(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 107849612 B
(45)授权公告日 2023.04.14

(21)申请号 201680030228.2
(22)申请日 2016.03.25
(65)同一申请的已公布的文献号
申请公布号 CN 107849612 A
(43)申请公布日 2018.03.27
(30)优先权数据
62/138,620 2015.03.26 US
62/253,908 2015.11.11 US
(85)PCT国际申请进入国家阶段日
2017.11.24
(86)PCT国际申请的申请数据
PCT/US2016/024319 2016.03.25
(87)PCT国际申请的公布数据
WO2016/154584 EN 2016.09.29

(73)专利权人 奎斯特诊断投资股份有限公司
地址 美国特拉华州
(72)发明人 C·埃尔津加
(74)专利代理机构 北京坤瑞律师事务所 11494
专利代理师 封新琴

(51)Int.Cl.
C12Q 1/6886(2018.01)
C12Q 1/6869(2018.01)
G16B 30/10(2019.01)
G06N 3/12(2023.01)
G06N 5/04(2023.01)

(56)对比文件
CN 104232761 A,2014.12.24
CN 103874767 A,2014.06.18

审查员 陈小燕

权利要求书3页 说明书25页 附图6页

(54)发明名称

比对和变体测序分析管线

(57)摘要

提供了用于分析来自下一代序列(NGS)平台的基因序列数据的系统和方法。还提供了制备通过NGS进行核酸序列分析的样品的方法。使用修改的GATK变体判读器进行变体判读。将读段映射到基因组参考序列是用Burrows Wheeler Aligner(BWA)进行的,且不包括软剪切。基因组参考序列是GRCh37.1人类基因组参考序列。测序方法包括乳液PCR(emPCR)、滚环扩增或固相扩增。在一些实施方案中,固相扩增是克隆桥扩增。

CN 107849612 B

1.一种通过一个或多个电子处理器处理原始测序数据的方法,其中所述方法包括:

(a) 从核酸测序仪获得原始测序数据,其中所述原始测序数据是从核酸样品产生的,其中使用一个或多个生物素化的RNA诱饵对所述核酸样品中的一个或多个感兴趣的基因进行外显子捕获;

(b) 从原始测序数据中去除未通过质量过滤器的低质量读段;

(c) 从经过滤的原始测序数据中修剪掉衔接子和分子鉴别序列;

(d) 将经过滤的原始测序数据映射到基因组参考序列以生成映射的读段;

(e) 对映射的读段进行排序和索引;

(f) 将所述读段添加到数据文件以生成经处理的序列文件;

(g) 创建重新比对目标;

(h) 进行经处理的序列文件的局部重新比对以生成重新比对的序列文件;

(i) 从重新比对的序列文件中去除重复读段;

(j) 分析感兴趣的编码区;和

(k) 基于步骤(j)中的分析,生成鉴定变体是否存在的报告,

其中使用改良的Smith-Waterman局部重新比对工具来进行步骤(g)和(h),所述改良的Smith-Waterman局部重新比对工具限于对所述一个或多个基因内的感兴趣核酸区域进行重新比对,其中所述感兴趣核酸区域由所述一个或多个基因的编码外显子+/-50个碱基组成,所述基因选自*BRCA1*或*BRCA2*,其中所述局部重新比对不包括碱基质量分数重新校准,其中使用改良的基因组分析工具包变体判读器进行变体判读,其中将读段映射到基因组参考序列是用Burrows Wheeler Aligner进行的,

其中所述通过一个或多个电子处理器处理原始测序数据的方法用于非诊断目的。

2.根据权利要求1所述的方法,其中所述分析感兴趣的编码区包括判读感兴趣区域中的每个位置处的变体。

3.根据权利要求1所述的方法,所述将读段映射到基因组参考序列不包括软剪切。

4.根据权利要求1所述的方法,其中所述基因组参考序列是GRCh37.1人类基因组参考序列。

5.根据权利要求1所述的方法,其中从核酸测序仪获得原始测序数据的方法包括使用乳液PCR、滚环扩增或固相扩增的步骤。

6.根据权利要求5所述的方法,其中所述固相扩增是克隆桥扩增。

7.根据权利要求1所述的方法,其中所述核酸是从生物样品中提取的。

8.根据权利要求7所述的方法,其中所述生物样品是流体或组织样品。

9.根据权利要求7所述的方法,其中所述生物样品是血液样品。

10.根据权利要求1所述的方法,其中所述核酸是基因组DNA。

11.根据权利要求1所述的方法,其中所述核酸是从mRNA反转录的cDNA。

12.根据权利要求1所述的方法,所述方法包括在测序前由核酸测序仪制备核酸,其中通过实施以下步骤制备所述核酸:

(a) 剪切核酸样品;

(b) 浓缩核酸样品;

(c)对剪切的核酸样品中的核酸分子进行大小选择;

(d) 使用DNA聚合酶修复样品中核酸分子的末端；

(e) 使核酸分子附接一个或多个衔接子序列；

(f) 扩增核酸以增加具有附接的衔接子序列的核酸的比例；

(g) 富集核酸样品中一个或多个感兴趣的基因；和

(h)在即将测序之前定量核酸样品。

13.根据权利要求12所述的方法，其中所述一个或多个衔接子序列包括用于引发测序反应和/或核酸扩增反应的核酸序列。

14.根据权利要求12所述的方法，其中所述一个或多个衔接子序列包含分子鉴别标签。

15.根据权利要求1-14中任一项所述的方法，其中确定*BRCA1*基因或*BRCA2*基因中的一种或多种变体。

16.根据权利要求1-14中任一项所述的方法，还包括通过Sanger测序证实所述变体的存在。

17.根据权利要求16所述的方法，其中分析感兴趣的编码区包括判读感兴趣区域中的每个位置的变体。

18.根据权利要求16所述的方法，其中使用改良的基因组分析工具包变体判读器来进行变体判读。

19.根据权利要求16所述的方法，其中将读段映射到基因组参考序列是用Burrows Wheeler Aligner进行的。

20.根据权利要求16所述的方法，其中将读段映射到基因组参考序列不包括软剪切。

21.根据权利要求16所述的方法，其中所述基因组参考序列是GRCh37.1人类基因组参考序列。

22.一种具有存储的指令的非瞬态计算机可读介质，所述指令包括：

(a) 从核酸测序仪获得原始测序数据的指令，其中原始测序数据是由核酸样品产生的，其中使用一个或多个生物素化的RNA诱饵对所述核酸样品中的一个或多个感兴趣的基因进行外显子捕获；

(b) 从所述原始测序数据去除未通过质量过滤器的低质量读段的指令；

(c) 从经过滤的原始测序数据中修剪掉衔接子和分子鉴别序列的指令；

(d) 将经过滤的原始测序数据映射到基因组参考序列以生成映射的读段的指令；

(e) 对映射的读段进行排序和索引的指令；

(f) 将所述读段添加到数据文件以生成经处理的序列文件的指令；

(g) 创建重新比对目标的指令；

(h)对经处理的序列文件进行局部重新比对以生成重新比对的序列文件的指令；

(i) 从重新比对的序列文件中去除重复读段的指令；和

(j) 分析感兴趣的编码区的指令；

其中使用改良的Smith-Waterman局部重新比对工具来进行步骤(g)和(h)的指令，所述改良的Smith-Waterman局部重新比对工具限于对一个或多个基因内的感兴趣核酸区域进行重新比对，其中所述感兴趣核酸区域由所述一个或多个基因的编码外显子+/-50个碱基组成，所述基因选自*BRCA1*或*BRCA2*，其中所述局部重新比对不包括碱基质量分数重新校准，其中使用改良的基因组分析工具包变体判读器进行变体判读，其中将读段映射到基因组参

考序列是用Burrows Wheeler Aligner进行的。

23.根据权利要求22所述的非瞬态计算机可读介质,其中分析感兴趣的编码区包括判读感兴趣区域中的每个位置的变体。

24.根据权利要求22所述的非瞬态计算机可读介质,其中将读段映射到基因组参考序列不包括软剪切。

25.根据权利要求22所述的非瞬态计算机可读介质,其中所述基因组参考序列是GRCh37.1人类基因组参考序列。

26.一种系统,其包括:

一个或多个电子处理器,其被配置为:

(a) 从核酸测序仪获得原始测序数据,其中原始测序数据是由核酸样品产生的,其中使用一个或多个生物素化的RNA诱饵对所述核酸样品中的一个或多个感兴趣的基因进行外显子捕获;

(b) 从原始测序数据中去除未通过质量过滤器的低质量读段;

(c) 从经过滤的原始测序数据中修剪掉衔接子和分子鉴别序列;

(d) 将经过滤的原始测序数据映射到基因组参考序列以生成映射的读段;

(e) 对映射的读段进行排序和索引;

(f) 将所述读段添加到数据文件以生成经处理的序列文件;

(g) 创建重新比对目标;

(h) 对经处理的序列文件进行局部重新比对以生成重新比对的序列文件;

(i) 从重新比对的序列文件中去除重复读段;和

(j) 分析感兴趣的编码区;

其中使用改良的Smith-Waterman局部重新比对工具来进行步骤(g)和(h),所述改良的Smith-Waterman局部重新比对工具限于对一个或多个基因内的感兴趣核酸区域进行重新比对,其中所述感兴趣核酸区域由所述一个或多个基因的编码外显子+/-50个碱基组成,所述基因选自*BRCA1*或*BRCA2*,其中所述局部重新比对不包括碱基质量分数重新校准,其中使用改良的基因组分析工具包变体判读器进行变体判读,其中将读段映射到基因组参考序列是用Burrows Wheeler Aligner进行的。

比对和变体测序分析管线

[0001] 对相关申请的交叉引用

[0002] 本申请要求2015年3月26日提交的美国临时申请No.62/138620和2015年11月11日提交的美国临时申请No.62/253908的优先权和权益,其内容各自通过提述完整并入本文。

[0003] 背景

[0004] 在美国,每年诊断出超过200000例乳腺癌新病例。其中约2％至5％与BRCA1或BRCA2基因中的功能丧失变体相关。在一般人群中估计的携带者频率BRCA1为1:300,BRCA2为1:800,例外是阿什肯纳兹犹太人(Ashkenazi-Jewish)女性,在她们中BRCA1和BRCA2中的3种始祖突变(founder mutation)的携带者频率为2％至5％。带有BRCA1或BRCA2基因中的有害突变的患者发生乳腺癌的终生风险有50％至80％,发生卵巢癌的终生风险有20％至40％。三阴性乳腺癌——不表达雌激素受体、孕激素受体或Her2/neu,特点是侵袭性更强——占所有乳腺癌的15％至20％;三阴性乳腺癌与BRCA突变相关,频率在4％至42％之间,取决于研究人群的特征(例如,阿什肯纳兹犹太人女性的比例)。

[0005] 美国国家综合癌症网络(NCCN)制定了帮助医疗保健提供者鉴定具有乳腺癌和卵巢癌高风险、并可能受益于癌症遗传风险评估的患者和家族成员的指南。遗传风险评估可能包括基因检测,但它是一个动态的咨询过程。确定乳腺癌女性是否为BRCA1/2阳性可有助于提供适当的关于增加监视以及关于对侧乳房切除术和/或输卵管卵巢切除术的风险与收益的咨询,而这两种手术都已被证明是对抗乳腺癌的防护措施。鉴定患者中的有害BRCA1/2变体也可能对家庭成员有帮助,家庭成员可能需要获取遗传咨询和测试的渠道来评估他们的癌症风险并确定适当的管理。美国乳腺外科医生学会建议对高风险人群个体进行BRCA1/2检测,包括具有以下情况的人群:早发型乳腺癌患者(50岁以前诊断);两处原发性乳腺癌,双侧或同侧;早发型乳腺癌家族史;男性乳腺癌;卵巢癌(特别是非粘液型)的个人或家族史;在新诊断的乳腺癌或乳腺癌家族史背景中的阿什肯纳兹(东欧)犹太血统;家族中以前鉴定出BRCA1或BRCA2突变;年龄≤60岁的三阴性乳腺癌;或与遗传性乳腺癌家族史和卵巢相关癌症相关联的胰腺癌。

[0006] 综合性的BRCA检测通常包括对BRCA1和BRCA2的所有编码外显子及剪接接合区进行测序,和对大基因重排进行分析。当扩增或测序引物序列中存在多态性时,基于PCR的测序方法,包括使用PCR扩增的Sanger测序和下一代测序(NGS)系统,可能由于等位基因脱扣(drop-out)而产生假阴性结果。

[0007] 因此,对于改进样品测序的方法以及准确高效地分析NGS测序数据的方法存在着需要。

[0008] 概述

[0009] 在本文某些实施方案中提供了处理由高通量测序方法(包括下一代测序平台)产生的测序数据的方法。示例性测序平台包括但不限于Illumina MiSeq系统和Life Technologies Ion Torrent个人化操作基因组测序仪。

[0010] 在本文某些实施方案中提供了确定受试者中一个或多个基因中的变体的存在的方法,所述方法包括:(a)使用核酸测序仪提供从来自受试者的核酸样品的核酸测序反应产

生的原始测序数据;(b)从所述原始测序数据中去除未通过质量过滤器的低质量读段;(c)从经过滤的原始测序数据中修剪衔接子和/或分子鉴别(MID)序列;(d)将经过滤的原始测序数据映射到基因组参考序列以生成经映射的读段;(e)对经映射的读段进行分类和索引;(f)将读段组添加到数据文件以生成经处理的序列文件;(g)创建重新比对(realigner)目标;(h)对经处理的序列文件进行局部重新比对以生成重新比对的序列文件;(i)从重新比对的序列文件中去除重复读段;(j)分析感兴趣的编码区;和(k)基于步骤(j)中的分析,生成鉴定变体是否存在的报告,其中使用限于含有该一个或多个感兴趣基因的核酸区域的改良的基因组比对工具(utility)进行步骤(g)和(j)。在一些实施方案中,该方法包括使用核酸测序仪对来自受试者的核酸样品进行核酸测序反应以产生步骤(a)的原始测序数据。在一些实施方案中,分析感兴趣的编码区包括判读(call)感兴趣的区域中的每个位置上的变体。在一些实施方案中,感兴趣的区域被额外的150个碱基填补(pad)。在一些实施方案中,使用改良的GATK变体判读器来进行变体判读。在一些实施方案中,将读段映射到基因组参考序列是用Burrows Wheeler Aligner(BWA)进行的。在一些实施方案中,将读段映射到基因组参考序列不包括软剪切(soft clipping)。在一些实施方案中,基因组参考序列是GRCh37.1人类基因组参考序列。在一些实施方案中,测序方法包括乳液PCR(emPCR)、滚环扩增或固相扩增。在一些实施方案中,固相扩增是克隆桥(clonal bridge)扩增。

[0011] 在一些实施方案中,用于序列分析的核酸从来自受试者的生物样品中提取。在一些实施方案中,生物样品是流体或组织样品。在一些实施方案中,生物样品是血液样品。在一些实施方案中,核酸是基因组DNA。在一些实施方案中,核酸是从mRNA反转录的cDNA。

[0012] 在一些实施方案中,其中核酸样品在测序前通过进行以下一种或多种方法制备:(a)剪切(shear)核酸;(b)浓缩核酸样品;(c)对经剪切的核酸样品中的核酸分子进行大小选择;(d)使用DNA聚合酶修复样品中核酸分子的末端;(e)附接一个或多个衔接子序列;(f)扩增核酸以增加具有附接的衔接子序列的核酸的比例;(g)富集核酸样品中的一个或多个感兴趣的基因;和/或(h)在即将测序之前定量核酸样品引物。在一些实施方案中,所述一个或多个衔接子序列包含用于引发测序反应和/或核酸扩增反应的核酸序列。在一些实施方案中,一个或多个衔接子序列包含分子鉴别(MID)标签。在一些实施方案中,富集核酸样品中的一个或多个感兴趣的基因包括使用一个或多个生物素化RNA诱饵的外显子捕获。在一些实施方案中,生物素化的RNA诱饵对于外显子区域、剪接接合位点或内含子区域或一个或多个感兴趣的基因是特异性的。在一些实施方案中,所述一个或多个生物素化RNA诱饵对于BRCA1基因和/或BRCA2基因是特异性的。

[0013] 在一些实施方案中,用于序列分析的核酸从哺乳动物受试者获得。在一些实施方案中,受试者是人类患者。在一些实施方案中,受试者是疑似患有癌症或疑似具有发生癌症的风险的人。在一些实施方案中,癌症是乳腺癌或卵巢癌。

[0014] 在一些实施方案中,通过本文中提供的方法确定与癌症相关的基因中的一个或多个变体。在一些实施方案中,确定了BRCA1基因或BRCA2基因中的一个或多个变体。在一些实施方案中,一个或多个变体选自表1中列出的变体。

[0015] 在一些实施方案中,所提供的方法进一步包括通过测序确认所述一个或多个变体的存在。

[0016] 本文还在某些实施方案中提供了包括一个或多个电子处理器的系统,所述电子处

理器配置为:(a)从原始测序数据中去除未通过质量过滤器的低质量读段;(b)从过滤的原始测序数据中修剪衔接子和/或分子鉴别(MID)序列;(c)将经过滤的原始测序数据映射到基因组参考序列以生成经映射的读段;(d)对映射的读段进行分类和索引;(e)将读段组添加到数据文件以生成经处理的序列文件;(f)创建重新比对目标;(g)对经处理的序列文件进行局部重新比对以生成重新比对的序列文件;(h)从重新比对的序列文件中去除重复读段;和(i)分析感兴趣的编码区。在一些实施方案中,分析感兴趣的编码区包括判读在感兴趣区域中的每个位置上的变体。在一些实施方案中,感兴趣的区域被另外150个碱基填补。在一些实施方案中,使用改良的GATK变体判读器来进行变体判读。在一些实施方案中,将读段映射到基因组参考序列是用Burrows Wheeler Aligner(BWA)进行的。在一些实施方案中,将读段映射到基因组参考序列不包括软剪切。在一些实施方案中,基因组参考序列是GRCh37.1人类基因组参考序列。

[0017] 本文还在某些实施方案中提供了具有存储在其上的指令的非瞬态计算机可读介质,所述指令包括:(a)去除未通过质量过滤器的低质量读段的指令;(b)从经过滤的原始测序数据中修剪衔接子和MID序列的指令;(c)将经过滤的原始测序数据映射到基因组参考序列以生成经映射的读段的指令;(d)对经映射的读段进行分类和索引的指令;(e)将读段组添加到数据文件以生成经处理的序列文件的指令;(f)创建重新比对目标的指令;(g)进行经处理的序列文件的局部重新比对以生成重新比对的序列文件的指令;(h)从重新比对的序列文件中去除重复读段的指令;和(i)分析感兴趣的编码区的指令。在一些实施方案中,分析感兴趣的编码区包括判读感兴趣区域中的每个位置上的变体。在一些实施方案中,感兴趣的区域被另外150个碱基填补。在一些实施方案中,使用改良的GATK变体判读器来进行变体判读。在一些实施方案中,将读段映射到基因组参考序列是用Burrows Wheeler Aligner(BWA)进行的。在一些实施方案中,将读段映射到基因组参考序列不包括软剪切。在一些实施方案中,基因组参考序列是GRCh37.1人类基因组参考序列。

[0018] 前述概述仅仅是说明性的,并且并不意图以任何方式进行限制。除了以上描述的说明性方面、实施方式以及特征之外,通过参考附图以和详细说明,其他方面、实施方式及特征是不言自明的。

[0019] 附图简述

[0020] 从以下的说明和随附的权利要求书可以更加明了前文所述的本公开的特征,以及本公开的其他特征。这些附图仅描绘了根据本公开的几个实施方式,因此不应被认为对本公开的范围具有限制性,在此理解的基础上,下面通过附图来更具体、更详细地说明本公开。

[0021] 图1展示的是:按照使用MiSeq和Personal Gene Machine,PGM平台的各种说明性实现方式,用于检测BRCA1和BRCA2变体的NGS测定法的示例性总体概览。对于MiSeq平台,首先使用供应商提供的MiSeq Reporter软件,然后使用Quest序列分析管线[Quest Sequencing Analysis Pipeline(QSAP)]进行变体判读。对于PGM平台,使用供应商提供的Torrent Suite变体判读软件。

[0022] 图2是按照各种说明性实施方式的QSAP生物信息学序列分析方法的流程图。

[0023] 图3显示了验证样品(validation sample)中BRCA1的40-bp缺失(缺失c.1175_1214del40)的示例性比对。综合基因组学查看器(Integrative Genomics Viewer,IGV)图

形报告显示,使用PGM平台结合Torrent Suite变体判读检测到突变(小图A),而使用MiSeq平台结合MiSeq Reporter检测不到突变(小图B)。使用QSAP结合MiSeq平台能够检测到该缺失(小图C)。

[0024] 图4显示了验证样品中64bp缺失(41246533-41246596del;c.952_1015del)的示例性比对。综合基因组学查看器(IGV)图形报告显示使用MiSeq平台结合QSAP检测到该缺失(小图A),而使用PGM平台结合Torrent Suite变体判读未检测到该缺失(小图B)。

[0025] 图5显示了示例性地使用下一代测序(NGS)确定顺式-反式取向。这份综合基因组学查看器(IGV)图形报告来自一名在单一DNA分子上具有2个相邻变体的患者,用NGS在MiSeq/QSAP平台上可视化。顺式方向清晰可见,因为每条链或是都含有两个突变,或是两个突变都不含。

[0026] 图6是可以结合本文提供的方法使用的计算系统的图解。

[0027] 下面的详细说明从始至终参考附图来进行。在图中,相似的符号通常标识相似的部件,除非上下文另外规定。详细说明、附图、以及权利要求书中所描述的说明性实施方式并不意图构成限制。可以利用其他实施方式,并且可以做出其他改变,而不脱离此处呈现的主题的精神或范围。容易理解的是,本公开的多个方面,如本文中一般性描述并且在附图中展示的,可以以各种各样的不同配置进行排列、取代、组合和设计,所有这些情况都被明确设想到并构成本公开的一部分。

[0028] 发明详述

[0029] 某些术语

[0030] 本说明书中采用的某些术语具有下面所定义的含义。未予定义的术语具有其领域公认的含义。也就是说,除非另外定义,本文中所用的所有技术术语和科学术语具有与本发明所属领域的普通技术人员的通常理解相同的含义。

[0031] 如本文中使用的,除非另有说明,当提及数值时,术语“约”表示所列举的值加上10%或减去10%。

[0032] 如本文中使用的,术语“分离的”、“纯化的”或“基本上纯化的”是指诸如核酸之类的分子,所述分子从它们的自然环境中移除、分离或分开,并且至少60%游离于,优选75%游离于,最优选90%游离于与它们天然结合的其他组分。因此,分离的分子是基本上纯化的分子。

[0033] 在基因片段或染色体片段的上下文中,“片段”是指具有至少约10个核苷酸、至少约20个核苷酸、至少约25个核苷酸、至少约30个核苷酸、至少约40个核苷酸、至少约50个核苷酸、至少约100个核苷酸、至少约250个核苷酸、至少约500个核苷酸、至少约1,000个核苷酸、至少约2,000个核苷酸的核苷酸残基序列。

[0034] 术语“同一性”和“相同”是指序列之间的一致性程度。可能存在部分同一性或完全同一性。部分相同的序列是与另一个序列小于100%相同的序列。部分相同的序列可以具有至少70%或至少75%、至少80%或至少85%、或至少90%或至少95%的总体同一性。

[0035] 如本文中使用的术语“扩增”(“amplification)或“扩增”(“amplify”)包括用于拷贝靶核酸从而增加选择的核酸序列的拷贝数的方法。扩增可以是指数或线性的。靶核酸可以是DNA或RNA。以这种方式扩增的序列形成“扩增产物”,也称为“扩增子”。虽然下文描述的示例性方法涉及使用聚合酶链式反应(PCR)的扩增,但是用于扩增核酸的许多其他方法(例

如等温法、滚环法等)是本领域中已知的。本领域技术人员将理解,这些其他方法可以代替PCR方法或与PCR方法一起使用。参见,例如,Saiki,“Amplification of Genomic DNA”in PCR Protocols,Innis等人编辑,Academic Press,San Diego,CA 1990,pp.13-20;Wharam等人,Nucleic Acids Res.,29(11):E54-E54,2001;Hafner等人,Biotechniques,30(4):852-56,858,860,2001;Zhong等人,Biotechniques,30(4):852-6,858,860,2001。

[0036] 如本文中使用的术语“可检测标记”是指与探针相关联的分子或化合物或一组分子或一组化合物,其用于识别杂交于基因组核酸或参考核酸的探针。

[0037] 如本文中使用的,术语“检测”是指观察来自可检测标记的信号以指示靶标的存在。更具体地说,“检测”用于“检测特定序列”的上下文中。

[0038] 如本文中使用的,术语“高通量、大规模并行测序”是指可以并行产生克隆方式扩增的(clonally amplified)分子的多个测序反应及单个核酸分子的多个测序反应的测序方法。这样可以提高数据的吞吐量和产率。这些方法在本领域中也称为下一代测序(NGS)方法。NGS方法包括例如使用可逆染料终止物的边合成边测序(sequencing-by-synthesis),以及边连接边测序(sequencing-by-ligation)。常使用的NGS平台的非限制性实例包括miRNA BeadArray(Illumina,Inc.)、Roche 454TM GS FLXTM-Titanium(Roche Diagnostics)、ABI SOLiDTM System(Applied Biosystems,Foster City,CA)、和HeliScope™Sequencing System(Helicos Biosciences Corp.,Cambridge MA)。

[0039] 如本文中使用的“测序深度”或“读取深度”是指序列已被测序的次数(测序的深度)。作为示例,读取深度可以通过比对多次测序运行结果,并计算具有一定大小(例如,100bp)的非重叠窗口中的开始位置的读段加以确定。可以使用本领域已知的方法,基于读取深度来确定拷贝数变化。例如,使用在Yoon等人,Genome Research 2009September;19(9):1586-1592;Xie等人,BMC Bioinformatics 2009Mar 6;10:80;或Medvedev等人,Nature Methods2009Nov;6(11Suppl):S13-20中描述的方法。这种类型的方法和分析的使用被称为“读段深度途径”。

[0040] 如本文中使用的“核酸片段读段”是指单一的短的连续信息块或序列数据段。读段可以具有任何适合的长度,例如约30个核苷酸至约1000个核苷酸之间的长度。长度通常取决于用于获得它的测序技术。在具体实施方案中,读段也可以更长,例如,2kb到10kb或更长。本方法一般性地涵盖任何读段或读长,并且不应被理解为限于当前能的读长,而是还包括该领域的进一步发展,例如,长读测序方法等的开发。

[0041] 如本文中使用的“核酸序列数据”可以是本领域技术人员已知的关于核酸分子的任何序列信息。序列数据可以包括必须转化成核酸序列的关于DNA或RNA序列、经修饰的核酸、单链或双链体序列或作为替代的氨基酸序列的信息。序列数据可以另外包括关于测序仪的信息、获取日期、读长、测序方向、测序的实体的来源、相邻序列或读段、重复的存在或本领域技术人员已知的任何其他适合的参数。序列数据可以作为本领域技术人员已知的任何适合的格式、档案、编码或文档呈现。数据可以例如是FASTQ、Qseq、CSFASTA、BED、WIG、EMBL、Phred、GFF、SAM、SRF、SFF或ABI-ABIF的格式。

[0042] 如本文中使用的,术语“从多个核酸片段读段获得序列数据”是指通过进行核酸测序反应来确定受试者或一组受试者的序列信息的过程。

[0043] 如本文中使用的术语“多重PCR”是指可供在同一反应容器内同时扩增和检测两种

或更多种靶核酸的测定法。每个扩增反应使用独特的一对引物来引发。在一些实施方案中，每个引物对中的至少一个引物以可检测部分标记。在一些实施方案中，多重反应可以进一步包括每个靶核酸的特异性探针。在一些实施方案中，特异性探针以不同的可检测部分可检测地标记。

[0044] 术语“巢式聚合酶链式反应”是聚合酶链式反应的一种改型，在本文的语境中，进行该反应来给扩增子添加序列。巢式聚合酶链式反应涉及在连续两轮聚合酶链式反应中使用两组引物，第二组引物旨在从第一轮的产物中扩增靶标。

[0045] 如本文中使用的，术语“寡核苷酸”是指由脱氧核糖核苷酸、核糖核苷酸或其任何组合组成的短聚合物。寡核苷酸的长度通常为约10、11、12、13、14、15、20、25或30至约150个核苷酸(nt)，更优选约10、11、12、13、14、15、20、25或30至约70个nt。

[0046] 术语“特异性”，如本文中对寡核苷酸引物使用的，是指当寡核苷酸和核酸比对时，引物的核苷酸序列与待扩增的核酸的一部分具有至少12个碱基的序列同一性。对核酸特异性的寡核苷酸引物是在严格杂交或洗涤条件下能够与感兴趣的靶标杂交并且基本上不与非感兴趣的核酸杂交的寡核苷酸引物。更高的序列同一性水平是优选的，并且包括至少75％、至少80％、至少85％、至少90％、至少95％、更优选至少98％的序列同一性。

[0047] 如本文中使用的，术语“受试者”或“个体”是指哺乳动物，比如人，但也可以是另一种动物，比如家畜(例如，狗、猫等)、农场动物(例如，牛、绵羊、猪、马等)或实验动物(例如，猴、大鼠、小鼠、兔、豚鼠等)。

[0048] 如本文中使用的术语“互补序列”、“互补的”或“互补性”是指与碱基配对规则相关的多核苷酸(即核苷酸序列，比如寡核苷酸或基因组核酸)。如本文中使用的，核酸序列的互补序列是指这样的寡核苷酸，当将其与所述核酸序列比对，使得一个序列的5’端与另一个序列的3’端配对时，其处于“反平行缔合”中。例如，序列5′-A-G-T-3′与序列3′-T-C-A-5′互补。在天然核酸中不常见的某些碱基可以被包括在本发明的核酸中，并且包括例如肌苷和7-脱氮鸟嘌呤。互补性不一定是完全的；稳定的双链体可能含有不匹配的碱基对或不匹配的碱基。核酸技术领域的技术人员可以凭经验考虑许多变量来确定双链体稳定性，所述变量包括例如寡核苷酸的长度、寡核苷酸的碱基组成和序列、错配碱基对的离子强度和发生率。互补性可以是“部分的”，其中只有一些核酸碱基根据碱基配对规则匹配。或者，在核酸之间可以存在“完整”、“总的”或“完全”互补。

[0049] “检测”基因或蛋白质中的突变可以通过进行适当的测定法来完成。为了检测生物样品中的基因或蛋白质中的突变，测定生物样品以确定突变基因或突变蛋白质的存在或不存在。测定可以包括从样品中提取核酸(例如，总基因组DNA和/或RNA)，并通过本领域已知的方法分析所提取的核酸。测定可涉及从生物样品中分离蛋白质并分析蛋白质。然而，测定不一定涉及核酸的提取或蛋白质的分离。也就是说，可以采用一些直接分析生物样品而不提取或分离核酸或蛋白质的测定法。

[0050] 如本文中使用的，术语“受试者”是指哺乳动物，比如人，但也可以是另一种动物，比如家畜(例如，狗、猫等)、农场动物(例如，牛、绵羊、猪、马等)或实验动物(例如，猴、大鼠、小鼠、兔、豚鼠等)。术语“患者”是指具有或疑似具有感兴趣的遗传多态性的“受试者”。

[0051] 概述

[0052] 利用Illumina MiSeq测序系统进行基因组序列分析，使用自带的MiSeq Reporter

软件检测BRCA1和BRCA2变体，并不能检出某些关键突变。具体地说，MiSeq测序系统对某些类型的变体，比如中等大小的插入或缺失，灵敏度较低。例如，如本文工作实施例中所述的，MiSeq测序系统未能鉴定两个具有大于9个碱基对(bp)的缺失的病理性BRCA1变体：40-bp缺失c.1175_1214del40和10-bp缺失c.3481_3491del10。开发了用于处理从下一代序列(NGS)平台生成的原始序列数据的序列数据分析管线(QSAP，代表“Quest序列分析管线”)，其用于检测BRCA1和BRCA2变体，并在本文中加以描述。如本文中所述的QSAP工作流程可以结合各种NGS平台用于比对和等位基因指配。在一些实施方案中，QSAP方法能够鉴定MiSeq Reporter软件遗漏的变体。

[0053] 就基因组参考序列而言，经剪切的基因组文库中的不同片段的开始和/或结束位置往往与其他片段的起始和/或末端位置不同。本文中提供的QSAP工作流程涉及通过去除表观PCR克隆来减轻扩增偏倚的可能影响，从而更好地恢复原始剪切基因组样品中的等位基因平衡。因此，生物信息学分析能够区分来自相同文库克隆的读段与来自不同文库克隆的读段。另外，QSAP工作流程涉及限于感兴趣靶区域的局部重新比对，从而提高了插入和缺失检测的灵敏度，并提供了更快更高效的分析。

[0054] 另外，当在扩增或测序引物序列中存在多态性时，使用诱饵小区(bait tile)文库捕获外显子再进行NGS，可以避免由于等位基因脱扣引起的假阴性检测的潜在原因。诱饵小区是长度为约125bp的生物素化RNA分子，用于捕获有关片段。由于诱饵小区比典型的PCR或测序引物长约100碱基，而且由于RNA/DNA杂交体比DNA/DNA杂交体更强，故多态性干扰外显子捕获的可能性降低。诱饵小区捕获相对于基于PCR的测序方法的第二个优点是可避免由于PCR或文库形成中的克隆偏倚引起的假阳性结果。

[0055] 靶基因和变体

[0056] 本文中提供的系统和方法可以应用于任何感兴趣基因中的变体的检测。示例性变体包括单核苷酸多态性、点突变、插入、缺失和易位。在一些示例性实施方案中，本文中提供的系统和方法用于检测大于5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50个或更多个碱基对的基因组缺失。

[0057] 在一些示例性实施方案中，感兴趣的基因是BRCA1基因或BRCA2基因。被测序和/或分析的BRCA1或BRCA2靶区段可以呈现BRCA1或BRCA2基因组DNA或cDNA的全部或部分。在一些实施方案中，对一个或多个BRCA1或BRACA2外显子或其部分测序。在一些实施方案中，对一个或多个BRCA1或BRACA2内含子或其部分测序。在一些实施方案中，对至少一个、二个、五个、10个或20个，乃至25个或27个外显子测序。在其他实施方案中，还评估BRCA1或BRCA2启动子区的全部或部分。

[0058] 在一些实施方案中，靶区段代表BRCA1或BRACA2基因的全部编码和非编码序列。在一个实施方案中，各个BRCA1或BRACA2靶区段当被合并时，呈现BRCA1或BRACA2编码区和所有内含子，加上第一外显子直接上游(在5’方向)的BRCA1或BRACA2启动子的约100、500、750、900或1000个上至约1200个核苷酸，加上BRCA1或BRACA2基因直接下游(在3’方向)的约50、100、150或200个上至约200、250、300或400个核苷酸。在一些实施方案中，相邻的上游区包括BRCA1或BRACA2启动子序列的全部或部分。在另一个实施方案中，呈现BRCA1或BRACA2的所有外显子、以及一个或多个内含子的一部分。

[0059] 在一些实施方案中，供检测的变体是BRCA1基因或BRCA2基因中的致病突变。在一

些实施方案中，致病突变选自表1中提供的突变。在一些实施方案中，供检测的变体是BRCA1基因中的c.1175_1214del40缺失或c.3481_3491del10缺失。

[0060] 表1.Coriell细胞系参考样品中的BRCA1和BRCA2变体

[0061]

	dbSNP HGVS名称	
样品	NM_007300.3	NP_009231.2
BRCA1		
GM13711	c.3119G>A	p.Ser1040Asn
GM13715	c.5326_5327insC	p.Ser1776delinsSerProfs
GM14634	c.4065_4068delTCAA	p.Asn1355_Gln1356delinsLysfs
GM14636	c.5621_5622insA	p.Tyr1874delinsTerProfs
GM14637	c.4327C>T	p.Arg1443Ter
GM14638	c.213-11T>G	-
GM14684	c.797_798delTT	p.Val266=fs
GM14090	c.66_67delAG	p.Leu22_Glu23delinsLeuValfs
GM14092	c.5201T>C	p.Val1734Ala
GM14093	c.1204delG	p.Glu402Serfs

[0062]

GM14094	c.1175_1214del40	p.Leu392_Ser405delinsGlnfs
GM14095	c.5200delG	p.Val1734Terfs
GM14096	c.3481_3491delGAAGATACTAG	p.Glu1161_Ser1164delinsPhefs
GM14097	c.181T>G	p.Cys61Gly
GM13714	c.5382_5383insC	p.Asn1795Glnfs
GM13713	c.3748G>T	p.Glu1250Ter
GM13712	c.2155_2156insA	p.Lys719delinsLysArgfs
GM13710	c.4327C>G	p.Arg1443Gly
GM13709	c.2068delA	p.Lys690=fs
GM13708	c.4752C>G	p.Tyr1584Ter
GM13705	c.3756_3759delGTCT	p.Leu1252_Ser1253delinsLeufs
BRCA2		
GM14170	c.5946delT	p.Ser1982Argfs
GM14622	c.6275_6276delTT	p.Leu2092Profs
GM14623	c.125A>G	p.Tyr42Cys
GM14624	c.5718_5719delCT	p.Asn1906_Ser1907=fs
GM14626	c.9976A>T	p.Lys3326Ter
GM14639	c.6198_6199delTT	p.Val2066_Ser2067delinsValHisfs

[0063] 所有突变均通过NGS结合PGM系统和MiSeq系统(使用QSAP变体判读)软件、以及Sanger测序法进行检测。

[0064] 用于分析的样品和样品制备

[0065] 本文中提供的方法可以适用于从生物样品获得的任何核酸。如本文中使用的,术语“生物样品”是指含有感兴趣核酸的样品。生物样品可以包括临床样品(即,直接从患者获得)或分离的核酸,并且可以是细胞或无细胞流体和/或组织(例如,活组织检查)样品。在一些实施方案中,从收集自受试者的组织或体液获得样品。样品来源包括但不限于:痰(经处理或未经处理)、支气管肺泡灌洗液(BAL)、支气管冲洗液(BW)、全血或任何类型的分离血细胞(例如,淋巴细胞)、体液、脑脊液(CSF)、尿、血浆、血清或组织(例如,活组织检查材料)。获得测试样品和参考样品的方法是本领域技术人员熟知的,包括但不限于吸出、组织切片、抽取血液或其他流体、外科或针吸活组织检查、采集石蜡包埋组织、采集体液、采集粪便等。在本文中,生物样品优选为血液、血清或血浆。如本文中使用的术语“患者样品”是指从寻求疾病的诊断和/或治疗或确定发生疾病的可能性的人获得的样品。

[0066] 用于制备基因组DNA样品的示例性方法包括但不适于:DNA分离、基因组DNA剪切、测量DNA浓度、DNA末端修复、衔接子连接、扩增、和富集方法,例如,诱饵捕获方法。用于制备DNA样品的示例性工艺流程提供在图1中,并在本文提供的实施例中进一步加以描述。

[0067] 在示例性实施方案中,从患者样品中分离基因组DNA,并将其随机剪切成平均大小

约250个碱基对的基因组DNA片段。在一些实施方案中，分离的基因组DNA可以进一步被纯化和浓缩。例如，可以在固体支持物、如固相可逆固定(SPRI)珠上将分离的基因组DNA纯化。

[0068] 在示例性实施方案中，将核酸衔接子添加至基因组DNA片段的5'和3'端。在一些实施方案中，至少一个衔接子包含用于识别个体DNA样品的独特索引序列(也称为索引标签、"条形码"或多重化标识符(MID))。可以将来自一个以上样品来源的经过索引的核酸单独定量，然后在测序之前合并。正因如此，索引序列的使用允许在每次测序运行中汇集多个样品(即，来自一个以上样品来源的样品)，并且随后基于索引序列确定样品来源。

[0069] 另外，衔接子可以包含用于引发扩增和/或测序反应的通用序列。在一些实施方案中，基因组DNA片段在测序前被扩增。在一些实施方案中，衔接子序列是被推荐用于Illumina测序仪(MiSeq和HiSeq)的P5和/或P7衔接子序列。参见，例如，Williams-Carrier等人，Plant J.，63(1):167-77(2010)。在一些实施方案中，衔接子序列是被推荐用于Life Technologies测序仪的P1或A衔接子序列。其他衔接子序列是本领域已知的。一些制造商建议使用提供的特定测序技术和装置的专用衔接子序列。

[0070] 衔接子可以通过连接反应来附接，或者使用连接有衔接子的和/或索引的引物来扩增来附接。当采用连接有衔接子的和/或索引的引物扩增靶片段时，衔接子序列和/或索引序列在扩增过程中被纳入扩增子(与靶标特异性引物序列一道)。

[0071] 在示例性实施方案中，对特定的序列靶物加以富集。有许多方法可用于序列选择。在一个实例中，将适合的核酸探针作为诱饵固定在固体支持物上以捕获具有互补序列的多核苷酸片段。例如，通过核酸剪切生成基因组DNA片段库，并用核酸探针(即诱饵)序列特异性捕获，可以富集基因组的选定靶区域。诱饵可以与位于感兴趣的一个或多个区域内的一个或多个外显子、内含子和/或剪接接合位点互补。在一些实施方案中，所述诱饵为RNA诱饵。在一些实施方案中，诱饵被加标签(例如，生物素化)，以便于基因组片段的纯化(例如，用链霉亲和素包被的珠子吸附生物素化的诱饵来进行纯化)。在一些实施方案中，所述诱饵为生物素化RNA诱饵。在一些实施方案中，所述诱饵为与位于BRCA1和/或BRCA2基因内的一个或多个外显子、内含子和/或剪接接合位点互补的生物素化RNA诱饵。

[0072] 在测序之前，可以通过使用对衔接子序列特异的引物进行扩增来将NGS测序所需的其他序列添加至5'和3'衔接子，NGS测序例如Illumina MiSeq™(Illumina，San Diego，CA)或Ion Torrent™Personal Gene Machine(PGM)(Life Technologies，Grand Island，NY)测序平台。

[0073] 可以采用来自在感兴趣的基因中具有已知有害变体的细胞系或核酸样品的对照样品，与含有未知序列的核酸的测试样品进行比较。

[0074] NGS测序平台

[0075] 通常使用高通量DNA测序系统，比如下一代测序(NGS)系统，来生成序列数据，高通量DNA测序系统采用DNA模板的大规模并行测序。用于生成核酸序列数据的示例性NGS测序平台包括但不限于：Illumina的边合成边测序技术(例如，Illumina MiSeq或HiSeq系统)、Life Technologies的Ion Torrent半导体测序技术(例如，Ion Torrent PGM或Ion质子系统)、Roche(454Life Sciences)GS系列和Qiagen(Intelligent Biosystems)Gene Reader测序平台。

[0076] 通常，使用两种方法来制备用于NGS反应的模板：源自单DNA分子的扩增模板、和单

DNA分子模板。对于无法检测单个荧光事件的成像系统，需要扩增DNA模板。三种最常见的扩增方法是乳液PCR(emPCR)、滚环和固相扩增。

[0077] 在克隆桥扩增方法中(该方法在Illumina HiSeq和MiSeq系统中使用)，正向和反向引物以高密度共价附接于流动池中的载玻片。支持物上引物与模板的比例限定了扩增簇的表面密度。将流动池暴露于用于基于聚合酶的延伸的试剂，当连接片段的游离端/远端与表面上的互补寡核苷酸“桥接”时，发生引发作用。重复进行变性和延伸，导致整个流动池表面上数百万个独立位置上的DNA片段发生局部扩增。固相扩增产生大约1200-1500万(MiSeq)个空间分离的模板簇(对于HiSeq，约为1-2亿个)，从而提供了游离端，以供接下来通用测序引物与之杂交以引发测序反应。然后通常使用可逆染料终止物方法进行测序，可逆染料终止物方法是使用可逆的结合有终止物的dNTP的循环方法，包括核苷酸掺入、荧光成像和切割。当每个dNTP加入时，荧光标记的终止物被成像，然后被切割掉，以允许下一个碱基掺入。这些核苷酸被化学封闭，使得每次掺入是唯一的事件。在每个碱基掺入步骤之后进行成像步骤，然后以化学方法去除封闭基团，以便于DNA聚合酶向每条链进行下一次掺入。这一系列步骤持续循环一定的次数，由用户自定义的仪器设置来决定。

[0078] 在乳液PCR方法中，首先通过基因组DNA的随机片段化生成DNA文库。使用衔接子或接头将单链DNA片段(模板)附接到珠子的表面，一个珠子与来自DNA文库的一个DNA片段附接。珠子的表面包含寡核苷酸探针，所述探针具有与结合DNA片段的衔接子互补的序列。然后将珠子在水-油乳液液滴中区室化。在水性水-油乳液中，每个捕获一个珠子的液滴是一个PCR微反应器，产生单个DNA模板的扩增拷贝。然后使用这些珠子来生成序列。在扩增溶液中使用的珠子覆盖有共价结合的寡核苷酸，所述寡核苷酸对文库的P1序列是反义的。由执行克隆扩增的Ion Torrent One Touch Instrument 2(OT2,Life Technologies,Grand Island,NY)创建微室。然后将由反义P1寡核苷酸延伸产生的DNA链的3’端与测序引物杂交，该测序引物与反义A寡核苷酸结合。向珠子中添加DNA聚合酶，然后将珠子沉积在计算机芯片状表面的微小孔中。然后使过量的四种dNTP一种一种地依次在芯片的表面上流过。当所需的核苷酸可用时，DNA聚合酶使生长链延伸。每当添加一个核苷酸时，释放一个氢分子，导致含有测序珠子的孔中的pH变化。pH变化的大小大致等于掺入的核苷酸数目，并与四种核苷酸中流过的那种核苷酸一起被检测和测量。

[0079] QSAP处理数据方法

[0080] 图2展示了根据用于序列数据分析(即，QSAP，请求序列分析流程)的各种说明性实施方式的流程图。图2所示的过程可以在计算设备上实现。在一个实施方式中，该过程被编码在包含指令的计算机可读介质上，计算设备执行所述指令时，指令使计算设备实施过程的操作。

[0081] 原始核酸测序数据来自核酸测序仪(例如来自NGS测序平台)，核酸测序仪从多个核酸片段读段生成多个核酸序列数据。优选地，数据或数据集以一种数据格式存在，更优选以统一的数据格式例如以FASTQ格式存在，并且它们的碱基质量呈Phred/Phrap或修改格式。优选的是，数据格式至少覆盖序列读段及其相关联的碱基质量。在一些实施方案中，多个原始序列数据可以被转换为统一格式。

[0082] 这些原始核酸测序数据通过本文提供的QSAP方法运行，所述方法包括改进的序列比对和变体判读程序，从而改善对序列变体的检测。QSAP方法使用高性能的计算基础设施，

它是开放源码和改良的测序工具的结合。该方法包括:Burrows Wheeler Aligner(BWA),用于对参考序列(例如hg19/GRCh37.1参考基因组)的比对映射;基因组分析工具包(GATK)的Queue,用于去重;改良的Smith-Waterman局部重新比对,和强制变体判读(例如,用于变体判读的改良的GATK)。在一个实施方案中,软剪切在比对映射过程中被关闭。例如,BWA工具可以在没有软剪切的情况下运行。被软剪切的序列是在部分映射读段中的不匹配的片段。在一些实施方案中,去除软剪切可消除由于被软剪切的读段的错误映射而导致的错误。局部重新比对可增加突变检测,包括大的插入和缺失的检测的灵敏度。该管线被设计为最大限度地提高变体判读的准确性,减少分析时间,并允许即时访问样本二进制比对/映射格式(BAM)和变体判读格式(VCF)文件。下文提供了QSAP方法的每个步骤的示例性实施方案。本文中使用的"参考序列"可以是覆盖该序列段的任何合适的预先存在的序列,其与新获得的序列数据或核酸片段读段相同或相似。

[0083] 在示例性方法中,首先将样品核苷酸序列文件(例如FASTQ文件)从核酸测序仪(例如,MiSeq仪)复制到用户运行数据(rundata)文件夹中。在仪器上利用索引读段对序列去多重化。原始FASTQ文件包含每个样品的至少2个读段,例如正向和反向读段。

[0084] 然后过滤样品FASTQ文件,以去除标记为未通过供应商质量过滤器的读段。在一些实施方案中,过滤器是碱基质量、覆盖度、周围区域的复杂性或错配过滤器的长度。可以使用序列排序(sequence sorting)工具,例如SAMtools(序列比对/映射工具)来执行过滤程序。

[0085] 然后从序列读段中修剪掉衔接子和分子标识符(MID)标签序列(即,在核酸样品制备过程期间连接到样品序列的标识符序列)。在一个实现方式中,使用FASTQ处理工具(比如fastq-mcf)来完成此过程,该工具扫描序列文件中的衔接子和标签序列,并且基于对数标度的阈值,确定一组剪切参数并执行剪切。这个步骤将产生一个包含经修剪的样品FASTQ读段的文件。经排序的样品具有独特的衔接子与MID序列组合。因此,根据衔接子及样品特异性MID序列的身份进行修剪,以确保在修剪过程中衔接子和MID序列都被去除。

[0086] 然后使用合适的参考比对算法将经修剪的读段映射到参考序列(例如,GRCh37.1人类基因组参考序列),进行排序和索引。该过程的这一部分可以使用Burrows-Wheeler Aligner(BWA)(例如版本0.7.5a-r405)(例如,使用命令:bwa mem-M-t 2)和SAMtools进行。在一些实施方案中,软剪切在比对映射过程中被关闭。也可以采用替代的基本比对工具(例如,bowtie)来进行初始比对。然后可以使用工具(如Picard)将读段组添加到BAM文件,该工具包括用于操作SAM和BAM格式文件的基于Java的命令行工具。这个步骤的输出为一个原始BAM文件,其包含每个样品的映射到基因组的所有读段。虽然SAMtools是对BAM文件进行排序和索引的标准工具,但可以使用其他可处理BAM文件的工具。实施这些比对算法的细节和方法是本领域技术人员已知的,或者可以从适合的文献来源获得,例如,Bao等人,Journal of Human Genetics,28Apr.2011,p.1-9,通过提述将该文献完整并入本文。本发明进一步设想使用这些算法的经优化或进一步开发的版本,或者使用遵循不同方案或算法逻辑的参考比对算法,包括还不可用的算法,只要可满足与本文描述的参考序列比对这一根本目的即可。

[0087] 该过程的下一步涉及管线化重新比对目标创建(pipelined realigner target creation)、局部重新比对和重复读段的标记。在一些实施方式中,这些步骤可以使用

Queue。在一个示例性实现方式中，Queue管线是基于Queue v.2.3-9分发版本中包括的DataProcessingPipeline.scala脚本的改良的Queue管线。在一些示例性实施方案中，改良包括以下的一者或多者：定制存储器和线程使用参数以适合计算平台；使用经改良的Smith-Waterman局部重新比对选项；去除碱基质量分数重新校准步骤(BQSR)，和/或Queue分析和BAM输出限于感兴趣的区域(例如，编码外显子+/-约50个碱基)，所述区域任选地填补有额外的碱基，例如约150个碱基。在一些实施方案中，使用Picard的MarkDuplicates应用程序执行明显的重复片段去除(即，去重)，这可以使用Queue来管线化。这些步骤的输出是一个经处理的序列BAM文件，其包含每个样品的所有读段，这些读段与感兴趣的区域重新比对，且去除了重复。

[0088] 在一些实施方案中，省略掉碱基质量分数重新校准(BQSR)以改善计算时间约束。在测序的感兴趣区域较小，例如为几个选定的感兴趣的基因的情况下，这样的改良是有用的。当关注较小的感兴趣区域时，BQSR的功能发挥不太好，因为它依赖于对基因组足够宽泛的采样，使得常见的变体(例如，如果参考序列包含稀有等位基因，则大多数样品将在该位置具有变体)不干扰重新校准。

[0089] 局部重新比对对于插入和缺失灵敏度是很重要的。虽然改良的Smith-Waterman比对是标准的比对工具，但它是计算密集的。在本文提供的方法中，使用改良的Smith-Waterman局部重新比对，其将重新比对限于感兴趣的区域，例如几个选定的感兴趣的基因(例如，BRCA1和BRCA2基因)。这种改良可减少重新比对目标创建和局部重新比对步骤的计算持续时间。这还使得系统更快速、更高效，因为可以减少运行时间。

[0090] 去重(deduplication)对于最小化扩增偏倚的影响是重要的。去重的功能是通过标记出在与参考序列比对时具有相同起点和终点的读段来去除明显的PCR重复。由于剪切是随机的，片段偶然具有相同起点和终点的几率极小，所以将这些片段从下游分析中去除(只保留表观克隆组的最高质量的片段)。在去重后，杂合变异体的等位基因平衡倾向于接近预期的50％，这对于灵敏度(例如，如果由于扩增偏倚而降低太低，则可能遗漏该变体)和接合性检测(例如，如果太高，则可能会被误认为是纯合的)是重要的。本文所述的用于通过在生物信息学过程中诱饵捕获和去重来制备样品的上游工艺方法有利于所需的等位基因平衡。

[0091] 然后在感兴趣区域(编码外显子+/-50个碱基)的每个位置判读变体。在一个实现方式中，可以使用GATK v 2.3-9中的UnifiedGenotyper。在一个具体实施方案中，采用GATK，其使用发现(discovery)模式，输出所有位点(emit all sites)，判读插入缺失和点突变，其中最大替代等位基因设置为2，并且indel空位开放罚分为30。判读所有变体的每一个位置可确保变体判读不被抑制(suppress)。确保不抑制变体判读对于本方法是重要的。例如，如果变体判读被抑制，则可能遗漏缺失，导致不太准确的结果。在每个位置强制变体判读后，不具有变体的位置被过滤掉，其余变体继续进行到下游处理。

[0092] GATK中的变体判读可以在发现模式或基因分型模式下执行，其中提供一系列变体用于判读存在或不存在。在一些实施方案中，可以进行这两种类型的变体判读，并将结果合并，以使灵敏度最大化。这样的方法可以解决在样品制备中使用不同扩增方案时遇到的问题(例如，无法除去PCR重复的情况)。

[0093] 在本发明的具体实施方案中，可以相应地实施过滤器或阈值，以区分显示可接受

的碱基质量的序列数据和显示不可接受的碱基质量的序列数据。如本文中使用的术语“可接受的碱基质量”是指约20和更高的phred样质量得分。phred样质量得分是Q分数，其为$-10\log_{10}(e)$，其中e是碱基判读错误的估计概率。该方法通常用于测量序列数据的准确性。较高质量得分表示碱基被错误判读的可能性较小。因此，质量得分20因此代表错误率为1/100，相应的判读准确度为99％。在本发明的进一步的具体实施方案中，可以实施过滤器或阈值以区分显示可接受的覆盖度的序列数据和显示不可接受的覆盖度的序列数据。如本文中使用的术语“可接受的覆盖度”是指约20×和以上的覆盖度。因此，在比对中覆盖某一碱基的读段的数量为约20或以上。

[0094] 在本发明的进一步的具体实施方案中，可以实施过滤器或阈值以区分显示周围区域的可接受的高复杂性的序列数据和显示周围区域的中等到低复杂性的序列数据。如本文中使用的术语“周围区域的高复杂性”是指存在重复的序列段，例如，存在重复的二聚体、三聚体，存在转座子残迹或源自转座子等的重复序列。

[0095] 在本发明的又另一个具体实施方案中，可以实施过滤器或阈值以区分显示可接受的错配长度的序列数据和显示不可接受的错配长度的序列数据。如本文中使用的术语“可接受的错配长度”是指不允许读段与参考序列完全匹配的空位。相应的匹配可以是连续匹配和约70％及以上的非连续匹配。

[0096] 二进制比对/映射格式

[0097] 图6是根据说明性实施方式的计算机系统的框图。该计算机系统可用于上文所述的序列数据分析，包括执行QSAP管线中的一个或多个或所有步骤。示例性计算系统600包括总线605或用于传送信息的其他通信组件，以及耦合到总线605以处理信息的处理器610或处理电路。计算系统600还可以包括耦合到总线以处理信息的一个或多个处理器610或处理电路。计算系统600还包括耦合到总线605以存储信息以及要由处理器610执行的指令的主存储器615，比如随机存取存储器(RAM)或其他动态存储装置。主存储器615还可以用于在处理器610执行指令期间存储位置信息、临时变量或其他中间信息。计算系统600可以进一步包括耦合到总线605的只读存储器(ROM)620或其他静态存储装置，用于存储处理器610的静态信息和指令的。存储装置625，例如固态器件、磁盘或光盘，耦合到总线605，用于持久存储信息和指令。

[0098] 计算系统600可以经由总线605耦合到显示器635，比如液晶显示器或有源矩阵显示器，用于向用户显示信息。输入设备630，诸如包括字母数字键和其他键的键盘，可以耦合到总线605，用于将信息和命令选择传送到处理器610。在另一个实现方式中，输入装置630具有触摸屏显示器635。输入装置630可以包括光标控制器，诸如鼠标、轨迹球或光标方向键，用于将方向信息和命令选择传送到处理器610，并用于控制显示器635上的光标移动。

[0099] 根据各种实现方式，本文描述的过程可以由计算系统600响应于处理器610执行包含在主存储器615中的指令安排(arrangement of instructions)来实现。这样的指令可以从另一计算机可读介质(比如存储装置625)读入主存储器615。包含在主存储器615中的指令安排被执行，使得计算系统600实施本文所述的例示性过程。也可以采用多处理安排(multi-processing arrangement)中的一个或多个处理器来执行包含在主存储器615中的指令。在替代实施方案中，可以用硬线电路代替软件指令或与软件指令组合，来实现例示性的实施方式。因此，实现方式不限于硬件电路和软件的任何特定组合。

[0100] 尽管在图6中已经描述了示例性计算系统，但是在本说明书中描述的实现方式可以在其他类型的数字电子电路中，或在计算机软件、固件或硬件中实现，包括本说明书中公开的结构及其结构等同物、或者它们中的一者或多者的组合。

[0101] 在本说明书中描述的实施方式可以在数字电子电路中，或在计算机软件、固件或硬件中实施，包括本说明书中公开的结构及其结构等同物、或者它们中的一者或多者的组合。在本说明书中描述的实施方式可以作为一个或多个计算机程序——即一个或多个计算机程序指令的模块，所述指令被编码在一个或多个计算机存储介质上，用于由数据处理设备执行或控制数据处理设备的操作——来实施。替代地或另外地，程序指令可以编码在人工生成的传播信号上，例如机器生成的电、光或电磁信号，这些信号被生成以编码用于传输到适合的接收器设备以供数据处理设备执行的信息。计算机存储介质可以是，或者包括于，计算机可读存储装置、计算机可读存储基板、随机或串行存取存储器阵列或装置，或是它们中一个或多个的组合。而且，虽然计算机存储介质不是传播信号，但是计算机存储介质可以是编码在人工生成的传播信号中的计算机程序指令的来源或目的地。计算机存储介质还可以是，或者包括于，一个或多个分离的物理组件或介质(例如，多张CD、磁盘或其他存储装置)。因此，计算机存储介质是有形的和非瞬态的。

[0102] 本说明书中描述的操作可以由数据处理设备对存储在一个或多个计算机可读存储装置的数据或从其他来源接收执行。术语“数据处理设备”或“计算装置”包括所有种类的用于处理数据的设备、装置和机器，举例来说，其包括可编程处理器、计算机、片上系统，或前述各项中的多项或前述各项的组合。所述设备可以包括特殊目的逻辑电路，例如，FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除硬件外，所述设备还可以包括创建所讨论的计算机程序的执行环境的代码，例如构成处理器固件、协议栈、数据库管理系统、操作系统、跨平台运行时环境、虚拟机，或者它们中的一者或多者的组合的代码。设备和执行环境可以实现各种不同的计算模型基础设施，如Web服务、分布式计算和网格计算基础设施。

[0103] 计算机程序(也称为程序、软件、软件应用程序、脚本或代码)可以使用任何形式的编程语言编写，包括编译或解释语言、说明性或过程语言，并且可以部署为任何形式，包括作为独立的程序或者作为模块、组件、子例程、对象或适合于在计算环境中使用的其他单元。计算机程序可以但不一定对应于文件系统中的文件。可以将程序存储于保持其他程序或数据的文件(例如，存储在标记语言文档的一个或多个脚本)的一部分、专用于所讨论的程序的单个文件、或多个协调文件(例如，存储一个或多个模块、子程序或部分代码的文件)中。计算机程序可以被部署为在一台计算机上或位于一个站点或分布在多个站点并通过通信网络互连的多台计算机上执行。

[0104] 举例来说，适于执行计算机程序的处理器包括通用和专用微处理器、以及任何类型的数字计算机的任何一个或多个处理器。通常，处理器接收来自只读存储器或随机存取存储器或两者的指令和数据。计算机的基本元件是：用于根据指令执行动作的处理器，以及一个或多个用于存储指令和数据的存储器装置。通常，计算机还将包括一个或多个用于存储数据的大容量存储装置，诸如磁盘、磁光盘或光盘，和/或与大容量存储装置可操作地耦合以接收数据或传输数据。然而，计算机不必具有这样的装置。而且，计算机可被嵌入到另一个装置中，仅举几例，例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储装置(例如，通用串行总线(USB)闪存驱动

器)。适于存储计算机程序指令和数据的装置包括所有形式的非易失性存储器、介质和存储器装置,例如包括半导体存储器装置,例如EPROM、EEPROM和闪存装置;磁盘,例如内部硬盘或可移动盘;磁光盘;以及CDROM和DVD-ROM盘。处理器和存储器可以由专用逻辑电路补充,或并入其中。

[0105] 数据报告/输出

[0106] 在变体判读的下游,可以使用度量报告生成器(例如,QC度量(例如读段数、平均覆盖度、最小覆盖度))来计算每样品的区域最小覆盖深度(计数感兴趣的区域中每个位置上具有至少Q30的读段),然后在专门的数据库中进行结果解析和编目,其中最小覆盖度低于20个读段的样品(去除重复读段之后)被标记为重复。

[0107] 在本发明的一些实施方案中,如上所述的序列分析管线可以与诊断决策支持系统相关联或连接。如本文中使用的"诊断决策支持系统"是指这样的系统,其包括用于提供受试者的序列数据的输入,以及在具体实施方案中任选地包含其功能读出(例如基因或非编码RNA表达、或蛋白质水平)。另外,该系统包括用于将包含核酸片段读段的核酸序列数据装配成连续核苷酸序列区段的程序元件或计算机程序或软件(当被处理器执行时,其被适配为进行如上文定义的方法的步骤)、以及用于输出受试者的连续核苷酸序列区段变异的输出、和用于存储输出信息的介质。优选地,输出信息能够指示基因组修饰的存在或不存在,更优选地指示受试者罹患疾病或具有疾病易感性。

[0108] 可以对变体判读文件数据进行自动初步评估,并作为注释文件返回(retuned)。初始自动评估可以包括变体分子分类(例如同义、错义、无义、移码)、全面临床证据(大部分由科学家从同行评议的文献中拣选)和基于证据的临床决策支持,其有助于观察到的变体的初始分类。初步自动评估可改善周转时间,并最大限度地增加用于变体的最终临床评估的信息。评估软件可以使用预定义的评分和分类规则进行配置,所述规则包括美国医学遗传学学院(American College of Medical Genetics)推荐的指南和/或证据的汇编,包括最新的文献。评估软件可以为审阅者提供对变体相关文献的直接访问,并提供关于自动评估如何获得的透明度,以便于审阅原始数据。也可以通过Web界面、另外的基因座特异性数据库和验证性查询,利用人工审查进行最终分类。

[0109] 可以将测序数据或评估分析向认证实验室、医师或直接向患者报告。定制可视化(customized visualization)可用于人工审阅提交给临床报告的结果。在本发明的又另一个实施方案中,诊断决策支持系统可以是电子图片/数据归档和通信系统。在一些实施方案中,测序数据或评估分析提供了患者患有疾病或病症的概率分数。在一些实施方案中,测序数据或评估分析提供了患者具有发生疾病或病症的风险的概率分数。

[0110] 可以根据本发明检测或诊断或预测的疾病或病症可以是本领域技术人员已知的任何可检测的疾病。在优选的实施方案中,所述疾病可以是遗传性疾病或病症,特别是可以基于基因组序列数据检测到的病症。这些病症包括但不限于例如在适当的科学文献、临床或医学出版物、合格的教科书、公共信息库、互联网资源或数据库中提到的病症,特别是在http://en.wikipedia.org/wiki/List_of_genetic_disorders中提到的一种或多种病症。

[0111] 在本发明的一个特别优选的实施方案中,被检测或诊断或预测的疾病是癌性疾病,例如,本领域技术人员已知的任何癌性疾病或肿瘤。在具体实施方案中,被检测或诊断或预测的疾病是乳腺癌、卵巢癌或前列腺癌。

[0112] 在特定实施方案中,所述诊断决策支持系统可以是分子肿瘤学决策工作站。决策工作站可优选地用于决定受试者的癌症治疗的开始和/或继续。进一步设想了对于不同疾病类型、例如对于上述提到的任何疾病的类似决策工作站。

实施例

[0113] 实施例1

[0114] 本公开描述了适合于临床实验室的用于检测BRCA1和BRCA2变体的快速高通量测序测定法的开发和验证。利用MiSeq/QSAP联合一式二份测试的初始1006份临床样品的结果显示没有差异性变体判读。在一个实施方案中,使用诱饵小区外显子捕获的基于NGS的测定法在参考实验室中检测BRCA1/2变体。在测试过程中,采用了两种不同的NGS平台:Illumina MiSeq系统和Life Technologies Ion Torrent个人化操作基因组测序仪。正如下面解释的,使用两个NGS平台获得前521份临床样品的结果,并且使用重复MiSeq运行获得另外1006个结果。

[0115] 图1展示了检测BRCA1和BRCA2变体的NGS测定法的总体概述。下面的说明描述了本发明的一个实现方式。

[0116] DNA样品

[0117] 来自具有在BRCA1(n=21;表1)或BRCA2(n=6;表1)中的已知有害变体的细胞系的DNA样品购自Coriell Mutant Cell Repository(Camden,NJ)。这些参考样品包含致病和非致病性变体。在一个实现方式中,获得了67名未受影响的、先前未测试过BRCA突变的个体的血液样品。进行Sanger测序以确定BRCA1或BRCA2序列变异的存在。在此实现方式中,在志愿者人群中鉴定出352个良性变体,并将它们用于技术验证。

[0118] DNA制备

[0119] 按照制造商的说明,使用Roche Molecular Systems(Indianapolis,IN)的Roche Magnapure™系统在96孔微量滴定板中从外周血细胞分离基因组DNA。根据制造商的说明,使用自适应聚焦声波技术(E220Focused Ultra-Sonicator,Covaris Inc.,Woburn,MA),将基因组DNA随机剪切至250个碱基对的平均大小。

[0120] 通过SPRI珠子和PEG/氯化钠混合物浓缩

[0121] 剪切后,将DNA浓缩2倍,除去大小不够的DNA分子。这是用SPRI(固相可逆固定)珠子(AMPure Beads,Agencourt,Beverley,MA)完成的。将珠子悬浮在EDTA-聚乙二醇(PEG)溶液中。

[0122] DNA末端修复

[0123] 在连接衔接子之前修复DNA分子的末端。这是使用同时具有5'至3'聚合酶活性和3'至5'核酸外切酶活性的DNA聚合酶完成的,从而填充5'突出端并去除3'突出端以产生平端。另外,DNA片段的5'端也在该过程中被磷酸化。

[0124] 衔接子连接和缺口修复

[0125] 每个5'衔接子含有独特的分子鉴别(MID)序列(条形码),用于标识个体的DNA样品。另外,它包含P5序列的一部分。3'衔接子对于所有标本是共同的,且包含P7MiSeq序列的一部分。两个衔接子都未被5'磷酸化。连接反应中还可以纳入与每个衔接子互补的短寡核苷酸,以确保衔接子仅与DNA片段连接,而不与自身连接。在连接过程中,两个衔接子的摩尔

比彼此相等，但与片段化的DNA相比是过量的。在这个程序之后，全部连接产物的大约一半是优选的种类，即：5'-(P5)-MID-BRCA_基因DNA-共同(P7)-3'。如上所述使用SPRI珠子清理样品，并通过DNA聚合酶修复连接位点处的缺口。聚合酶在缺口位点添加核苷酸，产生用于PCR扩增的引物结合位点。

[0126] 预杂交扩增

[0127] 为了提高加有衔接子的(adapted)DNA片段的比率，进行了非等位基因特异性PCR。所使用的引物与5'和3'衔接子序列互补。

[0128] 通过外显子捕获富集靶标

[0129] 将全部加有条形码的患者DNA片段合并以产生"文库"，加入到杂交反应混合物中并在65℃下温育12小时。该混合物含有生物素化的RNA诱饵。诱饵与BRCA1和BRCA2基因(外显子区和剪接接合位点、以及选定的内含子区)互补，以允许与适当的患者DNA片段杂交。杂交后，将文库与包被有链霉亲和素的珠子混合，以吸附生物素化的RNA诱饵。在70℃下洗涤文库-RNA诱饵杂交物，以去除非BRCA DNA。

[0130] 第二非特异性扩增

[0131] 利用融合引物，将Illumina MiSeq™(Illumina,San Diego,CA)或Ion Torrent™Personal Gene Machine(PGM)(Life Technologies,Grand Island,NY)测序平台所需的额外序列添加至5'和3'衔接子。将DNA文库分成两半。将其中的一半用具有与5'和3'衔接子互补的部分、并可添加用于MiSeq测序的额外序列的融合引物(P5和P7序列)扩增，另一半用可添加用于PGM测序的额外序列的一组引物(P1和A序列)扩增。

[0132] 通过Qubit定量DNA浓度

[0133] 使用高灵敏度Qubit试剂盒(Life Technologies)来定量DNA，该试剂盒使用基于插入染料的方法。

[0134] NGS测序

[0135] 稀释所制备的文库，以便从单个DNA分子扩增出良好分离的相同产物的簇(即克隆性扩增)。根据制造商的方案实施MiSeq和PGM NGS方案。

[0136] MiSeq

[0137] 将单链文库装入MiSeq测序盒的孔21。仪器使文库流过流动池，文库在流动池中与反义P5和P7寡核苷酸杂交，这些寡核苷酸与文库上的衔接子互补。稀释文库，以便从单个DNA分子扩增出良好分离的相同产物的簇(克隆性扩增)。这是通过等温桥扩增(isothermal bridge amplification)实现的。将荧光团标记的核苷三磷酸加到流动池中，然后用激光激发。MiSeq记录发射光谱，然后切割掉抑制进一步合成的核苷酸阻断剂，从而允许添加下一个核苷三磷酸。以这种方式对片段进行测序。

[0138] Ion Torrent™PGM

[0139] PGM采用乳液PCR，即在油中漂浮的微小水滴内进行的扩增。进行乳液PCR以在单个测序"珠子"上获得单个DNA分子的许多拷贝(即克隆性扩增)。然后使用这些珠子来生成序列。在扩增溶液中使用的珠子覆盖有共价结合的寡核苷酸，这些寡核苷酸是对文库的P1序列反义的。微室(micro-chambers)由实施克隆性扩增的Ion Torrent One Touch Instrument 2(OT2,Life Technologies,Grand Island,NY)创建。

[0140] 然后，由反义P1寡核苷酸延伸产生的DNA链在其3'端与测序引物杂交，该测序引物

与反义寡核苷酸结合。向珠子中加入DNA聚合酶，然后将珠子沉积在计算机芯片状表面的微小孔中。然后使过量的四种dNTP中一种一种地在芯片的表面上依次流过。当有需要的核苷酸可用时，DNA聚合酶使生长链延伸。每当添加一个核苷酸时，释放一个氢分子，导致含有测序珠子的孔中的pH变化。pH变化的大小与掺入的核苷酸数目大致相等，并与四种核苷酸中流过的那种核苷酸一起被检测和测量。

[0141] 生物信息学处理

[0142] 在测序反应之后，进行序列比对和等位基因指配。最初，对于MiSeq，使用随仪器提供的MiSeq Reporter™软件。然而，这个过程并未一致地鉴定出大于9bp的缺失。为了鉴定大于9bp的缺失，使用了以下的工作流程，该流程被称为QSAP。QSAP是一个生物信息学管线(bioinformatics pipeline)。QSAP是总工作流程的一个专门化的部分，整合了开源的、内部开发的、和许可授权的序列分析模块。该分析管线使用高性能计算基础设施，包括：用于映射到hg19/GRCh37.1参考基因组的Burrows Wheeler Aligner(BWA)、以及结合基因组分析工具包(GATK)使用的Queue，用于去重、改良的Smith-Waterman局部重新比对、和变体判读。该管线被设计为最大限度地提高变体判读的准确性，减少分析时间并允许即时访问样本二进制比对/映射格式(BAM)和变体判读格式(VCF)文件。

[0143] 具体地说，在从MiSeq仪中复制样品FASTQ核苷酸序列文件(已在仪器上利用索引读段去多重化)后，过滤样品FASTQ文件以去除被标记为未通过供应商质量过滤器的读段。该过程可以使用序列排序工具，例如SAMtools(序列比对/映射工具)来执行。修剪掉衔接子和分子标识符(MID)标签序列(即，在核酸样品制备过程中连接到样品序列的标识符序列)。在一个实现方式中，使用FASTQ处理工具(比如fastq-mcf)来完成此过程，该工具扫描衔接子的序列文件，并且基于对数标度的阈值，确定一组剪切参数并执行剪切。

[0144] 修剪后，将读段映射到基因组参考序列，进行排序和索引。这可以使用Burrows-Wheeler Aligner(例如版本0.7.5a-r405)(例如，使用命令：bwa mem-M-t 2)和SAMtools来完成。然后可以添加读段组。可以使用Picard添加读段组，Picard包括用于操作SAM文件的基于Java的命令行工具。该过程的接下来的步骤涉及管线重新比对目标创建、局部重新比对和重复读段的标记。在一些实现方式中，Queue可用于这些步骤。在一个实现方式中，Queue管线是基于Queue v.2.3-9分发版本中包含的DataProcessingPipeline.scala脚本。在该实现方式中进行了以下改良：定制了存储器和线程使用参数以适合计算平台，使用改良的Smith-Waterman局部重新比对选项，去除了碱基质量分数重新比对步骤，将Queue分析和BAM输出限于填补有150个额外碱基的感兴趣区域(编码外显子+/-50个碱基)。然后在感兴趣区域中(编码外显子+/-50个碱基)的每个位置判读变体。在一个实施方式中，采用GATK v 2.3-9中的UnifiedGenotyper，使用发现模式(discovery mode)，输出所有位点(emit all sites)，判读indel和点突变(calling indels and point mutations)，其中最大替代等位基因设置(maximum alternate alleles setting)为2，插入缺失空位开放罚分为30。

[0145] 对于PGM数据，使用随仪器提供的Torrent Suite™软件进行生物信息学分析。

[0146] 在变体判读的下游，可以使用度量报告生成器(例如，QC度量(例如读段数、平均覆盖度、最小覆盖度))来计算每个样品的区域最小覆盖深度(regional minimum coverage depths)(计数感兴趣的区域中每个位置上具有至少Q30的读段)，然后在专门目的的数据库中进行结果解析和编目，其中将最小覆盖度低于20个读段的样品(去除重复读段之后)标记

为重复。

[0147] 在一些实施方案中,从Illumina MiSeq FASTQ文件到提交给认证的实验室的最终报告的序列信息均由BRCA1/2高级测序生物信息学模块化工作流程来管理;认证的实验室例如由CLIA(Clincal Laboratory Improvement Amendments,美国临床实验室改进法案修正案)或CAP(College of American Pathologist,美国病理学家协会)认证的实验室。在一些实施方案中,所述工作流程利用定制可视化手段来对提交给临床报告的结果进行人工审阅。

[0148] 在一些实施方案中,去身份标识的(de-identified)VCF文件将经过自动初步评估,并生成注释文件。初始自动评估利用变体分子分类(例如同义、错义、无义、移码)、全面临床证据(大部分由科学家从同行评议的文献中拣选),并提供基于证据的临床决策支持,可有助于对观察到的变体的初始分选。初步自动评估可改善周转时间,并最大限度地增加用于变体的最终临床评估的信息。该自动评估具有另外两个基本特征。首先,该评估是使用Quest Diagnostics预定义的评分和分类规则进行的,这些规则以美国医学遗传学学会(American College of Medical Genetics)推荐的指南作为中心思想,结合包括最新的文献在内的证据汇总。其次,评估软件可以为审阅者提供对变体相关文献的直接访问,并提供关于自动评估如何获得的透明度,以便于审阅原始数据。随后借助Ingenuity的VCS Web界面、其他基因座特异性数据库和验证性查询进行人工审阅,以完成最终分类。对于PGM数据,使用随仪器提供的Torrent Suite™软件进行生物信息学分析。

[0149] 变体评估

[0150] 根据美国医学遗传学学会的指南,由一队变体科学家(VS)人工进行变体评估。通过软件程序Alamut(其提供基因组坐标)和SIFT体外功能分析来分析VCF文件。VS重新检查单独运行的质量度量,如果运行通过QC,则继续进行评估。此时,还对来自MLPA反应的缺失/重复结果进行审阅。

[0151] 然后,VS审阅被判读的变体,以确保与IGV数据一致。如果变体鉴定准确,则将变体从Alamut HT加载到被称为QuestIQ的专有数据库中。以下字段由Alamut HT软件界面自动填写:基因|变体、变体ID、参考序列、DNA水平、突变类型、代码解释、PUC、基因代码、外显子、核苷酸、变化、密码子、氨基酸、dbSNP rs号、dbSNP链接、SIFT、物种保护、到VUS分析文本的链接、到剪接报告的链接、和MolGen登录号。

[0152] 然后,VS利用基因特异性数据库(例如UMD、BIC、LOVD、IARC、ClinVar、ARUP、kConFab、HGMD、InSIGHT)搜索进一步的信息。然后,使用ESP和dbSNP评估变体频率。如果适用,使用翻译后预测性数据库,如NetPhosk,NetPhos,ScanSite:S、T、Y磷酸化预测,Yin o Yang:O连接的GlcNac预测。利用RefSeq数据库,使用Alamut HC中的链接软件进行剪接预测。然后使用功能预测程序SIFT和PolyPhen2。

[0153] 随后,通过Alamut、PubMed、ScienceDirect和BioMed Central,利用Google搜索,进行手工文献检索来确定是否进一步存在有关特定变体的支持数据。将所有相关结果输入到IQDB数据库中。

[0154] 根据ACMG指南进行最终变体分选,结果输入到IQDB数据库中。分选被评分为良性、可能良性、VUS、可能致病性、致病性。然后将此结果传送给指导者进行二次审阅和报告撰写。

[0155] 结果

[0156] 测定法开发

[0157] 在测定法开发的过程中，使用自带的MiSeq Reporter软件的MiSeq测序系统未能鉴定测试的Coriell样品中2种致病性BRCA1变体。两种变体均为>9bp的缺失：40bp缺失c.1175_1214del40和10bp缺失c.3481_3491del10。它们是这些样品中仅有的两种>9bp的缺失。然而，上述QSAP工作流程可以用于如上所述的比对和等位基因指配(QSAP)。图3显示了含有40bp缺失的样品的比对呈现。这种有害突变被PGM/Torrent Suite软件鉴定，但未被MiSeq/MiSeq Reporter软件鉴定出来。然而，当使用配备QSAP软件的MiSeq时，该缺失被清楚地鉴定出来(图3)。对于10bp缺失有同样的发现：MiSeq/MiSeq报告程序始终遗漏了该缺失，而PGM/Torrent Suite和MiSeq/QSAP总是鉴定出了该缺失(数据未显示)。MiSeq/QSAP组合和PGM/Torrent Suite组合对于验证集合中的BRCA1和BRCA2变体都显示出100％灵敏度。使用两个平台进行了技术验证。

[0158] 由于NGS测序错误可能由PCR或克隆性扩增错误引起，因此开发了QC度量来克服这些潜在问题，QC度量利用了这一事实：随机剪切产生具有不同的起始和终止位置的文库克隆。因此，生物信息学分析能够区分来自相同与不同的文库克隆的读段。这使我们能够开发一个最小QC度量，根据该度量，每个被靶向的碱基必须有来自至少20个独特克隆的高质量序列。然而，通常实现了335个独特读段的“平均读取深度”。

[0159] 技术验证：测定内精度

[0160] 通过分析从三个血液样品中提取的DNA，在每个平台上进行五次重复，证明了测定内精度。每个样品至少有一个BRCA1或BRCA2变体。在用MiSeq/QSAP组合进行检测时，在每个样品中鉴定出所有变体在五次重复中均100％一致。然而，PGM/Torrent Suite平台显示出低水平的随机测序错误。在PGM仪上的总测定内一致性(overall intra-assay concordance)仅为96.2％。在一个样品中，在第五个重复中检测到一个单碱基插入，而在其他重复中未检测到该单碱基插入。另外，在第四个重复中未鉴定出一个良性变体。在另一个样品中，第五个重复包含4个在任何其他重复中不存在的SNP，并且在重复3中判读出的一个SNP未在别处被判读。

[0161] 将MiSeq/QSAP等位基因判读与PGM/Torrent Suite判读进行比较，有几处不一致。一个样品显示在2个SNP位点处在两个平台之间不一致：PGM/Torrent Suite的一个重复在第13号染色体上的位置32906554处判读出一个A插入，但在PGM/Torrent Suite的其他重复以及在MiSeq/QSAP平台上都未鉴定出该A插入。PGM/Torrent Suite的一个重复中也未检测到17号染色体上的位置41245466处的rs1799949。结果，对于这个样品，在2个平台之间的样品内一致性为88％。第二个样品的所有变体判读在两个平台是一致的，故平台之间的一致性为100％。第三个样品由于PGM/Torrent Suite上的测序错误(每个SNP的5个重复中有1个)，11个SNP有5个不一致，一致率仅为64％。

[0162] 技术验证：测定间精度

[0163] 对于来自67个假定未受影响的个体的残留实验室样品的DNA、以及来自Coriell的27个DNA标本，以三个重复的实验设置进行了分析。每次运行中还纳入了两个阴性对照(质量控制空白[QCB])和无模板对照[NTC])。在MiSeq/QSAP和PGM/Torrent Suite平台上均检出了为每次运行准备的文库。PGM和MiSeq仪的3次重复运行中检测到的所有变体均使用IGV

(版本2.3.14)通过人工审阅加以验证。在2次以上重复运行中不合格的样品被排除在测定间变异性评估之外。

[0164] 在PGM/Torrent Suite平台上检测到的经验证变体较MiSeq/QSAP平台为少,原因是PGM/Torrent Suite平台上的测定不合格率更高(表2)。PGM/Torrent Suite的测定间精度为96.7%,MiSeq/QSAP的测定间精度为96.4%(表2)。值得注意的是,在MiSeq/QSAP平台上的3次重复中检测到的差异性判读代表假阳性结果;它们大多来自一个重复的一个样品,可能提示样品制备或孔污染的问题。

[0165] 表2.相对于Sanger的变体判读的测定间一致性

[0166]

	PGM/Torrent Suite	MiSeq/QSAP
一致的判读	1550	2188
有差异的判读	53	13

[0167]

总判读	1603	2201
一致性%	96.7%	99.4%

[0168] 所分析的样品包括27个具有已知有害突变的对照DNA样品和67个来自志愿者的样品,共有352个良性变体。

[0169] 不合格样品

[0170] 样品不合格定义为在任何外显子上均未达到>40x的平均覆盖深度。对于MiSeq仪,在重复1和2中没有不合格,而在重复3中有8个样品不合格。因此,对于重复3的总不合格率为8.5%(8/94),或总体上为2.8%(8/282)。对于PGM,重复1的不合格率为9.6%(9/94),重复2的不合格率为13.8%(13/94),重复3的不合格率为26.6%(25/94);总不合格率为16.7%(47/282)。具有低覆盖度的所有不合格样品都在来自给予同意的受试者的对照样品中,这可能反映出Coriell DNA样品中的DNA质量更高。

[0171] 在MiSeq/QSAP平台上的重复3中以及在PGM/Torrent Suite平台上的所有3次重复运行中,四个样品不合格。这个发现提示了样品质量问题,不过相同的样品在MiSeq/QSAP平台上在重复1和2中的所有区域中均被成功测序。所有在MiSeq/QSAP平台上不合格的样品均来自重复3。重复3在PGM/Torrent Suite平台上的不合格率也是最高的。这指向了该重复的样品制备问题,因为一直到杂交捕获步骤完成为止,两个平台的样品文库都是一起制备的。

[0172] 平台间一致性

[0173] 由于PGM运行中的不合格多于MiSeq运行中的不合格,所以仅验证了一部分来自MiSeq/QSAP平台的差异性变体判读,以及来自PGM/Torrent Suite平台的判读。来自MiSeq/QSAP的所有8个差异性变体判读都来自重复2。这8个中的4个在PGM/Torrent Suite平台上的重复2中同样被观察到。另外,8个变体判读中的5个是在一个样品中观察到的。

[0174] 分析灵敏度:检测限

[0175] 空白限(LOB)

[0176] 在整个测定过程中都带上了加有条形码的NTC和QCB样品,这些样品与所有其他样

品的处理相同。将映射到hg19基因组序列的读段数与每样品的平均比对读段数进行比较。对于MiSeq/QSAP平台，在重复1中，NTC或QCB没有映射到人类基因组的读段。在重复2中，QC空白没有读段，但NTC有1,486个读段。总计为板上读段平均数的1.1％，远低于等位基因判读的20％阈值。在重复3中，NTC有46个读段，占板1中的读段平均数的0.044％。

[0177] 对于PGM/Torrent Suite平台，NTC和QCB在板1中显示出0.255％和0.029％的比对读段。在板2中，NTC有平均比对读段的9.2％，而QC空白为零。第三重复板的值NTC和QCB分别为0.268％和0.063％。NTC和QCB在两个平台上都显示出可接受的低比对读段总数。比对读段或者是不可检出，或者是远低于我们对于变体判读的20％的截止阈值。

[0178] 检测限(LOD)

[0179] LOD被定义为在外显子区上的平均读取深度保持在每碱基≥40个读段时的最低DNA浓度(ng/μL)。为了确定LOD，可以采用以下实验。连续稀释两个Coriell DNA样品GM14094和GM14096、以及从没有致病性BRCA的67名对照个体中选择的一个随机DNA样品。在MiSeq/QSAP平台上，对照DNA未能达到所需的5ng/μL的平均读取/覆盖深度，表明MiSeq/QSAP平台的最小样品输入(LOD)必须大于5ng/μL(所有剪切反应在80μL体积中进行)。另外，对于高于这个下限的3个样品，所有变体在每个浓度下被一致地判读(即，100％一致)。在PGM/Torrent Suite平台上，样品在5ng/μL时不合格，表明最小样品输入也必须大于5ng/μL(400ng DNA)。在两个平台上，在所有浓度下都成功检出40bp和11bp的缺失突变。然而，对于非Coriell对照样品只有99.96％的判读变体是一致的。在15ng/μL时，使用PGM/Torrent Suite平台在对照DNA中判读出一个插入，该插入在任何其他浓度下均不存在，可能是由于测序错误所致。

[0180] 准确度

[0181] 在MiSeq/QSAP和PGM/Torrent Suite平台上，从Coriell获得的27个DNA样品包括在本验证研究中，分3次独立运行。在两个平台的所有三次验证运行中均成功地检出了样品中所有先前已知的BRCA1和BRCA2变体(即，癌症相关突变的准确度达100％)。另外，确定了在67个对照样品中检测到的352个良性序列变化的变体判读的总准确度。MiSeq/QSAP平台上只有一个遗漏的判读，系由于低读取深度(覆盖度)所致。此错误可以通过调整我们的QC度量中的最小深度要求来避免，因为这是在实施测定前进行的。PGM/Torrent Suite平台产生了两个假阳性判读、一个测序错误、和37个遗漏的变体判读，大多仅在三次验证运行中的一次观察到。然而，有4个变体未被Ion Reporter判读，这些变体通过人工审阅借助比对软件检出。总的来说，MiSeq/QSAP平台的错误率为<0.1％(1/1056)，PGM/Torrent Suite平台的错误率为3.7％(39/1056)。使用调整的QC参数，MiSeq/QSAP组合具有100％的灵敏度和几乎100％的特异性。通过人工审阅所有阳性样品，MiSeq/QSAP组合也实现了100％的特异性。

[0182] 前521个临床样品

[0183] 为了初始临床试验发布，使用MiSeq/QSAP平台和PGM/Torrent Suite平台变体判读软件进行了突变分析。对于在2个平台上有差异结果的样品，或者是人工审阅病例来确定差异的原因，或重新测试样品以确认。前521个报告病例中有35个差异，其中34个是由于PGM/Torrent Suite错误所致。单个MiSeq/QSAP平台测序的唯一一个错误是一个良性多态性的假阴性结果。比对的人工审阅揭示，这是由于链偏倚(19％变体)加上低覆盖度所致。可以随后调整QC参数，以利用随机剪切允许过滤重复读段这一事实。QC接受度量(QC

acceptance metric)要求每个测定中的每个碱基要从至少20个独立读段分析。由此,平均深度通常在从几百到几千的范围。使用经过调整的QC参数,MiSeq/QSAP组合具有100%的灵敏度和100%的特异性。对于所有阳性病例,人工审查比对,作为进一步的质量度量。

[0184] 在对QC度量进行这些调整后,MiSeq/QSAP平台在连续500多次分析中没有产生更多的错误。然而,PGM/Torrent Suite组合产生了2个针对致病性BRCA1变体的假阴性结果:一个10碱基对插入和一个64p缺失。这两种致病性变体均被MiSeq/QSAP平台检出。图4显示了针对该64bp缺失的QSAP比对。在此观察之后,可以停止使用PGM/Torrent Suite平台。在测试期间,为了确定我们的新质量度量能保证鉴定出所有变体,所有MiSeq/QSAP分析都重复进行。重复实验将确保由于链偏倚、低覆盖度或文库创建引起的任何假阳性或假阴性都被检出。在1006次连续重复的MiSeq/QSAP运行中检出了5000个以上的变体,在重复分析之间没有差异性结果(数据未显示)。因此,不需要重复运行。

[0185] NGS平台相比于标准Sanger测序的一个优势是它们能够确定2个SNP在取向上是顺式还是反式。如果两个变体被捕获在单个读段中(在这种情况下少于250个碱基),则表明它们为顺式。如果它们在分别的读段中被捕获,则表明它们为反式。图5显示了一名具有处于顺式的两个点突变的个体。在前521个临床样品中已经见到两名这样的连锁变体。另外,在例行操作中,MiSeq/QSAP平台比PGM/Torrent Suite平台稳健得多。因此,可以停止PGM/Torrent Suite平台试验,并且可以对每个病例重复执行MiSeq/QSAP运行,以确定是否存在由于文库形成所致的任何潜在的假阳性或假阴性问题。

[0186] 在初始的100个重复样品中注意到重复的MiSeq运行之间有一个差异。这是一个良性多态性,它在一次运行中被检出,但在第二次运行中未检出。对测序数据的检查显示,在第二次测序运行中存在显著的链偏倚,导致针对该变体的19%变体频率。此时我们的QC截止值为20%。为了防止此错误的再次发生,修改了变体判读的QC度量。DNA的随机片段化导致每个单独的DNA片段具有唯一的起始和终止核苷酸。这允许生物信息学唯一地识别被测序的克隆。将QC度量调整为要求来自至少20个不同克隆的每个编码外显子的每个碱基、以及所述外显子中的50个碱基对都被测序。这可消除由于测序偏倚所致的错误,因为来自过度呈现(overrepresented)的克隆的读段会被忽略,并且通过要求测序克隆的最小数量,杂合子具有接近50%的呈现。在作出这种调整之后,在连续1006次以上的分析中没有发现差异。

[0187] 数据表明,结合自带MiSeq Reporter软件的Illumina MiSeq测序仪和结合自带Torrent Suite软件的Life Technologies PGM两者都不适合于BRCA1和BRCA2的临床实验室测序。MiSeq系统不能检测到大于9bp的插入和缺失,使得其对于BRCA检测而言不可接受,因为许多已被描述的有害突变在该大小范围内。类似地,带有Torrent Suite软件的PGM不能检测到10个碱基对的插入和64bp的缺失,使该平台失去临床BRCA测试的资格。然而,通过将随机剪切与诱饵小区捕获、带有QSAP比对和等位基因判读软件的生物信息学的MiSeq平台、以及我们的质量度量相结合,本公开的测定法在我们的技术验证系列中对于BRCA1和BRCA2序列变异具有100%的灵敏度和特异性。现实世界的表现可能达不到这样的精度水平。

[0188] 使用NGS结合诱饵小区外显子捕获提供了几个优势。首先,NGS之前的诱饵小区外显子捕获降低了由于等位基因脱扣引起的假阴性结果的可能性,等位基因脱扣可能在扩增

或测序引物序列中存在多态性时在基于PCR的方法中发生。其次，使用5x冗余小区排布(tiling)，每个外显子被多个诱饵捕获，进一步降低了由于个体序列变异引起的假阴性结果的机会。诱饵小区捕获相对于基于PCR的靶富集方法的第三个优点是避免由于PCR或文库形成中的扩增子偏倚引起的假阳性结果。如果在早期PCR或文库扩增循环中发生碱基取代错误，则错误会被扩散(propagate)，在测序之前并产生混合群体(mixed population)。如果在单个扩增子中发生错误，并且扩增子被优先测序，可能导致假阳性结果。利用诱饵小区捕获方法，在诱饵小区捕获之前将基因组DNA随机剪切成大约250bp的片段。文库形成发生在捕获之后。因此，每个片段具有不同的5'和3'末端，并且序列比对软件可以检测2个读段是否从同一个片段生成。可以将过滤器设置为仅接受来自独特片段的读段，从而消除由于早期PCR或文库扩增错误所致的测序错误的可能性。所选择的质量控制度量要求来自至少20个不同克隆的读段，从而最大限度地降低NGS中假阳性测序的风险。

[0189] 相对于Sanger测序，NGS还具有检测大约在250bp内的SNP相位(即剪切后的基因组DNA片段的长度)的优点。由于这项技术对单个分子进行测序，所以处于顺式取向的2个SNP将一起出现在同一个读段中；如果取向为反式，则2个SNP将在分开的读段中出现。Sanger测序在不借助家系研究的情况下无法区分顺式和反式取向。

[0190] 总之，本文描述了适合于临床实验室的用于检测BRCA1和BRCA2变体的快速高通量测序测定法的开发和验证。利用MiSeq/QSAP组合重复测试了1006份初始临床样品，结果显示没有差异性变体判读。

[0191] 虽然本说明书包含许多具体的实施细节，但是这些细节不应被解释为对任何发明的范围或所要求保护的范围的限制，而是对特定发明的具体实施方式特有的特征的描述。在单独实施方式的上下文中，在本说明书中描述的某些特征也可以在单个实施方式中组合实现。相反，在单个实施方式的上下文中描述的各种特征也可以分开地或以任何适合的子组合在多个实施方式中实现。而且，虽然上文可以将特征描述为以某些组合的方式起作用，并且甚至最初因此而要求保护，但要求保护的组合的一个或多个特征在某些情况下可以从组合中去除，并且所要求保护的组合可以涉及子组合或子组合的变化。

[0192] 类似地，尽管在附图和表中以特定顺序描绘了操作，但是这不应被理解为要求以所示的特定顺序或按先后顺序执行这样的操作，或执行所有展示的操作以实现期望的结果。在某些情况下，多任务处理和并行处理可能是有利的。此外，上述实施方式中的各种系统组件的分开不应被理解为在所有实施方式中均要求这样的分开，并且应当理解，所描述的程序组件和系统通常可以集成在单个软件产品中或包装在多个软件产品中。

[0193] 因此，已经描述了本发明的特定实施方式。其他实施方式在以下的权利要求书的范围之内。在一些情况下，可以按照不同的顺序进行在权利要求中叙述的活动并且仍然实现期望的结果。另外，附图中所描绘的过程不一定需要所示的特定顺序或先后顺序来获得期望的结果。在某些实施方式中，多任务处理和并行处理可能是有利的。

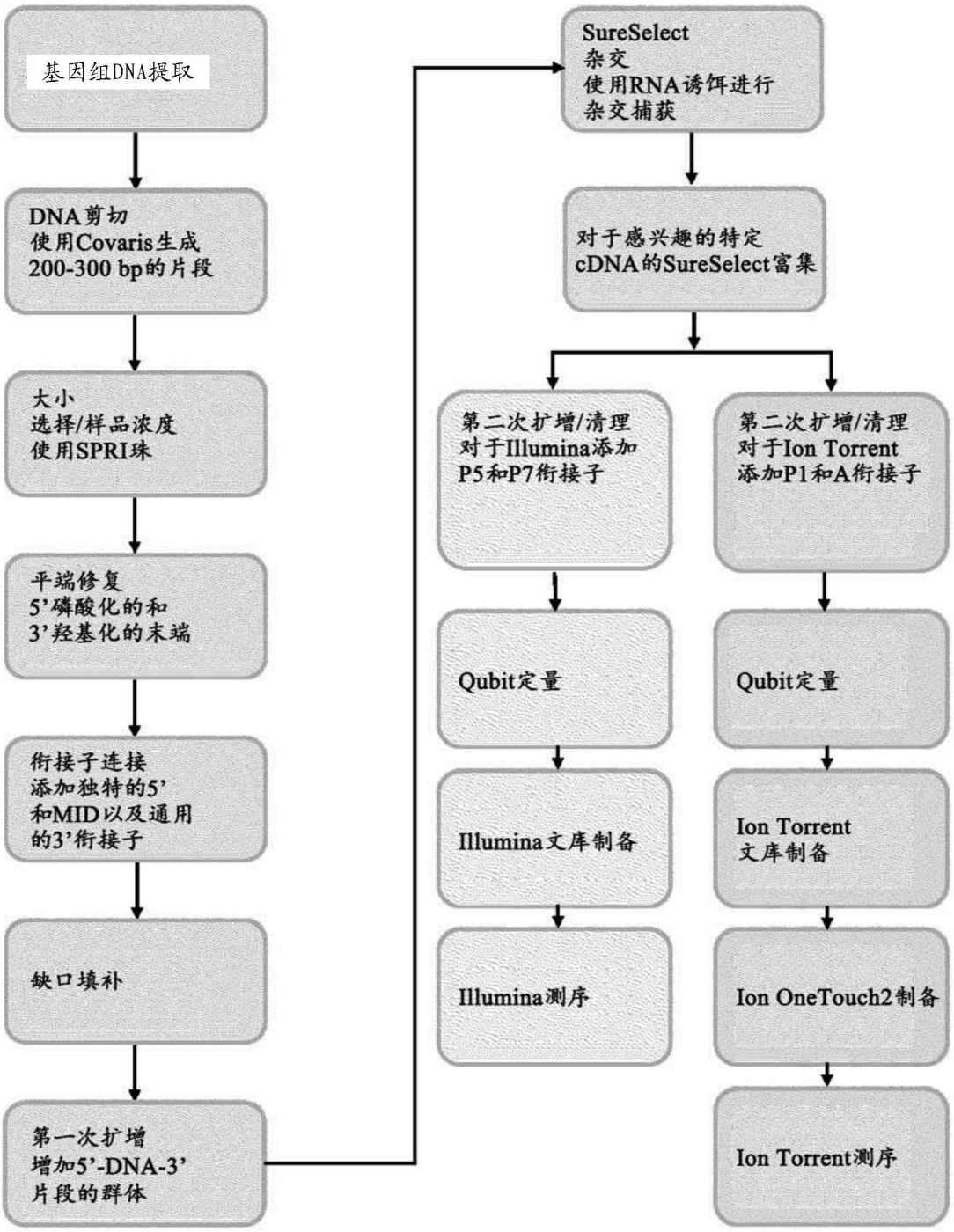
基因组DNA提取
DNA剪切
使用Covaris生成
200-300 bp的片段
大小
选择/样品浓度
使用SPRI珠
平端修复
5'磷酸化的和
3'羟基化的末端
衔接子连接
添加独特的5'
和MID以及通用
的3'衔接子
缺口填补
第一次扩增
增加5'-DNA-3'
片段的群体
SureSelect
杂交
使用RNA诱饵进行
杂交捕获
对于感兴趣的特定
cDNA的SureSelect富集
第二次扩增/清理
对于Illumina添加
P5和P7衔接子
第二次扩增/清理
对于Ion Torrent
添加P1和A衔接子
Qubit定量
Qubit定量
Illumina文库制备
Ion Torrent
文库制备
Illumina测序
Ion OneTouch2制备
Ion Torrent测序

图1

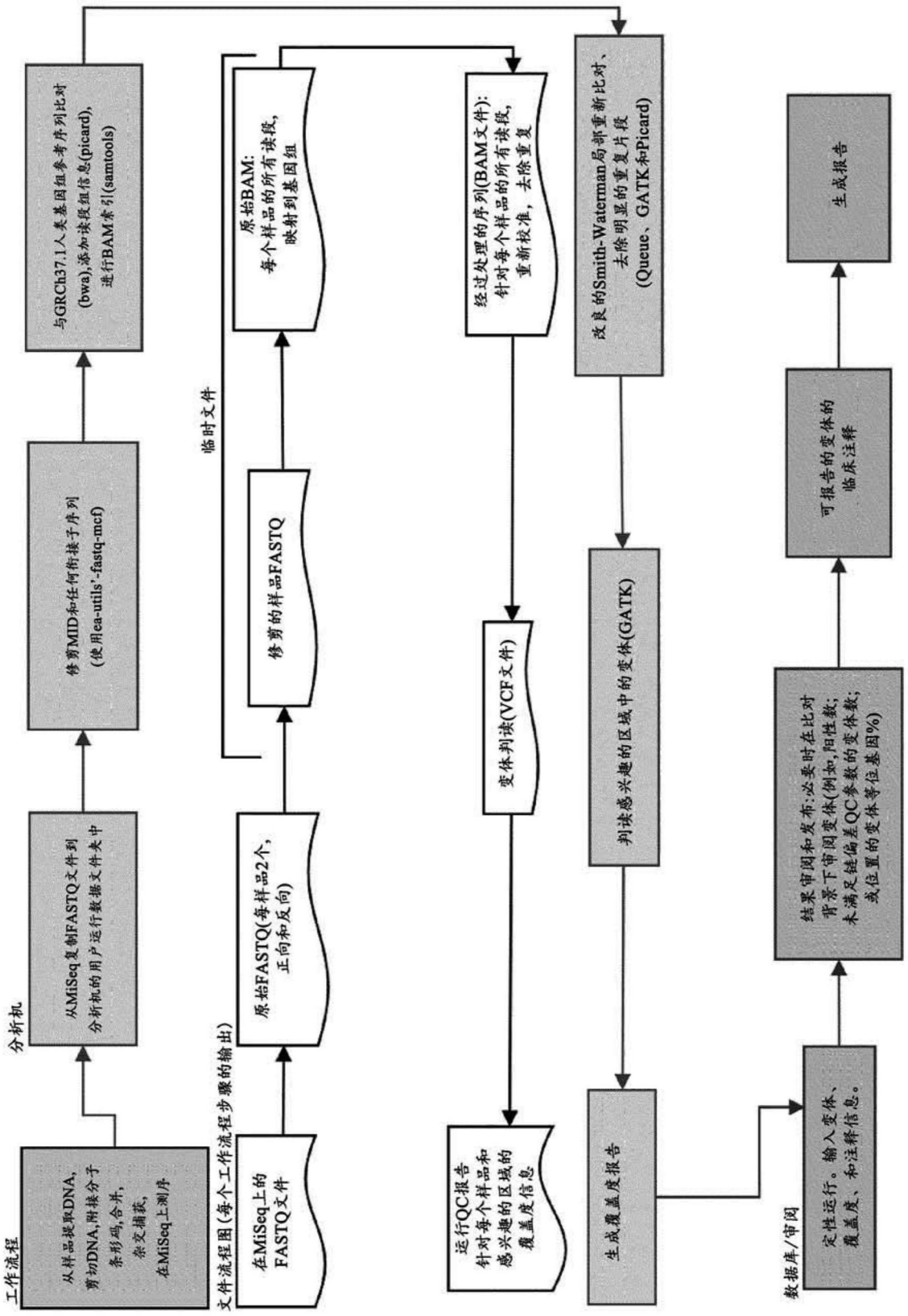
工作流程
从样品提取DNA,
剪切DNA,附接分子
条形码,合并,
杂交捕获,
在MiSeq上测序
分析机
从MiSeq复制FASTQ文件到
分析机的用户运行数据文件夹中
修剪MID和任何衔接子序列
(使用ea-utils'-fastq-mcf)
与GRCh37.1人类基因组参考序列比对
(bwa),添加读段组信息(picard),
进行BAM索引(samtools)
文件流程图(每个工作流程步骤的输出)
在MiSeq上的
FASTQ文件
原始FASTQ(每样品2个,
正向和反向)
修剪的样品FASTQ
临时文件
原始BAM:
每个样品的所有读段,
映射到基因组
运行QC报告
针对每个样品和
感兴趣的区域的
覆盖度信息
变体判读(VCF文件)
经过处理的序列(BAM文件):
针对每个样品的所有读段,
重新校准,去除重复
生成覆盖度报告
判读感兴趣的区域中的变体(GATK)
改良的Smith-Waterman局部重新比对、
去除明显的重复片段
(Queue、GATK和Picard)
数据库/审阅
定性运行。输入变体、
覆盖度、和注释信息。
结果审阅和发布:必要时在比对
背景下审阅变体(例如,阳性数;
未满足链偏差QC参数的变体数;
或位置的变体等位基因%)
可报告的变体的
临床注释
生成报告

图2

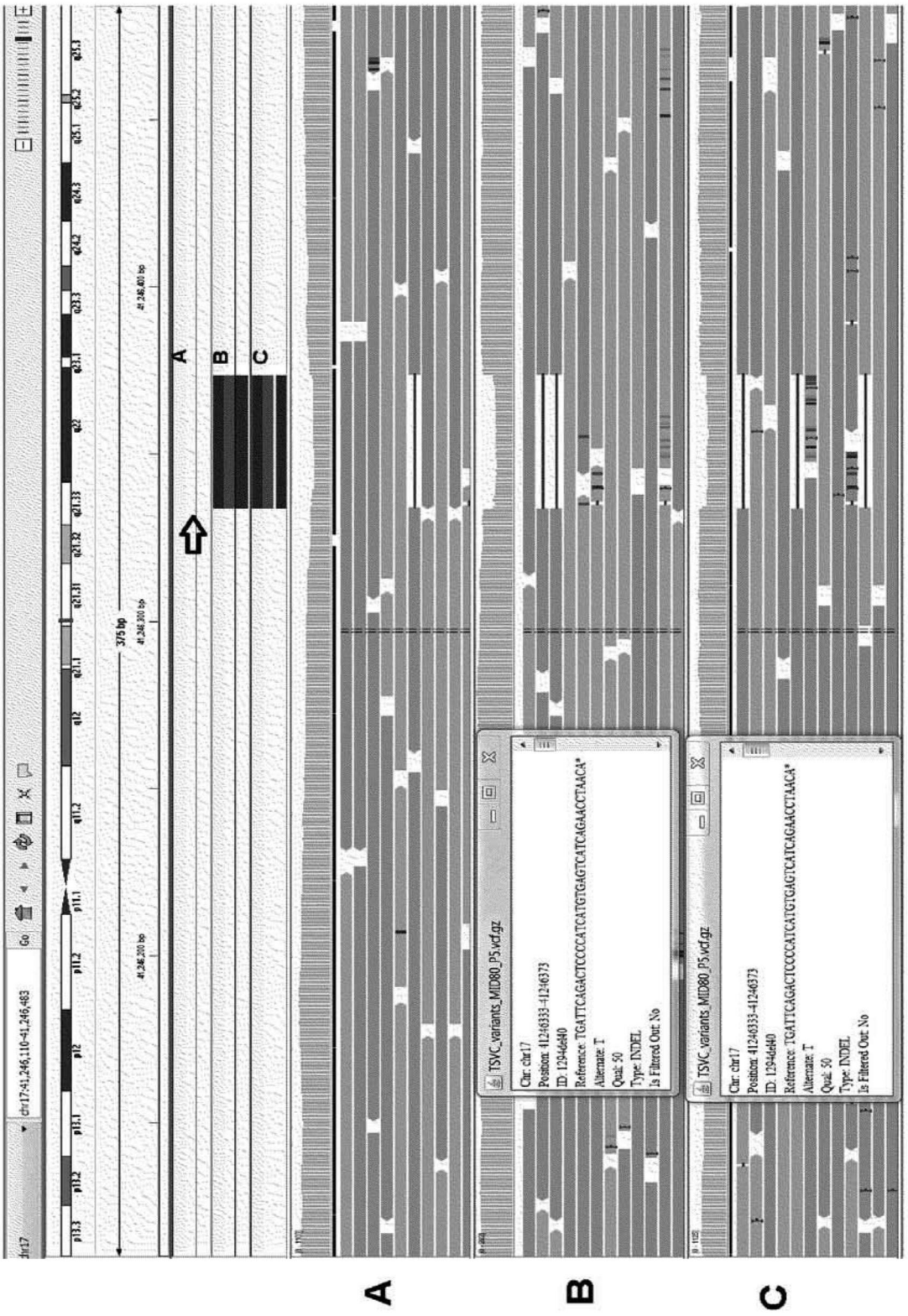
A
B
C
TSVC_variants_MID80_P5.vcf.gz
Char: chr17
Position: 41246333-41246373
ID: 12946d40
Reference: TGATTCAGACTCCCCATCATGTGAGTCATCAGAACCTAACA*
Alternate: T
Qual: 50
Type: INDEL
Is Filtered Out: No

图3

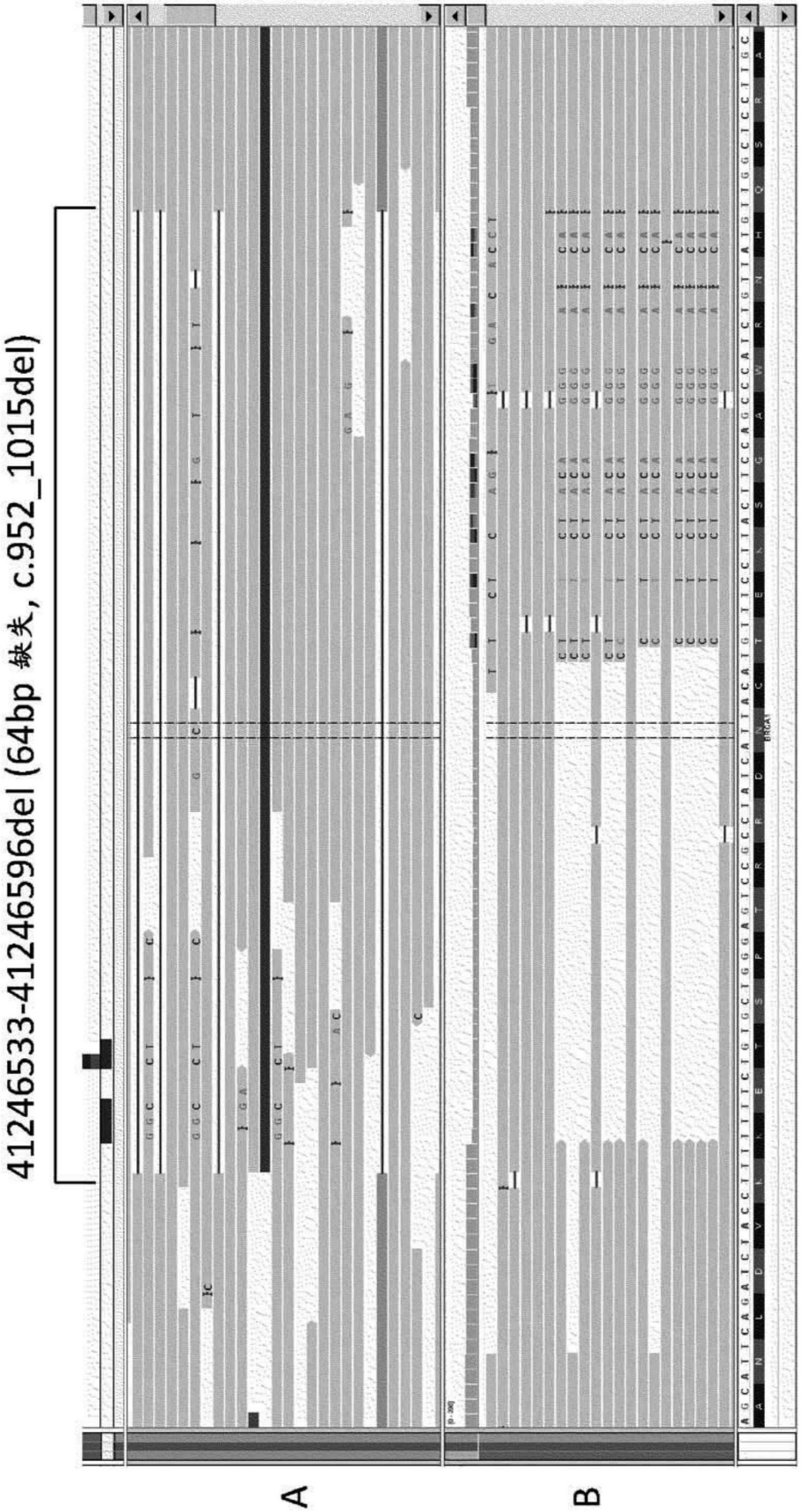

图4

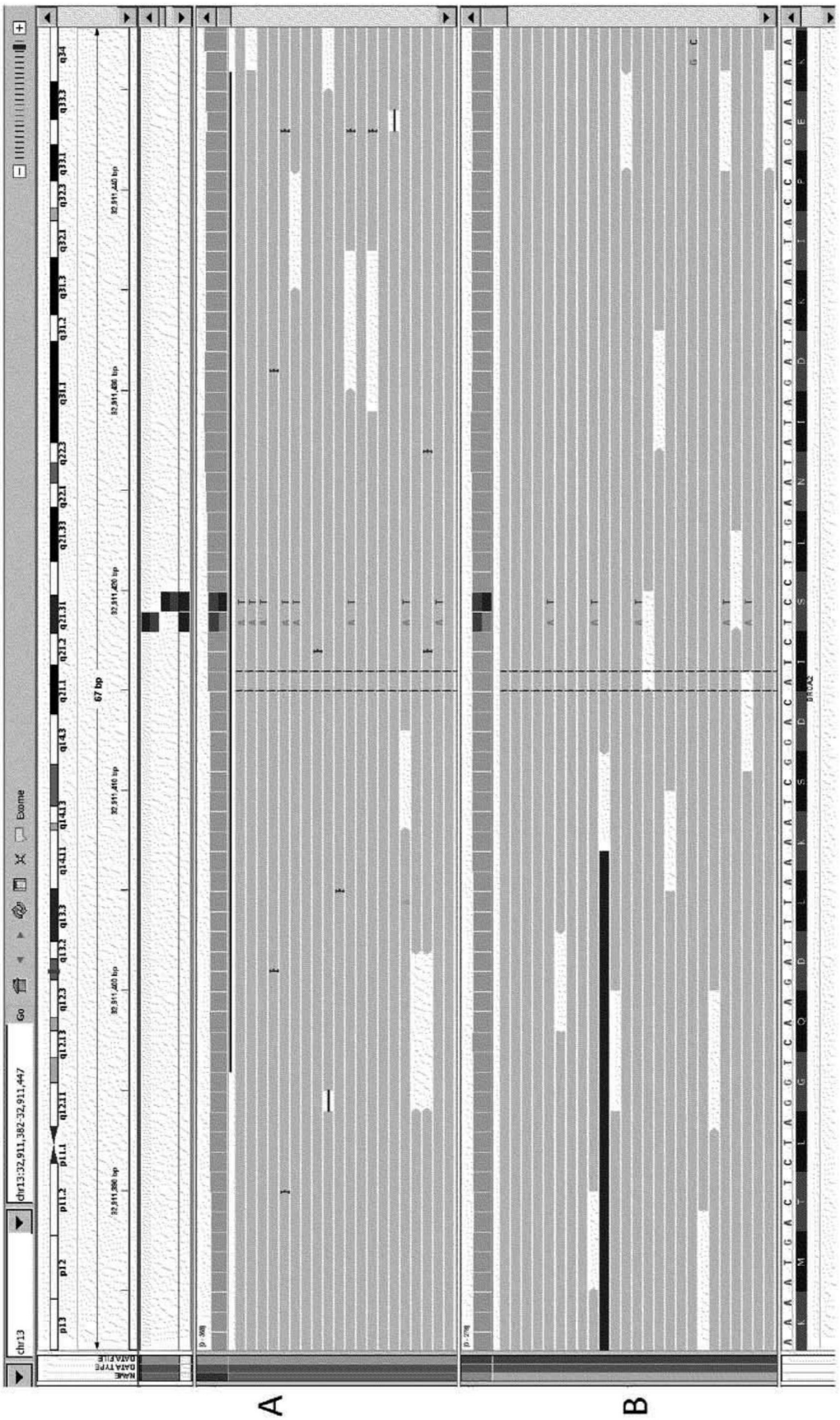

图5

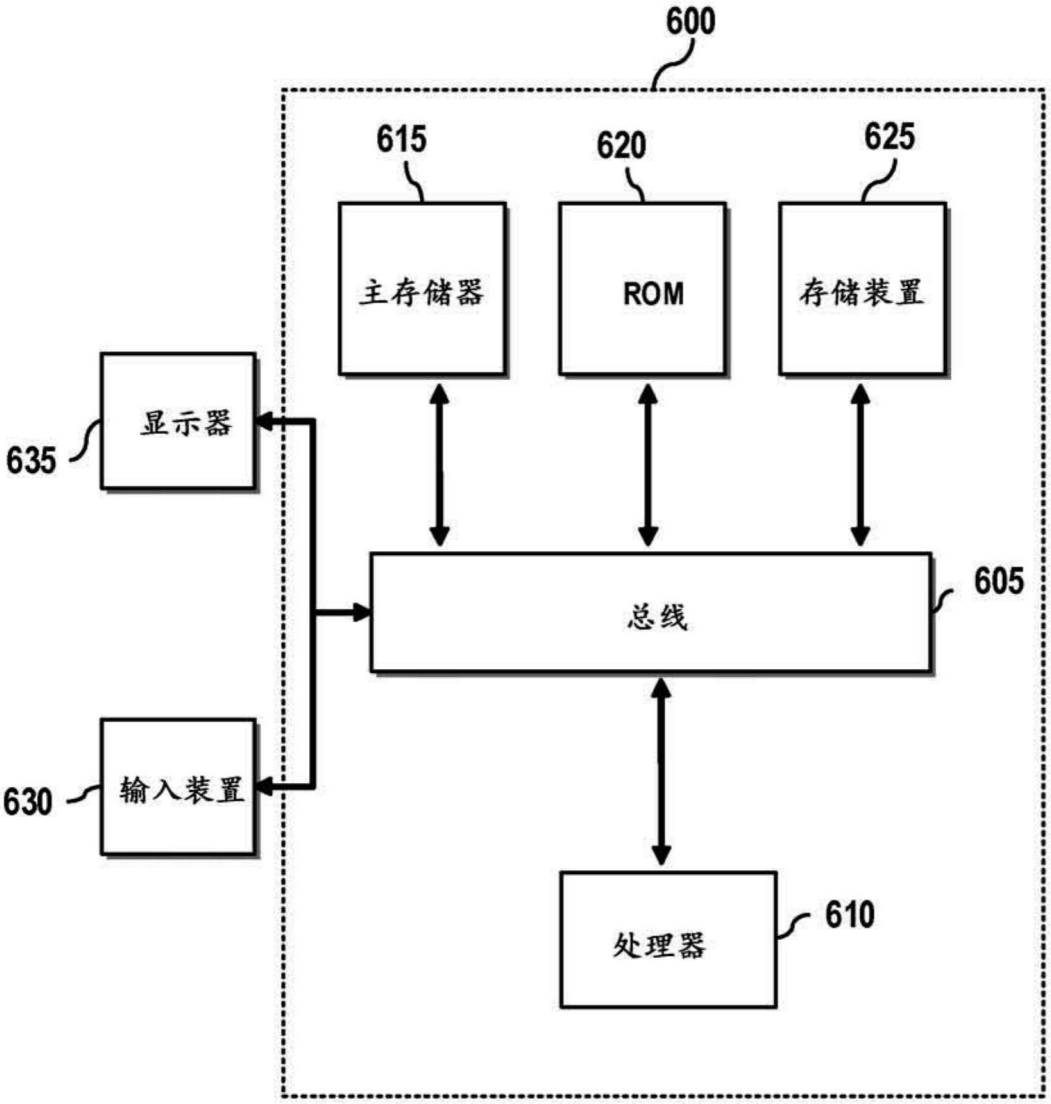

图6