



# (12)发明专利申请

(10)申请公布号 CN 107133491 A

(43)申请公布日 2017.09.05

(21)申请号 201710140534.8

(22)申请日 2017.03.08

(71)申请人 广州市达瑞生物技术股份有限公司

地址 510665 广东省广州市高新技术产业  
开发区荔枝山路6号

(72)发明人 胡天亮 欧阳国军 翁荣涛

杨学习 梁志坤

(74)专利代理机构 广州粤高专利商标代理有限

公司 44102

代理人 禹小明 凌衍芬

(51)Int.Cl.

G06F 19/10(2011.01)

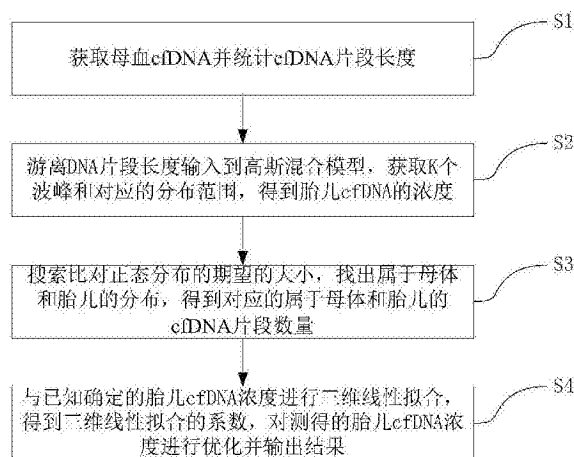
权利要求书2页 说明书4页 附图1页

## (54)发明名称

一种获取胎儿游离DNA浓度的方法

## (57)摘要

本发明公开一种获取胎儿游离DNA浓度的方法,对母体外周血中的cfDNA片段长度的数据,采用拥有K个正态分布的高斯混合模型,对母体和胎儿的cfDNA进行量化,自动准确的获取五个波峰和对应的分布范围,得到胎儿cfDNA的浓度,为产前无创检测(NIPT)提供更加合适和可靠的胎儿浓度,具有普适性和准确性。本方法通过动态确定母体和胎儿cfDNA片段长度的分布区域,对男胎和女胎都有效,并且对于不同胎儿以及胎龄的样本,能够自动获取并识别母体和胎儿cfDNA的分布,保证了胎儿浓度的准确度。



1. 一种获取胎儿游离DNA浓度的方法,其特征在于,包括以下步骤:

S1:获取母血游离DNA,并统计游离DNA片段长度;

S2:将统计游离DNA片段长度输入到高斯混合模型,采用拥有K个正态分布的高斯混合模型,对母体和胎儿的游离DNA进行量化,获取K个波峰和对应的分布范围,得到胎儿游离DNA的浓度;

S3:通过搜索比对正态分布的期望的大小,从K个正态分布分布中找出属于母体和胎儿的分布,得到对应的属于母体和胎儿的游离DNA片段数量 $N_i$ ,其中i表示1到K中属于母体和胎儿的编号;

S4:与已知确定的胎儿游离DNA浓度进行三维线性拟合,得到三维线性拟合的系数,对测得的胎儿游离DNA浓度进行优化并输出结果。

2. 根据权利要求1所述的获取胎儿游离DNA浓度的方法,其特征在于,步骤S2中,所述高斯混合模型表示为:

$$p(x) = \sum_{i=1}^K \pi_i N(x|\mu_i, \sigma_i), (i = 1, 2, \dots, K)$$

其中,K为正整数, $N(x|\mu_i, \sigma_i)$ 为正态分布, $\mu_i$ 表示期望, $\sigma_i$ 表示方差,样本x以 $\pi_i$ 的概率隶属于正态分布 $N(x|\mu_i, \sigma_i)$ 。

3. 根据权利要求2所述的获取胎儿游离DNA浓度的方法,其特征在于,步骤S2中,具体步骤包括:

S21:计算样本 $x_j$ 发生的概率 $p(x_j)$ ,其中 $j=1 \cdots n$ ,n为正整数, $p(x_j)$ 的公式为:

$$p(x_j) = \sum_{i=1}^K \pi_i N(x_j|\mu_i, \sigma_i)$$

其中 $\sum_{i=1}^K \pi_i = 1$ ;

S22:则样本 $x_j$ 存在,第k( $k=1, 2, \dots, K$ )个正态分布发生的概率为:

$$r_{jk} = \frac{\pi_k N(x_j|\mu_k, \sigma_k)}{\sum_{i=1}^K \pi_i N(x_j|\mu_i, \sigma_i)}$$

S23:目标函数为: $\text{Max } P(X) = \prod_{j=1}^n p(x_j)$ ,通过最大似然法求得属于母体和胎儿的游离DNA片段数量:

$$N_i = \sum_{j=1}^n r_{jk}$$

更新参数:

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} x_j$$

$$\sigma_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} (x_j - \mu_i)^2$$

$$\pi_i = \frac{N_i}{n}$$

S24: 返回步骤S21, 循环直到实现最大化结束。

4. 根据权利要求2所述的获取胎儿游离DNA浓度的方法, 其特征在于, 取K的值为:  $K=5$ 。

## 一种获取胎儿游离DNA浓度的方法

### 技术领域

[0001] 本发明涉及无创产前检测领域,更具体地,涉及一种获取胎儿游离DNA浓度的方法。

### 背景技术

[0002] 无创产前基因检测(NIPT)通过采集孕妇外周血,提取其中的胎儿游离DNA。利用基因测序技术并结合生物信息学分析手段,便可准确判断胎儿是否患有染色体病。当胎儿游离DNA比例过低时可能因为胎儿DNA量太少而不能被检测出来染色体是否有异常,所以胎儿游离DNA浓度检测是无创产前基因检测中至关重要的步骤。

[0003] 现有成熟的胎儿游离DNA浓度检测技术都是基于Y染色体在cfDNA中的含量来确定的,由于女胎不存在Y染色体,因此这个方法不适用于女胎。也有通过划分固定区间来统计母体和胎儿cfDNA片段的数量,以求得胎儿浓度的方法。但是,由于不同样本、不同胎龄的样本,它们的cfDNA片段长度的分布不是固定不变的,该方法无法动态获取准确的统计区间,得到的结果的准确度也受到严重的影响。

### 发明内容

[0004] 本发明为克服上述现有技术所述的至少一种缺陷,提供一种获取胎儿游离DNA浓度的方法。

[0005] 为解决上述技术问题,本发明的技术方案如下:

[0006] 一种获取胎儿游离DNA浓度的方法,包括以下步骤:

[0007] S1:获取母血游离DNA(cfDNA),并统计游离DNA片段长度;

[0008] S2:将统计游离DNA片段长度输入到高斯混合模型,采用拥有K个正态分布的高斯混合模型,对母体和胎儿的游离DNA进行量化,获取K个波峰和对应的分布范围,得到胎儿游离DNA的浓度;

[0009] S3:通过搜索比对正态分布的期望的大小,从K个正态分布分布中找出属于母体和胎儿的分布,得到对应的属于母体和胎儿的游离DNA片段数量 $N_i$ ,其中i表示1到K中属于母体和胎儿的编号;

[0010] S4:与已知确定的胎儿游离DNA浓度进行三维线性拟合,得到三维线性拟合的系数,对测得的胎儿游离DNA浓度进行优化并输出结果。

[0011] 在一种优选的方案中,步骤S2中,所述高斯混合模型表示为:

$$[0012] \quad p(x) = \sum_{i=1}^K \pi_i N(x|\mu_i, \sigma_i), (i = 1, 2, \dots, K)$$

[0013] 其中,K为正整数, $N(x|\mu_i, \sigma_i)$ 为正态分布, $\mu_i$ 表示期望, $\sigma_i$ 表示方差,样本x以 $\pi_i$ 的概率隶属于正态分布 $N(x|\mu_i, \sigma_i)$ ;

[0014] 在一种优选的方案中,步骤S2中,具体步骤包括:

[0015] S21:计算样本 $x_j$ 发生的概率 $p(x_j)$ ,其中 $j=1 \cdots n$ ,n为正整数, $p(x_j)$ 的公式为:

$$[0016] \quad p(x_j) = \sum_{i=1}^K \pi_i N(x_j | \mu_i, \sigma_i)$$

[0017] 其中  $\sum_{i=1}^K \pi_i = 1$ ;

[0018] S22: 则样本  $x_j$  存在, 第  $k$  ( $k=1, 2, \dots, K$ ) 个正态分布发生的概率为:

$$[0019] \quad r_{jk} = \frac{\pi_k N(x_j | \mu_k, \sigma_k)}{\sum_{i=1}^K \pi_i N(x_j | \mu_i, \sigma_i)}$$

[0020] S23: 目标函数为:  $\text{Max } P(X) = \prod_{j=1}^n p(x_j)$ , 通过最大似然法求得属于母体和胎儿的游离DNA片段数量:

$$[0021] \quad N_i = \sum_{j=1}^n r_{jk}$$

[0022] 更新参数:

$$[0023] \quad \mu_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} x_j$$

$$[0024] \quad \sigma_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} (x_j - \mu_i)^2$$

$$[0025] \quad \pi_i = \frac{N_i}{n}$$

[0026] S24: 返回步骤S21, 循环直到实现最大化结束。

[0027] 在一种优选的方案中, 取  $K$  的值为:  $K=5$ 。

[0028] 与现有技术相比, 本发明技术方案的有益效果是: 本发明提供一种获取胎儿游离DNA浓度的方法, 对母体外周血中的cfDNA片段长度的数据, 采用拥有  $K$  个正态分布的高斯混合模型, 对母体和胎儿的cfDNA进行量化, 自动准确的获取五个波峰和对应的分布范围, 得到胎儿cfDNA的浓度, 为产前无创检测 (NIPT) 提供更加合适和可靠的胎儿浓度, 具有普适性和准确性。本方法通过动态确定母体和胎儿cfDNA片段长度的分布区域, 对男胎和女胎都有效, 并且对于不同胎儿以及胎龄的样本, 能够自动获取并识别母体和胎儿cfDNA的分布, 保证了胎儿浓度的准确度。

## 附图说明

[0029] 图1为本发明获取胎儿游离DNA浓度的方法的流程图。

[0030] 图2为本发明步骤S2的具体流程图。

## 具体实施方式

[0031] 下面结合附图和实施例对本发明的技术方案做进一步的说明。

[0032] 实施例1

[0033] 如图1所示, 一种获取胎儿游离DNA浓度的方法, 包括以下步骤:

[0034] S1: 获取母血游离DNA (cfDNA), 并统计游离DNA片段长度;

[0035] S2:将统计游离DNA片段长度输入到高斯混合模型,采用拥有K个正态分布的高斯混合模型,对母体和胎儿的游离DNA进行量化,获取K个波峰和对应的分布范围,得到胎儿游离DNA的浓度;

[0036] S3:通过搜索比对正态分布的期望的大小,从K个正态分布分布中找出属于母体和胎儿的分布,得到对应的属于母体和胎儿的游离DNA片段数量 $N_i$ ,其中i表示1到K中属于母体和胎儿的编号;

[0037] S4:与已知确定的胎儿游离DNA浓度进行三维线性拟合,得到三维线性拟合的系数,对测得的胎儿游离DNA浓度进行优化并输出结果。

[0038] 在具体实施过程中,步骤S2中,所述高斯混合模型表示为:

$$[0039] \quad p(x) = \sum_{i=1}^K \pi_i N(x|\mu_i, \sigma_i), (i = 1, 2, \dots, K)$$

[0040] 其中,K为正整数, $N(x|\mu_i, \sigma_i)$ 为正态分布, $\mu_i$ 表示期望, $\sigma_i$ 表示方差,样本x以 $\pi_i$ 的概率隶属于正态分布 $N(x|\mu_i, \sigma_i)$ ;

[0041] 如图2所示,在具体实施过程中,步骤S2中,具体步骤包括:

[0042] S21:计算样本 $x_j$ 发生的概率 $p(x_j)$ ,其中 $j=1 \cdots n$ ,n为正整数, $p(x_j)$ 的公式为:

$$[0043] \quad p(x_j) = \sum_{i=1}^K \pi_i N(x_j|\mu_i, \sigma_i)$$

[0044] 其中 $\sum_{i=1}^K \pi_i = 1$ ;

[0045] S22:则样本 $x_j$ 存在,第k( $k=1, 2, \dots, K$ )个正态分布发生的概率为:

$$[0046] \quad r_{jk} = \frac{\pi_k N(x_j|\mu_k, \sigma_k)}{\sum_{i=1}^K \pi_i N(x_j|\mu_i, \sigma_i)}$$

[0047] S23:目标函数为: $\text{Max } P(X) = \prod_{j=1}^n p(x_j)$ ,通过最大似然法求得属于母体和胎儿的游离DNA片段数量:

$$[0048] \quad N_i = \sum_{j=1}^n r_{jk}$$

[0049] 更新参数:

$$[0050] \quad \mu_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} x_j$$

$$[0051] \quad \sigma_i = \frac{1}{N_i} \sum_{j=1}^n r_{jk} (x_j - \mu_i)^2$$

$$[0052] \quad \pi_i = \frac{N_i}{n}$$

[0053] S24:返回步骤S21,循环直到实现最大化结束。

[0054] 在具体实施过程中,取K的值为: $K=5$ 。

[0055] 本发明提供一种获取胎儿游离DNA浓度的方法,对母体外周血中的cfDNA片段长度

的数据,采用拥有K个正态分布的高斯混合模型,对母体和胎儿的cfDNA进行量化,自动准确的获取五个波峰和对应的分布范围,得到胎儿cfDNA的浓度,为产前无创检测(NIPT)提供更加合适和可靠的胎儿浓度,具有普适性和准确性。本方法通过动态确定母体和胎儿cfDNA片段长度的分布区域,对男胎和女胎都有效,并且对于不同胎儿以及胎龄的样本,能够自动获取并识别母体和胎儿cfDNA的分布,保证了胎儿浓度的准确度。

[0056] 显然,本发明的上述实施例仅仅是为清楚地说明本发明所作的举例,而并非是对本发明的实施方式的限定。对于所属领域的普通技术人员来说,在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,均应包含在本发明权利要求的保护范围之内。

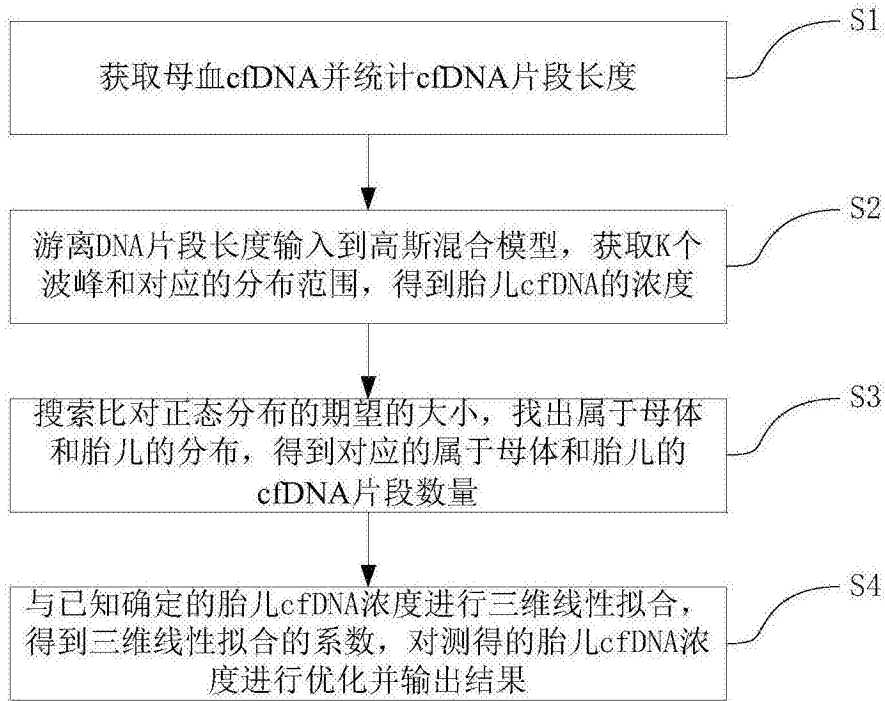


图1

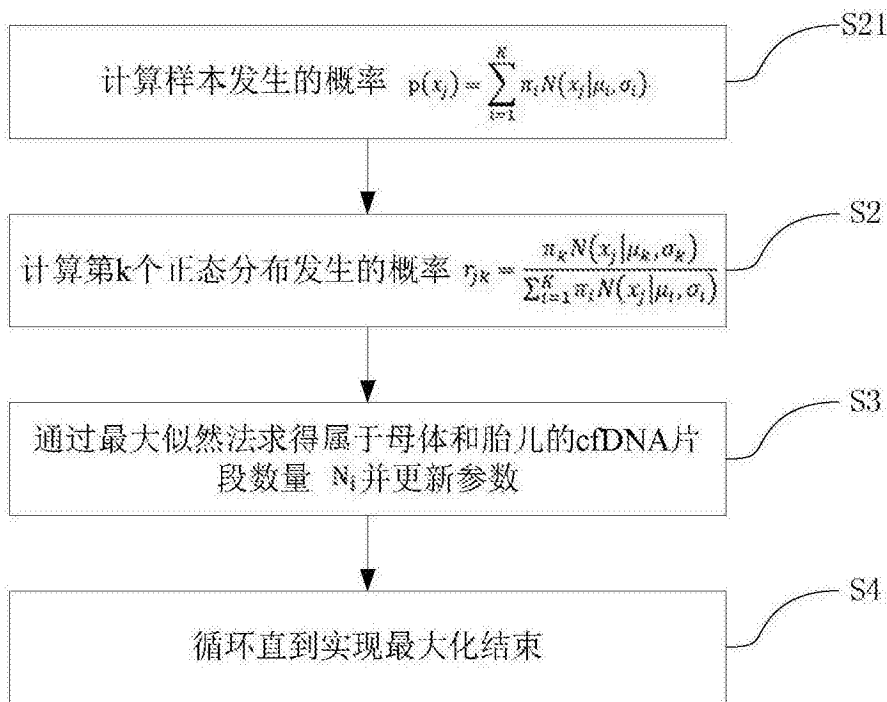


图2