



(12) 发明专利申请

(10) 申请公布号 CN 114634985 A

(43) 申请公布日 2022. 06. 17

(21) 申请号 202210267696.9

(22) 申请日 2016.03.03

(30) 优先权数据

62/128,463 2015.03.04 US

62/128,469 2015.03.04 US

62/238,893 2015.10.08 US

(62) 分案原申请数据

201680026050.4 2016.03.03

(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 M·帕甘 林竹芳

黄静 P·肖恩·沃尔什 松崎新

凯文·特拉弗斯 金洙延

(74) 专利代理机构 北京安信方达知识产权代理有限公司 11262

专利代理师 贺淑东 武晶晶

(51) Int.Cl.

C12Q 1/6886 (2018.01)

G16B 30/10 (2019.01)

G16H 50/30 (2018.01)

权利要求书2页 说明书29页 附图50页

(54) 发明名称

使用表达水平和序列变体信息评估疾病发生或复发风险的方法

(57) 摘要

本文提供了通过组合两个或更多个特征空间对从受试者获得的样品的疾病发生风险进行分层从而改善受试者管理个体化的方法、系统和试剂盒。

各风险组的性能

	中/高风险 (N=50)	低风险 (N=29)
分类为 “中/高风险”	41	6
分类为 “低风险”	9	23
灵敏度	82%	
特异性	79%	

途径或 基因本体	预期的 基因数	观察到的 基因数	FDR P值
细胞外基质	2	20	$1.28 \times 10^{-13}$
ECM受体 相互作用	1	11	$6.12 \times 10^{-08}$
粘着斑	2	12	$3.71 \times 10^{-05}$
酪氨酸 激酶活性	1	6	$4.77 \times 10^{-03}$
免疫系统 过程的调控	3	10	$6.83 \times 10^{-02}$
血管发育	1	7	$1.0 \times 10^{-02}$

用于分类的基因的抽样				
COX6C	FANCA	KCTD17	MPRIIP	TUBA1B
DKAKD	ICE2	MCM3AP	TNFRSF14	WSB2

1. 一种用于评估受试者的组织样品以确定所述受试者的疾病发生风险的方法,该方法包括:

(a) 获得与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平,所述第一组基因与所述受试者的所述疾病发生风险相关;

(b) 确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在,所述第二组基因与所述受试者的所述疾病发生风险相关;

(c) 分别将 (i) 在 (a) 中获得的所述表达水平以及 (ii) 在 (b) 中获得的所述核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与所述对照的比较,其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在;以及

(d) 使用采用经训练的算法编程的计算机处理器, (i) 分析所述比较,以及 (ii) 基于所述比较确定所述疾病的所述发生风险。

2. 根据权利要求1所述的方法,其中所述疾病为癌症。

3. 根据权利要求1所述的方法,其进一步包括在 (a) 之前,从所述受试者获得所述针抽吸样品。

4. 根据权利要求1所述的方法,其进一步包括在 (a) 之前,从所述针抽吸样品中的所述核酸样品确定所述表达水平。

5. 根据权利要求1所述的方法,其进一步包括在 (b) 之前,从所述针抽吸样品中的所述核酸样品确定所述核酸序列。

6. 根据权利要求5所述的方法,其进一步包括将所述核酸序列与所述参考序列进行比较,以鉴定所述一个或多个序列变体。

7. 根据权利要求6所述的方法,其中所述参考序列为来自所述受试者的管家基因。

8. 根据权利要求1所述的方法,其中所述第一组或第二组基因中的所述一个或多个基因包括多个基因。

9. 一种用于评估受试者的组织样品以确定所述受试者的疾病发生风险的系统,该系统包括:

一个或多个计算机存储器,该计算机存储器存储 (a) 与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达,所述第一组基因与所述受试者的所述疾病发生风险相关,以及 (b) 与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在的指示,所述第二组基因与所述受试者的所述疾病发生风险相关;以及

计算机处理器,其与所述一个或多个计算机存储器耦合并被编程为:

(i) 分别将 (1) 所述计算机存储器中的所述表达水平以及 (2) 所述核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与所述对照的比较,其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在;以及

(ii) 使用经训练的算法 (1) 分析所述比较以及 (2) 基于所述比较确定所述疾病的所述发生风险。

10. 一种非暂时性计算机可读介质,其包含机器可执行代码,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用于评估受试者的组织样品以确定所述受试者的疾病发生风险的方法,所述方法包括:

(a) 获得与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平,所述第一组基因与所述受试者的所述疾病发生风险相关;

(b) 确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在,所述第二组基因与所述受试者的所述疾病发生风险相关;

(c) 分别将 (i) 在 (a) 中获得的所述表达水平以及 (ii) 在 (b) 中获得的所述核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与所述对照的比较,其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在;以及

(d) 使用采用经训练的算法编程的计算机处理器, (i) 分析所述比较,以及 (ii) 基于所述比较确定所述疾病的所述发生风险。

## 使用表达水平和序列变体信息评估疾病发生或复发风险的方法

本申请是申请日为2016年03月03日、申请号为201680026050.4、发明名称为“使用表达水平和序列变体信息评估疾病发生或复发风险的方法”的中国专利申请(其对应PCT申请的申请日为2016年03月03日、申请号为PCT/US2016/020583)的分案申请。

### 交叉引用

[0001] 本申请要求于2015年3月4日提交的美国临时申请62/128,463、于2015年3月4日提交的美国临时申请62/128,469和于2015年10月8日提交的美国临时申请62/238,893的优先权,每个申请均通过引用而整体并入本文。

### 背景技术

[0002] 疾病治疗如甲状腺癌治疗的风险调整方法,除了改善疾病特异性生存之外,还可使疾病发生风险减至最小。目前,这种用于初始受试者管理的风险调整方法在很大程度上基于采用2009年美国甲状腺协会(ATA)分级系统进行的将受试者分类为高、中或低疾病复发风险的术后分类。虽然这种解剖学分级系统已被证明在临床上有用,但其不能在侵入性甲状腺切除术之前得到准确的评估,并且不包括疾病结果的任何分子预测因子。

### 发明内容

[0003] 本文提供了用于对疾病发生和/或复发的风险进行评估或分层的各种方法。在预诊断或诊断性评估如细针抽吸(FNA)期间获得的转录数据可改善疾病如甲状腺癌发生的术前风险预测,并且可进一步提供受试者疗法和治疗的个体化。本公开内容的方法可以以相对非侵入性方式以及使用低样品体积来提供针对于疾病发生和/或复发风险的评估。

[0004] 本公开内容的一个方面提供了一种用于评估受试者的组织样品以确定所述受试者的疾病发生风险的方法。所述方法包括(a)获得与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平,所述第一组基因与所述受试者的所述疾病发生风险相关;(b)确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在,所述第二组基因与所述受试者的所述疾病发生风险相关;(c)分别将(i)在(a)中获得的所述表达水平以及(ii)在(b)中获得的所述核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与所述对照的比较,其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在;以及(d)使用采用经训练的算法编程的计算机处理器,(i)分析所述比较,以及(ii)基于所述比较确定所述疾病的发生风险。

[0005] 在一些实施方案中,所述针抽吸样品为细针抽吸样品。在一些实施方案中,所述疾病为癌症。在一些实施方案中,所述方法进一步包括在(a)之前,从所述受试者获得所述针抽吸样品。在一些实施方案中,所述方法进一步包括在(a)之前,从所述针抽吸样品中的所述核酸样品确定所述表达水平。在一些实施方案中,所述方法进一步包括在(b)之前,从所述针抽吸样品中的所述核酸样品确定所述核酸序列。在一些实施方案中,所述方法进一步



包括将所述核酸序列与所述参考序列进行比较,以鉴定所述一个或多个序列变体。在一些实施方案中,所述参考序列为来自所述受试者的管家基因。在一些实施方案中,所述第一组或第二组基因中的所述一个或多个基因包括多个基因。

[0006] 在一些实施方案中,所述针抽吸样品已被发现是细胞学上不明确的或可疑的。在一些实施方案中,所述针抽吸样品具有约1微升或更小的体积。在一些实施方案中,所述针抽吸样品具有约9.0或更小的RNA完整性指数(RIN)值。在一些实施方案中,从针抽吸样品纯化的RNA具有约9.0或更小的RNA RIN值。在一些实施方案中,所述针抽吸样品具有约6.0或更小的RIN值。在一些实施方案中,所述RNA样品具有约6.0或更小的RIN值。

[0007] 在一些实施方案中,所述疾病的发生风险包括所述受试者的所述疾病的复发风险。在一些实施方案中,所述癌症的发生风险包括在所述受试者中转移的风险。在一些实施方案中,所述癌症的发生风险包括疾病进展加速的风险。在一些实施方案中,所述癌症的发生风险包括治疗失败的风险。

[0008] 在一些实施方案中,所述经训练的算法采用来自至少25名或至少100名已被诊断为患有所述疾病的受试者的组织样品进行训练。在一些实施方案中,所述经训练的算法采用来自至少200名已被诊断为患有所述疾病的受试者的组织样品进行训练。

[0009] 在一些实施方案中,(d)发生在术前。在一些实施方案中,(d)发生在所述受试者具有阳性疾病诊断之前。在一些实施方案中,(d)进一步包括将所述发生风险分层为低发生风险或中高发生风险,其中所述低发生风险具有约50%至约80%的发生概率,并且其中所述中高发生风险具有约80%至100%的发生概率。

[0010] 在一些实施方案中,所述方法进一步包括将一个或多个过滤器、一个或多个包装器、一个或多个嵌入式协议或其任意组合应用于所述比较。在一些实施方案中,将所述一个或多个过滤器应用于所述比较。在一些实施方案中,所述一个或多个过滤器包括t检验、方差分析(ANOVA)、贝叶斯框架、伽马分布、Wilcoxon秩和检验、类间内平方和检验、秩乘法、随机置换法、误分类数目阈值(TNoM)、双变量法、基于相关性的特征选择(CFS)法、最小冗余最大相关性(MRMR)法、Markov毯过滤法、非相关收缩重心法(uncorrelated shrunken centroid method)或其任意组合。在一些实施方案中,所述一个或多个序列变体包含点突变、融合基因、置换、缺失、插入、倒位、转变(conversion)、易位或其任意组合中的一种或多种。在一些实施方案中,所述一个或多个点突变为约5个至约4000个点突变。在一些实施方案中,所述一个或多个融合基因为至少两个融合基因。

[0011] 在一些实施方案中,所述分层具有约80%的准确度。在一些实施方案中,所述分层具有约80%的特异性。在一些实施方案中,所述第一组或第二组的所述一个或多个基因为少于约15个基因或少于约10个基因。在一些实施方案中,所述第一组或第二组的所述一个或多个基因为少于约75个基因。在一些实施方案中,所述第一组或第二组的所述一个或多个基因为约50个至约400个基因。

[0012] 在一些实施方案中,(b)中的获得包括对所述针抽吸样品中的核酸样品进行测序以获得核酸序列。在一些实施方案中,所述测序包括富集第二组基因的所述一个或多个基因或其变体。在一些实施方案中,(a)包括使用具有对所述第一组基因的所述一个或多个基因具有选择性的探针的微阵列。在一些实施方案中,(a)包括使用靶向测序平台(如Ion Torrent Ampliseq或Illumina TruSeq Custom Amplicon)。

[0013] 在一些实施方案中,所述组织样品为甲状腺组织样品。在一些实施方案中,所述第一组和第二组基因包含COL1A1、THBS2或其任意组合。在一些实施方案中,所述第二组基因包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROBO1、C6orf136、SPAG4、GALNT15、LUM、NCAM2、NUP210L、NR2F1、THBS2、PSORS1C1或其任意组合。在一些实施方案中,所述第一组基因包含COL1A1、TMEM92、C1orf87、SPAG4、EHF、COL3A1、GALNT15、NUP210L、PDZRN3、C6orf136、NA、NRXN3、COL6A3、RAPGEF5、PRICKLE1、LUM、ROBO1、BGN、AC019117.2、PRSS3P1或其任意组合。在一些实施方案中,所述第二组基因包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROBO1、C6orf136、SPAG4、GALNT15、LUM、NCAM2、SYNP02、NUP210L、AMZ1、NR2F1、THBS2、PSORS1C1、FTH1P24或其任意组合。在一些实施方案中,所述第二组基因包含AKAP9、SPRY3、SPRY3、CAMKK2、COL1A1、FITM2、COX6C、VSIG10L、CYC1、KDM1B、MAPK15、ARSG、PAXIP1、DAAM1、AVL9、DMGDH、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-DRB5、HLA-H、IRF1、MGAT1、P2RX1、PLEK、CCDC93、PPP1R12C、SLC41A3、METTL3、CCAR2、PTPRE、SRL、SLC30A5、BMP4、ZNF133、ICE2、DCAKD、TMX1、TNFSF12、PER2、MCM3AP或其任意组合。

[0014] 在一些实施方案中,所述第一组基因和所述第二组基因是不同的。在一些实施方案中,所述方法进一步包括鉴定所述疾病的新型遗传生物标志物。

[0015] 在一些实施方案中,(a)中的获得包括测定与所述一个或多个基因中的每一个相对应的表达水平。在一些实施方案中,所述测定包括使用针对所述一个或多个基因中的每一个所选择的标志物进行的阵列杂交、核酸测序或核酸扩增。在一些实施方案中,所述标志物为针对所述一个或多个基因中的每一个所选择的引物。

[0016] 在一些实施方案中,所述测定包括逆转录聚合酶链反应(PCR)。在一些实施方案中,所述确定包括测定所述核酸样品中所述第二组基因的所述一个或多个基因中的每一个。在一些实施方案中,所述测定包括使用针对所述一个或多个基因中的每一个所选择的标志物进行的阵列杂交、核酸测序或核酸扩增。在一些实施方案中,所述标志物为针对所述一个或多个基因中的每一个所选择的引物。在一些实施方案中,所述测定包括逆转录聚合酶链反应(PCR)。

[0017] 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计算机可读介质(例如,存储器),该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。

[0018] 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计算机可读介质的计算机系统。所述计算机可读介质可包含机器可执行代码,该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。

[0019] 通过下面的发明详述,本公开内容另外的方面和优点将变得对本领域技术人员而言显而易见,发明详述中仅示出并描述了本公开内容的说明性实施方案。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中进行修改,所有这些都脱离本公开内容。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

[0020] 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

[0021] 本发明的新特征在附加的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详细说明和附图(本文中也称为“图”),将会获得对本发明的特征和优点的更好的理解,附图中:

[0022] 图1示出了被分层为低癌症发生风险和中高癌症发生风险的样品群组的细胞学数据和专家组织病理学数据;

[0023] 图2示出了组织病理学风险特征以及每个特征的样品数目和百分比;

[0024] 图3示出了以真阳性率针对假阳性率绘图的交叉验证;

[0025] 图4示出了绘制预测值针对中高风险发生率的分类性能数据;

[0026] 图5示出了低风险组和中高风险组中的分类性能数据;

[0027] 图6示出了基于基因表达水平数据与甲状腺癌发生风险相关的基因的示例性列表;

[0028] 图7示出了基于从核糖核酸(RNA)测序获得的基因表达水平数据与甲状腺癌发生风险相关的基因的示例性列表;

[0029] 图8示出了基于序列变体数据与甲状腺癌发生风险相关的基因的示例性列表;

[0030] 图9示出了被编程或以其他方式配置为实现本文提供的方法的计算机控制系统;

[0031] 图10示出了确定准确的训练标记的流程图;

[0032] 图11A示出了以真阳性率针对假阳性率绘图的交叉验证;

[0033] 图11B示出了中/高风险组和低风险组中的分类性能数据;

[0034] 图12示出了由分类器以每种倍数选择的变体的基因的示例性列表;

[0035] 图13示出了由分类器以10倍选择8至10次的计数的基因的示例性列表;

[0036] 图14示出了五个点突变组和融合对的表格;

[0037] 图15示出了五组突变和融合对中的测试性能特异性和灵敏度的图;

[0038] 图16示出了按照细胞学,图14和图15中组3的突变性能的表格;

[0039] 图17示出了五组突变和融合对中的测试性能特异性和灵敏度的图;

[0040] 图18A示出了表示临床实验室改进修正(CLIA)细针抽吸(FNA)样品的突变频率的图示;图18B示出了其表格;

[0041] 图19A示出了表示FNA样品的突变频率的图示;图19B示出了其表格;以及

[0042] 图20A示出了表示组织样品的突变频率的图示;图20B示出了其表格。

## 具体实施方式

[0043] 尽管本文已经显示并描述了本发明的不同的实施方案,但对本领域技术人员而言显而易见的是,这些方案仅以举例的方式提供。在不脱离本发明的情况下,本领域技术人员可能想到许多变化、改变和替代。应理解,可采用本文描述的本发明实施方案的各种替代方案。

[0044] 如本文所用的术语“受试者”通常是指任何动物或活生物体。动物可以是哺乳动物,如人类、非人类灵长类动物、啮齿动物如小鼠和大鼠、狗、猫、猪、羊、兔及其他动物。动物可以是鱼、爬行动物或其他动物。动物可以是新生、幼体(infant)、成年前(adolescent)或成年动物。人类可以大于约1、2、5、10、20、30、40、50、60、65、70、75岁或约80岁。受试者可以患有或疑似患有疾病,如癌症。受试者可以是患者,如接受疾病治疗的患者,如癌症患者。受试者可能容易具有发展疾病如癌症的风险。受试者可以处于疾病缓解中,如癌症患者。受试者可以是健康的。

[0045] 如本文所用的术语“疾病”通常是指影响受试者的任何异常或病理状况。疾病的实例包括癌症,例如,甲状腺癌、甲状旁腺癌、肺癌、皮肤癌及其他癌症。疾病可以是可治疗的或不可治疗的。疾病可以是晚期的或非晚期的。疾病可以由遗传基因、环境暴露或其任意组合引起。疾病可以是癌症、遗传病、增生性病或如本文所述的其他疾病。

[0046] 如本文所定义的术语“疾病发生风险”通常是指与受试者的疾病发生相关的风险或概率。发生风险可包括疾病在受试者中首次发生,或者可以包括随后的发生,如第二次、第三次、第四次或随后发生。疾病发生风险可包括:a)首次发展疾病的风险,b)复发或再次发展疾病的风险,c)将来发展疾病的风险,d)在受试者的一生中易于发展疾病的风险,或e)作为幼体、成年前或成年易于发展疾病的风险。疾病如癌症发生的风险可包括癌症变为转移性的风险。疾病如癌症发生的风险可包括I期癌症、II期癌症、III期癌症或IV期癌症发生的风险。癌症发生风险可包括血癌、组织癌(例如肿瘤)发生的风险,或癌症从其他部位转移到一个或多个器官部位的风险。

[0047] 如本文所用的术语“序列变体”、“序列变异”、“序列改变”或“等位基因变体”通常是指与参考序列,如基因组脱氧核糖核酸(DNA)参考序列、编码DNA参考序列或蛋白质参考序列或其他参考序列有关的特定变化或变异。可从参考数据库获得参考DNA序列。序列变体可能影响功能。序列变体可能不影响功能。序列变体可在DNA水平下在一个或多个核苷酸、核糖核酸(RNA)水平下在一个或多个核苷酸、蛋白质水平下在一个或多个氨基酸或其任意组合中发生。可从数据库如NCBI参考序列数据库(RefSeq)获得参考序列。可构成序列变异的特定变化可包括一个或多个核苷酸或者一个或多个氨基酸中的置换、缺失、插入、倒位或转变。序列变体可以是点突变。序列变体可以是融合基因。融合对或融合基因可由序列变体,如易位、中间缺失(interstitial deletion)、染色体倒位或其任意组合产生。序列变异可构成重复序列数目的变异性,如三重复制、四重复制或其他数目复制。例如,序列变异可以是与给定序列相关的拷贝数的增加或减少(即拷贝数变异或CNV)。序列变异可包括不同等位基因中的两个或更多个序列变化或一个等位基因中的两个或更多个序列变化。序列变异可包括在一个等位基因中的一个位置处的两个不同的核苷酸,如镶嵌体。序列变异可包括在一个等位基因中的一个位置处的两个不同的核苷酸,如嵌合体。序列变体可存在于恶性组织中。序列变体可存在于良性组织中。变体的不存在可指示组织或样品是良性的。作为替代,变体的不存在可能不指示组织或样品是良性的。

[0048] 如本文所用的术语“突变组”通常是指指定用风险分类器进行检测(或询问)的特定数目的基因组位点和融合对的组。例如,突变组可包含待询问的9个基因组位点和3个融合对。通过增加检测的点突变和融合对的数目来增加风险分类器的灵敏度可降低风险分类器的灵敏度。

[0049] 突变组可包含一个或多个基因组位点和一个或多个融合对。突变组可包含多于约1、2、3、4或5个基因组位点。突变组可包含多于约15个基因组位点。突变组可包含多于约100个基因组位点。突变组可包含多于约200个基因组位点。突变组可包含多于约500个基因组位点。突变组可包含多于约1000个基因组位点。突变组可包含多于约2000个基因组位点。突变组可包含多于约3000个基因组位点。突变组可包含多于约1个或2个融合对。突变组可包含多于约5个融合对。突变组可包含多于约10个融合对。突变组可包含多于约15个融合对。突变组可包含多于约20个融合对。突变组可包含多于约25个融合对。

[0050] 如本文所用的术语“疾病诊断”通常是指诊断或筛查疾病,以对疾病发生风险进行分层,监测疾病的进展或缓解,制定疾病的治疗方案,或上述的任意组合。疾病诊断可包括a) 从来自受试者的一个或多个组织样品获得信息,b) 基于所获得的信息或组织样品,确定受试者是否具有特定疾病,c) 对受试者的疾病发生风险进行分层,d) 确认受试者是否患有所述疾病,是否正在发展所述疾病,或是否处于疾病缓解中,或其任意组合。疾病诊断可为疾病提示特定治疗或治疗性干预。疾病诊断还可诸如通过p值、校正的p值或统计学置信度指标来提供指示例如疾病如癌症的严重程度或等级或准确诊断可能性的得分。疾病诊断还可指示特定类型的疾病。例如,用于甲状腺癌的疾病诊断可指示亚型如滤泡性腺瘤(FA)、结节增生(NHP)、淋巴细胞性甲状腺炎(LCT)、许特尔细胞腺瘤(HA)、滤泡性癌(FC)、乳头状甲状腺癌(PTC)、乳头状癌的滤泡变型(FVPTC)、甲状腺髓样癌(MTC)、许特尔细胞癌(HC)、甲状腺未分化癌(ATC)、肾癌(RCC)、乳腺癌(BCA)、黑色素瘤(MMN)、B细胞淋巴瘤(BCL)、甲状旁腺(PTA)或增生性乳头状癌(HPC)。

#### 用于评估疾病发生或复发风险的方法

[0051] 本公开内容提供了用于评估受试者的组织样品以确定受试者的疾病发生或复发风险,并在一些情况下确定该疾病的新型遗传生物标志物的方法。这样的方法可包括获得与从受试者获得的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平。在一些情况下,使用具有对所述第一组基因的所述一个或多个基因具有选择性的探针的微阵列获得所述表达水平。可由受试者或另一个体如医学专业人员获得核酸样品。第一组基因可与受试者的疾病发生风险相关。在一些实例中,通过FNA、手术(例如,手术活检)或用于从受试者获得样品的其他方法获得核酸样品。核酸样品可在从受试者获得的组织样品(如甲状腺组织样品)、血液样品或流体样品中。在实例中,核酸样品可包含在从受试者获得的FNA样品中。

[0052] 接下来,确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在。第二组基因可与受试者的疾病发生风险相关。在一些实例中,通过对FNA样品中的核酸进行测序以获得核酸序列,来确定所述序列的存在。测序还可富集第二组基因的一个或多个基因或其变体。

[0053] 接下来,将所获得的表达水平和所获得的核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与对照的比较。所述核酸序列与所述对照中的参考序列的比较可指示针对于第二组基因的给定基因的一种或多种序列变体的存在。参考序列可以是,例如,从受试者获得的管家基因。

[0054] 接下来,对所述比较进行分析,并基于所述比较确定疾病发生或复发的风险。在一些实例中,使用由一个或多个编程的计算机处理器执行的算法来对所述比较进行分析并确

定疾病发生或复发的风险。该算法可以是经训练的算法(例如,在至少10、200、100或500个参考样品上进行训练的算法)。可从已经被诊断为患有疾病的受试者或从健康受试者获得参考样品。

[0055] 在一些实例中,可通过测定所述表达水平获得第一组基因的所述一个或多个基因中每一个的表达水平。在一些实例中,可通过测定所述一个或多个基因中的每一个来确定与第二组基因的所述一个或多个基因中的每一个相对应的核酸序列的存在。在这样的实例中,测定可包括阵列杂交、核酸测序、核酸扩增或其他测定。测定可包括测序,如DNA或RNA测序。这样的测序可以是下一代(NextGen)测序。测定可包括逆转录聚合酶链反应(PCR)。测定可利用针对第一组或第二组基因的所述一个或多个基因中的每一个所选择的标志物,如引物。

[0056] 可在获得与第一组基因的所述一个或多个基因相对应的表达水平之前,从受试者获得样品。还可在获得与第一组基因的所述一个或多个基因相对应的表达水平之前确定核酸样品的多个基因的表达水平。在一些情况下,可在确定第二组基因的核酸序列的存在之前,确定样品中所述多个基因的核酸序列。

[0057] 在一些实例中,所述疾病为癌症,如甲状腺癌、乳腺癌或其他癌症。确定发生或复发风险还可在非癌性疾病如遗传病、过度增生性病征或其他疾病中确定。

[0058] 从受试者获得的样品可以是细胞学上不明确的或可疑的(或不确定的)。在一些情况下,样品可提示疾病的存在。从受试者获得的样品的体积可以是少量的,如约100微升、50微升、10微升、5微升、1微升或更少。样品可包含低数量或低质量的多核苷酸,如具有降解或部分降解的RNA的组织样品。例如,FNA样品可产生低数量或低质量的多核苷酸。在这样的实例中,样品的RNA完整性指数(RIN)值可以为约9.0或更小。在一些实例中,RIN值可以为约6.0或更小。

[0059] 疾病发生风险可包括随后发生,如第二次、第三次、第四次或更往后发生的风险。疾病发生风险可包括以下的一种或多种:a)首次发展疾病的风险,b)复发或再次发展疾病的风险,c)将来发展疾病的风险,d)在受试者的一生中易于发展疾病的风险,e)作为婴儿、青少年或成年人易于发展疾病的风险。在疾病为癌症的情况下,发生风险可包括癌症变为转移性的风险。

[0060] 风险的确定可在术前,如在患者的手术之前完成。临床医师可建议,例如,如果患者被确定为具有乳头状甲状腺癌低风险,则继续对患者进行观察而不是建议手术。在一些情况下,如果患者被确定为具有乳头状甲状腺癌高风险,则临床医师更可能建议患者进行手术。可在受试者具有阳性疾病诊断之前,如当受试者疑似患有疾病时或在常规临床程序期间进行确定。

[0061] 风险的确定可进一步包括将风险分层为低发生风险或中高发生风险。在一些实例中,低风险可以是约50%至约80%的发生概率,而中高风险可以是约80%至100%的发生概率。

[0062] 将风险准确分层为低风险组和中高风险组可发生在约80%的分析样品中。可在约50%、60%、70%、75%、80%、85%、90%、95%、96%、97%、98%或约99%的分析样品,包括被鉴定为细胞学上不明确的或可疑的样品中准确地确定风险分层。将风险分层为低风险组和中高风险组可以为至少约80%特异的。在一些实例中,风险分层的特异性可以为约50%、

60%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%或更高,包括被鉴定为细胞学上不明确的或可疑的样品。

[0063] 第一组或第二组基因中的所述一个或多个基因可包括多个基因,如约2、10、20、40个基因或更多个基因。第一组或第二组中的所述一个或多个基因可以为少于约10个基因、20个基因、50个基因、60个基因或约75个基因。第一组或第二组的所述一个或多个基因可以为约50个至约400个基因。第一组基因可包含图6或图7的基因。第二组基因可包含图8的基因。

[0064] 第一组和第二组基因可以是相同的组。例如,第一组和第二组基因可包含COL1A1、THBS2或其任意组合。

[0065] 第一组和第二组基因可以是不同的组。第二组基因可包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROBO1、C6orf136、SPAG4、GALNT15、LUM、NCAM2、NUP210L、NR2F1、THBS2、PSORS1C1或其任意组合。第一组基因可包含COL1A1、TMEM92、C1orf87、SPAG4、EHF、COL3A1、GALNT15、NUP210L、PDZRN3、C6orf136、NA、NRXN3、COL6A3、RAPGEF5、PRICKLE1、LUM、ROBO1、BGN、AC019117.2、PRSS3P1或其任意组合。第二组基因可包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROBO1、C6orf136、SPAG4、GALNT15、LUM、NCAM2、SYNP02、NUP210L、AMZ1、NR2F1、THBS2、PSORS1C1、FTH1P24或其任意组合。第二组基因可包含AKAP9、SPRY3、SPRY3、CAMKK2、COL1A1、FITM2、COX6C、VSIG10L、CYC1、KDM1B、MAPK15、ARSG、PAXIP1、DAAM1、AVL9、DMGDH、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-DRB5、HLA-H、IRF1、MGAT1、P2RX1、PLEK、CCDC93、PPP1R12C、SLC41A3、METTL3、CCAR2、PTPRE、SRL、SLC30A5、BMP4、ZNF133、ICE2、DCAKD、TMX1、TNFSF12、PER2、MCM3AP或其任意组合。

#### 样品

[0066] 从受试者获得的样品可包括组织、细胞、细胞碎片、细胞器、核酸、基因、基因片段、表达产物、基因表达产物、基因表达产物片段或其任意组合。样品可以是异质的或均质的。样品可包括血液、尿液、脑脊液、精液、唾液、痰、粪便、淋巴液、组织或其任意组合。样品可以是组织特异性样品,如从甲状腺组织、皮肤、心脏、肺、肾脏、乳房、胰腺、肝脏、肌肉、平滑肌、膀胱、胆囊、结肠、肠、脑、食道或前列腺获得的样品。

[0067] 可通过各种方法,例如,细针抽吸(FNA)、芯针活检、真空辅助活检、切开活检、切除活检、钻取活检、刮取活检、皮肤活检或其任意组合获得本公开内容的样品。

[0068] FNA也称为细针抽吸活检(FNAB)或针抽吸活检(NAB),其为从受试者获得少量组织的方法。FNA可以比组织活检具有更小的侵入性,而为了获得组织活检可能需要对受试者进行手术和住院治疗。可将FNA方法的针插入受试者的组织块中,以获得用于进一步分析的样品量。在一些情况下,可将两根针插入组织块中。可通过穿过组织块的针的一个或多个通道获取从组织块获得的FNA样品。在一些情况下,FNA样品可包含少于约 $6 \times 10^6$ 、 $5 \times 10^6$ 、 $4 \times 10^6$ 、 $3 \times 10^6$ 、 $2 \times 10^6$ 、 $1 \times 10^6$ 个细胞或更少的细胞。可通过超声或其他成像装置将针引导至组织块。针可以是中空的,以允许通过抽吸或真空或其他抽吸技术利用针来回收FNA样品。

[0069] 使用本文公开的方法获得的样品,如FNA样品,可以包含低样品体积。样品体积可小于约500微升( $\mu\text{L}$ )、400 $\mu\text{L}$ 、300 $\mu\text{L}$ 、200 $\mu\text{L}$ 、100 $\mu\text{L}$ 、75 $\mu\text{L}$ 、50 $\mu\text{L}$ 、25 $\mu\text{L}$ 、20 $\mu\text{L}$ 、15 $\mu\text{L}$ 、10 $\mu\text{L}$ 、5 $\mu\text{L}$ 、1 $\mu\text{L}$ 、0.5 $\mu\text{L}$ 、0.1 $\mu\text{L}$ 、0.01 $\mu\text{L}$ 或更小。样品体积可小于约1 $\mu\text{L}$ 。样品体积可小于约5 $\mu\text{L}$ 。样品体积可小于约10 $\mu\text{L}$ 。样品体积可小于约20 $\mu\text{L}$ 。样品体积可以为约1 $\mu\text{L}$ 至约10 $\mu\text{L}$ 。样品体积可以为

约10uL至约25uL。

[0070] 使用本文公开的方法获得的样品,如FNA样品,可以包含低样品重量。样品重量,如组织重量,可小于约100毫克(mg)、75mg、50mg、25mg、20mg、15mg、10mg、9mg、8mg、7mg、6mg、5mg、4mg、3mg、2mg、1mg、0.5mg、0.1mg或更小。样品重量可小于约20mg。样品重量可小于约10mg。样品重量可小于约5mg。样品重量可以为约5mg至约20mg。样品重量可以为约1mg至约5mg。

[0071] 使用本文公开的方法获得的样品,如FNA,可包含低细胞数目。单个样品的细胞数目可少于约 $10 \times 10^6$ 、 $5.5 \times 10^6$ 、 $5 \times 10^6$ 、 $4.5 \times 10^6$ 、 $4 \times 10^6$ 、 $3.5 \times 10^6$ 、 $3 \times 10^6$ 、 $2.5 \times 10^6$ 、 $2 \times 10^6$ 、 $1.5 \times 10^6$ 、 $1 \times 10^6$ 、 $0.5 \times 10^6$ 、 $0.2 \times 10^6$ 、 $0.1 \times 10^6$ 个细胞或更少。单个样品的细胞数目可少于约 $5 \times 10^6$ 个细胞。单个样品的细胞数目可少于约 $4 \times 10^6$ 个细胞。单个样品的细胞数目可少于约 $3 \times 10^6$ 个细胞。单个样品的细胞数目可少于约 $2 \times 10^6$ 个细胞。单个样品的细胞数目可以为约 $1 \times 10^6$ 至约 $5 \times 10^6$ 个细胞。单个样品的细胞数目可以为约 $1 \times 10^6$ 至约 $10 \times 10^6$ 个细胞。

[0072] 使用本文公开的方法获得的样品,如FNA,可包含低脱氧核糖核酸(DNA)或核糖核酸(RNA)量。单个样品中DNA或RNA的量可少于约500纳克(ng)、400ng、300ng、200ng、100ng、75ng、50ng、45ng、40ng、35ng、30ng、25ng、20ng、15ng、10ng、5ng、1ng、0.5ng、0.1ng或更少。DNA或RNA的量可少于约40ng。DNA或RNA的量可少于约25ng。DNA或RNA的量可少于约15ng。DNA或RNA的量可以为约1ng至约25ng。DNA或RNA的量可以为约5ng至约50ng。

[0073] 样品的RNA产量或RNA量可以测得为纳克到微克的量。可用于在实验室中测定核酸产量的设备的实例是NANODROP®分光光度计、QUBIT®荧光计或QUANTUS™荧光计。NANODROP®测量的精确度在极低RNA浓度时可能显著降低。从本文所述方法获得的数据质量可能依赖于RNA量。可由通过NANODROP®测量具有低RNA浓度或检测不到的RNA浓度的样品产生有意义的基因表达数据或序列变体数据或其他数据。在一些情况下,可从具有检测不到的RNA浓度的样品产生基因表达数据或序列变体数据或其他数据。

[0074] 可以使用具有低数量或低质量的多核苷酸如DNA或RNA的样品进行本文所述的方法。具有低数量或低质量RNA的样品可以是例如降解或部分降解的组织样品。具有低数量或低质量RNA的样品可以是细针抽吸(FNA)样品。可以根据计算得到的RNA完整性指数(RIN)值来测定样品中的RNA质量。RIN值是用于将完整性值赋予RNA测量的算法。该算法可以赋予1到10的RIN值,其中RIN值10可以是完全完整的RNA。如本文所述的包含RNA的样品可具有约9.0、8.0、7.0、6.0、5.0、4.0、3.0、2.0、1.0或更小的RIN值。在一些情况下,包含RNA的样品可具有等于或小于约8.0的RIN值。在一些情况下,包含RNA的样品可具有等于或小于约6.0的RIN值。在一些情况下,包含RNA的样品可具有等于或小于约4.0的RIN值。在一些情况下,样品可具有小于约2.0的RIN值。

[0075] 可由另一个体或实体如医疗保健(或医学)专业人员或机器人从受试者获得样品如FNA样品。医学专业人员可包括医师、护士、医学技师或其他人员。在一些情况下,医师可以是专家,如肿瘤学家、外科医生或内分泌学家。医学技师可以是专家,如细胞学家、抽血技师、放射科医师、肺脏学家或其他专家。医学专业人员可从受试者获得用于测试的样品,或指引受试者去测试中心或实验室以提交样品。医学专业人员可指示测试中心或实验室对样品进行合适的测试或测定,如本公开内容的方法,包括确定基因序列数据、基因表达水平、



序列变体数据或其任意组合。

[0076] 在一些情况下,医学专业人员不必参与疾病的初始诊断或初始样品获得。个体如受试者可以另行通过使用非处方的试剂盒来获得样品。所述试剂盒可以包含如本文所述的用于获得所述样品的收集单元或装置,用于在样品分析之前储存所述样品的储存单元,以及指导使用所述试剂盒的说明书。

[0077] 可以a)在术前,b)在术后,c)在癌症诊断后,d)在疾病缓解或治愈后的常规筛查期间,e)在受试者疑似患有疾病时,f)在常规门诊就诊或临床筛查期间,g)遵循医学专业人员的要求,或以上述的任意组合获得样品。可在不同的时间从同一受试者获得多个样品,如在疾病治疗开始之前和治疗结束之后,如在一段时间进程内监测受试者。可在不同的时间从受试者获得多个样品,以监测受试者的疾病进展、消退或缓解的不存在或存在。

#### 细胞学分析

[0078] 如本文所述的方法(包括评估疾病发生风险)可包括样品的细胞学分析。细胞学分析的实例包括通过任何数目的方法和合适的试剂进行的细胞染色技术和/或显微镜检查,包括但不限于:曙红-天蓝(EA)染料、苏木精染料、CYTO-STAIN™、巴氏染料、曙红、尼斯尔染料(nissl stain)、甲苯胺蓝、银染料、偶氮胭脂红染料、中性红或詹纳斯绿。多于一种染料可与其他染料结合使用。在一些情况下,细胞完全没有染色。可在染色程序之前或期间用例如甲醇、乙醇、戊二醛或甲醛将细胞固定和/或通透化。在一些情况下,细胞可以不进行固定。染色程序还可用于测定样品的核酸含量,例如用溴化乙锭、苏木精、尼斯尔染料或任何其他核酸染料。

[0079] 样品中的细胞的显微镜检查可以包括通过用于细胞学检查的标准方法将细胞涂抹在载玻片上。可以使用基于液体的细胞学(LBC)方法。在一些情况下,LBC方法提供改善的细胞学载玻片制备方法、更均匀的样品、提高的灵敏性和特异性,或改善的样品处理效率,或其任意组合。在LBC方法中,可以将样品从受试者转移到含有LBC制剂溶液例如CYTYC THINPREP®、SUREPATH™或MONOPREP®或任何其他LBC制剂溶液的容器或小瓶中。另外,可以用LBC制剂溶液将样品从收集装置冲洗到容器或小瓶中,以确保基本上定量地转移样品。然后可以将LBC制剂溶液中含有样品的溶液储存和/或由机器或本领域技术人员处理,从而在载玻片上产生一层细胞。还可以用与常规细胞学制备相同的方法将样品染色并在显微镜下检查。

[0080] 可以通过免疫组织化学染色来分析样品。免疫组织化学染色可以通过利用抗体提供对样品(例如细胞或组织)中的特定分子或抗原的存在、位置和分布的分析。抗原可以是小分子、蛋白质、肽、核酸或能够被抗体特异性识别的任何其他分子。可以在有或者没有预先固定和/或透化步骤的情况下通过免疫组织化学方法对样品进行分析。在一些情况下,可通过使样品与抗原特异性抗体接触来检测目的抗原,然后通过一次或多次洗涤除去非特异性结合。然后通过抗体检测试剂例如标记的第二抗体或标记的抗生物素蛋白/链霉抗生物素蛋白来检测特异性结合的抗体。可以直接标记抗原特异性抗体。用于免疫组织化学的合适的标记包括但不限于:荧光团,如荧光素和若丹明;酶,如碱性磷酸酶和辣根过氧化物酶;或放射性核素,如<sup>32</sup>P和<sup>125</sup>I。可通过免疫组织化学染色检测的基因产物标志物包括但不限于Her2/Neu、Ras、Rho、EGFR、VEGFR、UbcH10、RET/PTC1、细胞角蛋白20、降钙素、GAL-3、甲状腺过氧化物酶或甲状腺球蛋白。

[0081] 与如本文所公开的疾病发生风险相关的指标,如第一组基因的基因表达水平或第二组基因的序列变体数据,不必是被发现包含疾病发生风险的样品的每个细胞的特征。因此,本文公开的方法可用于评估组织内疾病如癌症发生的风险,其中在该组织中,并非样品中的所有细胞都表现出基因表达水平或序列变体数据或指示疾病发生风险的其他数据的完整模式。基因表达水平、序列变体数据或其他数据可以完全存在于、部分存在于或不存在于受影响的细胞内以及样品中未受影响的细胞内。基因表达水平、序列变体数据或其他数据可以以可变的量存在于受影响的细胞内。基因表达水平、序列变体数据或其他数据可以以可变的量存在于未受影响的细胞内。在一些情况下,可阳性地检测到与疾病发生风险相关的第一组基因的基因表达水平或第二组基因中一种或多种序列变体的存在。在一些情况下,阳性检测可发生在从样品得到的至少70%、75%、80%、85%、90%、95%或100%的细胞中。在一些情况下,可能检测不到第一组基因的基因表达水平或第二组基因中一个或多个序列变体的存在。在一些情况下,检测不到可以发生在对应的正常非疾病样品的至少70%、75%、80%、85%、90%、95%或100%的细胞中。

[0082] 常规细胞学或其他测定可以指示样品为阴性的(无疾病)、诊断性的(疾病如癌症的阳性诊断)、不明确的或疑似的(提示疾病如癌症的存在)或非诊断性的(未提供充分的关于疾病的存在或不存在的信息)。如本文所述的方法可以确认从常规细胞学评估得到的结果,或者可在缺乏此种结果的情况下提供类似于常规细胞学评估的初始评估。如本文所述的方法可将样品分类为恶性的或良性的,包括被发现是不明确的或疑似的样品。所述方法可将样品如已知是恶性的样品进一步分层为疾病发生的低风险组和中高风险组,包括被发现是不明确的或疑似的样品。

#### 疾病

[0083] 如本文所公开的疾病可包括甲状腺癌。甲状腺癌可包括甲状腺癌的任何亚型,包括但不限于甲状腺的任何恶性肿瘤,如乳头状甲状腺癌(PTC)、滤泡性甲状腺癌(FTC)、乳头状甲状腺癌的滤泡变体(FVPTC)、甲状腺髓样癌(MTC)、滤泡性癌(FC)、许特尔细胞癌(HC)和/或甲状腺未分化癌(ATC)。在一些情况下,甲状腺癌可以是分化的。在一些情况下,甲状腺癌可以是未分化的。

[0084] 可使用本公开内容的方法将甲状腺组织样品分类为包括一种或多种良性或恶性组织类型(例如癌症亚型),包括但不限于滤泡性腺瘤(FA)、结节增生(NHP)、淋巴细胞性甲状腺炎(LCT),和许特尔细胞腺瘤(HA)、滤泡性癌(FC)、乳头状甲状腺癌(PTC)、乳头状癌的滤泡变型(FVPTC)、甲状腺髓样癌(MTC)、许特尔细胞癌(HC),和甲状腺未分化癌(ATC)、肾癌(RCC)、乳腺癌(BCA)、黑色素瘤(MMN)、B细胞淋巴瘤(BCL)或甲状旁腺(PTA)。

[0085] 本公开内容的其他类型的癌症可包括但不限于:肾上腺皮质癌、肛门癌、再生障碍性贫血、胆管癌、膀胱癌、骨癌、骨转移、中枢神经系统(CNS)癌、周围神经系统(PNS)癌、乳腺癌、Castleman病、宫颈癌、儿童非霍奇金淋巴瘤、淋巴瘤、结肠直肠癌、子宫内膜癌、食道癌、尤因肿瘤家族(例如尤因肉瘤)、眼癌、胆囊癌、胃肠道类癌瘤、胃肠道间质肿瘤、妊娠滋养细胞疾病、毛细胞白血病、霍奇金病、卡波西肉瘤、肾癌、喉癌和下咽癌、急性淋巴细胞性白血病、急性髓样白血病、儿童白血病、慢性淋巴细胞性白血病、慢性髓样白血病、肝癌、肺癌、肺类癌瘤、非何杰金淋巴瘤、男性乳腺癌、恶性间皮瘤、多发性骨髓瘤、脊髓增生异常综合征、骨髓增生性疾病、鼻腔和鼻旁癌、鼻咽癌、神经母细胞瘤、口腔和口咽癌、骨肉瘤、卵巢癌、胰

腺癌、阴茎癌、垂体瘤、前列腺癌、视网膜母细胞瘤、横纹肌肉瘤、唾液腺癌、肉瘤(成人软组织瘤)、黑色素瘤皮肤癌、非黑色素瘤皮肤癌、胃癌、睾丸癌、胸腺癌、子宫癌(例如子宫肉瘤)、阴道癌、外阴癌或瓦尔登斯特伦巨球蛋白血症。

[0086] 如本文所公开的疾病可包括过度增生性病症。可将恶性过度增生性病症分层为风险组,如低风险组和中高风险组。过度增生性病症可以包括但不限于癌症、增生或肿瘤。在一些情况下,过度增生性癌症可以是乳腺癌,如乳腺导管组织中的导管癌、髓样癌、胶质癌、小管癌和炎性乳腺癌;卵巢癌,包括上皮性卵巢肿瘤,比如卵巢中的腺癌和从卵巢转移到腹腔的腺癌;子宫癌;宫颈癌,比如宫颈上皮中的腺癌,包括鳞状细胞癌和腺癌;前列腺癌,比如选自以下的前列腺癌:腺癌或转移到骨的腺癌;胰腺癌,比如胰管组织中的上皮样癌和胰腺管中的腺癌;膀胱癌,比如膀胱中的移行细胞癌、尿路上皮癌(移行细胞癌)、内衬于膀胱的尿路上皮细胞中的肿瘤、鳞状细胞癌、腺癌和小细胞癌;白血病,比如急性髓样白血病(AML)、急性淋巴细胞白血病、慢性淋巴细胞白血病、慢性髓样白血病、毛细胞白血病、脊髓发育不良、骨髓增生性疾病、急性髓性白血病(AML)、慢性髓性白血病(CML)、肥大细胞增多症、慢性淋巴细胞白血病(CLL)、多发性骨髓瘤(MM)和骨髓增生异常综合征(MDS);骨癌;肺癌,如非小细胞肺癌(NSCLC),其分为鳞状细胞癌、腺癌和大细胞未分化癌,以及小细胞肺癌;皮肤癌,如基底细胞癌、黑色素瘤、鳞状细胞癌和光化性角化病,这是一种皮肤病,有时发展成鳞状细胞癌;眼睛视网膜母细胞瘤;皮肤或眼内(眼睛)黑色素瘤;原发性肝癌(起始于肝脏的癌症);肾癌;自身免疫缺陷综合征(AIDS)相关淋巴瘤,如弥漫性大B细胞淋巴瘤、B细胞免疫母细胞性淋巴瘤和小无裂细胞淋巴瘤;卡波西肉瘤;病毒诱发性癌症,包括乙型肝炎病毒(HBV)、丙型肝炎病毒(CBV)和肝细胞癌;1型人嗜淋巴细胞病毒(HTLV-1)和成人T细胞白血病/淋巴瘤;以及人乳头状瘤病毒(HPV)和宫颈癌;中枢神经系统(CNS)癌,如原发性脑肿瘤,其包括神经胶质瘤(星形细胞瘤、间变性星形细胞瘤或多形性胶质母细胞瘤)、少突神经胶质瘤、室管膜瘤、脑膜瘤、淋巴瘤、神经鞘瘤和髓母细胞瘤;周围神经细胞(PNS)癌,如听神经瘤和恶性周围神经鞘瘤(MPNST),包括神经纤维瘤和神经鞘瘤、恶性纤维细胞瘤、恶性纤维组织细胞瘤、恶性脑膜瘤、恶性间皮瘤和恶性混合Müllerian瘤;口腔癌和口咽癌,如下咽癌、喉癌、鼻咽癌和口咽癌;胃癌,如淋巴瘤、胃基质瘤和类癌瘤;睾丸癌,如生殖细胞肿瘤(GCT)(其包括精原细胞瘤和非精原细胞瘤)和性腺基质细胞瘤(其包括莱迪希细胞瘤和塞尔托利细胞瘤);胸腺癌,如胸腺瘤、胸腺癌和霍奇金病,非霍奇金淋巴瘤类癌瘤或类癌瘤;直肠癌;以及结肠癌。在一些情况下,通过本公开内容的方法分层、分类、表征或诊断的疾病包括但不限于甲状腺病,比如良性甲状腺病,包括但不限于滤泡性腺瘤、许特尔细胞腺瘤、淋巴细胞性甲状腺炎和甲状腺增生。在一些情况下,通过本公开内容的方法分层、分类、表征或诊断的疾病包括但不限于恶性甲状腺病,例如滤泡性癌、乳头状甲状腺癌的滤泡变型、髓样癌和乳头状癌。

[0087] 本公开内容的疾病可包括遗传病。遗传病是由基因或染色体异常引起的疾病。遗传病可以分为两类:单基因病及多因子和多基因(复杂)病。单基因病可以是单个突变基因的结果。单基因病的遗传可以包括但不限于常染色体显性、常染色体隐性、X连锁显性、X连锁隐性、Y连锁和线粒体遗传。仅基因的一个突变拷贝就可使人患上常染色体显性病症。常染色体显性类型的病症的实例可以包括但不限于亨廷顿氏病、神经纤维瘤病1型、马方综合征、遗传性非息肉性结肠直肠癌或遗传性多发性外生骨疣。在常染色体隐性病症中,必须突

变基因的两个拷贝以使受试者患上常染色体隐性病症。这一类型的病症的实例可以包括但不限于囊性纤维化、镰状细胞病(以及部分镰状细胞病)、泰-萨克斯(Tay-Sachs)病、Niemann-Pick病或脊髓性肌萎缩。X连锁显性病症由X染色体上的基因突变引起,如X连锁低血磷性佝偻症。一些X连锁显性病状如Rett综合征、2型色素失调症和艾卡迪(Aicardi)综合征可以是致死性的。X连锁隐性病症还由X染色体上的基因突变引起。这一类型的病症的实例可以包括但不限于血友病A、迪谢内(Duchenne)肌营养不良、红绿色盲、肌营养不良和雄激素性脱发。Y连锁病症由Y染色体上的突变引起。其实例可以包括但不限于男性不育症和耳廓多毛症(hypertrichosis pinnae)。线粒体遗传(又称为母体遗传)的遗传病可适用于线粒体DNA中的基因,如Leber遗传性视神经病变。

[0088] 遗传病症也可以是复杂的、多因子的或者多基因的。多基因遗传病可能与多基因的效应结合生活方式和环境因素有关。虽然复杂遗传病可以在家族中集中出现,但他们不具有明确的遗传模式。多因子或多基因病症可以包括心脏病、糖尿病、哮喘、孤独症、自身免疫性疾病如多发性硬化、癌症、纤毛疾病(ciliopathies)、腭裂、高血压、炎性肠病、精神发育迟滞或肥胖。

[0089] 其他遗传病可以包括但不限于1p36缺失综合征、21-羟化酶缺乏症、22q11.2缺失综合征、无铜蓝蛋白血症(aceruloplasminemia)、II型软骨成长不全、软骨发育不全、急性间歇性卟啉症、腺苷酸琥珀酸裂解酶缺陷、肾上腺脑白质营养不良、亚历山大病(Alexander disease)、黑尿酸尿症、 $\alpha$ -1抗胰蛋白酶缺陷、阿耳斯特雷姆(Alstrom)综合征、阿尔茨海默病(1、2、3和4型)、釉质形成不全、肌萎缩侧索硬化、2型肌萎缩侧索硬化、4型肌萎缩侧索硬化、雄激素不敏感综合征、贫血、Angelman综合征、阿佩尔(Apert)综合征、共济失调毛细血管扩张、比尔-史蒂文生皮肤回旋综合征(Beare-Stevenson cutis gyrata syndrome)、本杰明综合征、 $\beta$ 地中海贫血、生物素酶缺陷、Birt-Hogg-Dubé综合征、膀胱癌、布卢姆(Bloom)综合征、骨疾病、乳腺癌、屈肢骨发育不全(Camptomelic dysplasia)、卡纳万(Canavan)病、癌症、乳糜泻、慢性肉芽肿病(CGD)、夏科-马里-图斯(Charcot-Marie-Tooth)病、1型夏科-马里-图斯病、4型夏科-马里-图斯病、2型夏科-马里-图斯病、科凯恩(Cockayne)综合征、科-勒二氏(Coffin-Lowry)综合征、II型和XI型胶原病(collagenopathy)、结肠直肠癌、先天性输精管缺如、先天性双侧输精管缺如、先天性糖尿病、先天性红细胞生成性卟啉症、先天性心脏病、先天性甲状腺功能减退、结缔组织病、考登(Cowden)综合征、猫叫(Cri du chat)综合征、克罗恩病、纤维性狭窄病(fibrostenosing)、克鲁宗(Crouzon)综合征、Crouzonodermoskeletal综合征、囊性纤维化、德格罗契(De Grouchy)综合征、退行性神经病、登特(Dent)病、发育障碍、迪格奥尔格(DiGeorge)综合征、V型远端脊髓性肌萎缩、唐氏综合征、侏儒症、埃勒斯-丹洛斯(Ehlers-Danlos)综合征、关节松弛型埃勒斯-丹洛斯(Ehlers-Danlos)综合征、经典型埃勒斯-丹洛斯综合征、皮肤脆裂型埃勒斯-丹洛斯综合征、脊柱后凸侧弯型脉管型埃勒斯-丹洛斯综合征、红细胞生成性原卟啉症、法布瑞氏(Fabry)病、面部损伤和病症、因子V Leiden血栓形成倾向、家族性腺瘤性息肉病、家族性自主神经异常、范科尼(fanconi)贫血、FG综合征、脆性X染色体综合征、弗里德赖希共济失调(Friedreich ataxia)、弗里德赖希氏共济失调、G6PD缺陷、半乳糖血症、戈谢(Gaucher)病(1、2和3型)、遗传性脑病、甘氨酸脑病、2型血色素沉着症、4型血色素沉着症、丑角样鱼鳞病(Harlequin Ichthyosis)、头脑畸形、听力障碍和耳

聋、儿童听力问题、血色沉着病(新生儿、2型和3型)、血友病、肝红细胞生成性卟啉症、遗传性粪卟啉症、遗传性多发性外生骨疣、遗传性压迫易感性神经病、遗传性非息肉性结肠直肠癌、高胱氨酸尿症、亨廷顿病、早年衰老综合征(Hutchinson-Gilford Progeria syndrome)、原发性高草酸尿症、高苯丙氨酸血症、软骨形成不足、软骨发育不良、idic15、色素失调症、婴儿型戈谢病、婴儿-发作上升型遗传性痉挛性瘫痪(infantile-onset ascending hereditary spastic paralysis)、不育症、Jackson-Weiss综合征、朱伯特(Joubert)综合征、青少年原发性侧索硬化、肯尼迪(Kennedy)病、Klinefelter综合征、Kniest发育不全、Krabbe病、学习障碍(Learning disability)、Lesch-Nyhan综合征、脑白质营养不良、Li-Fraumeni综合征、家族性脂蛋白脂酶缺陷、男性生殖器障碍、马方综合征、McCune-Albright综合征、McLeod综合征、家族性地中海热、Menkes病、Menkes综合征、代谢障碍、 $\beta$ -球蛋白型高铁血红蛋白血症、先天性高铁血红蛋白血症、甲基丙二酸血症、Micro综合征、小头畸形、运动障碍、Mowat-Wilson综合征、粘多糖贮积病(MPS I)、Muenke综合征、肌营养不良、Duchenne和Becker型肌营养不良症、Duchenne和Becker型肌萎缩症、肌强直性营养不良、1型和2型肌强直性营养不良、新生儿血色病、神经纤维瘤病、神经纤维瘤病1、神经纤维瘤病2、I型神经纤维瘤病、II型神经纤维瘤病、神经病、神经肌肉障碍、Niemann-Pick病、非酮性高甘氨酸血症、非综合征型耳聋、常染色体隐性非综合征型耳聋、Noonan综合征、成骨不全(I型和III型)、耳脊椎骨骺发育不良(otospondylomegalapiphyseal dysplasia)、泛酸激酶-相关神经退行性变、Patau综合征(13染色体三体)、Pendred综合征、Peutz-Jeghers综合征、Pfeiffer综合征、苯丙酮尿症、卟啉症、迟发性皮肤卟啉症、Prader-Willi综合征、原发性肺动脉高压、朊病毒病、早衰、丙酸血症、蛋白C缺陷、蛋白S缺陷、假性戈谢病(pseudo-Gaucher disease)、弹性假黄瘤、视网膜病、视网膜母细胞瘤、视网膜母细胞瘤FA-弗里德赖希共济失调、Rett综合征、Rubinstein-Taybi综合征、Sandhoff病、感觉和自主神经病III型、镰状细胞性贫血、骨骼肌再生、皮肤色素沉着异常、Smith Lemli Opitz综合征、语言和交流障碍、脊髓性肌萎缩、脊髓延髓肌萎缩、脊髓小脑共济失调、Strudwick型脊椎干骺端发育不良、先天性脊椎干骺端发育不良、Stickler综合征、Stickler综合征COL2A1、Tay-Sachs病、四氢生物蝶呤缺陷、致死性发育不良、硫胺反应性巨幼细胞性贫血伴糖尿病和感觉神经性耳聋、甲状腺病、Tourette综合征、Treacher Collins综合征、X染色体三体综合征、结节性硬化、Turner综合征、Usher综合征、多样性卟啉病、von Hippel-Lindau病、Waardenburg综合征、Weissenbacher-Zweymüller综合征、Wilson病、Wolf-Hirschhorn综合征、着色性干皮病、X-连锁重度联合免疫缺陷、X-连锁铁粒幼细胞性贫血或X-连锁脊髓延髓肌萎缩。

#### 发生或复发风险的分层

[0090] 疾病发生风险可将样品分层为风险亚组。亚组可包含具有低疾病发生风险概率的样品和具有中高疾病发生风险概率的样品。亚组可包括低风险组、中风险组和高风险组。低风险可包含具有约1%、5%、10%、15%、20%、25%、30%、35%、40%或约45%的疾病发生风险概率的样品。低风险可包含具有约1%至约25%的疾病发生风险概率的样品。低风险可包含具有约1%至约30%的疾病发生风险概率的样品。低风险可包含具有约1%至约40%的疾病发生风险概率的样品。中高风险可包含具有约55%、60%、65%、70%、75%、80%、85%、90%、95%或100%的疾病发生风险概率的样品。中高风险可包含具有约50%至约

100%的疾病发生风险概率的样品。中高风险可包含具有约55%至约100%的疾病发生风险概率的样品。中高风险可包含具有约60%至约100%的疾病发生风险概率的样品。

[0091] 可以以至少50%、60%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%或更高的准确度将样品分层为低风险组或中高风险组,包括被鉴定为细胞学上不明确的或疑似的或不确定的样品。可以以至少70%的准确度对样品进行分层。可以以至少80%的准确度对样品进行分层。可以以至少90%的准确度对样品进行分层。可以以高于50%、60%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%或更高的准确度将样品鉴定为良性的、恶性的或非诊断性的,包括被鉴定为细胞学上不明确的或疑似的或不确定的样品。可以使用分类器计算准确度。

[0092] 可以以至少50%、60%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%或更高的特异性将样品分层为低风险组或中高风险组,包括被鉴定为细胞学上不明确的或疑似的或不确定的样品。可以以至少70%的准确度对样品进行分层。可以以至少80%的准确度对样品进行分层。可以以至少90%的准确度对样品进行分层。可以以高于50%、60%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%或更高的特异性将样品鉴定为良性的、恶性的或非诊断性的,包括被鉴定为细胞学上不明确的或疑似的或不确定的样品。可以使用分类器计算特异性。

[0093] 如本文所述的用于将样品分类为良性的、恶性的或非诊断性的疾病发生风险分层的方法可具有至少95%、95.5%、96%、96.5%、97%、97.5%、98%、98.5%、99%、99.5%或更大的阳性预测值;和/或至少95%、95.5%、96%、96.5%、97%、97.5%、98%、98.5%、99%、99.5%或更大的阴性预测值。疾病的阳性预测值(PPV),或准确率,或验后概率,可以是被正确诊断或正确分层为风险组的具有阳性测试结果的受试者的比例。其可以是重要的量度,因为它可以反映阳性测试反映出所测试的基础疾病的可能性。其值可依赖于可能发生变化的疾病的发生率。阴性预测值(NPV)可以是被正确诊断的具有阴性测试结果的受试者的比例。PPV和NPV量度可使用适当的疾病亚型发生率估计值获得。对于亚型特异性估计值,疾病发生率有时可能是无法计算的,因为可能没有任何可用的样品。

[0094] 样品可分类为以下的一种或多种:良性的(无疾病)、恶性的(疾病的阳性诊断)或非诊断性的(未提供充分的关于疾病的存在或不存在的信息)。被发现是恶性的样品可分层为疾病发生风险,如低疾病发生风险或中高疾病发生风险。样品可分类为良性的与疑似的(被怀疑为疾病阳性的)类别。可以诸如通过鉴定一种或多种癌症亚型的存在或不存在将样品进一步分类为疾病亚型。某一分子途径可以指示与疾病,或特定疾病的某一等级或阶段(如I、II、III或IV期癌症)有关。在一些情况下,所分层的发生风险可以告知适当的治疗性干预,如特定的药物方案,或外科手术介入,如甲状腺切除术或偏侧甲状腺切除术。

[0095] 本公开内容的分类器或经训练的算法可用于将样品分层为低风险组或中高风险组和/或将样品分类为良性的、恶性的、疑似的或非诊断性的或其他类别。可向分类器或经训练的算法单独提供或组合提供一个或多个选定的特征空间,如基因表达水平和序列变体数据。示例性的算法可包括但不限于减少变量数的方法,例如主成分分析算法、部分最小二乘法或独立成分分析算法。示例性的算法可包括直接处理大量变量的方法,例如统计方法或基于机器学习技术的方法。统计方法可包括惩罚逻辑回归、微阵列的预测分析(PAM)、基于收缩重心的方法、支持向量机分析或规范化线性判别分析。机器学习技术可包括装袋程

序(bagging procedure)、加速程序(boosting procedure)、随机森林算法或其组合。

[0096] 本公开内容的分类器或经训练的算法可包括两个或更多个特征空间。所述两个或更多个特征空间可以是独特的或彼此不同的。单个特征空间可包括针对于样品的信息类型,如基因表达水平数据或序列变体数据。相比于使用单个特征空间产生风险分层,在分类器中组合两个或更多个特征空间可以产生更高水平的风险分层或分类准确度。单个特征空间的动态范围可以不同,如至少1个或2个数量级差异。例如,基因表达水平特征空间的动态范围可以为0至约300,而序列变体特征空间的动态范围可以为0至约20。

[0097] 单个特征空间可包含一组基因,如第一特征空间的第一组基因和第二特征空间的第二组基因。单个特征空间的一组基因可与发生风险相关。第一组基因和第二组基因可以是相同的组。第一组基因和第二组基因可以是不同的组。第一组基因或第二组基因可包含少于约1000、500、400、300、200、100、75、70、65、60、55、50、45、40、35、30、25、20、15、10、5个基因或更少的基因。第一组基因或第二组基因可包含少于约10个基因。第一组基因或第二组基因可包含少于约50个基因。第一组基因或第二组基因可包含少于约75个基因。第一组基因或第二组基因可包含约50个至约400个基因。第一组基因或第二组基因可包含约50个至约200个基因。第一组基因或第二组基因可包含约10个至约600个基因。

[0098] 第一组基因可包含图6中列出的基因。第一组基因可包含图7中列出的基因。第一组基因可包含COL1A1、THBS2或其任意组合。第一组基因可包含COL1A1、TMEM92、C1orf87、SPAG4、EHF、COL3A1、GALNT15、NUP210L、PDZRN3、C6orf136、NA、NRXN3、COL6A3、RAPGEF5、PRICKLE1、LUM、ROB01、BGN、AC019117.2、PRSS3P1或其任意组合。

[0099] 第一组基因可包含图13中列出的基因。第一组基因可包含COL1A1、NUP210L、TMEM92、C6orf136、SPAG4、EHF、RAPGEF5、COL3A1、GALNT15、PRICKLE1、LUM、COL6A3、ROB01、SSC5D、PSORS1C1或其任意组合。第一组基因可选自COL1A1、NUP210L、TMEM92、C6orf136、SPAG4、EHF、RAPGEF5、COL3A1、GALNT15、PRICKLE1、LUM、COL6A3、ROB01、SSC5D、PSORS1C1及其任意组合。第一组基因可包含COL1A1。第一组基因可包含NUP210L。第一组基因可包含TMEM92。第一组基因可包含C6orf136。第一组基因可包含SPAG4。第一组基因可包含EHF。第一组基因可包含RAPGEF5。第一组基因可包含COL3A1。第一组基因可包含GALNT15。第一组基因可包含PRICKLE1。第一组基因可包含LUM。第一组基因可包含COL6A3。第一组基因可包含ROB01。第一组基因可包含SSC5D。第一组基因可包含PSORS1C1。

[0100] 第二组基因可包含图8中列出的那些基因。第二组基因可包含COL1A1、THBS2或其任意组合。第二组基因可包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROB01、C6orf136、SPAG4、GALNT15、LUM、NCAM2、NUP210L、NR2F1、THBS2、PSORS1C1或其任意组合。第二组基因可包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROB01、C6orf136、SPAG4、GALNT15、LUM、NCAM2、SYNP02、NUP210L、AMZ1、NR2F1、THBS2、PSORS1C1、FTH1P24或其任意组合。第二组基因可包含AKAP9、SPRY3、SPRY3、CAMKK2、COL1A1、FITM2、COX6C、VSIG10L、CYC1、KDM1B、MAPK15、ARSG、PAXIP1、DAAM1、AVL9、DMGDH、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-DRB5、HLA-H、IRF1、MGAT1、P2RX1、PLEK、CCDC93、PPP1R12C、SLC41A3、METTL3、CCAR2、PTPRE、SRL、SLC30A5、BMP4、ZNF133、ICE2、DCAKD、TMX1、TNFSF12、PER2、MCM3AP或其任意组合。

[0101] 第二组基因可包含图12中列出的基因。第二组基因可包含COL1A1、FITM2、AASDH、COX6C、COX10、VSIG10L、MAPK15、PAXIP1、AVL9、GIGYF2、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-

H、MGAT1、SLC41A3、PTPRE、SRL、SLC30A5、BMP4、ICE2、DCAKD、TMX1、HAVCR2、TNFSF12、PER2、MCM3AP或其任意组合。第二组基因可选自COL1A1、FITM2、AASDH、COX6C、COX10、VSIG10L、MAPK15、PAXIP1、AVL9、GIGYF2、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-H、MGAT1、SLC41A3、PTPRE、SRL、SLC30A5、BMP4、ICE2、DCAKD、TMX1、HAVCR2、TNFSF12、PER2、MCM3AP及其任意组合。第二组基因可包含COL1A1。第二组基因可包含FITM2。第二组基因可包含AASDH。第二组基因可包含COX6C。第二组基因可包含COX10。第二组基因可包含VSIG10L。第二组基因可包含MAPK15。第二组基因可包含PAXIP1。第二组基因可包含AVL9。第二组基因可包含GIGYF2。第二组基因可包含HLA-DQA1。第二组基因可包含HLA-DQB1。第二组基因可包含HLA-DRA。第二组基因可包含HLA-H。第二组基因可包含MGAT1。第二组基因可包含SLC41A3。第二组基因可包含PTPRE。第二组基因可包含SRL。第二组基因可包含SLC30A5。第二组基因可包含BMP4。第二组基因可包含ICE2。第二组基因可包含DCAKD。第二组基因可包含TMX1。第二组基因可包含HAVCR2。第二组基因可包含TNFSF12。第二组基因可包含PER2。第二组基因可包含MCM3AP。

[0102] 本公开内容的分类器或经训练的算法可使用一组样品如样品群组进行训练。样品群组可包含约5、10、20、30、40、50、60、70、80、90、100、150、200、250、300、350、400、450、500、600、700、800、900、1000、2000、3000、4000、5000个或更多个独立样品。样品群组可包含约100个独立样品。样品群组可包含约200个独立样品。样品群组可包含约100至约500个独立样品。独立样品可来自已被诊断为患有疾病如癌症的受试者，来自健康受试者，或其任意组合。

[0103] 样品群组可包含来自约5、10、20、30、40、50、60、70、80、90、100、150、200、250、300、350、400、450、500、600、700、800、900、1000个或更多个不同个体的样品。样品群组可包含来自约100个不同个体的样品。样品群组可包含来自约200个不同个体的样品。所述不同个体可以是已被诊断为患有疾病如癌症的个体、健康个体或其任意组合。

[0104] 样品群组可包含从居住在至少1、2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60、65、70、75或80个不同地理位置(例如，分布在国家如美国、洲或世界的地点)的个体获得的样品。地理位置包括但不限于测试中心、医疗设施、医务所、邮局地址、城市、县、州、国家或洲。在一些情况下，使用来自美国的样品群组训练的分类器可能需要重新训练以用于来自其他地理区域(例如，印度、亚洲、欧洲、非洲等)的样品群组。

[0105] 分类器或经训练的算法可在其每次运行时产生唯一的输出。例如，使用具有相同分类器的不同样品可在分类器每次运行时产生唯一的输出。使用具有相同分类器的相同样品可在分类器每次运行时产生唯一的输出。使用相同的样品训练分类器多于一次，可在分类器每次运行时产生唯一的输出。

[0106] 可将样品的特征与参考组的特征进行比较。所述比较可由分类器执行。可对样品的多于一个特征进行组合，以制定疾病发生风险。所述组合可由分类器执行。例如，可将从样品获得的序列与参考组进行比较，以确定样品中一个或多个序列变体的存在。在一些情况下，可将来自样品的一个或多个基因的基因表达水平与基因参考组的表达水平进行比较，以确定一个或多个基因的差异基因表达的存在。参考组可包含一个或多个管家基因。参考组可包含已知的序列变体或已知与特定疾病相关或已知与非疾病状态相关的基因的表达水平。分类器或经训练的算法可进行结果的比较、组合、统计评估或进一步分析或其任意



组合。可为不同的特征空间提供单独的参考组。例如,可将序列变体数据与序列变体数据参考组进行比较。可将基因表达水平数据与基因表达水平参考组进行比较。在一些情况下,可将多个特征空间与同一参考组进行比较。

[0107] 在一些情况下,特定基因的序列变体可能影响或可能不影响该相同基因的基因表达水平。特定基因的序列变体可影响一个或多个不同基因的基因表达水平,该一个或多个不同基因可位于具有序列变体的特定基因的附近和远端。一个或多个序列变体的存在可对一个或多个基因具有下游效应。特定基因的序列变体可扰乱一个或多个信号传导途径,可导致核糖核酸(RNA)转录调控变化,可导致脱氧核糖核酸(DNA)的扩增,可导致产生多个转录物拷贝,可导致产生过量的蛋白质,可导致单个碱基对、多个碱基对、部分基因或者一个或多个基因从序列中除去。

[0108] 可使用特征选择技术进一步分析来自所述方法的数据,如基因表达水平或序列变体数据,所述特征选择技术例如是可通过观察数据的本征性质来评估特定特征的相关性的过滤器技术;将模型假设嵌入特征子集检索内的包装器技术;或将最佳特征集的检索构建到分类器算法中的嵌入式协议。

[0109] 可用于本公开内容的方法的过滤器技术可包括(1)参数法,例如采用双样本t-检验、方差分析(ANOVA)分析、贝叶斯框架或伽马分布模型;(2)无模型法,例如采用Wilcoxon秩和检验、类间内平方和检验、秩乘法、随机置换法或误分类数目阈值(TNoM),TNoM包括设置两个数据集之间表达的倍数变化差异的阈值点,然后检测使误分类数目最小化的各基因中的阈值点;或者(3)多变量法,例如双变量法、基于相关性的特征选择法(CFS)、最小冗余最大相关法(MRMR)、Markov毯过滤法和非相关收缩重心法。可用于本公开内容的方法的包装器技术可包括顺序检索法、遗传算法或分布算法的评估。嵌入式协议可包括随机森林算法、支持向量机算法的权向量或逻辑回归算法的权重。

[0110] 从本文所述方法获得的结果的统计学评价可提供指示以下的一种或多种可能性的一个或多个定量值:风险评估准确性的可能性;诊断准确性的可能性;疾病如癌症的可能性;特定疾病如组织特异性癌症,例如甲状腺癌的可能性;以及特定治疗干预成功的可能性。因此,可能没有经过遗传学或分子生物学培训的医学专业人员不需要了解基因表达水平或序列变体数据结果。相反,数据可以以指导受试者医护或治疗的最有用的形式直接提供给医学专业人员。可通过分类器或经训练的算法来进行统计学评价、单独数据结果的组合和报告有用的结果。可使用多种方法进行结果的统计学评价,该方法包括但不限于: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、单向方差分析(ANOVA)、双向ANOVA等。可通过分类器或经训练的算法进行统计学评价。

[0111] 本文公开的方法可包括从来自受试者的一个或多个样品中提取并分析蛋白质或核酸(RNA或DNA)。可从获得的整个样品或可从一部分提取核酸。在一些情况下,未进行核酸提取的样品部分可以通过细胞学检查或免疫组织化学进行分析。从生物样品中提取RNA或DNA的方法可包括,例如,苯酚-氯仿提取(如硫氰酸胍苯酚-氯仿提取)、乙醇沉淀、基于旋转柱的纯化或其他方法。

[0112] 用于测定基因表达产物水平的一般方法可包括但不限于以下一种或多种:其他的细胞学分析、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

因表达的连续分析 (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、DNA测序 (例如,从RNA获得的互补脱氧核糖核酸 (cDNA) 的测序);下一代测序 (Next-Gen)、纳米孔测序、焦磷酸测序或Nanosttring测序。基因表达产物水平可以相对于内标如总信使核糖核酸 (mRNA) 或特定基因的表达水平进行归一化。彼此比较的样品之间,例如来自受试者的样品与参考样品之间的基因表达可存在特定差异或差异范围。基因表达水平的差异可以为至少5%、10%、15%、20%、25%、30%、35%、40%、45%或50%或更大。在一些情况下,基因表达水平的差异可以为至少2、3、4、5、6、7、8、9、10倍或更大。

[0113] RNA测序可产生两个或更多个特征空间,如特定样品的序列变体和基因表达的计数的存在。例如,RNA测序测量在特定组织或特定样品,如甲状腺组织或甲状腺结节中表达的基因的变体。下一代测序可提供特定样品的基因表达水平数据。可将测序结果,如RNA测序和下一代测序结果输入分类器中,该分类器可组合独特的特征空间,从而以比使用单个特征空间更高的准确度确定疾病发生风险。分类器或经训练的算法可包括使用已知恶性、良性和正常样品的参考组开发的算法。分类器或经训练的算法可包括使用已知低风险、中风险和高风险样品的参考组开发的算法。

#### 用于阵列杂交、测序、扩增的标志物

[0114] 用于进行阵列杂交、核酸测序、核酸扩增或其他扩增反应的合适试剂包括但不限于DNA聚合酶、标志物如正向和反向引物、脱氧核苷三磷酸 (dNTP) 以及一种或多种缓冲液。这样的试剂可包括针对感兴趣的给定序列,如第一组基因和/或第二组基因的一个或多个基因所选择的引物。

[0115] 在这样的扩增反应中,引物对中的一个引物可以是与靶多核苷酸分子 (例如,第一组或第二组的一个或多个基因) 的序列互补的正向引物,而引物对中的另一个引物可以是与靶多核苷酸分子的第二序列互补的反向引物,并且靶基因座可位于第一序列和第二序列之间。

[0116] 正向引物和反向引物的长度可取决于靶多核苷酸 (例如第一组或第二组的一个或多个基因) 和靶基因座的序列。在一些情况下,引物可以为大于或等于约5、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、65、70、75、80、85、90、95或约100个核苷酸的长度。作为替代方案,引物可以为少于约100、95、90、85、80、75、70、65、60、59、58、57、56、55、54、53、52、51、50、49、48、47、46、45、44、43、42、41、40、39、38、37、36、35、34、33、32、31、30、29、28、27、26、25、24、23、22、21、20、19、18、17、16、15、14、13、12、11、10、9、8、7、6个或其左右的核苷酸的长度。在一些情况下,引物可以为约15至约20、约15至约25、约15至约30、约15至约40、约15至约45、约15至约50、约15至约55、约15至约60、约20至约25、约20至约30、约20至约35、约20至约40、约20至约45、约20至约50、约20至约55、约20至约60、约20至约80或约20至约100个核苷酸的长度。

[0117] 可根据已知参数设计引物,以避免二级结构和自杂交,如引物二聚体对。不同的引物对可在大约相同的温度下,例如,在另一引物对的1℃、2℃、3℃、4℃、5℃、6℃、7℃、8℃、9℃或10℃之内退火和解链。

[0118] 靶基因座可以是多个模板多核苷酸从3'端或5'端开始的约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、100、150、200、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、510、520、530、540、550、560、570、580、590、600、650、700、750、800、850、900或1000个核苷酸。

[0119] 用于所述方法的标志物(即引物)可以是一个或多个相同的引物。在一些情况下,标志物可以是一个或多个不同的引物,如约2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000个或更多个不同的引物。在这样的实例中,所述一个或多个引物中的每个引物可包含不同的靶特异性区域或序列或者模板特异性区域或序列,如第一组或第二组的一个或多个基因。

[0120] 所述一个或多个引物可包括固定的引物组。所述一个或多个引物可包括至少一个或多个定制引物。所述一个或多个引物可包括至少一个或多个对照引物。所述一个或多个引物可包括至少一个或多个管家基因引物。在一些情况下,所述一个或多个定制引物与靶特异性区域或其互补区域退火。所述一个或多个引物可被设计用于扩增或进行引物延伸、逆转录、线性延伸、非指数扩增、指数扩增、PCR或者一种或多种靶多核苷酸或模板多核苷酸的任何其他扩增方法。

[0121] 引物可并入允许对引物进行检测或固定但不改变引物的基本性质(例如,充当DNA合成起始点)的其他特征。例如,引物可包含不与靶核酸杂交但有助于对扩增产物进行克隆或进一步扩增或测序的在5'端的核酸序列。例如,该序列可包含引物结合位点,如PCR引发序列、样品条形码序列或通用引物结合位点等。

[0122] 通用引物结合位点或序列可使通用引物与多核苷酸和/或扩增子连接。通用引物可包括-47F(M13F)、alfaMF、A0X3'、A0X5'、BGHr、CMV-30、CMV-50、CVMf、LACrmt、λgt10F、λgt10R、λgt11F、λgt11R、M13 rev、M13Forward(-20)、M13Reverse、male、p10SEQPpQE、pA-120、pet4、pGAP Forward、pGLRVpr3、pGLpr2R、pKLAC14、pQEFS、pQERS、pucU1、pucU2、reversA、seqIREStam、seqIRESzpet、seqori、seqPCR、seqpIRES-、seqpIRES+、seqpSecTag、seqpSecTag+、seqretro+PSI、SP6、T3-prom、T7-prom和T7-termInv。如本文所用的,连接可指共价相互作用和非共价相互作用两者或其任一者。通用引物与通用引物结合位点的连接可用于多核苷酸和/或扩增子的扩增、检测和/或测序。

#### 风险测定的应用

[0123] 可将分类器的结果如疾病发生风险或来自本文公开的方法的数据,如基因表达水平或序列变体数据输入数据库中以供分子谱分析企业的代表或代理人、个体、医学专业人员或保险提供者访问。可自动提供数据的计算机或算法分析。结果可呈现为计算机屏幕上的报告或作为纸件报告。在一些情况下,结果可自动上传至数据库或远程服务器。报告可包括但不限于如以下的一种或多种的信息:原始样品的适合性、差异表达的基因的名称和/或数目、具有序列变体的基因的名称和/或数目、序列变体的类型、差异表达的基因的表达水平、数值分类器得分、对受试者的诊断、诊断的统计学置信度、疾病发生风险、指定的疗法或其任意组合。

[0124] 可使用本文所述的方法在单个时间点或多个时间点对受试者进行监测。例如,可使用本文所述的方法诊断受试者患有疾病如癌症或遗传病。在一些情况下,这种初步诊断可能不涉及使用本文所述的方法。随后可为具有阳性疾病诊断如甲状腺癌的受试者指定治

疗性干预,如甲状腺切除术,或开始药物方案,如化疗。可通过使用本文所述的方法检测治疗性干预的效果,来持续地监测治疗性干预的结果。在另一个实例中,否则没有癌症的受试者可能被诊断为具有癌症发生的风险,并且可通过本文所述的方法来检测其健康状况的任何状态变化,来持续地对该受试者进行监测,以便确定癌症是否可能会在稍后的时间点出现或影响进行筛查方法的频率。

[0125] 如本文所述的方法还可用于在对受试者施用特定治疗性干预之前确定该特定治疗性干预的潜在效果。例如,受试者可被诊断为患有癌症。如本文所述的方法可指示已知与癌症恶性有关的基因产物(例如RAS癌基因)的基因表达的高水平。可以获得来自具有所述高水平的受试者的样品并将该样品在体外培养。然后可测试应用异常活化的或失调的途径的各种抑制剂或已知抑制该途径活性的药物对样品的肿瘤细胞的生长抑制。分子谱分析也可用于监测这些抑制剂对例如有关途径的下游靶标的作用。分子谱分析也可用于预测这些抑制剂的效果。

[0126] 本文所述的方法可用作研究工具,以鉴定用于诊断疾病如癌症的新标志物;监测药物或候选药物对于样品如肿瘤细胞、细胞系、组织或生物体的作用;或者揭示疾病进展或抑制如癌症发生和/或肿瘤抑制的新途径。

[0127] 本文所述的方法可提供:1) 包含少量和/或低质量的核酸的样品的基因表达分析;2) 假阳性和假阴性的明显减少,3) 导致所产生的病理学的基础遗传学、代谢或信号传导途径的确定,4) 将统计学概率赋予疾病如遗传病的诊断准确性的能力,5) 解析不明确的结果的能力,6) 区分疾病如癌症的亚型的能力,以及7) 区分低疾病发生风险与中高疾病发生风险的能力。

[0128] 预测可依赖于准确的训练标记。例如,如图10中所示,使用美国甲状腺协会(ATA)分级系统将在Afirma基因表达分类器(GEC)第1版中被标记或分类为组织学恶性的样品进一步标记或分类为低发生风险或中/高发生风险。对于被标记为低发生风险的样品,组织病理学报告可描述一个或多个风险特征的缺乏。对于被标记为中/高发生风险的样品,组织病理学报告可将一个或多个风险特征描述为阳性存在。风险特征可以是淋巴结转移、血管浸润、甲状腺外延伸或其任意组合。

[0129] 可使用包含特定癌症亚型的单个组织样品,例如,包含乳头状甲状腺癌(PTC)的组织样品来训练风险分类器。在一些情况下,使用包含两种、三种、四种或更多种癌症亚型,例如PTC、LCT、HA和FC的单个组织样品来训练风险分类器。在一些情况下,可使用多于一个组织样品,例如两个组织样品来训练风险分类器,其中该两个组织样品包含两种、三种、四种或更多种癌症亚型,例如PTC、LCT、HA和FC。

### 试剂盒

[0130] 疾病诊断企业、分子谱分析企业、制药企业或与患者医疗保健相关的其他企业可提供用于进行疾病发生风险确定的试剂盒。试剂盒可包括分类器、用于训练算法的样品群组,以及每个特征空间的基因列表,如第一组基因和第二组基因。在一些情况下,试剂盒可包括分类器和每个特征空间的基因列表。试剂盒可以是用于所有疾病类型的通用试剂盒。试剂盒可以是用于特定疾病如癌症的特定试剂盒,或针对疾病亚型如甲状腺癌的特定试剂盒。试剂盒可提供已使用试剂盒中未提供的样品群组进行训练的分类器。试剂盒可提供样品群组的定期更新或与分类器一起使用的特征空间的基因列表。试剂盒可提供自动形成可

由医学专业人员报告或展示或下载和/或输入数据库的结果概述的软件。所述结果概述可包括本文公开的任何结果,包括对患者治疗选择的建议和疾病发生风险。试剂盒还可提供用于从受试者获得样品的单元或装置(例如,具有与抽吸器耦合的针的装置)。试剂盒还可提供用于进行如本文所公开的方法的说明书,并且包括用于RNA测序和下一代(NextGen)测序的所有必需的缓冲液和试剂。试剂盒还可包括用于分析结果的说明书。这样的说明书可包括指导用户使用软件(例如,具有经训练的算法的软件)和数据库以分析结果。

#### 计算机控制系统

[0131] 本公开内容提供了计算机控制系统,其被编程为用以实现本公开内容的方法。图9显示了计算机系统9001,其被编程或以其他方式配置为实现本文提供的方法。计算机系统9001可以调节本公开内容的疾病发生风险分层的各个方面,例如,运行分类器或训练算法以及报告经分层的发生风险。计算机系统9001可以是用户或计算机系统的电子装置,该用户或计算机系统相对于该电子装置位于远处。该电子装置可以是移动电子装置。

[0132] 计算机系统9001包括中央处理器(CPU,本文还称为“处理器”和“计算机处理器”)9005,其可以是单核或多核处理器或多个处理器进行并行处理。计算机系统9001还包括存储器或存储器位置9010(例如,随机存取存储器、只读存储器、闪存存储器)、电子存储单元9015(例如,硬盘)、用于与一个或多个其他系统通信的通信接口9020(例如,网络适配器)和外围装置9025,外围装置9025诸如高速缓冲存储器、其他存储器、数据存储和/或电子显示适配器。存储器9010、存储单元9015、接口9020和外部装置9025通过通信总线(实线)与CPU 9005进行通信,所述通信总线诸如为母线。存储单元9015可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统9001在通信接口9020的辅助下可以操作地耦合到计算机网络(“网络”)9030。网络9030可以是因特网、互联网和/或外联网或与因特网通信的内联网和/或外联网。在一些情况下,网络9030是远程通信和/或数据网络。网络9030可以包括一个或多个计算机服务器,这可使得分布计算(诸如云计算)称为可能。在一些情况下,网络9030在计算机系统9001的辅助下可以实现对等网络,其可使得装置耦合到计算机系统9001以起到客户端或服务器的作用。

[0133] CPU 9005可以执行一系列嵌入在程序或软件中的机器可读指令。所述指令可存储在存储器位置如存储器9010中。所述指令可被导向CPU 9005,其随后可对CPU 9005进行编程或以其他方式配置以实现本公开内容的方法。由CPU 9005执行的操作的实例可包括取指令、解码、执行和写回。

[0134] CPU 9005可以是电路如集成电路的一部分。系统9001的一个或多个其他组件可包含在该电路中。在一些情况下,该电路是专用集成电路(ASIC)。

[0135] 存储单元9015可以存储文件,诸如存储驱动器、文库和已保存的程序。存储单元9015可以存储用户数据,例如存储用户偏好和用户程序。在一些情况下,计算机系统9001可以包括位于计算机系统9001外部的一个或多个附加的数据存储单元,所述数据存储单元诸如位于与计算机系统9001通过内联网或因特网通信的远程服务器上。

[0136] 计算机系统9001可以与一个或多个远程计算机系统通过网络9030进行通信。例如,计算机系统9001可以与用户(例如,服务提供者)的远程计算机系统通信。远程计算机系统的实例包括个人计算机(例如,便携式PC)、平板或平板电脑(例如,Apple® iPad、

Samsung® Galaxy Tab)、电话、智能电话(例如, Apple® iPhone、Android功能的装置, Blackberry®)或个人数码助手。用户可以通过网络9030访问计算机系统9001。

[0137] 如本文所述的方法可通过存储在计算机系统9001的电子存储单元上的机器(例如, 计算机处理器)可执行代码来实现, 所述机器可执行代码例如存储在存储器9010或电子存储单元9015上。该机器可执行或机器可读代码可以以软件的形式提供。在使用期间, 所述代码可由处理器9005执行。在一些情况下, 所述代码可以从存储单元9015中取回并存储到存储器9010中以供处理器9005访问。在一些情况下, 可以不包括电子存储单元9015, 而且机器可执行指令被存储在存储器9010上。

[0138] 所述代码可以进行预编译并配置用于与具有适于执行代码的处理器一起使用, 或者可以在运行过程中进行编译。所述代码可以以编程语言来提供, 所述编程语言可以被选择以使该代码能够以预编译或实时编译的方式来执行。

[0139] 本发明提供的系统和方法的各方面, 诸如计算机系统9001, 可以以编程来实施。所述技术的各个方面可以被认为是“产品”或“制品”, 一般以机器(或处理器)可执行代码和/或相关联数据的形式装载在或嵌入在一种类型的机器可读介质中。机器可执行代码可以被存储在电子存储单元, 所述电子存储单元诸如是存储器(例如, 只读存储器、随机存取存储器、闪存存储器)或硬盘。“存储”型介质可以包括计算机、处理器等等的任何或全部有形存储器或与其相关的模块, 诸如包括各种半导体存储器、磁带驱动器、磁盘驱动器等等, 这可以随时为软件编程提供非临时存储。该软件的全部或部分有时可以通过因特网或其他各种远程通信网络进行通信。这样的通信例如可以使软件能够从一台计算机或处理器加载到另一个, 例如从管理服务器或主机加载到应用程序服务器的计算机平台。因此, 可承载软件元件的另一种类型的介质包括光、电和电磁波, 诸如通过有线和光学陆线网络以及经由各种空中链接跨越本地装置之间的物理接口。携带这些波(诸如有线或无线链接、光学链接等等)的物理元件也可以被认为是承载所述软件的介质。如本文所用, 除非限于非临时的、有形的“存储”介质, 诸如计算机或机器“可读介质”等术语指的是参与向处理器提供执行指令的任何介质。

[0140] 因此, 诸如计算机可执行代码的机器可读介质可以采取许多形式, 其包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质例如包括光盘或磁盘, 光盘或磁盘诸如在任何计算机等等中的任何存储装置, 所述存储装置诸如可用于实现附图所示的数据库等等。易失性存储介质包括动态存储器, 诸如包括这种计算机平台的主存储器。有形传输介质包括同轴电缆; 铜线和光纤, 它们包括包含计算机系统内总线的导线。载波传输介质可采取诸如在射频(RF)和红外(IR)数据通信期间所生成的电或电磁信号或声波或光波的形式。因此, 计算机可读介质的常见形式包括例如: 软盘、柔性盘、硬盘、磁带、任何其他磁性介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、穿孔卡片纸带、具有孔图形的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储器芯片或盒、传送数据或指令的载波、传送这种载波的电缆或链接或计算机可以从其中读取程序代码和/或数据的任何其他介质。许多这些形式的计算机可读介质可参与将一个或多个指令的一个或多个序列装载至处理器以供执行。

[0141] 计算机系统9001可以包括电子显示器9035或与电子显示器9035通信, 电子显示器

9035包括用于例如提供分类器或经训练的算法的输出或读出的用户界面(UI) 9040。UI的实例包括但不限于图形用户界面(GUI) 和基于网络的用户界面。

[0142] 本公开内容的方法和系统可以以一个或多个算法的方式来实现。算法可以通过软件经中央处理器9005执行来实现。该算法可以,例如,对疾病发生风险进行分层或将样品分类为良性的、恶性的、疑似的或非诊断性的。

#### 实施例1:使用风险分类器对样品的风险分层

[0143] 甲状腺癌初始管理的现有风险调整方法是基于采用2009年美国甲状腺协会(ATA) 分级系统进行的将受试者分类为中高发生风险或低发生风险的术后分类。虽然这种解剖学分级系统可在临床上使用,但其不能在甲状腺切除术之前被准确地评估,并且不能包括受试者结果的任何分子预测因子。这项研究确定,是否可在甲状腺手术之前使用在恶性甲状腺结节诊断性细针抽吸(FNA) 期间获得的转录数据来提高风险分层。

[0144] 来自样品的FNA材料在术前收集( $n=79$ ),并由专家组在手术后诊断为乳头状甲状腺癌(PTC),包括经典的组织学亚型(图1和图2)。使用既定的发生风险分层指南将每名患者归类为“低风险”或“中高风险”。获得全基因组RNA序列(RNASeq) 数据(每个样品8000万个读取),并使用监督式学习来训练分类器;该分类器包括支持向量机(SVM)、随机森林(RF)、惩罚逻辑回归(PLR) 和这三者的集成。使用相同样品群组上的10倍交叉验证来测定分类器性能。

[0145] 使用320个基因和针对BRAF基因状态控制的开源软件DESeq模型构建分类器。对于最大接受者工作特性(ROC) 曲线下面积(AUC) 为0.86的支持向量机(SVM) 分类器,观察到“低风险”与“中高风险”的最大分类性能(图3和图4)。所有分类器均达到相似的AUC:RF 0.82, PLR 0.82和集成0.84。被发现可用于分类中的基因属于多种跨膜信号传导途径,包括ECM-受体相互作用、粘着斑和细胞粘附分子(图5)。所评估的分类器使用优化总准确度而不偏向灵敏度或特异性的阈值。当支持向量机(SVM) 分类器被应用于样品群组时,其正确地鉴定了79.3% (23/29) 的美国甲状腺协会(ATA) 低风险肿瘤和82.0% (41/50) 的ATA中高风险肿瘤(图5)。

#### 实施例2:交叉验证模型

[0146] 采用具有突变组的基因表达分类器(GEC) 测试不确定的甲状腺结节,以确定通过使用机器学习是否提高了术前风险分层。图10为示出训练标记确定的流程图。采用Afirma GEC第1版训练标记来区分组织学良性样品和组织学恶性样品。使用美国甲状腺协会(ATA) 风险训练标记将组织学恶性样品进一步区分为低发生风险和中/高发生风险。中/高风险特征包括淋巴结转移、血管浸润、甲状腺外延伸或其任意组合。风险训练样品群组在图1中示出。具有中/高组织学特征发生风险的样品的百分比在图2中示出。进行10倍交叉验证以评估包括线性支持向量机(SVM)、随机森林、GLMNet和集成分类器在内的不同学习模型的曲线下面积(AUC)。在该实施例中,最佳模型是集成分类器,其AUC为0.871(如图11A中所示),灵敏度为86%(如图11B中所示)且特异性为86%(如图11B中所示),阳性预测值为91.3%(PPV) 且阴性预测值为78.3%(NPV)。初始特征空间为850个初始特征,包括50个计数和800个变体。最佳性能是使用240个组合特征。来自自由分类器以每种倍数选择的变体的最优(top) 特征在图12中示出。来自自由分类器以10倍选择8至10次的计数的最优特征在图13中示出。

### 实施例3:突变分析

[0147] 收集细针抽吸 (FNA) 样品 ( $n=81$ ), 并由专家组在手术后诊断为恶性的 (乳头状甲状腺癌 (PTC)、多病灶乳头状甲状腺癌 (mPTC)、乳头状甲状腺癌的滤泡变型 (FVPTC)、具有高细胞特征的乳头状甲状腺癌 (PTC-TCV)、甲状腺髓样癌 (MTC)、分化良好但并未特化的癌 (well-differentiated carcinoma-not otherwise specified, WDC-NOS)、肝细胞癌 (HCC)、滤泡性癌 (FC)) 或良性的 (良性家族性嗜中性粒细胞减少症 (BFN)、纤维腺瘤 (FA)、肝细胞腺瘤 (HCA)、玻璃样变性梁状腺瘤 (HTA)、Leydig细胞瘤 (LCT))。还分析了具有组织病理学真实性的手术组织样品 ( $n=57$ )。还分析了来自不具有组织病理学的临床实验室改进修正 (CLIA) 实验室的连续系列的不确定 FNA ( $n=101$ )。对样品进行下一代测序 (NGS), 并且在五个不同突变组中用递增数目的探询基因组位点和融合对来评估 14 种基因 (图 14)。如图 14 中所示, 上方的表格示出了五个突变组中每一组的基因组位点数目和融合对数目。突变组 1 由 9 个基因组位点和 3 个融合对组成。突变组 2 由 19 个基因组位点和 25 个融合对组成。突变组 3 由 208 个基因组位点和 25 个融合对组成。突变组 4 由 929 个基因组位点和 25 个融合对组成。突变组 5 由 3670 个基因组位点和 25 个融合对组成。图 14 的下方表格示出了一个或多个突变组中靶向的 14 种基因。

[0148] 应用数个过滤器对数据进行评分。当没有融合或点突变存在时, 将样品评分为阴性。如果检测到至少一个融合或点突变 (除了鸟嘌呤核苷酸结合蛋白、 $\alpha$  刺激 (GNAS) 突变、被视为良性肿瘤的标志物的标志物), 则将样品评分为阳性。

[0149] 在所有样品群组中检测恶性肿瘤的灵敏度随着基因座数目的增加而提高。特异性显示出相反的趋势, 在所有样品群组中随着基因座数目的增加而降低。在图 15 的 FNA 样品中, 最小的 9 位点组达到了 53% 的灵敏度和 93% 的特异性。图 15 中的最大组 (3670 个位点) 达到了 100% 的灵敏度和 10% 的特异性。

[0150] 在图 17 的手术组织 ( $n=38$ ) 中, 观察到类似的趋势。对总计 57 个组织进行了评估。然而, 仅 38 个组织具有待用于测试性能计算的明确的组织学良性或组织学恶性病理学。在图 17 的最小的 9 位点组中, 89% 的特异性与 45% 的灵敏度相关联。在图 17 的最密集组 (3670 个位点) 中, 100% 的灵敏度与 0% 的特异性相关联。

[0151] 总体而言, 图 15 和图 17 的两个较大组将 87-90% 的组织学良性 FNA 错误地判定为恶性, 而图 15 和图 17 的两个较小组遗漏了 48-58% 的已知癌症。五个组中 CLIA FNA 样品的突变和融合频率分别为 13%、4%、21%、89% 和 92%。通过检测数目逐渐增多的点突变和融合而获得的灵敏度以特异性为代价, 并且在真正良性样品中具有过度评价为恶性的风险。

[0152] 具有 208 个位点的组 3 中的细胞学突变性能在图 16 中示出。该组由 Bethesda 细胞学分类进行划分, Bethesda 细胞学分类包括细胞学良性 (Cyto B)、未确定意义的异型性/未确定意义的滤泡性病变 (AUS/FLUS)、滤泡性肿瘤/疑似滤泡性肿瘤 (FN/SFN)、疑似恶性肿瘤 (SFM)、细胞学恶性 (Cyto M) 和所有样品。每个组的数个参数, 包括样品总数、组织学良性突变数/总数、组织学恶性突变数/总数、灵敏度、特异性, 在图 16 中示出。

[0153] 观察到的 CLIA FNA 样品的突变频率的图示在图 18A 中示出。突变阳性样品 (组 3) 以深灰色表示。GNAS 阳性结节以浅灰色表示。百分比突变频率被细分为不同的组, 包括整体组、AUS/FLUS 组和 FN/SFN 组。图 18B 示出了也在图 18A 中示出的各亚组中用组 3 检测到的基因和突变的表格。



[0154] 观察到的FNA样品的突变频率的图示在图19A中示出。突变阳性结节(组3)以深灰色表示。结节以大小比例示出,并且最小结节=1厘米(cm)。百分比突变频率被细分为不同的组,包括整体组、组织学恶性组和组织学良性组。图19B示出了也在图19A中示出的各亚组中用组3检测到的基因和突变的表格。

[0155] 观察到的组织样品的突变频率的图示在图20A中示出。突变阳性样品(组3)以深灰色表示。GNAS阳性结节以浅灰色表示。百分比突变频率被细分为不同的组,包括整体组、组织学恶性组、组织学良性组和组织学不符合要求组或非诊断组。图20B示出了也在图20A中示出的各亚组中用组3检测到的基因和突变的表格。

本发明提供了包括但不限于以下实施方式:

1. 一种用于评估受试者的组织样品以确定所述受试者的疾病发生风险的方法,该方法包括:

(a) 获得与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平,所述第一组基因与所述受试者的所述疾病发生风险相关;

(b) 确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在,所述第二组基因与所述受试者的所述疾病发生风险相关;

(c) 分别将(i)在(a)中获得的所述表达水平以及(ii)在(b)中获得的所述核酸序列与对照进行比较,以提供所述表达水平以及所述核酸序列与所述对照的比较,其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在;以及

(d) 使用采用经训练的算法编程的计算机处理器,(i)分析所述比较,以及(ii)基于所述比较确定所述疾病的所述发生风险。

2. 根据实施方式1所述的方法,其中所述疾病为癌症。

3. 根据实施方式1所述的方法,其进一步包括在(a)之前,从所述受试者获得所述针抽吸样品。

4. 根据实施方式1所述的方法,其进一步包括在(a)之前,从所述针抽吸样品中的所述核酸样品确定所述表达水平。

5. 根据实施方式1所述的方法,其进一步包括在(b)之前,从所述针抽吸样品中的所述核酸样品确定所述核酸序列。

6. 根据实施方式5所述的方法,其进一步包括将所述核酸序列与所述参考序列进行比较,以鉴定所述一个或多个序列变体。

7. 根据实施方式6所述的方法,其中所述参考序列为来自所述受试者的管家基因。

8. 根据实施方式1所述的方法,其中所述第一组或第二组基因中的所述一个或多个基因包括多个基因。

9. 根据实施方式1所述的方法,其中所述针抽吸样品已被发现是细胞学上不明确的或可疑的。

10. 根据实施方式1所述的方法,其中所述针抽吸样品具有约1微升或更小的体积。

11. 根据实施方式1所述的方法,其中所述针抽吸样品具有约9.0或更小的RNA完整性指数(RIN)值。

12. 根据实施方式10所述的方法, 其中所述针抽吸样品具有约6.0或更小的RIN值。

13. 根据实施方式1所述的方法, 其中所述疾病的所述发生风险包括所述受试者的所述疾病的复发风险。

14. 根据实施方式2所述的方法, 其中所述癌症的所述发生风险包括在所述受试者中转移的风险。

15. 根据实施方式1所述的方法, 其中所述经训练的算法采用来自至少25名已被诊断为患有所述疾病的受试者的组织样品进行训练。

16. 根据实施方式15所述的方法, 其中所述经训练的算法采用来自至少200名已被诊断为患有所述疾病的受试者的组织样品进行训练。

17. 根据实施方式1所述的方法, 其中 (d) 发生在术前。

18. 根据实施方式1所述的方法, 其中 (d) 发生在所述受试者具有阳性疾病诊断之前。

19. 根据实施方式1所述的方法, 其中 (d) 进一步包括将所述发生风险分层为低发生风险或中高发生风险, 其中所述低发生风险具有约50%至约80%的发生概率, 并且其中所述中高发生风险具有约80%至100%的发生概率。

20. 根据实施方式19所述的方法, 其中所述分层具有至少80%的准确度。

21. 根据实施方式19所述的方法, 其中所述分层具有至少80%的特异性。

22. 根据实施方式1所述的方法, 其进一步包括将一个或多个过滤器、一个或多个包装器、一个或多个嵌入式协议或其任意组合应用于所述比较。

23. 根据实施方式22所述的方法, 其进一步包括将所述一个或多个过滤器应用于所述比较。

24. 根据实施方式23所述的方法, 其中所述一个或多个过滤器包括t检验、方差分析 (ANOVA)、贝叶斯框架、伽马分布、Wilcoxon秩和检验、类间内平方和检验、秩乘积法、随机置换法、误分类数目阈值 (TNoM)、双变量法、基于相关性的特征选择 (CFS) 法、最小冗余最大相关性 (MRMR) 法、Markov毯过滤法、非相关收缩重心法或其任意组合。

25. 根据实施方式23所述的方法, 其中所述一个或多个序列变体包含点突变、融合基因、置换、缺失、插入、倒位、转变、易位或其任意组合中的一种或多种。

26. 根据实施方式25所述的方法, 其中所述一个或多个点突变为约5个至约4000个点突变。

27. 根据实施方式25所述的方法, 其中所述一个或多个融合基因为至少两个融合基因。

28. 根据实施方式1所述的方法, 其中所述第一组或第二组的所述一个或多个基因为少于约15个基因。

29. 根据实施方式1所述的方法, 其中所述第一组或第二组的所述一个或多个基因为少于约75个基因。

30. 根据实施方式1所述的方法, 其中所述第一组或第二组的所述一个或多个基因为约50个至约400个基因。

31. 根据实施方式1所述的方法, 其中 (b) 中的所述获得包括对所述FNA样品中的核酸样品进行测序以获得所述核酸序列。

32. 根据实施方式31所述的方法,其中所述测序包括富集第二组基因的所述一个或多个基因或其变体。

33. 根据实施方式1所述的方法,其中(a)包括使用具有对所述第一组基因的所述一个或多个基因具有选择性的探针的微阵列。

34. 根据实施方式1所述的方法,其中所述组织样品为甲状腺组织样品。

35. 根据实施方式34所述的方法,其中所述第一组和第二组基因包含COL1A1、THBS2或其任意组合。

36. 根据实施方式34所述的方法,其中所述第二组基因包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROB01、C6orf136、SPAG4、GALNT15、LUM、NCAM2、NUP210L、NR2F1、THBS2、PSORS1C1或其任意组合。

37. 根据实施方式34所述的方法,其中所述第一组基因包含COL1A1、TMEM92、C1orf87、SPAG4、EHF、COL3A1、GALNT15、NUP210L、PDZRN3、C6orf136、NA、NRXN3、COL6A3、RAPGEF5、PRICKLE1、LUM、ROB01、BGN、AC019117.2、PRSS3P1或其任意组合。

38. 根据实施方式34所述的方法,其中所述第二组基因包含EPHA3、COL1A1、EHF、RAPGEF5、PRICKLE1、TMEM92、ROB01、C6orf136、SPAG4、GALNT15、LUM、NCAM2、SYNP02、NUP210L、AMZ1、NR2F1、THBS2、PSORS1C1、FTH1P24或其任意组合。

39. 根据实施方式34所述的方法,其中所述第二组基因包含AKAP9、SPRY3、SPRY3、CAMKK2、COL1A1、FITM2、COX6C、VSIG10L、CYC1、KDM1B、MAPK15、ARSG、PAXIP1、DAAM1、AVL9、DMGDH、HLA-DQA1、HLA-DQB1、HLA-DRA、HLA-DRB5、HLA-H、IRF1、MGAT1、P2RX1、PLEK、CCDC93、PPP1R12C、SLC41A3、METTL3、CCAR2、PTPRE、SRL、SLC30A5、BMP4、ZNF133、ICE2、DCAKD、TMX1、TNFSF12、PER2、MCM3AP或其任意组合。

40. 根据实施方式1所述的方法,其中所述第一组基因和所述第二组基因是不同的。

41. 根据实施方式1所述的方法,其进一步包括鉴定所述疾病的新型遗传生物标志物。

42. 根据实施方式1所述的方法,其中(a)中的所述获得包括测定与所述一个或多个基因中的每一个相对应的所述表达水平。

43. 根据实施方式42所述的方法,其中所述测定包括使用针对所述一个或多个基因中的每一个所选择的标志物进行的阵列杂交、核酸测序或核酸扩增。

44. 根据实施方式43所述的方法,其中所述标志物为针对所述一个或多个基因中的每一个所选择的引物。

45. 根据实施方式43所述的方法,其中所述测定包括逆转录聚合酶链反应(PCR)。

46. 根据实施方式1所述的方法,其中所述确定包括测定所述核酸样品中所述第二组基因的所述一个或多个基因中的每一个。

47. 根据实施方式46所述的方法,其中所述测定包括使用针对所述一个或多个基因中的每一个所选择的标志物进行的阵列杂交、核酸测序或核酸扩增。

48. 根据实施方式47所述的方法,其中所述标志物为针对所述一个或多个基因中的每一个所选择的引物。

49. 根据实施方式47所述的方法,其中所述测定包括逆转录聚合酶链反应(PCR)。

50. 根据实施方式1所述的方法, 其中所述针抽吸样品为细针抽吸样品。

51. 一种用于评估受试者的组织样品以确定所述受试者的疾病发生风险的系统, 该系统包括:

一个或多个计算机存储器, 该计算机存储器存储 (a) 与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达, 所述第一组基因与所述受试者的所述疾病发生风险相关, 以及 (b) 与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在的指示, 所述第二组基因与所述受试者的所述疾病发生风险相关; 以及

计算机处理器, 其与所述一个或多个计算机存储器耦合并被编程为:

(i) 分别将 (1) 所述计算机存储器中的所述表达水平以及 (2) 所述核酸序列与对照进行比较, 以提供所述表达水平以及所述核酸序列与所述对照的比较, 其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在; 以及

(ii) 使用经训练的算法 (1) 分析所述比较以及 (2) 基于所述比较确定所述疾病的所述发生风险。

52. 一种非暂时性计算机可读介质, 其包含机器可执行代码, 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用于评估受试者的组织样品以确定所述受试者的疾病发生风险的方法, 所述方法包括:

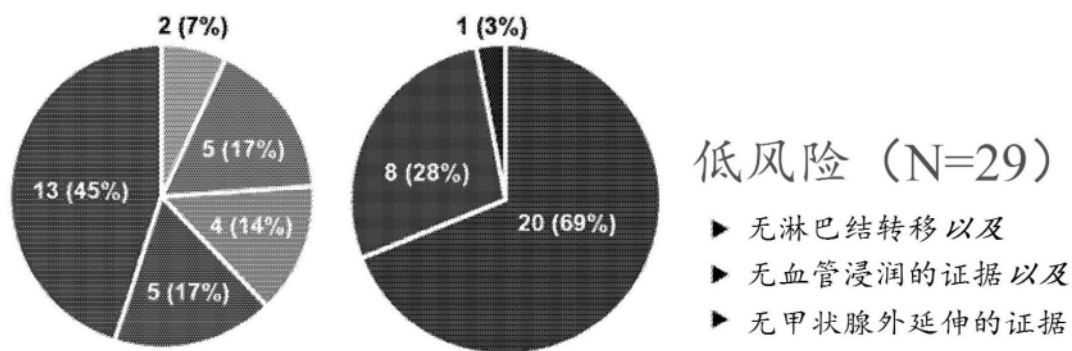
(a) 获得与从所述受试者获得的针抽吸样品中的核酸样品中第一组基因的一个或多个基因中的每一个相对应的表达水平, 所述第一组基因与所述受试者的所述疾病发生风险相关;

(b) 确定与所述核酸样品中第二组基因的一个或多个基因中的每一个相对应的核酸序列的存在, 所述第二组基因与所述受试者的所述疾病发生风险相关;

(c) 分别将 (i) 在 (a) 中获得的所述表达水平以及 (ii) 在 (b) 中获得的所述核酸序列与对照进行比较, 以提供所述表达水平以及所述核酸序列与所述对照的比较, 其中所述核酸序列与所述对照中的参考序列的比较指示针对于所述第二组基因的给定基因的一个或多个序列变体的存在; 以及

(d) 使用采用经训练的算法编程的计算机处理器, (i) 分析所述比较, 以及 (ii) 基于所述比较确定所述疾病的所述发生风险。

[0156] 尽管本文已经显示并描述了本发明的优选实施方案, 但对本领域技术人员而言显而易见的是, 这些方案仅以举例的方式提供。它并不意味着本发明受到本说明书中提供的具体实施例的限制。虽然本发明已经参照上述的说明书进行了描述, 本文实施方案的说明书和图示并不意味着以限制的意义来解释。在不脱离本发明的情况下, 本领域技术人员将会想到许多变化、改变和替代。此外, 应当理解的是, 本发明的所有方面不限于本文所述的具体的描述、构造或相对比例, 其依赖于各种条件和变量。应理解, 本文描述的本发明实施方案的各种替代方案可用于实施本发明。因此可以预期的是, 本发明还应当包括任何这样的替代、改变、变化或等同物。下列权利要求旨在限定本发明的范围, 从而涵盖这些权利要求的范围内的方法和结构及其等同物。

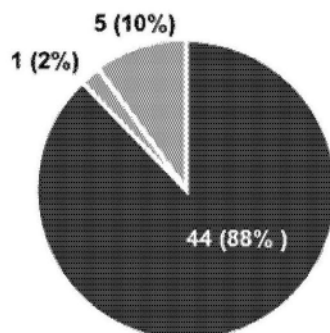
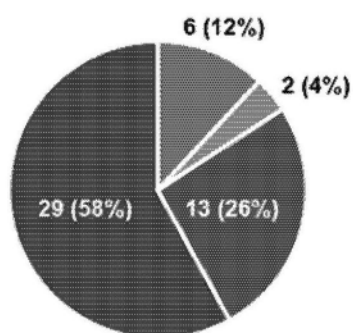


细胞学

专家组织病理学

■ B  
■ AUS/FLUS  
■ FN/SFN  
■ SFM  
■ M

■ PTC  
■ FVPTC  
■ WDC-NOS  
■ PDC  
■ PTC-TCV

**中/高风险 (N=50)**

- ▶ 存在淋巴结转移和/或
- ▶ 存在血管浸润和/或
- ▶ 存在甲状腺外延伸

细胞学

专家组织病理学

图1

组织病理学特征	样品数 (%)
仅存在淋巴结转移	11 (22%)
仅存在血管浸润	1 (2%)
仅存在甲状腺外延伸	15 (30%)
淋巴结转移+血管浸润	1 (2%)
淋巴结转移+甲状腺外延伸	17 (34%)
血管浸润+甲状腺外延伸	1 (2%)
淋巴结转移+血管浸润+甲状腺外延伸	4 (8%)
总计	50 (100%)

图2

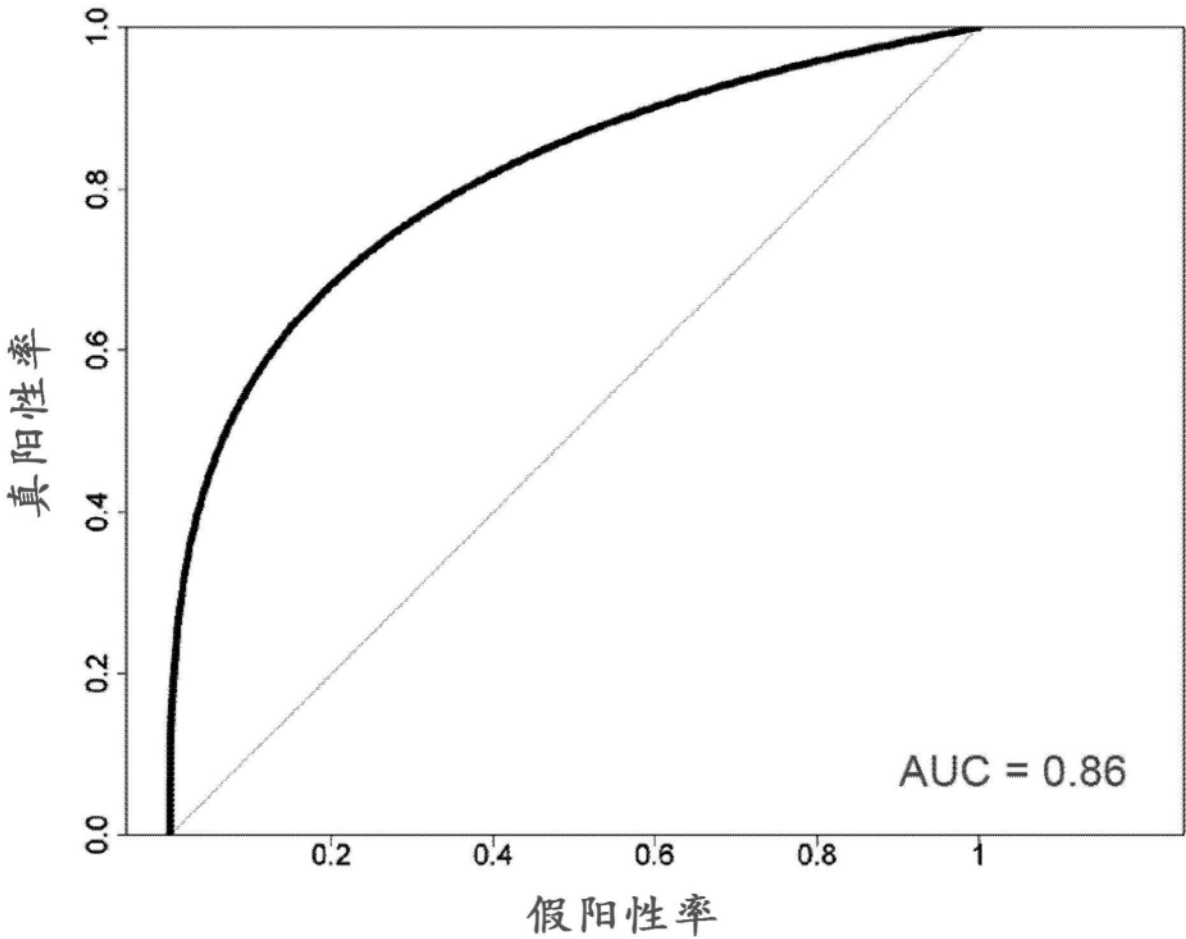


图3

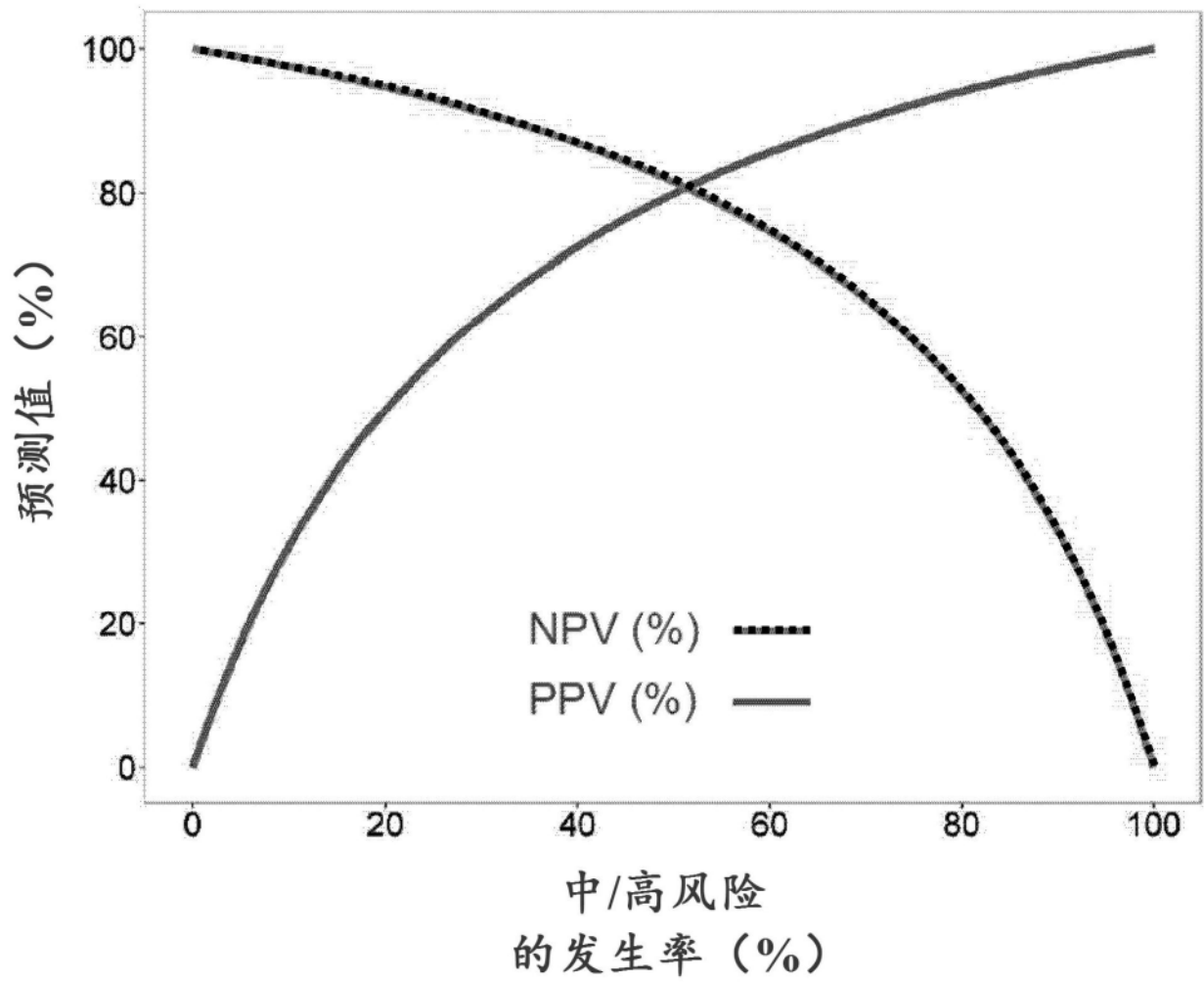


图4

## 各风险组的性能

	中/高风险 (N=50)	低风险 (N=29)
分类为 “中/高风险”	41	6
分类为 “低风险”	9	23
灵敏度	82%	
特异性	79%	

途径或 基因本体	预期的 基因数	观察到的 基因数	FDR P值
细胞外基质	2	20	$1.28 \times 10^{-13}$
ECM受体 相互作用	1	11	$6.12 \times 10^{-08}$
粘着斑	2	12	$3.71 \times 10^{-05}$
酪氨酸 激酶活性	1	6	$4.77 \times 10^{-03}$
免疫系统 过程的调控	3	10	$6.83 \times 10^{-02}$
血管发育	1	7	$1.0 \times 10^{-02}$

用于分类的基因的抽样				
COX6C	FANCA	KCTD17	MPRIIP	TUBA1B
DCAKD	ICE2	MCM3AP	TNFRSF14	WSB2

图5



Ensembl基因ID	ENTREZID	基因符号	说明	CV倍数 频率
ENSG00000044524	2042	EPHA3	EPH受体A3 [来源: HGNC符号; Acc:HGNC:3387]	10
ENSG00000108821	1277	COL1A1	胶原蛋白, I型, $\alpha 1$ [来源: HGNC符号; Acc:HGNC:2197]	10
ENSG00000135373	26298	EHF	ets同源因子 [来源: HGNC符号; Acc:HGNC:3246]	10
ENSG00000136237	9771	RAPGEF5	Rap鸟嘌呤核苷酸交换因子(GEF) 5 [来源: HGNC符号; Acc:HGNC:16862]	10
ENSG00000139174	144165	PRICKLE1	皮刺同源物1 (果蝇) [来源: HGNC符号; Acc:HGNC:17019]	10
ENSG00000167105	162461	TMEM92	跨膜蛋白92 [来源: HGNC符号; Acc:HGNC:26579]	10
ENSG00000169855	6091	ROBO1	迂回轴突导向受体同源物1 (果蝇) [来源: HGNC 符号; Acc:HGNC:10249]	10
ENSG00000204564	221545	C6orf136	染色体6开放阅读框136 [来源: HGNC符号; Acc:HGNC:21301]	10
ENSG00000061656	6676	SPAG4	精子相关抗原4 [来源: HGNC符号; Acc:HGNC:11214]	9
ENSG00000131386	117248	GALNT15	多肽N-乙酰氨基半乳糖基转移酶15 [来源: HGNC符号; Acc:HGNC:21531]	9
ENSG00000139329	4060	LUM	腔蛋白聚糖 [来源: HGNC符号; Acc:HGNC:6724]	9
ENSG00000154654	4685	NCAM2	神经细胞粘附分子2 [来源: HGNC符号; Acc:HGNC:7657]	9
ENSG00000172403	171024	SYNPO2	突触足蛋白2 [来源: HGNC符号; Acc:HGNC:17732]	9
ENSG00000143552	91181	NUP210L	核孔蛋白210kDa样 [来源: HGNC符号; Acc:HGNC:29915]	8
ENSG00000174945	155185	AMZ1	archaelysin家族金属肽酶1 [来源: HGNC符 号; Acc:HGNC:22231]	8
ENSG00000175745	7025	NR2F1	核受体亚家族2, F组, 成员1 [来源: HGNC符 号; Acc:HGNC:7975]	8
ENSG00000186340	7058	THBS2	血小板反应蛋白2 [来源: HGNC符号; Acc:HGNC:11786]	8
ENSG00000204540	170679	PSORS1C1	银屑病敏感性1候选物1 [来源: HGNC符号; Acc:HGNC:17202]	8
ENSG00000249302	NA	FTH1P24	铁蛋白, 重链多肽1假基因24 [来源: HGNC符 号; Acc:HGNC:37642]	8

图6

ENSG00000236039	NA	AC019117.2		7
ENSG00000158062	91544	UBXN11	UBX结构域蛋白11 [来源: HGNC符号; Acc:HGNC:30600]	6
ENSG00000178750	415117	STX19	突触融合蛋白19 [来源: HGNC符号; Acc:HGNC:19300]	6
ENSG00000196366	158055	C9orf163	染色体9开放阅读框163 [来源: HGNC符号; Acc:HGNC:26718]	6
ENSG00000254352	NA	RP11-578O24.2		6
ENSG00000269845	641367	RP11-420K14.6		6
ENSG00000132470	3691	ITGB4	整联蛋白, $\beta 4$ [来源: HGNC符号; Acc:HGNC:6158]	5
ENSG00000185567	113146	AHNAK2	AHNAK核蛋白2 [来源: HGNC符号; Acc:HGNC:20125]	5
ENSG00000232185	NA	CNOT7P2	CCR4-NOT转录复合体, 亚基7假基因2 [来源: HGNC符号; Acc:HGNC:44249]	5
ENSG00000232466	NA	RP11-337A23.5		5
ENSG00000000971	3075	CFH	补体因子H [来源: HGNC符号; Acc:HGNC:4883]	4
ENSG00000066248	25791	NGEF	神经元鸟嘌呤核苷酸交换因子 [来源: HGNC符号; Acc:HGNC:7807]	4
ENSG00000170454	9119	KRT75	角蛋白75 [来源: HGNC符号; Acc:HGNC:24431]	4
ENSG00000226824	100996437	RP4-756H11.3		4
ENSG00000238961	677828	SNORA47	小核仁RNA, H/ACA盒47 [来源: HGNC符号; Acc:HGNC:32640]	4
ENSG00000243020	NA	RPL7P39	核糖体蛋白L7假基因39 [来源: HGNC符号; Acc:HGNC:36214]	4
ENSG00000261649	728310	GOLGA6L7P	高尔基体蛋白A6家族样7, 假基因 [来源: HGNC符号; Acc:HGNC:37442]	4
ENSG00000141540	94015	TTYH2	tweety家族成员2 [来源: HGNC符号; Acc:HGNC:13877]	3
ENSG00000141748	390790	ARL5C	ADP核糖基化因子样5C [来源: HGNC符号; Acc:HGNC:31111]	3
ENSG00000149968	4314	MMP3	基质金属蛋白酶3 (溶基质蛋白酶1, 前明胶酶) [来源: HGNC符号; Acc:HGNC:7173]	3

图6 (续)

ENSG00000163359	1293	COL6A3	胶原蛋白, VI型, $\alpha 3$ [来源: HGNC符号; Acc:HGNC:2213]	3
ENSG00000204345	100131439	CD300LD	CD300分子样家族成员d [来源: HGNC符号; Acc:HGNC:16848]	3
ENSG00000269700	NA	NA	NA	3
ENSG00000168765	2948	GSTM4	谷胱甘肽S-转移酶mu4 [来源: HGNC符号; Acc:HGNC:4636]	2
ENSG00000175274	9537	TP53I11	肿瘤蛋白p53可诱导蛋白11 [来源: HGNC符号; Acc:HGNC:16842]	2
ENSG00000196569	3908	LAMA2	层粘连蛋白 $\alpha 2$ [来源: HGNC符号; Acc:HGNC:6482]	2
ENSG00000231172	101927884	AC007099.1		2
ENSG00000254708	NA	RP1-145M24.1		2
ENSG00000267337	101927921	LINC01478	长基因间非蛋白质编码RNA 1478 [来源: HGNC符号; Acc:HGNC:51121]	2
ENSG00000050344	9603	NFE2L3	核因子, 红细胞2样3 [来源: HGNC符号; Acc:HGNC:7783]	1
ENSG00000100362	5816	PVALB	小白蛋白 [来源: HGNC符号; Acc:HGNC:9704]	1
ENSG00000102780	160851	DGKH	二酰基甘油激酶, $\eta$ [来源: HGNC符号; Acc:HGNC:2854]	1
ENSG00000115414	2335	FN1	纤连蛋白1 [来源: HGNC符号; Acc:HGNC:3778]	1
ENSG00000117069	81849	ST6GALNAC5	ST6( $\alpha$ -N-乙酰神经胺基-2,3- $\beta$ -半乳糖基-1,3)-N-乙酰氨基半乳糖基 $\alpha$ -2,6-唾液酸转移酶5 [来源: HGNC符号; Acc:HGNC:19342]	1
ENSG00000137819	54852	PAQR5	孕酮和脂联素受体家族成员V [来源: HGNC符号; Acc:HGNC:29645]	1
ENSG00000145107	116211	TM4SF19	跨膜4 L六家族成员19 [来源: HGNC符号; Acc:HGNC:25167]	1
ENSG00000152894	5796	PTPRK	蛋白酪氨酸磷酸酶, 受体类型, K [来源: HGNC符号; Acc:HGNC:9674]	1
ENSG00000166016	25841	ABTB2	含有锚蛋白重复和BTB (POZ)域的2 [来源: HGNC符号; Acc:HGNC:23842]	1
ENSG00000196639	3269	HRH1	组胺受体H1 [来源: HGNC符号; Acc:HGNC:5182]	1
ENSG00000203585	100507175	RP11-542B15.1		1
ENSG00000235897	100874214	TM4SF19-AS1	TM4SF19反义RNA 1 [来源: HGNC符号; Acc:HGNC:41085]	1
ENSG00000240518	NA	RP11-515C16.1		1
ENSG00000241679	NA	RP11-80H8.4		1
ENSG00000250770	NA	RP5-1063M23.1		1
ENSG00000254429	NA	CTD-2562J17.7		1
ENSG00000267857	NA	RP5-1023B21.1		1

图6 (续)

Ensembl基因ID	entrez基因ID	外部基因名称	基因说明	排序	2倍变化的对数	p值	FDR p值
ENSG00000108821	1277	COL1A1	胶原蛋白, I型, $\alpha 1$	1	1.47	1.66E-07	2.44E-03
ENSG00000167105	162461	TMEM92	跨膜蛋白92	2	1.60	1.95E-07	2.44E-03
ENSG00000162598	127795	Clorf87	染色体1开放阅读框87	3	-1.58	9.19E-07	7.70E-03
ENSG00000061656	6676	SPAG4	精子相关抗原4	4	1.17	6.74E-06	2.85E-02
ENSG00000135373	26298	EHF	ets同源因子	5	1.24	6.80E-06	2.85E-02
ENSG00000168542	1281	COL3A1	胶原蛋白, III型, $\alpha 1$	6	1.39	5.86E-06	2.85E-02
ENSG00000131386	117248	GALNT15	多肽N-	7	1.41	9.33E-06	3.28E-02
ENSG00000143552	91181	NUP210L	核孔蛋白210kDa样	8	-1.25	1.04E-05	3.28E-02
ENSG00000121440	23024	PDZRN3	含有PDZ结构域的环指3	9	1.40	1.47E-05	4.11E-02
ENSG00000204564	221545	C6orf136	染色体6开放阅读框136	10	-0.57	1.84E-05	4.63E-02
ENSG00000250606	NA	NA	NA	11	1.43	2.14E-05	4.89E-02
ENSG00000021645	9369	NRXN3	神经元表面蛋白3	12	-1.41	2.53E-05	5.28E-02
ENSG00000163359	1293	COL6A3	胶原蛋白, VI型, $\alpha 3$	13	1.22	2.73E-05	5.28E-02
ENSG00000136237	9771	RAPGEF5	Rap鸟嘌呤核苷酸交换因子	14	0.77	4.43E-05	5.83E-02
ENSG00000139174	144165	PRICKLE1	皮刺同源物1 (果蝇)	15	0.87	4.22E-05	5.83E-02
ENSG00000139329	4060	LUM	腔蛋白聚糖	16	1.25	4.64E-05	5.83E-02
ENSG00000169855	6091	ROBO1	迂回轴突导向受体同源物1(果蝇)	17	0.91	3.94E-05	5.83E-02
ENSG00000182492	633	BGN	双糖链蛋白聚糖	18	1.19	3.52E-05	5.83E-02
ENSG00000236039	NA	AC019117.2		19	1.34	4.15E-05	5.83E-02
ENSG00000250591	NA	PRSS3P1	丝氨酸蛋白酶, 3假基因1	20	1.39	3.30E-05	5.83E-02
ENSG00000179954	284297	SSC5D	富含半胱氨酸的清道夫受体家族, 5	21	1.20	5.17E-05	6.18E-02
ENSG00000261649	728310	GOLGA6L7P	高尔基体蛋白A6家族样7, 假基因	22	-1.34	6.38E-05	7.28E-02
ENSG00000254352	NA	RP11-578O24.2		23	1.13	6.83E-05	7.46E-02
ENSG00000132470	3691	ITGB4	整合蛋白, $\beta 4$	24	1.00	8.14E-05	8.52E-02
ENSG00000204540	170679	PSORS1C1	银屑病敏感性1候选物1	25	1.09	9.00E-05	9.05E-02
ENSG00000044524	2042	EPHA3	EPH受体A3	26	1.14	1.09E-04	9.79E-02
ENSG00000154654	4685	NCAM2	神经细胞粘附分子2	27	1.14	1.07E-04	9.79E-02
ENSG00000231172	101927884	AC007099.1		28	1.13	1.03E-04	9.79E-02
ENSG00000136197	79020	C7orf25	染色体7开放阅读框25	29	0.87	1.28E-04	1.11E-01

图7

ENSG000000164100	9348	NDST3	N-脱乙酰酶N-磺基转移酶 (类肝素)	30	1.19	1.47E-04	1.14E-01
ENSG000000196353	131034	CPNE4	copine IV	31	1.24	1.43E-04	1.14E-01
ENSG000000230498	NA	RP4-564M11.2		32	1.27	1.50E-04	1.14E-01
ENSG000000263232	NA	ATP5A1P3	ATP合酶, H <sup>+</sup> 转运, 线粒体 F1复合体, $\alpha$ 亚基1假基因3	33	1.21	1.48E-04	1.14E-01
ENSG000000186340	7058	THBS2	血小板反应蛋白2	34	1.04	1.66E-04	1.22E-01
ENSG000000175745	7025	NR2F1	核受体亚家族2, F组	35	1.18	1.81E-04	1.30E-01
ENSG000000198261	NA	OR5BB1P	嗅觉受体, 家族5, 亚家族BB, 成员1假基因	36	1.25	2.11E-04	1.47E-01
ENSG000000115414	2335	FN1	纤连蛋白1	37	1.13	2.29E-04	1.56E-01
ENSG000000000971	3075	CFH	补体因子H	38	1.02	2.45E-04	1.62E-01
ENSG000000061337	11178	LZTS1	亮氨酸拉链, 推定肿瘤抑制 因子1	39	1.12	2.97E-04	1.72E-01
ENSG000000103485	23475	QPRT	喹啉磷酸核糖转移酶	40	-1.01	4.18E-04	1.72E-01
ENSG000000105894	5764	PTN	多营养因子	41	1.09	3.09E-04	1.72E-01
ENSG000000108448	147166	TRIM16L	含有三重基序的16样	42	-0.89	3.96E-04	1.72E-01
ENSG000000120875	1846	DUSP4	双特异性磷酸酶4	43	0.63	3.08E-04	1.72E-01
ENSG000000128656	1123	CHN1	嵌蛋白1	44	0.89	3.80E-04	1.72E-01
ENSG000000133318	10313	RTN3	浆膜蛋白3	45	-0.31	3.13E-04	1.72E-01
ENSG000000158062	91544	UBXN11	UBX结构域蛋白11	46	-0.47	3.66E-04	1.72E-01
ENSG000000164078	4486	MST1R	巨噬细胞刺激1受体(c-met相 关酪氨酸激酶)	47	0.88	4.12E-04	1.72E-01
ENSG000000166166	115708	TRMT61A	tRNA甲基转移酶61同源物A (S)	48	-0.49	4.13E-04	1.72E-01
ENSG000000170370	2018	EMX2	空呼吸孔同源框2	49	-1.12	4.14E-04	1.72E-01
ENSG000000170486	140807	KRT72	角蛋白72	50	-1.19	3.62E-04	1.72E-01
ENSG000000172403	171024	SYNPO2	突触足蛋白2	51	0.95	3.36E-04	1.72E-01
ENSG000000175197	1649	DDIT3	可诱导DNA损伤的转录物3	52	-0.75	3.86E-04	1.72E-01
ENSG000000178750	415117	STX19	突触融合蛋白19	53	1.01	3.87E-04	1.72E-01
ENSG000000187134	1645	AKR1C1	醛酮还原酶家族1, 成员C1	54	-1.12	2.74E-04	1.72E-01
ENSG000000196565	3048	HBG2	血红蛋白, $\gamma$ G	55	-1.20	3.28E-04	1.72E-01
ENSG000000196639	3269	HRH1	组胺受体H1	56	1.04	2.95E-04	1.72E-01
ENSG000000197977	54898	ELOVL2	ELOVL脂酸延伸酶2	57	1.18	3.95E-04	1.72E-01
ENSG000000198805	4860	PNP	嘌呤核苷酸磷酸化酶	58	0.94	4.14E-04	1.72E-01
ENSG000000214797	NA	RP11-1036E20.9		59	1.10	4.17E-04	1.72E-01

图7(续)

ENSG00000258947	10381	TUBB3	微管蛋白, $\beta 3$ , III类	60	1.05	3.59E-04	1.72E-01
ENSG00000269700	NA	NA	NA	61	0.86	3.81E-04	1.72E-01
ENSG00000116031	50489	CD207	CD207分子, 朗格汉斯蛋白	62	1.18	4.29E-04	1.73E-01
ENSG00000231852	1589	CYP21A2	细胞色素P450, 家族21, 亚家族A	63	1.18	4.33E-04	1.73E-01
ENSG00000124205	1908	EDN3	内皮素3	64	-1.18	4.66E-04	1.79E-01
ENSG00000174945	155185	AMZ1	archaealysin家族金属肽酶1	65	-0.99	4.63E-04	1.79E-01
ENSG00000204866	147920	IGFL2	IGF样家族成员2	66	1.16	4.69E-04	1.79E-01
ENSG00000137648	56649	TMPRSS4	跨膜蛋白酶, 丝氨酸4	67	1.06	4.88E-04	1.83E-01
ENSG00000049323	4052	LTBP1	潜伏转化生长因子 $\beta$	68	1.02	5.15E-04	1.84E-01
ENSG00000111799	1303	COL12A1	胶原蛋白, XII型, $\alpha 1$	69	1.01	5.20E-04	1.84E-01
ENSG00000188783	5549	PRELP	富含脯氨酸/精氨酸和富含亮氨酸的重复序列	70	1.10	5.10E-04	1.84E-01
ENSG00000223482	728190	NUTM2A-AS1	NUTM2A反义RNA 1	71	-0.42	5.17E-04	1.84E-01
ENSG00000105855	3696	ITGB8	整合蛋白, $\beta 8$	72	0.81	5.41E-04	1.89E-01
ENSG00000169908	4071	TM4SF1	跨膜4L六家族成员1	73	1.06	5.60E-04	1.93E-01
ENSG00000113361	1004	CDH6	钙粘着蛋白6, 2型, K-钙粘着蛋白(胎儿肾脏)	74	1.08	5.83E-04	1.96E-01
ENSG00000260774	NA	CTD-2083E4.4		75	0.85	5.84E-04	1.96E-01
ENSG00000254708	NA	RP1-145M24.1		76	0.98	5.93E-04	1.96E-01
ENSG00000206195	503637	DUXAP8	双同源框A假基因8	77	1.15	6.20E-04	2.02E-01
ENSG00000196139	8644	AKR1C3	醛酮还原酶家族1, 成员C3	78	-1.10	6.34E-04	2.04E-01
ENSG00000060718	1301	COL11A1	胶原蛋白, XI型, $\alpha 1$	79	1.12	6.98E-04	2.17E-01
ENSG00000124107	6590	SLPI	分泌型白细胞蛋白酶抑制素	80	1.01	6.88E-04	2.17E-01
ENSG00000261295	NA	RP11-524D16_A.3		81	1.10	6.94E-04	2.17E-01
ENSG00000103154	54550	NECAB2	N末端EF-手型钙结合蛋白2	82	-1.12	7.15E-04	2.17E-01
ENSG00000116106	2043	EPHA4	EPH受体A4	83	0.76	7.29E-04	2.17E-01
ENSG00000128641	4430	MYO1B	肌球蛋白IB	84	0.74	7.48E-04	2.17E-01
ENSG00000134873	9071	CLDN10	密蛋白10	85	1.13	7.64E-04	2.17E-01
ENSG00000184210	347516	DGAT2L6	二酰基甘油O-酰基转移酶2样6	86	1.13	7.61E-04	2.17E-01
ENSG00000185567	113146	AHNAK2	AHNAK核蛋白2	87	0.94	7.68E-04	2.17E-01
ENSG00000236510	NA	AC011284.3		88	1.13	7.64E-04	2.17E-01
ENSG00000249302	NA	FTH1P24	铁蛋白, 重链多肽1假基因24	89	1.00	7.39E-04	2.17E-01

图7 (续)

ENSG000000115602	9173	IL1RL1	白介素1受体样1	90	1.13	7.81E-04	2.18E-01
ENSG000000117069	81849	ST6GALNAC5	ST6( $\alpha$ -N-乙酰神经胺基-2,3- $\beta$ -半乳糖基-1,3)-N-乙酰氨基半乳糖苷	91	1.11	8.44E-04	2.33E-01
ENSG000000119681	4053	LTP2	潜伏转化生长因子 $\beta$	92	0.73	9.17E-04	2.41E-01
ENSG000000140464	5371	PML	前髓细胞性白血病	93	0.36	9.09E-04	2.41E-01
ENSG000000177453	167359	NIMIK	NIM1 丝氨酸/苏氨酸蛋白激酶	94	-1.10	9.35E-04	2.41E-01
ENSG000000183111	389337	ARHGEF37	Rho鸟嘌呤核苷酸交换因子	95	0.94	9.27E-04	2.41E-01
ENSG000000211670	NA	IGLV3-9	免疫球蛋白 $\lambda$ 变型3-9	96	1.11	8.84E-04	2.41E-01
ENSG000000234779	NA	RPI1-62F24.2		97	1.11	9.41E-04	2.41E-01
ENSG000000240498	100048912	CDKN2B-AS1	CDKN2B反义RNA 1	98	1.01	9.01E-04	2.41E-01
ENSG000000138316	140766	ADAMTS14	含血小板反应蛋白1型基序的ADAM金属肽酶14	99	1.10	9.63E-04	2.43E-01
ENSG000000267799	NA	MAN1A2P1	甘露糖苷酶, $\alpha$ , 1A类, 成员2	100	-1.05	9.72E-04	2.43E-01
ENSG000000272077	NA	RPI1-348P10.2		101	0.79	9.79E-04	2.43E-01
ENSG000000106819	54829	ASPN	asporin	102	0.99	1.04E-03	2.56E-01
ENSG000000120337	8995	TNFSF18	肿瘤坏死因子(配体)超家族	103	1.10	1.06E-03	2.56E-01
ENSG000000268223	644100	ARL14EPL	ADP核糖基化因子样14效应物	104	1.09	1.05E-03	2.56E-01
ENSG000000268460	93429	DKFZp434J0226	非典型的LOC93429	105	1.09	1.09E-03	2.60E-01
ENSG000000074527	59277	NTN4	引导蛋白4	106	0.91	1.13E-03	2.63E-01
ENSG000000144366	51454	GULP1	GULP, 吞入衔接子PTB域	107	1.00	1.15E-03	2.63E-01
ENSG000000220695	NA	RPI-121G13.3		108	0.90	1.13E-03	2.63E-01
ENSG000000245293	101929595	RPI1-286E11.1		109	0.74	1.14E-03	2.63E-01
ENSG000000261424	NA	ATPSF1P7	ATP合酶, H <sup>+</sup> 转运, 线粒体Fo复合体, 亚基B1假基因7	110	1.04	1.15E-03	2.63E-01
ENSG000000066248	25791	NGEF	神经元鸟嘌呤核苷酸交换因子	111	0.90	1.22E-03	2.69E-01
ENSG000000134201	2949	GSTM5	谷胱甘肽S-转移酶mu5	112	-1.07	1.23E-03	2.69E-01
ENSG000000150471	23284	LPHN3	蛛毒素受体3	113	1.06	1.21E-03	2.69E-01
ENSG000000150687	11098	PRSS23	蛋白酶, 丝氨酸, 23	114	0.90	1.23E-03	2.69E-01
ENSG000000164692	1278	COL1A2	胶原蛋白, I型, $\alpha$ 2	115	0.99	1.20E-03	2.69E-01
ENSG000000160191	5152	PDE9A	磷酸二酯酶9A	116	0.65	1.26E-03	2.72E-01
ENSG000000141540	94015	TTYH2	tweety家族成员2	117	-0.67	1.28E-03	2.75E-01
ENSG000000137100	11258	DCTN3	动力蛋白白激蛋白3 (p22)	118	-0.30	1.32E-03	2.78E-01
ENSG000000166183	374569	ASPG	天冬酰胺酶	119	1.03	1.32E-03	2.78E-01

图7 (续)

ENSG00000117152	5999	RGS4	G 蛋白信号传导调节因子4	120	1.06	1.45E-03	2.99E-01
ENSG00000182185	5890	RAD51B	RAD51 种内同源基因B	121	0.32	1.45E-03	2.99E-01
ENSG00000271318	NA	AP000654.5		122	1.07	1.44E-03	2.99E-01
ENSG00000184254	220	ALDH1A3	醛脱氢酶1家族, 成员	123	0.86	1.47E-03	3.00E-01
ENSG00000109255	10874	NMU	神经调节肽U	124	1.07	1.49E-03	3.01E-01
ENSG00000151914	667	DST	肌张力异常蛋白	125	0.73	1.50E-03	3.02E-01
ENSG00000273132	NA	RP11-350J20.12		126	0.94	1.55E-03	3.08E-01
ENSG00000156735	9530	BAG4	BCL2 相关永生基因4	127	-0.27	1.59E-03	3.10E-01
ENSG00000189182	374454	KRT77	角蛋白77	128	-0.98	1.62E-03	3.10E-01
ENSG00000196182	83931	STK40	丝氨酸/苏氨酸激酶40	129	-0.44	1.62E-03	3.10E-01
ENSG00000225802	NA	WARSP1	色氨酸-tRNA 合成酶假基因1	130	1.01	1.57E-03	3.10E-01
ENSG00000250770	NA	RP5-1063M23.1		131	1.04	1.60E-03	3.10E-01
ENSG00000099715	27328	PCDH11Y	Y-连锁原钙粘着蛋白11	132	0.94	1.65E-03	3.13E-01
ENSG00000050344	9603	NFE2L3	核因子, 红细胞2样3	133	0.62	1.69E-03	3.19E-01
ENSG00000102780	160851	DGKH	二酰基甘油激酶, $\eta$	134	0.47	1.76E-03	3.22E-01
ENSG00000155816	56776	FMN2	成蛋白2	135	-1.01	1.74E-03	3.22E-01
ENSG00000180155	66004	LYNX1	Ly6/神经毒素1	136	0.69	1.76E-03	3.22E-01
ENSG00000189283	2272	FHIT	脆性组氨酸三联体	137	-0.40	1.75E-03	3.22E-01
ENSG00000105246	10148	EBI3	爱泼斯坦-巴尔病毒诱导的3	138	0.83	1.77E-03	3.23E-01
ENSG00000105221	208	AKT2	v-akt 鼠胸腺瘤病毒癌基因	139	-0.38	1.81E-03	3.24E-01
ENSG00000180817	5464	PPA1	焦磷酸酶(无机) 1	140	0.43	1.81E-03	3.24E-01
ENSG00000123901	10888	GPR83	G 蛋白偶联受体83	141	-1.00	1.83E-03	3.26E-01
ENSG00000203585	100507175	RP11-542B15.1		142	0.98	1.84E-03	3.26E-01
ENSG00000233725	121838	LINC00284	长基因间非蛋白质编码RNA 284	143	1.05	1.86E-03	3.28E-01
ENSG00000073584	6605	SMARCE1	SWI/SNF 相关的基质结合肌 动蛋白依赖性染色质调节因 子亚家族	144	0.22	1.92E-03	3.28E-01
ENSG00000110944	51561	IL23A	白介素23, $\alpha$ 亚基p19	145	0.75	1.96E-03	3.28E-01
ENSG00000112936	730	C7	补体成分7	146	1.04	2.00E-03	3.28E-01
ENSG00000117983	727897	MUC5B	粘蛋白5B, 低聚粘液/凝胶形 成	147	1.03	1.99E-03	3.28E-01
ENSG00000138131	84171	LOXL4	赖氨酰氧化酶样4	148	-0.90	1.94E-03	3.28E-01
ENSG00000162444	116362	RBP7	视黄醇结合蛋白7, 细胞的	149	-0.61	2.03E-03	3.28E-01

图7 (续)



ENSG00000176222	342908	ZNF404	锌指蛋白404	150	-0.65	1.98E-03	3.28E-01
ENSG00000205189	65986	ZBTB10	含有锌指和BTB结构域的10	151	0.39	2.01E-03	3.28E-01
ENSG00000230730	NA	AC074011.2		152	-0.99	1.94E-03	3.28E-01
ENSG00000252130	NA	RNU6-1045P	RNA, U6小核1045, 假基因	153	0.98	1.96E-03	3.28E-01
ENSG00000256355	NA	NTANIP3	N末端天冬酰胺酰胺酶假基因3	154	1.00	2.01E-03	3.28E-01
ENSG00000261578	NA	RP11-21L23.2		155	-0.98	2.03E-03	3.28E-01
ENSG00000133083	9201	DCLK1	双皮质蛋白样激酶1	156	0.90	2.12E-03	3.30E-01
ENSG00000135925	80326	WNT10A	无翼型MMTV整合位点家族	157	0.86	2.11E-03	3.30E-01
ENSG00000185710	NA	SMG1P4	SMG1假基因4	158	0.74	2.08E-03	3.30E-01
ENSG00000223946	NA	RP11-533O20.2		159	0.97	2.06E-03	3.30E-01
ENSG00000237187	441094	NR2F1-AS1	NR2F1反义RNA 1	160	0.93	2.10E-03	3.30E-01
ENSG00000261521	NA	NA	NA	161	0.98	2.11E-03	3.30E-01
ENSG00000145423	6423	SFRP2	分泌性卷曲相关蛋白2	162	0.94	2.13E-03	3.31E-01
ENSG00000120278	57480	PLEKHG1	含有普列克底物蛋白同源结构域的, 家族G (具有RhoGef结构域) 成员1	163	0.66	2.16E-03	3.32E-01
ENSG00000060762	51660	MPC1	线粒体丙酮酸载体1	164	-0.37	2.23E-03	3.34E-01
ENSG00000101255	57761	TRIB3	三倍体假性激酶3	165	-0.84	2.24E-03	3.34E-01
ENSG00000104332	6422	SFRP1	分泌性卷曲相关蛋白1	166	-1.03	2.25E-03	3.34E-01
ENSG00000110042	23220	DTX4	84, E3泛素连接酶	167	0.67	2.20E-03	3.34E-01
ENSG00000196586	4646	MYO6	肌球蛋白 VI	168	0.65	2.24E-03	3.34E-01
ENSG00000213934	3047	HBG1	血红蛋白, $\gamma$ A	169	-1.03	2.20E-03	3.34E-01
ENSG00000115009	6364	CCL20	趋化因子(C-C基序)配体20	170	1.01	2.29E-03	3.36E-01
ENSG00000184292	4070	TACSTD2	肿瘤相关钙信号转导蛋白2	171	0.89	2.29E-03	3.36E-01
ENSG00000141337	22901	ARSG	芳香硫酸酯酶G	172	-0.37	2.35E-03	3.44E-01
ENSG00000227014	NA	AC007285.6		173	0.87	2.37E-03	3.44E-01
ENSG00000112249	10973	ASCC3	激活信号辅整合素1复合体	174	0.25	2.47E-03	3.57E-01
ENSG00000248727	102467147	CTC-236F12.4		175	0.73	2.49E-03	3.57E-01
ENSG00000079691	55604	LRRC16A	含有富含亮氨酸的重复序列的16A	176	0.37	2.52E-03	3.58E-01
ENSG00000125144	4495	MT1G	金属硫蛋白1G	177	0.89	2.53E-03	3.58E-01
ENSG00000259354	NA	RP11-519G16.3		178	0.87	2.53E-03	3.58E-01
ENSG00000196366	158055	C9orf163	染色体9开放阅读框163	179	0.89	2.57E-03	3.61E-01

图7 (续)

ENSG00000138685	2247	FGF2	成纤维细胞生长因子2 (碱性)	180	0.76	2.59E-03	3.62E-01
ENSG00000136859	23452	ANGPTL2	血管生成素样蛋白2	181	0.71	2.67E-03	3.67E-01
ENSG00000145147	9353	SLIT2	slit同源物2 (果蝇)	182	0.75	2.74E-03	3.67E-01
ENSG00000157514	1831	TSC22D3	TSC22结构域家族, 成员3	183	-0.51	2.74E-03	3.67E-01
ENSG00000162391	338094	FAM151A	具有序列相似性的家族151, 成员	184	-0.70	2.67E-03	3.67E-01
ENSG00000184258	1038	CDR1	小脑变性相关蛋白1	185	0.77	2.70E-03	3.67E-01
ENSG00000226824	10096437	RP4-756H11.3		186	-0.76	2.70E-03	3.67E-01
ENSG00000237523	439990	LINC00857	长基因间非蛋白质编码RNA 857	187	1.01	2.73E-03	3.67E-01
ENSG00000269305	NA	NA	NA	188	0.90	2.71E-03	3.67E-01
ENSG00000232185	NA	CNOT7P2	CCR4-非转录复合体, 亚基7	189	1.00	2.78E-03	3.69E-01
ENSG00000110427	25758	KIAA1549L	KIAA1549样	190	0.98	2.84E-03	3.72E-01
ENSG00000211959	NA	IGHV4-39	免疫球蛋白重链可变4-39	191	1.00	2.83E-03	3.72E-01
ENSG00000248180	NA	GAPDHP60	甘油醛-3-磷酸脱氢酶	192	-0.87	2.84E-03	3.72E-01
ENSG00000102901	80152	CENPT	着丝粒蛋白T	193	-0.51	2.89E-03	3.75E-01
ENSG00000145623	9180	OSMR	抑癌蛋白M受体	194	0.76	2.93E-03	3.75E-01
ENSG00000196569	3908	LAMA2	层粘连蛋白, $\alpha 2$	195	0.63	2.92E-03	3.75E-01
ENSG00000254471	NA	RP11-885L14.1		196	1.00	2.91E-03	3.75E-01
ENSG00000204022	142910	LIPJ	脂肪酶, 家族成员J	197	-1.00	2.94E-03	3.75E-01
ENSG00000173848	10276	NET1	神经上皮细胞转化因子1	198	0.49	3.00E-03	3.78E-01
ENSG00000176495	390195	OR5AN1	嗅觉受体, 家族5, 亚家族AN	199	1.00	2.99E-03	3.78E-01
ENSG00000197102	1778	DYNC1H1	动力蛋白, 细胞质1, 重链1	200	0.26	3.01E-03	3.78E-01
ENSG00000254959	NA	INMT-FAM188B	INMT-FAM188B通读(NMD)	201	0.98	3.04E-03	3.80E-01
ENSG00000101307	10326	SIRPB1	信号调节蛋白 $\beta 1$	202	-0.81	3.09E-03	3.81E-01
ENSG00000117139	10765	KDM5B	赖氨酸(K)特异的脱甲基酶 5B	203	0.28	3.16E-03	3.81E-01
ENSG00000135549	5570	PKIB	蛋白激酶(cAMP依赖性催化)	204	0.74	3.16E-03	3.81E-01
ENSG00000141428	83608	C18orf21	染色体18开放阅读框21	205	-0.34	3.20E-03	3.81E-01
ENSG00000169436	169044	COL22A1	胶原蛋白, XXII型, $\alpha 1$	206	0.99	3.15E-03	3.81E-01
ENSG00000172667	64393	ZMAT3	锌指, 3型基质	207	0.47	3.09E-03	3.81E-01
ENSG00000205269	100113407	TMEM170B	跨膜蛋白170B	208	-0.54	3.22E-03	3.81E-01
ENSG00000240541	100874091	TM4SF1-AS1	TM4SF1反义RNA 1	209	0.99	3.19E-03	3.81E-01

图7 (续)

ENSG00000244731	720	C4A	补体成分4A (Rodgers血液)	210	0.80	3.21E-03	3.81E-01
ENSG00000248429	101929911	RP11-597D13.9		211	-0.85	3.19E-03	3.81E-01
ENSG00000270599	NA	RP11-324O2.6		212	0.79	3.12E-03	3.81E-01
ENSG00000104823	1891	ECH1	烯酰辅酶A水合酶1, 过氧化 物酶体	213	-0.33	3.25E-03	3.82E-01
ENSG00000236257	NA	EI24P2	依托泊苷诱导的2.4假基因2	214	-0.58	3.25E-03	3.82E-01
ENSG00000254858	84769	MPV17L2	MPV17线粒体膜蛋白样	215	-0.35	3.32E-03	3.88E-01
ENSG00000269845	641367	RP11-420K14.6		216	0.78	3.34E-03	3.89E-01
ENSG00000110675	55531	ELMOD1	含有ELMO/CED-12结构域的 1	217	-0.82	3.40E-03	3.94E-01
ENSG00000136826	9314	KLF4	Kruppel样因子4 (肠)	218	-0.56	3.45E-03	3.94E-01
ENSG00000154175	25890	ABI3BP	ABI家族, 成员3 (NESH)结合 蛋白	219	-0.89	3.46E-03	3.94E-01
ENSG00000227036	400619	LINC00511	长基因间非蛋白质编码RNA 511	220	0.93	3.47E-03	3.94E-01
ENSG00000262678	NA	RP5-1050D4.4		221	0.84	3.44E-03	3.94E-01
ENSG00000166813	374654	KIF7	驱动蛋白家族成员7	222	0.76	3.51E-03	3.94E-01
ENSG00000175274	9537	TP53I1	可诱导肿瘤蛋白p53的蛋白11	223	-0.44	3.49E-03	3.94E-01
ENSG00000260534	NA	RP11-1006G14.4		224	0.46	3.51E-03	3.94E-01
ENSG00000125266	1948	EFNB2	肝配蛋白-B2	225	0.70	3.53E-03	3.94E-01
ENSG00000164296	81789	TIGD6	触发转座元件衍生的6	226	0.41	3.60E-03	4.00E-01
ENSG00000236699	54848	ARHGEF38	Rho鸟嘌呤核苷酸交换因子	227	0.80	3.61E-03	4.00E-01
ENSG00000126016	154796	AMOT	血管动蛋白	228	0.63	3.67E-03	4.04E-01
ENSG00000204262	1290	COL5A2	胶原蛋白, V型, $\alpha 2$	229	0.81	3.71E-03	4.04E-01
ENSG00000213871	NA	TAF9BP1	TAF9B RNA聚合酶II, TATA 盒结合蛋白(TBP)相关因子, 31kDa	230	0.67	3.71E-03	4.04E-01
ENSG00000227242	NA	NBPF13P	神经母细胞瘤断裂家族, 成员 13	231	0.78	3.71E-03	4.04E-01
ENSG00000100053	1417	CRYBB3	晶体蛋白, $\beta$ B3	232	0.92	3.76E-03	4.04E-01
ENSG00000137040	26953	RANBP6	RAN结合蛋白6	233	0.26	3.77E-03	4.04E-01
ENSG00000152894	5796	PTPRK	蛋白酪氨酸磷酸酶, 受体类 型, K	234	0.50	3.74E-03	4.04E-01
ENSG00000108932	9120	SLC16A6	溶质运载蛋白家族16, 成员6	235	-0.85	3.82E-03	4.05E-01
ENSG00000125872	164312	LRRN4	富含亮氨酸的神经元重复序 列4	236	0.92	3.84E-03	4.05E-01
ENSG00000197694	6709	SPTAN1	血影蛋白, $\alpha$ , 非红细胞的1	237	0.27	3.79E-03	4.05E-01
ENSG00000197808	92283	ZNF461	锌指蛋白461	238	0.34	3.80E-03	4.05E-01
ENSG00000233929	NA	MT1XP1	金属硫蛋白1X假基因1	239	0.84	3.85E-03	4.05E-01

图7 (续)

ENSG00000069020	375449	MAST4	微管相关的丝氨酸/苏氨酸激酶家族成员4	240	0.46	3.87E-03	4.05E-01
ENSG00000103034	65009	NDRG4	NDRG家族成员4	241	0.72	3.93E-03	4.08E-01
ENSG00000211972	NA	IGHV3-66	免疫球蛋白重链可变3-66	242	0.95	3.93E-03	4.08E-01
ENSG00000261553	NA	RP11-29G8.3		243	0.64	3.95E-03	4.08E-01
ENSG00000186416	55922	NKRF	NFKB阻遏因子	244	0.25	3.98E-03	4.10E-01
ENSG00000087916	NA	NA	NA	245	0.97	4.01E-03	4.10E-01
ENSG00000211979	NA	IGHV7-81	免疫球蛋白重链可变7-81 (非)	246	0.97	4.03E-03	4.10E-01
ENSG00000248626	NA	GAPDHP40	甘油醛3磷酸脱氢酶	247	0.85	4.01E-03	4.10E-01
ENSG00000241665	NA	RN7SL418P	RNA, 7SL, 细胞质的418, 假基因	248	0.95	4.08E-03	4.13E-01
ENSG00000249693	100506564	THEGL	the精子细胞蛋白样	249	-0.96	4.16E-03	4.20E-01
ENSG00000136110	11061	LECT1	白细胞衍生的趋化因子1	250	0.94	4.18E-03	4.20E-01
ENSG00000234521	NA	AC005041.11		251	0.96	4.20E-03	4.20E-01
ENSG00000171658	NA	RP11-443P15.2		252	-0.96	4.24E-03	4.23E-01
ENSG00000231564	NA	EIF4A1P11	真核生物翻译起始因子4A1	253	0.80	4.29E-03	4.25E-01
ENSG00000248213	NA	CICP16	回文转录阻遏物假基因	254	-0.96	4.29E-03	4.25E-01
ENSG00000115970	63892	THADA	与甲状腺腺瘤相关	255	0.20	4.41E-03	4.30E-01
ENSG00000173267	6623	SNCG	突触核蛋白, $\gamma$ (乳腺癌特异的)	256	0.88	4.41E-03	4.30E-01
ENSG00000206538	389136	VGLL3	残翅样家族成员3	257	0.92	4.42E-03	4.30E-01
ENSG00000232907	101926987	DLGAP4-AS1	DLGAP4反义RNA 1	258	0.77	4.37E-03	4.30E-01
ENSG00000157765	10568	SLC34A2	溶质运载蛋白家族34 (II型钠/磷酸盐协同转运蛋白), 成员2	259	0.88	4.48E-03	4.34E-01
ENSG00000255185	NA	PDXDC2P		260	0.77	4.49E-03	4.34E-01
ENSG00000091128	22798	LAMB4	层粘连蛋白, $\beta 4$	261	-0.95	4.66E-03	4.35E-01
ENSG00000124588	4835	NQO2	NAD(P)H脱氢酶, 醌2	262	-0.52	4.53E-03	4.35E-01
ENSG00000127603	23499	MACF1	微管-肌动蛋白交联因子1	263	0.32	4.62E-03	4.35E-01
ENSG00000162840	NA	MT2P1	金属硫蛋白2假基因1	264	0.71	4.59E-03	4.35E-01
ENSG00000168765	2948	GSTM4	谷胱甘肽S-转移酶mu4	265	-0.48	4.62E-03	4.35E-01
ENSG00000204869	400706	IGFL4	IGF样家族成员4	266	0.95	4.67E-03	4.35E-01
ENSG00000215796	NA	RP11-551G24.2		267	0.79	4.65E-03	4.35E-01
ENSG00000241679	NA	RP11-80H8.4		268	0.78	4.55E-03	4.35E-01
ENSG00000253641	101929191	RP11-981G7.2		269	0.93	4.65E-03	4.35E-01

图7 (续)

ENSG00000256092	100293704	MIR8072	微RNA 8072	270	0.73	4.58E-03	4.35E-01
ENSG00000105976	4233	MET	MET原癌基因, 受体酪氨酸激酶	271	0.72	4.74E-03	4.38E-01
ENSG00000225125	NA	RANP4	RAN, 成员RAS癌基因家族	272	-0.91	4.74E-03	4.38E-01
ENSG00000171812	1296	COL8A2	胶原蛋白, VIII型, $\alpha 2$	273	0.70	4.79E-03	4.40E-01
ENSG00000173227	91683	SYT12	突触结合蛋白XII	274	0.88	4.83E-03	4.43E-01
ENSG00000175087	149420	PDIK1L	PDLIM1相互作用激酶1样	275	0.44	4.86E-03	4.44E-01
ENSG00000125449	79637	ARMC7	含有抗徐重序列的7	276	-0.41	4.93E-03	4.46E-01
ENSG00000137819	54852	PAQR5	孕酮和脂联素受体家族成员	277	-0.86	4.99E-03	4.46E-01
ENSG00000142920	113451	AZIN2	抗酶抑制剂2	278	0.62	5.00E-03	4.46E-01
ENSG00000145247	132299	OCLAD2	含有OCIA结构域的2	279	0.42	4.99E-03	4.46E-01
ENSG00000157542	3763	KCNJ6	向内整流钾通道, 亚家族J, 成员6	280	0.94	4.99E-03	4.46E-01
ENSG00000182704	25987	TSKU	tsukushi, 富含小亮氨酸的蛋白聚糖	281	-0.70	4.97E-03	4.46E-01
ENSG00000255132	NA	RP11-63C8.1		282	-0.91	4.96E-03	4.46E-01
ENSG00000145332	57563	KLHL8	Kelch样家族成员8	283	-0.28	5.08E-03	4.49E-01
ENSG00000254632	NA	RP11-21L23.4		284	-0.94	5.07E-03	4.49E-01
ENSG00000166741	4837	NNMT	烟酰胺N-甲基转移酶	285	0.87	5.16E-03	4.54E-01
ENSG00000101542	28316	CDH20	钙粘着蛋白20, 2型	286	0.92	5.25E-03	4.56E-01
ENSG00000102699	143	PARP4	聚(ADP-核糖)聚合酶家族, 成员	287	0.31	5.37E-03	4.56E-01
ENSG00000121904	114784	CSMD2	CUB和Sushi多重结构域2	288	0.94	5.39E-03	4.56E-01
ENSG00000122756	1271	CNTFR	睫状神经营养因子受体	289	-0.86	5.27E-03	4.56E-01
ENSG00000130294	547	KIF1A	驱动蛋白家族成员1A	290	-0.93	5.44E-03	4.56E-01
ENSG00000162643	126820	WDR63	WD重复结构域63	291	-0.69	5.39E-03	4.56E-01
ENSG00000174839	201627	DENND6A	含有DENN/MADD结构域的6A	292	-0.29	5.30E-03	4.56E-01
ENSG00000176340	1351	COX8A	细胞色素c氧化酶亚基VIII A	293	-0.39	5.24E-03	4.56E-01
ENSG00000182404	NA	NA	NA	294	0.94	5.41E-03	4.56E-01
ENSG00000184156	3786	KCNQ3	钾电压门控通道, KQT样亚家族, 成员3	295	0.73	5.50E-03	4.56E-01
ENSG00000189136	388165	UBE2Q2P1	泛素缀合酶E2Q家族成员2假基因1	296	0.55	5.48E-03	4.56E-01
ENSG00000196208	9687	GREB1	乳腺中雌激素的生长调节	297	-0.89	5.43E-03	4.56E-01
ENSG00000207971	406911	MIR125B1	微RNA 125b-1	298	0.66	5.27E-03	4.56E-01
ENSG00000223419	NA	AC006022.4		299	0.93	5.19E-03	4.56E-01

图7 (续)

ENSG00000225731	101928284	AP001627.1			300	0.91	5.45E-03	4.56E-01
ENSG00000258844	NA	RP11-259K15.2			301	0.83	5.34E-03	4.56E-01
ENSG00000260647	NA	RP1-178F10.1			302	0.81	5.31E-03	4.56E-01
ENSG00000272365	NA	RP11-389C8.3			303	0.80	5.49E-03	4.56E-01
ENSG00000092607	6913	TBX15		T盒15	304	0.93	5.59E-03	4.57E-01
ENSG00000102962	6367	CCL22		趋化因子(C-C基序)配体22	305	0.89	5.63E-03	4.57E-01
ENSG00000104432	3574	IL7		白介素7	306	0.59	5.75E-03	4.57E-01
ENSG00000113083	4015	LOX		赖氨酰氧化酶	307	0.84	5.69E-03	4.57E-01
ENSG00000116785	10878	CFHR3		补体因子H相关的3	308	0.91	5.65E-03	4.57E-01
ENSG00000134013	4017	LOXL2		赖氨酰氧化酶样2	309	0.61	5.60E-03	4.57E-01
ENSG00000141441	64762	GAREM		GRB2相关的, MAPK1的调节因子	310	0.70	5.75E-03	4.57E-01
ENSG00000160211	2539	G6PD		葡萄糖-6-磷酸脱氢酶	311	-0.43	5.68E-03	4.57E-01
ENSG00000171992	11346	SYNPO		突触足蛋白	312	0.51	5.69E-03	4.57E-01
ENSG00000185664	6490	PMEL		前黑素体蛋白	313	0.90	5.72E-03	4.57E-01
ENSG00000232466	NA	RP11-337A23.5			314	0.75	5.75E-03	4.57E-01
ENSG00000248360	102724794	LINC00504		长基因间非蛋白质编码RNA 504	315	-0.86	5.70E-03	4.57E-01
ENSG00000256234	NA	RP11-283G6.4			316	0.75	5.69E-03	4.57E-01
ENSG0000010438	5646	PRSS3		蛋白酶, 丝氨酸, 3	317	0.93	5.84E-03	4.59E-01
ENSG00000110090	1374	CPT1A		肉碱棕榈酰转移酶1A (肝脏)	318	-0.53	5.84E-03	4.59E-01
ENSG00000151917	221336	BEND6		含有BEN结构域的6	319	0.85	5.85E-03	4.59E-01
ENSG00000249509	NA	RP11-402J6.1			320	-0.88	5.82E-03	4.59E-01
ENSG00000130558	10439	OLFM1		嗅介蛋白1	321	-0.81	5.89E-03	4.60E-01
ENSG00000254542	NA	NAV2-AS3		NAV2反义RNA 3	322	0.82	5.89E-03	4.60E-01
ENSG00000050555	10319	LAMC3		层粘连蛋白, γ3	323	-0.92	5.95E-03	4.62E-01
ENSG00000100453	3002	GZMB		粒酶B (粒酶2, 细胞毒性T淋巴细胞相关丝氨酸酯酶1)	324	-0.78	5.97E-03	4.62E-01
ENSG00000202569	574447	MIR146B		微RNA 146b	325	0.85	5.96E-03	4.62E-01
ENSG00000078098	2191	FAP		成纤维细胞活化蛋白, α	326	0.92	6.10E-03	4.64E-01
ENSG00000079739	5236	PGM1		葡萄糖磷酸变位酶1	327	0.33	6.05E-03	4.64E-01
ENSG00000129566	7011	TEP1		端粒酶相关蛋白1	328	0.35	6.05E-03	4.64E-01
ENSG00000175315	1474	CST6		半胱氨酸蛋白酶抑制剂E/M	329	0.80	6.08E-03	4.64E-01

图7 (续)

ENSG00000261049	NA	RP11-357N13.1			330	0.81	6.09E-03	4.64E-01
ENSG00000268734	NA	CTB-61M7.2			331	-0.82	6.13E-03	4.65E-01
ENSG00000001626	1080	CFTR		囊性纤维化跨膜传导调节蛋白 (ATP结合盒亚家族C)	332	0.87	6.22E-03	4.67E-01
ENSG00000121361	3764	KCNJ8		钾内向整流通道, 亚家族J, 成员8	333	-0.90	6.21E-03	4.67E-01
ENSG00000180662	NA	RPL2IP8		核糖体蛋白L21假基因8	334	0.66	6.18E-03	4.67E-01
ENSG00000237442	NA	HNRNP A1P57		核内不均一核糖核蛋白A1	335	0.92	6.23E-03	4.67E-01
ENSG000000006210	6376	CX3CL1		趋化因子(C-X3-C基序)配体1	336	0.72	6.31E-03	4.67E-01
ENSG00000132329	10267	RAMP1		受体(G蛋白偶联)活性	337	0.88	6.30E-03	4.67E-01
ENSG00000164808	23514	SPIDR		参与DNA修复的支架蛋白	338	-0.20	6.33E-03	4.67E-01
ENSG00000187045	164656	TMPRSS6		跨膜蛋白酶, 丝氨酸6	339	0.87	6.27E-03	4.67E-01
ENSG00000214018	NA	RRM2P3		核糖核甘酸还原酶M2多肽	340	-0.88	6.28E-03	4.67E-01
ENSG00000087128	28983	TMPRSS11E		跨膜蛋白酶, 丝氨酸11E	341	0.91	6.46E-03	4.76E-01
ENSG00000268568	NA	AC007228.9			342	0.64	6.52E-03	4.79E-01
ENSG00000082497	56256	SERTAD4		含有SERTA结构域的4	343	0.62	6.60E-03	4.80E-01
ENSG00000135205	57639	CCDC146		含有卷曲螺旋结构域的146	344	-0.58	6.61E-03	4.80E-01
ENSG00000188385	282973	JAKMIP3		詹纳斯激酶和微管相互作用	345	-0.91	6.59E-03	4.80E-01
ENSG00000203836	NA	NA		NA	346	-0.43	6.60E-03	4.80E-01
ENSG00000095787	51322	WAC		具有卷曲的含有WW结构域的衔接子	347	-0.20	6.82E-03	4.81E-01
ENSG00000104833	10382	TUBB4A		微管蛋白, $\beta$ 4AIVa类	348	-0.88	6.83E-03	4.81E-01
ENSG00000129009	3671	ISLR		含有免疫球蛋白超家族	349	0.70	6.84E-03	4.81E-01
ENSG00000132003	65249	ZSWIM4		锌指, 含有SWIM-类型的4	350	0.44	6.79E-03	4.81E-01
ENSG00000143499	56950	SMYD2		含有SET和MYND结构域的2	351	0.28	6.71E-03	4.81E-01
ENSG00000206579	114786	XKR4		XK, 凯尔血型复合体亚基相关家族, 成员4	352	0.91	6.72E-03	4.81E-01
ENSG00000228169	NA	PPIAP19		肽基脯氨酰异构酶A (杂环蛋白A)	353	0.74	6.73E-03	4.81E-01
ENSG00000237807	NA	RP11-400K9.4			354	0.90	6.83E-03	4.81E-01
ENSG00000243725	7268	TTC4		三角形四肽重复序列结构域4	355	0.58	6.77E-03	4.81E-01
ENSG00000251333	NA	RTN3P1		浆膜蛋白3假基因1	356	-0.37	6.86E-03	4.81E-01
ENSG00000255737	100130776	AGAP2-AS1		AGAP2反义RNA 1	357	-0.84	6.77E-03	4.81E-01
ENSG00000272777	NA	RP11-571L19.8			358	-0.55	6.78E-03	4.81E-01
ENSG00000177469	284119	PTRF		聚合酶I和转录物释放因子	359	0.52	6.91E-03	4.81E-01

图7 (续)



ENSG00000224389	721	C4B	补体成分4B (Chido血液)	360	0.70	6.92E-03	4.81E-01
ENSG00000231393	NA	RP11-331F9.3		361	0.64	6.89E-03	4.81E-01
ENSG000000005801	7748	ZNF195	锌指蛋白195	362	0.24	7.00E-03	4.84E-01
ENSG000000181038	124512	METTL23	甲基转移酶样23	363	-0.23	7.01E-03	4.84E-01
ENSG000000244952	102724126	RP11-1000B6.5		364	0.72	7.02E-03	4.84E-01
ENSG000000144834	29114	TAGLN3	转凝蛋白3	365	0.90	7.08E-03	4.87E-01
ENSG000000007171	4843	NOS2	一氧化氮合酶2, 可诱导的	366	0.90	7.29E-03	4.88E-01
ENSG000000067715	6857	SYT1	突触结合蛋白I	367	0.87	7.32E-03	4.88E-01
ENSG000000124493	2914	GRM4	谷氨酸受体, 代谢型4	368	0.88	7.34E-03	4.88E-01
ENSG000000126500	23769	FLRT1	富含亮氨酸的纤连蛋白跨膜	369	-0.89	7.27E-03	4.88E-01
ENSG000000135127	92558	CCDC64	含有卷曲螺旋结构域的64	370	0.51	7.35E-03	4.88E-01
ENSG000000136935	2800	GOLGA1	高尔基体蛋白A1	371	-0.28	7.51E-03	4.88E-01
ENSG000000148832	196743	PAOX	多胺氧化酶(exo-N4-氨基)	372	0.49	7.47E-03	4.88E-01
ENSG000000161544	114757	CYGB	细胞珠蛋白	373	-0.88	7.40E-03	4.88E-01
ENSG000000162545	55450	CAMK2N1	钙/钙调蛋白依赖的蛋白激酶	374	0.70	7.28E-03	4.88E-01
ENSG000000174197	23269	MGA	MGA, MAX二聚蛋白	375	0.17	7.45E-03	4.88E-01
ENSG000000196741	NA	CXorf24	染色体X开放阅读框24	376	0.49	7.33E-03	4.88E-01
ENSG000000237298	100506866	TTN-AS1	TTN反义RNA 1	377	0.57	7.27E-03	4.88E-01
ENSG000000239581	NA	NA	NA	378	0.83	7.44E-03	4.88E-01
ENSG000000246225	NA	RP11-17A1.3		379	0.72	7.32E-03	4.88E-01
ENSG000000254154	NA	RP4-798P15.3		380	0.82	7.30E-03	4.88E-01
ENSG000000254907	NA	RP11-484D2.2		381	0.67	7.49E-03	4.88E-01
ENSG000000256916	NA	RP11-817J15.2		382	0.89	7.30E-03	4.88E-01
ENSG000000261512	NA	RP11-46D6.1		383	0.50	7.38E-03	4.88E-01
ENSG000000262061	100506388	RP11-1260E13.4		384	0.83	7.44E-03	4.88E-01
ENSG000000264458	NA	RP11-220C2.1		385	0.68	7.28E-03	4.88E-01
ENSG000000269421	NA	ZNF92P3	锌指蛋白92假基因3	386	0.58	7.50E-03	4.88E-01
ENSG000000176920	2524	FUT2	岩藻糖基转移酶2 (分泌状态)	387	0.81	7.54E-03	4.90E-01
ENSG000000164074	80167	C4orf29	染色体4开放阅读框29	388	-0.25	7.63E-03	4.92E-01
ENSG000000167754	25818	KLK5	激肽释放酶相关肽酶5	389	-0.85	7.60E-03	4.92E-01

图7 (续)



ENSG00000178607	2081	ERN1	内质网向核的信号传导1	390	-0.35	7.64E-03	4.92E-01
ENSG00000130164	3949	LDLR	低密度脂蛋白受体	391	0.60	7.78E-03	4.93E-01
ENSG00000142207	9875	URB1	URB1核糖体生物发生1同源物(S)	392	0.28	7.76E-03	4.93E-01
ENSG00000150938	51232	CRIM1	富含半胱氨酸的跨膜BMP调节因子1	393	0.45	7.69E-03	4.93E-01
ENSG00000163082	130367	SGPP2	鞘氨醇-1-磷酸酶2	394	0.63	7.71E-03	4.93E-01
ENSG00000183401	126075	CCDC159	含有卷曲螺旋结构域的159	395	-0.32	7.76E-03	4.93E-01
ENSG00000187193	4501	MT1X	金属硫蛋白1X	396	0.66	7.76E-03	4.93E-01
ENSG00000263080	NA	RP11-485G7.5		397	0.85	7.83E-03	4.95E-01
ENSG00000131480	314	AOC2	胺氧化酶, 含有铜的2 (视网膜-速激肽受体1)	398	-0.83	7.89E-03	4.98E-01
ENSG00000115353	6869	TACR1	速激肽受体1	399	0.80	8.04E-03	4.99E-01
ENSG00000131148	10328	EMC8	ER膜蛋白复合体亚基8	400	-0.23	7.96E-03	4.99E-01
ENSG00000158008	2134	EXTL1	exostosin样糖基转移酶1	401	0.80	8.04E-03	4.99E-01
ENSG00000233483	NA	CTD-2020K17.4		402	0.70	7.99E-03	4.99E-01
ENSG00000244630	NA	RP11-241J12.1		403	0.82	7.97E-03	4.99E-01
ENSG00000258128	400058	MKRN9P	makorin环指蛋白9, 假基因	404	0.86	7.98E-03	4.99E-01
ENSG00000270109	NA	LINC01240	长基因间非蛋白质编码RNA 1240	405	0.62	8.01E-03	4.99E-01
ENSG00000157884	130106	CTB4	钙和整联蛋白结合家族成员4	406	-0.89	8.13E-03	5.01E-01
ENSG00000229955	NA	RP1-506.4		407	0.72	8.14E-03	5.01E-01
ENSG00000265396	100422824	MIR3128	微RNA 3128	408	0.74	8.13E-03	5.01E-01
ENSG000000000938	2268	FGR	FGR原癌基因, Src家族酪氨酸激酶, 成员5	409	-0.72	8.22E-03	5.04E-01
ENSG00000187116	353514	LILRA5	白细胞免疫球蛋白样受体, 亚家族A (具有TM结构域), 成员5	410	-0.72	8.27E-03	5.04E-01
ENSG00000215458	284837	AP001053.11		411	-0.66	8.28E-03	5.04E-01
ENSG00000228028	NA	AC069257.8		412	-0.83	8.24E-03	5.04E-01
ENSG00000249035	101927096	CTB-113P19.1	未表征的LOC101927096	413	0.57	8.22E-03	5.04E-01
ENSG00000126778	6495	SIX1	SIX同源框1	414	0.88	8.31E-03	5.04E-01
ENSG00000113525	3567	IL5	白介素5	415	0.85	8.35E-03	5.06E-01
ENSG00000197879	4641	MYO1C	肌球蛋白IC	416	0.36	8.49E-03	5.10E-01
ENSG00000244236	NA	RN7SL604P	RNA, 7SL, 细胞质的604, 假基因	417	-0.88	8.47E-03	5.10E-01
ENSG00000261485	100288730	PAN3-AS1	PAN3反义RNA 1	418	0.50	8.47E-03	5.10E-01
ENSG00000073282	8626	TP63	肿瘤蛋白p63	419	0.83	8.74E-03	5.13E-01

图7 (续)

ENSG00000100302	23551	RASD2	RASD家族, 成员2	420	0.74	8.83E-03	5.13E-01
ENSG00000102924	869	CBLN1	小脑肽1前体	421	0.88	8.86E-03	5.13E-01
ENSG00000108947	1949	EFNB3	肝配蛋白-B3	422	0.88	8.85E-03	5.13E-01
ENSG00000124766	6659	SOX4	SR Y (性别决定区域Y)-盒4	423	0.51	8.84E-03	5.13E-01
ENSG00000126267	1340	COX6B1	细胞色素c氧化酶亚基VIb多肽	424	-0.28	8.91E-03	5.13E-01
ENSG00000141101	28987	NOB1	NIN1/RPN12结合蛋白1同源物(S)	425	-0.23	8.92E-03	5.13E-01
ENSG00000143013	8543	LMO4	LIM结构域蛋白4	426	0.33	8.74E-03	5.13E-01
ENSG00000148795	1586	CYP17A1	细胞色素P450, 家族17, 亚家族A	427	-0.83	8.82E-03	5.13E-01
ENSG00000153790	136895	C7orf31	染色体7开放阅读框31	428	-0.32	8.92E-03	5.13E-01
ENSG00000154263	10349	ABCA10	ATP结合盒, 亚家族A (ABCI),	429	0.62	8.68E-03	5.13E-01
ENSG00000184500	5627	PROS1	蛋白S ( $\alpha$ )	430	0.63	8.70E-03	5.13E-01
ENSG00000186431	2204	FCAR	IgA的Fc片段, 受体	431	-0.76	8.82E-03	5.13E-01
ENSG00000196367	8295	TRRAP	转化/转录结构域	432	0.20	8.69E-03	5.13E-01
ENSG00000223350	NA	IGLV9-49	免疫球蛋白 $\lambda$ 变型9-49	433	0.87	8.79E-03	5.13E-01
ENSG00000249740	NA	OSMR-AS1	OSMR反义RNA 1 (头对头)	434	0.75	8.62E-03	5.13E-01
ENSG00000254850	NA	RP5-901A4.4		435	0.88	8.93E-03	5.13E-01
ENSG00000271225	NA	RP11-460N11.3		436	0.88	8.69E-03	5.13E-01
ENSG00000272761	100506621	LINC01279	长基因间非蛋白质编码RNA 1279	437	0.77	8.73E-03	5.13E-01
ENSG00000135773	10753	CAPN9	钙蛋白酶9	438	-0.82	8.96E-03	5.14E-01
ENSG00000104368	5327	PLAT	纤溶酶原激活物, 组织	439	0.81	9.02E-03	5.15E-01
ENSG00000188487	387755	INSC	inscuteable同源物(果蝇)	440	-0.88	9.01E-03	5.15E-01
ENSG00000005243	51226	COPZ2	外被体蛋白复合体, 亚基 $\zeta$ 2	441	-0.54	9.07E-03	5.16E-01
ENSG00000134086	7428	VHL	von Hippel-Lindau肿瘤抑制因子, E3泛素蛋白连接酶	442	-0.30	9.08E-03	5.16E-01
ENSG00000068394	27238	GPKOW	G膜片结构域和KOW基序	443	-0.24	9.20E-03	5.19E-01
ENSG00000213005	26255	PTTG3P	垂体肿瘤-转化3, 假基因	444	0.76	9.21E-03	5.19E-01
ENSG00000229116	101929465	RP11-20J15.3		445	0.86	9.21E-03	5.19E-01
ENSG00000233079	NA	RP11-90C4.3		446	-0.85	9.22E-03	5.19E-01
ENSG00000140511	145864	HAPLN3	透明质酸和蛋白聚糖连接蛋白3	447	0.63	9.24E-03	5.19E-01
ENSG00000178764	22882	ZHX2	锌指和同源框2	448	0.34	9.28E-03	5.21E-01
ENSG00000151422	2241	FER	fer (fps/fes相关的)酪氨酸激酶	449	0.28	9.38E-03	5.23E-01

图7 (续)

ENSG00000184106	340206	TREML3P	骨髓细胞样3上表达的触发受体, 假基因	450	-0.86	9.41E-03	5.23E-01
ENSG00000215840	NA	RP11-122G18.7		451	-0.75	9.46E-03	5.23E-01
ENSG00000243137	5672	PSG4	妊娠特异性 $\beta$ -1-糖蛋白4	452	0.84	9.45E-03	5.23E-01
ENSG00000267908	NA	ZSCAN5DP	含有锌指和SCAN结构域的5D	453	0.85	9.44E-03	5.23E-01
ENSG00000270196	NA	RP11-550A18.1		454	0.87	9.37E-03	5.23E-01
ENSG00000125149	80262	C16orf70	染色体16开放阅读框70	455	-0.50	9.48E-03	5.24E-01
ENSG00000163931	7086	TKT	转酮酶	456	-0.37	9.52E-03	5.24E-01
ENSG00000239445	100874207	ST3GAL6-AS1	ST3GAL6反义RNA 1	457	-0.83	9.54E-03	5.24E-01
ENSG00000086548	4680	CEACAM6	癌胚抗原相关的细胞粘附分子6 (非特异性交联)	458	0.85	9.60E-03	5.26E-01
ENSG00000153347	153643	FAM81B	具有序列相似性的家族81, 成员B	459	-0.67	9.60E-03	5.26E-01
ENSG00000221986	343263	MYBPHL	肌球蛋白结合蛋白H样	460	0.85	9.65E-03	5.27E-01
ENSG00000262999	101927131	CTD-3088G3.6		461	-0.86	9.67E-03	5.27E-01
ENSG00000145107	116211	TM4SF19	跨膜4 L六家族成员19	462	-0.87	9.77E-03	5.29E-01
ENSG00000243926	NA	TIPARP-AS1	TIPARP反义RNA 1	463	0.64	9.77E-03	5.29E-01
ENSG00000247095	100506211	MIR210HG	MIR210宿主基因(非蛋白质编码)	464	0.86	9.76E-03	5.29E-01
ENSG00000149575	6327	SCN2B	钠通道, 电压门控, II型, $\beta$	465	0.76	9.82E-03	5.29E-01
ENSG00000254799	NA	SLC25A47P1	溶质运载蛋白家族25, 成员47	466	0.84	9.81E-03	5.29E-01
ENSG00000108296	NA	NA	NA	467	-0.22	9.86E-03	5.30E-01
ENSG00000196704	51321	AMZ2	archaelysin家族金属肽酶2	468	-0.27	9.89E-03	5.31E-01
ENSG00000062524	4058	LTK	白细胞受体酪氨酸激酶	469	-0.87	9.93E-03	5.32E-01
ENSG00000253409	NA	TRBV7-4	T细胞受体 $\beta$ 型7-4	470	0.86	9.97E-03	5.33E-01
ENSG00000261222	101928343	CTD-2006K23.1		471	-0.86	1.00E-02	5.33E-01
ENSG00000173153	2101	ESRR	雌激素相关受体 $\alpha$	472	-0.39	1.00E-02	5.34E-01
ENSG00000256625	NA	RP11-582E3.4		473	0.65	1.01E-02	5.35E-01
ENSG00000169758	123591	C15orf27	染色体15开放阅读框27	474	-0.74	1.01E-02	5.36E-01
ENSG00000213516	494115	RBMXL1	RNA结合基序蛋白, X-连锁样1	475	0.22	1.02E-02	5.36E-01
ENSG00000231509	NA	RP11-574F11.3		476	0.75	1.02E-02	5.36E-01
ENSG00000100362	5816	PVALB	小白蛋白	477	-0.86	1.03E-02	5.39E-01
ENSG00000185483	4919	ROR1	受体酪氨酸激酶样孤独受体	478	0.75	1.03E-02	5.39E-01
ENSG00000212829	NA	RPS26P3	核糖体蛋白S26假基因3	479	0.65	1.03E-02	5.39E-01

图7 (续)

ENSG00000130876	56301	SLC7A10	溶质运载蛋白家族7(中性氨基酸转运蛋白轻链, asc系统), 成员	480	0.70	1.04E-02	5.43E-01
ENSG00000261505	NA	LA16c-358B7.3		481	0.60	1.05E-02	5.48E-01
ENSG00000157445	55799	CACNA2D3	钙通道, 电压依赖的, $\alpha 2/\delta$ 亚基3	482	-0.72	1.06E-02	5.51E-01
ENSG00000184226	5101	PCDH9	原钙粘着蛋白9	483	0.66	1.06E-02	5.51E-01
ENSG00000267375	NA	CTB-186G2.4		484	0.79	1.06E-02	5.52E-01
ENSG00000164379	94234	FOXQ1	叉头框Q1	485	0.70	1.07E-02	5.53E-01
ENSG00000187008	NA	NA	NA	486	0.54	1.07E-02	5.53E-01
ENSG00000245571	101927204	AP001258.4		487	0.43	1.07E-02	5.53E-01
ENSG00000187534	NA	PRR13P5	富含脯氨酸的13假基因5	488	-0.79	1.08E-02	5.55E-01
ENSG00000095059	1725	DHPS	赖氨酸合酶	489	-0.24	1.10E-02	5.61E-01
ENSG00000235605	NA	RP5-827C21.1		490	-0.45	1.09E-02	5.61E-01
ENSG00000213799	91664	ZNF845	锌指蛋白845	491	0.27	1.10E-02	5.62E-01
ENSG00000066136	4802	NFYC	核转录因子Y, $\gamma$	492	-0.20	1.13E-02	5.63E-01
ENSG00000109625	8532	CPZ	羧肽酶Z	493	0.83	1.13E-02	5.63E-01
ENSG00000115221	3694	ITGB6	整联蛋白, $\beta 6$	494	0.71	1.13E-02	5.63E-01
ENSG00000141738	2886	GRB7	生长因子受体结合蛋白7	495	0.60	1.13E-02	5.63E-01
ENSG00000144589	114790	STK11IP	丝氨酸/苏氨酸激酶11相互作用蛋白	496	-0.33	1.13E-02	5.63E-01
ENSG00000151790	6999	TDO2	色氨酸2,3-双加氧酶	497	-0.85	1.12E-02	5.63E-01
ENSG00000166016	25841	ABTB2	锚蛋白重复序列和BTB (POZ) 结构域	498	0.49	1.11E-02	5.63E-01
ENSG00000170866	NA	NA	NA	499	-0.74	1.12E-02	5.63E-01
ENSG00000180855	10224	ZNF443	锌指蛋白443	500	0.48	1.13E-02	5.63E-01

图7 (续)

ENTREZID	符号	基因名称	CV倍数 频率
10142	AKAP9	激酶(PRKA)锚定蛋白9	10
10251	SPRY3	sprouty同源物3 (果蝇)	10
10251	SPRY3	sprouty同源物3 (果蝇)	10
10645	CAMKK2	钙/钙调蛋白依赖的蛋白激酶	10
1277	COL1A1	胶原蛋白, I型, $\alpha 1$	10
128486	FITM2	诱导脂肪存储的跨膜蛋白2	10
1345	COX6C	细胞色素c氧化酶亚基VIc	10
147645	VSIG10L	含有V-set和免疫球蛋白结构域的10样	10
1537	CYC1	细胞色素c-1	10
221656	KDM1B	赖氨酸(K)特异的脱甲基酶1B	10
225689	MAPK15	丝裂原活化蛋白激酶15	10
22901	ARSG	芳香硫酸酯酶G	10
22976	PAXIP1	PAX相互作用(具有转录活化结构域)蛋白1	10
23002	DAAM1	形态发生的散乱蛋白相关激活物1	10
23080	AVL9	AVL9同源物(酿酒酵母)	10
29958	DMGDH	二甲基甘氨酸脱氢酶	10
3117	HLA-DQA1	主要组织相容性复合体, II类, DQ $\alpha$ 1	10
3119	HLA-DQB1	主要组织相容性复合体, II类, DQ $\beta$ 1	10
3122	HLA-DRA	主要组织相容性复合体, II类, DR $\alpha$	10
3127	HLA-DRB5	主要组织相容性复合体, II类, DR $\beta$ 5	10
3136	HLA-H	主要组织相容性复合体, I类, H(假基因)	10
3659	IRF1	干扰素调节因子1	10
4245	MGAT1	甘露糖基( $\alpha$ -1,3-)-糖蛋白 $\beta$ -1,2-N-乙酰葡萄糖胺基转移酶	10
5023	P2RX1	嘌呤能受体P2X, 配体门控离子通道, 1	10
5341	PLEK	普列克底物蛋白	10
54520	CCDC93	含有卷曲螺旋结构域的93	10
54776	PPP1R12C	蛋白磷酸酶1, 调节亚基12C	10
54946	SLC41A3	溶质运载蛋白家族41, 成员3	10
56339	METTL3	甲基转移酶样3	10

图8

57805	CCAR2	细胞周期和细胞凋亡调节因子2	10
5791	PTPRE	蛋白酪氨酸磷酸酶, 受体类型, E	10
6345	SRL	肌钙腔蛋白	10
64924	SLC30A5	溶质运载蛋白家族30 (锌转运蛋白), 成员5	10
652	BMP4	骨形态发生蛋白4	10
7692	ZNF133	锌指蛋白133	10
79664	ICE2	小延蛋白复合体ELL亚基2的相互作用物	10
79877	DCAKD	含有脱磷酸辅酶A激酶结构域	10
81542	TMX1	硫氧还蛋白相关的跨膜蛋白1	10
8742	TNFSF12	肿瘤坏死因子(配体)超家族, 成员12	10
8864	PER2	周期昼夜节律钟2	10
8888	MCM3AP	微型染色体维护复合体成分3相关蛋白	10
100462981	MTRNR2L2	MT-RNR2样2	9
10568	SLC34A2	溶质运载蛋白家族34 (II型钠/磷酸盐协同转运蛋白), 成员2	9
10626	TRIM16	含有三重基序的16	9
11276	SYNRG	synergyn, $\gamma$	9
126017	ZNF813	锌指蛋白813	9
1352	COX10	细胞色素c氧化酶装配同源物10 (酵母)	9
140685	ZBTB46	含有锌指和BTB结构域的46	9
163081	ZNF567	锌指蛋白567	9
2153	F5	凝血因子V (前加速素, 不稳定的因子)	9
23065	EMC1	ER膜蛋白复合体亚基1	9
23157	SEPT6	隔蛋白6	9
23164	MPRIP	肌球蛋白磷酸酶Rho相互作用蛋白	9
23379	ICE1	小延蛋白复合体ELL亚基1的相互作用物	9
26058	GIGYF2	GRB10相互作用GYF蛋白2	9
3728	JUP	连接斑珠蛋白	9
4318	MMP9	基质金属肽酶9 (明胶酶B, 92kDa明胶酶, 92kDa IV型胶原酶)	9
5518	PPP2R1A	蛋白磷酸酶2, 调节亚基A, $\alpha$	9
55200	PLEKHG6	含有普列克底物蛋白同源结构域的, 家族G (具有RhoGef结构域)成员6	9

图8 (续)

58191	CXCL16	趋化因子(C-X-C基序)配体16	9
629	CFB	补体因子B	9
633	BGN	双糖链蛋白聚糖	9
6990	DYNLT3	动力蛋白, 轻链, Tctex- 3型	9
7846	TUBA1A	微管蛋白, $\alpha$ 1a	9
79968	WDR76	WD重复序列结构域76	9
8544	PIR	pirin (铁结合核内蛋白质)	9
871	SERPINH1	丝氨酸蛋白酶抑制蛋白肽酶抑制剂, 进化枝H (热休克蛋白47), 成员1, (胶原结合蛋白1)	9
8773	SNAP23	突触体相关蛋白, 23kDa	9
914	CD2	CD2分子	9
91662	NLRP12	NLR家族, 含热蛋白域12	9
100130015	URHP	尿酸盐(羟基异构的) 水解酶, 假基因	8
10053	AP1M2	衔接子相关蛋白复合体1, $\mu$ 2亚基	8
10376	TUBA1B	微管蛋白, $\alpha$ 1b	8
10422	UBAC1	含有UBA结构域的1	8
1062	CENPE	着丝粒蛋白E, 312kDa	8
11022	TDRKH	含有tudor和KH域	8
11138	TBC1D8	TBC1结构域家族, 成员8 (具有GRAM结构域)	8
115361	GBP4	鸟苷酸结合蛋白4	8
1173	AP2M1	衔接子相关蛋白复合体2, $\mu$ 1亚基	8
150468	CKAP2L	细胞骨架相关蛋白2样	8
155061	ZNF746	锌指蛋白746	8
1595	CYP51A1	细胞色素P450, 家族51, 亚家族A, 多肽1	8
1677	DFFB	DNA断裂因子, 40kDa, $\beta$ 多肽(胱天蛋白酶活化的DNA酶)	8
199777	ZNF626	锌指蛋白626	8
221476	PI16	肽酶抑制物16	8
222235	FBXL13	F-盒和富含亮氨酸的重复序列蛋白13	8
222696	ZSCAN23	含有锌指和SCAN结构域的23	8
22822	PHLDA1	普列克底物蛋白同源样结构域, 家族A, 成员1	8
22900	CARD8	胱天蛋白酶招募结构域家族, 成员8	8

图8(续)

22948	CCT5	含有伴侣蛋白的TCP1, 亚基5 (ε)	8
2313	FLI1	Fli-1原癌基因, ETS转录因子	8
23426	GRIP1	谷氨酸受体相互作用蛋白1	8
25	ABL1	ABL原癌基因1, 非受体酪氨酸激酶	8
2532	ACKR1	非典型趋化因子受体1 (Duffy 血型)	8
25821	MT01	线粒体tRNA翻译优化	8
27102	EIF2AK1	真核生物翻译起始因子2α激酶1	8
27295	PDLIM3	PDZ和LIM结构域3	8
2842	GPR19	G蛋白偶联受体19	8
29882	ANAPC2	后期促进复合体亚基2	8
3337	DNAJB1	DnaJ (Hsp40)同源物, 亚家族B, 成员1	8
3486	IGFBP3	胰岛素样生长因子结合蛋白3	8
3712	IVD	异戊酰辅酶A脱氢酶	8
3821	KLRC1	杀伤细胞凝集素样受体超家族C, 成员1	8
4882	NPR2	利尿钠肽受体2	8
51084	CRYL1	晶体蛋白, λ1	8
51295	ECSIT	ECSIT信号传导整合体	8
55366	LGR4	含有富含亮氨酸的重复序列的G蛋白偶联受体4	8
56479	KCNQ5	钾电压门控通道, KQT样亚家族, 成员5	8
56906	THAP10	含有THAP结构域的10	8
56942	CMC2	含有C-x(9)-C基序的2	8
56999	ADAMTS9	具有血小板反应蛋白1型基序的ADAM金属肽酶, 9	8
57188	ADAMTSL3	ADAMTS样3	8
57613	KIAA1467	KIAA1467	8
58487	CREBZF	CREB/ATF bZIP转录因子	8
6338	SCNN1B	钠通道, 非电压门控1, β亚基	8
63894	VIPAS39	VPS33B相互作用蛋白, 顶端-基底外侧极性调节因子, spe-39同源物	8
65082	VPS33A	液泡蛋白分选33同源物A (酿酒酵母)	8
7053	TGM3	转谷氨酰胺酶3	8
7058	THBS2	血小板反应蛋白2	8

图8 (续)



7881	KCNAB1	钾电压门控通道, 振荡器相关的亚家族, $\beta$ 成员1	8
79143	MBOAT7	含有膜结合的O-酰基转移酶结构域的7	8
79152	FA2H	脂肪酸2-羟化酶	8
80298	MTERF2	线粒体转录终止因子2	8
84034	EMILIN2	弹性蛋白微纤维界面因子2	8
84951	TNS4	张力蛋白4	8
8578	SCARF1	清道夫受体F类, 成员1	8
8642	DCHS1	dachsous钙粘着蛋白相关1	8
8738	CRADD	具有死亡结构域的含有CASP2和RIPK1结构域的衔接子	8
8874	ARHGEF7	Rho鸟嘌呤核苷酸交换因子(GEF) 7	8
89932	PAPLN	papilin, 蛋白聚糖样硫酸化糖蛋白	8
9139	CBFA2T2	核心结合因子, runt结构域, $\alpha$ 亚基2; 位置转移至, 2	8
9194	SLC16A7	溶质运载蛋白家族16(单羧酸转运蛋白), 成员7	8
928	CD9	CD9分子	8
9957	HS3ST1	硫酸乙酰肝素(葡糖胺) 3-O-磺基转移酶1	8
100101467	ZSCAN30	含有锌指和SCAN结构域的30	7
10101	NUBP2	核苷酸结合蛋白2	7
10766	TOB2	ERBB2的转导物, 2	7
151194	METTL21A	甲基转移酶样21A	7
1975	EIF4B	真核生物翻译起始因子4B	7
1994	ELAVL1	ELAV样RNA结合蛋白1	7
221785	ZSCAN25	含有锌指和SCAN结构域的25	7
3305	HSPA1L	热休克70kDa蛋白1样	7
359821	MRPL42P5	线粒体核糖体蛋白L42假基因5	7
3880	KRT19	角蛋白19	7
441155	LOC441155	含有锌指CCCH型结构域样	7
54472	TOLLIP	toll相互作用蛋白	7
55005	RMND1	减数分裂核分裂1同源物(酿酒酵母)所需的	7
55827	DCAF6	DDB1和CUL4相关因子6	7
56992	KIF15	驱动蛋白家族成员15	7

图8(续)

5918	RARRES1	视黄酸受体响应物(他扎罗汀诱导的) 1	7
64236	PDLIM2	PDZ和LIM结构域2 (mystique)	7
79022	TMEM106C	跨膜蛋白106C	7
79168	LILRA6	白细胞免疫球蛋白样受体, 亚家族A (具有TM结构域), 成员6	7
80000	GREB1L	乳腺癌样中雌激素的生长调节	7
80219	COQ10B	辅酶Q10同源物B (酿酒酵母)	7
80279	CDK5RAP3	CDK5调节亚基相关蛋白	7
84131	CEP78	中心体蛋白78kDa	7
84984	CEP19	中心体蛋白19kDa	7
8925	HERC1	含有HECT和RLD结构域的E3泛素蛋白连接酶家族成员1	7
9559	VPS26A	液泡蛋白分选26同源物A (粟酒裂殖酵母)	7
994	CDC25B	细胞分裂周期25B	7
100506736	SLFN12L	schlafen家族成员12样	6
10260	DENND4A	含有DENN/MADD结构域的4A	6
10985	GCN1L1	氨基酸合成的GCN1一般控制1样1 (酵母)	6
165140	OXER1	氧桥十二烷(OXE)受体1	6
23294	ANKS1A	含有锚蛋白重复序列和不育 $\alpha$ 基序结构域的1A	6
3123	HLA-DRB1	主要组织相容性复合体, II类, DR $\beta$ 1	6
338692	ANKRD13D	锚蛋白重复序列结构域13家族, 成员D	6
3516	RBPJ	免疫球蛋白 $\kappa$ J区域的重组信号结合蛋白	6
3673	ITGA2	整联蛋白, $\alpha$ 2 (CD49B, VLA-2受体的 $\alpha$ 2亚基)	6
54434	SSH1	弹弓蛋白磷酸酶1	6
546	ATRX	$\alpha$ 地中海贫血/X-连锁的精神发育迟滞综合征	6
57335	ZNF286A	锌指蛋白286A	6
57511	COG6	寡聚的高尔基复合体6的组分	6
64848	YTHDC2	含有YTH结构域的2	6
653361	NCF1	嗜中性粒细胞胞质因子1	6
7572	ZNF24	锌指蛋白24	6
81027	TUBB1	微管蛋白, $\beta$ 1 VI类	6
84656	GLYR1	乙醛酸还原酶1同源物(拟南芥)	6

图8 (续)

84824	FCRLA	Fc受体样A	6
84868	HAVCR2	甲型肝炎病毒细胞受体2	6
9648	GCC2	含有GRIP和卷曲螺旋结构域的2	6
10482	NXF1	核RNA输出因子1	5
1106	CHD2	染色质域解旋酶DNA结合蛋白2	5
27237	ARHGEF16	Rho鸟嘌呤核苷酸交换因子(GEF) 16	5
3716	JAK1	詹纳斯激酶1	5
3831	KLC1	驱动蛋白轻链1	5
3965	LGALS9	凝集素, 半乳糖苷结合, 可溶的, 9	5
4052	LTBP1	潜伏转化生长因子 $\beta$ 结合蛋白1	5
440145	MZT1	有丝分裂纺锤体组织蛋白1	5
56107	PCDHGA9	原钙粘着蛋白 $\gamma$ 亚家族A, 9	5
6490	PMEL	前黑素体蛋白	5
80781	COL18A1	胶原蛋白, XVIII型, $\alpha$ 1	5
84334	APOPT1	致凋亡的1, 线粒体的	5
100532731	COMMD3-BMI1	COMMD3-BMI1通读	4
100532746	PPT2-EGFL8	PPT2-EGFL8通读(NMD候选物)	4
10521	DDX17	DEAD (Asp-Glu-Ala-Asp)盒解旋酶17	4
10569	SLU7	SLU7剪接因子同源物(酿酒酵母)	4
1066	CES1	羧酸酯酶1	4
10767	HBS1L	HBS1样翻译GTP酶	4
115677	NOSTRIN	一氧化氮合酶运输	4
149473	CCDC24	含有卷曲螺旋结构域的24	4
1606	DGKA	二酰基甘油激酶, $\alpha$ 80kDa	4
23031	MAST3	微管相关丝氨酸/苏氨酸激酶3	4
25824	PRDX5	过氧化物氧化蛋白5	4
26289	AK5	腺苷酸激酶5	4
27433	TOR2A	torsin家族2, 成员A	4
2820	GPD2	甘油-3-磷酸脱氢酶2(线粒体的)	4
282679	AQP11	水通道蛋白11	4

图8(续)

3073	HEXA	氨基己糖苷酶A ( $\alpha$ 多肽)	4
3159	HMGAI	高移动性型AT-钩1	4
3371	TNC	生腱蛋白C	4
4671	NAIP	NLR家族, 细胞凋亡抑制蛋白	4
468	ATF4	活化转录因子4	4
51154	MRT04	mRNA转换4同源物(酿酒酵母)	4
54439	RBM27	RNA结合基序蛋白27	4
54471	MIEF1	线粒体延伸因子1	4
54848	ARHGEF38	Rho鸟嘌呤核苷酸交换因子(GEF) 38	4
55780	ERMARD	ER膜相关的RNA降解	4
5669	PSG1	妊娠特异性 $\beta$ -1-糖蛋白1	4
6241	RRM2	核糖核苷酸还原酶M2	4
6423	SFRP2	分泌性卷曲相关蛋白2	4
64327	LMBR1	肢体发育膜蛋白1	4
64423	INF2	反向重复的成蛋白, 含有FH2和WH2结构域	4
6904	TBCD	微管蛋白折叠辅因子D	4
729830	FAM160A1	具有序列相似性的家族160, 成员A1	4
79065	ATG9A	自噬相关的9A	4
8348	HIST1H2BO	组蛋白簇1, H2bo	4
84513	PPAPDC1B	含有2型磷脂酸磷酸酶结构域的1B	4
8532	CPZ	羧肽酶Z	4
9099	USP2	泛素特异性肽酶2	4
9374	PPT2	棕榈酰-蛋白质硫酯酶2	4
100529209	RNASEK-C17orf49	RNASEK-C17orf49通读	3
100861437	NARR	九个氨基酸残基重复序列	3
10420	TESK2	睾丸特异性激酶2	3
10903	MTMR11	肌微管素相关蛋白11	3
11326	VSIG4	含有V-set和免疫球蛋白结构域的4	3
114786	XKR4	XK, 凯尔血型复合体亚基相关家族, 成员4	3
116534	MRGPRE	MAS相关GPR, 成员E	3

图8(续)

122773	KLHDC1	含有kelch结构域的1	3
1286	COL4A4	胶原蛋白, IV型, $\alpha$ 4	3
1491	CTH	胱硫醚 $\gamma$ 裂解酶	3
151648	SGOL1	shugoshin样1 (粟酒裂殖酵母)	3
157769	FAM91A1	具有序列相似性的家族91, 成员A1	3
162073	ITPRIPL2	肌醇1,4,5-三磷酸受体相互作用蛋白样2	3
166378	SPATA5	精子发生相关的5	3
1730	DIAPH2	透明相关的成蛋白2	3
197135	PATL2	与拓扑异构酶II相关的蛋白质同源物2 (酵母)	3
203523	ZNF449	锌指蛋白449	3
2051	EPHB6	EPH受体B6	3
22862	FNDC3A	含有纤连蛋白类型III结构域的3A	3
23158	TBC1D9	TBC1结构域家族, 成员9 (具有GRAM结构域)	3
23264	ZC3H7B	含有锌指CCCH-类型的7B	3
3303	HSPA1A	热休克70kDa蛋白1A	3
4689	NCF4	嗜中性粒细胞胞质因子4, 40kDa	3
55331	ACER3	碱性神经酰胺酶3	3
6929	TCF3	转录因子3	3
100131439	CD300LD	CD300分子样家族成员d	2
100526740	ATP5J2-PTCD1	ATP5J2-PTCD1通读	2
10380	BPNT1	3'(2'), 5'-二磷酸核苷酸酶1	2
10687	PNMA2	副肿瘤性Ma抗原2	2
107	ADCY1	腺苷酸环化酶1 (脑)	2
113146	AHNAK2	AHNAK核蛋白2	2
116443	GRIN3A	谷氨酸受体, 离子型, N-甲基-D-天冬氨酸3A	2
1211	CLTA	网格蛋白, 轻链A	2
127731	VWA5B1	含有von Willebrand因子A结构域的5B1	2
130162	CLHC1	含有网格蛋白重链接头结构域的1	2
131870	NUDT16	nudix (连接核苷二磷酸的部分X)-类型基序16	2
140691	TRIM69	含有三重基序的69	2

图8 (续)

149175	MANEAL	甘露糖苷酶, 内- $\alpha$ 样	2
158358	KIAA2026	KIAA2026	2
1622	DBI	地西泮结合抑制剂(GABA受体调节剂, 酰基辅酶A结合蛋白)	2
1979	EIF4EBP2	真核生物翻译起始因子4E结合蛋白2	2
2033	EP300	E1A结合蛋白p300	2
206358	SLC36A1	溶质运载蛋白家族36(质子/氨基酸同向转运蛋白), 成员1	2
221150	SKA3	纺锤体和着丝粒相关复合体亚基3	2
23148	NACAD	含有NAC $\alpha$ 结构域	2
23163	GGA3	高尔基相关, 含有 $\gamma$ 衔接蛋白, ARF结合蛋白3	2
23171	GPD1L	甘油-3-磷酸脱氢酶1样	2
23237	ARC	活性调节的细胞骨架相关蛋白	2
2327	FMO2	含有黄素的单加氧酶2(非功能性的)	2
23299	BICD2	双尾D同源物2(果蝇)	2
23365	ARHGEF12	Rho鸟嘌呤核苷酸交换因子(GEF) 12	2
23412	COMMD3	含有COMM结构域的3	2
246176	GAS2L2	生长停滞特异性2样2	2
25794	FSCN2	肌成束蛋白肌动蛋白集束蛋白2, 视网膜的	2
26024	PTCD1	三角状五肽重复序列结构域1	2
26145	IRF2BP1	干扰素调节因子2结合蛋白1	2
261734	NPHP4	肾消耗病4	2
283431	GAS2L3	生长停滞特异性2样3	2
284996	RNF149	环指蛋白149	2
286207	C9orf117	染色体9开放阅读框117	2
2907	GRINA	谷氨酸受体, 离子型, N-甲基-D-天冬氨酸相关蛋白1(谷氨酸结合)	2
2949	GSTM5	谷胱甘肽S-转移酶mu5	2
29929	ALG6	ALG6, $\alpha$ -1,3-葡糖基转移酶	2
339804	C2orf74	染色体2开放阅读框74	2
340543	TCEAL5	转录延伸因子A(SII)样5	2
342897	NCCRP1	非特异性细胞毒性细胞受体蛋白1同源物(斑马鱼)	2
348094	ANKDD1A	含有锚蛋白重复序列和死亡结构域的1A	2

图8(续)

348180	CTU2	胞质硫脲酶亚基2同源物(粟酒裂殖酵母)	2
3611	ILK	整联蛋白连接的激酶	2
387032	ZKSCAN4	具有KRAB和SCAN结构域的锌指4	2
387111	LINC00222	长基因间非蛋白质编码RNA 222	2
400673	VMAC	波形蛋白型中间丝相关卷曲螺旋蛋白	2
4171	MCM2	微型染色体维护复合体成分	2
4327	MMP19	基质金属肽酶19	2
494470	RNF165	环指蛋白165	2
5087	PBX1	前B细胞白血病同源框1	2
51559	NT5DC3	含有5'-核苷酸酶结构域的3	2
54756	IL17RD	白介素17受体D	2
55197	RPRD1A	含有核前mRNA结构域的1A的调控	2
55238	SLC38A7	溶质运载蛋白家族38, 成员7	2
55326	AGPAT5	1-酰基甘油-3-磷酸O-酰基转移酶5	2
55686	MREG	melanoregulin	2
55759	WDR12	WD重复序列结构域12	2
56898	BDH2	3-羟基丁酸脱氢酶, 2型	2
5891	MOK	MOK蛋白激酶	2
64174	DPEP2	二肽酶2	2
645121	CCNI2	细胞周期蛋白I家族, 成员2	2
692312	PPAN-P2RY11	PPAN-P2RY11通读	2
7355	SLC35A2	溶质运载蛋白家族35 (UDP-半乳糖转运蛋白), 成员A2	2
7465	WEE1	WEE1 G2检查点激酶	2
7978	MTERF1	线粒体转录终止因子1	2
9517	SPTLC2	丝氨酸棕榈酰转移酶, 长链碱性亚基2	2
9780	PIEZO1	压力型机械敏感离子通道组分1	2
100529240	ZNF816-ZNF321P	ZNF816-ZNF321P通读	1
10385	BTN2A2	嗜乳脂蛋白, 亚家族2, 成员A2	1
10523	CHERP	钙稳态内质网蛋白	1
10630	PDPN	平足蛋白	1

图8(续)

11060	WWP2	含有WW结构域的E3泛素蛋白连接酶2	1
11162	NUDT6	nudix (连接核苷二磷酸的部分X)-类型基序6	1
114904	C1QTNF6	C1q和肿瘤坏死因子相关蛋白6	1
116541	MRPL54	线粒体核糖体蛋白L54	1
1281	COL3A1	胶原蛋白, III型, $\alpha 1$	1
145173	B3GALT1	$\beta 1,3$ -半乳糖基转移酶样	1
146722	CD300LF	CD300分子样家族成员f	1
154796	AMOT	血管动蛋白	1
163259	DENN2C	含有DENN/MADD结构域的2C	1
165055	CCDC138	含有卷曲螺旋结构域的138	1
165215	FAM171B	具有序列相似性的家族171, 成员B	1
2125	EVPL	外被斑蛋白	1
222487	GPR97	G蛋白偶联受体97	1
2235	FECH	亚铁螯合酶	1
23216	TBC1D1	TBC1 (tre-2/USP6, BUB2, cdc16)结构域家族, 成员1	1
23240	KIAA0922	KIAA0922	1
23657	SLC7A11	溶质运载蛋白家族7 (阴离子型氨基酸转运蛋白轻链, xc系统), 成员11	1
23683	PRKD3	蛋白激酶D3	1
25800	SLC39A6	溶质运载蛋白家族39 (锌转运蛋白), 成员6	1
26278	SACS	sacsin分子伴侣	1
283284	IGSF22	免疫球蛋白超家族, 成员22	1
283578	TMED8	含有跨膜emp24蛋白转运结构域的8	1
283848	CES4A	羧酸酯酶4A	1
284759	SIRPB2	信号调节蛋白 $\beta 2$	1
2891	GRIA2	谷氨酸受体, 离子型, AMPA 2	1
2931	GSK3A	糖原合酶激酶3 $\alpha$	1
29841	GRHL1	grainyhead样1 (果蝇)	1
29964	PRICKLE4	皮刺同源物4 (果蝇)	1
30836	DNTTIP2	脱氧核苷酸转移酶, 末端, 相互作用蛋白2	1
3094	HINT1	组氨酸三联体核苷酸结合蛋白1	1

图8 (续)



3310	HSPA6	热休克70kDa蛋白6 (HSP70B')	1
340481	ZDHHC21	锌指, 含有DHH型的21	1
3431	SP110	SP110核质体蛋白	1
3613	IMPA2	肌醇(肌)-1(或4)-单磷酸酶2	1
376267	RAB15	RAB15, 成员RAS癌基因家族	1
3816	KLK1	激肽释放酶1	1
389677	RBM12B	RNA结合基序蛋白12B	1
399669	ZNF321P	锌指蛋白321, 假基因	1
400569	MED11	介质复合体亚基11	1
4058	LTK	白细胞受体酪氨酸激酶	1
440482	ANKRD20A5P	锚蛋白重复序列结构域20家族, 成员A5, 假基因	1
441108	C5orf56	染色体5开放阅读框56	1
4668	NAGA	N-乙酰氨基半乳糖苷酶, $\alpha$ -	1
4735	SEPT2	隔蛋白2	1
4853	NOTCH2	notch 2	1
5032	P2RY11	嘌呤能受体P2Y, G蛋白偶联, 11	1
51252	FAM178B	具有序列相似性的家族178, 成员B	1
5127	CDK16	细胞周期蛋白依赖性激酶16	1
51291	GMIP	GEM相互作用蛋白	1
5433	POLR2D	聚合酶(RNA) II (DNA指导)多肽D	1
54662	TBC1D13	TBC1结构域家族, 成员13	1
54977	SLC25A38	溶质运载蛋白家族25, 成员38	1
55001	TTC22	三角形四肽重复序列结构域22	1
55030	FBXO34	F盒蛋白34	1
55039	TRMT12	tRNA甲基转移酶12同源物(酿酒酵母)	1
55314	TMEM144	跨膜蛋白144	1
55593	OTUD5	OTU去泛素化酶5	1
55723	ASF1B	抗沉默功能1B组蛋白伴侣	1
55734	ZFP64	ZFP64锌指蛋白	1
5670	PSG2	妊娠特异性 $\beta$ -1-糖蛋白2	1

图8 (续)

56832	IFNK	干扰素, $\kappa$	1
56987	BBX	bobby sox同源物(果蝇)	1
57191	VN1R1	犁鼻1受体1	1
57455	REXO1	REX1, RNA外切核酸酶1同源物(酿酒酵母)	1
57458	TMCC3	跨膜和卷曲螺旋结构域家族3	1
57553	MICAL3	微管相关单加氧酶, 含有钙调理蛋白和LIM结构域的3	1
58526	MID1IP1	MID1相互作用蛋白1	1
5887	RAD23B	RAD23同源物B(酿酒酵母)	1
59067	IL21	白介素21	1
5970	RELA	v-rel鸟网状内皮组织增殖病毒癌基因同源物A	1
60490	PPCDC	磷酸泛酰半胱氨酸脱羧酶	1
6137	RPL13	核糖体蛋白L13	1
6273	S100A2	S100钙结合蛋白A2	1
637	BID	BH3相互作用结构域死亡激动剂	1
63916	ELMO2	吞入和细胞运动性2	1
64397	ZNF106	锌指蛋白106	1
64754	SMYD3	含有SET和MYND结构域的3	1
64761	PARP12	聚(ADP-核糖)聚合酶家族, 成员12	1
684	BST2	骨髓基质细胞抗原2	1
7022	TFAP2C	转录因子AP-2 $\gamma$ (活化增强子结合蛋白2 $\gamma$ )	1
7227	TRPS1	毛-鼻-指综合征I	1
7266	DNAJC7	DnaJ (Hsp40)同源物, 亚家族C, 成员7	1
729288	ZNF286B	锌指蛋白286B	1
730005	SEC14L6	SEC14样6(酿酒酵母)	1
7328	UBE2H	泛素缀合酶E2H	1
7564	ZNF16	锌指蛋白16	1
7866	IFRD2	干扰素相关发育调节因子2	1
79066	METTL16	甲基转移酶样16	1
79070	KDELC1	含有KDEL (Lys-Asp-Glu-Leu)的1	1
79661	NEIL1	nei内切核酸酶VIII样1(大肠杆菌)	1

图8(续)

79665	DHX40	DEAH (Asp-Glu-Ala-His)盒多肽40	1
79800	CARF	钙响应转录因子	1
79937	CNTNAP3	接触蛋白相关蛋白样3	1
80210	ARMC9	含有玃猢重复序列的9	1
84216	TMEM117	跨膜蛋白117	1
84255	SLC37A3	溶质运载蛋白家族37, 成员3	1
84989	JMJD1C-AS1	JMJD1C反义RNA 1	1
8570	KHSRP	KH型剪接调节蛋白	1
8702	B4GALT4	UDP-Gal:βGlcNAc β1,4-半乳糖基转移酶, 多肽4	1
8767	RIPK2	受体相互作用丝氨酸-苏氨酸激酶2	1
8897	MTMR3	肌微管素相关蛋白3	1
9013	TAF1C	TATA盒结合蛋白(TBP)相关因子, RNA聚合酶I, C, 110kDa	1
9033	PKD2L1	多囊性肾病2样1	1
9130	FAM50A	具有序列相似性的家族50, 成员A	1
91947	ARRDC4	含有抑制蛋白结构域的4	1
9249	DHRS3	脱氢酶/还原酶(SDR家族)成员3	1
93594	TBC1D31	TBC1结构域家族, 成员31	1
954	ENTPD2	胞外核苷三磷酸二磷酸水解酶2	1
9744	ACAP1	具有卷曲螺旋的ArfGAP, 锚蛋白重复序列和PH结构域1	1
9860	LRIG2	富含亮氨酸的重复序列和免疫球蛋白样结构域2	1
9936	CD302	CD302分子	1

图8(续)

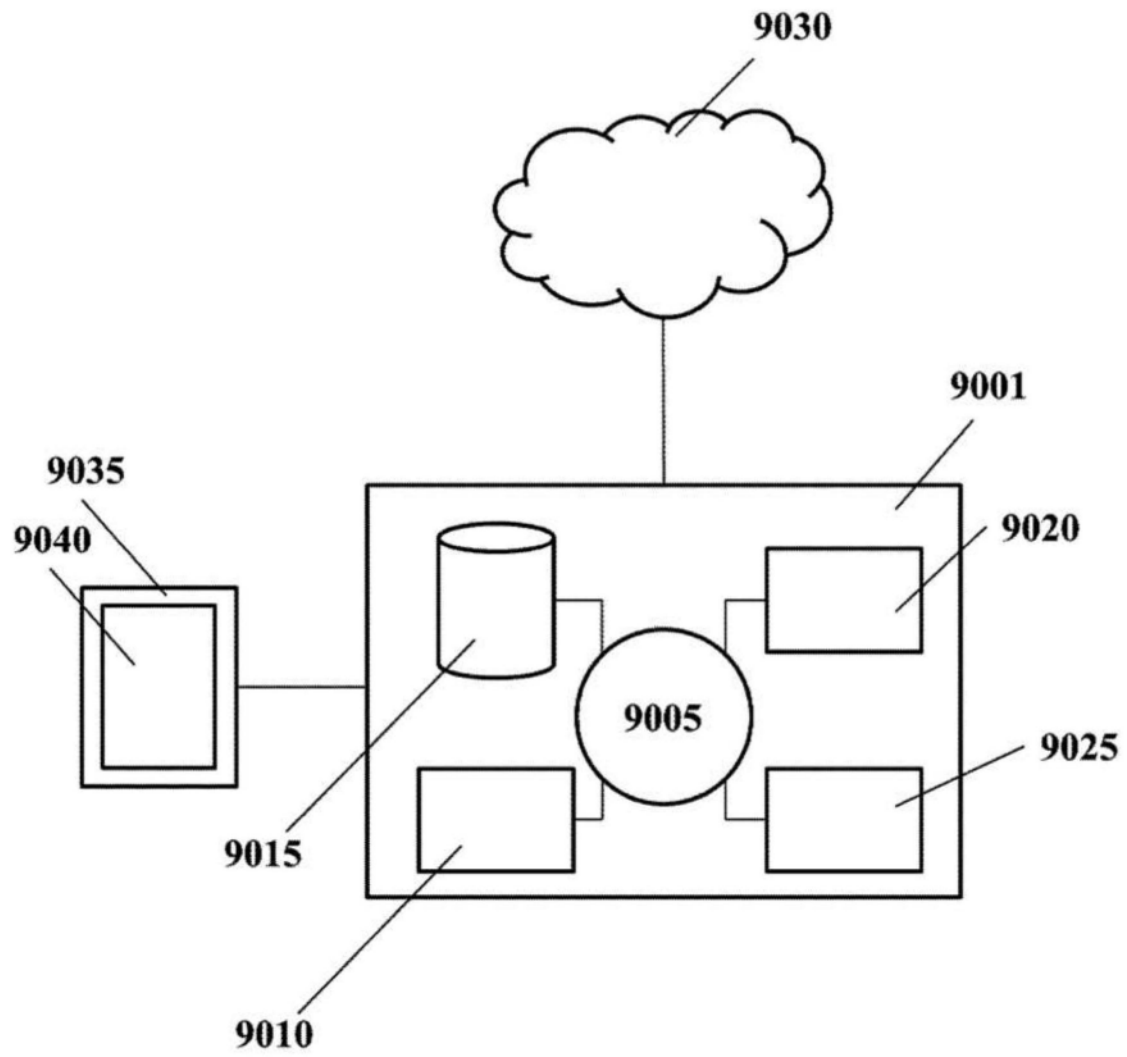


图9

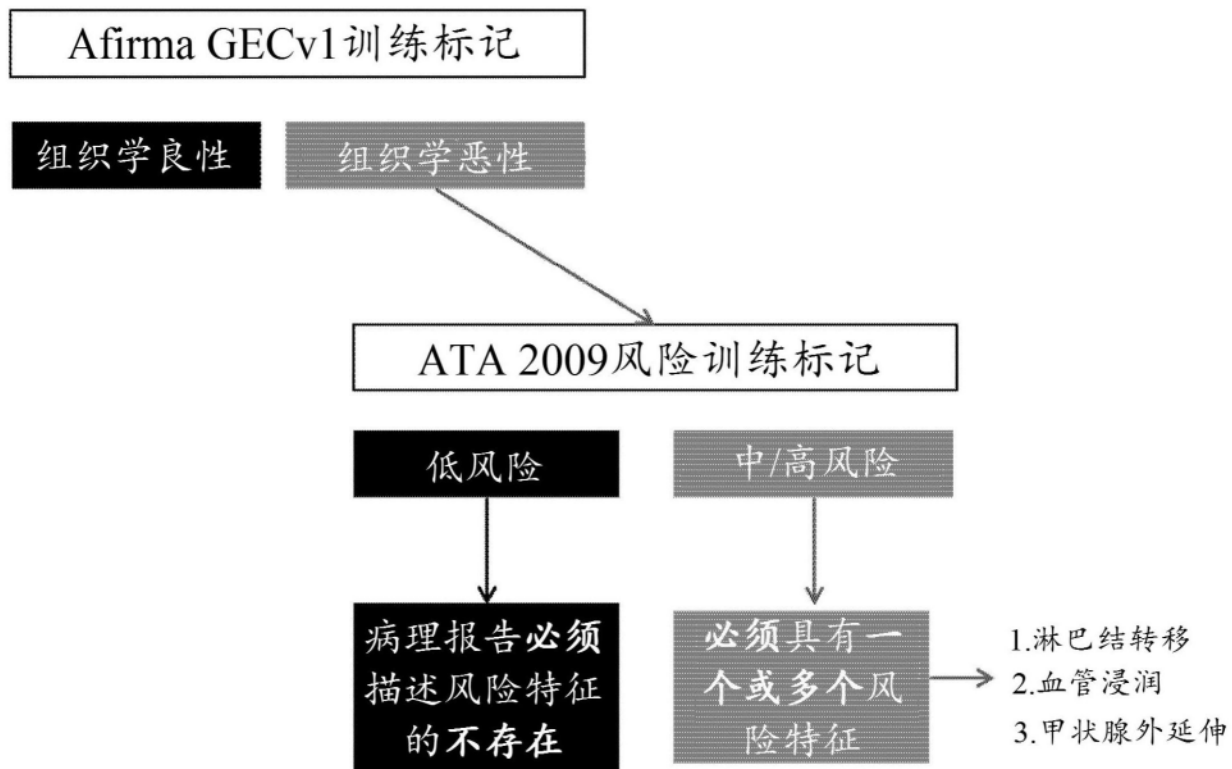


图10

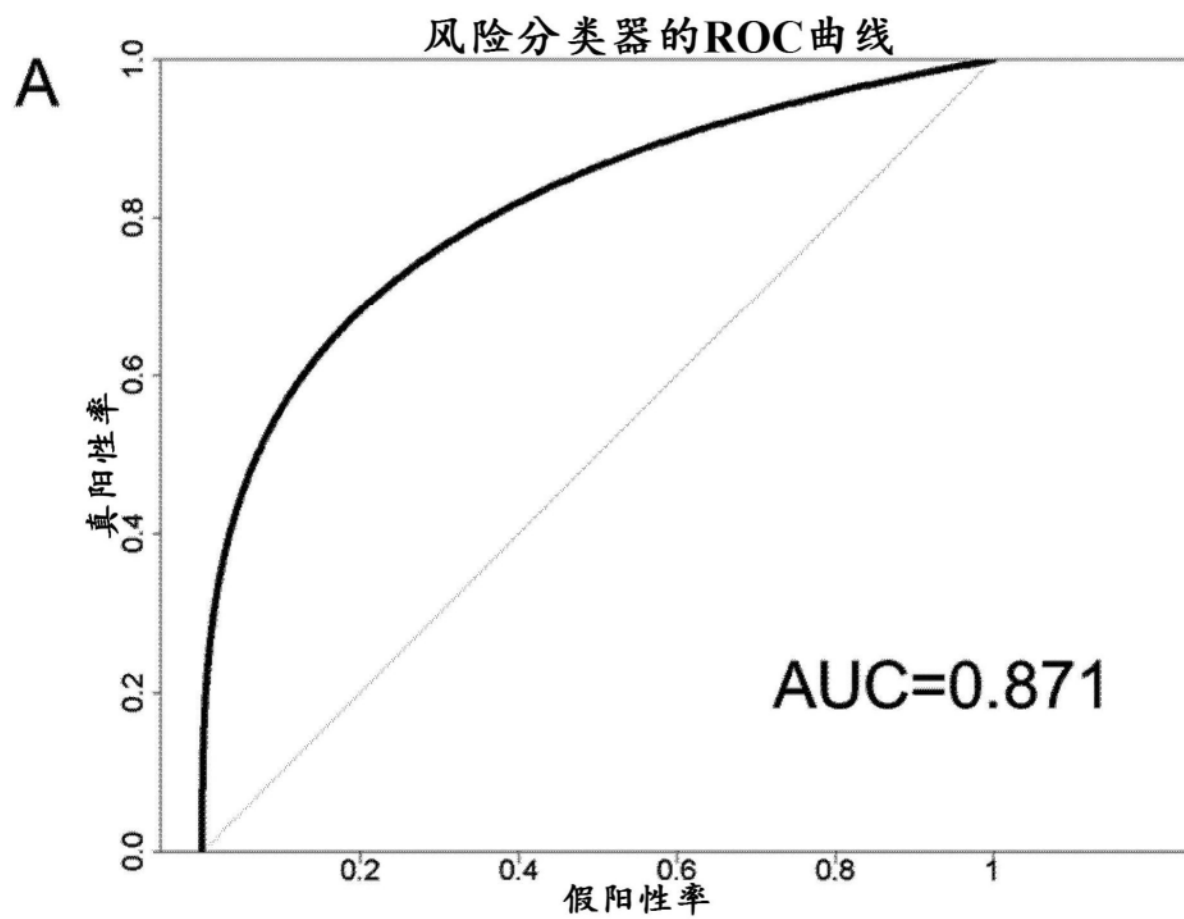


图11A

<b>B</b>		
	中/高风险 (n=50)	低风险 (n=29)
分类为 “中/高风险”	43	7
分类为 “低风险”	7	25
灵敏度= 86%		
特异性= 86%		

图11B

以10倍得到的频率	基因符号	说明
10	COL1A1	胶原蛋白, 1型, $\alpha 1$
10	FITM2	诱导脂肪储存的跨膜蛋白2
10	AASDH	氨基己二酸半醛脱氢酶
10	COX6C	细胞色素c氧化酶亚基VIc
10	COX10	细胞色素c氧化酶装配同源物10 (酵母)
10	VSIG10L	含有V-set和免疫球蛋白结构域的10样
10	MAPK15	丝裂原活化蛋白激酶15
10	PAXIP1	PAX相互作用(具有转录活化结构域)蛋白1
10	AVL9	AVL9同源物(酿酒酵母)
10	GIGYF2	GRB10相互作用GYF蛋白2
10	HLA-DQA1	主要组织相容性复合体, II类, DQ $\alpha 1$
10	HLA-DQB1	主要组织相容性复合体, II类, DQ $\beta 1$
10	HLA-DRA	主要组织相容性复合体, II类, DR $\alpha$
10	HLA-H	主要组织相容性复合体, I类, H(假基因)
10	MGAT1	甘露糖基( $\alpha$ -1,3-)-糖蛋白 $\beta$ -1,2-N-乙酰葡萄糖胺基转移酶
10	SLC41A3	溶质运载蛋白家族41, 成员3
10	PTPRE	蛋白酪氨酸磷酸酶, 受体类型, E
10	SRL	肌钙腔蛋白
10	SLC30A5	溶质运载蛋白家族30 (锌转运蛋白), 成员5
10	BMP4	骨形态发生蛋白4
10	ICE2	小延蛋白复合体ELL亚基2的交互器
10	DCAKD	含有脱磷酸辅酶A激酶结构域
10	TMX1	硫氧还蛋白相关的跨膜蛋白1
10	HAVCR2	甲型肝炎病毒细胞受体2
10	TNFSF12	肿瘤坏死因子(配体)超家族, 成员12
10	PER2	周期昼夜节律钟2
10	MCM3AP	微型染色体维护复合体成分3相关蛋白

图12



以10倍得到的频率	基因符号	说明
10	COL1A1	胶原蛋白, 1型, $\alpha 1$
10	NUP210L	核孔蛋白210kDa样
10	TMEM92	跨膜蛋白92
10	C6orf136	染色体6开放阅读框136
9	SPAG4	精子相关抗原4
9	EHF	ets同源因子
9	RAPGEF5	Rap鸟嘌呤核苷酸交换因子(GEF)5
9	COL3A1	胶原蛋白, III型, $\alpha 1$
8	GALNT15	多肽N-乙酰氨基半乳糖基转移酶15
8	PRICKLE1	皮刺同源物1 (果蝇)
8	LUM	腔蛋白聚糖
8	COL6A3	胶原蛋白, VI型, $\alpha 3$
8	ROBO1	迂回轴突导向受体同源物1(果蝇)
8	SSC5D	富含半胱氨酸的清道夫受体家族, 5结构域
8	PSORS1C1	银屑病敏感性1候选物1

图13

突变组	基因组位点	融合对
组1	9	3
组2	19	25
组3	208	25
组4	929	25
组5	3670	25

突变组中靶向的基因*						
AKT1	BRAF**	CTNNB1	EIF1AX	GNAS	HRAS**	KRAS**
NRAS**	PIK3CA	PTEN	RET	TERT***	TSHR***	TP53

\* 组2-5包括Nikiforov等人报道的这些基因和融合

\*\* 组1仅评估Beaudenon等人报道的基因和融合的亚组。

\*\*\* 仅包括在组5中

图14

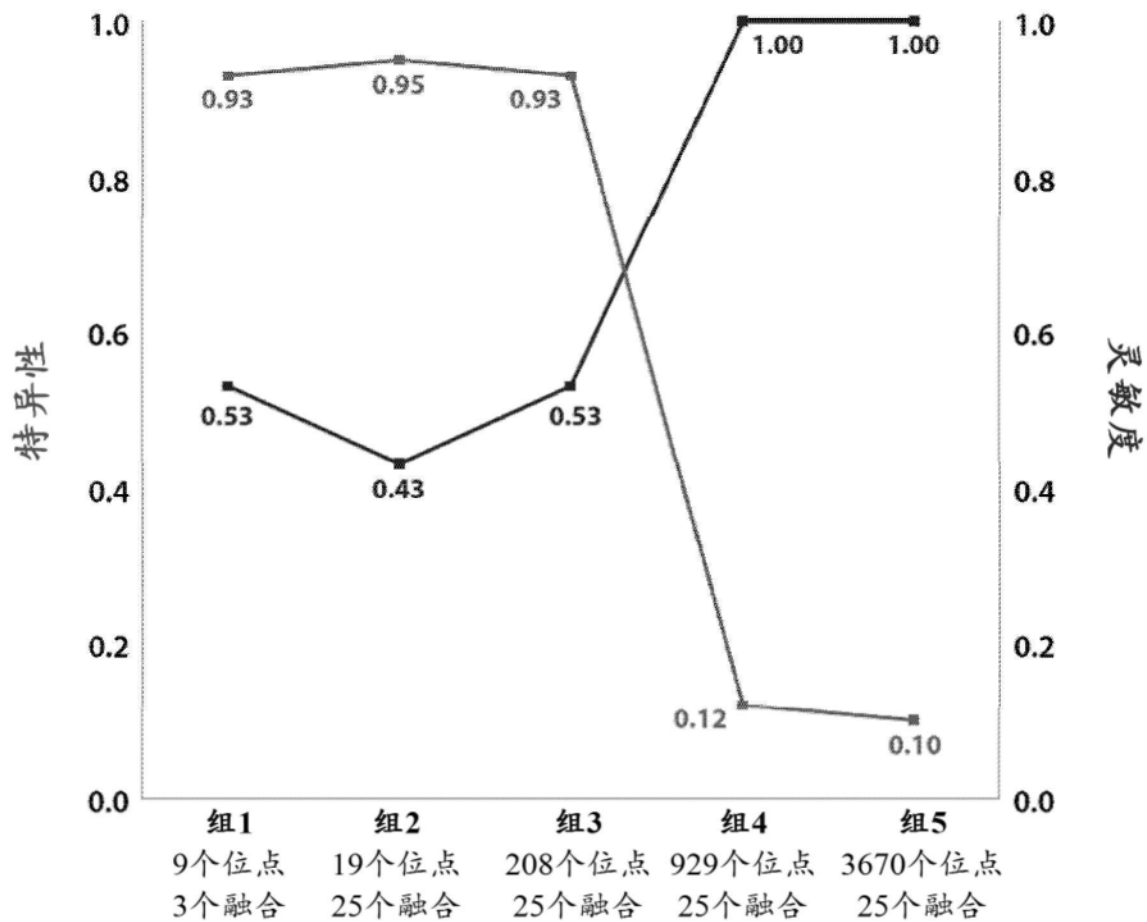
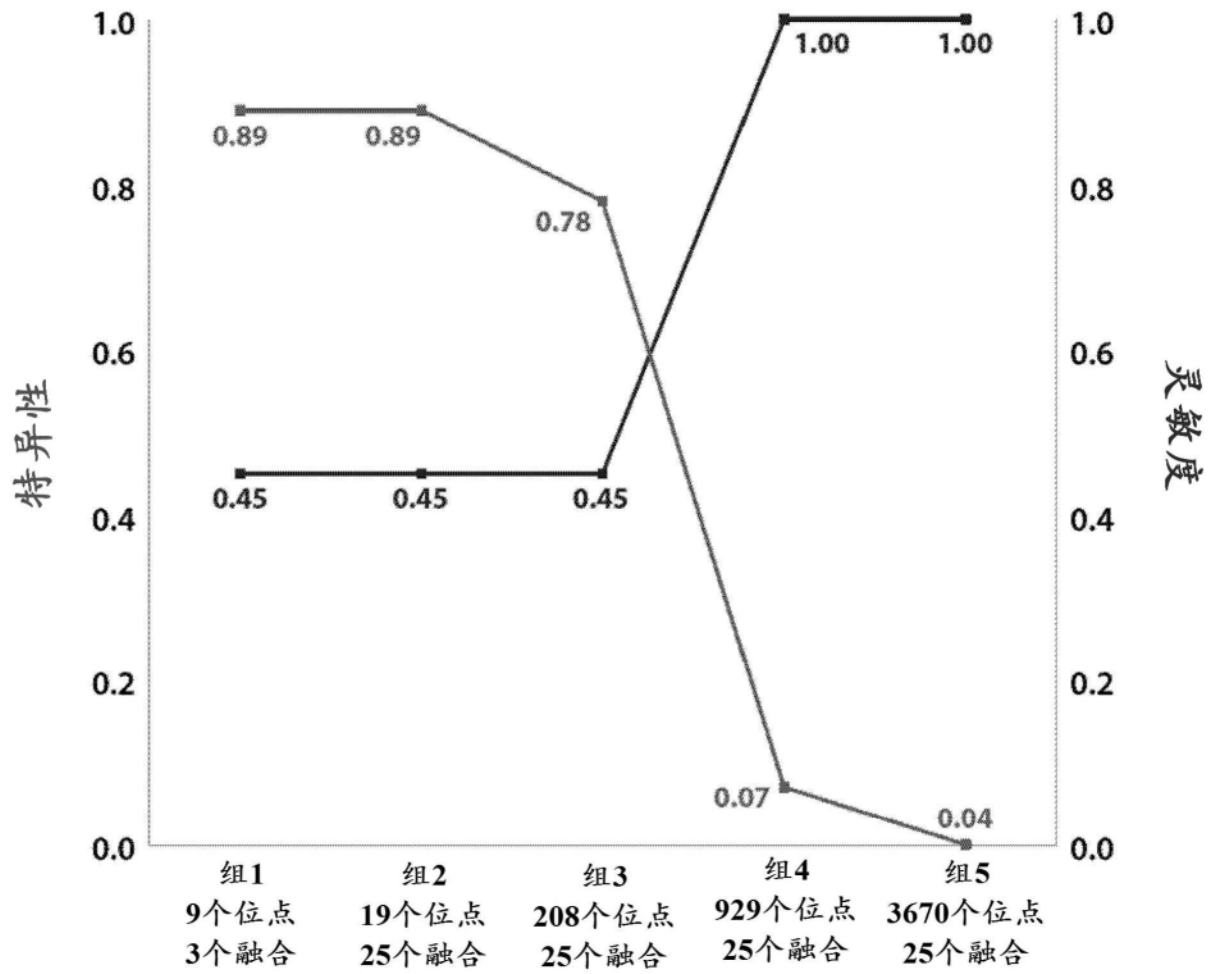


图15

Bethesda 细胞学分类	样品总数	组织学良性 突变(+)/ 总数	组织学恶性 突变(+)/ 总数	灵敏度 (95% CI)	特异性 (95% CI)
Cyto B	28	2/22	3/6	0.50 (0.12-0.88)	0.91 (0.71-0.99)
AUS/FLUS	12	1/11	0/1	0 (0-0.98)	0.91 (0.59-1.00)
FN/SFN	9	0/7	1/2	0.50 (0.01-0.99)	1.00 (0.59-1.00)
SFM	12	0/1	4/11	0.36 (0.11-0.69)	1.00(0.03-1.00)
Cyto M	20	0/0	13/20	0.65 (0.41-0.85)	NA
所有样品	81	3/41	21/40	0.53 (0.36-0.68)	0.93 (0.80-0.98)

图16



\*虽然对总计57个组织进行了评估，但仅38个组织具有待用于测试性能计算的明确的组织学良性或组织学恶性病理学。

图17

A

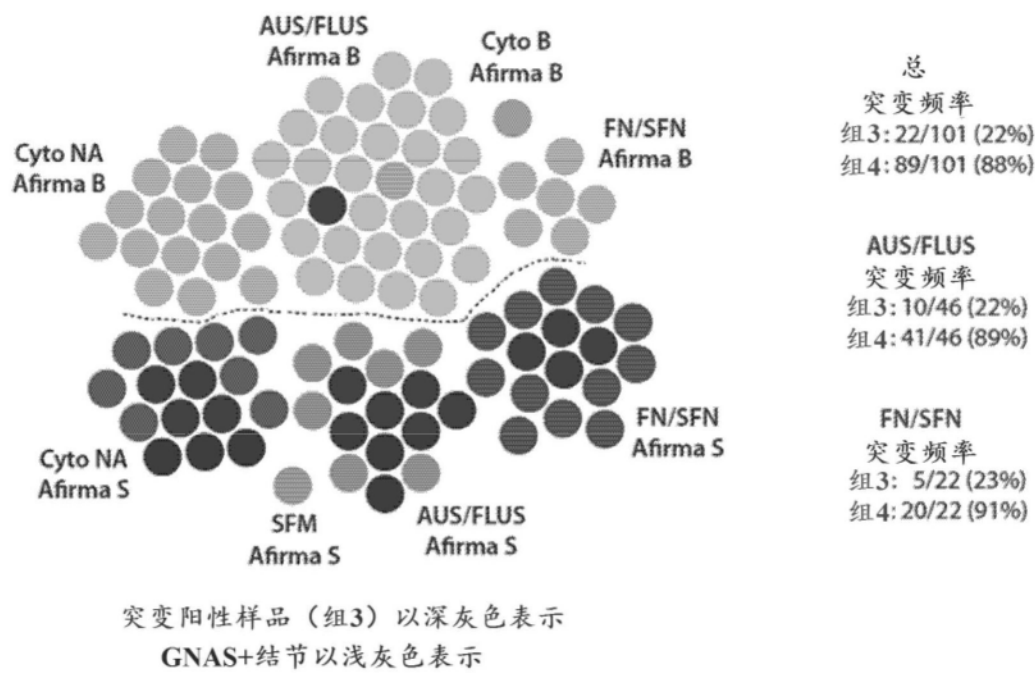


图18A

B

用组3检测		Afirma B (n=54)		Afirma S (n=47)	
基因	突变	AUS/FLUS (n=31)	AUS/FLUS (n=15)	FN/SFN (n=16)	NA (n=16)
BRAF	K601E				1
GNAS	R201H	1			
HRAS	Q61K	1			
HRAS	Q61R			1	
KRAS	G12D		1		
KRAS	Q61R		1	2	1
NRAS	G12V		1		
NRAS	Q61R		1	1	4
TP53	P152L		1		
TP53	R213*				1
融合	PAX8/PPARG		3		
突变总计		2	8	4	7

图18B

A

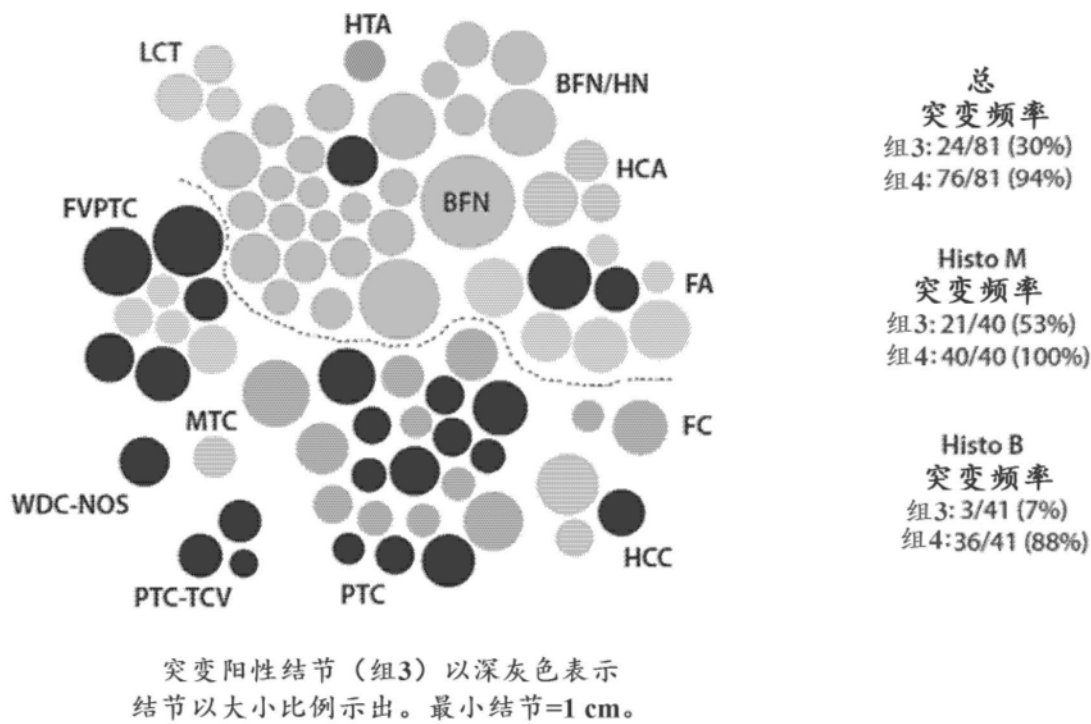


图19A

B

用组3检测		Histo B (n=41)		Histo M (n=40)	
基因	突变	Afirma B (n=31)	Afirma S (n=9)	Afirma B (n=4)	Afirma S (n=36)
BRAF	V600E				15
HRAS	Q61R		1		1
KRAS	G12D				1
NRAS	Q61K	1	1		
NRAS	Q61R				2
TP53	R248Q				1
融合	PAX8/PPARG				2
突变总计		1	2	0	22

图19B

A

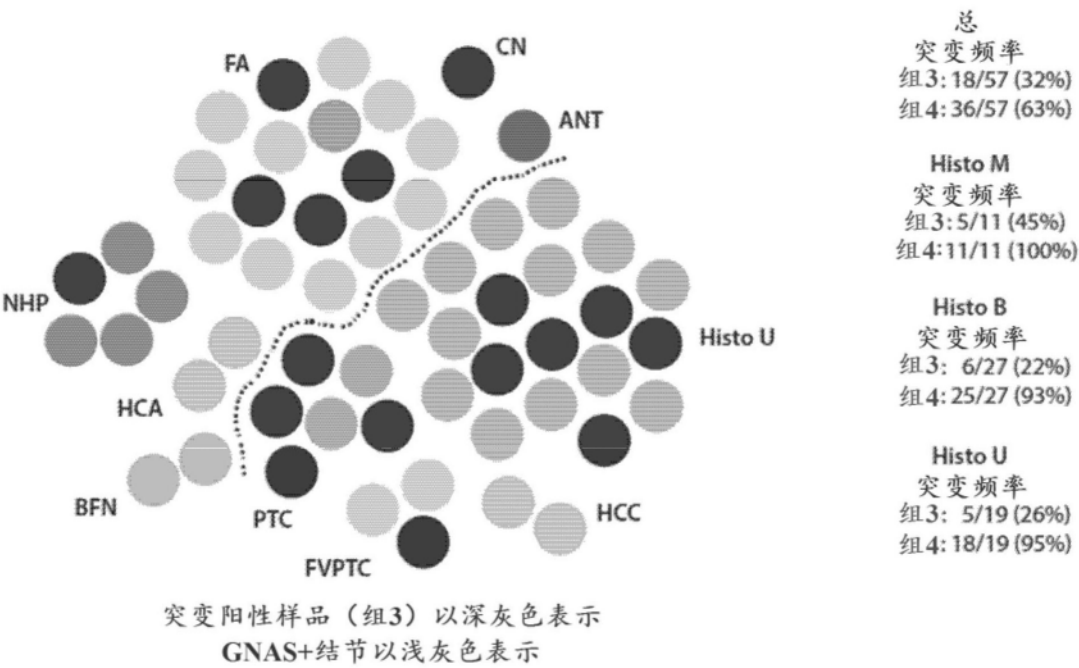


图20A

B

用组3检测		Histo B (n=27)			Histo M (n=11)	Histo U (n=19)
基因	突变	Afirma B (n=8)	Afirma S (n=18)	Afirma NA (n=1)	Afirma S (n=11)	Afirma S (n=16)
BRAF	V600E				5	
GNAS	Q227H		1			
NRAS	Q61K		1			3
NRAS	Q61R		1			2
TSHR	D633Y	2				
TSHR	I486F		1			
融合	PAX8/PPARG					1
融合	CCDC6/RET			1		
突变总计		2	4	1	5	6

图20B