(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117402950 A
(43)申请公布日 2024.01.16

(21)申请号 202311109971.5

(22)申请日 2015.07.27

(30)优先权数据
62/029,178 2014.07.25 US
62/087,619 2014.12.04 US

(62)分案原申请数据
201580052170.7 2015.07.27

(71)申请人 华盛顿大学
地址 美国华盛顿州

(72)发明人 杰·什杜尔 马修·斯奈德 马丁·基尔舍

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
专利代理师 邬玥 方挺

(51)Int.Cl.
C12Q 1/6869(2018.01)
C12Q 1/6881(2018.01)
C12Q 1/6883(2018.01)
C12Q 1/6886(2018.01)
G16B 20/00(2019.01)
G16B 20/10(2019.01)
G16B 20/30(2019.01)
G16B 30/00(2019.01)
G16B 40/00(2019.01)
G16B 40/10(2019.01)
G16B 45/00(2019.01)
G16H 50/20(2018.01)

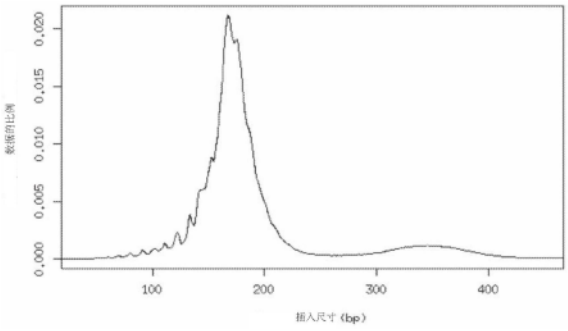
权利要求书2页 说明书48页 附图59页

(54)发明名称

确定导致无细胞DNA的产生的组织和/或细胞类型的方法以及使用其鉴定疾病或紊乱的方法

(57)摘要

本发明提供了确定对象的生物样品中对无细胞DNA(“cfDNA”)有贡献的一种或多种组织和/或细胞类型的方法。在一些实施方式中,本发明提供了根据来自对象的生物样品中一种或多种确定的对cfDNA有贡献的组织和/或细胞类型来鉴定该对象中疾病或紊乱的方法。



CN 117402950 A

1.一种确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型,所述方法包括:

从来自所述对象的生物样品中分离cfDNA,分离的cfDNA包含复数个cfDNA片段;

确定与所述复数个cfDNA片段的至少一部分相关的序列;

根据所述cfDNA片段序列确定所述复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;以及

根据所述至少一些cfDNA片段末端的基因组位置确定至少一些导致所述cfDNA片段的产生的组织和/或细胞类型。

2.一种鉴定对象中疾病或紊乱的方法,所述方法包括:

从来自所述对象的生物样品中分离无细胞DNA(cfDNA),分离的cfDNA包含复数个cfDNA片段;

确定与所述复数个cfDNA片段的至少一部分相关的序列;

根据所述cfDNA片段序列确定所述复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内基因组位置;

根据所述至少一些cfDNA片段末端的基因组位置确定至少一些导致所述cfDNA的产生的组织和/或细胞类型;以及

根据确定的导致所述cfDNA的产生的组织和/或细胞类型来鉴定所述疾病或紊乱。

3.一种确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,包括:

(i)通过以下步骤生成核小体图谱:从所述对象获得生物样品,从所述生物样品分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);

(ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象获得生物样品,从所述生物样品分离cfDNA,通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);以及

(iii)通过比较来源于cfDNA的所述核小体图谱与所述核小体图谱参考集合来确定导致所述cfDNA的产生的组织和/或细胞类型;

其中(a)、(b)和(c)是:

(a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;

(b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和

(c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

4.一种确定对象中导致无细胞DNA的产生的组织和/或细胞类型的方法,包括:

(i)通过以下步骤生成核小体图谱:从所述对象获得生物样品,从所述生物样品分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);

(ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象获得生物样品,从所述生物样品分离cfDNA,通过DNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c),所述DNA来源于使用微球菌核酸酶(MNase)消化染色质、DNA酶处理或ATAC-Seq;以及

(iii)通过比较来源于cfDNA的所述核小体图谱与所述核小体图谱参考集合来确定导致所述cfDNA的产生的组织和/或细胞类型;

其中(a)、(b)和(c)是:

(a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;

(b)人基因组碱基对中任意一对呈现为一对测序的片段终端的可能性分布;和

(c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

5.一种诊断对象中临床病症的方法,包括:

(i)通过以下步骤生成核小体图谱:从所述对象获得生物样品,从所述生物样品分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);

(ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对象获得生物样品,从所述生物样品分离cfDNA,通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);以及

(iii)通过比较来源于cfDNA的所述核小体图谱与所述核小体图谱参考集合来确定所述临床病症

其中(a)、(b)和(c)是:

(a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;

(b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和

(c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

6.一种诊断对象中临床病症的方法,包括:

(i)通过以下步骤生成核小体图谱:从所述对象获得生物样品,从所述生物样品分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c);

(ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象获得生物样品,从所述生物样品分离cfDNA,通过DNA的大规模平行测序和文库构建测量分布(a)、(b)和/或(c),所述DNA来源于使用微球菌核酸酶(MNase)消化染色质、DNA酶处理或ATAC-Seq;以及

(iii)通过比较来源于cfDNA的所述核小体图谱与所述核小体图谱参考集合来确定所述cfDNA的来源组织组成;

其中(a)、(b)和(c)是:

(a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;

(b)人基因组碱基对中任意一对呈现为一对测序的片段终端的可能性分布;和

(c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

确定导致无细胞DNA的产生的组织和/或细胞类型的方法以及使用其鉴定疾病或紊乱的方法

[0001] 本申请是2015年7月27日提交的申请号为“201580052170.7”,名称为“确定导致无细胞DNA的产生的组织和/或细胞类型的方法以及使用其鉴定疾病或紊乱的方法”的中国国家阶段专利申请的分案申请。

[0002] 优先权声明

[0003] 本申请要求2014年7月25日提交的美国临时申请号62/029,178和2014年12月4日提交的美国临时申请号62/087,619的优先权,其各自的主题通过引用纳入本文,如同其全文列于本文中一样。

[0004] 政府参与声明

[0005] 本发明在国家健康研究所(NIH)授予的批准号1DP1HG007811的政府支持下完成。政府在本发明中拥有某些权利。

技术领域

[0006] 本发明涉及确定一种或多种导致无细胞DNA的产生的组织和/或细胞类型的方法。在一些实施方式中,本发明提供了根据来自对象的生物样品中一种或多种确定的与无细胞DNA相关的组织和/或细胞类型鉴定该对象中疾病或紊乱的方法。

[0007] 背景

[0008] 无细胞DNA(“cfDNA”)存在于循环血浆、尿液和其他人体液中。cfDNA包含双链DNA片段,其相对较短(绝大多数小于200碱基对)且通常以低浓度存在(例如血浆中1-100ng/mL)。在健康个体的循环血浆中,cfDNA被认为主要来源于血细胞(即造血谱系的正常细胞)的凋亡。然而,在特定条件下,其他组织可对体液(例如循环血浆)中cfDNA的组成有显著贡献。

[0009] 虽然已在某些专门领域(例如,生殖医学、癌症诊断和移植医学)中使用cfDNA,但现有的基于cfDNA的测试依赖于两种或更多种细胞群体之间(例如,母系基因组对比胎儿基因组;正常基因组对比癌症基因组;移植受体基因组对比供体基因组等)基因型的区别(例如,一级序列或具体序列的拷贝数代表(copy number representation))。不幸的是,因为任何给定生物样品中发现的绝大部分cfDNA片段来源于有贡献的细胞群体之间序列相同的基因组区域,现有的基于cfDNA的测试极度受限于其应用范围。此外,许多疾病和紊乱伴随导致cfDNA的产生的组织和/或细胞类型的变化,例如来自与该疾病或紊乱相关的组织损伤或炎性过程。现有的基于cfDNA的诊断测试依赖于两个基因组之间一级序列或具体序列的拷贝数表现的区别,其无法检测这类变化。因此,虽然cfDNA提供有力的无活检切片检测方法的潜力巨大,仍然需要可应用于诊断多种疾病或紊乱的基于cfDNA的诊断方法。

发明内容

[0010] 本发明提供了确定对象的生物样品中一种或多种导致无细胞DNA(“cfDNA”)的产生的组织和/或细胞类型的方法。在一些实施方式中,本发明提供了根据来自对象的生物样

品中一种或多种经确定的与无细胞DNA相关的组织和/或细胞类型鉴定该对象中疾病或紊乱的方法。

[0011] 在一些实施方式中,本发明提供了确定对象的生物样品中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,该方法包括从来自该对象的生物样品中分离cfDNA,分离的cfDNA包含复数个cfDNA片段;确定与复数个cfDNA片段的至少一部分相关的序列;根据cfDNA片段序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;以及根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA片段的产生的组织和/或细胞类型。

[0012] 在其他实施方式中,本发明提供了鉴定对象中疾病或紊乱的方法,该方法包括从来自该对象的生物样品中分离无细胞DNA(cfDNA),分离的cfDNA包含复数个cfDNA片段;确定与复数个cfDNA片段的至少一部分相关的序列;根据cfDNA片段序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA的产生的组织和/或细胞类型;以及根据确定的导致cfDNA的产生的组织和/或细胞类型鉴定疾病或紊乱。

[0013] 在其他实施方式中,本发明提供了用于确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,该方法包括:(i)通过从该对象获得生物样品来生成核小体图谱,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);(ii)通过从一个或多个具有已知疾病的对照对象中获得生物样品来生成核小体图谱参考集合(reference set of nucleosome maps),从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);以及(iii)通过比较来源于来自该生物样品的cfDNA的核小体图谱与核小体图谱参考集合来确定来自该生物样品的导致cfDNA的产生的组织和/或细胞类型;其中(a)、(b)和(c)是:(a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;(b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和(c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

[0014] 在其他实施方式中,本发明提供了用于确定对象中导致cfDNA的产生的组织和/或细胞类型的方法,该方法包括:(i)通过从该对象获得生物样品来生成核小体图谱,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);(ii)通过从一个或多个具有已知疾病的对照对象中获得生物样品来生成核小体图谱参考集合,从该生物样品中分离cfDNA,以及通过对来源于染色质片段化的DNA进行大规模平行测序和文库构建来测量分布(a)、(b)和/或(c),该片段化使用酶(例如微球菌核酸酶、DNA酶或转座酶)进行;以及(iii)通过比较来源于来自该生物样品的cfDNA的核小体图谱与核小体图谱参考集合来确定来自该生物样品的导致cfDNA的产生的组织和/或细胞类型;其中(a)、(b)和(c)是:(a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;(b)人基因组碱基对中任意一对呈现为一对测序的片段终端的可能性分布;和(c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

[0015] 在其他实施方式中,本发明提供了用于诊断对象中临床病症的方法,该方法包括:(i)通过从该对象获得生物样品来生成核小体图谱,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);(ii)通过从一个或多个

具有已知疾病的对照对象中获得生物样品来生成核小体图谱参考集合,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);以及(iii)通过比较来源于来自该生物样品的cfDNA的核小体图谱与核小体图谱参考集合来确定临床病症;其中(a)、(b)和(c)是:(a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;(b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和(c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

[0016] 在其他实施方式中,本发明提供了用于诊断对象中临床病症的方法,该方法包括:(i)通过从该对象获得生物样品来生成核小体图谱,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);(ii)通过从一个或多个具有已知疾病的对照对象中获得生物样品来生成核小体图谱参考集合,从该生物样品中分离cfDNA,通过对来源于染色质片段化的DNA进行大规模平行测序和文库构建来测量分布(a)、(b)和/或(c),该片段化使用酶(例如微球菌核酸酶(MNase)、DNA酶或转座酶)进行;以及(iii)通过比较来源于来自该生物样品的cfDNA的核小体图谱与核小体图谱参考集合以确定来自该生物样品的cfDNA的来源组织组成;其中(a)、(b)和(c)是:(a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;(b)人基因组碱基对中任意一对呈现为一对测序的片段终端的可能性分布;和(c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

[0017] 下文中更详细地描述了这些和其他实施方式。

[0018] 附图简要说明

[0019] 图1显示以小基因组区域作为示例的三种信息类型,其将cfDNA片段化模式与核小体占位相关联。这些相同的信息类型也可通过染色质片段化产生,该片段化使用酶(例如微球菌核酸酶(MNase)、DNA酶或转座酶)进行。图1A显示测序的片段终端(即片段化的端点)处出现人基因组中任何特定碱基对的可能性分布;图1B显示人基因组碱基对中任意一对呈现为一对测序的片段终端(即导致单个分子的产生的连续对片段化端点)的可能性分布;且图1C显示人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段内(即相对覆盖)的可能性分布。

[0020] 图2显示典型cfDNA测序文库的插入尺寸分布;此处显示了来源于人血浆的混合cfDNA样品,该人血浆含有来自未知数目的健康个体的贡献(主体.cfDNA(bulk.cfDNA))。

[0021] 图3A显示在全部cDNA样品(血浆)、来自肿瘤患者样品的cfDNA(肿瘤)、来自妊娠女性个体的cfDNA(妊娠)、不同人细胞系的MNase(细胞系)和人DNA鸟枪测序文库(鸟枪)中的平均周期图强度,其来自映射至第一条(chr1)人常染色体的读数起始坐标(read start coordinate)的快速傅里叶变换(FFT)。

[0022] 图3B显示在全部cDNA样品(血浆)、来自肿瘤患者样品的cfDNA(肿瘤)、来自妊娠女性个体的cfDNA(妊娠)、不同人细胞系的MNase(细胞系)和人DNA鸟枪测序文库(鸟枪)中的平均周期图强度,其来自映射至最后一条(chr22)人常染色体的读数起始坐标的快速傅里叶变换(FFT)。

[0023] 图4显示在所有常染色体中的10千碱基对(kbp)区块中196碱基对(bp)周期处强度的前三种主成分(PC)。图4A显示PC 2对比PC 1;图4B显示PC 3对比PC 2。

[0024] 图5显示在所有常染色体中的10kbp区块中196bp周期处测量的强度的欧几里得距离的递阶聚类系统树图(hierarchical clustering dendogram)。

[0025] 图6显示在所有常染色体中的10kbp区块中181bp至202bp周期处强度的前三种主成分。图6A显示PC 2对比PC 1;图6B显示PC 3对比PC 2。

[0026] 图7显示在所有常染色体中的10kbp区块中181bp至202bp周期处测量的强度的欧几里得距离的递阶聚类系统树图。

[0027] 图8显示cfDNA数据集的在所有常染色体中的10kbp区块中181bp至202bp周期处强度的主成分分析(10个PC中的前7个):图8A显示PC 2对比PC 1;图8B显示PC 3对比PC 2;图8C显示PC 4对比PC 3;图8D显示PC 5对比PC 4;图8E显示PC 6对比PC 5;图8F显示PC 7对比PC 6。

[0028] 图9显示MNase数据集的在所有常染色体中的10kbp区块中181bp至202bp周期处强度的主成分分析:图9A显示PC 2对比PC 1;图9B显示PC 3对比PC 2;图9C显示PC 4对比PC 3;图9D显示PC 5对比PC 4;图9E显示PC 6对比PC 5。

[0029] 图10显示在所有合成的cfDNA和MNase数据集混合物中的代表性人常染色体(chr11)的平均周期图强度。

[0030] 图11显示合成的MNase数据集混合物的在所有常染色体中的10kbp区块中181bp至202bp周期处强度的前两个主成分。

[0031] 图12显示合成的cfDNA数据集混合物的在所有常染色体中的10kbp区块中181bp至202bp周期处强度的前两个主成分。

[0032] 图13显示合成的MNase和cfDNA混合数据集的在所有常染色体中的10kbp区块中181bp至202bp周期处强度的欧几里得距离的递阶聚类系统树图。

[0033] 图14显示具有至少100M读数的样品集合的23666个CTCF结合位点周围1kbp窗口中的读数起始密度。

[0034] 图15显示具有至少100M读数的样品集合的5644个c-Jun结合位点周围1kbp窗口中的读数起始密度。

[0035] 图16显示具有至少100M读数的样品集合的4417个NF-YB结合位点周围1kbp窗口中的读数起始密度。

[0036] 图17显示导致cfDNA片段的产生的过程的概述图。凋亡和/或坏死性细胞死亡导致天然染色质的接近完全的消化。通常与组蛋白或转录因子相关的蛋白质结合的DNA片段优先在消化后继续存在并被释放至循环系统中,而裸DNA丧失。可在蛋白酶处理后从外周血浆中回收片段。在健康个体中,cfDNA主要来源于骨髓样和淋巴样细胞谱系,而来自一种或多种额外组织的贡献物可存在于某些医学病症中。

[0037] 图18显示使用常规测序文库制备物观察到的cfDNA的片段长度。长度推导自双端(paired-end)测序读数的比对。167碱基对(bp)处(绿色虚线)片段长度的可重复峰与同染色质小体相关性一致。额外的峰印证了约10.4bp周期,对应于核小体核心上DNA的螺距。文库制备期间的酶促末端修复移除5'和3'突出并可模糊真正的切割位点。

[0038] 图19显示常规文库中167bp片段和侧翼基因组序列的二核苷酸(dinucleotide)组成。BH01文库中观察到的二核苷酸频率与来自模拟片段的预期频率相比较(匹配切割和接头连接偏爱所致末端偏好(endpoint biases))。

[0039] 图20显示cfDNA片段的单链文库制备方案的示意图。

[0040] 图21显示使用单链测序文库制备物观察到的cfDNA的片段长度。文库制备期间不进行酶促末端修复以对分子进行模板化。与常规文库相比50-120bp的短片段高度富集。而约10.4bp周期仍存在,其相位(phase)偏移了约3bp。

[0041] 图22显示单链文库中167bp片段和侧翼基因组序列的二核苷酸组成。IH02文库中观察到的二核苷酸频率与来自模拟片段的预期频率相比较,再次匹配末端偏好。BH01和IH02之间偏好的背景水平的明显差异涉及模拟物之间的差别,而非真实文库(数据未显示)。

[0042] 图23A显示使用常规方案制备的代表性cfDNA测序文库的凝胶图像。

[0043] 图23B显示使用单链方案制备的代表性cfDNA测序文库的凝胶图像。

[0044] 图24A显示cfDNA片段的单核苷酸切割偏好。

[0045] 图24B显示cfDNA片段的二核苷酸切割偏好。

[0046] 图25显示核苷酸定位推导的示意图。通过从完全横跨窗口的片段数目中扣除120bp窗口内的片段末端数目来计算逐一碱基(per-base)的窗口化保护评分(WPS)。高WPS值显示增加的对抗消化的DNA保护;低数值显示DNA未被保护。峰识别信号(call)鉴定到提高的WPS的连续区域。

[0047] 图26显示充分研究的α卫星阵列处强势定位的核小体。显示了第12号染色体上中心体周围基因座处长片段(120bp窗口;120-180bp读数)或短片段(16bp窗口;35-80bp读数)箱(bin)的来自样品CH01的覆盖、片段末端和WPS值。来自CH01的核小体识别信号(中间,蓝色盒)在所有基因座中都被有规律地隔开。还显示来自两项公开的研究的基于MNase消化的核小体识别信号(中间,紫色和黑色盒)。基因座与标注的α卫星阵列重叠。

[0048] 图27显示第9号染色体上DNA酶I超敏位点(DHS)周围推导的核小体定位。显示了长和短片段箱的来自样品CH01的覆盖、片段末端和WPS值。灰色高亮显示的超敏区域由长片段箱中降低的覆盖标记。与DHS相邻的来自CH01的核小体识别信号(中间,蓝色盒)具有比典型的相邻对更宽的间隔,与干扰序列对包括转录因子的调节蛋白质的可及性一致。与这类蛋白质相关的较短片段的覆盖在DHS处提高,其与若干标注的转录因子结合位点重叠(未显示)。来自两项公开的研究的基于MNase消化的核小体识别信号如图26所示。

[0049] 图28显示根据本发明的一个实施方式的峰识别和评分的示意图。

[0050] 图29通过GC含量显示CH01峰密度。

[0051] 图30显示逐一样品的相邻峰之间距离的曲线。测量从峰识别信号至相邻识别信号的距离。

[0052] 图31显示样品之间峰识别信号的比较。对于各对样品,计算具有较少峰的样品中的各峰识别信号与其他样品中最近峰识别信号之间的距离并将其可视化为箱尺寸为1的统计图。负值表明最近的峰是上游;正值表明最近的峰是下游。

[0053] 图32显示样品之间峰识别信号的比较。图32A显示IH01对比BH01;图32B显示IH02对比BH01;图32C显示IH02对比IH01。

[0054] 图33A显示真实的峰对比模拟的峰的核小体评分。

[0055] 图33B显示根据评分箱的评分箱内的中值峰偏移(左y轴)和各评分箱中的峰数目(右y轴)。

[0056] 图34显示样品和匹配的模拟物之间峰识别信号的比较。图34A显示BH01模拟物对比BH01实际;图34B显示IH01模拟物对比IH01实际;图34C显示IH02模拟物对比IH01实际。

[0057] 图35显示相邻峰之间的距离,样品CH01。黑色虚线表明分布的模式(185bp)。

[0058] 图36显示汇总并调节的22626个转录起始位点(TSS)周围的窗口化保护评分(WPS;120bp窗口)。调节链和转录方向后,TSS对齐在0位置处。通过加和相对于居中TSS的各位置处逐一TSS的WPS来针对真实数据和模拟数据对汇总的WPS进行制表。标绘的值表示真实和模拟的汇总的WPS之间的差异,如下文详述的那样针对局部背景对其进一步调节。较高的WPS值表示优先保护免于切割。

[0059] 图37显示汇总和调节的22626个起始密码子周围的WPS。

[0060] 图38显示汇总和调节的224910个剪接供体位点周围的WPS。

[0061] 图39显示汇总和调节的224910个剪接受体位点周围的WPS。

[0062] 图40显示汇总和调节的来自CH01的数据的多个基因特征周围的WPS,该数据包括真实数据、匹配的模拟数据及其差异。

[0063] 图41显示A/B分区(compartment)中的核小体间隔。在基因组范围上计算非重叠的100千碱基(kb)箱中的中值核小体间隔,各箱含有约500个核小体识别信号。同样是100kb分辨率的GM12878的A/B分区预测来自公开的来源。区室A与开放的染色质相关且区室B与关闭的染色质相关。

[0064] 图42显示第7和11号染色体上的核小体间隔和A/B分区。A/B区段(红色和蓝色条)很大程度上扼要说明了染色体G显带(G-banding)(模式图,灰色条)。中值核小体间隔(黑点)在100kb箱中计算并标绘在A/B区段上方。

[0065] 图43显示汇总和调节的针对长(顶部)和短(底部)组分(fraction)的93500个CTCF位点的WPS。

[0066] 图44显示汇总和调节的CTCF位点处短组分cfDNA的放大视图。浅红色条(和相应的图内阴影)显示已知的52bp CTCF结合基序的位置。该条的暗红色分段显示用于FIMO基序搜索的17bp基序的位置。

[0067] 图45显示以下CTCF位点周围计算的-1至+1核小体间隔:来源于聚类FIMO预测的CTCF位点(单纯基于基序的:518632个位点),与ENCODE ChIP-seq峰重叠的这些预测结果的子集(93530个位点)和通过实验观察到的在19种细胞系中都活跃的进一步子集(23723个位点)。最不严格的CFCT位点集合主要相隔与基因组范围平均值(约190bp)大致相同的距离。然而,在最高严格性处,大多数CTCF位点相隔宽得多的距离(约260bp),这与相邻核小体的重定位和活跃CTCF结合一致。

[0068] 图46-48显示核小体侧翼的CTCF占位复位:图46显示FIMO预测的518632个CTCF结合位点的三个最接近上游和三个最接近下游峰识别信号的峰间距离。图47显示如同图46中那样的FIMO预测的518632个CTCF结合位点的三个最接近上游和三个最接近下游峰识别信号的峰间距离,但其中已基于与ENCODE ChIP-seq峰的重叠过滤了相同的CTCF位点集合,剩余93530个位点。图48显示如同图47中那样的FIMO预测的93530个CTCF结合位点的三个最接近上游和三个最接近下游峰识别信号的峰间距离,但其中已基于与在19种细胞系中都通过实验观察到的活跃CTCF位点的集合的重叠过滤了CTCF位点集合,剩余23732个位点。

[0069] 图49显示,对于侧翼核小体间隔较宽(230-270bp)的推定的CTCF位点的子集,长

(顶部)和短(底部)组分都表现出严格性提高的CTCF位点子集的较强定位信号。对于关键的限定彩色线,参见图45。

[0070] 图50-52显示核小体侧翼的CTCF占位重定位:图50显示518632个位点的平均短组分WPS(顶部小图)和平均长组分WPS(底部小图),其被划分为显示针对各位点的分隔侧翼+1和-1核小体识别信号的碱基对数目的距离箱。图51显示图50的518632个位点的平均短组分WPS(顶部小图)和平均长组分WPS(底部小图),但其中已基于与ENCODE ChIP-seq峰的重叠过滤了相同的CTCF位点集合。图52显示图51的位点的平均短组分WPS(顶部小图)和平均长组分WPS(底部小图),但其中已基于与在19种细胞系中都通过实验观察到的活跃CTCF位点的集合的重叠过滤了相同的位点集合。图50中关键的限定彩色线与图51和图52相同。

[0071] 图53A-H显示来自短和长cfDNA片段的转录因子结合位点的足迹。将聚类FIMO结合位点预测结果与ENCODE ChIP-seq数据交集以获得针对额外因子集合的转录因子(TF)结合位点的可信集合。显示了长和短cfDNA片段的组分为所有TF结合位点的侧翼的汇总并调节的区域的WPS。显示针对所得TF结合位点集合侧翼区域的汇总和调节的WPS。较高的WPS值分别表示较高的核小体或TF占据可能性。图53A:AP-2;图53B:E2F-2;图53C:EBOX-TF;图53D:IRF;图53E:MYC-MAX;图53F:PAX5-2;图53G:RUNX-AML;图53H:YY1。

[0072] 图54显示汇总和调节的转录因子ETS的WPS(210798个位点)。显示由长(顶部)和短(底部)cfDNA组分计算的WPS。观察到与结合位点本身处TF保护一致的信号(短组分)以及周围核小体的组织(长组分)。额外TF的类似分析示于图53A-H。

[0073] 图55显示汇总和调节的转录因子MAFK的WPS(32159个位点)。显示由长(顶部)和短(底部)cfDNA组分计算的WPS。观察到与结合位点本身处TF保护一致的信号(短组分)以及周围核小体的组织(长组分)。额外TF的类似分析示于图53A-H。

[0074] 图56显示基于DNA酶超敏(DHS)位点的对无细胞DNA有贡献的细胞类型混合物的推导。来自116种多样化生物样品的DHS位点处核小体识别信号的峰至峰间隔的频率分布显示双模式(bimodal)分布,其中第二模式可能对应于干扰的转录因子结合所导致的活跃DHS位点处加宽的核小体间隔(约190bp→260bp)。淋巴样或骨髓样样品中鉴定的DHS位点具有最大比例的具有加宽核小体间隔的DHS位点,这与作为健康个体中cfDNA主要来源的造血细胞死亡相一致。

[0075] 图57显示将转录起始位点(TSS)周围的经调节WPS评分划分为针对NB-4(一种急性前髓细胞性白血病细胞系)限定的五种基因表达箱(五分位)如何揭示核小体的间隔和布局中的差异。高表达基因显示转录本体(transcript body)内的强核小体定相(phasing)。在TSS上游,-1核小体在各表达箱中都良好定位,但-2和-3核小体仅对中等至高度表达基因良好定位。

[0076] 图58显示,对于中等至高度表达基因,在TSS和-1核小体之间观察到短片段峰,这与转录活跃基因处转录预起始复合物(transcription preinitiation complex)或其一些成分的足迹一致。

[0077] 图59显示NB-4细胞系中测得的转录本体(transcript body)中的中值核小体距离与基因表达负关联(ρ=-0.17,n=19677个基因)。低表达至无表达的基因显示193bp的中值核小体距离,而对于表达的基因,该范围是186-193bp。在使用更多的核小体识别信号来确定更精确的中值距离时该负关联更强(例如需要至少60个核小体,ρ=-0.50;n=12344个基

因)。

[0078] 图60显示如何为了对多个分布进行去卷积而使用快速傅里叶变换(FFT)来定量各TSS处起始的基因体的前10kb的长片段WPS中特定频率分布的丰度(强度)。显示了在不同频率下具有这些强度的76种细胞系和基本组织中RNA表达之间关联的轨迹线。粗黑线标记的是NB-4细胞系。在193-199bp频率范围的强度量级中关联最强。

[0079] 图61显示健康状态和癌症中对无细胞DNA有贡献的细胞类型的推导。顶部小图显示多种cfDNA文库的具有193-199bp频率范围中平均强度的76种RNA表达数据集的关联排名,通过类型对其进行分类并按从最高排名(顶部行)至最低排名(底部行)排列。关联值和完整的细胞系或组织名称参见表3。所有三种健康样品(BH01、IH01和IH02,前三列)的所有最强关联都是与淋巴样和骨髓样细胞系以及骨髓。相反地,获自IV期癌症患者(IC15、IC17、IC20、IC35、IC37;后五列)的cfDNA样品显示与多种癌症细胞系的高关联,例如IC17(肝细胞癌,HCC)显示与HepG2(肝细胞癌细胞系)的最高关联;且IC35(乳腺导管癌,DC)显示与MCF7(转移性乳房腺癌(metastatic breast adenocarcinoma)细胞系)的最高关联。比较针对癌症样品与三种健康样品中每一种所观察到的细胞类型/组织排名和对排名变化进行平均时(底部小图),与彼此之间比较三种健康样品和对排名变化进行平均('对照')时所观察到的相比,最大排名变化要高超过2倍。例如,对于IC15(小细胞肺癌,SCLC)而言SCLC-21H(小细胞肺癌细胞系)提高了平均31个位置,对于IC20(鳞状细胞肺癌,SCC)而言SK-BR-3(转移性乳房腺癌)提高了平均21位排名,且对于IC37(结直肠腺癌,AC)而言HepG2提高了24位排名。

[0080] 图62显示定量非整倍性(aneuploidy)以选择具有最高循环肿瘤cfDNA负荷的样品,其基于覆盖(图62A)或等位基因平衡(图62B)。图62A显示:与假定无非整倍性的模拟样品(红点)相比,各样品(黑点)的基于观察到的和预期的测序读数数目计算的各染色体的Z评分总和。图62B显示各染色体单独评价的48800个常见SNP中每一个处的等位基因平衡,其针对经选择用于额外测序的样品子集。

[0081] 图63显示峰识别信号与公开的核小体识别信号集合的比较:图63A显示在三个公开数据集(Gaffney等2012,JS Pedersen等2014,和A Schep等2015)中的核小体峰识别信号与此处生成的识别信号之间的距离,包括匹配的CA01的模拟物。先前公开的数据集未显示典型的约185bp核小体距离处的限定的模式,这可能是由于其稀疏的选样或宽的识别信号范围。相反地,来自cfDNA的所有核小体读数都显示一种充分限定的模式。匹配的模拟数据集具有较短的模式(166bp)和较宽的分布。此外,用于生成识别信号的cfDNA数据的覆盖越高,分布模式所代表的识别信号比例越高。图63B显示与图63A相同的集合列表中每一个的核小体数目。cfDNA核小体识别信号呈现具有接近13M核小体峰识别信号的最广泛的识别信号集合。图63C显示IH01 cfDNA样品中各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离。图63D显示IH02cfDNA样品中各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离。图63E显示BH01 cfDNA样品中各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离。图63F显示CH01cfDNA样品中各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离。图63G显示CA01 cfDNA样品中各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离。负数表示最近峰是上游;正数表示最近峰是下游。随着cfDNA覆盖的增加,在较接近确定的核小体识别信号处发现了较高比例的先前公开的识别信号。最高的一致性出现在由

Gaffney等,PLoS Genet.,卷8,e1003036(2012)和A Schep等(2015)生成的识别信号中。图63H显示各峰识别信号与来自三个先前公开的数据集的最近峰识别信号之间的距离,但此时是针对匹配的CA01的模拟物。对于Gaffney等,PLoS Genet.,卷8,e1003036(2012)和JS Pedersen等,Genome Research,卷24,454-466页(2014)识别信号,最近的真实核小体位置倾向于远离模拟物中识别的峰。由A Schep等(2015)生成的识别信号似乎显示与模拟识别信号的一些重叠。

[0082] 发明详述

[0083] 本发明提供了确定对象的生物样品中一种或多种导致无细胞DNA的产生的组织和/或细胞类型的方法。在一些实施方式中,本发明提供了根据来自对象的生物样品中一种或多种确定的与cfDNA相关的组织和/或细胞类型鉴定该对象中疾病或紊乱的方法。

[0084] 本发明基于以下预测:源自不同细胞类型的cfDNA分子在以下方面是不同的:(a)cfDNA片段终端处出现人基因组中任何特定碱基对(即片段化的端点)的可能性分布;(b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端(即导致单个cfDNA分子的产生的连续对片段化端点)的可能性分布;和(c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中(即相对覆盖)的可能性分布。这在下文中称作分布(a)、(b)和(c),或统称为“核小体依赖性切割可能性图谱”、“切割可及性图谱”或“核小体图谱”(图1)。应注意,还可通过对以下片段进行测序来测量核小体图谱:这些片段来源于使用酶(例如微球菌核酸酶(MNase)、DNA酶或转座酶)进行的染色质片段化或优先对核小体或染色质小体的边界之间或边界处的基因组DNA进行分段的等价程序。

[0085] 在健康个体中,cfDNA绝大部分来源于血细胞,即造血谱系细胞的凋亡。随着这些细胞经历程序性细胞死亡,其基因组DNA被切割并释放至循环系统,其在循环系统中继续被核酸酶降解。cfDNA的长度分布以约10.5碱基对(bp)的周期波动,其对应于核小体周围围绕的螺旋DNA的螺距,且具有约167bp的显著峰,其对应于连接有连接子的单核小体相关的DNA长度(图2)。该证据导致以下假说:cfDNA与核小体的关连保护其免于循环系统中完全、快速的降解。另一种可能性是长度分布简单地来自凋亡本身期间DNA切割的模式,是直接受核小体定位的影响。无论如何,cfDNA的长度分布提供了清楚的证据,证明导致cfDNA的产生的片段化过程受核小体定位的影响。

[0086] 在一些实施方式中,本发明将核小体图谱定义为通过来自体液的cfDNA或来源于染色质片段化或等价程序的DNA的文库构建和大规模平行测序进行的分布(a)、(b)和/或(c)的测量,该片段化使用酶(例如微球菌核酸酶(MNase)、DNA酶或转座酶)进行,这些等价程序优先片段化染色质小体或核小体的边界之间或边界处的DNA片段。如下文所述,可‘变换’这些分布以聚集或汇总多个基因组子集内核小体定位的周期性信号(例如定量连续窗口中的周期),或者由转录因子结合位点、基因模式特征(如转录起始位点或基因体)、拓扑学相关结构域、组织表达数据或其他核小体定位关联物所定义的基因组的不连续子集内核小体定位的周期性信号。此外,这些可通过组织特异性数据定义。例如,可聚集或汇总组织特异性DNA酶I超敏位点周边的信号。

[0087] 本发明提供由血浆传播的cfDNA片段所推导的体内核小体保护的致密、基因组范围的图谱。来源于健康个体的cfDNA的CH01图谱包括核小体保护的接近13M均匀间隔的局部极大值,其跨越绝大部分的可映射(mappable)人参考基因组。虽然CH01中的峰数目基本饱

和,其他质量度量继续成为测序深度的函数(图33A-B)。因此针对本研究和其他工作构建了额外的基因组范围核小体图谱——通过相同的方法——其基于迄今为止发明人进行的几乎所有cfDNA测序('CA01',145亿(G)片段;700倍覆盖;13.0M峰)。虽然该图谱显示甚至更均匀的间隔和更高度支持的峰识别(图33A-B,63A-H),仍应注意其基于来自健康和非健康个体的cfDNA(表1,5)。

[0088] 本发明所述的核小体保护的致密、基因组范围的图谱接近了人参考基因组的可映射部分的饱和,且与先前生成核小体定位或保护的人基因组范围图谱的努力相比,峰与峰的间隔显著更均匀且与预期的核小体重复长度显著更一致(图63A-H)。与几乎所有先前努力相反,本发明观察到的片段是通过内源性生理过程生成的,且因此较不可能经历体外微球菌核酸酶消化相关的技术差异(technical variation)。该参考图谱中考虑的导致cfDNA的产生的细胞类型不可避免地是异质性的(例如,健康个体中淋巴样和骨髓样细胞类型的混合物)。然而,图谱的相对完整性可促进更深入地理解人细胞中支配核小体定位和间隔的过程,以及核小体与表观遗传调控、转录输出和细胞核架构的相互影响。

[0089] 确定对象的生物样品中cfDNA来源的方法

[0090] 如上文笼统讨论的那样,且如后文实施例中更具体所示,本技术可用于确定(如预测)对象的生物样品中对cfDNA有贡献的组织和/或细胞类型。

[0091] 因此,在一些实施方式中,本发明提供了确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,该方法包括:从来自该对象的生物样品中分离cfDNA,分离的cfDNA包含复数个cfDNA片段;确定与复数个cfDNA片段的至少一部分相关的序列;根据cfDNA片段序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;以及根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA片段的产生的组织和/或细胞类型。

[0092] 在一些实施方式中,该生物样品包括以下物质、基本由以下物质组成或由以下物质组成:全血、外周血浆、尿液或脑脊液。

[0093] 在一些实施方式中,确定至少一些导致cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较至少一些cfDNA片段末端的基因组位置或其分布的数学变换。本文中使用的术语"参考图谱"指任何类型或形式的以下数据:其可根据cfDNA序列所比对的基因组(例如参考基因组)内的坐标与对象的生物样品中的cfDNA的属性相关联或比较。参考图谱可通过任何合适方法与对象的生物样品中的cfDNA的属性相关联或比较。例如但不限于,该关联或比较可通过分析对象的生物样品中cfDNA末端的频率来实现,其可直接进行或在对参考基因组内其在所有窗口中的分布进行数学变换后进行,上述方法基于针对参考基因组的等价坐标由参考图谱限定的数值或任何其他状态。在另一个非限制性实施例中,该关联或比较可通过分析基于对象的生物样品的cfDNA的经测定的核小体间隔来实现,上述方法基于参考图谱中经测定的核小体间隔或与核小体间隔相关联的另一种特性。

[0094] 该参考图谱可以源自或来源于任何合适的数据源,包括例如公共的基因组信息数据库、公开的数据或针对参考对象的特定群体生成的数据(其各自可具有常见的属性,如疾病状态)。在一些实施方式中,该参考图谱包括DNA酶I超敏数据集。在一些实施方式中,该参考图谱包括RNA表达数据集。在一些实施方式中,该参考图谱包括包括染色体构象图谱。在一些实施方式中,该参考图谱包括染色质可及性图谱。在一些实施方式中,该参考图谱包括

由至少一种与疾病或紊乱相关的组织或细胞类型生成的数据。在一些实施方式中,该参考图谱包括组织或细胞类型中核小体和/或染色质小体的位置。在一些实施方式中,该参考图谱通过包括使用外源性核酸酶(如微球菌核酸酶)消化染色质的程序生成。在一些实施方式中,该参考图谱包括通过基于转座的方法(如ATAC-seq)测得的染色质可及性数据。在一些实施方式中,该参考图谱包括组织或细胞类型的与DNA结合和/或DNA占据蛋白质的位置相关的数据。在一些实施方式中,该DNA结合和/或DNA占据蛋白质是转录因子。在一些实施方式中,这些位置通过包括交联的DNA-蛋白质复合物的染色质免疫沉淀的过程测得。在一些实施方式中,这些位置通过包括使用核酸酶(如DNA酶I)处理组织或细胞类型相关DNA的过程测得。在一些实施方式中,该参考图谱通过对来自生物样品的cfDNA片段进行测序生成,该生物样品来自一个或多个具有已知疾病的个体。在一些实施方式中,生成参考图谱的生物样品收集自动物,其中已将人细胞或组织异种移植至该动物。

[0095] 在一些实施方式中,该参考图谱包括组织或细胞类型的对应于DNA结合或DNA占据蛋白质位置的生物特征。在一些实施方式中,该参考图谱包括对应于一种或多种基因的定量RNA表达的生物特征。在一些实施方式中,该参考图谱包括对应于存在或不存在一种或多种组蛋白标记的生物特征。在一些实施方式中,该参考图谱包括对应于对核酸酶切割超敏的生物特征。

[0096] 与一个或多个参考图谱比较至少一些cfDNA片段末端的基因组位置的步骤可以多种方式实现。在一些实施方式中,由生物样品生成的cfDNA数据(例如cfDNA片段的基因组位置、其末端、其末端的频率和/或由其分布推导的核小体间隔)与超过一种参考图谱比较。在这类实施方式中,生物样品中与cfDNA数据的关联程度最高的参考图谱相关的组织或细胞类型被视作有贡献(contributing)。例如但不限于,如果cfDNA数据包括可能的cfDNA末端及其在参考基因组内位置的列表,具有最类似cfDNA末端及其在参考基因组内位置的参考图谱可视作有贡献。作为另一个非限制性示例,与来自生物样品的cfDNA片段末端的分布的数学变换具有最高关联度(或增加的关联度,相对于来自健康对象的cfDNA)的参考图谱被视作有贡献。对应于那些被视作有贡献的参考图谱的组织类型和/或细胞类型随后被认为是分离自生物样品的cfDNA的潜在来源。

[0097] 在一些实施方式中,确定至少一些导致cfDNA片段的产生的组织和/或细胞类型的步骤包括对至少一些cfDNA片段末端的基因组位置的分布进行数学变换。适用于与本发明联用的数学变换的一个非限制性示例是傅里叶变换,如快速傅里叶变换(“FFT”)。

[0098] 在一些实施方式中,该方法还包括确定参考基因组的至少一些坐标中每一个的评分,其中根据至少复数个cfDNA片段末端及其基因组位置确定评分,且其中确定至少一些导致观察到的cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较这些评分。该评分可以是任何度量(例如数字排名或可能性),用于将相对或绝对值分配至参考基因组的坐标。例如,该评分可由可能性组成或与可能性相关,例如坐标代表cfDNA片段末端位置的可能性或坐标代表针对核酸酶切割而优先通过核小体或蛋白质结合进行保护的基因组位置的可能性。作为另一个示例,该评分可涉及基因组具体区域中的核小体间隔,如该区域内cfDNA片段末端的分布的数学变换所测定的那样。这类评分可通过任何合适的方式分配至坐标,包括例如通过对具体坐标相关的绝对或相对事件(例如cfDNA片段末端的数目)进行计数,或对该区域或基因组坐标中这类计数的值进行数学变换。在一些实施

方式中,坐标的评分与该坐标是cfDNA片段末端位置的可能性相关。在其他实施方式中,坐标的评分与该坐标代表针对核酸酶切割而优先通过核小体或蛋白质结合进行保护的基因组位置的可能性相关。在一些实施方式中,该评分与该坐标的基因组区域中的核小体间隔相关。

[0099] 本发明所述方法中提及的组织和/或细胞类型可以是导致cfDNA的产生的任何组织或细胞类型。在一些实施方式中,该组织或细胞类型是来自具有疾病或紊乱的对象的基本组织。在一些实施方式中,该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0100] 在一些实施方式中,该组织或细胞类型是来自健康对象的基本组织。

[0101] 在一些实施方式中,该组织或细胞类型是永生细胞系。

[0102] 在一些实施方式中,该组织或细胞类型是来自肿瘤的活检切片。

[0103] 在一些实施方式中,该参考图谱基于获自样品的序列数据,该样品获自至少一个参考对象。在一些实施方式中,该序列数据定义参考基因组内cfDNA片段末端的位置——例如,是否通过对来自具有已知疾病的对象的cfDNA进行测序来生成参考图谱。在其他实施方式中,作为参考图谱基础的该序列数据可包括以下任意一项或多项:DNA酶I超敏位点数据集、RNA表达数据集、染色体构象图谱、染色质可及性图谱,或使用微球菌核酸酶消化染色质生成的核小体定位图谱。

[0104] 在一些实施方式中,该参考对象是健康的。在一些实施方式中,该参考对象具有疾病或紊乱,其任选地选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0105] 在一些实施方式中,该参考图谱包括与组织或细胞类型相关的参考基因组的至少一部分坐标的评分。在一些实施方式中,该参考图谱包括这些评分的数学变换,如这些评分的傅里叶变换。在一些实施方式中,这些评分是基于组织或细胞类型的参考基因组坐标的标注(annotation)。在一些实施方式中,这些评分是基于核小体和/或染色质小体的位置。在一些实施方式中,这些评分是基于转录起始位点和/或转录终止位点。在一些实施方式中,这些评分是基于至少一种转录因子的预测结合位点。在一些实施方式中,这些评分是基于预测的核酸酶超敏位点。在一些实施方式中,这些评分是基于预测的核小体间隔。

[0106] 在一些实施方式中,这些评分与至少一种正交生物特征相关。在一些实施方式中,该正交生物特征与高表达的基因相关。在一些实施方式中,该正交生物特征与低表达的基因相关。

[0107] 在一些实施方式中,复数个评分中的至少一些具有阈(最小)值以上的值。在这类实施方式中,落入阈(最小)值以下的评分被排除在比较评分与参考图谱的步骤之外。在一些实施方式中,该阈值在确定导致cfDNA的产生的组织和/或细胞类型前确定。在其他实施方式中,该阈值在确定导致cfDNA的产生的组织和/或细胞类型后确定。

[0108] 在一些实施方式中,根据至少一些cfDNA片段末端的复数个基因组位置确定导致cfDNA的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱的一个或多个特征比较样品的至少一些cfDNA片段末端的基因组位置的分布的数学变换。适用于该目的的数

学变换的一个非限制性示例是傅里叶变换,如快速傅里叶变换(“FFT”)。

[0109] 在本发明所述任意实施方式中,该方法还可包括生成报告,该报告包括确定的导致分离的cfDNA的产生的组织和/或细胞类型的列表。该报告还可任选地包括:样品和/或对象的任意其他消息、生物样品的类型、从对象中获得生物样品的日期、进行cfDNA分离步骤的日期和/或可能不导致任何分离自生物样品的cfDNA的产生的组织和/或细胞类型。

[0110] 在一些实施方式中,该报告还包括推荐的治疗方案,其包括,例如但不限于,从对象获得额外诊断测试的建议、开始治疗方案的建议、修改患者的现有治疗方案的建议,和/或中止或终止现有治疗方案的建议。

[0111] 鉴定对象中疾病或紊乱的方法

[0112] 如上文笼统描述和下文实施例中更具体说明的那样,本技术可用于确定(例如预测)疾病或紊乱,或不存在疾病或紊乱,其至少部分基于对象的生物样品中对cfDNA有贡献的组织和/或细胞类型。

[0113] 因此,在一些实施方式中,本发明提供了鉴定对象中疾病或紊乱的方法,该方法包括从来自对象的生物样品中分离无细胞DNA(cfDNA),分离的cfDNA包含复数个cfDNA片段;确定与复数个cfDNA片段中至少一些相关的序列;根据cfDNA片段的序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA的产生的组织和/或细胞类型;以及根据确定的导致cfDNA的产生的组织和/或细胞类型来鉴定该疾病或紊乱。

[0114] 在一些实施方式中,该生物样品包括以下物质、基本由以下物质组成或由以下物质组成:全血、外周血浆、尿液或脑脊液。

[0115] 在一些实施方式中,确定导致cfDNA的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较至少一些cfDNA片段末端的基因组位置或其分布的数学变换。对于确定对象的生物样品中导致cfDNA的产生的组织和/或细胞类型的方法而言,与这些实施方式联用的术语“参考图谱”可具有与上文相同的意义。在一些实施方式中,该参考图谱可包括以下任意一种或多种:DNA酶I超敏位点数据集、RNA表达数据集、染色体构象图谱、染色质可及性图谱、由获自至少一个参考对象的样品生成的序列数据、对应于与疾病或紊乱相关的至少一种组织的酶介导的片段化数据,和/或组织或细胞类型中核小体和/或染色质小体的位置。在一些实施方式中,该参考图谱通过对来自生物样品的cfDNA片段进行测序来生成,该生物样品来自一个或多个具有已知疾病的个体。在一些实施方式中,生成参考图谱的生物样品收集自动物,其中已将人细胞或组织异种移植至该动物。

[0116] 在一些实施方式中,该参考图谱是通过使用外源性核酸酶(如微球菌核酸酶)消化染色质来生成的。在一些实施方式中,这些参考图谱包括通过基于转座的方法(如ATAC-seq)测定的染色质可及性数据。在一些实施方式中,这些参考图谱包括针对组织或细胞类型的与DNA结合和/或DNA占据蛋白质的位置相关的数据。在一些实施方式中,该DNA结合和/或DNA占据蛋白质是转录因子。在一些实施方式中,这些位置是通过交联的DNA-蛋白质复合物的染色质免疫沉淀测定的。在一些实施方式中,这些位置是通过使用核酸酶(如DNA酶I)处理组织或细胞类型相关DNA来测定的。

[0117] 在一些实施方式中,该参考图谱包括针对组织或细胞类型的对应于DNA结合或DNA占据蛋白质的位置的生物特征。在一些实施方式中,该参考图谱包括对应于一种或多种基

因的定量表达的生物特征。在一些实施方式中,该参考图谱包括对应于存在或不存在一种或多种组蛋白标记的生物特征。在一些实施方式中,该参考图谱包括对应于对核酸酶切割超敏的生物特征。

[0118] 在一些实施方式中,确定导致cfDNA的产生的组织和/或细胞类型的步骤包括对复数个cfDNA片段末端中至少一些的基因组位置的分布进行数学变换。在一些实施方式中,该数学变换包括傅里叶变换。

[0119] 在一些实施方式中,该方法还包括确定参考基因组的至少一些坐标中每一个的评分,其中该评分是根据至少复数个cfDNA片段末端及其基因组位置确定的,且其中确定至少一些导致观察到的cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较这些评分。该评分可以是任何度量(例如数字排名或可能性),用于将相对或绝对值分配至参考基因组的坐标。例如,该评分可由可能性组成或与可能性相关,例如坐标代表cfDNA片段末端位置的可能性或坐标代表针对核酸酶切割而优先通过核小体或蛋白质结合进行保护的基因组位置的可能性。作为另一个示例,该评分可涉及基因组具体区域中的核小体间隔,如同该区域内cfDNA片段末端的分布的数学变换所测定的那样。可通过任何合适方法将这类评分分配至坐标,例如通过计数与该具体坐标相关的绝对或相对事件(例如cfDNA片段末端的数目),或对该区域或基因组坐标中的这类计数的值进行数学变换。在一些实施方式中,坐标的评分与该坐标是cfDNA片段末端位置的可能性相关。在其他实施方式中,坐标的评分与该坐标代表以下基因组位置的可能性相关:该基因组位置优先地通过核小体或蛋白质结合而被保护免于核酸酶切割。在一些实施方式中,该评分与坐标的基因组区域中的核小体间隔相关。

[0120] 对于确定对象的生物样品中导致cfDNA的产生的组织和/或细胞类型的方法而言,与这些实施方式联用的术语“评分”可具有与上文相同的含义。在一些实施方式中,坐标的评分与该坐标是cfDNA片段末端位置的可能性相关。在其他实施方式中,坐标的评分与该坐标代表以下基因组位置的可能性相关:该基因组位置优先地通过核小体或蛋白质结合而被保护免于核酸酶切割。在一些实施方式中,该评分与坐标的基因组区域中的核小体间隔相关。

[0121] 在一些实施方式中,用于生成参考图谱的组织或细胞类型是来自具有疾病或紊乱的对象的基本组织。在一些实施方式中,该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、系统性自身免疫病、局部自身免疫病、炎症性肠病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0122] 在一些实施方式中,该组织或细胞类型是来自健康对象的基本组织。

[0123] 在一些实施方式中,该组织或细胞类型是永生细胞系。

[0124] 在一些实施方式中,该组织或细胞类型是来自肿瘤的活检切片。

[0125] 在一些实施方式中,该参考图谱基于获自样品的序列数据,该样品获自至少一个参考对象。在一些实施方式中,该序列数据限定参考基因组内cfDNA片段末端的位置——例如,是否通过对来自具有已知疾病的对象的cfDNA进行测序来生成参考图谱。在其他实施方式中,作为参考图谱基础的该序列数据可包括以下任意一种或多种:DNA酶I超敏位点数据集、RNA表达数据集、染色体构象图谱,或染色质可及性图谱,或通过使用微球菌核酸酶消化生成的核小体定位图谱。在一些实施方式中,该参考对象是健康的。在一些实施方式中,该

参考对象具有疾病或紊乱。在一些实施方式中,该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、系统性自身免疫病、炎症性肠病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0126] 在一些实施方式中,该参考图谱包括针对至少一部分与该组织或细胞类型相关的参考基因组的cfDNA片段末端可能性或与这类可能性相关联的量。在一些实施方式中,该参考图谱包括cfDNA片段末端可能性或与这类可能性相关联的量的数学变换。

[0127] 在一些实施方式中,该参考图谱包括与该组织或细胞类型相关的参考基因组的至少一部分坐标的评分。在一些实施方式中,该参考图谱包括这些评分的数学变换,如这些评分的傅里叶变换。在一些实施方式中,这些评分基于该组织或细胞类型的参考基因组坐标的标注。在一些实施方式中,这些评分基于核小体和/或染色质小体的位置。在一些实施方式中,这些评分基于转录起始位点和/或转录终止位点。在一些实施方式中,这些评分基于至少一个转录因子的预测结合位点。在一些实施方式中,这些评分基于预测的核酸酶超敏位点。

[0128] 在一些实施方式中,这些评分与至少一种正交生物特征相关。在一些实施方式中,该正交生物特征与高表达基因相关。在一些实施方式中,该正交生物特征与低表达基因相关。

[0129] 在一些实施方式中,复数个评分中的至少一些各自具有阈值以上的评分。在这类实施方式中,落入阈(最小)值以下的评分被排除在比较评分与参考图谱的步骤之外。在一些实施方式中,在确定导致cfDNA的产生的组织和/或细胞类型之前确定阈值。在其他实施方式中,在确定导致cfDNA的产生的组织和/或细胞类型之后确定阈值。

[0130] 在一些实施方式中根据至少一些cfDNA片段末端的复数个基因组位置确定导致cfDNA的产生的组织和/或细胞类型的步骤包括对具有一个或多个参考图谱的一个或多个特征的样品的至少一些cfDNA片段末端的基因组位置的分布进行数学变换。

[0131] 在一些实施方式中,该数学变换包括傅里叶变换。

[0132] 在一些实施方式中,该参考图谱包括对应于与疾病或紊乱相关的至少一种组织的酶介导的片段化数据。

[0133] 在一些实施方式中,该参考基因组与人相关。

[0134] 在本发明的一个方面中,本发明所述方法被用于对来自体液中的cfDNA分析的恶性肿瘤进行检测、监测和来源组织和/或细胞类型评估。目前充分记载的是,在具有恶性肿瘤的患者中,体液(例如循环血浆)中的一部分cfDNA可来源于肿瘤。本发明所述方法可潜在地用于检测和定量该肿瘤来源部分。此外,由于核小体占位图谱是细胞类型特异的,本发明所述方法可潜在地用于确定恶性肿瘤的来源组织和/或细胞类型。另外,如上文所强调的那样,已观察到癌症中循环血浆cfDNA浓度的大幅增长,可能与来自肿瘤本身的贡献不成比例。这暗示其他组织(例如基质、免疫系统)可能对癌症期间的循环血浆cfDNA有贡献。这类其他组织对cfDNA的贡献程度在给定癌症类型的患者之间一致时,上文所述方法可基于来自这些其他组织而非癌细胞本身的信号促进癌症检测、监测和/或来源组织和/或细胞类型分配。

[0135] 在本发明的另一个方面中,本发明所述方法被用于对来自体液中的cfDNA分析的组织损伤进行检测、监测和来源组织和/或细胞类型评估。预期许多病理过程将生成来源于

受损组织的体液(例如循环血浆)中的一部分cfDNA。本发明所述方法可潜在地用于检测和定量来源于组织损伤的cfDNA,包括鉴定相关来源组织和/或细胞类型。这可促进病理过程的诊断和/或监测,这些病理过程包括心肌梗死(心脏组织的急性损伤)、自身免疫病(多种组织的慢性损伤),和涉及急性或慢性组织损伤的许多其他过程。

[0136] 在本发明的另一个方面中,本发明所述方法被用于预测妊娠中cfDMA的胎儿组分和/或增强染色体或其他遗传畸形的检测。母体血浆传播cfDNA片段的相对浅的测序,加上上文所述的核小体图谱,可允许节约成本且迅速地预测男性和女性胎儿妊娠中的胎儿组分。此外,通过促进将非均匀的可能性分配至个体测序读数(相对于其来源于母体或胎儿基因组的可能性),这些方法还可通过母体体液中cfDNA的分析来增强针对检测染色体畸变(例如三染色体(trisomies))的测试的表现。

[0137] 在本发明的另一个方面中,本发明所述方法被用于定量(自体或同种异体)移植物对cfDNA的贡献-现有的急性同种异体排斥的早期和非侵入性检测方法涉及对血浆传播的DNA进行测序并鉴定来源于供体基因组的片段浓度的升高。该方法依赖于该片段池的相对深度测序以检测例如5-10%供体组分。基于供体器官的核小体图谱作为代替的方法可促进使用较浅测序进行类似预测,或使用等量测序进行更灵敏的预测。与癌症类似,除移植物本身以外的细胞类型也可能在移植排斥期间对cfDNA有贡献组成。转移排斥期间这类其他组织对cfDNA的贡献程度在患者之间一致时,上文所述方法可基于来自这些其他组织而非移植供体细胞本身的信号促进对移植排斥的监测。

[0138] 本发明的其他实施方式

[0139] 本发明还提供了使用由具有已知疾病或紊乱的对象生成的核小体参考图谱来诊断疾病或紊乱的方法。在一些这类实施方式中,该方法包括:(1)生成核小体图谱参考集合,其中各核小体图谱来源于来自具有限定的临床病症(例如正常、妊娠、癌症类型A、癌症类型B等)的个体的体液的cfDNA和/或来源于特定组织和/或细胞类型的染色质的消化的DNA;(2)通过比较来源于其cfDNA的核小体图谱与核小体图谱参考集合来预测来自个体体液的cfDNA的来源组织/细胞类型组成和/或临床病症。

[0140] 步骤1:生成核小体图谱参考集合,并汇总或总结来自核小体定位的信号。

[0141] 生成核小体图谱的优选方法包括对来自体液的cfDNA进行DNA纯化、文库构建(通过接头连接和可能的PCR扩增)和大规模平行测序。可用于本发明的环境中作为参考点或用于鉴定变异的主成分的替代性核小体图谱来源是来源于以下过程的DNA:使用微球菌核酸酶(MNase)消化染色质、DNA酶处理、ATAC-Seq或其中在分布(a)、(b)或(c)中捕获涉及核小体定位的信息的其他相关方法。这些分布(a)、(b)和(c)的描述提供于[0078]并图示于图1。

[0142] 原则上,使用这类文库的非常深度测序来定量基因组中特定坐标处汇总的对cfDNA有贡献的细胞类型中的核小体占位,但目前这种方法非常昂贵。然而,可在所有基因组的连续或不连续区域中总结或汇总核小体占据模式相关的信号。例如,在本发明所述实施例1和2中,在10千碱基对(kbp)连续窗口中对参考人基因组(对其测序读数起始位点图谱)中位点的分布,即分布(a),进行傅里叶变换,随后定量与核小体占位相关的频率范围的强度。这有效地总结了各10kbp窗口内核小体表现出结构化定位的程度。在本发明提供的实施例3中,我们在紧邻特定转录因子(TF)的转录因子结合位点(TFBS)处(其在该TF结合该TFBS时通常侧翼紧接核小体)定量参考人基因组(对其测序读数起始位点图谱)中位点的分

布,即分布(a)。这有效地总结了对cfDNA有贡献的细胞类型中因TF活性而形成的核小体定位。重要的是,许多相关途径中可有意义地总结核小体占位信号。这些包括汇总其他基因组地标(landmark)周围的来自分布(a)、(b)和/或(c)的信号,这些其他基因组地标是例如DNA酶I超敏位点、转录起始位点、拓扑结构域、其他表观遗传标记或其他数据集(如基因表达等)中相关联行为所定义的所有这类位点的子集。随着测序成本持续降低,也可以直接使用核小体占位的图谱(包括由与已知疾病相关的cfDNA样品生成的那些)作为对照图谱,即不使用汇总信号,目的是与未知cfDNA样品比较。在一些实施方式中,该生物样品收集自动物,其中已将人细胞或组织异种移植至该动物,且核小体占据的参考图谱由该生物样品制备。这样做的优势是映射至人基因组的测序的cfDNA片段将仅来源于异种移植的细胞或组织,而不是代表来源于感兴趣的细胞/组织以及造血谱系的cfDNA的混合物。

[0143] 步骤2:直接或在各图谱的数学变换后基于比较一个或多个新个体/样品的cfDNA来源的核小体图谱与核小体图谱参考集合来预测病理学、临床症状和/或来源组织/细胞类型组成。

[0144] 一旦生成核小体图谱参考集合后,存在多种用于比较额外的核小体图谱与参考集合的统计信号处理方法。在实施例1和2中,我们首先总结多种样品集合中沿基因组的10kbp窗口内的长程核小体排序,并随后进行主成分分析(PCA)以聚类样品(实施例1)或预测混合物比例(实施例2)。虽然我们已知这些实施例中使用的所有细胞系样品的来源组织/细胞类型和所有cfDNA样品的临床膨胀,原则上任一样品都可以是“未知的”,以及其在用于预测存在/不存在临床病症的PCA分析中的行为或基于其在PCA分析中的行为的其来源组织/细胞类型(相对于所有其他核小体图谱)。

[0145] 未知样品并非必须以1:1的方式精确匹配至参考集合的+1成员。而是,可定量其各自间的相似性(实施例1)或可将其核小体图谱建模为来自参考集合的2+样品的非均匀混合物(实施例2)。

[0146] 对于本发明的方法成功实施而言,无需预测或完全知晓各样品中cfDNA的来源组织/细胞类型组成。确切而言,本发明所述方法依赖于具体病理学或临床病症环境中cfDNA的来源组织/细胞类型组成的一致性。然而,直接通过分析来源于染色质消化的DNA并将添加至核小体图谱来研究大量组织和/或细胞类型的核小体图谱,能够预测对未知的cfDNA有贡献的组织和/或细胞类型所来源的样品。

[0147] 在本发明所述任意实施方式中,该方法还可包括生成报告,该报告包括鉴定疾病或紊乱的声明(statement)。在一些实施方式中,该报告还包括确定的导致分离的cfDNA的产生的组织和/或细胞类型的列表。在一些实施方式中,该报告还包括不太可能与对象相关的疾病或紊乱的列表。该报告还可任选地包括:样品和/或对象的任何其他信息、生物样品的类型、从对象中获得生物样品的日期、进行cfDNA分离步骤的日期和/或可能不导致任何分离自生物样品的cfDNA的产生的组织和/或细胞类型。

[0148] 在一些实施方式中,该报告还包括推荐的治疗方案,其包括,例如但不限于,从对象获得额外诊断测试的建议、开始治疗方案的建议、修改患者的现有治疗方案的建议,和/或中止或终止现有治疗方案的建议。

实施例

[0149] 实施例1.无细胞DNA核小体图谱的主成分分析

[0150] 检查来源于cfDNA提取和MNase消化实验的测序数据中读数起始位置的分布以评估涉及核小体定位的信号的存在。出于此目的,分析了混合的cfDNA样品(含有来自未知数目的健康个体的贡献物的人血浆;主体.cfDNA)、来自单个健康男性对照个体的cfDNA样品(MC2.cfDNA)、来自具有颅内肿瘤的患者的四份cfDNA样品(肿瘤.2349、肿瘤.2350、肿瘤.2351、肿瘤.2353)、来自五种不同人细胞系的六份MNase消化实验物(Hap1.MNase、HeLa.MNase、HEK.MNase、NA12878.MNase、HeLaS3、MCF.7)和来自不同妊娠女性个体的七份cfDNA样品(gm1matplas、gm2matplas、im1matplas、fgs002、fgs003、fgs004、fgs005)并与提取自女性类淋巴母细胞细胞系(NA12878)的DNA的常规鸟枪测序数据集对比。还包括了混合cfDNA样品的子集(26%,主体.cfDNA_部分)和单个健康男性对照个体的子集(18%,MC2.cfDNA_部分)作为单独的样品以探索测序深度的影响。

[0151] 提取读数起始坐标并使用方法部分所述的快速傅里叶变换(FFT)创建周期图。该分析确定读数起始位点的分布的不均匀性中有多少可通过特定频率/周期的信号来解释。我们关注于120-250bp的范围,其包括围绕单个核小体缠绕的DNA的长度范围(147bp)以及核小体连接子序列的额外序列(10-80bp)。图3显示在跨越人第1号染色体和人第22号染色体中的所有区块的各频率的平均强度。可以看到,MNase消化实验和cfDNA样品显示200bp周期以下清楚的峰。这类峰无法在人鸟枪数据中观察到。这些分析与cfDNA中片段边界的分布上的核小体定位的主要影响一致。

[0152] 还观察了样品之间精确峰频率中的差异。这可能是各细胞类型中连接子序列长度的不同分布的结果。峰来源于结合核小体的DNA加连接子序列的模式得到以下观察的支持:峰周围的侧翼不是对称的且高于峰的频率的强度比低于峰的频率要低。这暗示类似于图3所示那些的图表可用于进行cfDNA和MNase测序数据的质量控制。使用常规(鸟枪)DNA的cfDNA和MNase的随机片段化或污染将导致周期图中这些特征性强度模式的稀释或极端情况下的完全消失。

[0153] 在下文中,基于测量的196bp周期处强度以及针对181bp至202bp的频率范围测得的所有强度分析数据。选择较宽的频率范围以提供较高的分辨率,因为正在捕获的是较宽范围的连接子长度。选择这些强度作为单纯针对此处计算理由的中心;在相关实施方式中可使用不同的频率范围。图4和5探索在连续、非重叠的10kbp区块中的196bp处周期图强度的视图,这些区块瓷砖式覆盖人常染色体的全长(详细信息参见方法)。图4显示在前三种成分中的数据和预测的主成分分析(PCA)。主成分1(PC1)(28.1%方差(variance))捕获图3所示强度的差异并从而从基因组鸟枪数据中区分MNase和cfDNA样品。相反地,PC2(9.7%方差)捕获MNase和cfDNA样品之间的差异。PC3(6.4%方差)捕获单个样品之间的差异。图5显示基于强度向量的欧几里得距离的该数据的递阶聚类系统树图。我们注意到,两项HeLa S3试验在PCA和系统树图上紧密聚类,虽然数据在不同的实验室生成且遵循不同的实验方案。还聚类了“正常”cfDNA样品、肿瘤cfDNA样品和细胞系MNase样品的组。具体而言,源自相同肿瘤类型(多形性成胶质细胞瘤)的三种肿瘤样品似乎单独地从源自不同肿瘤类型的肿瘤.2351样品聚类(参见表1)。GM1和IM1样品单独地从获自妊娠妇女的其他cfDNA样品聚类。这符合这些样品中峰以下频率所观察到的较高强度(即图3中更显著的左肩)。这可能显示两

种样品集合之间cfDNA制备的细微差异,或未设置对照的生物差异(如妊娠年龄)。

[0154] 图6和7显示等价分析的结果,但其基于181bp至202bp的频率范围。比较这些图,这些结果对于较宽的频率范围而言很大程度上是稳定的;然而额外的频率可在更精确度量的分析中改善灵敏度。为进一步探索细胞类型来源特异性模式,使用该频率范围的强度的PCA单独分析cfDNA和MNase数据集。在以下分析集合中,排除了五份来自妊娠妇女的cfDNA样品,其显示图3中显著的左肩。图8显示cfDNA数据的前7种主成分且图9显示六种MNase数据集的全部六种主成分。虽然存在相关样品的聚类,但也存在相当大的变化(生物和技术变化)以区分各样品和剩余样品。例如,测序深度的影响是可观察的,这可以从本体.cfDNA和本体.cfDNA_部分以及MC2.cfDNA和MC2.cfDNA_部分的分离中看到。可使用读数抽样(read sampling)来校正该技术混淆(technical confounder)。

[0155] 该实施例的一些关键观察包括:

[0156] 1)cfDNA测序数据中的读数起始坐标捕获强核小体定位信号。

[0157] 2)跨越基因组的子集(如连续10kbp窗口)汇集的核小体定位信号的差异与样品来源相关。

[0158] 实施例2核小体图谱的混合比例预测

[0159] 在实施例1中,研究了从公共数据库生成或下载的样品的基本聚类。分析显示这些数据集中的读取起始坐标捕获强核小体定位信号(跨越获自2000万条序列至超过10亿条序列的测序深度的范围)且样品来源与该信号相关。针对该方法的目的,能够鉴定已知细胞类型的混合物以及在某种程度上定量来自该信号的各细胞类型的贡献也是有用的。出于此目的,该实施例探索了两种样品的合成的混合物(即基于序列读数)。我们对于两个MNase数据集(MCF.7和NA12878.MNase)和两个cfDNA数据集(肿瘤.2349和主体.cfDNA)以5:95、10:90、15:85、20:80、30:70、40:60、50:50、60:40、30:70、80:20、90:10和95:5的比率混合测序读数。从1亿9690万比对的读数的两个集合(其各自来自初始样品之一)中提取合成的MNase混合物数据集并从1亿8110万比对的读数的两个集合(其各自来自初始样品之一)中提取合成的cfDNA混合物数据集。

[0160] 图10显示第11号染色体的平均强度,相当于图3但针对这些合成的混合物。可从图10中看到不同的样品贡献如何导致全局性频率强度模式的偏移。可利用该信号推导合成的混合物比例。图11显示MNase数据集混合物的前两个主成分且图12显示cfDNA数据集混合物的前两个主成分。在两种情况下,第一个PC均直接捕获混合的数据集的组成。因此可直接理解在给定适当的参考集合并使用例如回归模型时,如何能够从频率强度数据的变换中预测两种或可能的更多种细胞类型的混合比例。图13显示全部两个数据集的系统树图,确认来源于类似样品比例的混合物样品的总体相似性以及cfDNA和MNase样品的区分。

[0161] 该实施例的关键观察之一是多种样品类型(cfDNA或细胞/组织类型)与未知样品的混合比例可通过核小体占位模式的建模来预测。

[0162] 实施例3:使用cfDNA测序数据相对于转录因子结合位点测量核小体占位虽然前述实施例显示可通过将基因组分割为连续不重叠的10kbp窗口来获得核小体定位的信号,也可使用正交方法生成切割可及性图谱且正交方法较不倾向于产生基于窗口尺寸和边界的人工现象。该实施例中详细探索的一种这类方法是核小体定位的推导,其通过观察到的转录因子(TF)结合位点周围读数起始的周期性实施。

[0163] 可以充分确定的是,局部核小体定位受周边TF占位的影响。对于染色质局部重塑的影响和因此对于周边核小体的稳定定位的影响在TF集合中都是不均匀的;给定TF的占位可对于核小体定位产生局部影响,其优先是结合位点的5'或3'且在特定细胞类型中延伸较大或较小的基因组距离。此外,且对于本发明的目的而言重要的是,具体细胞中体内占据的TF结合位点集合在组织和细胞类型之间变化,使得如果能够鉴定感兴趣的组织或细胞类型的TF结合位点占位图谱并针对一种或多种TF重复该过程,则可以通过鉴定一种或多种细胞类型或组织特异性TF结合位点占位概况的富集或消耗来鉴定对cfDNA有贡献群体的细胞类型和组织的混合物的成分。

[0164] 为证实该想法,使用TF结合位点周围的读数起始来从视觉上确认反映优先的局部核小体定位的切割偏好。ChIP-seq转录因子(TF)峰获自DNA元件百科全书("ENCODE")项目(国家人基因组研究所,国家健康研究所,马里兰州贝塞斯达)。因为这些峰的基因组区间较宽(平均200至400bp),所以通过使用保守p值截取($1x10^{-5}$,详细信息参见方法)信息化扫描相应结合基序的基因组来辨别这些间隔内的活跃结合位点。随后将这两种独立来源的预测的TF结合位点的集合的交集用于下游分析。

[0165] 在具有至少1亿条序列的样品中计算各候选TF结合位点的500bp内各位置处的读数起始数目。在各样品内。在各位置处加总所有的读数起始,产生总共每个TF每个样品1014至1019个位置,具体数值取决于TF识别序列的长度。

[0166] 图14显示多个不同样品中人基因组中24666个CTCF结合位点周围读数起始的分布,其中心在结合位点本身周围。CTCF是一种绝缘子结合蛋白质且在转录抑制中起重要作用。先前的研究暗示,CTCF结合位点锚定局部核小体定位,使得至少20个核小体在给定结合位点周围对称且被有规律地隔开,其大致的周期为185bp。对于图14中几乎所有样品都常见的一种惊人特征是结合位点上游和下游核小体定位的清晰周期,暗示在不同的cfDNA和MNase消化的样品中重现了体内CTCF结合的局部且基本对称作用。有趣的是,上游和下游峰的周期在所有样品集合中都是不均匀的;MNase消化的样品显示相对于结合位点的略宽的峰间隔,暗示不仅应用了峰强度,还应用了其周期。

[0167] 图15显示5644个c-Jun结合位点周围读数起始的分布。虽然对于该图中的若干样品而言可再次通过视觉鉴定到熟悉的周期,但该作用是不均匀的。应注意,MNase消化的样品中有三种(Hap1.MNase、HEK.MNase和NA12878.MNase)具有平缓得多的分布,其可表明这些细胞中c-Jun结合位点并非高度占位,或c-Jun结合对于局部染色质重塑的作用在这些细胞类型中较不显著。无论潜在的机制如何,读数起始的局部周边中的偏好在TF至TF和样品类型之间变化的观察重申了cfDNA样品中基于读数起始的核小体占位推导对于关联或去卷积组织来源组成的潜在作用。

[0168] 图16显示4417个NF-YB结合位点周围读数起始的分布。这些TF结合位点周边的读数起始分布显示偏离对称(a departure from symmetry):此时,下游作用(各图内右侧)似乎强于上游作用,证据是cfDNA样品中轻微向上的轨迹线。还应注意MNase消化的样品和cfDNA样品之间的区别:前者显示平均而言较平缓的概况,其中难以辨别峰;而后者同时具有能够更清晰辨别的周期和更能够鉴定的峰。

[0169] 实施例1-3的方法

[0170] 临床和对照样品

[0171] 全血在晚期妊娠产前护理期间提取自妊娠妇女fgs002、fgs003、fgs004和fgs005并迅速储存在含有EDTA的Vacutainer管(BD公司)中。来自妊娠妇女IM1、GM1和GM2的全血分别获自第18、13和10周妊娠期并迅速储存在含有EDTA的Vacutainer管(BD公司)中。收集来自神经胶质瘤患者2349、2350、2351和2353的全血作为大脑外科手术过程的一部分并在含有EDTA的Vacutainer管(BD公司)中储存小于三小时。来自健康成年男性的男性对照2(MC2)的全血收集在含有EDTA的Vacutainer管(BD公司)中。各个体可以使用4-10ml血液。通过4℃下1000x g离心10分钟从全血中分离血浆,之后收集上清液并再次4℃下2000x g离心15分钟。纯化的血浆在-80℃下以1ml等分试样储存直至使用。

[0172] 含有来自未知数目的健康个体的贡献物的主体人血浆获自干细胞技术公司(STEMCELL Technologies)(加拿大不列颠哥伦比亚省温哥华)并在-80℃下以2ml等分试样储存直至使用。

[0173] 血浆样品的处理

[0174] 使用前一刻在桌面上解冻冷冻的血浆等分试样。根据生产商的方案使用QiaAMP循环核酸试剂盒(凯杰公司(Qiagen),荷兰芬洛)从2ml的各血浆样品中纯化循环的cfDNA。使用Qubit荧光计(英杰公司(Invitrogen),加利福尼亚州卡尔斯巴德)和靶向人Alu序列的定制qPCR测定来定量DNA。

[0175] MNase消化

[0176] 使用标准方法生长各株系(GM12878、HeLa S3、HEK、Hap1)的约5000万个细胞。对生长介质通气并使用PBS清洗细胞。使用2x体积的CSS介质对细胞胰蛋白酶化和中和,随后通过4℃下1300rpm离心5分钟将其沉淀在锥形管中。使用添加有1x蛋白酶抑制剂混合物的12ml冰冷PBS重悬细胞沉淀物,计数并随后通过4℃下1300rpm离心5分钟沉淀。将细胞沉淀物重悬于RSB缓冲液(10mM Tris-HCl、10mM NaCl、3mM $MgCl_2$、0.5mM亚精胺、0.02% NP-40、1X蛋白酶抑制剂混合物)中至每毫升300万个细胞的浓度并在冰上轻柔颠倒下孵育10分钟。通过4℃下1300rpm离心5分钟沉淀细胞核。将沉淀的细胞核重悬在NSB缓冲液(25%甘油、5mM $MgAc_2$、5mM HEPES、0.08mM EDTA、0.5mM亚精胺、1mM DTT、1X蛋白酶抑制剂混合物)中至每毫升15M的终浓度。再次通过4℃下1300rpm离心5分钟沉淀细胞核并重悬在MN缓冲液(500mM Tris-HCl、10mM NaCl、3mM $MgCl_2$、1mM CaCl、1X蛋白酶抑制剂混合物)中至每毫升30M的终浓度。将细胞核分成200μl等分试样并在37℃下使用4U的微球菌核酸酶(沃新顿生物化学公司(Worthington Biochemical),美国新泽西州湖林市)消化五分钟。添加85μl的MNSTOP缓冲液(500mM NaCl、50mM EDTA、0.07% NP-40、1X蛋白酶抑制剂混合物)在冰上终止反应,随后在4℃下轻柔颠倒下孵育90分钟。使用苯酚:氯仿:异戊醇提取纯化DNA。使用标准方法用2%琼脂糖凝胶电泳对单核小体片段进行尺寸选择并使用Nanodrop光谱仪(赛默飞世尔科学公司(Thermo Fisher Scientific),美国马塞诸塞州沃尔瑟姆)定量。

[0177] 测序文库的制备

[0178] 使用ThruPLEX-FD或ThruPLEX DNA-seq 48D试剂盒(鲁比孔基因组公司(Rubicon Genomics),密歇根州安娜堡)制备条形码化的测序文库,包括专有的一系列末端修复、连接和扩增反应。使用3.0至10.0ng的DNA作为所有临床样品文库的输入物。对各文库使用30ng的输入物构建两个主体血浆cfDNA文库;各文库单独条形码化。对各文库的2ng的输入物构建来自MC2的两个文库;各文库单独条形码化。使用20ng的尺寸选择的输入DNA构建各MNase

消化的细胞系的文库。通过实时PCR监测所有样品的文库扩增以避免过度扩增。

[0179] 测序

[0180] 所有文库都在HiSeq 2000(伊露米娜公司(Illumina),美国加利福尼亚州圣地亚哥)仪器上测序,其中使用具有9bp索引读数(index read)的双端101bp读数。一条测序流通槽(lane)针对混合样品fgs002、fgs003、fgs004和fgs005进行,生成每份样品总共约$4.5x10^7$个读数对。在多个流通槽中都对样品IM1、GM1和GM2进行测序以分别生成个$1.2x10^9$、$8.4x10^8$和$7.6x10^7$个读数对。一条测序流通槽针对样品2349、2350、2351和2353中的每一个进行,生成每份样品约$2.0x10^8$个读数对。一条测序流通槽针对四种细胞系MNase消化的文库中的每一个进行,生成每个文库约$2.0x10^8$个读数对。四条测序流通槽针对两个重复的MC2文库之一进行且三条流通槽针对两个重复的主体血浆文库之一进行,分别产生每个文库共$10.6x10^9$和$7.8x10^8$个读数对。

[0181] cfDNA测序数据的处理

[0182] cfDNA和MNase文库的DNA插入尺寸倾向于较短(大部分数据为80bp至240bp);因此预测一些分子的阅读末端处的接头序列。修剪阅读末端处起始的接头序列,并将短原始分子的正向和反向双端("PE")读数数据合并为单个读数("SR");与至少11bp读数重叠的PE读数被合并为SR。弃去短于30bp或显示超过5个碱基的质量评分低于10的SR。使用快速比对工具(BWA-ALN或BWA-MEM)将剩余的PE和SR数据与人参考基因组(GRCh37,1000G发布v2)比对。使用SAMtools将所得SAM(序列比对/图谱)格式转化为分选的BAM(二元序列比对/图谱格式)。

[0183] 额外的公众可得数据

[0184] 下载公众可得的Hela-S3 MNase(登录号SRR633612、SRR633613)和MCF-7MNase实验(登录号SRR999659-SRR999662)的数据并如前文所述进行处理。

[0185] 公众可得的由伊露米娜剑桥有限公司(Illumina Cambridge)(英国艾塞克斯)生成的CEPH系谱(pedigree)146个单个NA12878的基因组鸟枪测序数据获自欧洲核苷酸档案库(European Nucleotide Archive)(ENA,登录号ERR174324-ERR174329)。该数据是Illumina HiSeq上使用2x101bp读数测序的PE且在测序前针对较长的插入尺寸选择文库。因此,未预测读数端处的接头序列;因此使用BWA-MEM对该数据进行直接比对。

[0186] 提取读数端信息

[0187] PE数据提供涉及测序文库制备中使用的DNA分子的两个物理端的信息。该信息使用SAMtools应用程序接口(API)提取自BAM文件。使用全部两种PE数据的比对坐标,其中两端均与相同染色体比对且使用具有相反取向的读数。对于未修剪的SR数据,仅一个读数端提供涉及原始DNA分子的物理端的信息。如果读数与参考基因组的正链(plus strand)比对,则使用最左侧的坐标。如果读数与反链(reverse strand)比对,则使用其最右侧的坐标代替。在通过接头修剪将PE数据转化为单读数数据的情况下,考虑两端坐标。如果SR测序实验修剪了至少五个接头碱基,也考虑两端坐标。

[0188] 对于人参考序列中的所有常染色体(第1-22号染色体),在10000碱基的窗口(区块)中提取所有位置处的覆盖和读数端数目。如果区块中没有比对读数,该区块被认为对于该特定样品而言是空的。

[0189] 平滑化周期图

[0190] 针对各样品的各非空区块计算覆盖和读数起始的比率。如果覆盖是0,则将比率设为0。这些比率被用于使用快速傅里叶变换(FFT,R统计编程环境中的spec.pgram)计算各区块的周期图,其中频率为1/500碱基至1/100碱基。任选地,使用对数据进行平滑化(3bp Daniell平滑器(smoother);移动平均给予端值一半权重)和去趋势化(如扣除系列的平均值并除去线性趋势)的参数。保存各区块的频率范围120-250bp的强度。

[0191] 平均染色体强度

[0192] 对于样品集合,鉴定在所有样品中的非空区块。在各常染色体中的各样品的所有区块对特定频率的强度进行平均化。

[0193] 主成分分析和系统树图

[0194] 收集在所有样品中的非空区块。使用主成分分析(PCA;R统计编程环境中的prcomp)来降低数据的维度并在二维空间中对其进行绘制。PCA鉴定到捕获数据的大部分变异的维度并构建正交维度,解释了数据中降低的变异量。

[0195] 计算样品强度之间的成对欧几里得距离并将其可视化为系统树图(R统计编程环境中的统计库(stats library))。

[0196] 转录因子结合位点预测

[0197] 通过对在多种细胞类型中生成的ChIP-seq数据进行分析而获得的推定的转录因子结合位点获自ENCODE项目。

[0198] 通过使用来自MEME软件包(版本4.10.0_1)的程序fimo扫描人参考基因组(GRCh37,1000G发布v2)来获得候选转录因子结合位点的独立集合。使用获自JASPAR_CORE_2014_vertebrates数据库的位置权重矩阵(positional weight matrices)来进行扫描,其中使用选项“--verbosity 1--thresh 1e-5”。使用的转录因子基序标识符是MA0139.1、MA0502.1和MA0489.1。

[0199] 使用bedtools v2.17.0对来自全部两个预测位点集合的染色质坐标进行交集。为保存作图中的任何对称性,仅使用“+”链上的预测结合位点。对各样品记录读数起始,前体是其落入预测结合位点的各端的500bp内,并通过在所有这类位点的位置中在样品内加总读数起始。仅将具有至少1亿总读数的样品用于该分析。

[0200] 实施例4:从cfDNA中确定正常/健康的来源组织

[0201] 为评价单个个体的cfDNA中观察到的片段化模式是否含有导致这些片段的产生的细胞的基因组组织的证据以及因此得到的cfDNA分子群体的来源组织的占位(即使在有贡献的细胞类型之间没有基因型差异的情况下),对cfDNA进行深度测序以更好地理解导致其产生的过程。所得数据用于建立基于他人先前工作的基因组范围的核小体占位图谱,但其全面得多。通过优化文库制备方案以恢复短片段,发现通过cfDNA也直接生成转录因子(TF)(例如CTCF)的体内占位的足迹。最后,发现健康个体中cfDNA测序所揭示的常规元件和基因体中的核小体占位与淋巴样和髓样细胞系中的基因表达和DNA酶超敏最强烈地相关。

[0202] cfDNA片段对应于染色质小体且含有大量DNA损伤

[0203] 通过对纯化自血浆的cfDNA片段进行末端修复和接头连接来制备常规测序文库,该血浆汇集自未知数目的健康个体(“BH01”)或来自单个个体(“IH01”)(图17;表1):

[0204] 表1.血浆样品的测序统计。

[0205]

样品名称	文库类型	读数	测序的片段	比对的	比对的Q30	覆盖	预测%重复	35-80bp	120-180bp
BH01	DSP	2x101	1489569204	97.20%	88.85%	96.32	6.00%	0.65%	57.64%
IH01	DSP	2x101	1572050374	98.58%	90.60%	104.92	21.00%	0.77%	47.83%
IH02	SSP	2x50, 43/42	779794090	93.19%	75.27%	30.08	20.05%	21.83%	44.00%
CH01	--	--	3841413668	96.95%	86.81%	231.32	14.99%	5.00%	50.85%

[0206] SSP,单链文库制备方案。DSP,双链文库制备方案。

[0207] 对于各样品,将测序相关统计数据制成表格:该测序相关统计数据包括测序的片段总数、读数长度、比对至具有或不具有映射质量阈值的参考物的这类序列的百分比、平均覆盖、重复率,和两个长度箱中测序的片段的比例。片段长度推导自双端读数的比对。由于读数长度较短,通过假定已读取完整片段来计算覆盖。预测的重复片段数目基于片段末端,其在高度固型(stereotyped)覆盖存在的情况下可能过度预测了真正的重复率。SSP,单链文库制备方案。DSP,双链文库制备方案。

[0208] 文库BH01和IH01分别测序至96和105倍覆盖(1.5G和1.6G片段)。由双端读数的比对所推导的片段长度分布具有约167bp处的主峰(符合染色质小体相关的DNA长度),和100-160bp长度范围中约10.4bp周期(图18)。这些分布符合下述模型:该模型中通过与蛋白质相关联而在细胞死亡前和后优先保护cfDNA片段免于核酸酶介切割(在这种情况下,通过核小体核心颗粒和连接子组蛋白),但其中发生一定程度的额外缺刻或切割,其与核小体连接的DNA的螺距相关。进一步支持该模型的是这些167bp片段的二核苷酸组成,其扼要说明了MNase来源、核小体相关片段的较早期研究的关键特征(例如针对二分体(dyad)处A/T二核苷酸的偏好)并支持以下概念:核小体核心颗粒相对于染色质小体对称定位(图19)。

[0209] 该cfDNA本体模型的预测是广泛的DNA损伤,例如单链缺刻以及5'和3'突出端。在常规的文库制备期间,不扩增缺刻的链,通过末端修复钝化突出端,且可代表总cfDNA中大部分的短双链DNA("dsDNA")分子可被简单地不充分恢复。为解决该问题,使用改编自Gansauge等的古老DNA研究的方案制备来自来源于额外健康个体('IH02')的血浆传播的cfDNA的单链测序文库,这些研究中已报道核小体周围广泛的DNA损伤和核酸酶切割。简言之,对cfDNA进行变性并将生物素偶联的单链接头连接至所得片段。随后对连接的片段进行第二链合成、末端修复和第二接头连接,同时将片段固定至链霉亲和素珠。最后,进行最少的PCT扩增以富集携带接头的分子,同时附加样品索引(index)(图20;表2)。

[0210] 表2.用于单链测序文库制备的合成的寡聚物。

[0211]

寡聚物名称	序列 (5'-3')	备注
CL9	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	HPLC纯化
接头2.1	CGACGCTCTTCCGATC/ddT/	HPLC纯化
接头2.2	/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAG*T*G*T*A	HPLC纯化
CL78	/5Phos/AGATCGGAAG/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/iSpC3/3BioTEG/	双重 HPLC纯化

[0212] 对于IH02,将所得文库测序至30倍覆盖(779M片段)。片段长度分布再次显示对应于染色质小体的约167bp处的主峰,但其相对于常规文库制备物显著地富集较短片段(图21、22、23A-B、24A-B)。虽然所有文库都显示约10.4bp周期,针对两种方法对片段尺寸偏移

3bp,这与损伤的或非齐平的输入分子一致,这些输入分子的真正末端更如实地呈现于单链文库中。

[0213] 基于深度cfDNA测序的体内核小体保护的基因组范围图谱

[0214] 为评估是否能够通过与一个或多个参考图谱比较经比对的片段末端的分布或其数学变换来推导对cfDNA有贡献的组织中在所有人基因组中的核小体的主要局部位置,开发了窗口化保护评分("WPS")。具体而言,预期cfDNA片段末端应聚集在核小体边界附近,同时还在核小体本身上所剩无几。为对其进行定量,开发了WPS,其代表以给定基因组坐标为中心的完全横跨120bp窗口的DNA片段数目减去在该同一窗口内具有末端的片段数目(图25)。如预期的那样,WPS值与强烈定位的阵列内核小体的位置关联,如同使用体外方法通过其他群体或古老DNA映射的那样(图26)。在其他位点处,WPS与基因组特征(例如DNA酶I超敏(DHS)位点)相关(例如,与远端调节元件侧翼核小体的重定位一致)(图27)。

[0215] 将启发式算法(heuristic algorithm)应用于BH01、IH02和IH02数据集的基因组范围WPS以分别鉴定12.6M、11.9M和9.7M核小体保护局部极大值(图25-31)。在各样品中,相邻峰之间距离的分布模式是低方差的185bp(图30),通常与人或小鼠细胞中核小体重复长度的先前分析一致。

[0216] 为确定在跨越样品中的峰识别信号的位置是否类似,计算样品中各峰至各其他样品中最近峰的基因组距离。观察到高一致性(图31;图32A-C)。从BH01峰识别信号至最近相邻IH01峰识别信号的中值(绝对)距离是总共23bp,但对于最高度评分的峰而言小于10bp(图33A-B)。

[0217] 因为通过核小体特异性或在文库制备期间导入的偏好可人为地促进核小体保护信号,还模拟了片段末端,匹配各样品的深度、尺寸分布和终端二核苷酸频率。随后计算基因组范围WPS,且通过相同启发式算法识别出10.3M、10.2M和8.0M局部极大值,其分别针对与BH01、IH01和IH02匹配的模拟数据集。与来自真实数据集的峰相比,来自模拟数据集的峰与更低的评分相关(图33A-B)。此外,从真实数据集识别的相对可重复的峰位置(图31;图32A-C)与从模拟数据集识别的峰位置(图31;图34A-C)不能很好地契合。

[0218] 为改善基因组范围核小体图谱的精确性和完整性,针对合并的231倍覆盖汇集并重新分析来自BH01、IH01和IH02的cfDNA测序数据('CH01';3.8B片段;表1)。针对该合并的样品计算WPS并识别12.9M峰。该峰识别信号集合与较高的评分相关且接近了峰数目方面的饱和(图33A-B)。考虑小于500bp的所有峰至峰距离(图35),CH01峰集合横跨人参考基因组的2.53十亿碱基(Gb)。

[0219] 已知核小体相对于基因调节的地标良好地定位,这些地标是例如转录起始位点和外显子-内含子边界。与该理解一致的是,也在该数据中观察到了类似的定位,其相对于转录、翻译和剪接的地标(图36-40)。基于核小体间隔与转录活性和染色质标记之间关联的过往观察,基于淋巴母细胞样细胞系中的长程相互作用(原位Hi-C)检测了100千碱基(kb)窗口内的中值峰至峰间隔,这些窗口已被分配为区室A(富集开放的染色质)或区室B(富集关闭的染色质)。与区室B中的核小体相比,区室A中的核小体表现出较紧的间隔(中值187bp(A)对比190bp(B)),同时某些子区室之间具有其他差别(图41)。沿着染色体的长度,没有观察到一般模式,中值染色体间隔在中心体周围区域中急剧下降除外,其受到跨域α卫星的阵列的强定位驱动(171bp单体长度;图42;图26)。

[0220] 短cfDNA片段直接生成CTCF和其他转录因子的足迹

[0221] 先前的DNA酶I切割模式研究鉴定了两种主要片段类别:与核小体之间切割相关的较长片段,和与转录因子结合位点(TFBS)相邻切割相关的较短片段。为评估体内来源的cfDNA片段是否也来自两种类别的核酸酶切割灵敏度,基于推导的片段长度划分序列读数(CHOI),且单独使用长片段(120-180bp;120bp窗口;对核小体识别而言在效果上与上文所述WPS相同)或短片段(35-80bp;16bp窗口)重新计算WPS(图26-27)。为在我们的数据中获得富活跃结合位点的充分限定的TFBS的集合,针对各TF使用来自ENCODE(TfbsClusteredV3)的ChIP-seq峰的一元化(unified)集合与聚类FIMO预测结果求交集。

[0222] 长组分WPS支持CTCF结合位点附近核苷酸的强组织(图43)。然而,也观察到短组分WPS中的强信号,其与CTCF结合位点本身一致(图44-45)。基于CTCF结合位点在体内结合的假定对CTCF结合位点进行分层(所有FIMO预测结果对比与ENCODE ChIP-seq交集的子集对比与似乎在19种细胞系中得到利用的那些交集的另一子集)。实验上得到充分支持的CTCF位点基于长组分WPS表现出侧翼-1和+1核小体之间显著较宽的间隔,这与其在CTCF结合后的重定位一致(约190bp→约260bp;图45-48)。此外,实验上得到充分支持的CTCF位点表现出对于短组分WPS而言比CTCF结合位点本身强得多的信号(图49-52)。

[0223] 针对额外的TF进行类似的分析,对于这些额外的TF而言FIMO预测结果和ENCODE CHiP-seq数据均是可得的(图53A-H)。对于许多这些TF,如ETS和MAFK(图54-55),观察到短组分足迹,伴随长组分WPS中的周期性信号。这与结合的TFBS周围核小体的强定位一致。总体而言,这些数据支持以下观点:通过单链方案明显更好恢复的单链cfDNA片段(图18,图21)直接生成包括CTCF及其他的DNA结合转录因子的体内占位的足迹。

[0224] 核小体间隔模式显示cfDNA来源组织

[0225] 为确定通过cfDNA测序所测量的体内核小体保护是否可用于推导健康个体中对cfDNA有贡献的细胞类型,检测了116种多样化生物样品中限定的DHS位点内核小体识别信号的峰至峰间隔。先前在调节元件处的-1和+1核小体之间观察到加宽的间隔(例如在DHS位点处分节性的(anecdotally)(图27)或在结合的CTCF位点处全局性的(图45))。与结合的CTCF位点类似,观察到DHS位点的子集内的核小体对的显著较宽间隔,似乎对应于核小体重定位的位点,该重定位是通过在导致cfDNA的产生的细胞类型中干扰转录因子结合实现的(约190bp→约260bp;图56)。实际上,根据所使用的细胞类型的DHS位点,加宽的核小体间隔(约260bp)的比例大幅变化。然而,该比例最高的所有细胞类型均为淋巴样或骨髓样来源(例如图56中的CD3_CB-DS17706等)。这与作为健康个体中cfDNA主要来源的造血细胞死亡相一致。

[0226] 随后重新检测转录起始位点周围核小体保护的信号(图36)。在基于淋巴样谱系细胞系NB-4中的基因表达对信号进行分层时,在高表达对比低表达的基因中观察到与TSS相关的核小体保护的位置或强度的显著差异(图57)。此外,短组分WPS表现出TSS直接上游处清晰的足迹,其强度也与表达水平密切相关(图58)。这可能反映了转录活跃基因处转录预起始复合物(transcription preinitiation complex)或其一些成分的足迹。

[0227] 这些数据表明cfDNA片段化模式实际上确实含有可用于推导导致cfDNA的产生的组织或细胞类型的信号。

[0228] 然而,挑战是基因组范围cfDNA文库中相对少的读数与DHS位点和转录起始位点重

叠。

[0229] 核小体间隔在细胞类型之间根据染色质状态和基因表达而变化。通常而言,开放的染色质和转录与较短的核小体重复长度相关,这与该实施例中区室A对比B的分析一致(图41)。该实施例的峰识别信号数据也显示在所有基因体中的核小体间隔与其表达水平之间的关联,较紧的间隔与较高的表达相关(图59;ρ=-0.17;n=19677个基因)。相对于邻近区域,该关联对于基因体本身最高(上游10kbρ=-0.08;下游10kbρ=-0.01)。如果分析限于跨域至少60个核小体识别信号的基因体,较紧的核小体间隔甚至更强烈地与基因表达相关联(ρ=-0.50;n=12344个基因)。

[0230] 利用信号(例如跨域基因体或其他结构域的核小体间隔)的一个优势是更大部分的cfDNA片段将是有信息量的(informative)。另一个潜在的优势是可检测多种对cfDNA有贡献的细胞类型所产生的信号混合物。为对其进行测试,对在所有基因体中的前10kb的长片段WPS并在逐基因(gene-by-gene)的基础上进行进一步数学变换,快速傅里叶变换(FFT)。FFT信号的强度与特定频率范围处的基因表达相关联,其中对于正关联在177-180bp处存在最大值且对于负关联在约199bp处存在最小值(图60)。在进行针对人细胞系和基本组织的76个表达数据集的数据集进行该分析时,最强的关联是与造血谱系(图60)。例如,对于三个健康样品(BH01、IH01、IH02)而言,平均强度在193-199bp频率范围的排名最高的负关联都是淋巴样细胞系、骨髓样细胞系或骨髓组织(图61;表3):

[0231] 表3.WPS FTT强度与基因表达数据集的关联

[0232]

RNA名称	类别	类型	描述	关联								排名差异					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	健康	IC15	IC20	IC17	IC37	IC35
A.431	皮肤	皮肤癌（鳞状细胞）	表皮样癌细胞系	-0.296	-0.188	-0.149	-0.200	-0.140	-0.176	-0.195	-0.178	2	3	-9	-9	-12	-21
A549	肺	肺癌	肺癌细胞系	-0.289	-0.185	-0.144	-0.202	-0.139	-0.172	-0.188	-0.170	3	-14	-12	-9	-2	-13
adipose_tissue	基本组织	脂肪组织	基本组织	-0.270	-0.169	-0.137	-0.169	-0.121	-0.153	-0.166	-0.148	1	12	5	0	14	12
adrenal_gland	基本组织	肾上腺	基本组织	-0.257	-0.158	-0.131	-0.173	-0.118	-0.145	-0.161	-0.138	-2	-11	-5	1	5	8
AN3.CA	乳腺/女性生殖	子宫癌	转移性子宫内膜腺癌细胞系	-0.303	-0.194	-0.157	-0.213	-0.147	-0.183	-0.195	-0.171	-4	-16	-13	-15	-8	-2
appendix	基本组织	阑尾	基本组织	-0.287	-0.185	-0.137	-0.166	-0.118	-0.148	-0.171	-0.152	6	24	20	23	8	9
BEWO	其他	子宫癌	转移性绒毛膜癌细胞系	-0.284	-0.184	-0.147	-0.193	-0.139	-0.173	-0.194	-0.173	-5	3	-12	-15	-19	-27
bone_marrow	基本组织	骨髓	基本组织	-0.343	-0.230	-0.165	-0.192	-0.142	-0.167	-0.193	-0.165	2	40	9	30	16	28
CACO.2	腹部	结肠腺癌	结肠腺癌细胞系	-0.281	-0.177	-0.137	-0.192	-0.128	-0.169	-0.184	-0.164	5	-5	-5	-14	-10	-9
CAPAN.2	腹部	胰腺腺癌	胰腺腺癌细胞系	-0.291	-0.187	-0.145	-0.202	-0.136	-0.176	-0.195	-0.175	3	-12	-2	-18	-19	-25
cerebral_cortex	基本组织	大脑皮层	基本组织	-0.225	-0.136	-0.120	-0.168	-0.108	-0.134	-0.142	-0.125	-1	-9	-3	0	0	0
colon	基本组织	结肠	基本组织	-0.261	-0.162	-0.124	-0.164	-0.111	-0.145	-0.168	-0.148	7	8	8	6	-7	1
Daudi	淋巴样	人伯基特淋巴瘤	人伯基特淋巴瘤细胞系	-0.321	-0.206	-0.153	-0.195	-0.133	-0.165	-0.189	-0.160	4	17	19	19	13	24
duodenum	基本组织	十二指肠	基本组织	-0.261	-0.164	-0.122	-0.159	-0.109	-0.144	-0.166	-0.144	10	10	10	7	-4	7
EFO.21	乳腺/女性生殖	卵巢癌	转移性卵巢浆液囊腺瘤细胞系	-0.287	-0.186	-0.149	-0.201	-0.140	-0.176	-0.188	-0.169	-7	-9	-14	-20	-1	-8
endometrium	基本组织	子宫内膜	基本组织	-0.257	-0.158	-0.132	-0.178	-0.119	-0.151	-0.166	-0.151	-3	-11	-4	-8	-3	-12
esophagus	基本组织	食道	基本组织	-0.237	-0.147	-0.124	-0.156	-0.116	-0.141	-0.158	-0.145	-3	1	-7	0	0	-7
fallopian_tube	基本组织	输卵管	基本组织	-0.247	-0.157	-0.129	-0.171	-0.114	-0.145	-0.161	-0.145	-4	-13	-2	-3	3	-2
gallbladder	基本组织	胆囊	基本组织	-0.249	-0.156	-0.119	-0.153	-0.103	-0.138	-0.154	-0.141	4	4	4	3	4	1
HaCaT	皮肤	角化细胞细胞系	角化细胞细胞系	-0.290	-0.186	-0.149	-0.194	-0.142	-0.173	-0.193	-0.173	-5	7	-18	-8	-8	-17
HDLM.2	淋巴样	霍奇金淋巴瘤	霍奇金淋巴瘤细胞系	-0.316	-0.200	-0.154	-0.201	-0.136	-0.173	-0.195	-0.171	1	6	11	1	-5	-5
heart_muscle	基本组织	心肌	基本组织	-0.246	-0.149	-0.126	-0.166	-0.113	-0.141	-0.155	-0.140	-3	-3	-3	0	3	2
HEK_293	其他	肾上腺前体细胞系	胚胎肾细胞系，由5型腺病毒转化	-0.292	-0.187	-0.150	-0.209	-0.139	-0.172	-0.189	-0.168	-4	-17	-4	-2	3	0
HEL	骨髓样	红白血病	红白血病细胞系（针对霍奇金病疾治疗反复发中的AML M6）	-0.324	-0.205	-0.161	-0.210	-0.140	-0.172	-0.194	-0.168	-1	-5	4	12	5	14
HeLa	乳腺/女性生殖	子宫颈癌	子宫颈上皮腺癌细胞系	-0.296	-0.186	-0.149	-0.203	-0.139	-0.172	-0.190	-0.171	1	-10	-5	-3	1	-8
Hep_G2	腹部	肝细胞癌	肝细胞癌细胞系	-0.294	-0.186	-0.152	-0.202	-0.145	-0.186	-0.196	-0.167	-4	-6	-18	-24	-17	2
HL.60	骨髓样	前髓细胞白血病	急性前髓细胞白血病(APL)细胞系	-0.332	-0.208	-0.161	-0.202	-0.137	-0.171	-0.197	-0.170	2	6	18	18	-1	11
HMC.1	骨髓样	肥大细胞白血病	肥大细胞白血病细胞系	-0.337	-0.228	-0.165	-0.212	-0.149	-0.181	-0.199	-0.180	0	-1	-2	3	0	-2
K.562	淋巴样	白血病	慢性骨髓样白血病(CML)细胞系	-0.317	-0.202	-0.158	-0.211	-0.143	-0.178	-0.195	-0.166	-3	-9	-5	-6	-1	13
Karpas.707	淋巴样	多发性骨髓瘤	多发性骨髓瘤细胞系	-0.325	-0.210	-0.155	-0.195	-0.136	-0.167	-0.188	-0.164	4	20	18	22	21	22
kidney	基本组织	肾	基本组织	-0.245	-0.150	-0.130	-0.168	-0.119	-0.153	-0.171	-0.147	-7	-4	-12	-19	-21	-6
liver	基本组织	肝	基本组织	-0.248	-0.148	-0.122	-0.150	-0.110	-0.150	-0.164	-0.138	1	4	-1	-13	-4	3
lung	基本组织	肺	基本组织	-0.264	-0.170	-0.133	-0.170	-0.121	-0.148	-0.167	-0.149	3	4	0	7	3	6

[0233]

RNA名称	类别	类型	描述	关联								排名差异					
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35	健康	IC15	IC20	IC17	IC37	IC35
lymph_node	基本组织	淋巴结	基本组织	-0.308	-0.195	-0.148	-0.162	-0.128	-0.155	-0.181	-0.156	7	24	17	25	14	22
MCF7	乳腺/女性生殖	乳腺癌	转移性乳腺癌细胞系	-0.298	-0.195	-0.154	-0.207	-0.145	-0.183	-0.196	-0.181	-3	-9	-12	-18	-11	-19
MOLT.4	淋巴样	白血病(ALL)	急性淋巴母细胞性白血病(T-ALL)细胞系	-0.323	-0.204	-0.163	-0.212	-0.144	-0.177	-0.197	-0.173	-3	-7	-2	-1	-5	-1
NB.4	骨髓样	前髓细胞性白血病	急性前髓细胞性白血病(APL)细胞系	-0.346	-0.228	-0.172	-0.211	-0.148	-0.182	-0.202	-0.171	0	4	3	5	2	13
NTERA.2	泌尿/男性生殖	泌尿系统癌	转移性胚胎癌细胞系，克隆自TERA-2	-0.269	-0.170	-0.137	-0.193	-0.117	-0.157	-0.169	-0.153	-2	-8	16	-2	5	0
ovary	基本组织	卵巢	基本组织	-0.266	-0.162	-0.135	-0.181	-0.120	-0.152	-0.166	-0.151	1	-7	2	-2	6	-1
pancreas	基本组织	胰腺	基本组织	-0.250	-0.159	-0.132	-0.170	-0.116	-0.150	-0.166	-0.150	-5	-5	1	-6	-5	-7
PC.3	泌尿/男性生殖	前列腺癌	转移性未充分分化的前列腺腺癌细胞系	-0.295	-0.190	-0.151	-0.204	-0.136	-0.174	-0.188	-0.173	-3	-10	2	-6	8	-12
placenta	基本组织	胎盘	基本组织	-0.266	-0.166	-0.134	-0.168	-0.126	-0.151	-0.166	-0.150	3	10	-7	1	9	6
prostate	基本组织	前列腺	基本组织	-0.248	-0.161	-0.133	-0.175	-0.123	-0.150	-0.165	-0.151	-8	-10	-11	-8	1	-12
rectum	基本组织	直肠	基本组织	-0.255	-0.154	-0.117	-0.159	-0.102	-0.136	-0.161	-0.142	6	0	5	4	-2	0
REH	淋巴样	白血病(ALL)	前B细胞白血病普通型 (ALL，首次复发	-0.330	-0.216	-0.165	-0.214	-0.150	-0.182	-0.204	-0.174	-2	-5	-5	-2	-4	1
RH.30	肉瘤	横纹肌肉瘤	转移性横纹肌肉瘤细胞系	-0.280	-0.165	-0.137	-0.194	-0.125	-0.158	-0.175	-0.158	2	-14	-3	-7	-7	-7
RPMI.8226	淋巴样	多发性骨髓瘤	多发性骨髓瘤细胞系	-0.322	-0.207	-0.155	-0.198	-0.138	-0.169	-0.190	-0.164	1	16	11	19	14	22
RT4	泌尿/男性生殖	膀胱癌	泌尿系统膀胱移行细胞癌细胞系	-0.282	-0.168	-0.145	-0.192	-0.136	-0.170	-0.191	-0.171	-5	-1	-12	-16	-19	-25
salivary_gland	基本组织	唾液腺	基本组织	-0.262	-0.166	-0.138	-0.177	-0.128	-0.154	-0.172	-0.155	-7	2	-9	-2	-5	-5
SCLC.21H	肺	小细胞肺癌	小细胞肺癌细胞系	-0.259	-0.160	-0.138	-0.201	-0.123	-0.157	-0.172	-0.146	-11	-31	-5	-12	-10	8
SH.SY5Y	大脑	神经母细胞瘤	转移性神经母细胞瘤，神经上皮瘤细胞系SK-N-SH的克隆亚系	-0.271	-0.170	-0.137	-0.201	-0.124	-0.157	-0.170	-0.151	2	-25	2	-5	1	6
SiHa	乳腺/女性生殖	子宫颈癌	子宫颈鳞状细胞癌细胞系，整合的HPV16的1-2份拷贝	-0.286	-0.180	-0.148	-0.201	-0.139	-0.176	-0.193	-0.175	-2	-7	-15	-19	-11	-27
SK.BR.3	乳腺/女性生殖	乳腺癌	转移性乳腺癌细胞系	-0.288	-0.176	-0.148	-0.195	-0.140	-0.176	-0.191	-0.169	-3	-4	-21	-22	-12	-11
SK.MEL.30	基本组织	黑素瘤	转移性恶性黑素瘤细胞系	-0.301	-0.187	-0.154	-0.208	-0.141	-0.174	-0.193	-0.171	-2	-12	-6	-3	-1	-6
skeletal_muscle	基本组织	骨骼肌	基本组织	-0.261	-0.166	-0.134	-0.179	-0.125	-0.150	-0.164	-0.145	-1	-7	-7	0	9	11
skin	皮肤	皮肤	基本组织	-0.259	-0.166	-0.134	-0.168	-0.127	-0.148	-0.167	-0.151	-4	8	-14	5	-1	-4
small_intestine	基本组织	小肠	基本组织	-0.260	-0.164	-0.121	-0.158	-0.107	-0.141	-0.166	-0.142	9	10	11	9	0	7
smooth_muscle	基本组织	平滑肌	基本组织	-0.259	-0.158	-0.127	-0.169	-0.113	-0.144	-0.161	-0.149	2	-6	3	4	4	-5
spleen	基本组织	脾	基本组织	-0.308	-0.202	-0.148	-0.180	-0.130	-0.155	-0.177	-0.154	7	27	15	25	20	25
stomach	基本组织	胃	基本组织	-0.264	-0.170	-0.131	-0.170	-0.117	-0.149	-0.169	-0.151	6	3	9	6	0	2
testis	基本组织	睾丸	基本组织	-0.215	-0.142	-0.109	-0.147	-0.093	-0.126	-0.133	-0.123	0	0	0	0	0	0
THP.1	骨髓样	单核细胞性白血病	急性单核细胞性白血病(AML)细胞系	-0.338	-0.218	-0.168	-0.206	-0.149	-0.182	-0.204	-0.176	-1	8	-1	1	-3	0

[0234]

RNA名称	类别	类型	描述	关联								健康	排名差异				
				BH01	IH01	IH02	IC15	IC20	IC17	IC37	IC35		IC15	IC20	IC17	IC37	IC35
thyroid_gland	基本组织	甲状腺	基本组织	-0.261	-0.158	-0.136	-0.178	-0.121	-0.153	-0.170	-0.161	-2	-7	-2	-6	-6	-19
TIME	其他	微血管内皮细胞系	嗜粒酸永生化人微血管内皮细胞系（混合型）	-0.296	-0.180	-0.147	-0.198	-0.134	-0.170	-0.186	-0.170	5	-3	3	-1	3	-11
tonsil	基本组织	扁桃体	基本组织	-0.282	-0.179	-0.141	-0.169	-0.125	-0.147	-0.173	-0.152	-1	20	8	23	4	9
U.138_MG	大脑	胶质母细胞瘤	胶质母细胞瘤细胞系	-0.288	-0.177	-0.144	-0.191	-0.126	-0.162	-0.177	-0.161	1	8	7	0	2	2
U.2_OS	肉瘤	骨肉瘤	骨肉瘤细胞系	-0.275	-0.175	-0.139	-0.192	-0.134	-0.159	-0.170	-0.160	-2	0	-11	-3	6	-3
U.2197	肉瘤	肉瘤	恶性纤维状组织细胞瘤细胞系	-0.290	-0.181	-0.146	-0.195	-0.129	-0.164	-0.180	-0.165	2	1	5	3	4	0
U.251_MG	大脑	胶质母细胞瘤	胶质母细胞瘤细胞系	-0.292	-0.178	-0.140	-0.197	-0.125	-0.160	-0.177	-0.165	9	-6	11	4	4	-4
U.266.70	淋巴样	多发性骨髓瘤	多发性骨髓瘤细胞系（1970，IL-6依赖性）	-0.320	-0.207	-0.157	-0.202	-0.135	-0.170	-0.191	-0.165	-1	4	19	15	12	17
U.266.84	淋巴样	多发性骨髓瘤	多发性骨髓瘤细胞系（1984，体外分化）	-0.326	-0.212	-0.162	-0.207	-0.139	-0.175	-0.194	-0.169	-1	2	11	8	10	14
U.698	淋巴样	B细胞淋巴瘤	B细胞淋巴瘤细胞系（淋巴母细胞性淋巴肉瘤）	-0.328	-0.212	-0.159	-0.203	-0.137	-0.170	-0.194	-0.166	2	5	18	20	6	20
U.87_MG	大脑	胶质母细胞瘤，星形细胞瘤	胶质母细胞瘤，星形细胞瘤细胞系	-0.285	-0.175	-0.143	-0.192	-0.127	-0.160	-0.174	-0.162	1	0	2	-2	2	-4
U.937	骨髓样	健单核细胞性组织细胞性淋巴瘤	健单核细胞性组织细胞性淋巴瘤细胞系	-0.346	-0.224	-0.167	-0.201	-0.146	-0.180	-0.199	-0.173	1	18	3	5	2	6
urinary_bladder	基本组织	泌尿膀胱	基本组织	-0.260	-0.158	-0.130	-0.165	-0.118	-0.146	-0.164	-0.150	3	5	-2	1	3	-6
WM.115	皮肤	黑素瘤	恶性黑素瘤细胞系	-0.284	-0.175	-0.144	-0.193	-0.130	-0.160	-0.178	-0.157	-1	-4	-4	-3	-3	2

[0235] 通过人蛋白质图集(Human Protein Atlas)在44种人细胞系和32种基本组织中针对19378个Ensembl基因标识符测量的FPKM表达值与转录起始位点下游前10kb中193-199bp频率的平均FTT(快速傅里叶变换)强度之间的关联值。表3也含有蛋白质图集提供的各表达样品的简要描述以及IH01、IH02和BH01样品的排名变换和排名差别。

[0236] 实施例5:从cfDNA确定非健康来源组织

[0237] 为测试是否可推导非健康状态中额外的有贡献的组织,对获自五名晚期癌症患者的cfDNA样品进行测序。这些样品中的核小体间隔模式揭示与非造血组织或细胞系最紧密关联的cfDNA的额外分布,这通常与患者癌症的解剖学来源匹配。

[0238] 癌症患者的cfDNA中的核小体间隔鉴定非造血贡献

[0239] 为确定是否可检测对非健康状态中循环cfDNA有贡献的非造血谱系的签名(signature),筛选了来自具有多种IV期癌症的临床诊断的个体的44种血浆样品并对制备自cfDNA的单链文库进行轻度测序(light sequencing)(表4;中值2.2倍覆盖):

[0240] 表4.癌症小组(panel)的临床诊断与cfDNA产率

[0241]

样品ID	临床疾病	阶段	cfDNA 产率 (ng/ml)	患者性别
IC01 †	肾癌（移行细胞）	IV	242	F
IC02	卵巢癌（不明确）	IV	22.5	F
IC03	皮肤癌（黑素瘤）	IV	12.0	M
IC04	乳腺癌（侵入性/浸染性导管）	IV	12.6	F
IC05	肺癌（腺癌）	IV	5.4	M
IC06	肺癌（间皮瘤）	IV	11.4	M
IC07 †	胃癌（不明确）	IV	52.2	M
IC08	子宫癌（不明确）	IV	15.0	F
IC09	卵巢癌（浆液肿瘤）	IV	8.4	F
IC10	肺癌（腺癌）	IV	11.4	F
IC11	结直肠癌（不明确）	IV	11.4	M
IC12	乳腺癌（侵入性/浸染性小叶）	IV	12.0	F
IC13	前列腺癌（不明确）	IV	12.3	M
IC14	头颈癌（不明确）	IV	27.0	M
IC15 §	肺癌（小细胞）	IV	22.5	M
IC16	膀胱癌（不明确）	IV	14.1	M
IC17 §	肝癌（肝细胞癌）	IV	39.0	M
IC18	肾癌（透明细胞）	IV	10.5	F
IC19	睾丸癌（精原细胞性）	IV	9.6	M
IC20 §	肺癌（鳞状细胞癌）	IV	21.9	M
IC21	胰腺癌（导管腺癌）	IV	35.4	M
IC22	肺癌（腺癌）	IV	11.4	F
IC23	肝癌（肝细胞癌）	IV	17.1	M
IC24	胰腺癌（导管腺癌）	IV	37.2	M
IC25	胰腺癌（导管腺癌）	IV	27.9	M
IC26	前列腺癌（腺癌）	IV	24.6	M
IC27	子宫癌（不明确）	IV	19.2	F
IC28	肺癌（鳞状细胞癌）	IV	33.3	M

[0242]

IC29	头颈癌（不明确）	IV	14.4	M
IC30	食管癌（不明确）	IV	10.5	M
IC31 †	卵巢癌（不明确）	IV	334.8	F
IC32	肺癌（小细胞）	IV	9.6	F
IC33	结直肠癌（腺癌）	IV	13.8	M
IC34	乳腺癌（侵入性/浸染性小叶）	IV	33.6	F
IC35 §	乳腺癌（导管癌，原位）	IV	16.2	F
IC36	肝癌（不明确）	IV	26.4	M
IC37 §	结直肠癌（腺癌）	IV	15.9	F
IC38	膀胱癌（不明确）	IV	6.6	M
IC39	肾癌（不明确）	IV	39.0	M
IC40	前列腺癌（腺癌）	IV	13.8	M
IC41	睾丸癌（精原细胞性）	IV	16.5	M
IC42	肺癌（腺癌）	IV	11.4	F
IC43	皮肤癌（黑素瘤）	IV	21.9	F
IC44	食管癌（不明确）	IV	25.8	F
IC45 †	结直肠癌（腺癌）	IV	3.0	M
IC46 **	乳腺癌（导管癌，原位）	IV	36.6	F
IC47	胰腺癌（导管腺癌）	IV	19.2	F
IC48 **	乳腺癌（侵入性/浸染性小叶）	IV	13.8	F

[0243] §:选择样品进行额外测序。

[0244] **:该样品仅可获得0.5ml血浆。

[0245] †:样品未通过QC且未用于进一步分析。

[0246] 表4显示48名患者的临床和组织学诊断,针对高肿瘤负荷的证据筛选了来自这些患者的血浆传播cfDNA,以及显示来自各个体1.0ml血浆的总cfDNA产率和相关临床协变量(covariaty)。在这48名患者中,44名通过QC并具有充足的材料。在这44名患者中,选择五名进行更深度的测序。通过Qubit Fluorometer 2.0(生命科技公司(Life Technologies))测定cfDNA产率。

[0247] 使用与实施例4的IH02相同的方案制备这些样品且许多样品在与实施例4的IH02相同的批次中。具有IV期癌症的临床诊断的52名个体的人外周血血浆(表4)获自康福赛特生物公司(Conversant Bio)或血浆实验室国际公司(PlasmaLab International)(美国华盛顿州埃弗雷特)并在-80℃下以0.5ml或1ml等分试样储存直至使用。具有系统性红斑狼疮的临床诊断的四名个体的人外周血血浆获自康福赛特生物公司并在-80℃下以0.5ml等分试样储存直至使用。使用前一刻在桌面上解冻冷冻的血浆等分试样。根据生产商的方案使用QiaAMP循环核酸试剂盒(凯杰公司)从2ml的各血浆样品中纯化循环的无细胞DNA。使用Qubit荧光计(英杰公司)定量DNA。为验证样品子集中的cfDNA产率,使用靶向多拷贝人Alu

序列的定制qPCR测定进一步定量纯化的DNA;发现这两个预测结果是一致的。

[0248] 由于匹配的肿瘤基因型不可得,在两个非整倍性(aneuploidy)的度量上对各样品进行评分以鉴定可能含有高比例肿瘤来源cfDNA的子集:首先,来源于各染色体的与预期读数比例的偏差(图62A);以及其次,一组常见单核苷酸多态性的泛染色体(per-chromosome)等位基因平衡概况(图62B)。基于这些度量,对来源于五个个体(具有小细胞肺癌、鳞状细胞肺癌、结直肠腺癌、肝细胞癌和导管癌原位乳腺癌)的单链文库测序至类似于实施例4中IH02的深度(表5;平均30倍覆盖):

[0249] 表5.CA01集合中包括的额外样品的测序统计

[0250]

样品名称	文库类型	读数	测序的片段	比对的	比对的 Q30	覆盖	预测% 重复	35-80bp	120-180bp
IH03	SSP	2x39	53292855	92.66%	72.37%	2.29	15.46%	11.05%	52.34%
IP01 †	DSP	2x101, 2x102	1214536629	97.22%	86.38%	76.11	0.55%	0.08%	62.77%
IP02 †	DSP	2x101, 2x102	855040273	97.16%	87.72%	52.46	0.83%	0.07%	68.10%
IA01	SSP	2x39	53934607	87.42%	68.30%	2.02	22.70%	15.20%	49.77%
IA02	SSP	2x39	42496222	95.42%	76.61%	1.95	4.74%	12.28%	59.00%
IA03	SSP	2x39	51278489	93.12%	71.33%	2.05	25.68%	14.27%	52.57%
IA04	SSP	2x39	50768476	90.30%	70.51%	2.14	7.83%	17.80%	36.76%
IA05	DSP	2x101	194985271	98.80%	90.61%	11.09	12.05%	2.24%	71.67%
IA06	DSP	2x101	171670054	98.90%	90.88%	9.90	5.41%	1.93%	71.26%
IA07	DSP	2x101	208609489	98.67%	90.34%	11.69	11.45%	2.59%	74.84%
IA08	DSP	2x101	193729556	98.81%	90.70%	10.84	11.96%	2.58%	76.24%
IC02	SSP	2x39	57913605	95.07%	75.57%	2.59	5.40%	12.98%	60.00%
IC03	SSP	2x39	63862631	95.78%	75.66%	2.79	8.32%	13.25%	62.20%
IC04	SSP	2x39	55239248	95.47%	76.26%	2.57	8.28%	10.98%	58.48%
IC05	SSP	2x39	39623850	89.80%	69.92%	1.60	9.24%	14.63%	50.33%
IC06	SSP	2x39	59679981	95.57%	74.90%	2.11	3.93%	24.30%	41.46%
IC08	SSP	2x39	46933688	94.38%	74.21%	1.92	5.92%	16.04%	45.25%
IC09	SSP	2x42	59639583	91.22%	71.15%	2.13	6.69%	21.39%	43.50%
IC10	SSP	2x42	53994406	93.73%	73.40%	1.83	2.00%	27.08%	37.62%
IC11	SSP	2x42	59225460	93.25%	72.51%	2.15	5.26%	21.30%	43.33%
IC12	SSP	2x42	57884742	93.52%	74.33%	2.34	2.66%	18.28%	46.58%
IC13	SSP	2x42	71946779	92.94%	72.47%	2.52	2.18%	23.51%	43.97%
IC14	SSP	2x42	61649203	94.54%	73.47%	2.20	3.23%	22.26%	43.37%
IC15	SSP	2x50, 43/42	908512803	95.49%	76.83%	29.77	10.66%	25.42%	38.47%
IC16	SSP	2x42	62739733	92.81%	72.85%	2.47	2.77%	17.71%	48.04%
IC17	SSP	2x50, 2x39	1072374044	96.02%	76.42%	42.08	12.16%	17.08%	50.02%
IC18	SSP	2x39	59976914	87.91%	68.67%	2.24	4.39%	18.85%	44.44%
IC19	SSP	2x39	51447149	89.38%	69.39%	2.02	8.24%	17.30%	46.33%
IC20	SSP	2x50, 2x39	640838540	96.30%	79.11%	23.38	12.43%	25.72%	39.87%
IC21	SSP	2x39	53000679	94.64%	74.57%	1.79	37.39%	29.89%	43.81%
IC22	SSP	2x39	58102606	94.08%	74.08%	2.51	6.24%	13.65%	58.41%
IC23	SSP	2x39	65859970	95.67%	75.67%	2.94	5.34%	11.09%	60.85%
IC24	SSP	43/42	66344431	94.63%	74.46%	2.48	2.00%	22.46%	46.31%
IC25	SSP	43/42	75066833	93.75%	73.66%	2.86	2.24%	21.30%	46.19%
IC26	SSP	43/42	79180860	92.59%	72.32%	2.97	2.93%	22.34%	40.42%
IC27	SSP	43/42	78037377	88.81%	67.04%	2.20	1.50%	31.31%	30.59%
IC28	SSP	43/42	61402081	95.24%	75.74%	2.60	2.46%	18.71%	46.44%
IC29	SSP	2x39	49989522	94.46%	73.36%	1.75	3.03%	25.82%	36.23%
IC30	SSP	2x39	58439504	93.52%	71.19%	1.75	17.35%	29.58%	30.47%
IC32	SSP	43/42	78233981	87.86%	66.80%	2.25	1.79%	30.12%	31.20%
IC33	SSP	43/42	62196185	87.26%	66.71%	1.93	1.93%	27.44%	36.92%
IC34	SSP	43/42	63572169	95.42%	76.74%	2.53	2.35%	19.64%	48.55%

[0251]

样品名称	文库类型	读数	测序的片段	比对的	比对的Q30	覆盖	预测%重复	35-80bp	120-180bp
IC35	SSP	43/42	618554393	86.47%	65.90%	18.22	5.23%	28.18%	35.24%
IC36	SSP	43/42	54402943	94.62%	74.73%	2.21	3.32%	17.02%	52.42%
IC37	SSP	2x50, 43/42	1175553677	93.00%	74.46%	38.22	10.15%	28.47%	35.11%
IC38	SSP	43/42	47981963	89.35%	69.45%	1.78	6.47%	18.59%	43.03%
IC39	SSP	43/42	61968854	95.29%	75.57%	2.62	2.54%	14.42%	57.28%
IC40	SSP	2x39	53228209	93.54%	71.69%	1.81	8.85%	24.88%	34.95%
IC41	SSP	43/42	78081655	87.11%	65.25%	2.26	1.61%	27.94%	35.21%
IC42	SSP	2x39	53017317	93.59%	74.33%	2.02	10.74%	19.04%	44.12%
IC43	SSP	43/42	76395478	88.41%	67.21%	2.40	1.56%	26.68%	37.76%
IC44	SSP	43/42	61354307	95.15%	74.88%	2.45	4.34%	19.10%	46.39%
IC46	SSP	2x39	60123123	94.51%	72.23%	2.13	10.37%	15.46%	50.93%
IC47	SSP	2x39	59438172	95.58%	73.84%	2.07	9.33%	21.67%	43.34%
IC48	SSP	43/42	55704417	91.35%	72.79%	2.01	13.87%	22.56%	38.68%
IC49	DSP	2x101	170489015	99.02%	90.53%	11.19	5.93%	2.41%	59.93%
IC50	DSP	2x101	203828224	98.72%	90.28%	10.82	2.83%	4.81%	66.23%
IC51	DSP	2x101	200454421	98.63%	90.53%	11.77	9.50%	2.58%	67.04%
IC52	DSP	2x101	186975845	98.97%	91.25%	11.37	2.57%	0.83%	68.96%

[0252] SSP,单链文库制备方案。DSP,双链文库制备方案。

[0253] †样品先前已公开(J.O.Kitzman等,Science Translational Medicine(2012))。

[0254] 表5以表格形式显示测序相关统计数据,包括针对各样品的测序的片段总数、读数长度、与具有和不具有映射质量阈值的参考物比对的这类片段百分比、平均覆盖、重复率,以及两个长度箱中测序的片段的比例。片段长度推导自双端读数的比对。由于读数长度较短,通过假定整个片段已读取来计算覆盖。预测的重复片段数目基于片段末端,其在高度固型覆盖存在的情况下可能过度预测了真实重复率。

[0255] 如上文所述,针对相同的76种人细胞系和基本组织的表达数据集,在所有基因体中都对长片段WPS值进行FTT,并将FTT与193-199bp频率范围中的平均强度相关联。与来自实施例4的健康个体的三种样品相反(其中前10中全部关联和前20中几乎全部关联都是与淋巴样或骨髓样谱系的关联),最高排名细胞系或组织中许多代表非造血谱系,其在一些情况下与癌症类型匹配(图61;表3)。例如,对于IC17,其中患者具有肝细胞癌,最高排名的关联是与HepG2,一种肝细胞癌细胞系。对于IC35,其中患者具有导管癌原位乳腺癌,最高排名的关联是与MCF7,一种转移性乳房腺癌细胞系。在其他情况中,在关联排名中表现出最大变化的细胞系或基本组织与癌症类型匹配。例如,对于IC15,其中患者具有小细胞肺癌,关联排名中最大的变化(-31)是针对一种小细胞肺癌细胞系(SCLC-21H)。对于IC20(一种肺鳞状细胞癌)和IC35(一种结直肠腺癌),在关联排名中有许多取代了淋巴样/骨髓样细胞系的非造血癌细胞系,但其与特定癌症类型的匹配较不清楚。可能这些癌症的特定分子概况在76种表达数据集中未被充分代表(例如,这些中无一是肺鳞状细胞癌;CACO-2是一种来源于结直肠腺癌的细胞系,但已知是高度异质性的)。

[0256] 一种贪婪(greedy)迭代方法被用于预测来源于生物样品的多种对cfDNA有贡献的细胞类型和/或组织的比例。首先,鉴定以下细胞类型或组织:其参考图谱(此处由76种RNA表达数据集限定)在给定cfDNA样品的基因体中都与WPS长片段值的193-199bp频率中的平均FFT强度具有最高关联。随后,拟合(fit)一系列“双组织”线性混合模型,包括具有最高关联的细胞类型或组织以及来自完整的参考图谱集合的其他剩余细胞类型或组织中的每一

种。对于后一种集合,具有最高系数的细胞类型或组织被保留作为存在贡献(contributory),除非该系数低于1%,此时程序终止且不纳入该最后的组织或细胞类型。重复该程序,即“三组织”、“四组织”等,直至基于混合模型的预测新添加的组织的贡献小于1%后终止。该混合模型的形式为:

[0257] argmax_{a,b,c,...}cor(Mean_FFTintensity_193-199,a*log2ExpTissue1+

[0258] b*log2Tissue2+c*log2Tissue3+...+(1-a-b-c-...)*log2ExpTissueN).例如,对于IC17,一种来源于具有晚期肝细胞癌的患者的cfDNA样品,该程序预测9种存在贡献的细胞类型,包括Hep_G2(28.6%)、HMC.1(14.3%)、REH(14.0%)、MCF7(12.6%)、AN3.CA(10.7%)、THP.1(7.4%)、NB.4(5.5%)、U.266.84(4.5%),和U.937(2.4%)。对于BH01,一种对应于健康个体混合物的cfDNA样品,该程序预测7种存在贡献的细胞类型或组织,包括骨髓(30.0%)、NB.4(19.6%)、HMC.1(13.9%)、U.937(13.4%)、U.266.84(12.5%)、Karpas.707(6.5%),和REH(4.2%)。应注意,对于来源于癌症患者的样品IC17,预测的贡献中的最高比例对应于与作为该cfDNA来源的患者中存在的癌细胞类型紧密相关的细胞系(Hep_G2和肝细胞癌)。相反地,对于BH01,该方式预测仅对应于主要与造血作用相关的组织或细胞类型的贡献,造血作用是健康个体中血浆cfDNA的主要来源。

[0259] 实施例6:实施例4-5的通用方法

[0260] 样品

[0261] 含有来自未知数目的健康个体的贡献物的主体人外周血血浆获自干细胞科技公司(STEMCELL Technologies)(加拿大不列颠哥伦比亚温哥华)并在-80℃下以2ml等分试样储存直至使用。来自匿名健康供体的单个人外周血血浆获自康福赛特生物公司(美国阿拉巴马州亨茨维尔)并在-80℃下以0.5ml等分试样储存直至使用。

[0262] 在第18和13妊娠周分别获得来自妊娠妇女IP01和IP02的全血并如前文所述41进行处理。

[0263] 具有IV期癌症临床诊断的52名个体的人外周血血浆(附表4)获自康福赛特生物公司或血浆实验室国际公司(美国华盛顿州埃弗雷特)并在-80℃下以0.5ml或1ml等分试样储存直至使用。具有系统性红斑狼疮临床诊断的四名个体的人外周血血浆获自康福赛特生物公司并在-80℃下以0.5ml等分试样储存直至使用。

[0264] 血浆样品的处理

[0265] 使用前一刻在桌面上解冻冷冻的血浆等分试样。根据生产商的方案使用QiaAMP循环核酸试剂盒(凯杰公司)从2ml的各血浆样品中纯化循环的无细胞DNA。使用Qubit荧光计(英杰公司)定量DNA。为验证样品子集中的cfDNA产率,随后使用靶向多拷贝人Alu序列的定制qPCR测定进一步定量纯化的DNA;发现两个预测值相一致。

[0266] 双链测序文库的制备

[0267] 使用ThruPLEX-FD或ThruPLEX DNA-seq 48D试剂盒(鲁比孔基因组公司)制备条形码化的测序文库,包括一系列专有的末端修复、连接和扩增反应。使用0.5ng至30.0ng的cfDNA作为所有临床样品文库的输入物。通过实时PCR监测所有样品的文库扩增以避免过度扩增且通常在4-6次循环后终止。

[0268] 单链测序文库的制备

[0269] 通过以下步骤制备接头2:合并4.5μl TE(pH 8)、0.5μl 1M NaCl、10μl 500uM寡聚

接头2.1和10μl 500μM寡聚接头2.2,在95℃下孵育10秒,并以0.1℃/秒的速率将温度降低至14℃。通过以下步骤对纯化的cfDNA片段进行去磷酸化:在20μl反应体积中合并2x CircLigase II缓冲液(表观中心公司(Epicentre))、5mM $MnCl_2$和1U FastAP碱性磷酸酶(赛默飞世尔公司)与0.5-10ng片段并在37℃下孵育30分钟。随后通过加热至95℃持续3分钟对片段进行变性,并将片段立即转移至冰浴中。对反应物附加生物素偶联的接头寡CL78(5pmol)、20% PEG-6000(w/v)和200U CircLigase II(表观中心公司公司),总体积为40μl,并在60℃下旋转孵育过夜,加热至95℃持续3分钟,并置于冰浴中。对各样品,在珠结合缓冲液(BBB)(10mM Tris-HCl[pH 8]、1M NaCl、1mM EDTA[pH 8]、0.05%吐温20和0.5% SDS)中对20μl MyOne C1珠(生命科技公司)清洗两次,并重悬于250μl BBB中。通过室温下旋转60分钟将连接有接头的片段连至珠。在磁力架上收集珠并弃去上清液。使用500ul清洗缓冲液A(WBA)(10mM Tris-HCl[pH 8]、1mM EDTA[pH 8]、0.05%吐温20、100mM NaCl、0.5% SDS)对珠清洗一次并使用500ul清洗缓冲液B(WBB)(10mM Tris-HCl[pH 8]、1mM EDTA[pH 8]、0.05%吐温20、100mM NaCl)清洗一次。在50μl的反应体积中将珠与1X等温扩增缓冲液(NEB公司)、2.5μM寡聚CL9、250μM(各)dNTP和24U Bst 2.0DNA聚合酶(NEB公司)合并,并通过以1℃/分钟将温度从15℃升高至37℃来轻柔震荡孵育,并在37℃下保持10分钟。在磁力架上收集后,使用200μl WBA对珠清洗一次,重悬于200μl严格清洗缓冲液(SWB)(0.1X SSC,0.1%SDS)中并在45℃下孵育3分钟。随后再次收集珠并使用200μl WBB清洗一次。随后将珠与1X CutSmart缓冲液(NEB公司)、0.025%吐温20、100μM(各)dNTP和5U T4 DNA聚合酶(NEB公司)合并并在室温下轻柔震荡孵育30分钟。如上文所述使用各WBA、SWB和WBB对珠清洗一次。随后将珠与1X CutSmart缓冲液(NEB公司)、5% PEG-6000、0.025%吐温20、2μM双链接头2和10U T4 DNA连接酶(NEB公司)混合并在室温下轻柔震荡孵育2小时。如上文所述使用各WBA、SWB和WBB对珠清洗一次并重悬于25μl TET缓冲液(10mM Tris-HCl[pH 8]、1mM EDTA[pH 8]、0.05%吐温20)中。通过以下方法将第二链从珠上洗脱:加热至95℃,在磁力架上收集珠并将上清液转移至新的管中。通过实时PCR监测所有样品的文库扩增以避免过度扩增且需要平均每个文库4-6次循环。

[0270] 测序

[0271] 在HiSeq 2000或NextSeq 500仪器(伊露米娜公司)上对所有文库进行测序。

[0272] 基本测序数据处理

[0273] 对条形码化的双末端(paired end,PE)伊露米娜公司测序数据进行拆分,允许条形码序列中最多一个取代。短于或等于读数长度的读数被一致性识别并修减接头。剩余的一致性单末端读数(SR)和单个PE读数被使用BWA v0.7.10中实施的ALN算法比对至人参考基因组序列(GRCh37,1000基因组2相技术参考物,下载自ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/)。使用BWA SAMPE进一步处理PE读数以解析读数对的模糊布局或通过定位的读数端周围更敏感的比对步骤来重新完成缺失的比对。使用SAMtools API将比对的SR和PE数据直接转化为分选的BAM格式。在所有泳道和测序轮次中都合并样品的BAM文件。

[0274] 使用FastQC(v0.11.2)进行质量控制,获得文库复杂性预测(Picard工具v1.113),测定接头二聚体的比例,推导的文库插入尺寸的分析,外部读数端处核苷酸和二核苷酸频率以及检查各文库的映射质量分布。

[0275] 模拟的读数数据集合

[0276] 针对人参考物(GRC37h)的所有主要染色体模拟比对的测序数据(如果比45bp短则SR,否则PE 45bp)。出于该目的,从全部两个读数端和全部两条链取向上的真实数据中测定二核苷酸频率。还针对全部两条链上的参考基因组记录二核苷酸频率。此外,针对1-500bp范围提取真实数据的插入尺寸分布。通过迭代模拟读数,该迭代通过主要参考染色体的序列进行。在各步骤处(即各位置处的一个或多个时间,其取决于所需覆盖),(1)随机选择链,(2)使用真实数据中二核苷酸频率相对于参考序列中频率的比率以随机决定是否考虑初始二核苷酸,(3)从提高的插入尺寸分布中取样插入尺寸以及(4)使用终端二核苷酸的频率比率来随机决定是否报告生成的比对。在除去PCR重复后将模拟的覆盖匹配至原始数据的覆盖。

[0277] 覆盖、读数起始和窗口保护评分

[0278] 本发明的数据提供了涉及测序文库制备中使用的DNA分子的两个物理末端的信息。我们使用SAMtools应用程序接口(API)从BAM文件提取该信息。作为读数起始,我们使用PE数据的全部两种外部比对坐标,其中全部两种读数均与相同染色体比对且读数具有相反的取向。在通过接头修剪(adapter trimming)将PE数据转化为单个读数数据的情况下,我们将SR比对的全部两种末端坐标视作读数起始。对于覆盖,我们考虑两个(推导的)分子末端之间的全部位置,包括这些末端位置。我们将窗口尺寸k的窗口化保护评分(WPS)定义为跨窗口的分子数减去被该窗口包围的任意碱基处起始的那些。我们将确定的WPS分配至窗口中心。对于35-80bp范围的分子(短组分),我们使用16的窗口尺寸,且对于120-180bp范围的分子(长组分),我们使用120的窗口尺寸。

[0279] 核小体峰识别

[0280] 从长组分WPS中识别核小体保护的局部极大值,我们对其局部调整至流动中值(running median)为零(1kb窗口)并使用Savitzky-Golay过滤器平滑化(窗口尺寸21,2阶多项式)。随后将WPS踪迹区段化为大于零区域(允许最多5个连续位置小于零)。如果所得区域长度为50-150bp,我们鉴定该区域的中值数值并搜寻该中值以上的最大总和连续窗口。我们报告该窗口的起始、终止和中心坐标。从这些中心坐标计算峰至峰距离等。将识别的评分确定为窗口中最大值和该区域周边两个相邻WPS极小值的平均值之间的距离。如果鉴定的区域长度为150-450bp,我们应用系统的上述中值连续窗口方法,但仅报告尺寸为50-150bp的那些窗口。对于来源于150-450bp区域的多个窗口的评分计算,我们假定该区域内的周边极小值是零。我们弃用比50bp短和比450bp长的区域。

[0281] 167bp片段的二核苷酸组成

[0282] 在样品内过滤具有推导的恰好167bp长度的片段(对应于片段尺寸分布的主峰(dominant peak))以除去重复。以链意识(strand-aware)的方式计算二核苷酸频率,其中在各位置处使用滑动的2bp窗口和参考等位基因,在一个片段末端的上游50bp开始并在另一端的下游50bp结束。比较各位置处观察到的二核苷酸频率与由模拟读数集合测得的预测二核苷酸频率,其反映以文库特异性方式计算的相同切割偏好(详细信息参见上文)。

[0283] 基因组特征和转录因子结合位点周围的WPS概况

[0284] 分析开始于(基于基序的)聚类FIMO区间的初始集合,其定义计算机预测的转录因子结合位点的集合。对于聚类转录因子的子集(AP-2-2、AP-2、CTCF_Core-2、E2F-2、EBF1、

Ebox-CACCTG、Ebox、ESR1、ETS、IRF-2、IRF-3、IRF、MAFK、MEF2A-2、MEF2A、MYC-MAX、PAX5-2、RUNX2、RUNX-AML、STAF-2、TCF-LEF、YY1),基于实验数据将位点的结合限定为更可信的活跃结合的转录因子结合位点。出于此目的,仅保留与来自公共可得ENCODE数据(从UCSC下载的TfbsClusteredV3集合)的ChIP-seq实验定义的峰重叠的预测结合位点。

[0285] 针对CHO1样品和相应模拟物提取这些位点周围的窗口化保护评分。相对于各结合位点的起始坐标计算各位置处各位点/特征的保护评分并汇总。移动CTCF结合位点的图使得x轴上的零坐标位于CTCF的已知52bp结合足迹的中心。随后从初始信号中扣除5kb提取的WPS信号的第一和最后500bp的平均值(其主要是扁平的且代表平均偏移)。仅对于长片段信号,使用200bp窗口计算滑动窗口平均值并从初始信号中扣除。最后,从校正的CHO1的WPS概况中扣除校正的模拟的WPS概况以校正作为连接偏好和片段长度产物的信号。对该最终概况作图并命名为“调整的WPS”。

[0286] 基因组特征(例如转录起始位点、转录终止位点、起始密码子、剪接供体和剪接受体位点)获自Ensembl Build版本75。针对转录因子结合位点如上文所述计算这些特征周围的调整的WPS并作图。

[0287] CTCF结合位点和相应WPS周围核小体间隔的分析

[0288] 用于该分析的CTCF位点首先包括CTCF结合位点的聚类FIMO预测结果(经由基序的计算机预测)。我们随后创建该集合的两个额外子集:1)与可通过ENCODE TfbsClusteredV3(参见上文)获得的CTCF ChIP-seq峰集合的交集,和2)与以下CTCF位点的集合的交集:这些CTCF位点通过实验观察为在19种组织中都活跃。

[0289] 针对各位点提取结合位点各侧上的10个核小体的位置。我们计算了所有相邻核小体之间的距离以获得各位点集合的核小体间距离的分布。-1至+1核小体的分布大幅变化,偏移为较大间隔,特别是在230-270bp范围中。这暗示真正活跃的CTCF位点朝-1至+1核小体之间的较大间隔偏移,且长和短读数组分的WPS中的差异可因此是明显的。因此,额外地计算相对于CTCF位点中心的各位置处的平均短和长片段WPS。为探索核小体间隔的作用,在-1至+1核小体间隔小于160、160-200、200-230、230-270、270-420、420-460且大于420bp的箱内提取该平均值。这些区间大致捕获感兴趣的间隔,如更确信的活跃位点的230-270bp处的形成峰(emerging peak)和主峰。

[0290] DNA酶I超敏位点(DHS)的分析

[0291] 从华盛顿大学编码数据库(University of Washington Encode database)中下载通过Maurano等(Science,卷337(6099),第1190-95页(2012);“all_fdr0.05_hot”文件,最后修改于2012年2月13日)的BED格式的349种基本组织和细胞系样品的DHS峰。从分析中移除包含这些峰集合中233个的来源于胎儿组织的样品,因为其在组织类型内表现得不一致,这可能是因为各组织样品内多种细胞类型的不等表现。代表多种细胞谱系的116种样品被保留以进行分析。对于具体集合中各DHS峰的中点,鉴定了CHO1识别集中最近的上游和下游识别信号,且计算了那两个识别信号的中心之间的基因组距离。所有这类距离的分布都针对各DHS峰识别信号集(callset)显示,其中使用针对0至500bp的距离计算的平滑化的强度预测值。

[0292] 基因表达分析

[0293] 该研究中使用FPKM表达值,其通过人蛋白质图集(Human Protein Atlas)

("ma.csv"文件)针对44种人细胞系和32种基本组织中的20344个Ensembl基因标识符测量。对于跨越组织的分析,排除了具有少于3个非零表达值的基因(19378个基因通过该筛选)。提供了表达数据集,其中对于FPKM值精确到小数点后一位。因此,零表达值(0.0)表示表达在0和小于0.05的数值之间。除非另有说明,否则在表达值的$\log_2$变换前将最小表达值设为0.04FPKM。

[0294] 平滑周期图和轨迹线的平滑化

[0295] 长片段WPS被用于使用快速傅里叶变换(FFT,R统计编程环境中的spec.pgram)计算基因组区域的周期图,其中频率为1/500碱基至1/100碱基。任选地额外使用对数据进行平滑化(3bp Daniell平滑器;移动平均给予端值一半权重)和去趋势化(即扣除系列的平均值和移除线性趋势)的参数。

[0296] 注明处,R统计编程环境中应用的递归时间系列过滤器被用于从轨迹线中除去高频率变化。使用24过滤器频率(1/seq(51004)),且使用前24个轨迹线值作为初始值。所得轨迹线中24值偏移的调节通过重复该轨迹线的后24个值来进行。

[0297] FTT强度和表达值的关联

[0298] 分析了120-280bp范围的基因表达环境中从平滑周期图(FFT)测得的强度值。观察到主要核小体间距离峰周围的基因表达值和FFT强度之间的S型皮尔森(Pearson)关联。在193-199bp范围中观察到显著的负关联。结果是,该频率范围中的强度与$\log_2$变换的表达值平均关联。

[0299] 其他实施例

[0300] 实施例7.一种确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,该方法包括:

[0301] 从来自该对象的生物样品中分离cfDNA,分离的cfDNA包含复数个cfDNA片段;

[0302] 确定与复数个cfDNA片段的至少一部分相关的序列;

[0303] 根据这些cfDNA片段序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;以及

[0304] 根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA片段的产生的组织和/或细胞类型。

[0305] 实施例8.实施例7的方法,其中确定至少一些导致cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较至少一些导致cfDNA片段末端的产生的基因组位置。

[0306] 实施例9.实施例7或实施例8的方法,其中确定至少一些导致cfDNA片段的产生的组织和/或细胞类型的步骤包括对至少一些cfDNA片段末端的基因组位置的分布进行数学变换。

[0307] 实施例10.实施例9的方法,其中该数学变换包括傅里叶变换。

[0308] 实施例11.前述实施例中任一项的方法,还包括确定该参考基因组的至少一些坐标中每一个的评分,其中该评分根据至少复数个cfDNA片段末端及其基因组位置确定,且其中确定至少一些导致观察到的cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较这些评分。

[0309] 实施例12.实施例11的方法,其中坐标的评分表示该坐标是cfDNA片段末端位置的

可能性或与其相关。

[0310] 实施例13.实施例8-12中任一项的方法,其中该参考图谱包括由至少一种细胞类型或组织生成的DNA酶I超敏位点图谱。

[0311] 实施例14.实施例8-13中任一项的方法,其中该参考图谱包括由至少一种细胞类型或组织生成的RNA表达图谱。

[0312] 实施例15.实施例8-14中任一项的方法,其中该参考图谱由来自动物的cfDNA生成,其中已将人组织或细胞异种移植至该动物。

[0313] 实施例16.实施例8-15中任一项的方法,其中该参考图谱包括由至少一种细胞类型或组织生成的染色体构象图谱。

[0314] 实施例17.实施例8-16中任一项的方法,其中该参考图谱包括由至少一种细胞类型或组织生成的染色质可及性图谱。

[0315] 实施例18.实施例8-17中任一项的方法,其中该参考图谱包括来自获自至少一个参考对象的样品的序列数据。

[0316] 实施例19.实施例8-18中任一项的方法,其中该参考图谱对应于至少一种与疾病或紊乱相关的细胞类型或组织。

[0317] 实施例20.实施例8-19中任一项的方法,其中该参考图谱包括组织或细胞类型中核小体和/或染色质小体的位置或间隔。

[0318] 实施例21.实施例8-20中任一项的方法,其中该参考图谱通过使用外源性核酸酶(如微球菌核酸酶)消化获自至少一种细胞类型或组织的染色质生成。

[0319] 实施例22.实施例8-21中任一项的方法,其中这些参考图谱包括来自至少一种细胞类型或组织的通过基于转座的方法(如ATAC-seq)测定的染色质可及性数据。

[0320] 实施例23.实施例8-22中任一项的方法,其中这些参考图谱包括与组织或细胞类型的DNA结合和/或DNA占据蛋白质的位置相关的数据。

[0321] 实施例24.实施例23的方法,其中该DNA结合和/或DNA占据蛋白质是转录因子。

[0322] 实施例25.实施例23或实施例24的方法,其中这些位置是通过交联的DNA-蛋白质复合物的染色质免疫沉淀测定的。

[0323] 实施例26.实施例23或实施例24的方法,其中这些位置是通过使用核酸酶(如DNA酶I)处理该组织或细胞类型相关DNA来测定的。

[0324] 实施例27.实施例8-26中任一项的方法,其中该参考图谱包括涉及组织或细胞类型内核小体、染色质小体或其他DNA结合或DNA占据蛋白质的位置或间隔的生物特征。

[0325] 实施例28.实施例27的方法,其中该生物特征是一种或多种基因的定量表达。

[0326] 实施例29.实施例27或实施例28的方法,其中该生物特征是存在或不存在一种或多种组蛋白标记。

[0327] 实施例30.实施例27-29中任一项的方法,其中该生物特征是对核酸酶切割超敏。

[0328] 实施例31.实施例8-30中任一项的方法,其中用于生成参考图谱的组织或细胞类型是来自具有疾病或紊乱的对象的基本组织。

[0329] 实施例32.实施例31的方法,其中该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0330] 实施例33.实施例8-30中任一项的方法,其中用于生成参考图谱的组织或细胞类型是来自健康对象的基本组织。

[0331] 实施例34.实施例8-30中任一项的方法,其中用于生成参考图谱的组织或细胞类型是永生细胞系。

[0332] 实施例35.实施例8-30中任一项的方法,其中用于生成参考图谱的组织或细胞类型来自肿瘤的活检切片。

[0333] 实施例36.实施例18的方法,其中该序列数据包括cfDNA片段末端的位置。

[0334] 实施例37.实施例36的方法,其中该参考对象是健康的。

[0335] 实施例38.实施例36的方法,其中该参考对象具有疾病或紊乱。

[0336] 实施例39.实施例38的方法,其中该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0337] 实施例40.实施例19-39中任一项的方法,其中该参考图谱包括与该组织或细胞类型相关的参考基因组的至少一部分坐标的参考评分。

[0338] 实施例41.实施例40的方法,其中该参考图谱包括这些评分的数学变换。

[0339] 实施例42.实施例40的方法,其中这些评分代表该组织或细胞类型的全部参考基因组坐标的子集。

[0340] 实施例43.实施例42的方法,其中该子集与核小体和/或染色质小体的位置或间隔相关。

[0341] 实施例44.实施例42或实施例43的方法,其中该子集与转录起始位点和/或转录终止位点相关。

[0342] 实施例45.实施例42-44中任一项的方法,其中该子集与至少一种转录因子的结合位点相关。

[0343] 实施例46.实施例42-45中任一项的方法,其中该子集与核酸酶超敏位点相关。

[0344] 实施例47.实施例40-46中任一项的方法,其中该子集额外地与至少一种正交生物特征相关。

[0345] 实施例48.实施例47的方法,其中该正交生物特征与高表达基因相关。

[0346] 实施例49.实施例47的方法,其中该正交生物特征与低表达基因相关。

[0347] 实施例50.实施例41-49中任一项的方法,其中该数学变换包括傅里叶变换。

[0348] 实施例51.实施例11-50中任一项的方法,其中复数个评分的至少一个子集具有阈值以上的评分。

[0349] 实施例52.实施例7-51中任一项的方法,其中根据至少一些cfDNA片段末端的复数个基因组位置确定导致cfDNA的产生的组织和/或细胞类型的步骤包括比较至少一些cfDNA片段末端的复数个基因组位置或其数学变换的傅里叶变换与参考图谱。

[0350] 实施例53.前述实施例中任一项的方法,还包括生成报告,该报告包括确定的导致cfDNA的产生的组织和/或细胞类型的列表。

[0351] 实施例54.一种鉴定对象中疾病或紊乱的方法,该方法包括:

[0352] 从来自该对象的生物样品中分离无细胞DNA(cfDNA),分离的cfDNA包含复数个cfDNA片段;

[0353] 确定与复数个cfDNA片段的至少一部分相关的序列;

[0354] 根据cfDNA片段序列确定复数个cfDNA片段的至少一些cfDNA片段末端的参考基因组内的基因组位置;

[0355] 根据至少一些cfDNA片段末端的基因组位置确定至少一些导致cfDNA的产生的组织和/或细胞类型;以及

[0356] 根据确定的导致cfDNA的产生的组织和/或细胞类型鉴定该疾病或紊乱。

[0357] 实施例55.实施例54的方法,其中确定导致cfDNA的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较至少一些cfDNA片段末端的基因组位置。

[0358] 实施例56.实施例54或实施例55的方法,其中确定导致cfDNA的产生的组织和/或细胞类型的步骤包括对复数个cfDNA片段末端中至少一些的基因组位置的分布进行数学变换。

[0359] 实施例57.实施例56的方法,其中该数学变换包括傅里叶变换。

[0360] 实施例58.实施例54-57中任一项的方法,还包括测定参考基因组的至少一些坐标中每一个的评分,其中该评分根据至少复数个cfDNA片段末端及其基因组位置测定,且其中确定至少一些导致观察到的cfDNA片段的产生的组织和/或细胞类型的步骤包括与一个或多个参考图谱比较这些评分。

[0361] 实施例59.实施例58的方法,其中坐标的评分表示该坐标是cfDNA片段末端的位置的可能性或与其相关。

[0362] 实施例60.实施例55-59中任一项的方法,其中该参考图谱包括DNA酶I超敏位点图谱、RNA表达图谱、表达数据、染色体构象图谱、染色质可及性图谱、染色质片段化图谱,或获自样品(该样品获自至少一个参考对象)且对应于与疾病或紊乱相关的至少一种细胞类型或组织的序列数据,和/或组织或细胞类型中核小体或染色质小体的位置或间隔。

[0363] 实施例61.实施例55-60中任一项的方法,其中该参考图谱通过使用外源性核酸酶(如微球菌核酸酶)消化来自至少一种细胞类型或组织的染色质生成。

[0364] 实施例62.实施例60或实施例61的方法,其中这些参考图谱包括染色质可及性数据,其通过向来自至少一种细胞类型或组织的细胞核或染色质应用基于转座的方法(如ATAC-seq)来测定。

[0365] 实施例63.实施例55-62中任一项的方法,其中这些参考图谱包括组织或细胞类型的与DNA结合和/或DNA占据蛋白质的位置相关的数据。

[0366] 实施例64.实施例63的方法,其中该DNA结合和/或DNA占据蛋白质是转录因子。

[0367] 实施例65.实施例63或实施例64的方法,其中这些位置是通过向至少一种细胞类型或组织应用交联的DNA-蛋白质复合物的染色质免疫沉淀来测定的。

[0368] 实施例66.实施例63或实施例64的方法,其中这些位置是通过使用核酸酶(如DNA酶I)处理与该组织或细胞类型相关的DNA来测定的。

[0369] 实施例67.实施例54-66中任一项的方法,其中该参考图谱包括与组织或细胞类型中核小体、染色质小体或其他DNA结合或DNA占据蛋白质的位置或间隔相关的生物特征。

[0370] 实施例68.实施例67的方法,其中该生物特征是一个或多个基因的定量表达。

[0371] 实施例69.实施例67或实施例68的方法,其中该生物特征是存在或不存在一种或多种组蛋白标记。

[0372] 实施例70.实施例67-69中任一项的方法,其中该生物特征是对核酸酶切割超敏。

[0373] 实施例71.实施例55-70中任一项的方法,其中用于生成参考图谱的组织或细胞类型是来自具有疾病或紊乱的对象的基本组织。

[0374] 实施例72.实施例71的方法,其中该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0375] 实施例73.实施例55-70中任一项的方法,其中用于生成参考图谱的组织或细胞类型是来自健康对象的基本组织。

[0376] 实施例74.实施例55-70中任一项的方法,其中用于生成参考图谱的组织或细胞类型是永生细胞系。

[0377] 实施例75.实施例55-70中任一项的方法,其中用于生成参考图谱的组织或细胞类型是来自肿瘤的活检切片。

[0378] 实施例76.实施例60的方法,其中获自样品的测序数据包括cfDNA片段末端可能性的位置,这些样品获自至少一个参考对象。

[0379] 实施例77.实施例76的方法,其中该参考对象是健康的。

[0380] 实施例78.实施例76的方法,其中该参考对象具有疾病或紊乱。

[0381] 实施例79.实施例78的方法,其中该疾病或紊乱选自下组:癌症、正常妊娠、妊娠并发症(如非整倍体妊娠)、心肌梗死、炎症性肠病、系统性自身免疫病、局部自身免疫病、具有排斥的异体移植、不具有排斥的异体移植、中风和局部组织损伤。

[0382] 实施例80.实施例60-79中任一项的方法,其中该参考图谱包括与该组织或细胞类型相关的至少一部分参考基因组的cfDNA片段末端可能性。

[0383] 实施例81.实施例80的方法,其中该参考图谱包括cfDNA片段末端可能性的数学变换。

[0384] 实施例82.实施例80的方法,其中这些cfDNA片段末端可能性代表该组织或细胞类型的所有参考基因组坐标的子集。

[0385] 实施例83.实施例82的方法,其中该子集与核小体和/或染色质小体的位置或间隔相关。

[0386] 实施例84.实施例82或实施例83的方法,其中该子集与转录起始位点和/或转录终止位点相关。

[0387] 实施例85.实施例82-84中任一项的方法,其中该子集与至少一个转录因子的结合位点相关。

[0388] 实施例86.实施例82-85中任一项的方法,其中该子集与核酸酶超敏位点相关。

[0389] 实施例87.实施例82-86中任一项的方法,其中该子集与至少一种正交生物特征相关。

[0390] 实施例88.实施例87的方法,其中该正交生物特征与高表达基因相关。

[0391] 实施例89.实施例87的方法,其中该正交生物特征与低表达基因相关。

[0392] 实施例90.实施例81-89中任一项的方法,其中该数学变换包括傅里叶变换。

[0393] 实施例91.实施例58-90中任一项的方法,其中复数个cfDNA片段末端评分的至少一个子集各自具有阈值以上的评分。

[0394] 实施例92.实施例54-91中任一项的方法,其中根据至少一些cfDNA片段末端的复数个基因组位置确定cfDNA的组织和/或细胞类型的步骤包括比较至少一些cfDNA片段末端的复数个基因组位置或其数学变换的傅里叶变换与参考图谱。

[0395] 实施例93.实施例54-92中任一项的方法,其中该参考图谱包括对应于与该疾病或紊乱相关的至少一种组织的DNA或染色质片段化数据。

[0396] 实施例94.实施例54-93中任一项的方法,其中该参考图谱与人相关。

[0397] 实施例95.实施例54-94中任一项的方法,还包括生成报告,该报告包括鉴定该疾病或紊乱的声明。

[0398] 实施例96.实施例95的方法,其中该报告还包括确定的分离的cfDNA的组织和/或细胞类型的列表。

[0399] 实施例97.前述实施例中任一项的方法,其中该生物样品包括以下物质、基本由以下物质组成或由以下物质组成:全血、外周血浆、尿液或脑脊液。

[0400] 实施例98.一种确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,包括:

[0401] (i)通过以下步骤获得核小体图谱:从该对象中获得生物样品,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);

[0402] (ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象中获得生物样品,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);以及

[0403] (iii)通过比较来源于cfDNA的核小体图谱与核小体图谱参考集合来确定导致cfDNA的产生的组织和/或细胞类型;

[0404] 其中(a)、(b)和(c)是:

[0405] (a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;

[0406] (b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和

[0407] (c)人基因组中任意特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

[0408] 实施例99.一种确定对象中导致无细胞DNA(cfDNA)的产生的组织和/或细胞类型的方法,包括:

[0409] (i)通过从该对象获得生物样品来生成核小体图谱,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);

[0410] (ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象中获得生物样品,从该生物样品中分离cfDNA,以及通过DNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c),该DNA来源于使用微球菌核酸酶(MNase)消化染色质、DNA酶处理或ATAC-Seq;以及

[0411] (iii)通过比较来源于来自该生物样品的cfDNA的核小体图谱与核小体图谱参考集合来确定来自导致cfDNA的产生的组织和/或细胞类型;

[0412] 其中(a)、(b)和(c)是:

[0413] (a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;

[0414] (b)人基因组碱基对中任意一对呈现为一对测序的片段终端的可能性分布;和

[0415] (c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

[0416] 实施例100.一种诊断对象中临床病症的方法,包括:

[0417] (i)通过以下步骤生成核小体图谱:从该对象中获得生物样品,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);

[0418] (ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象中获得生物样品,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);以及

[0419] (iii)通过比较来源于cfDNA的核小体图谱与核小体图谱参考集合来确定该临床病症;

[0420] 其中(a)、(b)和(c)是:

[0421] (a)cfDNA片段终端处出现人基因组中任何特定碱基对的可能性分布;

[0422] (b)人基因组碱基对中任意一对呈现为一对cfDNA片段终端的可能性分布;和

[0423] (c)人基因组中任何特定碱基对因差异性核小体占位而出现在cfDNA片段中的可能性分布。

[0424] 实施例101.一种诊断对象中临床病症的方法,包括:

[0425] (i)通过以下步骤获得核小体图谱:从该对象中获得生物样品,从该生物样品中分离cfDNA,以及通过cfDNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c);

[0426] (ii)通过以下步骤生成核小体图谱参考集合:从一个或多个具有已知疾病的对照对象中获得生物样品,从该生物样品中分离cfDNA,以及通过DNA的大规模平行测序和文库构建来测量分布(a)、(b)和/或(c),该DNA来源于使用微球菌核酸酶(MNase)消化染色质、DNA酶处理或ATAC-Seq;以及

[0427] (iii)通过比较来源于cfDNA的核小体图谱与核小体图谱参考集合来确定cfDNA的来源组织和/或细胞类型;

[0428] 其中(a)、(b)和(c)是:

[0429] (a)测序的片段终端处出现人基因组中任何特定碱基对的可能性分布;

[0430] (b)人基因组碱基对中任意一对呈现为一对测序片段终端的可能性分布;和

[0431] (c)人基因组中任何特定碱基对因差异性核小体占位而出现在测序的片段中的可能性分布。

[0432] 实施例102.实施例98-101中任一项的方法,其中该核小体图谱通过以下步骤生成:

[0433] 纯化分离自该生物样品的cfDNA;

[0434] 通过接头连接和任选的PCT扩增来构建文库;以及

[0435] 对所得文库进行测序。

[0436] 实施例103.实施例98-101中任一项的方法,其中该核小体图谱参考集合通过以下步骤生成:

[0437] 纯化分离自来自对照对象的生物样品的cfDNA;

[0438] 通过接头连接和任选的PCT扩增来构建文库;以及

[0439] 对所得文库进行测序。

[0440] 实施例104.实施例98-101中任一项的方法,其中在连续窗口中对分布(a)、(b)或(c)或这些分布之一的数学变换进行傅里叶变换,随后定量与核小体占位相关的频率范围的强度,从而总结核小体在各连续窗口内表现出结构化定位的程度。

[0441] 实施例105.实施例98-101中任一项的方法,其中在分布(a)、(b)或(c)或这些分布之一的数学变换中,我们定量参考人基因组中位点的分布,对于该参考人基因组对紧邻特定转录因子(TF)的转录因子结合位点(TFBS)的读数起始位点图谱进行测序,其在该TF结合该TFBS时通常侧翼紧接核小体,从而总结对cfDNA有贡献的细胞类型中因TF活性而形成核小体定位。

[0442] 实施例106.实施例98-101中任一项的方法,其中在其他基因组地标周围根据来自分布(a)、(b)和/或(c)或这些分布之一的数学变换的任一种汇总信号来总结核小体占位信号,这些其他基因组地标是例如DNA酶I超敏位点、转录起始位点、拓扑结构域、其他表观遗传标记或其他数据集(如基因表达等)中相关联行为所定义的所有这类位点的子集。

[0443] 实施例107.实施例98-101中任一项的方法,其中对这些分布进行变换以汇总或总结复数个基因组子集内核小体定位的周期信号,例如定量连续窗口中的周期,或者由转录因子结合位点、基因模式特征(如转录起始位点)、组织表达数据或核小体定位的其他关联物限定的不连续基因组子集中的周期。

[0444] 实施例108.实施例98-101中任一项的方法,其中这些分布通过组织特异性数据来限定,该组织特异性数据即汇总的组织特异性DNA酶I超敏位点周围信号。

[0445] 实施例109.实施例98-101中任一项的方法,还包括用于比较额外的核小体图谱与参考集合的统计信号处理步骤。

[0446] 实施例110.实施例109的方法,其中我们首先总结多种样品集合中沿基因组的连续窗口内的长程核小体排序,随后进行主成分分析(PCA)以对样品进行聚类或预测混合物成分。

[0447] 实施例111.实施例100或实施例101的方法,其中该临床病症是癌症,即恶性肿瘤。

[0448] 实施例112.实施例111的方法,其中该生物样品是含有cfDNA的循环血浆,其中一些成分来源于肿瘤。

[0449] 实施例113.实施例100或实施例101的方法,其中该临床病症选自组织损伤、心肌梗死(心脏组织的急性损伤)、自身免疫病(多种组织的慢性损伤)、妊娠、染色体畸变(如三染色体)和移植排斥。

[0450] 实施例114.前述实施例中任一项的方法,还包括对确定为对cfDNA有贡献的一种或多种组织或细胞类型中的每一种分配比例。

[0451] 实施例115.实施例114的方法,其中分配给一种或多种确定的组织或细胞类型中每一种的比例至少部分基于相对于来自一个或多个健康对象的cfDNA的一定程度的关联性或增加的关联性。

[0452] 实施例116.实施例114或实施例115的方法,其中关联性的程度至少部分基于来自该生物样品的cfDNA片段末端的分布的数学变换与经确定的组织或细胞类型相关参考图谱的比较。

[0453] 实施例117实施例114-116的方法,其中分配给一种或多种确定的组织或细胞类型中每一种的比例基于混合物模型。

[0454] 通过前文,应理解本发明的特定实施方式已在本发明中描述用于说明性目的,但可进行多种修改而不背离本发明的范围。因此,本发明除所附权利要求外不受限制。

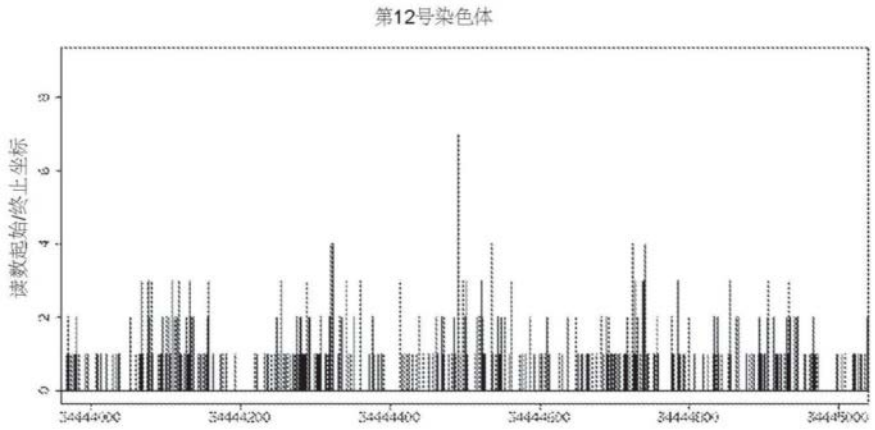

第12号染色体
读数起始/终止坐标
0
2
4
6
8
34444000
34444200
34444400
34444600
34444800
34445000

图1A

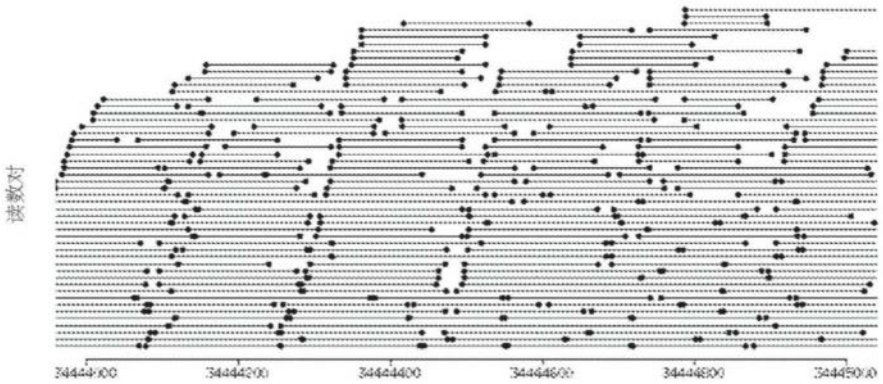

读数对
34444000
34444200
34444400
34444600
34444800
34445000

图1B

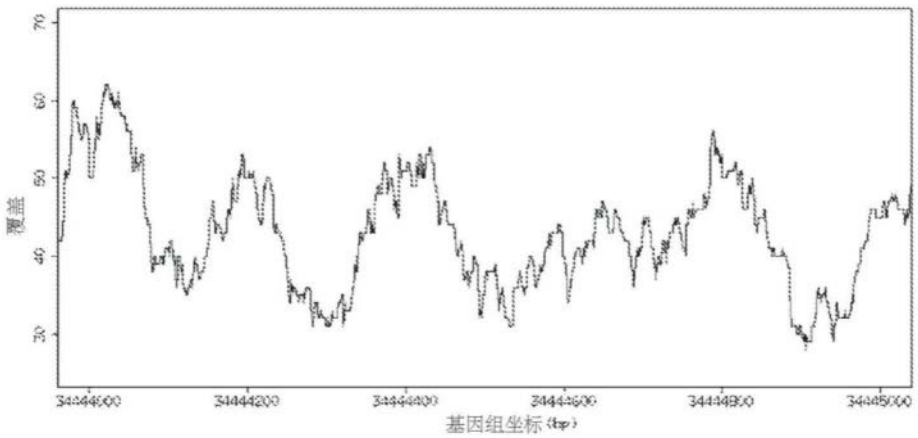

覆盖
30
40
50
60
70
34444000
34444200
34444400
34444600
34444800
34445000
基因组坐标 (bp)

图1C

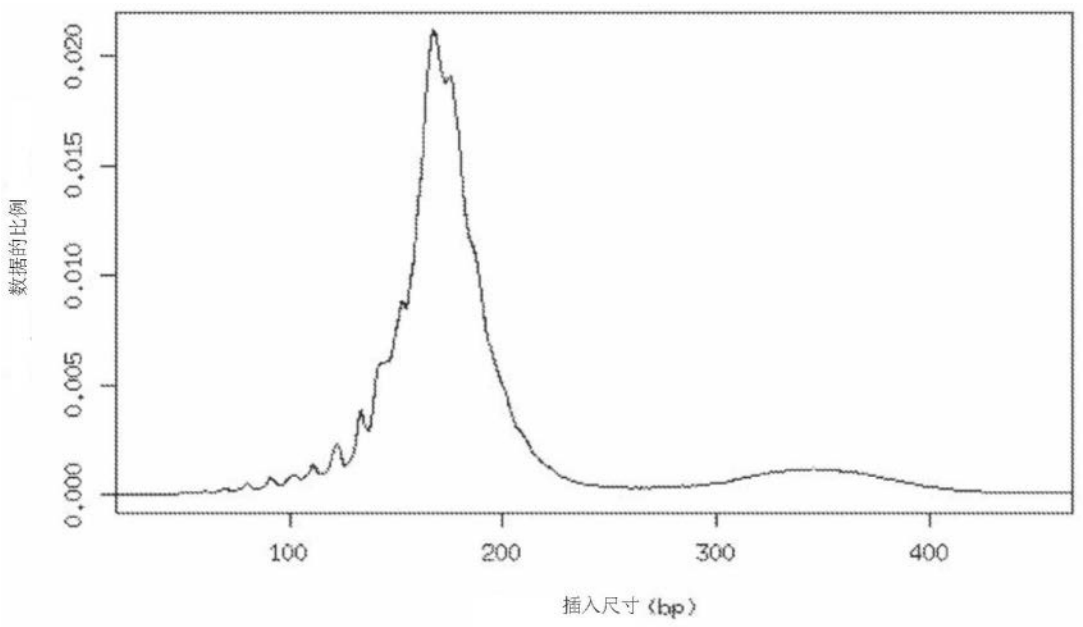


图2

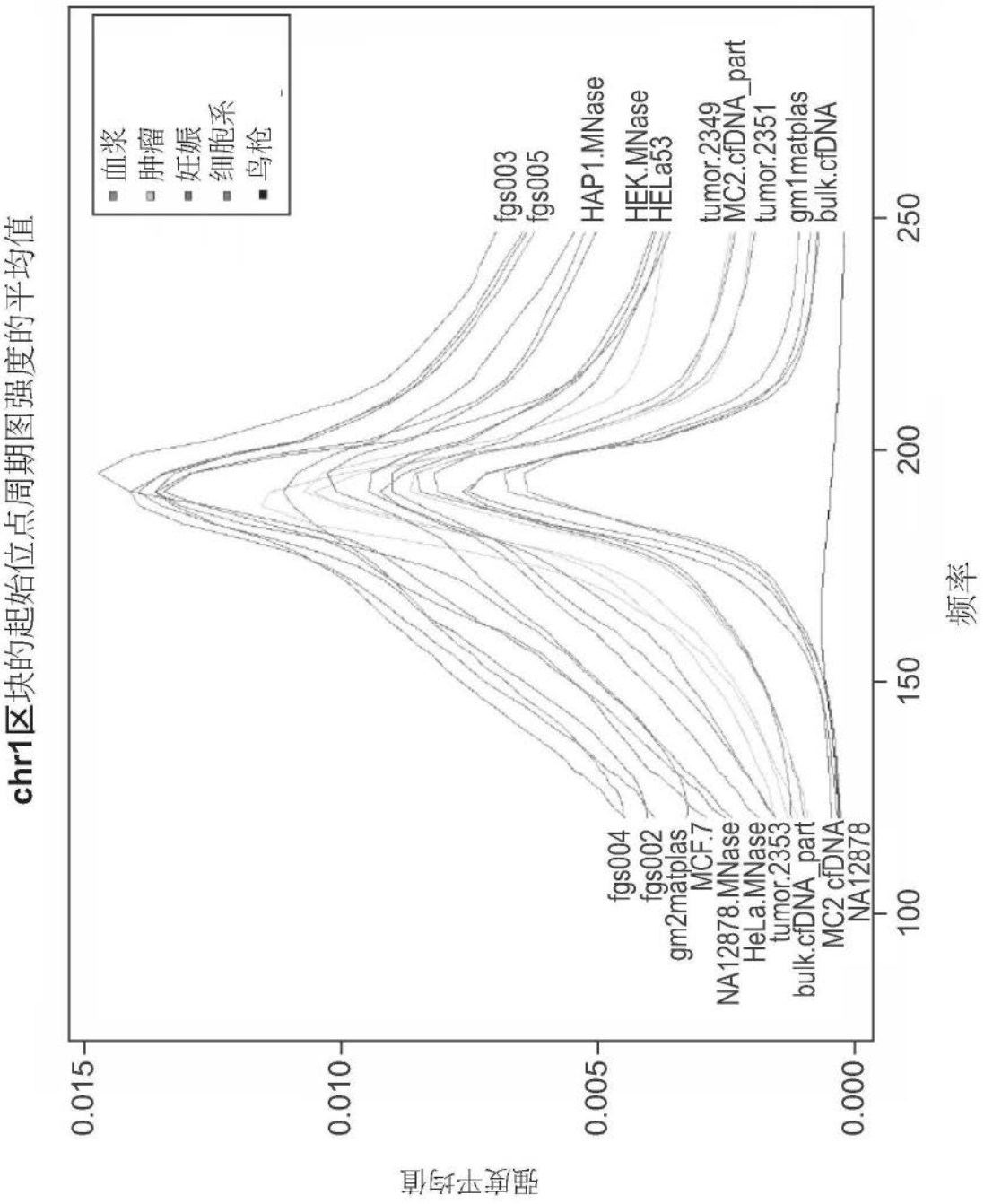

chr1区块的起始位点周期图强度的平均值
血浆
肿瘤
妊娠
细胞系
鸟枪
fgs003
fgs005
HAP1.MNase
HEK.MNase
HELa53
tumor.2349
MC2.cfDNA_part
tumor.2351
gm1matplas
bulk.cfDNA
fgs004
fgs002
gm2matplas
MCF.7
NA12878.MNase
HeLa.MNase
tumor.2353
bulk.cfDNA_part
MC2 cfDNA
NA12878
0.015
0.010
0.005
0.000
强度平均值
100
150
200
250
频率

图3A

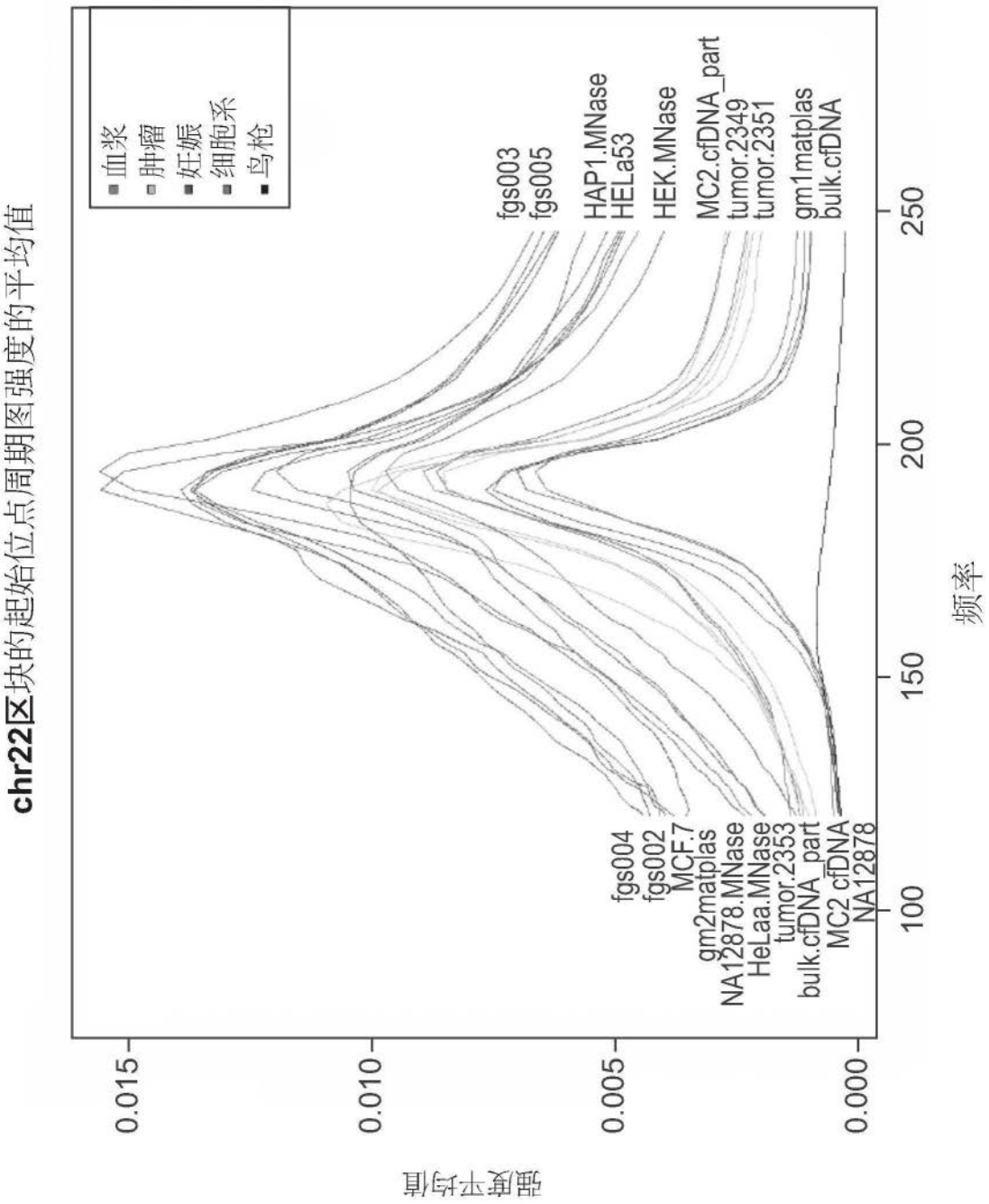

chr22区块的起始位点周期图强度的平均值
血浆
肿瘤
妊娠
细胞系
鸟枪
fgs003
fgs005
HAP1.MNase
HELa53
HEK.MNase
MC2.cfDNA_part
tumor.2349
tumor.2351
gm1matplas
bulk.cfDNA
fgs004
fgs002
MCF.7
gm2matplas
NA12878.MNase
HeLaa.MNase
tumor.2353
bulk.cfDNA_part
MC2 cfDNA
NA12878
0.015
0.010
0.005
0.000
强度平均值
100
150
200
250
频率

图3B

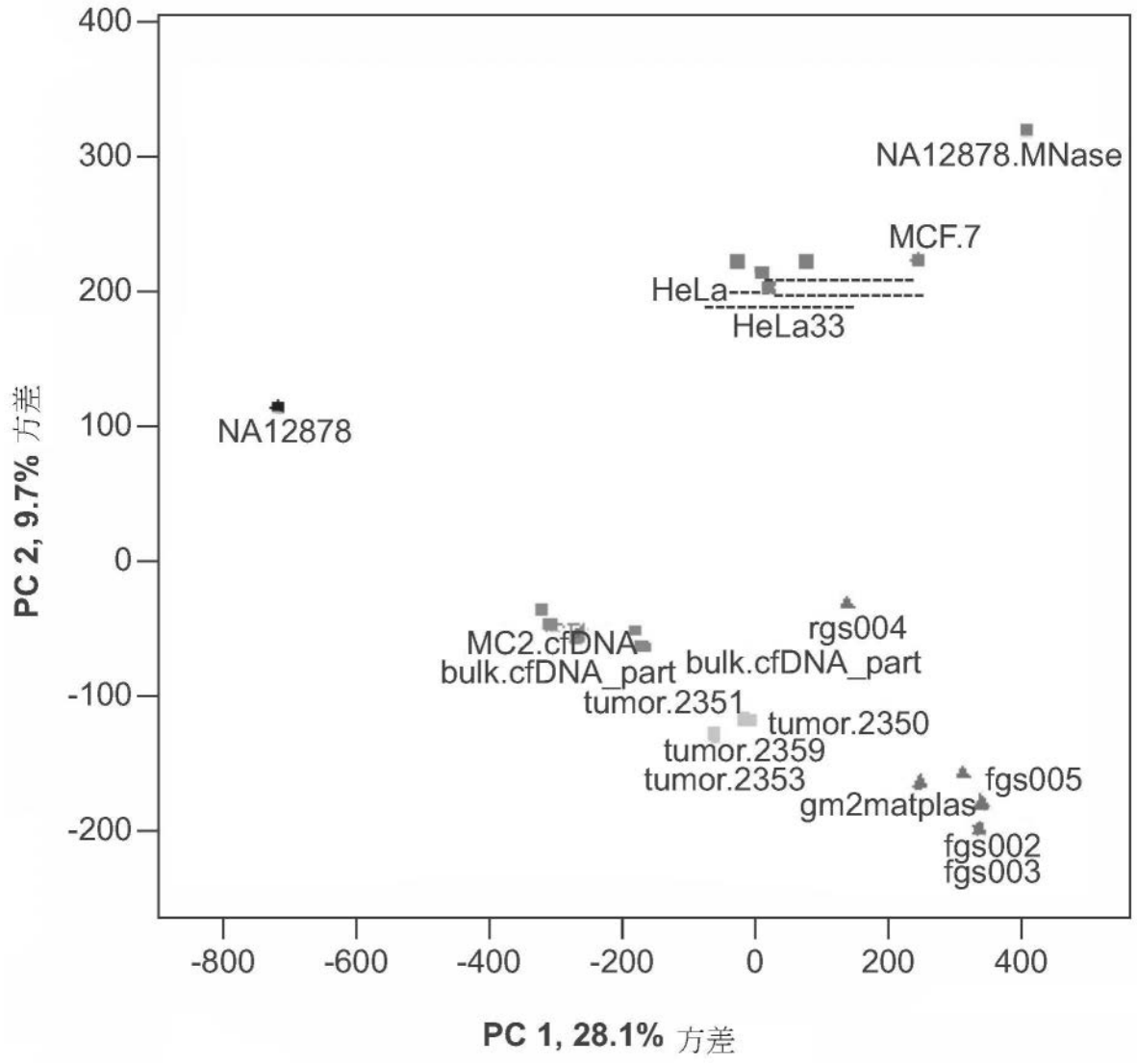

400
300
200
100
0
-100
-200
-800
-600
-400
-200
0
200
400
PC 2, 9.7% 方差
PC 1, 28.1% 方差
NA12878.MNase
MCF.7
HeLa
HeLa33
NA12878
MC2.cfDNA
bulk.cfDNA_part
bulk.cfDNA_part
rgs004
tumor.2351
tumor.2350
tumor.2359
tumor.2353
fgs005
gm2matplas
fgs002
fgs003

图4A

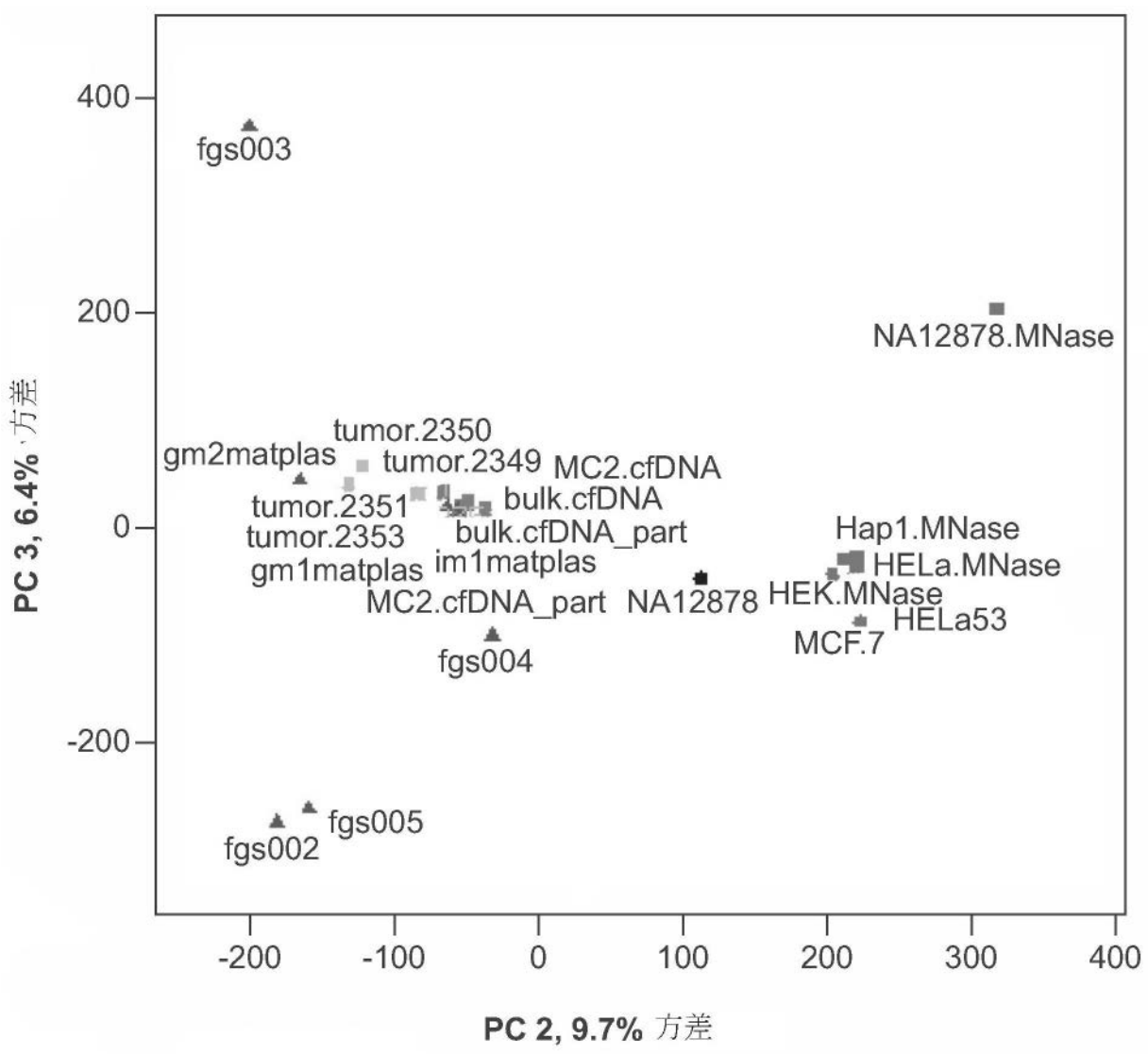

400
200
0
-200
PC 3, 6.4% 方差
fgs003
NA12878.MNase
tumor.2350
gm2matplas
tumor.2349
MC2.cfDNA
tumor.2351
bulk.cfDNA
tumor.2353
bulk.cfDNA_part
Hap1.MNase
gm1matplas
im1matplas
HELa.MNase
MC2.cfDNA_part
NA12878
HEK.MNase
HELa53
MCF.7
fgs004
fgs005
fgs002
-200
-100
0
100
200
300
400
PC 2, 9.7% 方差

图4B

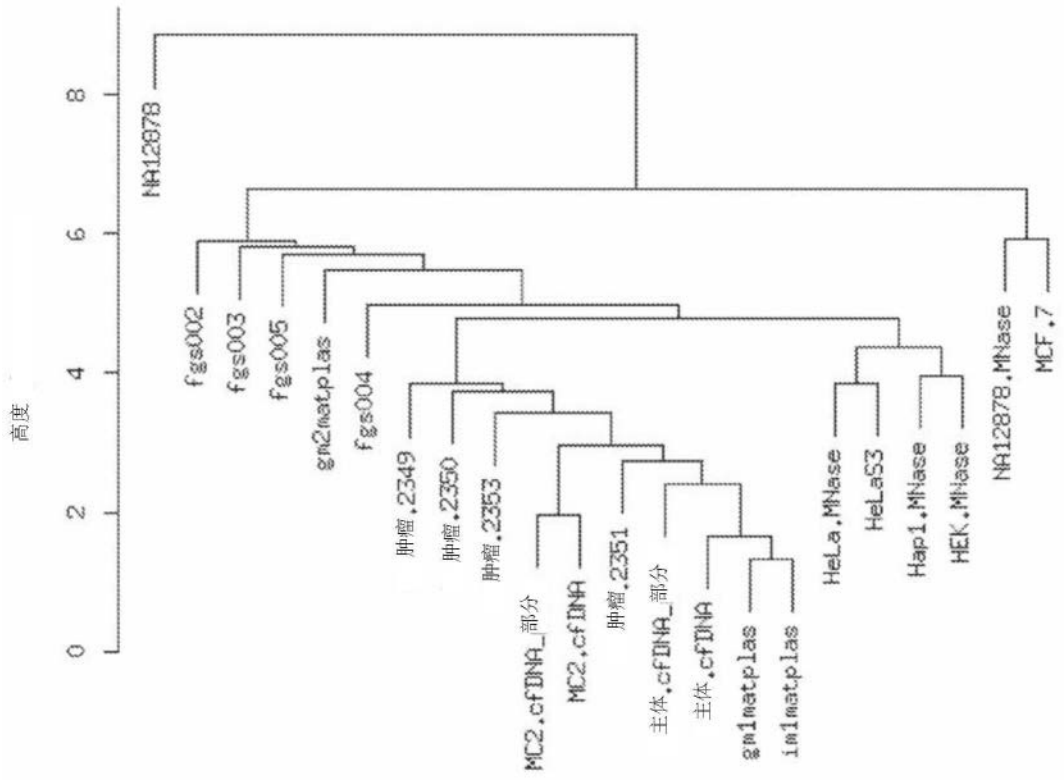

高度
0
2
4
6
8
NA12878
fgs002
fgs003
fgs005
gm2matplas
fgs004
肿瘤.2349
肿瘤.2350
肿瘤.2353
MC2.cfDNA_部分
MC2.cfDNA
肿瘤.2351
主体.cfDNA_部分
主体.cfDNA
gm1matplas
im1matplas
HeLa.MNase
HeLaS3
Hap1.MNase
HEK.MNase
NA12878.MNase
MCF.7

图5

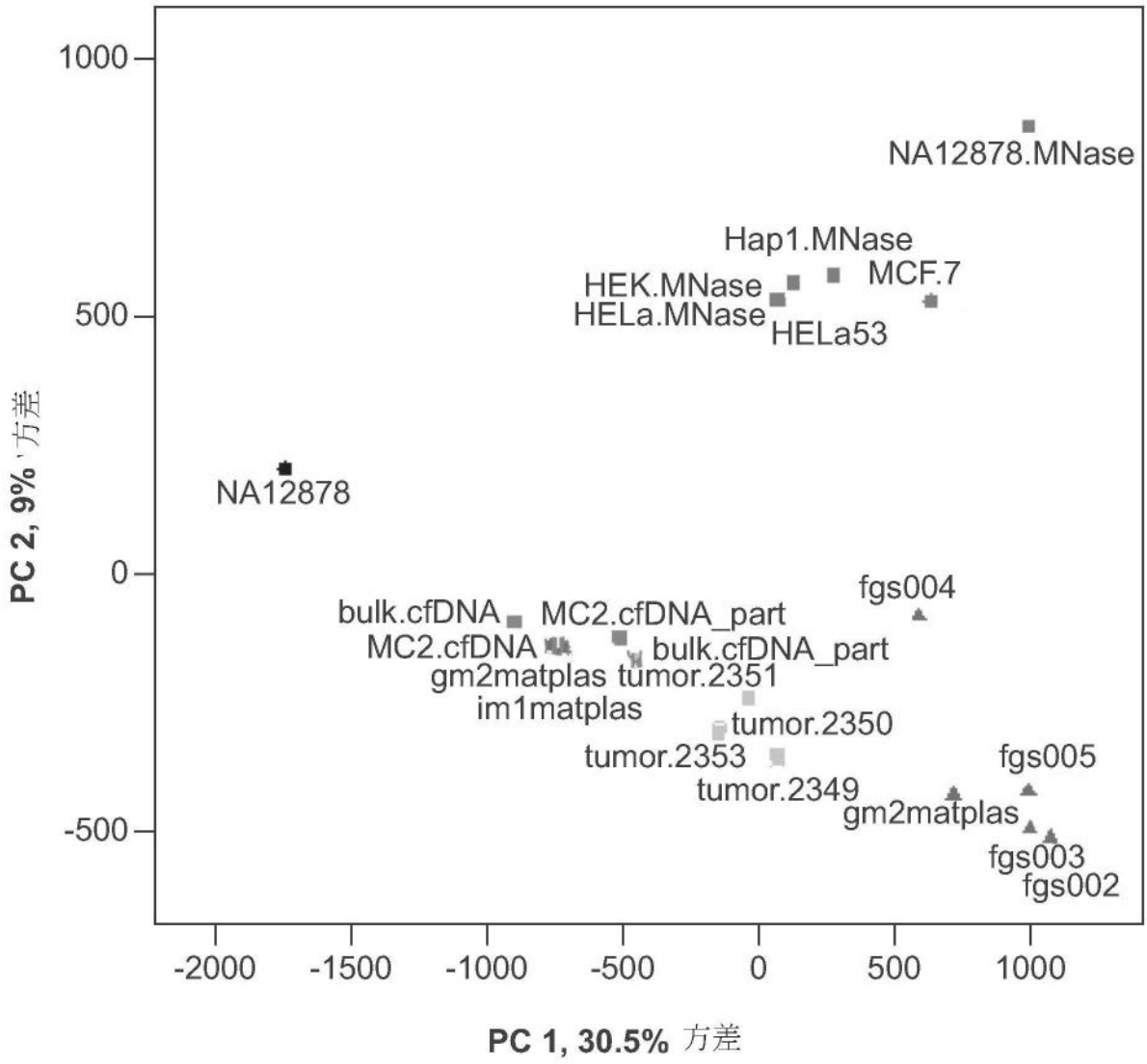

1000
500
0
-500
PC 2, 9% 方差
NA12878.MNase
Hap1.MNase
MCF.7
HEK.MNase
HELa.MNase
HELa53
NA12878
fgs004
bulk.cfDNA
MC2.cfDNA_part
MC2.cfDNA
bulk.cfDNA_part
gm2matplas
tumor.2351
im1matplas
tumor.2350
tumor.2353
fgs005
tumor.2349
gm2matplas
fgs003
fgs002
-2000
-1500
-1000
-500
0
500
1000
PC 1, 30.5% 方差

图6A

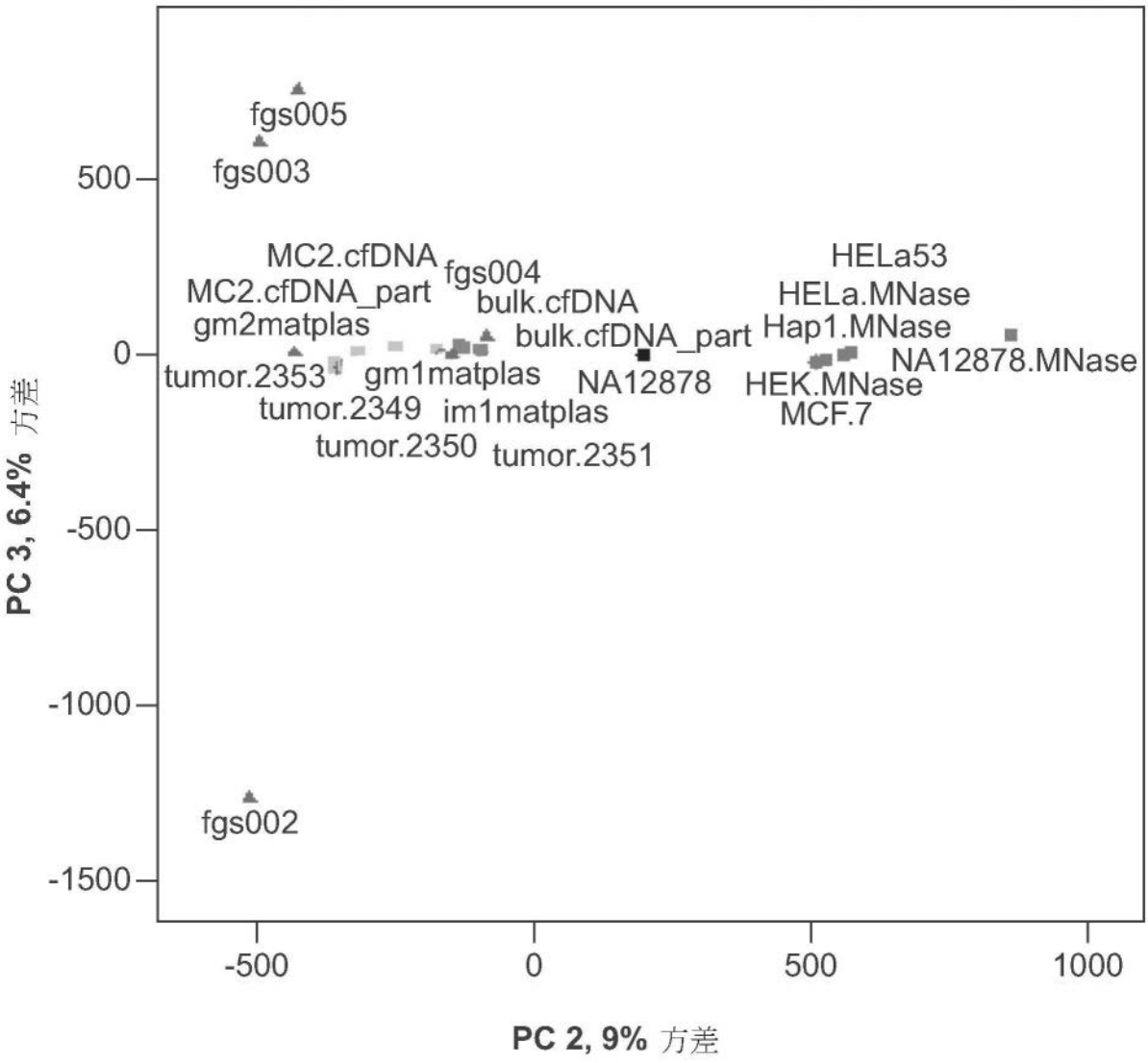

fgs005
fgs003
MC2.cfDNA
MC2.cfDNA_part
gm2matplas
tumor.2353
tumor.2349
tumor.2350
fgs004
bulk.cfDNA
bulk.cfDNA_part
gm1matplas
im1matplas
tumor.2351
NA12878
HELa53
HELa.MNase
Hap1.MNase
NA12878.MNase
HEK.MNase
MCF.7
fgs002
500
0
-500
-1000
-1500
-500
0
500
1000
PC 3, 6.4% 方差
PC 2, 9% 方差

图6B

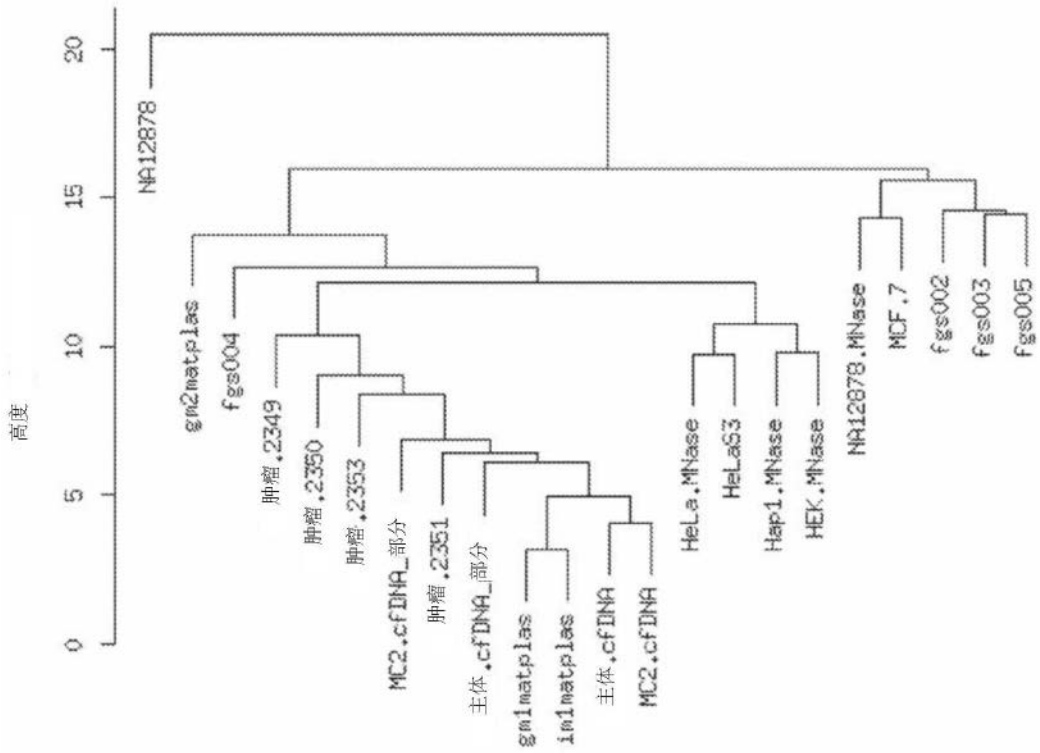

高度
0
5
10
15
20
NA12878
gm2matplas
fgs004
肿瘤.2349
肿瘤.2350
肿瘤.2353
MC2.cfDNA_部分
肿瘤.2351
主体.cfDNA_部分
gm1matplas
im1matplas
主体.cfDNA
MC2.cfDNA
HeLa.MNase
HeLaS3
Hap1.MNase
HEK.MNase
NA12878.MNase
MCF.7
fgs002
fgs003
fgs005

图7

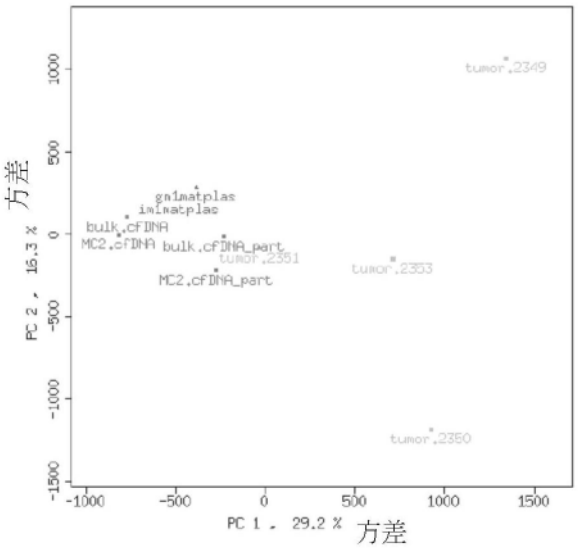

方差
PC 2 , 16.3 %
1000
500
0
-500
-1000
-1500
gm1matplas
im1matplas
bulk.cfDNA
MC2.cfDNA
bulk.cfDNA_part
tumor.2351
MC2.cfDNA_part
tumor.2349
tumor.2353
tumor.2350
-1000
-500
0
500
1000
1500
PC 1 , 29.2 %
方差

图8A

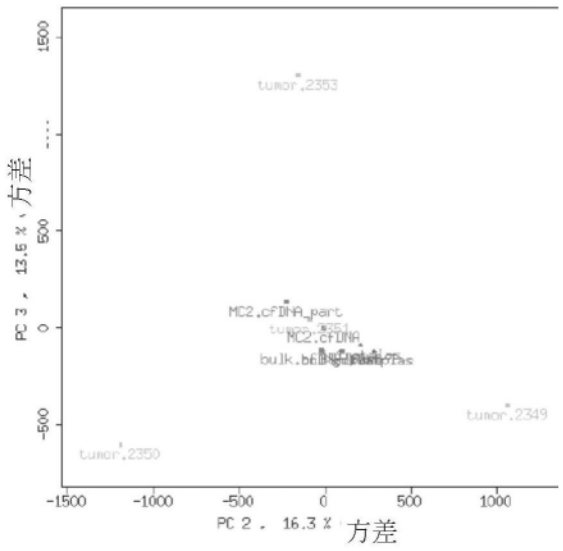

tumor.2353
MC2.cfDNA_part
MC2.cfDNA
tumor.2349
tumor.2350
PC 3 , 13.6 % 方差
PC 2 , 16.3 % 方差

图8B

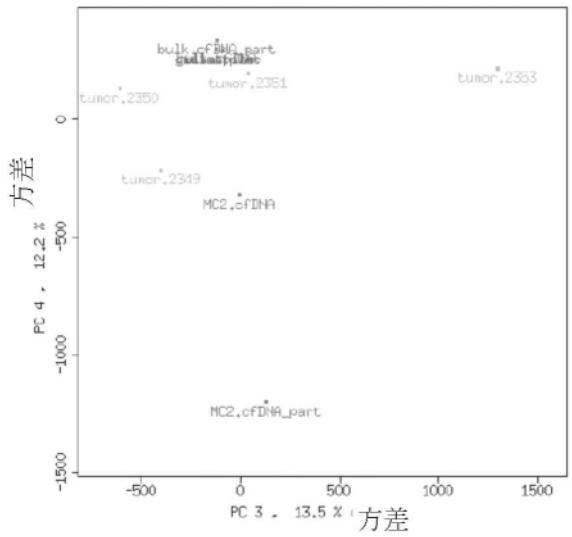

tumor.2351
tumor.2353
tumor.2350
tumor.2349
MC2.cfDNA
MC2.cfDNA_part
PC 4 , 12.2 % 方差
PC 3 , 13.5 % 方差

图8C

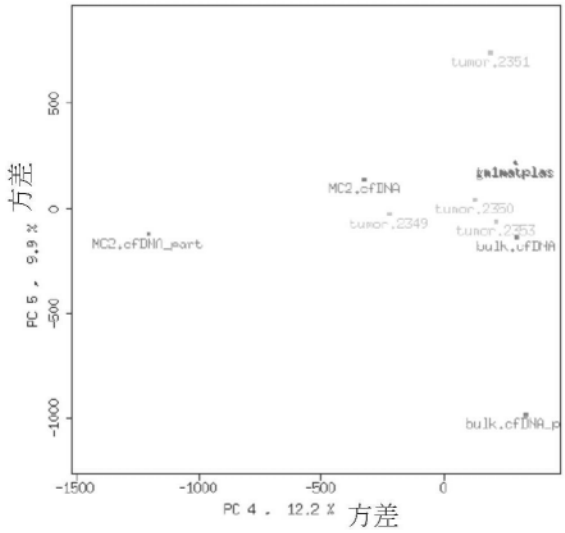

tumor.2351
gnlmatplas
MC2.cfDNA
tumor.2350
tumor.2349
tumor.2353
MC2.cfDNA_part
bulk.cfDNA
bulk.cfDNA_p
500
0
-500
-1000
PC 5 , 9.9 % 方差
-1500
-1000
-500
0
PC 4 , 12.2 % 方差

图8D

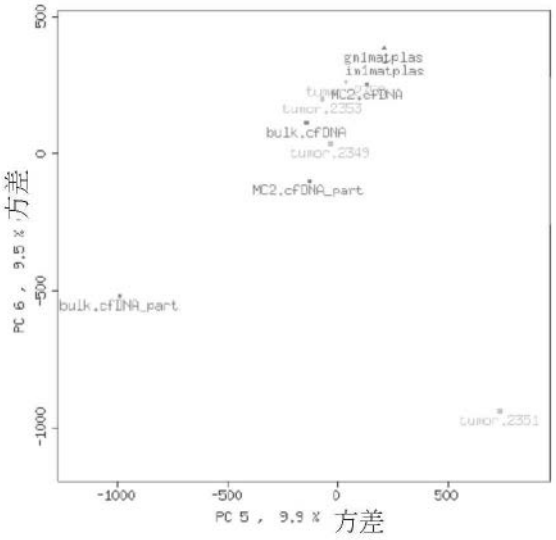

gnlmatplas
inlmatplas
MC2.cfDNA
tumor.2353
bulk.cfDNA
tumor.2349
MC2.cfDNA_part
bulk.cfDNA_part
tumor.2351
500
0
-500
-1000
PC 6 , 9.5 % 方差
-1000
-500
0
500
PC 5 , 9.9 % 方差

图8E

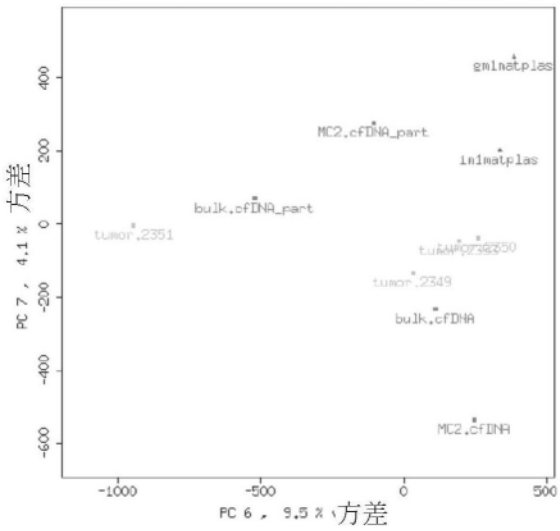

gmlmatplas
MC2.cfDNA_part
inlmatplas
bulk.cfDNA_part
tumor.2351
tumor.2349
bulk.cfDNA
MC2.cfDNA
400
200
0
-200
-400
-600
-1000
-500
0
500
PC 7 , 4.1 % 方差
PC 6 , 9.5 % 方差

图8F

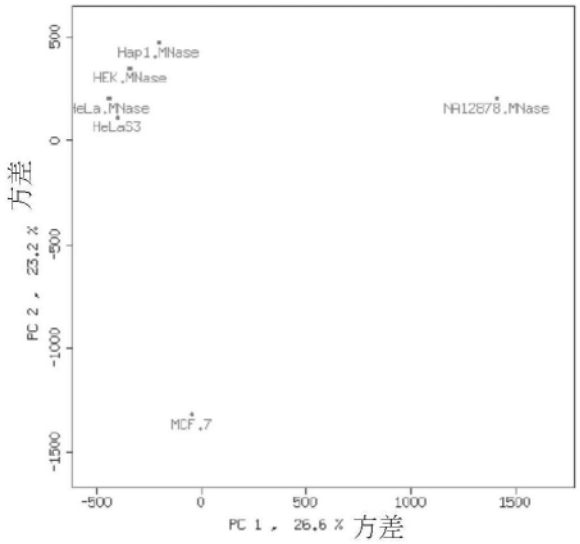

Hap1.MNase
HEK.MNase
HeLa.MNase
HeLaS3
NA12878.MNase
MCF.7
500
0
-500
-1000
-1500
-500
0
500
1000
1500
PC 2 , 23.2 % 方差
PC 1 , 26.6 % 方差

图9A

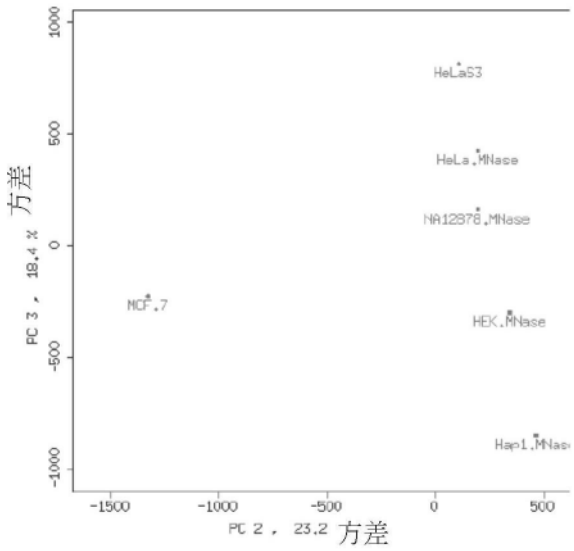

HeLaS3
HeLa.MNase
NA12878.MNase
MCF.7
HEK.MNase
Hap1.MNase
1000
500
0
-500
-1000
-1500
-1000
-500
0
500
PC 3 , 18.4 % 方差
PC 2 , 23.2 方差

图9B

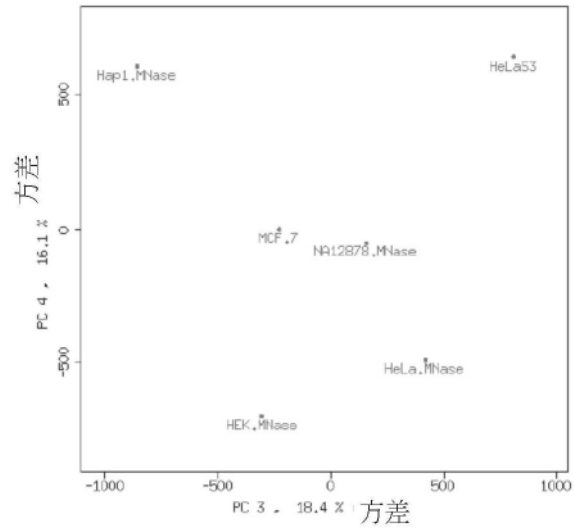

Hap1.MNase
HeLaS3
MCF.7
NA12878.MNase
HeLa.MNase
HEK.MNase
500
0
-500
-1000
-500
0
500
1000
PC 4 , 16.1 % 方差
PC 3 , 18.4 % 方差

图9C

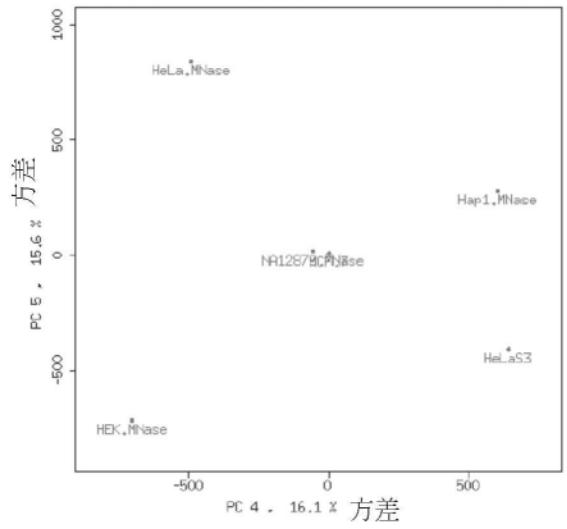

HeLa.MNase
Hap1.MNase
NA12878.MNase
HeLaS3
HEK.MNase
1000
500
0
-500
PC 5 , 15.6 % 方差
-500
0
500
PC 4 , 16.1 % 方差

图9D

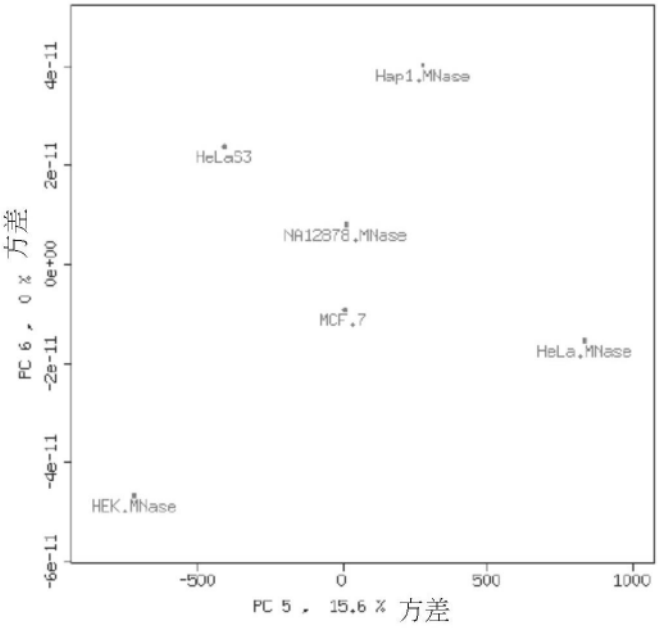

Hap1.MNase
HeLaS3
NA12878.MNase
MCF.7
HeLa.MNase
HEK.MNase
4e-11
2e-11
0e+00
-2e-11
-4e-11
-6e-11
PC 6 , 0 % 方差
-500
0
500
1000
PC 5 , 15.6 % 方差

图9E

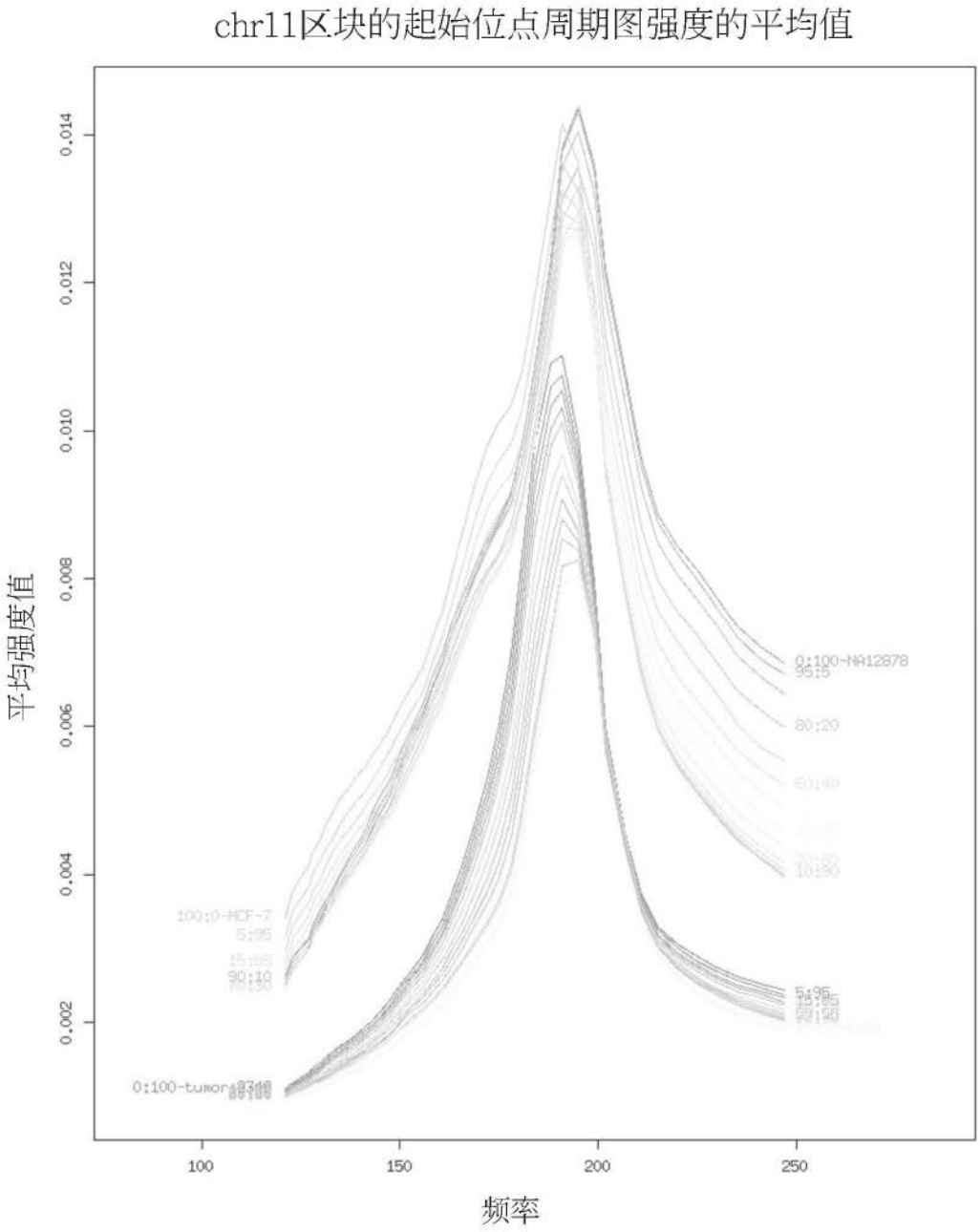

chr11区块的起始位点周期图强度的平均值
平均强度值
0.002
0.004
0.006
0.008
0.010
0.012
0.014
100
150
200
250
频率
0:100-NA12878
95:5
80:20
60:40
10:90
100:0-MCF-7
5:95
90:10
0:100-tumor

图10

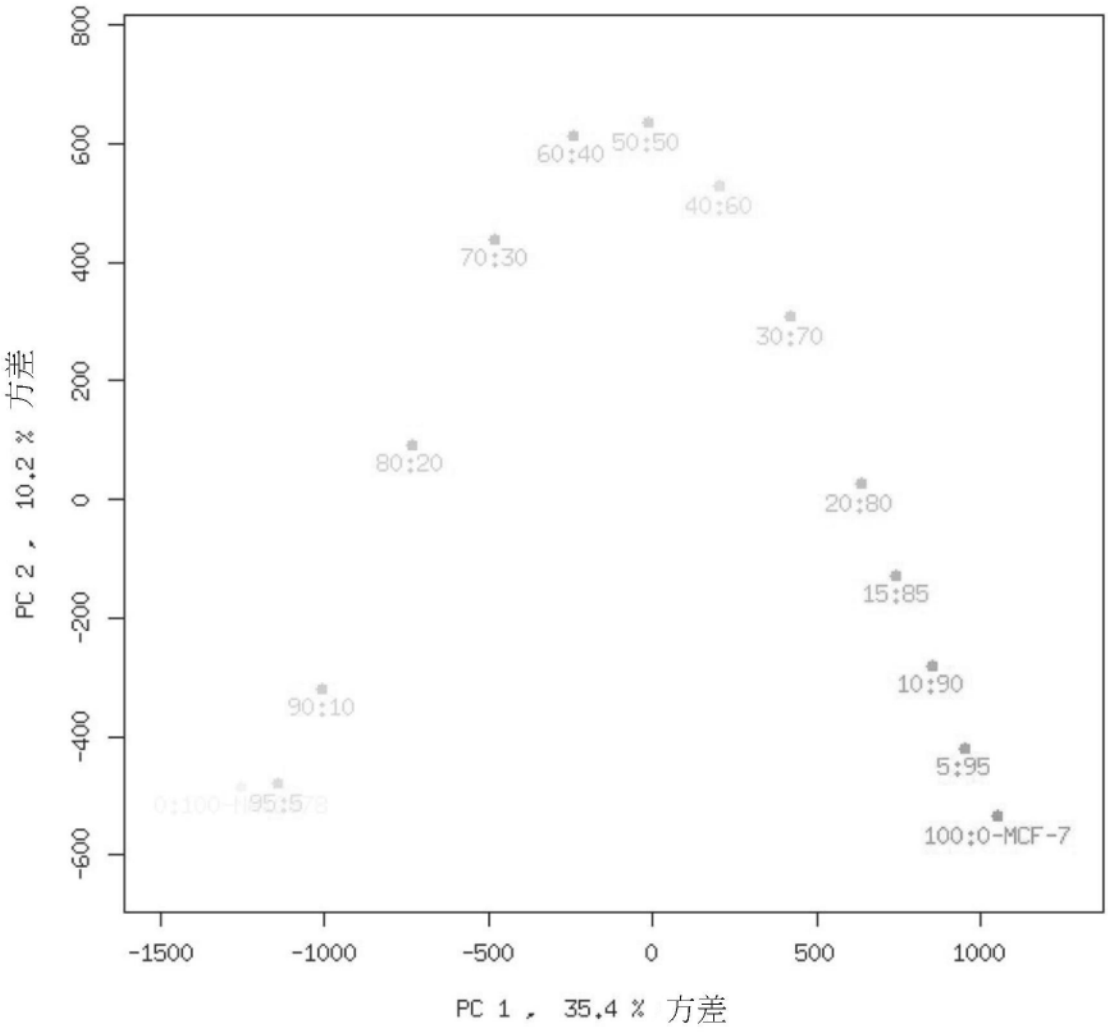

PC 2 , 10.2 % 方差
PC 1 , 35.4 % 方差
800
600
400
200
0
-200
-400
-600
-1500
-1000
-500
0
500
1000
60:40
50:50
40:60
70:30
30:70
80:20
20:80
15:85
10:90
90:10
5:95
100:0-MCF-7

图11

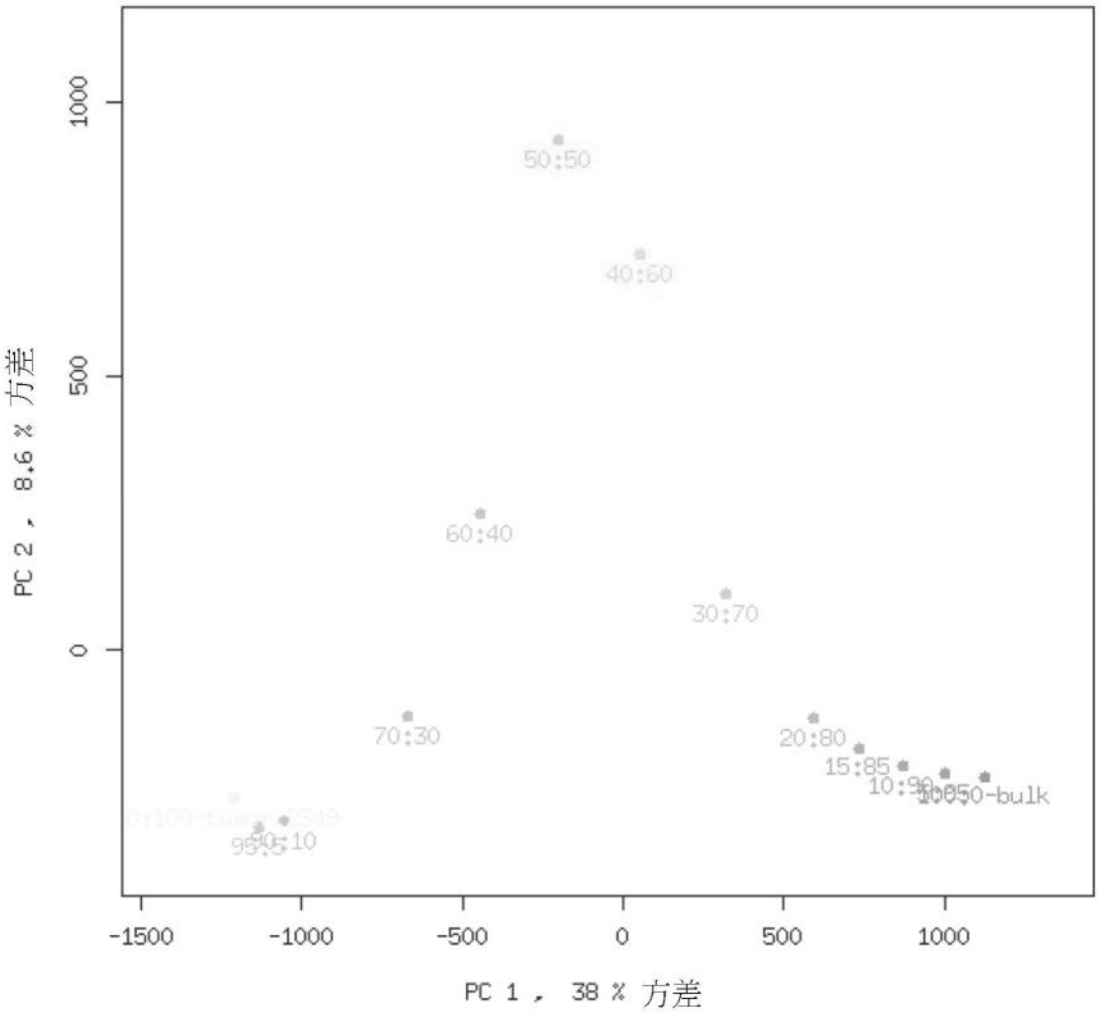

50:50
40:60
60:40
30:70
70:30
20:80
15:85
10:90
bulk
90:10
1000
500
0
-1500
-1000
-500
0
500
1000
PC 2 , 8.6 % 方差
PC 1 , 38 % 方差

图12

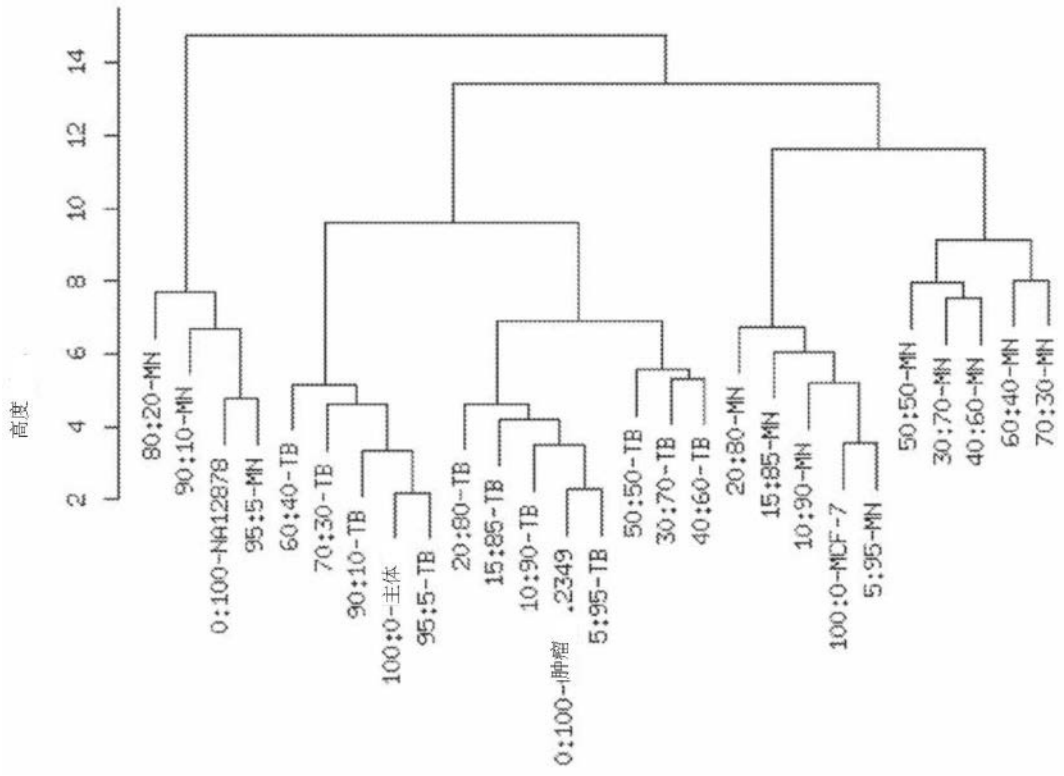

高度
2
4
6
8
10
12
14
80:20-MN
90:10-MN
0:100-HA12878
95:5-MN
60:40-TB
70:30-TB
90:10-TB
100:0-主体
95:5-TB
20:80-TB
15:85-TB
10:90-TB
0:100-肿瘤 .2349
5:95-TB
50:50-TB
30:70-TB
40:60-TB
20:80-MN
15:85-MN
10:90-MN
100:0-MCF-7
5:95-MN
50:50-MN
30:70-MN
40:60-MN
60:40-MN
70:30-MN

图13

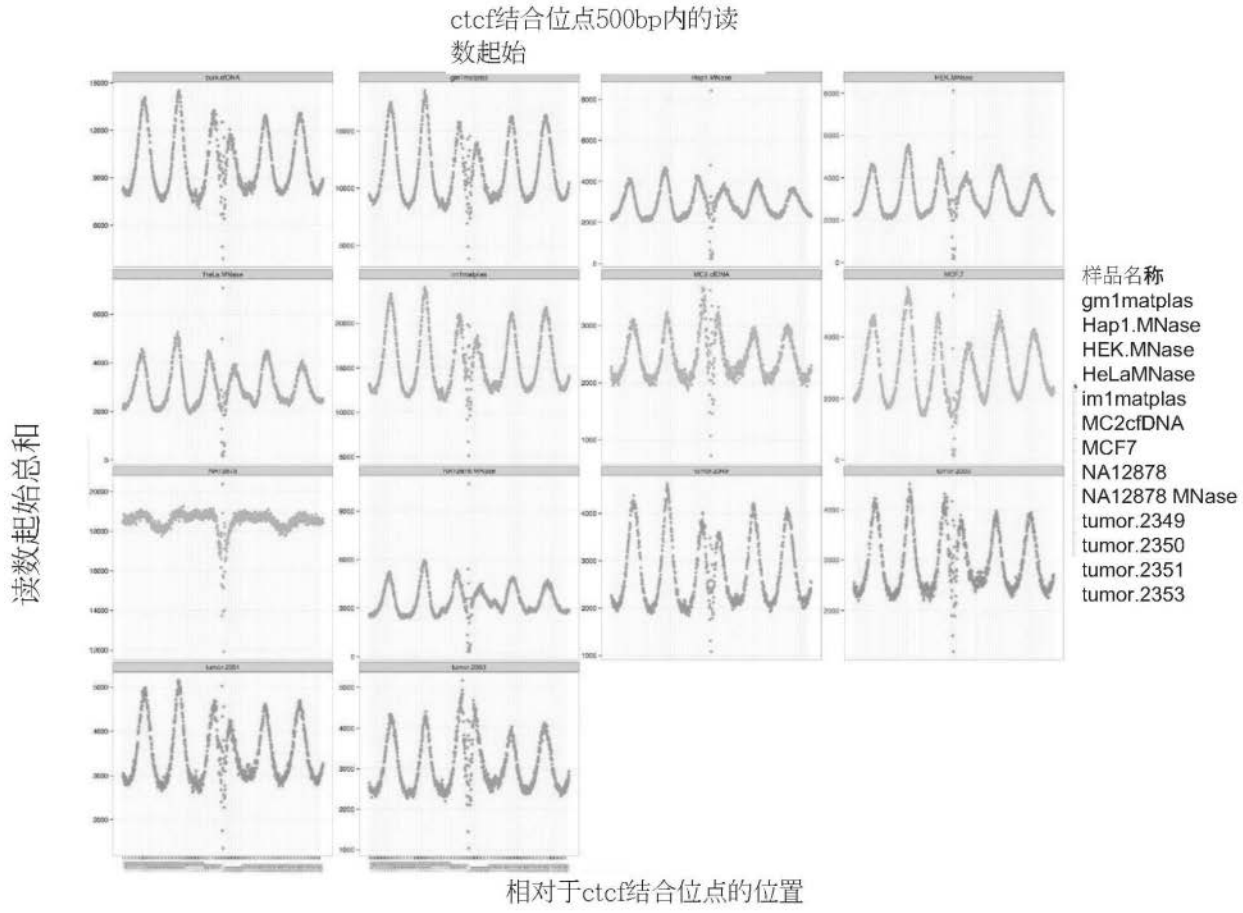

ctcf结合位点500bp内的读数起始
读数起始总和
相对于ctcf结合位点的位置
样品名称
gm1matplas
Hap1.MNase
HEK.MNase
HeLaMNase
im1matplas
MC2cfDNA
MCF7
NA12878
NA12878 MNase
tumor.2349
tumor.2350
tumor.2351
tumor.2353

图14

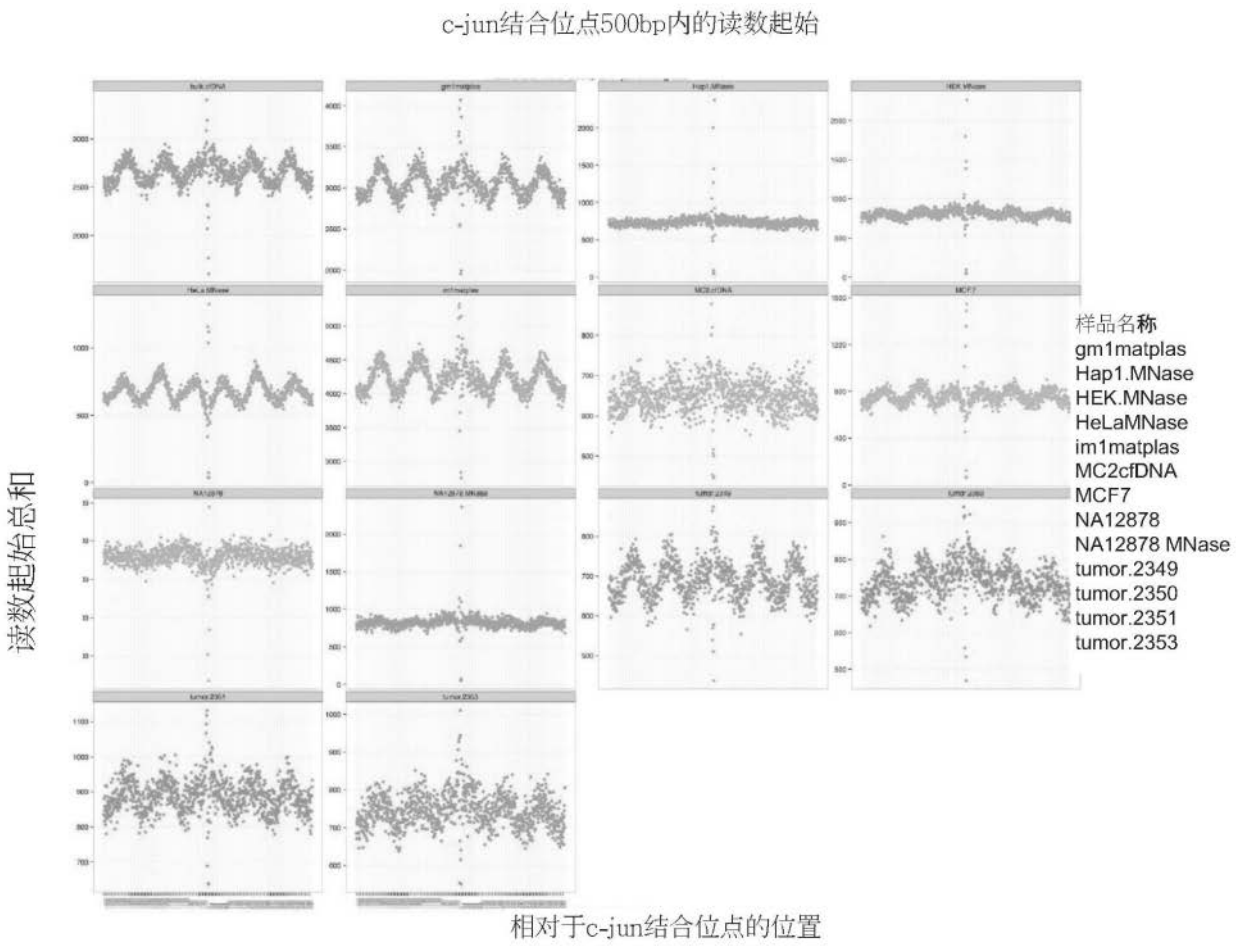

c-jun结合位点500bp内的读数起始
读数起始总和
相对于c-jun结合位点的位置
样品名称
gm1matplas
Hap1.MNase
HEK.MNase
HeLaMNase
im1matplas
MC2cfDNA
MCF7
NA12878
NA12878 MNase
tumor.2349
tumor.2350
tumor.2351
tumor.2353

图15

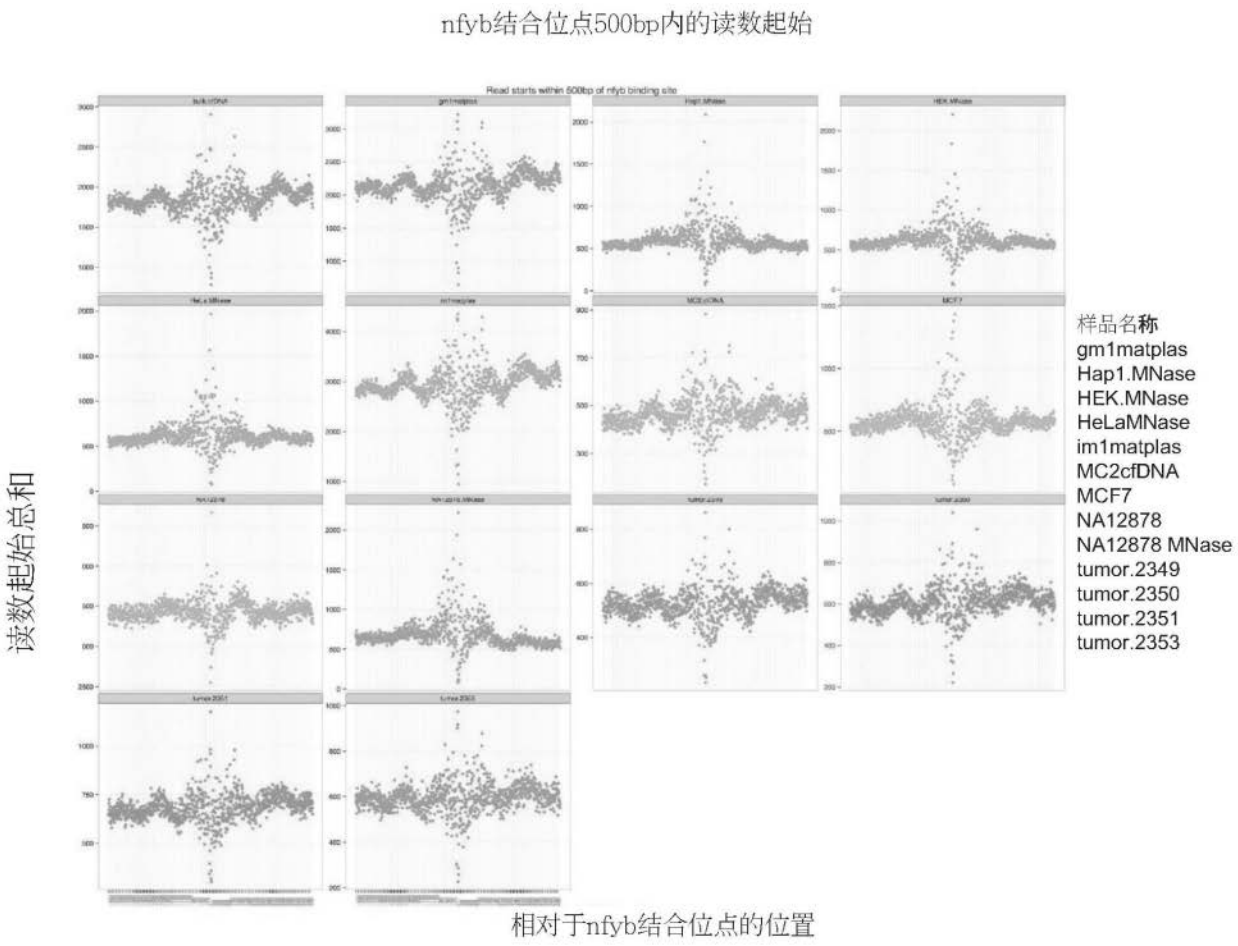

nfyb结合位点500bp内的读数起始
Read starts within 500bp of nfyb binding site
读数起始总和
相对于nfyb结合位点的位置
样品名称
gm1matplas
Hap1.MNase
HEK.MNase
HeLaMNase
im1matplas
MC2cfDNA
MCF7
NA12878
NA12878 MNase
tumor.2349
tumor.2350
tumor.2351
tumor.2353

图16

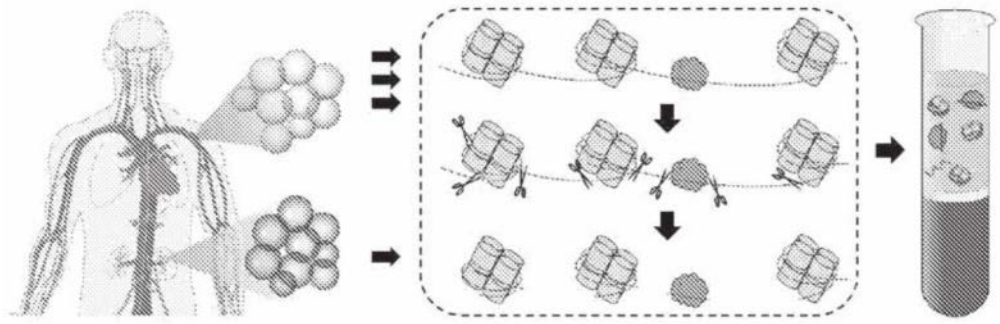


图17

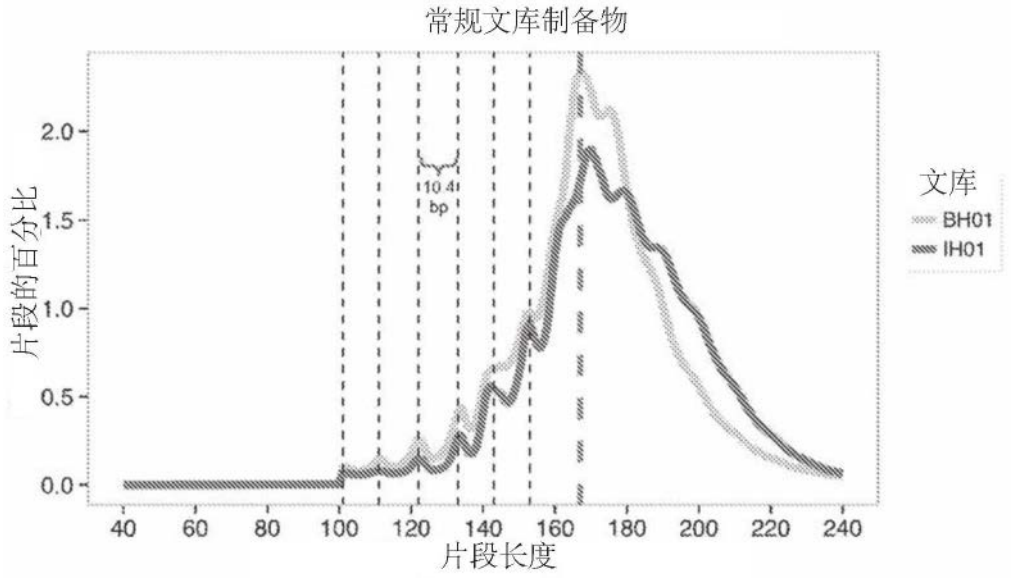

常规文库制备物
片段的百分比
2.0
1.5
1.0
0.5
0.0
10.4
bp
40
60
80
100
120
140
160
180
200
220
240
片段长度
文库
BH01
IH01

图18

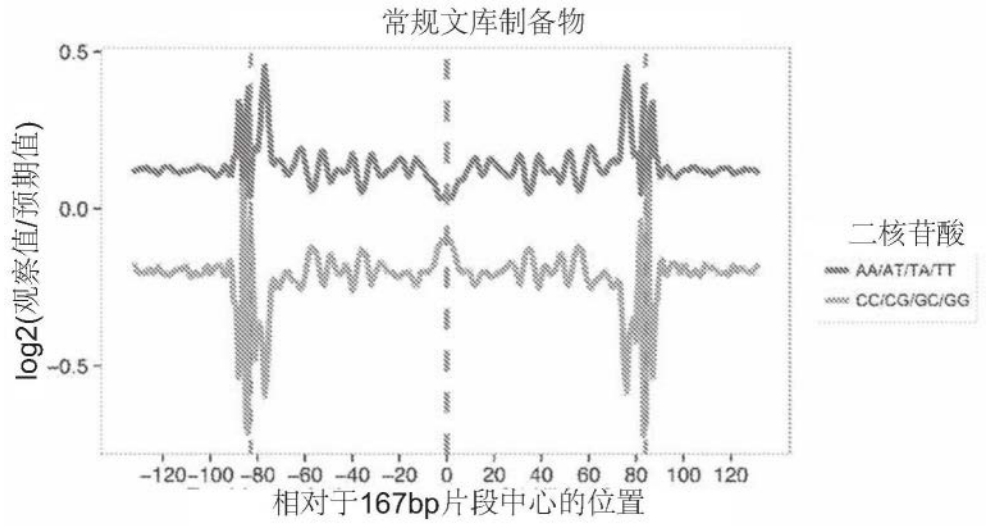

常规文库制备物
log2(观察值/预期值)
0.5
0.0
-0.5
-120
-100
-80
-60
-40
-20
0
20
40
60
80
100
120
相对于167bp片段中心的位置
二核苷酸
AA/AT/TA/TT
CC/CG/GC/GG

图19

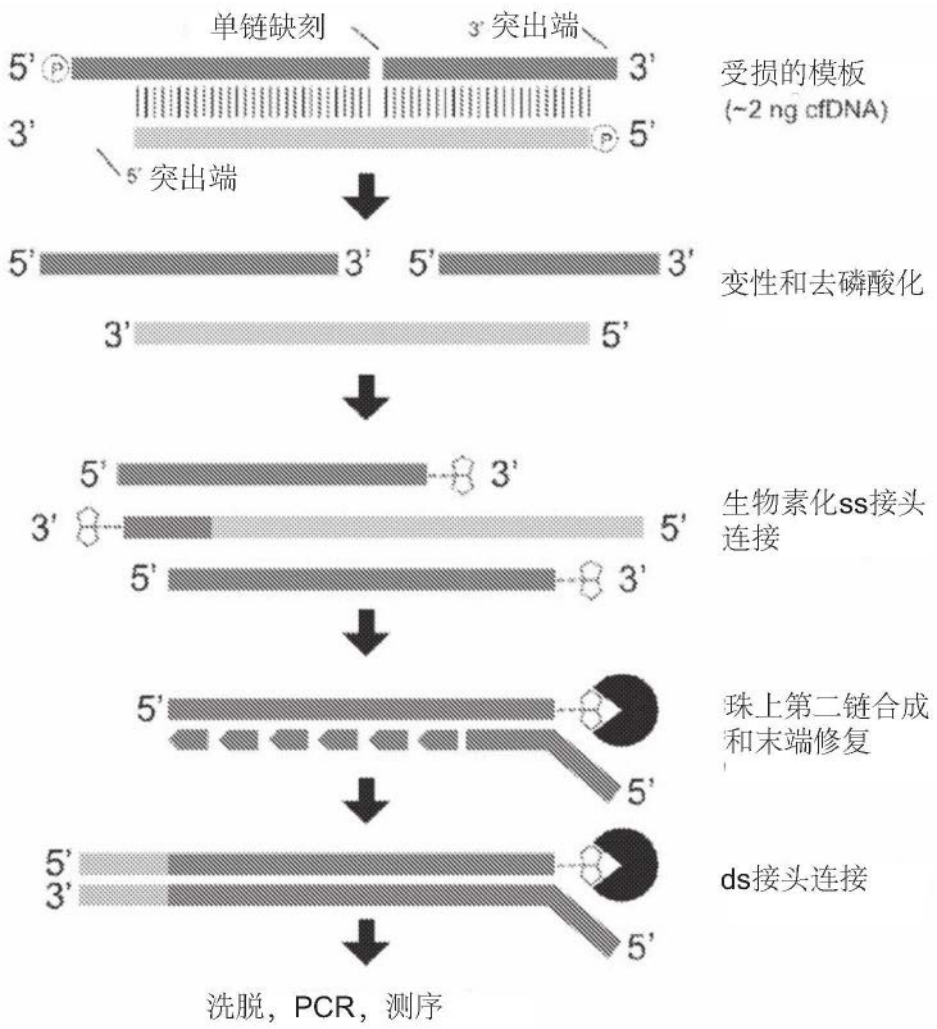

单链缺刻
3' 突出端
5'
3'
受损的模板
(~2 ng cfDNA)
3'
5'
5' 突出端
5'
3'
5'
3'
变性和去磷酸化
3'
5'
5'
3'
3'
5'
5'
3'
生物素化ss接头
连接
5'
3'
5'
珠上第二链合成
和末端修复
5'
5'
3'
ds接头连接
5'
洗脱，PCR，测序

图20

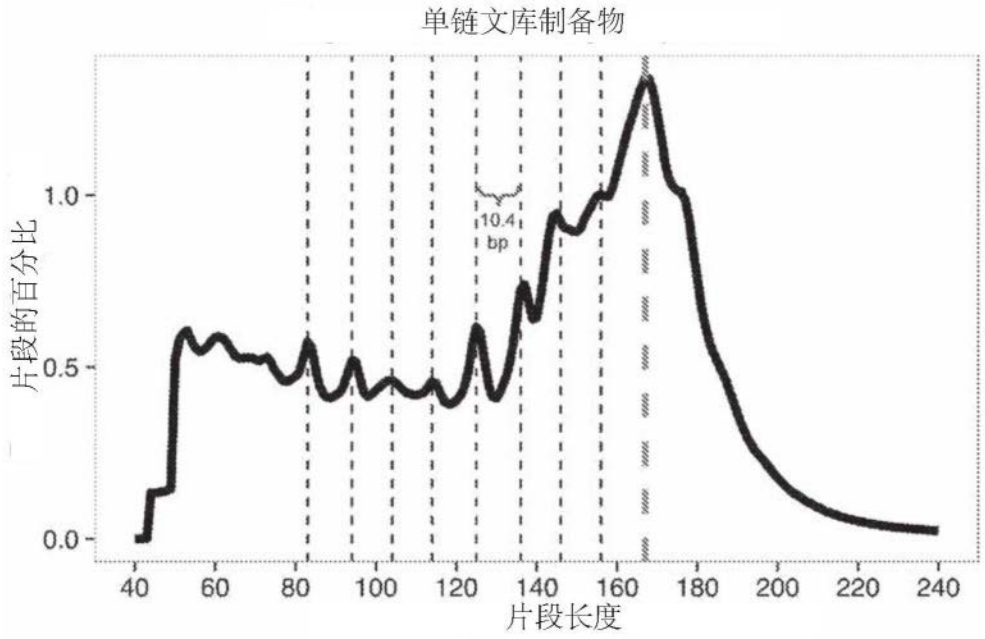

单链文库制备物
片段的百分比
1.0
0.5
0.0
10.4
bp
40
60
80
100
120
140
160
180
200
220
240
片段长度

图21

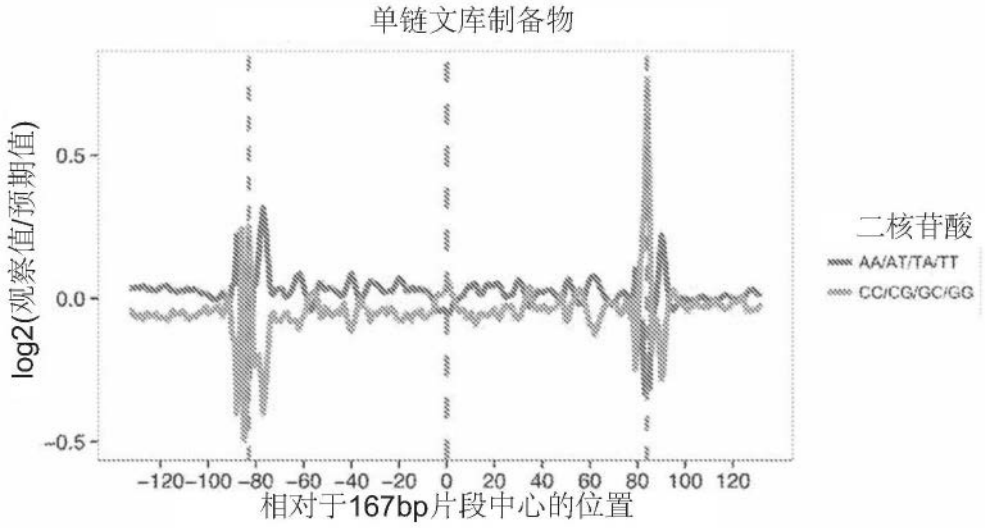

单链文库制备物
log2(观察值/预期值)
0.5
0.0
-0.5
-120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120
相对于167bp片段中心的位置
二核苷酸
AA/AT/TA/TT
CC/CG/GC/GG

图22

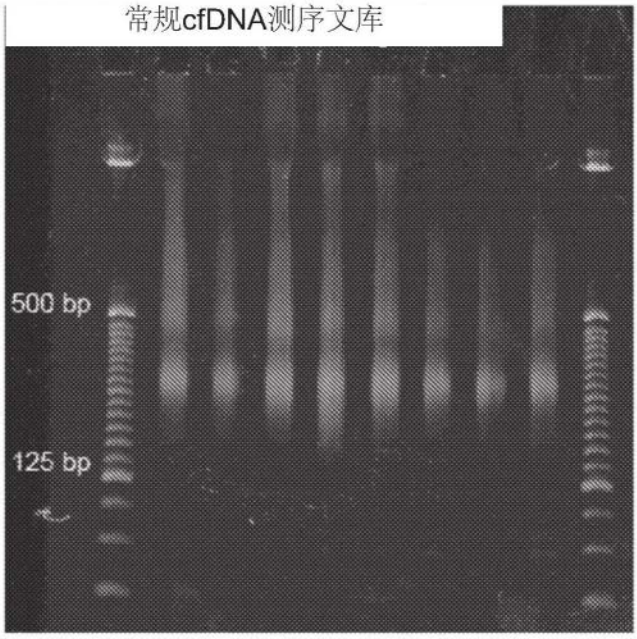


图23A

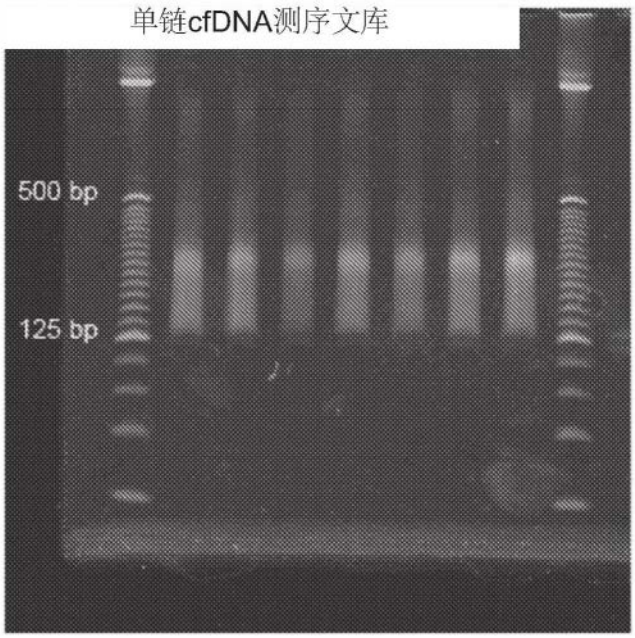


图23B

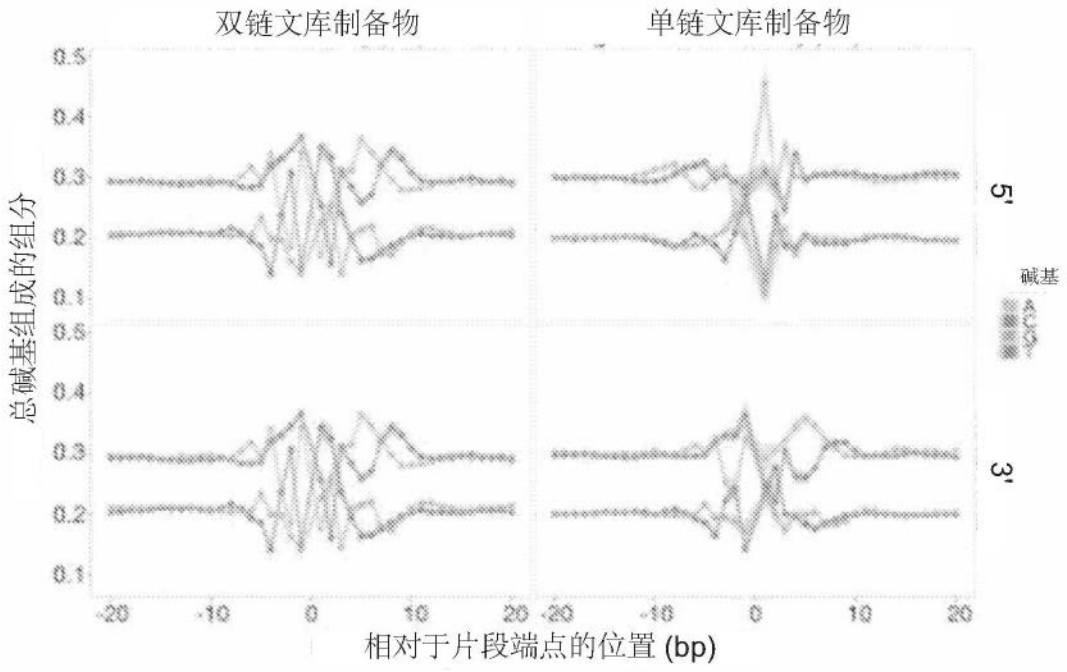

双链文库制备物
单链文库制备物
总碱基组成的组分
0.5
0.4
0.3
0.2
0.1
-20
-10
0
10
20
相对于片段端点的位置 (bp)
5'
3'
碱基

图24A

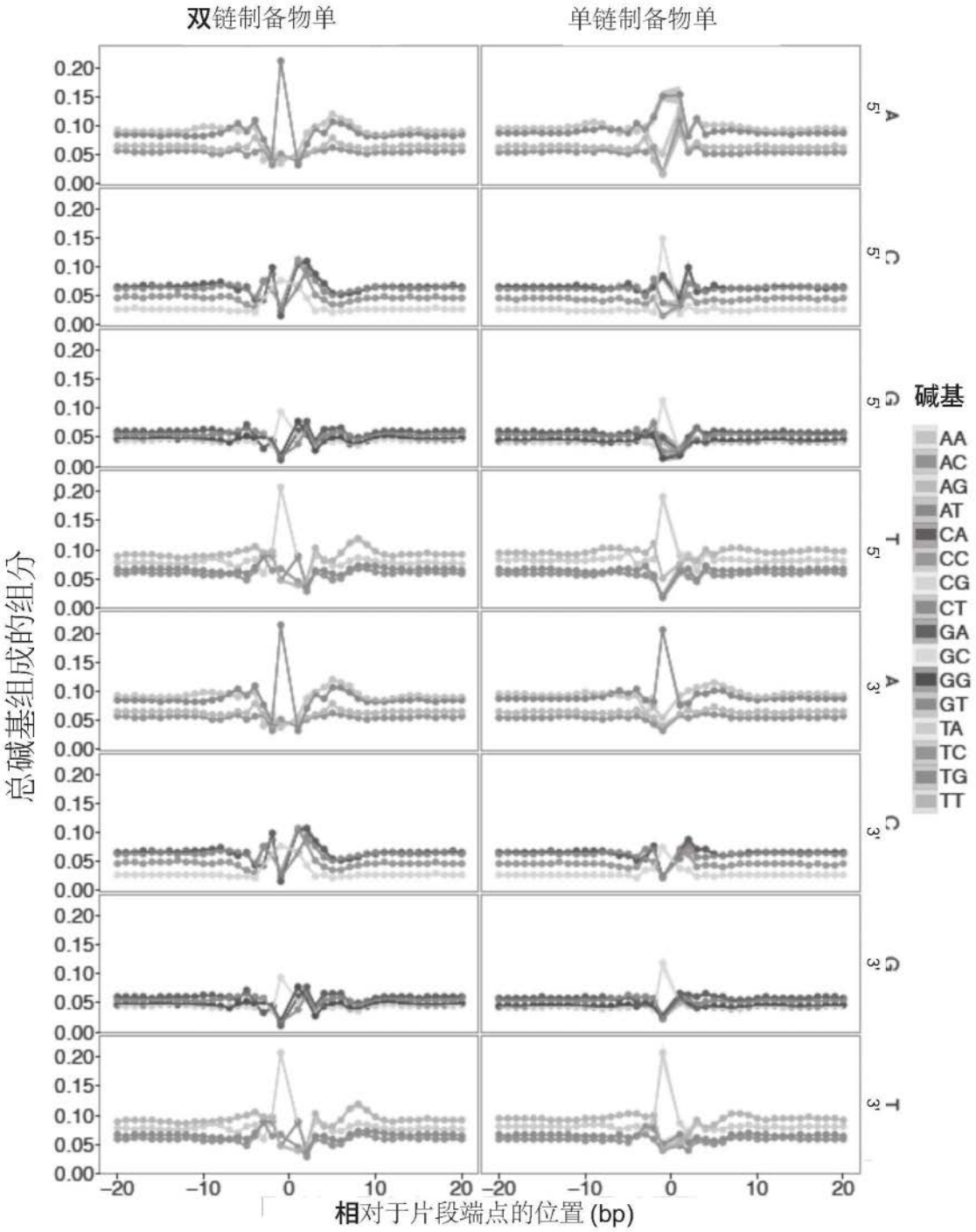

双链制备物单
单链制备物单
总碱基组成的组分
相对于片段端点的位置 (bp)
碱基
AA
AC
AG
AT
CA
CC
CG
CT
GA
GC
GG
GT
TA
TC
TG
TT
A 5'
C 5'
G 5'
T 5'
A 3'
C 3'
G 3'
T 3'

图24B

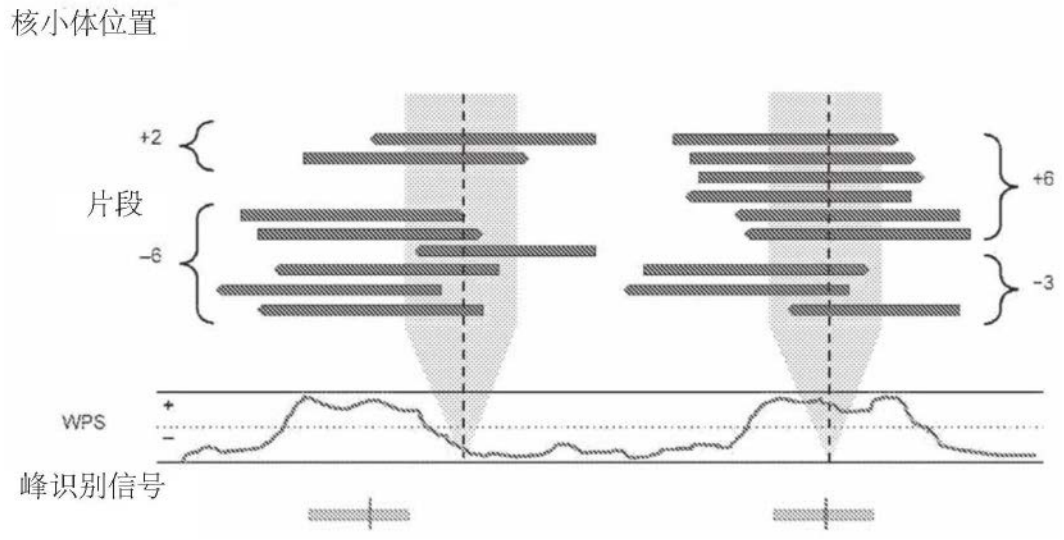

核小体位置
+2
片段
-6
+6
-3
WPS
+
-
峰识别信号

图25

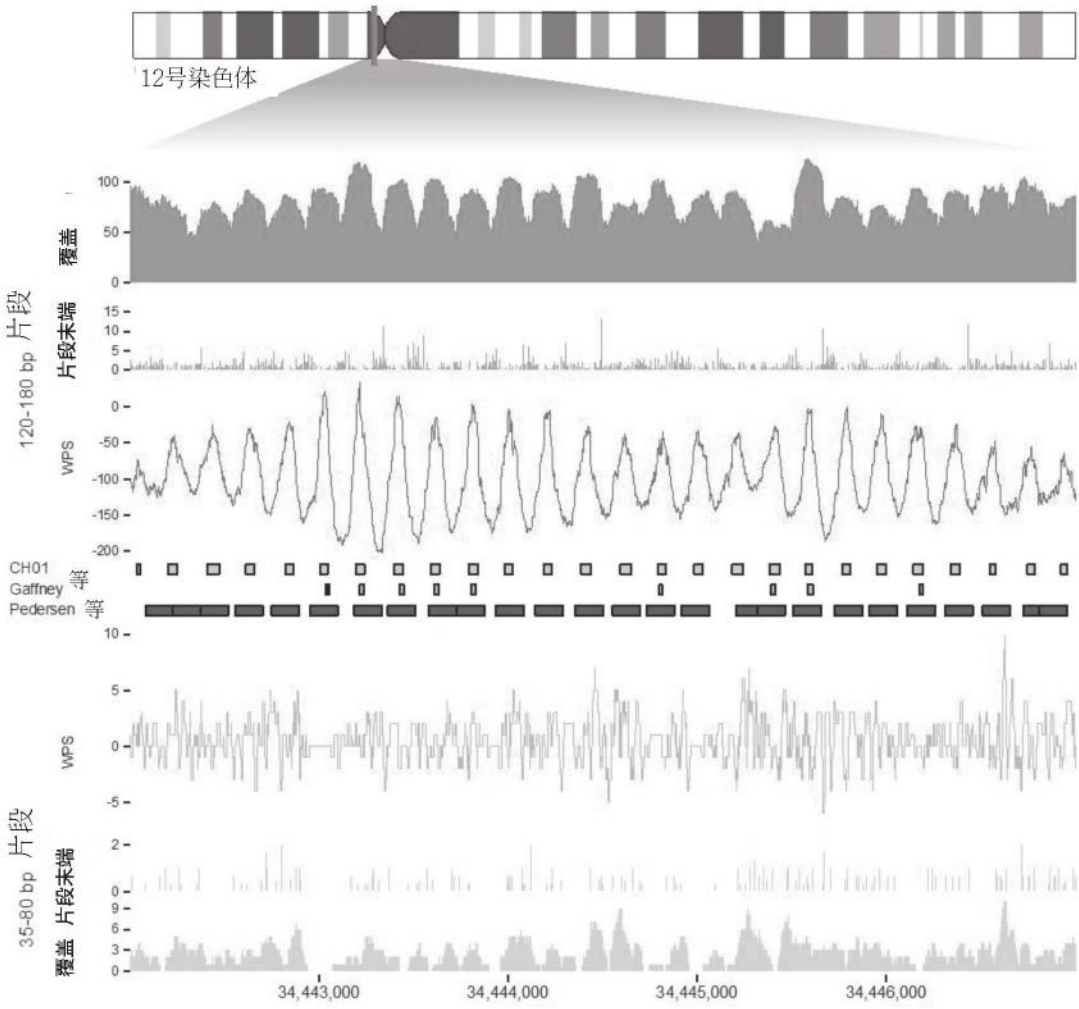

12号染色体
覆盖
片段末端
120-180 bp 片段
WPS
CH01
Gaffney 等
Pedersen 等
WPS
35-80 bp 片段
覆盖 片段末端
34,443,000
34,444,000
34,445,000
34,446,000

图26

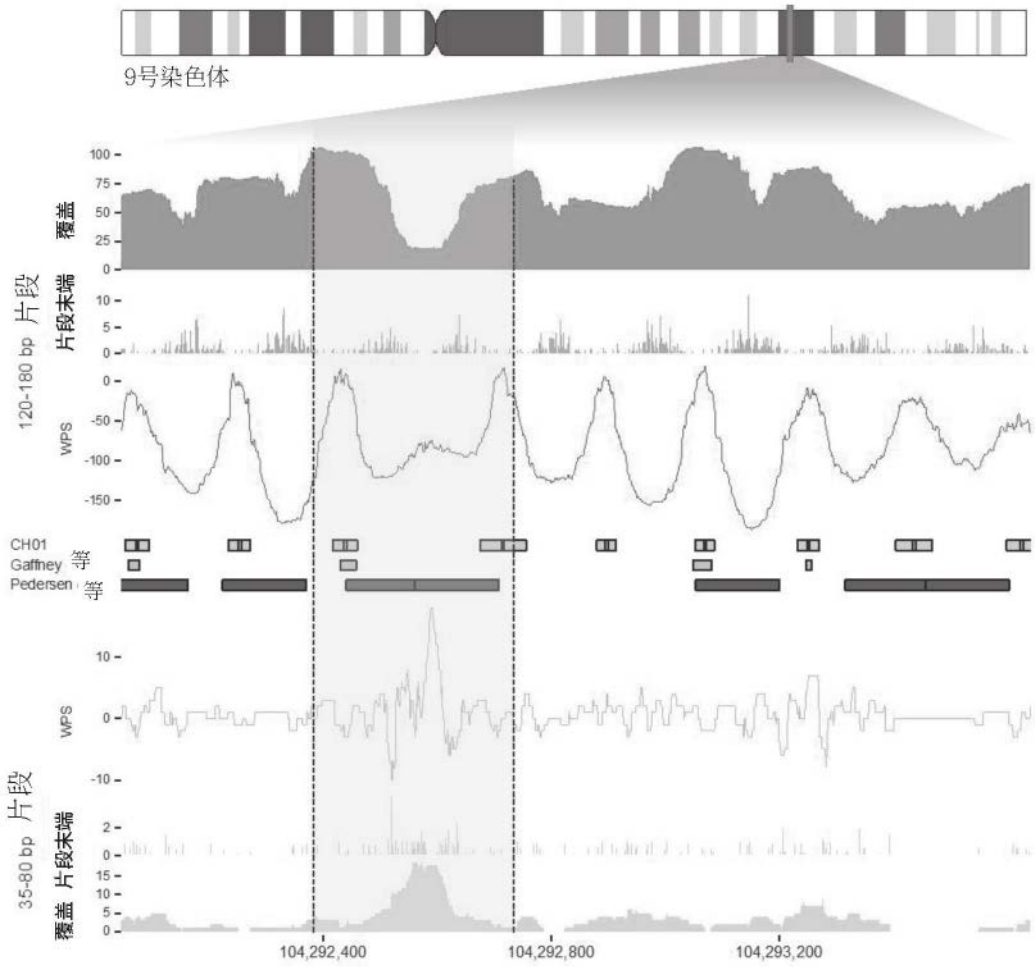

9号染色体
覆盖
片段末端
120-180 bp 片段
WPS
CH01
Gaffney 等
Pedersen 等
WPS
35-80 bp 片段
片段末端
覆盖
104,292,400
104,292,800
104,293,200

图27

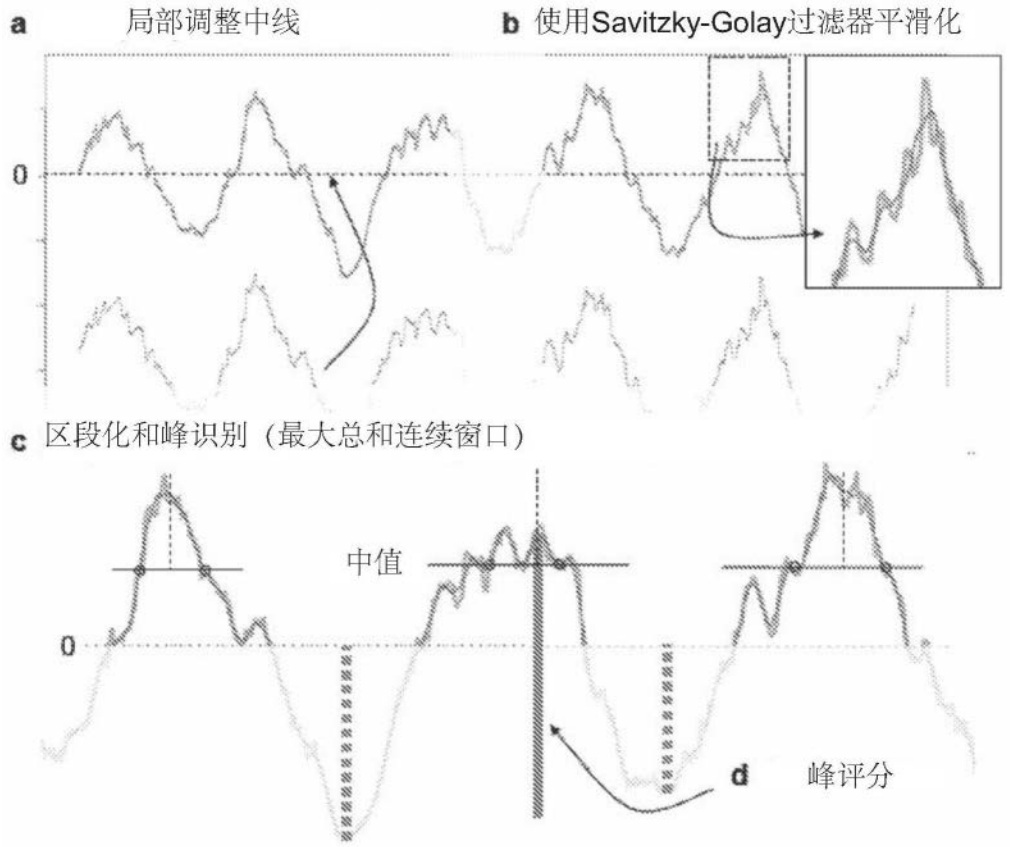

a 局部调整中线
b 使用Savitzky-Golay过滤器平滑化
0
c 区段化和峰识别（最大总和连续窗口）
中值
0
d 峰评分

图28

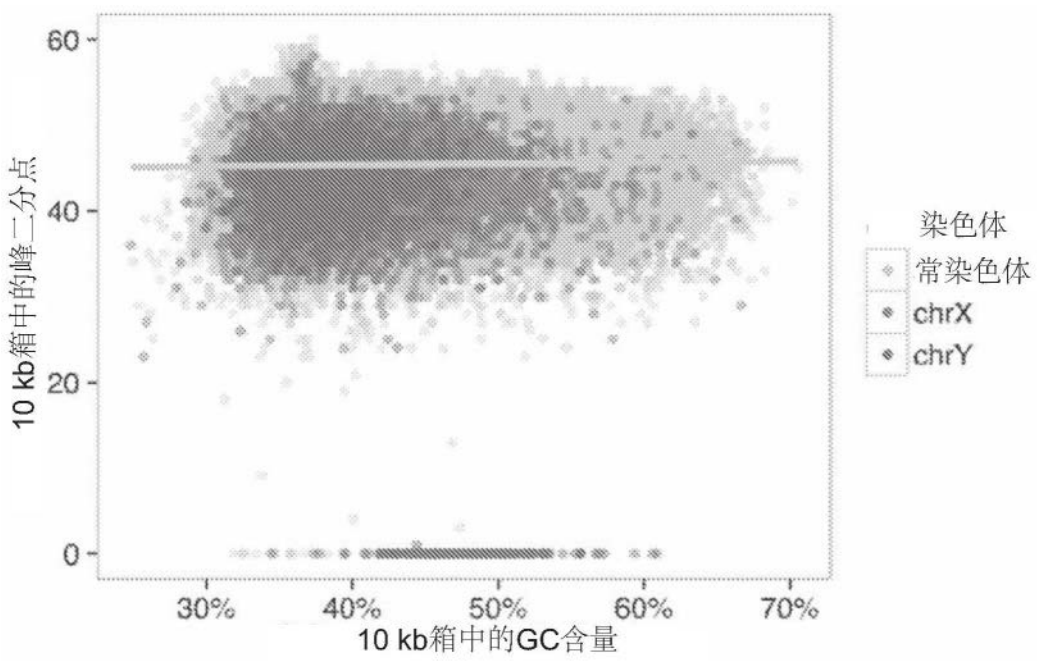

60
40
20
0
10 kb箱中的峰二分点
30%
40%
50%
60%
70%
10 kb箱中的GC含量
染色体
常染色体
chrX
chrY

图29

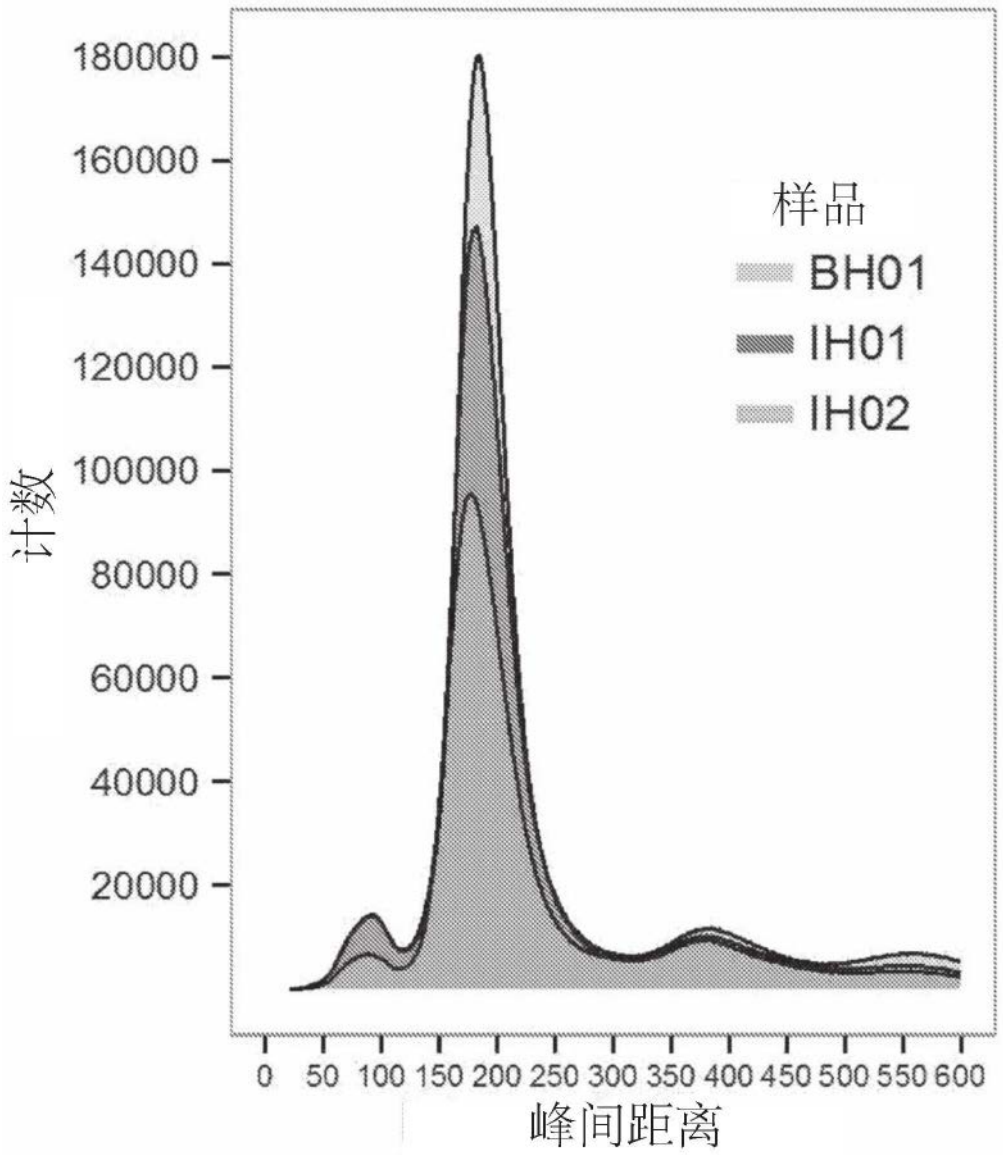

180000
160000
140000
120000
100000
80000
60000
40000
20000
计数
样品
BH01
IH01
IH02
0 50 100 150 200 250 300 350 400 450 500 550 600
峰间距离

图30

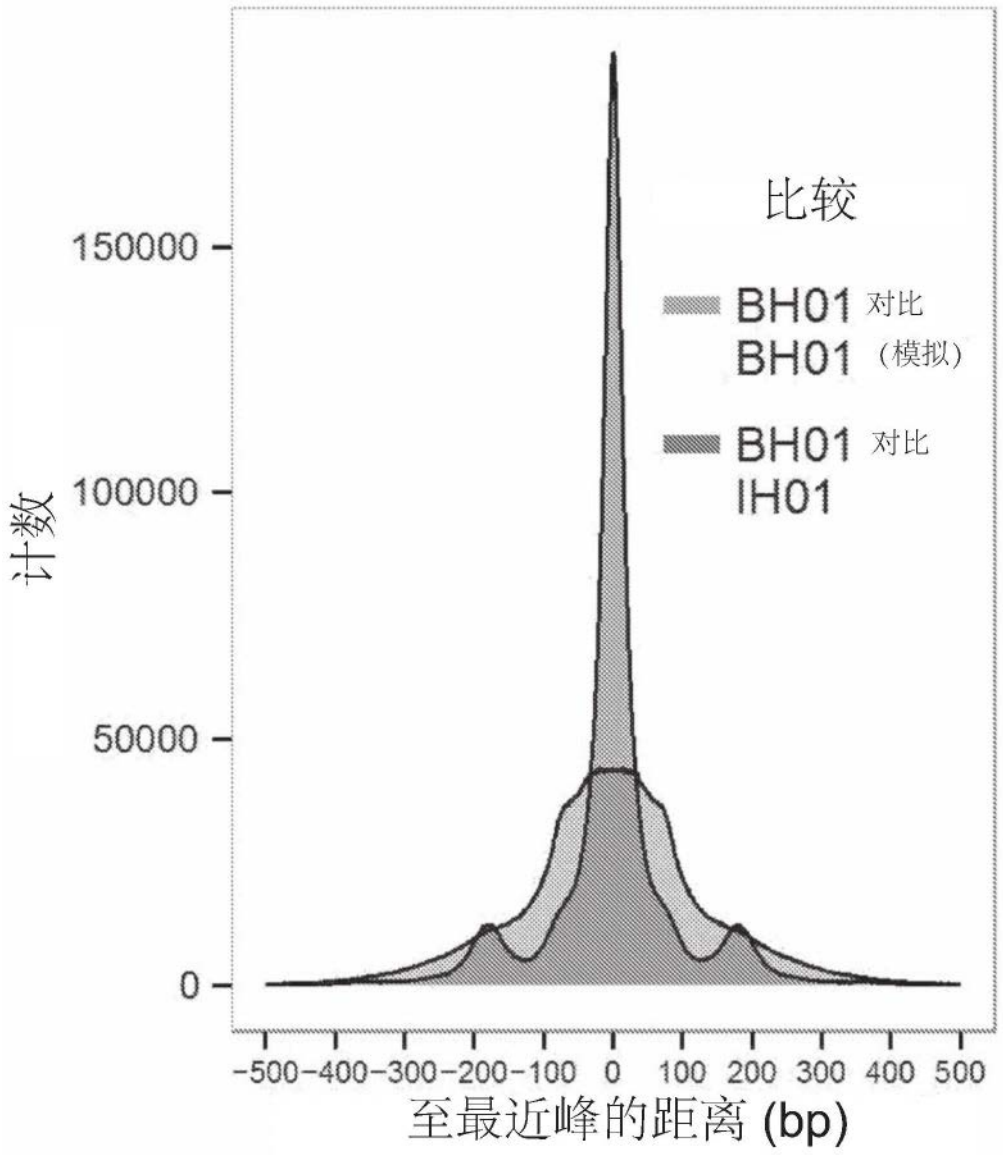

比较
BH01 对比
BH01 （模拟）
BH01 对比
IH01
150000
100000
50000
0
计数
-500 -400 -300 -200 -100 0 100 200 300 400 500
至最近峰的距离 (bp)

图31

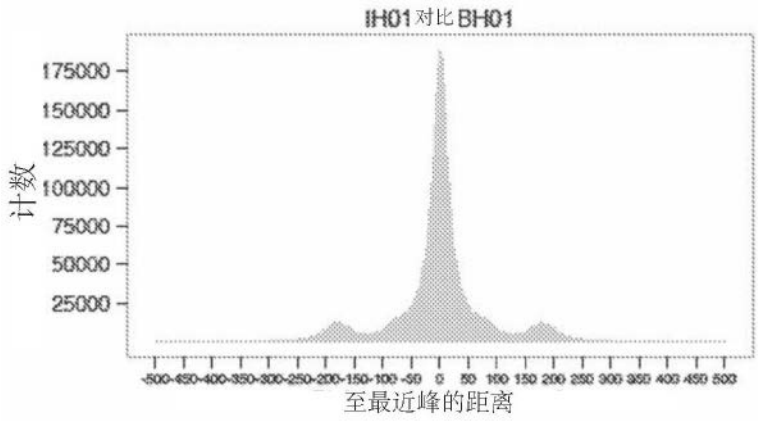

IH01 对比 BH01
175000
150000
125000
100000
75000
50000
25000
计数
至最近峰的距离

图32A

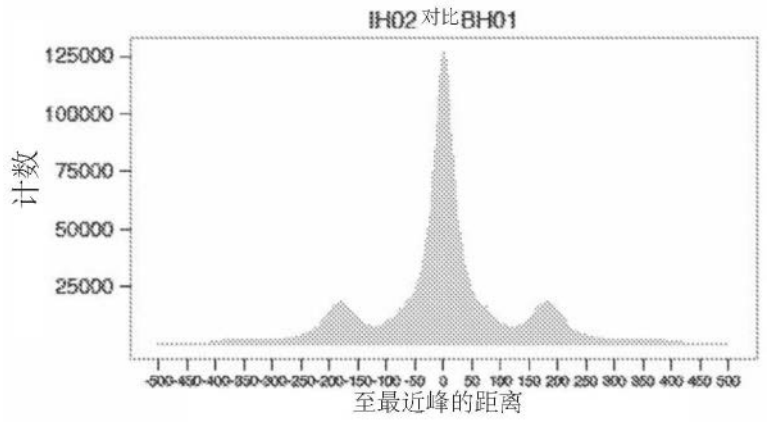


图32B

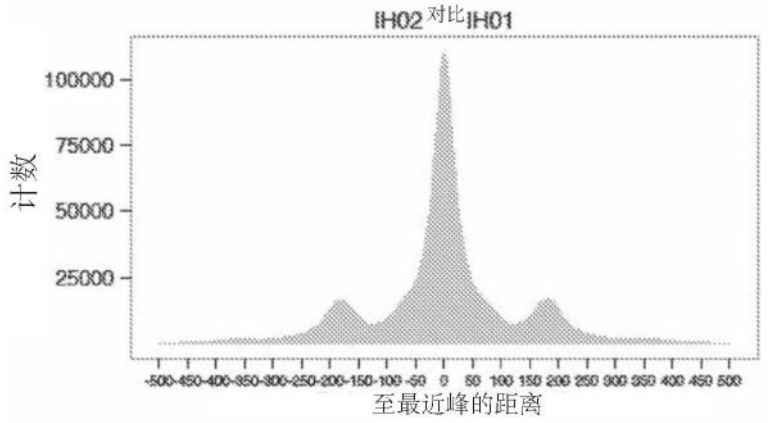


图32C

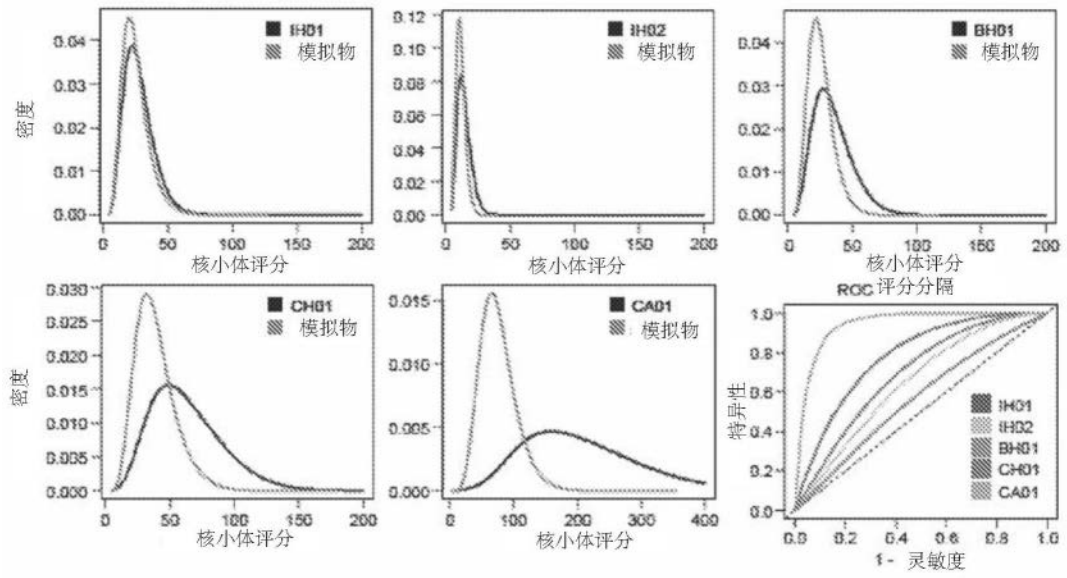


图33A

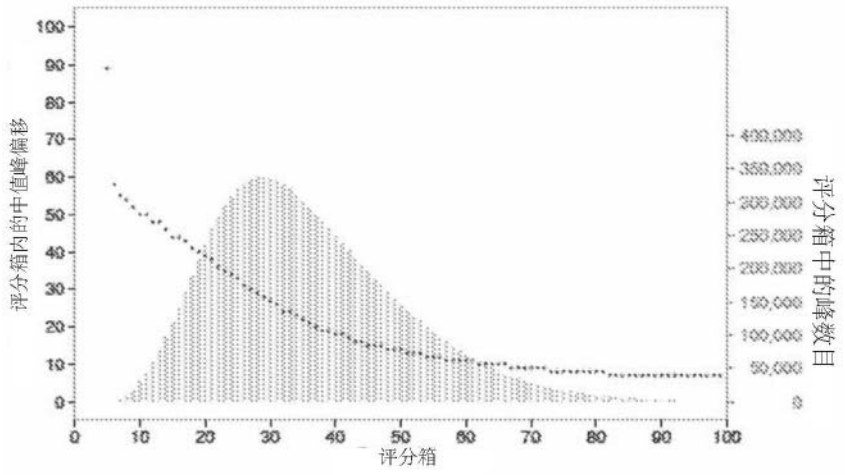


图33B

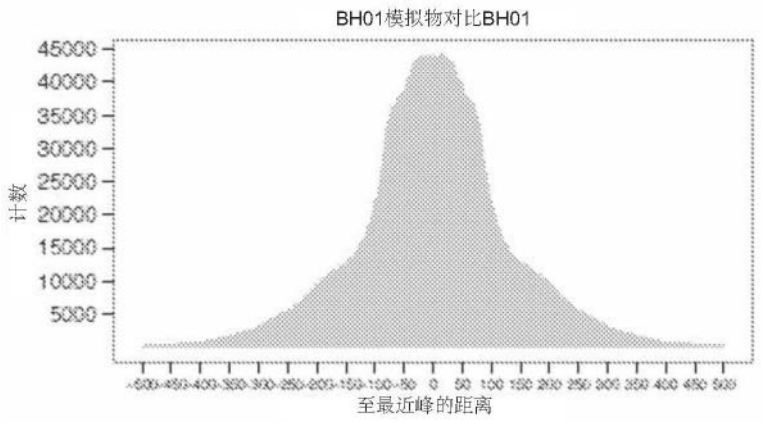


图34A

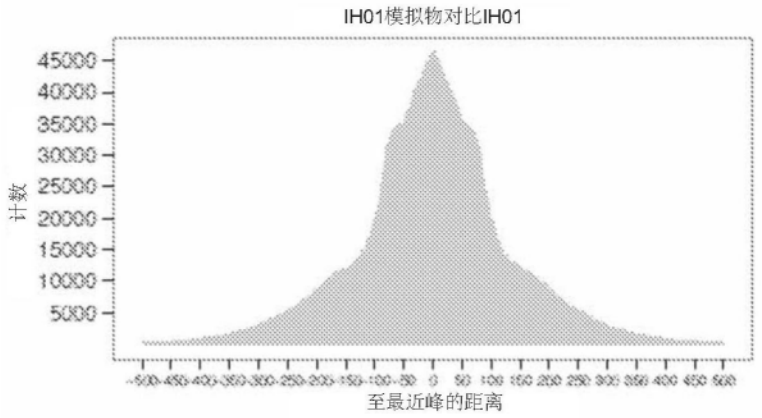


图34B

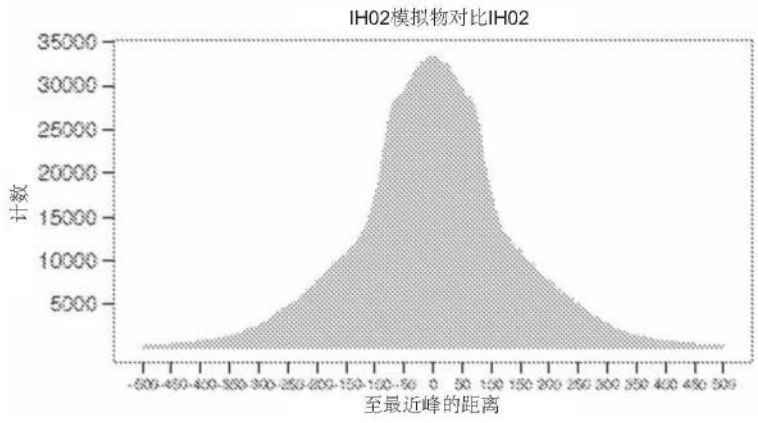

IH02模拟物对比IH02
35000
30000
25000
20000
15000
10000
5000
计数
至最近峰的距离

图34C

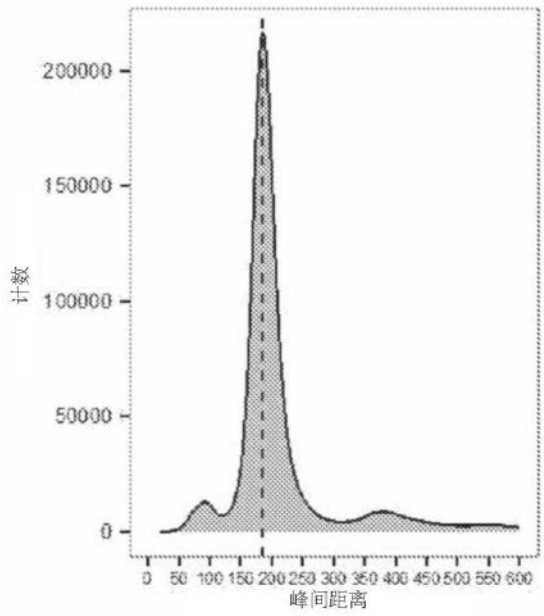

200000
150000
100000
50000
0
计数
0 50 100 150 200 250 300 350 400 450 500 550 600
峰间距离

图35

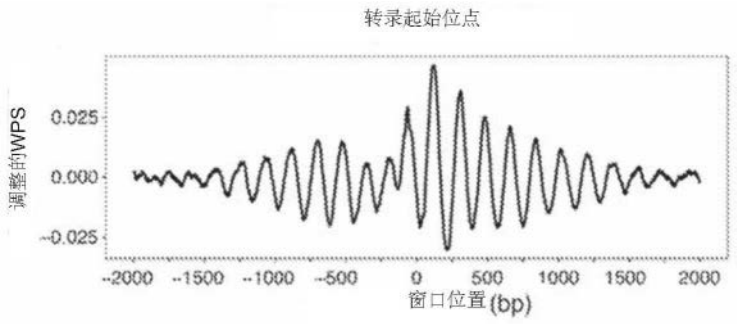

转录起始位点
调整的WPS
0.025
0.000
-0.025
-2000 -1500 -1000 -500 0 500 1000 1500 2000
窗口位置(bp)

图36

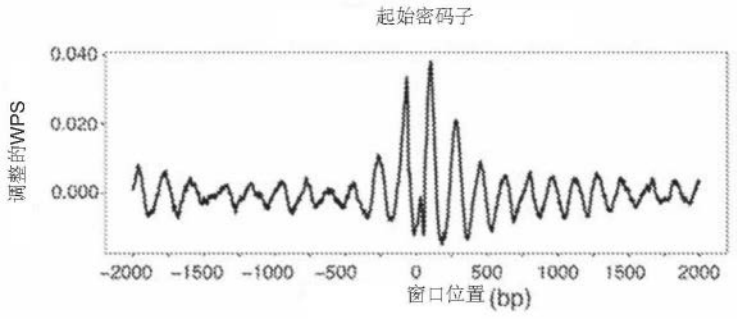

起始密码子
调整的WPS
0.040
0.020
0.000
-2000 -1500 -1000 -500 0 500 1000 1500 2000
窗口位置(bp)

图37

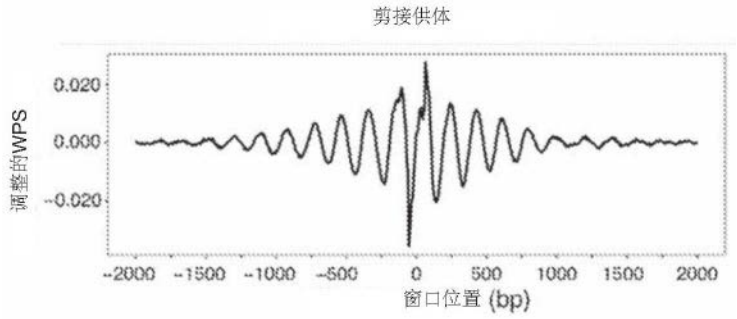

剪接供体
调整的WPS
0.020
0.000
-0.020
-2000 -1500 -1000 -500 0 500 1000 1500 2000
窗口位置 (bp)

图38

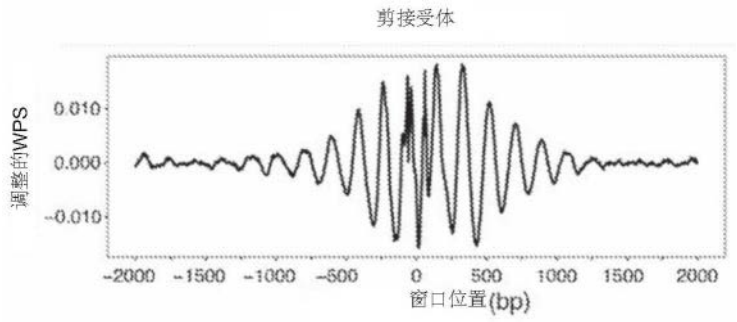

剪接受体
调整的WPS
0.010
0.000
-0.010
-2000 -1500 -1000 -500 0 500 1000 1500 2000
窗口位置(bp)

图39

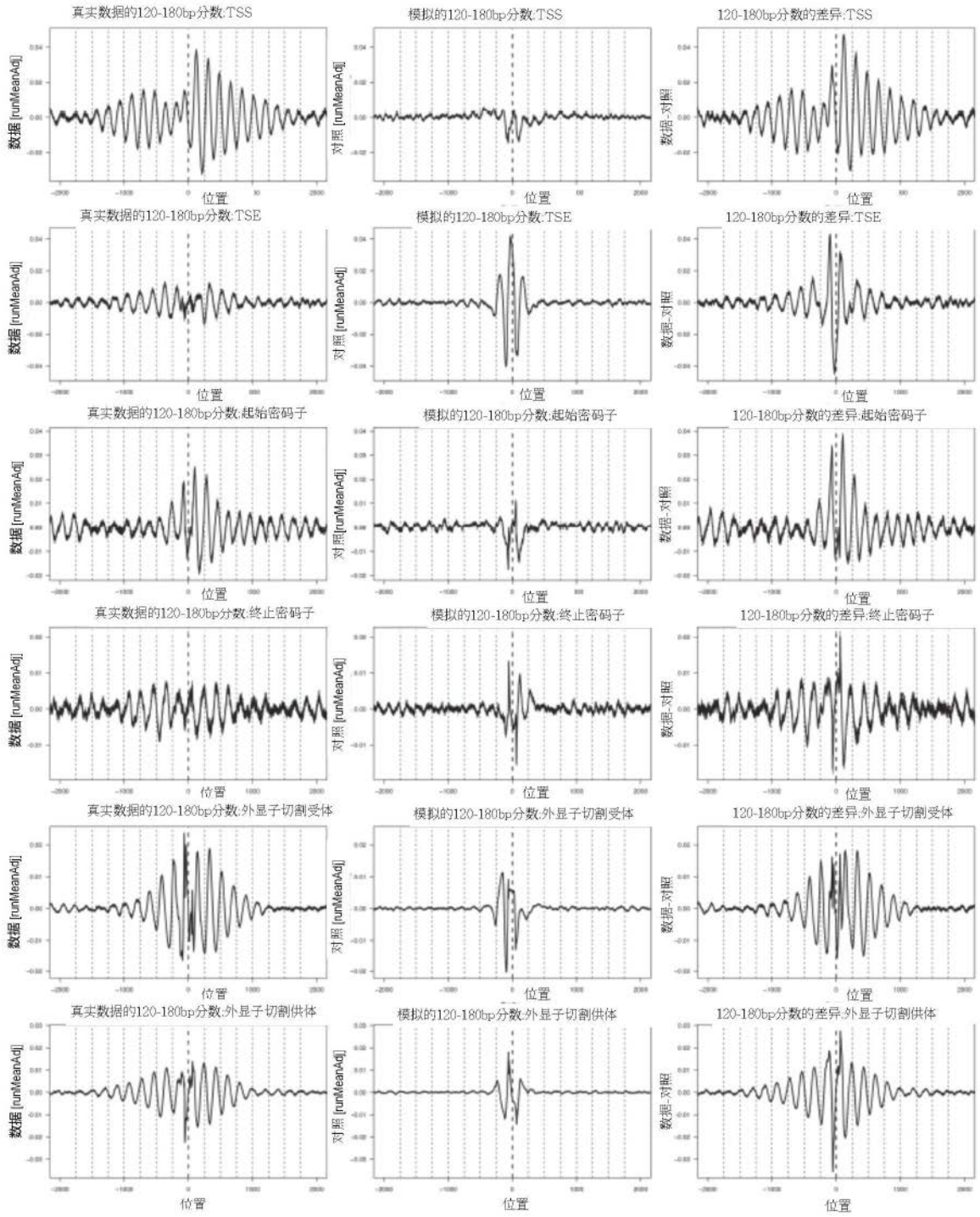

真实数据的120-180bp分数:TSS
模拟的120-180bp分数:TSS
120-180bp分数的差异:TSS
真实数据的120-180bp分数:TSE
模拟的120-180bp分数:TSE
120-180bp分数的差异:TSE
真实数据的120-180bp分数:起始密码子
模拟的120-180bp分数:起始密码子
120-180bp分数的差异:起始密码子
真实数据的120-180bp分数:终止密码子
模拟的120-180bp分数:终止密码子
120-180bp分数的差异:终止密码子
真实数据的120-180bp分数:外显子切割受体
模拟的120-180bp分数:外显子切割受体
120-180bp分数的差异:外显子切割受体
真实数据的120-180bp分数:外显子切割供体
模拟的120-180bp分数:外显子切割供体
120-180bp分数的差异:外显子切割供体
数据 [runMeanAdj]
对照 [runMeanAdj]
数据-对照
位置

图40

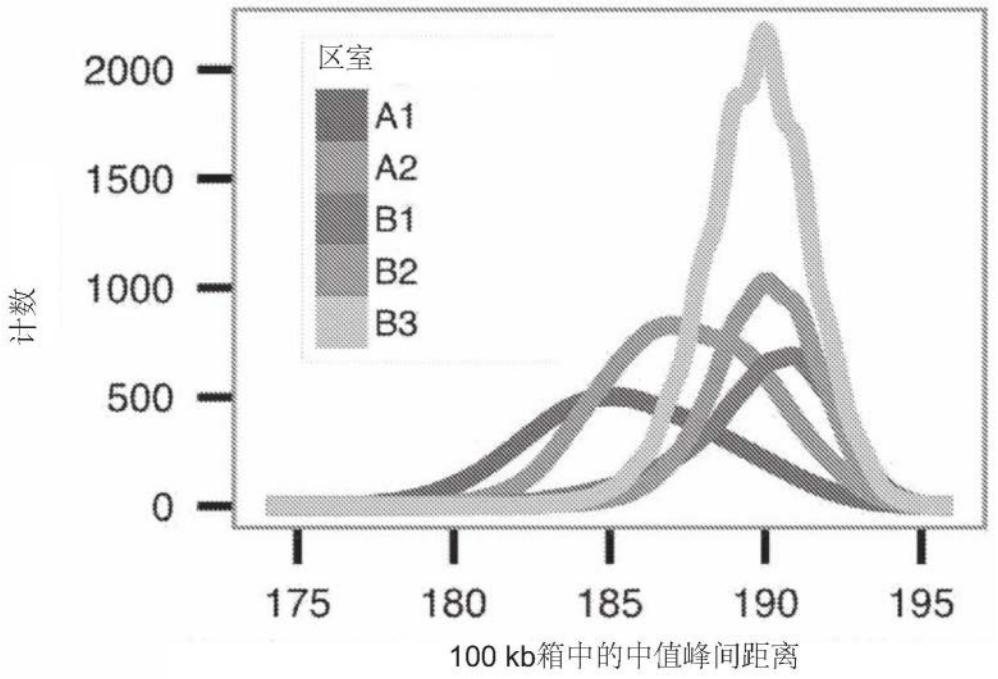

区室
A1
A2
B1
B2
B3
2000
1500
1000
500
0
计数
175
180
185
190
195
100 kb箱中的中值峰间距离

图41

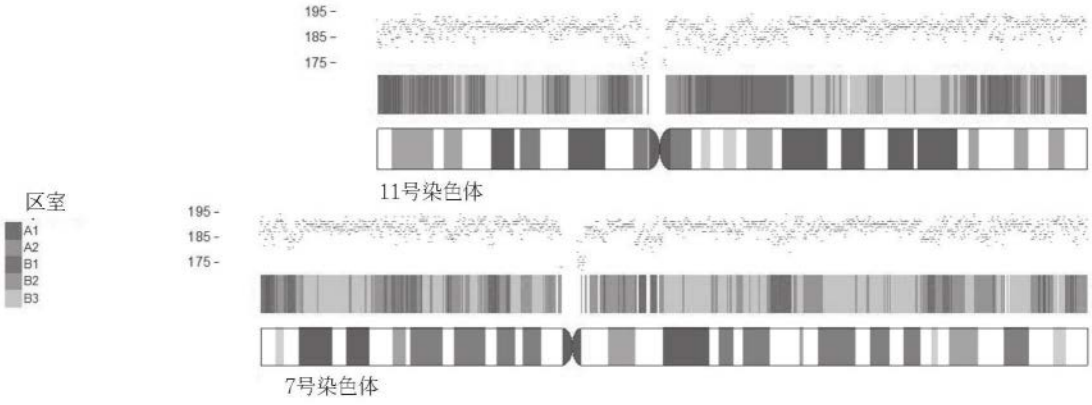

区室
A1
A2
B1
B2
B3
195
185
175
11号染色体
195
185
175
7号染色体

图42

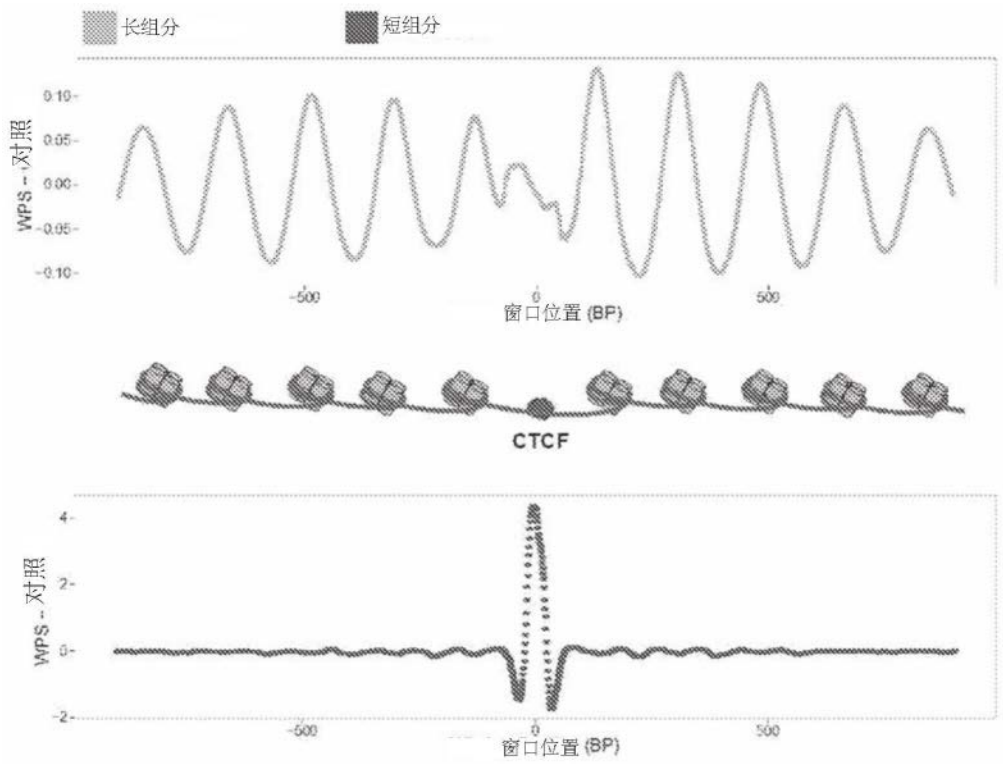

长组分
短组分
WPS - 对照
0.10
0.05
0.00
-0.05
-0.10
-500
0
500
窗口位置 (BP)
CTCF
WPS - 对照
4
2
0
-2
-500
0
500
窗口位置 (BP)

图43

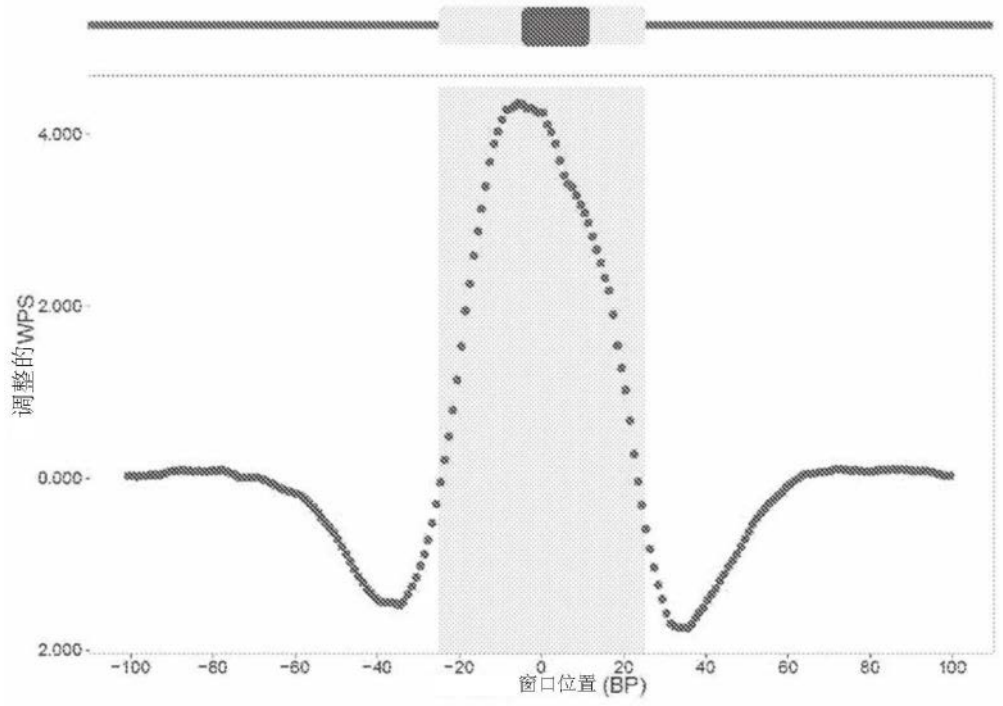

调整的WPS
4.000
2.000
0.000
2.000
-100
-80
-60
-40
-20
0
20
40
60
80
100
窗口位置 (BP)

图44

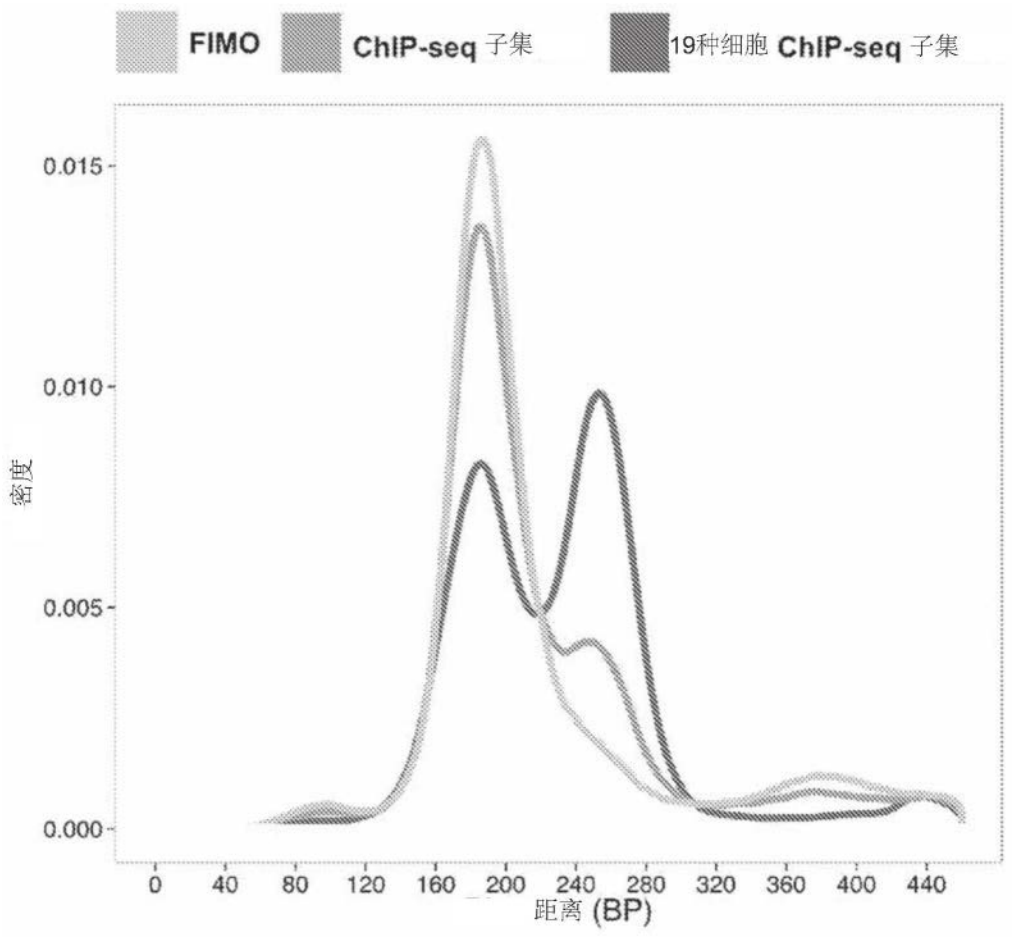

FIMO
ChIP-seq 子集
19种细胞 ChIP-seq 子集
0.015
0.010
0.005
0.000
密度
0
40
80
120
160
200
240
280
320
360
400
440
距离 (BP)

图45

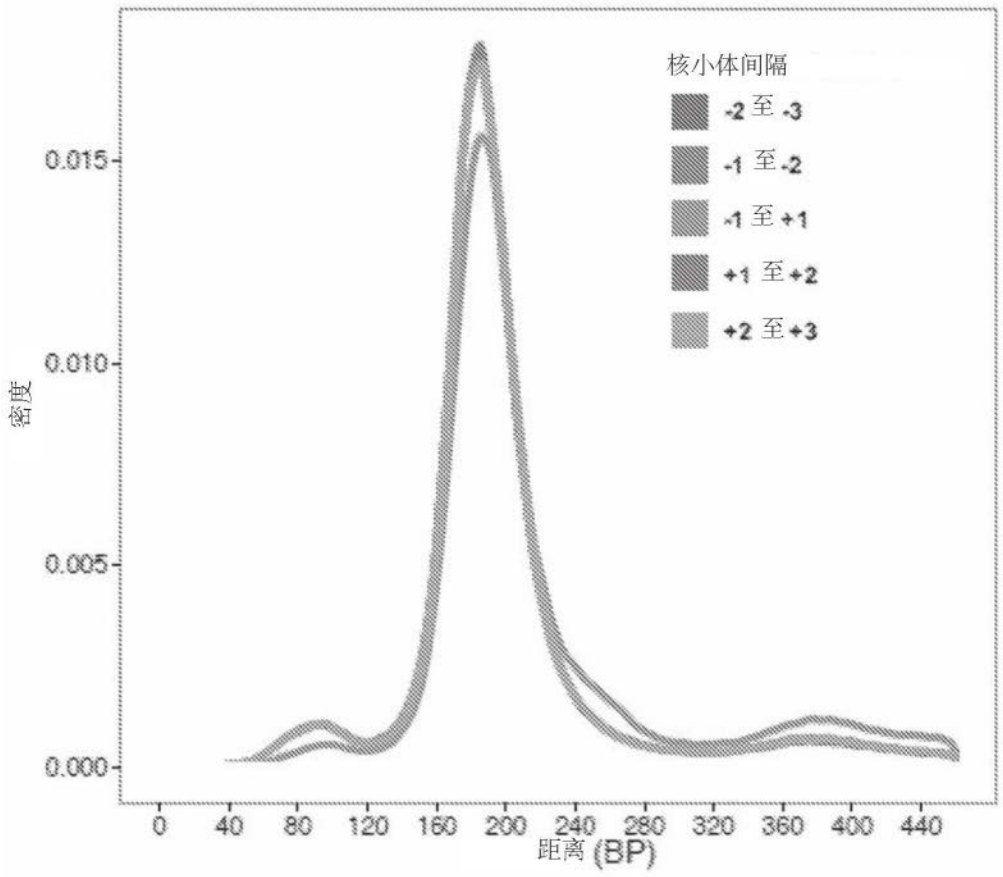

核小体间隔
-2 至 -3
-1 至 -2
-1 至 +1
+1 至 +2
+2 至 +3
密度
0.015
0.010
0.005
0.000
0 40 80 120 160 200 240 280 320 360 400 440
距离 (BP)

图46

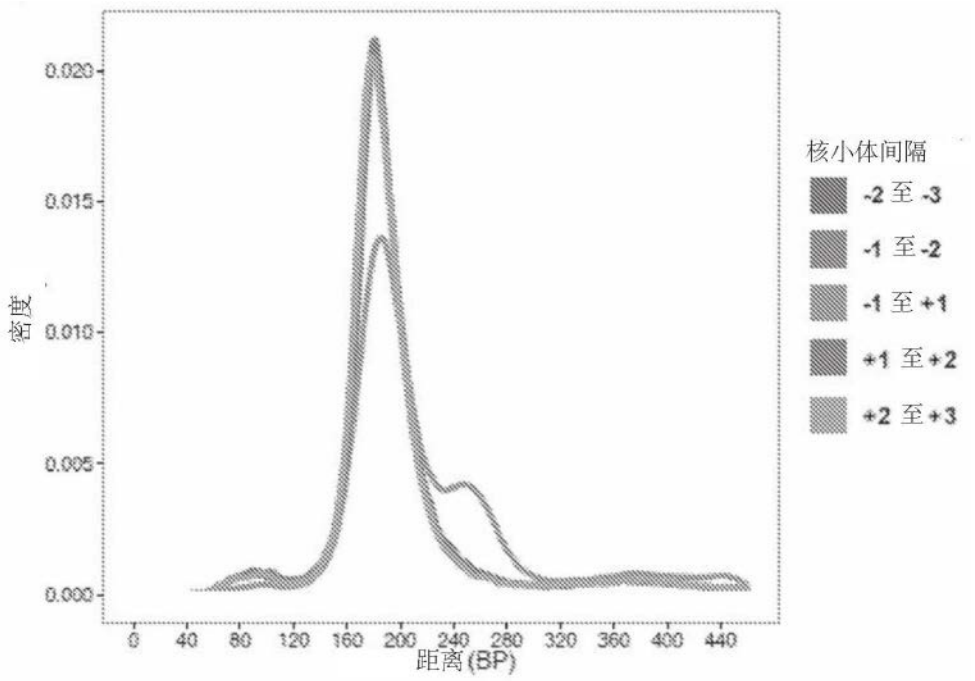

核小体间隔
-2 至 -3
-1 至 -2
-1 至 +1
+1 至 +2
+2 至 +3
密度
0.020
0.015
0.010
0.005
0.000
0 40 80 120 160 200 240 280 320 360 400 440
距离 (BP)

图47

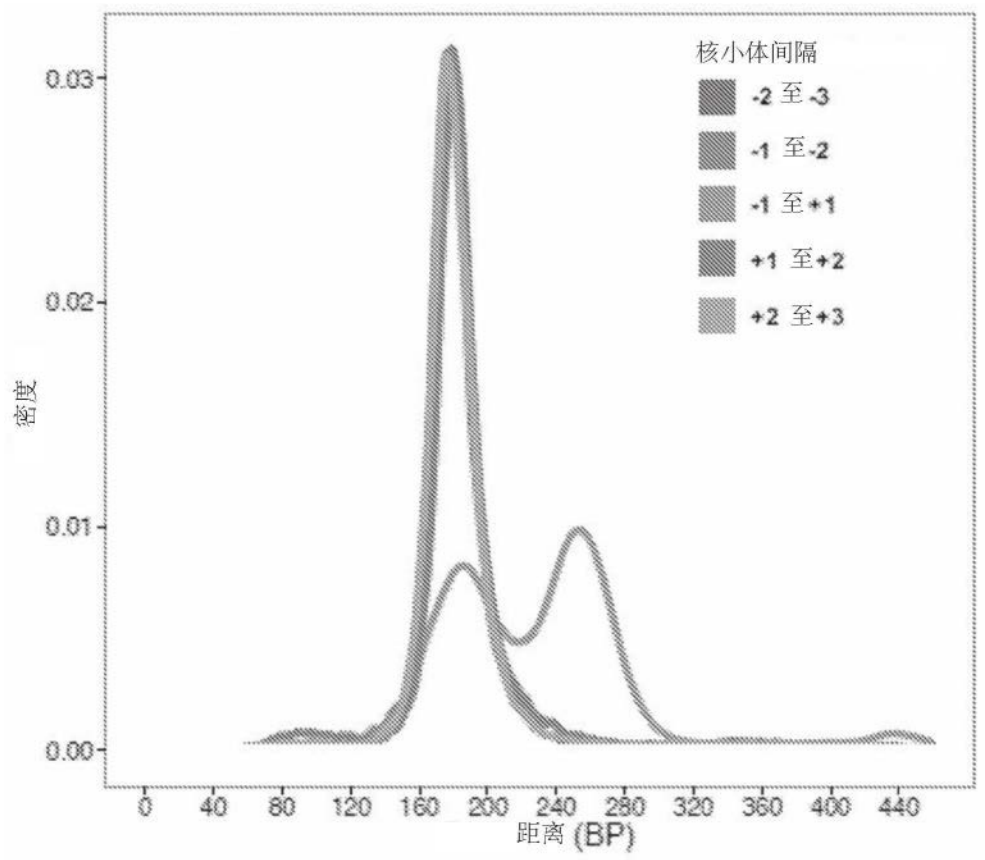

核小体间隔
-2 至 -3
-1 至 -2
-1 至 +1
+1 至 +2
+2 至 +3
密度
0.03
0.02
0.01
0.00
0
40
80
120
160
200
240
280
320
360
400
440
距离 (BP)

图48

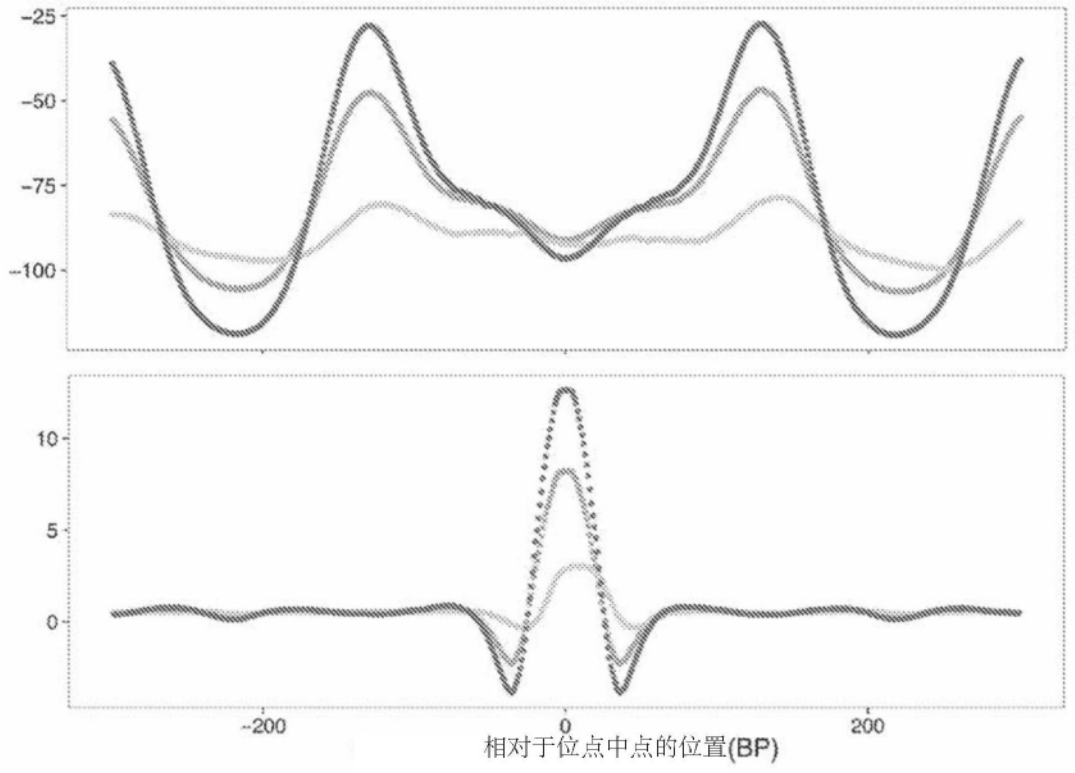


图49

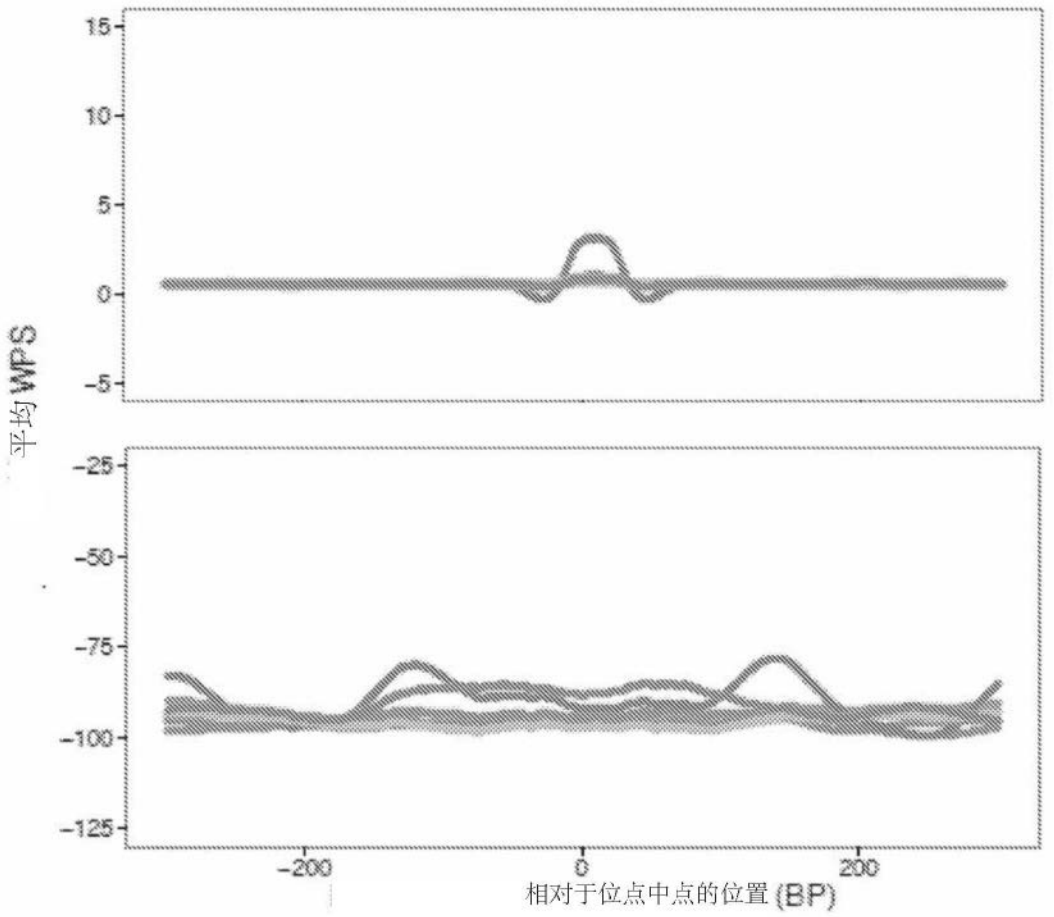

15
10
5
0
-5
-25
-50
-75
-100
-125
-200
0
200
WPS 均值
相对于位点中点的位置 (BP)

图50

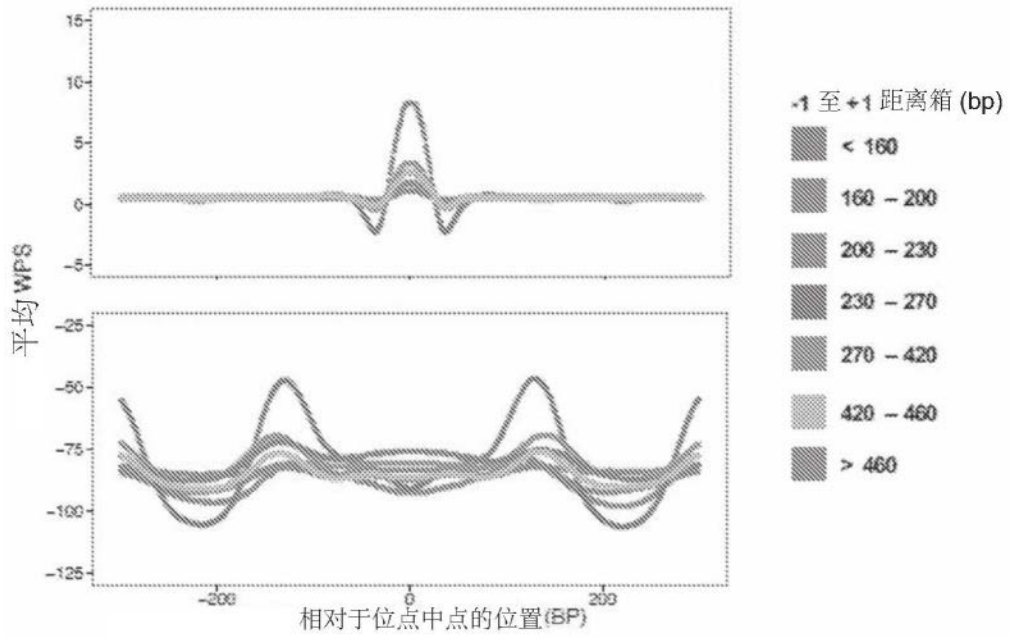

15
10
5
0
-5
-25
-50
-75
-100
-125
-200
0
200
WPS 均值
相对于位点中点的位置(BP)
-1 至 +1 距离箱 (bp)
< 160
160 – 200
200 – 230
230 – 270
270 – 420
420 – 460
> 460

图51

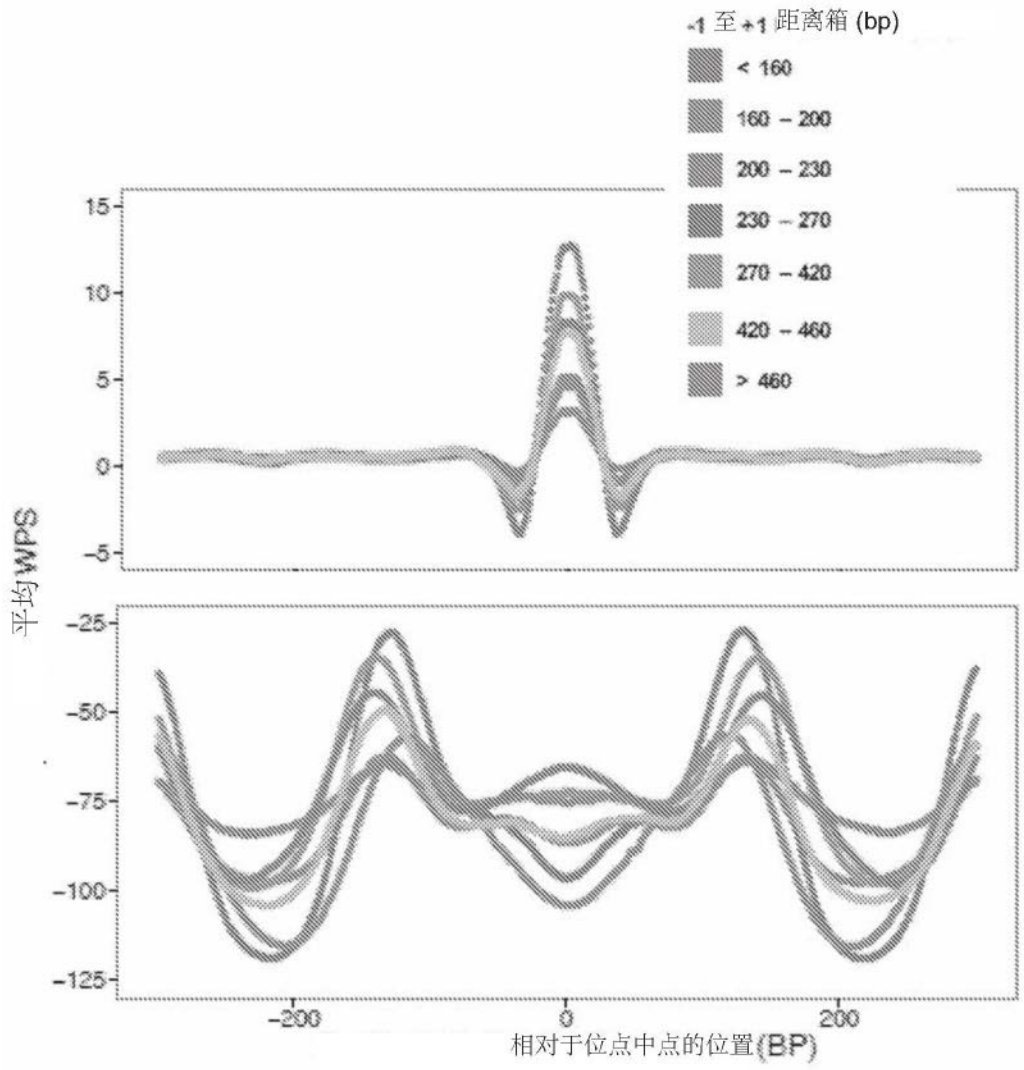

-1 至 +1 距离箱 (bp)
< 160
160 – 200
200 – 230
230 – 270
270 – 420
420 – 460
> 460
平均WPS
15
10
5
0
-5
-25
-50
-75
-100
-125
-200
0
200
相对于位点中点的位置(BP)

图52

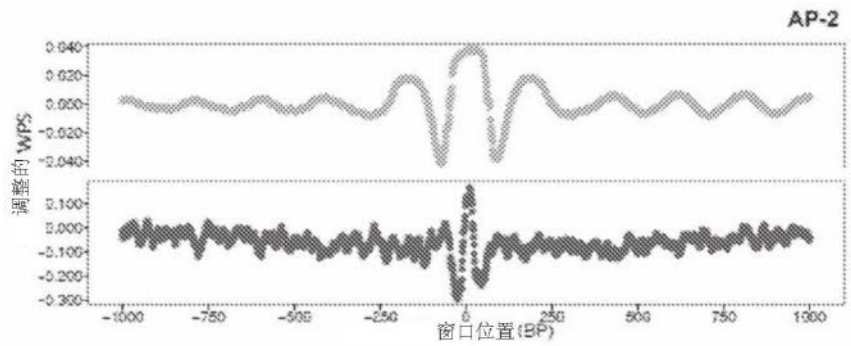

AP-2
调整的WPS
窗口位置(BP)

图53A

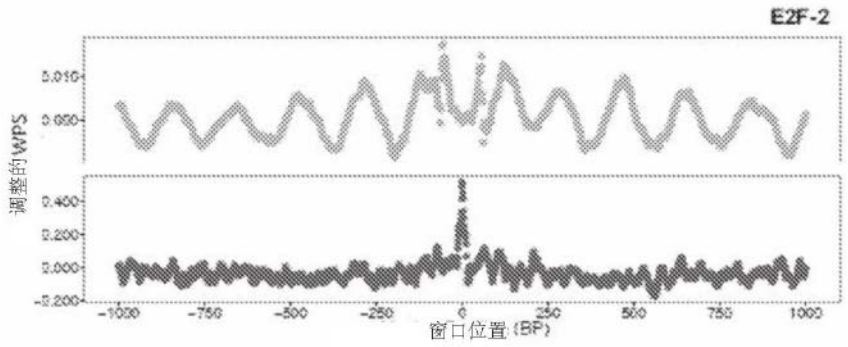


图53B

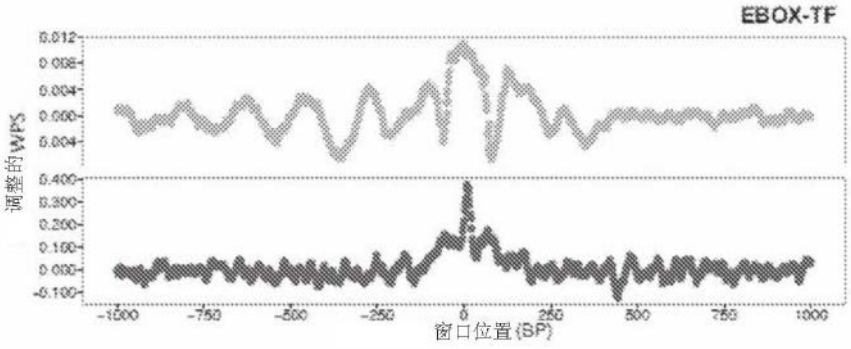


图53C

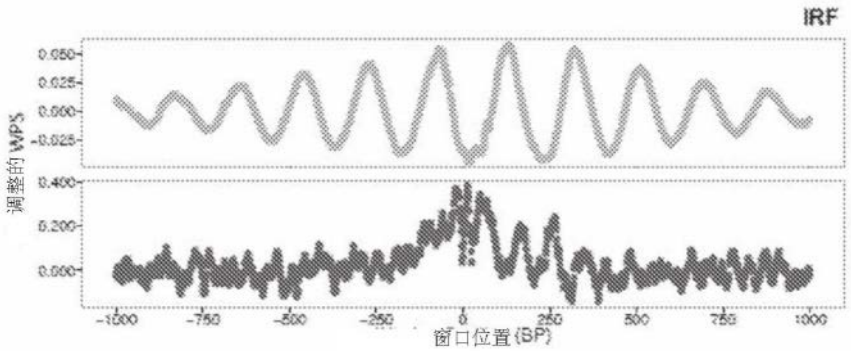


图53D

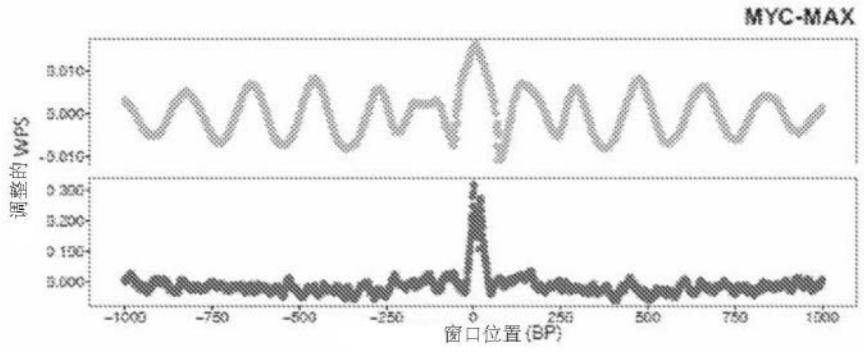


图53E

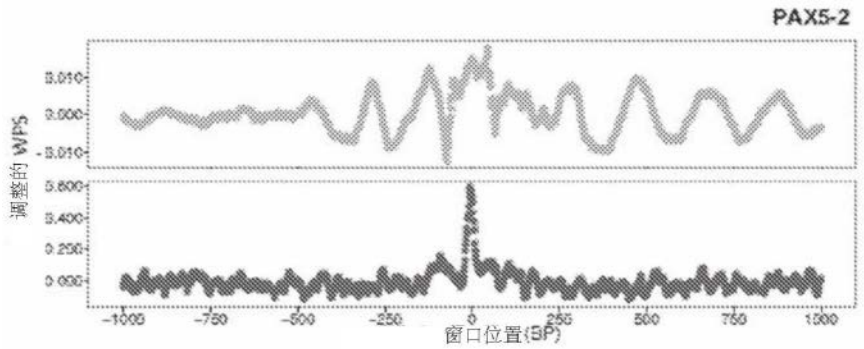


图53F

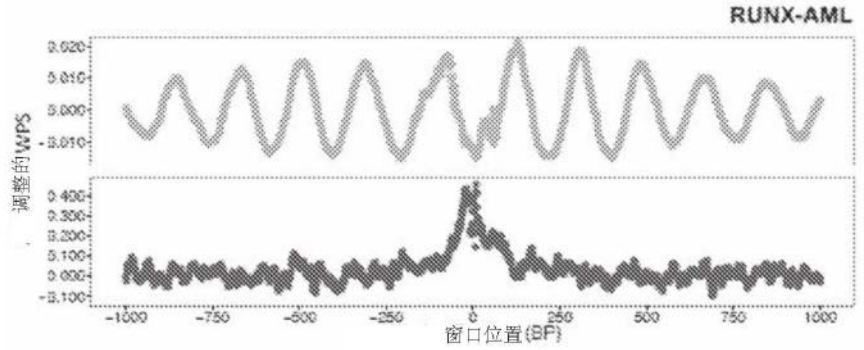


图53G

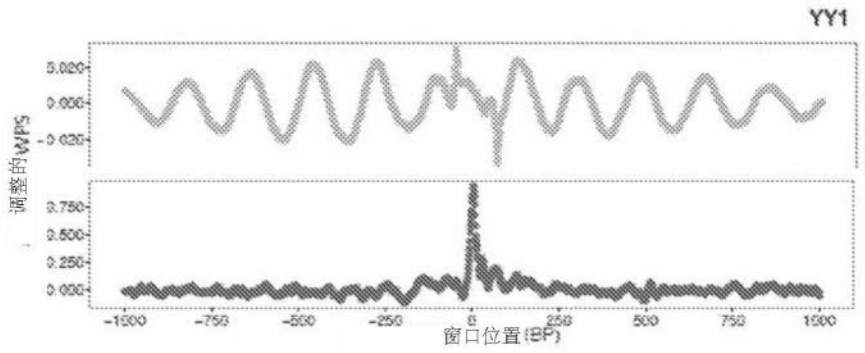


图53H

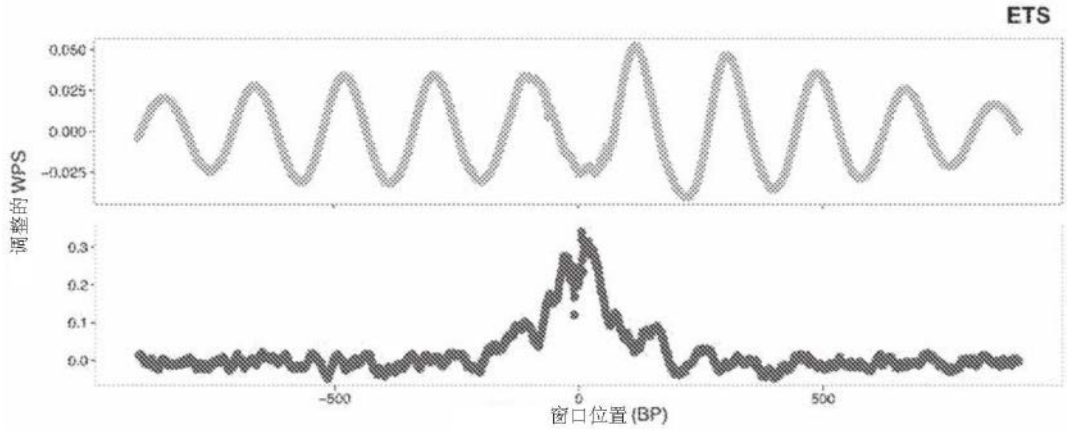


图54

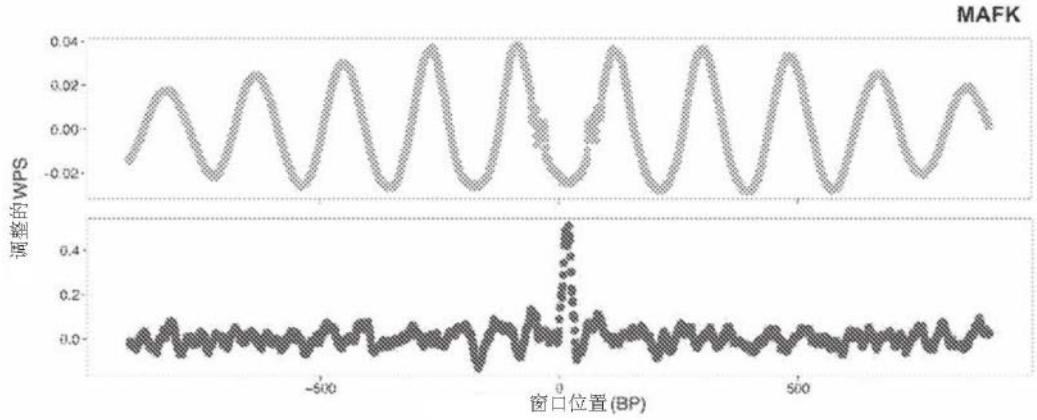


图55

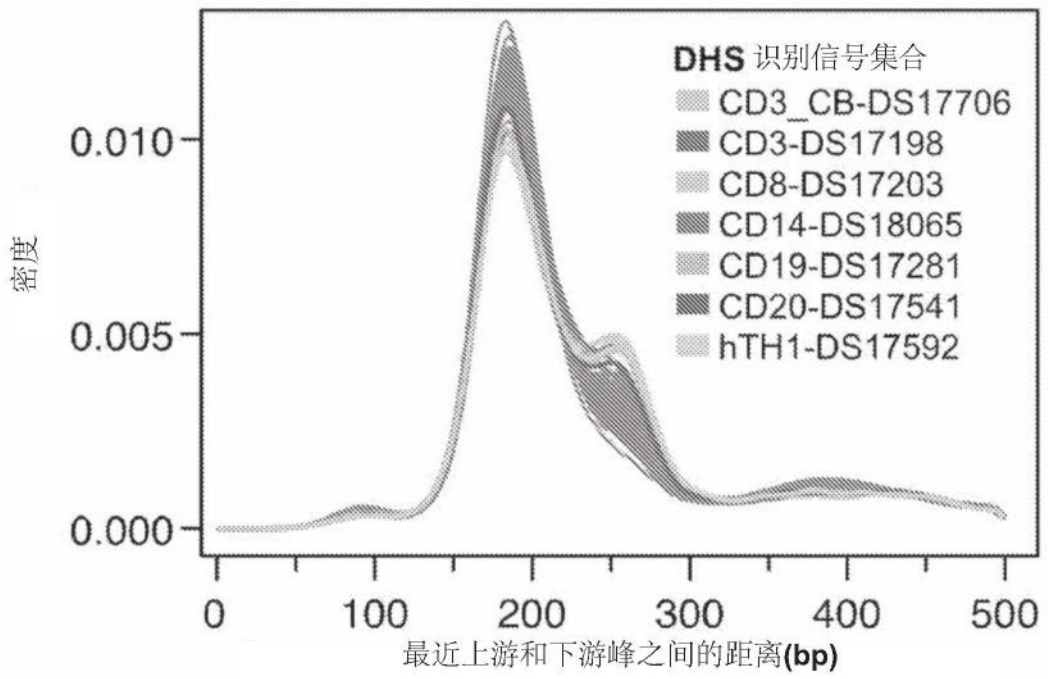

DHS 识别信号集合
CD3_CB-DS17706
CD3-DS17198
CD8-DS17203
CD14-DS18065
CD19-DS17281
CD20-DS17541
hTH1-DS17592
密度
0.010
0.005
0.000
0
100
200
300
400
500
最近上游和下游峰之间的距离(bp)

图56

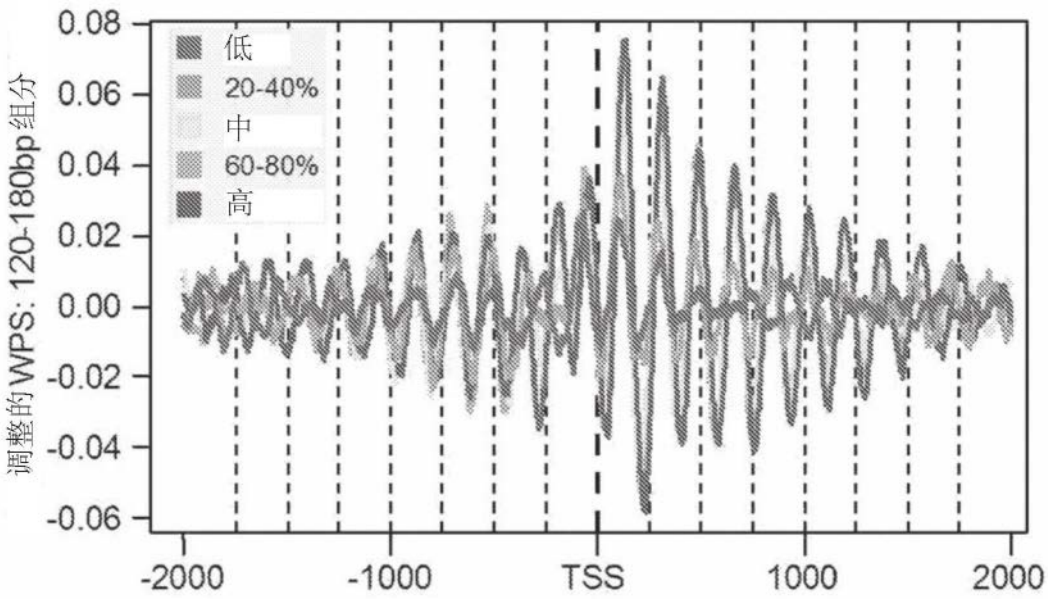

低
20-40%
中
60-80%
高
调整的WPS: 120-180bp组分
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-2000
-1000
TSS
1000
2000

图57

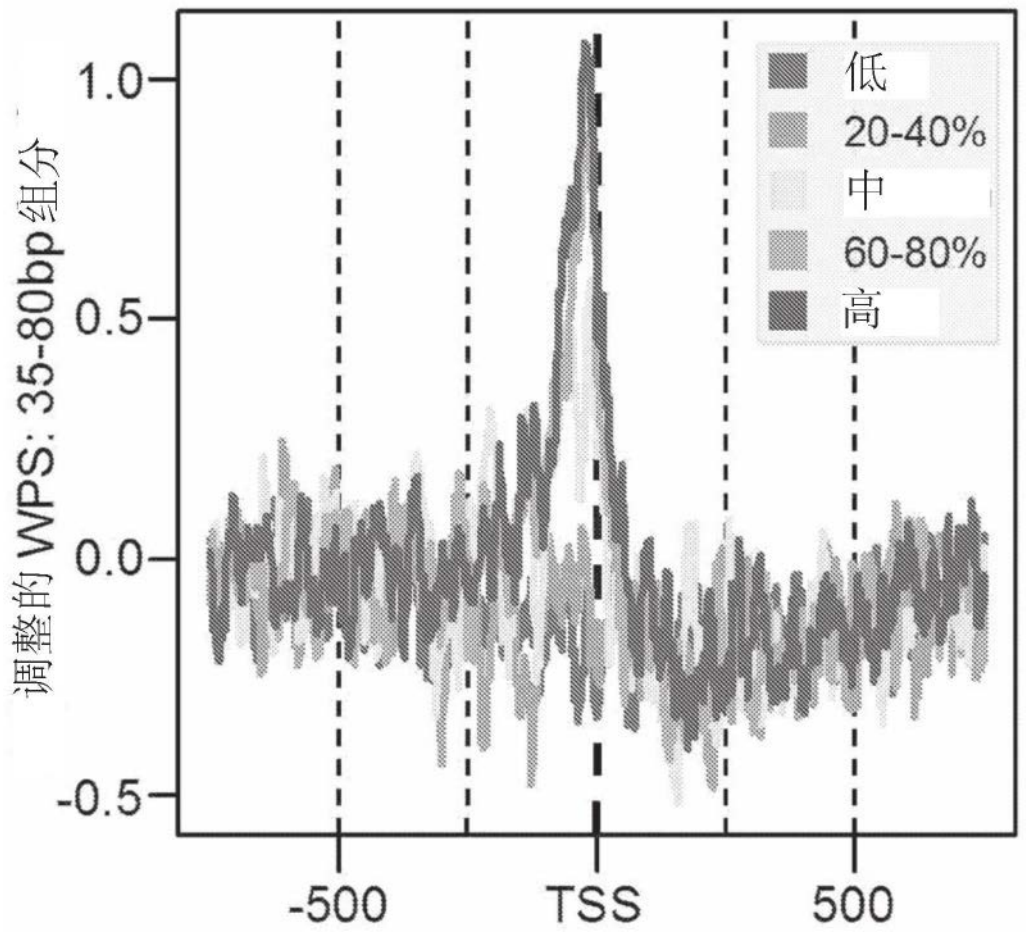


图58

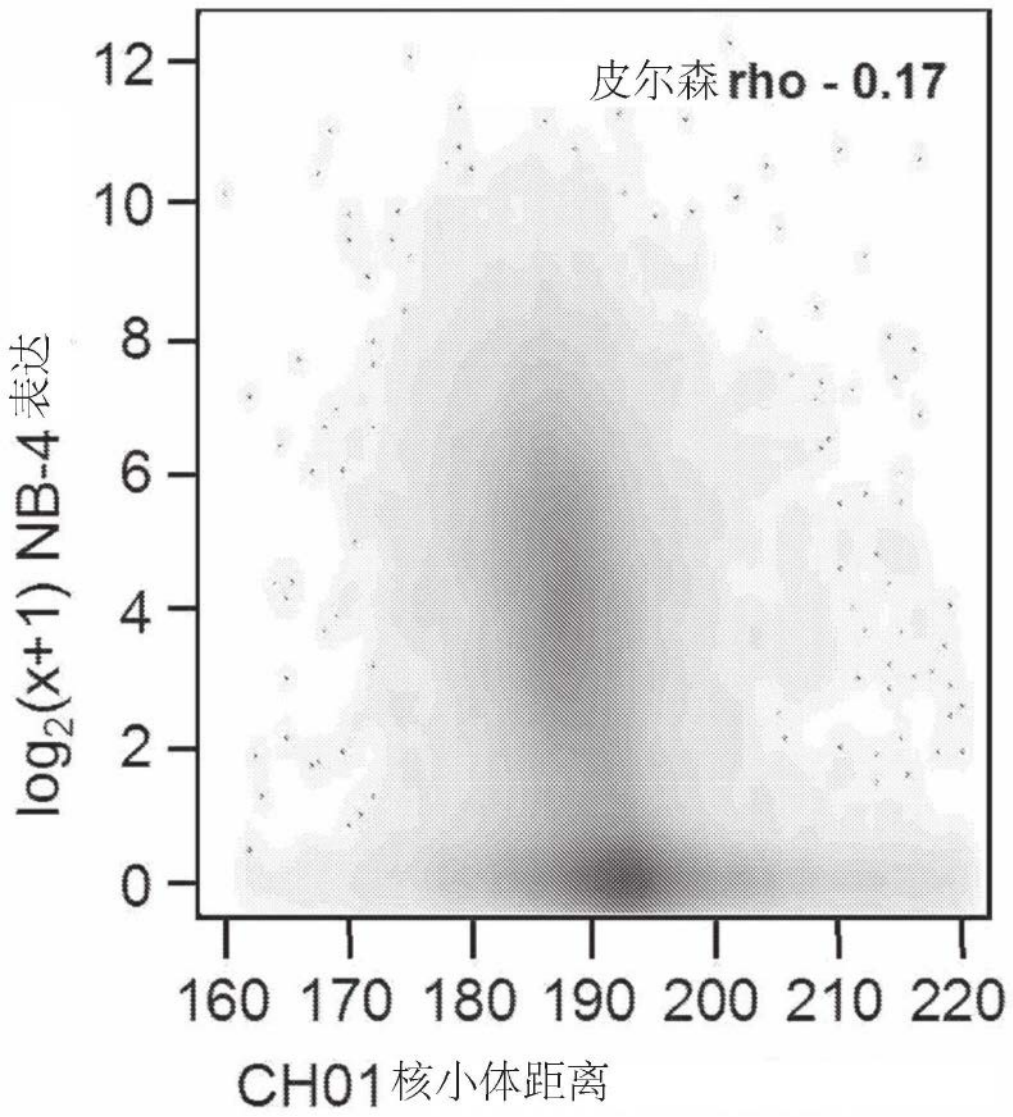

皮尔森 rho - 0.17
log2(x+1) NB-4 表达
0
2
4
6
8
10
12
160 170 180 190 200 210 220
CH01 核小体距离

图59

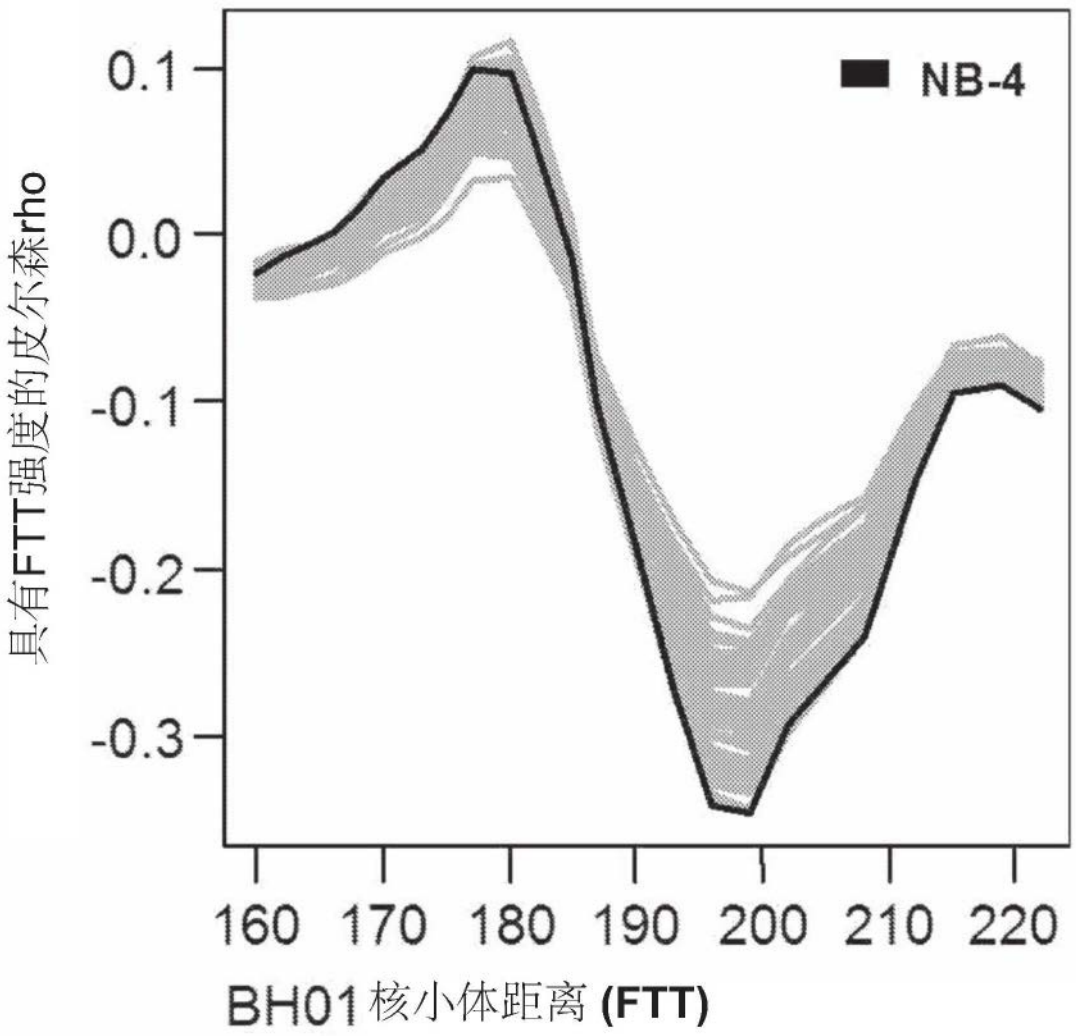
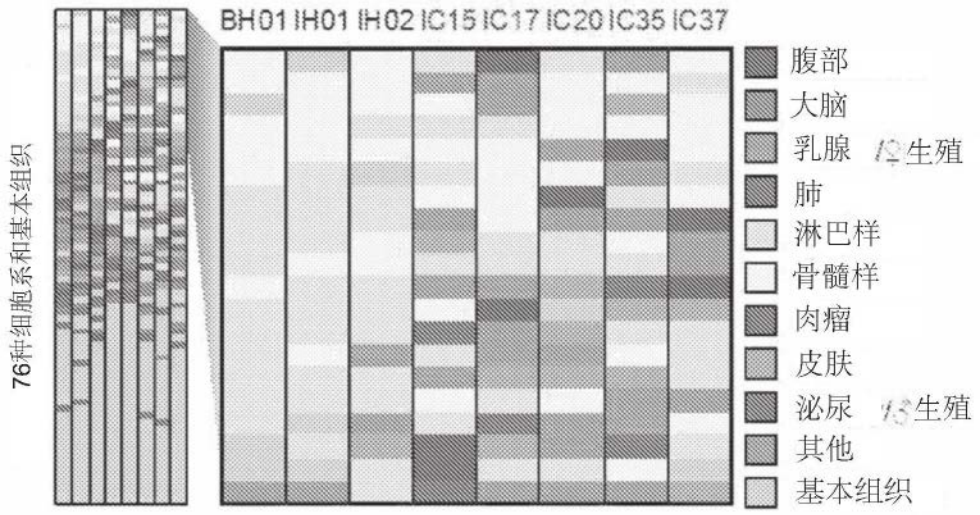

NB-4
0.1
0.0
-0.1
-0.2
-0.3
具有FTT强度的皮尔森rho
160
170
180
190
200
210
220
BH01核小体距离 (FTT)

图60



样品	癌症	前3位增加
IC15 ♂	肺 (SCLC)	SCLC-21H (31), SH-SY5Y (25), HEK 293 (17)
IC17 ♂	肝 (HCC)	肾 (21), CAPAN-2 (19), BEWO (19)
IC20 ♂	肺 (SCC)	SK-BR-3 (21), Hep G2 (18), HaCaT (18)
IC35 ♀	乳腺 (DC)	BEWO (27), SiHa (27), CAPAN-2 (25)
IC37 ♀	结直肠 (AC)	Hep G2 (24), SK-BR-3 (22), EFO-21 (20)
对照	-	十二指肠(10), U-251 MG (9), 小肠**(9)**

图61

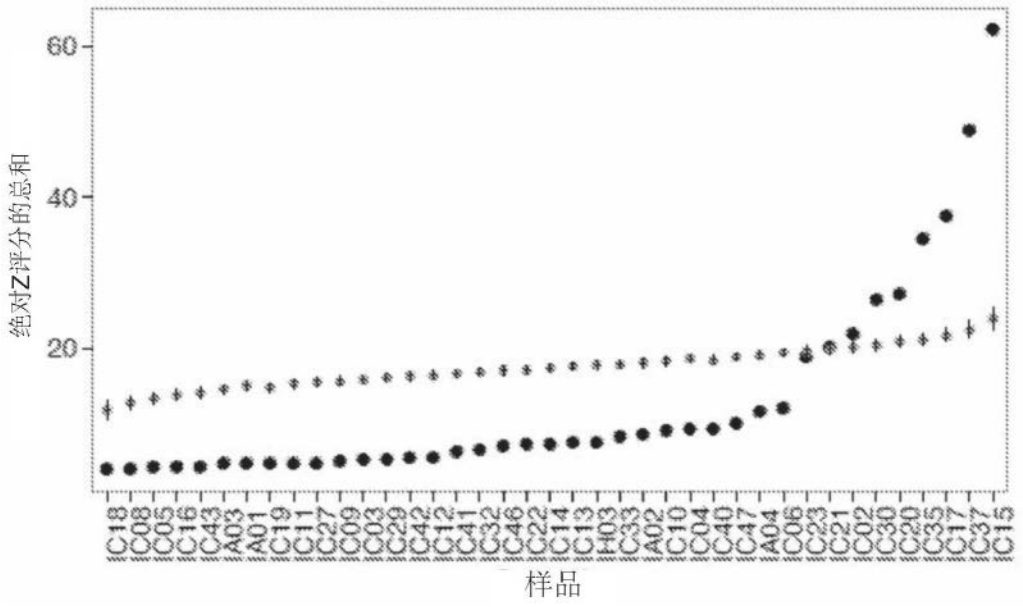

60
40
20
绝对Z评分的总和
C18
C08
C05
C16
C43
A03
A01
C19
C11
C27
C09
C03
C29
C42
C12
C41
C32
C46
C22
C14
C13
H03
C33
A02
C10
C04
C40
C47
A04
C06
C23
C21
C02
C30
C20
C35
C17
C37
C15
样品

图62A

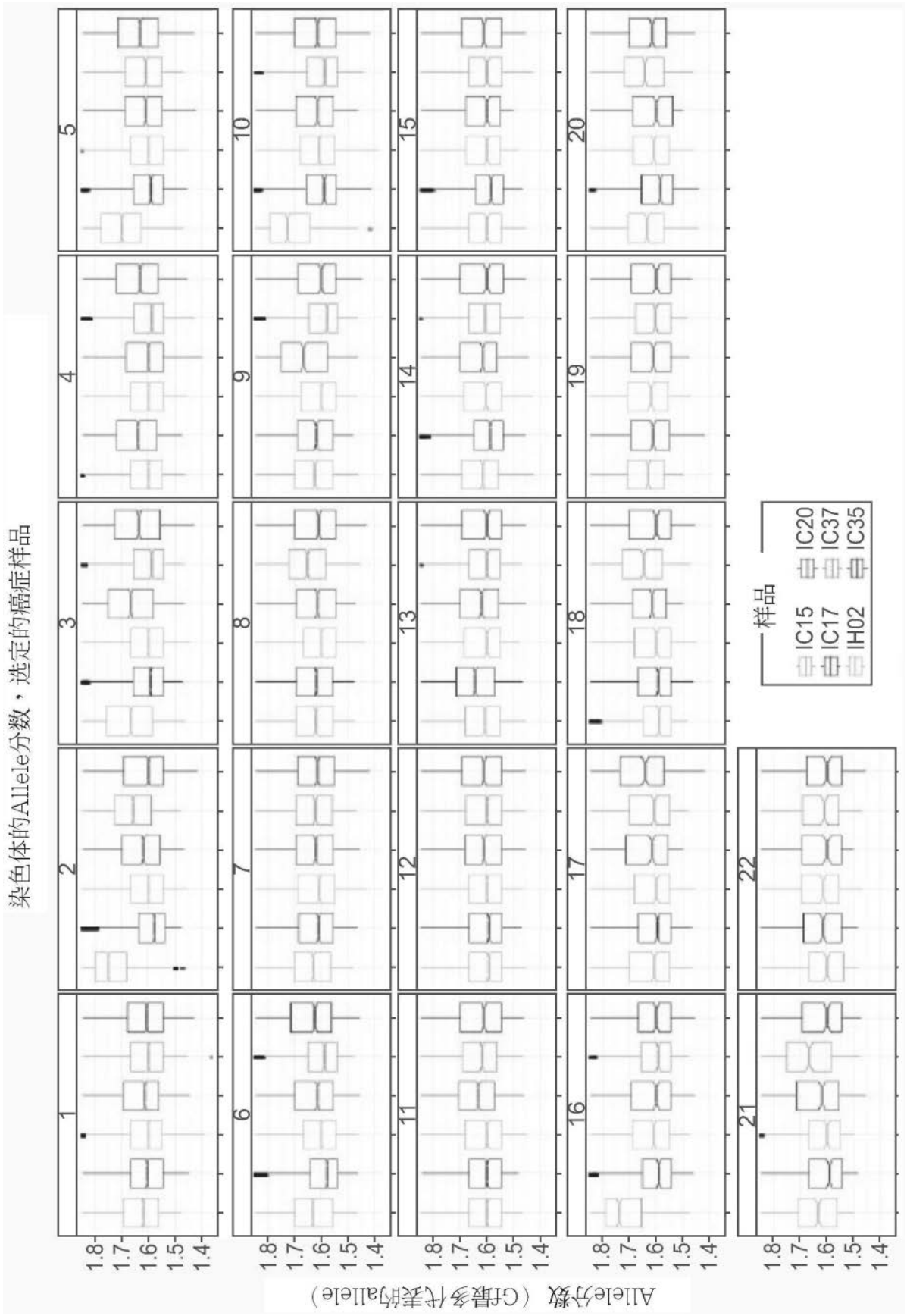

染色体的Allele分数，选定的癌症样品
Allele分数 （Gi展多个t表的allele）
样品
IC15
IC17
IH02
IC20
IC37
IC35
1.8
1.7
1.6
1.5
1.4

图62B

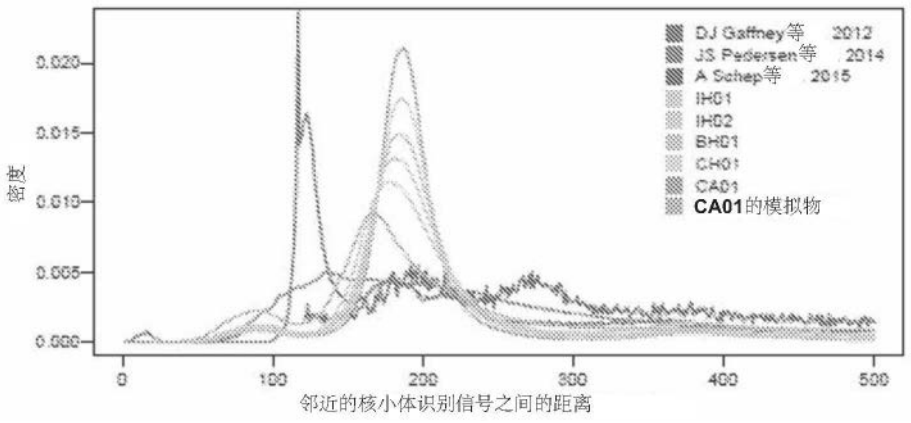


图63A

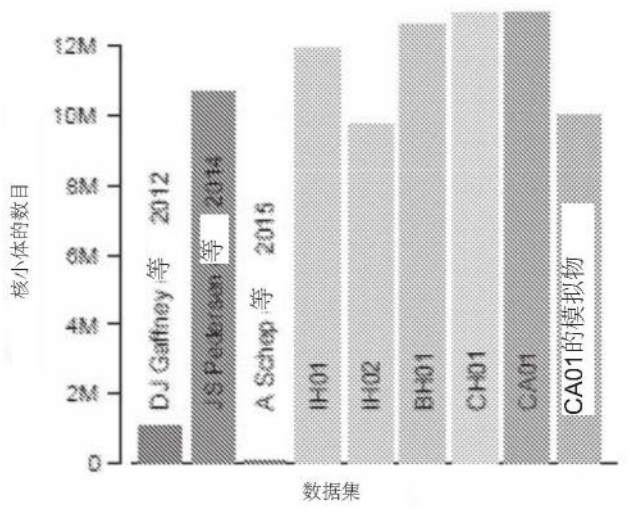


图63B

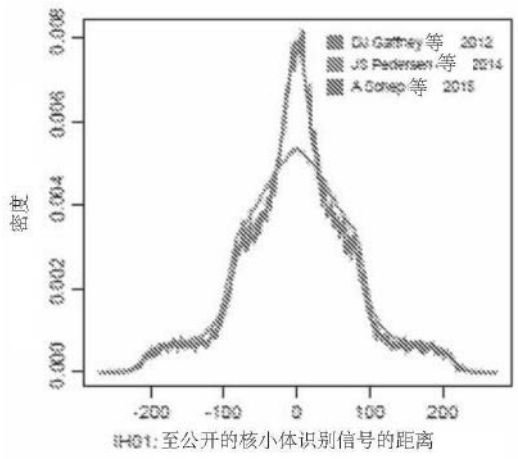


图63C

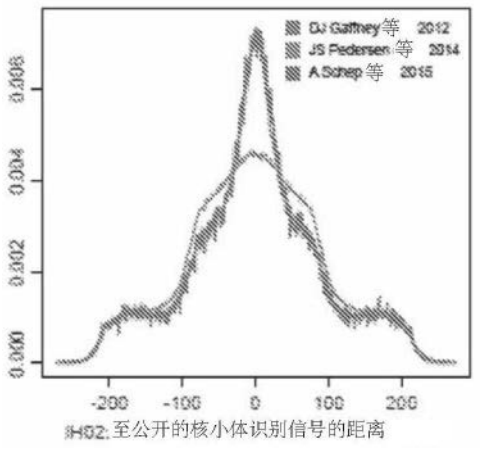


图63D

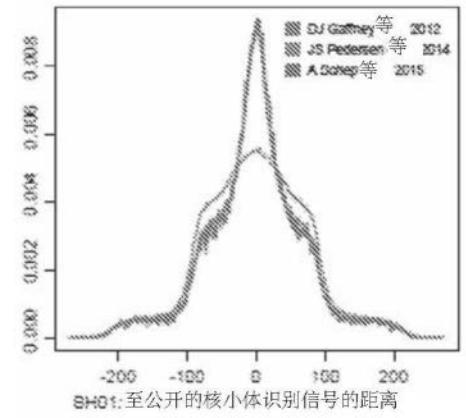


图63E

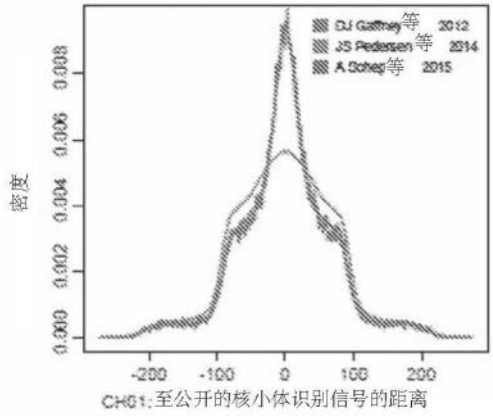


图63F

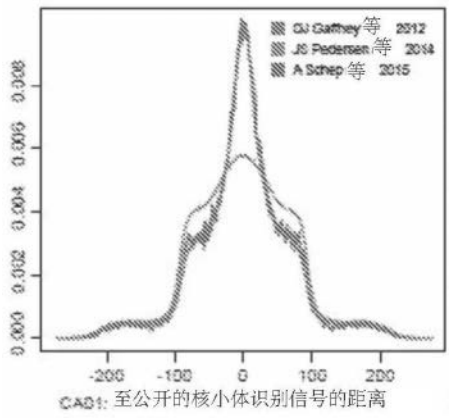

DJ Gaffney等 2012
JS Pedersen等 2014
A Schep等 2015
0.000
0.002
0.004
0.006
0.008
-200
-100
0
100
200
CA01: 至公开的核小体识别信号的距离

图63G

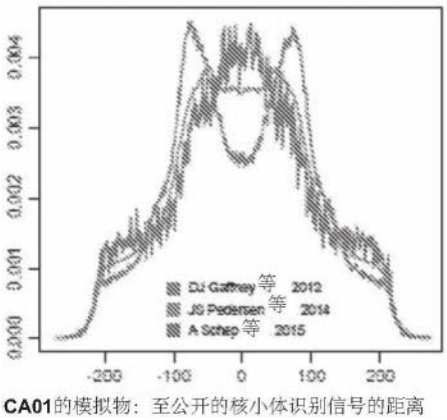

DJ Gaffney等 2012
JS Pedersen等 2014
A Schep等 2015
0.000
0.001
0.002
0.003
0.004
-200
-100
0
100
200
CA01的模拟物：至公开的核小体识别信号的距离

图63H