(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112740239 B
(45)授权公告日 2025.02.28

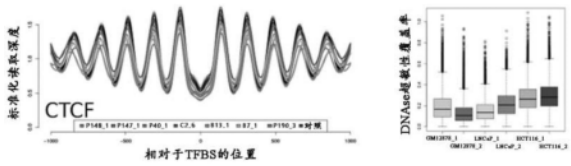
(21)申请号 201980061611.8
(22)申请日 2019.10.08
(65)同一申请的已公布的文献号
申请公布号 CN 112740239 A
(43)申请公布日 2021.04.30
(30)优先权数据
62/742,854 2018.10.08 US
62/752,270 2018.10.29 US
62/849,097 2019.05.16 US
(85)PCT国际申请进入国家阶段日
2021.03.19
(86)PCT国际申请的申请数据
PCT/US2019/055119 2019.10.08
(87)PCT国际申请的公布数据
WO2020/076772 EN 2020.04.16
(73)专利权人 福瑞诺姆控股公司
地址 美国加利福尼亚州
(72)发明人 彼得·乌尔兹
迈克尔·R·斯派克
埃伦·海策尔
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 徐爱文 武晶晶
(51)Int.Cl.
G16B 20/00(2019.01)
C12Q 1/6886(2018.01)
C12Q 1/6869(2018.01)
G16B 15/30(2019.01)
G16B 40/10(2019.01)
G16B 40/20(2019.01)
G16B 30/10(2019.01)
G06N 20/00(2019.01)
(56)对比文件
Matthew W. Snyder等.Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin.CELL.2016,第164卷(第1-2期),文章第57-68页,图1-6,图S1-S5,表S1-S5.
审查员 马晓晓
权利要求书7页 说明书49页 附图32页

(54)发明名称

转录因子分析

(57)摘要

本文公开的方法和系统可以改善基因组材料的分析能力。本文提供的方法可以检查转录因子结合位点可及性,以诊断疾病或监测受试者中疾病的进展。



CN 112740239 B

1.一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于检测人类受试者中癌症的存在或不存在的方法,所述方法包括:

(a)提供通过对从所述人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取的多个序列读取;

(b)将所述多个序列读取中的每个序列读取与人类参考基因组进行比对,以提供多个经比对的序列读取;

(c)对于转录因子的多个结合位点中的每个结合位点,选择所述多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供多个侧面基因组区域,其中相比于未患有所述癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,其中所述多个结合位点中的每一个均不同;

(d)对于所述多个侧面基因组区域中的每个侧面基因组区域,使用所述多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供多个读取深度覆盖模式;

(e)处理所述多个读取深度覆盖模式以提供信号,其中所述信号表征所述转录因子的所述多个结合位点的结合位点可及性,

(f)使用低通滤波器将所述多个读取深度覆盖模式中的每一个拆分为低频信号和高频信号,其中所述高频信号是一给定结合位点的结合位点可及性的量度,并且至少部分地基于所述高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分;和

(g)至少部分地基于所述可及性得分与参考可及性评分的偏差来检测所述人类受试者中所述癌症的所述存在或不存在。

2.根据权利要求1所述的计算机可读介质,其中所述方法进一步确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的所述可及性得分并将其输入机器学习模型中以训练能够区分健康受试者和癌症患者、癌症进展者和非进展者、多种癌症亚型、多个癌症分期、癌症治疗反应者和非反应者或其任何组合的分类器。

3.根据权利要求1所述的计算机可读介质,其中处理所述多个读取深度覆盖模式包括使用Savitzky-Golay滤波器。

4.根据权利要求1所述的计算机可读介质,其中所述转录因子是上皮转录因子。

5.根据权利要求1所述的计算机可读介质,其中所述转录因子是癌症特异性转录因子。

6.根据权利要求5所述的计算机可读介质,其中所述转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B、FOXA1、HNF-1A、HNF-4A、HNF-4G和DLX-2。

7.根据权利要求1所述的计算机可读介质,其中(g)还包括鉴定所述转录因子的较高可及性的信号。

8.根据权利要求1所述的计算机可读介质,其中所述方法以至少70%的准确度检测所述人类受试者中所述癌症的所述存在或不存在。

9.根据权利要求1所述的计算机可读介质,其中所述方法以至少80%的准确度检测所述人类受试者中所述癌症的所述存在或不存在。

10.根据权利要求1所述的计算机可读介质,其中所述方法以至少90%的准确度检测所

述人类受试者中所述癌症的所述存在或不存在。

11.根据权利要求1所述的计算机可读介质,其中所述方法以至少70％的灵敏度检测所述人类受试者中所述癌症的所述存在。

12.根据权利要求1所述的计算机可读介质,其中所述方法以至少80％的灵敏度检测所述人类受试者中所述癌症的所述存在。

13.根据权利要求1所述的计算机可读介质,其中所述方法以至少90％的灵敏度检测所述人类受试者中所述癌症的所述存在。

14.根据权利要求1所述的计算机可读介质,其中所述方法以至少70％的特异性检测所述人类受试者中所述癌症的所述不存在。

15.根据权利要求1所述的计算机可读介质,其中所述方法以至少80％的特异性检测所述人类受试者中所述癌症的所述不存在。

16.根据权利要求1所述的计算机可读介质,其中所述方法以至少90％的特异性检测所述人类受试者中所述癌症的所述不存在。

17.根据权利要求1所述的计算机可读介质,其中所述方法以至少70％的阳性预测值(PPV)检测所述人类受试者中所述癌症的所述存在。

18.根据权利要求1所述的计算机可读介质,其中所述方法以至少80％的阳性预测值(PPV)检测所述人类受试者中所述癌症的所述存在。

19.根据权利要求1所述的计算机可读介质,其中所述方法以至少90％的阳性预测值(PPV)检测所述人类受试者中所述癌症的所述存在。

20.根据权利要求1所述的计算机可读介质,其中所述方法以至少70％的阴性预测值(NPV)检测所述人类受试者中所述癌症的所述不存在。

21.根据权利要求1所述的计算机可读介质,其中所述方法以至少80％的阴性预测值(NPV)检测所述人类受试者中所述癌症的所述不存在。

22.根据权利要求1所述的计算机可读介质,其中所述方法以至少90％的阴性预测值(NPV)检测所述人类受试者中所述癌症的所述不存在。

23.根据权利要求1所述的计算机可读介质,其中所述方法以至少0.70的接受者操作特征下面积(AUROC)检测所述人类受试者中所述癌症的所述存在或不存在。

24.根据权利要求1所述的计算机可读介质,其中所述方法以至少0.80的接受者操作特征下面积(AUROC)检测所述人类受试者中所述癌症的所述存在或不存在。

25.根据权利要求1所述的计算机可读介质,其中所述方法以至少0.90的接受者操作特征下面积(AUROC)检测所述人类受试者中所述癌症的所述存在或不存在。

26.根据权利要求1所述的计算机可读介质,其中所述方法包括将训练过的分类器应用于所述信号以检测所述人类受试者中所述癌症的所述存在或不存在。

27.根据权利要求1所述的计算机可读介质,其中所述方法包括将训练过的分类器应用于所述转录因子的所述结合位点的所述可及性得分以检测所述人类受试者中所述癌症的所述存在或不存在。

28.根据权利要求27所述的计算机可读介质,其中所述训练过的分类器包括训练过的机器学习分类器。

29.根据权利要求28所述的计算机可读介质,其中所述训练过的机器学习分类器包含

监督机器学习算法。

30.根据权利要求29所述的计算机可读介质,其中所述监督机器学习算法包括以下中的一项或多项:回归、支持向量机、基于树的方法、神经网络和随机森林。

31.一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于监测人类受试者中癌症的进展或消退的方法,所述方法包括:

(a)提供通过对在第一时间从所述人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取的多个第一序列读取以及通过对在晚于所述第一时间的第二时间从所述人类受试者获得或者源自所述人类受试者的cfDNA进行全基因组测序而获取的多个第二序列读取;

(b)将所述多个第一序列读取和所述多个第二序列读取中的每个序列读取与人类参考基因组进行比对,以提供第一多个经比对的序列读取和第二多个经比对的序列读取;

(c)对于转录因子的多个结合位点中的每个结合位点,选择所述第一多个经比对的序列读取和所述第二多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供第一多个侧面基因组区域和第二多个侧面基因组区域,其中相比于未患有所述癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,其中所述多个结合位点中的每一个均不同;

(d)对于所述第一多个侧面基因组区域和所述第二多个侧面基因组区域中的每个侧面基因组区域,使用所述第一多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第一读取深度覆盖模式,并使用所述第二多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第二读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供第一多个读取深度覆盖模式和第二多个读取深度覆盖模式;

(e)处理所述第一多个读取深度覆盖模式以提供第一信号,并处理所述第二多个读取深度覆盖模式以提供第二信号,其中所述第一信号和所述第二信号表征所述转录因子的所述多个结合位点的结合位点可及性;

(f)使用低通滤波器将所述第一多个读取深度覆盖模式中的每一个拆分为第一低频信号和第一高频信号,所述第二多个读取深度覆盖模式中的每一个拆分为第二低频信号和第二高频信号,其中所述第一高频信号和所述第二高频信号是一给定结合位点的结合位点可及性的量度;

(g)至少部分地基于所述第一高频信号和所述第二高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分;以及

(h)至少部分地基于所述可及性得分与参考可及性评分的偏差来监测所述人类受试者中所述癌症的所述进展或消退。

32.根据权利要求31所述的计算机可读介质,其中处理所述第一多个读取深度覆盖模式包括使用Savitzky-Golay滤波器。

33.根据权利要求31所述的计算机可读介质,其中所述转录因子是癌症特异性转录因子。

34.根据权利要求33所述的计算机可读介质,其中所述转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B、FOXA1、HNF-1a、HNF-4a、HNF-4g和DLX-2。

35.根据权利要求31所述的计算机可读介质,进一步包括基于所监测的所述人类受试者中所述癌症的所述进展或消退,调整用于所述人类受试者的所述癌症的治疗方案。

36.一种用于监测人类受试者中癌症的进展或消退的系统,所述系统包括:

数据库,其包含通过对在第一时间从所述人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取多个第一序列读取,并通过对在晚于所述第一时间的第二时间从所述人类受试者获得或者源自所述人类受试者的cfDNA进行全基因组测序而获取多个第二序列读取,其中所述多个第一序列读取和所述多个第二序列读取与人类参考基因组进行比对,以提供第一多个经比对的序列读取和第二多个经比对的序列读取;和

一个或多个计算机处理器,其可操作地与所述数据库偶联,其中所述一个或多个计算机处理器被单独地或共同地编程为:

(a)对于转录因子的多个结合位点中的每个结合位点,选择所述第一多个经比对的序列读取和所述第二多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供第一多个侧面基因组区域和第二多个侧面基因组区域,其中相比于未患有所述癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,其中所述多个结合位点中的每一个均不同;

(b)对于所述第一多个侧面基因组区域和所述第二多个侧面基因组区域中的每个侧面基因组区域,使用所述第一多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第一读取深度覆盖模式,并使用所述第二多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第二读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供第一多个读取深度覆盖模式和第二多个读取深度覆盖模式;

(c)处理所述第一多个读取深度覆盖模式以提供第一信号,并处理所述第二多个读取深度覆盖模式以提供第二信号,其中所述第一信号和所述第二信号表征所述转录因子的所述多个结合位点的结合位点可及性;

(d)使用低通滤波器将所述第一多个读取深度覆盖模式中的每一个拆分为第一低频信号和第一高频信号,所述第二多个读取深度覆盖模式中的每一个拆分为第二低频信号和第二高频信号,其中所述第一高频信号和所述第二高频信号是一给定结合位点的结合位点可及性的量度;

(e)至少部分地基于所述第一高频信号和所述第二高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分;和

(f)至少部分地基于所述可及性得分与参考可及性评分的偏差来监测所述人类受试者中所述癌症的所述进展或消退。

37.一种用于确定肿瘤特异性转录因子结合位点(TFBS)模式的计算机实现的方法,所述方法包括:

(a)提供通过对在第一时间从人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取的多个第一序列读取和通过对在晚于所述第一时间的第二时间从所述人类受试者获得或者源自所述人类受试者的cfDNA进行全基因组测序而获取的多个第二序列读取;

(b)将所述多个第一序列读取和所述多个第二序列读取中的每个序列读取与人类参考

基因组进行比对,以提供第一多个经比对的序列读取和第二多个经比对的序列读取;

(c)对于转录因子的多个结合位点中的每个结合位点,选择所述第一多个经比对的序列读取和所述第二多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供第一多个侧面基因组区域和第二多个侧面基因组区域,其中相比于未患有癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,其中所述多个结合位点中的每一个均不同;

(d)对于所述第一多个侧面基因组区域和所述第二多个侧面基因组区域中的每个侧面基因组区域,使用所述第一多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第一读取深度覆盖模式,并使用所述第二多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第二读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供第一多个读取深度覆盖模式和第二多个读取深度覆盖模式;

(e)处理所述第一多个读取深度覆盖模式以提供第一信号,并处理所述第二多个读取深度覆盖模式以提供第二信号,其中所述第一信号和所述第二信号表征所述转录因子的所述多个结合位点的结合位点可及性;

(f)使用低通滤波器将所述第一多个读取深度覆盖模式中的每一个拆分为第一低频信号和第一高频信号,所述第二多个读取深度覆盖模式中的每一个拆分为第二低频信号和第二高频信号,其中所述第一高频信号和所述第二高频信号是一给定结合位点的结合位点可及性的量度;

(g)至少部分地基于所述第一高频信号和所述第二高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分;以及

(h)至少部分地基于所述可及性得分与参考可及性评分的偏差来确定所述人类受试者中所述肿瘤特异性TFBS模式。

38.一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于确定肿瘤特异性转录因子结合位点(TFBS)模式的方法,所述方法包括:

(a)提供通过对在第一时间从人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取的多个第一序列读取和通过对在晚于所述第一时间的第二时间从所述人类受试者获得或者源自所述人类受试者的cfDNA进行全基因组测序而获取的多个第二序列读取;

(b)将所述多个第一序列读取和所述多个第二序列读取中的每个序列读取与人类参考基因组进行比对,以提供第一多个经比对的序列读取和第二多个经比对的序列读取;

(c)对于转录因子的多个结合位点中的每个结合位点,选择所述第一多个经比对的序列读取和所述第二多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供第一多个侧面基因组区域和第二多个侧面基因组区域,其中相比于未患有癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,其中所述多个结合位点中的每一个均不同;

(d)对于所述第一多个侧面基因组区域和所述第二多个侧面基因组区域中的每个侧面基因组区域,使用所述第一多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第一读取深度覆盖模式,并使用所述第二多个经比对的序列读取的至少一部分来

生成所述转录因子的结合位点的第二读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供第一多个读取深度覆盖模式和第二多个读取深度覆盖模式;

(e)处理所述第一多个读取深度覆盖模式以提供第一信号,并处理所述第二多个读取深度覆盖模式以提供第二信号,其中所述第一信号和所述第二信号表征所述转录因子的所述多个结合位点的结合位点可及性;

(f)使用低通滤波器将所述第一多个读取深度覆盖模式中的每一个拆分为第一低频信号和第一高频信号,所述第二多个读取深度覆盖模式中的每一个拆分为第二低频信号和第二高频信号,其中所述第一高频信号和所述第二高频信号是一给定结合位点的结合位点可及性的量度;

(g)至少部分地基于所述第一高频信号和所述第二高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分;以及

(h)至少部分地基于所述可及性得分与参考可及性评分的偏差来确定所述人类受试者中所述肿瘤特异性TFBS模式。

39.一种用于确定肿瘤特异性转录因子结合位点(TFBS)模式的系统,所述系统包括:

数据库,其包含通过对在第一时间从人类受试者获得或者源自所述人类受试者的无细胞脱氧核糖核酸(cfDNA)进行全基因组测序而获取多个第一序列读取,并通过对在晚于所述第一时间的第二时间从所述人类受试者获得或者源自所述人类受试者的cfDNA进行全基因组测序而获取多个第二序列读取,其中所述多个第一序列读取和所述多个第二序列读取与人类参考基因组进行比对,以提供第一多个经比对的序列读取和第二多个经比对的序列读取;和

一个或多个计算机处理器,其可操作地与所述数据库偶联,其中所述一个或多个计算机处理器被单独地或共同地编程为:

(a)对于转录因子的多个结合位点中的每个结合位点,选择所述第一多个经比对的序列读取和所述第二多个经比对的序列读取中的位于所述结合位点侧面的基因组区域以提供第一多个侧面基因组区域和第二多个侧面基因组区域,其中相比于未患有癌症的第二人群,所述多个结合位点中的每个结合位点在患有所述癌症的第一人群中表现出对所述转录因子的靶增强子的差异可及性,

其中所述多个结合位点中的每一个均不同;

(b)对于所述第一多个侧面基因组区域和所述第二多个侧面基因组区域中的每个侧面基因组区域,使用所述第一多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第一读取深度覆盖模式,并使用所述第二多个经比对的序列读取的至少一部分来生成所述转录因子的结合位点的第二读取深度覆盖模式,从而分别为所述转录因子的所述多个结合位点提供第一多个读取深度覆盖模式和第二多个读取深度覆盖模式;

(c)处理所述第一多个读取深度覆盖模式以提供第一信号,并处理所述第二多个读取深度覆盖模式以提供第二信号,其中所述第一信号和所述第二信号表征所述转录因子的所述多个结合位点的结合位点可及性;

(d)使用低通滤波器将所述第一多个读取深度覆盖模式中的每一个拆分为第一低频信号和第一高频信号,所述第二多个读取深度覆盖模式中的每一个拆分为第二低频信号和第二高频信号,其中所述第一高频信号和所述第二高频信号是一给定结合位点的结合位点可

及性的量度;和

(e)至少部分地基于所述第一高频信号和所述第二高频信号确定所述转录因子的所述多个结合位点中的每一个的可及性得分,从而至少部分地基于所述可及性得分与参考可及性评分的偏差来确定所述肿瘤特异性TFBS模式。

转录因子分析

[0001] 交叉引用

[0002] 本申请要求于2018年10月8日提交的美国临时专利申请62/742,854、于2018年10月29日提交的美国临时专利申请62/752,270和于2019年5月16日提交的美国临时专利申请62/849,097的权益,其中每个申请均通过引用整体并入本文。

背景技术

[0003] 转录因子(TF)可调节其靶基因的表达,并可在发育和分化中起关键作用。基因组改变可导致TF的活化或失活,并且由此产生的基因调控紊乱可导致生理状况(例如衰老)或基础疾病(例如癌症)。为了结合调控性的脱氧核糖核酸(DNA),TF通常必须与核小体相互作用,这可能会影响其占据和定位。

[0004] 转录因子的改变可能是癌症中肿瘤发生的重要驱动力,并且TF核小体相互作用在很大程度上尚未被映射。但是,缺乏用于评估转录因子活性的非侵入性测定法。

[0005] 鉴于TF在调控染色质可及性和转录中的作用,了解遗传变异对TF结合的影响可以提供对发育和疾病的非编码遗传组分的见解。可以使用高通量、全基因组方法分别测定染色质可及性(“开放染色质”)、核小体定位和转录因子(TF)占据,来获得对染色质核蛋白结构内编码的表观遗传信息的主要见解。

[0006] 转录因子(TF)的失调可能是肿瘤发生的重要驱动力。为了使TF结合DNA,结合区可能需要是可及的。因此,TF和染色质重塑复合物移动并定位核小体以增强可及性。因此,需要一种剖析转录因子结合位点以推断核小体位置和染色质可及性的方法。还需要使用转录因子结合位点剖析和转录因子结合位点特征来推断疾病状态、疾病进展和治疗反应性的方法。

发明内容

[0007] 本公开内容提供了用于评估(例如,建模)转录因子(TF)结合位点(TFBS)和使用TFBS信息来检测、评估、诊断和分析疾病状态并鉴定治疗反应性的方法和系统。

[0008] 基于下一代测序的全基因组测定法可用于提供TF结合模式和相关的染色质架构。由于核小体和序列特异性TF以相互排斥的方式结合调控性的脱氧核糖核酸(DNA)区域,因此TF会与核小体竞争或相互作用,这影响其占据和定位。在给定的细胞群体中,核小体占据是指在指定的基因组区域内测得的核小体的平均数量,而核小体定位表示在特定基因组坐标处存在的核小体上参考点(通常为二分体(dyad),例如典型核小体的中点)的概率。

[0009] 无细胞循环核酸,例如无细胞DNA(cfDNA),可以为TFBS分析提供易于获得的核酸来源。此类cfDNA可以是消化过程的产物,该消化过程优先降解不受蛋白质(例如组蛋白复合物)保护的DNA。无细胞DNA覆盖模式可反映核小体定位和由转录因子主动结合基因组引起的占据。然后,通过cfDNA测量的这些核小体占据模式可用于推断正常基因组和肿瘤基因组中TF的活性。

[0010] TF可优先结合在开放染色质中,这可能影响核小体定位。来自血浆的循环无细胞

DNA可代表单核小体DNA,并且核小体血浆足迹可以是有关TFBS的信息。

[0011] 本公开内容提供了用于对来自cfDNA的核小体位置进行绘图的方法和系统,以提供关于TF的信息,以用于与疾病鉴定、预测、分期和/或鉴定治疗反应性有关的应用。本文描述了用于使用从核酸分子(例如cfDNA)中的核小体足迹确定的转录因子信息的方法和系统。来自核酸分子中的核小体足迹的信息可用于评价、评估、检测和诊断疾病,例如癌症。在一些示例中,该信息可以被特征化并用作在许多这些应用(例如疾病鉴定、预测、分期和鉴定治疗反应性)中有用的机器学习模型的输入。

[0012] 在一方面,本公开内容提供了一种用于确定来自受试者的核酸样品中的转录因子结合谱的计算机实现的方法,该方法包括:(a)提供来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;(b)使用所述序列读取集来产生转录因子的覆盖模式;(c)处理覆盖模式以提供信号;和(d)用参考信号处理信号,其中信号和参考信号具有不同的频率,从而确定样品中的转录因子结合谱。

[0013] 在一些示例中,DNA是无细胞DNA。

[0014] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-Golay滤波器。

[0015] 在一些示例中,受试者是人。

[0016] 在另一方面,本公开内容提供了一种用于检测受试者中疾病的存在或不存在的计算机实现的方法,该方法包括:(a)提供来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;(b)使用所述序列读取集来产生转录因子的覆盖模式;(c)处理覆盖模式以提供信号;和(d)用参考信号处理信号,其中信号和参考信号具有不同的频率,从而检测所述受试者中所述疾病的所述存在或不存在。

[0017] 在一些示例中,DNA是无细胞DNA。

[0018] 在一些示例中,疾病是癌症。

[0019] 在一些示例中,(b)包括将序列读取集与参考序列比对以提供经比对的序列模式,选择所述经比对的序列模式的与转录因子的结合位点相对应的区域,以及在该区域中对所述经比对的序列模式进行标准化。在一些示例中,(d)包括计算转录因子的每个结合位点的可及性得分。

[0020] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-Golay滤波器。

[0021] 在一些示例中,受试者是人。

[0022] 在一些示例中,转录因子是癌症特异性转录因子。在一些示例中,转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B、FOXA1、HNF-1A、HNF-4A、HNF-4G和DLX-2。

[0023] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分健康受试者和癌症患者、疾病进展者和非进展者、多种疾病亚型、多个疾病分期、疾病治疗反应者和非反应者或其任何组合的分类器。

[0024] 在一些示例中,转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B和FOXA1。

[0025] 在一些示例中,观察到患者特异性和/或肿瘤特异性模式,包括转录因子AR、

HOXB13和NKX3-1的推断的结合模式。

[0026] 在一些示例中,(d)包括鉴定转录因子的较高可及性的信号。在一些示例中,转录因子是上皮转录因子。在一些示例中,转录因子是GRH-L2。

[0027] 在一些示例中,转录因子GRHL2、FOXA1和ZNF121与乳腺癌患者中增加的可及性得分或开放染色质可及性相关。

[0028] 在一些示例中,转录因子GRHL2、FOXA1和ZNF121的开放可及性指示乳腺癌。

[0029] 在一些示例中,转录因子EVX2、DLX2、HNF1A、HNF4A、GRHL2和HNF4G与结肠癌患者中增加的可及性得分或开放染色质可及性相关。

[0030] 在一些示例中,转录因子EVX2、DLX2、HNF1A、GRH-L2、HNF4A和HNF4G的开放可及性指示结肠癌。

[0031] 在一些示例中,转录因子LYL1和PU.1与结肠癌患者中降低的可及性得分或封闭的染色质可及性相关。

[0032] 在一些示例中,转录因子LYL1和PU.1的封闭的可及性指示结肠癌。

[0033] 在一个示例中,转录因子tbx21或EOMES的开放可及性指示耗竭性$CD8^+$T细胞。

[0034] 在一个示例中,选自脱中胚蛋白(EOMES)、Ybx21、Gata3、Rora、Bcl6、Blimp-1、von Hippel-Lindau肿瘤抑制因子(VHL)、Foxo1、IRF4、BATF和NFATc1的转录因子的开放可及性指示耗竭性$CD8^+$T细胞。

[0035] 在一些示例中,该方法进一步包括以至少约70%的准确度检测受试者中疾病的存在或不存在。在一些示例中,该方法进一步包括以至少约80%的准确度检测受试者中疾病的存在或不存在。在一些示例中,该方法进一步包括以至少约90%的准确度检测受试者中疾病的存在或不存在。

[0036] 在一些示例中,该方法进一步包括以至少约70%的灵敏度检测受试者中疾病的存在。在一些示例中,该方法进一步包括以至少约80%的灵敏度检测受试者中疾病的存在。在一些示例中,该方法进一步包括以至少约90%的灵敏度检测受试者中疾病的存在。

[0037] 在一些示例中,该方法进一步包括以至少约70%的特异性检测受试者中疾病的不存在。在一些示例中,该方法进一步包括以至少约80%的特异性检测受试者中疾病的不存在。在一些示例中,该方法进一步包括以至少约90%的特异性检测受试者中疾病的不存在。

[0038] 在一些示例中,该方法进一步包括以至少约70%的阳性预测值(PPV)检测受试者中疾病的存在。在一些示例中,该方法进一步包括以至少约80%的阳性预测值(PPV)检测受试者中疾病的存在。在一些示例中,该方法进一步包括以至少约90%的阳性预测值(PPV)检测受试者中疾病的存在。

[0039] 在一些示例中,该方法进一步包括以至少约70%的阴性预测值(NPV)检测受试者中疾病的不存在。在一些示例中,该方法进一步包括以至少约80%的阴性预测值(NPV)检测受试者中疾病的不存在。在一些示例中,该方法进一步包括以至少约90%的阴性预测值(NPV)检测受试者中疾病的不存在。

[0040] 在一些示例中,该方法进一步包括以至少约0.70的接受者操作特征下面积(AUROC)检测受试者中疾病的存在或不存在。在一些示例中,该方法进一步包括以至少约0.80的接受者操作特征下面积(AUROC)检测受试者中疾病的存在或不存在。在一些示例中,该方法进一步包括以至少约0.90的接受者操作特征下面积(AUROC)检测受试者中疾病的存

在或不存在。

[0041] 在一些示例中,该方法进一步包括将训练过的分类器应用于信号以检测受试者中疾病的存在或不存在。在一些示例中,该方法进一步包括将训练过的分类器应用于转录因子的结合位点的可及性得分以检测受试者中疾病的存在或不存在。在一些示例中,训练过的分类器包括训练过的机器学习分类器。在一些示例中,训练过的机器学习分类器包含监督机器学习算法。在一些示例中,监督机器学习算法包括以下中的一项或多项:回归、支持向量机、基于树的方法、神经网络和随机森林。

[0042] 在另一方面,本公开内容提供了允许根据肿瘤类型对患者进行分类的方法,肿瘤类型包括例如肿瘤亚型(例如,前列腺癌、结直肠癌、乳腺癌、肺癌的亚型)或肿瘤分期,这可对患者管理(包括治疗计划和反应性)具有重要的临床意义。因此,本文提供用于基于患者样品(例如血液、血浆或血清样品)对体内肿瘤特异性转录因子结合进行映射的方法,从而使非编码基因组的关键部分适合于临床分析。

[0043] 在一些示例中,该方法包括区分疾病的亚型。

[0044] 在一些示例中,该方法包括区分癌症的亚型。

[0045] 在一些示例中,该方法包括区分前列腺癌、结直肠癌、乳腺癌和肺癌的亚型。

[0046] 在一些示例中,该方法包括例如在患有前列腺腺癌或小细胞神经内分泌前列腺癌的患者之中区分前列腺癌亚型。

[0047] 在一些示例中,该方法包括区分癌症的分期(例如,在I期、II期、III期和IV期癌症之中)。

[0048] 在一些示例中,该方法包括将I期和II期癌症与III期和IV期癌症区分开。

[0049] 在一些示例中,转录因子GRHL2、FOXA1、HOXB13、AR和NKX3-1与前列腺腺癌患者中增加的可及性得分或开放染色质可及性相关。

[0050] 在一些示例中,转录因子GRHL2、FOXA1、HOXB13、AR和NKX3-1的开放可及性指示前列腺腺癌。

[0051] 在一些示例中,转录因子REST、GRHL2、FOXA1、HOXB13、AR和NKX3-1与小细胞神经内分泌前列腺癌患者中降低的或封闭的染色质可及性相关。

[0052] 在一些示例中,转录因子REST、GRHL2、GRHL3、FOXA1、FOXA2、GATA2、GATA3、HOXB13、AR和NKX3-1的降低的可及性指示小细胞神经内分泌前列腺癌。

[0053] 在一些示例中,转录因子GLIS1、SOX2和SOX11的增加的可及性指示小细胞神经内分泌前列腺癌。

[0054] 在另一方面,本公开内容提供一种系统,该系统包括:计算设备,其包括至少一个计算机处理器;操作系统,其被配置为执行可执行指令;存储器;以及计算机程序,其包含可由计算设备执行以提供用于检测受试者中疾病的存在或不存在的计算机应用的指令,该计算机应用包括:序列模块,其被编程为获取来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;覆盖模块,其被编程为使用所述序列读取集来产生转录因子的覆盖模式;信号模块,其被编程为处理覆盖模式以提供信号;以及检测模块,其被编程为用参考信号处理信号,其中信号和参考信号具有不同的频率,从而检测受试者中疾病的存在或不存在。

[0055] 在另一方面,本公开内容提供了一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于检测受试者中疾

病的存在或不存在的方法,该方法包括:(a)提供来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;(b)使用所述序列读取集来产生转录因子的覆盖模式;(c)处理覆盖模式以提供信号;和(d)用参考信号处理信号,其中信号和参考信号具有不同的频率,从而检测所述受试者中所述疾病的所述存在或不存在。

[0056] 另一方面,本公开内容提供了一种用于检测受试者中疾病的存在或不存在的系统,该系统包括:数据库,其包含来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;和一个或多个计算机处理器,其可操作地与所述数据库偶联,其中,一个或多个计算机处理器被单独地或共同地编程为:(a)使用序列读取集来产生转录因子的覆盖模式;(b)处理覆盖模式以提供信号;(d)用参考信号处理信号,其中信号和参考信号具有不同的频率,从而检测受试者中疾病的存在或不存在。

[0057] 在另一方面,本公开内容提供了一种用于监测受试者中疾病的进展或消退的计算机实现的方法,该方法包括:(a)提供来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;(b)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;(c)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(d)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(e)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(f)基于用参考信号对第一信号和第二信号的处理,监测受试者中疾病的进展或消退。

[0058] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分疾病进展者和非进展者、多种疾病亚型、多个疾病分期或其任何组合的分类器。

[0059] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分疾病治疗反应者和非反应者的分类器。

[0060] 在一些示例中,第二覆盖模式指示在疾病过程中肿瘤的表型变化。

[0061] 在一些示例中,表型变化是癌症从雄激素依赖性到雄激素非依赖性分期的变化。

[0062] 在一些示例中,DNA是无细胞DNA。

[0063] 在一些示例中,疾病是癌症。

[0064] 在一些示例中,(b)包括将第一序列读取集和第二序列读取集与参考序列比对,以分别提供第一经比对的序列模式和第二经比对的序列模式,选择第一经比对的序列模式和第二经比对的序列模式的与转录因子的结合位点相对应的区域,以及在该区域中对第一经比对的序列模式和第二经比对的序列模式进行标准化。

[0065] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-Golay滤波器。

[0066] 在一些示例中,受试者是人。

[0067] 在一些示例中,转录因子是癌症特异性转录因子。

[0068] 在一些示例中,转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B和FOXA1。

[0069] 在一些示例中,转录因子选自HNF-1a、HNF-4a、HNF-4g、EVX-2和DLX-2。

[0070] 在一些示例中,该方法进一步包括基于(f)调整用于受试者的疾病的治疗方案。

[0071] 在另一方面,本公开内容提供一种系统,该系统包括:计算设备,其包括至少一个计算机处理器;操作系统,其被配置为执行可执行指令;存储器;以及计算机程序,其包含可由计算设备执行以提供用于监测受试者中疾病的进展或消退的计算机应用的指令,该计算机应用包括:序列模块,其被编程为获取来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;覆盖模块,其被编程为使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;信号模块,其被编程为处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;第一处理模块,其被编程为用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;第二处理模块,其被编程为用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;以及检测模块,其被编程为基于用参考信号对第一信号和第二信号的处理,监测受试者中疾病的进展或消退。

[0072] 在另一方面,本公开内容提供了一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于监测受试者中疾病的进展或消退的方法,该方法包括:(a)提供来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;(b)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;(c)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(d)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(e)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(f)基于用参考信号对第一信号和第二信号的处理,监测受试者中疾病的进展或消退。

[0073] 另一方面,本公开内容提供了一种用于监测受试者中疾病的进展或消退的系统,该系统包括:数据库,其包含来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;和一个或多个计算机处理器,其可操作地与所述数据库偶联,其中,一个或多个计算机处理器被单独地或共同地编程为:(a)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;(b)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(c)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(d)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(e)基于用参考信号对第一信号和第二信号的处理,监测受试者中疾病的进展或消退。

[0074] 在另一方面,本公开内容提供了一种用于确定来自受试者的核酸样品中的转录因子结合谱的系统,该系统包括处理器,该处理器被配置为:(a)分析来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集;(b)使用序列读取集来产生转录因子的覆盖模式;(c)处理覆盖模式以提供信号;和(d)用参考信号处理信号,其中信号和参考信号具有不同的频率,从而确定转录因子结合谱。

[0075] 在一些示例中,DNA是无细胞DNA。

[0076] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-

Golay滤波器。

[0077] 在一些示例中,受试者是人。

[0078] 在另一方面,本公开内容提供了一种用于检测受试者中疾病的存在或不存在的系统,其包括处理器,该处理器被配置为:(i)使用来自从受试者中提取的脱氧核糖核酸(DNA)的序列读取集来产生转录因子的覆盖模式;(ii)处理覆盖模式以提供信号,其中该信号具有与参考信号不同的频率;(iii)用参考信号处理该信号,从而检测受试者中疾病的存在或不存在。

[0079] 在一些示例中,本公开内容提供了一种用于根据肿瘤亚型或肿瘤分期对肿瘤进行分类的系统,其包括处理器,该处理器被配置为:(i)使用来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和在晚于第一时间的第二时间从来自受试者的DNA提取的第二序列读取集产生转录因子的第一覆盖模式和转录因子的第二覆盖模式;(ii)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号,其中第一信号和第二信号与参考信号具有不同的频率;和(iii)用参考信号处理第一信号,并用参考信号处理第二信号,以监测受试者中疾病的进展或消退。

[0080] 在另一方面,本公开内容提供了一种用于监测受试者中疾病的进展或消退的系统,其包括处理器,该处理器被配置为:(i)使用来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和在晚于第一时间的第二时间从来自受试者的DNA提取的第二序列读取集产生转录因子的第一覆盖模式和转录因子的第二覆盖模式;(ii)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号,其中第一信号和第二信号与参考信号具有不同的频率;和(iii)用参考信号处理第一信号,并用参考信号处理第二信号,以监测受试者中疾病的进展或消退。

[0081] 本公开内容的另一方面提供了一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现以上或本文其他地方的任何方法。

[0082] 本公开内容的另一方面提供了一种系统,该系统包括一个或多个计算机处理器和与其偶联的计算机存储器。该计算机存储器包含机器可执行代码,该机器可执行代码在由一个或多个计算机处理器执行时实现以上或本文其他地方的任何方法。

[0083] 在另一方面,本公开内容提供了一种用于确定肿瘤特异性TFBS模式的方法,该方法包括:(a)提供来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;(b)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;(c)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(d)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(e)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(f)基于用参考信号对第一信号和第二信号的处理,确定肿瘤特异性TFBS模式。

[0084] 在另一方面,本公开内容提供一种系统,该系统包括:计算设备,其包括至少一个计算机处理器;操作系统,其被配置为执行可执行指令;存储器;以及计算机程序,其包含可由计算设备执行以提供用于确定肿瘤特异性TFBS模式的计算机应用的指令,该计算机应用包括:序列模块,其被编程为获取来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的

第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;覆盖模块,其被编程为使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;信号模块,其被编程为处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;第一处理模块,其被编程为用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;第二处理模块,其被编程为用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;以及检测模块,其被编程为基于用参考信号对第一信号和第二信号的处理,确定肿瘤特异性TFBS模式。

[0085] 在另一方面,本公开内容提供了一种包含机器可执行代码的非暂时性计算机可读介质,所述机器可执行代码在由一个或多个计算机处理器执行时实现用于确定肿瘤特异性TFBS模式的方法,该方法包括:(a)提供来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;(b)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式序列读取;(c)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(d)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(e)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(f)基于用参考信号对第一信号和第二信号的处理,确定肿瘤特异性TFBS模式。

[0086] 另一方面,本公开内容提供了一种用于监测受试者中疾病的进展或消退的系统,该系统包括:数据库,其包含来自在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一序列读取集和来自在晚于第一时间的第二时间从受试者中提取的DNA的第二序列读取集;和一个或多个计算机处理器,其可操作地与所述数据库偶联,其中一个或多个计算机处理器被单独地或共同地编程为:(a)使用第一序列读取集产生转录因子的第一覆盖模式,并使用第二序列读取集产生转录因子的第二覆盖模式;(b)处理第一覆盖模式以提供第一信号,并处理第二覆盖模式以提供第二信号;(c)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(d)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;和(e)基于用参考信号对第一信号和第二信号的处理,确定肿瘤特异性TFBS模式。

[0087] 本公开内容的另一方面提供了一种包含机器可执行代码的非暂时性计算机可读介质,该机器可执行代码在由一个或多个计算机处理器执行时实现以上或本文其他地方的任何方法。

[0088] 本公开内容的另一方面提供了一种系统,该系统包括一个或多个计算机处理器和与其偶联的计算机存储器。该计算机存储器包含机器可执行代码,该机器可执行代码在由一个或多个计算机处理器执行时实现以上或本文其他地方的任何方法。

[0089] 通过以下在其中仅示出和描述了本公开内容的说明性示例的详细描述,本公开内容的其他方面和优点将会对本领域技术人员而言变得显而易见。如将认识到的,本公开内容能够具有其他的和不同的示例,并且其多个细节能够在各个明显的方面进行修改,所有这些均不脱离本公开内容。因此,附图和说明书在本质上将被视为是说明性而非限制性的。

[0090] 援引并入

[0091] 本说明书中提到的所有出版物、专利和专利申请均通过引用并入本文,其程度如同明确且单独地指出每个单独的出版物、专利或专利申请通过引用并入。在通过引用并入的出版物和专利或专利申请与说明书中包含的公开内容相抵触的程度上,本说明书旨在取

代和/或优先于任何此类相抵触的材料。

附图说明

[0092] 本发明的特征在所附的权利要求书中特别地阐述。通过参考以下对利用本方法和系统的原理的说明性示例加以阐述的详细描述以及附图,将获得对本方法和系统的特征和优点的更好的理解,在附图(本说明书中也称为“图”)中:

[0093] 图1示出了被编程或以其他方式配置为执行本公开内容的方法的计算机系统,诸如存储、处理、鉴定或解释受试者(例如,患者)数据、生物数据、生物序列、参考序列、转录因子(TF)结合位点(TFBS)数据或TFBS特征诸如z得分或TFBS可及性得分。

[0094] 图2A-图2E示出了由无细胞脱氧核糖核酸(cfDNA)建立的TF-核小体相互作用。图2A示出了具有高度组织化的(例如,定相的)核小体的区域导致振荡的读取深度模式,其中读取的峰指示二分体的位置(例如,典型核小体的中点)。不太明确的核小体定位产生相当浅的覆盖谱。图2B示出了从GTRD检索到676个TF的TFBS数据,并将其与已知的或可能的人类TF的精选列表比对。进行了三种不同的计算,每种计算具有增加的严格性。图2C示出了在所有分析的cfDNA中,CCCTC结合因子(CTCF)的覆盖模式相似,这与DNase超敏性数据一致,DNase超敏性数据显示血液(GM12878)和上皮组织(例如,前列腺(LNCaP)和结肠(HCT116))中的可及性大致相等。在此分图以及各自随后的分图中,由健康对照算出的谱以灰色示出,而患者来源的谱则以指示的颜色显示。图2D示出了通过DNA超敏性测定证实了TF(PU.1、LYL1、SPIB)的造血谱系特异性,并且与健康对照相比,在癌症患者血浆中它们的幅度降低。相反,来自癌症患者的cfDNA中上皮TFGRHL2的幅度增加。图2E示出了TFFOXA1的可及性图和DNase超敏性,显示了激素依赖性癌症(例如,前列腺癌和乳腺癌)患者的优先幅度变化。

[0095] 图3A-图3F示出了TFBS表征的可及性得分。图3A示出了如何确定TF可及性。为了测量TF可及性,使用不同的窗口大小,通过Savitzky-Golay滤波将观察到的原始覆盖信号(左分图的紫色和右分图的黑色)拆分为低频信号(红色)和高频信号(蓝色)。右分图显示了这三个信号的叠加。高频信号用作可及性的量度。图3B示出了高频信号(Y轴)的范围主要取决于TFBS的数量(X轴),因为具有较少结合位点的TF由于较小的平均而具有更多的噪声。拟合了LOESS模型(蓝色)以纠正此偏差。图3C示出了GRHL2的小波分析:沿着GRHL2的TFBS周围的区域的周期的热图(左分图)。颜色代码表示信号功率分布的分位数。转录因子GRHL2的周期的平均功率(右分图)。图3D示出了由小波分析得出的转录因子GRHL2的去趋势的原始(黑色)和重建的(红色)核小体覆盖谱。图3E示出了所有测试过程(左:>50%-TFBS、Savitzky-Golay滤波;中:功率总和、小波分析;右:1,000-msTFBS、Savitzky-Golay滤波)作为在血液中表达的转录因子的可及性的量度均显示出增加的值(大于10FPKM),而在没有表达或表达迹象低的基因中则没有(<0.1FPKM)。图3F示出了来自ENCODE项目的GM12878 DNase数据中平均DNase超敏性覆盖大于2的转录因子比在进行的所有三项分析中平均覆盖<1的因子具有更高的调整范围和更高的功率总和(左:>50%-TFBS、Savitzky-Golay滤波;中:功率总和、小波分析;右:1,000-msTFBS、Savitzky-Golay滤波)。

[0096] 图4A-图4F示出了前列腺谱系特异性TF、其可塑性和对肿瘤分类的适用性。图4A示出了前列腺腺癌是AR依赖性的,并因此具有经常增加的PSA(前列腺特异性抗原)水平和正常的NSE(神经元特异性烯醇酶)值。相反,t-SCNC不再依赖于AR,并通常具有较低的PSA和增

加的NSE水平。鉴定了从腺癌到t-SCNC的转分化过程中涉及的多种TF,并以箭头指示。图4B示出了前列腺谱系特异性同源盒TF HOXB13的可及性谱和前列腺癌细胞系LNCaP的各自的DNase超敏性测定。在此分图以及随后的分图中,由健康对照算出的谱以灰色示出,而患者来源的谱则以指示的颜色显示。图4C示出了NKX3-1的可及性模式和DNA超敏性测定,NKX3-1是前列腺上皮成熟过程中表达的最早基因之一。图4D示出了GTRD中所有AR结合位点和此外肿瘤中具有较高结合强度的AR结合位点(T-ARBS)和正常样品中具有高结合强度的位点(正常AR结合位点,N-ARBS)的AR可及性(Pomerantz等人,2015)。DNA超敏性测定证实了AR的良好建立的谱系特异性。图4E示出了从腺癌向神经内分泌癌的转分化过程中的覆盖模式变化,其由来自患者P148的两个血浆样品建立,针对激素依赖性(AR、FOXA1)、组织同一性特异性(HOXB13、NKX3-1)和神经内分泌重编程(REST、N-MYC)TF。图4F示出了来自神经内分泌前列腺癌患者的4个血浆样品的与图4A相同的TF的分析。

[0097] 图5示出了来自癌症患者的血浆样品中的体细胞拷贝数改变(SCNA)。在对来自四名患者(C2、P40、P147和P148)的8个血浆样品进行全基因组测序后鉴定了SCNA。

[0098] 图6A-图6E示出了具有可靠结合位点信息的676个高可信度TF的TF-核小体相互作用图。图6A示出了由来自健康对照的24个cfDNA样品建立的两个代表性TF(CREM和GATAD1)的TFBS-核小体覆盖谱,各自用单独的蓝线示出。从ENCODE获得的类淋巴母细胞细胞系GM12878的MNase-seq覆盖模式以红色显示。图17中显示了另外的MNase图。图6B示出了CTCF结合位点周围片段大小的热图,展示为每个测序读取的长度(Y轴)作为每个注释特征从片段中点到该位点中心的距离(X轴)的函数的图。图6C示出了单个CTCF结合位点和周围区域的热图。通过围绕TFBS的中心50个碱基对(bp)内的覆盖来对区域进行排序。计算了以TFBS为中心的1千碱基(kb)区域内cfDNA片段的空间密度并进行了分等级。图6D示出了TFBS之间的重叠矩阵(左分图:所有676个GRTD TF;右分图:具有1,000-msTFBS的505个TF)。每个点代表结合位点定义中的重叠百分比(约50bp以内)。图6E示出了用高分子量DNA(不是单核小体DNA)进行的TFBS分析,与血浆DNA(绿色)相比产生了均匀、无振荡的模式(蓝色)。

[0099] 图7示出了各种TF的TF-核小体相互作用图。示出了转录因子结合位点周围的cfDNA和MNase-seq的覆盖谱之间的其他比较。

[0100] 图8A-图8E示出了TFBS的形状。图8A示出了在分别针对TSS以内和以外的TFBS进行计算之后的TF AP-4和BCL-3的覆盖谱。图8B示出了用于TF ATF1、CREB、CREM和ATF-3的TFBS的分析可导致均匀间隔或TSS样覆盖模式,取决于是否包括GTRD中的所有组织,或是否更严格地仅包含那些由多于50%的最大数量的样品(>50%-TFBS)支持的峰。图8C示出了针对全部和>50%-TFBS(上分图)以及1,000-msTFBS(下分图)计算的TF-核小体谱的示例,显示了cfDNA中不同TF的可变核小体模式。图8D示出了对TFBS宽度的测量揭示了各种TFBS之间的巨大差异。图8E示出了显示CpG岛(左分图)和TSS(右分图)的重叠百分比的箱形图。

[0101] 图9示出了合并的浅覆盖cfDNA的分析。示出了上皮TF GRHL2和造血TF(PU.1、LYL1和SPIB)的来自前列腺(n=69)、结肠(n=100)和乳腺(n=60)癌病例的合并的cfDNA样品的可及性。也示出了谱系特异性TF AR、HOXB13和NKX3-1的前列腺癌cfDNA库内的可及性。

[0102] 图10A-图10B示出了从腺癌到t-SCNC的转分化中涉及的转录因子。图10A示出了来自患者P148的血浆样品P148_1和P148_3中的GRHL2可及性。图10B示出了来自患者P148的两个血浆样品中的GLIS1的分析。

[0103] 图11示出了来自患者P148的血浆样品P148_1和P148_3的下采样。将血浆样品P148_1(819,607,690个读取)和P148_3(768,763,081个读取)下采样至大约5000万个读取,并针对1,000-msTFBS(左列)以及全部和>50％-TFBS(右列)进行了分析。分析表明,TFBS数量少的TF优先受到噪声增加的影响。

[0104] 图12示出了系列分析中TFBS可及性的比较。来自患者C2、P147、P40和P148的系列样品之间的相关性图。X轴代表第一血浆样品,Y轴代表第二血浆样品。

[0105] 图13A-图13B示出了TF-核小体相互作用的建立。图13A示出了四个TF的TFBS-核小体谱,例如SP1和SP2,它们主要结合至基因组中的共同位点,并且还与NF-YA和NF-YB共结合。图13B示出了对于造血细胞谱系特异性TF PU.1、LYL1和SPIB以及上皮细胞特异性TF GRHL2所示出的以从血浆DNA建立的平均核小体占据谱描绘的TF-核小体相互作用。不同的幅度可以反映从造血细胞和上皮细胞释放的DNA对循环的不同贡献。

[0106] 图14A-图14B展示了CTCF是表征不同TFBS的出色示例,并展示了可及性得分是TFBS的表征。图14A显示了CTCF与TAD或TSS关联的各种结合位点。CTCF的覆盖模式分为与TAD边界重叠(红色)或在TAD边界之外(橙色)的CTCF位点、在TSS附近(例如,约2kbp以内;绿色)或远离TSS(大于2kbp;蓝色)的CTCF位点以及对于完整GTRD数据集超保守的CTCF位点(黑色)(左分图)以及仅那些由多于50％的最大数量的所分析的样品支持的峰(右分图)。图14B示出了TF-核小体谱,显示它们的模式的可变性。

[0107] 图15示出了致癌基因c-Jun(上分图)和JunD(下分图)仅在CRC患者C2中显示出增加的可及性,并且通过DNA超敏性测定证实了相对的结肠特异性。

[0108] 图16A-图16C示出了在前列腺癌的转分化过程中TF的的变化的可及性。特别地,图16A-图16C示出了样品P148中在神经内分泌分化后的覆盖模式改变。图16A示出了核小体定相在雄激素受体结合位点中在由GTRD定义的位点和肿瘤特异性AR结合位点(由Pomerantz等人定义)中显著变化。图16B示出了在AR轴的其他转录因子中核小体定相也急剧减少。定相在样品1中是显著的,但是在样品3中大部分消失。图16C示出了在样品3中在神经发生中起作用的抑制因子(ZNF644、REST)大部分失活。

[0109] 图17示出了具有可靠结合位点信息的676个高可信度TF的TF-核小体相互作用图。TF-核小体谱根据它们的可及性得分和TFBS的数量进行分类。

[0110] 图18A-图18B示出了具有串联复制子表型或染色体碎裂的前列腺癌病例的分析,包括四个具有串联复制子表型的cfDNA样品(P21_2、P111_1、P111_4和P166_1)和一个在10号染色体上具有染色体碎裂的病例(P143_3)。上皮TF FOXA1、GRHL1和GRHL2以及前列腺谱系特异性TF AR、HOXB13和NKX3-1的可及性。

[0111] 图19示出了来自P40的血浆样品的AR结合位点的分析。该前列腺癌患者接受了ADT治疗,并在样品P40_1和P40_2之间产生了高水平的AR扩增。

[0112] 图20示出了展示表观遗传控制区如何影响核小体定位的图。例示了组蛋白修饰和增强子。

[0113] 图21示出了所选TF的核小体定位。

[0114] 图22示出了所选TF的覆盖模式。

[0115] 图23示出了不同TF的重叠。这些重叠值对应于图2C的热图。

[0116] 图24示出了TFBS大小的影响。

[0117] 图25示出了来自24名健康个体的样品的REST和KLF16的核小体模式。每条线代表一个不同的个体。在24名健康个体中,在血细胞中活跃的转录因子的相同设置下,这些模式似乎几乎相同。

[0118] 图26A-图26C示出了晚期癌症样品的所选TF的核小体定位。所有样品中的CTCF模式都看上去相似。在癌症样品中,包括PU.1、Ly1-1和Spi-B在内的血液特异性TF的活性降低。包括GRH-L2(上皮标记物)、ASH-2和HOX-B13(雄激素受体轴的前列腺癌标记物)和EVX2(结肠癌标记物)在内的癌症特异性TF更具活性。

[0119] 图27包括可以在本文提供的方法和系统中使用的TF的列表。

[0120] 图28A-图28B包括可以在本文提供的方法和系统中使用的TF。图28A包括结合位点大于300bp的TF,而图28B包括结合位点接近双核小体大小(在312-352bp之间)的TF。

[0121] 图29A-图29E提供了来自癌症患者的血浆样品中可及性改变的转录因子的鉴定。图29A提供了来自健康供体(NPH001)的血浆样品的TFBS分析。每个点代表一个TF,y轴显示可及性值,x轴显示作为与正常对照样品的可及性偏差的量度的总体z得分。在来自健康供体的样品中(与其余所有健康供体相比),只有少数TF超过了±3的z得分(灰色虚线),并且没有TF超过±5的z得分(红线)阈值。与图29A一样,图29B提供了总体z得分图,但血浆样品源自前列腺癌患者(P40)。与图29A一样,图29C提供了总体z得分图,用于血浆样品C2_6。图29D提供了针对TF EVX2,来自健康对照的血浆DNA(灰色谱)和源自结肠癌患者C2的两个血浆样品(蓝色和红色)的核小体位置谱。图29E提供了合并的乳腺癌、前列腺癌和结肠癌库的总体z得分图的条形图。左分图显示在至少一个肿瘤实体中可及性增加的TF;右分图总结了造血相关的TF的可及性。

[0122] 图30A-图30B提供了示出基于TF的血浆分辨率极限和早期癌症检测的图。图30A提供了示出COAD群组子样品中基于其肿瘤部分所选TF的可及性比较的图。图30B提供了分别示出来自I期(左分图)和II期(右分图)结肠癌群组的样品的具有全部504个TF的逻辑回归的图。所有给出的结果均为交叉验证的测试设定值。

具体实施方式

[0123] 尽管本文中已经示出并描述了本发明的各种实施方案,但对于本领域技术人员将显而易见的是,这些实施方案仅以示例的方式提供。本领域技术人员在不脱离本发明的情况下可想到多种变化、改变和替代。应当理解,可采用本文所述的本发明实施方式实施方案的各种可替选方案。

[0124] 在将值描述为范围的情况下,将理解的是,这种公开包括在这些范围内的所有可能的子范围,以及落在这些范围内的特定数值的公开,而不管是否明确说明了特定数值或特定子范围。

[0125] 如本文所用,术语“可及性得分”通常是指每个转录因子(TF)结合位点的可及性的量度。由于转录因子结合可以打开或“准备(prime)”其靶增强子,而不必激活它们本身,因此分等级值称为“可及性得分”。可及性得分可用于客观比较来自同一个人或不同个体当中的系列分析中TFBS的可及性。该得分提供了对TFBS可及性的可靠评估,尤其适用于在临床诊断、癌症检测和治疗监测中使用cfDNA。

[0126] 如本文所用,术语“经比对序列模式”通常是指与参考基因组比对后的序列读取的

空间模式。

[0127] 如本文所用,术语“循环游离DNA”或“无细胞DNA”(cfDNA)通常是指脱氧核糖核酸(DNA),其于1948年在人类血浆中首次被检测到。(Mandel,P.Metais,P.,C R Acad.Sci.Paris,142,241-243(1948))。从那时起,它与疾病的联系已在若干领域确立。(Tong,Y.K.Lo,Y.M.,Clin Chim Acta,363,187-196(2006))。研究显示,血液中许多循环核酸来自坏死或凋亡细胞(Giacona,M.B.等人,Pancreas,17,89-97(1998)),并且在诸如癌症的疾病中观察到来自凋亡的核酸水平大幅升高。(Giacona,M.B.等人,Pancreas,17,89-97(1998);Fournie,G.J.等人,Cancer Lett,91,221-227(1995))。特别是对于癌症,其中循环DNA携带疾病的标志性信号,包括癌基因中的突变、微卫星改变,并且,对于某些癌症,血浆中的病毒基因组序列、DNA或RNA已被越来越多地研究作为疾病的潜在生物标志物。16266-16271(2008))。

[0128] 无细胞部分可以是血清或血浆。如本文所用,生物样品的术语“无细胞部分”通常是指生物样品的基本上不含细胞的部分。如本文所用,术语“基本上不含细胞”通常是指来自生物样品的制剂,其包含少于约20,000个细胞/mL、少于约2,000个细胞/mL、少于约200个细胞/mL或少于约20个细胞/mL。基因组DNA可以不从非细胞样品中排除,并且通常占样品中存在的核酸的约50%至约90%。

[0129] 如本文所用,术语“覆盖模式”通常是指与参考基因组比对后的序列读取的空间布局。覆盖模式鉴定下一代测序方法的覆盖范围和深度。

[0130] 如本文所用,术语“源自”通常是指起源或来源,并且可以包括天然存在的、重组的、未经纯化的或经纯化的分子。源自原始核酸的核酸可以部分或全部包含原始核酸,并且可以是原始核酸的片段或变体。源自生物样品的核酸可以从该样品中纯化。

[0131] 如本文所用,状态或结果的术语“诊断(diagnose)”或“诊断(diagnosis)”通常是指预测或诊断状态或结果、确定对状态或结果的易感性、监测受试者(例如,患者)的治疗、诊断受试者(例如,患者)的治疗反应以及状态或结果、进展和对特定治疗的反应的预后。

[0132] 如本文所用,术语“核酸”通常是指包含两个或更多个核苷酸的多核苷酸。它可以是DNA或RNA。核酸可以是任何长度的核苷酸(脱氧核糖核苷酸(dNTP)或核糖核苷酸(rNTP)或其类似物)的聚合形式。核酸可以具有任何三维结构,并且可以执行任何已知或未知的功能。核酸的非限制性示例包括脱氧核糖核酸(DNA)、核糖核酸(RNA)、基因或基因片段的编码或非编码区、从连锁分析中定义的基因座(基因位点)、外显子、内含子、信使RNA(mRNA)、转移RNA、核糖体RNA、短干扰RNA(siRNA)、短发夹RNA(shRNA)、微小RNA(miRNA)、核酶、cDNA、重组核酸、分支核酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。核酸可以包含一个或多个经修饰的核苷酸,例如甲基化的核苷酸和核苷酸类似物。对核苷酸结构的修饰若存在,则可以在核酸组装之前或之后进行。核酸的核苷酸序列可以被非核苷酸组分中断。核酸可以在聚合后被进一步修饰,例如通过与报告剂缀合或结合。“变体”核酸是除了具有至少一个被修饰(例如,分别被删除、插入或替换)的核苷酸之外,具有与其原始核酸的核苷酸序列相同的核苷酸序列的多核苷酸。变体可以具有与原始核酸的核苷酸序列至少约80%、90%、95%或99%同一性的核苷酸序列。

[0133] 如本文所用,术语“靶核酸”通常是指核酸分子起始群体中的核酸分子,该核酸分子具有核苷酸序列,该核苷酸序列的存在、量和/或序列或这些的一种或多种中的变化是希

望被确定的。靶核酸可以是任何类型的核酸,包括DNA、RNA及其类似物。如本文所用,“靶核糖核酸(RNA)”通常是指作为RNA的靶核酸。如本文所用,“靶脱氧核糖核酸(DNA)”通常是指作为DNA的靶核酸。

[0134] 如本文所用,术语“扩增”通常是指增加核酸分子的大小或量。核酸分子可以是单链或双链的。扩增可包括产生核酸分子的一个或多个拷贝或“扩增产物”。扩增可以例如通过延伸(例如,引物延伸)或连接来进行。扩增可包括进行引物延伸反应以产生与单链核酸分子互补的链,并且在一些情况下产生链和/或单链核酸分子的一个或多个拷贝。术语“DNA扩增”通常是指产生DNA分子的一个或多拷贝或“扩增DNA产物”。术语“逆转录扩增”通常是指通过逆转录酶的作用从核糖核酸(RNA)模板产生脱氧核糖核酸(DNA)。

[0135] 术语“转录因子”通常是指通过与特定DNA序列结合来控制遗传信息从DNA到信使RNA的转录速率的蛋白质。转录因子是与DNA调控序列(例如,增强子和沉默子)结合的蛋白质,通常位于靶基因的5’上游区域,以调节基因转录的速率。这可导致基因转录、蛋白质合成增加或减少,以及随后的细胞功能改变(例如,细胞响应环境(正常或病理)而变化,例如在萎缩、肥大、增生、化生或发育异常期间)。如本文所用,特定转录因子由命名法指称,尽管其他同义词也可以用于本文所记载的转录因子。

[0136] 术语“转录因子结合谱”通常是指给定转录因子的多因子信息谱,其包括组织贡献和生物过程。TFBP还包括“可及性得分”和z得分统计量,以客观比较不同血浆样品中TFBS可及性的显著变化。该谱可以允许鉴定适用于起源组织和起源肿瘤鉴定两者的谱系特异性TF。

[0137] 如本文所用,术语“受试者”通常是指具有或疑似具有可测试或可检测的遗传信息或材料的个体、实体或介质。受试者可以是人、个体或患者。受试者可以是脊椎动物,例如举例而言,哺乳动物。哺乳动物的非限制性示例包括人类、猿猴、农场动物、运动动物、啮齿动物和宠物。受试者可显示指示受试者的健康或生理状态或状况的症状,例如受试者的癌症或癌症的分期。作为可替选,受试者在这种健康或生理状态或状况方面可以是无症状的。

[0138] 如本文所用,术语“样品”通常是指获得自或源自一个或多个受试者的生物样品。生物样品可以是无细胞生物样品或基本上无细胞的生物样品,或者可以被处理或分级分离以产生无细胞生物样品。例如,无细胞生物样品可以包括无细胞核糖核酸(cfRNA)、无细胞脱氧核糖核酸(cfDNA)、无细胞蛋白质和/或无细胞多肽。生物样品可以是组织(例如,通过活组织检查获得的组织)、血液(例如,全血)、血浆、血清、汗液、尿液、唾液或其衍生物。可以使用乙二胺四乙酸(EDTA)收集管、无细胞RNA收集管(例如,Streck)或无细胞DNA收集管(例如,Streck)从受试者获得或衍生无细胞生物样品。无细胞生物样品可以通过分级分离从全血样品中获得。生物样品或其衍生物可含有细胞。例如,生物学样品可以是血液样品或其衍生物(例如,通过收集管收集的血液或血滴)、肿瘤样品、组织样品、尿液样品或细胞(例如,组织)样品。

[0139] 本公开内容提供了用于对转录因子(TF)结合位点(TFBS)建模并使用TFBS信息来检测、评估、诊断和分析疾病状态的方法和系统。cfDNA代表通过内源性生理过程产生的独特分析物,以通过全基因组测序产生核小体占据的体内图。转录因子结合位点(TFBS)处的核糖体占据可用于从释放其DNA到循环中的细胞推断表达的基因。cfDNA核小体占据可反映TF的足迹。

[0140] I.转录因子结合位点/核小体占据分析

[0141] 尽管下一代测序可以提供有关TF的重要信息,但仍需要一种非侵入性的方法来测量TF活性或其在治疗下的调节(例如,来自血液)。在癌症患者中还含有循环肿瘤DNA(ctDNA)的无细胞DNA(cfDNA)(例如来自血浆)可能为癌症患者的非侵入性诊断策略提供机会。由于cfDNA在酶促消化后可能会从凋亡细胞中释放,因此它可能主要以单核小体DNA的形式循环。因此,cfDNA片段的全基因组测序可以实现核小体图的生成,其中具有高核小体偏好的位点的二分体(例如,典型核小体的中点)导致读取的强峰,而位置较不优选的核小体的二分体示出降低的峰或根本没有峰。

[0142] 由于从cfDNA推断TF结合在癌症及其他方面具有巨大的诊断潜力,因此开发了改进和优化的生物信息学渠道。此过程能够解析那些涉及TFBS处核小体特征的成分,从而客观地评估和比较不同血浆样品中的TFBS可及性。可以从来自健康供体的血浆样品和从癌症(例如,转移性前列腺癌、结肠癌或乳腺癌)患者的血浆样品获得深度全基因组测序(WGS)数据。在一些示例中,cfDNA还包括循环肿瘤DNA(ctDNA)。此外,还可以从来自具有上述肿瘤实体的患者的血浆样品获得浅WGS数据。此方法可用于剖析单个TF,而不是如同根据其他方法使用由多种细胞类型产生的cfDNA信号的混合物并通过傅立叶变换进行分析来建立一般组织特异性模式。本文提供的方法和系统还有益地提供了组织贡献和生物过程两者的更细致入微的视角,其允许鉴定适用于起源组织和起源肿瘤分析的谱系特异性TF。

[0143] 某些谱系特异性TF可适用于确定血浆DNA的起源组织。但是,确定哪些TF可以用于在此类应用中,需要评价TF的可及性(例如在它们在cfDNA中的结合位点处)。常规方法可能缺乏评价在它们在cfDNA中的结合位点处的TF可及性作为它们活性的代表的能力。针对转录起始位点(TSS)以内和以外的TFBS分别进行计算。平均TFBS模式包含两个信号:TSS近端(在TSS的约2kb以内,导致“低频模式”)和TSS远端(距离TSS峰大于2kb,导致“高频模式”),对应于间隔更均匀的峰值信号。为了抑制并非由优先核小体定位造成的对覆盖范围的影响,并从核小体数据中消除局部偏差,可以使用滤波器进行去趋势化(例如,Savitzky-Golay滤波器)。然后可以将获得的低频信号用于标准化高频信号,并且随后可以记录高频信号的数据范围(数据值的最大值减去最小值,对应于幅度)。由于高频信号的范围取决于TFBS的数目(1,000-msTFBS除外),这些范围值可以通过平滑进行校正(因为它们取决于TFBS的数目),然后用于计算等级以作为每个TFBS的可及性的量度。

[0144] 为此分析开发的指标,称为“可及性得分”,可用于客观比较从同一个人或不同个体当中获得的样品的系列分析中TFBS的可及性。当TF结合打开或“准备”其靶增强子时,不必激活它们本身,等级值可以称为“可及性得分”。这些结果展示了评估TFBS可及性的可靠方法,尤其适用于在临床诊断中使用cfDNA。

[0145] 与可能利用使用由多种细胞类型产生的cfDNA信号的混合物并通过傅立叶变换进行分析的一般组织特异性模式的其他分析相对照,本公开内容的方法和系统可以剖析单个TF并由此建立了用于临床应用的谱系特异性TF。由于提高了TFBS分析的分辨率,从cfDNA监控TFBS的可及性可以是可能的,并且在一些示例中被证明了可用于揭示病程中的TF可塑性,例如,重编程为不同的细胞谱系。

[0146] 图20示出了展示表观遗传控制区如何影响核小体定位的图。例示了组蛋白修饰和增强子。

[0147] 图21示出了所选TF的核小体定位。

[0148] 图22示出了所选TF的覆盖模式。

[0149] 图23示出了不同TF的重叠。这些重叠值对应于图2C的热图。

[0150] 图24示出了TFBS大小的影响。

[0151] 图25示出了来自24名健康个体的样品的REST和KLF16的核小体模式。每条线代表一个不同的个体。在24名健康个体中,在血细胞中活跃的转录因子的相同设置下,这些模式似乎几乎相同。

[0152] 图26A-图26C示出了晚期癌症样品的所选TF的核小体定位。所有样品中的CTCF模式看上去相似。在癌症样品中,包括PU.1、Lyl-1和Spi-B在内的血液特异性TF的活性降低。包括GRH-L2(上皮标记物)、ASH-2和HOX-B13(雄激素受体轴的前列腺癌标记物)和EVX2(结肠癌标记物)在内的癌症特异性TF更具活性。

[0153] II.转录因子结合位点

[0154] 转录因子结合位点是使用本发明的方法和系统中使用的统计阈值从基因转录调控数据库(GTRD:a database on gene transcription regulation-2019年更新.I.S.Yevshin,R.N.Sharipov.S.K.Kolmykov,Y.V.Kondrakhin,F.A.Kolpakov.Nucleic Acids Res.2019年1月8日;47(D1):D100-D105)中鉴定的,并且对于机器学习模型和分类器生成是信息性的。在一些示例中,转录因子的相关通路和类别对于机器学习模型和分类器生成是类似有用的和信息性的。

[0155] 统计阈值用于鉴定两个或更多个患者组之间的差异TF,以进行分析(例如,健康与癌症、进展者与非进展者、多个分期之中的一个分期(例如,I、II、III或IV)、多个亚型中的一个亚型或治疗反应者与非反应者)。

[0156] 在一些示例中,可以使用本文描述的方法和系统来分析诸如在图27和图28中列出的那些转录因子。

[0157] 在一些示例中,转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B和FOXA1。

[0158] 在一些示例中,观察到患者特异性以及肿瘤特异性模式,包括转录因子AR、HOXB13和NKX3-1的推断的结合模式。

[0159] 在一些示例中,转录因子是上皮转录因子。在一些示例中,转录因子是GRHL2。

[0160] 在一些示例中,转录因子GRHL2、FOXA1和ZNF121与乳腺癌患者中增加的可及性得分或开放染色质可及性相关。

[0161] 在一些示例中,选自GRHL2、FOXA1和ZNF121的至少一种转录因子的开放可及性指示乳腺癌。

[0162] 在一些示例中,转录因子EVX2、DLX2、HNF1A、HNF4A、GRHL2和HNF4G与结肠癌患者中增加的可及性得分或开放染色质可及性相关。

[0163] 在一些示例中,选自EVX2、DLX2、HNF1A、GRHL2、HNF4A和HNF4G的至少一种转录因子的开放可及性指示结肠癌。

[0164] 在一些示例中,转录因子LYL1、EVI1、TAL1、Spi-B、TBX21和PU.1与结肠癌患者中降低的可及性得分或封闭的染色质可及性相关。

[0165] 在一些示例中,选自LYL1、EVI1、TAL1、Spi-B、TBX21和PU.1的至少一种转录因子的

封闭的可及性指示结肠癌。

[0166] 在一些示例中,转录因子GRHL2、FOXA1、HOXB13、AR和NKX3-1与前列腺腺癌患者中增加的可及性得分或开放染色质可及性相关。

[0167] 在一些示例中,选自GRHL2、FOXA1、HOXB13、AR和NKX3-1的至少一种转录因子的开放可及性指示前列腺腺癌。

[0168] 在一些示例中,转录因子REST、GRHL2、FOXA1、HOXB13、AR和NKX3-1与小细胞神经内分泌前列腺癌患者中降低的或封闭的染色质可及性相关。

[0169] 在一些示例中,选自REST、GRHL2、FOXA1、HOXB13、AR和NKX3-1的至少一种转录因子的的降低的可及性指示小细胞神经内分泌前列腺癌。

[0170] 在一个示例中,造血转录因子与组织特异性TF的可及性之间的相关性与疾病例如癌症的存在相关。

[0171] 在一个示例中,造血转录因子选自LYL1、SCL、Bcl11a、Hhex、Lmo2、Spi1和PU.1。在一个示例中,造血转录因子选自LYL1或PU.1。

[0172] 在一些示例中,造血转录因子,例如LYL1、SPIB和EVI1(转录调控子嗜亲性病毒整合位点1)的低可及性与前列腺癌相关。

[0173] 在一些示例中,转录因子选自GRH-L2、ASH-2、HOX-B13、EVX2、PU.1、Lyl-1、Spi-B和FOXA1。

[0174] 在一些示例中,转录因子选自HNF-1a、HNF-4A、HNF-4G、EVX-2和DLX-2。

[0175] 在一些示例中,与造血相关的TF,例如LYL1、TAL1(SCL/TAL1(干细胞白血病/T细胞急性淋巴母细胞白血病[T-ALL]1))、EVI1、TBX21(T-bet)和PU.1的低可及性与癌症相关。

[0176] 在慢性病毒感染或癌症中持续暴露于抗原的过程中,效应子$CD8^+$T细胞获得可替选的细胞分化命运,称为T细胞耗竭。它们不能像记忆细胞那样进行抗原非依赖性的自我更新,并且不能以分级的方式失去其效应子功能,这阻碍了这些抗原特异性$CD8^+$T细胞的病毒清除和肿瘤控制。

[0177] 在一个示例中,转录因子tbx21或EOMES的开放可及性指示耗竭性$CD8^+$T细胞。

[0178] 在一个示例中,转录因子脱中胚蛋白(EOMES)、Blimp-1、von Hippel-Lindau肿瘤抑制因子(VHL)、Foxo1、IRF4、BATF和NFATc1的开放可及性指示耗竭性$CD8^+$T细胞。

[0179] III.机器学习系统和模型

[0180] A.样品特征

[0181] 在一些示例中,TFBS可及性得分被用作机器学习模型中的输入特征,以找到序列组成与受试者(例如,患者)组之间的相关性。这样的患者组的示例包括疾病或病况的存在、分期、亚型、反应者与非反应者和进展者与非进展者。在一些示例中,生成特征矩阵以比较从具有已知病况或特性的个体获得的样品。在一些示例中,样品获得自健康个体或没有任何已知适应症的个体,以及来自已知患有癌症的患者的样品。

[0182] 如本文所使用的,当术语“特征”涉及机器学习和模式识别时,它指的是被观察的现象的单个可测量的属性或特性。特征通常是数字特征,但结构特征如字符串和图形可以被用于句法模式识别。“特征”的概念与统计技术中使用的解释变量的概念有关,例如但不限于线性回归。在一些示例中,特征是转录因子结合谱。在一些示例中,特征是从转录因子结合谱计算的可及性得分。

[0183] 在一些示例中,特征被输入到特征矩阵中以用于机器学习分析。

[0184] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分健康受试者和癌症患者、疾病进展者和非进展者的分类器。

[0185] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分多种疾病亚型或多种疾病分期的分类器。

[0186] 在一些示例中,确定至少2个或至少5个或至少10个或至少15个或至少20个或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中以训练能够区分疾病治疗反应者和非反应者的分类器。

[0187] 对于多个测定,系统鉴定要接受的特征集作为机器学习模型的输入。系统对每种分子类别进行测定,并根据测量值形成特征向量。系统将特征向量作为输入接受到机器学习模型中,并生成生物样品是否具有指定属性的输出分类。

[0188] 在一些示例中,机器学习模型生成能够区分在个体的群体或群体的特征中的个体或特征的两个或多个组或类别的分类器。例如,分类器可以是能够区分在个体的群体或群体的特征中的个体或特征的两个组或类别的二元分类器。作为另一个示例,分类器可以是能够区分在个体的群体或群体的特征中的个体或特征的多于两个的组或类别的多类别分类器。在一些示例中,分类器是训练过的机器学习分类器。

[0189] 在一些示例中,对癌症组织中的信息基因座或生物标志物的特征进行测定,以形成谱。在二元分类器的情况下,可以生成接收者操作特性(ROC)曲线,以绘制在区分两个群体(例如,对治疗剂有反应和无反应的个体)时特定特征(例如,本文所述的任何生物标志物和/或任何其他生物医学信息项目)的性能。在一些示例中,基于单个特征的值以升序对整个群体(例如,病例和对照)中的特征数据进行排序。

[0190] 在一些示例中,指定属性选自健康与癌症、多种疾病亚型之中的疾病亚型、多种疾病分期之中的疾病分期、进展者与非进展者、反应者与非反应者或其组合。

[0191] B.数据分析

[0192] 在一些示例中,本公开内容提供了在软件应用、计算硬件或两者中实现数据分析的系统、方法或套件。在一些示例中,分析应用或系统至少包括数据接收模块、数据预处理模块、数据分析模块(其可以对一种或多种类型的基因组数据进行操作)、数据解释模块或数据可视化模块。在一些示例中,数据接收模块可以包括将实验室硬件或仪器与处理实验室数据的计算机系统连接的计算机系统。在一些示例中,数据预处理模块可以包括硬件系统或计算机软件,其对数据执行操作以备用于分析。可以在预处理模块中应用于数据的操作示例包括仿射变换、去噪操作、数据清理、重新格式化或二次采样。可以被专门化以用于分析来自一种或多种基因组材料的基因组数据的数据分析模块,例如,可以获取已组装的基因组序列并进行概率和统计分析,以鉴定与疾病、病理、状态、风险、病况或表型相关的异常模式。数据解释模块可以使用例如从统计学、数学或生物学中得出的分析方法,以支持对所鉴定的异常模式与健康状况、功能状态、预后或风险之间的关系的理解。数据可视化模块可以使用数学建模、计算机图形学或渲染的方法来创建数据的视觉表示,该视觉表示可以促进(例如,用户如受试者(例如,患者)或医生或其他医疗保健提供者)对结果的理解或解

释。

[0193] 在一些示例中,应用机器学习方法来区分样品群体中的样品。在一些示例中,应用机器学习方法来区分健康和癌症(例如,晚期腺瘤)样品。

[0194] 在一些示例中,用于训练预测引擎的一个或多个机器学习操作包括以下中一项或多项:广义线性模型、广义加性模型、非参数回归运算、随机森林分类器、空间回归运算、贝叶斯回归模型、时间序列分析、贝叶斯网络、高斯网络、决策树学习运算、人工神经网络、递归神经网络、强化学习运算、线性或非线性回归运算、支持向量机、聚类运算和遗传算法运算。

[0195] 在一些示例中,计算机处理方法选自逻辑回归、多元线性回归(MLR)、降维、偏最小二乘(PLS)回归、主成分回归、自动编码器、变分自动编码器、奇异值分解、傅立叶基数(Fourier base)、小波、判别分析、支持向量机、决策树、分类和回归树(CART)、基于树的方法、随机森林、梯度提升树、逻辑回归、矩阵分解、多维缩放(MDS)、降维方法、t-分布随机邻域嵌入(t-SNE)、多层感知器(MLP)、网络聚类、神经模糊和人工神经网络。

[0196] 在一些示例中,本文公开的方法可包括对来自个体或来自多个个体的样品的核酸测序数据进行计算分析。

[0197] C.分类器生成

[0198] 在一方面,本公开内容提供了包括分类器的系统和方法,该分类器基于从来自cfDNA的生物样品的序列分析得到的特征信息而生成。分类器构成了预测引擎的一部分,该预测引擎可根据在生物样品(例如cfDNA)中鉴定出的序列特征来区分群体中的组。

[0199] 在一些示例中,通过将序列信息的相似部分格式化为统一格式和统一标度来标准化序列信息而创建分类器;将标准化的序列信息存储在列式数据库中;通过将一个或多个机器学习操作应用于存储的标准化的序列信息来训练预测引擎,该预测引擎针对特定群体映射一个或多个特征的组合;将预测引擎应用于所访问的字段信息,以鉴定与组相关的个体;并将个体分类到一个组中。

[0200] 训练过的分类器可以被配置为接受多个输入变量并基于多个输入变量产生一个或多个输出值。多个输入变量可以包括指示疾病、病症或异常病况(例如,癌症)的一个或多个数据集。例如,输入变量可以包含与一组疾病相关的基因组基因座相对应或比对上的许多核酸序列。多个输入变量还可包括受试者的临床健康数据。

[0201] 例如,临床健康数据可以包括受试者的一种或多种定量量度,例如年龄、体重、身高、体质指数(BMI)、血压、心率和葡萄糖水平。作为另一个示例,临床健康数据可以包括一种或多种分类度量,例如种族、民族、药物或其他临床治疗史、吸烟史、饮酒史、日常活动或健身水平、基因测试结果、血液测试结果和成像结果。

[0202] 本文提供的训练过的算法可以包括分类器,使得一个或多个输出值中的每一个都包括指示分类器对样品进行分类的固定数量的可能值中之一(例如,线性分类器、逻辑回归分类器等)。训练过的算法可以包括二元分类器,使得一个或多个输出值中的每一个都包括指示由分类器对样品进行分类的两个值(例如,{0、1}、{正、负}或{高风险、低风险})中的一个。训练过的算法可以是另一种类型的分类器,使得一个或多个输出值中的每一个都包括指示由分类器对样品进行分类的多于两个的值(例如,{0、1、2}、{正、负或不确定}或{高风险、中等风险或低风险})中的一个。输出值可以包括描述性标签、数值或其组合。一些输出

值可以包括描述性标签。这种描述性标签可以提供对受试者的疾病、病症或异常病况的评估的鉴定或指示,并且可以包括例如阳性、阴性、高风险、中风险、低风险或不确定。这种描述性标签可以提供对受试者的评估的疾病、病症或异常病况的治疗的鉴定,并且可以包括例如治疗干预、治疗干预的持续时间和/或适合于治疗疾病、病症或异常病况的治疗干预的剂量。这种描述性标签可以提供可能适合对受试者进行的二次临床测试的鉴定,并且可以包括例如影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。例如,这种描述性标签可以提供受试者的疾病、病症或异常病况的预后。作为另一个示例,这种描述性标签可以提供受试者的疾病、病症或异常病况的相对评估。一些描述性标签可以映射到数值,例如,通过将“正”映射为1并将“负”映射为0。

[0203] 一些输出值可以包括数值,例如二元、整数或连续值。这种二元输出值可以包括例如{0、1}、{正、负}或{高风险、低风险}。这种整数输出值可以包括例如{0、1、2}。这种连续输出值可以包括例如至少为0且不大于1的概率值。这种连续输出值可以包括例如至少为0的未标准化概率值。这种连续输出值可以指示受试者的疾病、病症或异常病况的预后。一些数值可以映射到描述性标签,例如,通过将1映射为“正”并将0映射为“负”。

[0204] 一些输出值可以基于一个或多个截断值来分配。例如,如果样品指示受试者具有至少50%患有疾病、病症或异常病况的概率,则样品的二元分类可分配“正”或1的输出值。例如,如果样品指示受试者具有小于50%患有疾病、病症或异常病况的概率,则样品的二元分类可分配“负”或0的输出值。在这种情况下,使用50%的单个截断值将样品分类为两个可能的二元输出值中之一。单个截断值的示例可包括约1%、约2%、约5%、约10%、约15%、约20%、约25%、约30%、约35%、约40%、约45%、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%和约99%。

[0205] 作为另一示例,如果样品指示受试者具有至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%或更大的患有疾病、病症或异常病况的概率,则样品的分类可分配“正”或1的输出值。如果样品指示受试者具有大于约50%、大于约55%、大于约60%、大于约65%、大于约70%、大于约75%、大于约80%、大于约85%、大于约90%、大于约91%、大于约92%、大于约93%、大于约94%、大于约95%、大于约96%、大于约97%、大于约98%或大于约99%的患有疾病、病症或异常病况的概率,则样品的分类可分配“正”或1的输出值。

[0206] 如果样品指示受试者具有小于约50%、小于约45%、小于约40%、小于约35%、小于约30%、小于约25%、小于约20%、小于约15%、小于约10%、小于约9%、小于约8%、小于约7%、小于约6%、小于约5%、小于约4%、小于约3%、小于约2%或小于约1%的患有疾病、病症或异常病况的概率,则样品的分类可分配“负”或0的输出值。如果样品指示受试者具有不大于约50%、不大于约45%、不大于约40%、不大于约35%、不大于约30%、不大于约25%、不大于约20%、不大于约15%、不大于约10%、不大于约9%、不大于约8%、不大于约7%、不大于约6%、不大于约5%、不大于约4%、不大于约3%、不大于约2%或不大于约1%的患有疾病、病症或异常病况的概率,则样品的分类可分配“负”或0的输出值。

[0207] 如果样品未被分类为“正”、“负”、“1”或“0”,则样品的分类可分配“不确定”或2的输出值。在这种情况下,将使用二个截断值的集将样品分类为三个可能的输出值之一。截断值集的示例可以包括{1%、99%}、{2%、98%}、{5%、95%}、{10%、90%}、{15%、85%}、{20%、80%}、{25%、75%}、{30%、70%}、{35%、65%}、{40%、60%}和{45%、55%}。类似地,可以使用n个截断值的集将样品分类为n+1个可能的输出值之一,其中n是任何正整数。

[0208] 可以用多个独立训练样品来训练训练过的分类器。每个独立训练样品可以包括来自受试者的样品、通过测定样品而获得的相关数据集(如本文别处所述)和与样品相对应的一个或多个已知输出值(例如,受试者的疾病、病症或异常病况的临床诊断、预后、不存在或治疗效果)。独立训练样品可以包括获得自或源自多个不同受试者的样品和相关数据集和输出。独立训练样品可以包括在多个不同时间点(例如,定期,如每周、每两周或每月)获得自同一受试者的样品和相关数据集和输出。独立训练样品可以与疾病、病症或异常病况的存在相关(例如,训练样品包括获得自或源自已知具有疾病、病症或异常病况的多个受试者的样品和相关数据集和输出)。独立训练样品可以与疾病、病症或异常病况的不存在相关(例如,训练样品包括获得自或源自已知以前没有疾病、病症或异常病况的诊断或已经收到疾病、病症或异常病况的阴性检测结果的多个受试者的样品和相关数据集和输出)。

[0209] 训练过的分类器可以用至少约5个、至少约10个、至少约15个、至少约20个、至少约25个、至少约30个、至少约35个、至少约40个、至少约45个、至少约50个、至少约100个、至少约150个、至少约200个、至少约250个、至少约300个、至少约350个、至少约400个、至少约450个或至少约500个独立训练样品进行训练。独立训练样品可以包括与疾病、病症或异常病况的存在相关的样品和/或与疾病、病症或异常病况的不存在相关的样品。训练过的分类器可以用不多于约500个、不多于约450个、不多于约400个、不多于约350个、不多于约300个、不多于约250个、不多于约200个、不多于约150个、不多于约100个或不多于约50个与疾病、病症或异常病况的存在相关的独立训练样品进行训练。在一些实施方案中,该样品独立于用于训练训练过的分类器的样品。

[0210] 训练过的分类器可以用与疾病、病症或异常病况的存在相关的第一数量的独立训练样品和与疾病、病症或异常病况的不存在相关的第二数量的独立训练样品进行训练。与疾病、病症或异常病况的存在相关的第一数量的独立训练样品可以不多于与疾病、病症或异常病况的不存在相关的第二数量的独立训练样品。与疾病、病症或异常病况的存在相关的第一数量的独立训练样品可以和与疾病、病症或异常病况的不存在相关的第二数量的独立训练样品同样多。与疾病、病症或异常病况的存在相关的第一数量的独立训练样品可以多于与疾病、病症或异常病况的不存在相关的第二数量的独立训练样品。

[0211] 训练过的分类器可被配置为以至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约81%、至少约82%、至少约83%、至少约84%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%或更高的准确度鉴定疾病、病症或异常病况的存在或不存在;对于至少约5个、至少约10个、至少约15个、至少约20个、至少约25个、至少约30个、至少约35个、至少约40个、至少约45个、至少约50个、至少约100个、至少约150个、至少约200个、至少约250个、至少约300个、至少约350个、至少约400个、至少约450个或至少约500个独立训练样本。由训练

过的算法鉴定疾病、病症或异常病况的存在或不存在的准确度可计算为正确鉴定或分类为具有或不具有疾病、病症或异常病况的独立测试样品(例如,已知患有该疾病、病症或异常病况的受试者,或该疾病、病症或异常病况的临床测试结果为阴性的受试者)的百分比。

[0212] 训练过的分类器可以被配置为以至少约5%、至少约10%、至少约15%、至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约81%、至少约82%、至少约83%、至少约84%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%或更高的阳性预测值(PPV)鉴定疾病、病症或异常病况的存在。使用训练过的分类器鉴定疾病、病症或异常病况的存在的PPV可以被计算为对应于真正患有该疾病、病症或异常病况的受试者的鉴定或分类为具有疾病、病症或异常病况的样品的百分比。

[0213] 训练过的分类器可以被配置为以至少约5%、至少约10%、至少约15%、至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约81%、至少约82%、至少约83%、至少约84%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%或更高的阴性预测值(NPV)鉴定疾病、病症或异常病况的不存在。使用训练过的分类器鉴定疾病、病症或异常病况的NPV可以被计算为对应于真正不患有该疾病、病症或异常病况的受试者的鉴定或分类为不具有疾病、病症或异常病况的样品的百分比。

[0214] 训练过的分类器可被配置为以至少约5%、至少约10%、至少约15%、至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约81%、至少约82%、至少约83%、至少约84%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%、至少约99%、至少约99.1%、至少约99.2%、至少约99.3%、至少约99.4%、至少约99.5%、至少约99.6%、至少约99.7%、至少约99.8%、至少约99.9%、至少约99.99%、至少约99.999%或更高的临床特异性鉴定疾病、病症或异常病况的不存在。如本文所用,特异性是指“在未患有该疾病的人之中测试为阴性的概率”。它等于测试为阴性的无病人数除以无病个体总数。使用训练过的分类器鉴定疾病、病症或异常病况的不存在的临床特异性可以被计算为被正确鉴定或分类为不患有该疾病、病症或异常病况的与疾病、病症或异常病况的不存在相关的独立测试样品(例如,疾病、病症或异常病况的临床测试结果为阴性的受试者)的百分比。在一些示例中,该模型、分类器或预测测试具有至少约40%、至少约45%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%或至少约99%的特异性。

[0215] 训练过的分类器可被配置为以至少约5%、至少约10%、至少约15%、至少约20%、至少约25%、至少约30%、至少约35%、至少约40%、至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约81%、至少约82%、至少约83%、

至少约84％、至少约85％、至少约86％、至少约87％、至少约88％、至少约89％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％、至少约99.1％、至少约99.2％、至少约99.3％、至少约99.4％、至少约99.5％、至少约99.6％、至少约99.7％、至少约99.8％、至少约99.9％、至少约99.99％、至少约99.999％或更高的临床灵敏度鉴定疾病、病症或异常病况的存在。如本文所用，灵敏度是指“在患有该疾病的人之中测试为阳性的概率”。它等于测试为阳性的患病个体数除以患病个体总数。

[0216] 在一些示例中，该模型、分类器或预测测试具有至少约40％、至少约45％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约85％、至少约90％、至少约95％或至少约99％的灵敏度。使用训练过的分类器鉴定疾病、病症或异常病况的存在的临床灵敏度可以计算为被正确鉴定或分类为患有该疾病、病症或异常病况的与该疾病、病症或异常病况的存在相关的独立测试样品(例如，已知患有该疾病、病症或异常病况的受试者)的百分比。

[0217] 训练过的分类器可以被配置为以至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.81、至少约0.82、至少约0.83、至少约0.84、至少约0.85、至少约0.86、至少约0.87、至少约0.88、至少约0.89、至少约0.90、至少约0.91、至少约0.92、至少约0.93、至少约0.94、至少约0.95、至少约0.96、至少约0.97、至少约0.98、至少约0.99或更大的接受者操作特征下面积(AUROC)鉴定疾病、病症或异常病况的存在或不存在。AUROC可以被计算为与训练过的分类器在将样本分类为患有或不患有疾病、病症或异常病况时相关的接收者操作特性(ROC)曲线的积分(例如，ROC曲线下的面积，或AUC)。

[0218] 训练过的分类器可以被进行调整或调优，以提高鉴定疾病、病症或异常病况的性能、准确度、PPV、NPV、临床灵敏度、临床特异性或AUC中的一项或多项。训练过的分类器可以通过调整训练过的分类器的参数(例如，如本文别处所描述的用于对样本进行分类的截断值的集，或神经网络的权重)来调整或调优。训练过的分类器可以在训练过程中或在训练过程结束后被不断进行调整或调优。

[0219] 在对训练过的分类器进行初步训练后，可以鉴定最有影响或最重要的输入的子集，以将其纳入用于进行高质量的分类。例如，可以鉴定最有影响或最重要的多个输入变量的子集，以将其纳入用于进行疾病、病症或异常病况的评估的高质量的分类或鉴定。可以基于分类指标对多个输入变量或其子集进行分等级，该分类指标指示每个输入变量关于对疾病、病症或异常病况的评估进行高质量分类或鉴定的影响或重要性。这种指标可用于(在某些情况下显著地)减少可用于将训练过的分类器训练到期望的性能水平(例如，基于期望的最小准确度、PPV、NPV、临床灵敏度、临床特异性、AUC或其组合)的输入变量(例如，预测变量)的数量。例如，如果在训练过的分类器中用包含几十个或几百个输入变量的多个来训练训练过的分类器产生的分类的准确度超过99％，那么仅用不多于约5个、不多于约10个、不多于约15个、不多于约20个、不多于约25个、不多于约30个、不多于约35个、不多于约40个、不多于约45个、不多于约50个或不多于约100个这种多个当中最有影响或最重要的输入变量的所选子集代为训练训练过的分类器，能够得到降低的但仍可接受的分类准确度(例如，至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、

至少约81％、至少约82％、至少约83％、至少约84％、至少约85％、至少约86％、至少约87％、至少约88％、至少约89％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％或至少约99％)。该子集可以通过对全部多个输入变量进行分等级排序并选择具有最佳分类指标的预定数量的(例如,不多于约5个、不多于约10个、不多于约15个、不多于约20个、不多于约25个、不多于约30个、不多于约35个、不多于约40个、不多于约45个、不多于约50个或不多于约100个)输入变量来选择。

[0220] D.数字处理装置

[0221] 在一些示例中,本文所述的主题可以包括数字处理装置或其用途。在一些示例中,数字处理装置可以包括执行该装置的功能的一个或多个硬件中央处理单元(CPU)、图形处理单元(GPU)或张量处理单元(TPU)。在一些示例中,数字处理装置可以包括被配置为执行可执行指令的操作系统。在一些示例中,数字处理装置可以连接到计算机网络。在一些示例中,数字处理装置可以连接到互联网。在一些示例中,数字处理装置可以连接到云计算基础设施。在一些示例中,数字处理装置可以连接到内联网。在一些示例中,数字处理装置可以连接到数据存储装置。

[0222] 合适的数字处理装置的非限制性示例包括服务器计算机、台式计算机、膝上型计算机、笔记本电脑、小型笔记本电脑、上网本电脑、上网板电脑(netpad computer)、机顶计算机、手持计算机、互联网电器、移动智能手机和平板电脑。合适的平板计算机可以包括,例如,具有小册子、平板和可转换配置的那些计算机。

[0223] 在一些示例中,数字处理装置可以包括被配置为执行可执行指令的操作系统。例如,操作系统可以包括软件(包括程序和数据),其管理装置的硬件并提供用于执行应用程序的服务。操作系统的非限制性示例包括Ubuntu、FreeBSD、OpenBSD、NetBSD®、Linux、Apple® Mac OS X Server®、Oracle® Solaris®、Windows Server® 和Novell® NetWare®。合适的个人计算机操作系统的非限制性示例包括Microsoft® Windows®、Apple® Mac OS X®、UNIX®和类似UNIX的操作系统(如GNU/Linux®)。在一些示例中,操作系统可以由云计算提供,并且云计算资源可以由一个或多个服务提供商提供。

[0224] 在一些示例中,装置可以包括存储和/或存储器装置。该存储和/或存储器装置可以是一个或多个用于临时或永久地存储数据或程序的物理装置。在一些示例中,装置可以是易失性存储器,并且需要电源来维持存储的信息。在一些示例中,装置可以是非易失性存储器,并且在数字处理装置未通电时保留所存储的信息。在一些示例中,非易失性存储器可以包括闪存。在一些示例中,非易失性存储器可以包括动态随机存取存储器(DRAM)。在一些示例中,非易失性存储器可以包括铁电随机存取存储器(FRAM)。在一些示例中,非易失性存储器可以包括相变随机存取存储器(PRAM)。在一些示例中,装置可以是存储装置,包括例如,CD-ROM、DVD、闪存装置、磁盘驱动器、磁带驱动器、光盘驱动器和基于云计算的存储装置。在一些示例中,存储和/或存储器装置可以是如本文公开的那些装置的组合。在一些示例中,数字处理装置可以包括显示器,用于将视觉信息发送给用户。在一些示例中,显示器可以是阴极射线管(CRT)。在一些示例中,显示器可以是液晶显示器(LCD)。在一些示例中,

显示器可以是薄膜晶体管液晶显示器(TFT-LCD)。在一些示例中,显示器可以是有机发光二极管(OLED)显示器。在一些示例中,OLED显示器可以是无源矩阵OLED(PMOLED)或有源矩阵OLED(AMOLED)显示器。在一些示例中,显示器可以是等离子显示器。在一些示例中,显示器可以是视频投影仪。在一些示例中,显示器可以是如本文公开的那些装置的组合。

[0225] 在一些示例中,数字处理装置可以包括输入装置以从用户接收信息。在一些示例中,输入装置可以是键盘。在一些示例中,输入装置可以是定点装置,包括例如鼠标、轨迹球、轨迹板、操纵杆、游戏控制器或指示笔。在一些示例中,输入装置可以是触摸屏或多点触摸屏。在一些示例中,输入装置可以是用于捕获语音或其他声音输入的麦克风。在一些示例中,输入装置可以是摄像机,用于捕获运动或视觉输入。在一些示例中,输入装置可以是如本文公开的那些装置的组合。

[0226] E.非暂时性计算机可读存储介质

[0227] 在一些示例中,本文公开的主题可以包括一种或多种用程序编码的非暂时性计算机可读存储介质,所述程序包括可由操作系统执行的指令。操作系统可以是联网数字处理装置的一部分。在一些示例中,计算机可读存储介质可以是数字处理装置的有形部件。在一些示例中,计算机可读存储介质可以从数字处理装置移除。在一些示例中,计算机可读存储介质可以包括,例如,CD-ROM、DVD、闪存装置、固态存储器、磁盘驱动器、磁带驱动器、光盘驱动器、云计算系统和服务等。在一些示例中,程序和指令可以被永久、基本上永久、半永久或非暂时地编码在介质上。

[0228] F.计算机系统

[0229] 本公开内容提供了被编程为实现本文所述方法的计算机系统。图1示出了计算机系统101,该计算机系统101被编程或以其他方式配置为执行本公开内容的方法,如存储、处理、鉴定或解释受试者(例如,患者)数据、生物数据、生物序列、参考序列、TFBS数据或TFBS特征,如z得分或TFBS可及性得分。计算机系统101可以处理本公开内容的受试者(例如,患者)数据、生物数据、生物序列或参考序列的各个方面。计算机系统101可以是用户的电子装置或相对于电子装置远程定位的计算机系统。电子装置可以是移动电子装置。

[0230] 计算机系统101包括中央处理单元(CPU,在本文也称为“处理器”和“计算机处理器”)105,其可以是单核或多核处理器,或者是用于并行处理的多个处理器。计算机系统101还包括存储器或存储器位置110(例如,随机存取存储器、只读存储器、闪存)、电子存储单元115(例如,硬盘)、用于与一个或多个其他系统进行通信的通信接口120(例如,网络适配器),以及外围装置125,如高速缓存、其他存储器、数据存储装置和/或电子显示适配器。存储器110、存储单元115、接口120和外围装置125通过诸如主板的通信总线(实线)与CPU 105通信。存储单元115可以是用于存储数据的数据存储单元(或数据存储库)。计算机系统101可以借助于通信接口120可操作地耦合到计算机网络(“网络”)130。网络130可以是因特网、互联网和/或外联网,或者内联网和/或与因特网通信的外联网。在一些示例中,网络130是电信和/或数据网络。网络130可以包括一个或多个计算机服务器,其可以实现分布式计算,如云计算。在一些示例中,借助于计算机系统101,网络130可以实现对等网络,该对等网络可以使耦合到计算机系统101的装置能够表现为客户端或服务器。

[0231] CPU 105可以执行一系列机器可读指令,这些机器可读指令可以体现在程序或软件中。指令可以存储在存储器位置,如存储器110中。指令可以被定向到CPU 105,其可以随

后对CPU 105进行编程或以其他方式配置CPU 105以实现本公开内容的方法。CPU 105执行的操作示例可以包括获取、解码、执行和写回。

[0232] CPU 105可以是电路(例如集成电路)的一部分。系统101的一个或多个其他部件可以被包括在电路中。在一些示例中,该电路是专用集成电路(ASIC)。

[0233] 存储单元115可以存储文件,如驱动程序、库和保存的程序。存储单元115可以存储用户数据,例如,用户偏好和用户程序。在一些示例中,计算机系统101可以包括在计算机系统101外部,例如位于通过内联网或因特网与计算机系统101通信的远程服务器上的一个或多个附加数据存储单元。

[0234] 计算机系统101可以通过网络130与一个或多个远程计算机系统通信。例如,计算机系统101可以与用户的远程计算机系统通信。远程计算机系统的示例包括个人计算机(例如,便携式PC)、平板电脑或平板计算机(例如,Apple® iPad、Samsung® Galaxy Tab)、电话、智能电话(例如,Apple® iPhone、支持Android的装置、Blackberry®)或个人数字助理。用户可以通过网络130访问计算机系统101。

[0235] 本文所述的方法可以通过存储在计算机系统101的电子存储位置,例如,存储器110或电子存储单元115上的机器(例如,计算机处理器)可执行代码来实现。机器可执行或机器可读代码可以以软件的形式提供。在使用过程中,代码可以由处理器105执行。在一些示例中,可以从存储单元115检索代码并将其存储在存储器110中,以供处理器105随时访问。在一些示例中,可以不包括电子存储单元115,并且将机器可执行指令存储在存储器110中。

[0236] 代码可以被预编译和配置,以用于具有适合于执行代码的处理器的机器,或者可以在运行时期间被解释或编译。可以以如下编程语言提供代码:所述编程语言可以被选择为使代码能够以预编译、解释或即时编译的方式执行。

[0237] 本文提供的系统和方法的方面,如计算机系统101,可以在编程中得到实施。技术的各个方面可以被认为“产品”或“制品”,通常呈在机器可读介质的类型上承载或以在机器可读介质的类型实施的机器(或处理器)可执行代码和/或关联数据的形式。机器可执行代码可以被存储在电子存储单元,例如存储器(例如,只读存储器、随机存取存储器、闪存)或硬盘上。“存储”类型的介质可以包括计算机、处理器等的任何或所有有形存储器,或其相关模块,例如可以在任何时候为软件编程提供非暂时性存储的各种半导体存储器、磁带驱动器、磁盘驱动器等。软件的全部或部分有时可以通过因特网或其他各种电信网络进行通信。例如,这种通信可以使得软件能够从一个计算机或处理器加载到另一个计算机或处理器,例如,从管理服务器或主机加载到应用服务器的计算机平台。因此,可以承载软件元件的另一种介质类型包括光波、电波和电磁波,例如通过有线和光学陆线网络以及经各种空中链路、在本地装置之间的物理接口之间所使用的光波、电波和电磁波。携带这种波的物理元件,例如有线或无线链路、光链路等也可以被视为承载软件的介质。如本文所用,除非限于非暂时性的、有形的“存储”介质,否则诸如计算机或机器“可读介质”的术语是指参与向处理器提供指令以供执行的任何介质。

[0238] 因此,机器可读介质,例如计算机可执行代码,可以采取多种形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括,例如,光盘或磁盘,例如

任何计算机中的任何存储装置等,其例如可用于实现附图所示的数据库等。易失性存储介质包括动态存储器,例如此类计算机平台的主存储器。有形的传输介质包括同轴电缆;铜线和光纤,包括构成计算机系统内总线的电线。载波传输介质可以采用电信号或电磁信号的形式,也可以采用声波或光波的形式,如在射频(RF)和红外(IR)数据通信期间生成的那些。因此,计算机可读介质的常见形式包括例如:软盘、柔性盘、硬盘、磁带、任何其他磁介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、穿孔卡纸带、带孔图案的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储芯片或盒,传输数据或指令的载波、传输此类载波的线缆或链路,或计算机可以从中读取编程代码和/或数据的任何其他介质。计算机可读介质的许多这些形式可以涉及将一个或多个指令的一个或多个序列载送给处理器以供执行。

[0239] 计算机系统101可以包括电子显示器135或与电子显示器135通信,该电子显示器135包括用户界面(UI)140,用于提供,例如,核酸序列、富集的核酸样品、转录因子结合谱、可及性得分、表达谱和表达谱分析。UI的示例包括,但不限于,图形用户界面(GUI)和基于web的用户界面。

[0240] 本公开内容的方法和系统可以通过一种或多种算法来实现。算法可以在由中央处理单元105执行时,通过软件来实现。例如,该算法可以探测多个调控元件、对核酸样品进行测序、富集核酸样品、确定核酸样品的表达谱、分析核酸样品的表达谱,以及存档或传播表达谱分析的结果。

[0241] 在一些示例中,本文公开的主题可以包括至少一个计算机程序或该计算机程序的用途。计算机程序可以执行一系列指令,这些指令可以在数字处理装置的CPU、GPU或TPU中执行,并被编写为执行指定任务。计算机可读指令可以实现为执行特定任务或实现特定抽象数据类型的程序模块,例如函数、对象、应用程序编程接口(API)、数据结构等。例如,可以用各种语言的各种版本来编写计算机程序。

[0242] 计算机可读指令的功能可以根据需要在各种环境中进行组合或分布。在一些示例中,计算机程序可以包括一个指令序列。在一些示例中,计算机程序可以包括多个指令序列。在一些示例中,可以从一个位置提供计算机程序。在一些示例中,可以从多个位置提供计算机程序。在一些示例中,计算机程序可以包括一个或多个软件模块。在一些示例中,计算机程序可以部分或全部包括一个或多个web应用程序、一个或多个移动应用程序、一个或多个独立运行的应用程序、一个或多个web浏览器插件、扩展、加载项或附加项或其组合。

[0243] 在一些示例中,计算机处理可以是统计学、数学、生物学或其任意组合的方法。在一些示例中,计算机处理方法包括降维法,包括例如,逻辑回归、降维、主成分分析、自动编码器、奇异值分解、傅立叶基数、奇异值分解、小波、判别分析、支持向量机、基于树的方法、随机森林、梯度提升树、逻辑回归、矩阵分解、网络聚类和神经网络。

[0244] 在一些示例中,计算机处理方法是监督机器学习方法,包括,例如,回归、支持向量机、基于树的方法和网络。

[0245] 在一些示例中,计算机处理方法是无监督的机器学习方法,包括,例如,聚类、网络、主成分分析和矩阵分解。

[0246] G.数据库

[0247] 在一些示例中,本文公开的主题可以包括一个或多个数据库,或者所述一个或多

个数据库用于存储受试者(例如,患者)数据、生物数据、生物序列或参考序列的用途。参考序列可以从数据库获得。例如,许多数据库可以适合存储和检索序列信息。在一些示例中,合适的数据库可以包括,例如,关系数据库、非关系数据库、对象定向性数据库、对象数据库、实体关系模型数据库、关联数据库和XML数据库。在一些示例中,数据库可以是基于互联网的。在一些示例中,数据库可以是基于web的。在一些示例中,数据库可以是基于云计算的。在一些示例中,数据库可以基于一个或多个本地计算机存储装置。

[0248] 使用来自基因转录调控数据库(GTRD;版本18.01)的676个TF;因为它们含有基于各种组织样品的ChIP-seq数据的详细TFBS信息。TF用1,639个已知或可能的人类TF的最新编目列表注释(图6A)。由于TF以可变频率结合的TFBS的潜在高数量,因此定义了三种不同的严格性标准(图6A):首先,GTRD中所有组织样品的所有TFBS;第二,由最大样品数的超过50%支持的那些峰值(随后称为“>50%-TFBS”;在这两个分析中,所有676个GTRD TF被包括在内);第三,由大多数样品支持的1,000个TFBS/TF(“1,000-msTFBS”;505个TF满足此标准)。

[0249] 在一些示例中,参考基因组选自GrCH38、GrCH37、hg19或hg38。

[0250] 在一些示例中,参考基因组数据库用于本文公开的方法的比对和映射步骤。

[0251] IV.使用方法

[0252] A.诊断和受试者表征方法和系统

[0253] 本文提供的方法和系统可以使用基于人工智能的方法来执行预测分析,以分析从受试者(例如,患者)获取的TFBS数据,以产生患有癌症(例如,结直肠癌、乳腺癌、前列腺癌)的受试者的评估(例如,诊断、预后、治疗选择、治疗监测、分期或亚型)的输出。例如,该应用可以将预测算法应用于所获取的TFBS数据,以生成患有癌症的受试者的评估(例如,诊断、预后、治疗选择、治疗监测、分期或亚型)。预测算法可以包括基于人工智能的预测器,例如基于机器学习的模型,其配置为处理所获取的TFBS数据以生成患有癌症的受试者的评估(例如,诊断、预后、治疗选择、治疗监测、分期或亚型)。

[0254] 机器学习预测器可以使用数据集(例如,通过对一组或多组患有癌症的患者群组的个体生物样品进行TFBS分析而生成的数据集)作为输入,并使用受试者的已知诊断(例如,分期和/或肿瘤分级,亚型,治疗反应者与非反应者,进展者与非进展者)结果作为机器学习预测器的输出来训练。

[0255] 训练数据集(例如,通过对个体的生物样品进行多分析物测定而产生的数据集)可以从,例如,具有共同特征(特征)和结果(标签)的一组或多组受试者中产生。训练数据集可以包括一组特征和对应于与诊断有关的特征的标签。特征可以包括特性,例如cfDNA测定测量的某些范围或类别,如z得分、可及性得分等。例如,在给定时间点从给定受试者收集的一组特征可以共同作为诊断标志,其可以指示在给定时间点已鉴定出受试者的癌症。特性还可以包括指示受试者的诊断结果的标签,如针对一种或多种癌症。

[0256] 标签可以包括结果,例如,受试者的已知诊断结果(例如,分期、亚型、肿瘤分级、或进展者与非进展者)。结果可包括与受试者癌症相关的特性。例如,特性可以指示受试者患有一种或多种癌症。

[0257] 训练集(例如,训练数据集)可以通过对与一组或多组受试者(例如,患有或不患有一种或多种癌症的受试者(例如,患者)的回顾性和/或前瞻性群组)相对应的一组数据进行

随机采样来选择。或者,训练集(例如,训练数据集)可以通过对与一组或多组受试者(例如,患有或不患有一种或多种癌症的受试者(例如,患者)的回顾性和/或前瞻性群组)相对应的一组数据进行比例采样来选择。训练集可以在与一组或多组受试者(例如,来自不同临床站点或试验的患者)相对应的数据集之间平衡。可以训练机器学习预测器,直到满足用于准确性或性能的某些预定条件,例如具有与诊断准确性测量相对应的最小期望值。例如,诊断准确性量度可以对应于对受试者中一种或多种癌症的诊断、分期或亚型的预测。

[0258] 诊断准确度量度的示例可以包括灵敏度、特异性、阳性预测值(PPV)、阴性预测值(NPV)、准确性以及与检测或预测癌症(例如,结直肠癌)的诊断准确性相对应的受试者工作特征(ROC)曲线的曲线下面积(AUC)。

[0259] 在一方面,本公开内容提供了一种用于检测受试者中的疾病的存在或不存在或诊断受试者中的疾病的计算机实现方法,该方法包括:(a)提供从受试者提取的脱氧核糖核酸(DNA)的一组序列读取;(b)使用该组序列读取以产生转录因子的覆盖模式;(c)处理覆盖模式以提供信号;(d)用参考信号处理所述信号,其中所述信号和参考信号具有不同的频率,从而检测受试者中疾病的存在或不存在或诊断受试者中的疾病。

[0260] 在一些示例中,DNA是无细胞DNA。

[0261] 在一些示例中,疾病是癌症。

[0262] 在一些示例中,(b)包括将序列读取与参考序列比对以提供经比对的序列模式,选择与转录因子的结合位点相对应的经比对序列模式的区域,以及在这些区域中将经比对的序列模式归一化。

[0263] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-Golay滤波器。

[0264] 在一些示例中,受试者是人。

[0265] 在一些示例中,转录因子是癌症特异性转录因子。

[0266] 在一些示例中,确定至少2、或至少5、或至少10、或至少15、或至少20、或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中,以训练能够区分健康受试者与癌症患者的分类器。

[0267] 在一些示例中,确定至少2、或至少5、或至少10、或至少15、或至少20、或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中,以训练能够区分疾病进展者与非进展者、多个疾病亚型之中的疾病亚型、多个疾病分期之中的疾病分期、或其任何组合的分类器。

[0268] 在一些示例中,确定至少2,或至少5,或至少10,或至少15,或至少20,或至少25个转录因子结合位点的可及性得分并将其输入机器学习模型中,以训练能够区分疾病治疗反应者与非反应者的分类器。

[0269] 在一方面,本文所述的方法允许按肿瘤类型对患者进行分类,肿瘤类型包括,例如,可能对患者管理(包括治疗计划和反应能力)具有重要临床意义的肿瘤亚型(例如,前列腺癌、结直肠癌、乳腺癌、肺癌的亚型)。因此,本文提供的用于基于患者样品(例如,血液、血浆或血清样品)在体内定位肿瘤特异性转录因子结合的方法,使得非编码基因组的关键部分适合于临床分析。

[0270] 在一些示例中,该方法区分疾病的亚型。

[0271] 在一些示例中,该方法区分癌症的亚型。

[0272] 在一些示例中,该方法区分前列腺癌、结直肠癌、乳腺癌和肺癌的亚型。

[0273] 在一些示例中,该方法区分患有前列腺腺癌或小细胞神经内分泌前列腺癌的前列腺癌亚型患者。

[0274] 在另一方面,本公开内容提供了用于监测受试者中疾病的进展或消退的计算机实现的方法,该方法包括:(a)提供在第一时间从受试者提取的脱氧核糖核酸(DNA)的第一组序列读取和在第二时间从受试者提取的DNA的第二组序列读取,所述第二时间晚于所述第一时间;(b)使用第一组序列读取产生转录因子的第一覆盖模式,并使用第二组序列读取产生转录因子的第二覆盖模式;(c)处理第一覆盖模式以提供第一信号,并且处理第二覆盖模式以提供第二信号;(d)用参考信号处理第一信号,其中第一信号和参考信号具有不同的频率;(e)用参考信号处理第二信号,其中第二信号和参考信号具有不同的频率;以及(f)基于用参考信号对第一信号和第二信号的处理,监测受试者中疾病的进展或消退。

[0275] 在一些示例中,DNA是无细胞DNA。

[0276] 在一些示例中,该疾病是癌症。

[0277] 在一些示例中,(b)包括将第一组序列读取和第二组序列读取与参考序列比对,以分别提供第一经比对序列模式和第二经比对序列模式,选择与转录因子的结合位点相对应的第一经比对序列模式和第二经比对序列模式的区域,并在所述区域中对第一经比对序列模式和第二经比对序列模式进行归一化。

[0278] 在一些示例中,(c)包括使用低通滤波器。在一些示例中,(c)包括使用Savitzky-Golay滤波器。

[0279] 在一些示例中,受试者是人。

[0280] 在一些示例中,转录因子是癌症特异性转录因子。

[0281] 在另一方面,本公开内容提供了一种用于检测或诊断受试者中的疾病的系统,其包括处理器,该处理器被配置为:(i)使用从受试者中提取的脱氧核糖核酸(DNA)的序列读取以生成转录因子的覆盖模式;(ii)处理覆盖模式以提供信号,其中所述信号具有与参考信号不同的频率;以及(iii)基于所述信号,为受试者提供对疾病的检测或诊断。

[0282] 在另一方面,本公开内容提供了一种用于在受试者的疗程期间或之后监测疾病的进展或消退的系统,其包括处理器,该处理器被配置为:(i)在治疗期间或之后,使用在第一时间从受试者中提取的脱氧核糖核酸(DNA)的第一组序列读取,和在晚于第一时间的第二时间从受试者中提取的DNA的第二组序列读取,以产生对应于第一组序列读取的转录因子的第一覆盖模式和对应于第二组序列读取的转录因子的第二覆盖模式;(ii)处理第一覆盖模式以提供第一信号,并且处理第二覆盖模式以提供第二信号,其中第一信号和第二信号具有与参考信号不同的频率;以及(iii)基于使用参考信号对第一信号和第二信号的处理,在受试者的疗程期间或之后监测疾病的进展或消退。

[0283] 在另一方面,本公开内容提供了一种用于检测或诊断受试者中的疾病的系统,其包括处理器,该处理器被配置为:(i)使用从受试者中提取的脱氧核糖核酸(DNA)的序列读取以生成转录因子的覆盖模式;(ii)处理覆盖模式以提供信号,其中所述信号具有与参考信号不同的频率;以及(iii)基于所述信号,为受试者提供对疾病的检测或诊断。

[0284] 在一些实施方案中,训练过的分类器可以以至少约5%、至少约10%、至少约15%、

至少约20％、至少约25％、至少约30％、至少约35％、至少约40％、至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约81％、至少约82％、至少约83％、至少约84％、至少约85％、至少约86％、至少约87％、至少约88％、至少约89％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％或更高确定受试者处于疾病、病症或异常病况(例如,癌症)的风险。

[0285] 训练过的分类器可以以至少约50％、至少约55％、至少约60％、至少约65％、至少约70％、至少约75％、至少约80％、至少约81％、至少约82％、至少约83％、至少约84％、至少约85％、至少约86％、至少约87％、至少约88％、至少约89％、至少约90％、至少约91％、至少约92％、至少约93％、至少约94％、至少约95％、至少约96％、至少约97％、至少约98％、至少约99％、至少约99.1％、至少约99.2％、至少约99.3％、至少约99.4％、至少约99.5％、至少约99.6％、至少约99.7％、至少约99.8％、至少约99.9％、至少约99.99％、至少约99.999％或更高的准确度确定受试者处于疾病、病症或异常病况的风险中。

[0286] 在将受试者鉴定为患有疾病、病症或异常病况后,可以向受试者提供治疗干预(例如,开出适当的疗程以治疗受试者的疾病、病症或异常病况)。治疗干预可以包括开具有效剂量的药物,疾病、病症或异常病况的进一步测试或评估,对疾病、病症或异常病况的进一步监测或其组合。如果受试者目前正在通过疗程治疗疾病、病症或异常病况,则该治疗干预可以包括随后的不同疗程(例如,由于当前疗程无效,而用于提高治疗效果)。

[0287] 治疗干预可以包括建议受试者进行二次临床测试,以确认对疾病、病症或异常病况的诊断或其他评估。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0288] 可以在一段时间内评估多个输入变量(例如,TFBS信息),以监测患者(例如,患有疾病、病症或异常病况或正因疾病、病症或异常病况接受治疗的受试者)。在这种情况下,患者样品的输入变量(例如,TFBS信息)可以在疗程中发生变化。例如,由于有效治疗而导致疾病、病症或异常病况风险降低的患者的TFBS信息可以移向健康受试者(例如,没有疾病、病症或异常病况的受试者)的谱或分布。相反,例如,由于无效治疗而导致疾病、病症或异常病况风险增加的患者的TFBS信息可以移向疾病、病症或异常病况风险较高,或疾病、病症或异常病况更晚期状态的受试者的谱或分布。

[0289] 可以通过监测治疗受试者的疾病、病症或异常病况的疗程来监测受试者的疾病、病症或异常病况。监测可以包括在两个或更多个时间点评估受试者的TFBS信息。评估可以至少基于在两个或更多个时间点中的每个时间点确定的TFBS信息。

[0290] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示一种或多种临床指征,如(i)对受试者的疾病、病症或异常病况的检测或诊断,(ii)受试者的疾病、病症或异常病况的预后,(iii)受试者的疾病、病症或异常病况的风险增加,(iv)受试者的疾病、病症或异常病况的风险降低,(v)用于治疗受试者的疾病、病症或异常病况的疗程有效,以及(vi)用于治疗受试者的疾病、病症或异常病况的疗程无效。

[0291] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示对受试者的疾病、病症或异常病况的诊断。例如,如果在较早的时间点未在受试者中检测到疾

病、病症或异常病况,但在较晚的时间点在受试者中检测到了疾病、病症或异常病况,则该差异指示对受试者的疾病、病症或异常病况的检测或诊断。基于对受试者的疾病、病症或异常病况的检测或诊断的这种指示可以做出临床行动或决定,例如,对受试者开出新的治疗干预。临床行动或决定可以包括建议受试者进行二次临床测试,以确认对疾病、病症或异常病况的诊断。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0292] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示受试者的疾病、病症或异常病况的预后。

[0293] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示受试者患有疾病、病症或异常病况的风险增加。例如,如果在较早的时间点和较晚的时间点都在受试者中检测到疾病、病症或异常病况,则该差异可以表明受试者患有该疾病、病症或异常病况的风险增加。基于对疾病、病症或异常病况的风险增加的这种指示可以做出临床行动或决定,例如,为受试者开出新的治疗干预或改变治疗干预(例如,结束当前治疗并开出新的治疗)。临床行动或决定可以包括建议受试者进行二次临床测试,以确认疾病、病症或异常病况的风险增加。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0294] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息中的差异可以指示受试者患有疾病、病症或异常病况的风险降低。例如,如果在较早的时间点和较晚的时间点都在受试者中检测到疾病、病症或异常病况,则该差异可以表明受试者患有该疾病、病症或异常病况的风险降低。基于对疾病、病症或异常病况的风险降低的这种指示可以做出临床行动或决定,例如,为受试者开出新的治疗干预或改变治疗干预(例如,继续或结束当前治疗)。临床行动或决定可以包括建议受试者进行二次临床测试,以确认疾病、病症或异常病况的风险增加。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0295] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示用于治疗受试者的疾病、病症或异常病况的疗程有效。例如,如果在较早的时间点在受试者中检测到疾病、病症或异常病况,但在较晚的时间点未在受试者中检测到,则该差异可以指示用于治疗受试者的疾病、病症或异常病况的疗程有效。基于疗程对受试者的疾病、病症或异常病况有效的这种指示可以做出临床行动或决定,例如,继续或结束对受试者的当前治疗干预。临床行动或决定可以包括建议受试者进行二次临床测试,以确认用于治疗疾病、病症或异常病况的疗程有效。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0296] 在一些示例中,在两个或更多个时间点之间确定的TFBS信息的差异可以指示用于治疗受试者的疾病、病症或异常病况的疗程无效。例如,如果在较早的时间点和较晚的时间点都在受试者中检测到疾病、病症或异常病况,并且如果在较早的时间点指示了有效的治

疗,则该差异可以指示用于治疗受试者的疾病、病症或异常病况的的疗程无效。基于疗程对用于治疗受试者的疾病、病症或异常病况无效的这种指示可以做出临床行动或决定,例如,结束当前的治疗干预和/或改变至(例如,开出)对受试者的另一种新的治疗干预。临床行动或决定可以包括建议受试者进行二次临床测试,以确认用于治疗疾病、病症或异常病况的疗程无效。该二次临床测试可以包括影像学测试、血液测试、计算机断层(CT)扫描、磁共振成像(MRI)扫描、超声扫描、胸部X射线、正电子发射断层(PET)扫描、PET-CT扫描、细胞学测定或其任何组合。

[0297] B.适应症说明

[0298] 通过所公开的方法可以推断出的癌症的非限制性示例包括急性淋巴母细胞白血病(ALL)、急性髓细胞性白血病(AML)、肾上腺皮质癌、卡波济肉瘤、肛门癌、基底细胞癌、胆管癌、膀胱癌、骨癌、骨肉瘤、恶性纤维组织细胞瘤、脑干神经胶质瘤、脑癌、颅咽管瘤、成室管膜细胞瘤、室管膜瘤、成神经管细胞瘤、髓上皮瘤(medulloepithelioma)、松果体实质肿瘤、乳腺癌、支气管肿瘤、伯基特淋巴瘤、非霍奇金淋巴瘤、类癌肿瘤、宫颈癌、脊索瘤、慢性淋巴细胞性白血病(CLL)、慢性髓性白血病(CML)、结肠癌、结直肠癌、皮肤T细胞淋巴瘤、原位导管癌、子宫内膜癌、食道癌、尤因肉瘤、眼癌、眼内黑色素瘤、视网膜母细胞瘤、纤维组织细胞瘤、胆囊癌、胃癌、神经胶质瘤、毛细胞白血病、头颈癌、心脏癌、肝细胞(肝)癌、霍奇金淋巴瘤、下咽癌、肾癌、喉癌、唇癌、口腔癌、肺癌、非小细胞癌、小细胞癌、黑色素瘤、口腔癌、骨髓增生异常综合症、多发性骨髓瘤、髓母细胞瘤、鼻腔癌、鼻旁窦癌、成神经细胞瘤、鼻咽癌、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、乳头瘤病、副神经节瘤、甲状旁腺癌、阴茎癌、咽癌、垂体瘤、浆细胞赘生物、前列腺癌、直肠癌、肾细胞癌、横纹肌肉瘤、唾液腺癌、Sezary综合征、皮肤癌、小肠癌、软组织肉瘤、鳞状细胞癌、睾丸癌、喉癌、胸腺瘤、甲状腺癌、尿道癌、子宫癌、子宫肉瘤、阴道癌、外阴癌、瓦尔登斯特伦巨球蛋白血症和维尔姆斯瘤。

[0299] 在各种示例中,肿瘤是选自结直肠癌、晚期腺瘤、溃疡性结肠炎、克罗恩氏病、肠易激综合症(IBS)的结直肠疾病。

[0300] 在一些示例中,结直肠癌按分期分类,例如0期、I期、IIA期、IIB期、IIC期、IIIA期、IIIB期、IIIC期、IVA期、IVB期或IVC期。

[0301] 实施例

[0302] 实施例1:从无细胞DNA推断细胞特异性转录因子结合使得能够进行肿瘤亚型预测和癌症早期检测

[0303] 根据本公开内容的方法和系统,开发了分析程序以确定转录因子结合位点的可及性,并且将该程序应用于来自前列腺癌、乳腺癌或结肠癌患者的244个cfDNA样品。

[0304] 从cfDNA推断TF结合在癌症及其他方面具有巨大的诊断潜力,并且开发了改进和优化的生物信息学渠道。此过程能够解析那些涉及TFBS处核小体特征的成分,从而客观评估和比较不同血浆样品中TFBS的可及性。为了验证该渠道的临床目的,从健康供体的24个血浆样品并从转移性前列腺癌、结肠癌或乳腺癌患者的15个血浆样品中获得了深度全基因组测序(WGS)数据,其中cfDNA还包括循环肿瘤DNA(ctDNA)。此外,生成了来自具有上述肿瘤实体的患者的229个血浆样品的浅WGS数据(总计超过185亿个映射的血浆序列读取)。还包括来自结肠癌(n=592)患者和健康对照(n=177)的额外769个血浆样品(提供约2380亿个映射的血浆序列读取)。此方法剖析了单个TF,而不是使用由多种细胞类型产生的cfDNA信

号混合物并按照其他方法通过傅里叶转化进行分析而建立一般组织特异性模式。本文提供的方法和系统提供了对组织贡献和生物过程两者的见解,这允许鉴定适用于起源组织和起源肿瘤分析的谱系特异性TF。此外,通过两个相关临床应用示例,证明了癌症患者的cfDNA中TFBS的可塑性以及TF用于前列腺癌亚型分类的潜力。首先,这些基于TF的cfDNA测定能够区分前列腺腺癌和小细胞神经内分泌前列腺癌,这一区分具有重要的治疗意义。其次,大型结肠癌群组能够准确确立分辨率极限,并能够进行使用基于TF的血浆分析检测早期癌症分期的探索。

[0305] 对一小组患有晚期癌症的个体(n=5)的分析表明,cfDNA片段化模式可以用于检测非造血特征。为了更详细探索从cfDNA映射TF-核小体相互作用的潜力,在来自健康对照的血浆样品中确认了已知的造血TF-核小体足迹。用最近发布的已知或可能的人TF列表对来自基因转录调控数据库(GTRD)的TFBS编目列表进行了注释,以从cfDNA生成676个TF的全面TFBS-核小体占据图。使用本文提供的生物信息学渠道,评价了不同的严格性标准,以测量TFBS处的核小体特征,并建立指标(称为"可及性得分")和z得分统计数据,以客观地比较不同血浆样品中TFBS可及性的显著变化。出于临床目的,使用一组谱系特异性TF鉴定cfDNA的起源组织或癌症患者的起源肿瘤。最后,使用可及性得分和z得分统计数据阐明癌症患者cfDNA中TFBS可及性的变化。

[0306] 了解基因组中核小体相对于TF结合位点(TFBS)的精确位置对于理解基因如何调控是很有用的。为此,对血浆中无细胞DNA(cfDNA)的分析(其在癌症患者中还含有循环肿瘤DNA(ctDNA)),为研究人体内非侵入性TFBS提供了改善的机会。由于cfDNA主要是在酶消化后从凋亡细胞中释放出来的,因此它主要以单核小体DNA的形式循环。因此,对cfDNA片段进行测序可生成核小体图,其中"完美定位"的核小体的二分体,例如,具有高核小体偏好的位点,会导致反映核小体定相的读取的强峰,而较不优先定位的核小体的二分体显示减弱的峰或根本没有。因此,cfDNA代表由内源性生理过程生成的独特分析物,允许通过全基因组测序生成核小体占据的体内图。这可以用于通过对转录起始位点(TSS)处核小体占据的详细分析来推断表达的基因。cfDNA核小体的占据可以反映TF的足迹。在一小组患有癌症的个体(n=5)中,将cfDNA片段模式与参考数据集相匹配,以检测非造血特征。然而,TF-核小体的相互作用在很大程度上仍未被绘制出来,并且需要在人体内在基因组尺度上获得TF实时动态的测量。

[0307] 来自cfDNA的核小体定位策略和生物信息学渠道用于解决以下问题:(1)cfDNA是否准确地反映已知的TF-核小体相互作用;(2)生成包括关于676个TF的数据的最全面的TF-核小体相互作用图;(3)建立改进的指标,称为"可及性得分",以客观比较同一个人或不同个体之间在系列分析中TFBS的可及性;以及(4)定义一组谱系特异性TF,其适合于鉴定cfDNA的起源组织或癌症患者的起源肿瘤。此外,本研究还检查了从癌症患者cfDNA追踪TFBS是否能够阐明TFBS的可及性和相关途径的改变。为此,从健康供体(12位男性和12位女性)的24份血浆样品以及从转移性前列腺癌、结肠癌或乳腺癌患者的16份血浆样品中获得了高覆盖全基因组测序(WGS)数据。此外,出于验证目的,将229位患者的cfDNA浅覆盖测序数据用于上述肿瘤实体,以生成总计超过183亿个映射的血浆序列读取,从而关于非编码基因组的重要部分提供了广泛的体内观点。

[0308] 从cfDNA推断的核小体占据显示特征性TF结合足迹

[0309] 制备TFBS处的核小体占据图,并测试健康个体和癌症患者之间的相似性和差异。为此,从24个健康对照(12位男性和12位女性)获得了高覆盖率的cfDNA样品,其中绝大多数(超过90%)的cfDNA来源于白细胞的凋亡,而其他组织的贡献很小,并且11份血浆样品来自7位具有3种常见肿瘤实体的患者(例如,4例前列腺癌(P40、P147、P148和P190),1例结直肠癌(CRC;C2)和2例乳腺癌(B7和B13),ctDNA分数范围为18-78%(图5))。

[0310] 使用来自基因转录调控数据库(GTRD;版本18.01)的676个TF;因为它们含有各种组织样品的基于ChIP-seq数据的详细TFBS信息。将TF用1,639个已知或可能的人类TF的最新编目列表注释(图6A)。由于TF以可变频率结合的TFBS的潜在高数量,因此定义了三种不同的严格性标准(图6A):首先,GTRD中所有组织样品的所有TFBS;第二,由最大样品数的超过50%支持的那些峰值(随后称为“>50%-TFBS”;在这两个分析中,所有676个GTRD TF被包括在内);第三,大多数样品支持的1,000个TFBS/TF(“1,000-msTFBS”;505个TF满足此标准)。

[0311] TF-核小体相互作用的建立

[0312] 图2A-图2E示出了由无细胞脱氧核糖核酸(cfDNA)建立的TF-核小体相互作用。图2A示出了具有高度组织化,例如,定相的核小体的区域导致振荡的读取深度模式,其中读取的峰指示二分体的位置,例如,典型核小体的中点。核小体的较不明确定位产生相当浅的覆盖谱。图2B示出了从GTRD检索到676个TF的TFBS数据,并将其与已知的或可能的人类TF的编目列表比对。进行了三种不同的计算,每种计算都具有增加的严格性。图2C示出了在所有分析的cfDNA中,CCTCC结合因子(CTCF)的覆盖模式均相似,这与DNase超敏性数据一致,显示血液(GM12878)和上皮组织(例如,前列腺(LNCaP)和结肠(HCT116))的可及性大致相同。在此组以及随后的各个组中,从健康对照计算出的谱以灰色显示,而患者来源的谱则以指示的颜色显示。图2D示出了通过DNA超敏性测定证实了TF(PU.1、LYL1、SPIB)的造血谱系特异性,并且与健康对照相比,癌症患者血浆中它们的振幅降低。相反,来自癌症患者的cfDNA中上皮TF GRHL2的振幅增加。图2E示出了TF FOXA1的可及性图和DNase超敏性,说明了激素依赖性癌症,例如,前列腺癌和乳腺癌患者的优先振幅变化。

[0313] 使用来自健康对照的24个cfDNA的样品,每个样品获得平均435,135,450(范围:352,904,231-556,303,420)个测序读取。TF结合位点通常位于一系列强定位的核小体的侧面,可见为周期性振荡模式(图6A-图6E)。相反,阴性对照正常的高分子量DNA用于观察TFBS上的均匀覆盖(图8A-图8E)。被强定位的核小体阵列包围的CTCF结合位点产生的振荡覆盖模式在所有分析的样品中保持相似,无论cfDNA是来自健康对照还是来自癌症患者(图2C)。这些结果与细胞系GM12878(来自欧洲血统的女性供体的B淋巴细胞细胞系)、LNCaP(对雄激素敏感的人前列腺腺癌细胞系)和HCT116(人结肠癌细胞系)的DNA元素百科全书(ENCODE)数据库中的DNase超敏性测定一致(图2C)。

[0314] 与健康对照相比,癌症患者血浆中的ctDNA改变了造血细胞与上皮细胞DNA之间的平衡,例如,导致源自癌症的样品的谱系限制性造血TF富含嘌呤的box1(PU.1)、LYL1(淋巴母细胞白血病1)和淋巴细胞谱系限制性转录因子SPIB的振幅降低,以及TF GRHL2(上皮细胞的先驱TF)的振幅增加(图2D)。还证实了这些TF的谱系特异性随着可公开获得的DNase超敏性测定的数据(图2D)。作为良好建立的TF的另一个示例,分析了FOXA1,它与内分泌驱动的乳腺和前列腺肿瘤中的核激素受体协同作用。与DNase超敏性测定一致,在前列腺癌和乳

腺癌患者的血浆样品中,观察到FOXA1优先增加的可及性。还进行了与ENCODE数据的比较,其中通过微球菌核酸酶(MNase)消化产生单核小体结合的DNA片段(图6A-图6E和图7)。进行了不依赖于覆盖率的分析(图6B),并且计算了与单个识别序列相关的cfDNA片段的空间密度(图6C)。序列特异性TF可以具有典型的基序和重要的次级基序,其可以与其他TF的基序对应。还生成了TFBS的目录,其可受到来自1,000-msTFBS的所有676个TF和505个TTF的多于一个TF的共结合的影响(图6D)。此外,使用纯化的高分子量DNA作为阴性对照,在TFBS上观察到均匀的覆盖(图6E)。因此,这些结果表明,相应的TFBS覆盖谱彼此非常相似,从而表明该方法具有高准确性,并且对于任何给定TF所获得的模式在所有样品中都是可重现的。

[0315] 由于序列特异性TF可以具有可以与其他TF的基序相对应的典型基序和重要的次级基序,所以计算了各种TFBS之间的重叠(图6D)。生成了TFBS列表,该列表可受到多于一个TF的共结合的影响(图27)。这种重叠影响的示例是TF SP1、SP2、NF-YA和NF-YB,其重叠范围在10%至36%之间,其中TFBS-核小体谱确实相似(图13A)。血液中cfDNA的主要起源特别反映在位于谱系限制性造血TF(如富含嘌呤的box1(PU.1)、LYL1(淋巴母细胞白血病1)和淋巴细胞谱系-限制性转录因子SPI-B)结合位点侧面的定位良好的核小体中(图13A)。相反,GRHL2(上皮细胞的先驱TF)的TFBS谱示出了显著降低的振幅(图13A)。

[0316] 应用上述三种不同严格性标准评价了被强定位核小体阵列包围的CTCF结合位点,并优先观察了>50%-TFBS和1,000-msTFBS的预期振荡模式(图2C-图2E)。此外,使用CTCF分别评价不同的结合位点(图17),并且作为额外的确认覆盖率进行了独立分析(图17),并计算了与单个识别序列有关的cfDNA片段的空间密度。产生的热图示出了在大部分分析位点,核小体定相是均匀的,这与覆盖谱一致。

[0317] 图7示出了各种TF的TF-核小体相互作用图。示出了cfDNA和MNase-seq在转录因子结合位点周围的覆盖谱之间的另外的比较。CTCF作为具有多个不同结合位点的TF的特别示例

[0318] 为了探索相同TF的不同TFBS,使用了CCCTC结合因子(CTCF)。CTCF在哺乳动物基因组中存在于55,000-65,000个结合位点处。这些位点中,约有5,000个是超保守的,约50%位于基因间区域,约15%位于启动子附近,约40%是基因内的。此外,染色体被划分为进化保守的高阶染色体结构,命名为拓扑关联结构域(TAD),并且它们的边界富集了CTCF和黏着素的结合位点。在哺乳动物中,基因组CTCF结合位点的15%位于TAD边界,而其他85%位于TAD内部。

[0319] 分别分析了重叠或超出TAD边界、TSS近端(例如,在约2kbp内)或远端(大于2kbp)的CTCF位点,以及超保守位点。对GTRD中所有组织类型进行分析,并得到不同的CTCF覆盖模式,其中超保守的CTCF位点振幅最大(图14A,左分图)。当分析限于那些在GTRD中在超过所有样品的50%中被调用的结合位点时,得到的谱彼此变得更加相似(图14A,右分图)。作为TFBS信号的不依赖于覆盖率的确认,将每个cfDNA片段的长度绘制为片段中点到CTCF结合位点距离的函数。由此得到的热图证实了与基于覆盖率的振荡模式一致的信号周期性(图6B)。此外,为了更严密地分析与单个识别序列相关的片段的分布,计算了以TFBS为中心的2kb区域内cfDNA片段的空间密度,并根据中心40bp的覆盖率对位点进行分等级。由此得到的热图示出了在所分析的大多数位点中的核小体定相是均匀的(图6C),这再次与覆盖谱一致。

[0320] “可及性得分”能够准确从cfDNA推断TF的结合

[0321] 结合位点(其中核小体通过干预TF结合而重新定位)确保相应的DNA对于蛋白质和转录和复制机制是可及的。一些TF示出均匀间隔的核小体峰,包括其结合位点(例如,图13B中的PU.1和GRHL2),而其他TF在其结合位点具有更宽的波谷(trough)(例如,图13A),类似于TSS的波谷。对于后者,测量了TF的基本结合位点宽度差异(图8D)。该测量鉴定了55个TFBS,其中TFBS超过300bp,其中26个具有接近二核小体大小(312-352bp)的结合位点(图28A-图28B)。为了测试这些模式是否是与CpG岛启动子结合的副作用,绘制了55个宽TF与具有狭窄定义结合位点的TF的CpG密度(箱形图)和与CpG岛的共定位的图(条形图/饼图)。

[0322] 某些谱系特异性TF适合于确定血浆DNA的起源组织。然而,要确定哪些TF在这种应用中可能有用,则需要评价TF的可及性(例如,在cfDNA中它们的结合位点处)。常规方法可能无法评价cfDNA中TF在其结合位点处的可及性,以作为其活性的代表。为了实现这种方法,研究了TF特异性核小体的覆盖谱。分别对转录起始位点(TSS)内和外的TFBS(图8A)和所有GTRD组织与>50%-TFBS(图8B)进行计算。这些分析表明,平均TFBS模式包括两个信号:TSS近端(在TSS的2kb内,产生“低频模式”)和TSS远端(与TSS峰相距超过2kb,产生“高频模式”),对应于更均匀间隔的峰值信号。为了抑制并非由优先核小体定位对覆盖率造成的影响,并从核小体数据中去除局部偏差,将Savitzky-Golay滤波器用于去趋势处理(图3A)。然后将获得的低频信号用于对高频信号进行归一化,随后记录高频信号的数据范围(数据值的最大值减去最小值,对应于振幅)。由于高频信号的范围取决于TFBS的数量(图3B)(1,000-ms TFBS除外),因此这些范围值通过LOESS平滑校正,因为它们取决于TFBS的数量(图3B),然后计算等级,作为每个TFBS的可及性的量度。图3C示出了GRHL2的小波分析:沿GRHL2的TFBS周围区域的周期热图(左分图)。颜色代码表示信号功率分布的分位数。转录因子GRHL2周期的平均功率(右分图)。图3D示出了由小波分析得出的转录因子GRHL2的去趋势的原始(黑色)和重构的(红色)核小体覆盖谱。

[0323] 为了测试TF可及性评估的潜在可替选方案,通过小波分析重构了在135和235bp之间的区段的无偏、去趋势信号,并且在2,000bp的侧面TFBS上对信号的功率求和(图3E-图3F)。为了对Savitzky-Golay滤波和小波分析的性能进行基准测试,使用了cfRNA数据,与表达的(例如,大于10FPKM)TF(>50%-TFBS;Savitzky-Golay滤波:$p=1.75\times10^{-13}$;功率总和(小波分析):$p=0.0004049$;1,000-msTFBS;Savitzky-Golay滤波:$p=1.254\times10^{-11}$;分别进行Mann-Whitney-U测试)相比,观察到未表达的TF(例如,<0.01FPKM[每千碱基百万个片段])的可及性显著降低(图14B)。当将调整范围与平均DNase覆盖率(>50%-TFBS;Savitzky-Golay滤波:$p<2.2\times10^{-16}$;功率总和(小波分析):$p<2.2\times10^{-16}$;1,000-msTFBS;Savitzky-Golay滤波:$p<2.2\times10^{-16}$;分别进行Mann-Whitney-U测试)比较时,这些差异也很显著。由于Savitzky-Golay滤波的性能略好,因此该方法是有利的,然后将TFBS可及性的检测阈值定义为偏离正常样品的标准偏差的±3平均值(z得分为3)。对于基于全部或>50%-TFBS的评估,归一化可及性得分的检测阈值对于1,000-msTFBS为±253和±88,其具有较少的可分析的TF(图8C)。

[0324] 此外,从cfDNA生成了676个GTRD TF的全面TF-核小体相互作用图(图14B;图17)。TF-核小体相互作用可通过,例如,使用来自ENCODE Consortium的ChIP-seq数据集进行映射,对119个人TF周围的染色质结构进行表征。根据这些努力,得出了以TF为中心的web存储

库Factorbook,其包括167个TF的数据。然而,这些数据基于离体组织样品,而本文研究了由内源性过程产生的体内可及性。

[0325] 这些结果证实了一种用于在临床诊断中特别地使用cfDNA评估TFBS可及性的可靠方法。

[0326] 跨多种细胞类型的cfDNA中的TFBS可及性

[0327] 来自3个常见肿瘤实体的血浆样品用于证明临床应用。这项研究首先分析了来自7位患者的11份血浆样品,例如,4例为前列腺癌(P40、P147、P148和P190),1例为结直肠癌(CRC;C2)和两例乳腺癌(B7和B13)(图4A-图4F)。在Illumina NovaSeq平台上对来自C2、P40、P147和P148的cfDNA进行了测序(以平均为688,482,254(范围:541,216,395-870,285,698)个测序读取),而B7(328,515,075个读取)和B13(379,733,061个读取)已在Illumina NextSeq平台上测序。

[0328] CTCF是一种特殊的转录因子,在每个组织中均具有活性,因为它调节在整个组织中都保守的染色体3D结构。无论cfDNA是源自健康对照还是源自癌症患者,CTCF的振幅在所有所分析的样品中均保持相似(图2C)。这与来自ENCODE数据库对细胞系GM12878、LNCaP(雄激素敏感的人前列腺腺癌细胞系)和HCT116(人结肠癌细胞系)的DNase超敏性测定一致,显示了CTCF结合位点在各种组织中的可及性增加(图2C)。然而,癌症患者的ctDNA比例增加,这改变了cfDNA内造血细胞与上皮细胞的DNA之间的平衡。因此,造血TF(PU.1、Lyl-1和Spi-B)的振幅降低,而上皮TF GRH-L2的振幅增加,说明造血系统的贡献被稀释,上皮细胞的贡献增加(图2D)。这些观察结果再次与DNase超敏性测定一致(图2D)。

[0329] 作为得到确认的TF的另一个示例,分析了FOXA1,这是在不同组织中广泛表达的TF,其中它控制细胞分化和器官功能。此外,FOXA1与内分泌驱动的乳腺肿瘤和前列腺肿瘤中的核激素受体协同作用,并且在前列腺中其表达与去势抗性前列腺癌(CRPC)相关。事实上,与DNase超敏性测定一致,在前列腺癌和乳腺癌患者的血浆样品中观察到FOXA1优先增加的可及性(图2E)。

[0330] 从cfDNA推断TF结合支持前列腺癌的分子亚分型

[0331] 在一些情况下,评估组织特异性TF适合于鉴定起源肿瘤和分子亚分型的程度可以是重要的。为此,前列腺癌是一种特别感兴趣的肿瘤实体,因为对靶向AR途径的新型药剂(如阿比特龙或恩杂鲁胺)产生治疗抗性的常见机制(约20%)是腺癌转分化为治疗引发的小细胞神经内分泌前列腺癌(t-SCNC)。这种转分化具有巨大的临床意义,因为它需要改变治疗方法,并且可以研究多种TF在这种转分化过程中的参与(图4A)。

[0332] 在一个肿瘤实体中检测到多种TF可及性增加,但在其他肿瘤实体中检测不到。例如,来自CRC患者C2的血浆样品显示对c-Jun和JunD致癌基因的可及性增加(图15),并且通过DNA超敏性测定证实了结肠偏好(图15)。

[0333] 对前列腺癌样品进行了另一种分析。针对人TF在组织以及(Lambert等人,2018)提供的各种细胞类型之间的表达以及可公开获得的人蛋白质图谱筛选数据,并确认了TFs AR、HOXB13和NKX3-1得到确认的前列腺谱系特异性,这也反映在前列腺癌细胞系LNCaP的DNase超敏性测定中(图4B-图4D)。

[0334] HOXB13是高度谱系特异性的同源框TF基因,其在前列腺发育中很重要,并且在正常前列腺中维持高表达水平到成年(图4B)。NKX3-1同源框基因是在前列腺上皮成熟过程中

最早表达的基因之一,并且对前列腺上皮分化至关重要,并且是前列腺肿瘤进展所必需的(图4C)。两种TF仅在患有前列腺癌的患者的cfDNA中显示出在其结合位点处的可及性增加,此外,通过DNase超敏性测定证实了组织特异性(图15和图4B-图4D)。

[0335] 因此,这些TF仅在患有前列腺癌的患者的cfDNA中显示出在其结合位点处增加的可及性。由于AR在前列腺癌中特别相关,不仅使用了GTRD定义的AR结合位点,而且还使用了(Pomerantz等人,2015)报道的那些结合位点,从而通过分析AR顺反组鉴定出在肿瘤中具有较高结合强度的9,179个肿瘤AR结合位点(肿瘤AR结合位点,T-ARBS)和在正常样品中具有高结合强度的2,690个正常AR结合位点(正常AR结合位点,N-ARBS)。事实上,尽管N-ARBS从对照组和患者都无法可及,但T-ARBS在患者血浆样品中显示了增加的可及性(图4D)。

[0336] 汇总样品中谱系特异性转录因子的确认

[0337] 这种方法也可以应用于覆盖率较小并且被在很大程度上重排的样品。为了测试这一点,在将P148_1(819,607,690个读取)和P148_3(768,763,081个读取)下采样至约5000万个读取后,重复TF分析。该比较表明,相同的TF被鉴定为可及性增加或减少,表明测序读取较少的样品适用这些分析。随后分析了4个具有串联复制子表型(Viswanathan等人,2018)的cfDNA样品(P21_2、P111_1、P111_4、P166_1)和染色体10上有染色体破碎的1例(P143_3)(平均:52,869,911;范围:41,780,819-84,049,593)(图18A-图18B)。在这些情况下,上皮TF FOXA1、GRHL1和GRHL2,以及前列腺谱系特异性TF AR、HOXB13和NKX3-1再次显示出可及性增加(图18A-图18B),表明即使在需求有碍的情况下,也可以得到结果,此外表明,这些TF的可及性改变似乎是前列腺癌的普遍特征。

[0338] 作为对cfDNA中谱系特异性TF的鲁棒性和可再现性的进一步确认,分析了浅覆盖率(<0.2x)生成的多个cfDNA样品库,表明这些TF在大多数或全部样品中具有增加的可及性,例如,谱系特异性的TF,具有更高的可及性得分,而其他则达到平均数。为此,分别汇总cfDNA样品用于前列腺癌病例(n=69)、结肠癌病例(n=100)和乳腺癌病例(n=60),并重复分析。上皮TF GRHL2以增加的可及性持续存在,而造血TF的可及性降低(图8A-图8E)。在前列腺癌cfDNA池中,谱系特异性TF AR(340;4.0)、HOXB13(712;8.4)和NKX3-1(253;3.0)显示出增加的可及性,表明这些TF可及性的改变是前列腺癌中的普遍特征(图8A-图8E),并且这些特征普遍存在于前列腺癌中,并且可以适合于从cfDNA鉴定起源肿瘤。

[0339] 图9示出了汇总的浅覆盖cfDNA的分析。显示了上皮TF GRHL2和造血TF(PU.1、LYL1和SPIB)的前列腺癌(n=69)、结肠癌(n=100)和乳腺癌(n=60)的汇总cfDNA样品的可及性。在前列腺癌cfDNA池中也显示了谱系特异性TF AR、HOXB13和NKX3-1的可及性。

[0340] 这些分析表明,cfDNA中这些TF结合位点的可及性也可用于鉴定起源肿瘤,因为一些谱系特异性TF通常分别在上皮癌和前列腺癌中发生变化。

[0341] 对于肿瘤亚分类,使用索引病例P148,其中对间隔12个月取的两个血浆样品(P148_1、P148_3)进行分析,在12个月期间前列腺腺癌转分化为t-SCNC。这两个样品显示了TFBS可及性的显著变化((Kendall's Tau:0.7573),具体反映在多个TF中。t-SCNC不再是前列腺癌的雄激素依赖性阶段,因此,在样品P148_3中不再观察到AR结合位点的可及性(图4E)。由于它与核激素受体密切协作,对FOXA1的可及性相应降低(图4E)。此外,随着对前列腺特异性谱系TF HOXB13和NKX3-1(图4E)和上皮TF GRHL2(图10A)的结合位点的可及性降低,细胞类型同一性的改变变得明显。与神经元发育相关的TF改变包括GLI-similar 1

(GLIS1)的可及性增加(图10B),GLIS1是一种在缺氧条件下表达显著增加的TF。讨论了缺氧用于促进前列腺腺癌向雄激素非依赖性状态发展,并进一步用于下调阻遏物因子1(RE-1)沉默转录因子(REST),这而诱导神经内分泌重编程,并观察到确实显著降低REST的可及性(图4E)。此外,N-MYC参与AR信号传导抑制和神经内分泌程序调节,这反映在增加的可及性方面(图4E)。这些观察表明,在一些癌症疾病阶段,TFBS可以具有高度塑性影响途径。

[0342] 为了证明基于来自cfDNA的TFBS的前列腺癌亚型分类是可行的,进一步分析了4个t-SCNC病例(P170_2、P179_4、P198_5和P240_1)的血浆样品。对于这些情况,表明该方法也适用于通过将血浆样品P148_1(819,607,690个读取)和P148_3(768,763,081个读取)下采样至约5000万个读取而以较小覆盖率测序的cfDNA。读取的减少导致噪声水平的增加,这取决于TFBS的数量,但对于具有超过1,000个TFBS的TF则可以忽略不计(图11),因此上述高度相关的TF的分析不受影响。对上述4个样品重复分析,每个样品用约为5000万个读取进行测序,再次观察到TF AR、FOXA1、HOX-B13和NKX3-1的可及性降低,或N-MYC的可及性增加(图4F)。仅在这四个病例中的两个病例显示REST的可及性降低(P170_2和P198_5;图4F),这与通常在50%的神经内分泌前列腺癌病例中观察到REST下调一致。仅在这两个病例中,GLIS1再次确实具有增加的可及性(z得分:P170_2:4.3;P198_5:4.4),表明这种缺氧相关的TF可能与REST下调有关。

[0343] TFBS的可及性在病程期间可以改变

[0344] 为了解决TF可及性是否随时间保持稳定的问题,从4位患者(P40、P147、P148、C2)获得并分析了系列样品。分析限于1,000msTFBS,并且四对血浆样品中的三对未显示出显著差异(对照:中值:0.8404±0.0196(IQR);P40:0.8620;P147:0.8370;C2:0.8719;各自为Kendall's Tau)(图12)。

[0345] 在P147_1和P147_3之间,形成了包括RET基因的新型高振幅扩增,而C2_7失去了包括KRAS的扩增,这在C2_6中观察到。前列腺癌中的RET和CRC中的KRAS都可以影响PI3K/AKT/mTOR途径,因此对下游靶标(如TF CREB)进行了研究;然而,可及性与对照血浆样品没有区别,而且保持不变。在P40_1和P40_2之间,已经形成了对雄激素剥夺治疗(ADT)的抗性,这反映在AR基因的高水平扩增方面。然而,如果AR扩展了其转录靶标库,则在上述的T-ARBS和N-ARBS处并不明显(图18A-图18B)。该分析采用保守方法,只有在一项分析与下一项分析之间的可及性得分相差≥100时,才观察到变化,并且可以解释这些样品之间减少或有限的差异。

[0346] 对于病例P148(Pearson相关性:0.777291),TF的可及性有显著变化,其中肿瘤从前列腺腺癌(P148_1)转分化为神经内分泌肿瘤(P148_3)。神经内分泌肿瘤不再是前列腺癌的雄激素依赖阶段,因此不再需要AR结合位点的可及性,这相应地反映在这些分析中(图16A-图16C)。由于失去了确定TF GRHL2和GRHL3的上皮细胞命运的结合位点的可及性,该前列腺癌病例的细胞类型同一性的改变是显而易见的(图16A-图16C)。此外,对于其他前列腺特异性谱系TF HOXB13、NKX3-1、FOXA1、GATA2和GATA3,也观察到了相似的可及性降低。

[0347] 重要的是,如上所述,还观察到与神经元发育有关的TF变化。缺氧在晚期实体瘤中频繁发生,可以促进前列腺腺癌发展为雄激素非依赖性状态,并可以诱导神经内分泌程序。事实上,观察到GLI-similar1(GLIS1)的可及性增加,GLIS1是一种在缺氧条件下表达显著增加的TF。此外,缺氧下调前列腺癌上皮中的阻遏物因子1(RE-1)沉默转录因子(REST),并

诱导神经内分泌重编程中涉及的神经元基因的表达。REST是由雄激素消耗引起的神经内分泌分化的关键中介物,并且确实观察到REST的可及性降低(图16A-图16C)。

[0348] 还观察到与干细胞特征有关的差异。TF SOX2和SOX11在神经内分泌转分化过程中被上调。观察到这两种TF的可及性增加;然而,这些在血浆样品P148_1中已经存在,因此先于其他变化(图6E)。该示例证实了这种实时分析的另一个特征,例如,可以确定事件的顺序。与干细胞相关的另一变化是FOXA2的可及性降低(图16A-图16C)。

[0349] 在特征不明显的TF中也观察到了变化,例如TF ZNF644(ZNF644是G9a/GLP复合物的核心亚基之一,它在特定的基因位点介导组蛋白H3的Lys9的单甲基化和二甲基化,这与转录抑制有关)或ZNF701(图16A-图16C),其在转分化过程中的潜在作用尚不清楚。

[0350] 基于来自cfDNA的TF的前列腺癌分类

[0351] 为了表明该方法适用于以较小覆盖率测序的样品,以及下采样至约5000万个读取的血浆样品P148_1(819,607,690个读取)和P148_3(768,763,081个读取)。读取的减少导致噪声水平的增加,这取决于TFBS的数量,并且对于TFBS超过1,000的TF可以忽略不计(图18A-图18B)。因此,对涉及转分化至神经内分泌癌的上述高度相关的TF的可及性分析不受影响。

[0352]

	PSA	NSE	REST
P170_2	3.5	133	下调(down)
P179_4	0.56	218	nl
P198_5	29.4	>370	下调
P240_1	3.2	542.4	nl

[0353] 讨论

[0354] 这项研究提供了用于从血液中无细胞DNA推断肿瘤细胞特异性转录因子的结合的方法和生物信息学软件渠道,与临床诊断和非侵入性肿瘤分类相关。尽管一些研究在评价身体获得性改变时采用了着重于以基因为中心的方法,但该分析使用了非编码基因组的重要组成部分,着重于TFBS。由于许多TF优先结合在开放的染色质中,因此必须与核小体相互作用,因此使用主要为单核小体的cfDNA,因为它允许核小体位置的映射。相比于与体外测定相关的技术变型,该方法的独特特征是用于通过内源性生理过程生成关于TFBS的体内数据。然而,这些数据与细胞系GM12878、LNCaP或HCT116的DNase I超敏性数据高度相关,从而证明了该方法的可靠性。

[0355] 与可以使用利用由多种细胞类型产生的cfDNA信号混合物的一般组织特异性模式,并通过傅里叶变换进行分析的其他分析相反,本公开的方法和系统可以剖析单个TF并由此为临床应用建立谱系特异性TF。由于提高了TFBS分析的分辨率,因此使得能够从cfDNA监测TFBS的可及性,并且可以揭示它们在疾病过程中的可塑性,例如重编程为不同的细胞谱系。此外,鉴于其他分析可能需要超过15亿个读取/样品,但本研究表明,约5,000万个读取足以进行深入的TF分析,从而使该方法对于临床应用更加有效和经济。重要的是,此cfDNA TFBS生物信息学渠道可以对肿瘤进行分类,从而填补例如前列腺癌患者管理中的重要诊断空白。

[0356] 这项工作对用于TF剖析的当前技术提供了一些实质性的改进。首先,使用cfDNA,来自GTRD的TFBS编目列表(用高可信度TF列表注释),可以从cfDNA中分析676个TF。其次,该

生物信息学渠道用于建立改进的指标(可及性得分),以允许比较不同cfDNA样品之间TFBS的可及性。第三,使用基于对照样品(例如,参考样品)和病例样品(例如,从受试者获得的测试样品)的比较的z得分统计数据,允许鉴定TFBS可及性的显著变化。第四,对于基于cfDNA的临床应用,显示了血液(PU.1、LYL1和SPIB)、上皮(GRHL2)和前列腺谱系(AR、HOXB13、NKX3-1)的谱系特异性TF的用途。这与其他方法相反,其他方法涉及由贡献cfDNA的多种细胞类型产生的信号混合物,和通过傅里叶变换的分析。目前对临床实用性的评估表明,这些TF广泛适用于鉴定患有上皮癌或前列腺癌的个体。证明了随着时间的推移监测TFBS可及性的能力,特别是在前列腺腺癌向神经内分泌肿瘤的转分化过程中,可以从外周血非侵入性评估涉及AR、上皮、前列腺和神经元谱系的剧烈变化。

[0357] TF核小体相互作用图可以是异质的,包括产生cfDNA的所有细胞类型的信号。可以使用看上去具有较大ctDNA负荷的个体的血浆样品,这可能会影响测量的灵敏度。此外,该方法使用覆盖率相对较高(约5000万个读取)的全基因组测序,这超过用于建立SCNA的浅测序血浆方法。

[0358] 然而,晚期前列腺癌(这里分析的肿瘤实体),是表征转移性癌的顽固性和因此致死性的经典示例。转移性病变的临床活检不按常规进行,因此缺少控制前列腺癌细胞存活和进展的分子机制的详细知识。的确,肿瘤研究缺乏动态模型,特别是临床样品的动态剖析,以探索途径之间的转换和相互作用。由于TF用于调控整个基因组中的基因转录的潜在性,以及其敏锐的谱系特异性方式,因此它们的详细分析为改善临床诊断提供了独特机会。通过改进的转录调控谱剖析方法,该数据还可为进一步仔细分析非编码基因组提供基础。

[0359] 方法

[0360] 受试者

[0361] 这项研究得到了格拉茨医科大学伦理委员会的批准(批准号21-227ex 09/10[乳腺癌],21-228ex 09/10[前列腺癌],21-229ex 09/10[结直肠癌]和29-272ex 16/17[血浆DNA的高分辨率分析],根据赫尔辛基宣言进行并分别从所有患者和健康先证者获得书面知情同意书。在其他研究中已经分析了一些血浆样品,例如,患者B7和B13以及P40、P147和P148的血浆样品。

[0362] B7和B13:这些研究通过全基因组测序和RNA-Seq分析了血浆DNA,以及从两个转移性乳腺癌病例(B7,B13)中匹配并同步获得的原发肿瘤。对血浆DNA以高覆盖率测序(B7:约4.11亿个读取;约8.2x;B13:约4.55亿个读取;约9.1x),并计算出拷贝数变化。鉴定了在乳腺癌中经常发生的病灶扩增,如在B7中11q13.3(包括CCND1的15个基因)的扩增或在B13中8p11(包括FGFR1的31个基因)和17q12(包括ERBB2的46个基因)的扩增。

[0363] P40:对患者P40的初始血浆DNA分析显示,大多数常染色体上存在多个拷贝数变化,而在X染色体上未观察到拷贝数变化(图3B)。在此治疗之前,患者经受了局部放疗治疗。由于疾病进展,治疗转换到第三代LHRH拮抗剂地加瑞克。然而,尽管转换为此治疗,但在10个月后仍出现了进展,并且反复的血浆分析显示,尽管常染色体的变化相同,但携带AR基因的Xq12染色体上有病灶扩增。

[0364] B7和B13:这些研究通过全基因组测序和RNA-Seq分析了血浆DNA,以及从两个转移性乳腺癌病例(B7,B13)中匹配并同步获得的原发肿瘤。对血浆DNA以高覆盖率测序(B7:约4.11亿个读取;约8.2x;B13:约4.55亿个读取;约9.1x),并计算出拷贝数变化。按照先前的

定义(Ulz等人,2016b)鉴定了在乳腺癌中频繁发生的病灶扩增,如在B7中11q13.3(包括CCND1的15个基因)的扩增或在B13中8p11(包括FGFR1的31个基因)和17q12(包括ERBB2的46个基因)的扩增。

[0365] P21:患者P21被诊断患有转移性去势抗性前列腺癌(CRPC)。用LHRH拮抗剂地加瑞克治疗4个月后,患者表现出临床进展的迹象,随后PSA值升高。在进展期(P21_2),观察到cfDNA谱具有串联复制子表型。

[0366] P40:在诊断时,患者P40被分类为患有去势敏感性前列腺癌(CSPC)。由于患者在先前的放疗中未表现出临床反应,因此将治疗转换为使用LHRH拮抗剂地加瑞克。最初,患者表现出对雄激素阻断的良好反应(PSA值从425.3ng/mL降至115.3ng/mL),但在10个月后,他进展为CRPC(PSA:656.0ng/mL)。进展期患者的cfDNA谱(P40_2)显示X染色体上有高水平AR扩增。

[0367] P40:对患者P40的初始血浆DNA分析显示,大多数常染色体上存在多个拷贝数变化,而在X染色体上未观察到拷贝数变化(图4B)。在此治疗之前,患者经受了局部放疗治疗。由于疾病进展,患者的治疗转换为第三代LHRH拮抗剂地加瑞克(Rick等人,2013)。然而,尽管转换为此治疗,但在10个月后仍出现了进展,并且反复的血浆分析显示,尽管常染色体上的变化相同,但携带AR基因的Xq12染色体上有病灶扩增。

[0368] P111:第一样品P111_1是在诊断前列腺癌时获得的。患者的骨骼、淋巴结和肾脏已经有多处恶性病变。一线治疗是GnRH-类似物戈舍瑞林,然后进行放疗。在两个样品P111_1和P111_4之间,患者对治疗反应良好(CSPC,PSA降至15.5ng/mL)。在P111_4采样前两个月,出现临床进展,并且患者发展为CRPC。在进展期分析cfDNA样品(P111_4),观察到部分AR扩增。此外,患者接受了化疗(多西他赛),但未发现进一步反应。

[0369] P143:在收集样品P143_3之前6年,患者P143被诊断患有转移性前列腺癌。此前,患者经受了不同的抗雄激素治疗,包括第二代抗雄激素(阿比特龙)。因此,当获得样品P143_3时,对他进行了很大程度的预处理。由于疾病进展以及用不同ADT多次治疗失败后,引入化疗(微管抑制剂-卡巴他赛)。

[0370] P147:第一份血液样品(P147_1)在诊断后5年获得。患者有多处骨转移,特征为CRPC。在这5年期间,他经受了放疗治疗,并接受了多种抗雄激素。在收集P147_1样品时,出现了新的PSA增加。cfDNA分析发现Xq12(AR)和5q14.3染色体上有高水平扩增。在化疗(多西他赛)和抗雄激素(阿比特龙和恩杂鲁胺)治疗6个月后,分析了新的cfDNA样品(P147_3)。该分析揭示了在染色体10上的新的RET扩增。新的病灶事件的发展和ctDNA含量的增加(P147_1ichorCNA:52%;147_3ichorCNA:73%)与临床进展相关。

[0371] 在患者P147中,前列腺切除术与第一血浆样品之间的时间间隔为56个月。手术后二十个月,发现PSA水平升高,并开始进行放疗。诊断后二十八个月,PSA水平再次升高。患者用非类固醇抗雄激素比卡鲁胺治疗了13个月,并且在随后的4个月中,另外施用了GnRH-类似物亮丙瑞林,并且后来最终由于检测到骨转移而添加了单克隆抗体地诺单抗。当PSA水平升高时,获得血浆样品,并在Xq12(AR)和5q14.3染色体上检测到新的高水平扩增。在第一血浆样品和第二血浆样品(例如,P147_1和P147_2)的收集之间,在染色体10q11.21上发生进一步的高水平病灶扩增;这两个样品之间的时间间隔为6个月。在此期间,患者经受了化疗治疗,例如,多西他赛。

[0372] P148:P148被诊断患有前列腺腺癌。患者在首次采血前的时段内接受了ADT(P148_1)。临床医生报告患有新的骨和淋巴结病变的疾病进展(PSA:694.4)。在收集P148_1样品时,引入了化疗(多西他赛)。鉴定了多个病灶事件(MYC扩增;PTEN缺失;FOXP1、RYBP、SHQ1缺失;TMPRSS2-ERG融合),包括AR扩增(患者以前被定性为CRPC)。

[0373] 首次收集样品六个月后,获得样品P147_3。在此期间,出现大量进展并伴有多发性肝和骨转移,PSA水平为52.0ng/mL,并且NSE值大于370ng/mL。有趣的是,在样品P147_3中未检测到AR扩增,这是腺癌向神经内分泌前列腺癌转分化的特征(如Ulz等人,2016,Belic等人,2018所述)。在用卡铂和依托泊苷进行姑息治疗的短暂反应后,出现疾病进展,并且患者于2个月后死亡。

[0374] 患者P148被诊断患有前列腺腺癌。在最初诊断后的16个月获得了第一份血浆样品,此时患者的疾病明显进展,并有至骨骼的增加转移和新诊断的淋巴结病。由于疾病进展(PSA:694.41ng/mL),患者经受多西他赛治疗7个月。在此期间的第二次血浆DNA分析确认了高水平AR扩增的存在。在最后一次多西他赛治疗后五个月,出现大量进展并伴有多发性肝和骨转移,PSA水平为52.0ng/mL,并且NSE值大于370ng/mL。患者经受了卡铂和依托泊苷的姑息治疗,其中最初部分反应持续了3个月。此后,他的疾病进展了,两个月后去世。

[0375] P166:患者P166的血液样品在最初诊断为转移性前列腺癌之后2年获得。他经受了抗雄激素比卡鲁胺治疗,但疾病进展(CRPC)。在样品P166_1中鉴定了X染色体上的AR扩增作为进展的标志。由于患者发展为CRPC,因此进一步引入了化疗(多西他赛)。化疗3个月后,患者表现出部分反应。

[0376] P190:在采血前五年,患者被诊断为局部前列腺癌。两年后,他出现了局部进展,并在接下来的两年期间发展为转移性疾病,其中主要是骨转移。在这些年期间,他经受了抗雄激素比卡鲁胺治疗。患者发生了骨转移和疾病进展,并且一些骨转移以神经内分泌表型为特征。因此,临床医生对他进行了基于卡铂的化疗(卡铂/依托泊苷)。卡铂/依托泊苷治疗开始后3个月进行cfDNA分析。在进行此分析时,患者表现出良好的反应,≥50%的PSA反应,以及NSE值正常化。

[0377] 串联复制子表型:

[0378] P21_2:59,849,368个读取

[0379] P111_1:58,258,680个读取

[0380] P111_4:61,085,342个读取

[0381] P166_1:52,829,575个读取

[0382] 染色体破碎(染色体10):

[0383] P143_3:111,958,416个读取(最少的PC特异性改变)

[0384] 血浆样品的成对比较

[0385] 为了解决TF可及性是否随时间的推移保持稳定,分析了各自来自患者P40、P147和C2的两个样品。然而,在非常严格的标准下,例如,通过将分析限制在1,000-msTFBS,在这些血浆样品对中未观察到显著差异(对照:中值:0.8404±0.0196(IQR);P40:0.8620;P147:0.8370;C2:0.8719;各自为Kendall's Tau)(图12)。

[0386] 在从患者P147收集的样品P147_1和P147_3之间,出现包括RET基因的新型高振幅扩增,而C2_7失去了包括KRAS的扩增,这在先前的样品C2_6中已经观察到。前列腺癌中的

RET和CRC中的KRAS均可以影响PI3K/AKT/mTOR途径,因此对下游靶标如TF CREB进行了研究;然而,可及性与对照血浆样品没有区别,而且保持不变。在患者P40的样品P40_1和P40_2之间,已经形成了对雄激素剥夺治疗(ADT)的抗性,这反映在AR基因的高水平扩增方面。然而,如果AR扩展了其转录靶标库,则在上述的T-ARBS和N-ARBS处并不明显(图19)。将分析限制在1,000-msTFBS的一种非常保守的方法,可以解释为什么未观察到这些样品之间的差异。

[0387] TFBS的形状

[0388] 研究了TF特异性核小体的覆盖谱,因为一些TF显示出均匀间隔的核小体峰,包括其结合位点(例如图2D中的PU.1和GRHL2),而其他TF在其结合位点处具有更宽的波谷(例如在图6A中的CREM),类似于对于TSS所观察到的那些。总共鉴定出55个TFBS,其中TFBS超过300bp,并且从这些中,26个具有接近二核小体大小(312-352bp;图8D)的结合位点。对于这些模式,对于大小大于300bp的TFBS,CpG岛($p=4.2\times10^{-11}$;Mann-Whitney U检验)和TSS($p=8.5\times10^{-12}$;Mann-Whitney U检验)均鉴定出重叠显著增加(图8E)。

[0389] CTCF作为具有多个不同结合位点的TF的特别示例

[0390] 为了探索相同TF的不同TFBS,使用了CCTC结合因子(CTCF)。CTCF在哺乳动物基因组中存在于55,000-65,000个结合位点处。这些位点中,约有5,000个是超保守的,约50%位于基因间区域,约15%位于启动子附近,且约40%是基因内的。此外,染色体被划分为进化保守的高阶染色体结构,命名为拓扑关联结构域(TAD),并且它们的边界富集了CTCF和黏着素的结合位点。在哺乳动物中,基因组CTCF结合位点的15%位于TAD边界处,而其他85%的基因组CTCF结合位点位于TAD内部。

[0391] 血液采样和文库制备

[0392] 外周血是从肿瘤科的转移性前列腺癌、乳腺癌和结肠癌患者以及格拉茨医科大学血液学科的没有已知的慢性或恶性疾病的匿名健康供体收集的。根据生产商的协议,使用QIAamp循环核酸试剂盒(QIAGEN,Hilden,Germany)从血浆中分离CfDNA。如前所述进行了WGS的文库制备(Heitzer等人,2013)。

[0393] 测序

[0394] 通过奥地利维也纳CeMM的生物医学测序设施,在Illumina NovaSeq S4流通池上以2×150bp对高覆盖率肿瘤样品和对照进行测序。对于对照样品,获得了平均435,135,450(范围:352,904,231-556,303,420)个双端读取。对于肿瘤样品(P40_1、P40_2、P147_1、P147_3、P148_1、P148_3、C2_6和C2_7),对平均688,482,253个读取(范围:541,216,395-870,285,698)进行了测序。分别使用Illumina NextSeq平台(B7_1、B13_1和P190_3;平均测序产量:296,733,931个读取;范围:181,953,656-379,733,061)和HiSeq平台(P21_2、P111_1、P111_4、P143_3和P166_1;平均测序产量:52,869,911个读取;范围:41,780819-84,049,593)对另外的样品进行测序。

[0395] 对用于创建单一实体库的低覆盖率肿瘤样品,在Illumina Next-Seq或MiSeq平台上进行测序。这从69个前列腺癌样品中产生382,306,130个读取,从60个乳腺癌样品中获得254,490,128个读取,并且从100个结肠癌样品中获得604,080,473个读取。

[0396] 血浆样品的表征

[0397] 分析了一些血浆样品,例如,患者B7和B13以及P40、P147和P148的血浆样品,并基

于ichorDNA算法收录了有关血浆样品的突变、特异性SCNA和肿瘤含量的信息。

[0398] ETS致癌转录因子家族(受(Sizemore等人,2017)启发)。约50％的局限性前列腺癌和约40％的转移性前列腺癌包含TMPRSS2-ETS融合。在约50％的原发性前列腺癌中观察到的、受雄激素调节的TMPRSS2的5'非翻译区与ERG的复发基因融合(TMPRSS2-ERG基因融合)导致AR劫持ETS表达和转录程序。

[0399] ERG也被发现阻断前列腺神经内分泌细胞分化。TMPRSS2-ERG介导的前列腺干细胞和祖细胞维持的一种可能机制是通过SOX9。

[0400] 考虑到它们作为转录因子的作用,ETS因子通过多种机制介导肿瘤发生也并不令人惊讶,所述多种机制从基本的生存线索到完整的表观遗传重编程。ETS因子也影响核苷酸、能量和类固醇代谢。

[0401] P40:BRCA1中的突变:NM_007294:Q975R;特异性SCNA:TMPRSS2-ERG融合;样品2中的AR扩增;chr12扩增(包含ARID2、HDAC7);肿瘤含量:P40_1:30％,P40_2:24％。在染色体15(包含SNORD(小核仁RNA,C/D盒)基因、16(2x)和19(BRD4)上另外的病灶扩增;P40_1ichorCNA:30％;P40_2ichorCNA:24％。

[0402] P147:突变:BRCA2:T298fs;TP53:F338I;特异性SCNA:样品3中的RET扩增;AR扩增;BRAF扩增(7q34);PTEN缺失;肿瘤含量:P147_1:52％;P147_3:73％。在染色体5(XRCC4)和21(RBM11)上另外的病灶扩增;P147_1ichorCNA:52％;147_3ichorCNA:73％。

[0403] P148:突变:TP53:R213X;特异性SCNA:MYC扩增;PTEN缺失;FOXP1、RYBP、SHQ1缺失;TMPRSS2-ERG融合;AR扩增(在P148_3中消失);肿瘤含量:P148_1:38％;148_3:49％。

[0404] C2:特异性SCNA:在C2_6中的染色体12(KRAS)上的高水平扩增,在C2_7中不可见;肿瘤含量:C2_6:18％;C2_7:28％。

[0405] 转录因子结合位点定义

[0406] 从GTRD数据库下载数据,并提取每个TF的单个BED文件。通过聚焦于meta集群具有最高ChIP-seq信号的报告点来重新计算位置。创建了另外的BED文件,该文件仅包括由针对该特定转录因子分析样品最大数量的超过50％支持的峰。然后,使用UCSC提供的liftOver工具将所有BED文件转换为hg19(来自原始的hg38)。

[0407] 转录因子结合位点重叠

[0408] 为了检查转录因子的结合位点是否重叠,使用bedtools斜率将来自GTRD的结合位点的区域(由超过50％的样品支持的位点)分别在两侧增加了25bp、50bp和100bp。随后,对于每个转录因子与每个其他转录因子,通过使用经由pybedtools的bedtools相交,计算重叠数。

[0409] 单端测序数据准备

[0410] 为了增强核小体信号,修剪了测序读取以去除与接头区域相关的测序读取部分。因此,将正向测序读取修剪为仅包含53-113碱基(这可以对应于166-bp片段的中心60bp)。然后使用Burrows-Wheeler比对仪(bwa)将读取与人hg19基因组比对,并使用samtools rmdup去除PCR重复。平均覆盖率由bedtools genomecov计算。

[0411] 双端测序数据准备

[0412] 使用bwa mem将双端测序读取与人hg19基因组比对,并用picard MarkDuplicates标记PCR重复。

[0413] MNase-seq数据准备

[0414] 从ENCODE门户下载GM12878的MNase-seq实验的BAM文件。使用pysam直接从BAM文件中修剪BAM文件中的测序读取。简言之,将BAM文件中最左侧的比对位置沿各自方向位移53bp,并将序列长度调整为60bp。然后以与修剪的无细胞DNA测序数据相同的方式计算覆盖模式。

[0415] 转录因子结合位点处的覆盖模式

[0416] 对于GTRD中的每个转录因子,计算覆盖模式。为此,使用pysam count_coverage在定义的结合位点周围±1000bp的区域中提取每个区域的覆盖率数据。通过区域拷贝数变化和平均覆盖率将每个位点处的覆盖率数据归一化。对于TFBS周围的每个位置,将覆盖率取平均值,并计算95%的置信区间。如果为转录因子定义了超过100,000个位点,则随机选择100,000个位点进行分析。

[0417] 在转录因子结合位点周围的插入物大小

[0418] 为了评估转录因子结合位点周围的片段大小是否有偏差,使用了来自双端分析的插入物大小数据。遍历距离结合位点-1000bp至1000bp的每个位置,并使用pysam获取中点周围中心3bp位于此位置的(单端)测序读取。同样,从相同的样品中获取双端比对,并且将插入物大小信息指定给各个读取。然后总结了相对于TFBS的特定位置的所有插入物大小,并采样1000个数据点,并对每个位置在距TFBS-1000bp至1000bp的范围内作图。

[0419] 测量转录因子结合位点大小

[0420] 为了测量转录因子结合位点的大小,使用三阶Savitzky-Golay滤波器(窗口大小:31)对各自的覆盖模式进行平滑处理。通过搜索大于任一侧相邻20个数据点的数据点来鉴定峰。如果峰位于假定的结合位点中心50bp以内,则将其除去。紧挨结合位点峰的最接近的峰之间的距离被指定为转录因子结合位点大小。

[0421] 由于结合位点估计仅在可检测到核小体同步时才合理,因此通过各种标准对信号进行滤波:

[0422] • 高频信号振幅大于0.1

[0423] • 中心100bp的平均归一化覆盖率小于1

[0424] • 峰数少于15

[0425] • 峰之间的中位距离大于150bp

[0426] • 结合位点集包括多于500个位点

[0427] 总共228个结合位点集通过这些滤波器,并用于结合位点估计。使用Savitzky-Golay滤波器测量转录因子可及性

[0428] 两种不同的信号组成覆盖模式,并且将两种不同频率的信号提取到较低和较高范围频率中。较低范围频率数据通过Savitzky-Golay滤波器(三阶多项式且窗口大小为1001)提取。高频信号通过不同的Savitzky-Golay滤波器(三阶多项式且窗口大小为51)提取。然后通过除以低频信号的结果来归一化高频信号。然后记录高频信号的数据范围。由于具有很少描述的结合位点的转录因子的覆盖谱固有噪声较多,因此在信号范围和描述结合位点的数量上进行了LOESS(局部加权平滑)。通过平滑的LOESS校正范围值,并计算调整范围的等级。利用小波变换测量转录因子可及性

[0429] 作为测量转录因子可及性的另外方法,通过使用R包“WaveletComp”应用小波变

换。对于每个信号,在功率谱中沿2bp和512bp之间的区段记录了峰值。135bp至235bp(185bp±50bp)范围内的最高峰用于在该特定区段重构去噪的高频核小体信号。此外,使用原始数据序列的去趋势去除了任何残余基线。分析了重构信号的三个参数:信号的最大振幅、信号功率之和(振幅平方)和沿转录因子结合位点周围2000bp的绝对振幅之和。

[0430] 为了比较肿瘤与正常样品,在每个转录因子的正常样品中记录各自参数的平均值和标准偏差,并且通过如下来计算Z得分:取癌症样品中的各个参数,减去正常值的平均值,然后除以标准偏差。

[0431] 比较肿瘤和对照样品

[0432] 为了比较肿瘤样品和对照样品,比较了调节范围中各个转录因子的等级。计算肿瘤样品和每个对照样品之间的等级差异,并记录平均等级差异。而且,通过如下根据可及性等级为每个转录因子计算z得分:取相应的等级,减去对照样品的平均等级,然后除以对照样品的该转录因子等级的标准偏差,。

[0433] DNase超敏性数据分析

[0434] 从GM12878、LNCaP和HCT116细胞系的ENCODE数据库下载来自DNase超敏性实验的BAM文件。转录因子的结合位点区域使用bedtools斜率在任一侧增加了25bp。使用mosdepth提取各个结合位点处的覆盖率,并通过1百万个映射读取/样品进行归一化。

[0435] 体细胞拷贝数改变(SCNA)分析

[0436] 对于对照数据,使用samtools视图对双端的比对进行二次采样,以仅包括2%的初始比对,并使用samtools fastq将其转换为FastQ。对于癌症样品,进行了单独的低覆盖率全基因组测序。将Plasma-Seq分别应用于二次采样的FastQ文件和癌症样品的低覆盖率数据。简言之,将测序读取与人hg19基因组比对,并在预先指定的箱中对测序读取进行计数。箱的大小由理论上可映射位置的数量决定,以解释整个基因组的可映射性的差异。将读取计数归一化为读取总量,并通过在GC谱上进行LOESS平滑校正箱的GC含量。此外,将校正后的读取计数通过非癌症对照的平均读取计数/箱进行归一化,以控制另外的位置变化。

[0437] 可及性得分能够从cfDNA准确推断TF结合

[0438] 来自健康供体的样品显示TF未超过±5z得分阈值(图29A);然而,在源自癌症患者的样品中观察到非常不同的模式。例如,在来自患者P40的前列腺样品P40_1中,除GRHL2、FOXA1之外,还包括可及性高于+5z得分阈值的TF,它们与核激素受体以及前列腺谱系特异性TF HOXB13、AR和NKX3-1协同作用于内分泌驱动的前列腺肿瘤和乳腺肿瘤(图29B)。相反,造血TF,如LYL1、SPIB和EVI1(转录调节因子嗜亲性病毒整合位点1(图29B))具有低可及性。这些结果与基于ATAC-seq数据的TF分等级非常一致。在乳腺癌样品B7和B13中,与GRHL2、FOXA1和ZNF121(锌指蛋白)的ATAC-seq数据一致,检测到的可及性增加,这与细胞增殖调节和乳腺癌的发展有关。

[0439] 在结肠癌患者C2的样品中,意外地观察到ATAC-seq数据具有分等级的EVX2,EVX2是与癌症没有密切关联的TF,在COAD中最可及。的确,在该分析中,EVX2的可及性等级最高(图29C),并且核小体位置图显示EVX2的可及性大大提高(图29D)。与ATAC-seq数据一致,对于TF HNF4A、GRHL2、DLX2、HNF4G和HNF1A也观察到增加的可及性(图29D)。

[0440] 此外,如通过评价ATAC-seq数据所预测的,造血相关TF的可及性,例如LYL1、TAL1(SCL/TAL1(干细胞白血病/T细胞急性淋巴母细胞白血病[T-ALL]1、EVI1、TBX21(T-bet)和

PU.1在所有肿瘤样品中均降低(图29B-图29C)。为进一步证实cfDNA中谱系特异性TF的稳健性和可重复性,在由浅覆盖(<0.2x)生成的多个cfDNA样品池中,显示在大多数样品中具有增加的可及性的TF具有增加的可及性得分,而其他可达到平均数。为此,分别汇总cfDNA样品用于前列腺癌(n=69)、结肠癌(n=100)和乳腺癌(n=60)的病例。当重复分析时,上皮TF GRHL2和造血TF在三个上皮谱系中分别重申了其增加和减少的可及性模式。在结肠cfDNA池中,EVX2、DLX2、HNF1A、HNF4A和HNF4G,以及前列腺癌cfDNA池中的TF AR和HOXB13具有增加的可及性,而FOXA1在前列腺和乳腺池中均超过了>5z得分阈值。这证实了来自ATAC-seq数据的TF可及性估计可以从血浆DNA核小体映射可靠地推断。

[0441] 图29E提供了合并的乳腺癌、前列腺癌和结肠癌池的整体z得分图的条形图。左分图显示在至少一个肿瘤实体中可及性增加的TF;右分图总结了造血相关TF的可及性。

[0442] 图30A-图30B提供的图示出了基于TF的血浆分辨率极限和早期癌症检测。图30A提供的图示出了基于其肿瘤分级的COAD群组子样品中选定TF的可及性比较。图30B提供的图示出了分别来自I期(左分图)和II期(右分图)结肠癌群组的样品的所有504TF的逻辑回归。所有呈现的结果均为交叉验证的测试设定值。

[0443] 尽管本文中已经示出并描述了方法和系统的一些示例,但本领域技术人员将认识到,这些仅以示例的方式提供,并且并非意图被说明书限制。本领域技术人员在不脱离本文所述的范围的情况下现将想到多种变化、改变和替代。此外,应当理解,所描述的方法和系统的所有方面不限于本文所阐述的依赖于各种条件和变量的具体描述、配置或相对比例,并且该说明书旨在包括这样的替选、修改、变化或等同方案。

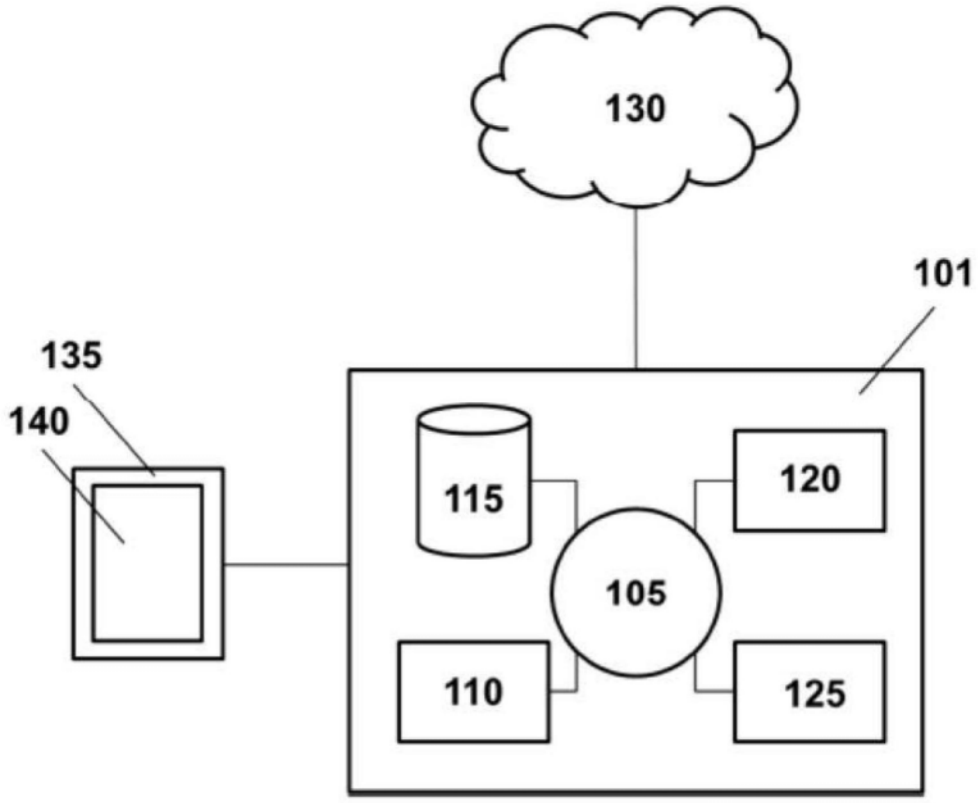


图1

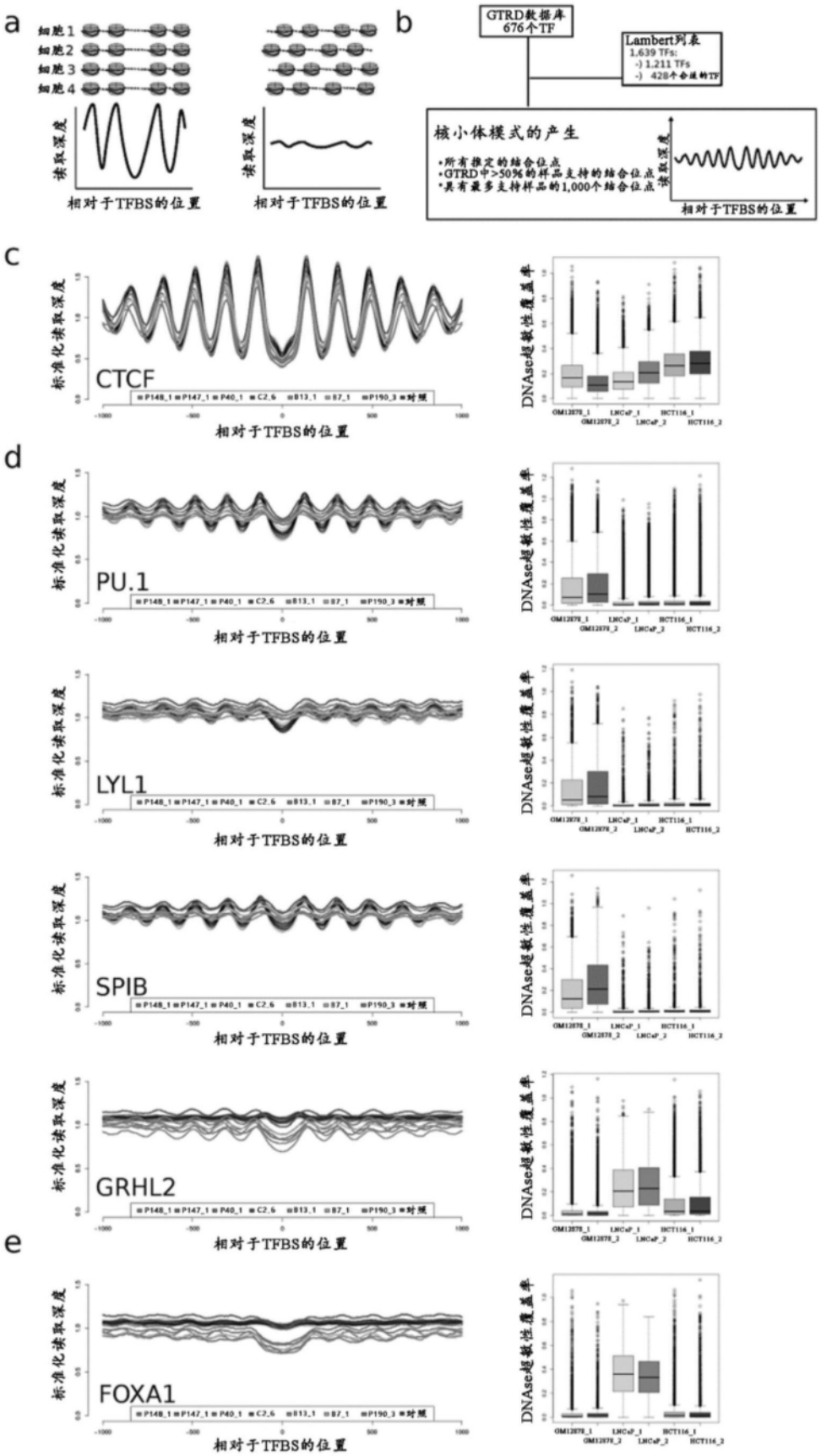

a
细胞1
细胞2
细胞3
细胞4
读取深度
相对于TFBS的位置
b
GTRD数据库
676个TF
Lambert列表
1,639 TFs:
-) 1,211 TFs
-) 428个合适的TF
核小体模式的产生
•所有推定的结合位点
•GTRD中>50%的样品支持的结合位点
•具有最多支持样品的1,000个结合位点
c
CTCF
d
PU.1
LYL1
SPIB
GRHL2
e
FOXA1
标准化读取深度
DNAse超敏性覆盖率

图2A-2E

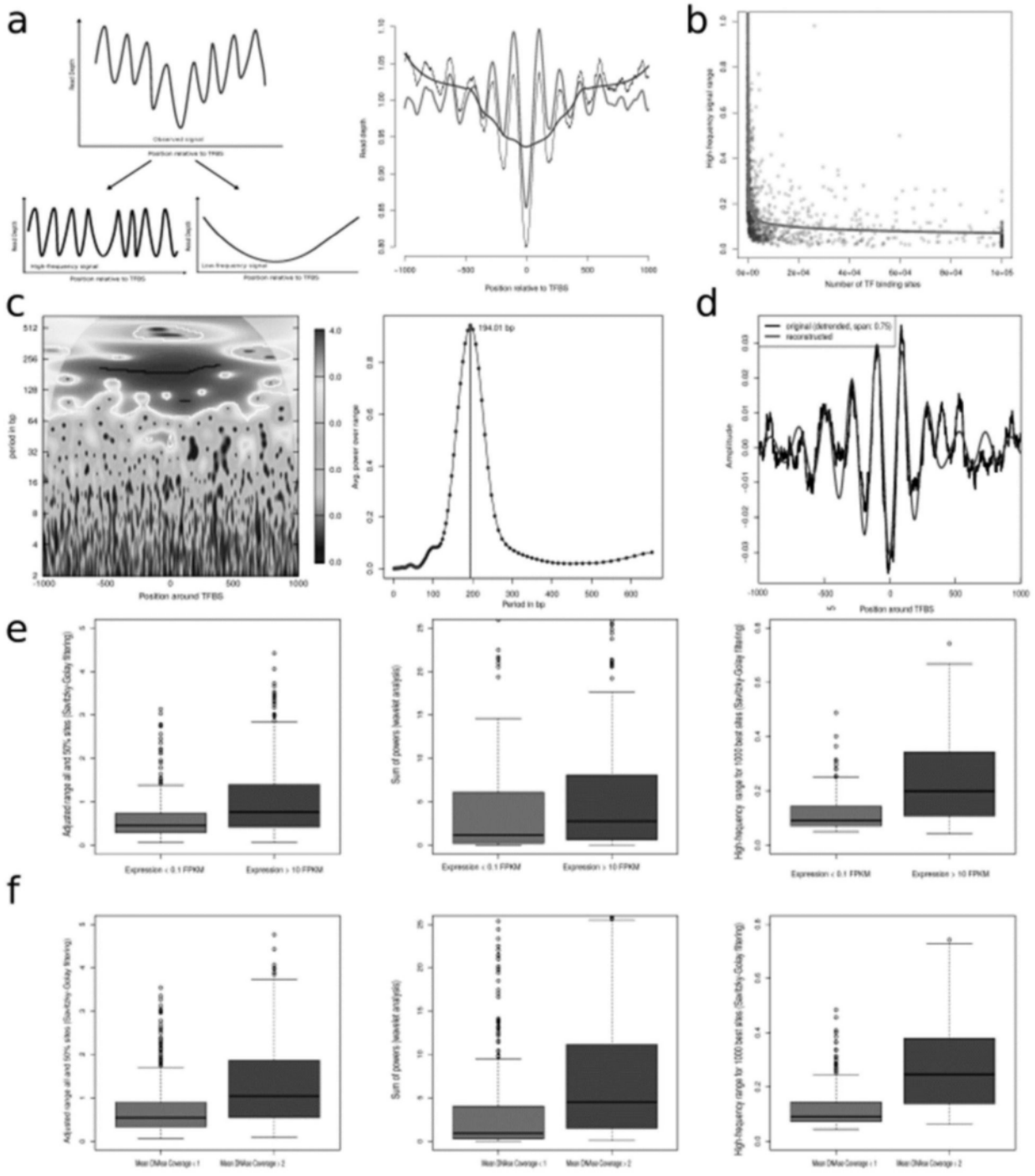

a
b
c
d
e
f
Position relative to TFBS
Number of TF binding sites
Position around TFBS
Period in bp
194.01 bp
original (detrended, span: 0.75)
reconstructed
Amplitude
Expression < 0.1 FPKM
Expression > 10 FPKM

图3A-3F

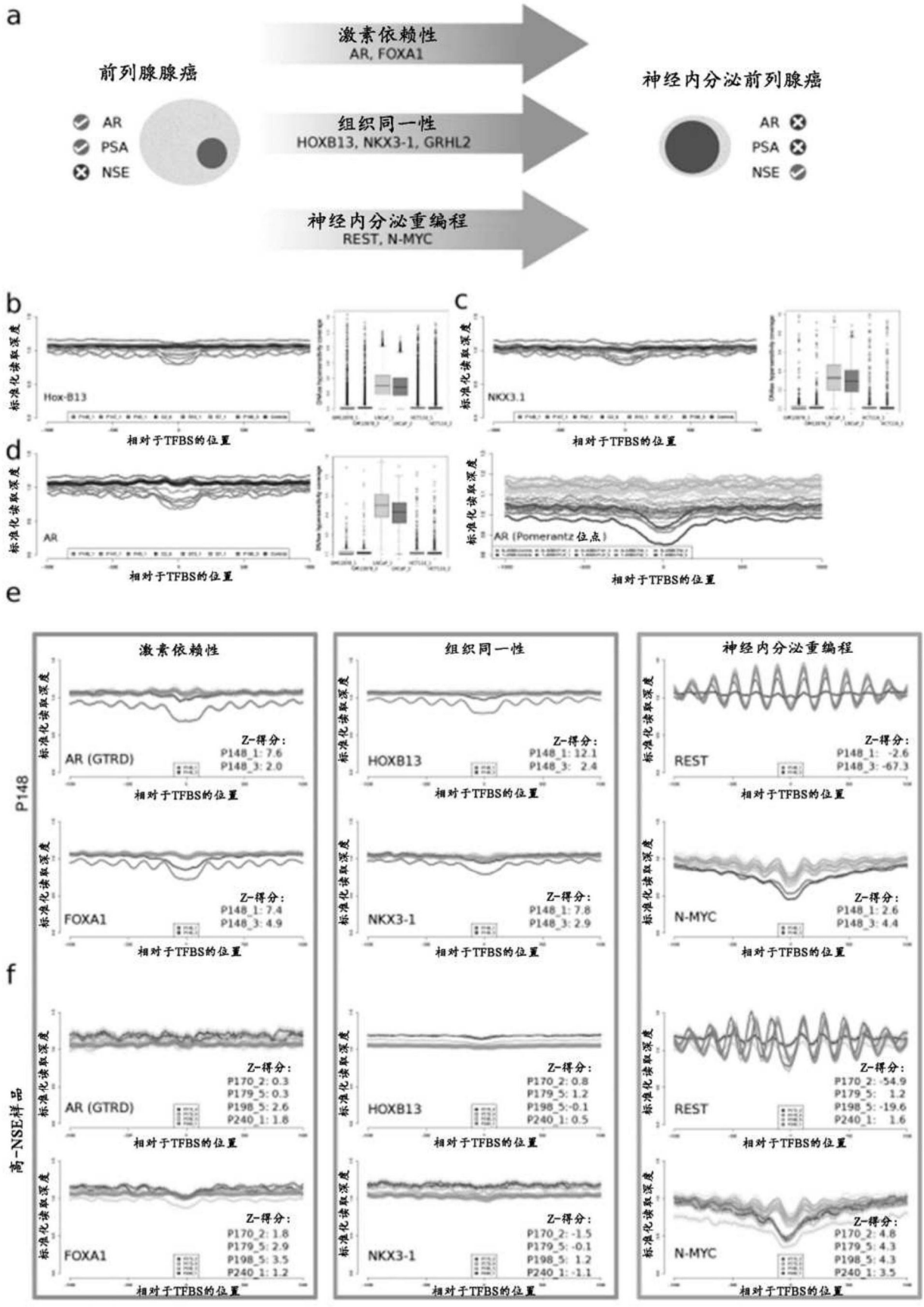

a
激素依赖性
AR, FOXA1
前列腺腺癌
AR
PSA
NSE
组织同一性
HOXB13, NKX3-1, GRHL2
神经内分泌前列腺癌
AR
PSA
NSE
神经内分泌重编程
REST, N-MYC
b
Hox-B13
c
NKX3.1
d
AR
AR (Pomerantz位点)
e
激素依赖性
组织同一性
神经内分泌重编程
P148
AR (GTRD)
FOXA1
HOXB13
NKX3-1
REST
N-MYC
f
高-NSE样品
AR (GTRD)
FOXA1
HOXB13
NKX3-1
REST
N-MYC
标准化读取深度
相对于TFBS的位置
Z-得分:

图4A-4F

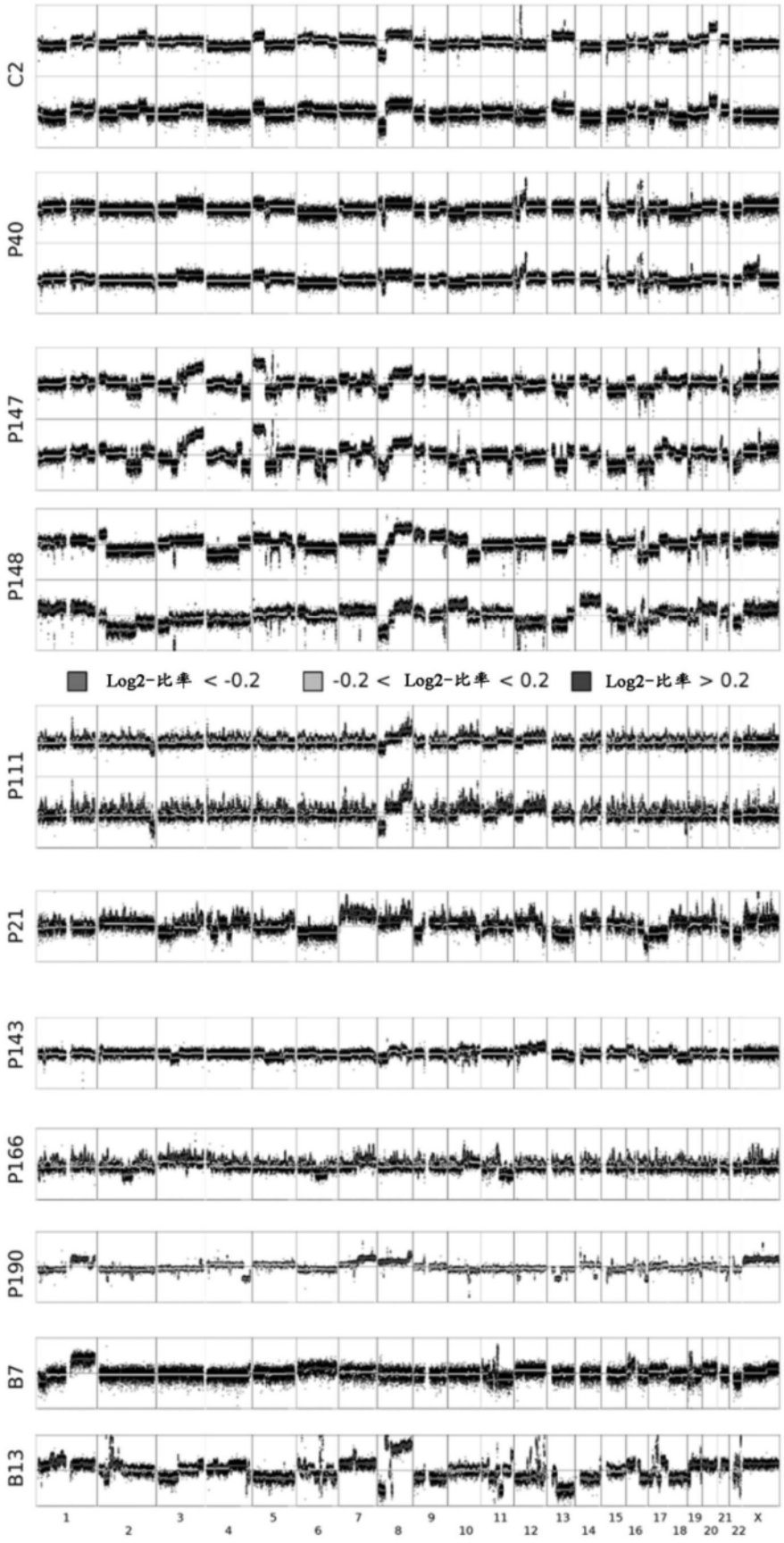

C2
P40
P147
P148
Log2-比率 < -0.2
-0.2 < Log2-比率 < 0.2
Log2-比率 > 0.2
P111
P21
P143
P166
P190
B7
B13
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X

图5

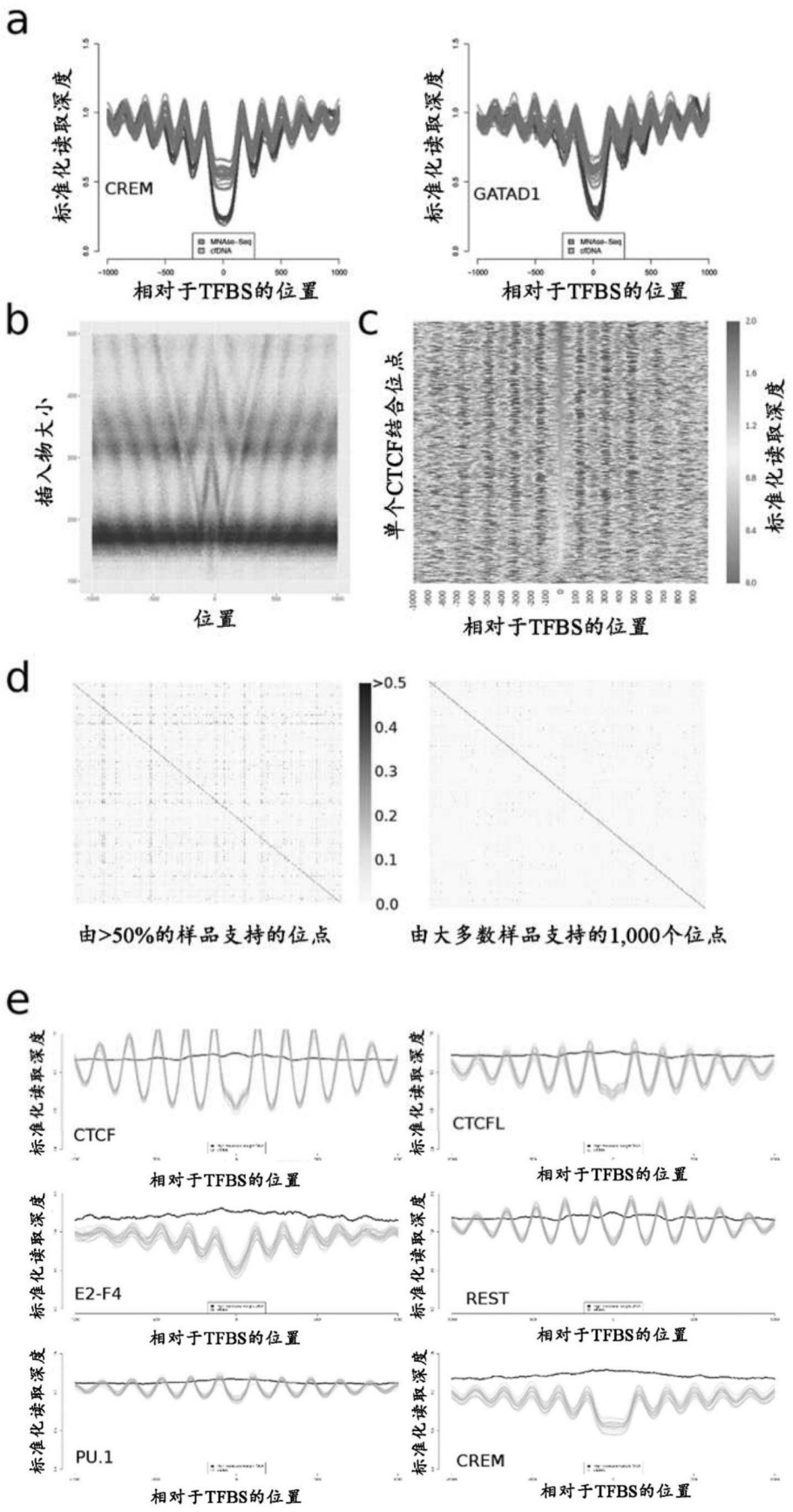

a
标准化读取深度
CREM
GATAD1
MNAse-Seq
cfDNA
相对于TFBS的位置
b
插入物大小
位置
c
单个CTCF结合位点
标准化读取深度
相对于TFBS的位置
d
由>50%的样品支持的位点
由大多数样品支持的1,000个位点
e
CTCF
CTCFL
E2-F4
REST
PU.1
CREM

图6A-6E

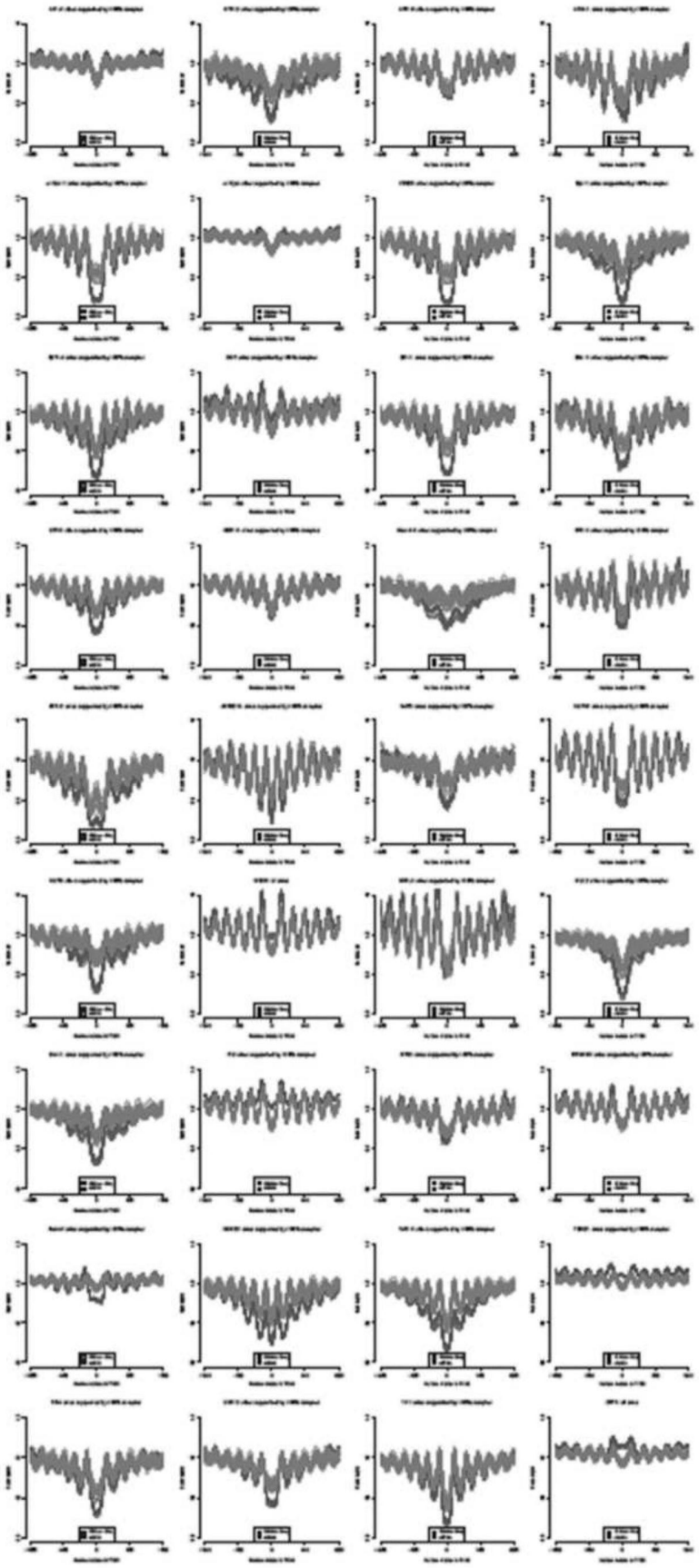


图7

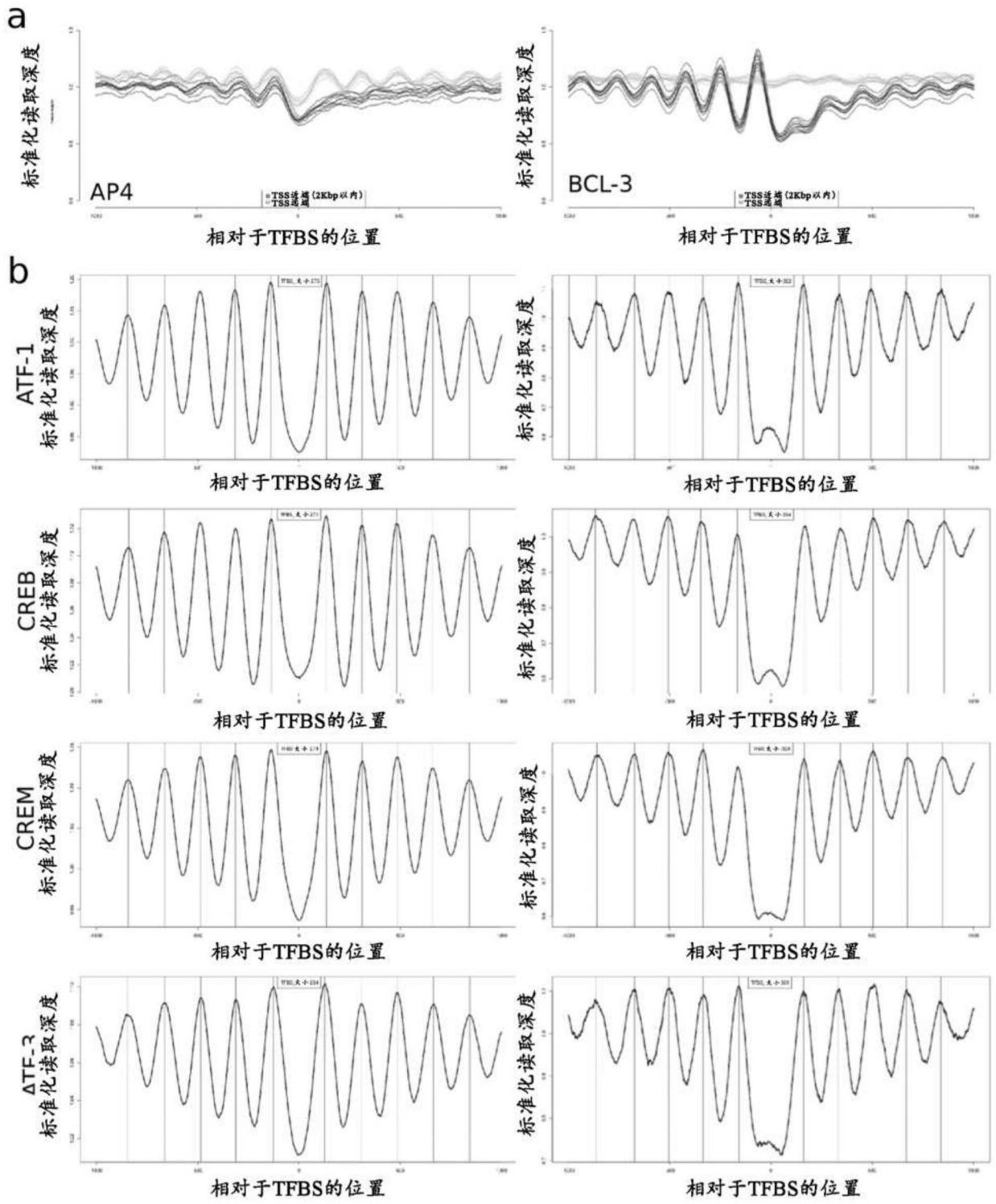

a
标准化读取深度
AP4
TSS远端（2Kbp以内）
TSS近端
相对于TFBS的位置
标准化读取深度
BCL-3
TSS远端（2Kbp以内）
TSS近端
相对于TFBS的位置
b
ATF-1
标准化读取深度
相对于TFBS的位置
标准化读取深度
相对于TFBS的位置
CREB
标准化读取深度
相对于TFBS的位置
标准化读取深度
相对于TFBS的位置
CREM
标准化读取深度
相对于TFBS的位置
标准化读取深度
相对于TFBS的位置
ΔTF-3
标准化读取深度
相对于TFBS的位置
标准化读取深度
相对于TFBS的位置

图8A-8B

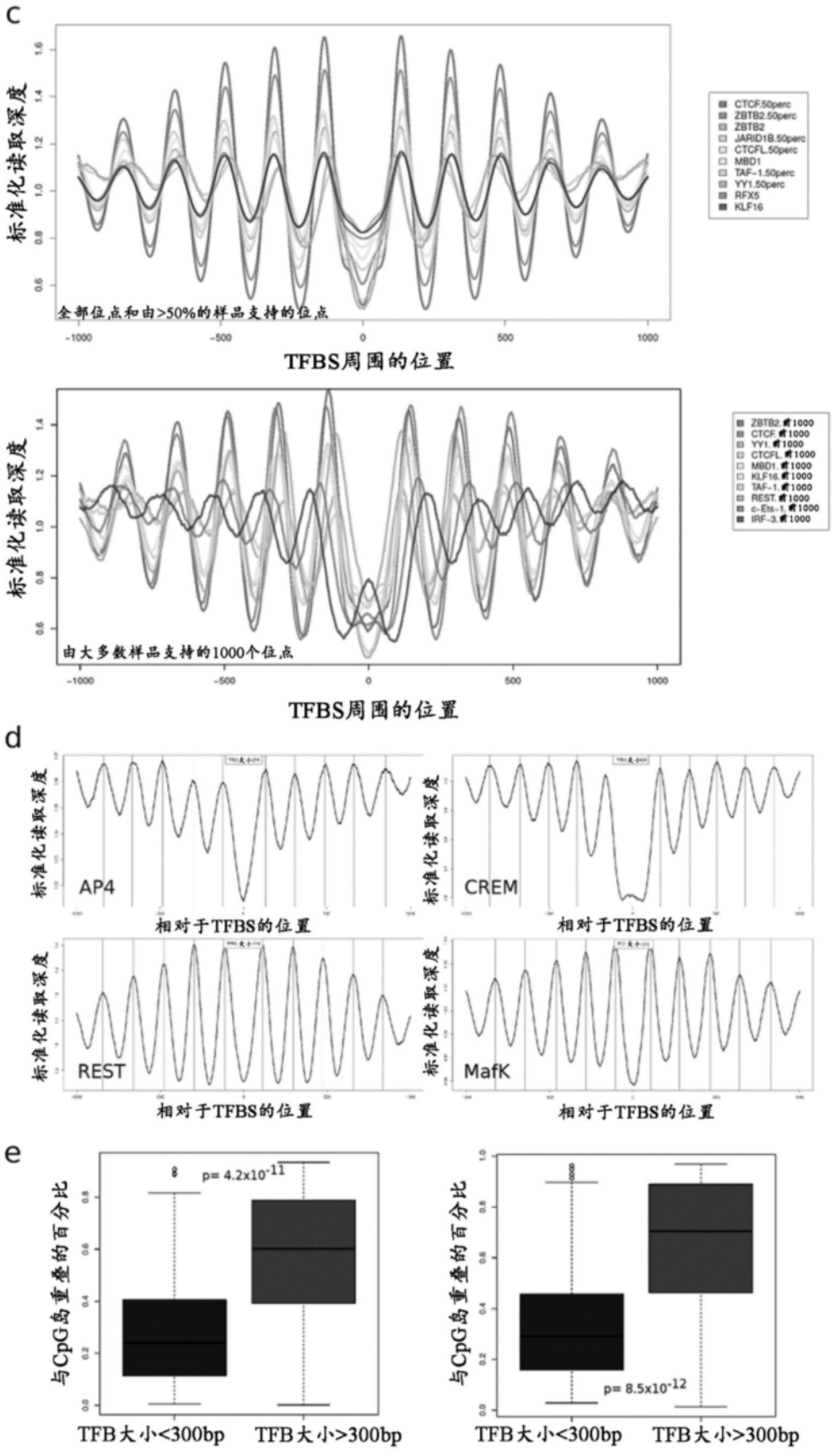

c
标准化读取深度
CTCF.50perc
ZBTB2.50perc
ZBTB2
JARID1B.50perc
CTCFL.50perc
MBD1
TAF-1.50perc
YY1.50perc
RFX5
KLF16
全部位点和由>50%的样品支持的位点
-1000
-500
0
500
1000
TFBS周围的位置
标准化读取深度
由大多数样品支持的1000个位点
TFBS周围的位置
d
AP4
CREM
REST
MafK
相对于TFBS的位置
e
与CpG岛重叠的百分比
p= 4.2x10-11
p= 8.5x10-12
TFB大小<300bp
TFB大小>300bp

图8C-8E

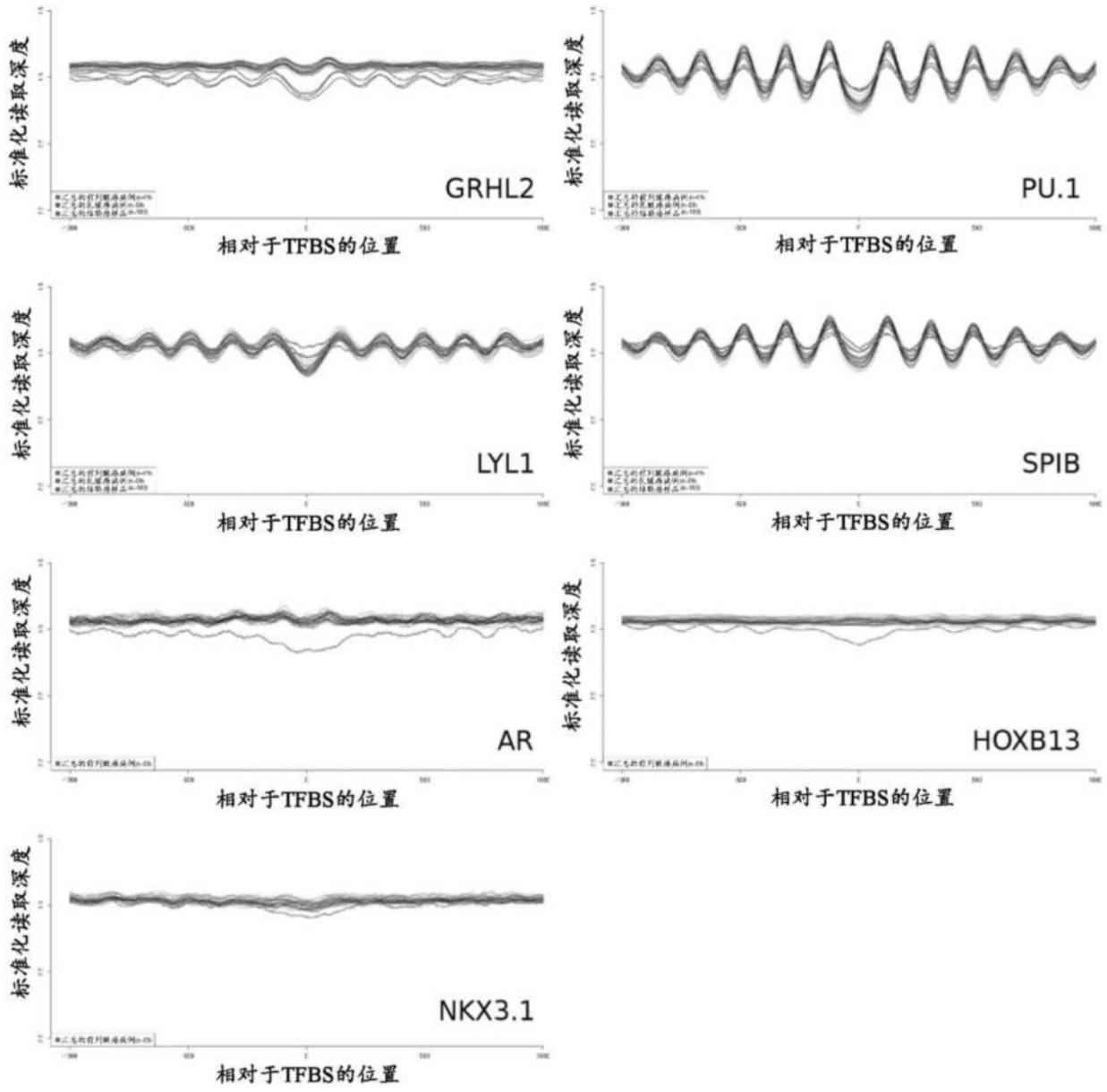

标准化读取深度
GRHL2
相对于TFBS的位置
标准化读取深度
PU.1
相对于TFBS的位置
标准化读取深度
LYL1
相对于TFBS的位置
标准化读取深度
SPIB
相对于TFBS的位置
标准化读取深度
AR
相对于TFBS的位置
标准化读取深度
HOXB13
相对于TFBS的位置
标准化读取深度
NKX3.1
相对于TFBS的位置

图9

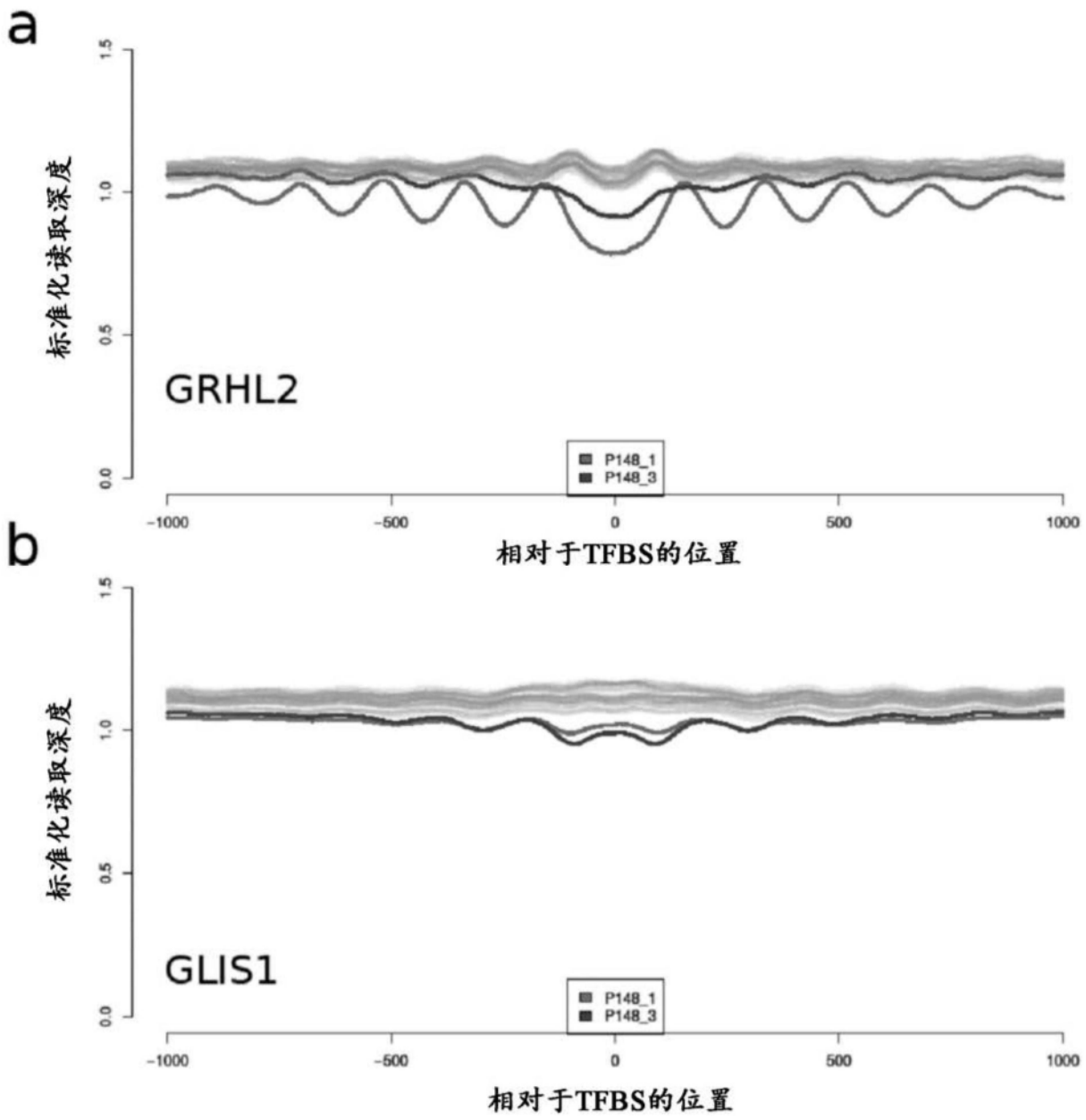

a
标准化读取深度
1.5
1.0
0.5
0.0
GRHL2
P148_1
P148_3
-1000
-500
0
500
1000
相对于TFBS的位置
b
标准化读取深度
1.5
1.0
0.5
0.0
GLIS1
P148_1
P148_3
-1000
-500
0
500
1000
相对于TFBS的位置

图10A-10B

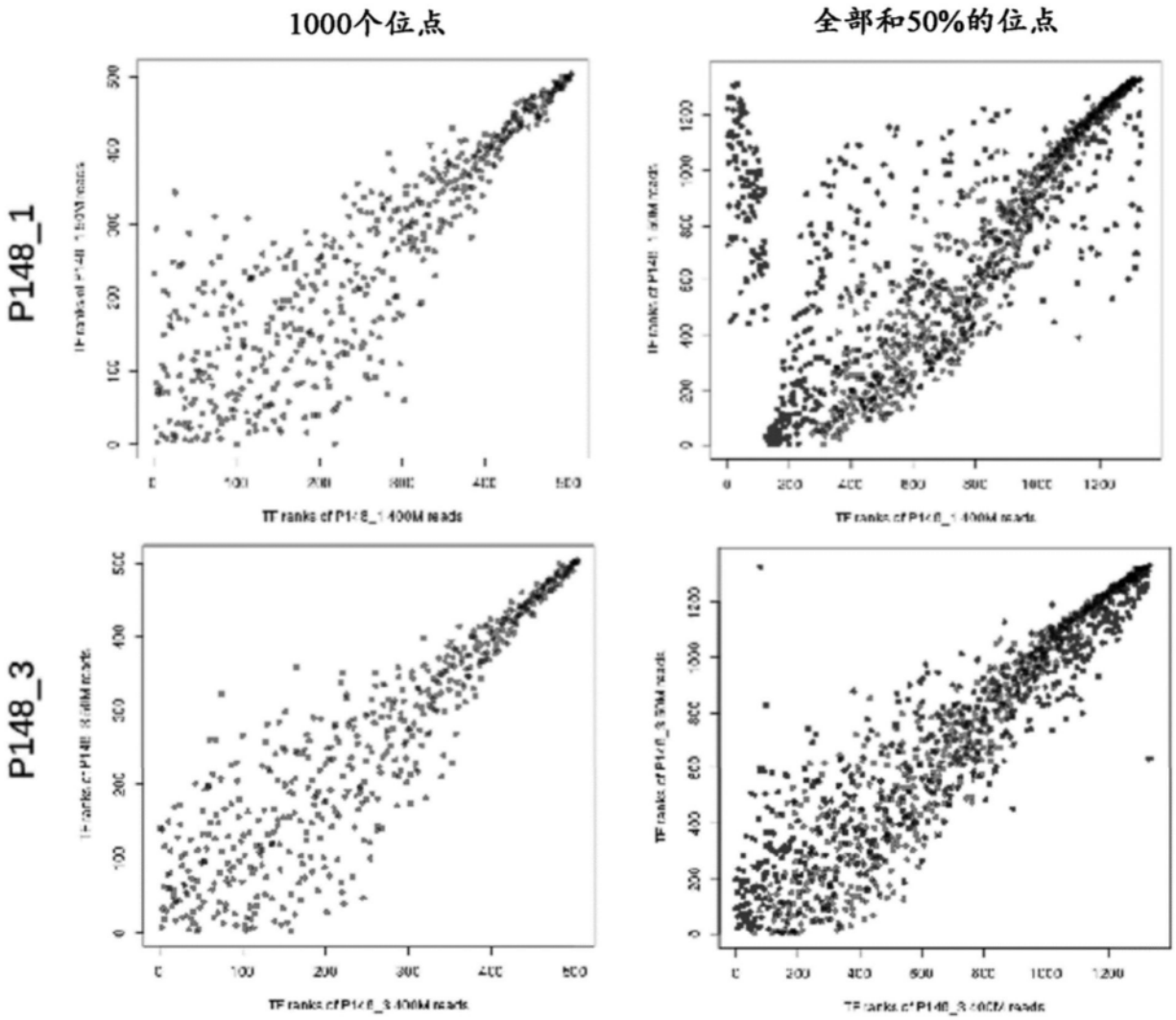


图11

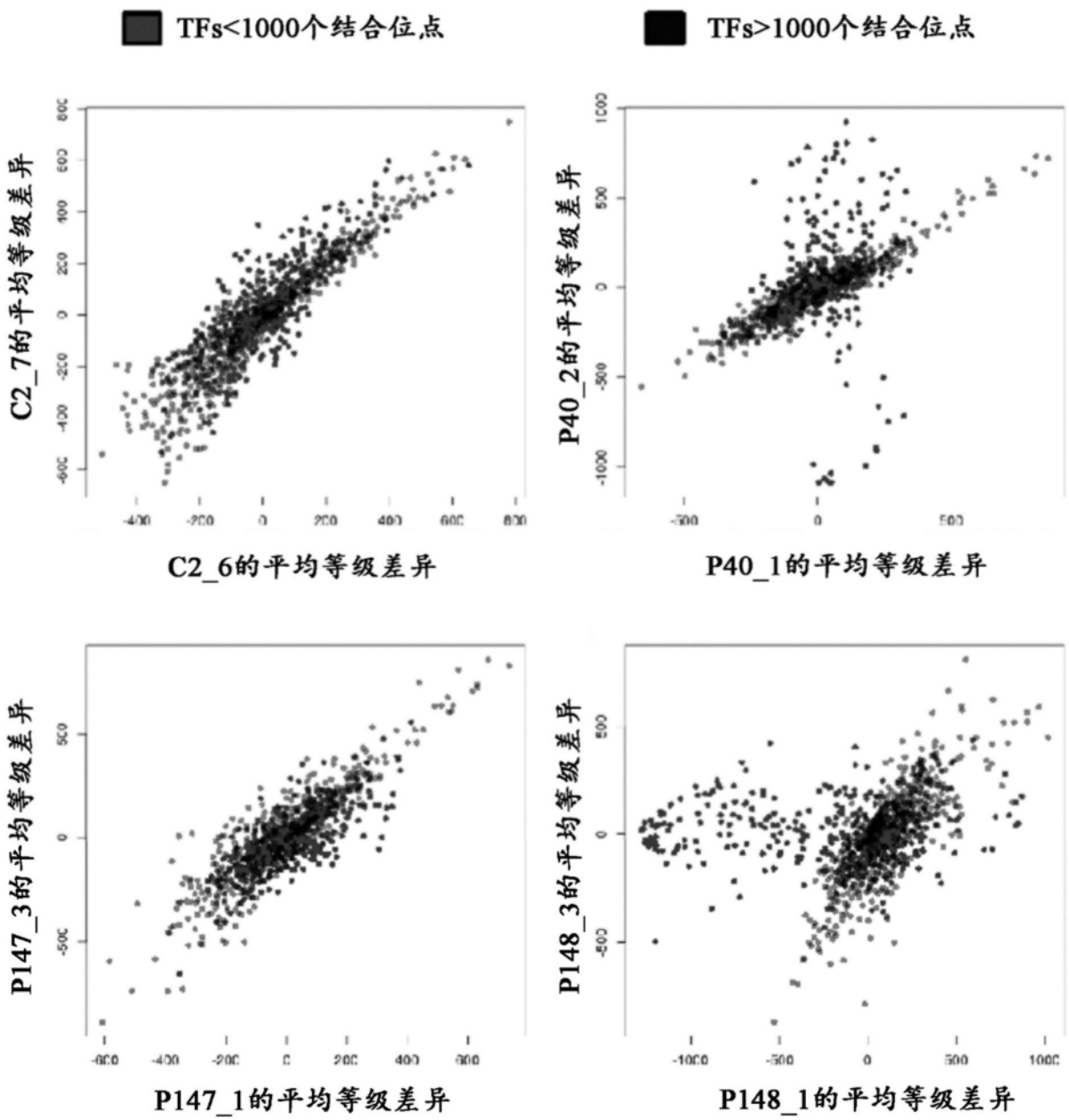

TFs<1000个结合位点
TFs>1000个结合位点
C2_7的平均等级差异
C2_6的平均等级差异
P40_2的平均等级差异
P40_1的平均等级差异
P147_3的平均等级差异
P147_1的平均等级差异
P148_3的平均等级差异
P148_1的平均等级差异

图12

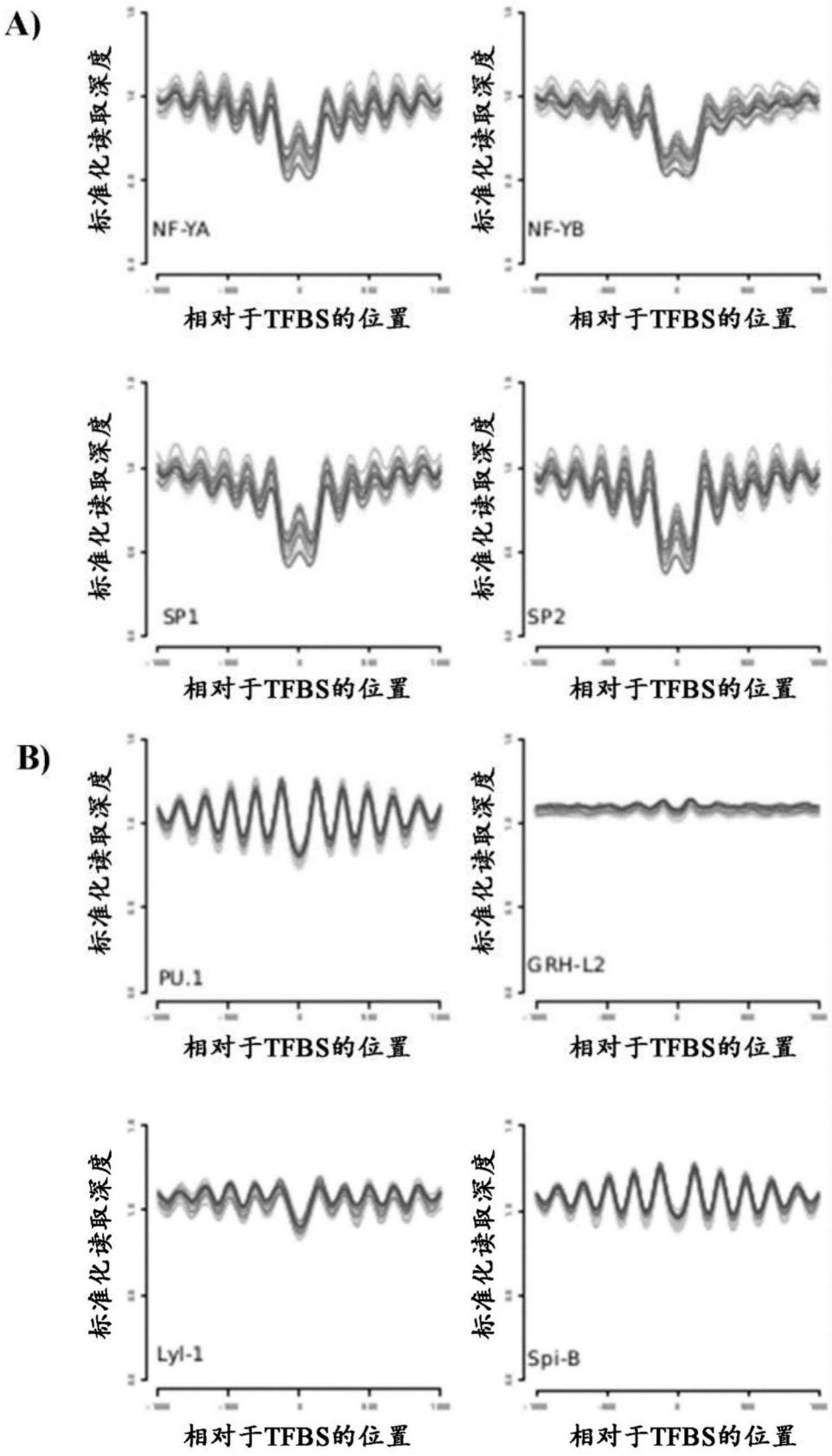

A)
标准化读取深度
NF-YA
相对于TFBS的位置
标准化读取深度
NF-YB
相对于TFBS的位置
标准化读取深度
SP1
相对于TFBS的位置
标准化读取深度
SP2
相对于TFBS的位置
B)
标准化读取深度
PU.1
相对于TFBS的位置
标准化读取深度
GRH-L2
相对于TFBS的位置
标准化读取深度
Lyl-1
相对于TFBS的位置
标准化读取深度
Spi-B
相对于TFBS的位置

图13A-13B

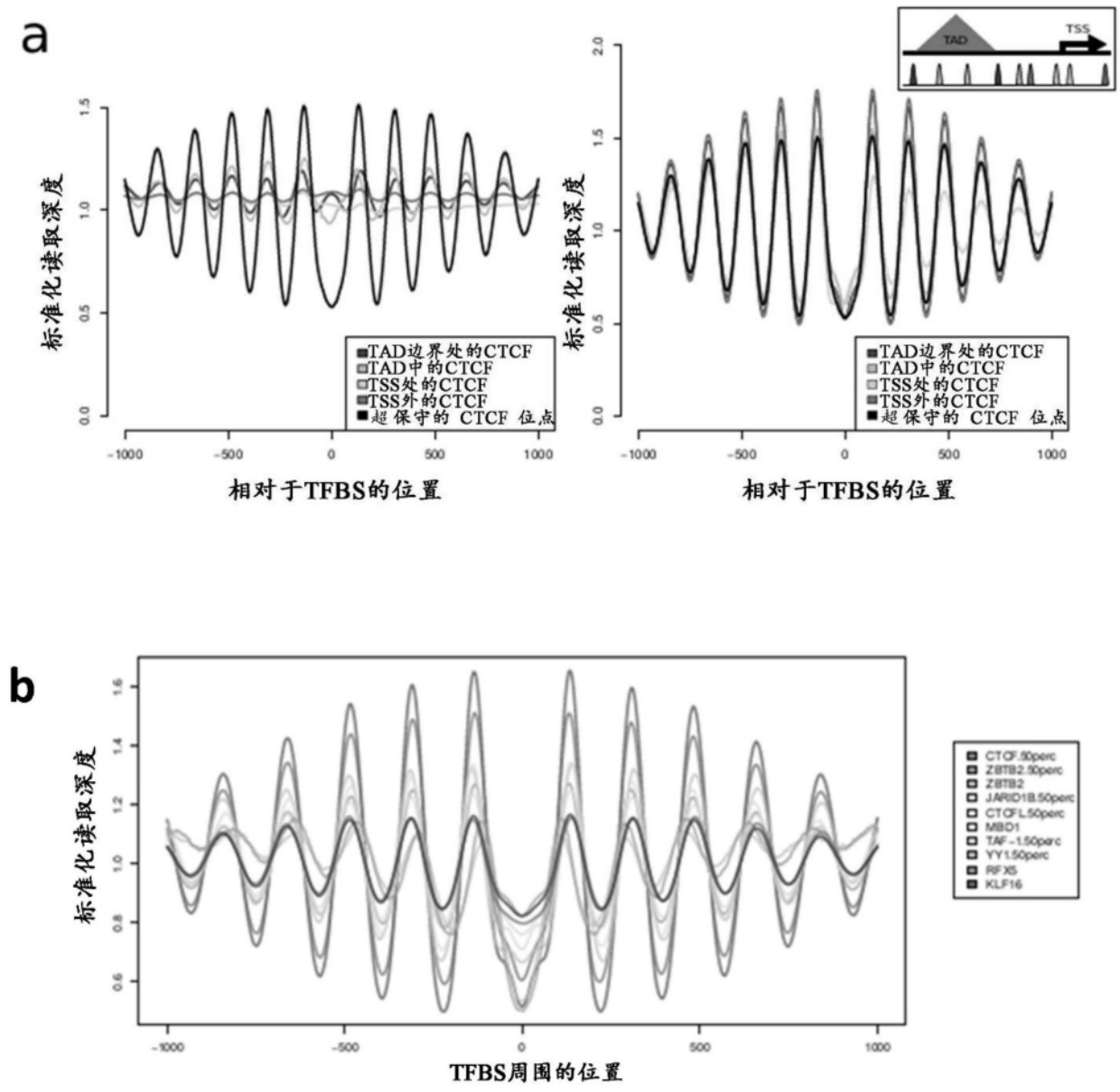
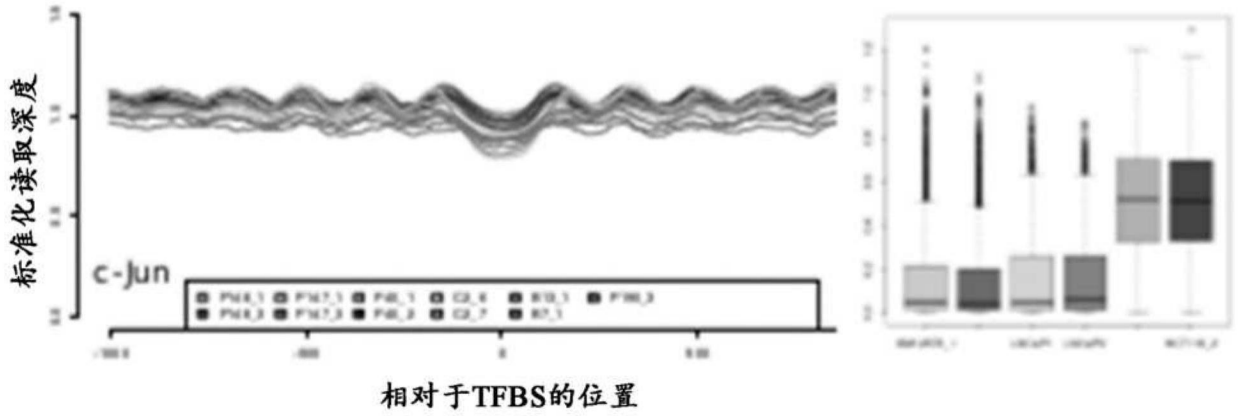

a
标准化读取深度
相对于TFBS的位置
TAD边界处的CTCF
TAD中的CTCF
TSS处的CTCF
TSS外的CTCF
超保守的 CTCF 位点
TAD
TSS
b
TFBS周围的位置
CTCF.50perc
ZBTB2.50perc
ZBTB2
JARID1B.50perc
CTCFL.50perc
MBD1
TAF-1.50perc
YY1.50perc
RFX5
KLF16

图14A-14B



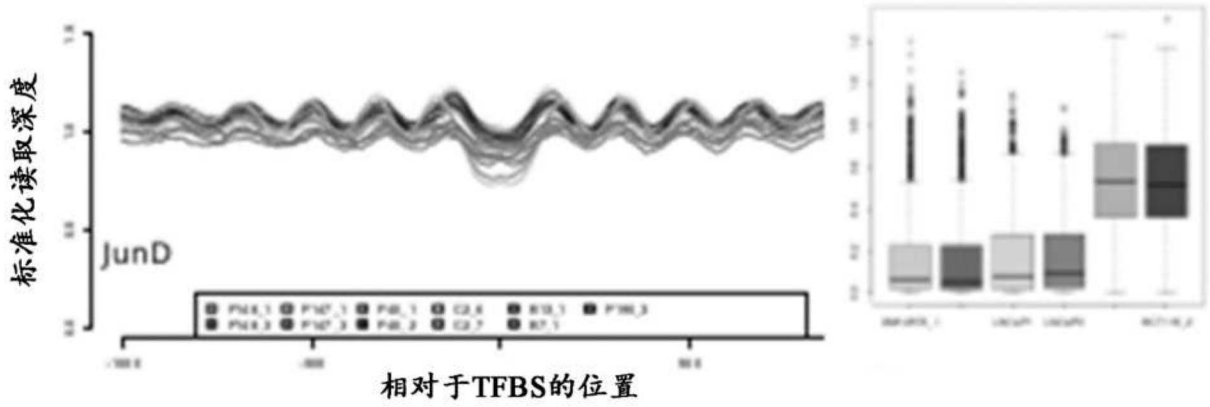


图15

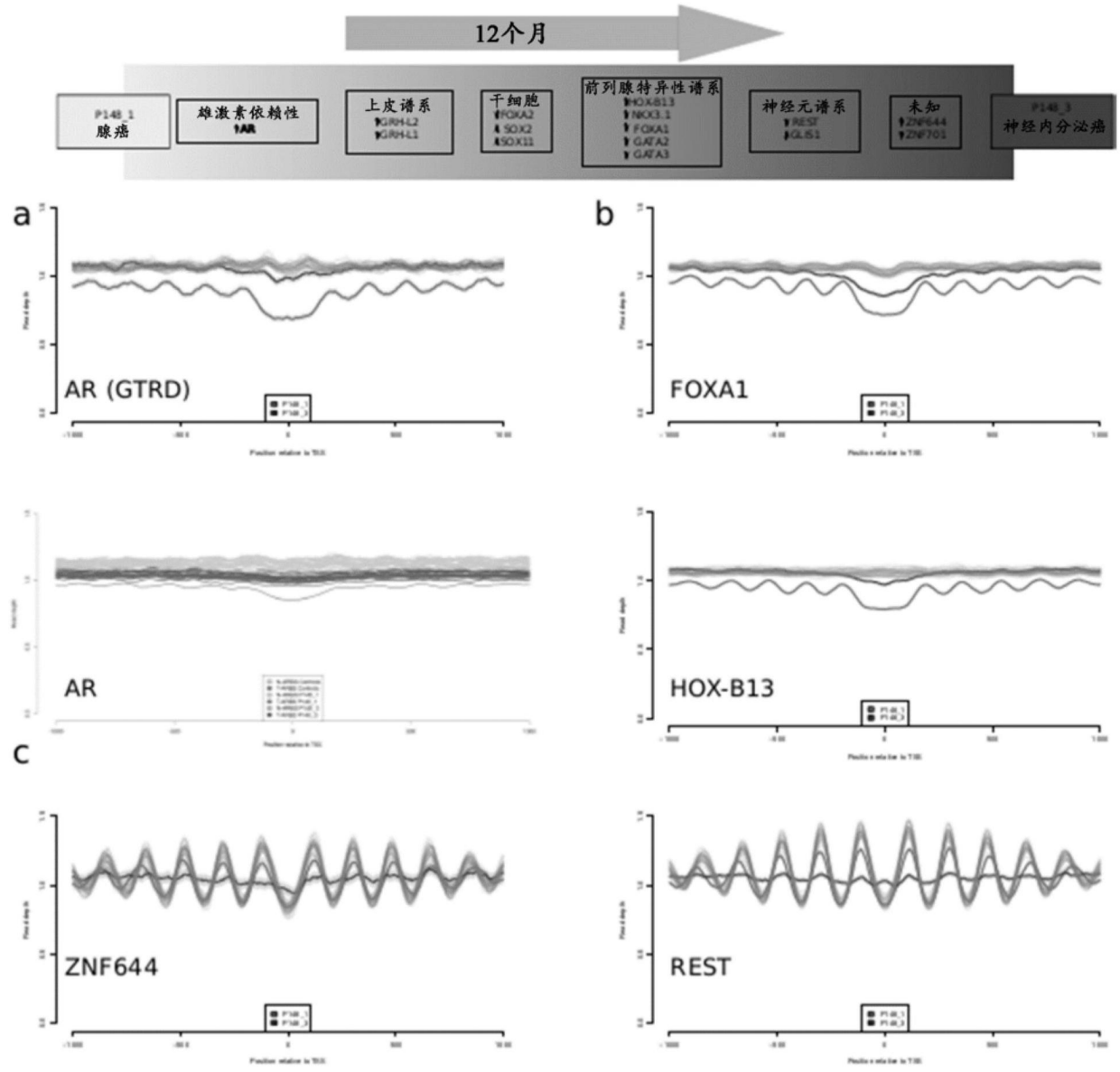

12个月
P148_1
腺癌
雄激素依赖性
↑AR
上皮谱系
↓GRHL2
↓GRHL1
干细胞
↑FOXA2
↓SOX2
↓SOX11
前列腺特异性谱系
↑HOXB13
↑NKX3.1
↑FOXA1
↑GATA2
↑GATA3
神经元谱系
↑REST
↓GLIS1
未知
↑ZNF644
↑ZNF701
P148_3
神经内分泌癌
a
AR (GTRD)
b
FOXA1
AR
HOX-B13
c
ZNF644
REST

图16A-16C

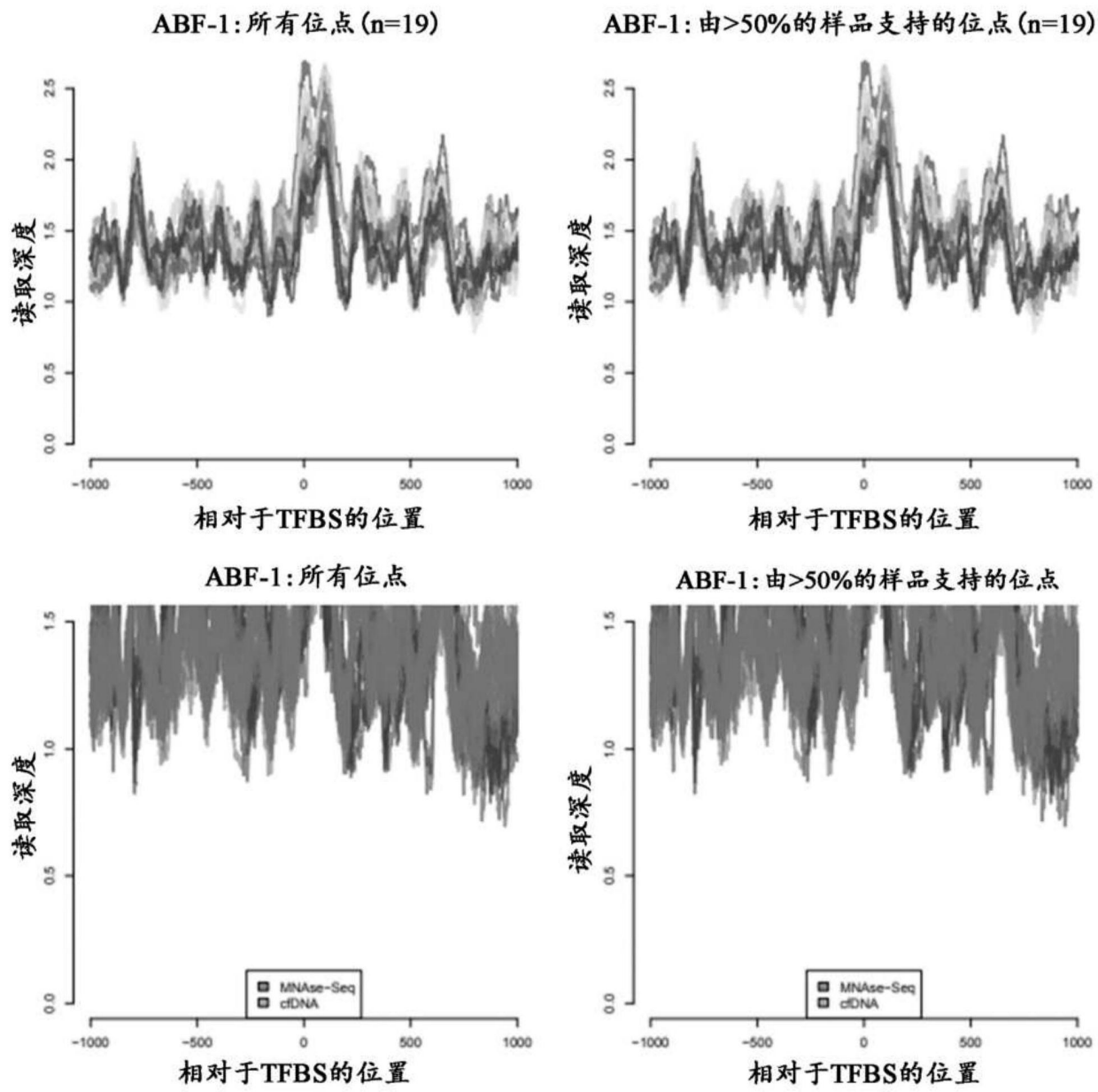

ABF-1:所有位点(n=19)
ABF-1:由>50%的样品支持的位点(n=19)
读取深度
相对于TFBS的位置
ABF-1:所有位点
ABF-1:由>50%的样品支持的位点
MNAse-Seq
cfDNA

图17

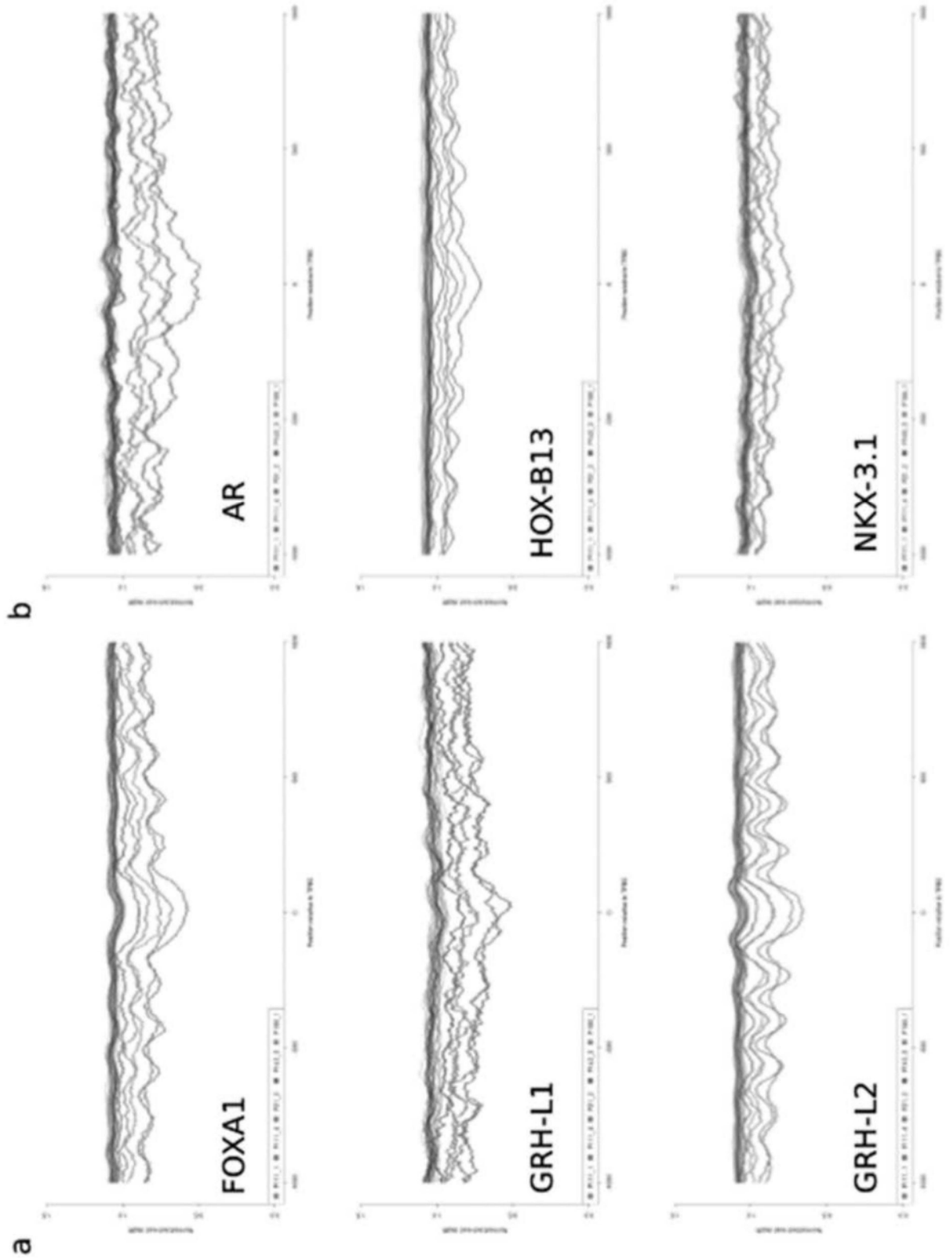


图18A-18B

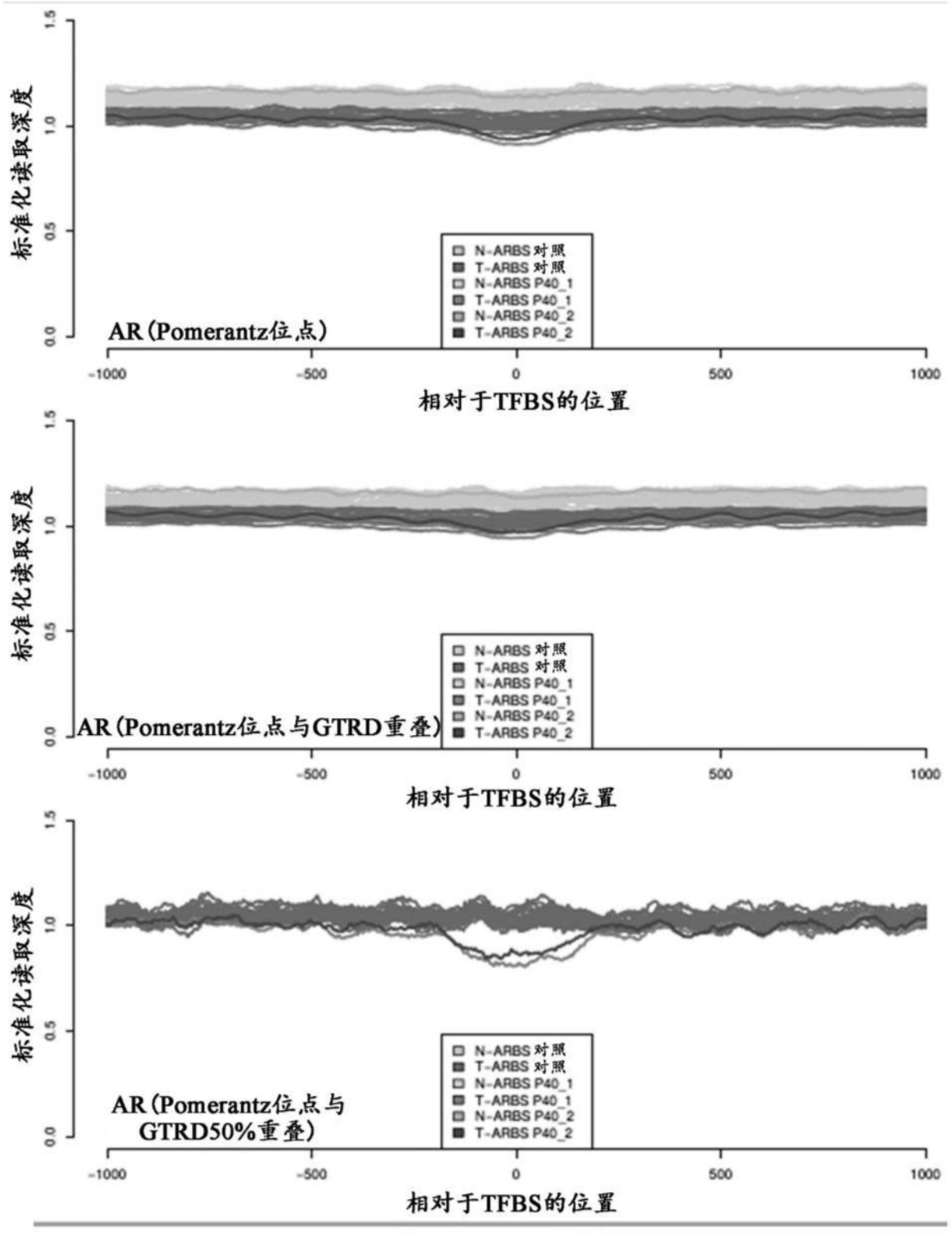
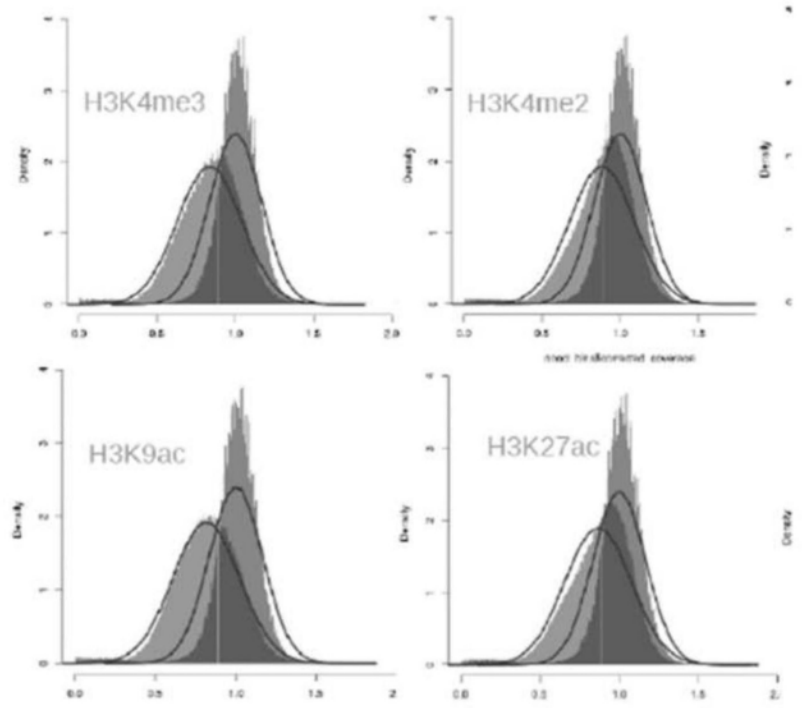

标准化读取深度
1.5
1.0
0.5
0.0
N-ARBS 对照
T-ARBS 对照
N-ARBS P40_1
T-ARBS P40_1
N-ARBS P40_2
T-ARBS P40_2
AR (Pomerantz位点)
-1000
-500
0
500
1000
相对于TFBS的位置
AR (Pomerantz位点与GTRD重叠)
AR (Pomerantz位点与
GTRD50%重叠)

图19


H3K4me3
H3K4me2
H3K9ac
H3K27ac
组蛋白修饰

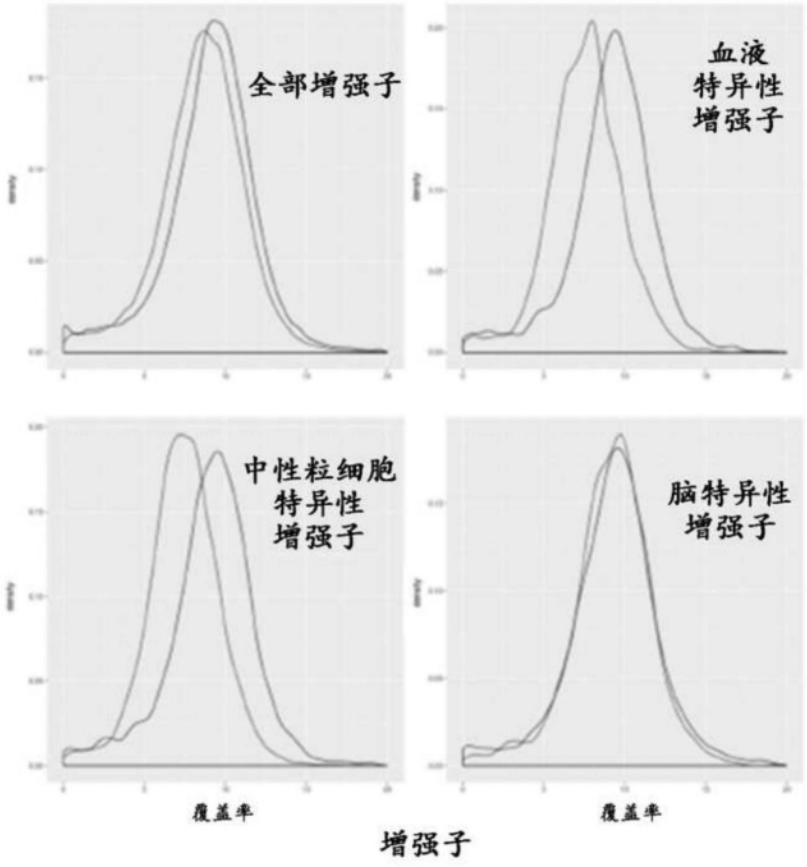

全部增强子
血液
特异性
增强子
中性粒细胞
特异性
增强子
脑特异性
增强子
覆盖率
覆盖率
增强子

图20

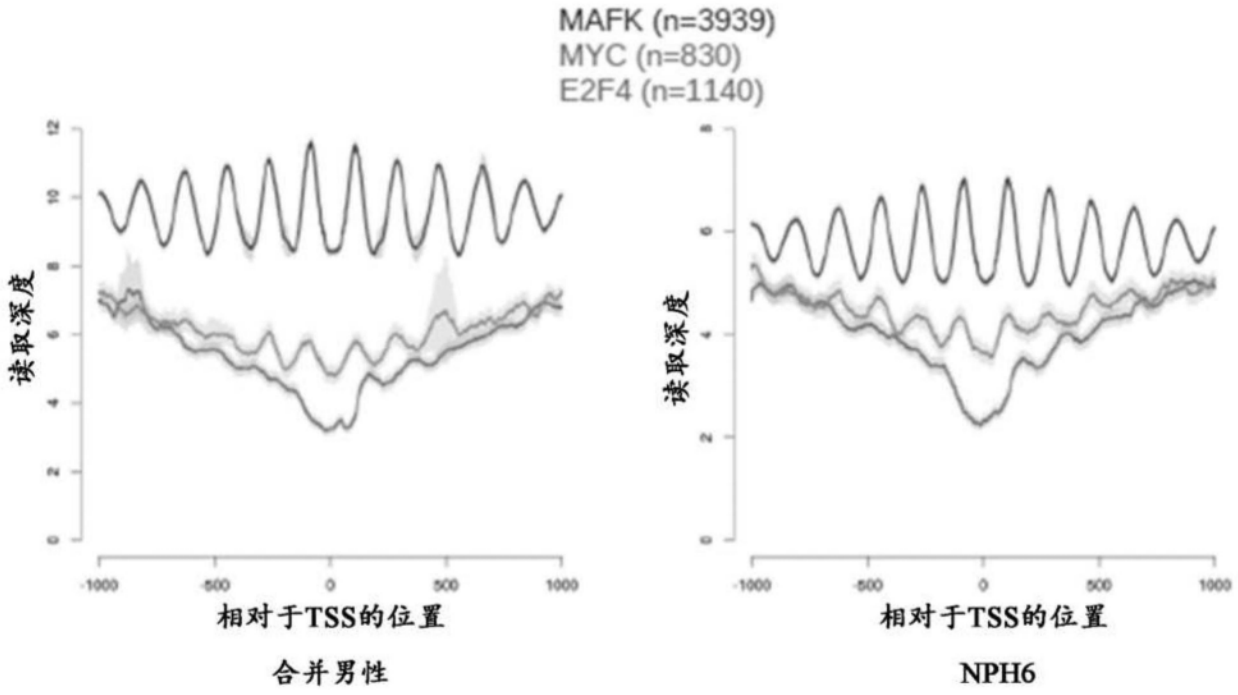

MAFK (n=3939)
MYC (n=830)
E2F4 (n=1140)
读取深度
相对于TSS的位置
合并男性
读取深度
相对于TSS的位置
NPH6

图21

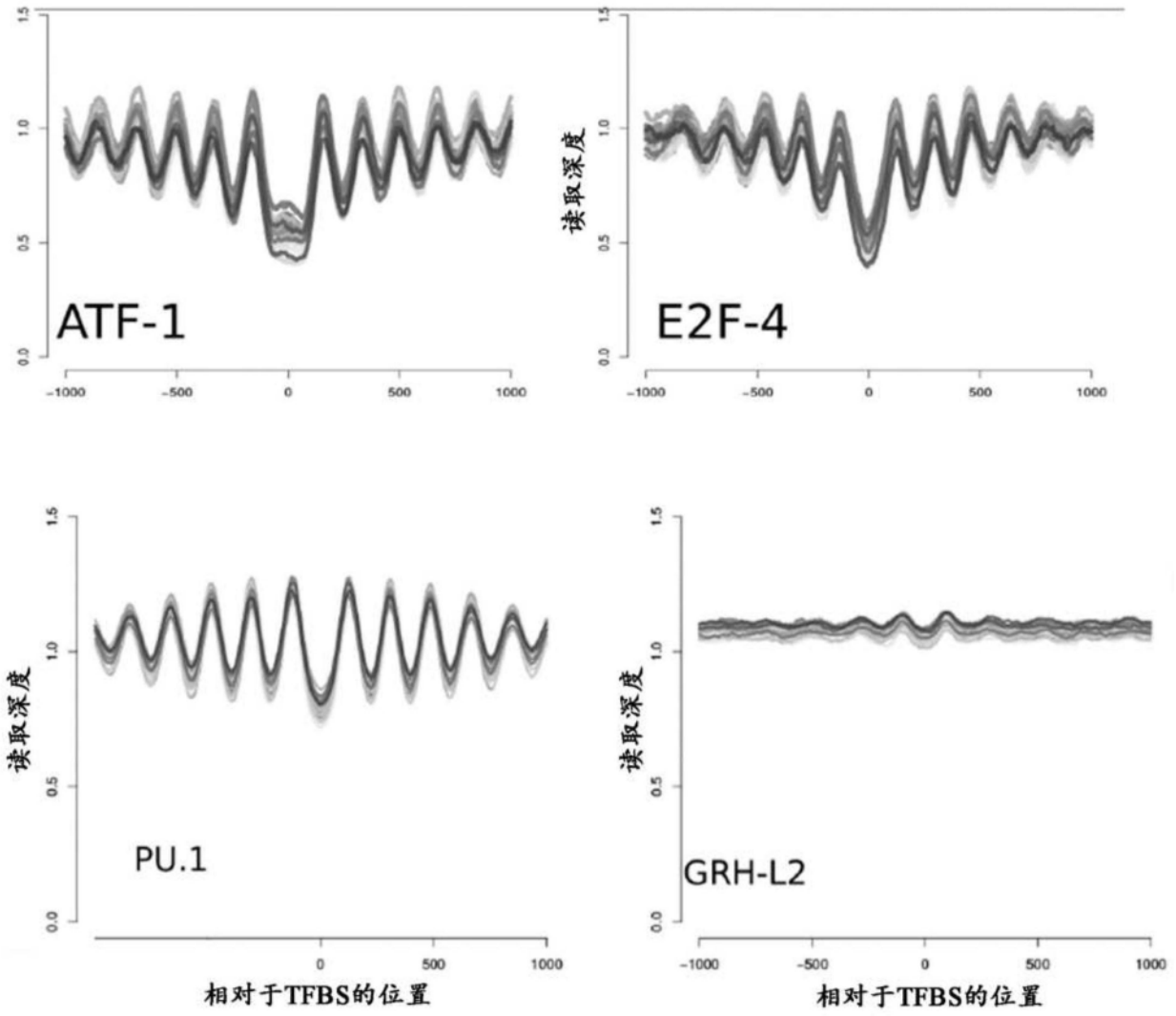


图22

TF1	TF2	重叠 (TF1的部分)
USF-2	USF-1	0.98
ZNF143	RBAK	0.95
REST	MIER2	0.94
AP-2A	AP-2C	0.92
ZNF75A	GABPA	0.91
MafK	MafF	0.87
MLL2	MYST1	0.86
JunD	Fra-2	0.86
HHEX	FOXA3	0.86
ZNF238	ZBTB42	0.85
CTCF	ZBTB2	0.85
SIX5	RBAK	0.84
ZNF143	ZNF76	0.84
c-Myc	Max	0.83
RERE	FOXA3	0.82
FOXA2	FOXA3	0.81
c-Jun	JunB	0.81
HMG20B	FOXA3	0.80
N-Myc	Max	0.78
HHEX	RERE	0.78

图23

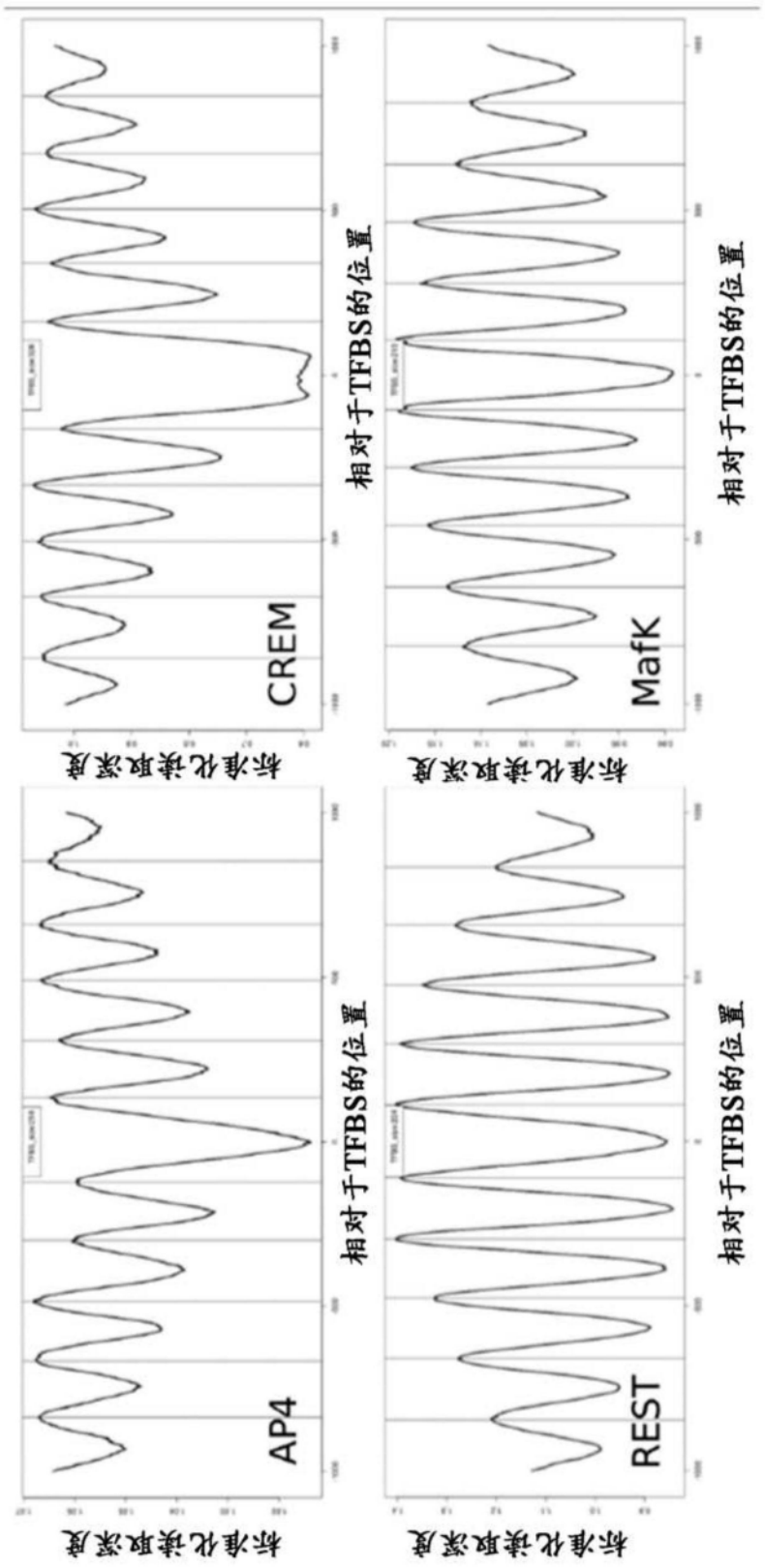


图24

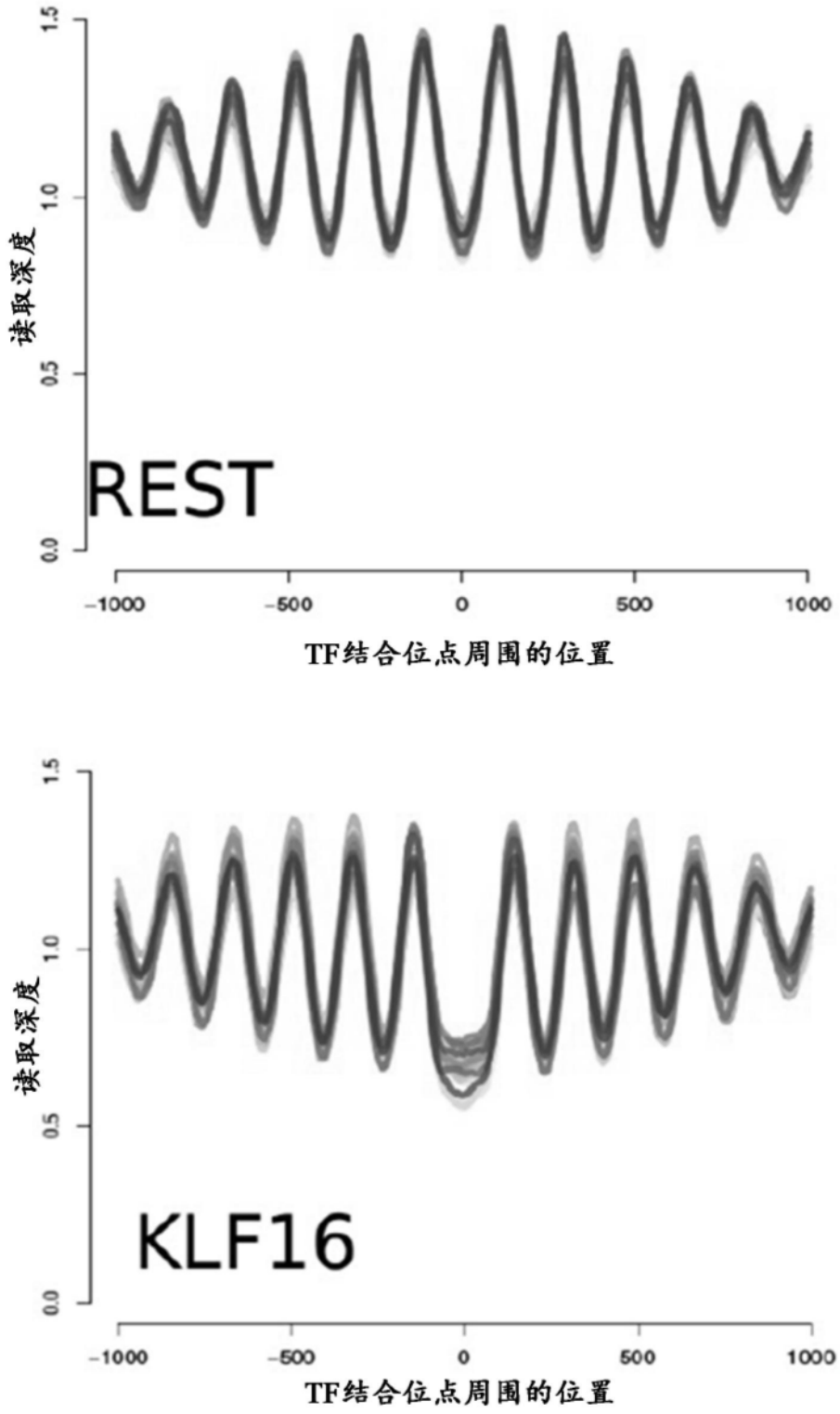

读取深度
0.0
0.5
1.0
1.5
REST
-1000
-500
0
500
1000
TF结合位点周围的位置
读取深度
0.0
0.5
1.0
1.5
KLF16
-1000
-500
0
500
1000
TF结合位点周围的位置

图25

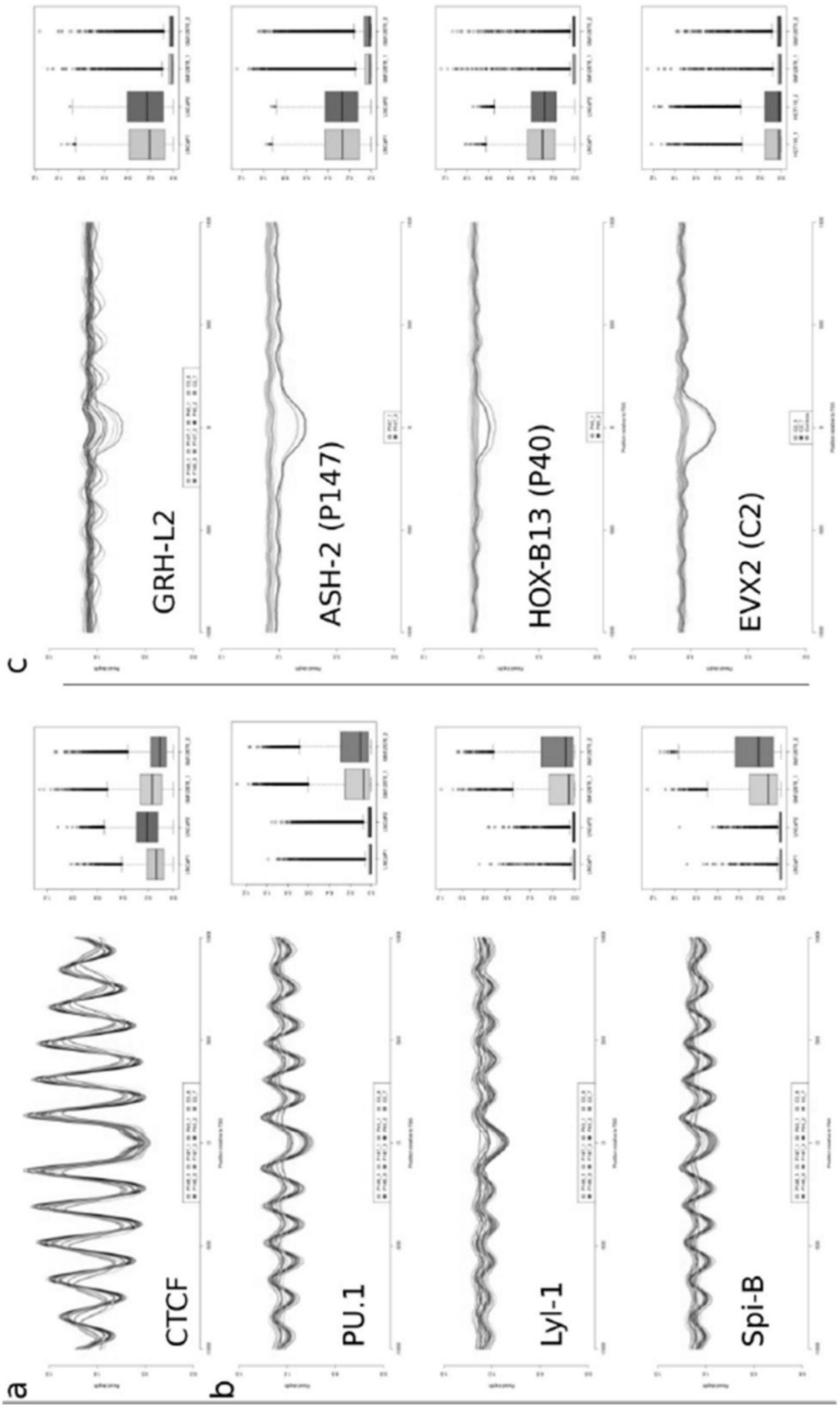
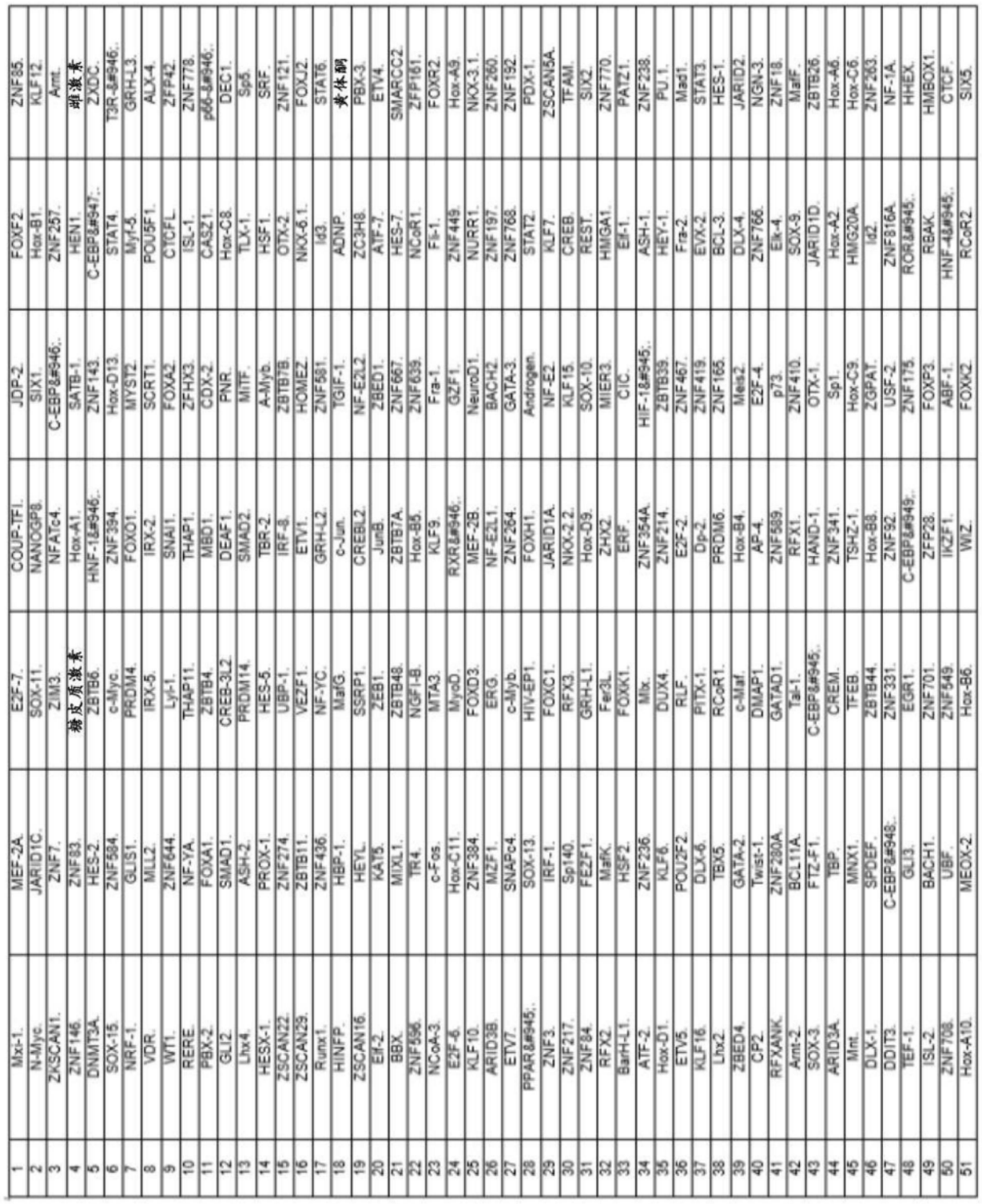


图26A-26C



	Mxi-1.	MEF-2A.	E2F-7.	COUP-TFI.	JDP-2.	FOXF2.	ZNF85.
1	Mxi-1.	MEF-2A.	E2F-7.	COUP-TFI.	JDP-2.	FOXF2.	ZNF85.
2	N-Myc.	JARID1C.	SOX-11.	NANOGP8.	SIX1.	Hox-B1.	KLF12.
3	ZKSCAN1.	ZNF7.	ZIM3.	NFATc4.	C-EBPβ.	ZNF257.	Amt.
4	ZNF146.	ZNF83.	绿皮质激素	Hox-A1.	SATB-1.	HEN1.	雌激素
5	DNMT3A.	HES-2.	ZBTB6.	HNF-1β.	ZNF143.	C-EBPγ.	ZXDC.
6	SOX-15.	ZNF584.	c-Myc.	ZNF394.	Hox-D13.	STAT4.	T3R-β
7	NRF-1.	GLIS1.	PRDM4.	FOXO1.	MYST2.	Myf-5.	GRH-L3.
8	VDR.	MLL2.	IRX-5.	IRX-2.	SCRT1.	POU5F1.	ALX-4.
9	Wt1.	ZNF644.	Lyl-1.	SNAI1.	FOXA2.	CTCFL.	ZFP42.
10	RERE.	NF-YA.	THAP11.	THAP1.	ZFHX3.	ISL-1.	ZNF778.
11	PBX-2.	FOXA1.	ZBTB4.	MBD1.	CDX-2.	CAS21.	p66-β
12	GLI2.	SMAD1.	CREB-3L2.	DEAF1.	PNR.	Hox-C8.	DEC1.
13	Lhx4.	ASH-2.	PRDM14.	SMAD2.	MITF.	TLX-1.	Sp5.
14	HESX-1.	PROX-1.	HES-5.	TBR-2.	A-Myb.	HSF1.	SRF.
15	ZSCAN22.	ZNF274.	UBP-1.	IRF-8.	ZBTB7B	OTX-2.	ZNF121.
16	ZSCAN29.	ZBTB11.	VEZF1.	ETV1.	HOMEZ.	NKX-6.1.	FOXJ2.
17	Runx1.	ZNF436.	NF-YC.	GRH-L2.	ZNF581.	Id3.	STAT6.
18	HINFP.	HBP-1.	Maf/G.	c-Jun.	TGIF-1.	ADNP.	类休酮
19	ZSCAN16.	HEYL.	SSRP1.	CREBL2.	NF-E2L2.	ZC3H8.	PBX-3.
20	Elf-2.	KAT5.	ZEB1.	JunB.	ZBED1.	ATF-7.	ETV4.
21	BBX.	MIXL1.	ZBTB48.	ZBTB7A.	ZNF667.	HES-7.	SMARCC2.
22	ZNF586.	TR4.	NGFI-B.	Hox-B5.	ZNF639.	NCoR1.	ZFP161.
23	NCoA-3.	c-Fos.	MTA3.	KLF9.	Fra-1.	Fli-1.	FOXR2.
24	E2F-6.	Hox-C11.	MyoD.	RXRβ	GZF1.	ZNF449.	Hox-A9.
25	KLF10.	ZNF384.	FOXD3.	MEF-2B.	NeuroD1.	NURR1.	NKX-3.1.
26	ARID3B.	MZF1.	ERG.	NF-E2L1.	BACH2.	ZNF197.	ZNF260.
27	ETV7.	SNAPc4.	c-Myb.	ZNF264.	GATA-3.	ZNF768.	ZNF192.
28	PPARα	SOX-13.	HIV-EP1.	FOXH1.	Androgen.	STAT2.	PDX-1.
29	ZNF3.	IRF-1.	FOXC1.	JARID1A.	NF-E2.	KLF7.	ZSCAN5A.
30	ZNF217.	Sp140.	RFX3.	NKX-2.2.	KLF15.	CREB.	TFAM.
31	ZNF84.	FEZF1.	GRH-L1.	Hox-D9.	SOX-10.	REST.	SIX2.
32	RFX2.	MafK.	Fer3L.	ZHX2.	MIER3.	HMGA1.	ZNF770.
33	BarH-L1.	HSF2.	FOXK1.	ERF.	CIC.	Elf-1.	PATZ1.
34	ATF-2.	ZNF236.	Mlx.	ZNF354A.	HIF-1α	ASH-1.	ZNF238.
35	Hox-D1.	KLF6.	DUX4.	ZNF214.	ZBTB39.	HEY-1.	PU.1.
36	ETV5.	POU2F2.	RLF.	E2F-2.	ZNF467.	Fra-2.	Mad1.
37	KLF16.	DLX-6.	PITX-1.	Dp-2.	ZNF419.	EVX-2.	STAT3.
38	Lhx2.	TBX5.	RCoR1.	PRDM6.	ZNF165.	BCL-3.	HES-1.
39	ZBED4.	GATA-2.	c-Maf.	Hox-B4.	Meis2.	DLX-4.	JARID2.
40	CP2.	Twist-1.	DMAP1.	AP-4.	E2F-4.	ZNF766.	NGN-3.
41	RFXANK.	ZNF280A.	GATAD1.	ZNF589.	p73.	Elk-4.	ZNF18.
42	Amt-2.	BCL11A.	Tal-1.	RFX1.	ZNF410.	SOX-9.	MafF.
43	SOX-3.	FTZ-F1.	C-EBPα	HAND-1.	OTX-1.	JARID1D.	ZBTB26.
44	ARID3A.	TBP.	CREM.	ZNF341.	Sp1.	Hox-A2.	Hox-A6.
45	Mnt.	MNX1.	TFEB.	TSHZ-1.	Hox-C9.	HMG20A.	Hox-C6.
46	DLX-1.	SPDEF	ZBTB44.	Hox-B8.	ZGPAT.	Id2.	ZNF263.
47	DDIT3.	C-EBPδ	ZNF331.	ZNF92.	USF-2.	ZNF816A.	NF-1A.
48	TEF-1.	GLI3.	EGR1.	C-EBPε	ZNF175.	RORα	HHEX.
49	ISL-2.	BACH1.	ZNF701.	ZFP28.	FOXP3.	RBAK.	HMBOX1.
50	ZNF708.	UBF.	ZNF549.	IKZF1.	ABF-1.	HNF-4α	CTCF.
51	Hox-A10.	MEOX-2.	Hox-B6.	WIZ.	FOXK2.	RCoR2.	SIX5.

图27

52	ZNF281.	ZNF382.	ZFP64.	ZNF22.	Pax-2.	Hox-A7.	ZHX1.
53	LRH-1.	ZNF680.	Lhx6.	IRF-9.	GATA-6.	GLI1.	IRF-2.
54	IRX-3.	TBX21.	MYNN.	ZNF574.	NKX-2.3.	GMEB-1.	E2F-8.
55	SOX-17.	SMARCA1.	DBP.	HIV-EP3.	Hox-B7.	NF-κB.	KLF13.
56	USF-1.	Id1.	ZNF250.	HNF-4γ.	ZNF407.	Lhx3.	ZNF184.
57	TEF-3.	B-Myb.	ZNF329.	FOXN3.	Elf-5.	B-ATF.	ZBTB42.
58	BARX-1.	Hox-B13.	WHSC1.	HLF.	HIF-3α.	AhR.	TCFL5.
59	FOXM1.	JunD.	FOXG1.	ZNF511.	GLIS2.	NANOG.	TSC22D4.
60	DEC2.	MLL.	NFAT5.	ZBTB17.	SNAI2.	T.	FOXO4.
61	EVX-1.	肌细胞生成素	Elk-1.	TBP-L1.	OVOL2.	Hox-D4.	NKX-2.8.
62	RAR-α.	SMARCA5.	Sp2.	ZNF193.	Id4.	TOX-4.	SFPQ.
63	GATA-4.	TBX3.	SOX-5.	EPAS-1.	ZNF224.	ATF-3.	ZNF418.
64	ZNF266.	Sp3.	EMX-1.	Mad3.	TEF-5.	AmtL2.	FOXQ1.
65	EGR2.	SOX-2	Dp-1.	AmtL.	MSX-1.	FosB.	E2F-1.
66	ZNF8.	Sp-B.	GLI4.	SMARCC1.	STAT5B.	PPARβ.	ARID1B.
67	Elf-3.	TAF-1.	ARID1A.	NCoA-1.	ZNF597.	GFI1B.	PBX-1.
68	Hox-B2.	ZNF595.	BCL11B.	EAR2.	EBF1.	AEBP2.	DMRT1.
69	B-ATF-3.	ZNF335.	ZFX.	IRF-3.	SALL4.	ZIC2.	ZNF37A.
70	AP-2α.	SMAD4.	ZNF350.	ZNF774.	TFE3.	IRF-4.	NFATc3.
71	ATF-α.	SREBP-2.	MA2.	ZFP82.	CLOCK.	FOXO2.	NF-IL3.
72	EBF3.	OSR2.	ZNF48.	GLIS3.	ZNF136.	ZNF280D.	ZNF76.
73	ZNF580.	ZNF41.	Pax-5.	SMAD3.	FOXP1.	ARID2.	Hox-A4.
74	ZFP1.	RXRα.	Myf-6.	p53.	ZNF792.	NFATc1.	CREB-3L1.
75	TGIF-2.	HSF4.	HMG20B.	ZNF134.	GTF3A.	TEF-4.	Lhx5.
76	TCF-7.	ZSCAN2.	ERR1.	POU3F2.	NF-E2L3.	ZNF711.	Lmx-1β.
77	HNF-1β.	RFX5.	ATH-1.	ATF-5.	MIER2.	ZNF490.	NGN-2.
78	Pax-6.	HES-4.	ZNF323.	PITX-3.	PPARγ.	ARID4B.	
79	ATRX.	Cux-1.	PRDM1.	KLF1.	RBP-Jκ.	NCoR2.	
80	FOXA3.	ZNF528.	DbpA.	GABPα.	ZNF554.	TCF-7L2.	
81	NF-1C.	NKX-2.1.	ZNF320.	ZNF317.	RelB.	KLF3.	
82	IKZF5.	EHF.	FOXP2.	ZNF669.	MEF-2C.	KLF4.	
83	Hox-C13.	SMARCE1.	MEF-2D.	Mad4.	EGR3.	PLAG1.	
84	DMRT2.	ZNF586.	c-Ets-2.	EVI-1.	ZNF618.	ZSCAN5D.	
85	CDX-1.	ZNF740.	HTF-4.	Hox-D11.	SMAD5.	DNMT3B.	
86	ZNF784.	IRF-5.	MSX-2.	FOXO3.	ZBTB16.	ZNF140.	
87	STAT1.	ETV6.	AP-2γ.	FOXF1.	GATA-1.	CREB-3L4.	
88	NF-YB.	JARID1B.	NF-1B.	HAND-2.	GMEB-2.	Runx2.	
89	E2F-3.	MafB.	ZNF30.	Runx3.	LXRβ.	ZNF454.	
90	POU2F1.	p63.	ATF-1.	NONO.	SOX-4.	PHOX-2B.	
91	ZFAT.	ZNF34.	BARX-2.	p66-β.	c-Ets-1.	ZBTB12.	
92	ZNF322A.	盐皮质激素	ZNF324A.	KLF11.	Hox-A11.	Max.	
93	E2A.	Pax-8.	ZBTB2.	Meis1.	PXR.	Elf-4.	
94	ZNF652.	ZNF35.	ZNF513.	Rev-ErbAα.	TCF-7L1.	STAT5A.	
95	E2F-5.	Hox-A5.	Sp4.	RFX7.	NeuroD2.	ZNF75A.	
96	LEF-1.	MYST1.	SEF2.	ZNF582.	FEV.	YY1.	
97	THAP4.	ZBTB24.	BCL6.	ZNF563.	ZSCAN5B.	Pax-7.	
98	CDC5L.	LXRα.	Hox-A13.	ZNF33A.	KLF5.	ZNF502.	
99	DLX-2.	RAX-2.	ATF-4.	ZNF547.	COUP-TFII.	HIC2.	
100	ZBTB33.	RAR-γ.	Hox-D12.	SREBP-1.	NKX-2.5.	XBP-1.	

图27续

大的结合位点(>300bp)

[1] ATF-1.50perc_hg19.tss ATF-3.50perc_hg19.tss BARX-2.50perc_hg19.tss
[4] c-Ets-1.50perc_hg19.tss CREB.50perc_hg19.tss CREM.50perc_hg19.tss
[7] DEAF1.50perc_hg19.tss DEAF1.hg19.tss E2F-1.50perc_hg19.tss
[10] Elf-1.50perc_hg19.tss Elk-1.hg19.tss ERG.50perc_hg19.tss
[13] ETV6.50perc_hg19.tss Fli-1.50perc_hg19.tss FOXP3.50perc_hg19.tss
[16] GATAD1.50perc_hg19.tss GLI2.hg19.tss GMEB-2.50perc_hg19.tss
[19] GMEB-2.hg19.tss IRF-3.hg19.tss KLF10.hg19.tss
[22] KLF11.50perc_hg19.tss KLF11.hg19.tss MAZ.50perc_hg19.tss
[25] MLL2.50perc_hg19.tss NF-YA.50perc_hg19.tss NF-YB.50perc_hg19.tss
[28] NF-YB.hg19.tss NF-YC.50perc_hg19.tss NF-YC.hg19.tss
[31] PBX-3.50perc_hg19.tss PHOX-2B.50perc_hg19.tss PHOX-2B.hg19.tss
[34] Runx1.50perc_hg19.tss SIX5.50perc_hg19.tss SMAD1.50perc_hg19.tss
[37] SOX-9.hg19.tss Sp1.50perc_hg19.tss Sp2.50perc_hg19.tss
[40] Sp4.50perc_hg19.tss Sp5.50perc_hg19.tss SREBP-1.hg19.tss
[43] SRF.50perc_hg19.tss STAT3.50perc_hg19.tss TR4.50perc_hg19.tss
[46] UBP-1.50perc_hg19.tss USF-2.50perc_hg19.tss ZBTB33.50perc_hg19.tss
[49] ZGPAT.50perc_hg19.tss ZHX2.50perc_hg19.tss ZNF75A.50perc_hg19.tss
[52] ZNF75A.hg19.tss ZNF76.50perc_hg19.tss ZNF76.hg19.tss
[55] ZSCAN16.50perc_hg19.tss

图28A

结合位点接近双核小体大小
(大小在312bp和352 bp之间)

[1] ATF-1.50perc_hg19.tss ATF-3.50perc_hg19.tss c-Ets-1.50perc_hg19.tss
[4] CREB.50perc_hg19.tss CREM.50perc_hg19.tss DEAF1.50perc_hg19.tss
[7] DEAF1.hg19.tss E2F-1.50perc_hg19.tss Elf-1.50perc_hg19.tss
[10] Elk-1.hg19.tss ERG.50perc_hg19.tss Fli-1.50perc_hg19.tss
[13] FOXP3.50perc_hg19.tss GMEB-2.50perc_hg19.tss GMEB-2.hg19.tss
[16] MAZ.50perc_hg19.tss NF-YC.hg19.tss SMAD1.50perc_hg19.tss
[19] SOX-9.hg19.tss SRF.50perc_hg19.tss STAT3.50perc_hg19.tss
[22] UBP-1.50perc_hg19.tss ZBTB33.50perc_hg19.tss ZGPAT.50perc_hg19.tss
[25] ZNF75A.50perc_hg19.tss ZNF75A.hg19.tss

图28B

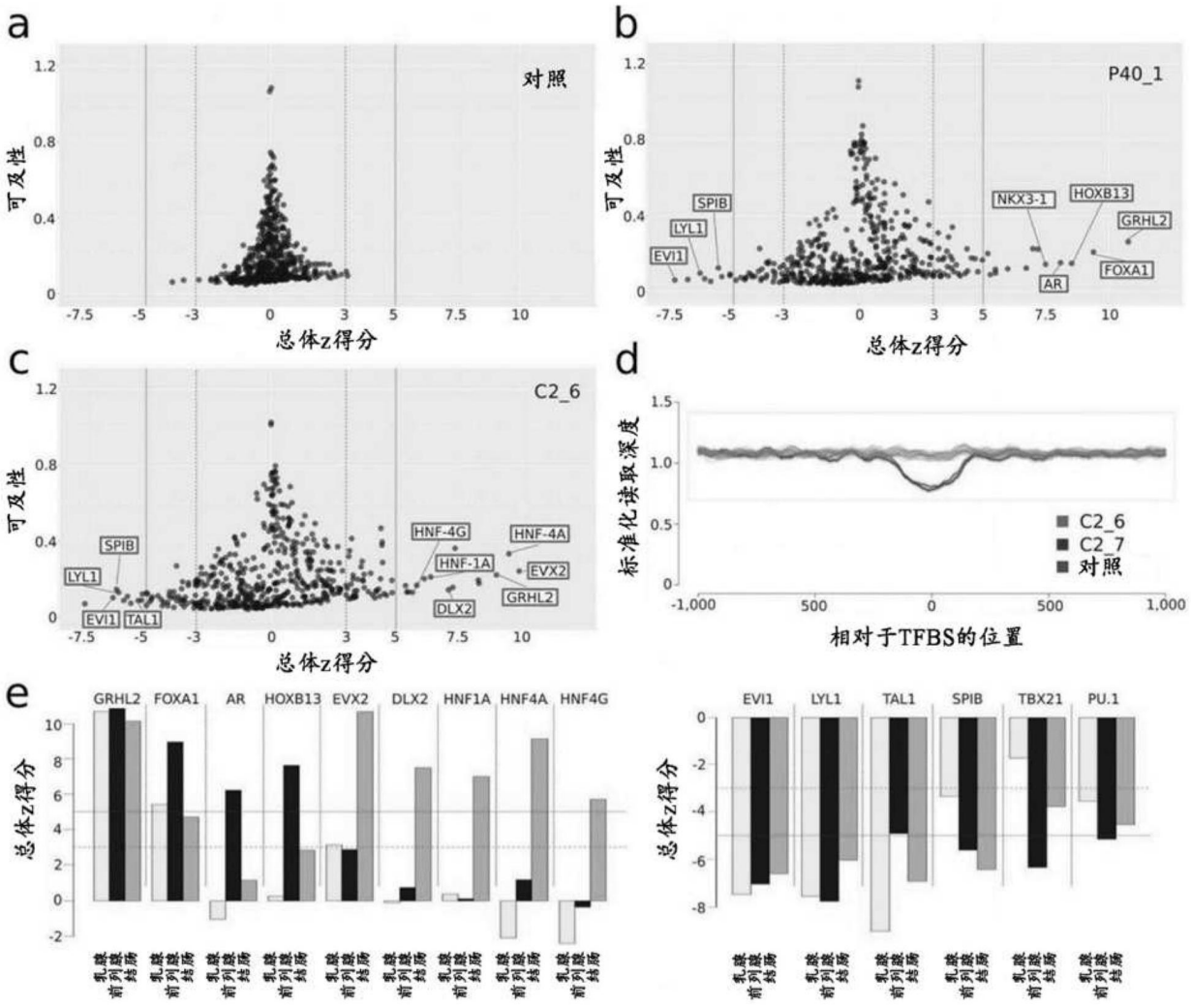

a
对照
可及性
总体z得分
b
P40_1
SPIB
LYL1
EVI1
NKX3-1
HOXB13
GRHL2
FOXA1
AR
c
C2_6
HNF-4G
HNF-4A
HNF-1A
EVX2
DLX2
TAL1
d
标准化读取深度
C2_6
C2_7
对照
相对于TFBS的位置
e
GRHL2 FOXA1 AR HOXB13 EVX2 DLX2 HNF1A HNF4A HNF4G
EVI1 LYL1 TAL1 SPIB TBX21 PU.1
乳腺 前列腺 结肠

图29A-29E

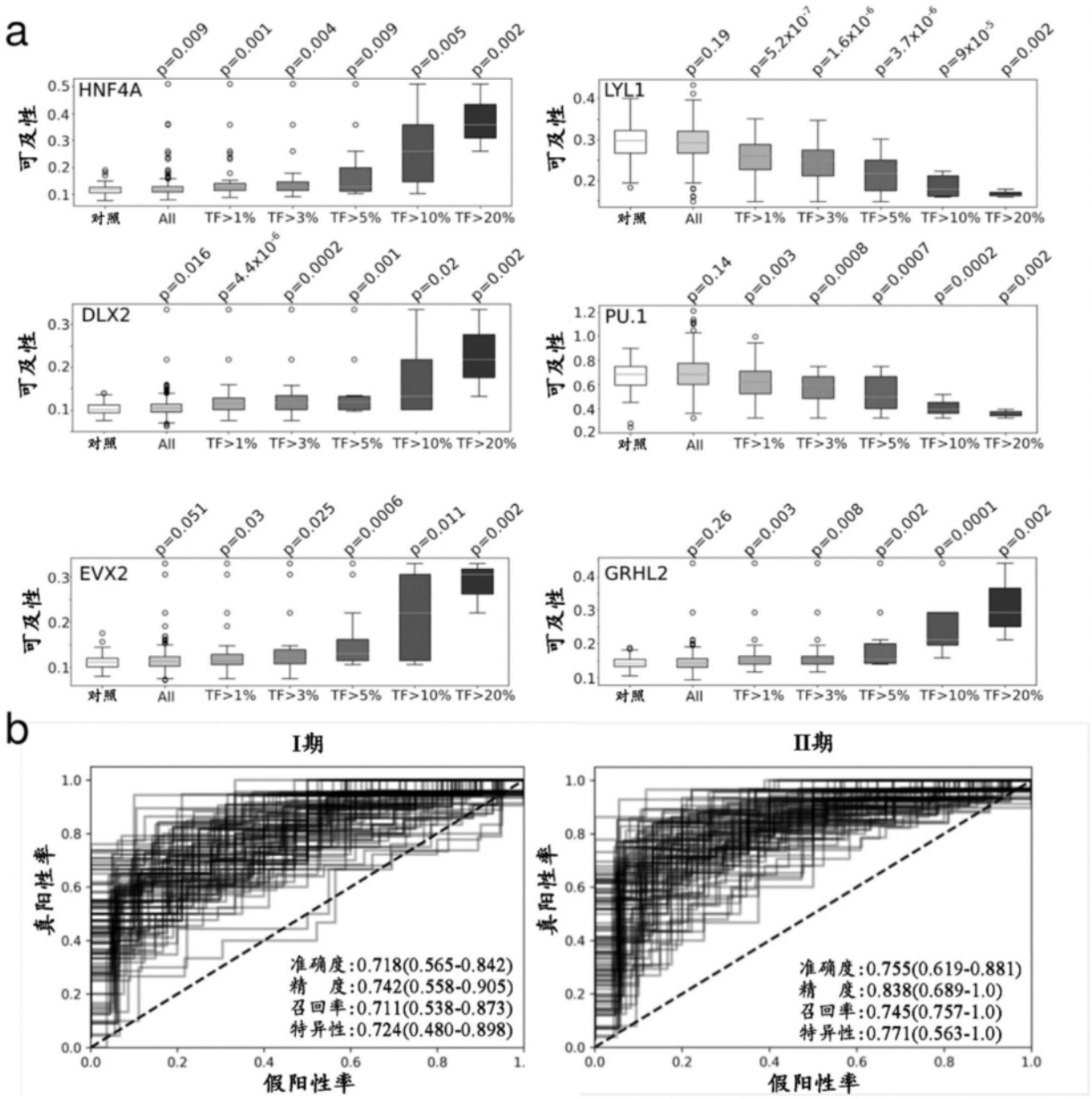

a
HNF4A
LYL1
DLX2
PU.1
EVX2
GRHL2
可及性
对照
All
TF>1%
TF>3%
TF>5%
TF>10%
TF>20%
b
I期
II期
真阳性率
假阳性率
准确度:0.718(0.565-0.842)
精 度:0.742(0.558-0.905)
召回率:0.711(0.538-0.873)
特异性:0.724(0.480-0.898)
准确度:0.755(0.619-0.881)
精 度:0.838(0.689-1.0)
召回率:0.745(0.757-1.0)
特异性:0.771(0.563-1.0)

图30A-30B