



(12) 发明专利申请

(10) 申请公布号 CN 112292458 A

(43) 申请公布日 2021.01.29

(21) 申请号 201980029736.2

孙坤

(22) 申请日 2019.05.03

(74) 专利代理机构 北京英赛嘉华知识产权代理

有限责任公司 11204

(30) 优先权数据

代理人 王达佐 洪欣

62/666,574 2018.05.03 US

62/732,509 2018.09.17 US

(51) Int.Cl.

(85) PCT国际申请进入国家阶段日

C12Q 1/6827 (2006.01)

2020.11.02

C12Q 1/6869 (2006.01)

(86) PCT国际申请的申请数据

C12Q 1/6883 (2006.01)

PCT/CN2019/085426 2019.05.03

C12Q 1/6886 (2006.01)

(87) PCT国际申请的公布数据

W02019/210873 EN 2019.11.07

(71) 申请人 香港中文大学

地址 中国香港新界

申请人 格里尔公司

(72) 发明人 卢煜明 赵慧君 陈君赐 江培勇

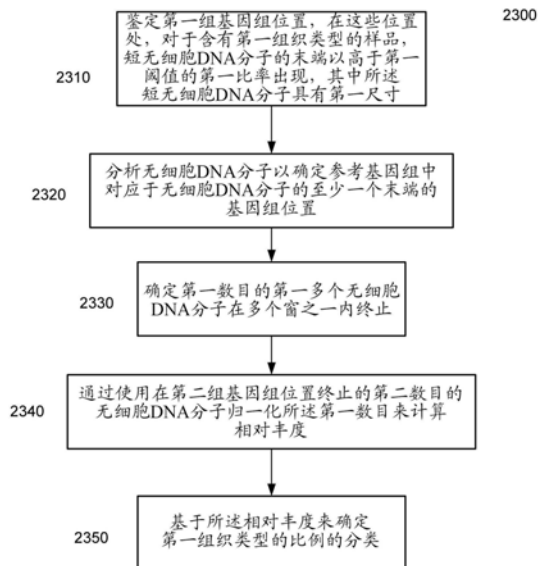
权利要求书6页 说明书57页 附图65页

(54) 发明名称

测量无细胞混合物特性的尺寸标记的优选末端和识别方向的分析

(57) 摘要

多种应用可以使用与无细胞DNA,例如血浆DNA和血清DNA相关的片段化模式。例如,DNA片段的末端位置可用于多种应用。短和长DNA分子的片段化模式可以与不同的称为尺寸标记的优选末端的优选DNA末端位置相关。在另一个实例中,分析了与组织特异性开放染色质区域有关的片段化模式。可以在来自不同组织类型的无细胞DNA的混合物中确定特定组织类型的比例贡献的分类。另外,可以确定特定组织类型的性质,例如,对于一组织类型,序列失衡是否存在于特定区域中,或者对于该组织类型,病状是否存在。



1. 分析包括来自多种组织类型的无细胞DNA分子的混合物的生物样品,以确定第一组织类型在所述混合物中的比例贡献的分类的方法,所述多种组织类型包括所述第一组织类型,所述方法包括:

鉴定第一组基因组位置,在所述位置处,对于含有所述第一组织类型的样品,短的无细胞DNA分子的末端以高于第一阈值的第一比率出现,其中所述短的无细胞DNA分子具有第一尺寸;

分析来自受试者的生物样品的第一多个无细胞DNA分子,其中分析无细胞DNA分子包括:

确定参考基因组中对应于所述无细胞DNA分子的至少一个末端的基因组位置;

基于对所述第一多个无细胞DNA分子的分析,确定第一数目的所述第一多个无细胞DNA分子在多个窗之一内终止,每个窗包括所述第一组基因组位置中的至少一个;

通过使用第二数目的无细胞DNA分子归一化所述第一数目的第一多个无细胞DNA分子来计算在所述多个窗之一内终止的所述第一多个无细胞DNA分子的相对丰度,其中所述第二数目的无细胞DNA分子包括在所述多个窗之外的第二组基因组位置处终止的无细胞DNA分子,所述多个窗包括所述第一组基因组位置;和

通过将所述相对丰度与从一个或多个校准样品确定的一个或多个校准值进行比较来确定第一组织类型的比例贡献的分类,所述一个或多个校准样品的第一组织类型的比例贡献是已知的。

2. 如权利要求1所述的方法,其中所述多个窗具有1bp的宽度。

3. 如权利要求1所述的方法,其中所述相对丰度包括所述第一数目和所述第二数目的比值。

4. 如权利要求1所述的方法,其中,所述比例贡献的分类对应于指定百分比以上的范围。

5. 如权利要求1所述的方法,其中所述第一组织类型是肿瘤,并且其中所述分类选自:所述受试者中肿瘤组织的量,所述受试者中肿瘤的尺寸,所述受试者中肿瘤的阶段,所述受试者中的肿瘤负荷,和所述受试者中肿瘤转移的存在。

6. 如权利要求1所述的方法,其中鉴定第一组基因组位置包括:

通过计算机系统分析来自至少一个另外的样品的第二多个无细胞DNA分子以鉴定所述第二多个无细胞DNA分子的终止位置,其中已知所述至少一个另外的样品包括第一组织类型并且为与所述生物样品相同的样品类型;和

对于多个基因组窗的每个基因组窗:

计算在所述基因组窗上终止的所述第二多个无细胞DNA分子的相应数目;和

将所述相应数目与参考值进行比较,以确定在所述基因组窗内的一个或多个基因组位置上终止的无细胞DNA分子的比率是否高于所述第一阈值。

7. 如权利要求6所述的方法,其中所述参考值由在所述基因组窗外的基因组位置处终止的所述第二多个无细胞DNA分子的数目确定。

8. 如权利要求7所述的方法,其中当相对于在特定基因组位置周围的窗内的基因组位置上终止的所述第二多个无细胞DNA分子的数目,所述特定基因组位置位于峰值时,所述特定基因组位置被鉴定为在所述第一组基因组位置中。

9. 如权利要求6所述的方法,其中所述参考值是使用在以所述基因组窗的特定基因组位置为中心的窗处终止的所述第二多个无细胞DNA分子的数目除以无细胞DNA分子的平均尺寸来确定的。

10. 如权利要求6所述的方法,其中所述参考值是根据所述至少一个另外的样品中的无细胞DNA分子的概率分布和平均长度,在所述基因组窗内终止的无细胞DNA分子的预期数目。

11. 如权利要求6所述的方法,其中所述至少一个另外的样品是所述一个或多个校准样品。

12. 如权利要求1所述的方法,其还包括:

鉴定第二组基因组位置,在所述位置处,长无细胞DNA分子的末端以高于第二阈值的第二比率出现,其中所述长无细胞DNA分子具有大于所述第一尺寸的第二尺寸。

13. 如权利要求12所述的方法,其中所述第一尺寸是第一尺寸范围,并且其中所述第二尺寸是第二尺寸范围。

14. 如权利要求13所述的方法,其中通过所述第一尺寸范围的第一最大值小于所述第二尺寸范围的第二最大值,所述第一尺寸范围小于所述第二尺寸范围。

15. 如权利要求14所述的方法,其中所述第一尺寸范围与所述第二尺寸范围重叠。

16. 如权利要求1所述的方法,其中所述第二组基因组位置包括对应于所述第一多个无细胞DNA分子中的至少一个的末端的所有基因组位置。

17. 如权利要求1所述的方法,其中将所述相对丰度与所述一个或多个校准值进行比较使用与校准点拟合的校准函数,所述校准点包括在多个校准样品中测量的所述第一组织类型的比例贡献和在所述多个校准样品中确定的相应相对丰度。

18. 分析包括来自多种组织类型的无细胞DNA分子的混合物的受试者的生物样品,以确定第一组织类型在所述无细胞DNA分子的混合物中是否展现出染色体区域的序列失衡的方法,所述多种组织类型包括所述第一组织类型,所述方法包括:

鉴定第一组基因组位置,在所述位置处,对于含有所述第一组织类型的样品,短无细胞DNA分子的末端以高于第一阈值的第一比率出现,其中所述短无细胞DNA分子具有第一尺寸;

通过计算机系统分析来自所述生物样品的第一多个无细胞DNA分子,其中分析无细胞DNA分子包括:

确定参考基因组中对应于所述无细胞DNA分子的至少一个末端的基因组位置;

基于对所述第一多个无细胞DNA分子的分析,鉴定在多个窗之一内终止的无细胞DNA分子的组,每个窗包括所述组的基因组位置中的至少一个并且位于所述染色体区域中;

确定所述无细胞DNA分子的组的值;和

基于所述无细胞DNA分子的组的值与参考值的比较,确定所述受试者的染色体区域内,所述第一组织类型中是否存在序列失衡的分类。

19. 如权利要求18所述的方法,其中所述参考值是从不具有序列失衡的一个或多个对照样品确定的。

20. 如权利要求18所述的方法,其中鉴定第一组基因组位置包括:

通过计算机系统分析来自至少一个另外的样品的第二多个无细胞DNA分子以鉴定所述

第二多个无细胞DNA分子的终止位置,其中已知所述至少一个另外的样品包括第一组织类型并且为与所述生物样品相同的样品类型;和

对于多个基因组窗的每个基因组窗:

计算在所述基因组窗上终止的所述第二多个无细胞DNA分子的相应数目;和

将所述相应数目与参考值进行比较,以确定在所述基因组窗内的一个或多个基因组位置上终止的无细胞DNA分子的比率是否高于所述第一阈值。

21. 如权利要求18所述的方法,其中所述无细胞DNA分子的组的值使用所述第一多个无细胞DNA分子的总数进行归一化。

22. 如权利要求18所述的方法,其中使用一个或多个参考区域的另一无细胞DNA分子的组的值对所述无细胞DNA分子的组的值进行归一化。

23. 如权利要求18所述的方法,其中所述序列失衡是非整倍性,扩增/缺失,或所述第一组织类型与所述多种组织类型的其它组织类型在所述染色体区域的基因座处的不同基因型。

24. 如权利要求23所述的方法,其中所述序列失衡是所述第一组织类型的基因型与所述多种组织类型的其它组织类型不同的结果,并且其中所述无细胞DNA分子的组的值是在所述基因座具有第一等位基因的所述组的第一数目的无细胞DNA分子和在所述基因座具有第二等位基因的第二数目的无细胞DNA分子之间的相对丰度。

25. 如权利要求24所述的方法,其中其它组织类型在所述染色体区域的基因座处是杂合的,并且其中序列失衡的分类是第一等位基因过多,这表明第一组织类型对于第一等位基因是纯合的。

26. 如权利要求24所述的方法,其中其它组织类型在所述染色体区域的基因座处是杂合的,并且其中分类是不存在失衡,这表明第一组织类型对于第一等位基因和第二等位基因是杂合的。

27. 如权利要求18所述的方法,其中所述无细胞DNA分子的组的值是无细胞DNA分子的组的量,所述无细胞DNA分子的组的尺寸分布的统计值,或所述无细胞DNA分子的组的甲基化水平。

28. 如权利要求27所述的方法,其中确定所述无细胞DNA分子的组的值包括:

鉴定在多个窗之一内终止的无细胞DNA分子的组的第一亚组,所述第一亚组对应于染色体区域中的第一单倍型;

确定所述第一亚组的无细胞DNA分子的第一单倍型值;

鉴定在多个窗之一内终止的无细胞DNA分子的组的第二亚组,所述第二亚组对应于染色体区域中的第二单倍型;

确定所述第二亚组的无细胞DNA分子的第二单倍型值;和

使用所述第一单倍型值和所述第二单倍型值确定分离值,所述分离值是无细胞DNA分子的组的值。

29. 如权利要求27所述的方法,其还包括:

通过以下确定参考值:

鉴定在多个参考窗之一内终止的无细胞DNA分子的参考组,每个参考窗包括所述组的基因组位置中的至少一个并且位于一个或多个参考染色体区域中;和

确定所述无细胞DNA分子的参考组的参考值,所述参考值是所述无细胞DNA分子的参考组的量,所述无细胞DNA分子的参考组的尺寸分布的统计值,或所述无细胞DNA分子的参考组的甲基化水平。

30. 如权利要求29所述的方法,其中将所述值与所述参考值比较包括:

使用所述无细胞DNA分子的组的值和所述无细胞DNA分子的参考组的参考值确定分离值;和

将所述分离值与截止值进行比较,所述截止值分离存在序列失衡和不存在序列失衡的分类。

31. 如权利要求18所述的方法,其中所述染色体区域是染色体。

32. 分析包括来自多种组织类型的无细胞DNA分子的混合物的生物样品,以确定第一组织类型在所述混合物中的比例贡献的分类的方法,所述多种组织类型包括所述第一组织类型,所述方法包括:

鉴定与对应于第一组织类型的一个或多个组织特异性开放染色质区域的中心具有指定距离的第一组基因组位置;

分析来自受试者的生物样品的第一多个无细胞DNA分子,其中分析无细胞DNA分子包括:

确定参考基因组中对应于所述无细胞DNA分子的两个末端的基因组位置;和

基于哪个末端具有基因组位置的较低值,将一个末端分类为上游末端,将另一个末端分类为下游末端;

确定第一数目的所述第一多个无细胞DNA分子在所述第一组基因组位置之一处具有上游末端;

确定第二数目的所述第一多个无细胞DNA分子在所述第一组基因组位置之一处具有下游末端;

计算所述第一数目与所述第二数目之间的分离值;和

通过将所述分离值与从一个或多个校准样品确定的一个或多个校准值进行比较来确定所述第一组织类型的比例贡献的分类,所述一个或多个校准样品的第一组织类型的比例贡献是已知的。

33. 如权利要求32所述的方法,其中将所述分离值与所述一个或多个校准值进行比较使用与校准点拟合的校准函数,所述校准点包括在多个校准样品中测量的所述第一组织类型的比例贡献和在所述多个校准样品中确定的相应相对丰度。

34. 如权利要求1、18或32中任一项所述的方法,其中所述第一组织类型是胎儿组织。

35. 分析包括来自多种组织类型的无细胞DNA分子的混合物的生物样品,以确定所述混合物中第一组织类型在是否存在病状的分类的方法,所述多种组织类型包括所述第一组织类型,所述方法包括:

鉴定与对应于第一组织类型的一个或多个组织特异性开放染色质区域的中心具有指定距离的第一组基因组位置;

分析来自受试者的生物样品的第一多个无细胞DNA分子,其中分析无细胞DNA分子包括:

确定参考基因组中对应于所述无细胞DNA分子的两个末端的基因组位置;和

基于哪个末端具有基因组位置的较低值,将一个末端分类为上游末端,将另一个末端分类为下游末端;

确定第一数目的所述第一多个无细胞DNA分子在所述第一组基因组位置之一处具有上游末端;

确定第二数目的所述第一多个无细胞DNA分子在所述第一组基因组位置之一处具有下游末端;

利用所述第一数目与所述第二数目计算分离值;和

基于所述分离值与参考值的比较,确定所述受试者的第一组织类型是否存在病状的分

36.如权利要求32或35所述的方法,其中所述一个或多个组织特异性开放染色质区域包括至少500个对应于所述第一组织类型的组织特异性开放染色质区域。

37.如权利要求32或35所述的方法,其中所述分离值包括比值和/或差值。

38.如权利要求32或35所述的方法,其中所述指定距离包括距离范围。

39.如权利要求38所述的方法,其中所述指定距离包括在所述中心之前的第一距离范围,并且包括在所述中心之后的第二距离范围。

40.如权利要求39所述的方法,其中对于所述第一范围,以第一方式确定对所述分离值的第一贡献,并且其中对于所述第二范围,以第二方式确定对所述分离值的第二贡献。

41.如权利要求40所述的方法,其中所述分离值被确定为

$$OCF = \sum_{-峰-仓}^{-峰+仓} (D - U) + \sum_{峰-仓}^{峰+仓} (U - D)$$

其中峰位置对应于与中心的偏移,并且仓值对应于峰位置周围的窗尺寸,并且其中所述第一数目是在所述第一组中的基因组位置之一的值U,并且其中所述第二数目是在所述第一组中的基因组位置之一的值D。

42.如权利要求35所述的方法,其中所述参考值是从不具有所述病状的一个或多个对照样品确定的。

43.如权利要求35所述的方法,其中所述参考值是从确实具有所述病状的一个或多个对照样品确定的。

44.如权利要求35所述的方法,其中所述病状是来自所述第一组织类型的无细胞DNA的异常高的浓度分数。

45.如权利要求35所述的方法,其中所述病状是移植器官的排斥。

46.如权利要求35所述的方法,其中所述病状是所述第一组织类型的癌症。

47.如权利要求46所述的方法,其中所述癌症是肝癌,结肠癌或肺癌。

48.如权利要求1、18、32或35所述的方法,其中所述第一组织类型是肿瘤。

49.如权利要求1、18、32或35所述的方法,其中所述第一组织类型是移植组织。

50.如权利要求1、18、32或35所述的方法,其中分析所述第一多个无细胞DNA分子包括:对所述第一多个无细胞DNA分子进行测序,以获得序列读取;和将所述序列读取与参考基因组比对以确定所述第一多个无细胞DNA分子的基因组位置。

51. 如权利要求1、18、32或35所述的方法,其中分析所述第一多个无细胞DNA分子包括:在所述第一组基因组位置处杂交捕获或扩增所述第一多个无细胞DNA分子。

52. 计算机产品,其包含计算机可读介质,所述计算机可读介质存储用于控制计算机系统以执行上述方法中的任何一项的操作的多个指令。

53. 系统,其包含:

权利要求52所述的计算机产品;和

一个或多个处理器,其用于执行存储在所述计算机可读介质上的指令。

54. 系统,其包含用于执行上述方法中任何一种的装置。

55. 系统,其包含被配置成执行上述方法中任何一种的一个或多个处理器。

56. 系统,其包含分别执行上述方法任何一种中的步骤的模块。

测量无细胞混合物特性的尺寸标记的优选末端和识别方向的分析

相关申请的交叉引用

[0001] 本申请要求2018年9月17日提交的标题为“测量无细胞混合物特性的尺寸标记的优选末端和识别方向的 (Orientation-Aware) 分析”的美国临时申请第62/732,509号和2018年5月3日提交的标题为“用于测量无细胞混合物性能的尺寸标记的优选末端”的美国临时申请第62/666,574号的优先权并且是它们的PCT申请,通过引用将其整体并入本文用于所有目的。

背景

[0002] 最早由Mandel和Metais报道了人血浆中循环无细胞DNA (cfDNA) 的存在 (86)。后来,在孕妇血浆中发现胎儿来源的DNA (82),在移植患者中发现供体来源的DNA (83) 和在癌症患者中发现肿瘤来源的DNA (100) 打开了基于血浆DNA的非侵入性产前测试 (108),移植监测 (97) 和癌症液体活检 (57,91,61) 的大门。因此,cfDNA已成为在全球范围内积极研究的生物标志物类别。

[0003] 采用人血浆中的循环无细胞DNA分析进行分子诊断和监测引起了全球关注。在孕妇血浆中发现胎儿DNA (1),在器官移植患者中发现供体特异性DNA (2) 和在癌症患者中发现肿瘤来源的DNA (3) 使得非侵入性产前测试,癌症液体活检,移植监测和器官损伤评估 (4-8) 成为可能。尽管有许多临床应用,血浆DNA的生物学特性尚未得到足够的研究关注。

发明内容

[0004] 多个实施方案涉及与无细胞DNA (例如血浆DNA和血清DNA) 相关的片段化模式的分析的应用 (例如,诊断应用)。例如,DNA片段 (分子) 的末端位置可用于多种应用。一些实施方案可以确定来自不同组织类型的无细胞DNA的混合物中特定组织类型的比例贡献的分类。例如,可以确定特定百分比,百分比范围或比例贡献是否高于指定百分比作为分类。在其它实施方案中,可以确定特定组织类型的性质,例如,对于一组织类型,序列失衡是否存在于特定区域中,或者对于该组织类型,病状是否存在。

[0005] 在一个实例中,分析了不同尺寸的无细胞DNA分子的片段化模式。短和长DNA分子可以与不同的称为尺寸标记的优选末端的优选DNA末端位置相关。短的优选的DNA末端位置与某些组织类型 (例如胎儿,肿瘤或移植组织) 相关。可以鉴定短 (和可能长) DNA分子的优选终止位置,并且在此类位置终止的DNA分子可以用于多种应用中。

[0006] 在一些实施方案中,在短DNA分子的优选终止位置终止的无细胞DNA分子的相对丰度可以用于确定测试混合物中第一组织类型的比例贡献,例如,通过与在已知比例贡献的校准样品中的类似测量值相比较。

[0007] 在其它实施方案中,可以分析在短DNA分子的优选终止位置和特定染色体区域中的位置终止的一组无细胞DNA分子,以确定该组的值 (例如,计数,尺寸分布的统计值或甲基化水平)。所述值可用于检测序列失衡 (例如,拷贝数畸变,例如非整倍性,缺失或扩增以及基因型差异)。当在染色体区域中存在序列失衡时,所述值将显示出与参考值的统计学上显

著的偏差。

[0008] 在另一个实例中,分析了与组织特异性开放染色质区域有关的片段化模式。可以使用相对于第一组织类型的组织特异性开放染色质区域的中心的一组基因组位置。具体地,可以在定量分析中使用关于DNA片段在这组基因组位置具有上游末端还是下游末端(例如相对于特定组织类型的开放染色质区域的中心)的知识。例如,可以使用在具有上游末端和下游末端的DNA分子的相应数目中的分离值(例如,差值或比值)。

[0009] 在一些实施方案中,分离值可以用于确定测试混合物中第一组织类型的比例贡献,例如,通过与已知比例贡献的校准样品中的类似测量值相比较。在其它实施方案中,例如当与参考值存在统计学上显著的偏差时,分离值可以用作第一组织类型中的病状的指标。这样的病状的实例包括来自第一组织类型的无细胞DNA的异常高的浓度分数,是第一组织类型的移植器官的排斥或癌症。

[0010] 下面详细描述本发明的这些和其它实施方案。例如,其它实施方案涉及与本文描述的方法相关的系统,装置和计算机可读介质。

[0011] 参考以下详细描述和附图,可以更好地理解本公开的实施方案的性质和优点。

附图的简要说明

[0012] 图1显示了根据本公开的实施方案的血浆DNA片段的片段末端位点的分析。

[0013] 图2显示了在24个母体血浆样品中覆盖Set S优选末端位点的血浆DNA读取的尺寸分布(红色)对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布(蓝色)。

[0014] 图3显示了根据本公开的实施方案,在一种母体血浆样品中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。

[0015] 图4A显示了26个母体血浆样品中具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)与胎儿DNA分数之间的相关性。图4B显示了26个母体血浆样品的尺寸比(短读取比长读取的数目)和胎儿DNA分数之间的相关性。

[0016] 图5A显示了根据本公开的实施方案,在对照病例与21三体病例之间的chr21读取的相对丰度的比较。图5B显示了根据本公开的实施方案,对于21三体测试,覆盖Set S优选末端位点的读取与随机读取之间的ROC比较。

[0017] 图6显示在24名健康受试者中,覆盖Set S优选末端位点的血浆DNA读取的尺寸分布对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布。

[0018] 图7A显示了根据本公开的实施方案,在健康受试者中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。图7B显示了根据本公开的实施方案,在孕妇和健康受试者中具有Set S优选末端位点的血浆DNA读取对比具有Set L优选末端位点的血浆DNA读取的相对丰度(S/L比值)的比较。

[0019] 图8显示了根据本公开的实施方案,在肝细胞癌(HCC)患者中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。

[0020] 图9显示了在代表性的一组24例肝细胞癌患者中,覆盖Set S优选末端位点的血浆DNA读取的尺寸分布对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布。

[0021] 图10显示了根据本公开的实施方案,在血浆中具有大于1%的肿瘤DNA分数的72例肝细胞癌患者中,具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)与肿瘤DNA分数之间的相关性。

[0022] 图11显示了健康受试者和肝细胞癌患者之间具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)。

[0023] 图12显示了根据本公开的实施方案,在健康受试者,没有或患有肝硬化的HBV携带者和HCC患者中覆盖chr1p上的Set S末端的归一化读取计数。

[0024] 图13显示了根据本公开的实施方案,在健康受试者,没有或患有肝硬化的HBV携带者和HCC患者中覆盖chr1q上的Set S末端的归一化读取计数。

[0025] 图14显示了根据本公开的实施方案,在健康受试者,没有或患有肝硬化的HBV携带者和HCC患者中覆盖chr8p上的Set S末端的归一化读取计数。

[0026] 图15显示了根据本公开的实施方案,在健康受试者,没有或患有肝硬化的HBV携带者和HCC患者中覆盖chr8q上的Set S末端的归一化读取计数。

[0027] 图16显示了根据本公开的实施方案,Set S和Set L优选末端位点中的任意两个最接近的优选末端位点之间的距离分布。

[0028] 图17A显示了根据本公开的实施方案,血浆DNA覆盖,Set S和Set L优选末端位点的快照。图17B显示了根据本公开的实施方案,围绕由胎盘组织和T细胞共有的共同的开放染色质区域的优选末端位点的分布。

[0029] 图18A显示了根据本公开的实施方案,妊娠血浆DNA中的尺寸标记的优选末端位点相对于核小体结构的分布。图18B显示了根据本公开的实施方案,尺寸标记的优选末端位点相对于由Straver等人(23)预测的核小体中心的分布。

[0030] 图19显示了根据本公开的实施方案,在健康的非妊娠受试者中短和长DNA分子的常染色体片段末端相对于核小体结构的分布。

[0031] 图20A显示了核小体结构的图示。图20B显示了核小体结构中胎儿和母体特异性的优选末端位点的分布。图20C显示了妊娠病例和健康男性受试者的chrY片段末端在核小体结构中的分布。图20D显示了在妊娠情况下短和长DNA分子的chrY片段末端在核小体结构中的分布。图20E显示了健康受试者中短和长DNA分子的chrY片段末端在核小体结构中的分布。

[0032] 图21A和21B显示了来自(A)血沉棕黄层样品和(B)胎盘组织的ATAC-seq数据的片段尺寸分布。

[0033] 图22显示了在短标记的终止位置上终止的无细胞DNA分子的相对丰度(例如短/长)与混合物中组织A对DNA的比例贡献(其通过分析来自组织A的两个或更多个具有已知的DNA比例浓度的校准样品确定)之间的关系。

[0034] 图23是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型的比例贡献的的方法的流程。

[0035] 图24是根据本公开的实施方案,分析生物学样品以确定第一组织类型是否无细胞DNA分子的混合物中的染色体区域中显示序列失衡的方法的流程。

[0036] 图25A-25F显示了根据本公开的实施方案的无细胞DNA(cfDNA)片段化分析的概念框架。图25A是具有包裹的DNA(黄线),接头(棕线)和活性调节元件(绿线)的核小体的图示。图25B显示了由凋亡DNA片段化产生的cfDNA的图示。图25C是两端的测序读取和提取的图示。红色和蓝色分别代表U(上游)和D(下游)血浆DNA末端。图25D显示了基因组覆盖。图25E显示了相对于基因组坐标的cfDNA的U和D片段末端概况。图25F显示了平滑的血浆DNA末端

信号和推导的核小体定位。

[0037] 图26A和26B显示了根据本公开的实施方案,在合并的健康非妊娠受试者的chr12p11.1区域中的血浆DNA片段化模式。图26A显示了原始信号。图26B显示了平滑的信号和推导的核小体定位。图26C显示了管家基因的活性启动子周围的血浆DNA覆盖和末端信号。图26D显示了非活性启动子周围的血浆DNA覆盖和末端信号。

[0038] 图27A,27B和27C显示了根据本公开的实施方案,在合并的健康非妊娠受试者中的血浆DNA片段化模式。图27A显示了T细胞和肝细胞共有的共同的开放染色质区域中的模式(还绘制了推导的核小体定位)。图27B显示了胚胎干细胞(ESC)特异性开放染色质区域中的模式。图27C是OCF(识别方向的cfDNA片段化)值的概念的图示。

[0039] 图28A-28G显示了根据本公开的实施方案,在健康受试者中的组织特异性开放染色质区域中的血浆DNA片段化模式。每幅图显示了来自与一种组织类型相对应的组织特异性开放染色质区域的结果:28A T细胞;28B胎盘;28C肝脏;28D肺;28E卵巢;28F乳房;28G肠。

[0040] 图29A显示了根据本公开的实施方案,在一例CRC患者的肠特异性开放染色质区域中的血浆DNA片段化模式。

[0041] 图29B显示了根据本公开的实施方案,在一例肺癌患者中的肺特异性开放染色质区域中的血浆DNA片段化模式。

[0042] 图30显示了根据本公开的实施方案,对在健康非妊娠受试者群体中的各种组织之间的血浆DNA片段化模式(OCF值)的定量。

[0043] 图31显示了根据本公开的实施方案,在健康个体中的组织类型的OCF值的表。

[0044] 图32A-32D显示了根据本公开的实施方案,血浆DNA片段化模式分析在非侵入性产前测试中的应用。图32A显示了在一例妊娠病例中胎盘特异性开放染色质区域中的血浆DNA片段化模式。图32B显示了健康的非妊娠受试者和孕妇之间T细胞的OCF值的比较。图32C显示了健康的非妊娠受试者和孕妇之间胎盘的OCF值的比较。图32D显示了在26名孕妇的群体中,胎盘的OCF值和胎儿DNA分数之间的相关性。

[0045] 图33显示了根据本公开的实施方案,妊娠受试者中的OCF值组织类型的表。

[0046] 图34显示了根据本公开的实施方案,肝脏移植患者中的OCF值组织类型的表。

[0047] 图35A,35B和35C显示了根据本公开的实施方案,血浆DNA片段化模式分析在肝脏移植和HCC患者中的应用。图35A显示了肝脏移植患者中肝脏的OCF值与供体DNA分数之间的相关性。图35B显示了HCC病例中的肿瘤DNA分数。图35C显示了健康受试者和HCC病例(根据血浆中的肿瘤DNA负荷分为两组)的T细胞的OCF值的比较。图35D显示了健康受试者和HCC病例(根据血浆中的肿瘤DNA负荷分为两组)的肝脏的OCF值的比较。

[0048] 图36A-36D显示了根据本公开的实施方案,肝细胞癌患者中的OCF值组织类型的表。

[0049] 图37A-37E显示了根据本公开的实施方案,血浆DNA片段化模式分析在CRC和肺癌患者中的应用。图37A显示了健康受试者与CRC患者之间T细胞的OCF值的比较。图37B显示了健康受试者和CRC患者之间肠的OCF值的比较。图37C显示了CRC患者中肠的OCF值与结肠DNA分数(通过血浆DNA组织映射法推导)之间的相关性。图37D显示了健康受试者和肺癌患者之间T细胞的OCF值的比较。图37E显示了健康受试者和肺癌患者之间的肺OCF值的比较。

[0050] 图38显示了根据本公开的实施方案,肺癌患者中的OCF值组织类型的表。

[0051] 图39显示了根据本公开的实施方案,结肠直肠癌患者中的OCF值组织类型的表。

[0052] 图40是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型的比例贡献的的分类的方法的流程。

[0053] 图41是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型是否存在病状的分类的方法的流程。

[0054] 图42示出了根据本公开的实施方案的测量系统。

[0055] 图43显示了可与根据本公开的实施方案的系统和方法一起使用的示例计算机系统的框图。

术语

[0056] “组织”对应于集合在一起作为功能单元的一组细胞。可以在单一组织中发现超过一种类型的细胞。不同类型的组织可以由不同类型的细胞(例如肝细胞、肺泡细胞或血细胞)组成,但也可以对应于来自不同生物体(母亲对比胎儿)的组织或对应于健康细胞对比肿瘤细胞。“参考组织”可以对应于用于确定组织特异性甲基化水平的组织。来自不同个体的相同组织类型的多个样品可以用于确定该组织类型的组织特异性甲基化水平。

[0057] “生物样品”是指从受试者(例如,人,例如孕妇,患有癌症的人或疑似患有癌症的人,器官移植受体或疑似患有涉及器官(例如,心肌梗塞的心脏、卒中的大脑或贫血的造血系统)的疾病过程的受试者)中采集的并且含有一种或多种目标核酸分子的任何样品。生物样品可以是体液,如血液、血浆、血清、尿液、阴道液、水囊肿(例如睾丸)液、阴道冲洗液、胸膜液、腹水液、脑脊髓液、唾液、汗液、泪液、痰液、支气管肺泡灌洗液、乳头排出液、来自身体不同部位(例如甲状腺、乳房)的吸入液等。也可以使用粪便样品。在多个实施方案中,已富集游离DNA的生物样品(例如经由离心方案获得的血浆样品)中的大部分DNA可以是游离的(例如超过50%、60%、70%、80%、90%、95%或99%的DNA可以是游离的)。离心方案可以包括例如3,000g×10分钟,获得流体部分,并且以例如30,000g再离心10分钟以去除残留的细胞。

[0058] 如本文中所使用,术语“单倍型”是指在同一染色体或染色体区域上一起被传递的多个基因座处的等位基因的组合。单倍型可指少至一对基因座或染色体区域或整个染色体。术语“等位基因”是指在相同物理基因组基因座处的可选DNA序列,其可能会或可能不会导致不同的表型性状。在任何特定的二倍体生物体中,每个染色体有两个拷贝(男性人类受试者中的性染色体除外),每个基因的基因型包括该基因座上存在的一对等位基因,其在纯合子中相同,而在杂合子中不同。生物体的群体或物种通常在各个个体的每个基因座上包含多个等位基因。在群体中发现一种以上等位基因的基因组基因座称为多态性位点。基因座处的等位基因变异可以被测量为存在的等位基因的数目(即多态性程度)或群体中杂合子的比例(即杂合率)。

[0059] 如本文中所使用,术语“片段”(例如DNA片段)可以指多核苷酸或多肽序列中包含至少3个连续核苷酸的部分。核酸片段可以保留亲本多肽的生物活性和/或一些特征。核酸片段可以是双链的或单链的、甲基化的或未甲基化的、完整的或切割的、与其它大分子(例如脂质粒子、蛋白质)复合或未复合的。片段可以源自特定的组织类型,例如胎儿、肿瘤、移植器官等。

[0060] 术语“分析法(assay)”通常是指用于确定核酸性质的技术。分析法(例如第一分析

法或第二分析法)通常是指用于确定以下的技术:样品中的核酸数目、样品中核酸的基因组身份、样品中核酸的拷贝数变异、样品中核酸的甲基化状态、样品中核酸的片段尺寸分布、样品中核酸的突变状态或样品中核酸的片段化模式。本领域技术人员已知的任何分析法都可以用于检测本文中提及的核酸的任一种性质。核酸的性质包括序列、数目、基因组身份、拷贝数、一个或多个核苷酸位置处的甲基化状态、核酸尺寸、一个或多个核苷酸位置处核酸中的突变以及核酸的片段化模式(例如核酸片段所在的核苷酸位置)。术语“分析法”可以与术语“方法”互换使用。分析法或方法可以具有特定的灵敏度和/或特异性,并且可以使用ROC-AUC统计来测量其作为诊断工具的相对有效性。

[0061] “序列读取”通常是指从核酸分子的任一部分或全部测序的核苷酸串。例如,序列读取可以是存在于生物样品中的整个核酸片段。还例如,序列读取可以是核酸片段测序的短核苷酸串(例如,20-150个碱基)、在核酸片段的一个或两个末端处的短核苷酸串,或生物样品中存在的整个核酸片段的测序。成对序列读取可以与参考基因组比对,这可以提供片段的长度。序列读取可以通过多种方式获得,例如使用测序技术或使用探针,例如通过杂交阵列或捕获探针或扩增技术,如聚合酶链式反应(PCR)或使用单引物的线性扩增或等温扩增,或基于生物物理的测量(例如质谱)。序列读取可从单分子测序获得。“单分子测序”是指对单个模板DNA分子进行测序以获得序列读取,而无需解读来自模板DNA分子克隆副本的碱基序列信息。单分子测序可以对整个分子或仅部分DNA分子进行测序。可以对DNA分子的大部分进行测序,例如大于50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%。

[0062] “临床相关的”DNA的实例包括母体血浆中的胎儿DNA和患者血浆中的肿瘤DNA。另一个实例包括对移植患者血浆中移植物相关DNA的量的测量。另一个实例包括对受试者血浆中造血DNA和非造血DNA的相对量的测量。后一个实施方案可用于检测或监测或预测涉及造血和/或非造血组织的病理学过程或损伤。

[0063] “终止位置”或“末端位置”(或仅仅“末端”)可以指游离DNA分子,例如血浆DNA分子的最外碱基(即在末端处)的基因组坐标或基因组身份或核苷酸身份。末端位置可以与DNA分子的任一末端对应。以此方式,如果一端是指DNA分子的起点和末端,那么两个都可以对应于终止位置。在实践中,一个末端位置是通过分析方法检测或确定的游离DNA分子的一个末端上的最外碱基的基因组坐标或核苷酸身份,所述分析方法例如(但不限于)大规模平行测序或下一代测序、单分子测序、双链或单链DNA测序文库制备方案、聚合酶链式反应(PCR)或微阵列。此类体外技术可以改变游离DNA分子的真实体内物理末端。因此,每个可检测末端可以表示生物学上的真实末端或所述末端是一个或多个朝内的核苷酸或一个或多个从分子的原始末端延伸的核苷酸,例如非平末端双链DNA分子的悬突通过克列诺片段(Klenow fragment)的5'钝化和3'填充。末端位置的基因组身份或基因组坐标可以从序列读取与参考基因组如hg19或其他参考基因组的比对结果获得。其可以来源于表示人类基因组的初始坐标的索引或代码的目录号。其可以指通过(但不限于)靶标特异性探针、微测序、DNA扩增读取的游离DNA分子上的位置或核苷酸身份。

[0064] “优选末端”(或“经常性的终止位置”)是指在具有生理学(例如妊娠)或病理学状态(例如,疾病)的生物样品中不具有这类状态的生物样品或比在相同的病理学或生理学状态下的不同时间点或阶段(例如在治疗之前或之后)的生物样品中具有更高的代表性或

更普遍(例如,如由比率测量的)的末端。因此,相对于其它状态,优选末端在相关生理学或病理学状态中有增加的可能性或概率被检测到。增加的概率可以在病理学状态与非病理学状态之间,例如在患有癌症与无癌症的患者之间比较且定量为似然比或相对概率。似然比可以基于检测测试样品中的至少阈值数目的优选末端的概率或基于检测患有这类病况的患者相比于无这类病况的患者中的优选末端的概率来确定。似然比的阈值的实例包括(但不限于)1.1、1.2、1.3、1.4、1.5、1.6、1.8、2.0、2.5、3.0、3.5、4.0、4.5、5、6、8、10、20、40、60、80以及100。这类似然比可以通过比较具有和不具有相关状态的样品的相对丰度值来测量。因为相关生理学或疾病状态下检测到优选末端的概率较高,所以可以在具有相同生理学或疾病状态下的一个以上个体中发现这类优选终止位置。随着概率的增加,即使所分析的游离DNA分子的数目远小于基因组的尺寸,也可以检测到多于一个游离DNA分子在同一个优选终止位置上终止。因此,优选或经常性的终止位置也称为“频繁终止位置”。在一些实施方案中,定量阈值可用于要求同一个样品或同一个样品等分试样中末端至少多次(例如,3、4、5、6、7、8、9、10、15、20或50次)被检测认为是优选末端。相关生理学状态可以包括当个体健康、无疾病或未患相关疾病时的状态。类似地,“优选的终止窗”对应于优选终止位置的连续集合。

[0065] 在位置上终止的DNA分子的“比率”与DNA分子在该位置上终止的频率有关。该比率可以基于相对于分析的多个DNA分子归一化的在该位置上终止的多个DNA分子。因此,该比率对应于在一个位置上终止多少个DNA分子的频率,而与在该位置上终止的DNA分子的数目中具有局部最大值的位置的周期性无关。

[0066] “校准样品”可以对应于这样的生物学样品,其组织特异性DNA部分是已知的或通过校准方法确定的,例如使用对组织特异的等位基因确定的。作为另一个实例,校准样品可以对应于可以从其确定优选终止位置的样品。校准样品可以用于这两个目的。

[0067] “校准数据点”包括“校准值”和感兴趣的DNA(即,特定组织类型的DNA)的测量或已知比例分布。校准值可以是针对校准样品确定的相对丰度,为此已知组织类型的比例分布。校准数据点可以包括校准值(例如,使用尺寸标记的终点位置或识别方向的片段化来测量的)和已知的(测量的)组织类型的比例分布。校准数据点可以以多种方式定义,例如,作为离散点或校准函数(也称为校准曲线或校准表面)。校准函数可以从校准数据点的附加数学转换中得出。校准函数可以是线性的或非线性的。

[0068] “位点”(也称为“基因组位点”)对应于单个位点,其可以是单个碱基位置或一组相关碱基位置,例如,优选尺寸的位点,CpG位点或更大的一组相关碱基位置。“基因座”可以对应于包括多个位点的区域。基因座可以仅包括一个位点,这将使基因座在该上下文中等同于位点。

[0069] 哺乳动物基因组中的“DNA甲基化”通常是指在CpG二核苷酸中的胞嘧啶残基(即5-甲基胞嘧啶)的5'碳上添加甲基基团。在其他情况下,胞嘧啶中可以发生DNA甲基化,例如CHG和CHH,其中H为腺嘌呤,胞嘧啶或胸腺嘧啶。胞嘧啶甲基化也可以是5-羟甲基胞嘧啶的形式。还报道了非胞嘧啶甲基化,例如N6-甲基腺嘌呤。

[0070] 每个基因组位点(例如,CpG位点)的“甲基化指数”可以指显示该位点处的甲基化的DNA片段(例如,根据序列读取或探针确定的)占覆盖该位点的读取总数的比例。“读取”可以对应于从DNA片段获得的信息(例如,位点处的甲基化状态)。可以使用优先与特定甲基化

状态的DNA片段杂交的试剂(例如引物或探针)获得读取。通常,这种试剂在用取决于其甲基化状态而差别修饰或差别识别DNA分子的方法(例如亚硫酸氢盐转化,或甲基化敏感性限制性酶,或甲基化结合蛋白,或抗甲基胞嘧啶抗体)处理后施加。在另一个实施方案中,识别甲基胞嘧啶和羟甲基胞嘧啶的单分子测序技术可以用于阐明甲基化状态和确定甲基化指数。

[0071] 区域的“甲基化密度”可以指在显示甲基化的区域内的位点处的读取的数目除以覆盖该区域中的位点的读取的总数。这些位点可以具有特定的特征,例如是CpG位点。因此,区域的“CpG甲基化密度”可以指显示CpG甲基化的读取的数目除以覆盖该区域CpG位点(例如,特定CpG位点,CpG岛内的CpG位点或更大的区域)的读取的总数。例如,可以从亚硫酸氢盐处理后在CpG位点未转化的胞嘧啶总数(对应于甲基化的胞嘧啶)确定人类基因组中每100-kb堆栈的甲基化密度,所述CpG位点作为映射到该100-kb区域的序列读取所覆盖的所有CpG位点的一部分。还可以针对其他堆栈尺寸执行此分析,例如500bp、5kb、10kb、50kb或1-Mb等。区域可以是整个基因组或染色体或染色体的一部分(例如,染色体臂)。当区域仅包含CpG位点时,该CpG位点的甲基化指数与该区域的甲基化密度相同。“甲基化胞嘧啶的比例”可以指显示被甲基化(例如,在亚硫酸氢盐转化后未转化)的胞嘧啶位点“C”的数目,相比在该区域中所分析的胞嘧啶残基(即包括CpG范围之外的胞嘧啶)的总数。甲基化指数,甲基化密度和甲基化胞嘧啶的比例是“甲基化水平”的实例,其可能包括涉及位点的甲基化读取的计数的其他比率。除亚硫酸氢盐转化外,本领域技术人员已知的其他方法可用于询问DNA分子的甲基化状态,包括但不限于对甲基化状态敏感的酶(例如甲基化敏感的限制性酶)、甲基化结合蛋白、使用对甲基化状态敏感的平台进行的单分子测序(例如,纳米孔测序(Schreiber等人,Proc Natl Acad Sci 2013;110:18910-18915))以及通过Pacific Biosciences单分子实时分析(Flusberg等人,NatMethods 2010;7:461-465))。

[0072] “可识别甲基化的测序”是指允许在测序过程中确定DNA分子的甲基化状态的任何测序方法,包括但不限于亚硫酸氢盐测序或甲基化敏感的限制性内切酶消化后的测序,使用抗甲基胞嘧啶抗体或甲基化结合蛋白进行的免疫沉淀或允许阐明甲基化状态的单分子测序。“可识别甲基化的分析法”或“甲基化敏感分析法”可包括基于测序和非测序的方法,例如MSP,基于探针的询问,杂交,限制性内切酶消化然后进行密度测量,抗甲基胞嘧啶免疫测定,质谱询问甲基化胞嘧啶或羟甲基胞嘧啶的比例,免疫沉淀后不进行测序等。

[0073] 术语“测序深度”是指基因座被与基因座对齐的序列读取覆盖的倍数。基因座可以小至核苷酸,或与染色体臂一样大,或与完整基因组一样大。测序深度可以表示为50x、100x等,其中“x”是指基因座被序列读取覆盖的倍数。测序深度也可以应用于多个基因座或整个基因组,在此情况下,x可以分别指对基因座或单倍基因组或整个基因组进行测序的平均倍数。超深测序可以指测序深度是至少100x。

[0074] “分离值”(或相对丰度)对应于涉及两个值,例如DNA分子的两个量,两个分数贡献或两个甲基化水平,例如样品(混合物)甲基化水平和参考甲基化水平的差值或比值。分离值可以是简单的差值或比值。作为示例, x/y 的直接比值以及 $x/(x+y)$ 是分离值。分离值可以包括其他因素,例如乘法因素。作为其他示例,可以使用值的函数的差值或比值,例如,两个值的自然对数(\ln)的差值或比值。分离值可以包括差值和/或比值。

[0075] “相对丰度”是一种使在基因组位置的一个窗内终止的游离DNA分子的量(一个值)与在基因组位置的另一窗内终止的游离DNA分子的量(另一个值)相关联的分离值。两个窗

可以重叠,但是可以具有不同的尺寸。在其它实施方式中,两个窗不重叠。此外,窗可以具有一个核苷酸的宽度,并且因此等效于一个基因组位置。“分离值”和“相对丰度”是参数(也称为度量)的两个实例,其提供不同分类(状态)之间有差异的样品的测量,并且因此可用于确定不同的分类。

[0076] 如本文中所使用,术语“分类”是指与样品的特定性质相关的任何数字或其它字符。举例来说,“+”符号(或词语“阳性”)可以表示样品归类为具有缺失或扩增。分类可以是二元的(例如阳性或阴性)或具有更多分类等级(例如1到10或0到1的标度)。

[0077] 术语“截止值”和“阈值”是指操作中所使用的预定数目。举例来说,截断尺寸可以指一种尺寸,高于所述尺寸则排除片段。阈值可以是一种值,高于或低于所述值则适用特定分类,例如病况的分类,例如受试者是否患有病况或病况的严重程度。截止值或阈值可以是“参考值”,也可以从代表特定分类或在两个或更多个分类之间进行区分的参考值得出。如本领域技术人员将理解的,可以以多种方式来确定这种参考值,例如,在测试数据的输出之后并基于测试数据的输出来选择。例如,可以针对具有不同的已知分类的两个不同群体的受试者确定度量,并且可以选择参考值来代表一个分类(例如,平均值)或在度量的两个聚类之间的值。因此,具有一种或多种病况的已知分类和测得的特征值(例如,甲基化水平,统计尺寸值或计数)的参考受试者可用于确定参考水平,以区分不同的病况和/或病况分类(例如,受试者是否患有该病况)。作为另一示例,可以基于样品的统计模拟来确定参考值。这些术语中的任何一个都可以在任何这些上下文中使用。如本领域技术人员将理解的,可以选择截止值以获得期望的灵敏度和特异性。

[0078] 如本文中所使用,术语“染色体非整倍性”是指染色体的定量数目与二倍体基因组的定量数目存在变化。变化可以是增加或损失。它可以涉及整个一条染色体或染色体的区域。染色体区域可以对应于整个一条染色体,染色体的臂或更小的区域,例如50kb、500kb、1Mb、2Mb、5Mb或10Mb。

[0079] 如本文中所使用,术语“序列失衡”或“畸变”是指在临床相关染色体区域(即,被测试的区域)的量中的至少一个截止值所定义与参考量的任何显著偏差。序列失衡可包括染色体剂量失衡,等位基因失衡,突变剂量失衡,拷贝数失衡,单倍型剂量失衡和其他类似失衡。例如,当肿瘤的基因的一个等位基因缺失或基因的一个等位基因扩增或其基因组中的两个等位基因差异扩增时,就可以发生等位基因失衡,从而在样品的特定基因座处产生失衡。作为另一个实例,患者可以在肿瘤抑制基因中具有遗传突变。然后,患者可以继续发展为肿瘤,其中肿瘤抑制基因的未突变等位基因缺失。因此,在肿瘤内,存在突变剂量失衡。当肿瘤将其DNA释放到患者血浆中时,肿瘤DNA将与患者血浆中的组成DNA(来自正常细胞)混合。通过使用本文所述的方法,可以检测该DNA混合物在血浆中的突变剂量失衡。畸变可包括染色体区域的缺失或扩增。

[0080] 术语“癌症水平”(或更一般地,“疾病水平”,“病状水平”或“病况水平”)可以指是否存在癌症(即存在或不存在)、癌症的阶段、肿瘤尺寸、是否存在转移、身体的总体肿瘤负担、癌症对治疗的反应和/或癌症严重程度的其他度量(例如癌症复发)。癌症水平可以是数字(例如,概率)或其他标记,例如符号,字母和颜色。水平可以为零。癌症水平还可以包括恶化前或癌前病况(状态)。癌症水平可以多种方式使用。例如,筛查可以检查以前不知道患有癌症的某人是否存在癌症。评估可以调查被诊断出患有癌症的某人,以监测癌症随时间的

进展,研究治疗的有效性或确定预后。在一个实施方案中,预后可以表示为患者死于癌症的机会,或在特定持续时间或时间之后癌症进展的机会,或癌症转移的机会。检测可以意指“筛查”,也可以意指检查具有癌症暗示特征(例如症状或其他阳性检查)的某人是否患有癌症。多个实施方案可以确定针对肝癌,肺癌,胰腺癌,脑癌,结直肠癌,鼻咽癌,卵巢癌,胃癌和血液癌的癌症水平。

[0081] 术语“对照”,“对样品”,“参考”,“参考样品”,“正常”和“正常样品”可以互换使用,以大体上描述不具有特定病况,或在其他方面是健康的样品。在一个实例中,本文公开的方法可以在患有肿瘤的受试者上进行,其中参考样品是取自受试者健康组织的样品。在另一个实例中,参考样品是取自患有疾病(例如,癌症或癌症的特定阶段)的受试者的样品。可以从受试者或数据库获得参考样品。参考通常是指参考基因组,其用于对从受试者的样品进行测序获得的序列读取进行映射。参考基因组通常是指单倍体或二倍体基因组,来自生物样品的序列读取和组成样品可以与之进行比对和比较。对于单倍体基因组,每个基因座只有一个核苷酸。对于二倍体基因组,可以鉴定杂合基因座,这样的基因座具有两个等位基因,其中任一等位基因都可以允许与基因座比对的匹配。

[0082] 如本文中所示使用,短语“健康”通常是指具有良好健康的受试者。这样的受试者表现出不存在任何恶性或非恶性疾病。“健康个体”可以患有与被检疾病无关的其他疾病或病况,通常不被视为“健康”。

[0083] 术语“癌症”或“肿瘤”可以互换使用,并且通常是指异常的组织块,其中该块的生长超过正常组织的生长并且与其不协调。根据以下特征,可以将癌症或肿瘤定义为“良性”或“恶性”:细胞分化程度,包括形态和功能,生长比率,局部浸润和转移。“良性”肿瘤通常分化良好,典型的比恶性肿瘤生长缓慢,并且仍然局限于起源部位。另外,良性肿瘤不具有浸润,侵袭或转移到远处的能力。“恶性”肿瘤通常分化较差(发育不良),典型的快速生长,伴随着周围组织的进行性浸润,侵袭和破坏。此外,恶性肿瘤具有转移至远处的能力。“阶段”可用于描述恶性肿瘤的进展程度。与后期恶性肿瘤相比,早期癌症或恶性肿瘤与体内较少的肿瘤负荷相关,通常伴有较少的症状,更好的预后和更好的治疗效果。后期或晚期癌症或恶性肿瘤通常与远处转移和/或淋巴扩散有关。

[0084] 术语“假阳性”(FP)可以指个体未患病况。假阳性通常指个体未患肿瘤、癌症、癌前病况(例如癌前病灶)、局部或转移性癌症、非恶性疾病,或在其它方面是健康的。术语假阳性通常指个体未患病况,但通过本公开的分析法或方法鉴别为患有该病况。

[0085] 术语“灵敏度”或“真阳性比率”(TPR)可以指真阳性的数目除以真阳性和假阴性的数目的总和。灵敏度可以表征分析法或方法正确鉴定真正患有病况的群体的比例的能力。例如,灵敏度可以表征一种方法正确鉴定受试者在患有癌症的群体内的数目的能力。在另一实例中,灵敏度可以表征一种方法正确鉴定指示癌症的一种或多种标志物的能力。

[0086] 术语“特异性”或“真阴性比率”(TNR)可以指真阴性的数目除以真阴性和假阳性的数目的总和。特异性可以表征分析法或方法正确鉴定真正未患有病况的群体的比例的能力。例如,特异性可以表征一种方法正确鉴定受试者在未患有癌症的群体内的数目的能力。在另一实例中,特异性可以表征一种方法正确鉴定指示癌症的一种或多种标志物的能力。

[0087] 术语“ROC”或“ROC曲线”可以指受体操作特征曲线。ROC曲线可以是二元分类器系统的性能的图形表示。对于任何给定方法,ROC曲线可以通过在多种阈值设置下针对特异性

绘制灵敏度来产生。用于检测受试者中是否存在肿瘤的方法的灵敏度和特异性可以在受试者的血浆样品中各种浓度的肿瘤衍生的核酸下确定。此外,只要提供三种参数(例如灵敏度、特异性和阈值设定)中的至少一种,并且ROC曲线可以确定任何未知参数的值或预期值。未知参数可以使用针对ROC曲线拟合的曲线测定。术语“AUC”或“ROC-AUC”通常是指受体操作特征曲线下面积。这一度量考虑方法的灵敏度和特异性,可以提供方法的诊断效用的测量值。通常,ROC-AUC在0.5到1.0范围内,其中更接近0.5的值表明方法具有有限的诊断效用(例如较低的灵敏度和/或特异性),并且更接近1.0的值表明方法具有较大的诊断效用(例如较高的灵敏度和/或特异性)。参见例如Pepe等人,“Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker”, *Am. J. Epidemiol* 2004, 159 (9): 882-890, 其通过引用全文并入本文中。使用似然函数、优势比、信息理论、预测值、校准(包括拟合优度)和再分类测量表征诊断效用的其它方法是根据Cook, “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction”, *Circulation* 2007, 115: 928-935概述, 其通过引用全文并入本文中。

[0088] 术语“约”或“大约”可以意指在如由本领域技术人员测定的具体值的可接受的偏差范围内,其将部分取决于所述值如何测量或测定,即,测量系统的限制。例如,根据所属领域中的实践,“约”可以意指在1或大于1个标准差内。或者,“约”可以意指给定值的多达20%、多达10%、多达5%或多达1%的范围。或者,尤其相对于生物学系统或方法,术语“约”或“大约”可以意指在数目级内,在值的5倍内且更优选在2倍内。在本申请和权利要求书中描述特定值的情况下,除非另有说明,否则应当假设术语“约”意指处于该特定值的可接受误差范围内。术语“约”可以具有如本领域技术人员通常理解的含义。术语“约”可以指±10%。术语“约”可以指±5%。

具体实施方式

[0089] 人血浆中的无细胞DNA是非随机片段化的并反映了全基因组内的核小体组织。具体地,cfDNA分子具有与其来源组织有关的信息。引起来自特定组织的细胞死亡的病状导致来自受累器官的DNA的相对分布的扰动。这种来源的组织分析在开发用于癌症,产前测试和移植监测的液体活检中是特别有用的。因此,以同时的方式准确地确定对血浆DNA池有贡献的组织的相对贡献是有价值的。

[0090] 非随机片段化的各种新颖方面可以被确定并用于实际应用,例如生物学测量。例如,测量片段化,包括DNA片段末端的优选位置,与DNA片段尺寸的关系。这种关系可用于实际应用,例如测量特定组织类型(例如胎儿,肿瘤或移植组织)的比例贡献和检测特定组织类型的染色体区域中的序列失衡。作为另一个实例,测量片段化和组织特异性开放染色质区域的关系,包括DNA片段的哪一端(上游或下游)位于组织特异性开放染色质区域附近。上游末端相对于下游末端的定量模式可用于实际应用,例如测量特定组织类型的比例贡献和检测特定组织类型中的病状。

[0091] 为了进行尺寸分析,我们对血浆DNA的片段化模式进行了深入的研究,并探讨了片段化机制是否与血浆DNA的尺寸概况有关。因此,我们研究了这种优选末端位点是否可能与血浆DNA的片段长度有任何关系。我们称这些末端位点为‘尺寸标记的优选末端’。我们鉴定了优先与长和短血浆DNA分子相关的优选末端位点。短和长血浆DNA分子通常与不同的优选

DNA末端位点相关。我们发现,这些‘尺寸标记的’末端在胎儿DNA分数估计(比例贡献)和增强的非侵入性胎儿三体21(序列失衡)测试中显示出改进的准确性,因为孕妇的血浆显示出具有优选末端位点的非随机片段化。这种‘尺寸标记的’末端可用于其它组织类型(例如肿瘤或移植)以估计特定组织类型的比例贡献或检测序列失衡。

[0092] 进一步的分析显示胎儿和母体的优选末端是从核小体结构内的不同位置产生的。胎儿DNA经常在核小体核心内切割,而母体DNA大部分在接头区内切割。我们进一步证明胎盘细胞中的核小体可及性高于白细胞,说明母体血浆中胎儿DNA的切割位置和短小的差异。有趣的是,覆盖从短读取中挖掘的优选末端的血浆DNA分子通常比覆盖从长读取中挖掘的优选末端的那些短,甚至在非妊娠健康受试者中也是如此。因为这些后面的样品不含有胎儿DNA,所以数据表明优选DNA末端,染色质可及性和血浆DNA尺寸概况的相互关系可能是一般性的,延伸到妊娠的背景之外。因此,血浆DNA片段末端模式在生产机理上已经变得更加清楚,并且在基于血浆DNA的非侵入性分子诊断的未来发展中显示出效用。

[0093] 我们还研究了DNA片段末端的定位与核小体结构的关系。在开放染色质区域中,cfDNA分子显示特征性片段化模式,其反映为测序覆盖失衡和不同相位的片段末端信号。后者是指与cfDNA分子的上游和下游末端的方向相对应的序列的读取密度相对于参考基因组的差异。这种cfDNA片段化模式优先发生在组织特异性开放染色质区域,其中相应的组织有助于DNA进入血浆。这些信号的定量分析允许测量各种组织对血浆DNA池的相对贡献,以及检测特定组织类型中的病状。通过从孕妇,器官移植受体和癌症患者获得的血浆DNA测序数据来验证这些发现。因此,在非侵入性产前测试,器官移植监测和癌症液体活检中,识别方向的血浆DNA片段化分析具有诊断应用。

I. 片段化与技术综述

[0094] 已经证明血浆DNA不是随机片段化的。高分辨率血浆DNA尺寸概况显示在166bp处的主要峰和150bp以下的10bp周期性(9)。已经提出这种尺寸概况与核小体结构密切相关(9)。在这方面,核小体由4种核心组蛋白的八聚体(形成“核小体核心”,其被147bp的具有~10bp螺旋重复的DNA包裹),接头组蛋白和接头DNA(平均尺寸约20bp)组成(10)。此外,已经发现母体血浆中的胎儿DNA(主要来源于胎盘组织(11))比母体DNA(主要来源于造血系统)短(12-14)。胎儿和母体DNA分子的尺寸差异已被用于非侵入性产前测试,允许胎儿DNA分数估计,胎儿染色体非整倍性检测和胎儿甲基化组分析(15-19)。然而,对循环胎儿DNA的这种相对缩短的机理基础仍然了解很少(9,14,20)。

[0095] 最近的研究进一步探索了血浆DNA的终止模式。孕妇血浆DNA的超深度测序揭示了胎儿和母体特异性优选末端位点的存在(21)。尽管这些优选末端位点显示出非侵入性产前测试的潜力,但是它们存在的分子基础在很大程度上是未知的。此外,血浆DNA被认为是从凋亡细胞中释放的(22),这表明片段化模式与核小体结构和染色质状态相关(23-25)。

[0096] 在本公开中,我们展示存在无细胞DNA的非随机片段化方法。非随机片段化方法可以在一定程度上在不同类型的生物样品中进行,所述生物样品含有无细胞DNA,例如血浆、血清、尿液、唾液、脑脊髓液、胸膜液、羊膜液、腹膜流体和腹水流体。此外,非随机片段化发生在不同尺寸的DNA片段上。无细胞DNA以短片段形式天然存在。无细胞DNA片段化是指当产生或释放无细胞DNA分子时,高分子量DNA(如细胞的细胞核中的DNA)裂解、破坏或消化成短片段的过程。

[0097] 并非所有无细胞DNA分子都具有相同长度。一些分子比其它分子短。已经表明,无细胞DNA,例如血浆DNA,在开放染色质结构域,包括转录起始位点周围,和在核小体核心之间的位置,例如在接头位置,通常较短和较不完整,即具有较差的完整概率或较差的完整性(Straver等人,Prenat Diagn 2016,36:614-621)。每种不同的组织具有其特有的基因表达概况,其依次通过包括染色质结构和核小体定位的方式调节。因此,某些基因组位置的完整概率或完整性的无细胞DNA模式,例如,如血浆DNA的完整概率或完整性的无细胞DNA模式是那些DNA分子的组织来源的特征或标志。类似地,当疾病过程,例如癌症改变细胞的基因组的基因表达概况和功能时,来源于患病细胞的无细胞DNA完整概率概况将反映那些细胞。因此,无细胞DNA概况将为疾病存在提供证据或作为疾病存在的标志。

[0098] 一些实施方案进一步增强研究无细胞DNA片段化概况的分辨率。代替仅对一段核苷酸的读取求和以鉴定具有较高或较低完整概率或完整性的区域,我们研究个别无细胞DNA分子,尤其是血浆DNA分子的实际终止位置或末端。值得注意的是,我们的数据表明切割无细胞DNA分子的具体位置是非随机的。在体外剪切或超声处理的高分子量基因组组织DNA显示DNA分子具有随机分散在基因组中的终止位置。然而,存在在样品(如血浆)内大量呈现的无细胞DNA分子的某些终止位置。这类终止位置的出现或呈现数目在统计学上显著高于单独偶然所预期的。这些数据使我们理解无细胞DNA片段化的一个步骤超过完整性的区域变化的一个步骤(Snyder等人,Cell 2016,164:57-68)。这里,我们显示无细胞DNA片段化的过程甚至被协调到切割或剪切的特定核苷酸位置。我们将无细胞DNA终止位置的这些非随机位置称为优选终止位置或优选末端。

[0099] 在本公开中,我们显示存在无细胞DNA终止位置,其通常出现在不同生理学状态或疾病状态的个体中,并且出现在某些尺寸的片段上。例如,存在由短DNA片段(例如,60-155个碱基),长DNA片段(例如,170-250个碱基),妊娠和非妊娠个体共有的,由妊娠和癌症患者共有的,以及由患有和未患有癌症的个体共有的共同的优选末端。另一方面,主要仅在短DNA片段,长DNA片段,孕妇,仅在癌症患者或仅在未患癌症的非妊娠个体中存在优选末端。有趣的是,这些妊娠特异性或癌症特异性或疾病特异性末端在具有相当生理学或疾病状态的其它个体中也是高度代表性的。例如,在一例孕妇的血浆中鉴定的优选末端在其它孕妇的血浆中是可检测的。

[0100] 这种优选末端(例如对于短片段)的占比量与其它孕妇的血浆中的胎儿DNA分数相关。这种优选末端确实与妊娠或胎儿有关,因为它们的量在非妊娠血浆样品中显著减少。类似地,在癌症中,在一例癌症患者的血浆中鉴定的优选末端在另一例癌症患者的血浆中是可检测的。此外,这种优选末端(例如,对于短片段)的占比量可以与其它癌症患者的血浆中的肿瘤DNA分数相关。这种优选末端与癌症相关,因为它们的量在癌症治疗,例如手术切除后减少。

[0101] 存在许多用于分析无细胞DNA尺寸优选(尺寸标记的)末端的应用或效用。它们可以提供关于妊娠中胎儿DNA分数以及因此胎儿健康的信息。例如,已经报道了与孕龄匹配的对照妊娠相比,许多妊娠相关病症(例如,子痫前期,早产,子宫内生长限制(IUGR),胎儿染色体非整倍性等)与胎儿DNA的浓度分数的扰动(也称为胎儿DNA分数,胎儿分数或来自胎儿组织的比例贡献)相关。因此,可以从这种对照妊娠确定胎儿DNA的浓度分数的阈值。可以将测量的新样品中胎儿DNA的浓度分数与阈值进行比较以确定妊娠相关病症的分类。因此,使

用尺寸优选的末端测量胎儿DNA分数可用于这种妊娠相关病症。

[0102] 与短DNA片段相关的无细胞血浆DNA优选末端也可揭示血浆样品中的肿瘤DNA分数或浓度分数。知道肿瘤DNA分数提供了关于癌症阶段,预后的信息,并有助于监测治疗效果或癌症复发。

[0103] 通过比较在具有不同生理学或病理学状态(或不同尺寸的片段)的个体中的优选末端的无细胞DNA概况,例如非妊娠样品对比妊娠样品,癌症样品对比非癌症样品,或未患癌症的孕妇的概况对比非妊娠癌症患者的概况,可以鉴定与具体生理学状态或病理学状态(或不同尺寸的片段)相关的优选末端的目录。另一种方法是比较生理学(例如妊娠)或病理学(例如癌症)过程中的不同时间优选末端的无细胞DNA概况。这类时间点的实例包括妊娠之前和之后,分娩胎儿之前和之后,跨越妊娠的不同孕龄收集的样品,治疗癌症之前和之后(例如靶向疗法、免疫疗法、化学疗法、手术)、在癌症诊断之后的不同时间点、在发展癌症之前和之后、在发展转移之前和之后、在疾病严重程度增加之前和之后或在发展并发症之前和之后。

[0104] 当优选末端具有在一种生理学或病理学状态下被检测到的高可能性或概率(比率)时,可以将所述优选末端视为与所述生理学或疾病状态(或某尺寸的片段)相关。在其它实施方案中,优选末端具有相比于其它状态,更可能在相关生理学或病理学状态下被检测到的某一概率。由于检测到相关生理学或疾病状态下的优选末端的概率较高,这类优选或经常性末端(或终止位置)将见于超过一个具有所述相同生理学或疾病状态的个体中。高概率也将使得这类优选或反复性末端在相同个体的相同无细胞DNA样品或等分试样中可以检测多次。在一些实施方案中,可以设定定量阈值以限制被视为优选末端的在相同样品或相同样品等分试样内检测至少指定次数(例如5、10、15、20等)的末端的纳入。

[0105] 在对于任何生理学或病理学状态(或不同尺寸)建立无细胞DNA优选末端的目录之后,靶向或非靶向方法可以用于检测其在无细胞DNA样品(例如血浆)或其它个体中的存在以确定具有类似健康、生理学或疾病状态的其它测试个体的分类。无细胞DNA优选末端可以通过随机非靶向测序来检测。需要考虑测序深度,以便可以实现鉴定全部或一部分相关优选末端的合理概率。或者,可以对无细胞DNA样品进行具有高密度的优选末端的基因座的杂交捕捉,以在不限于通过测序、微阵列或PCR进行的检测之后富集具有这类优选末端的无细胞DNA分子的样品。然而,或者,基于扩增的方法可以用于特异性扩增和富集具有优选末端的无细胞DNA分子,例如反向PCR、滚环扩增。扩增产物可以通过测序、微阵列、荧光探针、凝胶电泳和本领域技术人员已知的其它标准方法鉴定。

[0106] 在实践中,一个末端位置可以通过分析方法检测或确定的无细胞DNA分子的一个末端上的最外侧碱基的基因组坐标或核苷酸身份,所述分析方法是如(但不限于)大规模平行测序或下一代测序、单分子测序、双链或单链DNA测序文库制备方案、PCR、用于DNA扩增(例如等温扩增)的其它酶促方法或微阵列。这类体外技术可以改变无细胞DNA分子的真实体内物理末端。因此,每个可检测末端可以表示生物学上的真实末端或末端是一个或多个向内的核苷酸或一个或多个从分子的原始末端延伸的核苷酸。举例来说,克列诺片段用于通过5'悬突的钝化和3'悬突的填充在DNA测序文库构建期间产生平末端双链DNA分子。尽管这类程序可以展示不与生物末端相同的无细胞DNA末端位置,但仍可以建立临床相关性。这是因为与具体生理学或病理学状态相关或有关的优选末端的鉴定可以基于将在校准样品和

测试样品中对无细胞DNA末端产生一致和可再现改变的相同实验室方案或方法原理。多种DNA测序方案使用单链DNA文库(Snyder等人,Cell 2016,164:57-68)。单链文库的序列读取的末端可以比双链DNA文库的末端更向内或进一步延伸。

[0107] 末端位置的基因组身份或基因组坐标可以来源于序列读取与受试者的参考基因组(例如hg19和其它人类参考基因组)的比对结果。其可以来源于表示人类基因组的初始坐标的索引或代码的目录。尽管末端是无细胞DNA分子的一个或两个末端处的核苷酸,但末端的检测可以通过识别血浆DNA分子上的其它核苷酸或其它核苷酸区段来进行。举例来说,具有通过能结合扩增子的中间碱基的荧光探针检测的优选末端的血浆DNA分子的阳性扩增。举例来说,末端可以通过能结合血浆DNA分子的中间部分上的一些碱基的荧光探针的阳性杂交来鉴定,其中已知片段尺寸。通过这种方式,可以通过算出多少碱基在具有已知序列和基因组身份的荧光探针外部来确定末端的基因组身份或基因组坐标。换句话说,末端可以通过检测相同血浆DNA分子上的其它碱基来鉴定或检测。末端可以是通过(但不限于)靶标特异性探针、微测序和DNA扩增读取的无细胞DNA分子上的位置或核苷酸身份。进一步的细节可以在PCT公开W02017/012592中找到,其通过引用并入用于所有目的。

II. 短片段和长片段的片段化

[0108] 进行血浆DNA尺寸和优选DNA末端位点的整合分析。观察到短DNA片段和长DNA片段的终止位置之间的差异,从而说明尺寸标记的优选末端。可以使用各种定义的短和长DNA片段,例如可以使用各种长度范围。例如,短DNA片段对应于具有小于长DNA片段范围的最小值和/或最大值的最小值和/或最大值的范围。尽管实例可与血浆一起使用,但也可使用其它无细胞样品,因为样品中的无细胞DNA也是自然的片段化过程的结果。

A. 尺寸标记的优选末端位点

[0109] 母体血浆中胎儿来源的DNA分子通常比母体来源的DNA分子短(9,14)。母体血浆中DNA分子的尺寸概况分析使用配对末端测序和与参考基因组的比对进行,尽管可以进行整个DNA片段的测序。我们将先前公开的两个母体血浆样品的血浆DNA配对末端测序数据(20)汇集在一起以获得总共约470倍的人倍体基因组覆盖。我们将血浆DNA读取分成SHORT(短)和LONG(长)类别,如本文所述。然后,使用基于泊松分布的统计模型,我们确定人基因组中的某些位置是否在SHORT和/或LONG类别中具有存在于血浆DNA分子的末端的显著增加的概率,如下所述。可以使用其它分布,例如二项式分布,负二项式分布,正态分布和伽玛分布。

[0110] 图1显示了根据本公开的实施方案,对血浆DNA片段的片段末端位点的分析。Set S和Set L分别包括短和长血浆DNA分子的优选末端位点。中间的重叠组110包括短和长血浆DNA分子的优选末端位点。如下面更详细描述,可以使用对具有对应于Set S的终止位置的无细胞DNA分子的定量测量来表征特定组织类型,例如,确定组织类型的比例贡献或组织类型的序列失衡。

[0111] 我们分别获得了SHORT和LONG类别的8,832,009和12,889,647个优选末端。在这些优选末端中,发现1,649,575个末端由两个类别共有。然后我们收集了仅出现在SHORT类别($n=7,182,434$)或LONG类别($n=11,240,072$)中的基因组的优选末端,并将它们分别定义为Set S和Set L。这两个组含有尺寸标记的优选末端位点。可以使用Set S和/或Set L的子集。

[0112] 可以对其它类别的受试者,例如,患有癌症或具有移植器官的受试者进行类似的

过程,这样的受试者具有的组织类型(例如,肿瘤或移植物)通常比来自健康组织的DNA片段短。然而,尺寸优选的终止位点可以在受试者的类别中重新使用。对于不同类别的受试者,可以使用短和长的不同定义。

B. 优选终止位点的鉴定

[0113] 为了进行胎儿分析,我们将先前公开的两个孕妇的血浆DNA测序数据(21)汇集在一起,这实现了总共约470倍的人单倍体基因组覆盖。然后根据DNA分子的尺寸将测序读取分成两类:一类读取在60bp-155bp的尺寸范围内(表示为SHORT),另一类读取在170bp-250bp的尺寸范围内(表示为LONG)。尺寸范围设置的精确选择可以涉及在两个类别中的表观胎儿DNA分数的差异和两个类别的数据的测序深度之间的折衷。结果,对应于约140倍和165倍人单倍体基因组覆盖的汇集数据的~30%和~35%的读取分别属于SHORT和LONG类别。收集这些读取并用于以下分析中。

[0114] 短DNA分子的其它实例包括70-145bp、80-145bp、90-145bp、80-135bp、90-135bp等。长DNA分子的其它实例包括160-210bp、160-220bp、160-230bp、160-240bp、180-260bp、160-260bp等。此外,所述范围可以重叠,例如,短为60-155bp且长为150-230bp,或短为90-185bp且长为170-250bp。在这种重叠情况下,第一尺寸范围仍然小于第二尺寸范围,因为第一尺寸范围的第一最大值小于第二尺寸范围的第二最大值。作为另一个实例,长片段可以是所有片段长度。

[0115] 对于每个尺寸类别中的读取,我们以全基因组的方式筛选所有核苷酸位置,以搜索显示血浆DNA分子末端显著过度代表的基因座。对于每个核苷酸位置,我们对血浆DNA末端的存在进行计数,并将结果与来自该位置周围位置的结果进行比较,例如使用1,000bp的窗,尽管也可以使用其它的窗尺寸,例如500bp或更大。窗可以具有在被分析的位置处的中心。

[0116] 将计算基于泊松分布的p值,以确定特定位置是否具有作为读取的末端的显著增加的概率,即优选末端位点:

$$P值 = \text{泊松}(N_{实际}, N_{预测})$$

其中泊松()是泊松概率函数, $N_{实际}$ 是在特定核苷酸(基因组位置)终止的分子的实际数目, $N_{预测}$ 是相邻1,000bp的窗(例如以特定核苷酸为中心)内的读取的总数除以该窗中的DNA片段的平均片段尺寸(或通常在样品中的DNA片的平均尺寸)。在各种实例中,当整个片段在窗内或仅当片段部分地在窗内时,读取可以被定义为在窗内。在其它实施方式中,基因组位置的 $N_{预测}$ 可以是覆盖该位置的读取数目除以平均或预期片段尺寸。因此,实施方式可以确定全局参数,并将所有位点与全局参数进行比较,而不是与局部窗进行比较。 $N_{预测}$ 是用于确定在位置上终止的短(或长)DNA分子的比率是否高于阈值(例如,确定是否与参考值存在统计学显著差异)的参考值(参考比率)的实例。这些实例举例说明了使用以特定基因组位置为中心的窗处终止的多个DNA片段除以无细胞DNA分子的平均尺寸所确定的参考值。

[0117] 可以使用Benjamini方法进一步调整p值。使用 <0.01 的p值来指示统计学上显著的末端位点。这样的p值是用于确定在所述位置终止的无细胞DNA分子的比率是否足够高从而被认为是优选末端的阈值的实例。

[0118] 在其它实例中,可以追踪在一位置终止的短DNA分子的相对量,并且可以确定分布中的峰,例如,如稍后的图所示。对峰的追踪有效地比较了相对于在其它位置终止的数目

(其充当参考值)的在一位置终止的短DNA分子的数目。

[0119] 根据上述实例和本文的其它实例,参考值(也称为参考比率)可以从在特定基因组位置之外的基因组位置(或该位置周围的小窗)终止的第二多个无细胞DNA分子的数目来确定。以这种方式,可以确定更多的DNA片段以统计学上显著的量在特定位置而不是其它位置周围(例如,在该特定位置周围)终止。这将包括相对于在特定基因组位置周围的窗内的基因组位置处终止的DNA片段的数目鉴定在峰处的特定基因组位置。

[0120] 因此,在多个实例中,可以按以下方式鉴定某尺寸(例如短的)无细胞DNA分子的末端的出现率高于阈值的第一组基因组位置。第一组织类型可以与短DNA片段相关,因此也可以与短DNA片段的优选终止位置相关。可以按与测试样品类似的方式分析校准样品,其中已知两份相同类型的样品(例如血浆、血清、尿液等)和校准样品包括第一组织类型(例如来自孕妇的样品的胎儿组织或HCC患者的肝脏的肿瘤组织)。可以比较在基因组窗(例如宽度是一个或多个)中终止的无细胞DNA分子的数目与参考值以确定终止位置的比率是否超过所述位置的阈值。在一些实施方案中,如果比率超过参考值,那么当对应数目超过参考值时,第一基因组窗内的基因组位置中的每一个可以鉴定为具有高于阈值的比率。这类方法可以鉴定优选的终止窗,其包括优选的终止位置。

[0121] 参考值可以使得仅前N个基因组窗具有高于阈值的比率。举例来说,第一组基因组位置可以具有关于对应数目的最高N值。作为实例,N可以是至少10,000;50,000;100,000;500,000;1,000,000;或5,000,000。

[0122] 作为另一实例,参考值可以是根据样品中无细胞DNA分子的概率分布和平均长度,在基因组窗内终止的无细胞DNA分子的预期数目。可以使用相应数目和预期数目确定p值,其中阈值对应于截止p值(例如0.01)。p值小于截止p值指示比率高于阈值。作为另一个实例,参考值可以包括从被鉴定为具有减少量的第一组织类型的样品中测量的在基因组窗内终止的无细胞DNA分子的数目。

III. 尺寸标记的优选末端位点的胎儿使用

[0123] 优选的终止位点可用于测量临床相关DNA,例如胎儿DNA,肿瘤DNA或供体DNA,其具有与健康DNA不同的片段化模式。优选的终止位点可以从来自临床相关样品的历史数据集中挖掘出来。对随后的样品或样本的技术实践可以基于搜索在每个测试样品中的那些优选终止位点的存在或不存在或对其进行定量。本节描述了尺寸标记的优选末端位点在非侵入性产前测试中的应用。

[0124] 为了研究尺寸标记的优选末端位点在非侵入性产前测试中的潜在应用,我们重新分析了一个母体血浆DNA测序数据集,我们以前从26名早孕期孕妇中产生了该数据集(21)。对于每一种情况,我们分别检查在Set S和Set L优选末端上终止的读取。

[0125] 图2显示了在24个母体血浆样品中覆盖Set S优选末端位点的血浆DNA读取的尺寸分布(红色)对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布(蓝色)。X-轴表示片段尺寸(bp),Y-轴表示频率(%)。我们观察到,对于所有这些情况,覆盖Set S优选末端位点的血浆DNA读取比覆盖Set L优选末端位点的那些短。

[0126] 图3显示了根据本公开的实施方案,在一种母体血浆样品中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。对于图2,X轴表示片段尺寸(bp),Y轴表示频率(%)。覆盖Set S末端位点的读取的尺寸分布具有明确定义的周期性,其中峰和谷在尺寸为

约80bp至约150bp的峰之间。每个峰大约每10bp。

A. 确定胎儿分数

[0127] 图4A显示了26个母体血浆样品中具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)与胎儿DNA分数之间的相关性。相对丰度可以通过计数在Set S位点之一终止的无细胞DNA分子的第一数目并除以在Set L位点之一终止的无细胞DNA分子的第二数目来确定。每个校准数据点405对应于其相对丰度和胎儿DNA分数被确定的差异样品。胎儿DNA分数可以使用胎儿特异性标志物,例如父系特异性等位基因,Y染色体标志物或胎儿特异性表观遗传标志物,例如甲基化来确定。

[0128] 观察到血浆DNA的相对丰度(Set S对比Set L优选末端位点[表示为S/L比值])和胎儿DNA分数($R=0.79, P<0.001$, 皮尔森相关性)呈正相关。可使用相对丰度的其它值,例如,第一数目除以第一数目与第二数目之和,或第一数目除以所有读取。也可使用分离值的其它实例,例如,如上文术语部分中所定义的。

[0129] 为了确定新样品的胎儿DNA分数,系统可以确定与其它无细胞DNA分子(例如,在一组长优选末端位置终止的无细胞DNA分子)相比,在一组短优选末端位置终止的无细胞DNA分子的相对丰度。然后,可以将新测量的相对丰度与一个或多个校准数据点405进行比较。例如,校准函数410可以拟合校准数据点405,其中新测量的相对丰度可以用作校准函数410的输入,校准函数410提供胎儿DNA分数的输出。可以以类似的方式测量其它组织类型的比例贡献。

[0130] 值得注意的是,该R值高于使用基于SNP的方法挖掘的优选末端位点获得的R值(其为0.66)(21)。值得注意的是,尺寸标记的优选末端位点的挖掘不需要关于胎儿母体遗传多态性的知识。另一方面,我们的小组先前已经证明,单独的尺寸信息可以指示血浆DNA中的胎儿DNA分数(17)。因此,我们在没有选择具有特定末端的分子的情况下计算了母体血浆DNA的尺寸比,并评估了其于胎儿DNA分数的关系。

[0131] 图4B显示了26个母体血浆样品的尺寸比(短读取比长读取的数目)和胎儿DNA分数之间的相关性。尺寸比与胎儿DNA分数呈正相关($R=0.67, P<0.001$, 皮尔森相关)。尽管R值与先前研究的R值(17)相当,但它低于基于尺寸标记的优选末端的相关性。总之,结果表明,尺寸标记的优选末端允许改善血浆DNA中胎儿DNA分数的估计。

[0132] 因此,使用短DNA分子的优选末端位置可以通过将相对丰度与从一个或多个校准样品确定的一个或多个校准值进行比较来提供胎儿组织的比例贡献的分类,所述校准样品的胎儿组织的比例贡献是已知的。如本文所述,分类可以是特定的百分比或百分比范围。对于其它组织类型,例如肿瘤组织,分类可以是是否测量到任何肿瘤组织或至少可评估的量(例如,高于检测的最小阈值)。

[0133] 在一些实施方案中,尺寸标记的优选终止位置可以延伸到包括相邻的核苷酸。因此,一组短的优选终止位置可以包括扩展的终止位置Set S。在任一种情况下,可以使用第二数目的DNA片段(其中至少一些DNA片段在短优选集合之外的位置终止)将以短优选位置(Set S或扩展的Set S)终止的多个DNA片段归一化以获得相对丰度。第二数目可以包括短优选集合的第一数目。在一个实例中,基于窗的相对丰度(例如,比值)可以在窗A内终止的片段的数目(较小的)和在该窗之外终止的那些片段的数目之间或者在短的优选终止位置周围的较大窗B内终止的那些片段的数目之间取得,因此包括一些非优选位置。可以调整窗

A和窗B的尺寸,以获得所需的性能。可以通过实验获得不同窗尺寸的性能。可以设定窗A的尺寸,例如但不限于2bp、3bp、4bp、5bp、6bp、7bp、8bp、9bp、10bp、15bp、20bp、25bp和30bp。窗B的尺寸将大于窗A的尺寸并且可以被设定,例如但不限于20bp、25bp、30bp、40bp、50bp、60bp、70bp、80bp、100bp、120bp、140bp、160bp、180bp和200bp。

B. 胎儿非整倍性检测

[0134] 此外,我们研究了尺寸标记的优选末端位点是否可用于检测胎儿组织中的染色体区域的序列失衡,例如,用于检测拷贝数畸变。在尺寸标记的优选末端位点终止的DNA分子比随机选择任何DNA片段具有更高的来自胎儿的可能性。胎儿DNA的这种富集可以提高用于进行非侵入性产前测试的技术的准确性。作为实例,这种技术可以使用一定量的在短的优选末端位点终止的无细胞DNA分子,以及这种无细胞DNA分子的尺寸分布或甲基化水平的统计值,然后可以将其与参考值进行比较。

[0135] 为此,我们研究了尺寸标记的优选末端位点是否可以改善胎儿21三体的非侵入性产前测试。为此,我们从我们以前的研究中收集了包含36例21三体病例和108例对照病例的数据集(17)。我们利用覆盖Set S优选末端的读取进行该分析。值得注意的是,在这些样品中具有Set S优选末端的读取的中值数是133,702(范围:52,072-353,260)。

[0136] 一些实施方式可以使用基于Z分数的方法(26)将映射到chr21的第一数目的这种读取通过具有映射到所有常染色体的Set S优选末端的第二数目的读取归一化,以获得参数值,其可以与区分两个分类的参考值进行比较。在这种情况下,可以从整倍体病例确定参考值,其标准偏差为3或其它合适的偏差。因此,可以从对照样品确定参考值。由于可以分析不同数目的DNA分子,因此归一化可以解释样品(例如测试样品和对照样品)尺寸的差异。任何合适的归一化技术可以用于任何组织类型的任何应用,例如,通过分析跨样品的相同数目的序列读取。

[0137] 用于基于计数的技术的其它参数值可包括各种比值,其涉及第一数目,例如用于某区域的S/L比值除以一个或多个参考区域的第二数目(例如,S/L比值)。一个或多个参考区域可包括预期不具有序列失衡(例如,具有两个染色体拷贝)的至少一个其它区域。仅在短的优选末端终止的DNA片段的使用是富集胎儿DNA,并因此获得更高的准确性的方式,例如,因为胎儿DNA将是样品的更大百分比,并且将发生与参考值的更大百分比偏差。

[0138] 图5A显示了根据本公开的实施方案,在对照病例与21三体病例之间的chr21读取的相对丰度的比较。在该分析中仅考虑覆盖Set S优选末端位点的读取(中值读取数目:133,702)。如图5A所示,21三体病例与对照病例相比显示出显著升高的归一化的具有Set S优选末端的chr21读取($P < 0.001$, 曼-惠特尼秩和检验)。

[0139] 图5B显示了根据本公开的实施方案,对于21三体测试,覆盖Set S优选末端位点的读取与随机读取之间的ROC比较。随机读取分析仅使用任何读取,而不是针对优选的末端位点进行过滤。使用接收器操作特征(ROC)曲线分析,我们得到曲线下面积(AUC)值为0.97。为了在读取数目方面实现公平的比较,我们通过随机选择与覆盖Set S优选末端位点的读取相同数目的读取并重新计算降低采样数据集中的归一化chr21读取数目来降低采样每个样品的测序数据。结果,在21三体检测中,与覆盖Set S优选末端位点的读取相比,随机读取显示较低的AUC值(0.93) ($P = 0.033$, DeLong检验(27);图5B)。这些结果表明,Set S优选末端位点可以潜在地增强设计用于利用其特征的测定中的21三体测试(参见讨论)。

[0140] 除了由染色体拷贝的缺失或扩增引起的胎儿非整倍性之外,还可以检测其它拷贝数畸变,例如特定区域的扩增或缺失。例如,可以检测数个Mb的微缺失或微扩增。这种序列失衡发生在两个单倍型之间,例如,重复的单倍型导致其被过度代表或单倍型的缺失导致其代表不足。

C. 胎儿基因型测定

[0141] 考虑到短的优选末端位置可以与特定组织类型相关,在这些优选终止位置处终止的无细胞DNA分子具有来自该组织(例如胎儿,癌症或移植物)的高可能性。在一些情况下,无细胞DNA混合物中的特定组织类型相对于其它组织类型在特定基因组位置可以具有不同的基因型。例如,胎儿组织或肿瘤组织可以具有不同的基因型。由于在短优选位点终止的无细胞DNA分子具有来自目的组织类型的高可能性,所以可以分析在该位置终止的无细胞DNA分子以确定该位置的组织类型的基因型。以这种方式,尺寸优选的终止位置可以用作过滤器以鉴定来自该组织类型的DNA。

[0142] 关于无细胞DNA片段的尺寸优选的终止位置的信息(例如,从血浆测序)可以用于确定哪一个母体等位基因已经被来自孕妇的胎儿遗传。这里,我们使用一个假设的实例来说明该方法的原理。我们假定母亲,父亲和胎儿的基因型分别是AT,TT和TT。为了确定胎儿基因型,我们需要确定胎儿是从母亲遗传了A还是T等位基因。我们以前已经描述了一种称为相对突变剂量(RMD)分析的方法(Lun等人,Proc Natl Acad Sci USA 2008;105:19920-5)。在该方法中,比较母体血浆中两个母体等位基因的剂量。如果胎儿已经遗传了母体T等位基因,那么胎儿将是T等位基因的纯合子。在这种情况下,与A等位基因相比,母体血浆中的T等位基因将被过度代表。另一方面,如果胎儿已经从母亲遗传了A等位基因,则胎儿的基因型将是AT。在这种情况下,A和T等位基因在母体血浆中将以大致相同的剂量存在,因为母亲和胎儿对AT而言都是杂合的。因此,在RMD分析中,将比较母体血浆中两个母体等位基因的相对剂量。

[0143] 可以分析读取的终止位置以提高RMD方法的准确性。例如,可以对读取进行过滤,以仅包括在短的优选位点终止,并覆盖正被基因分型的位置的那些读取。

[0144] 在说明性实例中,在短的优选终止位置上终止的两个分子携带T等位基因(例如,在优选的终止位置或在被两个相应的读取覆盖的附近位置)。在一个实施方案中,当只有两个在短优选终止位置终止的分子被用于下游分析时,胎儿基因型将被推断为TT。因此,仅T相关读取的序列失衡(或高百分比,例如大于70%)可指示纯合基因型。序列平衡(例如,任一等位基因小于60%)可指示杂合基因型。

[0145] 在另一实施方案中,携带T等位基因的两个胎儿来源的分子在RMD分析中被赋予较高的权重,因为这两个分子在短的优选终止位置终止。可以将不同的权重给予在短的优选终止位置终止的分子,例如但不限于1.1、1.2、1.3、1.4、1.5、2、2.5、3和3.5。

[0146] 作为实例,用于确定基因座是否杂合的标准可以是两个等位基因的阈值,每个等位基因出现在至少预定百分比(例如,30%或40%)的与该基因座对齐的读取中。如果一个核苷酸以足够的百分比(例如,70%或更高)出现,则可以确定该基因座在特定组织中是纯合的。

[0147] 可以对患有肿瘤的受试者进行类似的技术。可以鉴定和分析在短优选终止位置终止的无细胞DNA分子。可以为该组中的每种无细胞DNA分子确定与该位置(或由DNA片段覆盖

的附近测试位置)对应(例如对齐)的碱基,并且可以为每种碱基计算总碱基的百分比。例如,可以测定在无细胞DNA分子上观察到的测试位置处的C的百分比,所述无细胞DNA分子在所述位置终止。如果在受试者的健康组织中未观察到C,则如果鉴定到足够数目的C,例如,高于阈值数目,这可取决于样品中测量的肿瘤DNA分数,则可将C鉴定为突变。

D. 健康受试者对比妊娠受试者的尺寸标记的优选末端

[0148] 上述分析表明,Set S优选末端位点确实反映了胎儿来源的DNA的片段化模式。然而,这些末端位点是从胎儿和母体DNA分子的混合物中挖掘的。因此,为了测试这些优选末端位点是否仅反映胎儿特异性片段化模式,我们从我们组的先前研究(28)中检索了包含32个健康(非妊娠的)受试者的数据集,并在这些样品中搜索携带Set S优选末端位点的血浆DNA读取。有趣的是,具有Set S优选末端位点的一些血浆DNA读取确实存在于健康受试者的血浆中,并且这种血浆DNA分子也比覆盖Set L优选末端位点的那些短。

[0149] 图6显示在24名健康受试者中,覆盖Set S优选末端位点的血浆DNA读取的尺寸分布对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布。红色和蓝色线是分别覆盖Set S和Set L优选末端位点的读取。X-轴表示片段尺寸(bp),Y-轴表示频率(%)。在Set S优选末端位点终止的无细胞DNA分子平均比在Set L终止的那些短。

[0150] 图7A显示了根据本公开的实施方案,在健康受试者中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。图7A显示了具有典型尺寸分布的情况。

[0151] 图7B显示了根据本公开的实施方案,在孕妇和健康受试者中具有Set S优选末端位点的血浆DNA读取对比具有Set L优选末端位点的血浆DNA读取的相对丰度(S/L比值)的比较。这些健康受试者与孕妇相比显示出较低的S/L比。因此,相对于其它终止位置组,例如Set L或整个基因组,在Set S终止的读取具有增加的胎儿DNA比例。

[0152] 这表明S/L用于序列失衡检测的增加准确性的参数值是可行的,例如,当相对于一个或多个参考区域的S/L归一化时。更一般地,Set S终止位置可以用作过滤器以仅使用某些鉴定的DNA分子,导致胎儿DNA的富集。在一个区域(胎儿DNA富含的)内的Set S处终止的DNA分子可用于检测胎儿DNA是否存在序列失衡。作为实例,参数值可包括测试区域的S/L与一个或多个参考区域的S/L的比值,或仅包括在测试区域的短优选末端终止的DNA分子的第一数目与在一个或多个参考区域的短优选末端终止的DNA分子的第二数目的比值。

[0153] 因此,数据表明,尺寸标记的优选末端位点是血浆中短和长DNA分子的一般足迹,而不管它们的起源(例如胎儿对比母体)如何。此外,与母体DNA相比,胎儿DNA分子显示较高比例的覆盖Set S优选末端位点的分子。因此,测试区域和一个或多个参考区域的S/L值的比值可以被用作参数值,该参数值与参考值进行比较以区分序列失衡的分类。

IV. 尺寸标记的优选末端位点的肿瘤使用

[0154] 可以对包括肿瘤DNA的样品进行类似的测量,如以下数据所示。例如,可以确定无细胞样品中肿瘤DNA的比例贡献,或者可以确定序列失衡。

A. 肿瘤DNA片段化

[0155] 图8显示了根据本公开的实施方案,在肝细胞癌(HCC)患者中覆盖Set S和Set L优选末端位点的血浆DNA读取的尺寸分布。X-轴表示片段尺寸(bp),Y-轴表示频率(%)。图8显示了具有典型尺寸分布的情况。尽管HCC被用作测试病例,但其它癌症也表现出短的无细胞DNA片段,因此该技术同样适用于其它类型的癌症。

[0156] 图9显示了在代表性的一组24例肝细胞癌患者中,覆盖Set S优选末端位点的血浆DNA读取的尺寸分布对比覆盖Set L优选末端位点的血浆DNA读取的尺寸分布。红色和蓝色线分别是覆盖Set S和Set L优选末端位点的读取。X-轴表示片段尺寸(bp),Y-轴表示频率(%)。总之,分析了90名HCC患者,其中90名患者具有与图9所示相似的尺寸分布。

B. 测定肿瘤分数

[0157] 图10显示了根据本公开的实施方案,在血浆中具有大于1%的肿瘤DNA分数的72例肝细胞癌患者中,具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)与肿瘤DNA分数之间的相关性。使用与图1相同的Set S和Set L位点。观察到具有Set S优选末端位点的血浆DNA与具有Set L优选末端位点的血浆DNA的相对丰度[表示为S/L比值]和肿瘤DNA分数之间正相关($R=0.58, P<0.001$, 皮尔森相关)。

[0158] 图10显示了与图4A类似的行为。例如,相对丰度可以通过计数在Set S位点之一终止的无细胞DNA分子的第一数目并除以在Set L位点之一终止的无细胞DNA分子的第二数目来确定。每个校准数据点1005对应于其相对丰度和胎儿DNA分数被确定的差异样品。可以使用肿瘤特异性标志物,例如肿瘤特异性等位基因,例如杂合性丧失(LOH)来确定肿瘤DNA分数。

[0159] 如同胎儿测量一样,为了确定新样品的肿瘤DNA分数,系统可以确定与其它无细胞DNA分子(例如,在一组长优选末端位置终止的无细胞DNA分子)相比,在一组短优选末端位置终止的无细胞DNA分子的相对丰度。然后,可以将新测量的相对丰度与一个或多个校准数据点1005进行比较。例如,校准函数1010可以拟合校准数据点1005,其中新测量的相对丰度可以用作校准函数1010的输入,校准函数1010提供肿瘤DNA分数的输出。

[0160] 组织类型(例如肿瘤组织)的比例贡献的分类可对应于除百分比或百分比范围以外的值。例如,分类可以对应于癌症的检测,更具体地,对应于肿瘤负荷。

[0161] 图11显示了健康受试者和肝细胞癌患者之间具有尺寸标记的优选末端位点的血浆DNA分子的相对丰度(S/L比值)。基于血浆中的肿瘤DNA分数,将肝细胞癌患者分成4组。S/L比值越高,肿瘤负荷越高。这4组对应于肿瘤DNA分数的不同百分比范围。<1组中的下降是由于小肿瘤,使得周围坏死组织中较长的DNA超过来自肿瘤的短DNA。

[0162] 因此,该分类可以是是否测量到任何肿瘤组织或者或至少可评估的量(例如,高于检测的最小阈值)。因此,比例贡献的分类可以是检测到癌症。根据灵敏度或特异性,实施方案可以使用约0.5、0.51、0.52或0.53的检测阈值作为实例。

[0163] 可以使用相对丰度的其它值(除了比值S/L之外),例如,如以上针对用于确定胎儿分数所描述的。例如,归一化可以使用所获得的读取的总数,这将包括在任何短优选窗之外的位置终止的读取。这样的总数是包括不在短的优选位置终止的读取的第二数目的读取的实例。分析从一个样品到另一个样品的相同数目的读取提供了与通过读取总数或其它第二数目进行归一化相同的结果,因此这种归一化被包括在内。

C. 检测肿瘤引起的序列失衡

[0164] 在肿瘤组织的染色体区域也可以检测到序列失衡。例如,扩增和缺失通常发生在肿瘤组织中。因此,序列失衡将发生,并导致一个单倍型相对于另一个单倍型过度代表。可以在不同尺寸的区域(例如染色体臂)的多个区域(例如,所有相同的尺寸,例如1Mb)中测试这样的拷贝数畸变。

[0165] 在下面的实例中,为了检测来自患有肿瘤的受试者的无细胞样品中的序列失衡,研究了染色体区域1p、1q、8p和8q,因为已知它们经常患有HCC中的CNA。在这些区域之一的短优选位置终止的第一数目的无细胞DNA分子可以用作检测该区域中序列失衡的参数值。在一个或多个参考区域的短优选位置终止的第二数目的无细胞DNA分子,可用于使第一数目归一化,例如,使得可考虑样品的尺寸。一个或多个第二区域可以已知或假定没有序列失衡。

[0166] 在下面的实例中,一个或多个参考区域包括所有常染色体,并因此包括在常染色体的短优选位点终止的所有DNA片段。因此,所有常染色体被组合用作对照,以便在Set S位置之一终止的读取的计数归一化。可以将特定位置组(例如,Set S)终止的DNA分子的归一化计数与参考值(例如,当不存在序列失衡时的期望值)进行比较,这可以包括与截止值进行比较以确定是否存在与参考值的统计学上显著的偏差。

[0167] 图12显示了根据本公开的实施方案,在健康受试者,患有或没有肝硬化的HBV携带者和HCC患者中覆盖chr1p上的Set S末端的归一化读取计数。图12显示了每一类受试者的框图,其中中间显示为条形,上四分位数和下四分位数显示为突出(whisker)。每个数据点对应于给定样品的chr1p区域的归一化读取计数,其中样品在四个类别中的一个中。归一化的读取计数可以被确定为在chr1p区域中的Set S末端之一处具有末端位置的读取的数目除以在Set S末端之一处具有末端位置的读取的总数。

[0168] 还并入了拷贝数畸变信息,因为某些样品被标记为表现出增加(扩增),损失(缺失)或正常。通常,期望在非癌症受试者中有相对较少的畸变,尽管在患有肝硬化的HBV受试者中很少,肝硬化可能是HCC的前兆。如图所示,具有拷贝数损失的区域通常具有低于中值的值。与中值的足够偏差或特定百分比值的偏差可以用作阈值或参考值,以确定该区域存在的序列失衡。使用(28)来确定区域的增加和损失。

[0169] 图13显示了根据本公开的实施方案,在健康受试者,患有或没有肝硬化的HBV携带者和HCC患者中覆盖chr1q上的Set S末端的归一化读取计数。还并入了拷贝数畸变信息(增加,损失或正常)。图13显示了与图12类似的图,但是拷贝数增加是chr1q的主要畸变,这与chr1p主要是损失相反。

[0170] 图14显示了根据本公开的实施方案,在健康受试者,患有或没有肝硬化的HBV携带者和HCC患者中覆盖chr8p上的Set S末端的归一化读取计数。还并入了拷贝数畸变信息。图14显示了与图12类似的图,其中拷贝数损失是chr8p的主要畸变。

[0171] 图15显示了根据本公开的实施方案,在健康受试者,患有或没有肝硬化的HBV携带者和HCC患者中覆盖chr8q上的Set S末端的归一化读取计数。还并入了拷贝数畸变信息。图15显示了与图12类似的图,但是拷贝数增加是chr8q的主要畸变,这与chr1p主要是损失相反。

[0172] 如第III.C节所述,序列失衡可涉及确定组织的基因型。可以鉴定出一组在短的优选位点终止的DNA分子,例如,通常对应于肿瘤DNA片段。可以分析被鉴定组的DNA片段覆盖的给定基因座上的等位基因以确定该基因座上的基因型。例如,可以确定具有第一等位基因的组中的第一数目的DNA片段与具有第二等位基因的组中的第二数目的DNA片段之间的差值或比值。差值或比值是所鉴定的无细胞DNA分子的组的值的实例。可以将该值与参考值进行比较以确定是否存在序列失衡,例如,如果不存在序列失衡,则该基因型对于肿瘤组织

中的两个等位基因是杂合的,并且当存在序列失衡时,该基因型对于优势等位基因(可能是组中仅有的等位基因)是纯合的。

V. 染色质中终止位点的位置

A. 尺寸标记的优选末端位点的基因组注释

[0173] 为了探索在基因组中如何产生尺寸标记的优选末端位点,我们分别研究了Set S和Set L中任何两个最接近的优选末端位点之间的分离(以bp为单位)。

[0174] 图16显示了根据本公开的实施方案,Set S和Set L优选末端位点中的任意两个最接近的优选末端位点之间的距离分布。该距离在Set S数据的最接近的S位点之间,以及该距离在Set L数据的最接近的Set L位点之间。对于Set S优选末端位点,存在高达~150bp的强10bp周期性。另一方面,对于Set L优选末端位点,在~170bp处存在一个峰,而没有观察到10bp的周期性。因此,这种分离模式与血浆DNA的尺寸特征和核小体结构高度一致,表明Set S优选末端位点可能位于核小体核心内,而Set L优选末端可能位于接头区内。

[0175] 为了探索这一假说,我们研究了尺寸标记的优选末端位点在具有良好定位的核小体的区域周围的分布。具体地,我们研究了Chr12p11.1中的优选末端概况,Chr12p11.1是已知在几乎所有组织类型中都具有良好定位的核小体的区域(29,30)。

[0176] 图17A显示了根据本公开的实施方案,血浆DNA覆盖,Set S和Set L优选末端位点的快照。显示了chr12p11.1区域上的核小体阵列的说明。核小体阵列1720显示具有核小体核心1705和接头区1710。DNA覆盖1730显示覆盖每个基因组位置的多个读取,横轴对应于基因组位置。如图17A所示,Set L优选末端主要位于接头区1710,而Set S优选末端主要位于核小体核心1705内,即使在核心的边缘。

[0177] 此外,由于也已知开放染色质区域(例如启动子和增强子)周围的核小体定位良好(30),我们研究了开放染色质区域周围的优选末端位点的定位。已知母体血浆中的胎儿和母体DNA分子主要分别来自胎盘组织和造血系统(12,31)。为此,我们从RoadMap Epigenomics项目中下载了胎盘和所选造血组织的DNaseI超敏概况(32)。值得注意的是,嗜中性粒细胞的DNaseI概况是不可用的。我们使用T-细胞概况作为其它造血细胞的代表,因为RoadMap项目揭示了几种造血细胞谱系(即T-细胞,B-细胞,自然杀伤细胞,单核细胞,嗜中性粒细胞和造血干细胞)之间的表观遗传学概况是相似的(32)。我们确定了由胎盘和T细胞共有的开放染色质区域(这些被称为共同的开放染色质区域)周围的尺寸标记的优选末端位点。

[0178] 图17B显示了根据本公开的实施方案,围绕由胎盘组织和T细胞共有的共同的开放染色质区域的优选末端位点的分布。显示了核小体位置的说明。由于数据是针对所有共同的开放染色质区域,优选末端位点的数目比图17A多得多,并且可以看到分布模式。

[0179] 在X-轴上绘制的比对的核小体位置是相对于表示为区域1770的共同开放染色质区域的中心。长优选位点的归一化末端计数显示为1750,短优选位点的归一化末端计数显示为1760。在图17B中,通过存在于共同的开放染色质区域,即图17B中所示的基因组坐标内的短和长优选位点的总数来归一化一个位置处的末端计数。因此,以相同的方式对两个数据集1750和1760进行归一化。

[0180] 如图17B所示,可以在任一数据集的峰之间观察到~190bp的周期性模式,这与核小体定相模式一致并代表核小体之间的距离(29)。此外,优选末端位点在开放染色质区域

的中心不太丰富。据报道,在开放染色质区域中存在转录因子结合的频繁占用(33),并因此可能阻止DNA切割。此外,Set S和Set L优选末端位点的峰不位于相同的位置。这些峰间隔约25bp,其约为接头区的尺寸。总之,这些数据表明,尺寸标记的优选末端位点的位置与核小体结构密切相关。因此,血浆DNA末端位点的位置与核小体结构有关。正好在开放染色质区域之后的第一核小体之后的高峰是由于开放染色质区域周围的两个核小体比附近的核小体更严格地良好定相,这使得优选末端在它们的接头中更可预测(即峰更高)。

[0181] 为了以全基因组方式进一步验证尺寸标记的优选末端位点和核小体结构的关系,我们从Snyder等人(24)下载了注释的“核小体轨迹”,其含有约13M核小体中心(即具有最大核小体保护的基因座)的位置,这是利用计算方法对所有组织推导出来的。对于Set S和Set L优选末端位点,我们将每个优选末端位点与其最近的核小体中心相关联。然后,我们描绘了优选末端位点到核小体中心的距离的分布。

[0182] 图18A显示了根据本公开的实施方案,妊娠血浆DNA中的尺寸标记的优选末端位点相对于核小体结构的分布。横轴是相对于核小体中心的基因组位置,纵轴是两类尺寸标记的优选末端的归一化末端计数,其中每一组值分别使用它们各自的尺寸优选末端位点的总数进行归一化。

[0183] 红色剪刀1805和蓝色剪刀1810分别表示将产生Set S和Set L优选末端位点的切割事件。如图18A所示,Set S和Set L优选末端位点分别在 $\pm 73\text{bp}$ 和 $\pm 95\text{bp}$ 处显示出主峰,其与基因组中包裹核小体核心的DNA的尺寸和核小体间隔模式相匹配。Straver等人(23)使用另一种计算推导的核小体轨迹进行注释显示了类似的结果。

[0184] 图18B显示了根据本公开的实施方案,尺寸标记的优选末端位点相对于由Straver等人(23)预测的核小体中心的分布。在X-轴上绘制的比对的核小体位置是相对于核小体中心。数据与图16一致,证明了Set S优选末端位点位于核小体核心内,而Set L优选末端位点位于接头区。图18B与图18A的不同之处在于,使用来自独立组的另一核小体位置来确认图18A中的结果。

[0185] 此外,我们还研究了健康受试者中所有常染色体的片段末端。

[0186] 图19显示了根据本公开的实施方案,在健康的非妊娠受试者中短和长DNA分子的常染色体片段末端相对于核小体结构的分布。红色剪刀1905和蓝色剪刀1910分别代表将产生短片段和长片段的切割事件。在X-轴上绘制的比对的核小体位置是相对于核小体中心(23)。

[0187] 归一化的末端计数是在特定位置终止的DNA片段的数目,例如短DNA片段1920的数目和长DNA片段1930的数目,除以对应尺寸类别的总读取数目。在73bp处出现短DNA的峰,在95bp处出现长DNA的峰。短DNA片段对应于60-155个碱基,长DNA片段对应于170-250个碱基。

[0188] 如图19所示,短DNA分子显示出与Set S优选末端相似分布,长DNA分子显示出与Set L优选末端相似分布。因此数据表明,在健康受试者中,短DNA分子大部分在核小体核心内切割,而长DNA分子大部分在接头区内切割。

B. 胎儿和母体特异性末端位点的特征

[0189] 考虑到从胎儿和母体DNA的混合物中挖掘出Set S和Set L优选末端位点,我们进一步研究了我们在以前的研究中(21)胎儿和母体特异性优选末端位点的核小体定位。这些优选末端位点是从携带胎儿特异性和母体特异性SNP等位基因的母体血浆中的DNA分子挖掘

的。因此,进行胎儿特异性,母体特异性血浆DNA末端位点和Chry片段末端位点的分析。

[0190] 图20A显示了核小体结构的图示。图20B显示了核小体结构中胎儿和母体特异性的优选末端位点的分布。图20C显示了妊娠病例和健康男性受试者的chrY片段末端在核小体结构中的分布。图20D显示了在妊娠情况下短和长DNA分子的chrY片段末端在核小体结构中的分布。图20E显示了健康受试者中短和长DNA分子的chrY片段末端在核小体结构中的分布。

[0191] 在X-轴上绘制的比对的核小体位置是相对于核小体中心(23)。纵轴是归一化的末端计数。每幅图显示两组数据,其为每个数据集提供归一化的终止或读取计数。

[0192] 如图20B所示,胎儿特异性优选末端位点主要位于核小体核心内,而母体特异性末端位点主要位于接头区内。使用胎儿和母体特异性SNP位点在先前的研究中挖掘这些胎儿和母体特异性优选末端(55)。这类似于主要位于核小体核心内的短优选末端位点(如图18A所示)和位于接头区中的长优选末端位点。归一化的末端计数对应于位置的数目除以给定组的总数目。因此,这两个组(胎儿优选的和母亲优选的)被分别归一化。

[0193] 在携带男性胎儿的孕妇的血浆中,chrY读取是胎儿来源的。另一方面,在健康男性受试者中,chrY读取主要来源于造血系统。在携带男性胎儿的孕妇的血浆和健康男性的血浆中研究所有chrY读取的末端位点。

[0194] 图20C显示了整个末端位点分布。归一化的末端计数对应于样品中在相对于核小体中心的位置终止的无细胞DNA片段的数目,归一化是基于样品中分析的DNA片段的总数。与从图20B得到的观察结果相似,妊娠样品中的chrY分子显示位于核小体核心内的更多末端位点,而健康男性受试者的血浆中的chrY分子显示超出核小体核心的更多末端位点。

[0195] 我们进一步将孕妇和健康男性受试者的chrY读取分成短和长类别。

[0196] 图20D和20E分别显示妊娠病例和健康受试者的末端位点的分布。有趣的是,妊娠和非妊娠样品中的短DNA分子显示出其末端位点的相似的核小体定位。这一观察结果表明类似的机制在产生这种短DNA分子中起作用的可能性。类似地,妊娠和非妊娠样品中的长DNA分子也显示出其末端位点的相似的核小体定位,因此可能在其产生中共有相似的机制。另一方面,产生短和长DNA分子的偏好似乎在胎儿和母体来源的DNA中不同。

[0197] 总之,在妊娠的情况下,胎儿DNA经常在核小体核心内被切割(即,Set S优选末端位点),并且母体DNA大部分在接头区内被切割(即,Set L优选末端位点)。

C. 胎盘和造血细胞中的核小体可及性

[0198] 我们想知道为什么胎儿DNA经常在核小体核心内被切割。在体细胞组织中,内切核酸酶切割核小体核心内的DNA比切割接头区更困难,因为核小体核心内的DNA与组蛋白结合(34)。因此,我们假设胎盘细胞与体细胞组织不同,因为核小体核心内的DNA更易接近,因此更容易被切割。

[0199] 为了检验这一假说,对两个胎盘组织样品(一个合胞体滋养层样品和一个细胞滋养层样品)和两个母体血沉棕黄层样品进行了ATAC-seq(使用测序对转座酶可及的染色质的测定)实验(35),该实验已经用于探索核小体可及性(36)。ATAC-seq实验利用切割无核小体DNA的转座酶来研究开放染色质区域和附近的核小体定位(35)。在先前对体细胞组织进行的ATAC-seq实验(35,37,38)中的DNA插入片段尺寸模式显示出约200bp的强周期性模式。这种模式表明开放染色质区域被200bp的区域分开并且可能与完整的核小体结合(35)。在

图21A和21B中显示了我们的ATAC-seq实验的插入片段尺寸分布。

[0200] 图21A和21B显示了来自(A)血沉棕黄层样品和(B)胎盘组织的ATAC-seq数据的片段尺寸分布。测量转座酶切割产生的DNA片段的尺寸,然后确定频率直方图。对于图21A和21B中的每一个标记染色质结构的不同部分。

[0201] 在血沉棕黄层样品中,转座酶主要切割非核小体结合的DNA(例如,接头区)。相反,转座酶能够在胎盘组织中的核小体内切割,表明胎盘组织中的核小体包装不如血沉棕黄层样品中的紧密。蓝色和红色剪刀分别指示血沉棕黄层样品和胎盘组织中可能的切割事件。

[0202] 血沉棕黄层样品(图21A)的插入片段尺寸分布类似于先前研究中观察到的那些(35,37,38)。尺寸概况中在~200和~400bp处的峰是由整数倍的核小体保护的DNA(37),这表明转座酶主要切割血沉棕黄层样品中的非核小体结合的DNA(例如,接头区)。另一方面,胎盘组织样品显示出显著改变的尺寸分布,因为在200bp附近的峰不存在(图21B)。相反,胎盘样品的ATAC-seq插入片段分布显示短得多的DNA分布,表明转座酶能够在核小体中切割,从而表明胎盘组织中的核小体包装不如血沉棕黄层样品中的紧密。结果,数据显示胎盘DNA与比血沉棕黄层DNA更易接近的染色质相关联。

VI. 使用尺寸标记的终止位置的技术

[0203] 如上所述,各实施方案可以使用短的优选终止位置来确定来自与短的无细胞DNA片段相关的特定组织类型(例如,肿瘤,移植物,或胎儿组织)的DNA片段的比例贡献。各实施方案还可以确定对于第一组织类型是否存在序列失衡。第一组织类型(例如,肿瘤、移植物或胎儿组织)可基于特定受试者来鉴定。例如,如果受试者先前患有肝癌,则可进行筛查以检查肝癌是否已经重新开始,这将导致来自肿瘤组织的比例贡献的增加。作为另一个实例,如果受试者是妊娠的女性,则第一组织类型可以是胎儿组织。这种选择标准适用于本文所述的其它方法。

A. 尺寸标记的优选末端的示例性结果概述

[0204] 我们对血浆DNA中的尺寸概况和优选DNA末端位点进行了综合分析。与使用基因型信息推断胎儿和母体特异性优选末端位点相比,在此描述的尺寸标记的方法允许我们挖掘尺寸优选的末端位点,这使得能够改进对血浆DNA中胎儿DNA分数的估计。如图4A和4B所示,为了估计胎儿DNA分数,这种尺寸标记的优选末端位点也显示出比单独使用尺寸概况(17)更好的性能。此外,我们显示覆盖尺寸标记的优选末端位点的读取在21三体的非侵入性产前测试中提供了比使用随机读取更好的性能(图5B)。这些数据为开发特异性富集具有尺寸标记的优选末端位点血浆DNA分子的靶向方法开辟了可能。这种富集方法将潜在地降低非侵入性胎儿非整倍性检测的测序深度要求。

[0205] 此外,我们将尺寸标记的优选末端位点在核小体结构的背景中的位置关联起来,例如,如图17A所示。我们发现,Set S优选末端位点位于核小体核心内,而Set L优选末端位点位于接头区内。有趣的是,我们发现,对于所有被研究的孕妇和健康非妊娠受试者,覆盖Set S优选末端位点的读取比覆盖Set L优选末端位点的读取短,如图2、3、6和7A中所示。这一观察结果表明,Set S和Set L优选末端位点与短的和长的血浆DNA分子相关,而与它们的来源组织无关,因为该相关也存在于健康的非妊娠受试者中。

[0206] 对孕妇血浆中chrY读取的进一步分析显示一致的结果。即使胎儿DNA在母体血浆中的相对短小在2004年首次报道(14),对这种现象的机理解释仍未解决。这里,我们已经提

出了胎盘组织中的核小体可及性高于母体体细胞组织(例如血细胞)的理论,从而允许内切核酸酶在细胞死亡过程(例如细胞凋亡)期间在核小体核心内切割。我们的ATAC-seq实验表明,与血细胞相比,胎盘细胞中的转座酶确实更容易接近核小体核心,如图21A和21B所示。尽管这种可及性的分子基础仍然不清楚,我们提出DNA甲基化可能是一个起作用的因素。在人基因组中,DNA甲基化概况在核小体结合的DNA中显示出10bp的周期性,这与血浆DNA的尺寸模式一致(39)。

[0207] 事实上,我们和其他人已经证明血浆DNA的片段尺寸与DNA甲基化水平正相关(40, 41)。此外,在妊娠期间,胎盘基因组的DNA甲基化逐渐增加,并且母体血浆中胎儿来源的DNA的片段尺寸也随孕龄而增加(42)。所有这些研究表明DNA甲基化可能通过改变染色质可及性而影响片段化过程。与体细胞组织相比,已知胎盘组织表现出全基因组的低甲基化(43)。先前的研究已经表明DNA甲基化可以诱导DNA在伴随的组蛋白(44)周围的更紧密的包裹,并且增加核小体的压缩,刚性和稳定性(45, 46)。此外,DNA甲基化还可以调节组蛋白修饰以及异染色质形成(47, 48),这与核小体展开,解体和稳定性有关(49)。所有这些研究表明,胎盘组织中较高的核小体可及性可能与其低甲基化有关。

[0208] 尽管我们使用循环的无细胞胎儿DNA和来自胎盘组织的DNA来获得对胎儿DNA片段化的机理认识,但该概念可适用于非胎儿来源的无细胞DNA。非妊娠个体的血浆中短和长DNA分子的优选末端位点显示出与核小体结构相同的定位模式,例如,如图20D和20E所示。这些数据表明,类似的一组机制可能有助于将短或长的DNA分子释放到妊娠个体和非妊娠个体的血浆中。然而,如图7B所示,在妊娠的样品中,短DNA分子与长DNA分子的比值高于来自非妊娠个体的血浆中的比值。此外,癌症患者和孕妇的血浆DNA概况之间存在显著的相似性。因此,血浆中的肿瘤来源的DNA分子较短(28),并且肿瘤基因组也表现出全基因组的低甲基化(50, 51)。因此,我们认为肿瘤来源的DNA的短小可能是由于类似的机制(52)。因此,如本文所述,尺寸标记的末端位点可用于非侵入性癌症测试。

[0209] 我们已经在挖掘无细胞DNA的优选末端位点中并入了尺寸特征,并证明了这种尺寸标记的位点非侵入性产前和癌症测试中的用途。我们进一步表明,优选末端与核小体结构高度相关,从而揭示了对母体血浆中无细胞DNA产生机制及胎儿DNA相对短小的机理认识。

[0210] 此外,我们使用短的尺寸和片段末端特征来富集临床相关的DNA分子。这里,实施方案使用这些特征来鉴定相关的无细胞DNA分子的子集。对于测试样品来说,不需要宽的和深的测序,并且宽的和深的测序可能只需要从历史样品中鉴定这些特征。用于临床相关DNA(例如胎儿,肿瘤和移植物)的这种富集的样品可用于以更高的准确性检测序列失衡。

B. 从特定组织类型确定DNA的分数

[0211] 图22显示了在短标记的终止位置上终止的无细胞DNA分子的相对丰度(例如短/长)与混合物中组织A对DNA的比例贡献(其通过分析来自组织A的两个或更多个具有已知的DNA比例浓度的校准样品确定)之间的关系。在所示的实例中,分析了具有 x_1 和 x_2 的组织A的比例贡献的两个样品。将两个样品的相对丰度值分别确定为 y_1 和 y_2 。相对丰度与A的比例贡献之间的关系可以基于 x_1 , x_2 , y_1 和 y_2 的值来确定。本文描述了在短标记的终止位置终止的无细胞DNA分子的相对丰度的各种实例。

[0212] 值 y_1 和 y_2 是校准值的实例。数据点 (x_1, y_1) 和 (x_2, y_2) 是校准数据点的实例。可以

将校准数据点拟合到函数以获得校准曲线(例如,1010,其可以是线性的)。当测量新样品的新的相对丰度时,可以将新的相对丰度与校准值中的至少一个进行比较,以确定新样品的比例贡献的分类。可以以各种方式进行与校准值的比较。例如,校准曲线可用于找到对应于新的相对丰度的比例贡献 x 。作为另一个实例,可以将新的相对丰度与第一校准数据点的校准值 y_1 进行比较,以确定新的样品是作为大于还是小于 x_1 的比例贡献。

[0213] 在其它实施方案中,可以类似地分析含有多于两种类型组织的混合物中的组织A的比例贡献,只要其它组织的相对丰度相对恒定即可。这种方法在实际中可用于分析不同的临床情况,例如但不限于癌症检测,移植监测,创伤监测,感染和产前诊断。

[0214] 对于胎儿分析,目标可以是提供比例贡献的定量值或确认存在最小百分比的胎儿DNA。例如,该方法可用于测定母体血浆中的胎儿DNA浓度。在母体血浆中,携带胎儿基因型的DNA分子通常来源于胎盘。

[0215] 对于癌症,可能需要其它分类。例如,可以测定短优选位置处的相对丰度,并与正常健康受试者进行比较。通过与类似于图22的校准曲线进行比较,可以确定特定组织(例如胎儿,肿瘤或移植物)的贡献。可以将测试病例的相对丰度值与健康受试者中肝脏的贡献范围进行比较。

[0216] 类似地,通过这种方法可以确定移植器官在已经接受器官移植的患者中的贡献。在先前的研究中,显示排斥患者将导致DNA从移植器官释放增加,导致血浆中移植器官DNA浓度升高。对移植器官相对丰度的分析将是检测和监测器官排斥的有用方法。用于这种分析的区域可以根据移植的器官而变化。

[0217] 图23是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型的比例贡献的分类的方法2003的流程。生物样品包括来自包括第一组织类型的多种组织类型的无细胞DNA分子的混合物。与这里描述的其它方法一样,方法2300可以使用计算机系统。第一组织类型的实例包括胎儿组织,移植组织和肿瘤组织。

[0218] 在框2310,鉴定第一组基因组位置,在这些位置处,对于含有第一组织类型的样品,短无细胞DNA分子的末端以高于第一阈值的第一比率出现。短无细胞DNA可具有指定的第一尺寸,例如60-155个碱基,本文所述的其它范围,或比长无细胞DNA片段小的其它范围。范围不必是连续的,例如60-120和125-155。作为实例,长DNA片段可以是170-250个碱基和本文所述的其它范围。可以在至少一个另外的样品(例如,在校准样品中)中确定较高的比率。关于框2310的进一步细节可以在上面的II.B部分和本公开的别处中找到。

[0219] 在一些实施方案中,鉴定第一组基因组位置可以包括分析来自至少一个另外的样品的第二多个无细胞DNA分子以鉴定第二多个无细胞DNA分子的终止位置。已知至少一个另外的样品包括第一组织类型并且具有与生物样品相同的样品类型。例如,另外的样品可以来自妊娠的女性,具有移植器官的受试者,或具有肿瘤的受试者。对于多个基因组窗中的每个基因组窗,可以计算在基因组窗上终止的第二多个无细胞DNA分子的相应数目,并将其与参考值进行比较,以确定在基因组窗内的一个或多个基因组位置上终止的无细胞DNA分子的比率是否高于阈值。

[0220] 在框2320,分析来自受试者的生物样品的第一多个无细胞DNA分子。无细胞DNA分子的分析可包括确定参考基因组中对应于无细胞DNA分子的至少一个末端的基因组位置(终止位置)。因此,可以确定两个终止位置,或仅确定无细胞DNA分子的一个终止位置。

[0221] 在一些实施方案中,分析第一多个无细胞DNA分子可包括对第一多个无细胞DNA分子进行测序以获得序列读取并将序列读取与参考基因组比对以确定第一多个无细胞DNA分子的基因组位置。在其它实施方案中,分析第一多个无细胞DNA分子可以包括在第一组基因组位置的第一多个无细胞DNA分子的杂交捕获或扩增。

[0222] 终止位置可以以多种方式确定,如本文所述。例如,可以对无细胞DNA分子进行测序以获得序列读取,并且可以将序列读取映射(比对)至参考基因组。如果生物体是人,则参考基因组将是可能来自特定亚群的参考人基因组。作为另一个实例,可以用不同的探针(例如,在PCR或其它扩增之后)分析无细胞DNA分子,其中每种探针对应于基因组位置,其可以覆盖至少一个基因组区域。

[0223] 可以分析统计学上显著数目的无细胞DNA分子,以便提供来自第一组织类型的比例贡献的精确测定。在一些实施方案中,分析至少1,000个无细胞DNA分子。在其它实施方案中,可以分析至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个无细胞DNA分子或更多。作为另一个实例,可以产生至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个序列读取。

[0224] 在框2330,确定第一数目的第一多个无细胞DNA分子在多个窗之一内终止。在方框2320中,可以基于对第一多个无细胞DNA分子的分析来进行测定。例如,无细胞DNA分子末端的基因组位置可以从分析(例如,特定探针的比对或使用)中获知。每个窗包括第一组基因组位置中的至少一个。如IIA部分所述,第一组基因组位置可以从初始组中鉴定,然后扩展到包括初始组周围的窗。因此,一组短的优选终止位置可以包括扩展的终止位置Set S。作为实例,窗的宽度可以是1bp、2bp、3bp、4bp、5bp、6bp、7bp、8bp、9bp、10bp、15bp、20bp、25bp和30bp。窗可以具有或不具有所有相同的宽度。提及bp和碱基可被认为是宽度或长度的等同单位。

[0225] 在框2340,计算在多个窗之一内终止的第一多个无细胞DNA分子的相对丰度。相对丰度可以通过使用第二数目的无细胞DNA分子将第一数目的第一多个无细胞DNA分子归一化来确定。第二数目的无细胞DNA分子可包括在包括第一组基因组位置的多个窗之外的第二组基因组位置终止的无细胞DNA分子。作为实例,相对丰度可以包括第一数目和第二数目的比值。

[0226] 在多个实施方案中,第二组基因组位置可以是长的无细胞DNA片段优选的终止位置或在生物样品中测定的任何终止位置。第二组基因组位置可以使得长的无细胞DNA分子的末端在至少一个另外的样品中以高于阈值的第二比率出现。长的无细胞DNA将具有大于第一尺寸的第二尺寸。第一尺寸可以具有第一尺寸范围,第二尺寸可以具有第二尺寸范围。第一尺寸范围可以小于第二尺寸范围,在于第一尺寸范围的第一最大值小于第二尺寸范围的第二最大值。如本文所述,第一尺寸范围可以与第二尺寸范围重叠。在另一个实施方式中,第二组基因组位置可以包括对应于第一多个无细胞DNA分子中的至少一个的末端的所有基因组位置,从而包括可能以随机方式采样的多种基因组位置。

[0227] 相对丰度值的另一个实例是在基因组窗上终止的无细胞DNA分子的比例,例如作为在优选终止位置上终止的测序DNA片段的比例来测量。因此,第二组基因组位置可以包括对应于第一多个无细胞DNA分子中的至少一个的末端的所有基因组位置。在另一个实例中,第二组基因组位置可以对应于这样的窗,其大于用于定义第一组基因组位置的窗,从而包

括不在第一组中的另外的基因组位置。可以调整两组窗的宽度以获得所需的性能。作为实例,第二组窗的宽度可以是20bp、25bp、30bp、40bp、50bp、60bp、70bp、80bp、100bp、120bp、140bp、160bp、180bp和200bp。

[0228] 在框2350,通过将相对丰度与从一个或多个校准样品确定的一个或多个校准值进行比较来确定第一组织类型的比例贡献的分类,所述一个或多个校准样品的第一组织类型的比例贡献是已知的。在图4A和4B中显示了胎儿组织作为第一组织类型的实例,并且在图10和11中显示了肿瘤DNA的实例。作为一个实例,比例贡献的分类可以对应于高于指定百分比的范围。作为另一个实例,分类可以对应于指定精度范围内的特定百分比或指定精度。作为进一步的实例,分类可以是对应于范围的文本分类,诸如低,中和高。

[0229] 如上所述,与校准值的比较可以通过已经使用校准样品中测量的校准数据点确定的校准函数进行,所述校准样品的比例贡献通过其它技术测量,例如使用组织特异性标志物(例如,用于胎儿,移植物或肿瘤组织的),例如组织特异性等位基因或组织特异性表观遗传标志物,例如在特定组织的特定部位相对于其它组织的低甲基化或高甲基化。因此,将相对丰度与一个或多个校准值进行比较可以使用与校准点拟合的校准函数,所述校准点包括在多个校准样品中测量的第一组织类型的比例贡献和在多个校准样品中确定的相应相对丰度。

[0230] 当第一组织类型是肿瘤时,分类可以选自:受试者中的肿瘤组织的量,受试者中的肿瘤尺寸,受试者中的肿瘤阶段,受试者中的肿瘤负荷,和受试者中的肿瘤转移的存在。

[0231] 对于癌症,如果比例贡献高,则可以进行进一步的动作,例如受试者的治疗干预或成像(例如,如果第一组织类型对应于肿瘤)。例如,调查可以使用受试者(整个受试者或身体的特定部分(例如胸腔或腹部),或具体地候选器官)的成像模式,例如计算机断层摄影(CT)扫描或磁共振成像(MRI),以确认或确定受试者中肿瘤的存在。如果肿瘤的存在被证实,则可以进行治疗,例如,手术(通过刀或通过放射)或化疗。

[0232] 可以根据确定的癌症水平,鉴定的突变和/或来源的组织提供治疗。例如,所鉴定的突变(例如,用于多态性实施)可以用特定的药物或化疗靶向。来源组织可用于指导手术或任何其它形式的治疗。并且,癌症的水平可以用于确定用任何类型的治疗有多大侵略性,这也可以基于癌症的水平来确定。

C. 确定序列失衡

[0233] 图24是根据本公开的实施方案,分析生物学样品以确定第一组织类型是否在没有细胞DNA分子的混合物的染色体区域中显示序列失衡的方法2400的流程。序列失衡可涉及染色体区域中的各种测量,例如非整倍性,扩增/缺失,或在区域中的基因座处对第一组织类型进行基因分型分析。例如,第一组织可以具有与多种组织类型中的其它组织类型不同的基因型。染色体区域可以是完整的染色体。第一组织类型的实例包括胎儿组织和肿瘤组织。

[0234] 在框2410,鉴定第一组基因组位置,在该位置,对于含有第一组织类型的样品,短无细胞DNA分子的末端以高于第一阈值的第一比率出现。短无细胞DNA分子可以具有第一尺寸,其可以是一个或多个范围。框2410可以以与图23的框2310相似的方式进行。

[0235] 在框2420,分析来自受试者的生物样品的第一多个无细胞DNA分子。分析无细胞DNA分子包括确定参考基因组中对应于无细胞DNA分子的至少一个末端的基因组位置。框2420可以以与图23的框2320相似的方式进行。

[0236] 在框2430,基于对第一多个无细胞DNA分子的分析来鉴定在多个窗之一内终止的无细胞DNA分子的组。每个窗包括基因组位置组中的至少一个,并且位于染色体区域中。通过选择在短DNA片段优选的这组基因组位置上终止的特定的无细胞DNA分子,这组无细胞DNA分子可以有效地富集用于第一组织类型,例如肿瘤DNA或胎儿DNA。此外,可以扩增或捕获覆盖或在基因组位置组终止的无细胞混合物中的DNA片段,以提供进一步的富集。

[0237] 框2430可以以与图23的框2330相似的方式进行,例如,关于在基因组位置组中的一个处终止的DNA分子的鉴定。通过在染色体区域内具有窗,该组无细胞DNA分子可以作为该染色体区域的代表性集合。因此,这组无细胞DNA分子(针对第一种组织类型而富集)可以使用现有技术进行分析以用于非侵入性分析。

[0238] 在多个实施方案中,可以为特定的单倍型选择组。在多个窗之一内终止的另一组无细胞DNA分子可以对应于另一个单倍型。或者,该组的一个亚组可以对应于一个单倍型,而该组的另一个亚组可以对应于另一个单倍型。对应于单倍型的DNA分子可以基于与特定单倍型的特定等位基因匹配的DNA分子的等位基因(例如,通过测序或探针确定)来确定。方法2400的后面的框可以分析这两个组以比较这两个单倍型的特性,例如从而确定序列失衡。

[0239] 在框2440,确定无细胞DNA分子的组的值。可以以各种方式确定该值。例如,组中的无细胞DNA分子的数目可以例如如美国专利公开号2009/0087847、2009/0029377、2011/010553、2013/0040824和2016/0201142中所述来确定。作为另一个实例,该值可以是无细胞DNA分子的组的尺寸分布的统计值,例如,如美国专利公开号2011/0276277、2013/0040824和2016/0201142中所述,其全部内容通过引用整体并入本文。作为另一个实例,该值可以是无细胞DNA分子的组的甲基化密度,例如,在被这些无细胞DNA分子覆盖的CpG位点的甲基化密度。因此,在多个实施方案中,无细胞DNA分子的组的值可以是无细胞DNA分子的组的量,无细胞DNA分子的组的尺寸分布的统计值,或无细胞DNA分子的组的甲基化水平。关于使用甲基化检测序列失衡的进一步细节可在PCT公开WO2017/012544中找到。

[0240] 无细胞DNA分子的组的值可以被归一化,例如,以解释不同样品中DNA分子的不同数目。例如,组的值可以通过来自一个或多个参考区域的另一组无细胞DNA分子的值或样品中无细胞DNA分子的总数进行归一化(例如除以)。作为另一个实例,可以分析相同数目的无细胞DNA分子,这是一种通过样品中无细胞DNA分子的总数进行归一化的类型。

[0241] 在框2450,基于该值与参考值的比较来确定受试者染色体区域内,第一组织类型中是否存在序列失衡的分类。参考值可以以各种方式确定,例如,从健康受试者,从患有癌症或妊娠的受试者,从样品中没有失衡的其它区域确定的一个或多个值,或从染色体区域中的另一个单倍型确定(例如,以确定基因型是什么)。基因型可以通过分析在一个基因座处的不同等位基因或单倍型的读取的不平衡来确定,例如,如部分III.C所述。比较可以包括确定所述值是否在统计学上不同于参考值(例如,超过截止值,例如从群体确定的特定数目的标准偏差)。

[0242] 作为一个实例,可以将第一染色体区域(待测试的临床相关区域)中的第一窗之一终止的第一数目的无细胞DNA分子与在一个或多个参考染色体区域中的第二窗之一终止的第二数目的无细胞DNA分子进行比较,其中第一和第二窗包括基因组位置组中的至少一个。这种比较可以包括使用第一目的和第二数目来确定分离值(例如,差值或比率),其中可

以将分离值与参考值进行比较以检测序列失衡。类似地,可以确定第一和第二单倍型的第一和第二数目。

[0243] 作为另一个实例,可以确定无细胞DNA分子的组的尺寸分布。可以确定尺寸分布的统计值,例如平均或中值尺寸,或短DNA分子比长DNA分子的量。可以确定染色体区域的第一统计值和一个或多个参考染色体区域的尺寸分布的第二统计值之间的分离值,其中可以将分离值与参考值进行比较以检测序列失衡。类似地,可以确定第一和第二单倍型的第一和第二统计值。

[0244] 作为另一个实例,甲基化水平可以使用无细胞DNA分子的组覆盖的多个位点处的甲基化状态(甲基化或非甲基化)来确定。可将该组的甲基化水平与对应于一个或多个参考染色体区域的另一组的另一甲基化水平进行比较。可以确定两个甲基化水平之间的分离值,其中可以将分离值与参考值进行比较以检测序列失衡。类似地,可以确定第一和第二单倍型的两个甲基化水平。在另一个实例中,可以确定区域中的不同位点的多个甲基化水平,并且可以使用如WO 2017/012544中的去卷积技术来确定分数贡献。分数贡献将是在框2440中确定的组的值的实例。

[0245] 因此,对于单倍型分析,可以使用对应于第一单倍型的第一亚组和对应于染色体区域中的第二单倍型的第二亚组来确定组的值。可以确定第一单倍型值和第二单倍型值(上面提供了实例)之间的分离值,并将其与参考值进行比较。

[0246] 为了在区域之间进行比较(如上所述),可以通过鉴定在多个参考窗之一内终止的无细胞DNA分子的参考组来确定参考值,每个参考窗包括基因组位置组中的至少一个,并且位于一个或多个参考染色体区域,其可以已知或假定没有序列失衡(例如扩增或缺失)。然后,可以从无细胞DNA分子的参考组确定参考值。参考值可以是与该值相同的类型(例如,量,统计尺寸值或甲基化水平)。然后将该值和参考值之间的分离值与截止值进行比较,所述截止值分离存在序列失衡和不存在序列失衡的分类,例如如图5A所示。

[0247] 例如,当序列失衡是第一组织类型与其它组织类型的不同基因型的结果时(例如,如部分III.C所述),无细胞DNA分子的组的值可以是该组中在一基因座具有第一等位基因的第一数目的无细胞DNA分子和在该基因座具有第二等位基因的第二数目的无细胞DNA分子之间的相对丰度。当其它组织类型在染色体区域中的基因座处是杂合时,序列失衡的分类可以是第一等位基因过多,这表明第一组织类型对于第一等位基因是纯合的。当其它组织类型在染色体区域中的基因座处是杂合时,分类可以是不存在失衡,这表明第一组织类型对于第一等位基因和第二等位基因是杂合的。

[0248] 如果序列失衡与癌症(扩增或缺失)相关,则可以确定癌症水平(例如,基于具有序列失衡的多个区域)。然后可提供治疗,例如如本文所述,如方法2300。

VII. 开放染色质区无定向血浆细胞DNA片段化分析

[0249] 最近的研究已经证实了cfDNA分析对敏感癌症筛查的临床可行性(56,57,61)。对于该领域的未来发展,开发用于在阳性液体活检之后定位肿瘤部位的稳健方法将是有益的。利用组织之间DNA甲基化模式的差异,我们以前已经证明母体血浆中的循环胎儿来源的DNA主要来自胎盘(58)。该工作基于母体血浆中作为胎盘标志物的未甲基化SERPINB5序列的检测(58)。最近,一种方法已经被应用于检测来源于脑(78),红系细胞(75),心脏(109)和肝脏(64,77)的cfDNA。

[0250] 我们进一步开发了一种基于DNA甲基化的通用方法,用于确定多种组织类型对cfDNA池的贡献,这是一种我们命名为“血浆DNA组织映射”的方法(102)。该原理也被其它研究者用来预测肿瘤的来源组织(72,79)。这些公开的方法使用全基因组亚硫酸氢盐测序(BS-seq)(80,54,85)。然而,BS-seq的缺点是亚硫酸氢盐转化与输入DNA的降解有关(65),并且还引入GC含量变化,这可能导致测序数据中的偏差(89)。

[0251] 近来的研究表明,cfDNA分子除了DNA甲基化外,还保留了其核小体来源的特征,这显示在166bp处具有主峰和10bp的周期性的尺寸分布(81)。已经显示cfDNA携带非随机的片段化模式,其提供了跨越基因组的表观遗传调节的窗(67)。考虑到跨越基因组的核小体定位与细胞身份高度相关(92),因此这种片段化模式具有追踪回cfDNA分子的来源组织的潜力。Snyder等人显示血浆DNA分子携带核小体足迹(98)。作者进一步构建了“核小体轨迹”,发现核小体间隔模式可用于推断cfDNA的组织来源。他们还证明了这种方法在预测癌症患者中的肿瘤来源方面的潜力。在另一个研究中,U1z等人报道了启动子中的血浆DNA覆盖可用于预测基因的表达(106)。我们的小组已经证实在cfDNA中存在组织特异性优选的终止位点,这在预测母体血浆中的胎儿DNA分数中显示出临床实用性(55)。

[0252] 在本公开中,我们进一步探索了片段化模式的临床潜力,特别是在追踪cfDNA分子的来源组织方面。我们首先对已知的良好定位的核小体阵列和开放染色质区域周围的覆盖和cfDNA片段末端特征进行了描述。在分析过程中,我们将血浆DNA片段末端分成两组,其中考虑了方向信息,即在血浆DNA片段相对于参考基因组的上游或下游侧的末端。我们显示在这些区域中,血浆DNA表现出特征性片段化模式,包括测序覆盖失衡以及上游和下游片段末端信号之间的差异。然后,我们分析各种组织特异性开放染色质区域中的血浆DNA片段化模式,并进一步定量各种临床情况中的片段化模式,以研究推断cfDNA的来源组织的可行性,包括预测癌症患者中的肿瘤位置。

A. 概念框架和命名法

[0253] 图25A-25F显示了我们的方法的概念框架。图25A显示了基因组中核小体定位的图示。核小体2505用DNA 2510(黄线)包裹。还显示了DNA的其它部分:接头DNA 2512(棕线),和活性调节元件2514(绿线),它们在开放染色质区域中。还显示了核小体定位的抽象概念和在细胞凋亡期间切割事件(剪刀)的图示。

[0254] 在真核染色质中,核小体是DNA包装的基本单元,其由包裹在组蛋白周围的DNA区段组成。核小体通常通过相对较短的接头DNA相互连接,除了在活性调节元件(例如,开放染色质区域)中之外,其中核小体被驱逐并且附近的核小体将通过更长的DNA节段连接。据信在细胞凋亡后释放了相当大比例的cfDNA分子(68,81)。在凋亡DNA片段化过程中,提出内切核酸酶优选切割核小体间DNA(94,103)。

[0255] 图25B显示了由凋亡DNA片段化产生的cfDNA的图示。包裹核小体的DNA部分2520被保留,同时接头和开放染色质区域中非常小的DNA片段2522被切割成这种小片段(灰线),这种小片段不能被有效测序。结果,当cfDNA分子进行测序时,包裹在组蛋白上的DNA部分2520被保留。另一方面,源自接头和活性调节元件的DNA,由于它们相对不受保护,将被切割成小DNA片段2522(灰线),并且可能不能被有效测序(图25C)(69,98,106)。

[0256] 图25C是测序读取和两个末端的提取的图示。红色末端2530和蓝色末端2532分别代表U(上游)和D(下游)血浆DNA末端。DNA片段2522没有显示,因为它们没有被测序。因此,

cfDNA的基因组覆盖在核小体中是高的,而在接头和开放染色质区域中是低的(图25D)。

[0257] 图25D显示了基因组覆盖。横轴对应于基因组坐标。纵轴对应于覆盖每个坐标(位置)的读取的数目。在这种理想化的描述中,在接头和开放染色质区域中的覆盖是零(或接近零),但是在核小体区域中是大量和均匀的。

[0258] 图25E显示了相对于基因组坐标的cfDNA的U和D片段末端概况。我们利用cfDNA片段末端的方向信息,并基于它们与参考基因组的比对来定义那些cfDNA片段末端。上游(U)末端2530代表在基因组坐标中具有较低值的末端,而下游(D)末端2532代表在基因组坐标中具有较高值的末端。因此,包裹在核小体上的DNA将分别在核小体的上游和下游边界处产生一对U和D末端。

[0259] DNA的上游末端2530和下游末端2532的示例性位置示于图25E中。上游U信号2550位于上游末端2530的终止位置。下游D信号2552位于下游末端2532的终止位置。U信号2550和D信号2552聚束在一起,显示出一些随机过程,因为不是每个片段都在相同的位置被切割。这种位置窗可以对应于上述用于尺寸优选的终止位置的窗。

[0260] 可以基于U信号2550和D信号2552来鉴定接头和开放染色质区域。对于接头或开放染色质区域,在它们的上游边界侧有D末端,在它们的下游边界侧有U末端。在这点上,U和D末端信号可用于推断核小体,接头和开放染色质区域的定位(图25F)。

[0261] 图25F显示了平滑的血浆DNA末端信号和推导的核小体定位。这种平滑的末端信号说明了真实的数据,因为DNA片段的末端将显示由于切割DNA所涉及的随机过程而引起的分布。上游分布2560以图25E中的U信号2550为中心。下游分布2562以图25E中的D信号2552为中心。

[0262] 在平滑的血浆DNA末端信号下鉴定不同的区域。紫色线2575代表核小体。棕色线2572代表接头区。绿色线2574代表开放染色质区域。

B. 显示差异定相的结果

[0263] 通过分析基因组的各个部分,例如管家基因的活性启动子,非活性启动子和组织特异性开放染色质区域,测试来自概念框架的假设。

1. 差异定相血浆DNA片段在核小体阵列终止

[0264] 为了说明在人基因组区域中的上述概念,我们首先检查chr12p11.1,一个已知在几乎所有人类组织类型中具有良好定位的核小体的区域(107,63,98)。为此,我们收集了来自我们以前研究的32名健康非妊娠受试者的血浆DNA数据(70),并对该区域的覆盖和片段末端进行了描述。

[0265] 图26A和26B显示了根据本公开的实施方案,在合并的健康非妊娠受试者的chr12p11.1区域中的血浆DNA片段化模式。图26A显示基因组覆盖2605,上游U末端位置2607和下游D末端位置2609的原始信号。X轴是基因组坐标。Y轴是基因组覆盖的归一化密度,因此在任何坐标下的平均值是1。基因组覆盖2605对应于与每个基因组比对了的读取的数目。上游终止位置2607和下游终止位置2609的数据是在这些位置终止的DNA片段的数目的归一化计数。由于我们只对不同位置上的末端的相对计数感兴趣,因此在该图中以拟合Y轴的方式对原始计数进行归一化。

[0266] 如图26A所示,血浆DNA覆盖2605显示~190bp的强周期性模式,并且具有较高和较低覆盖率的区域分别对应于核小体和接头(98)。U终止位置2607和D终止位置2609显示出类

似的周期性模式,并且两者都富集在接头中,即,接头区中的U和D末端比核小体中的U和D末端多。通过将原始信号除以该区域中的平均信号来归一化覆盖信号;将末端信号线性调整以拟合成该图。图26A,26B,26C和26D中覆盖和末端信号之间的这些非通用归一化程序仅用于说明片段化模式的目的。

[0267] 图26B显示了平滑的信号和推断的核小体定位。然后使用LOWESS(局部加权回归散点平滑)算法(60)对U和D端信号进行平滑,用于进一步分析。如图26B所示,任何D末端峰(例如2610)与其最近的上游U末端峰(例如2620)之间的距离为~170bp,其大致为核小体的尺寸(101)。任何D末端峰(例如,2610)与其最近的下游U末端峰(例如,2630)之间的距离为~20bp,其大致为接头的尺寸(101)。在该图下方,核小体2640和接头2650显示在对应于该图中的数据的位置。

[0268] 因此,这些数据与我们的概念框架高度一致(图25A-25F),并且显示出差异定相的血浆DNA片段末端确实反映了该区域中的核小体定位。值得注意的是,利用U和D末端的分离,我们能够解析核小体和接头两者的定位,这显示了相对于以前的研究的进展,这些研究主要集中于预测核小体中心的位置(即,具有最大核小体保护的基因座)(63,90,98)。

[0269] 除了chr12p11.1区域之外,还已知活性启动子周围的核小体被很好地定位(69)。为了探索活性启动子周围的片段化模式,从文献中获得了人类管家基因的列表(62)。

[0270] 图26C显示了管家基因的活性启动子周围的血浆DNA覆盖和末端信号。显示了位于Watson链上的管家基因的血浆DNA覆盖2660,U终止信号2662和D终止信号2664。X-轴是相对于管家基因的转录起始位点(TSS)的基因组坐标。Y轴是血浆DNA覆盖2660,U终止信号2662和D终止信号2664的归一化密度。TSS显示在两组核小体阵列之间的开放染色质区域2670的中心。

[0271] 位于Crick链上的管家基因显示出几乎相同的镜像模式。血浆DNA覆盖2660在启动子周围显示“V”形模式。然而,末端概况2662和2664在U和D末端之间显示出强的周期性和相位差异,这与转录起始位点(TSS)周围的核小体耗竭区域和附近的良好定位的核小体阵列一致。另外,在TSS和+1核小体2680(即,TSS下游的第一核小体)之间可以观察到~60bp的距离,这与人中的典型基因结构一致(69)。

[0272] 此外,我们还从Expression Atlas(73)中挖掘出在主要的人体细胞组织中不表达的基因列表,以研究无活性启动子(其中没有这样的核小体耗竭模式)周围的片段化模式。

[0273] 图26D显示了非活性启动子周围的血浆DNA覆盖和末端信号。在无活性启动子周围,发现血浆DNA末端均匀分布,并且在这些未表达基因的启动子周围没有显示任何特定的核小体定位模式。因此,特定类型细胞的非表达基因的启动子是无活性的,并且不具有指示开放染色质区域的结构。这些结果与先前对核小体定位的研究(在该研究中,研究了微球菌核酸酶或转座酶消化后的DNA片段末端)(96,95)一致。总之,我们的结果表明,差异定相的血浆DNA片段末端确实可以告知活性启动子中的核小体定位模式。

2. 组织特异性开放染色质区域中的差异定相的血浆DNA片段末端

[0274] 已知开放染色质区域是在中心缺乏核小体且侧翼为定相良好的核小体阵列的调节元件(63,95)。因此,我们假设来源于这种区域的cfDNA也可能显示出差异定相的片段末端信号。因此,我们首先研究了T细胞和肝脏共有的共同开放染色质区域,考虑到这些组织在各种临床情况下是血浆DNA池的重要贡献者。因此,来源于T细胞的DNA是从造血系统释放

的血浆DNA的一个实例(103),造血系统是健康个体血浆DNA的主要来源(84)。在健康个体以及肝脏移植受体和肝癌患者中,肝脏是血浆DNA的另一个主要来源(83,64,77)。

[0275] 我们从RoadMap Epigenomics项目(93)和ENCODE项目(104)获得了T细胞和肝脏的开放染色质数据(参见材料和方法)。我们将T细胞和肝脏共有的开放染色质区域鉴定为共同的开放染色质区域。然后我们对合并的血浆DNA数据中的这些区域进行片段化分析。

[0276] 图27A、27B和27C显示了根据本发明实施方案,合并的健康非妊娠受试者的血浆DNA片段化模式。使用上游和下游终止信号以及基因组覆盖分析开放染色质区域中和开放染色质区域附近的DNA片段化。

[0277] 图27A显示了T细胞和肝细胞共有的共同开放染色质区域中的模式(也绘制了推断的核小体定位)。X-轴是相对于共同开放染色质区域的中心的相对位置。Y轴是基因组覆盖2705,上游终止信号2707和下游终止信号2709的归一化密度。开放染色质区域2710在上面显示,在每一侧有两个核小体。覆盖和末端信号都通过除以它们相应的总信号进行归一化,然后通过恒定的数字因子1000扩大,使得覆盖和末端信号的平均值均匀地调整到5。将该归一化应用于所有显示开放染色质区域周围的覆盖和末端信号的附图(即,图27至29)。

[0278] 下游峰与核小体的下游末端一致,上游峰与核小体的上游末端一致。两个峰之间的差异程度表明在两个核小体之间是否存在接头或是否存在开放染色质区域。

[0279] 如图27A所示,可以观察到血浆DNA的特征性片段化模式,包括覆盖失衡和差异定相的片段末端。覆盖失衡由坐标0,即共同的开放染色质区域的中心处的覆盖减少来说明。差异定相的片段末端显示为接头区2716的峰之间的小间隔(例如2712),以及开放染色质区域2710的较大间隔(例如2714)。这些结果是开放染色质区域中心的核小体耗竭区域和相邻的定相良好的核小体的存在的结果。因此,这些结果表明,差异定相的血浆DNA片段末端可以告知开放染色质区域中的核小体定位模式。

[0280] 图27B显示胚胎干细胞(ESC)特异性开放染色质区域的模式。作为阴性对照,我们使用相同的数据集来分析对胚胎干细胞(ESC)特异的开放染色质区域周围的血浆DNA片段化模式。我们推断在健康成年人中没有来自ESC的血浆DNA。实际上,我们发现在ESC特异性开放染色质区域中不能看到核小体定位模式(例如,开放染色质区域中心的核小体耗竭)。

[0281] 我们进一步假设cfDNA仅在开放染色质区域显示片段化模式,在开放染色质区域,相应的组织将DNA贡献到血浆中。为了检验这一假说,除了T细胞和肝脏,我们为5个另外的主要人组织(即,胎盘,肺,卵巢,乳房和小肠)挖掘了组织特异性开放染色质区域(参见下面的材料和方法部分)。这些组织的选择是基于数据可用性和以前的知识,即它们将在选定的临床情况下将DNA贡献到血浆中。在以前的工作中,研究者已经表明,胎盘、肺、卵巢和乳房来源的DNA可以分别在孕妇,肺癌,卵巢癌和乳腺癌患者的血浆中发现(82,58,59,66,88)。此外,结肠DNA可以在结肠直肠癌患者的血浆中发现(99)。由于对于结肠组织没有可公开访问的开放染色质数据,我们在本项工作中使用来自小肠的数据来表示胃肠系统并且认为小肠特异性开放染色质区域作为结肠染色质的替代物。此后将这些开放染色质区域称为“肠特异性的”。我们相信我们的决定是正当的,因为小肠和结肠的表观遗传学概况共有许多相似性(93)。

[0282] 为每种组织类型总共获得~26,000个组织特异性开放染色质区域(范围:7,540-55,537)。组织特异性开放染色质区域可如后面部分所述进行鉴定。然后我们研究了健康个

体血浆中这些组织特异性开放染色质区域中的血浆DNA片段化模式。

[0283] 图28A-28F显示了根据本公开实施方案,健康受试者的组织特异性开放染色质区域中的血浆DNA片段化模式。每幅图显示对应于一种组织类型的组织特异性开放染色质区域的结果:图28A T-细胞;图28B肝脏;图28C胎盘;图28D肺;图28E卵巢;图28F:乳房;图28G 肠。X-轴显示相对于开放染色质区域的相应中心的位置。Y轴是基因组覆盖,U末端和D末端的归一化密度。

[0284] 正如所预期的,血浆DNA在T-细胞和肝脏特异性开放染色质区域中显示核小体耗竭和定相良好的核小体阵列,而在其它组织特异性开放染色质区域中则不显示。定相良好的核小体阵列可以指基因组中的区域,其中核小体的位置在相同组织类型的几乎所有细胞中是高度可重复的和可预测的。这些结果与在健康个体中造血系统和肝脏是血浆DNA的主要贡献者的事实是一致的(84,102,78)。

C. 血浆DNA片段化模式的定量

[0285] 探讨了开放染色质区域周围血浆DNA片段化模式的定量。为了定量组织特异性开放染色质区域周围的血浆DNA片段化模式,我们集中在中心处的核小体耗竭信号,因为它是该模式的关键特征之一(69)。在该核小体耗竭信号中,上游(U)和下游(D)末端在远离开放染色质区域的中心的不同方向上在偏移(例如,60bp)处表现出最高的读取密度(图27C)。

[0286] 图27C是OCF(识别方向的cfDNA片段化)值的概念的图示。X-轴是相对于开放染色质区域中心的相对位置。Y轴表示上游终止信号2727和下游终止信号2729的归一化密度。该分析集中在开放染色质区域中心的U和D末端,并测量阴影区域2737和2739中的U和D信号2727和2729之间的分离值(例如,差值或比值)作为组织特异性开放染色质区域中的OCF值。

[0287] 可以看出,D末端峰在左手侧,而U末端峰在右手侧。从图28A-28G和其它图中可以看出,组织类型的存在与上游和下游信号之间的定相差异有关。可以使用关于峰位置差异的信息来测量该定相差异,该关于峰位置差异的信息可以提供用于测量U和D末端的特定基因组位置。这种位置上的差异将导致在一个位置或位置窗(例如,在区域2737中)处出现比下游位置更多的上游末端。例如,在区域2737中,上游峰2747对应于该区域中比D末端信号2757多的U末端。类似地,在区域2739中,下游峰2749对应于该区域中比U末端信号2759多的D末端。考虑到大多数组织特异性开放染色质区域具有相似的尺寸,可以在相对于各种组织的中心对称的位置选择这些区域。

[0288] 在一些实例中,如下通过在峰周围的两个窗(例如,20bp)中的U和D末端的读取密度的差异定量定相差异:

$$OCF = \sum_{-峰-仓}^{峰+仓} (D - U) + \sum_{峰-仓}^{峰+仓} (U - D)$$

峰是距开放染色质区域的中心的距离,并且仓(bin)是该区域的宽度。如图27C所示,峰距中心60个碱基,宽约10个碱基。

[0289] 这类参数被称为OCF(识别方向的cfDNA片段化)值。在多个实施方案中,可以存在一个或两个项,并且可以使用不同的峰偏移值。在一些实施方式中,我们使用(但不限于)60bp作为峰和10bp作为仓尺寸,用于定量。峰偏移的其它示例值是40、45、50、55、65、70和75bp。窗的其它示例值是2、3、4、5、6、7、8、9、15、20、25和30bp。一个峰可以被鉴定为下游峰,

其中预期更多的下游终止位置。另一个峰可以被鉴定为上游峰,其中期望更多的上游终止位置。对于每种情况,分别使用其组织特异性开放染色质区域计算本研究中研究的7种组织类型的OCF值。

D. 应用

[0290] 上述结果表明,差异定相的血浆DNA片段末端可用于推断cfDNA的组织来源。并且,这些结果表明cfDNA片段化概况与开放染色质区域中的核小体定位有关。进一步的结果表明,可以使用特定组织特异性开放染色质区域的差异定相的血浆DNA片段末端的定量测量来检测该组织类型中的病状。也可以使用除血浆以外的其它无细胞样品。

1. 差异定相的血浆DNA片段末端的定量

[0291] 为了探索在推断血浆DNA池中各种组织的相对贡献的潜力,我们开发了一种新的测量组织特异性开放染色质区域中上游(U)和下游(D)片段末端的差异定相的方法。我们通常将这种策略称为识别方向的cfDNA片段化(OCF)分析,其中可以使用各种OCF值。OCF值可以基于在相对于相关开放染色质区域的中心的偏移位置处的U和D末端信号的差异,所述相关开放染色质区域出现在感兴趣的组织中。来自感兴趣组织的DNA越多,差异将越大,例如,在一个或多个偏移区域中下游峰2749和U末端信号2759之间的差异。

[0292] 如图27A所示,对于将DNA贡献到血浆中的组织,预期在相应的组织特异性开放染色质区域中心的核小体耗竭区域已经发生了大量的血浆DNA片段化。在这样的区域中,U和D末端在距中心~60bp处显示出最高的读取密度(即峰),U和D末端的峰分别位于右手侧和左手侧。在一些实例中,我们测量组织特异性开放染色质区域中的峰(例如,图27C中的阴影区域)周围的20bp窗中的U和D末端信号的差异作为相应组织的OCF值。相反,对于相应的组织没有将DNA贡献到血浆中的组织特异性开放染色质区域(例如,图27B中的ESC),这种模式将不会被预期。

[0293] 结果,对于将DNA贡献到血浆中的组织,预期相应的组织特异性开放染色质区域的正OCF值。否则,OCF值应该为零或负。当然,OCF值的不同定义可以具有相反的关系(即,如果测试组织存在,则预期负值)。使用正值为指示物的定义,负值可由噪声的末端信号产生,其可与测序偏倚(例如,GC偏倚)相关,当这些区域不具有开放染色质结构时,在这些区域中导致稍微更多的DNA。

[0294] 图30显示了根据本发明的实施方案,健康非妊娠受试者群体中的各种组织之间的血浆DNA片段化模式(OCF值)的定量。图31显示了根据本发明的实施方案,健康个体的组织类型的OCF值的表。

[0295] 在图30和图31中显示了在32名健康个体中7种组织类型的OCF值。所有受试者均显示出T细胞和肝脏的正OCF值;此外,在所有情况下T细胞的OCF值均高于肝脏的OCF值($P < 0.001$, 威尔克森符号秩检验)。其它组织类型的OCF值低得多并且接近于零或低于零。这些结果与以前的数据一致,表明在健康个体中,大部分血浆DNA来源于造血系统和肝脏,前者是最主要的来源(84, 102)。因此,我们的结果显示了OCF值在测量不同组织对cfDNA池的相对贡献中的效用。

2. 在非侵入性产前测试中的应用

[0296] 为了证明我们的方法在非侵入性产前测试中的效用,我们从先前的研究中取得了母体血浆DNA测序数据(55)。如之前所讨论的,孕妇血浆中的循环胎儿DNA主要来源于胎盘

(58)。图32A-32D显示了根据本发明的实施方案,血浆DNA片段化模式分析在非侵入性产前测试中的应用。图33显示了根据本发明的实施方案,妊娠受试者的OCF值组织类型的表。

[0297] 图32A显示了晚孕期妊娠病例中胎盘特异性开放染色质区域中的血浆DNA片段化模式。轴与类似的图类似。可以在健康的非妊娠个体中观察到类似于共同的开放染色质区域的强核小体定位模式(图27A)。这些观察结果表明血浆DNA片段化模式分析确实可以检测母体血浆中胎盘DNA的存在。

[0298] 我们使用来自26例早孕期妊娠病例的群体的先前公布的数据进一步研究血浆DNA片段化模式(55)。在该群体中的每一个病例都携带男性胎儿。因此,可以通过分析与Y染色体比对了的读取来确定血浆DNA中的胎儿DNA分数。我们分析了胎盘(妊娠病例较高)和T细胞的血浆DNA片段化,这在妊娠中应随着母亲百分比的降低而降低。

[0299] 图32B显示了健康非妊娠受试者和孕妇之间T细胞的OCF值的比较。图32C显示了健康非妊娠受试者和孕妇之间胎盘的OCF值的比较。总计25,223个开放染色质区域用于T细胞,而55,537个用于胎盘。当与来自非妊娠健康个体的结果比较时,在妊娠样品中T细胞的OCF值显著降低,而仅胎盘的OCF值显示显著升高(图32B和32C; $P<0.001$,曼-惠特尼秩和检验;图33)。只有胎盘的OCF值显示显著升高(图32C; $P<0.001$,曼-惠特尼秩和检验)。因此,OCF值和胎盘DNA之间的相关性表明OCF值可用于测量无细胞样品中的胎儿DNA分数。

[0300] 图32D显示了26位孕妇群体中胎盘的OCF值和胎儿DNA分数之间的相关性。观察到胎盘的OCF值和胎儿DNA分数之间的强正相关(图32D; $R=0.77$; $P<0.001$;皮尔森相关)。值得注意的是,该R值高于通过我们以前的胎儿特异性优选末端位点方法获得的值(其为0.66)(55)。胎儿DNA分数是控制非侵入性产前测试性能的最重要的参数之一。因此,这些结果证明了差异定相的血浆DNA片段末端在非侵入性产前测试中的潜在效用。

3. 肝脏移植与肝细胞癌患者

[0301] 为了研究血浆DNA片段化模式分析在预测肝脏组织的贡献中的性能,取得了来自先前报道的14例肝脏移植患者的群体的血浆DNA测序结果(64)。对于每一种情况,供体和受体都进行基因分型,从而可以鉴定供体特异性信息SNP位点以推断血浆中的供体-DNA分数(64)。供体特异性信息SNP位点具有对供体而不是受体特异的等位基因。图34显示了根据本发明实施方案,肝脏移植患者中的OCF值组织类型的表。最后一列显示使用供体特异性信息SNP位点测定的供体DNA分数。肝脏的OCF值与供体DNA分数之间存在相关性。

[0302] 图35A显示了肝脏移植患者中肝脏的OCF值与供体DNA分数之间的相关性。当对该数据集进行血浆DNA片段化模式分析时,可以观察到肝脏的OCF值与供体DNA分数之间的正相关($R=0.74$, $P=0.0022$,皮尔森相关)。

[0303] 此外,我们还从先前公布的肝细胞癌(HCC)患者群体中取得血浆DNA测序数据(70)。对于这些HCC患者,通过拷贝数畸变分析估计血浆DNA中的肿瘤DNA分数(70),尽管也可以使用其它技术,例如肿瘤特异性等位基因。通过这种分析,74个HCC血浆样品显示出血浆中存在肿瘤DNA的证据。值得注意的是,在这些HCC患者中,认为肿瘤来源的cfDNA分子起源于肝脏,因为它们仅在肝脏中具有肿瘤(102,64)。

[0304] 图35B显示了HCC病例中的肿瘤DNA分数。图36A-36D显示了根据本公开的实施方案,肝细胞癌患者中的OCF值组织类型的表。观察到肝脏的OCF值与肿瘤DNA分数之间的正相关($R=0.36$, $P=0.0017$,皮尔森相关)。

[0305] 此外,我们根据肿瘤DNA分数将HCC患者分成两个亚组:“低肿瘤DNA负荷”组包含肿瘤DNA负荷低于10%的HCC患者,而“高肿瘤DNA负荷”组则包含其余的病例。这种分离是基于肝脏在健康受试者中贡献约10%血浆DNA的知识(102)。

[0306] 图35C显示了健康受试者和HCC病例(根据血浆中的肿瘤DNA负荷分为两组)的T细胞的OCF值的比较。如图35C所示,当与健康受试者比较时,对于两个HCC患者组,T细胞的OCF值显著降低(对于低和高肿瘤DNA负荷组,分别为 $P=0.0035$ 和 $P<0.001$,曼-惠特尼秩和检验)。如本文所解释的,当来自其它组织(在这种情况下是肝脏)的贡献发生显著变化时,T细胞的贡献将下降。

[0307] 图35D显示了健康受试者和HCC病例(根据血浆中的肿瘤DNA负荷分为两组)的肝脏的OCF值的比较。图35D中肝脏的OCF值在低肿瘤DNA负荷组患者中没有显示统计学差异($P=0.080$,曼-惠特尼秩和检验),而在高肿瘤DNA负荷组患者中显著升高($P<0.001$,曼-惠特尼秩和检验)。总之,这些结果表明本发明的技术在肝脏移植监测和癌症测试中具有应用。

4. 结肠直肠癌和癌肺患者中的应用

[0308] 在本研究中新招募了11例结肠直肠癌(CRC)患者的群体。对于每种情况,对血浆DNA进行亚硫酸氢盐测序(参见材料和方法部分),使得可以使用血浆DNA组织映射方法测定结肠贡献。这些结果允许我们探索cfDNA片段化模式分析在BS-seq数据中的使用。在这些个体的血浆DNA中,我们观察到肠特异性开放染色质区域中的特征性片段化模式,其对应于中心处的核小体耗竭和附近的定向良好的核小体阵列(102)。

[0309] 图29A显示了根据本公开的实施方案,在一例CRC患者的肠特异性开放染色质区域中的血浆DNA片段化模式。当存在具有测试的开放染色质区域的组织时,基因组覆盖2905以及与图27A、28A和28B中类似的方式显示在开放染色质区域的中心处的减少。此外,U终止信号2907和D终止信号2909显示将导致正OCF值的定相差异。

[0310] 图37A显示了健康受试者与CRC患者之间T细胞的OCF值的比较。图37B显示了健康受试者和CRC患者之间肠的OCF值的比较。图39显示了根据本公开的实施方案,结肠直肠癌患者中的OCF值组织类型的表。结肠DNA贡献也在图39中提供。

[0311] 对于CRC患者,T细胞的OCF值降低,如当来自另一组织的贡献增加时所预期的。图37B显示了肠开放染色质区域(使用28,456个)的OCF值的相应增加。因此,当与健康受试者比较时,在CRC患者中T细胞的OCF值显著降低,而肠的OCF值显著升高(图37A和37B; $P<0.001$,曼-惠特尼秩和检验)。

[0312] 图37C显示了CRC患者中肠的OCF值与结肠DNA分数(通过血浆DNA组织映射方法推导)之间的相关性。使用血浆DNA组织映射方法测定结肠贡献(102)。可以观察到肠的OCF值和结肠贡献(如使用血浆DNA组织映射方法测量的(102))之间的正相关(图37C; $R=0.89$, $P<0.001$,皮尔森相关)。

[0313] 此外,从Snyder等人产生的数据集中取得9例肺癌患者的血浆DNA测序数据(98)。我们发现,血浆DNA显示特征性片段化,即在这些患者的肺特异性开放染色质区域中,中央核小体耗竭区域(侧翼为有相位良好的核小体阵列)的差异定相的末端特征。

[0314] 图29B显示了根据本公开的实施方案,在一例肺癌患者的肺特异性开放染色质区域中的血浆DNA片段化模式。当存在具有测试的开放染色质区域的组织时,基因组覆盖2955以及与图27A、28A和28B中类似的方式显示在开放染色质区域的中心处的减少。此外,U终止信

号2957和D终止信号2959显示将导致正OCF值的定相差异。

[0315] 图37D显示了健康受试者和肺癌患者之间T细胞的OCF值的比较。图37E显示了健康受试者和肺癌患者之间肺OCF值的比较。图38显示了根据本公开的实施方案,肺癌患者中的OCF值组织类型的表。

[0316] 对于肺癌患者,T细胞的OCF值降低,如当来自另一组织的贡献增加时所预期的。图37E显示了肺开放染色质区域(使用19,701个)的OCF值的相应增加。因此,与健康个体相比,T细胞的OCF值降低,而肺的OCF值升高(对于T细胞和肺,分别为 $P < 0.001$ 和0.025,曼-惠特尼秩和检验)。

E. 识别方向的技术

[0317] 如上所述,提供了使用开放染色质区域的识别方向的分析进行核小体定位概况分析的技术,以及通过这种片段化模式分析定量测定血浆DNA中各种组织的相对贡献。我们还证明了在非侵入性产前测试,器官移植监测以及癌症测试中使用组织特异性开放染色质区域的识别方向的分析的诊断能力。我们表明血浆DNA片段化模式分析在核小体耗竭区域和开放染色质区域周围的定相良好的核小体阵列中具有特征性概况。

1. 示例性结果识别方向的分析的概述

[0318] 追踪cfDNA的来源组织的能力在液体活检中,特别是在预测癌症患者中的来源肿瘤方面是非常感兴趣的。我们显示,通过定量癌症患者的血浆DNA片段化模式,T细胞的OCF值将降低,而肿瘤来源组织的OCF值将增加(例如,图32B、32C、35C、35D、37A、37B、37D和37E)。这些观察结果与以下事实是一致的:在这些患者中,肿瘤组织(和肿瘤周围组织)将DNA释放到血浆中,其:(i)将增加来自该癌症的来源组织的贡献,和(ii)将稀释造血系统的贡献。此外,对CRC病例的结果(图37C)显示,我们的方法与血浆DNA组织映射方法高度一致(102)。

[0319] 有趣的是注意到在亚硫酸氢盐转化的DNA中保持了血浆DNA片段化模式。这可能部分地与我们的文库制备方案有关,由此在亚硫酸氢盐处理之前首先将测序适配子连接到血浆DNA分子上(85)。一些实施方案可以通过以协同方式使用OCF测量和基于甲基化的组织映射来提供加和值,以进一步增强来源组织分析的性能。这里,我们证明OCF分析是一种在不依赖甲基化分析的情况下提供来源组织信息的方法。这可以节省成本。与亚硫酸氢盐测序(BS-seq)相比,标准DNA测序实验更便宜并且涉及更简单的方案。

[0320] 至于进一步的效率提高,U1z等人已经证明了血浆DNA覆盖模式分析在推断基因表达从而揭示癌症患者中肿瘤来源组织方面的潜力(105)。然而,这些作者估计为此目的可能需要血浆中75%的肿瘤DNA分数(105),这在大多数临床情况下是难以实现的。相比之下,本发明的技术可用于具有来自感兴趣组织的低得多的DNA分数的情况。例如,在CRC病例中,当结肠贡献仅为5%时,肠的OCF值高于健康个体的OCF值,如图37A、37B和39中可以看到。因此,这些结果表明这些技术可用于相对早期的癌症病例,其中血浆中的肿瘤DNA负荷可能不高。

[0321] 实施方案可以与靶向大规模平行测序技术整合(87)以分析血浆DNA。由于组织特异性开放染色质区域仅占人基因组的非常小的比例,通过设计杂交探针来捕获这些区域,可以大大降低成本。

[0322] 实施方案可以包括在确定患者的疾病或病况的水平之后治疗患者的疾病或病况。

治疗可以包括任何合适的疗法、药物、化疗、放射或手术,包括在本文提及的参考文献中描述的任何治疗。参考文献中关于治疗的信息通过引用并入本文中。

2. 确定组织类型的比例贡献

[0323] 图40是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型的比例贡献的分类的方法4000的流程。生物样品包括来自包括第一组织类型的多个组织类型的无细胞DNA分子的混合物。与这里描述的其它方法一样,方法4000可以使用计算机系统。第一组织类型的实例包括胎儿组织,肿瘤组织和来自移植器官的组织。方法4000的各方面可以以与方法2300和2400类似的方式进行。

[0324] 在框4010,鉴定第一组基因组位置,其与对应于第一组织类型的一个或多个组织特异性开放染色质区域的中心具有指定的距离。组织特异性开放染色质区域可以通过分析第一组织类型的组织样品来鉴定,例如肝脏、T细胞、结肠、卵巢、乳房等。该组基因组位置可以被指定为一个距离范围。作为实例,组织特异性开放染色质区域的数目可以是至少500、1000、2000、5000、10,000、20,000、30,000、40,000、50,000或更多。

[0325] 作为实例,指定距离可以是离中心的 $\pm X$ 个碱基对,包括数值范围(窗),如本文所述。因此,指定距离可以包括在中心之前的第一距离范围,并且包括在中心之后的第二距离范围。这种组可以由与中心的偏移和围绕该偏移的窗来定义。偏移的示例值是40、45、50、55、60、65、70和75bp。窗的其它示例值是2、3、4、5、6、7、8、9、10、15、20、25和30bp。范围可以是不对称的或对称的。

[0326] 在框4020,分析来自受试者的生物样品的第一多个无细胞DNA分子。无细胞DNA分子的分析可包括确定参考基因组中对应于无细胞DNA分子的两个末端的基因组位置(终止位置)。分析还可以包括基于哪个末端具有基因组位置的较低值,例如如参考基因组中所定义的,将一个末端分类为上游末端,将另一个末端分类为下游末端。可使用各种比对/映射程序来确定末端的基因组位置。框4020的各方面可以以与方法2300的框2320类似的方式进行。

[0327] 在框4030,确定第一数目的第一多个无细胞DNA分子在第一组基因组位置之一具有上游末端。基于对第一多个无细胞DNA分子的分析进行确定。考虑到第一组位置可以被定义为参考基因组中的特定基因组坐标,一旦DNA片段的序列读取被比对,上游末端位置可以与第一组进行比较以确定该末端位置是否落入第一组内。

[0328] 在框4040,确定第二数目的第一多个无细胞DNA分子在第一组基因组位置之一具有下游末端。基于对第一多个无细胞DNA分子的分析进行确定。考虑到第一组位置可以被定义为参考基因组中的特定基因组坐标,一旦DNA片段的序列读取被比对,下游末端位置可以与第一组进行比较以确定该末端位置是否落入第一组内。

[0329] 在框4050,使用第一数目和第二数目计算分离值。分离值可以以多种方式确定,并且可以包括比值和/或差值。分离值可以由多个贡献组成。在使用两个范围的实施方案中(例如,在对应于第一组织类型的组织特异性开放染色质区域的中心的任一侧),分离值可以具有对于第一范围以第一方式(例如,第一公式)确定的分离值的第一贡献,和对于第二范围以第二方式(例如,第二公式)确定的分离值的第二贡献。

[0330] 在一个实例中,分离值可以是OCF值,例如,如由以下定义的:

$$OCF = \sum_{\text{峰-仓}}^{\text{峰+仓}} (D-U) + \sum_{\text{峰-仓}}^{\text{峰+仓}} (U-D)$$

其中D是数字下游,U是数字上游。峰位置可以对应于与中心的偏移,并且仓对应于围绕峰的窗尺寸。这样的和可以在每个位置上进行。这样的和可以以任何顺序进行,例如,确定一个峰的D的总数和该峰的U的总数。可以确定围绕每个中心的一个或两个峰的贡献。一个峰可以被鉴定为下游峰,其中期望更多的下游终止位置。另一个峰可以被鉴定为上游峰,其中期望更多的上游终止位置。当使用两个峰时,可以确定和使用两个下游和两个上游数字,例如,如在上式中。作为另一个实例,可以利用用于该位置的指定公式来确定每个位置的分离值,例如,取决于该位置与哪个峰相关联,可以将不同的公式用于该位置。因此,第一组的每个位置可以具有由公式定义的贡献,该公式包括在该位置具有上游末端的第一数目的无细胞DNA片段和在该位置具有下游末端的第二数目的无细胞DNA片段。

[0331] 在具体的实施方案中,第一范围比中心小50至70个碱基,第二范围为50至70个碱基,并且其中分离值包括:

$$OCF = \sum_{-60-10}^{-60+10} (D-U) + \sum_{60-10}^{60+10} (U-D) \quad \text{其中U为第一数目且D为第二数目。}$$

[0332] 第一数目可以是在第一组中的一个位置(例如,第一范围或第二范围中的特定位置)处的数值U,并且第二数目可以是在相同位置处的数值D。作为另一个实例,第一数目可以是具有在第一范围内的上游末端(例如,对应于上游或下游峰)的无细胞DNA的数目的总和,并且第二数目可以是在相同的第一范围内的无细胞DNA的数目的总和。可以使用来自每个范围的数目对来确定分离值。例如,可以确定在第二范围内的位置处具有上游末端的第三数目的无细胞DNA(例如,以上OCF公式中的第二总和贡献),并且可以确定在第二范围内的位置处具有下游末端的第四数目的无细胞DNA。对分离值的第二贡献可以使用第三和第四数目来确定,例如,如上所述。

[0333] 其它示例分离值可以包括和的比值,而不是差的比值。例如,峰区域中的D末端之和除以峰区域的U末端之和,或两个数目的其它比值,例如分子或分母是在峰区域中具有任一末端的读取的总量)。例如,分离值可以包括第一数目和第二数目的比值。当使用一个以上的峰时,可以为每个峰不同地确定比值(或其它函数)。

[0334] 在框4060,通过将分离值与从一个或多个校准样品确定的一个或多个校准值进行比较来确定第一组织类型的比例贡献的分类,所述一个或多个校准样品的第一组织类型的比例贡献是已知的。在图32D中显示了胎儿组织作为第一组织类型的实例,在图35A中显示了来自移植的肝器官的供体DNA的实例,在图35B中显示了来自肝脏为第一组织类型的肿瘤DNA的实例。作为一个实例,比例贡献的分类可以对应于高于指定百分比以上的范围。另一个实例可以对应于癌症的存在以及本文提供的其他实例,例如,对于框2350,以及如本文所述的其他动作,例如治疗。框4060的各方面可以以与框2350相似的方式进行,例如,涉及用于分类的数值以及与校准值的比较,以及稍后的治疗步骤。

[0335] 图41是根据本公开的实施方案,分析生物样品以确定混合物中第一组织类型是否存在病状的方法4100的流程。生物样品包括来自包括第一组织类型的多个组织类型的无细胞DNA分子的混合物。与这里描述的其它方法一样,方法4100可以使用计算机系统。

第一组织类型的实例包括肿瘤组织和来自移植器官的组织。方法4100的各方面可以以与方法2300,2400和4100类似的方式进行。

[0336] 在框4110,鉴定第一基因组位置,其与对应于第一组织类型的一个或多个组织特异性开放染色质区域的中心具有指定的距离。框4110可以以与图40的框4010类似的方式进行。

[0337] 在框4120,分析来自受试者的生物样品的第一多个无细胞DNA分子。无细胞DNA分子的分析可包括确定参考基因组中对应于无细胞DNA分子的两个末端的基因组位置(终止位置)。分析还可以包括基于哪个末端具有基因组位置的较低值,例如如参考基因组中所定义的,将一个末端分类为上游末端,将另一个末端分类为下游末端。框4020的各方面可以以与方法2300的框2320类似的方式进行。

[0338] 在框4130,确定第一数目的第一多个无细胞DNA分子在第一基因组位置之一具有上游末端。框4130可以以与图40的框4030类似的方式进行。

[0339] 在框4140,确定第二数目的第一多个无细胞DNA分子在第一基因组位置之一具有下游末端。框4140可以以与图40的框4040类似的方式进行。

[0340] 在框4150,使用第一数目和第二数目计算分离值。框4150可以以与图40的框4050相似的方式进行。

[0341] 在框4160,基于分离值与参考值的比较来确定受试者的第一组织类型是否存在病状的分类。作为实例,框4160可以使用利用具有已知分类的训练样品确定的参考值,该训练样品的分离值(例如,OCF)已经被测量。图37B和37E提供了训练样品的示例组,其中病状是来自特定组织(即肺)的癌症。因此,病状可以是第一组织类型的癌症。还可以更具体地确定癌症的水平,例如,如图35C或35D所示。

[0342] 因此,可以根据不具有病状的一个或多个对照样品,和/或根据具有病状的一个或多个对照样品来确定参考值。

[0343] 病状的另一个实例是移植器官的排斥。如果移植的器官被排斥,来自该器官的DNA的浓度分数将增加到异常水平。病状的另一个实例是来自第一组织类型的无细胞DNA的异常高的浓度分数。其它示例性病状可以包括自身免疫攻击(例如,损伤肾脏的狼疮性肾炎)、炎性疾病(例如,肝炎)和缺血性组织损伤(例如,心肌梗塞)。受试者的健康状态可以被认为是无病状的分类。样品处理。

VIII. 材料和方法

A. 样品处理

[0344] 将外周血收集在含有EDTA的管中并在4°C以 $1,600 \times g$ 离心10分钟。将血浆部分在4°C以 $16,000 \times g$ 再离心10分钟以获得无细胞血浆并储存在-80°C。将白细胞和红细胞部分用ACK裂解缓冲液(Gibco)以1:10的比值在室温下处理5分钟以去除红细胞。将混合物在4°C以 $300 \times g$ 离心10分钟。丢弃具有裂解的红细胞的上清液,并用磷酸盐缓冲盐水(Gibco)洗涤白细胞团。将白细胞部分在4°C下以 $300 \times g$ 再离心10分钟以去除残留红细胞。将约50,000个细胞用于下游ATAC-seq文库制备。

[0345] 收集来自胎盘的组织并用磷酸盐缓冲盐水(Gibco)洗涤,然后通过Medimachine(BD Biosciences)解聚成单细胞溶液。分别用针对CD105的抗体(Miltenyi Biotec)和针对HAI-1的抗体(Abcam)处理来自胎盘组织的正选择的合胞体滋养层和细胞滋养层。通过用磷

酸盐缓冲液(Gibco)稀释MACS BSA储备溶液(Miltenyi Biotec),将匀浆的胎盘细胞重悬于80 μ L的0.5%牛血清白蛋白缓冲液中。为了分离合胞体滋养层,加入20 μ L CD105微珠(Miltenyi Biotec),并在4 $^{\circ}$ C下孵育15分钟。在合胞体滋养层结合到抗体包被的珠上之后,通过加入2mL缓冲液洗涤细胞,并在200 \times g下离心10分钟。将标记的细胞重悬于500 μ L缓冲液中用于分离步骤。为了分离细胞滋养层,将20 μ L的HAI-1抗体(Abcam)和80 μ L的缓冲液添加到匀浆的胎盘组织中并且在4 $^{\circ}$ C下孵育15分钟。孵育后,添加2mL的缓冲液,通过在200 \times g下离心10分钟来洗去过量的初级抗体。将细胞重悬于80 μ L缓冲液中,并加入20 μ L第二抗小鼠IgG微珠(Miltenyi Biotec),并在4 $^{\circ}$ C下孵育15分钟。与第一抗体类似,加入2mL缓冲液,通过在200 \times g下离心10分钟来洗去过量的初级抗体。将标记的细胞重悬于500 μ L缓冲液中用于分离步骤。每个细胞类型的每个样品使用一个MS柱(Miltenyi Biotec)。在施加标记的细胞之前,我们用500 μ L缓冲液冲洗柱子。通过将细胞施加到柱中,将标记的细胞附着到柱中的磁珠上,并将未标记的细胞留在流通中。洗涤柱3次,每次用500 μ L缓冲液。将分选的合胞体滋养层和细胞滋养层洗脱在1mL缓冲液中,并通过血细胞计数器计数,每份样品等分50,000个细胞用于ATAC-seq。

B. ATAC-seq文库的制备和测序

[0346] 如(35)所述进行ATAC-seq。简言之,将50,000个细胞在4 $^{\circ}$ C下以500 \times g旋转5分钟,然后使用冷裂解缓冲液(10mM Tris-HCl, pH7.4(Ambion), 10mM NaCl(Ambion), 3mM MgCl₂(Ambion)和0.1% IGEPAL CA-630(Sigma))进行细胞裂解。将混合物立即在4 $^{\circ}$ C下以500 \times g离心10分钟。将细胞核重悬于转座酶反应混合物中,所述转座酶反应混合物含有25 μ L 2 \times TD缓冲液、来自Nextera DNA文库制备试剂盒(Illumina)的2.5 μ L转座酶和22.5 μ L无核酸酶的水。转座和标记在37 $^{\circ}$ C进行30分钟。在转座后立即用Qiagen Min Elute试剂盒(Qiagen)按照制造商的说明书纯化样品。将纯化的DNA片段与1 \times NEBnext PCR主混合物(New England BioLabs)和1.25 μ M Nextera PCR引物1和2(IDT)混合,用于使用以下条件进行PCR扩增:72 $^{\circ}$ C 5分钟;98 $^{\circ}$ C 30秒;98 $^{\circ}$ C 10秒,63 $^{\circ}$ C 30秒和72 $^{\circ}$ C 1分钟,热循环15个循环。用Qiagen PCR清理试剂盒(Qiagen)纯化文库。用2100Bioanalyzer(Agilent)分析文库,并在测序前用KAPA文库定量试剂盒(Kapa Biosystems)定量。在Hi-Seq 2500(Illumina)上进行2 \times 75配对末端测序。

C. 测序数据的比对

[0347] 在实例中,使用配对末端模式的SOAP2比对器(53)将配对末端读取映射到参考人基因组(NCBI37/hg19),允许对于每个末端的比对有两个错配。只有两个末端以正确的方向与同一染色体比对了的,跨越 \leq 600bp的插入片段尺寸的配对末端读取用于下游分析。可以使用其它比对技术(软件),例如BLAST、BLAT、BWA、Bowtie、STAR等。如果整个DNA片段被测序,则不需要配对末端模式。此外,错配的数目可以根据期望的精度而变化。

D. 血浆DNA数据收集和可用性

[0348] 健康个体、HCC患者和妊娠病例的血浆数据从欧洲基因组-表型档案(EGA;登录号EGAS00001001024和EGAS00001001882)取得(70,55)。我们以前工作中描述的肝脏移植患者的血浆DNA测序数据(64)已经保藏在EGA(登录号EGAS00001003116)。从基因表达Omnibus(GEO;登录号GSE71378)获得肺癌病例的血浆DNA测序数据(98)。

[0349] 在本研究中新招募了结肠直肠癌患者。将外周血样品收集到含有EDTA的管中。将

血液样品在4°C以1,600×g离心10分钟。收集血浆部分,并在4°C以16,000×g再离心10分钟以去除血细胞。亚硫酸氢盐转化如之前所述进行(85)。使用KAPA HTP文库制备试剂盒(Kapa Biosystems)根据制造商的说明书(56)制备DNA文库,并在HiSeq 2000系统(Illumina)上以75×2(配对末端模式)循环模式用TruSeq SBS试剂盒v3(Illumina)测序。如之前所述(71, 102)进行BS-seq数据的分析,包括质量控制,序列比对,甲基化状态测定和结肠贡献推断。这些样品的中值测序深度为3.2倍(范围:0.6-6.4倍;图39)单倍体人基因组覆盖。

E. 组织特异性开放染色质区域

[0350] 开放染色质区域是基因组中重要的调节元件,并且是高度组织特异性的。活性启动子是一种类型的开放染色质区域。其它类型包括增强子和绝缘子。开放染色质区域可以使用感兴趣的组织的公共Dnase-seq数据来确定。Dnase-seq是使用DNaseI内切核酸酶处理细胞基因组DNA的实验程序,其优选切割非核小体结合的DNA。结果,开放染色质区域中的DNA被切割并收集用于测序。因此,我们可以将这些DNA坐标鉴定为开放染色质区域,例如,如图25D所示。对于每个区域,获得其开始和结束的基因组坐标,并且可以使用中间坐标(即(开始+结束)/2)作为中心。

[0351] 在从每种组织类型的Dnase-seq数据获得开放染色质区域之后,可以将开放染色质区域相互比较,并且只有那些一种组织类型特有的区域可以被保留并定义为“组织特异性”区域,用于进一步分析,如本文所述。对于这些组织特异性开放染色质区域,核小体仅在相应的组织类型中定位良好,从而允许确定血浆DNA中的比例贡献。除Dnase-seq外,鉴定开放染色质区域的其它示例方法包括CTCF转录因子上的FAIRE-seq、ATAC-seq、MNASE-seq和ChIP-seq。

[0352] 在一些实施方案中,我们使用可公开获得的DNase-seq(DNaseI高敏感位点测序)数据来挖掘开放染色质区域。从RoadMap Epigenomics项目获得T细胞,胎盘,肺,卵巢,乳房和小肠的DNase-seq数据(93)。从ENCODE项目获得肝脏和ESC的DNase-seq数据(104)。对于每种组织类型,下载原始测序数据并使用蝴蝶结比对软件(1.1.1版)与参考人基因组(UCSC hg19)比对(76)。然后,使用MACS(ChIP-Seq的基于模型的分析)软件(2.0.9版)确定开放染色质区域(110,74)。可以使用其它参考基因组和比对软件。

[0353] 对于这样的分析,ChIP-seq(染色质免疫沉淀,随后大规模平行DNA测序)输入数据用作阴性对照,并且0.01的Q值(即,反映错误发现率的调整的P值)用作呼叫峰的阈值。对于肺,分析IMR90(人胎儿的肺)和HLF(人肺成纤维细胞)细胞系的DNase-seq数据,并且仅鉴定存在于两个样品中的峰。然后,对于每种组织类型,我们将其峰与所有其它组织进行比较,并且仅保留该组织类型所特有的峰并且在50-200bp的尺寸范围内作为最终的组织特异性开放染色质区域。

IX. 实例系统

[0354] 图42说明根据本公开的实施方案的测量系统4200。所示系统包括样品4205,如样品保持器4210内的无细胞DNA分子,其中样品4205可以与分析器4208接触以提供物理特征4215的信号。样品保持器的实例可以是流动池,其包括分析器的探针和/或引物或液滴通过其移动的管(液滴包括于分析器中)。通过检测器4220检测来自样品的物理特征4215(如荧光强度、电压或电流)。检测器可以间隔地(例如,周期性间隔)进行测量以获得组成数据信号的数据点。在一个实施方案中,模数转换器多次将来自检测器的模拟信号转换成数字形

式。数据信号4225从检测器4220发送到逻辑系统4230。样品保持器4210和检测器4220可以形成测定装置,例如,根据本文所述的实施方案进行测序的测序装置。数据信号4225从检测器4220发送到逻辑系统4230。数据信号4225可以存储在局部存储器4235、外部存储器4240或存储装置4245中。

[0355] 逻辑系统4230可以是或可以包括计算机系统、ASIC、微处理器等。其还可以包括显示器(例如监测器、LED显示器等)和用户输入装置(例如鼠标、键盘、按钮等)或与其耦接。逻辑系统4230和其它组件可以是独立或网络连接的计算机系统的一部分,或者其可以直接连接或整合在包括检测器4220和/或样本保持器4210的装置(例如,测序装置)中。逻辑系统4230还可以包括在处理器4250中执行的软件。逻辑系统4230可以包括存储用于控制系统4200执行本文所述的任何方法的指令的计算机可读介质。例如,逻辑系统4230可以向包括样本保持器4210的系统提供命令,从而执行测序或其它物理操作。这样的物理操作可以以特定的顺序进行,例如,以特定的顺序加入和除去试剂。这种物理操作可以由机器人系统执行,例如,包括机器人臂,其可以用于获得样品和执行分析。

[0356] 本文中提及的任何计算机系统(例如逻辑系统4230)都可以利用任何合适数目的子系统。这类子系统的实例展示于计算机系统10中的图43中。在一些实施方案中,计算机系统包括单个计算机设备,其中子系统可以是计算机设备的组件。在其它实施方案中,计算机系统可以包括具有内部组件的多个计算机设备,每个计算机设备是子系统。计算机系统可以包括桌面计算机和膝上型计算机、平板计算机、移动电话和其它移动装置。

[0357] 图43中展示的子系统是通过系统总线75互连。展示其它子系统,如打印机74、键盘78、存储装置79、与显示适配器82耦接的监测器76等。耦接到I/O控制器71的外围装置和输入/输出(I/O)装置可以通过本领域中已知的任何数目的装置,如输入/输出(I/O)端口77(例如USB、FireWire[®])连接到计算机系统。举例来说,I/O端口77或外部接口81(例如以太网、Wi-Fi等)可以用于将计算机系统10连接到广域网,如因特网、鼠标输入装置或扫描仪。通过系统总线75的互连允许中央处理器73与每个子系统通信并且控制来自系统存储器72或存储装置79(例如固定磁盘,如硬盘驱动器或光盘)的多个指令的执行,以及子系统之间的信息交换。系统存储器72和/或存储装置79可以体现为计算机可读介质。另一种子系统是数据收集装置85,如相机、麦克风、加速计等。本文中提及的任何数据可以从一个组件输出到另一个组件且可以输出到用户。

[0358] 计算机系统可以包括例如通过外部接口81、通过内部接口或经由可移动存储装置(其可从一个组件连接到另一个组件并从其移除)连接在一起的多个相同组件或子系统。在一些实施方案中,计算机系统、子系统或设备可以通过网络进行通信。在这类情况下,一个计算机可以视为客户端且另一个计算机视为服务器,其中每一个可以是同一个计算机系统的一部分。客户端和服务器可以各自包含多个系统、子系统或组件。

[0359] 实施方案的各方面可以按使用硬件电路的控制逻辑(例如专用集成电路或现场可编程门阵列)形式实施,和/或借助于通用可编程处理器使用计算机软件以模块或集成的方式实施。如本文中所使用,处理器可以包括单核处理器、在同一集成芯片上的多核处理器,或在单个电路板上或网络化的多个处理单元以及专用硬件。基于本文中提供的公开和教导,本领域技术人员将知晓和理解使用硬件以及硬件和软件的组合实施本公开的实施方案的其它方式和/或方法。

[0360] 本申请中描述的任何软件组件或功能可以实施为使用任何适当计算机语言(例如Java、C、C++、C#、Objective-C、Swift)或脚本语言(如Perl或Python),使用例如传统或面向受试者技术由处理器执行的软件代码。软件代码可以存储为计算机可读介质上用于存储和/或传输的一系列指令或命令。适合的非暂时性计算机可读介质可以包括随机存取存储器(RAM)、只读存储器(ROM)、如硬盘驱动器或软盘等磁性媒体或如光盘(CD)或DVD(数字通用光盘)等光学介质、闪存等。计算机可读介质可以是这类存储或传输装置的任何组合。

[0361] 这类程序还可以使用适合于通过符合多种协议的有线、光学和/或无线网络(包括因特网)传送的载波信号来编码和传输。因此,计算机可读介质可以使用以这类程序编码的数据信号产生。以程序代码编码的计算机可读介质可以与兼容装置一起封装或其它装置分开提供(例如,通过因特网下载)。任何这类计算机可读介质可以驻留在单个计算机产品(例如,硬盘驱动器,CD或整个计算机系统)之上或之内,并且可以存在于系统或网络内的不同计算机产品之上或之内。计算机系统可以包括监视器、打印机,或用于向用户提供本文中提及的任何结果的其它适合的显示器。

[0362] 本文中所描述的任何方法可以完全或部分用计算机系统执行,所述计算机系统包括一或多个处理器,所述处理器可以经配置以执行所述步骤。因此,实施方案可以涉及经配置以执行本文中所描述的任何方法的步骤的计算机系统,所述计算机系统可能具有用于执行各步骤或各步骤组的不同组件。尽管以编号的步骤呈现,但本文中的方法的步骤可以同时或不同时或按不同的顺序执行。此外,一部分这些步骤可以与其它方法的一部分其它步骤一起使用。并且,所有或部分步骤可以是任选的。此外,任何方法中的任何步骤可以借助于用于执行这些步骤的模块、单元、电路或其它方法执行。

[0363] 在不脱离本发明实施方案的精神和范围的情况下,特定实施方案的具体细节可以以任何合适的方式组合。然而,本发明的其它实施方案可涉及与每个单独方面或这些单独方面的具体组合相关的具体实施方案。

[0364] 为了说明和描述的目的,已经给出了本发明的示例性实施方案的上述描述。并不是要穷举或将本发明限制于所描述的精确形式,并且根据以上教导,许多修改和变化是可能的。

[0365] 除非特别指出相反,否则“一个”、“一种”或“所述”的表述旨在表示“一个或多个”。除非特别指出相反,否则“或”的使用旨在表示“逻辑或”,而不是“互斥或”。提及“第一”部件不一定要提供第二部件。此外,除非明确说明,否则提及“第一”或“第二”组件并不将所提及的组件限制到特定位置。术语“基于”旨在表示“至少部分基于”。

[0366] 本文所提及的所有专利、专利申请、出版物和描述通过引用并入用于所有目的。任一个均未被认为是现有技术。

X. 参考文献

1. LO YMD, ET AL. (1997) PRESENCE OF FETAL DNA IN MATERNAL PLASMA AND SERUM. LANCET 350 (9076):485-487.

2. Lo YMD, et al. (1998) Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. Lancet 351 (9112):1329-1330.

3. Ulz P, Heitzer E, Geigl JB, & Speicher MR (2017) Patient monitoring through liquid biopsies using circulating tumor DNA. Int J Cancer 141 (5):887-896.

4. Cohen JD, et al. (2018) Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 359 (6378):926-930.
5. Schutz E, et al. (2017) Graft-derived cell-free DNA, a noninvasive early rejection and graft damage marker in liver transplantation: A prospective, observational, multicenter cohort study. *PLoS Med* 14 (4):e1002286.
6. Chan KCA, et al. (2017) Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer. *N Engl J Med* 377 (6):513-522.
7. Lehmann-Werman R, et al. (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* 113 (13):E1826-1834.
8. van Opstal D, et al. (2017) Origin and clinical relevance of chromosomal aberrations other than the common trisomies detected by genome-wide NIPS: results of the TRIDENT study. *Genet Med* Oct 2. doi:10.1038/gim.2017.132.
9. Lo YMD, et al. (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2 (61):61ra91.
10. Struhl K & Segal E (2013) Determinants of nucleosome positioning. *Nat Struct Mol Biol* 20 (3):267-273.
11. Chim SSC, et al. (2005) Detection of the placental epigenetic signature of the maspin gene in maternal plasma. *Proc Natl Acad Sci U S A* 102 (41):14753-14758.
12. Sun K, et al. (2015) Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci U S A* 112 (40):E5503-5512.
13. Lui YYN, et al. (2002) Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem* 48 (3):421-427.
14. Chan KCA, et al. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem* 50 (1):88-92.
15. Sun K, et al. (2018) Noninvasive reconstruction of placental methylome from maternal plasma DNA: potential for prenatal testing and monitoring. *Prenat Diagn* 38 (3):196-203.
16. Sun K, et al. (2017) COFFEE: control-free noninvasive fetal chromosomal examination using maternal plasma DNA. *Prenat Diagn* 37 (4):336-340.
17. Yu SCY, et al. (2014) Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc Natl Acad Sci U S A* 111 (23):8583-8588.
18. Cirigliano V, Ordonez E, Rueda L, Syngelaki A, & Nicolaides KH (2017) Performance of the neoBona test: a new paired-end massively parallel shotgun sequencing approach for cell-free DNA-based aneuploidy screening. *Ultrasound Obstet Gynecol* 49 (4):460-464.

19.Zhang L,Zhu Q,Wang H,&Liu S(2017)Count-based size-correction analysis of maternal plasma DNA for improved noninvasive prenatal detection of fetal trisomies 13,18,and 21.Am J Transl Res 9(7):3469-3473.

20.Yu SCY,et al.(2013)High-resolution profiling of fetal DNA clearance from maternal plasma by massively parallel sequencing.Clin Chem 59(8):1228-1237.

21.Chan KCA,et al.(2016)Second generation noninvasive fetal genome analysis reveals de novo mutations,single-base parental inheritance,and preferred DNA ends.Proc Natl Acad Sci U S A 113(50):E8159-E8168.

22.Jahr S,et al.(2001)DNA fragments in the blood plasma of cancer patients:quantitations and evidence for their origin from apoptotic and necrotic cells.Cancer Res 61(4):1659-1665.

23.Straver R,Oudejans CB,Sistermans EA,&Reinders MJ(2016)Calculating the fetal fraction for noninvasive prenatal testing based on genome-wide nucleosome profiles.Prenat Diagn 36(7):614-621.

24.Snyder MW,Kircher M,Hill AJ,Daza RM,&Shendure J(2016)Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin.Cell 164(1-2):57-68.

25.Ivanov M,Baranova A,Butler T,Spellman P,&Mileyko V(2015)Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation.BMC Genomics 16 Suppl 13:S1.

26.Chiu RWK,et al.(2008)Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma.Proc Natl Acad Sci U S A 105(51):20458-20463.

27.DeLong ER,DeLong DM,&Clarke-皮尔森DL(1988)Comparing the areas under two or more correlated receiver operating characteristic curves:a nonparametric approach.Biometrics 44(3):837-845.

28.Jiang P,et al.(2015)Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.Proc Natl Acad Sci U S A 112(11):E1317-1325.

29.Valouev A,et al.(2011)Determinants of nucleosome organization in primary human cells.Nature 474(7352):516-520.

30.Gaffney DJ,et al.(2012)Controls of nucleosome positioning in the human genome.PLoS Genet 8(11):e1003036.

31.Lam WKJ,et al.(2017)DNA of erythroid origin is present in human plasma and informs the types of anemia.Clin Chem 63(10):1614-1623.

32.Roadmap Epigenomics Consortium,et al.(2015)Integrative analysis of 111reference human epigenomes.Nature 518(7539):317-330.

33.Jiang C&Pugh BF(2009)Nucleosome positioning and gene regulation:

advances through genomics. *Nat Rev Genet* 10 (3):161–172.

34. Horlbeck MA, et al. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife* 5:e12677.

35. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, & Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10 (12):1213–1218.

36. Mueller B, et al. (2017) Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes Dev* 31 (5):451–462.

37. Buenrostro JD, Wu B, Chang HY, & Greenleaf WJ (2015) ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol* 109:21.29.1–9.

38. Schep AN, et al. (2015) Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* 25 (11):1757–1770.

39. Chodavarapu RK, et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466 (7304):388–392.

40. Jensen TJ, et al. (2015) Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biol* 16:78.

41. Lun FMF, et al. (2013) Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin Chem* 59 (11):1583–1594.

42. Jiang P, et al. (2017) Gestational age assessment by methylation and size profiling of maternal plasma DNA: a feasibility study. *Clin Chem* 63 (2):606–608.

43. Schroeder DI, et al. (2013) The human placenta methylome. *Proc Natl Acad Sci U S A* 110 (15):6037–6042.

44. Lee JY & Lee TH (2012) Effects of DNA methylation on the structure of nucleosomes. *J Am Chem Soc* 134 (1):173–175.

45. Choy JS, et al. (2010) DNA methylation increases nucleosome compaction and rigidity. *J Am Chem Soc* 132 (6):1782–1783.

46. Collings CK, Waddell PJ, & Anderson JN (2013) Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res* 41 (5):2918–2931.

47. Rose NR & Klose RJ (2014) Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta* 1839 (12):1362–1372.

48. Soppe WJ, et al. (2002) DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J* 21 (23):6549–

6559.

49.Simon M,et al.(2011)Histone fold modifications control nucleosome unwrapping and disassembly.Proc Natl Acad Sci U S A 108(31):12711-12716.

50.Ehrlich M(2009)DNA hypomethylation in cancer cells.Epigenomics 1(2):239-259.

51.Chan KCA,et al.(2013)Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.Proc Natl Acad Sci U S A 110(47):18761-18768.

52.Holtan SG,Creedon DJ,Haluska P,&Markovic SN(2009)Cancer and pregnancy: parallels in growth,invasion,and immune modulation and implications for cancer therapeutic agents.Mayo Clin Proc 84(11):985-1000.

53.Li R,et al.(2009)SOAP2:an improved ultrafast tool for short read alignment.Bioinformatics 25(15):1966-1967.

54.Chan KCA,Jiang P,Chan CW,Sun K,Wong J,Hui EP,Chan SL,Chan WC,Hui DS,Ng SS et al.2013a.Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing.Proc Natl Acad Sci U S A 110(47):18761-18768.

55.Chan KCA,Jiang P,Sun K,Cheng YK,Tong YK,Cheng SH,Wong AI,Hudecova I,Leung TY,Chiu RWK et al.2016.Second generation noninvasive fetal genome analysis reveals de novo mutations,single-base parental inheritance,and preferred DNA ends.Proc Natl Acad Sci U S A 113(50):E8159-E8168.

56.Chan KCA,Jiang P,Zheng YW,Liao GJ,Sun H,Wong J,Siu SS,Chan WC,Chan SL,Chan AT et al.2013b.Cancer genome scanning in plasma:detection of tumor-associated copy number aberrations,single-nucleotide variants,and tumoral heterogeneity by massively parallel sequencing.Clin Chem 59(1):211-224.

57.Chan KCA,Woo JKS,King A,Zee BCY,Lam WKJ,Chan SL,Chu SWI,Mak C,Tse IOL,Leung SYM et al.2017.Analysis of plasma Epstein-Barr virus DNA to screen for nasopharyngeal cancer.N Engl J Med 377(6):513-522.

58.Chim SSC,Tong YK,Chiu RW,Lau TK,Leung TN,Chan LY,Oudejans CB,Ding C,Lo YM.2005.Detection of the placental epigenetic signature of the maspin gene in maternal plasma.Proc Natl Acad Sci U S A 102(41):14753-14758.

59.Christie EL,Fereday S,Doig K,Pattnaik S,Dawson SJ,Bowtell DDL.2017.Reversion of BRCA1/2 germline mutations detected in circulating tumor DNA from patients with high-grade serous ovarian cancer.J Clin Oncol 35(12):1274-1280.

60.Cleveland WS.1979.Robust locally weighted regression and smoothing scatterplots.Journal of the American Statistical Association 74(368):829-836.

61.Cohen JD,Li L,Wang Y,Thoburn C,Afsari B,Danilova L,Douville C,Javed AA,Wong F,Mattox A et al.2018.Detection and localization of surgically

- resectable cancers with a multi-analyte blood test. *Science* 359(6378):926-930.
62. Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* 29(10):569-574.
63. Gaffney DJ, McVicker G, Pai AA, Fondufe-Mittendorf YN, Lewellen N, Michelini K, Widom J, Gilad Y, Pritchard JK. 2012. Controls of nucleosome positioning in the human genome. *PLoS Genet* 8(11):e1003036.
64. Gai W, Ji L, Lam WKJ, Sun K, Jiang P, Chan AWH, Wong J, Lai PBS, Ng SSM, Ma BBY et al. 2018. Liver- and colon-specific DNA methylation markers in plasma for investigation of colorectal cancers with or without liver metastases. *Clin Chem* (doi:10.1373/clinchem.2018.290304).
65. Grunau C, Clark SJ, Rosenthal A. 2001. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Res* 29(13):E65-65.
66. Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, Griffin C, Yang A, Huang P, Wrangle J et al. 2017. Early detection of lung cancer using DNA promoter hypermethylation in plasma and sputum. *Clin Cancer Res* 23(8):1998-2005.
67. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. 2015. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* 16 Suppl 13:S1.
68. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, Knippers R. 2001. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res* 61(4):1659-1665.
69. Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 10(3):161-172.
70. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, Wong VW, Wong GL, Chan SL, Mok TS, Chan HL et al. 2015. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A* 112(11):E1317-1325.
71. Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA, Chiu RWK, Lo YMD, Sun H. 2014. Methy-pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. *PLoS One* 9(6):e100360.
72. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, Grimes B, Krysan K, Yu M, Wang W et al. 2017. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* 18(1):53.
73. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. 2010. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res* 38(Database issue):D690-698.

74. Koohy H, Down TA, Spivakov M, Hubbard T. 2014. A comparison of peak callers used for DNase-Seq data. *PLoS One* 9(5):e96303.

75. Lam WKJ, Gai W, Sun K, Wong RSM, Chan RWY, Jiang P, Chan NPH, Hui WWI, Chan AWH, Szeto CC et al. 2017. DNA of erythroid origin is present in human plasma and informs the types of anemia. *Clin Chem* 63(10):1614-1623.

76. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.

77. Lehmann-Werman R, Magenheimer J, Moss J, Neiman D, Abraham O, Piyanzin S, Zemmour H, Fox I, Dor T, Grompe M et al. 2018. Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI Insight* 3(12).

78. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheimer J, Vaknin-Dembinsky A, Rubertsson S, Nellgard B, Blennow K, Zetterberg H et al. 2016. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A* 113(13):E1826-1834.

79. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, Liu CC, Matsuoka L, Sher L, Wong WH et al. 2018. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res* (doi:10.1093/nar/gky423).

80. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133(3):523-536.

81. Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FM, Zheng YW, Leung TY, Lau TK, Cantor CR et al. 2010. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med* 2(61):61ra91.

82. Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, Wainscoat JS. 1997. Presence of fetal DNA in maternal plasma and serum. *Lancet* 350(9076):485-487.

83. Lo YMD, Tein MS, Pang CC, Yeung CK, Tong KL, Hjelm NM. 1998. Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients. *Lancet* 351(9112):1329-1330.

84. Lui YYN, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YM. 2002. Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem* 48(3):421-427.

85. Lun FMF, Chiu RWK, Sun K, Leung TY, Jiang P, Chan KC, Sun H, Lo YM. 2013. Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. *Clin Chem* 59(11):1583-1594.

86. Mandel P, Metais P. 1948. Les acides nucléiques du plasma sanguin chez l'

homme.C R Seances Soc Biol Fil 142(3-4):241-243.

87.Mertes F,Elsharawy A,Sauer S,van Helvoort JM,van der Zaag PJ,Franke A, Nilsson M,Lehrach H,Brookes AJ.2011.Targeted enrichment of genomic DNA regions for next-generation sequencing.Brief Funct Genomics 10(6):374-386.

88.O'Leary B,Hrebien S,Morden JP,Beaney M,Fribbens C,Huang X,Liu Y, Bartlett CH,Koehler M,Cristofanilli M et al.2018.Early circulating tumor DNA dynamics and clonal selection with palbociclib and fulvestrant for breast cancer.Nat Commun 9(1):896.

89.Olova N,Krueger F,Andrews S,Oxley D,Berrens RV,Branco MR,Reik W.2018.Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNAmethylation data.Genome Biol 19(1):33.

90.Pedersen JS,Valen E,Velazquez AM,Parker BJ,Rasmussen M,Lindgreen S, Lilje B,Tobin DJ,Kelly TK,Vang S et al.2014.Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome.Genome Res 24(3):454-466.

91.Phallen J,Sausen M,Adeff V,Leal A,Hruban C,White J,Anagnostou V, Fiksel J,Cristiano S,Papp E et al.2017.Direct detection of early-stage cancers using circulating tumor DNA.Sci Transl Med 9(403).

92.Radman-Livaja M,Rando OJ.2010.Nucleosome positioning:how is it established,and why does it matter?Dev Biol 339(2):258-266.

93.Roadmap Epigenomics Consortium,Kundaje A,Meuleman W,Ernst J,Bilenky M, Yen A,Heravi-Moussavi A,Kheradpour P,Zhang Z,Wang J et al.2015.Integrative analysis of 111 reference human epigenomes.Nature 518(7539):317-330.

94.Samejima K,Earnshaw WC.2005.Trashing the genome:the role of nucleases during apoptosis.Nat Rev Mol Cell Biol 6(9):677-688.

95.Schep AN,Buenrostro JD,Denny SK,Schwartz K,Sherlock G,Greenleaf WJ.2015.Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions.Genome Res 25(11):1757-1770.

96.Schones DE,Cui K,Cuddapah S,Roh TY,Barski A,Wang Z,Wei G,Zhao K.2008.Dynamic regulation of nucleosome positioning in the human genome.Cell 132(5):887-898.

97.Schutz E,Fischer A,Beck J,Harden M,Koch M,Wuensch T,Stockmann M,Nashan B,Kollmar O,Matthaei J et al.2017.Graft-derived cell-free DNA,a noninvasive early rejection and graft damage marker in liver transplantation:A prospective,observational,multicenter cohort study.PLoS Med 14(4):e1002286.

98.Snyder MW,Kircher M,Hill AJ,Daza RM,Shendure J.2016.Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin.Cell 164(1-2):57-68.

99.Strickler JH,Loree JM,Ahronian LG,Parikh AR,Niedzwiecki D,Pereira AAL,McKinney M,Korn WM,Atreya CE,Banks KC et al.2018.Genomic landscape of cell-free DNA in patients with colorectal cancer.Cancer Discov 8(2):164-173.

100.Stroun M,Anker P,Maurice P,Lyautey J,Lederrey C,Beljanski M.1989.Neoplastic characteristics of the DNA found in the plasma of cancer patients.Oncology 46(5):318-322.

101.Struhl K,Segal E.2013.Determinants of nucleosome positioning.Nat Struct Mol Biol 20(3):267-273.

102.Sun K,Jiang P,Chan KCA,Wong J,Cheng YK,Liang RH,Chan WK,Ma ES,Chan SL,Cheng SH et al.2015.Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal,cancer,and transplantation assessments.Proc Natl Acad Sci U S A 112(40):E5503-5512.

103.Sun K,Jiang P,Wong AIC,Cheng YKY,Cheng SH,Zhang H,Chan KCA,Leung TY,Chiu RWK,Lo YMD.2018.Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing.Proc Natl Acad Sci U S A 115(22):E5106-E5114.

104.The ENCODE Project Consortium.2012.An integrated encyclopedia of DNA elements in the human genome.Nature 489(7414):57-74.

105.Ulz P,Heitzer E,Geigl JB,Speicher MR.2017.Patient monitoring through liquid biopsies using circulating tumor DNA.Int J Cancer 141(5):887-896.

106.Ulz P,Thallinger GG,Auer M,Graf R,Kashofer K,Jahn SW,Abete L,Pristauz G,Petru E,Geigl JB et al.2016.Inferring expressed genes by whole-genome sequencing of plasma DNA.Nat Genet 48(10):1273-1278.

107.Valouev A,Johnson SM,Boyd SD,Smith CL,Fire AZ,Sidow A.2011.Determinants of nucleosome organization in primary human cells.Nature 474(7352):516-520.

108.van Opstal D,van Maarle MC,Lichtenbelt K,Weiss MM,Schuring-Blom H,Bhola SL,Hoffer MJV,Huijsdens-van Amsterdam K,Macville MV,Kooper AJA et al.2017.Origin and clinical relevance of chromosomal aberrations other than the common trisomies detected by genome-wide NIPS:results of the TRIDENT study.Genet Med 20(5):480-485.

109.Zemmour H,Planer D,Magenheim J,Moss J,Neiman D,Gilon D,Korach A,Glaser B,Shemer R,Landesberg G et al.2018.Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA.Nat Commun 9(1):1443.

110.Zhang Y,Liu T,Meyer CA,Eckhoute J,Johnson DS,Bernstein BE,Nusbaum C,Myers RM,Brown M,Li W et al.2008.Model-based analysis of ChIP-Seq (MACS).Genome Biol 9(9):R137.

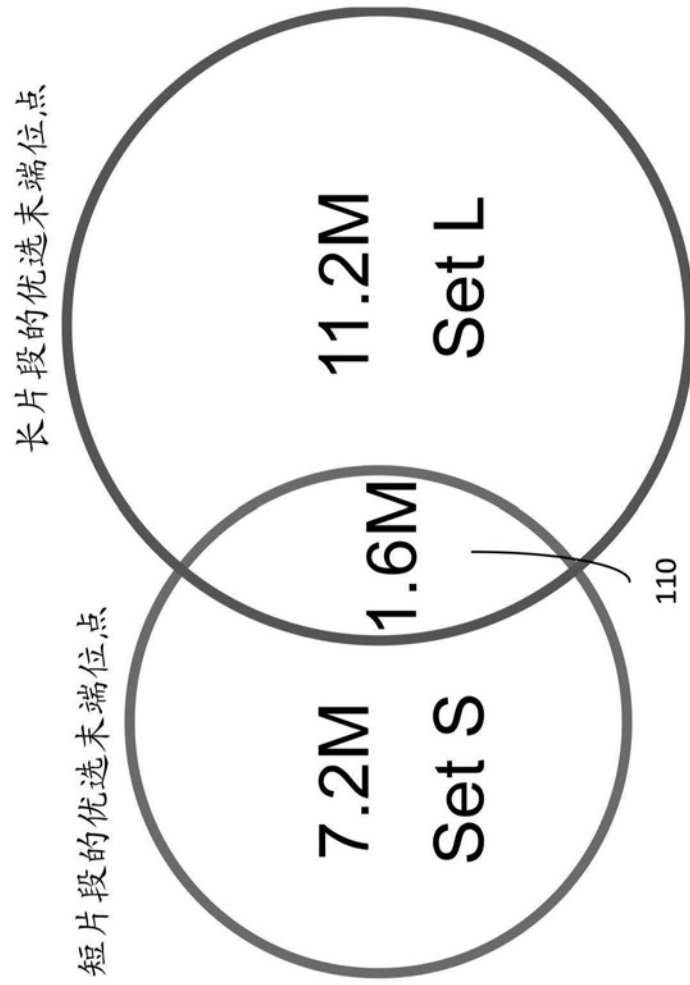


图1

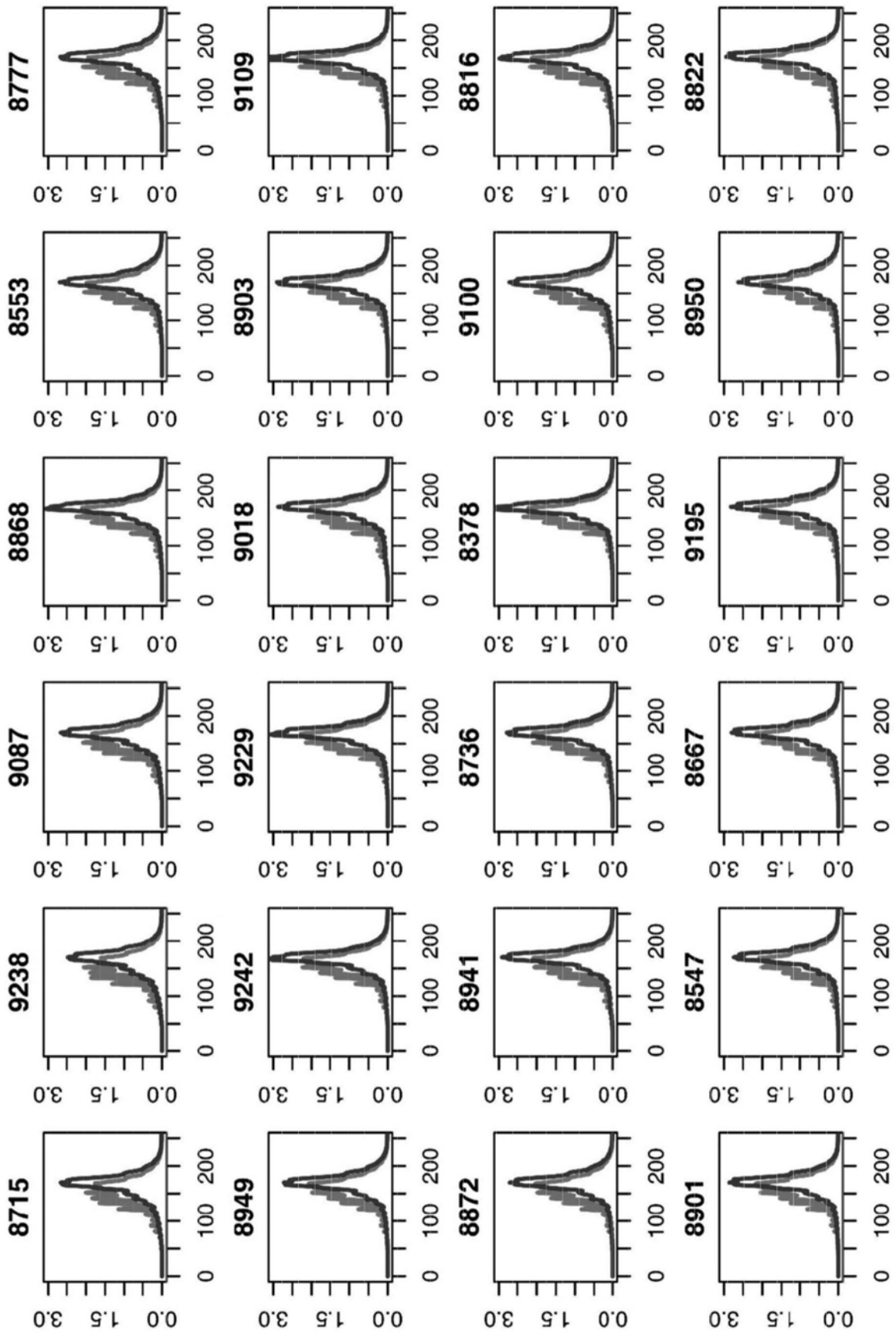


图2

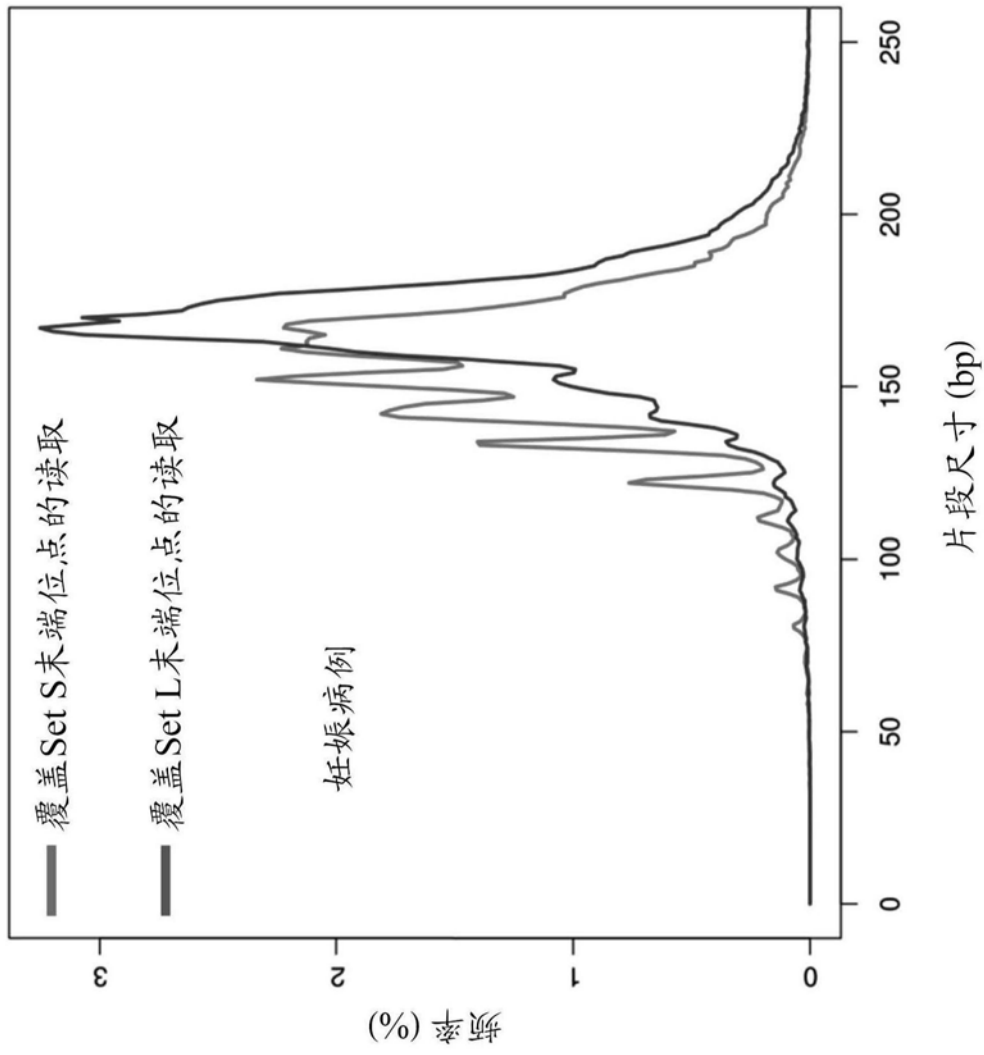


图3

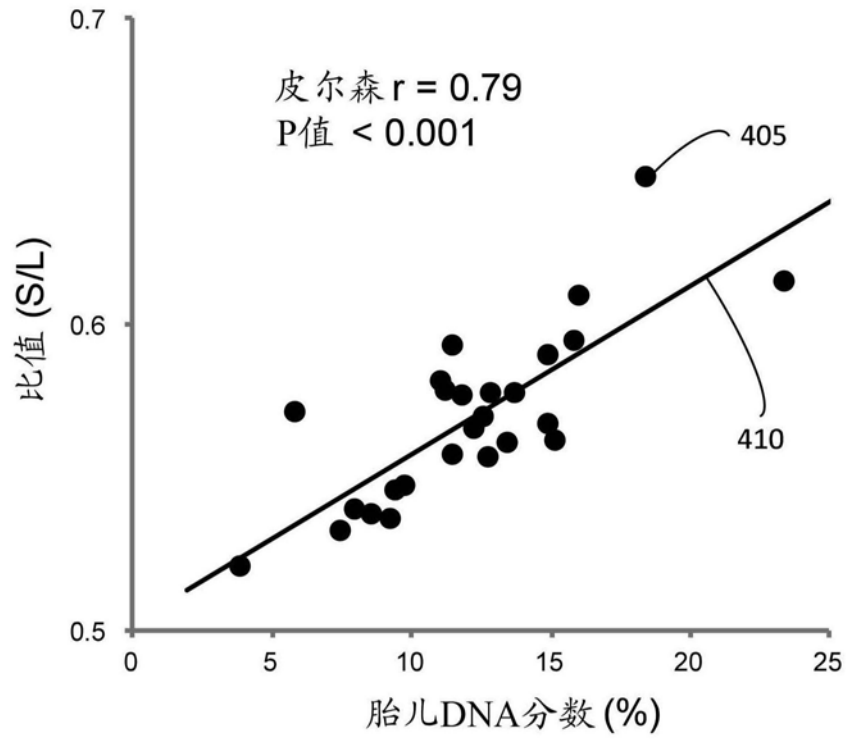


图4A

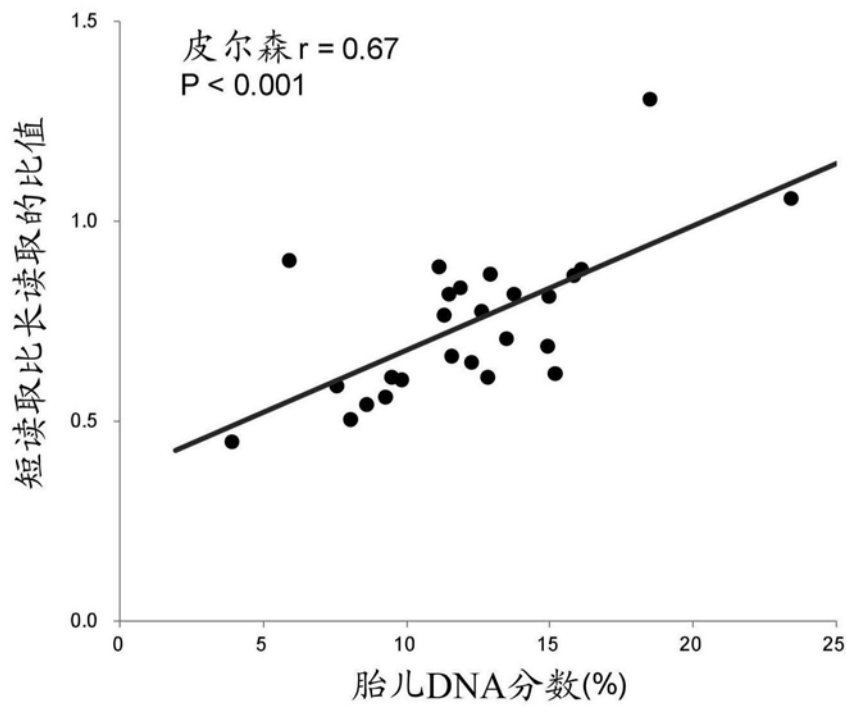


图4B

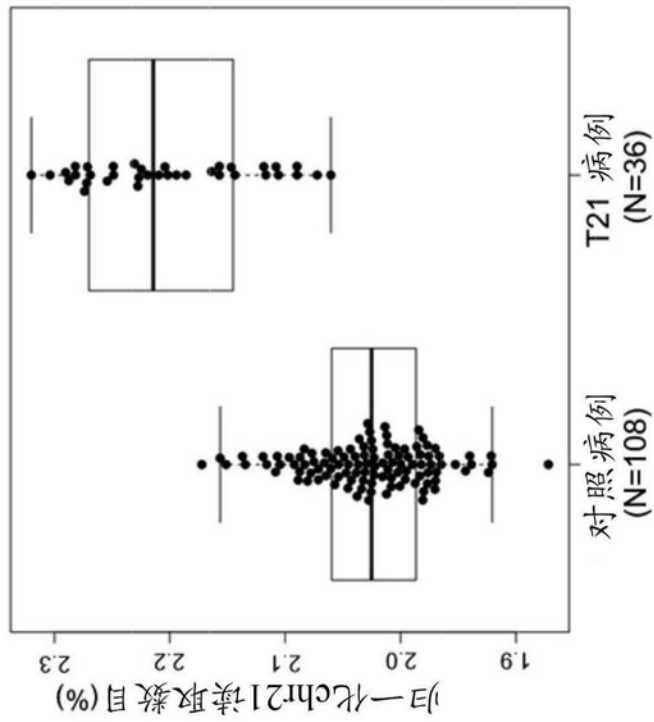


图5A

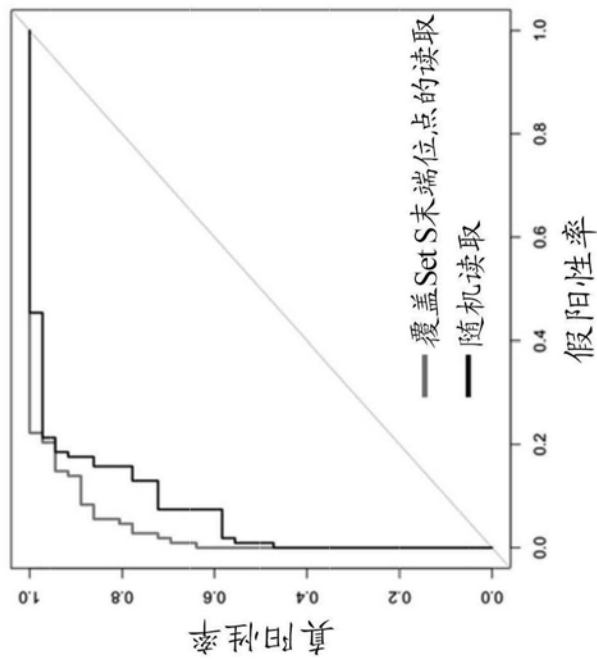


图5B

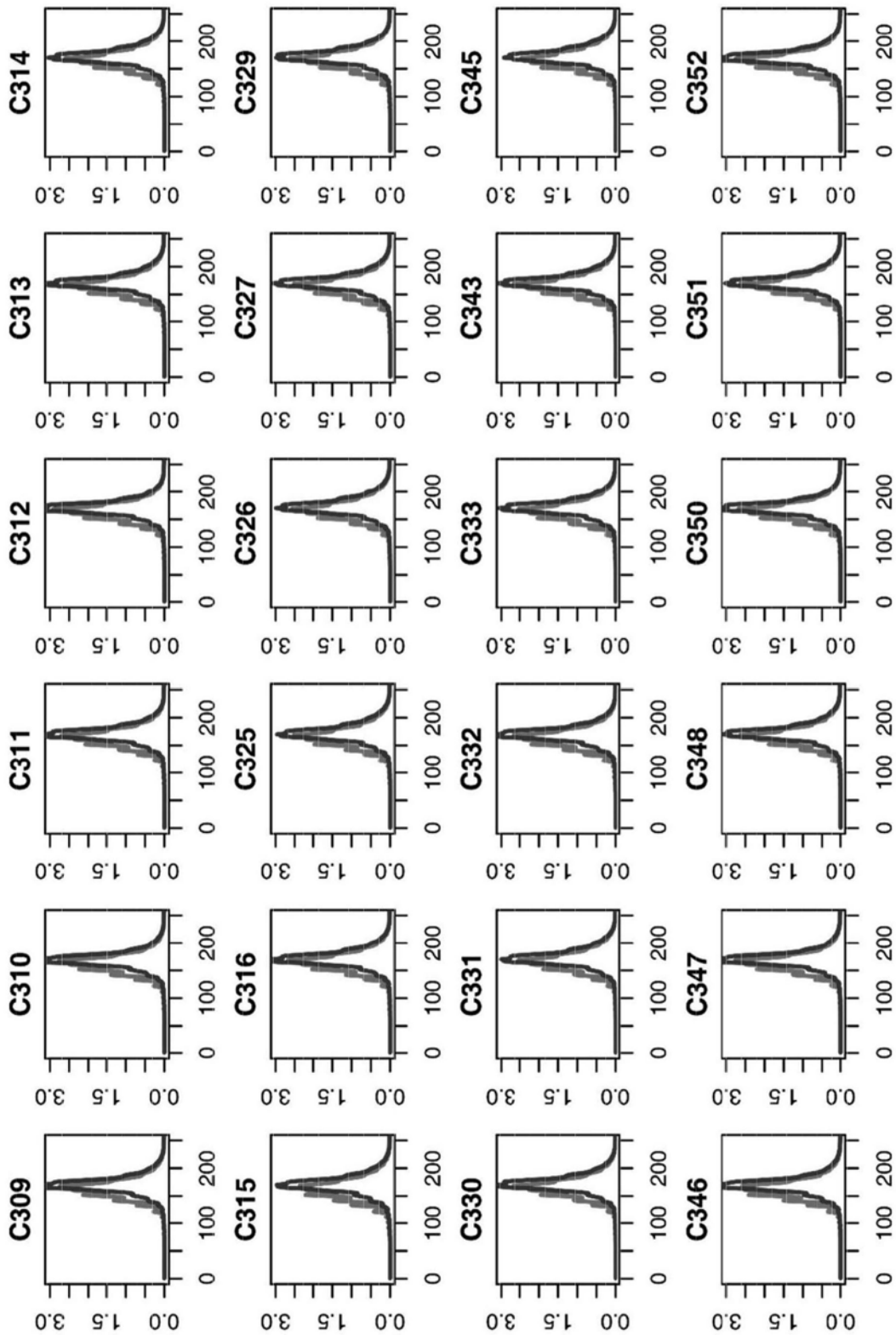


图6

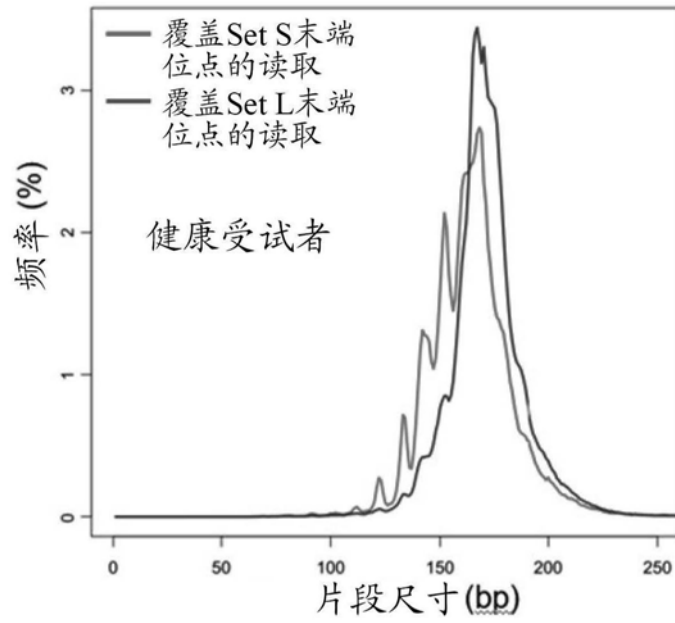


图7A

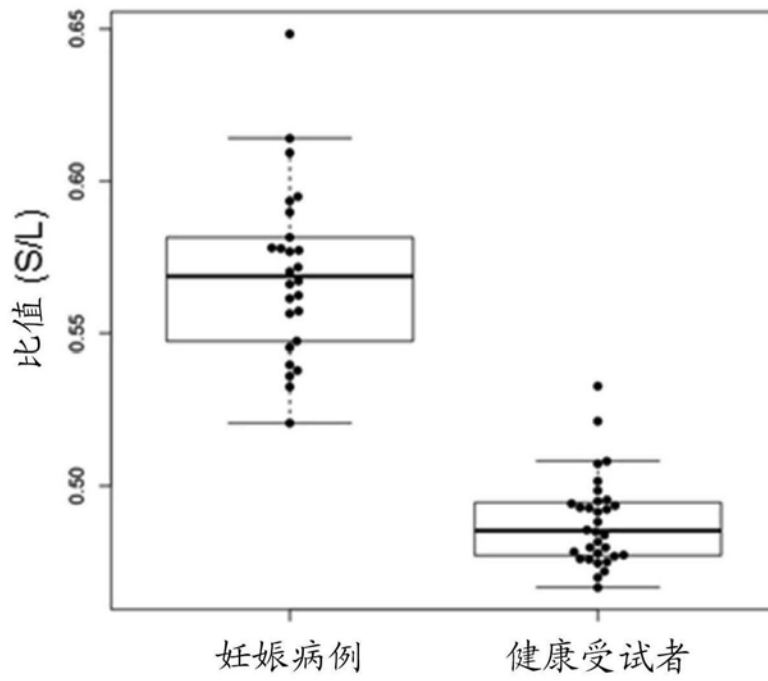


图7B

尺寸分布比较

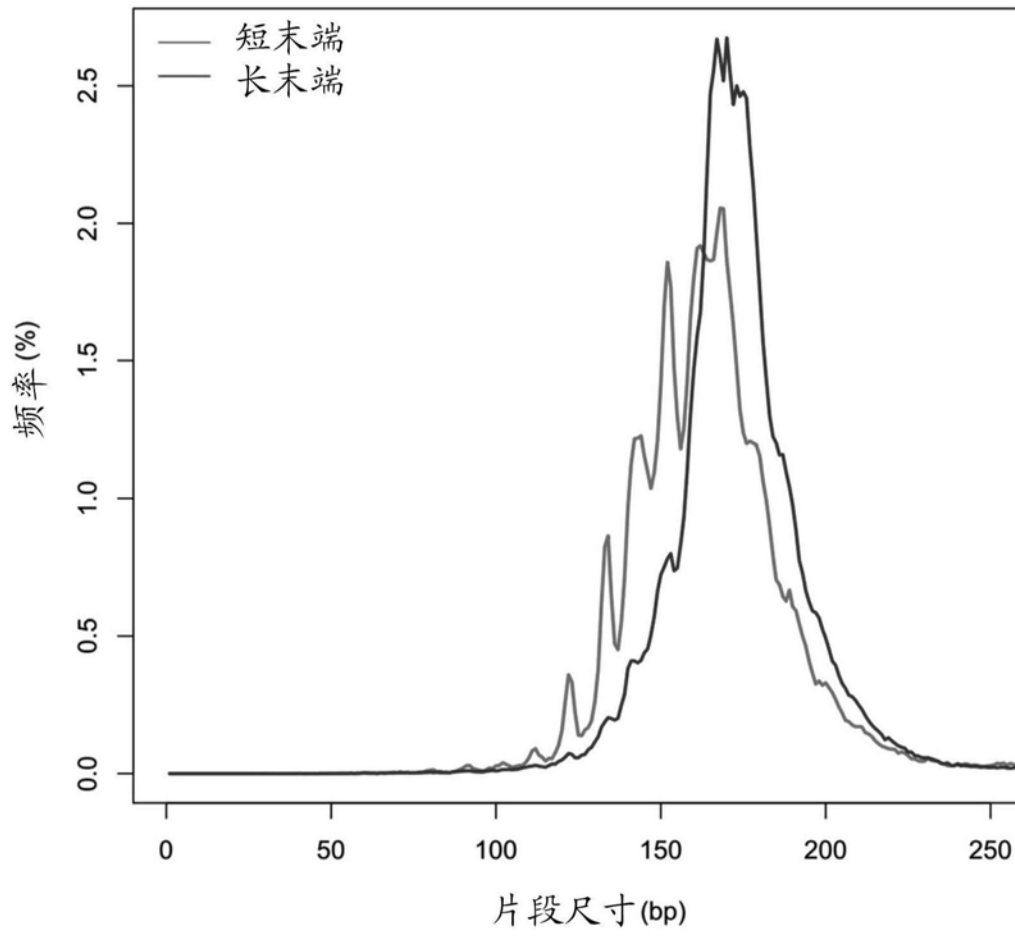


图8

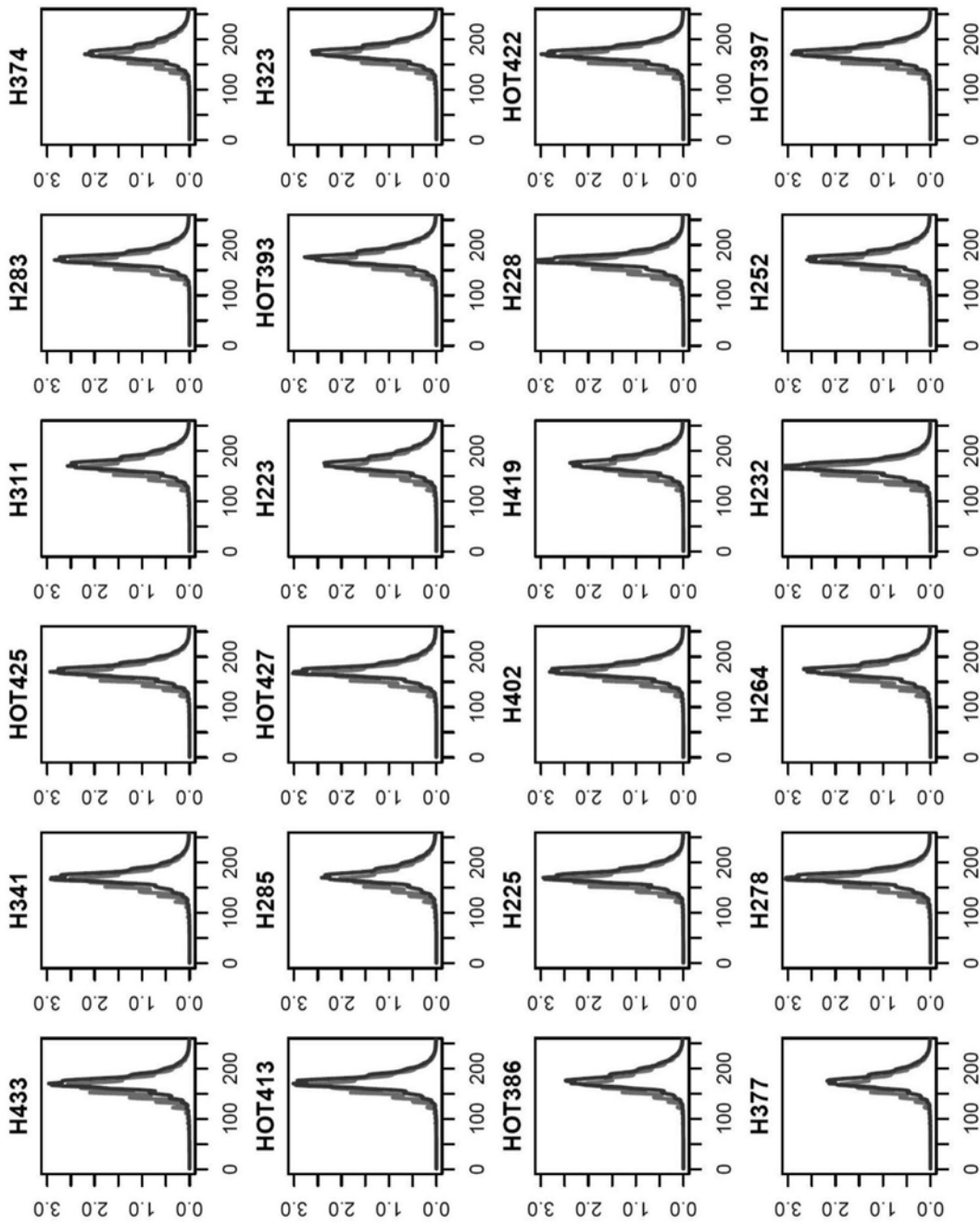


图9

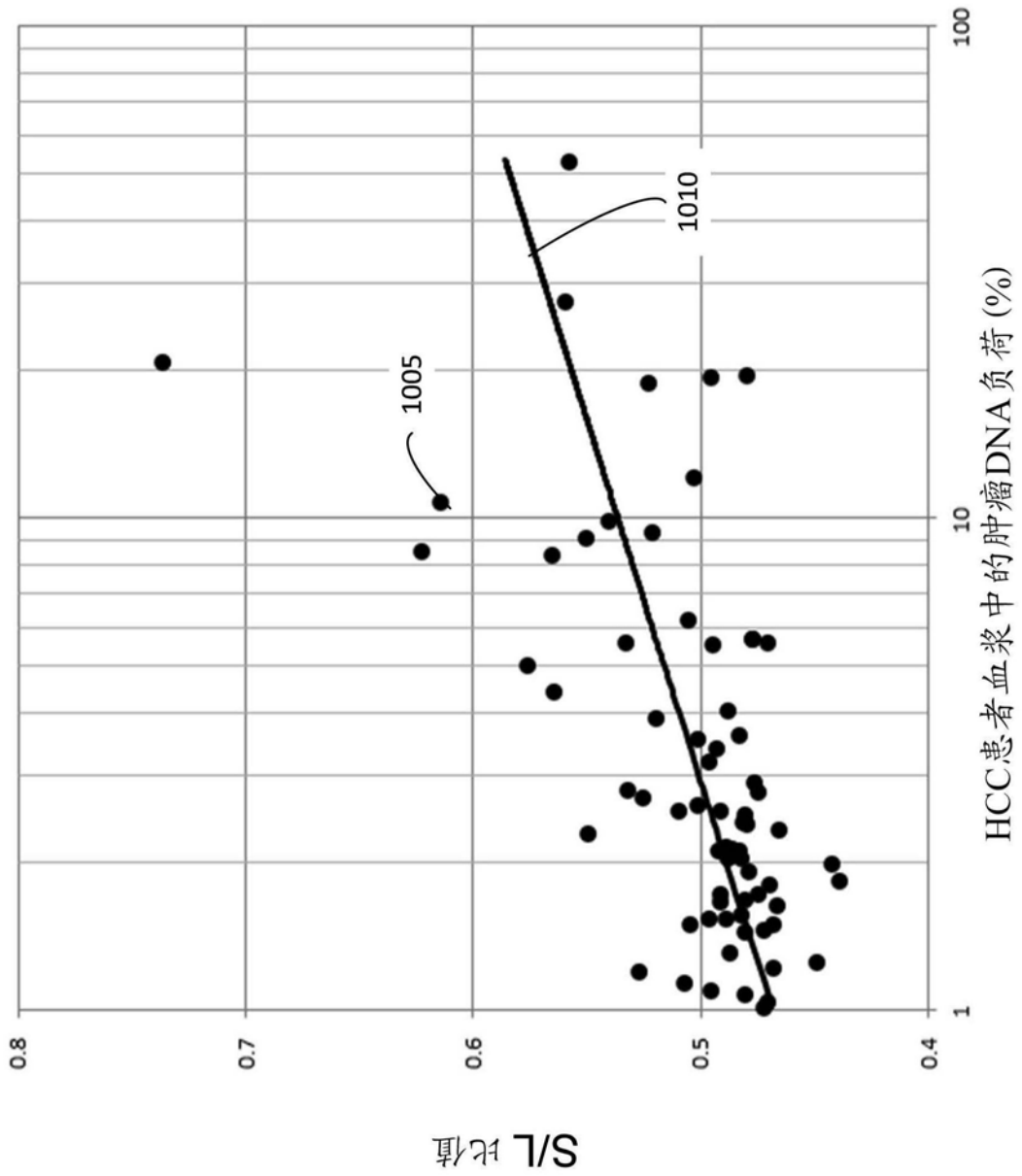


图10

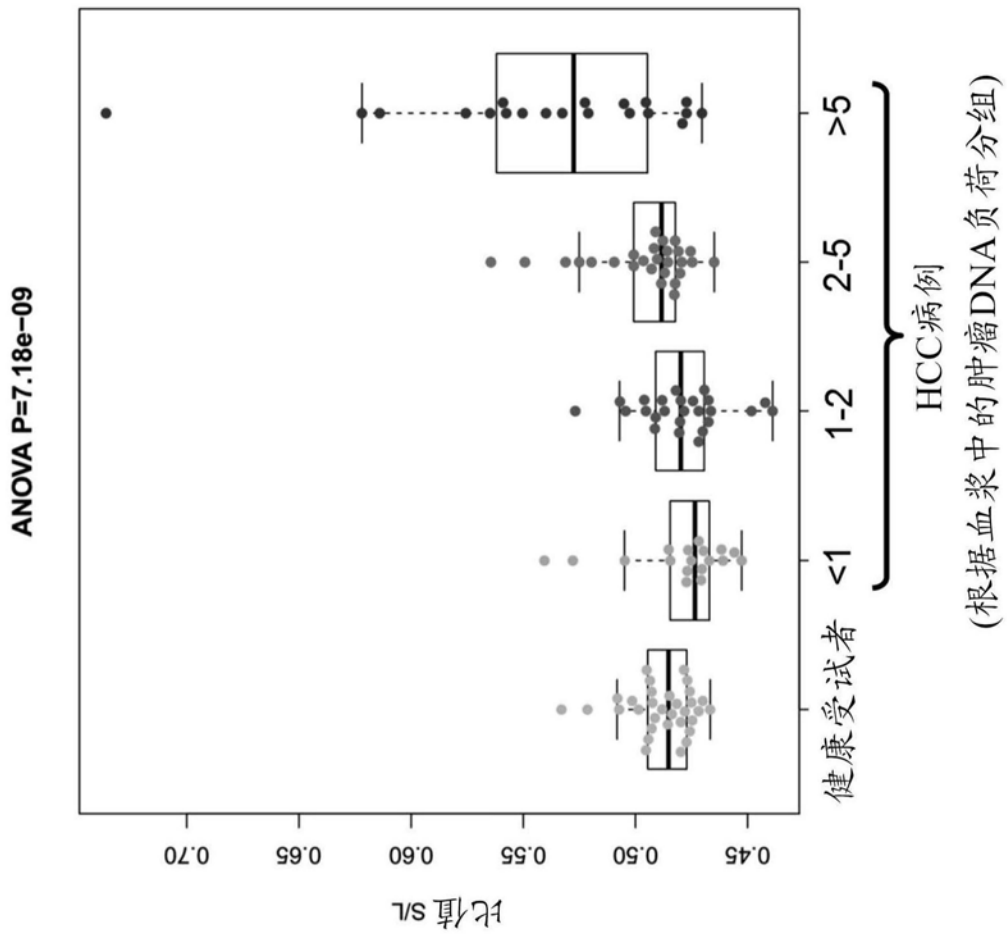


图11

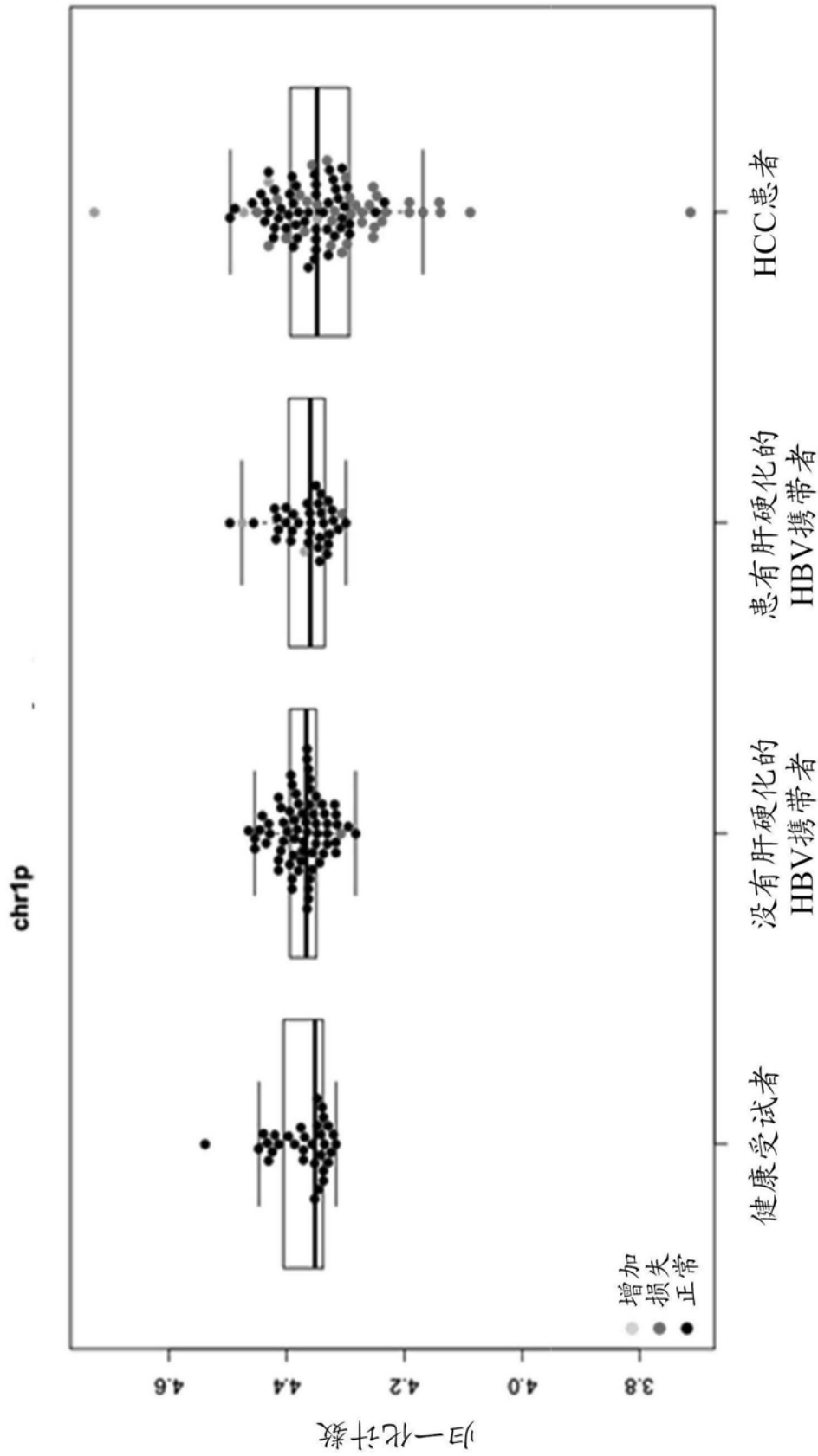


图12

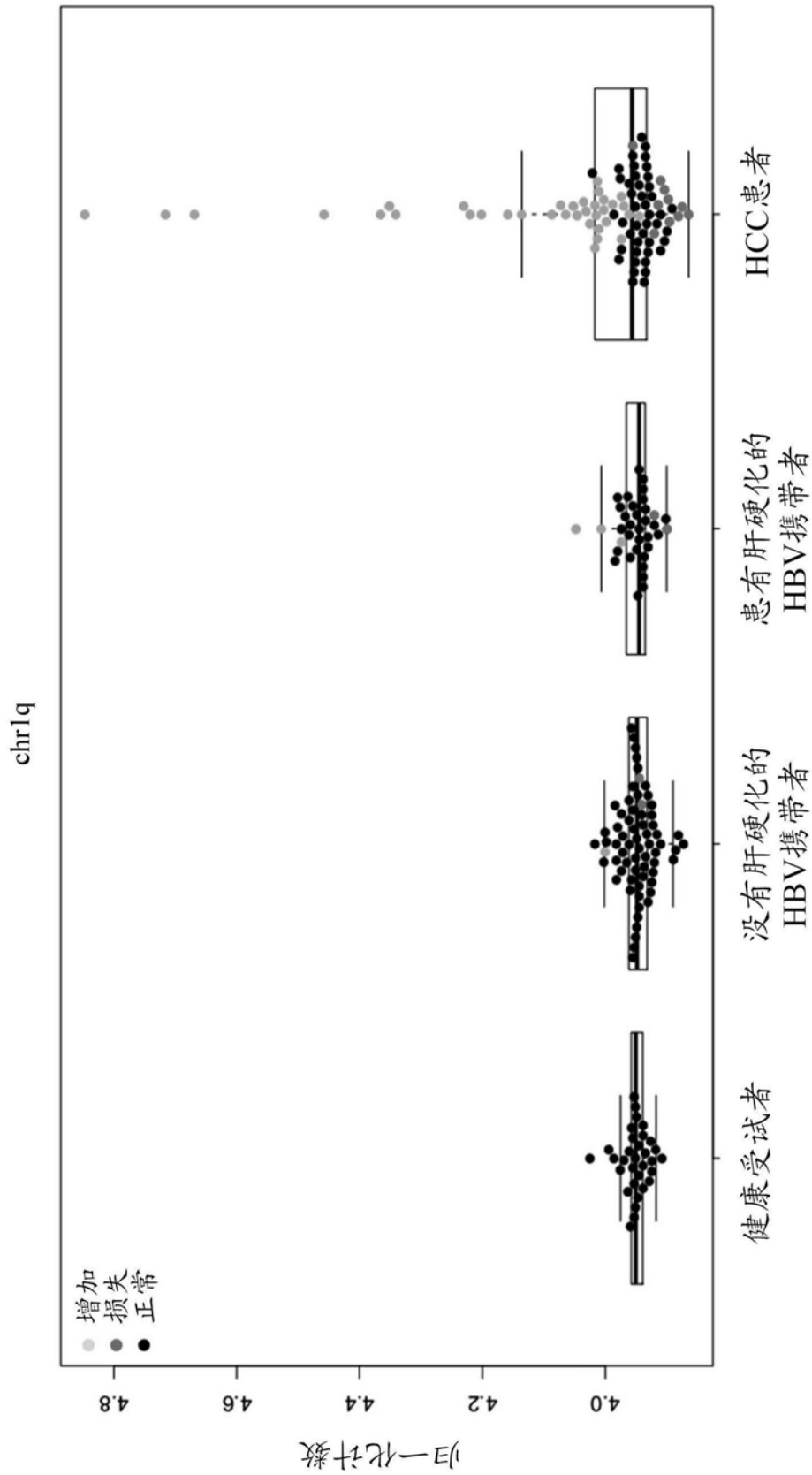


图13

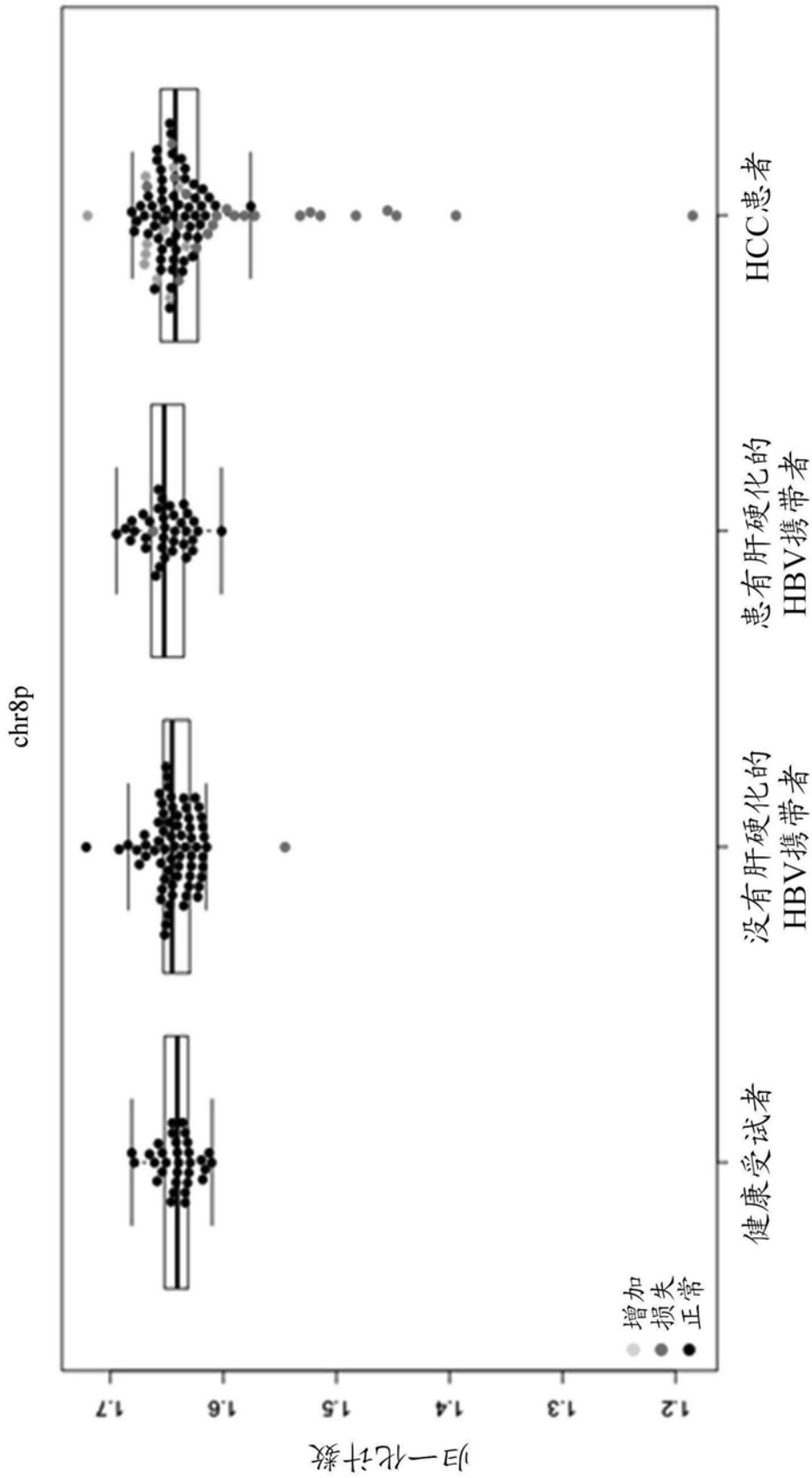


图14

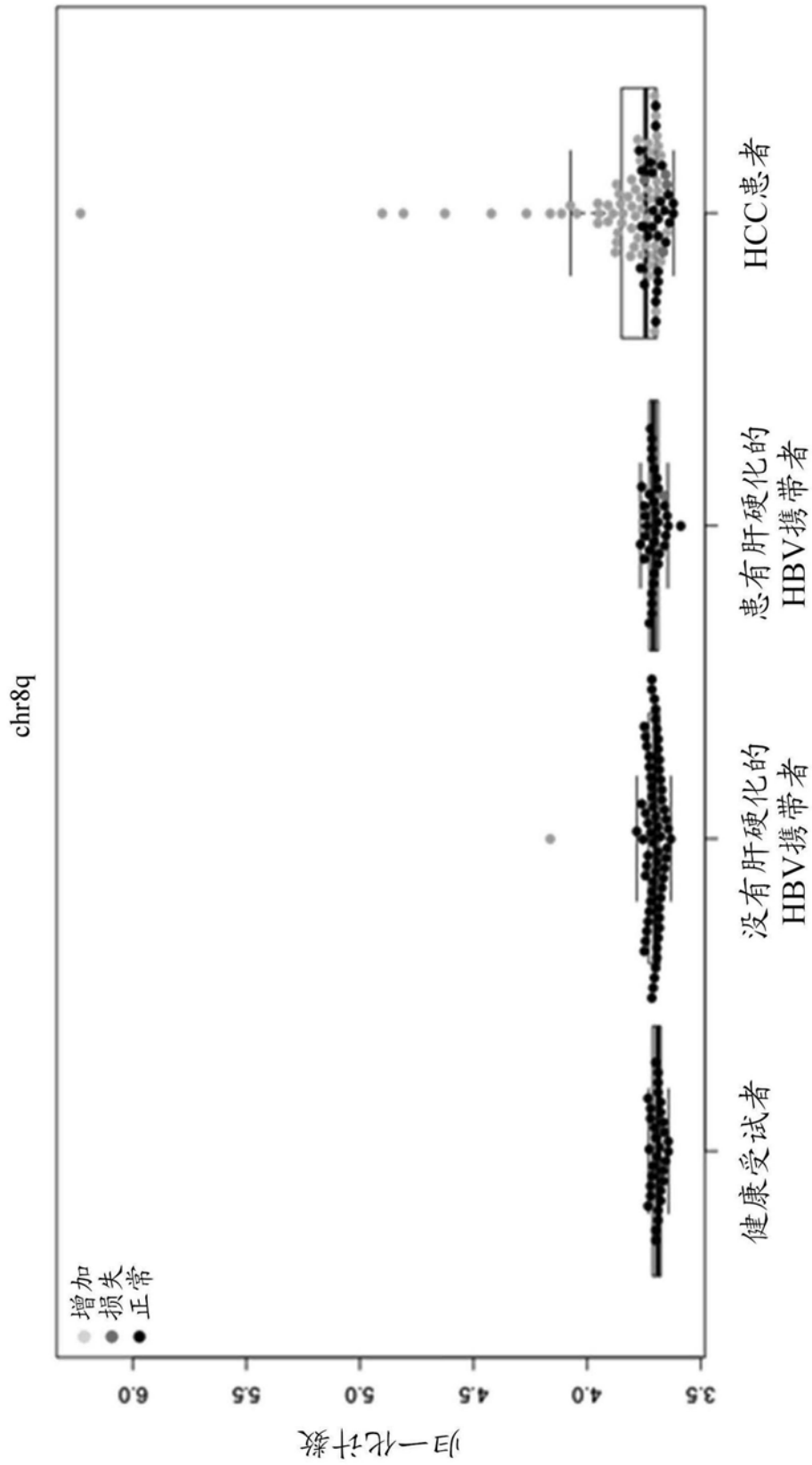


图15

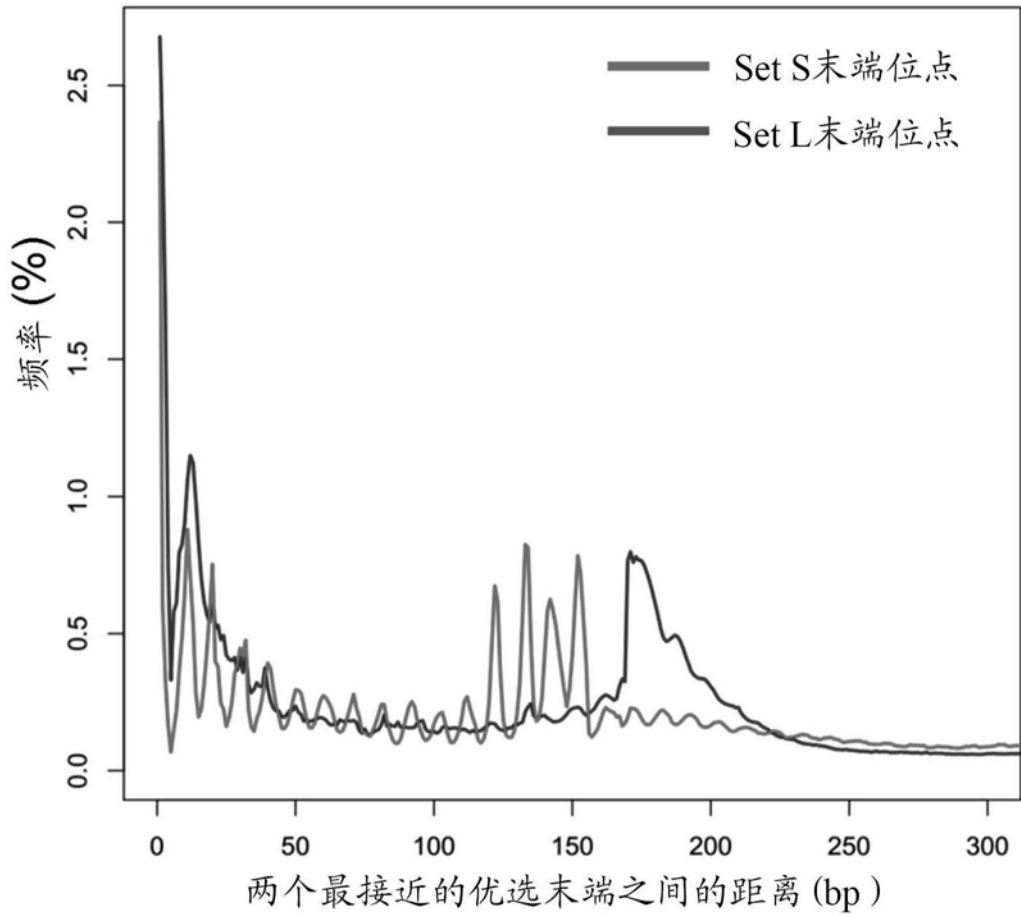


图16

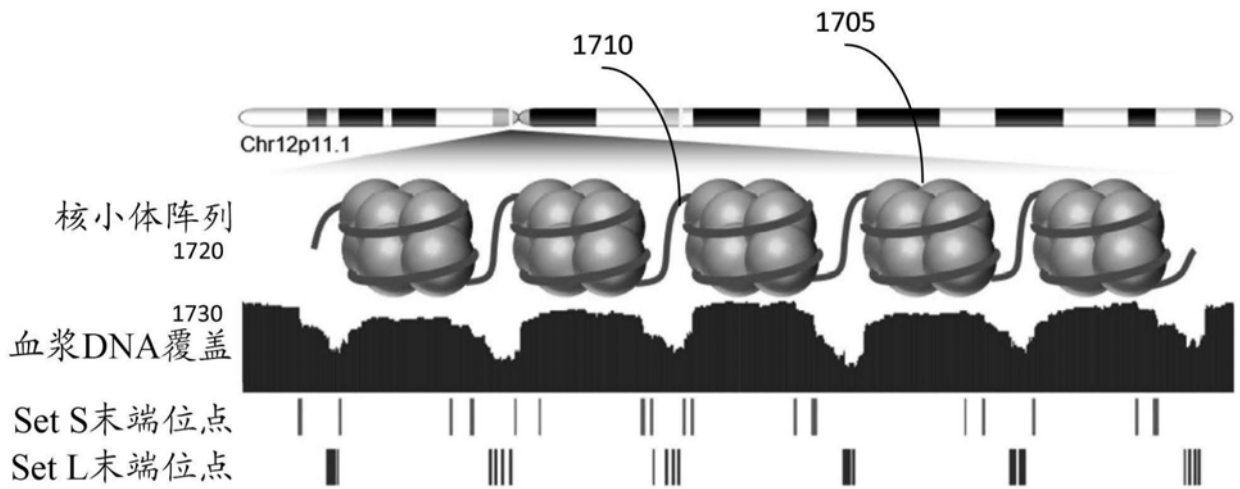


图17A

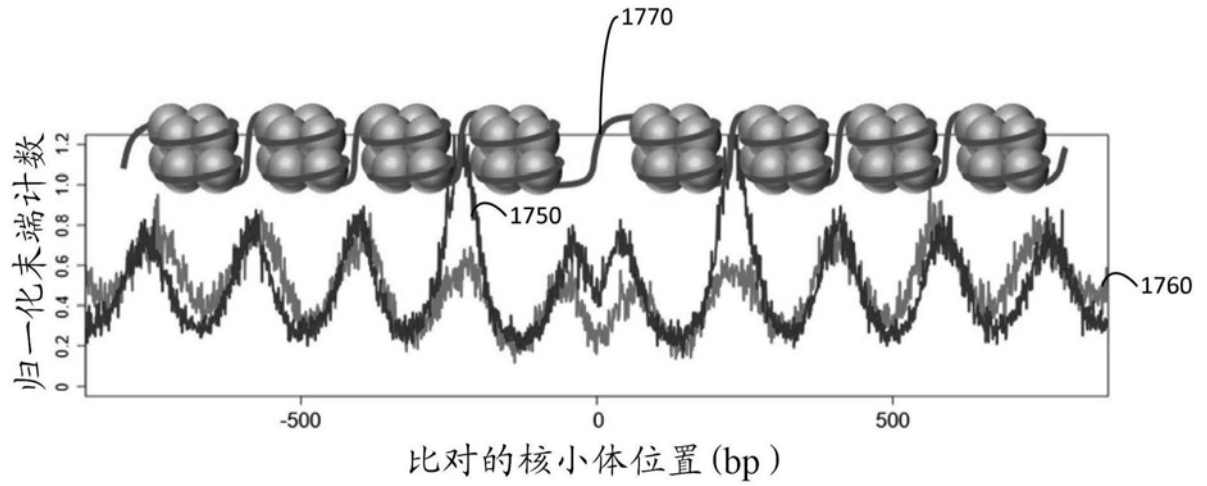


图17B

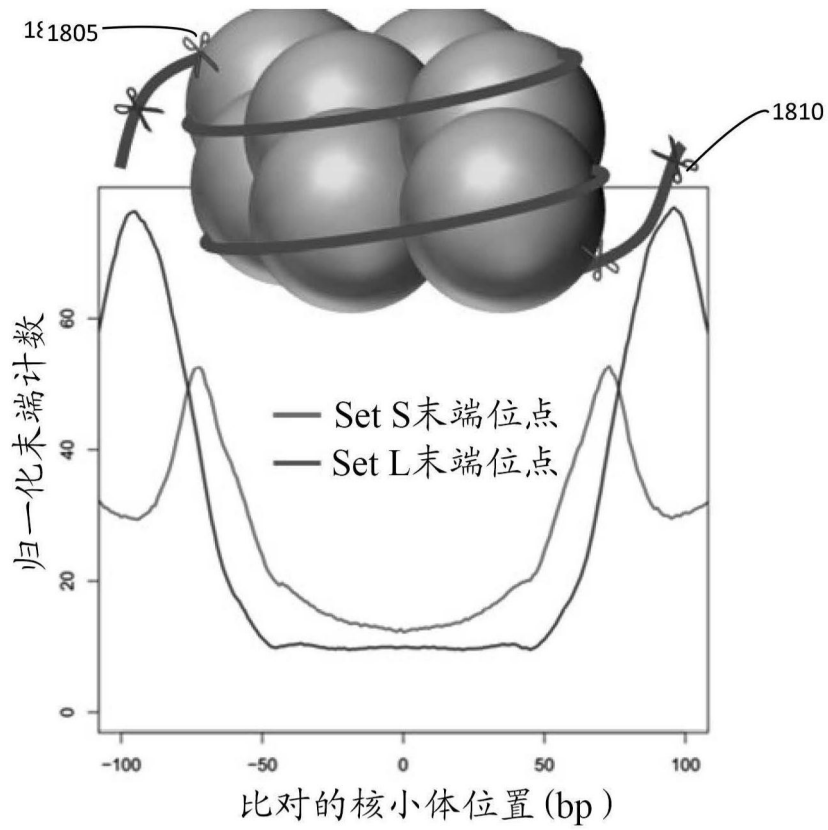


图18A

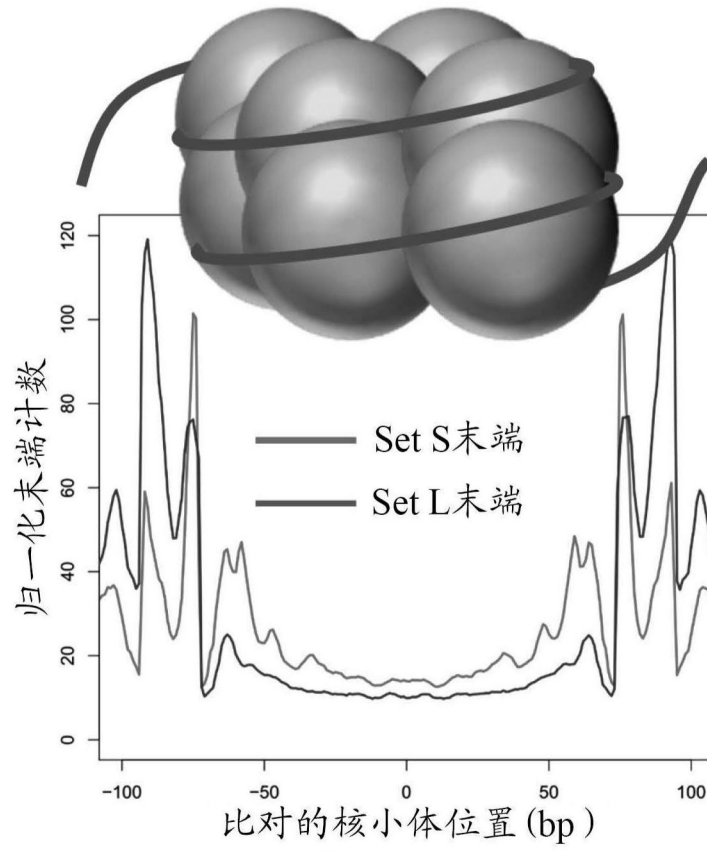


图18B

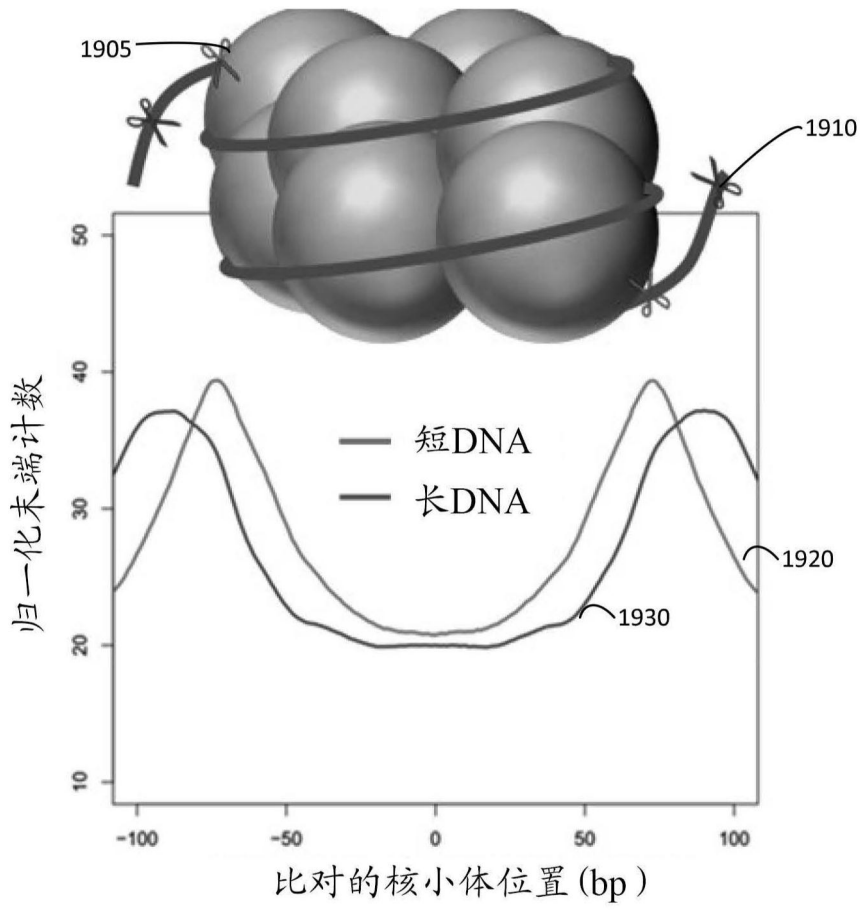


图19

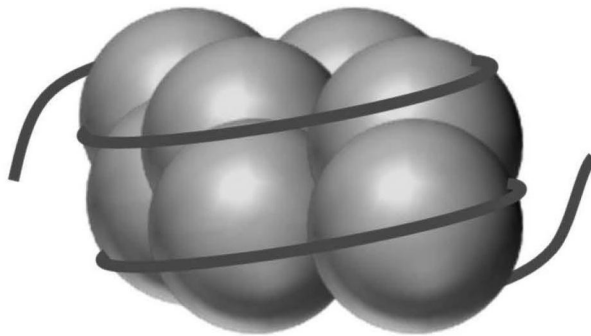


图20A

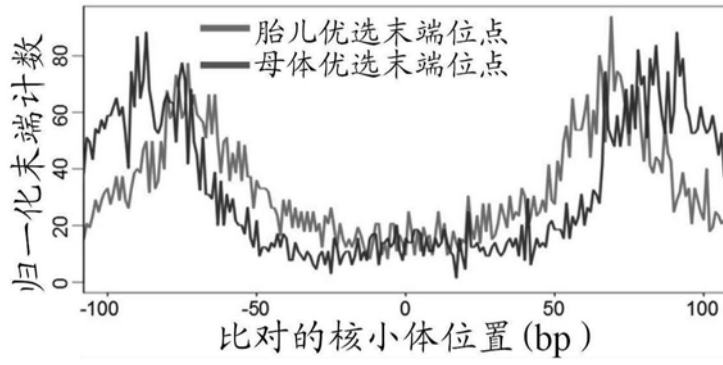


图20B

图 20C

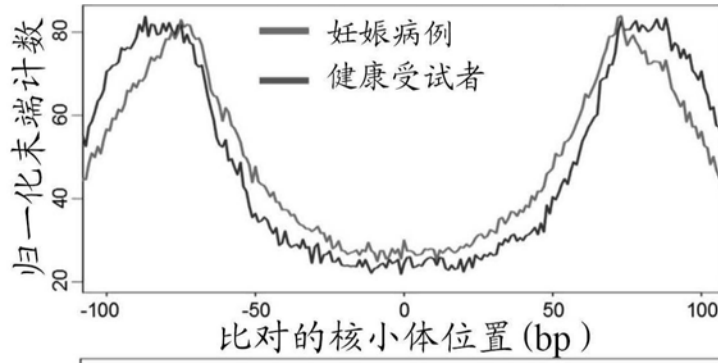
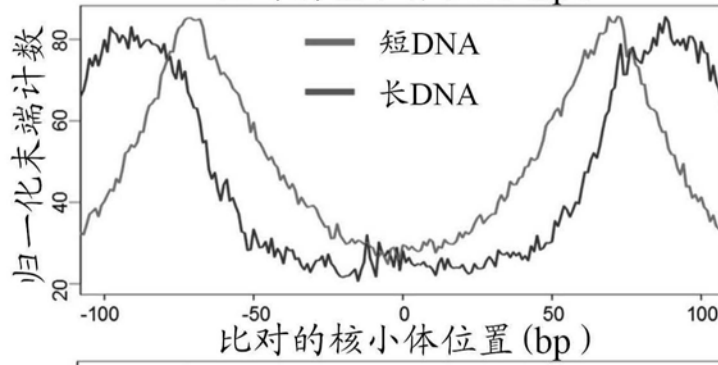
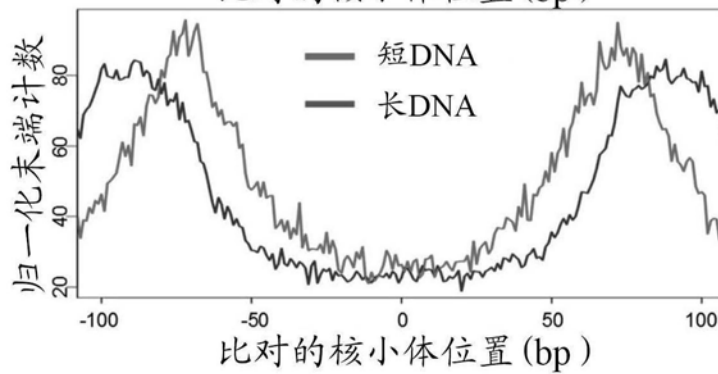


图 20D



妊娠

图 20E



非妊娠

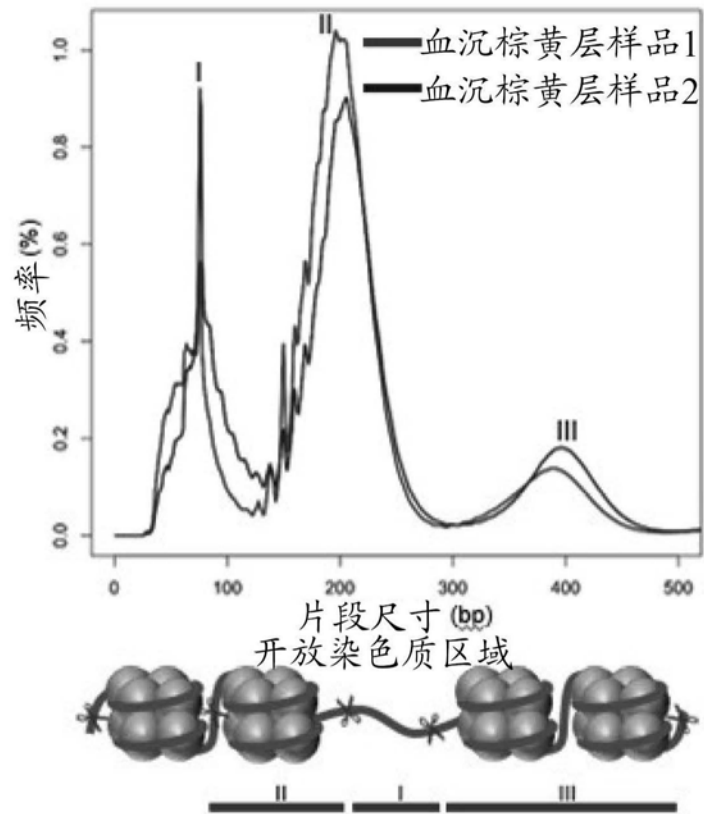


图21A

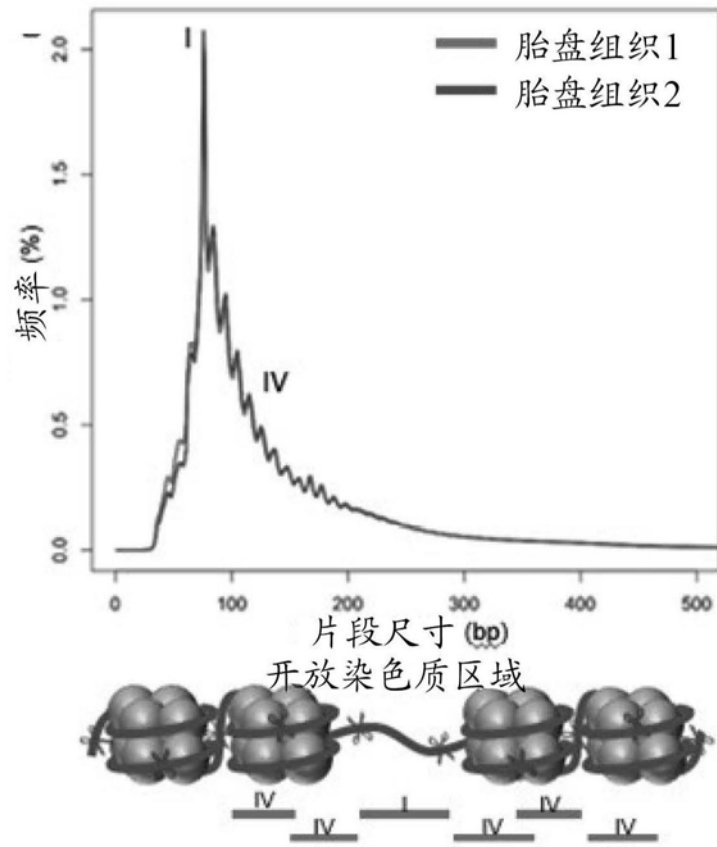


图21B

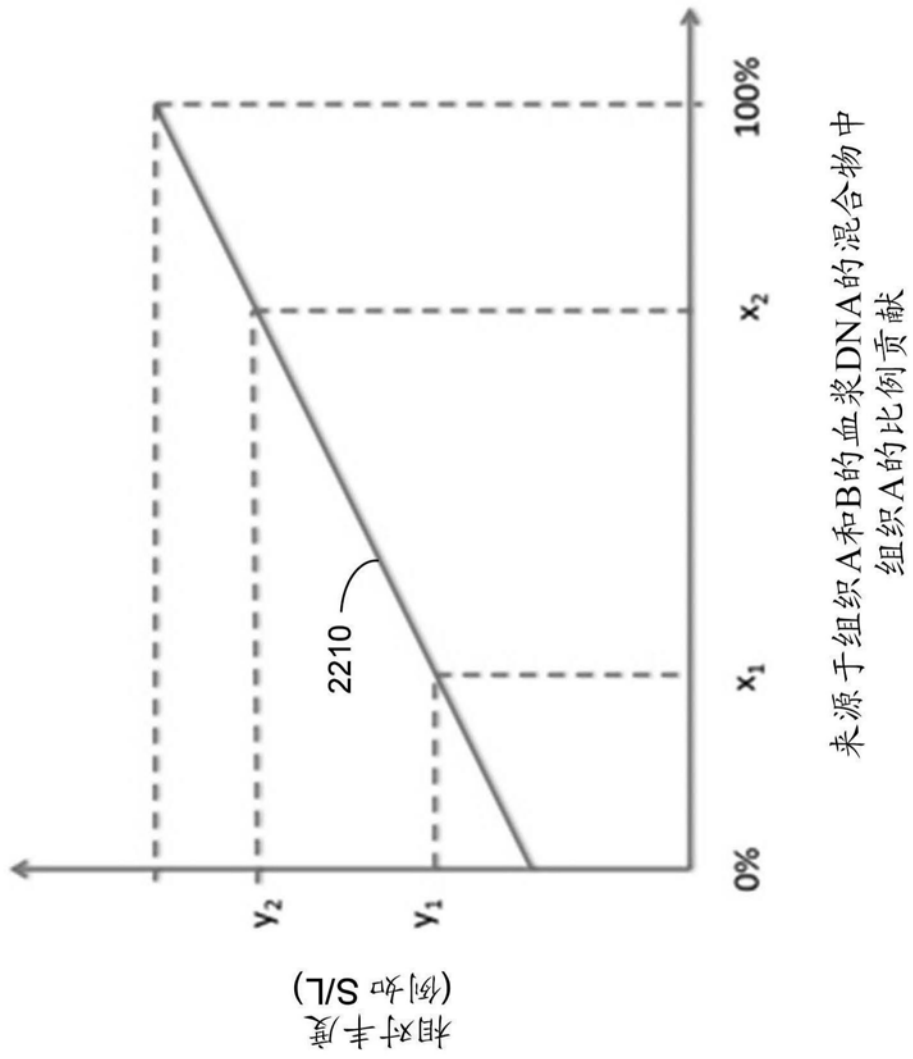


图22

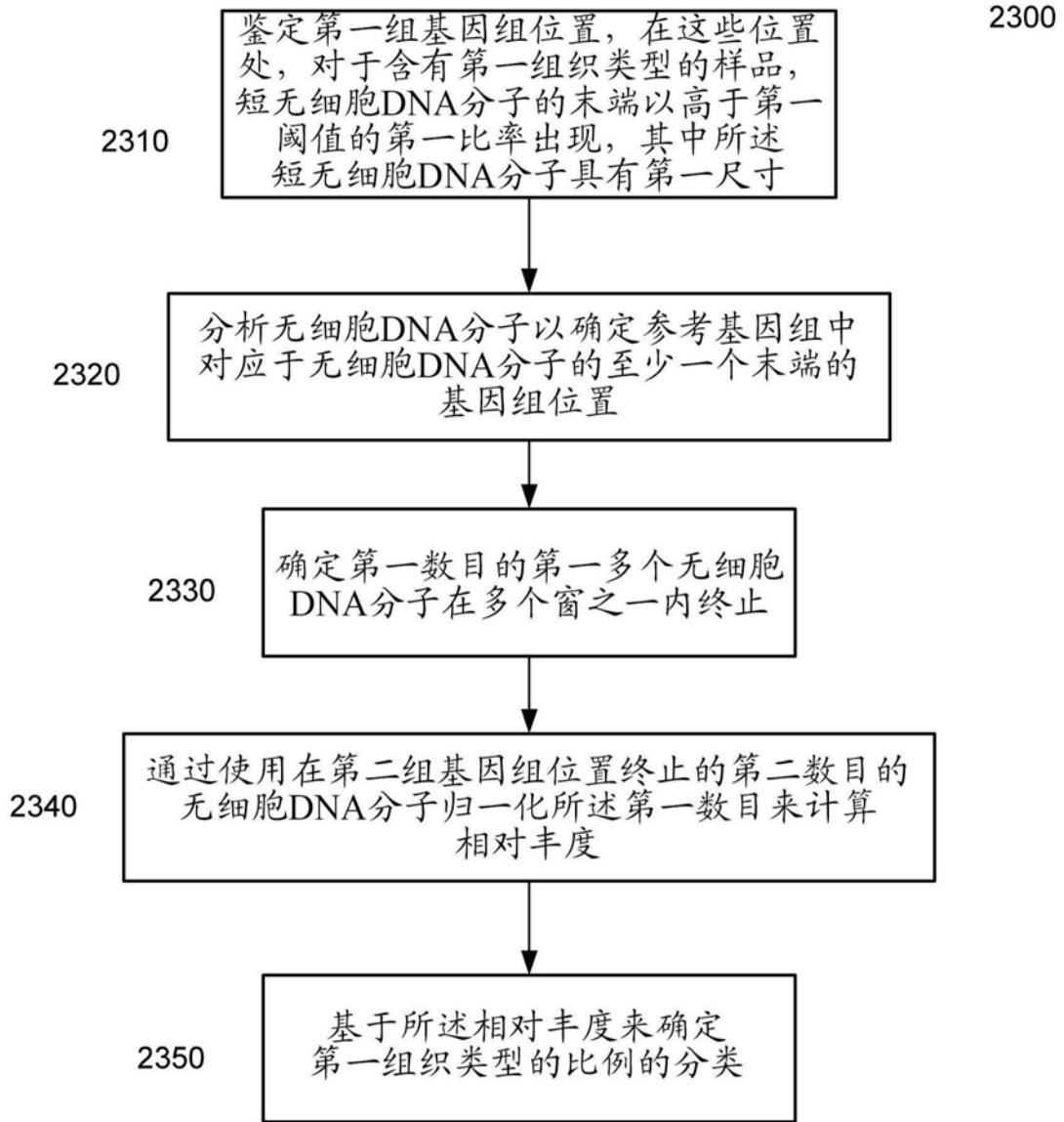


图23

2400

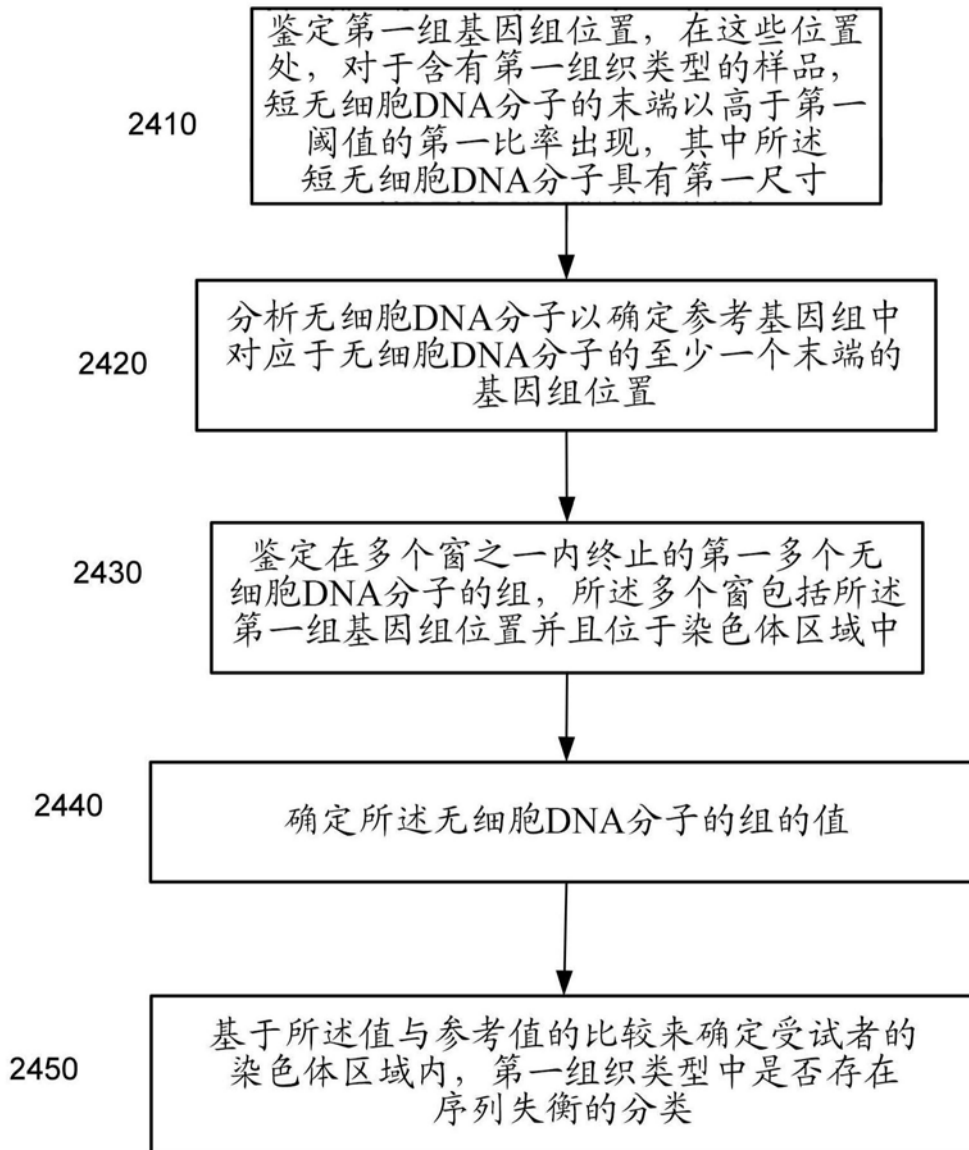
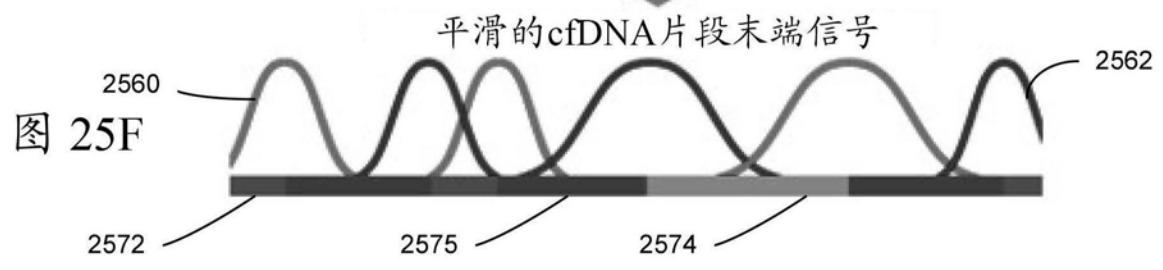
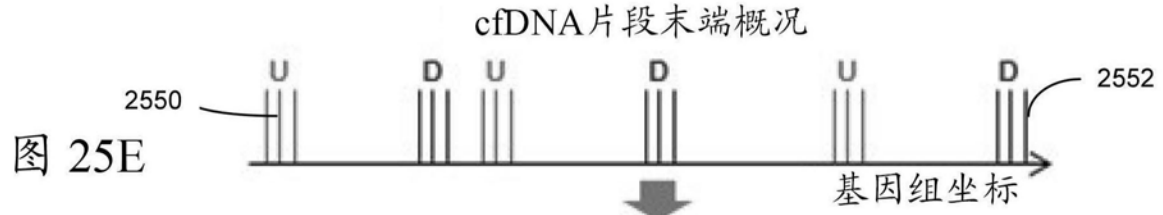
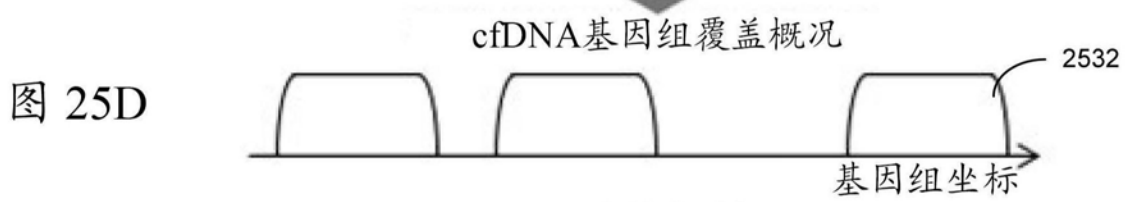
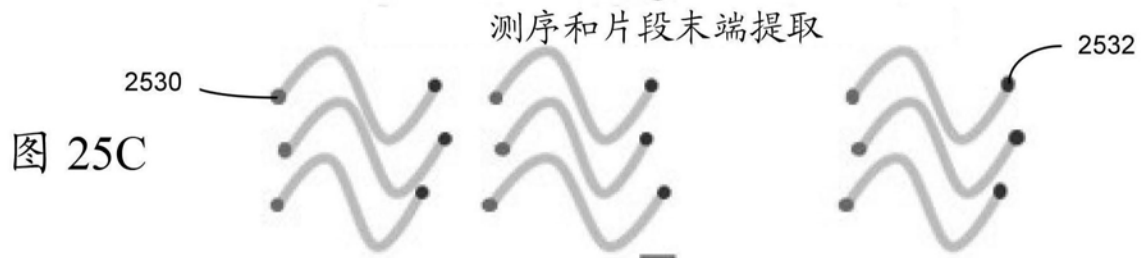
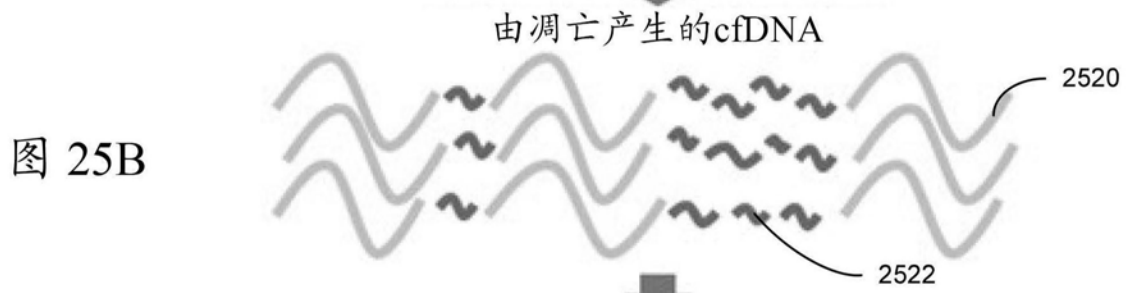
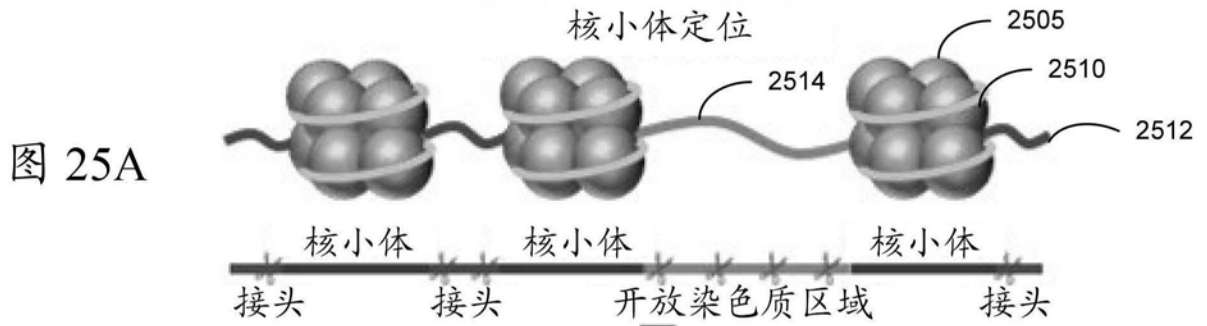


图24



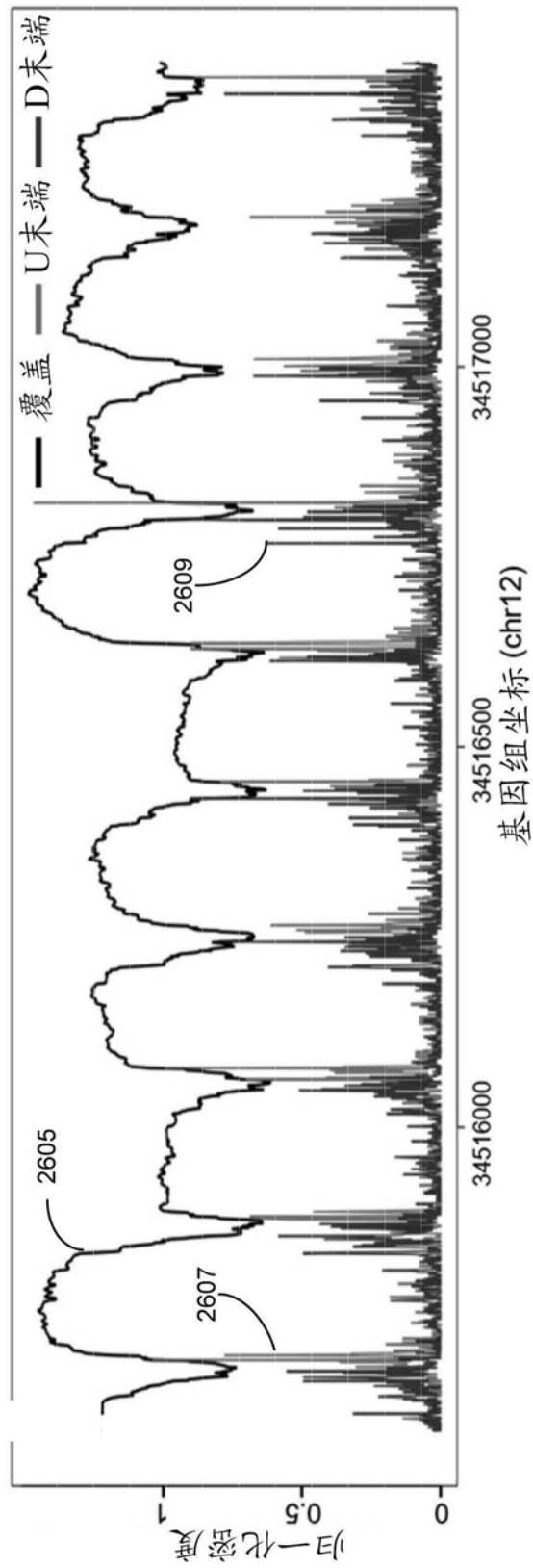


图26A

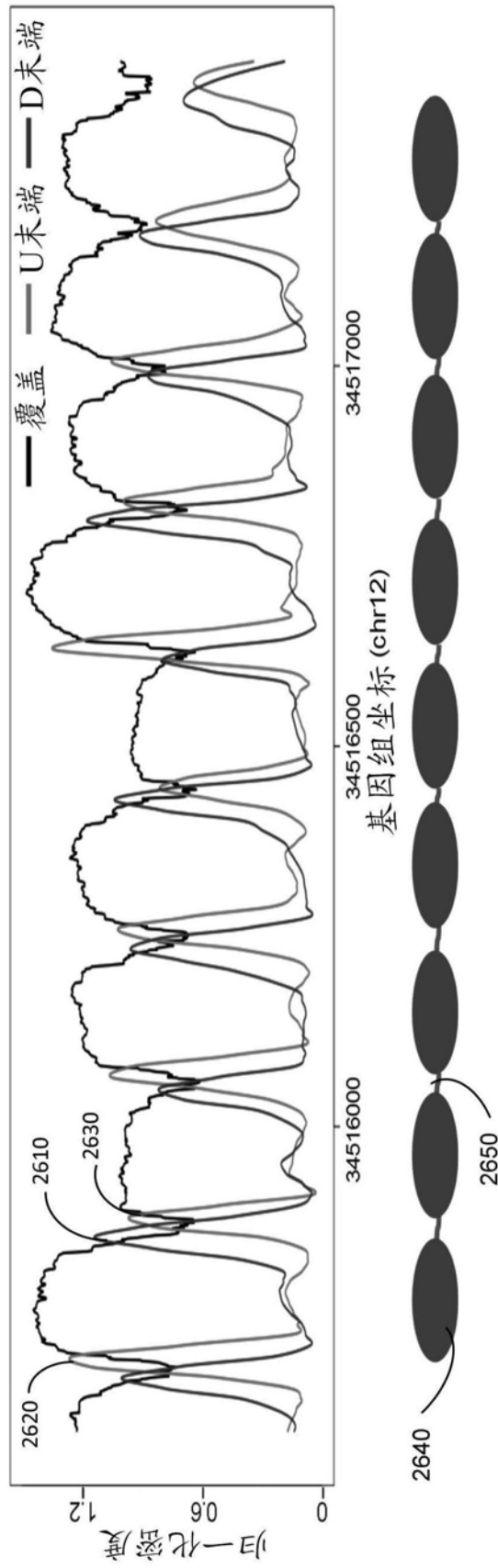


图26B

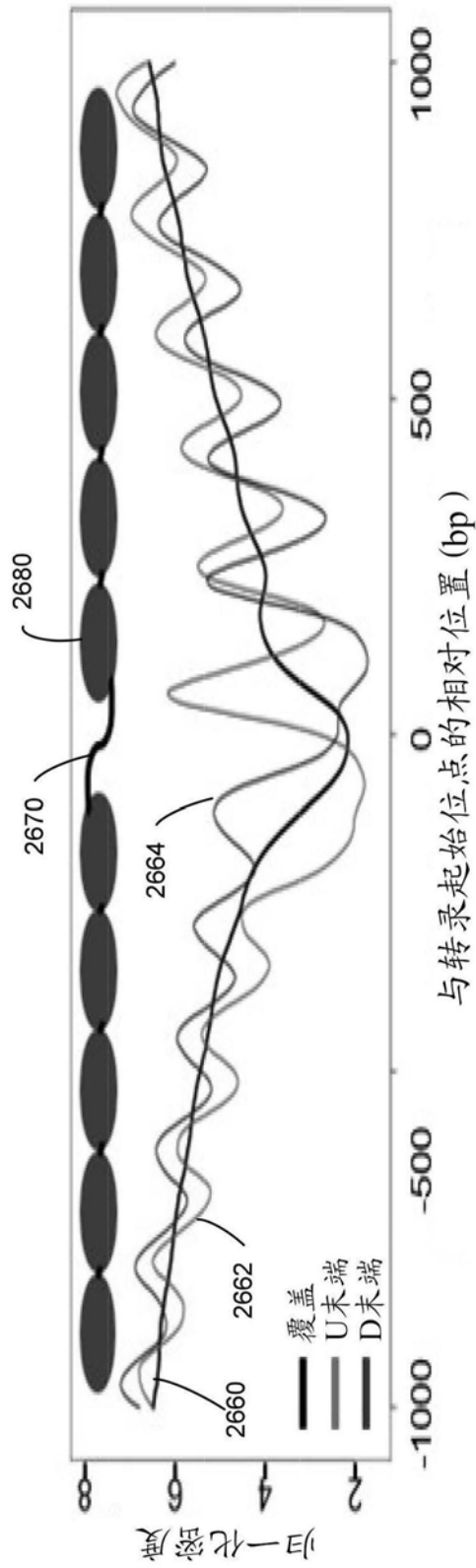


图26C

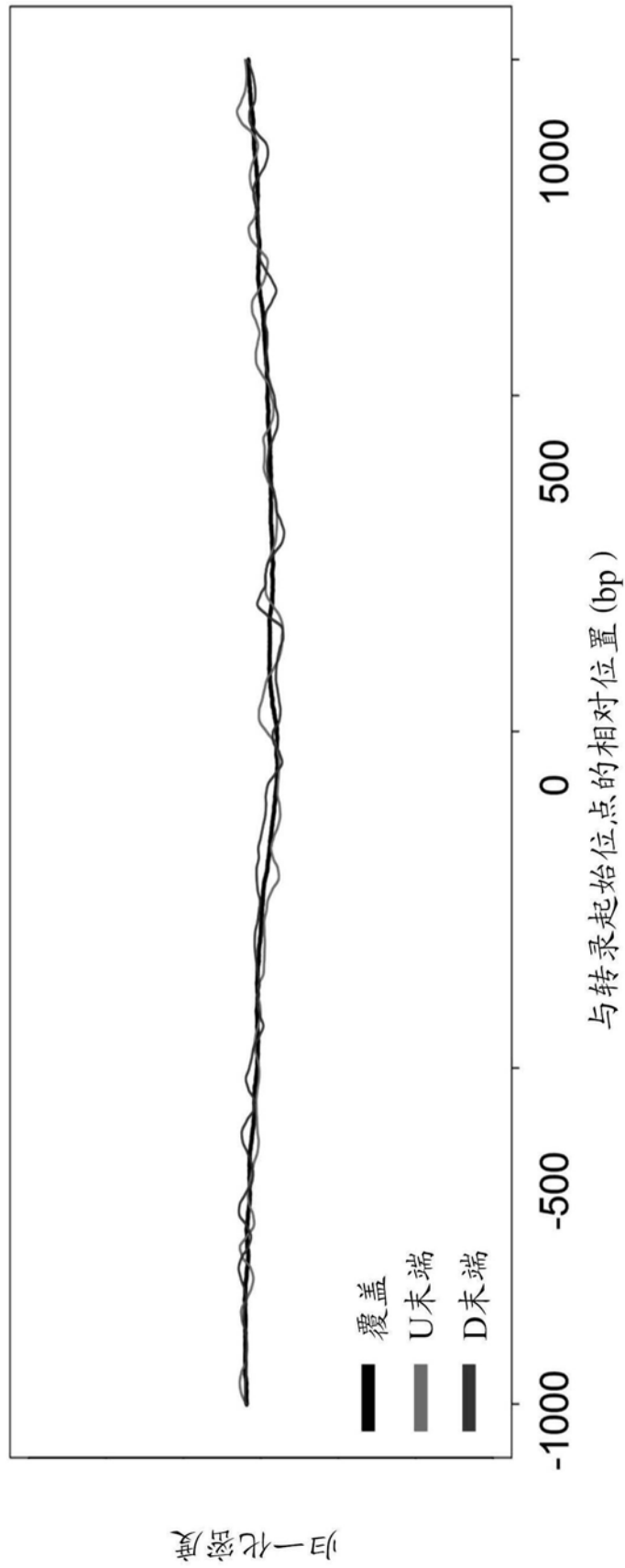


图26D

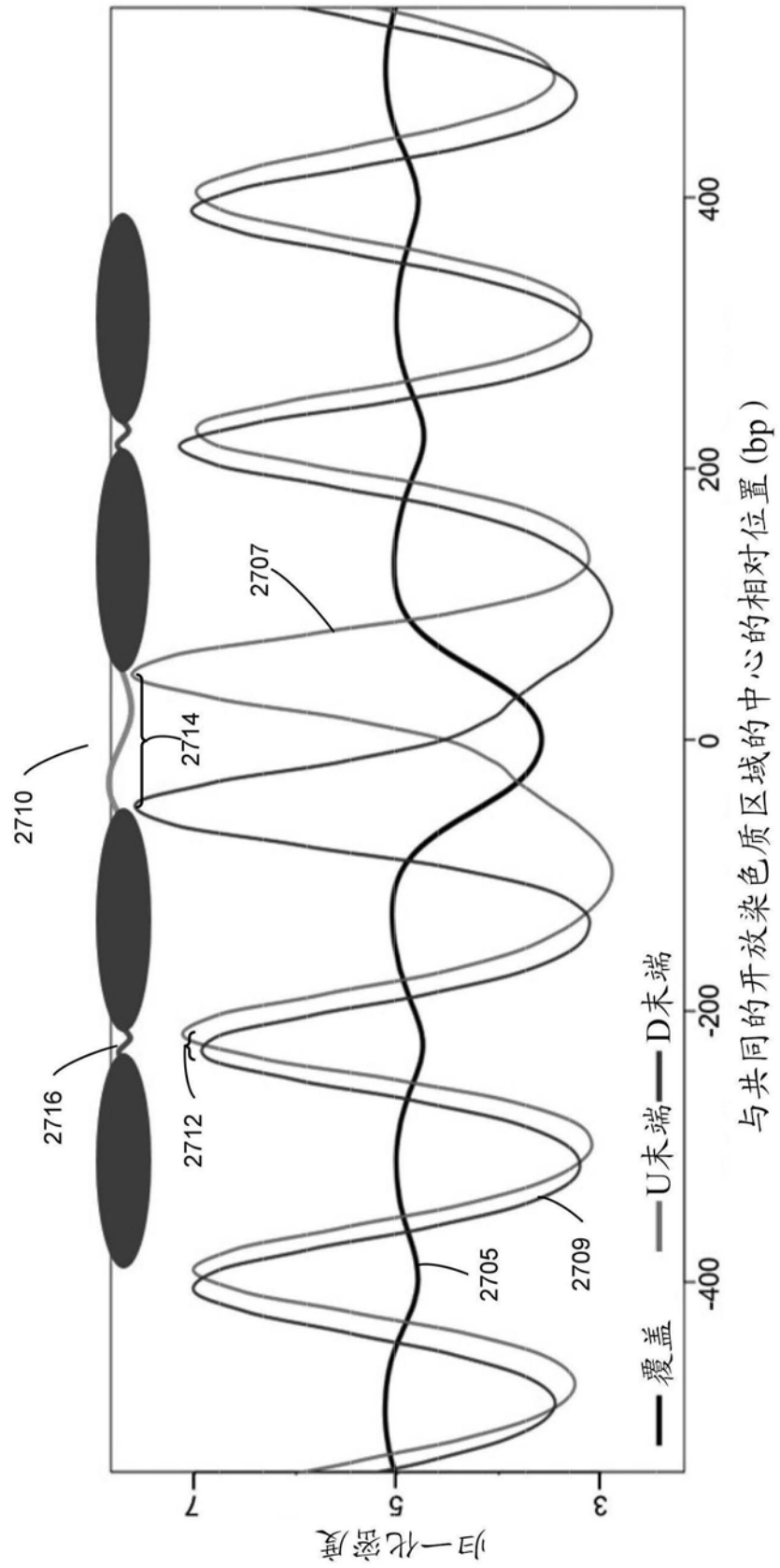


图27A

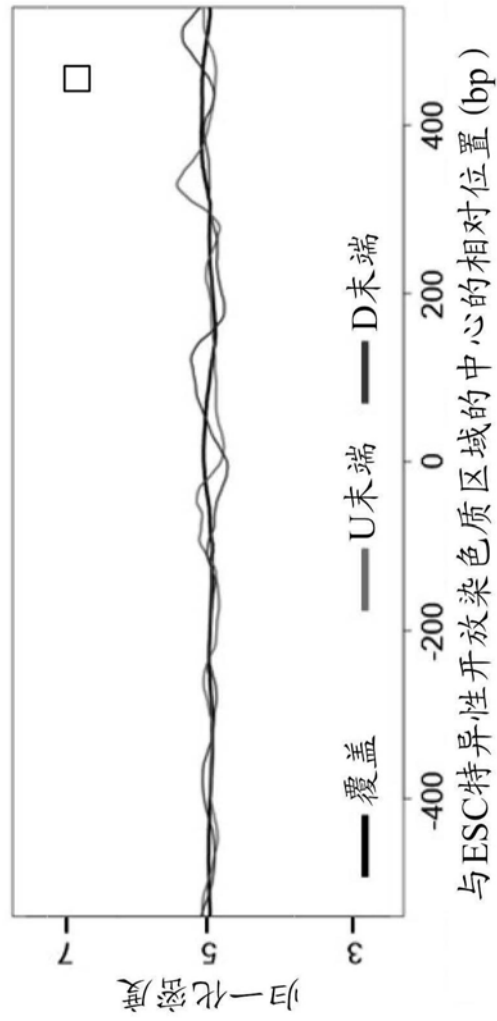


图27B

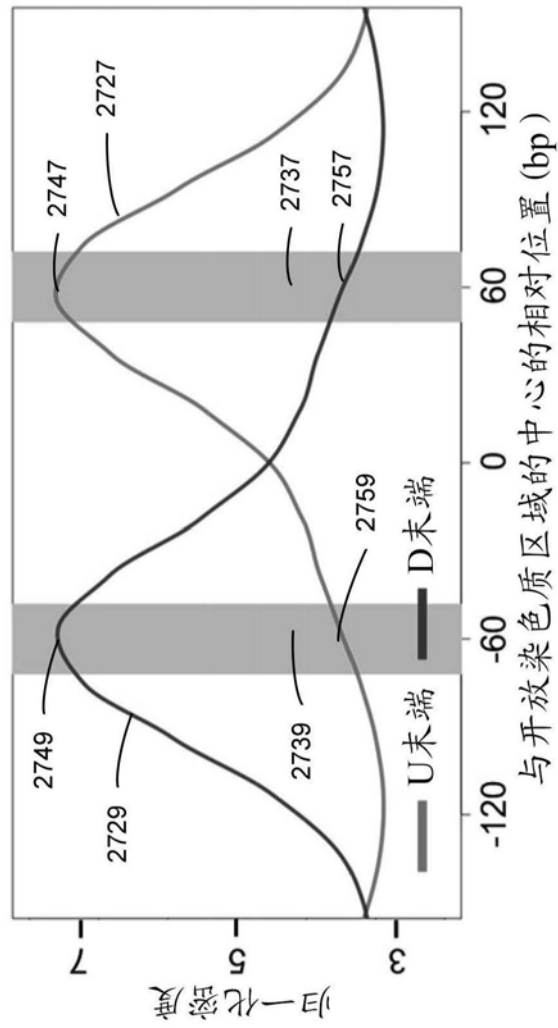


图27C

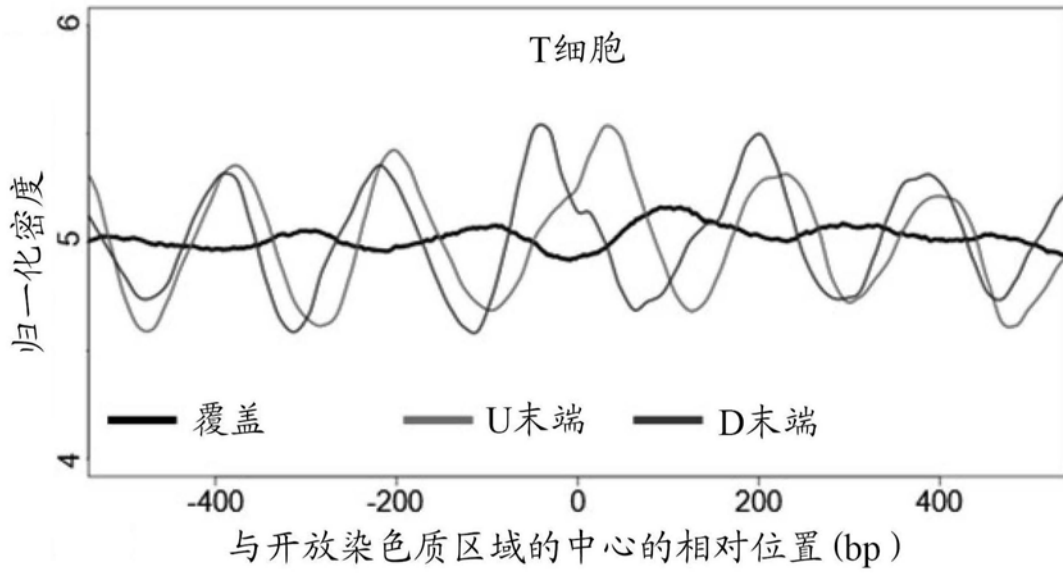


图28A

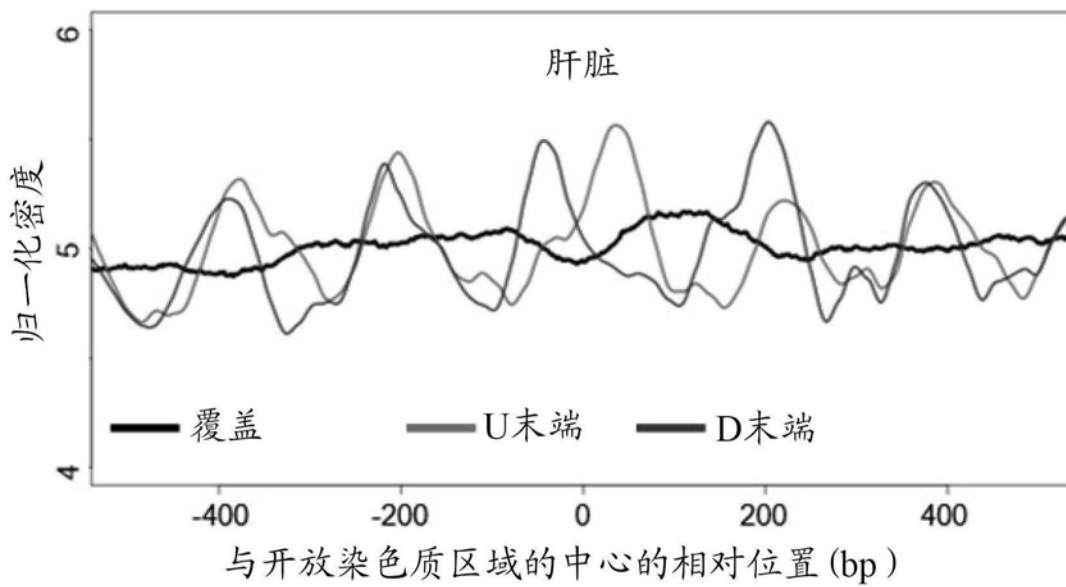


图28B

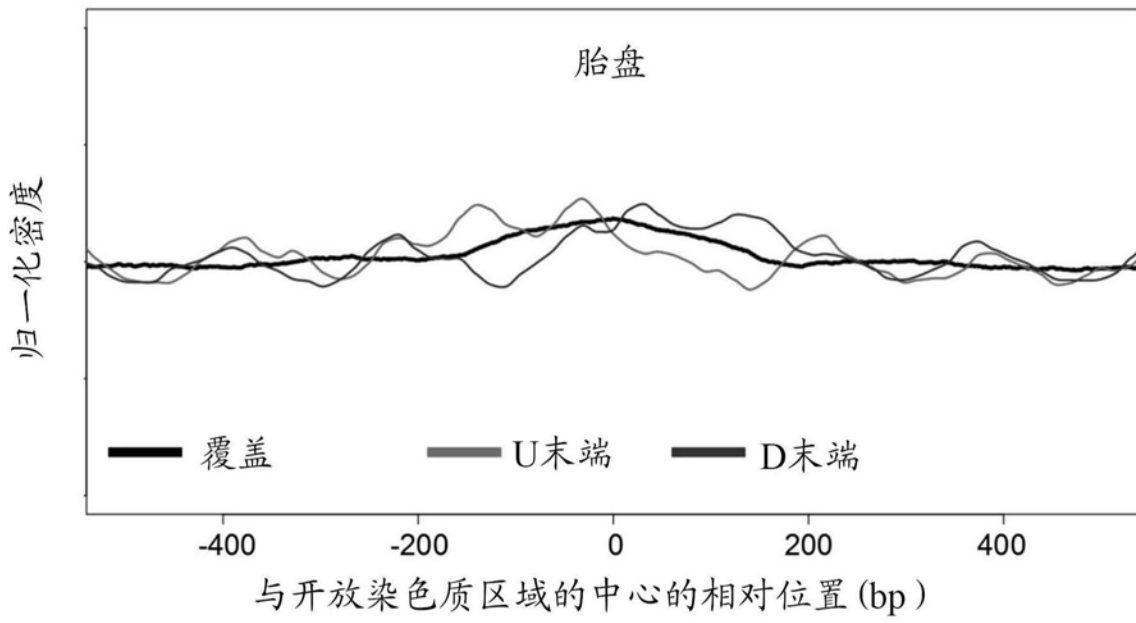


图28C

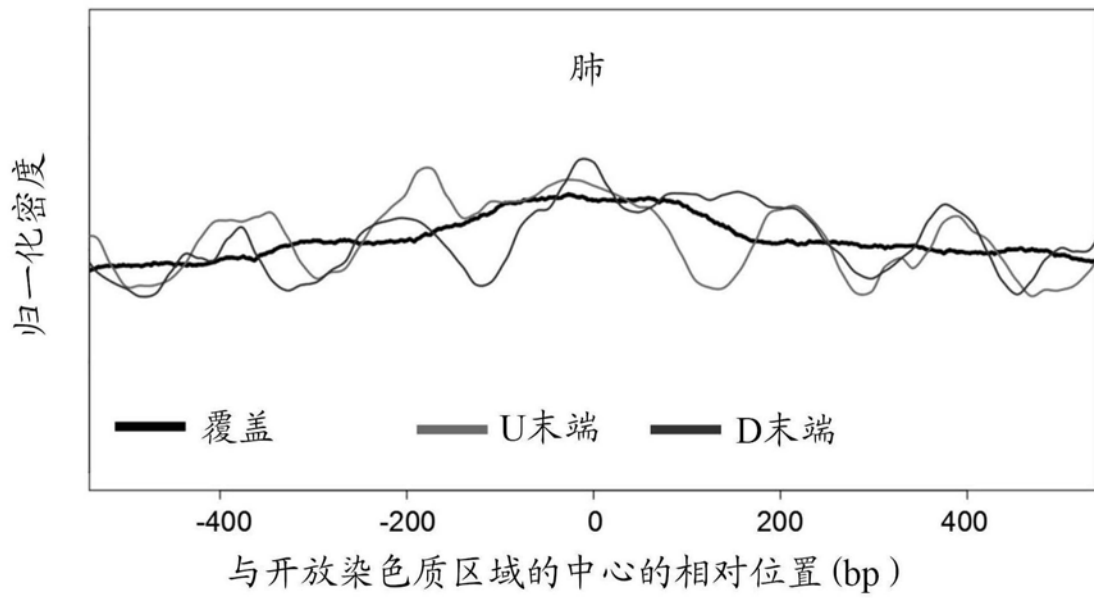


图28D

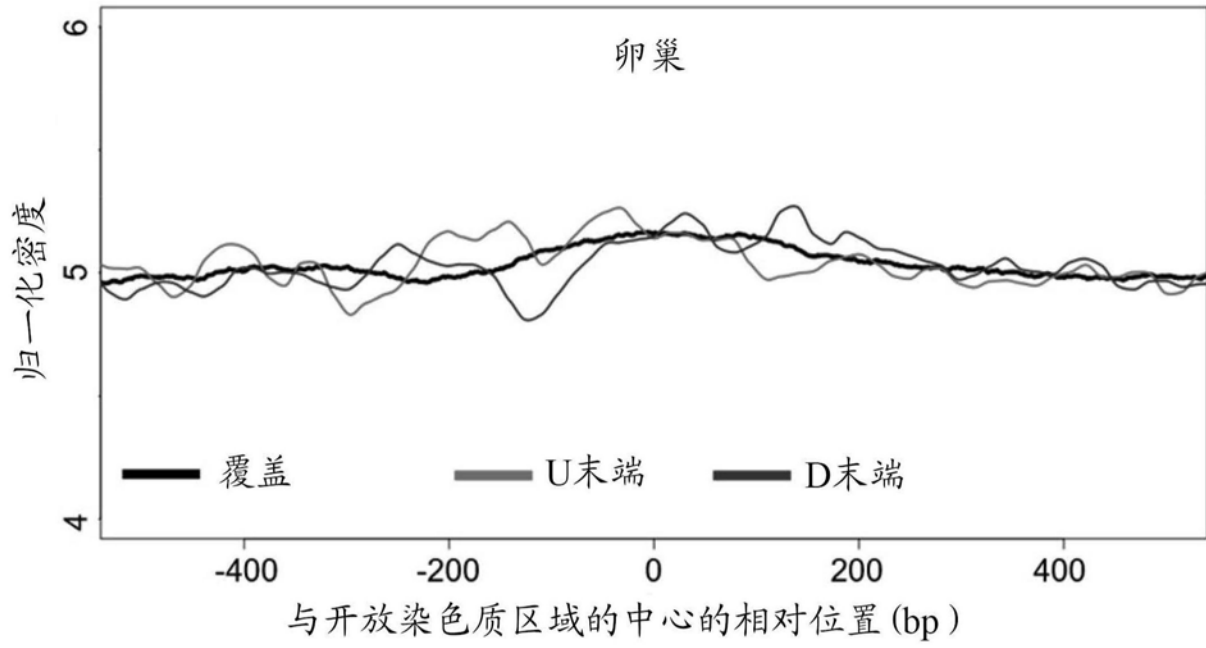


图28E

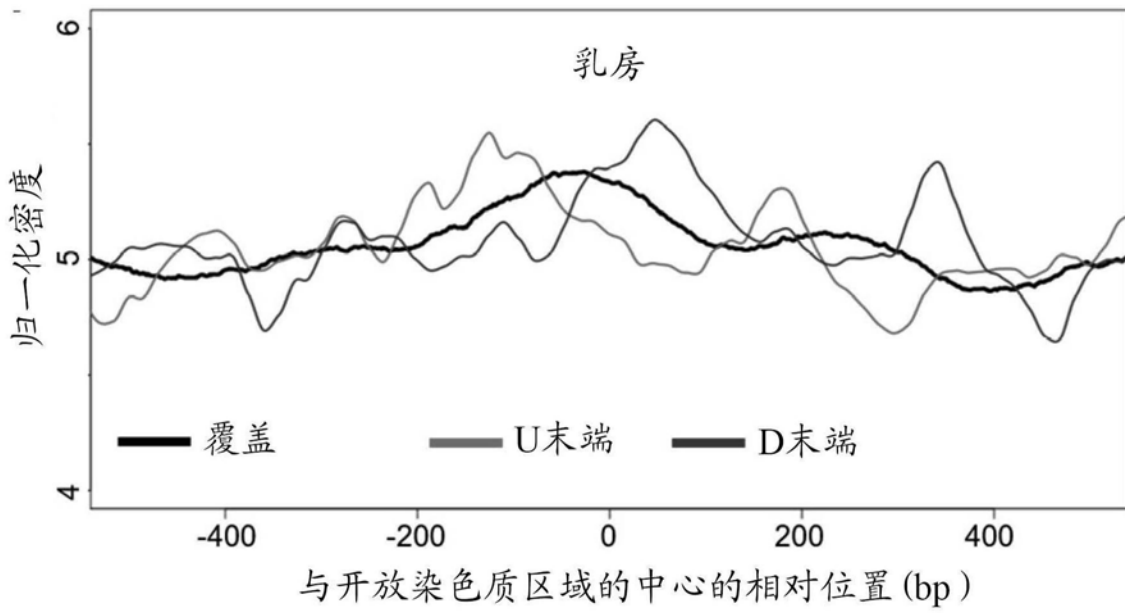


图28F

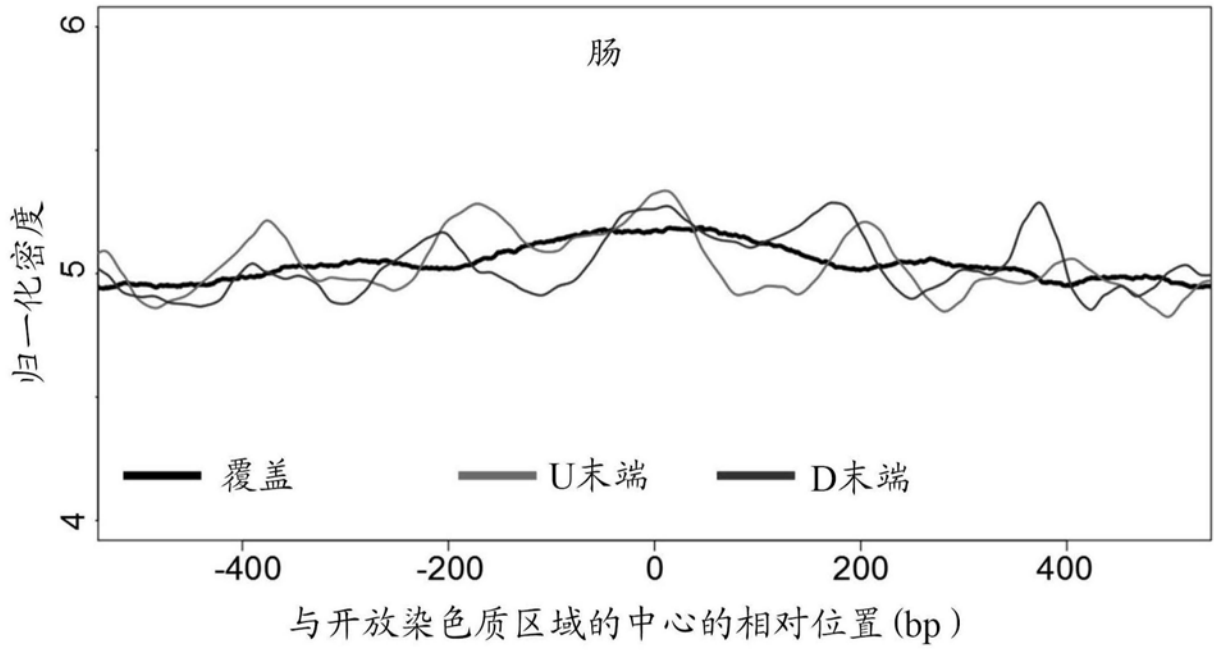


图28G

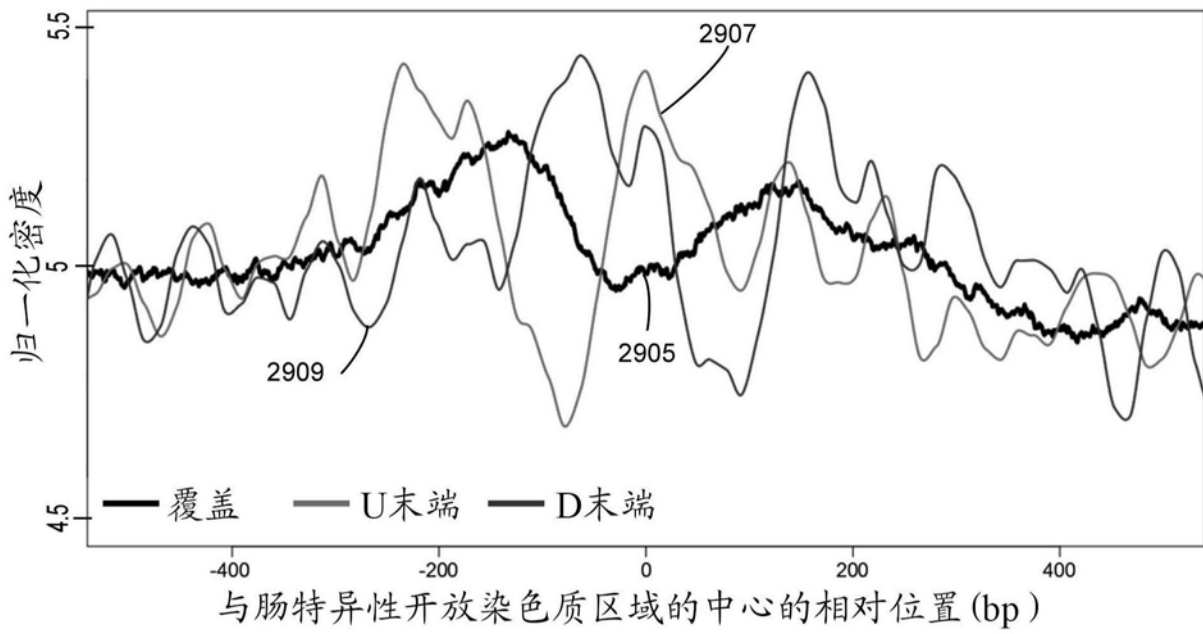


图29A

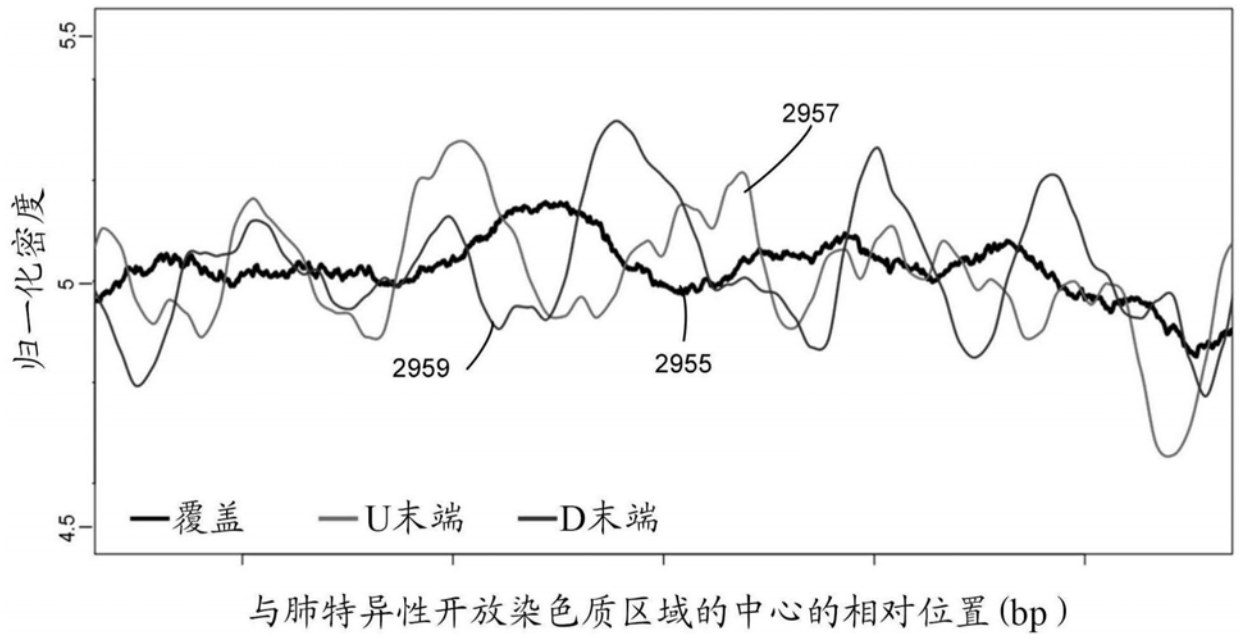


图29B

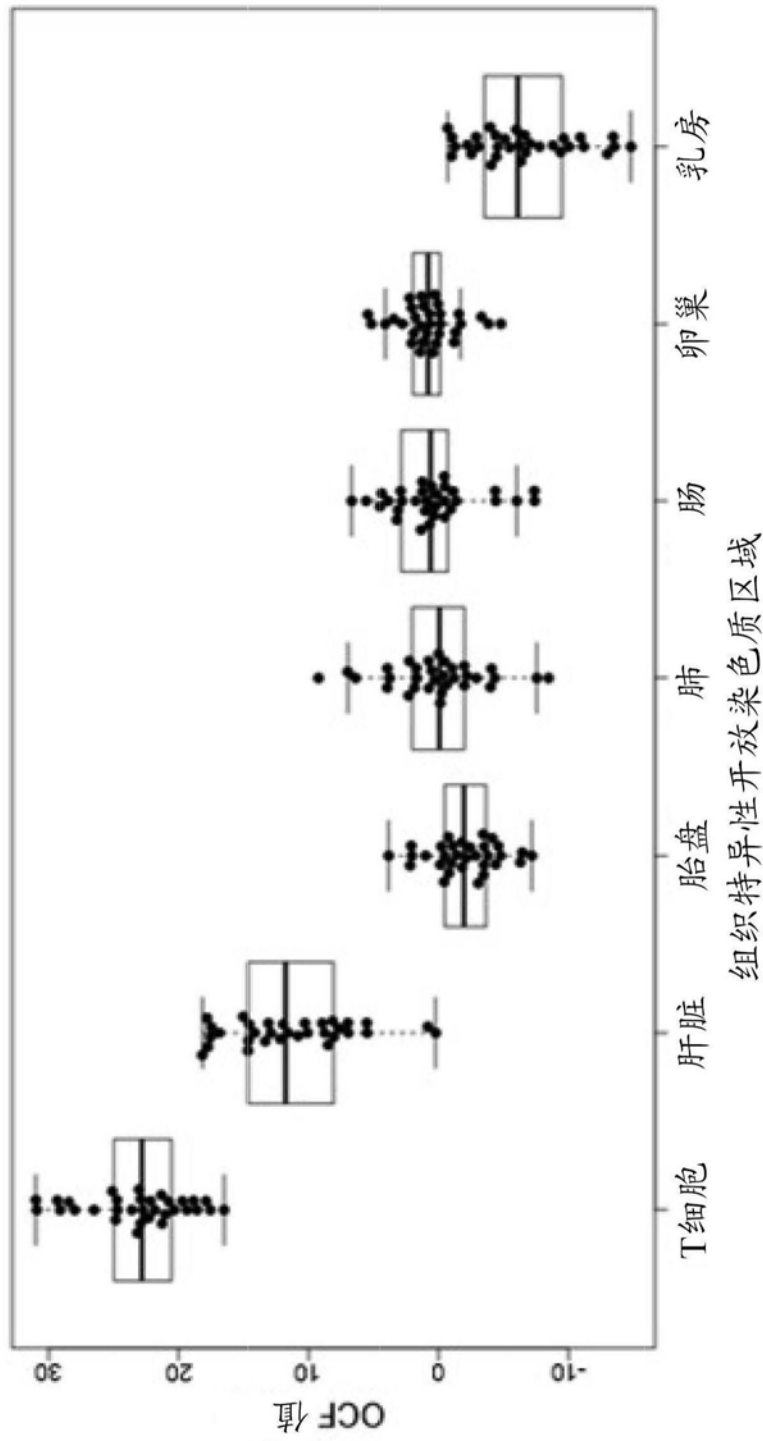


图30

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞
C309	-0.5	8.1	-4.0	0.7	-10.9	-6.3	24.8
C310	-7.4	17.8	-4.3	-0.1	-14.8	-4.4	21.0
C311	-4.4	8.0	-2.2	-1.3	-13.5	-6.5	21.8
C312	-7.4	8.8	0.8	0.4	-13.0	-3.4	18.8
C313	0.2	10.1	1.7	2.1	-3.1	-2.4	29.3
C314	-0.5	10.3	0.3	1.4	-1.0	-1.8	20.3
C315	1.2	14.7	-4.1	1.4	-4.4	1.0	29.1
C316	1.8	10.8	-2.0	1.8	-6.3	-2.1	22.2
C325	0.3	17.7	-1.1	4.1	-2.2	-0.7	23.0
C326	2.8	0.2	1.7	1.2	-9.4	-3.5	22.3
C327	1.0	8.9	-1.2	0.0	-5.1	-0.4	19.7
C329	0.1	17.6	-0.5	0.7	-4.5	-1.0	30.9
C330	5.6	11.6	9.2	0.9	-6.0	2.2	18.6
C331	2.9	5.5	2.2	-1.2	-5.5	2.1	28.0
C332	-0.4	12.2	-0.4	-0.1	-4.0	-3.5	21.3
C333	4.5	12.9	7.0	5.4	-2.5	3.8	25.1
C343	-6.0	0.8	-8.5	0.2	-6.7	-1.4	16.5
C345	-4.4	14.5	3.9	2.2	-1.3	-2.9	21.3
C346	0.5	7.6	-2.9	1.3	-10.0	-4.8	17.5
C347	4.4	14.6	6.4	3.5	-1.0	-4.1	31.0
C348	1.3	14.1	0.5	-3.8	-8.8	-7.2	20.7
C350	1.0	15.0	1.8	2.1	-6.6	-3.5	26.5
C351	6.7	8.5	3.8	-1.7	-7.0	-1.1	28.4
C352	0.7	5.5	-2.0	0.3	-4.6	-4.6	23.2
C353	3.9	17.5	0.0	-3.3	-11.2	-3.1	22.7
C354	-0.9	16.9	0.8	5.2	-6.2	2.1	22.9
C355	-1.4	7.0	4.0	-4.8	-9.6	-1.8	24.7
C356	-1.1	12.0	-0.2	2.8	-13.4	-0.8	19.4
C357	3.1	13.3	2.3	-1.6	-2.9	-0.5	17.9
C358	3.2	18.2	-0.6	1.9	-7.8	-3.8	24.7
C359	-0.5	13.1	-0.1	0.1	-4.0	-0.2	22.9
C360	1.2	6.9	-7.6	1.2	-0.7	-0.1	23.6

图31

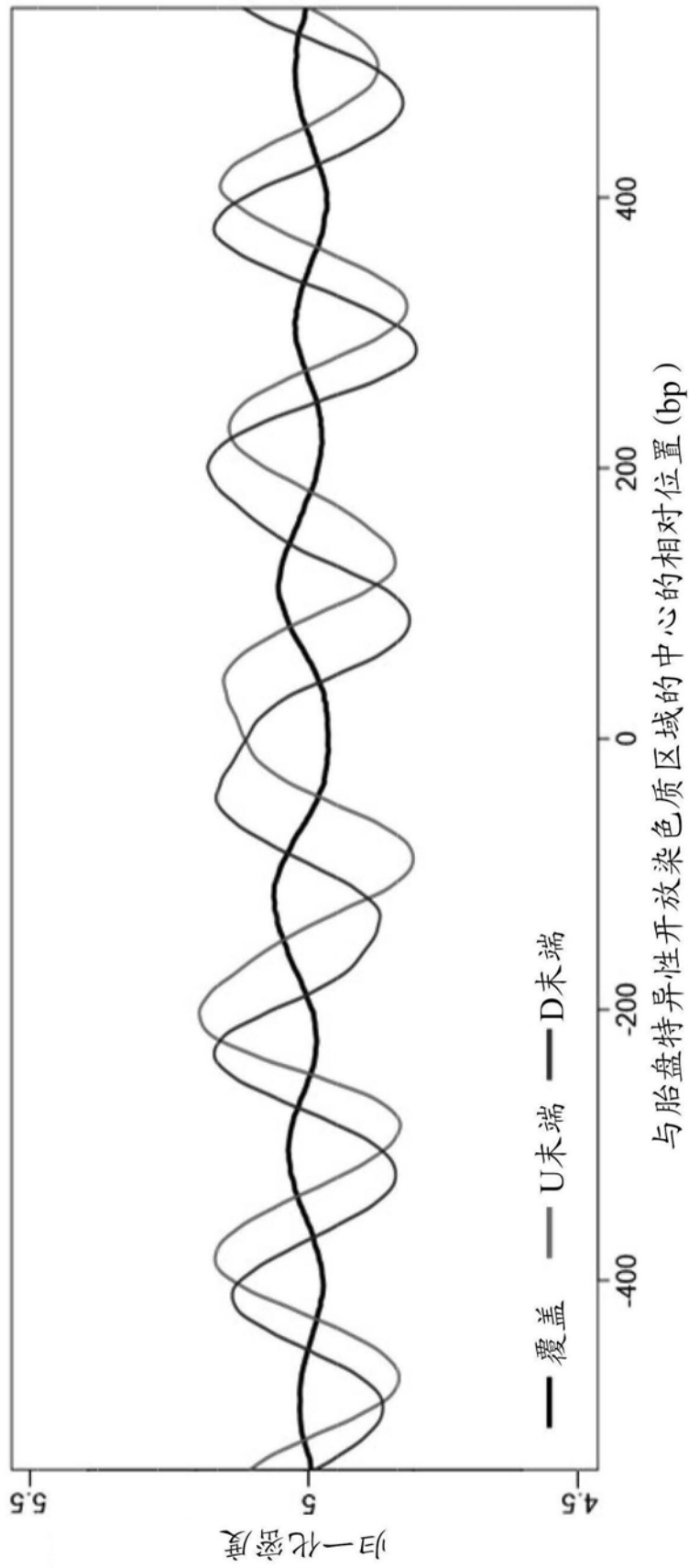


图32A

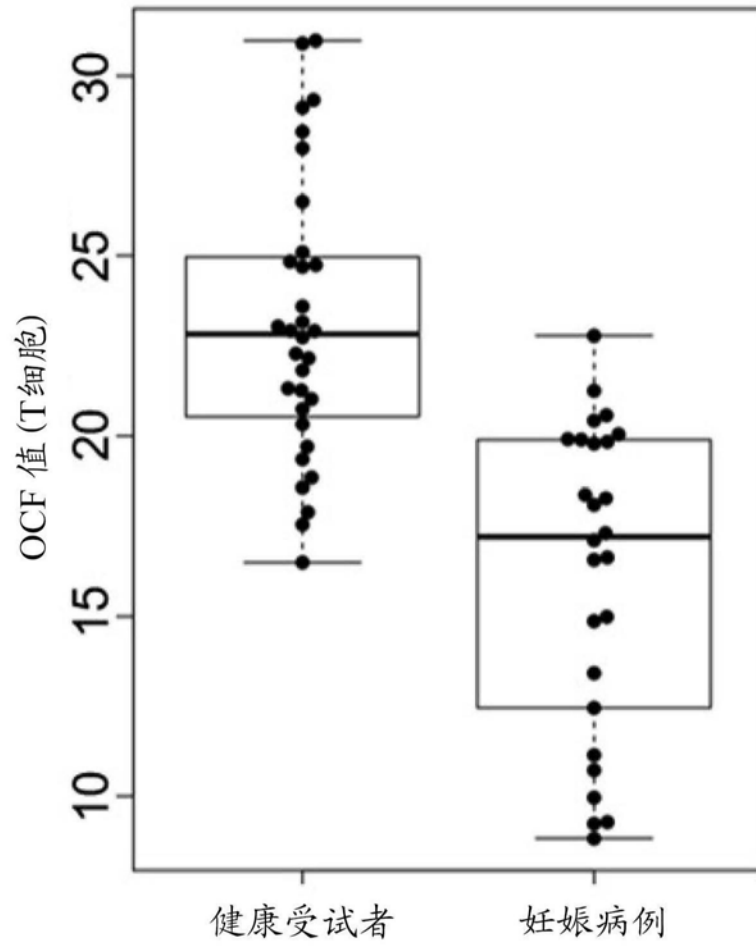


图32B

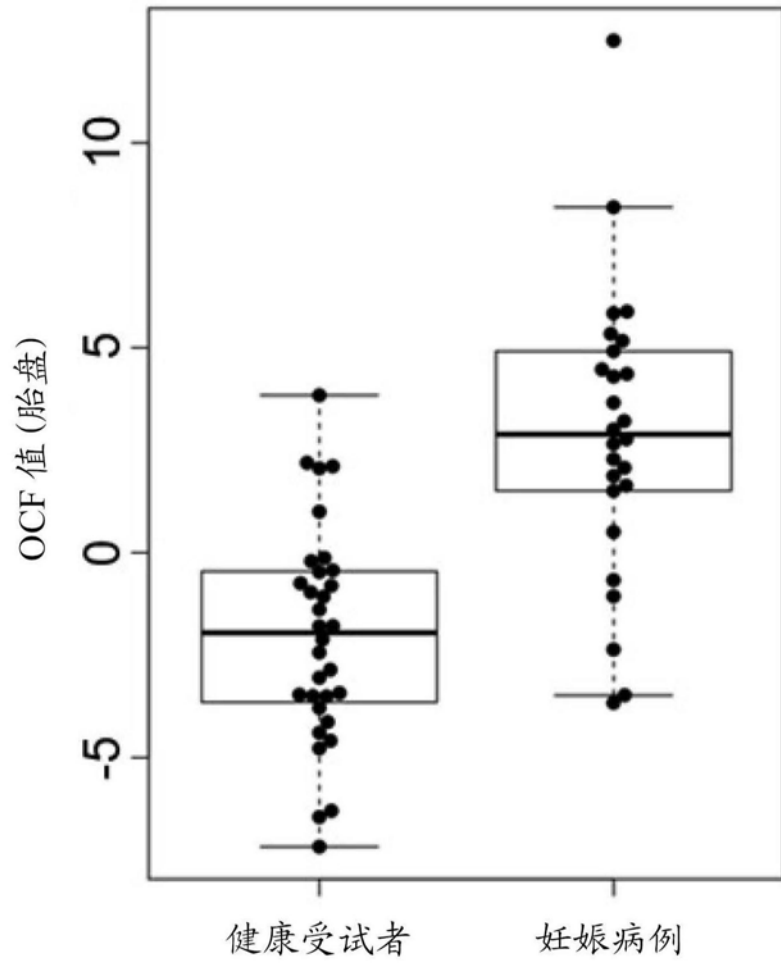


图32C

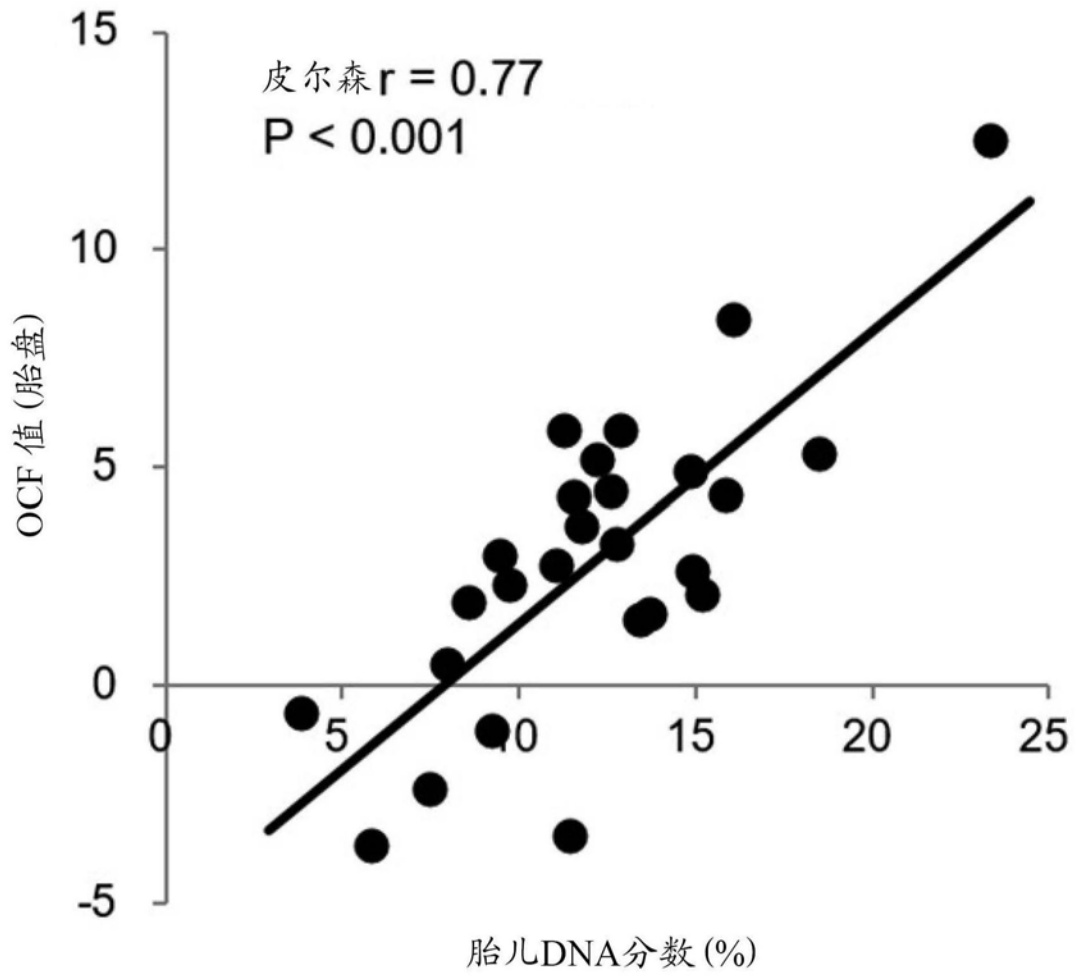


图32D

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞	胎儿DNA分数(%)
8378	-3.3	15.8	-6.2	5.1	-6.3	-3.5	17.1	11.5
8547	-0.6	7.8	13.5	-0.1	-4.7	3.0	18.4	9.5
8553	-2.6	11.0	-1.7	-3.4	-5.8	2.1	20.4	15.2
8667	-9.9	22.6	2.8	-4.1	-10.5	-1.1	19.9	9.3
8715	1.3	5.4	-1.3	5.9	-9.2	12.5	11.1	23.4
8736	-3.2	9.1	-4.7	-3.0	-2.7	4.3	19.9	11.6
8777	3.4	24.1	-1.5	2.0	-9.7	2.7	10.0	15.0
8816	3.4	2.8	3.3	1.8	-4.8	2.8	13.4	11.1
8822	2.2	0.1	-2.9	2.1	-9.6	-2.4	9.2	7.5
8868	3.2	19.4	2.2	2.1	-8.5	4.4	16.6	15.9
8872	-3.4	10.1	0.4	1.2	-7.6	5.2	19.8	12.3
8901	-1.6	10.3	2.2	-0.7	-10.5	2.3	18.1	9.8
8903	2.5	6.4	5.3	-3.5	-8.0	3.2	14.9	12.8
8941	1.6	21.8	-5.6	5.2	-2.8	3.7	19.8	11.9
8949	3.5	9.2	-5.2	5.9	-11.4	4.9	18.3	14.9
8950	-2.4	5.6	8.8	0.1	-7.4	0.5	20.6	8.0
8957	-4.2	18.2	6.4	0.8	-9.3	-3.7	12.5	5.9
9018	6.1	13.2	1.6	1.1	-18.9	5.9	8.8	12.9
9087	6.1	10.7	-3.2	3.2	-7.6	8.4	9.3	16.1
9100	0.4	21.2	-3.4	7.2	-13.6	5.8	10.7	11.3
9109	2.1	2.1	-6.3	-0.8	-15.7	4.5	22.8	12.6
9160	-3.7	16.4	-0.1	1.5	-8.8	-0.7	17.3	3.9
9195	1.9	10.5	7.4	2.6	-6.6	1.9	21.3	8.6
9229	-5.3	2.6	-4.4	-2.5	-11.9	1.5	15.0	13.5
9238	3.4	10.3	0.0	4.9	-4.8	5.3	20.0	18.5
9242	-7.4	12.7	-2.3	-4.0	-13.8	1.6	16.6	13.8

图33

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞	供体DNA分数 (%)
T99	-5.9	3.5	-6.3	-0.6	-12.2	-6.0	12.0	11.6
TBR1448	2.7	14.9	-0.1	0.9	-6.2	-1.8	15.9	2.6
TBR1449	0.9	3.4	4.5	-3.4	-12.8	-6.9	21.8	5.3
TBR1450	6.2	11.7	-3.8	-3.9	-11.0	-6.9	17.6	23.7
TBR1451	-4.0	6.6	-10.3	-3.3	-6.3	-3.5	14.3	3.3
TBR1452	-4.9	3.4	-5.7	-1.9	-9.8	-2.3	14.6	2.6
TBR1453	4.9	26.4	-0.3	-0.2	-6.1	-6.7	14.0	39.9
TBR1454	-3.4	3.2	-5.4	1.1	-4.1	-2.3	18.7	4.8
TBR1573	-5.2	1.9	-0.8	1.8	-3.6	-4.3	27.5	1.8
TBR1574	-2.5	13.8	-8.8	5.5	-2.2	-2.2	15.7	16.5
TBR1575	12.0	8.1	-6.0	3.2	-10.8	-5.3	17.9	13.4
TBR1576	0.1	20.4	-0.6	1.5	-8.0	-5.3	19.7	10.1
TBR1577	-3.6	1.5	-2.3	-1.0	-7.0	-2.1	20.5	2.1
TBR1578	4.7	15.4	-1.2	2.8	-3.8	-3.9	20.9	18.4

图34

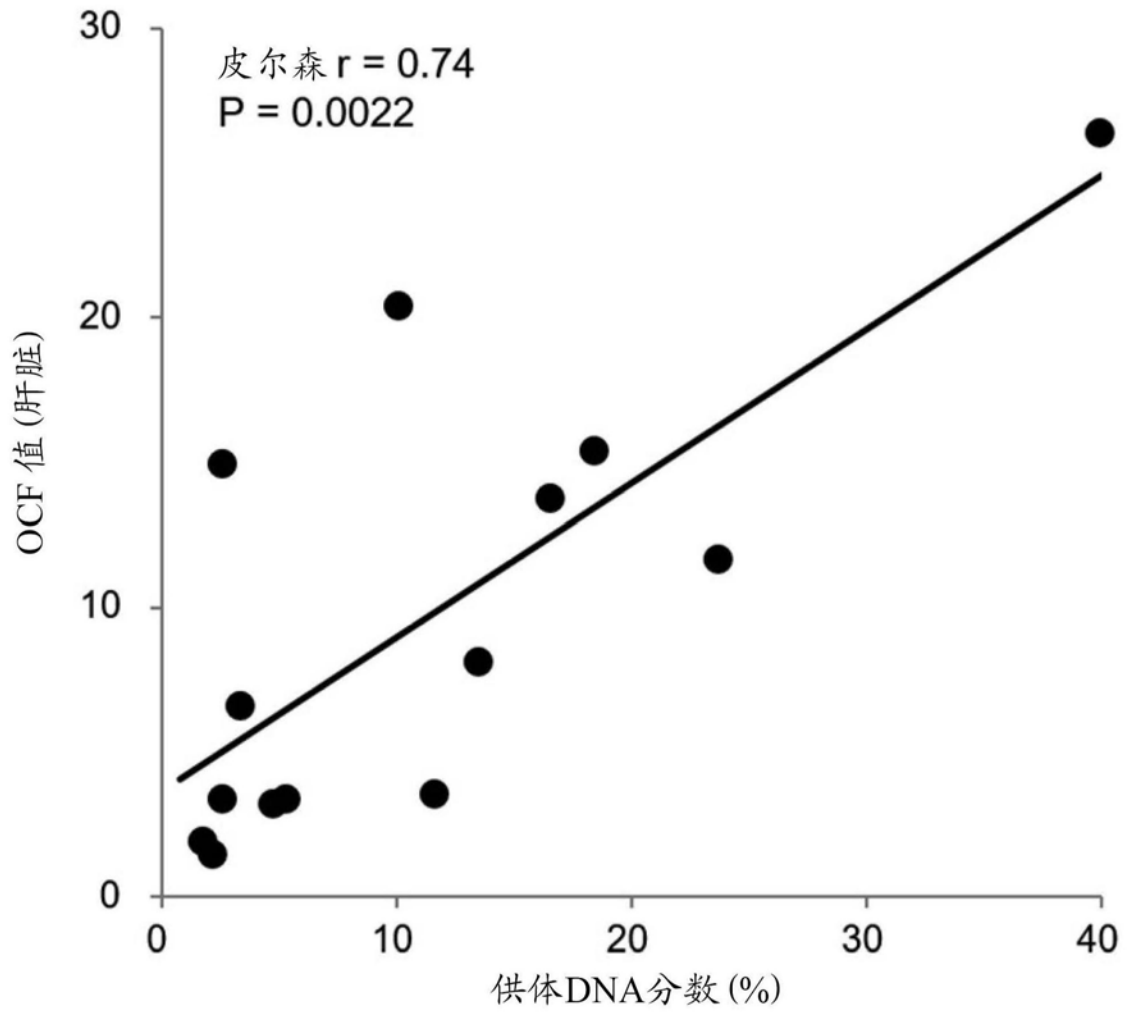


图35A

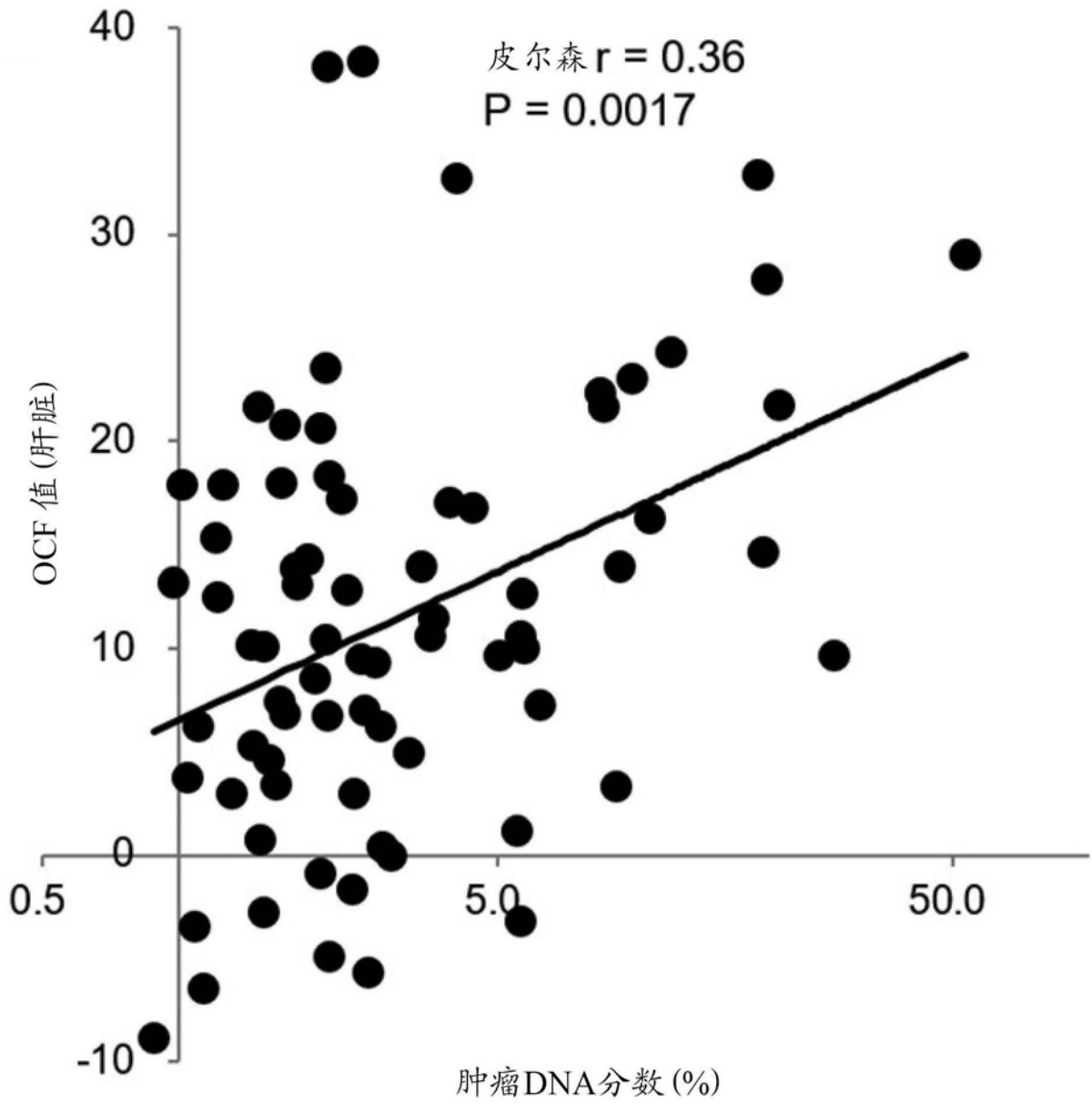


图35B

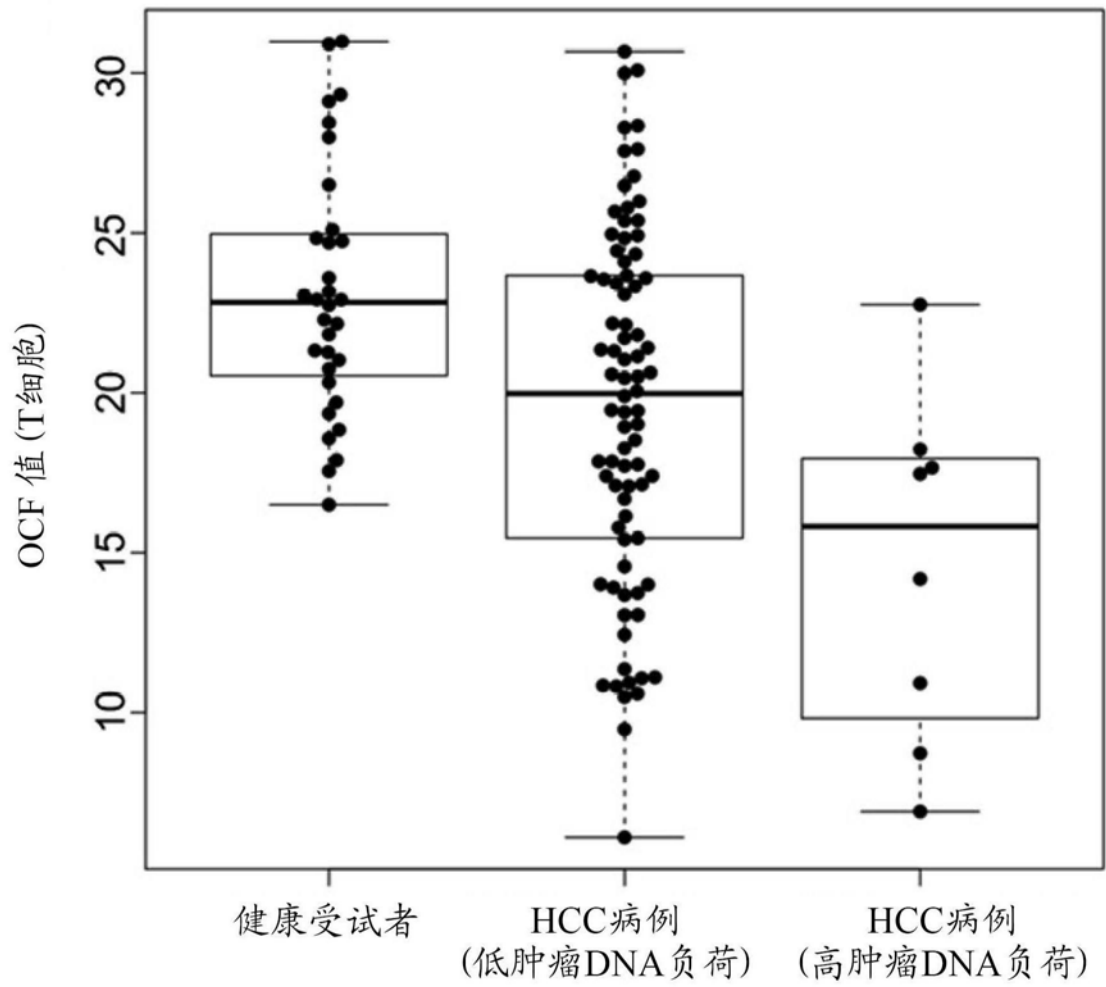


图35C

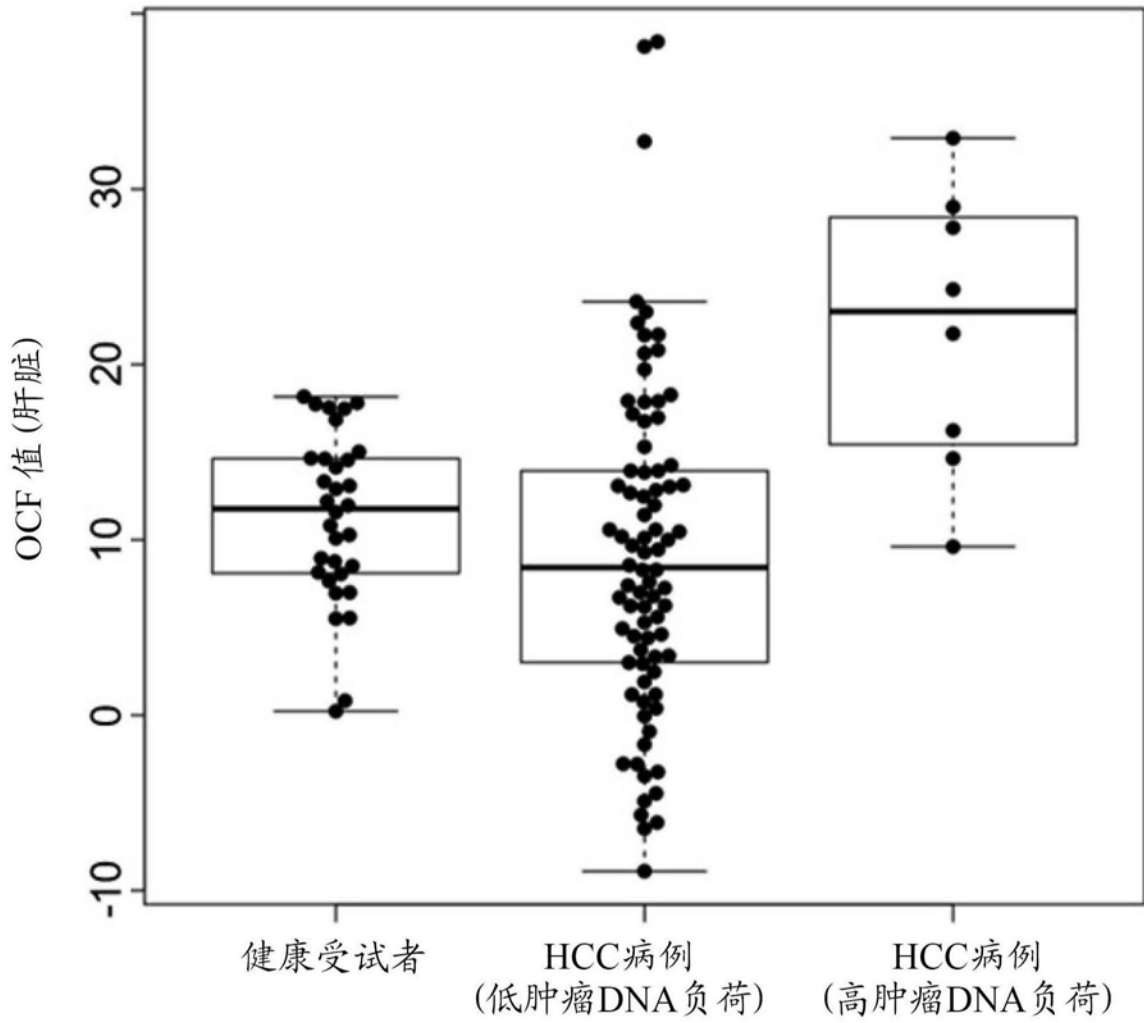


图35D

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞	肿瘤DNA分数(%)
H203	-0.2	12.0	-3.2	1.2	-13.9	-4.3	26.5	0
H209	-1.9	4.4	-3.2	-1.3	-3.6	-0.7	21.8	0
H246	-2.0	2.5	-9.8	-5.4	-2.9	-2.6	17.8	0
H247	5.4	6.2	-5.4	4.8	-4.7	1.7	25.7	0
H249	-0.8	5.6	-0.7	3.2	-13.5	-4.9	25.4	0
H253	1.1	7.6	-7.0	-0.8	-10.2	-3.7	22.2	0
H254	3.3	1.2	2.2	-5.8	-14.5	-5.1	21.7	0
H269	4.3	8.3	-1.5	-2.6	-0.8	-3.5	20.5	0
H275	0.2	8.3	-3.7	1.4	-8.2	1.8	25.0	0
H288	4.4	4.5	-3.1	-4.1	-5.2	-4.2	21.3	0
H315	1.0	13.0	-4.2	-3.5	-9.3	-1.5	17.9	0
H324	-10.0	-4.5	-11.6	-2.3	-15.6	-10.6	16.7	0
H342	-1.6	-6.1	-2.2	-2.7	-6.3	-4.3	21.3	0
H421	-3.0	1.9	-0.8	-4.5	-1.6	-4.5	20.6	0
H436	1.4	-2.8	-1.9	-4.2	-5.7	-4.9	18.3	0
HOT417	4.4	19.7	6.3	-0.2	-6.4	-1.6	24.9	0
H263	-11.9	-8.9	-17.1	-7.3	-21.5	-14.4	10.9	0.9
H220	-7.0	13.1	-11.9	-11.0	-12.3	-9.3	17.4	1.0
HOT432	7.1	17.9	2.3	-0.3	-8.3	0.4	27.6	1.0
H250	2.4	3.7	0.0	-1.4	-3.0	-0.2	23.7	1.0
H292	-3.1	-3.5	-4.7	1.8	-11.1	-7.4	26.0	1.1
H424	1.6	6.2	-5.1	-1.6	-5.3	-2.4	20.5	1.1

图36A

H261	-7.0	-6.5	-13.3	-3.9	-18.9	-13.8	16.1	1.1
H206	0.7	15.3	-6.0	1.2	-11.6	-6.5	18.5	1.2
H296	4.7	12.5	-0.8	-1.8	-1.3	-2.5	23.5	1.2
HOT410	4.3	17.9	0.9	1.2	-4.1	3.3	22.1	1.3
H300	-1.0	3.0	-3.7	-1.3	-7.0	-4.3	21.0	1.3
H196	0.3	10.2	-2.9	-1.9	-7.4	0.4	23.3	1.4
H218	-9.8	5.3	-7.5	5.0	-6.8	-6.0	13.7	1.5
HOT428	0.5	21.7	5.1	3.5	1.8	-1.2	23.7	1.5
H433	-2.1	0.8	-4.9	0.6	-12.4	-3.6	19.4	1.5
H341	-2.2	-2.8	-4.8	-0.9	-9.1	-9.3	24.3	1.5
HOT425	1.2	10.1	-0.2	9.3	-8.4	-3.8	26.8	1.5
H311	-1.7	4.6	-9.7	4.3	-13.0	-8.3	12.4	1.6
H283	-5.1	3.4	-7.5	-3.1	-12.8	-6.9	17.1	1.6
H374	-3.4	7.4	-4.2	-2.8	-15.3	-4.5	23.6	1.7
HOT413	4.8	17.9	2.1	6.3	-2.8	-3.3	18.9	1.7
H285	0.7	6.8	-9.4	-4.5	-7.6	-4.3	19.9	1.7
HOT427	3.6	20.8	4.9	-0.6	-2.0	6.1	30.7	1.7
H223	0.6	13.8	0.2	2.9	-9.3	-3.3	14.0	1.8
HOT393	-3.8	13.1	-4.3	2.1	-6.8	-0.7	10.5	1.8
H323	1.0	14.2	-9.2	4.5	-6.3	-6.1	11.1	1.9
HOT386	-1.6	8.5	-1.7	-2.9	-4.7	-5.3	21.4	2.0
H225	4.6	-0.9	5.7	5.8	-11.0	-5.2	14.0	2.0
H402	-0.1	20.6	-8.5	1.7	-11.5	-7.7	15.4	2.0
H228	1.4	10.5	4.5	-2.0	-9.3	-0.2	30.0	2.1
H419	4.5	23.6	-3.3	9.1	-7.1	-3.3	9.5	2.1
H377	4.6	6.7	-5.4	5.3	-7.1	-2.4	15.8	2.1
HOT422	2.5	38.1	5.2	5.7	-9.7	1.7	11.4	2.1
H278	4.3	18.3	-5.6	0.4	-9.4	-5.2	25.8	2.1

图36B

H264	-5.9	-4.9	-11.7	-13.0	-15.4	-15.2	13.0	2.1
H232	-3.0	17.2	-5.7	8.7	-8.9	-7.2	6.1	2.3
H252	3.6	12.8	1.9	-0.2	-10.8	1.8	27.6	2.3
HOT397	-0.4	-1.7	-14.3	4.4	-9.3	-6.4	13.0	2.4
H352	-6.8	2.9	-6.2	-4.4	-12.3	-13.3	19.5	2.4
H230	-0.4	9.4	0.9	1.0	-9.8	-3.6	23.1	2.5
H301	4.2	38.4	7.2	12.8	2.0	8.0	28.3	2.5
HOT394	-0.6	7.0	-11.0	1.9	-9.5	-7.8	17.4	2.6
H435	1.0	-5.7	-2.9	3.1	-6.5	-1.8	10.6	2.6
H321	1.8	9.3	-9.3	-3.2	-12.9	-9.3	17.1	2.7
H200	-1.7	6.2	-1.0	2.4	-2.1	-2.2	17.1	2.8
H256	-3.8	0.4	-7.8	-5.2	-16.2	-12.1	17.8	2.8
H344	-4.9	0.0	-8.5	-2.2	-6.1	-7.3	19.4	2.9
H226	3.5	4.9	0.1	-1.4	-6.7	-1.2	24.4	3.2
H214	3.0	13.9	-4.8	1.7	-9.2	-3.8	15.5	3.4
H390	-1.5	10.6	-5.6	-3.3	-11.1	-5.5	19.0	3.6
H217	-2.8	11.4	-1.5	0.5	-6.8	-5.2	25.4	3.6
H420	2.8	17.0	-7.0	4.0	-3.6	-5.2	10.8	3.9
HOT426	5.6	32.7	0.7	3.9	1.9	-2.3	14.6	4.1
H239	3.0	16.8	-1.4	-0.6	-11.6	-4.4	21.1	4.4
GM1100	0.9	9.7	-4.3	1.9	-4.6	-3.7	24.1	5.0

图36C

H381	-4.4	1.2	-9.9	-2.1	-13.1	-7.3	17.7		5.5
H221	0.7	10.6	-3.0	-2.4	-2.1	-1.3	24.8		5.6
H266	-9.2	-3.2	-9.3	-9.0	-24.8	-17.0	10.8		5.6
H199	6.0	12.7	6.5	-5.3	-4.5	-1.5	30.1		5.7
H267	1.9	10.0	3.0	-3.1	-9.1	1.4	28.4		5.7
H195	0.6	7.2	-0.3	2.0	-5.4	-0.5	23.4		6.2
H216	0.5	22.4	-1.7	-1.7	-11.9	-6.5	13.9		8.4
H210	5.4	21.7	-1.3	-0.3	-12.7	-6.1	13.7		8.6
H262	-3.7	3.3	-9.2	-7.3	-15.1	-11.9	11.1		9.1
H270	7.9	13.9	-4.4	4.1	-7.4	-0.5	20.1		9.3
H272	5.2	23.0	1.2	1.3	-6.0	-3.3	20.6		9.9
H235	3.9	16.2	-8.2	-5.5	-10.1	-5.3	10.9		10.8
H234	11.5	24.3	7.3	0.2	2.4	-0.1	18.2		12.1
HOT414	4.7	32.9	-8.3	2.0	-1.9	-0.1	14.2		18.8
H258	-0.4	14.6	-13.2	-7.1	-10.8	-8.2	22.8		19.3
HOT403	8.3	27.8	1.4	-3.7	-5.5	-3.2	17.7		19.5
H423	2.5	21.8	-2.5	-3.5	-7.4	-2.7	17.5		20.8
HOT412	6.8	9.6	-1.7	-0.5	-6.7	-2.6	8.7		27.5
H291	4.4	29.0	-13.3	-1.3	-9.3	-4.5	6.9		53.2

图36D

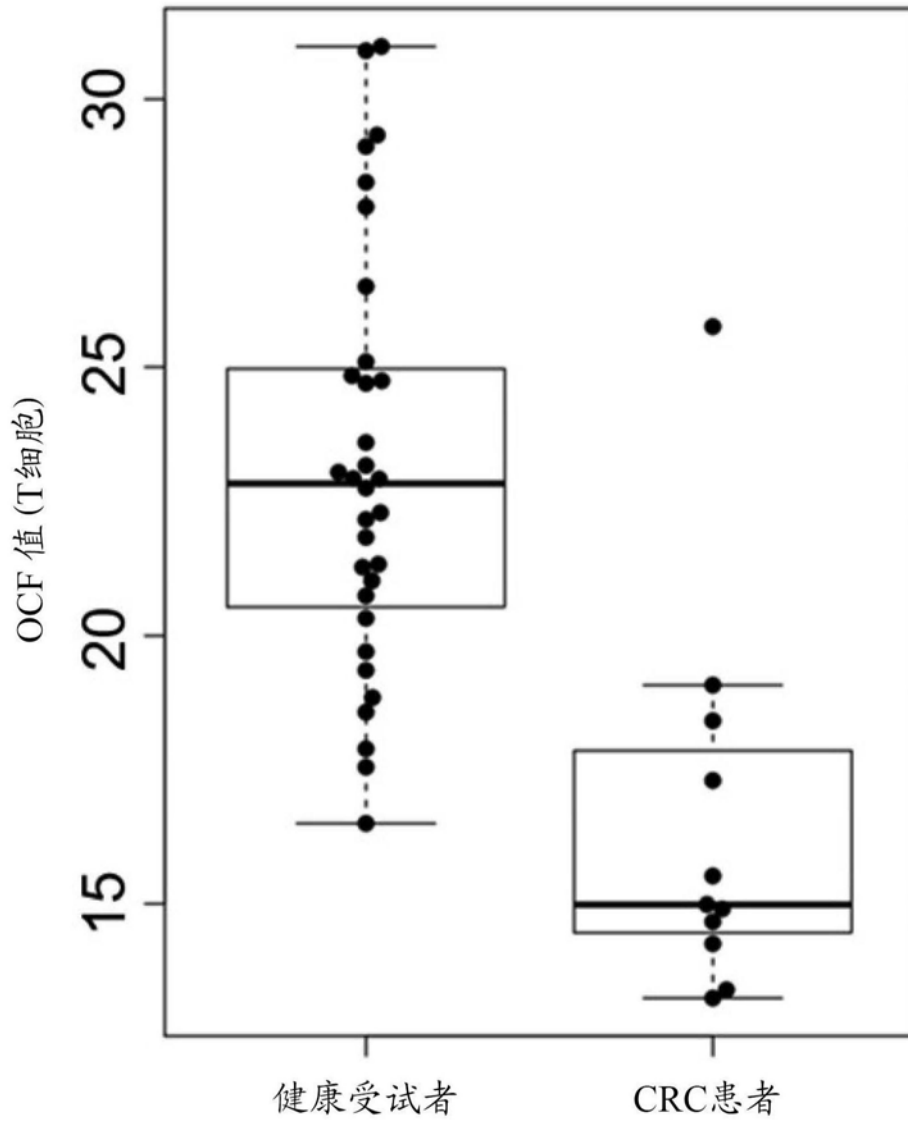


图37A

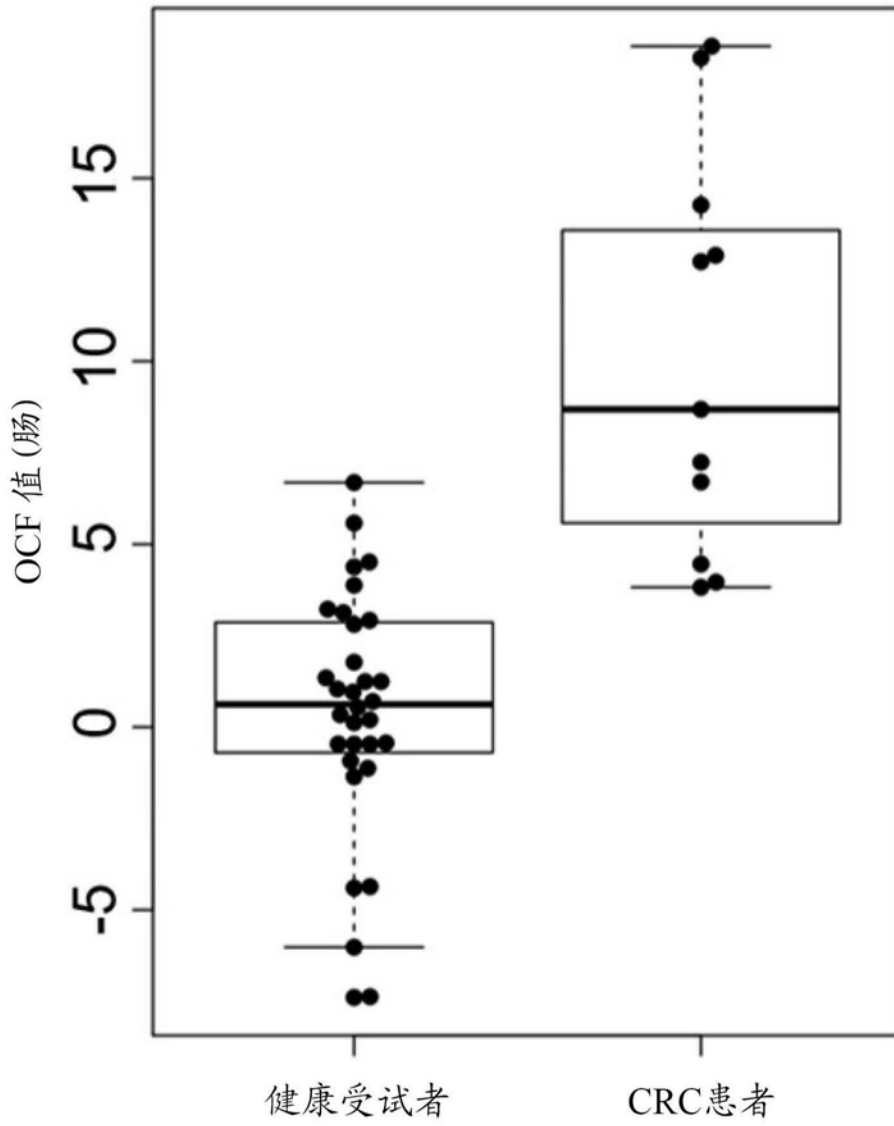


图37B

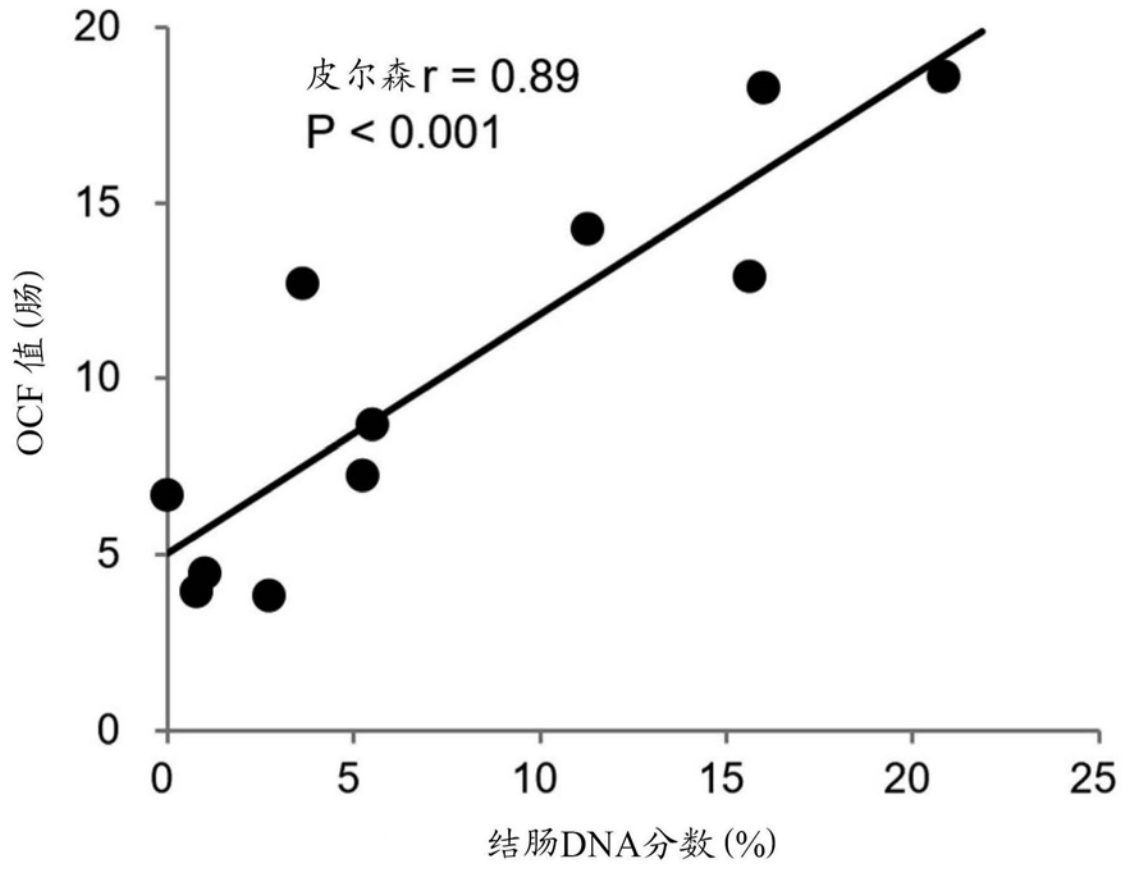


图37C

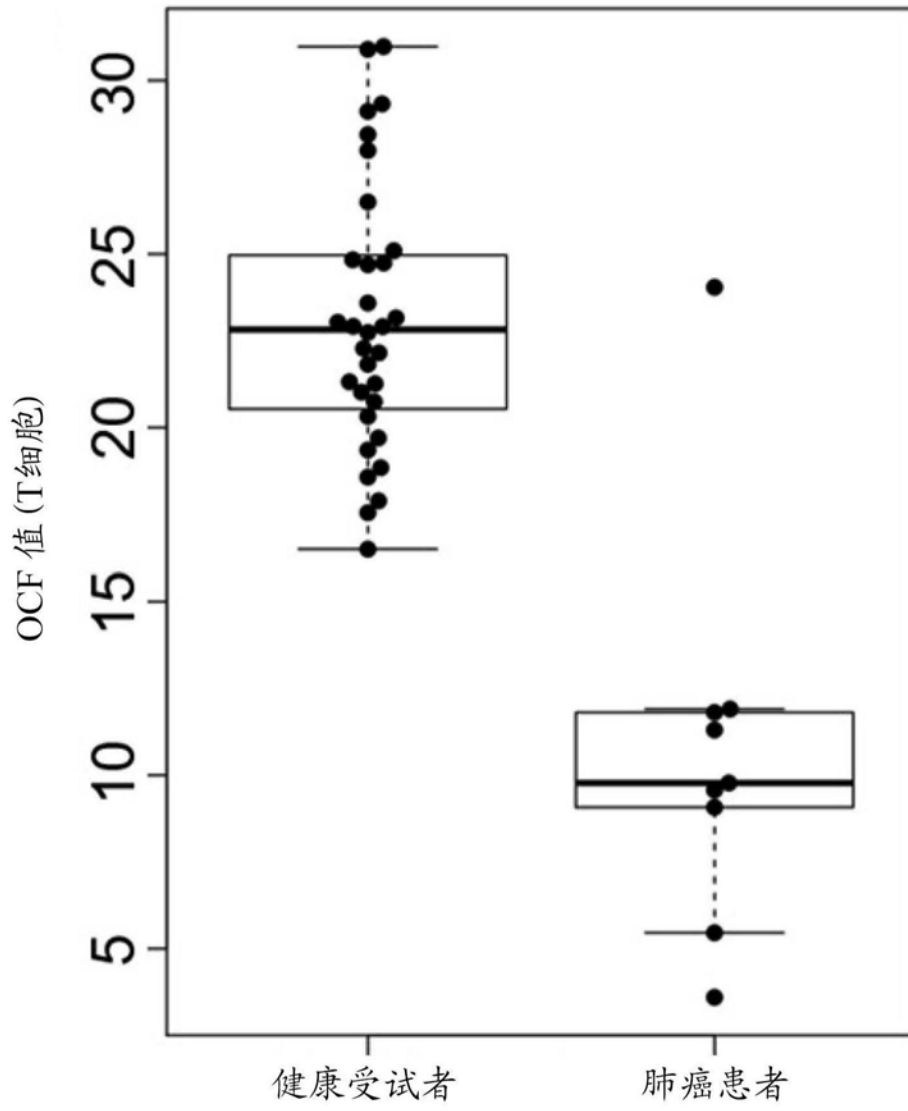


图37D

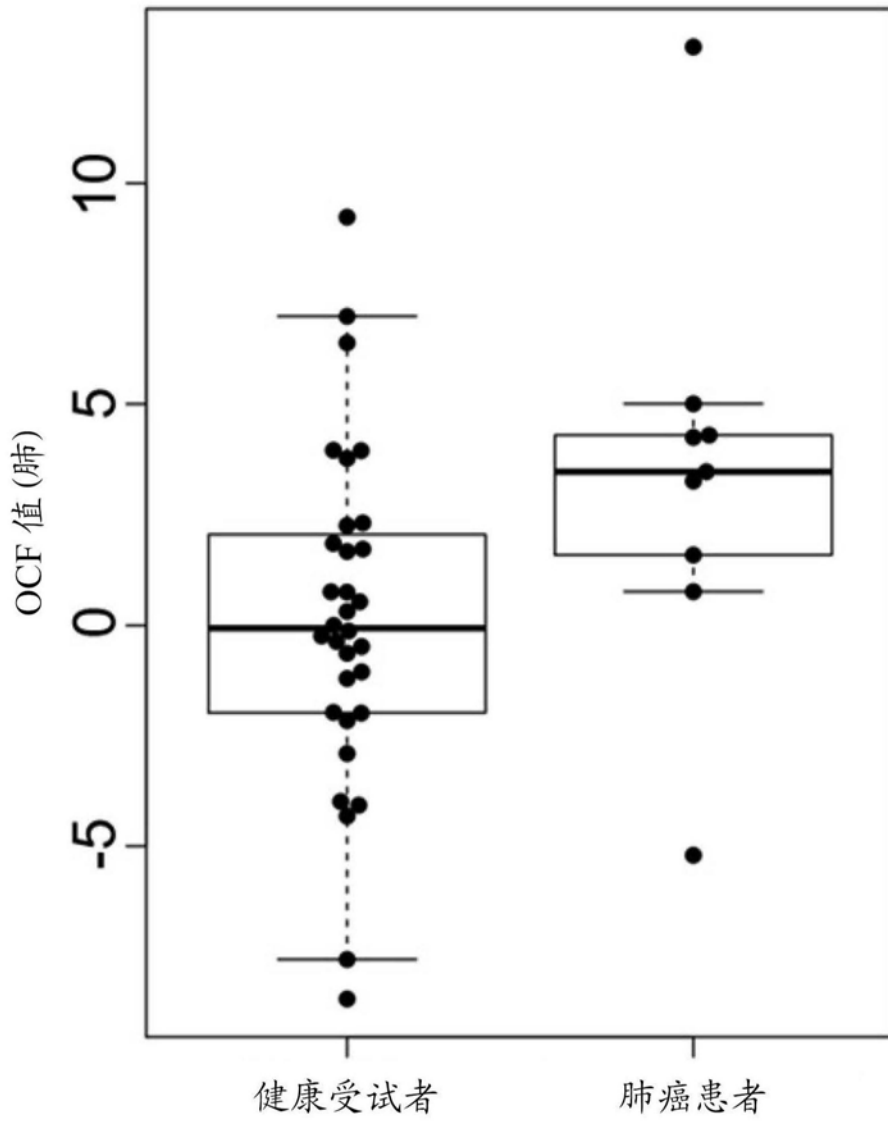


图37E

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞
IC05	5.1	23.0	13.1	4.1	1.8	4.2	24.0
IC06	3.9	30.0	4.2	3.7	-5.7	8.4	9.8
IC10	3.1	13.2	0.8	-3.2	-12.8	-5.3	9.1
IC15	0.9	14.7	-5.2	-0.3	-14.4	-6.7	3.6
IC20	3.2	13.8	4.3	1.2	-3.0	3.0	5.5
IC22	4.7	19.3	1.6	2.9	4.3	2.7	11.3
IC28	-0.2	18.6	5.0	-0.4	-3.1	1.2	11.8
IC32	1.0	1.7	3.5	0.7	-9.5	-6.6	11.9
IC42	2.9	13.0	3.3	-1.0	-8.3	4.0	9.6

图38

#Sid	肠	肝脏	肺	卵巢	乳房	胎盘	T细胞	结肠DNA贡献(%)
TBR901	18.6	43.6	13.3	6.2	0.4	9.0	25.8	20.9
TBR910	18.3	31.8	3.7	-6.5	1.0	14.0	17.3	16.0
TBR911	4.5	20.7	10.1	-1.1	-10.0	4.8	13.2	1.0
TBR912	12.9	26.8	6.5	5.1	-17.4	6.5	14.3	15.7
TBR914	6.7	17.9	11.1	3.0	-9.4	9.4	18.4	0.0
TBR916	8.7	43.7	15.0	4.9	-1.6	6.4	14.7	5.5
TBR917	3.8	28.2	3.2	-2.5	-7.9	8.4	19.1	2.7
TBR919	12.7	29.1	12.2	2.4	-5.5	9.7	13.4	3.6
TBR921	14.3	32.2	7.1	1.5	-9.6	9.4	15.5	11.3
TBR922	4.0	32.1	2.4	-5.6	-10.8	6.6	15.0	0.8
TBR924	7.2	21.5	5.7	-1.5	-8.5	5.2	14.9	5.2

图39

4000

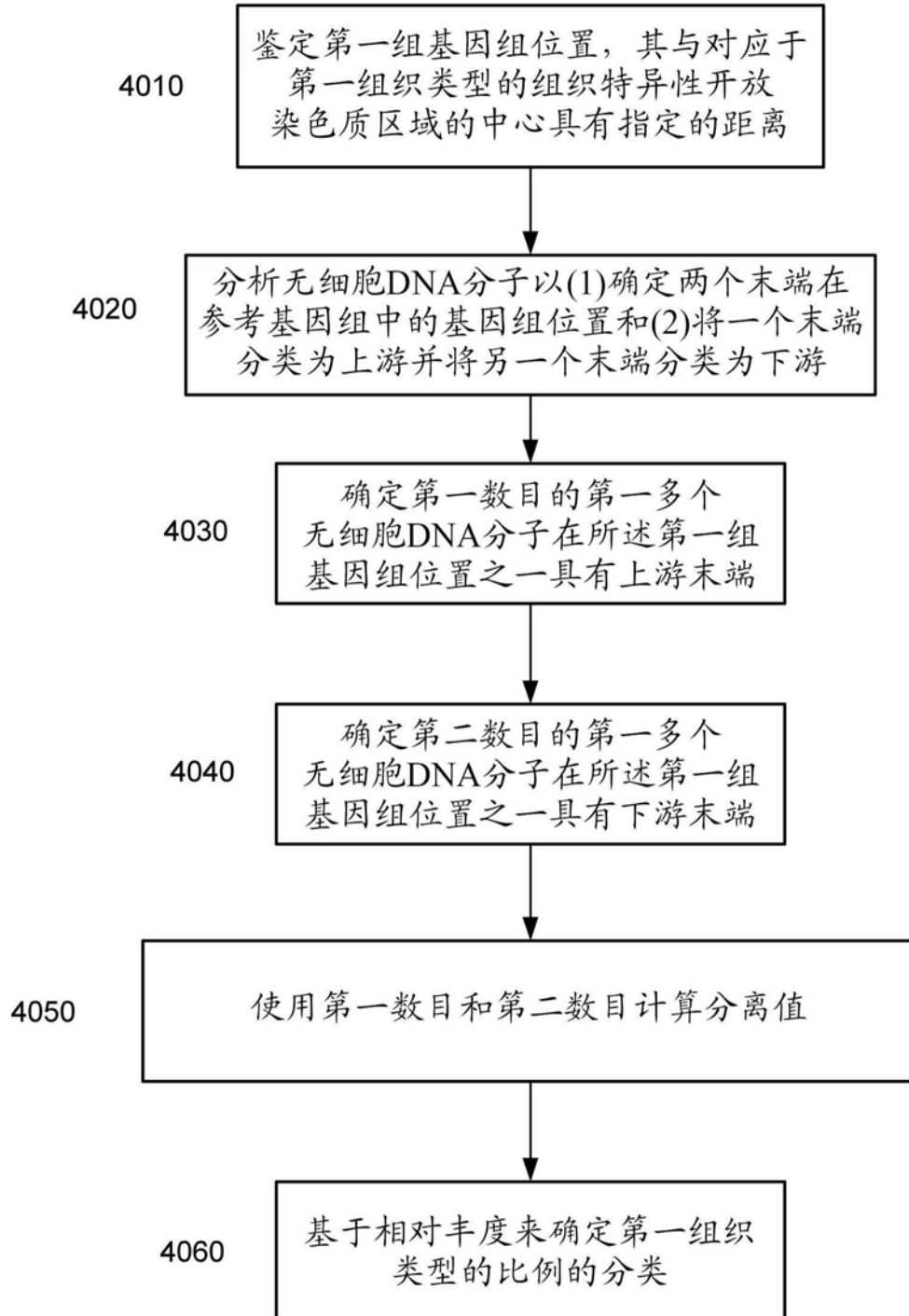


图40

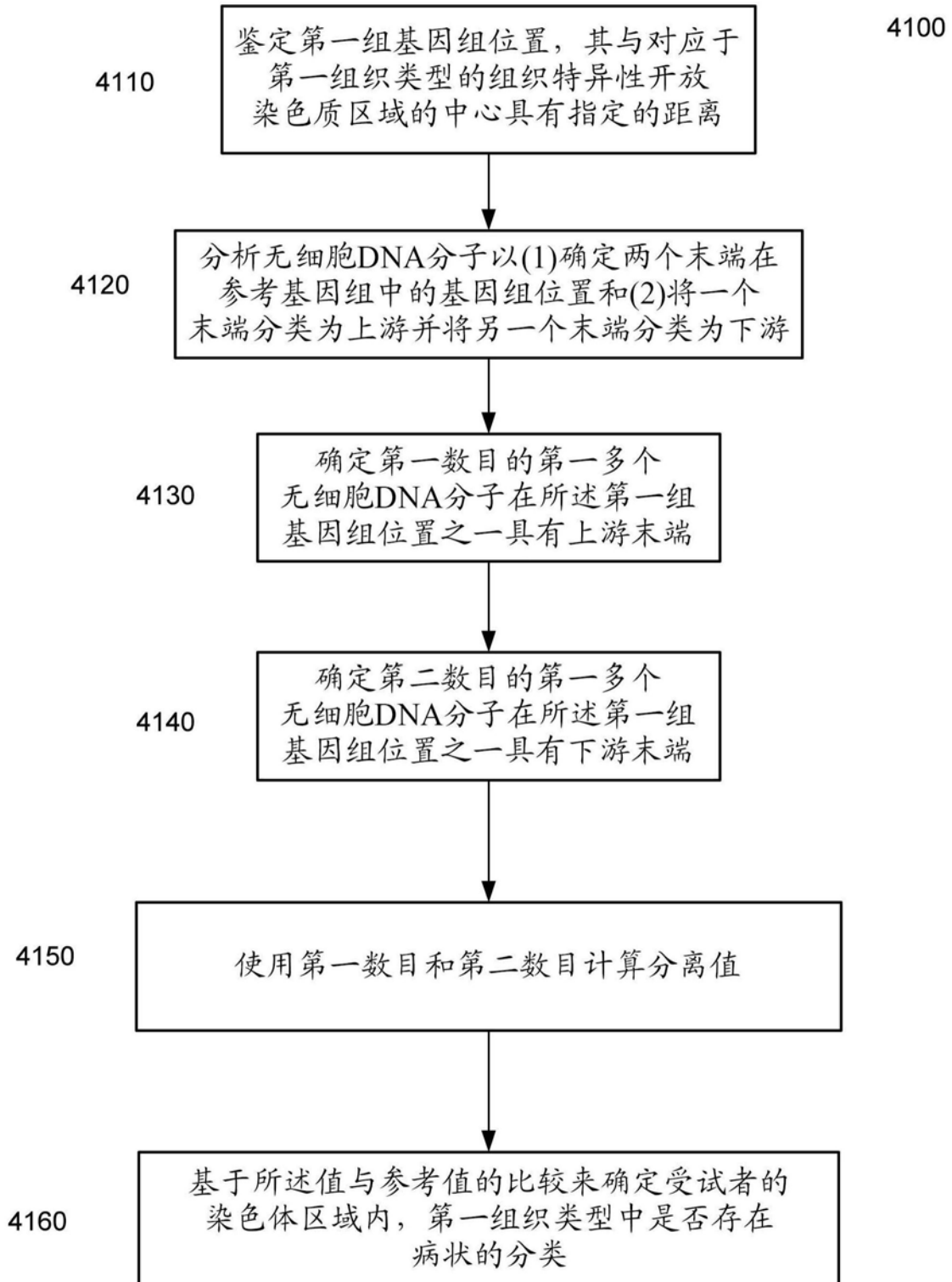


图41

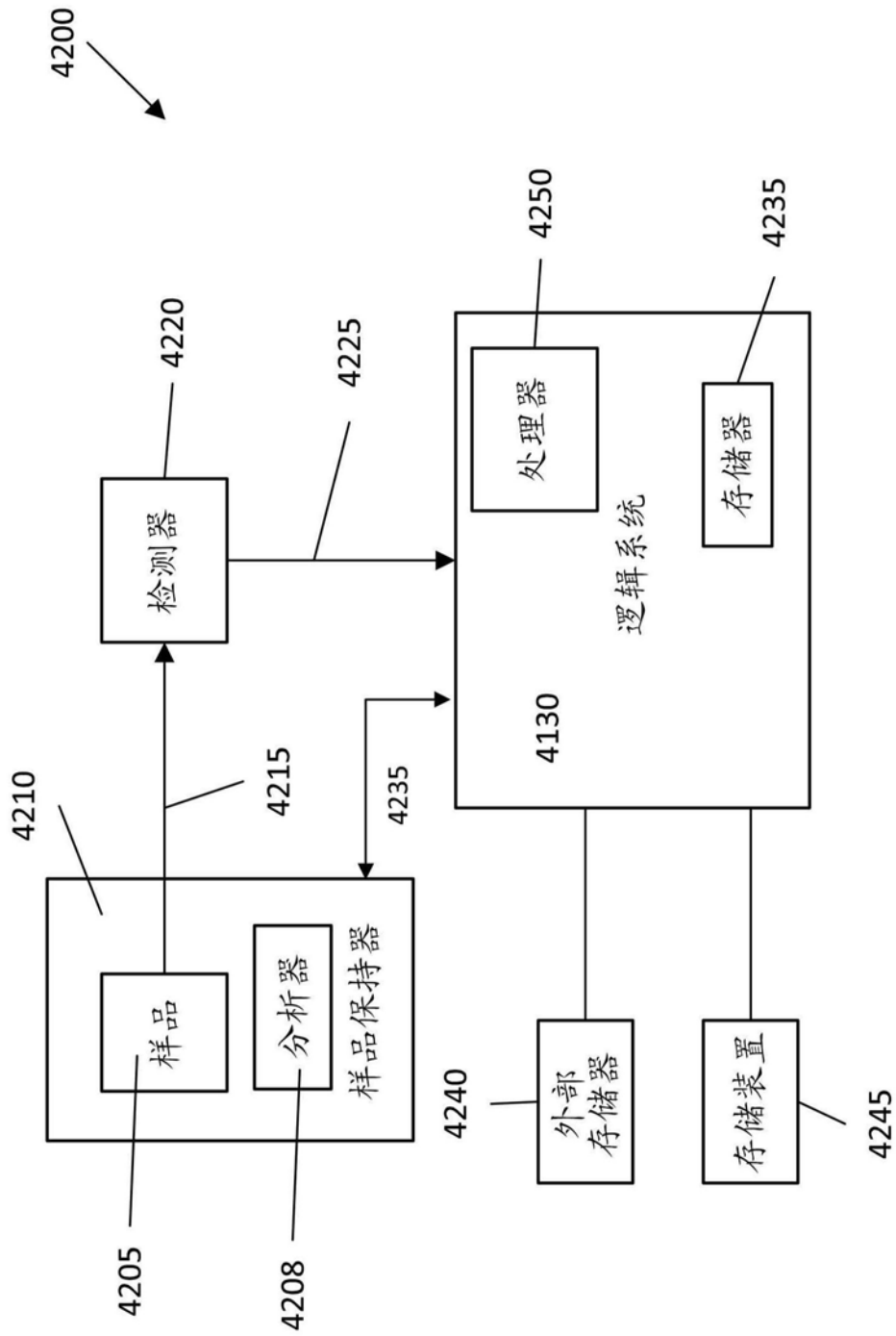


图42

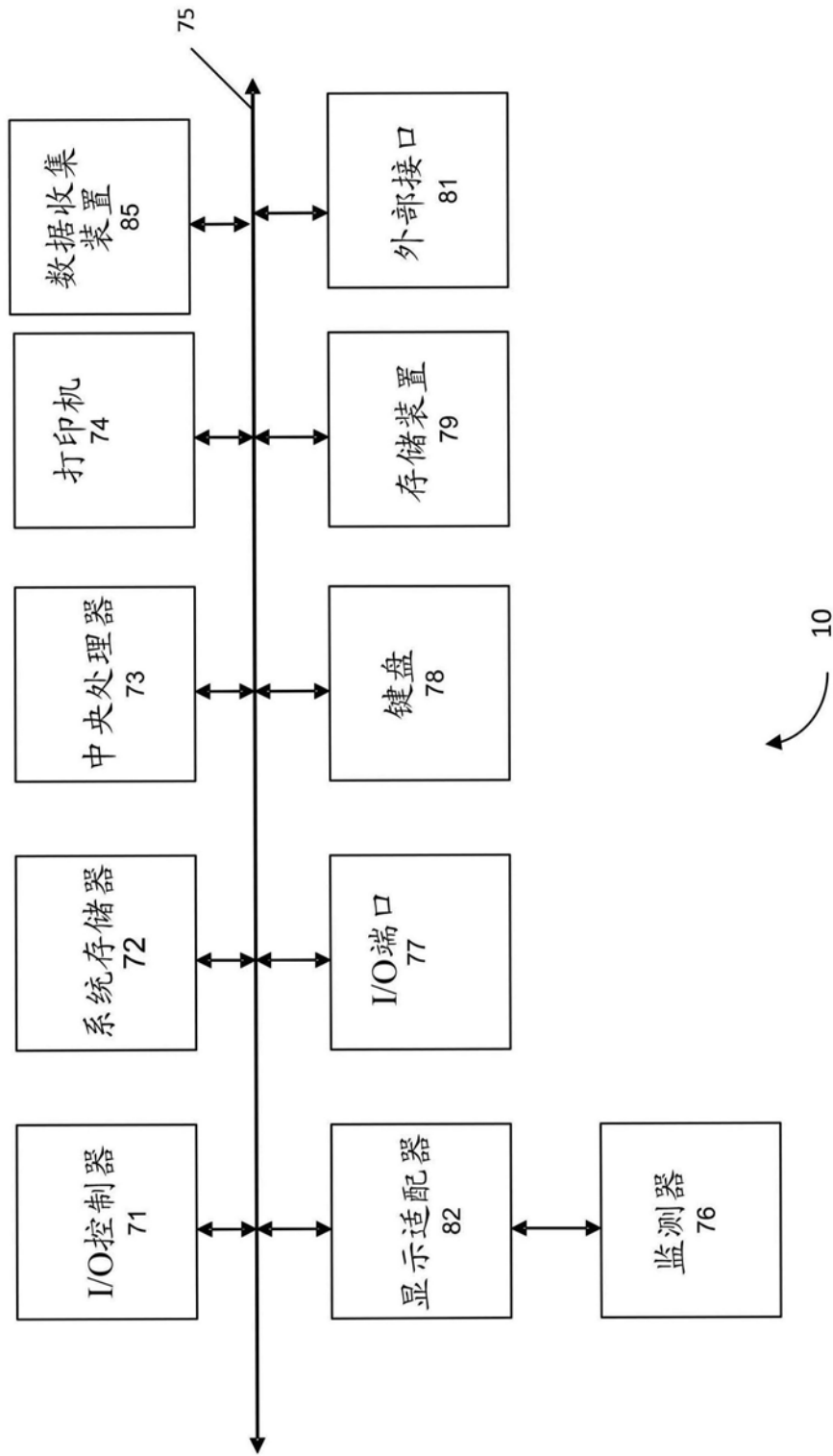


图43