



(12)发明专利申请

(10)申请公布号 CN 109415763 A

(43)申请公布日 2019.03.01

(21)申请号 201780032740.5

(22)申请日 2017.05.03

(30)优先权数据

62/331,070 2016.05.03 US

(85)PCT国际申请进入国家阶段日

2018.11.27

(86)PCT国际申请的申请数据

PCT/CA2017/000108 2017.05.03

(87)PCT国际申请的公布数据

WO2017/190215 EN 2017.11.09

(71)申请人 大学健康网络

地址 加拿大安大略省

(72)发明人 丹尼尔·迪尼兹·德·卡尔瓦霍

沈淑怡 拉贾特·辛加尼亚

(74)专利代理机构 北京康信知识产权代理有限公司 11240

代理人 沈敬亭 张英

(51)Int.Cl.

C12Q 1/6858(2018.01)

C12Q 1/6886(2018.01)

C12N 15/10(2006.01)

C40B 30/04(2006.01)

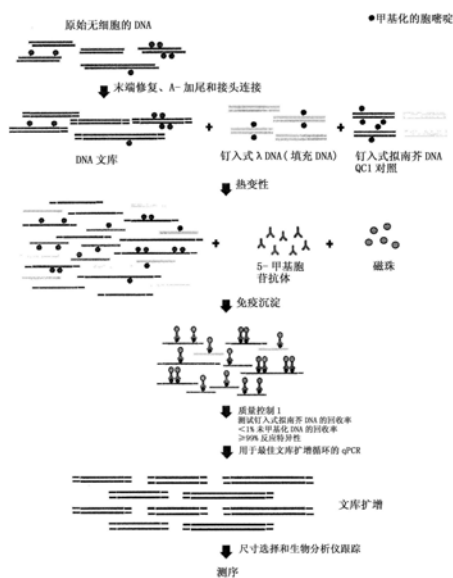
权利要求书2页 说明书24页 附图34页

(54)发明名称

捕获无细胞的甲基化DNA的方法及其用途

(57)摘要

本文描述了从具有少于100mg无细胞的DNA的样品中捕获无细胞的甲基化DNA的方法,包含以下步骤:对所述样品进行文库制备以允许随后所述无细胞的甲基化DNA的测序;向所述样品中添加第一量的填充DNA,其中所述填充DNA的至少一部分被甲基化;使所述样品变性;和使用对甲基化多核苷酸具有选择性的结合物捕获无细胞的甲基化DNA。



1. 一种从具有少于100ng无细胞的DNA的样品中捕获无细胞的甲基化DNA的方法,包含以下步骤:

- a. 对所述样品进行文库制备以允许随后所述无细胞的甲基化DNA的测序;
- b. 向所述样品中添加第一量的填充DNA,其中所述填充DNA的至少一部分被甲基化;
- c. 使所述样品变性;和
- d. 使用对甲基化多核苷酸具有选择性的结合物捕获无细胞的甲基化DNA。

2. 根据权利要求1所述的方法,进一步包括以下步骤:扩增捕获的无细胞的甲基化DNA并随后进行测序。

3. 根据权利要求1所述的方法,其中所述样品含有少于50ng的无细胞的DNA。

4. 根据权利要求1所述的方法,其中所述第一量的填充DNA包含约5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%或100%的甲基化的填充DNA,余量为未甲基化的填充DNA,并且优选地包含5%至50%、10%至40%、或15%至30%的甲基化的填充DNA。

5. 根据权利要求1所述的方法,其中所述第一量的填充DNA为20ng至100ng,优选30ng至100ng,更优选50ng至100ng。

6. 根据权利要求1所述的方法,其中来自所述样品的无细胞的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。

7. 根据权利要求1所述的方法,其中所述填充DNA的长度为50bp至800bp,优选长度为100bp至600bp,并且更优选长度为200bp至600bp。

8. 根据权利要求1所述的方法,其中所述填充DNA是双链的。

9. 根据权利要求1所述的方法,其中所述填充DNA是垃圾DNA。

10. 根据权利要求1所述的方法,其中所述填充DNA是内源DNA或外源DNA。

11. 根据权利要求10所述的方法,其中所述填充DNA是非人类DNA,优选λDNA。

12. 根据权利要求1所述的方法,其中所述填充DNA不能与人类DNA比对。

13. 根据权利要求1所述的方法,其中所述结合物是包含甲基-CpG-结合结构域的蛋白质。

14. 根据权利要求13所述的方法,其中所述蛋白质是MBD2蛋白质。

15. 根据权利要求1所述的方法,其中步骤(d)包括使用抗体免疫沉淀所述无细胞的甲基化DNA。

16. 根据权利要求15所述的方法,包括向所述样品中添加至少0.05μg所述抗体以进行免疫沉淀,并且优选为至少0.16μg。

17. 根据权利要求15所述的方法,其中所述抗体是5-MeC抗体或5-羟甲基胞嘧啶抗体。

18. 根据权利要求15所述的方法,进一步包括以下步骤:在步骤(b)之后向所述样品中添加第二量的对照DNA以确认免疫沉淀反应。

19. 根据权利要求1所述的方法,进一步包括以下步骤:在步骤(b)之后,向所述样品中添加第二量的对照DNA以确认捕获无细胞的甲基化DNA。

20. 权利要求1至19中任一项所述的方法用于测量样品中的DNA甲基化谱的用途。

21. 权利要求20中定义的DNA甲基化谱用于通过将所述谱与肿瘤组织的已知甲基化谱相关联以鉴别样品中存在来自癌细胞的无细胞的DNA的用途。

22. 权利要求20中定义的DNA甲基化谱用于通过将所述谱与特定组织的已知甲基化谱

相关联以鉴别样品中的无细胞的DNA的起源组织的用途。

23. 根据权利要求21所述的用途,进一步包括:权利要求22所述的用于鉴别样品中无细胞的DNA内的癌细胞的起源组织的用途。

24. 根据权利要求20至23中任一项所述是用途,用于监测免疫疗法。

25. 根据权利要求20至23中任一项所述的用途,用于诊断自身免疫病况。

26. 根据权利要求22所述的用途,用于确定从其中取样的受试者的细胞更新。

捕获无细胞的甲基化DNA的方法及其用途

[0001] 相关申请的引用

[0002] 本申请要求2016年5月3日提交的美国临时专利申请No.62/331,070的优先权,其通过引用以其整体并入本文。

技术领域

[0003] 本发明涉及无细胞的DNA领域,并且更具体地,涉及捕获无细胞的甲基化DNA的方法及其用途。

背景技术

[0004] DNA甲基化是一种DNA的共价修饰并且是一种稳定的基因调控机制,其在染色质结构中起重要作用。在人类中,DNA甲基化主要发生在CpG二核苷酸中的胞嘧啶残基上。与其他二核苷酸不同,CpG不是均匀分布在整个基因组中,而是集中在称为CpG岛的富含短CpG的DNA区域中。DNA甲基化可以通过两种主要机制导致基因抑制:1) 招募甲基结合结构域蛋白,其可以进而招募组蛋白脱乙酰酶(HDAC)和2) 阻止对转录因子(TF)的结合位点的访问,诸如c-MYC¹。

[0005] 一般而言,基因组中的大多数CpG位点是甲基化的,而大多数CpG岛在正常发育期间和分化组织中保持未甲基化¹。尽管如此,有可能在正常原发组织(primary tissue)中鉴别DNA甲基化的组织特异性模式²。此外,在恶性转化期间,经常观察到整体的DNA低甲基化(global DNA hypomethylation)和CpG岛处的局灶性高甲基化(focal hypermethylation)¹。事实上,在胶质母细胞瘤³、室管膜瘤(ependymomas)⁴、结肠直肠癌⁵、乳腺癌^{6,7}以及许多其它癌症类型中,DNA甲基化模式已被用于将癌症患者分层为具有预后值的临床相关亚组。

[0006] 由于其稳定性以及在正常分化和诸如癌症的疾病中的作用,DNA甲基化是一种良好的生物标志物,可用于表示肿瘤特征和表型状态,因此具有很高的个性化医疗潜力。许多样本类型适用于DNA甲基化作图(DNA methylation mapping)和生物标志物发现,包括新鲜和FFPE肿瘤组织、血细胞、尿液、唾液、粪便等⁸。最近,循环无细胞的DNA(cfDNA)作为生物标志物的使用正在获得动力,特别是在存在基因组区别的情况下,例如在癌症(体细胞突变)⁹、移植(供体相对于受体DNA)¹⁰和妊娠(胎儿相对于母体DNA)^{11,12}中。使用cfDNA的DNA甲基化作图作为生物标志物可能会产生重大影响,因为它可以允许以微创方式鉴别起源组织(tissue-of-origin)并对癌症患者进行分层。此外,它可以使得能够在不存在基因组区别的情况下将cfDNA用作生物标志物,例如监测免疫反应、神经退行性疾病或心肌梗塞,其中可以在cfDNA中检测到表观遗传畸变(epigenetic aberration)。

[0007] 此外,使用cfDNA的全基因组DNA甲基化作图(genome-wide DNA methylation mapping),可以克服在没有疾病射线照相证据的早期癌症患者中检测循环肿瘤DNA(ctDNA)中的关键敏感性(critical sensitivity)问题。现有的ctDNA检测方法是基于测序突变且其灵敏度有限,部分原因是由于可用于区分肿瘤和正常循环cfDNA的重复突变数量有限^{13,14}。另一方面,全基因组DNA甲基化作图利用大量表观遗传改变,其可用于区分循环肿瘤

DNA (ctDNA) 与正常循环的无细胞的DNA (cfDNA)。例如,一些肿瘤类型,例如室管膜瘤,可以具有广泛的DNA甲基化畸变而没有任何显著的重叠体细胞突变⁴。

[0008] 此外,来自癌症基因组图谱(The Cancer Genome Atlas) (TCGA) 的泛癌数据 (pan-cancer data) 显示,几乎所有肿瘤类型的肿瘤和正常组织之间都存在大量DMR¹⁵。因此,这些发现强调,成功从ctDNA中发现癌症特异性DNA甲基化改变的测定 (assay) 可以作为一种用于检测、分类和监测恶性疾病的非常敏感的工具,其具有低测序相关成本。

[0009] 然而,由于DNA的可用量较低以及cfDNA被片段化为长度小于200bp的事实,cfDNA中的全基因组作图DNA甲基化极具挑战性¹⁶。这使得无法进行传统的MeDIP-seq (其需要至少50-100ng DNA¹⁷) 或RRBS (简化表示亚硫酸氢盐测序 (Reduced Representation Bisulfite Sequencing)) (其需要非片段化的DNA¹⁸)。在cfDNA中作图DNA甲基化的另一个问题是正常cfDNA中感兴趣的DNA的低丰度¹⁹。这使得执行WGBS变得不切实际,因为以充足的深度捕获低丰度DNA的测序成本令人望而却步。另一方面,选择性富集易于甲基化的富含CpG特征的方法,可能使每次读取的有用信息量最大化、降低成本并减少DNA损失。

发明内容

[0010] 根据一个方面,提供了一种从具有小于100ng无细胞的DNA的样品捕获无细胞的甲基化DNA的方法,包含以下步骤:使所述样品经受文库制备以允许随后所述无细胞的甲基化DNA的测序;向所述样品中添加第一量的填充DNA (filler DNA),其中所述填充DNA的至少一部分被甲基化;使所述样品变性;和使用对甲基化多核苷酸具有选择性的结合物 (binder) 捕获无细胞的甲基化DNA。

附图说明

[0011] 通过参考以下描述和附图可以最好地理解本发明的实施方式。在附图中:

[0012] 图1显示cfDNA的甲基化组分析是一种在少量输入DNA中富集和检测ctDNA的高灵敏度方法。A) 计算机模拟用于检测至少一个表观突变的概率,作为ctDNA浓度 (列)、被调查的DMR数量 (行) 和测序深度 (x轴) 的函数。B) 来自片段化以模拟血浆cfDNA的HCT细胞系的1至100ng输入DNA的DNA甲基化信号之间的全基因组Pearson相关性。每种浓度具有两个生物学重复。C) 从不同浓度来自HCT116的输入DNA的cfMeDIP-seq获得的DNA甲基化图谱 (绿色迹线),加上获自ENCODE (ENCSR000DFS) 的RRBS (简化表示亚硫酸氢盐测序 (Reduced Representation Bisulfite Sequencing)) HCT116数据和获自GEO (GSM1465024) 的WGBS (全基因组亚硫酸氢盐测序 (Whole-Genome Bisulfite Sequencing)) HCT116数据。对于热图 (RRBS迹线),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连续稀释到多发性骨髓瘤 (MM) 细胞系MM1.S中。在纯HCT116 DNA (100% CRC), 纯MM1.S DNA (100% MM) 以及稀释到MM DNA中的10%、1%、0.1%、0.01%和0.001% CRC DNA中进行了cfMeDIP-seq。将所有DNA片段化以模拟血浆cfDNA。我们在观察值和预期值之间观察到了几乎完美的线性相关性 ($r^2=0.99$, $p<0.0001$) (D) DMR数量和 (E) 这些DMR中的DNA甲基化信号 (以RPKM计)。F) 在相同的稀释系列中,已知的体细胞突变仅可在1/100等位基因部分通过超深 (>10,000X) 靶向测序检测,高于背景测序仪和聚合酶错误率。显示的是在CRC细胞系中每个突变位点含有每个碱基或插入/缺失的读出部分 (fraction of read)。G) ctDNA

(人类)的频率,为在携带来自两名结直肠癌患者的患者来源异种移植物(PDX)的小鼠的血浆中总cfDNA(人类+小鼠)的百分比。

[0013] 图2显示了cfMeDIP-seq方案的示意图。

[0014] 图3显示了测序饱和度分析和质量控制。A) 该图显示了来自Bioconductor包装MEDIPS分析cfMeDIP-seq数据的饱和度分析的结果,所述cfMeDIP-seq数据来自被片段化以模拟血浆cfDNA的HCT116DNA的每个输入浓度的每个重复。B) 该方案是在四个HCT116细胞系的起始DNA浓度(100、10、5和1ng)的两个重复中进行测试。使用甲基化和未甲基化的钉入式(spiked-in)拟南芥DNA计算反应的特异性。使用片段化的HCT116DNA的基因组区域计算倍数富集比率(用于甲基化睾丸特异性H2B、TSH2B0和未甲基化的人类DNA区域的引物(GAPDH启动子))。水平虚线表示倍数富集比率阈值为25。误差棒代表 $\pm 1s.e.m.$ 。C) 测序样品的CpG富集评分显示,与输入对照相比,CpG在来自免疫沉淀样品的基因组区域内具有强烈富集。通过将区域的CpG的相对频率除以人类基因组的CpG的相对频率来获得CpG富集评分。误差棒代表 $\pm 1s.e.m.$ 。

[0015] 图4显示了来自cfMeDIP-seq的连续稀释的质量控制。A) 稀释至MMDNA(MM1.S)中的CRC DNA(HCT116)的示意图。B) 使用甲基化和未甲基化的钉入式拟南芥DNA计算每种稀释度的反应特异性。C) 测序样品的CpG富集评分显示了来自免疫沉淀样品的基因组区域内的CpG的强烈富集。通过将区域的CpG的相对频率除以人类基因组中CpG的相对频率来获得CpG富集评分。D) 该图显示了每个稀释点的饱和度分析结果。

[0016] 图5显示,cfMeDIP-seq方法可以鉴别从胰腺腺癌患者获得的循环cfDNA上的数千个差异甲基化的区域。A) 实验设计。B) 使用cfMeDIP-seq获得的来自胰腺癌(病例, $n=24$)与来自健康供体(对照, $n=24$)的循环cfDNA的火山图(Volcano plot)。红点指示在多次测试校正后达到显著性的窗口。C) 在来自健康供体和胰腺癌患者的血浆DNA中鉴别的38,085个DMR的热图。分层聚类方法(Hierarchical clustering method):Ward。D) 排列分析(Permutation analysis),以估计在血浆中鉴别的DMR(病例相对于对照)和在原发肿瘤组织中鉴别的癌症特异性DMC(原发肿瘤相对于正常组织)之间的预期重叠相对于观察到的重叠的频率。方框图代表重叠的空分布。菱形代表实验观察到的原发肿瘤组织与来自循环cfDNA的DNA甲基化之间的重叠数。红色菱形意味着观察到的重叠数量意外地显著高于预期的重叠数量。绿色菱形意味着所观察到的重叠数量意外地显著低于预期的重叠数量,并且蓝色菱形是非显著的。我们计算了四种可能的重叠:原发肿瘤组织中高甲基化和循环cfDNA中高甲基化(富集的, P 值: 6.4×10^{-22});肿瘤组织中的高甲基化和循环cfDNA中的低甲基化(耗尽的, P 值: 9.43×10^{-17});肿瘤组织中的低甲基化和循环cfDNA中的低甲基化(富集的, P 值: 1.88×10^{-283});肿瘤组织中的低甲基化和循环cfDNA中的高甲基化(P 值:0.105)。E) 排列分析(Permutation analysis),以估计在血浆中鉴别的DMR(病例相对于对照)和在原发肿瘤组织中鉴别的癌症特异性DMC(原发肿瘤相对于正常PBMC)之间的预期重叠相对于观察到的重叠的频率。

[0017] 图6显示了来自胰腺腺癌患者(病例)和健康供体(对照)的循环cfDNA的cfMeDIP-seq的质量控制。A-B) 使用甲基化和未甲基化的钉入式拟南芥DNA计算每个病例(A)和每个对照(B)样品的反应的特异性。由于可获得的DNA量非常有限,因此未计算倍数富集比率。C-D) 测序样品的CpG富集评分显示了在来自免疫沉淀样品的基因组区域内的CpG的强烈富集。

[0018] 图7显示了A)使用顶部百万个最可变的全基因组窗口对来自健康供体和早期胰腺腺癌患者的48种血浆cfDNA甲基化进行PCA。对于每个窗口,使用MAD(平均绝对偏差)度量来计算可变性,该度量是一个稳健的测量,其返回与数据中值的绝对偏差的中值;在这种情况下,数据是给定窗口的所有48个样品的RPKM值。示出了PC1相对于PC2(左)以及PC1相对于PC3(右)。B)每个主要分量的方差百分比。C)使用RRBS获得的来自胰腺腺癌患者的肿瘤相对于正常LCM组织的火山图。列出了识别的DMC(差异甲基化的CpG)的总数。红点指示在多次测试校正后达到显著性并且具有 >0.25 的绝对甲基化差异(绝对 $\Delta\beta$)的窗口。D)显示每个重叠窗口的DNA甲基化差异的显著性的散点图。X轴显示来自RRBS数据的原发胰腺腺癌肿瘤相对于正常组织的 $\log_{10}q$ 值。如果该区域在肿瘤中是高甲基化的,则以正标度显示显著性。低甲基化区域以负标度显示。Y轴显示来自cfMeDIP-seq数据的胰腺腺癌患者相对于健康供体的血浆cfDNA甲基化的 $\log_{10}q$ 值。蓝点在两者中都是显著的。红线显示趋势线。E)显示每个重叠窗口的DNA甲基化差异的散点图。X轴显示来自RRBS数据的原发胰腺腺癌肿瘤相对于正常组织的DNA甲基化差异。Y轴显示来自cfMeDIP-seq数据的胰腺腺癌患者相对于健康供体的血浆cfDNA甲基化的DNA甲基化差异。蓝线显示趋势线。F)使用RRBS获得的LCM胰腺腺癌组织相对于正常PBMC的火山图。列出了识别的DMC(差异甲基化的CpG)的总数。红点指示在多次测试校正后达到显著性并且具有 >0.25 的绝对甲基化差异(绝对 $\Delta\beta$)的窗口。G)显示每个重叠窗口的DNA甲基化差异的显著性的散点图。X轴显示来自RRBS数据的原发胰腺腺癌肿瘤相对于正常PBMC的 $\log_{10}q$ 值。如果该区域在肿瘤中是高甲基化的,则以正标度显示显著性。低甲基化区域以负标度显示。Y轴显示来自cfMeDIP-seq数据的胰腺腺癌患者相对于健康供体的血浆cfDNA甲基化的 $\log_{10}q$ 值。蓝点在两者是都是显著的。红线显示趋势线。H)显示每个重叠窗口的DNA甲基化差异的散点图。X轴显示来自RRBS数据的原发胰腺腺癌肿瘤相对于正常PBMC的DNA甲基化差异。Y轴显示来自cfMeDIP-seq数据的胰腺腺癌患者相对于健康供体的血浆cfDNA甲基化的DNA甲基化差异。

[0019] 图8显示循环cfDNA甲基化谱可以用于鉴别转录因子(TF)足迹并推断来源组织中的活性转录网络。A)所有TF($n=33$)表达谱,在多种人体组织中,所述TF的基序富集(使用软件HOMER²⁰)于来自健康供体的cfDNA中的低甲基化区域中(对照中为高甲基化足迹)。表达数据来自基因组-组织表达(GTE_x)项目²¹。鉴别了在造血系统中优先表达的几种TF(PU.1, Fli1, STAT5B, KLF1)。B)在对照中具有低甲基化基序的所有TF的表达谱($n=33$)相对于全血中33个TF的1000个随机组的表达谱(GTE_x数据)。C)所有TF($n=85$)的表达谱,其基序富集在来自胰腺腺癌患者的cfDNA低甲基化区域(病例中为低甲基化足迹)。鉴别了几种胰腺特异性或胰腺癌相关的TF。此外,还鉴别了驱动胰腺癌分子亚型的标志性TF。D)病例中具有低甲基化基序的所有TF的表达谱($n=85$)相对于正常胰腺中85个TF的1000个随机组的表达谱(GTE_x数据)。E)病例中具有低甲基化基序的所有TF的表达谱($n=85$)相对于胰腺腺癌组织中85个TF的1000个随机组的表达谱(TCGA数据)。

[0020] 图9显示了在免疫沉淀之前,在使用10ng、5ng和1ng起始无癌细胞的DNA量($n=3$)分别结合90ng、95ng和99ng的填充DNA或无填充DNA进行cfMeDIP-seq后,钉入式未甲基化的拟南芥DNA的%回收。在免疫沉淀至100ng之前,改变人工甲基化%相对于存在的%未甲基化的 λ DNA的组合物中所使用的填充DNA,以增加最终量。期望的钉入式未甲基化DNA的回收 $<1.0\%$,较低的回收导致较高的%反应特异性。

[0021] 图10显示了在免疫沉淀之前,使用10ng、5ng和1ng起始无癌细胞的DNA量(n=3)分别结合90ng、95ng和99ng填充DNA或无填充DNA进行cfMeDIP-seq后,钉入式甲基化的拟南芥DNA的%回收。在免疫沉淀至100ng之前,改变人工甲基化%相对于存在的%未甲基化的 λ DNA的组合物中所使用的填充DNA以增加最终量。期望的钉入式甲基化DNA的最小%回收为20%。

具体实施方式

[0022] 我们生物信息地模拟了具有从0.001%到10%不同比例的ctDNA的混合物(图1中的A,列侧面)。我们还模拟了与正常cfDNA相比,ctDNA具有1、10、100、1000或10000个DMR(差异甲基化的区域)的情况(图1中的A,行侧面)。然后在每个基因座(10X、100X、1000X和10000X)在不同测序深度对读数(read)进行取样(图1中的A,x轴)。我们发现,随着DMR数量的增加,检测到至少1例癌症特异性事件的概率增加(图1中的A),即使在低丰度的癌症ctDNA和浅覆盖率的情况下也是如此。

[0023] 为了克服这些挑战,我们开发了一种称为cfMeDIP-seq(无细胞的甲基化DNA免疫沉淀和高通量测序(cell-free Methylated DNA Immunoprecipitation and high-throughput sequencing))的新方法,以使用无细胞的DNA进行全基因组DNA甲基化作图。这里描述的cfMeDIP-seq方法是通过修改现有的低输入MeDIP-seq方案¹⁷(该方案可以稳定到100ng的输入DNA)开发的。然而,大多数血浆样品产生的DNA远低于100ng。为了克服这一挑战,我们将外源 λ DNA(填充DNA)添加到接头(adapter)连接的cfDNA文库中,以将起始DNA的量人工膨胀到100ng(图2)。这最小化了抗体的非特异性结合量,并且还最小化了由于与塑料器皿结合而损失的DNA量。填充DNA由与接头连接的cfDNA文库大小相似的扩增子组成,并且包含不同甲基化水平的未甲基化和体外甲基化DNA(图9和图10)。添加这种填充DNA也具有实际用途,因为不同的患者会产生不同量的cfDNA,从而使输入DNA的量标准化为100ng。这确保了下游方案对于所有样品保持完全相同,而不管可用的cfDNA的量是多少。

[0024] 根据一个方面,提供了一种从具有少于100ng无细胞的DNA的样品中捕获无细胞的甲基化DNA的方法,包含以下步骤:

[0025] a. 使所述样品经受文库制备以允许随后所述无细胞的甲基化DNA的测序;

[0026] b. 向所述样品中添加第一量的填充DNA,其中所述填充DNA的至少一部分被甲基化;

[0027] c. 使所述样品变性;和

[0028] d. 使用对甲基化多核苷酸具有选择性的结合物捕获无细胞的甲基化DNA。

[0029] 在一些实施方式中,这一方法进一步包含以下步骤:扩增所述捕获的无细胞的甲基化DNA并且随后进行测序。

[0030] 各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)之后Sanger测序。还可用的是新一代测序(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测试, 离子洪流: 质子/PGM测序, SOLiD测序。与先前使用的Sanger测序相比,NGS允许更快和更便宜地测序DNA和RNA。在一些实施方式中,所述测序针对短读测序进行了优化。

[0031] 无细胞的甲基化DNA是在血流中自由循环的DNA,并且在所述DNA的各种已知区域

被甲基化。可以采集样品(例如血浆样品)来分析无细胞的甲基化DNA。

[0032] 如本文所使用的,“文库制备”包括列表末端修复、A-加尾、接头连接或对无细胞的DNA进行的任何其他制备以允许随后的DNA测序。

[0033] 如本文所使用的,“填充DNA”可以是非编码DNA或它可以由扩增子组成。

[0034] 可以例如使用足够的热量使DNA样品变性。

[0035] 在一些实施方式中,样品含有少于50ng的无细胞的DNA。

[0036] 在一些实施方式中,所述第一量的填充DNA包含约10%、20%、30%、40%、50%、60%、70%、80%、90%或100%的甲基化填充DNA。在优选的实施方式中,所述第一量的填充DNA包含约50%的甲基化填充DNA。

[0037] 在一些实施方式中,所述第一量的填充DNA为20ng至100ng。在优选的实施方式中,30ng至100ng的填充DNA。在更优选的实施方式中,50ng至100ng的填充DNA。当将来自样品的无细胞的DNA和所述第一量的填充DNA组合在一起时,其包含至少50ng总DNA,且优选至少100ng总DNA。

[0038] 在一些实施方式中,填充DNA长度为50bp至800bp。在优选的实施方式中,长度为100bp至600bp;并且在更优选的实施方式中,长度为200bp至600bp。

[0039] 填充DNA是双链的。例如,填充DNA可以是垃圾DNA(junk DNA)。填充DNA也可以是内源DNA或外源DNA。例如,填充DNA是非人类DNA,在优选的实施方式中,是 λ DNA。如本文使用的,“ λ DNA”指肠杆菌噬菌体 λ DNA(Enterobacteria phage λ DNA)。在一些实施方式中,填充DNA与人类DNA不能比对。

[0040] 在一些实施方式中,结合物是包含甲基-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。如本文使用的,“甲基-CpG结合结构域(MBD)”指蛋白质和酶的某些结构域,所述蛋白质和酶长约70个残基并且与含有一个或多个对称甲基化CpG的DNA结合。MeCP2、MBD1、MBD2、MBD4和BAZ2的MBD介导与DNA的结合,并且在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包含通过在每个甲基-CpG结合结构域(MBD)中存在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特异性结合甲基化DNA。

[0041] 在其它实施方式中,所述结合物是抗体并且捕获无细胞的甲基化DNA包含使用该抗体免疫沉淀所述无细胞的甲基化DNA。如本文使用的,“免疫沉淀”是指使用特异性结合特定抗原(如多肽和核苷酸)的抗体从溶液中沉淀出所述抗原的技术。该方法可用于从样品中分离和浓缩特定蛋白质或DNA,并且要求在该程序的某些点处将抗体偶联到固体基底上。固体基底包括例如珠子,例如磁珠。其它类型的珠子和固体基底在本领域中是已知的。

[0042] 一种示例性抗体是5-MeC抗体。对于免疫沉淀程序,在一些实施方式中,向样品中加入至少0.05 μ g抗体;而在更优选的实施方式中,向样品中加入至少0.16 μ g抗体。为了确认免疫沉淀反应,在一些实施方式中,本文所述的方法还包括在步骤(b)后向样品中加入第二量的对照DNA的步骤。

[0043] 另一种示例性抗体是或5-羟甲基胞嘧啶抗体。

[0044] 在其它实施方式中,本文所述的方法进一步包含在步骤(b)之后向样品中添加第二量的对照DNA的步骤以确认无细胞的甲基化DNA的捕获。

[0045] 如本文使用的,“对照”可包括阳性和阴性对照,或至少阳性对照。

[0046] 根据进一步的方面,提供了本文所述的方法用于测量样品中的DNA甲基化谱的用途。

[0047] 根据进一步的方面,提供了本文所述的方法用于鉴别样品中存在来自癌细胞的无细胞的DNA的用途(通过将所述谱与肿瘤组织的已知甲基化谱相关联)。

[0048] 根据进一步的方面,提供了本文所述的DNA甲基化谱用于鉴别样品中无细胞的DNA的起源组织的用途(通过将所述谱与特定组织的已知甲基化谱相关联)。

[0049] 在一些实施方式中,所述用途进一步包含本文所述的用途,用于鉴别样品内无细胞的DNA中的癌细胞的起源组织。

[0050] 根据进一步的方面,提供了本文所述的用途,用于监测免疫疗法。

[0051] 根据进一步的方面,提供了本文所述的用途,用于诊断自身免疫病况。

[0052] 根据进一步的方面,提供了本文所述的用途,用于确定从其中取样的受试者内细胞更新(cell turnover)。

[0053] 以下实施例说明了本发明的各个方面,并且不限制本文公开的本发明的广泛方面。

[0054] 实施例

[0055] 方法

[0056] 供体招募和样品采集

[0057] 胰腺腺癌(PDAC)患者样品获自大学健康网络BioBank(University Health Network BioBank);通过位于加拿大多伦多的西奈山医院(Mount Sinai Hospital(MSH))的家庭医学中心(Family Medicine Centre)招募健康对照。经患者同意采集的所有样品均经过加拿大多伦多的大学健康网络(University Health Network)和西奈山医院的研究伦理委员会(Research Ethics Board)的机构批准而获得。

[0058] 样本处理-纯化的肿瘤和正常细胞

[0059] 对于原发性PDAC样品,在切除后立即处理样本,并使用代表性切片来确认诊断。在Leica LMD 7000仪器上进行新鲜液氮冷冻组织样品的激光捕获显微切割(LCM)。简言之,将保持在气相液氮中的冷冻组织包埋在OCT切割介质中,并在冷冻切片机中切成8 μ m厚的切片。将切片固定在PEN膜载玻片(Leica)上并且用苏木精轻微染色,以便于显微镜鉴别肿瘤区域。在切割切片的同一天进行LCM以使核酸降解最小化。

[0060] 通过重力将微切割的肿瘤细胞收集到无菌、无RNase的微量离心管的盖中。收集了大约150,000-200,000个肿瘤细胞用于DNA样品并且储存在-80 $^{\circ}$ C直至进一步处理。LCM通常每个病例需要1-2天来收集足够量的纯化肿瘤细胞。使用Qiagen细胞裂解缓冲液来提取基因组DNA。通过将玻璃载玻片上未染色的冷冻切片刮入适当的DNA提取缓冲液中,从冷冻十二指肠或胃粘膜中采集每个患者的匹配的正常、组织学评价的参考组织。

[0061] 样本处理-cfDNA

[0062] EDTA和ACD血浆样品获自BioBank和加拿大多伦多的西奈山医院(MSH)的家庭医学中心。将所有样品储存在-80 $^{\circ}$ C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血浆中提取无细胞的DNA。使用前通过Qubit定量所提取的DNA。

[0063] 样本处理-PDX cfDNA

[0064] 使用胶原酶A将经大学健康网络的研究伦理委员会批准、经大学健康网络Biobank

的患者同意所获得的人类结直肠癌肿瘤组织消化成单细胞。将单细胞皮下注射到4-6周龄的NOD/SCID雄性小鼠中。通过CO₂吸入使小鼠安乐死,之后通过心脏穿刺采集血液并储存在EDTA管中。从采集的血液样品中分离血浆并储存在-80℃。使用QIAamp循环核酸试剂盒(Qiagen)从0.3-0.7ml血浆中提取无细胞的DNA。所有动物工作均遵照大学健康网络的动物护理委员会批准的伦理规范而进行。

[0065] RRBS

[0066] 从来自同一患者(已经获得无细胞的DNA的患者)的富集LCM的肿瘤样品和正常样品提取的基因组DNA按照Gu et al., 2011¹⁸的方案稍作修改进行RRBS。简而言之,使用限制酶MspI消化10ng基因组DNA(通过Qubit确定),随后进行末端修复、A-加尾并且接头(adapter)连接到Illumina TruSeq甲基化的接头。然后使用Zymo EZ DNA甲基化试剂盒按照制造商的方案对制备的文库进行亚硫酸氢盐转化,然后通过凝胶尺寸选择160bp-300bp的片段。通过使用qPCR确定扩增每个纯化文库的最佳循环数,之后使用KAPA HiFi Uracil+ Mastermix (Kapa Biosystems)扩增样品并且使用AMPure珠(Beckman Coulter)进行纯化。对最终的文库进行BioAnalyzer分析,随后在UHN玛格丽特公主基因组中心(UHN Princess Margaret Genomic Centre)在Illumina HiSeq 2000上进行测序。

[0067] 外源性肠杆菌噬菌体λPCR产物的制备

[0068] 使用表1中所示的引物扩增肠杆菌噬菌体(Enterobacteria phage) λDNA (ThermoFischer Scientific),产生6种不同的PCR扩增子产物。按以下条件使用KAPA HiFi Hotstart ReadyMix进行PCR反应:在95℃下激活酶3分钟,98℃ 30个循环,持续20秒、60℃持续15秒、72℃持续30秒和在72℃下最终延伸持续1分钟。使用QIAquick PCR纯化试剂盒(Qiagen)纯化PCR扩增子,在凝胶上运行以验证大小并且扩增。使用CpG甲基移换酶(M.SssI) (ThermoFischer Scientific)将用于1CpG、5CpG、10CpG、15CpG和20CpGL的扩增子甲基化并且使用QIAquick PCR纯化试剂盒进行纯化。使用限制酶HpyCH4IV (New England Biolabs Canada)测试PCR扩增子的甲基化并且在凝胶上运行以确保其甲基化。使用picogreen测量未甲基化的(20CpGS)和甲基化的(1CpG, 5CpG, 10CpG, 15CpG, 20CpGL)扩增子的DNA浓度,随后汇集50%甲基化的和50%未甲基化的λPCR产物。

[0069] cfMeDIP-seq

[0070] 图2示出了cfMeDIP-seq方案的示意图。在cfMeDIP之前,使用Kapa Hyper Prep试剂盒(Kapa Biosystems)将DNA样品进行文库制备。遵循制造商方案,进行一些修改。简而言之,将感兴趣的DNA加入到0.2mL PCR管中并进行末端修复和A-加尾。随后使用NEBNext接头(来自NEBNext Multiplex Oligos的Illumina试剂盒, New England Biolabs)在0.181μM的最终浓度下进行接头连接(Adapter ligation),在20℃下温育20分钟并且用AMPure XP珠进行纯化。在MeDIP之前,使用USER酶(New England Biolabs加拿大)消化洗脱的文库,随后使用Qiagen MinElute PCR纯化试剂盒进行纯化。

[0071] 将制备的文库与汇集的甲基化的/未甲基化的λPCR产物组合至100ng的最终DNA量,并且使用来自Taiwo et al. 2012¹⁷的方案(稍作修改)进行MeDIP。对于MeDIP,使用Diagenode MagMeDIP试剂盒(Cat#C02010021)遵循制造商的方案(作一些修改)。在将0.3ng对照甲基化的和0.3ng对照未甲基化的拟南芥DNA、填充DNA(以将DNA的总量[cfDNA+填充物+对照]完成至100ng)和缓冲液添加到含有接头连接的DNA的PCR管中之后,将样品加热至95

℃保持10分钟,然后立即放入冰水浴中10分钟。将每个样品分入两个0.2mL PCR管:一个用于10%输入对照,另一个用于进行免疫沉淀的样品。将来自MagMeDIP试剂盒包括的5-mC单克隆抗体33D3 (Cat#C15200081) 稀释为1:15,随后产生稀释的抗体混合物并加入到样品中。还加入经洗涤的磁珠(遵循制造商说明),随后在4℃下温育17小时。使用Diagenode iPure试剂盒纯化样品并且在50μl缓冲液C中进行洗脱。通过qPCR验证反应(QC1)的成功,以检测钉入式拟南芥DNA的存在,确保未甲基化的钉入式DNA的%回收<1%并且反应的%特异性>99%(通过1-[钉入式未甲基化的对照DNA的回收率超过钉入式甲基化的对照DNA的回收率]来计算),随后进行下一步。通过使用qPCR确定扩增每个文库的最佳循环数,之后使用添加至0.3μM最终浓度的KAPA HiFi Hotstart Mastermix和NEBNext multiplex oligos扩增样品。用于扩增文库的PCR设置如下:在95℃下激活3分钟,随后进行98℃预定循环,持续20秒、65℃持续15秒、72℃持续30秒和在72℃下最终延伸持续1分钟。使用MinElute PCR纯化柱纯化扩增的文库,然后使用3%Nusieve GTG琼脂糖凝胶进行凝胶尺寸选择以除去任何接头二聚体(adapter dimer)。在进行测序之前,针对由经剪切以模拟无细胞的DNA的HCT116细胞系DNA(从ATCC获得的细胞系,无支原体)产生的MeDIP-seq和cfMeDIP-seq文库,确定甲基化的人类DNA区域(睾丸特异性H2B, TSH2B)和未甲基化的人类DNA区域(GAPDH启动子)的倍数富集。提交最终的文库进行BioAnalyzer分析,随后在UHN玛格丽特公主基因组中心在Illumina HiSeq 2000上进行测序。

[0072] 填充DNA中甲基化的差异%

[0073] 在如下方案的填充组分中使用不同%甲基化的与未甲基化的λDNA进行cfMeDIP-seq:

	%甲基化的填充物	%未甲基化的填充物
[0074] 在免疫沉淀之前使用填充(λ) DNA将DNA的最终量增加到100 ng,无细胞的DNA的起始量范围为10ng、5ng和1ng	100	0
	85	15
	70	30
	50	50
	30	70
	15	85
	0	100
	在免疫沉淀之前未使用填充DNA - 只有10ng、5ng或1ng的起始DNA	

[0075] 如图9和图10所示,在免疫沉淀之前用来将最终量增加到100ng的填充DNA(λDNA)应优选在其组成中具有一些人工甲基化的DNA(100%-15%),以便获得最小回收未甲基化的DNA(图9),同时仍在甲基化的DNA的回收方面获得良好的产量(图10)。在存在100%未甲基化的填充DNA或不存在填充DNA的样品中,虽然甲基化的DNA确实有较高的回收率,但是未甲基化DNA的%回收也较高。这表明,填充DNA中额外的甲基化的DNA有助于占据反应中存在

的过量抗体,最小化与样品中发现的未甲基化DNA的非特异性结合的量。鉴于在使用不同的无细胞的DNA样品的情况下优化抗体量不是非常经济或甚至不可行,因为不知道在整个样品中存在多少甲基化的DNA并且在样品与样品之间这可能会有很大差异,这种填充DNA有助于标准化不同的起始量,并允许以相同的方式(即使用相同量的抗体)处理不同的无细胞的DNA样品,同时仍然从中回收良好的甲基化数据。

[0076] 用于点突变检测的超深靶向测序

[0077] 我们使用QIAgen循环核酸试剂盒从来自具有在参加玛格丽特公主癌症中心(Princess Margaret Cancer Centre)的早期临床试验之前所产生的匹配的肿瘤组织分子谱数据的患者的~20mL血浆(4-5x 10mL EDTA血液管)中分离无细胞的DNA。使用PureGene Genra试剂盒从细胞系(稀释的CRC和MM细胞系)提取DNA,使用Covaris超声波仪片段化为~180bp,并且使用Ampure珠排除更大尺寸的片段以模拟无细胞的DNA的片段尺寸。使用KAPA Hyper Prep试剂盒(Kapa Biosystems,Wilmington,MA)利用NEXTflex-96DNA Barcode接头(Bio Scientific,Austin,TX)接头从83ng片段化的DNA构建DNA测序文库。为了分离含有已知突变的DNA片段,我们设计了生物素化的DNA捕获探针(xGen Lockdown Custom Probes Mini Pool,Integrated DNA Technologies,Coralville,IA),其靶向通过使用Illumina TruSeq Amplicon Cancer Panel的临床实验室测试的48个基因的突变热点。汇集条码文库并且按照制造商的说明应用常规混合捕获文库(IDT xGEN Lockdown方案,版本2.1)。使用Illumina HiSeq 2000仪器将这些片段测序至>10,000X读数(read)覆盖率。使用bwa-mem对比得到的读数并且使用samtools和muTect版本1.1.4检测突变。

[0078] 肿瘤特异性特征的数量与通过测序深度的检测概率之间的模型关系 (Modelling relationships)

[0079] 我们创建了145,000个模拟基因组,癌症特异性甲基化的DMR的比例分别设定为0.001%、0.01%、0.1%、1%和10%,并且分别由1、10、100、1000和10000个独立的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),并且进一步对每个基因座10、100、1000和10000个读数进行取样,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,重复该过程100次。我们估计了对于每个参数组合成功检测至少1个DMR的频率并且绘制了概率曲线(图1中的A),以直观地评估特征数量对成功检测概率的影响(以测序深度为条件)。

[0080] 从胰腺癌患者和健康供体的cfDNA计算和可视化差异甲基化区域

[0081] 使用MEDIPS R包²⁵计算了来自24个胰腺癌(PC)患者和24个健康供体的cfDNA样品之间的差异甲基化的区域(DMR)。对于每个样品,使用BAM比对(相对于人类基因组hg19)文件来创建MEDIPS R对象。随后,通过使用t检验比较来自两组样品的RPKM来计算DMR。使用Benjamini-Hochberg程序调整来自t检验的原始p值。然后将DMR定义为调整后的p值小于0.1的所有窗口;共发现38,085个DMR:胰腺癌患者中有6,651例Hyper(高)和31,544例Hypo(低)。将来自这些DMR的标度的(scale)RPKM值表示为热图(图5中的C)。该热图使用距离函数“欧几里德(euclidean)”制作,聚类函数(clustering function)“ward”用于列式聚类(column-wise clustering)且“平均(average)”用于行式聚类(row-wise clustering)。

[0082] 来自24个胰腺癌组织和5个正常PBMC的RRBS样品的比较

[0083] 从GEO下载了由RRBS配置(profiled)的五个正常PBMC样本(所以对照样品均是

Accession ID GSE89473下的样品),以将它们的甲基化谱与24个胰腺癌组织RRBS样品的甲基化谱进行比较。使用R methylKit包²⁶解析和处理下载的床文件 (bed file)。接下来将这五个样品与来自24位胰腺癌患者的类似处理的RRBS样品进行比较。使用自定义函数来提取存在于24个PC样品中的至少18个中,以及5个PBMC样品中的4个中的CpG,并且仅保留常染色体中的CpG,以产生1,806,808个CpG的背景组。其中,使用调整的 p 值 <0.01 和 $\Delta\beta>0.25$ 的Benjamini-Hochberg标准获得DMC,并且与PBMC相比,发现在胰腺癌中有134,021个DMC是Hyper (高)。类似地,使用相同的 q 值截止值和 $\Delta\beta<0.25$,我们获得了179,662个Hypo (低) DMC。在相应火山图中由红点表示总共313,683个DMC (图7中的F),其中相对于 $\Delta\beta$ 对 q 值的负 \log_{10} 作图 (负 $\log_{10}q$ 值=2的水平线表示调用DMC的 q 值截止值,并且虚线垂直线表示 $\Delta\beta$ 截止值)。

[0084] 来自原发性肿瘤相对于正常PBMC以及来自胰腺癌患者和健康供体的cfDNA的差异甲基化信号的重叠的评估

[0085] 进行排列分析 (Permutation analysis) 以比较在血浆中鉴别的DMR (使循环cfDNA经受我们的cfMeDIP-seq方案) 与在原发肿瘤组织中鉴别的癌症特异性DMC (使用RRBS) 之间的预期重叠和观察到的重叠的频率。我们研究了四种可能的情况:Hyper DMC与Hyper DMR重叠,Hyper DMC与Hypo DMR重叠,Hypo DMC与Hypo DMR重叠,和最后Hypo DMC与Hyper DMR重叠。对于每种情况,Hyper或Hypo DMC与Hyper或Hypo DMR重叠以获得“生物交叉点”的数量;然后将每组DMC随机洗牌到1,806,808个CpG的背景组中1000次,并与各组的DMR再次重叠。使用Z分数将这些随机和生物交叉点以相同的尺度放置,并分别用箱图 (boxplot) 和菱形显示 (图5中的E)。这些图中的虚线水平线表示与Bonferroni调整导出的 q 值0.05相关的截止Z分数。

[0086] 来自24个胰腺癌组织和24个正常组织的RRBS样品的比较以及来自这些组织和来自胰腺癌患者和健康供体的cfDNA的差异甲基化信号重叠的评估

[0087] 将与5个正常PBMC样品比较的24个PC样品再分别与来自相同患者的24个正常组织进行比较。使用相同的方法计算背景组 (763,874个CpG) 以及PC中的DMC Hyper和DMC Hypo (分别为34,013和11,160),并且还将其用来以相同的方式构建火山图 (图7中的C) 和箱图 (图5中的D)。

[0088] 对24个PC和24个健康cfDNA样品进行PCA绘图

[0089] 我们使用顶部百万个最可变的全基因组窗口以24个PC和24个健康cfDNA样品上的PCA (图7中的A-B) 进行了无监督聚类分析。对于每个窗口,使用MAD (平均绝对偏差) 度量来计算可变性。其是一个稳健的测量,它返回来自数据中值的绝对偏差的中值,其中所述数据是给定窗口的这些48个样品的RPKM值。

[0090] 具有与24个PC和24个健康cfDNA样品中低甲基化的基序相关的TF的GTEx表达谱的热图,

[0091] RNA-Seq数据从GTEx数据库以通过用于所有人类基因的组织的中值RPKM的形式获得 (获自<https://gtexportal.org/home/datasets>下的文件GTEx_Analysis_v6p_RNA-seq-RNA-SeQCv1.1.8_gene_median_rpk.m.gct.gz)。将感兴趣的TF与它们的基因名称匹配,并且使用在所有组织中评分的每个TF的中值RPKM构建热图 (图8中的A,C)。距离函数“曼哈顿 (manhattan)”和聚类函数“平均 (average)”用于行式和列式聚类。

[0092] 具有与24个PC和24个健康cfDNA样品中低甲基化的基序相关的TF的GTE_x表达谱的小提琴图(Violin plot)

[0093] 为了估计我们在病例与对照中检测到低甲基化区域中显著富集基序的TF是否在胰腺癌样品中显著上调,我们使用了随机化测试,将ssGSEA评分作为检验统计。对于每个样品,我们使用与低甲基化基序显著相关的85个TF以及85个TF的1,000个随机组来计算得分(所有人类TF的列表均获自<http://www.tfcheckpoint.org/data/>的文件TFCheckpoint_download_180515.txt);采用TCGA上178例胰腺腺癌患者的表达水平。

[0094] 可以在相关的小提琴图(图8中的E)中看到这些得分的分布。

[0095] 然后使用Wilcoxon秩和检验(Wilcoxon's Rank Sum test)来比较随机分布与观察到的分布,得到的p值 $<2.2e-16$ 。

[0096] 对正常胰腺的GTE_x数据进行了相同的分析(图8中的D)。还用TF($n=33$)在全血的GTE_x数据上重复了所述分析,所述TF的基序在健康供体的血浆cfDNA中被鉴别为低甲基化足迹(图8中的B)。

[0097] 结果/讨论

[0098] 适用于cfDNA甲基化作图的全基因组方法

[0099] 这里描述的cfMeDIP-seq方法是通过修改现有的低输入MeDIP-seq方案¹⁷(该方案可以稳定到100ng的输入DNA)开发的。然而,大多数血浆样品产生的DNA远低于100ng。为了克服这一挑战,我们将外源 λ DNA(填充DNA)添加到接头连接的cfDNA文库,以将起始DNA的量人工膨胀到100ng(图2)。这最小化了抗体的非特异性结合的量,并且还最小化了由于与塑料器皿结合而损失的DNA量。填充DNA由与接头连接的cfDNA文库大小相似的扩增子组成,并且由不同CpG密度的未甲基化和体外甲基化DNA组成。因为不同的患者会产生不同量的cfDNA,添加这种填充DNA也具有实际用途,允许使输入DNA的量标准化为100ng。这确保了下游方案对于所有样品保持完全相同,而不管可用的cfDNA的量是多少。

[0100] 我们首先使用来自人类结直肠癌细胞系HCT116的DNA验证了cfMeDIP-seq方案,所述DNA被剪切成与cfDNA中观察到的相似的片段大小。选择HCT116是因为公共DNA甲基化数据的可用性。我们使用100ng剪切的细胞系DNA进行了金标准MeDIP-seq方案¹⁷,并且同时使用10ng、5ng和1ng相同剪切的细胞系DNA进行了cfMeDIP-seq方案。对此进行了两个生物学重复。对于所有条件,我们获得了超过99%的反应特异性($1-[$ 钉入式未甲基化的对照DNA的回收率超过钉入式甲基化的对照DNA的回收率]),并且获得了已知甲基化区域超过未甲基化区域的非常高的富集(分别为TSH2B0和GAPDH)(图3中的B)。

[0101] 将文库测序至饱和(图3中的A),每个文库大约3千万到7千万个读数(表2)。将原始读数与人类基因组和 λ 基因组进行比对,发现实际上没有发现与 λ 基因组的比对(表3A和3B)。因此,添加外源 λ DNA作为填充DNA不会干扰测序数据的产生。最后,我们计算CpG富集评分作为免疫沉淀步骤的质量控制措施²⁵。所有文库显示出类似的CpG富集,而输入对照如预期的那样没有显示富集(图3中的C),即使在极低的输入(1ng)下也能验证我们的免疫沉淀。

[0102] 比较不同输入DNA水平的全基因组相关性估计显示,MeDIP-seq(100ng)和cfMeDIP-seq(10、5和1ng)方法都非常稳健,在任何两个生物重复之间的Pearson相关性至少为0.94(图1中的B)。该分析还表明,在5和10ng输入DNA的cfMeDIP-seq可以强有力地再现传统MeDIP-seq在100ng时获得的甲基化谱(成对Pearson相关性(Pairwise Pearson

correlation) 至少为0.9) (图1中的B)。与在100ng的MeDIP-seq相比,在1ng输入DNA的cfMeDIP-seq性能降低,但仍然显示出 >0.7 的强Pearson相关性(图1中的B)。我们还观察到,cfMeDIP-seq方案再现了使用金标准RRBS(简化表示亚硫酸氢盐测序(Reduced Representation Bisulfite Sequencing))和WGBS(全基因组亚硫酸氢盐测序(Whole-Genome Bisulfite Sequencing))的HCT116的DNA甲基化谱(图1中的C)。总之,我们的数据表明,cfMeDIP-seq是用于片段化和低输入DNA材料(例如循环cfDNA)的全基因组甲基化作图的稳健方案。

[0103] cfMeDIP-seq对肿瘤来源ctDNA的检测表现出高灵敏度

[0104] 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连续稀释到多发性骨髓瘤(MM)MM1.S细胞系DNA中,两者都被剪切以模拟cfDNA大小。我们稀释CRC DNA从100%、10%、1%、0.1%、0.01%、0.001%到0%并且在这些稀释液的每一个上进行cfMeDIP-seq(图4中的A-D)。我们还进行了超深(10,000X中值覆盖率)靶向测序,用于检测相同样品中的三个点突变。使用5%假发现率(FDR)阈值在每个CRC稀释点处相对于纯MM DNA观察到的鉴别的DMR数量与基于低至0.001%稀释的稀释因子的预期DMR数量几乎完全呈线性关系($r^2=0.99, p<0.0001$) (图1中的D)。另外,这些DMR中的DNA甲基化信号在观察信号与预期信号之间也显示出几乎完美的线性($r^2=0.99, p<0.0001$) (图1中的E)。相比之下,超过1%稀释时,超深靶向测序无法可靠地区分CRC特异性变体和由于PCR或测序误差引起的假变体(图1中的F)。因此,cfMeDIP-seq对癌症来源的DNA的检测表现出优异的灵敏度,超过了通过使用标准方案进行的超深靶向测序的变体检测性能。

[0105] 癌症DNA经常在富含CpG的区域高度甲基化¹。由于cfMeDIP-seq特异性地靶向富含甲基化的CpG的序列,我们假设在免疫沉淀过程中ctDNA会被优先富集。为了测试这一点,我们从两名结直肠癌患者产生了患者来源的异种移植物(PDX),并收集了小鼠血浆。肿瘤来源的人类cfDNA在输入样品的总cfDNA库中以低于1%的频率存在,并且在免疫沉淀后以2倍更大的丰度存在(图1中的G)。这些结果表明,通过ctDNA的偏置测序,cfMeDIP程序可以进一步提高ctDNA检测灵敏度。

[0106] 血浆cfDNA的甲基化分析区分早期胰腺腺癌患者与健康供体

[0107] 我们试图研究血浆cfDNA的甲基化组分析是否可用于检测早期癌症中的ctDNA。我们在24名早期胰腺癌患者(病例)和24名年龄和性别匹配的健康供体(对照)的手术前血浆中进行了甲基化组分析(表4A、4B和5)。对于每位患者,检查了具有高肿瘤纯度的激光捕获显微切割(LCM)肿瘤样品和正常组织样品。在循环cfDNA上进行cfMeDIP-seq并且在肿瘤和正常组织上进行RRBS(图5中的A和图6,表6A和6B)。使用t检验和Benjamini-Hochberg校正进行多次测试,我们在病例和对照cfDNA之间获得了38,085个DMR($p<0.01, q<0.1$) (图5中的B-C)。

[0108] 为了评估病例和对照之间cfDNA甲基化谱的差异是否是由于ctDNA的存在引起的,使用RRBS对原发肿瘤和正常组织(在手术切除后从同一患者获得的)的DNA甲基化模式进行作图。我们在肿瘤($n=24$)与正常($n=24$)组织之间鉴别了45,173个差异甲基化的CpG(DMC)(图7中的A-C)。

[0109] 通过检查肿瘤中的DMC和cfDNA中的DMR的组合(两者中的高甲基化,两者中的低甲基化,一方中的高甲基化和另一方中的低甲基化),对于相对于背景的富集,测试了cfDNA甲

甲基化谱在重现其原始肿瘤的甲基化谱中的应用。我们在cfDNA中观察到肿瘤特异性高甲基化和低甲基化位点在一致方向上的显著富集,而肿瘤特异性高甲基化位点在cfDNA低甲基化DMR中出现较少(under-represented)(图5中的D)。实际上,肿瘤中给定区域的DNA甲基化状态与血浆cfDNA中的甲基化谱之间存在相关性(图7中的D-E)。

[0110] 最后,因为癌症患者(特别是在早期阶段)中的大多数血浆cfDNA分子是非肿瘤来源的并且可能由血细胞释放¹⁴,我们评估了胰腺腺癌肿瘤组织与正常外周血单核细胞(PBMC)之间的DNA甲基化差异。我们在肿瘤(n=24)与PBMC(n=5)之间鉴别了313,683个DMC(图7中的F)。我们在cfDNA中观察到肿瘤特异性高甲基化和低甲基化位点一致方向上的显著富集,而肿瘤特异性高甲基化位点在cfDNA低甲基化DMR中出现较少(图5中的E)。同样,肿瘤中给定区域的DNA甲基化状态与血浆cfDNA中的甲基化谱之间存在相关性(图7中的G-H)。

[0111] 总之,这些结果表明,病例和对照之间循环cfDNA甲基化谱的差异很大程度上是由于循环系统中存在肿瘤来源的DNA引起的(图5中的D-E和图7中的C-H)。

[0112] 血浆cfDNA甲基化组允许推断肿瘤相关的活性转录因子网络

[0113] 由于对于肿瘤来源的DMR而言,病例和对照之间的DMR是高度富集的(图5中的D-E),我们假设cfDNA甲基化组将揭示与肿瘤特异性或组织相关的活性转录因子相关的基序的富集。这些cfDNA甲基化组可用于推断这些DNA分子的起源组织中的活性转录网络。为了推断活性转录网络,我们研究了cfDNA中的DMR是否能够揭露转录因子(TF)足迹的富集,因为大多数TF表现出基于靶序列的DNA甲基化状态的可变结合²⁸。使用HOMER软件²⁰分别对健康供体(图8中的A)和胰腺癌患者(图8中的C)的低甲基化DMR²⁰进行基序分析,以揭露潜在的TF足迹。

[0114] 与胰腺腺癌病例相比,我们在健康供体中鉴别出33个基序作为低甲基化足迹,并且与健康供体相比,我们在胰腺腺癌病例中鉴别出85个基序作为低甲基化足迹。

[0115] 在健康供体中被鉴别为低甲基化足迹的33个基序中,我们鉴别出几种优先在造血谱系中表达的TF,包括PU.1、Flil1、STAT5B和KLF1(图8中的A-B)。

[0116] 类似地,在胰腺腺癌病例中被鉴别为低甲基化足迹的85个基序中,我们鉴别了多种优先在胰腺中表达的TF,包括RBPJL、PTF1a、Onecut1(HNF6)和NR5A2(图8中的C-D)。在胰腺腺癌病例中被鉴别为低甲基化足迹的TF基序也经常来自TCGA的胰腺腺癌患者中过表达(图8中的E)。另外,我们能够鉴别胰腺腺癌病例中的多个低甲基化足迹,其对应于之前被鉴别为胰腺癌的各分子亚型的驱动因子(driver)的TF²⁴。它们包括c-MYC和HIF1a(鳞状亚型驱动因子)、NR5A2、MAFA、RBPJL和NEUROD1(ADEX驱动因子)以及最后FOXA2和HNF4A(胰祖细胞亚型)。

[0117] 总之,这些结果表明,基于差异甲基化的TF足迹,循环cfDNA的甲基化组分析可用于推断肿瘤内的活性转录网络,并可能鉴别健康供体和癌症患者之间免疫细胞群体的系统性变化。

[0118] 在此,我们提出了一种新的全基因组DNA甲基化方法,其适用于超低输入和片段化DNA,如循环无细胞的DNA。我们能够证明cfMeDIP-seq在低水平的输入DNA下非常稳健,并且允许快速生成文库。此外,由于我们的方法依赖于甲基化DNA的富集,因此为了对文库序列至饱和,每个文库仅需要大约3千万到7千万个读数(read),这使得不需要全基因组测序并显著降低了相关成本。除了相对较小的成本之外,快速周转时间(turnaround time)可以允

许将cfMeDIP-seq快速平移(translation)至临床环境。

[0119] 另外,因为cfMeDIP-seq依赖于表观遗传而不是基因组信息,其可潜在地被用于非侵入性地监测一系列广泛的非恶性疾病中的组织损伤。例如,它可用于监测对感染或癌症免疫疗法后的免疫应答;它可用于监测心肌梗死后循环中的心脏DNA或神经退行性疾病早期的脑DNA。

[0120] 最后,在肿瘤学的背景下,多种癌症类型已显示具有临床上不同的亚组。这些亚组可以通过不同的DNA甲基化谱进行分层,在胶质母细胞瘤³、室管膜瘤⁴、结肠直肠癌⁵、乳腺癌^{6,7}和胰腺癌²⁴(在众多其他癌症类型中)具有预后价值。最近的数据表明,胰腺癌患者可以分为由多种机制驱动四个亚组²⁴:鳞状,胰腺祖细胞,免疫原性和异常分化的内分泌外分泌(ADEX)。在胰腺癌患者的循环cfDNA甲基化组中,我们能够从驱动这些亚型的TF中鉴别出低甲基化足迹。例如,我们鉴别了MYC和HIF1 α (缺氧诱导因子1- α)(鳞状亚型中富含的两种路径²⁴)。我们还能够鉴别HNF4A和FOXA2(祖细胞亚型中富含的两种TF²⁴)。最后,我们能够鉴别NR5A2、RBPJL和MAFA(ADEX亚型中富含的三种TF²⁴)。这表明,cfMeDIP-seq也可用作生物标志物,以微创方法对癌症患者进行分层。

[0121] 已经参考具体实施方式描述了本发明。本领域技术人员应当清楚的是,在保持在本发明的精神和范围内的同时可以进行变化和改变。本文公开的具体实施方式不旨在限制保护范围,保护范围应仅由权利要求确定。本文公开的所有出版物和参考文献均通过引用以它们整体并入。

[0122] 表

[0123] 表1:用于产生来自Taiwo et al.,2012的肠杆菌噬菌体 λ PCR产物的PCR引物

[0124]

名称	正向引物	反向引物
1CpG	GAGGTGATAAAAATTAAGTGC	GGCTCTACCATATCTCCTA
5CpG	CATGTCCAGAGCTCATTC	GTTTAAAATCACTAGGCGA
10CpG	CTGACCATTTCCATCATTC	GTAACATAACAGGAGCCG
15CpG	ATGTATCCATTGAGCATTGCC	CACGAATCAGCGGTAAAGGT
20CpGL	GAGATATGGTAGAGCCGCAGA	TTTCAGCAGCTACAGTCAGAATTT
20CpGS	CGATGGGTAAATTCGCTCGTTGTGG	GCACAACGGAAAGAGCACTG

[0125] 表2:使用各种浓度的经剪切以模拟cfDNA的HCT116细胞系DNA制备的测序MeDIP-seq和cfMeDIP-seq文库的读数数量和作图效率(对于人类Hg19基因组和 λ 基因组)。对于低于100ng的起始浓度,在MeDIP之前用外部 λ DNA充满(topup)样品以将起始量人工增加至100ng。

[0126]

样品	原始读数#	与人类基因组(Hg19)比对的读数#	针对人类基因组(Hg19)的作图效率	与λ基因组比对的读数#	针对λ基因组的作图效率
输入	74,504,053	71,343,168	95.76	12	0.00
100 ng 重复 (replicate) 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

[0127] 表3A:使用来自胰腺癌患者的cfDNA制备的测序cfMeDIP-seq文库的读数数量和作图效率(对于人类Hg19基因组和λ基因组)。

[0128]

病例	原始读数#	比对读数#	针对人类基因组(Hg19)的作图效率	与λ基因组比对的读数#	针对λ基因组的作图效率
1	49,970,366	46,120,982	92.30	1	0.00
2	55,642,277	50,829,279	91.35	0	0.00
3	48,034,998	43,973,048	91.54	0	0.00
4	56,274,788	51,456,978	91.44	0	0.00
5	50,999,157	46,608,023	91.39	1	0.00
6	55,127,192	50,906,050	92.34	0	0.00
7	53,955,550	49,028,646	90.87	0	0.00

[0129]

8	55,482,420	51,194,364	92.27	1	0.00
9	57,626,914	53,446,965	92.75	0	0.00
10	58,397,946	53,904,793	92.31	0	0.00
11	59,048,723	54,529,214	92.35	0	0.00
12	54,514,051	49,619,174	91.02	1	0.00
13	70,758,028	63,842,005	90.23	0	0.00
14	60,317,171	55,124,638	91.39	0	0.00
15	45,187,670	41,341,336	91.49	0	0.00
16	50,755,880	46,309,615	91.24	2	0.00
17	51,582,370	47,256,012	91.61	1	0.00
18	52,081,710	47,763,766	91.71	0	0.00
19	53,845,426	48,602,035	90.26	0	0.00
20	48,133,619	43,658,379	90.70	0	0.00
21	54,237,903	49,286,048	90.87	0	0.00
22	58,627,625	53,654,439	91.52	1	0.00
23	52,748,183	49,179,765	93.23	0	0.00
24	60,072,026	55,046,660	91.63	0	0.00

[0130] 表3B: 使用来自健康供体的cfDNA制备的测序cfMeDIP-seq文库的读数数量和作图效率(对于人类Hg19基因组和λ基因组)。

[0131]

对照	原始读数 #	比对读数 #	针对人类基因组(Hg19)的作图效率	与λ基因组比对的读数#	针对λ基因组的作图效率
1	44,162,563	40,184,943	90.99	1	0.00
2	33,350,539	30,585,488	91.71	0	0.00
3	43,937,337	40,160,512	91.40	0	0.00
4	44,292,409	40,182,945	90.72	0	0.00
5	54,732,283	50,135,420	91.60	0	0.00
6	52,576,198	48,139,158	91.56	0	0.00
7	47,619,930	42,684,336	89.64	0	0.00
8	51,213,827	46,529,314	90.85	0	0.00

[0132]

9	48,935,317	44,487,195	90.91	1	0.00
10	56,055,796	51,971,671	92.71	0	0.00
11	54,769,256	50,075,252	91.43	0	0.00
12	50,625,023	45,869,378	90.61	1	0.00
13	51,015,277	46,252,402	90.66	0	0.00
14	58,266,281	53,191,317	91.29	1	0.00
15	55,663,935	50,447,271	90.63	0	0.00
16	44,004,381	39,533,012	89.84	1	0.00
17	55,459,971	50,786,298	91.57	0	0.00
18	52,610,463	48,293,244	91.79	0	0.00
19	46,806,153	42,677,554	91.18	0	0.00
20	63,267,324	57,692,667	91.19	0	0.00
21	51,213,236	46,637,533	91.07	0	0.00
22	54,380,628	49,267,997	90.60	0	0.00
23	44,608,650	40,685,531	91.21	0	0.00
24	52,118,117	47,734,519	91.59	0	0.00

[0133] 表4A:胰腺癌患者的相应性别和年龄

[0134]

病例	性别(1 = 男性, 2 = 女性)	年龄范围
1	1	70<75
2	2	65<70
3	1	50<55
4	1	65<70
5	2	60<65
6	2	80<85
7	1	60<65
8	2	70<75
9	2	60<65
10	1	50<55
11	1	70<75

[0135]	12	2	50<55
	13	1	60<65
	14	2	75<80
	15	2	55<60
	16	2	80<85
	17	1	55<60
	18	2	55<60
	19	2	65<70
	20	2	60<65
	21	1	65<70
	22	1	65<70
	23	2	<50
	24	1	65<70

[0136] 表4B: 健康供体的相应性别和年龄

对照	性别(1 = 男性, 2 = 女性)	年龄范围	
[0137]	1	1	60<65
	2	2	55<60
	3	1	70<75
	4	1	85+
	5	2	80<85
	6	2	<50
	7	1	80<85
	8	2	60<65
	9	2	85+
	10	1	60<65
	11	1	70<75
	12	2	55<60
	13	1	50<55

[0138]	14	2	70<75
	15	2	70<75
	16	2	80<85
	17	1	50<55
	18	2	70<75
	19	2	70<75
	20	2	70<75
	21	1	75<80
	22	1	<50
	23	2	60<65
	24	1	80<85

[0139] 表5:胰腺病例样品的腺癌的病理

[0140]

病理分期	病例数
I-II	23
III-IV	1

[0141] 表6A:使用来自胰腺癌患者的具有高肿瘤纯度的激光捕获显微切割富集 (LCM) 肿瘤样品制备的测序RRBS文库的读数数量、作图效率、亚硫酸氢盐转化效率和CpG覆盖率。

[0142]

来自病例的肿瘤	原始读数 #	与人类基因组 (Hg19)比对的读数#	针对人类基因组 (Hg19)的作图效率	亚硫酸氢盐转化效率	最少覆盖 10 个读数的 CpG 的#
1	36,348,593	22,050,682	60.7	99.65	1,152,313
2	56,788,591	35,512,606	62.5	98.29	1,790,326
3	55,115,742	35,229,781	63.9	98.94	1,630,272
4	58,503,145	37,373,806	63.9	99.48	1,777,824
5	54,364,708	34,274,401	63.0	98.84	2,053,577
6	57,291,238	34,912,515	60.9	99.23	2,001,726

[0143]

7	42,683,835	25,509,282	59.8	97.79	1,471,270
8	51,393,779	31,437,264	61.2	99.52	1,745,355
9	66,023,766	41,284,714	62.5	99.17	1,788,895
10	37,026,164	20,805,815	56.2	98.67	1,435,724
11	40,053,025	22,116,738	55.2	99.47	1,521,135
12	42,065,012	23,591,687	56.1	98.22	1,599,425
13	34,527,362	22,101,495	64.0	99.79	733,408
14	38,624,627	23,317,534	60.4	99.82	952,714
15	36,799,407	22,072,921	60.0	98.45	1,442,393
16	33,672,535	19,735,978	58.6	98.70	1,064,718
17	51,918,938	32,239,409	62.1	99.08	1,810,260
18	26,344,807	15,531,839	59.0	99.87	792,650
19	43,744,321	26,819,066	61.3	98.45	1,510,987
20	49,861,059	28,520,788	57.2	97.75	1,852,841
21	44,061,826	27,903,839	63.3	99.68	1,368,522
22	55,029,698	33,444,993	60.8	99.10	2,006,233
23	70,468,323	45,424,428	64.5	97.86	2,331,030
24	42,043,232	24,828,943	59.1	97.44	1,520,504

[0144] 表6B: 使用来自胰腺癌患者健康供体的激光捕获显微切割富集 (LCM) 正常组织样品制备的测序RRBS文库的读数数量、作图效率、亚硫酸氢盐转化效率和CpG覆盖率。

[0145]

来自病例的正常组织	原始读数 #	与人类基因组 (Hg19) 比对的读数 #	针对人类基因组 (Hg19) 的作图效率	亚硫酸氢盐转化效率	最少覆盖 10 个读数的 CpG#
1	44,991,458	25,708,044	57.1	98.88	1,688,179
2	50,603,038	30,001,913	59.3	99.81	1,767,046
3	50,790,690	31,585,415	62.2	99.61	1,733,966
4	49,480,382	30,017,286	60.7	99.60	1,895,462
5	46,587,496	29,742,054	63.8	98.94	1,648,785
6	55,402,162	33,193,299	59.9	99.10	2,149,673

[0146]

7	35,010,018	20,646,990	59.0	99.83	1,188,194
8	51,130,341	30,543,069	59.7	99.64	1,780,696
9	53,415,206	32,049,549	60.0	99.63	1,720,213
10	40,977,129	22,109,663	54.0	99.60	1,690,112
11	37,913,346	20,202,691	53.3	98.35	638,859
12	47,943,388	26,962,401	56.2	97.74	1,467,609
13	28,416,315	16,532,481	58.2	98.36	1,214,093
14	34,809,298	20,603,577	59.2	99.26	1,105,920
15	36,930,467	21,513,886	58.3	99.76	1,434,653
16	38,762,192	23,242,114	60.0	99.79	1,539,278
17	51,257,989	29,870,486	58.3	99.66	1,915,596
18	28,789,569	17,038,843	59.2	98.21	1,222,106
19	41,387,115	25,996,349	62.8	98.57	1,712,186
20	31,704,499	17,421,599	54.9	99.41	977,256
21	44,152,396	26,604,688	60.3	99.68	1,646,272
22	46,462,469	27,012,950	58.1	99.43	1,940,763
23	53,534,041	33,359,195	62.3	99.50	1,650,915
24	37,042,798	22,775,661	61.5	99.31	435,114

[0147] 参考文献

[0148] 1 Sharma,S.,Kelly,T.K.&Jones,P.A.Epigenetics in caneer.Carcinogenesis 31,27-36,doi:10.1093/carcin/bgp220 (2010) .

[0149] 2 Varley,,K.E.et al.Dynamic DNA methylation across diverse human cell lines and tissues.Genome Res 23,555-567,doi:10.1101/gr.147942.112 (2013) .

[0150] 3 Sturm,D.et al.Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma.Cancer Cell22,425-437, doi:10.1016/j.ccr.2012.08.024 (2012) .

[0151] 4 Mack,S.C.et al.Epigenomic alterations define lethal CIMP-positive ependymomas of infancy.Nature 506,445-450,doi:10.1038/nature13108 (2014) .

[0152] 5 Hinoue,T.et al.Genome-scale analysis of aberrant DNA methylation in colorectal cancer.GenomeRes 22,271-282,doi:10.1101/gr.117523.110 (2012) .

[0153] 6 Stirzaker,C.et al.Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value.Nat Commun 6,5899,doi:10.1038/ncomms6899 (2015) .

- [0154] 7 Fang, F. et al. Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci Transl Med* 3, 75ra25, doi:10.1126/scitranslmed.3001875 (2011).
- [0155] 8 Mikeska, T. & Craig, J. M. DNA methylation biomarkers: cancer and beyond. *Genes (Basel)* 5, 821–864, doi:10.3390/genes5030821 (2014).
- [0156] 9 Diaz, L. A., Jr. & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* 32, 579–586, doi:10.1200/JCO.2012.45.2011 (2014).
- [0157] 10 Snyder, T. M., Khush, K. K., Valentine, H. A. & Quake, S. R. Universal noninvasive detection of solid organ transplant rejection. *Proc Natl Acad Sci U S A* 108, 6229–6234, doi:10.1073/pnas.1013924108 (2011).
- [0158] 11 Chiu, R. W. et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci USA* 105, 20458–20463, doi:10.1073/pnas.0810641105 (2008).
- [0159] 12 Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 105, 16266–16271, doi:10.1073/pnas.0808319105 (2008).
- [0160] 13 Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* 20, 548–554, doi:10.1038/nm.3519 (2014).
- [0161] 14 Aravanis, A. M., Lee, M. & Klausner, R. D. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* 168, 571–574, doi:10.1016/j.cell.2017.01.030 (2017).
- [0162] 15 Hoadley, K. A. et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929–944, doi:10.1016/j.cell.2014.06.049 (2014).
- [0163] 16 Fleischhacker, M. & Schmidt, B. Circulating nucleic acids (CNAs) and cancer—a survey. *Biochim Biophys Acta* 1775, 181–232, doi:10.1016/j.bbcan.2006.10001 (2007).
- [0164] 17 Taiwo, O. et al. Methylome analysis using MeDIP-seq with low DNA concentrations. *Nat Protoc* 7, 617–636, doi:10.1038/nprot.2012.012 (2012).
- [0165] 18 Gu, H. et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* 6, 468–481, doi:10.1038/nprot.2010.190 (2011).
- [0166] 19 Hung, E. C., Chiu, R. W. & Lo, Y. M. Detection of circulating fetal nucleic acids: a review of methods and applications. *J Clin Pathol* 62, 308–313, doi:10.1136/jcp.2007.048470 (2009).
- [0167] 20 Heinz, S. et al. Simple combinations of lineage-determining

transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38,576–589,doi:10.1016/j.molcel.2010.05.004 (2010) .

[0168] 21 Consortium,G.T.Human genomics.The Genotype–Tissue Expression (GTEx) pilot analysis:multitissue gene regulation in humans.*Science* 348,648–660,doi:10.1126/science.1262110 (2015) .

[0169] 22 Su,A.l.et al.A gene atlas of the mouse and human protein-encodingtranscriptomes.*Proc Natl Acad Sci USA* 101,6062–6067,doi:10.1073/pnas.0400782101 (2004) .

[0170] 23 Wu,C.,Jin,X.,Tsueng,G.,Afrasiabi,C.&Su,A.I.BioGPS:building your ownmash-up of gene 8notations and expression profiles.*Nucleic Acids Res* 44, D313–316,doi:10.1093/nar/gkv1104 (2016) .

[0171] 24 Bailey,P.et al.Genomic analyses identify molecular subtypes of pancreatic cancer.*Nature*,doi:10.1038/nature16965 (2016) .

[0172] 25 Lienhard,M.,Grimm,C-,Morkel,M.,Herwig,R.&Chavez,L.MEDIPS:genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments.*Bioinformatics* 30,284–286,doi:10.1093/bioinformatics/btt650 (2014) .

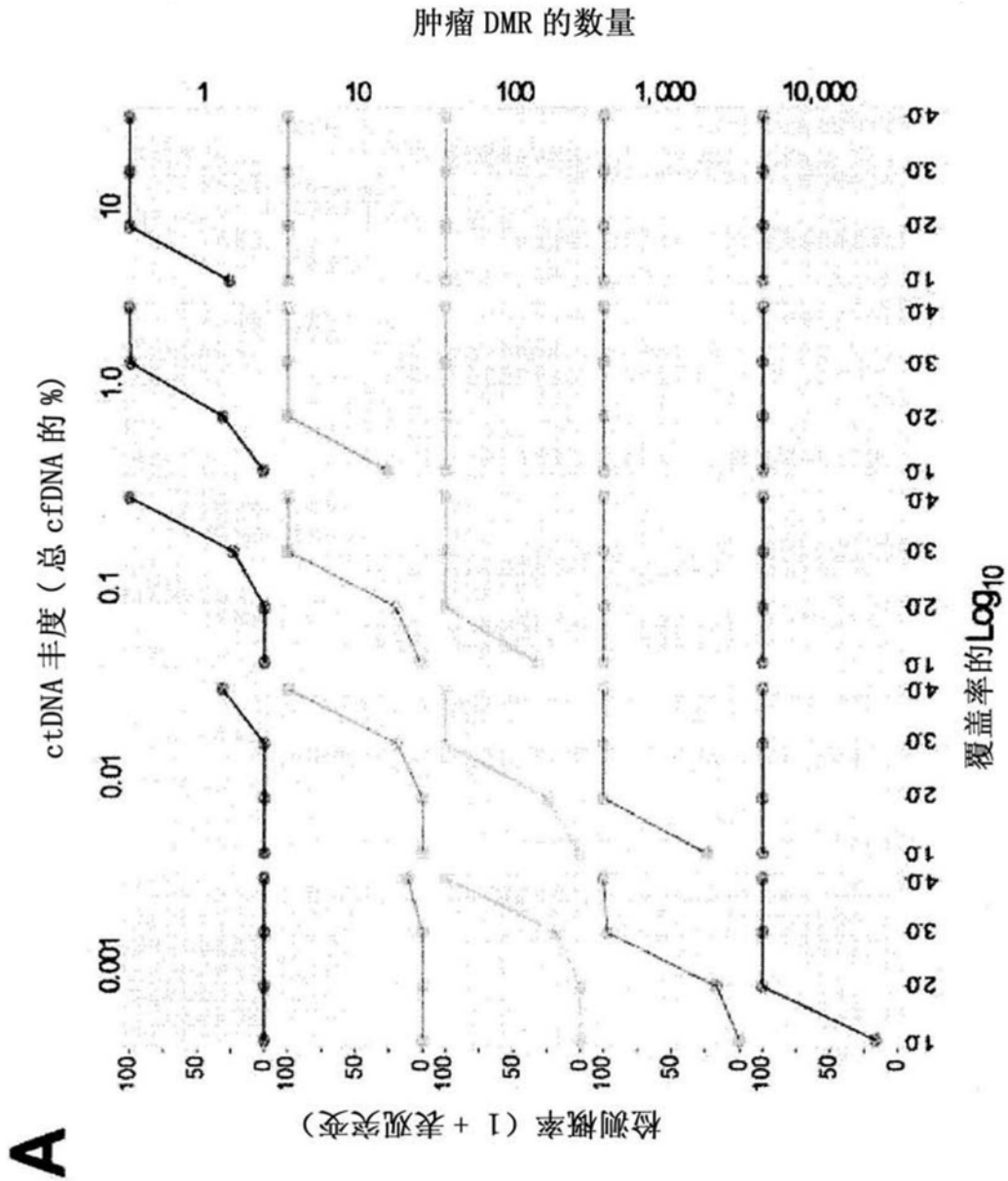
[0173] 26 Akalin,A.et al,methylKit:a comprehensive R package for the analysis of genome-wide DNA methylation profiles.*Genome Biol* 13,R87,doi:10.1186/gb-2012-13-10-r87 (2012) .

[0174] 27 Krueger,F.&Andrews,S.R.Bismark:a flexible aligner and methylation caller for Bisulfite-Seq applications.*Bioinformatics* 27,1571–1572,doi:10.1093/bioinformatics/btr167 (2011) .

[0175] 28 Hu,S.et al.DNA methylation presents distinct binding sites for human transcription factors.*Elife* 2,e00726,doi:10.7554/eLife.00726 (2013) .

[0176] 29 Lui,Y.Y.et al.Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation.*Clin Chem* 48,421–427 (2002) .

[0177] 30 Snyder,M.W.,Kircher,M.,Hill,A.J.,Daza,R.M.&Shendure,J.Cell-free DNA Compnses an In Vivo Nucleosome Footprint that Informs lts Tissues-Of-Origin.*Cell* 164,57–68,doi:10.1016/j.cell.2015.11.050 (2016) .

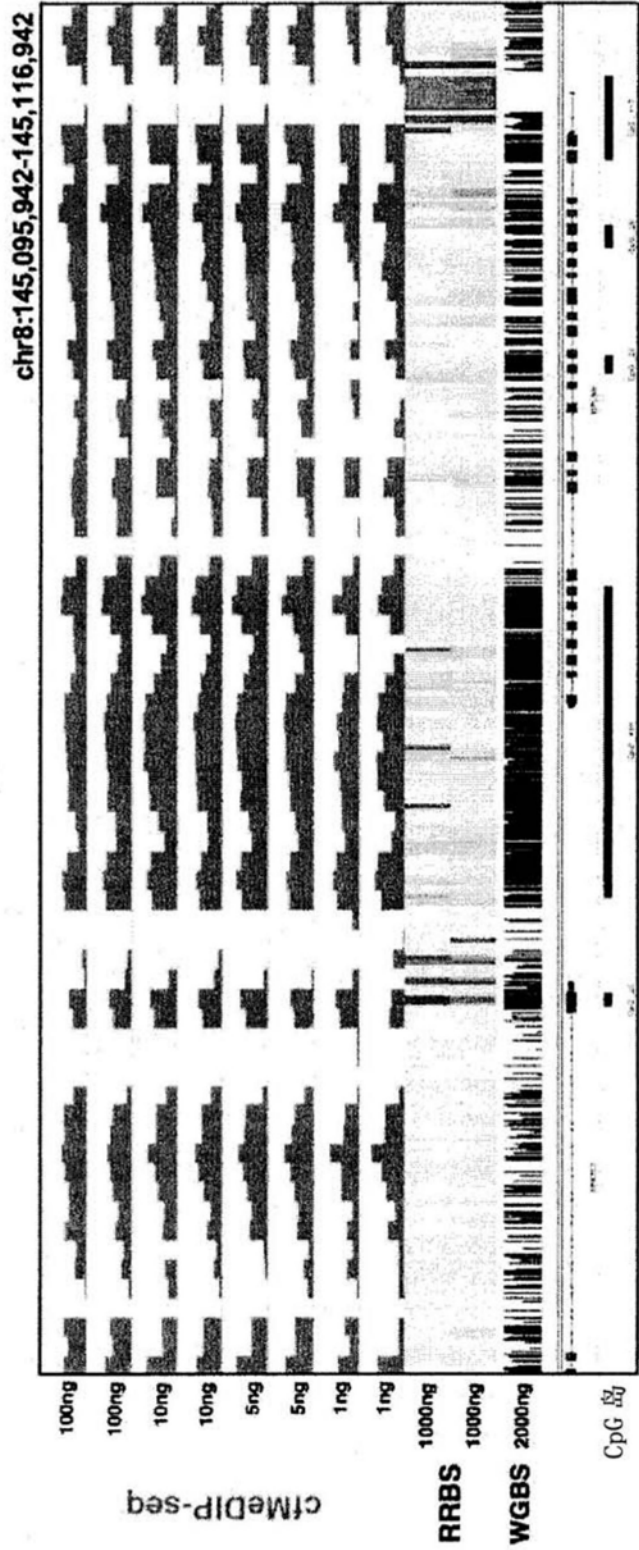


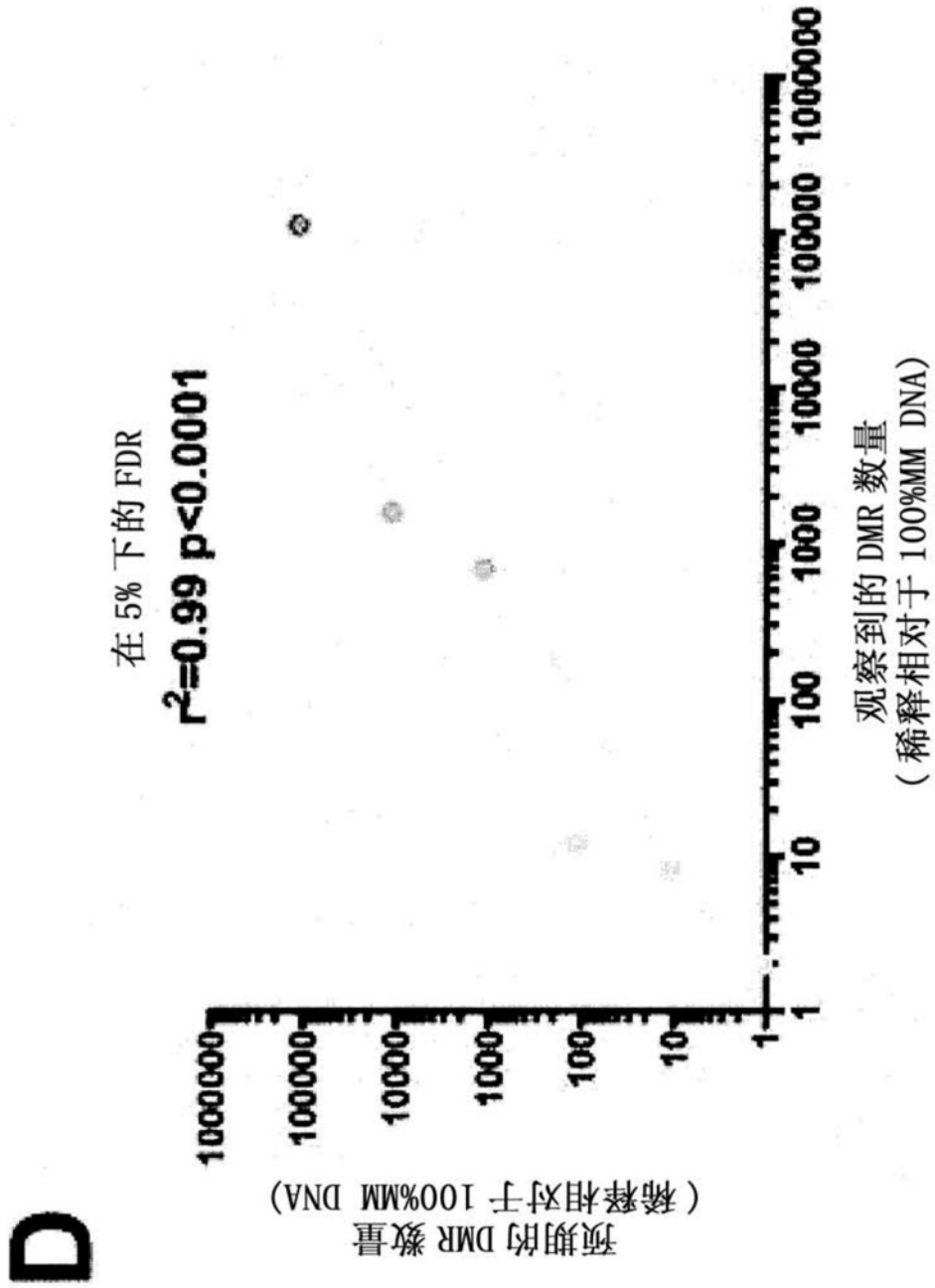
A

B

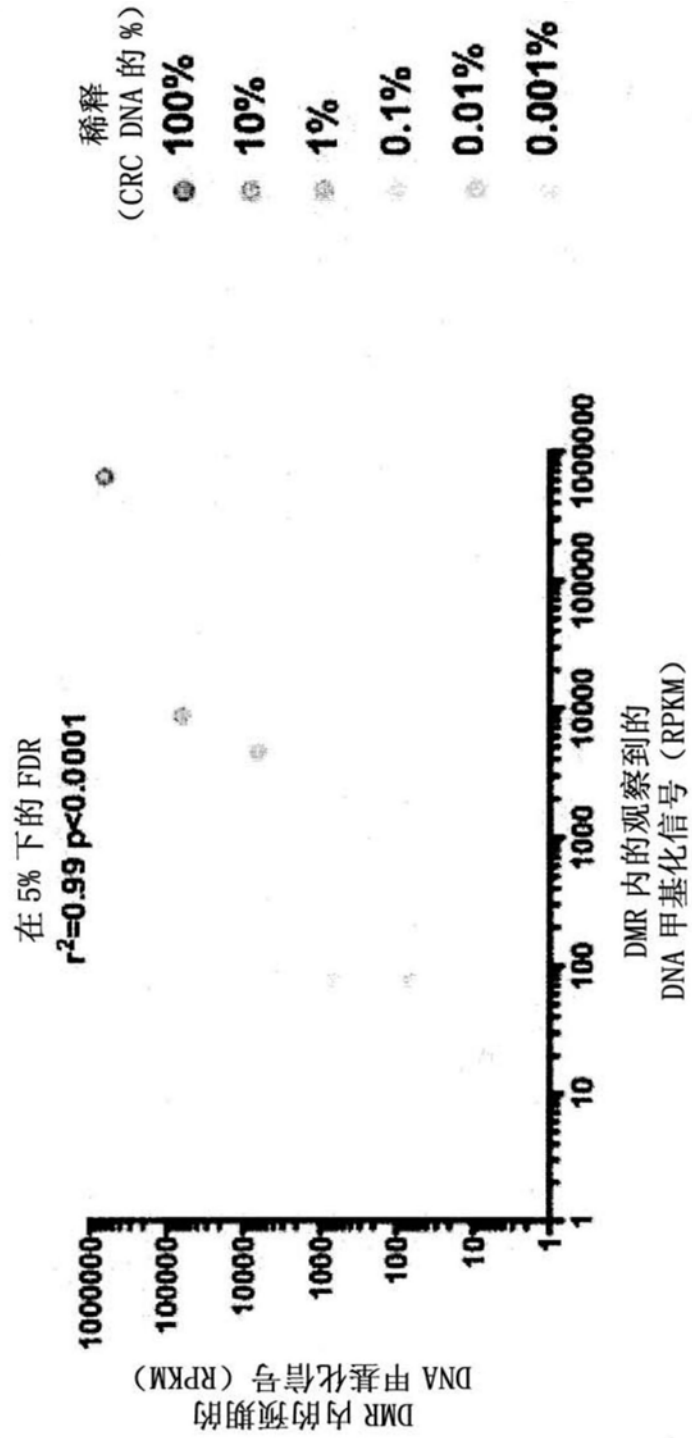
1	0.97	0.91	0.95	0.9	0.93	0.71	0.72	100ng
0.97	1	0.93	0.95	0.92	0.94	0.7	0.73	100ng
0.91	0.93	1	0.96	0.97	0.97	0.81	0.83	10ng
0.95	0.95	0.96	1	0.95	0.97	0.79	0.8	10ng
0.9	0.92	0.97	0.95	1	0.97	0.84	0.86	5ng
0.93	0.94	0.97	0.97	0.97	1	0.82	0.84	5ng
0.71	0.7	0.81	0.79	0.84	0.82	1	0.94	1ng
0.72	0.73	0.83	0.8	0.86	0.84	0.94	1	1ng
							1	100ng (输入)
100ng	100ng	10ng	10ng	5ng	5ng	1ng	1ng	100ng (输入)

C

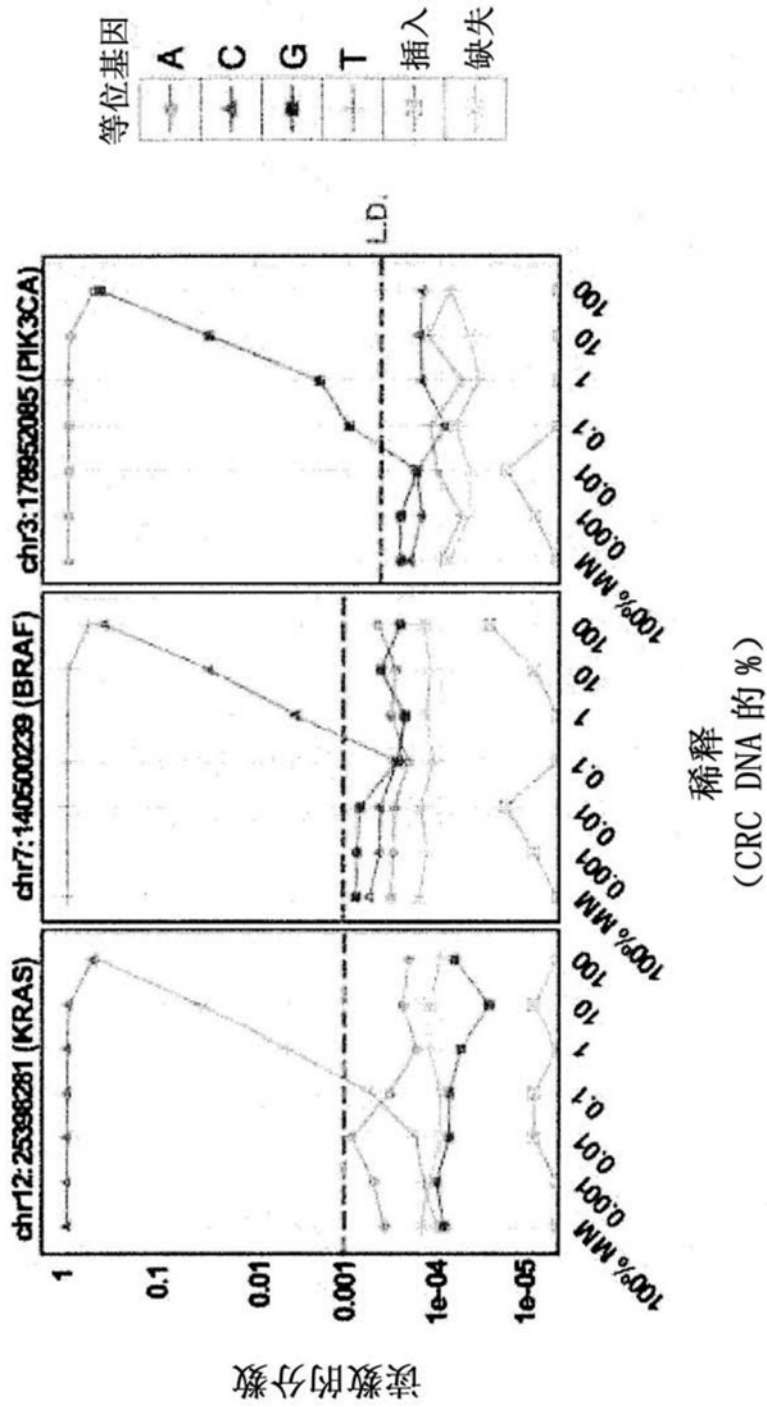




E



F



G

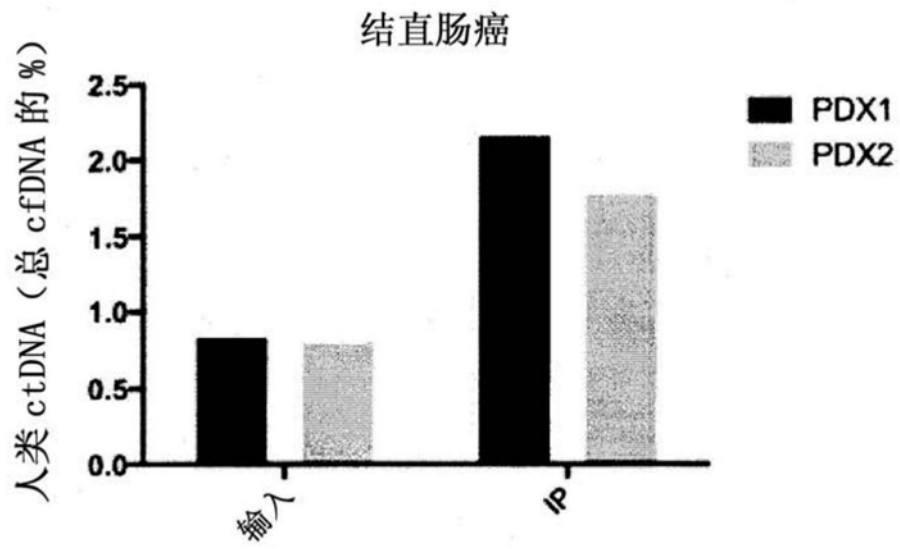


图1

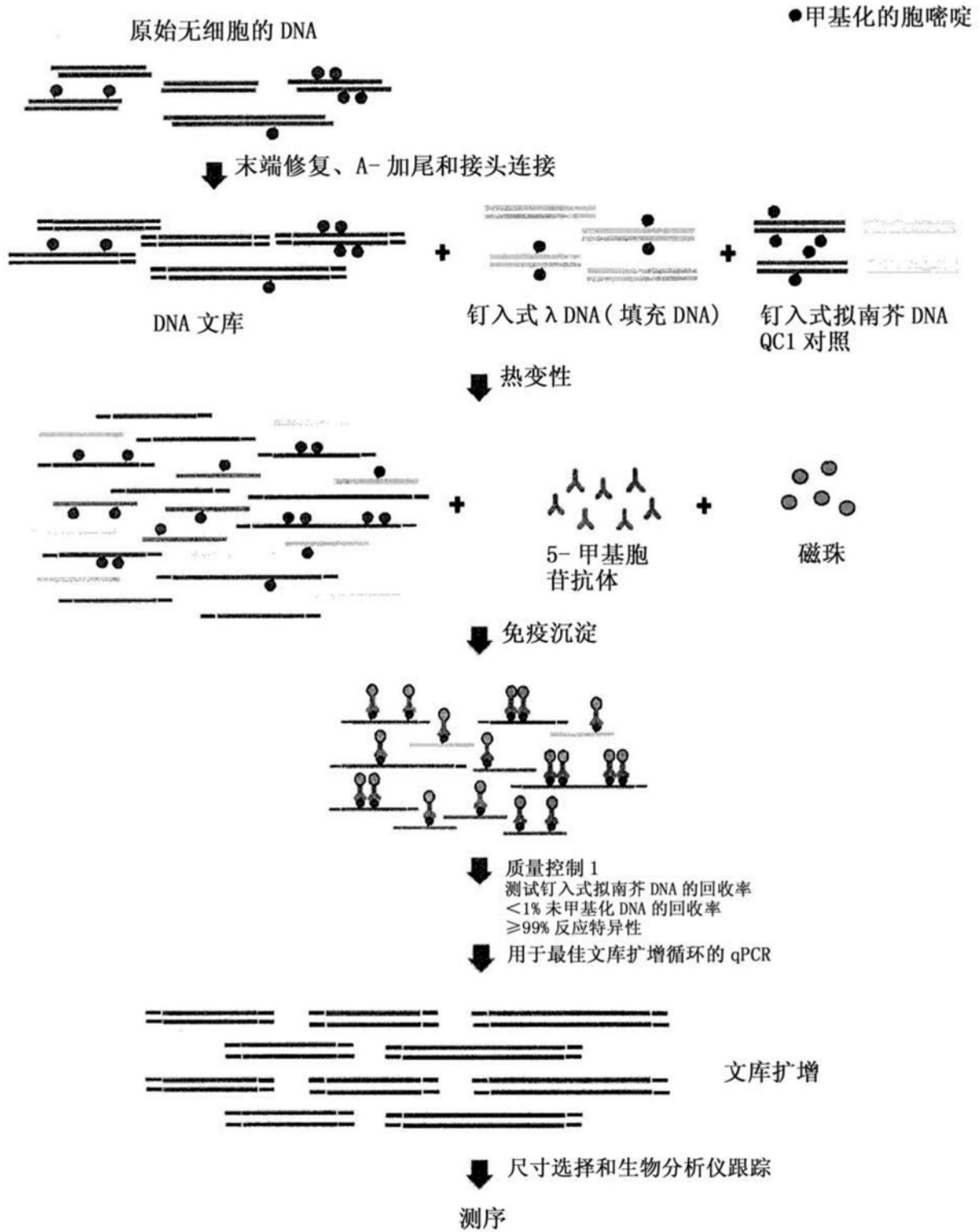
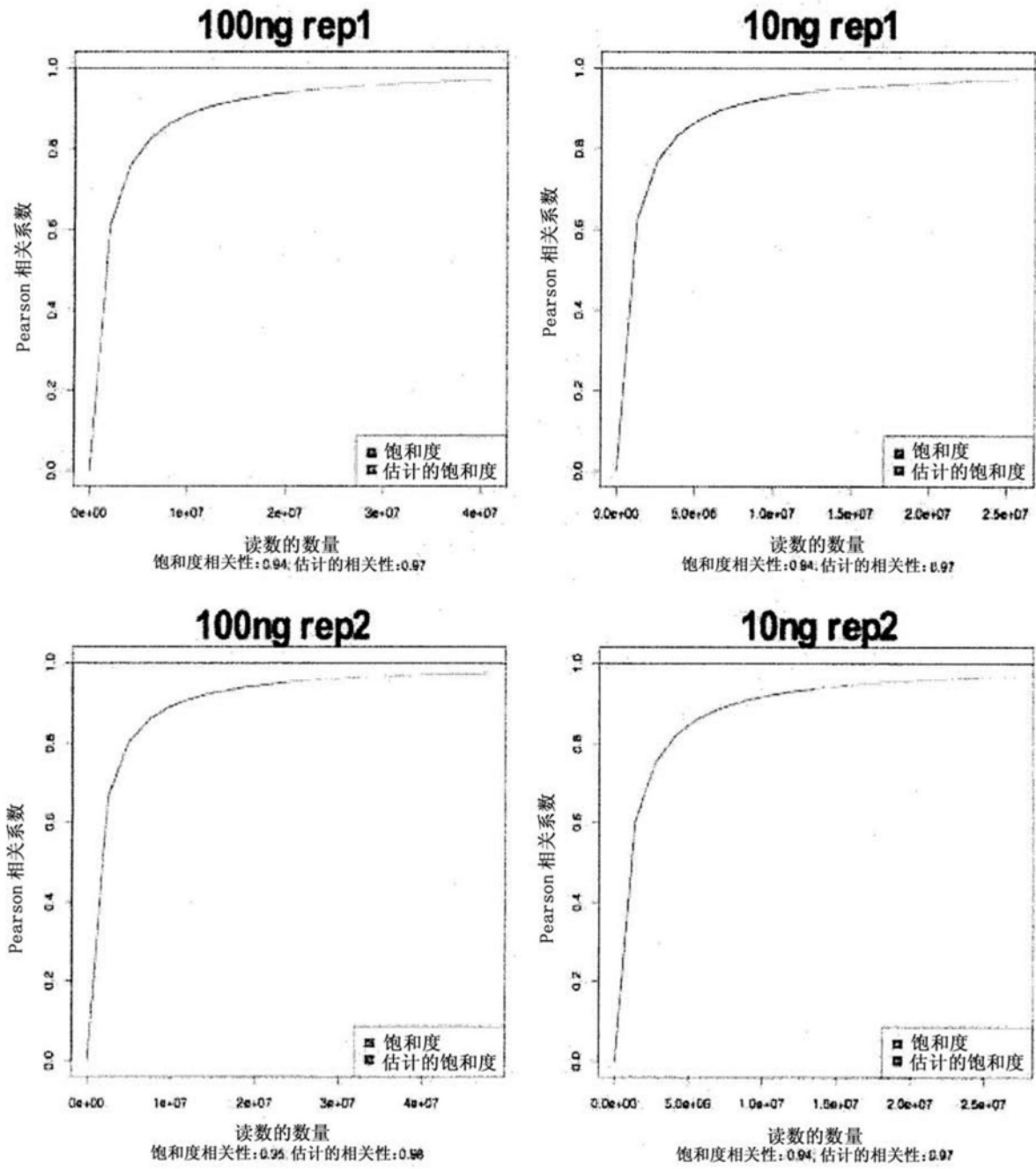
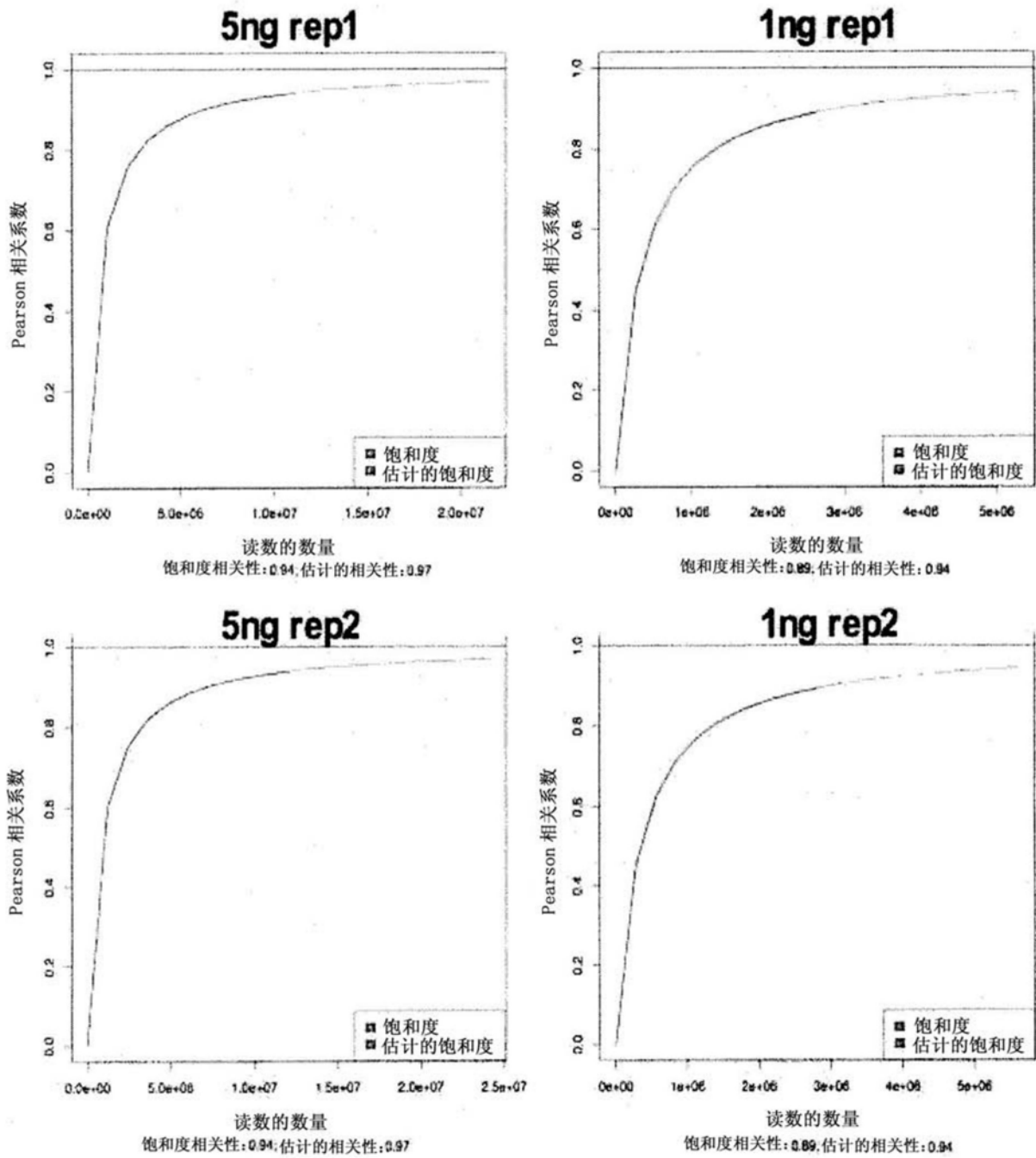


图2

A 饱和度分析

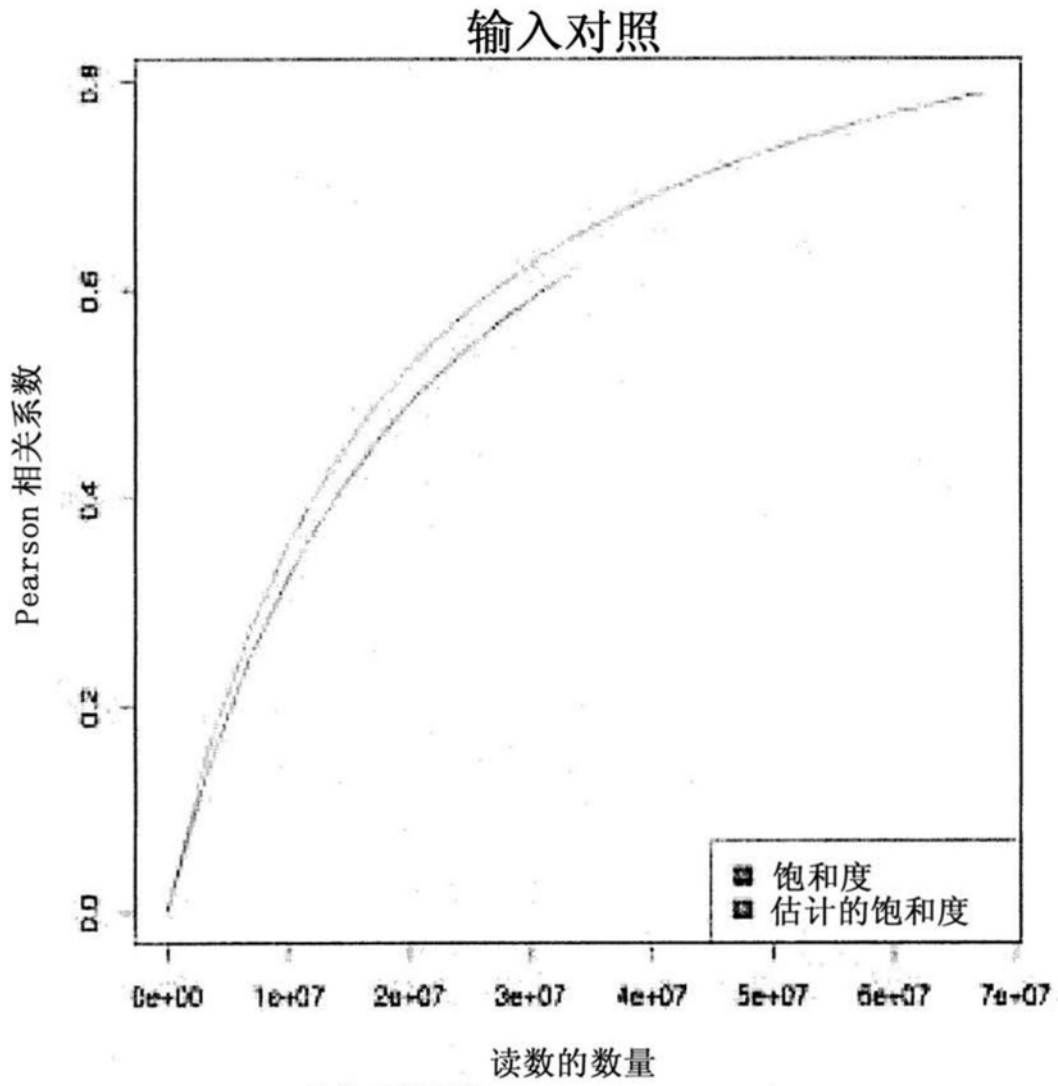


A (续) 饱和度分析



A

(续) 饱和度分析



饱和度相关性:0.62, 估计的相关性:0.79

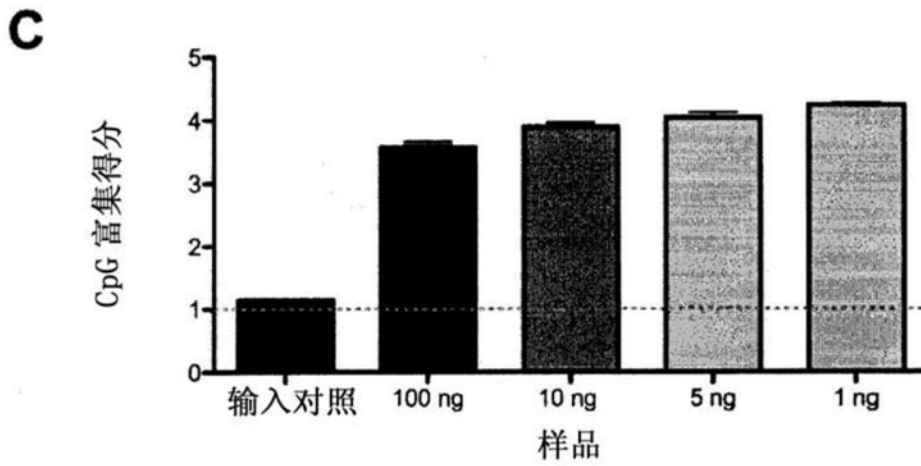
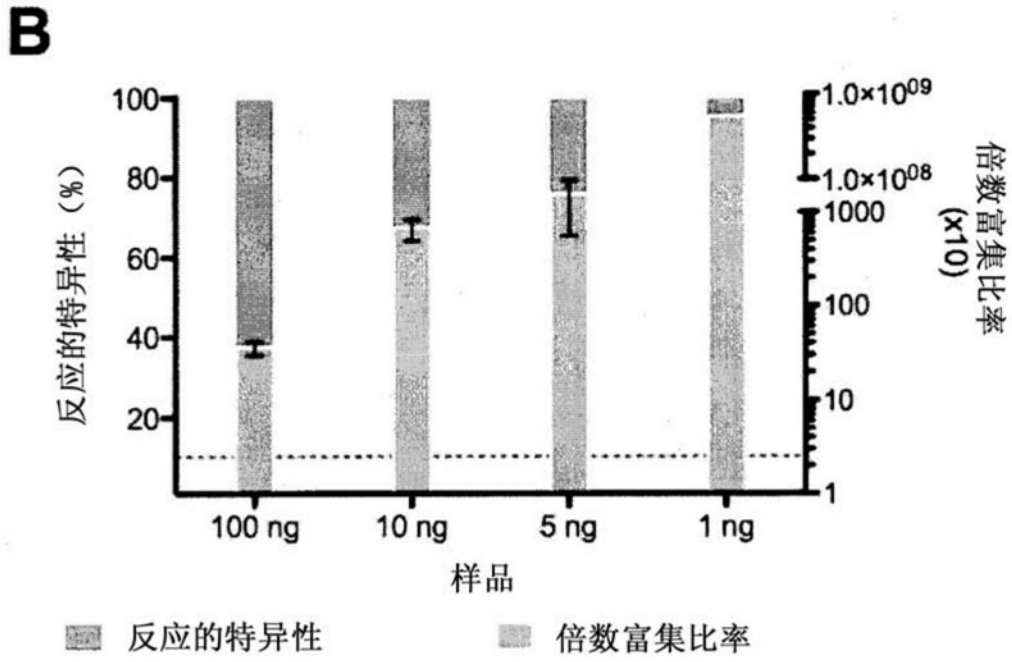
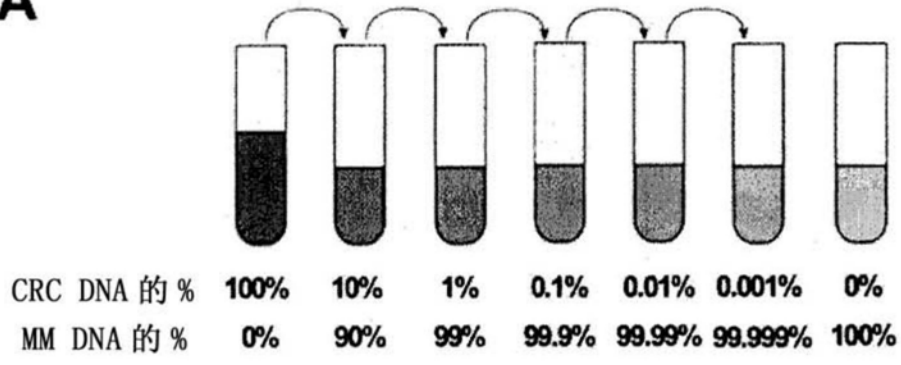
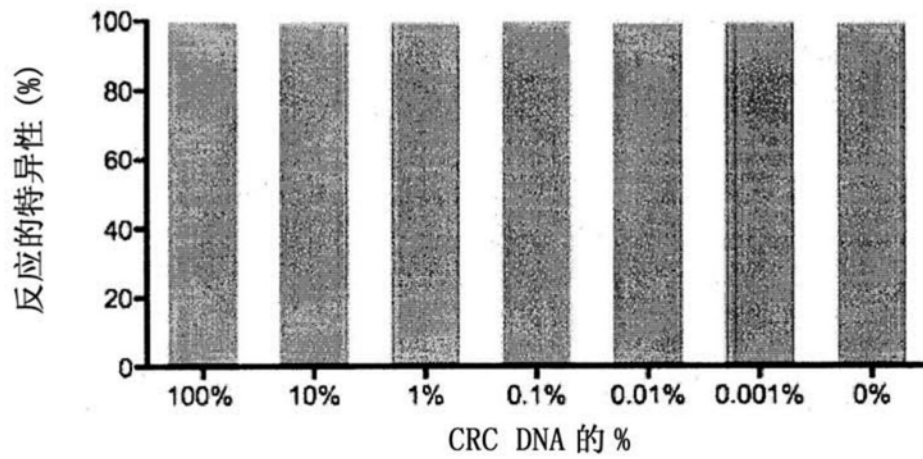


图3

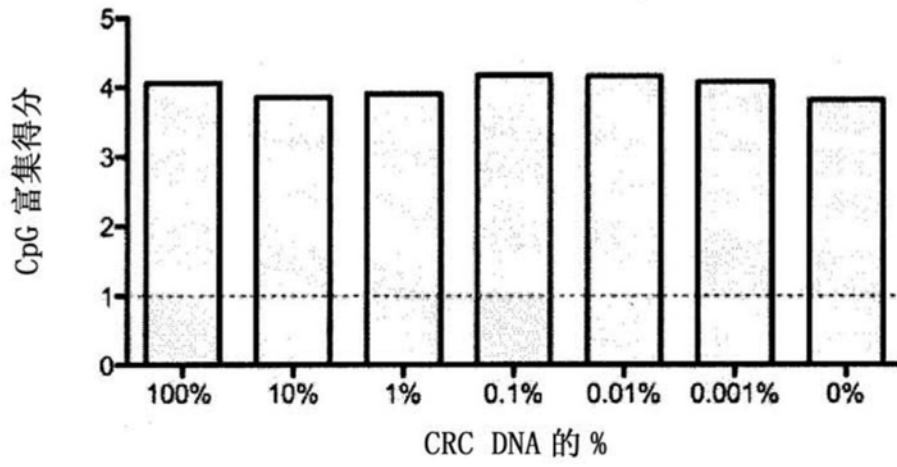
A



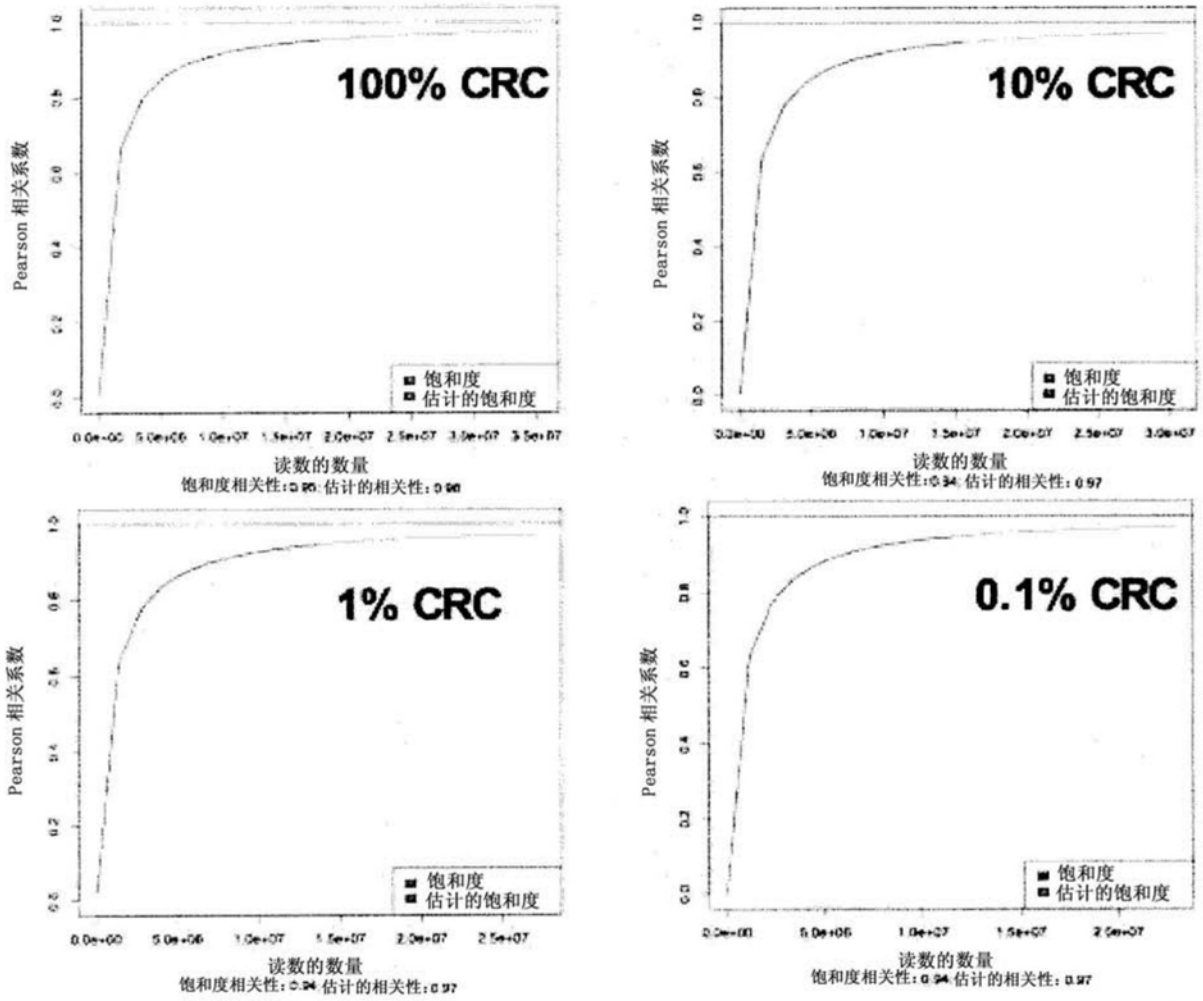
B



C



D



D (续)

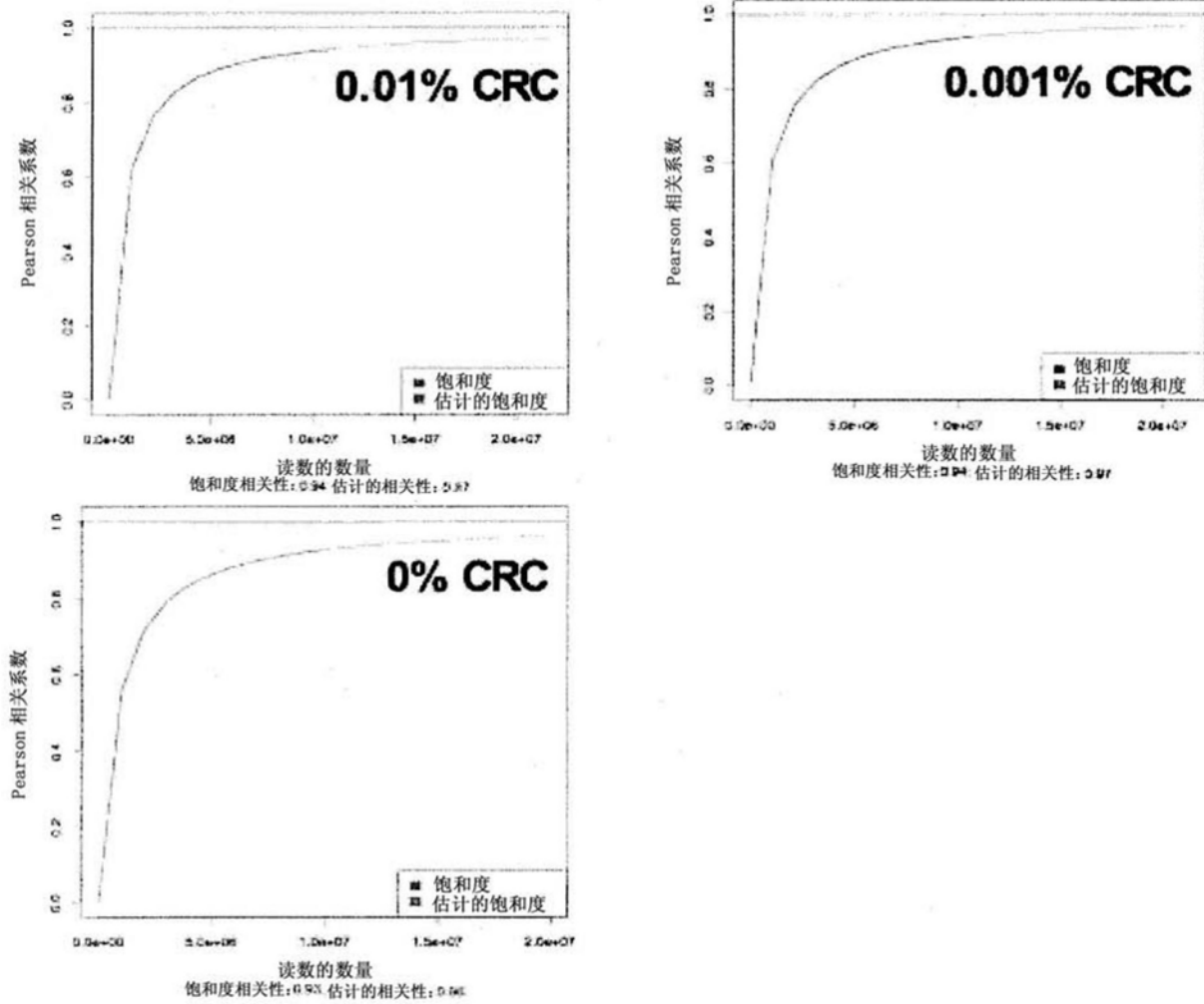
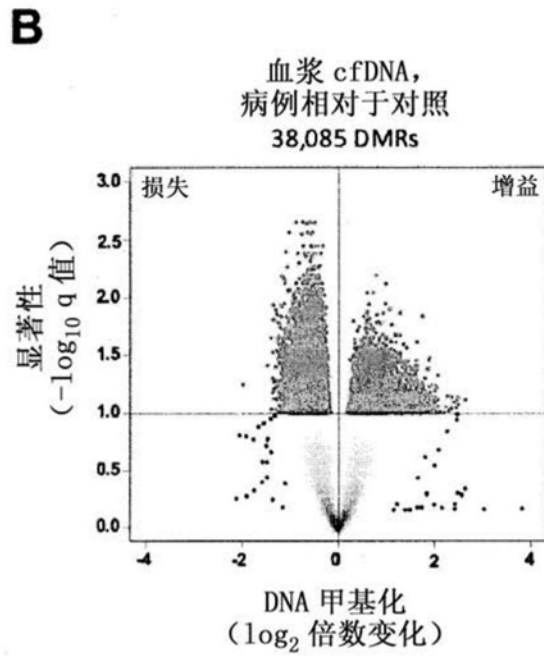
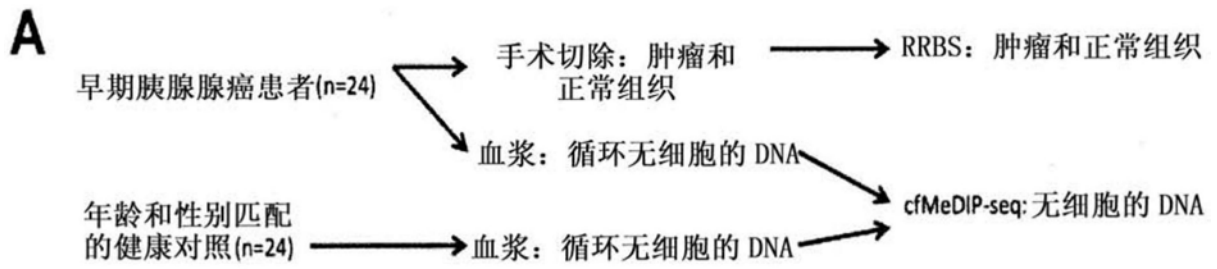
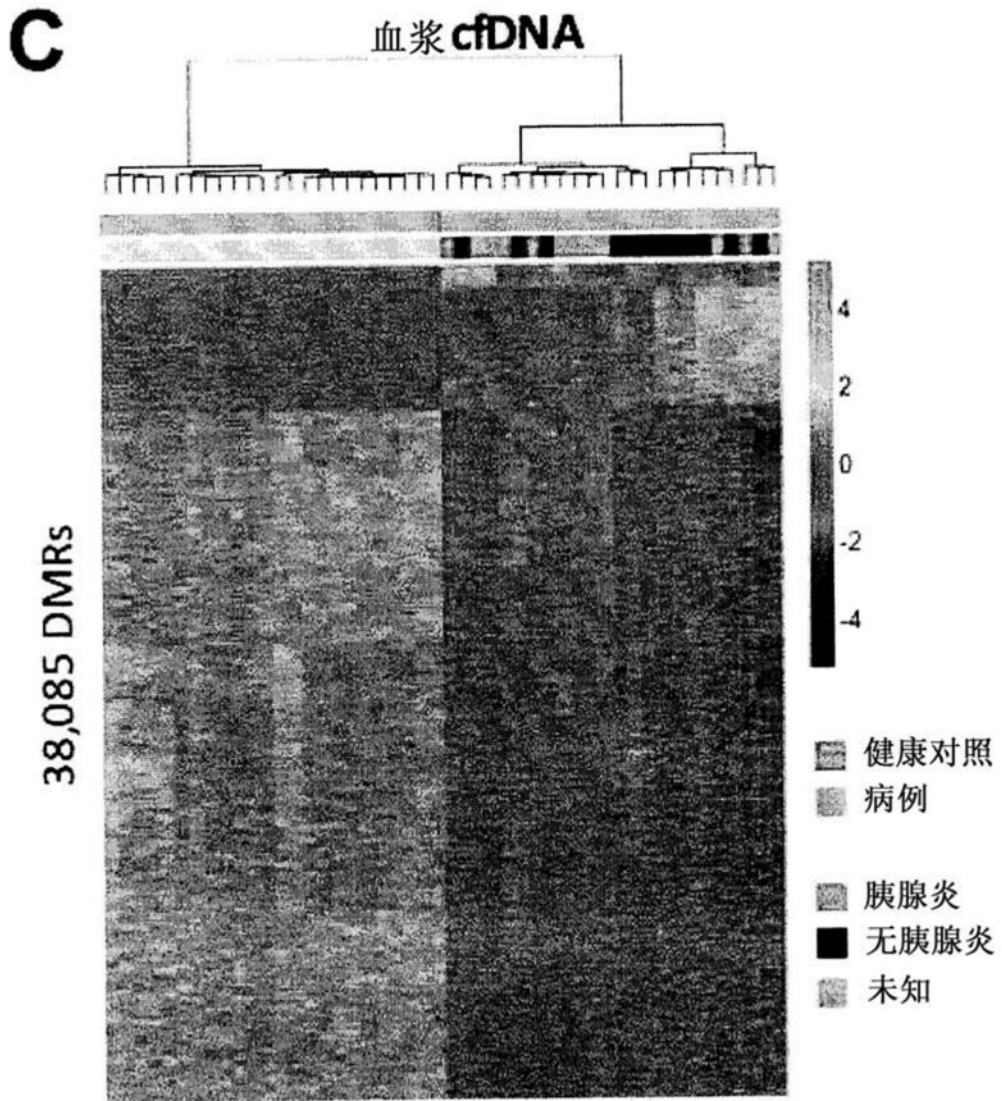


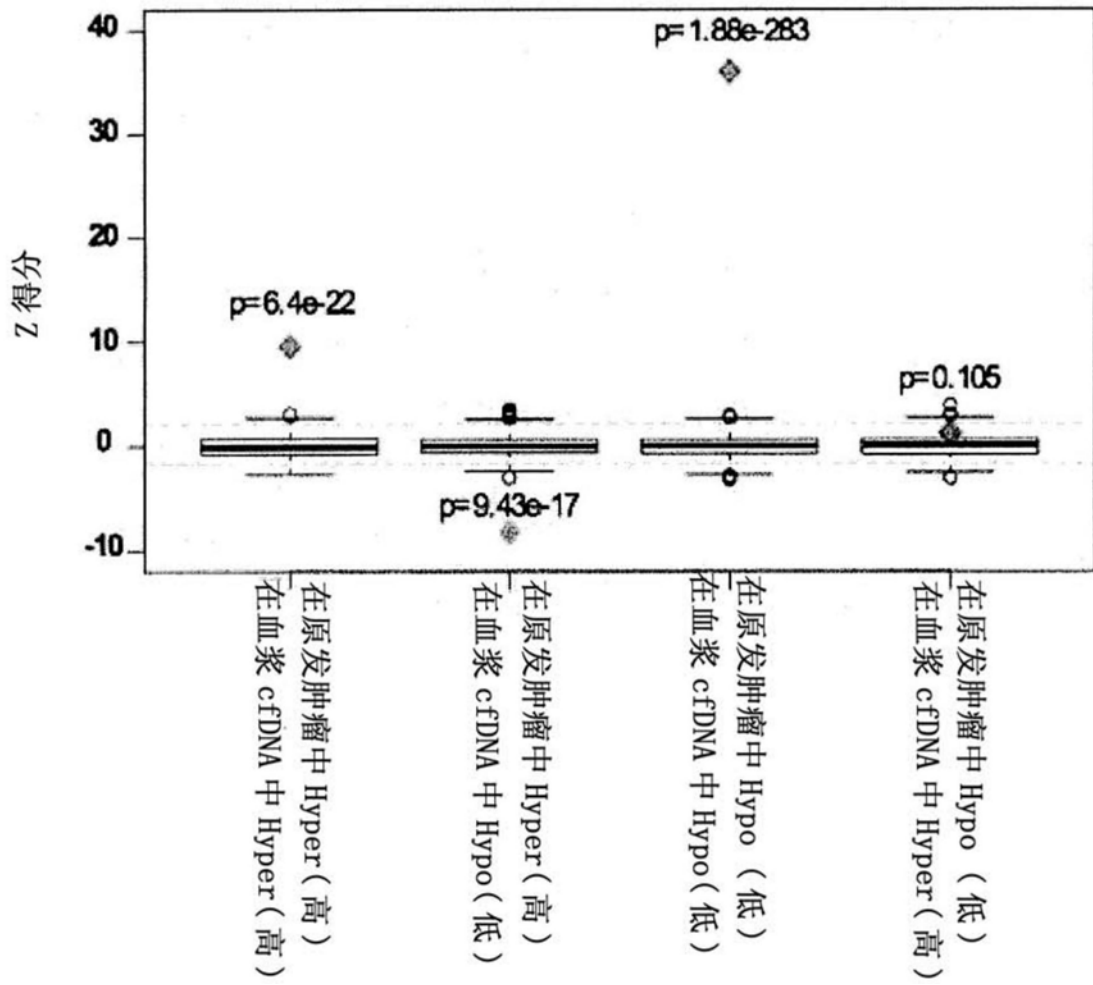
图4





D

重叠血浆 cfDNA (病例相对于对照)
和肿瘤 DNA (原发性肿瘤相对于正常组织)



E

重叠血浆 cfDNA (病例相对于对照)
和肿瘤 DNA (原发肿瘤相对于正常 PBMC)

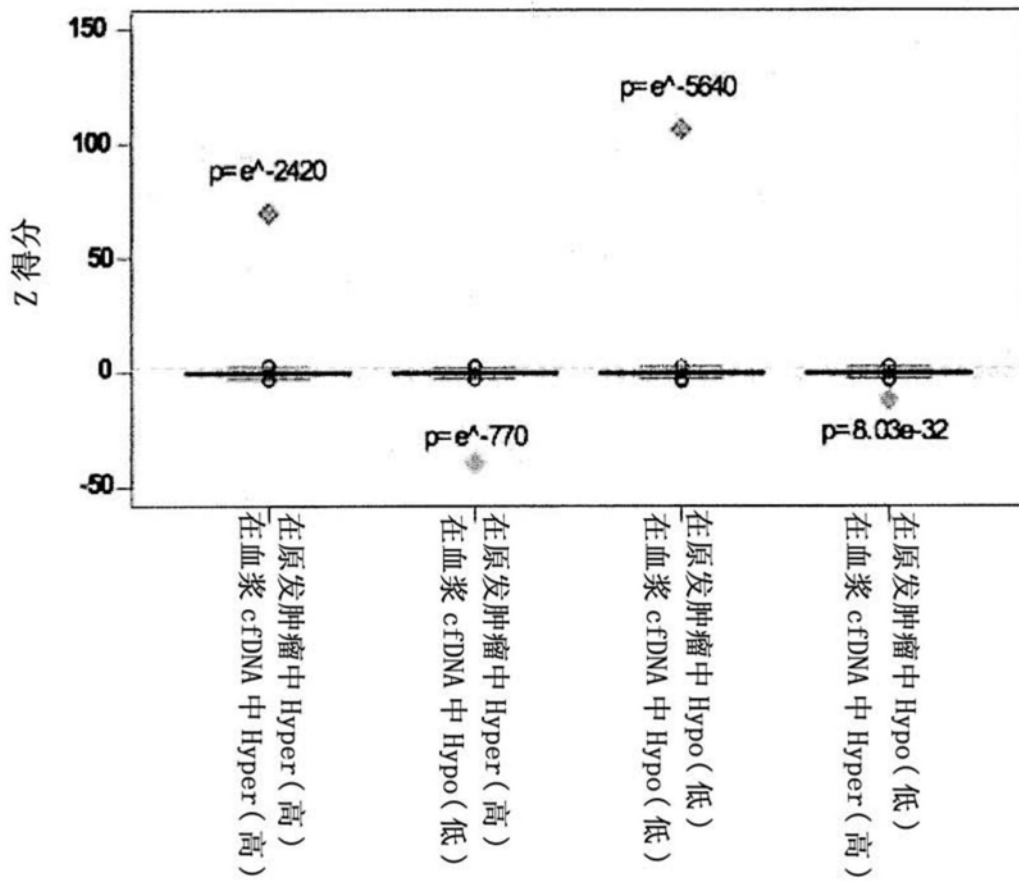
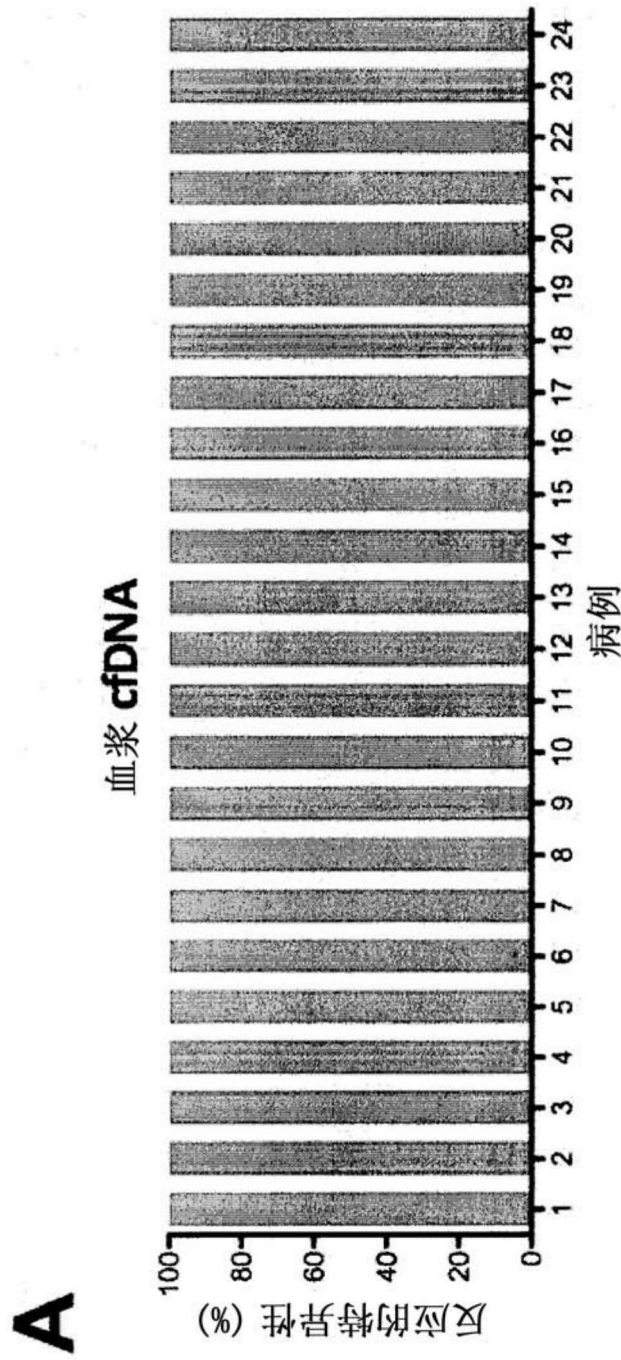
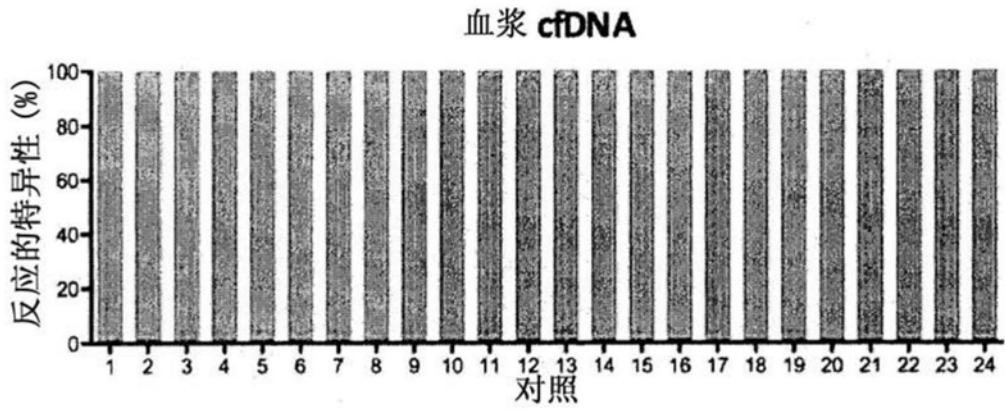


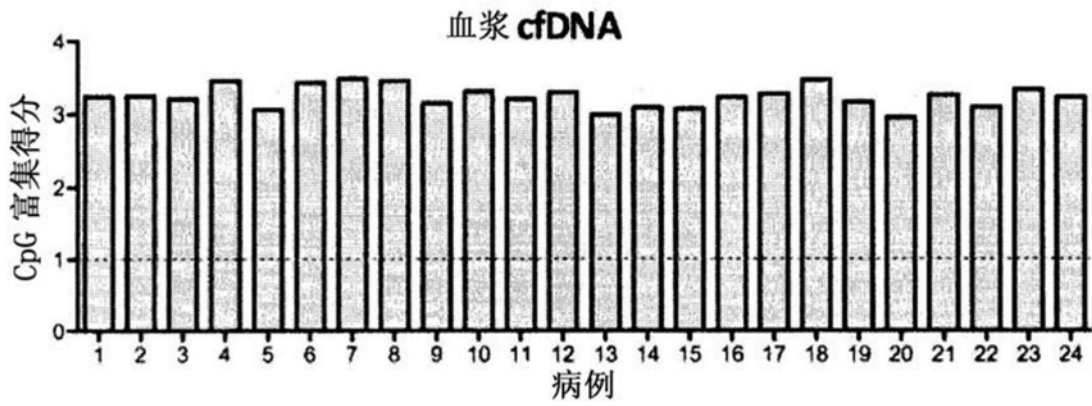
图5



B



C



D

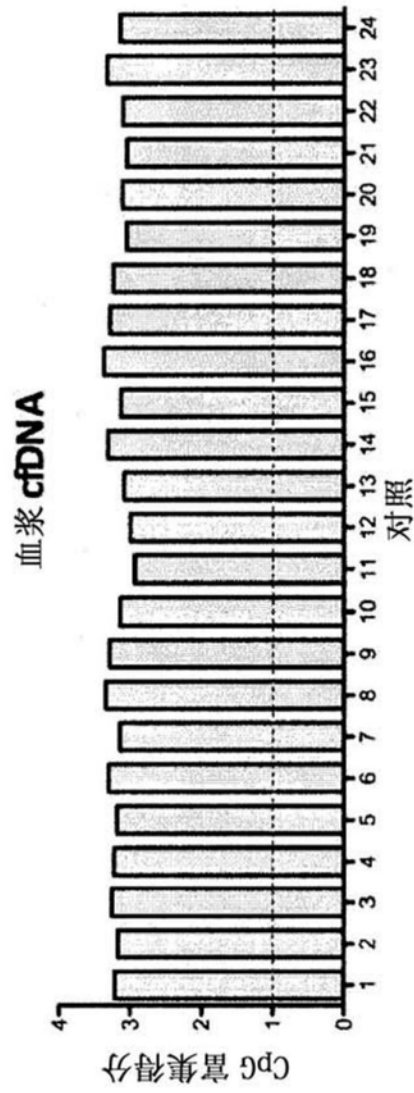
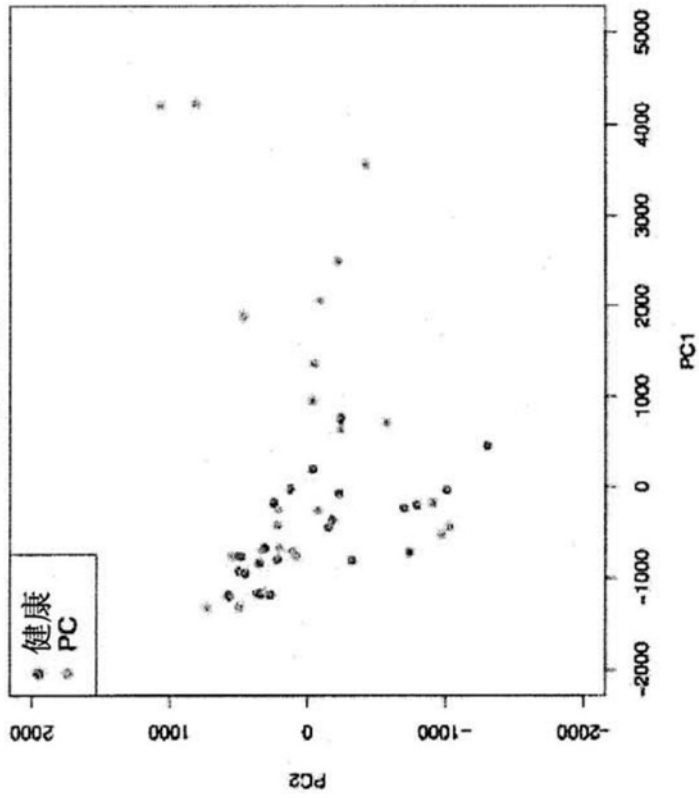


图6

A

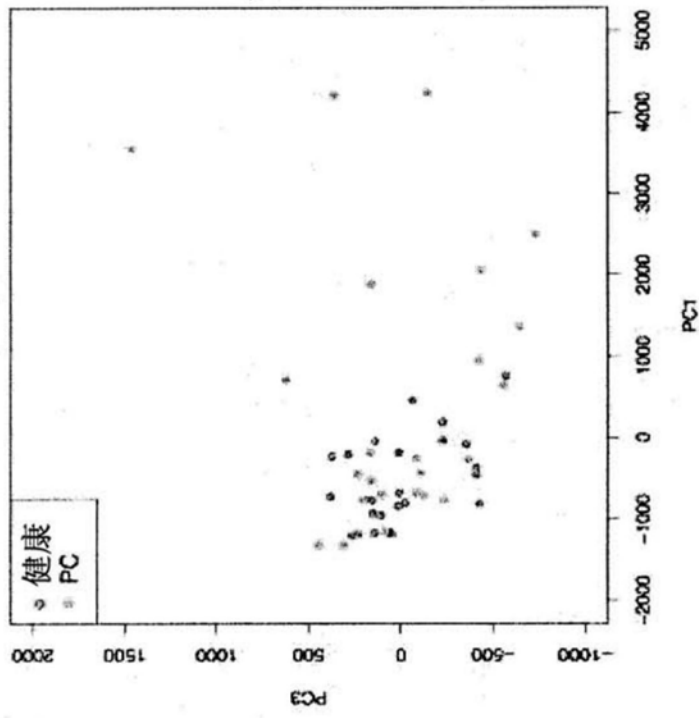
血浆 cfDNA 甲基化

100 万个窗口



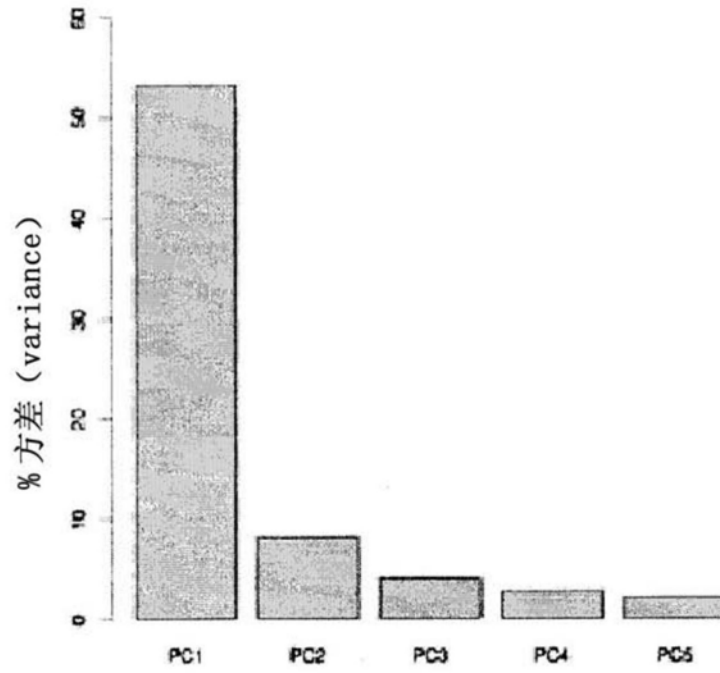
血浆 cfDNA 甲基化

100 万个窗口



B

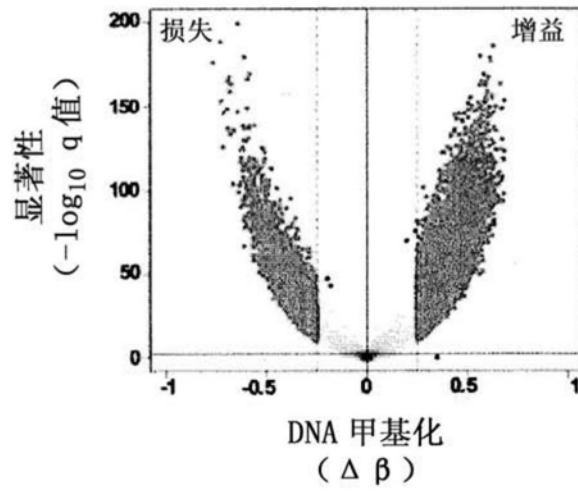
主要分量的方差扩散 (variance spread) %



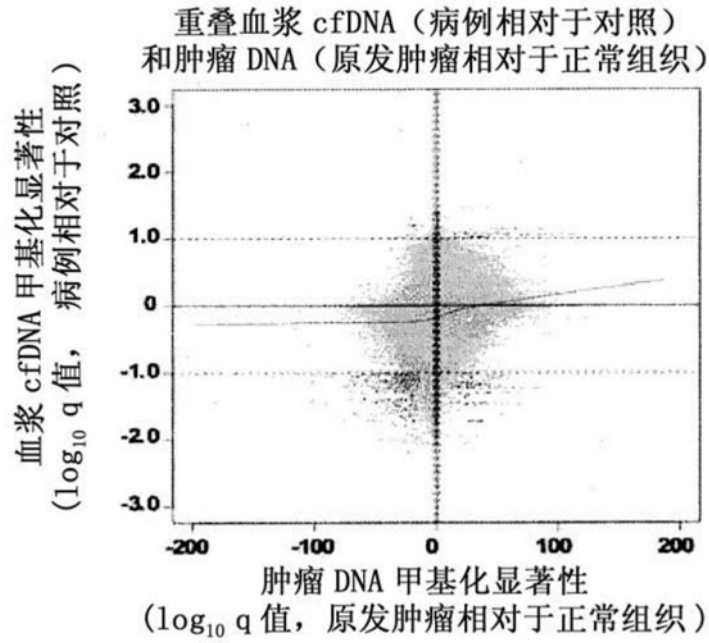
C

原发肿瘤相对于正常组织

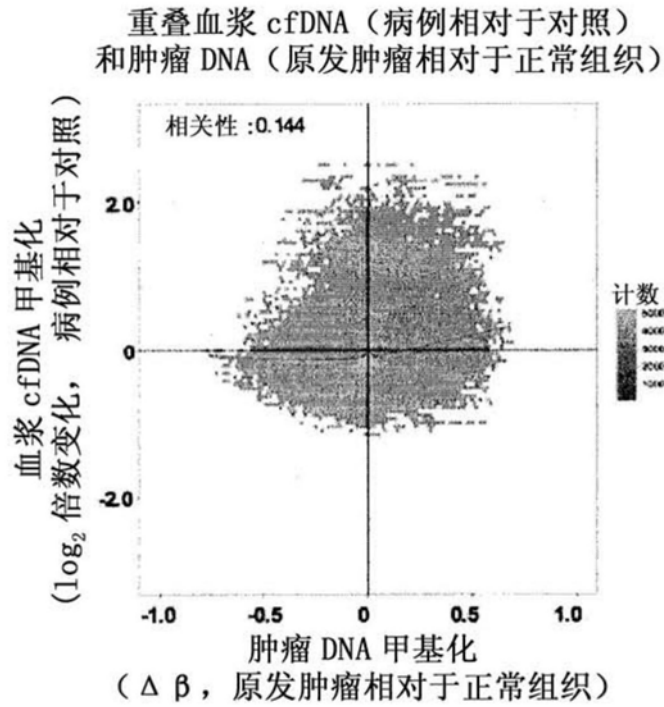
45,173 DMCs



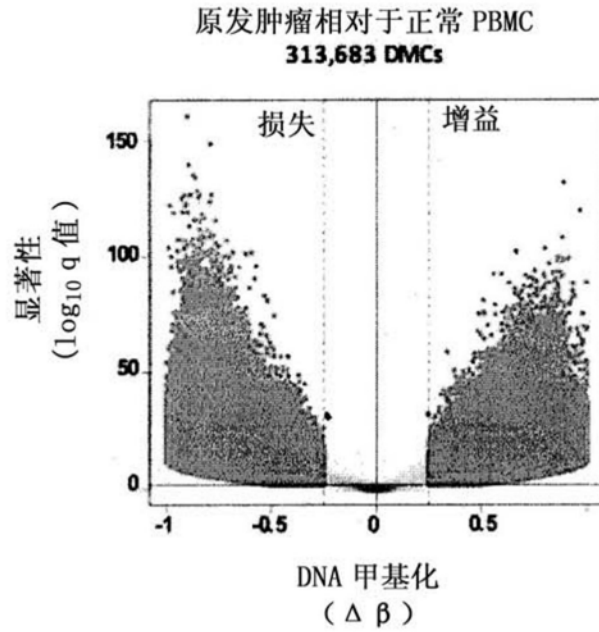
D



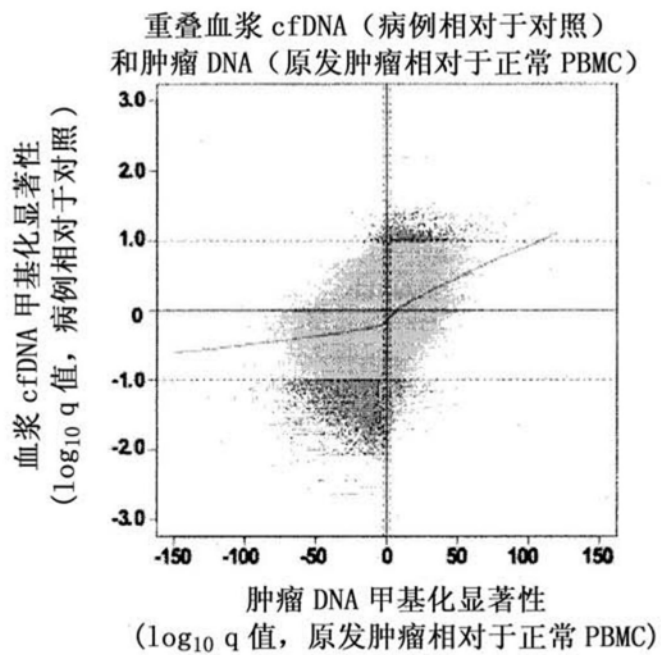
E



F



G



H

重叠血浆 cfDNA (病例相对于对照)
和肿瘤 DNA (原发性肿瘤相对于正常 PBMC)

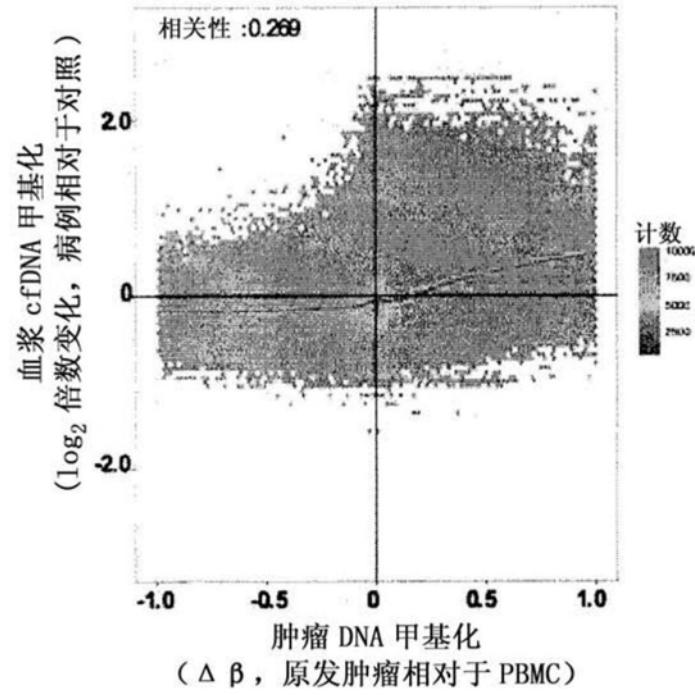
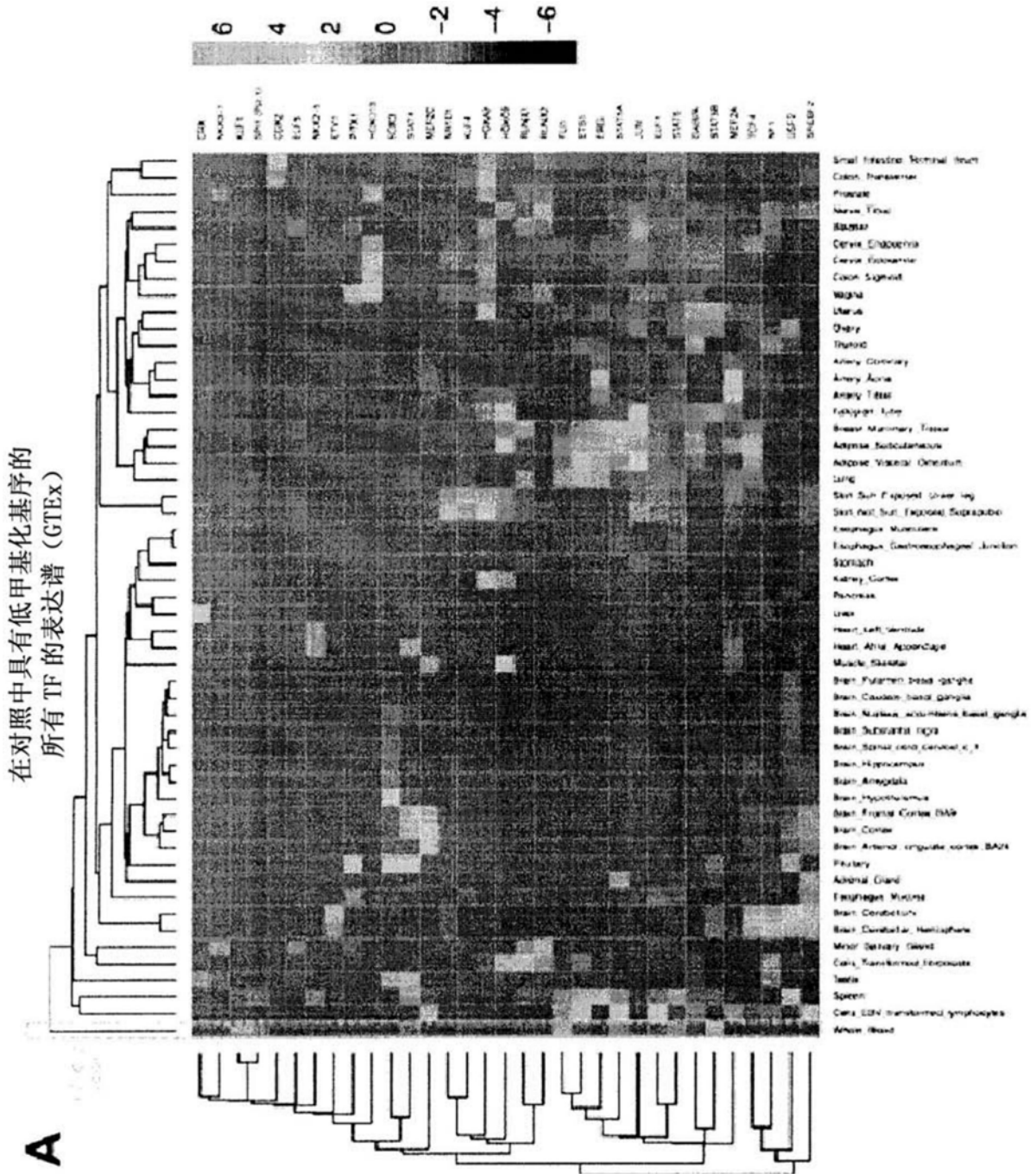
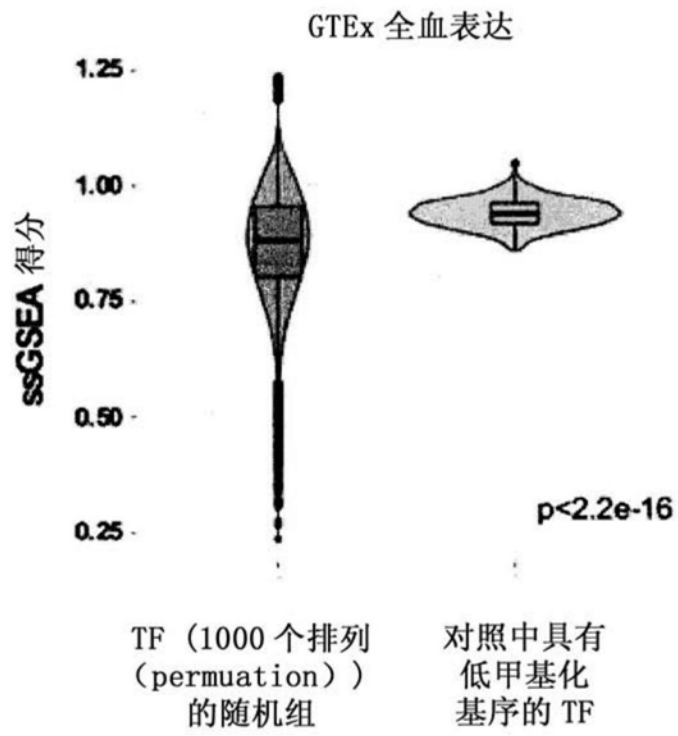


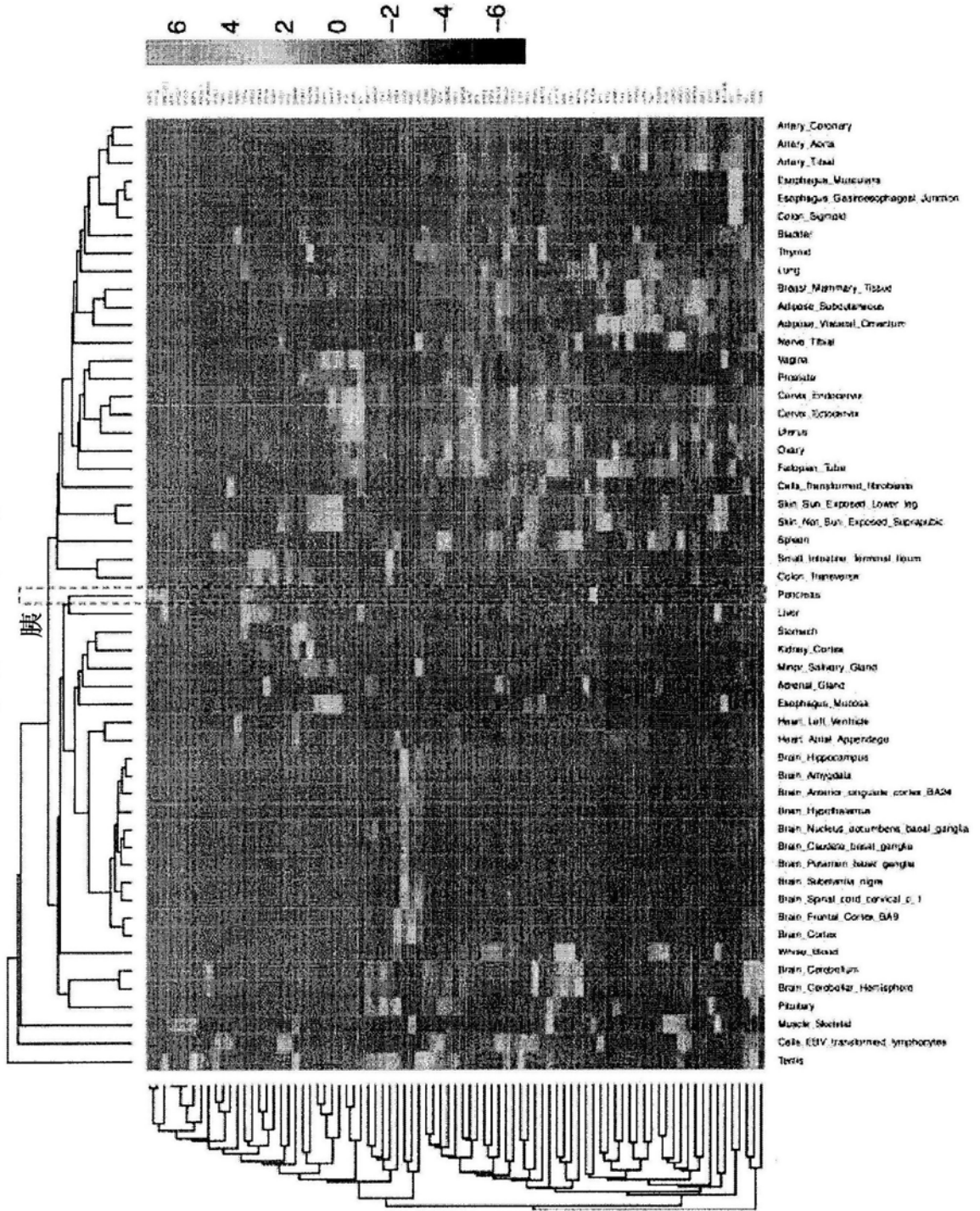
图7



B



在病例中具有低甲基化基序
的所有 TF 的表达谱 (GTEx)



C

图8

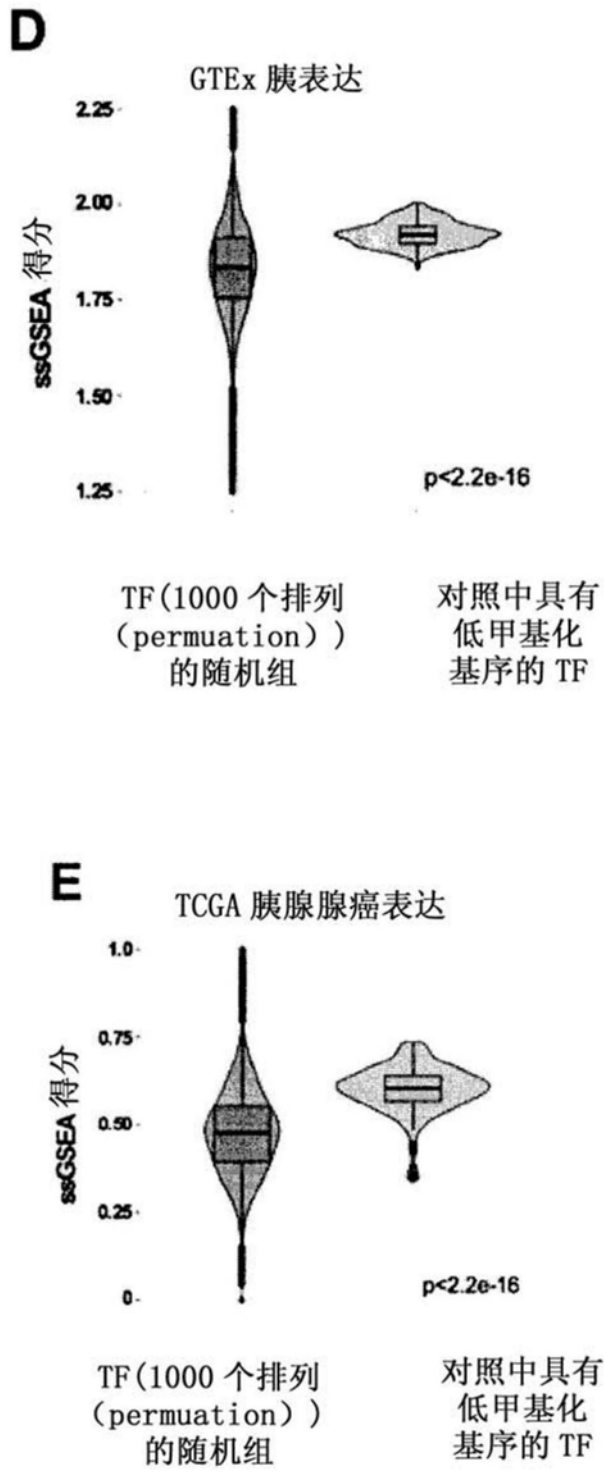


图8

未甲基化的钉入式 DNA 回收率

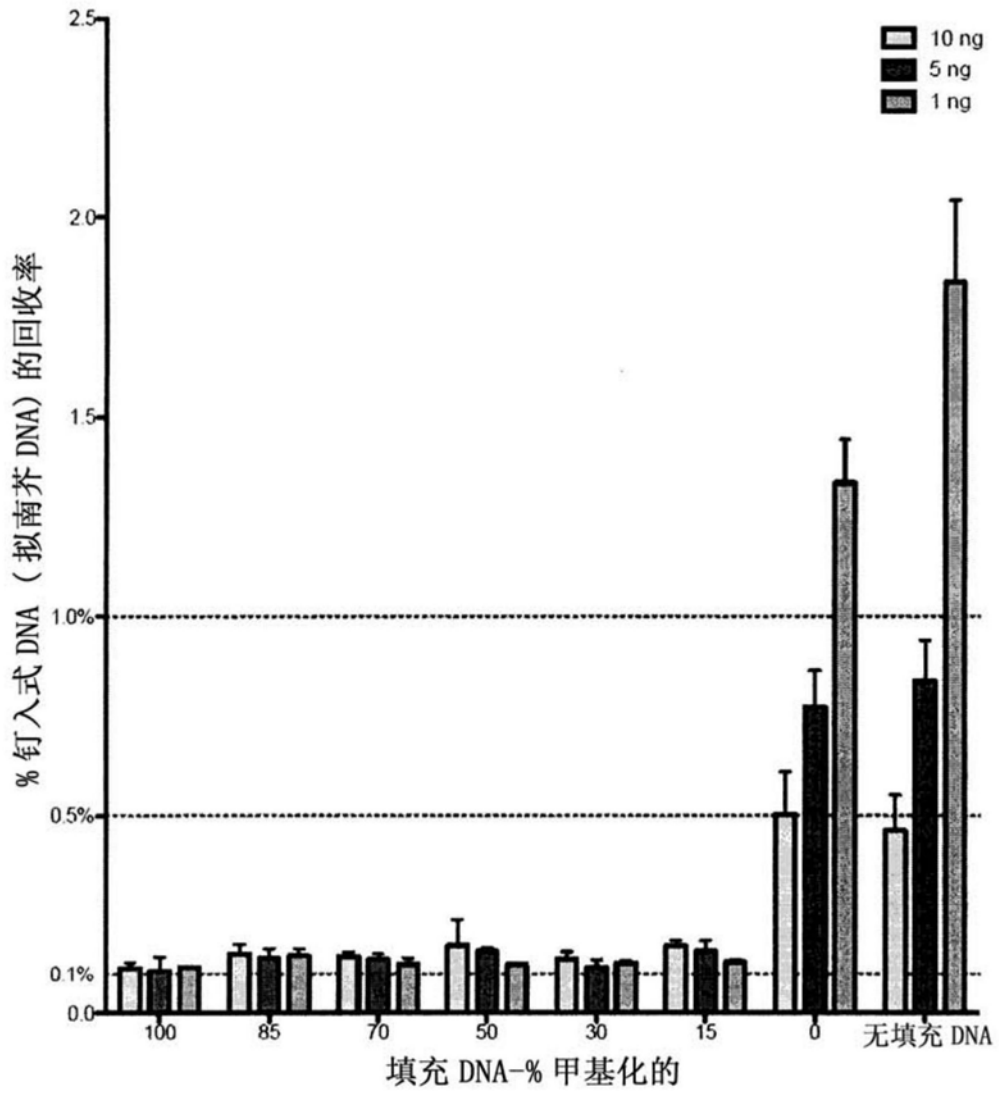


图9

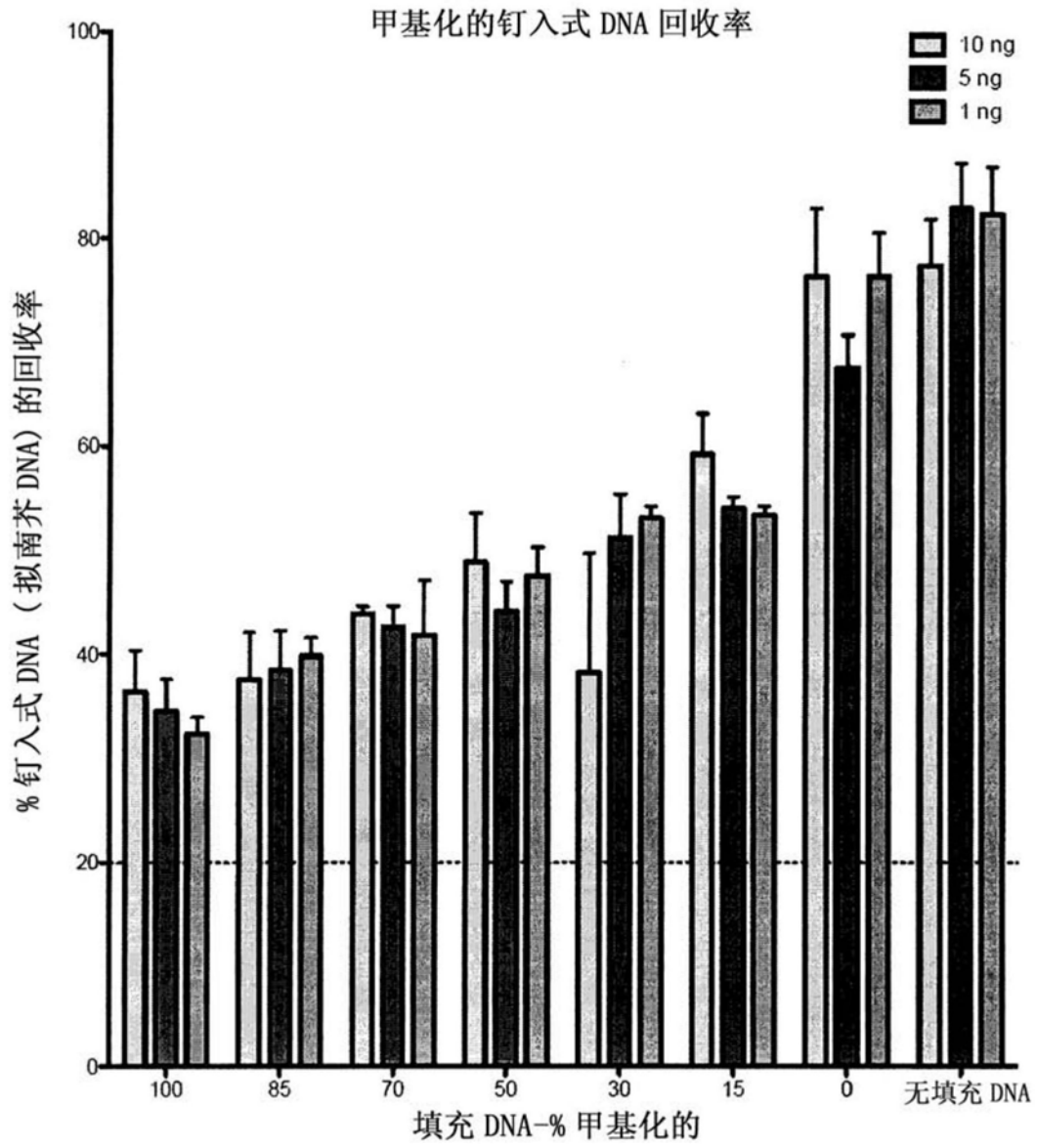


图10