(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109643584 A
(43)申请公布日 2019.04.16

(21)申请号 201780050613.8

(22)申请日 2017.05.30

(30)优先权数据
62/394,551 2016.09.14 US

(85)PCT国际申请进入国家阶段日
2019.02.19

(86)PCT国际申请的申请数据
PCT/EP2017/063073 2017.05.30

(87)PCT国际申请的公布数据
WO2018/050299 EN 2018.03.22

(71)申请人 菲利普莫里斯生产公司
地址 瑞士纳沙泰尔

(72)发明人 C·普森 V·贝尔卡斯特罗 F·马丁 S·布韦 M·C·派奇

(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 宋岩

(51)Int.Cl.
G16H 70/60(2018.01)
G16B 20/00(2019.01)
G16B 40/00(2019.01)
C12Q 1/68(2018.01)

权利要求书6页 说明书28页 附图19页

(54)发明名称

用于预测个体生物状态的系统、方法和基因标签

(57)摘要

用于评定受试对象的样本以预测所述受试对象的生物状态的系统和方法，所述生物状态例如吸烟者状态。所述计算机实施的方法包含通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集。所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5。所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分，其中所述得分基于少于40个基因，且指示所述受试对象的预测吸烟状态。

CN 109643584 A

1.一种用于评定从受试对象获得的样本的计算机实施的方法,其包括:

通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集,所述数据集包括小于全基因组的一组基因的定量表达数据,所述一组基因包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5;以及

通过所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分,其中所述得分基于少于40个基因,且指示所述受试对象的预测吸烟状态。

2.根据权利要求1所述的计算机实施的方法,其中所述一组基因还包括AK8、FSTL1、RGL1和VSIG4。

3.根据权利要求1到2中任一项所述的计算机实施的方法,其中所述一组基因还包括C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。

4.根据权利要求1到3中任一项所述的计算机实施的方法,其中所述得分是应用于所述数据集的分类方案的结果,其中所述分类方案基于所述数据集中的所述定量表达数据而确定。

5.根据权利要求1到4中任一项所述的计算机实施的方法,其还包括计算AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5中的每一者的倍数变化值。

6.根据权利要求5所述的计算机实施的方法,其还包括确定每个倍数变化值满足至少一个准则,所述准则要求对于至少两个独立群体数据集,每个相应的所计算倍数变化值超过预定阈值。

7.根据权利要求1所述的计算机实施的方法,其中所述一组基因由AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5组成。

8.一种包括计算机可读指令的计算机程序产品,所述计算机可读指令在包括至少一个处理器的计算机化系统中执行时使所述处理器执行根据权利要求1到7中任一项所述的方法的一个或多个步骤。

9.一种用于预测个体的吸烟者状态的试剂盒,其包括:

一组试剂,其检测具有少于40个基因的基因标签中的基因的表达水平,所述基因标签包括测试样本中的AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5;以及

使用所述试剂盒预测所述个体的吸烟者状态的说明书。

10.根据权利要求9所述的试剂盒,其中所述试剂盒用于评定吸烟产品的替代物对个体的作用。

11.根据权利要求10所述的试剂盒,其中所述吸烟产品的所述替代物是加热式烟草产品。

12.根据权利要求9到11中任一项所述的试剂盒,其中所述替代物对所述个体的所述作用是将所述个体归类为非吸烟者。

13.根据权利要求9到12所述的试剂盒,其中所述基因标签还包括AK8、FSTL1、RGL1和VSIG4。

14.根据权利要求9到13中任一项所述的试剂盒,其中所述基因标签还包括C15orf54、

CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。

15.一种用于评定从受试对象获得的样本的计算机实施的方法，其包括：

通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集，所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63；以及

通过所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分，其中所述得分基于少于40个基因，且指示所述受试对象的预测吸烟状态。

16.根据权利要求15所述的计算机实施的方法，其中所述得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的所述定量表达数据而确定。

17.根据权利要求15到16中任一项所述的计算机实施的方法，其还包括计算LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63中的每一者的倍数变化值。

18.根据权利要求17所述的计算机实施的方法，其还包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

19.根据权利要求15所述的计算机实施的方法，其中所述一组基因由LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63组成。

20.一种包括计算机可读指令的计算机程序产品，所述计算机可读指令在包括至少一个处理器的计算机化系统中执行时使所述处理器执行根据权利要求15到19中任一项所述的方法的一个或多个步骤。

21.一种用于预测个体的吸烟者状态的试剂盒，其包括：

一组试剂，其检测具有少于40个基因的基因标签中的基因的表达水平，所述基因标签包括测试样本中的LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63；以及

使用所述试剂盒预测所述个体的吸烟者状态的说明书。

22.根据权利要求21所述的试剂盒，其中所述试剂盒用于评定吸烟产品的替代物对个体的作用。

23.根据权利要求22所述的试剂盒，其中所述吸烟产品的所述替代物是加热式烟草产品。

24.根据权利要求21到23中任一项所述的试剂盒，其中所述替代物对所述个体的所述作用是将所述个体归类为非吸烟者。

25.一种用于获得用于预测生物状态的基因标签的计算机实施的方法，所述方法包括：

由计算机系统通过网络将训练数据集提供到多个用户装置，所述计算机系统包含通信端口和与至少一个非暂时性计算机可读媒体通信的至少一个计算机处理器，所述非暂时性计算机可读媒体存储包括所述训练数据集和测试数据集的至少一个电子数据库，其中：

所述训练数据集包含一组训练样本，且所述测试数据集包含一组测试样本，其中每个

训练样本和每个测试样本包含基因表达数据，且对应于具有选自一组生物状态的已知生物状态的患者；

从所述网络接收候选基因标签，所述候选基因标签各自通过基于所述训练数据集获得分类器而产生，其中每个候选基因标签包含被确定能判别所述训练数据集中的不同生物状态的一组基因；

基于相应候选基因标签预测所述测试样本的所述已知生物状态的表现，将得分指派给每个相应候选基因标签；

基于指派的所述得分，识别所述候选基因标签的子集；

在所述子集中识别包含在至少阈值数目的候选基因标签中的基因；以及

将所识别的所述基因存储为所述基因标签。

26.根据权利要求25所述的方法，其还包括将表示每个候选基因标签中允许的最大阈值数目的基因的数目提供到所述多个用户装置。

27.根据权利要求25或26所述的方法，其还包括通过所述网络将所述测试数据集的部分提供到所述多个用户装置，其中所述测试数据集的所述部分包含具有已知生物状态的患者的基因表达数据且不包含所述患者的所述已知生物状态。

28.根据权利要求27所述的方法，其还包括针对每个候选基因标签，接收所述测试数据集中的每个样本的置信水平。

29.根据权利要求28所述的方法，其中所述置信水平是指示所述测试数据集中的样本属于所述生物状态中的一个的预测可能性的值。

30.根据权利要求28或29所述的方法，其中所述得分至少部分地基于所述置信水平。

31.根据权利要求30所述的方法，其中所述得分至少部分地基于根据所述置信水平和所述测试数据集中的患者的所述已知生物状态而计算的精确度查全率下面积(AUPR)度量。

32.根据权利要求25到31中任一项所述的方法，其中所述得分至少部分地基于对应的候选基因标签是否会提供与所述测试数据集中的患者的所述已知生物状态一致的预测。

33.根据权利要求32所述的方法，其中使用马修斯相关系数(MCC)确定所述对应候选基因标签是否提供与所述测试数据集中的患者的所述已知生物状态一致的预测。

34.根据权利要求25到33中任一项所述的方法，其中所述候选基因标签根据至少两个不同度量进行排序以获得每个候选基因标签的第一排序和第二排序。

35.根据权利要求34所述的方法，其中将每个候选基因标签的所述第一排序和所述第二排序进行平均以获得每个相应候选基因标签的所述得分。

36.根据权利要求25到35中任一项所述的方法，其中所述一组生物状态包含吸烟者状态。

37.根据权利要求36所述的方法，其中所述吸烟者状态包含当前吸烟者和非吸烟者。

38.根据权利要求25到37中任一项所述的方法，其中所述基因标签小于全基因组，且包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5。

39.根据权利要求38所述的方法，其中所述基因标签还包括AK8、FSTL1、RGL1和VSIG4。

40.根据权利要求39所述的方法，其中所述基因标签还包括C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。

41.根据权利要求40所述的方法，其中所述基因标签还包括ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618。

42.根据权利要求25到37中任一项所述的方法，其中所述基因标签小于全基因组，且包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63。

43.根据权利要求42所述的方法，其中所述基因标签还包括DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1和GUCY1B3。

44.根据权利要求25到37中任一项所述的方法，其中所述基因标签小于全基因组，且包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21。

45.一种包括计算机可读指令的计算机程序产品，所述计算机可读指令在包括至少一个处理器的计算机化系统中执行时使所述处理器执行根据权利要求25到44中任一项所述的方法的一个或多个步骤。

46.一种用于评定从受试对象获得的样本的计算机实施的方法，其包括：

通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集，所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618；以及

通过所述至少一个硬件处理器基于接收到的所述数据集而产生得分，其中所述得分指示所述受试对象的预测吸烟状态。

47.根据权利要求46所述的计算机实施的方法，其中所述得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的所述定量表达数据而确定。

48.根据权利要求46到47中任一项所述的计算机实施的方法，其还包括计算AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618中的每一者的倍数变化值。

49.根据权利要求48所述的计算机实施的方法，其还包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

50.根据权利要求46到49中任一项所述的计算机实施的方法，其中所述一组基因由AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、

MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618组成。

51.一种包括计算机可读指令的计算机程序产品，所述计算机可读指令在包括至少一个处理器的计算机化系统中执行时使所述处理器执行根据权利要求46到50中任一项所述的方法的一个或多个步骤。

52.一种用于预测个体的吸烟者状态的试剂盒，其包括：

一组试剂，其检测测试样本中的基因标签中的基因的表达水平，所述基因标签包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618；以及

使用所述试剂盒预测所述个体的吸烟者状态的说明书。

53.根据权利要求52所述的试剂盒，其中所述试剂盒用于评定吸烟产品的替代物对个体的作用。

54.根据权利要求53所述的试剂盒，其中所述吸烟产品的所述替代物是加热式烟草产品。

55.根据权利要求52到54中任一项所述的试剂盒，其中所述替代物对所述个体的所述作用是将所述个体归类为非吸烟者。

56.一种用于评定从受试对象获得的样本的计算机实施的方法，其包括：

通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集，所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21；以及

通过所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分，其中所述得分基于少于40个基因，且指示所述受试对象的预测吸烟状态。

57.根据权利要求56所述的计算机实施的方法，其中所述得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的所述定量表达数据而确定。

58.根据权利要求56到57中任一项所述的计算机实施的方法，其还包括计算AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21中的每一者的倍数变化值。

59.根据权利要求58所述的计算机实施的方法，其还包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

60.根据权利要求56所述的计算机实施的方法，其中所述一组基因由AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21组成。

61.一种包括计算机可读指令的计算机程序产品，所述计算机可读指令在包括至少一个处理器的计算机化系统中执行时使所述处理器执行根据权利要求56到60中任一项所述

的方法的一个或多个步骤。

62.一种用于预测个体的吸烟者状态的试剂盒,其包括:

一组试剂,其检测测试样本中的基因标签中的基因的表达水平,所述基因标签包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21,所述基因标签包括少于40个基因;以及

使用所述试剂盒预测所述个体的吸烟者状态的说明书。

63.根据权利要求62所述的试剂盒,其中所述试剂盒用于评定吸烟产品的替代物对个体的作用。

64.根据权利要求63所述的试剂盒,其中所述吸烟产品的所述替代物是加热式烟草产品。

65.根据权利要求63到64中任一项所述的试剂盒,其中所述替代物对所述个体的所述作用是将所述个体归类为非吸烟者。

用于预测个体生物状态的系统、方法和基因标签

[0001] 相关申请的引用

[0002] 本申请根据35U.S.C.§119要求2016年9月14日提交的第62/394,551号美国临时专利申请的优先权,所述美国临时专利申请以全文引用的方式并入本文中。本申请涉及2014年12月11日提交的第PCT/EP2014/077473号PCT申请以及2014年8月12日提交的第PCT/EP2014/067276号PCT申请,每个PCT申请以全文引用的方式并入本文中。

背景技术

[0003] 人们不断暴露于可能触发有害分子变化的外部有毒物质(例如,香烟烟雾、杀虫剂)。21世纪毒理学的风险评估依赖于毒性机制的阐述以及来自高通量数据的暴露反应标志物的识别。全基因组微阵列等新技术已被纳入毒性测试中,以提高效率并提供更依据数据处理的暴露反应评估方法。对转录基因调节的基因组规模推断随着微阵列和RNA测序等高通量技术的出现而已成为可能,因为这些技术在许多测试实验条件下提供转录组的快照。

[0004] 生物医学研究群体一般对寻找用于疾病诊断的稳健标签感兴趣。一些证据表明,疾病的分子分类可能比形态分类更准确。然而,从主要暴露部位(例如在烟雾或空气污染物暴露情况下,呼吸道)进行样本采集通常是侵入性的,因此不便于进行暴露评估和监测。作为一种微创替代方案,可在一般人群中采用外周血取样以建立全身性生物标志物。血液因其含有的许多不同细胞亚群而分析起来较复杂。然而,它是调查标志物识别的高度相关组织,因为血液在更直接暴露于有毒物质的所有器官中循环,且血液易于获取。此外,即使未见组织学异常,也可检测到烟雾暴露的分子反应。

发明内容

[0005] 提供计算系统和方法以使用众包方法来识别基于血液的稳健基因标签,所述基因标签可用于预测个体的吸烟者状态。本文所描述的基因标签能够区分当前吸烟的受试对象与从不吸烟的受试对象,从而能够准确预测个体的吸烟者状态。

[0006] 在某些方面,本公开的系统和方法提供用于评定从受试对象获得的样本的计算机实施的方法。所述计算机实施的方法包含通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集。所述数据集包括小于全基因组的一组基因的定量表达数据,所述一组基因包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5。所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分,其中所述得分基于少于40个基因,且指示所述受试对象的预测吸烟状态。

[0007] 在某些实施方案中,所述一组基因还包括AK8、FSTL1、RGL1和VSIG4。在某些实施方案中,所述一组基因还包括C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。

[0008] 在某些实施方案中,所述得分是应用于所述数据集的分类方案的结果,其中所述

分类方案基于所述数据集中的定量表达数据而确定。在某些实施方案中，所述计算机实施的方法还包括计算AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5中的每一者的倍数变化值。所述计算机实施的方法还可包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

[0009] 在某些实施方案中，所述一组基因由AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5组成。

[0010] 在某些方面，本公开的系统和方法提供一种用于预测个体的吸烟者状态的试剂盒。所述试剂盒包含：一组检测具有少于40个基因的基因标签中的基因的表达水平的试剂，所述基因标签包括测试样本中的AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5；以及使用所述试剂盒来预测个体的吸烟者状态的说明书。

[0011] 在某些实施方案中，所述试剂盒用于评定吸烟产品的替代物对个体的作用。所述吸烟产品替代物可包含加热式烟草产品。所述替代物对个体的作用可以是将个体归类为非吸烟者。在某些实施方案中，所述基因标签还包括AK8、FSTL1、RGL1和VSIG4。在某些实施方案中，所述基因标签还包括C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。

[0012] 在某些方面，本公开的系统和方法提供用于评定从受试对象获得的样本的计算机实施的方法。所述计算机实施的方法包括通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集，所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63。所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分，其中所述得分基于少于40个基因，且指示所述受试对象的预测吸烟状态。

[0013] 在某些实施方案中，所述得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的定量表达数据而确定。

[0014] 在某些实施方案中，所述至少一个硬件处理器计算LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63中的每一者的倍数变化值。所述计算机实施的方法还可包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

[0015] 在某些实施方案中，所述一组基因由LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63组成。

[0016] 在某些方面，本公开的系统和方法提供一种用于预测个体的吸烟者状态的试剂盒。所述试剂盒包括：一组试剂，其检测具有少于40个基因的基因标签中的基因的表达水平，所述基因标签包括测试样本中的LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63；以及使用所述试剂盒来预测个体的吸烟者状态的说明书。

[0017] 在某些实施方案中，所述试剂盒用于评定吸烟产品的替代物对个体的作用。所述吸烟产品替代物可包含加热式烟草产品。所述替代物对个体的作用可以是将个体归类为非

吸烟者。

[0018] 在某些方面，本公开的系统和方法提供用于获得预测生物状态的基因标签的计算机实施的方法。所述计算机实施的方法包括通过计算机系统将训练数据集通过网络提供到多个用户装置，所述计算机系统包含通信端口和与至少一个非暂时性计算机可读媒体通信的至少一个计算机处理器，所述至少一个非暂时性计算机可读媒体存储包括所述训练数据集和测试数据集的至少一个电子数据库。所述训练数据集包含一组训练样本，且所述测试数据集包含一组测试样本。每个训练样本和每个测试样本包含基因表达数据，且对应于具有选自一组生物状态的已知生物状态的患者。所述计算机实施的方法还包括从网络接收候选基因标签，所述候选基因标签各自通过基于所述训练数据集获得分类器而产生，其中每个候选基因标签包含被确定能判别所述训练数据集中的不同生物状态的一组基因。基于相应候选基因标签对测试样本的已知生物状态的预测性能，将得分指派给每个相应候选基因标签。基于指派的得分，识别候选基因标签的子组(或候选基因标签的一部分，可包含整组的候选基因标签)，且在所述子组中识别出至少包含在阈值数目的候选基因标签中的基因。所识别基因存储为基因标签。

[0019] 在某些实施方案中，所述计算机实施的方法还包括向多个用户装置提供表示每个候选基因标签中允许的最大阈值数目的基因的数目。

[0020] 在某些实施方案中，所述计算机实施的方法还包括通过网络将测试数据集的一部分提供到多个用户装置，其中所述测试数据集的所述部分包含用于具有已知生物状态的患者的基因表达数据，且不包含患者的已知生物状态。所述计算机实施的方法还可包括针对每个候选基因标签，接收所述测试数据集中的每个样本的置信水平。所述置信水平可以是指示所述测试数据集中的样本属于所述生物状态中的一个的预测可能性的值。所述得分可至少部分地基于所述置信水平。具体地说，所述得分可至少部分地基于根据置信水平和所述测试数据集中的患者已知生物状态所计算的精确度查全率下面积(AUPR)度量。

[0021] 在某些实施方案中，所述得分至少部分地基于对应的候选基因标签是否会提供与所述测试数据集中的患者的已知生物状态一致的预测。可使用马修斯相关系数(Mathews correlation coefficient，MCC)确定对应的候选基因标签是否会提供与测试数据集中的患者的已知生物状态一致的预测。

[0022] 在某些实施方案中，候选基因标签根据至少两个不同度量进行排序以获得每个候选基因标签的第一排序和第二排序。每个候选基因标签的第一排序和第二排序可进行平均以获得每个相应候选基因标签的得分。

[0023] 在某些实施方案中，所述一组生物状态包含吸烟者状态。吸烟者状态可包含当前吸烟者和非吸烟者。

[0024] 在某些实施方案中，所述基因标签小于全基因组，且包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B和TLR5。另外，所述基因标签还可包括AK8、FSTL1、RGL1和VSIG4。另外，所述基因标签还可包括C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG和PTGFRN。另外，所述基因标签还可包括ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618。在一些实施方案中，所述基因标签可能限于阈值数目的基因，例如10、15、20、25、30、35、40或小于全基因组中的基因数目的任何其它合

适的基因数目。

[0025] 在某些实施方案中，所述基因标签小于全基因组，且包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63。另外，所述基因标签还可包括DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1和GUCY1B3。在一些实施方案中，所述基因标签可能限于阈值数目的基因，例如10、15、20、25、30、35、40或小于全基因组中的基因数目的任何其它合适的基因数目。

[0026] 在某些实施方案中，所述基因标签小于全基因组，且包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21。在一些实施方案中，所述基因标签可能限于阈值数目的基因，例如10、15、20、25、30、35、40或小于全基因组中的基因数目的任何其它合适的基因数目。

[0027] 在某些方面，本公开的系统和方法提供用于评定从受试对象获得的样本的计算机实施的方法。所述计算机实施的方法包括通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集。所述数据集包括小于全基因组的一组基因的定量表达数据，所述一组基因包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618。所述至少一个硬件处理器基于接收到的数据集而产生得分，其中所述得分指示受试对象的预测吸烟状态。

[0028] 在某些实施方案中，所述得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的定量表达数据而确定。

[0029] 在某些实施方案中，所述计算机实施的方法还包括计算AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618中的每一者的倍数变化值。所述计算机实施的方法还可包括确定每个倍数变化值满足至少一个准则，所述准则要求对于至少两个独立群体数据集，每个相应的所计算倍数变化值超过预定阈值。

[0030] 在某些实施方案中，所述一组基因由AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618组成。

[0031] 在某些方面，本公开的系统和方法提供一种用于预测个体的吸烟者状态的试剂盒。所述试剂盒包括：一组试剂，其检测测试样本中的基因标签的基因的表达水平，所述基因标签包括AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、

LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3和ZNF618;以及使用所述试剂盒来预测个体的吸烟者状态的说明书。

[0032] 在某些实施方案中,所述试剂盒用于评定吸烟产品的替代物对个体的作用。所述吸烟产品替代物可包含加热式烟草产品。所述替代物对个体的作用可以是将个体归类为非吸烟者。

[0033] 在某些方面,本公开的系统和方法提供用于评定从受试对象获得的样本的计算机实施的方法。所述计算机实施的方法包括通过包含至少一个硬件处理器的计算机系统接收与所述样本相关联的数据集,所述数据集包括小于全基因组的一组基因的定量表达数据,所述一组基因包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21。所述至少一个硬件处理器基于所接收的所述数据集中的所述一组基因的所述定量表达数据产生得分,其中所述得分基于少于40个基因,且指示所述受试对象的预测吸烟状态。

[0034] 在某些实施方案中,所述得分是应用于所述数据集的分类方案的结果,其中所述分类方案基于所述数据集中的定量表达数据而确定。

[0035] 在某些实施方案中,所述计算机实施的方法还包括计算AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21中的每一者的倍数变化值。所述计算机实施的方法还可包括确定每个倍数变化值满足至少一个准则,所述准则要求对于至少两个独立群体数据集,每个相应的所计算倍数变化值超过预定阈值。

[0036] 在某些实施方案中,所述一组基因由AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21组成。

[0037] 在某些方面,本公开的系统和方法提供一种用于预测个体的吸烟者状态的试剂盒。所述试剂盒包括:一组试剂,其检测测试样本中的基因标签中的基因的表达水平,所述基因标签包括AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1和TBX21,所述基因标签包括少于40个基因;以及使用所述试剂盒来预测个体的吸烟者状态的说明书。

[0038] 在某些实施方案中,所述试剂盒用于评定吸烟产品的替代物对个体的作用。所述吸烟产品替代物可包含加热式烟草产品。所述替代物对个体的作用可以是将个体归类为非吸烟者。

附图说明

[0039] 本发明的其它特征、其性质和各种优势在考虑结合附图进行的以下详细描述后将显而易见,

[0040] 在附图中,相似参考字符始终指代相似部分,且

[0041] 其中:

[0042] 图1是使用众包执行基因标签的识别的计算机化系统的框图。

[0043] 图2是可用于实施本文所描述的任一计算机化系统中的任一组件的示范性计算装置的框图。

[0044] 图3是使用众包识别基因标签以用于预测个体生物状态的过程的流程图。

[0045] 图4A和4B是指示跨越不同小组的人类数据(图4A)和无关物种的数据(图4B)的共同出现率的表。

[0046] 图5是评定指示受试对象的预测吸烟状态的得分的过程的流程图。

[0047] 图6是概括用于不同研究的样本群组/类别、大小和特性的表。

[0048] 图7A是简图,其示出识别来自人类和小鼠全血基因表达数据的化学暴露反应标志物以及将这些标志物用作计算模型中的标签以用于对作为暴露或非暴露群组的部分的新血液样本的预测性分类。

[0049] 图7B是示出开发稳健和稀疏人类(子挑战1,SC1)和无关物种的(子挑战2,SC2)基于血液的基因标签分类模型以(i)区分吸烟者与非当前吸烟者(任务1)以及随后(ii)将非当前吸烟者分类为曾吸烟者和从不吸烟者(任务2)的图。

[0050] 图8是示出发布血液基因表达数据的训练数据集、测试数据集和验证数据集的图。

[0051] 图9A是示出吸烟者与非吸烟者之间的清晰分隔的箱线图。

[0052] 图9B包含两个箱线图,其示出吸烟群组在0到5天的戒烟中无显著差异,但Cess组和切换组与其在0天时的相应基线相比显著下降。

[0053] 图10包含两个表,示出用于类别预测的基因标签分类模型的类别预测性能。

[0054] 图11A和11B是示出参与者针对测试和验证数据集进行的血液样本类别预测的箱线图。

[0055] 图12包含示出针对验证数据集在限制下的第0天与第5天之间的群体对数几率比的箱线图。

[0056] 图13是示出按组/类别以及暴露于pMRTP或候选MRTP的时间所分割的或在切换到pMRTP或候选MRTP之后的群体对数几率分布的箱线图。

[0057] 图14和15是MCC和AUPR得分的图,所述得分用以通过基于ML的类别预测来估计长度为2到18的标签的所有可能组合的性能。

具体实施方式

[0058] 本文描述用于识别稳健基因标签的计算系统和方法,所述基因标签可用来预测个体的生物状态。具体地说,生物状态可对应于个体的吸烟暴露反应状态。本文所述的基因标签能够区分当前吸烟的受试对象与从不吸烟或已戒烟的受试对象。尽管本文所描述的实例主要涉及吸烟者状态或吸烟暴露反应状态,但所属领域的技术人员应理解,本公开的系统和方法适用于使用众包方法来识别用于预测个体生物状态的基因标签,其中所述生物状态可指吸烟暴露反应状态、吸烟者状态、疾病状态、生理状态、化学暴露状态或与个体生物学数据相关联的任何其它合适的个体状态或状况。

[0059] 如本文所使用,个体生物状态可表示各种分子变化,所述分子变化可能在疾病中或响应于暴露于一种或多种有毒物质、药物、环境改变(例如温度、微重力、压力和辐射等)或其任何合适的组合而出现。限定预测性分类模型的基准且将其用在预测性分类模型的开发和训练的计算分析中。提取区分类别的特征并将其嵌入分类模型中以用于类别预测。如本文所使用,分类器包含判别特征和用于类别预测的规则。

[0060] 本文所描述的众包方法可用于识别稳健基因标签以预测个体暴露于一种或多种

化学物质的状态。下文相对于实例1所描述的研究涉及一种用于识别基因标签以用于预测个体暴露于烟雾的此类众包方法的示范性图解。下文描述的实例1中的研究识别从群体(例如多个挑战参与者)获得的基于人血的吸烟暴露反应基因标签的基因列表,以及从所述群体获得的无关物种的基于血液的吸烟暴露反应基因标签的基因列表。本文所描述的基因标签可应用于一个或多个分类模型,所述分类模型可应用于新人类(人类标签)或人类和啮齿动物(无关物种的标签)血液基因表达样本数据以预测个体是否已暴露于烟雾。本文所描述的系统和方法可扩展到识别基因标签和一个或多个分类模型以预测个体是否已暴露于一种或多种化学物质。尽管下文相对于实例1所描述的研究涉及识别基于血液的基因标签,但所属领域的技术人员应理解,本公开的系统和方法适用于使用众包方法来识别不仅仅是基于血液的基因标签。替代地,本公开适用于识别基于组织和例如蛋白质和甲基化改变等其它特征的基因标签。

[0061] 本公开的系统和方法可用于识别能够预测暴露于有毒物质的标志物。实际上,应用于新样本的基于稳健标志物的分类模型可实现(i)预测受试对象是否已暴露或未暴露于化学物质,以及(ii)允许在产品测试或撤回期间监测随着时间推移的暴露反应量值。

[0062] 如本文所使用,“稳健”基因标签是在研究、实验室、样品来源以及其它人口因素中维持强大性能的基因标签。重要的是,稳健标签应即使是在包含较大个体变化的一组群体数据中也可检测。跨越数据集的稳健性还应适当地进行验证以避免标签性能的过于乐观的报告。

[0063] 系统生物学旨在形成对生物系统借以对外部刺激(例如药物、营养和温度)和基因修饰(例如突变、表观遗传修饰)作出反应或调整的机制的详细理解。通过分析和整合使用组学(omics)或高内涵筛选等优势技术产生的大量分子和功能数据获得新的机理见解。当应用于毒理学领域时,称为系统毒理学的总体方法能够量化由外源性物质(例如杀虫剂、化学物质)所触发的生物系统扰动、阐明毒性动作模式以及估计相关联风险。系统毒理学有可能将短期观测结果外推到长期结果,并将从实验系统中识别的潜在风险转化到人类,从而表明其应用可成为用于风险评估和决策制定的新标准。系统毒理学数据的分析以及预测性毒理学结果和风险估计值的外推和转化需要开发高级的计算方法。为了展示新计算方法的改进性能和可靠性,研究人员可能根据先进的方法对其自身技术进行基准测试,但通常落入称作“自我评估陷阱”中,从而导致有偏倚的评估。此外,在系统生物学/毒理学中产生和分析的泛滥数据会使审裁者对发布的结果和结论的审查变得繁重。尽管审核人原则上可以访问已存储在公共存储库中的原始数据,但他们自己通常难以再现整个分析。因此,明确需要涉及外部第三方的方法和数据的独立和客观评估或验证。本公开的系统和方法解决此需要且提供一种众包方法,所述众包方法接收来自研究人员的提交、识别最佳执行技术以及将其结果汇总以形成用于预测生物状态的稳健基因标签。

[0064] 图1描绘可用于实施本文公开的系统和方法的计算机网络和数据库结构的实例。图1是根据说明性实施方案的用于使用众包执行基因标签的识别的计算机化系统100的框图。系统100包含服务器104和通过计算机网络102连接到服务器104的两个用户装置108a和108b(统称为用户装置108)。服务器104包含处理器105,且每个用户装置108包含处理器110a或110b以及用户界面112a或112b。如本文所用,术语“处理器”或“计算装置”是指一个或多个计算机、微处理器、逻辑装置、服务器或配置有硬件、固件和软件以执行本文所述的

计算化技术中的一种或多种的其它装置。处理器和处理装置还可包含用于存储输入、输出和当前在处理的数据的一个或多个存储装置。下文参考图2详细描述说明性计算装置200，其可用于实施本文所描述的处理器和服务器中的任一个。如本文所使用，“用户界面”包含但不限于一个或多个输入装置(例如小键盘、触摸屏、轨迹球、语音识别系统等)和/或一个或多个输出装置(例如视觉显示器、扬声器、触觉显示器、打印装置等)的任何合适的组合。如本文所用，“用户装置”包含但不限于配置有硬件、固件和软件以执行本文中所描述的一个或多个计算机化动作或技术的一个或多个装置的任何合适的组合。用户装置的实例包含但不限于个人计算机、笔记本电脑和移动装置(例如智能电话、平板电脑等)。图1中仅示出一个服务器、一个数据库和两个用户装置以免使图复杂化，但所属领域的技术人员应理解，系统100可支持多个服务器和任何数目的数据库或用户装置。

[0065] 计算机化系统100可用于利用群体智慧来识别用于预测个体生物状态的基因标签。如上文所描述，研究系统生物学的科学家通常落入自我评估陷阱，从而导致有偏倚的评估。本文所描述的众包方法通过设计挑战、使其向科学界开放(例如通过使关于基因表达的数据和已知生物状态数据库106可供用户装置108使用)、(例如从用户装置108a和108b)接收来自独立科学家或群组的提交以及将最佳执行结果或预测汇总而有助于避免这些偏倚。为确保广泛参与，所述挑战可能旨在解决与共同关注的科学问题相关的问题，例如识别基于血液的基因标签以用于预测个体生物状态或吸烟者状态。

[0066] 所述挑战使与从个体群组获得的血液样本数据相关联的某些数据可供科学界使用。具体地说，基因表达和已知生物状态数据库106(统称为数据库106)是包含表示一组个体和基因表达数据(从来自此组患者的血液样本获得)的已知生物状态的数据的数据库。所述一组个体(其血液样本数据存储在数据库106中)里的每个个体可随机指派为训练样本或测试样本。在一些实施方案中，将个体指派为训练或测试样本可能不是完全随机的。在此情况下，可在指派期间使用一个或多个准则，例如确保具有不同生物状态的类似数目的个体处于每个训练和测试数据集中。一般来说，任何合适的方法可用于将个体指派为训练或测试样本，同时确保生物状态的分布在训练数据集和测试数据集中在某种程度上是类似的。

[0067] 每个训练样本和测试样本包含从个体血液样本以及个体已知生物状态(例如个体已知吸烟者状态)所测量的基因表达水平。训练样本构成训练数据集，且测试样本构成测试数据集。整个训练数据集从数据库106提供到用户装置108，而仅一部分测试数据集提供到用户装置108。具体地说，将来自测试样本的所测量基因表达水平提供给用户装置108，但使对应于测试样本的已知生物状态对用户装置108保持隐藏。

[0068] 用户装置108处的科学家可分析训练样本以尝试识别所测量基因表达水平与训练数据集中的个体生物状态之间的相依性、关联或相关性。所识别相关性可具有候选基因标签和分类器的形式。候选基因标签包含针对与不同生物状态(例如当前吸烟者与非当前吸烟者)相关联的样本有差异地表达的基因列表。科学家可使用任何合适的计算技术、使用筛选器、包装器和嵌入法等任何特征选择技术来识别候选基因标签。所提取特征在使用机器学习方法训练过的分类模型中组合，所述机器学习方法例如判别分析、支持向量机、线性回归、逻辑回归、决策树、朴素贝叶斯(naive Bayes)、k最近邻法、K均值、随机森林或任何其它合适的技术。分类器包含决策规则或使用候选基因标签中的基因的表达水平将样本指派到某一类别的映射，所述类别可指个体的预测生物状态。以此方式，每个用户装置108处的每

个科学家基于训练数据集识别候选基因标签和分类器。

[0069] 用户装置108处的科学家使用其候选基因标签和分类器来预测测试数据集中的测试样本的生物状态。候选基因标签以及针对每个测试样本所获得的结果从用户装置108通过网络102提供到服务器104。来自科学家的提交可以是匿名的。在一个实例中，每个测试样本的结果包含对应于对应的测试样本属于预测生物状态的可能性或概率的置信水平。图3中相对于步骤308详细地描述所述置信水平。在另一实例中，所述结果不包含置信水平，而实际上仅包含每个测试样本的预测生物状态。

[0070] 接着，服务器104可通过比较针对每个测试样本所获得的结果与每个测试样本的已知生物状态来识别表现最佳的候选基因标签。一般来说，表现最佳候选基因标签具有密切匹配已知生物状态的结果。接着，服务器104跨越最佳执行候选基因标签进行汇总以获得可用于预测个体的生物状态的稳健基因标签。图3中相对于步骤314、316和318更详细地描述此过程。

[0071] 图1的系统100的组件可按数种方式中的任一种进行布置、分布和组合。例如，可使用将系统100的组件分布在通过网络102连接的多个处理和存储装置上的计算机化系统。此类实施方案可适用于在多个通信系统上的分布式计算，所述多个通信系统包含共享对共同网络资源的接入权的无线和有线通信系统。在一些实施方案中，系统100实施于云计算环境中，其中一个或多个组件由通过互联网或其它通信系统连接的不同处理和存储服务提供。服务器104可以是例如在云计算环境中实例化的一个或多个虚拟服务器。在一些实施方案中，服务器104与数据库106组合成一个组件。

[0072] 图3是使用众包识别基因标签以用于预测个体生物状态的方法300的流程图。方法300可由服务器104执行，且包含向一组用户装置提供包含基因表达数据和已知生物状态的训练数据集的步骤(步骤302)、向所述一组用户装置提供包含基因表达数据的测试数据集的步骤(步骤304)、接收包含被确定能判别训练数据集中的不同生物状态的一组基因的候选基因标签的步骤(步骤306)，以及针对每个候选基因标签，接收测试数据集中的每个样本的置信水平的步骤(步骤308)。方法300另外包含：基于置信水平与测试数据集中的已知生物状态之间的比较，根据第一性能度量对候选基因标签进行排序(步骤310)；针对每个候选基因标签，使用置信水平将测试数据集中的每个样本指派到预测生物状态(步骤312)；基于预测生物状态是否匹配测试数据集中的已知生物状态，根据第二性能度量对候选基因标签进行排序(步骤314)；基于步骤310和314中指派的排序，根据第三性能度量对候选基因标签进行排序(步骤316)；以及识别排名最靠前的候选基因标签中至少阈值数目个候选基因标签中包含的基因(步骤318)。

[0073] 在步骤302，将包含一组训练样本的基因表达数据和已知生物状态的训练数据集提供到一组用户装置108。如相对于图1所描述，在步骤302提供的训练数据集包含训练样本，所述训练样本包含从个体血液样本测得的基因表达水平以及所述个体的已知生物状态。用户装置108处的科学家接收训练数据集，且使用训练数据集训练提供所测量基因表达水平与已知生物状态之间的映射的分类器。在步骤304，将包含基因表达数据的测试数据集提供到此组用户装置108。如相对于图1所描述，在步骤304提供的测试数据集包含测试样本，所述测试样本仅包含从个体血液样本测得的基因表达水平，但不包含所述个体的已知生物状态。换句话说，测试样本的已知生物状态对用户装置108处的科学家保持隐藏。

[0074] 在步骤306，接收包含被确定能判别训练数据集中的不同生物状态的一组基因的候选基因标签。用户装置108处的每个科学家或科学家小组可将候选基因标签提供到服务器104，其中科学家确定，对于一个或多个准则(例如训练数据集中的样本的生物状态或暴露反应状态)，候选基因标签中的基因表达水平的组合具有判别力。借以提供训练数据集的用户装置可与科学家借以提供候选基因标签的用户装置相同或不同。

[0075] 在步骤308，对于每个候选基因标签，接收测试数据集中的每个测试样本的置信水平。所述置信水平可以是表示对应的测试样本属于特定生物状态的可能性的介于零与一之间的值。在一个实例中，当存在两个生物状态(例如第一生物状态和第二生物状态)时，所述置信水平可对应于值p，其指代特定测试样本属于第一生物状态的可能性。在此情况下，值1-p可指特定测试样本属于第二生物状态的可能性。一般来说，在存在多于两个生物状态时，可针对每个测试样本且针对每个候选基因标签提供多个置信水平。

[0076] 在步骤310，服务器104基于(在步骤308接收到的)置信水平与测试数据集中的已知生物状态之间的比较而根据第一性能度量对(在步骤306接收到的)候选基因标签进行排序。在步骤310执行的排序使每个候选基因标签被指派第一排序值。

[0077] 一种估计候选基因标签的性能的方式是在表中呈现预测结果，表的行包含预测生物状态，而表的列包含实际生物状态。下文所示表1是一种呈现预测结果的方式的实例。表的第一行指示被预测与第一生物状态(例如预测当前吸烟者)相关联的实际具有第一生物状态(例如真实当前吸烟者)的个体数目以及实际具有第二生物状态(例如非当前吸烟者)的个体数目。表的第二行指示被预测与第二生物状态(例如预测非当前吸烟者)相关联的实际具有第一生物状态(例如真实当前吸烟者)的个体数目以及实际具有第二生物状态(例如非当前吸烟者)的个体数目。

[0078] 表1

[0079]

	实际生物状态1	实际生物状态2

[0080]

预测生物状态1	真阳性	假阳性
预测生物状态2	假阴性	真阴性

[0081] 完美的预测器将使实际具有第一生物状态的所有个体准确预测为具有第一生物状态(真阳性将为100％，且假阴性将为0％)，以及实际具有第二生物状态的所有个体将准确预测为具有第二生物状态(真阴性将为100％且假阳性将为0％)。如本文所描述，可将个体分类成多个生物状态，例如吸烟状态(例如当前吸烟者、非当前吸烟者、曾吸烟者、从不吸烟者等)，但总的来说，所属领域的技术人员应理解，本文所描述的系统和方法适用于任何分类方案。

[0082] 为了估计预测器(例如分类器和候选基因标签)的力度，可使用基于预测结果表中的值的各种度量。在第一实例中，一个度量在本文中称为“敏感度”或“查全率”，其为被准确分类为第一生物状态(例如当前吸烟者)的个体在实际具有第一生物状态的一组个体中的比例。换句话说，敏感度(或查全率)度量等于真阳性的数目除以真阳性和假阴性的总和，或TP/(TP+FN)。敏感度值一指示实际属于第一生物状态的每个样本被正确地预测为属于第一生物状态，但未提供关于有多少其它样本被不当地预测为属于第一生物状态(FP)的信息。

[0083] 在第二实例中，一个度量在本文中称为“特异度”，其为被准确分类为第二生物状态(例如非当前吸烟者)的个体在实际具有第二生物状态的一组个体中的比例。换句话说，所述特异度度量等于真阴性的数目除以真阴性和假阳性的总和，或TN/(TN+FP)。特异度值一指示实际属于第二生物状态的每个样本被正确地预测为属于第二生物状态，但未提供关于具有第一生物状态的被不当预测为具有第二生物状态(FN)的样本数目的信息。

[0084] 在第三实例中，一个度量在本文中称为“精确度”，其为被准确分类为第一生物状态(例如当前吸烟者)的个体在预测具有第一生物状态的一组个体中的比例。换句话说，精确度度量等于真阳性的数目除以真阳性和假阳性的总和，或TP/(TP+FP)。精确度值一指示预测属于特定类别(例如生物状态)的每个样本实际属于该类别，但未提供关于具有第一生物状态的被不当预测为具有第二生物状态(FN)的样本数目的信息。

[0085] 若被视为强大的预测器，可能需要高值敏感度和特异度、高值敏感度和精确度或高值敏感度、特异度和精确度。尽管本文可能使用敏感度、特异度和精确度度量来评估候选基因标签的性能，但总的来说，也可在不脱离本公开的范围的情况下使用任何其它度量，例如阴性测试的预测值(TN/(TN+FN))。

[0086] 在实例中，第一性能度量涉及曲线下面积(AUC)度量。具体地说，所述曲线可对应于接受者工作特征(ROC)曲线或精确度-查全率(PR)曲线。ROC曲线的轴线对应于敏感度(或真阳性率:TP/(TP+FN))和假阳性率(FP/(FP+TN))。PR曲线的轴线对应于敏感度(TP/(TP+FN))和精确度(TP/(TP+FP))。在一个实例中，PR曲线下面积(AUPR)用作获得特定候选基因标签的第一排序的第一性能度量。在另一实例中，ROC曲线下面积用作第一性能度量。尽管PR曲线和/或ROC曲线可能是连续的，但本公开可使用离散值(因为阈值是变化的)，且一种或多种插值技术可用于计算曲线下面积。

[0087] 在步骤312，对于每个候选基因标签，服务器104使用置信水平将测试数据集中的每个样本指派到预测生物状态。具体地说，对于来自科学家的每个提交，基于所述提交中的置信水平将每个测试样本指派到预测生物状态。在一个实例中，当存在两个生物状态(第一生物状态和第二生物状态)时，所述置信水平可具有值p，其为测试样本属于第一生物状态的可能性。在此情况下，值1-p可对应于测试样本属于第二生物状态的可能性。一般来说，科学家在存在多个生物状态时可提交多个置信水平，且特定候选基因标签的预测生物状态可对应于具有最高置信水平的生物状态。

[0088] 在步骤314，服务器基于预测生物状态(在步骤312获得)是否匹配测试数据集中的已知生物状态而根据第二性能度量对候选基因标签进行排序。在步骤314处执行的排序使每个候选基因标签被指派第二排序值。

[0089] 在另一实例中，第二性能度量可对应于马修斯相关系数(MCC)度量。所述MCC度量将所有的真/假阳性和阴性率组合，且因此提供单值公平度量。MCC是可用作复合性能得分的性能度量。MCC是介于-1与+1之间的值，且基本上是介于已知的二元分类与预测的二元分类之间的相关系数。MCC可使用以下方程式计算：

[0090] $$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

[0091] 其中TP:真阳性;FP:假阳性;TN:真阴性;FN:假阴性。然而，一般来说，用于基于一组性能度量产生复合性能度量的任何合适的技术可用于评估候选基因标签的性能和其对

应的预测。MCC值+1指示模型获得完美预测，MCC值0指示模型预测几乎就是随机的，且MCC值-1表示模型预测完全不准确。MCC的优势在于，在以仅类别预测可用的方式编码分类器函数时，能够被容易地计算。一般来说，解释TP、FP、TN和FN的任何度量可用作根据本公开的第二性能度量。

[0092] 在步骤316，服务器104基于在步骤310和314处指派的排序而根据第三性能度量对候选基因标签进行排序。具体地说，在步骤310基于测试样本的原始置信水平与已知生物状态之间的比较而获得第一排序，且在步骤314基于测试样本的预测生物状态(根据置信水平评估)与已知生物状态之间的比较而获得第二排序。第一和第二排序可平均化(或以某种程度组合)以获得第三性能度量。

[0093] 在步骤318，服务器104识别N个排名最靠前的候选基因标签中至少阈值数目(例如M)个候选基因标签中包含的一组基因。在实例中，确定根据第三性能度量的N个最高排序候选基因标签。在这些N个候选基因标签中的至少M个中出现的任何基因包含于在步骤318识别的基因中，其中M小于N。在一些实施方案中，(N,M)＝(3,2)、(4,3)、(4,2)、(5,4)、(5,3)、(5,2)、(6,5)、(6,4)、(6,3)、(6,2)或N和M的值的任何其它合适的组合，其中N是范围从2到候选基因标签总数目的整数，且M是范围从2到N的整数。

[0094] 实例1-简介

[0095] 本文描述实例研究，其中众包方法用于获得稳健基因标签以准确预测个体吸烟者状态。实例研究的一个目标是通过对用于预测吸烟和戒烟状态的人类和无关物种的血液暴露反应标志物和模型的计算方法进行基准测试来识别血液中的化学暴露反应标志物。

[0096] 实例1-研究群体和设计

[0097] 在临床和活体研究期间将全血样本收集在PAXgene™管中，或从Biobank库购买全血样本。在图6所示的表中概述不同研究的样本群组/类别、大小和特性。简单来说，从以下方面获得人类血液样本：(i)在英国伦敦的Queen Ann Street Medical Center(QASMC)处进行并在ClinicalTrials.gov注册且识别码为NCT01780298的临床案例-对照研究；(ii)生物样本库(美国马里兰州贝茨维尔BioServe Biotechnologies Ltd.)(数据集BLD-SMK-01)。来自这些源的样本包含基于明确限定的纳入标准所选的吸烟者(S)、曾吸烟者(FS)和从不吸烟者(NS)(图6)；以及(iii)对应于随机化、对照性、3组平行群组和单中心研究的临床ZRHR降低暴露(REX)C-03-EU和-04-JP研究。所述REX研究旨在展示与限制的连续5天使用常规香烟(吸烟者)相比较，暴露于吸烟中的选定烟雾成分的降低切换到候选修改风险烟草产品(“MRTP”)或吸烟节制/戒烟(“Cess”)的健康受试对象。总的来说，MRTP可以是加热式烟草产品。如本文所使用，加热式烟草产品包含在使用期间通过在不燃烧或烧灼烟草的情况下加热烟草或包含烟草的混合物以生成气溶胶的产品。小鼠血液样本从两个独立香烟烟雾(“CS”)吸入研究中获得，所述研究分别利用雌性C57BL/6和ApoE$^{-/-}$小鼠进行7个月和8个月。研究包含随机化成五个组的小鼠：假(暴露于空气)、3R4F(暴露于来自参照香烟3R4F的CS)、原型/候选MRTP(暴露于来自原型/候选MRTP的主流气溶胶，所述原型/候选MRTP的尼古丁水平与3R4F的尼古丁水平匹配)、戒烟(Cess)和在暴露于3R4F达2个月之后切换到原型/候选MRTP(切换)。在不同时间点收集血液样本。

[0098] 实例1-血液转录组学数据集

[0099] 转录物学数据集由收集在PAXgene™管中的全血样本产生。

[0100] 人类和小鼠血液样本的数据产生

[0101] 总RNA使用PAXgene血液试剂盒隔离。使用UV分光光度计(NanoDrop® 1000或Nanodrop 8000;美国马萨诸塞州沃尔瑟姆赛默飞世尔科技)通过测量230、260和280nm处的吸光度来确定RNA样本的浓度和纯度。还使用Agilent 2100生物分析仪(安捷伦科技(Agilent Technologies),美国加州圣克拉拉)检查RNA完整性。仅对具有大于6的RNA完整性数目的RNA进行处理以进一步分析。

[0102] 根据制造商的说明书(Qiagen)在PAXgene™管中将总RNA与样本隔离。在使用Ovation® 全血试剂和Ovation RNA扩增系统V2(Nugen,AC Leek,荷兰)制靶后所提取RNA的质量和cDNA质量以及片段化(例如使用电泳图监测最终片段化和生物素化产物的大小分布)使用Agilent 2100生物分析仪(美国加州圣克拉拉)进行检查。利用SpectraMax® 384Plus微板读数仪(Molecular Devices,美国加州森尼韦尔测量cDNA的量。)通过使用片段分析仪(Advanced analytical,美国爱荷华州Ankeny)评定未片段化cDNA的大小来确定cDNA质量。在片段化和标记之后,根据制造商指南将cDNA片段与GeneChip Human Genome U133 Plus 2.0阵列(昂飞(Affymetrix))杂交。从微阵列图像分析获得原始转录组学数据。对于QASMC研究,由AROS Applied Biotechnology AS(丹麦奥尔胡斯)产生血液转录组学数据。

[0103] 数据处理

[0104] 使用冷冻稳健微阵列分析FRMA v1.1在R环境(v3.1.2)中处理和归一化来自每个数据集的原始数据(CEL文件)。frma和GNUSE函数使用冷冻参数向量人类(hgu133plus2frmavecs v1.3.0)。用于人类的自定义brainarray cdf文件(hgu133plus2hsentrezgcdf v16.0.0)用于昂飞探针到Entrez Gene ID映射,且对于一个基因关系,产生一个探针集。

[0105] 所述数据通过质量检查步骤,此步骤移除所有不符合本文所描述的基准的一个下述截断值的CEL文件。首先,对于给定探针集j,归一化未缩放标准误差(NUSE)提供其对给定阵列i相对于其它阵列的表达估计值的精确度的量度标准。有问题的阵列导致标准误差(SE)高于中值SE。如果任一NUSE中值超过1或阵列具有较大四分位距(IQR),则怀疑阵列质量不佳。将NUSE值高于1.05的阵列移除。其次,相对对数表达(RLE)针对每个阵列,比较给定探针的强度等级相对于该探针在所有j个阵列中的中值强度等级。RLE的阵列特异性分布用于确定特定阵列是否具有过低或过高表达的特征。未接近零的中值RLE指示上调基因的数目并非约等于下调基因的数目,且较大RLE IQR指示大多数基因的表达不同。具有中值RLE>0.1(按绝对值)的阵列被视为异常值且被移除。第三,具有中值绝对RLE(MARLE)大于所有阵列数据集MARLE除以0.01的平方根的中值绝对偏差(或中值(MARLE)/(1.4826*mad(MARLE))>1/sqrt(0.01))的阵列被视为具有不良质量的芯片且被移除。

[0106] 针对小鼠和人类的自定义Brainarray CDF文件用于昂飞探针到Entrez Gene ID映射,从而对于一个基因关系,产生一个探针集(分别为HGU133Plus2_Hs_ENTREZG v16.0、Mouse4302_Mm_ENTREZG v16.0)。所述质量检查排除不符合最小质量基准的CEL文件。为了促进数据集处理,人类和小鼠基因表达数据集同时具有人类基因符号。使用NCBI/HCOP映射文件将小鼠基因与人类基因相应。在小鼠基因映射到多个人类基因的情况下,仅保留匹配

所用小鼠基因的人类基因。

[0107] 实例1-挑战概述

[0108] 对于所述挑战，将来自吸烟者(S)和非当前吸烟者(NCS)受试对象的血液的基因表达谱例如通过相对于图1所描述的网络102提供给科学界。将一组基因表达谱均匀分成训练集和测试集。在发布测试数据集(没有关于受试对象生物状态的信息)之前发布训练数据集(具有关于受试对象生物状态的全部信息：吸烟者、曾吸烟者、从不吸烟者类别)。将135位注册科学家分组成61个小组。所述61个小组中的23个小组提供与挑战规则一致的提交，且所述23个小组中的12个小组提供符合条件的提交。图7A示出挑战的目的是从人类和小鼠全血基因表达数据中识别化学暴露反应标志物，并将这些标志物用作计算模型中的标签以用于作为暴露或非暴露群组的部分的新血液样本的预测分类。

[0109] 数据从独立临床和活体研究中收集的血液样本获得，所述研究与人类和啮齿动物的CS暴露和戒烟相关。实验群组还包含暴露于原型/候选MRTP或在暴露于CS一段时间之后切换到原型/候选MRTP的个体。要求参与者基于由血液样本产生的受试对象的基因表达谱来开发用以预测吸烟暴露的模型。具体地说，要求参与者解决两个任务：(1)识别吸烟者与非当前吸烟者受试对象，以及(2)对于预测为非当前吸烟者的每个受试对象，识别所述受试对象是曾吸烟者(FS)还是从不吸烟者(NS)受试对象。为了符合评分条件，对于这两个任务，需要小组提交预测(例如每个测试样本的置信水平)和候选基因标签(包含最大40个基因)当挑战结束时，根据与外部专家委员会建立的管线对匿名化预测评分。挑战中的最佳表现者实现几近完美预测以区分吸烟者与非当前吸烟者。

[0110] 挑战目标和规则

[0111] 要求参与者开发稳健和稀疏人类(子挑战1，SC1)和无关物种的(子挑战2，SC2)基于血液的基因标签分类模型以(i)区分吸烟者与非当前吸烟者(任务1)以及随后(ii)将非当前吸烟者分类为曾吸烟者和从不吸烟者(任务2，图7B)作为第一约束条件，要求预测模型为归纳式(与直推式相反)，能够预测单个新个体血液样本属于哪个类别，而不需要重新训练/优化模型或使用半监督方法组合训练和测试数据集来预测样本类别。作为第二约束条件，标签可包含不超过40个基因。

[0112] 数据发布为训练、测试和验证数据集

[0113] 图8示出发布血液基因表达数据的训练数据集、测试数据集和验证数据集的方法。在血液样本处理和基因表达数据产生之后，将来自独立研究的数据分成训练、测试和验证数据集。来自训练数据集的数据和类别标记被提供用于开发和训练基于血液的基因标签分类模型。所训练的模型无针对性地应用于随机化测试和验证基因表达数据集以用于血液样本的类别预测。

[0114] 具体地说，将来自QASMC临床(图7B，数据集H1)和小鼠C57BL/6吸入(图7B，数据集M1a)研究的归一化基因表达数据和类别标记提供为训练数据集。人类BLD-SMK-01和小鼠ApoE$^{-/-}$数据(图7B，分别是数据集H2和M2a)用作测试数据集。来自REX C-03-EU(图7B，数据集H3)/-04-JP(图7B，数据集H4)临床研究以及小鼠C57BL/6(图7B，数据集M1b)和ApoE$^{-/-}$(图7B，数据集M2b)吸入研究的数据发布为验证数据集。来自测试和验证集的样本数据完全随机化且分成依序发布用于类别标签预测的两个类别平衡子集(图8)。来自测试数据集的样本用于对每个子挑战中的参与者预测进行评分以及评估小组表现。验证集用于估计参与者

是将样本预测为更接近吸烟者还是非当前吸烟者。分别针对SC1和SC2发布仅人类数据以及人类和小鼠数据(图7B)。

[0115] 预测性基因标签分类模型

[0116] 为了避免选择偏倚或为了减弱维度通常影响基于全阵列的基因标签的性能的僵局,两个公共独立数据集用于导引筛选和基因选择。通过基于所述两个研究的N个最高倍数变化(按绝对值)的交集中的基因来评估(对于每一者,N≥1)线性判别模型,共同使用来自独立研究的最高倍数变化基因。通过5倍交叉验证(重复100次)选择最佳N,且产生11基因标签。

[0117] 对于所述挑战,参与者使用各种特征选择和机器学习方法来识别鉴别特征(基因)且对样本进行分类。随机森林、偏最小二乘判别分析(partial least square discriminant analysis)、线性判别分析(LDA)和逻辑回归是表现最佳的三个小组在两个子挑战中使用的分类方法。对于来自测试和验证数据集的每个样本,要求参与者提供样本属于类别1(例如吸烟者)的置信值值P(介于0与1之间),以及对应于样本属于类别2(例如非当前吸烟者)的置信值值的置信值值1-P。要求P和1-P不相等。

[0118] 表现评估评分

[0119] 将存在于测试数据集而非验证数据集中的样本用于对每个子挑战中的小组表现进行评估。使用马修斯相关系数和精确度查全率曲线下面积度量对匿名化参与者的类别预测进行评分。总体小组表现基于跨越度量和任务(任务1:吸烟者相对于非当前吸烟者;任务2:曾吸烟者相对于从不吸烟者)所计算的平均排序。评分结果和最终排序由所属领域的外部独立专家评分审查组进行审查和核准。为了估计此公开案的验证数据集的小组表现,使用来自REX研究的吸烟者和曾吸烟者(Cess)样本应用相同的评分方案。

[0120] 挑战后分析

[0121] 对应于血液样本是否属于吸烟者或3R4F组的置信值值转换为对数几率(log(P/(1-P)))。个体前三个小组(使用验证数据集重新评分)的或汇总为所有合格小组的中值的对数几率分布按类别在箱线图上显现。针对关键比较(即,所有组与其对应的吸烟者/3R4F组相比较)执行配对(纵向REX研究的第0天相对于第5天)和韦尔奇t检验(Welch t-test)。使用R软件v3.1.2完成所有统计和图形可视化。

[0122] 实例1-结果

[0123] 本实例中的案例分析报告了与MRTP评估相关的系统毒理学中的方法和数据的独立验证的结果。本研究的一个目标是评估用于开发能够预测吸烟暴露或戒烟状态的基于血液的人类和无关和物种的基因表达标签分类模型的计算方法(图7)。参与者无针对性地将其训练的模型应用于包含吸烟者/3R4F和非当前吸烟者(曾吸烟者/Cess和从不吸烟者/假)数据以及来自已暴露于原型/候选MRTP的小鼠或在暴露于常规CS之后切换到候选MRTP的人类受试对象和小鼠的数据的独立基因表达数据集。对于每个样本,参与者提交样本是属于烟雾暴露组还是非当前烟雾暴露组的置信值值。

[0124] 使用人类吸烟暴露基因标签分类模型减小来自5天戒烟和切换到候选MRTP组的样本与吸烟者(S)组的关联

[0125] 针对包括吸烟者、曾吸烟者和从不吸烟者的QASMC数据集训练人类吸烟暴露反应基因标签分类模型。所识别标签包含一组11个基因:LRRN3、SASH1、TNFRSF17、DDX43、RGL1、

DST、PALLD、CDKN1C、IFI44L、IGJ和LPAR1。为了测试标签区分吸烟者与非当前吸烟者的能力，将模型应用于测试数据集(BLD-SMK-01)，且针对每个样本计算具有样本属于吸烟者组的概率的LDA得分。样本属于吸烟者组(P)和NCS组(1-P)的概率被计算和转换为对数几率(P/(1-P))，从而量化样本与吸烟者组或非当前吸烟者组的关联。每组/类别的对数几率分布在箱线图上显现(图9A，其中韦尔奇t测试p值3*＜0.001相对于S组)。吸烟者类别的对数几率分布的中值大约为+3.0，而曾吸烟者类别和从不吸烟者类别的中值分别大约为-3.8和-5.8。吸烟者与非当前吸烟者类别之间的中值差越大，基因标签分类模型越具可辨别性。箱线图示出一侧的吸烟者与另一侧被定义为非当前吸烟者的曾吸烟者和从不吸烟者之间的清晰分隔(图9A)。

[0126] 相同的模型和程序直接应用于验证数据集(REX C-03-EU和REX C-04-JP)以确定来自切换或Cess受试对象的数据是分类成更接近吸烟者还是非当前吸烟者(图9A)。具体地说，切换受试对象是切换到候选MRTP的受试对象，且Cess受试对象是在限制下戒烟5天的受试对象。与吸烟者组相比较，在仅5天戒烟或切换之后，与这些组相关的对数几率显著地减小，而Cess组与切换组之间未发现差异(图9A)。对于吸烟组，未发现0天与5天之间有显著差异(对数几率比)，而对于Cess和切换组，与其在0天时相应的基线相比，观测到显著减小(图9B，配对t测试p值3*＜0.001)。

[0127] 众包数据验证确认来自5天戒烟和切换到候选MRTP组的血液样本属于吸烟者组的置信值减小的预测

[0128] 在训练其人类吸烟暴露反应基因标签分类模型之后，参与者将其模型应用于随机化测试和验证数据集，且计算每个受试对象他/她属于吸烟者组的置信值值(概率)。在挑战结束之后，对仅包含吸烟者、曾吸烟者和从不吸烟者的测试数据集执行评分。仅针对验证同期组群对参与者的预测提交进行重新评分，且小组225、264和257被识别为SC1的前三个小组(图10中示出的表)。用于类别预测的基因标签分类模型的类别预测性能使用作为最高标准的吸烟者和Cess(对于性能评估，被视为曾吸烟者)真类别标记进行评估，且发现前三个表现最佳的小组的AUPR曲线值至少为0.90(图10中示出的表)。

[0129] 图11示出参与者针对测试和验证数据集进行的人类和小鼠血液样本类别预测。具体地说，参与者训练人类(图11A)和无关物种的(图11B)基于血液的吸烟暴露基因标签模型以区分烟雾暴露(代表人类的S或代表小鼠的3R4F)人类受试对象和小鼠与非当前烟雾(NCS)暴露(曾吸烟者FS/Cess和从不吸烟者NS/假)人类受试对象和小鼠。对于每个样本，要求参与者提供样本属于S/3R4F组的置信值值P以及样本属于NCS组的置信值值1-P。置信值值转换为对数几率(log(P/(1-P)))，且通过计算所有12个具备资格的小组中的每个样本的中值进行汇总且显示为如箱线图的按类别分布(图11A)。对于测试数据集，所有结果示出吸烟者与非当前吸烟者(曾吸烟者和从不吸烟者)之间的清晰区分。对于验证数据集，使用所述模型所获得的来自5天Cess和切换组的样本与吸烟者组关联减小的观测通过产生类似结果的个体或汇总的参与者预测明显确认(图11A)。韦尔奇t测试p值为*＜0.05、2*＜0.01、3*＜0.001相对于S/3R4F组。这种朝向曾/从不类别的置信值值下降反映出已发生标签基因表达修饰且在5天的戒烟或切换到候选MRTP之后已经可以在血细胞中检测到。

[0130] 众包技术基准测试识别表现最佳的无关于人类和啮齿动物物种的血液样本类别预测的吸烟暴露模型。

[0131] 对于SC2,要求参与者开发用于类别预测的直接适用于人类和啮齿动物数据的无关物种吸烟暴露反应基因标签模型。使用验证数据集对参与者的预测提交进行重新评分将小组219、250和264识别为SC2的前三个小组(图10中的表)。对于SC1,由表现最佳的小组或在汇总所有小组值之后获得的置信值值显现为按类别的对数几率分布(图11B)。在针对人类和小鼠的箱线图上可观测到暴露于CS/3R4F与未暴露(从不吸烟者/假以及曾吸烟者/Cess)的同期组群之间的清晰分隔,指示所述模型能够对血液样本进行分类而无关于物种(图10、图11B中示出的表)。当将模型无针对性地应用于来自两个独立小鼠活体研究的验证样本时,对应于暴露于原型MRTP(pMRTP)或候选MRTP的组的样本具有与分别针对小鼠和人类数据集的假组和从不吸烟者对照组类似水平的对数几率值(图11B)。

[0132] 图12示出针对验证数据集在限制下的第0天与第5天之间的群体对数几率比。对于Cess和切换组,第0天与第5天之间的对数几率比显著不同,但正如所料,对于吸烟者组来说并无显著不同(配对t测试p值3*<0.001)。

[0133] 图13示出按组/类别分割以及暴露于pMRTP或候选MRTP的时间或在切换到pMRTP或候选MRTP之后分割的群体对数几率分布。具体地说,在CS暴露于pMRTP 2个月后切换之后,在根据时间点分类时,随着时间推移观测到对数几率值的逐渐减小(例如切换3、切换5和切换7对应于暴露于pMRTP 1个月、3个月和4个月),这指示随着时间推移在血细胞中发生逐渐的基因表达改变。

[0134] 血液中预测吸烟暴露状态的人类和无关物种的反应标志物示出共性且包含跨越小组高度一致的核心基因子集

[0135] 通过提取跨越前三个小组和PMI标签至少共同出现两次的基因(图4)来识别吸烟暴露核心基因子集。编码细胞周期素依赖性激酶抑制因子1C(CDKN1C)、富亮氨酸重复神经元3(LRRN3)和含有1的SAM和SH3域(SASH1)的基因是人类标签中最频繁出现的基因(图4A),且编码芳烃受体抑制子(AHRR)、嘧啶受体P2Y6(P2RY6)的基因在无关物种的标签中具有最高共同出现率(图4B)。两个核心基因子集之间的比较揭示编码LRRN3、SASH1、AHRR和P2RY6的一组共同的四个基因(图4)。

[0136] 实例1--来自前六个小组的基因标签长度、基因表达共线性等级和分类方法的基于人类的吸烟暴露共有标签影响的所有基因组合的性能分析

[0137] 方法

[0138] 考虑来自共有标签的所有可能的基因组合。由于此分析所需计算机密集计算所施加的限制,基于18个基因的人类吸烟暴露共有标签的提取限于前六个小组(而非12个合格小组)。血液中包含DSC2、FSTL1、GPR63、GSE1、GUCY1A3、RGL1、CTTNBP2、F2R、SEMA6B、CDKN1C、CLEC10A、GPR15、LINC00599、P2RY6、PID1、SASH1、AHRR和LRRN3的基于18个基因的共有标签通过选择跨越前六个小组的标签至少共同出现两次的基因进行识别。研究了基因标签大小和共线性等级对分类性能的影响。分别使用五倍交叉验证训练(10次重复)和来自SC1的测试数据集进行所述分析。所述挑战中最广泛应用的机器学习(ML)方法包含随机森林(RF)、具有线性核的支持向量机(svmLinear)、偏最小二乘判别分析(PLS)、朴素贝叶斯、k最近邻法(kNN)、线性判别分析(LDA)和逻辑回归(LR)。产生长度为2到18的18个基因的所有可能的组合(即262,125个基因集)。将七个ML方法中的每一者应用于每个基因集会产生总计1,834,875个测试分类策略。基因集内基因的共线性等级反映为限制于该基因集的表达矩阵

的第一主分量的差异百分比。通过计算MCC和AUPR得分来评估1,834,875个基因集-ML预测(称作“最前”)的性能。将这些“最前”基因集的性能与在差异表达基因(DEG;错误发现率或FDR＜＝0.5)或HG-U133_Plus_2芯片上表示的所有基因当中随机选择的基因集(2--18个基因)的性能相比较。针对每个基因集大小重复1,000次所述取样过程,从而产生总计17,000个随机“DEG”或″所有基因″基因集。

[0139] 结果:来自前六个小组的基于18个基因的共有标签的基因集组合信息量大且在吸烟暴露状态类别预测方面胜过“DEG”和“所有基因”导出的基因集

[0140] 使用来自前六个小组的预测的基于18个基因的共有标签探索基因标签大小和共线性等级对吸烟暴露状态类别预测性能的影响。计算MCC和AUPR得分以通过基于ML的类别预测来估计长度为2到18的标签的所有可能组合的性能(图14和15)。图14和15显示MCC得分(图14)和AUPR得分(图15)的结果。在两个图中,图区A描绘得分与交叉验证和测试数据集的基因标签大小。特征选自以下列表:(i)“最前”基因(即,被参与者频繁地选为标签的部分的基因);(ii)“DEG”,差异表达基因列表;(iii)“所有基因”,所有所测量基因。在两个图中,图区B描绘得分与标签中的基因之间的相似系数。测试七个不同机器学习分类器:随机森林(RF)、具有线性核的支持向量机(svmLinear)、偏最小二乘判别分析(PLS)、朴素贝叶斯(NB)、k最近邻法(kNN)、线性判别分析(LDA)和逻辑回归(LR)。在两个图中,图区C描绘CV和测试集数据中的得分分布加上“最前”(顶部)、“DEG”(中间)和“所有基因”(底部)选择的差异的分布。

[0141] 如图14和15中的数据所指示,预测性能随着基因集大小增大且在较长集的情况下逐渐稳定,所述较长集包含训练(交叉验证,CV)(对于CV,大小＝2时,MCC＝0.57,且大小＝18时,MCC＝0.91)和测试集(对于测试,大小＝2时,MCC＝0.42,且大小＝18时,MCC＝0.77)中多达18个基因(图14A)。预测性能在“最前”基因集中的基因的共线性等级(通过由第一主分量表示的差异百分比反映,所述第一主分量根据基因集表达矩阵计算出来)范围介于50％与60％之间时达到最大值,且接着随着共线性的增加而减小(图14B)。考虑到“最前”基因集由来自不同小组的标签基因构成且已经非常不同,将某一程度上共线的基因进行组合可加强预测。性能随着来自DEG的基因集内的基因的共线性增加而减小(图14B)。一般来说,来自“最前”、“DEG”和“所有基因”的基因集分别产生最佳、中等和最差性能(图14)。另外,源自CV的性能胜过针对测试集所计算的性能(图14)。通过各种ML方法获得的性能度量示出类似图案(图14B),且因此进行汇总以促进结果的可视化(图14A和图14C)。总体上,结果指示来自基于18个基因的共有标签的血液基因信息量大且在组合时对吸烟暴露状态具有较高预测力。

[0142] 实例1-论述

[0143] 在此实例研究中所获得的结果提供暴露于候选MRTP的受试对象、或在常规CS暴露之后切换到候选MRTP的受试对象的血液样本属于烟雾暴露组或是非当前烟雾暴露组的预测置信值。

[0144] 所述结果清晰地分开吸烟者和非当前吸烟者。挑战参与者成功地开发出无关物种的基于血液的基因标签模型,其示出极好的吸烟暴露状态预测性能而无关于人类和小鼠物种。在人类测试数据集中,曾吸烟者组,尽管极接近于从不吸烟者组,但仍处于吸烟者组与从不吸烟者组中间,从而指示曾吸烟者的基因标签中的基因的表达可能无法完全逆转回到

从不吸烟者的表达水平。改变的逆转很可能取决于吸烟史和戒烟持续时间，这在受试对象之间有所不同，从而还解释了此组的预测的较高变化性。对于曾吸烟者的血细胞，DNA甲基化水平(例如F2RL3基因)可能取决于吸烟指数和自戒烟以来的时间。

[0145] 在小鼠数据集中，Cess组的表达水平达到假手术组水平，从而表明小鼠品系的血细胞的标签基因表达改变的逆转与遗传基因有关且在实验上更均匀。有趣的是，此逆转随着时间推移逐渐发生，就如在基于戒烟持续时间分组时观测到的那样。这表明所述基因标签分类方法不仅适用于二元分类，还可以更定量的方式使用(例如LDA得分等模型参数的量值或相关联置信值值)以遵循在产品测试或撤回后在血液中出现的改变的量值和动力学。实际上，这是来自验证人类REX数据集的切换组和Cess组的案例，与吸烟者组相比较，其示出朝向从不吸烟者组的值的显著对数几率减小。此观测指示在切换到候选MRTP或戒除常规香烟仅5天之后，在血细胞中发生吸烟暴露标签基因反映的分子改变。这些结果与在临床“每天香烟减少”限制研究一周之后所测量的剂量反应性暴露生物标志物的降低一致。对于小鼠验证数据集，3R4F组与原型/候选MRTP或切换组(类似假手术组的水平)之间的对数几率差甚至更为重要，因为这可通过在切换之后较长(数月)暴露于候选MRTP或pMRTP来阐述，且与常规CS相比较，反映出MRTP对血细胞的生物作用降低。

[0146] 尽管用于开发和训练基于血液的吸烟暴露反应分类模型的计算方法不同，但通过表现最优的小组获得的样本分类性能较高。识别跨越小组高度一致的核心基因标签，从而指示烟雾暴露所引发的基因表达改变提供充足信息且一贯地选择共同建立特异性和稳健血液标志物的基因，所述标志物仅预测人类或人类和小鼠(无关物种的标签)的吸烟暴露状态。

[0147] 血细胞类型特异性转录组分析，类似于报告的来自吸烟者和非吸烟者的细胞特异性白细胞的DNA甲基化分析，可帮助更好地理解每个血细胞类型对吸烟暴露反应标签的贡献。一些基因可与特定血细胞亚群相关。总体上，作为核心标签的部分的这些吸烟暴露相关联基因构成一组稳健的血液标志物，其可用于监测以及有可能量化与常规香烟的影响相比较的候选MRTP等新产品的影响。

[0148] 相对于实例1所描述的研究示出可如何利用群体的力量来评估计算方法以及验证系统毒理学中的数据。除补充传统同行互审过程之外，对产品风险评估数据的独立和无偏倚评估可用于确认和提供科学结论的置信值，且可支持管理机构进行决策。尽管本文所描述的实例主要涉及使用众包方法来识别用于预测个体吸烟者状态的稳健基因标签，但所属领域的技术人员应理解，本公开的系统和方法可应用于获得用于预测个体生物状态的基因标签，所述生物状态包含吸烟者状态、疾病状态、生理状态、暴露状态或与个体生物状态相关联的任何其它合适的个体状态或状况。

[0149] 下文表2包含根据实例1进行的研究的结果。具体地说，表2中示出的结果从人类吸烟标签中得来，且在第一列中列出一组基因。第二列列出在其标签中包含对应基因的(共12个)小组或参与者的数目。第三列列出在其标签中包含对应基因的前3个小组(根据测试数据集评估)的数目。第四列列出在其标签中包含对应基因的前3个小组(根据验证数据集评估)的数目。第五列列出第三和第四列中的值的均值。

[0150] 表2

[0151]

评分测试集	总计（共 12 组）	总计前 3 测试集	总计前 3 验证集	平均集+验证
LRRN3	9	3	3	3
AHRR	9	3	3	3
CDKN1C	9	3	3	3
PID1	8	3	3	3
SASH1	7	3	3	3
GPR15	7	3	3	3
P2RY6	6	3	3	3
LINC00599	6	2	3	2.5
CLEC10A	6	3	2	2.5
SEMA6B	5	2	3	2.5
F2R	5	2	2	2
DSC2	5	1	0	0.5
TLR5	5	0	1	0.5
RGL1	4	1	2	1.5
FSTL1	4	1	0	0.5
VSIG4	4	0	0	0
AK8	4	0	0	0
CTTNBP2	3	2	2	2
GUCY1A3	3	1	1	1
GSE1	3	1	0	0.5
MIR4697HG	3	0	0	0
PTGFRN	3	0	0	0
LOC200772	3	0	0	0

[0152]

FANK1	3	0	0	0
C15orf54	3	0	0	0
MARC2	3	0	0	0
GPR63	2	2	1	1.5
TPPP3	2	1	1	1
ZNF618	2	1	1	1
PTGFR	2	1	0	0.5
GUCY1B3	2	0	1	0.5
P2RY1	2	0	0	0
TMEM163	2	0	0	0
ST6GALNAC1	2	0	0	0
SH2D1B	2	0	0	0
CYP4F22	2	0	0	0
PF4	2	0	0	0
FUCA1	2	0	0	0
MB21D2	2	0	0	0
NLK	2	0	0	0
B3GALT2	2	0	0	0
ASGR2	2	0	0	0
NR4A1	2	0	0	0
RTN1	1	1	1	1
MAFB	1	1	1	1
ARHGEF10L	1	1	1	1
CLDN23	1	1	1	1
TGFBI	1	1	1	1
LOC284837	1	1	1	1
SYCE1L	1	1	1	1
SEZ6L	1	1	1	1
KLF4	1	1	1	1
NOD1	1	1	1	1
FAM225A	1	1	1	1
CRACR2B	1	1	0	0.5

[0153] 在一些实施例中，用于确定吸烟暴露反应状态的基因标签包含在表2中列出的基因，其对应于在表现最佳的三个基因标签中的至少两个基因标签中出现的基因。在根据测试数据集(例如表2的第三列中示出)进行评估时，这包含LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63。在根据验证数据集(例如表2的第四列中示出)进行评估时，这包含LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、RGL1和CTTNBP2。在根据测试和验证数据集之间的均值(例如表2的第五列中示出)进行评估时，这包含LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R和CTTNBP2。

[0154] 在一些实施例中，用于确定吸烟暴露反应状态的基因标签包含在表2中列出的基因，其对应于在十二个候选基因标签中的至少M个基因标签中出现的基因，其中M是1、2、3、4、5、6、7、8或9。例如，当M是9时，基因标签包含在第二列中具有至少为9的值的那些基因，即：LRRN3、AHRR和CDKN1C。作为另一实例，当M是8时，基因标签包含在第二列中具有至少为8

的值的那些基因，即：LRRN3、AHRR、CDKN1C和PID1。作为另一实例，当M是7时，基因标签包含在第二列中具有至少为7的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1和GPR15。作为另一实例，当M是6时，基因标签包含在第二列中具有至少为6的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599和CLEC10A。作为另一实例，当M是5时，基因标签包含在第二列中具有至少为5的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、DSC2和TLR5。作为另一实例，当M是4时，基因标签包含在第二列中具有至少为4的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、DSC2、TLR5、RGL1、FSTL1、VSIG4和AK8。作为另一实例，当M是3时，基因标签包含在第二列中具有至少为3的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、CTTNBP2、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54和MARC2。作为另一实例，当M是2时，基因标签包含在第二列中具有至少为2的值的那些基因，即：LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、CTTNBP2、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、GPR63、TPPP3、ZNF618、PTGFR、GUCY1B3、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2和NR4A1。作为另一实例，当M是1时，基因标签包含上文表2中列出的所有基因。

[0155] 下文表3包含根据实例1进行的研究的结果。具体地说，表2中示出的结果从无关物种的吸烟标签中得来，且在第一列中列出一组基因。第二列列出在其标签中包含对应基因的(共12个)小组或参与者的数目。第三列列出在其标签中包含对应基因的前3个小组(根据测试数据集评估)的数目。第四列列出在其标签中包含对应基因的前3个小组(根据验证数据集评估)的数目。第五列列出第三和第四列中的值的均值。

[0156] 表3

[0157]

评分测试集	总计（共 12 组）	总计前 3 测试集	总计前 3 验证集	平均集+验证
AHRR	5	3	3	3
P2RY6	4	3	3	3
COX6B2	2	2	2	2
DSC2	2	2	2	2
KLRG1	3	2	2	2
LRRN3	3	2	2	2
SASH1	2	2	2	2
TBX21	2	2	2	2
ADORA3	1	1	1	1
AF529169	1	1	1	1
AKAP5	1	1	1	1
ASGR2	1	1	1	1
B3GALT2	1	1	1	1
BCL3	1	1	1	1
BIRC2	1	1	1	1
CCR4	1	1	1	1
CDKN1C	1	1	1	1
CLEC10A	1	1	1	1
CLEC5A	1	1	1	1
CNNM1	1	1	1	1
COL6A3	1	1	1	1
COX6C	1	1	1	1
CRACR2B	1	1	1	1
CTNNAL1	1	1	1	1
CTTNBP2	2	1	1	1
DCAF8	1	1	1	1
EIF5A2	1	1	1	1
ELOVL7	1	1	1	1
ENDOU	1	1	1	1
ERI1	1	1	1	1
ESAM	1	1	1	1
EVA1B	1	1	1	1
F2R	2	1	1	1
FANK1	1	1	1	1
FKRP	1	1	1	1
FSTL1	1	1	1	1
GGT7	1	1	1	1

[0158]

GLCCI1	1	1	1	1
GNAZ	1	1	1	1
GNPDA2	1	1	1	1
GP1BA	1	1	1	1
GPR63	1	1	1	1
GSE1	1	1	1	1
GUCY1B3	2	1	1	1
HES1	1	1	1	1
HPGD	1	1	1	1
HSPB6	1	1	1	1
IRF7	1	1	1	1
JARID2	1	1	1	1
KCNQ1OT1	1	1	1	1
KISS1R	1	1	1	1
LIMS1	1	1	1	1
LRRK1	1	1	1	1
LTBP1	1	1	1	1
MBTD1	1	1	1	1
MCEMP1	1	1	1	1
MKNK1	1	1	1	1
MPP2	1	1	1	1
MRAS	1	1	1	1
MT2	2	1	1	1
NDUFA3	1	1	1	1
NGFRAP1	2	1	1	1
NR4A1	1	1	1	1
PF4	1	1	1	1
PGRMC1	1	1	1	1
PHACTR3	1	1	1	1
PID1	1	1	1	1
PTGFR	1	1	1	1
R3HDM4	1	1	1	1
RBM43	1	1	1	1
REEP6	2	1	1	1
REXO2	1	1	1	1
RUNDC3A	1	1	1	1
SAMD11	1	1	1	1
SDR16C5	1	1	1	1
SIAH1A	1	1	1	1
SLPI	1	1	1	1
SPINK2	1	1	1	1
STAR	1	1	1	1
SYTL4	1	1	1	1
TCEAL8	1	1	1	1
TLR2	1	1	1	1
TMEM163	1	1	1	1
TRIB3	1	1	1	1
UBE2B	1	1	1	1

[0159]

VCAN	1	1	1	1
VSIG4	1	1	1	1
WDFY1	1	1	1	1
ZFP704	1	1	1	1

[0160] 在一些实施例中，用于确定吸烟暴露反应状态的基因标签包含在表3中列出的基因，其对应于在表现最佳的三个基因标签中的至少两个基因标签中出现的基因。如表3中所示，不论这是根据测试数据集(例如表3的第三列中示出)、验证数据集(例如表3的第四列中示出)还是根据测试和验证数据集之间的均值(例如表3的第五列中示出)进行的评估，这包含AHRR、P2RY6、COX6B2、DSC2、KLRG1、LRRN3、SASH1 TBX21。

[0161] 在一些实施例中，用于确定吸烟暴露反应状态的基因标签包含表3中列出的基因，其对应于在所提交的12个基因标签中的至少M个基因标签中出现的基因，其中M是1、2、3、4或5。例如，当M是5时，基因标签包含在第二列中具有至少为5的值的那些基因，即：AHRR。作为另一实例，当M是4时，基因标签包含在第二列中具有至少为4的值的那些基因，即：AHRR和P2RY6。作为另一实例，当M是3时，基因标签包含在第二列中具有至少为3的值的那些基因，即：AHRR、P2RY6、KLRG1和LRRN3。作为另一实例，当M是2时，基因标签包含在第二列中具有至少为2的值的那些基因，即：AHRR、P2RY6、KLRG1、LRRN3、COX6B2、DSC2、SASH1、TBX21、CTTNBP2、F2R、GUCY1B3、MT2、NGFRAP1和REEP6。作为另一实例，当M是1时，基因标签包含上文表3中列出的所有基因。

[0162] 在一些实施例中，本文所描述的基因标签限于具有最大数目的基因，例如10、11、12、13、14、15、20、25、30、35、40或小于全基因组中的基因数目的任何其它合适的数目。此处所描述的基因标签限于与全基因组相比相对较少数目的基因。在较长基因标签与训练数据集过拟合的情况下，较长基因标签可能表现得比较短基因标签差。在此情况下，较长基因标签可能描述训练数据集中的随机误差或噪声。当用于预测测试数据集中的类别时，较短基因标签可能胜于过拟合的较长基因标签。本文所描述的任一基因标签，包含相对于表2和3所描述的基因标签，可限于具有特定最大数目的基因。

[0163] 图5是根据本公开的说明性实施例的用于评定从受试对象获得的样本的过程500的流程图。过程500包含以下步骤：接收与样本相关联的数据集，所述数据集包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63的定量表达数据(步骤502)；以及基于接收到的数据集产生得分，其中所述得分指示受试对象的预测吸烟状态(步骤504)。在一些实施例中，在步骤502接收到的数据集还包括任何数目的以下基因的定量表达数据：DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1和GUCY1B3。在一些实施例中，在步骤502接收到的数据集还包括相对于上文表2和3所描述的任何基因标签或本文所描述的任何其它基因标签的定量表达数据。

[0164] 在步骤504产生的得分是应用于所述数据集的分类方案的结果，其中所述分类方案基于所述数据集中的定量表达数据而确定。具体地说，在本文所描述的实例中，使用机器学习技术训练的分类器可应用于在502接收到的数据集以确定个体的预测分类。

[0165] 本文所述的基因标签可在计算机实施的方法中用于评定从受试对象获得的样本。具体地说，可获得与所述样本相关联的数据集，且所述数据集可包含用于核心基因标签的LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63的定量表达数据。总的来说，相对于表2和3所描述的任何基因标签可用作核心基因标签。核心基因标签包含小于全基因组中的基因数目的数个基因，且包含在一起视为整体时提供用于预测吸烟状态等生物状态的信息的一组基因。可基于接收到的数据集中的基因标签产生得分，其中所述得分指示受试对象的预测吸烟状态。具体地说，所述得分可基于使用本文所描述的众包方法构建的分类器。所述数据集还可包括可包含在扩展基因标签中的额外标志物DSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1和GUCY1B3的任何合适组合的定量表达数据。所述数据集还可包括相对于上文表2和3所描述的任何基因标签的定量表达数据。

[0166] 在一些实施例中，所述数据集包含以下一组标志物LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63的任何子集的任何数目。所述子集可包含不到全部的这些所识别基因。一个或多个准则可应用于将包含在标签中的标志物，所述标签例如包含以下核心集中的至少三个(或任何其它合适的数目，例如4、5、6、7、8、9、10、11或12个)标志物:LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63，以及相对于表2或3所描述的基因标签中的任何标志物中的至少两个(或任何其它合适的数目，例如2、3、4、5、6、7、8、9、10、11或12个)标志物。如上文所描述，在一些实施例中，所述标签限于小于全基因组中的基因数目的数个基因，且可能限于最大数目的基因，例如10、11、12、13、14、15、20、25、30、35、40或小于全基因组中的基因数目的任何其它合适的数目。总的来说，在不脱离本公开的范围的情况下，使用这些标志物的组合的任何标签可用于预测受试对象的生物状态，例如吸烟状态。

[0167] 在一些实施例中，本文所描述的标签中的基因用于装配用于预测个体的吸烟者状态的试剂盒。具体地说，所述试剂盒包含:一组试剂，其检测测试样本中基因标签中的基因的表达水平;以及使用所述试剂盒预测个体的吸烟者状态的说明书。所述试剂盒可用于评估戒烟或吸烟产品的替代物——例如HTP——对个体的作用。

[0168] 图2是用于执行本文所描述的过程中的任一种，例如相对于图1和2所描述的过程，或用于存储本文所描述的核心基因标签、扩展基因标签或任何其它基因标签的计算装置的框图。具体地说，存储在计算机可读媒体上的基因标签包含LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63的表达数据。在另一实例中，计算机可读媒体包含基因标签，所述基因标签包含选自以下组的至少4、5、6、7、8、9、10、11或12个标志物的表达数据:LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63。在另一实例中，计算机可读媒体包含与本文所描述的任何基因标签或标志物集合相关的数据。

[0169] 在某些实施方案中，可跨越若干计算装置200实施组件和数据库。计算装置200包括至少一个通信界面单元、输入/输出控制器210、系统存储器和一个或多个数据存储装置。所述系统存储器包含至少一个随机存取存储器(RAM 202)和至少一个只读存储器(ROM

204)。所有这些元件都与中央处理单元(CPU 206)通信以促进计算装置200的操作。可按许多不同方式配置计算装置200。例如,计算装置200可以是常规的独立计算机,或者,计算装置200的功能可分布在多个计算机系统和架构中。计算装置200可配置成执行建模、评分和汇总操作中的一些或全部操作。在图2中,计算装置200通过网络或局域网连接到其它服务器或系统。

[0170] 计算装置200可配置成分布式架构,其中数据库和处理器容纳在分开的单元或位置中。一些此类单元执行主要的处理功能,且至少含有通用控制器或处理器和系统存储器。在此类方面,这些单元中的每一个通过通信界面单元208附接到通信集线器或端口(未示出),所述集线器或端口用作与其它服务器、客户端或用户计算机和其它相关装置的主要通信链路。所述通信集线器或端口自身可具有最低的处理能力,其主要用作通信路由器。各种通信协议可以是系统的部分,包含但不限于:Ethernet、SAP、SAS™、ATP、BLUETOOTH™、GSM和TCP/IP。

[0171] CPU 206包括处理器,例如一个或多个常规的微处理器和用于分担CPU 206的工作负荷的数学协处理器等一个或多个补充协处理器。CPU 206与通信界面单元208和输入/输出控制器210通信,CPU 206通过所述通信界面单元和输入/输出控制器与其它服务器、用户终端或装置等其它装置通信。通信界面单元208和输入/输出控制器210可包含多个通信信道以用于与例如其它处理器、服务器或客户终端同时通信。彼此通信的装置无需不断地彼此发送。相反,此类装置仅需要在必要时彼此发送,实际上可以在大部分时间避免交换数据,且可能需要执行若干步骤以在装置之间建立通信链路。

[0172] CPU 206还与数据存储装置通信。所述数据存储装置可包括磁性、光学或半导体存储器的适当组合,且可包含例如RAM 202、ROM 204、快闪驱动器、压缩光盘或硬盘或驱动器等光学光盘。CPU 206和数据存储装置各自可例如完全位于单个计算机或其它计算装置内;或通过通信媒体彼此连接,所述通信媒体例如USB端口、串口线、同轴电缆、以太网类型线缆、电话线、射频收发器或其它类似无线或有线媒体,或前述各者的组合。例如,CPU 206可通过通信界面单元208连接到数据存储装置。CPU 206可配置成执行一个或多个特定处理功能。

[0173] 所述数据存储装置可存储例如:(i)用于计算装置200的操作系统212;(ii)一个或多个应用程序214(例如计算机程序代码或计算机程序产品),其适于根据本文所述系统和方法且尤其根据相对于CPU 206详细描述的过程来指导CPU 206;或(iii)适于存储信息的数据库216,可用来存储程序所需的信息。在一些方面,所述数据库包含存储实验数据和公布的文献模型的数据库。

[0174] 操作系统212和应用程序214可例如以压缩、未编译和加密的格式存储,且可包含计算机程序代码。程序的指令可从数据存储装置之外的计算机可读媒体,例如从ROM 204或从RAM 202,读取到处理器的主存储器中。尽管程序中的指令的序列的执行会使CPU 206执行本文所描述的过程步骤,但硬接线电路系统可替代或结合软件指令来用于实施本公开的过程。因此,所描述的系统和方法不限于硬件和软件的任何特定组合。

[0175] 可提供合适的计算机程序代码来执行本文所描述的一个或多个功能。所述程序还可包含操作系统212、数据库管理系统和“装置驱动器”等允许处理器通过输入/输出控制器210与计算机外围装置(例如视频显示器、键盘、计算机鼠标等)介接的程序元件。

[0176] 如本文所使用，术语“计算机可读媒体”是指提供或参与提供指令到计算装置200的处理器(或本文所描述的装置的任何其它处理器)以供执行的任何非暂时性媒体。此类媒体可采用许多形式，包含但不限于非易失性媒体和易失性媒体。非易失性媒体包含例如光学、磁性或光学磁盘或集成电路存储器，例如快闪存储器。易失性媒体包含动态随机存取存储器(DRAM)，其通常构成主存储器。常见形式的计算机可读媒体包含例如软盘、软磁盘、硬盘、磁带、任何其它磁性媒体、CD-ROM、DVD、任何其它光学媒体、穿孔卡片、纸带、带有孔图案的任何其它物理媒体、RAM、PROM、EPROM或EEPROM(电可擦除可编程只读存储器)、FLASH-EEPROM、任何其它存储芯片或盒，或计算机可从中读取的任何其它非暂时性媒体。

[0177] 各种形式的计算机可读媒体可涉及将一个或多个指令的一个或多个序列传输到CPU 206(或本文所描述装置的任何其它处理器)以供执行。例如，所述指令初始可承载于远程计算机(未示出)的磁盘上。远程计算机可将指令加载到其动态存储器中，且通过以太网连接、电缆线路或甚至使用调制解调器的电话线发送所述指令。计算装置200本地的通信装置(例如服务器)可在相应的通信线路上接收数据，并将数据置于用于处理器的系统总线上。系统总线将数据传输到主存储器，处理器从主存储器检索和执行指令。由主存储器接收的指令可任选地在由处理器执行之前或之后存储在存储器中。另外，指令可通过通信端口接收为电信号、电磁信号或光信号，这些信号是传输各种类型的信息的无线通信或数据流的示范形式。

[0178] 本文引用的每篇参考文献均以引用方式将其相应全文并入本文中。

[0179] 虽然已参考具体实例特别地示出和描述了本公开的实施方案，但所属领域的技术人员应理解，在不脱离由所附权利要求限定的本公开的范围的情况下，可对这些实施方案做出形式和细节上的各种改变。因此，本公开的范围由所附权利要求书指示，且因此，属于权利要求书的等同涵义和范围内的所有改变都旨在被涵盖。

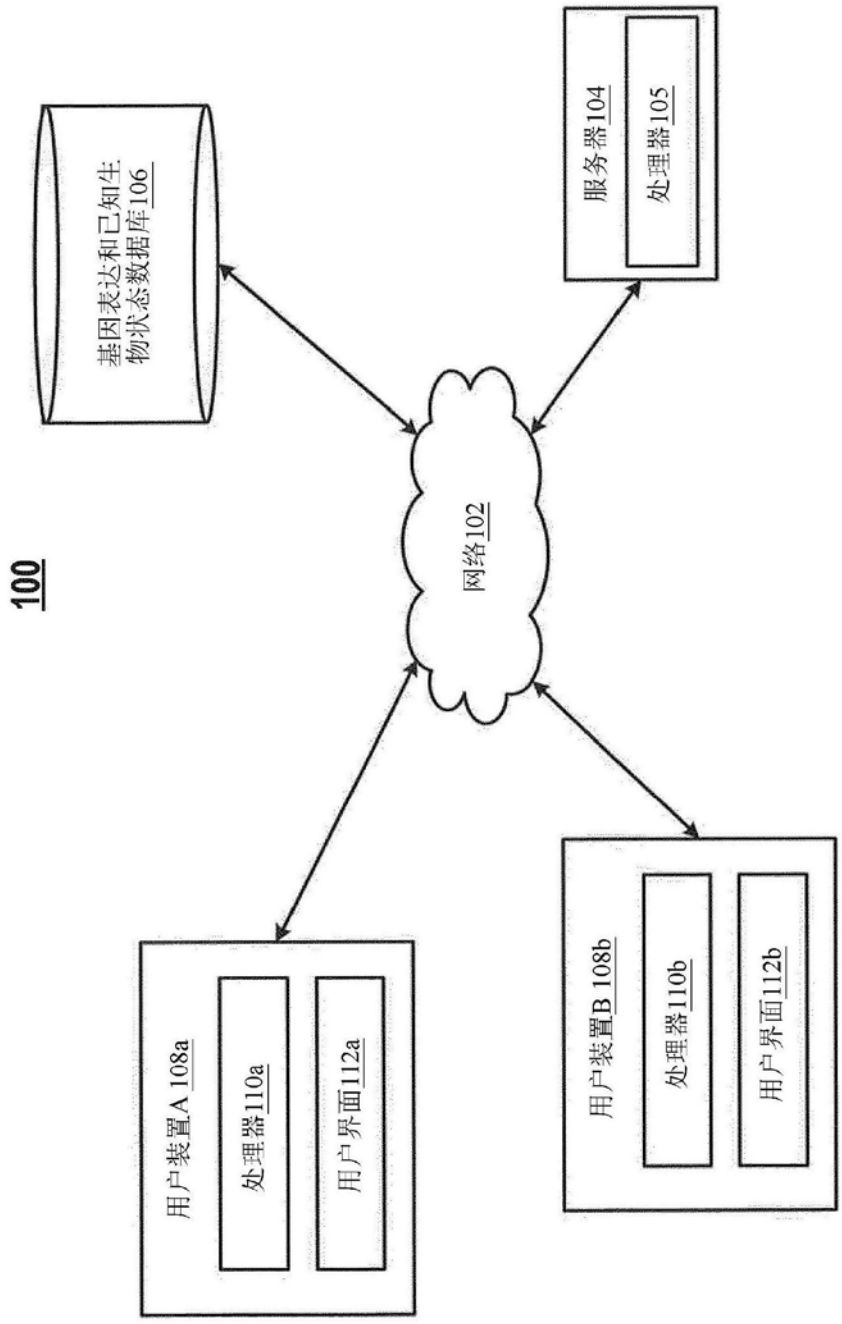
100
基因表达和已知生物状态数据库106
服务器104
处理器105
网络102
用户装置A 108a
处理器110a
用户界面112a
用户装置B 108b
处理器110b
用户界面112b

图1

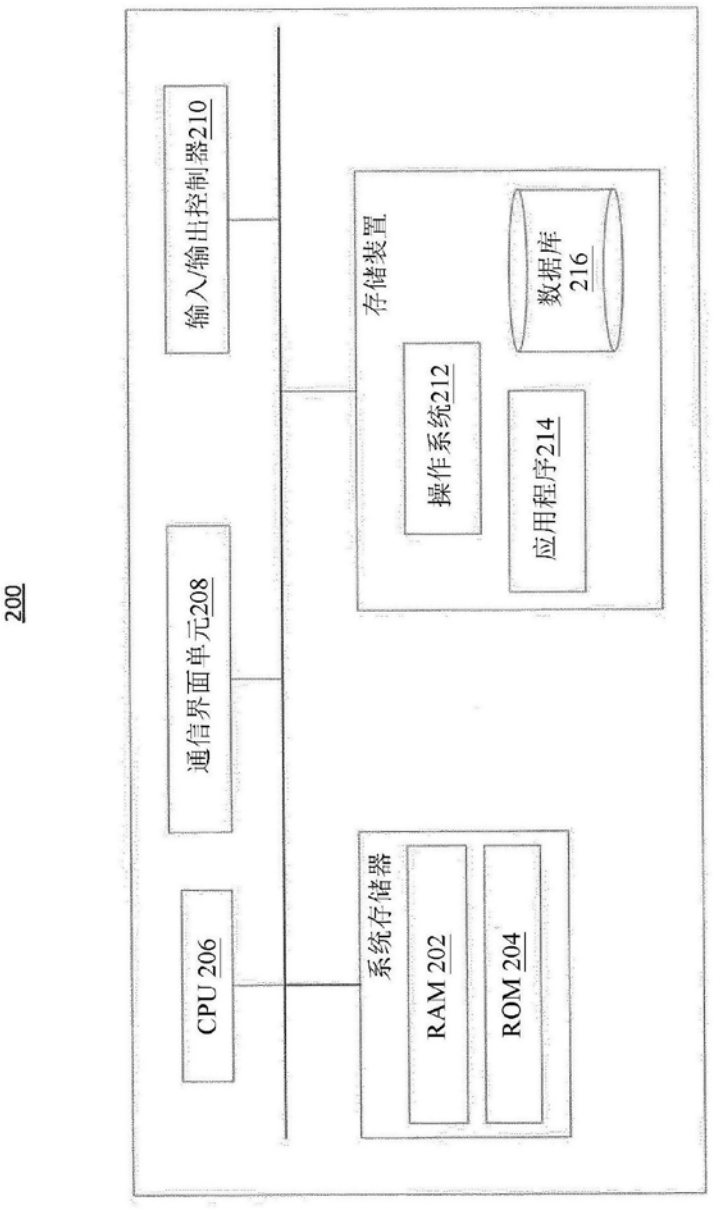
200
输入/输出控制器210
通信界面单元208
CPU 206
系统存储器
RAM 202
ROM 204
存储装置
操作系统212
应用程序214
数据库
216

图2

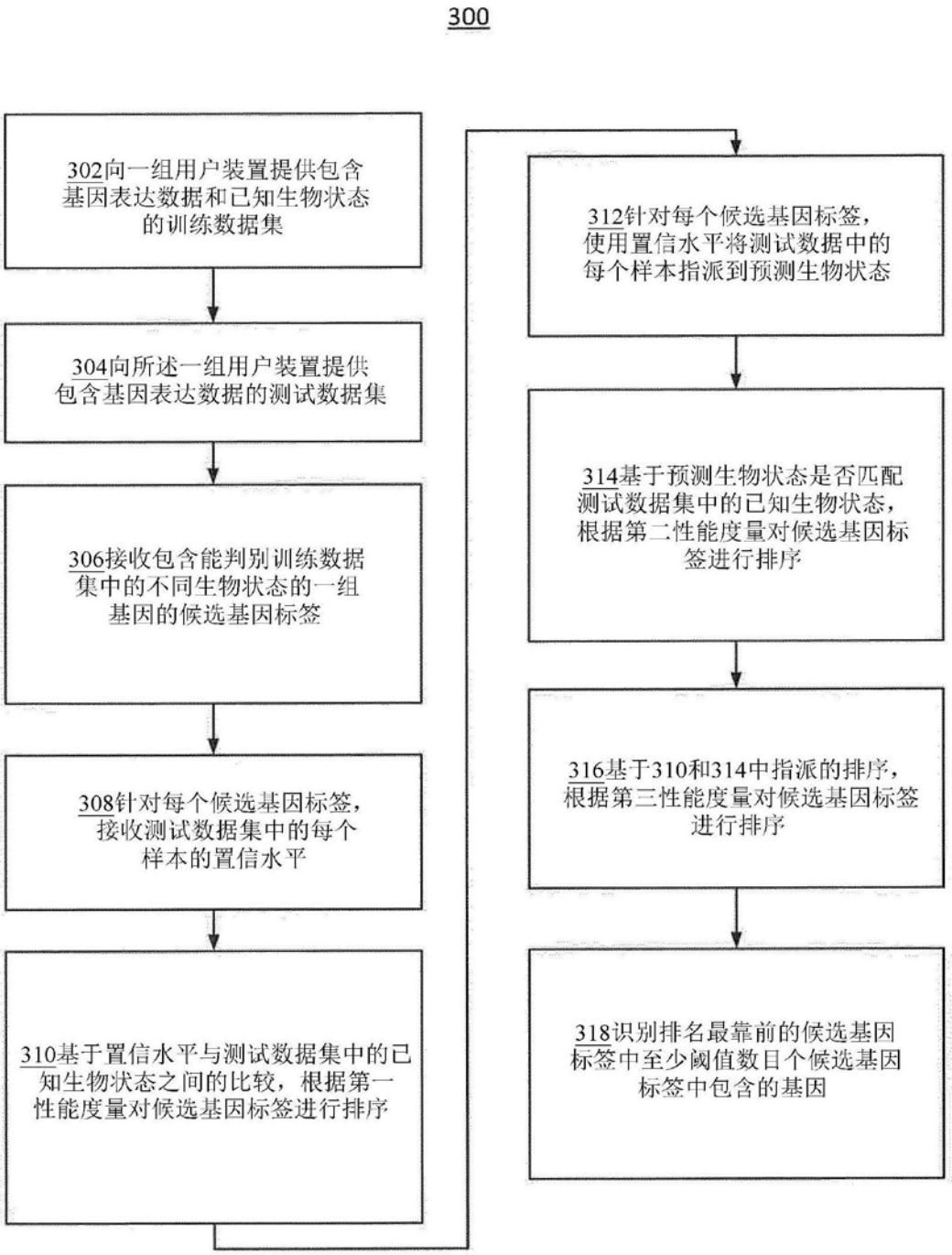
300
302向一组用户装置提供包含基因表达数据和已知生物状态的训练数据集
304向所述一组用户装置提供包含基因表达数据的测试数据集
306接收包含能判别训练数据集中的不同生物状态的一组基因的候选基因标签
308针对每个候选基因标签，接收测试数据集中的每个样本的置信水平
310基于置信水平与测试数据集中的已知生物状态之间的比较，根据第一性能度量对候选基因标签进行排序
312针对每个候选基因标签，使用置信水平将测试数据中的每个样本指派到预测生物状态
314基于预测生物状态是否匹配测试数据集中的已知生物状态，根据第二性能度量对候选基因标签进行排序
316基于310和314中指派的排序，根据第三性能度量对候选基因标签进行排序
318识别排名最靠前的候选基因标签中至少阈值数目个候选基因标签中包含的基因

图3

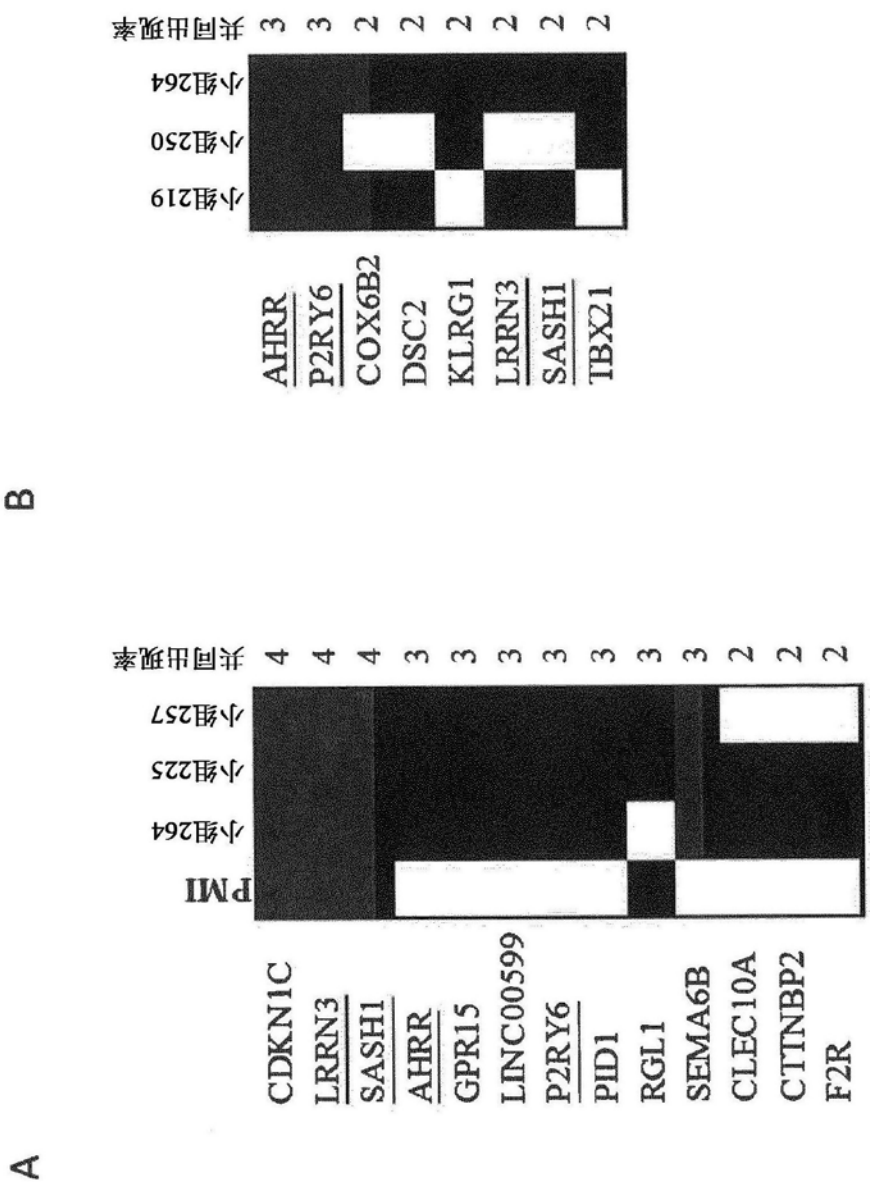
B
共同出现率
3 3 2 2 2 2 2 2
小组219
小组250
小组264
AHRR
P2RY6
COX6B2
DSC2
KLRG1
LRRN3
SASH1
TBX21
A
共同出现率
4 4 4 3 3 3 3 3 3 3 2 2 2
小组257
小组225
小组264
PMI
CDKN1C
LRRN3
SASH1
AHRR
GPR15
LINC00599
P2RY6
PID1
RGL1
SEMA6B
CLEC10A
CTTNBP2
F2R

图4

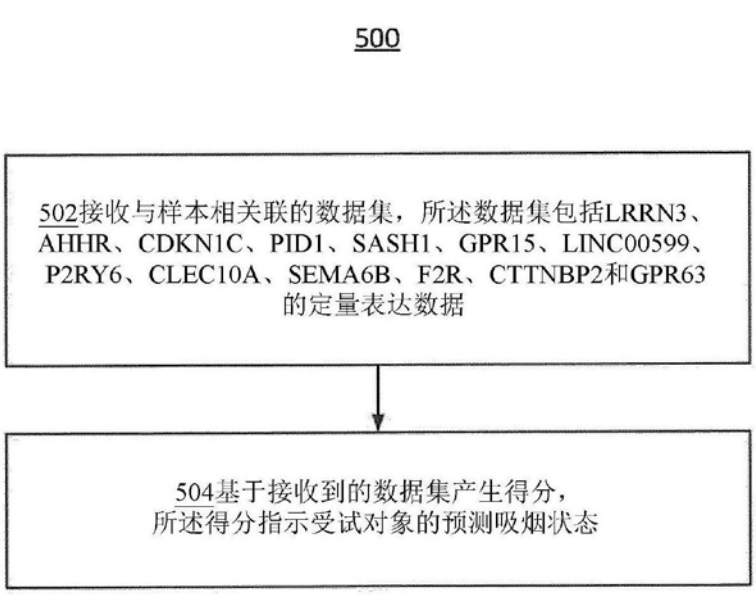
500
502接收与样本相关联的数据集，所述数据集包括LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2和GPR63的定量表达数据
504基于接收到的数据集产生得分，所述得分指示受试对象的预测吸烟状态

图5

数据集名称	数据集代码	总体描述	吸烟者	非当前吸烟者		其它组	
			S/3R4F	FS/CESS	NS/SHAM	P/CMRTP	SWITCH
QASMC NCT01780298	H1	临床案例-对照研究每组60个40到70岁的男（58%）女（42%）受试对象	N=109 COPD黄金阶段1和2 ≥10包/年吸烟史	N=57匹配吸烟史戒烟至少1年	N=58按年龄和种族匹配	-	-
BLD-SMK-01	H2	BIOBANK血液样本23到65岁排除：病史和药物治疗	N=27至少3年中每天≥10支香烟	N=26戒烟至少2年	N=28按年龄和性别匹配	-	-
REX C-03-EU NCT01959932	H3	限制下的临床随机化对照研究。	N=180	N=37戒烟5天	-	-	N=70吸烟者切换到THS2.2达5天（随意）
REX C-04-JP NCT01970982	H4		N=176	N=31	-	-	N=63
小鼠 C57BL/6 IS	M1AM1B	每日暴露于3R4F烟雾或MRTP气溶胶：四次（C57BL/6）或三次（APOE-/-）1小时封闭穿插新鲜空气暴露间歇。 暴露于3R4F达2个月后发生戒烟或切换	N=40	N=27	N=45	PMRTP N=45	N=28
小鼠 APOE-/-IS	M2AM2B		N=12	N=8	N=13	CMRTP (THS2.2) N=9	N=8

缩写： REX，临床ZRHR降低暴露；IS，吸入研究；S，吸烟者；FS，曾吸烟者；NS，从不吸烟；P/CMRTP，潜在/候选修改风险烟草产品；CESS，戒烟；SWITCH，切换。以NCT起始的编号对应于在CLINICAL TRIALS.GOV.注册的临床研究的唯一标识符。

颜色代码：训练数据集 测试数据集 验证数据集

图6

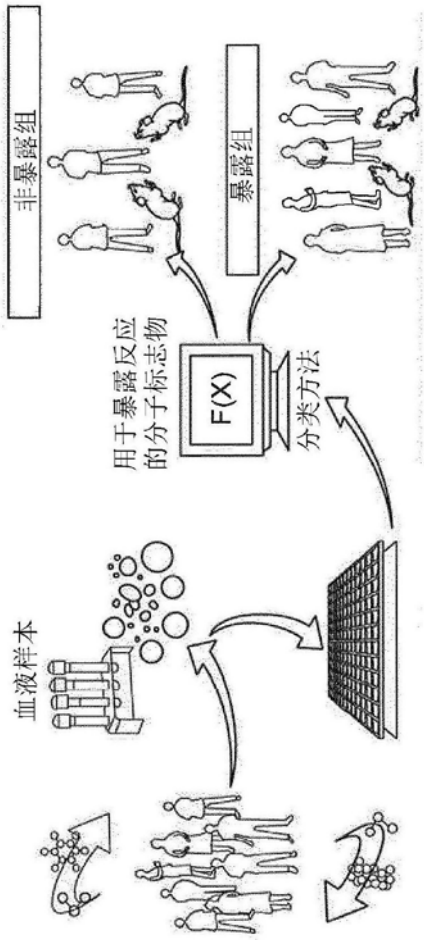

图7A

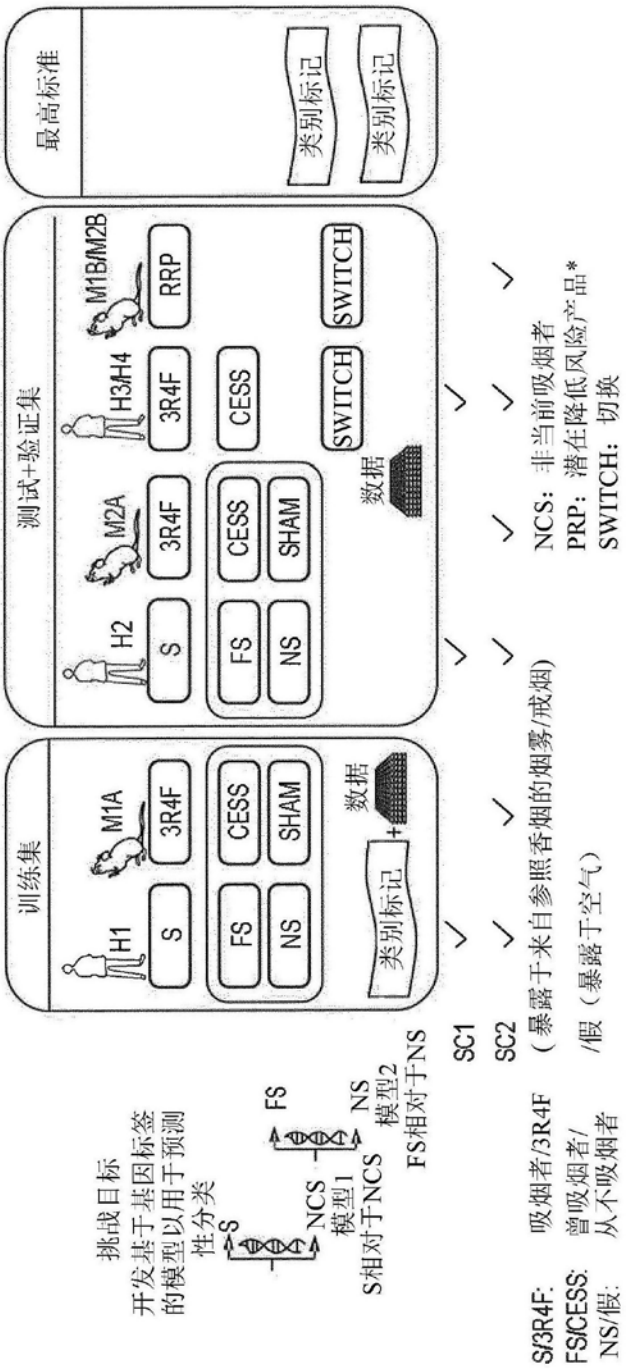
训练集
测试+验证集
最高标准
H1
M1A
H2
M2A
H3H4
M1B/M2B
S
3R4F
FS
CESS
NS
SHAM
RRP
SWITCH
数据
类别标记
SC1
SC2
NCS：非当前吸烟者
PRP：潜在降低风险产品*
SWITCH：切换
S/3R4F：吸烟者/3R4F（暴露于来自参照香烟的烟雾/戒烟）
FS/CESS：曾吸烟者/
NS/假：从不吸烟者/假（暴露于空气）
挑战目标
开发基于基因标签
的模型以用于预测
性分类
S
NCS
模型1
S相对于NCS
FS
NS
模型2
FS相对于NS

图7B

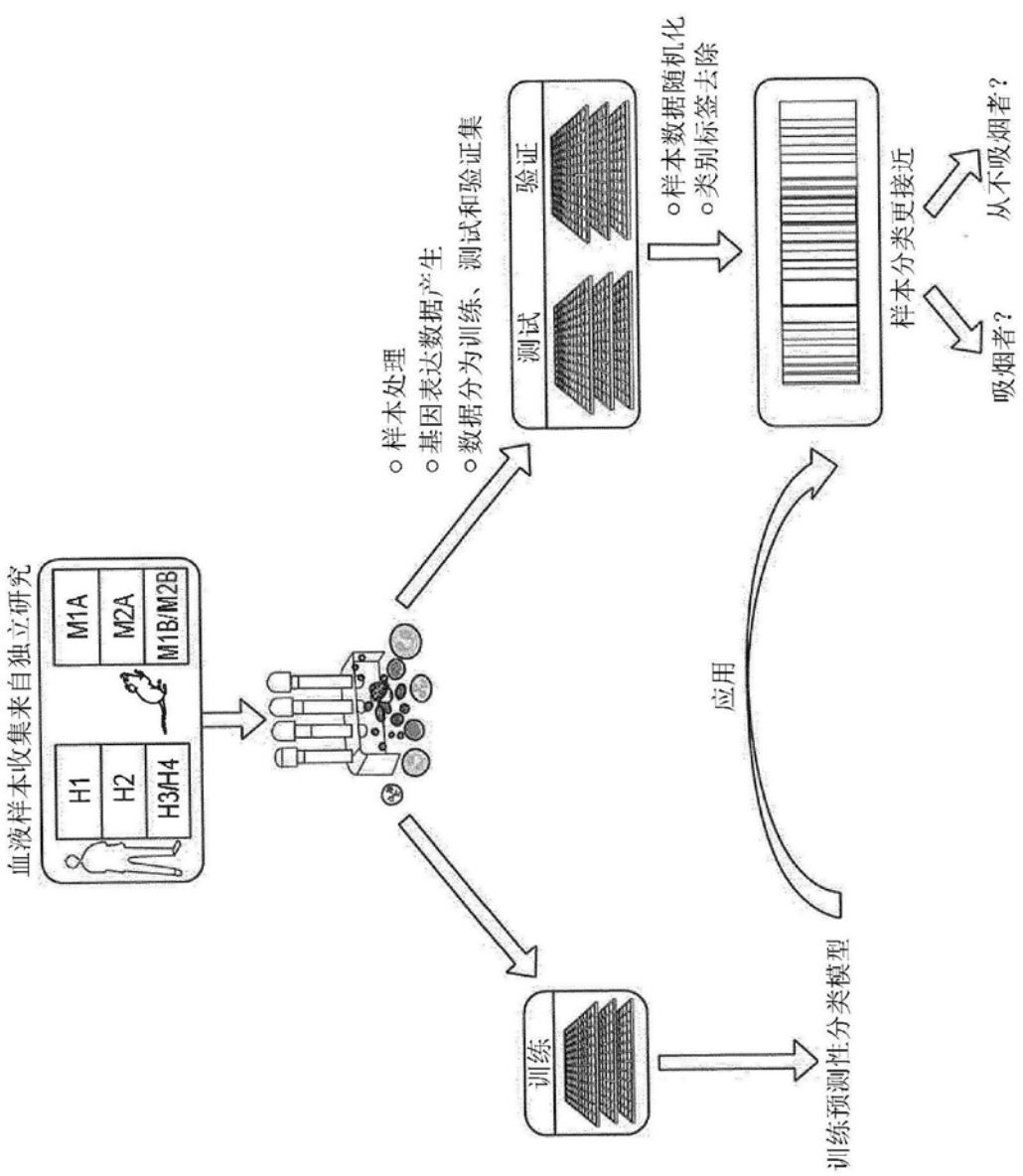
血液样本收集来自独立研究
H1
H2
H3/H4
M1A
M2A
M1B/M2B
样本处理
基因表达数据产生
数据分为训练、测试和验证集
训练
测试
验证
样本数据随机化
类别标签去除
应用
样本分类更接近
吸烟者？
从不吸烟者？
训练预测性分类模型

图8

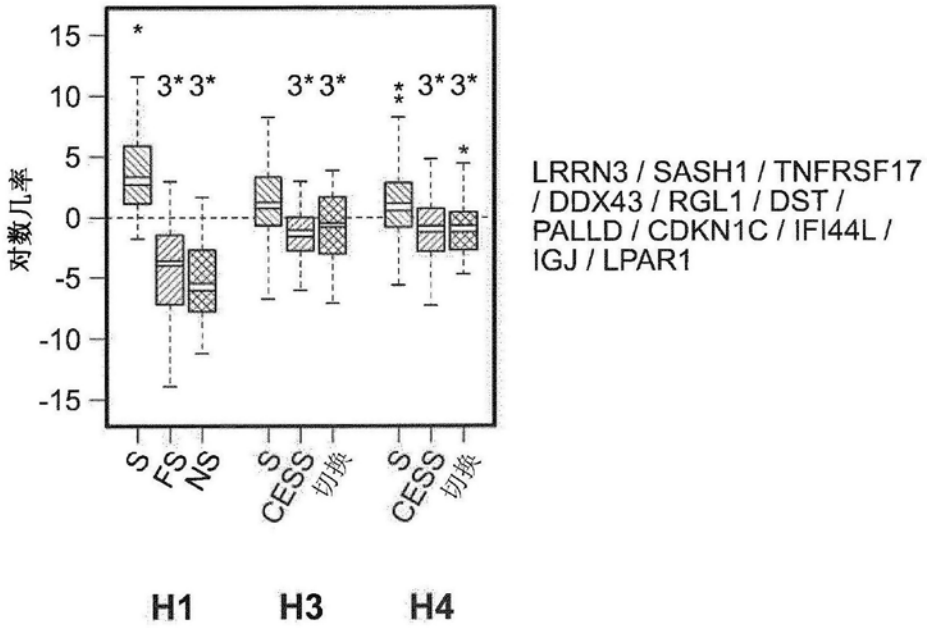
对数几率
15
10
5
0
-5
-10
-15
3* 3*
3* 3*
3* 3*
S
FS
NS
S
CESS
切换
S
CESS
切换
H1
H3
H4
LRRN3 / SASH1 / TNFRSF17 / DDX43 / RGL1 / DST / PALLD / CDKN1C / IFI44L / IGJ / LPAR1

图9A

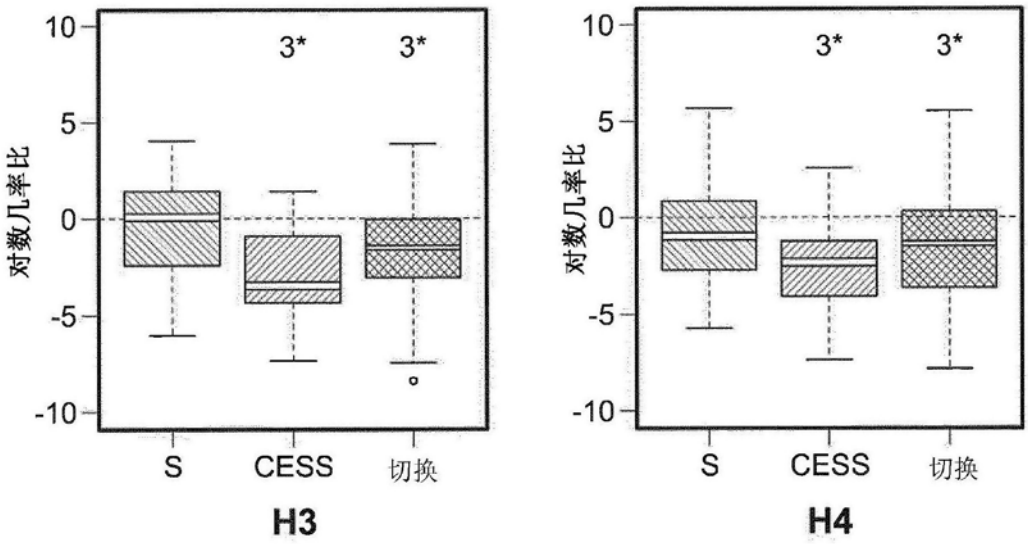
对数几率比
10
5
0
-5
-10
3*
3*
S
CESS
切换
H3
对数几率比
10
5
0
-5
-10
3*
3*
S
CESS
切换
H4

图9B

SC1

小组	AUPR	MCC	平均排序
225	**0.94**	**0.24**	**1**
264	**0.92**	**0.18**	**2**
257	**0.91**	**0.16**	**3**
259	0.90	0.15	4
269	0.90	0.10	6.5
222	0.89	0.13	7
250	0.89	0.12	7
247	0.90	0.09	8
283	0.90	0.08	8
290	0.87	0.11	8.5
221	0.85	0.06	11
215	0.82	-0.07	12
PMI	0.93	0.24	

SC2

小组	AUPR	MCC	平均排序
219	**0.99**	**0.87**	**1**
250	**0.96**	**0.81**	**2**
264	**0.96**	**0.75**	**3**
225	0.73	0.38	4
247	0.62	0.20	5.5
221	0.46	0.39	5.5

图10

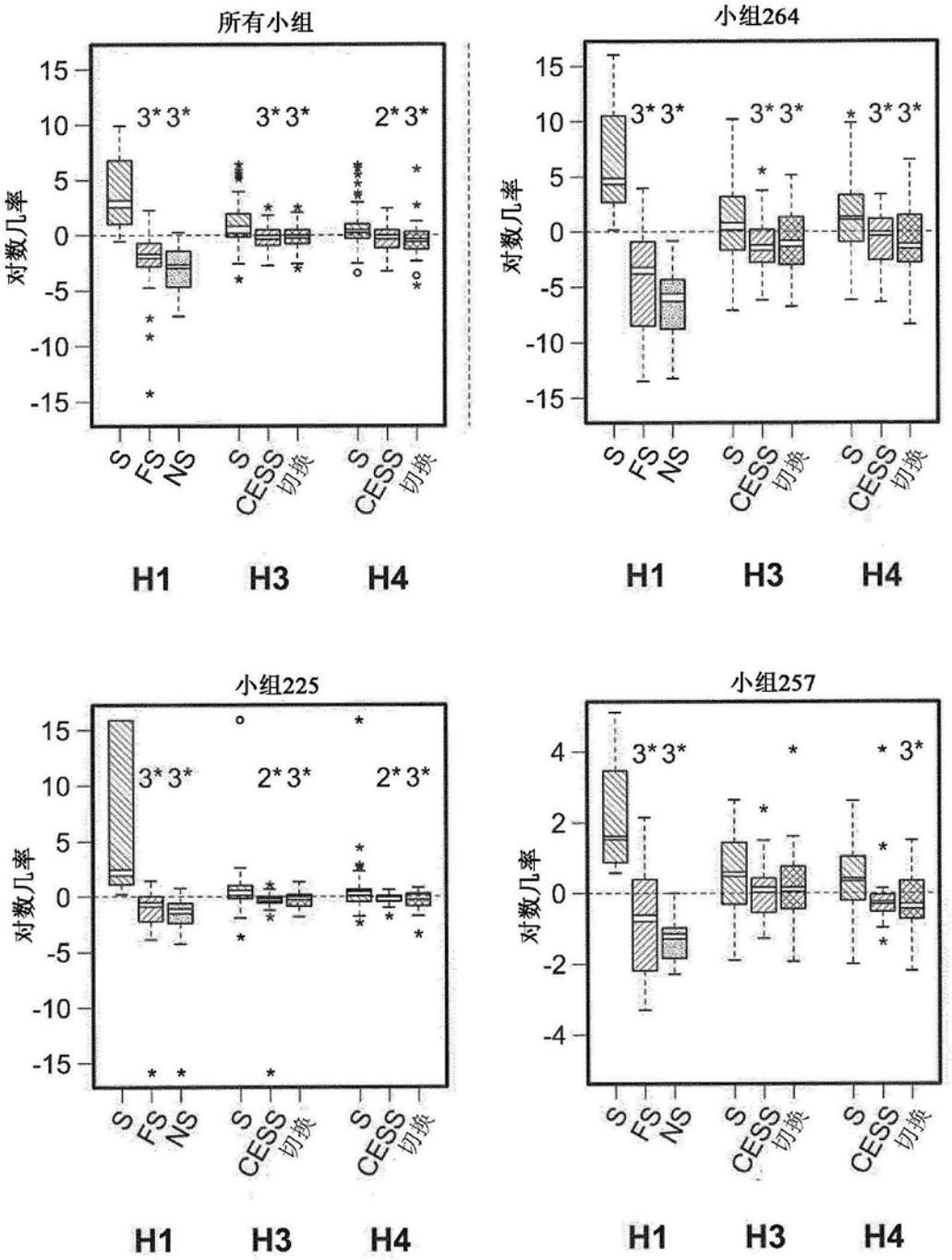

图11A

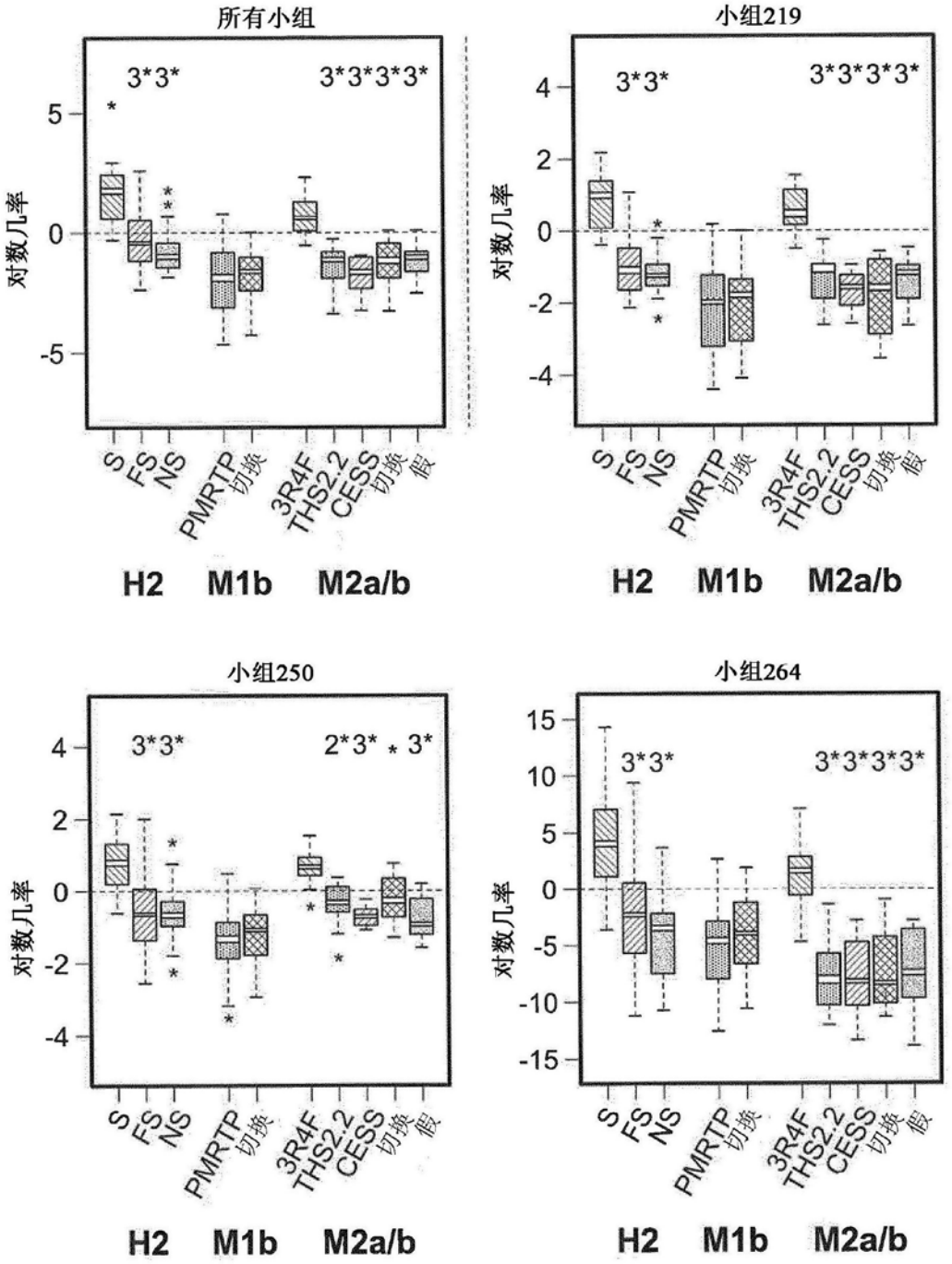
所有小组
小组219
小组250
小组264
对数几率
3*3*
3*3*3*3*
2*3*
S
FS
NS
PMRTP
切换
3R4F
THS2.2
CESS
切换
假
H2
M1b
M2a/b

图11B

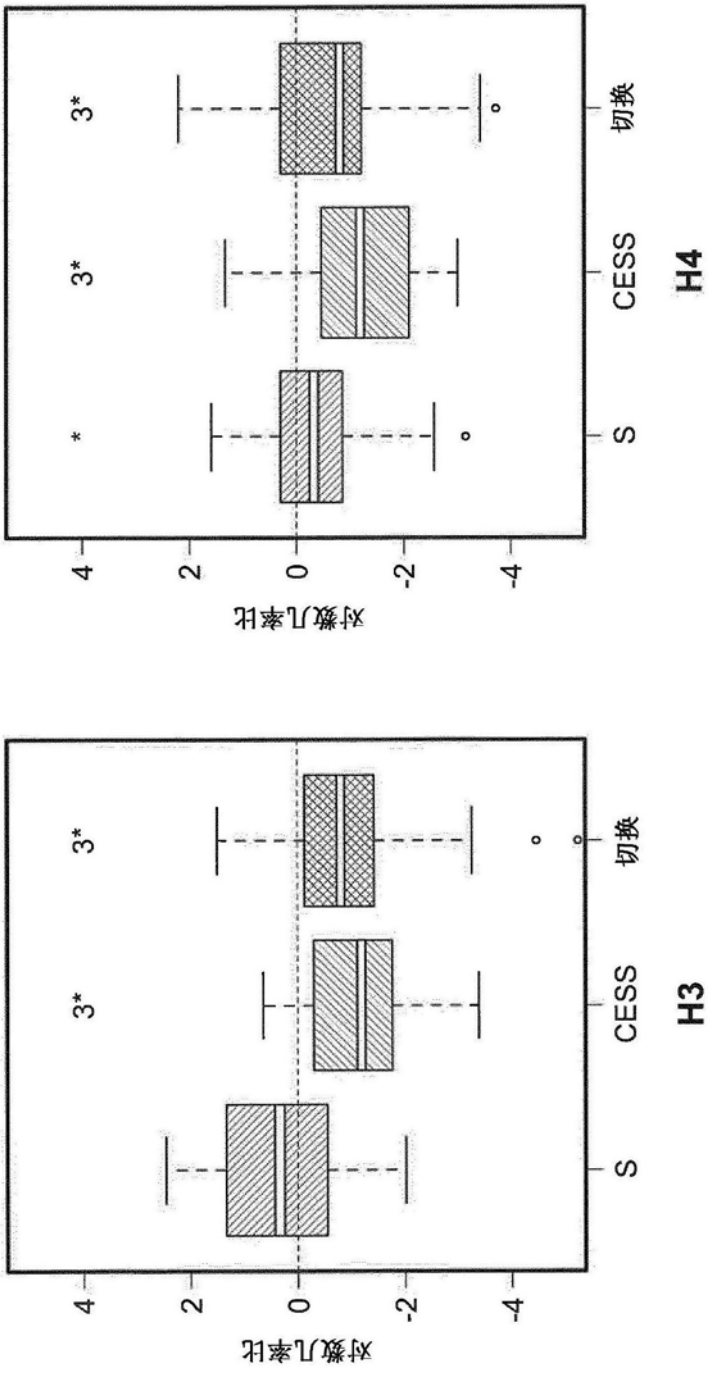

图12

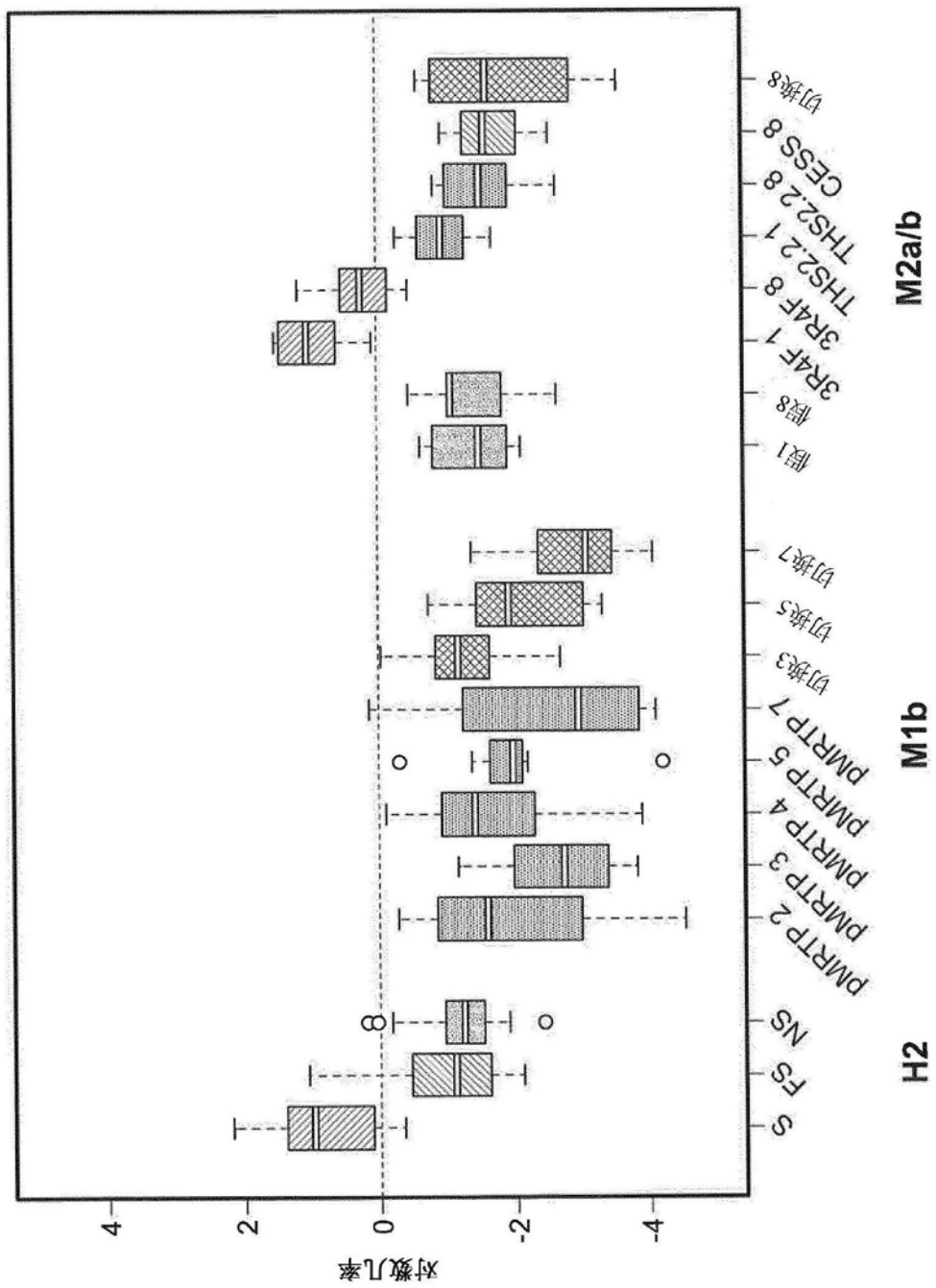
对数几率
4
2
0
-2
-4
S
FS
NS
H2
PMRTP 2
PMRTP 3
PMRTP 4
PMRTP 5
PMRTP 7
切换3
切换5
切换7
M1b
假1
假8
3R4F 1
3R4F 8
TH52.2 1
TH52.2 8
CESS 8
切换8
M2a/b

图13

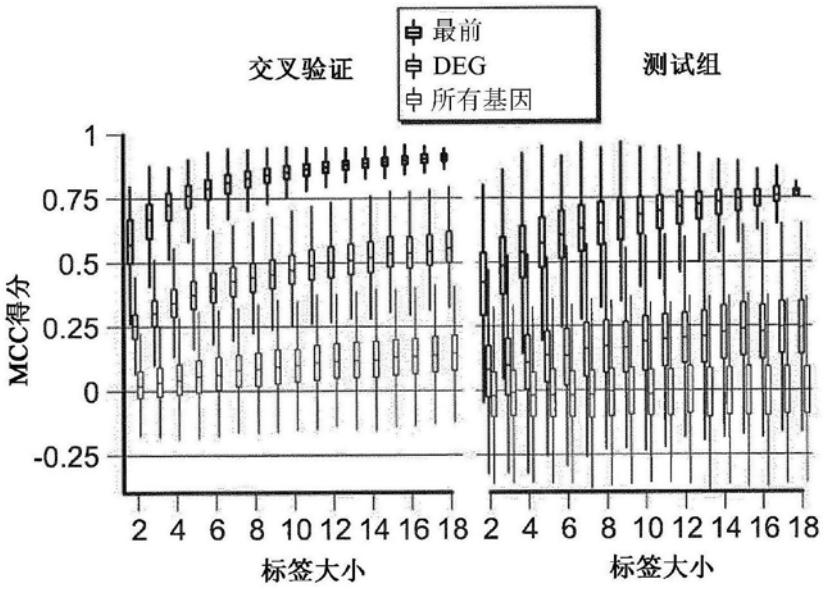
交叉验证
测试组
最前
DEG
所有基因
MCC得分
1
0.75
0.5
0.25
0
-0.25
2 4 6 8 10 12 14 16 18
标签大小
2 4 6 8 10 12 14 16 18
标签大小

图14A

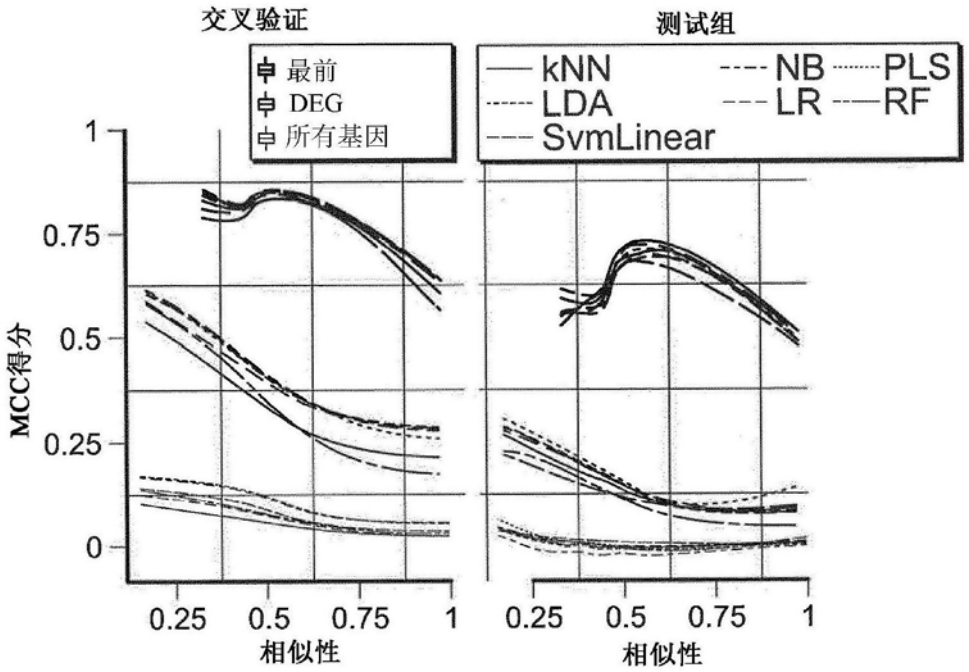
交叉验证
测试组
最前
DEG
所有基因
kNN
LDA
SvmLinear
NB
LR
PLS
RF
MCC得分
1
0.75
0.5
0.25
0
0.25 0.5 0.75 1
相似性
0.25 0.5 0.75 1
相似性

图14B

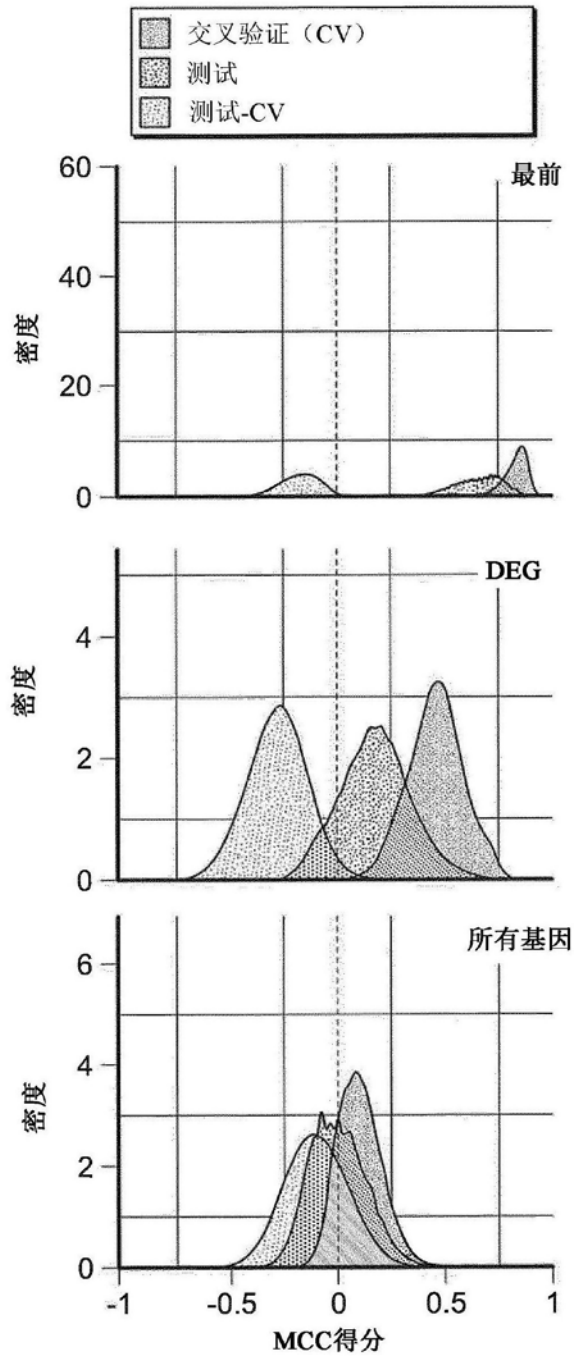

图14C

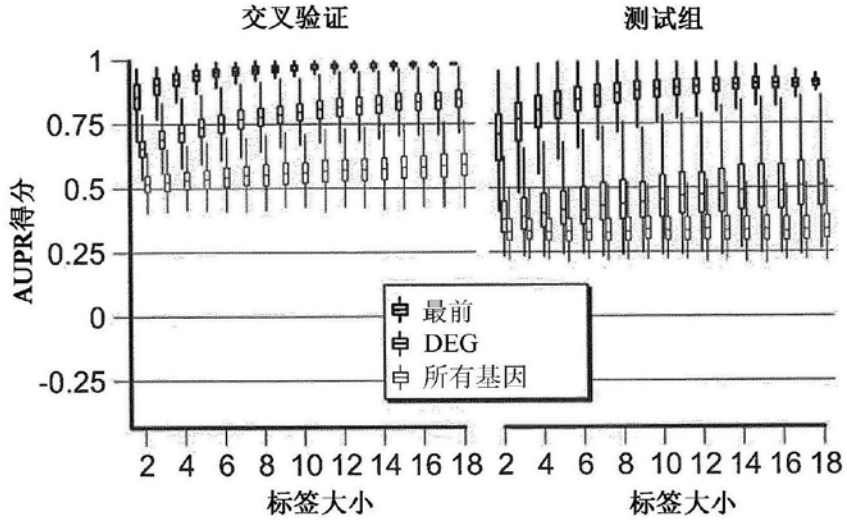
交叉验证
测试组
AUPR得分
1
0.75
0.5
0.25
0
-0.25
最前
DEG
所有基因
2 4 6 8 10 12 14 16 18
标签大小

图15A

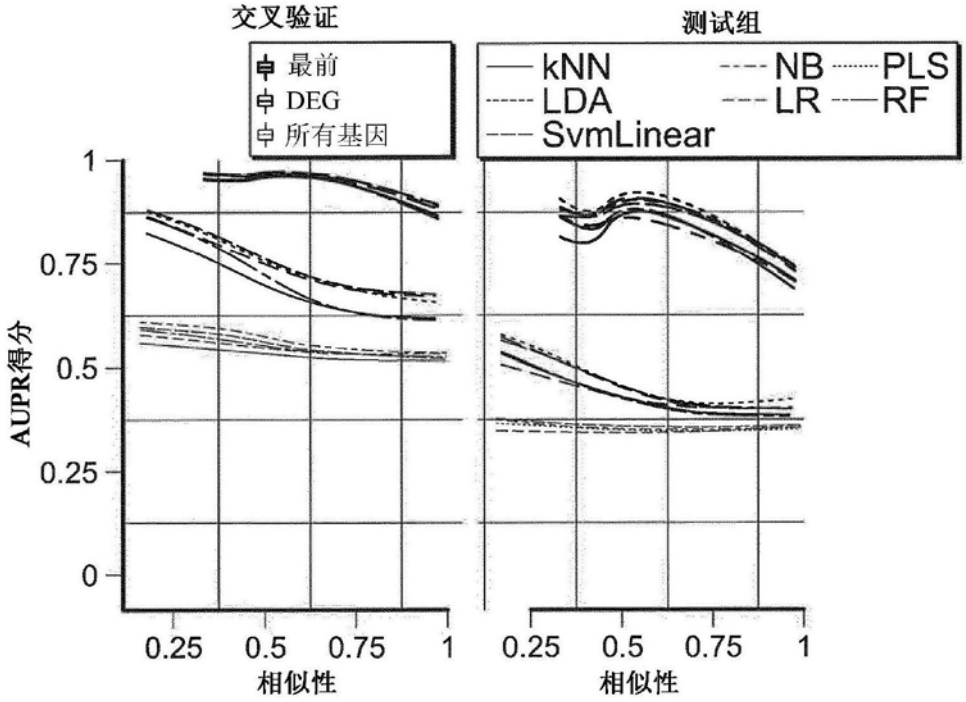
交叉验证
测试组
最前
DEG
所有基因
kNN
NB
PLS
LDA
LR
RF
SvmLinear
AUPR得分
1
0.75
0.5
0.25
0
0.25 0.5 0.75 1
相似性

图15B

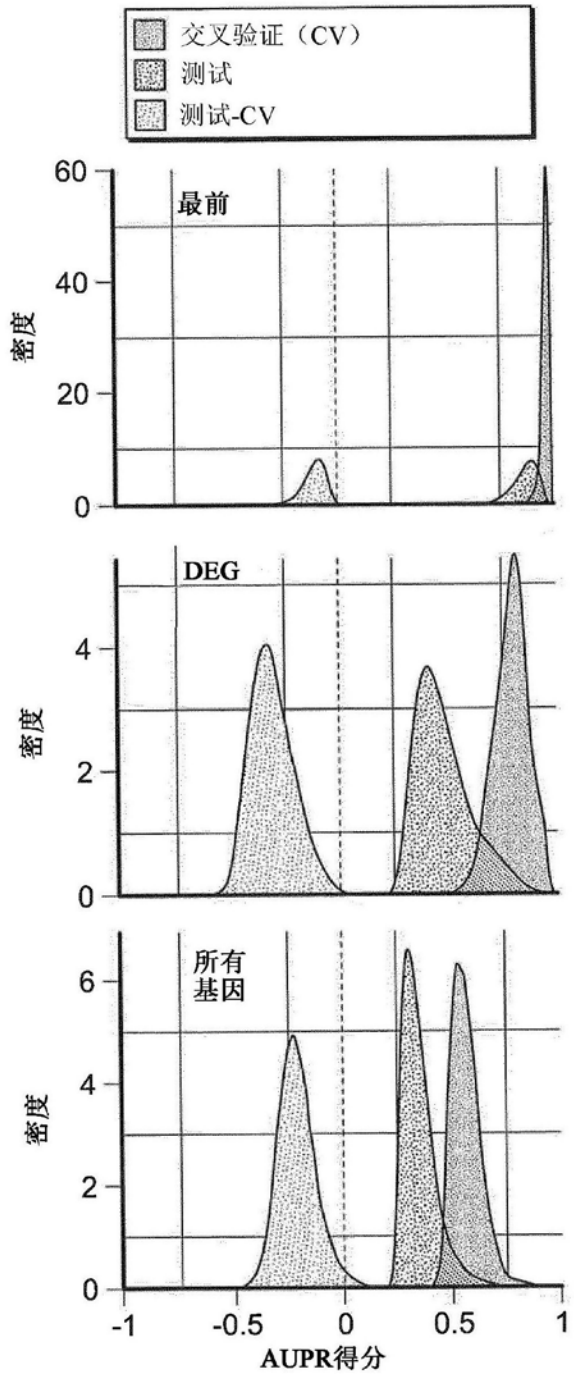

图15C