(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
2023 年 3 月 9 日 (09.03.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/030233 A1

(51) 国际专利分类号:
C12Q 1/6869 (2018.01) *G16B 20/30* (2019.01)
G16B 20/20 (2019.01) *G16B 30/10* (2019.01)

(21) 国际申请号: PCT/CN2022/115447

(22) 国际申请日: 2022 年 8 月 29 日 (29.08.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111002171.4 2021年8月30日 (30.08.2021) CN
202111095132.3 2021年9月17日 (17.09.2021) CN

(71) 申请人: 广州燃石医学检验所有限公司 (GUANGZHOU BURNING ROCK DX CO., LTD.) [CN/CN]; 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。

(72) 发明人: 刘成林 (LIU, Chenglin); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 赵宇 (ZHAO, Yu); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 旷婷 (KUANG, Ting); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 张之宏 (ZHANG, Zhihong); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 张振 (ZHANG, Zhen); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 张光亮 (ZHANG, Guangliang); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 张周 (ZHANG, Zhou); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 揣少坤 (CHUAI, Shaokun); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。 汉雨生 (HAN, Yusheng); 中国广东省广州市黄埔区广州国际生物岛螺旋四路7号第二层201、202单元, Guangdong 510300 (CN)。

(74) 代理人: 上海巅石知识产权代理事务所（普通合伙） (SHANGHAI DIANSHI PARTNERS, P.C.); 中国上海市浦东新区盛夏路608号2号103室/张琤、孙宋暄, Shanghai 201210 (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, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI,

(54) **Title:** COPY NUMBER VARIATION DETECTION METHOD AND APPLICATION THEREOF

(54) 发明名称: 一种拷贝数变异的检测方法及其应用

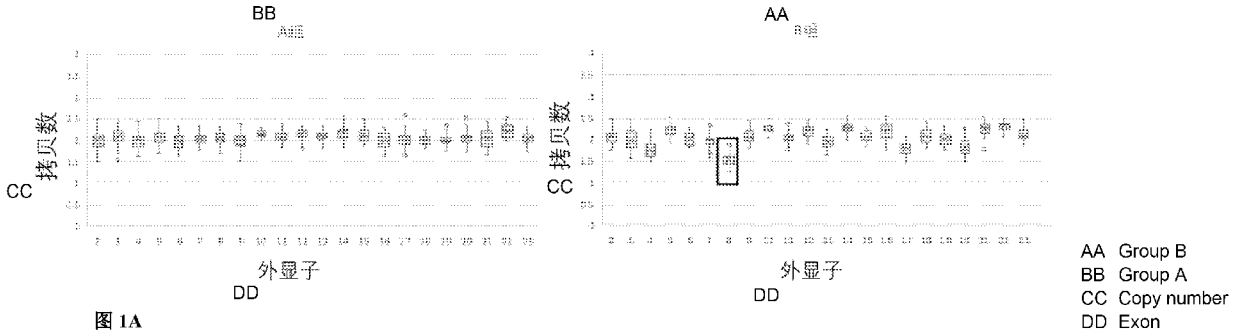


图 1A

(57) **Abstract:** Provided are a copy number variation detection method and an application thereof. Also provided is a copy number status analysis method, comprising: dividing a target interval of a sample to be tested into several window areas, acquiring sequencing data of a control window area in a sample group to be tested, and determining the copy number status of a target gene of the sample to be tested based on the sequencing data of the control window area.

(57) 摘要: 提供了一种拷贝数变异的检测方法及其应用。还提供了一种拷贝数状态分析方法，包含将待测样本的目标区间划分为若干个窗口区域，获取待测样本组中的对照窗口区域的测序数据，基于对照窗口区域的测序数据，确定待测样本的目标基因的拷贝数状态。

WO 2023/030233 A1

[见续页]

GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, ST, 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, KM, 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))。

一种拷贝数变异的检测方法及其应用

技术领域

本申请涉及生物信息领域，具体的涉及一种拷贝数变异的检测方法及其应用。

背景技术

拷贝数变异（CNV）是人类基因组中常见的变异类型之一。拷贝数变异包括基因拷贝数的扩增和缺失两种变异类型。对于基因拷贝数变异的检测可以用于监测受试者的基因组状态的，也可以用于发现特定疾病与某些基因组变异的关联性关系。例如，基因发生拷贝数变异可能导致多种常见遗传病，如 BRCA1/2 基因缺失可能导致遗传性乳腺癌的风险；基因发生拷贝数变异可能影响肿瘤的发生和发展，如 HER2 基因扩增不仅与肿瘤的发生发展相关，还是一个重要的临床治疗监测及预后指标，并且是肿瘤靶向治疗的一个重要靶点。因此，拷贝数变异的检测可以在受试者的基因组状态监测、全基因组关联研究、遗传病的预防以及肿瘤的精准治疗起着至关重要的作用。例如，携带某些特定的拷贝数变异的受试者与普通人群相比，可以具有更高的患有疾病（例如肿瘤）的终生风险。因此，拷贝数变异检测方法可以用于筛选出具有较高风险的受试者，该受试者可以接受疾病的个体化监测，从而可以达到早诊早治的目的。

传统的拷贝数变异检测方法如 ddPCR（微滴式数字 PCR）、MLPA（多重连接探针扩增技术）、FISH（荧光原位杂交技术）等，一次只能检测一个或少数几个基因的拷贝数状态，或者只能检测特定基因的拷贝数状态，不能对基因组进行全局分析，具有通量低、成本高的特点。目前，基于高通量技术检测拷贝数变异的方法有很多，但是不同的检测方法结果存在着很大的差异，检测的敏感性和特异性具有一定的局限。一方面，高通量测序技术在建库及测序过程中存在着一定的批次效应和技术误差，另一方面，由于肿瘤样本的复杂性，导致拷贝数检测结果的稳定性存在很大的挑战，使得基于高通量测序技术检测拷贝数变异检测在精准医疗领域有很大的难度。本领域急需一种可以降低批次效应、误差和/或提高拷贝数检测结果稳定性的分析方法。

发明内容

本申请的目的就是针对上述现有技术不足，提供一种针对基因拷贝数异常检测的方法。该方法至少可以降低批次效应、误差和/或提高拷贝数检测结果稳定性，对检测拷贝数异常相

关的驱动事件以及解读肿瘤基因组演化信息有着重要的意义。本申请提供了一种拷贝数变异的检测方法及其应用。

一方面，本申请提供了一种拷贝数状态分析方法，所述方法包含，将待测样本的目标区间划分为若干个窗口区域，获取待测样本组中的对照窗口区域的测序数据，基于所述对照窗口区域的测序数据，确定所述待测样本的目标基因的拷贝数状态；任选地，所述对照窗口区域包含覆盖波动水平低的窗口区域。

一方面，本申请提供了一种拷贝数状态分析装置，包含以下模块：接收模块，用于获取待测样本组的测序数据；确定模块，用于确定待测样本中的目标基因；判断模块，用于根据所述待测样本组的测序数据确定所述待测样本中的目标基因的拷贝数状态。

一方面，本申请提供了一种储存介质，其记载可以运行本申请所述的方法的程序。

本领域技术人员能够从下文的详细描述中容易地洞察到本申请的其它方面和优势。下文的详细描述中仅显示和描述了本申请的示例性实施方式。如本领域技术人员将认识到的，本申请的内容使得本领域技术人员能够对所公开的具体实施方式进行改动而不脱离本申请所涉及发明的精神和范围。相应地，本申请的附图和说明书中的描述仅仅是示例性的，而非为限制性的。

附图说明

本申请所涉及的发明的具体特征如所附权利要求书所显示。通过参考下文中详细描述的示例性实施方式和附图能够更好地理解本申请所涉及发明的特点和优势。对附图简要说明如下：

图 1A-1B 显示的是基于构建参考基线方法和本申请方法的检测结果图。每个箱线图表示 30 例样本的 *BRCA1* 基因外显子拷贝数值分布。A 组和 B 组分别代表不同批次探针捕获。图 1A 显示基于构建参考基线方法计算 *BRCA1* 基因各外显子的拷贝数分布。图 1B 显示本申请方法计算 *BRCA1* 基因各外显子的拷贝数分布。

图 2A-2B 显示的是基于构建参考基线方法和本申请方法对 NGS 建库方法差异的样本检测结果图。横坐标：染色体坐标；纵坐标：评估的拷贝数（CN）值。图 2A 显示基于构建参考基线方法检测拷贝数变异结果。图 2B 显示本申请方法检测拷贝数变异结果。方框表示检出的拷贝数变异。

图 3A-3B 显示的是基于筛选稳定窗口不同设定阈值后的样本检测结果。横坐标：染色体坐标；纵坐标：评估的拷贝数（CN）值。图 3A 显示设定阈值为 0.05 的样本检测拷贝数变异结果。图 3B 显示设定阈值为 0.15 的样本检测拷贝数变异结果。方框表示检出的拷贝数变异。

图 4A-4J 显示的是 10 例模拟阳性拷贝数变异样本构建批次基线后的检测结果。横坐标：染色体坐标；纵坐标：评估的拷贝数（CN）值。图中方框表示表示检出的拷贝数变异。

图 5A-5F 显示的是，本申请对于模拟样本的检测结果部分数据的拷贝数分布图示例。

图 6A-6C 显示的是，本申请对于标准样本的检测结果部分数据的拷贝数分布图示例。

图 7A-7C 显示的是，本申请对于真实样本的检测结果部分数据的拷贝数分布图示例。

图 8A-8F 显示的是，本申请对于标准样本 1 使用不同的基线检测结果的拷贝数分布图示例。

具体实施方式

以下由特定的具体实施例说明本申请发明的实施方式，熟悉此技术的人士可由本说明书所公开的内容容易地了解本申请发明的其他优点及效果。

在本申请中，术语“二代基因测序”、高通量测序”或“下一代测序”通常是指第二代高通量测序技术及之后发展的更高通量的测序方法。下一代测序平台包括但不限于已有的 Illumina 等测序平台。随着测序技术的不断发展，本领域技术人员能够理解的是还可以采用其他方法的测序方法和装置用于本方法。例如，二代基因测序可以具有高灵敏度、通量大、测序深度高、或低成本的优势。根据发展历史、 影响力、测序原理和技术不同等，主要有以下几种：大规模平行签名测序（Massively Parallel Signature Sequencing，MPSS）、聚合酶克隆（Polony Sequencing）、454 焦磷酸测序（454 pyro sequencing）、Illumina（Solexa）sequencing、离子半导体测序（Ion semi conductor sequencing）、DNA 纳米球测序（DNA nano-ball sequencing）、Complete Genomics 的 DNA 纳米阵列与组合探针锚定连接测序法等。所述二代基因测序可以使对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序（deep sequencing）。例如，本申请的方法同样可以应用于一代基因测序、二代基因测序、三代基因测序或单分子测序（SMS）。在本申请中，术语“数据库”通常是指相关数据的有组织实体，而不管数据或有组织实体的表示方式。例如，所述相关数据的有组织实体可以采取表、映射、网格、分组、数据报、文件、文档、列表的形式或任何其他形式。在本申请中，所述数据库可以包括以计算机可存取的方式来收集并保存的任何数据。

在本申请中，术语“计算模块”通常是指用于计算的功能模块。所述计算模块可以根据输入值计算输出值或得到结论或结果，例如计算模块可以主要是用于计算输出值。计算模块可以是有形的，例如电子计算机的处理器、带有处理器的计算机或电子设备或计算机网络，也可以是存储在电子介质上的一段程序、命令行或软件包。

在本申请中，术语“处理模块”通常是指用于数据处理的功能模块。所述处理模块可以

根据将输入值处理为有统计学意义的数据，例如可以是用于输入值的数据的分类。处理模块可以是有形的，例如用于存储数据的电子或磁介质，以及电子计算机的处理器、带有处理器的计算机或电子设备或计算机网络，也可以是存储在电子介质上的一段程序、命令行或软件包。

在本申请中，术语“判断模块”通常是指用于获得相关判断结果的功能模块。在本申请中，所述判断模块可以根据输入值计算输出值或得到结论或结果，例如判断模块可以主要是用于得到结论或结果。判断模块可以是有形的，例如电子计算机的处理器、带有处理器的计算机或电子设备或计算机网络，也可以是存储在电子介质上的一段程序、命令行或软件包。

在本申请中，术语“样品获得模块”通常是指用于获得受试者的所述样本的功能模块。例如，所述样品获得模块可以包括用以获得所述样本（例如组织样本、血液样本、唾液、胸腔积液、腹腔积液、脑脊液等）所需的试剂和/或仪器。例如，可以包括采血针、采血管和/或血液样本运输箱。例如，本申请的装置可以不含或包含 1 个或以上的所述样品获得模块，并可以可选地具有输出本申请所述的样本的测量值的功能。

在本申请中，术语“接收模块”通常是指用于获得所述样本中所述测量值的功能模块。在本申请中，所述接收模块可以输入本申请所述样本（例如组织样本、血液样本、唾液、胸腔积液、腹腔积液、脑脊液等）。在本申请中，所述接收模块可以输入本申请所述样本（例如组织样本、血液样本、唾液、胸腔积液、腹腔积液、脑脊液等）的测量值。所述接收模块可以对所述样本的状态进行检测。例如，所述数据接收模块可以可选地对所述样本进行本申请所述的基因测序（例如二代基因测序）。例如，所述数据接收模块可以可选地包括用以进行所述基因测序所需的试剂和/或仪器。所述数据接收模块可以可选地检测出测序深度、测序读长计数或拷贝数。

在本申请中，术语“拷贝数变异”通常是指目标区间、目标基因或目标基因中的目标区间的拷贝数的扩增或缺失。例如，本申请提供的拷贝数变异分析方法可以是用于治疗或诊断目的。例如，本申请提供的拷贝数变异分析方法可以是用于非治疗或诊断目的，例如通过测序结果判定是否存在着拷贝数变异现象。

在本申请中，术语“滑窗法”通常是指一种窗口区域的划分方法，例如，可以按照相同或不同的窗口区域长度，将全长区域的划分为多个窗口。例如，可以按照相同或不同步长，将全长区域的划分为多个窗口。例如，可以按照相同的窗口区域长度，并可以按照相同的步长，将全长区域的划分为多个窗口。

在本申请中，术语“质量合格的样品”通常是通过质量控制标准的样本。例如，质量合格的样品可以是指平均测序深度、最小测序深度和/或覆盖均一度合格的样本。例如，平均测序深度合格可以是指平均测序深度为约 100x 或以上的样本。例如，最小测序深度合格的样本可以是指最小测序深度为约 30x 或以上的样本。例如，覆盖均一度合格的样本可以是指大于或等于样本平均测序深度的 20%的碱基数量占样本中总碱基数量的百分比为约 90%或以上的样本。

在本申请中，术语“不合格目标区间”通常是指测序质量低的区间。例如，不合格的区间可以不适合用作拷贝数变异的分析。例如，不合格的区间可以不适合用作参考或基线的构建。在一些情况下，筛除不合格区间可以提高检测结果的准确性；在另一些情况下，不筛除不合格区间也可以得到具有一定准确性的检测结果。例如，不合格的区间可以是指测序深度低的区间；例如，不合格的区间可以是指该区间在不同的样本中变化大的区间。

在本申请中，术语“捕获效率低区间”通常是指不容易被用到的探针捕获的区间。例如，某区间的序列存在特定的序列组合时，可能难以被核酸探针捕获。例如，捕获效率低区间可以是指测序深度低的区间。例如，捕获效率低区间可以是指测序读长计数为约 5 或更低的区间。

在本申请中，术语“不稳定区间”通常是指不同样本中测序结果差异大的区间。例如，可以是在同一样本中多次测序结果差异大的区间。例如，可以是在不同样本的同一批次中测序结果差异大的区间。例如，可以是在不同样本的不同批次中测序结果差异大的区间。例如，可以是在不同参考样本中测序结果差异大的区间。例如，确定不稳定区间的方法可以是计算某个区间在不同样本的测序深度的标准差与均值的比值，并确定该比值是否大于某阈值的方法，如阈值可以是 0.8，或本领域技术人员根据实际测序情况进行调整。

在本申请中，术语“待测样本” 通常是指需要进行检测，并判定该样本上的一个或者多个基因区域是否存在有拷贝数变异的样本。例如，待测样本或其数据可以在进行检测之前预先储存在储存器中。

在本申请中，术语 “人类参考基因组” 通常是指可以在基因测序中发挥参照功能的人类基因组。所述人类参考基因组的信息可以参考 UCSC（University of California，Santa Cruz）。所述人类参考基因组可以有不同的版本，例如，可以为 hg19、GRCH37 或 ensembl 75。

在本申请中，术语“GC 含量”通常是指在基因序列(碱基序列)中鸟嘌呤 G 和胞嘧啶 C 所占该序列汇总所有核苷酸的比率。

在本申请中，术语“靶向测序面板”或“panel”通常是指一组/一套检测对象。例如，在测序的过程中，对一个或多个目标区间通过设计一个或多个探针进行捕获并检测，这样的一个或多个探针可以组成靶向测序面板。例如，靶向测序面板可以针对目标基因、目标区间或感兴趣的区域进行任意设计，例如也可以针对若干个外显子区域。例如，探针可以是指在研究中与目标区间的寡聚核苷酸或靶核酸互补的寡聚核苷酸。例如，目标区间是设计探针时所针对的区间。

在本申请中，术语“测序深度”通常是指特定区域（例如特定基因、特定区间、特定碱基）被检测的次数。测序深度可以是指通过测序检测的一段碱基序列。例如，通过将测序深度比对到人类参考基因组，并可选地去重，可以确定和统计特定基因、特定区间或特定碱基位置上测序读长的数量，作为测序深度。在一些情况下，测序深度可以与测序深度相关。例如，测序深度可以受到拷贝数状态的影响。

在本申请中，术语“测序数据”通常是指测序后获得的短序列的数据。例如，测序数据包含测序短序列（测序读长）的碱基序列、测序读长的数目等。

在本申请中，术语“测序偏差量”通常是指不同区间产生测序数据偏差量。例如，区间的序列的特殊排列方式或者碱基比例可以对该区间的测序读长计数产生影响。例如，当区间含有较高或较低 GC 含量时，该区间的测序读长计数相对于 GC 含量接近 50%的区间可以有偏差。

在本申请中，术语“分布相似程度”可以是指两组数据的分布相似程度。例如，本申请中分布相似程度可以是指在一个或多个区间上参考样本组与待测样本的测序读长计数的相似程度。

在本申请中，术语“统计距离”可以是指两组数据的数据值的距离。例如，本申请中统计距离可以是指在一个或多个区间上参考样本组与待测样本的测序读长计数的差值的统计量。例如，可以通过欧式距离，切比雪夫距离，马氏距离等计算统计距离。

在本申请中，术语“统计值”可以是指根据样本的数据值计算的分析数值。例如，本申请的统计值可以是指平均值、方差、标准差、中位数值、众数值等。本领域技术人员根据实际情况选择一个或多个统计值用于分析数据。

在本申请中，术语“概率分布”通常是指随机变量取值的分布规律。例如，根据随机变量所属类型的不同，概率分布可以取不同的表现形式。例如，正态分布可以作为一种随机变量的概率分布。

在本申请中，术语“平滑化”通常是指使一个以上的本申请所述的差值之间的偏差减小的数据处理的方法。例如，可以是指将散点数据拟合到平滑的线的方法。例如，可以通过局部加权回归的方法分析和进行平滑化处理。例如，平滑化处理后，可以通过消除某一变量（例如 GC 含量）对于样本测序数据的固有影响，而消除或减弱该变量（例如 GC 含量））对于样本测序数据造成的偏差。例如，所述平滑化处理可以包括获得一定数量的本申请所述差值的平均值。例如，所述平滑化处理可以包括根据一定的间隔长度，选择不同长度所对应的数据值，计算不同数据值的差值。例如，所述平滑化处理可以包括将一定长度范围内，所述差值的累加值再除以间隔长度以获得比值。例如，所述比值可以被认为是该长度范围的所述差值的平均差值。

在本申请中，术语“回归”通常是指变量之间关系的统计分析方法。例如，本申请通过回归分析，可以得出样本测序数据与某一变量（例如 GC 含量）的线性或非线性关系。例如，可以通过局部加权回归得到样本的测序数据与某一变量（例如 GC 含量）的关系，并通过这一关系调整/矫正该样本的测序数据。例如，本申请的矫正可以是指根据样本的测序数据与某一变量的关系，对该样本的测序数据进行处理以消除或减弱该变量对该样本的测序数据造成的偏差。

在本申请中，术语“局部加权回归”通常是指在输入变量和目标变量的回归分析中，在局部引入权重的回归分析方法。例如，局部加权回归可以通过（loess(X~Y)）的算法根据 Y 对 X 进行局部加权回归分析和处理。

在本申请中，术语“降噪化”通常是指对于数据中噪声数据的去除或减少。例如，根据噪声数据一般表现为高频信号，可以通过变换分析、主成分分析算法、奇异值分解和/或高斯滤波等方法对含噪音的数据提取有用信号进行降噪化处理。

在本申请中，术语“聚类分析”通常是指把相似的对象通过分类的方法分成不同的组别，使得在同一个组别中的成员对象都有相似的一些属性。

在本申请中，术语“K 均值聚类”通常是指一种聚类分析的方法。例如，通过 K 均值聚类可以根据 K 个聚类中心将一组数据分类为若干个（K 个）类别的聚类分析的方法，每个数据距离其最近聚类中心的距离和最小。

在本申请中，术语“变换分析”通常是指一种分析数据的方法。例如，变换分析可以通过将数据的原始分布变换为容易求解或处理的变换域的分布，对数据进行分析和用于进一步的处理。例如，变换分析可以包含离散小波变换。

在本申请中，术语“离散小波变换”通常是指对基本小波的尺度和平移进行离散化。例如，离散小波变换可以作为一种降噪化的方法。

在本申请中，术语“标准化”或“归一化”通常是指一种对数据变换的方式。例如，标准化可以是指将不同组的数据变换到某一固定范围的处理。例如，标准化可以是指将不同组的数据变换到同一中位值的处理。例如，本申请的标准化可以是指将不同样本的测序数据变换为的中位值相近水平的数据的处理方式。

在本申请中，术语“显著性检验”通常是指判断样本与假设分布之间的差异是否显著的方式。例如，通过显著性检验可以判断待测样本的拷贝数变异是否属于显著的差异。

在本申请中，术语“正态性概率分布”通常是指随机变量的概率分布。例如，通过正态性概率分布以及正态性概率分布密度函数可以确定一个随机变量发生的概率。例如，可以基于参考样本组的测序数据，通过正态性概率分布，确认待测样本的目标区间拷贝数变异的存在的概率。

在本申请中，术语“格拉布斯检验法”通常是指判断和/或筛除异常值的方法。例如，通过判断某一数值是否符合整体的分布范围，可以确定该数值是否属于异常值。

在本申请中，术语“T 检验”通常是指一种有学生 t 分布的统计假设检验的方式。例如，通过 T 检验可以确认待测样本的某一目标基因的拷贝数变异具有显著性。

在本申请中，术语“包含”通常是指包括明确指定的特征，但不排除其他要素。

在本申请中，术语“约”通常是指在指定数值以上或以下 0.5%-10%的范围内变动，例如在指定数值以上或以下 0.5%、1%、1.5%、2%、2.5%、3%、3.5%、4%、4.5%、5%、5.5%、6%、6.5%、7%、7.5%、8%、8.5%、9%、9.5%、或 10%的范围内变动。

发明详述

一方面，本申请提供一种拷贝数状态的分析方法

一方面，本申请提供一种拷贝数状态的分析方法，可以包含获取待测样本组的测序数据；确定待测样本中的目标基因；根据所述待测样本组的测序数据确定所述待测样本中的目标基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态的分析方法，可以包含以下步骤：

（S1）将所述目标基因所在区域划分为若干个窗口区域，获取所述待测样本组中的对照窗口区域的测序数据；

（S2）基于所述对照窗口区域的测序数据，确定所述待测样本的目标基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态的分析方法，可以包含以下步骤：

（S1）将所述目标基因所在区域划分为若干个窗口区域，获取所述待测样本组中的对照窗口区域的测序数据；

（S2）基于所述对照窗口区域的测序数据，确定所述待测样本的目标基因的拷贝数状态，将所述质量合格的样品的窗口区域按照所述覆盖波动水平从低到高排序，所述对照窗口区域可以包含覆盖波动水平前 4 个或更多所述窗口，可以基于所述质量合格的样品的窗口区域的测序数据的绝对离差中位数与中位数的比值，确定所述覆盖波动水平，或者所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值可以为约 0.15 或更小。

一方面，本申请提供一种拷贝数状态的分析方法，可以包含以下步骤：

步骤（S1-1）：获取所述待测样本组中所有样本的窗口区域的测序数据；步骤（S1-2）：获取所述待测样本组中质量合格的样本，所述质量合格的样本可以包含平均测序深度、最小测序深度和/或覆盖均一度合格的样本；步骤（S1-3）：使所述待测样本组中所有样本的窗口区域的所述测序数据标准化；

（S2）基于所述对照窗口区域的测序数据，确定所述待测样本的目标基因的拷贝数状态，将所述质量合格的样品的窗口区域按照所述覆盖波动水平从低到高排序，所述对照窗口区域可以包含覆盖波动水平前 4 个或更多所述窗口，可以基于所述质量合格的样品的窗口区域的测序数据的绝对离差中位数与中位数的比值，确定所述覆盖波动水平，或者所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值可以为约 0.15 或更小；步骤（S2-1）：基于所述对照窗口区域的测序数据，确定归一化系数；步骤（S2-2）：基于所述归一化系数，确定待测样本的每一个窗口区域的拷贝数；步骤（S2-3）：基于待测样本的每一个窗口区域的测序数据以及相应窗口区域的待测样本组中其它样本的测序数据，确定待测样本的拷贝数变异显著性。

例如，所述测序数据可以包含测序深度。例如，所述拷贝数状态可以包含拷贝数扩增和/或缺失。例如，所述拷贝数状态可以包含外显子拷贝数状态。

例如，所述待测样本组可以包含约 10 个或更多的样本。例如，所述待测样本组可以包含约 10 个或更多、约 12 个或更多、约 15 个或更多、约 20 个或更多、约 25 个或更多、约 50 个或更多、或约 100 个或更多的样本。例如，本申请可以不要求同一批次有较多的样本数量。例如，所述待测样本组可以包含约 10 个或更少、约 12 个或更少、约 15 个或更少、约 20 个或更少、约 25 个或更少、或约 50 个或更少的样本。例如，本申请的拷贝数状态分析方法，可以对待测样本的拷贝数变异水平具有较高的容忍度。例如，包含约 30%拷贝数变异的样本，

可以通过本申请的分析方法进行评估。例如，包含 10%或更低、15%或更低、20%或更低、25%或更低、或 30%或更低拷贝数变异的样本，可以通过本申请的分析方法进行评估。例如，本申请的样本来源可以是任意的含有核酸的样本，例如组织、血液、唾液、胸腔积液、腹腔积液、脑脊液等。

例如，本申请方法所述步骤（S1）还可以包含步骤（S1-1）：获取所述待测样本组中所有样本的窗口区域的测序数据。例如，本申请的基因测序可以包含任选的高通量测序方法或模块、装置。例如，测序可以选自下组：Solexa 测序技术、454 测序技术、SOLiD 测序技术、Complete Genomics 测序方法和半导体(Ion Torrent)测序技术及其对应的装置。

例如，本申请方法所述步骤（S1-1）可以包含通过滑窗法将所述目标基因所在区域划分为所述窗口区域。例如，所述划窗法的步长可以为约 24 个碱基。例如，所述窗口区域的长度可以为约 120 个碱基。

例如，本申请方法所述步骤（S1-1）可以包含获取每一个所述窗口区域去除重复测序片段后的平均测序深度。

例如，本申请方法所述步骤（S1）还可以包含步骤（S1-2）：获取所述待测样本组中质量合格的样本，所述质量合格的样本可以包含平均测序深度、最小测序深度和/或覆盖均一度合格的样本。例如，所述平均测序深度合格的样本包含平均测序深度可以为约 100x 或以上的样本。例如，所述最小测序深度合格的样本包含最小测序深度可以为约 30x 或以上的样本。例如，质量合格的各个阈值可以根据测序情况进行调整。

例如，所述覆盖均一度可以与所述样本每个碱基的测序深度有关。例如，所述覆盖均一度可以通过大于或等于样本平均测序深度的 20%的碱基数量占样本中总碱基数量的百分比计算。例如，所述覆盖均一度合格的样本可以包含覆盖均一度为约 90%或以上的样本。例如，所述覆盖均一度合格的样本可以包含覆盖均一度为约 90%或以上的样本。例如，所述覆盖均一度合格的样本可以包含覆盖均一度为约 90%或以上、约 92%或以上、约 95%或以上、约 97%或以上、或约 99%或以上的样本。

例如，所述待测样本组中质量合格的样本数量可以为 10 个或更多。

例如，本申请方法所述步骤（S1）还可以包含步骤（S1-3）：使所述待测样本组中所有样本的窗口区域的所述测序数据标准化。

例如，所述标准化可以包含基于所述样本所有窗口区域的平均测序深度对所述样本的每一个窗口区域的所述测序数据标准化、和/或基于所述样本的每一个窗口区域的 GC 含量对所述样本的每一个窗口区域的所述测序数据标准化。

例如，所述标准化可以包含使所述样本的每一个窗口区域上所述测序数据除以所述样本的所有窗口区域上所述测序数据的总和，再乘以因子。例如，所述因子可以根据所有区间的大小进行设置。例如，所述因子可以任选为 1E+07。例如，所述因子可以任选为 1E+100、1E+20、1E+10、1E+09、1E+08、1E+07、1E+06、1E+05、1E+04、1E+03、或 1E+02。

例如，所述标准化可以包含基于 GC 含量，通过回归的方法使所述样本的每一个窗口区域的所述测序数据标准化。例如，所述回归可以包含局部加权回归。

例如，所述对照窗口区域可以包含覆盖波动水平低的窗口区域。

例如，可以基于所述质量合格的样品的窗口区域的测序数据统计值，确定所述覆盖波动水平。例如，可以基于所述质量合格的样品的窗口区域的测序数据的离差，确定所述覆盖波动水平。例如，可以基于所述质量合格的样品的窗口区域的测序数据的绝对离差中位数和/或中位数，确定所述覆盖波动水平。例如，可以基于所述质量合格的样品的窗口区域的测序数据的绝对离差中位数与中位数的比值，确定所述覆盖波动水平。

例如，将所述质量合格的样品的窗口区域按照所述覆盖波动水平从低到高排序，所述对照窗口区域可以包含覆盖波动水平前 2 个或更多所述窗口。

例如，将所述质量合格的样品的窗口区域按照所述覆盖波动水平从低到高排序，所述对照窗口区域可以包含覆盖波动水平前 4 个或更多所述窗口。

例如，所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值可以为约 0.15 或更小。例如，所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值可以为约 0.15 或更小、约 0.14 或更小、约 0.13 或更小、约 0.12 或更小、约 0.11 或更小、约 0.10 或更小、约 0.09 或更小、约 0.08 或更小、约 0.07 或更小、约 0.06 或更小、或约 0.05 或更小。例如，所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值可以为约 0.05 至约 0.15、约 0.07 至约 0.15、约 0.10 至约 0.15、约 0.12 至约 0.15、约 0.05 至约 0.12、约 0.07 至约 0.12、约 0.10 至约 0.12、约 0.05 至约 0.10、约 0.07 至约 0.10、或约 0.05 至约 0.07。

例如，本申请所述步骤（S2）还可以包含步骤（S2-1）：基于所述对照窗口区域的测序数据，确定归一化系数。

例如，可以通过计算所述对照窗口区域的所有所述质量合格的样品的测序数据平均值，确定所述归一化系数。

例如，在确定所述归一化系数前，可以筛除对照窗口区域的异常样本的覆盖水平值。例如，所述异常覆盖水平值可以为每个所述对照窗口区域通过异常值分析方法判断为异常样本

的覆盖水平值。例如，所述异常值分析方法可以包含格拉布斯检验法（Grubbs test）。例如，每个窗口可以包含了批次内质量合格样本在该窗口的覆盖水平值，然后可以采用格拉布斯检验法检验判断这些覆盖水平值是否含有异常值，如果含有，可以将异常值去除。然后对剩余的覆盖水平值，可以继续重复采用格拉布斯检验法检验判断是否有异常，直至无异常值出现。例如，也可以当剩余覆盖水平值数量少于质量合格样本数的 60%或更少、50%或更少、或 40%或更少时，停止去除异常值，可以将剩余数值都用于确定所述归一化系数。

例如，筛除所述异常样本后剩余的样本数量可以为筛除前样本数量的 40%或更多、70%或更多、80%或更多、90%或更多、95%或更多、或 99%或更多。

例如，本申请所述步骤（S2）还可以包含步骤（S2-2）：基于所述归一化系数，确定待测样本的每一个窗口区域的拷贝数。

例如，本申请所述步骤（S2-2）可以包含基于所述归一化系数，通过使所述待测样本的每一个窗口区域的所述测序数据归一化的方法，确定待测样本的每一个窗口区域的所述拷贝数。

例如，所述归一化方法可以包含将所述窗口区域的待测样本的测序数据除以所述窗口区域的归一化系数，乘以倍性。例如，对于男性 X 染色体来说，所述倍性可以为 1。当受试者为多倍体是，所述倍性可以根据具体情况进行调整。例如，所述倍性可以为 2。

例如，本申请所述步骤（S2）还可以包含步骤（S2-3）：基于待测样本的每一个窗口区域的测序数据以及相应窗口区域的待测样本组中其它样本的测序数据，确定待测样本的拷贝数变异显著性。

例如，本申请所述步骤（S2-3）可以包含基于待测样本的每一个窗口区域的所述拷贝数，确定拷贝数变异候选区域。

例如，可以通过区域分割的方法，确定所述拷贝数变异候选区域。例如，所述区域分割可以包含通过循环二元分割算法，确定所述拷贝数变异候选区域的前后端点。

例如，本申请所述步骤（S2-3）可以包含基于待测样本的所述拷贝数变异候选区域中窗口区域的测序数据以及相应窗口区域的待测样本组中其它样本的测序数据，确定拷贝数变异的显著性。例如，可以通过显著性检验的方法，确定所述拷贝数变异的显著性。例如，所述显著性检验可以包含 T 检验。

另一方面，本申请还提供了一种拷贝数状态分析装置，可以包含以下模块：接收模块，用于获取待测样本组的测序数据；确定模块，用于确定待测样本中的目标基因；判断模块，用于根据所述待测样本组的测序数据确定所述待测样本中的目标基因的拷贝数状态。

例如，本申请的拷贝数状态分析装置，其中的模块可以被配置为基于存储在所述储存介

质中的程序执行以实现本申请所述的拷贝数状态分析方法。

拷贝数状态分析方法

一方面，本申请提供一种拷贝数状态分析方法，可以包含以下步骤：（S1）获取待测样本的测序数据和/或多个参考样本的测序数据；（S2）将所述参考样本分为两个或以上参考样本组；（S3）确定与所述待测样本最接近的参考样本组；（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态分析装置，可以包含以下模块：（M1）接收模块，用于获取待测样本的测序数据和/或多个参考样本的测序数据；（M2）处理模块，用于将所述参考样本分为两个或以上参考样本组；（M3）计算模块，用于确定与所述待测样本最接近的参考样本组；（M4）判断模块，用于基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态分析方法，可以包含以下步骤：

（S1）获取待测样本的测序数据和/或多个参考样本的测序数据；步骤（S1-1）：通过基因测序获取所述待测样本和/或所述参考样本的所述测序数据；步骤（S1-2）：矫正所述待测样本和/或参考样本的所述测序数据；

（S2）将所述参考样本分为两个或以上参考样本组；步骤（S2-1）：使所述参考样本分组；步骤（S2-2）：确认所述参考样本组的所述测序数据的统计值；

（S3）确定与所述待测样本最接近的参考样本组；

（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态步骤（S4-1）：确定所述待测样本的目标基因在所述目标区间 i 上的拷贝数 CN_i；步骤（S4-3）：确定所述待测样本在目标基因上的拷贝数 CN_g；步骤（S4-4）：确定在所述目标区间上待测样本的拷贝数变异的存在的概率；步骤（S4-5）：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio；步骤（S4-6）：确定在所述目标基因上所述待测样本的拷贝数变异的存在的统计检验参数；通过以下内容确定所述待测样本的目标基因的拷贝数状态：当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值。

一方面，本申请提供一种拷贝数状态分析装置，可以包含实现本申请拷贝数状态分析方

法的模块。

一方面，本申请提供一种拷贝数状态分析方法，可以包含以下步骤：

（S1）获取待测样本的测序数据和/或多个参考样本的测序数据；步骤（S1-1）：通过基因测序获取所述待测样本和/或所述参考样本的所述测序数据；步骤（S1-2）：矫正所述待测样本和/或参考样本的所述测序数据；

（S2）将所述参考样本分为两个或以上参考样本组；步骤（S2-1）：使所述参考样本分组；步骤（S2-2）：确认所述参考样本组的所述测序数据的统计值；

（S3）确定与所述待测样本最接近的参考样本组；

（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态步骤（S4-1）：确定所述待测样本的目标基因在所述目标区间 i 上的拷贝数 CN_i；步骤（S4-2）：使所述待测样本的所述目标区间上的拷贝数降噪化；步骤（S4-3）：确定所述待测样本在目标基因上的拷贝数 CN_g；步骤（S4-4）：确定在所述目标区间上待测样本的拷贝数变异的存在的概率；步骤（S4-5）：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio；步骤（S4-6）：确定在所述目标基因上所述待测样本的拷贝数变异的存在的统计检验参数；通过以下内容确定所述待测样本的目标基因的拷贝数状态：当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值。

一方面，本申请提供一种拷贝数状态分析装置，可以包含实现本申请拷贝数状态分析方法的模块。

例如，本申请的测序数据可以包含测序读长计数。例如，本申请的测序数据可以包含目标基因或目标区间上测序读长（read）的数量。

例如，本申请的所述步骤（S1）或模块（M1）可以包含步骤（S1-1）或模块（M1-1）：通过基因测序获取所述待测样本和/或所述参考样本的所述测序数据。例如，所述基因测序可以包含二代基因测序（NGS）。例如，本申请的基因测序可以包含任选的高通量测序方法或模块、装置。例如，测序可以选自下组：Solexa 测序技术、454 测序技术、SOLiD 测序技术、Complete Genomics 测序方法和半导体(Ion Torrent)测序技术及其对应的装置。

例如，所述待测样本和/或所述参考样本可以包含含有核酸的样本。例如，本申请的样

本来源可以是任意的含有核酸的样本，例如组织、血液、唾液、胸腔积液、腹腔积液、脑脊液等。

例如，所述步骤（S1-1）或模块（M1-1）可以包含获取所述待测样本和/或参考样本在目标区间中每个碱基的所述测序数据。

例如，所述目标区间可以包含靶向测序面板序列对应的区间。例如，所述目标区间的长度可以为约 20 至约 500 个碱基量。例如，所述目标区间的长度可以为约 20 至约 500 个碱基量、约 50 至约 500 个碱基量、约 100 至约 500 个碱基量、约 200 至约 500 个碱基量、约 20 至约 200 个碱基量、约 50 至约 200 个碱基量、约 100 至约 200 个碱基量、约 20 至约 100 个碱基量、约 50 至约 100 个碱基量、或约 20 至约 50 个碱基量。

例如，所述目标区间的数量可以至少为约 100 个。例如，所述目标区间的数量可以至少为约 100 个、至少为约 200 个、至少为约 500 个、至少为约 1000 个、或至少为约 10000 个。

例如，所述步骤（S1）或模块（M1）可以包含步骤（S1-2）或模块（M1-2）：矫正所述待测样本和/或参考样本的所述测序数据。例如，本申请的方法也可以不包含步骤（S1-2）或仅包含步骤（S1-2）的部分步骤。例如，本申请的装置也可以不包含模块（M1-2）或仅包含模块（M1-2）的部分模块。例如，本申请的方法步骤（S1-2）的以下步骤顺序可以是任意的：使所述待测样本和/或参考样本的所述测序数据标准化、使所述待测样本和/或参考样本的所述测序数据平滑化和筛除 GC 含量异常的所述目标区间。例如，本申请的装置的模块（M1-2）的以下模块顺序可以是任意的：使所述待测样本和/或参考样本的所述测序数据标准化、使所述待测样本和/或参考样本的所述测序数据平滑化和筛除 GC 含量异常的所述目标区间。

例如，所述步骤（S1-2）或模块（M1-2）可以包含：使所述待测样本和/或参考样本的所述测序数据标准化或均一化。例如，所述标准化或均一化可以包含使所述目标区间上所述测序数据除以所述目标区间对应的样本的所有目标区间上所述测序数据的总和，再乘以因子。例如，所述因子可以根据所有区间的大小进行设置。例如，所述因子可以任选为 1E+07。例如，所述因子可以任选为 1E+100、1E+20、1E+10、1E+09、1E+08、1E+07、1E+06、1E+05、1E+04、1E+03、或 1E+02。

例如，所述步骤（S1-2）或模块（M1-2）可以包含：使所述待测样本和/或参考样本的所述测序数据平滑化。例如，所述平滑化可以包含基于测序偏差量，通过回归的方法或记载该程序的装置使所述待测样本和/或参考样本的所述测序数据平滑化。例如，所述回归可以

包含局部加权回归。

例如，所述测序偏差量可以包含所述目标区间上覆盖的探针数。

例如，所述测序偏差量可以包含所述目标区间的 GC 含量。

例如，所述步骤（S1-2）或模块（M1-2）可以任选地包含：筛除 GC 含量异常的所述目标区间。

例如，所述 GC 含量异常的所述目标区间可以包含 GC 含量为约 25%或更低的所述目标区间和/或 GC 含量为约 75%或更高的所述目标区间。

例如，所述步骤（S2）或模块（M2）可以包含步骤（S2-1）或模块（M2-1）：使所述参考样本分组。例如，所述参考样本可以源自所述待测样本，也可以源自所述待测样本以外的样本。例如，可以将待测样本中的一部分划分出去作为参考样本。例如，所述参考样本可以进行更新，例如每次分析新样品的测序数据后，可以将该新样品的数据添加到已有的数据库中，并进行数据库再次建立处理。

例如，所述分组可以包含基于所述目标区间的所述测序数据使所述参考样本分组。

例如，所述分组可以包含通过聚类分析的方法或记载该程序的装置使所述参考样本分组。

例如，所述聚类分析的方法可以包含 K 均值聚类、层次聚类、密度聚类、网格聚类、概率模型聚类、或神经网络模型聚类等。例如，所述聚类分析的方法或记载该程序的装置可以包含任意的聚类、分类和分组的方法或记载该程序的装置。

例如，所述参考样本的数量可以为约 30 个或更多。例如，所述参考样本的数量可以为约 50 个或更多。例如所述参考样本的数量可以为约 30 个或更多、约 40 个或更多、约 50 个或更多、约 60 个或更多、约 70 个或更多、约 80 个或更多、约 90 个或更多、约 100 个或更多、约 200 个或更多、约 300 个或更多、约 400 个或更多、约 500 个或更多、或约 1000 个或更多。

例如，所述分组可以包含分为约 2 个组或更多。例如，当全部参考样本的测序数据较为近似时，可以仅分为 1 组。例如，所述分组包含分为约 2 个或更多、约 3 个或更多、约 4 个或更多、约 5 个或更多、约 6 或更多、约 7 个或更多、约 8 个或更多、约 9 个或更多、约 10 个或更多、约 20 个或更多、约 30 个或更多、约 40 个或更多、约 50 个或更多、约 60 个或更多、约 70 个或更多、约 80 个或更多、约 90 个或更多、或约 100 个或更多。

例如，每组中所述参考样本的数量可以为约 30 个或更多。例如，每组中的参考样本的数量可以为约 30 个或更多、约 40 个或更多、约 50 个或更多、约 60 个或更多、约 70 个或

更多、约 80 个或更多、约 90 个或更多、约 100 个或更多、约 200 个或更多、约 300 个或更多、约 400 个或更多、约 500 个或更多、或约 1000 个或更多。

例如，所述步骤（S2）或模块（M2）可以包含步骤（S2-2）或模块（M2-2）：确认所述参考样本组的所述测序数据的统计值。例如，所述参考样本组的测序数据的统计值可以作为各个候选基线。例如，所述确认统计值可以包含计算在所述目标区间上每组中所述参考样本的均值和/或标准差。

例如，所述步骤（S2）或模块（M2）可以包含步骤（S2-3）或模块（M2-3）：筛除所述参考样本中不合格目标区间。例如，所述不合格目标区间可以包含捕获效率低区间和/或不稳定区间。

例如，所述不合格目标区间可以包含测序读长计数为约 5 或更低的目标区间。例如，所述不合格目标区间可以包含测序读长计数为约 30 或更低、约 20 或更低、约 10 或更低、约 5 或更低、约 4 或更低、约 3 或更低、约 2 或更低、约 1 或更低、或约 0 或更低的目标区间。

例如，所述不合格目标区间可以包含变异系数为约 0.8 或更高的目标区间，所述变异系数为所述目标区间上每组中所述参考样本的所述测序数据的标准差与均值的比值。例如，所述不合格目标区间可以包含变异系数为约 0.8 或更高、约 0.9 或更高或约 1.0 或更高。例如，捕获效率低区间和/或不稳定区间的各个阈值可以根据测序情况进行调整。

例如，所述步骤（S3）或模块（M3）可以包含确认所述待测样本与所述参考样本组的相似度。

例如，所述确认相似度可以包含基于在所述目标区间上所述参考样本组与所述待测样本的所述测序数据，确认所述参考样本组与所述待测样本的分布相似程度。

例如，所述相似度可以包含在所述目标区间上所述参考样本组与所述待测样本的所述测序数据的近似程度。

例如，所述确认相似度可以包含通过计算统计距离、相似度算法的方法或记载该程序的装置，确认所述参考样本组与所述待测样本的分布相似程度。例如，所述统计距离可以包含所述目标区间上所述参考样本组与所述待测样本的所述测序数据的差值的统计值。例如，所述统计距离可以包含所述目标区间上所述参考样本组与所述待测样本的所述测序数据的差值的绝对值的统计值。例如，所述统计距离可以包含所述目标区间上所述参考样本组与所述待测样本的所述测序数据的差值的绝对值的 p 次方的统计值，所述 p 为 1 或更大。例如，所述统计值可以包含求和值。

例如，所述相似度高可以包含在所述目标区间上所述参考样本组与所述待测样本的所述统计距离短。

例如，所述统计距离可以包含闵可夫斯基距离。例如，所述统计距离可以包含欧式距离、曼哈顿距离、切比雪夫距离、闵可夫斯基距离（当 p=1 时，为曼哈顿距离；当 p=2 时，为欧式距离；当 p 接近于无穷大是，为切比雪夫距离）等。例如，所述相似度算法可以包含余弦相似度、皮尔森相关系数、斯皮尔曼相关系数、对数似然相似度、交叉熵等。

例如，所述待测样本的目标基因的拷贝数状态可以包含述待测样本的目标基因的拷贝数变异的存在和/或数量。

例如，所述拷贝数变异可以包含拷贝数的扩增和/或缺失。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-1）或模块（M4-1）：确定所述待测样本的目标基因在所述目标区间 i 上的拷贝数 CN_i。

例如，所述确定所述 CN_i 可以包含，使所述待测样本的目标基因所述目标区间上的测序数据的均值除以所述与待测样本最接近的参考样本组在相应目标区间上的测序数据的均值，再乘以倍性，得到所述 CN_i。

例如，所述倍性可以为 2。例如，对于男性 X 染色体来说，所述倍性可以为 1。当受试者为多倍体是，所述倍性可以根据具体情况进行调整。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-2）或模块（M4-2）：使所述待测样本的所述目标区间上的拷贝数降噪化。

例如，所述降噪化可以包含通过变换分析、主成分分析算法、奇异值分解和/或高斯滤波的方法或记载该程序的装置使所述待测样本的所述目标区间上的拷贝数降噪化。

例如，所述降噪化可以包含通过离散小波变换的方法或记载该程序的装置使所述待测样本的所述目标区间上的拷贝数降噪化。例如，所述降噪化可以包含通过变换分析、主成分分析算法、奇异值分解和/或高斯滤波等方法或记载该程序的装置使所述待测样本的所述目标区间上的拷贝数降噪化。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-3）或模块（M4-3）：确定所述待测样本在目标基因上的拷贝数 CN_g。

例如，所述目标基因可以包含待确定的拷贝数变异发生的基因。

例如，所述目标基因可以包含选自以下组基因：ABL1、ABL2、ABRAXAS1、ACVR1、ACVR1B、AKT1、AKT2、AKT3、ALK、ALOX12B、AMER1、APC、AR、ARAF、ARFRP1、ARID1A、ARID1B、ARID2、ARID5B、ASXL1、ASXL2、ASXL3、

ATG5、ATM、ATR、ATRX、AURKA、AURKB、AXIN1、AXIN2、AXL、B2M、BAP1、BARD1、BBC3、BCL10、BCL2、BCL2L1、BCL2L11、BCL2L2、BCL6、BCOR、BCORL1、BIRC3、BLM、BMPR1A、BRAF、BRCA1、BRCA2、BRD4、BRD7、BRINP3、BRIP1、BTG1、BTG2、BTK、CALR、CARD11、CASP8、CBFB、CBL、CCND1、CCND2、CCND3、CCNE1、CD274、CD28、CD58、CD74、CD79A、CD79B、CDC73、CDH1、CDH18、CDK12、CDK4、CDK6、CDK8、CDKN1A、CDKN1B、CDKN1C、CDKN2A、CDKN2B、CDKN2C、CEBPA、CENPA、CHD1、CHD2、CHD4、CHD8、CHEK1、CHEK2、CIC、CIITA、CREBBP、CRKL、CRLF2、CRYBG1、CSF1R、CSF3R、CSMD1、CSMD3、CTCF、CTLA4、CTNNA1、CTNNB1、CUL3、CUL4A、CXCR4、CYLD、CYP17A1、CYP2D6、DAXX、DCUN1D1、DDR1、DDR2、DDX3X、DICER1、DIS3、DNAJB1、DNMT1、DNMT3A、DNMT3B、DOT1L、DPYD、DTX1、DUSP22、EED、EGFR、EIF1AX、EIF4E、EMSY、EP300、EPCAM、EPHA2、EPHA3、EPHA5、EPHA7、EPHB1、EPHB4、ERBB2、ERBB3、ERBB4、ERCC1、ERCC2、ERCC3、ERCC4、ERCC5、ERG、ERRFI1、ESR1、ETV4、ETV5、ETV6、EWSR1、EZH2、EZR、FANCA、FANCC、FANCD2、FANCE、FANCF、FANCG、FANCI、FANCL、FANCM、FAS、FAT1、FAT3、FBXW7、FGF10、FGF12、FGF14、FGF19、FGF23、FGF3、FGF4、FGF6、FGF7、FGFR1、FGFR2、FGFR3、FGFR4、FH、FLCN、FLT1、FLT3、FLT4、FOXA1、FOXL2、FOXO1、FOXO3、FOXP1、FRS2、FUBP1、FYN、GABRA6、GALNT12、GATA1、GATA2、GATA3、GATA4、GATA6、GEN1、GID4、GLI1、GNA11、GNA13、GNAQ、GNAS、GPS2、GREM1、GRIN2A、GRM3、GSK3B、H3F3A、H3F3B、H3F3C、HDAC1、HDAC2、HGF、HIST1H1C、HIST1H2BD、HIST1H3A、HIST1H3B、HIST1H3C、HIST1H3D、HIST1H3E、HIST1H3G、HIST1H3H、HIST1H3I、HIST1H3J、HIST2H3D、HIST3H3、HLA-A、HLA-B、HLA-C、HNF1A、HOXB13、HRAS、HSD3B1、HSP90AA1、ICOSLG、ID3、IDH1、IDH2、IFNGR1、IGF1、IGF1R、IGF2、IGHD、IGHJ、IGHV、IKBKE、IKZF1、IL10、IL7R、INHA、INHBA、INPP4A、INPP4B、INSR、IRF2、IRF4、IRS1、IRS2、ITK、ITPKB、JAK1、JAK2、JAK3、JUN、KAT6A、KDM5A、KDM5C、KDM6A、KDR、KEAP1、KEL、KIR2DL4、KIR3DL2、KIT、KLF4、KLHL6、KLRC1、KLRC2、KLRK1、KMT2A、KMT2C、KMT2D、KRAS、LATS1、LATS2、LMO1、LRP1B、LTK、LYN、MAF、MAGI2、MALT1、MAP2K1、MAP2K2、MAP2K4、MAP3K1、MAP3K13、MAP3K14、

MAPK1、MAPK3、MAX、MCL1、MDC1、MDM2、MDM4、MED12、MEF2B、MEN1、MERTK、MET、MFHAS1、MGA、MIR21、MITF、MKNK1、MLH1、MLH3、MPL、MRE11、MSH2、MSH3、MSH6、MST1、MST1R、MTAP、MTOR、MUTYH、MYC、MYCL、MYCN、MYD88、MYOD1、NAV3、NBN、NCOA3、NCOR1、NCOR2、NEGR1、NF1、NF2、NFE2L2、NFKBIA、NKX2-1、NKX3-1、NOTCH1、NOTCH2、NOTCH3、NOTCH4、NPM1、NRAS、NRG1、NSD1、NSD2、NSD3、NT5C2、NTHL1、NTRK1、NTRK2、NTRK3、NUP93、NUTM1、P2RY8、PAK1、PAK3、PAK5、PALB2、PALLD、PARP1、PARP2、PARP3、PAX5、PBRM1、PCDH11X、PDCD1、PDCD1LG2、PDGFRA、PDGFRB、PDK1、PGR、PHOX2B、PIK3C2B、PIK3C2G、PIK3C3、PIK3CA、PIK3CB、PIK3CD、PIK3CG、PIK3R1、PIK3R2、PIK3R3、PIM1、PLCG2、PLK2、PMS1、PMS2、PNRC1、POLD1、POLE、POM121L12、PPARG、PPM1D、PPP2R1A、PPP2R2A、PPP6C、PRDM1、PREX2、PRKAR1A、PRKCI、PRKDC、PRKN、PTCH1、PTEN、PTPN11、PTPRD、PTPRO、PTPRS、PTPRT、QKI、RAB35、RAC1、RAD21、RAD50、RAD51、RAD51B、RAD51C、RAD51D、RAD52、RAD54L、RAF1、RARA、RASA1、RB1、RBM10、RECQL4、REL、RET、RHEB、RHOA、RICTOR、RIT1、RNF43、ROS1、RPA1、RPS6KA4、RPS6KB2、RPTOR、RSPO2、RUNX1、RUNX1T1、SDC4、SDHA、SDHAF2、SDHB、SDHC、SDHD、SETD2、SF3B1、SGK1、SH2B3、SH2D1A、SHQ1、SLC34A2、SLIT2、SLX4、SMAD2、SMAD3、SMAD4、SMARCA4、SMARCB1、SMARCD1、SMO、SNCAIP、SOCS1、SOX10、SOX17、SOX2、SOX9、SPEN、SPI1、SPOP、SPTA1、SRC、SRSF2、STAG2、STAT3、STAT4、STAT5A、STAT5B、STAT6、STK11、STK40、SUFU、SYK、TAF1、TBX21、TBX3、TCF3、TCF7L2、TEK、TENT5C、TERC、TERT、TET1、TET2、TGFBR1、TGFBR2、TIPARP、TMEM127、TMPRSS2、TNFAIP3、TNFRSF14、TOP1、TOP2A、TP53、TP63、TP73、TRAF2、TRAF3、TRAF7、TRIM58、TRPC5、TSC1、TSC2、TSHR、TYRO3、U2AF1、UGT1A1、VEGFA、VEGFB、VEGFC、VHL、WISP3、WRN、WT1、XIAP、XPO1、XRCC2、XRCC3、YAP1、YES1、ZAP70、ZBTB16、ZBTB2、ZNF217、ZNF703 和 ZNRF3。例如，所述目标基因可以包含选自以下组基因：ALK（转录本编号可以是 NM_004304.4）、ERBB2（转录本编号可以是 NM_004448.3）、EGFR（转录本编号可以是 NM_005228.3）、FGFR1（转录本编号可以是 NM_023110.2）、FGFR2（转录本编号可以是 NM_000141.4）、CDK4（转录本编号可以是 NM_000075.3）和 MET（转录本编号可以是

NM_000245.3）。

例如，本申请中样本选自以下组：组织样本、血液样本、唾液、胸腔积液、腹膜积液和脑脊液。

例如，所述步骤（S4-3）或模块（M4-3）可以包含基于所述待测样本的所述目标基因的外显子的长度以及所述待测样本的所述目标区间 i 上的拷贝数 CN_i，确定所述 CN_g。

例如，所述步骤（S4-3）或模块（M4-3）可以包含基于下式确定所述 CN_g，

$$CN_g = \frac{\sum_j^m (\frac{\sum_i^n (CN_i)}{n} * Len_j)}{\sum_j^m Len_j}$$

其中，i 可以表示目标区间，j 可以表示目标外显子，n 可以表示目标外显子 j 上的目标区间个数，m 可以表示目标外显子个数，CN_i可以表示目标区间 i 的拷贝数，Len_j可以表示目标外显子 j 的长度。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-4）或模块（M4-4）：确定在所述目标区间上待测样本的拷贝数变异的存在的概率。

例如，所述拷贝数变异的存在的概率可以包含在所述目标区间上所述待测样本发生拷贝数扩增的概率（p_a）和/或缺失的概率（p_d）。

例如，所述步骤（S4-4）或模块（M4-4）可以包含，基于所述待测样本在所述目标区间 i 上的测序数据、以及在相应目标区间上所述与待测样本最接近的参考样本组的测序数据的均值与标准差，通过概率分布的方法或记载该程序的装置确认所述拷贝数变异的存在的概率。

例如，所述概率分布可以包含正态性概率分布。例如，所述概率分布可以包含任意常见的概率分布。例如，所述概率分布可以包含任意的离散的概率分布。例如，所述概率分布可以包含任意的连续的概率分布。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-5）或模块（M4-5）：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio。

例如，所述步骤（S4-5）或模块（M4-5）可以包含，使所述目标基因上发生显著性拷贝数变异的目标区间数量除以所述目标基因上所有目标区间数量，得到所述 sigRatio。

例如，所述发生显著性拷贝数变异的目标区间可以包含所述拷贝数变异的比例为约 30%或更高的所述目标区间。例如，所述发生显著性拷贝数变异的目标区间可以包含所述拷贝数变异的比例为约 30%或更高、约 40%或更高、约 50%或更高、约 60%或更高、约 70%或更高、约 80%或更高、约 90%或更高、约 95%或更高、或约 95%或更高的所述目标区间。

例如，所述步骤（S4）或模块（M4）可以包含步骤（S4-6）或模块（M4-6）：确定在所述目标基因上所述待测样本的拷贝数变异的存在的统计检验参数。

例如，所述统计检验参数可以包含通过显著性检验确定的 p 值。

例如，所述显著性检验可以包含 T 检验。例如，所述显著性检验可以任意的显著性检验方式，以及根据实际情况进行变形的显著性检验方式。

例如，所述步骤（S4-6）或模块（M4-6）可以包含，基于在所述目标基因上所述待测样本的所述目标区间的数量、在所述目标基因上所述待测样本的各个所述目标区间的测序数据、在所述目标基因上所述待测样本的各个所述目标区间的测序数据的标准差、以及在相应目标基因上述与待测样本最接近的参考样本组的目标区间上的测序数据的均值和标准差，通过 T 检验的方法或记载该程序的装置确认 p 值 p_{ttest}。

例如，所述步骤（S4）或模块（M4）可以通过以下内容确定所述待测样本的目标基因的拷贝数状态：

当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；

当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；

当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值。

例如，其中 CN_{thA} 可以为约 2.25 至约 4。例如，其中 CN_{thA} 可以为约 2.25、约 2.50、约 2.75、约 3.00、约 3.25、约 3.50、约 3.75、或约 4.00。

例如，其中 CN_{thD} 可以为约 1.0 至约 1.75。例如，其中 CN_{thD} 可以为约 0.25、约 0.50、约 0.75、约 1.00、约 1.25、约 1.50、约 1.75。

例如，其中 $sigRatio_{th}$ 可以为约 0.3 至约 1。例如，其中 $sigRatio_{th}$ 可以为约 0.3、约 0.4、约 0.5、约 0.6、约 0.7、约 0.8、约 0.9、或约 1.0。

例如，其中 p_{th} 可以为约 0.05 至约 0.00001。例如，其中 p_{th} 可以为约 0.05、约 0.01、约 0.001、约 0.0001、约 0.00001、约 0.000001、或约 0.0000001。

数据库建立

一方面，本申请提供一种数据库建立方法，可以包含获取多个参考样本的测序数据，以及将所述参考样本分为两个或以上参考样本组。

例如，所述数据库建立方法可以包含（S1）获取待测样本的测序数据和/或多个参考样

本的测序数据；（S2）将所述参考样本分为两个或以上参考样本组。

一方面，本申请提供一种数据库建立装置，可以包含以下模块：接收模块，用于获取待测样本的测序数据和/或多个参考样本的测序数据；处理模块，用于将所述参考样本分为两个或以上参考样本组。

例如，所述数据库建立装置可以包含（M1）接收模块，用于获取待测样本的测序数据和/或多个参考样本的测序数据；（M2）处理模块，用于将所述参考样本分为两个或以上参考样本组。

一方面，本申请提供一种数据库建立方法，可以包含以下步骤：

（S1）获取待测样本的测序数据和/或多个参考样本的测序数据；步骤（S1-1）：通过基因测序获取所述待测样本和/或所述参考样本的所述测序数据；步骤（S1-2）：矫正所述待测样本和/或参考样本的所述测序数据；

（S2）将所述参考样本分为两个或以上参考样本组；步骤（S2-1）：使所述参考样本分组；步骤（S2-2）：确认所述参考样本组的所述测序数据的统计值。

一方面，本申请提供一种数据库建立装置，可以包含实现本申请数据库建立方法的模块。

一方面，本申请提供一种数据库建立方法，可以包含以下步骤：

（S1）获取待测样本的测序数据和/或多个参考样本的测序数据；步骤（S1-1）：通过基因测序获取所述待测样本和/或所述参考样本的所述测序数据；步骤（S1-2）：矫正所述待测样本和/或参考样本的所述测序数据；

（S2）将所述参考样本分为两个或以上参考样本组；步骤（S2-1）：使所述参考样本分组；步骤（S2-2）：确认所述参考样本组的所述测序数据的统计值。

一方面，本申请提供一种数据库建立装置，可以包含实现本申请数据库建立方法的模块。

另一方面，本申请提供一种数据库建立装置，可以包含以下模块：（M1）接收模块，用于获取待测样本的测序数据和/或多个参考样本的测序数据；（M2）处理模块，用于将所述参考样本分为两个或以上参考样本组。

拷贝数状态分析方法

一方面，本申请根据已有数据库的信息提供一种拷贝数状态分析方法， 可以包含从两个或以上参考样本组确定与待测样本最接近的参考样本组，以及基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

例如，所述拷贝数状态分析方法可以包含（S3）确定与所述待测样本最接近的参考样本组；（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标

基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态分析装置，可以包含以下模块：计算模块，用于从两个或以上参考样本组确定与所述待测样本最接近的参考样本组；判断模块，用于基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

例如，所述拷贝数状态分析装置可以包含（M3）计算模块，用于确定与所述待测样本最接近的参考样本组；（M4）判断模块，用于基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

一方面，本申请提供一种拷贝数状态分析方法，可以包含以下步骤：

（S3）确定与所述待测样本最接近的参考样本组；

（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态步骤（S4-1）：确定所述待测样本的目标基因在所述目标区间 i 上的拷贝数 CN_i；步骤（S4-3）：确定所述待测样本在目标基因上的拷贝数 CN_g；步骤（S4-4）：确定在所述目标区间上待测样本的拷贝数变异的存在的概率；步骤（S4-5）：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio；步骤（S4-6）：确定在所述目标基因上所述待测样本的拷贝数变异的存在的统计检验参数；通过以下内容确定所述待测样本的目标基因的拷贝数状态：当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值。

一方面，本申请提供一种拷贝数状态分析装置，可以包含实现本申请拷贝数状态分析方法的模块。

一方面，本申请提供一种拷贝数状态分析方法，可以包含以下步骤：

（S3）确定与所述待测样本最接近的参考样本组；

（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态步骤（S4-1）：确定所述待测样本的目标基因在所述目标区间 i 上的拷贝数 CN_i；步骤（S4-2）：使所述待测样本的所述目标区间上的拷贝数降噪化；步骤（S4-3）：确定所述待测样本在目标基因上的拷贝数 CN_g；步骤（S4-4）：确定在所述目标区间上待测样本的拷贝数变异的存在的概率；步骤（S4-5）：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio；步骤（S4-6）：确定在所述目标基因上所述待

测样本的拷贝数变异的存在的统计检验参数；通过以下内容确定所述待测样本的目标基因的拷贝数状态：当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值。

一方面，本申请提供一种拷贝数状态分析装置，可以包含实现本申请拷贝数状态分析方法的模块。

数据库、设备和应用方法

一方面，本申请提供一种数据库，其根据本申请所述的拷贝数状态分析方法或数据库建立方法建立。

另一方面，本申请还提供了一种储存介质，其记载可以运行本申请所述的方法的程序。

另一方面，本申请还提供了一种设备，其可以包含本申请所述的储存介质。例如，所述非易失性计算机可读存储介质可以包括软盘、柔性盘、硬盘、固态存储（SSS）（例如固态驱动（SSD））、固态卡（SSC）、固态模块（SSM））、企业级闪存驱动、磁带或任何其他非临时性磁介质等。非易失性计算机可读存储介质还可以包括打孔卡、纸带、光标片（或任何其他具有孔型图案或其他光学可识别标记的物理介质）、压缩盘只读存储器（CD-ROM）、可重写式光盘（CD-RW）、数字通用光盘（DVD）、蓝光光盘（BD）和/或任何其他非临时性光学介质。

例如，本申请的设备还可以包含耦接至所述储存介质的处理器，所述处理器可以被配置为基于存储在所述储存介质中的程序执行以实现本申请所述的方法。

另一方面，本申请还提供了一种本申请的方法，可以在疾病诊断、预防和/或治疗中的应用。

另一方面，本申请还提供了一种本申请的方法，可以在目标基因的拷贝数状态监测中的应用。

另一方面，本申请还提供了一种本申请的方法，可以在全基因组关联研究中的应用。

在本申请中，所述方法可以用于判断所述受试者是否具有拷贝数变异。例如，本申请的任一个或多个方法可以是非诊断目的的。例如，本申请的任一个或多个方法可以是诊断目的的。

在本申请中，所述方法可以用于通过检测所述拷贝数变异，用于临床实践（例如可以推

测某些特定的肿瘤治疗方式是否适于该受试者）。在某些情况下，所述方法检测出的拷贝数变异水平可以与本领域已知的生物标志物联合使用于临床实践。

不欲被任何理论所限，下文中的实施例仅仅是为了阐释本申请的方法和用途等，而不用于限制本申请发明的范围。

实施例

实施例 1

1.1 数据准备

选取 30 例阴性外周血样本，然后采用同一批实验试剂，从外周血中提取 DNA，并通过片段化、加接头及 PCR 扩增等实验步骤制备全基因组预文库。接着，将制备好的预文库分成两份，分别用不同批次的探针，记为 A 批次和 B 批次，与预文库杂交，特异性地捕获人类基因组中 *BRCA1* 基因，获得终文库 A 和终文库 B。用测序仪对两份终文库进行高通量测序。最后，将测序数据与人类基因组标准序列 hg19 进行比对，获得比对后的 BAM 文件。

1.2 基于构建参考基线的传统方法检测拷贝数变异

事先，用前期收集的足量（例如：50 例及以上）拷贝数正常的阴性样本作为参考集构建参考基线。随后，将两组实验样本采用该参考集构建的基线计算 *BRCA1* 基因上各外显子的拷贝数值及检测拷贝数变异。从计算的外显子拷贝数结果看（如图 1A 所示），A 批次探针捕获的实验数据均一性更好，与理论拷贝数值 2 更接近，而 B 批次探针捕获的结果相对更差，特别是 *BRCA1* 基因 8 号外显子，所有样本都有明显偏低的偏好。另一方面，从拷贝数变异检出结果看，采用 B 批次探针实验组中，30 例样本检出 2 例来自 *BRCA1* 的假阳性的拷贝数变异。说明由于可能的探针批次差异，采用传统的基于参考基线方式很容易导致拷贝数变异检出的准确性下降。

1.3 基于本申请方法检测拷贝数变异

因此，接下来采用本申请方法检测拷贝数变异。

（1）数据准备

本申请的拷贝数变异检测算法，可以选取足够的样本数量，例如可以是 15 例来自相同样本类型和相同实验方法学，且尽可能保证实验中使用试剂批次、实验设备等是一致的样本数据。每个参与的样本数据需要来自 NGS 测序数据比对后的 BAM 文件。

（2）基于 BAM 文件去重和 DNA 序列片段覆盖深度的统计

对每个样本 BAM 文件，首先，可以去除由于 NGS 建库中 PCR 引入的重复 DNA 序列片

段，获取唯一比对 DNA 片段。然后，根据要检测的目标 DNA 区域，采用滑窗法，每次滑动 24bp，将该区域划分成探针固定长度为 120bp 的窗口区域，并统计每个窗口内唯一比对的 DNA 片段的平均覆盖水平。

（3）样本测序覆盖度质控

任选地，对每个样本进行质量控制，判断平均测序深度、最小测序深度和覆盖均一性是否满足要求。其中，要求平均测序深度≥100X，最小测序深度≥30X，覆盖均一性≥90%（指碱基测序深度≥样本平均测序深度 20%的碱基占比，公式如下：覆盖均一度=（大于或等于样本平均测序深度的 20%的碱基数量/样本中总碱基数量）×100%。如果样本数据质量不满足要求，可以不用于矫正基线的构建。本申请的检测方法可以对满足质量合格的样本数至少为 10 例进行检测。

（4）数据校正及归一化处理

为了减低噪音和系统性偏差对拷贝数变异检测结果的影响，可以对每个窗口区域的覆盖水平进行矫正，包括覆盖水平初步矫正（基于样本平均覆盖水平）、GC 矫正和批次矫正。

（5）覆盖水平初步矫正

为了矫正不同样本测序覆盖深度的差异，覆盖水平初步矫是将批次内所有样本的覆盖水平矫正到同一指定的覆盖水平。具体地，对批次内样本的每一个窗口区域，测序所得平均覆盖水平除以样本内所有窗口区域的平均覆盖水平的总和，再乘以固定因子（因子为 1E+07）。

（6）GC 矫正

为了矫正由于 GC 偏好导致测序覆盖深度差异，GC 矫正通过计算每个窗口 GC 含量，然后利用 loess 回归方法对样本内每一个窗口区域的覆盖水平进行 GC 偏好矫正。

（7）批次矫正

i. 获取批次内所有质控合格样本的 GC 矫正数据。

ii. 计算参与构建批次基线的样本在每个窗口内的覆盖水平的中位值（median）和绝对离差中位值（MAD）。如果 MAD/median ＞ 设定阈值（例如，设定阈值可以为约 0.05 至约 0.15），表明该窗口覆盖水平不稳定，需要被剔除。

iii. 保留 MAD/median<小于设定阈值的窗口，或者保留 MAD/median 最小的前 4 个窗口，作为覆盖水平稳定的窗口区域。

iv. 接着，对保留的每个覆盖水平稳定的窗口区域，采用 grubbs 检验去除该窗口内异常的覆盖水平值，然后计算剩余覆盖水平值的平均水平，作为批次矫正参考系数。

v. 最后，对于每个待测样本，基于上述计算的批次矫正参考系数，对每个窗口区域的覆

盖水平进行归一化处理，并计算拷贝数 CN 值，每个窗口拷贝数 CN 值计算公式如下：

$$\text{CN 值} = \frac{\textit{待测样本在该窗口区域的覆盖水平}}{\textit{该窗口的批次矫正参考系数}} * 2$$

（8）拷贝数变异识别

利用 CBS 算法识别样本目标区域上的断点位置，获取候选的拷贝数变异区域。然后，对每个候选的拷贝数变异区域进行显著性检验，具体地，通过 T 检验，判断待测样本在候选拷贝数变异区域上的窗口覆盖水平与批次内其他样本在该区域的覆盖水平是否具有显著性差异，从而判断候选拷贝数变异的可靠性。

其中 *BRCA1* 基因外显子拷贝数分布如图 1B 所示，可以看出，相比基于构建参考基线的传统方法，本申请方法获得拷贝数结果均一性更好，特别是对批次差异大的 B 组探针，效果更明显，并且两组实验数据都没有检出任何假阳性拷贝数变异。

实施例 2

选取 20 例细胞系样本，其中 19 例为阴性样本，1 例为已知的外显子拷贝数变异（LGR）样本（*BRCA1*: Exon 12 amp）。实验采用仪器自动化建库方式获取高通量测序数据。最后，将测序数据与人类基因组标准序列 hg19 进行比对，获得比对后的 BAM 文件。将样本 BAM 文件分别用基于构建参考基线的传统方法和本申请的方法检测拷贝数变异。其中，基于构建参考基线方法采用的基线可以是用早期手动化建库方式的样本数据建立（例如实施例 1 中所用的参考基线）。

含有拷贝数变异的阳性样本结果如图 2A-2B 所示，基于构建参考基线传统方法检测结果（图 2A 所示）数据背景噪音极大，无法检出拷贝数变异，而本申请方法数据背景噪音明显更小，可以检出拷贝数变异（图 2B 所示），说明不同实验方法产生的 NGS 数据可能有很大差异，基于手动化建库方式的数据构建的基线不适用于自动化建库数据。说明当实验方法发生改变时，如果采用参考基线传统方式，需要事先收集足够量使用该实验方法的样本数据，然后人工构建新的基线，这大大增加了实验成本和人力浪费。

实施例 3

选取 696 例外周血样本检测 *BRCA1* 和 *BRCA2* 的外显子拷贝数变异（LGR），实验采用 RNA 探针特异性捕获 *BRCA1* 和 *BRCA2* 基因区域，然后经高通量测序，将测序数据与人类基因组标准序列 hg19 进行比对，获得比对后的 BAM 文件。随后，分别采用基于构建参考基线的方法和本申请方法检测拷贝数变异。同时，所有样本拷贝数变异都经 *BRCA* MASTR Plus

Dx 试剂盒（基于多重 PCR 捕获方法学）确认，一共包含 17 例 LGR 阳性样本和 679 例阴性样本。

以 *BRCA* MASTR Plus Dx 试剂盒检出结果为基准，分别得到该 696 例外周血样本基于构建参考基线传统方法检测结果和本申请方法检测结果的灵敏度和特异性，分别如表 1 和表 2 所示。

表 1. 基于构建参考基线方法的检测结果

		基于 MASTR 检测结果		检测指标
		阳性	阴性	
基于参考基线检测结果	阳性	12	167	阳性预测值 67%
	阴性	5	512	阴性预测值 99%
检测指标		灵敏度 70.6%	特异性 75.4%	准确性 75.3%

表 2. 本申请方法的检测结果

		基于 MASTR 检测结果		检测指标
		阳性	阴性	
本申请检测结果	阳性	12	3	阳性预测值 80%
	阴性	5	676	阴性预测值 99.3%
检测指标		灵敏度 70.6%	特异性 99.6%	准确性 98.9%

对比表 1 和表 2 可以看出，相比传统构建基线方法，通过本申请方法在不损失灵敏度的情况下，可以大大降低样本假阳性，检测准确性可以从 75.3%提高到 98.9%。

实施例 4

选取 14 例细胞系样本测序比对后的数据构建批次基线，构建批次基线过程中，对于描述窗口覆盖波动水平的阈值，分别设为 0.05 和 0.15，构建 2 条批次基线。然后分别用 2 条批次基线对该 14 例样本中已知 LGR 拷贝数变异的样本（*BRCA1*: exon4-6del）进行批次矫正，再检测拷贝数变异。

含有拷贝数变异的阳性样本结果如图 3A-3B 所示，基于不同窗口覆盖波动水平的阈值构建的批次基线都可以明显地检测到拷贝数变异，说明本申请筛选稳定区间的阈值范围都可以实现对于拷贝数变异的检测。

实施例 5

选取 10 例阴性细胞系样本作为模拟样本本底，然后选择 10 例已被文献报道的 *BRCA1* 和

BRCA2 基因的 LGR 拷贝数变异作为待模拟的突变（表 3 所示），包含 5 种拷贝数扩增变异和 5 种拷贝数缺失变异。经过模拟将上述拷贝数扩增变异和拷贝数缺失变异人为加入到模拟样本本底数据中，最终得到 10 例阳性 LGR 模拟样本数据。

将 10 例模拟阳性样本构建批次基线，再使用构建的批次基线对该 10 例模拟样本进行批次矫正和拷贝数变异识别。10 例模拟样本的结果如图 4A-4J 所示，模拟的 10 例拷贝数变异都能准确检出，说明本申请对于任意区域的拷贝数变异都可以实现准确的检测。

表 3：10 例模拟的拷贝数变异类型

附图名称	基因	拷贝数变异
BRCA1 外显子 16-19 拷贝数扩增（图 4A）	*BRCA1*	外显子 16-19_Dup
BRCA1 外显子 3-7 拷贝数扩增（图 4B）	*BRCA1*	外显子 3-7_Dup
BRCA2 外显子 14-18 拷贝数扩增（图 4C）	*BRCA2*	外显子 14-18_Dup
BRCA2 外显子 3 拷贝数扩增（图 4D）	*BRCA2*	外显子 3_Dup
BRCA2 外显子 25-27 拷贝数扩增（图 4E）	*BRCA2*	外显子 25-27_Dup
BRCA1 外显子 15-22 拷贝数缺失（图 4F）	*BRCA1*	外显子 15-22_Del
BRCA1 外显子 2 拷贝数缺失（图 4G）	*BRCA1*	外显子 2_Del
BRCA1 外显子 7-12 拷贝数缺失（图 4H）	*BRCA1*	外显子 7-12_Del
BRCA2 外显子 15-16 拷贝数缺失（图 4I）	*BRCA2*	外显子 15-16_Del
BRCA2 外显子 8-11 拷贝数缺失（图 4J）	*BRCA2*	外显子 8-11_Del

实施例 6

本申请的方法，使用聚类方法，将大量的真实样本数据，根据测序深度的趋势聚类情况划分成不同的样本集合，分别构建基线（平均深度及深度波动范围），根据样本与基线的相似性动态筛选背景基线，消除批次效应的同时，提高检测的特异性和敏感性。同时可以任选地使用离散小波变换方法对拷贝数进行平滑降噪，提高测序数据的信噪比。

为了提供一种高通量测序中可以实现高敏感性和精准度的拷贝数变异检测的方法和介质，本申请的方法基于样本覆盖特征特异性的差异化来检测拷贝数变异。具体地，本申请基于大规模样本的聚类分析，构建多条对照组基线，可以避免了测序中由于实验和样本差异导致覆盖深度特征不一致引起的基线不匹配问题；并且可以综合多种覆盖深度矫正策略，降低样本特异性的数据差异；最终可以通过定量分析、统计学差异评估，保证结果的准确性和稳定性。本申请拷贝数变异检测的方法不仅可以适用于特定基因面板（panel）的靶向捕获测序数据，

也可以适用于全外显子组的捕获测序数据。为了上述问题，本申请的数据库建立方法可以包含如下步骤：

1. 数据准备模块，包括：

a) 序列比对：将高通量测序原始 fastq 数据回帖到人类参考基因组，确定待测样本、参考样本的目标区间与人类参考基因组参考序列匹配的序列；

b) 重复序列去除：将 PCR 扩增过程中产生的重复序列去掉；

c) 覆盖深度计算：计算目标区间上每个碱基的测序深度 RD_{Base}；

2. 覆盖深度矫正模块，包括 3 方面独立的，顺序任选的矫正：

a) 使所述测序数据标准化：样本总测序深度矫正，具体地可以是，对每个目标区间各个位点上的覆盖深度根据样本总测序深度进行标准化，以消除不同样本之间的测序数据量差异，获得 RD_{normD}；

$$RD_{normD} = \frac{RD_i}{\sum_{i=1}^{n} RD_i} * R(R = 1e + 07)$$

其中，i 表示目标区间上的位点，n 表示所有目标区间上位点的总数，RD_i表示目标区间上位点 i 的测序深度，R 为一个常数，可以根据所有区间的大小进行设置，以保证待测样本矫正后的深度和参考样本组矫正后的深度在相同的水平。

b) 使所述测序数据平滑化 1：探针铺设特征矫正，具体可以是，根据探针设计中不同区间探针铺设乘数差异，例如区间上覆盖的探针数，将区间分割，每个目标区间长度可以为约 24 个碱基对，并计算各个目标区间平均覆盖深度 RD，根据每个目标区间上覆盖的探针数 ProbeN，对目标区间上测序深度进行局部加权回归（loess(RD~ProbeN)）矫正，得到探针矫正后的测序深度 RD_{normP};

c) 使所述测序数据平滑化 2：GC 矫正，具体可以是，将用于覆盖深度计算的目标区间按照侧翼延伸至总长大于 200bp 长度，计算平均 GC 占比，根据各区间的 GC 含量 GC，对测序深度 RD 进行局部加权回归（loess(RD~GC)）矫正，得到 GC 矫正后的测序深度 RD_{normGC}；

d) 可选的，可以进行筛选 GC 含量均衡的区间，将 GC 极度不均衡的区域（CG 含量低于 0.25，或 GC 含量高于 0.75 的区域）过滤掉，使用矫正后的覆盖深度进行拷贝数变异检测。

3. 基线构建模块，具体包含以下步骤：

a) 样本聚类：已有方法一般将所有参考样本作为一个类别来构建基线。本申请的方法对于参考样本进行分组，具体是基于各参考样本覆盖深度在目标区间上变化的一致性，例如在目标区间上参考样本的所述测序的近似程度，进行聚类分析，将参考样本分成不同类别的参

考样本组，聚类方法可以例如 K 均值聚类、层次聚类等方法；

b) 基线构建：对于每个参考样本组，分别构建一条基线。具体为，分别计算每个参考样本组中，各个目标区间上所有参考样本的平均测序深度（$MeanRD^{i}_{Baseline}$）和测序深度的标准差($SdRD^{i}_{Baseline}$)，作为基线（$Baseline_i$），其中 i={1,2,3,4,…}。例如，每个参考样本组中需要有足够样本数量才具有统计学意义，每个参考样本组的样本数不低于 30 例。参考样本组数的设定需要考虑到肿瘤样本特征及测序质量，根据捕捉到的特征数确定参考样本组数，例如参考样本组数量可以为 2 个以上，如 2-10 条。

c) 可选的，可以进行区间筛选：计算每个区间上测序深度的变异系数 cv，去掉在样本中波动程度大的不稳定区间，其中：

$$cv^{i} = \frac{SdRD^{i}_{Baseline}}{MeanRD^{i}_{Baseline}}$$

其中，$MeanRD^{i}_{Baseline}$和$SdRD^{i}_{Baseline}$分别表示各个目标区间上所有参考样本的平均测序深度和测序深度的标准差。当 cv>0.8 时，该区间被认为是不稳定区域而被过滤掉；同时矫正后的测序深度低于 5 时，被认为是捕获效率低的区域而被过滤掉，最终保留的区间作为稳定的区间。

最终，根据本实施例可以得到本申请的数据库，包含在目标区间上变化具有一致性的两个或以上参考样本组。与现有技术相比，本申请的数据库的优势在于：通过大规模样本的聚类分析，将参考样本划分为不同类别的参考样本组，分别构建样本特异性的背景基线，极大的降低了高通量测序数据在拷贝数变异检测中批次效应产生的假阳性，增加结果的稳定性。同时本申请消除批次效应的方法不需要同批次中保证有足够数量的相同基因 panel 的样本，极大降低了实际应用中的困难。

实施例 7

为了解决测序中由于实验和样本差异导致覆盖深度特征不一致引起的基线不匹配问题，本申请还提供了一种拷贝数状态分析方法。本申请的拷贝数状态分析方法可以包含如下步骤：

a) 根据所述待测样本与所述参考样本组的相似度，确定与所述待测样本最接近的参考样本组，即动态筛选基线：通过统计距离的计算方法，例如闵可夫斯基距离等，比较各个目标区间上待测样本的测序深度与该目标区间上各个参考样本组的测序深度，确认所述参考样本组与所述待测样本的统计距离：

$$L_p(RD_{Sample}, RD_{Baseline}) = \left(\sum_{i=1}^{n} |RD^i_{Sample} - RD^i_{Baseline}|^p\right)^{\frac{1}{p}}$$

其中，L_p 值表示统计距离，i 表示目标区间，n 表示目标区间个数，RD^i_{Sample} 表示待测样本的目标区间 i 的测序深度，$RD^i_{Baseline}$ 表示参考样本组的目标区间 i 的测序深度，其中 p 任选地大于或等于 1。选择与待测样本具有最小统计距离（L_p 值最小，表示相似度最高）的参考样本组作为该待测样本的背景基线（$Baseline_x$）。

b) 基因拷贝数检测，具体如下：

i. 各待测样本的目标区间拷贝数评估：计算待测样本各个目标区间的拷贝数 CN_i，公式如下：

$$CN_i = 2 * \frac{RD_{sample}}{RD^x_{Baseline}}$$

其中，RD_{sample} 表示待测样本的各个目标区间的测序深度，$RD^x_{Baseline}$ 表示与待测样本最接近的参考样本组的各个目标区间的测序深度，其中倍性可以为 2。

ii. 可选的，可以进行各区间拷贝数平滑降噪：使用降噪算法对各个区间的 *CNi* 进行平滑降噪，提高数据的信噪比。降噪方法可以使用离散小波变换（Discrete Wavelet Transformation，DWT）、主成分分析算法、奇异值分解和/或高斯滤波的方法进行平滑降噪。DWT 是将信号分成高频信号和低频信号，分别通过低通滤波器和高通滤波器，将离散信号作离散小波变换，将连续的小波及其小波变换离散化，达到数据降噪的目的。这样可以得到降噪后的 *CNi*;

iii. 各目标基因的拷贝数评估：计算样本各个目标基因的加权平均拷贝数 CN_g，并使用目标外显子的长度对 CN_i 进行矫正，例如：

$$CN_g = \frac{\sum_j^m (\frac{\sum_i^n (CN_i)}{n} * Len_j)}{\sum_j^m Len_j}$$

其中，i 表示目标区间，j 表示目标外显子，n 表示目标外显子 j 上的目标区间个数，m 表示目标外显子个数，CN_i 表示目标区间 i 的拷贝数，Len_j 表示目标外显子 j 的长度。

iv. 确定在所述目标区间上待测样本的拷贝数变异的存在的概率，例如可以是多个区间的测序深度与筛选出的背景基线的分布进行正态性检验，公式如下：

$$p_a = pnorm(RD^i_{Sample}, mean = MeanRD^x_{Baseline}, sd = SdRD^x_{Baseline}, lower.tail = F)$$

$$p_d = pnorm(RD^i_{Sample}, mean = MeanRD^x_{Baseline}, sd = SdRD^x_{Baseline}, lower.tail = T)$$

其中，RD^{i}_{Sample}表示待测样本的目标区间 i 的测序深度，$MeanRD^{x}_{Baseline}$表示与待测样本最接近的参考样本组的各个目标区间的测序深度的均值，$SdRD^{x}_{Baseline}$表示与待测样本最接近的参考样本组的各个目标区间的测序深度的标准差，$lower.tail = F$计算右尾概率，p_a表示该区间发生拷贝数扩增的概率值，$lower.tail = T$计算左尾概率，p_d表示该区间发生拷贝数缺失的概率值。

v. 各目标基因显著性比例评估：分别计算各个目标基因显著性扩增或缺失的比例 *sigRatio*，公式如下：

$$sigRatio = \frac{\text{发生显著性拷贝数变异的目标区间的个数}}{\text{所述目标基因上所有目标区间的个数}}$$

其中，所述发生显著性拷贝数变异的目标区间包含拷贝数变异的比例为约 30%或更高的所述目标区间。

vi. 各目标基因整体水平显著性检验：对每个目标基因上各个目标区间矫正后的测序深度基于与待测样本最接近的参考样本组的各个区间的平均测序深度，进行 T 检验，判断样本与基线的差异是否显著，得到 p_{ttest}。

c) 确定所述待测样本的目标基因的拷贝数状态，判断标准为：

$$\text{拷贝数状态} = \begin{cases} \text{扩增：} CN_g \geq CN_{thA}\ AND\ sigRatio \geq sigRatio_{th}\ AND\ p_{ttest} \leq p_{th} \\ \text{缺失：} CN_g \leq CN_{thD}\ AND\ sigRatio \geq sigRatio_{th}\ AND\ p_{ttest} \leq p_{th} \\ \text{正常：} CN_{thD} < CN_g < CN_{thD}\ OR\ sigRatio < sigRatio_{th}\ OR\ p_{ttest} > p_{th} \end{cases}$$

各个阈值可通过使用大规模样本训练得到。其中：CN_{thA}表示拷贝数扩增的阈值，取值可以任选为 2.25～4；CN_{thD}表示拷贝数缺失的阈值，取值可以任选为 1.0~1.75；$sigRatio_{th}$表示显著性扩增/缺失比例的阈值，取值可以任选为 0.3～1；p_{th}表示显著性 T 检验的阈值，取值可以任选为 0.05～0.00001。

本申请的拷贝数状态分析方法根据待测样本与参考样本组的相似性动态筛选与待测样本最接近的参考样本组作为背景基线，可以消除批次效应，也可以提高检测的特异性和敏感性。

实施例 8

数据库建立：使用 655 例参考样本构建基线，采用本申请的数据库构建方法，例如使用 k 均值聚类算法，将参考样本划分为 5 个参考样本组，构建 5 条不同候选基线作为数据库。

构建模拟数据：使用 varBen 肿瘤突变数据模拟软件（github.com/nccl-jmli/VarBen），以良性组织样本为基础，通过插入基因的读段的方式，在测序数据中插入目标基因的读段，梯度模拟目标基因不同拷贝数扩增的情况，模拟样本列表如表 4 所示。

表 4：模拟样本列表

目标基因	模拟样本数	模拟拷贝数梯度
ALK	20	2.5，2.75，3.0，3.5，4.0
ERBB2	20	2.5，2.75，3.0，3.5，4.0
FGFR1	20	2.5，2.75，3.0，3.5，4.0
FGFR2	20	2.5，2.75，3.0，3.5，4.0

根据本申请的拷贝数状态分析方法检测模拟样本，检测的结果如表 5 所示。

表 5：本申请在模拟样本中的检测结果

目标基因	模拟拷贝数	本申请扩增检出比例	本申请检测的拷贝数平均值	本申请检测的95%置信区间的拷贝数
ALK	2.5	100%	2.58	(2.41, 2.76)
	2.75	100%	2.84	(2.64, 3.04)
	3.0	100%	3.11	(2.94, 3.28)
	3.5	100%	3.41	(3.19, 2.63)
	4.0	100%	3.96	(3.69,4.22)
ERBB2	2.5	100%	2.58	(2.36, 2.81)
	2.75	100%	2.8	(2.6,3.18)
	3.0	100%	3.18	(2.89,3.47)
	3.5	100%	3.50	(3.14, 3.86)
	4.0	100%	4.08	(3.68, 4.48)
FGFR1	2.5	100%	2.55	(2.40, 2.70)
	2.75	100%	2.84	(2.66, 3.02)
	3.0	100%	3.12	(2.94, 3.30)
	3.5	100%	3.40	(3.21, 3.58)
	4.0	100%	3.98	(3.78, 4.19)
FGFR2	2.5	100%	2.6	(2.38, 2.82)
	2.75	100%	2.92	(2.69, 3.15)
	3.0	100%	3.25	(2.97, 3.53)

	3.5	100%	3.57	(3.33, 3.82)
	4.0	100%	4.17	(3.89, 4.45)

图 5A-5F 显示的是，本申请的检测结果部分数据的拷贝数分布图示例。每个点表示基因的一个区间，灰色的点表示拷贝数正常的基因，黑色的点表示拷贝数发生扩增或缺失的基因，同时标注出相应的基因名。横轴表示基因所在的染色体位置，纵轴表示基于本申请方法计算得到的拷贝数（中间横线表示正常基因的拷贝数），灰色背景表示背景基线（与待测样本最接近的参考样本组）中各目标区间的波动范围。图 5A–5C 分别为模拟 ERBB2 基因发生不同程度的拷贝数扩增，图 5D–5F 分别为 FGFR1 基因发生不同层度的拷贝数扩增，模拟拷贝数梯度为 2.5，2.75 和 3.0。结果表明，本申请的拷贝数状态分析方法用于模拟样本中，所有模拟基因及不同梯度的拷贝数扩增均能稳定检测，且拷贝数预测准确。

实施例 9

阳性标准样本：本申请测试包括 30 例阳性标准样本，来源于 NCI-BL2009 细胞系，使用质粒转染的方式将相应比例的目标基因转染至细胞系中，获得 CNV 阳性数据，并使用微滴式数字 PCR（ddPCR）定量基因的拷贝数。质粒编号为：Life RPCI11.C-433C10 BAC-EGFR，Life RPCI11.C-936I7 BAC-CDK4，Life RPCI11.C-163C9 BAC-MET，Life RPCI11.C-909L6 BAC-ERBB2，Life RPCI11.C-957P17 BAC-FGFR1。阳性标准样本列表如表 6 所示。

表 6：阳性标准样本列表

基因	标准样本数	ddPCR 标定拷贝数
CDK4	10	3, 5, 8
ERBB2	10	3, 5, 8
EGFR	10	3, 5, 8
FGFR1	10	3, 5, 8
MET	10	3, 5, 8

数据库建立：使用 655 例参考样本构建基线，采用本申请的数据库构建方法，例如使用 k 均值聚类算法，将参考样本划分为 5 个参考样本组，构建 5 条不同候选基线作为数据库。

根据本申请的拷贝数状态分析方法检测拷贝数扩增阳性标准样本的拷贝数状态，检测的结果如表 7 所示。

表 7：本申请在阳性标准样本的检测结果

基因	ddPCR 标定拷贝数	本申请扩增检出比例	本申请检测的拷贝数平均值	本申请检测的95%置信区间的拷贝数
CDK4	3	100%	2.84	(2.68, 3.00)
	5	100%	4.86	(4.70, 5.02)
	8	100%	7.60	(7.30, 7.90)
ERBB2	3	100%	2.70	(2.70, 2.70)
	5	100%	4.17	(4.03, 4.32)
	8	100%	6.51	(6.24, 6.78)
EGFR	3	100%	2.70	(2.70, 2.70)
	5	100%	4.43	(4.28, 4.57)
	8	100%	7.03	(6.80, 7.26)
FGFR1	3	100%	2.90	(2.90, 2.90)
	5	100%	4.77	(4.63, 4.92)
	8	100%	7.53	(7.24, 7.81)
MET	3	100%	2.91	(2.80, 3.03)
	5	100%	4.66	(4.50, 4.82)
	8	100%	7.26	(7.02, 7.49)

图 6A-6C 显示的是，本申请的检测结果部分数据的拷贝数分布图示例。图 6A-6C 表示质粒转染的 CNV 阳性细胞系标准样本的检测结果，ddPCR 标定拷贝数分别为 3，5 和 8。结果表明，本发明方法用于质粒转染的细胞系标准品中，所有的基因以及不同的拷贝数状态下，均能稳定检出，且拷贝数预测准确。

实施例 10

真实数据：本申请测试的真实样本包括第三方免疫组化法（IHC）验证的 20 例 ERBB2 扩增阳性样本，真实样本列表如表 8 所示。

表 8：真实样本列表

基因	样本数	IHC 结果
ERBB2	20	拷贝数：3+

数据库建立：使用 443 例参考样本构建基线，采用本申请的数据库构建方法，例如使用 k 均值聚类算法，将参考样本划分为参考样本组，构建不同候选基线作为数据库。

根据本申请的拷贝数状态分析方法检测真实样本的拷贝数状态，检测的结果如表 9 所示。

表 9：本申请在真实样本中的检测结果

基因	本申请扩增检出比例	本申请检测的拷贝数范围	本申请检测的拷贝数平均值
ERBB2	100%	3.27～6.66	4.7485

图 7A-7C 显示的是，本申请的检测结果部分数据的拷贝数分布图示例。图 7A-7C 表示真实 ERBB2 阳性样本的检测结果。结果表明，本申请方法用于真实样本中，20 例 IHC 结果为 HER2 阳性的样本，均能稳定检出。

实施例 11

阳性标准样本：本申请测试包括 3 例阳性标准样本，同实施例 9 中来源，用于检测不同基线的检测结果。阳性标准样本列表如表 10 所示。

表 10：阳性标准样本列表

样本	**基因**	**ddPCR 标定拷贝数**
标准样本 1	CDK4，ERBB2，EGFR，FGFR1，MET	3
标准样本 2	CDK4，ERBB2，EGFR，FGFR1，MET	5
标准样本 3	CDK4，ERBB2，EGFR，FGFR1，MET	8

数据库建立：使用 655 例参考样本构建基线，采用本申请的数据库构建方法，例如使用 k 均值聚类算法，将参考样本划分为 5 个参考样本组，构建 5 条不同候选基线作为数据库。同时，不使用聚类方法，将所有的参考样本构建 1 条基线。

分别使用聚类算法得到的 5 条基线和不聚类的 1 条基线，作为参考对照，检测拷贝数扩增标准样本的拷贝数状态。基线的选择和样本的波动情况如表 11 所示，检测的结果如表 12 所示。

表 11：不同基线检测结果的波动情况

样本	**基线**	**样本与基线的距离**	**样本整体拷贝数的 SD**
标准样本 1	基线 1（本申请方法）	34.655	0.2378
	基线 2	66.018	0.4711
	基线 3	43.609	0.3223

	基线 4	60.49	0.5293
	基线 5	48.74	0.3685
	所有参考样本构建 1 条基线	57.993	0.5107
标准样本 2	基线 1（本申请方法）	35.198	0.2356
	基线 2	66.625	0.4628
	基线 3	44.022	0.3172
	基线 4	61.076	0.5260
	基线 5	49.181	0.3573
	所有参考样本构建 1 条基线	58.544	0.5095
标准样本 3	基线 1（本申请方法）	34.94	0.2378
	基线 2	65.714	0.4670
	基线 3	44.089	0.3150
	基线 4	60.32	0.5297
	基线 5	48.457	0.3597
	所有参考样本构建 1 条基线	57.749	0.5044

表 12：不同基线的检测结果

基因	**基线**	**标准样本 1**		**标准样本 2**		**标准样本 3**	
		检测拷贝数	**检出情况**	**检测拷贝数**	**检出情况**	**检测拷贝数**	**检出情况**
CDK4	基线 1（本申请方法）	2.8	扩增	4.8	扩增	7.7	扩增
	基线 2	2.5	正常	4.2	扩增	6.7	扩增
	基线 3	2.8	正常	4.6	扩增	7.2	扩增
	基线 4	2.9	正常	4.8	扩增	7.6	扩增
	基线 5	2.7	正常	4.5	扩增	7.2	扩增
	所有参考样本构建 1 条基线	3.1	正常	5.2	扩增	8.3	扩增
ERBB2	基线 1（本申请方法）	2.7	扩增	4.2	扩增	6.4	扩增

	基线 2	2.3	正常	3.7	扩增	5.6	扩增
	基线 3	2.5	正常	4	扩增	6.1	扩增
	基线 4	2	正常	3.2	扩增	4.9	扩增
	基线 5	2.4	正常	3.9	扩增	5.9	扩增
	所有参考样本构建 1 条基线	2.5	正常	4	扩增	6.1	扩增
EGFR	基线 1（本申请方法）	2.7	扩增	4.4	扩增	7.1	扩增
	基线 2	2.5	正常	4.2	扩增	6.8	扩增
	基线 3	2.5	正常	4.2	扩增	6.6	扩增
	基线 4	3.1	正常	5.2	扩增	8.2	扩增
	基线 5	2.5	正常	4.1	扩增	6.7	扩增
	所有参考样本构建 1 条基线	3.1	正常	5.2	扩增	8.2	扩增
FGFR1	基线 1（本申请方法）	2.9	扩增	4.8	扩增	7.6	扩增
	基线 2	2.8	正常	4.6	扩增	7.2	扩增
	基线 3	2.9	正常	4.7	扩增	7.5	扩增
	基线 4	3.1	正常	5.1	扩增	8	扩增
	基线 5	2.7	正常	4.5	扩增	7.2	扩增
	所有参考样本构建 1 条基线	3.5	扩增	5.8	扩增	9.1	扩增
MET	基线 1（本申请方法）	2.9	扩增	4.6	扩增	7.3	扩增
	基线 2	3.2	扩增	5.1	扩增	8.1	扩增
	基线 3	2.8	正常	4.4	扩增	7	扩增
	基线 4	3.2	扩增	5.2	扩增	8.1	扩增
	基线 5	2.9	扩增	4.6	扩增	7.2	扩增
	所有参考样本构建 1 条基线	2.9	正常	4.6	扩增	7.3	扩增

图 8A-8F 显示的是，标准样本 1 使用不同的基线检测结果的拷贝数分布示例图。

结果表明，使用本申请发明方法匹配到的最优基线与待测样本最相近（样本与基线的距离值最小），样本整体拷贝数的波动（SD）最低，拷贝数分布图最稳定，噪音最小，说明本方法检测结果更加稳定。本申请方法，在所有基因以及不同的拷贝数状态下，均能稳定检出，其他基线在拷贝数为 3 时无法稳定检出。

前述详细说明是以解释和举例的方式提供的，并非要限制所附权利要求的范围。目前本申请所列举的实施方式的多种变化对本领域普通技术人员来说是显而易见的，且保留在所附的权利要求和其等同方案的范围内。

权 利 要 求 书

1. 一种拷贝数状态分析方法，

所述方法包含，将待测样本的目标区间划分为若干个窗口区域，获取待测样本组中的对照窗口区域的测序数据，基于所述对照窗口区域的测序数据，确定所述待测样本的目标基因的拷贝数状态；任选地，所述对照窗口区域包含覆盖波动水平低的窗口区域。

2. 如权利要求 1 所述的方法，所述方法还包含以下步骤：

（S1）获取所述待测样本的测序数据和/或多个参考样本的测序数据；

（S2）将所述参考样本分为两个或以上参考样本组；

（S3）确定与所述待测样本最接近的参考样本组；

（S4）基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

3. 如权利要求 1-2 中任一项所述的方法，

所述方法包含，将质量合格的样品的窗口区域按照覆盖波动水平从低到高排序，所述对照窗口区域包含覆盖波动水平前 2 个或更多、或前 4 个或更多所述窗口，或者所述对照窗口区域的所有所述质量合格的样品的所述测序数据的绝对离差中位数与中位数的比值为约 0.15 或更小。

4. 如权利要求 3 所述的方法，

所述方法包含，基于所述质量合格的样品的窗口区域的测序数据统计值，确定所述覆盖波动水平；任选地，基于所述质量合格的样品的窗口区域的测序数据的绝对离差中位数与中位数的比值，确定所述覆盖波动水平。

5. 如权利要求 3-4 中任一项所述的方法，

所述方法包含，基于所述对照窗口区域的测序数据，确定归一化系数；任选地，通过计算所述对照窗口区域的所有所述质量合格的样品的测序数据平均值，确定所述归一化系数。

6. 如权利要求 5 所述的方法，

所述方法包含，基于所述归一化系数，确定待测样本的每一个窗口区域的拷贝数；任选地，所述归一化包含将所述窗口区域的待测样本的测序数据除以所述窗口区域的归一化系数，乘以倍性。

7. 如权利要求 1-6 中任一项所述的方法，

所述方法包含，基于待测样本的每一个窗口区域的测序数据以及相应窗口区域的待测样本组中其它样本的测序数据，确定待测样本的拷贝数变异显著性；任选地，通过 T 检验

显著性检验的方法，确定所述拷贝数变异的显著性。

8. 如权利要求 1-7 中任一项所述的方法，

所述待测样本选自以下组：组织样本、血液样本、唾液、胸腔积液、腹膜积液和脑脊液。

9. 如权利要求 2-8 中任一项所述的方法，

所述步骤（S2）包含步骤（S2-1）：使所述参考样本分组，所述分组包含基于目标区间的所述测序数据通过聚类分析的方法使所述参考样本分组；优选地，所述聚类分析的方法包含 K 均值聚类和/或层次聚类；

所述步骤（S2）包含步骤（S2-2）：确认所述参考样本组的所述测序数据的统计值；优选地，所述确认统计值包含计算在所述目标区间上每组中所述参考样本的均值和/或标准差。

10. 如权利要求 2-9 中任一项所述的方法，

所述步骤（S3）包含基于在目标区间上所述参考样本组与所述待测样本的所述测序数据，通过计算统计距离的方法，确认所述参考样本组与所述待测样本的分布相似程度；优选地，所述分布相似程度高包含在所述目标区间上所述参考样本组与所述待测样本的所述统计距离短。

11. 如权利要求 10 所述的方法，

所述统计距离包含所述目标区间上所述参考样本组与所述待测样本的所述测序数据的差值的绝对值的 p 次方的统计值，所述 p 为 1 或更大，优选地，所述统计值包含求和值；优选地，所述统计距离包含闵可夫斯基距离。

12. 如权利要求 2-11 中任一项所述的方法，

所述步骤（S4）包含基于所述待测样本的所述目标基因的外显子的长度以及所述待测样本的目标区间 i 上的拷贝数 CN_i，确定待测样本在目标基因上的拷贝数 CN_g；优选地，所述步骤（S4）包含基于下式确定所述 CN_g，

$$CN_g = \frac{\sum_j^m (\frac{\sum_i^n (CN_i)}{n} * Len_j)}{\sum_j^m Len_j}$$

其中，i 表示目标区间，j 表示目标外显子，n 表示目标外显子 j 上的目标区间个数，m 表示目标外显子个数，CN_i表示目标区间 i 的拷贝数，Len_j表示目标外显子 j 的长度。

13. 如权利要求 2-12 中任一项所述的方法，

所述步骤（S4）包含：确定在目标区间上待测样本的拷贝数变异的存在的概率，所述拷

贝数变异的存在的概率包含在目标区间上所述待测样本发生拷贝数扩增的概率（p_a）和/或缺失的概率（p_d）；优选地，所述步骤（S4）包含，基于所述待测样本在目标区间 i 上的测序数据、以及在相应目标区间上所述与待测样本最接近的参考样本组的测序数据的均值与标准差，通过概率分布的方法确认所述拷贝数变异的存在的概率；优选地，所述概率分布包含正态性概率分布。

14. 如权利要求 2-13 中任一项所述的方法，

所述步骤（S4）包含：确定在所述目标基因上所述待测样本的显著性拷贝数扩增或缺失的存在的比例 sigRatio；优选地，使所述目标基因上发生显著性拷贝数变异的目标区间数量除以所述目标基因上所有目标区间数量，得到所述 sigRatio；所述发生显著性拷贝数变异的目标区间包含所述拷贝数变异的比例为约 30%或更高的目标区间；

所述步骤（S4）还包含：确定在所述目标基因上所述待测样本的拷贝数变异的存在的统计检验参数；优选地，基于在所述目标基因上所述待测样本的所述目标区间的数量、在所述目标基因上所述待测样本的各个目标区间的测序数据、在所述目标基因上所述待测样本的各个目标区间的测序数据的标准差、以及在相应目标基因上述与待测样本最接近的参考样本组的目标区间上的测序数据的均值和标准差，通过 T 检验的方法确认 p 值 p_{ttest}。

15. 如权利要求 14 所述的方法，

所述步骤（S4）通过以下内容确定所述待测样本的目标基因的拷贝数状态：

当 $CN_g \geq CN_{thA}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数扩增；

当 $CN_g \leq CN_{thD}$，$sigRatio \geq sigRatio_{th}$，且 $p_{ttest} \leq p_{th}$ 时，确认所述待测样本的目标基因发生拷贝数缺失；

当 $CN_{thA} < CN_g < CN_{thD}$，或 $sigRatio < sigRatio_{th}$，或 $p_{ttest} > p_{th}$ 时，确认所述待测样本的目标基因拷贝数正常，其中 CN_{thA}，CN_{thD}，$sigRatio_{th}$，和 p_{th} 各自独立地为阈值；优选地，其中 CN_{thA} 为约 2.25 至约 4；优选地，其中 CN_{thD} 为约 1.0 至约 1.75；优选地，其中 $sigRatio_{th}$ 为约 0.3 至约 1；优选地，其中 p_{th} 为约 0.05 至约 0.00001。

16. 如权利要求 2-15 中任一项所述的方法，

所述目标基因包含选自以下组基因：ABL1、ABL2、ABRAXAS1、ACVR1、ACVR1B、AKT1、AKT2、AKT3、ALK、ALOX12B、AMER1、APC、AR、ARAF、ARFRP1、ARID1A、ARID1B、ARID2、ARID5B、ASXL1、ASXL2、ASXL3、

ATG5、ATM、ATR、ATRX、AURKA、AURKB、AXIN1、AXIN2、AXL、B2M、BAP1、BARD1、BBC3、BCL10、BCL2、BCL2L1、BCL2L11、BCL2L2、BCL6、BCOR、BCORL1、BIRC3、BLM、BMPR1A、BRAF、BRCA1、BRCA2、BRD4、BRD7、BRINP3、BRIP1、BTG1、BTG2、BTK、CALR、CARD11、CASP8、CBFB、CBL、CCND1、CCND2、CCND3、CCNE1、CD274、CD28、CD58、CD74、CD79A、CD79B、CDC73、CDH1、CDH18、CDK12、CDK4、CDK6、CDK8、CDKN1A、CDKN1B、CDKN1C、CDKN2A、CDKN2B、CDKN2C、CEBPA、CENPA、CHD1、CHD2、CHD4、CHD8、CHEK1、CHEK2、CIC、CIITA、CREBBP、CRKL、CRLF2、CRYBG1、CSF1R、CSF3R、CSMD1、CSMD3、CTCF、CTLA4、CTNNA1、CTNNB1、CUL3、CUL4A、CXCR4、CYLD、CYP17A1、CYP2D6、DAXX、DCUN1D1、DDR1、DDR2、DDX3X、DICER1、DIS3、DNAJB1、DNMT1、DNMT3A、DNMT3B、DOT1L、DPYD、DTX1、DUSP22、EED、EGFR、EIF1AX、EIF4E、EMSY、EP300、EPCAM、EPHA2、EPHA3、EPHA5、EPHA7、EPHB1、EPHB4、ERBB2、ERBB3、ERBB4、ERCC1、ERCC2、ERCC3、ERCC4、ERCC5、ERG、ERRFI1、ESR1、ETV4、ETV5、ETV6、EWSR1、EZH2、EZR、FANCA、FANCC、FANCD2、FANCE、FANCF、FANCG、FANCI、FANCL、FANCM、FAS、FAT1、FAT3、FBXW7、FGF10、FGF12、FGF14、FGF19、FGF23、FGF3、FGF4、FGF6、FGF7、FGFR1、FGFR2、FGFR3、FGFR4、FH、FLCN、FLT1、FLT3、FLT4、FOXA1、FOXL2、FOXO1、FOXO3、FOXP1、FRS2、FUBP1、FYN、GABRA6、GALNT12、GATA1、GATA2、GATA3、GATA4、GATA6、GEN1、GID4、GLI1、GNA11、GNA13、GNAQ、GNAS、GPS2、GREM1、GRIN2A、GRM3、GSK3B、H3F3A、H3F3B、H3F3C、HDAC1、HDAC2、HGF、HIST1H1C、HIST1H2BD、HIST1H3A、HIST1H3B、HIST1H3C、HIST1H3D、HIST1H3E、HIST1H3G、HIST1H3H、HIST1H3I、HIST1H3J、HIST2H3D、HIST3H3、HLA-A、HLA-B、HLA-C、HNF1A、HOXB13、HRAS、HSD3B1、HSP90AA1、ICOSLG、ID3、IDH1、IDH2、IFNGR1、IGF1、IGF1R、IGF2、IGHD、IGHJ、IGHV、IKBKE、IKZF1、IL10、IL7R、INHA、INHBA、INPP4A、INPP4B、INSR、IRF2、IRF4、IRS1、IRS2、ITK、ITPKB、JAK1、JAK2、JAK3、JUN、KAT6A、KDM5A、KDM5C、KDM6A、KDR、KEAP1、KEL、KIR2DL4、KIR3DL2、KIT、KLF4、KLHL6、KLRC1、KLRC2、KLRK1、KMT2A、KMT2C、KMT2D、KRAS、LATS1、LATS2、LMO1、

LRP1B、LTK、LYN、MAF、MAGI2、MALT1、MAP2K1、MAP2K2、MAP2K4、MAP3K1、MAP3K13、MAP3K14、MAPK1、MAPK3、MAX、MCL1、MDC1、MDM2、MDM4、MED12、MEF2B、MEN1、MERTK、MET、MFHAS1、MGA、MIR21、MITF、MKNK1、MLH1、MLH3、MPL、MRE11、MSH2、MSH3、MSH6、MST1、MST1R、MTAP、MTOR、MUTYH、MYC、MYCL、MYCN、MYD88、MYOD1、NAV3、NBN、NCOA3、NCOR1、NCOR2、NEGR1、NF1、NF2、NFE2L2、NFKBIA、NKX2-1、NKX3-1、NOTCH1、NOTCH2、NOTCH3、NOTCH4、NPM1、NRAS、NRG1、NSD1、NSD2、NSD3、NT5C2、NTHL1、NTRK1、NTRK2、NTRK3、NUP93、NUTM1、P2RY8、PAK1、PAK3、PAK5、PALB2、PALLD、PARP1、PARP2、PARP3、PAX5、PBRM1、PCDH11X、PDCD1、PDCD1LG2、PDGFRA、PDGFRB、PDK1、PGR、PHOX2B、PIK3C2B、PIK3C2G、PIK3C3、PIK3CA、PIK3CB、PIK3CD、PIK3CG、PIK3R1、PIK3R2、PIK3R3、PIM1、PLCG2、PLK2、PMS1、PMS2、PNRC1、POLD1、POLE、POM121L12、PPARG、PPM1D、PPP2R1A、PPP2R2A、PPP6C、PRDM1、PREX2、PRKAR1A、PRKCI、PRKDC、PRKN、PTCH1、PTEN、PTPN11、PTPRD、PTPRO、PTPRS、PTPRT、QKI、RAB35、RAC1、RAD21、RAD50、RAD51、RAD51B、RAD51C、RAD51D、RAD52、RAD54L、RAF1、RARA、RASA1、RB1、RBM10、RECQL4、REL、RET、RHEB、RHOA、RICTOR、RIT1、RNF43、ROS1、RPA1、RPS6KA4、RPS6KB2、RPTOR、RSPO2、RUNX1、RUNX1T1、SDC4、SDHA、SDHAF2、SDHB、SDHC、SDHD、SETD2、SF3B1、SGK1、SH2B3、SH2D1A、SHQ1、SLC34A2、SLIT2、SLX4、SMAD2、SMAD3、SMAD4、SMARCA4、SMARCB1、SMARCD1、SMO、SNCAIP、SOCS1、SOX10、SOX17、SOX2、SOX9、SPEN、SPI1、SPOP、SPTA1、SRC、SRSF2、STAG2、STAT3、STAT4、STAT5A、STAT5B、STAT6、STK11、STK40、SUFU、SYK、TAF1、TBX21、TBX3、TCF3、TCF7L2、TEK、TENT5C、TERC、TERT、TET1、TET2、TGFBR1、TGFBR2、TIPARP、TMEM127、TMPRSS2、TNFAIP3、TNFRSF14、TOP1、TOP2A、TP53、TP63、TP73、TRAF2、TRAF3、TRAF7、TRIM58、TRPC5、TSC1、TSC2、TSHR、TYRO3、U2AF1、UGT1A1、VEGFA、VEGFB、VEGFC、VHL、WISP3、WRN、WT1、XIAP、XPO1、XRCC2、XRCC3、YAP1、YES1、ZAP70、ZBTB16、ZBTB2、ZNF217、ZNF703 和 ZNRF3。

17. 一种拷贝数状态分析装置，包含以下模块：接收模块，用于获取待测样本组的测序数据；确定模块，用于确定待测样本中的目标基因；判断模块，用于根据所述待测样本组的测序数据确定所述待测样本中的目标基因的拷贝数状态。

18. 如权利要求 17 所述的拷贝数状态分析装置，包含以下模块：（M1）接收模块，用于获取待测样本的测序数据和/或多个参考样本的测序数据；（M2）处理模块，用于将所述参考样本分为两个或以上参考样本组；（M3）计算模块，用于确定与所述待测样本最接近的参考样本组；（M4）判断模块，用于基于所述与待测样本最接近的参考样本组的测序数据，确定所述待测样本的目标基因的拷贝数状态。

19. 一种储存介质，其记载可以运行权利要求 1-16 中任一项所述的方法的程序。

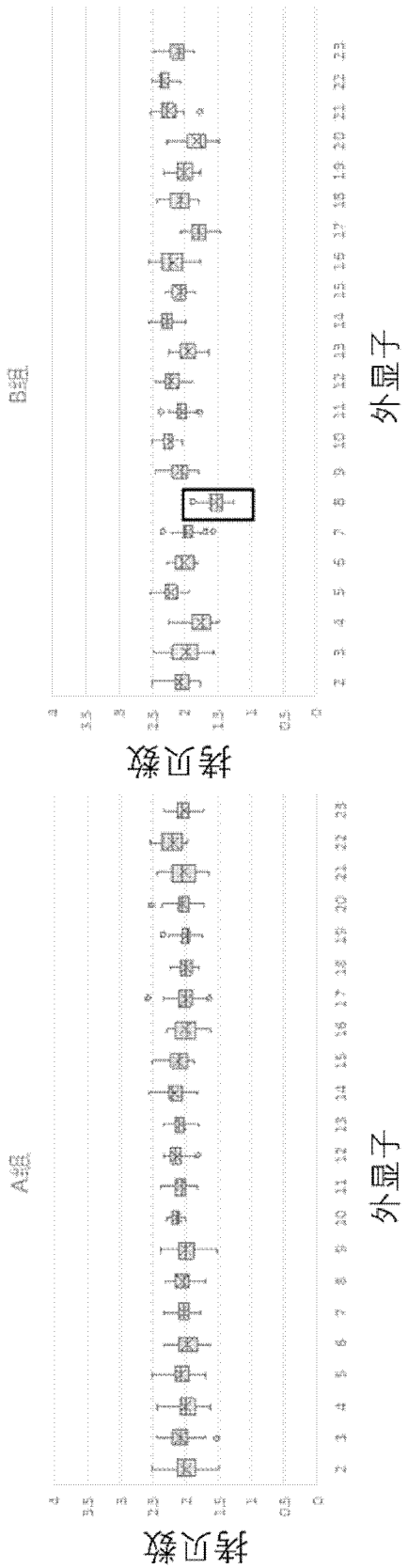


图 1A

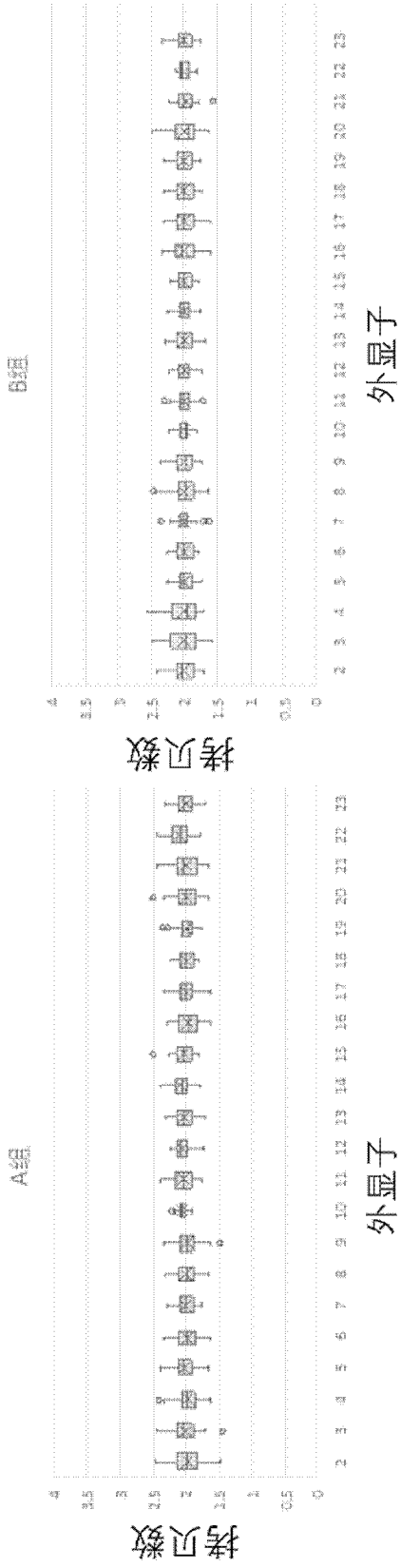


图 1B

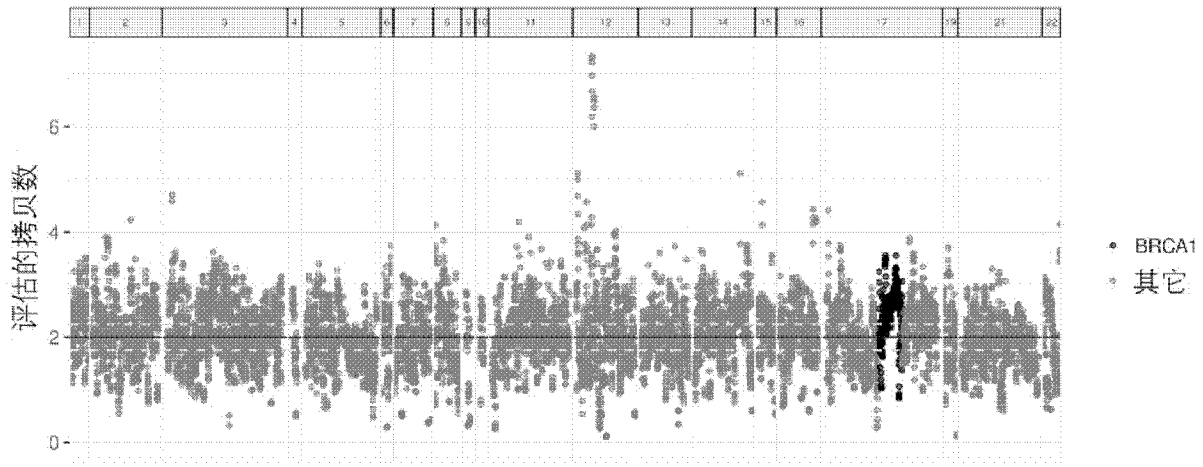


图 2A

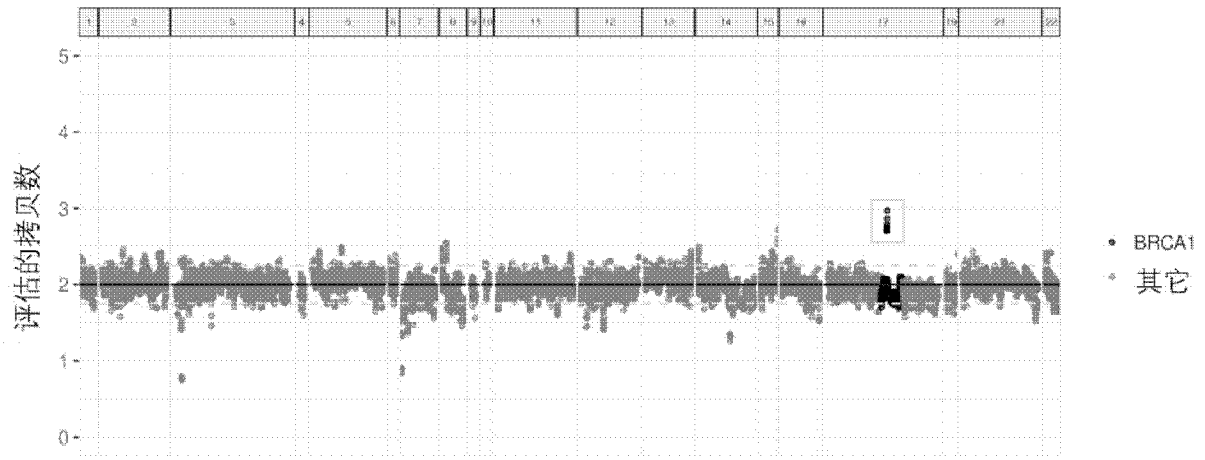


图 2B

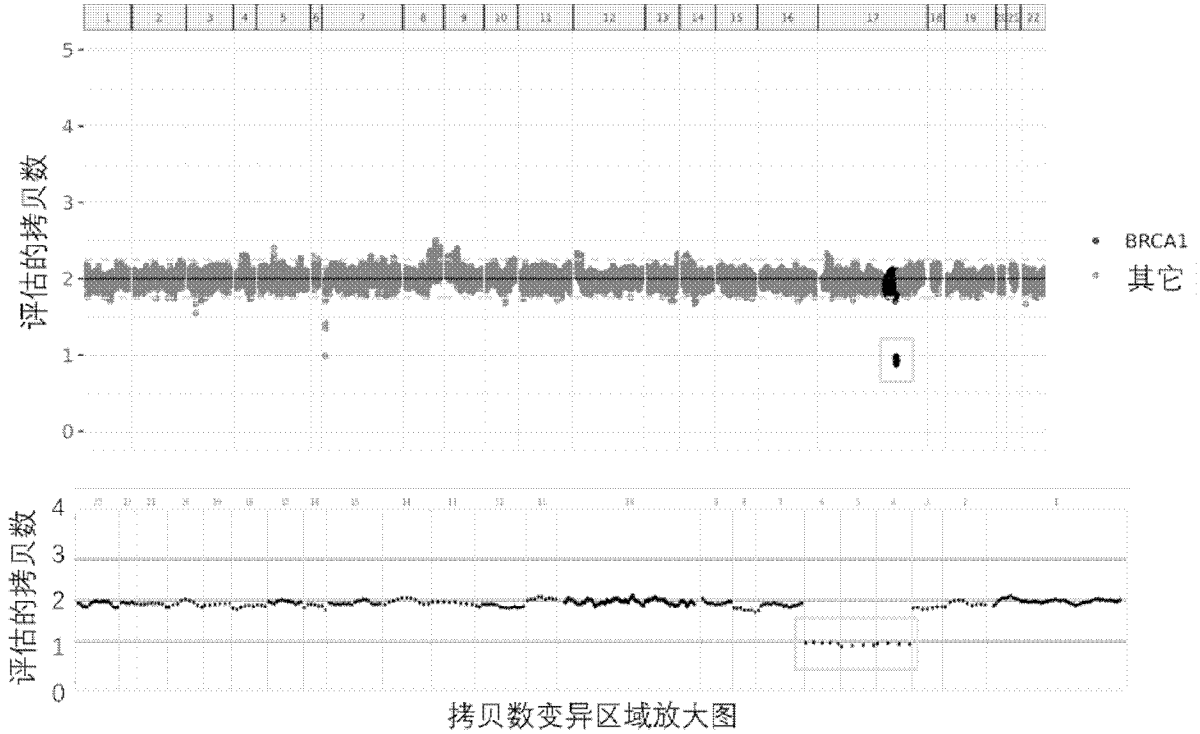

评估的拷贝数
BRCA1
其它
评估的拷贝数
拷贝数变异区域放大图

图 3A

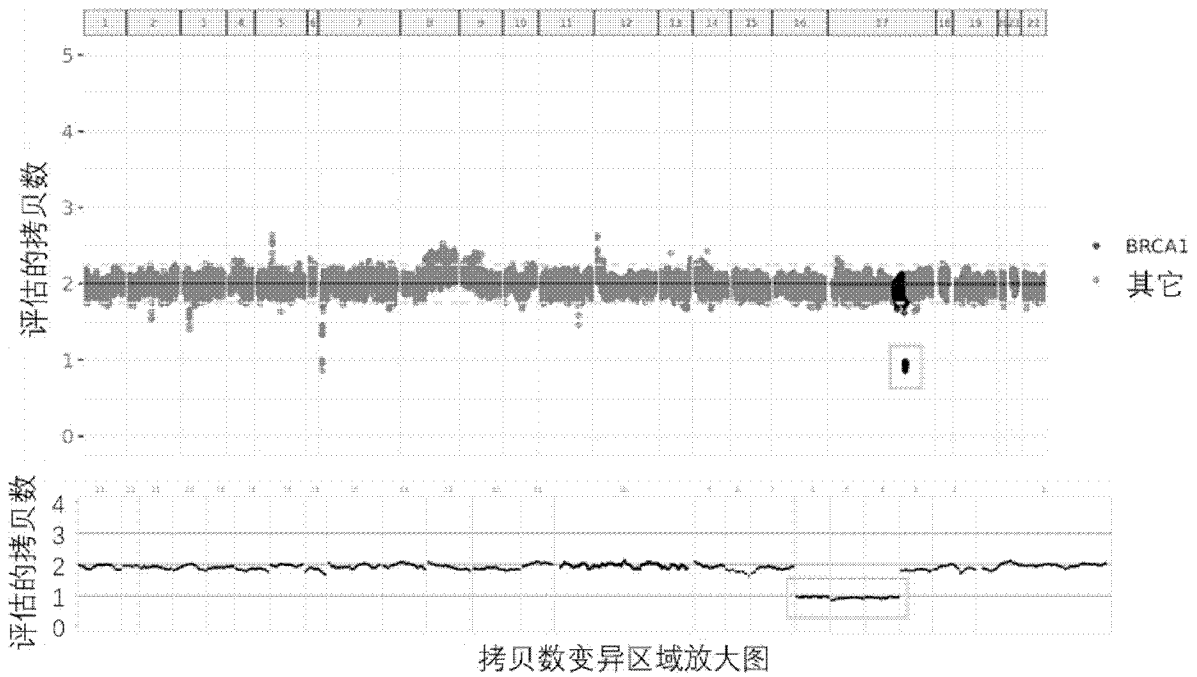
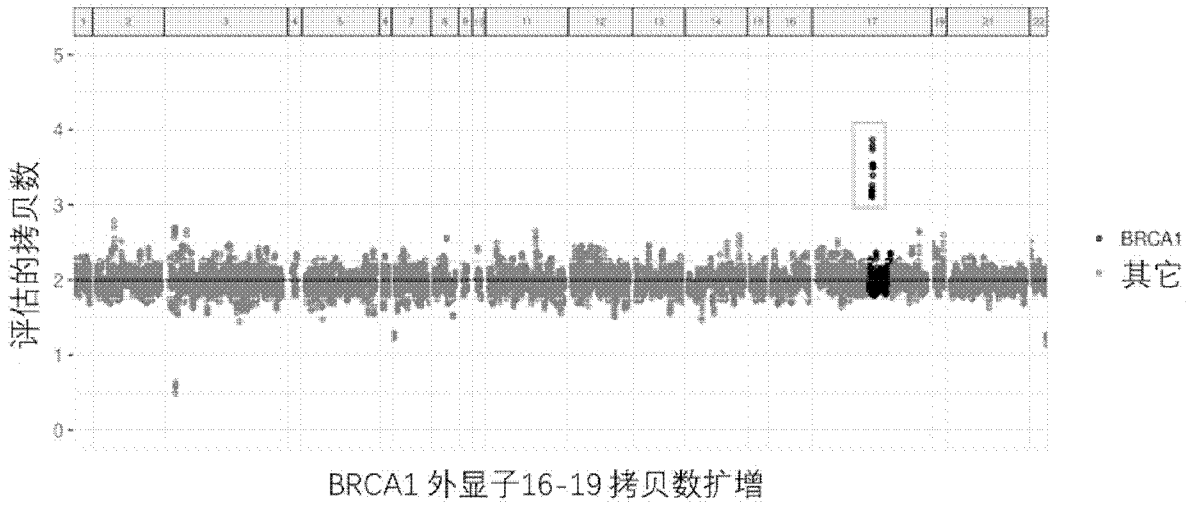
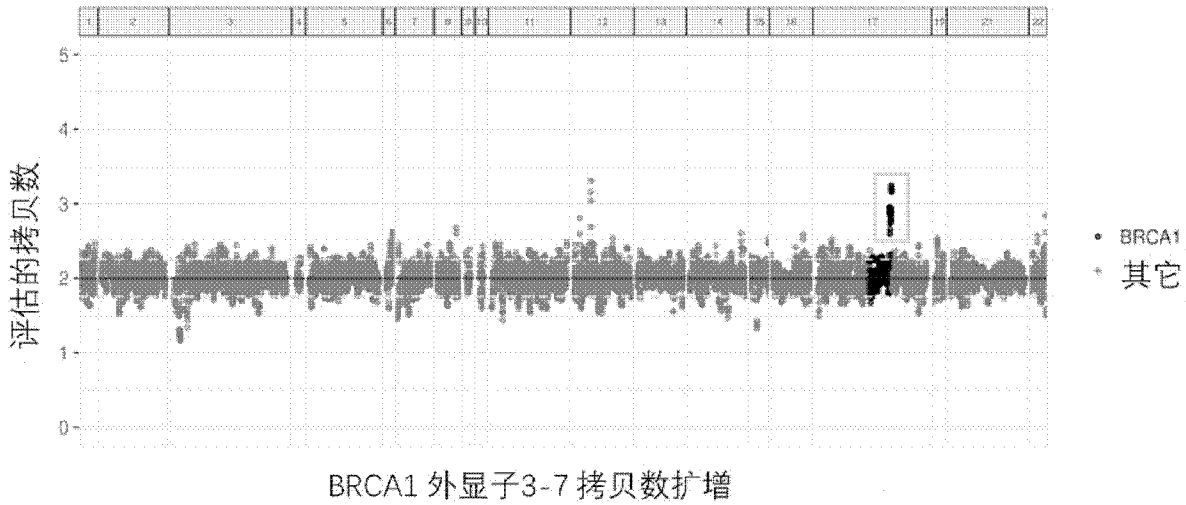

评估的拷贝数
BRCA1
其它
评估的拷贝数
拷贝数变异区域放大图

图 3B



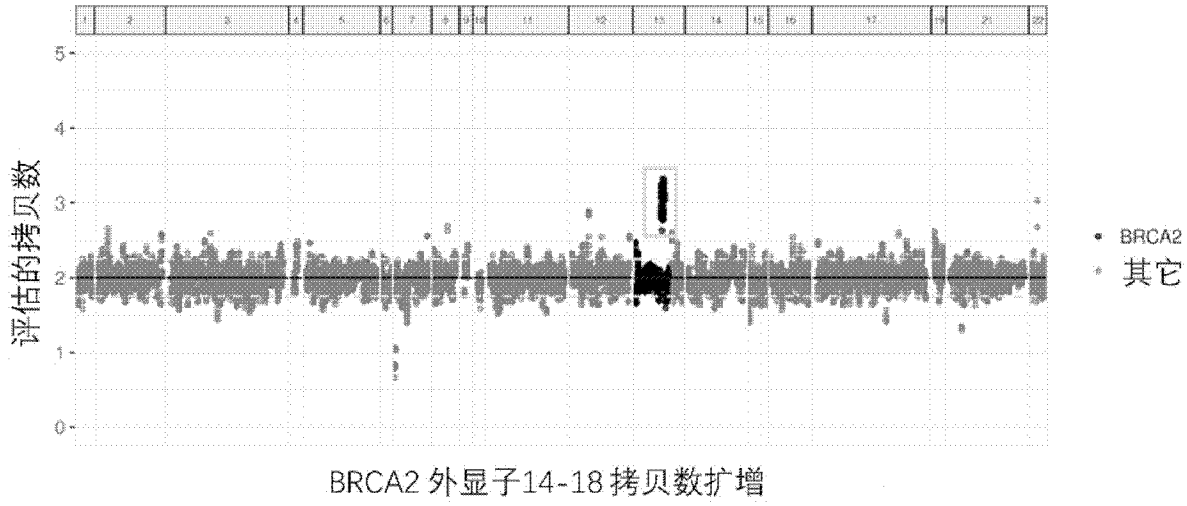
BRCA1 外显子16-19 拷贝数扩增

图 4A



BRCA1 外显子3-7 拷贝数扩增

图 4B



BRCA2 外显子14-18 拷贝数扩增

图 4C

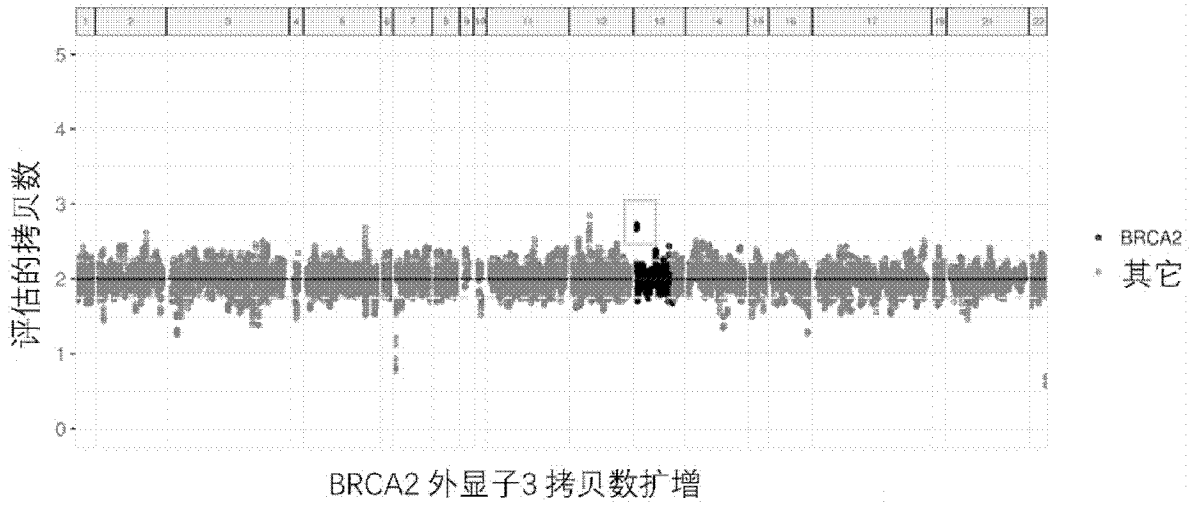

评估的拷贝数
BRCA2
其它
BRCA2 外显子3 拷贝数扩增

图 4D

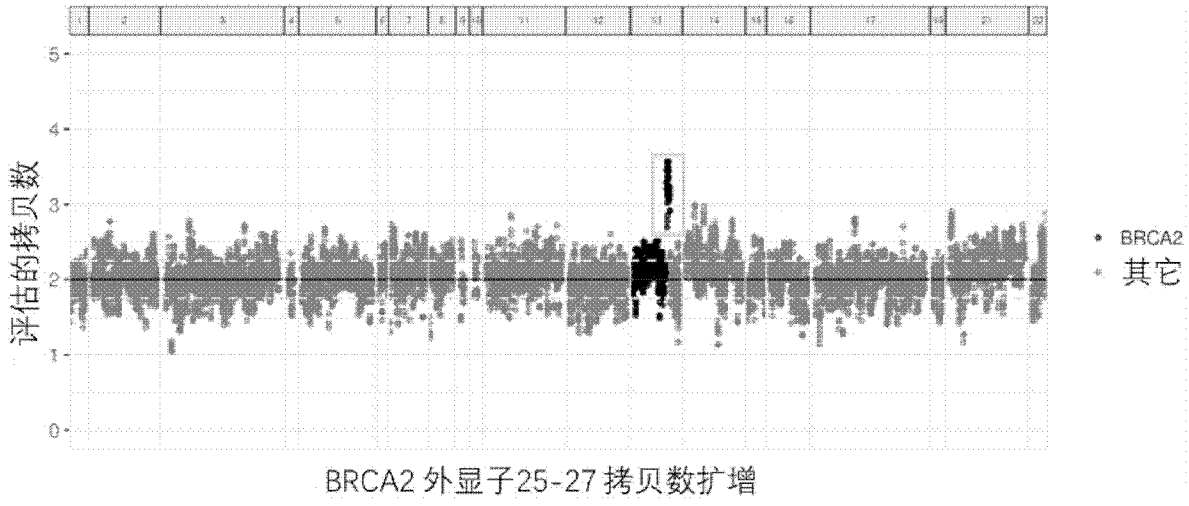

评估的拷贝数
BRCA2
其它
BRCA2 外显子25-27 拷贝数扩增

图 4E

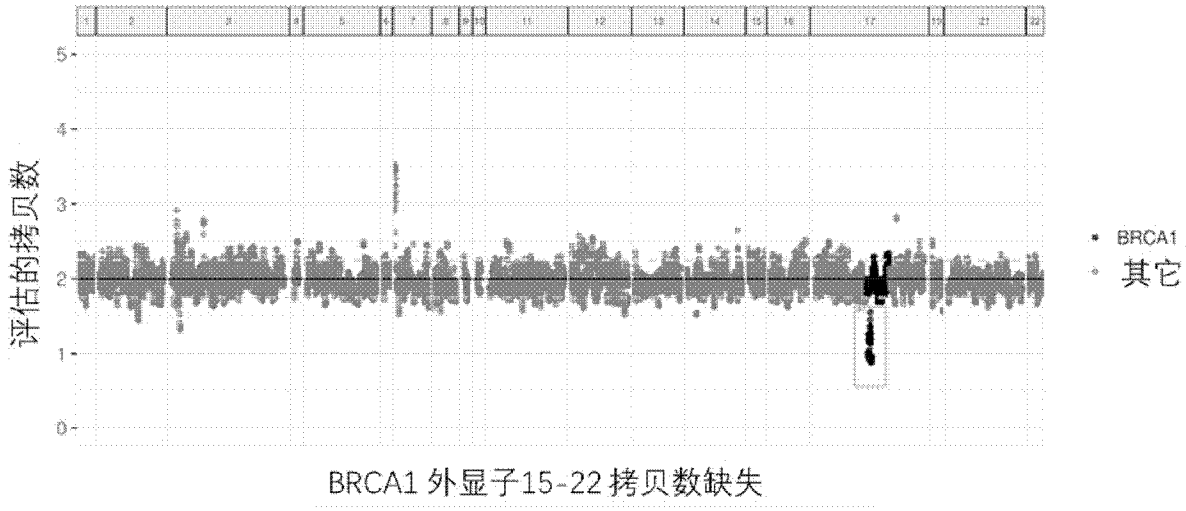

评估的拷贝数
BRCA1
其它
BRCA1 外显子15-22 拷贝数缺失

图 4F

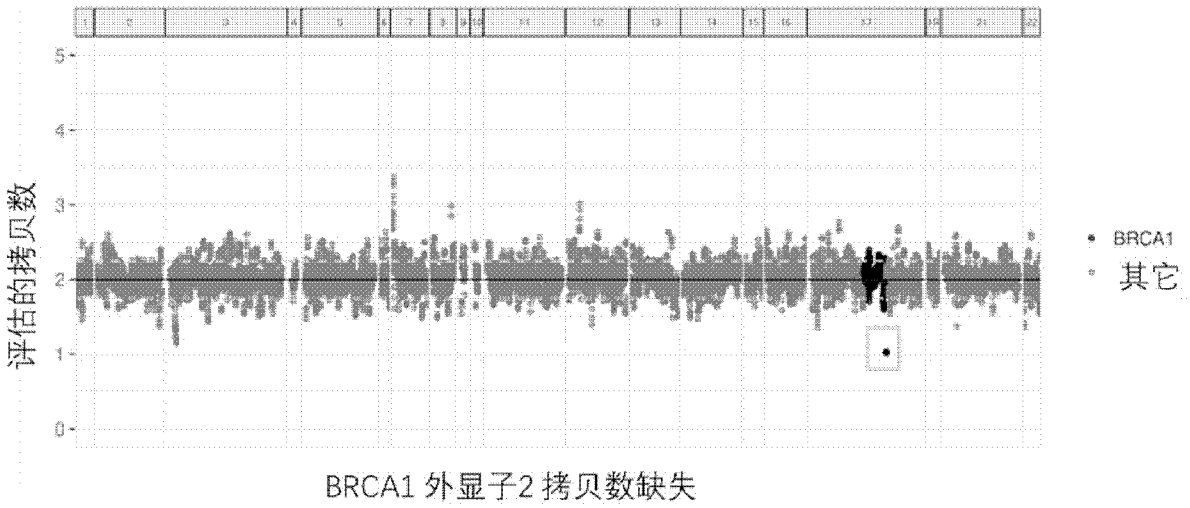

评估的拷贝数
BRCA1
其它
BRCA1 外显子2 拷贝数缺失

图 4G

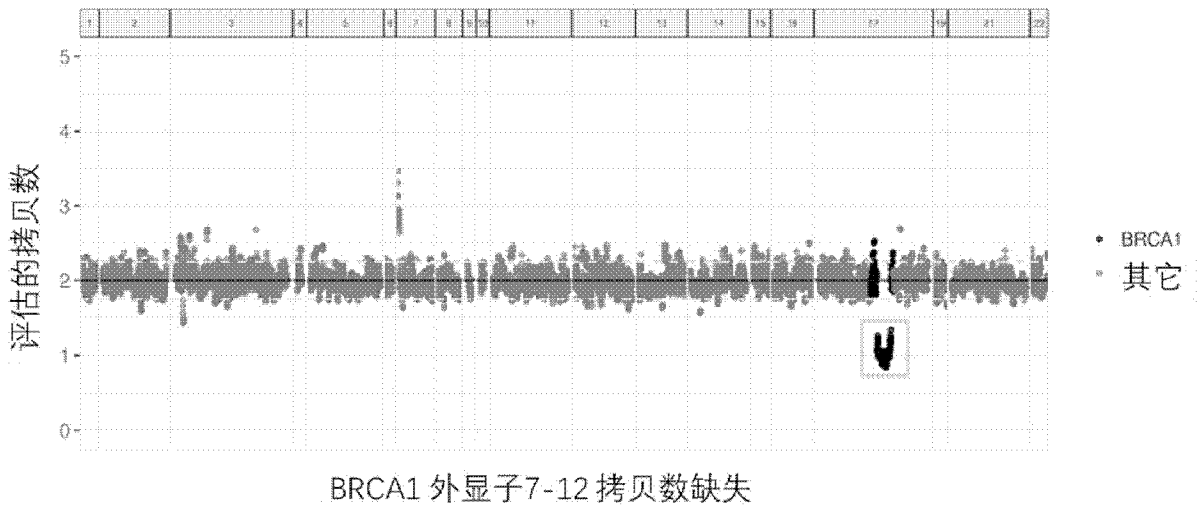

评估的拷贝数
BRCA1
其它
BRCA1 外显子7-12 拷贝数缺失

图 4H

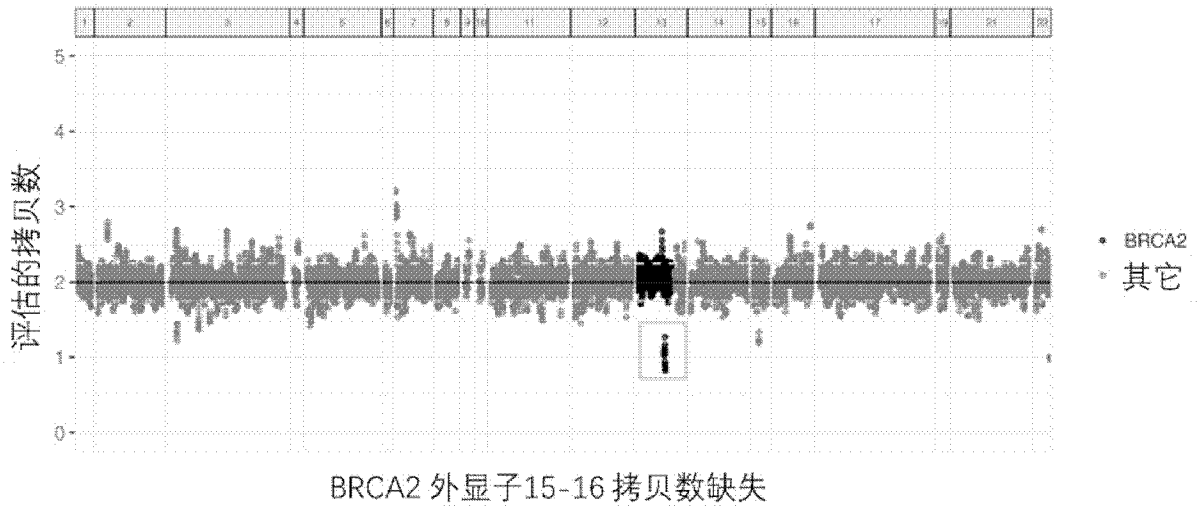

评估的拷贝数
BRCA2
其它
BRCA2 外显子15-16 拷贝数缺失

图 4I

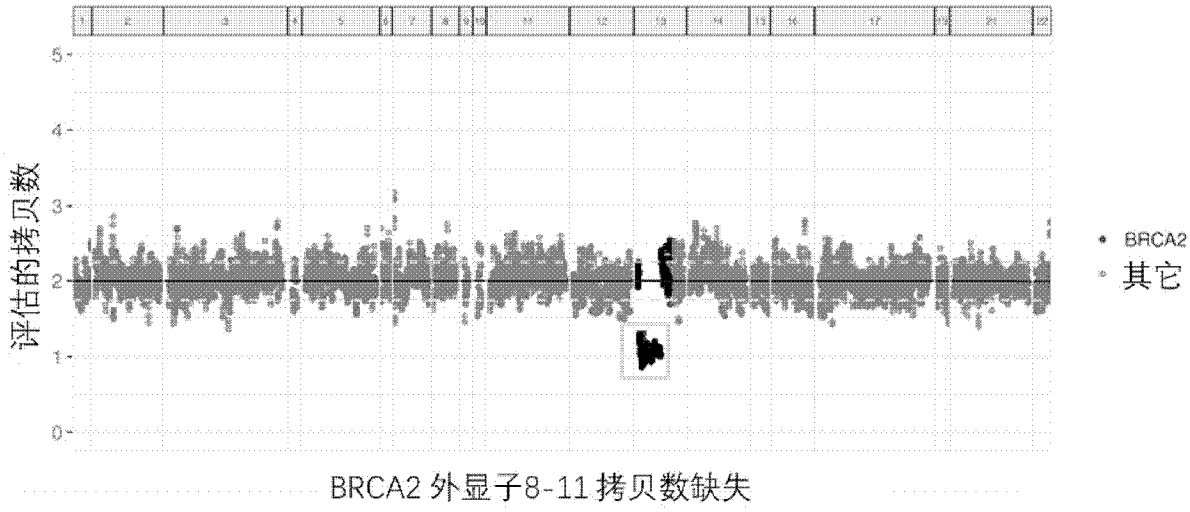


图 4J

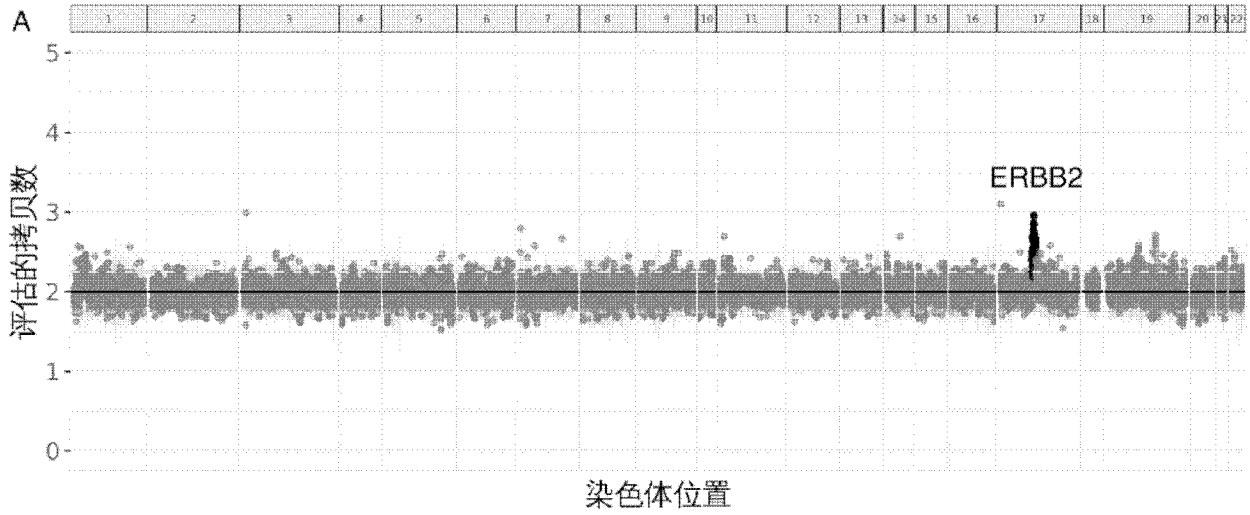

A
评估的拷贝数
ERBB2
染色体位置

图 5A

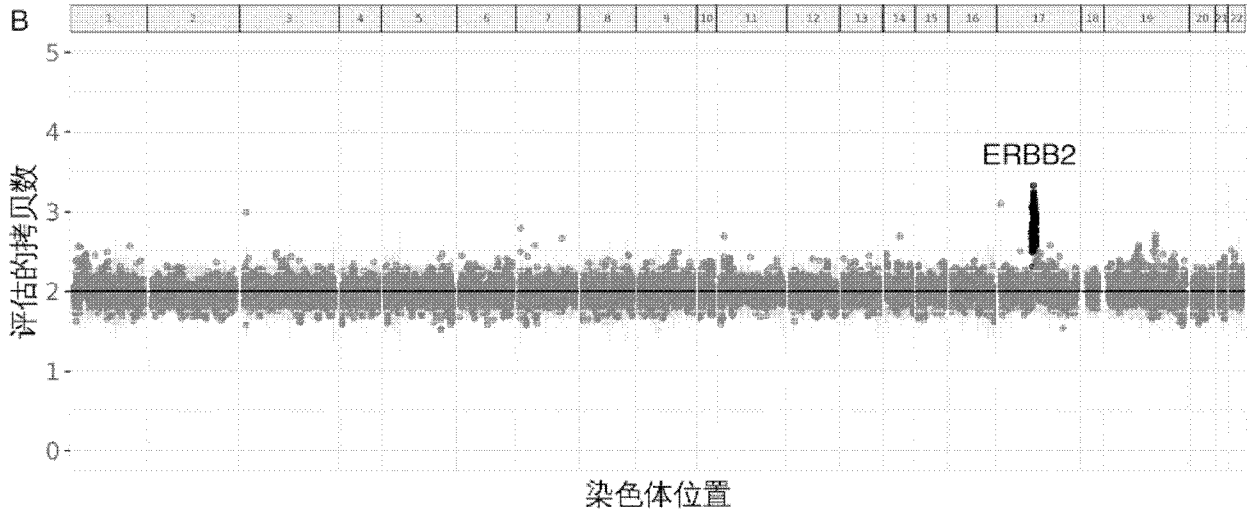

B
评估的拷贝数
ERBB2
染色体位置

图 5B

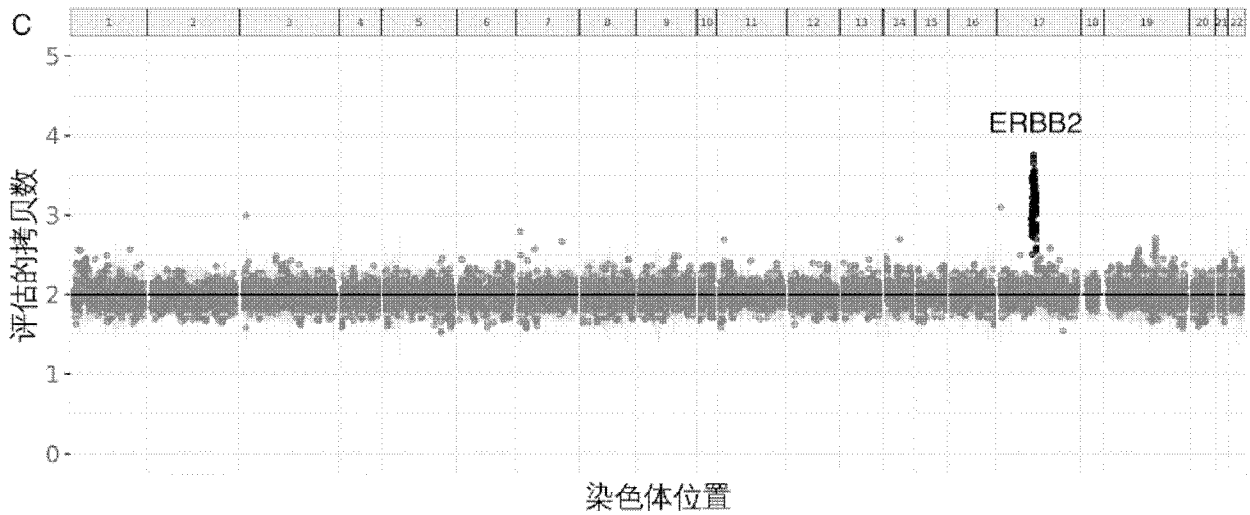

C
评估的拷贝数
ERBB2
染色体位置

图 5C

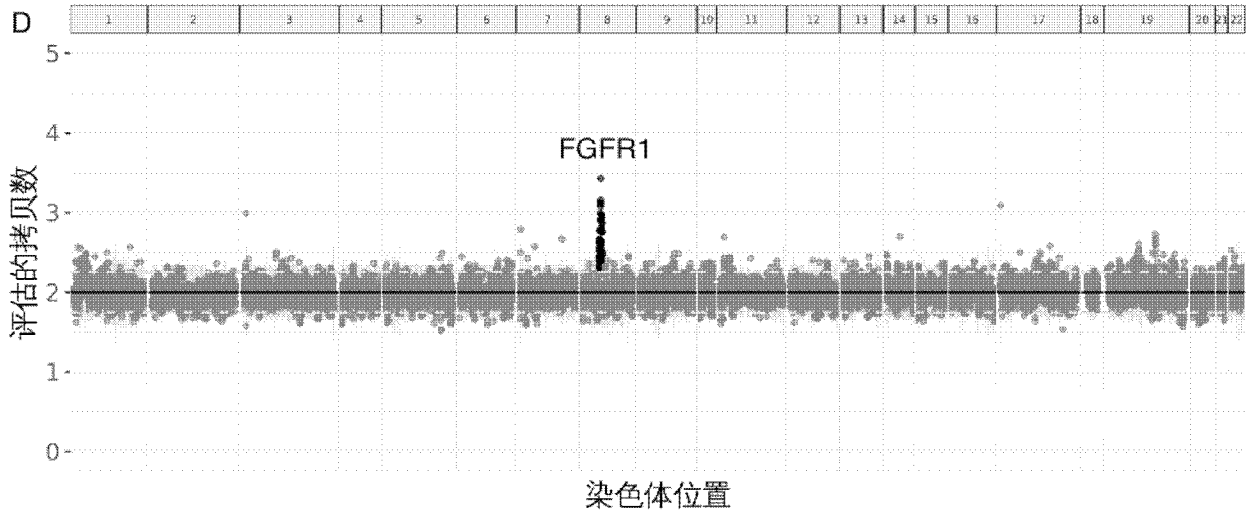

D
FGFR1
评估的拷贝数
染色体位置

图 5D

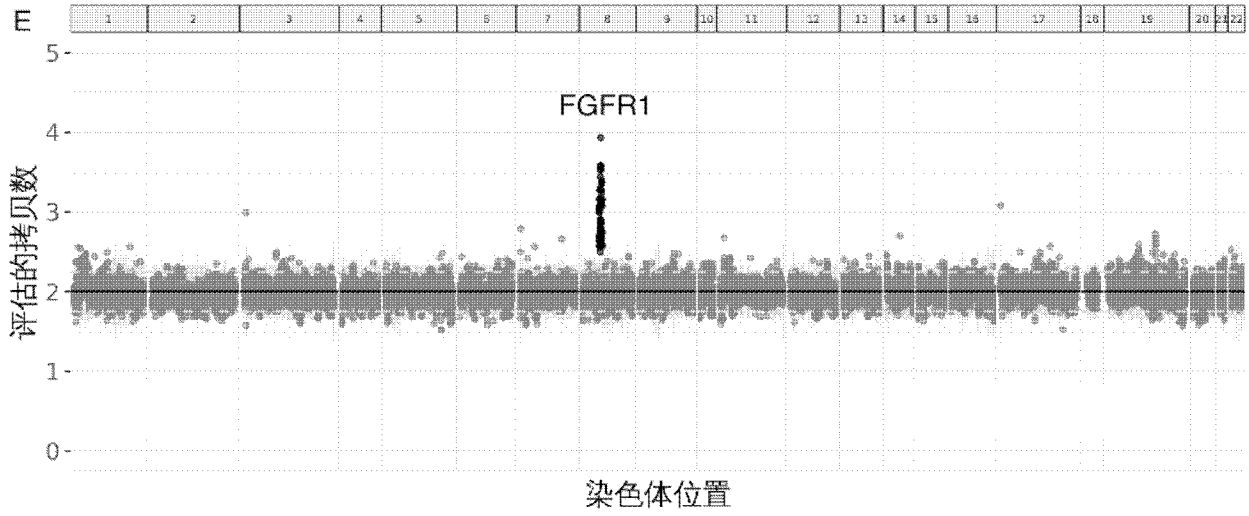

E
FGFR1
评估的拷贝数
染色体位置

图 5E

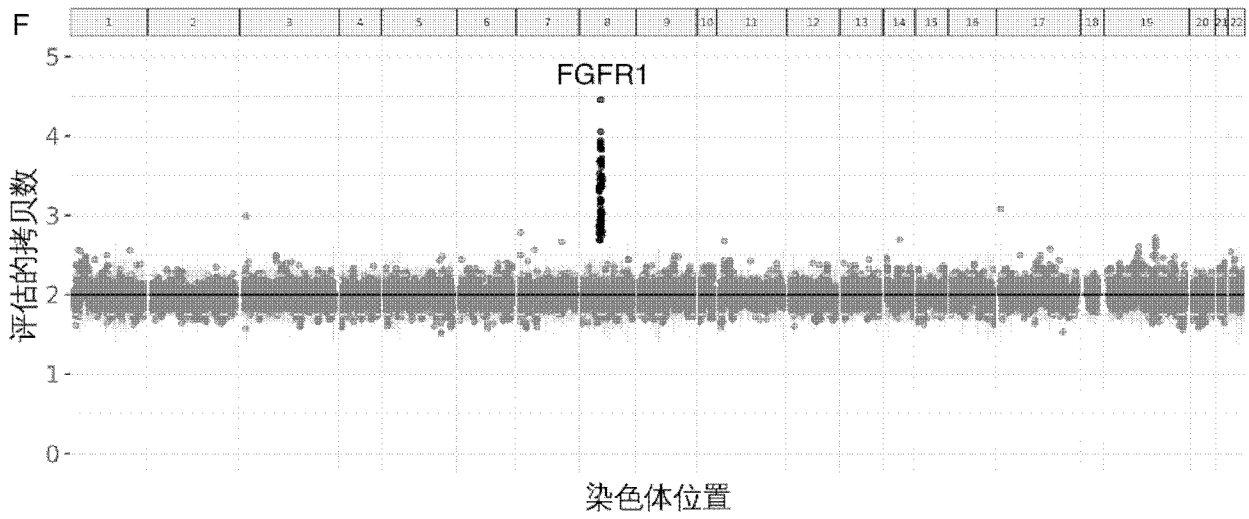

F
FGFR1
评估的拷贝数
染色体位置

图 5F

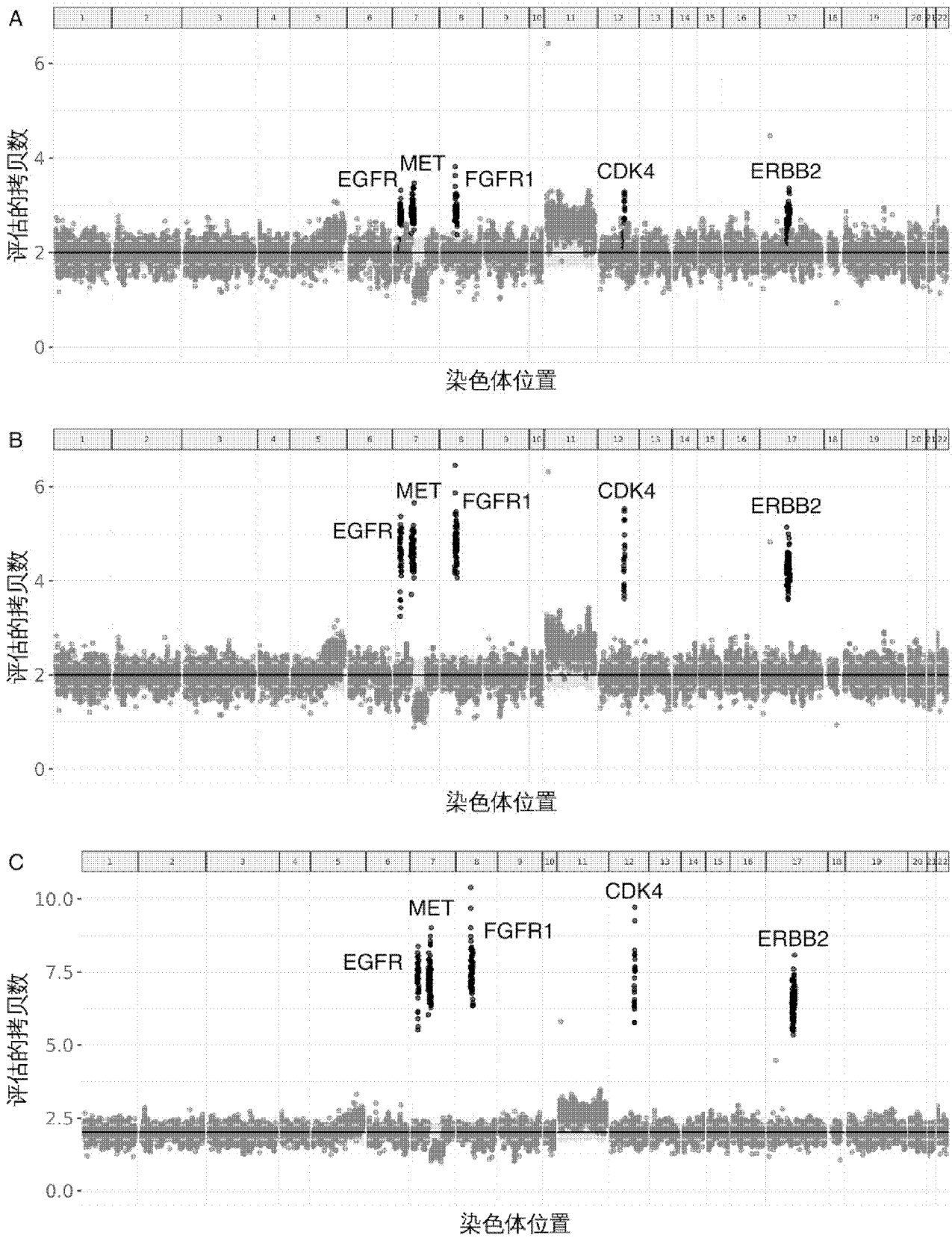

A
评估的拷贝数
染色体位置
EGFR
MET
FGFR1
CDK4
ERBB2
B
评估的拷贝数
染色体位置
EGFR
MET
FGFR1
CDK4
ERBB2
C
评估的拷贝数
染色体位置
EGFR
MET
FGFR1
CDK4
ERBB2

图 6

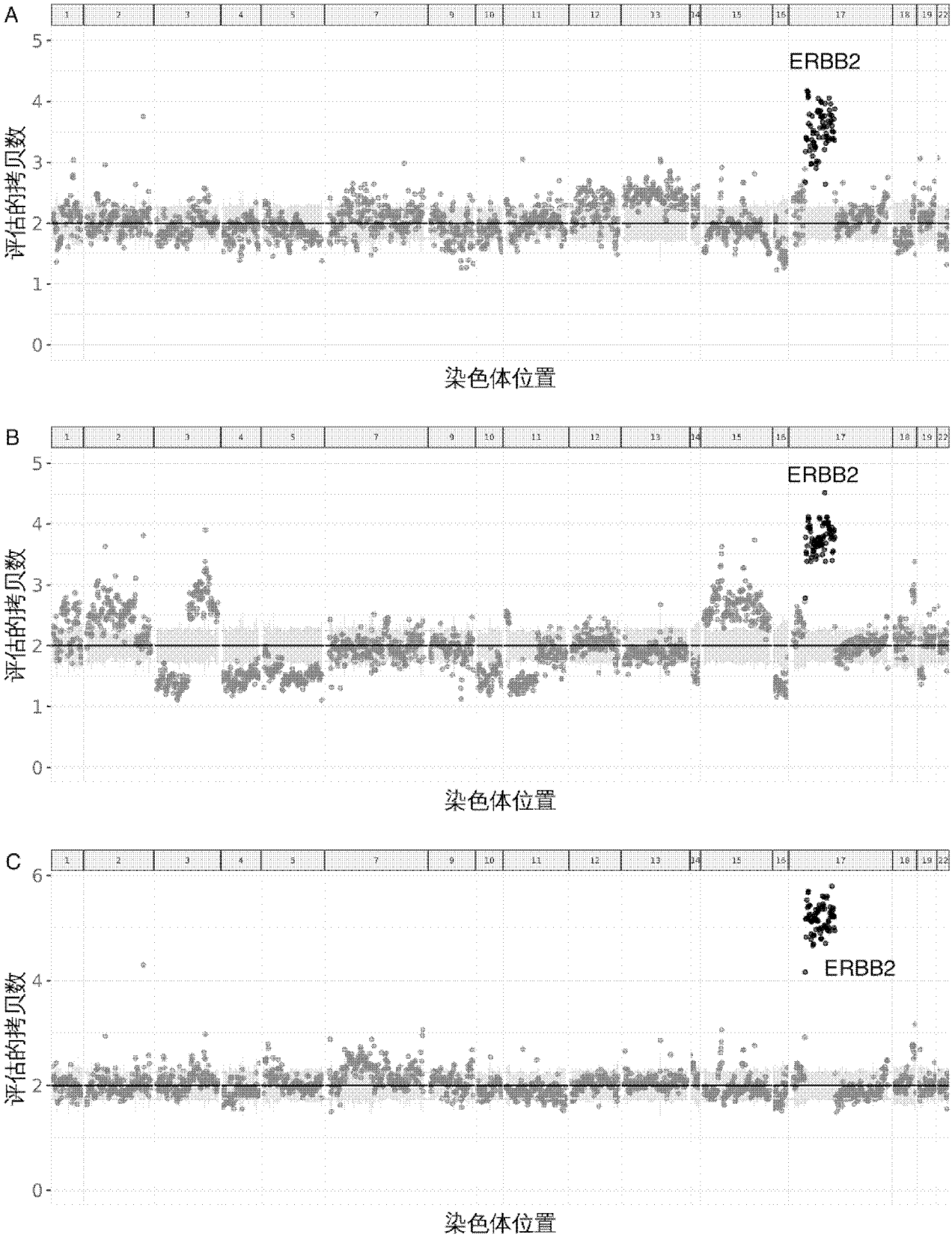


图 7

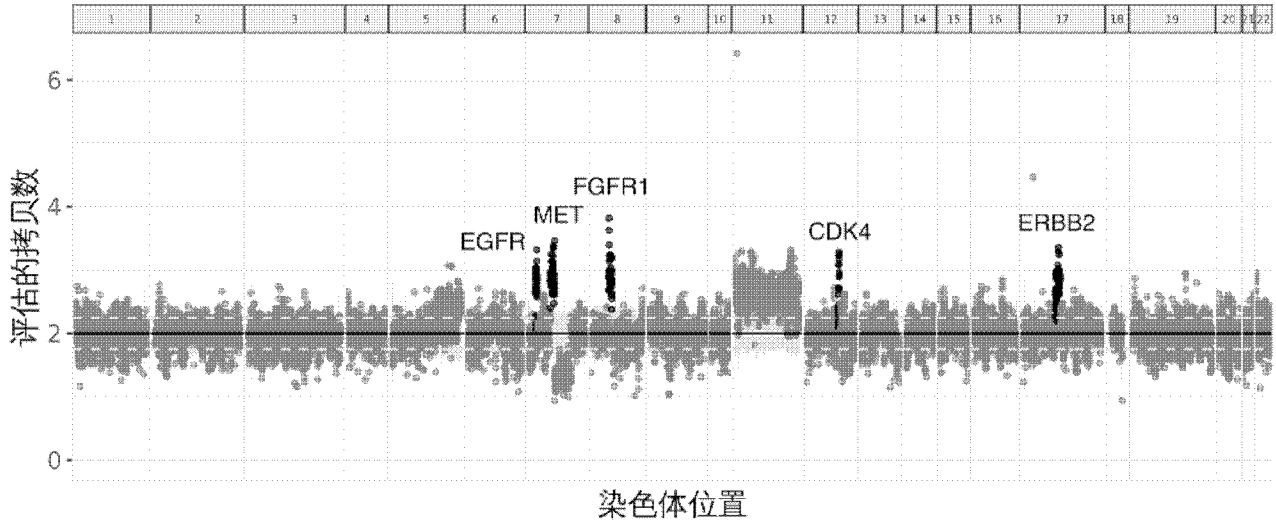

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
6
4
2
0
评估的拷贝数
EGFR
MET
FGFR1
CDK4
ERBB2
染色体位置

图 8A

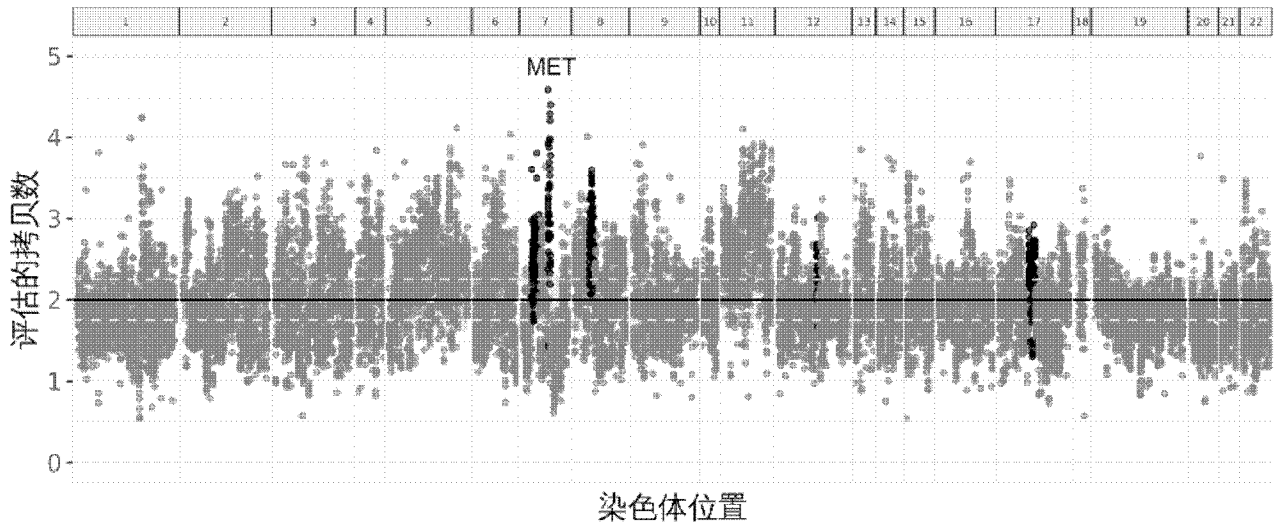

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
5
4
3
2
1
0
评估的拷贝数
MET
染色体位置

图 8B

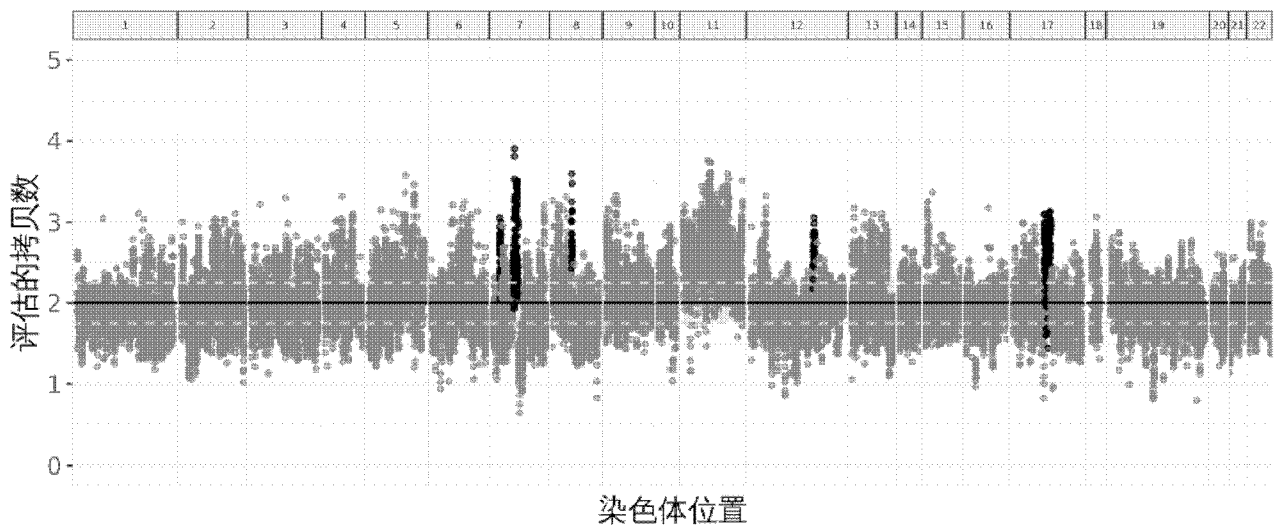

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
5
4
3
2
1
0
评估的拷贝数
染色体位置

图 8C

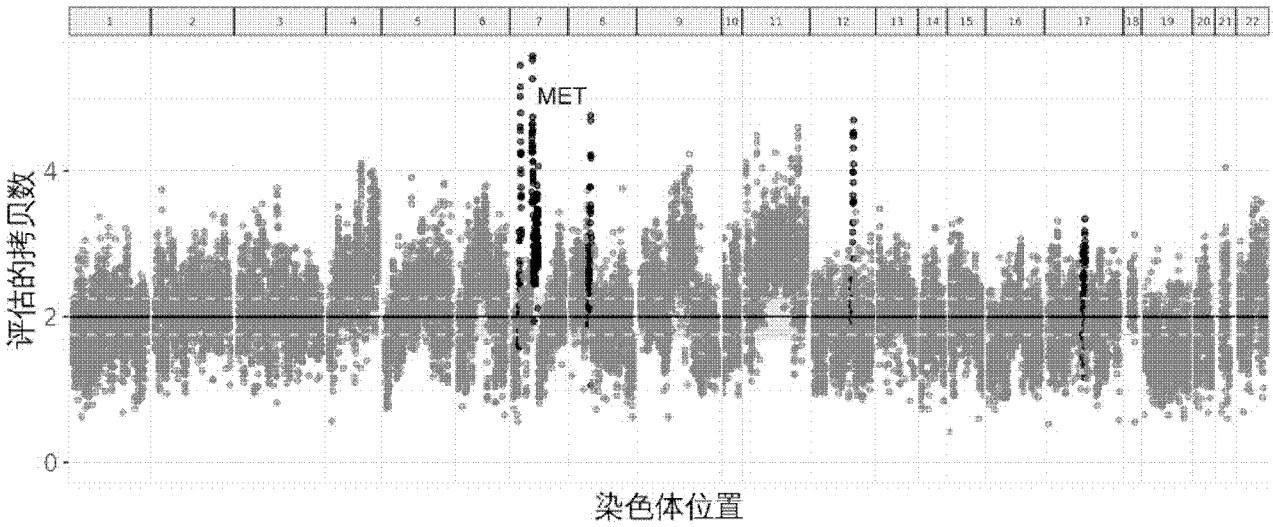


图 8D

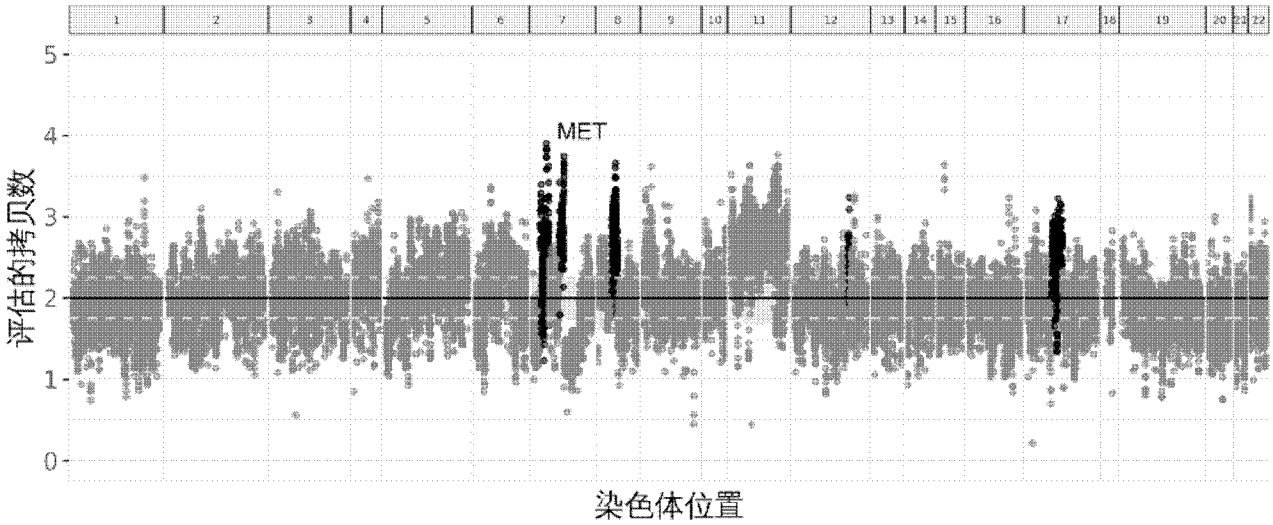


图 8E

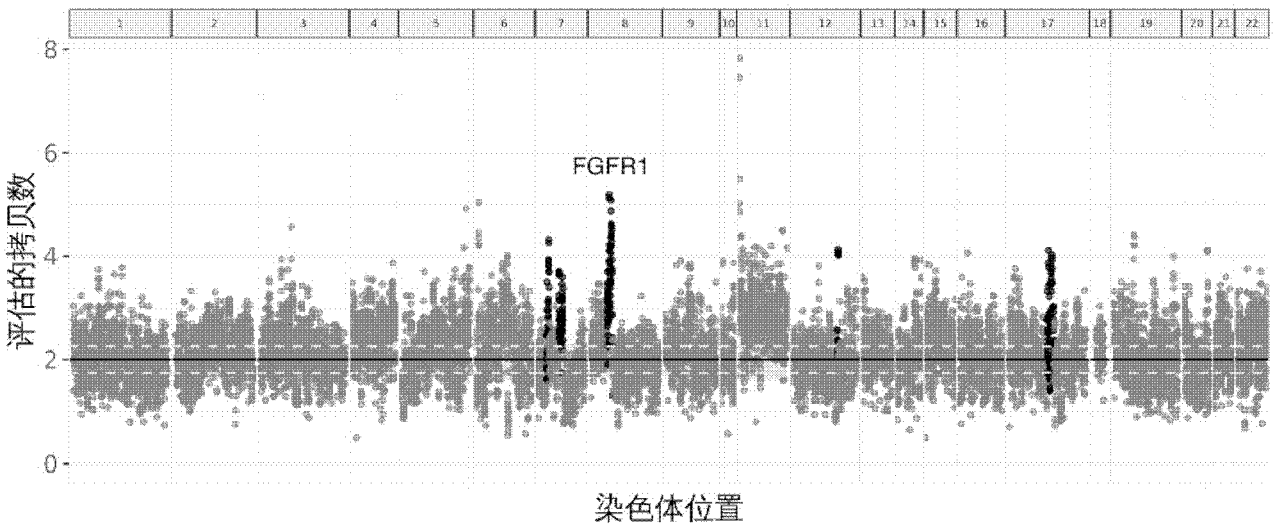


图 8F

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/115447

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/6869(2018.01)i; G16B 20/20(2019.01)i; G16B 20/30(2019.01)i; G16B 30/10(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C12Q; G16B

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)

CNTXT, CNABS, DWPI, WPABS, VEN, ENTXT, USTXT, CNKI, CJFD, IEEE: 拷贝数, 变异, 基因, 扩增, 缺失, 窗口, 均值, 标准差, 中位数, 距离, 相似度, 测序数据, 显著性, 检验, 校正, 矫正, 比对, 对照, 比较, 概率, 比例, 区间, 归一化, Copy number, Variation, gene, amplificat+, delet+, window?, mean, standard deviation, median, distance, similarity, sequencing data, significance, test+, correct+, alignment, control+, compar+, probability, proportion, interval, Normalizat+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113789371 A (GUANGZHOU BURNING ROCK MEDICAL EXAMINATION INSTITUTE CO., LTD.) 14 December 2021 (2021-12-14) paragraphs 45-85	1, 3-7, 19
PX	CN 113674803 A (GUANGZHOU BURNING ROCK MEDICAL EXAMINATION INSTITUTE CO., LTD.) 19 November 2021 (2021-11-19) claim 8, and paragraphs 163-168	17-18
X	CN 108256292 A (ANNOROAD GENE TECHNOLOGY (BEIJING) CO., LTD. et al.) 06 July 2018 (2018-07-06) paragraphs 50-112	1-19
X	CN 106650312 A (ANNOROAD GENE TECHNOLOGY (BEIJING) CO., LTD.) 10 May 2017 (2017-05-10) paragraphs 7-59	1-19
X	US 2013316915 A1 (HALPERN AARON et al.) 28 November 2013 (2013-11-28) paragraphs 45-274	1-19

☑ 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
19 October 2022	**25 November 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111968701 A (BEIJING GENE+ TECHNOLOGY CO., LTD. et al.) 20 November 2020 (2020-11-20) paragraphs 53-71	17-18
Y	CN 113789371 A (GUANGZHOU BURNING ROCK MEDICAL EXAMINATION INSTITUTE CO., LTD.) 14 December 2021 (2021-12-14) paragraphs 45-85	2, 8-16, 19
Y	CN 113674803 A (GUANGZHOU BURNING ROCK MEDICAL EXAMINATION INSTITUTE CO., LTD.) 19 November 2021 (2021-11-19) claim 8, and paragraphs 163-168	2, 8-16, 19
A	CN 106951737 A (SOUTHERN MEDICAL UNIVERSITY) 14 July 2017 (2017-07-14) entire document	1-19
A	CN 108427864 A (GENESEEQ TECHNOLOGY INC.) 21 August 2018 (2018-08-21) entire document	1-19
A	CN 104133914 A (XIAMEN VANGENES BIOTECHNOLOGY CO., LTD.) 05 November 2014 (2014-11-05) entire document	1-19
A	CN 106715711 A (SHENZHEN BGI GENOMICS CO., LTD.) 24 May 2017 (2017-05-24) entire document	1-19
A	US 2016300013 A1 (AGILENT TECHNOLOGIES INC.) 13 October 2016 (2016-10-13) entire document	1-19

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/115447

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113789371	A	14 December 2021	None			
CN	113674803	A	19 November 2021	None			
CN	108256292	A	06 July 2018	CN	108256292	B	02 November 2021
CN	106650312	A	10 May 2017	CN	106650312	B	17 May 2022
US	2013316915	A1	28 November 2013	US	2016117444	A1	28 April 2016
CN	111968701	A	20 November 2020	None			
CN	106951737	A	14 July 2017	CN	106951737	B	14 June 2019
CN	108427864	A	21 August 2018	WO	2019157791	A1	22 August 2019
				CN	108427864	B	29 January 2019
CN	104133914	A	05 November 2014	CN	104133914	B	08 March 2017
CN	106715711	A	24 May 2017	WO	2016000267	A1	07 January 2016
				CN	106715711	B	17 September 2021
US	2016300013	A1	13 October 2016	US	2021174901	A1	10 June 2021

国际检索报告

国际申请号
PCT/CN2022/115447

A. 主题的分类

C12Q 1/6869(2018.01)i; G16B 20/20(2019.01)i; G16B 20/30(2019.01)i; G16B 30/10(2019.01)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)

C12Q; G16B

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）)

CNTXT, CNABS, DWPI, WPABS, VEN, ENTXT, USTXT, CNKI, CJFD, IEEE:拷贝数，变异，基因，扩增，缺失，窗口，均值，标准差，中位数，距离，相似度，测序数据，显著性，检验，校正，矫正，比对，对照，比较，概率，比例，区间，归一化,Copy number, Variation, gene, amplificat+, delet+, window?, mean, standard deviation, median, distance, similarity, sequencing data, significance, test+, correct+, alignment, control+, compar+, probability, proportion, interval,Normalizat+

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
PX	CN 113789371 A (广州燃石医学检验所有限公司) 2021年12月14日 (2021-12-14) 第45-85段	1，3-7，19
PX	CN 113674803 A (广州燃石医学检验所有限公司) 2021年11月19日 (2021-11-19) 权利要求8，第163-168段	17-18
X	CN 108256292 A (安诺优达基因科技北京有限公司等) 2018年7月6日 (2018-07-06) 第50-112段	1-19
X	CN 106650312 A (安诺优达基因科技北京有限公司) 2017年5月10日 (2017-05-10) 第7-59段	1-19
X	US 2013316915 A1 (HALPERN AARON等) 2013年11月28日 (2013-11-28) 第45-274段	1-19
X	CN 111968701 A (北京吉因加科技有限公司等) 2020年11月20日 (2020-11-20) 第53-71段	17-18
Y	CN 113789371 A (广州燃石医学检验所有限公司) 2021年12月14日 (2021-12-14) 第45-85段	2，8-16，19

☑ 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
2022年10月19日	2022年11月25日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 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 陈安安 电话号码 (86-27)59371773

国际检索报告

国际申请号
PCT/CN2022/115447

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	CN 113674803 A (广州燃石医学检验所有限公司) 2021年11月19日 (2021-11-19) 权利要求8，第163-168段	2，8-16，19
A	CN 106951737 A (南方医科大学) 2017年7月14日 (2017-07-14) 全文	1-19
A	CN 108427864 A (南京世和基因生物技术有限公司) 2018年8月21日 (2018-08-21) 全文	1-19
A	CN 104133914 A (厦门万基生物科技有限公司) 2014年11月5日 (2014-11-05) 全文	1-19
A	CN 106715711 A (深圳华大基因股份有限公司) 2017年5月24日 (2017-05-24) 全文	1-19
A	US 2016300013 A1 (AGILENT TECHNOLOGIES INC) 2016年10月13日 (2016-10-13) 全文	1-19

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/115447

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	113789371	A	2021年12月14日	无			
CN	113674803	A	2021年11月19日	无			
CN	108256292	A	2018年7月6日	CN	108256292	B	2021年11月2日
CN	106650312	A	2017年5月10日	CN	106650312	B	2022年5月17日
US	2013316915	A1	2013年11月28日	US	2016117444	A1	2016年4月28日
CN	111968701	A	2020年11月20日	无			
CN	106951737	A	2017年7月14日	CN	106951737	B	2019年6月14日
CN	108427864	A	2018年8月21日	WO	2019157791	A1	2019年8月22日
				CN	108427864	B	2019年1月29日
CN	104133914	A	2014年11月5日	CN	104133914	B	2017年3月8日
CN	106715711	A	2017年5月24日	WO	2016000267	A1	2016年1月7日
				CN	106715711	B	2021年9月17日
US	2016300013	A1	2016年10月13日	US	2021174901	A1	2021年6月10日