

(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
2024 年 10 月 24 日 (24.10.2024)

WIPO | PCT

(10) 国际公布号
WO 2024/217010 A1

(51) 国际专利分类号:

G16B 40/00 (2019.01) G16H 50/70 (2018.01)
C12Q 1/6883 (2018.01)

(21) 国际申请号:

PCT/CN2023/137633

(22) 国际申请日:

2023 年 12 月 8 日 (08.12.2023)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

202310461017.6 2023年4月18日 (18.04.2023) CN

(71) 申请人: 中国科学院深圳先进技术研究院 (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(72) 发明人: 王中昊 (WANG, ZhongHao); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。殷鹏 (YIN, Peng); 中国

广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。黄华振 (HUANG, HuaZhen); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。朱木春 (ZHU, MuChun); 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(74) 代理人: 北京中巡通大知识产权代理有限公司 (BEIJING ZHONG XUN TONG DA INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国北京市西城区西直门外大街 1 号院 2 号楼 7 层 7C1, Beijing 100044 (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, 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, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,

(54) Title: GENE AND DISEASE ASSOCIATION ANALYSIS METHOD AND APPARATUS, COMPUTER DEVICE AND STORAGE MEDIUM

(54) 发明名称: 基因与疾病关联分析方法、装置、计算机设备及存储介质

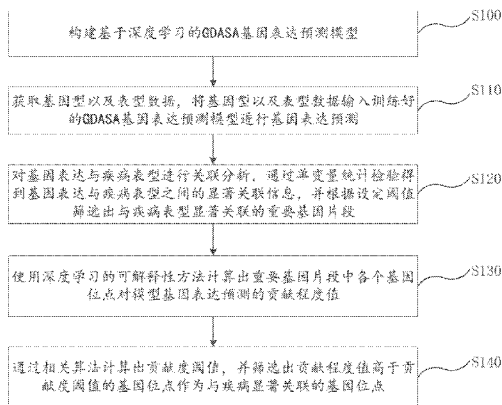


图1

- S100 Construct a deep learning-based GDASA gene expression prediction model
S110 Obtain genotype and phenotype data, and input the genotype and phenotype data into a trained GDASA gene expression prediction model for gene expression prediction
S120 Perform association analysis on gene expression and disease phenotype, obtain significant association information between the gene expression and the disease phenotype by means of univariate statistical testing, and according to a set threshold, screen out important gene segments significantly associated with the disease phenotype
S130 Calculate contribution values of gene loci in the important gene segments on model gene expression prediction by using an interpretability method of deep learning
S140 Calculate a contribution threshold by means of a related algorithm, and screen out the gene loci with the contribution values higher than the contribution threshold as gene loci significantly associated with the disease

(57) Abstract: Provided are a gene and disease association analysis method and apparatus, a computer device and a storage medium. The method comprises: constructing a deep learning-based GDASA gene expression prediction model; inputting genotype and disease phenotype data into a trained GDASA gene expression prediction model for gene expression prediction; performing association analysis on a gene expression prediction result and the disease phenotype data, obtaining significant association information between the gene expression prediction result and the disease phenotype data by means of univariate statistical testing, and according to a set threshold, screening out important gene segments significantly associated with the disease phenotype data; and calculating contribution values of gene loci in the important gene segments by using an interpretability method of deep learning, and screening out the gene loci with the contribution values higher than a set contribution threshold as gene loci significantly associated with the disease phenotype data. According to the method, the gene expression prediction precision is higher, and the accuracy of model interpretable analysis is improved.

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, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, 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, ME, 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))。

(57) 摘要: 提供一种基因与疾病关联分析方法、装置、计算机设备以及存储介质。所述方法包括: 构建基于深度学习的GDASA基因表达预测模型; 将基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测; 对基因表达预测结果与疾病表型数据进行关联分析, 通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息, 并根据设定阈值筛选出与疾病表型数据显著关联的重要基因片段; 使用深度学习的可解释性方法计算重要基因片段中各个基因位点的贡献程度值, 并筛选出贡献程度值高于设定贡献度阈值的基因位点作为与疾病表型数据显著关联的基因位点。所述方法使基因表达预测精度更高, 并提高了模型可解释分析的精确度。

基因与疾病关联分析方法、装置、计算机设备及存储介质

技术领域

本申请属于基因数据分析技术领域，特别涉及一种基因与疾病关联分析方法、装置、计算机设备以及存储介质。

背景技术

GWAS (Genome-wide association study, 全基因组关联研究) 是指在人类全基因组范围内找出存在的序列变异, 即SNP(single nucleotide polymorphism, 单核苷酸多态性), 从中筛选出与疾病相关的SNPs。随着高通量基因分型成本的快速下降, GWAS已被广泛用于寻找导致个体罹患复杂疾病或性状的遗传风险因素。然而, 由于遗传变异的多基因效应以及较小的效应大小, GWAS显示出有限的统计能力。此外, 遗传变异与这些性状之间的潜在分子机制尚不清楚。而遗传变异影响性状的一种方式是通过调节基因表达, 表达定量性状基因座 (eQTL) 研究表明了基因表达调控的重要性, 基因表达和性状之间的这种关系可以通过测量同一研究对象的遗传变异和转录组数据来研究。当转录组数据在GWAS样本中不可用时, TWAS (transcriptome-wide association, 转录组广泛关联研究) 为复杂的疾病机制提供了新的见解, 并作为执行基于基因的关联分析的替代方法而受到欢迎。

目前, TWAS的分析方法包括PrediXcan、MultiXcan以及UTMOST等, 其中PrediXcan是一种两阶段式方法, 它的预测效果可归因于基因表达预测模型的准确性以及基因表达和表型之间的关联。然而, 目前的方法并没有充分利用转录组资源的多组织性质和具有调控要素的综合图谱, 忽略了组织共享遗传调控的存在。MultiXcan整合了多组织研究中的信息, 通过回归预测跨组织的预测表达数据的主成分来提高关联分析的统计能力, 但是, MultiXcan是一种多组织关联分析方法, 其目的不是改善每个组织中基因表达的预测, 且每个主成分的效应大小和方向不容易解释。UTMOST是一种跨组织TWAS方法, 通过使用组惩罚项来做变量选择以提高预测性能。但是, 它没有利用组织之间的相似性。

日前所述, 现有的TWAS研究方法在训练基因表达预测模型时大多选用的是传统的线性回归模型, 预测精度不高, 同时在之后基因调控网络的解释中也很难做到精确。

发明内容

本申请提供了一种基因与疾病关联分析方法、装置、计算机设备以及存储介质，旨在至少在在一定程度上解决现有技术中的上述技术问题之一。

为了解决上述问题，本申请提供了如下技术方案：

一种基因与疾病关联分析方法，包括：

构建基于深度学习的GDASA基因表达预测模型；

获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；

使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

本申请实施例采取的技术方案还包括：所述基于深度学习的GDASA基因表达预测模型使用GTEx数据库第8版中的基因型及匹配基因表达数据进行训练得到。

本申请实施例采取的技术方案还包括：所述GDASA基因表达预测模型包括特征控制门、线性和非线性模块、先验知识融合模块和多任务模块，所述特征控制门用于过滤噪声和不显著的基因位点，所述线性和非线性模块用于集成基因位点之间的线性关联和非线性交互，所述先验知识融合模块用于融合先验知识，所述多任务模块用于利用主成分分析算法构建联合组织二级任务标签，对基因表达预测任务施加影响。

本申请实施例采取的技术方案还包括：所述获取基因型以及疾病表型数据具体为：

根据疾病类型在GTEx.v8数据库中挑选设定数量的特定组织，并获取每个特定组织的全转录组的基因型以及疾病表型数据。

本申请实施例采取的技术方案还包括：所述对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段具体为：

使用单变量最小二乘回归算法对特定组织中基因表达和疾病表型数据的相关性进行建模为：

$$D = \delta_i + \hat{Y}_i \alpha_i + \varepsilon_i$$

上述公式中， D 是疾病表型数据， δ_i 是截距， $\hat{Y}_i$ 是特定组织的第 i 个基因表达， α_i 是OLS回归模型的系数， ε_i 是残差；构造用于表示表达性状效应大小的 t 统计值：

$$t_i = \frac{\hat{\alpha}_i - \alpha_i}{SE(\hat{\alpha}_i)}$$

上述公式中， t_i 表示第 i 个表达性状效应大小的 t 统计值， SE 表示样本标准差，根据 t 值计算显著性关联的 P 值为：

$$p_i \equiv P(|T| > |t_i|)$$

上述公式中， p_i 表示表达性状效应大小的第 i 个 P 值，将 P 值作为评价指标筛选出与疾病表型显著关联的重要基因片段。

本申请实施例采取的技术方案还包括：所述使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值具体为：

针对每个重要基因片段，使用SHAP算法计算该重要基因片段的SHAP Value；

对所有计算出来的SHAP Value值，取其绝对值表示对应重要基因片段的贡献程度值；

针对每一个重要基因片段的SHAP Value矩阵的样本维度做平均化处理，得到该重要基因片段上每个基因位点对应的贡献程度值。

本申请实施例采取的技术方案还包括：所述筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点具体为：

针对设定数量的特定组织，使用统计学中的bootstrap抽样方法在每个特定组织的所有基因片段中随机有放回地抽取设定数量的基因片段，得到 N 个基因片段；

计算 N 个基因片段中所有基因位点的SHAP Value分布，取所述SHAP Value分布的置信区间95%分位点的值作为贡献度阈值；

筛选出所述贡献程度值高于贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

本申请实施例采取的另一技术方案为：一种基因与疾病关联分析装置，包括：

模型构建模块：用于构建基于深度学习的GDASA基因表达预测模型；

基因表达预测模块：用于获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

关联分析模块：用于对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据

之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；

基因位点筛选模块：用于使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

本申请实施例采取的又一技术方案为：一种计算机设备，所述计算机设备包括处理器、与所述处理器耦接的存储器，其中，

所述存储器存储有用于实现所述基因与疾病关联分析方法的程序指令；

所述处理器用于执行所述存储器存储的所述程序指令以控制基因与疾病关联分析方法。

本申请实施例采取的又一技术方案为：一种存储介质，存储有处理器可运行的程序指令，所述程序指令用于执行所述基因与疾病关联分析方法。

相对于现有技术，本申请实施例产生的有益效果在于：本申请实施例的基因与疾病关联分析方法、装置、计算机设备以及存储介质通过构建基于深度学习的GDASA基因表达预测模型，利用GDASA基因表达预测模型进行基因表达预测，对基因表达与疾病表型进行关联分析，通过单变量统计检验得到基因表达与疾病表型之间的显著关联信息，并根据设定阈值筛选出与疾病表型显著关联的重要基因片段，使用深度学习的可解释性方法计算出重要基因片段中各个基因位点对模型基因表达预测的贡献程度值，并筛选出贡献程度值高于设定贡献度阈值的基因位点作为与疾病显著关联的基因位点。本申请实施例的GDASA基因表达预测模型基于深度学习框架，使基因表达的预测精度更高，并大大提高了模型可解释分析的精确度。

附图说明

图1是本申请第一实施例的基因与疾病关联分析方法的流程图；

图2为本申请实施例的GDASA框架示意图；

图3是本申请第二实施例的基因与疾病关联分析方法的流程图；

图4为本申请实施例针对基因表达与疾病表型的关联分析结果进行的TWAS分析及绘制的曼哈顿图；

图5为本申请实施例的基因与疾病关联分析装置结构示意图；

图6为本申请实施例的计算机设备结构示意图；

图7为本申请实施例的存储介质的结构示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅是本申请的一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请中的术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”、“第三”的特征可以明示或者隐含地包括至少一个该特征。本申请的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。本申请实施例中所有方向性指示（诸如上、下、左、右、前、后……）仅用于解释在某一特定姿态（如附图所示）下各部件之间的相对位置关系、运动情况等，如果该特定姿态发生改变时，则该方向性指示也相应地随之改变。此外，术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或计算机设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或计算机设备固有的其它步骤或单元。

在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

请参阅图1，是本申请第一实施例的基因与疾病关联分析方法的流程图。本申请第二实施例的基因与疾病关联分析方法包括以下步骤：

S100：构建基于深度学习的GDASA基因表达预测模型；

本步骤中，本申请实施例基于基因表达预测模型设计了一套检验基因-疾病表型关联关系的GDASA(Gene and disease association study with AI, 基于人工智能的基因与疾病相关性研究)框架。具体如图2所示，为本申请实施例的GDASA框架示意图。所述GDASA框架在现有基因表达预测模型的基础上构建了一个基于深度学习的基因序列预测基因表达量的GDASA基因表达预测模型，GDASA基因表达预测模型通过使用GTEx数据库（第8版）中49个组织的基因型和匹配基因表达数据进行训练得到，通过基因SNP序列预测基因表达，再对基因表达与疾病表型进行关联分析，得到与疾病表型显著关联的重要基因片段，并对重要基因片段做模型可解释分析，最终得到与疾病表型数据显著关联的基因位点。本申请实施例中，为了整合多组织信息、先验知识和基因组位点的非线性相互作用，开发了一个带注释的多任务深度学习架构

GDASA, 突破了线性可加预测模型的局限性, 提升基因表达的预测效果。

进一步地, GDASA基因表达预测模型包括特征控制门 (FCG)、线性和非线性模块、先验知识融合模块和多任务模块四个部分, 其中, FCG用于过滤噪声和不显著的基因位点, 提高基因表达的预测性能。基于神经网络的因子分解机 (NFM) 思想, 本申请实施例通过引入线性和非线性模块的组合架构, 可以集成SNP位置之间的线性关联和非线性交互。先验知识融合模块充分考虑先验知识并进行融合, 多任务模块利用主成分分析 (PCA) 算法构建联合组织二级任务标签, 期望对基因表达预测任务施加积极影响。

S110: 获取基因型以及疾病表型数据, 将基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测;

本步骤中, 本申请实施例使用GTEx.v8数据库中的基因型以及疾病表型数据进行基因表达预测, 根据疾病类型在GTEx.v8数据库中挑选一定数量的特定组织, 并获取每个特定组织的全转录组的基因型以及疾病表型数据。

S120: 对基因表达预测结果与疾病表型数据进行关联分析, 通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联, 并根据设定阈值筛选出与疾病表型数据显著关联的重要基因片段;

本步骤中, 为了提取与疾病表型数据显著关联的重要基因片段, 需要预先计算基因表达预测结果与疾病表型数据之间的显著性关联信息, 显著性关联信息是指基因表达量与疾病表型数据单变量检验所得到的P值、AIC值、BIC值等各种统计量。具体的, 以P值为例, 基因表达与疾病表型数据之间的显著性关联信息计算方式为: 使用单变量最小二乘回归 (OLS) 算法对某些特定组织中基因表达和疾病表型数据的相关性进行建模为:

$$D = \delta_i + \hat{Y}_i \alpha_i + \varepsilon_i \quad (1)$$

公式 (1) 中, D 是疾病表型数据, δ_i 是截距, $\hat{Y}_i$ 是特定组织的第 i 个基因表达, α_i 是OLS回归模型的系数, ε_i 是残差。因此, 可以构造用于表示表达性状效应大小的 t 统计值:

$$t_i = \frac{\hat{\alpha}_i - \alpha_i}{SE(\hat{\alpha}_i)} \quad (2)$$

公式 (2) 中, t_i 表示第 i 个表达性状效应大小的 t 统计值, SE 表示样本标准差。因此, 根据 t 值计算显著性关联的P值为:

$$p_i \equiv P(|T| > |t_i|) \quad (3)$$

公式(3)中, p_i 表示表达性状效应大小的第*i*个P值, 将P值作为评价指标, 使用P值<0.05作为设定阈值筛选出与疾病表型显著关联的重要基因片段。

S130: 使用深度学习的可解释性方法计算出重要基因片段中各个基因位点对模型基因表达预测的贡献程度值;

本步骤中, 为了获得重要基因片段中的基因位点, 本申请实施例使用SHAP (Shapley Additive explanations, 即沙普利加和解释, 一个有效的模型可解释方法) 来提取基因位点对模型基因表达预测的贡献程度值。SHAP属于训练后的模型解释方法, 其基本思想是计算所有特征序列组合中每个特征的不同边际贡献的平均值, 作为代表每个贡献程度值的SHAP值。

进一步地, SHAP算法具体为: 针对每个重要基因片段, 用SHAP算法计算其SHAP Value来表示其对模型基因表达预测的贡献程度值, SHAP算法对于每个重要基因片段都会计算出其SHAP Value值, SHAP Value值有大于0或小于0的情况, SHAP Value大于0时表示该基因位点对应样本下对基因表达预测的贡献为正面贡献, SHAP Value小于0则表示该基因位点对应样本下对基因表达预测的贡献为负面贡献, 无论是正面贡献还是负面贡献, 都表示该基因位点对基因表达预测具有调控作用, 因此对所有计算出来的SHAP Value值, 取其绝对值来表示每个基因位点的贡献程度值, 同时针对每一个重要基因片段, 对计算出的SHAP Value矩阵的样本维度做平均化处理, 以得到该重要基因片段上每个基因位点对应的贡献程度值, 根据每个基因位点对应的贡献程度值提取出重要基因片段上具有高贡献程度的基因位点。

S140: 通过相关算法计算出贡献度阈值, 并筛选出贡献程度值高于贡献度阈值的基因位点作为与疾病表型数据显著关联的基因位点;

本步骤中, 由于每个基因位点都有一个贡献度, 为了筛选出对模型表达预测贡献度较高的基因位点, 需要设定一个贡献度阈值对基因位点进行筛选。具体的, 本申请实施例中的贡献度阈值计算方式为: 使用统计学中的bootstrap抽样方法在所有基因片段上进行有放回抽样, 针对设定数量的特定组织, 在每个组织的所有基因片段中随机有放回地抽取一定数量的基因片段, 得到N个基因片段, 然后计算N个基因片段中所有基因位点的SHAP Value分布, 取该SHAP Value分布的置信区间95%分位点的值作为贡献度阈值, 根据贡献度阈值筛选出对应组织的重要基因片段中与疾病表型显著关联的基因位点。本申请实施例中, 最终计算出的贡献度阈值为SHAPValue>0.05642。

基于上述, 本申请第一实施例的基因与疾病关联分析方法通过构建基于深度学习的GDASA基因表达预测模型, 利用GDASA基因表达预测模型进行基因表达预测, 对基因表达与疾病表型进行关联分析, 通过

单变量统计检验得到基因表达与疾病表型之间的显著关联信息，并根据设定阈值筛选出与疾病表型显著关联的重要基因片段，使用深度学习的可解释性方法计算出重要基因片段中各个基因位点对模型基因表达预测的贡献程度值，并筛选出贡献程度值高于设定贡献度阈值的基因位点作为与疾病显著关联的基因位点。本申请实施例的GDASA基因表达预测模型基于深度学习框架，使基因表达的预测精度更高，并大大提高了模型可解释分析的精确度，对疾病的靶点预测以及药物发现具有重要意义。

请参阅图3，是本申请第二实施例的基因与疾病关联分析方法的流程图。在本申请第二实施例中，将所述基因与疾病关联分析方法应用于抑郁症疾病表型的靶标分析中，在13个脑部组织中检测出了与抑郁症有显著关联的539个重要基因片段以及对应的基因位点，为抑郁症疾病的靶点发现与药物研发提供依据。具体的，本申请第二实施例的基因与疾病关联分析方法包括以下步骤：

S200：获取ROSMAP的原始数据，并对原始数据进行过滤和质量控制处理，得到用于进行基因表达预测的基因型以及表型数据；

本步骤中，所述疾病表型为与抑郁症相关的疾病表型，本申请实施例使用GTEx.v8的基因型与表达量数据进行GDASA基因表达预测模型的表达预测，针对抑郁症疾病在GTEx.v8数据集中挑选了包括大脑杏仁核组织、大脑下丘脑组织、大脑海马体组织等13个脑部组织的全转录组基因表达量数据。

S210：将基因型以及表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

S220：对基因表达与疾病表型进行关联分析，通过单变量统计检验得到基因表达与疾病表型之间的显著性关联信息，并根据设定阈值筛选出与疾病表型显著关联的重要基因片段；

本步骤中，针对基因表达与疾病表型的关联分析结果进行了TWAS分析并绘制曼哈顿图，具体如图4所示。显著性关联信息是指基因表达量与疾病表型单变量检验所得到的P值、AIC值、BIC值等各种统计量。本申请实施例选择P值作为评价指标，使用P值<0.05作为设定阈值筛选出与疾病表型显著关联的重要基因片段，同时对单变量统计检验得到的P值与设定阈值进行函数变换，变换公式如下：

$$P_N = -\text{Log}(P_O) \quad (4)$$

公式(4)中， P_O 代表变换前的P值， P_N 代表变换后的P值。

通过阈值筛选，最终在13个脑部组织中获得了总共539个与抑郁症疾病显著关联的重要基因片段，每个组织中与疾病显著关联的重要基因片段个数如表1所示：

表1 每个脑部组织与疾病表型显著关联的重要基因片段个数

组织名	显著关联的重要基因片段个数
Brain_Amygdala	22
Brain_Anterior_cingulate_cortex_BA24	30
Brain_Caudate_basal_ganglia	48
Brain_Cerebellar_Hemisphere	57
Brain_Cerebellum	99
Brain_Cortex	67
Brain_Frontal_Cortex_BA9	43
Brain_Hippocampus	30
Brain_Hypothalamus	27
Brain_Nucleus_accumbens_basal_ganglia	47
Brain_Putamen_basal_ganglia	37
Brain_Spinal_cord_cervical_c.1	20
Brain_Substantia_nigra	12

如表1所示，大部分组织中与疾病表型显著关联的重要基因片段个数都超过30个，只有大脑黑质组织、大脑脊髓颈组织、大脑杏仁核组织以及大脑下丘脑组织由于对应组织样本的数量限制等原因导致iGene个数较少，并导致能够被检测出来的基因更少，其中大脑黑质组织显著的基因个数最少仅仅只有12个基因片段，而大脑黑质组织以及大脑杏仁核组织在GETx数据集上由GDASA表达预测模型检测出的iGene个数也是比较少的只分别有2575个和2581个iGene。

S230: 使用深度学习的可解释性方法计算出重要基因片段中各个基因位点的贡献程度值，并筛选出贡献程度值高于设定贡献度阈值的基因位点作为与疾病表型显著关联的基因位点；

本步骤中，在进行了基因片段与疾病表型的关联关系分析后，需要研究与疾病表型有显著关联的重要基因片段中有多少致病的基因位点。本申请实施例使用SHAP Value对训练好的GDASA基因表达预测模型进行可解释性分析，得到对模型基因表达预测贡献度较高的基因位点。具体的，针对TWAS分析中得到的每个组织中与疾病表型有显著关联的重要基因片段，在GTEx.v8数据集上做GDASA基因表达预测模型的训练，并且用SHAP方法计算其SHAP Value来表示其对模型基因表达预测的贡献程度，由于SHAP算法对于每个样本都会计算出其SHAP Value值，SHAP Value值有大于0或小于0的情况，SHAP Value大于0时表示该基因位点在对应样本下对基因表达预测的贡献为正面贡献，SHAP Value小于0则表示该基因位点在对应样本下对基因表达预测的贡献为负面贡献，无论是正面贡献还是负面贡献，都表示该基因位点对基因表达预测具有调控作用，因此对所有计算出来的SHAP Value值取绝对值来表示该基因位点的贡献值大小，同时针对每一个重要基因片段，

对计算出的SHAP Value矩阵的样本维度做平均化处理，以得到该重要基因片段上每个基因位点对应的贡献程度值，这样就可以提取出与疾病显著关联的重要基因片段上具有高贡献度的基因位点。具体的，使用贡献度阈值筛选出的13个脑部组织中所有与抑郁症疾病表型显著关联的基因位点个数如表2所示：

表213个脑部组织检测出的与抑郁症显著关联的基因位点个数

组织名	显著的基因位点个数
Brain_Amygdala	1348
Brain_Anterior_cingulate_cortex_BA24	845
Brain_Caudate_basal_ganglia	1129
Brain_Cerebellar_Hemisphere	1236
Brain_Cerebellum	1008
Brain_Cortex	1224
Brain_Frontal_Cortex_BA9	1157
Brain_Hippocampus	1046
Brain_Hypothalamus	1166
Brain_Nucleus_accumbens_basal_ganglia	803
Brain_Putamen_basal_ganglia	1061
Brain_Spinal_cord_cervical_c.1	1738
Brain_Substantia_nigra	1639

如表2所示，在13个脑部组织中总共提取出了15400个可能与抑郁症有显著关联的基因位点，其中大脑前扣带回皮层组织和大脑伏隔基底神经节脑核组织所检测出的显著基因位点个数最少分别只有845个和803个，其余组织所检测出的基因位点个数均超过1000个，最多的两个组织大脑脊髓颈组织和大脑脑黑质组织分别检测出了1738个和1639个显著的基因位点。

请参阅图5，为本申请实施例的基因与疾病关联分析装置结构示意图。本申请实施例的基因与疾病关联分析装置40包括：

模型构建模块41：用于构建基于深度学习的GDASA基因表达预测模型；

基因表达预测模块42：用于获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

关联分析模块43：用于对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；

基因位点筛选模块44：用于使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

请参阅图6，为本申请实施例的计算机设备结构示意图。该计算机设备50包括：

存储有可执行程序指令的存储器51；

与存储器51连接的处理器52；

处理器52用于调用存储器51中存储的可执行程序指令并执行以下步骤：构建基于深度学习的GDASA基因表达预测模型；获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

其中，处理器52还可以称为CPU（Central Processing Unit，中央处理单元）。处理器52可能是一种集成电路芯片，具有信号的处理能力。处理器52还可以是通用处理器、数字信号处理器（DSP）、专用集成电路（ASIC）、现成可编程门阵列（FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

请参阅图7，图7为本申请实施例的存储介质的结构示意图。本申请实施例的存储介质存储有能够实现以下步骤的程序指令61：构建基于深度学习的GDASA基因表达预测模型；获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。其中，该程序指令61可以以软件产品的形式存储在上述存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者

网络计算机设备等)或处理器(processor)执行本申请各个实施方式方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM, Read-Only Memory)、随机存取存储器(RAM, Random Access Memory)、磁碟或者光盘等各种可以存储程序指令的介质,或者是计算机、服务器、手机、平板等终端计算机设备。其中,服务器可以是独立的服务器,也可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络(Content Delivery Network, CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

在本申请所提供的几个实施例中,应该理解到,所揭露的系统,装置和方法,可以通过其它的方式实现。例如,以上所描述的系统实施例仅仅是示意性的,例如,单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

另外,在本申请各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。以上仅为本申请的实施方式,并非因此限制本申请的专利范围,凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1. 一种基因与疾病关联分析方法，其特征在于，包括：

构建基于深度学习的GDASA基因表达预测模型；

获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；

使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

2. 根据权利要求1所述的基因与疾病关联分析方法，其特征在于，所述基于深度学习的GDASA基因表达预测模型使用GTEx数据库第8版中的基因型及匹配基因表达数据进行训练得到。

3. 根据权利要求2所述的基因与疾病关联分析方法，其特征在于，所述GDASA基因表达预测模型包括特征控制门、线性和非线性模块、先验知识融合模块和多任务模块，所述特征控制门用于过滤噪声和不显著的基因位点，所述线性和非线性模块用于集成基因位点之间的线性关联和非线性交互，所述先验知识融合模块用于融合先验知识，所述多任务模块用于利用主成分分析算法构建联合组织二级任务标签，对基因表达预测任务施加影响。

4. 根据权利要求3所述的基因与疾病关联分析方法，其特征在于，所述获取基因型以及疾病表型数据具体为：

根据疾病类型在GTEx.v8数据库中挑选设定数量的特定组织，并获取每个特定组织的全转录组的基因型以及疾病表型数据。

5. 根据权利要求1至4任一项所述的基因与疾病关联分析方法，其特征在于，所述对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段具体为：

使用单变量最小二乘回归算法对特定组织中基因表达和疾病表型数据的相关性进行建模为：

$$D = \delta_i + \hat{Y}_i \alpha_i + \varepsilon_i$$

上述公式中， D 是疾病表型数据， δ_i 是截距， $\hat{Y}_i$ 是特定组织的第 i 个基因表达， α_i 是OLS回归模型的系数， ε_i 是残差；构造用于表示表达性状效应大小的 t 统计值：

$$t_i = \frac{\hat{\alpha}_i - \alpha_i}{SE(\hat{\alpha}_i)}$$

上述公式中， t_i 表示第*i*个表达性状效应大小的*t*统计值，*SE*表示样本标准差，根据*t*值计算显著性关联的*P*值为：

$$p_i \equiv P(|T| > |t_i|)$$

上述公式中， p_i 表示表达性状效应大小的第*i*个*P*值，将*P*值作为评价指标筛选出与疾病表型显著关联的重要基因片段。

6. 根据权利要求5所述的基因与疾病关联分析方法，其特征在于，所述使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值具体为：

针对每个重要基因片段，使用SHAP算法计算该重要基因片段的SHAP Value；

对所有计算出来的SHAP Value值，取其绝对值表示对应重要基因片段的贡献程度值；

针对每一个重要基因片段的SHAP Value矩阵的样本维度做平均化处理，得到该重要基因片段上每个基因位点对应的贡献程度值。

7. 根据权利要求6所述的基因与疾病关联分析方法，其特征在于，所述筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点具体为：

针对设定数量的特定组织，使用统计学中的bootstrap抽样方法在每个特定组织的所有基因片段中随机有放回地抽取设定数量的基因片段，得到*N*个基因片段；

计算*N*个基因片段中所有基因位点的SHAP Value分布，取所述SHAP Value分布的置信区间95%分位点的值作为贡献度阈值；

筛选出所述贡献程度值高于贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

8. 一种基因与疾病关联分析装置，其特征在于，包括：

模型构建模块：用于构建基于深度学习的GDASA基因表达预测模型；

基因表达预测模块：用于获取基因型以及疾病表型数据，将所述基因型以及疾病表型数据输入训练好的GDASA基因表达预测模型进行基因表达预测；

关联分析模块：用于对所述基因表达预测结果与疾病表型数据进行关联分析，通过单变量统计检验得到基因表达预测结果与疾病表型数据之间的显著关联信息，并根据设定阈值筛选出与所述疾病表型数据显著关联的重要基因片段；

基因位点筛选模块：用于使用深度学习的可解释性方法计算所述重要基因片段中各个基因位点对所述GDASA基因表达预测模型进行基因表达预测的贡献程度值，并筛选出所述贡献程度值高于设定贡献度阈值的基因位点作为与所述疾病表型数据显著关联的基因位点。

9. 一种计算机设备，其特征在于，所述计算机设备包括处理器、与所述处理器耦接的存储器，其中，

所述存储器存储有用于实现权利要求1-7任一项所述的基因与疾病关联分析方法的程序指令；

所述处理器用于执行所述存储器存储的所述程序指令以控制基因与疾病关联分析方法。

10. 一种存储介质，其特征在于，存储有处理器可运行的程序指令，所述程序指令用于执行权利要求1至7任一项所述基因与疾病关联分析方法。

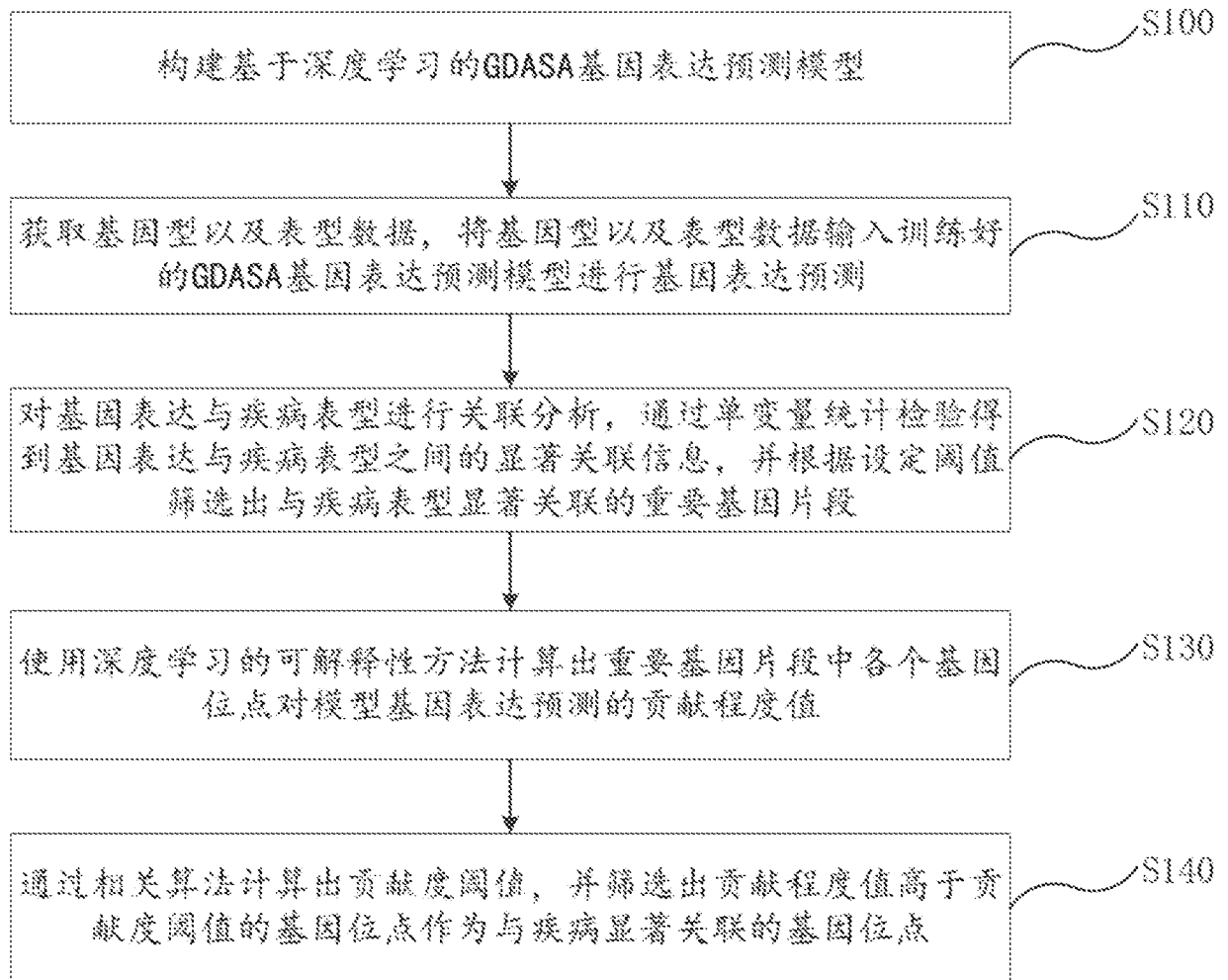


图1

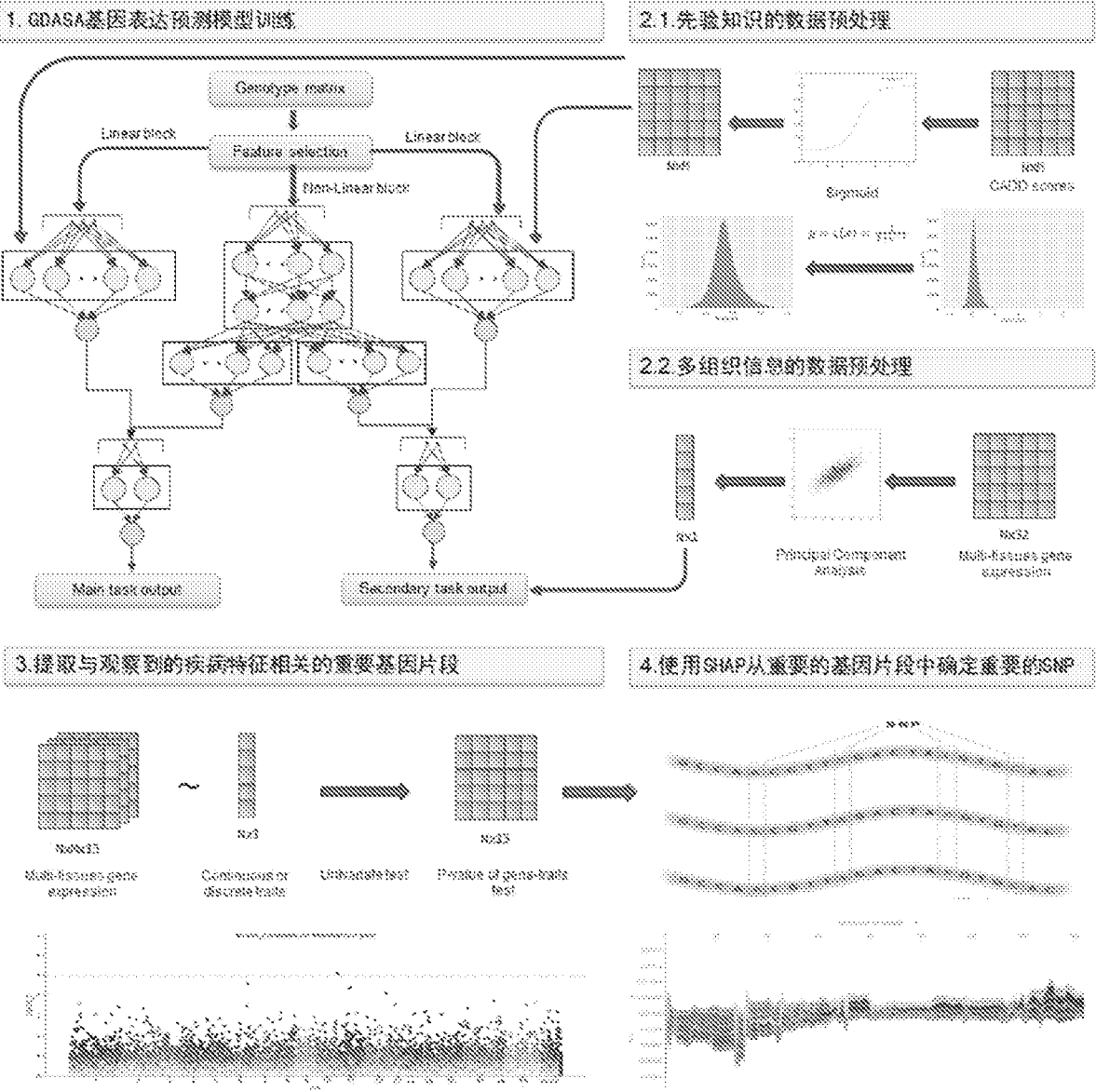


图2

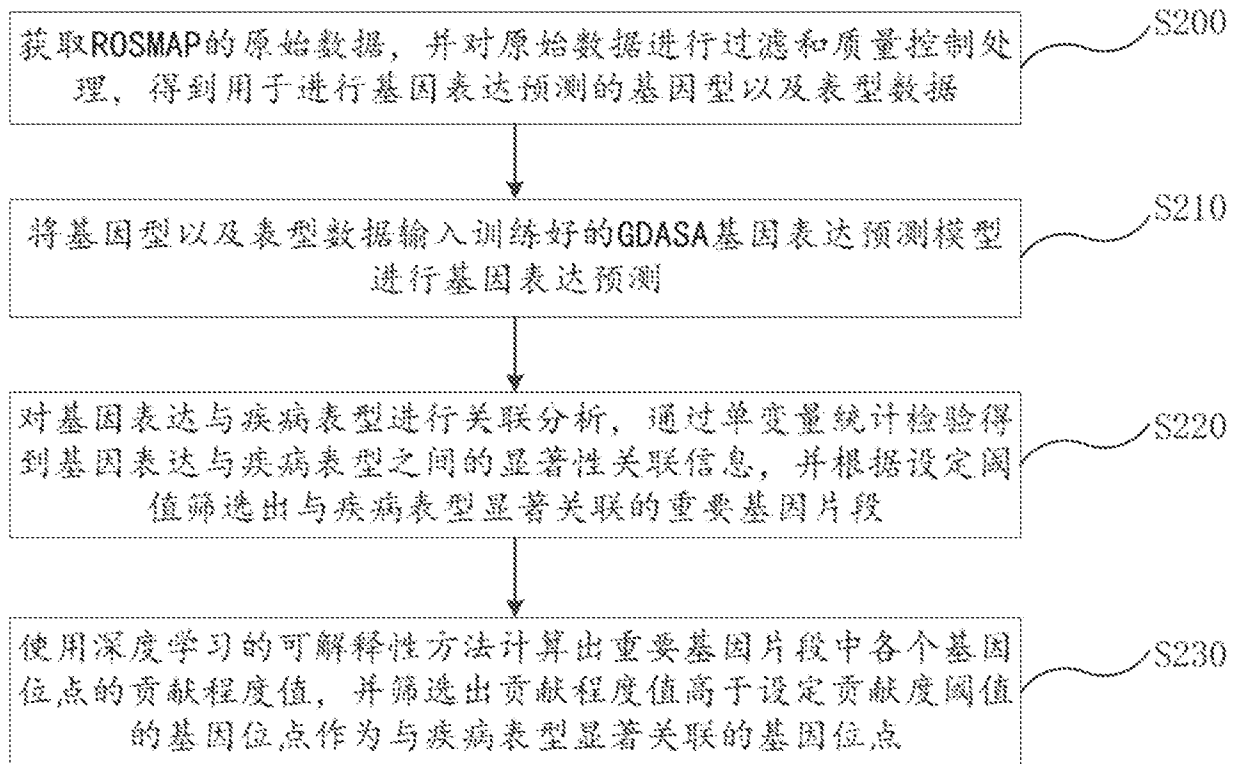


图3

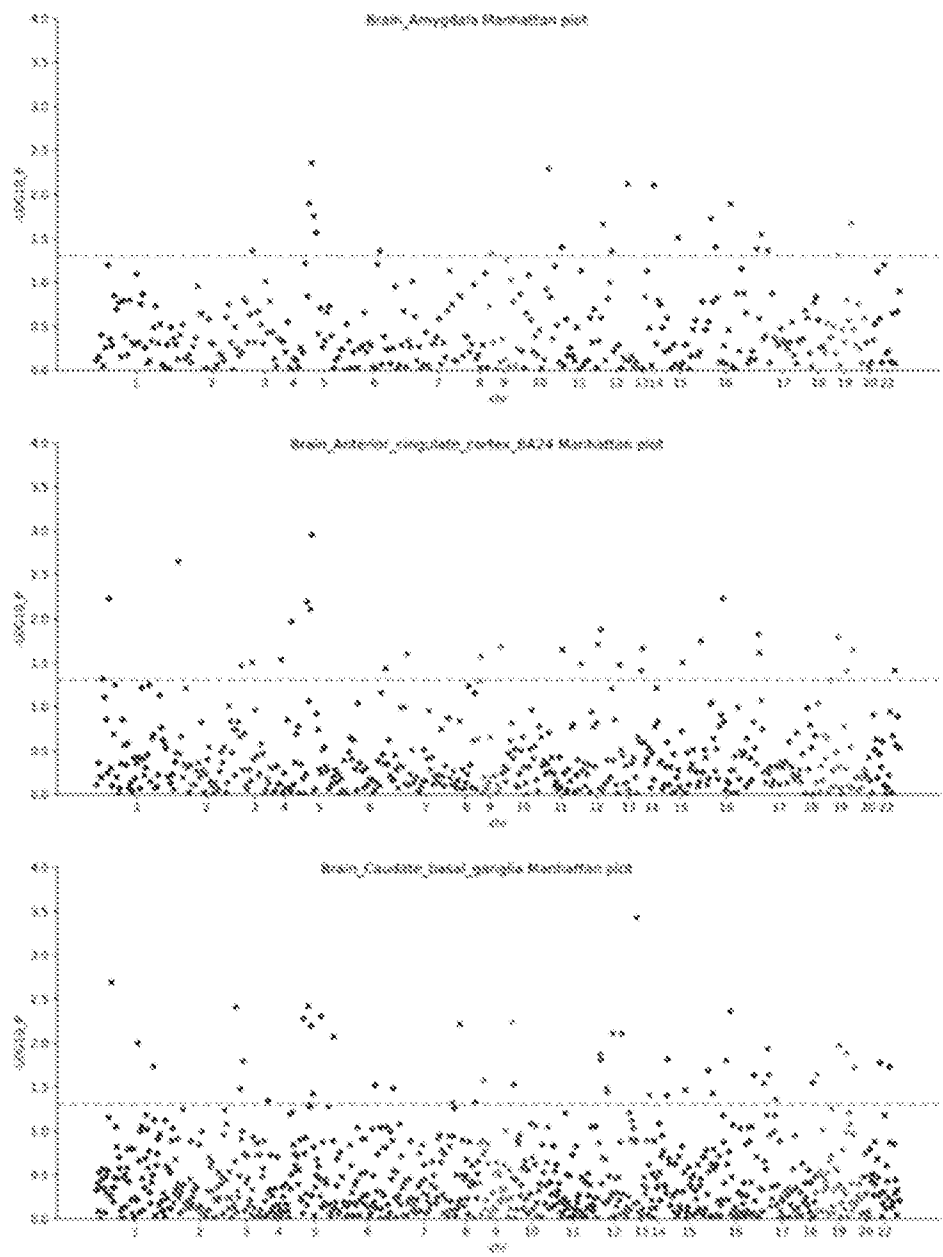


图4

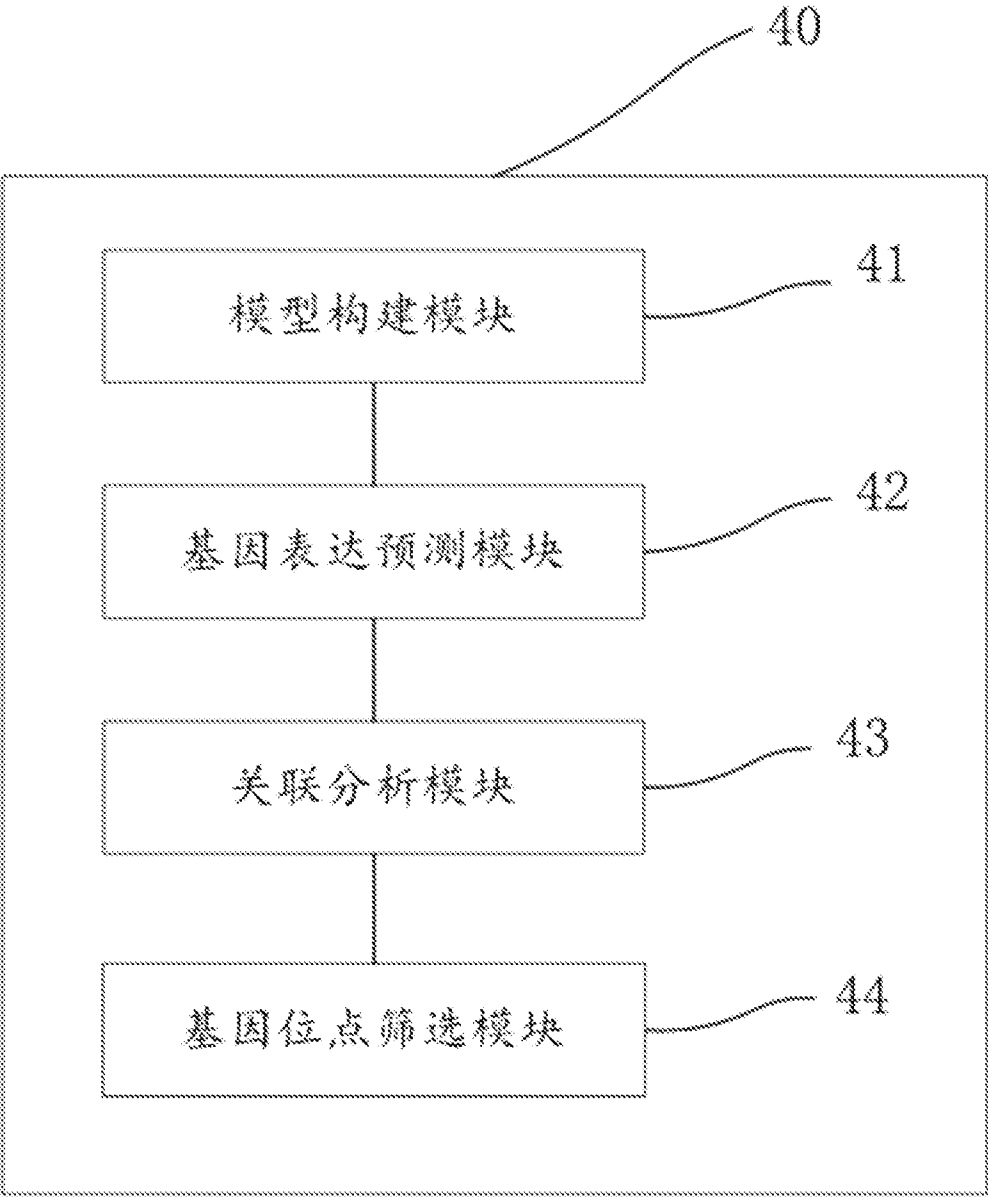


图5

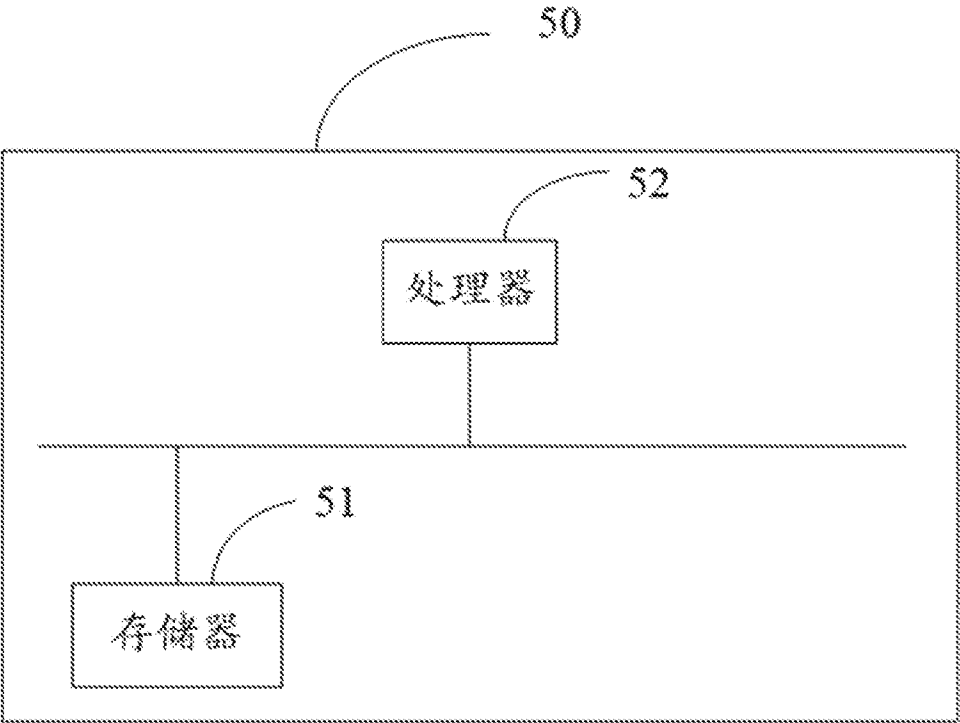


图6

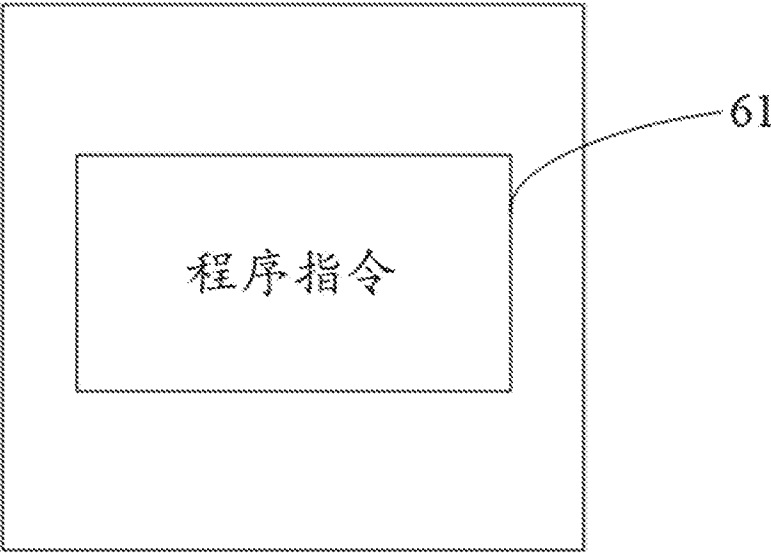


图7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/137633

A. CLASSIFICATION OF SUBJECT MATTER

G16B 40/00(2019.01)i; C12Q 1/6883(2018.01)i; G16H 50/70(2018.01)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,C12Q,G16H

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)

CNXTX, WPABSC, WPABS, ENTXTC, ENTXT, 超星读秀学术, DUXIU ACADEMIC, CNKI, 万方, WANFANG, ISI web of knowledge, Elsevier Science Direct, Springerlink, pubmed, 百度, BAIDU, 百度学术, BAIDU SCHOLAR, patentics: 基因, 疾病, 基因型, 表型, 关联, 深度学习, 机器学习, 人工智能, 模型, 预测, 单变量, 筛选, 可解释性方法, 贡献度, GTE_x, GDASA, gene, disease, genotype, screening, AI, prediction, association, machine learning, model, deep learning, SHAP

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SHU, Juan et al. "Disease Gene Prediction with Privileged Information and Heteroscedastic Dropout" <i>Bioinformatics</i> , Vol. 37, 31 December 2021 (2021-12-31), pages i410-i417	1-10
A	CN 115698335 A (INSITRO, INC.) 03 February 2023 (2023-02-03) entire document	1-10
A	WO 2023033329 A1 (BASGENBIO CO., LTD.) 09 March 2023 (2023-03-09) entire document	1-10
A	CN 113096816 A (XI'AN JIAOTONG UNIVERSITY) 09 July 2021 (2021-07-09) entire document	1-10
A	CN 114927163 A (ZHEJIANG UNIVERSITY) 19 August 2022 (2022-08-19) entire document	1-10
A	US 2021158967 A1 (NATIONAL CENTRAL UNIVERSITY) 27 May 2021 (2021-05-27) entire document	1-10



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
 "D" document cited by the applicant in the international application
 "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

20 February 2024

Date of mailing of the international search report

06 March 2024

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
China No. 6, Xitucheng Road, Jimenqiao, Haidian District,
Beijing 100088

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/137633

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	115698335	A	03 February 2023	WO	2021237117	A1	25 November 2021
WO	2023033329	A1	09 March 2023	KR	102351306	B1	14 January 2022
				KR	102351306	B9	16 June 2022
CN	113096816	A	09 July 2021	None			
CN	114927163	A	19 August 2022	None			
US	2021158967	A1	27 May 2021	TWI	709904	B	11 November 2020
				TW	202121223	A	01 June 2021

A. 主题的分类

G16B 40/00(2019.01)i; C12Q 1/6883(2018.01)i; G16H 50/70(2018.01)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,C12Q,G16H

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))

CNTEXT, WPABSC, WPABS, ENTXTC, ENTXT, 超星读秀学术, CNKI, 万方, ISI web of knowledge, Elsevier Science Direct, Springerlink, pubmed, 百度, 百度学术, patents: 基因, 疾病, 基因型, 表型, 关联, 深度学习, 机器学习, 人工智能, 模型, 预测, 单变量, 筛选, 可解释性方法, 贡献度, GTE_x, GDASA, gene, disease, genotype, screening, AI, prediction, association, machine learning, model, deep learning, SHAP

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	SHU, J.等. "Disease gene prediction with privileged information and heteroscedastic dropout" Bioinformatics, 第37卷, 2021年12月31日 (2021 - 12 - 31), 第i410-i417页	1-10
A	CN 115698335 A (因斯特罗公司) 2023年2月3日 (2023 - 02 - 03) 全文	1-10
A	WO 2023033329 A1 (BASGENBIO CO., LTD.) 2023年3月9日 (2023 - 03 - 09) 全文	1-10
A	CN 113096816 A (西安交通大学) 2021年7月9日 (2021 - 07 - 09) 全文	1-10
A	CN 114927163 A (浙江大学) 2022年8月19日 (2022 - 08 - 19) 全文	1-10
A	US 2021158967 A1 (NAT. CENTRAL UNIV.) 2021年5月27日 (2021 - 05 - 27) 全文	1-10

☐ 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
“D” 申请人在国际申请中引证的文件
“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

2024年2月20日

国际检索报告邮寄日期

2024年3月6日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

王慧

电话号码 (+86) 010-53962103

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/137633

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	115698335	A	2023年2月3日	WO	2021237117	A1	2021年11月25日
WO	2023033329	A1	2023年3月9日	KR	102351306	B1	2022年1月14日
				KR	102351306	B9	2022年6月16日
CN	113096816	A	2021年7月9日	无			
CN	114927163	A	2022年8月19日	无			
US	2021158967	A1	2021年5月27日	TWI	709904	B	2020年11月11日
				TW	202121223	A	2021年6月1日