

(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
2020 年 10 月 8 日 (08.10.2020)



(10) 国际公布号
WO 2020/199337 A1

- (51) 国际专利分类号:
G16B 20/20 (2019.01) **G16B 30/00** (2019.01)
- (21) 国际申请号: PCT/CN2019/089504
- (22) 国际申请日: 2019 年 5 月 31 日 (31.05.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910252747.9 2019 年 3 月 29 日 (29.03.2019) CN
- (71) 申请人: 北京市商汤科技开发有限公司(BEIJING SENSETIME TECHNOLOGY DEVELOPMENT CO., LTD.) [CN/CN]; 中国北京市海淀区中关村东路 1 号院 3 号楼 7 层 710-712 房间, Beijing 100084 (CN)。
- (72) 发明人: 胡志强(HU, Zhiqiang); 中国北京市海淀区中关村东路 1 号院 3 号楼 7 层 710-712 房间, Beijing 100084 (CN)。
- (74) 代理人: 北京林达刘知识产权代理事务所(普通合伙)(LINDA LIU & PARTNERS); 中国北京

市东城区北三环东路 36 号北京环球贸易中心 C 座 16 层, Beijing 100013 (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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, 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)。

(54) Title: GENOVARIATION IDENTIFICATION METHOD AND DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 一种基因变异识别方法、装置和存储介质

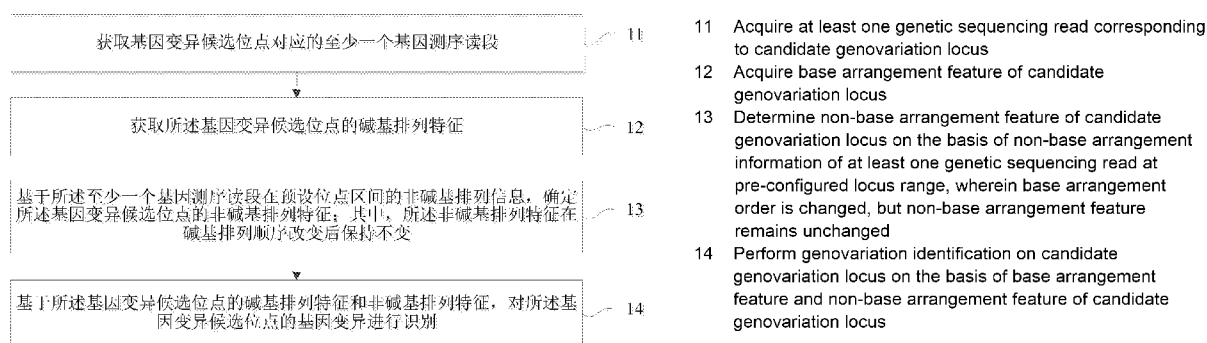


图 1

(57) Abstract: A genovariation identification method and device, and a storage medium. The method comprises: acquiring at least one genetic sequencing read corresponding to a candidate genovariation locus (11); acquiring a base arrangement feature of the candidate genovariation locus (12); determining a non-base arrangement feature of the candidate genovariation locus on the basis of non-base arrangement information of the at least one genetic sequencing read at a pre-configured locus range, wherein after changing a base arrangement order, the non-base arrangement feature remains unchanged (13); and performing genovariation identification on the candidate genovariation locus on the basis of the base arrangement feature and the non-base arrangement feature of the candidate genovariation locus (14). The method takes into account the fact that a non-base arrangement feature is not affected by a base arrangement order, thereby eliminating pseudo genovariations caused by germline genovariations and interference such as noise and errors, and improving the accuracy of genovariation identification.

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种基因变异识别方法、装置和存储介质，其中，该方法包括：获取基因变异候选位点对应的至少一个基因测序读段(11)；获取所述基因变异候选位点的碱基排列特征(12)；基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征；其中，所述非碱基排列特征在碱基排列顺序改变后保持不变(13)；基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别(14)。上述方法考虑非碱基排列特征不受碱基排列顺序制约的特点，更好地筛除由于胚系基因变异以及噪声、错误等干扰造成的伪基因变异，更好地对基因变异进行识别，提高基因变异识别的准确性。

一种基因变异识别方法、装置和存储介质

本公开要求在2019年3月29日提交中国专利局、申请号为201910252747.9、申请名称为“一种基因变异识别方法、装置和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

5 本公开涉及计算机技术领域，尤其涉及一种基因变异识别方法、装置和存储介质。

背景技术

随着生物技术的发展，通过基因测序技术可以测定人类基因的序列，碱基序列的分析可以作为进一步基因研究和改造的基础。目前，基因的二代测序技术相比于一代测试技术而言，极大地提高了基因测序的效率，降低了基因测序的成本，并且保持了基因测序的准确性。第一代测试技术如果完成
10 一个人类基因组的测序可能需要3年的时间，而使用二代测序技术则可以将时间缩短为仅仅1周。

发明内容

有鉴于此，本公开提出了一种基因变异识别技术方案。

根据本公开的一方面，提供了一种基因变异识别方法，所述方法包括：

获取基因变异候选位点对应的至少一个基因测序读段；获取所述基因变异候选位点的碱基排列特征；基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征；其中，所述非碱基排列特征在碱基排列顺序改变后保持不变；基于所述基因变异
15 候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别。

在一种可能的实现方式中，所述获取所述基因变异候选位点的碱基排列特征，包括：

确定所述基因变异候选位点所在的预设位点区间；根据参考基因组在所述预设位点区间的碱基排列信息，获取所述基因变异候选位点的碱基排列特征；其中，所述碱基排列特征用于表征碱基排列顺序。
20

在一种可能的实现方式中，所述基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息；基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。
25

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；根据所述预设位点区间中每个位点对应的第一基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。
30

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；在所述预设位点区间中的每个位点，确定所述第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量，作为第一基因测序读段的变异数量；根据所述第一基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。
35

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致

的第二基因测序读段；根据所述预设位点区间中每个位点对应的第二基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

- 5 在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；在所述预设位点区间中的每个位点，确定所述第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量，作为第二基因测序读段的变异数量；根据所述第二基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

- 10 在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；根据所述预设位点区间中每个位点对应的第三基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

- 15 在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

- 20 确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；在所述预设位点区间中的每个位点，确定所述第三基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量；根据所述第三基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

- 25 确定所述至少一个基因测序读段中来源于正常细胞的基因测序读段；基于所述正常细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

- 30 确定所述至少一个基因测序读段中来源于病变细胞的基因测序读段；基于所述病变细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别，包括：

- 35 根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别。

在一种可能的实现方式中，所述根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位

点的基因变异进行识别, 包括:

根据所述基因变异候选位点的特征矩阵, 得到所述基因变异候选位点的基因发生变异的变异值; 在所述变异值大于或等于预设阈值的情况下, 确定所述基因变异候选位点的基因存在变异。

在一种可能的实现方式中, 所述根据所述基因变异候选位点的碱基排列特征和非碱基排列特征, 得到所述基因变异候选位点的特征矩阵, 包括:

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征, 生成所述预设位点区间的每个第一维度特征的特征向量; 确定所述特征向量中碱基排列特征形成的碱基排列特征向量; 对所述碱基排列特征向量进行随机排序, 得到所述基因变异候选位点的特征矩阵。

在一种可能的实现方式中, 获取基因变异候选位点对应的至少一个基因测序读段, 包括:

获取由体细胞基因进行基因测序得到的基因测序读段; 将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对, 得到比对结果; 根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点; 获取所述基因变异候选位点对应的至少一个基因测序读段。

根据本公开的另一方面, 提供了一种基因变异识别装置, 所述装置包括:

第一获取模块, 用于获取基因变异候选位点对应的至少一个基因测序读段; 第二获取模块, 用于获取所述基因变异候选位点的碱基排列特征; 确定模块, 用于基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息, 确定所述基因变异候选位点的非碱基排列特征; 其中, 所述非碱基排列特征在碱基排列顺序改变后保持不变; 识别模块, 用于基于所述基因变异候选位点的碱基排列特征和非碱基排列特征, 对所述基因变异候选位点的基因变异进行识别。

在一种可能的实现方式中, 所述第二获取模块, 包括:

第一确定子模块, 用于确定所述基因变异候选位点所在的预设位点区间; 第二确定子模块, 用于根据参考基因组在所述预设位点区间的碱基排列信息, 获取所述基因变异候选位点的碱基排列特征; 其中, 所述碱基排列特征用于表征碱基排列顺序。

在一种可能的实现方式中, 所述确定模块, 包括:

第一获取子模块, 用于获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息; 第三确定子模块, 用于基于所述预设位点区间中每个位点的非碱基排列信息, 确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中, 所述第三确定子模块, 具体用于,

在所述基因测序读段中, 确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段; 根据所述预设位点区间中每个位点对应的第一基因测序读段的数量, 确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中, 所述第三确定子模块, 具体用于,

在所述基因测序读段中, 确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段; 在所述预设位点区间中的每个位点, 确定所述第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量, 作为第一基因测序读段的变异数量; 根据所述第一基因测序读段的变异数量, 确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中, 所述第三确定子模块, 具体用于,

在所述基因测序读段中, 确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段; 根据所述预设位点区间中每个位点对应的第二基因测序读段的数量, 确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；在所述预设位点区间中的每个位点，确定所述第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量，作为第二基因测序读段的变异数量；根据
5 所述第二基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；根据所述预设位点区间中每个位点对应的第三基因
10 测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；在所述预设位点区间中的每个位点，确定所述第三
15 基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量；根据所述第三基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述至少一个基因测序读段中来源于正常细胞的基因测序读段；基于所述正常细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列
20 特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述至少一个基因测序读段中来源于病变细胞的基因测序读段；基于所述病变细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列
25 特征。

在一种可能的实现方式中，所述识别模块，包括：

生成子模块，用于根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；识别子
30 模块，用于根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别。

在一种可能的实现方式中，所述识别子模块，具体用于，

根据所述基因变异候选位点的特征矩阵，得到所述基因变异候选位点的基因发生变异的变异值；在所述变异值大于或等于预设阈值的情况下，确定所述基因变异候选位点的基因存在变异。

在一种可能的实现方式中，所述生成子模块，具体用于，

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，生成所述预设位点区间的每个第一维度特征的特征向量；确定所述特征向量中碱基排列特征形成的碱基排列特征向量；对所述碱基排列特征向量进行随机排序，得到所述基因变异候选位点的特征矩阵。
35

在一种可能的实现方式中，所述第一获取模块，包括：

第二获取子模块，用于获取由体细胞基因进行基因测序得到的基因测序读段；对比子模块，用于

将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对,得到比对结果;第四确定子模块,用于根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点;第三获取子模块,用于获取所述基因变异候选位点对应的至少一个基因测序读段。

根据本公开的另一方面,提供了一种基因变异识别装置,包括:处理器;用于存储处理器可执行指令的存储器;其中,所述处理器被配置为执行上述方法。

根据本公开的另一方面,提供了一种非易失性计算机可读存储介质,其上存储有计算机程序指令,其中,所述计算机程序指令被处理器执行时实现上述方法。

本公开实施例提供的基因变异识别方案,可以获取基因变异候选位点对应的至少一个基因测序读段,获取基因变异候选位点的碱基排列特征,基于至少一个基因测序读段在预设位点区间的碱基排列信息,确定基因变异候选位点的非碱基排列特征,从而可以基于基因变异候选位点的碱基排列特征和非碱基排列特征,对基因变异候选位点的基因变异进行识别。这里,非碱基排列特征在碱基排列顺序改变后保持不变,即可以认为非碱基排列特征具有碱基排列不变性的性质,因此,在对基因变异候选位点的基因变异进行识别时,可以考虑基因变异候选位点的基因变异不受碱基排列顺序制约的特点,更好地筛选由于胚系基因变异以及噪声、错误等干扰造成的伪基因变异,从而可以更好地对基因变异进行识别,提高基因变异识别的准确性,

应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,而非限制本公开。

根据下面参考附图对示例性实施例的详细说明,本公开的其它特征及方面将变得清楚。

附图说明

包含在说明书中并且构成说明书的一部分的附图与说明书一起示出了本公开的示例性实施例、特征和方面,并且用于解释本公开的原理。

图1示出根据本公开一实施例的基因变异识别方法的流程图。

图2示出根据本公开一实施例的获取基因变异候选位点对应的至少一个基因测序读段的流程图。

图3示出根据本公开一实施例的基因变异候选位点的碱基排列特征过程的流程图。

图4示出根据本公开一实施例的基因变异候选位点的非碱基排列特征过程的流程图。

图5示出根据本公开一实施例的识别基因变异候选位点的基因变异过程的流程图。

图6示出根据本公开一实施例的得到基因变异候选位点的特征矩阵过程的流程图。

图7示出根据本公开一实施例的得到基因变异候选位点的特征矩阵过程的流程图。

图8示出根据本公开一实施例的得到基因变异候选位点的特征矩阵过程的流程图。

具体实施方式

以下将参考附图详细说明本公开的各种示例性实施例、特征和方面。附图中相同的附图标记表示功能相同或相似的元件。尽管在附图中示出了实施例的各种方面,但是除非特别指出,不必按比例绘制附图。

在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中术语“至少一个”表示多种中的任意一个或多个中的至少两个的任意组合,例如,包括A、B、C中的至少一个,可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。

另外,为了更好地说明本公开,在下文的具体实施方式中给出了众多的具体细节。本领域技术人

员应当理解，没有某些具体细节，本公开同样可以实施。在一些实例中，对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述，以便于凸显本公开的主旨。

本公开实施例提供的基因变异识别方案，可以获取基因变异候选位点对应的至少一个基因测序读段，从而可以利用至少一个基因测序读段对基因变异候选位点的基因变异进行识别。在基因变异识别过程中，可以确定基因变异候选位点的碱基排列特征，并根据至少一个基因测序读段在预设位点区间的碱基排列信息，确定基因变异候选位点的非碱基排列特征，然后通过碱基排列特征和非碱基排列特征对基因变异候选位点的基因变异进行识别。这里的非碱基排列特征在碱基排列顺序改变后保持不变，即可以认为基因变异候选位点的基因变异是否为真变异不受碱基排列顺序的影响，从而对基因变异候选位点的基因变异进行识别时，考虑基因数据的碱基排列不变性，提高基因变异识别的准确性。

在相关技术中，通常是利用支持向量机、随机森林等传统机器学习方法进行基因变异识别，这种方式虽然实现简单，但基因变异识别的效果在基因数据量增加到一定程度之后会陷入瓶颈。还有一些相关技术采用深度学习方法，利用神经网络对基因变异进行识别。但是，神经网络提取的特征通常与碱基排列顺序相关，碱基排列顺序稍有不同就可能得到不同的识别结果，造成神经网络过度拟合的问题。而本公开实施例提供的基因变异识别方案，考虑了基因数据的碱基排列不变性，利用基因变异识别模型提取基因变异候选位点的非碱基排列特征，使得到的识别结果不受碱基排列顺序的影响，提高基因变异识别模型的鲁棒性，缓解过度拟合的问题，减小基因变异识别模型训练的难度。下述实施例将会对基因变异识别过程作详细说明。

图1示出根据本公开一实施例的基因变异识别方法的流程图。该基因变异识别方法可以由基因变异识别装置或其它处理设备执行，其中，基因变异识别装置可以为用户设备（User Equipment, UE）、移动设备、用户终端、终端、蜂窝电话、无绳电话、个人数字处理（Personal Digital Assistant, PDA）、手持设备、计算设备、车载设备、可穿戴设备等，或者，基因变异识别装置可以为服务器。在一些可能的实现方式中，该基因变异识别方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。如图1所示，该基因变异识别方法包括：

步骤11，获取基因变异候选位点对应的至少一个基因测序读段。

在本公开实施例中，基因变异识别装置可以获取由基因测序得到的基因测序读段，然后在基因测序得到的基因测序读段中，获取基因变异候选位点对应的至少一个基因测序读段。这里的基因测序读段可以理解为经过基因测序后标注有碱基类型的碱基序列，每个基因测序读段的长度可以相同也可以不同。在长度不同的情况下，每个基因测序读段的长度可以在预设长度范围内，从而可以保证每个基因测序读段的长度比较接近。碱基类型可以包括胞嘧啶（C）、鸟嘌呤（G）、腺嘌呤（A）、胸腺嘧啶（T），从而基因测序读段可以包括AGCT的碱基序列。这里的基因变异候选位点可以是碱基序列存在异常的位点。碱基序列的位点可以表示碱基序列的位置，针对每个位点，可以存在至少一个基因测序读段，即，在同一个位点可以存在由基因测序得到的至少一个基因测序读段。相应地，基因变异候选位点对应至少一个基因测序读段，其中，这至少一个基因测序读段都覆盖这一位点。基因变异候选位点可以为至少一个，每个基因变异候选位点可以对应至少一个基因测序读段。为了便于理解，本公开实施例以一个基因变异候选位点进行说明。

步骤12，获取所述基因变异候选位点的碱基排列特征。

在本公开实施例中，可以利用基因变异识别模型，根据基因变异候选位点的基因排列信息，提取基因变异候选位点的碱基排列特征。这里的碱基排列信息可以是与碱基排列顺序相关的信息，例如，某个基因测序读段在某个位点区间的碱基序列依次为A、C、G、T，则碱基排列信息可以为ACGT。

碱基排列信息可以包括预设位点区间内参考基因组的碱基类型、每种碱基类型的基因数量、每种碱基类型的缺失基因数量、每种碱基类型的插入基因数量等信息。由碱基排列信息得到的碱基排列特征与碱基排列顺序相关。

步骤13, 基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息, 确定所述基因变异候选位点的非碱基排列特征; 其中, 所述非碱基排列特征在碱基排列顺序改变后保持不变。

在本公开实施例中, 在获取基因变异候选位点对应的至少一个基因测序读段之后, 可以在预设位点区间, 提取该基因变异候选位点对应的至少一个基因测序读段的碱基排列信息, 并根据提取的碱基排列信息生成该基因变异候选位点的非碱基排列特征。非碱基排列信息可以是不受到碱基排列顺序限制的信息。从而可以根据至少一个基因测序读段在预设位点区间的非碱基排列信息, 确定基因变异候选位点的非碱基排列特征。这里, 非碱基排列信息可以包括位点处对应的基因测序读段的数量、在该位点发生变异的基因测序读段的数量等具有碱基排列不变性的信息。

这里, 在提取非碱基排列信息时, 可以随机选择该基因变异候选位点对应的若干个基因测序读段, 提取随机选择的若干个基因测序读段的非碱基排列信息; 还可以提取该基因变异候选位点对应的每个基因测序读段的非碱基排列信息。在提取至少一个基因测序读段在预设位点区间的非碱基排列信息时, 可以提取至少一个基因测序读段在该预设位点区间内每个位点的非碱基排列信息, 还可以随机选择该预设位点区间内若干个相邻位点, 提取至少一个基因测序读段在若干个相邻位点的非碱基排列信息。在确定所述基因变异候选位点的非碱基排列特征时, 可以利用基于神经网络训练得到的基因变异识别模型。

步骤14, 基于所述基因变异候选位点的碱基排列特征和非碱基排列特征, 对所述基因变异候选位点的基因变异进行识别。

在本公开实施方式中, 在确定碱基排列特征和非碱基排列特征之后, 可以由碱基排列特征和非碱基排列特征得到基因变异候选位点的特征矩阵中, 利用该特征矩阵对该基因变异候选位点的基因变异进行识别, 例如, 可以利用上述基因变异识别模型判断该基因变异候选位点的基因是否是由于病变引起的真变异, 还是由于噪声等原因而导致的碱基序列异常的假变异。这里, 得到的基因变异候选位点的特征矩阵可以为二维特征矩阵, 特征矩阵的尺寸可以是特征向量的个数 $\times$ 预设位点区间的大小, 其中的特征向量可以是基于碱基排列特征和非碱基排列特征生成的。由于变异候选位点的基因变异是否为病变引起的真基因变异不受碱基排列顺序的影响, 更多地受到基因变异候选位点所在的基因环境的影响, 例如, 受到基因变异候选位点附近的其他位点存在变异基因等基因环境的影响, 从而得到的特征矩阵中对应于碱基排列特征的特征向量的排列顺序可以不受限制, 碱基排列特征的特征向量在特征矩阵中的排列顺序可以随机变动, 提高基因变异识别的效率和准确率。

本公开实施例中可以根据基因变异候选位点的碱基排列特征和非碱基排列特征对基因变异候选位点的基因变异进行识别, 从而可以考虑基因变异的碱基排列不变性, 更好地对基因变异进行识别。在对基因变异候选位点的基因变异进行识别时, 可以获取基因变异候选位点对应的至少一个基因测序读段。本公开实例还提供了一种获取基因变异候选位点对应的至少一个基因测序读段的过程。

图2示出根据本公开一实施例的获取基因变异候选位点对应的至少一个基因测序读段的流程图。在一种可能的实现方式中, 获取基因变异候选位点对应的至少一个基因测序读段, 可以包括以下步骤:

步骤111, 获取由体细胞基因进行基因测序得到的基因测序读段。

这里, 通过体细胞基因进行基因测序可以得到至少一个基因测序读段, 基因测序读段可以是对体细胞基因进行碱基类型标注的序列。体细胞基因在进行基因测序之后, 不仅可以得到基因测序读段中

每个基因的碱基类型，还可以得到基因测序读段中每个基因所在位点的基因位置信息。同一个位点可以对应至少一个基因测序读段。

在一种可能的实现方式中，通过体细胞基因进行基因测序可以得到至少一个基因测序读段，可以对基因测序得到的基因测序读段进行预处理，这里的预处理方式可以包括交叉污染筛选、测序质量筛选、比对质量筛选、读段长度异常筛选等。通过预处理，可以筛选掉交叉污染的基因测序读段，以及筛选掉测序质量和比对质量较低、读段长度异常的基因测序读段。

步骤112，将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对，得到比对结果。

在本公开实施例中，在获取由体细胞基因进行基因测序得到的基因测序读段之后，可以将获取的基因测序读段的碱基序列与相同位点的参考基因组的碱基序列进行比对，得到对比结果。举例来说，可以将每个进行基因测序得到的基因测序读段与相同位点的参考基因组的碱基序列进行对比，确定基因测序读段的碱基序列与参考基因组的碱基序列不同的位点。还可以将具有相同位点的至少一个基因测序读段与相同位点的参考基因组的碱基序列进行对比，确定至少一个基因测序读段的碱基序列与参考基因组的碱基序列不同的位点。这里，参考基因组可以是标注有正确碱基序列的碱基序列。

步骤113，根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点。

在本公开实施例中，可以根据比对结果确定基因测序读段与参考基因组的碱基序列不同的位点，如果该位点对应的至少一个基因测序读段中，在该位点发送变异的基因测序读段的比例大于预设比例，则可以确定该位点为基因变异候选位点，否则，可以认为该位点不是基因变异候选位点。基因测序读段在该位点与参考基因组的碱基序列不同，可能是因为测序错误导致的不同，通过这种方式，可以减少由于基因测序失误引起的碱基序列异常现象。

步骤114，获取所述基因变异候选位点对应的至少一个基因测序读段。

在本公开实施例中，在确定基因变异候选位点之后，可以获取基因变异候选位点对应的至少一个基因测序读段。其中，每个基因变异候选位点对应的至少一个基因测序读段，在该基因变异候选位点的碱基序列与相同位点的参考基因组的碱基序列可以不同。这里的基因变异候选位点可以为至少一个。

通过上述获取基因变异候选位点对应的至少一个基因测序读段的过程，不仅可以较为准确地确定基因变异候选位点，还可以在基因测序得到的基因测序读段中确定基因变异候选位点对应的至少一个基因测序读段。

本公开实施例中可以根据基因变异候选位点对应的至少一个基因测序读段的碱基排列信息，确定该基因变异候选位点的碱基排列特征，从而在识别基因变异候选位点的基因变异时，可以根据该碱基排列特征对基因识别进行数据增强处理。下面通过一示例对确定基因变异候选位点的碱基排列特征的过程进行详细说明。

图3示出根据本公开一实施例的基因变异候选位点的碱基排列特征过程的流程图。如图3所示，上述步骤12可以包括以下步骤：

步骤121，确定所述基因变异候选位点所在的预设位点区间；

步骤122，根据参考基因组在所述预设位点区间的碱基排列信息，获取所述基因变异候选位点的碱基排列特征；其中，所述碱基排列特征用于表征碱基排列顺序。

在本公开实施例的示例中，每一个基因变异候选位点可以存在至少一个基因测序读段。为了提高基因变异识别的准确度，不仅可以考虑该基因变异候选位点的碱基排列信息，还可以考虑该基因变异候选位点附近的位点的碱基排列信息。这里，碱基排列信息可以包括候选基因组的碱基排列信息，在碱基排列信息为候选基因组的碱基排列信息的情况下，可以认为每个基因测序读段的碱基排列信息相

同，均为候选基因组的碱基排列信息。从而可以根据基因变异候选位点的基因位置信息，确定该基因变异候选位点所在的预设位点区间，例如，可以将基因变异候选位点前后150个碱基形成的区间作为基因变异候选位点所在的预设位点区间。然后可以针对该预设位点区间内的每个位点，获取参考基因组在预设位点区间的碱基排列信息，由参考基因组在预设位点区间的碱基排列信息生成基因变异候选位点的碱基排列特征。碱基排列信息可以参考基因组在预设位点区间中每个位点的碱基序列组成，例如，预设位点区间包括4个碱基序列，分别为A、C、G、T，则碱基排列信息可以为ACGT的碱基排列顺序。碱基排列特征可以用碱基排列特征向量进行表示，可以是基因变异候选位点的特征矩阵的一部分，例如，如果表征碱基排列信息的碱基排列特征向量为4个，分别为a1、a2、a3和a4，则a1、a2、a3和a4可以为特征矩阵的前4维特征。

本公开实施例的示例中不仅在对基因变异候选位点的基因变异进行识别时，考虑了基因变异候选位点所对应的碱基排列特征，还考虑了基因变异候选位点具有碱基排列不变性的非碱基排列特征。下面通过一示例对确定基因变异候选位点的非碱基排列特征的过程进行详细说明。

图4示出根据本公开一实施例的基因变异候选位点的非碱基排列特征过程的流程图。如图4所示，上述步骤13可以包括以下步骤：

步骤131，获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息；
步骤132，基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在本公开实施例的示例中，考虑到基因数据具有碱基排列不变性的性质，从而可以在基因变异识别过程中，获取至少一个基因测序读段在预设位点区间中每个位点的非碱基排列信息。这里，非碱基排列信息可以是具有碱基排列不变性的信息，例如，位点处对应的基因测序读段的数量、变异数量。非碱基排列信息可以为多种，相应地，每种非碱基排列信息生成的非碱基排列特征可以形成一个非碱基排列特征向量，非碱基排列特征向量可以为一个或多个。

本公开实施例提供的基因变异识别方案可以应用于已经确诊为患有癌症的病人，通过基因变异识别可以为病人指导用药。因此，基因测序读段中的一部分基因测序读段可以来源于正常细胞，正常细胞可以认为是没有发生病变的细胞。还有一部分基因测序读段可以来源于病变细胞。从而在确定基因变异候选位点的非碱基排列特征时，可以分别基于来源于正常细胞的基因测序读段和来源于病变细胞的基因测序读段，确定基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，确定基因变异候选位点的非碱基排列特征时，可以确定至少一个基因测序读段中来源于正常细胞的基因测序读段，然后基于正常细胞的基因测序读段在预设位点区间中每个位点的非碱基排列信息，确定基因变异候选位点的非碱基排列特征。这样，可以基于来源于正常细胞的基因测序读段确定基因变异候选位点的非碱基排列特征。

下面提供了基于正常细胞的基因测序读段确定基因变异候选位点的非碱基排列特征的几个示例。

在该公开实施例的一个示例中，在确定基因变异候选位点的非碱基排列特征时，可以在基因测序读段中，确定在基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段，然后根据预设位点区间中每个位点对应的第一基因测序读段的数量，确定基因变异候选位点的非碱基排列特征。

在该示例中，可以在基因测序读段中选择在基因变异候选位点未发生基因变异的第一基因测序读段，针对预设位点区间中的每个位点，可以统计第一基因测序读段在该位点的数量。换言之，可以统计有多少个第一基因测序读段包含该位点。其中，包含某一位点的第一基因测序读段可认为是该位点对应的第一基因测序读段。由于每个基因测序读段的长度可能不同，基因变异候选位点相对于每个基

因测序读段的位置不同,例如,基因变异候选位点可以位于基因测序读段的中间位置,还可以位于基因测序读段的边缘位置,从而预设位点区间中的每个位点所对应的基因测序读段的数量不同。由每个位点对应的第一基因测序读段的数量,可以生成非碱基排列特征对应的一个非碱基排列特征向量,该非碱基排列特征向量中的每个特征元素可以对应相应位点的第一基因测序读段的数量。

5 在该公开实施例的另一个示例中,在确定基因变异候选位点的非碱基排列特征时,可以在基因测序读段中,确定在基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段,然后在预设位点区间中的每个位点,确定第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量,作为第一基因测序读段的变异数量,根据第一基因测序读段的变异数量,确定基因变异候选位点的非碱基排列特征。

10 在该示例中,可以在基因测序读段中选择在基因变异候选位点未发生基因变异的第一基因测序读段,针对预设位点区间中的每个位点,可以统计第一基因测序读段在该位点发生基因变异的变异数量。这里,虽然基因测序读段在基因变异候选位点未发生基因变异(即在基因变异候选位点与参考基因组的碱基类型一致),但是可能在基因变异候选位点之外的其他位点发生基因变异(即在其他位点与参考基因组的碱基类型不一致),从而可以针对预设位点区间的每个位点,统计在该位点的第一基因测序读段中发生变异的变异数量。换言之,针对每个位点,可以统计包含该位点的第一基因测序读段中,
15 有多少个第一基因测序读段在该位点发生变异。由每个位点对应的第一基因测序读段中发生变异的变异数量,可以生成非碱基排列特征对应的一个非碱基排列特征向量,该非碱基排列特征向量中的每个特征元素可以对应相应位点的第一基因测序读段的变异数量,换言之,包含该相应位点且在该相应位点发生变异的第一基因测序读段的数量。

20 举例来说,针对来源于正常细胞的基因测序读段,可以确定正常细胞的基因测序读段中在基因变异候选位点未发生变异的第一基因测序读段,然后针对预设位点区间中的每个位点,统计每个位点对应的第一基因测序读段的数量和在该位点发生变异的数量,这两个信息可以对应于上述特征矩阵中的第5维特征和第6维特征。

 在该公开实施例的另一个示例中,在确定基因变异候选位点的非碱基排列特征时,可以在所述基因
25 因测序读段中,确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段,然后根据所述预设位点区间中每个位点对应的第二基因测序读段的数量,确定所述基因变异候选位点的非碱基排列特征。在该示例中,可以在基因测序读段中选择与基因变异候选位点变异一致的第二基因测序读段,针对预设位点区间中的每个位点,可以统计第二基因测序读段在该位点的数量。由每个位点对应的第二基因测序读段的数量,生成非碱基排列特征对应的一个非碱基排列特征向量,
30 该非碱基排列特征向量中的每个特征元素可以对应相应位点的第二基因测序读段的数量。

 在该公开实施例的另一个示例中,在确定基因变异候选位点的非碱基排列特征时,可以在所述基因
 因测序读段中,确定在基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段,然后在预设位点区间中的每个位点,确定第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量,作为第二基因测序读段的变异数量,根据第二基因测序读段的变异
35 数量,确定基因变异候选位点的非碱基排列特征。在该示例中,可以在基因测序读段中选择与基因变异候选位点变异一致的第二基因测序读段(基因变异候选位点的变异碱基类型可通过基因测序得到),针对预设位点区间中的每个位点,统计第二基因测序读段在该位点发生基因变异的变异数量,换言之,统计包含该位点且在该位点发生变异的第二基因测序读段的数量。每个位点对应的第二基因测序读段中发生变异的变异数量,可以生成非碱基排列特征对应的一个非碱基排列特征向量,该非碱

基排列特征向量中的每个特征元素可以对应相应位点的第二基因测序读段的变异数量。

举例来说，针对来源于正常细胞的基因测序读段，可以在正常细胞的基因测序读段中选择与基因变异候选位点变异一致的第二基因测序读段，然后针对预设位点区间中的每个位点，统计每个位点对应的第二基因测序读段的数量和在该位点发生变异的数量，这两个信息可以对应于上述特征矩阵中的第7维特征和第8维特征。

在该公开实施例的另一个示例中，在确定所述基因变异候选位点的非碱基排列特征时，可以确定基因测序读段中的第三基因测序读段，然后根据预设位点区间中每个位点对应的第三基因测序读段的数量，确定基因变异候选位点的非碱基排列特征。这里，第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致，即，第三基因测序读段是基因测序读段中除去第一基因测序读段和第二基因测序读段的剩余基因测序读段。第三基因测序读段可以是在基因变异候选位点存在插入基因、缺失基因等情况的基因测序读段。在该示例中，可以在基因测序读段中确定剩余的第三基因测序读段，针对预设位点区间中的每个位点，可以统计第三基因测序读段在该位点的数量。由每个位点对应的第三基因测序读段的数量，生成非碱基排列特征对应的一个非碱基排列特征向量，该非碱基排列特征向量中的每个特征元素可以对应相应位点的第三基因测序读段的数量。

在该公开实施例的另一个示例中，在确定基因变异候选位点的非碱基排列特征时，可以确定基因测序读段中的第三基因测序读段，然后在预设位点区间中的每个位点，确定第三基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量，根据第三基因测序读段的变异数量，确定基因变异候选位点的非碱基排列特征。这里，第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致，即，第三基因测序读段是基因测序读段中除去第一基因测序读段和第二基因测序读段的剩余基因测序读段。在该示例中，可以在基因测序读段中确定剩余的第三基因测序读段，针对预设位点区间中的每个位点，统计第三基因测序读段在该位点发生基因变异的变异数量。每个位点对应的第三基因测序读段中发生变异的变异数量，可以生成非碱基排列特征对应的一个非碱基排列特征向量，该非碱基排列特征向量中的每个特征元素可以对应相应位点的第三基因测序读段的变异数量。

举例来说，针对来源于正常细胞的基因测序读段，可以在正常细胞的基因测序读段中选择除第一基因测序读段和第二基因测序读段之外的第三基因测序读段，然后针对预设位点区间中的每个位点，统计每个位点对应的第三基因测序读段的数量和在该位点发生变异的数量，这两个信息可以对应于上述特征矩阵中的第9维特征和第10维特征。

在一种可能的实现方式中，确定基因变异候选位点的非碱基排列特征时，可以确定至少一个基因测序读段中来源于病变细胞的基因测序读段，然后基于病变细胞的基因测序读段在预设位点区间中每个位点的非碱基排列信息，确定基因变异候选位点的非碱基排列特征。这样，可以基于来源于病变细胞的基因测序读段确定基因变异候选位点的非碱基排列特征。

在该实现方式中，基于病变细胞的基因测序读段确定基因变异候选位点的非碱基排列特征的过程，可以参见上述正常细胞的基因测序读段确定非碱基排列特征的过程。举例来说，针对来源于病变细胞的基因测序读段，可以在病变细胞的基因测序读段中确定第一基因测序读段、第二基因测序读段和第三基因测序读段，然后针对预设位点区间中的每个位点，统计每个位点对应的第一基因测序读段的数量和变异数量、第二基因测序读段的数量和变异数量和第三基因测序读段的数量和变异数量，这些信

息可以对应于上述特征矩阵中的第11至16维特征。

通过上述方式，可以针对至少一个基因测序读段在预设位点区间与碱基排列相关的非碱基排列信息，确定基因变异候选位点的非碱基排列特征，从而可以在基因变异识别时考虑基因数据的碱基排列不变性，使基因变异识别更加容易、准确。下面通过一示例对基因变异候选位点的基因变异进行识别的过程进行说明。

图5示出根据本公开一实施例的识别基因变异候选位点的基因变异过程的流程图。如图5所示，上述步骤14可以包括以下步骤：

步骤141，根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；

步骤142，根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别。

在本公开实施例的示例中，在确定基因变异候选位点的碱基排列特征和非碱基排列特征之后，可以利用基于神经网络得到的基因变异识别模型，对碱基排列特征和非碱基排列特征进行特征整合，将碱基排列特征形成的碱基排列特征向量与非碱基排列特征形成的非碱基排列特征向量合成一个特征矩阵。特征矩阵的第一维度特征对应于碱基排列信息和非碱基排列信息，第二维度特征对应于所述预设位点区间的位点。特征矩阵的尺寸是特征向量的个数 $\times$ 预设位点区间的大小。举例来说，若特征向量的个数为16，预设位点区间包括150个位点，则特征矩阵的尺寸可以为 16×150 ，其中，第一维度特征对应于16维特征向量，第1至4为可以对应于碱基排列特征，第5至16维特征向量可以对应于非碱基排列特征，具有碱基排列不变性。然后可以利用上述基因变异识别模型根据该特征矩阵对变异候选位点的基因变异进行识别。通过这种方式，可以利用神经网络模型整合基因变异候选位点对应的碱基排列信息和非碱基排列信息，从而可以更加全面地对基因测序数据进行分析，使基因变异识别更加准确。

在一种可能的实现方式中，根据所述基因变异候选位点的整合特征，对所述基因变异候选位点的基因变异进行识别，可以包括：根据所述基因变异候选位点的特征矩阵，得到所述基因变异候选位点的基因发生变异的变异值，在所述变异值大于或等于预设阈值的情况下，确定所述基因变异候选位点的基因存在变异。这里，基因发生变异的变异值可以是表征该基因变异候选位点发生真变异的可能性，举例来说，如果变异值越大，该基因变异候选位点发生真变异的可能性越大。可以利用上述基因变异识别模型对得到的二维的特征矩阵进行处理得到变异值，并根据变异值判断基因变异候选位点的基因变异是否为真变异。在一种可能的实现方式中，变异值可以在0至1之间。预设阈值可以根据应用场景进行设置，例如，0.3、0.5，如果变异值大于预设阈值，则可以认为该基因变异候选位点的基因变异为真变异，即为病变引起的基因变异；否则，可以为该基因变异候选位点的基因变异为假变异，即是干扰形成的基因异常。

本公开实施例可以利用基因变异识别模型对基因变异候选位点的基因变异进行识别，该基因变异识别模型在训练过程中，可以利用基因数据的碱基排列不变性，将基因变异识别模型提取的特征矩阵进行矩阵变换，从而可以在模型训练过程中进行数据增强处理，使训练的基因变异识别模型具有更好地鲁棒性，减少过度拟合等问题。

图6示出根据本公开一实施例的得到基因变异候选位点的特征矩阵过程的流程图。

在本公开实施例中，可以将碱基排列信息的数据增强应用在基因变异识别模型的训练过程中。如图6所示，根据基因变异候选位点的碱基排列特征和非碱基排列特征，得到基因变异候选位点的特征

矩阵，可以包括：

步骤1411，根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，生成所述预设位点区间的每个第一维度特征的特征向量；

步骤1412，确定所述特征向量中碱基排列特征形成的碱基排列特征向量；

5 步骤1413，对所述碱基排列特征向量进行随机排序，得到所述基因变异候选位点的特征矩阵。

这里，第一维度特征对应于所述至少一个基因测序读段在预设位点区间的碱基排列信息，第一维度特征的特征向量可以包括由碱基排列特征形成的碱基排列特征向量和由非碱基排列特征形成的非碱基排列特征向量。由于非碱基排列特征具有碱基排列不变性，从而在碱基排列特征向量的排列顺序改变之后，非碱基排列特征不会受到影响。因此，可以将特征向量中碱基排列特征形成的碱基排列特征向量进行随机排序，得到基因变异候选位点的特征矩阵，实现碱基排列信息的数据增强处理，使训练后得到的基因变异识别模型考虑碱基排列不变性的性质，具有更优越的性能。

举例来说，若特征向量的个数为16，第一维度特征对应于16维特征向量，第1至4为可以对应于碱基排列特征，第5至16维特征向量可以对应于非碱基排列特征，则可以将第1至4的特征向量进行随机排序，形成多个特征矩阵。

15 本公开实施例通过提取基因变异候选位点的碱基排列特征和非碱基排列特征，从而在基因变异进行识别时可以考虑基因数据的碱基排列不变性，使基因变异进行识别的识别结果更加准确，筛掉胚系基因变异以及由于噪声和错误带来的干扰，提高基因变异识别的准确率。

本领域技术人员可以理解，在具体实施方式的上述方法中，各步骤的撰写顺序并不意味着严格的执行顺序而对实施过程构成任何限定，各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

20 图7示出根据本公开实施例的基因变异识别装置的框图，如图7所示，所述基因变异识别装置包括：第一获取模块71，用于获取基因变异候选位点对应的至少一个基因测序读段；

第二获取模块72，用于获取所述基因变异候选位点的碱基排列特征；

确定模块73，用于基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征；其中，所述非碱基排列特征在碱基排列顺序改变后保持不变；

25 识别模块74，用于基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别。

在一种可能的实现方式中，所述第二获取模块72，包括：

第一确定子模块，用于确定所述基因变异候选位点所在的预设位点区间；

30 第二确定子模块，用于根据参考基因组在所述预设位点区间的碱基排列信息，获取所述基因变异候选位点的碱基排列特征；其中，所述碱基排列特征用于表征碱基排列顺序。

在一种可能的实现方式中，所述确定模块73，包括：

第一获取子模块，用于获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息；

35 第三确定子模块，用于基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；根据所述预设位点区间中每个位点对应的第一基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；在所述预设位点区间中的每个位点，确定所述第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量，作为第一基因测序读段的变异数量；根据所述第一基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；根据所述预设位点区间中每个位点对应的第二基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；在所述预设位点区间中的每个位点，确定所述第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量，作为第二基因测序读段的变异数量；根据所述第二基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；根据所述预设位点区间中每个位点对应的第三基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；在所述预设位点区间中的每个位点，确定所述第三基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量；根据所述第三基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述至少一个基因测序读段中来源于正常细胞的基因测序读段；基于所述正常细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述第三确定子模块，具体用于，

确定所述至少一个基因测序读段中来源于病变细胞的基因测序读段；基于所述病变细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

在一种可能的实现方式中，所述识别模块74，包括：

生成子模块，用于根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；

识别子模块，用于根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异

进行识别。

在一种可能的实现方式中，所述识别子模块，具体用于，

根据所述基因变异候选位点的特征矩阵，得到所述基因变异候选位点的基因发生变异的变异值；
在所述变异值大于或等于预设阈值的情况下，确定所述基因变异候选位点的基因存在变异。

5 在一种可能的实现方式中，所述生成子模块，具体用于，

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，生成所述预设位点区间的每个第一维度特征的特征向量；确定所述特征向量中碱基排列特征形成的碱基排列特征向量；对所述碱基排列特征向量进行随机排序，得到所述基因变异候选位点的特征矩阵。

在一种可能的实现方式中，所述第一获取模块，包括：

10 第二获取子模块，用于获取由体细胞基因进行基因测序得到的基因测序读段；对比子模块，用于将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对，得到比对结果；第四确定子模块，用于根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点；第三获取子模块，用于获取所述基因变异候选位点对应的至少一个基因测序读段。

15 在一些实施例中，本公开实施例提供的装置具有的功能或包含的模块可以用于执行上文方法实施例描述的方法，其具体实现可以参照上文方法实施例的描述，为了简洁，这里不再赘述。

图8是根据一示例性实施例示出的一种用于基因变异识别装置1900的框图。例如，装置1900可以被提供为一服务器。参照图8，装置1900包括处理组件1922，其进一步包括一个或多个处理器，以及由存储器1932所代表的存储器资源，用于存储可由处理组件1922的执行的指令，例如应用程序。存储器1932中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件1922被配置为执行指令，以执行上述方法。

20 装置1900还可以包括一个电源组件1926被配置为执行装置1900的电源管理，一个有线或无线网络接口1950被配置为将装置1900连接到网络，和一个输入输出（I/O）接口1958。装置1900可以操作基于存储在存储器1932的操作系统，例如Windows Server™，Mac OS X™，Unix™，Linux™，FreeBSD™或类似。

25 在示例性实施例中，还提供了一种非易失性计算机可读存储介质，例如包括计算机程序指令的存储器1932，上述计算机程序指令可由装置1900的处理组件1922执行以完成上述方法。

本公开可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本公开的各个方面的计算机可读程序指令。

30 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是一—但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、
35 以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以

包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

用于执行本公开操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等，以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本公开的各个方面。

这里参照根据本公开实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本公开的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本公开的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

权 利 要 求 书

1、一种基因变异识别方法，其特征在于，所述方法包括：

获取基因变异候选位点对应的至少一个基因测序读段；

获取所述基因变异候选位点的碱基排列特征；

基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征；其中，所述非碱基排列特征碱基排列顺序改变后保持不变；

基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别。

2、根据权利要求1所述的方法，其特征在于，所述获取所述基因变异候选位点的碱基排列特征，包括：

确定所述基因变异候选位点所在的预设位点区间；

根据参考基因组在所述预设位点区间的碱基排列信息，获取所述基因变异候选位点的碱基排列特征；其中，所述碱基排列特征用于表征碱基排列顺序。

3、根据权利要求1或2所述的方法，其特征在于，所述基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息；

基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

4、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；

根据所述预设位点区间中每个位点对应的第一基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

5、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；

在所述预设位点区间中的每个位点，确定所述第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量，作为第一基因测序读段的变异数量；

根据所述第一基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

6、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；

根据所述预设位点区间中每个位点对应的第二基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

7、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致

的第二基因测序读段；

在所述预设位点区间中的每个位点，确定所述第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量，作为第二基因测序读段的变异数量；

根据所述第二基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

- 5 8、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；

- 10 根据所述预设位点区间中每个位点对应的第三基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

9、根据权利要求3所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

- 15 确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；

在所述预设位点区间中的每个位点，确定所述第三基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量；

根据所述第三基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

- 20 10、根据权利要求3至9中任意一项所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

确定所述至少一个基因测序读段中来源于正常细胞的基因测序读段；

基于所述正常细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

- 25 11、根据权利要求3至9中任意一项所述的方法，其特征在于，所述基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征，包括：

确定所述至少一个基因测序读段中来源于病变细胞的基因测序读段；

基于所述病变细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

- 30 12、根据权利要求1至11中任意一项所述的方法，其特征在于，所述基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别，包括：

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；

- 35 根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别。

13、根据权利要求12所述的方法，其特征在于，所述根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别，包括：

根据所述基因变异候选位点的特征矩阵，得到所述基因变异候选位点的基因发生变异的变异值；在所述变异值大于或等于预设阈值的情况下，确定所述基因变异候选位点的基因存在变异。

14、根据权利要求12所述的方法，其特征在于，所述根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵，包括：

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，生成所述预设位点区间的每个第一维度特征的特征向量；

5 确定所述特征向量中碱基排列特征形成的碱基排列特征向量；

对所述碱基排列特征向量进行随机排序，得到所述基因变异候选位点的特征矩阵。

15、根据权利要求1至14中任意一项所述的方法，其特征在于，获取基因变异候选位点对应的至少一个基因测序读段，包括：

获取由体细胞基因进行基因测序得到的基因测序读段；

10 将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对，得到比对结果；

根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点；

获取所述基因变异候选位点对应的至少一个基因测序读段。

16、一种基因变异识别装置，其特征在于，所述装置包括：

第一获取模块，用于获取基因变异候选位点对应的至少一个基因测序读段；

15 第二获取模块，用于获取所述基因变异候选位点的碱基排列特征；

确定模块，用于基于所述至少一个基因测序读段在预设位点区间的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征；其中，所述非碱基排列特征在碱基排列顺序改变后保持不变；

识别模块，用于基于所述基因变异候选位点的碱基排列特征和非碱基排列特征，对所述基因变异候选位点的基因变异进行识别。

20 17、根据权利要求16所述的装置，其特征在于，所述第二获取模块，包括：

第一确定子模块，用于确定所述基因变异候选位点所在的预设位点区间；

第二确定子模块，用于根据参考基因组在所述预设位点区间的碱基排列信息，获取所述基因变异候选位点的碱基排列特征；其中，所述碱基排列特征用于表征碱基排列顺序。

18、根据权利要求16或17所述的装置，其特征在于，所述确定模块，包括：

25 第一获取子模块，用于获取所述至少一个基因测序读段在所述预设位点区间中每个位点的非碱基排列信息；

第三确定子模块，用于基于所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

19、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

30 在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；

根据所述预设位点区间中每个位点对应的第一基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

20、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

35 在所述基因测序读段中，确定在所述基因变异候选位点与参考基因组的碱基类型一致的第一基因测序读段；

在所述预设位点区间中的每个位点，确定所述第一基因测序读段的碱基类型与参考基因组的碱基类型不一致的第一基因测序读段的数量，作为第一基因测序读段的变异数量；

根据所述第一基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

21、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；

根据所述预设位点区间中每个位点对应的第二基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

22、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

在所述基因测序读段中，确定在所述基因变异候选位点与基因变异候选位点的变异碱基类型一致的第二基因测序读段；

在所述预设位点区间中的每个位点，确定所述第二基因测序读段的碱基类型与参考基因组的碱基类型不一致的第二基因测序读段的数量，作为第二基因测序读段的变异数量；

根据所述第二基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

23、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；

根据所述预设位点区间中每个位点对应的第三基因测序读段的数量，确定所述基因变异候选位点的非碱基排列特征。

24、根据权利要求18所述的装置，其特征在于，所述第三确定子模块，具体用于，

确定所述基因测序读段中的第三基因测序读段；其中，所述第三基因测序读段在基因变异候选位点的碱基类型与参考基因组的碱基类型不一致，并且，第三基因测序读段在基因变异候选位点的碱基类型与基因变异候选位点的变异碱基类型不一致；

在所述预设位点区间中的每个位点，确定所述第三基因测序读段的碱基类型与参考基因组的碱基类型不一致的第三基因测序读段的数量，作为所述第三基因测序读段的变异数量；

根据所述第三基因测序读段的变异数量，确定所述基因变异候选位点的非碱基排列特征。

25、根据权利要求18至24中任意一项所述的装置，其特征在于，所述第三确定子模块，具体用于，确定所述至少一个基因测序读段中来源于正常细胞的基因测序读段；

基于所述正常细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

26、根据权利要求18至24中任意一项所述的装置，其特征在于，所述第三确定子模块，具体用于，

确定所述至少一个基因测序读段中来源于病变细胞的基因测序读段；

基于所述病变细胞的基因测序读段在所述预设位点区间中每个位点的非碱基排列信息，确定所述基因变异候选位点的非碱基排列特征。

27、根据权利要求16至26中任意一项所述的装置，其特征在于，所述识别模块，包括：

生成子模块，用于根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，得到所述基因变异候选位点的特征矩阵；其中，所述特征矩阵的第一维度特征对应于所述基因变异候选位点的碱基排列特征和非碱基排列特征，所述特征矩阵的第二维度特征对应于所述预设位点区间的位点；

识别子模块，用于根据所述基因变异候选位点的特征矩阵，对所述基因变异候选位点的基因变异进行识别。

28、根据权利要求27所述的装置，其特征在于，所述识别子模块，具体用于，

根据所述基因变异候选位点的特征矩阵，得到所述基因变异候选位点的基因发生变异的变异值；
在所述变异值大于或等于预设阈值的情况下，确定所述基因变异候选位点的基因存在变异。

29、根据权利要求27所述的装置，其特征在于，所述生成子模块，具体用于，

根据所述基因变异候选位点的碱基排列特征和非碱基排列特征，生成所述预设位点区间的每个第

5 一维度特征的特征向量；

确定所述特征向量中碱基排列特征形成的碱基排列特征向量；

对所述碱基排列特征向量进行随机排序，得到所述基因变异候选位点的特征矩阵。

30、根据权利要求16至29中任意一项所述的装置，其特征在于，所述第一获取模块，包括：

第二获取子模块，用于获取由体细胞基因进行基因测序得到的基因测序读段；

10 对比子模块，用于将所述基因测序读段的碱基序列与参考基因组的碱基序列进行比对，得到比对结果；

第四确定子模块，用于根据所述比对结果确定所述体细胞基因的基因存在异常的基因变异候选位点；

第三获取子模块，用于获取所述基因变异候选位点对应的至少一个基因测序读段。

15 31、一种基因变异识别装置，其特征在于，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为：执行权利要求1至15中任意一项所述的方法。

20 32、一种非易失性计算机可读存储介质，其上存储有计算机程序指令，其特征在于，所述计算机程序指令被处理器执行时实现权利要求1至15中任意一项所述的方法。

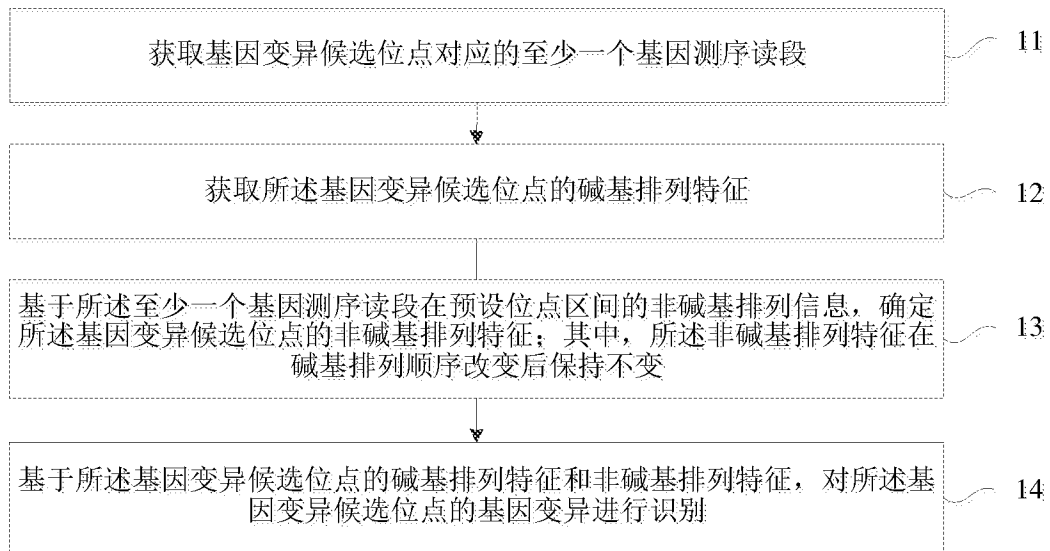


图 1

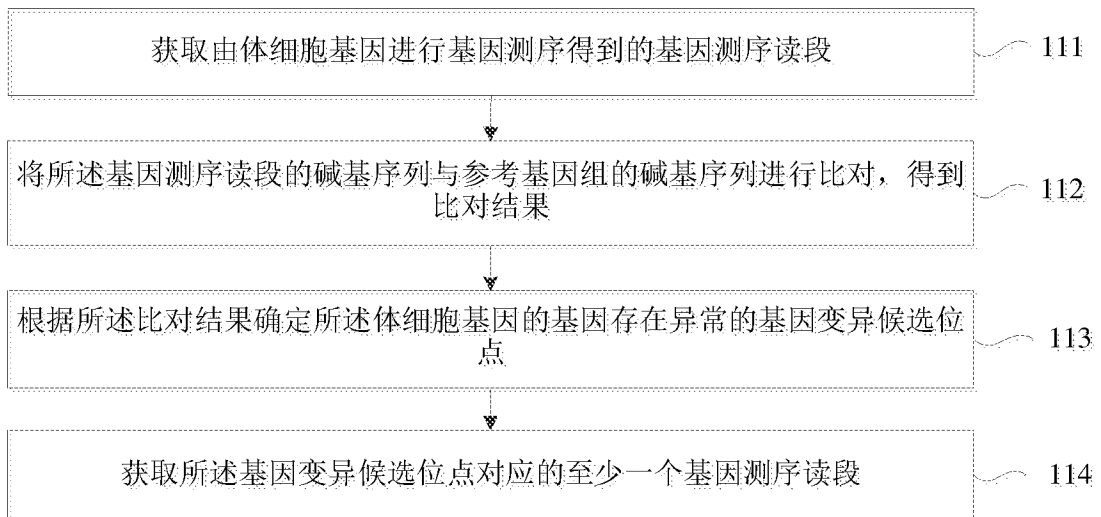


图 2

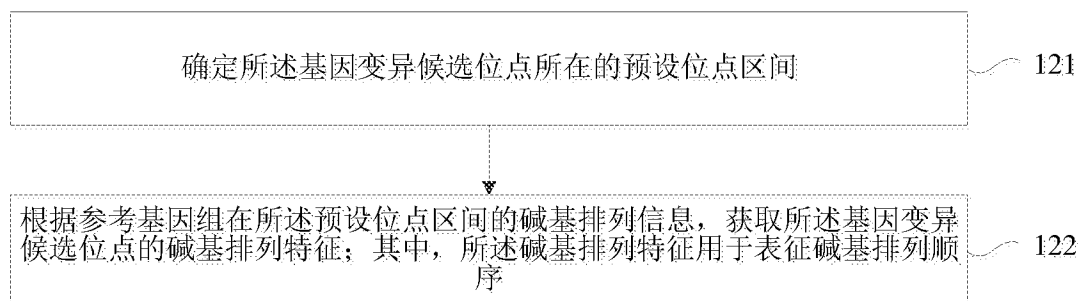


图 3

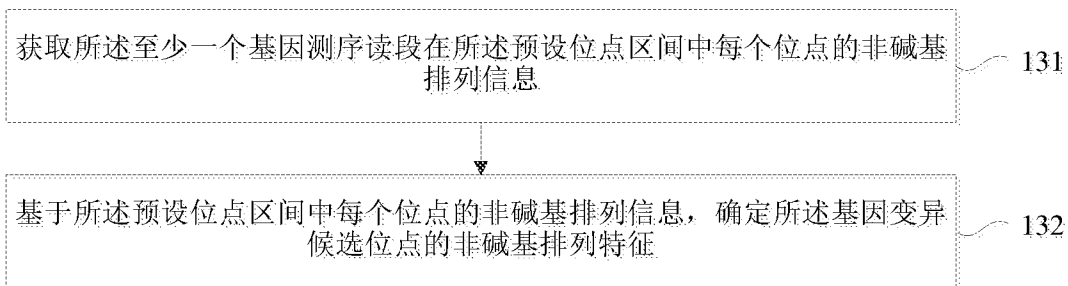


图 4

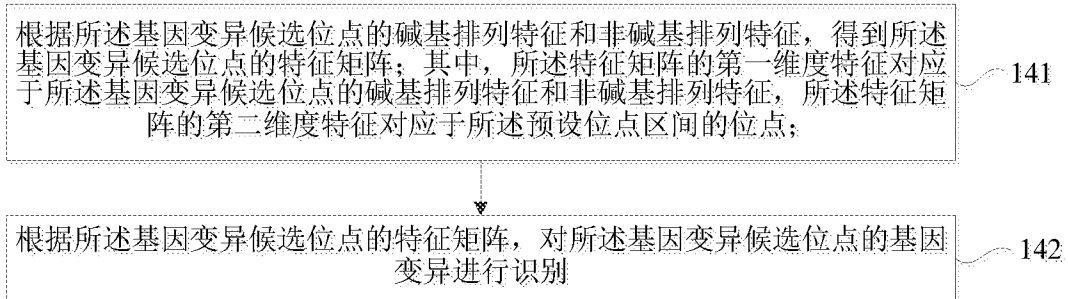


图 5

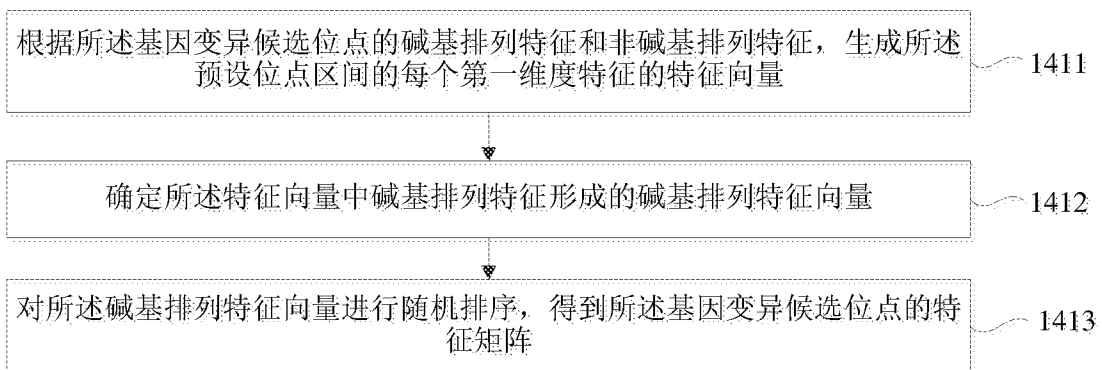


图 6

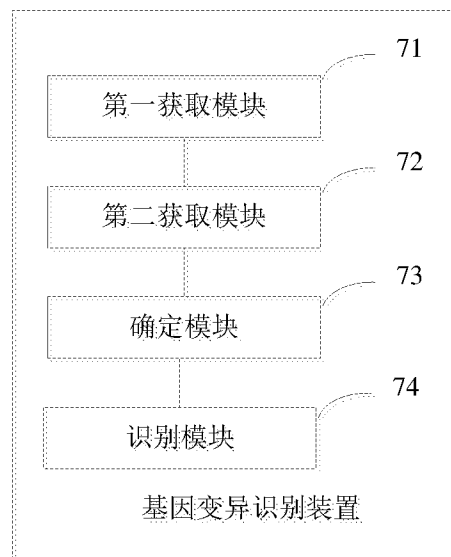


图 7

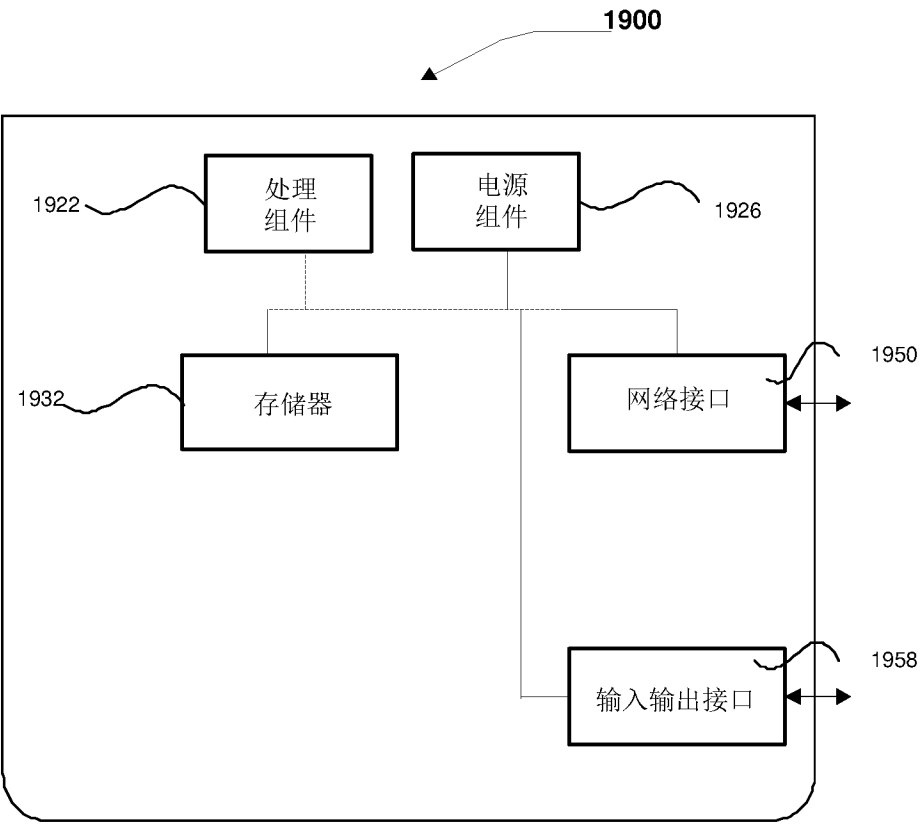


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/089504

A. CLASSIFICATION OF SUBJECT MATTER

G16B 20/20(2019.01)i; G16B 30/00(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)

G16B; G06F

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)

CNPAT; WPI; EPODOC; CNKI; IEEE: 筛选, 筛查, 检测, 识别, 基因, 异常, 变异, 位点, 排列, 数量, 碱基, 非碱基, 读段, detect, identify, gene, mutation, variation, base, non-base, number, arrangement, read

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 106611106 A (UNITED ELECTRONICS CO., LTD.) 03 May 2017 (2017-05-03) see description, paragraphs [0043]-[0070], figure 1	1-32
A	CN 108595912 A (SHENZHEN DIRECT GENOMICS CO., LTD.) 28 September 2018 (2018-09-28) entire document	1-32
A	WO 2015062184 A1 (ACCURASCIENCE, LLC) 07 May 2015 (2015-05-07) entire document	1-32

☐ 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

18 December 2019

Date of mailing of the international search report

03 January 2020

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

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/089504

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	106611106	A	03 May 2017	None			
CN	108595912	A	28 September 2018	None			
WO	2015062184	A1	07 May 2015	CN	105874460	A	17 August 2016
				US	2016026757	A1	28 January 2016
				US	2018232485	A1	16 August 2018

国际检索报告

国际申请号

PCT/CN2019/089504

A. 主题的分类

G16B 20/20(2019.01)i; G16B 30/00(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)

G16B; G06F

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))

CNPAT; WPI; EPDOC; CNKI; IEEE:筛选, 筛查, 检测, 识别, 基因, 异常, 变异, 位点, 排列, 数量, 碱基, 非碱基, 读段, detect, identify, gene, mutation, variation, base, non-base, number, arrangement, read

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
X	CN 106611106 A (北京荣之联科技股份有限公司) 2017年 5月 3日 (2017 - 05 - 03) 参见说明书第[0043]-[0070]段, 图1	1-32
A	CN 108595912 A (深圳市瀚海基因生物科技有限公司) 2018年 9月 28日 (2018 - 09 - 28) 全文	1-32
A	WO 2015062184 A1 (ACCURASCIENCE, LLC) 2015年 5月 7日 (2015 - 05 - 07) 全文	1-32

☐ 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

2019年 12月 18日

国际检索报告邮寄日期

2020年 1月 3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

授权官员

于白

电话号码 86-(10)-53961381

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/089504

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	106611106	A	2017年 5月 3日	无			
CN	108595912	A	2018年 9月 28日	无			
WO	2015062184	A1	2015年 5月 7日	CN	105874460	A	2016年 8月 17日
				US	2016026757	A1	2016年 1月 28日
				US	2018232485	A1	2018年 8月 16日