(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118600004 A
(43)申请公布日 2024.09.06

(21)申请号 202410710233.4
(22)申请日 2018.07.11
(30)优先权数据
62/531,527 2017.07.12 US
(62)分案原申请数据
201880059089.5 2018.07.11
(71)申请人 大学健康网络
地址 加拿大安大略省
申请人 西奈卫生系统公司
(72)发明人 丹尼尔·迪尼兹·德·卡尔瓦霍
斯科特·维克托·布拉特曼
拉贾特·辛加尼亚
阿库尔·拉维纳拉亚纳·查克拉瓦西
沈淑怡
(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 王玉桂 武晶晶
(51)Int.Cl.
C12Q 1/6886(2018.01)
C12Q 1/6869(2018.01)
C12Q 1/6827(2018.01)
G16B 40/00(2019.01)
G06F 18/241(2023.01)

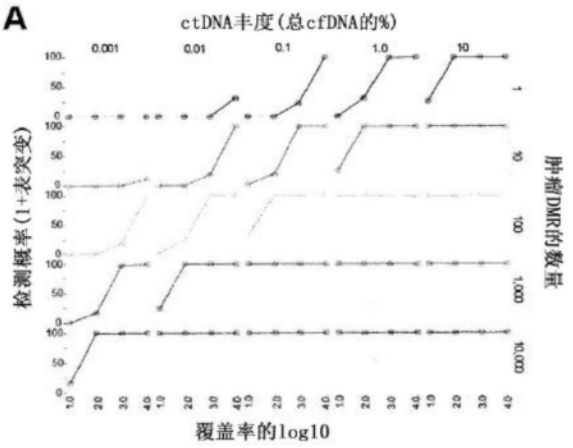
权利要求书2页 说明书21页 附图22页

(54)发明名称

使用甲基化组分析进行癌症检测和分类

(57)摘要

本申请描述了一种在受试者中检测来自癌细胞的DNA的存在的方法,包括:提供来自受试者的无细胞DNA的样品;使样品进行文库制备以允许随后对无细胞甲基化DNA进行测序;向样品中加入第一量的填充DNA,其中至少一部分的填充DNA是甲基化的,然后可选地使样品变性;使用对甲基化多核苷酸具有选择性的结合剂捕获无细胞甲基化DNA;对捕获的无细胞甲基化DNA进行测序;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体以及来自具有不同癌症类型和亚型的个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在。



CN 118600004 A

1.一种在受试者中检测来自癌细胞的DNA的存在的方法,包括:

(a)提供来自受试者的无细胞DNA的样品;

(b)使所述样品进行文库制备以允许随后对所述无细胞甲基化DNA进行测序;

(c)向所述样品中加入第一量的填充DNA,其中至少一部分的所述填充DNA是甲基化的,然后可选地使所述样品变性;

(d)使用对甲基化多核苷酸具有选择性的结合剂捕获无细胞甲基化DNA;

(e)对所捕获的无细胞甲基化DNA进行测序;

(f)比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列;

(g)如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在。

2.根据权利要求1所述的方法,其中所述样品来自所述受试者的血液或血浆。

3.根据权利要求1或2所述的方法,其中比较步骤(f)基于使用统计分类器的拟合。

4.根据权利要求3所述的方法,其中所述分类器是机器学习得出的。

5.一种检测来自癌细胞的DNA的存在并识别癌症亚型的方法,所述方法包括:

a.从受试者样品接收无细胞甲基化DNA的测序数据;

b.比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列;

c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在;以及

d.如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

6.一种检测来自癌细胞的DNA的存在并识别癌症亚型的计算机实现的方法,所述方法包括:

a.至少一个处理器从受试者样品接收无细胞甲基化DNA的测序数据;

b.在所述至少一个处理器处,比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列;

c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则在所述至少一个处理器处识别来自癌细胞的DNA的存在;并且如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

7.一种与具有处理器和连接到所述处理器的存储器的通用计算机结合使用的计算机程序产品,所述计算机程序产品包括其上编码有计算机机制的计算机可读存储介质,其中可以将所述计算机程序机制加载到所述计算机的存储器中,并使所述计算机执行权利要求6所述的方法。

8.一种计算机可读介质,其上存储有用于存储权利要求7所述的计算机程序产品的数据结构。

9.一种用于检测来自癌细胞的DNA的存在并识别癌症亚型的设备,所述设备包括:

至少一个处理器;以及

与所述至少一个处理器通信的电子存储器,所述电子存储器存储处理器可执行代码,所述代码在所述至少一个处理器处执行时,使所述至少一个处理器:

a.从受试者样品接收无细胞甲基化DNA的测序数据;

b.比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列;

c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在,并且如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

10.一种检测来自癌细胞的DNA的存在并确定来自两个或多个可能器官的所述癌细胞所处的癌症的位置的方法,所述方法包括:

(a)提供来自受试者的无细胞DNA的样品;

(b)使用对甲基化多核苷酸具有选择性的结合剂从所述样品中捕获无细胞甲基化DNA;

(c)对所捕获的无细胞甲基化DNA进行测序;

(d)比较所捕获的无细胞甲基化DNA的序列模式与对照个体的两个或多个群体的DNA序列模式,所述两个或多个群体中的每一个在不同器官中具有局限性癌症;

(e)基于所述无细胞DNA的甲基化模式与所述两个或多个群体中一个之间的统计学上显著的相似性,确定所述癌细胞来自哪个器官。

使用甲基化组分析进行癌症检测和分类

[0001] 本申请是申请日为2018年07月11日、申请号为201880059089.5、发明名称为“使用甲基化组分析进行癌症检测和分类”的中国专利申请(其对应PCT申请的申请日为2018年07月11日、申请号为PCT/CA2018/000141)的分案申请。

[0002] 相关申请

[0003] 本申请要求2017年7月12日提交的美国临时专利申请号62/531527的优先权,其通过引用并入文中。

技术领域

[0004] 本发明涉及癌症检测和分类,更具体地说,涉及甲基化组分析用于癌症检测和分类的用途。

背景技术

[0005] 循环无细胞DNA(cfDNA)作为生物标记物来源的用途在肿瘤学中迅速获得发展[1]。cfDNA的DNA甲基化映射作为生物标记物的用途可能会对液体活检领域产生重大影响,因为它可以以微创方式识别起源的组织[2],可以进行癌症类型和亚型分类并分层癌症患者[3]。此外,使用cfDNA的全基因组DNA甲基化映射可以克服在没有疾病的放射线证据下在早期癌症患者中检测循环肿瘤DNA(ctDNA)的关键灵敏度问题。现有的ctDNA检测方法基于测序突变,并且灵敏度有限,部分原因是可用于区分肿瘤和正常循环cfDNA的有限数量的复发性突变[4,5]。另一方面,全基因组DNA甲基化映射利用了可用于区分循环肿瘤DNA(ctDNA)与正常循环无细胞DNA(cfDNA)的大量的表观遗传学改变。例如,某些类型的肿瘤(例如室管膜瘤)可能具有广泛的DNA甲基化畸变,而没有任何明显的复发性体细胞突变[6]。

[0006] 捕获无细胞甲基化DNA的某些方法在WO 2017/190215中有所描述,其通过引用并入本文。

发明内容

[0007] 在一方面,提供了一种在受试者中检测来自癌细胞的DNA的存在的方法,包括:提供来自受试者的无细胞DNA的样品;使样品进行文库制备以允许随后对无细胞甲基化DNA进行测序;向样品中加入第一量的填充DNA(filler DNA),其中至少一部分的填充DNA是甲基化的,然后可选地使样品变性;使用对甲基化多核苷酸具有选择性的结合剂捕获无细胞甲基化DNA;对捕获的无细胞甲基化DNA进行测序;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在。

[0008] 在一方面,提供了一种检测来自癌细胞的DNA的存在并识别癌症亚型的方法,该方法包括:从受试者样品接收无细胞甲基化DNA的测序数据;将捕获的无细胞甲基化DNA的序

列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较进一步识别起源的癌细胞组织和癌症亚型。

[0009] 在一方面,提供了一种计算机实现的方法,其检测来自癌细胞的DNA的存在并识别癌症亚型,该方法包括:至少一个处理器从受试者样品接收无细胞甲基化DNA的测序数据;在至少一个处理器处,将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则在至少一个处理器处识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较进一步识别起源的癌细胞组织和癌症亚型。

[0010] 在一方面,提供了一种与具有处理器和连接到该处理器的存储器的通用计算机结合使用的计算机程序产品,该计算机程序产品包括其上编码有计算机机制的计算机可读存储介质,其中可以将计算机程序机制加载到计算机的存储器中,并使计算机执行本文所描述的方法。

[0011] 在一方面,提供了一种计算机可读介质,其上存储有用于存储本文所描述的计算机程序产品的数据结构。

[0012] 在一方面,提供了一种用于检测来自癌细胞的DNA的存在并识别癌症亚型的设备,该设备包括:至少一个处理器;以及与至少一个处理器通信的电子存储器,电子存储器存储处理器可执行代码,该代码在至少一个处理器处执行时,使至少一个处理器:从受试者样品接收无细胞甲基化DNA的测序数据;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较进一步识别起源的癌细胞组织和癌症亚型。

[0013] 在一方面,提供了一种检测来自癌细胞的DNA的存在并确定来自两个或多个可能器官的癌细胞所处的癌症的位置的方法,该方法包括:提供来自受试者的无细胞DNA的样品;使用对甲基化多核苷酸具有选择性的结合剂从所述样品中捕获无细胞甲基化DNA;对捕获的无细胞甲基化DNA进行测序;比较捕获的无细胞甲基化DNA的序列模式与对照个体的两个或多个群体的DNA序列模式,所述两个或多个群体中的每一个在不同器官中具有局限性癌症;基于无细胞DNA的甲基化模式与所述两个或多个群体中一个之间的统计学上显著的相似性,确定癌细胞来自哪个器官。

附图说明

[0014] 在下面的详细描述中,本发明的优选实施例的这些和其他特征将变得更加明显,在这些详细描述中参考了附图,其中:

[0015] 图1A-1G显示,对cfDNA的甲基化组分析是一种高灵敏度的方法,可以在少量输入DNA中富集和检测ctDNA。A)作为ctDNA浓度(列),所研究的DMR数量(行)和测序深度(x轴)函数的检测至少一个表突变的概率的计算机模拟。B)1至100ng来自被片段化以模拟血浆

cfDNA的HCT116细胞系的输入DNA的DNA甲基化信号之间的全基因组Pearson相关性。每个浓度具有两个生物学重复。C)从不同浓度的来自HCT116的输入DNA的cfMeDIP-seq获得的DNA甲基化图谱(绿色轨道)加上从ENCODE(ENCSR000DFS)获得的RRBS(简化代表性亚硫酸氢盐测序)HCT116数据和从GEO(GSM1465024)获得的WGBS(全基因组亚硫酸氢盐测序)HCT116数据。对于热图(RRBS轨道),黄色表示甲基化,蓝色表示未甲基化,灰色表示无覆盖。D-E)将CRC细胞系HCT116在多发性骨髓瘤(MM)细胞系MM1.S中系列稀释。cfMeDIP-seq在纯HCT116 DNA(100%CRC),纯MM1.SDNA(100%MM)和在MM DNA中稀释的10%、1%、0.1%、0.01%和0.001%CRC DNA中进行。将所有DNA片段化以模拟血浆cfDNA。我们观察到DMR数量(D)以及(E)这些DMR中的DNA甲基化信号(以RPKM表示)的观测值与预期值之间几乎完美的线性相关性(r^2=0.99,p<0.0001)。F)在相同的稀释系列中,只有通过超出背景测序仪和聚合酶错误率的超深(>10,000X)靶向测序,才能在1/100等位基因分数下检测到已知的体细胞突变。显示的是在CRC细胞系中每个突变位点包含每个碱基或插入/缺失的读取分数。G)携带来自两名结肠直肠癌患者的患者源性异种移植物(PDX)的小鼠的血浆中ctDNA(人类)占总cfDNA(人类+小鼠)的百分比。

[0016] 图2A-2D显示,对血浆cfDNA的甲基化组分析可以进行肿瘤分类。A)示意性说明了用于癌症分类的机器学习分类器构建方法。B)包含在多类弹性网机器学习分类器中的DMR的热图。对来自健康供体(n=24)、肺癌(n=25)、乳腺癌(n=25)、结肠直肠癌(n=23)、急性髓系白血病(AML)(n=28)和多形性胶质母细胞瘤(GBM)(n=71)的血浆DNA样品进行了分类器训练。层次聚类方法:Ward。C)通过tSNE(t分布随机邻域嵌入)对在10%或25%的模型中识别的癌症类型相关DMR的2D可视化。D)基于血浆cfDNA甲基化的多癌分类器的性能指标。生成50倍弹性网机器学习分类器后,Y轴上显示每种癌症类型和健康供体的接受者工作曲线下的面积(auROC)。

[0017] 图3A-3B显示了多癌症分类器对单独队列的验证。A)显示了多癌症分类器对肺癌(LUC)(n=55LUC相对于其他n=97)、AML(n=35AML相对于其他n=117)和健康供体队列(n=62个健康供体相对于其他n=90)的单独验证的ROC曲线。B)显示了多癌症分类器对早期LUC(n=32个I-II期LUC相对于其他n=97)和晚期LUC(n=23个III-IV期LUC相对于其他n=97)的单独验证的ROC曲线。

[0018] 图4A-4E显示了,对血浆cfDNA的甲基化组分析可以进行肿瘤亚型分类。A)通过tSNE(t分布随机邻域嵌入)对癌症亚型相关DMR的2D可视化。乳腺癌亚型表现出区分具有不同基因表达模式和转录因子活性(ER状态)以及不同肿瘤拷贝数畸变(HER2状态)的肿瘤患者的能力。AML亚型显示出区分具有不同重排(FLT3状态)的肿瘤患者的能力。多形性胶质母细胞瘤(GBM)亚型显示出区分具有不同点突变(IDH基因突变状态)的肿瘤患者的能力。肺癌亚型显示出区分具有不同组织学的的肿瘤患者的能力(腺癌相对于鳞状癌相对于小细胞癌),这具有预后和治疗意义。B)热图(Heatmap)显示了,顶部DMR可准确区分乳腺癌血浆样品中的三种乳腺癌亚型。C)热图显示了,顶部DMR可准确区分AML患者血浆样品中的FLT3-ITD状态。D)热图显示了,顶部DMR可准确区分多形性胶质母细胞瘤(GBM)患者血浆样品中的IDH基因突变状态。E)热图显示了,顶部DMR可准确区分肺癌血浆样品中的三种肺癌组织学。

[0019] 图5显示了适当配置的计算机设备以及相关联的通信网络、设备、软件和固件,以提供用于实现本文所描述的一个或多个实施例的平台。

[0020] 图6A-6C显示了测序饱和度分析和质量控制。A)该图显示了,Bioconductor软件包MEDIPS分析来自被片段化以模拟血浆cfDNA的HCT116 DNA的每个输入浓度的每次重复的cfMeDIP-seq数据的饱和度分析结果。B)对HCT116细胞系的四个起始DNA浓度(100、10、5和1ng)进行两次重复测试。使用甲基化和未甲基化的加入的拟南芥DNA计算反应的特异性。使用片段化的HCT116 DNA的基因组区域(甲基化睾丸特异性H2B、TSH2B0和未甲基化的人类DNA区域(GAPDH启动子)的引物)计算富集倍数比。水平虚线表示富集倍数比阈值为25。误差条表示±1s.e.m.。C)测序样品的CpG富集得分显示,与输入对照相比,从免疫沉淀样品中在基因组区域内大量富集CpG。CpG富集得分通过将区域中CpG的相对频率除以人类基因组中CpG的相对频率来获得。误差条表示±1s.e.m.。

具体实施方式

[0021] 在以下描述中,阐述了许多具体细节以提供对本发明的透彻理解。然而,应理解,可以在没有这些具体细节的情况下实施本发明。

[0022] DNA甲基化图谱是细胞类型特异性的,并在癌症中被破坏。使用被设计成对微量循环无细胞DNA(cfDNA)进行甲基化组分析的稳健而灵敏的方法,我们识别了上千个差异性甲基化区域(DMR),它们将多种肿瘤类型彼此以及与健康个体区分开。对cfDNA的甲基化组分析高度灵敏并适用于检测早期患者的循环肿瘤DNA(ctDNA)。使用cfDNA甲基化组的机器学习衍生分类器能够根据交叉验证正确分类来自患有5种癌症类型的患者和健康供体的196个血浆样品。在单独验证中,使用在血浆cfDNA中识别的相同DMR,分类器能够正确分类AML、肺癌、和健康供体以及早期和晚期肺癌。因此,对cfDNA的甲基化组分析可用于ctDNA的无创早期检测,并对癌症类型进行可靠分类。

[0023] 在一方面,提供了一种在受试者中检测来自癌细胞的DNA的存在的方法,包括:提供来自受试者的无细胞DNA的样品;使样品进行文库制备以允许随后对无细胞甲基化DNA进行测序;向样品中加入第一量的填充DNA,其中至少一部分的填充DNA是甲基化的,然后可选地使样品变性;使用对甲基化多核苷酸具有选择性的结合剂捕获无细胞甲基化DNA;对捕获的无细胞甲基化DNA进行测序;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在。

[0024] 申请人的共同拥有的申请:2016年5月3日提交的美国临时专利申请第62/331,070号和2017年5月3日提交的国际专利申请第PCT/CA2017/000108号描述了捕获无细胞甲基化DNA的方法并通过引用并入文中。

[0025] 传统上,癌症已按来源组织进行了分类—例如,结肠直肠癌、乳腺癌、肺癌等。在临床肿瘤学的现代实践中,能够通过各种分子、发育以及功能基础区分癌症亚型变得越来越重要。治疗决策通常取决于癌症的精确亚型,临床医生可能有必要在开始治疗之前识别亚型。可能影响治疗决策的癌症亚型的实例包括(但不限于)阶段(例如,手术治疗的早期肺癌相对于化疗治疗的晚期肺癌)、组织学(例如,肺癌中的小细胞癌相对于腺癌相对于鳞状细胞癌)、基因表达模式或转录因子活性(例如乳腺癌中的ER状态)、拷贝数畸变(例如乳腺癌中的HER2状态)、特定的重排(例如AML中的FLT3)、特定的基因点突变状态(例如,IDH基因点

突变)和DNA甲基化模式(例如,脑癌中的MGMT基因启动子甲基化)。

[0026] 本文描述的方法适用于多种癌症,包括但不限于肾上腺癌、肛门癌、胆管癌、膀胱癌、骨癌、脑/cns肿瘤、乳腺癌、Castleman病、宫颈癌、结肠癌/直肠癌、子宫内膜癌、食道癌、尤文家族肿瘤、眼癌、胆囊癌、胃肠道类癌、胃肠道间质瘤(gist)、妊娠滋养细胞疾病、霍奇金病、卡波西肉瘤、肾癌、喉和下咽癌、白血病(急性淋巴细胞、急性髓样、慢性淋巴细胞、慢性髓样、慢性粒单核细胞)、肝癌、肺癌(非小细胞、小细胞、肺类癌)、淋巴瘤、皮肤淋巴瘤、恶性间皮瘤、多发性骨髓瘤、骨髓增生异常综合征、鼻腔和鼻旁窦癌、鼻咽癌、神经母细胞瘤、非霍奇金淋巴瘤、口腔和口咽癌、骨肉瘤、卵巢癌、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-成人软组织癌、皮肤癌(基底和鳞状细胞、黑素瘤、默克尔细胞)、小肠癌、胃癌、睾丸癌、胸腺癌、甲状腺癌、子宫肉瘤、阴道癌、外阴癌、Waldenstrom巨球蛋白血症、Wilms瘤。

[0027] 多种测序技术是本领域技术人员已知的,例如聚合酶链式反应(PCR),然后进行Sanger测序。还可以使用下一代测序(NGS)技术,也称为高通量测序,其包括各种测序技术,包括:Illumina(Solexa)测序、Roche 454测序、Ion torrent:质子/PGM测序、SOLiD测序。NGS可以比以前使用的Sanger测序更快且更便宜地对DNA和RNA进行测序。在一些实施方式中,对所述测序优化用于短读取测序。

[0028] 如本文所用,术语“受试者”是指动物界的任何成员,优选为人类,最优选为患有,曾患有或疑似患有前列腺癌的人类。

[0029] 无细胞甲基化DNA是在血液中自由循环的DNA,并在DNA的各种已知区域被甲基化。可以采集样品(例如血浆样品)来分析无细胞甲基化DNA。因此,在一些实施方式中,样品是受试者的血液或血浆。

[0030] 如本文所用,“文库制备”包括列表末端修复、A-加尾、衔接子连接或对无细胞DNA进行以允许DNA的后续测序的任何其他制备。

[0031] 如本文所用,“填充DNA”可以是非编码DNA或可以由扩增子组成。

[0032] 可以例如使用足够的热量使DNA样品变性。

[0033] 在一些实施方式中,比较步骤基于使用统计分类器的拟合。使用DNA甲基化数据的统计分类器可用于将样品分配给特定的疾病状态,例如癌症类型或亚型。出于癌症类型或亚型分类的目的,分类器将由统计模型内的一个或多个DNA甲基化变量(即特征)组成,并且统计模型的输出将具有一个或多个阈值以区分不同的疾病状态。统计分类器中使用的特定特征和阈值可以从癌症类型或亚型的先验知识、可能最有用的特征的先验知识、机器学习、或两种或多种这些方法的组合得到。

[0034] 在一些实施方式中,分类器是机器学习得到的。优选地,分类器是弹性网分类器、套索、支持向量机、随机森林或神经网络。

[0035] 被分析的基因组空间可以是全基因组的,或者优选地限于调节区域(即,FANTOM5增强子、CpG岛、CpG海岸和CpG架子)。

[0036] 优选地,回收的spike-in甲基化DNA的百分比作为协变量包括在内,以控制下拉效率变化。

[0037] 对于能够彼此区分多种癌症类型(或亚型)的分类器,分类器将优选地由来自感兴趣的每种类型(或亚型)的成对比较的差异性甲基化区域组成。

[0038] 在一些实施方式中,来自健康和癌症个体的对照无细胞甲基化DNA序列包含在健康和癌症个体之间的差异性甲基化区域(DMR)的数据库中。

[0039] 在一些实施方式中,来自健康和癌症个体的对照无细胞甲基化DNA序列限于在源自体液的无细胞DNA的DNA中的健康和癌症个体之间是差异性甲基化的那些对照无细胞甲基化DNA序列,例如来自血清、脑脊髓液、尿液、痰、胸膜液、腹水、眼泪、汗液、巴氏涂片液、内窥镜刷液等,优选地来自血浆。

[0040] 在一些实施方式中,样品具有小于100ng、75ng或50ng的无细胞DNA。

[0041] 在一些实施方式中,第一量的填充DNA包含约5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%或100%的甲基化填充DNA,其余为未甲基化的填充DNA,优选地为5%至50%、10%至40%或15%至30%的甲基化填充DNA。

[0042] 在一些实施方式中,第一量的填充DNA为20ng至100ng,优选地30ng至100ng,更优选地50ng至100ng。

[0043] 在一些实施方式中,来自样品的无细胞DNA和第一量的填充DNA一起包含至少50ng的总DNA,优选地至少100ng的总DNA。

[0044] 在一些实施方式中,填充DNA的长度为50bp至800bp,优选地100bp至600bp,并且更优选地200bp至600bp。

[0045] 在一些实施方式中,填充DNA是双链的。填充DNA是双链的。例如,填充DNA可以是垃圾DNA。填充DNA也可以是内源性或外源性DNA。例如,填充DNA是非人类DNA,并且在优选的实施例中,是λDNA。如本文所用,“λDNA”是指肠杆菌噬菌体λDNA。在一些实施方式中,填充DNA与人类DNA没有比对。

[0046] 在一些实施方式中,结合剂是包含甲基-CpG结合结构域的蛋白质。一种这样的示例性蛋白质是MBD2蛋白质。如本文所用,“甲基-CpG结合结构域(MBD)”是指蛋白质和酶的某些结构域,其长度约70个残基,并结合至包含一个或多个对称甲基化CpG的DNA。MeCP2、MBD1、MBD2、MBD4和BAZ2的MBD介导与DNA的结合,在MeCP2、MBD1和MBD2的情况下,优先与甲基化CpG结合。人类蛋白质MECP2、MBD1、MBD2、MBD3和MBD4包含存在于每个甲基CpG结合结构域(MBD)中而相关的核蛋白家族。除MBD3外,这些蛋白质中的每一种都能够与甲基化DNA特异性结合。

[0047] 在其他实施方式中,结合剂是抗体,并且捕获无细胞甲基化DNA包括使用抗体将无细胞甲基化DNA免疫沉淀。如本文所用,“免疫沉淀”是指使用与该特定抗原特异性结合的抗体从溶液中沉淀出抗原(例如多肽和核苷酸)的技术。此过程可用于从样品中分离和浓缩特定的蛋白质或DNA,并要求在该过程中的某个时刻将抗体偶联到固体底物上。固体底物包括例如珠子,例如磁珠。其他类型的珠子和固体底物是本领域已知的。

[0048] 一种示例性抗体是5-MeC抗体。对于免疫沉淀过程,在一些实施方式中,将至少0.05μg抗体添加至样品;而在更优选的实施例中,将至少0.16μg抗体加入样品中。为了确认免疫沉淀反应,在一些实施方式中,本文所描述的方法还包括将第二量的对照DNA添加至样品中的步骤。

[0049] 在一些实施方式中,该方法还包括将第二量的对照DNA添加至样品中以确认免疫沉淀反应的步骤。

[0050] 如本文所用,“对照”可包括阳性和阴性对照,或至少阳性对照。

[0051] 在一些实施方式中,该方法还包括将第二量的对照DNA添加至样品中以确认无细胞甲基化DNA的捕获的步骤。

[0052] 在一些实施方式中,识别来自癌细胞的DNA的存在还包括识别起源的癌细胞组织。

[0053] 在一些情况下,肿瘤组织采样可能具有挑战性或具有重大风险,在这种情况下,可能期望在不需要肿瘤组织采样的情况下诊断和/或分型癌症。例如,肺肿瘤组织采样可能需要侵入性手术,例如纵隔镜检查、开胸手术或经皮穿刺活检;这些手术可能需要住院、胸管、机械通气、抗生素或其他医疗介入。由于医学上的并存病或由于偏爱,因此一些个体可能未进行肿瘤组织采样所需的侵入性手术。在某些情况下,获取肿瘤组织的实际程序可能取决于疑似的癌症亚型。在其他情况下,癌症亚型可能在同一个体内随时间进展。通过侵入性肿瘤组织采样程序的连续评估通常是不切实际的,患者对此不能很好地耐受。因此,通过血液测试的非侵入性癌症分型在临床肿瘤学实施中可能具有许多有利的应用。

[0054] 因此,在一些实施方式中,识别起源的癌细胞组织还包括识别癌症亚型。优选地,癌症亚型基于阶段(例如,手术治疗的早期肺癌相对于化疗治疗的晚期肺癌)、组织学(例如,肺癌中的小细胞癌相对于腺癌相对于鳞状细胞癌)、基因表达模式或转录因子活性(例如乳腺癌中的ER状态)、拷贝数畸变(例如乳腺癌中的HER2状态)、特定的重排(例如AML中的FLT3)、特定的基因点突变状态(例如,IDH基因点突变)和DNA甲基化模式(例如,脑癌中的MGMT基因启动子甲基化)来区分癌症。

[0055] 在一些实施方式中,步骤(f)中的比较在全基因组范围内进行。

[0056] 在其他实施方式中,步骤(f)中的比较从全基因组限制到特定的调节区域,例如但不限于FANTOM5增强子、CpG岛、CpG海岸、CpG架子或前述的任何组合。

[0057] 在一些实施方式中,某些步骤由计算机处理器执行。

[0058] 在一方面,提供了一种检测来自癌细胞的DNA的存在并识别癌症亚型的方法,该方法包括:从受试者样品接收无细胞甲基化DNA的测序数据;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较步骤进一步识别起源的癌细胞组织和癌症亚型。

[0059] 在一方面,提供了一种检测来自癌细胞的DNA的存在并确定来自两个或多个可能器官的癌细胞所处的癌症的位置的方法,该方法包括:提供来自受试者的无细胞DNA的样品;使用对甲基化多核苷酸具有选择性的结合剂从所述样品中捕获无细胞甲基化DNA;对捕获的无细胞甲基化DNA进行测序;比较捕获的无细胞甲基化DNA的序列模式与对照个体的两个或多个群体的DNA序列模式,所述两个或多个群体中的每一个在不同器官中具有局限性癌症;基于无细胞DNA的甲基化模式与所述两个或多个群体中一个之间的统计学上显著的相似性,确定癌细胞来自哪个器官。

[0060] 可以在各种实施方式中实施本发明系统和方法。适当配置的计算机设备以及相关联的通信网络、设备、软件和固件可以提供用于实现如上所述的一个或多个实施例的平台。通过实例,图5示出了通用计算机设备100,其可以包括连接到存储单元104和随机存取存储器106的中央处理单元(“CPU”)102。CPU102可以处理操作系统101,应用程序103和数据123。操作系统101,应用程序103和数据123可以存储在存储单元104中,并根据需要加载到存储

器106中。计算机设备100可以进一步包括图形处理单元(GPU)122,其可操作地连接到CPU102和存储器106,以从CPU102卸载密集的图像处理计算并且与CPU102并行地运行这些计算。使用通过视频接口105连接的视频显示器108,以及通过I/O接口109连接的各种输入/输出设备,例如键盘115、鼠标112和磁盘驱动器或固态驱动器114,操作者107可以与计算机设备100交互。通过已知方式,鼠标112可以被配置为控制光标在视频显示器108中的移动,并使用鼠标按钮来操作出现在视频显示器108中的各种图形用户界面(GUI)控件。磁盘驱动器或固态驱动器114可以被配置为接受计算机可读介质116。计算机设备100可以经由网络接口111形成网络的一部分,从而允许计算机设备100与其他适当配置的数据处理系统(非显示)通信。一种或多种不同类型的传感器135可以用于从各种来源接收输入。

[0061] 本发明系统和方法实际上可以在任何形式的计算机设备上实施,其包括台式计算机、膝上型计算机、平板计算机或无线手持设备。本发明系统和方法还可以被实现为包括计算机程序代码的计算机可读/可用介质,以使一个或多个计算机设备能够实现根据本发明的方法中的各个处理步骤中的每一个。在执行整个操作的多于一个计算机设备的情况下,将计算机设备联网以分配操作的各个步骤。应理解,术语计算机可读介质或计算机可用介质包括程序代码的物理实施例的任何类型中的一个或多个。特别地,计算机可读/可用介质可以包括体现在一个或多个便携式存储制品(例如光盘、磁盘、磁带等)上,在一个或多个计算设备的数据存储部分(例如与计算机和/或存储系统关联的存储器)上的程序代码。

[0062] 在一方面,提供了一种计算机实现的方法,其检测来自癌细胞的DNA的存在并识别癌症亚型,该方法包括:至少一个处理器从受试者样品接收无细胞甲基化DNA的测序数据;在至少一个处理器处,将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则在至少一个处理器处识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较步骤进一步识别起源的癌细胞组织和癌症亚型;

[0063] 在一方面,提供了一种与具有处理器和连接到该处理器的存储器的通用计算机结合使用的计算机程序产品,该计算机程序产品包括其上编码有计算机机制的计算机可读存储介质,其中可以将计算机程序机制加载到计算机的存储器中,并使计算机执行本文所描述的方法。

[0064] 在一方面,提供了一种计算机可读介质,其上存储有用于存储本文所描述的计算机程序产品的数据结构。

[0065] 在一方面,提供了一种用于检测来自癌细胞的DNA的存在并识别癌症亚型的设备,该设备包括:至少一个处理器;以及与至少一个处理器通信的电子存储器,电子存储器存储处理器可执行代码,该代码在至少一个处理器处执行时,使至少一个处理器:从受试者样品接收无细胞甲基化DNA的测序数据;将捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列相比较;如果捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在;并且如果识别出来自癌细胞的DNA,则基于比较步骤进一步识别起源的癌细胞组织和癌症亚型。

[0066] 如本文所用,“处理器”可以是任何类型的处理器,诸如,例如,任何类型的通用微

处理器或微控制器(例如,Intel™x86、PowerPC™、ARM™处理器等)、数字信号处理(DSP)处理器、集成电路、现场可编程门阵列(FPGA)或其任意组合。

[0067] 如本文所用,“存储器”可以包括位于内部或外部的任何类型的计算机存储器的适当组合,诸如,例如,随机存取存储器(RAM)、只读存储器(ROM)、光盘只读存储器(CDROM)、电光存储器、磁光存储器、可擦除可编程只读存储器(EPROM)和电可擦除可编程只读存储器(EEPROM)等。存储器102的部分可以使用常规文件系统来组织,并通过控制设备的整体操作的操作系统来控制和管理。

[0068] 如本文所用,“计算机可读存储介质”(也称为其中体现有计算机可读程序代码的机器可读介质、处理器可读介质或计算机可用介质)是能够以计算机或机器可读的格式存储数据的介质。机器可读介质可以是任何合适的有形、非暂时性介质,包括磁、光或电存储介质,包括磁盘、光盘只读存储器(CD-ROM)、存储设备(易失性或非易失性)、或类似的存储机制。所述计算机可读存储介质可以包含指令、代码序列、配置信息或其他数据的各种集合,这些集合在被执行时,使处理器执行根据本公开的实施例的方法中的步骤。本领域普通技术人员将意识到,实现所述实现方式所必需的其他指令和操作也可以存储在计算机可读存储介质上。存储在计算机可读存储介质上的指令可以由处理器或其他合适的处理设备执行,并且可以与电路交互以执行所描述的任务。

[0069] 如本文所用,“数据结构”是在计算机中组织数据以便可以被有效地使用的特定方式。数据结构可以实现一种或多种特定的抽象数据类型(ADT),它们指定可以对数据结构执行的操作以及这些操作的计算复杂性。相比之下,数据结构是ADT提供的规范的具体实现。

[0070] 本发明的优点通过以下实例进一步说明。本文阐述的实例及其具体细节仅用于举例说明,不应解释为对本发明权利要求的限制。

[0071] 本发明提供了包括但不限于以下实施方式:

[0072] 1.一种在受试者中检测来自癌细胞的DNA的存在的方法,包括:

[0073] (a)提供来自受试者的无细胞DNA的样品;

[0074] (b)使所述样品进行文库制备以允许随后对所述无细胞甲基化DNA进行测序;

[0075] (c)向所述样品中加入第一量的填充DNA,其中至少一部分的所述填充DNA是甲基化的,然后可选地使所述样品变性;

[0076] (d)使用对甲基化多核苷酸具有选择性的结合剂捕获无细胞甲

[0077] 基化DNA;

[0078] (e)对所捕获的无细胞甲基化DNA进行测序;

[0079] (f)比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症

[0080] 个体的对照无细胞甲基化DNA序列;

[0081] (g)如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在。

[0082] 2.根据实施方式1所述的方法,其中所述样品来自所述受试者的血液或血浆。

[0083] 3.根据实施方式1或2所述的方法,其中比较步骤(f)基于使用统计分类器的拟合。

[0084] 4.根据实施方式3所述的方法,其中所述分类器是机器学习得出的。

[0085] 5.根据实施方式4所述的方法,其中所述分类器是弹性网分类器、套索、支持向量机、随机森林或神经网络。

[0086] 6.根据实施方式1至5中任一项所述的方法,其中来自健康和癌症个体的所述对照无细胞甲基化DNA序列包含在健康和癌症个体之间的差异性甲基化区域(DMR)的数据库中。

[0087] 7.根据实施方式1至6中任一项所述的方法,其中来自健康和癌症个体的所述对照无细胞甲基化DNA序列限于在源自无细胞DNA的DNA中的在健康和癌症个体之间是差异性甲基化的那些对照无细胞甲基化DNA序列。

[0088] 8.根据实施方式7所述的方法,其中所述对照无细胞甲基化DNA序列在源自血浆的DNA中在健康和癌症个体之间是差异性甲基化的。

[0089] 9.根据实施方式1至8中任一项所述的方法,其中所述样品具有小于100

[0090] ng、75ng或50ng的无细胞DNA。

[0091] 10.根据实施方式1至9中任一项所述的方法,其中所述第一量的填充DNA

[0092] 包含约5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%或100%的甲基化填充DNA,其余为未甲基化的填充DNA,优选地为5%至50%、10%至40%或15%至30%的甲基化填充DNA。

[0093] 11.根据实施方式1至9中任一项所述的方法,其中所述第一量的填充DNA

[0094] 为20ng至100ng,优选为30ng至100ng,更优选为50ng至100ng。

[0095] 12.根据实施方式1至11中任一项所述的方法,其中来自所述样品的所述无细胞DNA和所述第一量的填充DNA一起包含至少50ng的总DNA,优选至少100ng的总DNA。

[0096] 13.根据实施方式1至12中任一项所述的方法,其中所述填充DNA的长度为50bp至800bp,优选为100bp至600bp,并且更优选为200bp

[0097] 至600bp。

[0098] 14.根据实施方式1至13中任一项所述的方法,其中所述填充DNA是双链的。

[0099] 15.根据实施方式1至2中任一项所述的方法,其中所述填充DNA是垃圾DNA。

[0100] 16.根据实施方式1至3中任一项所述的方法,其中所述填充DNA是内源性或外源性DNA。

[0101] 17.根据实施方式16所述的方法,其中所述填充DNA是非人类DNA,优选为λDNA。

[0102] 18.根据实施方式1至17中任一项所述的方法,其中所述填充DNA与人类DNA没有比对。

[0103] 19.根据实施方式1至18中任一项所的方法,其中所述结合剂是包含甲基-CpG结合结构域的蛋白质。

[0104] 20.根据实施方式1至19中任一项所述的方法,其中所述蛋白质是MBD2蛋白质。

[0105] 21.根据实施方式1至20中任一项所述的方法,其中步骤(d)包括使用抗体将所述无细胞甲基化DNA免疫沉淀。

[0106] 22.根据实施方式21所述的方法,包括将至少0.05μg的所述抗体添加至所述样品用于免疫沉淀,并且优选地至少0.16μg。

[0107] 23.根据实施方式21所述的方法,其中所述抗体是5-MeC抗体。

[0108] 24.根据实施方式21所述的方法,还包括在步骤(c)之后向所述样品中添加第二量的对照DNA以确认所述免疫沉淀反应的步骤。

[0109] 25.根据实施方式1至23中任一项所述的方法,还包括在步骤(c)之后向所述样品中添加第二量的对照DNA以确认所述无细胞甲基化DNA的捕获的步骤。

[0110] 26.根据实施方式1至25中任一项所述的方法,其中识别来自癌细胞的DNA的存在还包括识别起源的癌细胞组织。

[0111] 27.根据实施方式26所述的方法,其中识别所述起源的癌细胞组织还包括识别癌症亚型。

[0112] 28.根据实施方式27所述的方法,其中所述癌症亚型基于阶段、组织学、基因表达模式、拷贝数畸变、重排或点突变状态来区分所述癌症。

[0113] 29.根据实施方式1至28中任一项所述的方法,其中步骤(f)中的所述比较在全基因组范围内进行。

[0114] 30.根据实施方式1至28中任一项所述的方法,其中步骤(f)中的所述比较从全基因组限制到特定的调节区域。

[0115] 31.根据实施方式30所述的方法,其中所述调节区域是FANTOM5增强子、

[0116] CpG岛、CpG海岸、CpG架子或前述的任何组合。

[0117] 32.根据实施方式1至31中任一项所述的方法,其中步骤(f)和(g)通过计算机处理器执行。

[0118] 33.一种检测来自癌细胞的DNA的存在并识别癌症亚型的方法,所述方法

[0119] 包括:

[0120] a.从受试者样品接收无细胞甲基化DNA的测序数据;

[0121] b.比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个

[0122] 体的对照无细胞甲基化DNA序列;

[0123] c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则

[0124] 识别来自癌细胞的DNA的存在;以及

[0125] d.如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

[0126] 34.一种检测来自癌细胞的DNA的存在并识别癌症亚型的计算机实现的方法,所述方法包括:

[0127] a.至少一个处理器从受试者样品接收无细胞甲基化DNA的测序

[0128] 数据;

[0129] b.在所述至少一个处理器处,比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个体的对照无细胞甲基化DNA序列;

[0130] c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则在所述至少一个处理器处识别来自癌细胞的DNA的存在;并且如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

[0131] 35.一种与具有处理器和连接到所述处理器的存储器的通用计算机结合使用的计算机程序产品,所述计算机程序产品包括其上编码有计算机机制的计算机可读存储介质,其中可以将所述计算机程序机制加载到所述计算机的存储器中,并使所述计算机执行实施方式34所述的方法。

[0132] 36.一种计算机可读介质,其上存储有用于存储实施方式35所述的计算机程序产

品的数据结构。

[0133] 37.一种用于检测来自癌细胞的DNA的存在并识别癌症亚型的设备,所述

[0134] 设备包括:

[0135] 至少一个处理器;以及

[0136] 与所述至少一个处理器通信的电子存储器,所述电子存储器存储处理器可执行代码,所述代码在所述至少一个处理器处执行时,使所述至

[0137] 少一个处理器:

[0138] a.从受试者样品接收无细胞甲基化DNA的测序数据;

[0139] b.比较所捕获的无细胞甲基化DNA的序列与来自健康和癌症个

[0140] 体的对照无细胞甲基化DNA序列;

[0141] c.如果所捕获的无细胞甲基化DNA与来自癌症个体的无细胞甲基化DNA序列的一个或多个序列之间在统计学上具有显著相似性,则识别来自癌细胞的DNA的存在,并且如果在步骤c中识别出来自癌细胞的DNA,则基于步骤b中的所述比较进一步识别所述起源的癌细胞组织和癌症亚型。

[0142] 38.一种检测来自癌细胞的DNA的存在并确定来自两个或多个可能器官的所述癌细胞所处的癌症的位置的方法,所述方法包括:

[0143] (a)提供来自受试者的无细胞DNA的样品;

[0144] (b)使用对甲基化多核苷酸具有选择性的结合剂从所述样品中

[0145] 捕获无细胞甲基化DNA;

[0146] (c)对所捕获的无细胞甲基化DNA进行测序;

[0147] (d)比较所捕获的无细胞甲基化DNA的序列模式与对照个体的两个或多个群体的DNA序列模式,所述两个或多个群体中的每一个在

[0148] 不同器官中具有局限性癌症;

[0149] (e)基于所述无细胞DNA的甲基化模式与所述两个或多个群体中一个之间的统计学上显著的相似性,确定所述癌细胞来自哪个器官。

[0150] 实施例

[0151] 方法与材料

[0152] 供体招募和样品采集

[0153] CRC、乳腺癌和GBM样品是从大学健康网络生物库获得的;AML样品是从大学健康网络白血病生物库获得的;最后,通过加拿大多伦多西奈山医院(MSH)的家庭医学中心招募健康对照。在获得患者同意的情况下收集的所有样品均获得了研究伦理委员会、大学健康网络和加拿大多伦多西奈山医院的机构批准。

[0154] 试样处理-cfDNA

[0155] EDTA和ACD血浆样品是从生物库和加拿大多伦多西奈山医院(MSH)的家庭医学中心获得的。所有样品都保存在-80℃或气相液氮中直至使用。使用QIAamp循环核酸试剂盒(Qiagen)从0.5-3.5ml血浆中提取无细胞DNA。使用前,通过Qubit对提取的DNA进行定量。

[0156] 试样处理-PDX cfDNA

[0157] 经大学健康网络研究伦理委员会批准,在患者同意下从大学健康网络生物库获得的人结肠直肠肿瘤组织使用胶原酶A消化成单细胞。将单细胞皮下注射到4-6周龄的NOD/

SCID雄性小鼠中。在通过心脏穿刺采血之前,通过吸入二氧化碳使小鼠安乐死,并将其存储在EDTA管中。从收集的血样中分离血浆并存储在-80℃下。使用QIAamp循环核酸试剂盒(Qiagen)从0.3-0.7ml血浆中提取无细胞DNA。所有动物工作均按照大学健康网络动物保护委员会批准的道德规定进行。

[0158] cfMeDIP-seq

[0159] cfMeDIP-seq方案的示意图显示在WO2017/190215中。在cfMeDIP之前,使用Kapa Hyper Prep试剂盒(Kapa Biosystems)对DNA样品进行文库制备。遵循制造商方案进行了一些修改。简而言之,将感兴趣的DNA添加到0.2mL PCR管中,并进行末端修复和A-加尾。使用NEBNext衔接子(来自New England Biolabs的NEBNext Multiplex Oligos for Illumina试剂盒)进行衔接子连接,最终浓度为0.181μM,在20℃孵育20分钟并用AMPure XP珠子纯化。使用USER酶(加拿大New England Biolabs)消化洗脱的文库,然后在MeDIP之前使用Qiagen MinElute PCR纯化试剂盒进行纯化。

[0160] 将制备的文库与合并的甲基化/未甲基化λPCR产物合并,最终DNA量为100ng,并在进行一些修改后使用Taiwo等人2012[7]的方案进行MeDIP。简而言之,对于MeDIP,在进行一些修饰后按照制造商的方案使用Diagenode MagMeDIP试剂盒(编号C02010021)。在将0.3ng对照甲基化和0.3ng对照未甲基化拟南芥DNA,填充DNA(使DNA[cfDNA+填充剂+对照]的总量达到100ng)和缓冲液添加到含有衔接子连接的DNA的PCR管中之后,将样品加热到95℃,持续10分钟,然后立即放入冰水浴中10分钟。将每个样品分成两个0.2mL PCR管:一个用于10%输入对照,另一个用于要进行免疫沉淀的样品。将MagMeDIP试剂盒中包含的5-mC单克隆抗体33D3(编号C15200081)进行1:15稀释,然后生成稀释的抗体混合物,并添加到样品中。还添加洗涤过的磁珠(按照制造商的说明),然后在4℃孵育17小时。使用Diagenode iPure试剂盒纯化样品,并用50μl缓冲液C洗脱。通过qPCR验证了反应的成功(QC1),以检测加入的拟南芥DNA的存在,确保未甲基化加入的DNA的回收百分比<1%,反应的特异性百分比>99%(根据1-[加入的未甲基化对照DNA的回收率超过加入的甲基化对照DNA的回收率]计算),然后进行下一步。通过使用qPCR确定扩增每个文库的最佳循环数,然后使用KAPA HiFi Hotstart Mastermix扩增样品,并添加NEBNext多重寡核苷酸至终浓度0.3μM。用于扩增文库的PCR设置如下:在95℃下活化3分钟,然后为98℃进行20秒,65℃进行15秒和72℃进行30秒的预定循环,并在72℃下进行1分钟的最终延伸。使用MinElute PCR纯化柱纯化扩增的文库,然后用3% Nusieve GTG琼脂糖凝胶进行凝胶大小选择,以除去任何衔接子二聚体。在提交测序之前,确定从被剪切以模拟无细胞DNA的HCT116细胞系DNA(细胞系获自ATCC,无支原体)生成的MeDIP-seq和cfMeDIP-seq文库的甲基化人DNA区(睾丸特异性H2B,TSH2B)和未甲基化人DNA区(GAPDH启动子)的富集倍数。将最终的文库提交用以BioAnalyzer分析,然后在UHN玛格丽特公主基因组中心(UHN Princess Margaret Genomic Centre)通过Illumina HiSeq 2000进行测序。

[0161] 用于点突变检测的超深度靶向测序

[0162] 我们使用QIAgen循环核酸试剂盒以从约20mL血浆(4-5x 10mL EDTA血液管)中分离出无细胞DNA,这些血浆来自具有匹配肿瘤组织分子概况分析数据的患者,该数据在入组早期临床试验前在玛格丽特公主癌症中心产生。使用PureGene Gentra试剂盒从细胞系(CRC和MM细胞系稀释液)中提取DNA,使用Covaris超声仪将其片段化至约180bp,并使用

Ampure珠排除较大的片段以模拟无细胞DNA的片段大小。利用NEXTflex-96 DNA条形码衔接子(Bio Scientific,Austin,TX)衔接子,使用KAPA Hyper Prep试剂盒(Kapa Biosystems,Wilmington,MA)由83ng片段化DNA构建DNA测序文库。为了分离出包含已知突变的DNA片段,我们设计了生物素化DNA捕获探针(xGen Lockdown Custom Probes Mini Pool,Integrated DNA Technologies,Coralville,IA),其靶向由临床实验室使用Illumina TruSeq扩增子癌症组测试的48个基因的突变热点。将条形码库合并,然后按照制造商的说明(IDT xGEN Lockdown方案2.1版)应用自定义混合捕获库。使用Illumina HiSeq 2000仪器将这些片段测序到>10,000X读取覆盖率。使用bwa-mem将所得读取对齐,并使用samtools和muTect 1.1.4版检测突变。

[0163] 肿瘤特异性特征数量与通过测序深度检测的概率之间的关系的建模

[0164] 我们创建了145,000个模拟基因组,其中癌症特异性甲基化DMR的比例分别设置为0.001%、0.01%、0.1%、1%和10%并由1、10、100、1000和10000个独立DMR组成。我们从这些原始混合物中取样了14,500个二倍体基因组(代表100ng DNA),并进一步对每个基因座取样了10、100、1000和10000个读数,以表示在这些深度的测序覆盖率。对于覆盖率、丰度和特征数量的每种组合,重复此过程100次。我们估计了每种参数组合的至少1个DMR的成功检测频率,并绘制了概率曲线(图1A),以可视方式评估特征数量对取决于测序深度的成功检测概率的影响。

[0165] 组织区别特征的推导,多组织分类器的开发以及450k数据的验证

[0166] 使用MEDIPS R软件包[8]对cfDNAMeDIP图谱进行了定量,将其转换为RPKM,然后转换为log2每百万计数。随后,使用limma-trend[9]将线性模型拟合在映射到FANTOM5增强子、CpG岛、CpG海岸和CpG架子的特征矩阵上,其中回收的spike-in甲基化DNA的百分比作为协变量包括在内,以控制下拉效率变化。对每对组织类型评估成对对比,并选择前150个和后150个DMR进行弹性网分类器训练并验证癌症类型的特异性。性能指标是通过交叉验证中具有最高Kappa值的模型的折外呼叫的多数票投票得出的,Chakravarthy等人[10]曾采用过这种启发式方法。

[0167] 机器学习分析用于评估分类准确性

[0168] 对发现队列的模型训练和评估

[0169] 为了在没有高计算成本的情况下评估cfMeDIP数据在肿瘤分类中的性能,我们将可能的候选特征的初始集合减少到包含CpG岛、CpG海岸、CpG架子和FANTOM5增强子(在此标记为“调节特征”)的窗口,从而产生196个样品和505,027个特征的矩阵。然后,我们使用插入符号R软件包以80%-20%的方式将发现队列数据划分为50个独立训练和测试集合(图2A)。进行拆分,同时保持发现队列中的类别比例。然后,我们相对于其他类别对每个类别使用limma-trend,通过训练数据分区上的中度t统计量(150个高甲基化,150个低甲基化)选择了前300个DMR。然后使用这些DMR(最多300个DMR x 7个其他类别=2100个特征)训练了二项式GLMnet,并使用10倍交叉验证(CV)的3次迭代来优化混合参数(α,值=0、0.2、0.5、0.8和1)和惩罚(λ,值=0-0.05,以0.01为增量)的值,以Cohen的Kappa作为性能指标。对于每个训练集,产生了6个一类对其他类二项式分类器的集合。

[0170] 然后,我们使用AUROC(接受者工作特性曲线下的区域)估算对保留测试集的分类性能。这些估计值代表了分类的无偏度量,因为保留测试集样品未用于DMR预选或GLMnet训

练和调整。50个独立训练和测试集还由于训练集偏差而允许将乐观估计最小化。

[0171] 对验证队列的模型评估

[0172] 对于每个验证队列cfMeDIP样品,我们估计了在发现队列内的50个不同训练集上训练的AML、LUC和正常的一对多二项式分类器的分类概率。将来自50个模型的概率取平均值,以产生单个得分,然后将其用于AUROC估计。我们还将早期(I和II期)或晚期LUC样品(III和IV期)留给了一对多分类器,通过估计AUROC来评估疾病阶段是否影响性能。

[0173] 结果和讨论

[0174] 我们用生物信息学方法模拟了从0.001％到10％的ctDNA的不同比例的混合物(图1A,列面)。我们还模拟了ctDNA与正常cfDNA相比具有1、10、100、1000或10000个DMR(差异性甲基化区域)的情况(图1A,行面)。然后在每个基因座上以不同的测序深度(10X、100X、1000X和10000X)对读数进行采样(图1A,x轴)。我们发现,随着DMR数量的增加,检测到至少1个癌症特异性事件的概率越来越高(图1A),甚至在低丰度的癌ctDNA且浅覆盖率下也如此。

[0175] 此外,来自癌症基因组图谱(TCGA)的全癌数据显示在几乎所有肿瘤类型中,肿瘤与正常组织之间存在大量DMR[11]。因此,这些发现强调了,一种成功地从ctDNA中回收癌症特异性DNA甲基化改变的测定法可以作为一种非常灵敏的工具,来以低测序相关成本检测、分类和监测恶性疾病。

[0176] 然而,由于循环中DNA的数量且片段化非常少,因此血浆cfDNA中DNA甲基化的全基因组映射具有挑战性[12]。因此,以前对cfDNA进行甲基化图谱分析的努力主要限于基于基因座特异性PCR的测定[2,3],例如FDA批准的用于结肠直肠癌筛查的SEPT9甲基化测定[13]。尽管最近已经对片段化cfDNA进行了全基因组亚硫酸氢盐测序[14-16],但是CpGs在全基因组内的低丰度可能会减少可从测序获得的有用的甲基化相关信息量。因此,WGBS对血浆DNA的主要问题是成本高,效率低以及与亚硫酸氢盐转化相关的DNA损失。另一方面,选择性富集易于甲基化的富含CpG的特征的方法可能会使每次读取可用的有用信息量最大化,降低成本并减少DNA损失。

[0177] 适用于cfDNA甲基化映射的全基因组方法

[0178] 我们开发了一种称为cfMeDIP-seq(无细胞甲基化DNA免疫沉淀和高通量测序)的新方法,以使用无细胞DNA进行全基因组DNA甲基化映射。文中所描述的cfMeDIP-seq方法是通过修改现有的低输入MeDIP-seq方案[7]而开发的,根据我们的经验,该方案在低至100ng输入DNA的情况下仍非常稳健。但是,大多数血浆样品产生的DNA远远少于100ng。为了克服这一挑战,我们将外源λDNA(填充DNA)添加到了衔接子连接的cfDNA文库中,以人为地将起始DNA的量增加到100ng。这将抗体的非特异性结合量最小化,并且还由于与塑料制品结合而将DNA损失量最小化。填充DNA由大小与衔接子连接的cfDNA文库相似的扩增子组成,并由不同CpG密度的未甲基化和体外甲基化DNA组成。这种填充DNA的添加也有实际用途,因为不同的患者会产生不同量的cfDNA,从而可以将输入DNA量归一化为100ng。这样确保所有样品的下游方案均保持完全相同,而与可用cfDNA的量无关。

[0179] 我们首先使用来自被剪切成类似于cfDNA中观察到的片段大小的人结肠直肠癌细胞系HCT116的DNA验证了cfMeDIP-seq方案。选择HCT116是因为可获得公共DNA甲基化数据。我们同时使用100ng剪切的细胞系DNA进行黄金标准MeDIP-seq方案[7],并使用10ng、5ng和1ng相同的剪切细胞系DNA进行cfMeDIP-seq方案。这是在两个生物学重复中进行的。在所有

条件下,我们都获得了超过99%的反应特异性(1-[加入的未甲基化对照DNA的回收率超过加入的甲基化对照DNA的回收率]),以及已知的甲基化区域相对于非甲基化区域的非常大量的富集(分别为TSH2B0和GAPDH)(图6B)。

[0180] 以每个文库约30至7,000万个读取(补充表1)将该文库测序至饱和(图6A)。将原始读取与人类基因组和λ基因组比对,发现与λ基因组几乎没有比对(补充表1)。因此,添加外源λDNA作为填充DNA不会干扰测序数据的产生。最后,我们计算CpG富集得分作为免疫沉淀步骤的质量控制量度[8]。所有文库都显示出类似的CpGs富集,而输入对照如预期的那样没有富集(图6C),即使在极低的输入(1ng)下也能验证我们的免疫沉淀。

[0181] 比较不同输入DNA水平的全基因组相关性估计表明,MeDIP-seq(100ng)和cfMeDIP-seq(10、5和1ng)方法都非常稳健,任何两个生物学重复之间的Pearson相关性为至少0.94(图1B)。分析还表明,5和10ng输入DNA的cfMeDIP-seq可以稳健地概括通过100ng的传统MeDIP-seq获得的甲基化图谱(Pairwise Pearson相关性为至少0.9)(图1B)。与100ng的MeDIP-seq相比,1ng输入DNA的cfMeDIP-seq的性能有所降低,但仍表现出>0.7的强Pearson相关性(图1B)。我们还观察到,cfMeDIP-seq方案使用黄金标准RRBS(简化代表性亚硫酸氢盐测序)和WGBS(全基因组亚硫酸氢盐测序)概括了HCT116的DNA甲基化图谱(图1C)。总而言之,我们的数据表明,cfMeDIP-seq是用于片段化和低输入DNA物质(例如循环cfDNA)的全基因组甲基化映射的可靠方案。

[0182] cfMeDIP-seq对检测肿瘤来源的ctDNA显示出高灵敏度

[0183] 为了评估cfMeDIP-seq方案的灵敏度,我们将结肠直肠癌(CRC)HCT116细胞系DNA系列稀释到多发性骨髓瘤(MM)MM1.S细胞系DNA中,二者均被剪切为模拟cfDNA大小。我们将CRC DNA从100%、10%、1%、0.1%、0.01%、0.001%稀释到0%,并对每种稀释液进行cfMeDIP-seq。我们还执行了超深(10,000X中值覆盖率)靶向测序,以检测同一样品中的三点突变。使用5%错误发现率(FDR)阈值观察到的在每个CRC稀释点相对于纯MM DNA识别到的DMR数量与基于降至0.001%稀释度的稀释因子的DMR预期数量几乎是完全线性的(r^2=0.99,p<0.0001)(图1D)。此外,这些DMR中的DNA甲基化信号在观察到的信号与预期信号之间也显示出几乎完美的线性关系(r^2=0.99,p<0.0001)(图1E;补充表2B)。相比之下,超出1%稀释度,由于PCR或测序错误,超深层靶向测序无法可靠地区分CRC特异性变异和伪变体(图1F;补充表2A)。因此,cfMeDIP-seq对癌症来源的DNA的检测显示出出色的灵敏度,超过了超深度靶向测序使用标准方案检测变体的性能。

[0184] 癌症DNA经常在富含CpG的区域超甲基化[17]。由于cfMeDIP-seq特异性靶向甲基化的富含CpG的序列,因此我们假设ctDNA在免疫沉淀过程中会被优先富集。为了测试这一点,我们从两名结肠直肠癌患者中产生了患者源性异种移植物(PDX),并收集了小鼠血浆。肿瘤来源的人cfDNA在输入样品中以低于1%频率存在于总cfDNA池中,并且在免疫沉淀后的丰度提高了2倍(图1G;补充表3)。这些结果表明,通过对ctDNA进行偏向测序,cfMeDIP程序可以进一步提高ctDNA检测灵敏度。

[0185] 循环血浆cfDNA甲基化图谱可区分多种癌症类型和健康供体

[0186] DNA甲基化模式具有组织特异性,已用于将癌症患者分为胶质母细胞瘤[18]、室管膜瘤[6]、结肠直肠癌[19]和乳腺癌[20,21]等许多癌症类型的临床相关疾病亚组。我们询问是否可以使用cfDNA相关图谱来识别多种肿瘤类型的起源组织。为此,我们分析了来自5

种不同肿瘤类型的196个样品以及来自早期和晚期肿瘤的正常对照。我们使用线性建模来识别每个成对比较中映射到CpG海岸、CpG架子、CpG岛和FANTOM5增强子的前300个DMR,总共产生了2,100个独特DMR(图2A)。基于这些特征的甲基化状态,对196个血浆样品进行的基于t分布随机邻域嵌入(tSNE)的密度聚类[22]显示了样品基于起源组织和肿瘤类型的不同聚类(图2B,图2C)。使用适合这些特征的弹性网多癌症分类器(图2A),我们观察到了不同肿瘤类型之间的高度准确的区分(图2D)。

[0187] 疾病亚型的区分

[0188] 我们评估了cfDNA MeDIP图谱在五种不同情况下区分疾病亚型的能力-基因表达模式(乳腺癌中的ER状态)、拷贝数畸变(乳腺癌中的HER2状态)、重排(AML中的FLT3 ITD状态)、点突变(GBM中的IDH突变)、最后是肺癌的组织学。在每种情况下,如前所述,将线性模型用于选择特征并对其进行排名。在每种情况下,使用层次聚类来评估样品的分组。基于所选特征的甲基化状态的基于t分布随机邻域嵌入(tSNE)的密度聚类[22]显示,基于这五个癌症亚型分类的不同实例中的每一个对样品进行不同聚类。

[0189] 使用机器学习检测癌症并分类癌症类型

[0190] 为了严格评估cfMeDIP图谱检测癌症并进一步分类癌症类型的能力,我们随后对发现队列进行了一组机器学习分析。为了加快计算分析,我们最初将我们的cfMeDIP发现队列缩减为映射到CpG岛、CpG海岸、CpG架子和FANTOM5增强子的特征(n=505,027个窗口)。然后,我们对我们的发现队列样品实施了一项策略,以得出对性能的无偏估计,同时考虑了训练集偏向。

[0191] 在文中,我们将发现队列分为均衡的训练和测试集(训练集为80%,测试集为20%)。仅使用训练集中的样品,我们根据limma-trend测试统计数据,对每个类别(样品类型)相对于其他类别选择前300个DMR,并使用这些特征对训练集数据训练一系列一类对其他类别的GLMnet。训练程序由3轮对α和λ值网格的10倍交叉验证(CV)组成,并对Cohen的Kappa进行了优化。希望利用更多随机化来进行更通用的模型调整,从而促使使用多轮10倍CV。

[0192] 然后使用从测试集样品(在DMR选择和随后的GLMnet训练/调整步骤中保留)得出的AUROC(接收者工作特性曲线下的区域)评估性能。通过将发现队列50次分为不同训练集和测试集来重复进行此过程,以减轻训练集偏差的影响。对于每个类别对其他类别的比较,最终收集了50个模型(总共480个模型)。因此,我们将此模型集合称为E50。

[0193] 随后,我们通过生成另外152个血浆样品的验证队列来评估各批次的性能:AML(n=35),肺癌(n=55)和健康对照(n=62)样品。对于每个类别,我们在E50中对模型输出的类别概率取平均值,并估计一个类别相对于所有其他类别的AUROC(图3A)。对于AML相对于其他(0.993),LUC相对于其他(0.943)和正常相对于其他(1.000)的分类,分类器显示出高AUROC值。这进一步证实了cfMeDIP-seq与机器学习方法相结合以准确地检测和分类肿瘤类型的能力。最后,我们观察到,分类器在早期样品(0.950)中与在晚期样品(0.934)中一样准确(图3B),表明该方法适用于癌症早期检测以及早期和晚期癌症的检测。

[0194] 使用cfMeDIP-seq进行cfDNA甲基化组图谱分析的其他优点

[0195] cfDNA甲基化模式准确代表起源组织的能力也克服了基于突变的测定法的局限性,其中由于不同组织中许多癌症的潜在驱动基因突变的复发性质,因此对起源组织的特

异性可能较低[23]。基于突变的测定法也可能由于肿瘤的克隆结构而变得不灵敏,在这种情况下,由于在ctDNA中的丰度较低,因此亚克隆驱动子可能更难检测到[24]。基于突变的ctDNA方法也容易受到良性组织中驱动子突变的潜在混淆的影响,这已被观察到[25],并有文献证明阳性选择的证据[26]。

[0196] 综上所述,我们发现-基于迄今获得的最大的癌症cfDNA甲基化组-将cfMeDIP-seq建立为一种有效且具有成本效益的工具,其具有影响癌症管理和早期检测的潜力。cfMeDIP-seq的准确性和多功能性可用于在耐药性与表观遗传学改变相关的环境中(例如前列腺癌中对雄激素受体抑制的敏感性)提供治疗决策[27]。早期诊断和筛查的潜在机会在肺癌中尤为明显,在肺癌中,这种筛查已经显示出临床实用性,但是对于现有的筛查测试(即低剂量CT扫描)存在明显的局限性,例如电离辐射暴露和高假阳性率。

[0197] 总之,我们的研究结果强调了cfDNA甲基化图谱的实用性,将其作为非侵入性、具有成本效益、灵敏、高度准确的早期肿瘤检测、多癌分类和癌症亚型分类的基础。

[0198] 尽管文中已经描述了本发明的优选实施例,但是本领域技术人员将理解,在不脱离本发明的精神或所附权利要求的范围的情况下,可以对其进行变化。本文公开的所有文件,包括以下参考文献列表中的那些文件,均通过引用并入本文。

[0199] 表1:使用被剪切以模拟cfDNA的HCT116细胞系DNA的各种起始输入制备的测序MeDIP-seq(100ng Rep 1和Rep 2)和cfMeDIP-seq(10ng、5ng和1ng,Rep 1和Rep 2)文库与人(Hg19)基因组和λ基因组的读取次数和映射效率。两个生物学重复用于起始输入DNA。对于小于100ng的起始输入,在MeDIP之前向样品加入外源λDNA以将起始量人工地增加至100ng。

[0200]

样品	原始读取数	与人类基因组(Hg19)的比对读取数	映射到人类基因组(Hg19)的效率	与λ基因组的比对读取数	映射到λ基因组的效率
输入	74,504,053	71,343,168	95.76	12	0.00
100 ng 重复 1	55,396,238	50,472,273	91.11	0	0.00
100 ng 重复 2	66,569,209	60,770,277	91.29	1	0.00
10 ng 重复 1	70,054,607	64,020,441	91.39	0	0.00
10 ng 重复 2	58,297,539	53,308,777	91.44	0	0.00
5 ng 重复 1	65,845,430	60,540,743	91.94	1	0.00
5 ng 重复 2	64,750,879	59,358,412	91.67	0	0.00
1 ng 重复 1	35,102,361	32,258,451	91.90	0	0.00
1 ng 重复 2	33,881,118	31,194,711	92.07	0	0.00

[0201] 表2A:使用CRC细胞系HCT116 DNA在MM细胞系MM1.SDNA中的系列稀释的超深度靶向变体测序的平均覆盖率

[0202]

稀释度(%CRC DNA)	未折叠的读数平均目标覆盖率	SSCS（单链共有序列）平均目标覆盖率	DCS（双重共有序列）平均目标覆盖率
100	155,964	4284	655
10	154,657	4877	654
1	154,419	4890	654
0.1	183,271	5674	887
0.01	238,291	8068	1602
0.001	199,766	7337	1299
0.0001	187,695	6891	1181
0	216,434	7721	1412

[0203] 表2B:CRC细胞系HCT116 DNA在MM细胞系MM1.SDNA中的系列稀释所观察到的DMR和DNA甲基化信号

[0204]

稀释度(%CRC DNA)	所观察到的 DMR 数	所观察到的 DNA 甲基化信号(DMR 内的总 RPKM)
100	111,472	645,683.90
10	1,597	8,775.61
1	692	4,521.60
0.1	12	75.71
0.01	8	79.73
0.001	2	22.42

[0205] 表3:PDX和输入对照样品的cfMeDIP-seq文库在与人类(Hg19)基因组比对后的读取数和映射效率

[0206]

样品	原始读取数的#	与人类基因组（Hg19）的比对读取数的#	映射到人类基因组（Hg19）的效率
输入对照 1	46,857,633	389,073	0.83
输入对照 2	35,658,454	283,799	0.80
PDX 1	49,997,949	1,080,277	2.16
PDX 2	34,802,767	614,988	1.77

[0207] 参考文献

[0208] 1.Diaz,L.A.,Jr.and A.Bardelli,Liquid biopsies:genotyping circulating tumor DNA.J ClinOncol,2014.32(6):p.579-86.

[0209] 2.Lehmann-Werman,R.,et al.,Identification of tissue-specific cell death using methylationpatterns of circulating DNA.Proc Natl Acad Sci U S A, 2016.113(13):p.E1826-34.3.Visvanathan,K.,et al.,Monitoring of Serum DNA Methylation as an Early IndependentMarker of Response and Survival in Metastatic Breast Cancer:TBCRC 005ProspectiveBiomarker Study.J Clin Oncol, 2016:p.JCO2015662080.

[0210] 4.Newman,A.M.,et al.,An ultrasensitive method for quantitating circulating tumor DNAwith broad patient coverage.Nat Med,2014.20(5):p.548-54.

[0211] 5.Aravanis,A.M.,M.Lee,and R.D.Klausner,Next-Generation Sequencing of CirculatingTumor DNA for Early Cancer Detection.Cell,2017.168(4):p.571-574.

[0212] 6.Mack,S.C.,et al.,Epigenomic alterations define lethal CIMP-positive

ependymomas ofinfancy.Nature,2014.506(7489):p.445-50.

[0213] 7.Taiwo,O.,et al.,Methylome analysis using MeDIP-seq with low DNA concentrations.NatProtoc,2012.7(4):p.617-36.

[0214] 8.Lienhard,M.,et al.,MEDIPS:genome-wide differential coverage analysis of sequencingdata derived from DNA enrichment experiments.Bioinformatics,2014.30(2):p.284-6.9.Law,C.W.,et al.,voom: Precision weights unlock linear model analysis tools for RNA-

[0215] seq read counts.Genome Biol,2014.15(2):p.R29.

[0216] 10.Chakravarthy,A.,et al.,Human Papillomavirus Drives Tumor Development Throughoutthe Head and Neck:Improved Prognosis Is Associated With an Immune ResponseLargely Restricted to the Oropharynx.J Clin Oncol,2016.34 (34):p.4132-4141.11.Hoadley,K.A.,et al.,Multiplatform analysis of 12cancer types reveals molecularclassification within and across tissues of origin.Cell,2014.158(4):p.929-44.

[0217] 12.Fleischhacker,M.and B.Schmidt,Circulating nucleic acids(CNAs)and cancer--a survey.

[0218] Biochim Biophys Acta,2007.1775(1):p.181-232.

[0219] 13.Potter,N.T.,et al.,Validation of a real-time PCR-based qualitative assay for the detectionof methylated SEPT9 DNA in human plasma.Clin Chem, 2014.60(9):p.1183-91.14.Legendre,C.,et al.,Whole-genome bisulfite sequencing of cell-free DNA identifiessignature associated with metastatic breast cancer.Clin Epigenetics,2015.7:p.100.15.Sun,K.,et al.,Plasma DNA tissue mapping by genome-wide methylation sequencing fornoninvasive prenatal,cancer, and transplantation assessments.Proc Natl Acad Sci U SA,2015.112(40):p.E5503-12.

[0220] 16.Chan,K.C.,et al.,Noninvasive detection of cancer-associated genome-widehypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.ProcNatl Acad Sci U S A,2013.110(47):p.18761-8.

[0221] 17.Sharma,S.,T.K.Kelly,and P.A.Jones,Epigenetics in cancer.Carcinogenesis,2010.

[0222] 31(1):p.27-36.

[0223] 18.Sturm,D.,et al.,Hotspot mutations in H3F3A and IDH1 define distinct epigenetic andbiological subgroups of glioblastoma.Cancer Cell, 2012.22(4):p.425-37.

[0224] 19.Hinoue,T.,et al.,Genome-scale analysis of aberrant DNA methylation in colorectalcancer.Genome Res,2012.22(2):p.271-82.

[0225] 20.Stirzaker,C.,et al.,Methylome sequencing in triple-negative breast cancer revealsdistinct methylation clusters with prognostic value.Nat Commun, 2015.6:p.5899.21.Fang,F.,et al.,Breast cancer methylomes establish an

epigenomic foundation formetastasis.Sci Transl Med,2011.3(75):p.75ra25.

[0226] 22.Laurens van der Maaten,G.H.,Visualizing Data using t-SNE.Journal of MachineLearning Research,2008.9:p.2579-2605.

[0227] 23.Kandoth,C.,et al.,Mutational landscape and significance across 12 major cancer types.

[0228] Nature,2013.502(7471):p.333-9.

[0229] 24.McGranahan,N.,et al.,Clonal status of actionable driver events and the timing ofmutational processes in cancer evolution.Sci Transl Med,2015.7(283):p.283ra54.25.Zauber,P.,S.Marotta,and M.Sabbath-Solitare,KRAS gene mutations are morecommon in colorectal villous adenomas and in situ carcinomas than in carcinomas.Int JMol Epidemiol Genet,2013.4(1):p.1-10.

[0230] 26.Martincorena,I.,et al.,Tumor evolution.High burden and pervasive positive selection ofsomatic mutations in normal human skin.Science,2015.348(6237):p.880-6.

[0231] 27.Beltran,H.,et al.,Divergent clonal evolution of castration-resistant neuroendocrineprostate cancer.2016.22(3):p.298-305。

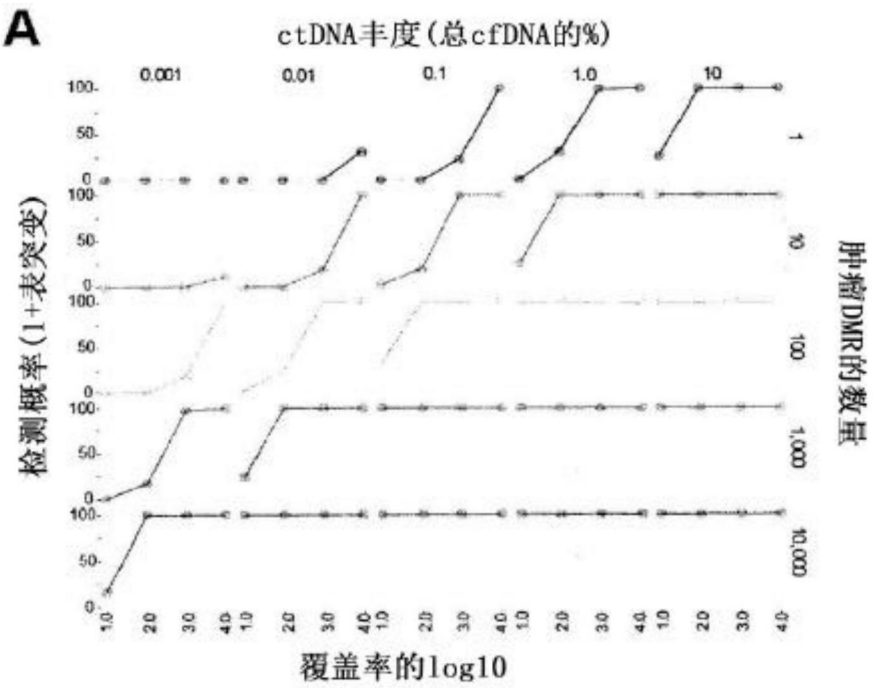

A
ctDNA丰度(总cfDNA的%)
0.001
0.01
0.1
1.0
10
检测概率(1+表突变)
肿瘤DMR的数量
覆盖率的log10

图1A

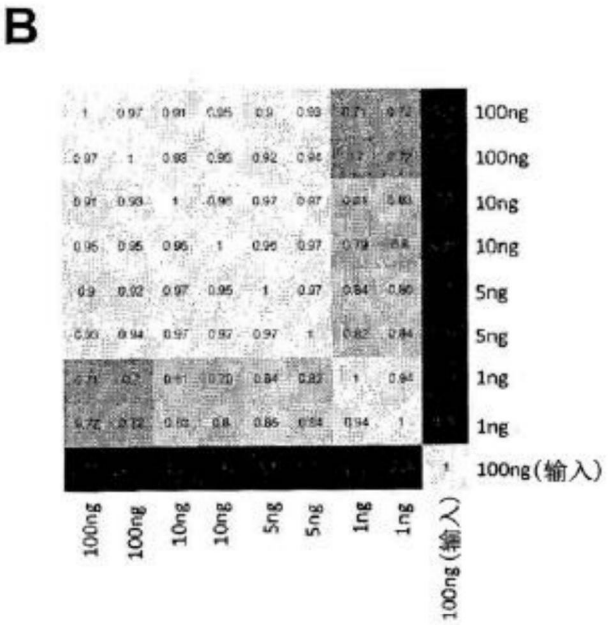
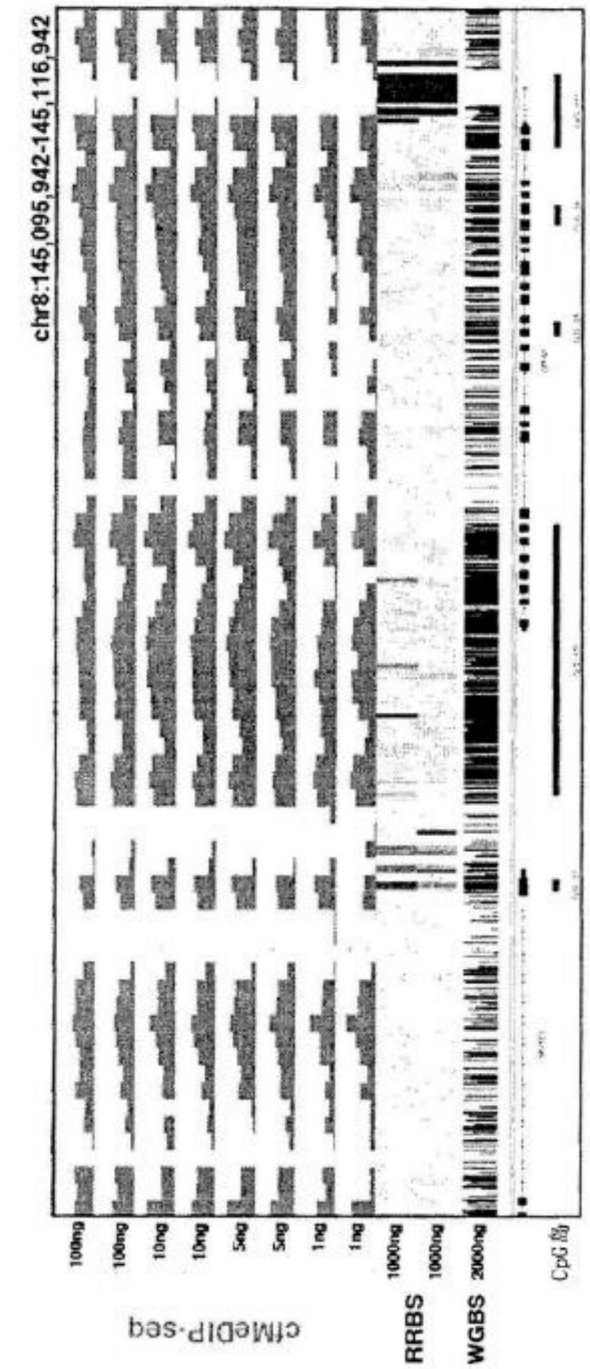

B
100ng
100ng
10ng
10ng
5ng
5ng
1ng
1ng
100ng(输入)
100ng(输入)

图1B


chr8:145,095,942-145,116,942
100ng
100ng
10ng
10ng
5ng
5ng
1ng
1ng
1000ng
1000ng
2000ng
cfMeDIP-seq
RRBS
WGBS
CpG 岛

C

图1C

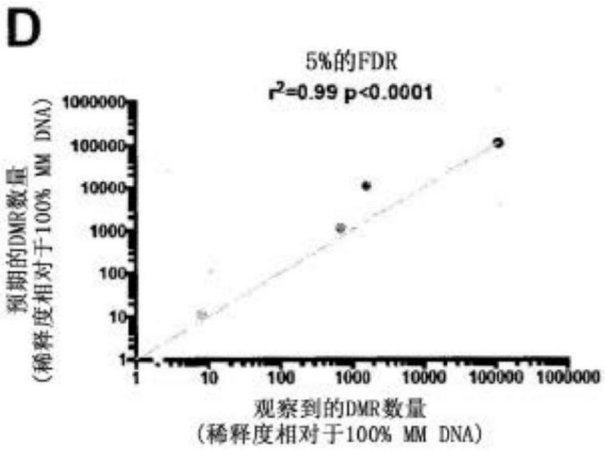

D
5%的FDR
r²=0.99 p<0.0001
预期的DMR数量
(稀释度相对于100% MM DNA)
观察到的DMR数量
(稀释度相对于100% MM DNA)

图1D

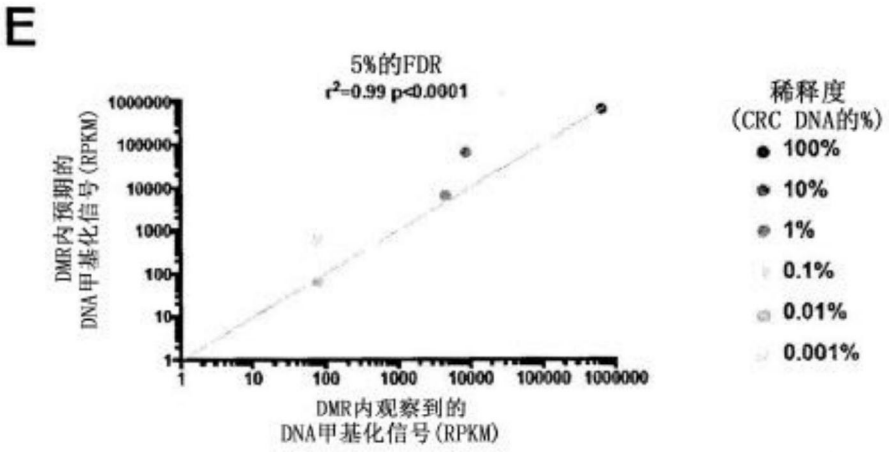

E
5%的FDR
r²=0.99 p<0.0001
DMR内预期的
DNA甲基化信号(RPKM)
DMR内观察到的
DNA甲基化信号(RPKM)
稀释度
(CRC DNA的%)
100%
10%
1%
0.1%
0.01%
0.001%

图1E

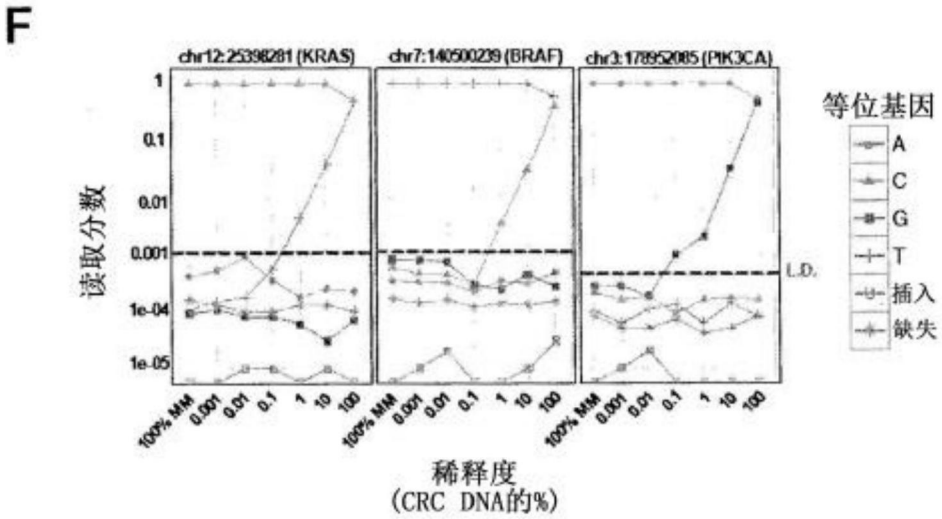

F
chr12:25398281(KRAS)
chr7:140500239(BRAF)
chr3:178952085(PIK3CA)
读取分数
L.D.
等位基因
A
C
G
T
插入
缺失
稀释度
(CRC DNA的%)

图1F

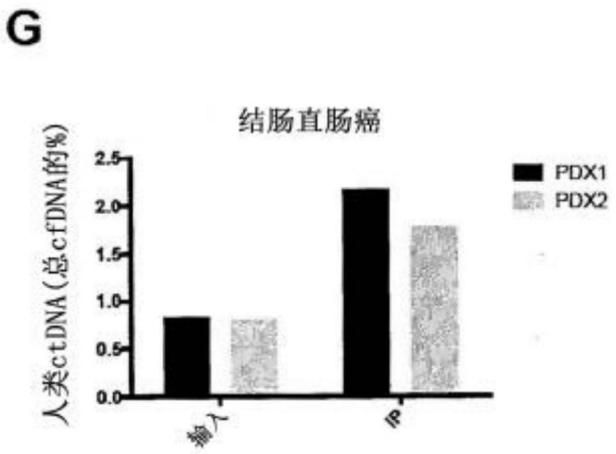


图1G

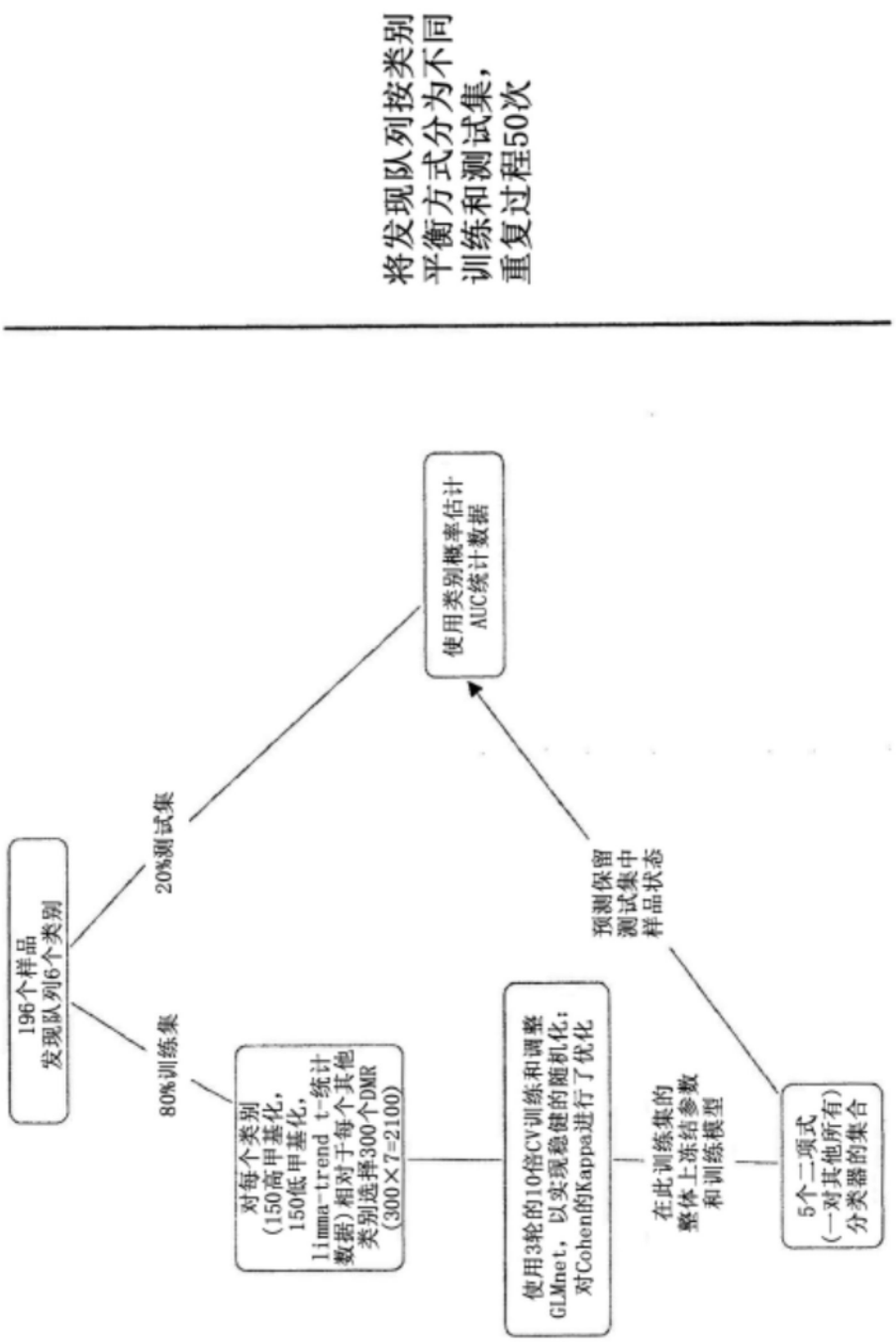

196个样品
发现队列6个类别
20%测试集
80%训练集
使用类别概率估计
AUC统计数据
预测保留
测试集中
样品状态
对每个类别
(150高甲基化,
150低甲基化,
limma-trend t-统计
数据)相对于每个其他
类别选择300个DMR
(300×7=2100)
使用3轮的10倍CV训练和调整
GLMnet，以实现稳健的随机化;
对Cohen的Kappa进行了优化
在此训练集的
整体上冻结参数
和训练模型
5个二项式
(一对其他所有)
分类器的集合
将发现队列按类别
平衡方式分为不同
训练和测试集,
重复过程50次

图2A

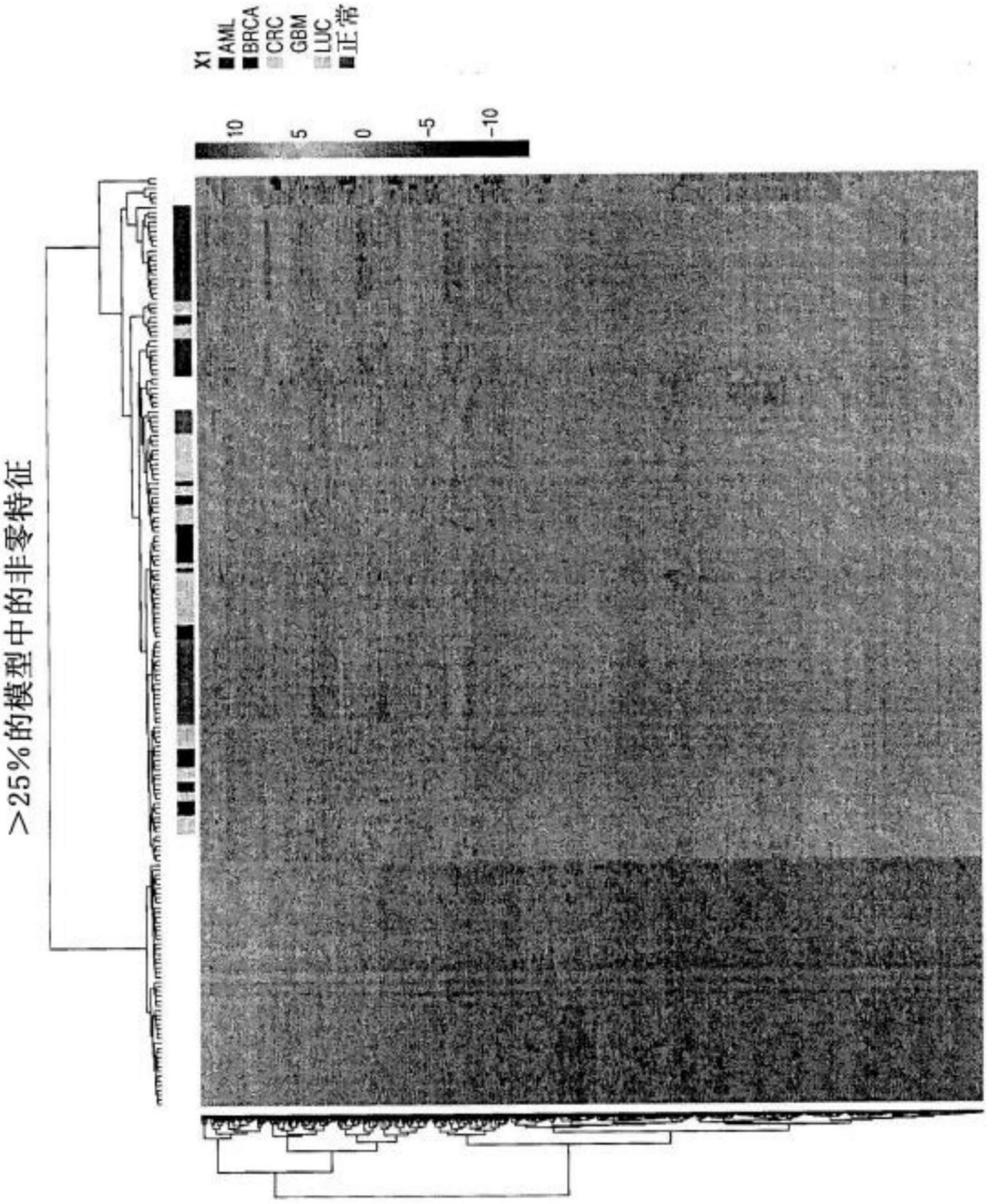


图2B

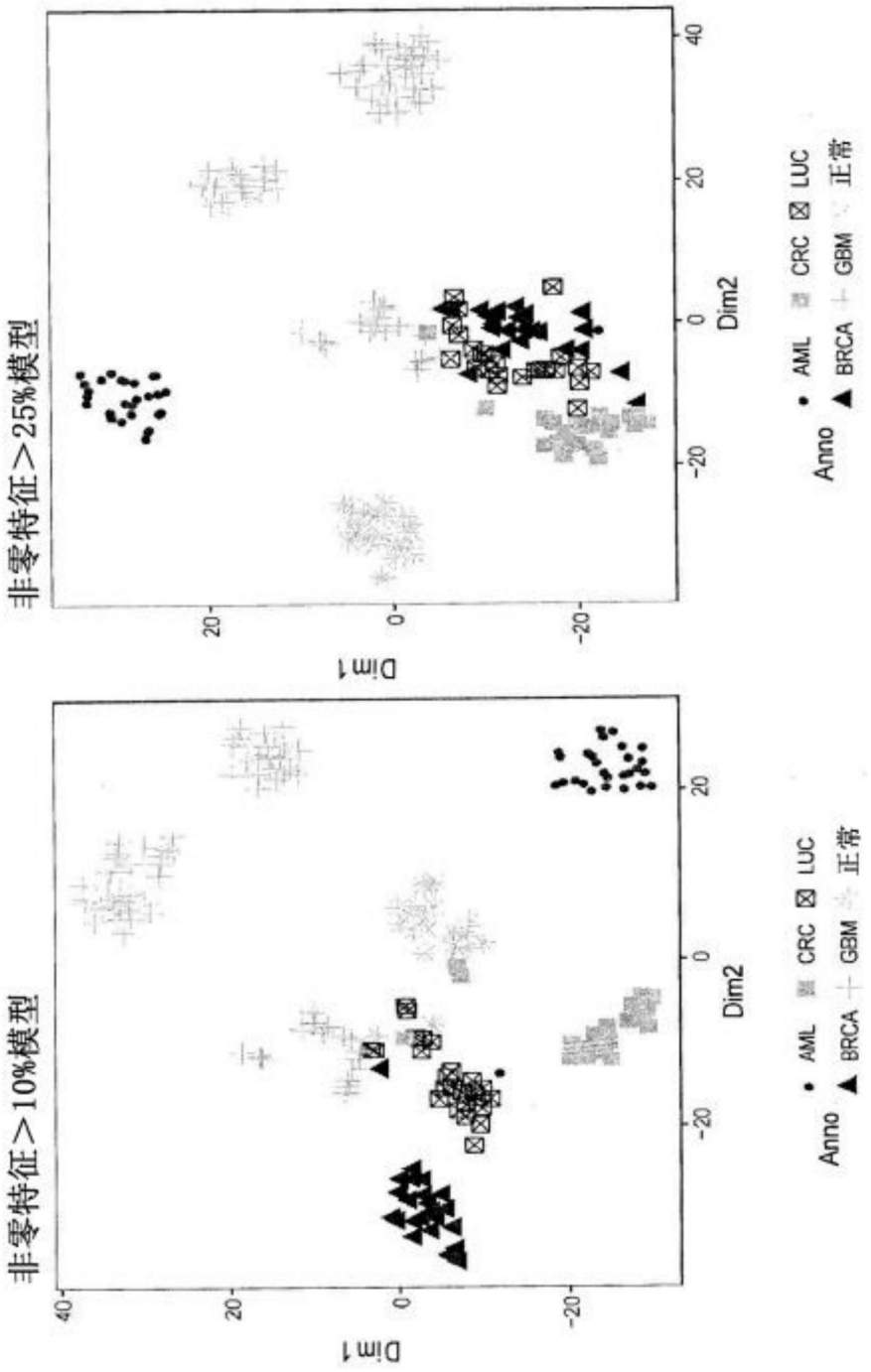


图2C

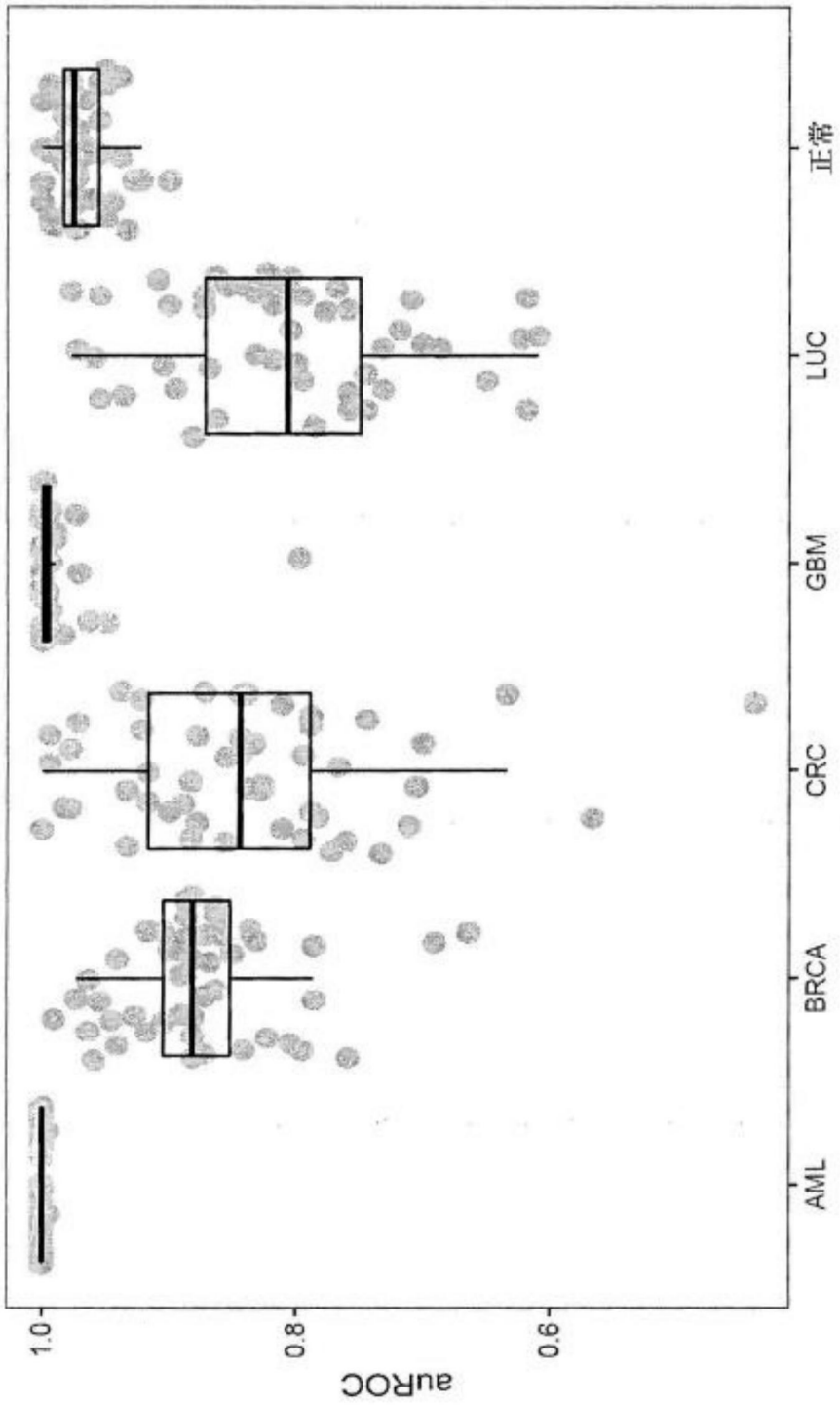


图2D

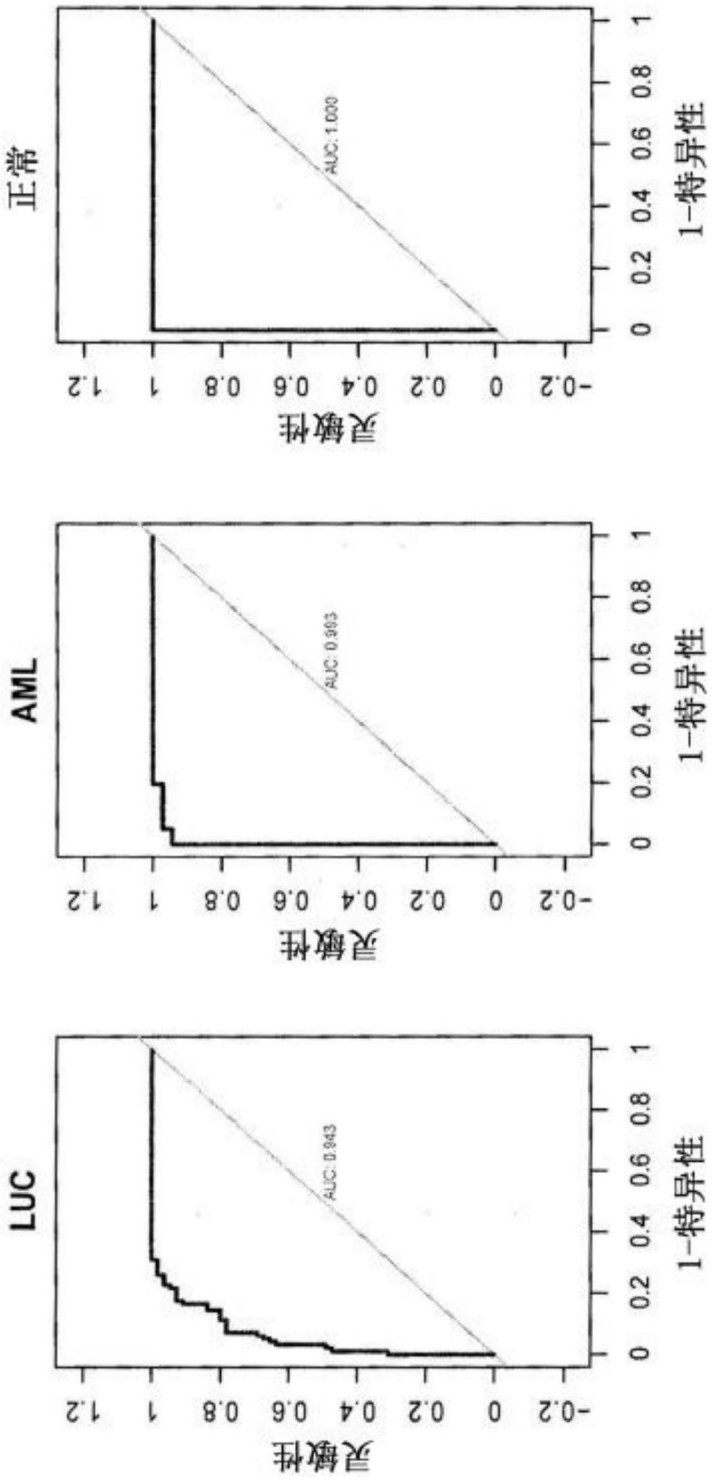

正常
AUC: 1.000
灵敏性
1-特异性
AML
AUC: 0.993
灵敏性
1-特异性
LUC
AUC: 0.943
灵敏性
1-特异性

图3A

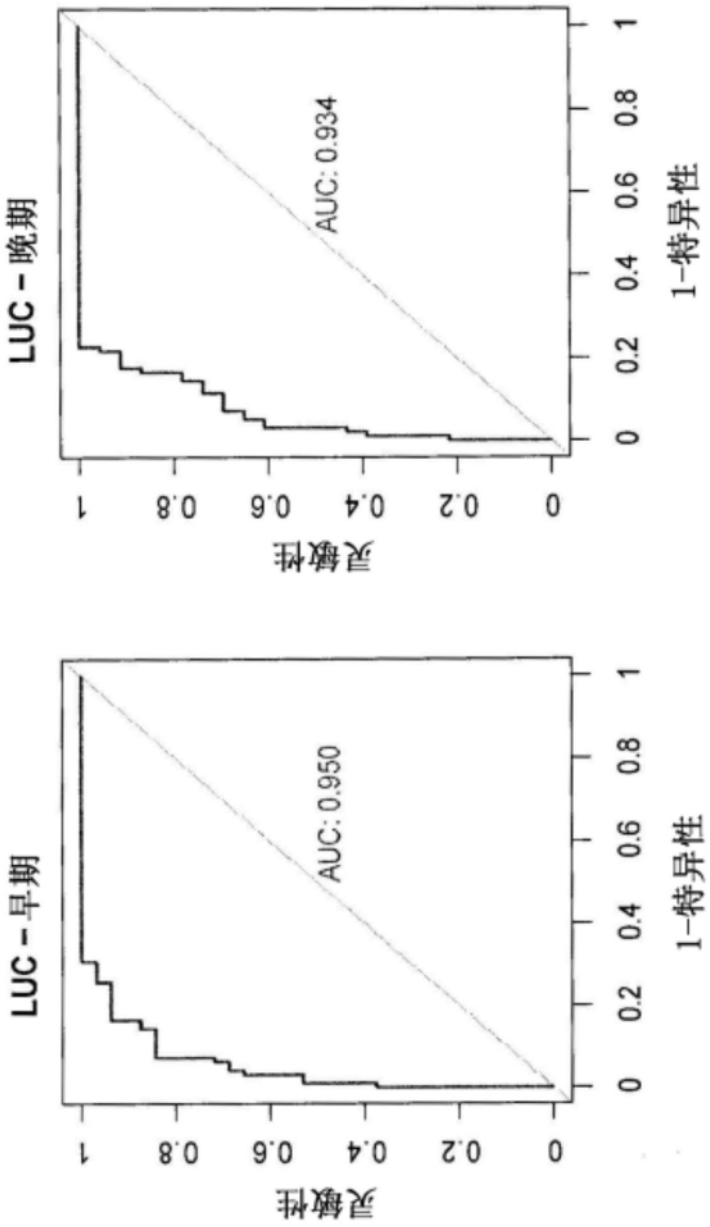

LUC – 晚期
AUC: 0.934
灵敏性
1–特异性
0
0.2
0.4
0.6
0.8
1
LUC – 早期
AUC: 0.950
灵敏性
1–特异性
0
0.2
0.4
0.6
0.8
1

图3B

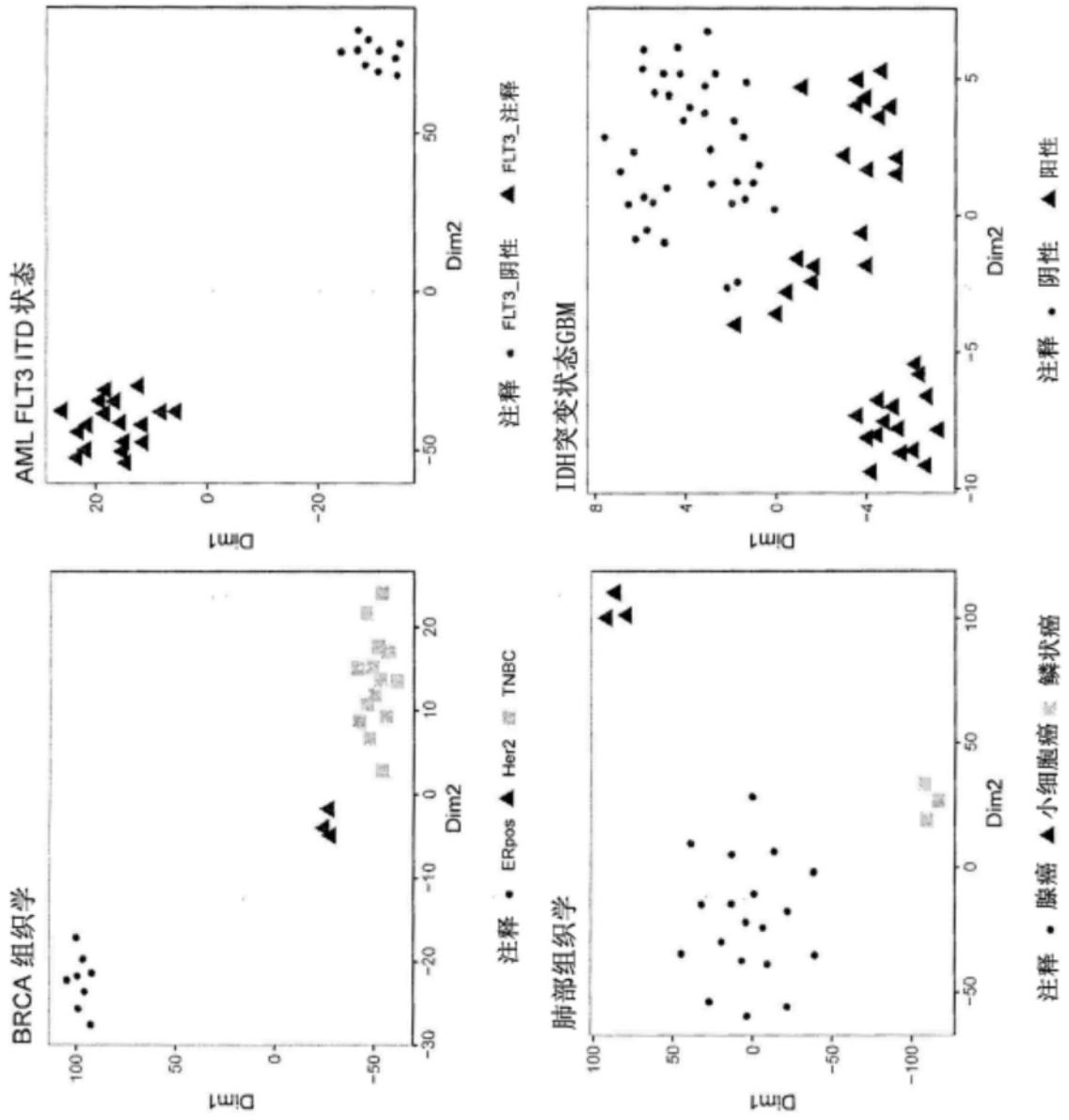

AML FLT3 ITD 状态
注释 • FLT3_阴性 ▲ FLT3_注释
IDH突变状态GBM
注释 • 阴性 ▲ 阳性
BRCA 组织学
注释 • ERpos ▲ Her2 TNBC
肺部组织学
注释 • 腺癌 ▲ 小细胞癌 鳞状癌
Dim1
Dim2

图4A

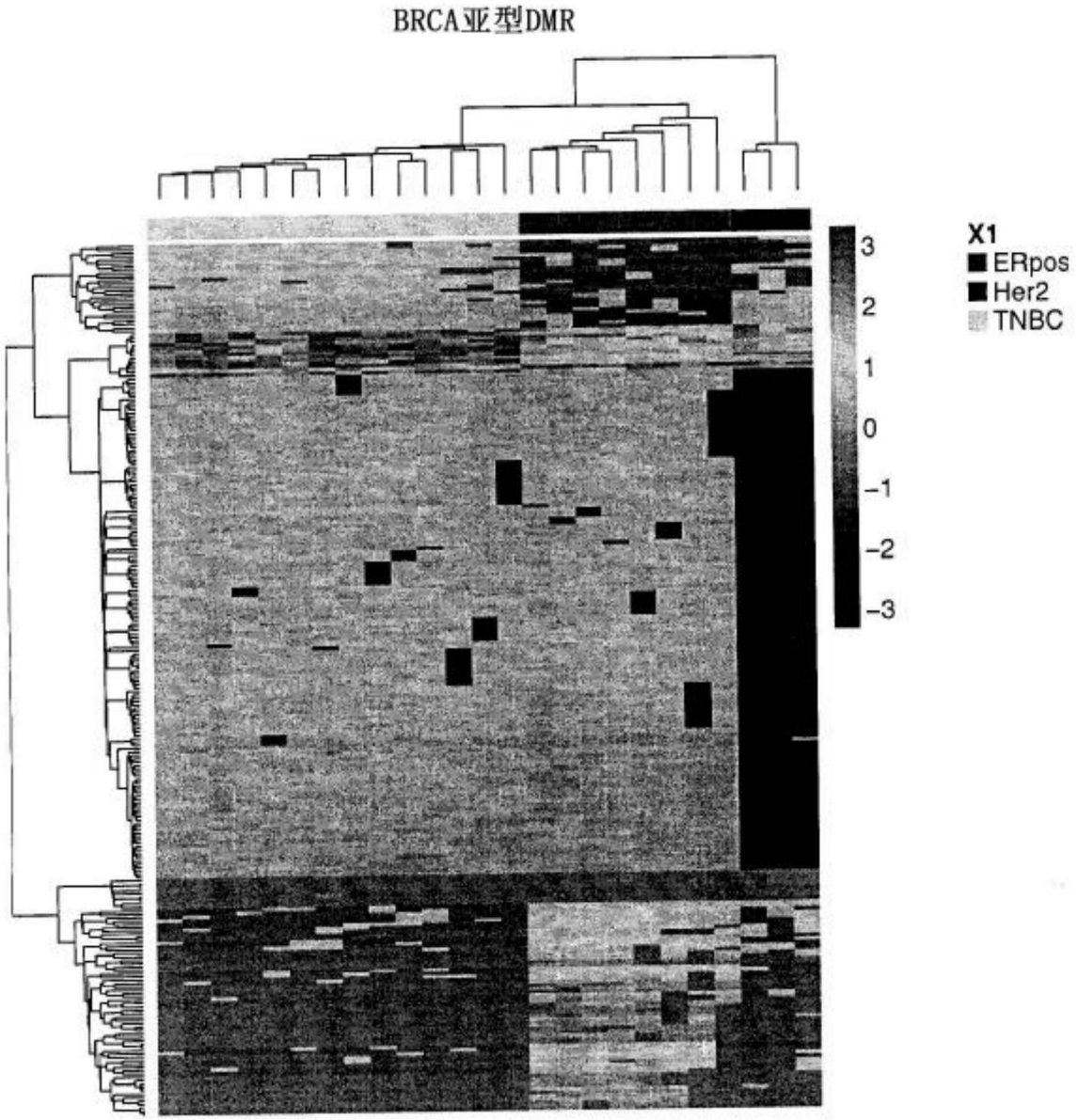


图4B

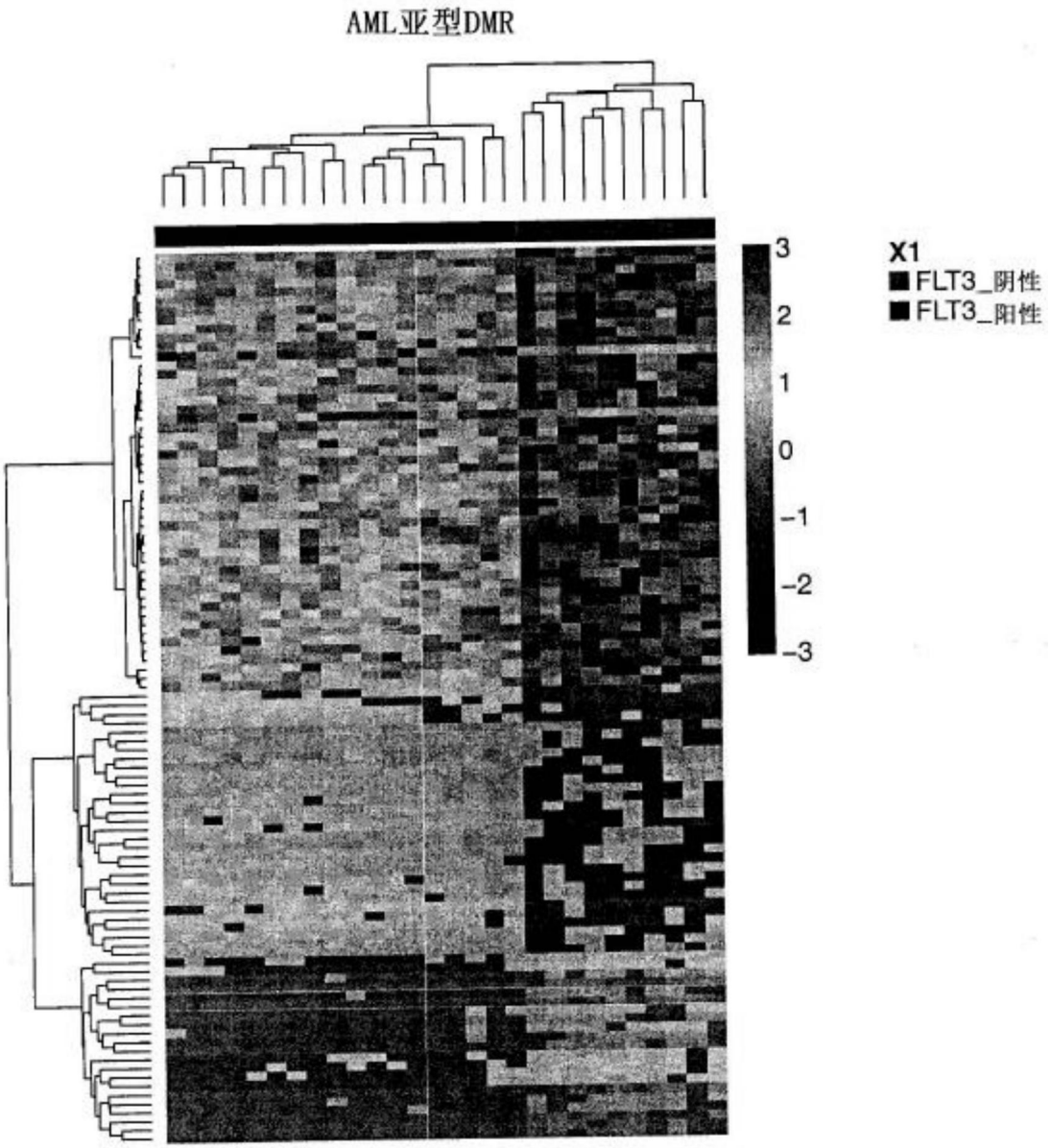


图4C

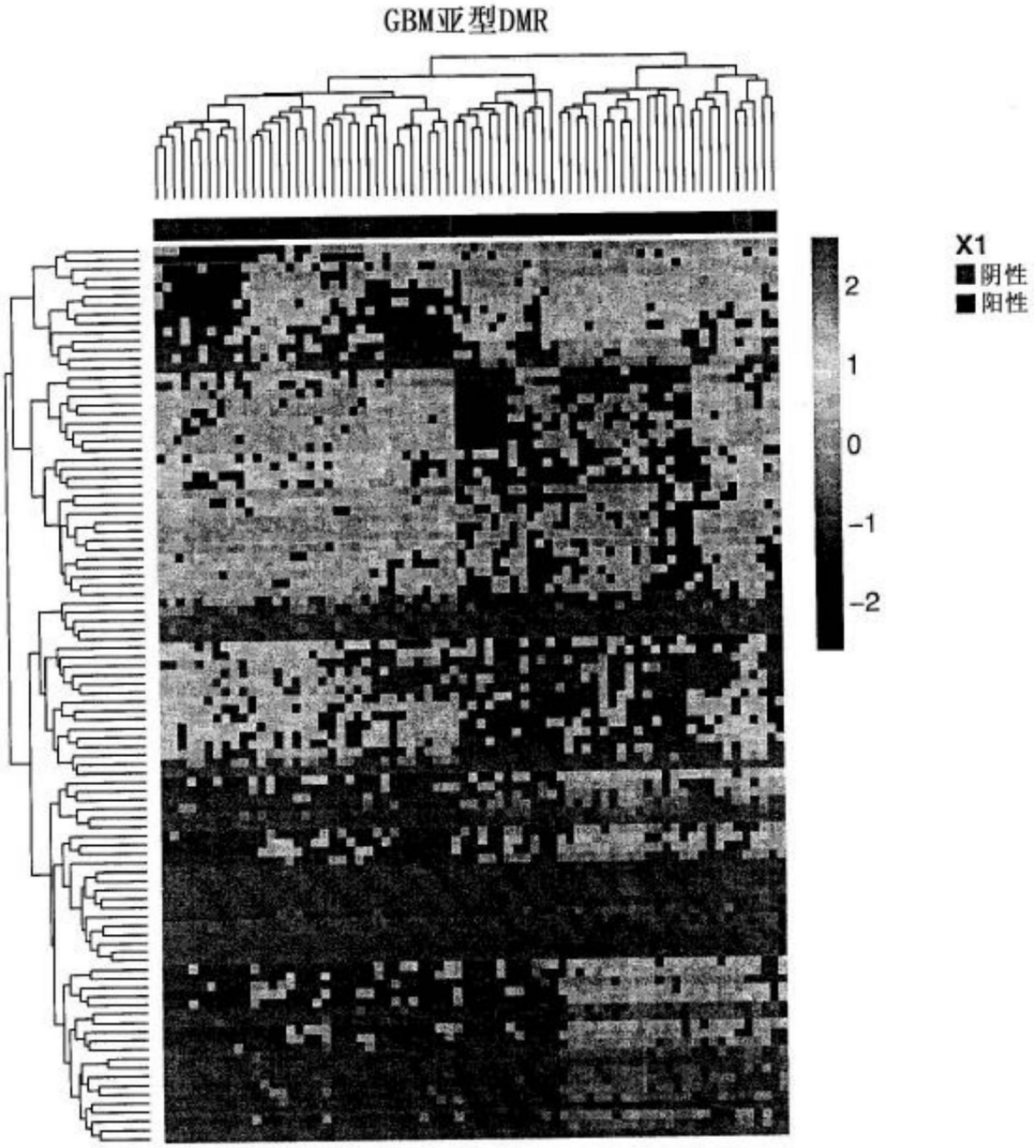


图4D

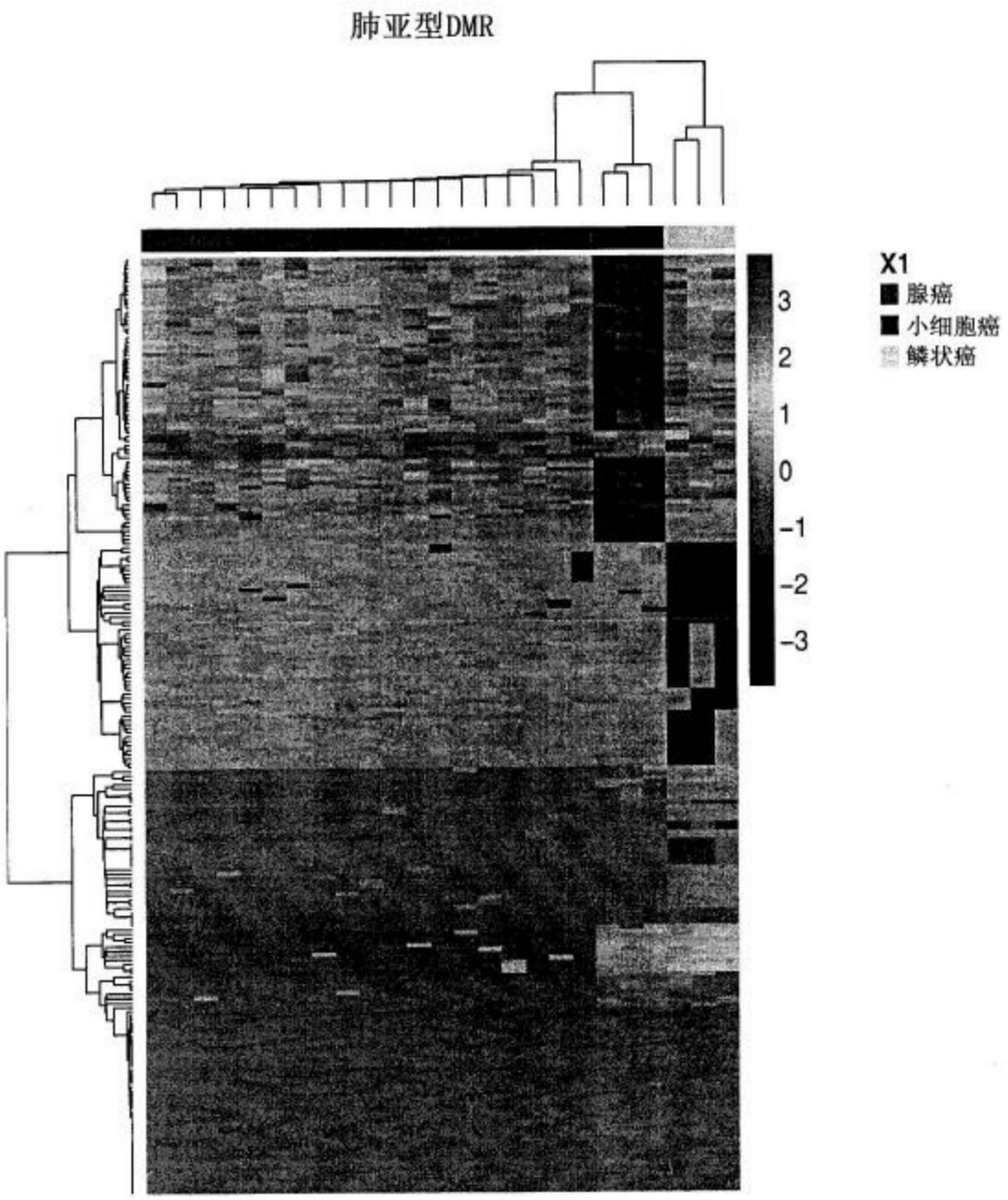


图4E

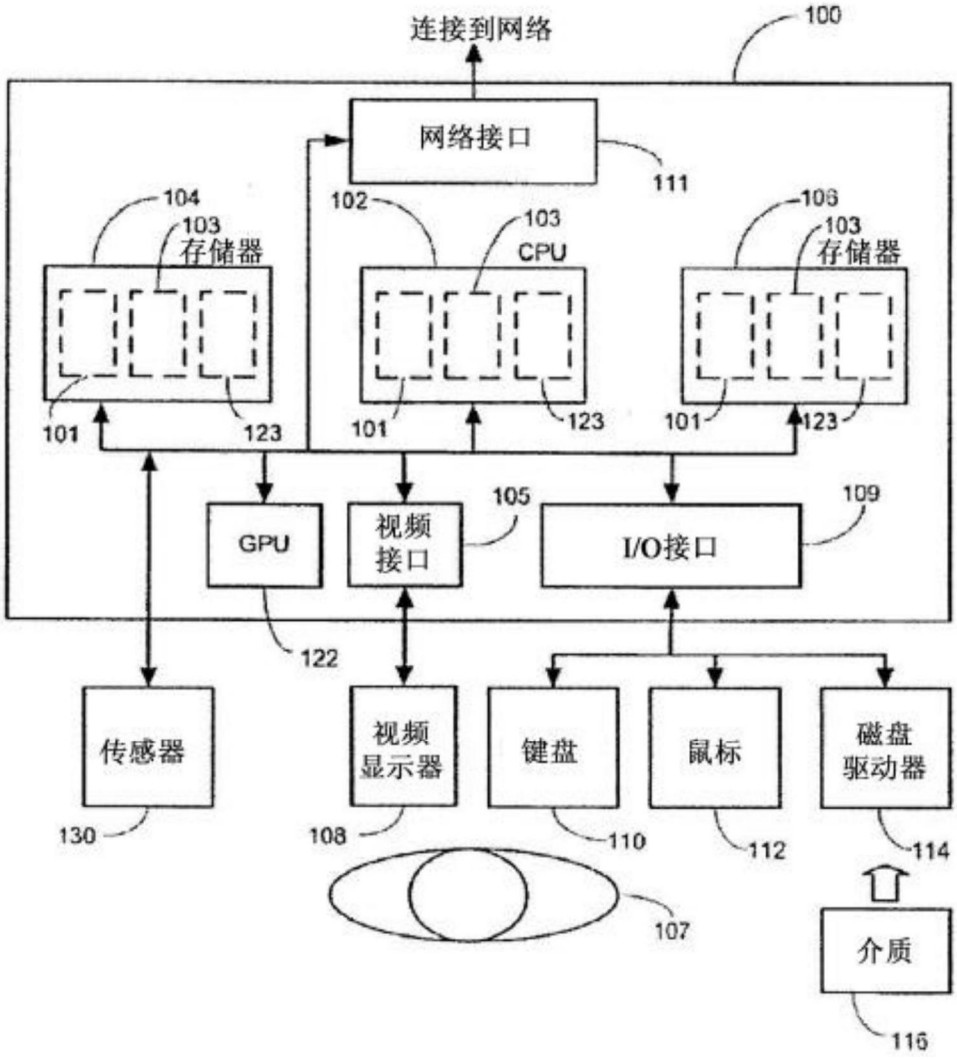

连接到网络
100
网络接口
111
102
103
CPU
104
103
存储器
106
103
存储器
101
123
101
123
101
123
GPU
122
视频
接口
105
I/O接口
109
传感器
130
视频
显示器
108
键盘
110
鼠标
112
磁盘
驱动器
114
107
介质
116

图5

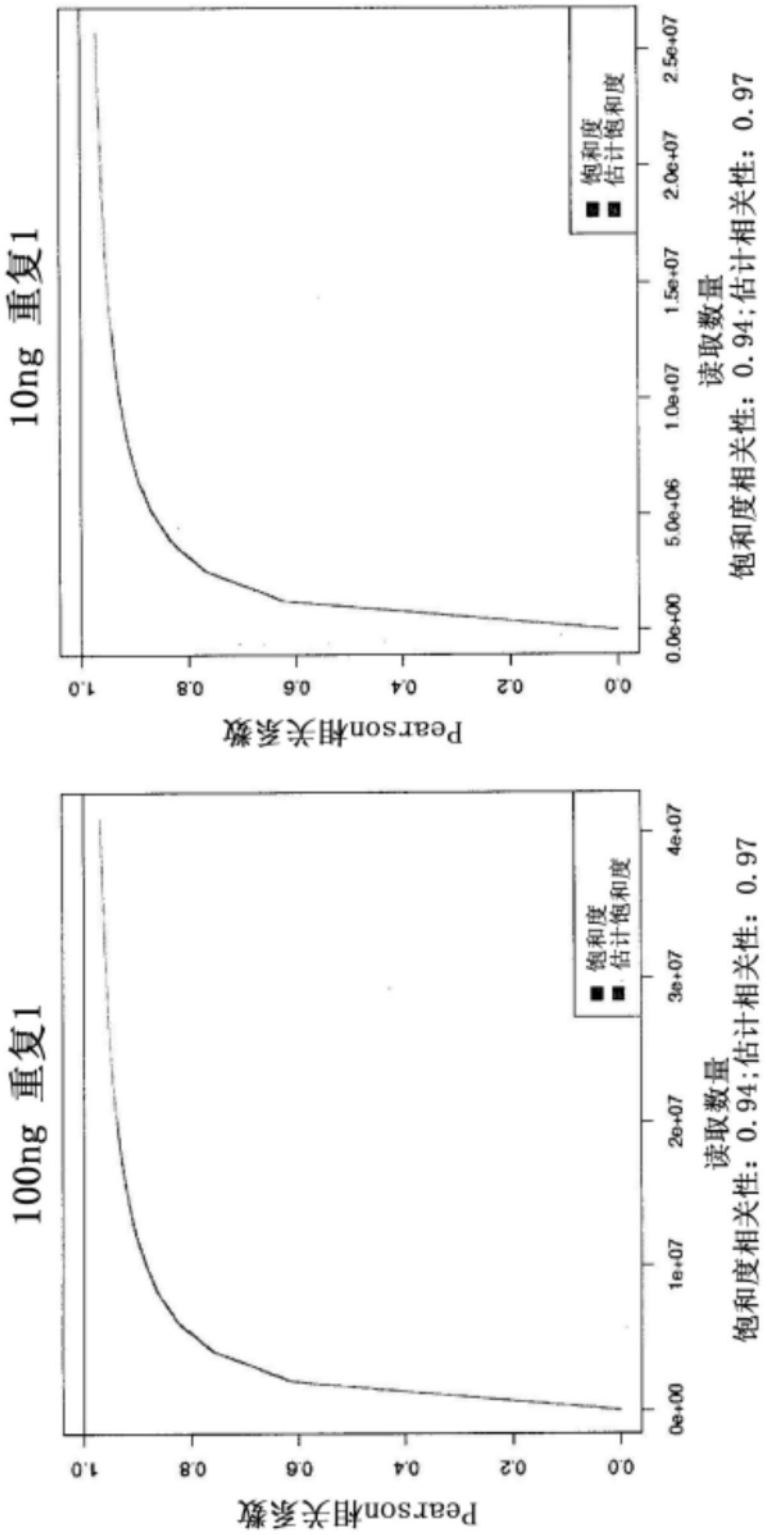

10ng 重复1
Pearson相关系数
读取数量
饱和度
估计饱和度
饱和度相关性：0.94;估计相关性：0.97
100ng 重复1
Pearson相关系数
读取数量
饱和度
估计饱和度
饱和度相关性：0.94;估计相关性：0.97

图6A

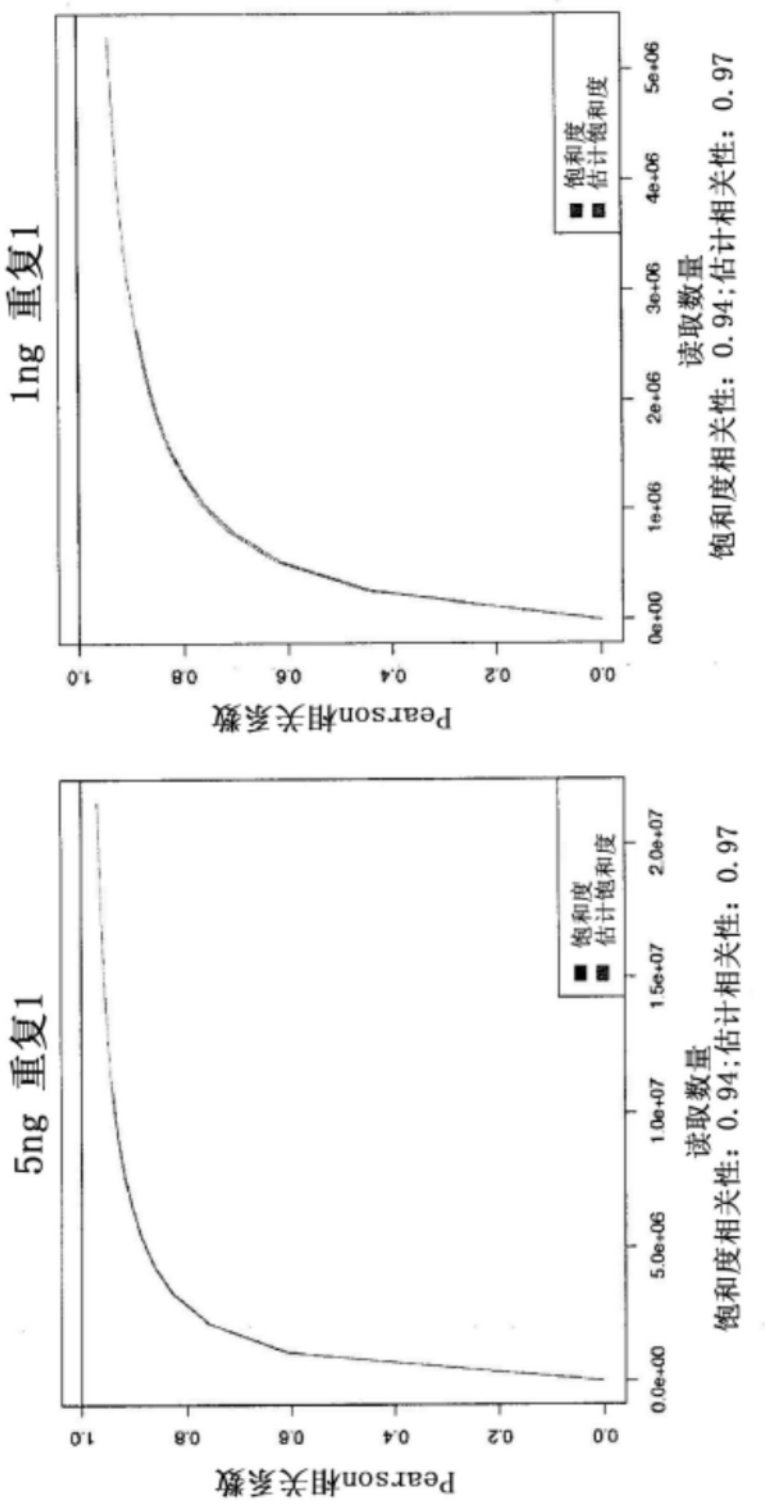

1ng 重复1
Pearson相关系数
读取数量
饱和度相关性：0.94;估计相关性：0.97
饱和度
估计饱和度
5ng 重复1
Pearson相关系数
读取数量
饱和度相关性：0.94;估计相关性：0.97
饱和度
估计饱和度

图6A续

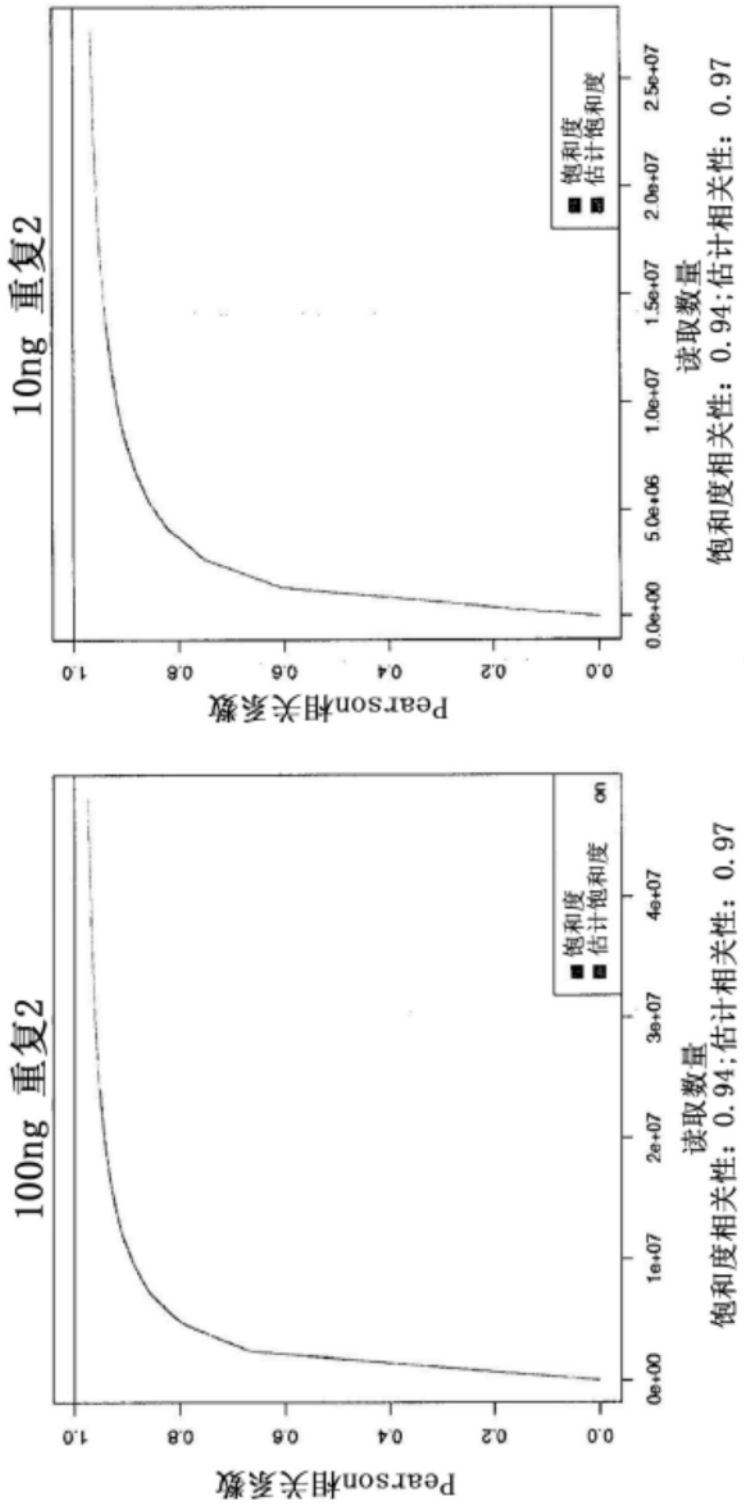

10ng 重复2
Pearson相关系数
读取数量
饱和度相关性：0.94;估计相关性：0.97
饱和度
估计饱和度
100ng 重复2
Pearson相关系数
读取数量
饱和度相关性：0.94;估计相关性：0.97
饱和度
估计饱和度

图6A续

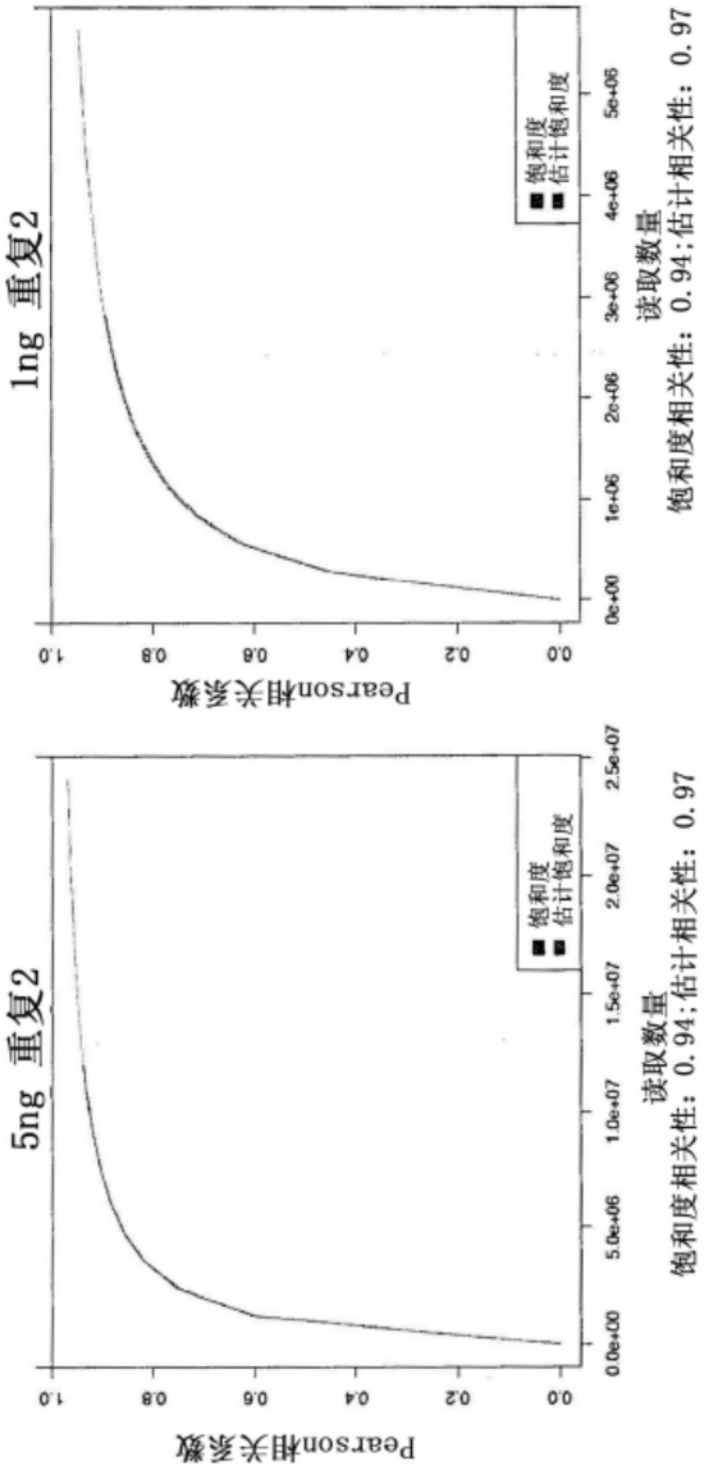

1ng 重复2
Pearson相关系数
读取数量
饱和度
估计饱和度
饱和度相关性：0.94;估计相关性：0.97
5ng 重复2
Pearson相关系数
读取数量
饱和度
估计饱和度
饱和度相关性：0.94;估计相关性：0.97

图6A续

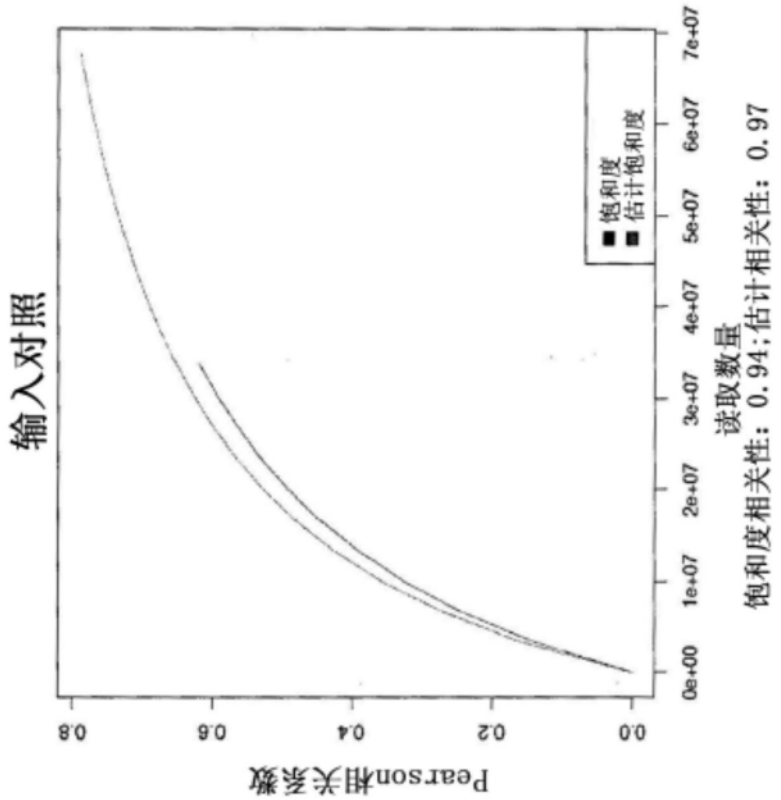

输入对照
Pearson相关系数
0.0
0.2
0.4
0.6
0.8
读取数量
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
7e+07
饱和度
估计饱和度
饱和度相关性：0.94;估计相关性：0.97

图6A续

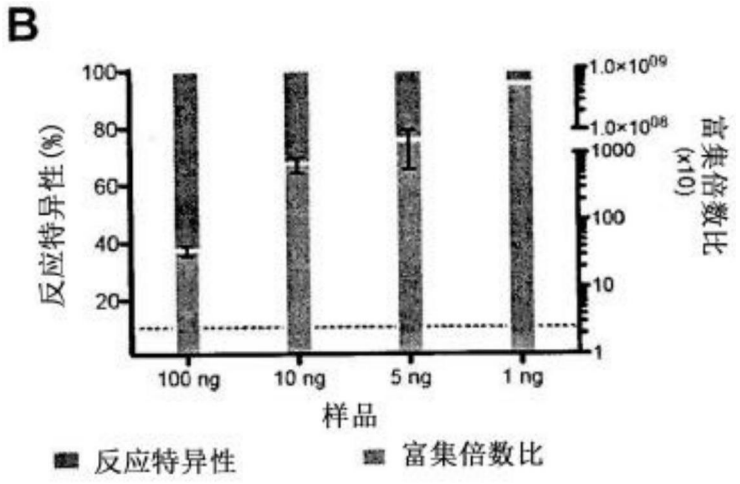

B
反应特异性(%)
100
80
60
40
20
富集倍数比
(x10)
1.0×10^09
1.0×10^08
1000
100
10
1
100 ng
10 ng
5 ng
1 ng
样品
反应特异性
富集倍数比

图6B

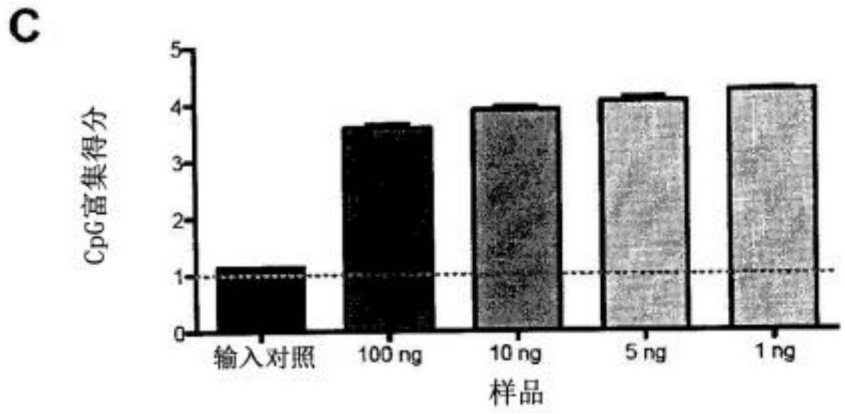


图6C