



(10) 授权公告号 CN 112470229 B

(45) 授权公告日 2024. 11. 15

(21) 申请号 201980028750.0

(22) 申请日 2019.02.26

(65) 同一申请的已公布的文献号  
申请公布号 CN 112470229 A

(43) 申请公布日 2021.03.09

(30) 优先权数据  
1803202.9 2018.02.27 GB

(85) PCT国际申请进入国家阶段日  
2020.10.27

(86) PCT国际申请的申请数据  
PCT/GB2019/050525 2019.02.26

(87) PCT国际申请的公布数据  
W02019/166792 EN 2019.09.06

(73) 专利权人 基因组学公开有限公司  
地址 英国牛津

(72) 发明人 克里斯托弗·查尔斯·艾伦·斯宾塞  
杰拉德·安东·伦特  
彼得·詹姆斯·唐纳利  
文森特·雅恩·玛丽·普雷格诺

(74) 专利代理机构 北京同立钧成知识产权代理有限公司 11205  
专利代理师 李璐 臧建明

(51) Int.Cl.  
G16B 40/20 (2006.01)  
G16B 20/40 (2006.01)

(56) 对比文件  
US 2014032122 A1, 2014.01.30  
SAEED MOHAMMAD ED ET AL..NOVEL  
LINKAGE DISEQUILIBRIUM CLUSTERING  
ALGORITHM IDENTIFIES NEW LUPUGS GENES ON  
META-ANALYSIS OF GWAS  
DATASETS.IMMUNOGENETICS.2017,第69卷(第5  
期),295-302.  
NIEUWBOER HAROLD ET AL..GWIS:GENOME-  
WIDE INFERRED STATISTICS FOR FUNCTIONS OF  
MULTIPLE PHENOTYPES.AMERICAN JOURNAL OF  
HUMAN GENETICS.2016,第99卷(第4期),917-  
927.

(续)

审查员 贾小丽

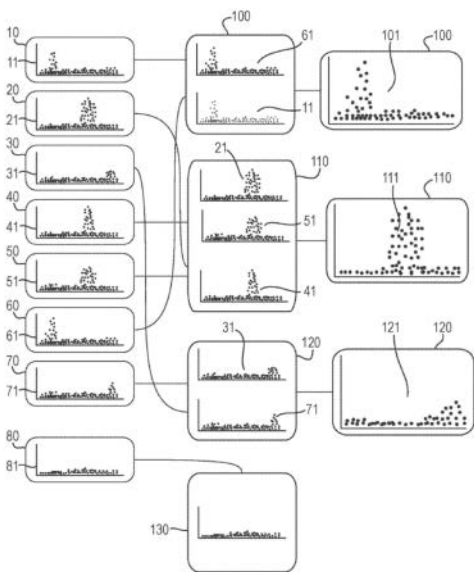
权利要求书4页 说明书17页 附图11页

(54) 发明名称

分析关于生物体的基因数据的计算机实现的方法

(57) 摘要

公开了用于分析关于生物体的基因数据的方法。在一种配置中,输入单元来自于提供关于基因变异(genetic variants)和表型之间的关联信息的研究。基于对输入单元共享基因变异的程度的评估,所述基因变异影响对应于每一个输入单元的表型的任何方面或所述表型的任何潜在生物学机制,将输入单元分配至多个集群中的一个,从而鉴定共享潜在生物学机制的表型。



CN 112470229 B

[接上页]

**(56) 对比文件**

HANA LANGO ALELEN ET AL..HUNDREDS OF VARIANTS CLUSTERED IN GENOMIC LOCI AND BIOLOGICAL PATHWAYS AFFECT HUMAN HEIGHT.NATURE.2010,第467卷(第7317期),832-838.

YUE LI ET AL..A probabilistic framework to dissect functional cell-type-specific regulatory elements and risk loci underlying the genetics of complex traits.<http://www.biorxiv.org/content/biorxiv/early/2017/10/24/059345.full.pdf>.2016,14、19、25-29.

Claudia Giambartolomei et al..A Bayesian Framework for Multiple Trait Colocalization from Summary Association Statics.<http://www.biorxiv.org/content/biorxiv/early/2018/02/13/155481.full.pdf>.2018,全文.

JOSEPH K PICKRELL ET AL..Detection and interpretation of shared genetic influences on 42 human traits.NATURE GENETICS.2016,第48卷(第7期),709-717.

KYLE KAI-HOW FARH ET AL..Genetic and epigenetic fine mapping of causal autoimmune disease variants.NATURE.2014,第518卷(第7539期),337-343.

1. 一种分析关于生物体的基因数据的计算机实现的方法,包括:

接收包括多个输入单元的输入数据,其中每个输入单元源自提供概括统计数据的基因关联研究,所述概括统计数据包括:沿着所述生物体的基因组的多个基因变异的每一个针对与所述输入单元相对应的表型的推断效应量和所述推断效应量的标准误差;

每个输入单元包括基于所述概括统计数据确定的特征组,所述特征组包括概率指标,所述概率指标量化了所述多个基因变异的每一个和与所述输入单元相对应的表型具有因果关系的证据,然后

对所述生物体的所述基因组选择一个或多个区域;然后

对于每个选定的区域,将每个所述输入单元分配至多个集群中的一个或多个,所述分配包括:

确定下列之间的相似度:i)所述输入单元的所述特征组,和ii)所述多个集群中的每一个的特征组,其中,每个所述集群的所述特征组是预先确定的,或者是通过组合已经分配到所述集群的输入单元的特征组来计算的;以及

a)以取决于所述相应相似度的概率,将所述输入单元分配到所述多个集群中的一个或多个现有集群;或

b)在所述多个集群中创立新集群,并以取决于所述输入单元的所述特征组和所述多个集群的所述现有集群的所述特征组的概率,将所述输入单元分配到所述新集群;

该方法还包括:

对与分配到相同集群的输入单元相对应的表型共享潜在生物学机制进行鉴定。

2. 根据权利要求1所述的方法,其中,所述每个输入单元到所述多个集群中的一个或多个的分配是基于采用贝叶斯模型对所述分配的评估。

3. 根据权利要求1所述的方法,其中,所述每个输入单元到所述多个集群中的一个或多个的分配是迭代进行的。

4. 根据权利要求1所述的方法,其中,通过马尔科夫链蒙特卡洛算法来探索可能的集群分配的空间,并且将所述每个输入单元分配到所述多个集群中的一个或多个是基于马尔科夫链所走路径的分析。

5. 根据权利要求4所述的方法,其中:

所述马尔科夫链蒙特卡洛算法的一个步骤包括将所述输入单元中的一个以取决于所述输入单元的所述特征组和所述集群的所述特征组之间的相似度的概率,分配到多个集群中的一个。

6. 根据权利要求1所述的方法,其中,所述多个集群包括空集群。

7. 根据权利要求6所述的方法,其中,所述空集群的特征是预先确定的。

8. 根据权利要求1所述的方法,其中,每个所述输入单元到所述多个集群中的一个或多个的分配还包括在集群的数量和/或大小上使用先验分布。

9. 根据权利要求8所述的方法,其中,所述集群的数量和/或大小上的先验分布遵循中国餐馆过程。

10. 根据权利要求1所述的方法,其中,确定所述输入单元的所述特征组包括:计算所述输入单元中的所述多个基因变异中的每一个的贝叶斯因子。

11. 根据权利要求10所述的方法,其中,计算所述集群的所述特征组包括计算分配至所

述集群的所述输入单元的贝叶斯因子的乘积。

12. 根据权利要求1所述的方法, 其中, 计算集群的所述特征组使用所述多个基因变异中每一个都是因果关系的概率的先验分布。

13. 根据权利要求12所述的方法, 其中, 所述概率的先验分布结合了与基因变异因果关系相关的功能变异的有关已有信息。

14. 根据权利要求1所述的方法, 其中, 计算集群的所述特征组还包括考虑提供用于导出输入单元的信息的研究之间的相关性或其他已知关系。

15. 根据权利要求1所述的方法, 其中, 每个所述输入单元到所述多个集群中的一个或多个的分配还包括在分配到相同集群的输入单元对上或更大集合上使用先验分布。

16. 根据权利要求1所述的方法, 还包括输出在集群成员的多个方面上的概率分布, 并且表征围绕集群分配的确定性。

17. 根据权利要求1所述的方法, 还包括迭代地重复将所述输入单元分配到所述多个集群中的一个或多个, 直到达到预定的收敛阈值或已经执行了预定的迭代次数。

18. 根据上述任一项权利要求所述的方法, 还包括, 对于所述多个集群中的一个或多个中的每一个, 从所述多个基因变异中鉴定可能与所述集群对应的一个或多个表型有因果关系的一个或多个重要基因变异。

19. 根据权利要求18所述的方法, 还包括:

计算一个或多个重要基因变异对目标表型的影响大小;

确定相对于一个或多个重要基因变异中的每一个的个体生物体的基因型;

根据一个或多个重要基因变异对所述目标表型的影响大小, 计算所述个体生物体的基因型对个体生物体的所述目标表型的贡献。

20. 根据权利要求19所述的方法, 还包括将所述个体生物体的基因型对所述目标表型的贡献与关于所述个体生物体的其他临床和/或基因数据相结合, 以改进所述个体生物的目标表型的确定。

21. 根据权利要求18所述的方法, 其中针对一个或多个集群中的每一个鉴定两个或更多个重要基因变异, 并且所述两个或更多个重要基因变异的鉴定取决于基因变异之间的关联矩阵的计算。

22. 根据权利要求18所述的方法, 其中针对一个或多个集群中的每一个鉴定两个或更多个重要基因变异, 并且对两个或更多个重要基因变异的鉴定包括:

计算所述集群的特征组量化与所述集群有因果关系的所述每个基因变异的概率;

根据所述集群的所述特征组确定所述集群的第一重要基因变异;

更新所述集群的特征组以考虑所述第一个重要基因变异的影响;

根据更新的特征组确定所述集群的第二重要基因变异。

23. 根据权利要求1所述的方法, 其中, 所述多个基因变异选自所述基因组的选定区域中的基因变异。

24. 根据权利要求23所述的方法, 其中, 所述选定区域包括预定数目的碱基对并且包括目的基因。

25. 根据权利要求23所述的方法, 其中通过最小化所述选定区域中的基因变异之间的相关性以及在所述基因组的一个或多个其他区域内的基因变异间的相关性来选择所述选

定区域。

26. 根据权利要求23中任一项所述的方法, 其中, 对所述基因组的多个选定区域中的每一个执行所述方法, 从而提供包含分配了输入单元的多个集群组, 每个集群组对应于多个选定区域中不同的一个选定区域。

27. 根据权利要求26所述的方法, 还包括: 通过与集群的第一子集内的输入单元相对应的表型与集群的一个或多个第二子集内的输入单元相对应的表型之间的相似的评估, 将从第一组的一个或多个集群的第一子集与一个或多个其他组的每一个的一个或多个集群的第二子集联系起来。

28. 根据权利要求26所述的方法, 还包括鉴定对应所述输入单元的表型组, 所述输入单元在与在频率高于预定频率阈值的多个组中被分配到相同集群。

29. 根据权利要求1所述的方法, 其中, 一个或多个所述输入单元还包括在所述基因组的所述选定区域内的一个或多个基因变异和对应于所述输入单元的表型间的关联的其他信息, 所述其他信息通过模型化所述一个或多个基因变异和对应于所述输入单元的所述表型之间的关联而获得。

30. 根据权利要求1所述的方法, 其中, 对应于所述输入单元的所述表型包括以下的一个或多个: 基因表达水平; 基因表达的调节; 表观遗传特征; 蛋白质或肽的丰度水平; 蛋白质或肽的功能和/或分子结构; 生物体的分子数量; 生化和代谢过程的特征; 包括形态和功能的细胞特性; 包括形态和功能的组织特性; 包括形态和功能的器官和器官系统特性; 对外部刺激或多个刺激的任何反应; 对暴露于某种物质或病原体的任何反应; 行为和生活方式特征; 生殖和生命历程的特征和功能; 疾病或状况的发生、发展轨迹和预后; 可测量的解剖学特征; 可测量的生理或功能特征; 以及可测量的心理或认知特征。

31. 根据权利要求1所述的方法, 其中, 对彼此直接或间接有因果相关的表型的鉴定包括对以下表型组合的一个或多个的鉴定:

i) 疾病或状况的发生以及生物体中基因表达水平, 蛋白质或肽表达水平, 生物分子数量中的一个或多个;

ii) 生物体的可测量特征以及所述生物体中基因表达水平, 蛋白质或肽的表达水平, 生物分子的数量中的一个或多个;

i) 所述生物体疾病或状况的发生以及可测量的特征;

ii) 所述生物体对物质的输入的反应, 以及所述生物体的疾病或状况的发生, 所述生物体的可测量特征, 所述生物体中生物分子的数量中的一个或多个;

iii) 所述生物体中基因的表达水平以及不同基因的表达水平, 蛋白质或肽的表达水平, 生物分子的数量中的一个或多个。

32. 根据权利要求1所述的方法, 其中, 所述方法, 基于对彼此直接或间接因果相关的表型的鉴定, 包括以下一项或多项:

(i) 确定所述生物体的临床支持策略;

(ii) 确定所述生物体发生疾病或状况的概率;

(iii) 在临床或临床前试验中对所述生物体进行分层; 或

(iv) 鉴定所述生物体中的生物途径。

33. 根据权利要求1所述的方法, 其中, 对应于所述输入单元的表型包括生物体对输入

一种或多种新的或现有药物的反应,并且所述方法,基于彼此直接或间接因果相关的表型的鉴定,包括以下一项或多项:

- (i) 确定一种或多种药物的作用机理;
- (ii) 确定所述生物体对一种或多种药物的临床和/或不良反应;
- (iii) 确定一种或多种药物的生物标志物;或
- (iv) 确定为达到特定的临床或安全结果应该或不应该将药物给予哪些患者子集。

34. 根据权利要求1所述的方法,其中对应于所述输入单元的表型包括生物体对两种或更多种新药或现有药物的输入的反应,并且所述方法,基于彼此直接或间接因果相关表型的鉴定,包括以下一项或多项:

- (i) 确定两种或多种药物之间的相互作用;或
- (ii) 确定两种或多种药物之间相互作用对生物体的影响。

35. 根据权利要求1所述的方法,其中,将所述方法应用于所述基因组的不同区域以鉴定生物学特性、特征或机制,获得是集群组的进一步分析,其在多于一个集群组间共享以鉴定以下任何一项:

- (i) 在达到特定疾病结果或结果组的过程中出现的中间分子、细胞或其他表型;
- (ii) 将上述 (i) 中鉴定的表型用作治疗靶点;
- (iii) 将上述 (i) 或 (ii) 中鉴定的表型用作药物开发分析的读数或用作药物开发中的生物标记物;
- (iv) 使用上述 (i) - (iii) 中任何一项鉴定的表型对临床试验中的患者进行分层;或
- (v) 使用上述 (i) - (iv) 中任何一项所鉴定的表型来确定为了获得特定结果,将特定治疗或处理或治疗和处理的结合应当或不应当给与多个患者子集。

## 分析关于生物体的基因数据的计算机实现的方法

### 技术领域

[0001] 本申请涉及一种分析和说明关于生物体的大量基因和表型数据以便理解其生物学的方法。

### 背景技术

[0002] 基因关联研究 (GAS) 评估自然发生的基因变异与给定表型之间的关系。自二十世纪中期以来,已经针对成千上万个(主要是人类)表型进行了GAS(通常是全基因组关联研究:GWAAS,或针对单变异的关联研究,或所述基因组区域内变异的关联研究,或限制于所述基因组的特定区域内的GWAS),在数百万个个体中,产生了数十亿个基因型和表型间的潜在联系。然后通常会简化所得的原始数据以生成概括统计数据。对每个基因变异(无论是估算或观察),GAS概括统计数据包括,所述基因变异对所述GAS的所述表型的推断效应量( $\hat{\beta}$ )和所述推断效应量的标准误差(SE)。在下文中,我们将一个表型称为单项研究的同义词。但是,通常情况是,可以从针对相同或相似表型的多个不同研究中获得数据,或者从测量多个不同表型的单个队列中获得数据。

[0003] 已经开发出许多不同的方法来使用来自GAS的数据以检测与多个表型或性状相关的基因变异(有时称为多效性)。索洛维耶夫等人于(2013)<sup>1</sup>以及哈金格和齐吉尼(2017)<sup>2</sup>在人类基因学领域提供了有用的论述。许多现有的“单点”方法分别分析每个基因变异位点。哈金格和齐吉尼(2017)的表1中提供了此类单点方法的示例,其中特别包括GPA<sup>3</sup>,MTAG<sup>4</sup>和EPS<sup>5</sup>。其他相关方法则侧重于所谓的“表达数量性状基因座”(eQTLs)。eQTL是一种解释了单个基因表达表型的所述变异的一部分的基因变异。值得注意的与eQTL相关的方法包括MT-eQTL<sup>6</sup>,aSPU<sup>7</sup>和SMR<sup>8</sup>。另一类相关方法是应用于原始GAS数据(而不是GAS概括统计数据)的单点方法。哈金格和齐吉尼(2017)的表2中提供了此类单点原始数据方法的示例,其中包括BUHMBOX<sup>9</sup>。

[0004] 尽管GAS提供了有关所述基因变异与表型之间的关系的有价值的信息,但要在基因变异、基因或生物学途径与一种或多种表型之间建立强有力的因果联系,仍然存在相当大的挑战。一个困难是,所述基因组的给定区域内的基因变异通常相互关联,这种现象称为连锁不平衡(LD)。LD的出现使得很难将对表型具有因果效应的基因变异(“因果变异”)和与其相关的其他基因变异区分开来。当正在研究的基因组的所述区域内存在两个或多个因果基因变异,并且来自这些不同因果基因变异的关联信号“渗出”以在其他基因变异处产生明显但虚假的信号(有时是指“虚假的多效性”)时,就会出现进一步的困难。当试图在多个表型的GAS中汇总信息以最大限度提高检测真实效应的统计能力时,这些问题又特别相关。虽然将许多(可能成百上千)GAS汇集在一起可以提高所述统计能力,但这提出了一个巨大的挑战,因为只有在表型已知共享相同因果变异时,汇集才在统计学上对所述表型适用。实际上,这种信息通常是不可用的,并且肯定不能在大量的表型中使用,其结果是在联合分析中不清楚应汇集哪些表型。即使在汇集也可能是基于所述表型的所述相似性的先验信息的情况下,表型之间因果变异的确切共享模式随所述基因组的不同区域而变化。对这些共享的

因果变异进行推断对于获得对生物体的生物学的详细和有意义的见解(insights)至关重要。

[0005] 上述任何一种单点方法都不能直接解决跨表型鉴定共享因果变异的问题,或者只能通过提议任意关联规则和LD强度以特别的方式(ad-hocway)进行鉴定。这些局限性可以通过在所述基因组的相同区域内使用多个变异以明确推断不同表型的关联信号与相同因果变异是否有潜在一致性来克服(有时称为“共定位”)。

[0006] 迄今为止,几乎所有共定位和相关方法一次只能应用两个GAS数据集。采用这种成对方法的方法包括coloc<sup>10</sup>, gwas-pw<sup>11</sup>, eCAVIAR<sup>12</sup>, enloc<sup>13</sup>和JLIM<sup>14</sup>。从这种成对的方法推断多个表型之间的关系是有问题的。两个主要限制是,首先,成对分析的数量随着GAS数量的增加呈二次方增长,因此,如果添加更多GAS(数据由成千上万的GAS提供),则无法很好地扩展;其次,当尝试检测由两个以上表型共享的信号时,这些方法的统计功能将会丢失。

[0007] 最近,已经提出了使用GAS概括统计数据的方法,可以对两个以上的表型进行联合分析。2016年的出版物提出了RiVIERA-MT<sup>15</sup>方法,2017年的另一出版物提出了mcoloc<sup>16</sup>,类似于RiVIERA-MT的贝叶斯框架。但是,2017年出版物的作者指出“可能的组合数量随着性状数量的增加而呈指数增长,因此计算时间是一个限制因素,实际上,它最多可以适用于四个性状”。

[0008] 考虑到现有的在先技术,仍然需要一种能够组合来自许多GAS的数据的方法(例如,大于50,优选大于100,更优选大于500,更优选大于1000,更优选大于5000)来推断在计算有效框架中共享因果关系变异的表型组。

[0009] 当将GAS数据应用于不同类别的表型(涵盖从基因表达,通过内表型到二元和纵向端点的范围)时,这种方法将获得对生物体的生物学的更完整的了解(通过在变异和表型之间,基因和表型之间,以及生物学途径和表型之间建立联系)。在整个基因组中,对共享因果变异的表型组的高效鉴定,不仅提供对个体生物学机制的具体见解(specificinsights)(通过确定基因变异以这样的方式干扰生物学,从而对已鉴定的表型组产生因果影响),而且还可以获得共享生物学机制的全面理解(通过分析整个基因组的表型分组模式)。在应用到人类GAS数据的情况下,这种计算机模拟方法将有助于洞悉生物学机制,否则将需要对人类受试者进行直接的,通常是侵入性的实验。

## 发明内容

[0010] 本发明的目的是提供一种分析关于生物体的大量基因和表型数据的计算机实现的方法,该方法是可以扩展的,而没有过多的增加计算需求。

[0011] 根据本发明的一个方面,提供了一种用于分析关于生物体的基因数据的计算机实现的方法,包括:接收包括多个输入单元的输入数据,其中每个输入单元源自提供沿着基因组的多个基因变异的每一个和对应于所述输入单元的表型之间的关联信息的研究;然后对生物体的所述基因组选择一个或多个区域;然后对于每个选定的区域,将所述每个输入单元分配至多个集群中的一个,所述输入单元的所述分配是基于在选定区域内输入单元共享基因变异的程度的评估,所述基变异影响对应于每一个输入单元的表型的任何方面或所述表型的任何潜在生物学机制,从而基于所述输入单元如何被分配至所述集群,来鉴定共享潜在生物学机制的表型。

[0012] 通过分析大量的表型,可以表明许多因果基因变异会影响多个表型。这样的发现可以有效地将输入数据表示为一组隐藏的表型集群,这些表型沿着基因组变化。因为假定每个集群具有相似的因果变化模式,所述方法在集群内的表型之间建立了生物学联系。因此,提供一种在使用合理的计算资源同时使用来自许多GAS (涵盖从基因表达,通过内在表型到二元和纵向端点的范围) 数据的方法,从而提供对生物学机制的详细见解。通过有效地集群(clustering)许多表型,所述方法可以同时提供改进的检测变异与表型,基因与表型,和/或生物学途径与表型之间的关系的能力,这反过来也提供关于基因因果关系和生物体的详细生物学机制的进一步见解。通过用已发表的生物学研究数据来证实发现,所述方法的输出结果仍然可以变得更加有力。另外,本公开的实施例的统计能力可以允许对真实生物过程进行详细和可靠的检查和对个体生物几乎没有风险,以及在许多情况下,以较高的时间和成本效率的情况下确定分子水平扰动(perturbing)对细胞和组织功能的影响。这是可能的,因为所述输入数据的规模超出了其他方法的能力;在一些实施例中,全部输入数据会最终从人类研究的参与者中获得。

[0013] 由于单个GAS通常检查单个表型,因此GAS可以类似地被集群,关联或分配到具有功能等效含义的集群,集群的部分或集群的成员。另外,源自GAS的输入单元可以等同地被集群,关联或分配到集群,集群的部分或集群的成员。另外,我们使用词组因果变异来表示影响所考虑的表型的潜在生物学的基因变异。但是,我们不一定需要假设因果变异在输入数据中,因为存在连锁不平衡,当因果变异不包含在输入数据中时,附近的基因变异可以作为所述因果变异的代理(proxies)。在下文中,当出现在输入数据中时,我们使用“因果变异”来指代真实因果变异;或者当不存在时,指代代理。

[0014] 将在下面更详细地描述,这种方法的所述输出可以以各种方式使用,特别是:(i) 对生物体的生物学机制的详细了解;(ii) 在疗法的发现和开发中;(iii) 计算个人患上特定疾病,状况或其他任何目标表型的机会;(iv) 协助个人或其医疗保健专业人员或两者做出临床决策。所述方法的这些应用可以利用所述输入(通常是有关生物体的基因信息),并且更广泛地可以利用所述输出(即有关生物学机制的详细信息)。

[0015] 在一个实施例中,所述每个所述输入单元到多个集群中的一个的分配还包括在集群的数量和/或大小上使用先验分布(prior distribution)。

[0016] 所述实施例的优点在于,它使得能够使用计算上有效的方法来推断集群成员,而不会事先限制所允许的集群数量。于是,所述方法可以自动适应所述数据中证据的强度。这种内在的适应性使所述方法可以应用于各种类型和数据的大小,而无需对表型之间的基因和生物学关系施加强假设。

[0017] 施加先验分布,将较低的概率分配到更多的集群上,具有防止“过度拟合”的好处(当函数过于紧密地与有限的数据点集合对齐时会发生建模错误)。对真实集群的改进的检测和防止对错误集群的检测增强了所述方法对生物学机制的准确解释的整体实用性。

[0018] 在一个实施例中,所述多个集群包括空集群。

[0019] 所述实施例的优点是明确包括特定表型的数据不包含任何因果变异的可能性。所述方法得益于以下事实:可以对所述空集群的特征预先确定,因此不需要对空集群内其他表型的信息进行更新,从而在规模上提高计算效率。有证据表明不存在因果变异(即通过对所述空集群的强分配)的表型的鉴定对于定义表型之间的生物学关系很重要。没有因

果变异效应的证据对于所述输出的许多下游应用至关重要(例如,调节基因或通路的效应不会影响特定的表型)。

[0020] 在一个实施例中,所述每个输入单元到多个集群中的一个的分配是基于采用概率模型对所述分配的评估。

[0021] 所述实施例在概率模型内实现了所述集群,该概率模型描述了所述数据和包括集群成员的标签的未观察到的目标参数的联合分布。所述方法受益于能够表征集群分配的确定性。另外,可以通过马尔科夫链蒙特卡洛算法(Markov chain Monte Carlo algorithm)来探索可能的集群分配的空间。这种概率模型在基于所述方法的所述输出进行决策的过程中具有广泛的应用。

[0022] 在一个实施例中,所述方法计算每个输入单元和每个集群之间的相似性度量。这种相似性度量可以直接用于在推断时,特别是在潜在因果变异上,边缘化集群分配中的不确定性,或者可以用于表征所述算法的所述输出。可以通过多种方式来计算这种相似性度量。在一个实施例中,重复地执行输入单元到多个集群中的一个的分配,导致多个配置共同反映输入单元到集群的分配的由有限量的可用输入数据造成的固有不确定性。在以预设次数重复执行所述分配或直到满足预定条件之后,将所述相似性度量计算为,在所述过程中每个输入单元已经成为每个集群的成员的频率。在另一个实施例中,所述相似性度量被计算为首先通过用于将输入单元分配到集群的过程将输入单元分配到集群的概率。实际上,计算输入单元分配到多个集群中的一个的概率是为了对所述集群分配进行采样。

[0023] 在一个实施例中,所述方法还包括,对于所述多个集群中的一个或多个中的每一个,使用所述集群的所述特征组来从所述多个基因变异中鉴定一个或多个重要基因变异,所述多个基因变异可能与对应于所述集群的一个或多个表型有因果关系。该实施例增加了生物学解释和提高了通过使用集群类内的所有表型表征基因变异组来鉴定集群的所述统计能力,所述基因变异可能与对应于所述集群的一个或多个表型有因果关系。现有方法试图从与给定单表型显著相关的证据中明确鉴定出因果变异。该实施例可以将这样的信息视为“干扰变量(nuisance variable)”,并且为了有效地鉴定集群的目的和同时保持使用所述算法的所述输出来推断因果变异的能力,而将其分析性地整合。该方法的实用性是进一步增强将集群内所述表型与所述基因组的所述功能联系起来的能力,并且通常有助于鉴定基因,因此提供对由因果变异干扰的(perturbed)生物学机制的进一步了解。

[0024] 在一个实施例中,所述方法还包括:鉴定与输入单元相对应的表型组,所述输入单元在频率高于预定频率阈值的多个组中被分配到相同集群。

[0025] 该方法的优点是能够更容易地鉴定潜藏在目标表型下的分子机制。例如,有可能鉴定出与目标表型相关的集群中富集或耗尽的所有表型。该信息允许对所分析表型的共享的或不同遗传基础进行详细探索。

[0026] 在该实施例中,所述集群可用于鉴定区域内的一组因果变异,每个因果变异可能对应于所述区域内的多于一个集群或所有集群,并且每个集群可能对应于所述区域内的多于一个因果变异,从中可以获得因果变异组对每个研究的影响的联合估计,并根据需要调整估计的影响的相关性。这些联合估计提供了有关输入单位生成中正在研究的表型之间关系的进一步生物学信息。然后,可以通过直接比较不同研究之间和集群之间的所述效应大小的评估来确定集群之间的相似性。

[0027] 此外,在一个具有特征为允许针对每个集群鉴定出多于一个因果(或重要)基因变异的可能性的实施例中,集群之间的所述表型的比较(跨多个集群组每种情况下)可以基于对共享集群成员的评估或通过输入单元中已鉴定的因果变异的效应大小(effect size)一致性的评估来进行。

[0028] 一个实施例还包括考虑提供用于导出输入单元的信息的研究之间的相关性或其他已知关系。

[0029] 该实施例的优点是能够考虑到关于生物体中生物机制的预先存在的信息或知识,以提高结果的准确性和可靠性。该信息可以包括关于与一般表型或特定表型具有因果关系的特定变异似然性的信息,或者与表型具有共享集群成员的似然性有关的信息。该信息还可以包括通过所述方法获得的超出目标区域的包括从全基因组数据获得的信息,然后使所述方法能够得知并受益于研究之间或输入单元之间的相关性。

[0030] 一个实施例还包括对基因组的多个选定区域的每一个执行所述方法,从而提供包含分配了输入单元的多组集群,每个集群组对应于多个选定区域中不同的一个选定区域,并且通过与集群的所述第一子集内的输入单元相对应的表型与集群的一个或多个第二子集内的输入单元相对应的表型之间的相似性的评估,将从第一组的一个或多个集群的第一子集与一个或多个其他组的每一个的一个或多个集群的第二子集联系起来。

[0031] 该实施例的优点在于,可以进一步分析鉴定出的集群组以提供对共享生物学机制的见解。用确定性或概率性方法对集群进行聚类(clustering),可以根据生物扰动对生物体的表型的影响,在不同的生物扰动之间建立关系,从而提供对关联的生物学机制和过程的见解(insights)。例如,显示与许多基因组区域中的与目标疾病相关联的集群的子集可以进一步探究以寻找共享的和独特的致病机理。

## 附图说明

[0032] 本发明实施例将通过示例并参考附图的方式进行进一步描述。

[0033] 图1示出了基因组中的基因和区域;

[0034] 图2示出了基因组区域内的基因变异,以及使用研究信息导出输入单元的示意图;

[0035] 图3示出了一实施例中输入单元到集群的分配的示意图;

[0036] 图4示出了所述方法的一实施例中的步骤流程图;

[0037] 图5示出了具有相关的重要变异的集群;

[0038] 图6示出了所述方法的一实施例的输出;

[0039] 图7示出了现有技术方法与本公开的一实施例之间的比较;

[0040] 图8示出了所述方法的一实施例使用先前发布的GAS数据的输出的示例。

## 具体实施方式

[0041] 本方法旨在聚类表型,并对集群的共享生物学机制提供解释。表型可以说是被集群的(被聚类的),关联的或分配到具有等效含义的集群,集群一部分或集群的成员。

[0042] GAS研究的表型示例包括:基因(和相关核苷酸序列)的表达水平,和表达调节;表观遗传特征(例如核苷酸修饰、染色体构象);蛋白质或肽的丰度水平;蛋白质或肽的功能和/或分子结构;生物体的分子数量(例如药物、激素、DNA分子或RNA分子、代谢产物、维生

素);生化和代谢过程的特征(例如基础代谢率、凝血酶原时间、活化的部分凝血活酶时间);细胞形态和功能(例如,红细胞平均血球体积,绝对中性粒细胞计数);组织形态和功能(例如,骨矿物质密度、头发颜色);器官和器官系统的形态和功能(例如,左心室射血分数、最大肺活量);对外部刺激或多个刺激的任何反应(例如光、声音、触摸或任何其他感官输入);对暴露于某种物质或病原体的任何反应(例如饮食输入、药物、气体、病毒、细菌);行为和生活方式特征(例如,吸烟、饮酒、职业);生殖和生命历程的特征和功能(例如初潮年龄、胎盘重量、受教育年限);疾病或状况的发生、发展轨迹和预后(例如糖尿病、心血管疾病、肥胖症);可测量的解剖学特征(例如,体质指数、瘦肌肉质量、身体脂肪百分比);可测量的生理或功能特征(例如,心率、血压、智力);以及可测量的心理或认知特征(例如,流体智力指标、精神病性症状)。这些测量中的任何一个都可能是绝对的或相对的。

[0043] 本公开的一个实施例使用一种基于模型的集群形式,其可以被直接解释为在集群内鉴定共享的因果基因变异。基于距离指标或可扩展模式识别方法学的用于对大数据组进行聚类的其他方法也可以在这种情况下应用。

[0044] 在其最高水平上,本公开提供了分析关于生物体的基因数据的计算机实现的方法,其中输入单元来自于提供关于基因变异与表型之间的关联的信息的研究。然后,基于对输入单元共享基因变异的程度的评估,将输入单元分配到多个集群中的一个,从而鉴定共享潜在生物学机制的表型,所述基变异影响对应于每一个输入单元的表型的任何方面或所述表型的任何潜在生物学机制。所述评估可以基于多个基因变异中的一些或全部的关联度量或这种度量模式之间的相似性的评估。

[0045] 示例性实施例将采用所述方法的各发面的特定实施例来进行描述。所述示例实施例包括使用GAS数据来鉴定以下各项的组合的方法:(i)分配了表型(以及所述集群的GAS,表型是对象)的集群组;(ii)对于每个集群,在选定区域中至少有一个重要(即推定的因果)基因变异。所述示例实施例包括描述所述输入单元的联合分布和模型的潜在变量的概率模型(更具体地是贝叶斯统计模型),其包括聚类分配。所述实施例使用马尔科夫链蒙特卡洛算法来探索该模型的后验分布,并因此探索可能的集群分配的空间。所述马尔科夫链蒙特卡洛算法从该后验分布中产生一个样本,并且该样本用于产生最终的集群(final clustering)。通过这样做,所述实施例实现了鉴定表型集群和共享潜在生物学机制的相关基因变异的目标。

[0046] 在示例实施例中,所述方法具有六个主要部分:即:(i)鉴定所述基因组的选定区域以进行分析的方法;(ii)指定在选定区域中存在的表型集群的数量和/或大小的先验分布的方法;(iii)计算一个特定基因变异或多个特定基因变异与给定GAS/表型有因果关系的概率的似然函数的方法;(iv)将多个研究之间的这些似然性结合起来的以进行荟萃分析(metaanalysis)的方法;(v)将所述GAS分配到集群的方法;(vi)迭代地执行步骤(v)直到满足某些收敛标准,并且在每次迭代中使用分配系列来获得最终的集群分配的系统。在这些步骤中,(ii)-(v)可被视为所述概率模型的说明书。在使用贝叶斯模型的示例实施例中,这些步骤被设计为通过马尔科夫链蒙特卡洛算法从该模型的所述后验分布中采样。

[0047] 所述示例实施例的第一部分包括形成用于分析的所述基因组的选定区域。选定区域可以任意选择。然而,选定区域的所述形成期望在以下两者之间取得平衡:(i)最小化所述选定区域大小以避免捕获对单个集群有因果关系的多个基因变异;(ii)最大化所述选定

区域的大小以包括彼此相关的所有基因变异,从而最小化由于其与选定区域之外的其他基因变异相关而将因果关系错误地归因于某个基因变异的可能。有很多方法可以做到这一点。在一个替代实施例中,可以将选定区域围绕特定目标的基因或所述基因组的其他功能部分周围设置。参照图1,在示例实施例中,选定区域2包括特定的目标基因4,并且包括围绕所述基因的预定数目的碱基对的窗口(在 $10^2$ 和 $10^{11}$ 个碱基对之间,例如1兆碱基对)。如图2所示,所述选定区域2包括多个基因变异3a-3h,其中一些(变异3a-3d)在目标基因4内,而其中一些(变异3e-3h)在目标基因4之外。在一替代实施例中,可以通过在所述基因组1上平铺窗口来形成选定区域2。这些窗口的大小可以通过例如物理距离(即所述区域中碱基对的数目)或基因距离(其与重组率有关)来决定。另一个选择是针对选定区域2的不同选择自适应地评估跨所述基因组1的表型的集群分配的变化。

[0048] 所述示例实施例的第二部分包括在所述集群的数量和/或大小上生成先验分布的方法。该示例实施例利用了由所谓的中国餐馆过程(CRP)引起的分配。简而言之,所述CRP是一种特定的用于将连续的单元分配到集群或组的随机模型,使得组本身是随机的并且没有预先指定<sup>17</sup>。该示例实施例通过添加具有预定的先验概率 $p$ 的单个“空”集群扩展了所述CRP,而由所述CRP诱发的其他集群被共同分配了剩余的先验概率(所述CRP是特殊情况,其中 $p=0$ )。因此,集群的数量是从基于具有先验参数 $\alpha$ 的CRP的数据评估的。另一种生成先验分布的方法是固定集群的所述数量,并通过狄利克雷(Dirichlet)分布分配先验概率。

[0049] 所述示例实施例的第三部分包括生成似然函数的方法,该似然函数描述了给定GAS表型的一个假定的集群分配时所观察到的GAS概括统计数据的所述概率。这对应于确定输入单元的所述特征。

[0050] 参照图2,GAS 5、6至少包含与所述GAS相关的所述表型的概括统计数据 $\hat{\beta}$ ,SE。每个GAS 5、6用于导出输入单元10、20,所述输入单元10、20的所述特征包括一组指标,这些指标量化了在选定区域2中的每个基因变异3a-3h在每个输入单元10、20中对输入单元10、20的表型具有因果关系的证据强度。输入单元的所述特征还可以包括所述输入单元的除概率指标之外的其他函数。在所述示例实施例中,所述推断的效应大小 $\hat{\beta}$ 和推断效应大小的标准误差SE是用于导出输入单元10、20的最重要的所述GAS概括统计数据。

[0051] 在所述示例实施例中,所述似然函数被选择作为概括统计 $\hat{\beta}$ ,SE的每个变量似然性(per-variant likelihood)。对于占据非空集群(non-null cluster)的基因变异,由替代模型( $H1: \beta \sim N(0, SE^2 + \sigma^2)$ )给出,而对于分配至所述空集群的基因变异是由所述空模型给出( $H0: \beta \sim N(0, SE^2)$ )。在这两种情况下, $N(x, y)$ 是均值为 $x$ 和方差为 $y$ 的正态分布。不直接处理这些似然性,而是将模型下所述似然性除以空模型下的似然性,以获得“缩放(scaled)”似然性。在所述示例实施例中,所述替代模型的所述缩放似然性采用贝叶斯因子BF的形式,从而用于确定相似程度的输入单元的特征包括在与对应于所述输入单元的表型成因果关系的在所述选定区域的多个基因变异的每一个的贝叶斯因子。在所述示例实施例中,所述贝叶斯因子由下式给出:

$$[0052] \quad BF(SE, \sigma^2) = \frac{p(H_1)}{p(H_0)} = \frac{p(H_1)}{p(H_0)} = \sqrt{\frac{SE_i^2}{SE_i^2 + \sigma^2}} \exp\left(\frac{\hat{\beta}_i^2}{2 SE_i^2 (SE_i^2 + \sigma^2)}\right)。$$

[0053] 第二个等式合理的条件是,假设所述区域没有其他变异是因果关系的,基因变异是因果关系的(参阅马勒等人<sup>18</sup>),对于最后一个等式参阅克菲尔德<sup>19</sup>。对于所述空模型(nullmodel),缩放的似然性只是一个因子1。由于这些贝叶斯因子仅取决于 $\hat{\beta}_i, SE_i$ ,因此可以针对每个GAS中包含的每个基因变异进行预先计算。在所述示例实施例中,通过使用因果基因变异上的先验分布将因果基因变异i边缘化来计算“边缘化”的似然性。这意味着在所述示例实施例中,当使用多个基因变异中的每一个都是因果关系的似然性的先验分布来计算集群的所述特征时,输入单元被组合。在所述示例实施例中,对于每个基因变异,使用均等的似然性和统一的先验分布。

[0054] 替代实施例可以通过在计算中加入一个信息丰富的函数先验,利用跨区域的基因变异的功能注释而不是在统一的先验分布上来边缘化,将所述基因组的所述选定区域内那些基因变异是可能为因果变异的其他信息纳入。如果使用信息丰富的函数先验,所述似然性的先验分布包含了与基因变异因果相关的功能变异的预先存在的信息。函数先验在其他方法中是已知的;例如PAINTOR方法<sup>20</sup>,但之前尚未应用于此处描述的此类型方法中。

[0055] 在所有基因变异都具有关于它们在每个GAS有关联的完整信息的假设下,则所述边缘化似然性是每一个基因变异使用每个变异是因果变异的概率性的所述先验分布以进行加权时的所述BF的总和。

[0056] 所述示例实施例的第四部分包括用于对分配到集群的GAS进行荟萃分析的方法,以便计算每个基因变异是因果基因变异的概率。这包括通过组合分配到所述集群的输入单元来计算所述集群的所述特征。在所述示例实施例中,假设GAS是独立的,并且所述效应大小被假定为不相关,从而可以在整个GAS上累积证据,作为针对每个GAS计算的贝叶斯因子的乘积。然后,集群的所述特征的计算包括计算分配到所述集群的所述输入单元的所述贝叶斯因子的乘积。所有这一切组合成完整的集群似然性,它表示在已分配GAS的情况下集群的似然性。在示例实施例中,其计算如下。

[0057] 设 $BF_{ij} = BF(SE_{ij}, \sigma^2)$ 为与变量i和GAS j相关的所述贝叶斯因子,令 $C_s$ 为分配到集群s的GAS组,其中, $C_0$ 表示分配到空集群(其中被认为在选定区域未携带因果基因变异)的所述GAS。令K为(非空)集群的数量,令 $k = (|C_1|, \dots, |C_K|)$ 为集群大小的向量,而N为GAS的总数。最后,令p和 $\alpha$ 为集群的数量和大小的先验分布的参数,其中,所述参数p确定将GAS分配到所述空集群的先验概率,因此假定在所述选定区域中没有携带因果基因变异。首先,基因变量体i与集群s的表型组有因果关系的综合证据定义为:

$$[0058] \quad meta_{is} = \prod_{j \in C_s} BF_{ij},$$

[0059] 知道 $BF_{i0} = 1$ ,因此, $meta_{i0} = 1$ 。然后,整个集群似然性变为:

$$[0060] \quad L = p^{|C_0|} (1-p)^{N-|C_0|} \Pr(\alpha) \prod_{s=1}^K \left( \sum_i \Pr(i) meta_{is} \right).$$

[0061] 此处, $\Pr(i)$ 是基因变异i在集群中有因果关系的先验分布,在所述示例实施例中是统一分布(即 $\Pr(i) = \frac{1}{G}$ ,其中,G是所述选定区域的基因变异的数量)。最后,在所述示例实施例中, $\Pr(\alpha)$ 是具有浓度参数 $\alpha$ 的中国餐馆过程,或者明确地:

$$[0062] \quad \Pr(\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+|k|)} \alpha^K \prod_{i=1}^K \Gamma(k_i),$$

[0063] 其中,  $|k| = k_1 + k_2 + \dots + k_K = N - |C_0|$ 。

[0064] 所述示例实施例的第五部分是将GAS分配到集群的过程。通过确定所述输入单元的特征和集群的所述特征之间的相似度概率,将源自GAS的输入单元分配集群。该相似度可以用于确定将所述输入单元分配到所述集群的概率。在所述示例实施例中,每个GAS  $j$ 最初被分配到所述空集群。然后,使用马尔科夫链蒙特卡洛算法,依次考虑每个GAS  $j$ ,将其从其当前集群中移除,并计算将其分配到以下各项的先验和似然性:(i)所述空集群;(ii)一个新的集群;(iii)每个现有的集群。然后,该算法随机进行新分配,以与紧接在上述(i)至(iii)中描述的GAS  $j$ 分配到每个所述集群的似然性成比例的概率。分配输入单元后,可以重新计算分配到它的集群的所述特征。

[0065] 更详细地,在所述示例实施例中,首先将考虑的GAS  $j$ 从其当前集群 $s$ 中移除(如果已将其分配到空集群以外的集群),然后针对保持分配到该集群的所述GAS重新计算该集群 $meta_{is}$ 的特征。接下来,计算相似度,在所述示例实施例中,相似度是将研究 $j$ 分配到任何现有集群 $s'$ ,新集群 $s'' = K+1$ ,或所述空聚类( $s' = 0$ )的似然性。这些可能性中的每一个的似然性是:

[0066] 空集群: $p$ ;

[0067] 新集群 $s'' = K+1$ :  $\frac{(1-p)\alpha}{|k|-1+\alpha} \sum_i BF_{ij} \Pr(i)$ ;

[0068] 现有集群 $s'$ :  $\frac{(1-p)k'_s}{|k|-1+\alpha} \sum_i BF_{ij} \frac{\Pr(i)meta'_{is}}{\Pr(i)meta'_{is}}$ 。

[0069] 在所述示例实施例中,然后以与该可能性的所述似然性成比例的概率来选择这些可能性中的一个。一些实施例可能不包括所述空集群,其在以上等式中相当于设置 $p=0$ 。

[0070] 上述的所述算法通过将每个研究重复一定次数迭代地执行,例如在100到1,000,000次之间,但通常为10,000次,或直到达到预定的收敛阈值。然后,所述输出可用于描述研究到集群的所述分配和有关可能的因果变异的信息,以及潜在的数量可信度。这些输出可直接用于鉴定共享的生物学机制。

[0071] 在由于达到预设的收敛标准或者经过预定数量的迭代之后,而停止马尔科夫链蒙特卡洛算法之后,可以对所述算法每次迭代时的所述集群分配进行后处理,以提供各种形式的单个集群分配方法。在所述示例实施例中,所述算法以固定的迭代次数运行,并且最终研究分配是基于马尔科夫链蒙特卡洛算法的最后迭代时的分配。其他自然的可能性(natural possibilities)是,例如,在链的步长(steps of the chain)的子集上的适当数量求平均值,比如集群的共现,并使用这些值来确定最终的集群分配。另一种方法是计算每个集群共享一个集群的频率,并形成最大化集群内共享频率和最小化集群内共享频率的集群。

[0072] 在所述示例实施例的集群的分配研究可以参考图3和图4而可视化,其中,图3表征了图4中概述的算法的可能实现。在图4的步骤S0中,包括多个输入单元10-80单元(如图3所示)的输入数据被接收。在步骤S1中,确定包含多个基因变异3a-3h的选定区域2(如图2所示)。在步骤S2中,为多个输入单元10-80中的每一个计算特征组11-81(如图3所示)。对于特征组11-81中的每组,每个点都显示多个变异中的一个与对应于所述输入单元的表型的关联,其中x轴显示所述变异在选定区域2内的位置(如图1和图2所示),y轴显示关联强度。在步骤S3中,选择具有特征11的输入单元10。最初,没有形成非空集群,因此没有计算集群特

征。在图3所示的示例实现中,将输入单元10分配到新的集群100。过程返回到步骤S3,选择具有特征21的输入单元20。基于分配到集群100的输入单元10的特征11计算集群100的特征组101。输入单元20的特征21与集群100的所述特征101之间的相似度很小,因此输入单元20还分配到新集群110。重复步骤S3至S7,直到达到预定收敛阈值或执行预定迭代次数。

[0073] 在所述分配过程结束时,将输入单元10-70分配到集群100、110和120。例如,在图3中,将输入单元80分配到空集群130。从分配到每个集群的输入单元10-70的特征组11-71的组合计算出集群100,110和120的特征组101、111和121。例如,集群110的所述特征组111是由输入单元20、40和50的所述特征组21、41和51的组合形成的。图5说明了使用集群110的所述特征组111以确定所述重要变异112,它最可能与所述集群110有因果关系。

[0074] 所述方法的输出在图6中示出,显示重要变异112与被识别为集群110的成员的性状(包括基因表达度量和其他表型)的关联。中心零线(centralnull line) 200右侧的效果估计值202指示与重要变异112正关联,而左侧的效果估计值201则与重要变异112负关联。效果估计上的条(Bars)表示不确定性。条210针对每个性状指示集群110成员的后验概率的大小221和所述空集群130(在图3,假定分配到其他集群100和120的后验概率大约为0)成员后验概率的大小220。这些描述了围绕集群分配的所述确定性,可以使用马尔科夫链蒙特卡罗算法的所述输出来确定。

[0075] 在所述示例实施例的所述选择允许在所述分配过程的每个阶段进行特别有效的计算。特别是,基因变异*i*对GAS集群是因果关系的所述(缩放的)似然性仅仅是使用这些GAS计算出的贝叶斯因子的乘积;然后,通过使用这些(缩放的)似然性对*i*进行边际化以及对基因变异进行先验分布,从而获得全部似然性。

[0076] 这有效地提高了使用此方法可以分析的所述数据的范围。现有方法最多只能同时分析大约十二个GAS中的数据。相反,本公开的实施例是完全可扩展的,并且可以在整个可获得的表型范围内考虑来自成千上万个GAS的数据。

[0077] 此外,现有方法并不试图明确聚类(explicitly cluster)表型和在所述基因组的选定区域中鉴定出一个或多个重要变异,而是将其推论基于“因果矩阵”*C*,如果基因变异*i*对表型*j*有因果关系,其中 $C_{ij}$ 为1。这种方法仅对所述集群进行隐性编码,而且在计算上分析起来比较困难。

[0078] 图7给出了本方法与先前方法之间的视觉比较。常规的全基因组关联分析301鉴定了与疾病风险相关的基因变异。但是,他们很少能对所述观察到的关联的真正原因,其潜在的生物学机制或潜在药理学调控的靶标提供深入的了解。一系列性状的基因关联分析之间的比较302可以提供一些机械的见解。然而,在区分离散的生物学通路时,这些常常受到分析挑战的阻碍。因此,这些观察可能会产生误导。使用本公开303的实施方式一起分析大量性状,允许在其各自的通路中对机械步骤进行分组和排序。通过这样做,还可以鉴定生物学的“压力点”304和305(与代表有价值的治疗靶标的疾病病因密切相关的病理通路的特征)。

[0079] 与已知方法相比,本公开实施例的增强的效率和功率赋予了显著的优点。至关重要的是,所述方法能够鉴定所述区域内的大量不同相关和不相关关联信号组。例如,当这些表型都由相同的基因变异的原因驱动时,在同一集群同时考虑,例如基因表达水平、蛋白质表达、生物标记和疾病终点等表型是有益的。此外,有益的是在对表型单独进行集群,所述表型在所述基因组的相同选定区域中与不同的基因变异有因果相关。这两个方面与成千上

万的GAS的分析特别相关,因为在相同的选定区域中发生相关和不相关关联的所述增加的机会。

[0080] 因此,本公开实施例在区域内将如此广泛的GAS集群的能力意味着其可以用于:(i)在集群内的表型之间建立机械的因果关系(即证明机械联系或鉴定生物学通路);(ii)鉴定和辨别基因变异与许多目标表型之间的机械独立重叠关联,否则这些关联可能被认为是通过相同的机械关系起作用的;(iii)通过在关联模式中鉴定相似性而由基因变异与目标表型之间弱的关联信号鉴定真实关联,所述弱的关联信号可能会被应用到在本发明方法中的不形成集群的个体弱的关联信号的基于阈值的方法丢失。

[0081] 本方法的潜在输入范围的广度和性质使得用户可以确信将呈现所述相关生物学的非常完整的图像的输出生。因此,所述输出可以包括共享潜在生物机制的表型的所述鉴定,和/或针对表型的集群重要变异的鉴定,可以应用于许多不同的,有价值的应用中,其可以使用所述输入(通常是关于个体的所述基因信息),并且也可以更广泛地,利用所述输出(即,关于生物学机理的详细信息)。

[0082] 所述方法最重要的应用是能够生成关于生物体中不同基因、细胞、组织、分子机制和生物学通路之间的关系的详细了解。例如,输出结果使用户可以了解到,当某种特定基因受到某种形式的干扰时,很可能将会发生特定的后果,从而可以得出有关相关疾病和状况的有价值的结论。所述方法的许多后续应用都考虑这些复杂的生物学见解。

[0083] 在说明所述方法揭示这种生物学见解的能力的示例中,图8描述了将先前描述的示例实施例应用于成千上万个先前发布的GAS的数据的结果。在TWIST1基因周围的区域中,通过所述示例实施例鉴定了多个包括链接至TWIST1的两个不同集群的集群。在图8a中显示了极有可能成为所述第一个集群成员的性状(顶部的刻度给出了针对每个性状集群成员的后验概率)。对该集群成员的性状进行的荟萃分析会生成一个组合信号图,如图8b所示(y轴刻度为所述区域每个常见变异处进行关联的 $\log_{10}$ 贝叶斯因子),而这依次生成一组可靠的因果变异,具有变异rs2107595作为其主要成员(图8c)。极有可能成为该集群的成员的性状包括收缩压和冠状动脉疾病、高血压和大血管缺血性中风的风险(回想,在所述聚类过程中未使用任何有关表型的背景)。该集群的另一个目标性状是在大动脉的基因TWIST1的表达水平。

[0084] 因此,该分析提供了充分的证据,表明脉管系统中TWIST1的水平在冠心病、高血压和大血管缺血性中风的风险中起因果生物学作用。图8a左侧的箭头也指示了作用的方向:脉管系统中TWIST1水平的增加与这三种状况的风险增加有关。相反,对该基因组区域的广泛研究在2012<sup>21</sup>年被确定为大血管缺血性中风的危险位点后,未能查明因果基因。已经将TWIST1建立为可能的因果基因,最近的一项研究<sup>22</sup>表明TWIST1在内皮增殖中作为对切应力的血管反应,这为它的作用提供了合理的机制。

[0085] 为了进一步说明所述方法揭示复杂生物学见解的能力,图8d显示了高概率属于TWIST1基因周围同一区域中鉴定出的单独集群的性状。对作为该集群成员的性状进行的荟萃分析生成了一个组合信号图,如图8e所示,这又生成了一组可靠的因果变异,具有变异rs67248060作为其主要成员(图8f)。极有可能成为该集群成员的性状包括男性型秃顶,以及脑和甲状腺中基因TWIST1的表达水平。

[0086] 总之,图8a-8f中描述的所述分析说明了所述方法揭示新的生物学见解的能力。所

述分析解释了生物学驱动所述先前发布的中风GAS信号,将其与其他心脏表型联系起来,并提供了对具有不同生物学后果的同一基因进行组织特异性调节的示例。所有这些观察都是新的。这些观察结果也提供了明确的治疗假设:降低脉管系统中TWIST1蛋白的水平,将是治疗冠心病、高血压和大血管缺血性中风的药物的理想靶点和作用方式。

[0087] 所述方法特别适合于治疗剂的发现和开发,特别是用于人类的治疗。所述方法的输出可用于提出,支持和验证可支持药物靶标发现,药物靶标验证以及药物重组和重新规划项目的全部或部分的假设。这样的假设可以包括治疗假设,例如,抑制一个被命名的基因或会有一定的生物学后果的基因产物,这反过来将有助于被命名的疾病的治疗。同样,所述方法可用于辅助研究药物-药物相互作用、靶标安全效应、生物标志物的鉴定、患者选择、临床试验设计、临床试验中人类患者的分层以及剂量方案。

[0088] 本公开的实施例可以应用于有兴趣去了解生物体的基因型与其在其一生中表现出的性状和表型之间的关系的任何情况。在人类中,有关个体基因型和性状之间关系的信息可能具有内在的意义,并直接提供给个体。所述信息可能与他们的健康和生活方式选择有关。

[0089] 本公开的实施例还可以应用于更多的临床情况。所述方法可用于改进以预测生物体发展为特定疾病或状况或其他任何目标的表型的机会的计算。本发明的进一步临床应用是可能的,例如使用所述方法的输入和输出来协助医师在个体或个体群水平上进行患者管理。从基因变异预测性状在医疗保健中的特定用途包括但不限于根据个体患特定疾病的风险对个体进行分层;协助疾病诊断或鉴定疾病亚型;告知最佳的行为或医学干预措施;改进筛选程序;改进用于个体比较的参考分布;并筛选参加临床试验的个体。在每种情况下,本公开的实施例都可以在个体,个体群或人群的水平上应用。它们可能适用于特定性状或目标疾病,但也可能适用于构成个体或生物体的所述自然史的任何临床或生物学事件。

[0090] 有关个体患疾病或表现出性状的风险的信息可用于其他环境中,包括:计算保险费;用于植入前后的诊断;针对个体提供商品、服务或机会;或确实在目标性状具有基因基础且可获得有关个体基因型的信息的任何情况下。

[0091] 示例实施例将使用所述方法来估计所述感兴趣表型(目标表型)的所述因果变异和效应大小,并使用这些来对基于个体的基因型进行计算预测。例如,可以通过以下方式实现:

[0092] 1、在大量研究中应用所述全基因组方法,包括但不限于所述目标表型,以鉴定一种或多种因果变异(或因果变异的后验分布)。

[0093] 2、在与所述目标表型相关的所述研究中,估计这些鉴定出的变异或具有较高后验概率的变异的效应大小。这些可以通过适当地调整研究和因果变异之间的效应大小估计之间的相关性来改善。

[0094] 3、获得个体的所述基因型,并对所述表型进行基因贡献计算。

[0095] 4、根据这些因果变异和个体基因型的估计效应大小,计算所述因果变异对所述目标表型的所述贡献。一种方法是适当地组合与它们在每个因果变异上携带的等位基因相关的效应大小,或在因果变异及其影响的后验分布上平均这些效应大小。代表这些对个体风险的影响的所述组合的度量有时称为多基因风险评分(PRS)。

[0096] 5、将所述PRS用作风险的定量度量(例如,与典型个体相比,作为所述相对风险的

估计)或与适当的个体群(cohort)评估其位置(例如,可以鉴定某个个体位于整个个体群的经验分布的前10%)。

[0097] 在另一个示例实施例中,效应大小的所述估计是建立在不同种族的研究基础上的。这种方法的好处是双重的。首先,它完善了影响所述目标表型的因果变异组的推论。其次,所述效应大小估计值结合了来自每个独立种群的值,因此针对目标表型提供了群体的特定估计,通过在群体间部分共享基因效应而使估计值变得更加准确的。

[0098] 在另一个示例实施例中,对预测模型的效应大小的所述估计利用了对不同但相关的性状的研究所的优势。可以使用全基因组信息从数据的性状之间推断相关矩阵或相关矩阵的集合。然后,这些全基因组范围的估计值可用于优化对目标研究的每个因果变异的效应大小的所述估计。最终目的是获得更精确的能够转化出更准确的预测的效应大小估计。应用本文描述的方法的实用性在于,它们允许所有的输入研究有助于集群的所述鉴定,因此也有助于因果变异的鉴定,并允许在推断出的共享变异的研究中估计真正因果效应。

[0099] 在另一个实施例中,可以将上述描述的所述基因预测与其他重要的协变量整合并对其进行调整,包括但不限于:人口统计学信息(例如年龄、性别、身高);基因背景的概括(例如血统、基因纯合性、多基因风险的进一步测量);以及针对基因组特定区域的潜在有针对性的分析(例如,主要组织相容性抗原复合体),其中存在有关基因变异与目标性状之间关系的信息。

[0100] 如果临床数据可用于计算基因对表型的贡献的个体队列,那么对于任何给定的(可能调整的)基因预测,也有可能根据经验来校准特定临床事件风险的增加。在另一个示例实施例中,基因预测(例如,PRS)可以与有关个体的任何其他数据相组合,这些数据可以提供关于感兴趣的未来事件的信息。例如,在临床上,基因信息,例如PRS,可以作为用于预测未来与健康相关事件的更广泛的生物医学数据的一部分。

[0101] 特别是,筛查测试所提供的测量值可以在个人所涉及疾病的基因风险的背景下解释,因此优先考虑筛查分数接近阈值且具有高基因风险的个体进行进一步的临床研究。同样,可以根据高风险个体的基因评分,可以优先进行筛查。这可以通过使用基因风险来计算发展为目标疾病的先验似然性,然后使用筛选数据来计算疾病状态的后验概率来实现。

[0102] 通过将有关个体的这些其他类型的临床和/或基因数据与先前确定的基因型贡献结合起来,可以改善对特定个体当前或将来可能的表型的确定。

[0103] 所述示例性实施例可以与其他利用基因变异和目标表型之间关系的信息,通过关联证据,或通过链接有关所述基因组的功能的信息,结合使用。

[0104] 对于理解人类生物学和人类疾病的治疗的特定示例应用而言,产生的结果特别重要,如果不使用本方法,则只能通过直接的体内实验才有可能获得结果,这可能需要侵入性手段。之所以可行,是因为输入数据(当被研究的生物体是人类时)最终将从人类研究参与者那里获得可以在超出其他方法能力的范围内进行分析。

[0105] 在一些实施方案中,所述方法包括针对选定区域内的表型的一个或多个表型集群中的每一个鉴定两个或多个因果基因变异。这涉及到考虑区域内基因变异之间的相关性,通常用“LD矩阵”来概括,在位置 $i, j$ 上基因型 $g_i, g_j$ 的相关性矩阵 $r_{ij}$ ,通常从参考面板,例如1000个基因组联盟或单倍型参考联盟的亚群获得。在此假设下,唯一需要的更改是使用不同的似然计算。所需的适当似然性遵循类似于FINEMAP方法<sup>23</sup>,但适用于本方法的方法。

[0106] 一个这样的实施例将包括第一步,以鉴定可能如上所述的集群中的表型可能有因果关系的第一基因变异(即第一重要基因变异)。在第二步骤中,所述方法除了将每个单独变异可能是表型的潜在的因果基因变异之外,还将考虑由第一步中鉴定的所述第一基因变异,以及所述区域内的其他基因变异构成的变异对,都作为所述基因型的因果基因变异。此方法可通过在后续步骤中包括第三,第四等潜在因果变异(即其他重要基因变异)进一步扩展。这种逐步的方法的优点是它避免了一次考虑所有对(三倍体,四倍体等)的需要,尽管在某些情况下仍然可行,但计算成本会很高。

[0107] 另一个这样的实施例将通过更新概括统计以鉴定其他因果变异来考虑集群内已经鉴定的因果变异的效应,然后评估其他因果变异的剩余证据。通过建议添加或删除最多一个因果变变异,可以迭代地应用该方法来探索因果变异在集群内部或集群之间的空间。在给定一组具有效应大小 $\alpha$ 的因果变异的情况下,此类方法利用观察到的效应大小估计分布的结果:

$$[0108] \quad MVN(\hat{\beta}; SRS^{-1}\alpha, \Sigma)^{24} \quad (1)$$

[0109] 其中S和R是等位基因的方差和相关性,假定等位基因的影响大小估计为 $\hat{\beta}$ ,而 $\Sigma$ 描述这些估计中的协方差。向量 $\alpha$ 的非零元素表示变异被指定为因果的(在基本实施例中,仅假设其中之一为非零)。可以通过减去这些变异的影响来获得关联的残余证据,因此,例如,所述残余效应大小估计为 $\hat{\beta}' = \hat{\beta} - SRS^{-1}\alpha$ 。这些可用于计算每个变异的等效贝叶斯因子。如果确定了新的因果变异i,则很容易获得其效应大小的估计值 $\alpha_i$ 作为最可能的估计值,或者通过从其后验分布中采样而容易地获得 $\alpha_i$ ,这两种方法都可以包含上述的效应大小的先验信息(例如,对于均值为0,方差为 $\sigma^2$ ,通过假设真实效果大小为正态分布)。在评估研究的集群分配时,可以仅针对向量 $\alpha$ 的非零元素计算 $MVN(\hat{\beta}; SRS^{-1}\alpha, \Sigma)$ 密度。

[0110] 如果对于每个GAS,概括统计的信息性不尽相同,则可以使用其他替代实施例。发生这种情况的原因可能多种多样,包括:(i) 由于基因分型阶段出现问题而偶尔丢失数据;(ii) 有必要推定某些GAS在特定基因分型阶段未测量或未充分测量某些基因变异;(iii) 由于局部基因变异和附近的“标签”基因变异之间连锁不平衡(LD)水平的波动,整个基因组的填补质量(imputationquality)也有所不同。在极端的情况下,甚至没有在GAS j中观察到特定的基因变异i或估算出该基因变异i,即使GAS j实际上是由变异i驱动的并且连锁基因变异的概括统计与此一致,也不会选择该GAS参与以重要变异i为特征的集群。在这种情况下,可以通过对一种或多种基因变异与对应于输入单位的表型之间的关联进行建模,来将关于基因变异与输入单位的表型之间的其他关联信息添加至一个或多个所述输入单位中。例如,可以在基因变异i是因果变异的假设下而不是在上述示例实施例中所处理的 $|\beta_i| > 0$ 的证据强度的假设下,对研究j的预期 $\beta$ 进行建模。这是通过在假定所述基因变异i为因果变异的情况下,对向量 $\hat{\beta}$ ,取值 $\hat{\sigma}^2$ ,和LD矩阵 $\Sigma$ 建模来实现的。

[0111] 在这样的解决差异信息的实施例中,对于正被测试因果关系的给定基因变异i,将需要鉴定对变异i的基因型状态最有信息的另一变异j。可以将其量化为所述模型的预期对数贝叶斯因子,其中i是因果关系模型,而不是i不是因果关系模型,其中期望取值于因果关系模型下观察到的统计值,并且该模型描述了观察过程不完整。对于观察模型的特定选择,

在这种情况下,这种期望首先是 $pr^2$ 的函数,其中 $p$ 是观察到的基因型的分数,而 $r^2$ 是LD矩阵 $\Sigma$ 的 $i, j$ 项,这样最大化 $pr^2$ 的方式可以最大化基因变异 $j$ 提供的有关基因变异 $i$ 的基因型的信息。

[0112] 另一个这样的实施例将通过首先基于关于效应大小 $\hat{\beta}$ 和/或它们的测量精度的真实相关性的信息通过调整GAS概括统计数据来应用似然函数。可以通过在原始研究的理想版本: $\beta^*$ 和 $SE^*$ 中,针对观测数据,对观察数据、GAS概括统计、真实效应大小及其不确定性定义联合模型来获得调整后的概括统计。联合分布可以通过多元正态分布来近似,从中可以估计出原始研究中真实效应的大小。然后这些可以在似然函数中替代原始的 $\hat{\beta}$ ,  $SE$ 。一种方法是通过等式(1)中的所述矩阵 $\Sigma = UDU'$  (因为 $\Sigma$ 是对称的)进行奇异值分解来获得具有对角协方差矩阵: $MVN(D^{-1/2}U'\hat{\beta}; D^{-1/2}U'SRS^{-1}\alpha, I)$ 的多元正态密度。这是具有独立观测值和单位方差的标准回归形式,其中 $Y = D^{-1/2}U'\hat{\beta}$ 和 $X = D^{-1/2}U'SRS^{-1}$ ,因此可以使用普通最小二乘估计来获得对于变量 $i$ ,  $\beta_i^* = (X_i'X_i)^{-1}X_i'Y$ 和 $SE_i^* = 1/(X_i'X_i)^{-1/2}$ 的每个变异的边际调整后的概括统计。可以使用矩阵运算对所有变异进行有效地计算。

[0113] 另一个替代实施例优化了基因组分割以形成选定区域。可以通过将基因组与固定数量的不重叠的对接块或部分重叠的块进行平铺来创建选定区域,以使块之间的基因变异最小化。最小化可以通过动态编程算法来实现。优化分段的另一种方法将分段过程视为“多个更改点过程”,并使用动态编程算法来有效地考虑所有或大量可能的分段。优选的实施方式使用基于粒子滤波的算法来实现该动态编程算法为具有支撑状态稀疏选择的隐马尔科夫模型,例如如Fearnhead等人<sup>25</sup>所述的。这样的算法需要给定特定分割的观察的似然性,可以针对其使用本公开的实施例。

[0114] 在另一个替代实施例中,可以针对多个选定区域重复所述方法,所述基因的不同选定区域的集群组的鉴定可以进一步通过评估每个集群的表型组和跨集群组的其他集群的表型组之间的相似性而进行分析。进行此类分析可以为共享生物学提供强大的见解。例如,表现出与多个选定区域中的目标疾病或病状相关的集群子集可以进一步分析以寻找共享的和独特的致病机制。该全基因组聚类分析的结果可被利用以进一步提高上述实施例的统计能力,通过将共同出现在聚类中某些多个表型的频率作为先验,通过其这些表型被提供给集群。这可以通过多种方式来实现。一种方法是使用Ising模型,为选定区域 $r$ 和给定聚类 $C$ 定义一个“聚类成员”向量 $m_r$ ,当研究是非空聚类 $C$ 的一部分时,坐标为1;如果在选定区域 $r$ ,它是空集群或任何其他集群的一部分,则坐标为0,然后可以通过包含一个项来对GAS共聚进行建模

$$[0115] \quad L(k) = \frac{1}{Z(J, h)} \exp \left[ \sum_{j \neq i} J_{ij} m_{ri} m_{rj} + \sum_i h_i m_{ri} \right]$$

[0116] 在模型中,其中, $J$ 和 $h$ 是建模参数, $Z(J, h)$ 是归一化常数,其值难以计算,但更新方程式不需要。由于存在线性项 $\sum_i h_i m_{ri}$ ,其模拟了GAS参与选定区域 $r$ 中的任何集群的边际概率,该项分别替代了空集群和非空集群的关系的因子 $p$ 和 $1-p$ 。可替代地,由于所述线性项 $\sum_i h_i m_{ri}$ 对研究 $i$ 参与集群的边际概率进行建模,因此可以如标准实施例中那样忽略该项并且使用因子 $p$ 和 $1-p$ 。所述建模参数 $J$ 和 $h$ 是通过在推断的集群成员向量 $m$ 上优化总集群似然

性 $Z(J, h)$ 来找到的。

[0117] 在另一个实施例中,在计算集群的特征时使用的集群 $s$ 中变异 $i$ 的荟萃分析函数 $meta_{is}$ 不假定研究是独立的,而是考虑了研究之间的相关性。该实施例既可以考虑GAS(或表型)之间的效应大小估计中的隐性相关性,又可以做出关于研究之间真实效应的大小和相关性的特定假设。由于对一组重叠的个体执行了不同表型的关联GAS,因此研究之间可能会产生隐性的相关性。在这种情况下,评估关联证据的自然方法是使用多变量正态分布对真实效果和采样误差进行建模,以获得多变量边际似然性(例如,参见Band等人的补充方法<sup>26</sup>)的形式

[0118]  $MVN(\hat{\beta}; 0, \Sigma + V)$ .

[0119] 所述MVN函数是多元正态密度,在 $\hat{\beta}_{is}$ 处评估,是观察到的效应大小估计的向量,平均向量为 $0s$ ,并且方差-协方差由 $V = (SE_{is}) R_{\text{隐}} (SE_{is})^T$ 和 $\Sigma = \sigma R_{\text{真}}$ 来确定。 $R_{\text{真}}$ 是真实效应中的假定相关性,当假定效应的大小和方向不相关时,它可能是一个单位矩阵。 $R_{\text{隐}}$ 是GAS之间beta估计中不确定性的假定相关性,可以从每对研究之间的beta估计中的经验相关性进行估计(理想情况下,但不一定是全基因组范围内)。如上所述, $\sigma$ 是对真实效应的先验分布的方差,在无关联的无效假设下, $\sigma = 0$ 。然后,基于边际似然性比率的贝叶斯因子对与变异相关联的证据进行荟萃分析

[0120]  $meta_{is} = \frac{MVN(\hat{\beta}; 0, \Sigma + V)}{MVN(\hat{\beta}; 0, V)}$

[0121] 注意,当 $\Sigma$ 和 $V$ 对角线,则研究都被认为是独立的,并且在示例实施例恢复 $meta_{is}$ 。一种计算有效的近似方法涉及在独立性假设下进行聚类,然后执行多元正态元分析作为进行集群分配的最后一步。替代的计算有效实施例涉及更新矩阵 $\Sigma^{-1}$ 和 $(V + \Sigma)^{-1}$ ,以及他们的决定因子,其需要使用Sherman-Morrison公式Matrix Determinant Lemma来计算 $meta_{is}$ 。

[0122] 参考引用

[0123] 1.Nat Rev Genet.2013Jul;14(7):483-95

[0124] 2.Open Biol.2017Nov;7(11).pii:170125

[0125] 3.PLoS Genet.2014Nov 13;10(11):e1004787

[0126] 4.Nat Genet.2018Feb;50(2):229-237

[0127] 5.Bioinformatics.2016Jun 15;32(12):1856-64

[0128] 6.Biostatistics.kxx048,https://doi.org/10.1093/biostatistics/kxx048  
Preprint published online:2017Sep 25

[0129] 7.Bioinformatics.2017Jan 1;33(1):64-71

[0130] 8.Nat Genet.2016May;48(5):481-7

[0131] 9.Nature Genetics volume 48,pages 803810(2016)

[0132] 10.PLoS Genet.2014May 15;10(5):e1004383

[0133] 11.Nat Genet.2016Jul;48(7):709-17

[0134] 12.Am J Hum Genet.2016Dec 1;99(6):1245-1260

[0135] 13.PLoS Genet.2017Mar 9;13(3):e1006646

[0136] 14.Nat Genet.2017Apr;49(4):600-605

- [0137] 15.doi:<https://doi.org/10.1101/059345>Preprint published online 2016Nov 3;<https://yueli-compbio.github.io/RiVIERA/>
- [0138] 16.doi:<https://doi.org/10.1101/155481>Preprint published online 2017Oct 21
- [0139] 17.奥尔德斯 • D. • J. (1985). "Exchangeability and related topics". *École d'Été de Probabilités de Saint-Flour XIII-1983. Lecture Notes in Mathematics*. 1117. pp.1-1. doi:10.1007/BFb0099421. ISBN 978-3-540-15203-3.
- [0140] 18. *Nat Genet*. 2012Dec; 44(12):1294-1301.
- [0141] 19. *Gen. Epid*. 33:79-86 2009
- [0142] 20. *PLoS Genet* 10(10):e1004722
- [0143] 21. *Nat Genet*. 2012Feb; 44(3):328333.
- [0144] 22. *Circ Res*. 2016Jul; 119(3):450-62
- [0145] 23. 本纳 • C等., FINEMAP, *Bioinformatics* 2016
- [0146] 24. *Ann. Appl. Stat. Vol. 11, Number 3 (2017)*, 1561-1592 (<https://projecteuclid.org/euclid.aoas/1507168840>)
- [0147] 25. *P. Stat Comput* (2006) 16:203
- [0148] 26. *PLoS Genet* 9(5):e1003509

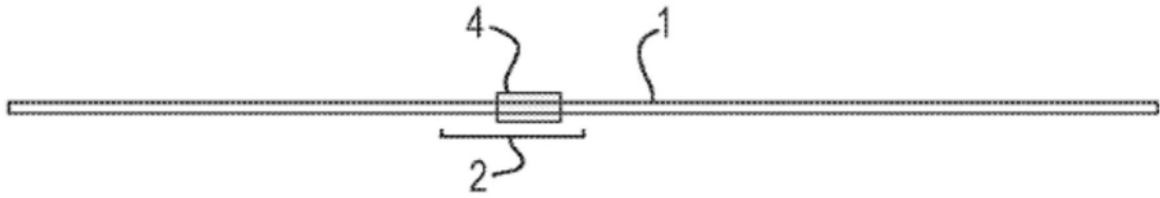


图1

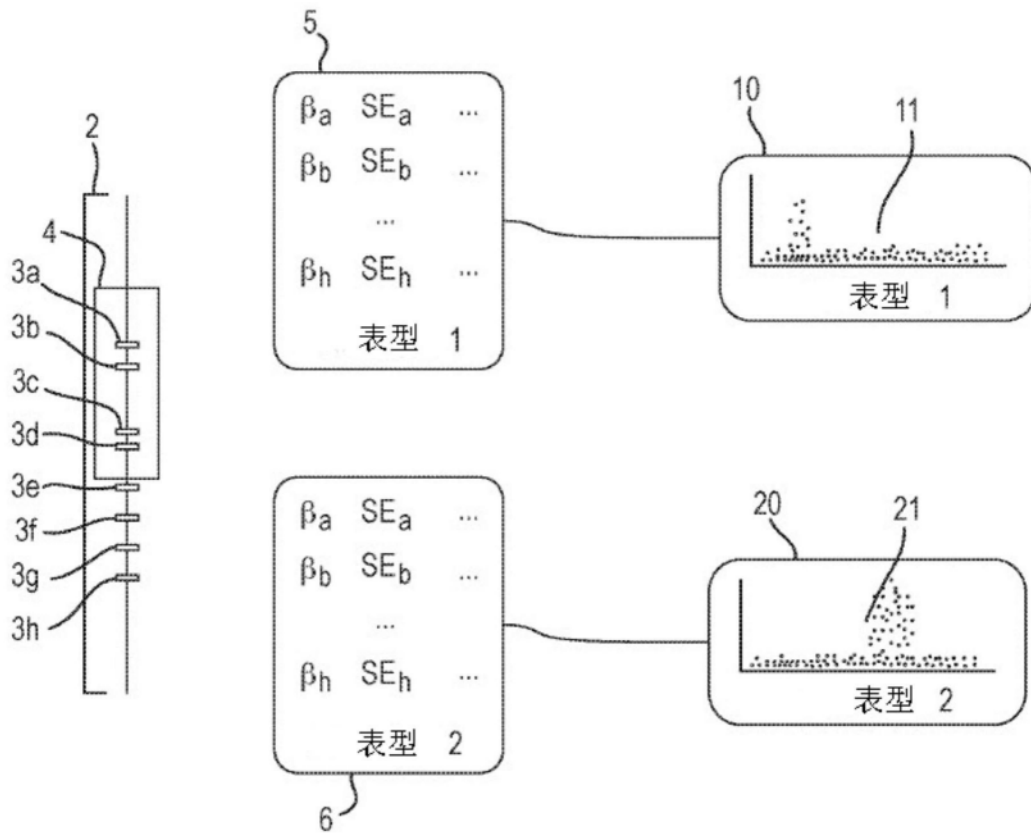


图2

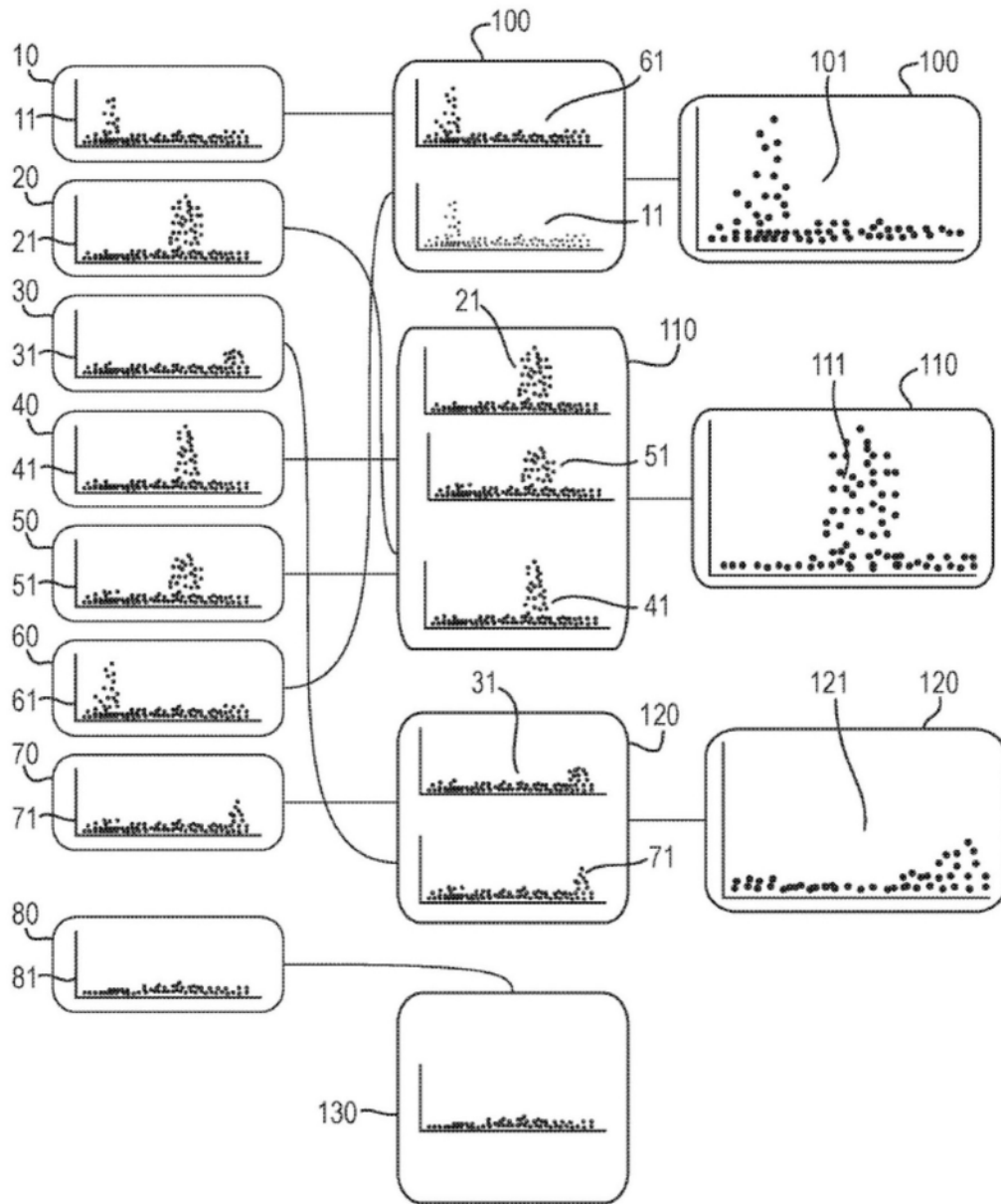


图3

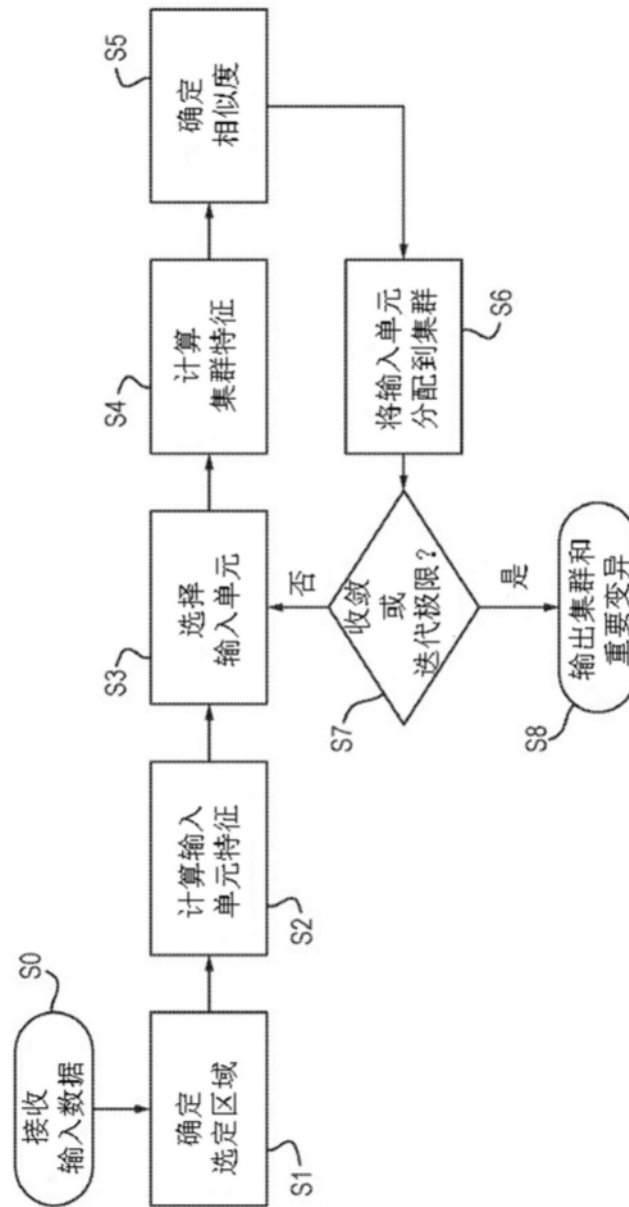


图4

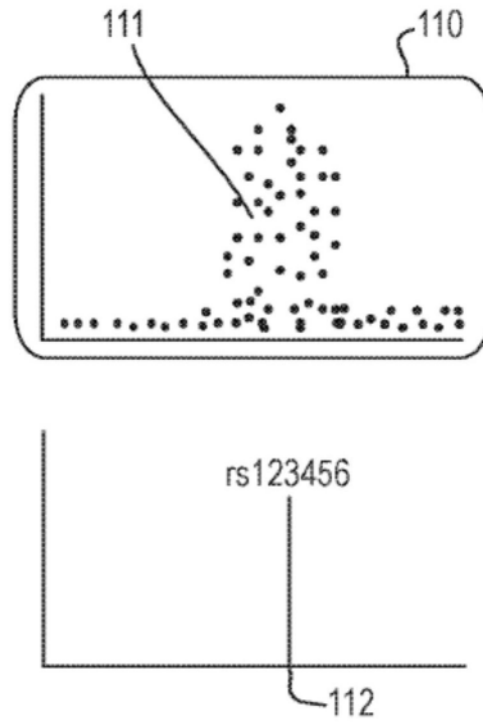


图5

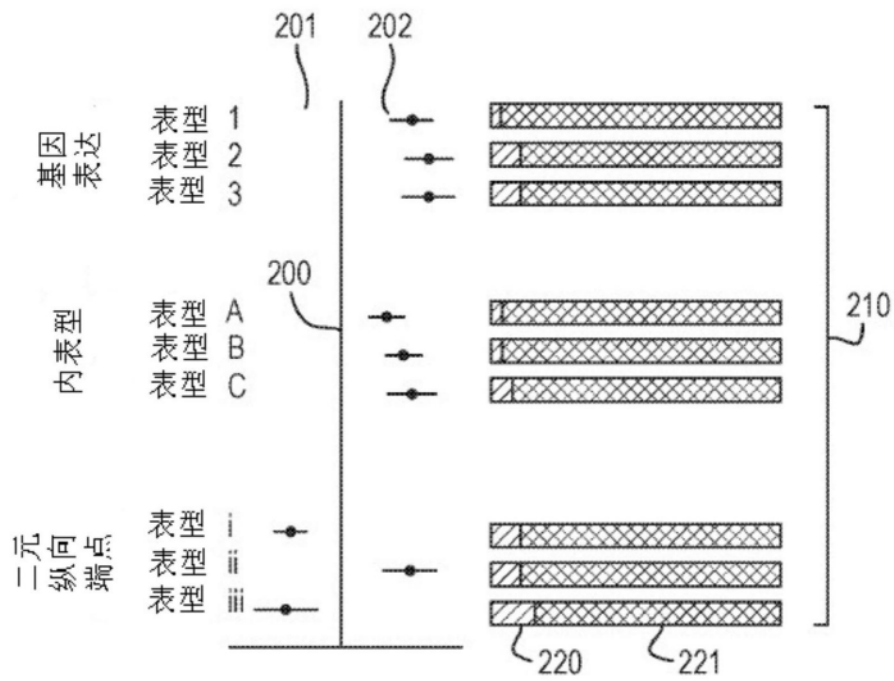


图6

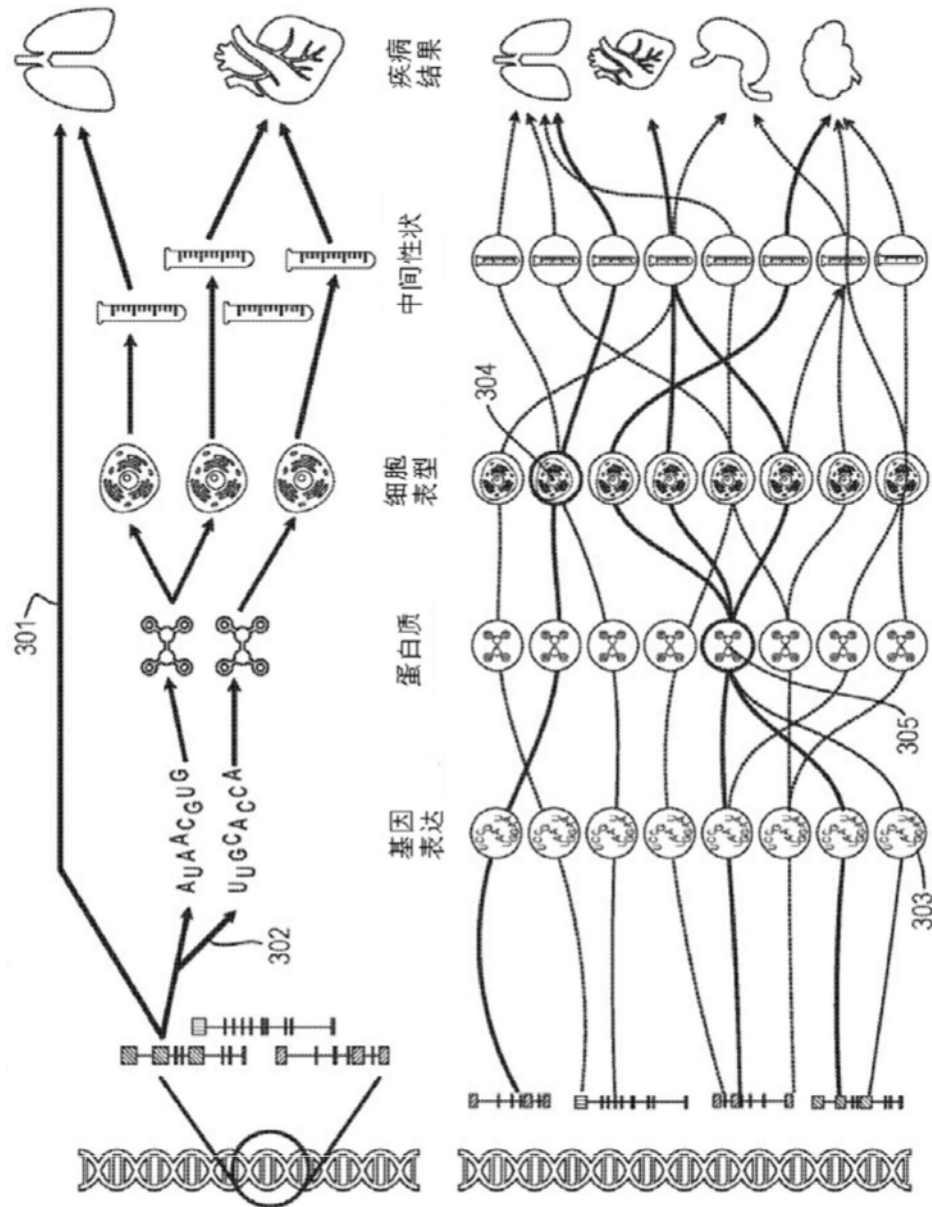


图7

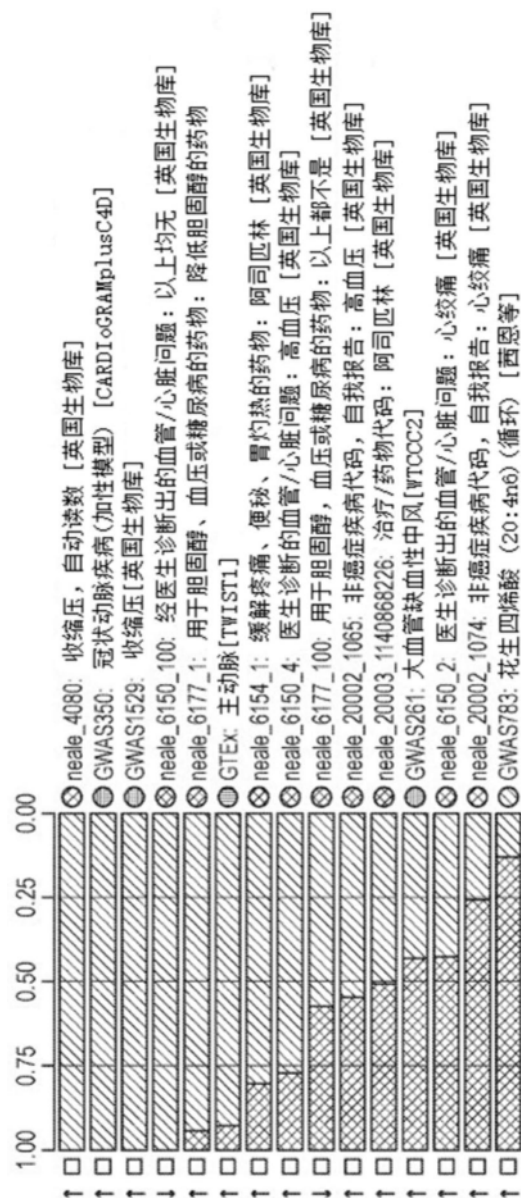


图8a

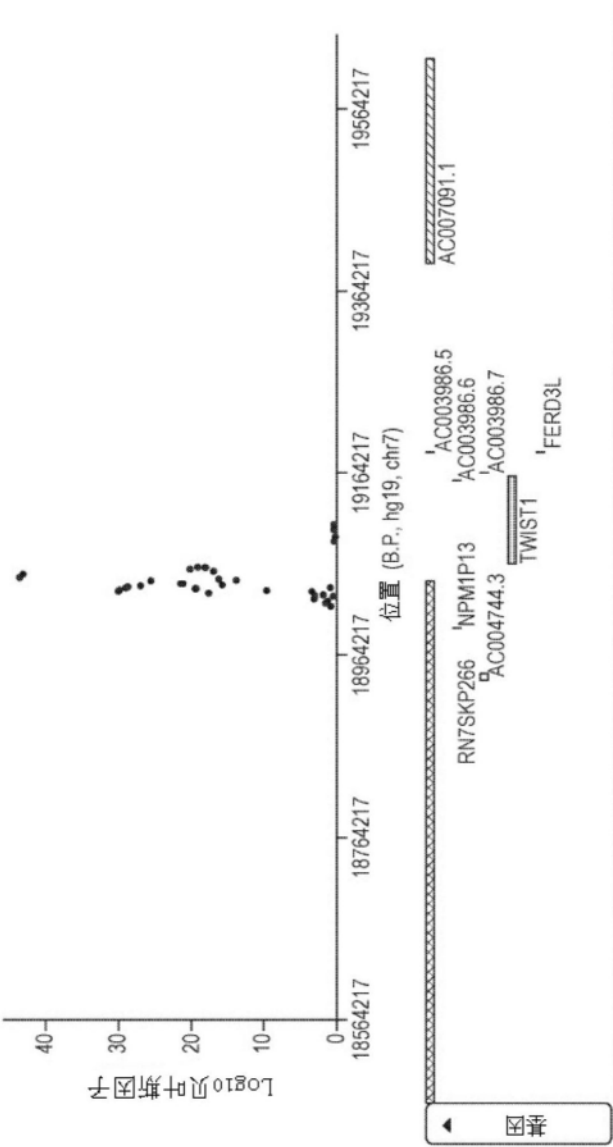


图8b

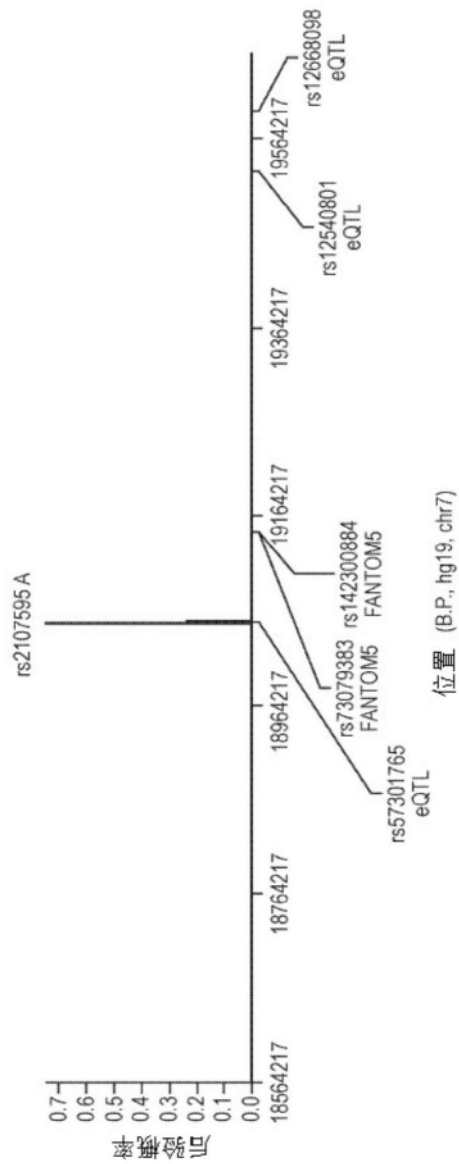


图8c

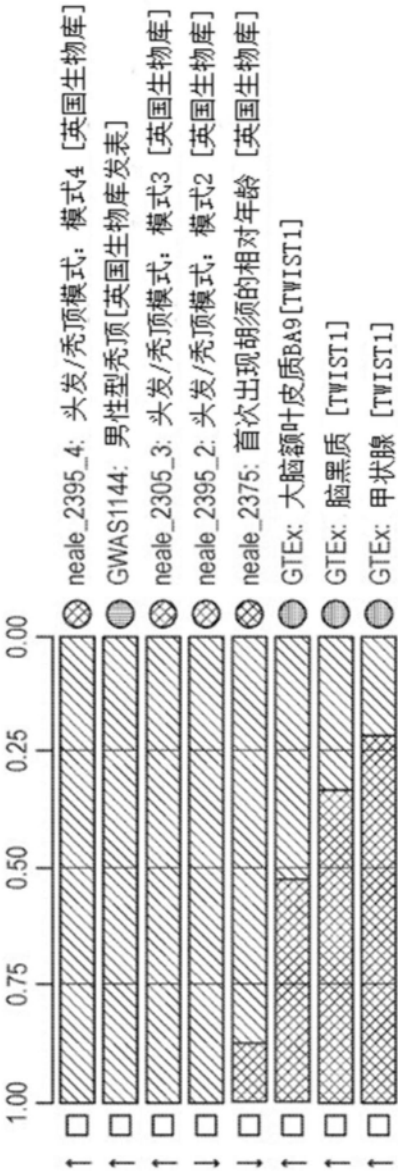


图8d

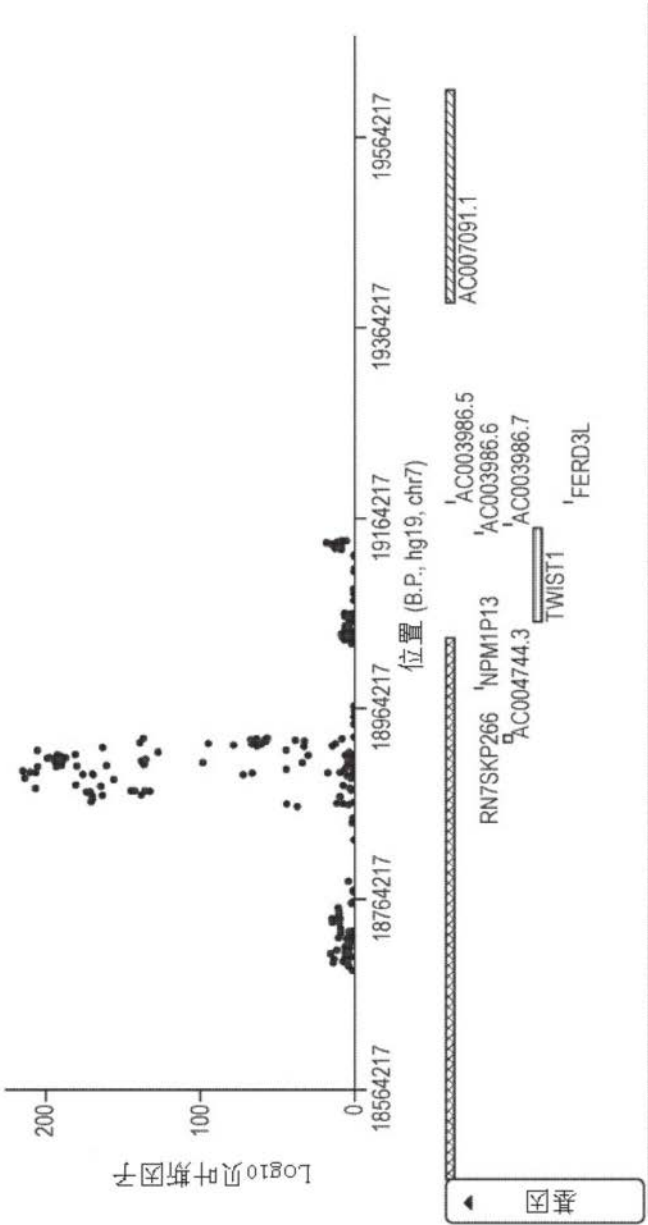


图8e

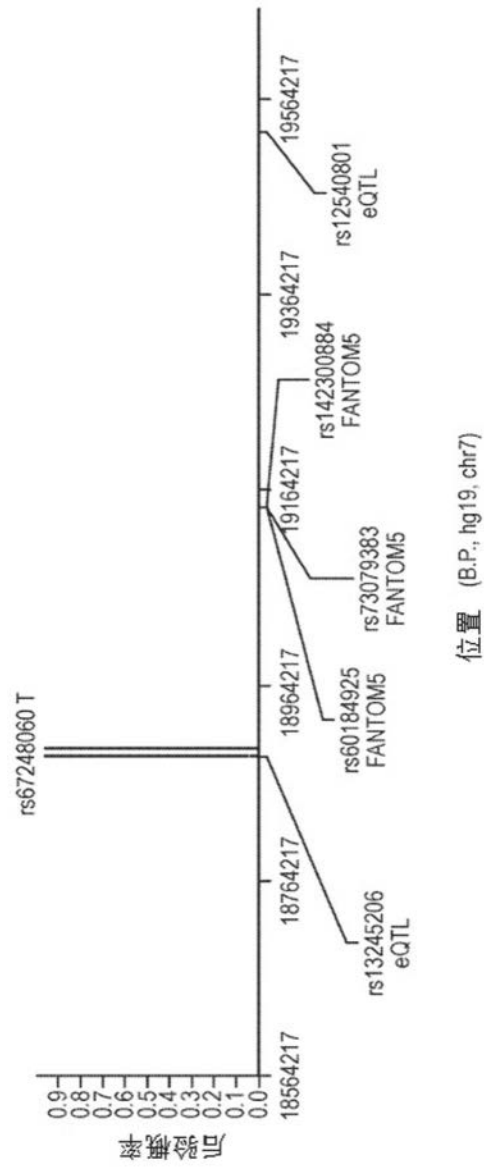


图8f