(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 109689892 A
(43)申请公布日 2019.04.26

(21)申请号 201780055050.1
(22)申请日 2017.09.06
(30)优先权数据
62/384,609 2016.09.07 US
62/528,899 2017.07.05 US
(85)PCT国际申请进入国家阶段日
2019.03.07
(86)PCT国际申请的申请数据
PCT/US2017/050358 2017.09.06
(87)PCT国际申请的公布数据
WO2018/048960 EN 2018.03.15
(71)申请人 威拉赛特公司
地址 美国加利福尼亚州
(72)发明人 G·C·肯尼迪 黄静 Y·崔
丹尼尔·潘克拉茨
帕特里克·肖恩·沃尔什
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 贺淑东
(51)Int.Cl.
C12Q 1/6883(2018.01)
G16B 40/00(2019.01)

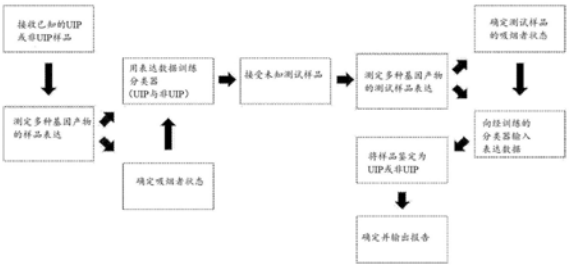
权利要求书7页 说明书99页 附图33页

(54)发明名称

用于检测寻常型间质性肺炎的方法和系统

(57)摘要

本公开内容提供了用于区分寻常型间质性肺炎(UIP)样品与非UIP样品的系统、方法和分类器。



CN 109689892 A

1.一种检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:

(a)测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表1和/或表15中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表1和/或表15中列出的任一基因的一个或多个序列;以及

(b)将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(i)对应于所述第一组的表达水平的升高和/或(ii)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(i)对应于所述第二组的表达水平的升高和/或(ii)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

2.一种检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:

(a)通过测序、阵列杂交或核酸扩增来测定来自受试者肺组织的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表1和/或表15中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表1和/或表15任一者中列出的任一基因的一个或多个序列;以及

(b)将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(i)对应于所述第一组的表达水平的升高和/或(ii)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(i)对应于所述第二组的表达水平的升高和/或(ii)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

3.一种检测肺组织样品是UIP阳性还是非UIP阳性的方法,包括:

(a)测定在测试样品中表达的两种或更多种转录物的表达水平;以及

(b)使用计算机生成的分类器将所述样品分类为UIP或非UIP;

其中使用包括HP、NSIP、结节病、RB、细支气管炎和机化性肺炎(OP)在内的非UIP病理学亚型的非均质谱训练所述分类器;并且

其中在所述测试样品中表达的所述两种或更多种转录物选自表1和/或表15中列出的两个或更多个基因的转录物,或SEQ ID NO:1-320中的任两个或更多个。

4.根据前述权利要求中的任一项所述的方法,其中所述测试样品是从所述受试者获得的多个样品的池。

5.根据权利要求1-3中的任一项所述的方法,其中所述方法包括汇集从所述受试者获得的多个单独样品的表达水平数据。

6.根据权利要求1-3中的任一项所述的方法,其包括在测定所述表达水平之前从cDNA合成双链cDNA。

7.根据权利要求1-3中的任一项所述的方法,其包括在测定所述表达水平之前从所述双链cDNA合成非天然RNA。

8.根据权利要求1-2中的任一项所述的方法,其进一步包括使用吸烟状态作为(1)或(2)的分类步骤的协变量。

9.根据权利要求8所述的方法,其中通过检测指示所述受试者的吸烟者状态的表达谱来确定吸烟状态。

10.根据前述权利要求中的任一项所述的方法,其中所述样品的分类包括检测对吸烟者状态偏差敏感的一种或多种转录物的表达水平,并且其中与对吸烟者偏差不敏感的转录物不同地,对所述对吸烟者状态偏差敏感的转录物进行加权。

11.根据前述权利要求中的任一项所述的方法,其中所述样品的分类包括检测对吸烟者状态偏差敏感的一种或多种转录物的表达水平,并且其中从所述分类步骤中排除所述对吸烟者状态偏差敏感的转录物。

12.根据前述权利要求中的任一项所述的方法,其中所述分类步骤进一步包括检测所述测试样品中的序列变体,并将所述序列变体与参考样品中的相应序列进行比较,以将所述样品分类为UIP或非UIP。

13.根据前述权利要求中的任一项所述的方法,其中用于将所述样品分类为UIP或非UIP的表达数据包括选自SEQ ID NO:1-320的基因的至少两种转录物的表达数据。

14.根据权利要求1-3中的任一项所述的方法,其进一步包括(i)从所述受试者获得样品,(ii)使所述样品的第一部分经受细胞学分析,该细胞学分析指示所述样品的所述第一部分是模糊的或不确定的,以及(iii)将所述样品的第二部分作为测试样品进行测定。

15.根据权利要求14所述的方法,其中所述第一部分和所述第二部分是不同的部分。

16.根据权利要求14所述的方法,其中所述第一部分和所述第二部分是相同的部分。

17.根据权利要求1或2中的任一项所述的方法,其中使用经多个样品训练的经训练算法来执行(b),其中所述测试样品独立于所述多个样品。

18.一种治疗特发性肺纤维化(IPF)未确诊的受试者的方法,包括,

(a)通过阵列、测序或qRT-PCR来测量从受试者的气道获得的一个或多个样品中至少两个基因的表达水平,其中所述基因选自表1和/或表15中列出的基因,并且其中所述方法包括:

(i)在所述测量步骤之前汇集至少两个样品;

(ii)汇集从两个单独的样品独立地测量的至少两个表达数据集;或

(iii)(i)与(ii)的组合;

(b)如果存在以下情况,则施用有效治疗IPF的化合物:

(i)相比于相应转录物的参考表达水平,所述至少两个基因中的每一个的表达水平升高;和/或

(ii)相比于相应转录物的参考表达水平,所述至少两个基因中的每一个的表达水平降低;和/或

(iii)相比于相应转录物的参考表达水平,所述至少两个基因中的至少一个的表达水平升高,并且相比于相应转录物的参考表达水平,所述至少两个基因中的至少一个的表达水平降低。

19.一种检测汇集的肺组织测试样品是UIP阳性还是非UIP阳性的方法,包括:

(a)测定在测试样品中表达的一种或多种转录物的表达水平;以及

(b)使用计算机生成的经训练的分类器将所述测试样品分类为UIP或非UIP;

其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有UIP或非UIP的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

20.根据权利要求19所述的方法,其中所述分类器训练使用在表1和/或表15中列出的一种或多种转录物的表达水平。

21.根据权利要求19所述的方法,其中所述分类器训练使用在表1和/或表15中列出的全部基因的转录物的表达水平。

22.根据权利要求19-21中的任一项所述的方法,其中所述计算机生成的经训练的分类器基于在表1和/或表15中列出的基因的一种或多种转录物的表达水平将所述测试样品分类为UIP或非UIP。

23.一种检测汇集的肺组织测试样品是否对疾病或病况呈阳性的方法,包括:

(a)测定在测试样品中表达的一种或多种转录物的表达水平;以及

(b)使用计算机生成的经训练的分类器将所述测试样品分类为对所述疾病或病况呈阳性或呈阴性;

其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有对所述疾病或病况呈阳性或阴性的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

24.根据权利要求23所述的方法,其中所述疾病或病况选自:肺部病症、肺癌、间质性肺病(ILD)、特发性肺纤维化(IPF)、非特异性间质性肺炎(NSIP)、Favor NSIP、寻常型间质性肺炎(UIP)或非寻常型间质性肺炎(非UIP)、急性肺损伤、细支气管炎、脱屑性间质性肺炎、弥漫性肺泡损伤、肺气肿、嗜酸性粒细胞性肺炎、非特异性间质性肺炎(包括细胞亚型、混合亚型或Favor亚型)、肉芽肿病、过敏性肺炎(HP)、Favor亚型过敏性肺炎(Favor HP)、机化性肺炎、肺孢子虫肺炎、肺动脉高压、呼吸性细支气管炎、肺结节病、吸烟相关的间质纤维化、慢性阻塞性肺病(COPD)、烟雾暴露史、长期烟雾暴露、短期烟雾暴露以及慢性间质纤维化。

25.一种用有效治疗特发性肺纤维化(IPF)的治疗剂治疗有需要的受试者的方法,包括,

向所述有需要的受试者施用有效剂量的有效治疗IPF的化合物;其中如通过计算机生成的经训练的分类器所确定的,所述有需要的受试者的表1和/或表15中一个或多个基因的表达水平表明所述受试者需要IPF治疗。

26.根据权利要求25所述的方法,其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有UIP或非UIP的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

27.根据权利要求25所述的方法,其中所述计算机生成的经训练的分类器将从所述受试者获得的样品鉴定为UIP。

28.根据权利要求25所述的方法,其中所述计算机生成的经训练的分类器将从所述受

试者获得的样品鉴定为IPF。

29.一种检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:

(a)测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表5中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表5中列出的任一基因的一个或多个序列;以及

(b)将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(i)对应于所述第一组的表达水平的升高和/或(ii)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(i)对应于所述第二组的表达水平的升高和/或(ii)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

30.一种检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:

(a)通过测序、阵列杂交或核酸扩增来测定来自受试者肺组织的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表5中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表5中列出的任一基因的一个或多个序列;以及

(b)将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(i)对应于所述第一组的表达水平的升高和/或(ii)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(i)对应于所述第二组的表达水平的升高和/或(ii)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

31.一种检测肺组织样品是UIP阳性还是非UIP阳性的方法,包括:

(a)测定在测试样品中表达的两个或更多个基因的表达水平;以及

(b)使用计算机生成的分类器将所述样品分类为UIP或非UIP;

其中使用包括HP、NSIP、结节病、RB、细支气管炎和机化性肺炎(OP)在内的非UIP病理学亚型的非均质谱训练所述分类器;并且

其中在所述测试样品中表达的所述两个或更多个基因选自在表5中列出的任两个或更多个基因。

32.根据权利要求29-31中的任一项所述的方法,其中所述测试样品是从所述受试者获得的多个样品的池。

33.根据权利要求29-31中的任一项所述的方法,其中所述方法包括汇集从所述受试者获得的多个单独样品的表达水平数据。

34.根据权利要求29-33中的任一项所述的方法,其中所述测试样品是活检样品或支气管肺泡灌洗样品。

35.根据权利要求29-34中的任一项所述的方法,其中所述活检样品是经支气管活检样

品。

36.根据前述权利要求中的任一项所述的方法，其中使用qRT-PCR、DNA微阵列杂交、RNAseq或其组合来完成测定所述表达水平。

37.根据权利要求29-36中的任一项所述的方法，其包括在测定所述表达水平之前从在所述测试样品中表达的RNA合成cDNA。

38.根据权利要求37所述的方法，其包括在测定所述表达水平之前从所述cDNA合成双链cDNA。

39.根据权利要求38所述的方法，其包括在测定所述表达水平之前从所述双链cDNA合成非天然RNA。

40.根据权利要求29-36中的任一项所述的方法，其包括在测定所述表达水平之前扩增所述核苷酸。

41.根据权利要求29-36中的任一项所述的方法，其中所述转录物中的一种或多种被标记。

42.根据权利要求29-32中的任一项所述的方法，其进一步包括测量所述测试样品中的至少一种对照核酸的表达水平。

43.根据权利要求29-32中的任一项所述的方法，其中所述肺组织被分类为间质性肺病(ILD)、特定类型的ILD、非ILD或非诊断性中的任一者。

44.根据权利要求29-30中的任一项所述的方法，其进一步包括使用吸烟状态作为(1)或(2)的分类步骤的协变量。

45.根据权利要求44所述的方法，其中通过检测指示所述受试者的吸烟者状态的表达谱来确定吸烟状态。

46.根据权利要求3或31所述的方法，其进一步包括使用吸烟状态作为所述分类步骤的协变量。

47.根据权利要求9、44或46中的任一项所述的方法，其中所述方法在所述分类步骤之前使用吸烟状态作为协变量。

48.根据前述权利要求中的任一项所述的方法，其包括执行使用选自基因表达、变体、突变、融合、杂合性丢失(LOH)和生物学途径效应的一个或多个特征训练的分类器。

49.根据前述权利要求中的任一项所述的方法，其中所述分类具有至少约90％的特异性和至少约70％的灵敏度。

50.根据前述权利要求中的任一项所述的方法，其中用于将所述样品分类为UIP或非UIP的表达数据包括与选自表5所列基因的基因相对应的至少两种转录物的表达数据。

51.根据权利要求50所述的方法，其中所使用的表达数据包括表5中列出的每一个基因。

52.一种治疗特发性肺纤维化(IPF)未确诊的受试者的方法，包括，

(a)通过阵列、测序或qRT-PCR来测量从受试者的气道获得的一个或多个样品中至少两个基因的表达水平，其中所述基因选自表5中列出的基因，并且其中所述方法包括：

(i)在所述测量步骤之前物理汇集至少两个样品；

(ii)汇集从两个单独的样品独立地测量的至少两个表达数据集；或

(iii)(i)与(ii)的组合；

(b)如果存在以下情况,则施用有效治疗IPF的化合物:

(i)相比于相应转录物的参考表达水平,所述至少两个基因中的每一个的表达水平升高;和/或

(ii)相比于相应转录物的参考表达水平,所述至少两个基因中的每一个的表达水平降低;和/或

(iii)相比于相应转录物的参考表达水平,所述至少两个基因中的至少一个的表达水平升高,并且相比于相应转录物的参考表达水平,所述至少两个基因中的至少一个的表达水平降低。

53.根据权利要求52所述的方法,其中所述施用步骤仅在(i)中的所述升高和/或(ii)中的所述降低显著时进行。

54.一种检测汇集的肺组织测试样品是UIP阳性还是非UIP阳性的方法,包括:

(a)测定在测试样品中表达的一种或多种转录物的表达水平;以及

(b)使用计算机生成的经训练的分类器将所述测试样品分类为UIP或非UIP;

其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有UIP或非UIP的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

55.根据权利要求54所述的方法,其中所述分类器训练使用在表5中列出的一个或多个基因的表达水平。

56.根据权利要求54-55中的任一项所述的方法,其中所述分类器训练使用在表5中列出的所有基因的表达水平。

57.根据权利要求54-56中的任一项所述的方法,其中所述计算机生成的经训练的分类器基于在表5中列出的一个或多个基因的表达水平将所述测试样品分类为UIP或非UIP。

58.根据权利要求57所述的方法,其中所述分类器基于在表5中列出的所有基因的转录物的表达水平将所述测试样品分类为UIP或非UIP。

59.一种检测汇集的肺组织测试样品是否对疾病或病况呈阳性的方法,包括:

(a)测定在测试样品中表达的一种或多种转录物的表达水平;以及

(b)使用计算机生成的经训练的分类器将所述测试样品分类为对所述疾病或病况呈阳性或呈阴性;

其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有对所述疾病或病况呈阳性或阴性的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

60.一种用有效治疗特发性肺纤维化(IPF)的治疗剂治疗有需要的受试者的方法,包括,

向所述有需要的受试者施用有效剂量的有效治疗IPF的化合物;其中如通过计算机生成的经训练的分类器所确定的,所述有需要的受试者的表5中一个或多个基因的表达水平表明所述受试者需要IPF治疗。

61.根据权利要求60所述的方法,其中使用在从多个受试者获得的多个单独的训练样

品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器，每个训练样品具有UIP或非UIP的确认诊断，其中至少两个所述训练样品从单个受试者获得；并且其中在所述分类之前汇集所述测试样品。

62.根据权利要求61所述的方法，其中所述计算机生成的经训练的分类器将从所述受试者获得的样品鉴定为UIP。

63.根据权利要求61所述的方法，其中所述计算机生成的经训练的分类器将从所述受试者获得的样品鉴定为IPF。

64.一种鉴定受试者是否对肺部病症呈阳性的方法，包括：

(a)获得所述受试者的组织样品；

(b)使所述组织样品的第一部分经受细胞学测试，该细胞学测试指示所述第一部分是模糊的或可疑的；

(c)在鉴定所述第一部分是模糊的或可疑的之后，测定所述组织样品的第二部分中与所述肺部病症相关的一种或多种标志物的表达水平；

(d)用经训练的算法处理所述表达水平，从而以至少约90％的准确性生成所述组织样品对所述肺部病症呈阳性的分类，其中所述经训练的算法用包含多个训练样品的训练集来训练，并且其中所述组织样品独立于所述多个训练样品；以及

(e)电子地输出所述分类，由此鉴定所述受试者是否对所述肺部病症呈阳性。

65.根据权利要求64所述的方法，其中所述组织样品是肺组织样品。

66.根据权利要求64所述的方法，其中所述组织样品是非肺组织样品。

67.根据权利要求66所述的方法，其中所述非肺组织样品是呼吸上皮样品。

68.根据权利要求67所述的方法，其中所述呼吸上皮样品来自所述受试者的鼻或口。

69.根据权利要求64所述的方法，其中所述表达水平是与UIP相关的多种标志物的表达水平。

70.根据权利要求64所述的方法，其中所述准确性为至少约95％。

71.根据权利要求64所述的方法，其中所述分类以至少约90％的特异性生成。

72.根据权利要求64所述的方法，其中所述分类以至少约70％的灵敏度生成。

73.根据权利要求64所述的方法，其中所述经训练的算法被配置用于在至少100个独立的测试样品中以至少约90％的准确性将组织样品分类。

74.根据权利要求64所述的方法，其中所述分类电子地输出在用户的电子显示器的图形用户界面上。

75.根据权利要求64所述的方法，其中所述肺部病症是寻常型间质性肺炎(UIP)或非寻常型间质性肺炎(非UIP)。

76.根据权利要求64所述的方法，其中所述第一部分不同于所述第二部分。

用于检测寻常型间质性肺炎的方法和系统

交叉引用

[0001] 本申请要求于2016年9月7日提交的美国临时专利申请序列号62/384,609和于2017年7月5日提交的美国临时专利申请序列号62/528,899的优先权，二者通过引用整体并入本文。

发明背景

[0002] 间质性肺病(ILD)是一组非均一的急性和慢性双侧实质肺病，其具有类似的临床表现，但严重性和后果较为宽泛，包括不同的疾病进展、治疗反应和存活期[1]。其中，特发性肺纤维化(IPF)是一种最常见(每年北美发病率为万分之1.4至万分之6)且最严重的ILD，其特征在于进行性纤维化、肺功能恶化和死亡[3-6]。在适当的临床环境中，根据在HRCT和/或SLB上存在寻常型间质性肺炎(UIP)模式来确定IPF[8]。较长的诊断时间加上疾病的快速进程迫使人们需要新的工具来尽量减少患者在未确诊期间的痛苦。大多数被诊断为患有IPF的患者在其初次诊断后的五年内死亡[7,8]。然而，最近可用的已在稳定IPF疾病进展方面显示出前景的两种新抗纤维化药物吡非尼酮(pirfenidone)和尼达尼布(nintedanib)，以及开发中的其他治疗剂，可能会改变这种情况[9-11]，因此准确的诊断对于适当的治疗干预至关重要[5,12]。

[0003] 鉴于这些用肺移植和/或抗纤维化口服化合物进行治疗的新可能性，区分IPF与其他纤维化IIP的诊断具有重要意义[2]。此外，许多经常与IPF混淆的病症用免疫抑制剂进行治疗。由于联合免疫抑制治疗IPF已被证明是有害的，因此选择正确的治疗方法至关重要[2,33]。

[0004] IPF可能难以诊断。国际公认的指南建议在ILD的诊断和管理中对临床、放射学和病理疾病特征进行多学科评价。IPF的诊断方法需要排除其他间质性肺炎，以及结缔组织疾病以及环境和职业暴露[3-6]。疑似患有IPF的患者通常接受胸部的高分辨率计算机断层扫描(HRCT)，但这只有在寻常型间质性肺炎(UIP)模式非常明显时才能以高特异性确认该疾病[5,13]。因此，对于大约三分之一的ILD患者，可以在不进行SLB的情况下实现对IPF的可靠诊断[34-36]。对于那些在HRCT上没有可靠的UIP模式诊断的患者(例如，可能UIP，以及很可能UIP的工作类别)，据估计存在组织学UIP的阳性预测值(PPV)约为60％[35,36]，这个水平被认为仍需进行SLB的确认[8]。因此，由于HRCT结果常常是非决定性的，大量患者需要进行侵入性诊断性外科肺活检(SLB)来阐明间质性肺炎和/或UIP模式的组织病理学特征[5,14]，而从症状发作到诊断IPF的典型时长可能为1-2年[15]。由于针对冷冻活检报告的高手术并发症发生率[37]，以及与SLB相关的院内和90天死亡率分别达到1.7％和3.9％[38]，本领域非常需要一种侵入性较小的诊断IPF的方法。

[0005] 在经支气管活检(TBB)中可靠鉴定UIP病理学的难点在于对肺泡状肺实质的充分采样和非均一的疾病分布。病理学家之间会出现不一致，并且正确的诊断可能取决于个体的经验[16]。尽管进行了组织病理学评价，可能仍然难以得到明确的诊断。在具有高TBB采样充分率的回顾性研究中，具有与UIP一致的临床和影像学特征的患者中有30％-43％被确认为UIP[11,12]，而第三项研究报告确认率<10％[13]。这导致许多人去评价可能提供更大肺泡采样的替代支气管镜研究[14,15]。目前这些研究受到可用性和缺乏大型多中心研究的限制[16]。当肺病

学家、放射科医师和病理学家的多学科团队(MDT)合议时,诊断准确性有所提高[17];但遗憾的是,并非所有患者及其医师都能得到由经验丰富的MDT进行的这一级别的专家评审。这种评审非常耗时,并且需要患者前往具有公认专业知识的地区中心。

[0006] 因此,需要更有效的诊断IPF的方法,例如,在支气管镜采样中检测UIP但不依赖于肺泡的充分采样的更稳定的方法。此外,还需要区分UIP与非UIP的方法。

[0007] 虽然科技文献中的基因表达谱分析研究已经报道了IPF与其他ILD亚型之间的差异表达[18,19],但是除了我们的在先申请PCT/US2015/059309(通过引用整体并入本文)之外,没有人尝试在包含常常作为临床医生区分诊断的一部分而存在的其他亚型的数据集中进行UIP分类。此外,没有人利用过实际或计算机模拟的样品汇集来实现差异诊断的更高灵敏度和/或特异性。此外,没有人报告过不受细胞异质性影响的分类器。

[0008] 本文描述的方法令人惊讶地能够通过利用患者样品的物理或计算机汇集来获得区分诊断的更高特异性和/或灵敏度。此外,尽管现有技术表明需要细胞同质性,但本文所述的方法令人惊讶地不受细胞异质性的影响。因此,本公开内容提供了对现有技术的显著改进,用于使用差异基因表达来区分IPF与其他ILD亚型。

发明内容

[0009] 本公开内容提供了使用分类器来区分寻常型间质性肺炎(UIP)样品与非UIP样品的方法和系统。已使用专家病理学诊断作为真值标签确认了本文所述的方法的准确性。因此,本文所述的方法提供了病理学替代测试,其准确地区分样品例如经支气管活检物(TBB)中的UIP模式与非UIP模式。

[0010] 在一些实施方案中,本公开内容提供了用于检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法和/或系统。在一些实施方案中,提供了用于确定肺组织样品是否是寻常型间质性肺炎(UIP)阳性的方法,包括检测表1、表5、表15或其组合中列出的一个或多个基因在生物样品中的mRNA表达水平。在特定实施方案中,本公开内容提供了用于检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法和/或系统,包括检测表5中列出的一个或多个基因在生物样品中的mRNA表达水平。在一些实施方案中,所述方法包括检测表5中列出的所有基因。在一些实施方案中,所述方法进一步包括将上文确定的表达水平(例如,表5中列出的一个或多个基因的表达水平)转化成UIP评分,该UIP评分指示受试者患有IPF(例如,相对于另一ILD)的可能性。在一些实施方案中,根据具有大于70%的阴性预测值(NPV)的模型来确定风险评分,以排除UIP。在一些实施方案中,根据具有大于80%的阳性预测值(NPV)的模型来确定风险评分,以诊断UIP。在一些实施方案中,提供了一种方法,其用于:测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括在UIP中超表达并在表1和/或表15任一者中列出的任一个或多个基因,而所述第二组转录物包括在UIP中低表达并在表1和/或表15任一者中列出的任一个或多个基因。在一些实施方案中,提供了一种方法,其用于:测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括在UIP中超表达并在表5中列出的任一个或多个基因,而所述第二组转录物包括在UIP中低表达并在表5中列出的任一个或多个基因。在一些实施方案中,所述方法进一步提供将所述第一组转录物和所述第二组转录物中每一

者的表达水平与相应转录物的参考表达水平进行比较，以便(1)如果相比于所述参考表达水平，存在(a)对应于所述第一组的表达水平的升高或(b)对应于所述第二组的表达水平的降低，则将肺组织分类为寻常型间质性肺炎(UIP)，或者(2)如果相比于所述参考表达水平，存在(c)对应于所述第二组的表达水平的升高或(d)对应于所述第一组的表达水平的降低，则将肺组织分类为非寻常型间质性肺炎(非UIP)。在一些实施方案中，所述方法进一步提供对在表1和/或表15中列出的所述一个或多个基因中的任一者的序列变体进行确定和/或比较。在一些实施方案中，所述方法提供对在表5中列出的所述一个或多个基因中的任一者的序列变体进行确定和/或比较。

[0011] 在一些实施方案中，本公开内容提供了检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法，包括：测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平，其中所述第一组转录物包括对应于在UIP中超表达并在表1和/或表15中列出的任一基因的一个或多个序列，而所述第二组转录物包括对应于在UIP中低表达并在表1和/或表15中列出的任一基因的一个或多个序列；以及将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较，以便(1)如果相比于所述参考表达水平，存在(a)对应于所述第一组的表达水平的升高和/或(b)对应于所述第二组的表达水平的降低，则将所述肺组织分类为寻常型间质性肺炎(UIP)，或者(2)如果相比于所述参考表达水平，存在(c)对应于所述第二组的表达水平的升高和/或(d)对应于所述第一组的表达水平的降低，则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

[0012] 在一些实施方案中，本公开内容提供了检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法，包括：测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平，其中所述第一组转录物包括对应于在UIP中超表达并在表5中列出的任一基因的一个或多个序列，而所述第二组转录物包括对应于在UIP中低表达并在表5中列出的任一基因的一个或多个序列；以及将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较，以便(1)如果相比于所述参考表达水平，存在(a)对应于所述第一组的表达水平的升高和/或(b)对应于所述第二组的表达水平的降低，则将所述肺组织分类为寻常型间质性肺炎(UIP)，或者(2)如果相比于所述参考表达水平，存在(c)对应于所述第二组的表达水平的升高和/或(d)对应于所述第一组的表达水平的降低，则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。在一些实施方案中，本公开内容提供了检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法，包括：测定受试者的测试样品中第一组转录物和第二组转录物中每一者的表达水平，其中所述第一组转录物包括对应于在UIP中超表达并在表1、表5和/或表15中列出的任一基因的一个或多个序列，而所述第二组转录物包括对应于在UIP中低表达并在表1、表5和/或表15中列出的任一基因的一个或多个序列；以及将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较，以便(1)如果相比于所述参考表达水平，存在(a)对应于所述第一组的表达水平的升高和/或(b)对应于所述第二组的表达水平的降低，则将所述肺组织分类为寻常型间质性肺炎(UIP)，或者(2)如果相比于所述参考表达水平，存在(c)对应于所述第二组的表达水平的不变或升高和/或(d)对应于所述第一组的表达水平的不变或降低，则将所述肺

组织分类为非寻常型间质性肺炎(非UIP)。

[0013] 在一些实施方案中,本公开内容提供了检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:通过测序、阵列杂交或核酸扩增来测定来自受试者肺组织的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表1和/或表15中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表1和/或表15任一者中列出的任一基因的一个或多个序列;以及将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(a)对应于所述第一组的表达水平的升高和/或(b)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(c)对应于所述第二组的表达水平的升高和/或(d)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

[0014] 在一些实施方案中,本公开内容提供了检测肺组织样品是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法,包括:通过测序、阵列杂交或核酸扩增来测定来自受试者肺组织的测试样品中第一组转录物和第二组转录物中每一者的表达水平,其中所述第一组转录物包括对应于在UIP中超表达并在表5中列出的任一基因的一个或多个序列,而所述第二组转录物包括对应于在UIP中低表达并在表5中列出的任一基因的一个或多个序列;以及将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较,以便(1)如果相比于所述参考表达水平,存在(a)对应于所述第一组的表达水平的升高和/或(b)对应于所述第二组的表达水平的降低,则将所述肺组织分类为寻常型间质性肺炎(UIP),或者(2)如果相比于所述参考表达水平,存在(c)对应于所述第二组的表达水平的升高和/或(d)对应于所述第一组的表达水平的降低,则将所述肺组织分类为非寻常型间质性肺炎(非UIP)。

[0015] 在一些实施方案中,所述第一组包括2种或更多种不同的转录物,或3种或更多种、4种或更多种、5种或更多种、10种或更多种、15种或更多种、20种或更多种,或超过20种不同的转录物。

[0016] 在一些实施方案中,所述第二组包括2种或更多种不同的转录物,或3种或更多种、4种或更多种、5种或更多种、10种或更多种、15种或更多种、20种或更多种,或超过20种不同的转录物。

[0017] 在一些实施方案中,本公开内容提供了检测肺组织样品是UIP阳性还是非UIP阳性的方法,包括:测定在测试样品中表达的两种或更多种转录物的表达水平;以及使用计算机生成的分类器将所述样品分类为UIP或非UIP;其中使用包括HP、NSIP、结节病、RB、细支气管炎和机化性肺炎(OP)在内的非UIP病理学亚型的非均质谱训练所述分类器;并且其中在所述测试样品中表达的所述两种或更多种转录物选自表1和/或表15中列出的任两个或更多个序列,或SEQ ID NO:1-151中的任两个或更多个。

[0018] 在一些实施方案中,本公开内容提供了检测肺组织样品是UIP阳性还是非UIP阳性的方法,包括:测定在测试样品中表达的两种或更多种转录物的表达水平;以及使用计算机生成的分类器将所述样品分类为UIP或非UIP;其中使用包括HP、NSIP、结节病、RB、细支气管炎和机化性肺炎(OP)在内的非UIP病理学亚型的非均质谱训练所述分类器;并且其中在所

述测试样品中表达的所述两种或更多种转录物选自表5中列出的任两个或更多个序列。

[0019] 在一些实施方案中，所述测试样品是从所述受试者获得的多个样品的池。在一些实施方案中，所述池包括2、3、4或5个从所述受试者获得的样品。

[0020] 在一些实施方案中，所述方法包括汇集从所述受试者获得的多个单独样品的表达水平数据。在一些实施方案中，汇集来自2、3、4或5个从所述受试者获得的样品的表达水平数据。

[0021] 在一些实施方案中，所述测试样品是活检样品或支气管肺泡灌洗样品。在一些实施方案中，所述活检样品是经支气管活检样品。在一些实施方案中，所述测试样品是新鲜冷冻或固定的。

[0022] 在一些实施方案中，使用RT-PCR、DNA微阵列杂交、RNASeq或其组合来完成测定所述表达水平。在一些实施方案中，所述表达水平通过检测所述测试样品中表达的核苷酸来测定，或从所述测试样品中表达的核苷酸合成。在一些实施方案中，所述方法包括在测定所述表达水平之前从所述测试样品中表达的RNA合成cDNA。在一些实施方案中，所述方法包括在测定所述表达水平之前从cDNA合成双链cDNA。在一些实施方案中，所述方法包括在测定所述表达水平之前从所述双链cDNA合成非天然RNA。在一些实施方案中，所述非天然RNA是cDNA。在一些实施方案中，所述非天然RNA被标记。在一些实施方案中，标签包括测序衔接子或生物素分子。在一些实施方案中，所述方法包括在测定所述表达水平之前扩增所述核苷酸。

[0023] 在一些实施方案中，所述方法包括标记所述转录物中的一种或多种。在一些实施方案中，所述方法进一步包括测量所述测试样品中的至少一种对照核酸的表达水平。

[0024] 在一些实施方案中，所述方法包括将所述肺组织分类为间质性肺病(ILD)、特定类型的ILD、非ILD或非诊断性中的任一者。在一些实施方案中，所述肺组织被分类为特发性肺纤维化(IPF)或非特异性间质性肺炎(NSIP)。在一些实施方案中，所述方法包括使用吸烟状态作为所述分类步骤的协变量。在一些实施方案中，通过检测指示所述受试者的吸烟者状态的表达谱来确定吸烟状态。

[0025] 在一些实施方案中，所述样品的分类包括检测对吸烟者状态偏差敏感的一种或多种转录物的表达水平，其中与对吸烟者偏差不敏感的转录物不同地，对所述对吸烟者状态偏差敏感的转录物进行加权。

[0026] 在一些实施方案中，所述样品的分类包括检测对吸烟者状态偏差敏感的一种或多种转录物的表达水平，其中从所述分类步骤中排除所述对吸烟者状态偏差敏感的转录物。

[0027] 在一些实施方案中，所述方法包括执行使用选自基因表达、变体、突变、融合、杂合性丢失(LOH)和生物学途径效应的一个或多个特征训练的分类器。在一些实施方案中，使用包括基因表达、序列变体、突变、融合、杂合性丢失(LOH)和生物学途径效应在内的特征来训练所述分类器。

[0028] 在一些实施方案中，所述分类步骤进一步包括检测所述测试样品中的序列变体，并将所述序列变体与参考样品中的相应序列进行比较，以将所述样品分类为UIP或非UIP。

[0029] 在一些实施方案中，本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法进一步包括如果所述样品被分类为UIP，则用能够治疗IPF的化合物治疗所述受试者。在一些实施方案中，所述化合物是抗纤维化药。在一些实施方案中，所述化合物选自吡非尼

酮、尼达尼布、其药学上可接受的盐，以及它们的组合。

[0030] 在一些实施方案中，在本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法中进行的分类具有至少约90％的特异性和至少约70％的灵敏度。

[0031] 在一些实施方案中，本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法包括测定选自SEQ ID NO:1-320的至少两种转录物的表达数据。在一些实施方案中，本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法包括测定SEQ ID NO:1-320中每一者的表达数据。

[0032] 在一些实施方案中，本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法包括测定选自表5所列基因的至少两个基因的表达数据。在一些实施方案中，本文公开的用于检测肺组织样品是UIP阳性还是非UIP阳性的方法包括测定表5中列出的基因中每一者的表达数据。

[0033] 在一些实施方案中，本文公开的方法进一步包括(i)从受试者获得样品，(ii)使所述样品的第一部分经受细胞学分析，该细胞学分析指示所述样品的所述第一部分是模糊的或不确定的，以及(iii)将所述样品的第二部分作为测试样品进行测定。在一些实施方案中，所述第一部分和第二部分是不同的部分。

[0034] 在一些实施方案中，所述将所述第一组转录物和所述第二组转录物中每一者的表达水平与相应转录物的参考表达水平进行比较使用经训练的算法来进行，所述经训练的算法用多个样品来训练，其中所述测试样品独立于所述多个样品。

[0035] 在一些实施方案中，本公开内容提出了治疗特发性肺纤维化(IPF)未确诊的患者的方法，包括，(A)通过阵列、测序或qRT-PCR来测量从受试者的气道获得的一个或多个样品中至少两个基因的表达水平，其中所述基因选自表1和/或表15中列出的基因，并且其中所述方法包括(i)在所述测量步骤之前汇集至少两个样品；(ii)汇集从两个单独的样品独立地测量的至少两个表达数据集；或(i)与(ii)的组合；以及(B)如果存在以下情况，则施用有效治疗IPF的化合物：(i)相比于相应转录物的参考表达水平，所述至少两个基因中的每一个的表达水平升高；和/或(ii)相比于相应转录物的参考表达水平，所述至少两个基因中的每一个的表达水平降低；和/或(iii)相比于相应转录物的参考表达水平，所述至少两个基因中的至少一个的表达水平升高，并且相比于相应转录物的参考表达水平，所述至少两个基因中的至少一个的表达水平降低。

[0036] 在一些实施方案中，所述施用步骤仅在(i)中的所述升高和/或(ii)中的所述降低显著时进行。

[0037] 在一些实施方案中，本公开内容提出了治疗特发性肺纤维化(IPF)未确诊的患者的方法，包括，(A)通过阵列、测序或qRT-PCR来测量从受试者的气道获得的一个或多个样品中至少两个基因的表达水平，其中所述基因选自表5中列出的基因，并且其中所述方法包括(i)在所述测量步骤之前汇集至少两个样品；(ii)汇集从两个单独的样品独立地测量的至少两个表达数据集；或(i)与(ii)的组合；以及(B)如果存在以下情况，则施用有效治疗IPF的化合物：(i)相比于相应转录物的参考表达水平，所述至少两个基因中的每一个的表达水平升高；和/或(ii)相比于相应转录物的参考表达水平，所述至少两个基因中的每一个的表达水平降低；和/或(iii)相比于相应转录物的参考表达水平，所述至少两个基因中的至少一个的表达水平升高，并且相比于相应转录物的参考表达水平，所述至少两个基因中的至

少一个的表达水平降低。

[0038] 在一些实施方案中,所述施用步骤仅在(i)中的所述升高和/或(ii)中的所述降低显著时进行。

[0039] 在一些实施方案中,本公开内容提供了检测汇集的肺组织测试样品是UIP阳性还是非UIP阳性的方法,包括:(A)测定在测试样品中表达的一种或多种转录物的表达水平;以及(B)使用计算机生成的经训练的分类器将所述测试样品分类为UIP或非UIP;其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有UIP或非UIP的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

[0040] 在一些实施方案中,所述汇集包括物理汇集。在一些实施方案中,所述汇集包括计算机汇集。

[0041] 在一些实施方案中,所述分类器训练使用在表1和/或表15中列出的一种或多种转录物的表达水平。在一些实施方案中,所述分类器训练使用在表5中列出的一个或多个基因的表达水平。在一些实施方案中,所述分类器训练使用在表1中列出的所有转录物的表达水平。在一些实施方案中,所述分类器训练使用在表15中列出的所有转录物的表达水平。在一些实施方案中,所述分类器训练使用在表5中列出的所有转录物的表达水平。在一些实施方案中,所述分类器训练使用在表5中列出的所有转录物以及在表1或表15中列出的一个或多个额外基因的表达水平。在一些实施方案中,所述分类器训练使用在表1和表15中列出的所有转录物的表达水平。在一些实施方案中,所述计算机生成的经训练的分类器基于在表1和/或表15中列出的一种或多种转录物的表达水平将所述测试样品分类为UIP或非UIP。在一些实施方案中,所述分类器基于在表1中列出的所有转录物的表达水平将所述测试样品分类为UIP或非UIP。在一些实施方案中,所述分类器基于在表15中列出的所有转录物的表达水平将所述测试样品分类为UIP或非UIP。在一些实施方案中,所述分类器基于在表1和表15中列出的所有转录物的表达水平将所述测试样品分类为UIP或非UIP。在一些实施方案中,所述分类器训练使用在表5中列出的所有基因的表达水平。在一些实施方案中,所述计算机生成的经训练的分类器基于在表5中列出的一种或多种转录物的表达水平将所述测试样品分类为UIP或非UIP。在一些实施方案中,所述分类器基于在表5中列出的所有转录物的表达水平将所述测试样品分类为UIP或非UIP。

[0042] 在一些实施方案中,本公开内容提供了检测汇集的肺组织测试样品是否对疾病或病况呈阳性的方法,包括:(A)测定在测试样品中表达的一种或多种转录物的表达水平;以及(B)使用计算机生成的经训练的分类器将所述测试样品分类为对所述疾病或病况呈阳性或呈阴性;其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器,每个训练样品具有对所述疾病或病况呈阳性或阴性的确认诊断,其中至少两个所述训练样品从单个受试者获得;并且其中在所述分类之前汇集所述测试样品。

[0043] 在一些实施方案中,所述汇集包括物理汇集。在一些实施方案中,所述汇集包括计算机汇集。在一些实施方案中,所述分类器基于在表5中列出的一个或多个基因的表达水平将所述样品分类。在特定实施方案中,所述分类器基于在表5中列出的所有基因的表达水平将所述样品分类。

[0044] 在一些实施方案中，所述疾病或病况选自：肺部病症、肺癌、间质性肺病(ILD)、特发性肺纤维化(IPF)、寻常型间质性肺炎(UIP)或非寻常型间质性肺炎(非UIP)、急性肺损伤、细支气管炎、脱屑性间质性肺炎、弥漫性肺泡损伤、肺气肿、嗜酸性粒细胞性肺炎、非特异性间质性肺炎(NSIP)(包括细胞亚型、混合亚型或Favor亚型)、肉芽肿病、过敏性肺炎(HP)、Favor亚型过敏性肺炎(Favor HP)、机化性肺炎、肺孢子虫肺炎、肺动脉高压、呼吸性细支气管炎、肺结节病、吸烟相关的间质纤维化、慢性阻塞性肺病(COPD)、烟雾暴露史、长期烟雾暴露、短期烟雾暴露以及慢性间质纤维化。

[0045] 在一些实施方案中，本公开内容提供了用有效治疗特发性肺纤维化(IPF)的治疗剂治疗有需要的受试者的方法，包括向有需要的受试者施用有效剂量的有效治疗IPF的化合物，其中如通过计算机生成的经训练的分类器所确定的，所述有需要的受试者的表5中一个或多个基因的表达水平表明所述受试者需要IPF治疗。

[0046] 在一些实施方案中，使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平来训练所述计算机生成的经训练的分类器，每个训练样品具有UIP或非UIP的确认诊断，其中至少两个所述训练样品从单个受试者获得；并且其中在所述分类之前汇集所述测试样品。在特定实施方案中，所述计算机生成的经训练的分类器将从所述受试者获得的样品鉴定为UIP。在特定实施方案中，所述计算机生成的经训练的分类器将从所述受试者获得的样品鉴定为IPF。

[0047] 在一些实施方案中，本公开内容提供了用于鉴定受试者是否对肺部病症呈阳性的方法，包括：(a)获得所述受试者的组织样品；(b)使所述组织样品的第一部分经受细胞学测试，该细胞学测试指示所述第一部分是模糊的或可疑的；(c)在鉴定所述第一部分是模糊的或可疑的之后，测定所述组织样品的第二部分中与所述肺部病症相关的一种或多种标志物的表达水平；(d)用经训练的算法处理所述表达水平，从而以至少约90％的准确性生成所述组织样品对所述肺部病症呈阳性的分类，其中所述经训练的算法用包含多个训练样品的训练集来训练，并且其中所述组织样品独立于所述多个样品；以及(e)电子地输出所述分类，由此鉴定所述受试者是否对所述肺部病症呈阳性。

[0048] 在一些实施方案中，所述组织样品是肺组织样品。在一些实施方案中，所述组织样品是非肺组织样品。在一些实施方案中，所述非肺组织样品是呼吸上皮样品。在一些实施方案中，所述呼吸上皮样品来自所述受试者的鼻或口。

[0049] 在一些实施方案中，所述表达水平是与UIP相关的多种标志物的表达水平。

[0050] 在一些实施方案中，所述准确性为至少约95％。

[0051] 在一些实施方案中，所述分类以至少约90％的特异性生成。在一些实施方案中，所述分类以至少约70％的灵敏度生成。

[0052] 在一些实施方案中，所述经训练的算法被配置用于在至少100个独立的测试样品中以至少约90％的准确性将肺组织样品分类。

[0053] 在一些实施方案中，所述分类电子地输出在用户的电子显示器的图形用户界面上。

[0054] 在一些实施方案中，所述肺部病症是寻常型间质性肺炎(UIP)或非寻常型间质性肺炎(非UIP)。

[0055] 在一些实施方案中，所述第一部分不同于所述第二部分。

[0056] 在一些实施方案中，本公开内容提供了鉴定受试者是寻常型间质性肺炎(UIP)阳性还是非寻常型间质性肺炎(非UIP)阳性的方法，包括：(a)获得所述受试者的组织样品；(b)使所述组织样品的第一部分经受细胞学测试，该细胞学测试指示所述第一部分是模糊的或可疑的；(c)在鉴定所述第一部分是模糊的或可疑的之后，测定所述组织样品的第二部分中与UIP相关的一种或多种标志物的表达水平；(d)用经训练的算法处理所述表达水平，从而以至少约90％的准确性生成所述组织样品为UIP阳性或非UIP阳性的分类，其中所述经训练的算法用包含多个训练样品的训练集来训练，并且其中所述组织样品独立于所述多个样品；以及(e)电子地输出所述分类，由此鉴定所述受试者是UIP阳性还是非UIP阳性。

[0057] 本公开内容的另一个方面提供了一种包含机器可执行代码的非暂时性计算机可读介质，所述机器可执行代码在由一个或多个计算机处理器执行时实现本文上面或其他地方所述的任何方法。

[0058] 本公开内容的另一个方面提供了一种包含一个或多个计算机处理器和与所述一个或多个计算机处理器耦合的非暂时性计算机可读介质的计算机系统。所述非暂时性计算机可读介质包含机器可执行代码，所述机器可执行代码在由所述一个或多个计算机处理器执行时实现本文上面或其他地方所述的任何方法。

[0059] 通过仅示出并描述本发明的说明性实施方案的以下具体实施方式，本公开内容的其他方面和优点将变得对本领域技术人员而言显而易见。应当认识到，本公开内容能够具有其他和不同的实施方案，并且其若干细节能够在各个明显的方面进行修改，所有这些都不脱离本公开内容。因此，附图和详述将被视为在本质上是说明性的，而不是限制性的。

援引并入

[0060] 在本说明书中所提及的所有出版物、专利和专利申请都通过引用并入本文，其程度犹如特别地和单独地指出每个单独的出版物、专利或专利申请通过引用而并入。在通过引用并入的出版物和专利或专利申请与本说明书中包含的公开内容相矛盾时，本说明书旨在替代和/或优先于任何这类矛盾的材料。

附图说明

[0061] 本公开内容的新特征在所附权利要求书中详细阐明。通过参考对利用本发明原理的说明性实施方案加以阐述的以下具体实施方式和附图(本文中也称为“图”)，将会对本公开内容的特征和优点获得更好的理解，附图中：

[0062] 图1。用两个样品(样品A和样品B)进行的假想患者的中心病理学诊断过程。三名专业病理学家参与评审过程。对于样品水平上的诊断，每个样品的载玻片由每位病理学家评审(病理学家缩写为Path.)。对于患者水平上的诊断，将来自所有样品的载玻片(本练习中为两个)集中并由每个病理学家一起评审。样品水平和患者水平上的诊断都经历相同的评审过程。使用多数票决作为最终诊断，除非专业病理学家甚至在商议后也未能达成一致，在这种情况下，由于对诊断缺乏信心而略过该样品。在所有库存组织(n＝128)中仅观察到一例这样的情况。

[0063] 图2。样品排除/纳入程序。图2示出了经筛选用于本研究的113名患者和相关TBB样品的流程图。该图阐明了在每个连续处理步骤中患者和样品的队列(中央方形)、处理步骤(不规则四边形)和排除(外侧方形)。

[0064] 图3。分类器表现。图3A-3D示出了单样品分类表现。使用经53名患者训练的分类器通过交叉验证对训练中使用的各个TBB样品进行评分(图3A,图3B),并对来自31名患者的独立测试队列的TBB样品进行前瞻性评分(图3C,图3D)。对于训练集(图3A)和验证集(图3C)中的每个TBB样品,绘制按患者垂直排列的分类评分(y轴)。各个样品根据肺叶水平的病理学诊断进行着色,其中符号表示来源肺叶(图例)。在下x轴上提供患者水平的病理学诊断,并且在每个图的上x轴上提供放射学诊断。在训练集上的交叉验证中确定并且前瞻性地应用于测试集的决策边界以水平虚线示出。在为所有样品给出评分时的总体表现总结,在训练集上的交叉验证中提供(图3B),并且前瞻性地在验证集上提供(图3D)。每个队列中的真阳性、真阴性、假阳性和假阴性样品的总数被总结成2x 2表。列出了受试者工作特征曲线下面积(ROC-AUC)、灵敏度和特异性,以及相关的90％置信区间。图3中使用的病理学和放射学首字母缩略词:ACL,急性肺损伤;BR,细支气管炎;CIF,NOC,慢性间质纤维化,未另外分类;DIP,脱屑性间质性肺炎;DAD,弥漫性肺泡损伤;EMP,肺气肿;EO-PN,嗜酸性粒细胞性肺炎;NA,不可用/缺失;ND,非诊断性;NSIP,非特异性间质性肺炎;NSIP-C,细胞NSIP;NSIP-F,Favor NSIP;GR,肉芽肿病;HP,过敏性肺炎;HP-F,Favor HP;OP,机化性肺炎;OTHR,其他;PN-PN,肺孢子虫肺炎;PL-HY,肺动脉高压;RB,呼吸性细支气管炎;SRC,结节病;SRIF,吸烟相关的间质纤维化;UIP,寻常型间质性肺炎;UIP-C,典型UIP;UIP-D,困难UIP;UIP-F,倾向UIP(Favor UIP);UIP-DE,明确UIP;UIP-P,很可能UIP。

[0065] 图4:来自相同患者的TBB混合物中UIP的分类。图4A示出了来自8名患者(x轴)的TBB样品,其在体外作为单独样品处理并由84-患者分类器(蓝色方形)评分(y轴)。示出了来自每名患者的各个TBB样品的平均评分以用于比较(深蓝色三角形)。图4B示出了通过单样品TBB数据的随机采样对整个84-患者队列的多个(每名患者2-5个)TBB样品的混合物进行的计算机模拟。用84-患者UIP分类器对混合物进行评分,并且生成了100倍的针对跨整个队列的分类表现的ROC-AUC点估值,并针对每种混合物条件绘图。箱形图表示每个采样条件下的中值ROC-AUC。图4C示出了图4B中所示的表现,表示为在目标特异性为90％时在混合物中的测试灵敏度。随着可变性的降低,测试灵敏度提高到约72％。水平红色虚线示出了单样品分类器的ROC-AUC,作为参考点。图4D示出了一组33名受试者中的混合物模拟,其中每名受试者有两个肺上叶TBB和三个肺下叶TBB可用。当采样限于肺上叶或肺下叶时,表现没有改善。

[0066] 图5。图5A示出了使用24种标志物(44种标志物的子集)根据主成分的无监督聚类;TBB样品为蓝色,SLB样品为橙色。图5B示出了TBB群体内9个基因的双峰表达:SFTPB、SFTPC、SFTPD、ABCA3、CEBPA、AGER、GPRC5A、HOPX和SFTPA1;TBB表达以蓝色计数,SLB表达以橙色计数。图5C示出了SFTPA1、SFTPB、SFTPC和SFTPD之间的相关的、方向一致的表达,但在PDPN与AQP5之间或在这两个组的成员之间没有这样的表达;TBB表达以蓝色计数,SLB表达以橙色计数。

[0067] 图6:经支气管活检物中肺泡基因表达的分布。图6A示出了多种组织、细胞系和肿瘤类型(x轴)的I型肺泡细胞的两种标志物的表达总和(I型肺泡统计学)(y轴)。还示出了在当前研究中正常肺组织、肺肿瘤、外科肺活检物(SLB,n＝22)和经支气管活检物中的表达以用于比较(n＝283)。图6B示出了针对每个TBB样品绘制的I型肺泡统计学,其根据相对于病理学真值标签的分类正确性(例如,真阴性、假阴性、真阳性和假阳性)进行分组。图6C示出

了多种组织、细胞系和肿瘤类型(x轴)的四种肺泡细胞标志物的表达总和(II型肺泡统计学)(y轴)。还示出了在当前研究中正常肺组织、肺肿瘤、外科肺活检物(SLB,n＝22)和经支气管活检物中的表达以用于比较(n＝283)。图6D示出了针对每个TBB样品绘制的II型肺泡统计学,其根据相对于病理学真值标签的分类正确性(例如,真阴性、假阴性、真阳性和假阳性)进行分组。从3名被诊断为IPF的患者(患者P1、P2和P3)获得的外植体样品的成对相关性。每个样品都标明了位置(上部或下部,中央或外围)。使用将IPF样品与正常肺样品分开的前200个差异表达基因来计算成对Pearson相关系数,并绘制为热图,其中以品红色表示较高相关性,以绿色表示较低相关性。正常肺样品之间的相关性以及与正常肺样品的相关性在0.7的范围内(未示出)。

[0068] 图7。计算机系统;处理器;以及用于训练和利用本文公开的分类器的计算机可执行过程。图7A示出了可用于实现本文公开的方面的计算机系统的图示。图7B示出了图7A的计算机系统的处理器的详细图示。图7C示出了本公开内容的一种非限制性方法的详细图示,其中使用已知的UIP和非UIP样品的基因产物表达数据来训练分类器(例如,使用分类器训练模块)以区分UIP与非UIP,其中该分类器在一些情况下将吸烟者状态视为协变量,并且其中将来自未知样品的基因产物表达数据输入到经训练的分类器中,以将未知样品鉴定为UIP或非UIP,并且其中经由分类器的分类结果得到确定,并经由报告而输出。

[0069] 图8。阐明从88名BRAVE研究受试者衍生出Envisia最终验证组和二级分析组的流程图(实施例10)。

[0070] 图9。在实施例10研究中用来确定研究受试者的参考标签的中心病理学评审过程的流程图。

[0071] 图10。Envisia基因组分类器的验证表现。图10A示出了Envisia对49名受试者的最终验证组的ROC-AUC曲线,其中预先指定的决策边界在ROC曲线上用星号标出。图10B示出了最终验证组的Envisia分类结果的2x 2表。

[0072] 图11。Envisia验证组中49名受试者的分类评分。按照逐渐增加的分类评分(y轴)从左到右对受试者进行划分,其中中心病理学诊断在下x轴上而中心放射学诊断(如果可获得)在上x轴上。实心圆表示具有UIP参考标签的受试者,空心圆表示具有非UIP参考标签的受试者。测试决策边界用虚线表示。

[0073] 图12。由放射学定义的受试者亚组中的Envisia表现。示出了46个可获得放射学的最终验证受试者的中心和局部放射学诊断的2x2表,以病理学作为参考标准。对于放射学与UIP(明确UIP、很可能UIP和可能UIP)一致以及与UIP不一致的受试者子集,示出了针对病理学的Envisia测试表现。分别针对中心和局部放射学诊断来评价Envisia测试结果。

[0074] 图13。针对受试者临床因素的Envisia测试表现的亚组分析。UIP受试者用红色实心圆圈标记,非UIP受试者用空心或蓝色圆圈标记。图13A:作为验证队列受试者年龄的函数的Envisia分类评分。图13B:受试者年龄与分类评分之间没有显著相关性。图13C:作为受试者性别的函数的Envisia评分。男性UIP患者更容易被Envisia测试遗漏(17名患有病理学UIP的男性中有10名被Envisia判定为UIP;41％灵敏度,而UIP女性中是6/7)。图13D:作为受试者吸烟史的函数的Envisia评分。相比于非吸烟者,具有吸烟史的男性UIP患者被Envisia错误分类的比率更高。

[0075] 图14。针对样品技术因素的Envisia测试表现的亚组分析。UIP受试者用红色实心

圆圈标记，非UIP受试者用空心或蓝色圆圈标记。图14A和14B：用于估计肺泡含量的总和基因表达统计学。图14A示出了肺泡I型细胞含量(x轴)，而图14B示出了肺泡II型含量(x轴)^{E10}，各自相对于Envisia评分(y轴)作图。图14C：按照肺泡II型含量绘制Envisia测试真阳性(TP)、假阴性(FN)、真阴性(TN)和假阳性(FP)。在被Envisia测试误判的受试者中没有低肺泡II型含量的富集。图14D:Envisia分类评分与样品质量(RIN或DV200)的相关性，由UIP参考标签分隔。在非UIP样品中，在更强(更负)的分类评分与更高的样品质量之间存在相关性，这一点在UIP样品中不明显。

具体实施方式

[0076] 尽管本文中已经示出并描述了本公开内容的各个实施方案，但对于本领域技术人员显而易见的是，这些实施方案仅以示例的方式提供。本领域技术人员在不脱离本公开内容的情况下可想到多种变化、改变和替代。应当理解，可采用本文所述的本公开内容实施方案的各种替代方案。

[0077] 如本文所用的“间质性肺病”或“ILD”(也称为弥漫性实质性肺病(DPLD))是指影响间质组织(肺部气囊周围的组织和空间)的一组肺病。ILD可根据疑似或已知的病因进行分类，或可能是特发性的。例如，ILD可分类为由吸入物质(无机或有机)引起、药物(例如，抗生素、化疗药物、抗心律失常剂、他汀类药物)诱导的、与结缔组织疾病(例如，系统性硬化病、多发性肌炎、皮肌炎、系统性红斑狼疮、类风湿性关节炎)相关、与肺部感染(例如，非典型性肺炎、肺孢子虫肺炎(PCP)、结核病、沙眼衣原体、呼吸道合胞病毒)相关、与恶性肿瘤(例如，淋巴管癌病)相关，或者可以是特发性的(例如，结节病、特发性肺纤维化、哈-里综合征(Hamman-Rich syndrome)、抗合成酶综合征)。

[0078] 如本文所用的“ILD炎症”是指特征在于潜在炎症的炎性ILD亚型的分析分组。这些亚型可共同用作针对IPF和/或任何其他非炎症肺病亚型的比较物。“ILD炎症”可包括HP、NSIP、结节病和/或机化性肺炎。

[0079] “特发性间质性肺炎”或“IIP”(也称为非感染性肺炎)是指一类ILD，其包括例如脱屑性间质性肺炎、非特异性间质性肺炎、淋巴样间质性肺炎、隐原性机化性肺炎和特发性肺纤维化。

[0080] 如本文所用的“特发性肺纤维化”或“IPF”是指特征在于肺的支撑框架(间质)纤维化的慢性进行性肺病形式。根据定义，当肺纤维化的原因未知时使用该术语(“特发性”)。在显微镜下，来自患有IPF的患者的肺组织显示出一组特征性的组织学/病理学特征，称为寻常型间质性肺炎(UIP)，其为IPF的病理对应物。

[0081] “非特异性间质性肺炎”或“NSIP”是特发性间质性肺炎的一种形式，其特征在于细胞模式由具有一致的或斑块状胶原蛋白沉积的慢性炎性细胞所定义，并且纤维化模式由弥散性斑块状纤维化所定义。与UIP相反，没有蜂窝状外观和成纤维细胞病症等寻常型间质性肺炎的特征。

[0082] “过敏性肺炎”或“HP”也称为外源性变应性肺泡炎(EAA)，是指由吸入抗原(例如，有机粉尘)导致的过度免疫应答超敏反应所引起的肺内的肺泡炎症。

[0083] “肺结节病”或“PS”是指涉及可形成结节的慢性炎性细胞的异常聚集(肉芽肿)的综合征。HP的炎性过程通常涉及肺泡、小支气管和小血管。在急性和亚急性HP病例中，体检

通常会显示干性罗音。

[0084] 术语“微阵列”是指可杂交阵列元件，优选多核苷酸探针在基底上的有序排列。

[0085] 当以单数或复数形式使用时，术语“多核苷酸”通常是指任何多核糖核苷酸或多脱氧核糖核苷酸，其可以是未修饰的RNA或DNA或者修饰的RNA或DNA。因此，例如，如本文所定义的多核苷酸包括但不限于单链和双链DNA、包括单链和双链区域的DNA、单链和双链RNA、以及包括单链和双链区域的RNA、包含可以是单链的或更通常是双链的或者包括单链和双链区域的DNA和RNA的杂合分子。另外，如本文所用的术语“多核苷酸”是指包含RNA或DNA或者RNA和DNA二者的三链区域。这样的区域中的链可来自相同分子或来自不同分子。该区域可包括一个或多个分子中的全部，但更典型地仅包括一些分子的区域。三螺旋区域分子中的一种通常是寡核苷酸。术语“多核苷酸”还可包括含有一个或多个修饰碱基(例如，以提供可检测的信号，如荧光团)的DNA(例如，cDNA)和RNA。因此，具有为了稳定性或其他原因而修饰的骨架的DNA或RNA是“多核苷酸”，如该术语在本文所意指的。此外，包含非常见碱基如肌苷或修饰碱基如氚化碱基的DNA或RNA包括如在本文定义的术语“多核苷酸”内。通常，术语“多核苷酸”包括未修饰的多核苷酸的所有化学、酶促和/或代谢修饰形式，以及病毒和细胞(包括简单和复杂细胞)特有的DNA和RNA的化学形式。

[0086] 术语“寡核苷酸”是指相对较短的多核苷酸(例如，100个、50个、20个或更少的核苷酸)，包括但不限于单链脱氧核糖核苷酸，单链或双链核糖核苷酸、RNA:DNA杂合体和双链DNA。寡核苷酸，如单链DNA探针寡核苷酸，通常通过化学方法例如使用可商购的自动化寡核苷酸合成仪合成。然而，寡核苷酸可通过各种其他方法制备，包括体外重组DNA介导的技术以及通过在细胞和生物体中表达DNA。

[0087] 术语“基因产物”或“表达产物”在本文可互换使用，以指基因的RNA转录产物(RNA转录物)，包括mRNA，以及这样的RNA转录物的多肽翻译产物。基因产物可以是例如多核苷酸基因表达产物(例如，未剪接的RNA、mRNA、剪接变体mRNA、微小RNA、片段化RNA等)或蛋白质表达产物(例如，成熟多肽、翻译后修饰的多肽、剪接变体多肽等)。在一些实施方案中，基因表达产物可以是包含突变、融合、杂合性丢失(LOH)和/或生物学途径效应的序列变体。

[0088] 应用于基因表达产物的术语“归一化表达水平”是指基因产物相对于一种或多种参考(或对照)基因表达产物归一化的水平。

[0089] 应用于基因表达产物的术语“参考表达水平”是指一种或多种参考(或“对照”)基因表达产物的表达水平。应用于基因表达产物的术语“参考归一化表达水平”是指一种或多种参考(或对照)基因表达产物的归一化表达水平值(即归一化参考表达水平)。在一些实施方案中，参考表达水平是正常样品中的一种或多种基因产物的表达水平，如本文所述。在一些实施方案中，参考表达水平通过实验确定。在一些实施方案中，参考表达水平是历史表达水平，例如，正常样品中的参考表达水平的数据库值，该样品指示单个参考表达水平或多个参考表达水平的总结(例如，(i)来自单个样品的参考表达水平的重复分析的两个或更多个，优选三个或更多个参考表达水平的平均值；(ii)来自多个不同样品(例如，正常样品)的参考表达水平分析的两个或更多个，优选三个或更多个参考表达水平的平均值；(iii)以及上述步骤(i)和(ii)的组合(即，由多个样品分析的参考表达水平的平均值，其中至少一个参考表达水平被重复分析))。在一些实施方案中，“参考表达水平”是序列变体的表达水平，该序列变体例如在通过其他方法明确确定为UIP或非UIP(即确认的病理诊断)的样品中。

[0090] 应用于基因表达产物的术语“参考表达水平值”是指一种或多种参考(或对照)基因表达产物的表达水平值。应用于基因表达产物的术语“归一化参考表达水平值”是指一种或多种参考(或对照)基因表达产物的归一化表达水平值。

[0091] 杂交反应的“严格性”可由本领域普通技术人员容易地确定,并且通常是取决于探针长度、洗涤温度和盐浓度的经验计算。通常,较长的探针需要较高的温度来进行适当的退火,而较短的探针需要较低的温度。当互补链存在于低于其解链温度的环境中时,杂交通常取决于变性DNA再退火的能力。探针与可杂交序列之间所需的同源性程度越高,可使用的相对温度越高。因此,结果是较高的相对温度可能倾向于使反应条件更严格,而较低的温度则不那么严格。关于杂交反应的严格性的其他细节和解释,参见Ausubel等人,Current Protocols in Molecular Biology,(Wiley Interscience,1995)。

[0092] 如本文所定义的“严格条件”或“高严格性条件”通常:(1)使用低离子强度溶液和高温进行洗涤,例如0.015M氯化钠/0.0015M柠檬酸钠/0.1%十二烷基硫酸钠在50℃下;(2)在杂交期间用变性剂如甲酰胺,例如50%(v/v)甲酰胺与0.1%牛血清白蛋白/0.1%Ficoll/0.1%聚乙烯吡咯烷酮/pH 6.5的50mM磷酸钠缓冲液,伴有750mM氯化钠、75mM柠檬酸钠,在42℃下;或者(3)使用50%甲酰胺、5x SSC(0.75M NaCl、0.075M柠檬酸钠)、50mM磷酸钠(pH 6.8)、0.1%焦磷酸钠、5x Denhardt溶液、超声处理的鲑精DNA(50μg/ml)、0.1%SDS和10%硫酸葡聚糖在42℃下,并在55℃的0.2x SSC(氯化钠/柠檬酸钠)和50%甲酰胺中在42℃下洗涤,然后是由含有EDTA的0.1x SSC组成的在55℃下的高严格性洗涤。

[0093] “中等严格条件”可如Sambrook等人,Molecular Cloning:A Laboratory Manual(Cold Spring Harbor Press,1989)所述进行鉴定,并包括使用不如上述严格的洗涤溶液和杂交条件(例如,温度、离子强度和%SDS)。中等严格条件的实例是在37℃下在包含20%甲酰胺、5x SSC(150mM NaCl、15mM柠檬酸三钠)、50mM磷酸钠(pH 7.6),5x Denhardt溶液、10%硫酸葡聚糖和20mg/ml变性的剪切鲑精DNA的溶液中过夜温育,然后在1x SSC中约37-50℃下洗涤滤器。本领域技术人员将认识到如何根据需要调节温度、离子强度等以适应诸如探针长度等因素。

[0094] 如本文所用,“敏感性”是指测试的总数中实际具有目标病症的真阳性的比例(即,患有目标病症的具有阳性测试结果的患者的比例)。如本文所用,“特异性”是指测试的所有患者中实际不具有目标病症的真阴性的比例(即,未患目标病症的具有阴性测试结果的患者的比例)。

[0095] 在本公开内容的上下文中,对任何特定基因组中列出的基因的“至少一个”、“至少两个”、“至少五个”等的提及意指所列基因中的任一个或者任何和所有组合。

[0096] 术语“剪接”和“RNA剪接”可互换使用,并且是指去除内含子并连接外显子以产生移动到真核细胞的细胞质中的具有连续编码序列的成熟mRNA的RNA加工。

[0097] “治疗有效量”或“治疗有效剂量”是指当施用于受试者(例如,优选哺乳动物,更优选人)时,足以实现动物的疾病或病况的如下所定义的治疗的本发明化合物的量。构成“治疗有效量”的本发明化合物的量将根据化合物、病况及其严重程度、给药方式和待治疗的受试者的年龄而变化,但可以由本领域普通技术人员考虑到其自身知识和本公开内容而常规确定。因此,当以“有效剂量”施用化合物时,这意指该化合物能够以这样的剂量对受试者中的疾病或病况(例如,IPF)实现如下所定义的治疗。

[0098] 如本文所用的“治疗”或“处理”包括在患有感兴趣的疾病或病况的受试者(优选人)中治疗感兴趣的疾病或病况(例如,IPF),并且包括:(i)预防或抑制该疾病或病况在受试者中发生,特别是当这样的受试者易患该病况但尚未被诊断患有该病况时;(ii)抑制该疾病或病况,即阻止其发展;(iii)缓解该疾病或病况,即导致该疾病或病况消退;或(iv)缓解由该疾病或病况引起的症状。如本文所用,术语“疾病”、“病症”和“病况”可互换使用,或者可以是不同的,因为特定的疾病、损伤或病况可能不具有已知的病原体(因此病因学尚未被解决),因此,其不被认为是一种损伤或疾病,而仅仅是一种临床医生已鉴别出或多或少的特定一组症状的不期望的状况或综合征。

[0099] 术语“外显子”是指在成熟RNA产物中表达的间断基因的任何区段(B.Lewin,Genes 7V(Cell Press,1990))。理论上,术语“内含子”是指被转录但通过将其两侧的外显子剪接在一起而从转录物中去除的任何DNA区段。在操作上,外显子序列出现在由参考SEQ ID号定义的基因的mRNA序列中。在操作上,内含子序列是基因的基因组DNA内的间插序列,外显子序列在其两侧,并且通常在其5’和3’边界处具有GT和AG剪接共有序列。

[0100] “基于计算机的系统”是指用于分析信息的硬件、软件和数据存储介质的系统。基于患者计算机的系统的硬件可包括中央处理单元(CPU),以及用于数据输入、数据输出(例如,显示)和数据存储的硬件。数据存储介质可包括包含如上所述的现有信息的记录的任何制造品,或者可访问这样的制造品的存储器访问设备。

[0101] 如本文所用,术语“模块”是指任何组装件和/或一组可操作地耦合的电子组件,其可包括例如存储器、处理器、电气线路、光学连接器、软件(在硬件中执行)和/或诸如此类。例如,在处理器中执行的模块可以是能够执行与该模块相关联的一个或多个特定功能的基于硬件的模块(例如,现场可编程门阵列(FPGA)、专用集成电路(ASIC)、数字信号处理器(DSP))和/或基于软件的模块(例如,存储在存储器中和/或在处理器处执行的计算机代码的模块)的任何组合。

[0102] 在计算机可读介质上“记录”数据、编程或其他信息是指使用各种方法存储信息的过程。可基于用于访问所存储的信息的方法来选择任何方便的数据存储结构。各种数据处理器程序和格式均可用于存储,例如,文字处理文本文件、数据库格式等。

[0103] “处理器”(或“计算机处理器”)是指将会执行需要其的功能的任何硬件和/或软件组合。例如,合适的处理器可以是可编程数字微处理器,诸如可以以电子控制器、主机、服务器或个人计算机(台式或便携式)的形式提供。在处理器可编程的情况下,合适的编程可从远程位置传送到处理器,或者先前保存在计算机程序产品(如便携式或固定式计算机可读存储介质,无论是基于磁、光还是固态设备的)中。例如,磁介质或光盘可携带编程,并且可由与其相应站点处的每个处理器通信的合适读取器读取。

[0104] “测试样品”是一个或多个细胞的样品,优选从受试者获得的组织样品(例如,肺组织样品,如经支气管活检(TBB)样品)。在一些实施方案中,测试样品是活检样品,其可通过各种方法(例如,手术)获得。在特定实施方案中,测试样品是通过电视辅助胸腔镜手术(VATS)、支气管肺泡灌洗(BAL)、经支气管活检(TBB)或低温经支气管活检获得的样品。可通过辅助性支气管镜程序如擦刷(诸如通过细胞刷、组织刷)、支气管活检、支气管灌洗或针吸获得测试样品。可通过口腔清洗、触摸制备或痰收集来获得样品。可基于患者所呈现的临床体征和症状(例如,呼吸短促(通常通过劳累而加重)、干咳),并且在一些情况下基于成像检

查(例如,胸部X光、计算机断层扫描(CT))、肺功能检查(例如,肺量测定法、血氧测定法、运动压力测试)、肺组织分析(例如,通过支气管镜检查、支气管肺泡灌洗、手术活检获得的样品的组织学和/或细胞学分析)中的一种或多种,从疑似患有肺病例如ILD的患者获得测试样品。在一些实施方案中,从受试者的呼吸上皮获得测试样品。呼吸上皮可来自口、鼻、咽、气管、支气管、细支气管或肺泡。然而,还可使用其他呼吸上皮来源。在一些实施方案中,测试样品是汇集的样品。

[0105] 术语"汇集"在本文中用于描述(i)"物理汇集",即样品实际混合在一起,或(ii)"计算机汇集",即汇集样品中所检测的一种或多种基因的表达值的方法。可以如何进行这样的计算机汇集的非限制性实例概述于实施例6中。术语"计算机混合"和"计算机汇集"在本文可互换使用。包含已经历物理汇集的多个样品的样品(例如,测试样品)在本文中可称为"汇集样品"。

[0106] 如本文所用,术语"受试者"通常是指哺乳动物。通常,受试者是人。然而,该术语包括其他物种,例如猪、小鼠、大鼠、狗、猫或其他灵长类动物。在某些实施方案中,受试者是实验受试者,如小鼠或大鼠。受试者可以是雄性(男性)或雌性(女性)。受试者可以是婴儿、幼儿、儿童、年轻人、成人或老年人。受试者可以是吸烟者、曾经吸烟者或非吸烟者。受试者可具有ILD的个人史或家族史。受试者可具有无ILD的个人史或家族史。受试者可表现出ILD或另一种肺部病症(例如,癌症、肺气肿、COPD)的一种或多种症状。例如,受试者可表现出呼吸短促(通常由于劳累而加重)和/或干咳,并且在一些情况下可能已经获得成像检查(例如,胸部X光、计算机断层扫描(CT))、肺功能检查(例如,肺量测定法、血氧测定法、运动压力测试)、肺组织分析(例如,通过支气管镜检查、支气管肺泡灌洗、手术活检获得的样品的组织学和/或细胞学分析)中的一种或多种的结果,其指示可能存在ILD或其他肺部病症。在一些实施方案中,受试者患有或已被诊断患有慢性阻塞性肺病(COPD)。在一些实施方案中,受试者未患有或未被诊断患有COPD。受医生或其他医疗保健提供者照顾的受试者可被称为"患者"。

[0107] "基因特征"是一种基因表达模式(即,一个或多个基因或其片段的表达水平),其指示一些特性或表型。在一些实施方案中,基因特征是指一个基因、多个基因、基因片段或者一个或多个基因的多个片段的表达(和/或缺乏表达),该表达和/或缺乏表达指示UIP、非UIP、吸烟者状态或非吸烟者状态。

[0108] 如本文所用,"是吸烟者"是指当前吸烟的受试者或过去曾经吸烟的人,或者具有当前吸烟或过去曾经吸烟的人的基因特征的人。

[0109] 如本文所用,当用于描述在训练本公开内容的分类器期间使用的特征时,"变体"是指可变剪接变体。

[0110] 如本文所用,当用于描述在训练本公开内容的分类器期间使用的特征时,"突变"是指与已知的正常参考序列有偏差的序列。在一些实施方案中,该偏差是与公认的原始基因序列的偏差,该原始基因序列是根据公众可访问的数据库,如UniGene数据库(Pontius JU,Wagner L,Schuler GD.UniGene:a unified view of the transcriptome,见NCBI手册,Bethesda(MD):国家生物技术信息中心;2003,并入本文)、RefSeq(NCBI手册[因特网],Bethesda(MD):国家医学图书馆(美国),国家生物技术信息中心;2002年10月,第18章,The Reference Sequence(RefSeq)Project,可在万维网地址ncbi.nlm.nih.gov/refseq/获

得)、Ensembl(EMBL,可在万维网地址ensembl.org/index.html获得)等等。在一些实施方案中,突变包括参考序列中存在的序列残基的添加、缺失或置换。

[0111] 缩写包括:HRCT,高分辨率计算机断层扫描;VATS,电视辅助胸腔镜手术;SLB,外科肺活检;TBB,经支气管活检;RB,呼吸性细支气管炎;OP,机化性肺炎,DAD,弥漫性肺泡损伤,CIF/NOC,未另外分类的慢性间质纤维化;MDT,多学科团队;CV,交叉验证;LOPO,留一患者(leave-one-patient-out);ROC,受试者工作特征;AUC,曲线下面积;RNASeq,通过下一代测序进行RNA测序的技术;NGS,下一代测序技术;H&E,苏木精和伊红;FDR,错误发现率;IRB,机构审查委员会;ATS,美国胸腔学会;COPD,慢性阻塞性肺病;KEGG,京都基因与基因组百科全书;CI,置信区间。

[0112] 当提供数值范围时,应当理解,在此范围的上限与下限之间的每个中间值(精确到下限单位的十分之一,除非上下文另有明确指出)以及在所述范围内的任何其他所指出的值或中间值都包括在本公开内容内。这些较小范围的上限和下限可独立地包括在较小范围中,而且也包括在本公开内容内,所述范围内任何具体排除的限值除外。当所述范围包括限值中的一个或两个时,排除掉这些所含限值中的任一个或两个的范围也被包括在本公开内容中。如本文所用,"约"意指指示值加或减10%。

检测寻常型间质性肺炎(UIP)的方法

[0113] 本文公开了使用分子特征来区分UIP与非UIP的方法和/或系统。在专业病理学不可用的情况下,由样品(例如,从患者获得的样品)对UIP的准确诊断通过加速诊断而有益于ILD患者,从而促进治疗决策并降低患者的手术风险和医疗保健系统的费用。

[0114] 本文还公开了使用受试者的吸烟者或非吸烟者状态来使用分子特征改善UIP与其他ILD亚型的区分的方法和/或系统。

[0115] 因此,本文公开的方法和/或系统提供了分类器,其可在先前不了解临床或人口统计信息的情况下基于转录数据(例如,高维转录数据)区分UIP与非UIP模式。

[0116] 在一些实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法,该分类器包含表1和/或表15中提供的一个或多个序列或其片段或者来自表1和/或表15的至少一个序列或其片段,或者由其组成。在一些实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法,该分类器包含表5中提供的一个或多个基因或者来自表5的至少一个序列或其片段,或者由其组成。在一些实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法,该分类器包含表1和/或表15中提供的一个或多个序列或者来自表1和/或表15的至少一个序列,或者由其组成。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表1和/或表15中提供的至少1、2、3、4、5、6、7、8、9、10个或更多个序列或者由其组成。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表5中提供的至少1、2、3、4、5、6、7、8、9、10个或更多个序列或者由其组成。例如,在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表1中提供的至少11、12、13、14、15、20、30、50、100、125、150或151个序列或者由其组成,包括之间的所有整数(例如,16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表1的约1-10个序列、约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-151个序列等)。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表15中提供的至少11、12、13、14、15、20、30、50、100、125、150或169个序列或者由其组成,包括之间的所有整数(例如,来自表

15的16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表15的约1-10个序列,来自表15的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-169个序列等)。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表5中提供的至少11、12、13、14、15、20、30、50、100、125、150、160、170、180、181、182、183、184、185、186、187、188、189或190个基因或者由其组成,包括之间的所有整数(例如,来自表5的16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表1的约1-10个序列,来自表5的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-169个序列、60-190个序列等)。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表1和表15中的一者或两者中提供的至少11、12、13、14、15、20、30、50、100、125、150、200、250、300或320个序列或者由其组成,包括之间的所有整数(例如,来自表1和/或表15的16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表1和/或表15的约1-10个序列,来自表1和/或表15的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-200个序列、75-250个序列、100-300个序列等)。在一些实施方案中,本公开内容提供了使用分类器的方法,该分类器包含表1、表5和表15中的一者、两者或三者中提供的至少11、12、13、14、15、20、30、50、100、125、150、200、250、300、320、350个或更多个基因或者由其组成,包括之间的所有整数(例如,来自表1、表5和/或表15的16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表1、表5和/或表15的约1-10个序列,来自表1、表5和/或表15的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-200个序列、75-250个序列、100-300个序列等)。

表 1					
SEQ ID No	基因 id	基因生物型	SEQ ID No	基因 id	基因生物型
1.	ENSG00000162408	编码蛋白质	2.	ENSG00000163872	编码蛋白质
3.	ENSG00000116285	编码蛋白质	4.	ENSG00000197701	编码蛋白质
5.	ENSG00000219481	编码蛋白质	6.	ENSG00000168826	编码蛋白质
7.	ENSG00000204219	编码蛋白质	8.	ENSG00000178988	编码蛋白质
9.	ENSG00000117751	编码蛋白质	10.	ENSG00000178177	编码蛋白质
11.	ENSG00000159023	编码蛋白质	12.	ENSG00000109618	编码蛋白质
13.	ENSG00000116761	编码蛋白质	14.	ENSG00000250317	编码蛋白质
15.	ENSG00000117226	编码蛋白质	16.	ENSG00000081041	编码蛋白质
17.	ENSG00000163386	编码蛋白质	18.	ENSG00000145284	编码蛋白质
19.	ENSG00000186141	编码蛋白质	20.	ENSG00000163644	编码蛋白质
21.	ENSG00000122497	编码蛋白质	22.	ENSG00000163110	编码蛋白质
23.	ENSG00000203832	编码蛋白质	24.	ENSG00000138795	编码蛋白质
25.	ENSG00000143379	编码蛋白质	26.	ENSG00000205403	编码蛋白质
27.	ENSG00000143367	编码蛋白质	28.	ENSG00000153404	编码蛋白质
29.	ENSG00000163220	编码蛋白质	30.	ENSG00000206077	编码蛋白质
31.	ENSG00000007933	编码蛋白质	32.	ENSG00000145736	编码蛋白质
33.	ENSG00000143322	编码蛋白质	34.	ENSG00000145730	编码蛋白质
35.	ENSG00000174307	编码蛋白质	36.	ENSG00000168938	编码蛋白质
37.	ENSG00000143466	编码蛋白质	38.	ENSG00000113621	编码蛋白质
39.	ENSG00000135766	编码蛋白质	40.	ENSG00000120738	编码蛋白质
41.	ENSG00000163029	编码蛋白质	42.	ENSG00000253953	编码蛋白质
43.	ENSG00000115828	编码蛋白质	44.	ENSG00000261934	编码蛋白质
45.	ENSG00000135625	编码蛋白质	46.	ENSG00000155846	编码蛋白质
47.	ENSG00000115317	编码蛋白质	48.	ENSG00000186470	编码蛋白质
49.	ENSG00000228325	编码蛋白质	50.	ENSG00000026950	编码蛋白质
51.	ENSG00000074582	编码蛋白质	52.	ENSG00000137331	编码蛋白质
53.	ENSG00000123983	编码蛋白质	54.	ENSG00000244731	编码蛋白质
55.	ENSG00000144712	编码蛋白质	56.	ENSG00000240065	编码蛋白质
57.	ENSG00000168036	编码蛋白质	58.	ENSG00000204252	编码蛋白质
59.	ENSG00000187094	编码蛋白质	60.	ENSG00000137309	编码蛋白质
61.	ENSG00000179152	编码蛋白质	62.	ENSG00000137166	编码蛋白质
63.	ENSG00000173402	编码蛋白质	64.	ENSG00000124702	编码蛋白质
65.	ENSG00000163412	编码蛋白质	66.	ENSG00000112299	编码蛋白质

表 1					
SEQ ID No	基因 id	基因生物型	SEQ ID No	基因 id	基因生物型
67.	ENSG00000227124	编码蛋白质	68.	ENSG00000111962	编码蛋白质
69.	ENSG00000184500	编码蛋白质	70.	ENSG00000112110	编码蛋白质
71.	ENSG00000181458	编码蛋白质	72.	ENSG00000048052	编码蛋白质
73.	ENSG00000034533	编码蛋白质	74.	ENSG00000006625	编码蛋白质
75.	ENSG00000198585	编码蛋白质	76.	ENSG00000075303	编码蛋白质
77.	ENSG00000172667	编码蛋白质	78.	ENSG00000158457	编码蛋白质
79.	ENSG00000078070	编码蛋白质	80.	ENSG00000050327	编码蛋白质
81.	ENSG00000033050	编码蛋白质	82.	ENSG00000072310	编码蛋白质
83.	ENSG00000105983	编码蛋白质	84.	ENSG00000108448	编码蛋白质
85.	ENSG00000164821	编码蛋白质	86.	ENSG00000141068	编码蛋白质
87.	ENSG00000012232	编码蛋白质	88.	ENSG00000196712	编码蛋白质
89.	ENSG00000130958	编码蛋白质	90.	ENSG00000242384	编码蛋白质
91.	ENSG00000041982	编码蛋白质	92.	ENSG00000073605	编码蛋白质
93.	ENSG00000136861	编码蛋白质	94.	ENSG00000167941	编码蛋白质
95.	ENSG00000136933	编码蛋白质	96.	ENSG00000154263	编码蛋白质
97.	ENSG00000160447	编码蛋白质	98.	ENSG00000161533	编码蛋白质
99.	ENSG00000148357	编码蛋白质	100.	ENSG00000181045	编码蛋白质
101.	ENSG00000170835	编码蛋白质	102.	ENSG00000211563	miRNA
103.	ENSG00000130653	编码蛋白质	104.	ENSG00000132199	编码蛋白质
105.	ENSG00000165997	编码蛋白质	106.	ENSG00000154655	编码蛋白质
107.	ENSG00000120539	编码蛋白质	108.	ENSG00000075643	编码蛋白质
109.	ENSG00000156113	编码蛋白质	110.	ENSG00000101000	编码蛋白质
111.	ENSG00000138166	编码蛋白质	112.	ENSG00000130005	编码蛋白质
113.	ENSG00000148925	编码蛋白质	114.	ENSG00000130513	编码蛋白质
115.	ENSG00000171714	编码蛋白质	116.	ENSG00000213965	编码蛋白质
117.	ENSG00000149090	编码蛋白质	118.	ENSG00000006659	编码蛋白质
119.	ENSG00000254761	lincRNA	120.	ENSG00000086544	编码蛋白质
121.	ENSG00000137474	编码蛋白质	122.	ENSG00000104812	编码蛋白质
123.	ENSG00000149289	编码蛋白质	124.	ENSG00000167757	编码蛋白质
125.	ENSG00000120647	编码蛋白质	126.	ENSG00000198464	编码蛋白质
127.	ENSG00000111679	编码蛋白质	128.	ENSG00000022556	编码蛋白质
129.	ENSG00000139197	编码蛋白质	130.	ENSG00000083814	编码蛋白质
131.	ENSG00000110900	编码蛋白质	132.	ENSG00000093072	编码蛋白质
133.	ENSG00000123358	编码蛋白质	134.	ENSG00000185133	编码蛋白质
135.	ENSG00000172789	编码蛋白质	136.	ENSG00000198792	编码蛋白质

表 1					
SEQ ID No	基因 id	基因生物型	SEQ ID No	基因 id	基因生物型
137.	ENSG00000073910	编码蛋白质	138.	ENSG00000189306	编码蛋白质
139.	ENSG00000083544	编码蛋白质	140.	ENSG00000100376	编码蛋白质
141.	ENSG00000187630	编码蛋白质	142.	ENSG00000154642	编码蛋白质
143.	ENSG00000157379	编码蛋白质	144.	ENSG00000100557	编码蛋白质
145.	ENSG00000100592	编码蛋白质	146.	ENSG00000100650	编码蛋白质
147.	ENSG00000119711	编码蛋白质	148.	ENSG00000128891	编码蛋白质
149.	ENSG00000140718	编码蛋白质	150.	ENSG00000182810	编码蛋白质
151.	ENSG00000103044	编码蛋白质			

[0117] 本文列出的ENSG标识符(即,基因id)是指可在万维网地址ensembl.org获得的Ensembl数据库的基因标识符,该数据库的内容通过引用整体并入本文。

[0118] 在一些特定实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统,该分类器包含表1和/或表15中列出的一个或多个序列或其片段,或者由其组成。在特定方面,该分类器可包含1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面,该分类器可略去这些基因中的1、2、3、4、5、6、7、8个或更多个,而在一些情况下包括其他基因。

[0119] 在一些特定实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统,该分类器包含表5中列出的一个或多个序列或其片段,或者由其组成。在特定方面,该分类器可包含1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面,该分类器可略去这些基因中的1、2、3、4、5、6、7、8个或更多个,而在一些情况下包括其他基因。在某些实施方案中,本公开内容提供了使用Envisia分类器区分UIP与非UIP的方法和/或系统,该分类器可包含表5中列出的所有基因。

[0120] 在一些实施方案中,本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统,该分类器包含以下序列中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150或151个,或者由其组成:单独的或任意组合的ENSG00000162408;ENSG00000116285;ENSG00000219481;ENSG00000204219;ENSG00000117751;ENSG00000159023;ENSG00000116761;ENSG00000117226;ENSG00000163386;ENSG00000186141;ENSG00000122497;ENSG00000203832;ENSG00000143379;ENSG00000143367;ENSG00000163220;ENSG00000007933;ENSG00000143322;ENSG00000174307;ENSG00000143466;ENSG00000135766;ENSG00000163029;ENSG00000115828;ENSG00000135625;ENSG00000115317;ENSG00000228325;ENSG00000074582;ENSG00000123983;

ENSG00000144712;ENSG00000168036;ENSG00000187094;ENSG00000179152;
ENSG00000173402;ENSG00000163412;ENSG00000227124;ENSG00000184500;
ENSG00000181458;ENSG00000034533;ENSG00000198585;ENSG00000172667;
ENSG00000078070;ENSG00000033050;ENSG00000105983;ENSG00000164821;
ENSG00000012232;ENSG00000130958;ENSG00000041982;ENSG00000136861;
ENSG00000136933;ENSG00000160447;ENSG00000148357;ENSG00000170835;
ENSG00000130653;ENSG00000165997;ENSG00000120539;ENSG00000156113;
ENSG00000138166;ENSG00000148925;ENSG00000171714;ENSG00000149090;
ENSG00000254761;ENSG00000137474;ENSG00000149289;ENSG00000120647;
ENSG00000111679;ENSG00000139197;ENSG00000110900;ENSG00000123358;
ENSG00000172789;ENSG00000073910;ENSG00000083544;ENSG00000187630;
ENSG00000157379;ENSG00000100557;ENSG00000100592;ENSG00000100650;
ENSG00000119711;ENSG00000128891;ENSG00000140718;ENSG00000182810;
ENSG00000103044;ENSG00000163872;ENSG00000197701;ENSG00000168826;
ENSG00000178988;ENSG00000178177;ENSG00000109618;ENSG00000250317;
ENSG00000081041;ENSG00000145284;ENSG00000163644;ENSG00000163110;
ENSG00000138795;ENSG00000205403;ENSG00000153404;ENSG00000206077;
ENSG00000145736;ENSG00000145730;ENSG00000168938;ENSG00000113621;
ENSG00000120738;ENSG00000253953;ENSG00000261934;ENSG00000155846;
ENSG00000186470;ENSG00000026950;ENSG00000137331;ENSG00000244731;
ENSG00000240065;ENSG00000204252;ENSG00000137309;ENSG00000137166;
ENSG00000124702;ENSG00000112299;ENSG00000111962;ENSG00000112110;
ENSG00000048052;ENSG00000006625;ENSG00000075303;ENSG00000158457;
ENSG00000050327;ENSG00000072310;ENSG00000108448;ENSG00000141068;
ENSG00000196712;ENSG00000242384;ENSG00000073605;ENSG00000167941;
ENSG00000154263;ENSG00000161533;ENSG00000181045;ENSG00000211563;
ENSG00000132199;ENSG00000154655;ENSG00000075643;ENSG00000101000;
ENSG00000130005;ENSG00000130513;ENSG00000213965;ENSG00000006659;
ENSG00000086544;ENSG00000104812;ENSG00000167757;ENSG00000198464;
ENSG00000022556;ENSG00000083814;ENSG00000093072;ENSG00000185133;
ENSG00000198792;ENSG00000189306;ENSG00000100376;ENSG00000154642。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0121] 在一些实施方案中，本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统，该分类器包含所有下列序列或者由其组成：ENSG00000162408;ENSG00000116285;
ENSG00000219481;ENSG00000204219;ENSG00000117751;ENSG00000159023;
ENSG00000116761;ENSG00000117226;ENSG00000163386;ENSG00000186141;
ENSG00000122497;ENSG00000203832;ENSG00000143379;ENSG00000143367;

ENSG00000163220;ENSG00000007933;ENSG00000143322;ENSG00000174307;ENSG00000143466;ENSG00000135766;ENSG00000163029;ENSG00000115828;ENSG00000135625;ENSG00000115317;ENSG00000228325;ENSG00000074582;ENSG00000123983;ENSG00000144712;ENSG00000168036;ENSG00000187094;ENSG00000179152;ENSG00000173402;ENSG00000163412;ENSG00000227124;ENSG00000184500;ENSG00000181458;ENSG00000034533;ENSG00000198585;ENSG00000172667;ENSG00000078070;ENSG00000033050;ENSG00000105983;ENSG00000164821;ENSG00000012232;ENSG00000130958;ENSG00000041982;ENSG00000136861;ENSG00000136933;ENSG00000160447;ENSG00000148357;ENSG00000170835;ENSG00000130653;ENSG00000165997;ENSG00000120539;ENSG00000156113;ENSG00000138166;ENSG00000148925;ENSG00000171714;ENSG00000149090;ENSG00000254761;ENSG00000137474;ENSG00000149289;ENSG00000120647;ENSG00000111679;ENSG00000139197;ENSG00000110900;ENSG00000123358;ENSG00000172789;ENSG00000073910;ENSG00000083544;ENSG00000187630;ENSG00000157379;ENSG00000100557;ENSG00000100592;ENSG00000100650;ENSG00000119711;ENSG00000128891;ENSG00000140718;ENSG00000182810;ENSG00000103044;ENSG00000163872;ENSG00000197701;ENSG00000168826;ENSG00000178988;ENSG00000178177;ENSG00000109618;ENSG00000250317;ENSG00000081041;ENSG00000145284;ENSG00000163644;ENSG00000163110;ENSG00000138795;ENSG00000205403;ENSG00000153404;ENSG00000206077;ENSG00000145736;ENSG00000145730;ENSG00000168938;ENSG00000113621;ENSG00000120738;ENSG00000253953;ENSG00000261934;ENSG00000155846;ENSG00000186470;ENSG00000026950;ENSG00000137331;ENSG00000244731;ENSG00000240065;ENSG00000204252;ENSG00000137309;ENSG00000137166;ENSG00000124702;ENSG00000112299;ENSG00000111962;ENSG00000112110;ENSG00000048052;ENSG00000006625;ENSG00000075303;ENSG00000158457;ENSG00000050327;ENSG00000072310;ENSG00000108448;ENSG00000141068;ENSG00000196712;ENSG00000242384;ENSG00000073605;ENSG00000167941;ENSG00000154263;ENSG00000161533;ENSG00000181045;ENSG00000211563;ENSG00000132199;ENSG00000154655;ENSG00000075643;ENSG00000101000;ENSG00000130005;ENSG00000130513;ENSG00000213965;ENSG00000006659;ENSG00000086544;ENSG00000104812;ENSG00000167757;ENSG00000198464;ENSG00000022556;ENSG00000083814;ENSG00000093072;ENSG00000185133;ENSG00000198792;ENSG00000189306;ENSG00000100376;ENSG00000154642。在特定方面，该分类器包含1、2、3、4、5、6、7、8个或更多个额外的基因。

[0122] 在一些实施方案中，本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统，该分类器单独地或组合地包含2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、

70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个以下基因，或者由其组成：ENSG00000005381；ENSG00000005955；ENSG00000007908；ENSG00000007933；ENSG00000010379；ENSG00000012232；ENSG00000022556；ENSG00000026950；ENSG00000033050；ENSG00000038295；ENSG00000048052；ENSG00000054803；ENSG00000054938；ENSG00000060688；ENSG00000071909；ENSG00000072310；ENSG00000073605；ENSG00000078070；ENSG00000079385；ENSG00000081041；ENSG00000081985；ENSG00000082781；ENSG00000083814；ENSG00000086544；ENSG00000089902；ENSG00000092295；ENSG00000099251；ENSG00000099974；ENSG00000100376；ENSG00000100557；ENSG00000101544；ENSG00000102837；ENSG00000103044；ENSG00000103257；ENSG00000104812；ENSG00000105255；ENSG00000105559；ENSG00000105696；ENSG00000105784；ENSG00000105983；ENSG00000106018；ENSG00000106178；ENSG00000107929；ENSG00000108312；ENSG00000108551；ENSG00000109205；ENSG00000110092；ENSG00000110900；ENSG00000110975；ENSG00000111218；ENSG00000111321；ENSG00000111328；ENSG00000112164；ENSG00000112299；ENSG00000112852；ENSG00000114248；ENSG00000114923；ENSG00000115415；ENSG00000115607；ENSG00000116285；ENSG00000116761；ENSG00000119711；ENSG00000119725；ENSG00000120217；ENSG00000120738；ENSG00000120903；ENSG00000121380；ENSG00000121417；ENSG00000122497；ENSG00000124205；ENSG00000124702；ENSG00000124935；ENSG00000125255；ENSG00000128016；ENSG00000128266；ENSG00000128791；ENSG00000128891；ENSG00000130164；ENSG00000130487；ENSG00000130598；ENSG00000131095；ENSG00000131142；ENSG00000132199；ENSG00000132204；ENSG00000132915；ENSG00000132938；ENSG00000133636；ENSG00000133794；ENSG00000134028；ENSG00000134245；ENSG00000135148；ENSG00000135447；ENSG00000135625；ENSG00000136881；ENSG00000136883；ENSG00000136928；ENSG00000136933；ENSG00000137285；ENSG00000137463；ENSG00000137573；ENSG00000137709；ENSG00000137968；ENSG00000138166；ENSG00000138308；ENSG00000140274；ENSG00000140279；ENSG00000140323；ENSG00000140450；ENSG00000140465；ENSG00000140505；ENSG00000140718；ENSG00000141279；ENSG00000142178；ENSG00000142661；ENSG00000143185；ENSG00000143195；ENSG00000143320；ENSG00000143322；ENSG00000143367；ENSG00000143379；ENSG00000143603；ENSG00000144655；ENSG00000145248；ENSG00000145284；ENSG00000145358；ENSG00000145736；ENSG00000148541；ENSG00000148700；ENSG00000148702；ENSG00000149043；ENSG00000149289；

ENSG00000151012;ENSG00000151572;ENSG00000152672;ENSG00000153404;ENSG00000154227;ENSG00000154451;ENSG00000156414;ENSG00000157103;ENSG00000157680;ENSG00000158457;ENSG00000159231;ENSG00000159674;ENSG00000161609;ENSG00000162594;ENSG00000163029;ENSG00000163110;ENSG00000163285;ENSG00000163412;ENSG00000163635;ENSG00000163644;ENSG00000163735;ENSG00000163817;ENSG00000163884;ENSG00000164604;ENSG00000164821;ENSG00000165948;ENSG00000165973;ENSG00000165983;ENSG00000166923;ENSG00000167748;ENSG00000168004;ENSG00000168036;ENSG00000168062;ENSG00000168394;ENSG00000168661;ENSG00000168938;ENSG00000169248;ENSG00000170113;ENSG00000170442;ENSG00000170509;ENSG00000170837;ENSG00000171016;ENSG00000171408;ENSG00000171649;ENSG00000171714;ENSG00000172137;ENSG00000172183;ENSG00000172215;ENSG00000172667;ENSG00000173809;ENSG00000173812;ENSG00000173926;ENSG00000175764;ENSG00000175806;ENSG00000176046;ENSG00000177182;ENSG00000177294;ENSG00000178187;和ENSG00000178229。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0123] 在一些实施方案中，本公开内容提供了使用分类器区分UIP与非UIP的方法和/或系统，该分类器单独地或组合地包含2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个以下基因，或者由其组成：MPO;GGNBP2;SELE;FMO3;SLC6A13;EXTL3;NLRP2;BTN3A1;ABCF2;TLL1;HDAC9;CBLN4;CHRDL2;SNRNP40;MYO3B;SREBF1;GSDMB;MCCC1;CEACAM1;CXCL2;IL12RB2;ITGB5;ZNF671;ITPKC;RCOR1;TGM1;HSD17B7P2;DDTL;FAM118A;C14orf105;ADNP2;OLFM4;HAS3;SLC7A5;GYS1;FSD1;PLEKHA4;TMEM59L;RUNDC3B;LMBR1;VIPR2;CCL24;LARP4B;UBTF;RASD1;ODAM;CCND1;TSPAN11;SYT10;PRMT8;LTBR;CDK2AP1;GLP1R;VNN1;PCDHB2;LRRC31;SLC4A3;STAT1;IL18RAP;ERRFI1;CTH;ALDH6A1;ZNF410;CD274;EGR1;CHRNA2;BCL2L14;ZNF211;NBPF14;EDN3;KLHDC3;SCGB1D2;SLC10A2;ZFP36;GNAZ;TWSG1;C15orf57;LDLR;KLHDC7B;TNNI2;GFAP;CCL25;ENOSF1;LINC00470;PDE6A;MTUS2;NTS;ARNTL;ADAMDEC1;WNT2B;TRAFD1;PPP1R1A;EGR4;BAAT;KIF12;GABBR2;RABEPK;TUBB2B;MGARP;SULF1;POU2F3;SLC44A5;DUSP5;PLA2G12B;DUOXA2;DUOX2;DISP2;ARRDC4;CYP1A1;CYP1A2;FTO;NPEPPS;SIK1;MYOM3;XCL2;ILDR2;CRABP2;ABL2;TUFT1;SETDB1;KCNN3;CSRNP1;

SLC10A4;SCD5;DDIT4L;GTF2H2;FAM13C;ADD3;HABP2;SYT8;ZC3H12C;SLC7A11;ANO4;CLEC4F;PLEKHG4B;CERS3;GBP5;TDRD9;SLC6A1;DGKI;TSPAN33;CBR3;SPON2;CCDC155;IL23R;SMC6;PDLIM5;GABRG1;EIF4E3;ATXN7;PPM1K;CXCL5;SLC6A20;KLF15;GPR85;DEFA4;IFI27L1;NELL1;PTER;GREM1;KLK1;HRASLS5;CTNNB1;BATF2;TAP1;ZNF30;PPIC;CXCL11;NIPA1;KRT86;HSD17B13;GPR27;PYGO1;PDE7B;ZIK1;ANO5;CALB2;ISG20;CXCR6;ZMAT3;TDRD12;EIF1;MARCH3;TTLL11;MSRA;NUPR1;CLVS1;FBXO39;ZNF454;和ZNF543。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0124] 在一些实施方案中，本公开内容提供了使用本文所述的分类器区分UIP与非UIP的方法和/或系统，其中该方法进一步包括运行将受试者分类为吸烟者或非吸烟者的分类器。在一些情况下，这样的吸烟者状态分类可在运行UIP与非UIP分类器之前运行，或者吸烟者状态分类步骤可作为在训练(例如，使用分类器训练模块)本公开内容的UIP与非UIP分类器期间使用的协变量构建在其中。

[0125] 在特定实施方案中，本公开内容提供了使用Envisia分类器区分UIP与非UIP的方法和/或系统，其中该方法进一步包括运行将受试者分类为吸烟者或非吸烟者的分类器。在一些情况下，这样的吸烟者状态分类可在运行Envisia分类器之前运行，或者吸烟者状态分类步骤可作为在重新训练(例如，使用分类器训练模块)根据本公开内容的包含表5中所列基因的UIP与非UIP分类器期间使用的协变量构建到Envisia分类器中。

[0126] 在一些实施方案中，替代地或附加地，使用本文所述的分类器(例如，Envisia分类器)区分UIP与非UIP的方法和/或系统进一步包括排除某些基因或其变体或者向某些基因或其变体分配差异权重的步骤，该基因或其变体在训练(例如，使用分类器训练模块)或运行UIP与非UIP分类器期间对吸烟者状态偏差敏感。如本文所用，“吸烟者状态偏差”是指在非吸烟者患者中在UIP与非UIP患者中差异表达，但在吸烟者(或曾经吸烟者)中在UIP与非UIP患者中检测不到差异表达的基因或其变体。

[0127] 在一些实施方案中，本公开内容的方法和/或系统包括分层分类器，其包括至少第一分类器和第二分类器，其中对第一分类器进行训练(例如，使用分类器训练模块)以识别区分吸烟者与非吸烟者的基因特征，并且对第二分类器进行训练(例如，使用分类器训练模块)以分别区分吸烟者或非吸烟者中的UIP与非UIP。在一些这样的实施方案中，第二分类器是Envisia分类器。

[0128] 在一些实施方案中，替代地或附加地，使用本文所述的分类器来区分UIP与非UIP的方法和/或系统包括汇集从受试者获得的多个样品，然后测定汇集样品中存在的一组转录物的表达水平的步骤。在一些实施方案中，所述多个样品等于2、3、4或5个样品。在一些实施方案中，所述多个样品等于多于5个样品。在一些实施方案中，所述分类器包含SEQ ID NO:1-151中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、

119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150或151个或其任何组合，或者由其组成。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。在一些实施方案中，该分类器包含表5中所列基因中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个，或者由其组成。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0129] 在一些实施方案中，替代地或附加地，使用本文所述的分类器来区分UIP与非UIP的方法和/或系统包括计算机汇集从受试者获得的多个样品，然后测定多个样品中的每一个中存在的一组转录物的表达水平的步骤。这样的计算机汇集的一个实例描述于实施例6中。在一些实施方案中，计算机汇集的非限制性实例包括以下步骤：(i)测定从个体受试者获得的多个样品的第一样品中存在的一组转录物的表达水平；(ii)测定从个体受试者获得的多个样品的第二个样品中存在的相同或重叠的一组转录物的表达水平；(iii)在一些情况下，测定从个体受试者获得的多个样品中的一个或多个额外样品中相同或重叠的一组转录物(与第一和第二样品相比)的表达水平；(iv)缩放表达水平；(v)对表达水平取平均以产生“计算机汇集的”表达水平；(vi)对平均缩放表达水平进行方差稳定化转化(VST)，(vii)使用计算机汇集表达的VST进行评分；以及(viii)将评分与决策边界进行比较，并指定UIP/非UIP预测标签。

[0130] 在一些实施方案中，通过计算机汇集而汇集的多个样品中包含的来自受试者的样品数等于2、3、4或5个样品。在一些实施方案中，所述多个样品中的样品数等于多于5个样品。在一些实施方案中，所述分类器包含SEQ ID NO:1-151中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150或151个或其任何组合，或者由其组成。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些

前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0131] 在一些实施方案中，通过计算机汇集而汇集的多个样品中包含的来自受试者的样品数等于2、3、4或5个样品。在一些实施方案中，所述多个样品中的样品数等于多于5个样品。在一些实施方案中，所述分类器包含表5中所列基因中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个，或者由其组成。在特定方面，这样的分类器包含额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0132] 在一些特定实施方案中，使用从多个受试者获得的多个单独训练样品中表达的一种或多种转录物的表达水平来训练用于区分UIP与非UIP的计算机生成的分类器，每个训练样品具有UIP或非UIP的确认诊断(即，如本文所公开的“分类标签”或“真值标签”(参见例如图1))，其中至少两个训练样品从单个受试者获得。在一些实施方案中，本公开内容提供了使用这样的分类器(例如，Envisia分类器)检测汇集的肺组织测试样品是UIP呈阳性还是非UIP阳性的方法，其中该方法包括(A)测定测试样品中表达的一种或多种转录物的表达水平；以及(B)使用计算机生成并训练的分类器将测试样品分类为UIP或非UIP，其中通过物理汇集或通过计算机汇集来汇集测试样品。

[0133] 在一些实施方案中，通过分别训练所有样品实现最大表示和采样多样化，并且减小可用样品的先验亚采样偏差。此外，在一些实施方案中，通过使用汇集的样品进行分类步骤，减轻了采样影响。因此，在一些实施方案中，使用对单个(非汇集的)样品进行训练的分类器与已经汇集(物理汇集或通过计算机汇集)的测试样品提供了区分UIP与非UIP的改善的准确性。

[0134] 因此，在一个实施方案中，本公开内容提供了使用Envisia分类器检测来自受试者的汇集的肺组织测试样品是UIP阳性还是非UIP阳性的方法，该方法包括(A)测定在来自受试者的测试样品中表达的一种或多种转录物的表达水平；以及(B)使用计算机生成的Envisia分类器将测试样品分类为UIP或非UIP，其中测试样品包括已通过物理汇集或通过计算机汇集而汇集的来自受试者的多个样品。在一些实施方案中，所述多个样品包括2、3、4或5个样品。在一些实施方案中，所述多个样品包括多于5个样品。

[0135] 在一些实施方案中，使用本文描述的分类器(例如，Envisia分类器)区分UIP与非UIP的方法和/或系统包括在具有可变细胞组成的样品(例如，单个样品或样品池)中区分UIP与非UIP。在一些实施方案中，具有可变细胞组成的样品(例如，单个样品或样品池)包括1型肺泡细胞、2型肺泡细胞、细支气管细胞、肺祖细胞或其组合。在一些实施方案中，用于区

分UIP与非UIP的分类器的准确性不依赖于被分类的样品或者汇集样品的肺泡含量。如本文所用,术语“不受细胞组成影响”用于指示这样的分类器,其中用于区分UIP与非UIP的分类器的准确性不依赖于被分类的样品(例如,单个样品或样品池)的肺泡含量。

[0136] 在一些实施方案中,本公开内容提供了不受细胞组成影响的分类器,该分类器表现出分类器精确性与样品或汇集样品的肺泡含量之间的Pearson相关性小于约0.1;0.09;0.08;0.07;0.06;0.05;0.04;0.03;0.02;或小于约0.01。在一些实施方案中,本公开内容提供了不受细胞组成影响的分类器,该分类器表现出分类器精确性与样品或汇集样品的肺泡含量之间的Pearson相关性大于约-0.1;-0.09;-0.08;-0.07;-0.06;-0.05;-0.04;-0.03;-0.02;或大于约-0.01。在一些实施方案中,不受细胞组成影响的分类器是Envisia分类器。

[0137] 可通过任何合适的方法检测样品中的可变细胞组成。在一些实施方案中,使用细胞含量的半定量基因组测量来确定可变细胞组成。在一些实施方案中,细胞含量的半定量基因组测量确定样品中肺泡细胞的相对丰度。

[0138] 在一些实施方案中,这样的肺泡含量的半定量基因组测量包括能够确定样品中的肺泡1型细胞的相对丰度的度量(“肺泡1型细胞度量”)。在一些实施方案中,肺泡1型细胞度量包括一种或多种肺泡特异性基因。在一些实施方案中,所述一种或多种肺泡特异性基因是主要在肺泡1型细胞中表达的基因。在某些实施方案中,所述主要在肺泡1型细胞中表达的一种或多种肺泡特异性基因选自AQP5、PDPN或其组合。在特定实施方案中,AQP5、PDPN或其组合的表达与样品中肺泡1型细胞的丰度相关。在特定实施方案中,所述方法包括检测AQP5和PDPN的表达水平,在一些情况下对该表达水平进行归一化,并对这些基因的表达水平求和,其中高表达水平表明样品中的较高肺泡1型细胞含量;低表达水平表明样品中的较低肺泡1型细胞含量;并且中等表达水平表明样品中的中等肺泡1型细胞含量。

[0139] 在特定实施方案中,本公开内容提供了确定存在于两个或更多个样品中的1型肺泡细胞的相对丰度的方法,包括(i)测定从单个受试者获得的第一样品中主要在肺泡1型细胞中表达的肺泡特异性基因的一种或多种转录物的表达水平;(ii)测定从单个受试者获得的第二样品中主要在肺泡1型细胞中表达的肺泡特异性基因的相同的一种或多种转录物的表达水平;(iii)以及比较两个样品之间的所述一种或多种转录物的表达水平,以确定样品中存在的1型肺泡细胞的相对丰度。在一些实施方案中,所述主要在肺泡1型细胞中表达的一种或多种肺泡特异性基因选自AQP5、PDPN或其组合。在一些实施方案中,所述主要在肺泡1型细胞中表达的一种或多种肺泡特异性基因包括AQP5、PDPN或其组合。在一些实施方案中,所述主要在肺泡1型细胞中表达的一种或多种肺泡特异性基因包括AQP5和PDPN两者。在一些实施方案中,第一样品和第二样品从不同的受试者获得。在一些实施方案中,第一样品和第二样品从相同的受试者获得。在一些实施方案中,所述方法进一步包括测定从单个受试者获得的至少一个另外的样品中主要在肺泡1型细胞中表达的肺泡特异性基因的一种或多种转录物的表达水平,然后比较所述至少一个另外的样品中的表达水平与第一和/或第二样品中的表达水平,以确定样品中存在的1型肺泡细胞的相对丰度。在一些实施方案中,至少两个样品从相同的受试者获得。在一些实施方案中,至少3、4、5个或更多个样品从相同的受试者获得。在一些实施方案中,所有两个样品从不同的受试者获得。

[0140] 在一些实施方案中,本公开内容提供了肺泡含量的半定量基因组测量,其包括能够确定样品中的肺泡2型细胞的相对丰度的度量(“肺泡2型细胞度量”)。在一些实施方案

中，该度量包括一种或多种肺泡特异性基因。在一些实施方案中，所述一种或多种肺泡特异性基因是主要在肺泡2型细胞中表达的基因。在某些实施方案中，所述主要在肺泡2型细胞中表达的一种或多种肺泡特异性基因选自SFTPB、SFTPC、SFTPD或其组合。在某些实施方案中，所述主要在肺泡2型细胞中表达的一种或多种肺泡特异性基因包括SFTPB、SFTPC、SFTPD或其组合。在某些实施方案中，所述主要在肺泡2型细胞中表达的一种或多种肺泡特异性基因包括SFTPB、SFTPC和SFTPD。在一些实施方案中，肺泡2型细胞度量还包括在肺泡1型细胞和肺泡2型细胞中均表达的一种或多种肺泡特异性基因。在某些实施方案中，在肺泡1型细胞和肺泡2型细胞中均表达的基因是SFTPA1。在特定实施方案中，所述度量包括主要在肺泡2型细胞中表达的一种或多种肺泡特异性基因以及在肺泡1型细胞和肺泡2型细胞中均表达的一种或多种基因。在特定实施方案中，所述度量包括SFTPB、SFTPC、SFTPD、SFTPA1或其组合。

[0141] 在特定实施方案中，本公开内容提供了确定样品中的2型肺泡细胞的相对丰度的方法，包括检测FTPB、SFTPC、SFTPD、SFTPA1或其组合的表达水平，在一些情况下对表达水平进行归一化，并对这些基因的表达水平求和，其中高表达水平表明样品中的较高肺泡2型细胞含量；低表达水平表明样品中的较低肺泡2型细胞含量；并且中等表达水平表明样品中的中等肺泡2型细胞含量。

[0142] 在特定实施方案中，本公开内容提供了确定存在于两个或更多个样品中的2型肺泡细胞的相对丰度的方法；该方法包括(i)测定从单个受试者获得的第一样品中主要在肺泡2型细胞中表达的肺泡特异性基因的一种或多种转录物的表达水平；(ii)测定从单个受试者获得的第二样品中主要在肺泡2型细胞中表达的肺泡特异性基因的相同的一种或多种转录物的表达水平；(iii)以及比较两个样品之间的所述一种或多种转录物的表达水平，以确定样品中存在的2型肺泡细胞的相对丰度。在一些这样的实施方案中，所述主要在肺泡2型细胞中表达的一种或多种肺泡特异性基因选自SFTPB、SFTPC和SFTPD，及其组合。在特定实施方案中，所述方法包括测定第一和第二样品中的SFTPB、SFTPC和SFTPD中每一种的表达。替代地或附加地，在各个实施方案中，该方法进一步包括测定第一样品和第二样品中的一种或多种额外的基因的表达水平。在一些实施方案中，所述一种或多种额外的基因包括主要在肺泡细胞中表达的基因。在一些实施方案中，所述额外的基因在肺泡1型细胞和肺泡2型细胞中均有表达。在特定实施方案中，所述额外的基因为SFTPA1。在一些实施方案中，第一样品和第二样品从不同的受试者获得。在一些实施方案中，第一样品和第二样品从相同的受试者获得。在一些实施方案中，所述方法进一步包括测定从单个受试者获得的至少一个另外的样品中主要在肺泡1型细胞和/或肺泡1型细胞和肺泡2型细胞两者中表达的肺泡特异性基因的一种或多种转录物的表达水平，然后比较至少一个另外的样品中的表达水平与第一和/或第二样品中的表达水平，以确定样品中存在的2型肺泡细胞的相对丰度。在一些实施方案中，至少两个样品从相同的受试者获得。在一些实施方案中，至少3、4、5个或更多个样品从相同的受试者获得。在一些实施方案中，所有两个样品从不同的受试者获得。

[0143] 本文公开的方法可包括将信息基因的表达水平与一种或多种适当参考进行比较。“适当参考”是指示已知的肺ILD状态(即，UIP与非UIP；IPF与非IPF)的特定信息基因的表达水平(或表达水平范围)。适当参考可由所述方法的实施者通过实验确定，或者可以是预先存在的值或值范围。适当参考表示指示UIP/非UIP状态的表达水平(或表达水平范围)。例

如,适当参考可代表已知表达UIP的参考(对照)生物样品中的信息基因的表达水平。当适当参考指示UIP时,从需要表征或诊断UIP的受试者确定的表达水平与适当参考之间缺乏可检测的差异(例如,缺乏统计上显著的差异)可指示受试者中的UIP。当适当参考指示UIP时,从需要表征或诊断UIP的受试者确定的表达水平与适当参考之间的差异可指示受试者无UIP(即,非UIP)。

[0144] 或者,适当参考可以是指示受试者无UIP(即非UIP)的基因的表达水平(或表达水平范围)。例如,适当参考可代表从已知无UIP的受试者获得的参考(对照)生物样品中的特定信息基因的表达水平。当适当参考指示受试者无UIP时,从需要表征或诊断UIP的受试者确定的表达水平与适当参考之间的差异可指示受试者中的UIP。或者,当适当参考指示受试者无UIP时,从需要诊断UIP的受试者确定的表达水平与适当参考水平之间缺乏可检测的差异(例如,缺乏统计上显著的差异)可指示受试者无UIP。

[0145] 在一些实施方案中,参考标准提供阈值变化水平,使得如果样品中基因的表达水平在阈值变化水平内(根据特定标记增大或减小),则将受试者鉴定为无UIP,但如果该水平高于阈值,则将受试者鉴定为有患UIP的风险。

[0146] 在一些实施方案中,所述方法涉及将信息基因的表达水平与参考标准进行比较,该参考标准代表被鉴定为不具有UIP的对照受试者中的信息基因的表达水平。该参考标准可以是例如被鉴定为不具有UIP的对照受试者群体中的信息基因的平均表达水平。

[0147] 表达水平与适当参考之间的统计上显著的差异大小可有所变化。例如,当生物样品中信息基因的表达水平比该基因的适当参考高或低至少1%、至少5%、至少10%、至少25%、至少50%、至少100%、至少250%、至少500%或至少1000%时,可检测到指示UIP的显著差异。类似地,当生物样品中信息基因的表达水平高达该基因的适当参考的以下倍数或比其低以下倍数时,可检测到显著差异:至少1.1倍、1.2倍、1.5倍、2倍、至少3倍、至少4倍、至少5倍、至少6倍、至少7倍、至少8倍、至少9倍、至少10倍、至少20倍、至少30倍、至少40倍、至少50倍、至少100倍或者更高或更低时。在一些实施方案中,信息基因与适当参考之间的表达的至少20%至50%的差异是显著的。通过使用适当的统计检验可鉴定显著差异。统计显著性检验的实例提供于Applied Statistics for Engineers and Scientists by Petruccelli,Chen and Nandram1999重印版,其通过引用整体并入本文。

[0148] 应当理解,可将多种表达水平与多种适当参考水平进行比较,例如,基于逐个基因,以评估受试者的UIP状态。可以以矢量差异进行比较。在这样的情况下,多变量检验例如Hotelling T2检验可用于评估观察到的差异的显著性。这样的多变量检验的实例提供于Applied Multivariate Statistical Analysis by Richard Arnold Johnson and Dean W.Wichern Prentice Hall;第6版(2007年4月2日),其通过引用整体并入本文。

分类方法

[0149] 所述方法还可涉及将从受试者获得的生物样品中的信息基因的表达水平集(称为表达模式或表达谱)与多个参考水平集(称为参考模式)进行比较,每个参考模式与已知的UIP状态相关联,鉴定最接近类似于表达模式的参考模式,并将参考模式的已知UIP状态与表达模式相关联,从而对受试者的UIP状态进行分类(表征)。

[0150] 所述方法还可涉及构建或构造预测模型,该模型也可被称为可用于对受试者的疾病状态进行分类的分类器或预测器。如本文所用,“UIP分类器”是预测模型,其基于在从受

试者获得的生物样品中确定的表达水平来表征受试者的UIP状态。通常，使用已经确定了分类(UIP状态)的样品构建模型。一旦模型(分类器)得到构建，其随后可应用于从UIP状态未知的受试者的生物样品获得的表达水平，以便预测受试者的UIP状态。在特定实施方案中，该UIP分类器是Envisia分类器。因此，所述方法可以涉及将UIP分类器(例如，Envisia分类器)应用于表达水平，使得UIP分类器基于表达水平表征受试者的UIP状态。可基于预测的UIP状态由例如医疗保健提供者进一步治疗或评价受试者。在一些实施方案中，可基于预测的UIP状态(例如，基于通过将分类器应用于来自从受试者获得的测试样品的基因表达数据而确定的UIP的分类)，用选自吡非尼酮、尼达尼布或其药学上可接受的盐的化合物治疗受试者。测试样品可包括来自受试者的多个物理或计算机汇集的样品(例如，至少1、2、3、4、5个或更多个样品)。

[0151] 分类方法可涉及将表达水平转化为UIP风险评分，该评分指示受试者患有UIP的可能性。在一些实施方案中，例如，当使用诸如GLMNET等弹性网络回归模型时，可获得UIP风险评分作为加权表达水平的组合(例如，总和、乘积或其他组合)，其中表达水平按照它们对预测患有UIP的可能性增加的相对贡献来加权。

[0152] 可使用各种预测模型作为UIP分类器。例如，UIP分类器可包括选自逻辑回归、偏最小二乘、线性判别分析、二次判别分析、神经网络、朴素贝叶斯、C4.5决策树、k最近邻、随机森林、支持向量机或其他适当方法的算法。

[0153] UIP分类器可用包括从被鉴定为患有UIP的多个受试者获得的生物样品中的多个信息基因的表达水平的数据集进行训练。例如，UIP分类器可用包括从基于组织学发现被鉴定为患有UIP的多个受试者获得的生物样品中的多个信息基因的表达水平的数据集进行训练。训练集通常还将包括被鉴定为不具有UIP的对照受试者。如本领域技术人员将理解的，训练数据集的受试者群体可通过设计而具有各种特性，例如，群体的特性可取决于使用分类器的诊断方法可能有用的受试者的特性。例如，群体可由全部男性、全部女性组成，或可由男性和女性组成。群体可由具有癌症史的受试者、没有癌症史的受试者或来自两个类别的受试者组成。群体可包括吸烟者、曾经吸烟者和/或非吸烟者。

[0154] 还可测量类别预测强度以确定模型对生物样品进行分类的置信度。该置信度可用作受试者属于由模型预测的特定类别的可能性的估计。

[0155] 因此，预测强度传达样品分类的置信度，并评估样品何时不能被分类。可存在样品被测试但不属于或无法可靠地分配给特定类别的情况。这可通过例如利用阈值或范围来实现，其中评分高于或低于确定的阈值或者在特定范围内的样品不是可被分类的样品(例如，“无分类(no call)”)。

[0156] 一旦建立了模型，就可使用各种方法测试模型的有效性。测试模型有效性的一种方法是通过数据集的交叉验证。为了执行交叉验证，消除样品中的一个或子集，并且如上所述在没有所消除的样品的情况下构建模型，形成“交叉验证模型”。然后如本文所述，根据该模型对消除的样品进行分类。该过程用初始数据集的所有样品或子集完成，并确定错误率。然后评估该模型的准确性。对于已知的类别或先前已确定的类别，该模型以高精确性对待测试的样品进行分类。验证模型的另一种方法是将模型应用于独立的数据集，诸如具有未知UIP状态的新生物样品。

[0157] 如本领域技术人员将会理解的，可以用各种参数来评估模型的强度，该参数包括

但不限于准确性、灵敏度和特异性。本文描述了用于计算准确性、灵敏度和特异性的各种方法(参见例如实施例)。UIP分类器可具有至少60％、至少65％、至少70％、至少75％、至少80％、至少85％、至少90％、至少95％、至少99％或更高的准确性。UIP分类器可具有约60％至70％、70％至80％、80％至90％或90％至100％的准确性。UIP分类器可具有至少60％、至少65％、至少70％、至少75％、至少80％、至少85％、至少90％、至少95％、至少99％或更高的灵敏度。UIP分类器可具有约60％至70％、70％至80％、80％至90％或90％至100％的灵敏度。UIP分类器可具有至少60％、至少65％、至少70％、至少75％、至少80％、至少85％、至少90％、至少95％、至少99％或更高的特异性。UIP分类器可具有约60％至70％、70％至80％、80％至90％或90％至100％的特异性。

[0158] 阴性预测值(NPV)可大于或等于40％、41％、42％、43％、44％、45％、46％、47％、48％、49％、50％、51％、52％、53％、54％、55％、56％、57％、58％、59％、60％、61％、62％、63％、64％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％或99％，以在意向使用群体(例如，受试者，诸如患者)中排除UIP。当UIP被排除时，可划入非UIP。

[0159] UIP分类器可具有大于或等于40％、41％、42％、43％、44％、45％、46％、47％、48％、49％、50％、51％、52％、53％、54％、55％、56％、57％、58％、59％、60％、61％、62％、63％、64％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％或99％的阳性预测值(PPV)，以划入UIP。当UIP被划入时，可排除非UIP。

[0160] 意向使用群体可具有等于或约为40％、41％、42％、43％、44％、45％、46％、47％、48％、49％、50％、51％、52％、53％、54％、55％、56％、57％、58％、59％、60％、61％、62％、63％、64％、65％、66％、67％、68％、69％、70％、71％、72％、73％、74％、75％、76％、77％、78％、79％、80％、81％、82％、83％、84％、85％、86％、87％、88％、89％、90％、91％、92％、93％、94％、95％、96％、97％、98％或99％的癌症患病率。

[0161] 在一些实施方案中，本公开内容的方法和/或系统包括：从测试样品(例如，肺组织)提取核酸(例如RNA，例如总RNA)；扩增核酸以产生经表达的核酸文库(例如，通过聚合酶链反应介导的cDNA(在一些情况下为标记的cDNA)扩增，其中可通过逆转录(RT-PCR)从一个或多个RNA样品产生cDNA)；通过阵列(例如，微阵列)或通过直接测序(例如，RNAseq)检测核酸文库中存在的一种或多种核酸的表达(例如，通过测量由RT-PCR产生的cDNA种类来检测RNA表达谱)；以及使用本文所述的经训练的分类器(例如，Envisia分类器)确定测试样品是UIP还是非UIP。

[0162] 在一些实施方案中，本公开内容的方法和/或系统进一步包括将吸烟者状态并入训练练习中。在某些实施方案中，吸烟者状态在一些情况下以以下方式之一并入：

(i)通过在训练(例如，使用分类器训练模块)期间使用吸烟状态作为UIP或非UIP分类器中的协变量。

(ii)通过鉴定多个对吸烟者状态偏差敏感的基因，并且在UIP或非UIP分类器训练(例如，使用分类器训练模块)期间排除此类基因，或在一些情况下对此类基因与对此类偏差不

敏感的基因不同地进行加权。

(iii)通过构建分层分类,其中被训练(例如,使用分类器训练模块)以识别区分吸烟者与非吸烟者的基因特征的初始分类器被用于基于测试样品的基因特征将测试样品预分类为“吸烟者”或“非吸烟者”;然后,在预分类之后,运行被训练(例如,使用分类器训练模块)以区分吸烟者或非吸烟者中的UIP与非UIP的不同分类器。例如,如果预分类器确定测试样品来自吸烟者,则使用利用来自吸烟者的UIP和非UIP样品训练(例如,使用分类器训练模块)的分类器执行UIP与非UIP分类。相反,如果预分类器确定测试样品来自非吸烟者,则使用利用来自非吸烟者的UIP和非UIP样品训练(例如,使用分类器训练模块)的分类器执行UIP与非UIP分类。在一些实施方案中。这样的吸烟者或非吸烟者特异性分类器至少部分地由于降低了在分类器训练中包含对吸烟者状态偏差敏感的基因引起的背景噪声而提供改善的诊断性能。

[0163] 因此,本公开内容还提供了在区分UIP与非UIP的方法中使用的合适的分类器,如本文所公开的(例如,Envisia分类器)。在各个实施方案中,本公开内容提供了适用于区分UIP与非UIP的分类器,其中使用来自对应于由病理学专家确定的一种或多种组织病理学标签的样品(例如,单个样品或汇集样品)的微阵列、qRT-PCR或测序数据训练(例如,使用分类器训练模块,例如Envisia分类器)分类器。在一些实施方案中,样品被标记为UIP或非UIP。

[0164] 在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表1和/或表15中提供的一个或多个序列或其片段或者来自表1和/或表15的至少一个序列或其片段,或者由其组成。在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表1和/或表15中的任一个或多个或者全部中提供的至少1、2、3、4、5、6、7、8、9、10个或更多个序列或者由其组成。例如,在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表1中提供的至少11、12、13、14、15、20、30、50、100、150、151个序列或者由其组成,包括之间的所有整数(例如,16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表5、7、8、9、10、11或12中的任一个或多个或者全部的约1-10个序列,来自表1和/或表15中的任一个或多个或者全部的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-151个序列)。在一个实施方案中,本公开内容提供了一种分类器,其包含表1和/或表15中提供的所有序列或者由其组成。

[0165] 在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表5中提供的一个或多个序列或其片段或者来自表5的至少一个序列或其片段,或者由其组成。在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表5中提供的至少1、2、3、4、5、6、7、8、9、10个或更多个序列或者由其组成。例如,在一些实施方案中,本公开内容提供了一种分类器,该分类器包含表5中提供的至少11、12、13、14、15、20、30、50、100、150、160、170、180或190个序列或者由其组成,包括之间的所有整数(例如,16、17、18、19、21、22、23、24、25个序列等)和范围(例如,来自表5、7、8、9、10、11或12中的任一个或多个或者全部的约1-10个序列,来自表5的约10-15个序列、10-20个序列、5-30个序列、5-50个序列、10-100个序列、50-150个序列、60-190个序列)。在一个实施方案中,本公开内容提供了一种分类器,其包含表5中提供的所有序列或者由其组成。

[0166] 在一些特定实施方案中,本公开内容提供了用于区分UIP与非UIP的分类器,其中该分类器包含SEQ ID NO:1-151中的一个或多个或其片段或其任何组合,或者由其组成。在

一个实施方案中，该分类器包含上述序列中的全部151个或者由其组成。在一些实施方案中，本公开内容提供了用于区分UIP与非UIP的分类器，其中该分类器包含上述151个序列中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150或151个，或者由其组成。在特定方面，该分类器包含1、2、3、4、5、6、7、8个或更多个额外的基因或其片段。在其他方面，该分类器略去上述151个序列中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。在其他方面，所述151个基因中的每一个可与其他基因中的任一个或多个或至多20个组合使用。

[0167] 在一些特定实施方案中，本公开内容提供了用于区分UIP与非UIP的分类器，其中该分类器包含表5中所列基因中的一个或多个或其片段或其任何组合，或者由其组成。在一个实施方案中，该分类器包含表5中所列基因中的全部190个或者由其组成。在一些实施方案中，本公开内容提供了用于区分UIP与非UIP的分类器，其中该分类器包含表5中所列的上述190个基因中的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个，或者由其组成：在特定方面，该分类器包含表5中所列的190个基因以及1、2、3、4、5、6、7、8个或更多个额外的基因或其片段。在其他方面，该分类器略去表5中所列的上述190个基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。在其他方面，所述190个基因中的每一个可与其他基因中的任一个或多个或至多20个组合使用，以根据本文公开的方法将样品分类为UIP或非UIP。

[0168] 在某些实施方案中，本公开内容提供了改善肺组织样品中疾病或病况的检测的方法，该方法包括(A)测定在测试样品中表达的一种或多种转录物的表达水平；(B)使用计算机生成的经训练的分类器(例如，Envisia分类器)将测试样品分类为对疾病或病况呈阳性或阴性；其中使用在从多个受试者获得的多个单独的训练样品中表达的一种或多种转录物的表达水平训练所述计算机生成的经训练的分类器，每个训练样品具有对疾病或病况呈阳性或阴性的确认诊断，其中至少两个训练样品从单个受试者获得；并且其中在分类之前汇集测试样品。

组织样品

[0169] 在主题分析或诊断方法中使用的肺组织样品可以是活检样品(例如，通过电视辅

助胸腔镜手术VATS获得的活检样品);支气管肺泡灌洗(BAL)样品;经支气管活检物;低温经支气管活检物;等等。用于分析的肺组织样品可在合适的保存溶液中提供。在一些实施方案中,通过辅助性支气管镜程序如擦刷(诸如通过细胞刷、组织刷)、支气管活检、支气管灌洗或针吸获得组织样品。在一些实施方案中,可通过口腔清洗、触摸制备或痰收集来获得组织样品。在一些实施方案中,从受试者的呼吸上皮获得组织样品。呼吸上皮可来自口、鼻、咽、气管、支气管、细支气管或肺泡。然而,还可使用其他呼吸上皮来源。

[0170] 可基于患者所呈现的临床体征和症状(例如,呼吸短促(通常通过劳累而加重)、干咳),并且在一些情况下基于成像检查(例如,胸部X光、计算机断层扫描(CT))、肺功能检查(例如,肺量测定法、血氧测定法、运动压力测试)和/或肺组织分析(例如,通过支气管镜检查、支气管肺泡灌洗、手术活检获得的样品的组织学和/或细胞学分析)中的一种或多种,从疑似患有肺病例如ILD的患者获得组织样品。在一些情况下,对于肺病的存在或不存在,组织样品的细胞学或组织学分析可能是模糊的或可疑的(或不确定的)。

[0171] 肺组织样品可以以多种方式中的任一种进行处理。例如,肺组织样品可进行细胞裂解。肺组织样品可保存在RNA保护溶液(抑制RNA降解的溶液,例如,抑制RNA的核酸酶消化的溶液)中,并随后进行细胞裂解。可从肺组织样品富集或分离诸如核酸和/或蛋白质等组分,并且可以在主题方法中使用富集或分离的组分。可使用富集和分离诸如核酸等组分的各种方法。可使用分离RNA以供表达分析的各种方法。

测定表达产物水平的体外方法

[0172] 用于测定基因表达产物水平的方法可包括但不限于以下一种或多种:其他细胞学分析、对特定蛋白质或酶活性的分析、对包括蛋白质或RNA或特定RNA剪接变体在内的特定表达产物的分析、原位杂交、全基因组或部分基因组表达分析、微阵列杂交分析、基因表达的系列分析(SAGE)、酶联免疫吸附测定、质谱法、免疫组织化学、印迹法、测序、RNA测序(例如,外显子组富集RNA测序)、DNA测序(例如,从RNA获得的cDNA的测序);下一代测序、纳米孔测序、焦磷酸测序或Nanostring测序。例如,可根据Kim等人(Lancet Respir Med.2015Jun;3(6):473-82,整体并入本文,包括所有补充)描述的方法测定基因表达产物水平。如本文所用,术语“测定”或“检测”或“确定”在指代测定基因表达产物水平时可互换使用。在实施方案中,上述测定基因表达产物水平的方法适用于检测或测定基因表达产物水平。基因表达产物水平可相对于内标如总mRNA或特定基因的表达水平进行归一化,该内标包括但不限于甘油醛-3-磷酸脱氢酶或微管蛋白。

[0173] 在各个实施方案中,样品包括从组织样品(例如,肺组织样品,如TBB样品)收获的细胞。可使用各种技术从样品收获细胞。例如,可通过离心细胞样品并重悬浮沉淀的细胞来收获细胞。可将细胞重悬浮于缓冲溶液如磷酸盐缓冲盐水(PBS)中。在离心细胞悬浮液以获得细胞沉淀物后,可裂解细胞以提取核酸,例如信使RNA(mRNA)。从受试者获得的所有样品,包括经受任何程度的进一步处理的那些样品,被认为是从受试者获得的。

[0174] 在一个实施方案中,在如本文所述进行基因表达产物的检测之前,进一步处理样品。例如,细胞或组织样品中的mRNA可与样品的其他组分分离。当mRNA不在其天然环境中时,可浓缩和/或纯化样品从而以非天然状态分离mRNA。例如,研究已经表明,体内mRNA的高级结构不同于相同序列的体外结构(参见例如,Rouskin等人(2014).Nature505,第701-705页,出于所有目的整体并入本文)。

[0175] 在一个实施方案中，来自样品的mRNA与合成DNA探针杂交，该合成DNA探针在一些实施方案中包含检测部分(例如，可检测标签、捕获序列、条形码报告序列)。因此，在这些实施方案中，最终制备非天然mRNA-cDNA复合物并用于检测基因表达产物。在另一个实施方案中，来自样品的mRNA直接用可检测标签例如荧光团进行标记。在进一步的实施方案中，非天然的标记mRNA分子与cDNA探针杂交，并检测复合物。

[0176] 在一个实施方案中，一旦从样品中获得mRNA，就在杂交反应中将其转化为互补DNA(cDNA)，或与一种或多种cDNA探针一起用于杂交反应。cDNA不存在于体内，因此是非天然分子。此外，cDNA-mRNA杂合体是合成的并且不存在于体内。除了cDNA不存在于体内之外，cDNA必然不同于mRNA，因为其包含脱氧核糖核酸而不是核糖核酸。然后例如通过聚合酶链反应(PCR)或其他扩增来扩增cDNA。例如，可采用的其他扩增方法包括连接酶链反应(LCR)(Wu和Wallace,Genomics,4:560(1989)，Landegren等人,Science,241:1077(1988)，出于所有目的通过引用整体并入)、转录扩增(Kwoh等人,Proc.Natl.Acad.Sci.USA,86:1173(1989)，出于所有目的通过引用整体并入)、自动维持序列复制(Guatelli等人，Proc.Nat.Acad.Sci.USA,87:1874(1990)，出于所有目的通过引用整体并入)，以及基于核酸的序列扩增(NASBA)。用于选择PCR扩增引物的指南的实例提供于McPherson等人,PCR Basics:From Background to Bench,Springer-Verlag,2000，出于所有目的通过引用整体并入。该扩增反应的产物，即扩增的cDNA，也必然是非天然产物。首先，如上所述，cDNA是非天然分子。其次，在PCR的情况下，扩增过程用于为起始材料的每个单个cDNA分子产生数亿个cDNA拷贝。产生的拷贝数与体内存在的mRNA拷贝数相差甚远。

[0177] 在一个实施方案中，用引物扩增cDNA，该引物将额外的DNA序列(例如，衔接子、报道基因、捕获序列或部分、条形码)引入片段上(例如，通过使用衔接子特异性引物)，或者mRNA或cDNA基因表达产物序列与包含额外序列(例如，衔接子、报道基因、捕获序列或部分、条形码)的cDNA探针直接杂交。因此，mRNA的扩增和/或与cDNA探针的杂交用于通过引入额外序列并形成非天然杂合体从非天然单链cDNA或mRNA产生非天然双链分子。此外，扩增程序具有与它们相关的错误率。因此，扩增引入了对cDNA分子的进一步修饰。在一个实施方案中，在用衔接子特异性引物扩增期间，将可检测标签(例如，荧光团)添加到单链cDNA分子中。因此，扩增也可用于产生自然界中不存在的DNA复合物，至少是因为(i)cDNA不存在于体内，(i)将衔接子序列添加到cDNA分子的末端以产生不存在于体内的DNA序列，(ii)与扩增相关的错误率进一步产生不存在于体内的DNA序列，(iii)与天然存在的结构相比，cDNA分子的结构不同，以及(iv)将可检测标签化学添加至cDNA分子。

[0178] 在一些实施方案中，通过检测非天然cDNA分子而在核酸水平上检测感兴趣的基因表达产物的表达。

[0179] 本文所述的基因表达产物包括包含任何感兴趣的核酸序列的全部或部分序列的RNA，或在逆转录反应中体外经合成获得的其非天然cDNA产物。术语“片段”意指多核苷酸的一部分，其通常包含至少10、15、20、50、75、100、150、200、250、300、350、400、450、500、550、600、650、700、800、900、1,000、1,200或1,500个连续核苷酸，或至多本文公开的全长基因表达产物多核苷酸中存在的核苷酸数。基因表达产物多核苷酸的片段通常会编码至少15、25、30、50、100、150、200或250个连续氨基酸，或至多本公开内容的全长基因表达产物蛋白质中存在的氨基酸总数。

[0180] 在某些实施方案中,基因表达谱可通过全转录组鸟枪法测序("WTSS"或"RNAseq";参见例如,Ryan等人,BioTechniques 45:81-94)获得,其利用高通量测序技术对cDNA进行测序以获得关于样品的RNA含量的信息。一般而言,由RNA产生cDNA,对cDNA进行扩增,并对扩增产物进行测序。

[0181] 扩增后,可使用任何方便的方法对cDNA或其衍生物进行测序。例如,可使用Illumina的可逆终止子法、Roche的焦磷酸测序法(454)、Life Technologies的连接测序(SOLiD平台)或Life Technologies的Ion Torrent平台对片段进行测序。此类方法的实例描述于以下参考文献中:Margulies等人(Nature 2005 437:376-80);Ronaghi等人(Analytical Biochemistry 1996 242:84-9);Shendure(Science 2005 309:1728);Imelfort等人(Brief Bioinform.2009 10:609-18);Fox等人(Methods Mol Biol.2009;553:79-108);Appleby等人(Methods Mol Biol.2009;513:19-39)和Morozova(Genomics.2008 92:255-64),其通过引用并入,关于对方法的一般描述和方法的具体步骤,包括每个步骤的所有起始产物、试剂和最终产物。显而易见的是,可在扩增步骤期间将与选择的下一代测序平台相容的正向和反向测序引物位点添加到片段的末端。

[0182] 在其他实施方案中,可用纳米孔测序对产物进行测序(例如,如Soni等人,Clin Chem 53:1996-2001 2007所述,或如Oxford Nanopore Technologies所述)。纳米孔测序是一种单分子测序技术,其中单个DNA分子在穿过纳米孔时被直接测序。纳米孔是直径为1纳米的小孔。将纳米孔浸入导电流体中并在其上施加电势(电压)导致由于离子通过纳米孔的传导而产生的轻微电流。流动的电流的量对纳米孔的大小和形状敏感。当DNA分子穿过纳米孔时,DNA分子上的每个核苷酸以不同程度阻塞纳米孔,从而以不同程度改变通过纳米孔的电流的大小。因此,当DNA分子穿过纳米孔时,这种电流变化代表DNA序列的读取。纳米孔测序技术公开于美国专利号5,795,782、6,015,714、6,627,067、7,238,485和7,258,838,以及美国专利申请公开US2006003171和US20090029477中。

[0183] 在一些实施方案中,主题方法的基因表达产物是蛋白质,并且使用来源于从样品队列获得的蛋白质数据的分类器来分析特定生物样品中的蛋白质的量。可通过以下一种或多种方法测定蛋白质的量:酶联免疫吸附测定(ELISA)、质谱法、印迹法或免疫组织化学。

[0184] 在一些实施方案中,可通过使用例如Affymetrix阵列、cDNA微阵列、寡核苷酸微阵列、点样微阵列或来自Biorad、Agilent或Eppendorf的其他微阵列产品的微阵列分析来确定基因表达产物标志物和可变剪接标志物。微阵列提供特别的优点,因为它们可以包含可以在单个实验中测定的大量基因或可变剪接变体。在一些情况下,微阵列装置可以包含允许综合评价基因表达模式、基因组序列或可变剪接的整个人类基因组或转录物组或其大部分。可使用如Sambrook,Molecular Cloning a Laboratory Manual 2001以及Baldi,P.和Hatfield,W.G.,DNA Microarrays and Gene Expression 2002描述的标准分子生物学和微阵列分析技术发现标志物。

[0185] 微阵列分析通常开始于使用各种方法从生物样品(例如活检物或细针抽吸物)中提取和纯化核酸。对于表达和可变剪接分析,可以有利地从DNA提取和/或纯化RNA。此外可能有利的是从其他形式的RNA例如tRNA和rRNA中提取和/或纯化niRNA。

[0186] 例如,可以通过逆转录、聚合酶链反应(PCR)、连接、化学反应或其他技术,用荧光标记、放射性核素或化学标记如生物素、地高辛配基或地高辛来进一步标记纯化的核酸。标

记可以是直接或间接的，其可能进一步需要偶联阶段。偶联阶段可以发生在杂交之前，例如，使用氨基烯丙基-UTP和NHS氨基反应性染料(如花青染料)，或在杂交之后，例如，使用生物素和标记的链霉抗生物素蛋白。在一个实例中，以低于正常核苷酸的速率酶促添加修饰的核苷酸(例如以1aaUTP:4TTP的比例)，从而通常导致每60个碱基中有1个修饰的核苷酸(用分光光度计测量)。然后，可用例如柱或渗滤装置纯化aaDNA。氨基烯丙基是附接到与反应性标签(例如荧光染料)反应的核碱基上的长接头上的胺基团。

[0187] 标记的样品然后可与杂交溶液混合，该杂交溶液可包含十二烷基硫酸钠(SDS)、SSC、硫酸葡聚糖、封闭剂(如COT1DNA、鲑精DNA、小牛胸腺DNA、聚A或聚T)、Denhardt溶液、甲酰胺(formamine)或其组合。

[0188] DNA杂交探针是可变长度的DNA或RNA片段，其用于检测DNA或RNA样品中与探针中的序列互补的核苷酸序列(DNA靶标)的存在。因此所述探针与单链核酸(DNA或RNA)杂交，该单链核酸的碱基序列由于探针与靶标之间的互补性而允许探针-靶碱基配对。标记的探针首先(通过加热或在碱性条件下)变性成单链DNA，然后与靶DNA杂交。

[0189] 为了检测探针与其靶序列的杂交，用分子标记物标示(或标记)所述探针；常用的标记物是32P或地高辛配基，后者是非放射性的基于抗体的标记物。然后通过经由放射自显影或其他成像技术使杂交的探针可视化来检测与探针具有中等至高度序列互补性(例如，至少70％、80％、90％、95％、96％、97％、98％、99％或更高的互补性)的DNA序列或RNA转录物。具有中等或高度互补性的序列的检测取决于应用多严格的杂交条件—高严格性，例如高杂交温度和杂交缓冲液中的低盐，仅允许高度相似的核酸序列之间的杂交，而低严格性，例如较低温度和高盐，允许序列相似度较低时的杂交。DNA微阵列中使用的杂交探针是指与惰性表面如包被的载玻片或基因芯片共价连接且移动的cDNA靶标与之杂交的DNA。

[0190] 然后可通过热或化学方法使包含将与阵列上的探针杂交的靶核酸的混合物变性，并将其添加到微阵列中的口中。然后可以密封孔口，并且微阵列例如在杂交烘箱中杂交，其中通过旋转或在混合器中混合微阵列。杂交过夜后，可洗去非特异性结合(例如用SDS和SSC)。然后可以干燥微阵列，并在包含激发染料的激光器和测量染料发射的检测器的机器中进行扫描。可用模板栅格覆盖图像，并可对特征(例如，包含几个像素的特征)的强度进行定量。

[0191] 各种试剂盒可用于主题方法的核酸扩增和探针产生。可在本公开内容中使用的试剂盒的例子包括但不限于Nugen WT-Ovation™FFPE试剂盒，带有Nugen外显子模块和Frag/Label模块的cDNA扩增试剂盒。NuGEN WT-Ovation™FFPE System V2是全转录物组扩增系统，使得能够对来源于FFPE样品的降解的小RNA的大量存档物进行全面基因表达分析。该系统由扩增少至50ng总FFPE RNA所需的试剂和方案组成。该方案可用于qPCR、样品存档、片段化和标记。可以将扩增的cDNA在不到两小时内片段化并标记，以供使用NuGEN的FL-Ovation™cDNA生物素模块V2进行GeneChip™3’表达阵列分析。对于使用Affymetrix GeneChip™Exon和Gene ST阵列的分析，扩增的cDNA可以与WT-Ovation外显子模块一起使用，然后使用FL-Ovation™cDNA生物素模块V2进行片段化并标记。对于Agilent阵列上的分析，可使用NuGEN的FL-Ovation™cDNA荧光模块对扩增的cDNA进行片段化并标记。

[0192] 在一些实施方案中，可使用Ambion™WT-表达试剂盒。Ambion WT-表达试剂盒允许直接扩增总RNA，而无需单独的核糖体RNA(rRNA)消耗步骤。采用Ambion™WT-表达试剂盒，可

在Affymetrix™GeneChip™人、小鼠和大鼠外显子和基因1.0ST阵列上分析少至50ng总RNA的样品。除较低的输入RNA要求以及Affymetrix™方法与Taqman™实时PCR数据之间的高度一致性之外，Ambion™WT表达试剂盒还提供灵敏度的明显提高。例如，由于信噪比增加，用Ambion™WT表达试剂盒可以在外显子水平上获得检测到的高于背景的更大量的探针集。Ambion™WT-表达试剂盒可以与其他Affymetrix™标记试剂盒组合使用。在一些实施方案中，AmpTec™Trinucleotide Nano mRNA扩增试剂盒(6299-A15)可以在主题方法中使用。ExpressArt™TRinucleotide™mRNA扩增Nano试剂盒适用于大范围的、从1ng到700ng的输入总RNA。根据输入总RNA的量和所需的aRNA的产量，其可以用于1轮(输入量>300ng总RNA)或2轮(最小输入量为1ng总RNA)，其中RNA产量在>10μg的范围内。AmpTec的专有TRinucleotide™引发技术导致与针对rRNA的选择相组合的mRNA的优先扩增(与通用的真核3'-聚(A)-序列无关)。该试剂盒可以与cDNA转化试剂盒和Affymetrix™标记试剂盒结合使用。

[0193] 然后可以例如通过减去背景强度，随后再除使得各通道上的特征总强度相等的强度或参考基因的强度，对原始数据进行归一化，之后可以计算所有强度的t值。更复杂的方法包括z比、局部加权最小二乘(loess)和局部加权(lowess)回归以及例如用于Affymetrix芯片的RMA(强化多芯片分析)。

[0194] 在一些实施方案中，上述方法可用于测定用于训练(例如，使用分类器训练模块)分类器以区分受试者是患有UIP还是非UIP的转录物表达水平。在一些实施方案中，上述方法可用于确定用于输入到能够区分样品是UIP还是非UIP的分类器模块中的转录物表达水平。

数据分析

(i)样品与正常的比较

[0195] 在一些实施方案中，对来自受试者的样品(“测试样品”)进行的分子谱分析的结果可以与已知或怀疑为正常的生物样品(“正常样品”)进行比较。在一些实施方案中，正常样品是不包含或期望不包含ILD或所评价的病况的样品，或者在分子谱分析中对于一种或多种所评价的ILD测试呈阴性的样品。在一些实施方案中，正常样品是没有或期望没有任何ILD的样品，或在分子谱分析中对于任何ILD均可测试呈阴性的样品。正常样品可以来自与正测试的受试者不同的受试者，或来自同一受试者。在一些情况下，正常样品例如是从受试者如正测试的受试者获得的肺组织样品。正常样品可以与测试样品同时测定或在不同的时间测定。在一些实施方案中，正常样品是已知或疑似来自非吸烟者的样品。在特定实施方案中，正常样品是已经由至少两名病理学专家确认为非UIP样品的样品。在特定实施方案中，正常样品是已经由至少两名病理学专家确认为非IPF样品的样品。

[0196] 测试样品的测定结果可与具有已知疾病状态(例如，正常、受选定ILD(例如，IPF、NSIP等)影响、吸烟者、非吸烟者、非UIP、UIP)的样品的相同测定的结果进行比较。在一些情况下，正常样品的测定结果来自于数据库或参考文献。在一些情况下，正常样品的测定结果是本领域技术人员普遍接受的值或值的范围。在一些情况下，这种比较是定性的。在另一些情况下，这种比较是定量的。在一些情况下，定性或定量比较可以涉及但不限于以下一种或多种：比较荧光值、斑点强度、吸光度值、化学发光信号、柱状图、临界阈值、统计显著性值、基因产物表达水平、基因产物表达水平变化、替代外显子使用(alternative exon usage)、

替代外显子使用的变化、蛋白质水平、DNA多态性、拷贝数变异、一种或多种DNA标志物或区域的存在或不存在的指示，或者核酸序列。

(ii)结果评价

[0197] 在一些实施方案中，使用用于将基因产物表达水平或替代外显子使用与特定表型相关联的各种方法评价分子谱分析结果，所述表型例如是特定ILD或正常(例如无疾病或病况)。在一些情况下，可以确定规定的统计学置信水平以提供诊断置信水平。例如，可以确定大于90％的置信水平是存在ILD或者吸烟者或非吸烟者状态的有用预测器。在其他实施方案中，可以选择更严格或更不严格的置信水平。例如，可以选择大约或至少约50％、60％、70％、75％、80％、85％、90％、95％、97.5％、99％、99.5％或99.9％的置信水平作为有用的表型预测器。在一些情况下，所提供的置信水平可与样品质量、数据质量、分析质量、所用的具体方法和/或所分析的基因表达产物的数目有关。用于提供诊断的规定置信水平可基于假阳性或假阴性的期望数目和/或成本来选择。为了达到规定的置信水平而选择参数或鉴定具有诊断能力的标志物的方法包括但不限于受试者工作特征(ROC)曲线分析、双正态ROC、主成分分析、部分最小二乘法分析、奇异值分解、最小绝对收缩和选择算子分析、最小角回归和阈值梯度定向正则化方法。

(iii)数据分析

[0198] 在一些情况下，可通过应用为归一化和/或提高数据可靠性而设计的方法和/或过程来改进原始基因表达水平和可变剪接数据。在本公开内容的一些实施方案中，由于需处理大量个别的数据点，数据分析需要计算机或其他装置、机器或设备以应用本文所述的多种方法和/或过程。“机器学习分类器”是指用于表征基因表达谱的基于计算的预测数据结构或方法。例如通过外显子组富集的RNA测序或基于微阵列的杂交分析获得的对应于某些表达水平的信号通常运行所述分类器，从而对表达谱进行分类。监督的学习通常包括“训练”分类器以识别各类别之间的区别，然后“测试”分类器对独立测试集的准确性。对于新的未知样品，分类器可用于预测样品所属的类别。在各个实施方案中，这样的训练使用例如分类器训练模块来实现。

[0199] 在一些情况下，强化多阵列平均(RMA)法可用于对原始数据进行归一化。RMA法开始于计算多个微阵列上各匹配细胞的背景校正强度。背景校正的值被限制为正值，如Irizarry等人,Biostatistics 2003年4月4(2):249-64所述。背景校正后，获得各背景校正的匹配细胞强度的以2为底的对数。然后使用分位数归一化方法将各微阵列上的背景校正的、对数转化的匹配强度归一化，在该方法中，对于每个输入阵列和每个探针表达值，用所有阵列百分点的平均值替换阵列百分位探针值，该方法由Bolstad等人,Bioinformatics 2003更充分地描述。分位数归一化后，归一化的数据可以拟合线性模型以获得每个微阵列上的每个探针的表达量度。然后可利用Tukey中位数平滑算法(Tukey,J.W.,Exploratory Data Analysis.1977)确定归一化的探针集数据的对数级表达水平。

[0200] 可实现各种其他软件和/或硬件模块或过程。在某些方法中，特征选择和模型估计可使用glmnet(Friedman J,Hastie T,Tibshirani R.Regularization Paths for Generalized Linear Models via Coordinate Descent.Journal of statistical software 2010;33(1):1-22)通过具有套索罚分(lasso penalty)的逻辑回归来进行。可使用TopHat(Trapnell C,Pachter L,Salzberg SL.TopHat:discovering splice junctions

with RNA-Seq.Bioinformatics 2009;25(9):1105-11)比对原始读取。可使用HTSeq(Anders S,Pyl PT,Huber W.HTSeq-a Python framework to work with high-throughput sequencing data.Bioinformatics 2014)获得基因计数,并使用DESeq(Love MI,Huber W,Anders S.Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2;2014)进行归一化。在方法中,使用排名靠前的特征(N范围从10到200)利用e1071文库(Meyer D.Support vector machines:the interface to libsvm in package e1071.2014)来训练线性支持向量机(SVM)(Suykens JAK,Vandewalle J.Least Squares Support Vector Machine Classifiers.Neural Processing Letters 1999;9(3):293-300)。可使用pROC包计算置信区间(Robin X,Turck N,Hainard A等人pROC:an open-source package for R and S+to analyze and compare ROC curves.BMC bioinformatics 2011;12:77)。

[0201] 另外,可以进一步过滤数据以去除可能认为是可疑的数据。在一些实施方案中,得自具有少于约4、5、6、7或8个鸟苷和胞嘧啶核苷酸的微阵列探针的数据由于其异常杂交倾向或二级结构问题而可能被认为是不可靠的。类似地,得自具有超过约12、13、14、15、16、17、18、19、20、21或22个鸟苷和胞嘧啶核苷酸的微阵列探针的数据由于其异常杂交倾向或二级结构问题而可能被认为是不可靠的。

[0202] 在一些情况下,可以通过相对于一系列参考数据集对探针集可靠性进行排序而选择不可靠的探针集以从数据分析中排除。例如,RefSeq或Ensembl(EMBL)被认为是质量非常高的参考数据集。在一些情况下,来自与RefSeq或Ensembl序列匹配的探针集的数据由于其预期的高可靠性而可以特别地包括在微阵列分析实验中。类似地,来自匹配可靠性较低的参考数据集的探针集的数据可从进一步的分析中排除,或视情况而定包括在进一步的分析中。在一些情况下,可单独地或共同地使用Ensembl高通量cDNA(HTC)和/或mRNA参考数据集来确定探针集可靠性。在其他情况下,可以对探针集的可靠性进行排序。例如,可将与所有参考数据集如RefSeq、HTC、HTSeq和mRNA完全匹配的探针和/或探针集排序为最可靠的(1)。此外,可将与三分之二参考数据集匹配的探针和/或探针集排序为次最可靠的(2),可将与三分之一参考数据集匹配的探针和/或探针集排序为下一级(3),并且可将不与参考数据集匹配的探针和/或探针集排序为最后(4)。然后可以根据其排序从分析中包括或排除探针和/或探针集。例如,可以选择包括来自1、2、3和4类探针集,1、2和3类探针集,1和2类探针集,或1类探针集的数据用于进一步分析。在另一个实例中,可根据与参考数据集项错配的碱基对数目对探针集进行排序。应当理解,存在许多本领域已知的、用于评估给定探针和/或探针集在分子谱分析中的可靠性的方法,并且本公开内容的方法包括这些方法中的任一种及其组合。

[0203] 在本公开内容的一些实施方案中,如果来自探针集的数据不表达或以检测不到的水平(不高于背景)表达,则可从分析中将其排除。如果任何组满足以下情况,则探针集被判断为高于背景表达:

标准正态分布的T0到无穷大的积分<显著性(0.01)

其中:T0=Sqr(GroupSize)(T-P)/Sqr(Pvar);GroupSize=组中的CEL文件数;T=探针集中探针评分的平均值;P=GC含量的背景探针平均值的平均值,并且Pvar=背景探针变异之和/(探针集中的探针数)2。

[0204] 这允许这样的探针集:其中组中探针集的平均值高于作为探针集背景中心的与该探针集具有类似GC含量的背景探针的平均表达,并且使得能够从背景探针集变异导出探针集离差。

[0205] 在本公开内容的一些实施方案中,不显示变异或显示低变异的探针集可从进一步的分析中排除。低变异探针集经由卡方(Chi-Square)检验从分析中排除。如果探针集的转化变异在具有(N-1)自由度的卡方分布的99%置信区间的左侧,则认为它是低变异的。(N-1)*探针集变异/(基因探针集变异)约为卡方(N-1),其中N是输入CEL文件数,(N-1)是卡方分布的自由度,"基因探针集变异"是基因间的探针集变异的平均值。在本公开内容的一些实施方案中,如果给定基因或转录物簇的探针集包含少于最小数目的通过了之前描述的针对GC含量、可靠性、变异等的过滤器步骤的探针,则它们可从进一步的分析中排除。例如在一些实施方案中,如果给定基因或转录物簇的探针集包含少于约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15个或少于约20个探针,则它们可从进一步的分析中排除。

[0206] 基因表达水平或可变剪接的数据分析方法还可以包括使用本文提供的特征选择方法和/或过程。在本公开内容的一些实施方案中,通过利用LIMMA软件包(Smyth,G.K.(2005).Limma:linear models for microarray data.In:Bioinformatics and Computational Biology Solutions using R and Bioconductor,R.Gentleman,V.Carey,S.Dudoit,R.Irizarry,W.Huber(eds.),Springer,New York,397-420页)提供特征选择。

[0207] 基因表达水平和/或可变剪接的数据分析方法还可以包括使用预分类器方法和/或过程(例如,通过预分类器分析模块实现)。例如,方法和/或过程可利用细胞特异性分子指纹根据其组成对样品进行预分类,然后应用校正/归一化因子。然后可将该数据/信息输入最终分类方法和/或过程中,该方法和/或过程可整合该信息以帮助最终诊断。

[0208] 在某些实施方案中,本公开内容的方法包括使用预分类器方法和/或过程(例如,通过预分类器分析模块实现),其在应用本公开内容的UIP/非UIP分类器之前使用分子指纹将样品预分类为吸烟者或非吸烟者。

[0209] 基因表达水平和或可变剪接的数据分析方法还可以包括使用本文提供的分类器方法和/或过程(例如,通过预分类器分析模块实现)。在本公开内容的一些实施方案中,提供对角线线性判别分析、k-最近邻分类器、支持向量机(SVM)分类器、线性支持向量机、随机森林分类器或基于概率模型的方法或其组合以用于微阵列数据的分类。在一些实施方案中,基于目标类别之间表达水平差异的统计显著性来选择能区分样品(例如,UIP与非UIP、第一ILD与第二ILD、正常与ILD)或区分亚型(例如,IPF与NSIP)的鉴定标志物。在一些情况下,通过将Benjamini Hochberg程序或另一种校正应用于错误发现率(FDR)来调整统计显著性。

[0210] 在一些情况下,分类器可以补充有荟萃分析法,例如由Fishel和Kaufman等人,2007Bioinformatics 23(13):1599-606描述的方法。在一些情况下,分类器可以补充有荟萃分析法,例如再现性分析。在一些情况下,所述再现性分析选择出现在至少一种预测性表达产物标志物集中的标志物。

[0211] 用于导出后验概率并将后验概率应用于微阵列数据分析的方法的实例提供于Smyth,G.K.2004Stat.Appl.Genet.Mol.Biol.3:Article 3中,其通过引用整体并入本文。在一些情况下,后验概率可用于对由分类器提供的标志物进行排序。在一些情况下,可以根

据其后验概率对标志物进行排序，并且可以选择通过了所选阈值的那些标志物作为其差异表达指示或诊断例如UIP或非UIP的样品的标志物。示例性的阈值包括0.7、0.75、0.8、0.85、0.9、0.925、0.95、0.975、0.98、0.985、0.99、0.995或更高的先验概率。

[0212] 分子谱分析结果的统计学评价可以提供但不限于提供指示以下一种或多种可能性的一个或多个定量值：诊断准确性的可能性；样品是UIP的可能性；样品是非UIP的可能性；ILD的可能性；特定ILD的可能性；特定治疗性干预成功的可能性；受试者是吸烟者的可能性；以及受试者是非吸烟者的可能性。因此，可能没有经过遗传学或分子生物学培训的医师不需要了解原始数据。相反，所述数据以指导患者医护的最有用的形式直接提供给医师。分子谱分析的结果可使用许多方法进行统计学评价，包括但不限于：students T检验、双侧T检验、皮尔森秩和分析、隐马尔可夫模型分析、q-q图分析、主成分分析、单向ANOVA、双向ANOVA、LIMMA等。

[0213] 在本公开内容的一些实施方案中，单独使用分子谱分析或者与细胞学分析结合使用分子谱分析可以提供约85％的准确性到约99％或约100％的准确性的分类、鉴定或诊断。在一些情况下，分子谱分析方法和/或细胞学分析提供准确性为大约或至少约85％、86％、87％、88％、90％、91％、92％、93％、94％、95％、96％、97％、97.5％、98％、98.5％、99％、99.5％、99.75％、99.8％、99.85％或99.9％的ILD的分类、鉴定、诊断。在一些实施方案中，分子谱分析方法和/或细胞学分析提供准确性为大约或至少约85％、86％、87％、88％、90％、91％、92％、93％、94％、95％、96％、97％、97.5％、98％、98.5％、99％、99.5％、99.75％、99.8％、99.85％或99.9％的特定ILD类型(例如，IPF；NSIP；HP)的存在的分类、鉴定或诊断。

[0214] 在一些情况下，可通过随着时间的推移追踪受试者来确定初始诊断的准确性，从而确定其准确性。在其他情况下，可通过确定性的方式或者使用统计学方法确定准确性。例如，可利用受试者工作特征(ROC)分析确定最优分析参数，从而实现特定水平的准确性、特异性、阳性预测值、阴性预测值和/或错误发现率。

[0215] 在本公开内容的一些实施方案中，可以选择如下所述的基因表达产物和编码此类产物的核苷酸组合物用作本公开内容的分子谱分析试剂，所述基因表达产物和编码此类产物的核苷酸组合物经测定在UIP与非UIP、UIP与正常之间和/或吸烟者与非吸烟者之间表现出表达水平的最大差异或可变剪接的最大差异。这样的基因表达产物由于提供比其他方法更宽的动态范围、更大的信噪比、改善的诊断能力、更低的假阳性或假阴性可能性或更高的统计学置信水平而可能特别地有用。

[0216] 在本公开内容的其他实施方案中，与使用本领域所用的标准细胞学技术相比，单独使用分子谱分析或者与细胞学分析结合使用分子谱分析可以使评定为非诊断性的样品的数目减少约或至少约100％、99％、95％、90％、80％、75％、70％、65％或约60％。在一些情况下，与本领域使用的标准细胞学方法相比，本公开内容的方法可以使评定为不确定或疑似的样品的数目减少约或至少约100％、99％、98％、97％、95％、90％、85％、80％、75％、70％、65％或约60％。

[0217] 在一些情况下，将分子谱分析的结果输入数据库中以供分子谱分析企业、个体、医疗提供者或保险提供者的代表或代理人访问。在一些情况下，测定结果包括企业的代表、代理人或顾问如医学专业人员的样品分类、鉴定或诊断。在其他情况下，自动提供数据的计算

机分析。在一些情况下，分子谱分析企业可以就以下一项或多项服务向个体、保险提供者、医疗提供者、研究人员或政府机构收费：所进行的分子谱分析、咨询服务、数据分析、结果报告或数据库访问。

[0218] 在本公开内容的一些实施方案中，分子谱分析结果作为计算机屏幕上的报告或作为纸件报告提供。在一些情况下，所述报告可以包括但不限于以下一种或多种信息：差异表达的基因数、原始样品的适合性、显示差异可变剪接的基因数、诊断、诊断的统计学置信度、受试者是吸烟者的可能性、ILD的可能性和指定的疗法。

(iv)基于分子谱分析结果的样品分类

[0219] 分子谱分析的结果可被分类为例如以下之一：吸烟者、非吸烟者、ILD、特定类型的ILD、非ILD或非诊断性的(提供关于ILD是否存在的不充分信息)。在一些情况下，分子谱分析的结果可分为IPF与NSIP类别。在特定情况下，结果分类为UIP或非UIP。

[0220] 在本公开内容的一些实施方案中，使用经训练的分类器对结果进行分类。经训练的分类器可以是经训练的算法。本公开内容的经训练的分类器实现已使用已知UIP和非UIP样品的参考组开发的方法和/或过程。在一些实施方案中，训练(例如，使用分类器训练模块)包括将来自UIP样品的第一组生物标志物中的基因表达产物水平与来自非UIP样品的第二组生物标志物中的基因表达产物水平进行比较，其中第一组生物标志物包含至少一种不在第二组中的生物标志物。在一些实施方案中，训练(例如，使用分类器训练模块)包括将来自是非UIP的第一ILD的第一组生物标志物中的基因表达产物水平与来自是UIP的第二ILD的第二组生物标志物中的基因表达产物水平进行比较，其中第一组生物标志物包含至少一种不在第二组中的生物标志物。在一些实施方案中，训练(例如，使用分类器训练模块)进一步包括将来自是吸烟者的第一受试者的第一组生物标志物中的基因表达产物水平与来自是非吸烟者的第二受试者的第二组生物标志物中的基因表达产物水平进行比较，其中第一组生物标志物包含至少一种不在第二组中的生物标志物。在一些实施方案中，可使用分类组内的生物标志物组与分类器中所使用的所有其他生物标志物组(或所有其他生物标志物特征)的表达水平的比较来训练(例如，使用分类器训练模块)整个分类器或分类器的部分。在一些实施方案中，可使用在包含从单个受试者获得的至少2、3、4、5个或更多个单独的样品的汇集样品中测量的表达水平的比较来训练(例如，使用分类器训练模块)整个分类器或分类器的部分。在一些实施方案中，可使用计算机汇集的表达水平的比较来训练(例如，使用分类器训练模块)整个分类器或分类器的部分，如本文所述，其中计算机汇集的表达水平包括来自从单个受试者获得的至少2、3、4、5个或更多个单独的样品的汇集表达水平。在一些实施方案中，如本段所述的经训练的分类器比较测试样品与参考样品或参考样品组之间的SEQ ID NO:1-151中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150或151个或其任何组合，以确定测试样品是UIP还是非UIP。在特定方面，这样的分类器比较额外的基因，例如1、

2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0221] 在一些实施方案中，如本文所述的训练的分类器比较测试样品与参考样品或参考样品组之间的表5中所列基因中的1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100、101、102、103、104、105、106、107、108、109、110、111、112、113、114、115、116、117、118、119、120、121、122、123、124、125、126、127、128、129、130、131、132、133、134、135、136、137、138、139、140、141、142、143、144、145、146、147、148、149、150、151、152、153、154、155、156、157、158、159、160、161、162、163、164、165、166、167、168、169、170、171、172、173、174、175、176、177、178、179、180、181、182、183、184、185、186、187、188、189或190个的基因表达水平，以确定测试样品是UIP还是非UIP。在特定方面，这样的分类器比较额外的基因，例如1、2、3、4、5、6、7、8个或更多个额外的基因。在其他方面，该分类器略去某些前述基因，例如这些基因中的1、2、3、4、5、6、7、8个或更多个，而在一些情况下包括其他基因。

[0222] 适用于样品分类的分类器包括但不限于k最近邻分类器、支持向量机、线性判别分析、对角线性判别分析、updown、朴素贝叶斯分类器、神经网络分类器、隐马尔可夫模型分类器、遗传分类器或其任何组合。

[0223] 在一些情况下，本公开内容的经训练的分类器可以整合除了基因表达数据或可变剪接数据以外的数据，例如但不限于DNA多态性数据、测序数据、本公开内容的细胞学家或病理学家的评分或诊断、由本公开内容的预分类器方法和/或过程提供的信息或本公开内容受试者的医疗史信息。

[0224] 当为了ILD(例如，用UIP)诊断而对生物样品进行分类时，二元分类器通常出现两种可能的结果。类似地，当为了吸烟者诊断而对生物样品进行分类时，二元分类器通常出现两种可能的结果。当二元分类器与实际真值(例如，来自生物样品的值)相比时，通常存在四种可能的结果。如果预测结果为p(其中“p”是阳性分类器输出，诸如特定ILD)，并且实际值也为p，那么它被称为真阳性(TP)；然而如果实际值为n，那么它被称为假阳性(FP)。相反，当预测结果和实际值二者都为n时(其中“n”是阴性分类器输出，诸如非ILD，或不存在本文所述的特定病变组织)出现真阴性(例如，TN)，而当预测结果为n而实际值为p时则出现假阴性(FN)。在一个实施方案中，考虑试图确定人是否患有某种疾病的诊断性测试。当该人测试为阳性，但实际上未患该疾病的情况时，出现假阳性(FP)。另一方面，当该人测试为阴性(提示他们是健康的)，但他们实际上确实患有该疾病时，出现FN。在一些实施方案中，可通过以相应比例对可用样品上获得的误差进行重采样而产生假设亚型的真实世界流行度的受试者工作特征(ROC)曲线。

[0225] 疾病的阳性预测值(PPV)，或准确率，或验后概率，是正确诊断的具有阳性测试结果的患者的比例。它是诊断方法的最重要的判断标准，因为它反映了阳性测试反映出所测试的基础病状的可能性。然而，它的值取决于疾病的患病率，这可能有所不同。假阳性率(α)＝FP/(FP+TN)-特异性；假阴性率(β)＝FN/(TP+FN)-灵敏度；能力＝灵敏度＝1-β；似然比阳性＝灵敏度/(1-特异性)；似然比阴性＝(1-灵敏度)/特异性。

[0226] 阴性预测值是正确诊断的具有阴性测试结果的患者的比例。PPV和NPV量度可使用适当的疾病亚型流行度估计值获得。合并的疾病流行度的估计值可由通过外科手术大致分类为B与M的不确定结果的合并库来计算。在一些实施方案中，对于亚型特异性估计值，疾病流行度有时可能是无法计算的，因为没有任何可用的样品。在这些情况下，亚型疾病流行度可以用合并的疾病流行度估计值来代替。

[0227] 在一些实施方案中，表达产物或替代外显子使用的水平指示以下一种或多种：IPF、NSIP、HP、UIP、非UIP。

[0228] 在一些实施方案中，表达产物或替代外显子使用的水平指示受试者是吸烟者或非吸烟者。

[0229] 在一些实施方案中，主题方法的表达分析结果提供了给定诊断为正确的统计置信水平。在一些实施方案中，此统计置信水平为至少约或大于约85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％、99.5％或更高。

报告

[0230] 主题方法和/或系统可包括生成报告，该报告提供样品(肺组织样品)是UIP样品的指示(例如，使用报告模块)。主题方法和/或系统可包括生成报告，该报告提供样品(肺组织样品)是非UIP样品的指示(例如，使用报告模块)。主题方法和/或系统可包括生成报告，该报告提供样品(肺组织样品)是ILD样品的指示(例如，使用报告模块)。主题诊断方法可包括生成报告，该报告提供关于被测试的个体是否患有ILD的指示。主题诊断方法可包括生成报告，该报告提供关于被测试的个体是否是吸烟者的指示。主题方法(或报告模块)可包括生成报告，该报告提供关于被测试的个体是否具有IPF的指示(例如，是ILD而不是IPF；例如，该报告可指示该个体患有IPF而不是NSIP)。

[0231] 在一些实施方案中，诊断UIP与非UIP的主题方法涉及生成报告(例如，使用报告模块)。这样的报告可包含诸如以下信息：患者患有UIP的可能性；患者患有非UIP的可能性；患者患有IPF的可能性；患者是吸烟者的可能性；关于进一步评价的建议；关于治疗药物和/或设备干预的建议；等等。

[0232] 例如，本文公开的方法可进一步包括生成或输出提供主题诊断方法的结果的报告的步骤，该报告可以以电子介质(例如，计算机监视器上的电子显示器)的形式提供，或以有形介质(例如，印在纸上或其他有形介质上的报告)的形式提供。关于主题诊断方法的结果的评估(例如，患者患有UIP的可能性；患者患有非UIP的可能性；患者患有IPF的可能性；个体患有ILD的可能性；个体患有IPF的可能性；个体是吸烟者的可能性)可被称为“报告”或简称为“评分”。制备报告的人或实体(“报告生成者”)也可执行诸如样品收集、样品处理等步骤。或者，除报告生成者之外的实体可执行诸如样品收集、样品处理等步骤。可向用户提供诊断评估报告。“用户”可以是健康专业人员(例如，临床医生、实验室技术人员、医师(例如，心脏病学家)等)。

[0233] 主题报告可进一步包含以下中的一种或多种：1)服务提供者信息；2)患者数据；3)关于给定基因产物或基因产物组的表达水平、评分或分类器决定的数据；4)后续评价建议；5)治疗干预或建议；以及6)其他特征。

进一步评价

[0234] 基于给定基因产物或基因产物组的表达水平和/或基于报告(如上所述)，医师或

其他合格医疗人员可确定是否需要对测试受试者(患者)进行进一步评价。进一步评价可包括例如肺量测定法。

治疗干预

[0235] 基于给定基因产物或基因产物组的表达水平和/或基于报告(如上所述),医师或其他合格医疗人员可确定是否建议进行适当的治疗干预。治疗干预包括基于药物的治疗干预、基于装置的治疗干预和手术干预。当报告表明个体患有UIP和/或IPF的可能性时,基于药物的治疗干预包括例如向个体施用有效量的吡非尼酮、泼尼松、硫唑嘌呤和/或N-乙酰半胱氨酸。手术干预包括例如动脉旁路手术。

计算机实现的方法、系统和装置

[0236] 本公开内容的方法可以是计算机实现的,使得方法步骤(例如,测定、比较、计算等)全部或部分自动化。

[0237] 因此,本公开内容提供了与促进间质性肺病的诊断(例如,UIP、非UIP、IPF、NSIP、HP等的诊断)的计算机实现的方法相关的方法、计算机系统、装置等,该诊断包括区分诊断。

[0238] 本公开内容进一步提供了与促进确定吸烟者状态(例如,吸烟者与非吸烟者)的计算机实现的方法相关的方法、计算机系统、装置等。

[0239] 本公开内容进一步提供了与促进间质性肺病的诊断(例如,UIP、非UIP、IPF、NSIP、HP等的诊断)的计算机实现的方法相关的方法、计算机系统、装置等,该诊断包括区分诊断,其中该方法进一步包括确定受试者的吸烟者状态(吸烟者与非吸烟者)并将吸烟者状态并入受试者间质性肺病诊断的确定中。在一些实施方案中,(i)将吸烟者状态作为训练(例如,使用分类器训练模块)期间使用的模型中的协变量并入间质性肺病诊断中。该方法提高了信噪比,特别是在吸烟者的数据中(其中噪声较高),并允许来自吸烟者和非吸烟者的数据进行结合并同时使用。在一些实施方案中,(ii)通过在间质性肺病诊断分类器训练期间鉴定对吸烟者状态偏差敏感的一个或多个基因并排除此类基因或者对此类基因与对吸烟者状态不敏感的基因不同地进行加权,将吸烟者状态并入间质性肺病诊断中。在一些实施方案中,(iii)通过构建分层分类将吸烟者状态并入间质性肺病诊断中,其中对初始分类器进行训练以识别区分吸烟者与非吸烟者的基因特征(例如,使用分类器训练模块)。一旦患者样品被预分类为“吸烟者”或“非吸烟者”(例如,使用预分类器分析模块),可运行被各自训练以分别区分吸烟者或非吸烟者中的UIP与非UIP的不同分类器以诊断间质性肺病。在更进一步的实施方案中,包括将吸烟者状态并入受试者间质性肺病诊断的确定中的步骤的此类方法包括上述此类并入方法中的一种或多种的组合(即,本段中实施方案(i)至(iii)中的两种或更多种的组合)。

[0240] 例如,包括获得生物标志物水平的值、将归一化的生物标志物(基因)表达水平与对照水平进行比较、计算UIP或非UIP的可能性(并且在一些情况下,计算受试者是吸烟者的可能性)、生成报告等在内的方法步骤可由计算机程序产品完全或部分地执行。获得的值可以以电子方式存储在例如数据库中,并且可经受由编程计算机执行的分类器(例如,使用分类器分析模块)。

[0241] 例如,本公开内容的方法和/或系统可涉及将生物标志物水平(例如,归一化的基因产物表达水平)输入到分类器分析模块中以执行方法和/或过程以进行本文所述的比较和计算步骤,并且通过例如在计算机本地或远程的位置处向输出设备显示或打印报告来生

成如本文所述的报告(例如,使用报告模块)。报告的输出可以是代表数值或数值范围的评分(例如,数值评分(代表数值)或非数值评分(例如,非数值输出(例如,“IPF”、“无IPF证据”)))。在其他方面,输出可指示“UIP”与“非UIP”。在其他方面,输出可指示“吸烟者”与“非吸烟者”。

[0242] 因此,本公开内容提供了计算机程序产品,其包括其上存储有软件和/或硬件模块的计算机可读存储介质。当由处理器执行时,软件和/或硬件模块可基于从来自个体的一个或多个生物样品(例如,肺组织样品)的分析获得的值来执行相关计算。计算机程序产品在其中存储有用于执行计算的计算机程序。

[0243] 本公开内容提供了用于执行上述程序的系统,该系统通常包括:a)执行软件和/或硬件模块的中央计算环境或处理器;b)可操作地连接到计算环境以接收患者数据的输入设备,其中该患者数据可包括例如从使用来自患者的生物样品的测定所获得的生物标志物水平或其他值,如上所述;c)连接到计算环境以向用户(例如,医疗人员)提供信息的输出设备;以及d)由中央计算环境(例如,处理器)执行的方法和/或过程,其中该方法和/或过程基于由输入设备接收的数据来执行,并且其中该方法和/或过程计算值,其中该值指示受试者具有UIP、非UIP、ILD或IPF的可能性,如本文所述。

[0244] 本公开内容还提供了用于执行上述程序的系统,该系统通常包括:a)执行软件和/或硬件模块的中央计算环境或处理器;b)可操作地连接到计算环境以接收患者数据的输入设备,其中该患者数据可包括例如从使用来自患者的生物样品的测定所获得的生物标志物水平或其他值,如上所述;c)连接到计算环境以向用户(例如,医疗人员)提供信息的输出设备;以及d)由中央计算环境(例如,处理器)执行的方法和/或过程,其中该方法和/或过程基于由输入设备接收的数据来执行,其中该方法和/或过程计算值,其中该值指示受试者具有UIP、非UIP、ILD或IPF的可能性,如本文所述,并且其中该方法和/或过程使用吸烟状态(吸烟者与非吸烟者)作为训练期间使用的模型中的协变量。在一些实施方案中,该方法和/或过程在分类器训练期间排除对吸烟者状态偏差敏感的一个或多个基因或对其不同地进行加权,以用不受吸烟状态混淆或影响的基因来丰富用于训练的特征空间。

[0245] 在更进一步的实施方案中,本公开内容提供了用于执行上述程序的系统,该系统通常包括:a)执行软件和/或硬件模块的中央计算环境或处理器;b)可操作地连接到计算环境以接收患者数据的输入设备,其中患者数据可包括例如从使用来自患者的生物样品的测定获得的生物标志物水平或其他值,如上所述;c)连接到计算环境以向用户(例如,医疗人员)提供信息的输出设备;以及d)由中央计算环境(例如,处理器)执行的第一方法和/或过程,其中第一方法和/或过程基于由输入设备接收的数据执行,其中第一方法和/或过程计算值,该值指示受试者是吸烟者或非吸烟者的可能性,如本文所述,其中受试者作为吸烟者或非吸烟者的状态导致第一方法和/或过程应用被特别训练(例如,使用分类器训练模块)以分别区分吸烟者或非吸烟者中的UIP与非UIP的第二方法和/或过程;以及e)其中第二方法和/或过程由中央计算环境(例如,处理器)执行,其中第二方法和/或过程基于由输入设备接收的数据执行,并且其中第二方法和/或过程计算值,该值指示受试者具有ILD的可能性,如本文所述。

计算机系统

[0246] 图7A图示了处理系统100,其包括经由总线或总线组110耦合在一起的至少一个处

理器102或者处理单元或多个处理器、存储器104、至少一个输入设备106和至少一个输出设备108。处理系统可在任何合适的设备上实现，例如，主机设备、个人计算机、手持式或膝上型设备、个人数字助理、多处理器系统、基于微处理器的系统、可编程的消费电子设备、小型计算机、服务器计算机、网络服务器计算机、主计算机和/或包括任何上述系统或设备的分布式计算环境。

[0247] 在某些实施方案中，输入设备106和输出设备108可以是相同的设备。还可提供接口112以供将处理系统100耦合到一个或多个外围设备，例如，接口112可以是PCI卡或PC卡。还可提供容纳至少一个数据库116的至少一个存储设备114。

[0248] 存储器104可以是任何形式的存储器设备，例如，易失性或非易失性存储器、固态存储设备、磁性设备等。例如，在一些实施方案中，存储器104可以是随机存取存储器(RAM)、存储缓冲器、硬盘驱动器、只读存储器(ROM)、可擦除可编程只读存储器(EPROM)、数据库和/或诸如此类。

[0249] 处理器102可包含多于一个不同的处理设备，例如以在处理系统100内处理不同功能。处理器100可以是被配置用于运行或执行一组指令或代码(例如，存储在存储器中)的任何合适的处理设备，诸如通用处理器(GPP)、中央处理单元(CPU)、加速处理单元(APU)、图形处理单元(GPU)、专用集成电路(ASIC)和/或诸如此类。这样的处理器100可运行或执行存储在存储器中的一组指令或代码，该指令或代码与使用个人计算机应用、移动应用、因特网浏览器、蜂窝和/或无线通信(例如，经由网络)和/或诸如此类相关联。更具体地，如本文所述，处理器可执行存储在存储器104中的一组指令或代码，该指令或代码与分析和分类数据相关联。

[0250] 输入设备106接收输入数据118并且可包括例如键盘、指针设备如笔类设备或鼠标、用于语音控制激活的音频接收设备如麦克风、数据接收器或天线如调制解调器或无线数据适配器、数据采集卡等。输入数据118可来自不同来源，例如键盘指令结合经由网络接收的数据。

[0251] 输出设备108产生或生成输出数据120，并且可包括例如显示设备或监视器(在这种情况下输出数据120是可视的)、打印机(在这种情况下输出数据120是打印的)、端口例如USB端口、外围组件适配器、数据发送器或天线如调制解调器或无线网络适配器等。输出数据120可以是不同的并且来自不同的输出设备，例如监视器上的视觉显示结合发送到网络的数据。用户可在例如监视器上或使用打印机来查看数据输出或数据输出的解释。

[0252] 在一些实施方案中，输入设备106和/或输出设备108可以是被配置用于经由网络发送和/或接收数据的通信接口。更具体地，在这样的实施方案中，处理系统100可充当一个或多个客户端设备(图7A中未示出)的主机设备。如此，处理系统100可将数据发送到客户端设备(例如，输出数据120)并从客户端设备接收数据(例如，输入数据118)。这样的通信接口可以是任何合适的模块和/或设备，其可使处理系统100与客户端设备如一个或多个网络接口卡等进行通信。这样的网络接口卡可包括例如可经由网络等使客户端设备150与主机设备110等进行通信的以太网端口、WiFi®无线电、蓝牙®无线电、近场通信(NFC)无线电和/或蜂窝无线电。

[0253] 存储设备114可以是任何形式的数据或信息存储系统或方法，例如，易失性或非易失性存储器、固态存储设备、磁性设备等。例如，在一些实施方案中，存储设备114可以是随

机存取存储器(RAM)、存储缓冲器、硬盘驱动器、只读存储器(ROM)、可擦除可编程只读存储器(EPROM)、数据库和/或诸如此类。

[0254] 在使用中，处理系统100适于允许数据或信息经由有线或无线通信系统或方法存储在至少一个数据库116中和/或从其中检索。接口112可允许处理单元102和可用于专用目的的外围组件之间的有线和/或无线通信。通常，处理器102可经由输入设备106接收指令作为输入数据118，并且可通过利用输出设备108向用户显示处理的结果或其他输出。可提供多于一个输入设备106和/或输出设备108。处理系统100可以是任何合适形式的终端、服务器、专用硬件等。处理系统100可以是网络化通信系统的一部分。

[0255] 处理系统100可连接到网络，例如，局域网(LAN)、虚拟网络如虚拟局域网(VLAN)、广域网(WAN)、城域网(MAN)、全球微波互联接入(WiMAX)、蜂窝网络、因特网以及/或者实现为有线和/或无线网络的任何其他合适的网络。例如，当在LAN网络环境中使用时，计算系统环境100通过网络接口或适配器连接到LAN。当在WAN网络环境中使用时，计算系统环境通常包括用于在WAN上建立通信的调制解调器或者其他系统或方法，诸如因特网。调制解调器可以是内部的或外部的，可经由用户输入接口或经由另外的适当机制连接到系统总线。在网络化环境中，相对于计算系统环境100描述的程序模块或其部分可存储在远程存储器存储设备中。应当理解，图7所示的网络连接是实例，并且可使用在多个计算机之间建立通信链路的其他系统和方法。

[0256] 输入数据118和输出数据120可经由网络传送至其他设备。可使用有线或无线的通信系统和方法来实现通过网络的信息和/或数据传输。服务器可促进网络与一个或多个数据库之间的数据传输。服务器与一个或多个数据库提供信息源的实例。

[0257] 因此，图7A中所示的处理计算系统环境100可通过使用与一个或多个远程计算机的逻辑连接而在网络化环境中运行。远程计算机可以是个人计算机、服务器、路由器、网络PC、对等设备或其他公共网络节点，并且通常包括上述元件中的许多或全部。

[0258] 图7B更详细地图示了图7A的处理器102。处理器102可被配置用于执行特定模块。模块可以是例如存储在存储器104中和/或在处理器102中执行的硬件模块、软件模块，和/或其任何组合。例如，如图7B所示，处理器102包括和/或执行预分类器分析模块130、分类器训练模块132、分类器分析模块134和报告模块136。如图7B所示，预分类器分析模块130、分类器训练模块132、分类器分析模块134和报告模块136可连接和/或电耦合。如此，可在预分类器分析模块130、分类器训练模块132、分类器分析模块134和报告模块136之间发送信号。

[0259] 分类器训练模块132可被配置用于接收数据语料库(例如，基因表达数据、测序数据)并训练分类器。例如，来自先前被鉴定为UIP和非UIP(例如，由专家鉴定)的样品的临床注释数据可由输入设备106接收并且由分类器训练模块132使用以鉴定先前被鉴定为UIP和非UIP的样品之间的相关性。例如，可获得专家TBB组织病理学标签(即，UIP或非UIP)、专家HRCT标签和/或专家患者水平临床结果标签，并单独地或组合地使用以使用微阵列和/或测序来训练分类器数据。使用的特征空间可包括基因表达、变体、突变、融合、杂合性丢失(LOH)、生物学途径效应和/或可作为用于训练机器学习算法的特征而提取的数据的任何其他维度。在一些实施方案中，用于训练UIP与非UIP分类器、吸烟者与非吸烟者分类器、或UIP与非UIP和吸烟者与非吸烟者分类器的训练的特征空间包括基因表达、变体、突变、融合、杂合性丢失(LOH)和生物学途径效应。在一些实施方案中，用于训练UIP与非UIP分类器、吸烟

者与非吸烟者分类器、或UIP与非UIP和吸烟者与非吸烟者分类器的训练的特征空间包括基因表达和变体维度。

[0260] 在一些实施方案中，分类器训练模块132可基于与所接收的样品是否与吸烟者或非吸烟者相关联的指示来训练吸烟者分类器和非吸烟者分类器。在其他实施方案中，吸烟者/非吸烟者可用作训练单个分类器的属性(模型协变量)。在训练该分类器后，其可用于鉴定和/或分类如本文所述的新接收的和未知的样品。

[0261] 预分类器分析模块130可鉴定样品是否与吸烟者或非吸烟者相关联。具体地，预分类器分析模块130可使用任何合适的方法来将样品鉴定和/或分类为来自吸烟(或具有过去的重度吸烟史)的个体与不吸烟(或没有吸烟史)的个体。分类可以以任何合适的方式进行，诸如，接收来自用户的指示、鉴定对吸烟者状态偏差敏感的基因、使用机器学习分类器和/或本文所述的任何其他合适的方法。

[0262] 分类器分析模块134可将样品输入到分类器中以将所接收的样品鉴定和/或分类为与UIP和非UIP相关联。具体地，分类器分析模块134可使用训练的分类器来鉴定样品指示UIP还是非UIP。在一些实施方案中，分类器分析模块134可指示与UIP或非UIP相关联的样品的百分比或置信度评分。在一些实施方案中，分类器分析模块134可执行两个单独的分类器：一个针对吸烟者样品，另一个针对非吸烟者样品(通过预分类器分析模块130确定)。在其他实施方案中，针对伴随有吸烟者状态输入的吸烟者和非吸烟者样品执行单个分类器。

[0263] 报告模块136可被配置用于基于分类器分析模块134的结果生成任何合适的报告，如本文进一步详细描述的。在一些情况下，该报告可包括但不限于诸如以下的一种或多种信息：差异表达的基因数、原始样品的适合性、显示差异可变剪接的基因数、诊断、诊断的统计学置信度、受试者是吸烟者的可能性、ILD的可能性和指定的疗法。

[0264] 图7C图示了本公开内容的一个非限制性实施方案的流程图，其中使用已知的UIP和非UIP样品的基因产物表达数据来训练(例如，使用分类器训练模块)分类器以区分UIP与非UIP，其中分类器在一些情况下将吸烟者状态视为协变量，并且其中来自未知样品的基因产物表达数据被输入训练的分类器中以将未知样品鉴定为UIP或非UIP，并且其中通过分类器分类的结果被确定并通过报告输出。

[0265] 可参考由一个或多个计算设备如图7A的计算系统环境100执行的动作和操作的符号表示来描述某些实施方案。如此，将会理解，这样的有时被称为计算机执行的动作和操作包括通过计算机的处理器对表示结构化形式的数据的电信号的操纵。该操纵变换数据或将它们保持在计算机的存储器系统中的位置处，其以本领域技术人员理解的方式重新配置或以其他方式改变计算机的操作。将数据保持在其中的数据结构是存储器的物理位置，其具有由数据格式定义的特定属性。然而，虽然在前述上下文中描述了实施方案，但是并非意在限制，因为本领域技术人员将理解，下文描述的动作和操作也可在硬件中实现。

[0266] 实施方案可用许多其他通用或专用的计算设备以及计算系统环境或配置来实现。可适用于实施方案的其他计算系统、环境和配置的示例包括但不限于个人计算机、手持式或膝上型设备、个人数字助理、多处理器系统、基于微处理器的系统、可编程的消费电子产品、网络、小型计算机、服务器计算机、网络服务器计算机、主计算机和包括任何上述系统或设备的分布式计算环境。

[0267] 可在计算机可执行指令如硬件和/或软件模块的一般上下文中描述实施方案。还

可在分布式计算环境中实践实施方案，其中任务由通过通信网络链接的远程处理设备执行。在分布式计算环境中，程序模块可位于包括存储器存储设备的本地和远程计算机存储介质中。

计算机程序产品

[0268] 本公开内容提供了计算机程序产品，该计算机程序产品当在诸如以上参考图7描述的可编程计算机上执行时可执行本公开内容的方法。如上所述，取决于期望的配置，本文所述的主题可体现在系统、装置、方法和/或物品中。这些各种实现可包括在可编程系统上可执行和/或可解释的一个或多个计算机程序中的实现，该可编程系统包括至少一个可编程处理器，其可以是专用的或通用的，被耦合以从存储系统、至少一个输入设备(例如，摄像机、麦克风、操纵杆、键盘和/或鼠标)以及至少一个输出设备(例如，显示监视器、打印机等)接收数据和指令，并向其发送数据和指令。

[0269] 计算机程序(也称为程序、软件、软件应用、应用、组件或代码)包括用于可编程处理器的指令，并且可用高级程序性和/或面向对象的编程语言以及/或者用汇编/机器语言来实现。如本文所用，术语“机器可读介质”是指用于向可编程处理器提供机器指令和/或数据的任何计算机程序产品、装置和/或设备(例如，磁盘、光盘、存储器等)，包括接收机器指令作为机器可读信号的机器可读介质。

[0270] 从该描述中将会显而易见的是，本公开内容的方面可至少部分地以软件、硬件、固件或其任何组合来体现。因此，本文所述的技术不限于硬件电路和/或软件的任何特定组合，或者不限于由计算机或其他数据处理系统执行的指令的任何特定来源。相反，这些技术可在计算机系统或其他数据处理系统中响应于一个或多个处理器如微处理器来执行，从而执行存储在存储器或其他计算机可读介质中的指令序列，该计算机可读介质包括任何类型的ROM、RAM、高速缓冲存储器、网络存储器、软磁盘、硬盘驱动器(HDD)、固态设备(SSD)、光盘、CD-ROM和磁光盘、EPROM、EEPROM、闪速存储器或适用于以电子格式存储指令的任何其他类型的介质。

[0271] 此外，处理器可以是或可以包括一个或多个可编程的通用或专用微处理器、数字信号处理器(DSP)、可编程控制器、专用集成电路(ASIC)、可编程逻辑器件(PLD)、可信平台模块(TPM)等，或此类器件的组合。在替代实施方案中，专用硬件如逻辑电路或其他硬连线电路可与软件指令组合使用以实现本文所述的技术。

阵列和试剂盒

[0272] 本公开内容提供了用于执行受试者评价方法或受试者诊断方法的阵列和试剂盒。

阵列

[0273] 主题阵列可包含多个核酸，每个核酸与在组织样品中存在的细胞中差异表达的基因杂交，该组织样品从测试UIP、非UIP、IPF或ILD的个体获得。

[0274] 主题阵列可包含多个核酸，每个核酸与在组织样品中存在的细胞中差异表达的基因杂交，该组织样品从测试吸烟者状态的个体获得。

[0275] 主题阵列可包含多个核酸，每个核酸与在组织样品中存在的细胞中差异表达的基因杂交，该组织样品从测试吸烟者状态以及UIP、非UIP、IPF或ILD的个体获得。

[0276] 主题阵列可包含多个成员核酸，其中每个成员核酸与不同的基因产物杂交。在一些情况下，两个或更多个成员核酸与相同的基因产物杂交；例如，在一些情况下，2、3、4、5、

6、7、8、9、10个或更多个成员核酸与相同的基因产物杂交。成员核酸可具有约5个核苷酸(nt)至约100nt的长度，例如，5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、20-25、25-30、30-40、40-50、50-60、60-70、70-80、80-90或90-100nt。核酸可具有一种或多种磷酸骨架修饰。

[0277] 主题阵列可包含约10至约10^5个独特成员核酸，或者多于10^5个独特成员核酸。例如，主题阵列可包含约10至约10^2个、约10^2至约10^3个、约10^3至约10^4个、约10^4至约10^5个或者多于10^5个独特成员核酸。试剂盒

[0278] 本公开内容的试剂盒可包含阵列，如上所述；以及用于分析基因产物的表达水平的试剂。

[0279] 用于分析核酸基因产物的表达水平的试剂包括，例如，适用于对核酸进行测序的试剂；适用于扩增核酸的试剂；以及适用于核酸杂交的试剂。

[0280] 该试剂盒可包含：缓冲液；可检测标签；用于产生可检测标签的组分(例如，其中核酸探针包含可检测标签)；等等。试剂盒的各种组分可存在于单独的容器中，或者某些相容的组分可根据需要预先在单个容器中组合。

[0281] 除了上述组分之外，主题试剂盒可包含用于使用试剂盒的组分来实践主题方法的说明。用于实践主题方法的说明通常记录在合适的记录介质上。例如，说明可印刷在基板上，如纸或塑料等。因此，说明可作为包装插入物存在于试剂盒中、存在于试剂盒的容器或其组件的标签中(即，与包装或分装相关联)等。在其他实施方案中，说明作为存在于合适的计算机可读存储介质上的电子存储数据文件而存在，该计算机可读存储介质例如光碟只读存储器(CD-ROM)、数字通用光盘(DVD)、软盘等。在其他实施方案中，实际说明不存在于试剂盒中，而是提供例如通过互联网从远程来源获得说明的方法。该实施方案的实例是包含网址的试剂盒，在该网址中可查看说明和/或可从其下载说明。与说明相同，用于获得说明的该方法被记录在合适的基板上。

缩写

adj.P.Value.edgeR:	使用 edgeR 分析的经错误发现率调整的 RNAseq 基因表达数据的 p 值
adj.P.Value.microarray	使用微阵列分析的经错误发现率调整的 RNAseq 基因表达数据的 p 值
adj.P.Value.npSeq:	使用 npSeq 分析的经错误发现率调整的 RNAseq 基因表达数据的 p 值
BRONCH:	细支气管炎
CIF-NOC	未另外分类的慢性间质纤维化
edgeR:	用于测序数据的显著性分析的 R 包

Ensembl ID:	来自 Ensembl Genome Browser 数据库的基因标识符
FDR:	错误发现率，限制由于同时评价的大量基因而导致结果随机的可能性的经调整的 p 值
Gene Symbol:	来自 HUGO Gene Nomenclature Committee 的基因标识符
logFC.edgeR:	使用 edgeR 分析的 RNAseq 基因表达数据的 Log2 倍数变化
logFC.microarray:	使用 LIMMA 微阵列分析的 RNAseq 基因表达数据的 Log2 倍数变化
logFC.npSeq:	使用 npSeq 分析的 RNAseq 基因表达数据的 Log2 倍数变化
microarray:	使用诸如来自 Affymetrix 的基因阵列进行基因表达分析
NML:	正常肺，通常从基本上未经移植的人肺供体组织获得
npSeq:	用于测序数据的显著性分析的 R 包
NSIP:	非特异性间质性肺炎
OP:	机化性肺炎
P.value.edgeR:	使用 edgeR 分析的 RNAseq 基因表达数据的 p 值
P.value.microarray:	使用 LIMMA 微阵列分析的 RNAseq 基因表达数据的 p 值
P.value.npSeqp:	使用 npSeq 分析的 RNAseq 基因表达数据的值
RB:	呼吸性细支气管炎
REST:	除了与之进行比较的子类型的所有其他 ILD 的组合。通常是 HP 和 NSIP、BRONCH、CIF-NOC、OP、RB 和 SARC。
SARC:	结节病
SQC:	鳞状细胞癌
TCID:	“TCID”或“转录物簇集标识符”是指所有 Affymetrix 微阵列使用的基因水平标识符。每个 TCID 与固定的参考编号相关联，该参考编号识别具有特定基因序列的一组特异性探针。此类特异性探针存在于可从 Affymetrix 商购的给定阵列上。因此，TCID 编号是指特定基因的基因产物，并且可在例如以下万维网地址：affymetrix.com/中找到，其中探针和基因产物的序列在此以其整体并入本文。
UIP:	寻常型间质性肺炎；在 IPF 中观察到的 HRCT 或组织病理学模式
LIMMA:	微阵列数据的线性模型；用于微阵列数据的显著性分析的 R 包

[0282] “ENSEMBL ID”是指来自Ensembl Genome Browser数据库的基因标识符号码(参见万维网地址:ensembl.org/index.html,其通过引用整体并入本文)。每个标识符以字母ENSG开头,表示“Ensembl Gene”。每个ENSEMBL ID号(即Ensembl数据库中的每个“基因”)是指由特定人染色体上的特定起始和终止位置定义的基因,因此定义了人类基因组的特定基因座。本领域技术人员可充分理解,本文公开的所有基因符号均指基因序列,其可在可公开获得的数据库中容易地获得,例如UniGene数据库(Pontius JU,Wagner L,Schuler

GD.UniGene:a unified view of the transcriptome,见NCBI手册,Bethesda(MD):国家生物技术信息中心;2003,可在万维网地址:ncbi.nlm.nih.gov/unigene中获得,并入本文)、RefSeq(NCBI手册[互联网],Bethesda(MD):国家医学图书馆(美国),国家生物技术信息中心;2002年10月,第18章,The Reference Sequence(RefSeq)Project,可在万维网地址ncbi.nlm.nih.gov/refseq/获得,并入本文)、Ensembl(EMBL,可在万维网地址ensembl.org/index.html获得,并入本文)等等。本文公开的基因序列通过其基因符号、Ensembl ID和Entrez ID以其整体并入本文。

[0283] 本文引用的所有参考文献、专利和专利申请均出于所有目的以其整体并入。

实施例

[0284] 鉴于弥漫性实质病症的复杂性,ILD的诊断方法仍然颇具挑战。诊断方法强调临床、放射学和病理学数据的多学科评价。后者传统上强调SLB以最大化肺组织采样时的产量。可以作为诊断替代物的分子标志物的开发是令人感兴趣的。为了在临床上可用于诊断ILD,病理学的替代测试需要区分UIP与类似但病理上不同的疾病过程。

[0285] 我们假设,基因组分类器可以在多样化的患者群体中以高准确性检测TBB中的UIP基因表达特征。在以下实施例中,我们在外显子组富集的转录数据上使用机器学习来训练分类器以区分UIP与临床实践中遇到的各种ILD。然后,我们证明了此分类器在独立的多中心验证队列中准确地预测了UIP的存在。此外,出人意料地,我们证明了样品汇集能够改善诊断的灵敏度和特异性,并且分类器表现不受细胞异质性的影响。这是令人惊讶的,因为先前的研究表明,IPF中感兴趣的细胞是肺泡细胞;所以可以预期所有生物学都包含在肺泡细胞内。然而,我们的结果表明,肺泡细胞外的信号足以告知IPF分类,而这在之前尚未被描述。

[0286] 因此,本文公开的基因组分类器可以减少ILD诊断中对外科肺活检的需要,并且最终可以用于告知患有IPF的患者的诊断和治疗。

实施例1

样品收集、病理学诊断和标记

[0287] 前瞻性收集视频辅助胸腔镜手术(VATS)样本,以作为由Veracyte,Inc.(South San Francisco,CA)赞助的机构评审委员会(IRB)所批准的正在进行的多中心临床方案—新型基因组测试的支气管样品收集(BRAVE)—的一部分。从库存来源获得额外的VATS和外科肺活检样本。当可行时,在常规临床护理期间收集的高分辨率计算机断层扫描(HRCT)的扫描结果由专业放射科医师评审。根据ATS指南(Raghu G,等人,Am J Respir Crit Care Med 2011,183:788-824,其通过引用整体并入本文)总结放射学诊断。病理学诊断由专业病理学家(A-LK、TC、JM和SG)根据集中评审过程来确定。

[0288] 手术后,由研究所从外科肺活检物(SLB)、支气管镜肺冷冻活检物(BLC)或经支气管活检物(TBB)制备组织学载玻片,去掉身份信息,并提交给两位病理学家进行盲法的独立专家病理学评审。扫描选定的载玻片以构建显微图像的永久数字文件(Aperio,Vista,CA)。根据Kim SY等人,The Lancet Respiratory Medicine2015;3:473-482中描述的集中评审过程来评价载玻片,该文献其通过引用整体并入本文。

[0289] 每个病理学家整体地确定对患者的诊断(患者水平)并确定针对病理学采样的特定肺叶的诊断(样品或肺叶水平)。对诊断进行评价,其中一致的情况被定义为亚型一致性

(concordance)。在一致的情况下,对分类的UIP或非UIP“真值”标签进行定义,而在其他情况下,采用由第三位病理学家进行的盲法评审来达到2/3(“决胜局”)共识。在不一致的情况下,使用非盲法的合议过程。该过程也在图1中描述,从而产生样品水平和患者水平的病理学诊断。

[0290] 使用对来自相同肺叶的外科肺活检物(SLB)进行的病理学诊断将用于算法训练和发展的真值标签分配给TBB。将病理学亚型翻译成UIP或非UIP的样品和患者标签,以供用于如上的Kim SY等人所述的算法训练和验证,除非在肺下叶中检测到三个UIP模式但是在肺上叶中分配了非UIP或非诊断性标签的三个患者被分配患者水平的UIP标签(表14)。

[0291] 从每名患者收集多达5个TBB样品(两个肺上叶,三个肺下叶)用于分子测试。采样根据主治医师的判断进行,遵循从病理学采样附近的区域获得可见组织的指导。将UIP或非UIPD标签以肺叶级别分辨率分配给TBB样品,以供用于算法训练和样品评分。患者可以具有多个样品水平的诊断(即,每名患者每个VATS样品一个,最常见的是来自右肺下叶和上叶各一个),但只能具有一个患者水平诊断。对于混合物(参见实施例6),从样品标签推断出真值标签,以便对训练中的所有患者进行评分。

[0292] 在17个临床场所从84名患者总共收集了283个TBB样品,并将其用于本文报道的研究中。出于算法训练和评分的目的,以下病理学诊断被定义为非UIP:急性肺损伤、细支气管炎、脱屑性间质性肺炎、弥漫性肺泡损伤、肺气肿、嗜酸性粒细胞性肺炎、非特异性间质性肺炎(NSIP)(包括细胞亚型、混合亚型或Favor亚型)、肉芽肿病、过敏性肺炎(包括Favor亚型)、机化性肺炎、肺孢子虫肺炎、肺动脉高压、呼吸性细支气管炎、结节病以及吸烟相关的间质纤维化。

[0293] 出于算法训练和评分的目的,UIP被定义为任何UIP亚型(典型UIP、困难UIP、Favor UIP或UIP)。

[0294] 诊断一致性被定义为非UIP病理学或UIP的任何UIP亚型的亚型一致性。在亚型不一致的情况下(例如,Favor HP与HP,Favor NSIP与NISP),在会诊后接受共识诊断(例如,分别为HP和NSIP)。慢性间质纤维化的诊断——未另外分类、非诊断性或“其他”——未被分配训练标签并被排除在训练之外。

[0295] 如上所述,来自具有一致的UIP或非UIP诊断的患者的跨肺叶混合物被分配UIP或非UIP标签以用于混合物评分。3名具有肺下叶UIP模式但肺上叶为非UIP或非诊断性标签的患者被分配UIP标签以用于混合评分目的。

[0296] 大多数诊断术语遵循美国胸科学会(American Thoracic Society,ATS)的2011或2013指南[5,6],但专业病理学家小组作出了一些改变以更好地表现肺叶水平的特征。特别是,其中包括“典型UIP”和“困难UIP”,而不包括ATS 2011指南中描述的“明确UIP”和“很可能UIP”。未另外分类的慢性间质纤维化(CIF/NOC)对应于无法分类的纤维化ILD。CIF/NOC的三个子类别—“Favor UIP”、“Favor NSIP”和“Favor HP”—被定义为指定无法分类的纤维化病例,其在专业病理学小组的判断中,表现出暗示UIP、非特异性间质性肺炎(NSIP)或过敏性肺炎(HP)的特征。还包括吸烟相关的间质纤维化(SRIF)的诊断[20]。

[0297] 为了分类,将样品水平病理学诊断转换为二元分类标签(UIP和非UIP)。在病理学诊断类别中,该“UIP”类别包括(1)UIP、(2)典型UIP、(3)困难UIP,以及(4)Favor UIP。除非诊断性(ND)之外的所有其他病理学诊断被分配至“非UIP“类别。

实施例2

样品处理

[0298] 从患者收集术前或术中经支气管活检样本用于分子测试，在4℃下在核酸防腐剂中包装和运输，并在-80℃下的Veracyte设施中长期储存直至处理。简而言之，使用Tissue-Tek O.C.T.介质(Sakura Finetek U.S.A.)来固定冷冻组织样品用于切片，并使用CM1800低温恒温器(Leica Biosystems,Buffalo Grove,Illinois)来产生2×20μm切片。将组织卷立即浸入RNAprotect(QIAGEN,Valencia,California)中，在4℃下温育过夜并在-80℃下储存直至提取。只要有可能，就将相邻的5μm组织卷固定于载玻片上，并遵循标准程序进行苏木精和伊红(H&E)染色。

[0299] 使用改良的AllPrep™ Micro Kit(QIAGEN,Valencia,CA)程序从保存的TBB样品中提取核酸。简言之，根据制造商的说明书(QIAGEN)，在将DNA级分与RNA级分的基于柱进行分离之前使用TissueLyzer™和QIAshredder™将TBB组织彻底破坏并均质化。分别使用QuantiFluor™ RNA系统(Promega,Madison,WI)和Agilent RNA 6000Pico测定法(Agilent Technologies,Santa Clara,CA)来确定总RNA样品量和质量。我们还获得了源自人脑、心脏、肺、胎盘和睾丸(Life Technologies,Carlsbad,CA)、甲状腺和肺肿瘤(Takara Bio USA,Mountain View,CA)的总RNA(Asterand USA;Cooperative Human Tissue Network)，以及肺上皮细胞系(HBEC,NL-20,Beas2b;由Avrum Spira博士惠赠)。此外，还使用从22名BRAVE I患者的外科肺活检物中提取的总RNA(Kim SY等人，同上)。

[0300] 使用TruSeq™ RNA Access文库制备试剂盒(Illumina,San Diego,CA)根据制造商的说明书来制备富集外显子序列的RNA文库。简言之，在升高的温度下使用二价阳离子将RNA样品片段化成小段，并使用随机六聚体引物经由逆转录酶将片段化的RNA转化成cDNA。随后将cDNA文库用作第二链合成的模板；由此产生双链cDNA文库，根据制造商的方案将其与测序衔接子连接。最后，根据制造商的方案，通过两轮PCR扩增、验证和捕获探针杂交来产生富集的高特异性的衔接子连接的cDNA文库。

实施例3

下一代RNA测序

[0301] 在该实施例中，根据制造商的说明书，使用NextSeq™ 500仪器(Illumina)对满足过程中PCR产量标准的选定样品进行外显子组富集的下一代RNA测序，目标读取深度为每个样品至多2500万个配对末端读数，并且在过滤数据质量后，将17,601个Ensembl基因的表达计数进行归一化并输入机器学习算法。使用机器学习来训练弹性网络逻辑回归模型。通过交叉验证并针对31名患者的独立组来评价表现。测序和算法开发如下进行。

[0302] 简言之，使用Ovation™ RNASeq System v2(NuGEN,San Carlos,California)来扩增10ng总RNA，并制备TruSeq™(Illumina,San Diego,California)测序文库，并根据制造商的说明书将其在Illumina HiSeq上测序(如实施例2所述)。使用STAR RNAseq比对器软件(Dobin A,等人,Bioinformatics 2013Jan 1;29(1):15-21,其通过引用整体并入本文)将原始测序(FASTQ)文件与人参考组装物37(Genome Reference Consortium)比对。使用HTSeq(Anders S,等人,Bioinformatics 2015;31:166-169,其通过引用整体并入本文)来确定高达26,268个Ensembl注释的基因水平特征的读取计数。

[0303] 使用RNA-SeQC(DeLuca DS,等人,Bioinformatics2012;28:1530-153222,其通过

引用整体并入本文)生成测序数据质量度量。针对总读取、映射的独特读取、平均每碱基覆盖率、碱基重复率、与编码区对齐的碱基百分比、碱基错配率和基因内覆盖均匀性的接受度量来评价每次复制中的质量度量。将测序数据过滤以排除未被靶向用于通过文库测定来富集的特征,以及排除在Ensembl中注释为假基因的基因、T细胞受体或免疫球蛋白基因中的非表达外显子或rRNA,从而产生具有高特异性富集可信度的17,601个Ensembl基因。

[0304] 对于84名患者分类器(参见图2),还排除了在多个测定中具有可变表达的基因(总测定间SD>0.3),这导致在运行之间有14,811个基因具有可再现的表达。在下游分析之前,通过基因分散功能对表达计数数据进行评分并将其VST转化。使用"使用分并将其现的表函数(https://www.r-project.org)在R中进行主成分分析。模型特征选择和参数估计通过具有弹性净罚分的逻辑回归来进行,如Friedman J等人,Journal of statistical software 2010;33:1-22中所述,其通过引用整体并入本文。通过留一患者的交叉验证(LOPO CV)来确定参数调整和表现评价。

实施例4

患者队列特征

[0305] 作为BRAVE研究的一部分(参见实施例9)的来自在18个临床场所登记的113名ILD患者的样品被筛选用于开发针对ILD的分子测试。图2示出了筛选用于本研究的113名患者和相关TBB样品的流程图,并且其在处理的每个连续步骤中示出了患者和样品的队列(中心方块)、处理步骤(不规则四边形)和排除(外侧方块)。在病理学诊断可用之前,将患者前瞻性地分配至训练集和测试集。在算法锁定和评分之后,实验室和分析人员仍然不知晓病理学诊断和测试集的标签。

[0306] 我们使用实施例1中所述的中心病理学评审过程,对这些患者中的95名获得了对肺的各个肺叶特异的病理学诊断。

[0307] 我们排除了需要非盲法评审(即,合议)的诊断和一名被诊断患有肺癌的患者,得到89名在至少一个肺叶中具有高置信度ILD病理的患者。

[0308] 我们从收集自113名患者的496个TBB样品中提取总RNA,并最终为来自108名患者的407个样品生成了高质量的RNAseq数据。

[0309] 诊断患者和高质量样品数据的结合代表来自84名患者(52名UIP和32名非UIP)的283个样品(图2,表2)。

[0310] 我们前瞻性地将53名患者分配至算法训练,并将31名患者分配至验证队列,目标是训练队列与测试队列之间的等效UIP患病率(表2)。

表 2：人口统计学和 UIP 患病率

	训练集	测试集	总计
受试者数目	53	31	84
临床因素			
年龄，中值(范围)	63.5 (31-88)	62 (18-78)	63 (18-88)
男性，数目(%)	26 (49%)	14 (45%)	40 (48%)
吸烟史，是，否(%)	34 (64%)	19 (61%)	53 (63%)
通过病理学获得的 UIP 患病率			
通过外科肺活检，UIP 数目(%)	26/38 (68%)	17/22 (77%)	43/60 (72%)
典型 UIP	11	6	17
UIP	9	6	15
困难 UIP	5	5	10
Favor UIP	1	0	1
通过冷冻活检，UIP 数目(%)	6/11 (55%)	2/6 (33%)	8/17 (47%)
UIP	2	0	2
困难 UIP	0	1	1
Favor UIP	4	1	5
通过经支气管活检,UIP 数目(%)	1/4 (25%)	0/3 (0%)	1/7 (14%)
困难 UIP	1	0	1
总 UIP 患病率，UIP 数目(%)	33/53 (62%)	19/31 (61%)	52/84 (62%)
通过放射学获得的 UIP 患病率			
明确 UIP	4	2	6
UIP	4	2	6
很可能 UIP	0	1	1
总 UIP 患病率，UIP 数目(%)	8/52 (15%)	5/27 (19%)	13/79 (16%)

[0311] 由于我们的预期集合中有几种罕见的非UIP ILD,一些亚型由患者队列中的单个病例表示(表14,图3)。将单个病例的细胞NSIP、Favor HP、肺气肿和肺孢子虫肺炎分配至训练集,而将单个病例的弥漫性肺泡损伤、肺动脉高压和嗜酸性粒细胞性肺炎分配至测试集。在这些患者中普遍存在的ILD亚型的多样性和缺乏性说明了在临床实践中遇到的在平衡ILD谱上训练基因组分类器的挑战。

[0312] 作为常规临床护理的一部分而对我们的患者队列进行的放射学提供了UIP患病率的独立估计。我们对可用的HRCT扫描进行了专家评审,并根据UIP模式的ATS标准对放射学结果进行了总结(Raghu G.,2011,同上)。我们队列中HRCT UIP模式的患病率为16%,而所有病理活检类型的该患病率为62%(表2)。相比于支气管镜活检,SLB中UIP的患病率更高(72%比47%[冷冻活检]比14%[经支气管活检]),通常在SLB中鉴定出所确定的UIP(表2)。

实施例5

使用各个TBB样品进行的分类器开发和表现

[0313] 我们使用多个基因组和临床特征,在我们来自53名患者的170个TBB样品的训练集上评价了多种归一化方案、特征选择和机器学习算法。在交叉验证中,我们从具有在表达计数数据上训练的弹性净罚分的逻辑回归模型中观察到最高和最稳定的分类表现,其使用

169个基因作为特征(表15)。该模型基于样品水平数据在训练集上的交叉验证中获得了0.85的受试者工作特征曲线下面积(ROC-AUC)(图3A,图3B)。

[0314] 我们定义了靶向高(92%)特异性的决策边界,并观察到相应的65%的灵敏度(图3A,图3B)。

[0315] 使用该分类器,对来自31名患者的113个样品的独立测试集的TBB样品进行了前瞻性评分,并且分类器显示基于样品水平数据的ROC-AUC为0.86,灵敏度为63%[95%CI:43-87],特异性为86%[95%CI:73-97](图3C,图3D)。推广到验证队列的交叉验证表现表明,尽管队列大小相对不太大,但仍然实现了稳健的训练。

[0316] 在来自84名患者的283个TBB样品的组合队列上再训练的算法在来自每名患者的所有TBB样品均被单独评分时得到交叉验证的ROC-AUC为0.87[CI:0.82-0.91](灵敏度为63%[CI:54-72],特异性为91%[CI:80-98])。在更大的样品集上,类似的交叉验证结果(如本案例中所观察到的)是有希望的,并且将需要在另外的独立测试集上进行评价,目前计划在BRAVE研究中积累预期患者。

[0317] 因此,我们已经证明使用基因表达特征的UIP基因组分类器可有效地将UIP特有的空间和时间异质性纤维化疾病模式与同免疫应答(RB-ILD/DIP、嗜酸性粒细胞性肺炎、肉芽肿性疾病)、炎症(NSIP,HP)相关或作为对损伤的急性反应的均一且典型的活性纤维化区分开[10]。

[0318] 使用R版本3.0.1[21]进行所有统计学分析。对于微阵列分类器,通过limma [26]对UIP与非UIP类别之间差异表达的基因进行排序,然后将具有最低错误发现率(FDR)(<0.0003)的前200个基因作为模型构建的候选基因来继续推进。使用不同方法构建了几个模型,并选择具有最低误差的模型。使用glmnet [27]通过具有lasso罚分的逻辑回归进行特征选择和模型估计。对于RNAseq分类器,通过FDR对基因进行排序,该FDR是根据原始计数数据而在DESeq2 [22]包中实施的Wald样式检验所得出的。将排名靠前的特征(N从10到200)用于在归一化表达数据上使用e1071文库[24]来训练线性支持向量机(SVM)[23]。

[0319] 通过CV和独立测试集(若可用)来评价分类器表现。为了最大限度地减少过度拟合,在定义训练/测试集和CV分区时,将单个患者保持为最小单位;即,属于同一患者的所有样品在训练/测试集或CV分区中作为一组保持在一起。所使用的CV方法包括留一患者(LOPO)和10倍患者水平的CV。

[0320] 图3示出了单样品分类表现的结果。

[0321] 在给定的评分阈值下,将表现报告为曲线下面积(AUC)和特异性(1.0-假阳性率)和灵敏度(1.0-假阴性率)。我们将评分阈值设置为要求至少>90%的特异性。对于每次表现测量,使用2000个分层的自举重复和pROC包[22]来计算95%置信区间,并报告为[CI下限-上限]。

实施例6

使用汇集的样品进行的分类器开发和表现

[0322] 尽管实施例5中实现的总体单一样品表现是优异的,但是分类器在来自一些UIP患者的一些样品中未检测到UIP(FN)。由于在来自同一患者的其他样品中经常检测到UIP,因此采样效应(组织采样不足或疾病异质性)疑似为FN的来源。

[0323] 我们没有观察到假阴性样品中肺泡含量的系统性降低,从而排除了肺泡组织的不

充分采样的原因(参见实施例7)。因此,疾病异质性或技术样品质量影响仍然是假阴性的可能解释。重要的是,我们选择专家病理学评论作为UIP存在的参考标准。尽管已知操作者之间的不一致性[4,29,30],但我们在83%的患者中实现了在亚型水平下的两位专业病理学家之间的盲法一致性。

[0324] 通过设计,我们的临床研究收集每名患者的多个TBB样品,通常每个肺叶2至3个样品,以减轻可能的疾病和采样异质性影响,该影响可能导致训练错误或错误的测试判定。对于大多数患者,我们的样品水平分类器在每名患者的超过一个可用TBB样品中正确检测疾病,与整体的高样品水平的测试准确性一致(图3)。这提高了通过汇集每名患者的多个样品而使基于多个TBB样品的混合物的UIP的患者水平报告可行的可能性,并且我们假设这样的混合物可以在患者水平下总体提高检测准确性。因此,我们评价了涉及多个TBB样品的组合的测试设计。

[0325] 我们首先使用计算机模拟方法来推导出模拟将来自同一患者的多个TBB样品进行混合以产生单个测试结果的模型。该方法在此被称为"计算机混合"或可互换地称为"计算机汇集"。通过在方差稳定化转化(VST)之前对缩放的基因计数数据进行平均,从多个样品来模拟计算机患者自身(within-patient)混合物。在每种条件下进行100次模拟,在VST水平上加入基因水平的技术变异性。

[0326] 然后将来自计算机模拟混合物的评分与8名患者在体外产生的实际混合物以及相应的各个(例如未混合的)TBB样品的评分进行比较(图4A)。结果表明,我们的计算机模拟合理地近似于从实际混合物和各个TBB样品中观察到的评分。

[0327] 然后,我们使用该分析方法模拟每名患者2至5个TBB的混合物,该TBB在每名患者中随机选择(图4B)。通过模拟,每名患者两个或三个样品的混合物显示出相对于随机选择的每名患者的单个样品的分类准确性增加(图4B)。此外,4或5个TBB的混合物在表现估计中表现出降低的可变性(即,更高的置信度),且具有相似的最大准确性(图4B)。在约90%的目标特异性下,混合物中的测试灵敏度提高至约72%,变异性降低(AUC=0.90[CI 0.88-0.93],在90%特异性[CI 60-81]下灵敏度=72%)(图4C)。在每名受试者可获得两个肺上叶和三个肺下叶TBB的一组33名受试者中,当采样限于肺上叶或肺下叶时,混合物模拟显示表现没有改善(图4D)。该分析表明,使用单一分子测试,每名患者多达5个TBB样品的混合物可最大化UIP模式的准确检测。这样的结果可能是令人惊讶的,因为预期汇集会由于细胞异质性而引入更多的可变性。

[0328] 因此,物理或计算机混合研究表明,对每名患者的多次采样进行组合促使准确性提高。

[0329] 通过分别训练所有样品(即,如实施例5中所述),我们使表示和采样多样性最大化,并减轻了可用样品的先验子采样偏差。正如测试准确性提高所证明的,通过对样品混合物进行测试,我们似乎可以减轻采样影响。

实施例7

采样异质性和表现

[0330] 鉴于ILD患者的肺中存在确定的疾病异质性[4,21-23],可通过肺的可变采样来获得较强分类器表现的这一发现引起了在TBB程序期间是否需要充分的肺泡采样的问题。我们假设如果UIP与非UIP的准确分类需要来自肺泡细胞的基因信号,则那些缺乏肺泡细胞的样品

应该比具有更高肺泡含量的样品产生更多的分类器错误(特别是FN)。为了解分类器准确性是否取决于充分的肺泡采样,我们测试了分类器准确性与肺泡特异性基因之间的相关性。

[0331] 具体地,我们首先开发了TBB中肺泡含量的半定量基因组测量,然后使用该度量来确定它是否与分类器准确性相关。评价了TBB样品中44种肺特异性基因的表达,这些基因在文献中被报道为是细支气管、肺泡和肺祖细胞的标志物[E5-E9](表16)。使用44种标志物通过主要成分的无监督聚类表明,该TBB队列代表所采样的肺组织的连续谱,其中一部分与外科肺活检重叠(图5A;TBB样品为蓝色,SLB样品为橙色)。

[0332] 我们开发了两种肺泡量度,一种用于I型肺泡细胞,另一种用于II型肺泡细胞。

[0333] 对于I型肺泡统计,我们总结了两个基因—PDPN和AQP5—的表达。这些基因在样品集中显示出连续的基因表达模式。

[0334] 我们用于II型肺泡量度的第二种方法是检查在TBB群体内显示出双峰表达的证据的标志物。这种模式见于9个基因,其中5个(SFTPB、SFTPC、SFTPD、ABCA3、CEBPA)是肺泡II型(ATII)特异性的,3个(AGER、GPRC5A、HOPX)是肺泡I型(ATI)特异性的,且1个(SFTPA1)同时见于I型和II型细胞(图5B;TBB表达计数为蓝色,SLB表达计数为橙色)。在SFTPA1、SFTPB、SFTPC和SFTPD之间观察到相关的、方向一致的表达,但在PDPN与AQP5之间,或在这两组的成员之间没有观察到该现象(图5C;TBB表达计数为蓝色,SLB表达计数为橙色)。因此,我们选择四种表面活性蛋白SFTPA1、SFTPB、SFTPC和SFTPD作为II型肺泡含量的标志物,并将它们在样品中的表达相加来作为每个样品中肺泡含量的代表性量度。

[0335] 虽然这些量度显示了各种样品类型中的宽范围的I型和II型肺泡特异性基因表达,其中在SLB和许多TBB中具有高表达,而在多种非肺组织类型和三种支气管上皮细胞系(Beas2b、HBEC和NL-20)中具有低表达(图6A),但这些转录物的表达与分类器准确性无关(Pearson相关性,0.03,p-值=0.61)。因此,这些结果表明,假阴性和假阳性误差都与较低的I型或II型肺泡I含量无关,这暗示可以在具有可变细胞组成的TBB样品中实现准确的分类结果(图6B)。

[0336] 我们还发现单个样品的分类器准确性与TBB肺泡基因表达、RNA质量或RNA产量之间没有显著相关性(表3)。

表3:TBB样品性质与分类准确性的Pearson相关性

样品性质	相关性	p-值
肺泡 I 表达统计学	-0.07	0.27
肺泡 II 表达统计学	0.03	0.61
RNA 质量		
RIN	-0.07	0.24
DV_{200}	-0.10	0.09
RNA 产量, 纳克	0.06	0.32

[0337] 这些结果表明,在具有可变肺泡含量的TBB样品中可以实现准确的分类结果。

实施例8

与分类器所使用的基因相关的生物学途径

PANTHER™途径分析

[0338] 我们使用DESeq2[19]来鉴定来自84名具有病理学真实性的患者的UIP与非UIP TBB之间的差异表达。在经错误发现率(FDR)调整的p值≤0.05下在UIP(n=926)和非UIP(n=

1330)中显著上调的Ensembl基因被用作PANTHER™分类系统的输入，用于途径超表现分析(网络版本11.0,2016-07-15发布)(Mi H,Lazareva-Ulitsky B.等人,Nucleic Acids Res 2005;33:D284-D288,其通过引用整体并入本文)。PANTHER™途径被筛选(curated)以去除一般或冗余的途径分类，并按重要性排序(表4)。我们发现具有UIP的TBB显著富集了细胞代谢、粘附和发育过程的标志物的表达，而非UIP TBB则显示出免疫激活、脂质代谢、应激反应和细胞死亡的证据(表4)。发育途径和细胞增殖的异常重新激活是IPF的标志[24-27]。

表 4：在 UIP 和非 UIP TBB 样品中超表现的生物学过程

生物学过程	预期数目	观察到的数目	增加倍数	P-值
在 UIP 中超表现				
细胞-细胞粘附	13	44	3.4	<0.0001
细胞组分形态发生	23	53	2.3	<0.0001
神经系统发育	29	63	2.2	<0.0001
转录，DNA 依赖性	65	122	1.9	<0.0001
RNA 代谢过程	88	144	1.6	<0.0001
核碱基代谢过程	135	189	1.4	0.0002
氮化合物代谢过程	86	129	1.5	0.0008
外胚层发育	17	39	2.3	0.0010
视觉感知	8	23	2.8	0.0036
中胚层发育	19	37	1.9	0.0371
肌肉收缩	7	18	2.7	0.0398
在非 UIP 中超表现				
抗原处理和呈递	4	20	5.3	<0.0001
细胞防御反应	13	39	3.0	<0.0001
脂质代谢过程	33	68	2.1	<0.0001
免疫系统过程	79	131	1.7	<0.0001
胆固醇代谢过程	5	19	3.8	0.0003
类固醇代谢过程	11	30	2.7	0.0004
免疫反应	44	74	1.7	0.0054
凋亡过程	26	49	1.9	0.0057
含磷酸盐的化合物代谢	77	114	1.5	0.0076
I-kappaB 激酶/ NF-kappaB 级联	4	15	3.4	0.0157
对压力的反应	53	83	1.6	0.0172
跨膜酪氨酸激酶信号传导	12	28	2.3	0.0213
分解代谢过程	49	77	1.6	0.0222
造血	6	17	2.9	0.0328

实施例9

BRAVE研究设计

[0339] BRAVE(用于新型基因组测试的支气管样品收集)研究的目的在于收集支气管镜样本、临床数据和相关病理学载玻片以供外部检查,从而优化分子谱分析测试,该分子谱分析测试将提供许多关于间质性肺病(ILD)的诊断和预后信息。

[0340] 将BRAVE分为三组:BRAVE-1旨在登记将诊断性外科肺活检(SLB)计划作为其常规护理临床诊断的一部分的患者。BRAVE-2仅旨在用于计划进行诊断性支气管镜检查的患者。BRAVE-3旨在用于计划进行诊断性冷冻活检的患者。

[0341] 还收集了支气管肺泡灌洗液、血液、血清和口腔拭子。持续登记受试者,直到收集到足够数目的样品以满足ILD分子测试的开发和前瞻性验证的功效和样品大小要求。

[0342] 在样品采集后跟踪受试者长达一年以评估疾病的进展。年龄小于18岁且SLB未经医学指明的患者,或因非ILD医学状况而接受过SLB的患者不符合研究登记的资格。具有是进行支气管镜活检的禁忌症的医学状况的患者或者不建议或难以进行支气管镜采样的患者也被排除在BRAVE研究之外。

实施例10

Envisia分类器的生成

[0343] 在已证明机器学习可以检测通过SLB和TBB获得的肺组织中的UIP组织病理学模式的情况下(参见实施例1-9),我们试图在更大和更多样化的患者组中扩展分类器训练,并在一组独立的/前瞻性收集的受试者上验证锁定算法。

方法

[0344] 共有201名受试者在18个美国和欧洲的场所的前瞻性多中心研究中登记。我们收集了每名受试者多达5个TBB,在肺叶水平下与标准护理肺组织活检样品配对。获得了针对139名受试者由三位专业病理学家的小组得到的组织学模式诊断。在汇集的TBB上进行外显子组富集的RNA测序,比对所得到的序列,并提取xyx基因的转录物计数。我们使用大约90名患者来训练并锁定机器学习算法——Envisia基因组分类器,然后在具有组织学参考标签的49名受试者的独立组上验证该测试。我们优化了测试决策边界以提供高特异性,即减少假阳性,因为这可能通过过度判定出UIP模式而造成损害,可能导致IPF治疗的不必要的风险和费用。我们锁定了所有分类器参数,并定义了在测试指示范围内的患者和样品特征。我们在此报告了Envisia基因组分类器在来自49名受试者的独立队列的TBB中的前瞻性临床验证,并将其分类表现与HRCT进行比较。

研究设计和监督

[0345] 对于该独立验证研究,在三个独立的BRAVE研究中登记总计88名受试者(图1)。在BRAVE-1中,受试者接受临床指示的SLB(n=43);BRAVE-2受试者接受临床指示的TBB(n=9);而BRAVE-3受试者接受临床指示的冷冻活检(n=36)。BRAVE-2受试者仅有TBB用于组织病理学评价,BRAVE-1受试者和BRAVE-3受试者分别通过SLB或冷冻活检来接受诊断。

[0346] 收集多达五个专门的经支气管活检物(TBB)(通常为两个肺上叶和三个肺下叶),用于从由参与医师鉴定的用于组织病理学诊断的临床指示活检物的相同肺叶进行分子测试。向Veracyte提供了研究指示的TBB样本;研究场所制备的组织病理学载玻片和去除患者身份信息的临床数据;胸部HRCT;局部临床诊断;以及一年和两年的随访(若可行)。未向参与的医师提供分子测试的结果,也未将其用于告知患者诊断或治疗。

[0347] HRCT扫描由专业胸部放射科医师(D.Lynch)进行评审并分类为明确UIP、很可能

UIP或可能UIP9、脱屑性间质性肺炎(DIP)、过敏性肺炎(HP)、Langerhans细胞组织细胞增生症(LCH)、非特异性间质性肺炎(NSIP)、机化性肺炎(OP)、呼吸性细支气管炎(RB)、结节病或“其他”(无法分类)。Veracyte临床人员使用相同的标准来评审和解释研究场所的放射学描述,但用“与UIP不一致”来代替特定的非UIP诊断[35]。

[0348] 来自临床指示的SLB、冷冻活检或TBB的组织病理学载玻片由对患者临床信息不知情的两位或三位专业肺病理学家独立评审,如先前描述的[E1,E10]。每位病理学家独立地确定每个采样的肺叶的组织病理学模式诊断。我们将共识定义为三位评审病理学家中的两位或者两位评审病理学家中的两位达成组织学模式水平下的盲法一致,或者如果未达成盲法一致,在三位病理学家(合议)之间进行非盲会诊之后达成一致。

[0349] 根据以下类别,Veracyte人员基于肺叶水平诊断的共识将UIP或非UIP的参考标签分配至每个研究受试者(图9)。如果任何肺叶通过病理学诊断为UIP,则该受试者被分配UIP标签[4]。如果所有其他肺叶也是非UIP或非诊断性的,则在任何肺叶中诊断为非UIP病理学的受试者被分配非UIP参考标签(图9)。在测试和算法开发过程中,所有Veracyte实验室和分析人员都对参考标签不知情。

实验室测试程序

[0350] 将来自88名BRAVE患者的研究指示的TBB收集到专用核酸防腐剂(RNAprotect,QIAGEN,Valencia,CA)中,在现场冷藏至多14天并运送到Veracyte进行处理。我们使用改良的AllPrep Micro程序(QIAGEN)提取总RNA,然后使用RNA结合染料荧光(QuantiFluor,Promega,Madison,WI)进行定量。我们预先指定,每名受试者至少3个、最多5个TBB,其中各自产生至少31ng总RNA,是研究纳入所必需的。由于样品数目或RNA产量不足,排除了9名受试者(图8)。此外,还前瞻性地排除了含有异物(牙签,一名受试者)的样本,递送至Veracyte但缺少防腐剂的样本(一名受试者),以及超过48小时的集装箱冷却限制的装运样本(5名受试者)(图8)。因此,72名受试者的各个TBB满足了我们预先指定的研究纳入标准。

[0351] 将每名受试者的汇集的RNA输入至部分自动化的TruSeq RNA Access文库制备程序(Illumina,San Diego,CA)以富集表达的外显子序列,并在NextSeq 500仪器(Illumina)上测序至≥25M的成对末端读取的目标深度。针对经测序和独特映射的读取的总数、经映射读取和外显子读取的总比例、平均单碱基覆盖率、碱基覆盖度的均匀性,以及碱基重复和错配率的标准来评价计数数据。来自一名受试者的数据不符合这些标准并且被排除,留下71名受试者(图8)。表达计数数据相对于测序深度(比例因子)归一化,并在分类之前使用DESeq216通过方差稳定化转化进行转换。

算法开发

[0352] 将来自先前在2012年12月至2015年7月的BRAVE研究[E10]中登记的90名受试者的354个个体TBB样品专门用于训练机器学习算法(分类模型)。通过使用弹性网络逻辑回归的算法进行特征选择和超参数优化。使用受试者工作特征曲线下面积(ROC-AUC)在训练集中评价模型的表现,该面积由留一患者的交叉验证(CV)确定。选择优化训练集中的特异性(最小化UIP假阳性判定)的测试决策边界。因此定义了使用190个基因作为特征且具有锁定的决策边界的罚分逻辑分类器(Envisia基因组分类器)(表5)。Envisia报告了每个TBB池的UIP或非UIP的分子诊断。分类评分高于决策边界的受试者被Envisia判定为UIP,而评分等于或低于决策边界的受试者判定为非UIP。在揭示参考标签之前,由不参与测试开发的第三方内

部地且独立地对评分进行验证。

表5:Envisia基因组分类器所使用的190个基因

基因 ID	基因符号
ENSG00000005381	MPO
ENSG00000005955	GGNBP2
ENSG00000007908	SELE
ENSG00000007933	FMO3
ENSG00000010379	SLC6A13
ENSG00000012232	EXTL3
ENSG00000022556	NLRP2

ENSG00000026950	BTN3A1
ENSG00000033050	ABCF2
ENSG00000038295	TLL1
ENSG00000048052	HDAC9
ENSG00000054803	CBLN4
ENSG00000054938	CHRDL2
ENSG00000060688	SNRNP40
ENSG00000071909	MYO3B
ENSG00000072310	SREBF1
ENSG00000073605	GSDMB
ENSG00000078070	MCCC1
ENSG00000079385	CEACAM1
ENSG00000081041	CXCL2
ENSG00000081985	IL12RB2
ENSG00000082781	ITGB5
ENSG00000083814	ZNF671
ENSG00000086544	ITPKC
ENSG00000089902	RCOR1
ENSG00000092295	TGM1
ENSG00000099251	HSD17B7P2
ENSG00000099974	DDTL
ENSG00000100376	FAM118A
ENSG00000100557	C14orf105
ENSG00000101544	ADNP2
ENSG00000102837	OLFM4
ENSG00000103044	HAS3
ENSG00000103257	SLC7A5
ENSG00000104812	GYS1
ENSG00000105255	FSD1
ENSG00000105559	PLEKHA4
ENSG00000105696	TMEM59L
ENSG00000105784	RUNDC3B
ENSG00000105983	LMBR1
ENSG00000106018	VIPR2
ENSG00000106178	CCL24
ENSG00000107929	LARP4B
ENSG00000108312	UBTF
ENSG00000108551	RASD1
ENSG00000109205	ODAM
ENSG00000110092	CCND1
ENSG00000110900	TSPAN11
ENSG00000110975	SYT10
ENSG00000111218	PRMT8
ENSG00000111321	LTBR
ENSG00000111328	CDK2AP1
ENSG00000112164	GLP1R
ENSG00000112299	VNN1
ENSG00000112852	PCDHB2
ENSG00000114248	LRRC31

ENSG00000114923	SLC4A3
ENSG00000115415	STAT1
ENSG00000115607	IL18RAP
ENSG00000116285	ERRFI1
ENSG00000116761	CTH
ENSG00000119711	ALDH6A1
ENSG00000119725	ZNF410
ENSG00000120217	CD274
ENSG00000120738	EGR1
ENSG00000120903	CHRNA2
ENSG00000121380	BCL2L14
ENSG00000121417	ZNF211
ENSG00000122497	NBPF14
ENSG00000124205	EDN3
ENSG00000124702	KLHDC3
ENSG00000124935	SCGB1D2
ENSG00000125255	SLC10A2
ENSG00000128016	ZFP36
ENSG00000128266	GNAZ
ENSG00000128791	TWSG1
ENSG00000128891	C15orf57
ENSG00000130164	LDLR
ENSG00000130487	KLHDC7B
ENSG00000130598	TNNI2
ENSG00000131095	GFAP
ENSG00000131142	CCL25
ENSG00000132199	ENOSF1
ENSG00000132204	LINC00470
ENSG00000132915	PDE6A
ENSG00000132938	MTUS2
ENSG00000133636	NTS
ENSG00000133794	ARNTL
ENSG00000134028	ADAMDEC1
ENSG00000134245	WNT2B
ENSG00000135148	TRAFD1
ENSG00000135447	PPP1R1A
ENSG00000135625	EGR4
ENSG00000136881	BAAT
ENSG00000136883	KIF12
ENSG00000136928	GABBR2
ENSG00000136933	RABEPK
ENSG00000137285	TUBB2B
ENSG00000137463	MGARP
ENSG00000137573	SULF1
ENSG00000137709	POU2F3
ENSG00000137968	SLC44A5
ENSG00000138166	DUSP5
ENSG00000138308	PLA2G12B
ENSG00000140274	DUOXA2

ENSG00000140279	DUOX2
ENSG00000140323	DISP2
ENSG00000140450	ARRDC4
ENSG00000140465	CYP1A1
ENSG00000140505	CYP1A2
ENSG00000140718	FTO
ENSG00000141279	NPEPPS
ENSG00000142178	SIK1
ENSG00000142661	MYOM3
ENSG00000143185	XCL2
ENSG00000143195	ILDR2
ENSG00000143320	CRABP2
ENSG00000143322	ABL2
ENSG00000143367	TUFT1
ENSG00000143379	SETDB1
ENSG00000143603	KCNN3
ENSG00000144655	CSRNP1
ENSG00000145248	SLC10A4
ENSG00000145284	SCD5
ENSG00000145358	DDIT4L
ENSG00000145736	GTF2H2
ENSG00000148541	FAM13C
ENSG00000148700	ADD3
ENSG00000148702	HABP2
ENSG00000149043	SYT8
ENSG00000149289	ZC3H12C
ENSG00000151012	SLC7A11
ENSG00000151572	ANO4
ENSG00000152672	CLEC4F
ENSG00000153404	PLEKHG4B
ENSG00000154227	CERS3
ENSG00000154451	GBP5
ENSG00000156414	TDRD9
ENSG00000157103	SLC6A1
ENSG00000157680	DGKI
ENSG00000158457	TSPAN33
ENSG00000159231	CBR3
ENSG00000159674	SPON2
ENSG00000161609	CCDC155
ENSG00000162594	IL23R
ENSG00000163029	SMC6
ENSG00000163110	PDLIM5
ENSG00000163285	GABRG1
ENSG00000163412	EIF4E3
ENSG00000163635	ATXN7
ENSG00000163644	PPM1K
ENSG00000163735	CXCL5
ENSG00000163817	SLC6A20
ENSG00000163884	KLF15

ENSG00000164604	GPR85
ENSG00000164821	DEFA4
ENSG00000165948	IFI27L1
ENSG00000165973	NELL1
ENSG00000165983	PTER
ENSG00000166923	GREM1
ENSG00000167748	KLK1
ENSG00000168004	HRASLS5
ENSG00000168036	CTNNB1
ENSG00000168062	BATF2
ENSG00000168394	TAP1
ENSG00000168661	ZNF30
ENSG00000168938	PPIC
ENSG00000169248	CXCL11
ENSG00000170113	NIPA1
ENSG00000170442	KRT86
ENSG00000170509	HSD17B13
ENSG00000170837	GPR27
ENSG00000171016	PYGO1
ENSG00000171408	PDE7B
ENSG00000171649	ZIK1
ENSG00000171714	ANO5
ENSG00000172137	CALB2
ENSG00000172183	ISG20
ENSG00000172215	CXCR6
ENSG00000172667	ZMAT3
ENSG00000173809	TDRD12
ENSG00000173812	EIF1
ENSG00000173926	MARCH3
ENSG00000175764	TTLL11
ENSG00000175806	MSRA
ENSG00000176046	NUPR1
ENSG00000177182	CLVS1
ENSG00000177294	FBXO39
ENSG00000178187	ZNF454
ENSG00000178229	ZNF543

统计学分析

[0353] 使用3.2.3版的R软件(https://www.r-project.org)进行统计学分析。通过学生t检验比较连续变量,并通过卡方检验比较分类变量。除非另有说明,否则所有置信区间均为双侧95%。我们使用预测准确性的标准量度来评估测试表现。我们使用先前开发的肺泡I型和II型基因表达评分[14]来评估测试准确性是否与肺泡细胞基因表达相关。我们使用在http://genetrail.bioinf.uni-sb.de/上可得的GeneTrail软件对在训练队列TBB(UIP与非UIP)中差异表达的前1000个基因与190个分类基因的组合进行了生物学途径分析。

结果

研究受试者的人口统计学和病理学特征

[0354] 于2014年8月至2016年5月之间在18个美国和欧洲临床场所处进行的3个BRAVE研究之一中登记了总计88名受试者(图8)。由于样本处理不当或材料不足,我们在分析测试之

前排除了16名受试者，并且排除了一名在测试期间未能通过分析QC的受试者。总计71名受试者满足研究纳入标准。其中，两名组织病理学载玻片缺失的受试者和一名患有肺腺癌的受试者随后被排除，留下总计68名受试者进行病理学检查。我们无法将UIP/非UIP病理学参考标签分配给具有非诊断性病理学的12名受试者和7名具有无法分类的纤维化的受试者，因此未将他们包括在最终验证中(图8)。确定最终的组织病理学模式诊断，并为剩余的49名受试者提供UIP/非UIP参考标签。这49名受试者成为最终验证组(图8)。

[0355] 与88名登记的受试者和39名排除的受试者相比，最终验证组中的49名受试者在受试者年龄、性别或吸烟状态方面没有显示出显著差异(表6)。最终验证集包括多种UIP亚型以及临床实践中可能遇到的非UIP ILD(表7)。

表6：研究受试者的临床特征。

	研究合格组 N, (%)	最终验证组 N, (%)	P-值
性别–n (%)			1
女性	38 (43%)	21 (43%)	
男性	50 (57%)	28 (57%)	
平均年龄(SD)–岁	63.0 (11.7)	64.1 (10.3)	0.56
吸烟状态 –n (%)			0.73
是	56 (64%)	33 (67%)	
否	28 (32%)	15 (31%)	
未知	4 (5%)	1 (2%)	
场所–n (%)			0.51
学校	36 (41%)	20 (41%)	
社区	37 (42%)	24 (49%)	
欧洲	15 (17%)	5 (10%)	
研究 –n (%)			0.45
BRAVE 1	43 (49%)	26 (53%)	
BRAVE 2	9 (10%)	2 (4%)	
BRAVE 3	36 (41%)	21 (43%)	
通过病理学的 UIP 患病率, n (%)	N/A	24 (49%)	
通过放射学的 UIP 患病率, n/n (%)	N/A	9/46 (20%)	
放射学错判, n (%)	N/A	3 (6%)	
受试者总数	88	49	

表7：验证中呈现的ILD病理学模式

病理学模式诊断	最终验证组 N，（%）
UIP 分类标签	
UIP（典型 UIP，困难 UIP 或 Favor UIP）	19（39%）
伴有 Favor HP 的 UIP；伴有 CIF、NOC 的 UIP；伴有 NSIP 的 UIP；伴有肺动脉高压的 UIP	5（10%）
UIP 总计	24（49%）
非 UIP 分类标签	
OP；伴有 CIF、NOC 的 OP；伴有急性肺损伤的 OP	1（2%）
呼吸性细支气管炎；SRIF；伴有 SRIF 的 RB；伴有 CIF、NOC 的 RB	6（12%）
细支气管炎；伴有 Favor 细支气管炎的细支气管炎	1（2%）
肉样瘤病	3（6%）
NISP；细胞 NISP；Favor NISP；伴有 Favor HP 的细胞 NSIP；伴有 Favor NISP 的 NISP；伴有 CIF、NOC 的 Favor NISP	3（6%）
过敏性肺炎；Favor HP	4（8%）
DAD；伴有血黄素蛋白沉着症的 DAD	2（4%）
嗜酸性粒细胞性肺炎	1（2%）
机化性肺泡出血	1（2%）
外源性脂质性肺炎	1（2%）
淀粉样或轻链沉积	1（2%）
肺气肿；伴有潜在感染的肺气肿；伴有 RB 的肺气肿	1（2%）
非 UIP 总计	25（51%）
总计	49

*在肺叶上有不同诊断的受试者被记为“伴有”。

Envisia基因组分类器表现

[0356] 用于UIP分子诊断的Envisia基因组分类器在验证组中实现88％的高特异性[CI：68％-97％]和67％的中等灵敏度[CI:45％-84％]。ROC-AUC为0.85(图10)。当分析仅限于具有完全来自外科肺活检的病理学的26名受试者或具有来自冷冻活检的病理学的21名受试者时，测试表现保持在置信区间内(数据未显示)。在21名冷冻活检受试者中，5名来自单一欧洲研究场所，1名、7名和8名受试者分别来自3个美国研究场所。

[0357] 在检查由分类器产生的错误时，25名具有非UIP病理学的受试者中的三名被分类为分子UIP(FP)(图11)。一例FP患有晚期小气道疾病和滤泡性细支气管炎的病理模式诊断，但研究场所得到的临床诊断为很可能是IPF。第二例FP最初通过中心放射学诊断为HP并通过中心病理学诊断为细胞NSIP，但是在长期随访中通过HRCT显示出致密的纤维化。通过放射学诊断并且在组织病理学上记录到表现出淀粉样或轻链沉积的具有严重肺气肿的NSIP病例也称为分子UIP(表8)。

表8:与三名Envisia基因组分类器假阳性受试者相关的临床因素

FP 受试者	Envisia 判定	放射学 局部 IPF	放射学 中心 其他(UIP)	放射学 局部长期随访 N/A	病理学 局部	病理学 中心	局部临床诊断 初始诊断	局部临床诊断 更新的诊断
1	UIP	双侧肺底浸润物无蜂窝状	HP	N/A	CIF,NOC	SLB: 细支气管炎(肺上叶) 细支气管炎(肺下叶)	细胞 NSIP	致密纤维化
2	UIP	肺纤维化	其他(肺气肿)	严重	UIP	SLB: NSIP(肺下叶) 冷冻活检: 其他-淀粉样或轻链沉积(肺上叶)	NSIP	NSIP
3	UIP							

[0358] 通过Envisia测试,将8名患有病理性UIP的受试者分类为分子非UIP(FN)(图11)。虽然这些受试者的最终参考标签是UIP,但这些病例中有一半通过研究场所的病理学、放射

学或临床诊断的非UIP诊断。研究场所诊断包括HP、RB、NSIP/DIP(后来更新为SRIF)和无法分类的ILD(表9)。其余4例的研究场所诊断为IPF,其中2例的HRCT模式诊断为UIP,1例为NSIP,并且1例为与可能的潜在自身免疫性疾病相关的HP(表9)。

表9:与八名Envisia基因组分类器FN受试者相关的临床细节

		放射学			病理学		局部临床诊断	
FN 受试者	Envisia 判定	局部 非特异性 肺泡炎	中心 机化性 肺炎	局部长期随访 SLB相关的 变化	局部 NSIP	中心 SLB: 典型 UIP（肺中叶） 典型 UIP（肺下叶） 冷冻活检: UIP（肺下叶） SLB: UIP（肺上叶） UIP（肺中叶） 典型 UIP（肺下叶） 冷冻活检: Favor UIP（肺上叶） UIP（肺中叶） 无诊断（肺下叶） SLB: UIP（肺上叶） UIP（肺下叶） SLB: 典型 UIP（肺上叶） SLB: UIP（肺上叶） 典型 UIP（肺中叶） 典型 UIP（肺下叶） 冷冻活检: 非诊断性（肺上叶） 非诊断性（肺中叶） Favor UIP（肺下叶）	初始诊断 NSIP/DIP	更新的诊断 SRIF
1	非 UIP							
2	非 UIP	可能 UIP	NSIP	N/A	UIP		可能 IPF	N/A
3	非 UIP	双侧肺 纤维化	HP	N/A	UIP		可能 IPF	UIP + NSIP 伴有 自身免疫 疾病 （可能 RA）
4	非 UIP	UIP	明确 UIP	N/A	UIP		明确 IPF	N/A
5	非 UIP	NSIP	HP	N/A	慢性间质性纤维化，未另外分类		无法分类 ILD	N/A
6	非 UIP	IPF	很可能 UIP	UIP	UIP		明确 IPF	明确 IPF
7	非 UIP	慢性 ILD 的 急性加重	HP	N/A	慢性 HP		HP	N/A
8	非 UIP	UIP	N/A	N/A	SRIF		RB	N/A

[0359] 与指南一致，使用HRCT评价可疑ILD患者，目标是评估UIP模式的存在或不存在(例如“HRCT-UIP”)。在没有HRCT确定的UIP诊断的情况下，应考虑对患者进行SLB以获得UIP或

非UIP的组织病理学模式诊断[5]。为了建立Envisia分子UIP判定的表现基线，我们使用组织病理学UIP作为参考标准评价了HRCT-UIP的预测值。我们检查了来自专家评审(D.Lynch)的HRCT模式诊断以及研究场所的模式诊断。在最终验证集中，中心放射学显示出完美的特异性和阳性预测值(PPV)，以及边际灵敏度(图12)。这与先前针对专家HRCT-UIP的高特异性低灵敏度的报告一致[17]。该组患者的局部放射学特异性和PPV(分别为70％和67％)显著低于专家中心评审(图12)。

[0360] 具有中心HRCT-UIP的受试者中的分子UIP的PPV是100％，类似于专家放射学的整体PPV(相对于病理学)，但远远优于研究场所HRCT-UIP的总体PPV(图12)。在研究部位HRCT-UIP的受试者中，分子UIP的PPV降低至73％，但在中心HRCT-UIP病例中分子UIP仍为100％(图12)。有趣的是，分子UIP判定在研究场所放射学诊断为与UIP不一致的受试者中高度准确，显示PPV为100％，NPV为89％，类似于中心放射学观察到的100％PPV(图12)。此外，分子UIP显示出比专家放射学更高的灵敏度——67％比41％。在特定中心放射学诊断为HP的15名受试者中，9名具有UIP组织病理学模式(表10)。分子UIP在具有UIP组织病理学模式的9名HP患者中的6名中正确鉴定了组织病理学UIP(表10)，表明通过Envisia测试的分子诊断可以帮助鉴定HP患者中组织病理学UIP的存在。

表10:对于15名经中心放射学诊断为过敏性肺炎的受试者，Envisia相对于病理学的表现

受试者	病理学	放射学	Envisia 判定
1	UIP	HP	UIP
2	NSIP	HP	UIP
3	UIP	HP	UIP
4	UIP	HP	UIP
5	UIP	HP	UIP
6	UIP	HP	UIP
7	UIP	HP	UIP
8	UIP	HP	非 UIP
9	DAD	HP	非 UIP
10	UIP	HP	非 UIP
11	UIP	HP	非 UIP
12	DAD	HP	非 UIP
13	SRIF	HP	非 UIP
14	HP	HP	非 UIP
15	HP	HP	非 UIP

[0361] 实现诊断病理学的公认挑战意味着可能无法确定19名受试者的UIP或非UIP标签(表11)。可能在临床中遇到与这些受试者相似的患者，因此可能由Envisia进行测试。因此，我们将Envisia测试结果与关于这些受试者的可用临床信息进行了比较。在由Envisia进行分子UIP的6名受试者中，有两名具有HRCT-UIP模式，两名具有IPF临床诊断(表11)。在通过Envisia确定为分子非UIP的13名受试者中，7名具有HRCT非UIP模式；其中三人的临床诊断为非UIP病况(表11)。

表11:具有非诊断性病理学或无法分类纤维化的受试者的Envisia分类(二级分析组)。

受试者	Envisia 评分	Envisia 判定	中心病理学	中心放射学	局部临床诊断	局部病理学诊断
1	-2.05	**非 UIP**	CIF，NOC	**其他-吸入性肺炎**	**细支气管炎**	
2	-1.43	**非 UIP**	非诊断性	**HP**	其他	
3	-0.91	**非 UIP**	非诊断性	**其他-吸入性肺炎**	其他	其他
4	-0.49	**非 UIP**	非诊断性	**石棉肺**		
5	-0.39	非 UIP	CIF、NOC			
6	-0.25	**非 UIP**	非诊断性	**其他-吸入性肺炎**	其他	
7	0.13	非 UIP	非诊断性			
8	0.15	**非 UIP**	CIF、NOC	**HP**	**混合 NISP**	
9	0.17	非 UIP	CIF、NOC	明确 UIP	混合 NISP	
10	0.36	非 UIP	非诊断性	很可能 UIP	其他	其他
11	0.56	**非 UIP**	非诊断性	**HP**	**HP**	非诊断性
12	0.74	非 UIP	非诊断性		很可能 IPF	其他
13	0.87	非 UIP	CIF、NOC	明确 UIP	很可能 IPF	CIF，NOC
14	0.98	UIP	非诊断性	RB	其他	非诊断性
15	0.98	**UIP**	CIF、NOC	HP	**很可能 IPF**	
16	1.26	UIP	非诊断性			其他
17	1.39	**UIP**	非诊断性	**可能 UIP**	Favor NSIP	
18	1.42	**UIP**	非诊断性	**明确 UIP**		
19	2.59	**UIP**	非诊断性		**明确 IPF**	

[0362] 在Envisia基因组分类器所使用的190个基因中，124个在UIP TBB与非UIP TBB之间差异表达的前1000个基因中。在UIP中上调的分类器特征和基因富集有四种生物学途径的成员，其中三种先前已使用微阵列基因表达平台在SLB中鉴定[13](表12)。

表12:389个在TBB中的UIP中上调的基因和92个Envisia基因组分类器基因的途径富集分析(55个基因是两个集所共有的)。字体加粗的途径在外科肺活检物中是显著富集的(Kim SY等人,2015)。

类别	预期的基因数目	观察到的基因数目	P-值（经校正）	富集方向
扩张型心肌病	1.9	11	0.000126	途径基因超出预期
肥厚型心肌病（HCM）	1.7	9	0.003087	途径基因超出预期
黏着斑	4.1	13	0.012032	途径基因超出预期
神经活性配体-受体相互作用	5.5	15	0.021914	途径基因超出预期

[0363] KEGG扩张和肥大性心肌病网络包括参与细胞外基质相互作用、生长因子应答和细胞骨架重塑的基因，据报道所有这些都在IPF中被上调[18,19]。

[0364] 类似地，在非UIP TBB中上调的特征和基因富集有在非UIP SLB中也上调的多种途径，包括免疫应答、细胞-细胞信号传导和发育途径(表13)。与IPF相比，HP显示出细胞增殖、免疫应答基因的差异上调[20]，尽管一些基因在这些疾病中是共同调节的[21]。

表13:611个在TBB中的非UIP中上调的基因和98个Envisia基因组分类器基因的途径富集分析(69个基因是两个集所共有的)。字体加粗的途径在外科肺活检物中是显著富集的(Kim SY等人,2015)。

类别	预期的基因数目	观察到的基因数目	P-值（经校正）	富集方向
抗原处理和呈递	3.6	22	4.05E-10	途径基因超出预期
利什曼病	3.4	20	7.47E-09	途径基因超出预期
移植物抗宿主病	2.0	15	4.75E-08	途径基因超出预期
I 型糖尿病	2.1	15	9.94E-08	途径基因超出预期
同种异体移植物排斥	1.8	14	1.22E-07	途径基因超出预期
病毒性心肌炎	3.5	18	5.90E-07	途径基因超出预期
Toll 样受体信号传导途径	4.8	21	7.52E-07	途径基因超出预期
自身免疫性甲状腺疾病	2.5	14	1.39E-05	途径基因超出预期
吞噬体	7.4	23	0.000118	途径基因超出预期
嗅觉转导	18.1	2	0.000145	途径基因低于预期
细胞因子-细胞因子受体相互作用	12.4	29	0.001704	途径基因超出预期
查加斯病	4.8	16	0.002929	途径基因超出预期
细胞粘附分子（CAM）	6.3	18	0.006574	途径基因超出预期
NOD 样受体信号传导途径	2.9	11	0.015921	途径基因超出预期
趋化因子信号传导途径	8.8	21	0.023401	途径基因超出预期

讨论

[0365] 临床背景和在胸部的HRCT扫描中看到的放射学模式的组合未能对接受ILD评估的患者提供确信的诊断是常见的。虽然来自SLB的组织病理学模式诊断可以对这些患者提供确定的诊断，但是许多患者不愿意或因病得太重而不能进行外科诊断程序。即使在这样做的情况下，与活检结果的病理解释相关的挑战也可能留下重大的临床不确定性。与最小风险相关并且不依赖于有经验的肺病理学家的视觉和主观技能来确认组织学UIP的存在的准确且可用的测试可能是非常有用的。

[0366] 在积累有意义数目的患者和样品以支持开发Envisia所需的机器学习努力中的重大挑战反映了临床医生在调查新诊断的ILD患者时所面临的挑战。在通过我们的BRAVE样品采集研究所累计的201名受试者中，尽管使用了一组专家肺病理学家，我们仅确定了140名具有诊断性组织病理学结果的受试者。这种低产率突出了社区临床医生所面临的实现诊断病理学的挑战。我们使用前90名受试者训练并锁定了Envisia基因组分类器，并使用随后累积的受试者验证了该测试。这种针对传统TBB中的UIP的基因组分类器在两个队列中均显示出高表现。

[0367] 具有未被研究场所放射科医师预测的组织学UIP模式的25名受试者中的分子UIP判定的准确性较高，其中成功鉴定了78％的具有UIP组织病理学模式的受试者，没有假阳性。在这组受试者中，Envisia作为一项真正的划入(rule-in)测试，其可以在5个无法通过HRCT鉴定的UIP组织病理学模式病例中还原近4个。此外，该亚组富集有HP患者，其中60％(9/15)在本研究中存在晚期纤维化疾病的证据。Envisia在HP患者中以67％的灵敏度检测到UIP，该灵敏度与总体上在49名受试者的验证队列中检测到UIP的灵敏度相同。在放射学模式诊断为与UIP不一致的受试者中，Envisia分子UIP判定的NPV>80％，这表明对于阳性和阴性的Envisia测试结果都具有实质性效用。

实施例11

样品临床和技术因素以及Envisia表现

[0368] Envisia测试表现显示出与受试者临床和样品技术因素的一些相关性。在男性受

试者和有吸烟史的受试者中,UIP疾病被漏判的概率较高(图13)。与肺泡II型细胞一致的基因表达与Envisia测试准确性不大相关(图14),表明肺泡采样对于测试表现并不重要,这与先前在90个ILD受试者队列中观察到的一致[E10]。在非UIP样品中,更强(更负)的分类评分与由样品大小和RNA质量定义的更高样品质量之间存在轻微的相关性(图14),而这在UIP样品中不明显。

[0369] 可以组合上述各个实施方案以提供进一步的实施方案。本说明书中提及和/或在申请数据表中列出的所有美国专利、美国专利申请公开、美国专利申请、外国专利、外国专利申请和非专利出版物均通过引用整体并入本文。如果需要,可以修改实施方案的各方面以采用各个专利、申请和出版物的概念来提供进一步的实施方案。

[0370] 根据以上详细描述,可以对实施方案进行这些和其他改变。通常,在以下权利要求中,所使用的术语不应被解释为将权利要求限制于说明书和权利要求中公开的特定实施方案,而是应该被解释为包括所有可能的实施方案,以及这些权利要求享有的实施方案的等同方案的全部范围。因此,权利要求不受本公开内容的限制。

[0371] 本文描述的一些实施方案涉及具有非暂时性计算机可读介质(也可以称为非暂时性处理器可读介质)的计算机存储产品,其上具有用于执行各种计算机实现的操作的指令或计算机代码。计算机可读介质(或处理器可读介质)是非暂时性的,这意味着其不包括暂时传播信号本身(例如,在诸如空间或电缆等传输介质上携带信息的传播电磁波)。介质和计算机代码(也可以称为代码)可以是为特定目的而设计和构造的代码。非暂时性计算机可读介质的实例包括但不限于:磁存储介质,如硬盘、软盘和磁带等;光盘存储介质,如光碟/数字影碟(CD/DVD)、光盘只读存储器(CD-ROM)和全息设备;磁光存储介质,如光盘;载波信号处理模块;专门配置用于存储和执行程序代码的硬件设备,如专用集成电路(ASIC)、可编程逻辑器件(PLD)、只读存储器(ROM)和随机存取存储器(RAM)器件。本文描述的其他实施方案涉及计算机程序产品,其可以包括例如本文所讨论的指令和/或计算机代码。

[0372] 本文描述的一些实施方式和/或方法可以由软件(在硬件上执行)、硬件或其组合来执行。硬件模块可以包括例如通用处理器、现场可编程门阵列(FPGA)和/或专用集成电路(ASIC)。软件模块(在硬件上执行)可以以多种软件语言(例如,计算机代码)表示,包括C、C++、Java™、Ruby、Visual Basic™、R和/或其他面向对象的、程序性的、统计的或其他编程语言和开发工具。计算机代码的实例包括但不限于微代码或微指令、如由编译器产生的机器指令、用于产生web服务的代码,以及包含由采用解释程序的计算机所执行的更高级指令的文件。例如,可以使用命令式编程语言(例如,C、FORTRAN等)、函数编程语言(例如,Haskell、Erlang等)、逻辑编程语言(例如,Prolog)、面向对象的编程语言(例如,Java、C++等)、统计编程语言和/或环境(例如,R等)或其他合适的编程语言和/或开发工具来实现实施方案。计算机代码的附加实例包括但不限于控制信号、加密代码和压缩代码。

参考文献

所有以下参考文献和本文所引用的所有参考文献以其整体并入本文。

1.Travis WD,Costabel U,Hansell DM,King TE,Lynch DA,Nicholson AG,Ryerson CJ,Ryu JH,Selman M,Wells AU,Behr J,Bouros D,Brown KK,Colby TV,Collard HR,Cordeiro CR,Cottin V,Crestani B,Drent M,Dudden RF,Egan J,Flaherty K,Hogaboam C,Inoue Y,Johkoh T,Kim DS,Kitaichi M,Loyd J,Martinez FJ,Myers J,Protzko S,

Raghu G,Richeldi L,Sverzellati N,Swigris J,Valeyre D.An Official American Thoracic Society/European Respiratory Society Statement:Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias.Am J Respir Crit Care Med 2013;188:733-748.

2.Raghu G,Rochwerg B,Zhang Y,Garcia CAC,Azuma A,Behr J,Brozek JL,Collard HR,Cunningham W,Homma S,Johkoh T,Martinez FJ,Myers J,Protzko SL,Richeldi L,Rind D,Selman M,Theodore A,Wells AU,Hoogsteden H,Schünemann HJ.An Official ATS/ERS/JRS/ALAT Clinical Practice Guideline:Treatment of idiopathic pulmonary fibrosis.An update of the 2011clinical practice guideline.Am J Respir Crit Care Med 2015;192:e3-e19.

3.Bjoraker JA,Ryu JH,Edwin MK,Myers JL,Tazelaar HD,Schroeder DR,Offord KP.Prognostic significance of histopathologic subsets in idiopathic pulmonary fibrosis.Am J Respir Crit Care Med 1998;157:199-203.

4.Flaherty KR,Travis WD,Colby TV,Toews GB,Kazerooni EA,Gross BH,Jain A,Strawderman RL,Flint A,Lynch JP,Martinez FJ.Histopathologic variability in usual and nonspecific interstitial pneumonias.Am J Respir Crit Care Med 2001;164:1722-1727.

5.Flaherty KR,Toews GB,Travis WD,Colby TV,Kazerooni EA,Gross BH,Jain A,Strawderman RL,Paine R,Flint A,Lynch JP,Martinez FJ.Clinical significance of histological classification of idiopathic interstitial pneumonia.Eur Respir J 2002;19:275-283.

6.Flaherty K,Thwaite E,Kazerooni E,Gross B,Toews G,Colby T,Travis W,Mumford J,Murray S,Flint A,Lynch J,Martinez F.Radiological versus histological diagnosis in UIP and NSIP:Survival implications.Thorax 2003;58:143-148.

7.Katzenstein A-LA,Mukhopadhyay S,Myers JL.Diagnosis of usual interstitial pneumonia and distinction from other fibrosing interstitial lung diseases.Hum Pathol 2008;39:1275-1294.

8.Raghu G,Collard HR,Egan JJ,Martinez FJ,Behr J,Brown KK,Colby TV,Cordier J-F,Flaherty KR,Lasky JA,Lynch DA,Ryu JH,Swigris JJ,Wells AU,Ancochea J,Bouros D,Carvalho C,Costabel U,Ebina M,Hansell DM,Johkoh T,Kim DS,King TE,Kondoh Y,Myers J,Müller NL,Nicholson AG,Richeldi L,Selman M,Dudden RF,Griss BS,Protzko SL,Schünemann HJ.An Official ATS/ERS/JRS/ALAT Statement:Idiopathic Pulmonary Fibrosis:Evidence-based guidelines for diagnosis and management.Am J Respir Crit Care Med2011;183:788-824.

9.American Thoracic S,European Respiratory S.American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias.This joint statement of the American Thoracic Society(ATS)and the European Respiratory Society

(ERS)was adopted by the ATS board of directors,June 2001 and by the ERS Executive Committee,June 2001.Am J Respir Crit Care Med 2002;165:277-304.10.Katzenstein A-LA,Myers JL.Idiopathic Pulmonary Fibrosis.Am J Respir Crit Care Med1998;157:1301-1315.

11.Berbescu EA,Katzenstein A-LA,Snow JL,Zisman DA.Transbronchial biopsy in usual interstitial pneumonia.Chest 2006;129:1126-1131.

12.Tomassetti S,Cavazza A,Colby TV,Ryu JH,Nanni O,Scarpi E,Tantalocco P,Buccioli M,Dubini A,Piciucchi S,Ravaglia C,Gurioli C,Casoni GL,Gurioli C,Romagnoli M,Poletti V.Transbronchial biopsy is useful in predicting UIP pattern.Respir Res 2012;13:96-96.

13.Shim HS,Park MS,Park IK.Histopathologic findings of transbronchial biopsy in usual interstitial pneumonia.Pathol Int 2010;60:373-377.

14.Tomassetti S,Wells AU,Costabel U,Cavazza A,Colby TV,Rossi G,Sverzellati N,Carloni A,Carretta E,Buccioli M,Tantalocco P,Ravaglia C,Gurioli C,Dubini A,Piciucchi S,Ryu JH,Poletti V.Bronchoscopic lung cryobiopsy increases diagnostic confidence in the multidisciplinary diagnosis of idiopathic pulmonary fibrosis.Am J Respir Crit Care Med2016;193:745-752.

15.Dhooria S,Sehgal IS,Aggarwal AN,Behera D,Agarwal R.Diagnostic yield and safety of cryoprobe transbronchial lung biopsy in diffuse parenchymal lung diseases:Systematic review and meta-analysis.Respir Care 2016;61:700-712.

16.Poletti V,Ravaglia C,Gurioli C,Piciucchi S,Dubini A,Cavazza A,Chilosi M,Rossi A,Tomassetti S.Invasive diagnostic techniques in idiopathic interstitial pneumonias.Respirology2016;21:44-50.

17.Kim SY,Diggans J,Pankratz D,Huang J,Pagan M,Sindy N,Tom E,Anderson J,Choi Y,Lynch DA,Steele MP,Flaherty KR,Brown KK,Farah H,Bukstein MJ,Pardo A,Selman M,Wolters PJ,Nathan SD,Colby TV,Myers JL,Katzenstein A-LA,Raghu G,Kennedy GC.Classification of usual interstitial pneumonia in patients with interstitial lung disease:Assessment of a machine learning approach using high-dimensional transcriptional data.Lancet Respir Med 2015;3:473-482.

18.Friedman J,Hastie T,Tibshirani R.Regularization paths for generalized linear models via coordinate descent.J Stat Softw 2010;33:1-22.

19.Love MI,Huber W,Anders S.Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.Genome Biol 2014;15:550.

20.Mi H,Lazareva-Ulitsky B,Loo R,Kejariwal A,Vandergriff J,Rabkin S,Guo N,Muruganujan A,Doremieux O,Campbell MJ,Kitano H,Thomas PD.The PANTHER database of protein families,subfamilies,functions and pathways.Nucleic Acids Res2005;33:D284-D288.

21.Katzenstein A-LA,Zisman DA,Litzky LA,Nguyen BT,Kotloff RM.Usual

interstitial pneumonia:Histologic study of biopsy and explant specimens.Am J Surg Pathol2002;26:1567-1577.

22.Trahan S,Hanak V,Ryu JH,Myers JL.Role of surgical lung biopsy in separating chronic hypersensitivity pneumonia from usual interstitial pneumonia/idiopathic pulmonary fibrosis:Analysis of 31 biopsies from 15 patients.Chest 2008;134:126-132.

23.Akashi T,Takemura T,Ando N,Eishi Y,Kitagawa M,Takizawa T,Koike M,Ohtani Y,Miyazaki Y,Inase N,Yoshizawa Y.Histopathologic analysis of sixteen autopsy cases of chronic hypersensitivity pneumonitis and comparison with idiopathic pulmonary fibrosis/usual interstitial pneumonia.Am J Clin Pathol 2009;131:405-415.

24.Selman M,Pardo A,Barrera L,Estrada A,Watson SR,Wilson K,Aziz N,Kaminski N,Zlotnik A.Gene expression profiles distinguish idiopathic pulmonary fibrosis from hypersensitivity pneumonitis.Am J Respir Crit Care Med 2006;173:188-198.

25.Lockstone HE,Sanderson S,Kulakova N,Baban D,Leonard A,Kok WL,McGowan S,McMichael AJ,Ho LP.Gene set analysis of lung samples provides insight into pathogenesis of progressive,fibrotic pulmonary sarcoidosis.Am J Respir Crit Care Med 2010;181:1367-1375.

26.Selman M,Pardo A.Revealing the pathogenic and aging-related mechanisms of the enigmatic idiopathic pulmonary fibrosis.An integral model.Am J Respir Crit Care Med2014;189:1161-1172.

27.Bauer Y,Tedrow J,de Bernard S,Birker-Robaczewska M,Gibson KF,Guardela BJ,Hess P,Klenk A,Lindell KO,Poirey S,Renault B,Rey M,Weber E,Nayler O,Kaminski N.A novel genomic signature with translational significance for human idiopathic pulmonary fibrosis.Am J Respir Cell Mol Biol 2015;52:217-231.

28.Jonigk D,Izykowski N,Rische J,Braubach P,Kuhnel M,Warnecke G,Lippmann T,Kreipe H,Haverich A,Welte T,Gottlieb J,Laenger F.Molecular profiling in lung biopsies of human pulmonary allografts to predict chronic lung allograft dysfunction.Am J Pathol2015;185:3178-3188.

29.Nicholson AG,Fulford LG,Colby TV,du Bois RM,Hansell DM,Wells AU.The relationship between individual histologic features and disease progression in idiopathic pulmonary fibrosis.Am J Respir Crit Care Med 2002;166:173-177.

30.Walsh SL,Wells AU,Desai SR,Poletti V,Piciucchi S,Dubini A,Nunes H,Valeyre D,Brillet PY,Kambouchner M,Morais A,Pereira JM,Moura CS,Grutters JC,van den Heuvel DA,van Es HW,van Oosterhout MF,Seldenrijk CA,Bendstrup E,Rasmussen F,Madsen LB,Gooptu B,Pomplun S,Taniguchi H,Fukuoka J,Johkoh T,Nicholson AG,Sayer C,Edmunds L,Jacob J,Kokosi MA,Myers JL,Flaherty KR,Hansell

DM.Multicentre evaluation of multidisciplinary team meeting agreement on diagnosis in diffuse parenchymal lung disease:A case-cohort study.Lancet Respir Med 2016;4:557-565.

31.Flaherty KR,King TE,Raghu G,Lynch JP,Colby TV,Travis WD,Gross BH,Kazerooni EA,Toews GB,Long Q,Murray S,Lama VN,Gay SE,Martinez FJ.Idiopathic Interstitial Pneumonia.Am J Respir Crit Care Med 2004;170:904-910.

32.Tominaga J,Sakai F,Johkoh T,Noma S,Akira M,Fujimoto K,Colby TV,Ogura T,Inoue Y,Taniguchi H,Homma S,Taguchi Y,Sugiyama Y.Diagnostic certainty of idiopathic pulmonary fibrosis/usual interstitial pneumonia:The effect of the integrated clinico-radiological assessment.Eur J Radiol 2015;84:2640-2645.

33.The Idiopathic Pulmonary Fibrosis Clinical Research Network.Prednisone,azathioprine,and n-acetylcysteine for pulmonary fibrosis.N Engl J Med 2012;366:1968-77.

34.Sumikawa H,Johkoh T,Colby TV,Ichikado K,Suga M,Taniguchi H,Kondoh Y,Ogura T,Arakawa H,Fujimoto K,Inoue A,Mihara N,Honda O,Tomiyama N,Nakamura H,Muller NL.Computed tomography findings in pathological usual interstitial pneumonia.Am J Respir Crit Care Med 2008;177:433-439.

35.Chung JH,Chawla A,Peljto AL,Cool CD,Groshong SD,Talbert JL,McKean DF,Brown KK,Fingerlin TE,Schwarz MI,Schwarz DA,Lynch DA.CT scan findings of probable usual interstitial pneumonitis have a high predictive value for histologic usual interstitial pneumonitis.Chest 2015;147:450-459.

36.Brownell R,Moua T,Henry TS,Elicker BM,White D,Vittinghoff E,Jones KD,Urisman A,Aravena C,Johannson KA,Golden JA,King TE Jr,Wolters PJ.Collard HR,Ley B.The use of pretest probability increases the value of high-resolution CT in diagnosing usual interstitial pneumonia.Thorax 2017;72(5):424-429.

37.DiBardino DM,Haas AR,Lanfranco AR,Litzky LA,Sterman D,Bessich JL.High complication rate after introduction of transbronchial cryobiopsy into clinical practice at an academic medical center.Annals Am Thorac Soc 2017;14(6):851-857.

38.Hutchinson JP,McKeever TM,Fogarty AW,Navaratnam V,Hubbard RB.Surgical lung biopsy for the diagnosis of interstitial lung disease in England:1997-2008.Eur Respir J2016;48:1453-61.

E1.Kim SY,Diggans J,Pankratz D,Huang J,Pagan M,Sindy N,Tom E,Anderson J,Choi Y,Lynch DA,Steele MP,Flaherty KR,Brown KK,Farah H,Bukstein MJ,Pardo A,Selman M,Wolters PJ,Nathan SD,Colby TV,Myers JL,Katzenstein A-LA,Raghu G,Kennedy GC.Classification of usual interstitial pneumonia in patients with interstitial lung disease:assessment of a machine learning approach using high-dimensional transcriptional data.Lancet Respir Med 2015;3:473-482.

E2.Dobin A,Davis CA,Schlesinger F,Drenkow J,Zaleski C,Jha S,Batut P,

Chaisson M,Gingeras TR.STAR:ultrafast universal RNA-seq aligner.Bioinformatics 2012.

E3.Anders S,Pyl PT,Huber W.HTSeq—a Python framework to work with high-throughput sequencing data.Bioinformatics 2015;31:166-169.

E4.DeLuca DS,Levin JZ,Sivachenko A,Fennell T,Nazaire M-D,Williams C,Reich M,Winckler W,Getz G.RNA-SeQC:RNA-seq metrics for quality control and process optimization.Bioinformatics 2012;28:1530-1532.

E5.Wuenschell CW,Sunday ME,Singh G,Minoo P,Slavkin HC,Warburton D.Embryonic mouse lung epithelial progenitor cells co-express immunohistochemical markers of diverse mature cell lineages.J Histochem Cytochem 1996;44:113-123.

E6.Nielsen S,King LS,Christensen BM,Agre P.Aquaporins in complex tissues.II.Subcellular distribution in respiratory and glandular tissues of rat.Am J Physiol1997;273:C1549-1561.

E7.Kim CF,Jackson EL,Woolfenden AE,Lawrence S,Babar I,Vogel S,Crowley D,Bronson RT,Jacks T.Identification of bronchioalveolar stem cells in normal lung and lung cancer.Cell2005;121:823-835.

E8.Zemke AC,Snyder JC,Brockway BL,Drake JA,Reynolds SD,Kaminski N,Stripp BR.Molecular staging of epithelial maturation using secretory cell-specific genes as markers.Am J Respir Cell Mol Biol 2009;40:340-348.

E9.Treutlein B,Brownfield DG,Wu AR,Neff NF,Mantalas GL,Espinoza FH,Desai TJ,Krasnow MA,Quake SR.Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq.Nature 2014;509:371-375.

E10.Pankratz DG,Choi Y,Imtiaz U,Fedorowicz GM,Anderson JD,Colby TV,Myers JL,Lynch DA,Brown KK,Flaherty KR,Steele MP,Groshong SD,Raghu G,Barth NM,Walsh PS,Huang J,Kennedy GC,Martinez FJ.Usual interstital pneumonia can be detected in transbronchial biopsies using machine learning.Annals Am Thorac Soc 2017.

[0373] 本说明书中提及和/或在申请数据表或上述参考文献列表中列出的所有上述美国专利、美国专利申请公开、美国专利申请、外国专利、外国专利申请和非专利出版物通过引用以其整体并入本文。

表14.本研究中评价的113名患者及其相关的样品

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
01-202	A	上	74	F	是	否	24	HP	困难 UIP	困难 UIP	UIP	训练	0.84	N/A	1.91	N/A
01-202	B	上								困难 UIP	UIP	训练	1.03	N/A	1.97	N/A
02-101	A	上								UIP	UIP	训练	0.61	N/A	0.08	N/A
02-101	B	上	76	F	是	否	45	HP	UIP	UIP	UIP	训练	-0.10	N/A	-1.38	N/A
02-101	C	下								Favor UIP	UIP	训练	-0.68	N/A	1.40	N/A
02-101	E	下								Favor UIP	UIP	训练	1.33	N/A	1.96	N/A
02-102	A	上								UIP	UIP	训练	0.22	N/A	-0.29	N/A
02-102	C	下	74	F	是	否	24	HP	UIP	UIP	UIP	训练	-0.80	N/A	-0.84	N/A
02-102	D	下								UIP	UIP	训练	2.41	N/A	-0.01	N/A
02-102	E	下								UIP	UIP	训练	1.29	N/A	0.21	N/A
02-103	C	下	59	F	否	N/A	N/A	丢失	Favor HP	Favor HP	非 UIP	训练	0.44	N/A	-2.73	N/A
02-103	E	下								Favor HP	非 UIP	训练	-0.95	N/A	-3.16	N/A
02-104	A	上								肺气肿	非 UIP	训练	0.86	N/A	2.46	N/A
02-104	B	上	53	F	是	否	56	其他	肉芽肿病	肺气肿	非 UIP	训练	-0.11	N/A	0.16	N/A
02-104	C	下								肺气肿	非 UIP	训练	-0.88	N/A	-0.10	N/A
02-104	D	下								肺气肿	非 UIP	训练	-3.24	N/A	-4.21	N/A
03-102	A	上								典型 UIP	UIP	训练	1.27	N/A	5.12	N/A
03-102	B	上	65	F	是	No	20	HP	UIP	典型 UIP	UIP	训练	2.49	N/A	6.50	N/A
03-102	C	下								困难 UIP	UIP	训练	1.59	N/A	4.59	N/A
03-102	D	下								困难 UIP	UIP	训练	0.61	N/A	3.18	N/A
05-101	A	上								Favor UIP	UIP	训练	4.89	N/A	1.46	N/A
05-101	B	上								Favor UIP	UIP	训练	2.54	N/A	1.39	N/A
05-101	C	下	54	F	是	否	60	HP	Favor UIP	UIP	UIP	训练	2.76	N/A	1.23	N/A
05-101	D	下								UIP	UIP	训练	2.43	N/A	1.15	N/A
05-101	E	下								UIP	UIP	训练	2.30	N/A	0.04	N/A
05-102	A	上								UIP	UIP	训练	0.07	N/A	-0.06	N/A
05-102	B	上								UIP	UIP	训练	0.17	N/A	-0.57	N/A
05-102	C	下	68	M	是	否	100	明确 UIP	UIP	困难 UIP	UIP	训练	2.20	N/A	1.29	N/A
05-102	D	下								困难 UIP	UIP	训练	1.68	N/A	1.47	N/A
05-102	E	下								困难 UIP	UIP	训练	-1.91	N/A	-1.82	N/A
05-103	A	上								HP	非 UIP	训练	2.02	N/A	3.43	N/A
05-103	B	上								HP	非 UIP	训练	-0.35	N/A	-0.09	N/A
05-103	C	下	37	M	否	N/A	N/A	HP	HP	HP	非 UIP	训练	1.10	N/A	2.35	N/A
05-103	D	下								HP	非 UIP	训练	-0.39	N/A	1.33	N/A
05-103	E	下								HP	非 UIP	训练	-0.56	N/A	1.98	N/A
06-301	A	上	70	F	否	N/A	N/A	NSIP	无诊断	OP	非 UIP	训练	4.11	N/A	1.80	N/A
06-301	B	上								OP	非 UIP	训练	0.86	N/A	0.89	N/A
06-316	C	下								RB	非 UIP	训练	0.12	N/A	-4.05	N/A
06-316	D	下	62	F	是	是	60	RB	无诊断	RB	非 UIP	训练	-0.36	N/A	-0.54	N/A
06-316	E	下								RB	非 UIP	训练	-1.65	N/A	1.71	N/A
08-101	C	下	60	M	是	否	36	嗜酸性粒细胞性肺炎	DAD	RB	非 UIP	训练	-0.02	N/A	-0.72	N/A
08-101	E	下								RB	非 UIP	训练	-3.38	N/A	-7.17	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
08-102	C	下	76	F	是	否	45	嗜酸性粒细胞性肺炎	细胞 NSIP	细胞 NSIP	非UIP	训练	-1.05	N/A	-3.32	N/A
08-102	D	下								细胞 NSIP	非UIP	训练	0.76	N/A	-1.58	N/A
08-103	A	上								困难 UIP	UIP	训练	2.87	N/A	1.99	N/A
08-103	B	上	78	M	是	否	60	NSIP	典型 UIP	困难 UIP	UIP	训练	3.52	N/A	6.41	N/A
08-103	E	下								困难 UIP	UIP	训练	-0.20	N/A	2.39	N/A
08-104	C	下	71	F	是	否	48	UIP	典型 UIP	典型 UIP	UIP	训练	1.25	N/A	0.68	N/A
08-104	E	下								典型 UIP	UIP	训练	2.54	N/A	1.78	N/A
08-106	A	上								UIP	UIP	训练	2.16	N/A	4.54	N/A
08-106	B	上	58	M	否	N/A	N/A	NSIP	UIP	UIP	UIP	训练	4.18	N/A	5.82	N/A
08-106	C	下								UIP	UIP	训练	7.93	N/A	9.65	N/A
08-106	D	下								UIP	UIP	训练	4.01	N/A	4.94	N/A
08-107	A	上								典型 UIP	UIP	训练	1.15	N/A	2.97	N/A
08-107	B	上	74	F	否	N/A	N/A	UIP	典型 UIP	典型 UIP	UIP	训练	2.93	N/A	4.22	N/A
08-107	C	下								典型 UIP	UIP	训练	3.96	N/A	3.36	N/A
08-107	D	下								典型 UIP	UIP	训练	4.14	N/A	5.46	N/A
08-108	A	中	71	M	是	否	74	嗜酸性粒细胞性肺炎	困难 UIP	UIP	UIP	训练	1.36	N/A	-3.17	N/A
08-108	B	中								UIP	UIP	训练	1.72	N/A	0.80	N/A
08-112	A	上								典型 UIP	UIP	训练	-0.78	N/A	-1.29	N/A
08-112	B	上	48	F	是	是	31	HP	典型 UIP	典型 UIP	UIP	训练	0.93	N/A	1.34	N/A
08-112	C	下								UIP	UIP	训练	-0.24	N/A	-1.37	N/A
08-112	E	下								UIP	UIP	训练	-2.10	N/A	-1.48	N/A
08-114	C	下	61	M	是	是	46	HP	困难 UIP	困难 UIP	UIP	训练	2.97	N/A	3.56	N/A
08-114	D	下								困难 UIP	UIP	训练	2.13	N/A	4.03	N/A
08-114	E	下								困难 UIP	UIP	训练	1.71	N/A	3.10	N/A
08-116	A	上								困难 UIP	UIP	训练	4.46	N/A	5.36	N/A
08-116	B	上								困难 UIP	UIP	训练	5.01	N/A	5.22	N/A
08-116	C	下	72	M	否	N/A	N/A	明确 UIP	困难 UIP	困难 UIP	UIP	训练	6.37	N/A	6.26	N/A
08-116	D	下								困难 UIP	UIP	训练	7.04	N/A	8.26	N/A
08-116	E	下								困难 UIP	UIP	训练	6.32	N/A	6.41	N/A
08-117	C	下	73	M	是	否	51	其他	CIF,NOC	UIP	UIP	训练	4.37	N/A	3.16	N/A
08-117	D	下								UIP	UIP	训练	5.01	N/A	4.28	N/A
08-117	E	下								UIP	UIP	训练	3.50	N/A	4.90	N/A
08-118	A	上								OP	非UIP	训练	0.26	N/A	-5.24	N/A
08-118	B	上								OP	非UIP	训练	-1.30	N/A	-0.15	N/A
08-118	C	下	69	F	是	否	Unk.	嗜酸性粒细胞性肺炎	OP	OP	非UIP	训练	-0.44	N/A	-0.35	N/A
08-118	D	下								OP	非UIP	训练	1.26	N/A	-0.15	N/A
08-118	E	下								OP	非UIP	训练	-0.62	N/A	-0.96	N/A
08-120	A	中								典型 UIP	UIP	训练	3.94	N/A	2.95	N/A
08-120	B	中	83	M	是	否	62	HP	典型 UIP	典型 UIP	UIP	训练	2.61	N/A	1.06	N/A
08-120	D	下								典型 UIP	UIP	训练	1.94	N/A	1.24	N/A
08-120	E	下								典型 UIP	UIP	训练	2.56	N/A	1.79	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	LOPO 评分	测试评分	LOPO 评分	测试评分
08-123	C	下								HP	非 UIP	训练	2.75	N/A	1.43	N/A
08-123	D	下	69	F	是	否	2	HP	HP	HP	非 UIP	训练	3.09	N/A	-0.81	N/A
08-123	E	下								HP	非 UIP	训练	3.15	N/A	-0.06	N/A
08-125	A	中								典型 UIP	UIP	训练	2.74	N/A	1.87	N/A
08-125	B	中								典型 UIP	UIP	训练	2.17	N/A	1.20	N/A
08-125	C	下	71	M	是	否	30	明确 UIP	典型 UIP	典型 UIP	UIP	训练	1.86	N/A	5.56	N/A
08-125	D	下								典型 UIP	UIP	训练	2.33	N/A	3.34	N/A
08-125	E	下								典型 UIP	UIP	训练	2.60	N/A	0.92	N/A
08-201	A	上								RB	非 UIP	训练	-1.38	N/A	-3.67	N/A
08-201	B	上	46	M	是	是	51	RB	无诊断	RB	非 UIP	训练	-2.50	N/A	-2.39	N/A
08-206	D	下								肺孢子虫肺炎	非 UIP	训练	-2.05	N/A	-1.11	N/A
08-206	E	下	53	M	是	是	Unk.	其他	肺孢子虫肺炎	肺孢子虫肺炎	非 UIP	训练	-1.47	N/A	-1.49	N/A
10-101	C	下								细支气管炎	非 UIP	训练	-2.15	N/A	-0.97	N/A
10-101	D	下	56	F	是	否	17	HP	细支气管炎	细支气管炎	非 UIP	训练	-0.24	N/A	-0.91	N/A
10-101	E	下								细支气管炎	非 UIP	训练	0.30	N/A	1.92	N/A
11-101	C	下	56	M	否	N/A	N/A	UIP	典型 UIP	典型 UIP	UIP	训练	3.26	N/A	1.90	N/A
13-101	C	下								UIP	UIP	训练	5.50	N/A	7.21	N/A
13-101	E	下	67	M	是	否	80	其他	国难 UIP	UIP	UIP	训练	3.66	N/A	6.10	N/A
13-102	A	上								典型 UIP	非 UIP	训练	2.70	N/A	2.13	N/A
13-102	B	上								典型 UIP	非 UIP	训练	5.21	N/A	3.44	N/A
13-102	C	下	61	F	是	否	12	UIP	典型 UIP	典型 UIP	非 UIP	训练	4.26	N/A	4.11	N/A
13-102	D	下								典型 UIP	非 UIP	训练	3.03	N/A	3.04	N/A
13-105	A	上								典型 UIP	UIP	训练	1.94	N/A	2.26	N/A
13-105	B	上								典型 UIP	UIP	训练	2.89	N/A	5.20	N/A
13-105	C	下	57	M	是	否	30	HP	典型 UIP	典型 UIP	UIP	训练	4.38	N/A	5.66	N/A
13-105	D	下								典型 UIP	UIP	训练	2.99	N/A	3.47	N/A
13-105	E	下								典型 UIP	UIP	训练	3.58	N/A	4.57	N/A
13-106	A	上								细支气管炎	非 UIP	训练	0.85	N/A	2.16	N/A
13-106	B	上								细支气管炎	非 UIP	训练	0.45	N/A	2.73	N/A
13-106	C	下	65	F	是	否	14	其他	细支气管炎	细支气管炎	非 UIP	训练	1.43	N/A	2.96	N/A
13-106	D	下								细支气管炎	非 UIP	训练	1.17	N/A	2.47	N/A
13-106	E	下								细支气管炎	非 UIP	训练	1.79	N/A	2.45	N/A
13-110	A	上								NSIP	UIP	训练	0.68	N/A	4.83	N/A
13-110	B	上	52	M	否	N/A	N/A	NSIP	国难 UIP	NSIP	UIP	训练	0.64	N/A	3.90	N/A
13-110	C	下								国难 UIP	UIP	训练	2.83	N/A	4.31	N/A
13-110	D	下								国难 UIP	UIP	训练	1.25	N/A	5.89	N/A
13-111	A	上								NSIP	非 UIP	训练	-0.05	N/A	-0.24	N/A
13-111	B	上								NSIP	非 UIP	训练	0.01	N/A	0.36	N/A
13-111	C	下	70	M	否	N/A	N/A	HP	NSIP	Favor NSIP	非 UIP	训练	-0.60	N/A	0.21	N/A
13-111	D	下								Favor NSIP	非 UIP	训练	1.11	N/A	0.22	N/A
13-111	E	下								Favor NSIP	非 UIP	训练	1.25	N/A	-1.46	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
13-112	A	上								Favor UIP	UIP	训练	2.67	N/A	4.79	N/A
13-112	B	上	68	M	否	N/A	N/A	HP	典型 UIP	Favor UIP	UIP	训练	1.97	N/A	6.37	N/A
13-112	C	下								UIP	UIP	训练	1.76	N/A	5.25	N/A
13-112	D	下								UIP	UIP	训练	4.15	N/A	6.85	N/A
13-201	A	上	49	M	是	否	30	结节病	结节病	肉样瘤病	非 UIP	训练	0.41	N/A	-2.31	N/A
13-201	B	上								肉样瘤病	非 UIP	训练	-2.25	N/A	-3.63	N/A
14-101	B	上								UIP	UIP	训练	2.65	N/A	5.98	N/A
14-101	C	下	80	M	是	否	Unk.	HP	UIP	典型 UIP	UIP	训练	2.51	N/A	3.89	N/A
14-101	D	下								典型 UIP	UIP	训练	3.20	N/A	2.04	N/A
14-101	E	下								典型 UIP	UIP	训练	2.27	N/A	2.68	N/A
15-302	B	中								UIP	UIP	训练	3.70	N/A	6.87	N/A
15-302	C	下	70	F	是	否	33	HP	UIP	UIP	UIP	训练	5.22	N/A	5.76	N/A
15-302	E	下								UIP	UIP	训练	4.38	N/A	5.59	N/A
15-303	C	下	63	F	否	N/A	N/A	NSIP	Favor UIP	Favor UIP	UIP	训练	2.54	N/A	2.39	N/A
15-304	B	中								Favor UIP	UIP	训练	1.45	N/A	0.28	N/A
15-304	C	下	52	M	否	N/A	N/A	HP	Favor UIP	Favor UIP	UIP	训练	1.78	N/A	-0.36	N/A
15-304	D	下								Favor UIP	UIP	训练	4.37	N/A	6.18	N/A
15-305	C	下	58	M	是	否	Unk.	HP	CIF,NOC	Favor UIP	UIP	训练	1.95	N/A	4.20	N/A
15-305	D	下								Favor UIP	UIP	训练	2.33	N/A	3.05	N/A
18-101	C	下	67	F	否	N/A	N/A	结节病	结节病	结节病	非 UIP	训练	-0.42	N/A	1.37	N/A
18-102	A	上	46	F	是	否	1.5	结节病	结节病	结节病	非 UIP	训练	-0.89	N/A	0.30	N/A
18-102	B	上								结节病	非 UIP	训练	-2.07	N/A	0.04	N/A
18-112	C	下								UIP	UIP	训练	-1.01	N/A	0.15	N/A
18-112	D	下	61	F	否	N/A	N/A	NSIP	UIP	UIP	UIP	训练	-1.32	N/A	-0.94	N/A
18-112	E	下								UIP	UIP	训练	0.17	N/A	0.38	N/A
19-301	A	上	66	M	是	否	30	DIP	OP	OP	非 UIP	训练	0.20	N/A	-0.21	N/A
19-301	B	上								OP	非 UIP	训练	-0.28	N/A	-0.82	N/A
19-306	C	下								Favor UIP	UIP	训练	3.66	N/A	8.16	N/A
19-306	D	下	64	F	否	N/A	N/A	HP	Favor UIP	Favor UIP	UIP	训练	2.41	N/A	3.36	N/A
19-306	E	下								Favor UIP	UIP	训练	2.17	N/A	5.25	N/A
32-304	A	上	58	F	否	N/A	N/A	HP	结节病	结节病	非 UIP	训练	-0.03	N/A	-2.05	N/A
32-304	B	上								结节病	非 UIP	训练	-0.24	N/A	-2.81	N/A
32-309	A	上								NSIP	非 UIP	训练	-0.63	N/A	-2.12	N/A
32-309	B	上								NSIP	非 UIP	训练	-0.92	N/A	-0.31	N/A
32-309	C	下	31	F	是	是	12	HP	NSIP	NSIP	非 UIP	训练	1.64	N/A	-1.00	N/A
32-309	D	下								NSIP	非 UIP	训练	0.40	N/A	-0.67	N/A
32-309	E	下								NSIP	非 UIP	训练	-0.10	N/A	-2.39	N/A
32-311	C	下								Favor UIP	UIP	训练	1.37	N/A	2.25	N/A
32-311	D	下	67	M	否	N/A	N/A	明魂 UIP	UIP	Favor UIP	UIP	训练	1.74	N/A	3.90	N/A
32-311	E	下								Favor UIP	UIP	训练	3.46	N/A	4.34	N/A
36-101	A	上	53	F	否	N/A	N/A	HP	UIP	UIP	UIP	训练	-0.87	N/A	-1.73	N/A
36-101	B	上								UIP	UIP	训练	-0.50	N/A	-1.02	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
36-102	A	上								典型 UIP	UIP	训练	1.67	N/A	3.34	N/A
36-102	B	上	88	M	否	N/A	N/A	其他	典型 UIP	典型 UIP	UIP	训练	0.79	N/A	3.04	N/A
36-102	C	下								困难 UIP	UIP	训练	1.69	N/A	1.74	N/A
36-102	E	下								困难 UIP	UIP	训练	1.41	N/A	2.81	N/A
01-206	B	上	42	M	是	否	1	丢失	NSIP	NSIP	非 UIP	测试	N/A	0.86	-2.88	N/A
01-207	A	上	42	F	是	是	Unk.	丢失	DAD	DAD	非 UIP	测试	N/A	-0.61	3.01	N/A
01-207	B	上								DAD	非 UIP	测试	N/A	-0.91	2.97	N/A
06-314	A	上								Favor UIP	UIP	测试	N/A	4.80	5.65	N/A
06-314	B	上								Favor UIP	UIP	测试	N/A	3.51	4.79	N/A
06-314	C	下	76	F	否	N/A	N/A	明确 UIP	Favor UIP	Favor UIP	UIP	测试	N/A	2.95	3.90	N/A
06-314	D	下								Favor UIP	UIP	测试	N/A	3.29	6.67	N/A
06-314	E	下								Favor UIP	UIP	测试	N/A	4.56	6.72	N/A
06-318	A	上								RB	非 UIP	测试	N/A	-0.27	-0.66	N/A
06-318	B	上								RB	非 UIP	测试	N/A	-1.25	-1.56	N/A
06-318	C	下	45	F	是	是	37.5	丢失	非诊断性	RB	非 UIP	测试	N/A	-0.41	-3.03	N/A
06-318	D	下								RB	非 UIP	测试	N/A	1.52	-1.31	N/A
06-318	E	下								RB	非 UIP	测试	N/A	0.06	-1.43	N/A
08-105	A	上								OP	非 UIP	测试	N/A	-0.95	-2.85	N/A
08-105	B	上	40	F	否	N/A	N/A	OP	OP	OP	非 UIP	测试	N/A	-1.00	-2.91	N/A
08-105	C	下								OP	非 UIP	测试	N/A	-0.27	-0.32	N/A
08-105	E	下								OP	非 UIP	测试	N/A	-0.81	-0.77	N/A
08-109	A	上								Favor UIP	UIP	测试	N/A	1.91	4.16	N/A
08-109	C	下	74	M	是	否	46	结节病	困难 UIP	Favor UIP	UIP	测试	N/A	5.33	10.36	N/A
08-109	D	下								Favor UIP	UIP	测试	N/A	3.84	4.98	N/A
08-109	E	下								Favor UIP	UIP	测试	N/A	3.74	5.82	N/A
08-110	C	下								典型 UIP	UIP	测试	N/A	2.86	5.24	N/A
08-110	D	下	72	M	是	否	52	UIP	典型 UIP	典型 UIP	UIP	测试	N/A	2.56	2.82	N/A
08-110	E	下								典型 UIP	UIP	测试	N/A	6.32	6.69	N/A
08-111	A	上								NSIP	UIP	测试	N/A	5.07	6.02	N/A
08-111	B	上								NSIP	UIP	测试	N/A	2.85	1.72	N/A
08-111	C	下	54	F	是	否	10	HP	UIP	典型 UIP	UIP	测试	N/A	2.18	1.02	N/A
08-111	D	下								典型 UIP	UIP	测试	N/A	3.81	3.49	N/A
08-111	E	下								典型 UIP	UIP	测试	N/A	1.47	2.37	N/A
08-119	A	上								结节病	非 UIP	测试	N/A	-0.98	-7.24	N/A
08-119	B	上	43	F	是	否	10.5	OP	结节病	结节病	非 UIP	测试	N/A	-1.88	-6.17	N/A
08-119	C	下								结节病	非 UIP	测试	N/A	-0.07	-1.88	N/A
08-119	E	下								结节病	非 UIP	测试	N/A	-1.88	-7.24	N/A
08-121	A	上								UIP	UIP	测试	N/A	0.02	0.25	N/A
08-121	B	上	64	F	是	否	Unk.	明确 UIP	UIP	UIP	UIP	测试	N/A	0.96	3.32	N/A
08-121	C	下								UIP	UIP	测试	N/A	0.75	2.23	N/A
08-121	E	下								UIP	UIP	测试	N/A	1.24	2.69	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
08-122	A	上								HP	非UIP	测试	N/A	-1.27	0.32	N/A
08-122	B	上								HP	非UIP	测试	N/A	-1.51	-0.09	N/A
08-122	C	下	50	M	否	N/A	N/A	HP	HP	HP	非UIP	测试	N/A	-2.09	0.10	N/A
08-122	D	下								HP	非UIP	测试	N/A	0.81	2.15	N/A
08-122	E	下								HP	非UIP	测试	N/A	1.24	1.51	N/A
08-124	A	中								Favor UIP	UIP	测试	N/A	2.29	2.73	N/A
08-124	B	中	68	F	是	否	10	HP	UIP	Favor UIP	UIP	测试	N/A	1.74	2.77	N/A
08-124	D	下								Favor UIP	UIP	测试	N/A	5.57	5.14	N/A
08-124	E	下								Favor UIP	UIP	测试	N/A	1.85	-2.84	N/A
08-127	C	下								UIP	UIP	测试	N/A	3.19	3.71	N/A
08-127	D	下	71	M	是	否	25	DIP	UIP	UIP	UIP	测试	N/A	3.15	4.54	N/A
08-127	E	下								UIP	UIP	测试	N/A	4.92	3.10	N/A
08-128	A	上								UIP	UIP	测试	N/A	4.78	6.44	N/A
08-128	B	上								UIP	UIP	测试	N/A	4.59	5.83	N/A
08-128	C	下	75	M	是	否	Unk.	HP	典型UIP	典型UIP	UIP	测试	N/A	2.29	6.47	N/A
08-128	D	下								典型UIP	UIP	测试	N/A	4.98	7.19	N/A
08-128	E	下								典型UIP	UIP	测试	N/A	4.95	7.99	N/A
08-129	A	中								UIP	UIP	测试	N/A	0.44	1.08	N/A
08-129	B	中								UIP	UIP	测试	N/A	0.98	1.02	N/A
08-129	C	下	64	M	是	否	30	很可能UIP	UIP	困难UIP	UIP	测试	N/A	0.01	-0.74	N/A
08-129	D	下								困难UIP	UIP	测试	N/A	1.48	2.58	N/A
08-129	E	下								困难UIP	UIP	测试	N/A	-0.49	-0.04	N/A
08-203	A	上	24	F	是	否	4	其他	嗜酸性粒细胞性肺炎	嗜酸性粒细胞性肺炎	非UIP	测试	N/A	-1.49	-6.21	N/A
08-203	B	上								嗜酸性粒细胞性肺炎	非UIP	测试	N/A	-2.19	-3.19	N/A
10-102	A	中								困难UIP	UIP	测试	N/A	-0.17	-0.26	N/A
10-102	B	中								困难UIP	UIP	测试	N/A	-0.05	-1.37	N/A
10-102	C	下	27	M	否	N/A	N/A	NSIP	困难UIP	困难UIP	UIP	测试	N/A	0.46	3.38	N/A
10-102	D	下								困难UIP	UIP	测试	N/A	-1.57	-1.78	N/A
10-102	E	下								困难UIP	UIP	测试	N/A	0.17	-3.19	N/A
13-103	A	上								UIP	UIP	测试	N/A	2.74	0.32	N/A
13-103	C	下	75	F	否	N/A	N/A	HP	典型UIP	典型UIP	UIP	测试	N/A	0.85	4.06	N/A
13-103	D	下								典型UIP	UIP	测试	N/A	1.23	-0.43	N/A
13-103	E	下								典型UIP	UIP	测试	N/A	-1.71	-0.71	N/A
13-104	C	下	66	F	否	N/A	N/A	NSIP	困难UIP	典型UIP	UIP	测试	N/A	1.37	1.85	N/A
13-104	E	下								典型UIP	UIP	测试	N/A	-1.15	1.16	N/A
13-107	A	上								典型UIP	UIP	测试	N/A	3.02	3.64	N/A
13-107	B	上								典型UIP	UIP	测试	N/A	2.61	3.73	N/A
13-107	C	下	69	M	是	否	13.5	HP	典型UIP	UIP	UIP	测试	N/A	3.60	3.52	N/A
13-107	D	下								UIP	UIP	测试	N/A	4.64	5.13	N/A
13-107	E	下								UIP	UIP	测试	N/A	4.23	3.68	N/A

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
13-108	A	上								UIP	UIP	测试	N/A	1.07	1.24	N/A
13-108	B	上								UIP	UIP	测试	N/A	1.25	1.22	N/A
13-108	C	下	78	F	否	N/A	N/A	HP	典型 UIP	典型 UIP	UIP	测试	N/A	3.32	2.60	N/A
13-108	D	下								典型 UIP	UIP	测试	N/A	5.06	4.21	N/A
13-108	E	下								典型 UIP	UIP	测试	N/A	2.02	2.69	N/A
13-109	A	上	71	M	是	否	106	NSIP	OP	OP	非 UIP	测试	N/A	3.09	0.32	N/A
13-109	B	上								OP	非 UIP	测试	N/A	2.63	1.81	N/A
13-113	B	上	73	M	是	否	45	其他	细支气管炎	细支气管炎	非 UIP	测试	N/A	2.16	-3.20	N/A
13-113	E	下							细支气管炎	细支气管炎	非UIP	测试	N/A	1.81	-2.39	N/A
13-115	A	上								典型 UIP	UIP	测试	N/A	4.99	7.50	N/A
13-115	B	上								典型 UIP	UIP	测试	N/A	3.80	5.44	N/A
13-115	C	下	60	M	否	N/A	N/A	很可能UIP	典型 UIP	UIP	UIP	测试	N/A	3.74	10.28	N/A
13-115	D	下								UIP	UIP	测试	N/A	3.92	6.65	N/A
13-115	E	下								UIP	UIP	测试	N/A	4.34	7.11	N/A
18-114	E	下	77	M	是	否	18	HP	困难 UIP	Favor UIP	UIP	测试	N/A	0.84	1.52	N/A
28-302	E	下	62	M	是	否	40	HP	困难 UIP	UIP	UIP	测试	N/A	1.13	3.92	N/A
32-301	C	下								Favor NSIP	非 UIP	测试	N/A	-2.35	-4.05	N/A
32-301	D	下	18	F	否	N/A	N/A	DIP	Favor NSIP	Favor NSIP	非UIP	测试	N/A	-0.50	-4.92	N/A
32-301	E	下								Favor NSIP	非UIP	测试	N/A	-0.68	-3.31	N/A
32-313	A	上	48	F	是	否	15	结节病	RB	RB	非 UIP	测试	N/A	-2.09	-3.79	N/A
32-313	B	上								RB	非UIP	测试	N/A	-4.54	-8.06	N/A
32-318	A	上								OP	非 UIP	测试	N/A	0.56	1.40	N/A
32-318	B	上								OP	非UIP	测试	N/A	1.09	0.24	N/A
32-318	C	下	38	F	Unk.	N/A	N/A	OP	OP	OP	非 UIP	测试	N/A	2.53	2.23	N/A
32-318	D	下								OP	非UIP	测试	N/A	0.67	-0.79	N/A
32-318	E	下								OP	非UIP	测试	N/A	2.85	3.29	N/A
36-103	A	上								UIP	UIP	测试	N/A	1.05	3.25	N/A
36-103	B	上								UIP	UIP	测试	N/A	0.69	3.43	N/A
36-103	C	下	62	F	否	N/A	N/A	HP	UIP	UIP	UIP	测试	N/A	0.76	2.98	N/A
36-103	D	下								UIP	UIP	测试	N/A	-1.01	-0.01	N/A
36-103	E	下								UIP	UIP	测试	N/A	-0.04	3.12	N/A
47-103	A	上								肺动脉高压	UIP	测试	N/A	3.50	3.75	N/A
47-103	B	上								肺动脉高压	UIP	测试	N/A	3.54	4.50	N/A
47-103	C	下	72	F	否	N/A	N/A	丢失	困难 UIP	困难 UIP	UIP	测试	N/A	2.56	3.78	N/A
47-103	D	下								困难 UIP	UIP	测试	N/A	3.71	5.37	N/A
47-103	E	下								困难 UIP	UIP	测试	N/A	3.33	4.84	N/A
01-201	C	N/A	56	F	是	否	50	丢失	N/A	N/A	N/A	排除	N/A	3.00	N/A	2.51
01-203	C	N/A								N/A	N/A	排除	N/A	2.46	N/A	1.04
01-203	D	N/A	50	M	是	否	2.3	丢失	N/A	N/A	N/A	排除	N/A	2.19	N/A	-1.32
01-203	E	N/A								N/A	N/A	排除	N/A	1.50	N/A	-1.95

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
02-103	A	上								Favor HP	非 UIP	排除	N/A	N/A	N/A	N/A
02-103	B	上	59	F	否	N/A	N/A	丢失	Favor HP	Favor HP	非 UIP	排除	N/A	N/A	N/A	N/A
02-103	D	下								Favor HP	非 UIP	排除	N/A	N/A	N/A	N/A
02-104	E	下	53	F	是	否	56	其他	肉芽肿疾病	肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
03-101	B	N/A								N/A	N/A	排除	N/A	-2.25	N/A	-5.60
03-101	C	N/A	54	F	是	否	40	NSIP	N/A	N/A	N/A	排除	N/A	3.48	N/A	3.51
03-101	D	N/A								N/A	N/A	排除	N/A	3.50	N/A	3.44
03-101	E	N/A								N/A	N/A	排除	N/A	2.60	N/A	1.89
03-201	A	N/A								N/A	N/A	排除	N/A	2.47	N/A	-4.34
03-201	B	N/A	85	M	是	否	50	丢失	N/A	N/A	N/A	排除	N/A	-2.99	N/A	-5.50
03-201	D	N/A								N/A	N/A	排除	N/A	3.12	N/A	0.57
03-201	E	N/A								N/A	N/A	排除	N/A	2.58	N/A	-6.49
06-301	D	下	70	F	否	N/A	N/A	NSIP	非诊断性	OP	非 UIP	排除	N/A	N/A	N/A	N/A
06-301	E	下								OP	非 UIP	排除	N/A	N/A	N/A	N/A
06-302	A	N/A	72	F	否	N/A	N/A	其他	N/A	N/A	N/A	排除	N/A	1.31	N/A	-1.30
06-302	C	N/A								N/A	N/A	排除	N/A	0.22	N/A	-3.40
06-303	A	N/A								N/A	N/A	排除	N/A	4.88	N/A	2.33
06-303	B	N/A								N/A	N/A	排除	N/A	1.49	N/A	-1.00
06-303	C	N/A	67	M	是	是	150	RB	N/A	N/A	N/A	排除	N/A	2.78	N/A	-6.93
06-303	D	N/A								N/A	N/A	排除	N/A	-0.87	N/A	-0.30
06-303	E	N/A								N/A	N/A	排除	N/A	4.53	N/A	2.64
06-304	A	N/A								N/A	N/A	排除	N/A	2.43	N/A	2.83
06-304	B	N/A								N/A	N/A	排除	N/A	1.60	N/A	1.18
06-304	C	N/A	77	M	否	N/A	N/A	HP	N/A	N/A	N/A	排除	N/A	1.38	N/A	1.56
06-304	D	N/A								N/A	N/A	排除	N/A	5.05	N/A	-0.14
06-304	E	N/A								N/A	N/A	排除	N/A	2.82	N/A	4.50
06-305	A	N/A								N/A	N/A	排除	N/A	1.88	N/A	5.56
06-305	B	N/A								N/A	N/A	排除	N/A	2.27	N/A	-0.75
06-305	C	N/A	77	F	否	N/A	N/A	HP	N/A	N/A	N/A	排除	N/A	-0.51	N/A	1.99
06-305	D	N/A								N/A	N/A	排除	N/A	0.61	N/A	-0.87
06-305	E	N/A								N/A	N/A	排除	N/A	2.80	N/A	2.75
06-306	A	N/A								N/A	N/A	排除	N/A	0.75	N/A	-3.18
06-306	B	N/A								N/A	N/A	排除	N/A	1.40	N/A	1.63
06-306	C	下	41	F	是	是	30	其他	非诊断性	SRIF	非 UIP	排除	N/A	N/A	N/A	N/A
06-306	D	下								SRIF	非 UIP	排除	N/A	N/A	N/A	N/A
06-306	E	下								SRIF	非 UIP	排除	N/A	N/A	N/A	N/A
06-312	A	N/A								N/A	N/A	排除	N/A	-0.85	N/A	0.34
06-312	B	N/A								N/A	N/A	排除	N/A	2.05	N/A	2.55
06-312	C	N/A	70	F	是	否	丢失	丢失	N/A	N/A	N/A	排除	N/A	1.60	N/A	0.45
06-312	D	N/A								N/A	N/A	排除	N/A	2.47	N/A	3.41
06-312	E	N/A								N/A	N/A	排除	N/A	0.98	N/A	-1.17

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
06-316	A	上	62	F	是	是	60	RB	无诊断	SRIF	非 UIP	排除	N/A	N/A	N/A	N/A
06-316	B	上								SRIF	非UIP	排除	N/A	N/A	N/A	N/A
08-101	A	上	60	M	是	否	36	嗜酸性粒细胞性肺炎	DAD	RB	非UIP	排除	N/A	N/A	N/A	N/A
08-102	B	上	76	F		否	45	嗜酸性粒细胞性肺炎	细胞 NSIP	细胞 NSIP	非UIP	排除	N/A	N/A	N/A	N/A
08-103	C	下	78	M	是	否	60	NSIP	典型 UIP	困难 UIP	UIP	排除	N/A	N/A	N/A	N/A
08-104	A	N/A	71	F	是	否	48	UIP	典型 UIP	N/A	N/A	排除	N/A	3.78	N/A	3.89
08-104	B	N/A								N/A	N/A	排除	N/A	2.01	N/A	-0.69
08-106	E	下	58	M	否	N/A	N/A	NSIP	UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
08-108	C	N/A								N/A	N/A	排除	N/A	-0.99	N/A	-1.39
08-108	D	N/A	71	M	是	否	74	嗜酸性粒细胞性肺炎	困难 UIP	N/A	N/A	排除	N/A	1.26	N/A	0.24
08-108	E	N/A								N/A	N/A	排除	N/A	2.22	N/A	2.81
08-109	B	上	74	M	是	否	46	结节病	困难 UIP	Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
08-110	A	N/A	72	M	是	否	52	UIP	典型 UIP	N/A	N/A	排除	N/A	2.65	N/A	1.55
08-110	B	N/A								N/A	N/A	排除	N/A	1.70	N/A	-0.23
08-112	D	下	48	F	是	是	31	HP	典型 UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
08-114	A	N/A	61	M	是	是	46	HP	困难 UIP	N/A	N/A	排除	N/A	-10.23	N/A	-14.59
08-114	B	N/A								N/A	N/A	排除	N/A	1.62	N/A	0.71
08-117	A	N/A	73	M	是	否	51	其他	CIF,NOC	N/A	N/A	排除	N/A	5.57	N/A	6.32
08-117	B	N/A								N/A	N/A	排除	N/A	3.25	N/A	2.34
08-121	D	下	64	F	是	否	Unk.	明确 UIP	UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
08-123	A	N/A	69	F	是	否	2	HP	HP	N/A	N/A	排除	N/A	2.50	N/A	3.48
08-123	B	N/A								N/A	N/A	排除	N/A	2.90	N/A	1.32
08-124	C	下	68	F	是	否	10	HP	UIP	Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
08-126	A	N/A								N/A	N/A	排除	N/A	1.63	N/A	-0.12
08-126	B	N/A								N/A	N/A	排除	N/A	2.62	N/A	2.16
08-126	C	N/A	74	F	否	N/A	N/A	丢失	N/A	N/A	N/A	排除	N/A	2.39	N/A	3.01
08-126	D	N/A								N/A	N/A	排除	N/A	1.88	N/A	2.30
08-126	E	N/A								N/A	N/A	排除	N/A	0.28	N/A	1.63
08-127	A	N/A	71	M	是	否	25	DIP	UIP	N/A	N/A	排除	N/A	0.74	N/A	1.06
08-127	B	N/A								N/A	N/A	排除	N/A	-3.26	N/A	-5.85
08-201	D	N/A	46	M	是	是	51	RB	非诊断性	N/A	N/A	排除	N/A	1.15	N/A	-0.89
08-201	E	N/A								N/A	N/A	排除	N/A	3.13	N/A	3.36
08-203	C	N/A								N/A	N/A	排除	N/A	-2.65	N/A	-5.88
08-203	D	N/A	24	F	是	否	4	其他	嗜酸性粒细胞性肺炎	N/A	N/A	排除	N/A	3.94	N/A	0.27
08-203	E	N/A								N/A	N/A	排除	N/A	3.25	N/A	2.20
08-204	A	上								肺癌	非UIP	排除	N/A	N/A	N/A	N/A
08-204	B	上								肺癌	非 UIP	排除	N/A	N/A	N/A	N/A
08-204	C	N/A	80	M	是	否	80	其他	肺癌	N/A	N/A	排除	N/A	1.79	N/A	0.50
08-204	D	N/A								N/A	N/A	排除	N/A	-0.63	N/A	1.22
08-204	E	N/A								N/A	N/A	排除	N/A	4.20	N/A	4.57

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
08-205	A	上								OP	非 UIP	排除	N/A	N/A	N/A	N/A
08-205	B	上								OP	非 UIP	排除	N/A	N/A	N/A	N/A
08-205	C	N/A	58	M	是	否	9	HP	OP	N/A	N/A	排除	N/A	2.98	N/A	0.50
08-205	D	N/A								N/A	N/A	排除	N/A	-3.14	N/A	-6.00
08-205	E	N/A								N/A	N/A	排除	N/A	-2.48	N/A	-2.48
08-206	A	N/A								N/A	N/A	排除	N/A	-0.06	N/A	-3.60
08-206	B	N/A	53	M	是	是	未知	其他	肺孢子虫肺炎	N/A	N/A	排除	N/A	2.93	N/A	-0.46
08-206	C	下								肺孢子虫肺炎	非 UIP	排除	N/A	N/A	N/A	N/A
10-101	A	上	56	F	是	否	17	HP	细支气管炎	Favor 细支气管炎	非 UIP	排除	N/A	N/A	N/A	N/A
10-101	B	上								Favor 细支气管炎	非 UIP	排除	N/A	N/A	N/A	N/A
11-101	D	下	56	M	否	N/A	N/A	UIP	典型 UIP	典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
11-102	B	上	65	F	是	否	30	UIP	UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
11-102	D	下								UIP	UIP	排除	N/A	N/A	N/A	N/A
11-103	C	下	69	F	是	否	15	UIP	典型 UIP	典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
11-103	D	下								典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
13-101	B	上	67	M	是	否	80	其他	困难 UIP	困难 UIP	UIP	排除	N/A	N/A	N/A	N/A
13-101	D	下								UIP	UIP	排除	N/A	N/A	N/A	N/A
13-102	E	下	61	F	是	否	12	UIP	典型 UIP	典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
13-103	B	上	75	F	否	N/A	N/A	HP	典型 UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
13-104	A	N/A								N/A	N/A	排除	N/A	-0.87	N/A	-1.54
13-104	B	N/A	66	F	否	N/A	N/A	NSIP	困难 UIP	N/A	N/A	排除	N/A	-4.49	N/A	-5.06
13-104	D	下								典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
13-109	C	N/A								N/A	N/A	排除	N/A	3.82	N/A	2.87
13-109	D	N/A	71	M	是	否	106	NSIP	OP	N/A	N/A	排除	N/A	4.09	N/A	2.75
13-109	E	N/A								N/A	N/A	排除	N/A	-0.39	N/A	1.38
13-110	E	下	52	M	否	N/A	N/A	NSIP	困难 UIP	困难 UIP	UIP	排除	N/A	N/A	N/A	N/A
13-112	E	下	68	M	否	N/A	N/A	HP	典型 UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
13-113	A	上								细支气管炎	非 UIP	排除	N/A	N/A	N/A	N/A
13-113	C	下	73	M	是	否	45	其他	细支气管炎	细支气管炎	非 UIP	排除	N/A	N/A	N/A	N/A
13-113	D	下								细支气管炎	非 UIP	排除	N/A	N/A	N/A	N/A
13-201	C	N/A								N/A	N/A	排除	N/A	3.00	N/A	-0.97
13-201	D	N/A	49	M	是	否	30	结节病	结节病	N/A	N/A	排除	N/A	-2.81	N/A	-6.40
13-201	E	N/A								N/A	N/A	排除	N/A	3.36	N/A	2.59
14-101	A	上	80	M	是	否	未知	HP	UIP	UIP	UIP	排除	N/A	N/A	N/A	N/A
14-102	A	N/A								N/A	N/A	排除	N/A	-0.25	N/A	0.75
14-102	C	N/A	45	F	否	N/A	N/A	HP	N/A	N/A	N/A	排除	N/A	4.35	N/A	3.25
14-102	D	N/A								N/A	N/A	排除	N/A	-1.86	N/A	-2.89
14-102	E	N/A								N/A	N/A	排除	N/A	1.09	N/A	0.32
15-301	C	N/A								N/A	N/A	排除	N/A	-0.37	N/A	-0.99
15-301	D	N/A	72	F	是	否	2.5	UIP	N/A	N/A	N/A	排除	N/A	-0.77	N/A	-3.90
15-301	E	N/A								N/A	N/A	排除	N/A	3.87	N/A	0.48
15-303	A	N/A	63	F	否	N/A	N/A	NSIP	Favor UIP	N/A	N/A	排除	N/A	4.04	N/A	2.01

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
15-304	A	中	52	M	No	N/A	N/A	HP	Favor UIP	Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
15-305	A	N/A								N/A	N/A	排除	N/A	0.65	N/A	0.95
15-305	B	N/A	58	M	是	否	未知	HP	CIF,NOC	N/A	N/A	排除	N/A	4.36	N/A	2.34
15-305	E	下								Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
15-306	A	N/A								N/A	N/A	排除	N/A	2.09	N/A	1.51
15-306	B	N/A								N/A	N/A	排除	N/A	2.46	N/A	1.65
15-306	C	下	78	F	否	N/A	N/A	其他	UIP	Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
15-306	D	下								Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
15-306	E	下								Favor UIP	UIP	排除	N/A	N/A	N/A	N/A
18-101	A	N/A								N/A	N/A	排除	N/A	-0.12	N/A	-1.70
18-101	B	N/A	67	F	否	N/A	N/A	结节病	结节病	N/A	N/A	排除	N/A	4.01	N/A	4.54
18-101	D	下								结节病	非 UIP	排除	N/A	N/A	N/A	N/A
18-104	A	下	83	M	是	否	40	HP	典型 UIP	典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
18-104	D	下								典型 UIP	UIP	排除	N/A	N/A	N/A	N/A
18-106	D	N/A	75	M	是	否	20	RB	N/A	N/A	N/A	排除	N/A	0.08	N/A	0.65
18-106	E	N/A								N/A	N/A	排除	N/A	2.72	N/A	1.66
18-108	C	N/A								N/A	N/A	排除	N/A	-1.04	N/A	-4.09
18-108	D	N/A	70	M	否	N/A	N/A	其他	N/A	N/A	N/A	排除	N/A	1.04	N/A	0.79
18-108	E	N/A								N/A	N/A	排除	N/A	-2.38	N/A	-4.27
18-109	C	下								RB	非 UIP	排除	N/A	N/A	N/A	N/A
18-109	D	下	55	F	是	是	21.5	HP	RB	RB	非 UIP	排除	N/A	N/A	N/A	N/A
18-109	E	下								RB	非 UIP	排除	N/A	N/A	N/A	N/A
18-110	C	下								肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-110	D	下	61	M	是	是	30	细支气管炎	肺气肿	肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-110	E	下								肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-113	C	N/A	39	F	否	N/A	N/A	NSIP	N/A	N/A	N/A	排除	N/A	1.66	N/A	-0.98
18-115	C	下								肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-115	D	下	50	F	是	缺失	74	HP	肺气肿	肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-115	E	下								肺气肿	非 UIP	排除	N/A	N/A	N/A	N/A
18-116	C	N/A	63	M	否	N/A	N/A	结节病	N/A	N/A	N/A	排除	N/A	-4.05	N/A	-5.59
18-116	D	N/A								N/A	N/A	排除	N/A	6.28	N/A	5.43
18-117	C	N/A								N/A	N/A	排除	N/A	-2.00	N/A	-5.47
18-117	D	N/A	71	F	否	N/A	N/A	其他	N/A	N/A	N/A	排除	N/A	2.44	N/A	0.64
18-117	E	N/A								N/A	N/A	排除	N/A	1.04	N/A	0.79
19-306	A	N/A	64	F	否	N/A	N/A	HP	Favor UIP	N/A	N/A	排除	N/A	-0.85	N/A	-1.76
19-306	B	N/A								N/A	N/A	排除	N/A	1.41	N/A	1.02
20-303	D	N/A	73	F	是	否	1	缺失	N/A	N/A	N/A	排除	N/A	4.28	N/A	2.39
20-303	E	N/A								N/A	N/A	排除	N/A	-0.11	N/A	-2.97
28-301	A	上	58	M	是	否	25	嗜酸性粒细胞性肺炎	非诊断性	含铁血黄素沉着症	N/A	N/A	N/A	N/A	N/A	N/A
28-301	B	上								含铁血黄素沉着症	N/A	N/A	N/A	N/A	N/A	N/A
32-301	A	N/A	18	F	否	N/A	N/A	DIP	Favor NSIP	N/A	N/A	排除	N/A	3.37	N/A	3.96
32-301	B	N/A								N/A	N/A	排除	N/A	3.15	N/A	0.90

患者	TBB	肺叶	年龄	性别	吸烟史	当前吸烟者	吸烟指数	中心放射学 Dx	患者病理学 Dx	样品途径 Dx	UIP 标签	队列	53pt LOPO 评分	53pt 测试评分	84pt LOPO 评分	84pt 测试评分
32-304	C	N/A								N/A	N/A	排除	N/A	-0.72	N/A	-0.23
32-304	D	N/A	58	F	否	N/A	N/A	HP	结节病	N/A	N/A	排除	N/A	0.79	N/A	-0.46
32-304	E	N/A								N/A	N/A	排除	N/A	0.56	N/A	-0.40
32-311	A	N/A	67	M	否	N/A	N/A	明确 UIP	UIP	N/A	N/A	排除	N/A	2.14	N/A	1.91
32-311	B	N/A								N/A	N/A	排除	N/A	0.94	N/A	1.40
32-313	C	N/A								N/A	N/A	排除	N/A	-0.34	N/A	-0.20
32-313	D	N/A	48	F	是	否	15	结节病	RB	N/A	N/A	排除	N/A	-4.23	N/A	-6.70
32-313	E	N/A								N/A	N/A	排除	N/A	-0.97	N/A	0.00
36-101	C	N/A	53	F	否	N/A	N/A	HP	UIP	N/A	N/A	排除	N/A	1.65	N/A	-0.99
36-102	D	下	88	M	否	N/A	N/A	其他	典型 UIP	困难 UIP	UIP	排除	N/A	N/A	N/A	N/A

N/A：数据或信息不可用
排除：从训练和测试集中排除
pn.：肺炎
Pulm.：肺

表15:分类中所用的169个Ensembl基因ID(53-患者分类器)。

SEQ ID NO	基因id	基因生物型	SEQ ID NO	基因id	基因生物型	SEQ ID	基因id	基因生物型
152.	ENSG00000189339	编码蛋白质	153.	ENSG00000105991	编码蛋白质	154.	ENSG00000248713	编码蛋白质
155.	ENSG00000116285	编码蛋白质	156.	ENSG00000136275	编码蛋白质	157.	ENSG00000138795	编码蛋白质
158.	ENSG00000219481	编码蛋白质	159.	ENSG00000146707	编码蛋白质	160.	ENSG00000172399	编码蛋白质
161.	ENSG00000204219	编码蛋白质	162.	ENSG00000221305	miRNA	163.	ENSG00000109471	编码蛋白质
164.	ENSG00000142661	编码蛋白质	165.	ENSG00000012232	编码蛋白质	166.	ENSG00000151005	编码蛋白质
167.	ENSG00000157131	编码蛋白质	168.	ENSG00000104381	编码蛋白质	169.	ENSG00000145736	编码蛋白质
170.	ENSG00000116761	编码蛋白质	171.	ENSG00000204844	lincRNA	172.	ENSG00000168938	编码蛋白质
173.	ENSG00000134245	编码蛋白质	174.	ENSG00000136928	编码蛋白质	175.	ENSG00000169194	编码蛋白质
176.	ENSG00000122497	编码蛋白质	177.	ENSG00000136881	编码蛋白质	178.	ENSG00000113621	编码蛋白质
179.	ENSG00000159164	编码蛋白质	180.	ENSG00000136883	编码蛋白质	181.	ENSG00000253910	编码蛋白质
182.	ENSG00000232671	编码蛋白质	183.	ENSG00000148200	编码蛋白质	184.	ENSG00000261934	编码蛋白质
185.	ENSG00000143367	编码蛋白质	186.	ENSG00000148339	编码蛋白质	187.	ENSG00000145888	编码蛋白质
188.	ENSG00000143320	编码蛋白质	189.	ENSG00000176919	编码蛋白质	190.	ENSG00000055163	编码蛋白质
191.	ENSG00000143195	编码蛋白质	192.	ENSG00000107929	编码蛋白质	193.	ENSG00000184845	编码蛋白质
194.	ENSG00000007908	编码蛋白质	195.	ENSG00000207937	miRNA	196.	ENSG00000234284	编码蛋白质
197.	ENSG00000171806	编码蛋白质	198.	ENSG00000188234	编码蛋白质	199.	ENSG00000198518	编码蛋白质
200.	ENSG00000007933	编码蛋白质	201.	ENSG00000148541	编码蛋白质	202.	ENSG00000261839	lincRNA
203.	ENSG00000162782	编码蛋白质	204.	ENSG00000204020	编码蛋白质	205.	ENSG00000235109	编码蛋白质
206.	ENSG00000177489	编码蛋白质	207.	ENSG00000148702	编码蛋白质	208.	ENSG00000204701	编码蛋白质
209.	ENSG00000138075	编码蛋白质	210.	ENSG00000149043	编码蛋白质	211.	ENSG00000204632	编码蛋白质
212.	ENSG00000135625	编码蛋白质	213.	ENSG00000130598	编码蛋白质	214.	ENSG00000204110	lincRNA
215.	ENSG00000115317	编码蛋白质	216.	ENSG00000171987	编码蛋白质	217.	ENSG00000124641	编码蛋白质
218.	ENSG00000183281	编码蛋白质	219.	ENSG00000166796	编码蛋白质	220.	ENSG00000124702	编码蛋白质
221.	ENSG00000144057	编码蛋白质	222.	ENSG00000183908	编码蛋白质	223.	ENSG00000112818	编码蛋白质
224.	ENSG00000257207	编码蛋白质	225.	ENSG00000166004	编码蛋白质	226.	ENSG00000174156	编码蛋白质
227.	ENSG00000144320	编码蛋白质	228.	ENSG00000183560	编码蛋白质	229.	ENSG00000118402	编码蛋白质
230.	ENSG00000188282	编码蛋白质	231.	ENSG00000149289	编码蛋白质	232.	ENSG00000112299	编码蛋白质
233.	ENSG00000074582	编码蛋白质	234.	ENSG00000254842	lincRNA	235.	ENSG00000048052	编码蛋白质
236.	ENSG00000054356	编码蛋白质	237.	ENSG00000010379	编码蛋白质	238.	ENSG00000129204	编码蛋白质
239.	ENSG00000114923	编码蛋白质	240.	ENSG00000111321	编码蛋白质	241.	ENSG00000129221	编码蛋白质
242.	ENSG00000115009	编码蛋白质	243.	ENSG00000212126	编码蛋白质	244.	ENSG00000108551	编码蛋白质
245.	ENSG00000181798	加工的转录物	246.	ENSG00000110900	编码蛋白质	247.	ENSG00000108342	编码蛋白质
248.	ENSG00000144712	编码蛋白质	249.	ENSG00000139211	编码蛋白质	250.	ENSG00000131095	编码蛋白质
251.	ENSG00000168329	编码蛋白质	252.	ENSG00000187166	编码蛋白质	253.	ENSG00000167105	编码蛋白质
254.	ENSG00000168036	编码蛋白质	255.	ENSG00000086159	编码蛋白质	256.	ENSG00000258890	编码蛋白质
257.	ENSG00000179152	编码蛋白质	258.	ENSG00000170374	编码蛋白质	259.	ENSG00000141562	编码蛋白质

SEQ ID NO	基因id	基因生物型	SEQ ID NO	基因id	基因生物型	SEQ ID NO	基因id	基因生物型
260.	ENSG00000256097	编码蛋白质	261.	ENSG00000221479	miRNA	262.	ENSG00000128791	编码蛋白质
263.	ENSG00000227124	编码蛋白质	264.	ENSG00000139352	编码蛋白质	265.	ENSG00000170558	编码蛋白质
266.	ENSG00000184500	编码蛋白质	267.	ENSG00000122966	编码蛋白质	268.	ENSG00000075643	编码蛋白质
269.	ENSG00000206531	编码蛋白质	270.	ENSG00000125255	编码蛋白质	271.	ENSG00000166573	编码蛋白质
272.	ENSG00000163884	编码蛋白质	273.	ENSG00000134905	编码蛋白质	274.	ENSG00000256463	编码蛋白质
275.	ENSG00000180697	编码蛋白质	276.	ENSG00000187630	编码蛋白质	277.	ENSG00000125827	编码蛋白质
278.	ENSG00000198685	编码蛋白质	279.	ENSG00000257365	编码蛋白质	280.	ENSG00000182931	编码蛋白质
281.	ENSG00000034533	编码蛋白质	282.	ENSG00000133997	编码蛋白质	283.	ENSG00000198768	编码蛋白质
284.	ENSG00000172667	编码蛋白质	285.	ENSG00000119725	编码蛋白质	286.	ENSG00000101188	编码蛋白质
287.	ENSG00000078070	编码蛋白质	288.	ENSG00000198208	编码蛋白质	289.	ENSG00000131142	编码蛋白质
290.	ENSG00000159674	编码蛋白质	291.	ENSG00000258945	编码蛋白质	292.	ENSG00000086544	编码蛋白质
293.	ENSG00000174123	编码蛋白质	294.	ENSG00000169918	编码蛋白质	295.	ENSG00000188293	编码蛋白质
296.	ENSG00000109158	编码蛋白质	297.	ENSG00000198838	编码蛋白质	298.	ENSG00000167748	编码蛋白质
299.	ENSG00000145248	编码蛋白质	300.	ENSG00000140323	编码蛋白质	301.	ENSG00000189013	编码蛋白质
302.	ENSG00000035720	编码蛋白质	303.	ENSG00000167014	编码蛋白质	304.	ENSG00000022556	编码蛋白质
305.	ENSG00000081041	编码蛋白质	306.	ENSG00000137875	编码蛋白质	307.	ENSG00000273311	内含子意义
308.	ENSG00000145284	编码蛋白质	309.	ENSG00000067141	编码蛋白质	310.	ENSG00000183066	编码蛋白质
311.	ENSG00000170509	编码蛋白质	312.	ENSG00000095917	编码蛋白质	313.	ENSG00000189306	编码蛋白质
314.	ENSG00000170502	编码蛋白质	315.	ENSG00000155714	编码蛋白质	316.	ENSG00000142192	编码蛋白质
317.	ENSG00000163644	编码蛋白质	318.	ENSG00000166848	编码蛋白质			
319.	ENSG00000163110	编码蛋白质	320.	ENSG00000166509	编码蛋白质			

表16:本研究中所用的44种细支气管和肺泡细胞文献标志物

基因	基因名称	细胞类型	证据	Ensembl 基因 ID
SFTPC	表面活性蛋白 C	上皮细胞前体，肺泡 II 型	IHC[2], qPCR[2], 原位杂交[4], 时间进程[2]	ENSG00000168484
PDPN	平足蛋白	上皮细胞前体，肺泡 I 型	IHC[4]	ENSG00000162493
CGRP	CGRP 受体组分	上皮细胞前体	IHC[1,2]	ENSG00000241258
CD34	CD34 分子	上皮细胞前体	IHC[3]	ENSG00000174059
ATXN1	Ataxin 1	上皮细胞前体	IHC[3]	ENSG00000124788
SOX11	SRY 框 11	上皮细胞前体	RNAseq[4]	ENSG00000176887
TUBA1A	微管蛋白 α 1a	上皮细胞前体	RNAseq[4]	ENSG00000167552
FOXJ1	叉头框 J1	纤毛细支气管上皮细胞	IHC[2,4]	ENSG00000129654
AQP4	水通道蛋白 4	纤毛细支气管上皮细胞	IHC[5]	ENSG00000171885
ITGB4	整联蛋白 β4 亚基	纤毛细支气管上皮细胞	qPCR[4]	ENSG00000132470
TOP2A	拓扑异构酶 DNA II α	纤毛细支气管上皮细胞	qPCR[4]	ENSG00000131747
SCGB1A1	分泌球蛋白家族 1A 成员 1	纤毛细支气管上皮细胞，克拉拉细胞	损伤时间进程[2], IHC[1,2,4]	ENSG00000149021
CLDN10	Claudin 10	纤支气管克拉拉细胞	损伤时间进程[2], IHC[2]	ENSG00000134873
KRT15	角蛋白 15	纤支气管克拉拉细胞	IHC[4]	ENSG00000171346
AQP3	水通道蛋白 3	纤支气管克拉拉细胞	原位 EM[5]	ENSG00000165272
CYP2F2P	细胞色素 P450 家族 2 亚家族 F 成员 2，假基因	纤支气管克拉拉细胞	损伤时间进程[2]	ENSG00000237118
FMO3	含黄素单加氧酶 3	纤支气管克拉拉细胞	损伤时间进程[2]	ENSG00000007933
PON1	对氧磷酶 1	纤支气管克拉拉细胞	损伤时间进程[2]	ENSG00000005421
AOX3P	醛氧化酶 3，假基因	纤支气管克拉拉细胞	损伤时间进程[2]	ENSG00000244301
SCGB3A2	分泌球蛋白家族 3A 成员 2	纤支气管克拉拉细胞	微阵列[2]	ENSG00000164265
CES1	羧酸酯酶 1	纤支气管克拉拉细胞	微阵列[2]	ENSG00000198848
GABRP	γ-氨基丁酸 A 型受体 pi 亚基	纤支气管克拉拉细胞	微阵列[2]	ENSG00000094755
SFTPA1	表面活性蛋白 A1	肺泡 I 型和 II 型	IHC[1]	ENSG00000122852
HOPX	HOP 同源框	肺泡 I 型	Tg-IF[4]	ENSG00000171476
AGER	晚期糖基化终产物特异性受体	肺泡 I 型	IHC[4]	ENSG00000204305
AQP5	水通道蛋白 5	肺泡 I 型	qPCR[4], RNAseq[4], IHC[5]	ENSG00000161798
VEGFA	血管内皮生长因子 A	肺泡 I 型	qPCR[4], RNAseq[4]	ENSG00000112715
HES1	Hes 家族 bHLH 转录因子 1	肺泡 I 型	RNAseq[4]	ENSG00000114315

基因	基因名称	细胞类型	证据	Ensembl 基因 ID
SEMA3A	脑信号蛋白 3A	肺泡 I 型	RNAseq[4]	ENSG00000075213
TGFB1	转化生长因子 β1	肺泡 I 型	RNAseq[4]	ENSG00000105329
GPRC5A	G 蛋白偶联受体 C 类 5 组成员 A	肺泡 I 型	RNAseq[4]	ENSG00000013588
EGFL6	EGF 样结构域 6	肺泡 II 型	RNAseq[4], 原位杂交[4]	ENSG00000198759
ABCA3	ATP 结合盒亚家族 A 成员 3	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000167972
MUC1	细胞表面相关的粘蛋白 1	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000185499
LYZ	溶菌酶	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000090382
SFTPB	表面活性蛋白 B	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000168878
CFTR	囊性纤维化跨膜传导调节蛋白	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000001626
CEBPA	CCAAT/增强子结合蛋白 α	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000245848
SFTPD	表面活性蛋白 D	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000133661
ID2	DNA 结合抑制 2，HLH 蛋白	肺泡 II 型	qPCR[4], RNAseq[4]	ENSG00000115738
SOX9	SRY 框 9	肺泡 II 型	RNAseq[4]	ENSG00000125398
CITED2	Cbp/p300 相互作用反式激活蛋白，具有富 Glu/Asp 羧基末端结构域 2	肺泡 II 型	RNAseq[4]	ENSG00000164442
CMTM8	CKLF 样 MARVEL 跨膜结构域 8	肺泡 II 型	RNAseq[4]	ENSG00000170293
FGFR2	纤维母细胞生长因子受体 2	肺泡 II 型	RNAseq[4]	ENSG00000066468

[1]Wuenschell 1996;2 Zemke 2009;3 Kim 2005;4 Treutlein 2014;5 Nielsen 1997。

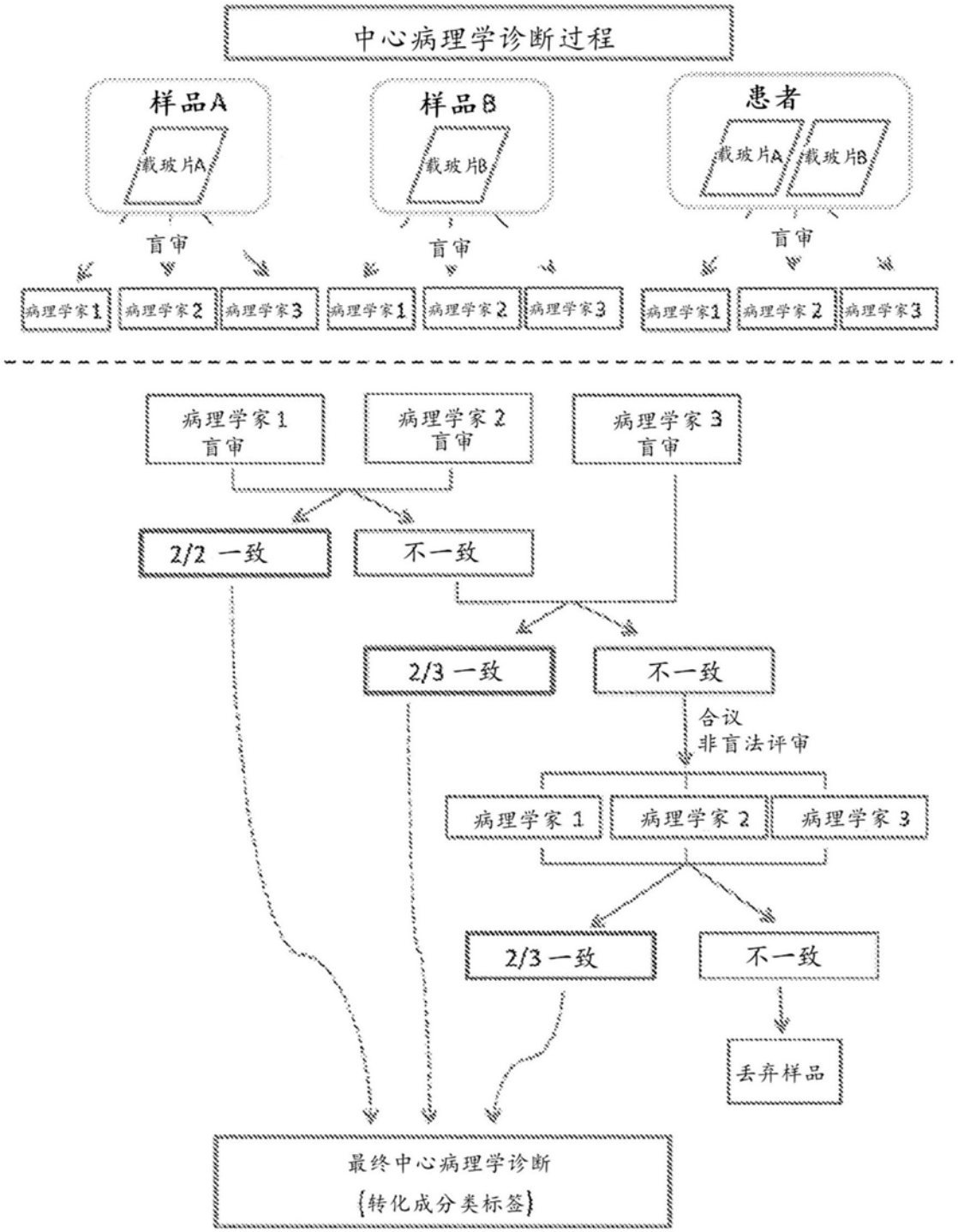

中心病理学诊断过程
样品A
载玻片A
样品B
载玻片B
患者
载玻片A
载玻片B
盲审
盲审
盲审
病理学家1
病理学家2
病理学家3
病理学家1
病理学家2
病理学家3
病理学家1
病理学家2
病理学家3
病理学家1 盲审
病理学家2 盲审
病理学家3 盲审
2/2 一致
不一致
2/3 一致
不一致
合议
非盲法评审
病理学家1
病理学家2
病理学家3
2/3 一致
不一致
丢弃样品
最终中心病理学诊断
{转化成分类标签}

图1

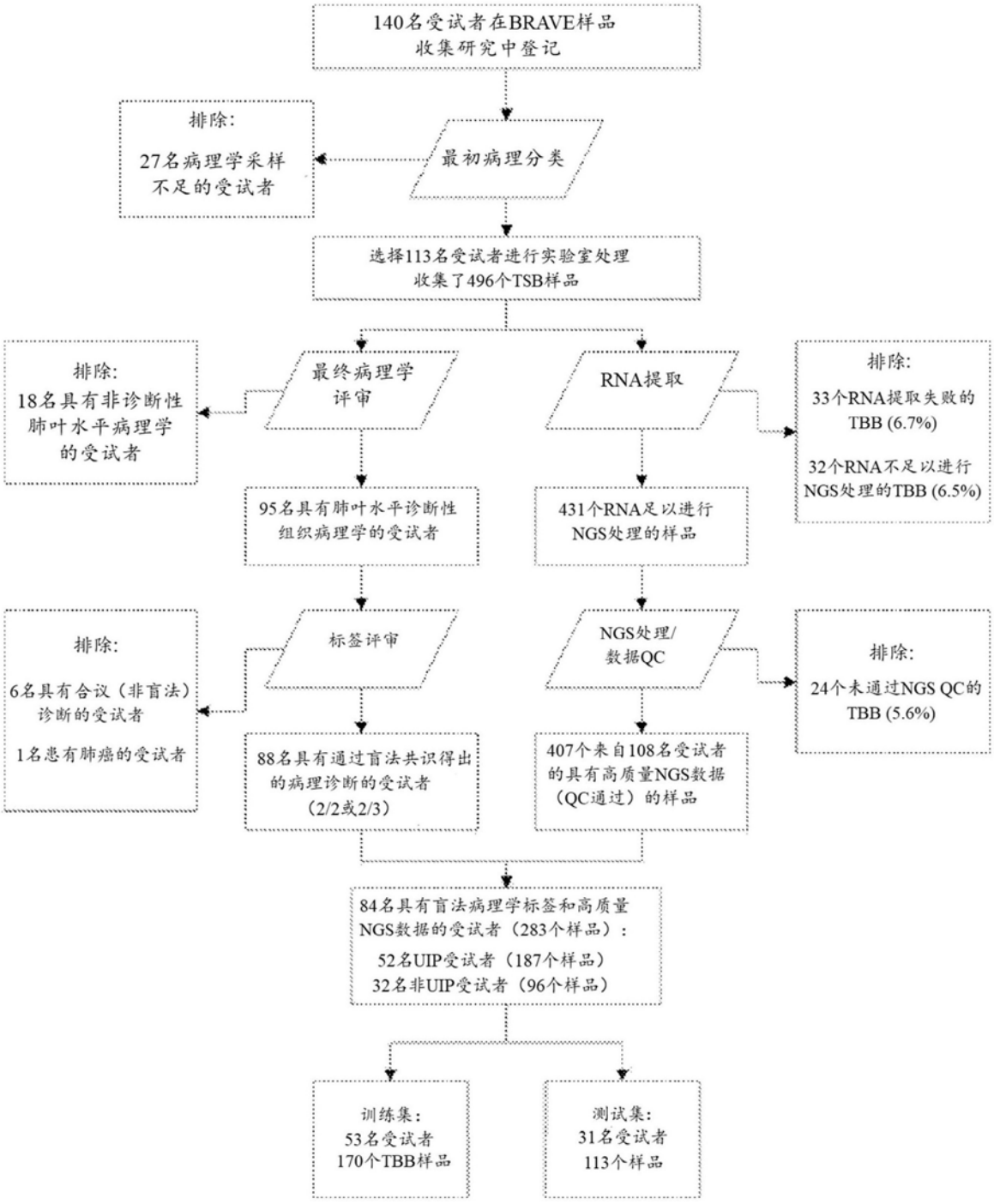

140名受试者在BRAVE样品收集研究中登记
最初病理分类
排除:
27名病理学采样不足的受试者
选择113名受试者进行实验室处理
收集了496个TSB样品
最终病理学评审
RNA提取
排除:
18名具有非诊断性肺叶水平病理学的受试者
排除:
33个RNA提取失败的TBB (6.7%)
32个RNA不足以进行NGS处理的TBB (6.5%)
95名具有肺叶水平诊断性组织病理学的受试者
431个RNA足以进行NGS处理的样品
标签评审
NGS处理/数据QC
排除:
6名具有合议（非盲法）诊断的受试者
1名患有肺癌的受试者
排除:
24个未通过NGS QC的TBB (5.6%)
88名具有通过盲法共识得出的病理诊断的受试者（2/2或2/3）
407个来自108名受试者的具有高质量NGS数据（QC通过）的样品
84名具有盲法病理学标签和高质量NGS数据的受试者（283个样品）：
52名UIP受试者（187个样品）
32名非UIP受试者（96个样品）
训练集:
53名受试者
170个TBB样品
测试集:
31名受试者
113个样品

图2

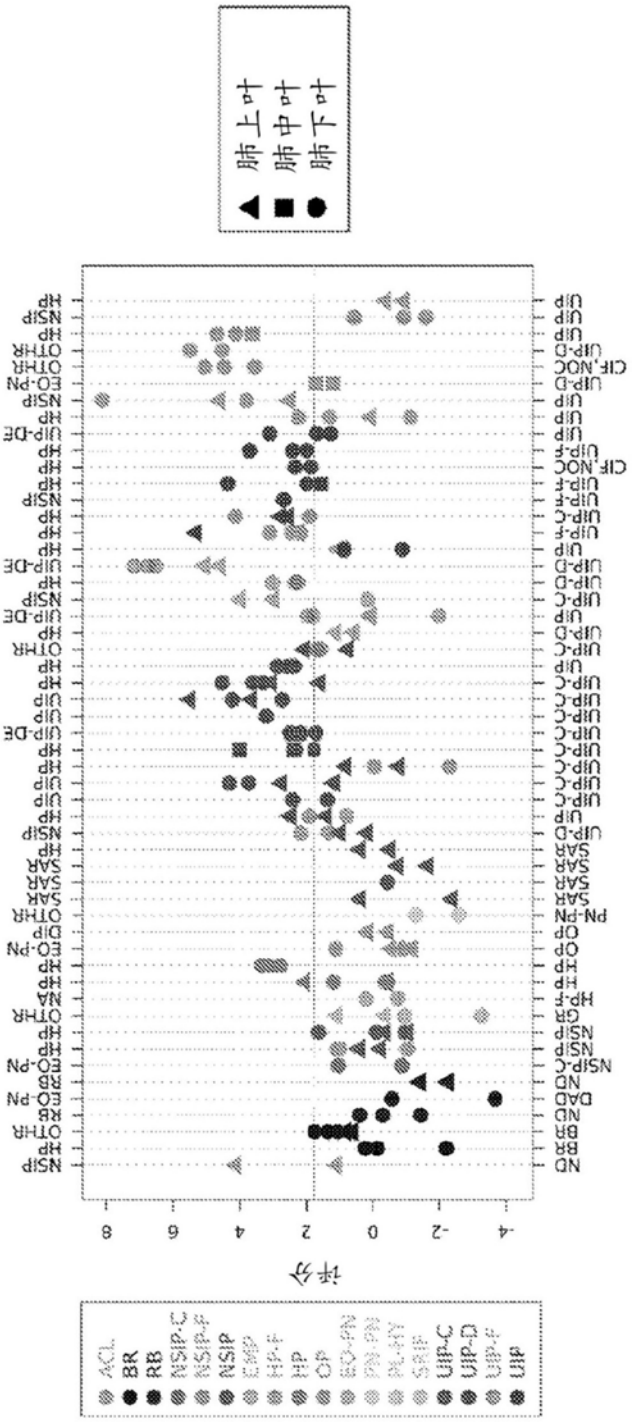

肺上叶
肺中叶
肺下叶
评分
ACL
BR
RB
NSIP-C
NSIP-F
NSIP
UIP-F
HP
OP
EO-PN
PN-PN
AL-LY
SAR
UIP-C
UIP-D
UIP-F
UIP

图3A

		分类结果	
		非UIP	UIP
病理学标签	非 UIP	54	5
	UIP	39	72
AUC		0.85 [0.78-0.91]	
灵敏度		0.65 [0.25, 0.82]	
特异性		0.92 [0.81, 0.97]	

图3B

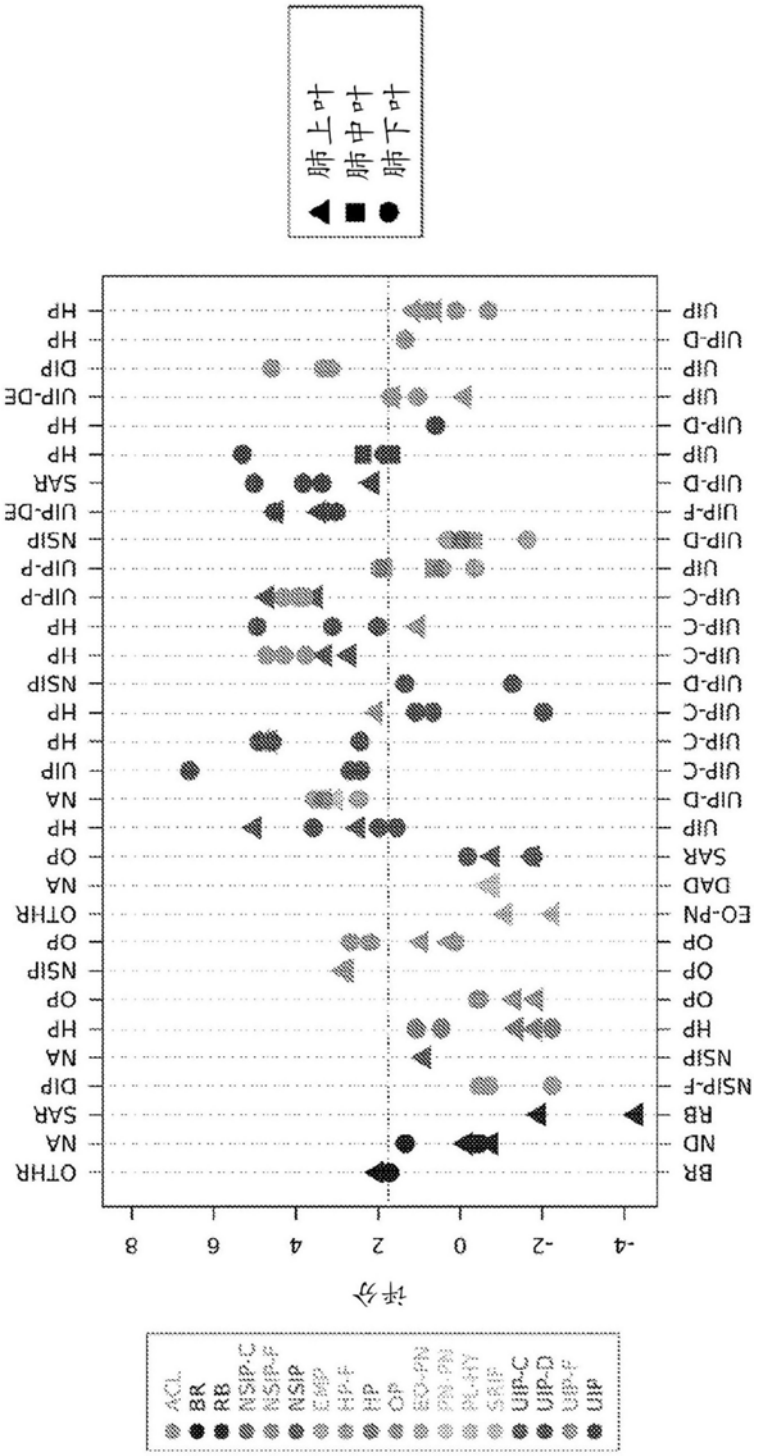

肺上叶
肺中叶
肺下叶
BR
ND
RB
NSIP-F
NSIP
HP
OP
OP
OP
EO-PN
DAD
SAR
UIP
UIP-D
UIP-C
UIP-C
UIP-C
UIP-C
UIP-C
UIP-D
UIP-C
UIP-C
UIP-C
UIP
UIP-D
UIP
UIP-F
UIP-D
UIP
UIP-D
UIP
UIP
UIP-D
UIP
HP
HP
DIP
UIP-DE
HP
HP
SAR
UIP-DE
NSIP
UIP-P
UIP-P
HP
HP
NSIP
HP
HP
UIP
NA
HP
OP
NSIP
OP
OTHR
NA
SAR
DIP
NA
HP
OP
NSIP
OP
OTHR
NA
-4
-2
0
2
4
6
8
评分
ACL
BR
RB
NSIP-C
NSIP-F
NSIP
CHP
HP-F
HP
OP
EO-PN
PN-HY
PL-HY
SRP
UIP-C
UIP-D
UIP-F
UIP

图3C

		分类结果	
		非UIP	UIP
病理学标签	非UIP	32	5
	UIP	28	48
AUC		0.86 [0.79-0.93]	
灵敏度		0.63 [0.43, 0.87]	
特异性		0.86 [0.73, 0.97]	

图3D

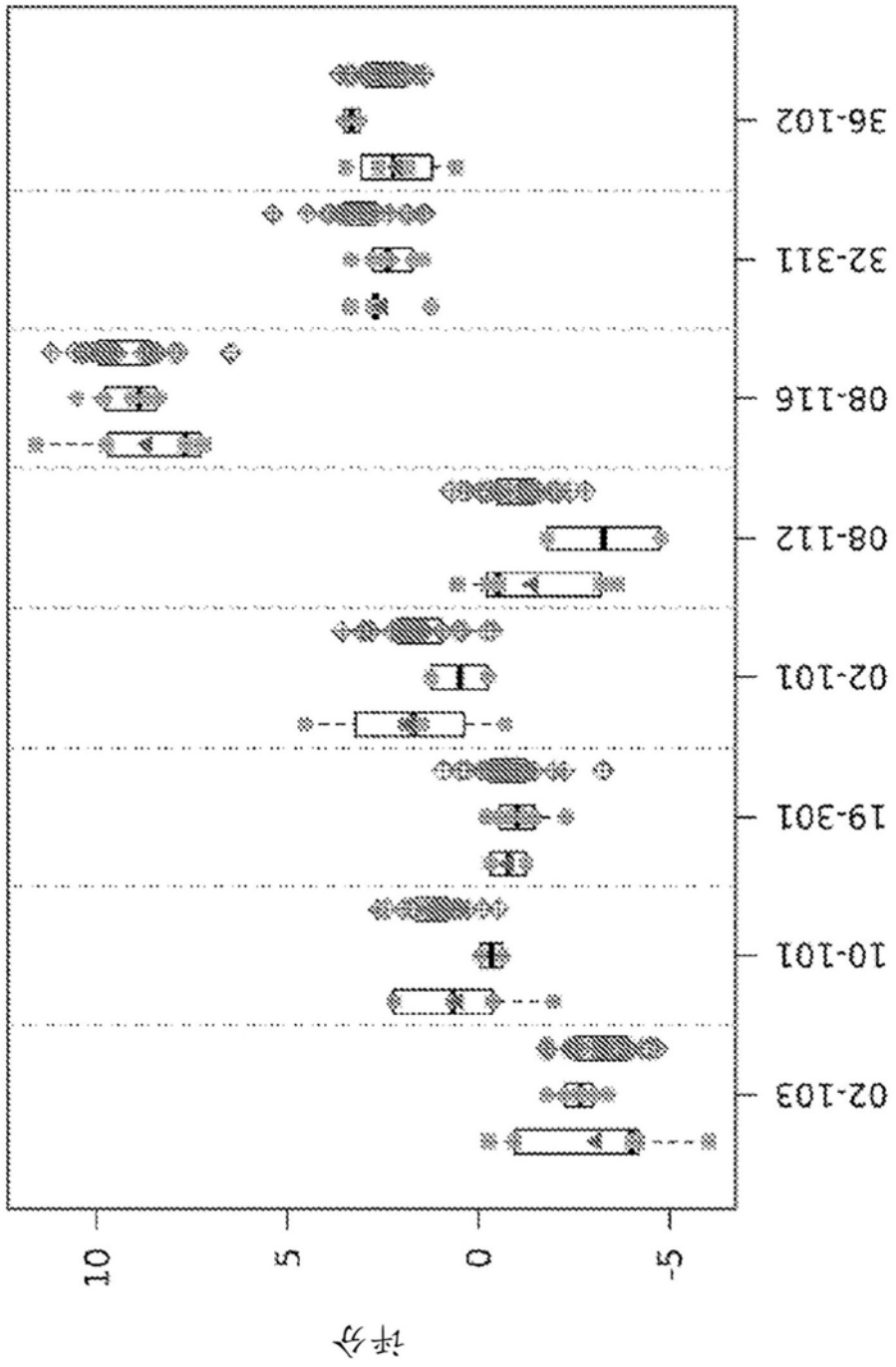


图4A

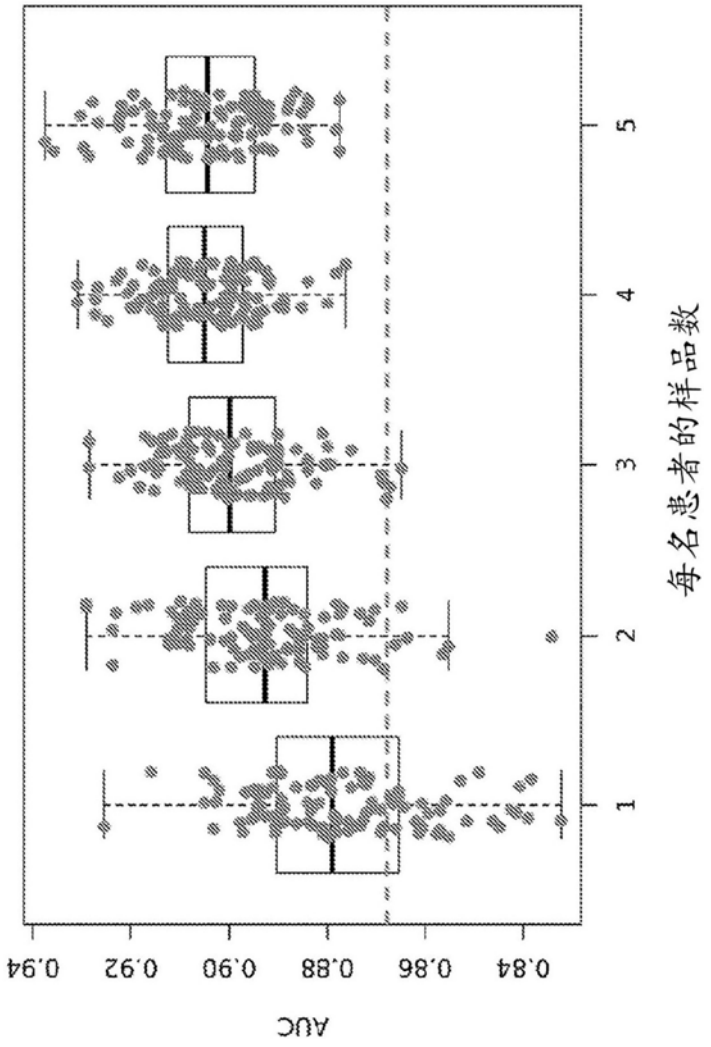


图4B

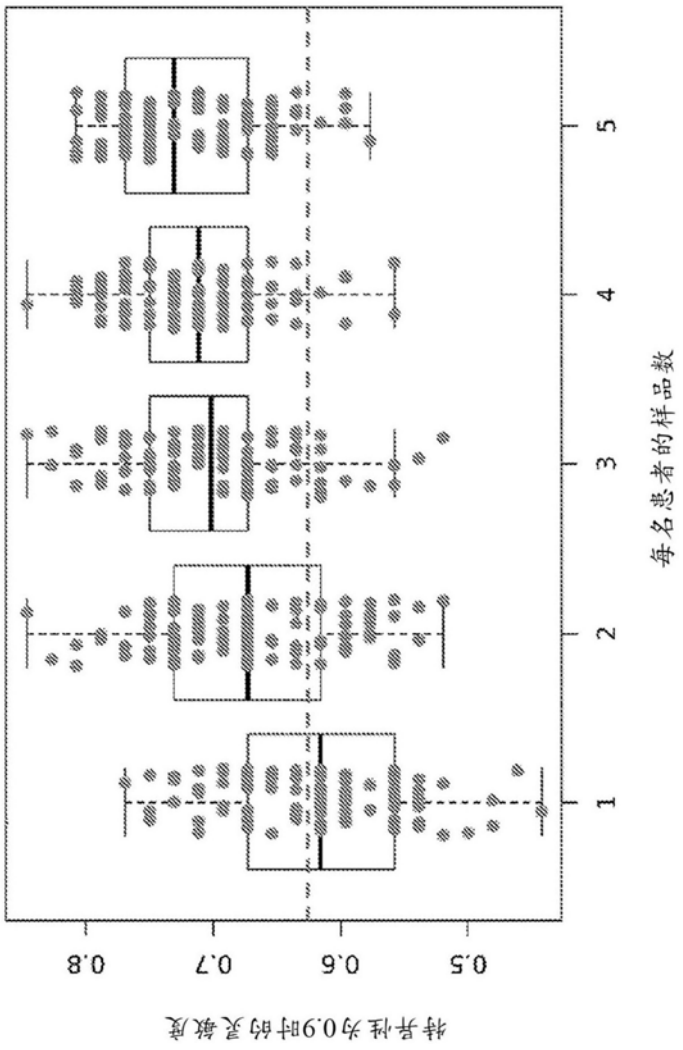


图4C

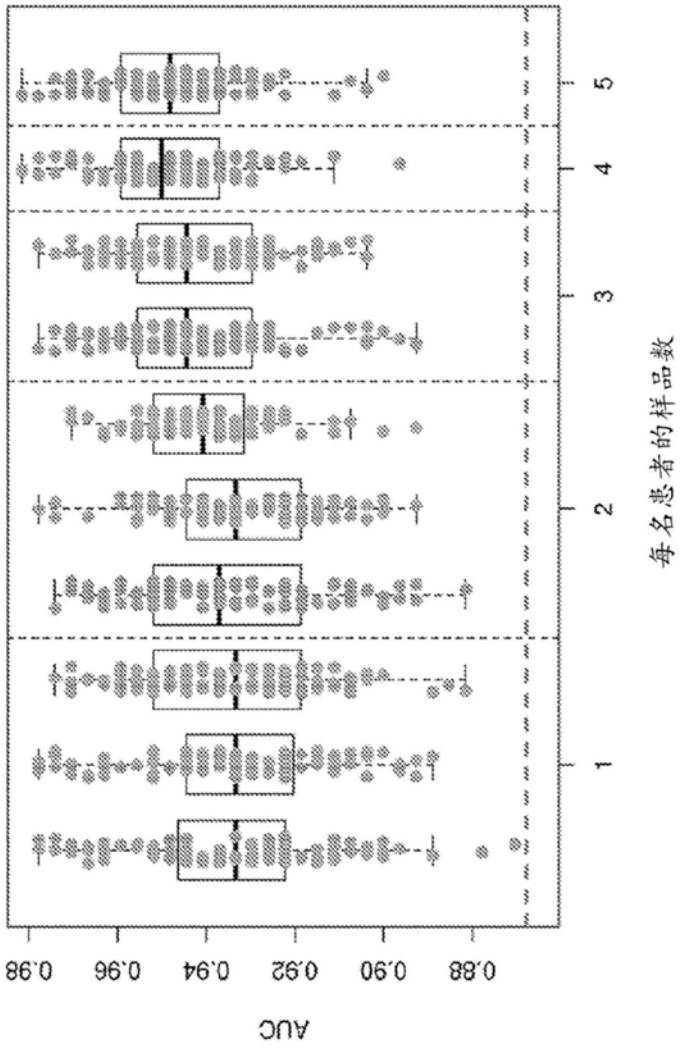


图4D

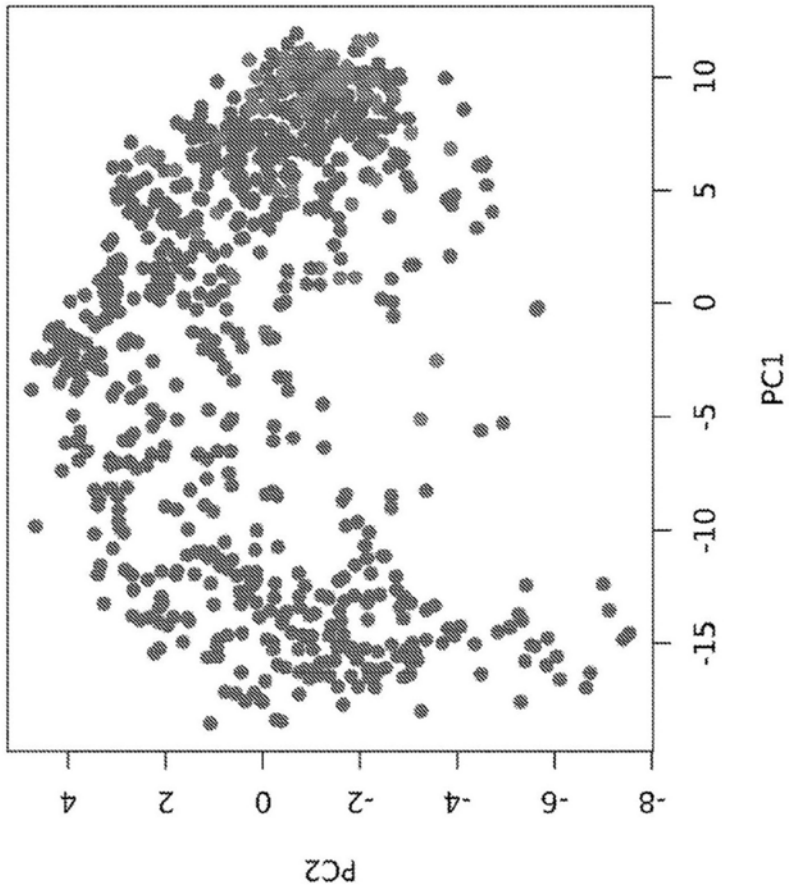


图5A

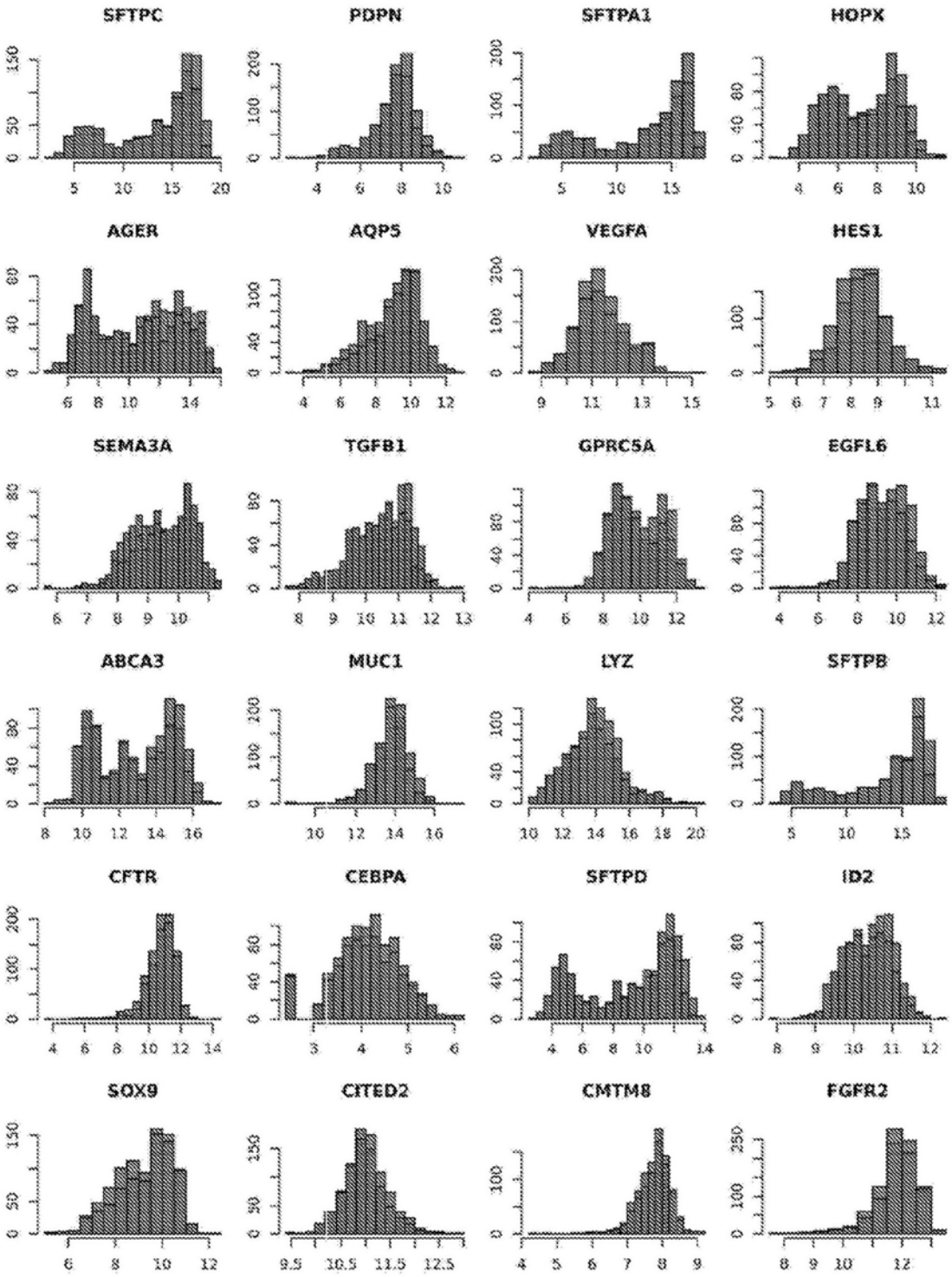

SFTPC
PDPN
SFTPA1
HOPX
AGER
AQP5
VEGFA
HES1
SEMA3A
TGFB1
GPRC5A
EGFL6
ABCA3
MUC1
LYZ
SFTPB
CFTR
CEBPA
SFTPD
ID2
SOX9
CITED2
CMTM8
FGFR2

图5B

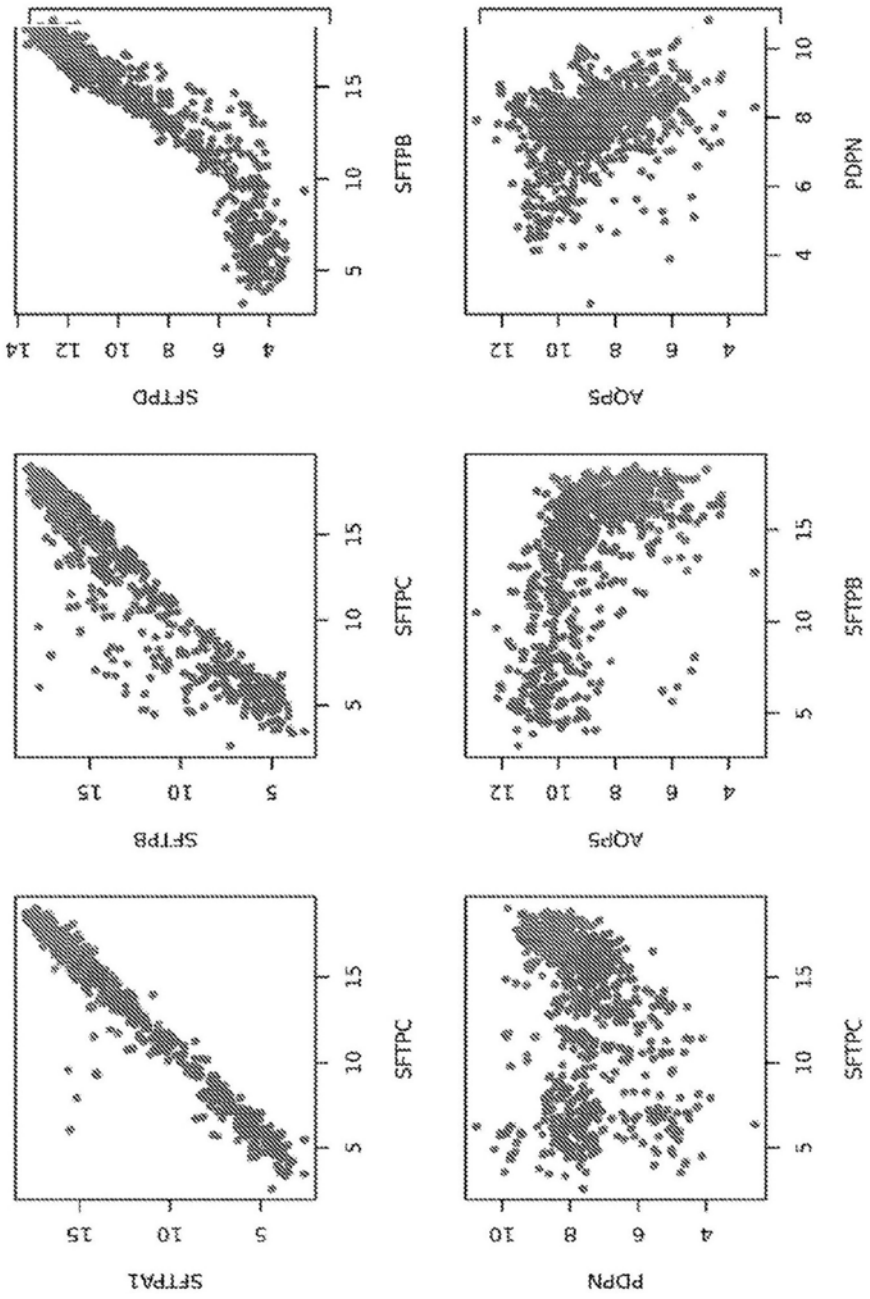

SFTPB
PDPN
SFTPD
AQP5
SFTPC
SFTPB
SFTPB
AQP5
SFTPC
SFTPC
SFTPA1
PDPN

图5C

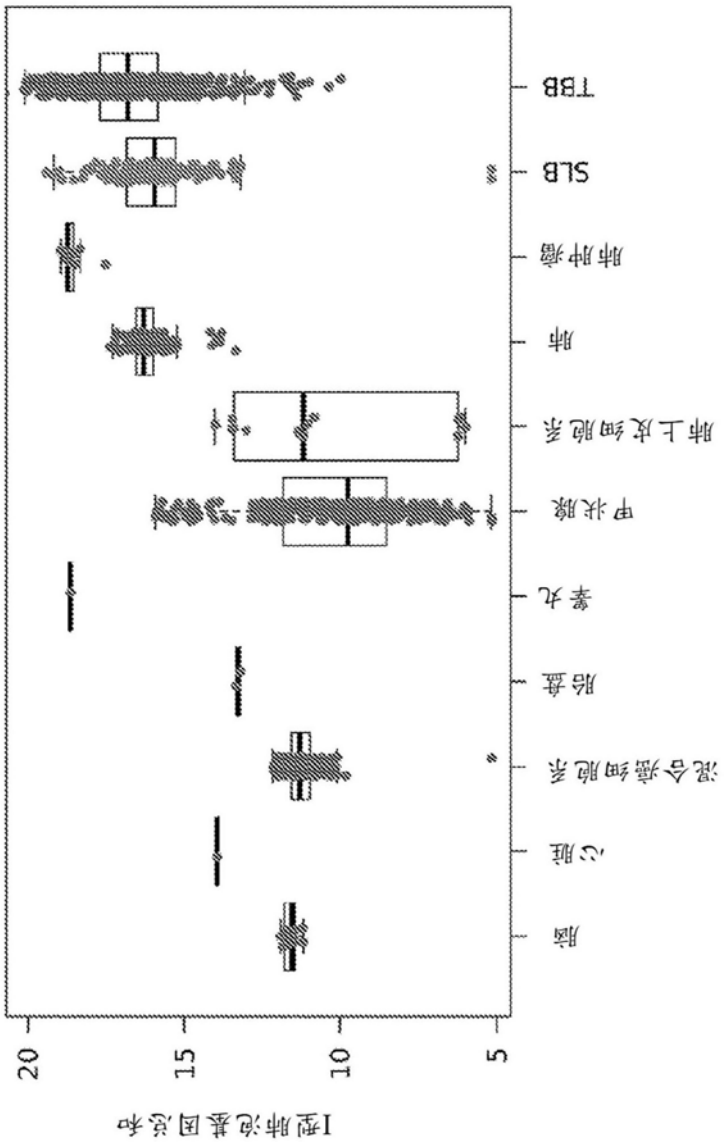


图6A

图6B

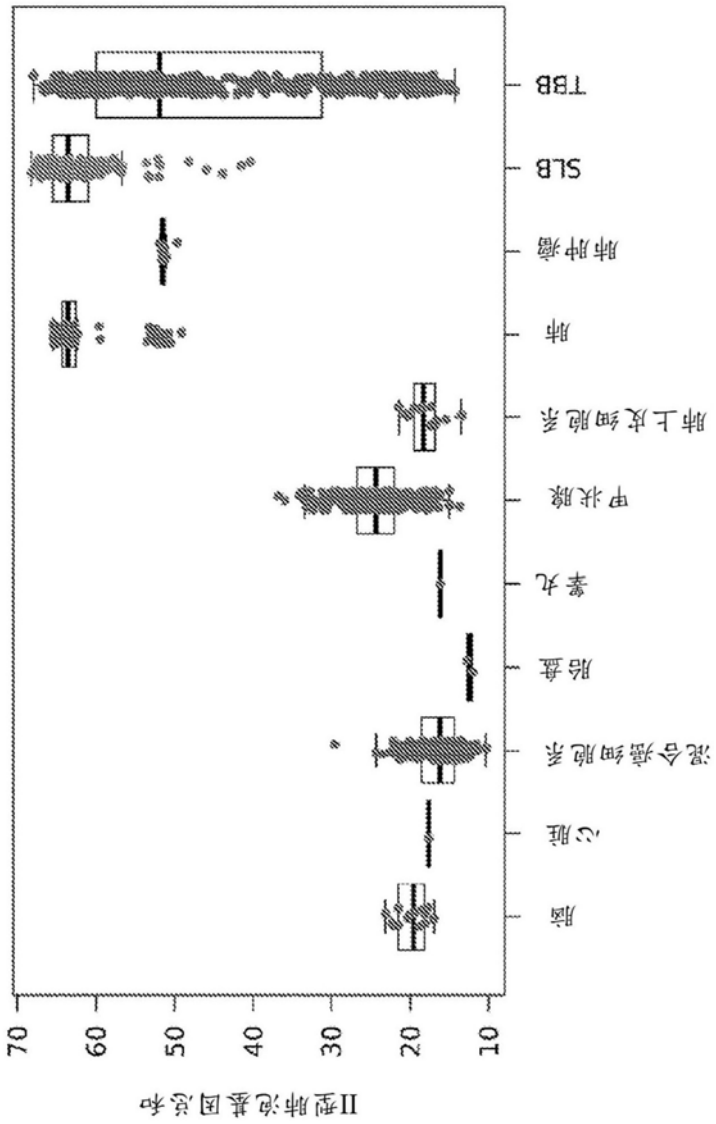

II型肺泡基因总和
70
60
50
40
30
20
10
脑
心脏
混合癌细胞系
胎盘
睾丸
甲状腺
肺上皮细胞系
肺
肺肿瘤
SLB
TBB

图6C

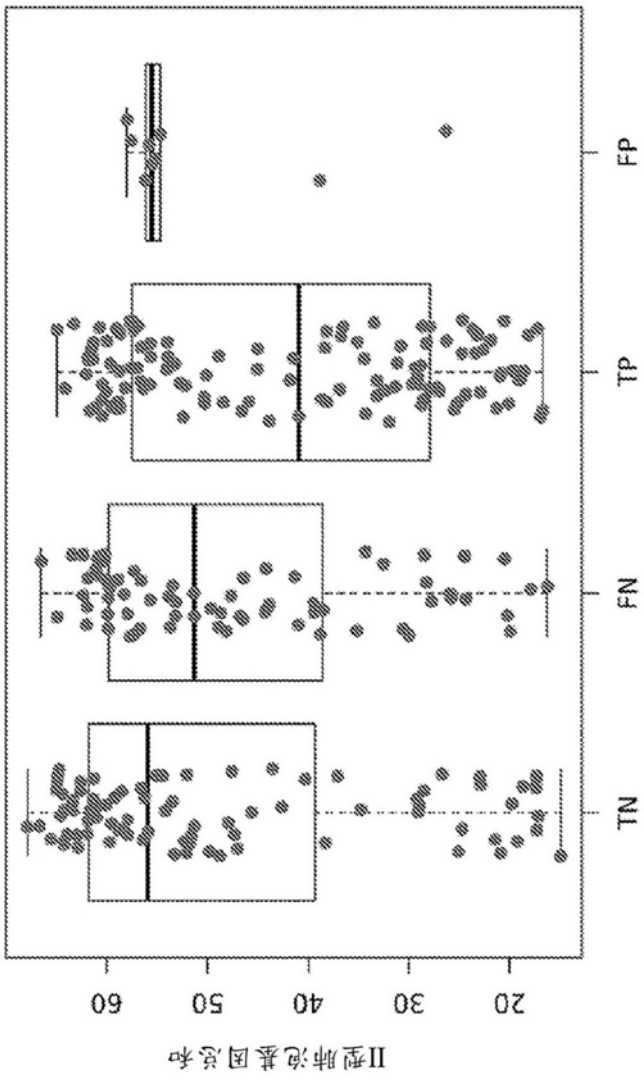


图6D

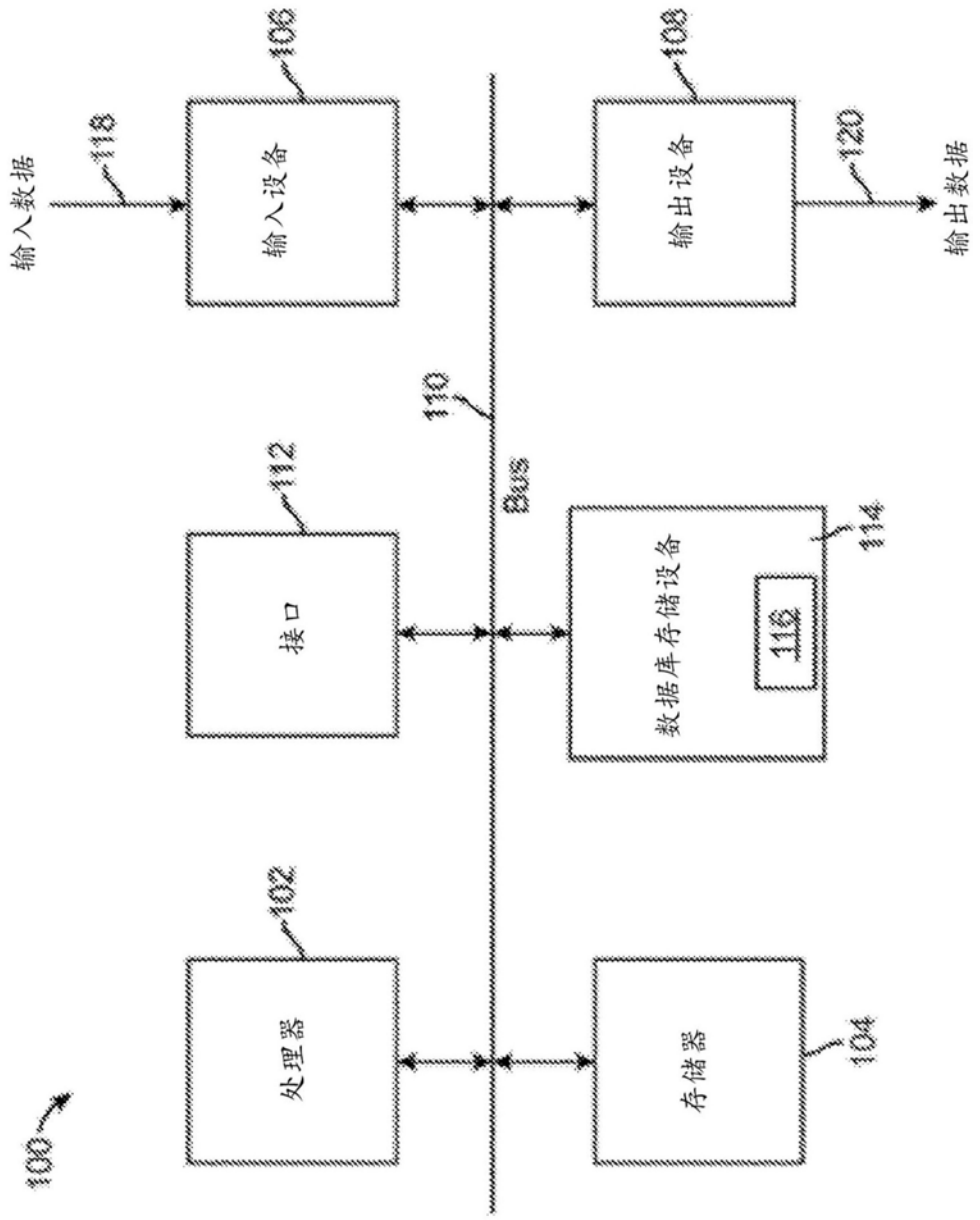

100
处理器
102
接口
112
Bus
110
输入数据
118
输入设备
106
输出设备
108
输出数据
120
数据库存储设备
114
116
存储器
104

图7A

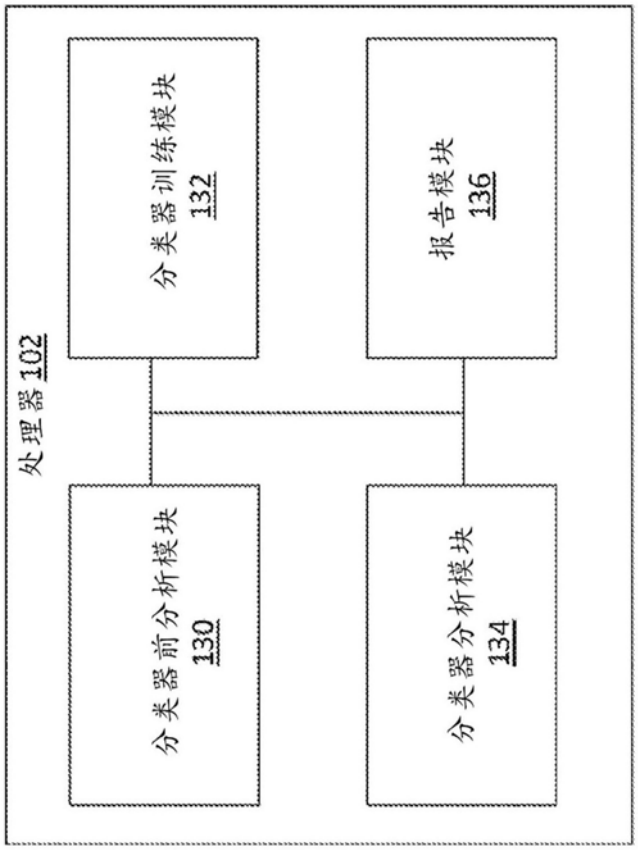

处理器 102
分类器训练模块 132
报告模块 136
分类器前分析模块 130
分类器分析模块 134

图7B

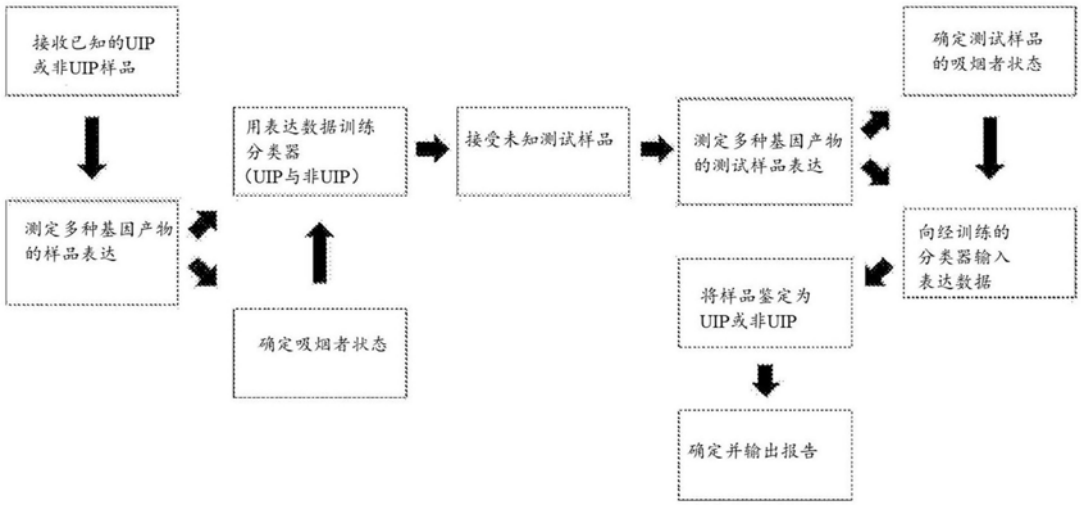

接收已知的UIP或非UIP样品
测定多种基因产物的样品表达
用表达数据训练分类器（UIP与非UIP）
确定吸烟者状态
接受未知测试样品
测定多种基因产物的测试样品表达
确定测试样品的吸烟者状态
向经训练的分类器输入表达数据
将样品鉴定为UIP或非UIP
确定并输出报告

图7C

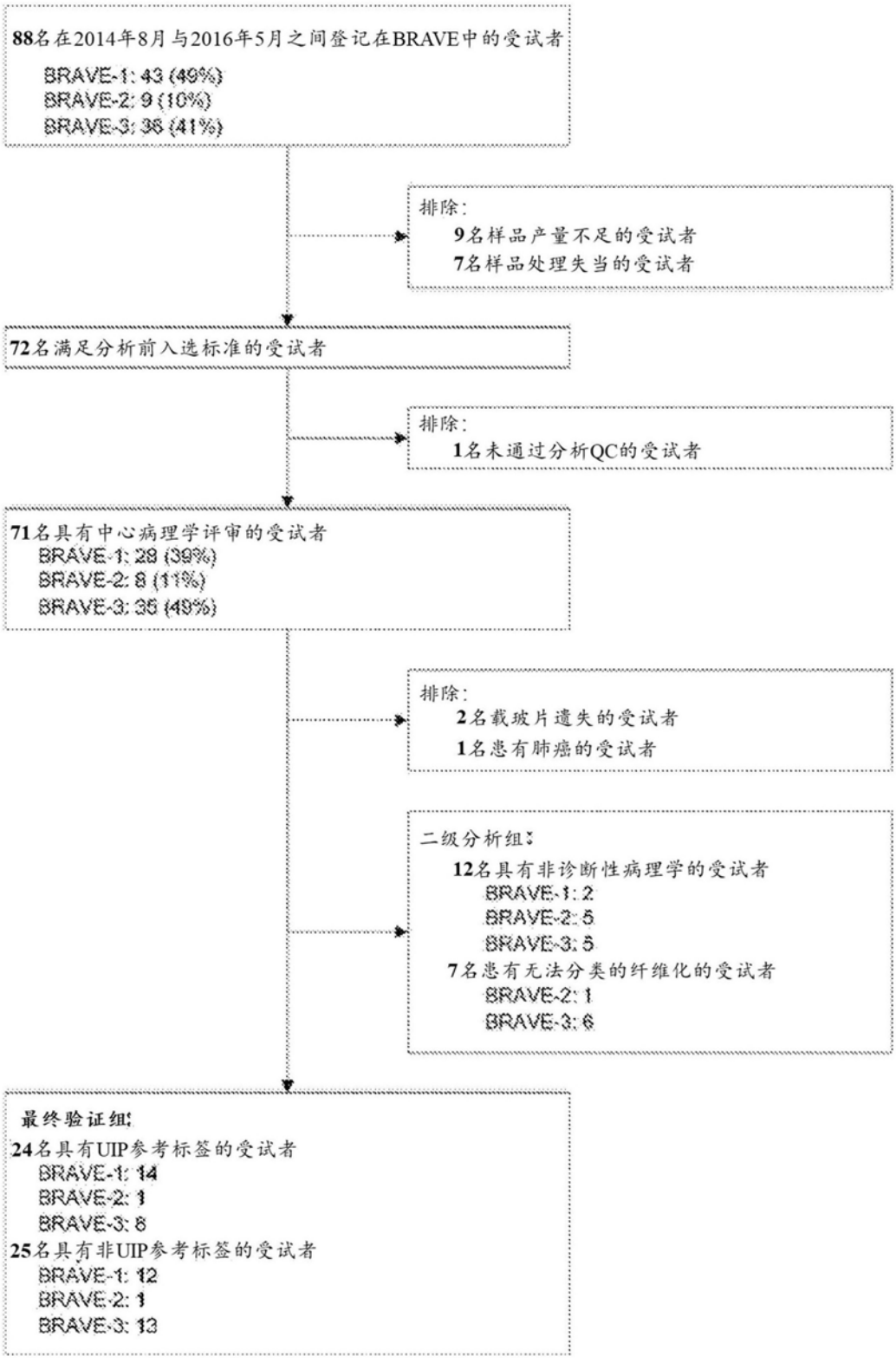

88名在2014年8月与2016年5月之间登记在BRAVE中的受试者
BRAVE-1: 43 (49%)
BRAVE-2: 9 (10%)
BRAVE-3: 36 (41%)
排除:
9名样品产量不足的受试者
7名样品处理失当的受试者
72名满足分析前入选标准的受试者
排除:
1名未通过分析QC的受试者
71名具有中心病理学评审的受试者
BRAVE-1: 28 (39%)
BRAVE-2: 8 (11%)
BRAVE-3: 35 (49%)
排除:
2名载玻片遗失的受试者
1名患有肺癌的受试者
二级分析组:
12名具有非诊断性病理学的受试者
BRAVE-1: 2
BRAVE-2: 5
BRAVE-3: 5
7名患有无法分类的纤维化的受试者
BRAVE-2: 1
BRAVE-3: 6
最终验证组:
24名具有UIP参考标签的受试者
BRAVE-1: 14
BRAVE-2: 1
BRAVE-3: 8
25名具有非UIP参考标签的受试者
BRAVE-1: 12
BRAVE-2: 1
BRAVE-3: 13

图8

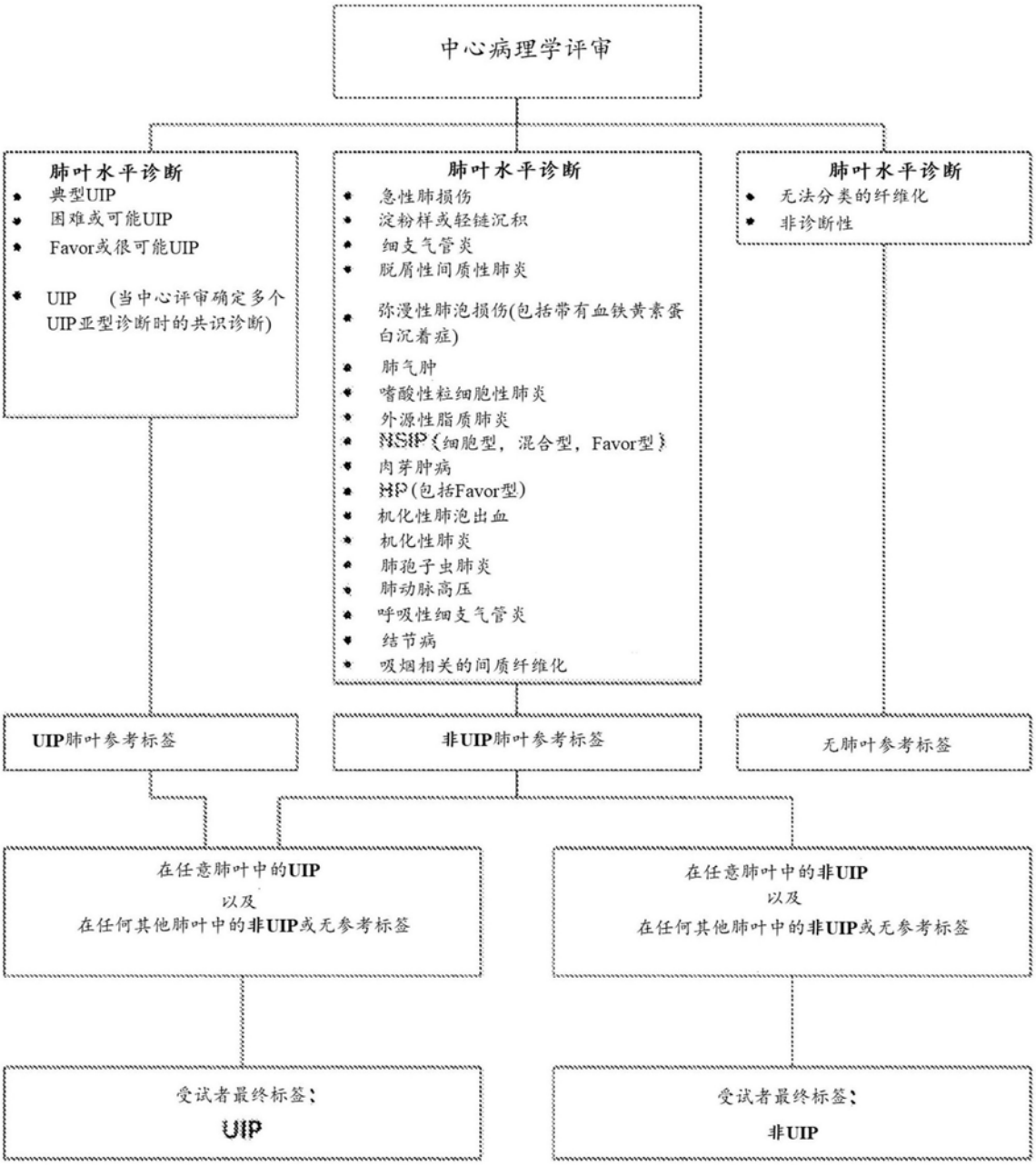

中心病理学评审
肺叶水平诊断
典型UIP
困难或可能UIP
Favor或很可能UIP
UIP (当中心评审确定多个UIP亚型诊断时的共识诊断)
肺叶水平诊断
急性肺损伤
淀粉样或轻链沉积
细支气管炎
脱屑性间质性肺炎
弥漫性肺泡损伤(包括带有血铁黄素蛋白沉着症)
肺气肿
嗜酸性粒细胞性肺炎
外源性脂质肺炎
NSIP(细胞型，混合型，Favor型)
肉芽肿病
HP(包括Favor型)
机化性肺泡出血
机化性肺炎
肺孢子虫肺炎
肺动脉高压
呼吸性细支气管炎
结节病
吸烟相关的间质纤维化
肺叶水平诊断
无法分类的纤维化
非诊断性
UIP肺叶参考标签
非UIP肺叶参考标签
无肺叶参考标签
在任意肺叶中的UIP
以及
在任何其他肺叶中的非UIP或无参考标签
在任意肺叶中的非UIP
以及
在任何其他肺叶中的非UIP或无参考标签
受试者最终标签：
UIP
受试者最终标签：
非UIP

图9

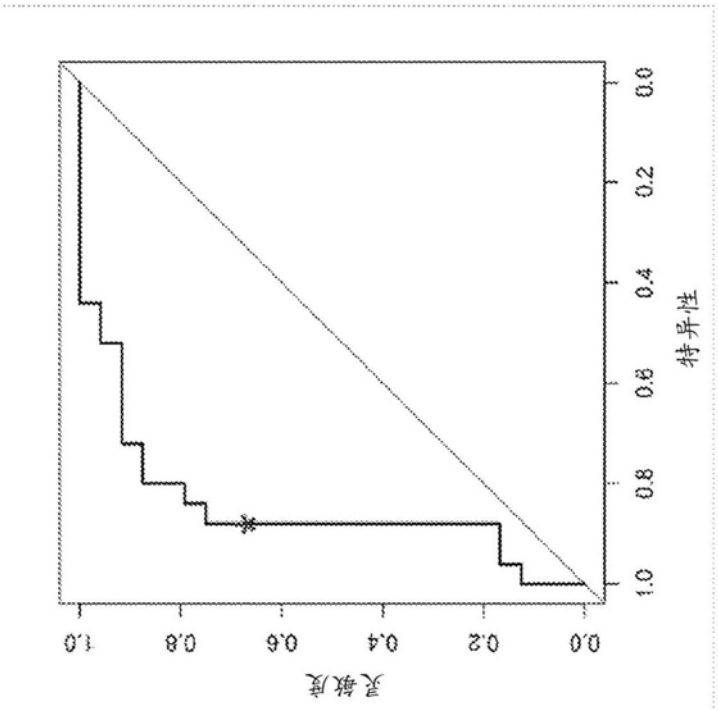


图10A

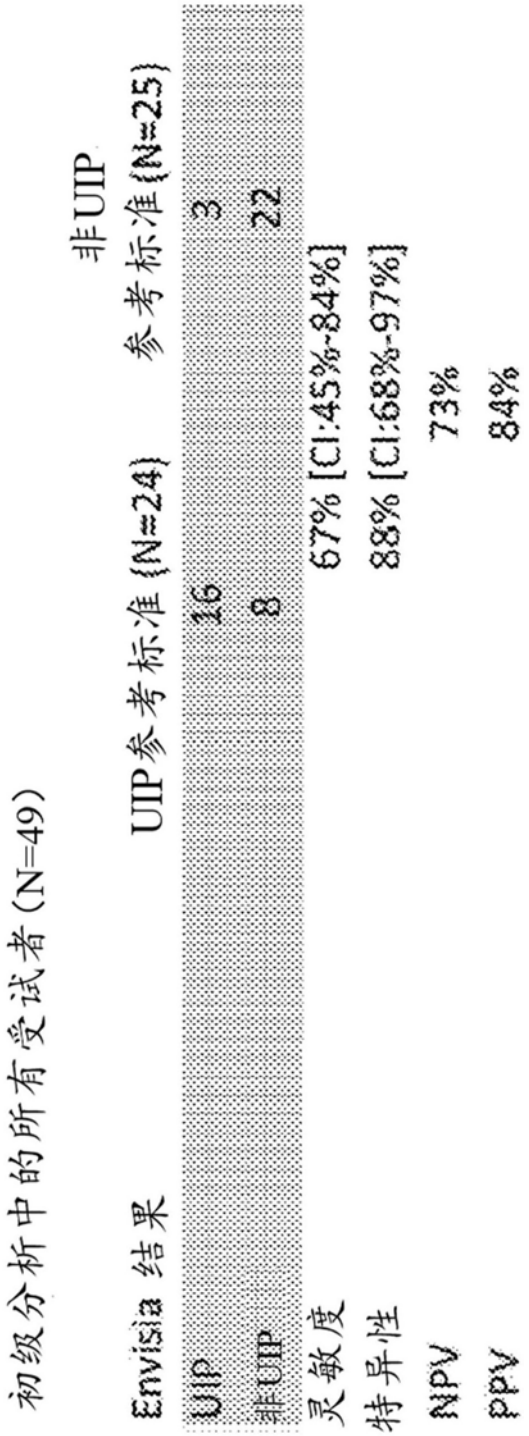

初级分析中的所有受试者（N=49）
非UIP
Envisia 结果
UIP参考标准{N=24}
参考标准{N=25}
UIP
16
3
非UIP
8
22
灵敏度
67% [CI:45%-84%]
特异性
88% [CI:68%-97%]
NPV
73%
PPV
84%

图10B

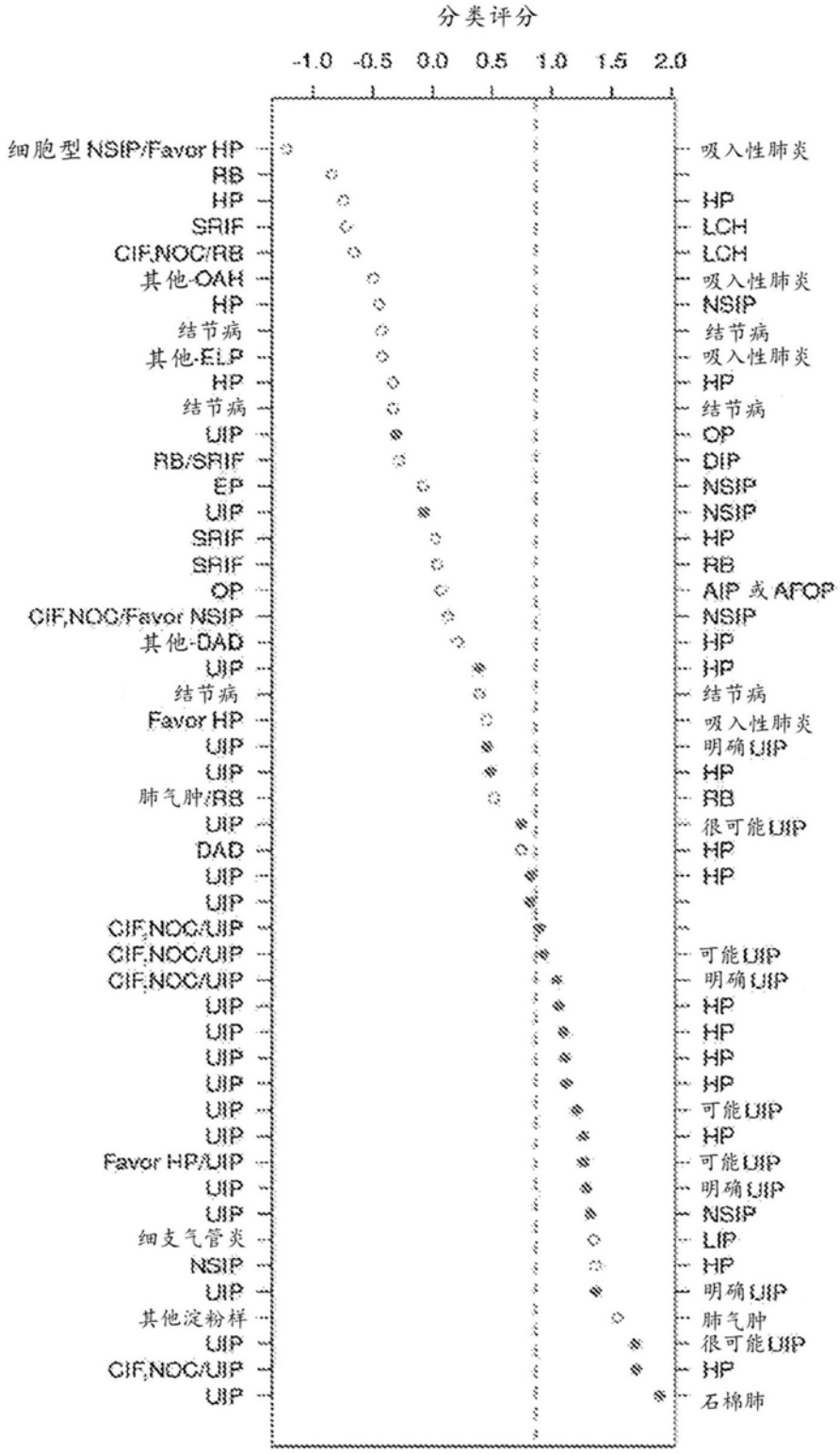

分类评分
-1.0 -0.5 0.0 0.5 1.0 1.5 2.0
细胞型 NSIP/Favor HP
RB
HP
SRIF
CIF,NOC/RB
其他-OAH
HP
结节病
其他-ELP
HP
结节病
UIP
RB/SRIF
EP
UIP
SRIF
SRIF
OP
CIF,NOC/Favor NSIP
其他-DAD
UIP
结节病
Favor HP
UIP
UIP
肺气肿/RB
UIP
DAD
UIP
UIP
CIF,NOC/UIP
CIF,NOC/UIP
CIF,NOC/UIP
UIP
UIP
UIP
UIP
UIP
UIP
Favor HP/UIP
UIP
UIP
细支气管炎
NSIP
UIP
其他淀粉样
UIP
CIF,NOC/UIP
UIP
吸入性肺炎
HP
LCH
LCH
吸入性肺炎
NSIP
结节病
吸入性肺炎
HP
结节病
OP
DIP
NSIP
NSIP
HP
RB
AIP 或 AFOP
NSIP
HP
HP
结节病
吸入性肺炎
明确UIP
HP
RB
很可能UIP
HP
HP
可能UIP
明确UIP
HP
HP
HP
HP
可能UIP
HP
可能UIP
明确UIP
NSIP
LIP
HP
明确UIP
肺气肿
很可能UIP
HP
石棉肺

图11

中心放射学结果	UIP参考标准 (N=22)	非UIP 参考标准 (N=24)
明确/很可能/可能UIP	9	0
与UIP不一致	13	24
灵敏度	41% [21-64%]	
特异性	100% [86-100%]	
NPV	65% [47-80%]	
PPV	100% [66-100%]	

局部放射学结果	UIP参考标准 (N=23)	非UIP 参考标准 (N=23)
明确/很可能/可能UIP	14	7
与UIP不一致	9	16
灵敏度	61% [39-80%]	
特异性	70% [47-87%]	
NPV	64% [43-82%]	
PPV	67% [43-85%]	

通过中心放射学确定的患有明确/很可能/可能UIP的受试者

Envisia 结果	UIP参考标准 (N=9)	非UIP 参考标准 (N=0)
UIP	7	0
非UIP	2	0
灵敏度	78% [40-97%]	
特异性	N/A	
NPV	N/A	
PPV	100% [59-100%]	

通过局部放射学确定的患有明确/很可能/可能UIP的受试者

Envisia 结果	UIP参考标准 (N=14)	非UIP 参考标准 (N=7)
UIP	8	3
非UIP	6	4
灵敏度	57% [29-82%]	
特异性	57% [18-90%]	
NPV	40% [12-74%]	
PPV	73% [39-94%]	

通过中心放射学确定的患有不一致UIP的受试者

Envisia 结果	UIP参考标准 (N=13)	非UIP 参考标准 (N=24)
UIP	8	3
非UIP	5	21
灵敏度	62% [32-86%]	
特异性	88% [68-97%]	
NPV	81% [61-93%]	
PPV	73% [39-94%]	

通过局部放射学确定的患有不一致UIP的受试者

Envisia 结果	UIP参考标准 (N=9)	非UIP 参考标准 (N=16)
UIP	7	0
非UIP	2	16
灵敏度	78% [40-97%]	
特异性	100% [79-100%]	
NPV	89% [65-99%]	
PPV	100% [59-100%]	

图12

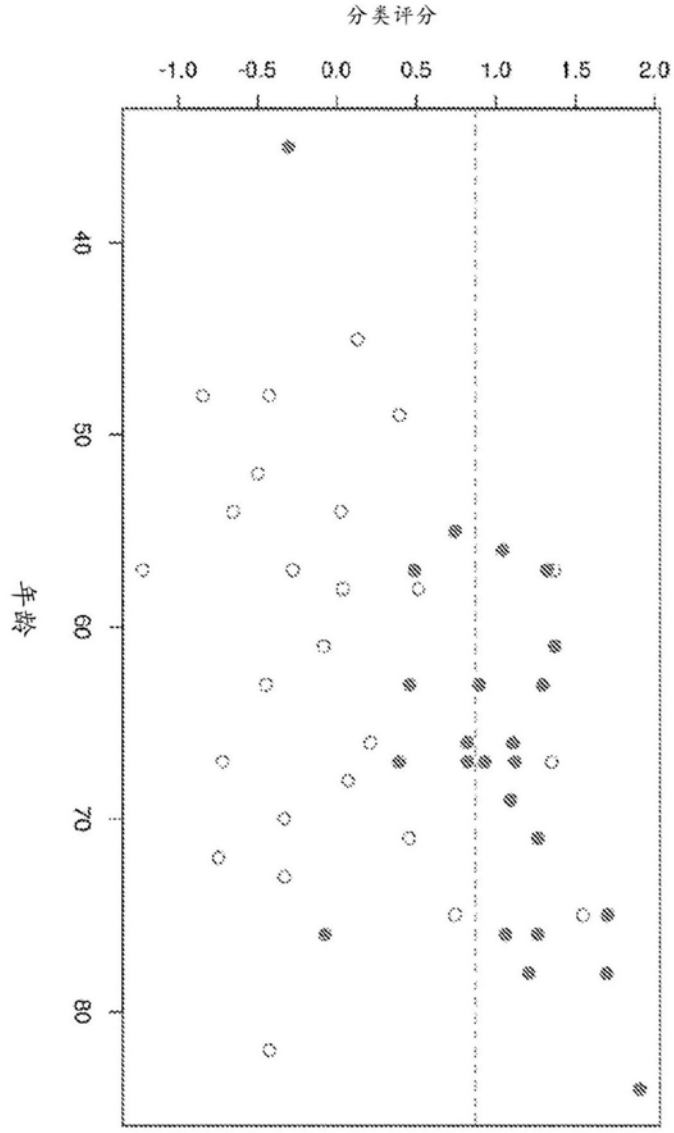
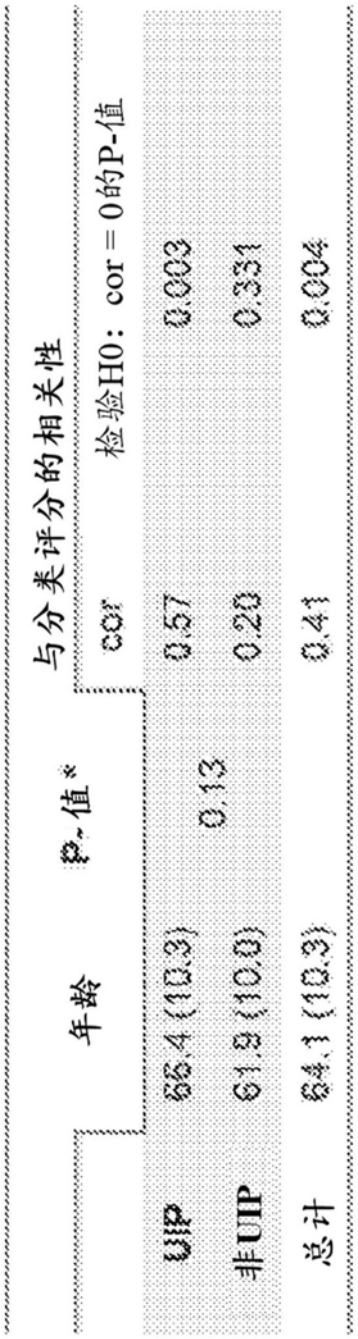


图13A



	年龄	P-值*	与分类评分的相关性	
			cor	检验H0：cor = 0的P-值
UIP	65.4 (10.3)	0.13	0.57	0.003
非UIP	61.9 (10.0)		0.20	0.331
总计	64.1 (10.3)		0.41	0.004

图13B

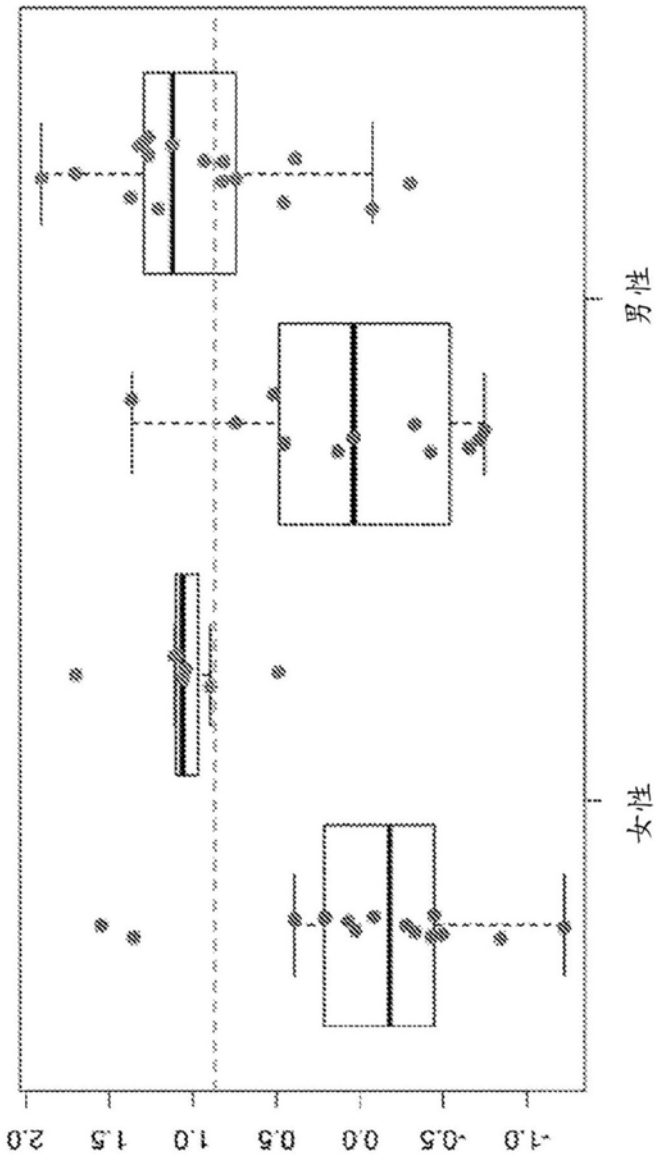


图13C

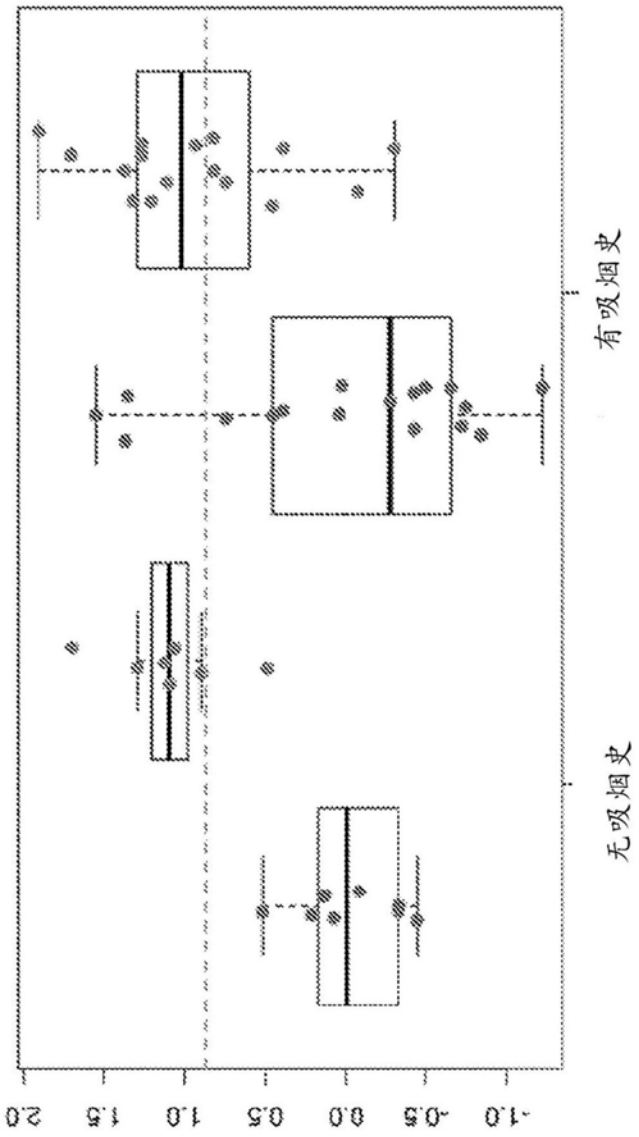


图13D

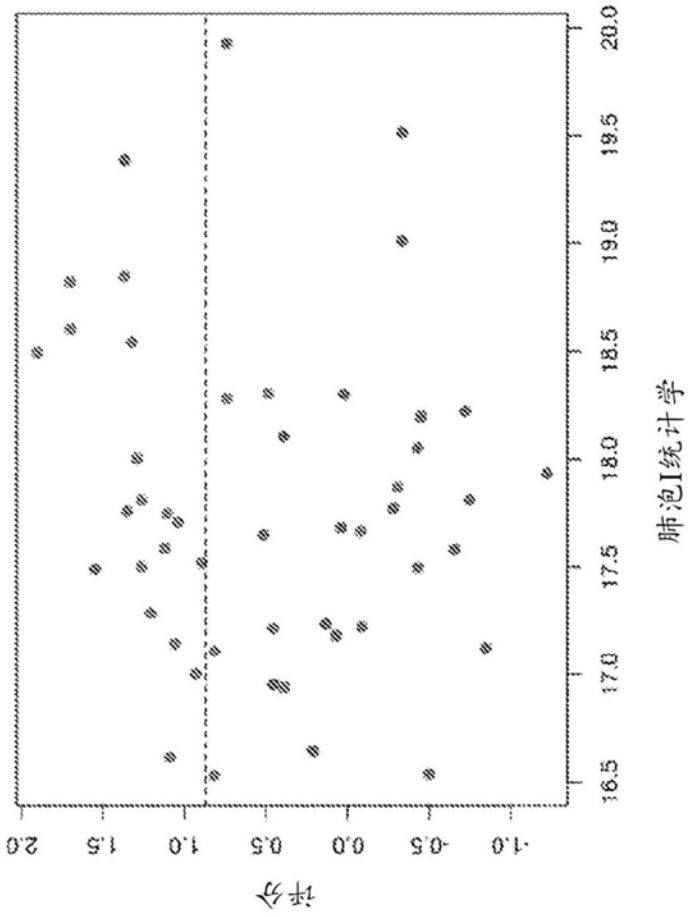


图14A

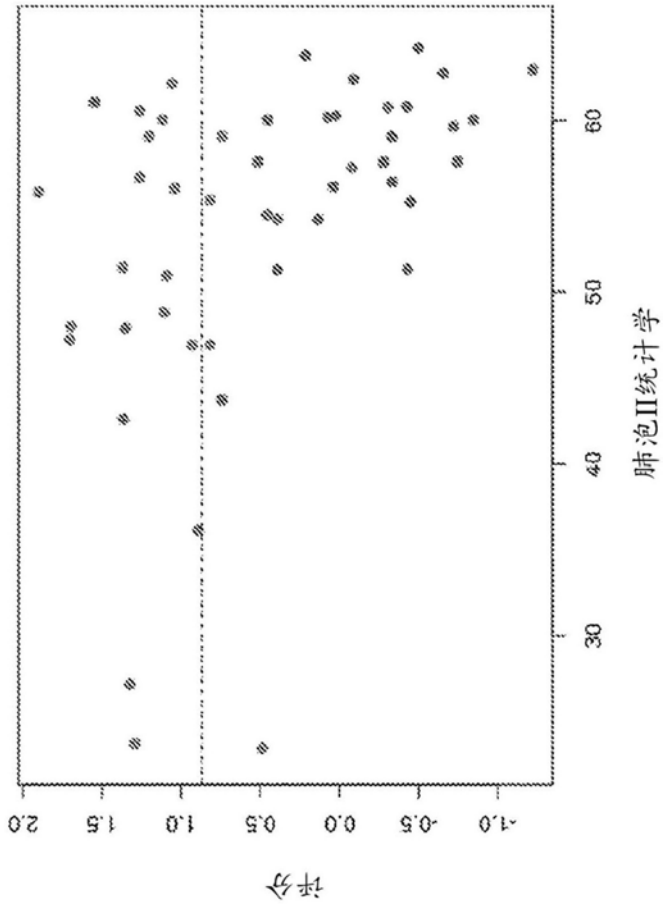


图14B

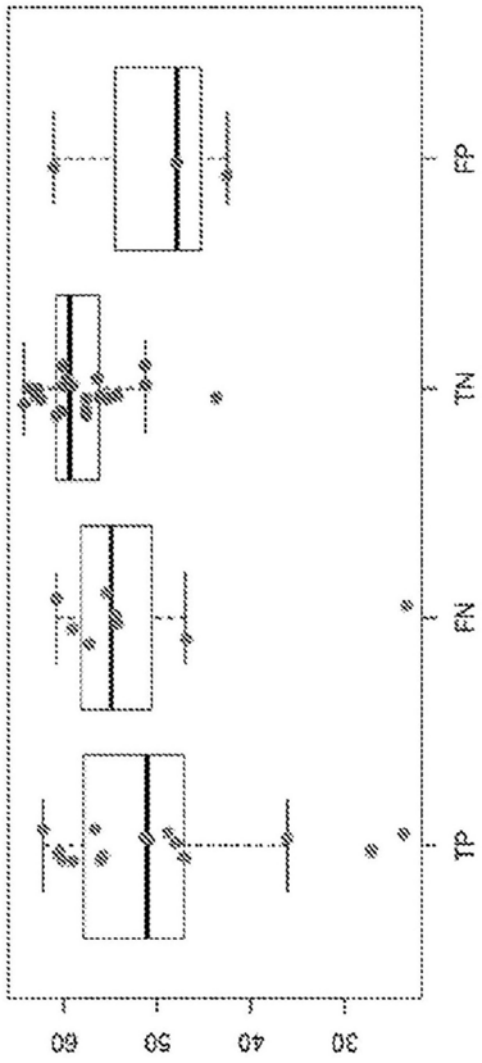


图14C

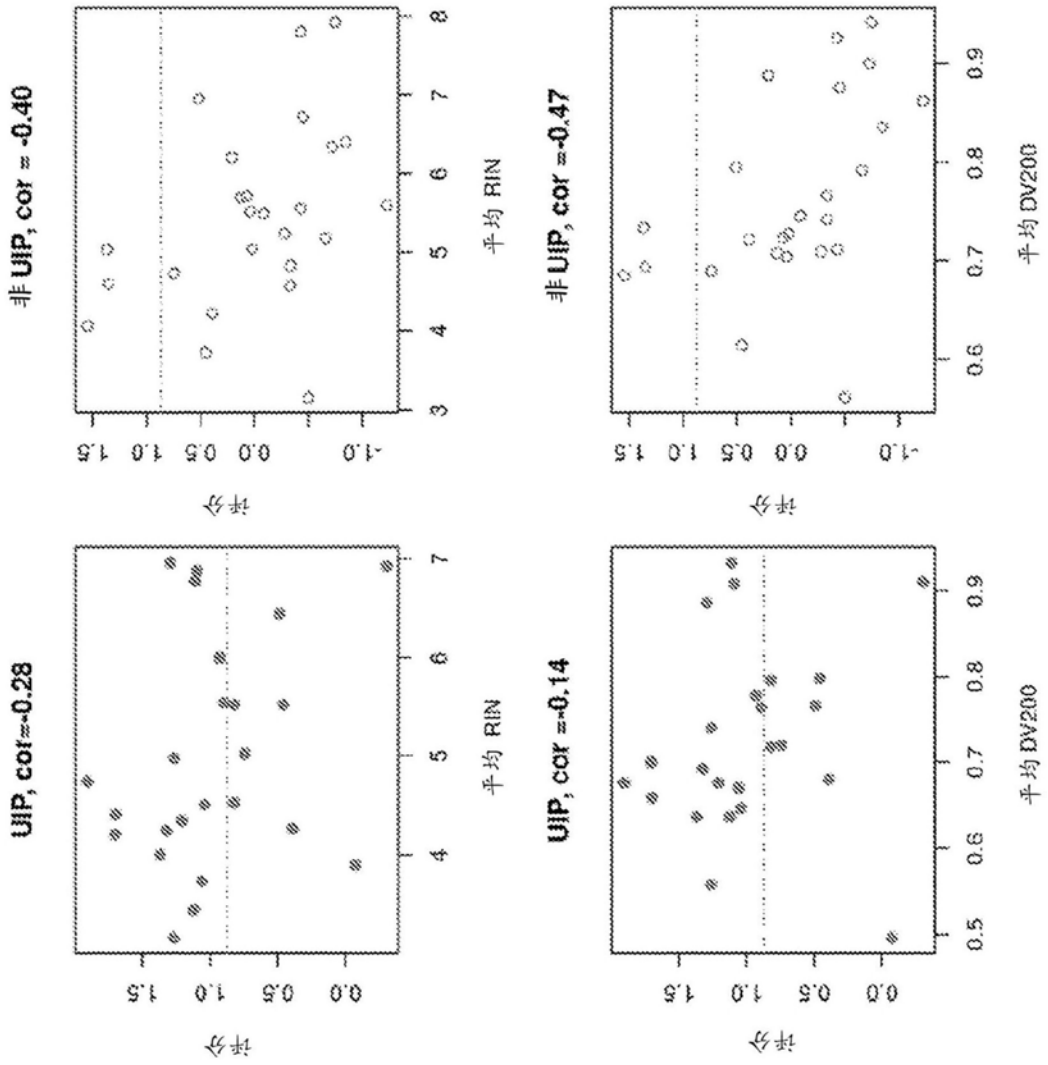

非UIP, cor = -0.40
UIP, cor=-0.28
非UIP, cor =-0.47
UIP, cor =-0.14
评分
平均 RIN
平均 DV200

图14D