反应体系：

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

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

接头连接反应体系：

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

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。

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

反应程序如下:

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

使用 PCR Purification Kit(QIAGEN)回收 PCR 产物。样品最后溶于 50μl 的 EB 缓冲液中。

构建好的文库经 Agilent®Bioanalyzer 2100 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）。

（5）父母基因组测序分型:

a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

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。

c. 提取出标记位点的基因型。

（6）父母单倍体型推断:

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。

b. 使用 BEAGLE 对父母的单倍型进行推断。

（7）胎儿单倍体型推断:

a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

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。

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）。

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）。

d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

e. 确定最终状态状态，并回溯最优路径。

f. 输出。

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:

			母亲						合计		
			纯合			杂合					
			位点数	准确数	准确率	位点数	准确数	准确率	位点数	准确数	准确率
常染色体	父亲	纯合	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		杂合	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
X染色体			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

工业实用性

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。

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。

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。

权利要求书

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。

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，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。

8、根据权利要求 7 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，

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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:

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

其中，

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

。

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。

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，所述分析装置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。

17、根据权利要求 16 所述的系统，其特征在于，所述惠特比算法采用 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，

采用

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

作为重组转移矩阵，其中，p_r=re/N。

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：

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

其中，

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

。

19、一种计算机可读介质，其特征在于，所述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。

20、根据权利要求 19 所述的计算机可读介质，其特征在于，所述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。

21、根据权利要求 20 所述的计算机可读介质，其特征在于，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，

采用

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

作为重组转移矩阵，其中，p_r=re/N。

22、根据权利要求 19 所述的计算机可读介质，其特征在于，所述指令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。

23、根据权利要求 22 所述的计算机可读介质，其特征在于，所述指令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:

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

其中，

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

。

24、根据权利要求 19 所述的计算机可读介质，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。

25、根据权利要求 24 所述的计算机可读介质，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。

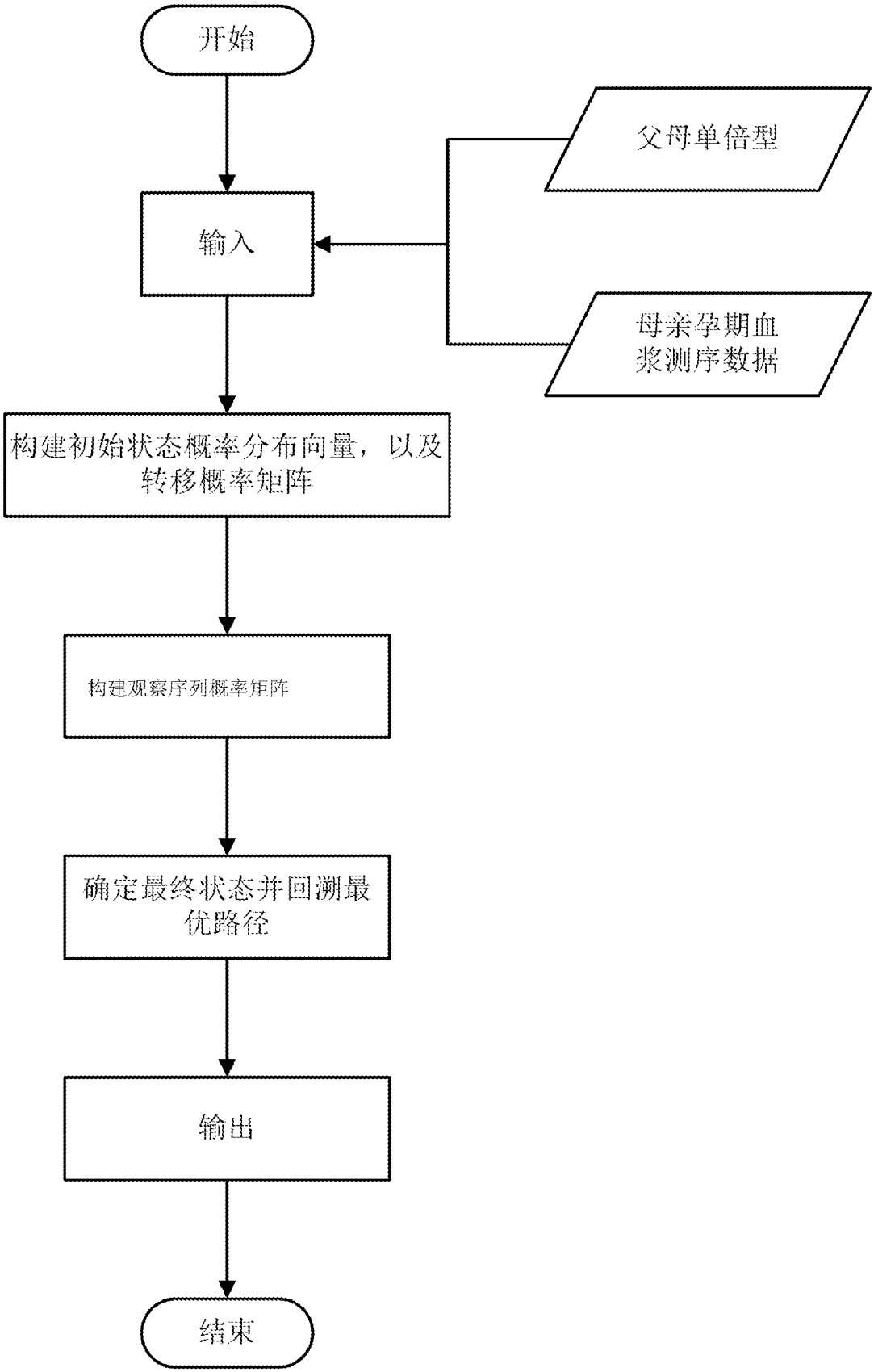
开始
父母单倍型
输入
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

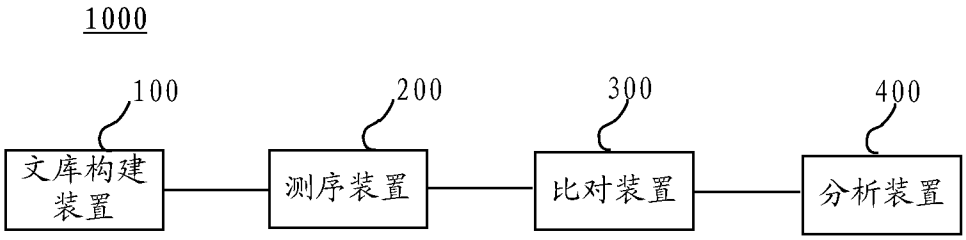

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

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)

IPC: 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)

CNABS, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, GOOGLE SCHOLAR, Elsevier Science, PubMed: hidden Markov model, HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) , claims 1-7, description, page 3, paragraph [0035] to page 4, paragraph [0050]	1-18
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA INC.) 11 January 2012 (11.01.2012) , claims 1 and 2, description, page 32, paragraph [0003] to page 33, paragraph [0001]	1-18

☐ 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
27 January 2013 (27.01.2013)	14 February 2013 (14.02.2013)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer ZHAO, Yanhao Telephone No. (86-10) 62411043

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:19-25
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 19-25 are directed to mere presentations of information (PCT Rule 39. 1(v)).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/075478

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102127818 A	20.07.2011	None	
CN 102317473 A	11.01.2012	WO 2010/068289 A2	17.06.2010
		US 2011/183320 A1	28.07.2011
		US 2010/221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010/068289 A3	14.10.2010
		AU 2009/325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011

国际检索报告

国际申请号
PCT/CN2012/075478

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNABS,CPRSABS,SIPOABS,DWPI,CNTXT,WOTXT,EPTXT,USTXT,CNKI,GOOGLE SCHOLAR, Elsevier Science, PubMed：隐马尔可夫模型，惠特比算法，文库，测序，隐马可夫, 隐马氏，隐 MarkoV, 胎儿，hidden Markov model,HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN102127818 A (张康) 20.7 月 2011（20.07.2011）,权利要求 1-7，说明书第 3 页第 35 段-说明书第 4 页第 50 段	1-18
Y	CN102317473 A (加利福尼亚太平洋生物科学股份有限公司) 11.1 月 2012（11.01.2012）,权利要求 1-2，说明书第 32 页第 3 段-第 33 页第 1 段	1-18

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
27.1 月 2013(27.01.2013)	**14.2 月 2013 (14.02.2013)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 赵彦豪 电话号码：(86-10) **62411043**

国际检索报告

国际申请号
PCT/CN2012/075478

第II栏　某些权利要求被认为是不能检索的意见(续第 1 页第 2 项)

根据条约第 17 条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：19-25

因为它们涉及不要求本单位进行检索的主题，即：

权利要求 19-25 涉及单纯的信息提供（PCT 细则 39.1 (v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则 6.4(a)第 2 句和第 3 句的要求撰写。

第III栏　缺乏发明单一性的意见(续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。
具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明：
☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。
☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。
☐ 缴纳附加检索费时未提交异议书。

国际检索报告

关于同族专利的信息

国际申请号
PCT/CN2012/075478

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN102127818 A	20.07.2011	无	
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011