

序列变异分析方法、系统以及存储介质

技术领域

本公开涉及基因检测技术领域，尤其涉及一种序列变异分析方法、系统以及存储介质。

背景技术

基因缺陷是导致早期流产、死胎、围产儿死亡、婴幼儿死亡和先天残疾的主要原因。随着二代测序技术的应用和推广，人们对出生缺陷防控的重视程度越来越强，一次性对多种疾病进行快速、准确、低价的筛查或诊断成为可能。然而，如何客观、准确、高效地解读 NGS（Next Generation Sequencing，下一代测序）测序结果以指导临床应用一直是需要优化的重要问题。

基因检测程序一般包含样本处理、基因测序、变异识别、变异注释、变异解读、变异验证、检测报告等。目前，基因解读主要存在以下问题：

1) 基因检测数据需要依赖专业人员手工查询各种数据库和文献，以解读基因变异的致病性，这种解读往往耗费大量人力和时间；并且，基因解读对从业人员的要求高，需要从业人员有基因、变异、疾病等方面的知识储备，且由于解读是基于从业人员的经验逻辑，故因人而异解读可能会有不同的结果。

2) 大多变异注释工具无法识别复杂变异，导致注释信息错误，继而可能会引起解读错漏。例如，对于变异的临床意义解读，通常采用常见变异类型（如 SNV/InDel）的排序技术和致病性机器学习模型，该技术主要是利用公开数据库/软件数据信息，进行机器学习训练，或是结合一定的人工经验逻辑，从而得到变异临床意义排序和分类算法。然而，由于信息来源不够全面，且局限于训练量，该技术的准确性无法保证，对于部分变异的解读结果可能会有错漏。

3) 检测获得的数据，经人工或自动化解读后的结果可能存在人工主观差异或解读错漏，故需对结果进行复核。由于不能快速准确的定位需重点复核的变异，导致复核工作量大，造成不必要的人工成本。

4) 由于 NGS 测序技术均一性问题，部分基因变异检测结果需要进一步进行实验验证，由于变异位点缺乏有效的质量分类标准，故对检测结果的验证主要依赖于人工主观判断，对于解读致病的变异大都需要变异验证以确保解读结果的准确性。常见的变异验证是针对变异位点所在区段人工进行引物设计、PCR（Polymerase Chain Reaction，聚合酶链式反应）扩增、一代测序、结果分析，会耗费大量人力、时间和检测成本。

发明内容

本公开提出了一种序列变异分析方法、系统以及存储介质，以实现快速定位需复核的变异，降低人工成本，提升复核效率，提高基因检测报告解读的准确性。

第一方面，本公开提出了一种序列变异分析方法，包括以下步骤：获取待分析序列变异数据；对所述待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集；将所述第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，并将所述第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果；取所述第一表型关系预测结果和所述第二表型关系预测结果的并集，获得第三表型关系预测结果。

第二方面，本公开提出了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时，实现上述的序列变异分析方法。

第三方面，本公开提出了一种序列变异分析系统，包括：获取模块，用于获取待分析序列变异数据；第一分析模块，用于对所述待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集，并将所述第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，将所述第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果，以及取所述第一表型关系预测结果和所述第二表型关系预测结果的并集，得到第三表型关系预测结果。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本公开的其它特征、目的和优点将会变得更加明显：

图 1 是本公开第一个实施例的序列变异分析方法的流程图；

图 2 是本公开一个实施例的预测模型的结构和应用示意图；

图 3 是本公开第二个实施例的序列变异分析方法的流程图；

图 4 是本公开第三个实施例的序列变异分析方法的流程图；

图 5 是本公开第四个实施例的序列变异分析方法的流程图；

图 6 是本公开一个示例的变异类型预测分析结果的示意图；

图 7 是本公开一个示例的表型关系预测结果的示意图；

图 8 是本公开实施例的序列变异分析系统的结构框图。

具体实施方式

为了使本公开的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本

公开进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本公开，并不用于限定本公开。为了说明本公开的技术方案，下面通过具体实施例来进行说明。

图1是本公开第一个实施例的序列变异分析方法的流程图。

如图1所示，序列变异分析方法包括以下步骤：

S1，获取待分析序列变异数据。

在本公开的实施例中，为了便于分析，使用待分析序列变异数据中的各变异唯一身份标识进行数据查找和处理。该唯一身份标识与变异具有确定的且唯一的映射关系。如本领域惯常使用的“染色体号-基因组坐标-参考序列-改变序列”（chr-pos-ref-alt，简称 cpra）作为变异的唯一身份标识；其中，chr 为染色体号，pos 为参考序列物理位置，ref 为参考序列，alt 为变异序列。

所述待分析序列变异数据包括至少一个变异位点信息。变异位点信息至少包括变异位点和变异位点的唯一身份标识。具体地，该变异位点信息可通过生物信息分析获得，生物信息分析可包括多种常规步骤，如质控、过滤、比对等。对获取该数据的技术手段不做限制，例如具体的序列信息可通过各种测序技术获得。

S2，对待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集。

在一优选的实施方式中，第一变异特征集中的各第一变异特征值为公开的变异位点的表型关系判定结果，第二变异特征集中的各第二变异特征值为变异位点等位基因频率数据以及功能预测数据。

具体地，当本方案应用于人类孟德尔遗传病相关变异分析时，前述表型关系是指变异与临床表型的关系，更确切的是指变异的致病性。

第一变异特征值可来自于各公开数据库的收录信息，例如 HGMD（<http://www.hgmd.cf.ac.uk/ac/index.php>）、ClinVar（<https://www.ncbi.nlm.nih.gov/clinvar/>）、LOVD（<https://www.lovd.nl/3.0/search>）等。在此，并不限制第一变异特征值的来源，包括具体的数据库及数据库的数目。各数据库所记录的表型关系判定结果通常基于个性化的判定细则，其判定结果也常采取个性化的字段进行描述。以 HGMD 为例，其表型关系判定结果包括 DM（disease-causing mutation）、DM?（likely disease-causing mutation）、DP（disease-associated polymorphism）、FP（in vitro or in vivo functional polymorphism）、DFP（disease-associated polymorphism with additional functional evidence）、R（retired record）。在本公开的实施例中，可利用待分析序列变异数据中的各变异唯一身份标识（如 cpra），在公开数据库中提取第一变异特征。

第二变异特征值中变异位点等位基因频率数据，可来自于各公开数据库的收录信息，例如 gnomAD (<https://gnomad.broadinstitute.org/>)、EVS (<http://evs.gs.washington.edu/EVS>)、1000 Genomes Project (<http://browser.1000genomes.org>) 等。在此，并不限制第二变异特征值中变异位点等位基因频率数据的来源，包括具体的数据库及数据库的数目。第二变异特征值中功能预测数据，可来自于现有多种功能预测软件的预测结果，从预测结果中可选取一个或多个具体指标作为特征值。此外，若公开数据库对功能预测数据进行了收录，则可直接从公开数据库中提取对应的功能预测数据。在此，并不限制第二变异特征值中功能预测数据的来源，包括具体的软件以及软件的数目、具体的指标以及指标的数目。所述的功能预测软件的预测数据可包括蛋白保守性预测数据、核酸保守性预测数据、剪切危害性预测数据、变异位点有害程度预测数据至少之一但不限于此。在本公开的实施例，可利用待分析序列变异数据中的各变异唯一身份标识（如 cpra），在公开数据库中提取第二变异特征。

S3，将第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，并将第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果。

S4，取第一表型关系预测结果和第二表型关系预测结果的并集，获得第三表型关系预测结果。

作为一优选实施方案，在步骤 S4 后还可以包括：

将所述第三表型关系预测结果与对应的变异解读结果进行比较；当所述第三表型关系预测结果与所述变异解读结果一致时，判定所述变异解读结果可信；当所述第三表型关系预测结果与所述变异解读结果不一致时，判定所述变异解读结果不可信。

其中，所述变异解读结果是指与所述第三表型关系预测结果采取不同方法获得的对所述待分析序列变异数据各变异位点的表型关系判定结果。人工解读结果是所述变异解读结果的常见类型。此外，可采取本领域的惯常手段，分别对判定为可信的变异解读结果、判定为不可信的变异解读结果添加自定义的标识，以便后续数据处理。本方案可以快速的对大量变异解读结果进行可信、不可信的判定，并在后续环节中按类型分别进行处理。例如，可对所述变异解读结果判定为不可信的变异位点进行复核。通常采取人工复核的方式进行确认。本方案可以极大的减少后续环节的工作量，降低人工成本，提高解读效率。

不同于现有技术中使用单一模型进行预测，包含步骤 S1 至 S4 的技术方案根据数据类型的不同，使用了两个相对独立的模型（即第一表型关系预测模型和第二表型关系预测模型）进行预测，两模型具有互补作用。第一表型关系预测模型能够弥补第二表型关系预测模型没有变异位点功能数据、共分离数据和新发数据等可能导致的假阴性，第二表型关系

预测模型能够弥补第一表型关系预测模型中因变异没有公开的变异位点表型关系判定结果导致的假阴性。

第一表型关系预测模型和第二表型关系预测模型，均可基于机器学习模型构建，构建方法为：

S31，构建数据集，包括第一数据集和第二数据集，第一数据集为选定变异位点的第一变异特征及该变异位点的可信表型关系判定结果，第二数据集为选定变异位点的第二变异特征以及该变异位点的可信表型关系判定结果。

其中，选定变异位点是指，包括全部表型关系判定结果类型的多个变异位点。在一优选实施方案中，所述可信表型关系判定结果可以是根据行业指南或共识人工对表型关系判定的结果；所述可信表型关系判定结果也可以是多个公开的变异位点表型关系判定结果均一致的结果。

S32，用所述第一数据集对第一机器学习模型进行训练和评估，用第二数据集对第二机器学习模型进行训练和评估。

在一优选实施方案中，第一机器学习模型、第二机器学习模型可选自逻辑回归、朴素贝叶斯、支持向量机和人工神经网络至少之一。

S33，根据所述第一机器学习模型的评估结果确定第一表型关系预测模型，根据第二机器学习模型的评估结果确定第二表型关系预测模型。

在一优选实施方案中，第一表型关系预测模型为逻辑回归模型，第二表型关系预测模型为神经网络模型。

当本方案应用于人类孟德尔遗传病相关变异分析时，具体过程如下：

在该实施例中，第一表型关系预测模型用于快速对公共数据库有记录的变异以统一标准实现致病性判定。第一表型关系预测模型的构建过程包括：

A2，构建第一数据集，第一数据集包括公共变异数据库中记录的致病性判定结果以及解读专家做出的致病性判定结果。

其中，公共变异数据库可包括 ClinVar、HGMD、LOVD、UMD (<http://umd-predictor.eu/>) 等数据库中的至少一个。

具体地，在从公共变异数据库收集变异数据之后，可利用变异注释工具对从各个公共变异数据库收集的变异数据进行注释，得到其cpra信息，并将该信息作为变异的唯一身份标识。以变异的cpra为索引，将各个公共变异数据库的变异数据进行汇总。

同时，由解读专家根据ACMG(The American College of Medical Genetics and Genomics, 美国医学遗传学与基因组学学会) 变异解读指南，结合人群频率数据库、预测软件和医学文献等信息，对相同变异进行致病性判定，该判定结果即为可信表型关系判定结果，并将

该结果作为训练数据的标准结果。

进一步地，需建立公共变异数据库中记录的致病性判定结果、解读专家做出的致病性判定结果与满足机器学习模型输入要求字符的唯一映射关系。

第一数据集的结构和内容如表1所示。

表1

染色体号-基因组坐标-参考序列-改变序列	ClinVar致病性判定	HGMD致病性判定	LOVD致病性判定	UMD致病性判定	解读专家致病性判定
chr8-90976737-T-C	Likely_pathogenic	DM	pathogenic	5-Causal	pathogenic

由于各个公共变异数据库和解读专家的致病性判定结果均为定序的离散型数据，因此本公开将第一数据集中的字符型数据转化为数值型数据，如表2所示。

表2

染色体号-基因组坐标-参考序列-改变序列	ClinVar致病性判定	HGMD致病性判定	LOVD致病性判定	UMD致病性判定	解读专家致病性判定
8-90976737-T-C	4	5	5	5	5

B2，将第一数据集中的数据按照第一预设比例随机分为第一训练集和第一测试集。

其中，第一预设比例可以是 8:2。

C2，利用第一训练集中的数据对第一机器学习模型进行训练，并利用第一测试集对训练得到的第一机器学习模型进行评估。

其中，第一机器学习模型以选择逻辑回归、朴素贝叶斯、支持向量机和人工神经网络等机器学习模型中的至少一个。

D2，根据评估结果从训练得到的多个第一机器学习模型中确定第一表型关系预测模型。

具体地，第一表型关系预测模型要实现的功能是根据多个不同的公共变异数据库的变异致病性判定输出唯一的变异致病性判定，而对变异进行致病性判定是一个特征较少的多分类问题，故可以选择逻辑回归、朴素贝叶斯、支持向量机和人工神经网络等机器学习模型来实现该功能。

作为一个实施方式，可选择多个不同的机器学习模型，并通过第一训练集对选择的各个机器学习模型各自独立进行训练，并通过第一测试集对训练的各个机器学习模型的性能进行评估，评估指标可包括准确率、精准率、召回率等。根据各个模型的性能评估结果，选择性能最优的一个模型作为本公开所使用的第一表型关系预测模型，例如，可将训练得

到的准确率最高的模型作为第一表型关系预测模型。

在利用第一表型关系预测模型预测表型关系时，可提取各变异的第一变异特征值；将第一变异特征值输入第一表型关系预测模型，即可得到第一表型关系预测结果。

需要说明的是，第一表型关系预测模型是对公共数据库有收录的变异、按预设的表型关系、输出预测结果。以人类孟德尔遗传病相关变异致病性判定为例，预设的表型关系通常有三类（致病的或可能致病的、良性的或可能良性的、意义不明确的）、五类（致病的、可能致病的、良性的、可能良性的、意义不明确的）的不同分法，对此并不做限制。

相应地，第二表型关系预测模型的构建过程可包括：

A3，构建第二数据集，第二数据集包括变异位点等位基因频率数据、功能预测数据以及可信表型关系判定结果。

具体地，等位基因频率数据可来源于 gnomAD、千人数据库、ExAC 数据库等；功能预测数据来源于各预测软件的数据，如 SIFT、Polyphen2、MutationTaster、GERP++、DANN 等。由此可见，与第一表型关系预测模型相区别的是，第二表型关系预测模型是基于判断表型关系的原始数据得到，而第一表型关系预测模型是基于公开的表型关系判定结果得到。

第二数据集中的变异位点需在多个公开的变异位点表型关系判定结果均一致，该一致的判定结果即为可信表型关系判定结果，并将该结果作为训练数据的标准结果。

进一步地，需建立等位基因频率数据、功能预测数据以及可信表型关系判定结果与满足机器学习模型输入要求字符的唯一映射关系。

B3，将第二数据集中的数据按照第二预设比例随机分为第二训练集和第二测试集。

其中，第二预设比例可以是 8:2。

C3，利用第二训练集中的数据对第二机器学习模型进行训练，并利用第二测试集对训练得到的第二机器学习模型进行评估。

其中，第二机器学习模型可以选择逻辑回归、朴素贝叶斯、支持向量机和人工神经网络等机器学习模型。

D3，根据评估结果确定训练得到最优第二机器学习模型，并将最优第二机器学习模型作为第二表型关系预测模型。

在利用第二表型关系预测模型预测表型关系时，可提取各变异的第二变异特征值；将第二变异特征值输入第二表型关系预测模型，即可得到第二表型关系预测结果。

在本公开的一个实施例中，实际应用中，基于待分析序列变异数据可通过人工解读的方式获得变异解读结果。同时，将相同变异位点的对应数据输入训练好的第一表型关系预测模型，获得第一表型关系预测结果；输入训练好的第二表型关系预测模型，获得第二表型关系预测结果；取第一表型关系预测结果与第二表型关系预测结果的并集作为第三表型

关系预测结果。可将变异解读结果与对应的第三表型关系预测结果进行比较，若两者结果一致则判定变异解读结果可信，否则判定不可信。不可信的变异解读结果需进一步人工复核。

需要说明的是，上述用于训练的可信表型关系判定结果是被认为准确的、少量的；用于与第三表型关系预测结果进行比较的解读结果可以是采取与第三表型关系预测结果任何不同方法获得的对待分析序列变异数据各变异位点的表型关系判定结果，如本领域技术人员的人工解读结果。

在本公开的一个实施例中，如图3所示，序列变异分析方法还可包括：

S5，过滤待分析序列变异数据中的第一变异位点，第一变异位点位于参考序列的简单重复区、与表型改变弱相关、且为非单碱基变异。

其中，简单重复区是指由1-5碱基重复单元构成的重复区域，例如AAA、CAACAACAACAA。与表型改变弱相关可以是指，变异位于非编码区或内含子区，且变异的等位基因频率大于0.05。变异位于非编码区或内含子区可以是指变异的功能注释为-（表示无功能注释）、intron（表示内含子）、3'-UTR（表示3'端的非翻译区）、5'-UTR（表示5'端的非翻译区）、nochange（与参考mRNA序列相比未发生改变）。

作为一个示例，第一变异位点还满足如下任一条件：

条件一：第一变异位点的总检出次数大于第一预设值，且低质量检出次数大于第二预设值；

条件二：位于满足条件一的变异位点参考序列物理位置上下游第三预设值内。

作为一个示例，可预先建立包含第一变异位点的第一变异位点数据集，其构建步骤如下：

A1，获取多个样本待分析序列变异数据。

具体地，该步骤与前述S1步骤相同。更具体地，可收集基因测序仪（如MGISEQ-2000测序仪）上固定测序芯片（如LCY171芯片）上的样本测序数据，进而从样本测序数据中得到变异位点和对应的变异唯一身份标识。

需要说明的是，实际使用中，第一变异位点数据集是随着样本的累计不断更新，实现对待分析序列变异数据的过滤。

B1，判断所得到的各变异位点是否位于参考序列的简单重复区，若是则记录该变异位点对应的变异唯一身份标识。

C1，对B1结果进行过滤，获得满足预设过滤条件的变异位点。

其中，所满足预设过滤条件的变异位点可为：功能注释为-（表示无功能注释）、intron（表示内含子）、3'-UTR（表示 3'端的非翻译区）、5'-UTR（表示 5'端的非翻译区）、nochange，且频率（表示变异在人群中的概率）大于 0.05 的变异。

具体地，可基于变异功能及变异频率进行过滤，过滤条件为：功能注释为-、intron、3'-UTR、5'-UTR、nochange，且频率大于 0.05。该过滤条件的设置是为了挑选出大概率不会引起基因表型改变的变异，即位于非外显子区的高频变异。

D1，统计 C1 结果中各变异位点的低质量检出次数和总检出次数。

具体地，经步骤C1处理后，以变异唯一身份标识为索引统计低质量（low quality）检出次数，高质量（high quality）检出次数以及总（total）检出次数。其中，总检出次数为低质量检出次数与高质量检出次数的加和。

E1，利用总检出次数大于第一预设值，低质量检出次数大于第二预设值，且为非单碱基变异的变异，构建第一变异位点数据集。

具体地，统计规则为：1）总检出次数 $\geq$ 第一预设值（如8），低质量检出次数 $\geq$ 第二预设值（如1），非单碱基变异的变异位点；2）查找1）中变异位点参考序列物理位置上下游第三预设值（如3bp）内的、仍位于简单重复区的、非单碱基变异的变异位点。需要说明的是，在简单重复区，变异位点与参考序列比对的位置会出现波动，不同数据库或分析软件的记录方式不同，但实质仍为同一变异。例如，参考序列ATATAT，变异位点序列AT，变异位点序列可分别与参考序列的第一、三、五位进行匹配。因此设置了统计规则2）以确保第一变异位点的全面收录。由满足上述统计规则1）和2）中任一个的变异位点组成第一变异位点数据集。

进一步地，在过滤待分析序列变异数据中的第一变异位点时，可先得到待分析序列变异数据中各变异位点的变异唯一身份标识，并基于该变异唯一身份标识进行过滤。分析过程具体可为：调用第一变异位点数据集，并基于待分析序列变异数据的变异唯一身份标识遍历第一变异位点数据集。若存在一致的变异唯一身份标识，则说明待分析序列变异数据中的该变异位点为第一变异位点，需被过滤；否则需保留。

作为一个示例，用于表型关系预测的数据，可以是经步骤S5过滤操作之后的数据，即步骤S2-S4可在步骤S5之后执行，并利用步骤S5过滤后的数据。由此，通过减少表型关系预测的数据，可避免不必要的预测，提高预测效率。

在本公开的一个实施例中，如图 4 所示，序列变异分析方法，方法还可包括：

S6，获取待分析序列变异数据的氨基酸水平注释结果。

S7，根据氨基酸水平注释结果判断待分析序列变异数据是否存在复杂变异。

其中，复杂变异可包括插入和替换复合事件、缺失和缺失复合事件、缺失和替换复合事件，且各类复杂变异对各具体变异的数目无限制。此外，可采取本领域的惯常手段，对判定为存在复杂变异和/或不存在复杂变异的位点添加自定义的标识，以便后续数据处理。

作为一个示例，存在复杂变异的变异位点可满足如下条件：

至少两个导致氨基酸改变的变异位点参考序列坐标重叠；或者，

至少两个导致氨基酸改变的变异位点参考序列坐标相邻且影响同一个编码氨基酸。

在该实施例中，考虑到利用变异检测工具（如GATK软件）检测变异时，只会单纯的检测到替换事件或插入事件，不能检测到包含插入和替换的复合事件。为弥补该缺陷，本公开对变异检测工具检测到的变异注释，进行复杂变异判断，具体为判断相同基因编码区相邻位置的变异是否影响同一个氨基酸残基编码。若存在复杂变异，说明需要进行复核，例如通过人工进行合并解读；反之亦然。

具体而言，获取待分析序列变异数据的氨基酸水平注释结果，通常采取HGVS命名规则对注释结果进行命名。对存在氨基酸改变的变异，则在满足如下任意条件时，判定存在复杂变异：1）存在至少两个相同基因的变异发生参考序列坐标重叠；2）存在至少两个相同基因的变异发生参考序列坐标相邻且影响同一个氨基酸编码。条件1）中，“坐标”是指变异对应的参考序列的位置，“重叠”是指 $\geq 1\text{bp}$ 的参考序列坐标重叠；该条件可以尽可能全面的筛选到复杂变异。条件2）中，“相邻”是指参考序列上下游4bp范围内；该条件可以准确的筛选到复杂变异。由此，通过复杂变异的判断能够弥补现有变异检测工具的缺陷，进而将复杂变异进行合并解读，可提升变异的解读效率。

作为一个示例，步骤S6中的待分析序列变异数据可以是步骤S1中的待分析序列变异数据，也可以是经第三表型关系预测结果与对应的变异解读结果进行比较后、带有判定结果的待分析序列变异数据。

在本公开的一个实施例中，如图5所示，序列变异分析方法还可包括：

S8，对待分析序列变异数据进行特征提取，得到第三变异特征集。

其中，第三变异特征集中的第三变异特征值可包括变异支持数据（当待分析序列变异数据是通过测序手段获得时，该变异支持数据可包括质量值、测序深度、支持变异的 reads 比值）、等位基因频率（如 ESP6500_MAF、G1000_AF）。

S9，将第三变异特征集输入训练好的变异类型预测模型，得到待分析序列变异数据中各变异位点所属各变异类型的预测概率。

其中，变异类型可包括纯合变异、杂合变异、无变异。无变异可为参考序列在此位点的纯合基因型。

S10，根据预测概率确定对应变异位点的变异类型预测结果。

具体地，确定变异类型预测结果的方法为：将各变异类型的预测概率进行比较，并确定预测概率最大的变异类型；将最大的预测概率与预设阈值进行比较；如果预测概率大于预设阈值，则确定该最大的预测概率所对应的变异类型为该变异位点的变异类型预测结果。

作为一个示例，变异类型预测模型的构建过程可包括：

A4，构建第三数据集，第三数据集包括质量值、测序深度、支持变异的 reads 比值、等位基因频率以及可信的变异类型判定结果。

其中，可信的变异类型判定结果可以是使用行业内共识金标准手段验证所获得的结果，例如，sanger 验证结果。

B4，将第三数据集中的数据按照第三预设比例随机分为第三训练集和第三测试集。

其中，第三预设比例可以是 8:2。

C4，利用第三训练集中的数据对第三机器学习模型进行训练，并利用第三测试集对训练得到的第三机器学习模型进行评估。

具体地，以计算贝叶斯分类器所需的先验概率和条件概率，构建变异类型预测模型。

D4，根据评估结果从训练得到的多个第三机器学习模型中确定第三表型关系预测模型。

通过第三测试集中的数据对变异类型预测模型进行模型测试，通过变异类型预测模型判断测试集数据的所属分类，并且根据判断正确的数量来求得变异类型预测模型的准确度。

具体地，挑选经过sanger验证的变异位点作为训练样本，收集其相关数据组成第三训练集，这些训练样本一共被分为 K 个变异类型 c_1 、 c_2 、...、 c_K ；预处理获得的变异数据，获取变异信息相关特征值，记为 x_i ；基于选取的训练集统计出每个特征值的先验概率，该先验概率可通过第三训练集中的各类样本出现的次数来估计，以变异类型为 c_1 时特征值 x_1 为例，其先验概率为：

$$P(x_i = x_1 | c_k = c_1) = \frac{x_i = x_1 \text{ 的数据总量}}{\text{数据总量}}$$

对于每个类别 c_k ，基于第三训练集计算先验概率 $P(c_k)$ 和类条件概率，利用贝叶斯公式估计给定训练特征 x 的条件概率 $P(c_k|x)$ ，即

$$P(c_k|x) = \frac{P(c_k) \prod_{i=1}^n P(x_i|c_k)}{P(x)}$$

其中， $P(x)$ 是用于归一化的证据因子，对所有类别的取值为固定值。

通过第三测试集对以上机器学习模型的性能进行评估，评估指标可包括准确率、精准率，召回率。

在利用该模型时，将待测变异经过构建的变异类型预测模型进行预测，获得每个类别的概率 $P(c_k|x)$ 。取 $P_{\max}(c_k|x)$ ，判定数据 x 属于对应的类别 c_k 。或者，设定质控阈值，当且仅当 $P_{\max}(c_k|x)$ 大于预设阈值时，判定数据 x 属于对应的类别 c_k 。

作为一个示例，步骤 S8 中的待分析序列变异数据可以是步骤 S1 中的待分析序列变异数据，也可以是经第三表型关系预测结果与对应的变异解读结果进行比较后、带有判定结果的待分析序列变异数据，还可以是带有是否存在复杂变异结果的待分析序列变异数据。

作为一优选实施方案，在步骤 S10 后还可以包括：当所述变异类型预测结果仅包含参考序列基因型，则判定该变异位点不可信；当所述变异类型预测结果包含至少一个非参考序列基因型，则判定该变异位点可信。此外，可采取本领域的惯常手段，对判定为不可信和/或可信的变异位点添加自定义的标识，以便后续数据处理。本方案可以快速的对大量变异位点进行可信、不可信的判定，并在后续环节中按类型分别进行处理。例如，仅对判定为不可信的变异位点进行验证。又如，仅对判定为不可信的、与表型改变相关的变异位点进行验证，其中与表型改变相关的变异位点可为致病变异。通常采取本领域公认的金标准方式进行验证，例如一代测序验证。本方案可以极大的减少后续验证环节的工作量，降低人工成本，提高解读效率。

由此，本公开的序列变异分析方法，可快速判定该变异是否需要复核，是否需要实验验证等。对整个解读环节，一定程度上减少了变异实验验证成本，且对于人工解读或自动化解读结果进行分析，快速输出需复核的变异，减少了人工审核成本，确保结果的准确性。

下面结合一个具体实施方式描述本公开实施例的序列变异分析方法：

在该具体实施方式中，以人类孟德尔遗传病变异分析为例，从临床数据中选择599例样本的7937个变异解读数据，作为待分析序列变异数据，该临床数据需是合法合规获取的。具体实施如下：

1 第一变异位点过滤

1.1 构建第一变异位点数据集

1.1.1 持续收集MGISEQ-2000测序仪上LCY171测序芯片上的样本测序数据，总共收集3038个样本的数据，将位于简单重复区的变异标记为yes。例如，chr1-241663902-T-TGAGAGAGA，参考序列存在“GA”重复，标记为yes。

1.1.2 基于变异功能及等位基因频率对标记为yes的变异位点进行过滤。实施例中可设置的过滤条件为：①功能注释为“-”、“intron”、“3-UTR”、“5-UTR”、“nochange”的变异；②频率大于0.05的变异，具体可通过G1000、ESP6500、genomAD等人群数据库进行判定。

例如，变异chr11-108150207-CT-C，在G1000、ESP6500、genomAD三个数据库中记录

的变异的等位基因频率均大于0.05，因此该变异被过滤。

1.1.3 对1.1.2过滤后保留的变异位点分别统计低质量（low quality）检出次数、高质量（high quality）检出次数以及总（total）检出次数。其中，总检出次数为低质量检出次数与高质量检出次数的加和。

1.1.4 构造第一变异位点数据集

收录满足以下两条件之一的变异位点进行第一变异位点数据集的构建。第一变异位点数据集中至少需记录被收录变异位点的cpri信息。

①标签为yes，总检出次数 ≥ 8 ，低质量检出次数 ≥ 1 ，非单碱基变异（即ref碱基数 $\neq$ alt碱基数）的变异；

②查找①中附近3bp内的ref碱基数 $\neq$ alt碱基数，标签为yes的变异。

1.2 查找及过滤第一变异位点

提取待分析数据的变异cpri信息，遍历第一变异位点数据集。若cpri与数据集中某一cpri一致，则输出“Yes”，说明存在第一变异位点，是需过滤的变异；否则输出“No”，说明不存在第一变异位点，是需保留的变异。

2 复杂变异判定

具体实施如下：

2.1 根据是否存在pHGVs点的结果判断变异是否存在氨基酸改变。

在一些实施方式中，复杂变异判定还可以处理接收到的数据（例如，以VCF格式、FASTQ格式等的输入数据），利用生信软件（如GATK、ANNOVAR、Alamut等）进行氨基酸水平的注释，确定变异是否包含pHGVs点的结果。例如，变异chr7-142458526-A-G，氨基酸水平的注释结果为p.Asn54Ser，即包含pHGVs点。

2.2 将存在氨基酸改变的变异进行位置判定，判断是否存在 ≥ 2 个编码区相邻的同一基因变异。例如，检出变异同时包含chr7-142458526-A-G（c.161A>G，p.Asn54Ser）和chr7-142458527-C-G（c.162C>G，p.Asn54Lys），前者为cDNA 161位后者为162位，两者相邻。

2.3 根据相邻变异的变异类型及变异氨基酸改变位置判定是否存在复杂变异。以上两个变异（即chr7-142458526-A-G，chr7-142458527-C-G）都为SNP（Single Nucleotide Polymorphism，单核苷酸多态性），氨基酸改变的位置按照pHGVs均导致第54位氨基酸Asn的改变，两者具有相同的氨基酸改变位置，判定为存在复杂变异。

对于判定为存在复杂变异的变异位点，使用自定义字符进行标识。后续可进行合并解读，从而可提高解读效率。

3 表型关系预测

3.1 第一表型关系预测模型

3.1.1 搭建

3.1.1.1 训练数据收集

从公共变异数据库ClinVar和HGMD获取总共3万多个变异及其致病性判定作为训练数据。

3.1.1.2 训练数据预处理和特征提取

①训练数据预处理

利用变异注释工具，对训练数据进行注释，得到其cpra信息（基因组版本为hg19），并将该信息作为变异的唯一身份标识。以变异的cpra为索引，将公共变异数据库的变异及其致病性判定整理汇总到一个训练汇总数据库中。由解读专家根据ACMG变异解读指南，结合人群频率数据库、预测软件和医学文献等，对汇总数据库中的变异人工进行致病性判定，并将判定结果作为训练数据的标准结果。经预处理后的训练汇总数据库结构和内容如表3所示。

表3

染色体号-基因组坐标-参考序列-改变序列	ClinVar致病性判定	HGMD致病性判定	解读专家致病性判定
13-32968886-G-A	Pathogenic	DM	Pathogenic/ Likely Pathogenic
13-32893467-G-A	Pathogenic/Likely_pathogenic	DM	Pathogenic/ Likely Pathogenic
13-32971167-G-C	Benign/Likely_benign	DM?	VUS
14-75485594-C-T	Uncertain_significance	DM?	VUS
7-6031639-T-C	Conflicting_interpretations_of_pathogenicity (submission without pathogenic)	DP	VUS
2-48033981-T-TT TGA	Likely_benign	FP	VUS
16-89849480-C-T	Benign	DP	Benign/ Likely_benign
2-47643457-G-A	Benign	FP	Benign/

			Likely benign
...	...	...	...

②训练数据特征提取

由于公共变异数据库和解读专家的致病性判定结果均为定序的离散型数据，故可以将汇总数据库中的字符型数据根据表4~表6所示的对应关系，转化为数值型数据，完成第一训练集的构建。经特征提取后的第一训练集的结构和内容如表8所示。

表4

HGMD致病性判定（字符型）	HGMD致病性判定（数值型）
Null（该数据库无该变异记录）	0
R	1
FP	2
DP	3
DFP	4
DM?	5
DM	6

表5

ClinVar致病性判定（字符型）	ClinVar致病性判定（数值型）
Null(该数据库无该变异记录)	0
not_provided	1
risk_factor	1
drug_response	1
other	1
Benign	2
Benign/Likely_benign	3
Likely_benign	4
Conflicting_interpretations_of_pathogenicity (submission without pathogenic)	5
Uncertain_significance	6
Conflicting_interpretations_of_pathogenicity (submission with pathogenic)	7

Likely_pathogenic	8
Likely_pathogenic, risk_factor	8
Pathogenic/Likely_pathogenic	9
Pathogenic/Likely_pathogenic, risk_factor	9
Pathogenic	10

表6

解读专家致病性判定（字符型）	解读专家致病性判定（数值型）
Pathogenic/Likely Pathogenic	3
VUS	2
Benign/Likely Benign	1

表7

染色体号-基因组坐标 -参考序列-改变序列	ClinVar致病性判定	HGMD致病性判定	解读专家致病性判定
13-32968886-G-A	10	6	3
13-32893467-G-A	9	6	3
13-32971167-G-C	3	5	2
14-75485594-C-T	6	5	2
7-6031639-T-C	5	3	2
2-48033981-T-TTTGA	4	2	2
16-89849480-C-T	2	3	1
2-47643457-G-A	2	2	1
...	...	...	...

3.1.1.3 模型选取和训练

以第一预设比例如8:2，将第一训练集随机拆分为第一训练集和第一测试集。通过逻辑回归模型，使用第一训练集进行训练。

3.1.1.4 模型评估和确定

通过第一训练集对逻辑回归模型进行训练，经第一测试集评估其准确率为91.0%，测试性能数据如表8所示。

表8

性能指标	精准率	召回率	f1得分
------	-----	-----	------

1	0.91	0.84	0.87
2	0.87	0.92	0.90
3	0.97	0.97	0.97
宏平均	0.92	0.91	0.91
加权平均	0.91	0.91	0.91

3.1.2 应用

3.1.2.1 预测数据收集

收集待预测变异，如NM_000267.3(NF1):c.1722-2A>G，NM_000057.4(BLM):c.893C>T (p.Thr298Met)和NM_000244.3(MEN1):c.670-6C>T，以及其在公共变异数据库ClinVar和HGMD的致病性判定作为待预测数据。可选地，待预测变异可取自上述第一变异位点过滤后的变异位点。

需要说明的是，若待预测变异在这2个公共变异数据库均无结果，则该变异无法通过本公开实现致病性判定。

3.1.2.2 训练数据预处理和特征提取

利用变异注释工具对待预测变异进行注释，得到其cpva信息(基因组版本为hg19)，并将该信息作为变异的唯一标识。以变异的cpva为索引，将待预测变异在公共变异数据库ClinVar和HGMD的致病性判定整理汇总到一个数据库中。因为公共变异数据库的致病性判定结果均为定序的离散型数据，根据表8所示转化逻辑将预测汇总数据库中的字符型数据转化为数值型数据，完成预测数据处理。经预处理和特征提取后的预测汇总数据库结构和内容如表9所示。

表9

染色体号-基因组坐标-参考序列-改变序列	ClinVar致病性判定	HGMD致病性判定
17-29550460-A-G	6	10
15-91295110-C-T	5	3
11-64575158-G-A	5	5

3.1.2.3 模型结果输出

将预测汇总数据库中的数据转换成模型接受的输入格式，并将相应数据输入训练好的第一表型关系预测模型，即可得到上述3个待预测变异的变异致病性判定，即第一表型关系预测结果，结果如表10所示。通过表4和表5所示的数值型和字符型数据的转换可以得到

本模型。将NM_000267.3(NF1):c.1722-2A>G、NM_000057.4(BLM):c.893C>T (p.Thr298Met)和NM_000244.3(MEN1):c.670-6C>T分别预测为Pathogenic/Likely Pathogenic、Benign/Likely benign和VUS。

表10

染色体号-基因组坐标-参考序列-改变序列	ClinVar 致病性判定	HGMD 致病性判定	模型预测致病性判定
17-29550460-A-G	6	10	3
15-91295110-C-T	5	3	1
11-64575158-G-A	5	5	2

3.2 第二表型关系预测模型

3.2.1 搭建

3.2.1.1 获取经专家解读具有明确致病性结果的变异数据。

具体地，变异数据可取自ClinVar数据库，ClinVar数据库(VCF_GRCH37，fileDate=2017-10-29)中包含近30万个人类遗传变异，其涵盖了从致病变异到良性变异在内的各种类别的变异。该ClinVar数据库中包含明确致病位点46585个，筛选在HGMD数据库中为被认为是“DM”，且在dbNSFP (<https://sites.google.com/site/jpopgen/dbNSFP>) 数据库中的变异位点共21105个，选择其中的一半共10552个作为第二训练集中的致病变异，另10553个致病位点组成第二测试集的第一测试子集。该ClinVar数据库包含良性位点23892个，选择在dbNSFP数据库中的位点共4664个作为第二训练集中的良性变异。对dbNSFP数据库中缺失值的处理，为直接赋值为0。

3.2.1.2 预处理第二训练集和第一测试子集中的变异数据，以变异的cpra为变异的唯一标识及索引，获取用于训练的变异信息特征值 x_n 。具体实施例中，该模型基于等位基因频率数据库1000 Genomes Project、ESP6500，以及多个功能预测软件数据。具体地，功能预测软件可分为，用于蛋白保守性预测例如SIFT、Polyphen2、MutationTaster，用于核酸保守性预测例如GERP++，用于剪切危害性预测例如dbSCNV，用于变异位点有害程度预测例如DANN。提取各功能预测软件的一个或多个技术指标作为特征值，共获取变异的多个特征值（计为 x_n ），具体特征值分别为：SIFT_converted_rankscore，Polyphen2_HDIV_rankscore，Polyphen2_HVAR_rankscore，LRT_converted_rankscore，LRT_Omega，MutationTaster_converted_rankscore，MutationAssessor_score_rankscore，FATHMM_converted_rankscore，PROVEAN_converted_rankscore，MetaSVM_rankscore，

MetaLR_rankscore , REVEL_score , CADD_raw_rankscore , DANN_rankscore , GERP++_RS_rankscore , splicing_consensus_ada_score , splicing_consensus_rf_score , phyloP100way_vertebrate_rankscore , phyloP20way_mammalian_rankscore , phastCons100way_vertebrate_rankscore , phastCons20way_mammalian_rankscore , SiPhy_29way_logOdds_rankscore, 1000Gp3_AF, ESP6500_AA_AF。可选的各具体特征的详情情况可参见 https://drive.google.com/file/d/1Vse3b_qw_E46eDcsuLegF5HxpodAZ6Og/view 、 https://drive.google.com/file/d/12-sSAB_hP9fWNxvJZ6IpMZvGXCvYmfns/view。

3.2.1.3 构造第二机器学习模型。基于变异信息相关特征值 x_n 及变异的唯一致病性判定, 可以选择逻辑回归、朴素贝叶斯、支持向量机和人工神经网络等机器学习模型作为第二机器学习模型来实现预测功能。

在本公开的一个具体实施例中, 可选择神经网络模型作为第二机器学习模型。构建的神经网络模型可包含三个隐藏层, 节点个数分别为16、128、16; 权值初始化函数为uniform, 激活函数为hard_sigmoid, 优化器采用Adadelta方法的默认参数, 神经元失活概率dropout_rate为0.05, 训练迭代次数800, 一次性训练所选取的样本数batch_size为64。

3.2.a4 模型结果评估

可通过2个独立的测试子集对训练得到的第二机器学习模型进行性能评估, 以测试及验证模型的准确性。测试子集1即为上述的第一测试子集, 测试子集2可以是基于临床解读积累的通过ACMG解读出来的变异结果, 筛选临床明确认为是Pathogenic(致病变异)和Benign(良性变异)的位点, 其中, 对于多次解读到并且解读结果不一致位点选择解读结果次数最多的结果。筛选在dbNSFP数据库中的变异, 去除训练集中的变异, 得到共728个变异, 其中致病变异618个, 良性变异108个。

通过以上2个测试子集对训练得到的第二机器学习模型的性能进行评估, ClinVar数据库中10553个致病位点, 准确性为98.77%; 临床728个位点, 准确性为97.79%。ROC(Receiver Operating Characteristic, 受试者工作特征曲线)结果如图6所示, 对临床728个位点预测, AUC可达0.99。类似方法中REVEL测试最好效果为AUC=0.957; MLP测试最好效果为AUC=0.94; DANN测试最好效果为AUC=0.9459。显然, 本公开的预测子模型较以往方法AUC可提升4~5个百分点, 可以辅助遗传分析进行快速致病性审核, 用以进行变异致病性质控, 可达较高的准确性。

3.2.b 应用

3.2.b1 收集待预测变异的特征值, 即收集待预测变异在训练数据所用的各特征值 x_n , 具体地与3.2.1.2所述相同的多个特征值。

其中，待预测变异可以是取自上述第一变异位点过滤后的变异位点。

3.2.b2 将变异的特征值输入第二表型关系预测模型，即可得到待预测变异的表型关系预测结果，即第二表型关系预测结果。

3.3 取第一表型关系预测结果和第二表型关系预测结果的并集，获得第三表型关系预测结果。将第三表型关系预测结果与相应的变异解读结果进行比较，当所述第三表型关系预测结果与所述变异解读结果一致时，判定所述变异解读结果可信；当所述第三表型关系预测结果与所述变异解读结果不一致时，判定所述变异解读结果不可信。其中，变异解读结果不可信的变异需进一步人工复核。

4 变异类型预测

4.1 模型搭建

4.1.1 搭建训练集。从经过sanger验证变异结果的历史测序数据中选取843个变异位点，将其按照4:1的比例随机分为第三训练集和第三测试集，第三训练集用来对模型进行训练，第三测试集用来对模型进行测试。具有674个变异位点的训练集具体数据分布为：269个无变异位点，212个纯合变异位点，193个杂合变异位点。

4.1.2 预处理获得的674个变异位点数据，获取变异信息相关特征值 x_n 。选取的特征值可包括 x_1 质量值、 x_2 测序深度、 x_3 支持变异的reads比值、 x_4 ESP6500_MAF（等位基因频率）、 x_5 G1000_AF（等位基因频率）这五维特征，例如，变异chr2-223086092-G-A，sanger验证为杂合变异，其特征值预处理为： $x_1 = 99$ ； $x_2 = 250$ ； $x_3 = 0.4480$ ； $x_4 = 0.000077$ ； $x_5 = 0.001597$ ；

4.1.3 针对变异信息相关特征值 x_n ，基于选取的第三训练集统计出每个特征值类的先验概率，通过第三训练集中各类数据出现的次数估计。以变异类型为c1时特征值 x_1 为例，其先验概率即：

$$P(x_i = x_1 | c_k = c_1) = \frac{x_i = x_1 \text{ 的数据总量}}{\text{数据总量}}$$

4.1.4 训练。基于674个变异位点的第三训练集计算先验概率 $P(c_k)$ 和类条件概率。第三训练集中的样本一共被分为 K 个类别，如 $K=3$ ，类别为 c_1, c_2, c_3 。对于类别 c_k ($k=1, 2, 3$)，可令 $c_1 = 0$ 代表无变异位点， $c_2 = 1$ 代表纯合变异位点， $c_3 = 2$ 代表杂合变异位点，利用贝叶斯公式估计在给定训练数据 x 时的条件概率 $P(c_k|x)$ ，即

$$P(c_k|x) = \frac{P(c_k) \prod_{i=1}^n P(x_i|c_k)}{P(x)}$$

以变异 chr2-223086092-G-A 为例，训练计算的结果为：
 $P(c_1|x) = 0.00190984495430601$ ； $P(c_2|x) = 3.63E - 07$ ； $P(c_3|x) = 0.998089791607098$ 。

4.1.5 模型结果评估及质控阈值确定

通过第三测试集对上述训练得到的机器学习模型的性能进行评估，评估指标可包括准确率、精准率、召回率。利用上述选取的169个变异数据作为第三测试集。选择大于预设阈值且最大预测概率所对应的变异类型为该变异位点的变异类型预测结果，其中预设阈值设定为0.8。具体测试结果详见表11。需要说明的是，变异类型预测模型的预设阈值可根据实际情况调整。

表11

	精 确 率 (Precision)	回报率 (Recall)	准确率 (Accuracy)
无变异	95.56%	89.58%	95.86%
纯合变异	88.24%	100.00%	96.45%
杂合变异	100.00%	93.75%	98.22%

4.2 应用

4.2.1 构建的变异类型预测模型预测待测变异位点属于每个类别的概率 $P(c_k|x)$ ，其中，待测变异位点可以是取自上述第一变异位点过滤后和经第一表型预测模型、第二表型预测模型判定为可信的不需人工复核的变异位点。

4.2.2 变异类型预测结果判定方法具体为，选择大于预设阈值 0.8 且最大预测概率所对应的变异类型为该变异位点的变异类型预测结果。以变异解读结果为致病变异的位点为例，当变异类型预测结果为纯合变异或杂合变异时，则判定该变异位点可信，无需进行实验验证；当变异类型预测结果为无变异类型时，则判定该变异位点不可信，需进行实验验证。

本公开整体性能评估：

选取599例样本的7937个变异位点作为待分析序列变异数据。对此变异数据通过人工解读的方式获得解读结果，将人工解读结果与本公开各环节结果进行比较。经第一变异位点过滤，过滤掉690个第一变异位点，进而可减少8.69%的复核工作量。经复杂变异判定，确定存在34个复杂变异位点，需要合并后进行解读，可弥补变异检测工具的缺陷。经第一表型关系预测模型和第二表型关系预测模型的分析结果（如图7所示），相比于其他单纯的致病性预测模型，两者相互结合可获得解读结果中与两个模型不一致的变异位点2041个，确定7247个变异位点中有5206个的解读结果可信无需复核，可减少71.84%复核工作量，且得到较全面的人工复核量。经变异类型预测模型分析，可进一步确认76个致病变异中的61个无需进行变异验证，减少80.26%变异验证工作量。

综上，本公开实施例的序列变异分析方法，能够快速、准确、全面的确定需过滤、复核、验证的变异。只需对判定为不可信的解读结果进行审核和维护，对判定为不可信的致

病变进行验证。从而大大降低了人工成本，提升了基因变异位点的复核、验证效率，缩短了基因检测报告的出具时间，同时也提高了基因检测报告解读的准确性，优化了整个解读分析过程。

基于上述的序列变异分析方法，本公开提出了一种计算机可读存储介质。

在该实施例中，计算机可读存储介质上存储有计算机程序，计算机程序被处理器执行时，实现上述的序列变异分析方法。

图 8 是本公开序列变异分析系统的具体实施方式之一的结构框图。

序列变异分析系统 100 包括：获取模块 110 和第一分析模块 120。

其中，获取模块 110 用于获取待分析序列变异数据；第一分析模块 120 用于对待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集，并将第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，将第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果，以及取第一表型关系预测结果和第二表型关系预测结果的并集，得到第三表型关系预测结果。

在本公开的一个实施例中，序列变异分析系统 100 还可包括：第二分析模块 130。

其中，第二分析模块 130 用于过滤待分析序列变异数据中的第一变异位点，第一变异位点位于参考序列的重复区、与表型改变弱相关、且为非单碱基变异。

在本公开的一个实施例中，序列变异分析系统 100 还可包括：第三分析模块 140。

其中，第三分析模块 140 用于获取待分析序列变异数据的氨基酸水平注释结果，并根据氨基酸水平注释结果判断待分析序列变异数据是否存在复杂变异。

在本公开的一个实施例中，序列变异分析系统 100 还可包括：第四分析模块 150。

其中，第四分析模块 150 用于对待分析序列变异数据进行特征提取，得到第三变异特征集，并将第三变异特征集输入训练好的变异类型预测模型，得到待分析序列变异数据中各变异位点所属各变异类型的预测概率，以及根据预测概率确定对应变异位点的变异类型预测结果。

需要说明的是，本公开实施例的序列变异分析系统 100 其他具体实施方式，可参见本公开上述实施例的序列变异分析方法的具体实施方式。

综上，本公开实施例的序列变异分析系统，能够快速、准确、全面的确定需过滤、复核、验证的变异。只需对判定为不可信的解读结果进行审核和进行维护，对判定为不可信的致病变异进行验证。从而大大降低了人工成本，提升了基因变异位点的复核、验证效率，缩短了基因检测报告的出具时间，同时也提高了基因检测报告解读的准确性，优化了整个解读分析过程。

以上实施例仅用以说明本公开的技术方案，而非对其限制；尽管参照前述实施例对本

公开进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本公开各实施例技术方案的精神和范围，均应包含在本公开的保护范围之内。

权利要求书

1、一种序列变异分析方法，其特征在于，包括以下步骤：

获取待分析序列变异数据；

对所述待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集；

将所述第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，并将所述第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果；

取所述第一表型关系预测结果和所述第二表型关系预测结果的并集，获得第三表型关系预测结果。

2、如权利要求 1 所述的序列变异分析方法，其特征在于，在所述获得第三表型关系预测结果后还包括：

将所述第三表型关系预测结果与对应的变异解读结果进行比较；

当所述第三表型关系预测结果与所述变异解读结果一致时，判定所述变异解读结果可信；

当所述第三表型关系预测结果与所述变异解读结果不一致时，判定所述变异解读结果不可信。

3、如权利要求 1 所述的序列变异分析方法，其特征在于，所述方法还包括：

过滤所述待分析序列变异数据中的第一变异位点，所述第一变异位点位于参考序列的简单重复区、与表型改变弱相关、且为非单碱基变异。

4、如权利要求 3 所述的序列变异分析方法，其特征在于，所述与表型改变弱相关是指，变异位于非编码区或内含子区，且所述变异的等位基因频率大于 0.05。

5、如权利要求 3 所述的序列变异分析方法，其特征在于，所述第一变异位点还满足如下任一条件：

条件一：总检出次数大于第一预设值，且低质量检出次数大于第二预设值；

条件二：位于满足所述条件一的变异位点的参考序列物理位置上下游第三预设值内。

6、如权利要求 1 所述的序列变异分析方法，其特征在于，所述方法还包括：

获取所述待分析序列变异数据的氨基酸水平注释结果；

根据所述氨基酸水平注释结果判断所述待分析序列变异数据是否存在复杂变异。

7、如权利要求 6 所述的序列变异分析方法，其特征在于，存在复杂变异的变异位点满足如下条件：

至少两个导致氨基酸改变的变异位点参考序列坐标重叠；或者，

至少两个导致氨基酸改变的变异位点参考序列坐标相邻且影响同一个编码氨基酸。

8、如权利要求 1 所述的序列变异分析方法，其特征在于，所述方法还包括：

对所述待分析序列变异数据进行特征提取，得到第三变异特征集；

将所述第三变异特征集输入训练好的变异类型预测模型，得到所述待分析序列变异数据中各变异位点所属各变异类型的预测概率；

根据所述预测概率确定对应变异位点的变异类型预测结果。

9、如权利要求 8 所述的序列变异分析方法，其特征在于，所述根据所述预测概率确定对应变异位点的变异类型预测结果，具体为：

选择大于预设阈值且最大所述预测概率所对应的变异类型为该变异位点的变异类型预测结果。

10、如权利要求 9 所述的序列变异分析方法，其特征在于，确定对应变异位点的变异类型预测结果后还包括：

当所述变异类型预测结果仅包含参考序列基因型，则判定该变异位点不可信；

当所述变异类型预测结果包含至少一个非参考序列基因型，则判定该变异位点可信。

11、如权利要求 8 所述的序列变异分析方法，其特征在于，所述第三变异特征值包括变异支持数据以及等位基因频率，所述变异类型包括纯合变异、杂合变异、无变异。

12、如权利要求 1 所述的序列变异分析方法，其特征在于，所述第一变异特征集中的各第一变异特征值为公开的变异位点的表型关系判定结果，所述第二变异特征集中的各第二变异特征值为变异位点等位基因频率数据以及功能预测数据。

13、如权利要求 12 所述的序列变异分析方法，其特征在于，所述第二变异特征值中的所述功能预测数据包括蛋白保守性预测数据、核酸保守性预测数据、剪切危害性预测数据、变异位点有害程度预测数据至少之一。

14、一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时，实现如权利要求 1-13 中任一项所述的序列变异分析方法。

15、一种序列变异分析系统，其特征在于，包括：

获取模块，用于获取待分析序列变异数据；

第一分析模块，用于对所述待分析序列变异数据进行特征提取，得到第一变异特征集、第二变异特征集，并将所述第一变异特征集输入训练好的第一表型关系预测模型，得到第一表型关系预测结果，将所述第二变异特征集输入训练好的第二表型关系预测模型，得到第二表型关系预测结果，以及取所述第一表型关系预测结果和所述第二表型关系预测结果的并集，得到第三表型关系预测结果。

16、如权利要求 15 所述的序列变异分析系统，其特征在于，所述系统还包括：

第二分析模块，用于过滤所述待分析序列变异数据中的第一变异位点，所述第一变异位点位于参考序列的重复区、与表型改变弱相关、且为非单碱基变异。

17、如权利要求 15 所述的序列变异分析系统，其特征在于，所述系统还包括：

第三分析模块，用于获取所述待分析序列变异数据的氨基酸水平注释结果，并根据所述氨基酸水平注释结果判断所述待分析序列变异数据是否存在复杂变异。

18、如权利要求 15 所述的序列变异分析系统，其特征在于，所述系统还包括：

第四分析模块，用于对所述待分析序列变异数据进行特征提取，得到第三变异特征集，并将所述第三变异特征集输入训练好的变异类型预测模型，得到所述待分析序列变异数据中各变异位点所属各变异类型的预测概率，以及根据所述预测概率确定对应变异位点的变异类型预测结果。