(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111394426 B
(45)授权公告日 2024.05.10

(21)申请号 202010205546.6

(22)申请日 2014.05.20

(65)同一申请的已公布的文献号
申请公布号 CN 111394426 A

(43)申请公布日 2020.07.10

(30)优先权数据
61/826,728 2013.05.23 US

(62)分案原申请数据
201480028601.1 2014.05.20

(73)专利权人 斯坦福大学托管董事会
地址 美国加利福尼亚州

(72)发明人 P·吉里西 J·D·比恩罗斯特罗 张元豪 W·J·格林里夫

(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 张小勇

(51)Int.Cl.
C12Q 1/6806(2018.01)
C12Q 1/6869(2018.01)
C12Q 1/6874(2018.01)
G16B 20/00(2019.01)
G16B 30/00(2019.01)
G16H 50/20(2018.01)

(56)对比文件
CN 1612931 A,2005.05.04
WO 2012106546 A2,2012.08.09
黄春等.转座子及其相关技术的研究.世界华人消化杂志.2006,(第17期),第1714-1720页.

审查员 白晓岩

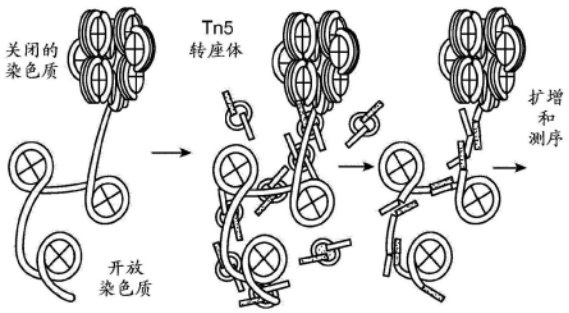
权利要求书4页 说明书32页
序列表6页 附图24页

(54)发明名称

用于个人表观基因组学的至天然染色质的转座

(57)摘要

本文提供了用于分析多核苷酸例如基因组DNA的方法。在某些实施方案中,所述方法包括:(a)用插入酶复合物处理分离自细胞群的染色质以产生基因组DNA的标记片段;(b)测序标记片段的一部分以产生多个序列读数;和(c)通过将获自序列读数的信息映射至细胞的基因组的区域而制作所述细胞的基因组的该区域的表观遗传图谱。还提供了用于执行所述方法的试剂盒。



CN 111394426 B

1.用于核酸处理或分析的非诊断目的的方法,其包括:

(a)裂解多个细胞以提供多个细胞核,其中所述多个细胞核包含染色质;及

(b)将所述多个细胞核的细胞核与插入酶复合物接触,使得所述细胞核的多核苷酸在开放染色质区域标签片段化,以产生多个标记片段,

其中所述插入酶复合物包含:

转座酶,

第一核酸插入元件,所述第一核酸插入元件包含第一衔接子序列,和

第二核酸插入元件,所述第二核酸插入元件包含第二衔接子序列,并且

其中所述标记片段中的一个或多个包含:

所述第一衔接子序列,和

所述第二衔接子序列。

2.根据权利要求1所述的方法,其还包括对所述标记片段进行一个或多个核酸反应以产生测序文库。

3.根据权利要求2所述的方法,其还包括对所述测序文库进行测序以产生多个序列读数。

4.根据权利要求2所述的方法,其中所述一个或多个核酸反应包括核酸扩增反应。

5.根据权利要求4所述的方法,

其中所述核酸扩增反应经配置以将一个或多个功能性序列添加至所述标记的核酸分子或其衍生物,

其中所述一个或多个功能性序列与所选的下一代测序平台兼容。

6.根据权利要求1所述的方法,其中所述多个标记片段中的标记片段包含对应于所述标记片段的核苷酸序列。

7.根据权利要求6所述的方法,其中所述标记片段还包含引物序列。

8.根据权利要求1所述的方法,其中所述插入酶复合物包含Tn5转座酶或来源于Tn5转座酶的转座酶。

9.根据权利要求1所述的方法,其中所述第一衔接子序列包含第一测序衔接子序列。

10.根据权利要求1所述的方法,其中所述第一衔接子序列包含条形码序列。

11.根据权利要求1所述的方法,其中所述第一衔接子序列包含第一引物序列。

12.根据权利要求1所述的方法,其中所述第二衔接子序列包含第二测序衔接子序列。

13.根据权利要求1所述的方法,其中所述第二衔接子序列包含条形码序列。

14.根据权利要求1所述的方法,其中所述第二衔接子序列包含第二引物序列。

15.根据权利要求3所述的方法,其还包括分析所述序列读数以产生沿所述细胞核的染色质可接近性的特征谱。

16.根据权利要求3所述的方法,其还包括分析所述序列读数以确定针对所述细胞核的核苷酸中一个或多个DNA结合蛋白结合位点的DNA结合蛋白的占据。

17.根据权利要求3所述的方法,其还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个转录起始位点的位置。

18.根据权利要求1所述的方法,其中所述插入酶复合物不含特异性针对作为染色质的一部分的蛋白的抗体。

19.根据权利要求3和15～17之任一项所述的方法,其还包括分析所述序列读数以产生代表所述细胞核的多核苷酸的一个或多个表观遗传特征的表观遗传图谱。

20.根据权利要求3和15～17之任一项所述的方法,其还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个核小体的位置。

21.用于核酸处理或分析的非诊断目的的方法,其包括:

(a)裂解多个细胞以分离多个细胞核,其中所述多个细胞核包含染色质;及

(b)将所述多个细胞核的细胞核与Tn5转座酶复合物接触,使得所述细胞核的多核苷酸在开放染色质区域标签片段化,以产生多个标记片段,

其中所述Tn5转座酶复合物包含第一测序衔接子序列和第二测序衔接子序列,

其中所述Tn5转座酶复合物不含特异性针对作为染色质的一部分的蛋白的抗体,并且

其中所述多个标记片段中的标记片段包含:

(i)对应于开放染色质区域的核苷酸序列,

(ii)所述第一测序衔接子序列,和

(iii)所述第二测序衔接子序列。

22.根据权利要求21所述的方法,其还包括对所述标记片段进行一个或多个核酸反应以产生测序文库。

23.根据权利要求22所述的方法,其还包括对所述测序文库进行测序以产生多个序列读数。

24.根据权利要求23所述的方法,其还包括分析所述序列读数以产生沿所述细胞核的染色质可接近性的特征谱。

25.根据权利要求23所述的方法,其还包括分析所述序列读数以确定针对所述细胞核的核苷酸中一个或多个DNA结合蛋白结合位点的DNA结合蛋白的占据。

26.根据权利要求23所述的方法,其还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个转录起始位点的位置。

27.根据权利要求23～26之任一项所述的方法,其还包括分析所述序列读数以产生代表所述细胞核的多核苷酸的一个或多个表观遗传特征的表观遗传图谱。

28.根据权利要求23～26之任一项所述的方法,其还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个核小体的位置。

29.插入酶复合物在制造用于核酸处理或分析的试剂盒中的用途,所述核酸处理或分析通过包括下列步骤的方法实施:

(a)裂解多个细胞以提供多个细胞核,其中所述多个细胞核包含染色质;及

(b)将所述多个细胞核的细胞核与插入酶复合物接触,使得所述细胞核的多核苷酸在开放染色质区域标签片段化,以产生多个标记片段,

其中所述插入酶复合物包含:

转座酶,

第一核酸插入元件,所述第一核酸插入元件包含第一衔接子序列,和

第二核酸插入元件,所述第二核酸插入元件包含第二衔接子序列,并且

其中所述标记片段中的一个或多个包含:

所述第一衔接子序列,和

所述第二衔接子序列。

30.根据权利要求29所述的用途,其中所述方法还包括对所述标记片段进行一个或多个核酸反应以产生测序文库。

31.根据权利要求30所述的用途,其中所述方法还包括对所述测序文库进行测序以产生多个序列读数。

32.根据权利要求30所述的用途,其中所述一个或多个核酸反应包括核酸扩增反应。

33.根据权利要求32的用途,

其中所述核酸扩增反应经配置以将一个或多个功能性序列添加至所述标记的核酸分子或其衍生物,

其中所述一个或多个功能性序列与所选的下一代测序平台兼容。

34.根据权利要求29所述的用途,其中所述多个标记片段中的标记片段包含对应于所述标记片段的核苷酸序列。

35.根据权利要求34所述的用途,其中所述标记片段还包含引物序列。

36.根据权利要求29所述的用途,其中所述插入酶复合物包含Tn5转座酶或来源于Tn5转座酶的转座酶。

37.根据权利要求29所述的用途,其中所述第一衔接子序列包含第一测序衔接子序列。

38.根据权利要求29所述的用途,其中所述第一衔接子序列包含条形码序列。

39.根据权利要求29所述的用途,其中所述第一衔接子序列包含第一引物序列。

40.根据权利要求29所述的用途,其中所述第二衔接子序列包含第二测序衔接子序列。

41.根据权利要求29所述的用途,其中所述第二衔接子序列包含条形码序列。

42.根据权利要求29所述的用途,其中所述第二衔接子序列包含第二引物序列。

43.根据权利要求31所述的用途,其中所述方法还包括分析所述序列读数以产生沿所述细胞核的染色质可接近性的特征谱。

44.根据权利要求31所述的用途,其中所述方法还包括分析所述序列读数以确定针对所述细胞核的核苷酸中一个或多个DNA结合蛋白结合位点的DNA结合蛋白的占据。

45.根据权利要求31所述的用途,其中所述方法还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个转录起始位点的位置。

46.根据权利要求29所述的用途,其中所述插入酶复合物不含特异性针对作为染色质的一部分的蛋白的抗体。

47.根据权利要求31和43～45之任一项所述的用途,其中所述方法还包括分析所述序列读数以产生代表所述细胞核的多核苷酸的一个或多个表观遗传特征的表观遗传图谱。

48.根据权利要求31和43～45之任一项所述的用途,其中所述方法还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个核小体的位置。

49.Tn5转座酶复合物在制造用于核酸处理或分析的试剂盒中的用途,所述核酸处理或分析通过包括下列步骤的方法实施:

(a)裂解多个细胞以分离多个细胞核,其中所述多个细胞核包含染色质;及

(b)将所述多个细胞核的细胞核与Tn5转座酶复合物接触,使得所述细胞核的多核苷酸在开放染色质区域标签片段化,以产生多个标记片段,

其中所述Tn5转座酶复合物包含第一测序衔接子序列和第二测序衔接子序列,

其中所述Tn5转座酶复合物不含特异性针对作为染色质的一部分的蛋白的抗体,并且

其中所述多个标记片段中的标记片段包含:

(i)对应于开放染色质区域的核苷酸序列,

(ii)所述第一测序衔接子序列,和

(iii)所述第二测序衔接子序列。

50.根据权利要求49所述的用途,其中所述方法还包括对所述标记片段进行一个或多个核酸反应以产生测序文库。

51.根据权利要求50所述的用途,其中所述方法还包括对所述测序文库进行测序以产生多个序列读数。

52.根据权利要求51所述的用途,其中所述方法还包括分析所述序列读数以产生沿所述细胞核的染色质可接近性的特征谱。

53.根据权利要求51所述的用途,其中所述方法还包括分析所述序列读数以确定针对所述细胞核的核苷酸中一个或多个DNA结合蛋白结合位点的DNA结合蛋白的占据。

54.根据权利要求51所述的用途,其中所述方法还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个转录起始位点的位置。

55.根据权利要求51～54之任一项所述的用途,其中所述方法还包括分析所述序列读数以产生代表所述细胞核的多核苷酸的一个或多个表观遗传特征的表观遗传图谱。

56.根据权利要求51～54之任一项所述的用途,其中所述方法还包括分析所述序列读数以确定所述细胞核的多核苷酸中的一个或多个核小体的位置。

用于个人表观基因组学的至天然染色质的转座

[0001] 本申请是中国专利申请CN201480028601.1的分案申请。

[0002] 政府支持

[0003] 本发明是在国立卫生研究院授予的合同AI057229、HG000044和NS073015下由政府支持作出的。政府具有本发明的某些权利。

[0004] 交叉引用

[0005] 本申请要求2013年5月23日提交的美国临时申请系列号61/826728的利益,该申请通过引用以其整体并入本文。

[0006] 背景

[0007] 真核生物基因组分层次地包装成染色质,并且此包装的性质在基因调控中起着中心作用。对编码在染色质的核蛋白结构中的表观遗传信息的主要认知来自于高通量的全基因组方法,其用于单独测定染色质可接近性(“开放染色质”)、核小体定位和转录因子(TF)占据。虽然存在已公开的方案,但这些方法需要数百万个细胞作为起始材料、复杂和费时的样品制备并不能同时探测核小体定位、染色质可接近性和TF结合的相互作用。这些限制在三个主要方面存在问题:第一,目前的方法可平均和“淹没”细胞群的异质性。第二,细胞通常必须离体生长以获得足够的生物材料,从而扰乱体内背景并且以未知的方式调节表观遗传状态。第三,输入要求通常会阻止这些测定应用于明确定义的临床样品,从而妨碍诊断时间尺度上“个人表观基因组学”的产生。本文提供的是可以克服这些限制的方法,其用于分析多核苷酸包括其可接近性及其结构。还提供的是单细胞方法,其可以提供较高的灵敏度和对染色质可接近性的进一步信息,包括细胞间变异性,以潜在地使其用作生物标志物。

[0008] 概述

[0009] 本文提供了用于分析多核苷酸例如基因组DNA的方法。在某些实施方案中,该方法包括:(a)用转座酶和分子标签处理分离自细胞群的染色质以产生多核苷酸的标记片段;(b)测序标记片段的一部分以产生多个序列读数;和(c)通过将获自序列读数的信息映射至细胞的基因组的区域而制作所述细胞的基因组的该区域的表观遗传图谱。

[0010] 在一些情况下,信息通过使用在序列读数的开头的核苷酸序列和任选末端上的核苷酸序列获得。在某些情况下,在(c)中映射的信息选自下列的一种或多种:(i)转座酶的切割位点;(ii)在步骤(a)中产生的片段的大小;(iii)序列读数长度;(iii)确定长度范围的序列读数的位置;和(iv)序列读数丰度。在一些情况下,确定大小范围的片段是无核小体的片段。

[0011] 在一些情况下,表观遗传图谱显示下列的一种或多种:(i)沿该区域的染色质可接近性的特征谱;(ii)该区域中结合位点的DNA结合蛋白的占据;(iii)该区域中的无核小体的DNA;(iv)沿该区域的核小体定位;和/或(v)染色质状态。在一些情况下,该方法还可包括测量DNA结合蛋白对于结合位点的总体占据。DNA结合蛋白可以例如是转录因子。

[0012] 在一些情况下,细胞群可以包括约500至100,000个细胞。细胞可以分离自个体,例如分离自该个体的血液。在一些实例中,细胞可以是相同的细胞类型。在一些实例中,细胞可以是FACS选择的细胞。

[0013] 在一些情况下,处理步骤(a)可以包括:从细胞群分离细胞核;和将分离的细胞核与插入酶复合物组合,其中所述组合导致细胞核裂解以释放染色质,以及导致产生基因组DNA的标记片段。在一些实例中,转座酶可来源于Tn5转座酶。在其它实例中,转座酶可来源于MuA转座酶。在进一步的实例中,转座酶可来源于Vibhar转座酶(例如来源于哈氏弧菌(Vibrio harveyi))。

[0014] 本公开内容还提供了用于比较两种样品的方法,其包括:(a)分析第一细胞群以产生第一表观遗传图谱;和(b)分析第二细胞群以产生第二表观遗传图谱;以及(c)比较第一表观遗传图谱与第二表观遗传图谱。例如,第一细胞群和第二细胞群可以是从相同个体在不同的时间收集的。或者,第一细胞群和第二细胞群可以是从不同个体收集的不同细胞群。

[0015] 本公开内容还提供了一种诊断方法,其包括:分析来自患者的染色质以产生表观遗传图谱;和基于表观遗传图谱提供诊断或预后。

[0016] 本公开内容提供了用于测定多核苷酸在某位点的可接近性的方法,其中所述多核苷酸来自细胞样品,所述方法包括:(a)用插入酶将多个分子标签插入多核苷酸;和(b)使用所述分子标签来测定所述位点上的可接近性。该方法还可包括使用所测定的可接近性来鉴定在所述位点上结合至多核苷酸的一种或多种蛋白。在一些情况下,所述蛋白的至少一种是转录因子。该方法还可包括使用分子标签来产生多核苷酸的可接近性图。

[0017] 本公开内容还提供了用于分析来自细胞样品的多核苷酸的三维结构的方法,包括:(a)用插入酶将多个分子标签插入多核苷酸;和(b)使用分子标签来分析所述多核苷酸的三维结构。在一些情况下,插入酶可包含两个或更多个酶部分,其中各个酶部分将共同的序列插入多核苷酸。酶部分可以连接在一起。共同的序列可包括共同的条形码。酶部分可包括转座酶。多核苷酸可以在步骤(a)过程中被分割成多个片段,其中包含共同的条形码的片段被测定为在多核苷酸的三维结构中是靠近的。

[0018] 多核苷酸可以在插入过程中被分割成多个片段。该方法还可包括扩增所述片段。可接近性可以通过对片段测序从而产生多个测序读数来测定。片段可以例如通过高通量测序技术测序。该方法还可包括基于插入酶的序列插入偏好标准化测序读数。测序读数的长度还可用于确定染色质状态注释(chromatin stae annotation)。

[0019] 细胞样品可以经透化以允许插入酶进入。在一些情况下,细胞样品中的细胞核在透化期间被最小限度地扰乱。细胞样品可以使用透化剂来透化,所述透化剂包括但不限于NP40、洋地黄皂苷、吐温、链球菌溶血素和/或阳离子脂质。细胞样品还可以使用低渗休克和/或超声处理来透化。

[0020] 该方法还可包括基于特定位点的可接近性来分析受试者中的疾病状态,其中细胞样品获自所述受试者。细胞样品和/或多核苷酸还可被划分成多个部分,其可任选地基于分子标签来划分。该方法还可包括分析细胞样品的表型。在一些情况下,表型可以与位点的可接近性相关。

[0021] 插入可以通过加入一种或多种二价阳离子来促进。在一些情况下,所述一种或多种二价阳离子可以包括镁。在一些情况下,所述一种或多种二价阳离子可以包括锰。

[0022] 细胞样品可获自原始来源。细胞样品可以由少于约500,000个细胞组成,或甚至是单个细胞。多核苷酸可以结合至多个关联分子。关联分子可以包括蛋白质,例如组蛋白。插入酶可以是转座酶。在一些情况下,转座酶可来源于Tn5转座酶。在其它情况下,转座酶可来

源于MuA转座酶。在其它情况下,转座酶可来源于Vibhar转座酶(例如来源于哈氏弧菌)。在一些情况下,分子标签可包含测序衔接子,其还可包含条形码标记。条形码标记可包括独特的序列。在其它情况下,分子标签可包括荧光标签。插入酶还可包含亲和标签,其可任选地是结合转录因子、修饰的核小体和/或修饰的核酸的抗体。修饰的核酸可以例如是甲基化或羟甲基化的DNA。亲和标签还可以是单链核酸,其可任选地结合至靶核酸。插入酶还可包含核定位信号。

[0023] 本公开内容还提供了组合物。该组合物可包含多核苷酸、插入酶和插入元件,其中:插入元件包括包含预先确定的序列的核酸;并且插入酶还包含亲和标签。组合物还可包含多核苷酸、插入酶和插入元件,其中:插入酶包含两个或更多个酶部分;并且所述酶部分连接在一起。亲和标签可以是抗体,其可任选地结合至转录因子、修饰的核小体和/或修饰的核酸。修饰的核酸可以是例如甲基化或羟甲基化的DNA。亲和标签还可以是单链核酸,其可任选地结合至靶核酸。插入元件可以结合至插入酶并且插入酶结合至多核苷酸。多核苷酸还可结合至多个关联分子。关联分子可以包括蛋白质,例如组蛋白。

[0024] 本公开还提供了试剂盒。该试剂盒可包含:(a)用于从细胞群分离细胞核的试剂;(b)插入酶复合物,和(c)转座酶反应缓冲液,在一些情况下,试剂盒的组分可以被配置为使得反应缓冲液、转座子标签和衔接子与细胞核的体外组合导致细胞核裂解以释放染色质,以及导致产生基因组DNA的标记片段。试剂盒还可包含:细胞裂解缓冲液;包含亲和标签的插入酶;和包含核酸的插入元件,其中所述核酸包含预先确定的序列。试剂盒还可包含:细胞裂解缓冲液;包含两个或更多个酶部分的插入酶,其中所述酶部分连接在一起;和(c)插入元件。亲和标签可以是抗体,其可任选地结合至转录因子、修饰的核小体和/或修饰的核酸。修饰的核酸可以是例如甲基化或羟甲基化的DNA。亲和标签还可以是单链核酸,其可任选地结合至靶核酸。

[0025] 本教导内容的这些和其它特征示于本文中。

[0026] 通过引用并入

[0027] 在本说明书中提及的所有出版物、专利和专利申请均通过引用并入本文,其程度如同每个单独的出版物、专利或专利申请被明确地和单独地指明通过引用并入。

[0028] 附图简述

[0029] 本领域技术人员将理解下文描述的附图仅用于举例说明的目的。附图并不意图以任何方式限制本教导内容的范围。

[0030] 图1A-1C:ATAC-seq是开放染色质状态的灵敏的、准确的探针。(a)ATAC-seq反应原理图。装载有测序衔接子(红色和蓝色)的转座酶(绿色)仅插入开放染色质(灰色的核小体)的区域,并生成可PCR扩增的测序文库片段。(b)开放染色质分析的全基因组方法的近似报告输入材料和样品制备时间要求。(c)在GM12878淋巴母细胞样细胞中的基因座上ATAC-seq与其它开放染色质测定的比较,显示高度一致性。靠下的ATAC-seq轨迹由500个FACS分选的细胞产生。

[0031] 图2A-2B:ATAC-seq提供染色质紧密态的全基因组信息。(a)从GM12878细胞核(红色)产生的ATAC-seq片段大小指示具有与核小体一致的空间频率的染色质依赖的周期性,以及与小于200bp片段的DNA螺旋螺距一致的高频率周期性。(插图)对数转换的直方图显示出明显的持续至6个核小体的周期性。(b)之前确定的7类染色质状态的标准化读数富集。

[0032] 图3A-3E:ATAC-seq提供了关于调节区域中核小体定位的全基因组信息。(a)含有两个转录起始位点(TSS)的示例基因座,显示无核小体读数轨迹、计算的核小体轨迹(“方法”)、以及用于比较的DNase、MNase和H3K27ac、H3K4me3以及H2A.Z轨迹。(b)针对所有活性TSS(n=64,836)所显示的ATAC-seq(1.98x 10^6个配对读数)和MNase-seq(来自ref 23的4x 10^9个单末端读数)核小体信号,TSS通过CAGE表达分选。(c)TSS针对无核小体的片段富集,并且在-2、-1、+1、+2、+3和+4位置上显示与MNase-seq所见的相似的定相核小体。(d)TSS和远端位点中核小体关联的相对于无核小体的(NFR)碱基的相对分数(见“方法”)。(e)可接近的染色质内相对于最近的核小体二分体的DNA结合因子位置的层次聚类揭示了不同类别的DNA结合因子。与核小体强烈关联的因子针对染色质重塑体富集。

[0033] 图4A-4C:ATAC-seq测定全基因组因子占据。(a)在chr1上的特定基因座上,在ATAC-seq和DNase-seq数据中观察到的CTCF足迹。(b)在基因组中的结合位点上产生的针对CTCF(所显示的基序)的集合ATAC-seq足迹。(c)从ATAC-seq数据、针对CTCF基序的位置权重矩阵(PWM)得分和进化保守性(PhyloP)推断的CTCF预测结合概率。最右边的列是针对该GM12878细胞系的CTCF ChIP-seq数据(ENCODE),显示出与预测的结合概率的高度一致性。

[0034] 图5A-5D:ATAC-seq实现实时的个人表观基因组学。(a)从标准抽血的工作流。(b)来自先证者T细胞的三天内连续ATAC-seq数据。(c)ATAC-seq数据(绿色轨迹)用于对候选TF药物靶标区分优先次序的应用实例。在鉴定的靠近细胞因子基因IL2(可以由FDA批准的药物靶向)的TF结合位点中,仅NFAT结合先证者T细胞。ATAC-seq足迹预测通过与公开的NFAT ChIP-seq数据(蓝色轨迹,来自ref[35]的数据)比对来确认。(d)与GM 12878B细胞系比较的来自先证者T细胞的细胞类型特异性调节网络。每行或列是TF相对于相同细胞类型中的所有其它TF的足迹特征谱。颜色表示T相对于B细胞的相对相似性(黄色)或差异性(蓝色)。NFAT是一个最高差异调节的TF(红色框),而典型CTCF结合在T细胞和B细胞中基本相似。

[0035] 图6:ATAC-seq峰强度与DNase-seq峰强度良好相关。Duke DNase-seq(向下采样至60x 10^6个读数)、UW DNase-seq(40x 10^6个读数)和ATAC-seq数据(60x 10^6个配对末端读数)中的峰使用ZINBA(Rashid等人Genome Biol.2011 12:R67)来调用。由于每个数据集有不同的读数长度,我们选择过滤可映射区域内的峰(Duke DNase-seq=20bp读数,UW DNase-Seq=36bp读数,ATAC-Seq=配对末端50bp读数)。对于(A)Duke DNase-seq和ATAC-seq,(B)UW DNase-seq和ATAC-seq,以及(C)UW DNAse-seq和Duke DNase-seq比较log10(读数强度)。ATAC-seq数据的技术重现性显示于D中。

[0036] 图7:ATAC-seq捕获DNase识别峰的很大一部分。对于所有数据集使用ZINBA调用峰。维恩图显示每个方法之间的峰调用重叠。下图:大多数ATAC-seq读数在与Duke和UW DNase-seq峰相交的强峰中。显示了从ATAC-seq、UW DNase-seq和Duke DNase-seq调用的峰内的读数的总分数,以及这些数据的交集。所有三种方法中超过65%的读数被发现在三种方法的峰的交集中,提示通过所有方法检测到良好定型的峰。表单元格颜色与读数的分数成比例。

[0037] 图8:相对于一组背景区域,与GM12878细胞中通过Duke DNase、UW DNase和FAIRE鉴定的一组开放染色质区域重叠的读数数量的图,其中对于检测开放染色质位点灵敏性和特异性所需的读数深度的测定在不同的读数深度上评估,包括50k、100k、500k、1x 10^7和5x 10^7个读数。底部图显示ATAC-seq在GM12878细胞中的表现通过使用500、5,000或50,000个

细胞作为起始材料来进行评估。

[0038] 图9:基因组DNA和染色质中的Tn5插入偏好。核苷酸频率得分代表针对每个碱基所观察到的核苷酸频率,将核苷酸频率针对1进行标准化。x＝0的位置表示读数开始,虚线表示Tn5二聚体的对称轴。我们在纯化的基因组DNA和人染色质之间没有看到Tn5插入偏好之间的实质性差别,这表明染色质中的局部插入偏好与裸基因组DNA中所发现的相同。所报告的这些序列偏好类似于以前报道过的那些(ref.11的正文)。

[0039] 图10:在每个ATAC-seq峰上各特征的每碱基平均强度的图;所有ENCODE ChIP数据针对输入进行标准化;数据已使用200个峰的滑动窗口进行了处理。

[0040] 图11:各种细胞数量的ATAC-seq。对于ATAC-seq来自不同起始数量细胞的数据的代表性UCSC基因组浏览器轨迹。此相同的基因组还显示于正文的图1b中。按顺序:使用FACS分离500个细胞,和通过从细胞培养物的简单稀释实现500个细胞和5,000个细胞的一式两份重复。为了比较,底部轨迹代表50,000个细胞,还显示于图1b中。此图证实,我们能够从少至500个细胞捕获开放染色质位点。

[0041] 图12:将核小体峰拟合至ATAC-seq片段大小分布以实现核小体占据测量。所观察到的片段分布被划分为四个读数群-预期源自开放DNA的读数,和跨越1、2或3个推定核小体的读数。为了实现数据的这种划分,将ATAC-seq片段分布拟合至下列的总和:1)低于1个核小体的插入片段大小上片段分布模式的指数函数和2)从一、二、三、四和五个核小体的保护产生的分布的5个高斯。所显示的这些拟合的总和(黑色虚线)类似于所观察到的片段分布(蓝线)。垂直虚线是鉴定为源自无核小体(<100bp)、1-核小体、2-核小体和3-核小体区域的片段的边界。虚线被设置为确保<10%的片段从邻近起源,如由我们的拟合所限定的。

[0042] 图13:GM 12878细胞中通过ATAC-seq检测的转录因子足迹的选择组。对于所指示的转录因子,使用CENTIPEDE在匹配对应基序的全基因组位点集上计算ATAC-seq读数的集合信号。在基序边界的区域+/-100bp中计算读数。垂直虚线指示基序的边界。

[0043] 图14:使用ATAC-seq和DNase足迹利用CENTIPEDE预测CTCF结合位点。CTCF结合位点的预测使用通过由CENTIPEDE报告的后验概率分选的全基因组CTCF基序集来评估。那些重叠CTCF ChIP-seq峰用作阳性集并且所有其它的被认为是阴性集。这产生0.92的曲线下面积(AUC),其提示CTCF的特异性和灵敏性结合推断。Duke DNase和UW DNase数据在相同的CENTIPEDE设置下使用,并显示了ROC图。ATAC-seq数据由198x 10^6个配对读数组成,Duke DNase包含245x 10^6个读数并且UW DNase包含48x 10^6个读数。

[0044] 图15:T细胞特异性NFAT调节:通过ATAC-seq预测并通过与NFAT ChIP-seq(来自ref 35正文的数据)比对确认的T细胞特异性NFAT靶基因的实例。

[0045] 图16:来自人血的FACS纯化的细胞群的ATAC-seq。(A)从标准抽血,我们使用荧光激活细胞分选(FACS)纯化CD4+T细胞、CD8+T细胞和CD14+单核细胞。每个群体产生成功的ATAC-seq数据(B)并揭示了已知的谱系特异性基因上的细胞类型特异性染色质开放位点。

[0046] 图17:使用ATAC-seq检测GM12878细胞中的等位基因特异性开放染色质。通过可公开获得的变体数据,我们测量了假定的杂合基因座上开放染色质区域中的等位基因频率。由于潜在的虚假杂合位点,我们需要多于两个读数来验证等位基因的杂合性。红点(n＝167)是$p<10^{-5}$的候选等位基因特异性开放染色质位点,而灰色(n＝900)代表P<0.01的候选物。使用由Audic等人(Genome Research 1997 7,986-995)开发的贝叶斯模型来计算P值。

[0047] 图18:转座酶可用作开放染色质染料。通过用荧光标记的DNA衔接子装载Tn5转座酶,以绿色显示的转座事件主要定位于细胞核,并表现出与高阶组织一致的点状图案。

[0048] 图19:相较于50,000个细胞,来自单个细胞核的单细胞ATAC-seq数据(蓝色)在全基因组开放染色质的预期位置上显示明显的峰。

[0049] 图20:单细胞插入片段长度分布与来自50,000个细胞的分布相匹配,显示因核小体存在的周期性。

[0050] 定义

[0051] 除非本文另外定义,否则本文使用的所有技术和科学术语具有与本发明所属领域中普通技术人员所通常理解的相同的含义。虽然类似或等同于本文描述的方法和材料的任何方法和材料可用于本发明的实践或测试,但对优选的方法和材料进行了描述。

[0052] 本文引用的所有专利和出版物,包括这些专利和出版物中公开的所有序列,明确地通过引用并入。

[0053] 数字范围包括限定该范围的数字。除非另外指出,否则分别地,核酸以5'至3'方向从左到右书写;氨基酸序列以氨基至羧基方向从左至右书写。

[0054] 本文提供的标题不是对本发明的各个方面或实施方案的限制。因此,下文即将定义的术语通过参考整个说明书会更加充分地定义。

[0055] 除非另有定义,否则本文使用的所有技术和科学术语具有与本发明所属领域中普通技术人员所通常理解的相同的含义。Singleton等人,DICTIONARY OF MICROBIOLOGY AND MOLECULAR BIOLOGY,2D ED.,John Wiley and Sons,New York(1994)以及Hale&Markham,THE HARPER COLLINS DICTIONARY OF BIOLOGY,Harper Perennial,N.Y.(1991)为技术人员提供了本文所用的许多术语的一般含义。尽管如此,为了清楚和便于参考的目的,在下文定义了某些术语。

[0056] 如本文所用的术语"样品"涉及材料或材料的混合物,其通常含有一种或多种目标分析物。在一个实施方案中,如在其最广泛的意义上使用该术语,是指含有DNA或RNA的任何植物、动物或病毒材料,例如,从个体分离的组织或液体(包括但不限于血浆,血清,脑脊髓液,淋巴,泪液,唾液和组织切片)或从体外细胞培养成分分离的组织或液体,以及来自环境的样品。

[0057] 如本文所用的术语"核酸样品"表示含有核酸的样品。本文所用的核酸样品可以是复杂的,因为它们包含多个不同的包含序列的分子。来自哺乳动物(例如小鼠或人)的基因组DNA样品是复杂样品的类型。复杂样品可具有超过约10^4、10^5、10^6或10^7、10^8、10^9或10^{10}个不同的核酸分子。DNA靶可源自任何来源例如基因组DNA或人工DNA构建体。本文可使用含有核酸的任何样品,例如来自组织培养细胞的基因组DNA或组织样品。

[0058] 如本文所用的术语"混合物"是指元素的组合,所述元素是散布的并且不处于任何特定的次序。混合物是异质性的并且不可空间分离成其不同的成分。元素的混合物的实例包括溶解于相同水溶液中的许多不同元素以及在随机位置上(即没有特定的次序)连接至固体支持物的许多不同的元素。混合物是不可寻址的。为了通过实例说明,如在本领域中通常已知的在空间上分离的表面结合的多核苷酸的阵列不是表面结合的多核苷酸的混合物,因为表面结合的多核苷酸的种类是空间上独特的并且阵列是可寻址的。

[0059] 术语"核苷酸"意欲包括不仅包含已知的嘌呤和嘧啶碱基还包含已被修饰的其它

杂环碱基的那些部分。这样的修饰包括甲基化的嘌呤或嘧啶、酰化的嘌呤或嘧啶、烷基化的核糖或其它杂环。此外,术语“核苷酸”包括含有半抗原或荧光标记的那些部分并且可不仅包含常规的核糖和脱氧核糖糖类还包含其它糖类。修饰的核苷或核苷酸还包括在糖部分上的修饰,例如其中一个或多个羟基被替换成卤素原子或脂族基团,或被官能化为醚、胺或类似的。

[0060] 术语“核酸”和“多核苷酸”在本文可互换使用来描述任何长度的聚合物,例如大于约2个碱基、大于约10个碱基、大于约100个碱基、大于约500个碱基、大于1000个碱基、大于10,000个碱基、大于100,000个碱基、大于约1,000,000、多至约10^{10}或更多碱基组成的核苷酸,例如脱氧核糖核苷酸或核糖核苷酸,并且可以酶促或合成产生(例如,如在美国专利号5,948,902和其中引用的参考文献中描述的PNA),其可以与天然存在的核酸以序列特异性方式(类似于两个天然存在的核酸的方式)杂交,例如可以参与Watson-Crick碱基配对相互作用。天然存在的核苷酸包括鸟嘌呤、胞嘧啶、腺嘌呤、胸腺嘧啶、尿嘧啶(分别地G、C、A、T和U)。DNA和RNA分别具有脱氧核糖和核糖的糖主链,而PNA的主链包括通过肽键连接的N-(2-氨基乙基)-甘氨酸重复单元。在PNA中各种嘌呤和嘧啶碱基通过亚甲基羰基键连接至主链。通常被称为不可接近RNA的锁核酸(LNA)是一种修饰的RNA核苷酸。LNA核苷酸的核糖部分被连接2’氧和4’碳的额外桥修饰。该桥将核糖“锁定”在3’-内型(North)构象,其常见于A-型双链体中。当需要时,可将LNA核苷酸与寡核苷酸中的DNA或RNA残基混合。术语“非结构化核酸”或“UNA”是包含以降低的稳定性彼此结合的非天然核苷酸的核酸。例如,非结构化核酸可以含有G’残基和C’残基,其中这些残基对应于非天然存在的形式,即G和C的类似物,其以降低的稳定性彼此碱基配对但保留分别与天然存在的C和G残基碱基配对的能力。非结构化核酸描述于US20050233340中,其对于UNA的公开内容通过引用并入本文。

[0061] 如本文所用的术语“寡核苷酸”表示约2至200个核苷酸、多至500个核苷酸长的核苷酸单链多聚体。寡核苷酸可以是合成的或者可以酶促制备,并且在一些实施方案中,为30至150个核苷酸长。寡核苷酸可以包含核糖核苷酸单体(即,可以是寡核糖核苷酸)或脱氧核糖核苷酸单体或核糖核苷酸单体和脱氧核糖核苷酸单体两者。例如,寡核苷酸可以是10至20、21至30、31至40、41至50、51至60、61至70、71至80、80至100、100至150或150至200个核苷酸长。

[0062] “引物”意指天然或合成的寡核苷酸,其能够在与多核苷酸模板形成双链体后用作核酸合成的起始点并从其3’末端沿着模板延伸以使得形成延伸的双链体。在延长过程中添加的核苷酸的序列由模板多核苷酸的序列确定。通常引物通过DNA聚合酶延伸。引物的长度通常与其在引物延伸产物合成中的使用兼容,并且通常在8至100个核苷酸的范围内,例如10至75、15至60、15至40、18至30、20至40、21至50、22至45、25至40等。典型的引物可以在10-50个核苷酸长的范围内,例如15-45、18-40、20-30、21-25等以及在所述范围之间的任何长度。在一些实施方案中,引物通常不超过约10,12,15,20,21,22,23,24,25,26,27,28,29,30,35,40,45,50,55,60,65或70个核苷酸长。

[0063] 引物通常是单链的以用于最大效率的扩增,但可选择地可以是双链的。如果是双链的,引物通常首先在用于制备延伸产物之前进行处理以分开其链。此变性步骤通常通过加热实现,但可选择地可以使用碱来进行,随后进行中和。因此,“引物”与模板互补,并通过氢键合或杂交与模板复合以产生引物/模板复合物用于起始通过聚合酶的合成,其通过在

其3'末端于DNA合成过程中互补于模板连接的共价键合的碱基的添加来延伸。

[0064] 术语“杂交”或“使杂交”是指其中核酸链的区域在正常杂交条件下退火并与第二互补核酸链形成稳定的双链体(无论是同源双链体或异源双链体),并且在相同的正常杂交条件下与不相关的核酸分子不形成稳定的双链体的过程。双链体的形成通过在杂交反应中退火两个互补核酸链区域来完成。杂交反应可以通过调整在其下发生杂交反应的杂交条件(通常称为杂交严格性)而成为高度特异性的,使得两条核酸链不会形成稳定的双链体,例如在正常严格条件下保持双链型区域的双链体,除非这两条核酸链包含基本上或完全互补的特定序列中的一定数量的核苷酸。“正常杂交或正常严格条件”对于任何给定的杂交反应可容易地确定。参见例如,Ausubel等人,Current Protocols in Molecular Biology,John Wiley&Sons,Inc.,New York或Sambrook等人,Molecular Cloning:A Laboratory Manual,Cold Spring Harbor Laboratory Press。如本文所用,术语“杂交的”或“杂交”是指核酸链通过碱基配对与互补链结合的任何过程。

[0065] 如果两个序列在中等至高严格性杂交和洗涤条件下彼此特异性杂交,则核酸被认为是与参考核酸序列“可选择性杂交的”。中等和高严格性杂交条件是已知的(参见例如,Ausubel等人,Short Protocols in Molecular Biology,3rd ed.,Wiley&Sons 1995以及Sambrook等人,Molecular Cloning:A Laboratory Manual,Third Edition,2001Cold Spring Harbor,N.Y.)。高严格条件的一个实例包括在约42℃在50％甲酰胺、5×SSC、5×Denhardt溶液、0.5％SDS和100μg/ml变性载体DNA中杂交随后在2X SSC和0.5％SDS中于室温洗涤两次和在0.1×SSC和0.5％SDS中在42℃下洗涤额外两次。

[0066] 如本文所用的术语“双链体”或“双链的”描述碱基配对即杂交在一起的两个互补的多核苷酸区域。

[0067] 如本文所用的术语“扩增”是指合成与模板核酸的一条或两条链互补的核酸分子的过程。扩增核酸分子可包括使模板核酸变性,在低于引物的解链温度的温度下将引物退火至模板核酸,以及从引物酶促延伸以产生扩增产物。变性、退火和延伸步骤各自可进行一次或多次。在某些情况下,变性、退火和延伸步骤进行多次,使得扩增产物的量增加,常常指数倍增,尽管指数扩增不是本方法所需的。扩增通常需要存在脱氧核苷三磷酸、DNA聚合酶和合适的缓冲液和/或用于聚合酶的最佳活性的辅因子。术语“扩增产物”是指从如本文所定义的扩增过程产生的核酸。

[0068] 术语“确定”、“测量”、“评估”、“评价”、“测定”和“分析”在本文可互换使用,其指任何形式的测量并且包括确定元素是否存在。这些术语包括定量和/或定性测定。评估可以是相对或绝对的。“评估…的存在”包括测定某物质的存在量以及确定其是存在或不存在。

[0069] 术语“使用”具有其常规含义,并因此,意指采用(例如使投入使用)方法或组合物以实现目的。例如,如果使用程序来创建文件,则执行程序以制作文件,该文件通常是该程序的输出。在另一个实例中,如果使用计算机文件,则其通常被存取、读取并且存储在该文件中的信息被用来实现目的。类似地,如果使用独特的标识符,例如条形码,则该独特的标识符通常被读取以鉴定例如与该独特的标识符相关联的对象或文件。

[0070] 如本文所用的术语“连接(ligating)”是指第一DNA分子的5'末端上的末端核苷酸与第二DNA分子的3'末端上的末端核苷酸的酶促催化连接。

[0071] “多个”包含至少2个成员。在某些情况下,“多个”可具有至少2个、至少5个、至少10

个、至少100个、至少100个、至少10,000个、至少100,000个、至少10^6个、至少10^7个、至少10^8个或至少10^9个或更多个成员。

[0072] 如果两个核酸是“互补的”,则它们在高严格条件下彼此杂交。术语“完全互补”用于描述其中一个核酸的每个碱基与另一个核酸中的互补核苷酸碱基配对的双链体。在许多情况下,互补的两个序列具有互补的至少10个例如至少12个或15个核苷酸。

[0073] “寡核苷酸结合位点”是指寡核苷酸在靶多核苷酸中杂交的位点。如果寡核苷酸“提供”针对引物的结合位点,则该引物可以杂交至该寡核苷酸或其互补体。

[0074] 如本文所用的术语“链”是指由通过共价键(例如磷酸二酯键)共价连接在一起的核苷酸构成的核酸。在细胞中,DNA通常以双链形式存在,并因此具有核酸的两条互补链,其在本文中称为“顶部”和“底部”链。在某些情况下,染色体区域的互补链可以被称为“正”和“负”链、“第一”和“第二”链、“编码”和“非编码”链、“沃森”和“克里克”链或“有义”和“反义”链。链作为顶部或底部链的分配是任意的,并不意味着任何特定的方向、功能或结构。几个示例性哺乳动物染色体区域(例如,BAC、组装体、染色体等)的第一链的核苷酸序列是已知的,并且可见于例如NCBI’s Genbank数据库。

[0075] 如本文所用的术语“顶部链”是指核酸的任一链但不是核酸的两条链。当寡核苷酸或引物结合或退火至“仅顶部链”时,其仅结合至一条链而不结合至另一条链。如本文所用的术语“底部链”是指与“顶部链”互补的链。当寡核苷酸结合或退火至“仅一条链”时,其仅结合至一条链例如第一或第二链,但不结合至另一条链。

[0076] 如本文所用的术语“测序”是指通过其获得对多核苷酸的至少10个连续核苷酸的识别(例如,识别至少20、至少50、至少100或至少200个或更多个连续核苷酸)的方法。

[0077] 术语“下一代测序”或“高通量测序”是指目前由Illumina、Life Technologies和Roche等采用的所谓并行合成测序或连接测序平台。下一代测序方法还可包括纳米孔测序方法或基于电子检测的方法,例如由Life Technologies商业化的Ion Torrent技术或由Pacific Biosciences商业化的基于单分子荧光的方法。

[0078] 如本文所用的术语“条形码序列”或“分子条形码”是指用于a)鉴定和/或示踪反应中多核苷酸的来源和/或b)对初始分子被测序的次数进行计数(例如,在其中样品中的基本上每个分子用不同的序列标记,然后将样品扩增的情况下)的核苷酸的独特序列。条形码序列可以是在寡核苷酸的5’末端、3’末端或在中间。条形码序列可在大小和组成上差别很大;下面的参考文献提供了用于选择适合用于具体实施方案的条形码序列集的指导:Brenner,美国专利号5,635,400;Brenner等人,Proc.Natl.Acad.Sci.,97:1665-1670(2000);Shoemaker等人,Nature Genetics,14:450-456(1996);Morris等人,欧洲专利申请0799897A1;Wallace,美国专利号5,981,179等。在具体的实施方案中,条形码序列可具有4至36个核苷酸或6至30个核苷酸或8至20个核苷酸范围内的长度。

[0079] 术语“体外”是指在具有分离的部件的容器中而不是在细胞中发生的反应。

[0080] 在沿着靶核酸分子的长度上分布的切割位点的上下文中,术语“分布”是指沿着靶核酸分子的长度上彼此间隔的插入。不需要所有插入以相同的量间隔开。相反,插入之间的间距可以是随机的、半随机的或不是随机的。

[0081] 如本文所用的术语“染色质”是指包含蛋白质和多核苷酸(例如DNA、RNA)的分子的复合物,如发现于真核细胞的细胞核中的。染色质部分地由形成核小体的组蛋白、基因组

DNA和通常结合至基因组DNA的其它DNA结合蛋白(例如转录因子)组成。

[0082] 如本文所用的术语“处理”是指在导致反应(例如切割)的条件(例如,合适的温度、时间和条件)下的组合。

[0083] 如本文所用的术语“分离自细胞群的染色质”是指被使得成为可用的染色质的来源。分离的细胞核(其可被裂解以产生染色质)以及分离的染色质(即,裂解的细胞核的产物)均被认为是分离自细胞群的染色质类型。

[0084] 如本文所用的术语“转录因子”是指可以自身地或与至少一种其它多肽组合地起作用以调节基因表达水平的任何多肽。该术语包括但不限于,直接结合DNA序列的多肽。转录因子可以增加或抑制表达水平。转录因子的实例包括但不限于Myc/Max,AP-1(Jun,Fos,ATF)CREB,SMAD,HIF,ETS,ERG,ELK,STAT,雌激素受体(ER),雄激素受体(AR),糖皮质激素受体(GR),孕激素受体(PR),NFκB,p53,OCT,SOX和PAX。转录因子可以是通过序列分析鉴定的转录因子,或是先前未被表征为转录因子的天然存在的阅读框序列。多肽还可以是人工产生的或经化学或酶修饰的多肽。

[0085] 如本文所用的术语“插入酶复合物(insertional enzyme complex)”是指包含插入酶和两个衔接分子(“转座子标签”)的复合物,其与多核苷酸组合以分割多核苷酸并将衔接子添加至多核苷酸。这样的系统描述于各种出版物中,包括Caruccio(Methods Mol.Biol.2011 733:241-55)和US20100120098,其以引用的方式并入本文。

[0086] 如本文所用的术语“标记片段”是指连接至标签的多核苷酸片段。

[0087] 如本文所用的术语“区域”是指生物体基因组中连续长度的核苷酸。染色体区域可以在1bp至整个染色体长度的范围内。在一些情况下,区域可具有至少200bp、至少500bp、至少为1kb、至少10kb或至少100kb或更多(例如,多至1Mb或10Mb或更多)的长度。基因组可来自任何真核生物,例如动物或植物基因组,例如人、猴、大鼠、鱼或昆虫的基因组。

[0088] 如本文所用的术语“表观遗传图谱”是指表观遗传特征的任何表示法,所述特征为例如核小体、无核小体区域的位点、转录因子的结合位点等。图谱可以物理展示在例如计算机显示器上。示例性表观遗传图谱显示于图1C、3A、4A、4B、5B和5C中。

[0089] 如本文所用的术语“映射信息”是指将实验获得的关于区域的信息组装至该区域的物理图谱。

[0090] 如本文所用的术语“序列读数丰度”是指特定序列或核苷酸在一批序列读数中观察到的次数。

[0091] 如本文所用的术语“无核小体的片段”是指相对贫乏或缺乏核小体(即核小体之间)的基因组DNA的片段。

[0092] 如本文所用的术语“染色质可接近性(chromatin accessibility)”是指多核苷酸例如基因组DNA内的核酸位点可接近的程度,即染色质“开放”的程度。与多肽关联的核酸位点,例如核小体中的基因组DNA,通常是不可接近的。未与多肽复合的核酸位点通常是可接近的,例如核小体之间的基因组DNA(除与转录因子和其它DNA结合蛋白复合的核酸位点以外)。

[0093] 如本文所用的术语“DNA结合蛋白的占据”是指针对序列特异性DNA结合蛋白的结合位点(例如,针对转录因子的结合位点)是否由DNA结合蛋白占据。DNA结合蛋白的占据可以定量或定性测量。

[0094] 如本文所用的术语“总体占据”是指是否有多个分布在整个基因组中的针对DNA结合蛋白的不同结合位点(例如,针对转录因子的结合位点)被DNA结合蛋白结合。DNA结合蛋白的占据可以定量或定性测量。

[0095] 如本文所用的术语“诊断”是指测定受试者是否患有特定疾病或病状。

[0096] 如本文所用的术语“预后”是指预测临床结果例如疾病复发、从疾病恢复、死亡,以及预测患有特定疾病或病状的受试者如何响应特定治疗。

[0097] 术语的其它定义可以出现在整个说明书中。

[0098] 示例性实施方案的描述

[0099] 在一个方面,提供了用于分析染色质的方法。在某些实施方案中,该方法包括:(a)用插入酶复合物处理分离自细胞群的染色质以产生基因组DNA的标记片段。在该步骤中,染色质通过使用插入酶来标签片段化(tagmented)(即,在相同的反应中切割和标记),所述插入酶为例如Tn5或MuA,其在染色质的开放区域中切割基因组DNA并将衔接子添加至片段的两个末端。用于标签片段化分离的基因组DNA的方法是本领域已知的(参见例如,CaruccioMethods Mol.Biol.2011 733:241-55;Kaper等人,Proc.Natl.Acad.Sci.2013 110:5552-7;Marine等人,Appl.Environ.Microbiol.2011 77:8071-9和US20100120098)并且可商购自Illumina(San Diego,CA)及其它供应商。这样的系统可容易地适用于本发明。在一些情况下,可以调整条件以获得染色质中插入的期望水平(例如,插入在开放区域中以平均每50至200个碱基对出现)。在该方法中使用的染色质可以通过任何合适的方法来制备。在一些实施方案中,可分离、裂解细胞核,并可进一步例如从核膜纯化染色质。在其它实施方案中,染色质可以通过将分离的细胞核与反应缓冲液接触来分离。在这些实施方案中,分离的细胞核在与反应缓冲液(其包含插入酶复合物和其它必需的试剂)接触时可裂解,这允许插入酶复合物接近染色质。在这些实施方案中,该方法可以包括从细胞群中分离细胞核;并将分离的细胞核与转座酶和衔接子组合,其中所述组合导致细胞核裂解以释放所述染色质,以及产生基因组DNA的加衔接子标签的片段。染色质不需要如其它方法(例如ChIP-SEQ方法)中的交联。

[0100] 在染色质已被分割和标记以产生基因组DNA的标记片段后,对至少一些加衔接子标签的片段进行测序以产生多个序列读数。所述片段可以使用任何方便的方法进行测序。例如,片段可以使用Illumina可逆终止法、Roche焦磷酸测序法(454)、Life Technologies连接测序(SOLiD平台)或Life Technologies Ion Torrent平台来进行测序。这样的方法的实例描述于下列参考文献中:Margulies等人(Nature 2005 437:376-80);Ronaghi等人(Analytical Biochemistry 1996 242:84-9);Shendure等人(Science 2005 309:1728-32);Imelfort等人(Brief Bioinform.2009 10:609-18);Fox等人(Methods Mol Biol.2009;553:79-108);Appleby等人(Methods Mol Biol.2009;513:19-39)和Morozova等人(Genomics.2008 92:255-64),其通过引用并入方法的一般描述和方法的具体步骤,包括所有的起始产物、文库制备的方法、试剂、以及每个步骤的最终产物。如将是显而易见的,与所选的下一代测序平台兼容的正向和反向测序引物位点可在扩增步骤中被添加至片段的末端。在某些实施方案中,片段可以使用杂交至已被添加至所述片段的标签的PCR引物来扩增,其中用于PCR的引物具有与特定测序平台兼容的5’尾。在某些情况下,所使用的引物可以含有分子条形码(“索引”),使得不同的库可以在测序之前合并在一起,并且序列读数

可以用该条形码序列追溯至特定样品。

[0101] 在另一个方面,本公开内容提供了用于测定多核苷酸在某位点的可接近性的方法,其中所述多核苷酸来自细胞样品,所述方法包括:用插入酶将多个分子标签插入多核苷酸和使用所述分子标签来测定所述位点上的可接近性。细胞样品可以来自原始来源。细胞样品可以由单个细胞组成。细胞样品可以由有限数量的细胞(例如少于约500,000个细胞)组成。

[0102] 该方法还可包括使用所测定的可接近性来鉴定在该位点结合至多核苷酸的一种或多种蛋白。在一些情况下,蛋白的至少一种是转录因子。此外,该方法可以包括使用分子标签来产生多核苷酸的可接近性图谱。

[0103] 多核苷酸在分子标签的插入过程中可被分割成多个片段。在一些情况下,片段可被扩增。在某些情况下,片段可被测序以产生多个测序读数。这可用来测定任何给定多核苷酸在某位点的可接近性。片段可以使用高通量测序技术进行测序。在一些情况下,测序读数可基于插入酶的序列插入偏好来标准化。测序读数的长度用于确定染色质状态注释。

[0104] 多核苷酸可结合至多个关联分子。关联分子可以是例如蛋白质、核酸或糖。在一些情况下,关联分子可包括组蛋白。在其它情况下,关联分子可包括适体。

[0105] 插入酶可以是能够将核酸序列插入多核苷酸的任何酶。在一些情况下,插入酶可以以基本上序列非依赖性方式将核酸序列插入多核苷酸。插入酶可以是原核生物的或真核生物的。插入酶的实例包括但不限于转座酶、HERMES和HIV整合酶。转座酶可以是Tn转座酶(例如Tn3,Tn5,Tn7,Tn10,Tn552,Tn903)、MuA转座酶、Vibhar转座酶(例如来自哈氏弧菌)、Ac-Ds、Ascot-1、Bs1、Cin4、Copia、En/Spm、F因子、hobo、Hsmar1、Hsmar2、IN(HIV)、IS1、IS2、IS3、IS4、IS5、IS6、IS10、IS21、IS30、IS50、IS51、IS150、IS256、IS407、IS427、IS630、IS903、IS911、IS982、IS1031、ISL2、L1、Mariner、P因子、Tam3、Tc1、Tc3、Tel、THE-1、Tn/O、TnA、Tn3、Tn5、Tn7、Tn10、Tn552、Tn903、Tol1、Tol2、Tn10、Ty1、任何原核转座酶或与上面所列的那些相关的和/或来源于其的任何转座酶。在某些情况下,与亲代转座酶相关的和/或来源于其的转座酶可以包括与亲代转座酶的相应肽片段具有至少约50%、约55%、约60%、约65%、约70%、约75%、约80%、约85%、约90%、约91%、约92%、约93%、约94%、约95%、约96%、约97%、约98%、或约99%氨基酸序列同源性的肽片段。肽片段的长度可以是至少约10、约15、约20、约25、约30、约35、约40、约45、约50、约60、约70、约80、约90、约100、约150、约200、约250、约300、约400、或约500个氨基酸。例如,来源于Tn5的转座酶可包含长度为50个氨基酸并且与亲代Tn5转座酶的相应片段约80%同源的肽片段。在一些情况下,插入可以通过添加一种或多种阳离子来促进和/或触发。阳离子可以是二价阳离子,例如,Ca^{2+}、Mg^{2+}和Mn^{2+}。

[0106] 分子标签可以包括测序衔接子、锁核酸(LNA)、拉链核酸(ZNA)、RNA、亲和反应分子(例如生物素、dig)、自身互补分子、硫代磷酸修饰、叠氮化物或炔基。在一些情况下,测序衔接子还可包括条形码标记。此外,条形码标记可包含独特的序列。独特的序列可用于鉴定个体插入事件。任何所述标签还可包括荧光标签(例如荧光素、罗丹明、Cy3、Cy5、噻唑橙等)。

[0107] 此外,插入酶还可包含亲和标签。在一些情况下,亲和标签可以是抗体。抗体可以结合至例如转录因子、修饰的核小体或修饰的核酸。修饰的核酸的实例包括但不限于甲基化或羟甲基化的DNA。在另一些情况下,亲和标签可以是单链核酸(例如ssDNA、ssRNA)。在一些实例中,单链核酸可结合于靶核酸。在其它情况下,插入酶还可包含核定位信号。

[0108] 在一些情况下,细胞样品可以经透化以允许插入酶进入。可以以最小限度地扰乱细胞样品中的细胞核的方式来进行透化。在一些情况下,细胞样品可以使用透剂来透化。透化剂的实例包括但不限于NP40、洋地黄皂苷、吐温、链球菌溶血素和阳离子脂质。在其它情况下,细胞样品可以使用低渗休克和/或超声处理来透化。在其它情况下,插入酶可以是带高电荷的,这可允许其通过细胞膜渗透化。

[0109] 在另一个方面,本公开内容提供了用于分析来自细胞样品的多核苷酸的三维结构的方法,包括:用插入酶将多个分子标签插入多核苷酸;和使用分子标签来分析所述多核苷酸的三维结构。插入酶可包含包含两个或更多个酶部分,其可任选地连接在一起。酶部分可以通过使用任何合适的化学合成或生物缀合方法来连接。例如,酶部分可以通过酯/酰胺键、巯基至马来酰亚胺的添加、天然化学连接(NCL)技术、点击化学(即炔-叠氮化物对)、或生物素-抗生蛋白链菌素对。在一些情况下,每个酶部分可将共同的序列插入多核苷酸。共同的序列可包含共同的条形码。酶部分可包括转座酶或其衍生物。在一些实施方案中,多核苷酸可在插入过程中被分割成多个片段。包含共同的条形码的片段可被测定为在多核苷酸的三维结构中是靠近的。

[0110] 多核苷酸可以是基因组DNA。多核苷酸还可结合至蛋白质例如组蛋白,并且可任选地包装在染色质的形式中。在特定情况下,对应于基因组的一个或多个区域(例如2个或更多、10个或更多、50个或更多、100个或更多、多至1000个或更多个区域)的DNA片段可以在测序之前通过杂交富集(即选择)。在这些实施方案中,不需要对整个文库测序。取决于所期望的结果和所选择的区域的长度(如果已进行了选择步骤),该方法的此步骤可导致至少1000个测序(例如,至少10,000、至少100,000、至少500,000、至少10^6、至少$5x10^6$、多至10^7或更多个测序读数)。序列读数通常存储在计算机存储器中。

[0111] 方法的一些实施方案包括制作细胞基因组的区域的表观遗传图谱。此步骤可通过将获自序列读数的信息映射至该区域来完成。在这些实施方案中,对序列读数进行计算分析以产生许多被映射到目标区域的表示法(例如,图形表示法)的数值输出。如将在下文更详细地解释的,可对许多类型的信息进行映射,包括但不限于:(i)转座酶的切割位点;(ii)在步骤a)中产生的片段的大小;(iii)片段长度;(iii)确定长度范围的序列读数的位置;和(iv)序列读数丰度。

[0112] 例如,可以对序列读数进行计算分析以鉴定片段的末端(从其可推断转座子切割位点)。在这些实施方案中,片段的一个末端可以通过位于测序读数的开头的序列限定并且片段的另一末端可以通过位于第二测序读数的开头的序列限定,其中第一和第二测序读数通过配对末端测序(例如,使用Illumina的测序平台)获得。相同的信息可以从检查较长序列读数的开头和末端(其理论上应具有两个衔接子的序列;一个在一个末端上,另一个在另一末端上)获得。在这些实施方案中,单个序列读数可包括两个衔接子序列,在此情况下可以从单个序列读数来推断片段的两个末端(其对应于两个单独的转座酶的两个切割位点)。片段的长度可以通过例如将片段末端映射至目标区域的核苷酸序列并对那些位置之间的碱基对数目计数来计算。信息可以通过使用在序列读数的开头和/或末端上的核苷酸序列获得。

[0113] 在某些情况下,可以将序列读数按长度分组。在一些实施方案中,一些序列可基于其大小被注释为无核小体的序列(即,被预测为在核小体之间的片段的序列)。还可鉴定与

单核小体、双核小体和三核小体关联的读数。这些截断值可使用图12中所示的数据来确定。片段长度(其提供与序列读数长度相同的信息)也可以以同样的方式处理。在某些情况下,可以计算序列读数丰度,即,基因组区域中的特定序列被表示在序列读数中的次数。

[0114] 所得的表观遗传图谱可提供目标区域中的染色质的分析。例如,取决于所映射的信息,图谱可以显示以下的一种或多种:沿区域的染色质可接近性的特征谱;区域中位点的DNA结合蛋白(例如,转录因子)占据;区域中的无核小体的DNA;沿区域的核小体定位;以及沿着区域的染色质状态的特征谱。在一些实施方案中,方法还可包括例如通过综合一个DNA结合蛋白在该蛋白结合的多个位点上的数据来测量DNA结合蛋白的结合位点的总体占据。在某些情况下,图谱还可以用序列信息以及关于该序列的信息(例如,启动子、内含子、外显子、已知的增强子、转录起始位点、非翻译区、终止子等的位置)来注释,以使得表观遗传信息可以在该注释的情境下查看。

[0115] 在某些实施方案中,表观遗传图谱可以提供关于活性调节区和/或结合至调节区的转录因子的信息。例如,核小体位置可以从所产生的测序读数的长度来推断。可选择地,转录因子结合位点可以从所产生的测序读数的大小、分布和/或位置来推断。在一些情况下,新的转录因子结合位点可以从所产生的测序读数推断。在其它情况下,新的转录因子可以从所产生的测序读数推断。

[0116] 在测定中使用的细胞群可以包括任何数目的细胞,例如约500至约10^6或更多的细胞、约500至约100,000个细胞、约500至约50,000个细胞、约500至约10,000个细胞、约50至1000个细胞、约1至500个细胞、约1至100个细胞、约1至50个细胞、或单个细胞。在一些情况下,细胞样品可以由少于约1000、约2000、约3000、约4000、约5000、约6000、约7000、约8000、约9000、约10,000、约15,000、约20,000、约25,000、约30,000、约40,000、约50,000、约60,000、约70,000、约80,000、约90,000、约100,000、约120,000、约140,000、约160,000、约180,000、约200,000、约250,000、约300,000、约350,000、约400,000、约450,000、约500,000、约600,000、约700,000、约800,000、约900,000、或约1,000,000个细胞组成。在其它情况下,细胞样品可以由多于约1000、约2000、约3000、约4000、约5000、约6000、约7000、约8000、约9000、约10,000、约15,000、约20,000、约25,000、约30,000、约40,000、约50,000、约60,000、约70,000、约80,000、约90,000、约100,000、约120,000、约140,000、约160,000、约180,000、约200,000、约250,000、约300,000、约350,000、约400,000、约450,000、约500,000、约600,000、约700,000、约800,000、约900,000、或约1,000,000个细胞组成。

[0117] 细胞可以来自任何来源。在某些情况下,细胞可以获自细胞的培养物,例如细胞系。在其它情况下,细胞可以分离自个体(例如,患者或类似的)。细胞可以分离自软组织或体液或体外生长的细胞培养物。在具体的实施方案中,染色质可以分离自软组织,例如脑、肾上腺、皮肤、肺、脾、肾、肝、脾、淋巴结、骨髓、膀胱胃、小肠、大肠或肌肉等。体液包括血液、血浆、唾液、粘液、痰、脑脊髓液、胸膜液、泪液、阴道管液(lactal duct fluid)、淋巴液、痰液、脑脊液、滑膜液、尿液、羊水和精液等。

[0118] 在一些实施方案中,在方法中使用的多核苷酸(例如基因组DNA,染色体DNA)可以来自血细胞,其中血细胞是指全血样品或全血中的细胞亚群。全血中的细胞亚群包括血小板、红血细胞(红细胞)、血小板和白血细胞(即外周血白细胞,其由嗜中性粒细胞、淋巴细胞、嗜酸性粒细胞、嗜碱性粒细胞和单核细胞构成)。这五种类型的白血细胞可以被进一步

分为两组:粒细胞(也称为多形核白细胞并且包括嗜中性粒细胞、嗜酸性粒细胞和嗜碱性粒细胞)和单核白细胞(其包括单核细胞和淋巴细胞)。淋巴细胞可进一步分为T细胞、B细胞和NK细胞。外周血细胞发现于血液循环池并且不隔离在淋巴系统、脾、肝或骨髓内。可分离存在于血液中的其它细胞。如果血液首先与试剂接触,然后将血液样品用于测定,那么部分或全部的接触血液可用于测定。

[0119] 在某些实施方案中,细胞样品可以直接分离自原始来源。例如,细胞样品可以直接分离自新鲜组织。在其它情况下,细胞样品可以直接分离自冷冻组织。在另一些情况下,细胞样品可以直接分离自经固定的组织。细胞样品的原始来源的其它实例包括但不限于,从组织解离的细胞、血液细胞、FFPE组织、细菌、病毒、线粒体、叶绿体、体外组装的蛋白DNA复合物、嗜中性粒细胞胞外陷阱(neutrophil extracellular trap)。

[0120] 通过使用本公开内容中提供的方法,可以基于获自受试者的细胞样品中的多核苷酸位点的可接近性来分析该受试者的疾病状态。例如,任何给定位点上的转录因子占据可导致位点上可接近性的缺乏。基于转录因子占据,受试者随后可以用合适的试剂(例如转录因子抑制剂)治疗。

[0121] 在某些情况下,细胞样品可进一步进行表型分析。例如,细胞样品可以使用荧光激活细胞分选(FACS)和/或激光捕获显微切割(LCM)来分析。在一些情况下,细胞样品和/或多核苷酸可以被划分成多个部分。可以基于分子标签(例如荧光标签)划分部分。在一些情况下,细胞样品和/或多核苷酸可以进行分选。分选可以在分子标签被插入多核苷酸之后进行。分选可以在片段测序之前进行。还可以用技术例如荧光原位杂交(FISH)分析细胞样品的基因转录。染色质可接近性可与表型、转录或翻译分析相关联。

[0122] 在一些实施方案中,细胞是相同的细胞类型。在这些实施方案中,细胞群可以通过MACS或FACS使用针对细胞表面标志物的标记抗体经由已知的方法从细胞的异质群体例如血液中选择。使用这些方法可以分离广泛多样的细胞,包括干细胞、癌症干细胞和血细胞的子集。在具体的实施方案中,可以从血液通过FACS或MACS分离下列细胞;T细胞($CD3^+CD4^+CD8^+$),B细胞($CD19^+CD20^+$),树突状细胞($CD11c^+CD20^+$),NK细胞($CD56^+$),干细胞/前体细胞($CD34^+$;仅造血干细胞),巨噬细胞/单核细胞($CD14^+CD33^+$),粒细胞($CD66b^+$),血小板($CD41^+CD61^+CD62^+$),红细胞($CD235a^+$),内皮细胞($CD146^+$)和上皮细胞($CD326^+$)。这些细胞的子集可以使用针对其它细胞表面标志物的抗体来分离。

[0123] 在一些实施方案中,方法可以用来比较两种样品。在这些实施方案中,方法可以包括使用上文所述方法分析第一细胞群,以产生第一表观遗传图谱;和使用上文所述方法分析第二细胞群,以产生第二表观遗传图谱;以及比较第一表观遗传图谱与第二表观遗传图谱,例如,以查看例如染色质开放性或转录因子占据是否有任何变化。

[0124] 在一些实施方案中,第一细胞群和第二细胞群是从相同个体在不同的时间收集的。在其它实施方案中,第一细胞群和第二细胞群是从组织或不同个体收集的不同细胞群。

[0125] 可用于所述方法的示例性细胞类型包括,例如分离自组织活检的细胞(例如来自具有疾病例如结肠癌、乳腺癌、前列腺癌、肺癌、皮肤癌或受病原体感染的组织等),以及来自相同组织例如来自同一患者的正常细胞;在组织培养物中生长的细胞,其是永生的(例如,具有增殖性突变或永生化转基因的细胞)、受病原体感染的、或经处理的(例如,用环境或化学剂如肽、激素,改变的温度、生长条件、物理应激、细胞转化等处理),以及正常细胞

(例如,除它们不是永生化的、未经感染或处理等之外,在其它方面与实验细胞相同的细胞);分离自患有癌症、疾病的哺乳动物、衰老的哺乳动物、或暴露于条件的哺乳动物的细胞,和来自相同物种例如来自同一家族的健康或年轻哺乳动物的细胞;以及来自相同哺乳动物的分化的细胞和未分化的细胞(例如,作为例如哺乳动物中其它细胞的祖先的一个细胞)。在一个实施方案中,可比较不同类型的细胞例如神经元和非神经元细胞,或不同状态的细胞(例如,在对细胞刺激之前和之后)。在另一个实施方案中,实验材料是易受病原体(例如病毒,例如人类免疫缺陷病毒(HIV)等)感染的细胞,并且对照材料是耐病原体感染的细胞。在本发明的另一个实施方案中,未分化的细胞例如干细胞和分化的细胞代表样品对。来自酵母、植物和动物例如鱼类、鸟类、爬行类、两栖类和哺乳动物的细胞用于本发明的方法。在某些实施方案中,可使用哺乳动物细胞,即来自小鼠、兔、灵长类动物、或人类、或其培养的衍生物的细胞。

[0126] 在一些示例性实施方案中,方法可用于鉴定测试试剂例如药物的效应,或者用于测定两种或更多种不同的测试试剂的效应是否存在差异。在这些实施方案中,可以制备两个或更多个相同的细胞群,并且取决于如何进行实验,细胞群的一个或多个可以用测试试剂温育确定的时期。在用测试试剂温育后,可使用上文所示的方法分析细胞群的染色质,并可对结果进行比较。在具体的实施方案中,细胞可以是血细胞,并且细胞可以用测试试剂离体温育。这些方法可用于确定测试试剂的作用方式,例如以鉴定染色质结构或转录因子占据响应于药物的变化。

[0127] 上文所述方法还可用作诊断法(该术语意欲包括提供诊断的方法以及提供预后的方法)。这些方法可以包括例如使用上文所述方法分析来自患者的染色质,以产生表观遗传图谱;和基于表观遗传图谱提供诊断或预后。

[0128] 本文所示的方法可用于提供针对与改变的染色质或DNA结合蛋白的占据相关的任何病况的可靠诊断。该方法可以应用于由表观遗传模式(例如,染色质可接近性或DNA结合蛋白的占据的模式)表征的病状的表征、分类、区分、分级、分期、诊断或预后。例如,该方法可用于确定来自怀疑受疾病或病状影响的个体的样品的表观遗传图谱相较于关于该疾病或病状被认为是“正常”的样品是否是相同或不同的。在具体的实施方案中,该方法可涉及诊断具有由测试样品中特定基因座上的表观遗传模式表征的病状的个体,其中该模式与病状相关联。该方法还可用于预测个体对病状的易感性。

[0129] 适合于使用本文所示的方法分析的示例性病状可以是例如,细胞增殖性病症或对细胞增殖病症的倾向;代谢功能失常或障碍;免疫功能失常、损伤或障碍;CNS功能失常、损伤或疾病;攻击症状或行为障碍;脑损伤的临床、心理和社会结果;精神障碍和人格障碍;痴呆或相关症状;心血管疾病、功能失常和损伤;胃肠道的功能失常、损伤或疾病;呼吸系统的功能失常、损伤或疾病;病变、炎症、感染、免疫和/或恢复期;作为发育过程中的异常的身体功能失常、损伤或疾病;皮肤、肌肉、结缔组织或骨的功能失常、损伤或疾病;内分泌和代谢功能失常、损伤或疾病;头痛或性功能失常,以及它们的组合。

[0130] 在一些实施方案中,该方法可以提供预后,例如以确定患者是否处于发生复发的风险。癌症复发是关于多种类型的癌症的担忧。预后方法可用于鉴定可能经历癌症复发的经手术治疗的患者,以便可给他们提供另外的治疗选择,包括术前或术后辅助系统,例如化学疗法、放射、生物调节剂和其它合适的疗法。该方法对于确定在检查或手术时没有显示可

测量的转移的患者的转移风险是特别有效的。

[0131] 该方法还可用于确定针对患有疾病或病状的患者(例如患有癌症的患者)的适当治疗过程。治疗过程是指在诊断后或在治疗后采取的针对患者的治疗措施。例如,对于复发、扩散或患者存活的可能性的确定可帮助确定是否应采取更保守或更激进的方法来治疗,或者治疗方式是否应组合。例如,当癌症可能复发时,可有利的是在手术治疗之前或之后进行化学疗法、放射、免疫疗法、生物调节剂疗法、基因疗法、疫苗等,或者调整患者治疗的时间跨度。

[0132] 在具体的实施方案中,实验室将接收来自远程位置(例如,医师办公室或医院)的样品(例如,血液),该实验室将如上所述分析样品中的细胞以产生数据,并且该数据可转移至远程位置用于分析。

[0133] 组合物

[0134] 在一个方面,本公开内容提供了与本文提供的方法相关的组合物。组合物可以包含多核苷酸、插入酶和插入元件,其中:插入元件可包括包含预先确定的序列的核酸并且插入酶还可包含亲和标签。多核苷酸还可结合至多个关联分子。关联分子可以是蛋白质(例如组蛋白)或核酸(例如适体)。亲和标签可以是抗体。在一些情况下,抗体可以结合至转录因子。在其它情况下,抗体可以结合至修饰的核小体。在其它情况下,抗体可以结合至修饰的核酸。修饰的核酸的实例包括但不限于甲基化或羟甲基化的DNA。亲和标签还可以是单链核酸(例如ssDNA,ssRNA)。在一些情况下,单链核酸可结合于靶核酸。在一些情况下,插入酶还可包含核定位信号。

[0135] 组合物可以包含多核苷酸、插入酶和插入元件,其中:插入酶包含两个或更多个酶部分并且酶部分连接在一起。插入元件可以结合至插入酶。插入酶还可结合至多核苷酸。在一些情况下,多核苷酸还可结合至多个关联分子。关联分子可以是蛋白质(例如组蛋白)或核酸(例如适体)。

[0136] 试剂盒

[0137] 在另一个方面,本公开内容提供了包含如上所述的用于实施本发明方法的试剂的试剂盒。本发明的试剂盒可以包含:(a)用于从细胞群分离细胞核的试剂;(b)转座酶和转座子标签,和(c)转座酶反应缓冲液,其中试剂盒的组分被配置为使得反应缓冲液、转座酶和衔接子与细胞核的体外组合导致细胞核裂解以释放染色质,以及产生基因组DNA的加衔接子标签的片段。

[0138] 在一些情况下,试剂盒可以包含:(a)细胞裂解缓冲液;(b)包含亲和标签的插入酶;和(c)包含核酸的插入元件,其中所述核酸包含预先确定的序列。插入酶可以是例如转座酶。插入酶还可包含连接在一起的两个或更多个酶部分。在一些情况下,亲和标签可以是抗体。抗体可结合至转录因子、修饰的核小体或修饰的核酸。修饰的核酸的实例包括但不限于甲基化或羟甲基化的DNA。在另一些情况下,亲和标签可以是单链核酸(例如ssDNA,ssRNA)。

[0139] 试剂盒可任选地含有其它成分,例如:如上所述的PCR引物、PCR试剂如聚合酶、缓冲液、核苷酸等。根据需要,试剂盒的各个组分可以存在于分开的容器中或者某些相容成分可以预组合在单个容器中。

[0140] 除了上述组分,本发明的试剂盒还可包含使用试剂盒组分以实施本发明方法的说

明书,即样品分析说明书。用于实施本发明方法的说明书通常记录在合适的记录介质上。例如,说明书可以被打印在基质上,例如纸或塑料等。如此,说明书可存在于试剂盒中作为包装说明书,存在于试剂盒或其组分的容器的标记中(即与包装或分装关联)等。在其它实施方案中,说明书以存在于合适的计算机可读存储介质(例如,CD-ROM、磁盘等)上的电子存储数据文件存在。在其它实施方案中,实际的说明书不存在于试剂盒中,但提供用于从远程来源例如经由互联网获得说明书的装置。此实施方案的实例是包含网址的试剂盒,通过该网址可以查看说明书和/或下载说明书。与说明书一样,用于获得说明书的该装置被记录在合适的基质上。

[0141] 实施方案

[0142] 提供了映射染色质的方法。在一些实施方案中,该方法包括以下步骤:用将测序衔接子插入染色质内的多核苷酸中的转座酶分割稀少或大量细胞的染色质,和扩增并测序所述片段以产生细胞特异性图谱。

[0143] 在某些实施方案中,细胞特异性图谱提供关于活性调节区和结合至所述调节区的转录因子的信息。

[0144] 在某些实施方案中,所述稀少细胞的数目介于1和100,000之间。

[0145] 在某些实施方案中,转座酶来源于Tn5转座酶。

[0146] 在某些实施方案中,转座酶来源于MuA转座酶。

[0147] 在某些实施方案中,从所产生的测序读数的长度推断核小体位置。

[0148] 在某些实施方案中,从所产生的测序读数的长度推断转录因子结合位点。

[0149] 在某些实施方案中,染色质直接分离自新鲜组织。

[0150] 在某些实施方案中,染色质直接分离自冷冻组织。

[0151] 在某些实施方案中,染色质直接分离自经固定的组织。

[0152] 在某些实施方案中,对于多路复用,将序列添加至测序衔接子以独特地鉴定片段(加条形码)。

[0153] 在某些实施方案中,将亲和标签用于将转座酶靶向至特定的目标大分子。

[0154] 在某些实施方案中,对于多路复用,将序列添加至测序衔接子以独特地鉴定片段(加条形码),并且将亲和标签用于将转座酶靶向至特定的目标大分子。

[0155] 在某些实施方案中,亲和标签是靶向至转录因子的抗体。

[0156] 在某些实施方案中,亲和标签是靶向至修饰的核小体的抗体。

[0157] 在某些实施方案中,特定基因组基因座上的插入片段大小分布用于推断染色质开放性。

[0158] 在某些实施方案中,插入片段大小分布和插入的位置用于推断转录因子结合。

[0159] 在某些实施方案中,获得的测序读数的数量通过所测量的转座酶的序列插入偏好标准化。

[0160] 在某些实施方案中,新的转录因子结合位点从所产生的测序读数推断。

[0161] 在某些实施方案中,新的转录因子从所产生的测序读数推断。

[0162] 在某些实施方案中,因果性变体可以通过查看测序读数的等位基因特异性产生来推断。

[0163] 在某些实施方案中,染色质状态注释从测序读数长度的分布推断。

实施例

[0164] 本教导内容的各方面可根据下列实施例进一步理解,下列实施例不应被解释为以任何方式限制本教导内容的范围。

[0165] 实施例1:使用测序对转座酶可接近的染色质的测定(ATAC-seq)

[0166] 本文描述了使用测序对转座酶可接近的染色质的测定(ATAC-seq)-基于测序衔接子向天然染色质中的直接体外转座-作为用于综合表观基因组分析的快速和灵敏的方法。ATAC-seq使用简单的2步方案从500至50,000个细胞捕获开放染色质位点,并揭示开放染色质的基因组位置、DNA结合蛋白、个体核小体以及调节区域上的高阶紧密态与核苷酸解析之间的相互作用。发现了严格避免、可容忍或倾向于与核小体重叠的DNA结合因子种类。通过使用ATAC-seq,从先证者经由标准抽血对静息人T细胞的连续每日表观基因组进行了测量和评价,显示出在临床时间尺度上阅读个人表观基因组用于监测健康和疾病的可行性。

[0167] 材料和方法

[0168] ATAC-seq方案的示例性实现具有三个主要步骤:

[0169] 1)制备细胞核:为了制备细胞核,将50,000个细胞在500x g下离心5分钟,随后使用50μL冷的1x PBS洗涤并在500x g下离心5分钟。将细胞用冷的裂解缓冲液(10mM Tris-Cl,pH 7.4,10mMNaCl,3mM $MgCl_2$和0.1%IGEPAL CA-630)裂解。裂解后立即使用冷冻离心机将细胞核在500x g下离心10分钟。为了避免在细胞核制备过程中丢失细胞,使用固定角离心机并且在离心后将它们从沉淀物小心地吸出。

[0170] 2)转座和纯化:在细胞核制备后立即将沉淀重悬于转座酶反应混合物(25μL 2x TD缓冲液,2.5μL转座酶(Illumina)和22.5μL不含核酸酶的水)。转座反应在37℃下进行30分钟。在转座后直接使用QiagenMinelute试剂盒纯化样品。

[0171] 3)PCR:在纯化后,我们使用1x NEBnext PCR预混物和1.25μM的定制Nextera PCR引物1和2(见下表)来扩增文库片段,使用下面的PCR条件:72℃下5分钟,98℃下30秒,随后在98℃下10秒、63℃下30秒和72℃下1分钟进行热循环。为了减少PCR中的GC和大小偏倚,使用qPCR监测PCR反应以在饱和之前停止扩增。为此,将整个文库扩增5个循环,在5个循环之后取出PCR反应的等分试样并以0.6x的终浓度加入至具有Sybr Green的10μl PCR混合物。我们将该反应运行20个循环,以测定其余45μL反应所需的额外循环数。使用Qiagen PCR净化试剂盒纯化文库,产生在20μL中~30nM的最终文库浓度。将文库扩增总共10-12个循环。

[0172]

Ad1_noMX:	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG (SEQ ID NO:1)
Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:2)
Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:3)
Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:4)
Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:5)
Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:6)
Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:7)
Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:8)
Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:9)
Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:10)
Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:11)
Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:12)
Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:13)

[0173]

Ad2.13_GTCGTGAT	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:14)
Ad2.14_ACCACTGT	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:15)
Ad2.15_TGGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:16)
Ad2.16_CCGTTTGT	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:17)
Ad2.17_TGCTGGGT	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:18)
Ad2.18_GAGGGGTT	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:19)
Ad2.19_AGGTTGGG	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:20)
Ad2.20_GTGTGGTG	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:21)
Ad2.21_TGGGTTTC	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:22)
Ad2.22_TGGTCACA	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:23)
Ad2.23_TTGACCCT	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:24)
Ad2.24_CCACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT (SEQ ID NO:25)

[0174] 低细胞数目的方案:为了制备500和5,000细胞反应,使用相同的方案,除了一些显著的例外:转座反应在5μL而非50μL的反应物中进行。另外,不进行PCR之前的Qiagen Minelute纯化而替代地在转座后立即取出该5μL反应物直接加入50μL PCR。

[0175] 文库QC和定量:在ATAC-seq方案中,避免了大小选择步骤以最大化文库复杂性。测序的插入片段大小分布在40bp至1kb之间,平均值为~120bp。从生物分析仪和凝胶,我们观察到>2kb的片段,这会使得Qubit和其它基于质量的定量方法难以解释。为此,我们使用基于qPCR的方法定量我们的文库。

[0176] 从外周血富集$CD4^+$:在斯坦福大学IRB批准的协议下从1个正常志愿者在72小时的时间内三次获得一绿顶管的全血。获得了知情同意书。使用RosetteSep人$CD4^+$T细胞富集混合物(StemCell Technology),对每个时间点的5mL血液针对CD4+细胞进行阴性选择。将RosetteSep混合物与血液在50μL/mL下温育20分钟,在等体积的具有2%FBS的PBS中稀释,并置于15mL Ficol-PaquePlus(GE)上。将血液在1200x g下不间断地离心20分钟,从密度介质:血浆界面移出阴性选择的细胞,并将细胞在具有2%FBS的PBS中洗涤2次。

[0177] FACS分选外周血白细胞和GM细胞:将GM 12878细胞用DAPI NucBlue固定细胞染料(分子探针)染色并使用FACSAria(BD Biosciences)使用100μm管口分选活细胞。一个外周血样品(血沉棕黄层)用BD Bioscience抗体CD14-A-488(M5E2,1:20)、CD3-PE-Cy7(SK7,1:20)、CD4-APC-Cy7的(RPA-T4,1:20)和CD8(RPA-T8,1:20)在室温下于黑暗中染色20分钟。将细胞用以1:10稀释于diH_2O中的BDpharmLyse(BD)裂解15分钟,离心5分钟,用PBS 2%FBS洗涤两次,并重悬于具有2%FBS的PBS中。将50,000个$CD3^+CD8^+$、$CD3^+CD4^+$和$CD14^+$细胞群分选至具有10%FBS的PBS中。

[0178] 数据分析

[0179] 原始数据处理:使用来自MiSeq的34×8×34读数或HiSeq上的50×8×50读数收集数据。使用BOWTIE(Langmead等人Genome Biol.2009 10,R25)采用参数-X2000和-m1将读数比对至hg19。这些参数确保允许比对多至2kb的片段(-X2000)并且仅收集独特的比对读数(-m1)。对于所有的数据文件使用Picard去除重复项。

[0180] 对于峰识别和足迹法,将读数起始位点调整为表示转座子结合事件的中心。Tn5转座酶的先前描述显示转座子以二聚体结合并插入间隔9bp的两个衔接子(Adey,A.等人Genome Biol201011:R119)。因此,比对至+链的所有读数偏移+4bp,并且比对至-链的所有读数偏移-5bp。

[0181] ATAC-seq峰识别:我们使用ZINBA来调用本文中所报告的所有ATAC-seq峰。ZINBA使用300bp的窗口大小和75bp偏移来运行。将可比对性用于对背景和富集组分的零膨胀组分和ATAC-seq读数计数进行建模。富集区域被鉴定为具有后验概率>0.8的那些区域。

[0182] 染色质注解内的ATAC-seq插入片段大小富集分析:首先计算重叠每个染色质状态的配对末端测序片段大小的分布(参见ensemble.org网站)。随后将分布标准化至每个状态内的最大百分比并相对于全基因组的片段大小集合计算富集。

[0183] 核小体定位:为了产生核小体位置数据轨迹,我们选择将读数分解为多个箱(bin)。低于100bp的读数被认为是无核小体的,180和247bp之间的读数被认为是单核小体,315和473bp之间的读数被认为是双核小体并且558和615bp之间的读数被认为是三核小体(测定截断值见图12)。双核小体读数被分解为两个读数,三核小体读数被分解为三个读数。读数使用Danpos和Dantools利用参数-p 1、-a 1、-d 20、-clonalcut 0进行分析。所用的背景是无核小体读数(小于100bp的读数),从而允许这些读数的有效负加权。此分析允许调用多个重叠核小体。虽然使用简单的插入片段大小截断值来产生核小体轨迹可能因其它的核小体大小特征即增强体(enhanaceosome)而产生假阳性,但我们观察到我们如实地概括了全基因组核小体位置的总体特征。

[0184] ChIP-seq峰调用和聚类:ChIP-seq数据下载自UCSC ENCODE库。使用GEM调用峰(Guo等人,PLoSComput.Biol.2012 8:e1002638),所用的参数是-k_min 6-k_max 20。输入

用作峰调用的对照。结合事件通过与10bp箱中最近的二分体的距离来注释。随后使用欧几里得距离法对因子层次聚类,并通过基因标准化和通过平均值集中。(Eisen等人Proc.Natl.Acad.Sci.1998 95:14863-14868)。

[0185] 使用CENTIPEDE的足迹法:全基因组基序集获自ENCODE基序库(在broadinstitute.org的网站上)。针对CENTIPEDE的输入包括匹配基序的每个基因组区域的+/-100bp内的PWM得分、保守性(PhyloP)和ATAC-seq计数。ChIP-seq数据获自UCSC ENCODE库。

[0186] 转录因子调节网的比较:通过比较GENCODE v14基因与通过CENTIPEDE对于各细胞类型评估的全基因组后验概率集来构建转录因子调节网。转录因子调节每个基因的程度通过对映射至相同染色体的给定转录因子的加权后验概率取总和来确定。对于每个映射的基序,基于与每个基因的转录起始位点的距离来加权后验概率。转录因子调节网的比较被计算为给定细胞类型中的每个转录因子与另一细胞类型中的所有转录因子的关联性。所得的关联矩阵使用Pearson相关系数和完全连锁来层次聚类。

[0187] 候选IL2增强子分析:对UCSC基因组浏览器上的ENCODE数据进行检查以鉴定可响应于FDA批准的免疫调节药物的一种或多种细胞类型中的推定IL2增强子。针对(i)增强子相关的组蛋白标记(H3K4me1和H3K27ac)、(ii)如通过ChIP-seq确认的一种或多种TF的结合、和(iii)可由人治疗剂靶向的TF途径,我们扫描了hg19中IL2上游的基因间区域。此分析鉴定了IRF4和STAT3结合位点以及已知的NFAT响应性元件。

[0188] 结果

[0189] ATAC-seq用转座子探测染色质可接近性

[0190] 体外装载了用于高通量DNA测序的衔接子的高活性的Tn5转座酶(Goryshin,J Biol Chem.1998 273:7367-7374;Adey,A.等人GenomeBiol 2010 11:R119)可以同时分割基因组并用测序衔接子标记基因组(前文描述为“标签片段化”)。据推测,在少量未固定的真核细胞核上通过纯化的Tn5(一种原核转座酶)的转座会询问可接近的染色质的区域。描述了对于转座酶可接近的染色质的测定和随后的高通量测序(ATAC-seq)。ATAC-seq使用Tn5转座酶来将其衔接子负载整合至可接近的染色质的区域,而空间位阻较不可接近的染色质使得转座较不可能发生。因此,适合于高通量测序的可扩增DNA片段优选在开放染色质的位置上产生(图1a)。整个测定和文库构建可以在包括Tn5插入和PCR的简单两步过程中进行。与此相反,公开的用于测定染色质可接近性的DNase-和FAIRE-seq方案包括多步骤方案和许多潜在的损失多发步骤,例如衔接子连接、凝胶纯化和交联逆转。例如,公开的DNase-seq方案要求约44个步骤和两次过夜温育,而公开的FAIRE-seq方案需要在至少3天内进行两次过夜温育。此外,这些方案需要1-50x 10^6个细胞(FAIRE)或5x 10^7个细胞(DNase-seq),可能是因为这些复杂的工作流(图1b)。相较于已建立的方法,ATAC-seq能够实现快速和有效的文库生成,因为测定和文库制备在单个酶促步骤中进行。

[0191] 深入的分析表明,ATAC-seq提供了全基因组染色质可接近性的准确和灵敏测量。ATAC-seq在分离自GM12878淋巴母细胞系的50,000和500个未经固定的细胞核上进行(ENCODE Tier 1)用于与染色质可接近性数据集(包括DNase-seq和FAIRE-seq)比较和验证。在先前由他人着重标示的基因座上(图1c),ATAC-seq具有类似于从多约3至5个数量级的细胞产生的DNase-seq的信噪比。峰强度在技术性重复之间是高度可重现的(R=0.98),

并且在ATAC-seq和DNase-seq之间是高度相关的(R＝0.79和R＝0.83,图6),并且注意到大部分的峰内读数来自DNase和ATAC-seq峰的交集(图7)。通过将我们的数据与ENCODE DNase-seq数据中鉴定的DHS进行比较,受试者工作特征(ROC)曲线显示与DNase-seq相似的灵敏度和特异度(图8)。还注意到,ATAC-seq峰强度与活性染色质的标志物良好相关,但不与转座酶序列偏好良好相关(图9和10)。高度灵敏的开放染色质检测即使在使用5,000或500个人细胞核作为起始材料时仍得到维持(图8和图11),尽管在所用的条件下对于较少量的输入材料灵敏度减小,如可见于图1c。

[0192] ATAC-seq插入片段大小揭示核小体位置

[0193] 发现ATAC-seq配对末端读数产生关于核小体包装和定位的详细信息。来自人类染色体的测序片段的插入片段大小分布具有约200碱基对的明显周期性,提示许多片段受到整数倍核小体的保护(图2a)。该片段大小分布还显示出等于DNA的螺旋间距的明显周期性。通过根据由以前的模型(Hoffman等人Nucleic Acids Res.2013 41:827-841)所定义的染色质功能种类划分插入片段大小分布,并标准化至总体插入片段分布,我们观察在此插入片段大小分布上明显的种类特异性富集(图2b),这表明染色质的这些功能状态具有可以用ATAC-seq读出的可接近性“指纹”。这些差异性分割模式与这些种类的推定功能状态是一致的,因为CTCF结合的区域针对DNA的短片段富集,而转录起始位点对于单-、双-和三-核小体关联的片段是差异性贫乏的。转录的和启动子侧翼区域针对较长的多核小体片段富集,这表明它们可能代表更紧密形式的染色质。最后,先前的研究已经表明,某些DNA序列耐核酸酶消化并被释放为大型的、多核小体大小的片段;随后的研究表明,这样的片段是浓缩的异染色质。事实上,被抑制的区域被发现对于短片段贫乏并且对于定相的多核小体插入片段富集,与其预期的不可接近的状态一致。这些数据表明,ATAC-seq揭示了染色质的差异性可接近形式,其已被长期推测在体内存在。

[0194] 为了探索GM 12878细胞系中可接近的染色质内的核小体定位,数据被划分为从DNA的推定无核小体区域产生的读数和可能来源于核小体关联的DNA的读数(见图12)。通过使用对核小体关联的片段正向加权并对无核小体的片段负向加权的简单启发法(见“方法”),我们计算了用于调用可接近的染色质的区域内的核小体位置的数据轨迹(Chen,K.等人Genome Research 2013 23,341-351)。示例基因座(图3a)含有推定的双向启动子,其中CAGE数据显示间隔～700bp的两个转录起始位点(TSS)。事实上ATAC-seq揭示了两个不同的无核小体区域,其由单个良好定位的单核小体隔开(图3a)。相较于MNase-seq,ATAC-seq数据更适合于检测推定的调节区内的核小体,因为大多数读数集中在染色质的可接近区域内(图3b)。通过平均所有活性TSS上的信号,注意到无核小体的片段富集在重叠TSS的典型无核小体启动子区域上,而核小体信号富集在活性TSS的上游和下游,并显示上游和下游核小体的特征性定相(图3c)。由于ATAC-seq读数集中在开放染色质的区域,因而强烈的核小体信号见于+1核小体,其在+2、+3和+4核小体上减弱,与此相反,MNase-seq核小体信号在离TSS较远的距离上增加,可能是由于更可接近的核小体的过度消化。此外,MNase-seq($4x10^9$个读数)测定所有的核小体,而从ATAC-seq产生的读数($1.98x10^8$个配对读数)集中于调节性核小体(图3b,c)。通过使用核小体识别,推定的远端调节区域和TSS被进一步划分成无核小体的区域和被预测为是核小体结合的区域。注意到当与倾向于保持丰富的核小体的远端元件相比较时,TSS针对无核小体区域富集(图3d)。这些数据表明ATAC-seq可以提供全基因

组调节元件中核小体关联的和无核小体的区域的高分辨率读出。

[0195] ATAC-seq揭示核小体-TF间隔的模式

[0196] ATAC-seq高分辨率调节核小体图谱可以用来理解核小体和DNA结合因子之间的关系。通过使用ChIP-seq数据,我们绘制了各种DNA结合因子相对于最近的核小体的二分体的位置。无监督层次聚类(图3e)揭示了相对于邻近核小体的结合的主要种类,包括1)在离最近的核小体二分体～180碱基处发生结合事件的因子的强核小体避开组(包括C-FOS,NFYA和IRF3),2)精确地"依偎"核小体DNA接触的预期末端的因子种类,其主要包括染色质循环因子CTCF和凝聚复杂亚基RAD21和SMC3;3)具有分级的核小体避开或核小体重叠结合行为的一大类主要转录因子,以及4)其结合位点趋于重叠核小体关联的DNA的种类。有趣的是,该最后一类包括染色质重塑因子例如CHD1和SIN3A以及RNA聚合酶II,其似乎富集在核小体边界上。精确核小体定位和DNA结合因子的位置之间的相互作用立即提示机理研究的特定假设,这是ATAC-seq的潜在优势。

[0197] ATAC-seq足迹推断全基因组因子占据

[0198] ATAC-seq使得能够准确推断全基因组DNA结合因子占据。由DNA结合蛋白直接占据的DNA序列应受到保护而避免转座;所得的序列"足迹"揭示了每个位点上DNA结合蛋白的存在,类似于DNase消化足迹。在染色体1的特异性CTCF结合位点上,我们在CTCF基序的精确位置上观察到明显的足迹(ATAC-seq信号的深凹口),类似于通过DNase-seq所见的足迹,所述CTCF基序的精确位置与GM12878细胞中CTCF ChIP-seq信号的最高峰相同(图4a)。ATAC-seq信号在基因组内的所有预期的CTCF位置上进行平均并观察到良好定型的"足迹"(图4b)。对于各种常见的TF获得了类似的结果(例如参见图13)。我们从基序一致得分、进化保守性和ATAC-seq插入数据推断CTCF结合概率以产生所有基因座上的CTCF的后验概率(图4c)(Pique-Regi等人Genome Research 2011 21 447-455)。使用ATAC-seq的结果接近地概括此细胞系中的ChIP-seq结合数据并且有利地相比于基于DNase的因子占据推断(见图14),表明可从这些ATAC-seq数据提取因子占据数据从而允许重建调节网络。

[0199] ATAC-seq实现临床时间尺度上的表观基因组分析

[0200] ATAC-seq是快速的、信息丰富的且兼容于小数量的细胞,并且可以作为强大的工具用于临床的个性表观基因组学。具体而言,可以设想"个人表观基因组学"为在临床时间尺度上从来自标准临床样品的个体产生的关于染色质的基因组范围的信息。ATAC-seq被应用于经由标准连续抽血测定健康志愿者的个人T细胞表观基因组,以证明能够在临床时间尺度上产生ATAC-seq文库的工作流。通过使用快速的T细胞富集和样品处理方案,从抽血到测序所需的总时间为约275分钟(图5a)。当结合对测序和分析周转时间的持续改进时,ATAC-seq可以提供个人表观基因组图谱的每日周转时间可能性。为了探索这种可能性,连续三天通过从单一个体的标准抽血来进行ATAC-seq(图5b)。作为考虑个人表观基因组图谱可如何包含个性化调节信息的训练,我们研究了IL2基因座上的ATAC-seq特征谱。IL-2是驱动T细胞生长并在炎症和自身免疫疾病中发挥功能的关键细胞因子。此外,不同的药物抑制不同转录因子的活性,所述转录因子以背景依赖性方式结合推定的IL2增强子。原则上,可能希望鉴定因果性转录因子途径以合理靶向抑制而不使患者暴露于不太可能实现IL-2阻断的治疗目标的药物。ATAC-seq显示,在先证者的T细胞中,仅NFAT而非其它两种药物靶标结合IL2(图5c),从而提供了关于该个体的调节状态的临床相关信息。

[0201] 通过使用ATAC-seq足迹,产生了先证者T细胞中89个转录因子的占据特征谱,使得能够系统性重建调节网络。通过利用此个性化调节图谱,我们比较了相同的89个转录因子在GM 12878和先证者CD4+T细胞之间的基因组分布。在T细胞和B细胞之间的分布上表现出较大变化的转录因子针对T细胞特异性因子富集(图5d)。该分析显示NFAT被差异调节,而典型的CTCF占据在这两个细胞类型中高度相关(图5d)。支持这种解释的是,注意到其中NFAT位于已知的T细胞特异性基因例如CD28和新型lincRNA RP11-229C3.2附近的特定基因座(图15)。此外,$CD4^+$和$CD8^+$T细胞以及通过荧光激活细胞分选(FACS)从单个抽血样品分离的单核细胞的ATAC-seq生成了个人表观基因组的解释框架,并证明了ATAC-seq与使用表面标志物的细胞富集兼容(图16)。单独地,等位基因特异性染色质可接近性已被显示与我们对人类疾病的理解特别相关。作为原理的证明,我们还使用ATAC-seq来鉴定GM12878细胞系内的候选等位基因特异性开放染色质区域(图17)。这些结果表明从临床样品产生详细的个性化基因调节网络的可行性,从而为未来的诊断应用打开大门。

[0202] 染色质可接近性的表观基因组研究已产生了非常多的生物学见解,但目前受到其复杂工作流和大细胞数量要求的应用的限制。虽然现有方法的改进可以使它们能够达到类似的表现,但在某些情况下ATAC-seq可因其快速性、简单性和低输入细胞数要求而提供优于现有技术的显著优势。ATAC-seq是信息丰富的测定,允许同时询问因子占据、调节位点中的核小体位置和全基因组染色质可接近性。这些见解来源于转座反应中插入的位置和插入片段长度的分布。虽然现存的方法如DNase-和MNase-seq可以提供ATAC-seq中的一些信息子集,但它们各自需要具有大细胞数量的单独测定,其增加了时间、成本并限制对许多系统的适用性。ATAC-seq还提供了生物学相关的基因组区域的插入片段大小“指纹”,提示其捕获了关于染色质紧密态的信息。ATAC-seq可具有广泛的适用性,显著添加至基因组学工具包,并改善我们对基因调控的理解,特别是当与其它强大的稀少细胞技术结合时,例如FACS、激光捕获显微切割(LCM)和最近RNA-seq的进展。

[0203] ATAC-seq可用于在与临床决策兼容的时间尺度上产生“个人基因组”特征谱。优化的程序可以在275分钟内将临床血液样品转化为完成的测序文库。减少的输入要求和快速的工作流,当结合最近引入的快速周转高通量测序仪器例如MiSeq和HiSeq2500时,应使得能够在实验室和临床上研究所选择的组织的个性化表观遗传景观。ATAC-seq与FACS兼容,从而可实现对来自原始组织的经仔细分选且稀少的亚群的研究。在发育和衰老以及人类疾病包括癌症、自身免疫和神经精神障碍的不同点上选择的细胞亚群是可行的应用。

[0204] 实施例2:单细胞ATAC-seq

[0205] 单细胞染色质可接近性数据集通过使用ATAC-seq方案获得。为了确保转座酶分子对开放染色质位点的比率几乎保持恒定,在初始转座反应后通过操作个体细胞来进行单细胞ATAC-seq测定。

[0206] 转座酶可用作开放染色质染色

[0207] 观察到在体外插入测序衔接子后,Tn5转座酶保持与DNA的紧密结合并形成高亲和力大分子复合物,其阻止所产生的ATAC-seq DNA片段的解离。为了支持此观察结果,用荧光标记的DNA衔接子装载Tn5的转座酶,并允许个体细胞的细胞核内的开放染色质区域的可视化(图18)。另外的电泳迁移率变动测定也表明转座酶在转座后保持与DNA关联。

[0208] 单细胞ATAC-seq提供了染色体DNA的独特读数特征

[0209] 由于该荧光信号定位于细胞核并且即使在转座后仍是可检测的,因而通过在随后的分选和细胞选择步骤中将转座的片段保持在细胞核中来进行单细胞ATAC-seq实验。将一组细胞透化,并用Tn5转座酶使染色体DNA转座。细胞被保持在阻止所得的ATAC-seq片段离开细胞核的条件下,(即二价阳离子未螯合),并且如上所述,将个体细胞分选至独立的PCR反应中用于文库制备。此工作流显著简化了单细胞分析的工作流并且提供了两个额外的优势。首先,这消除了分选过程对染色质状态的任何效应,因为转座在分选前进行。其次,它提供了更强健的ATAC-seq信号,因为细胞被直接分选至PCR预混合物中并进行扩增。通过使用此工作流,产生了每细胞~2,000-5,000个独特的ATAC-seq读数。这些读数针对GM 12878细胞中已知的开放染色质位点富集(图19)并且展示出指示核小体的特征周期性富集(图20)。

[0210] 实施例3:质量控制

[0211] 转座酶可接近染色质的测定(ATAC-seq)已显示与用于细胞收集的许多方法兼容,并且还在许多细胞类型和物种上有效地运行。然而,下列方案针对人类淋巴母细胞样细胞进行了优化。较小的变化(即细胞数、离心速度和裂解条件)可以针对特定的应用进行优化。

[0212] I.细胞制备

[0213] 1.收获细胞(无固定),方案由用户定义。

[0214] 2.离心50,000个细胞,在500×g下5分钟,4℃。

[0215] 3.用50μL冷的1x PBS缓冲液洗涤一次,在500×g 4℃下离心5分钟。

[0216] 4.轻轻吸取以将细胞沉淀重悬于50μL冷的裂解缓冲液(10mM Tris-HCl,pH 7.4,10mMNaCl,3mM $MgCl_2$和0.1%IGEPAL CA-630)。立即在500×g 4℃下离心10分钟。

[0217] 5.弃去上清液,并立即进行转座反应。

[0218] II.转座反应和纯化

[0219] 1.确保细胞沉淀置于冰上。

[0220] 2.为了制备转座反应混合物,组合以下成分:

[0221] 25μL 2x TD缓冲区(Illumina Cat#FC-121-1030)

[0222] 2.5μL Tn5转座(Illumina Cat#FC-121-1030)

[0223] 22.5μL不含核酸酶的H_2O

[0224] 总计50μl

[0225] 3.轻轻吸取以将细胞核重悬于转座反应混合物。

[0226] 4.在37℃下温育转座反应物30分钟。

[0227] 5.在转座后立即使用Qiagen MinElute试剂盒纯化。

[0228] 6.将转座的DNA洗脱在10μL洗脱缓冲液(10mM Tris缓冲液,pH 8)中。

[0229] 7.纯化的DNA可储存在-20℃。

[0230] III.PCR扩增

[0231] 1.为了扩增转座的DNA片段,将以下成分组合在PCR管中:

[0232] 10μL转座的DNA

[0233] 9.7μL不含核酸酶的H_2O

[0234] 2.5μL 25μM的定制Nextera PCR引物1*

[0235] 2.5μL 25μM的定制Nexttera PCR引物2*[条形码]

[0236] 0.3μL 100x SYBR Green I**(Invitrogen Cat#S-7563)

[0237] 25μL NEBNext高保真2x PCR预混合物(New England Labs Cat#M0541)

[0238] 总计50μl

[0239] *引物的完整列表如上文所示。

[0240] **10,000x SYBR Green I稀释在10mM Tris缓冲液,pH 8中以制备100x工作溶液。

[0241] 2.循环如下:

[0242] (1)72℃,5分钟

[0243] (2)98℃,30秒

[0244] (3)98℃,10秒

[0245] (4)63℃,30秒

[0246] (5)72℃,1分钟

[0247] (6)重复步骤3-5,4x

[0248] (7)保持在4℃

[0249] 3.为了减少PCR中的GC和大小偏倚,使用qPCR监测PCR反应以在饱和之前停止扩增。为了运行qPCR副反应,组合以下成分:

[0250] 5μL 5个循环PCR扩增的DNA

[0251] 4.44μL不含核酸酶的H_2O

[0252] 0.25μL 25μM的定制Nextera PCR引物1*

[0253] 0.25μL 25μM的定制Nextera PCR引物2*

[0254] 0.06μL 100x SYBR Green I

[0255] 5μL NEBNext高保真2X PCR预混合物

[0256] 总计15μl

[0257] *引物的完整列表可在本方案的第VI部分获得

[0258] 4.qPCR循环如下:

[0259] (1)98℃,30秒

[0260] (2)98℃,10秒

[0261] (3)63℃,30秒

[0262] (4)72℃,1分钟

[0263] (5)重复步骤2-4,19x

[0264] (6)保持在4℃下。

[0265] 5.其余45μL PCR反应物所需的额外循环数量如下确定:

[0266] (1)对线性Rn相对于循环映射

[0267] (2)设置5000RF阈值

[0268] (3)计算对应于四分之一最大荧光强度的

[0269] 如果待增加的循环#介于两个循环之间,则该#通过采取待增加的循环#的较小整数(即,蓝色和粉红色样品)来确定

[0270] 如果两个样品具有相似的Ct值但荧光强度不同,则使用具有较低荧光强度的样品(即,红色和蓝色的样品)计算循环#

[0271] 6.运行其余45μL PCR反应物以校正循环#。循环如下:

[0272] (1)98℃,30秒

[0273] (2)98℃,10秒

[0274] (3)63℃,30秒

[0275] (4)72℃,1分钟

[0276] (5)重复步骤2-4,x次

[0277] (6)保持在4℃

[0278] 7.使用Qiagen PCR Cleanup试剂盒纯化文库。将纯化的文库洗脱在20μL洗脱缓冲液(10mM Tris缓冲液,pH 8)中。确保在添加洗脱缓冲液之前柱是干燥的。

[0279] IV.使用凝胶电泳的文库QC

[0280] 1.用10mM Tris缓冲液,pH8以1:20稀释100bp NEB DNA梯度液。

[0281] 2.每5μL的稀释梯度液加入0.6μL 10×SYBR Green I。

[0282] 3.用2x DNA上样染料以1:1混合稀释梯度液。

[0283] 4.用2x DNA上样染料以1:1混合扩增的文库。

[0284] 5.将扩增的文库在5%Bio-Rad Mini-Protean TBE预制胶(保存在4℃)上运行。加载5μL稀释梯度液/DNA上样染料混合物。加载10μL扩增文库/DNA上样染料混合物。

[0285] 6.在~100mV运行45分钟。

[0286] 7.SYBR Green I染料在~488nm上具有最大激发并且在~520nm上具有最大发射。用SYBR Green I染料染色的DNA可以使用蓝光源或装备有在488nm下发射的激光的成像系统来可视化。我们通常使用来自Amersham Biosciences的Typhoon TRIO可变模式成像仪来进行可视化。图像通过利用520nm带通发射滤光器以筛除反射和散射激发光和背景荧光在100微米像素大小分辨率下进行数字化来最佳获得。

[0287] V.文库定量

[0288] 我们使用基于qPCR的方法来定量我们的ATAC-seq文库。我们已发现其它方法例如生物分析仪和Qubit,可以因插入片段大小的较大分布而产生误导性和不准确的结果。我们推荐使用用于Illumina测序平台的KAPA Library Quant试剂盒(KAPABiosystems)来定量文库。

[0289] 虽然前述实施方案已以举例说明和实例的方式为了清楚理解的目的在一定细节上进行了描述,但根据上述教导对于本领域普通技术人员显而易见的是可对其进行某些变化和修改而不脱离所附权利要求的精神或范围。

[0290] 本发明还涉及以下实施方案。

[0291] 1.用于分析染色质的方法,包括:

[0292] (a)用插入酶复合物处理分离自细胞群的染色质以产生基因组DNA的标记片段;

[0293] (b)测序至少一些标记片段以产生多个序列读数;和

[0294] (c)通过将获自序列读数的信息映射至细胞的基因组的区域而制作所述细胞的基因组的该区域的表观遗传图谱。

[0295] 2.实施方案1的方法,其中所述信息通过使用在序列读数的开头的核苷酸序列和任选末端的核苷酸序列获得。

[0296] 3.实施方案1-2之任一项的方法,其中在(c)中映射的所述信息选自下列的一种或多种:

[0297] (i)转座酶的切割位点;

[0298] (ii)在步骤(a)中产生的片段的大小;

[0299] (iii)序列读数长度;

[0300] (iii)确定长度范围的序列读数的位置;和

[0301] (iv)序列读数丰度。

[0302] 4.实施方案3的方法,其中确定大小范围的片段是无核小体的片段。

[0303] 5.实施方案1-4之任一项的方法,其中所述表观遗传图谱显示下列的一项或多项:

[0304] (i)沿所述区域的染色质可接近性的特征谱;

[0305] (ii)针对所述区域中结合位点的DNA结合蛋白的占据;

[0306] (iii)所述区域中的无核小体的DNA;

[0307] (iv)沿所述区域的核小体定位;

[0308] (v)染色质状态。

[0309] 6.实施方案5的方法,其还包括测量所述DNA结合蛋白对结合位点的总体占据。

[0310] 7.实施方案5的方法,其中所述DNA结合蛋白是转录因子。

[0311] 8.实施方案1-7之任一项的方法,其中所述细胞群包括500至100,000个细胞。

[0312] 9.实施方案1-8之任一项的方法,其中所述细胞分离自个体。

[0313] 10.实施方案1-9之任一项的方法,其中所述细胞分离自所述个体的血液。

[0314] 11.实施方案1-10的方法,其中所述细胞是相同的细胞类型。

[0315] 12.实施方案11的方法,其中所述细胞是FACS选择的细胞。

[0316] 13.实施方案1-12之任一项的方法,其中所述处理步骤(a)包括:

[0317] 从细胞群分离细胞核;和

[0318] 将分离的细胞核与所述插入酶复合物组合,其中所述组合导致细胞核裂解以释放所述染色质,以及导致产生基因组DNA的所述标记片段。

[0319] 14.实施方案1-13之任一项的方法,其中所述转座酶来源于Tn5转座酶。

[0320] 15.实施方案1-14之任一项的方法,其中所述转座酶来源于MuA转座酶。

[0321] 16.用于比较两种样品的方法,包括:

[0322] (a)使用实施方案1的方法分析第一细胞群,以产生第一表观遗传图谱;和

[0323] (b)使用实施方案1的方法分析第二细胞群,以产生第二表观遗传图谱;和

[0324] (c)比较所述第一表观遗传图谱与所述第二表观遗传图谱。

[0325] 17.实施方案16的方法,其中所述第一细胞群和所述第二细胞群是在不同的时间从相同个体收集的。

[0326] 18.实施方案16的方法,其中所述第一细胞群和所述第二细胞群是从不同个体收集的不同细胞群。

[0327] 19.诊断方法,其包括:

[0328] 使用实施方案1的方法分析来自患者的染色质,以产生表观遗传图谱;和

[0329] 基于所述表观遗传图谱提供诊断或预后。

[0330] 20.试剂盒,其包含:

[0331] (a)用于从细胞群分离细胞核的试剂;

[0332] (b)插入酶复合物,和

[0333] (c)转座酶反应缓冲液,

[0334] 其中所述试剂盒的组分被配置为使得反应缓冲液、转座子标签和衔接子与细胞核的体外组合导致细胞核裂解以释放染色质,以及导致产生基因组DNA的标记片段。

[0335] 21.用于测定多核苷酸在某位点上的可接近性的方法,其中所述多核苷酸来自细胞样品,所述方法包括:

[0336] (a)用插入酶将多个分子标签插入所述多核苷酸;和

[0337] (b)使用所述分子标签来测定所述位点上的可接近性。

[0338] 22.实施方案21的方法,其还包括使用所测定的所述可接近性来鉴定在所述位点上结合至所述多核苷酸的一种或多种蛋白。

[0339] 23.实施方案22的方法,其中至少一种所述蛋白是转录因子。

[0340] 24.实施方案21的方法,其还包括使用所述分子标签来产生所述多核苷酸的可接近性图谱。

[0341] 25.用于分析来自细胞样品的多核苷酸的三维结构的方法,包括:

[0342] (a)用插入酶将多个分子标签插入所述多核苷酸;和

[0343] (b)使用所述分子标签来分析述多核苷酸的三维结构。

[0344] 26.实施方案21或25的方法,其中所述细胞样品获自原始来源。

[0345] 27.实施方案21或25的方法,其中所述细胞样品由少于约500,000个细胞组成。

[0346] 28.实施方案27的方法,其中所述细胞样品是单个细胞。

[0347] 29.实施方案21或25的方法,其中所述多核苷酸在步骤(a)过程中被分割成多个片段。

[0348] 30.实施方案29的方法,其还包括扩增所述片段。

[0349] 31.实施方案29的方法,其中通过对所述片段测序从而产生多个测序读数来测定所述可接近性或分析所述三维结构。

[0350] 32.实施方案31的方法,其中所述片段通过高通量测序技术测序。

[0351] 33.实施方案31的方法,其还包括基于所述插入酶的序列插入偏好标准化所述测序读数。

[0352] 34.实施方案31的方法,其中所述测序读数的长度用于确定染色质状态注释。

[0353] 35.实施方案21或25的方法,其还包括透化所述细胞样品以允许所述插入酶进入。

[0354] 36.实施方案35的方法,其中所述细胞样品中的细胞核在所述透化期间被最小限度地扰乱。

[0355] 37.实施方案35的方法,其中所述细胞样品使用透化剂来透化。

[0356] 38.实施方案37的方法,其中所述透化剂选自NP40、洋地黄皂苷、吐温、链球菌溶血素和阳离子脂质。

[0357] 39.实施方案35的方法,其中所述细胞样品使用低渗休克和/或超声处理来透化。

[0358] 40.实施方案21或25的方法,其中所述插入酶还包含核定位信号。

[0359] 41.实施方案21或25的方法,其中所述插入通过加入一种或多种二价阳离子来促进。

[0360] 42.实施方案41的方法,其中所述一种或多种二价阳离子包括镁。

[0361] 43.实施方案41的方法,其中所述一种或多种二价阳离子包括锰。

[0362] 44.实施方案21或25的方法,其还包括基于所述特定位点的所述可接近性或所述

多核苷酸的所述三维结构来分析受试者的疾病状态,其中所述细胞样品获自所述受试者。

[0363] 45.实施方案21或25的方法,其还包括将所述细胞样品或所述多核苷酸划分成多个部分。

[0364] 46.实施方案45的方法,其中基于所述分子标签划分所述部分。

[0365] 47.实施方案21或25的方法,其还包括分析所述细胞样品的表型。

[0366] 48.实施方案47的方法,其中所述表型与所述位点的所述可接近性或所述多核苷酸的所述三维结构相关。

[0367] 49.实施方案21或25的方法,其中所述插入酶包含两个或更多个酶部分。

[0368] 50.实施方案49的方法,其中每一个所述酶部分将共同的序列插入所述多核苷酸。

[0369] 51.实施方案50的方法,其中所述共同的序列包括共同的条形码。

[0370] 52.实施方案49的方法,其中所述酶部分连接在一起。

[0371] 53.实施方案49的方法,其中所述酶部分包括转座酶。

[0372] 54.实施方案21或25的方法,其中所述多核苷酸在步骤(a)过程中被分割成多个片段,并且其中包含所述共同的条形码的所述片段被测定为在所述多核苷酸的三维结构中是靠近的。

[0373] 55.实施方案21或25的方法,其中所述分子标签包含测序衔接子。

[0374] 56.实施方案55的方法,其中所述测序衔接子还包含条形码标记。

[0375] 57.实施方案55的方法,其中所述条形码标记包括单一序列。

[0376] 58.实施方案21或25的方法,其中所述分子标签包括荧光标签。

[0377] 59.组合物,其包含多核苷酸、插入酶和插入元件,其中:

[0378] (a)所述插入元件包括包含预先确定的序列的核酸;和

[0379] (b)所述插入酶还包含亲和标签。

[0380] 60.组合物,其包含多核苷酸、插入酶和插入元件,其中:

[0381] (a)所述插入酶包含两个或更多个酶部分;和

[0382] (b)所述酶部分连接在一起。

[0383] 61.试剂盒,其包含:

[0384] (a)细胞裂解缓冲液;

[0385] (b)包含亲和标签的插入酶;和

[0386] (c)包含核酸的插入元件,其中所述核酸包含预先确定的序列。

[0387] 62.试剂盒,其包含:

[0388] (a)细胞裂解缓冲液;

[0389] (b)包含两个或更多个酶部分的插入酶,其中所述酶部分连接在一起;和

[0390] (c)插入元件。

[0391] 63.实施方案21、25、59、60、61或62的方法、组合物或试剂盒,其中所述多核苷酸结合至多个关联分子。

[0392] 64.实施方案63的方法、组合物或试剂盒,其中所述关联分子是蛋白质。

[0393] 65.实施方案64的方法、组合物或试剂盒,其中所述蛋白质包括组蛋白。

[0394] 66.实施方案21、25、59、60、61或62的方法、组合物或试剂盒,其中所述插入酶是转座酶。

[0395] 67.实施方案66的方法、组合物或试剂盒,其中所述转座酶来源于Tn5转座酶。

[0396] 68.实施方案66的方法、组合物或试剂盒,其中所述转座酶来源于MuA转座酶。

[0397] 69.实施方案66的方法、组合物或试剂盒,其中所述转座酶来源于Vibhar转座酶。

[0398] 70.实施方案21、25、59、60、61或62的方法、组合物或试剂盒,其中所述插入酶还包含亲和标签。

[0399] 71.实施方案70的方法、组合物或试剂盒,其中所述亲和标签是抗体。

[0400] 72.实施方案71的方法、组合物或试剂盒,其中所述抗体结合至转录因子。

[0401] 73.实施方案71的方法、组合物或试剂盒,其中所述抗体结合至修饰的核小体。

[0402] 74.实施方案71的方法、组合物或试剂盒,其中所述抗体结合至修饰的核酸。

[0403] 75.实施方案74的方法、组合物或试剂盒,其中所述修饰的核酸是甲基化或羟甲基化的DNA。

[0404] 76.实施方案70的方法、组合物或试剂盒,其中所述亲和标签是单链核酸。

[0405] 77.实施方案76的方法、组合物或试剂盒,其中所述单链核酸结合至靶核酸。

[0001] 序列表

[0002] <110> Buenrostro, Jason

[0003] Chang, Howard Y

[0004] Greenleaf, William J

[0005] Giresi, Paul

[0006] <120> 用于个人表观基因组学的至天然染色质中的转座

[0007] <130> STAN-1111WO

[0008] <150> US 61/826,728

[0009] <151> 2013-05-23

[0010] <160> 25

[0011] <170> PatentIn version 3.5

[0012] <210> 1

[0013] <211> 50

[0014] <212> DNA

[0015] <213> 人工序列

[0016] <220>

[0017] <223> 合成的多核苷酸

[0018] <400> 1

[0019] aatgatacgg cgaccaccga gatctacact cgtcggcagc gtcagatgtg 50

[0020] <210> 2

[0021] <211> 53

[0022] <212> DNA

[0023] <213> 人工序列

[0024] <220>

[0025] <223> 合成的多核苷酸

[0026] <400> 2

[0027] caagcagaag acggcatacg agattcgcct tagtctcgtg ggctcggaga tgt 53

[0028] <210> 3

[0029] <211> 53

[0030] <212> DNA

[0031] <213> 人工序列

[0032] <220>

[0033] <223> 合成的多核苷酸

[0034] <400> 3

[0035] caagcagaag acggcatacg agatctagta cggtctcgtg ggctcggaga tgt 53

[0036] <210> 4

[0037] <211> 53

[0038] <212> DNA

[0039] <213> 人工序列

[0040] <220>

[0041] <223> 合成的多核苷酸

[0042] <400> 4

[0043] caagcagaag acggcatacg agatttctgc ctgtctcgtg ggctcggaga tgt 53

[0044] <210> 5

[0045] <211> 53

[0046] <212> DNA

[0047] <213> 人工序列

[0048] <220>

[0049] <223> 合成的多核苷酸

[0050] <400> 5

[0051] caagcagaag acggcatacg agatgctcag gagtctcgtg ggctcggaga tgt 53

[0052] <210> 6

[0053] <211> 53

[0054] <212> DNA

[0055] <213> 人工序列

[0056] <220>

[0057] <223> 合成的多核苷酸

[0058] <400> 6

[0059] caagcagaag acggcatacg agataggagt ccgtctcgtg ggctcggaga tgt 53

[0060] <210> 7

[0061] <211> 53

[0062] <212> DNA

[0063] <213> 人工序列

[0064] <220>

[0065] <223> 合成的多核苷酸

[0066] <400> 7

[0067] caagcagaag acggcatacg agatcatgcc tagtctcgtg ggctcggaga tgt 53

[0068] <210> 8

[0069] <211> 53

[0070] <212> DNA

[0071] <213> 人工序列

[0072] <220>

[0073] <223> 合成的多核苷酸

[0074] <400> 8

[0075] caagcagaag acggcatacg agatgtagag aggtctcgtg ggctcggaga tgt 53

[0076] <210> 9

[0077] <211> 53

[0078] <212> DNA
[0079] <213> 人工序列
[0080] <220>
[0081] <223> 合成的多核苷酸
[0082] <400> 9
[0083] caagcagaag acggcatacg agatcctctc tggtctcgtg ggctcggaga tgt 53
[0084] <210> 10
[0085] <211> 53
[0086] <212> DNA
[0087] <213> 人工序列
[0088] <220>
[0089] <223> 合成的多核苷酸
[0090] <400> 10
[0091] caagcagaag acggcatacg agatagcgta gcgtctcgtg ggctcggaga tgt 53
[0092] <210> 11
[0093] <211> 53
[0094] <212> DNA
[0095] <213> 人工序列
[0096] <220>
[0097] <223> 合成的多核苷酸
[0098] <400> 11
[0099] caagcagaag acggcatacg agatcagcct cggtctcgtg ggctcggaga tgt 53
[0100] <210> 12
[0101] <211> 53
[0102] <212> DNA
[0103] <213> 人工序列
[0104] <220>
[0105] <223> 合成的多核苷酸
[0106] <400> 12
[0107] caagcagaag acggcatacg agattgcctc ttgtctcgtg ggctcggaga tgt 53
[0108] <210> 13
[0109] <211> 53
[0110] <212> DNA
[0111] <213> 人工序列
[0112] <220>
[0113] <223> 合成的多核苷酸
[0114] <400> 13
[0115] caagcagaag acggcatacg agattcctct acgtctcgtg ggctcggaga tgt 53
[0116] <210> 14

[0117] <211> 53

[0118] <212> DNA

[0119] <213> 人工序列

[0120] <220>

[0121] <223> 合成的多核苷酸

[0122] <400> 14

[0123] caagcagaag acggcatacg agatatcacg acgtctcgtg ggctcggaga tgt 53

[0124] <210> 15

[0125] <211> 53

[0126] <212> DNA

[0127] <213> 人工序列

[0128] <220>

[0129] <223> 合成的多核苷酸

[0130] <400> 15

[0131] caagcagaag acggcatacg agatacagtg gtgtctcgtg ggctcggaga tgt 53

[0132] <210> 16

[0133] <211> 53

[0134] <212> DNA

[0135] <213> 人工序列

[0136] <220>

[0137] <223> 合成的多核苷酸

[0138] <400> 16

[0139] caagcagaag acggcatacg agatcagatc cagtctcgtg ggctcggaga tgt 53

[0140] <210> 17

[0141] <211> 53

[0142] <212> DNA

[0143] <213> 人工序列

[0144] <220>

[0145] <223> 合成的多核苷酸

[0146] <400> 17

[0147] caagcagaag acggcatacg agatacaaac gggtctcgtg ggctcggaga tgt 53

[0148] <210> 18

[0149] <211> 53

[0150] <212> DNA

[0151] <213> 人工序列

[0152] <220>

[0153] <223> 合成的多核苷酸

[0154] <400> 18

[0155] caagcagaag acggcatacg agatacccag cagtctcgtg ggctcggaga tgt 53

[0156] <210> 19

[0157] <211> 53

[0158] <212> DNA

[0159] <213> 人工序列

[0160] <220>

[0161] <223> 合成的多核苷酸

[0162] <400> 19

[0163] caagcagaag acggcatacg agataacccc tcgtctcgtg ggctcggaga tgt 53

[0164] <210> 20

[0165] <211> 53

[0166] <212> DNA

[0167] <213> 人工序列

[0168] <220>

[0169] <223> 合成的多核苷酸

[0170] <400> 20

[0171] caagcagaag acggcatacg agatcccaac ctgtctcgtg ggctcggaga tgt 53

[0172] <210> 21

[0173] <211> 53

[0174] <212> DNA

[0175] <213> 人工序列

[0176] <220>

[0177] <223> 合成的多核苷酸

[0178] <400> 21

[0179] caagcagaag acggcatacg agatcaccac acgtctcgtg ggctcggaga tgt 53

[0180] <210> 22

[0181] <211> 53

[0182] <212> DNA

[0183] <213> 人工序列

[0184] <220>

[0185] <223> 合成的多核苷酸

[0186] <400> 22

[0187] caagcagaag acggcatacg agatgaaacc cagtctcgtg ggctcggaga tgt 53

[0188] <210> 23

[0189] <211> 53

[0190] <212> DNA

[0191] <213> 人工序列

[0192] <220>

[0193] <223> 合成的多核苷酸

[0194] <400> 23

[0195] caagcagaag acggcatacg agattgtgac cagtctcgtg ggctcggaga tgt 53

[0196] <210> 24

[0197] <211> 53

[0198] <212> DNA

[0199] <213> 人工序列

[0200] <220>

[0201] <223> 合成的多核苷酸

[0202] <400> 24

[0203] caagcagaag acggcatacg agatagggtc aagtctcgtg ggctcggaga tgt 53

[0204] <210> 25

[0205] <211> 53

[0206] <212> DNA

[0207] <213> 人工序列

[0208] <220>

[0209] <223> 合成的多核苷酸

[0210] <400> 25

[0211] caagcagaag acggcatacg agataggagt gggtctcgtg ggctcggaga tgt 53

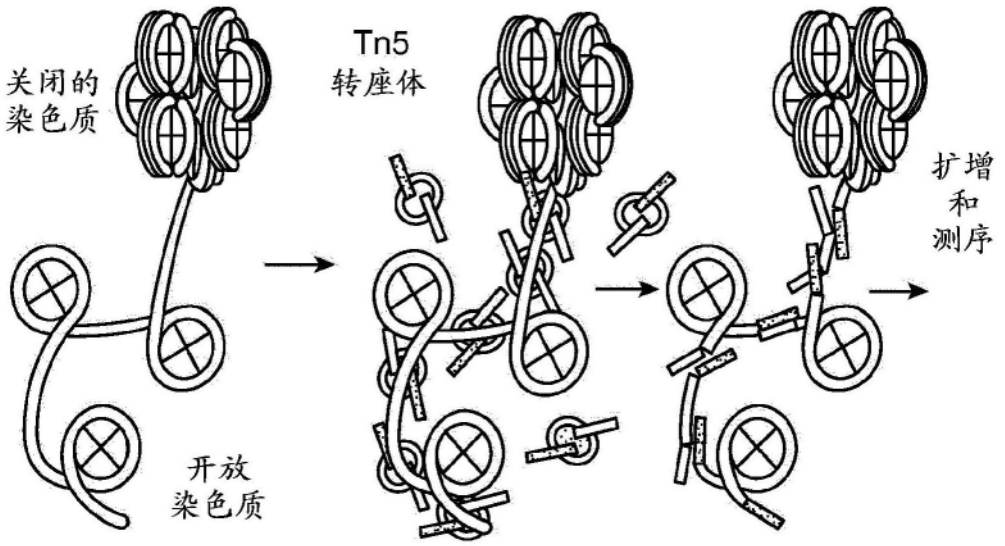

关闭的
染色质
Tn5
转座体
扩增
和
测序
开放
染色质

图1A

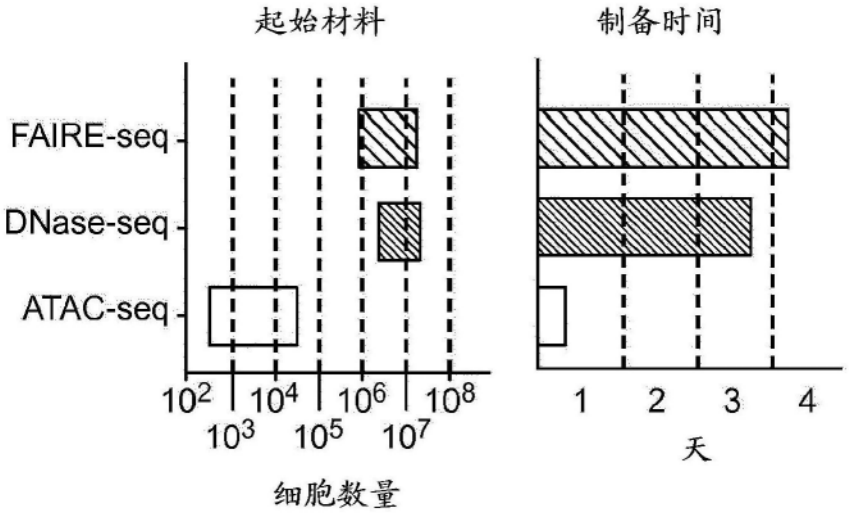

起始材料
制备时间
FAIRE-seq
DNase-seq
ATAC-seq
10^2 10^4 10^6 10^8
10^3 10^5 10^7
细胞数量
1 2 3 4
天

图1B

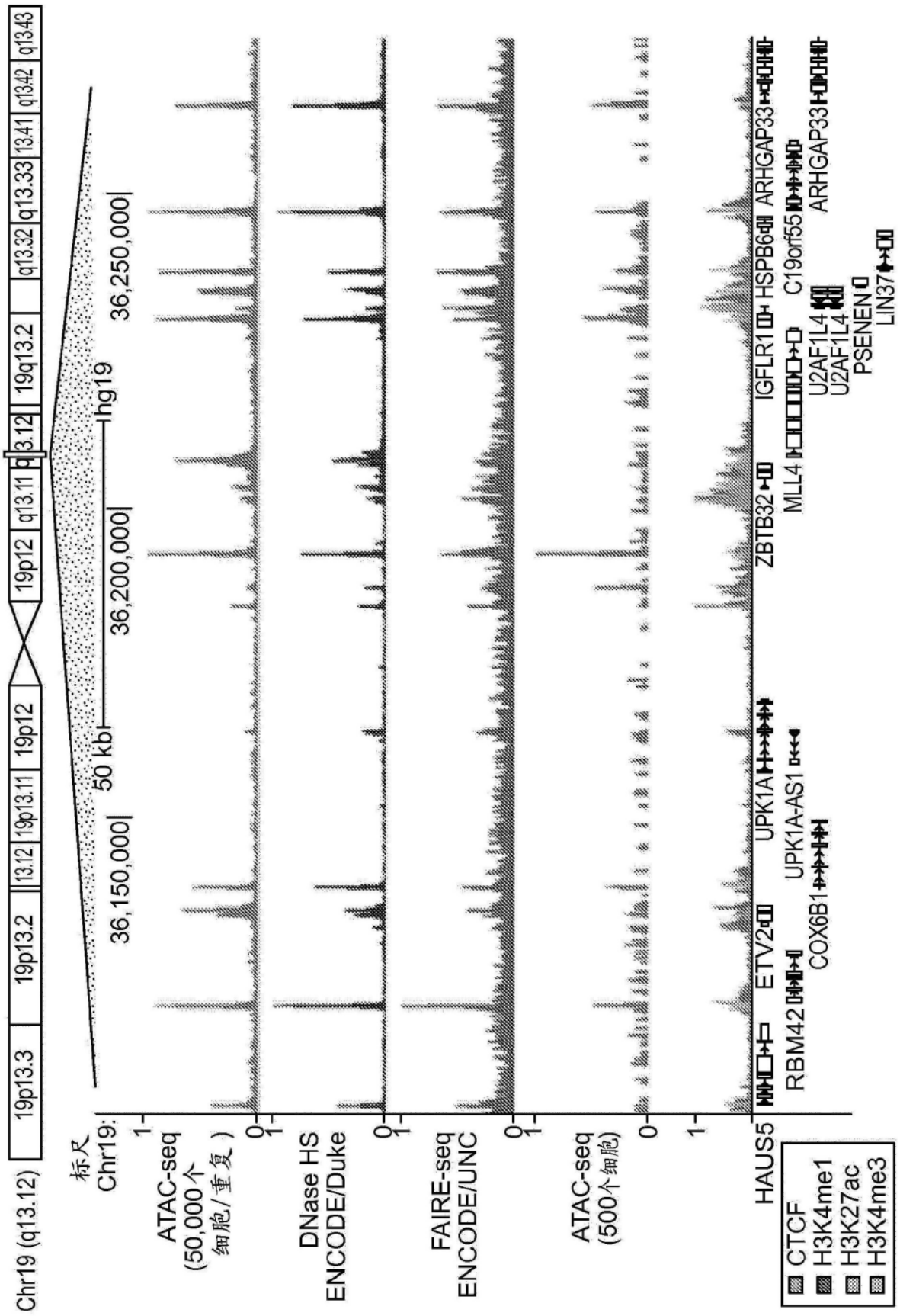

Chr19 (q13.12)
19p13.3
19p13.2
13.12
19p13.11
19p12
19p12
q13.11
13.12
19q13.2
q13.32
q13.33
13.41
q13.42
q13.43
标尺
Chr19:
36,150,000
50 kb
36,200,000
hg19
36,250,000
ATAC-seq
(50,000个细胞/重复)
DNase HS
ENCODE/Duke
FAIRE-seq
ENCODE/UNC
ATAC-seq
(500个细胞)
HAUS5
ETV2
UPK1A
ZBTB32
IGFLR1
HSPB6
ARHGAP33
RBM42
UPK1A-AS1
MLL4
C19orf55
U2AF1L4
U2AF1L4
ARHGAP33
COX6B1
PSENEN
LIN37
CTCF
H3K4me1
H3K27ac
H3K4me3

图1C

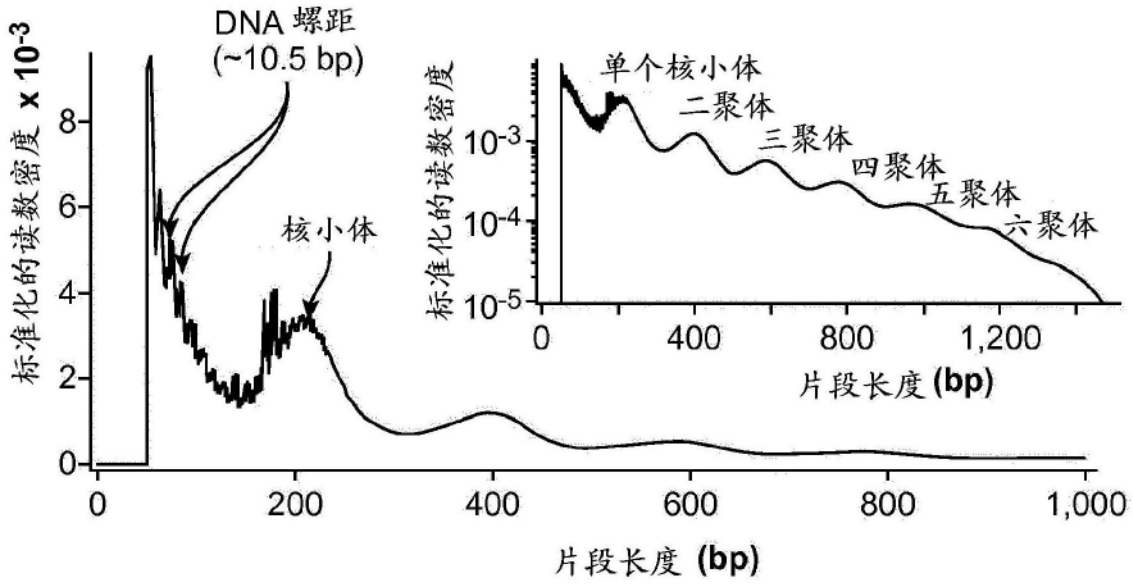

DNA 螺距
(~10.5 bp)
核小体
标准化的读数密度 x 10-3
片段长度 (bp)
单个核小体
二聚体
三聚体
四聚体
五聚体
六聚体
标准化的读数密度
片段长度 (bp)

图2A

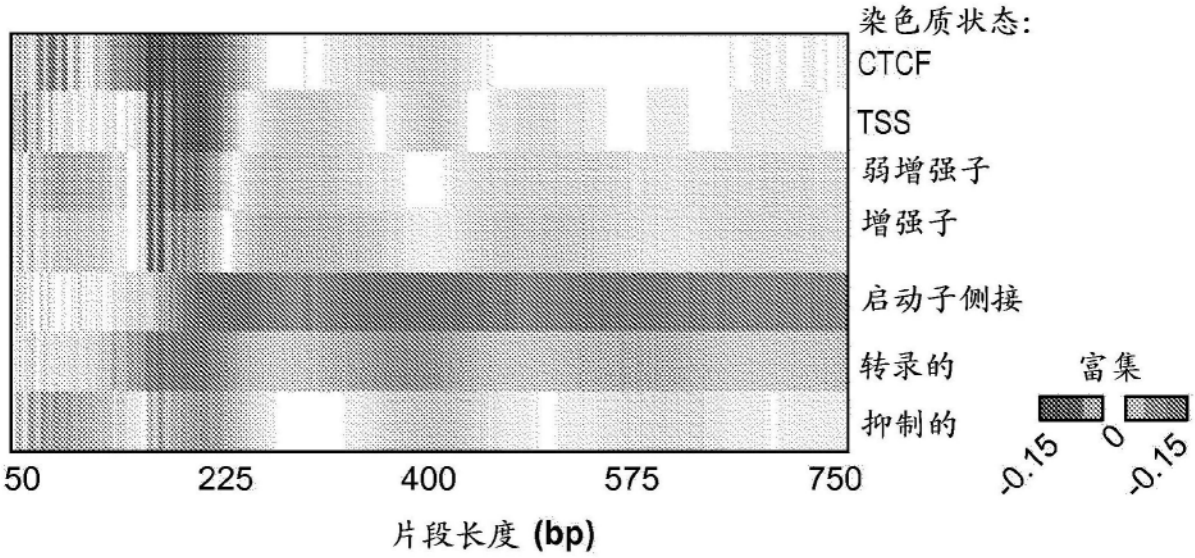

染色质状态:
CTCF
TSS
弱增强子
增强子
启动子侧接
转录的
抑制的
富集
片段长度 (bp)

图2B

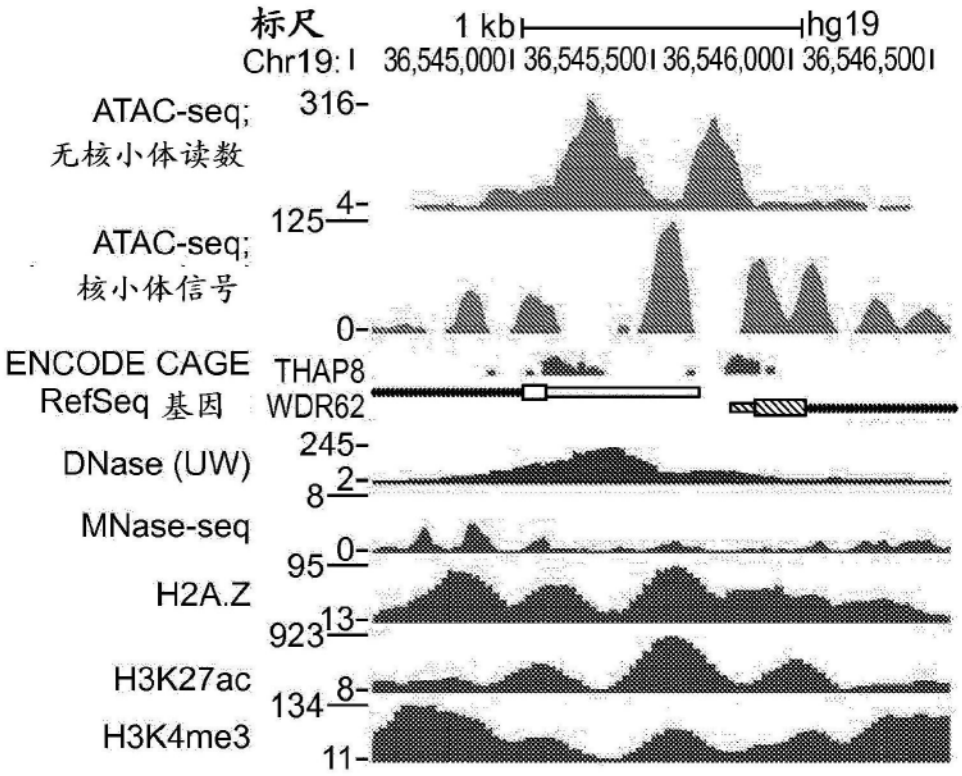

标尺
1 kb
hg19
Chr19: I
36,545,000I 36,545,500I 36,546,000I 36,546,500I
ATAC-seq;
无核小体读数
316-
125
4-
ATAC-seq;
核小体信号
0-
ENCODE CAGE
RefSeq 基因
THAP8
WDR62
DNase (UW)
245-
8
2-
MNase-seq
95
0-
H2A.Z
923
13-
H3K27ac
134
8-
H3K4me3
11-

图3A

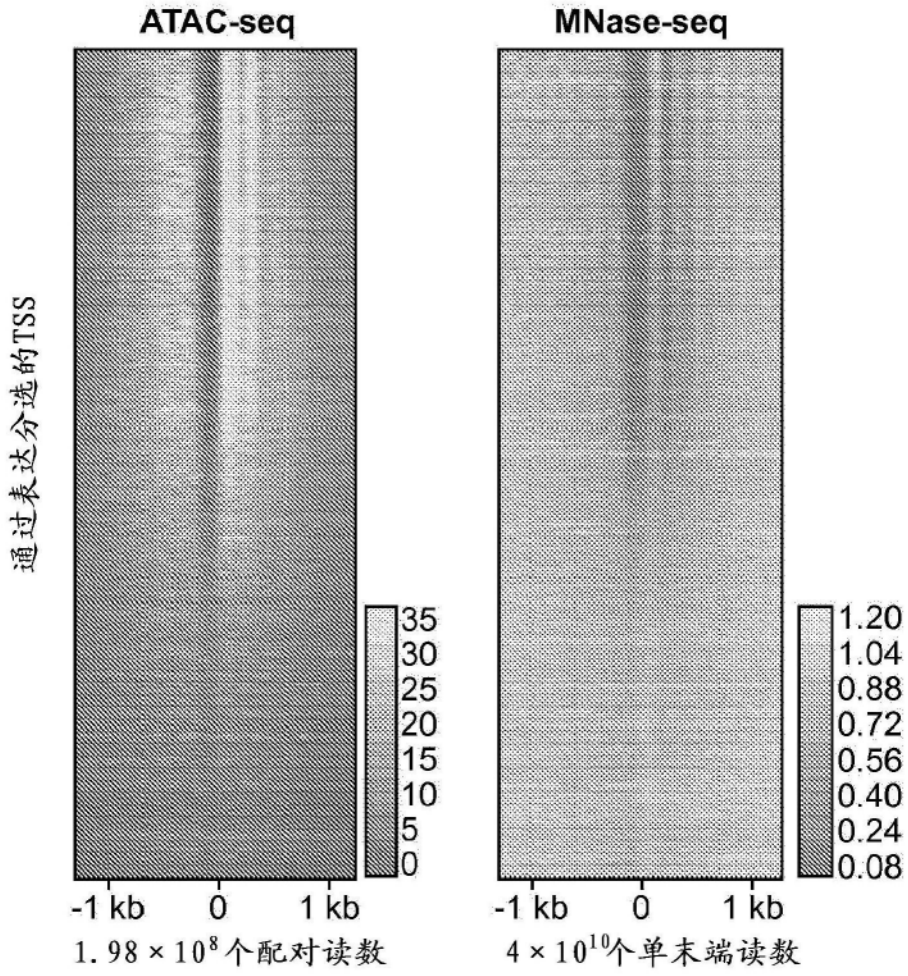

ATAC-seq
MNase-seq
通过表达分选的TSS
35
30
25
20
15
10
5
0
1.20
1.04
0.88
0.72
0.56
0.40
0.24
0.08
-1 kb
0
1 kb
-1 kb
0
1 kb
1.98 × 10^8个配对读数
4 × 10^10个单末端读数

图3B

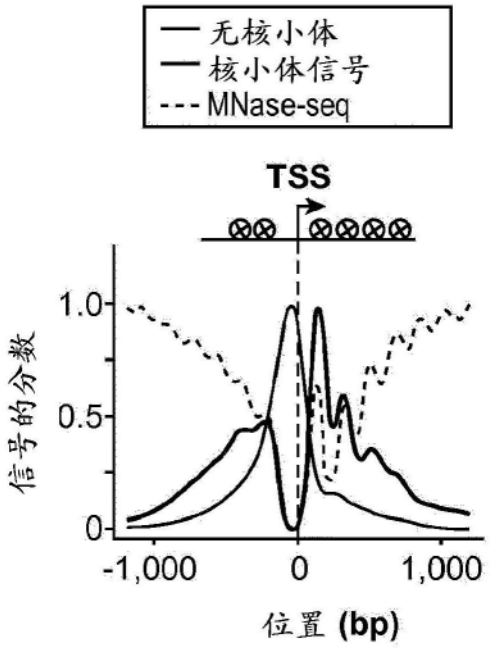

无核小体
核小体信号
MNase-seq
TSS
1.0
0.5
0
信号的分数
-1,000
0
1,000
位置 (bp)

图3C

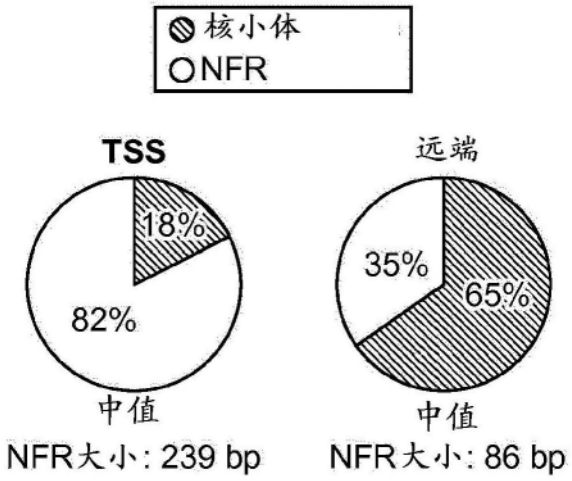

核小体
NFR
TSS
18%
82%
中值
NFR大小: 239 bp
远端
35%
65%
中值
NFR大小: 86 bp

图3D

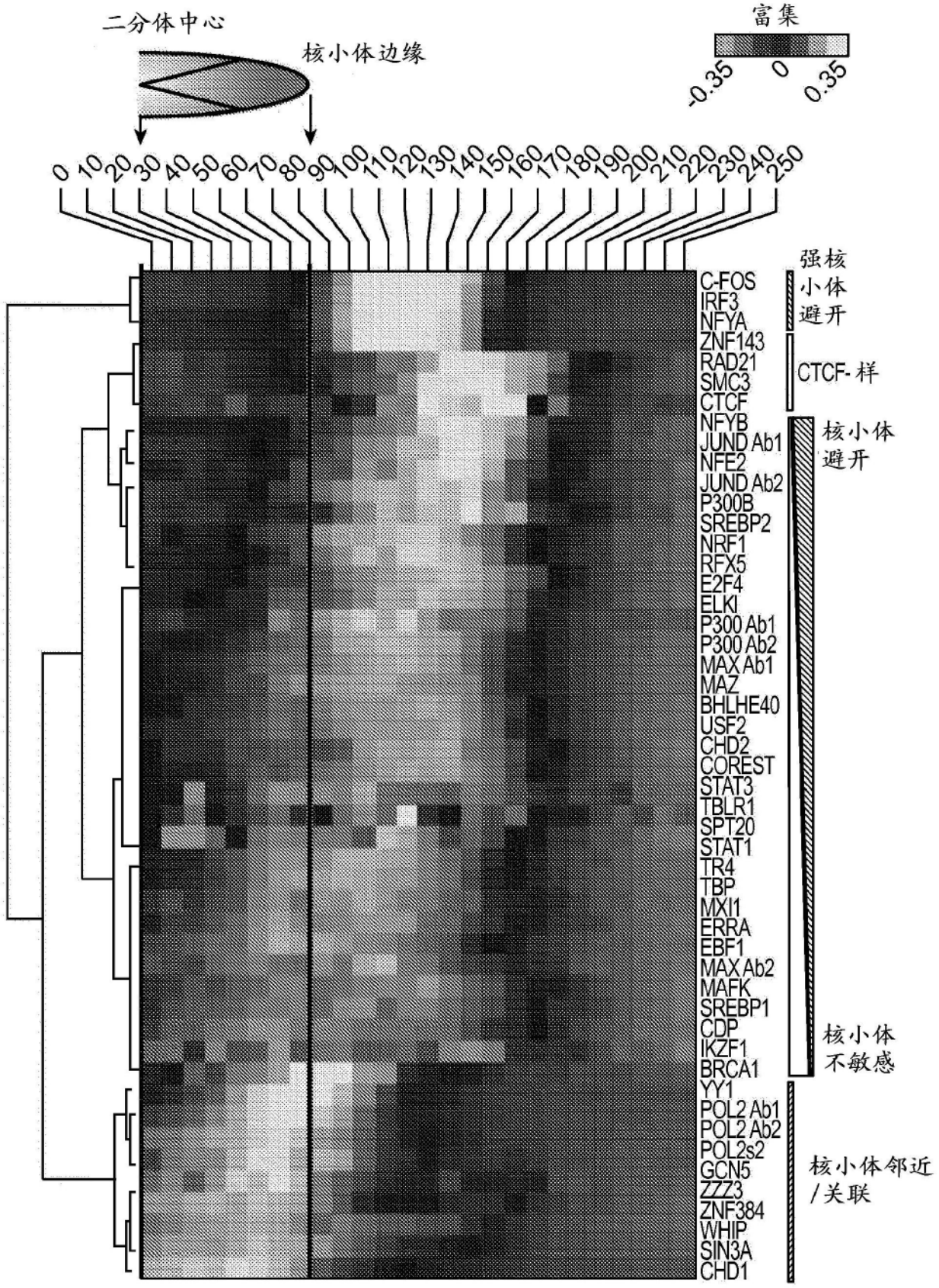

二分体中心
核小体边缘
富集
-0.35
0
0.35
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
C-FOS
IRF3
NFYA
ZNF143
RAD21
SMC3
CTCF
NFYB
JUND Ab1
NFE2
JUND Ab2
P300B
SREBP2
NRF1
RFX5
E2F4
ELKI
P300 Ab1
P300 Ab2
MAX Ab1
MAZ
BHLHE40
USF2
CHD2
COREST
STAT3
TBLR1
SPT20
STAT1
TR4
TBP
MXI1
ERRA
EBF1
MAX Ab2
MAFK
SREBP1
CDP
IKZF1
BRCA1
YY1
POL2 Ab1
POL2 Ab2
POL2s2
GCN5
ZZZ3
ZNF384
WHIP
SIN3A
CHD1
强核小体避开
CTCF-样
核小体避开
核小体不敏感
核小体邻近/关联

图3E

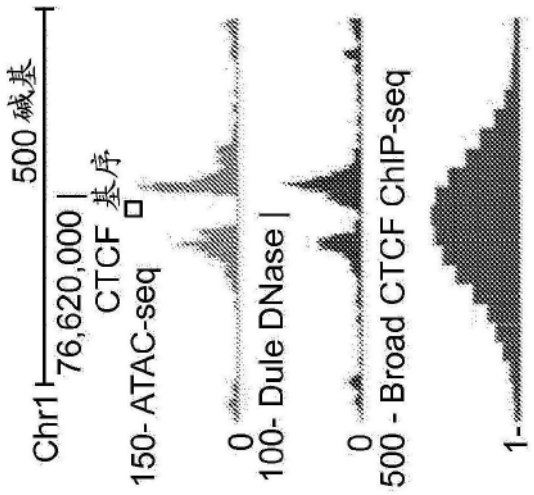

500 碱基
Chr1
76,620,000
CTCF 基序
150- ATAC-seq
0
100- Dule DNase
0
500 - Broad CTCF ChIP-seq
1-

图4A

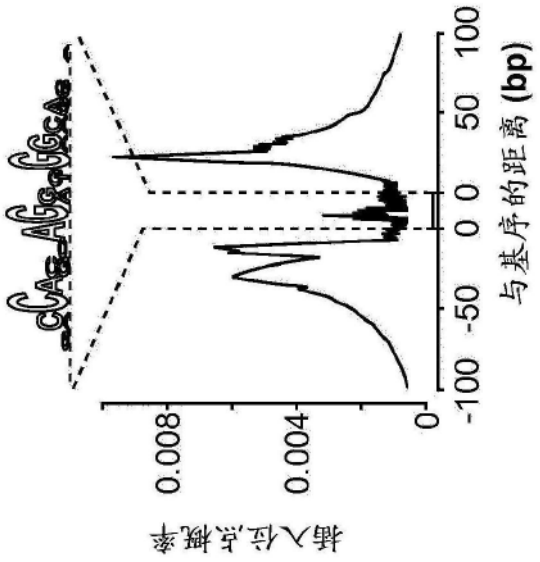

插入位点概率
0.008
0.004
0
-100 -50 0 0 50 100
与基序的距离 (bp)

图4B

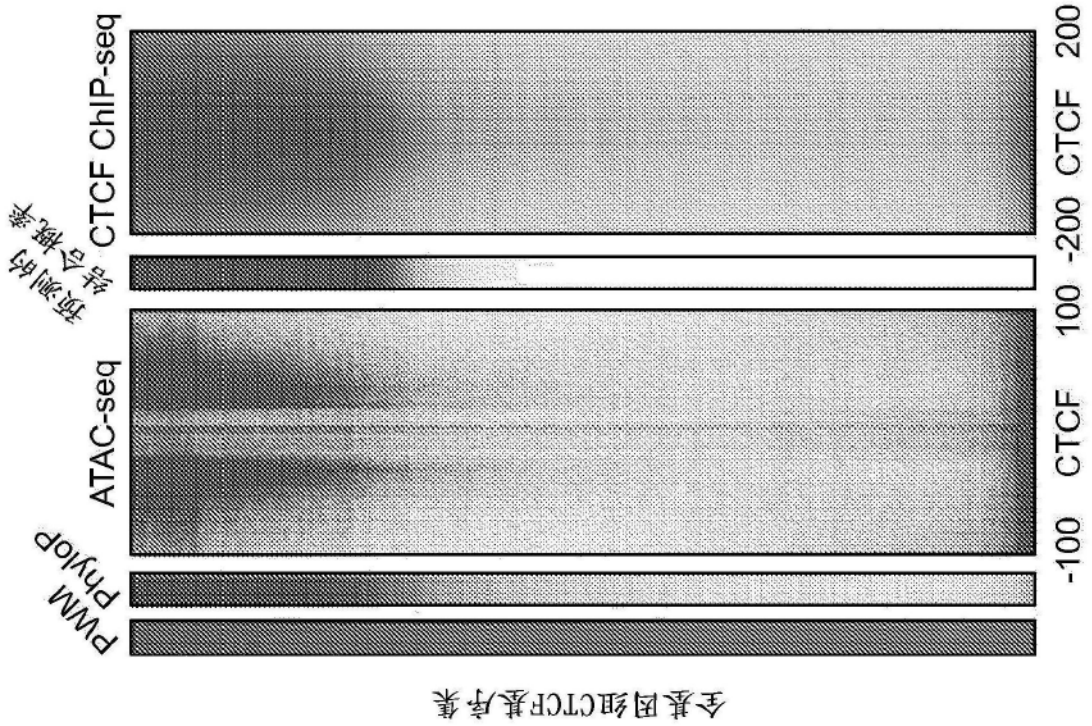

CTCF ChIP-seq
预测的结合概率
ATAC-seq
PWM Physiop
200
CTCF
100
-200
CTCF
-100
全基因组CTCF基序集

图4C


抽血
5 min
CD4+ T细胞纯化
90 min
转座 & 扩增
180 min
测序
240 min-120 h

图5A

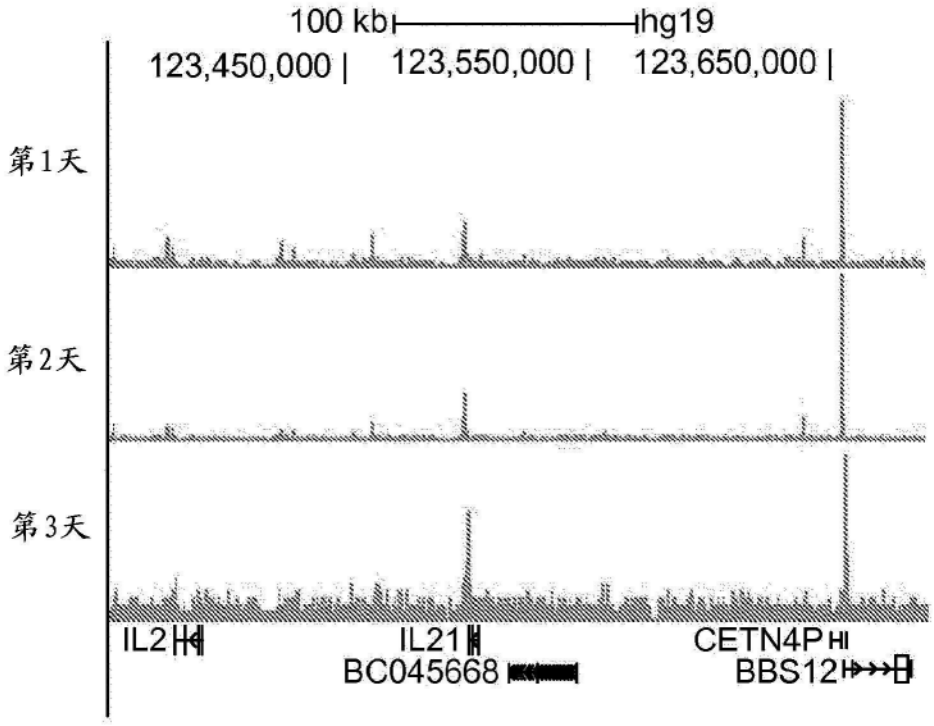

100 kb
hg19
123,450,000
123,550,000
123,650,000
第1天
第2天
第3天
IL2
IL21
BC045668
CETN4P
BBS12

图5B

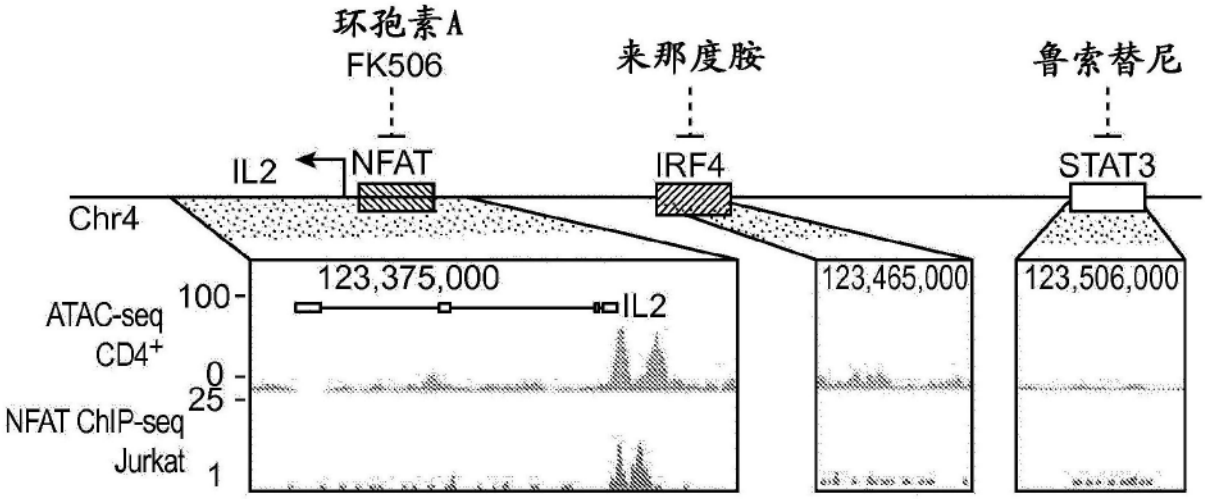

环孢素A
FK506
来那度胺
鲁索替尼
IL2
NFAT
IRF4
STAT3
Chr4
123,375,000
IL2
123,465,000
123,506,000
ATAC-seq
CD4+
100
0
NFAT ChIP-seq
Jurkat
25
1

图5C

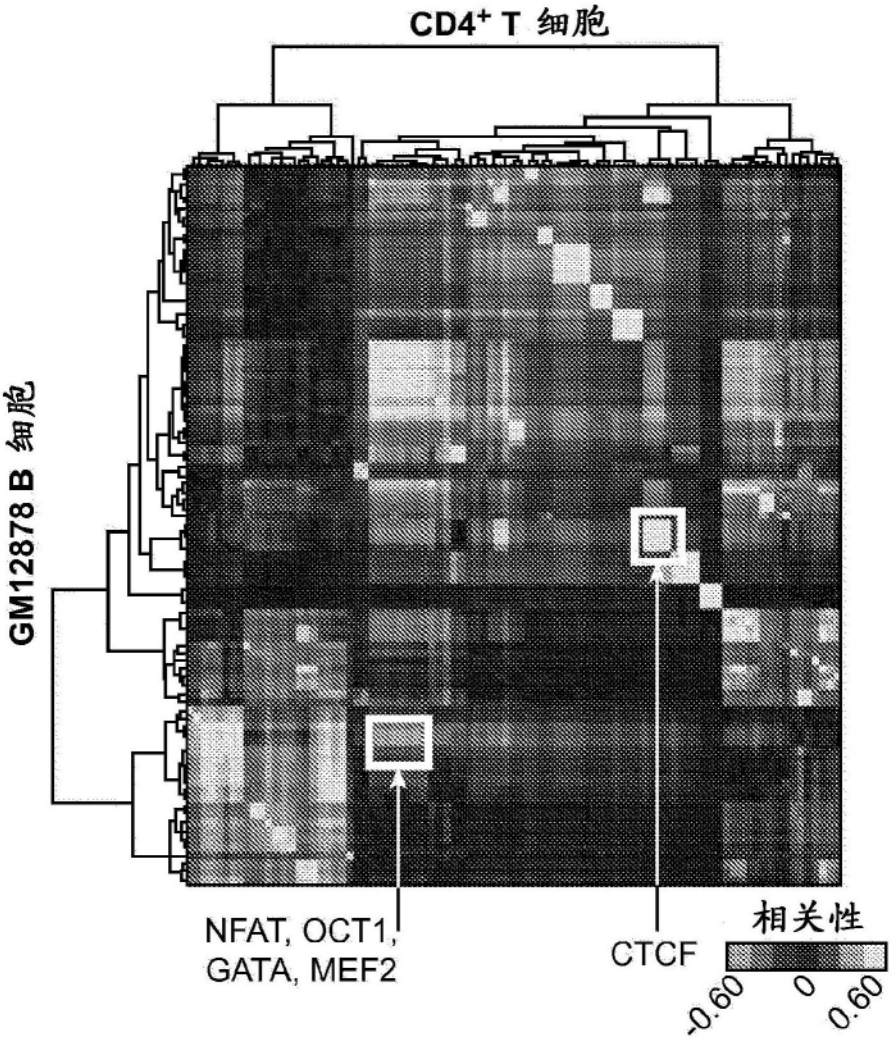


图5D

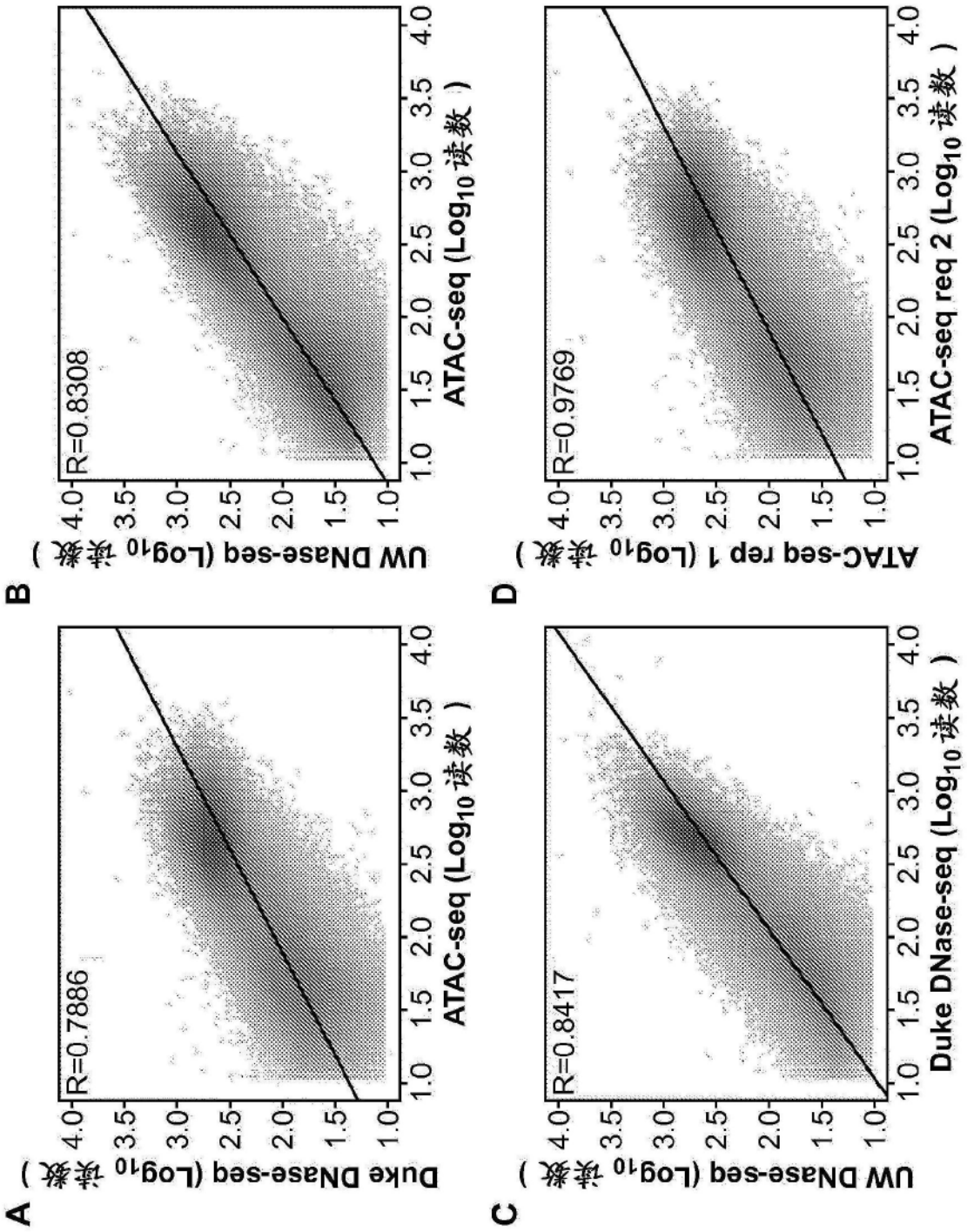
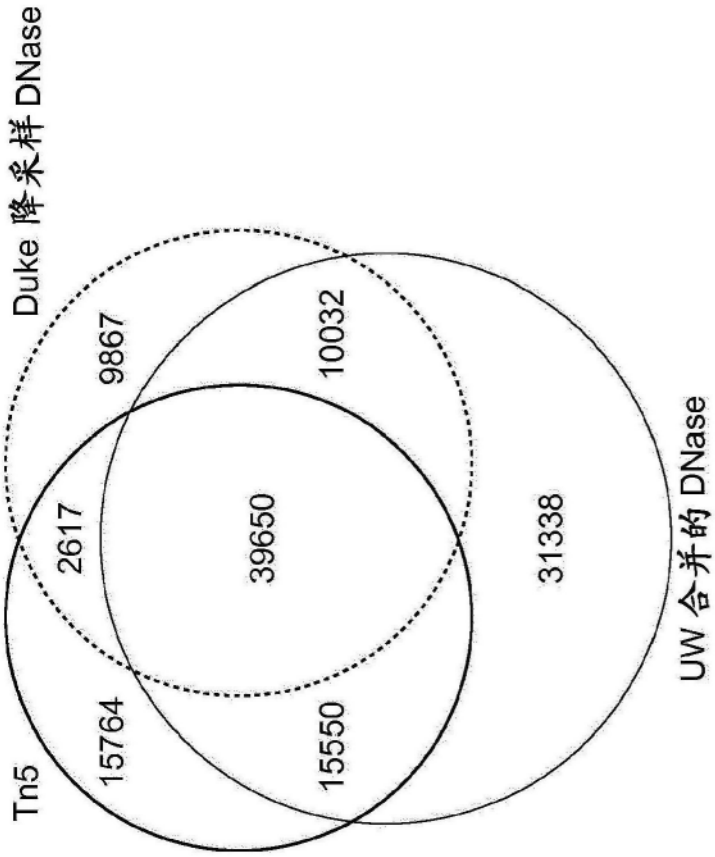

A
Duke DNase-seq (Log10 读数)
ATAC-seq (Log10 读数)
R=0.7886
B
UW DNase-seq (Log10 读数)
ATAC-seq (Log10 读数)
R=0.8308
C
UW DNase-seq (Log10 读数)
Duke DNase-seq (Log10 读数)
R=0.8417
D
ATAC-seq rep 1 (Log10 读数)
ATAC-seq req 2 (Log10 读数)
R=0.9769
1.0
1.5
2.0
2.5
3.0
3.5
4.0

图6



	唯一的ATAC	唯一的UW	唯一的Duke	ATAC和UW	ATAC和Duke	UW和Duke	交集
ATAC-Seq	6.30%	4.36%	1.08%	10.83%	1.53%	2.09%	73.80%
UW DNase	1.32%	8.13%	0.68%	7.57%	0.33%	4.34%	77.62%
Duke DNase	2.10%	5.27%	14.80%	4.02%	1.44%	6.91%	65.46%

图7

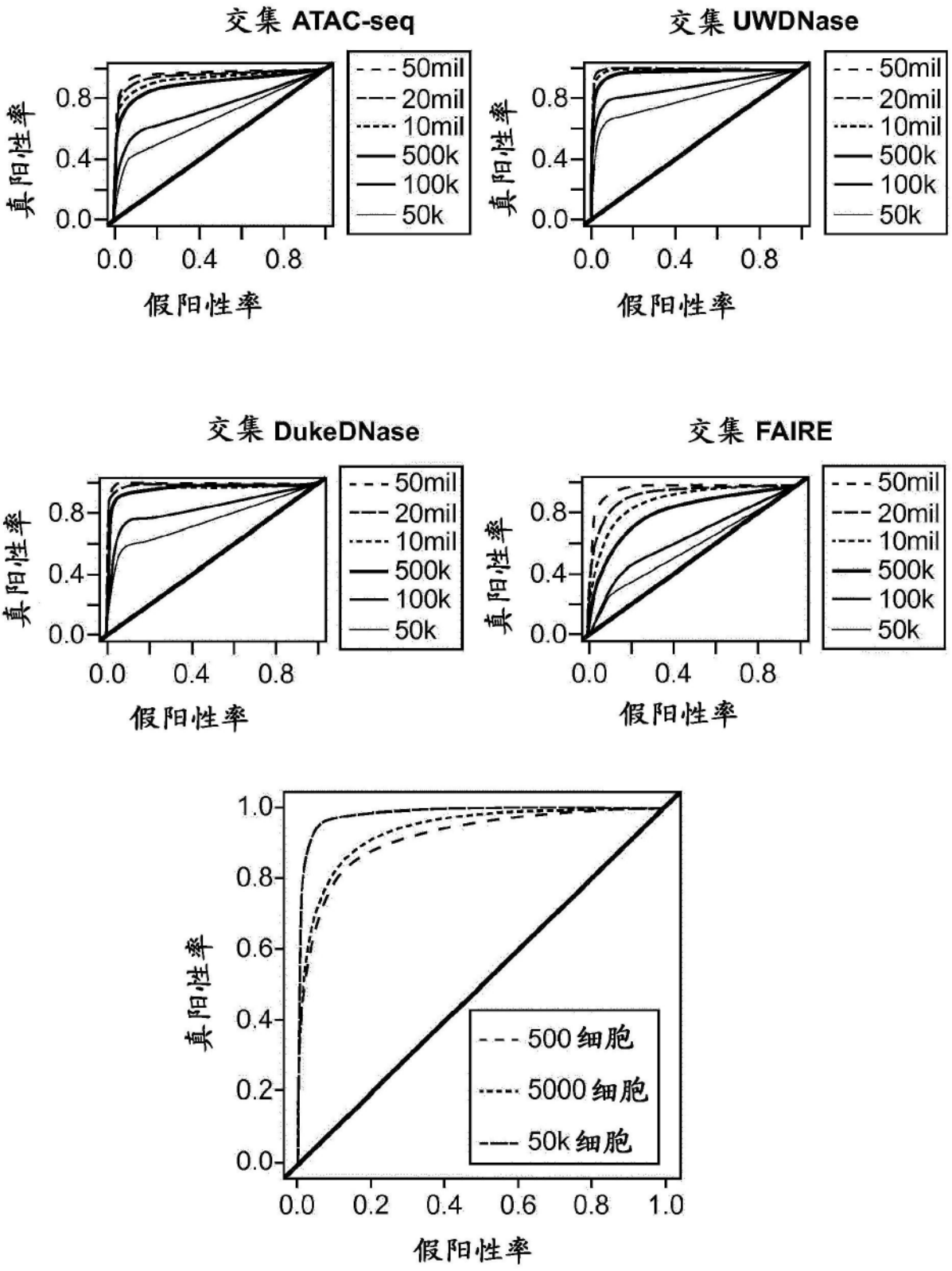

交集 ATAC-seq
交集 UWDNase
交集 DukeDNase
交集 FAIRE
真阳性率
假阳性率
50mil
20mil
10mil
500k
100k
50k
0.0
0.4
0.8
500 细胞
5000 细胞
50k 细胞
0.2
0.6
1.0

图8

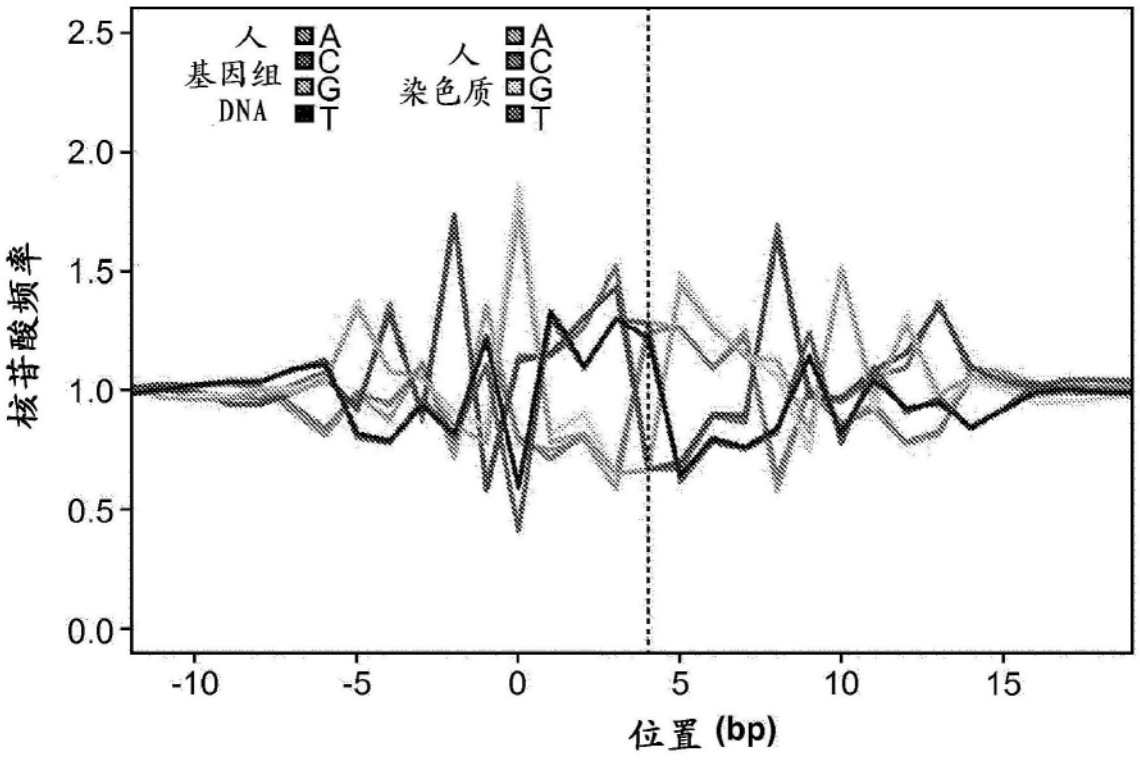


图9

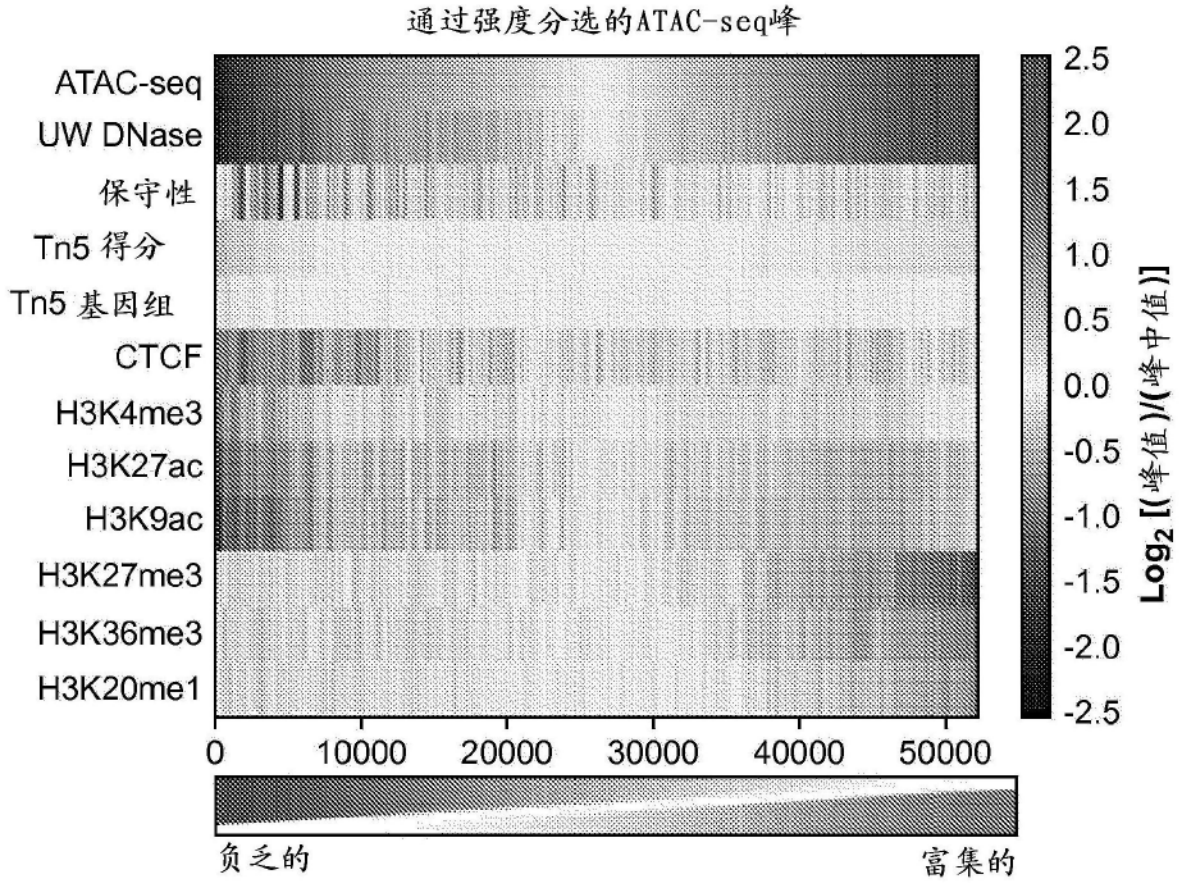

通过强度分选的ATAC-seq峰
ATAC-seq
UW DNase
保守性
Tn5 得分
Tn5 基因组
CTCF
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H3K36me3
H3K20me1
0
10000
20000
30000
40000
50000
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
Log2 [(峰值)/(峰中值)]
负乏的
富集的

图10

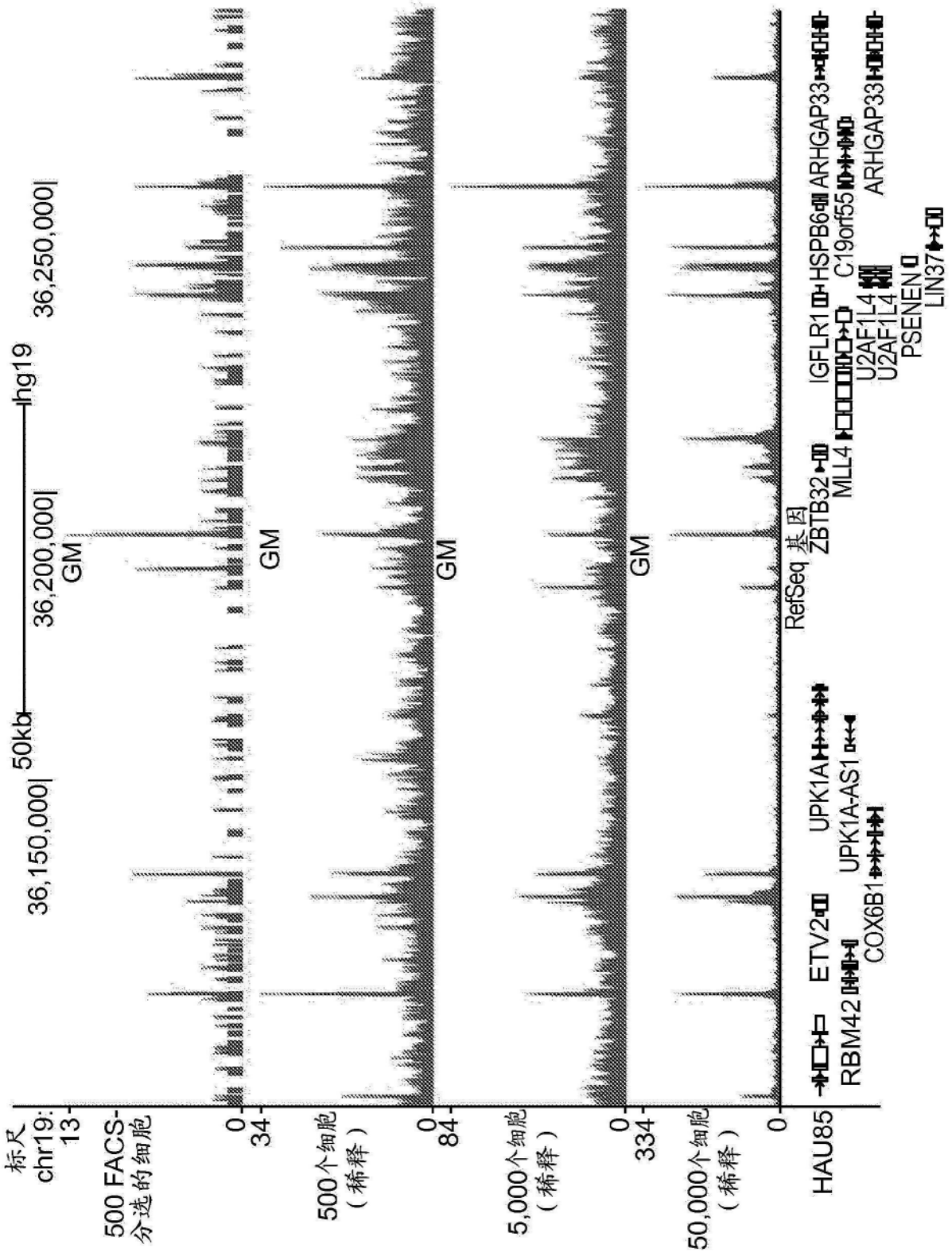

标尺
chr19:
13
500 FACS-分选的细胞
0
34
500个细胞（稀释）
0
84
5,000个细胞（稀释）
0
334
50,000个细胞（稀释）
0
36,150,000
50kb
hg19
36,200,000
36,250,000
GM
RefSeq 基因
HAU85
ETV2
RBM42
COX6B1
UPK1A
UPK1A-AS1
ZBTB32
MLL4
IGFLR1
HSPB6
ARHGAP33
C19orf55
U2AF1L4
PSENEN
LIN37

图11

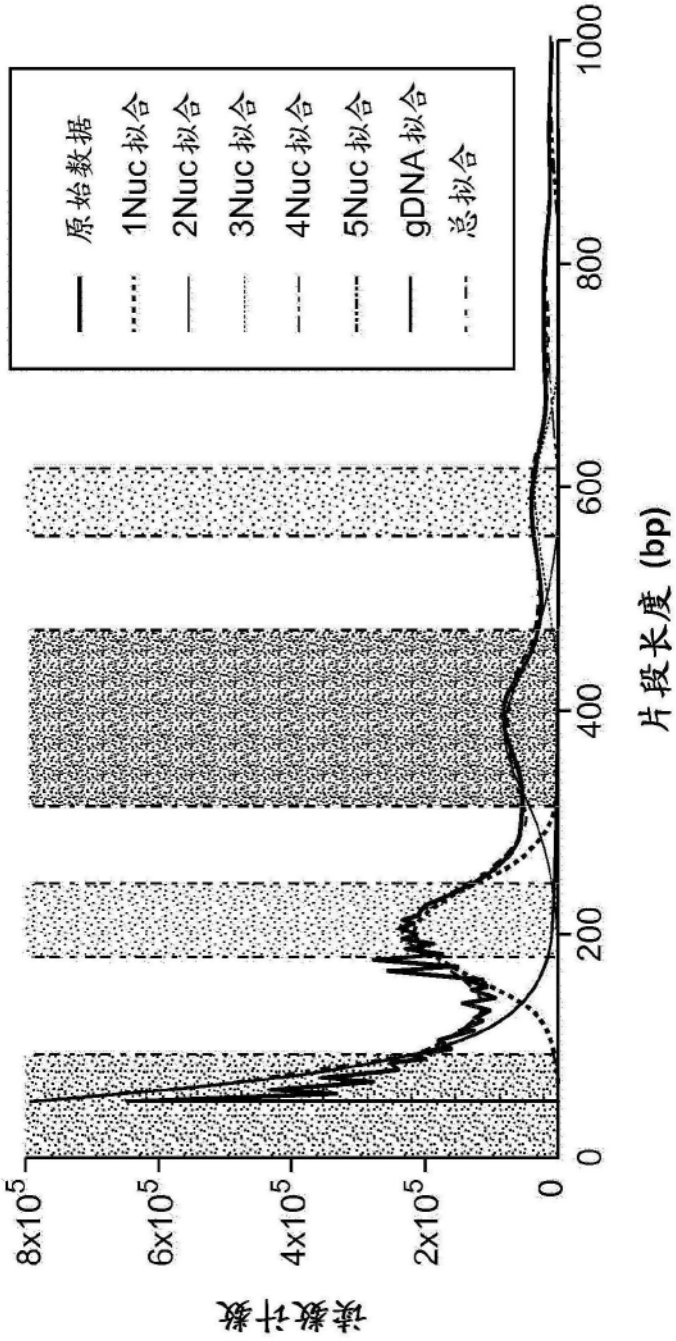

原始数据
1Nuc 拟合
2Nuc 拟合
3Nuc 拟合
4Nuc 拟合
5Nuc 拟合
gDNA 拟合
总拟合
读数计数
8x10^5
6x10^5
4x10^5
2x10^5
0
0
200
400
600
800
1000
片段长度 (bp)

图12

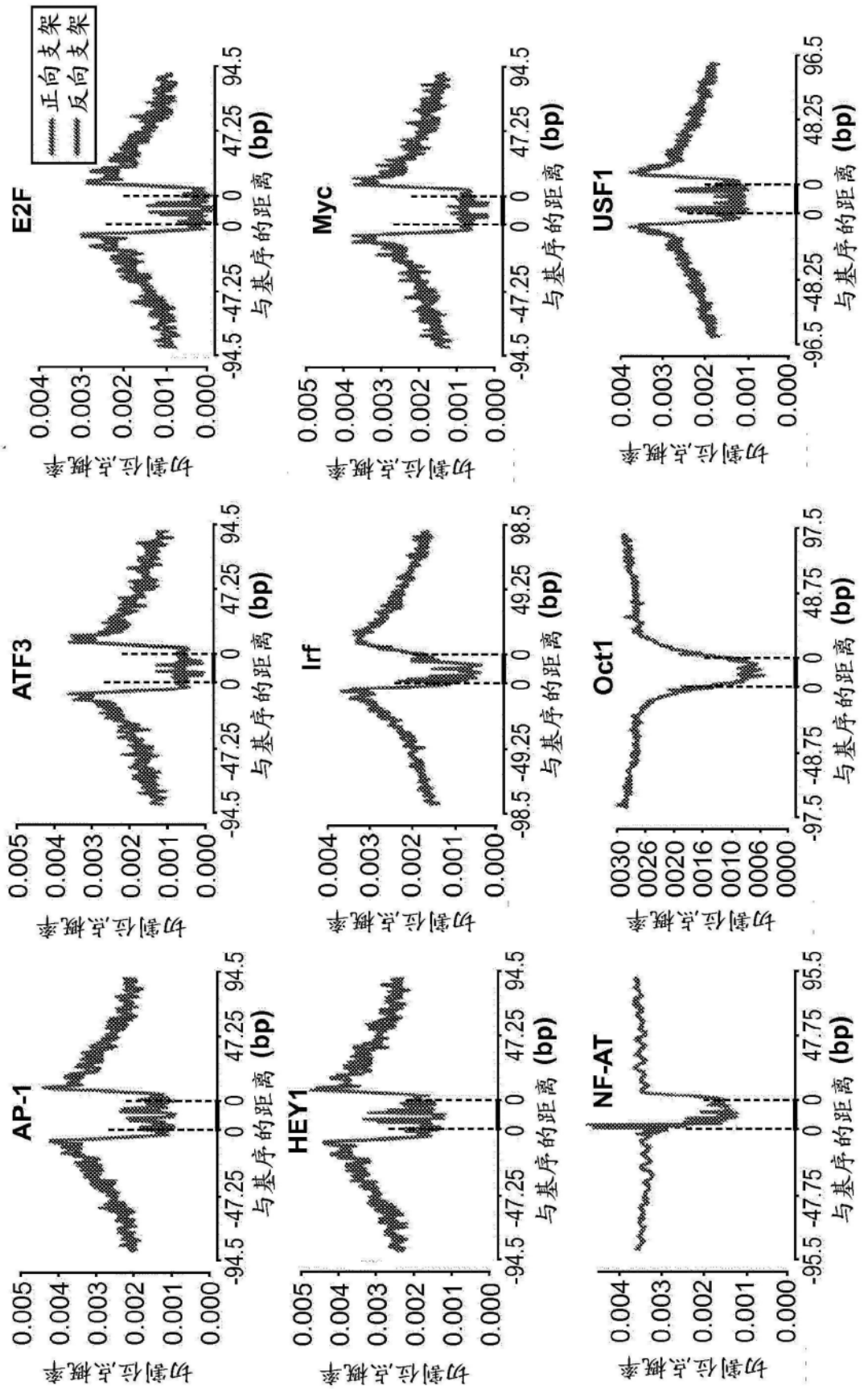

E2F
正向支架
反向支架
切割位点概率
与基序的距离 (bp)
Myc
USF1
ATF3
Irf
Oct1
AP-1
HEY1
NF-AT

图13

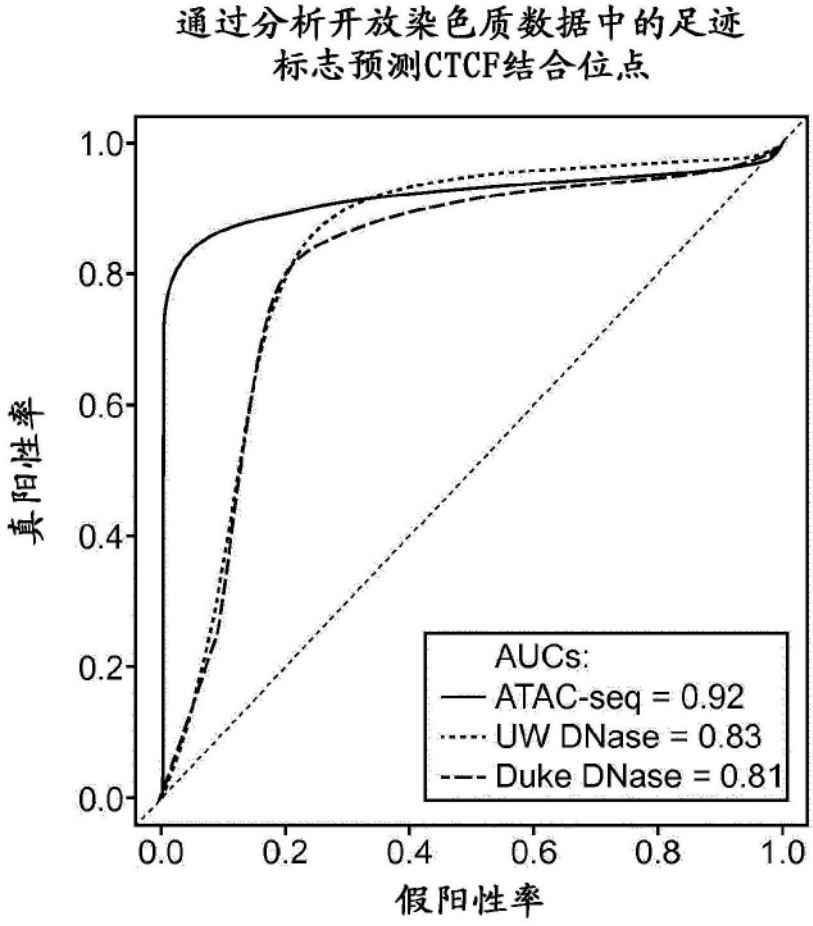

通过分析开放染色质数据中的足迹
标志预测CTCF结合位点
真阳性率
假阳性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
AUCs:
ATAC-seq = 0.92
UW DNase = 0.83
Duke DNase = 0.81

图14

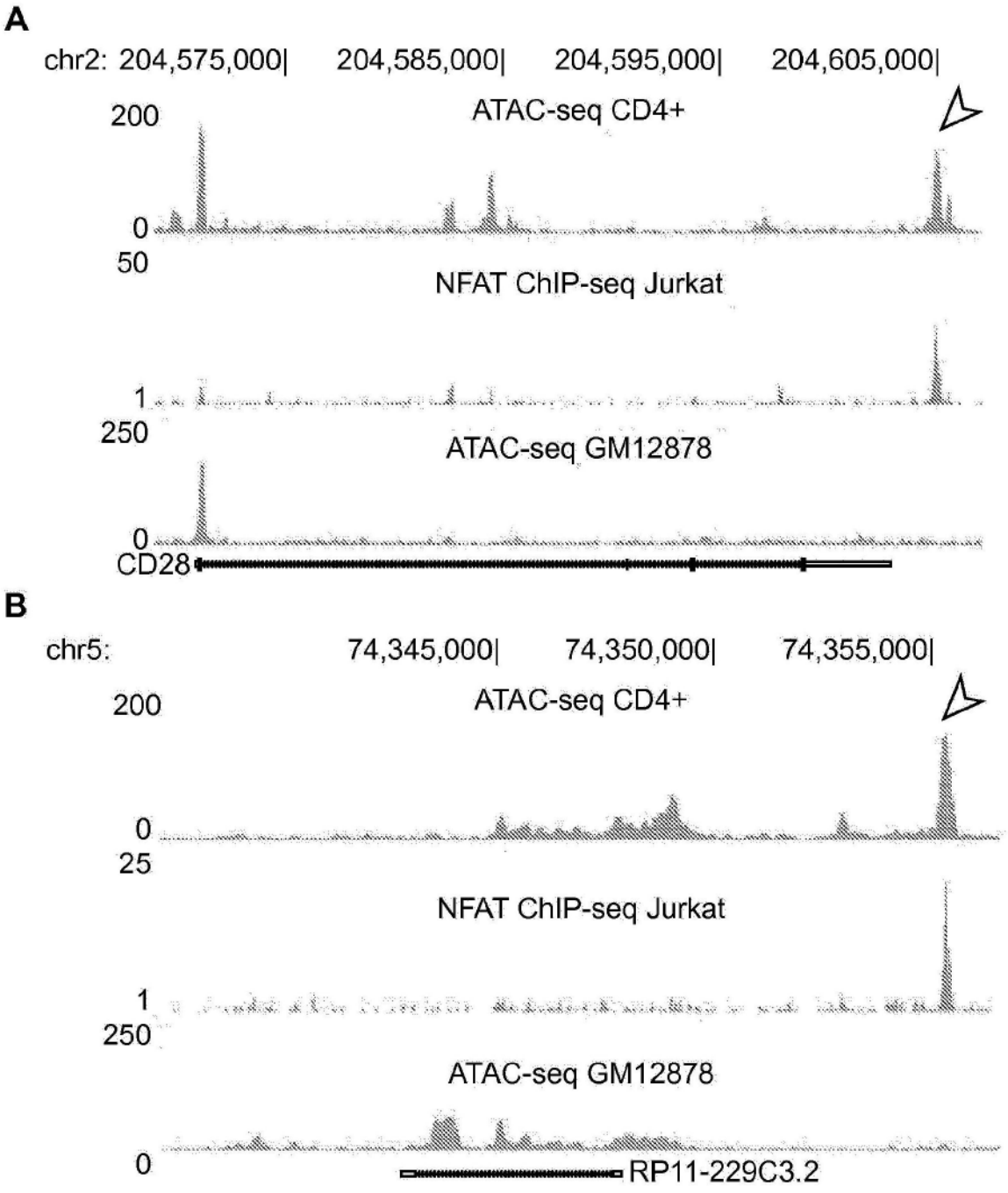

A
chr2: 204,575,000| 204,585,000| 204,595,000| 204,605,000|
200
ATAC-seq CD4+
0
50
NFAT ChIP-seq Jurkat
1
250
ATAC-seq GM12878
0
CD28
B
chr5: 74,345,000| 74,350,000| 74,355,000|
200
ATAC-seq CD4+
0
25
NFAT ChIP-seq Jurkat
1
250
ATAC-seq GM12878
0
RP11-229C3.2

图15

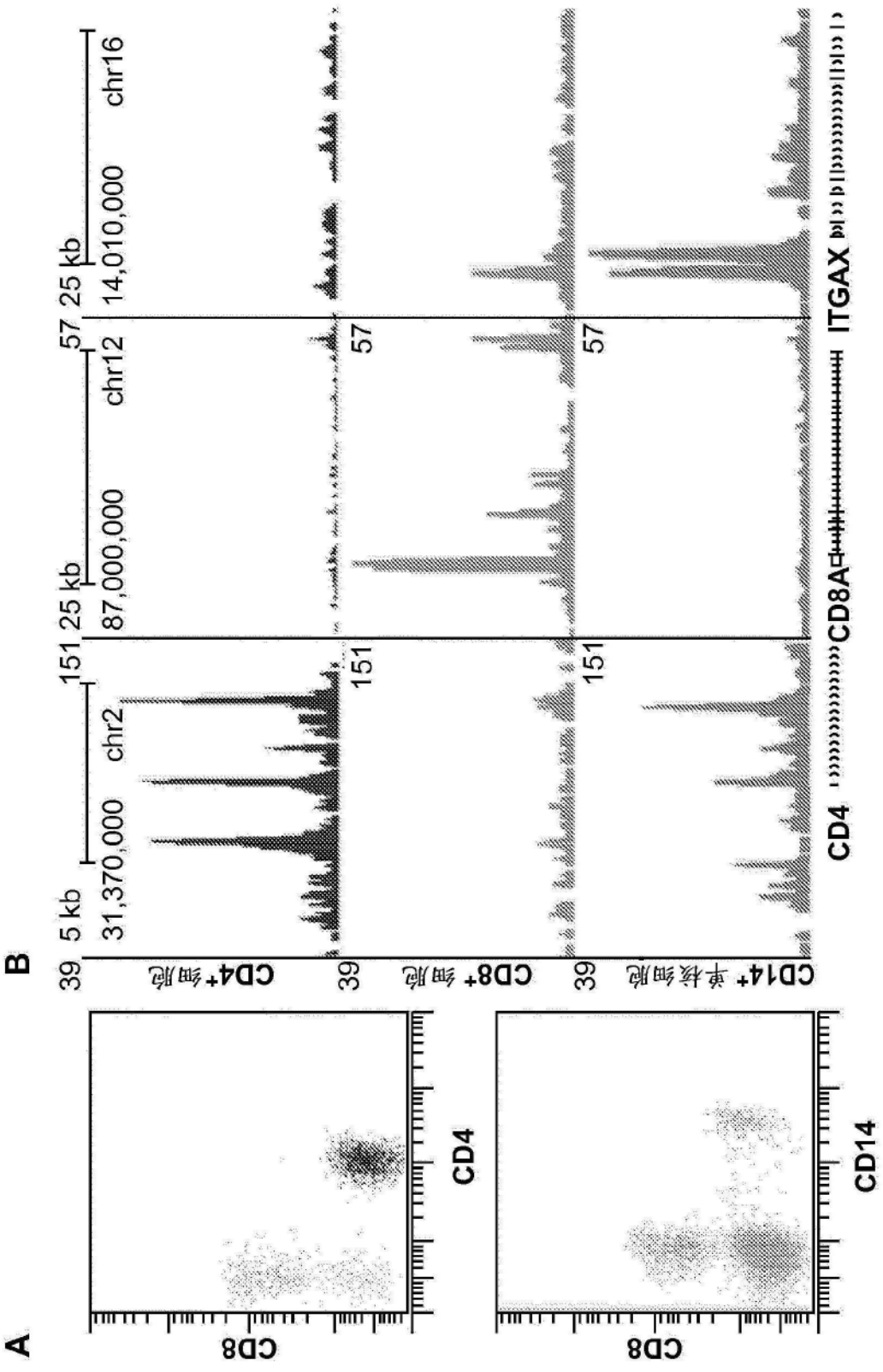

A
CD8
CD4
CD14
B
CD4+ 细胞
CD8+ 细胞
CD14+ 单核细胞
39
151
5 kb
31,370,000
chr2
25 kb
87,000,000
chr12
57
25 kb
14,010,000
chr16
CD4
CD8A
ITGAX

图16

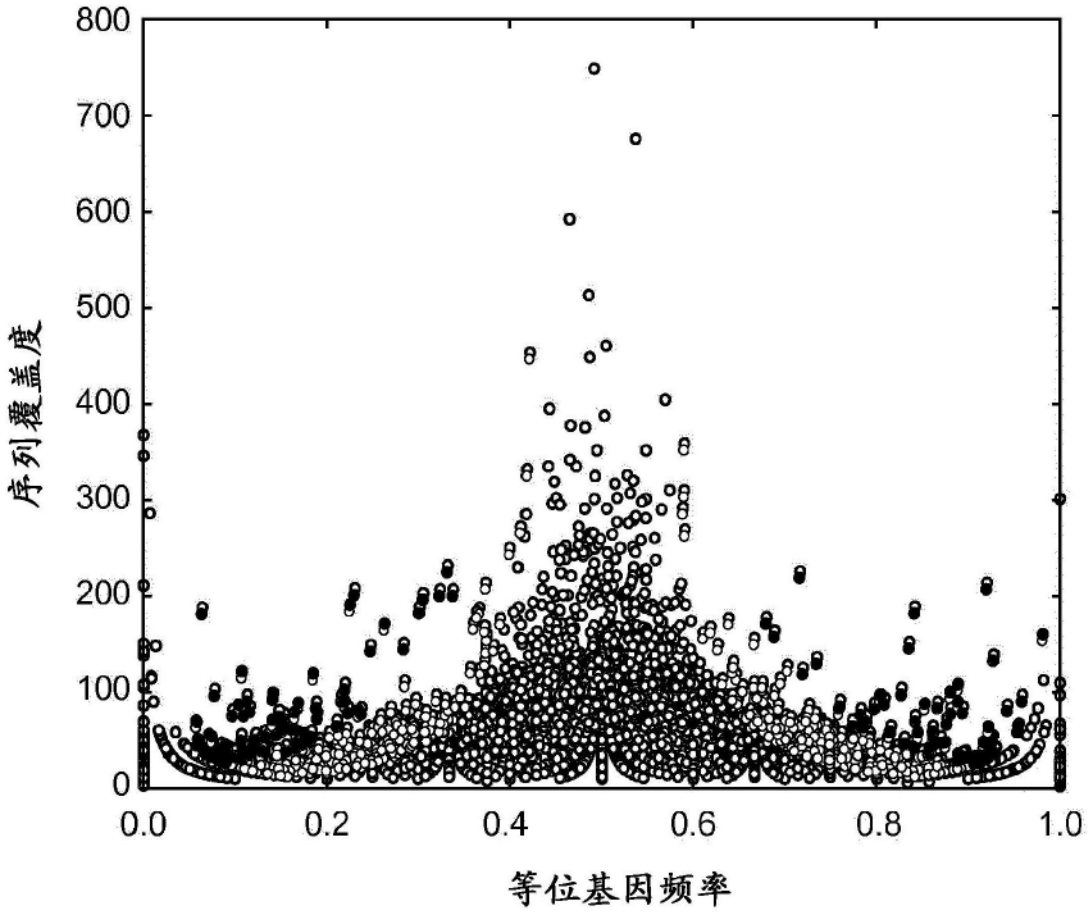

800
700
600
500
400
300
200
100
0
序列覆盖度
0.0
0.2
0.4
0.6
0.8
1.0
等位基因频率

图17

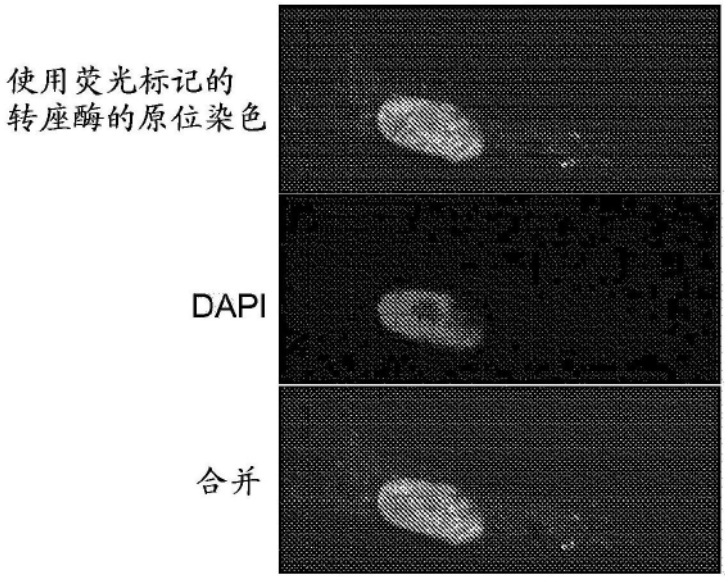

使用荧光标记的
转座酶的原位染色
DAPI
合并

图18

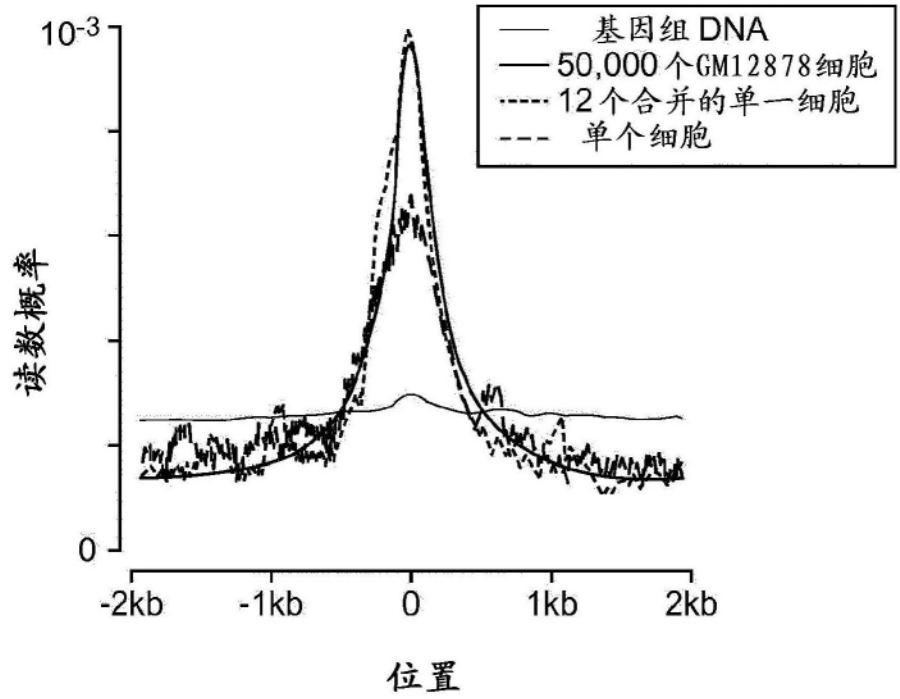

基因组 DNA
50,000个GM12878细胞
12个合并的单一细胞
单个细胞
10-3
0
读数概率
-2kb
-1kb
0
1kb
2kb
位置

图19

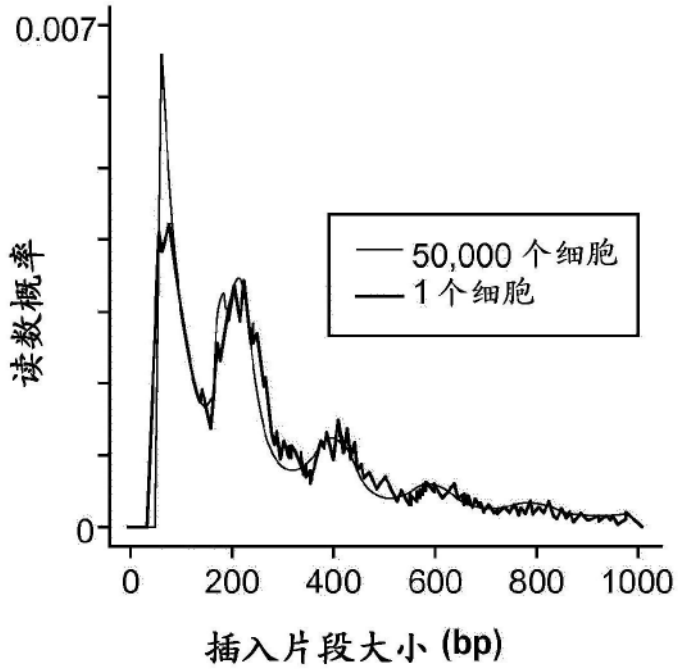

0.007
0
读数概率
50,000 个细胞
1 个细胞
0
200
400
600
800
1000
插入片段大小 (bp)

图20