(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105648045 B
(45)授权公告日 2019.10.11

(21)申请号 201410639577.7

(22)申请日 2014.11.13

(65)同一申请的已公布的文献号
申请公布号 CN 105648045 A

(43)申请公布日 2016.06.08

(73)专利权人 天津华大基因科技有限公司
地址 300308 天津市滨海新区空港经济区环河北路80号空港商务园东区3号楼101、201室
专利权人 深圳华大基因科技有限公司

(72)发明人 袁媛 王垚燊 朱红梅 易鑫

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东

(51)Int.Cl.
C12Q 1/6869(2018.01)
G16B 30/00(2019.01)
C12M 1/34(2006.01)

(56)对比文件
CN 102770558 A,2012.11.07,
WO 2011/041485 A1,2011.04.07,
洪萍.检测孕妇血浆中游离胎儿DNA行产前诊断的研究进展.《临床检验杂志》.2006,第24卷(第3期),第225-227页.
Cristina Conzalez-Gonzalez.Application of fetal DNA detection in maternal plasma:A prenatal diagnosis unit experience.《Journal of Histochemistry》.2005,第53卷(第3期),第307-314页.

审查员 游浩峰

权利要求书2页 说明书12页 附图3页

(54)发明名称

确定胎儿目标区域单体型的方法和装置

(57)摘要

本发明提供了一种确定胎儿目标区域单体型的方法及其装置。确定胎儿目标区域单体型的方法包括:对孕妇体液中游离核酸的目标区域进行序列测定,以便获得第一测序数据;对胎儿家系成员的相同目标区域进行序列测定,以便获得第二测序数据、第三测序数据和第四测序数据,其中,第二测序数据为胎儿母亲的测序数据,第三测序数据为胎儿父亲的测序数据,第四测序数据为先证者的测序数据;基于第一、第二以及任选的第三测序数据,确定述孕妇体液中的胎儿核酸含量;基于第二、第三和第四测序数据,分别构建胎儿母亲的目标区域单体型和胎儿父亲的目标区域单体型;以及基于胎儿母亲、父亲的目标区域单体型以及胎儿核酸含量,确定胎儿的目标区域单体型。

CN 105648045 B

1.一种确定胎儿目标区域单体型的方法,所述方法用于非疾病诊断目的,其特征在于,包括,

对孕妇体液中游离核酸的所述目标区域进行序列测定,以便获得第一测序数据;

对所述胎儿的家系成员的所述目标区域进行序列测定,以便获得第二测序数据、第三测序数据和第四测序数据,其中,所述第二测序数据为胎儿母亲的测序数据,所述第三测序数据为胎儿父亲的测序数据,所述第四测序数据为先证者的测序数据;

基于所述第一测序数据、第二测序数据以及任选的第三测序数据,确定所述孕妇体液中的胎儿核酸含量;

基于所述第二测序数据、第三测序数据和第四测序数据,分别构建所述胎儿母亲的目标区域单体型和所述胎儿父亲的目标区域单体型;以及

基于所述胎儿母亲的目标区域单体型、所述胎儿父亲的目标区域单体型以及所述胎儿核酸含量,确定所述胎儿的目标区域单体型;

其中,所述胎儿核酸含量是通过下列步骤确定的:

确定在所述第二测序数据和所述第三测序数据中均为不同纯合基因型的位点,其中,RR和rr表示不同的纯合基因型,R和r为一对等位基因,

基于公式f=g/(g+h)确定所述胎儿核酸含量,

其中,

g为所述第一测序数据中支持等位基因r的读段数目,h为所述第一测序数据中支持等位基因R的读段数目;

所述确定胎儿目标区域单体型,包括,

利用多个在父亲目标区域单体型上为杂合、在母亲目标区域单体型上为纯合的位点确定胎儿遗传到的父亲目标区域单体型,利用多个在父亲目标区域单体型上为纯合、在母亲目标区域单体型上为杂合的位点以及胎儿核酸含量确定胎儿遗传到的母亲目标区域单体型;

其中,对于所述多个在父亲目标区域单体型上为纯合、在母亲目标区域单体型上为杂合的位点,若有多个这样的位点符合R/r=(1+x%)/(1-x%),则判定胎儿遗传了母亲等位基因R所在的目标区域单体型,若有多个这样的位点符合R/r=1,则判定胎儿遗传了母亲等位基因r所在的目标区域单体型,R和r表示一对等位基因,x%表示胎儿核酸含量,R/r=第一测序数据中支持R的读段数目/第一测序数据中支持r的读段数目。

2.权利要求1的方法,其特征在于,对孕妇体液中游离核酸的所述目标区域进行序列测定包括:

利用探针对所述游离核酸进行捕获,所述探针特异性识别所述目标区域。

3.权利要求2的方法,其特征在于,所述探针是以芯片形式提供的。

4.权利要求2的方法,其特征在于,所述探针包括SNP位点探针,所述SNP位点探针在参考基因组上是唯一比对的。

5.权利要求2的方法,其特征在于,所述探针的GC含量为40-50%。

6.确定胎儿目标区域单体型的装置,包括,

测序单元,用于对孕妇体液中游离核酸的所述目标区域进行序列测定,以便获得第一测序数据,以及,对所述胎儿的家系成员的所述目标区域进行序列测定,以便获得第二测序

数据、第三测序数据和第四测序数据，其中，所述第二测序数据为胎儿母亲的测序数据，所述第三测序数据为胎儿父亲的测序数据，所述第四测序数据为先证者的测序数据；

胎儿核酸含量确定单元，与所述测序单元连接，用于基于所述第一测序数据、第二测序数据以及任选的第三测序数据，确定所述孕妇体液中的胎儿核酸含量；

父母单体型确定单元，与所述测序单元连接，用于基于所述第二测序数据、第三测序数据和第四测序数据，分别构建所述胎儿母亲的目标区域单体型和所述胎儿父亲的目标区域单体型；以及

胎儿单体型确定单元，与所述胎儿核酸含量确定单元和所述父母单体型确定单元相连，用于基于所述胎儿母亲的目标区域单体型、所述胎儿父亲的目标区域单体型以及所述胎儿核酸含量，确定所述胎儿的目标区域单体型；

其中，所述胎儿核酸含量确定单元通过下列步骤确定所述胎儿核酸含量：

确定在所述第二测序数据和所述第三测序数据中均为不同纯合基因型的位点，其中，RR和rr表示不同的纯合基因型，R和r为一对等位基因，

基于公式f＝g/(g+h)确定所述胎儿核酸含量，

其中，

g为所述第一测序数据中支持等位基因r的读段数目，h为所述第一测序数据中支持等位基因R的读段数目；

所述确定胎儿目标区域单体型，包括，

利用多个在父亲目标区域单体型上为杂合、在母亲目标区域单体型上为纯合的位点确定胎儿遗传到的父亲目标区域单体型，利用多个在父亲目标区域单体型上为纯合、在母亲目标区域单体型上为杂合的位点以及胎儿核酸含量确定胎儿遗传到的母亲目标区域单体型；

其中，对于所述多个在父亲目标区域单体型上为纯合、在母亲目标区域单体型上为杂合的位点，若有多个这样的位点符合R/r＝(1+x％)/(1-x％)，则判定胎儿遗传了母亲等位基因R所在的目标区域单体型，若有多个这样的位点符合R/r＝1，则判定胎儿遗传了母亲等位基因r所在的目标区域单体型，R和r表示一对等位基因，x％表示胎儿核酸含量，R/r＝第一测序数据中支持R的读段数目/第一测序数据中支持r的读段数目。

7.权利要求6的装置，其特征在于，所述目标区域包括SMN1基因的外显子区。

8.权利要求7的装置，其特征在于，所述目标区域还包括SMN1基因内部及SMN1基因上下游各3M区域中的次等位碱基频率为0.3-0.5的SNP位点。

确定胎儿目标区域单体型的方法和装置

技术领域

[0001] 本发明涉及生物信息领域,特别地,涉及确定胎儿目标区域单体型的方法和装置。

背景技术

[0002] 脊髓性肌萎缩症(Spinal Muscular Atrophy,SMA)是一组常见的常染色体隐性遗传病,居致死性常染色体隐性遗传病的第二位,活产婴儿中患者发生率为1/6000～1/10000。目前的研究已经表明,SMA的病因主要是SMN基因缺失:其中SMN1为决定基因,表达完整而稳定的SMN功能蛋白,而SMN2为SMA的修饰基因。据报道,98.7％(226/229)儿童型患者存在SMN1基因缺失,其中约90％的SMA病人显示纯合SMN1外显子7和/或8缺失。SMA神经肌肉疾病病情严重,目前临床上无有效治疗手段。产前诊断是预防该出生缺陷的重要手段。

[0003] 随着孕妇外周血浆中胎儿游离DNA存在的发现,为无创产前检测胎儿基因型提供了可能。然而目前尚未见到有关通过孕妇血浆游离DNA进行无创胎儿SMA检测的报道。已有的SMA检测报道,多是通过诊断SMN17号外显子设计QPCR引物及探针,实现对缺失型SMN1突变的检测,如徐湘民等人公开的“一种诊断人类脊髓性肌萎缩症的荧光定量PCR试剂盒”(公开号CN103614477A)。然而由于母体血浆中胎儿DNA含量相对较低,而QPCR的灵敏度不足以在高母体背景下实现对胎儿SMN1基因突变情况的检测。

[0004] 因此,发展一种可以无创检测胎儿SMN1基因型的检测方法,将对该疾病的产前诊断起到十分重要的作用。

发明内容

[0005] 依据本发明的一方面,提供一种确定胎儿目标区域单体型的方法,该方法包括以下步骤:对孕妇体液中游离核酸的所述目标区域进行序列测定,以便获得第一测序数据;对所述胎儿的家系成员的所述目标区域进行序列测定,以便获得第二测序数据、第三测序数据和第四测序数据,其中,所述第二测序数据为胎儿母亲的测序数据,所述第三测序数据为胎儿父亲的测序数据,所述第四测序数据为先证者的测序数据;基于所述第一测序数据、第二测序数据以及任选的第三测序数据,确定所述孕妇体液中的胎儿核酸含量;基于所述第二测序数据、第三测序数据和第四测序数据,分别构建所述胎儿母亲的目标区域单体型和所述胎儿父亲的目标区域单体型;以及,基于所述胎儿母亲的目标区域单体型、所述胎儿父亲的目标区域单体型以及所述胎儿核酸含量,确定所述胎儿的目标区域单体型。其中,第一、第二、第三和第四测序数据的获得没有必需遵循的先后关系,可同时获得,也可一个个获得或几个几个一起获得;胎儿核酸含量的确定步骤和父母单体型的构建步骤也没有先后关系。

[0006] 依据本发明的另一方面,提供一种确定胎儿目标区域单体型的装置,该装置能够执行本发明一方面提供的方法的部分或全部步骤,该装置包括:测序单元,用于对孕妇体液中游离核酸的所述目标区域进行序列测定,以便获得第一测序数据,以及,对所述胎儿的家系成员的所述目标区域进行序列测定,以便获得第二测序数据、第三测序数据和第四测序

数据，其中，所述第二测序数据为胎儿母亲的测序数据，所述第三测序数据为胎儿父亲的测序数据，所述第四测序数据为先证者的测序数据；胎儿核酸含量确定单元，与所述测序单元连接，用于基于所述第一测序数据、第二测序数据以及任选的第三测序数据，确定所述孕妇体液中的胎儿核酸含量；父母单体型确定单元，与所述测序单元连接，用于基于所述第二测序数据、第三测序数据和第四测序数据，分别构建所述胎儿母亲的目标区域单体型和所述胎儿父亲的目标区域单体型；以及胎儿单体型确定单元，与所述胎儿核酸含量确定单元和所述父母单体型确定单元相连，用于基于所述胎儿母亲的目标区域单体型、所述胎儿父亲的目标区域单体型以及所述胎儿核酸含量，确定所述胎儿的目标区域单体型。

[0007] 本发明的一方面的方法和/或装置，提供了一种基于目标区域捕获及家系目标区域单体型连锁分析的方法，通过连锁分析从孕妇体液样本比如孕妇外周血浆DNA测序数据中推断胎儿目标区域基因型，可用于判断或辅助判断胎儿是否患有目标区域变异相关疾病或异常。本发明的方法或装置通过利用连锁单体型信息极大的降低了假阳性及假阴性的发生。该方法和/或装置的应用可以极大的避免由于单个位点测量比例不准，单个位点测序错误等方面带来的假阴性和假阳性结果，使得检测结果更加准确可靠。通过该方法在SMN1患病高风险家庭的应用，可以有效检出患病儿，并减少不必要的羊水穿刺等有创取样手术。

附图说明

[0008] 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方式的描述中将变得明显和容易理解，其中：

[0009] 图1是本发明的一个具体实施方式中的确定胎儿目标区域单体型的装置的示意图；

[0010] 图2是本发明的一个具体实施方式中的胎儿基因型判断的整体技术线路示意图；

[0011] 图3是本发明的一个具体实施方式中的胎儿基因型判断结果，图3A为胎儿从父亲处所遗传到的单体型的判断结果图，图3B为胎儿从母亲处所遗传的单体型的判断结果图；图中，点表示一个snp位点遗传自Hap0的概率与遗传自Hap1的概率的差值，圈线为组合判断结果。

具体实施方式

[0012] 依据本发明的一种实施方式，提供一种确定胎儿目标区域单体型的方法，包括以下步骤：

[0013] 步骤一：获得第一、第二、第三和第四测序数据。

[0014] 获得孕妇体液中的游离核酸，捕获目标区域，对所述捕获得的目标区域进行序列测定，获得第一测序数据。孕妇体液样本为包含胎儿核酸的样本，比如孕妇外周血血浆包含胎儿核酸，提取的外周血游离核酸是孕妇和胎儿核酸的混合物，混合物是高度片段化的。依据现有测序平台，通过对从孕妇外周血样本提取的游离核酸进行测序文库构建，利用探针或芯片或液相探针捕获获得目标区域测序文库，对目标区域测序文库进行上机测序，获得第一测序数据，第一测序数据是孕妇核酸和胎儿核酸混合物的混合数据。测序平台包括但不限于CG(Complete Genomics)、Illumina/Solexa、Life Technologies ABI SOLiD和Roche 454，可根据所选用的测序平台进行相应的测序文库制备，可选择单端或双端测序，

由此获得的各个测序数据由多个短序列组成，将各个短序列称为读段。捕获所用的芯片是由固相基质和固定在其上的多个探针组成的，探针能够特性识别目标区域，目标区域可以是待测样本基因组DNA的一部分也可以是整个基因组，在本发明的一个具体实施方式中，目标捕获区域包括SMN1基因外显子区，表1显示各个外显子区域在参考基因组HG19上的位置，目标区域还包括SMN1基因内部及其上下游3M区域内高杂合率的SNP位点，表2为这些SNP在各个区域的数量分布，这些SNP的次等位基因频率(MAF)在0.3-0.5之间。这些区域以及位点信息有利于判断分析胎儿单体型，目标基因上下游的3M区域的捕获使得重组概率减少到万分之一以下，使得后续能够准确的进行单体型构建或确定，而且上述高杂合率的SNP位点的捕获，使得容易获得来源于胎儿自身的特定位点或序列，利用来自胎儿本身的位点或序列能够估算在混合DNA中胎儿的核酸含量。设计能够特异性识别上述区域的探针时，为保证捕获的特性性、检测的准确性，使包含至少一个上述SNP位点的探针在参考基因组上是唯一比对的，这样能增强探针捕获目标位点的特异性。在探针设计时，使每条探针的GC含量为40-50％，这样有利于在同一个体系中整组探针一起特异性结合目标区域、在同一个反应体系中能够一起洗脱下来。

[0015] 表1 SMN1基因区域捕获范围

[0016]

区域(Region)	染色体编号(chr)	起始位置(start)	终止位置(end)
1	chr5	70220738	70221835
2	chr5	70222126	70223263
3	chr5	70223351	70223620
4	chr5	70224046	70224569
5	chr5	70224596	70225332
6	chr5	70225421	70227146
7	chr5	70227276	70229560
8	chr5	70229641	70230603
9	chr5	70230671	70231084
10	chr5	70231091	70231402
11	chr5	70231511	70232075
12	chr5	70232161	70232534
13	chr5	70233276	70233724
14	chr5	70234111	70235041
15	chr5	70235136	70235933
16	chr5	70236016	70236631

[0017]

17	chr5	70236716	70239101
18	chr5	70239196	70239701
19	chr5	70239786	70241034
20	chr5	70241131	70242428
21	chr5	70242496	70242844

22	chr5	70243026	70243331
23	chr5	70243681	70244193
24	chr5	70244286	70244815
25	chr5	70245011	70245717
26	chr5	70247436	70248868

[0018] 表2 SMN1区单体型分析所用SNP位点分别情况

[0019]

region	SNP位点数
upstream10M-3M	7
upstream3M-2.5M	1
upstream2.5M-2M	14
upstream2M-1.5M	98
upstream1.5M-1M	52
upstream1M-500K	71
upstream500K-0K	66
Gene±1M	1629
downstream0K-500K	67
downstream500K-1M	26
downstream1M-1.5M	42
downstream1.5M-2M	78
downstream2M-2.5M	87
downstream2.5M-3M	0
downstream3M-10M	7

[0020] 获取胎儿家系成员的样本，包括胎儿生物学母亲(孕妇)、胎儿生物学父亲以及先证者的核酸样本，提取各个家系成员样本中的核酸，参考上述获取第一测序数据的方式，捕获胎儿家系成员核酸中的同样目标区域，对各个家系成员的同样目标区域进行序列测定，获得家系成员测序数据，所述家系成员测序数据包括第二、第三和第四测序数据，分别对应胎儿生物学母亲、胎儿生物学父亲和先证者的同样目标区域的测序数据。其中第二测序数据，即母亲测序数据的获得，可以通过分离上述获得第一测序数据的孕妇外周血样本，分离孕妇外周血样本获得孕妇外周血血浆样本和孕妇血细胞，从孕妇血细胞，比如白细胞，可以获得母亲基因组核酸，进而获得第二测序数据。先证者该家系中是确定带有目标区域相关变异的的成员，在这里，先证者是与待测胎儿同样生物学父母的胎儿的兄弟姐妹，包括出生的和未出生，包括体外培养的胚胎或受精卵，包括在世和不在世的。另外，在其他具体实施方式中，先证者也可以是待测胎儿的父母的兄弟姐妹，比如胎儿的舅舅、叔叔、姑姑等，这时，胎儿的家系成员的测序数据还应包括胎儿的祖父母和/或外祖父母，这样能够利用父母的兄弟姐妹的测序数据以及父母的测序数据构建祖父母或外祖父母的目标区域单体型，进而判断父母的遗传到的目标区域单体型。第一、第二、第三和第四测序数据的获得没有必需遵循的先后关系，可同时获得，比如利用标签标记多个样本，对多个样本核酸混合建库混合上机测序同时获得多个样本的测序数据，也可一个个获得或几个几个获得核酸样本的测序

数据。

[0021] 步骤二:确定胎儿核酸含量。

[0022] 基于第一和第二测序数据,或者基于第一、第二和第三测序数据,确定所述孕妇体液样本中的胎儿核酸含量。

[0023] 其中,基于第一和第二测序数据确定孕妇体液样本中的胎儿核酸含量,是这样进行的:首先是筛选出在第一测序数据中有两种基因型以及在第二测序数据中只有一种基因型的位点。位点的筛选可以通过比对来进行,比对可以利用SOAP(Short OligonucleotideAnalysis Package),bwa,samtools等软件进行,本实施方式对此不作限制,比对的进行也可以识别出多态性位点。比对所使用的参考序列是已知序列,可以是预先获得的目标个体所属生物类别中的任意的参考模板。例如,若目标个体是人类,参考序列可选择NCBI数据库提供的HG19。进一步地,也可以预先配置包含更多参考序列的资源库,在进行序列比对前,先依据目标个体的性别、人种、地域等因素选择或是测定组装出更接近的序列来作为参考序列,有助于获得更准确的检测分析结果。在比对过程中,根据比对参数的设置,各测序数据中的每条或每对读段(reads或一对末端读段pair-end reads)最多允许有n个碱基错配(mismatch),n优选为1或2,若reads中有超过n个碱基发生错配,则视为该条/对reads无法比对到参考序列。一个位点,假设在参考序列上该位点是A,第二测序数据的比对结果表明第二测序数据即母亲测序数据中比对上到参考序列该位点的碱基都是A,但是第一测序数据即母亲与胎儿的测序数据的比对结果表明第一测序数据中比对到参考序列该位点的碱基是A和另外一种非A的碱基,非A碱基比如T、C或G,由于第一测序数据中是母亲和胎儿核酸的混合测序数据,而从第二测序数据的比对结果可知母亲的该位点为AA,那么就可判断出第一测序数据中该位点非A碱基来源于胎儿,这样筛选出所有这样的位点,基于这些位点在混合测序数据中占的比例,就能反映出混合核酸中胎儿核酸的含量。类似的,若第二测序数据的比对结果表明母亲某位点的基因型为杂合的,比如AG,而第一测序数据比对结果显示支持该位点AG和AA两种基因型,这样基于第一测序数据中A碱基的数量、含量或比例,也能估算获得孕妇外周血样本中的胎儿核酸含量。当像上面前者情况,在第二测序数据中只有纯合基因型、而在第一测序数据中除有一样的纯合基因型还有杂合基因型时,胎儿核酸含量f=2d/(c+d),而当像上面后者情况,在第二测序数据中只有杂合基因型、而在第一测序数据中除有那杂合基因型还有纯合基因型,胎儿核酸含量f=(c-d)/(c+d),公式中的c为第一测序数据中支持等位基因A的读段数目,d为第一测序数据中支持非A等位基因的读段数目。

[0024] 基于第一、第二和第三测序数据确定孕妇体液样本中的胎儿核酸含量,是通过以下进行的:筛选出在第二测序数据和第三测序数据中为不同纯合基因型的位点,比如该位点在第二和第三测序数据中的基因型分别为RR和rr,这样以遗传角度,胎儿核酸中该位点的基因型为Rr,基于多个这种类型的位点计算孕妇外周血样本中胎儿核酸含量,胎儿核酸含量f=g/(g+h),g为第一测序数据中支持等位基因r的读段数目,h为第一测序数据中支持等位基因R的读段数目。位点的筛选涉及的比对,比对参数的设置、比对结果等可参照前面基于第一和第二测序数据估算胎儿核酸含量的描述进行。

[0025] 步骤三:构建父母的目标区域单体型。

[0026] 基于第二、第三和第四测序数据构建母亲和父亲的目标区域单体型,即基于父母

各自的测序数据和已知的该对父母的目标区域带变异的子女(先证者)的测序数据,来构建父母各自的单体型。将父母各自的测序数据以及先证者的测序数据分别与参考序列比对,利用软件比如SOAPsnp、GATK、bowtite等识别出父母以及先证者目标区域中的SNP和获得各个SNP的基因型,由于先证者的两条单体型(两组SNP集合)是由父亲和母亲的各一条单体型组成的,所以依据孟德尔遗传规律,依据父母及先证者的各个SNP所在位点的基因型,比如利用多个区分型SNP,区分型SNP指该位点父母为不同基因型能够提供给下一代能区分单体型来源的SNP,构建父亲和母亲的单体型。单体型倾向作为一个遗传单元遗传给子代,在这里,单体型是一组SNP的集合。

[0027] 需要说明的是,本发明的实施方式对步骤二和步骤三的进行没有先后顺序限制,可以先进行步骤二再进行步骤三,或者先进行步骤三获得父母目标区域单体型再进行步骤二确定胎儿核酸含量。

[0028] 步骤四:确定胎儿目标区域单体型。

[0029] 基于母亲和父亲的目标区域单体型以及胎儿核酸含量,确定所述胎儿目标区域单体型。具体地,利用多个在父亲目标区域单体型上为杂合、在母亲目标区域单体型上为纯合的位点确定胎儿遗传到的父亲目标区域单体型,这是由于若胎儿某SNP位点为杂合的,由于源自母亲的只可能为一种类型的碱基,所以就可确定该位点的另一碱基来自父亲,利用多个这样的位点,比如可以确定超过10个这样的位点的等位基因源自父亲的一条单体型,就能确定胎儿两条单体型中的源自父亲的那条单体型。而对于胎儿另一条单体型的确定,可类似的利用多个在父亲目标区域单体型上为纯合、在母亲目标区域单体型上为杂合的位点来确定,但由于胎儿核酸样本,即母体外周血样本混有大量的母体DNA,单从以上类型SNP没法判断胎儿遗传了R还是r所在的母亲单体型,因为该位点任何的等位碱基也都可能就只是母体的,在这里我们结合胎儿核酸含量来确定胎儿遗传到的母亲的单体型。对于多个在父亲单体型上为纯合、母亲单体型上为杂合的多态性位点,这样的位点在母体外周血样本中每个都可表示为Rr,若多个这样的位点都符合R/r=(1+x%)/(1-x%),则判定胎儿遗传了母亲等位基因R所在的单体型,若多个这样的位点都符合R/r=1,则判定胎儿遗传了母亲等位基因r所在的单体型,R和r表示一对等位基因,x%表示胎儿核酸含量,R/r=比对后第一测序数据中支持R的读段数目/比对后第一测序数据中支持r的读段数目。由此,确定胎儿的单体型。

[0030] 本领域普通技术人员可以理解,上述实施方式中各种方法的全部或部分步骤可以通过程序来指令相关硬件完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:只读存储器、随机存储器、磁盘或光盘等。

[0031] 依据本发明的另一个实施方式,提供一种确定胎儿目标区域单体型的装置,该装置能够用以完成本发明一个实施方式中的方法的部分或全部步骤,如图1所示,该装置1000包括:测序单元100,用以获得孕妇体液中的游离核酸,捕获目标区域,对所述捕获得的目标区域进行序列测定,获得第一测序数据,用以捕获胎儿家系成员核酸中的同样目标区域,对所述家系成员的同样目标区域进行序列测定,获得家系成员测序数据,所述家系成员测序数据包括第二、第三和第四测序数据,分别对应胎儿母亲、胎儿父亲和先证者的同样目标区域的测序数据;胎儿核酸含量确定单元200,与所述测序单元100相连,用于基于第一和第二测序数据,或者基于第一、第二和第三测序数据,以确定所述孕妇体液样本中的胎儿核酸含

量;父母单体型确定单元300,与所述测序单元100相连,用于基于第二、第三和第四测序数据构建母亲和父亲的目标区域单体型;胎儿单体型确定单元400,与所述胎儿核酸含量确定单元200和所述父母单体型确定单元300相连,用于基于母亲和父亲的目标区域单体型以及胎儿核酸含量,确定所述胎儿目标区域单体型。对本发明的一个实施方式中的方法的技术特征和优点的描述,同样适用本发明这一实施方式的装置,在此不再赘述。

[0032] 以下结合对具体样本依据本发明的方法进行目标区域单体型的确定、基因型的确定、单体型或基因型确定后的用途进行详细的描述及结果展示。下面示例,仅用于解释本发明,而不能理解为对本发明的限制。在本发明中所使用的“第一”、“第二”、“第三”等仅用于方便描述目的,而不能理解为指示或暗示相对重要性,也不能理解为之间有先后顺序关系。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。

[0033] 除另有交待,以下实施例中涉及的未特别交待的试剂、序列(接头、标签和引物)、软件及仪器,都是常规市售产品或者公开的,比如购自Illumina公司的hiseq2000测序平台建库相关试剂盒来进行测序文库构建等。

[0034] 一般方法:

[0035] 1.目标捕获区域的选择及探针的设计

[0036] 目标捕获区域包括SMN1基因外显子区,SMN1基因内部及其上下游3M区域内高杂合率SNP位点的捕获测序。SNP的选择参考dbSNP数据库,选择其中参考染色体数大于100条、MAF在0.3-0.5之间的SNP位点。同时,为了保证检测的准确性,保证SNP位点所在序列63mer碱基序列在基因组上为唯一比对,且GC含量在40％-50％。SMN1区域捕获区域如表1及表2所示

[0037] 2.家系致病单体型的获得

[0038] 通过生物信息分析,对孕妇、孕妇丈夫及先证者在目标基因及其上下游区域的SNP位点基因型进行判断。通过对三者的SNP基因型进行连锁分析,以确定与致病突变紧密连锁的SNP位点的基因信息,并进一步获得与致病突变连锁的单体型信息。整体技术路线如图2所示。

[0039] (1)从孕妇、孕妇丈夫及先证者的外周血中抽提基因组DNA,并使用电泳及OD对获得的DNA进行质量检测。

[0040] (2)使用质量检测合格的基因组DNA进行目标区域捕获文库的制备。文库制备是将1μg基因组DNA打断成主带为200-300bp小片段DNA,然后将打断后DNA片段进行末端补平,在3’端加碱基“A”,使得DNA片段能与3’端带有“T”碱基的特殊接头连接,经Non-Captured PCR(未捕获前PCR)构建完成的文库,通过SMN1基因目标区域捕获探针选取的特定基因的Exon及侧翼±30bp区域进行富集,再通过PCR扩增富集后产物,最后通过杂交前后PCR产物QPCR检测获得序列捕获杂交效率。

[0041] (3)使用高通量测序仪对获得的样品文库进行测序。使得目标区域平均测序深度达到200×以上。

[0042] (4)通过生物信息分析,对测序信息进行分析和研究,以得到相关基因的单核苷酸变异(SNV)、少数碱基的插入和缺失(InDel)等遗传变异信息。并明确与目标待检致病突变相连锁遗传的SNP信息,即致病单体型。假设先证者分别从父母双方得到一个致病突变,若,

[0043] 1)假设先证者致病基因外某一位点的基因型为AA,父亲为AC,母亲为AA。则可知:

先证者从父亲处获得了A,从母亲处获得了一个A,且这两个SNP位点均与致病突变相连锁遗传。而在父亲中C与非致病等位基因(allele)连锁;

[0044] 2)假设先证者致病基因外某一位点的基因型为AC,父亲为AC,母亲为AA。则可知:先证者从父亲处获得了C,从母亲处获得了一个A,且这两个SNP位点均与致病突变相连锁遗传。而在父亲中C与非致病allele连锁;

[0045] 3)假设先证者致病基因外某一位点的基因型为AC,父亲为AA,母亲为AC。则可知:先证者从父亲处获得了A,从母亲处获得了一个C,且这两个SNP位点均与致病突变相连锁遗传。而在母亲中C与非致病allele连锁;

[0046] 将上述推测方法应用到SMN1基因及两侧3M区域的SNP位点,则可获得者一范围内的单体型信息,获知在这一区域内与致病突变连锁的单体型信息。从而并可进一步推断出与非致病allele紧密连锁的SNP信息。

[0047] 3.孕妇血浆DNA目标区域捕获测序

[0048] 对孕妇血浆DNA进行目标区域捕获测序,并进行生物信息学SNP/indel分析。以亲缘关系是否正确及胎儿DNA含量为质控环节,仅对质控合格的样品进行后续分析。对孕妇的血浆游离DNA测序数据进行genotyping,并结合该家系单体型进行连锁分析,判断胎儿是否遗传了夫妇的致病单体型。

[0049] (1)从1.2ml孕妇血浆中抽提细胞游离DNA,并使用Qubit定量DNA进行质量检测。

[0050] (2)使用质量检测合格的基因组DNA进行目标区域捕获文库的制备。首先对DNA片段进行末端补平,在3'端加碱基“A”,使得DNA片段能与3'端带有“T”碱基的特殊接头连接,经Non-Captured PCR构建完成的文库,通过SMN1目标区域捕获探针选取的特定基因的Exon及侧翼±30bp区域进行富集,再通过PCR扩增富集后产物,最后通过杂交前后PCR产物QPCR检测获得序列捕获杂交效率。

[0051] (3)使用高通量测序仪对获得的样品文库进行测序。使得目标区域平均测序深度达到500×以上。

[0052] 4.胎儿基因型推测

[0053] (1)通过生物信息分析,对测序信息进行分析和研究,以得到相关基因的单核苷酸变异(SNV)、少数碱基的插入和缺失(InDel)等遗传变异信息。

[0054] (2)对血浆游离DNA中胎儿DNA的含量进行计算,计算方式如下

[0055] a)假设母亲白细胞DNA基因型为AA,胎儿基因组DNA为AT,则此时血浆中可观察到的基因型为A和T,若支持A的reads数为c,支持C的reads数为d,则此时f=2d/(c+d);

[0056] b)假设母亲白细胞DNA基因型为AT,胎儿基因组DNA为AA,则此时血浆中可观察到的基因型为A和T,若支持A的reads数为c,支持T的reads数为d,则此时f=(c-d)/(c+d)。

[0057] 若胎儿DNA含量>3%则认为质控合格,进入后续实验。

[0058] (3)判断胎儿从父亲处遗传的基因型,计算方式如下:

[0059] a)选择母亲为纯合,而父亲为杂合的位点进行父亲遗传单体型的判断。假设某一SNP位点母亲基因型为AA,父亲基因型为AC,若血浆测序数据call SNP结果为A,C,且C的含量符合估计的胎儿浓度。则表明胎儿从处获得SNP C所在的allele;

[0060] b)将SMN1捕获区域内所有满足a)条件的SNP用于判断胎儿从父亲处所获得的SNP信息,构成胎儿从父亲处获得的单体型信息。并根据2-(4)中的信息,明确该单体型是否与

致病突变相连锁，从而获知胎儿是否从父亲处获得致病allele。

[0061] (4)判断胎儿从母亲处遗传的基因型，计算方式如下

[0062] 选择母亲为杂合，而父亲为纯合的位点进行母亲遗传单体型的判断。假设某一SNP位点母亲基因型为AC，父亲基因型为AA，若血浆测序数据call SNP结果为A和C，若胎儿从母亲处遗传了A等位基因，胎儿的基因型为AA，则可观察到A/C近似与(1+f)/(1-f)；若胎儿遗传了C等位基因，胎儿的基因型为AC，则可观察到A/C近似为0.5。对等位基因的reads支持数构建二项分布模型分别计算出遗传A、C的概率后得到相对概率Pa、Pc(Pa+Pc＝1)并将所有SNP各点概率构建HMM模型用Viterbi算法(Lawrence R.Rabiner,PROCEEDINGS OF THE IEEE,Vol.77,No.2,1989年2月)判断胎儿从母亲处获得的单体型信息，并根据单体型是否与致病突变相连锁，得知胎儿是否从母亲处获得致病allele。

[0063] (5)综合(3)和(4)的结果，获得胎儿的基因型信息。

[0064] 实施例

[0065] 对1例具有生育SMN1患病二胎高风险的孕妇(天津妇幼保健院)进行无创产前基因检测。孕妇及其丈夫均为为SMN1基因7号外显子缺失突变的杂合携带者，并生育过一个SMN1纯合突变的患者。现第二次怀孕，抽取孕妇外周血并及时分离血浆，而后通过血浆DNA及孕妇、孕妇丈夫、先证者的基因组DNA进行捕获测序，对本胎胎儿的基因情况进行分析。

[0066] 用盐析法提取标本DNA，大片段DNA进行超声打断，目前使用样品打断方法为Covaris打断法，将样品DNA打碎至100-700bp范围的片段。(注：打断效果一般以所要求制备文库Insert片段主带位置在200-250bp位置较为理想，若打断效果不理想则需要进行重新打断。)

[0067] 用盐析法提取血浆游离DNA，使用Qubit定量后直接进行文库构建。

[0068] 1.文库制备

[0069] 1.1 末端修复和纯化

[0070]

试剂名称	体积（μ L）
10X Polynucleotide Kinase Buffer(B904)	10
dNTP Solution Set	4
T4 DNA Polymerase	5
Klenow Fragment	1
T4 Polynucleotide Kinase(T4 PNK)	5

[0071] 将配置好的mix震荡混匀后，每个反应加入25μL酶反应混合液。

[0072] 反应条件：20℃，30min

[0073] 使用180μL Ampure Beads进行产物纯化，回收的DNA溶于30μL(其中1.9μL为损耗)的水中。

[0074] 1.2 末端加“A”(A-Tailing)

[0075]

试剂名称	体积（μ L）
10X Blue buffer	3.5
dATP(5mM)	1.4
Klenow(3’-5’ exo-)	2

[0076] 将配置好的mix震荡混匀后，每管加入6.9μL酶反应混合液。

[0077] 反应条件：20℃，30min

[0078] 注:末端加“A”后不纯化

[0079] 1.3 Adapter的连接和纯化

[0080]

试剂名称	体积（μ L）
10X Ligation buffer	1.5
Index PE Adapter(40μ M)	1
ATP(10mM)	3.5
T4 DNA Ligase	3
ddH_2O	6

[0081] 将配置好的mix震荡混匀,每个反应加入15μL酶反应混合液。

[0082] 反应条件:16℃,12-16h(过夜)

[0083] 使用75μL Ampure Beads进行产物纯化,回收的DNA溶于35μL(其中2μL为损耗)的水中。

[0084] 1.4Non-Captured样品Pre-LM-PCR

[0085]

试剂名称	体积（μ L）
Index P1(10μ M公用引物)	8
10XPfx Amplification Buffer	10
dNTP (10mM)	4
$MgSO_4$ (50mM)	4
PCR Index primer2.0 (10 pmol/μ L)	4
ddH2O	34

[0086] PCR程序：

[0087]

94℃ 2min;

94℃ 15s, 62℃ 30s，72℃ 30s，4cycles;

72℃ 5min;

4℃ forever

[0088] 2.芯片杂交,目标区域捕获富集

[0089] 本实验中参照NimbleGen使用说明书进行杂交洗脱,获取目的基因并PCR富集。

[0090] 3.上机测序

[0091] 本实验采用hiseq2000或hiseq2500PE101+8+101程序进行上机测序。

[0092] 4.信息分析

[0093] 测序仪获取原始短序列;

[0094] 去除测序数据中的接头和低质量数据;

[0095] 短序列定位到人类基因组数据相应的位置上;

[0096] 统计测序结果信息,短序列数量、目标区域覆盖大小、平均测序深度等;

[0097] 过滤低质量值和低覆盖度的单核苷酸;

[0098] 注释,确定突变位点发生的基因、坐标、氨基酸改变等;

[0099] 确定SMN1捕获区域内各SNP的基因型。

[0100] 5.结果分析

[0101] 1)数据产出情况

[0102] 如表3所示,所测样品在目标区域平均测序深度均在100X以上,血浆测序深度达到

271x。

[0103] 表3 数据产出情况表

[0104]

原始样品名	数据量（Data production，Gb)	捕获特异性（Capture specificity，%)	目标区域平均测序深度（mean depth of target region）	目标区域深度>=20X所占比例(Fraction of target covered >=20X，%)
孕妇	2.69	0.33	146.68	0.79
孕妇丈夫	2.01	0.38	99.93	0.76

[0105]

先证者	2.06	0.43	114.49	0.77
孕妇血浆	34.33	0.45	103.84	0.83

[0106] 2)SNP phasing情况

[0107] 我们使用父亲、母亲及先证者在SMN1基因上下游1M以内的SNP位点进行先证者单体型构建。表4统计了该区域成功判断所属单体型的SNP的数目(phased SNP)。这些phased SNP后续用于父亲遗传单体型判断(SNP used for Pat-Hap)及用于母亲遗传单体型判断(SNP used for Mat-Hap)

[0108] 表4 SMN1基因相关区域phase SNP情况统计

[0109]

Region	SMA1	
	SNP calling	Phased SNP
SMN2_upstream 500kb~1M	\	\
SMN2_upstream 200~500kb	\	\
SMN2_upstream 200kb	\	\
SMN2_gene	\	\
SMN2_downstream 200kb	1	\
SMN2_downstream 200~500kb	\	\
SMN2_SMN1_overlap150kb 200~500kb	\	\
SMN1_upstream 200~500kb	2	\
SMN1_upstream 200kb	\	\
SMN1_gene	\	\
SMN1_downstream 200kb	2	\
SMN1_downstream 200~500kb	13	8
SMN1_downstream 500kb~1M	201	137
Total	219	145

[0110] 3)血浆中胎儿DNA含量分析

[0111] 选择父亲为杂合而母亲为纯合的点，对血浆中胎儿DNA含量进行估计：假设母亲基因型为AA，胎儿基因型为AT，若测得为A的reads数为a，为C的reads数为b，则血浆中胎儿DNA含量c＝2b/(a+b)。结果显示该血浆样品中胎儿DNA含量为0.0930。

[0112] 4)胎儿基因型判断

[0113] 对SMA 1家系中孕妇外周血浆数据进行分析，利用HMM算法推测本次怀孕胎儿SMN1基因情况，具体地，将胎儿的单体型Hap0和Hap1作为隐含状态(hidden states)，将成功判断所属单体型的SNP作为观测序列(observations)，根据snp位点的位置和遗传图谱计算的

相邻snp间的重组概率推算出状态转移概率(transition probabilities),根据reads支持数计算每个snp位点支持Hap0、Hap1的相对概率(Emission_probability),然后通过惠特比算法(Viterbi algorithm)可以推断出SNP总体支持的单体型排列,获得最可能的胎儿单体型组合。可参考Chen S1,Ge H2,Wang X,et al.Haplotype-assisted accurate non-invasive fetal whole genome recovery through maternal plasma sequencing.Genome Med.2013,5(2):18进行。

[0114] 为了避免重复序列区域对分析结果的影响,仅使用unique序列区域进行分析。结果如图3所示,图上的每个点表示一个snp位点遗传自父/母Hap0的概率与遗传自父/母Hap1的概率的差值,每个小圈是一个组合判断结果,小圈形成的线在中间基线上游表示最终判断遗传自Hap0,小圈形成的线在中间基线下面表示最终判断遗传自Hap 1。从图3可看出,Pat-Hap 0及Mat-Hap 0分别父母双方带有致病突变的单体型,Pat-Hap 1及Mat-Hap 1分别父母双方不携带致病突变的单体型。推断结果显示胎儿从其父母处获得了Pat-Hap1及Mat-Hap1,即不携带SMN1致病突变的染色体。表明胎儿不存在SMN1缺失。

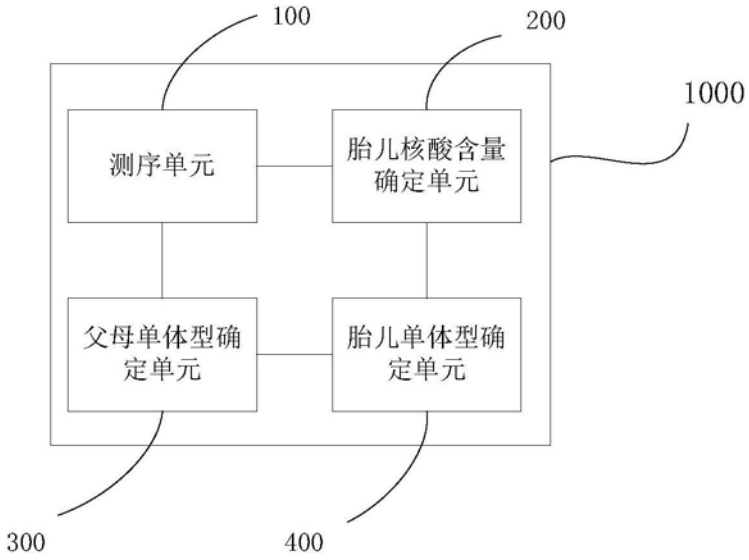

图1

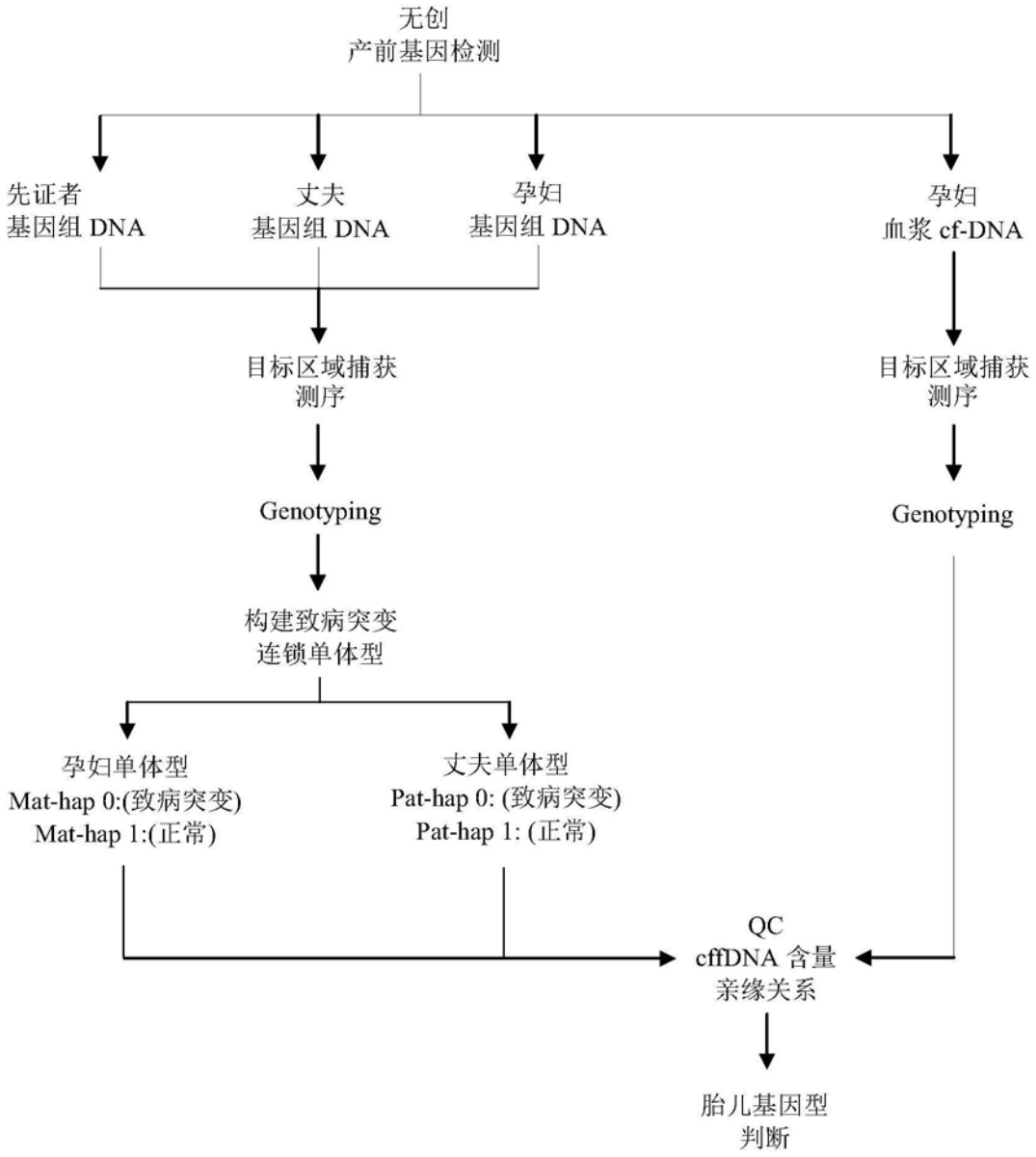
无创
产前基因检测
先证者
基因组 DNA
丈夫
基因组 DNA
孕妇
基因组 DNA
孕妇
血浆 cf-DNA
目标区域捕获
测序
目标区域捕获
测序
Genotyping
Genotyping
构建致病突变
连锁单体型
孕妇单体型
Mat-hap 0:(致病突变)
Mat-hap 1:(正常)
丈夫单体型
Pat-hap 0: (致病突变)
Pat-hap 1: (正常)
QC
cffDNA 含量
亲缘关系
胎儿基因型
判断

图2

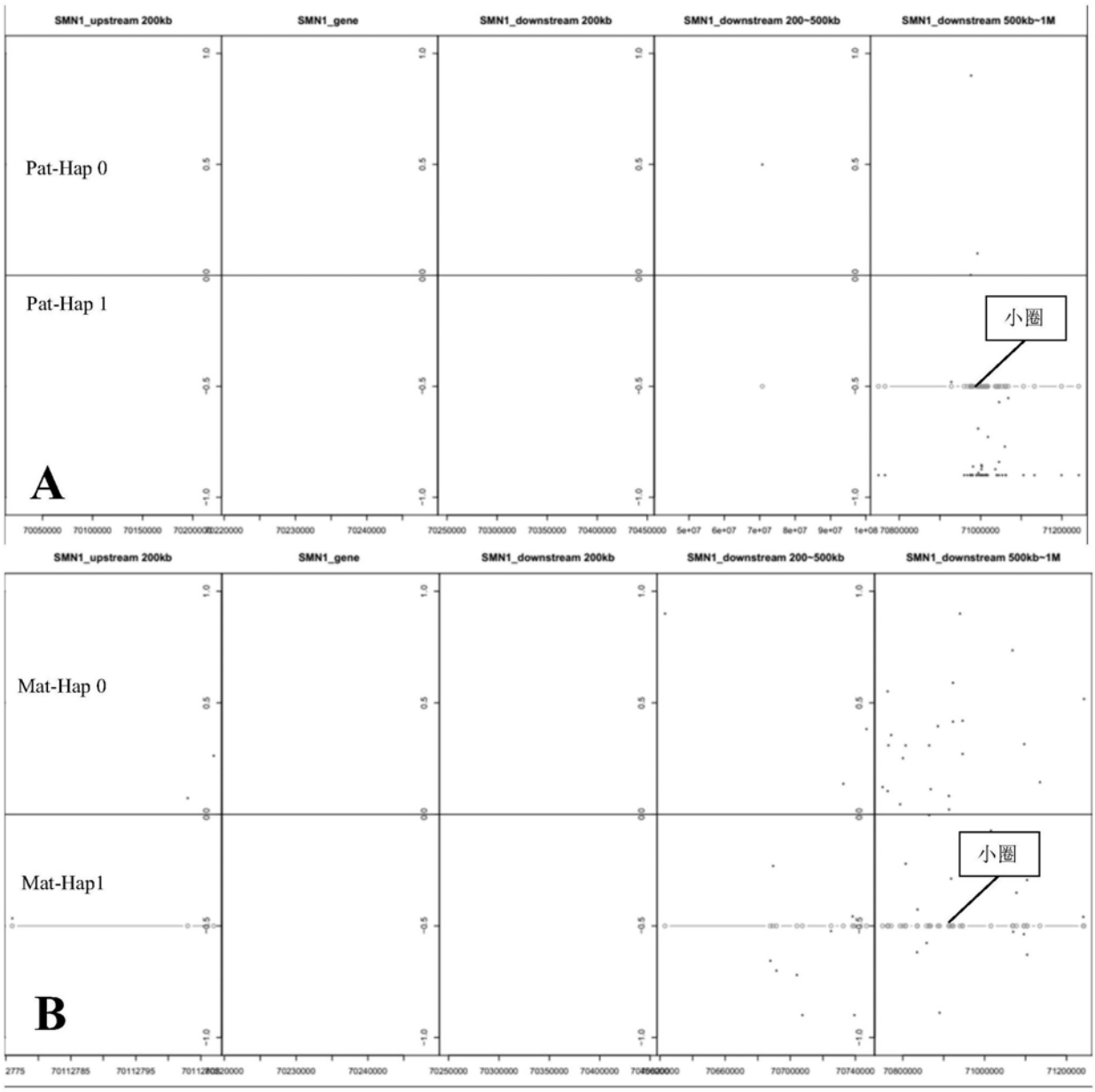
SMN1_upstream 200kb
SMN1_gene
SMN1_downstream 200kb
SMN1_downstream 200~500kb
SMN1_downstream 500kb~1M
Pat-Hap 0
Pat-Hap 1
小圈
A
Mat-Hap 0
Mat-Hap1
小圈
B

图3