(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109689891 B
(45)授权公告日 2024.06.18

(21)申请号 201780054825.3
(22)申请日 2017.07.06
(65)同一申请的已公布的文献号
申请公布号 CN 109689891 A
(43)申请公布日 2019.04.26
(30)优先权数据
62/359,151 2016.07.06 US
62/420,167 2016.11.10 US
62/437,172 2016.12.21 US
62/489,399 2017.04.24 US
(85)PCT国际申请进入国家阶段日
2019.03.06
(86)PCT国际申请的申请数据
PCT/US2017/040986 2017.07.06
(87)PCT国际申请的公布数据
WO2018/009723 EN 2018.01.11

(73)专利权人 夸登特健康公司
地址 美国加利福尼亚州
(72)发明人 戴安娜·阿布杜伊瓦
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 贺淑东

(51)Int.Cl.
C12Q 1/6883(2018.01)
G16B 20/00(2019.01)
G16B 30/00(2019.01)
G16B 40/00(2019.01)
(56)对比文件
WO 2016015058 A2,2016.01.28
CN 101901345 A,2010.12.01
US 2016040229 A1,2016.02.11
审查员 李莎莎
权利要求书3页 说明书57页 附图31页

(54)发明名称

用于无细胞核酸的片段组谱分析的方法

(57)摘要

本公开内容构思了无细胞DNA的各种用途。本文提供的方法可以在具有或没有体细胞变体信息的情况下以宏观和全局的方式使用序列信息,以评估可以代表起源组织、疾病、进展等的片段组谱。在一个方面,本文公开了一种用于确定来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的方法,该方法包括:(a)构建DNA片段跨基因组中多于一个碱基位置的多参数分布;以及(b)在不考虑第一基因座中每个碱基位置的碱基身份的情况下,使用多参数分布以确定受试者中第一基因座中遗传畸变的存在或不存在。

$$\mathrm{cfDNA} = \sum_{\substack{\text{组织}\\(\text{包括血液})}} \begin{bmatrix} \text{良性全身响应} \\ \text{肿瘤全身响应} \\ \text{肿瘤微环境} \\ \text{肿瘤} \end{bmatrix}$$

CN 109689891 B

1.一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时实现一种方法,所述方法用于确定来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中拷贝数变异的存在或不存在,所述方法包括:

(a)通过计算机构建所述DNA片段跨基因组中多于一个碱基位置的多参数分布,其中所述多参数分布包括指示以下参数:(i)与所述基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与所述基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在所述基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;并且

使用所述多参数分布来确定分布评分,其中所述分布评分指示所述拷贝数变异的突变负荷,其中所述分布评分包括指示具有双核小体保护的DNA片段的数量和具有单核小体保护的DNA片段的数量中的一个或更多个的值;以及

(b)在不考虑第一基因座中每个碱基位置的碱基身份的情况下,使用所述多参数分布来确定所述受试者中所述第一基因座中拷贝数变异的存在或不存在。

2.一种计算机可读介质,其存储有用于实现分类器的程序,所述分类器用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定所述测试受试者中的拷贝数变异,所述分类器包括:

(a)从多于一个受试者中的每一个获得的一个或更多个无细胞DNA群体中的每一个的一组分布评分的输入,其中每个分布评分至少基于以下生成:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量,并且其中所述分布评分包括指示具有双核小体保护的DNA片段的数量和具有单核小体保护的DNA片段的数量中的一个或更多个的值;以及

(b)所述测试受试者中一种或更多种拷贝数变异的分类的输出。

3.一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时实现一种方法,所述方法用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定所述测试受试者中的拷贝数变异,所述方法包括:

(a)提供计算机实施的分类器,所述分类器被配置为使用来自从测试受试者获得的无细胞DNA的DNA片段确定所述测试受试者中的拷贝数变异,所述分类器使用训练集训练;

(b)向所述分类器提供所述测试受试者的一组分布评分作为输入,其中每个分布评分指示:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量,其中所述分布评分包括指示具有双核小体保护的DNA片段的数量和具有单核小体保护的DNA片段的数量中的一个或更多个的值;以及

(c)通过计算机使用所述分类器生成所述测试受试者中拷贝数变异的分类。

4.一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时实现一种方法,所述方法用于分析源自受试者的无细胞脱氧核糖核酸(DNA)片段,所述方法包括:

对所述无细胞DNA片段进行文库制备和高通量测序,以生成代表来自所述受试者的所述无细胞DNA片段的序列信息,包括:

(i)用独特或非独特的分子标签将所述无细胞DNA片段加标签;

(ii)扩增加标签的无细胞DNA片段,和

(iii)通过追踪所述标签来追踪子代序列;以及

使用所述序列信息对多于一个数据集进行多参数分析以生成代表所述无细胞DNA片段的多参数模型,其中所述多参数模型包括三个或更多个维度,所述方法还包括生成分布评分,所述分布评分包括指示具有双核小体保护的DNA片段的数量或具有单核小体保护的DNA片段的数量的值。

5.一种生成分类器的方法,所述分类器用于确定受试者属于一个或更多个具有临床意义的类别的似然,所述方法包括:

a)提供训练集,所述训练集包括针对所述一个或更多个具有临床意义的类别中的每一个的来自属于所述具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体以及来自不属于所述具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体;

b)对来自所述无细胞DNA群体的无细胞DNA片段进行测序以产生多于一个DNA序列;

c)针对每个无细胞DNA群体,将所述多于一个DNA序列映射到所述物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;

d)针对每个无细胞DNA群体制备数据集以产生训练集,所述数据集包括针对多于一个所述遗传基因座中的每一个的指示选自以下的至少一个特性的定量量度的值:

(i)映射到所述遗传基因座的DNA序列,(ii)在所述基因座处起始的DNA序列,以及(iii)在所述遗传基因座处终止的DNA序列,

所述定量量度包括具有所选特征的DNA序列的大小分布,所述大小分布包括指示具有双核小体保护的DNA片段的数量和/或具有单核小体保护的DNA片段的数量的值;以及

e)针对所述训练集训练基于计算机的机器学习系统,从而生成用于确定所述受试者属于一个或更多个具有临床意义的类别的似然的分类器。

6.一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时实现一种确定受试者中异常生物学状态的方法,所述方法包括:

a)对来自所述受试者的无细胞DNA的无细胞DNA片段进行测序以产生DNA序列;

b)将所述DNA序列映射到所述受试者的物种的参考基因组中一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;

c)针对多于一个遗传基因座中的每一个制备数据集,所述数据集包括指示选自以下的至少一个特征的定量量度的值:

(i)映射到所述遗传基因座的DNA序列,(ii)在所述基因座处起始的DNA序列,以及(iii)在所述遗传基因座处终止的DNA序列,

所述定量量度包括具有所选特征的DNA序列的大小分布,所述大小分布包括指示具有双核小体保护的DNA片段的数量和/或具有单核小体保护的DNA片段的数量的值;以及

d)基于所述数据集,确定所述异常生物学状态的似然。

7.一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时实现一种方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中拷贝数变异的存在或不存在的输出,所述方法包括:

(a)通过计算机,基于包括以下的参数构建来自所述无细胞DNA的所述DNA片段跨基因组中多于一个碱基位置的分布:所述无细胞DNA的所述DNA片段的长度和起始位置;以及

(b)通过计算机计算一个或更多个遗传基因座中的每一个的指示以下的定量量度:(1)具有与来自所述一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段的数量,与(2)具有与所述遗传基因座相关的单核小体保护的DNA片段的数量的比率,或反之亦然;以及

(c)使用所述一个或更多个遗传基因座中的每一个的所述定量量度来确定指示所述受试者中所述一个或更多个遗传基因座中拷贝数变异的存在或不存在的所述输出。

8.权利要求7所述的计算机可读介质,所述参数还包括:测序的DNA片段的终止位置和DNA片段覆盖率的一种或两种。

9.根据权利要求7所述的计算机可读介质,其中至少基于指示以下的定量量度来确定指示所述拷贝数变异的存在或不存在的所述输出:与双核小体峰相关的第一峰值和与单核小体峰相关的第二峰值的比率,或反之亦然。

用于无细胞核酸的片段组谱分析的方法

[0001] 交叉引用

[0002] 本申请要求2016年7月6日提交的美国临时申请第62/359,151号、2016年11月10日提交的美国临时申请第62/420,167号、2016年12月21日提交的美国临时申请第62/437,172号和2017年4月24日提交的美国临时申请第62/489,399号的优先权,其每一个通过引用整体并入本文。

[0003] 背景

[0004] 无细胞核酸(例如,DNA或RNA)的癌症诊断测定的当前方法集中于检测肿瘤相关的体细胞变体,包括单核苷酸变体(SNV)、拷贝数变异(CNV)、融合和插入/缺失(indel)(即,插入或缺失),这些都是液体活检的主流靶标。越来越多的证据表明,由于核小体定位而出现的新型结构变体可以被鉴定和测量以获得肿瘤相关信息,所述肿瘤相关信息当与体细胞突变调用组合时,可以对肿瘤状态做出比单独使用任何一种方法可得的更为全面的评估。通过分析受染色质组织影响的核酸片段分布的潜在非随机模式,可以以独立于体细胞变体的方式在样品中观察到这组新的结构变体,并且甚至事实上在未检测到体细胞变体的样品中被观察到。

[0005] 概述

[0006] 核小体定位(nucleosome positioning)是有助于基因表达的表观遗传控制的关键机制,是高度组织特异性的,并且指示各种表型状态。本公开内容描述了用于使用无细胞核酸(例如cfDNA)进行核小体谱分析(nucleosome profiling)的方法、系统和组合物。这可用于鉴定新的驱动基因,确定拷贝数变异(CNV),鉴定体细胞突变和结构变异,诸如融合和插入/缺失,以及鉴定可用于多重测定以检测上述变异中的任一种的区域。

[0007] 本公开内容提供了无细胞核酸(例如,DNA或RNA)的各种用途。这些用途包括对患有或怀疑患有健康状况,诸如疾病(例如癌症)的受试者进行检测、监测和确定用于所述受试者的治疗。本文提供的方法可以在具有或没有体细胞变体信息的情况下以宏观和全局方式使用序列信息,以评估可以代表起源组织(tissue of origin)、疾病、进展等的片段组谱(fragmentome profile)。

[0008] 在一个方面,本文公开了一种计算机实施的方法,所述方法用于确定来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变(genetic aberration)的存在或不存在,该方法包括:(a)通过计算机构建DNA片段跨基因组中多于一个碱基位置的多参数分布(multi-parametric distribution);以及(b)在不考虑第一基因座中每个碱基位置的碱基身份的情况下,使用多参数分布来确定受试者中第一基因座中遗传畸变的存在或不存在。

[0009] 在一些实施方案中,遗传畸变包括序列畸变。在一些实施方案中,序列畸变包括单核苷酸变体(SNV)。在一些实施方案中,序列畸变包括插入或缺失(插入/缺失)或基因融合。在一些实施方案中,序列畸变包括选自由以下组成的组的两种或更多种不同成员:(i)单核苷酸变体(SNV),(ii)插入或缺失(插入/缺失),和(iii)基因融合。在一些实施方案中,遗传畸变包括拷贝数变异(CNV)。

[0010] 在一些实施方案中,多参数分布包括指示与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度的参数。在一些实施方案中,多参数分布包括指示与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量的参数。在一些实施方案中,多参数分布包括指示在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量的参数。在一些实施方案中,多参数分布包括指示以下中的两种或更多种的参数:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量。在一些实施方案中,多参数分布包括指示以下的参数:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量。

[0011] 在一些实施方案中,使用分布包括通过计算机将多参数分布应用于分类器,所述分类器具有DNA片段跨基因组中多于一个碱基位置的多于一个其他多参数分布的输入,该其他多参数分布从选自以下的组获得:(a)患有组织特异性癌症的受试者,(b)患有特定阶段的癌症的受试者,(c)患有炎性状况的受试者,(d)对于癌症是无症状的但患有将进展为癌症的肿瘤的受试者,和(e)对疗法具有正面或负面响应的受试者。

[0012] 在一些实施方案中,分类器包括机器学习引擎。在一些实施方案中,分类器还包括在基因组的一个或更多个基因座处的一组遗传变体的输入。在一些实施方案中,该组遗传变体包含被报告的肿瘤标志物的一个或更多个基因座。

[0013] 在一些实施方案中,方法还包括使用多参数分布来确定分布评分。在一些实施方案中,分布评分指示遗传畸变的突变负荷。在一些实施方案中,分布评分包括指示具有双核小体保护的DNA片段的数量和具有单核小体保护的DNA片段的数量中的一个或更多个的值。

[0014] 在一些实施方案中,方法还包括使用多参数分布来估计多模态密度(multimodal density),以及使用多模态密度来确定遗传畸变的存在或不存在。在一些实施方案中,使用多模态密度包括从多模态密度生成判别评分,并将判别评分与截止值进行比较以确定遗传畸变的存在或不存在。在一些实施方案中,方法还包括通过计算残差密度估计来估计与遗传畸变相关的基因的表达。在一些实施方案中,方法还包括通过计算单核小体中的残差密度来估计与遗传畸变相关的基因的拷贝数。

[0015] 在另一方面,本文公开了一种计算机实施的分类器,所述分类器用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定测试受试者中的遗传畸变,所述分类器包括:(a)从多于一个受试者中的每一个获得的一个或更多个无细胞DNA群体中的每一个的一组分布评分的输入,其中每个分布评分至少基于以下中的一个或更多个生成:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;以及(b)测试受试者中一种或更多种遗传畸变的分类的输出。

[0016] 在一些实施方案中,分类器还包括机器学习引擎。在一些实施方案中,分类器还包括在基因组的一个或更多个基因座处的一组遗传变体的输入。在一些实施方案中,该组遗传变体包含被报告的肿瘤标志物的一个或更多个基因座。

[0017] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定测试受试者中的遗传畸变,该方法包括:(a)提供计算机实施的分类器,该分类器被配置为使用来自从测试受试者获得的无细胞DNA的DNA片段来确定测试受试者中的遗传畸变,该分类器使用训练集训练;(b)向分类器提供测试受试者的一组分布评分作为输入,其中每个分布评分指示以下中的一个或更多个:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;以及(c)通过计算机使用分类器生成测试受试者中的遗传畸变的分类。

[0018] 在一些实施方案中,方法还包括在(a)之前执行:(i)提供训练集,所述训练集包括:(1)来自多于一个对照受试者中的每一个的一个或更多个无细胞DNA群体中的每一个的一组参考分布评分,其中每个参考分布评分指示以下中的一个或更多个:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;(2)来自具有观察到的表型的多于一个受试者中的每一个的一个或更多个无细胞DNA群体中的每一个的一组表型分布评分,其中每个表型分布评分指示以下中的一个或更多个:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;(3)从对照受试者获得的每个无细胞DNA群体的一组参考分类;(4)从具有观察到的表型的受试者获得的每个无细胞DNA群体的一组表型分类;以及(ii)通过计算机使用训练集训练分类器。

[0019] 在一些实施方案中,对照受试者包括无症状的健康个体。在一些实施方案中,具有观察到的表型的受试者包括(a)患有组织特异性癌症的受试者,(b)患有特定阶段的癌症的受试者,(c)患有炎性状况的受试者,(d)对于癌症是无症状的但患有将进展为癌症的肿瘤的受试者,或者(e)对疗法具有正面或负面响应的患有癌症的受试者。

[0020] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于分析源自受试者的无细胞脱氧核糖核酸(DNA)片段,该方法包括:获得代表无细胞DNA片段的序列信息;以及使用序列信息对多于一个数据集进行多参数分析以生成代表无细胞DNA片段的多参数模型,其中该多参数模型包括三个或更多个维度。

[0021] 在一些实施方案中,数据集选自由以下组成的组:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段的末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,(g)测序的DNA片段的序列基序,(h)GC含量,(i)测序的DNA片段长度分布,以及(j)甲基化状态。在一些实施方案中,序列基序是位于DNA片段的末端的长度为2至8个碱基对的序列。在一些实施方案中,多参数分析包括将选自由以下组成的组的一个或更多个分布映射到基因组的多于一个碱基位置或区域中的每一个:(i)含有覆盖基因组中可映射位置的序列的独特无细胞DNA片段的数量的分布,(ii)无细胞DNA片段的至少一些中的每一个的片段长度的分布,使得DNA片段含有覆盖基因组中可映射位置的序列,以及(iii)可映射碱基对位置将出现在测序的

DNA片段末端处的似然分布。在一些实施方案中,基因组的多于一个碱基位置或区域包括与表1中列出的一种或更多种基因相关的至少一个碱基位置或区域。在一些实施方案中,基因组的多于一个碱基位置或区域中的每一个的长度在2和500个碱基对之间。在一些实施方案中,通过以下鉴定基因组的多于一个碱基位置或区域:(i)提供一个或更多个基因组分区图谱,以及(ii)从基因组分区图谱选择基因组的多于一个碱基位置或区域,基因组的每个碱基位置或区域映射到目的基因。在一些实施方案中,映射包括将来自多于一个数据集中的每一个的多于一个值映射到基因组的多于一个碱基位置或区域中的每一个。在一些实施方案中,多于一个值中的至少一个是选自由以下组成的组的数据集:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,或者(g)测序的DNA片段的序列基序。

[0022] 在一些实施方案中,多参数分析包括通过计算机应用一个或更多个数学变换以生成多参数模型。在一些实施方案中,数学变换包括分水岭变换。在一些实施方案中,多参数模型是选自由以下组成的组的多于一个变量的联合分布模型:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,以及(g)测序的DNA片段的序列基序。

[0023] 在一些实施方案中,方法还包括在多参数模型中鉴定一个或更多个峰,每个峰具有峰分布宽度和峰覆盖率。在一些实施方案中,方法还包括掺入由受试者中存在的种系或体细胞单核苷酸多态性引起的变异性。在一些实施方案中,方法还包括检测代表无细胞DNA片段的多参数模型与参考多参数模型之间的一个或更多个偏差。在一些实施方案中,偏差选自由以下组成的组:(i)核小体区域外的读段数量的增加,(ii)核小体区域内的读段数量的增加,(iii)相对于可映射基因组位置更宽的峰分布,(iv)峰位置的偏移,(v)新峰的鉴定,(vi)峰的覆盖深度的变化,(vii)峰周围的起始位置的变化,以及(viii)与峰相关的片段大小的变化。在一些实施方案中,参考多参数模型源自健康的无症状个体。在一些实施方案中,参考多参数模型源自不同时间点的受试者。

[0024] 在一些实施方案中,参考多参数模型源自从来自受试者的肿瘤周围微环境的基质组织获得的DNA。在一些实施方案中,参考多参数模型源自来自健康无症状个体的经剪切的基因组DNA。在一些实施方案中,参考多参数模型源自给定组织类型的核小体占位谱。在一些实施方案中,组织类型是选自由以下组成的组的正常组织:乳腺、结肠、肺、胰腺、前列腺、卵巢、皮肤和肝脏。在一些实施方案中,参考多参数模型源自具有共有特性的个体群组(cohort)。在一些实施方案中,共有特性选自由以下组成的组:肿瘤类型、炎性状况、凋亡状况、坏死状况、肿瘤复发和对治疗的耐受性。在一些实施方案中,凋亡状况选自由以下组成的组:感染和细胞周转。在一些实施方案中,坏死状况选自由以下组成的组:心血管状况、败血症和坏疽。

[0025] 在一些实施方案中,方法还包括确定归因于无细胞DNA起源的细胞中的凋亡过程的多参数模型的贡献。在一些实施方案中,方法还包括确定归因于无细胞DNA起源的细胞中

的坏死过程的多参数模型的贡献。在一些实施方案中,方法还包括对来自受试者的身体样品进行以下测定中的一种或更多种:(i)起源组织分析,(ii)基因表达分析,(iii)转录因子结合位点(TFBS)占据分析,(iv)甲基化状态分析,(v)体细胞突变检测,(vi)可检测体细胞突变水平的测量,(vii)种系突变检测,以及(viii)可检测种系突变水平的测量。

[0026] 在一些实施方案中,方法还包括进行多参数分析以测量无细胞DNA片段的RNA表达。在一些实施方案中,方法还包括进行多参数分析以测量无细胞DNA片段的反向甲基化(reverse methylation)。在一些实施方案中,方法还包括进行多参数分析以测量无细胞DNA片段的反向核小体映射。在一些实施方案中,方法还包括进行多参数分析以鉴定无细胞DNA片段中一个或更多个体细胞单核苷酸多态性的存在。在一些实施方案中,方法还包括进行多参数分析以鉴定无细胞DNA片段中一个或更多个种系单核苷酸多态性的存在。在一些实施方案中,方法还包括生成分布评分,该分布评分包括指示具有双核小体保护的DNA片段的数量和/或具有单核小体保护的DNA片段的数量的值。在一些实施方案中,方法还包括估计受试者的突变负荷。在一些实施方案中,方法还包括估计多模态密度,并使用多模态密度来鉴定无细胞DNA片段中一种或更多种遗传畸变的存在。在一些实施方案中,方法还包括映射规范核小体架构。在一些实施方案中,映射包括执行二元正态混合物的拓扑建模(topographic modeling)。

[0027] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于分析源自受试者的无细胞脱氧核糖核酸(DNA)片段,该方法包括:获得代表无细胞DNA片段的多参数模型;并且利用计算机进行统计分析以将多参数模型分类为与代表不同群组的一个或更多个核小体占位谱相关。

[0028] 在一些实施方案中,统计分析包括提供一个或更多个基因组分区图谱,所述基因组分区图谱列出代表目的基因的相关基因组间隔以用于进一步分析。在一些实施方案中,统计分析还包括基于基因组分区图谱选择一个或更多个局部基因组区域(localized genomic region)的组。在一些实施方案中,统计分析还包括分析该组中的一个或更多个局部基因组区域以获得一个或更多个核小体图谱中断(nucleosomal map disruption)的组。在一些实施方案中,统计分析包括以下中的一个或更多个:模式识别、深度学习和无监督学习。在一些实施方案中,通过以下构建基因组分区图谱:(a)提供来自群组中的两个或更多个受试者的无细胞DNA群体;(b)对每个无细胞DNA群体进行多参数分析,以生成每个样品的多参数模型;以及(c)分析多参数模型以鉴定一个或更多个局部基因组区域。在一些实施方案中,其中核小体图谱中断中的至少一个与驱动突变(driver mutation)相关,其中驱动突变选自由以下组成的组:体细胞变体、种系变体和DNA甲基化。在一些实施方案中,核小体图谱中断中的至少一个被用于将多参数模型分类为与代表不同群组的一个或更多个核小体占位谱相关。

[0029] 在一些实施方案中,局部基因组区域中的至少一个是范围为约2至约200个碱基对的短DNA区域,其中该区域含有显著结构变异的模式。在一些实施方案中,局部基因组区域中的至少一个是范围为约2至约200个碱基对的短DNA区域,其中该区域含有显著结构变异的簇。在一些实施方案中,结构变异是选自由以下组成的组的核小体定位的变异:插入、缺失、易位、基因重排、甲基化状态、微卫星、拷贝数变异、拷贝数相关的结构变异、或指示差异的任何其他变异。在一些实施方案中,簇是局部基因组区域内的热点区域,其中热点区域包

含一个或更多个显著波动或峰。在一些实施方案中,局部基因组区域中的至少一个是范围为约2至约200个碱基对的短DNA区域,其中该区域含有显著不稳定性的模式。在一些实施方案中,分析一个或更多个局部基因组区域包括检测代表无细胞DNA片段的多参数模型与选自以下的一个或更多个参考多参数模型之间的一个或更多个偏差:(i)与健康对照的一个或更多个群组相关的一个或更多个健康参考多参数模型,以及(ii)与患病受试者的一个或更多个群组相关的一个或更多个患病参考多参数模型。

[0030] 在一些实施方案中,方法还包括选择一组结构变异,其中结构变异的选择是以下中的一个或更多个的函数:(i)一个或更多个健康参考多参数模型;(ii)靶向结构变异的一种或更多种探针的效率;(iii)关于基因组中结构变异的预期频率高于跨基因组的结构变异的平均预期频率的部分的先前信息。

[0031] 在一些实施方案中,核小体占位谱中的至少一个与选自由以下组成的组的一种或更多种评估相关:肿瘤适应症、癌症的早期检测、肿瘤类型、肿瘤严重性、肿瘤侵袭性、肿瘤对治疗的耐受性、肿瘤克隆性、肿瘤可药性、肿瘤进展和血浆失调评分(plasma dysregulation score)。在一些实施方案中,通过观察样品中跨无细胞DNA片段的核小体图谱中断的异质性来确定肿瘤克隆性的评估。在一些实施方案中,确定对两个或更多个克隆中的每一个的相对贡献的评估。

[0032] 在一些实施方案中,方法还包括确定疾病的疾病评分,其中根据以下中的一项或更多项确定疾病评分:(i)与疾病相关的一种或更多种核小体占位谱;(ii)与未患有该疾病的群组相关的一个或更多个健康参考多参数模型;(iii)与患有该疾病的群组相关的一个或更多个患病参考多参数模型。

[0033] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于创建经训练的分类器,包括:(a)提供多于一个不同的类别,其中每个类别代表具有共有特性的一组受试者;(b)针对从每个类别获得的多于一个无细胞DNA群体中的每一个,提供代表来自无细胞DNA群体的无细胞脱氧核糖核酸(DNA)片段的多参数模型,从而提供训练数据集;以及(c)通过计算机针对训练数据集训练学习算法以创建一个或更多个经训练的分类器,其中每个经训练的分类器被配置为将来自测试受试者的无细胞DNA的测试群体分类为多于一个不同类别中的一个或更多个。

[0034] 在一些实施方案中,学习算法选自由以下组成的组:随机森林、神经网络、支持向量机和线性分类器。在一些实施方案中,多于一个不同类别的每一个选自由以下组成的组:健康、乳腺癌、结肠癌、肺癌、胰腺癌、前列腺癌、卵巢癌、黑素瘤和肝癌。

[0035] 在一个方面,本文公开了一种对来自受试者的测试样品进行分类的方法,所述方法包括:(a)提供代表来自受试者的无细胞DNA测试群体的无细胞脱氧核糖核酸(DNA)片段的多参数模型;以及(b)使用经训练的分类器对无细胞DNA测试群体进行分类。

[0036] 在一些实施方案中,方法还包括基于无细胞DNA群体的分类对受试者进行治疗性干预。

[0037] 在另一方面,本文公开了一种计算机实施的方法,所述方法包括:(a)通过计算机生成来自受试者的无细胞DNA片段的序列信息;(b)通过计算机基于序列信息将无细胞DNA片段映射到参考基因组;以及(c)通过计算机分析映射的无细胞DNA片段,以确定参考基因组中多于一个碱基位置的每一个处的选自由以下组成的组的多于一种量度:(i)映射到碱

基位置的无细胞DNA片段的数量,(ii)映射到碱基位置的每个无细胞DNA片段的长度,(iii)映射到碱基位置的无细胞DNA片段的数量随无细胞DNA片段的长度的变化;(iv)在碱基位置处起始的无细胞DNA片段的数量;(v)在碱基位置处终止的无细胞DNA片段的数量;(vi)在碱基位置处起始的无细胞DNA片段的数量随长度的变化,以及(vii)在碱基位置处终止的无细胞DNA片段的数量随长度的变化。在一些实施方案中,序列信息是无细胞DNA片段的完整或部分序列。

[0038] 在另一方面,本文公开了一种分析源自受试者的无细胞DNA片段的计算机实施的方法,该方法包括:(a)通过计算机接收代表无细胞DNA片段的序列信息,以及(b)进行每个可映射碱基位置或基因组位置的包括以下的多于一个的分析:(i)在碱基位置或基因组位置处起始或终止的序列片段的数量,(ii)碱基位置或基因组位置处的序列或片段长度,(iii)碱基位置或基因组位置处的片段或序列覆盖率,以及(iv)碱基位置或基因组位置处的序列基序分布。

[0039] 在一些实施方案中,方法还包括检测来自受试者的无细胞DNA与一个或更多个无细胞DNA参考群体之间的偏差,其中该偏差指示受试者中状况或属性的存在。在一些实施方案中,分析包括由以下组成的组中的一种或更多种:(i)起源组织分析,(ii)基因表达分析,(iii)转录因子结合位点(TFBS)占据分析,(iv)甲基化状态分析,(v)体细胞突变检测,(vi)可检测体细胞突变水平的测量,(vii)种系突变检测,以及(viii)可检测种系突变水平的测量。

[0040] 在一些实施方案中,状况或属性是由以下组成的组中的一种或更多种:(i)癌症的存在,(ii)组织异常的存在,(iii)特定组织特异性异常的存在,(iv)表观遗传调节或功能的变异的存在,以及(v)表观遗传调节或功能的变异的存在。在一些实施方案中,分析还包括检测由以下组成的组中的一种或更多种:(i)单核苷酸变体,(ii)拷贝数变体,(iii)插入,(iv)缺失,(v)基因重排,(vi)甲基化状态,以及(vii)杂合性丢失。

[0041] 在另一方面,本文公开了一种生成分类器的方法,该分类器用于确定受试者属于一个或更多个具有临床意义的类别的似然,所述方法包括:a)提供训练集,该训练集包括针对一个或更多个具有临床意义的类别中的每一个的来自属于具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体以及来自不属于具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体;b)对来自无细胞DNA群体的无细胞DNA片段进行测序以产生多于一个DNA序列;c)针对每个无细胞DNA群体,将多于一个DNA序列映射到物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;d)针对每个无细胞DNA群体制备数据集以产生训练集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一种特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列;以及e)针对训练集训练基于计算机的机器学习系统,从而生成用于确定受试者属于一个或更多个具有临床意义的类别的似然的分类器。

[0042] 在一些实施方案中,具有临床意义的类别指示一种或更多种遗传变体的存在或不存在。在一些实施方案中,具有临床意义的类别指示一种或更多种癌症的存在或不存在。在一些实施方案中,具有临床意义的类别指示一种或更多种非癌症疾病、紊乱或异常生物学状态的存在或不存在。在一些实施方案中,具有临床意义的类别指示一种或更多种规范驱

动突变的存在或不存在。在一些实施方案中,具有临床意义的类别指示一种或更多种癌症亚型的存在或不存在。在一些实施方案中,具有临床意义的类别指示对癌症治疗的响应的似然。在一些实施方案中,具有临床意义的类别指示拷贝数变异(CNV)的存在或不存在。在一些实施方案中,具有临床意义的类别指示起源组织。在一些实施方案中,定量量度包括具有所选特性的DNA序列的大小分布。

[0043] 在另一方面,本文公开了一种确定受试者中异常生物学状态的方法,该方法包括:a)对来自受试者的无细胞DNA的无细胞DNA片段进行测序以产生DNA序列;b)将DNA序列映射到受试者物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;c)制备数据集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一个特征的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列;以及d)基于数据集,确定异常生物学状态的似然。

[0044] 在一些实施方案中,参考基因组包含人类的参考基因组。在一些实施方案中,定量量度包括具有所选特征的DNA序列的大小分布。在一些实施方案中,大小分布包括指示具有双核小体保护的DNA片段的数量和/或具有单核小体保护的DNA片段的数量的值。在一些实施方案中,定量量度还包括具有所选特征的DNA序列的大小分布的比率。在一些实施方案中,数据集还包含指示针对多于一个遗传基因座的在内含子或外显子中的位置的值。在一些实施方案中,定量量度是归一化量度。在一些实施方案中,确定异常状态包括确定异常程度。在一些实施方案中,方法还包括施用治疗性干预以治疗异常生物学状态。

[0045] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中的多于一个碱基位置的分布;以及(b)针对一个或更多个遗传基因座中的每一个,通过计算机计算指示以下的定量量度:(1)具有与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段的数量,与(2)具有与遗传基因座相关的单核小体保护的DNA片段的的数量的比率,或反之亦然;以及(c)使用一个或更多个遗传基因座中的每一个的定量量度来确定指示受试者中一个或更多个遗传基因座中遗传畸变的存在或不存在的所述输出。在一些实施方案中,分布包括一个或更多个多参数分布。

[0046] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;以及(b)使用该分布确定指示受试者中遗传畸变的存在或不存在的所述输出,其中所述存在或不存在是在(i)不将DNA片段的分布与来自受试者基因组外部来源的参考分布进行比较,(ii)不将源自DNA片段的分布的参数与参考参数进行比较,以及(iii)不将DNA片段的分布与来自受试者的对照的参考分布进行比较的情况下确定的。

[0047] 在一些实施方案中,遗传畸变包括拷贝数变异(CNV)。在一些实施方案中,遗传畸变包括单核苷酸变体(SNV)。在一些实施方案中,分布包括一个或更多个多参数分布。

[0048] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于对来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段的分布进行去卷积,该方法包括:(a)通过计算

机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的覆盖率的分布;以及(b)针对一个或更多个遗传基因座中的每一个,通过计算机对覆盖率的分布进行去卷积,从而生成与选自由拷贝数(CN)组分、细胞清除组分和基因表达组分组成的组的一个或更多个成员相关的分数贡献。

[0049] 在一些实施方案中,计算包括计算与选自由拷贝数(CN)组分、细胞清除组分和基因表达组分组成的组的两个或更多个成员相关的DNA片段覆盖率的分布的分数贡献。在一些实施方案中,计算包括计算与拷贝数组分、清除组分和表达组分相关的DNA片段覆盖率的分布的分数贡献。

[0050] 在一些实施方案中,方法还包括至少基于分数贡献的一部分生成指示遗传畸变的存在或不存在的输出。在一些实施方案中,分布包括一个或更多个多参数分布。

[0051] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;(b)通过计算机鉴定DNA片段的分布中在多于一个碱基位置的一个或更多个碱基位置处的一个或更多个峰,其中每个峰包含峰值和峰分布宽度;以及(c)通过计算机至少基于(i)一个或更多个碱基位置、(ii)峰值和(iii)峰分布宽度确定受试者中遗传畸变的存在或不存在。

[0052] 在一些实施方案中,一个或更多个峰包含双核小体峰或单核小体峰。在一些实施方案中,一个或更多个峰包含双核小体峰和单核小体峰。在一些实施方案中,至少基于指示以下的定量量度来确定指示遗传畸变的存在或不存在的所述输出:与双核小体峰相关的第一峰值和与单核小体峰相关的第二峰值的比率,或反之亦然。在一些实施方案中,分布包括一个或更多个多参数分布。

[0053] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;(b)通过计算机分析DNA片段在一个或更多个遗传基因座处的分布,该分析包括检测DNA片段的分布与选自以下的多于一个参考分布之间的偏差:(i)与健康对照的一个或更多个群组相关的一个或更多个健康参考分布,以及(ii)与患病受试者的一个或更多个群组相关的一个或更多个患病参考分布;以及(c)通过计算机至少基于在(b)中检测到的偏差确定指示受试者中遗传畸变的存在或不存在的所述输出。

[0054] 在一些实施方案中,分布包括一个或更多个多参数分布。在一些实施方案中,分析包括计算一个或更多个差量信号(delta signal),每个差量信号包括DNA片段的分布与多于一个参考分布的参考分布之间的差异。

[0055] 在另一方面,本文公开了一种用于处理受试者的生物样品的方法,所述方法包括:(a)获得所述受试者的所述生物样品,其中所述生物样品包含脱氧核糖核酸(DNA)片段;(b)测定所述生物样品以生成指示具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与遗传基因座相关的单核小体保护的DNA片段的存在或不存在的信号;以及(c)使用所述信号生成指示具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与遗传基因座相关的单核小体保护的所述DNA片段的

存在或不存在的输出。

[0056] 在一些实施方案中,测定包括富集所述生物样品以获得一个或更多个遗传基因座的组的DNA片段。在一些实施方案中,测定包括将所述生物样品的所述DNA片段测序。

[0057] 在另一方面,本文公开了一种用于分析生物样品的方法,所述生物样品包含源自受试者的无细胞DNA片段,其中该方法包括检测来自相同遗传基因座的对应于单核小体保护和双核小体中的每一个的DNA片段。

[0058] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于确定来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在,该方法包括:(a)通过计算机构建DNA片段跨基因组中多于一个碱基位置的多参数分布;以及(b)在不考虑第一基因座中每个碱基位置的碱基身份的情况下,使用多参数分布来确定受试者中第一基因座中遗传畸变的存在或不存在。

[0059] 在一些实施方案中,遗传畸变包括序列畸变或拷贝数变异(CNV),其中序列畸变选自由以下组成的组:(i)单核苷酸变体(SNV),(ii)插入或缺失(插入/缺失),和(iii)基因融合。在一些实施方案中,多参数分布包括指示以下中的一种或更多种的参数:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量。在一些实施方案中,方法还包括使用多参数分布来确定分布评分,其中该分布评分指示遗传畸变的突变负荷。在一些实施方案中,分布评分包括指示具有双核小体保护的DNA片段的数量和具有单核小体保护的DNA片段的数量中的一个或更多个的值。

[0060] 在另一方面,本文公开了一种计算机实施的分类器,所述分类器用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定测试受试者中的遗传畸变,所述分类器包括:(a)从多于一个受试者中的每一个获得的一个或更多个无细胞DNA群体中的每一个的一组分布评分的输入,其中每个分布评分至少基于以下中的一个或更多个生成:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;以及(b)测试受试者中一种或更多种遗传畸变的分类的输出。

[0061] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于使用来自从测试受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段确定测试受试者中的遗传畸变,该方法包括:(a)提供计算机实施的分类器,该分类器被配置为使用来自从测试受试者获得的无细胞DNA的DNA片段来确定测试受试者中的遗传畸变,该分类器使用训练集训练;(b)向分类器提供测试受试者的一组分布评分作为输入,其中每个分布评分指示以下中的一个或更多个:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量;以及(c)通过计算机使用分类器生成测试受试者中的遗传畸变的分类。

[0062] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于分析源自受试者的无细胞脱氧核糖核酸(DNA)片段,该方法包括:获得代表无细胞DNA片段的序列信息;以及

使用序列信息对多于一个数据集进行多参数分析以生成代表无细胞DNA片段的多参数模型,其中多参数模型包括三个或更多个维度。

[0063] 在一些实施方案中,数据集选自由以下组成的组:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段的末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,(g)测序的DNA片段的序列基序,(h)GC含量,(i)测序的DNA片段长度分布,和(j)甲基化状态。在一些实施方案中,多参数分析包括将选自由以下组成的组的一个或更多个分布映射到基因组的多于一个碱基位置或区域中的每一个:(i)含有覆盖基因组中可映射位置的序列的独特无细胞DNA片段的数量的分布,(ii)无细胞DNA片段的至少一些中的每一个的片段长度的分布,使得DNA片段含有覆盖基因组中可映射位置的序列,以及(iii)可映射碱基对位置将出现在测序的DNA片段末端处的似然分布。在一些实施方案中,基因组的多于一个碱基位置或区域包括与表1中列出的一种或更多种基因相关的至少一个碱基位置或区域。在一些实施方案中,映射包括将来自多于一个数据集中的每一个的多于一个值映射到基因组的多于一个碱基位置或区域中的每一个。在一些实施方案中,多于一个值中的至少一个是选自由以下组成的组的数据集:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,或者(g)测序的DNA片段的序列基序。在一些实施方案中,多参数分析包括通过计算机应用一个或更多个数学变换以生成多参数模型。在一些实施方案中,多参数模型是选自由以下组成的组的多于一个变量的联合分布模型:(a)测序的DNA片段的起始位置,(b)测序的DNA片段的终止位置,(c)覆盖可映射位置的独特测序的DNA片段的数量,(d)测序的DNA片段的长度,(e)可映射碱基对位置将出现在测序的DNA片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的DNA片段内的似然,以及(g)测序的DNA片段的序列基序。

[0064] 在一些实施方案中,方法还包括在多参数模型中鉴定一个或更多个峰,每个峰具有峰分布宽度和峰覆盖率。在一些实施方案中,方法还包括检测代表无细胞DNA片段的多参数模型与参考多参数模型之间的一个或更多个偏差。在一些实施方案中,偏差选自由以下组成的组:(i)核小体区域外的读段数量的增加,(ii)核小体区域内的读段数量的增加,(iii)相对于可映射基因组位置更宽的峰分布,(iv)峰位置的偏移,(v)新峰的鉴定,(vi)峰的覆盖深度的变化,(vii)峰周围的起始位置的变化,以及(viii)与峰相关的片段大小的变化。

[0065] 在一些实施方案中,方法还包括确定归因于以下的多参数模型的贡献:(i)无细胞DNA起源的细胞中的凋亡过程或(ii)无细胞DNA起源的细胞中的坏死过程。在一些实施方案中,方法还包括进行多参数分析以(i)测量无细胞DNA片段的RNA表达,(ii)测量无细胞DNA片段的甲基化,(iii)测量无细胞DNA片段的核小体映射,或者(iv)鉴定无细胞DNA片段中一个或更多个体细胞单核苷酸多态性的存在或无细胞DNA片段中一个或更多个种系单核苷酸多态性的存在。在一些实施方案中,方法还包括生成分布评分,该分布评分包括指示具有双核小体保护的DNA片段的数量或具有单核小体保护的DNA片段的数量的值。在一些实施方案

中,方法还包括估计受试者的突变负荷。

[0066] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于分析源自受试者的无细胞脱氧核糖核酸(DNA)片段,该方法包括:获得代表无细胞DNA片段的多参数模型;并且利用计算机进行统计分析以将多参数模型分类为与代表不同群组的一个或更多个核小体占位谱相关。

[0067] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于创建经训练的分类器,包括:(a)提供多于一个不同的类别,其中每个类别代表具有共有特性的一组受试者;(b)针对从每个类别获得的多于一个无细胞DNA群体中的每一个,提供代表来自无细胞DNA群体的无细胞脱氧核糖核酸(DNA)片段的多参数模型,从而提供训练数据集;以及(c)通过计算机针对训练数据集训练学习算法以创建一个或更多个经训练的分类器,其中每个经训练的分类器被配置为将来自测试受试者的无细胞DNA测试群体分类为多于一个不同类别中的一个或更多个。

[0068] 在另一方面,本文公开了一种对来自受试者的测试样品进行分类的方法,所述方法包括:(a)提供代表来自受试者的无细胞DNA测试群体的无细胞脱氧核糖核酸(DNA)片段的多参数模型;以及(b)使用经训练的分类器对无细胞DNA测试群体进行分类。

[0069] 在另一方面,本文公开了一种计算机实施的方法,所述方法包括:(a)通过计算机生成来自受试者的无细胞DNA片段的序列信息;(b)通过计算机基于序列信息将无细胞DNA片段映射到参考基因组;以及(c)通过计算机分析映射的无细胞DNA片段,以确定在参考基因组中的多于一个碱基位置的每一个处的选自由以下组成的组的多于一种量度:(i)映射到碱基位置的无细胞DNA片段的数量,(ii)映射到碱基位置的每个无细胞DNA片段的长度,(iii)映射到碱基位置的无细胞DNA片段的数量随无细胞DNA片段的长度的变化;(iv)在碱基位置处起始的无细胞DNA片段的数量;(v)在碱基位置处终止的无细胞DNA片段的数量;(vi)在碱基位置处起始的无细胞DNA片段的数量随长度的变化,以及(vii)在碱基位置处终止的无细胞DNA片段的数量随长度的变化。

[0070] 在另一方面,本文公开了一种分析源自受试者的无细胞DNA片段的计算机实施的方法,该方法包括:(a)通过计算机接收代表无细胞DNA片段的序列信息,以及(b)进行每个可映射碱基位置或基因组位置的包括以下的多于一个的分析:(i)在碱基位置或基因组位置处起始或终止的序列片段的数量,(ii)碱基位置或基因组位置处的序列或片段长度,(iii)碱基位置或基因组位置处的片段或序列覆盖率,以及(iv)碱基位置或基因组位置处的序列基序分布。在另一方面,本文公开了一种生成分类器的方法,该分类器用于确定受试者属于一个或更多个具有临床意义的类别的似然,该方法包括:a)提供训练集,该训练集包括针对一个或更多个具有临床意义的类别中的每一个的来自属于具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体以及来自不属于具有临床意义的类别的物种的多于一个受试者中的每一个的无细胞DNA群体;b)对来自无细胞DNA群体的无细胞DNA片段进行测序以产生多于一个DNA序列;c)针对每个无细胞DNA群体,将多于一个DNA序列映射到物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;d)针对每个无细胞DNA群体制备数据集以产生训练集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一种特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因

座处终止的DNA序列;以及e)针对训练集训练基于计算机的机器学习系统,从而生成用于确定受试者属于一个或更多个具有临床意义的类别的似然的分类器。

[0071] 在另一方面,本文公开了一种确定受试者中异常生物学状态的方法,该方法包括:a)对来自受试者的无细胞DNA的无细胞DNA片段进行测序以产生DNA序列;b)将DNA序列映射到受试者的物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;c)制备数据集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一个特征的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列;以及d)基于数据集,确定异常生物学状态的似然。

[0072] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;以及(b)针对一个或更多个遗传基因座中的每一个,通过计算机计算指示以下的定量量度:(1)具有与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段的数量,与(2)具有与遗传基因座相关的单核小体保护的DNA片段的数量的比率,或反之亦然;以及(c)使用针对一个或更多个遗传基因座中的每一个的定量量度来确定指示受试者中一个或更多个遗传基因座中遗传畸变的存在或不存在的所述输出。

[0073] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;以及(b)使用该分布确定指示受试者中遗传畸变的存在或不存在的所述输出,其中所述存在或不存在是在(i)不将DNA片段的分布与来自受试者基因组外部来源的参考分布进行比较,(ii)不将源自DNA片段分布的参数与参考参数进行比较,以及(iii)不将DNA片段的分布与来自受试者的对照的参考分布进行比较的情况下确定的。在一些实施方案中,遗传畸变包括拷贝数变异(CNV)或单核苷酸变体(SNV)。

[0074] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于对来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段的分布进行去卷积,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段在基因组中多于一个碱基位置的覆盖率的分布;以及(b)针对一个或更多个遗传基因座中的每一个,通过计算机对该覆盖率的分布进行去卷积,从而生成与选自由拷贝数(CN)组分、细胞清除组分和基因表达组分组成的组的一个或更多个成员相关的分数贡献。在一些实施方案中,方法还包括至少基于分数贡献的一部分生成指示遗传畸变的存在或不存在的输出。

[0075] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;(b)通过计算机鉴定DNA片段的分布中在多于一个碱基位置的一个或更多个碱基位置处的一个或更多个峰,其中每个峰包含峰值和峰分布宽度;以及(c)通过计算机至少基于(i)一个或更多个碱基位置、(ii)峰值和(iii)峰分布宽度,确定受试者中遗传畸变的存在或不存在。

[0076] 在一些实施方案中,一个或更多个峰包含双核小体峰或单核小体峰。在一些实施方案中,至少基于指示以下的定量量度来确定指示遗传畸变的存在或不存在的所述输出:与双核小体峰相关的第一峰值和与单核小体峰相关的第二峰值的比率,或反之亦然。

[0077] 在另一方面,本文公开了一种计算机实施的方法,所述方法用于生成指示来自从受试者获得的无细胞DNA的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出,该方法包括:(a)通过计算机构建来自无细胞DNA的DNA片段跨基因组中多于一个碱基位置的分布;(b)通过计算机分析DNA片段在一个或更多个遗传基因座处的分布,该分析包括检测DNA片段的分布与选自以下的多于一个参考分布之间的偏差:(i)与健康对照的一个或更多个群组相关的一个或更多个健康参考分布,以及(ii)与患病受试者的一个或更多个群组相关的一个或更多个患病参考分布;以及(c)通过计算机至少基于在(b)中检测到的偏差确定指示受试者中遗传畸变的存在或不存在的所述输出。在一些实施方案中,分析包括计算一个或更多个差量信号,每个差量信号包括DNA片段的分布与多于一个参考分布的参考分布之间的差异。

[0078] 在另一方面,本文公开了一种用于处理受试者的生物样品的方法,所述方法包括:(a)获得所述受试者的所述生物样品,其中所述生物样品包含脱氧核糖核酸(DNA)片段;(b)测定所述生物样品以生成指示具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与遗传基因座相关的单核小体保护的DNA片段的存在或不存在的信号;以及(c)使用所述信号生成指示具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与遗传基因座相关的单核小体保护的DNA片段的所述存在或不存在的输出。在一些实施方案中,测定包括(i)富集所述生物样品以获得一个或更多个遗传基因座的组的DNA片段或(ii)将所述生物样品的所述DNA片段测序。

[0079] 在另一方面,本文公开了一种用于分析生物样品的方法,所述生物样品包含源自受试者的无细胞DNA片段,该方法包括检测来自相同遗传基因座的对应于单核小体保护和双核小体保护中的每一个的DNA片段。

[0080] 在另一方面,本文公开了一种用于分析生物样品的方法,所述生物样品包含源自受试者的无细胞DNA片段,该方法包括检测具有与遗传基因座相关的双核小体保护的DNA片段。在一些实施方案中,遗传基因座包含ERBB2、TP53或NF1。在一些实施方案中,遗传基因座包含表1中列出的基因。

[0081] 在另一方面,本公开内容提供了一种生成分类器的方法,该分类器用于确定受试者属于一个或更多个具有临床意义的类别的似然,该方法包括:a)提供训练集,该训练集包括针对一个或更多个具有临床意义的类别中的每一个的来自属于具有临床意义的类别的物种的多于一个受试者中的每一个的生物样品以及来自不属于具有临床意义的类别的物种的多于一个受试者中的每一个的生物样品,b)对来自生物样品的无细胞脱氧核糖核酸(cfDNA)分子进行测序以产生多于一个脱氧核糖核酸(DNA)序列;c)针对每个生物样品,将多于一个DNA序列映射到物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;d)针对每个样品制备数据集以产生训练集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一种特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列;以及e)针对训练集训练基于计算机的机器学习系统,从而生成

用于确定受试者属于一个或更多个具有临床意义的类别的似然的分类器。在一个实施方案中,定量量度包括具有所选特性的DNA序列的大小分布。

[0082] 在另一方面,一种确定受试者中异常生物学状态的方法包括:a)对来自受试者的生物样品的cfDNA分子进行测序以产生DNA序列;b)将DNA序列映射到受试者的物种的参考基因组中的一个或更多个基因组区域中的每一个,每个基因组区域包含多于一个遗传基因座;c)制备数据集,该数据集包括针对多于一个遗传基因座中的每一个的指示选自以下的至少一种特征的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列;以及d)基于数据集,确定异常生物学状态的似然。在一个实施方案中,方法还包括施用治疗性干预以治疗异常生物学状态。因此,用于施用治疗性干预以治疗异常生物学状态的方法可包括如本文所公开的确定受试者中的异常生物学状态,随后施用治疗性干预。

[0083] 在一个实施方案中,定量量度包括具有所选特征的DNA序列的大小分布。在一个实施方案中,大小分布包括指示具有双核小体保护的片段的数量和/或具有单核小体保护的片段的数量的值。在一个实施方案中,定量量度还包括具有所选特征的DNA序列的大小分布的比率。在一个实施方案中,数据集还包含针对多于一个遗传基因座的指示在内含子或外显子中的位置的值。

[0084] 另一方面提供了一种计算机可读介质,所述计算机可读介质包括机器可执行的代码,所述机器可执行的代码当被一个或更多个计算机处理器执行时,实施一种用于基于输入数据集输出数据集的异常状态类别的似然的方法,该方法包括针对多于一个遗传基因座中的每一个的指示一个或更多个特征的定量量度的值,所述一个或更多个特征源自片段组谱分析并且选自:(i)映射到遗传基因座的DNA序列,(ii)在基因座处起始的DNA序列,以及(iii)在遗传基因座处终止的DNA序列。

[0085] 本公开内容的另一方面提供了一种方法,所述方法包括向具有异常生物学状态的受试者施用有效量的被设计为治疗异常生物学状态的治疗,该受试者的特征在于具有指示异常生物学状态的片段组谱。

[0086] 本公开内容的另一方面提供了一种药物,所述药物有效治疗异常生物学状态,所述药物用于在包括以下的方法中使用:将药物施用至具有异常生物学状态或怀疑具有异常生物学状态的受试者,该受试者的特征在于具有指示异常生物学状态的片段组谱。

[0087] 本公开内容还提供了一种药物,所述药物有效治疗异常生物学状态,所述药物用于在制备用于治疗具有异常生物学状态或怀疑具有异常生物学状态的受试者的药剂中使用,该受试者的特征在于具有指示异常生物学状态的片段组谱。

[0088] 在另一方面,本文提供了一种方法,所述方法包括:提供来自多于一个训练受试者(例如,至少50个训练受试者)的训练数据,所述训练受试者包括来自第一类别的多于一个受试者和来自第二类别的多于一个受试者,并且其中该训练数据包括来自每个训练受试者的训练样品的映射到一个或更多个所选基因组基因座的cfDNA分子的多参数分布;并且训练机器学习算法以开发分类模型,该分类模型基于来自测试受试者的测试样品的测试数据将受试者分类为患有癌症或未患有癌症,所述测试数据包括映射到所选基因组基因座的cfDNA分子的多参数分布。在一些实施方案中,分类模型是概率模型。

[0089] 在一些实施方案中,第一类别和第二类别选自:患有癌症和未患有癌症,对疗法具

有响应和对疗法无响应,以及第一阶段的癌症和第二阶段的癌症。在一些实施方案中,多参数分布包括分子大小、分子起始位置和/或分子终止位置。在一些实施方案中,所选基因组基因座包括跨越多于一个癌基因(例如来自表1的目的基因)中的每一个的至少一个双核小体距离。

[0090] 在另一方面,本文提供了一种方法,所述方法包括:提供来自测试受试者的测试样品的测试数据,所述测试数据包括映射到一个或更多个所选基因组基因座的cfDNA分子的多参数分布;并且使用基于计算机的分类模型将测试受试者分类为属于第一类别或第二类别,所述分类模型基于来自多于一个训练受试者的训练数据,所述训练受试者包括来自第一类别的多于一个受试者和来自第二类别的多于一个受试者,并且其中该训练数据包括来自每个训练受试者的训练样品的映射到一个或更多个所选基因组基因座的cfDNA分子的多参数分布。在一些实施方案中,分类模型被选择为具有至少90%、至少95%、至少98%、至少99%或至少99.8%的阳性预测值。

[0091] 在另一方面,本文提供了一种方法,所述方法包括:使用如本文所描述的分类方法将受试者分类为患有癌症,并且对如此分类的受试者施用治疗性治疗。在另一方面,本文提供了一种方法,所述方法包括:向通过本文所描述的方法被分类为患有癌症的受试者施用治疗癌症的治疗性治疗。

[0092] 通过以下详细描述,本公开内容的其他方面和优点对于本领域技术人员将变得明显,其中仅示出和描述了本公开内容的说明性实施方案。如将认识到的,本公开内容能够具有其他和不同的实施方案,并且其若干细节能够在各种明显的方面进行修改,所有这些都不脱离本公开内容。因此,附图和描述本质上被认为是说明性的,而不是限制性的。

[0093] 通过引用并入

[0094] 本说明书中提及的所有出版物、专利和专利申请均通过引用并入本文,其程度如同每个单独的出版物、专利或专利申请被具体地和单独地指示为通过引用并入一样。

[0095] 附图简述

[0096] 在所附权利要求中具体阐述了本公开内容的新颖特征。通过参考以下详细描述以及附图(本文也为“图(Figure)”和“图(FIG.)”)将获得对本公开内容的特征和优点的更好的理解,所述详细描述阐述了其中利用本公开内容的原理的说明性实施方案,在附图中:

[0097] 图1A示出了具有一个或更多个组分的片段组信号的实例。

[0098] 图1B示出了具有一个或更多个组分的片段组信号的实例,每个组分受到清除因子的影响。

[0099] 图1C示出了转录起始位点(TSS)的变化,如通过对比正常样品,恶性(晚期肺癌)中双核小体复合物的存在所指示的。

[0100] 图1D示出了在相同区域中单变量片段起始密度的有限分辨率。

[0101] 图1E示出了在临床样品中观察到的无细胞DNA(cfDNA)的片段长度分布。

[0102] 图2示出了cfDNA片段跨片段长度和基因组位置的热图(heat plot)的实例,即三维多参数分析。

[0103] 图3A至图3D显示了4个转换的多参数热图谱(heat map)的实例,该转换的多参数热图谱显示了三个不同基因组位置(两个来自PIK3CA,并且一个来自EGFR)的血浆失调度量。

[0104] 图3A显示了对应于PIK3CA|2238基因组位置的热图谱,其中外显子归一化的10bp(碱基对)片段起始覆盖率(x-轴)的值范围为从约0至约0.10,并且居中中值10bp片段大小(y-轴)的值范围为从约148bp至约172bp。

[0105] 图3B显示了对应于PIK3CA|2238基因组位置的热图谱,其中外显子归一化的10bp片段起始覆盖率(x-轴)的值范围为从约0.014至约0.035,并且居中中值10bp片段大小(y-轴)的值范围为从约150bp至约185bp。

[0106] 图3C显示了对应于PIK3CA|2663基因组位置的热图谱,其中外显子归一化的10bp片段起始覆盖率(x-轴)的值范围为从约0.028至约0.075,并且居中中值10bp片段大小(y-轴)的值范围为从约155bp至约185bp。

[0107] 图3D显示了对应于EGFR|6101基因组位置的热图谱,其中外显子归一化的10bp片段起始覆盖率(x-轴)的值的范围为从约0.01至约0.061,并且居中中值10bp片段大小(y-轴)的值范围为从约145bp至约186bp。每个临床样品用如下的纯色圆圈表示:健康对照以深绿色显示,并且患有癌症的受试者以范围从蓝色、青色、黄色、橙色和红色的颜色显示(分别对应于0.1%至93%的最大突变体等位基因分数(最大MAF)值)。实际上,蓝色圆圈可对应于谱图(spectrum)(例如,患有癌症的受试者群组中最大MAF值的范围)的最小值或最低值端,而红色圆圈可对应于谱图(例如,患有癌症的受试者群组中的最大MAF值的范围)的最大值或最高值端。

[0108] 图4显示了血浆失调评分的样品,因为它在给定临床样品中跨基因组片段随位置而变化(下图)。上图显示了测定的相关基因列表以及在这些基因中发现的任何改变(SNV或CNV)。

[0109] 图5显示了由在5,000个样品中跨多个基因组区域的血浆失调评分的无监督聚类所生成的热图,每个样品来自不同的非小细胞肺癌(NSCLC)患者。Y-轴反映5,000个患者样品中的每一个。X-轴反映分析的一组基因组位置。颜色反映了每个样品的每个基因组位置的血浆失调评分。

[0110] 图6显示了跨小范围的基因组位置,例如KRAS基因生成的热图谱。在这种情况下,血浆失调评分具有10bp的分辨率,例如,每10bp计算一次。Y-轴提供2,000个临床样品的信息。X-轴以10bp的分辨率提供跨KRAS基因的血浆失调评分。

[0111] 图7示出了可以在碱基对之间切割双链DNA的酶的实例:微球菌核酸酶。

[0112] 图8示出了多参数模型的一个方面,特别是基因组范围内每个基因组位置处的片段频率的图。

[0113] 图9示出了多参数模型的一个方面,特别是基因组范围内每个基因组位置处的片段频率的图。

[0114] 图10示出了多参数模型的两个方面,特别是在基因组范围内的每个基因组位置处的归一化分子计数和归一化片段大小(即,长度)的图。

[0115] 图11示出了多参数模型的两个方面,特别是在基因组范围内的每个基因组位置处的归一化分子计数和归一化片段大小(即,长度)的图。

[0116] 图12示出了多参数模型的三个方面,特别是在基因组范围内的每个基因组位置处的归一化分子计数、归一化片段大小(即,长度)以及归一化双链的百分比。

[0117] 图13示出了多参数模型的一个方面,特别是基因组范围内的每个基因组位置(x-

轴)处的读段计数(y-轴)。

[0118] 图14示出了可以作为多参数分析的一部分被执行以生成多参数模型的数学变换的实例。

[0119] 图15示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。

[0120] 图16示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。

[0121] 图17示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。

[0122] 图18示出了在基因组的给定区域中的核小体组织(nucleosomal organization)对比基因组位置的实例。

[0123] 图19示出了在基因组的给定区域中的核小体组织对比基因组位置的实例。

[0124] 图20示出了用于确定绝对拷贝数(CN)的过程的实例。

[0125] 图21A和21B示出了通过血浆DNA的全测序使用片段组谱分析推断拷贝数扩增基因的活化的实例。图21A显示了2,076个临床样品中ERBB2中归一化的双核小体与单核小体计数的比率的图。图21B显示了图21A的图的放大部分。

[0126] 图22显示了被编程或以其他方式配置以实施本文提供的方法的计算机系统。

[0127] 图23显示了跨肿瘤类型的单核小体分辨率片段化模式(例如,来自片段组谱分析或“片段组学”分析)。

[0128] 图24显示了源自包含768名患有晚期肺腺癌的患者的群组的片段组谱分析(“片段组学”)的特征的实例。

[0129] 图25显示了可以用于使用片段组信号进行异常检测的K-组分混合物模型的实例。

[0130] 图26A显示了拟合到二元正态混合物模型以鉴定异常cfDNA片段组信号的椭圆形包络的实例。

[0131] 图26B示出了通过跨5个不同群组(结肠直肠癌术后、结肠直肠癌术前、肺癌术后、肺癌术前和正常)的cfDNA样品的片段组分析而生成的失调评分的分布的实例。

[0132] 图27A示出了多参数模型的实例,该多参数模型包括在与TP53基因,外显子#7相关的基因组区域中的受试者的片段大小(例如,片段长度)和基因组位置。

[0133] 图27B显示了20个样品的四个聚合的晚期乳腺癌群组中的ERBB2启动子区域的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(ii)包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(iii)包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组,以及(iv)包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组。

[0134] 图27C显示了20个样品的四个聚合的晚期乳腺癌群组中的ERBB2增强子区域的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(ii)包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(iii)包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组,以及(iv)包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组。

[0135] 图28A显示了对齐的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)从单个样品(来自ERBB2阳性受试者)生成的ERBB2增强子区域(右上)的热图谱,(ii)从多于一个健康对照生成的聚合群组热图谱,以及(iii)从多于一个高ERBB2CN和低突变负荷受试者生成的聚合群组热图谱。此外,示出了在4个不同的基因组区域(例如,

对应于TP53、NF1、ERBB2和BRCA1基因)处的单核小体和双核小体计数(例如,测试样品中计数的在该基因组位置处起始的片段的数量)的覆盖图。

[0136] 图28B显示了对齐的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)从单个样品(来自ERBB2阴性受试者)生成的ERBB2增强子区域(右上)的热图谱,(ii)从多于一个健康对照生成的聚合群组热图谱,以及(iii)从多于一个高ERBB2CN和低突变负荷受试者生成的聚合群组热图谱。此外,示出了在4个不同的基因组区域(例如,对应于TP53、NF1、ERBB2和BRCA1基因)处的单核小体和双核小体计数的覆盖图。

[0137] 图29A和29B显示了ERBB2和NF1外显子结构域的2D核小体映射(未扩增)。在每个图的底部,显示了2D密度估计和图像处理。在每个图的顶部,显示了跨30个近二倍体ERBB2临床病例观察到的规范结构域的核小体掩模(nucleosomal mask)。

[0138] 图30显示了跨4个不同群组的推断的染色体17肿瘤负荷的图,这4个不同群组先前已通过液体活检测定法测定最大MAF:(i)最大MAF范围为(0,0.5]的群组,(ii)最大MAF范围为(0.5,5]的群组,(iii)最大MAF范围为(5,20]的群组,以及(iv)最大MAF范围为(20,100]的群组。

[0139] 图31A显示了ERBB2表达组分对比ERBB2拷贝数的图。

[0140] 图31B示出了使用ERBB2阴性训练集的2D阈值化的图,其经由构建方差-协方差矩阵,对方差-协方差矩阵求逆以及生成椭圆判别函数来执行。

[0141] 图32A显示了跨2360个晚期癌症受试者和43个健康对照MPL基因结构域中的双核小体片段的相对富集的图。

[0142] 图32B和图32C显示了MPL基因的可变转录物中残差双核小体比率信号中的断点的实例。图32C显示了图32B的放大部分。

[0143] 详细描述

[0144] 虽然本文已经显示和描述了本发明的优选实施方案,但是对于本领域技术人员明显的是,此类实施方案仅以示例性的方式提供。本领域技术人员现在将想到许多变化、改变和替换形式,而不偏离本发明。应理解的是,在实践本发明时,可以采用本文所描述的本发明实施方案的各种替代方案。

[0145] 如本文所用,术语“生物样品”通常指源自受试者的组织或流体样品。可以直接从受试者获得生物样品。生物样品可以是或可以包括一种或更多种核酸分子,诸如脱氧核糖核酸(DNA)或核糖核酸(RNA)分子。生物样品可以源自任何器官、组织或生物流体。生物样品可包括例如体液或实体组织样品。实体组织样品的实例是肿瘤样品,例如来自实体肿瘤活检的肿瘤样品。体液包括例如血液、血清、血浆、肿瘤细胞、唾液、尿液、淋巴液、前列腺液、精液、乳汁、痰液、粪便、泪液及其衍生物。

[0146] 如本文所用,术语“受试者”通常指任何动物、哺乳动物或人。受试者可具有、潜在具有或怀疑具有选自癌症、与癌症相关的症状、对于癌症无症状或未确诊(例如,未被诊断为癌症)的一种或更多种特性。受试者可患有癌症,受试者可表现出与癌症相关的症状,受试者可以没有与癌症相关的症状,或者受试者可以未被诊断出患有癌症。在一些实施方案中,受试者是人。

[0147] 如本文所用,术语“无细胞DNA”(或“cfDNA”)通常指在受试者的血流中自由循环的DNA片段。无细胞DNA片段可具有双核小体保护(例如,至少240个碱基对(“bp”)的片段大

小)。具有双核小体保护的这些cfDNA片段可能不在核小体之间被切割,导致片段长度较长(例如,具有以334bp为中心的典型大小分布)。无细胞DNA片段可具有单核小体保护(例如,小于240碱基对("bp")的片段大小)。具有单核小体保护的这些cfDNA片段可能在核小体之间被切割,导致片段长度较短(例如,具有以167bp为中心的典型大小分布)。本文论述的cfDNA可以不具有胎儿起源,并且受试者通常可以是未妊娠的。

[0148] 如本文所用,术语"DNA序列"通常指"原始序列读段"和/或"共有序列"。原始序列读段是DNA测序仪的输出物,并且通常包括同一亲本分子的冗余序列(例如在扩增后)。"共有序列"是源自亲本分子的冗余序列的序列,旨在代表原始亲本分子的序列。共有序列可以通过投票(其中序列中的每个多数核苷酸,例如在给定碱基位置处最常观察到的核苷酸,是共有核苷酸)或其他方法诸如与参考基因组进行比较而产生。通过用独特或非独特的分子标签将原始亲本分子加标签可以产生共有序列,这允许通过追踪标签和/或使用序列读段内部信息来追踪子代序列(例如,在扩增后)。加标签或条形码化的实例以及标签或条形码的使用提供在例如美国专利公布第2015/0368708号、第2015/0299812号、第2016/0040229号和第2016/0046986号中,其全部内容通过引用并入本文。

[0149] 测序方法可以是第一代测序方法,例如Maxam-Gilbert或Sanger测序,或高通量测序(例如,下一代测序或NGS)方法。高通量测序方法可以同时(或基本上同时)将至少10,000、100,000、1百万、1000万、1亿、10亿或更多个多核苷酸分子测序。测序方法可包括但不限于:焦磷酸测序、合成测序、单分子测序、纳米孔测序、半导体测序、通过连接测序、通过杂交测序、数字基因表达(Helicos)、大量平行测序,例如Helicos、克隆单分子阵列(Solexa/Illumina)、使用PacBio、SOLiD、Ion Torrent或纳米孔(Nanopore)平台的测序。

[0150] 如本文所用,术语"参考基因组"(有时称为"组装体(assembly)")通常指一种核酸序列数据库,其由遗传数据组装并且旨在代表物种的基因组。通常,参考基因组是单倍体。通常,参考基因组不代表该物种的单个个体的基因组,而是若干个体的基因组的镶嵌图。参考基因组可以是公众可获得的或私有参考基因组。人类参考基因组包括例如hg19或NCBI Build 37或Build 38。

[0151] 如本文所用,术语"参考序列"通常指与受试者的核苷酸序列进行比较的核苷酸序列。通常,参考序列源自参考基因组。

[0152] 如本文所用,术语"映射"通常指基于序列同源性将DNA序列与参考序列对齐。可以使用对齐算法例如,Needleman-Wunsch算法(参见例如,可在URL ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html获得的EMBOSS Needle对齐器,任选地具有默认设置)、BLAST算法(参见例如,可在URL blast.ncbi.nlm.nih.gov/Blast.cgi获得的BLAST对齐工具,任选地具有默认设置)或Smith-Waterman算法(参见例如,可在URL ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html获得的EMBOSS Water对齐器,任选地具有默认设置)来执行对齐。可以使用所选算法的任何合适参数,包括默认参数来评估最佳对齐。

[0153] 如本文所用,术语"基因组区域"通常指基因组的任何区域(例如,碱基对位置的范围),例如整个基因组、染色体、基因或外显子。基因组区域可以是连续的或非连续的区域。"遗传基因座"(或"基因座")可以是基因组区域的一部分或全部(例如,基因、基因的一部分或基因的单个核苷酸)。

[0154] 如本文所用,术语"定量量度"通常指绝对或相对量度。定量量度可以是但不限于

数字、统计测量值(例如,频率、平均值、中值、标准偏差或分位数),或程度或相对量(例如,高、中和低)。定量量度可以是两种定量量度的比率。定量量度可以是定量量度的线性组合。定量量度可以是归一化量度。

[0155] 如本文所用,术语“异常生物学状态”通常指在某种程度上偏离正常的生物系统的状态。异常状态可以在生理或分子水平上发生。例如但不限于,异常生理状态(疾病、病理)或遗传畸变(突变、单核苷酸变体、拷贝数变体、基因融合、插入/缺失等)。疾病状态可以是癌症或癌前期。异常生物学状态可以与异常程度(例如,指示距正常状态的距离的定量量度)相关。

[0156] 如本文所用,术语“似然”通常指概率、相对概率、存在或不存在或程度。

[0157] 如本文所用,术语“机器学习算法”通常指通过计算机执行的算法,该算法自动化分析模型构建,例如用于聚类、分类或模式识别。机器学习算法可以是监督的或无监督的。学习算法包括,例如人工神经网络(例如,反向传播网络)、判别分析(例如,贝叶斯分类器(Bayesian classifier)或Fischer分析)、支持向量机、决策树(例如,递归分区过程,诸如CART-分类和回归树,或随机森林)、线性分类器(例如,多元线性回归(MLR)、偏最小二乘(PLS)回归和主成分回归)、分层聚类和聚类分析。机器学习算法针对其进行学习的数据集可以称为“训练数据”。

[0158] 如本文所用,术语“分类器”通常指算法计算机代码,其接收测试数据作为输入,并且产生输入数据属于一个类别或另一个类别的分类作为输出。

[0159] 如本文所用,术语“数据集”通常指表征系统的要素的值的集合。系统可以是例如来自生物样品的cfDNA。此类系统的要素可以是遗传基因座。数据集(dataset)(或“数据集(data set)”)的实例包括指示选自以下的特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在遗传基因座处起始的DNA序列,(iii)在遗传基因座处终止的DNA序列;(iv)DNA序列的双核小体保护或单核小体保护;(v)位于参考基因组的内含子或外显子中的DNA序列;(vi)具有一个或更多个特性的DNA序列的大小分布;以及(vii)具有一个或更多个特性的DNA序列的长度分布等。

[0160] 如本文所用,术语“值”通常指数据集中的条目,可以是表征该值所指的特征的任何事物。这包括但不限于数字、单词或短语、符号(例如,+或-)或度数。

[0161] 如本文所用,术语“液体活检”通常指非侵入性或微创性实验室测试或测定(例如,生物样品或无细胞DNA的实验室测试或测定)。此类“液体活检”测定可以报告一种或更多种肿瘤相关的标志物基因的测量值(例如,次要等位基因频率、基因表达或蛋白质表达)。此类液体活检测定可以是商业上可获得的,例如来自Guardant Health的循环肿瘤DNA测试,来自Fluxion Biosciences的Spotlight 59oncology panel,来自Agena Bioscience的UltraSEEK lung cancer panel,来自Foundation Medicine的FoundationACT液体活检测定以及来自Personal Genome Diagnostics的PlasmaSELECT测定。此类测定可报告一组遗传变体(例如,SNV、CNV、插入/缺失和/或融合)中的每一种的次要等位基因分数(MAF)值的测量值。

[0162] 如本文所用,术语“多模态密度(multimodal density)”通常指跨多个参数的密度或密度分布。多模态密度可包括多元混合的分布。

[0163] 引言

[0164] 癌症形成和进展可以源于脱氧核糖核酸(DNA)的遗传和表观遗传修饰。本公开内容提供了分析DNA,诸如无细胞DNA(cfDNA)的表观遗传修饰的方法。此类“片段组”分析可单独使用或与现有技术组合使用,以确定疾病或状况的存在或不存在、诊断的疾病或状况的预后、诊断的疾病或状况的治疗性治疗,或预测的疾病或状况的治疗结果。

[0165] 循环无细胞DNA(cfDNA)可以主要是从垂死的组织细胞脱落到诸如外周血(血浆或血清)的体液中的短DNA片段(例如,具有约100至400个碱基对的长度,众数为约165bp)。除了癌症相关的遗传变体之外,cfDNA的分析还揭示表观遗传印迹和垂死细胞的吞噬去除的特征,这可产生存在的恶性肿瘤(例如肿瘤)以及它们的微环境组分的聚合核小体占位谱。

[0166] 一种、两种或更多种组分或因子可以促成血浆片段组信号(例如,从cfDNA片段的分析获得的信号),包括(i)在DNA拆解期间的细胞死亡类型和相关的染色质凝聚事件,(ii)清除机制,其可涉及被受试者免疫系统调节的各种类型的吞噬机制,以及(iii)血液组成的非恶性变异,其可受循环中细胞类型的潜在组合的影响,(iv)给定类型的器官或组织中非恶性细胞死亡的多种来源或原因,以及(v)癌症中细胞类型的异质性,因为恶性实体瘤包括肿瘤相关的正常细胞、上皮细胞和基质细胞、免疫细胞和血管细胞,所有这些细胞中的任何一种都可促成并且在cfDNA样品(例如,其可以从受试者的体液获得)中被体现。

[0167] 呈组蛋白保护的复合物形式的无细胞DNA可以被各种宿主细胞释放,包括嗜中性粒细胞、巨噬细胞、嗜酸性粒细胞以及肿瘤细胞。循环DNA通常具有短的半衰期(例如,约10分钟至15分钟),并且肝脏通常是从血液循环除去循环DNA片段的主要器官。cfDNA在循环中的积累可能是由于细胞死亡和/或活化增加、cfDNA清除受损和/或内源性DNA酶水平降低所致。在受试者血流中循环的无细胞DNA(cfDNA)通常可以包装成膜包被的结构(例如凋亡小体)或与生物聚合物的复合物(例如组蛋白或DNA结合血浆蛋白)。可以分析DNA片段化和随后的运输的过程以确定它们对无细胞DNA信号特性的影响,如由片段组分析检测到的。

[0168] 在细胞核(例如人类的细胞核)中,DNA通常存在于核小体中,核小体被组织成包含缠绕在核心组蛋白八聚体周围的约145个DNA碱基对(bp)的结构。DNA和组蛋白二聚体的静电和氢键相互作用可导致DNA在蛋白质表面上的在能量上不利的弯曲。此类弯曲可在空间上禁止其他DNA结合蛋白,并且因此可用于调节细胞核中对DNA的接近。细胞中的核小体定位可以动态地波动(例如,随着时间的推移并且跨各种细胞状态和条件动态地波动),例如,自发地部分解缠绕(unwrap)和重新缠绕(rewrap)。由于片段组信号可以反映源自受核小体单元影响的构型的组蛋白保护的DNA片段,因此核小体稳定性和动力学可影响此类片段组信号。这些核小体动力学可来源于多种因素,诸如:(i)ATP依赖性重塑复合物,其可利用ATP水解能量滑动核小体并从染色质纤维交换或驱逐出组蛋白,(ii)组蛋白变体,其可具有不同于规范组蛋白的属性,并在染色质纤维内产生局部特异性结构域,(iii)组蛋白伴侣,其可控制游离组蛋白的供应,并与组蛋白沉积和驱逐中的染色质重构器配合,以及(iv)组蛋白的翻译后修饰(PTM)(例如,乙酰化、甲基化、磷酸化和泛素化),其可直接或间接地影响染色质结构。

[0169] 因此,cfDNA中的片段化信号或模式可以指示聚合的cfDNA信号,所述聚合的cfDNA信号源于与跨基因组的染色质组织中的异质性相关的多个事件。此类染色质组织可取决于诸如全局细胞身份、代谢状态、区域调节状态、垂死细胞中的局部基因活性和DNA清除机制等因素而不同。此外,无细胞DNA片段组信号可仅部分归因于贡献细胞的潜在染色质架构。

此类cfDNA片段组信号可指示细胞死亡期间染色质压缩以及免于酶促消化的DNA保护的更复杂印迹。因此,特定于给定细胞类型或细胞谱系类型的染色质图谱可仅部分地促成由于在细胞死亡或碎片运输的各个阶段的核小体稳定性、构象和组成的变化的DNA可及性的固有异质性。结果,一些核小体可能变得优先存在或不存在于无细胞DNA中(例如,可存在影响cfDNA清除并释放到血液循环中的过滤机制),这可取决于诸如死亡以及细胞尸体清除的模式和机制等因素。

[0170] 由于细胞过程诸如细胞凋亡和坏死期间的核DNA片段化,片段组信号可以在细胞中生成并且作为cfDNA释放到血液循环中。此类片段化可由于不同的核酸酶作用于不同阶段的细胞中的DNA而产生,导致序列特异性DNA裂解模式,其可以在cfDNA片段组信号中被分析。对此类清除模式进行分类可以是细胞环境(例如,肿瘤微环境、炎症、疾病状态、肿瘤发生等)的临床相关标志。

[0171] 可以通过将cfDNA片段分类成不同的组分来分析片段组信号,这些组分对应于cfDNA片段所来源的不同染色质状态。例如,片段组信号可以被表示为代表不同潜在染色质状态的组分(例如,良性全身响应、肿瘤全身响应、肿瘤微环境和肿瘤)的总和,如图1A所示。该"染色质状态清除"模型可以通过将组分乘以清除因子(clearance factor)来修改,因为每种染色质状态可以具有不同的潜在清除机制(例如,特定于组织类型、器官类型或肿瘤类型的清除机制)。如图1B所示,片段组信号可以被建模为一个或更多个组分的总和,其中每个组分受到清除因子的影响(例如,乘以清除因子)。此类组分和清除因子可以代表可以被用于区分相似或相同染色质状态的非变体标志。片段组分析可以使用此"染色质状态清除"模型通过鉴定特定区域(或特征)来进行,其中染色质状态中的一种或更多种或其清除机制中的一种或更多种足够不同以被用作例如遗传畸变或疾病状态的标志指标物。此类遗传畸变可包括SNV、CNV、插入/缺失、融合。

[0172] 片段组分析可以揭示染色质组织或结构中的规范或非规范变异,其可能是DNA中基因组畸变和/或表观遗传变化的结果。此类测量值可以揭示例如以下中的一个或更多个:(i)癌症特异性肿瘤微环境,(ii)对生理应激的基质响应,所述对生理应激的基质响应导致癌症特异性的基质脱落特性,(iii)响应于免疫活性癌症片段的微小存在的血细胞组成变化,和/或(iv)响应于与芽殖肿瘤小生境形成相关的细微组织免疫谱变异的血液组成。可以通过片段组分析测量或推断的遗传畸变可以包括表观遗传变体或变化。

[0173] 包括局灶扩增和/或非整倍性的体细胞拷贝数变体(CNV)代表在许多癌症,特别是转移性癌症中通常观察到的一组遗传畸变。通常,拷贝数指特定基因或DNA序列的每个细胞的拷贝数。然而,当对异质性多克隆肿瘤环境进行谱分析时,此类拷贝数(CN)的解释可能变得不太准确。此类肿瘤细胞跨异质性肿瘤细胞群体可具有宽范围的CN。

[0174] 体细胞获得的染色体重排诸如缺失和重复,特别是局灶重排,可以导致基因表达水平的改变--一种称为基因剂量效应的现象。

[0175] 微阵列技术被广泛用于CNV检测中,诸如阵列比较基因组杂交(阵列CGH)和单核苷酸多态性(SNP)微阵列。在传统阵列CGH中,参考和测试DNA被荧光标记并与阵列杂交,并且信号比率用作拷贝数(CN)比率的估计值。SNP微阵列也基于杂交,但是在每个微阵列上处理单个样品,并且通过将调查中的样品的强度与参考样品的集合或与所研究的所有其他样品进行比较来形成强度比率。虽然微阵列/基因分型阵列对于大型CNV检测是有效的,但它们

对于检测短基因或DNA序列(例如,长度小于约50千碱基(kb))的CNV不太灵敏。

[0176] 通过提供基因组的逐个碱基视图,下一代测序(NGS)可以检测可能保持未被阵列检测到的小的或新颖的CNV。合适的NGS方法的实例可包括全基因组(WGS)、全外显子组测序(WES)或靶向外显子组测序(TES)。然而,在开发用于检测来自个体测序样品的CNV(例如,拷贝数扩增(CNA))的计算算法方面仍存在挑战,部分原因是由于杂交引入的偏倚以及整个基因组中的稀疏和不均匀覆盖。

[0177] 获得肿瘤组织(例如,通过昂贵和侵入性活检程序)的困难和相关的健康风险促使基于血液的微创测定的开发。对血液进行谱分析可提供若干实际优点,包括样品采集的微创性质、采样方案归一化的相对容易性,以及随时间获得重复样品的能力。以前的研究已经鉴定了患有不同癌症类型的患者血浆中的癌症相关的变体,包括微卫星改变和基因突变。在血浆中大量非肿瘤DNA的存在下检测癌症变体可能在拷贝数检测中提出新的挑战。

[0178] 此外,血浆来源的无细胞DNA保留了先前在染色质结构的全基因组分析中(特别是在微球菌核酸酶测序中,或'MNase-seq',测定中)注意到的特性,特别是那些通过检查cfDNA中观察到的DNA片段化模式而确定的与人类组织的表观遗传景观相关的特性。图7示出了可以在碱基对之间切割双链DNA的酶的实例:微球菌核酸酶(MNase)。1:3稀释的微球菌核酸酶可以在任何碱基对位置处裂解而对特定序列没有特异性。微球菌核酸酶可以消化染色质,并且从而提供有关DNA链中核小体位置的信息。对各种模式生物和人类细胞系的研究揭示,核小体在DNA上的定位是可变的并且是组织特异性的,使得依赖于参考信号的传统拷贝数方法对于短CNV变体的血浆来源的DNA拷贝数检测是次优的。特别地,cfDNA片段拷贝数可取决于潜在细胞或组织类型的核小体定位、细胞清除和/或基因表达,其可随时间和细胞状态而变化。已经观察到无细胞DNA信号根据在组织中观察到的核小体定位表现,使得核小体消耗发生在活跃表达的基因的转录起始位点(TSS)处,并且因此TSS内某些DNA片段的盛行直接反映了造血细胞的表达特征。

[0179] 即使当基因被活跃转录(例如,通过DNA聚合酶II(Pol II))时,也可以存在核小体。然而,核小体定位通常在细胞中随时间改变,并且当诱导转录时一些核小体可能丢失。例如,在许多真核基因上,Pol II在转录模板的最初50bp至100bp后暂停。在涉及DNA环化的中等水平转录期间,原始组蛋白可能保留在DNA上,而当多个转录复合物置换组蛋白时,在强烈转录期间可能发生更显著的重塑。结果,DNA片段的单核小体和双核小体性质之间的判别可以有助于鉴定和确定转录起始位点(TSS)周围的潜在调节,例如,在可变TSS启动子使用的情况下,如图1C所示,其中片段起始覆盖的单变量分析未显示存在双核小体复合物(例如,其可指示可变转录起始,如图1D中所示)。

[0180] 尽管最近在阐明无细胞DNA的起源方面取得了进展,但仍然需要核小体感知的体细胞变体检测算法。核小体感知变体检测方法可以扩展我们对核小体定位如何影响cfDNA片段模式和信号的理解,并且可以集中于在转录因子结合位点和转录起始位点之外的基于核小体的无细胞DNA片段化模式(片段组学)分析的扩展。

[0181] 本公开内容提供了使用单参数分析或多参数分析来确定血浆失调评分。单参数分析可以包括对具有一个独立参数的分布函数的分析。多参数分析可以包括对具有两个或更多个独立参数的分布函数的分析。血浆失调评分可以跨基因组(例如,跨基因组位置)变化。该变化可以基于例如与多于一个碱基位置中的每个碱基位置重叠的片段的数量。多于一个

碱基位置可以选自基因组的一部分或全部。该变化可以基于例如与基因组的一部分或全部的每个位置重叠的片段的长度的分布。

[0182] 在一个方面,确定血浆失调评分可以包括绘制样品中在一组基因组位置中的每一个处具有特定长度的cfDNA片段的数量(例如,通过NGS或其他测序方法检测)。这可以通过多参数分析来完成,例如,创建三维(3-D)图,其中第一个轴可以代表与基因组的一个或更多个区域(例如,多于一个碱基对位置的连续跨度,或如表1中给出的一组基因组区域)重叠的多于一个基因组位置。3-D图的第二个轴可以代表样品中一组可能的片段的长度(例如,0bp至400bp)中的每一个。3-D图的第三个轴可以代表在片段长度的每一个处与独特基因组位置重叠的片段的数量。

[0183] 当在此类3-D矩阵中绘制数据时,所得到的多参数分布图可用于确定评分。该评分可以是血浆失调评分,如本文其他地方所描述。

[0184] 在另一方面,确定血浆失调评分可包括单参数分析,例如,创建二维(2-D)图,其中第一个轴可以代表与基因组的一个或更多个区域(例如,多于一个碱基对位置的连续跨度,或如表1中给出的一组基因组区域)重叠的多于一个基因组位置。2-D图的第二个轴可以代表样品中具有特定长度并且与多于一个基因组位置中的每一个重叠的cfDNA片段的数量。

[0185] 片段组分析可包括上述一个或更多个单参数或多参数分析。片段组分析可包括使用无细胞核酸的核小体谱分析,将核小体谱分析的模式与特定表型诸如疾病或状况相关,或配置分类器以帮助将样品分类为一个或更多个相关类别。例如,分类器使用内含子-外显子边界信息,其包括参考基因组中的内含子-外显子边界的位置和片段组信息(例如,一个或更多个多参数模型或单参数模型),其包含指示内含子或外显子中的或接近内含子-外显子边界的位置的值。此类内含子-外显子边界信息可以提供用于判别遗传变体或异常生物学状态的信息。还可以使用片段组分析,例如,以鉴定可用于选择性富集基因组的独特部分以检测相关表型的探针、引物和引诱物(bait)。

[0186] 序列信息

[0187] 本文的片段组谱分析利用源自无细胞核酸分子样品的序列信息。有许多方法可以确定序列信息。实例包括使用HiSeq(Illumina)或Ion Torrent(Thermo Fisher)进行测序。特别地,配对末端测序可用于测量血浆中单个DNA分子的邻接性,例如,以研究将染色质DNA裂解成核小体间片段的内源性内切核酸酶的活化模式。由于核小体占位模式,这些cfDNA片段长度被观察为分布,如图1E所示。横轴是片段长度(以碱基对,“bp”计),而纵轴显示具有给定片段长度的cfDNA片段的数量。在片段长度分布中的峰在约167bp观察到,其对应于缠绕在组蛋白八聚体核心周围的约147bp的DNA和接头DNA区段。在约334bp(例如,在167bp的片段长度的两倍处)也观察到较小的峰,其对应于缠绕组蛋白八聚体核心两次的DNA(例如,围绕单个组蛋白或围绕两个连续组蛋白的两次)和相关的接头DNA。通过观察沿着多参数热图的一个或更多个轴间隔约167bp的一个或更多个周期性峰,在多参数分析期间约167bp的该片段长度分布的峰可以是明显的。

[0188] 在cfDNA信号中观察到的凋亡DNA片段化的存在下,配对末端测序允许确定DNA结合的核小体和转录因子的位置和占据二者。反过来,这种方法允许人们甚至以亚核小体分辨率区分来自不同染色质架构谱的分子群体。检查cfDNA片段如何跨基因组起始与片段长度空间变化可导致热图可视化,如图2中所示。

[0189] 在从无细胞核酸样品获得序列数据后,可以对齐序列数据并将其叠并成独特的分子读段。对齐方法包括ClustalW2、Clustal Omega和MAFFT。

[0190] 可任选地叠并本文来源的测序信息以确定独特的分子和/或独特的序列读段。用于叠并成独特分子的方法描述于例如Population Genetics的VeriTag和Johns Hopkins University的SafeSeqS。

[0191] 用于将cfDNA测序和映射到参考基因组的技术是本领域已知的,例如参见Chandrananda等人(2015)BMC Medical Genomics 8:29。

[0192] 单参数建模

[0193] 本公开内容提供了用于单参数建模的方法。单参数模型可以包括对2-D分布例如片段计数分布执行2-D分析。单参数模型可以包括基因组中的一组位置。基因组可以是人类基因组。基因组可包含报告的肿瘤标志物的一个或更多个基因座。2-D片段计数分布可包括基因组中的一组位置和与基因组中的该组位置中的每个位置对齐的一组片段的数量。此类建模可以与分类器一起使用(如本文更详细描述的),用于鉴定与状况或状况的状态相关的模式或特征,或者确定测试受试者中的遗传畸变(例如,SNV、CNV、融合或插入/缺失)。单参数模型的其他实例包括但不限于针对2-D起始位置分布、针对2-D终止位置分布或针对2-D片段长度分布的2-D分析。

[0194] 2-D起始位置分布可包括基因组中的一组位置和在基因组中的该组位置中的每个位置处起始的一组片段的数量。

[0195] 2-D终止位置分布可包括基因组中的一组位置和在基因组中的该组位置中的每个位置处终止的一组片段的数量。

[0196] 第一2-D片段长度分布可以包括基因组中的一组位置和与基因组中的该组位置中的每个位置重叠的一组片段的长度。

[0197] 第二2-D片段长度分布可包括一组长度和具有该组长度中的长度的一组片段的数量(例如,如图1E所示)。

[0198] 在一个实例中,单参数模型用于检测来自受试者的无细胞DNA中的SNV。首先,从患有肺癌的受试者的体液样品获得无细胞DNA。对cfDNA片段进行测序以产生片段的多于一个序列读段。每个序列读段被映射到来自人类基因组的一组多于一个参考序列。针对该组参考序列中的每个碱基位置,计数映射到该碱基位置的序列读段的数量,从而产生该组参考序列的2-D片段计数分布。在该组参考序列中,鉴定一个参考序列,使得2-D片段计数分布在该参考序列处异常低(相对于该组中的其他参考序列)。这在生物学上被解释为含有具有上调的基因表达的基因座的参考序列。该参考序列含有EGFR L858R单核苷酸多态性基因座。因此,单参数模型进行“无变体”检测EGFR L858R SNV的存在而不使用参考序列中碱基位置的碱基身份(即,不通过序列中的核苷酸身份变异直接检测SNV)。然后,该SNV检测可用于确定临床诊断、预后、疗法选择、疗法预测、疗法监测等。

[0199] 多参数建模

[0200] 在生成来自样品的序列数据之后,可以执行序列数据的多参数分析以生成多参数模型。多参数分析指同时使用多个参数(数据集)的任何分析。例如,多参数分析可以包括具有n个自变量(具有值$x_1, x_2, \cdots, x_n$)的分布函数(具有函数值y),其中n是为至少2的整数。例如,在一个实例中,多参数分析可以包括沿着基因组生成分布图,该分布图在可映射逐个碱

基轴上(例如,跨基因组的多于一个基因组位置中的每一个)指定跨越该碱基的独特分子的数量和在该碱基处起始的独特分子的数量。作为另一实例,多参数分析可以包括生成与每个输入向量[x_1,x_2,…,x_n]相关的片段的数量(例如,函数值y)的分布图,其中每个x_i是跨序列读段数据的自变量(多于一个n自变量中的一个自变量)。此输入向量的实例可以是这样的输入向量:其中x_1是由cfDNA片段跨越的可映射碱基位置(例如,跨基因组的多于一个此类基因组位置中的可映射碱基位置),并且x_2是cfDNA片段以碱基计的长度(例如,"片段长度")。DNA片段的数量的覆盖值(例如,计数)可以被归一化或未被归一化,因为片段组分析通常包括分析片段的相对分布(例如,相对于不同的受试者、在不同时间点提取的样品、不同的基因组位置或基因基因座等)。

[0201] 参数可以指示以下一个或更多个:(i)与基因组中多于一个碱基位置的每一个对齐的DNA片段的长度,(ii)与基因组中多于一个碱基位置的每一个对齐的DNA片段的数量,以及(iii)在基因组中多于一个碱基位置的每一个处起始或终止的DNA片段的数量。多参数模型可包括两个或更多个此类参数。此类参数可以是归一化的值或未被归一化的值。

[0202] 与单参数建模一样,多参数建模可以产生指示基因组结构变异或不稳定性的簇或区域的模式(例如,作为核小体占位或定位的结果)。

[0203] 可以通过从无细胞核酸样品生成一个或更多个多参数或单参数模型来进行片段组谱分析,从而生成无细胞核酸样品的片段组谱。可以对一个或更多个片段组谱(或片段组数据)进行无监督聚类以揭示一类或更多类不同的异常生物学状态。可以将一个或更多个片段组谱(或片段组数据)合并到分类器中(例如,使用机器学习技术)以确定受试者属于一个或更多个具有临床意义的类别的似然。具有临床意义的类别可以是例如指示异常生物学状态或遗传变体的分类。具有临床意义的类别的实例包括(i)一种或更多种遗传变体的存在或不存在,(ii)一种或更多种癌症的存在或不存在,(iii)一种或更多种规范驱动突变的存在或不存在,(iv)一种或更多种疾病亚型(例如,肺癌分子亚型)的存在或不存在,(v)对癌症或其他疾病、紊乱或异常生物学状态的治疗(例如,药物或疗法)的响应的似然,(vi)拷贝数变异(CNV)(例如,ERBB2扩增)的存在或不存在,或(vii)源自肿瘤微环境的信息(例如,对应于cfDNA片段的起源组织)。

[0204] 可以将一个或更多个片段组谱(或片段组数据)并入分类器中以确定一个或更多个规范驱动突变的存在或不存在的似然。驱动突变可以是一种通过增加克隆的存活或繁殖,为克隆在其微环境中提供选择优势的突变。驱动突变可以是与癌症或另一种异常生物学状态相关的体细胞突变。驱动突变的存在可以指示癌症诊断、具有癌症亚型的受试者的分层、肿瘤负荷、组织或器官中的肿瘤、肿瘤转移、治疗功效或对治疗的耐受性。规范的驱动突变可以是本领域公知的突变,例如,Catalogue of Somatic Mutations in Cancer(COSMIC)(可在URL cancer.sanger.ac.uk/cosmic获得)中列出的突变。规范驱动突变的实例包括肺癌中的表皮生长因子受体(EGFR)外显子19缺失、EGFR外显子19插入、EGFR G719X、EGFR外显子20插入、EGFR T790M、EGFR L858R和EGFR L861Q。关于一个或更多个规范驱动突变的存在或不存在的似然的此类信息可用于诊断受试者(例如,患有肺癌),对具有诊断的受试者进行分层(例如,肺癌的分子亚型),选择用于治疗患有疾病或其他异常生物学状态的受试者的治疗(例如,给定剂量的药物诸如靶向治疗),停止用于治疗患有疾病或其他异常生物学状态的受试者的治疗,改变用于治疗患有疾病或其他异常生物学状态的受试者的

治疗(例如,从第一药物到第二药物,或从第一剂量到第二剂量),或者或对受试者进行进一步的医学测试(例如,成像或活检)。

[0205] 可以将一个或更多个片段组谱(或片段组数据)并入分类器中以确定一种或更多种疾病亚型(例如,受试者中的肺癌分子亚型)的存在或不存在的似然。例如,EGFR T790M和EGFR L858R是肺癌的两种分子亚型。关于一种或更多种疾病亚型的存在或不存在的似然的此类信息可用于诊断受试者(例如,患有肺癌),对具有诊断的受试者进行分层(例如,肺癌的分子亚型),选择用于治疗患有疾病或其他异常生物学状态的受试者的治疗(例如,给定剂量的药物诸如靶向治疗),停止用于治疗患有疾病或其他异常生物学状态的受试者的治疗,改变用于治疗患有疾病或其他异常生物学状态的受试者的治疗(例如,从第一药物到第二药物,或从第一剂量到第二剂量),或者对受试者进行进一步的医学测试(例如,成像或活检)。

[0206] 可将一个或更多个片段组谱(或片段组数据)并入分类器中以确定对受试者的治疗(例如,用于癌症或其他疾病、紊乱或异常生物学状态的药物或疗法)做出响应的似然。例如,治疗可以是靶向治疗,诸如被设计用于治疗EGFR阳性肺癌的酪氨酸激酶抑制剂(TKI)。TKI的实例是厄洛宁(erlonitib)和吉非替尼(gefinitib)。关于对受试者的治疗做出响应的似然的此类信息可用于选择用于治疗患有疾病或其他异常生物学状态的受试者的治疗(例如,给定剂量的药物诸如靶向治疗),停止用于治疗患有疾病或其他异常生物学状态的受试者的治疗,改变用于治疗患有疾病或其他异常生物学状态的受试者的治疗(例如,从第一药物到第二药物,或从第一剂量到第二剂量),或者对受试者进行进一步的医学测试(例如,成像或活检)。

[0207] 可以将一个或更多个片段组谱(或片段组数据)并入分类器中以确定源自肿瘤微环境的信息(例如,对应于cfDNA片段的起源组织)的似然。由于片段组谱可包含来自血液中循环核酸的特征信号(或特征),因此此类特征可包括来自肿瘤细胞、白细胞和其他背景细胞以及肿瘤微环境的聚合信号。肿瘤的细胞生物学和微环境都可能在影响肿瘤生物学和活性方面发挥作用。因此,关于源自肿瘤微环境的信息的似然的此类信息可用于鉴定起源组织(例如,肿瘤效应(tumor activity)在组织或器官中盛行)。可以对此类信息进行去卷积以鉴定子组分(例如,发炎的器官、白细胞、肿瘤、正常的凋亡细胞)。此类子组分信息可用于确定肿瘤所在的组织和/或器官。

[0208] 多参数分析可以由2-D密度图(例如,热图或热图谱)代表,其实例在图2中示出。横轴可以是第一自变量(例如,跨基因组中多于一个基因组区域的基因组位置)。纵轴是第二自变量(例如,cfDNA片段长度)。热图具有多于一种颜色,这些颜色代表跨分布函数值范围的分布函数值(例如,函数值y)的不同分位数。例如,热图可以包括六种颜色(蓝色、青色、绿色、黄色、橙色和红色)中的多于一种颜色,该组中的每种连续颜色分别代表分布函数值范围的第一分位数、第二分位数、第三分位数、第四分位数、第五分位数和第六分位数中的分布函数值。可选地,热图可以包括多于一种离散颜色(例如,蓝色、青色、绿色、黄色、橙色和红色)的连续组合,根据分布函数值范围内的每个热图点的函数值的相对百分位数,每种颜色代表多于一种离散颜色的线性加权组合。此类热图可以是三维的(3-D)。然而,可以使用许多其他用于生成多维的方法。在一些情况下,多参数分析包括同时分析的2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20或多于20个维度。

[0209] 如图2所示,此热图可以揭示由于cfDNA片段分布中的典型模式(图1E),跨基因组位置或片段长度的周期性。在热图的横轴或纵轴上,该周期性可以是约167bp。

[0210] 一个多参数分析生成多参数模型,诸如作为一个实例的热图谱,数据挖掘工具可用于鉴定非随机的系统模式。此类模式可以包括峰高度或峰宽度的关联,其与群组的表型相关,诸如被诊断患有状况(例如,心血管状况、感染、炎症、自身免疫紊乱、癌症、被诊断患有特定类型的癌症、被诊断患有特定癌症阶段等)的群组的表型相关。

[0211] 一旦生成了多参数热图谱,就可以以多种不同方式中的一种方式,例如,使用多元机器学习技术或相对于非恶性群组的2-D密度图的残差变化的直接建模来转换该空间(如图3所示)。例如,人们可以在多参数分析中建立血浆失调的度量(分布函数值y)作为给定基因组位置处的片段丰度(x_1)和片段长度(x_2)的函数。此类功能形式可以简单到(1)归一化覆盖率和片段长度空间中的L2范数,或者可以表示为(2)阴性对照和/或健康供体参考集的二元正态近似。作为后者(2)的实例,血浆失调度量可以是二元正态密度的对数的负值,其中概率轮廓椭圆由数据的一阶矩和二阶矩,例如,使用稳健多元位置和具有高崩溃点的规模估计(也称为快速最小协方差行列式估计器(Fast Minimum Covariance Determinant estimator))确定。

[0212] 为了示出数据转换的一个实施方案,图3A至图3D示出了4种不同的转换的多参数热图谱的实例,该转换的多参数热图谱显示了三组不同基因组位置(两个来自PIK3CA,一个来自EGFR)的血浆失调度量。通过将片段起始和宽度密度转换为跨多于两千个临床样品的血浆失调度量来生成每个热图谱。横轴可以表示外显子归一化的10bp片段起始覆盖率。纵轴可以表示居中中值10bp片段大小。每个临床样品用如下的纯色圆圈表示:健康对照以深绿色显示,并且患有癌症的受试者显示颜色范围为蓝色、青色、黄色、橙色和红色(分别对应于0.1%至93%的最大突变等位基因分数(MAF)值)。实际上,蓝色圆圈可对应于谱图(例如,患有癌症的受试者群组中最大MAF值的范围)的最小值或最低值端,而红色圆圈可对应于谱图(例如,患有癌症的受试者群组中的最大MAF值的范围)的最大值或最高值端。

[0213] 从图3A和图3B中,我们观察到与健康对照(例如,用绿色圆圈表示)相比,对于PIK3CA|2238组基因组位置,具有高的最大MAF的癌症受试者(例如,用红色圆圈表示)倾向于具有居中中值10bp片段大小的较低值和外显子归一化的10bp片段起始覆盖率的较高值。从图3C,我们也观察到与健康对照(例如,用绿色圆圈表示)相比,对于PIK3CA|2663组基因组位置,具有高的最大MAF的癌症受试者(例如,用红色圆圈表示)倾向于具有居中中值10bp片段大小的较高值和外显子归一化的10bp片段起始覆盖率的较低值。从图3D,我们也观察到与健康对照(例如,用绿色圆圈表示)相比,对于EGFR|6101组基因组位置,具有高的最大MAF的癌症受试者(例如,用红色圆圈表示)倾向于具有居中中值10bp片段大小的较高值和外显子归一化的10bp片段起始覆盖率的较高值。对于这3组基因组位置中的每一组,与健康对照相比,在癌症受试者群组中观察到(1)居中中值10bp片段大小的分布的偏移以及(2)外显子归一化的10bp片段起始覆盖率的分布的偏移(例如,x-轴和y-轴上的偏移)。这些由于癌症状态引起的多参数分布的分布偏移的观察结果明显独立于序列读段数据分析(例如,生物信息学分析),并且可以被用作鉴定单核苷酸变体(SNV)、拷贝数变异(CNV)、插入和缺失(插入/缺失)或其他常规遗传畸变的基础(例如,单独或与其他临床观察数据联合使用)。

[0214] 在一个实例中,多参数模型用于通过分析来自受试者的无细胞DNA来检测癌症。首

先,从来自一组患有癌症的多个受试者和未患有癌症的受试者的体液样品获得无细胞DNA。对cfDNA片段进行测序以产生片段的多于一个序列读段。将每个序列读段映射到来自人类基因组的多于一个参考序列的组。如下生成多参数模型:对于一组居中中值10bp片段大小值中的每个值(第一变量),对于一组外显子归一化的10bp片段起始覆盖率值中的每个值(第二变量),以及对于PIK3CA|2663组基因组位置中的每个基因组位置(第三变量),将每个未患有癌症的健康对照受试者的MAF绘制成绿色,并将每个患有癌症的受试者的MAF绘制在代表MAF的色谱上(例如,从蓝色到黄色到橙色到红色增加)。在该多参数模型中,观察到与健康对照(例如,用绿色圆圈表示)相比,具有高的最大MAF的癌症受试者(例如,由红色圆圈表示)倾向于具有居中中值10bp片段大小的较高值和外显子归一化的10bp片段起始覆盖率的较低值。接下来,对于具有未知癌症状态的第一个测试受试者和第二个测试受试者重复上述相同的程序。与第一个测试受试者相关的圆圈落在代表健康对照的范围内(例如,具有一簇绿色圆圈的区域),因此基于该测试,第一个测试受试者被诊断为癌症阴性。与第二个测试受试者相关的圆圈落入代表具有90%的非常高的MAF的患有癌症的受试者的范围内(例如,具有一簇红色圆圈的区域),因此第二个测试受试者被诊断为癌症阳性或基于该测试提交进行进一步的活检测试。由此对来自受试者的cfDNA样品进行多参数模型以检测这些受试者中的癌症。

[0215] 在获得计算的血浆失调度量之前或者在建立血浆失调度量之后,可以将一种或更多种多重过滤技术应用于多参数分布数据。过滤技术可以创建近似函数,该近似函数尝试捕获一组数据(例如,一组粒度数据)中的重要信息、趋势或参数,同时省去噪声或其他精细尺度现象。对于样品,过滤技术可以使得能够从一组数据中提取更多信息或者实现灵活或稳健的分析。样品过滤技术包括移动平均值、全局多项式、样条、数字平滑(例如,巴特沃斯过滤器(Butterworth filter)、傅立叶平滑(Fourier smoothing)等)、Wigner变换、连续小波变换(CWT)和离散小波变换(DWT)。过滤技术还可以涉及经由减去与测定偏倚,例如与靶向捕获相关的富集相关偏倚相关的预定义片段起始覆盖率来去除测定特异性噪声。可以测定代表均匀片段分布的人为样品,并且在此类人为样品中观察到的片段长度富集可以用于校正临床样品信号(例如,通过拟合和/或减去信号的测定相关组分)。可选地或另外地,可以进一步归一化片段计数以校正来自血浆DNA降解的偏倚。此降解可以源于例如处理和储存,并且可以导致预期的片段长度分布的变化和/或污染的基因组DNA的存在。

[0216] 作为实例,图4显示了血浆失调评分的样品,因为它在给定临床样品中跨基因组片段随位置而变化(下图)。上图显示了测定的相关基因列表以及在这些基因中发现的任何改变(SNV或CNV)。血浆失调评分可以是代表局部基因组区域的血浆失调的值。血浆失调评分可以指示其中观察到源自健康细胞的大多数DNA片段组信号的规范包络(canonical envelope)(例如,多参数分布的区域(例如,地区))。可以通过使用非恶性健康对照受试者(没有目的疾病)的训练集并对来自训练集的每个受试者的cfDNA样品进行多参数分析来生成血浆失调评分。接下来,可以鉴定在群组中以指定频率(例如,90%、95%、96%、97%、98%、99%、99.9%、99.99%、99.999%或99.995%)观察到片段的区域。接下来,可以掩蔽这些区域,从而鉴定这些区域之外的密度。接下来,可以将这些密度聚合(或对其求和)以获得血浆失调评分。该血浆失调评分可以指示例如突变负荷、肿瘤负荷或疾病负荷。

[0217] 血浆失调评分的实例可以是无变体覆盖(VCF)评分,其指示覆盖给定基因组区域

或碱基位置的DNA片段的数量。血浆失调评分的低值可指示在局部基因组区域的血浆失调水平相对较低。血浆失调评分的高值可指示在局部基因组区域的血浆失调水平相对较高。血浆失调评分可以由不同颜色代表以指示相对差异(例如,对于在血浆失调评分范围内的多于一个分位数中的每个不同分位数的不同颜色),例如,如在单参数热图(或热图谱)或多参数热图(或热图谱)中所观察到的。

[0218] 再次参考图4,可以观察到血浆失调评分中的许多不同的峰,其对应于许多公认的癌症标志物基因(例如,PIK3CA、MYC、CDKN2A、CCND1、CCND2、KRAS、CDK4、RB1和ERBB2)。血浆失调评分中的不同峰可以与已知的肿瘤标志物相关,所述已知的肿瘤标志物例如在癌症体细胞突变目录(COSMIC)中报告的体细胞突变。

[0219] 通过在大量(例如,数百至数千或更多)临床样品上生成多参数模型,此类多参数模型可以产生包括可以与特定癌症类型相关或者被分析的经验性特征的度量(例如,血浆失调评分),以发现体细胞或其他类型的变体。然后可以将此类信息并入无变体的体细胞变体分类器中。例如,可以分析5,000个非小细胞肺癌(NSCLC)患者样品中跨多个基因组区域的血浆失调评分的无监督聚类,并将其可视化为热图。

[0220] 例如,图5显示了通过在5,000个样品中跨多个基因组区域的血浆失调评分的无监督聚类所生成的热图,每个样品来自不同的非小细胞肺癌(NSCLC)患者。y-轴反映5,000个患者样品中的每一个。x-轴反映了被分析的一组基因组位置。颜色反映了每个样品的每个基因组位置的血浆失调评分。使用无监督聚类算法对整个数据集进行聚类。基于该热图谱,我们可以使用该数据来鉴定可以用作患者的无变体分类的热点的区域。该分类可用于鉴定待被包括在临床试验中、待被给予某种疗法、待被取消疗法处理等的患者。

[0221] 横(较长)轴可以表示跨基因组中多于一个基因组位置的基因组位置。纵(较短)轴可以表示临床样品(例如,每行示出来自一个临床样品的数据)。此类热图可以揭示相对高的血浆失调的区域(例如,在红色区域、橙色区域和黄色区域中)和相对低的血浆失调的区域(例如,在蓝色区域和绿色区域中)。

[0222] 作为多参数模型的另一个实例,可以跨基因组位置(例如,以10碱基对(“bp”)分辨率)生成热图谱,以跨大量临床样品(例如,2000个)可视化单个基因(例如,KRAS),如图6(A部分)所示。横轴可以表示跨基因组中多于一个基因组位置(例如,跨越KRAS基因)的基因组位置。纵轴可以表示临床样品(例如,每行示出来自一个临床样品的数据)。在该分析中,将具有至少一种报告的变体的KRAS无变体覆盖值(VFC)在热图中可视化(图6(A部分))。最高var(可变)分箱(bin)以基因组顺序放置并且覆盖有转录物同种型和mRNA谱(图6(B部分))。

[0223] 从跨大量临床样品的一个或更多个单参数和/或多参数模型生成的血浆失调评分的观察到的特征可以并入在公知的体细胞突变检测和定量方法处理中,以提高此类体细胞突变检测和定量方法的检测灵敏度。例如,在用于检测和定量无细胞核酸例如cfDNA中的拷贝数变异(例如,CNV)的当前方法中,典型的覆盖度量(例如,包含变体的分子数与没有变体的分子的参考数的计算比率)可以通过对应于多参数模型中的偏移的度量来调整或取代。

[0224] 可以对从跨越大量临床样品的一个或更多个单参数和/或多参数模型生成的血浆失调评分的观察到的特征进行聚类,并进行富集分析以产生具有潜在体细胞变化相关的血浆谱。该方法可以导致通过使用无变体血浆失调评分来计算或确定一个或更多个体细胞突变的组(例如,已知肿瘤标志物)存在于从其获得cfDNA样品的患者中的概率似然。

[0225] 可以将从受试者的无细胞DNA样品生成的一个或更多个单参数模型并入分类器(例如,机器学习引擎)中,该分类器被训练为将所述样品分类为具有或不具有一组单核苷酸变体(SNV)或其他遗传变体中的每一个。这些SNV或其他遗传变体可以在选自表1的一种或更多种基因中找到。该分类器可以是无变体分类器(例如,不基于体细胞突变鉴定进行分类)。该分类器可以是变体感知分类器(例如,基于体细胞突变鉴定进行分类)。

[0226] 无变体分类器可确定基因组中基因座处序列畸变的存在或不存在,而不考虑基因组的任何基因座或亚基因座(sub-locus)中的多于一个碱基位置的每一个处的碱基身份,其中所述多于一个碱基身份指示已知的体细胞突变。亚基因座可以是多于一个连续的碱基位置,使得所述多于一个是基因组中基因座的子集。无变体分类器可以使用单参数或多参数分析来确定受试者中基因座中序列畸变的存在或不存在。该基因座可以是报告的肿瘤标志物。该基因座可以是先前未报告的肿瘤标志物。

[0227] 变体感知分类器可以通过考虑在基因组的一个或更多个基因座或亚基因座中的多于一个碱基位置中的每一个处的碱基身份来确定基因组中第一基因座处的序列畸变的存在或不存在,其中所述多于一个碱基身份指示已知的体细胞突变,并且其中第一基因座不在基因组的一个或更多个基因座或亚基因座中。换言之,变体感知分类器可以通过合并关于在基因组中的任何其他基因座处检测到的已知体细胞突变的信息来鉴定给定基因座处的序列畸变。

[0228] 可选地,可以将从受试者的无细胞DNA样品生成的一个或更多个多参数模型并入分类器(例如,机器学习引擎)中,该分类器被训练以将所述样品分类为具有或不具有一组单核苷酸变体(SNV)或其他遗传变体中的每一个。这些SNV或其他遗传变体可以选自表1。该分类器可以是无变体分类器(例如,不基于体细胞突变鉴定进行分类)。该分类器可以是变体感知分类器(例如,基于体细胞突变鉴定进行分类)。多参数模型可以包括一个或更多个数据集,该数据集包括与一个或更多个遗传基因座相关的任何信息,例如,指示选自以下的特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在遗传基因座处起始的DNA序列,(iii)在遗传基因座处终止的DNA序列;(iv)DNA序列的双核小体保护或单核小体保护;(v)位于参考基因组的内含子或外显子中的DNA序列;(vi)具有一个或更多个特性的DNA序列的大小分布;(vii)具有一个或更多个特性的DNA序列的长度分布,或(viii)其任何组合。

[0229] 可选地,可以将从受试者的无细胞DNA样品生成的一个或更多个单参数模型和一个或更多个多参数模型并入分类器(例如,机器学习引擎)中,该分类器被训练以将所述样品分类为具有或不具有一组单核苷酸变体(SNV)或其他遗传变体中的每一个。这些SNV或其他遗传变体可以选自表1。该分类器可以是无变体分类器(例如,不基于体细胞突变鉴定进行分类)。该分类器可以是变体感知分类器(例如,基于体细胞突变鉴定进行分类)。单参数模型可以包括一个或更多个数据集,该数据集包括与一个或更多个遗传基因座相关的任何信息,例如,指示选自以下的特性的定量量度的值:(i)映射到遗传基因座的DNA序列,(ii)在遗传基因座处起始的DNA序列,(iii)在遗传基因座处终止的DNA序列;(iv)DNA序列的双核小体保护或单核小体保护;(v)位于参考基因组的内含子或外显子中的DNA序列;(vi)具有一个或更多个特性的DNA序列的大小分布;(vii)具有一个或更多个特性的DNA序列的长度分布,或(viii)其任何组合。

[0230] 除了诸如血浆失调评分的度量之外,多参数分析还可以揭示受试者的肿瘤相关信

息。在一个实例中,基因组中任何给定位置中的读段的数量可以产生对从其获取无细胞核酸样品的受试者的肿瘤状态的见解,诸如起源组织、肿瘤负荷、肿瘤侵袭性、肿瘤可药性、肿瘤进化和克隆性以及肿瘤对治疗的耐受性。

[0231] 在另一个实例中,基因组中任何给定位置中的读段的数量介入(interposed with)与基因组中该位置处的读段长度,并且可以产生对从其获取无细胞DNA样品的受试者的肿瘤状态的见解,诸如起源组织、肿瘤负荷、肿瘤侵袭性、肿瘤可药性、肿瘤进化和克隆性以及肿瘤对治疗的耐受性。

[0232] 模型中的模式例如,峰的高度、峰的宽度、新峰的出现、峰的偏移和/或拖尾(smear)可以用作表型的指示。在一些情况下,将个体的核小体谱与参考多参数模型或模式进行比较,以确定表型或表型变化。

[0233] 在一个方面,本文公开了一种方法,所述方法用于生成指示来自从受试者获得的无细胞样品(或无细胞DNA)的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出。方法可以包括构建(例如,通过计算机)来自无细胞样品(或无细胞DNA)的DNA片段跨基因组中多于一个碱基位置的分布。接下来,可以使用该分布来确定指示受试者中遗传畸变的存在或不存在的输出。存在或不存在可以在(i)不将DNA片段的分布与来自受试者基因组外部的来源的参考分布进行比较,(ii)不将来源于DNA片段的分布的参数与参考参数进行比较,和/或(iii)不将DNA片段的分布与来自受试者的对照的参考分布进行比较的情况下确定。在一些实施方案中,遗传畸变包括拷贝数变异(CNV)和/或单核苷酸变体(SNV)。在一些实施方案中,分布包括一个或更多个多参数分布。

[0234] 在一个方面,本文公开了一种方法,所述方法用于处理受试者的生物样品的具有双核小体保护的DNA片段和/或具有单核小体保护的DNA片段。该处理可以包括获得受试者的生物样品。生物样品可包含脱氧核糖核酸(DNA)片段。测定可以包括生成指示以下的存在或不存在的信号:(i)具有与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段和/或(ii)具有与该遗传基因座相关的单核小体保护的DNA片段。此类生成的信号可以用于生成指示以下的存在或不存在的输出:(i)具有与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段和/或(ii)具有与该遗传基因座相关的单核小体保护的DNA片段。测定可以包括富集生物样品,以获得一个或更多个遗传基因座的DNA片段的组。此类遗传基因座可包含肿瘤相关的遗传基因座和/或非肿瘤相关的遗传基因座。测定可以包括将生物样品的DNA片段测序。

[0235] 在另一个方面,本文公开了一种方法,所述方法用于生成指示来自从受试者获得的无细胞样品(或无细胞DNA)的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出。生成可以包括构建(例如,通过计算机)来自无细胞样品(或无细胞DNA)的DNA片段的分布(例如,跨基因组中多于一个碱基位置)。接下来,针对一个或更多个遗传基因座中的每一个,可以计算指示以下的定量量度(例如,通过计算机):(1)具有与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护的DNA片段的数量,以及(2)具有与遗传基因座相关的单核小体保护的DNA片段的的数量的比率,或反之亦然。接下来,可以生成指示受试者中一个或更多个遗传基因座中遗传畸变的存在或不存在的输出。该生成可以对一个或更多个遗传基因座中的每一个使用定量量度。在一些实施方案中,分布包括一个或更多个多参数分布。

[0236] 参考模型

[0237] 参考多参数模型可以源自在不同时间点从同一受试者获得的不同样品。此类样品中的一些或全部可包含无细胞DNA。可选地,这些样品中的一种或更多种可以直接源自肿瘤(例如,经由活检或细针抽吸)。源自此类样品的模型可用于监测患者的癌症,观察癌症中的克隆性,检测新的突变和耐药性。

[0238] 参考多参数模型可以源自来自受试者的周围肿瘤微环境的基质组织。例如,可以在活检期间获得用于此类模型的DNA。源自基质组织的模型可用于创建基线多参数模型。这可以允许早期观察肿瘤衍生的无细胞DNA中的新变异。

[0239] 参考多参数模型可以源自来自健康无症状个体的剪切的基因组(非无细胞)DNA。剪切的DNA可用于模拟健康个体的无细胞DNA样品。例如,此类剪切的DNA样品可用于片段组信号的归一化。例如,可以生成剪切的DNA并用于实验中以验证和优化一组一种或更多种探针的捕获效率(例如,在靶向测定中)。

[0240] 参考多参数模型可以源自来自给定组织类型的片段组(例如,核小体)谱。核小体占位谱分析技术的实例包括Statham等人,Genomics Data,第3卷,2015年3月,第94页至96页(2015)。

[0241] 使用参考样品的多参数模型,人们可以确定与凋亡过程和坏死过程相关的片段组(例如,核小体)模式或谱。然后可以独立地或组合地使用此类模式的检测来监测受试者的状况。例如,随着肿瘤扩大,肿瘤微环境中坏死与细胞凋亡的比率可能变化。可以使用本文所述的方法使用片段组谱分析来检测坏死和/或细胞凋亡的此类变化。

[0242] 距离函数可以通过计算(1)受试者的单参数模型或多参数模型与(2)参考单参数模型或多参数模型(例如,健康群体的典型)之间的差异,从片段组谱推导出。

[0243] 片段组特征

[0244] 在一个实例中,具有一种表型的受试者群组(例如,无症状健康个体或具有特定类型癌症的个体)可以使用本文的方法测定其片段组谱。分析群组成员的片段组谱并确定群组的片段组特征。从头测试的受试者可以通过经训练的分类器(经训练的数据库)使用两个或更多个群组的片段组特征将他们的谱分类为一个或更多个类别。

[0245] 个体的群组可以都具有共有特性。该共有特性可选自由以下组成的组:肿瘤类型、炎性状况、凋亡状况、坏死状况、肿瘤复发和对治疗的耐受性。与健康受试者相比,凋亡状况可以是例如通过细胞凋亡导致细胞死亡的似然高于坏死的疾病或状况。凋亡状况可选自由以下组成的组:感染和细胞周转。与健康受试者相比,坏死状况可以是例如通过坏死导致细胞死亡的似然高于细胞凋亡的疾病或状况。坏死状况可选自由以下组成的组:心血管状况、败血症和坏疽。

[0246] 在一些情况下,群组包括具有特定类型癌症(例如,乳腺癌、结肠直肠癌、胰腺癌、前列腺癌、黑素瘤、肺癌或肝癌)的个体。为了获得此类癌症的核小体特征,每个此类个体提供了血液样品。从这些血液样品获得无细胞DNA。对此类群组的无细胞DNA进行测序(有或没有选择性富集来自基因组的一组区域)。将来自测序反应的序列读段形式的序列信息映射到人类基因组。任选地,在映射操作之前或之后,将分子叠并成独特的分子读段。

[0247] 由于给定样品中的无细胞DNA片段代表产生无细胞DNA的细胞混合物,因此来自每种细胞类型的差异性核小体占位可导致对代表给定无细胞DNA样品的数学模型的贡献。例

如,由于跨不同细胞类型或者跨肿瘤对比非肿瘤细胞的核小体保护差异,可能出现片段长度的分布。方法可用于基于序列数据的单参数、多参数和/或统计分析来开发一组临床上有用的评估。

[0248] 该模型可以用于组配置(panel configuration)以选择性地富集区域(例如,片段组谱相关的区域)并确保跨越特定突变的大量读段,也可考虑重要的染色质居中事件如转录起始位点(TSS)、启动子区域、交界位点和内含子区域。

[0249] 例如,在内含子和外显子的连接处(或边界)处或附近发现片段组谱的差异。一个或更多个体细胞突变的鉴定可以与一个或更多个多参数模型或单参数模型相关,以揭示cfDNA片段分布的基因组位置。该相关性分析可揭示片段组谱中断最明显的一个或更多个内含子-外显子连接。例如,片段组谱中断可能是由于不同的蛋白质同种型被表达,导致结合位点被改变,从而改变了cfDNA片段的核小体保护,这可以凭经验被观察为在内含子-外显子连接处cfDNA片段的差异特征和分布,此处内含子-外显子连接的特定位置与同种型的起始相关。内含子-外显子边界可以被包括在组配置中以选择性地富集这些区域,这可以给出疾病或其他异常生物学状态的更好的判别(例如,差异似然的确定)。这种方法可以通过关注外显子-内含子连接来代替整个外显子区域,或者除了整个外显子区域之外关注外显子-内含子连接,改进组设计。

[0250] 片段组谱可以与现有的体细胞突变组组合。在一些情况下,将SNV信息与片段组谱分析组合使用可以增加SNV调用的灵敏度或准确度。例如,如果某个SNV主要存在于比平均更短的片段中(例如,长度小于155bp、154bp、153bp、152bp、151bp、150bp、149bp或148bp),则SNV更可能是体细胞突变。如果发现SNV主要存在于比平均更长的片段中(例如,超过155、156、157、158、159、160、161、162、163、164、165或166),则SNV更可能是种系SNV。因此,本公开内容的测定可包括确定来自无细胞DNA样品的独特分子中的SNV以及每个独特分子的片段大小,并且基于包括SNV的独特分子的大小分布调整体细胞SNV的调用的置信度评分。

[0251] 片段组谱分析分析可以包括进行代表受试者的无细胞DNA的单参数分析或多参数分析。根据给定受试者的序列数据,可以生成跨参考基因组中的每个碱基位置的一个或更多个预期分布,其中每个预期分布描述以下中的一个或更多个:映射到给定位置的读段的数量、映射到给定位置的无细胞DNA片段长度、在给定位置处起始的无细胞DNA片段的数量、以及在给定位置处终止的无细胞DNA片段的数量。

[0252] 通过在基因组的给定基因座处进行样品和参考之间的逐个碱基对比较,对该模式的任何偏差的观察(例如,在给定碱基位置处读段的数量比预期的增加或减少,或者分布的偏移)揭示肿瘤相关信息,诸如肿瘤负荷、肿瘤类型、肿瘤克隆性或异质性、肿瘤侵袭性等。此类偏差是核小体定位变异和细胞过程的下游后果。

[0253] 例如,诸如感染、炎症和肿瘤生长和侵袭性的异常细胞过程影响凋亡和坏死途径的相对贡献以将DNA脱落到血流中,在血流中无细胞DNA片段循环并作为血液样品的一部分被收集,用于液体活检应用。由于细胞凋亡过程跨核小体切割,这些过程可产生其中存在核小体的更长的读段(例如,更长的片段)。由于肿瘤细胞中的核小体保护不同于正常细胞,因此可以跨群组观察到不同的数据模式,例如癌症与正常之间或者两种肿瘤类型之间。

[0254] 为了进行片段组谱分析的分析,可以从收集自受试者的血液样品提供无细胞DNA分子的集合。无细胞DNA可以是短片段的形式(其中大多数长度小于200个碱基对)。可以对

无细胞DNA进行文库制备和高通量测序,以生成代表来自样品的无细胞DNA分子的序列信息。在对齐后,可以对对齐的序列信息进行多参数分析,以生成代表来自样品的无细胞DNA分子的多参数模型。

[0255] 可以使用所述序列信息对两个数据集的组执行单参数分析,以生成代表来自样品的无细胞DNA分子的单参数模型,其中单参数模型具有二个维度。数据集可以包括定量值的向量。单参数模型可以包括两个数据集,例如,使得一个数据集构成y-轴而一个数据集构成x-轴。

[0256] 可以使用所述序列信息对三个或更多个数据集的多于一个进行多参数分析,以生成代表来自样品的无细胞DNA分子的多参数模型,其中多参数模型具有三个或更多个维度。多参数模型可以包括三个数据集,例如,使得一个数据集构成z-轴(或阴影颜色),一个数据集构成y-轴,并且一个数据集构成x-轴。

[0257] 选择用于单参数或多参数分析的数据集可以选自由以下组成的组:(a)测序的片段的起始位置,(b)测序的片段的终止位置,(c)覆盖可映射位置的独特的测序的片段的数量,(d)片段长度,(e)可映射碱基对位置将出现在测序的片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的片段内的似然,以及(g)测序的片段的序列基序。序列基序是位于片段的末端处的长度为2至8个碱基对的序列,其可用于鉴定序列信息中的模式,并且可以被并入分类方案中。

[0258] 单参数分析可以包括将一个参数映射到基因组的两个或更多个位置或区域中的每一个。该参数可以选自由以下组成的组:(a)测序的片段的起始位置,(b)测序的片段的终止位置,(c)覆盖可映射位置的独特的测序的片段的数量,(d)片段长度,(e)可映射碱基对位置将出现在测序的片段的末端处的似然,以及(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的片段内的似然。基因组的这两个或更多个位置或区域可包括至少一个与表1中列出的目的基因中的一个或更多个相关的区域。

[0259] 多参数分析可以包括将两个或更多个参数映射到基因组的两个或更多个位置或区域中的每一个。这些参数可以选自由以下组成的组:(a)测序的片段的起始位置,(b)测序的片段的终止位置,(c)覆盖可映射位置的独特的测序的片段的数量,(d)片段长度,(e)可映射碱基对位置将出现在测序的片段的末端处的似然,以及(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的片段内的似然。基因组的这两个或更多个位置或区域可包括至少一个与表1中列出的目的基因中的一个或更多个相关的区域。

[0260] 表1

[0261]

点突变 (SNV)						扩增 (CNV)		融合	插入/缺失
AKT1	ALK	APC	AR	ARAF	ARID1A	AR	BRAF	ALK	EGFR (外显子 19 & 20)
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	CCND1	CCND2	FGFR2	
CCNE1	CDH1	CDK4	CDK6	CDKN2A	CDKN2B	CCNE1	CDK4	FGFR3	
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	CDK6	EGFR	NTRK1	ERBB2 (外显子 19 & 20)
FGFR1	FGFR2	FGFR3	GATA3	GNA11	GNAQ	ERBB2	FGFR1	RET	
GNAS	HNF1A	HRAS	IDH1	IDH2	JAK2	FGFR2	KIT	ROS1	
JAK3	KIT	KRAS	MAP2K1	MAP2K2	MET	KRAS	MET		MET (外显子 14 跳跃)
MLH1	MPL	MYC	NF1	NFE2L2	NOTCH1	MYC	PDGFRA		
NPM1	NRAS	NTRK1	PDGFRA	PIK3CA	PTEN	PIK3CA	RAF1		
PTPN11	RAF1	RB1	RET	RHEB	RHOA				
RIT1	ROS1	SMAD4	SMO	SRC	STK11				
TERT	TP53	TSC1	VHL						

[0262] 无细胞DNA可包含代表其潜在染色质组织的印迹,其可捕获以下中的一种或更多种:表达控制核小体占位、RNA聚合酶II暂停、细胞死亡特异性DNA酶超敏性和细胞死亡期间染色质凝聚。此类印迹可以携带细胞碎片清除和运输的特征,例如,通过细胞凋亡垂死的细胞中由半胱天冬酶活化的DNA酶(CAD)进行的DNA片段化,但也可以在垂死细胞被吞噬后由溶酶体DNA酶II进行,产生不同的裂解图谱。基因组分区图谱可以经由将重要窗口聚合成目的区域,通过对在与上述染色质特性相关的恶性和非恶性状况中不同染色质状态进行基因组范围鉴定来构建。这些目的区域通常称为基因组分区图谱。

[0263] 基因组的两个或更多个位置或区域可以通过以下来鉴定:(i)提供一个或更多个基因组分区图谱,以及(ii)从基因组分区图谱选择基因组的位置或区域,基因组的每个此类位置或区域映射到目的基因。基因组的两个或更多个位置或区域的长度可以各自在2至500个碱基对之间。基因组的这些位置或区域代表与目的基因相关的局部基因组区域,用于进一步分析。

[0264] 多参数分析可以包括生成基因组的两个或更多个区域的热图谱。该热图谱可以给出两个或更多个参数如何跨给定基因组的位置变化的视觉代表。基因组的两个或更多个区域可以包括选自表1中列出的一个或更多个基因的至少一个区域。代表群组内或跨群组的大量(例如,超过100个)受试者的热图谱可以被组合以生成代表受试者所属的给定群组或群组的组的一个或更多个参考热图谱。例如,群组可以包括共有特性的受试者,所述共有特性例如,诊断的疾病(例如,肿瘤类型)、共同的疾病状态(例如,健康对照)或共同的疾病结果(例如,肿瘤复发或对治疗的耐受性)。

[0265] 多参数分析还可包括应用一个或更多个数学变换以生成多参数模型。多参数模型

可以是选自由以下组成的组的两个或更多个变量的联合分布模型:(a)测序的片段的起始位置,(b)测序的片段的终止位置,(c)覆盖可映射位置的独特的测序的片段的数量,(d)片段长度,(e)可映射碱基对位置将出现在测序的片段末端处的似然,(f)由于差异性核小体占位,可映射碱基对位置将出现在测序的片段内的似然,以及(g)序列基序。根据多参数模型,可以鉴定一个或更多个峰。每个此类峰可以具有峰分布宽度和峰覆盖率。

[0266] 代表群组内或跨群组的大量(例如,至少50、100、200、300、500、700、1000、2000、3000、5000或更多)受试者的单参数模型或多参数模型可以被组合以分别生成代表受试者所属的给定群组或群组的组的一个或更多个参考单参数模型或多参数模型。例如,群组可包括具有诊断的共同疾病(例如,肿瘤类型)、共同疾病状态(例如,健康对照)或共同疾病结果(例如,肿瘤复发)的受试者。

[0267] 单参数分析或多参数分析还可包括测量无细胞DNA分子的RNA表达。单参数分析或多参数分析还可包括测量无细胞DNA分子的甲基化。单参数分析或多参数分析还可包括测量无细胞DNA分子的核小体映射。由于差异性核小体占位与测序的片段的鸟嘌呤-胞嘧啶(GC)含量有关,因此可以间接评估甲基化水平,例如,通过检查可以从核小体占位推断甲基化抑制的TSS区域。在这些区域中,由于甲基化(例如,由于组蛋白周围的不同缠绕),可以观察到峰的覆盖率和/或宽度的变化。类似地,可以间接评估cfDNA分子的核小体映射。

[0268] 单参数分析或多参数分析还可包括鉴定无细胞DNA分子中一个或更多个体细胞单核苷酸变体(SNV)的存在。单参数或多参数分析还可包括鉴定无细胞DNA分子中一个或更多个种系单核苷酸变体(SNV)的存在。

[0269] 可以将一个基因组参数并入单参数分析中。可以将一个或更多个基因组参数并入多参数分析中。基因组参数可以选自:(i)组织类型,(ii)基因表达模式,(iii)转录因子结合位点(TFBS)占据,(iv)甲基化位点,(v)可检测体细胞突变的组,(vi)可检测体细胞突变水平,(vii)可检测种系突变的组,以及(viii)可检测种系突变水平。

[0270] 可以检测与参考单参数或多参数模型的偏差。此类偏差可能包括:(i)核小体区域外的读段数量的增加,(ii)核小体区域内的读段数量的增加,(iii)相对于可映射基因组位置的更宽的峰分布,(iv)峰位置的偏移,(v)新峰的鉴定,(vi)峰的覆盖深度的变化,(vii)峰周围的起始位置的变化,以及(viii)与峰相关的片段大小的变化。这些偏差可以指示代表源自样品的无细胞DNA的核小体图谱中断。

[0271] 局部基因组区域是基因组的短区域,其长度范围可为约2至约200个碱基对。每个局部基因组区域可含有显著结构变异或不稳定性的模式或簇。可以提供基因组分区图谱以鉴定相关的局部基因组区域。局部基因组区域可含有显著结构变异或结构不稳定性的模式或簇。簇是局部基因组区域内的热点区域。热点区域可含有一个或更多个显著波动或峰。结构变异是核小体定位的变异。结构变异可以选自由以下组成的组:插入、缺失、易位、基因重排、甲基化状态、微卫星、拷贝数变异、拷贝数相关的结构变异或任何其他指示分化的变异。

[0272] 可以通过以下方式获得基因组分区图谱:(a)提供来自群组中的两个或更多个受试者的无细胞DNA样品,(b)对每个无细胞DNA样品进行多参数分析以为所述样品中的每一个生成多参数模型,以及(c)分析多参数模型以鉴定一个或更多个局部基因组区域,该基因组区域中的每一个含有显著结构变异或不稳定性的模式或簇。

[0273] 提供了一种用于分析包含源自受试者的无细胞DNA的样品的方法,其中获得了代

表来自样品的无细胞DNA分子的序列信息,并对所述序列信息进行统计分析以将一个或更多个单参数模型的组分类为与代表不同群组的一个或更多个核小体占位谱相关。

[0274] 提供了一种用于分析包含源自受试者的无细胞DNA的样品的方法,其中获得了代表来自样品的无细胞DNA分子的序列信息,并对所述序列信息进行统计分析以将多参数模型分类为与代表不同群组的一个或更多个核小体占位谱相关。

[0275] 统计分析可以包括提供一个或更多个基因组分区图谱,其列出代表目的基因的相关基因组间隔以供进一步分析。统计分析还可包括基于基因组分区图谱选择一个或更多个局部基因组区域的组。统计分析还可包括分析该组中的一个或更多个局部基因组区域以获得一个或更多个核小体图谱中断的组。统计分析可以包括以下中的一个或更多个:模式识别、深度学习和无监督学习。

[0276] 核小体图谱中断是根据生物学相关信息表征给定的局部基因组区域的测量值。核小体图谱中断可以与选自由以下组成的组的驱动突变相关:野生型、体细胞变体、种系变体和DNA甲基化。

[0277] 可以使用一个或更多个核小体图谱中断来将单参数模型或多参数模型分类为与代表不同群组的一个或更多个核小体占位谱相关。这些核小体占位谱可以与一个或更多个评估相关。评估可以被认为是治疗性干预的一部分(例如,治疗选项、治疗选择、通过活检和/或成像的进一步评估)。

[0278] 评估可以选自由以下组成的组:适应症、肿瘤类型、肿瘤严重程度、肿瘤侵袭性、肿瘤对治疗的耐受性和肿瘤克隆性。可以通过观察跨样品中无细胞DNA分子的核小体图谱中断的异质性来确定肿瘤克隆性的评估。确定两个或更多个克隆中的每一个的相对贡献的评估。

[0279] 可以将疾病评分确定为从其获得无细胞DNA样品的受试者的健康状态指标。该疾病评分可以根据以下中的一个或更多个来确定:(i)一种或更多种评估,(ii)与疾病相关的一种或更多种健康参考多参数模型,以及(iii)与疾病相关的一种或更多种患病参考多参数模型。

[0280] 基因组分区图谱可以应用于选择一组结构变异。结构变异的选择可以根据以下中的一个或更多个:(i)与一种或更多种疾病相关的一种或更多种参考多参数模型,(ii)一种或更多种探针靶向结构变异的效率,以及(iii)关于基因组部分的先前信息,其中结构变异的预期频率高于跨基因组的结构变异的平均预期频率。

[0281] 分析一种或更多种无细胞DNA样品的方法可以被应用于配置多模块组(multi-modular panel)。该多模块化的组配置可以包括分析以下中的一个或更多个:(i)一个或更多个体细胞突变,(ii)人类基因组中核小体位置分布的信息,以及(iii)关于源自正常组织或细胞类型以及源自含有体细胞突变的组织或细胞类型的无细胞DNA分子中的覆盖偏倚的先前信息。在上述分析之后,多模块组配置还可以包括选择包括以下的一个或更多个的组以包含在多模块组中:(i)一个或更多个结构变异,其中至少一个指示在从其获取无细胞DNA样品的受试者中存在一种或更多种疾病的似然增加,(ii)一种或更多种体细胞突变,其中至少一种指示在从其获取无细胞DNA样品的受试者中存在一种或更多种疾病的似然增加,以及(iii)一个或更多个染色质居中事件。该染色质居中事件可包括转录起始位点、启动子区域、交界位点和内含子区域中的一个或更多个。

[0282] 分析一种或更多种无细胞DNA样品的方法可用于检测或监测状况。状况的此类检测或监测可包括从样品中获得代表无细胞DNA分子的序列信息;并且使用与所述分子有关的宏观尺度信息(例如,除碱基身份之外的信息)来检测或监测所述状况。

[0283] 分析一种或更多种无细胞DNA样品的方法可以应用于基于多参数模型检测绝对拷贝数(CN)相关的结构变异。CN相关的结构变异代表基于基因组分区图谱,多参数模型的相对更高或更低偏差的区域。CN相关的结构变异可以代表一种或更多种核小体图谱中断,以确定一种或更多种评估,例如肿瘤负荷或肿瘤类型。通过适当的健康参考单参数模型或多参数模型和患病参考单参数模型或多参数模型,受试者的单参数模型或多参数模型中的偏差可以被解释为核小体图谱中断。可以组合这些核小体图谱中断中的一个或更多个以确定一种或更多种评估,例如肿瘤异质性。

[0284] 组配置

[0285] 本文所述的片段组谱分析技术可进一步被用于模块组配置。此类模块组配置允许设计一组探针或引诱物,其选择性地富集与核小体谱分析相关的基因组区域。通过并入这种“片段组认知”或“核小体认知”,可以收集来自许多个体的序列数据以优化模块组配置的程序,例如,确定靶向哪些基因组位置以及用于这些基因组位置的探针的最佳浓度。

[0286] 例如,染色质结构的变化,例如转录起始位点(TSS)处的核小体重新定位或拓扑相关的结构域架构的中断,可能在基因转录的调节中发挥不可或缺的作用,并且已经与人类健康的许多方面包括疾病相关。因此,比较非恶性群组与恶性群组之间的全基因组染色质可及性可以允许鉴定发展伴随疾病的工具性表观遗传变化的位置。例如,通过对核小体占位、染色质可及性、转录因子结合位点和DNA酶灵敏度图谱的公共图集的研究,以及在非恶性和恶性病例(例如受试者)的代表性群组中直接发现从头的差异性染色质架构(例如,经由全基因组测序(WGS)),可以产生富含染色质标志物的聚焦印迹。此类染色质标志物可以对某些组织、细胞类型、细胞死亡类型和恶性类型(例如,肿瘤类型)具有特异性,并且可以经由靶向富集测定以足够的分辨率和覆盖率进行靶向。

[0287] 通过并入体细胞变异和结构变异以及不稳定性的知识,探针、引诱物或引物的组可以被配置成用结构变异或不稳定性的已知模式或簇来靶向基因组的特定部分(“热点”)。例如,序列数据的统计分析揭示了一系列累积的体细胞事件和结构变异,并且从而使克隆进化研究成为可能。数据分析揭示了重要的生物学见解,包括跨群组的差异性覆盖率、指示存在某些肿瘤子集的模式、具有高体细胞突变负荷的样品中的外来结构事件,以及归因于血细胞与肿瘤细胞的差异性覆盖。

[0288] 在另一个实例中,片段组谱分析可以被应用于对一个或更多个基因生成低多重聚合酶链式反应(PCR))组。可以通过以下生成低多重PCR组:(a)提供一个或更多个基因组分区图谱;(b)提供覆盖一个或更多个基因组分区图谱中的一个或更多个局部基因组区域的多于一个探针;以及(c)从多于一个探针选择一种或更多种具有最佳PCR性能的探针,其中所述探针中的每一个覆盖与每个基因相关的给定局部基因组区域。

[0289] 通过与每个基因相关的探针的最大覆盖深度来测量最佳PCR性能的评估。因此,针对每个基因,可以选择一种或更多种最佳探针以包含在PCR组中。

[0290] 在一个实例中,低多重PCR组包含至少1、2、3、4、5或6个基因,其中该组的任何子集可以同时组合成单个多重PCR测定。低多重PCR组可用于对无细胞DNA或无细胞RNA分子进行

选自由以下组成的组的测定:数字PCR、液滴数字PCR、定量PCR和逆转录PCR。由于低多重PCR测定不具有跨给定目的基因拼贴多个探针和引物的能力,因此使用此类优化的组将确保选择少量探针的最佳集合用于包含在PCR组中。

[0291] 分类

[0292] 本文的方法和系统可以应用于分类器。分类器可以是经训练的或未经训练的。分类器用于鉴定与状况或状况的状态相关的模式。分类器可以在计算机上实施。

[0293] 在一个方面,分类器可以使用来自从测试受试者获得的无细胞样品(或无细胞DNA)的DNA来确定测试受试者中的遗传畸变。该分类器可以包括(a)来自受试者的一个或更多个样品(或无细胞DNA)中的每一个的一组分布评分的输入,其中每个分布评分代表映射到基因组中多于一个位置中的每一个的存在于来自受试者的无细胞样品(或无细胞DNA)的DNA中的碱基的数量;以及(b)一种或更多种遗传畸变的分类的输出。

[0294] 分类器可以包括机器学习引擎。分布评分可以代表碱基位置从其映射的每个分子的长度。分布评分可以代表与碱基位置重叠的每个分子的计数。分布评分可以代表在碱基位置处起始的每个分子的计数。分布评分可以代表在碱基位置处终止的每个分子的计数。

[0295] 分类器可以用于通过为测试受试者提供一组分布评分并使用分类器生成测试受试者的分类,来使用来自从测试受试者获得的无细胞样品(或无细胞DNA)的DNA确定测试受试者中的遗传畸变。

[0296] 可以通过训练集训练分类器。训练集可以包括针对来自受试者的多于一个样品中的每一个的一组分布评分和针对多于一个样品中的每一个的一组分类。该组分布评分可以包括(a)来自对照受试者的多于一个样品中的每一个的一组参考分布评分,其中每个参考分布评分代表来自对照受试者的无细胞样品(或无细胞DNA)的DNA中存在的碱基的数量,所述碱基映射到基因组中的多于一个位置中的每一个,或者(b)来自具有观察到的表型的受试者的多于一个样品中的每一个的一组表型分布评分,其中每个表型分布评分代表来自具有观察到的表型的受试者的无细胞样品(或无细胞DNA)的DNA中存在的碱基的数量,所述碱基映射到基因组中多于一个位置中的每一个。该组分类可以包括(c)来自对照受试者的多于一个样品中的每一个的一组参考分类,或者(d)来自具有观察到的表型的受试者的多于一个样品中的每一个的一组表型分类。

[0297] 与该组参考分布评分或该组参考分类相关的对照受试者可以是无症状的健康个体。具有与该组表型分布评分或该组表型分类相关的观察到的表型的受试者可包括(a)患有组织特异性癌症的受试者,(b)患有特定癌症阶段的受试者,(c)患有炎性状况的受试者,(d)对癌症无症状但患有将进展为癌症的肿瘤的受试者,或者(e)患有对特定药物或药物方案具有阳性或阴性响应的癌症的受试者。

[0298] 分类器还可以包括在基因组的一个或更多个基因座处的一组遗传变体的输入。该组遗传变体可包含报告的肿瘤标志物(例如,COSMIC中报告的肿瘤标志物)的一个或更多个基因座。

[0299] 提供了一种用于创建训练分类器的方法,包括:(a)提供多于一个不同的类别,其中每个类别代表具有共有特性的一组受试者(例如,来自一个或更多个群组);(b)提供代表来自属于每个类别的多于一个样品中的每一个的无细胞DNA分子的单参数模型或多参数模型,从而提供训练数据集;以及(c)在训练数据集上训练学习算法以创建一个或更多个经训

练的分类器,其中每个经训练的分类器将测试样品分类为多于一个类别中的一个或更多个。

[0300] 作为实例,经训练的分类器可以使用选自由以下组成的组的学习算法:随机森林、神经网络、支持向量机和线性分类器。多于一种不同类别中的每一个可选自由以下构成的组:健康、乳腺癌、结肠癌、肺癌、胰腺癌、前列腺癌、卵巢癌、黑素瘤和肝癌。

[0301] 可以将经训练的分类器应用于对来自受试者的样品进行分类的方法。该分类方法可以包括:(a)提供代表来自受试者的测试样品的无细胞DNA分子的一组一个或更多个单参数模型;以及(b)使用经训练的分类器对测试样品进行分类。在将测试样品分类为一个或更多个类别之后,基于样品的分类对受试者进行治疗性干预。

[0302] 可以将经训练的分类器应用于对来自受试者的样品进行分类的方法。该分类方法可以包括:(a)提供代表来自受试者的测试样品的无细胞DNA分子的多参数模型;以及(b)使用经训练的分类器对测试样品进行分类。在将测试样品分类为一个或更多个类别之后,基于样品的分类对受试者进行治疗性干预。

[0303] 图8和图9各自示出了可以并入多参数模型中的一个方面,特别是基因组范围内的每个基因组位置处的片段频率的图。在每个图中,由于核小体定位差异,片段频率随着基因组位置而波动。在图8中,半周期线显示了基因组位置(x-轴)上的平均片段频率(y-轴),这示出了由于差异性核小体占位而导致的变化的片段组信号。在图9中,两条半周期线分别显示了基因组位置(x-轴)上的规范片段起始分布(y-轴)和源自给定位置处的片段的中值肿瘤负荷(y-轴),这示出了由于差异性核小体占位导致的变化的片段组信号和源自在较低规范片段起始分布的位置处的给定位置处的片段的较高中值肿瘤负荷两者。

[0304] 图10和图11示出了多参数模型的两个方面,特别是在基因组范围内的每个基因组位置处的归一化分子计数(上图)和归一化片段大小(即,长度;下图)的图。在每个图中,由于核小体定位差异,分子的归一化计数和归一化的片段大小二者随着基因组位置而波动。

[0305] 图12示出了多参数模型的三个方面,特别是在基因组范围内的每个基因组位置处的归一化分子计数、归一化片段大小(即,长度)以及归一化双链的百分比。由于核小体定位差异,多参数模型的所有三个方面随着基因组位置而波动。特别地,该波动在多参数模型中显示出一定的周期性。该周期通常为约10.5个碱基对。

[0306] 图13示出了多参数模型的一个方面,特别是基因组范围内的每个基因组位置(x-轴)处的读段计数(y-轴)。该基因组范围对应于几种肿瘤相关基因,包括NF1、ERBB2、BRCA1、MET、SMO、BRAF、EGFR和COK6。

[0307] 图14示出了可以作为多参数分析的一部分被执行以生成多参数模型的数学变换的实例。特别地,应用快速傅立叶变换(FFT)以通过基因组范围内的每个基因组位置处的起始位置生成读段计数的图。该基因组范围对应于几种肿瘤相关基因,包括NF1、ERBB2、BRCA1和TP53。如所示,特别地,ERBB2基因表现出比所指示的其他基因显著更高(约两倍或更多倍)的读段计数值,这指示可能存在ERBB2突变。

[0308] 图15示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。特别地,该基因组区域对应于肿瘤相关基因TP53。从对应于具有肿瘤的受试者(下图)的多参数模型(在这种情况下,热图谱),可以观察到相对于没有肿瘤的受试者(上图)的偏差,特别是在外显子9标记的区域附近。此类偏差包括热图谱的不太平滑的拓扑图和更多可变区

域(例如,峰)的存在。

[0309] 图16示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。特别地,该基因组区域对应于肿瘤相关基因NF1。从对应于具有肿瘤的受试者(下图)的多参数模型(在这种情况下,热图谱),可以观察到相对于没有肿瘤的受试者(上图)的偏差。此类偏差包括热图谱的不太平滑的拓扑图和更多可变区域(例如,峰)的存在。

[0310] 图17示出了在基因组的给定区域中的两个不同受试者的两个多参数模型的实例。特别地,该基因组区域对应于肿瘤相关基因ERBB2。从对应于具有肿瘤的受试者(下图)的多参数模型(在这种情况下,热图谱),可以观察到相对于没有肿瘤的受试者(上图)的偏差。此类偏差包括热图谱的不太平滑的拓扑图和更多可变区域(例如,峰)的存在。

[0311] 图18和图19示出了基因组的给定区域中核小体组织与基因组位置的关系的实例。特别地,每个图示出了跨不同的受试者(y-轴)上测量的不同人类染色体(图18中的染色体19和图19中的染色体20)中的核小体组织(由阴影颜色表示的覆盖率)与基因组位置(x-轴)的关系。图18和图19示出了,可以跨群组中的不同受试者中观察到片段组信号的类似簇,而不管这些基因组区域中的碱基身份。

[0312] 图20示出了用于确定绝对拷贝数(CN)的过程的实例。首先,定位核小体位置并将其与正常群组中预期的匹配。然后,对于FGFR中的每个核小体窗口,确定超保守的非chr10核小体位点的集合并确定超保守的chr10核小体位点的集合。最后,对FGFR核小体位点的位置与插入物大小密度进行整合。

[0313] 图21A和图21B示出了通过血浆DNA的全测序使用片段组谱推断拷贝数扩增基因的活化的实例。图21A显示了2,076个临床样品中ERBB2中归一化的双核小体与单核小体计数的比率的图。通过对该热图谱的视觉检查,可以在正常至低扩增活性的背景(例如,以绿色2102显示)下观察到高扩增活性的区域(例如,以黄色2104和红色2106显示)。图21B显示了图21A的图右侧的放大部分,显示了在绿色或蓝色2112的背景下富含高振幅CNV调用的簇(例如,如以黄色2114和红色2116显示)。图21B的下图显示通过相似的片段组信号聚合在一起的基因组区域(例如,由于对应于共同基因基因座的基因组区域的连续部分)。

[0314] 对于每个临床样品,仅切下ERBB2片段(例如,映射至ERBB2基因的cfDNA片段)并进行片段组谱分析。众所周知,ERBB2是某些类型癌症诸如乳腺癌和胃癌的标志物,并且为患有癌症的受试者的治疗耐受性的标志物。对于每个临床样品,跨ERBB2基因组结构域(例如基因组区域)通过以下确定双核小体对单核小体的计数比率:(1)计数具有双核小体保护的片段(例如,至少240个碱基对(“bp”)的片段大小)的数量,(2)计数具有单核小体保护的片段(例如,小于240碱基对(“bp”)的片段大小)的数量,(3)获取(1)与(2)的比率,以及(4)将该比率对样品中值(例如,跨样品的此类比值的中值)归一化。然后,对于每个临床样品,用与该样品相关的CNV测量值绘制样品的双核小体与单核小体计数比率(例如,每个扩增调用显示为紫色点;上图)。

[0315] 在正常至低扩增活性的背景(例如,以绿色2102显示)下,该数据图跨2,076个临床样品的无监督聚类揭示存在3个高扩增活性簇(如由读段计数表示的最高片段组信号所指示)(例如,以黄色2104和红色2106显示),其中右侧的一个看起来最明显。该簇富含高幅度CNV调用,而其他簇则跨中间的簇拖尾,并且跨右侧的簇则较少。簇可以被解释为拷贝数扩增的基因(例如,与ERBB2相关的基因)已经被活化用于与可见簇相关的临床样品的指示(例

如,红色和黄色)。因此,片段组谱(例如,在ERBB2中)可以与扩增状态相关。甚至对于没有相关的高振幅CNV调用的基因组区域也可以进行此类观察(可能是因为循环肿瘤DNA(例如,ctDNA)的低灵敏度,这使得只能进行有限的检测)。这些观察结果可以解释为指示那些基因组区域活跃转录片段组谱分析的基因(例如ERBB2)的似然更高。可以将此类片段组谱分析并入现有的CNV检测方法中(例如,通过进行液体活检测定)以提高灵敏度和特异性。可以跨多于一个基因进行类似的分析,以观察多于一个基因中拷贝数扩增的相对高和低的活化。

[0316] 图21A和图21B的结果显示,cfDNA片段可通过进行包括分析片段大小和片段位置的片段组谱分析揭示对癌细胞的肿瘤微环境的见解。在这种情况下,在肿瘤微环境中从细胞主动脱落的拷贝数扩增基因(例如,ERBB2)的活化可以被观察为独立于进行高振幅CNV调用的ERBB2双核小体保护特征。该方法可能优于现有的CNV检测和调用方法,因为后者很难在循环肿瘤DNA(例如ctDNA)中灵敏地检测到,因为通常在循环中存在低等位基因分数。此类片段组方法也可适用于测量和预测其他遗传变异体诸如SNV、插入/缺失和融合的存在,特别是当此类遗传变体不会导致可观察到的表型差异时。跨患有共同疾病的群组中的受试者的片段组谱分析,例如,相对于正常样品的不同维度(片段长度、位置)的位置、片段长度或距离函数的结合可揭示群组内的分子亚型(例如,肺癌患者群组中肺癌的不同分子亚型),从而对群组中的受试者进行分层。

[0317] 核小体片段长度中的差异的测定

[0318] 本文公开了一种用于处理受试者的生物样品的方法,包括(a)获得所述受试者的所述生物样品,其中所述生物样品包含脱氧核糖核酸(DNA)片段;(b)测定所述生物样品以生成指示DNA片段存在或不存在的信号,所述DNA片段具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与该遗传基因座相关的单核小体保护;以及(c)使用所述信号生成指示具有(i)与来自一个或更多个遗传基因座的遗传基因座相关的双核小体保护,以及(ii)与该遗传基因座相关的单核小体保护的DNA片段存在或不存在的输出。

[0319] 该方法可以包括富集生物样品以获得一个或更多个遗传基因座的组的DNA片段。

[0320] 本文还公开了用于分析生物样品的方法,所述生物样品包含源自受试者的无细胞DNA片段,其中所述方法包括检测来自相同遗传基因座的对应于单核小体保护和双核小体保护中的每一个的DNA片段。

[0321] 本文还公开了用于分析受试者的生物样品的方法,其中该方法包括:(i)将样品中的cfDNA片段测序,以提供DNA序列;(ii)将(i)中获得的DNA序列映射到受试者物种的参考基因组中的一个或更多个基因组区域;以及(iii)针对具有映射的DNA序列的一个或更多个基因组区域,计算对应于单核小体的序列的数量和对应于双核小体的序列的数量。可以比较(iii)中获得的单核小体和双核小体序列的数量。

[0322] 因此,一般而言,分别测定对应于一个或更多个相同遗传基因座的单核小体保护和双核小体保护的cfDNA片段。如本文所示,这些片段的测量水平的变化可以揭示受试者体内生物学状态的变化,例如,图27B显示具有高ERBB2拷贝数的乳腺癌患者样品中的双核小体片段的增加。因此,所述方法可以包括使用检测到或计算出的信号(例如,使用分类器,如本文其他地方所论述的)来评估从中提取样品的受试者的生物学状态(例如,以诊断疾病)的另外步骤。特别地,单核小体或双核小体片段的量的变化可用于评估受试者的生物学状

态。

[0323] 可以以各种方式测定片段,例如,通过如本文其他地方所论述的对cfDNA片段进行测序,或通过按大小(例如,在琼脂糖凝胶上)分离cfDNA片段并对它们进行定量。

[0324] 这些方法可以考虑在基因座处观察到的单核小体和双核小体片段的定量比率(例如,该比率可以随着生物学状态的变化而变化)、在基因座处观察到的片段的量(例如,尽管比率保持不变,两种类型的片段的水平可以增加),或者片段的出现或消失(例如,双核小体片段在一种生物学状态中可能无法检测到,但在另一种状态中可检测到)。在该方法中可以考虑这些信号中的每一个。

[0325] 所述方法可以集中于一个或更多个目的特定遗传基因座,例如,已知其根据生物学状态表现出单核小体和/或双核小体信号的变化。然而,在其他实施方案中,该方法可以检测信号,该信号然后可以与生物学状态的变化相关。例如,可以对cfDNA进行测序,并且可以将序列映射到参考基因组上,如本文其他地方所论述的。在一些实施方案中,对于其中单核小体和/或双核小体信号的变化已经与生物学状态的差异(例如,患病对比非患病,或突变对比野生型,或低拷贝数对比高拷贝数等)相关的基因座,可以评估这些基因座处的信号(例如,使用分类器,如本文其他地方所论述的)。在其他实施方案中,可以将一个或更多个基因座处的单核小体/双核小体信号与取自具有不同生物学状态的受试者的样品中的相同基因座处的信号进行比较,并且可以评估任何差异(例如,使用来自另外受试者的样品)以查看它们是否与生物学状态的差异相关或者构建分类器,如本文其他地方所论述的。

[0326] 因此,一种方法可以包括将单核小体/双核小体片段的量与从参考样品获得的值进行比较的步骤。此类比较可以使用如本文其他地方所述的分类器。

[0327] 用这些方法考虑的基因座通常可以在单个基因或单个基因的启动子区域内。

[0328] 除了考虑双核小体片段之外,这些方法可以另外(或替代地)考虑其他寡核小体片段(三核小体、四核小体等),尽管如图1E所示,此类片段不太丰富,因此不容易检测到。寡核小体片段(双核小体、三核小体等)可以单独或共同考虑。

[0329] 用于单核小体DNA片段和寡核小体DNA片段的测定是本领域已知的。例如,Cell Death Detection ELISA^{PLUS}产品是可商购的,并已应用于血清中的cfDNA(Holdenrieder等人,2005),但它没有在DNA片段的长度之间或不同基因座处的片段之间区分。

[0330] 计算机系统

[0331] 本公开内容提供了被编程为实施本公开内容的方法的计算机系统。图22显示计算机系统2201,其被编程或以其他方式配置成分析包含源自受试者的无细胞核酸的样品。计算机系统2201可以调节本公开内容的方法的各个方面。计算机系统2201可以是用户的电子设备或相对于电子设备远程定位的计算机系统。电子设备可以是移动电子设备。

[0332] 计算机系统2201包括中央处理单元(CPU,本文也称为“处理器”和“计算机处理器”)2205,其可以是单核或多核处理器,或者是用于并行处理的多于一个处理器。计算机系统2201还包括存储器或存储器位置2210(例如,随机存取存储器、只读存储器、闪存)、电子存储单元2215(例如,硬盘)、用于与一个或更多个其他系统通信的通信接口2220(例如,网络适配器)以及外围设备2225,诸如高速缓存、其他存储器、数据存储和/或电子显示适配器。存储器2210、存储单元2215、接口2220和外围设备2225通过诸如母板的通信总线(实线)与CPU 2205通信。存储单元2215可以是用于存储数据的数据存储单元(或数据存储库)。计

算机系统2201可借助于通信接口2220可操作地耦合到计算机网络(“网络”)2230。网络2230可以是因特网(Internet)、互联网(internet)和/或外联网,或与因特网通信的内联网和/或外联网。在一些情况下,网络2230是电信和/或数据网络。网络2230可以包括一个或更多个计算机服务器,其可以实现分布式计算,诸如云计算。在一些情况下,借助于计算机系统2201,网络2230可以实施对等网络,该对等网络可以使耦合到计算机系统2201的设备能够充当客户端或服务器。

[0333] CPU 2205可以执行一系列机器可读指令,这些指令可以体现在程序或软件中。指令可以存储在存储器位置中,诸如存储器2210。指令可以指向CPU 2205,该指令随后可以编程或以其他方式配置CPU 2205以实施本公开内容的方法。由CPU 2205执行的操作的实例可以包括获取、解码、执行和回写。

[0334] CPU 2205可以是电路诸如集成电路的一部分。系统2201的一个或更多个其他组件可以被包括在电路中。在某些情况下,该电路是专用集成电路(ASIC)。

[0335] 存储单元2215可以存储文件,诸如驱动程序、库和保存的程序。存储单元2215可以存储用户数据,例如用户偏好和用户程序。在一些情况下,计算机系统2201可以包括在计算机系统2201外部的一个或更多个另外的数据存储单元,诸如位于通过内联网或因特网与计算机系统2201通信的远程服务器上。

[0336] 计算机系统2201可以通过网络2230与一个或更多个远程计算机系统通信。例如,计算机系统2201可以与用户的远程计算机系统通信。远程计算机系统的实例包括个人计算机(例如,便携式PC)、平板或平板PC(例如,Apple® iPad、Samsung® Galaxy Tab)、电话、智能电话(例如,Apple® iPhone、支持Android的设备、Blackberry®))或个人数字助理。用户可以经由网络2230访问计算机系统2201。

[0337] 本文所述的方法可以通过存储在计算机系统2201的电子存储位置上,例如,存储在存储器2210或电子存储单元2215上的机器(例如,计算机处理器)可执行代码来实施。机器可执行代码或机器可读代码可以以软件的形式提供。在使用期间,代码可以由处理器2205执行。在一些情况下,代码可以从存储单元2215检索并存储在存储器2210上以供处理器2205随时访问。在一些情况下,电子存储单元2215可以被排除,并且机器可执行指令存储在存储器2210中。

[0338] 代码可以被预编译和配置为与具有适于执行代码的处理器的机器一起使用,或者可以在运行时间期间编译。代码可以用编程语言供应,可以选择该编程语言以使代码能够以预编译或编译的方式执行。

[0339] 本文提供的系统和方法的各方面,诸如计算机系统2201,可以在编程中体现。该技术的各个方面可以被认为是“产品”或“制品”,通常是机器(或处理器)可执行代码和/或相关的数据的形式,这些可执行代码和/或相关的数据被承载在一种机器可读介质上或体现在一种机器可读介质中。机器可执行代码可以存储在电子存储单元上,诸如存储器(例如,只读存储器、随机存取存储器、闪存)或硬盘上。“存储”型介质可以包括计算机、处理器等或其相关的模块的任何或所有有形存储器,诸如各种半导体存储器、磁带驱动器、磁盘驱动器等,它们可以在任何时候为软件编程提供非暂时性存储。软件的全部或部分可以通过因特网或各种其他电信网络进行通信。例如,此类通信可以使软件能够从一台计算机或处理器

加载到另一台计算机或处理器中,例如,从管理服务器或主计算机加载到应用服务器的计算机平台中。因此,可以承载软件元件的另一种类型的介质包括光波、电波和电磁波,诸如通过有线和光学陆线网络以及通过各种空中链路,在本地设备之间的物理接口上使用。承载此类波的物理元件,诸如有线或无线链路、光学链路等,也可以被认为是承载软件的介质。如本文所使用的,除非限于非暂时性的有形“存储”介质,否则诸如计算机或机器“可读介质”的术语指参与向处理器提供指令以供执行的任何介质。

[0340] 因此,诸如计算机可执行代码的机器可读介质可以采用许多形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括例如光盘或磁盘,诸如任何计算机等中的任何存储设备,诸如可用于实施附图中所示的数据库等。易失性存储介质包括动态存储器,诸如此类计算机平台的主存储器。有形传输介质包括同轴电缆;铜线和光纤,包括构成计算机系统内总线的导线。载波传输介质可以采用电信号或电磁信号的形式或者声波或光波的形式,诸如在射频(RF)和红外(IR)数据通信期间生成的那些。因此,常见形式的计算机可读介质包括:软盘(floppy disk)、软性磁盘(flexible disk)、硬盘、磁带、任何其他磁介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、穿孔卡纸磁带、具有孔模式的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储器芯片或盒、传输数据或指令的载波、传输此类载波的缆线或链路、或计算机可以从中读取编程代码和/或数据的任何其他介质。这些形式的计算机可读介质中的许多可以参与将一个或更多个指令的一个或更多个序列传送到处理器以供执行。

[0341] 计算机系统2201可包括电子显示器2235或与电子显示器2235通信,电子显示器2235包括用户界面(UI)2240,用户界面2240用于提供例如与源自受试者的包含无细胞核酸的样品的分析相关的信息。UI的实例包括但不限于图形用户界面(GUI)和基于网络的用户界面。

[0342] 可以通过一种或更多种算法来实施本公开内容的方法和系统。算法可以在由中央处理单元2205执行时通过软件实施。

[0343] 虽然本文已经显示和描述了本发明的优选实施方案,但是对于本领域技术人员明显的是,这些实施方案仅以实例的方式提供。本发明并不旨在受说明书中提供的具体实例的限制。虽然已经参考前述说明书描述了本发明,但是本文的实施方案的描述和说明并不意指以限制意义来解释。本领域技术人员现在将想到许多变化、改变和替换,而不偏离本发明。此外,应理解的是,本发明的所有方面不限于本文所述的具体描述、配置或相对比例,其取决于各种条件和变量。应理解的是,本文所述的本发明实施方案的各种替代方案可在实践本发明中采用。因此,预期本发明还应涵盖任何此类替代、修改、变化或等同物。以下权利要求旨在限定本发明的范围,并且由此覆盖这些权利要求范围内的方法和结构及其等同物。

[0344] 实施例1:无细胞DNA片段化模式揭示了与原发性肿瘤中的体细胞突变相关的变化并且提高了体细胞变体检测的灵敏度和特异性

[0345] 从循环血浆分离的无细胞DNA(cfDNA)包括在垂死细胞清除和血流运输中存活下来的DNA片段。在癌症中,这些片段携带肿瘤体细胞变异以及它们的微环境的印迹,使得能够在临床实践中实现非侵入性的基于血浆的肿瘤基因分型。然而,癌症衍生的DNA的分数通常较低,对早期阶段的准确检测提出了挑战,并促使人们寻找与癌症状态相关的正交体细

胞无变体模式。由于已显示cfDNA片段的基因组分布反映了造血细胞中的核小体占位,因此进行了一项实验(a)以观察癌症中cfDNA定位的异质模式与患者肿瘤中的不同突变的关联以及(b)以将cfDNA定位整合到现有的分析方法可以允许提高检测的灵敏度和特异性。

[0346] 通过靶向70个基因的高度准确、深度覆盖(15,000x)ctDNA NGS测试来确定超过15,000名患有晚期临床癌症的患者的cfDNA片段长度和位置的分布以及相关的体细胞基因组谱。进行无变体片段组谱分析的综合分析,并使用统计学方法测试片段组谱与检测到的体细胞改变的关联。观察到不同类别的片段组亚型(例如,通过视觉观察、聚类或其他方法揭示的具有差异性片段组谱的亚型)在具有充分表征的驱动改变和基因组分子亚型的样品中显著富集。对具有已知HER2免疫组织化学状态的独立样品群组进行询问,以证实发现的cfDNA定位和HER2扩增的模式之间的关联。

[0347] 总体而言,片段组谱分析显示ERBB2(例如,HER2)扩增特征与肿瘤的HER2免疫组织化学(IHC)状态显著相关联,导致HER2扩增检测的灵敏度增加42%并且HER2扩增检测的特异性增加7%。观察到的肺腺癌片段组亚型与相互排斥的基因组改变和先前描述的肺癌的内在分子亚型共同发生。总之,这些结果表明cfDNA片段化景观的综合分析可能有助于针对各种人类状况进一步开发基于cfDNA的生物标志物。因此,片段组谱分析可以实现癌症cfDNA的分类,并且可以为观察到的体细胞变异和潜在的肿瘤微环境提供独立的证据,导致变体检测的灵敏度和准确性更高。这表明了为综合检测临床相关类别提供了一条途径,这些类别具有不同的癌症亚型发病机制和疗法选择。

[0348] 实施例2:无细胞DNA片段化模式(片段组谱分析或“片段组学”分析)揭示与肿瘤相关的体细胞突变相关的变化

[0349] 从循环血浆分离的无细胞DNA(cfDNA)包括在垂死细胞清除和血流运输中存活下来的DNA片段。在癌症中,这些片段携带肿瘤体细胞变异以及它们的微环境的印迹,使得能够在临床实践中实现非侵入性的基于血浆的肿瘤基因分型。然而,癌症衍生的DNA的分数通常较低,对早期阶段的准确检测提出了挑战,并促使人们寻找与癌症状态相关的正交体细胞无变体模式。由于已显示cfDNA片段的基因组分布反映了造血细胞中的核小体占位,因此进行了一项实验(a)以观察癌症中cfDNA定位的异质模式与患者肿瘤中的不同突变的关联以及(b)以将cfDNA定位整合到现有的分析。此类方法可以允许提高检测的灵敏度和特异性。

[0350] 通过靶向70个基因的高度准确、深度覆盖(>15,000X)ctDNA NGS测试来确定超过15,000名患有晚期临床癌症的患者的cfDNA片段长度和位置的分布以及相关的体细胞基因组谱。进行无变体片段组谱分析(“片段组学分析”)的综合分析,并使用统计学方法测试片段组谱与检测到的体细胞改变的关联。观察到不同类别的片段组亚型(例如,通过视觉观察、聚类或其他方法揭示的具有差异性片段组谱的亚型)在具有充分表征的驱动改变和基因组分子亚型的样品中显著富集。

[0351] 使用cfDNA片段化模式的信号去卷积,产生跨肿瘤类型的单核小体分辨率片段化模式,如图23中对EGFR基因所见。如在a部分中所见,存在可含有用于癌症检测的肿瘤相关标志物(例如,其可通过液体活检进行测定)的EGFR基因的多个基因组区域。如在b部分中所见,“无序列片段组学”分析揭示了跨EGFR基因的基因组区域的变体,包括良性变体、非体细胞变体和体细胞变体。如在c部分中所见,此类EGFR DNA变体可包含突变(SNV)和扩增(例如

CNV)。如在d部分中所见,通过片段组分析检测包括SNV和CNV的变体来指示总突变负荷。

[0352] 对来自768名患有晚期(late-stage)(晚期(advanced stage))肺腺癌的患者的验证群组的独立样品群组进行询问,以评估片段组学谱并证实发现的cfDNA定位模式与肺癌特异性核小体特征之间的关联。对来自晚期肺腺癌患者的验证群组的所生成的片段组谱进行最小冗余特征选择(例如,如Ding等人,J Bioinform Comput Biol 2005Apr;3(2):185-205中所描述的)。该无监督聚类分析鉴定了肺癌特异性特征的子集(包括与EGFR、KRAS、FGFR2、ALK、EML4、TSC1、RAF1、BRCA2和KIT基因相关的体细胞突变),如图24所示。每行(y-轴)表示从患者抽取的768个cfDNA样品中的一个,并且每列(x-轴)表示对应于不同基因的不同基因组位置。特别地,片段组模式揭示了EGFR、KRAS和FGFR2中的显著的体细胞突变簇(通常在肺腺癌和其他类型的肺癌患者中观察到,例如通过基因分型分析)。因此,片段组谱证实了发现的cfDNA定位的模式(片段组学)和肺癌特异性核小体特征之间的关联。

[0353] 实施例3:无细胞DNA片段化模式(片段组谱分析或“片段组学”分析)可以被建模为异常检测的密度

[0354] 片段组谱可以在3D坐标空间中建模为与特定状况(例如,恶性或非恶性,其中恶性状况代表异常情况)相关的观察到的片段起始和长度的密度。可以使用多种测定方法获得此类片段组谱,所述测定方法诸如数字液滴聚合酶链式反应(ddPCR)、定量聚合酶链式反应(qPCR)和基于阵列的比较基因组杂交(CGH)。此类“液体活检”测定可以是商业上可获得的,例如来自Guardant Health的循环肿瘤DNA测试,来自Fluxion Biosciences的Spotlight 59oncology panel,来自Agena Bioscience的UltraSEEK lung cancer panel,来自Foundation Medicine的FoundationACT液体活检测定以及来自Personal Genome Diagnostics的PlasmaSELECT测定。此类测定可报告一组遗传变体(例如,SNV、CNV、插入/缺失和/或融合)中的每一种的次要等位基因分数(MAF)值的测量值。

[0355] 可以通过异常检测算法对片段组谱进行分析以鉴定异常状况(例如,受试者中的恶性癌症)。异常检测广泛用于数据挖掘,并且可以使用混合物模型和期望最大化(EM)算法来执行。异常检测可以包括混合物建模,这是一种常见的概率聚类技术,其中片段起始和长度的分布可以正式地描述为K-组分(代表K种不同的染色质配置)混合物模型,如图25所示。

[0356] 在上述模型下,可以处理cfDNA起始位置(“起始”)和长度信号(例如,多于一个cfDNA片段中的每一个的起始和长度)以定义界定与特定染色质单元相关的DNA片段子集(例如,在细胞死亡和细胞清除中存活下来的那些)的非恶性观察分布的轮廓的边界。如果进一步的观察位于此类边界界定的子空间内,则这些观察点被认为源自与初始观察结果相同的非恶性群体。否则,位于边界之外的进一步观察结果可以指示异常(例如,源自恶性群体)细胞状态。可以用给定的置信水平确定该异常指示。各种数据分析技术可用于将混合物模型应用于将异质观察集合中的子群体成簇,包括:单类SVM[Estimating the support of a high-dimensional distribution Schölkopf, Bernhard,等人.Neural computation 13.7(2001):1443-1471.]、拟合椭圆包络[Rousseeuw,P.J.,Van Driessen,K.“A fast algorithm for the minimum covariance determinant estimator”Technometrics 41(3),212(1999)]、以及分离森林(Isolation Forest)[Liu,Fei Tony,Ting,Kai Ming和Zhou,Zhi-Hua.“Isolation forest.”Data Mining,2008.ICDM‘08.Eighth IEEE International Conference on.],其每一个都通过引用并入本文。

[0357] 拟合椭圆包络的方法可以应用于上面定义的二元正态混合物(并且在图25中示出)。第一操作包括建立与来自相同组蛋白保护的DNA单元的片段相关的轮廓线。下面描述多元正态中的等线(iso-line)的此类推导,并将轮廓线建立为椭圆体。给定一组非恶性对照血浆样品,基因组空间可以细分为非重叠区段,该区段定义在cfDNA片段的群体中观察到的受保护DNA的簇。接下来,构建二元正态或二元t分布模型P(x)以获得特定片段来自非恶性细胞的概率。如果概率p低于阈值ε,则认为此类片段是异常的。将跨所有基因组区段的异常片段的密度(适当注意染色体X和染色体Y)求和产生恶性负荷(例如,肿瘤负荷)的定量量度,其代表源自非恶性染色质构型之外的cfDNA片段(即,起源异常的cfDNA片段)的分数。如果训练集包括从多于一个非恶性对照(例如,健康对照受试者)获得的生理学上不同的cfDNA样品集,则任何检测到的恶性贡献(例如,检测到的异常)可以指示癌症起源。通过将椭圆形包络拟合到二元正态混合物(如图26A所示),可以执行此类恶性负荷确定,使得:

[0358] $(x-\mu)^T\Sigma^{-1}(x-\mu)=c$

[0359] 其中Σ是协方差矩阵。该等式代表一个椭圆。在一个简单的情况下,其中μ=(0,0)且Σ是对角线,得到以下等式:

[0360] $(x/\sigma_x)^2+(y/\sigma_y)^2=c$

[0361] 在Σ不是对角线的情况下,可以执行对角化以得到相同的结果。对角化技术在例如[Hyndman,R.J.(1996).Computing and graphing highest density regions.The American Statistician,50(2),120-126.]中描述,其通过引用并入本文。

[0362] 使用来自参考样品(例如,健康对照)的cfDNA群体,进行以下算法以训练和测试二元正态混合物模型。

[0363] 首先,使用包含40个非恶性成人血浆样品的数据集进行训练。针对每个人类染色体,忽略片段长度并使用统计软件包R中的“密度”函数计算内核密度估计值。算法(1)将经验分布函数的质量分散跨至少5000个点的规则网格,然后(2)使用快速傅里叶变换将该近似与内核的离散化版本进行卷积,并且然后(3)使用线性近似来评价指定点处的密度。内核密度估计方法描述于例如[Venables,W.N.和Ripley,B.D.(2002)Modern Applied Statistics with S.New York:Springer.]中,其通过引用并入本文。

[0364] 接下来,以计算的密度建立谷,以便建立染色质保护单元的边界。谷被定义为发生方向变化的系列中的最低值。接下来,针对每个定义的区段,使用统计软件包R中的KernSmooth包计算2D分箱内核密度估计值。例如,KernSmooth算法描述于例如[Wand,M.P.(1994).Fast Computation of Multivariate Kernel Estimators.Journal of Computational and Graphical Statistics,3,433-445.]中,其通过引用并入本文。接下来,在每个坐标方向上产生一组网格点(基因组位置作为x-轴并且片段长度作为y-轴)。接下来,跨由网格点引起的网格计算密度估计矩阵。

[0365] 使用的内核是标准的二元正态密度。对于预定义网格上的每个(x_1,x_2)对,二元高斯内核居中在该位置上,并且在每个数据点处对通过带宽缩放的内核的高度求和。可以根据需要将网格稀疏地定义(例如,每3bp、5bp等)。两个方向的15bp的网格大小用于最小化内存使用。带宽指内核带宽平滑参数,带宽值越大使估计值越平滑,并且带宽值越小使估计值越不平滑。通过检查12p11.1区域中的不同带宽性能,进行了带宽为30bp的启发式调谐,该区域包含超过400个强烈定位的核小体谱(即,跨多个组织、细胞谱系和生物体保持相同核

小体结构的那些谱)。此类强烈定位的核小体谱描述于例如[Gaffney,D.J.等人.Controls of nucleosome positioning in the human genome.PLoS Genet.8,e1003036(2012)],其通过引用并入本文。可选地,可以使用正式带宽估计(可从URL www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf获得)来最小化平均积分平方误差。

[0366] 接下来,使用估计的均值和协方差,使用统计软件包R中的mvtnorm库建立99.995%的椭圆包络。该算法包括使用solve()函数来对方差-协方差矩阵求逆,并且使用ellipse()函数将高度度量计算为二元正态密度的对数的负数。可以使用椭圆形包络的其他值,例如,至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%、至少99.9%、至少99.99%、至少99.999%或至少99.9995%。

[0367] 上述训练操作已经在3D片段起始位置和长度空间中建立了区域,这些区域以99.995%的置信度代表非恶性簇。接下来,使用包含从肺癌和结肠癌患者群组获得的cfDNA样品的数据集进行二元正态混合物模型的测试,其中cfDNA样品来源于切除前和切除后血液抽取。与训练类似,算法的测试部分包括计算2D内核密度估计值。接下来,将恶性负荷(恶性负载、肿瘤负荷或肿瘤负载)计算为非恶性椭圆形包络之外的密度的加权和。将权重设置为非恶性训练集的2D内核密度估计值的倒数。

[0368] 图26B显示通过使用上述二元正态混合物模型对5个不同群组(结肠直肠癌术后、结肠直肠癌术前、肺癌术后、肺癌术前和正常)中的cfDNA样品的片段组分析生成的失调评分的分布的实例。“术后”指在手术切除操作之后进行从血液抽取中分析其cfDNA的受试者。“术前”指在手术切除操作之前进行从血液抽取中分析其cfDNA的受试者。注意,结肠直肠癌术后和肺癌术后群组的失调评分(以及因此的恶性负荷)具有较低的值并且与正常(例如,健康)群组的那些相似。相反,结肠直肠癌术前和肺癌术前群组的失调评分(以及因此的恶性负荷)具有显著高于正常(例如,健康)群组的值。此外,与其他三个群组(结肠直肠癌术后、肺癌术后和正常受试者)相比,结肠直肠癌术前和肺癌术前群组的失调评分(以及因此的恶性负荷)在这些群组内的变异显著较高。

[0369] 实施例4:无细胞DNA片段化模式(片段组谱分析或“片段组学”分析)揭示与肿瘤相关的拷贝数变异(CNV)相关的变化

[0370] 从循环血浆分离的无细胞DNA(cfDNA)包括在垂死细胞清除和血流运输中存活下来的DNA片段。在癌症中,这些片段携带肿瘤拷贝数变异以及它们的微环境的印迹,使得能够在临床实践中实现非侵入性的基于血浆的肿瘤基因分型。然而,癌症衍生的DNA的分数通常较低,对早期阶段的准确检测提出了挑战,并促使人们寻找与癌症状态相关的正交拷贝数无变体模式。由于已显示cfDNA片段的基因组分布反映了造血细胞中的核小体占位,因此进行了一项实验(a)以观察癌症中cfDNA定位的异质模式与患者肿瘤中的不同CNV的关联以及(b)以将cfDNA定位整合到现有的分析。此类方法可以允许提高检测的灵敏度和特异性。

[0371] 通过进行液体活检测定以测量晚期靶向外显子组的MAF研究了ERBB2核小体动力学。包含DNA片段大小对比DNA片段起始位置的2D热图谱的多参数模型(例如,DNA片段覆盖率作为第三个维度)用于通过经由线性分箱的起始位置、经由FFT的离散卷积和二元高斯内核拟合导出对片段计数的普通内核密度估计值的分箱近似,其结果如图27A所示。

[0372] 图27A示出了多参数模型的实例,该多参数模型包括在与TP53基因,外显子编号7

(其中在z-轴上的片段计数由颜色阴影表示)相关的基因组区域中的受试者的片段大小(例如,片段长度)(y-轴)和基因组位置(x-轴)。该多参数模型可用于可视化无细胞核小体定位的效果。从对应于患有肿瘤的受试者的多参数模型(在这种情况下,热图谱),可以观察到两个峰,该两个峰间隔大约180个碱基位置(例如,沿着对应于位置的横轴)。另外,可以观察到对应于单核小体保护的三个峰(例如,对应于约160至约180个碱基位置(bp)范围内的片段大小)。另外,可以观察到对应于双核小体保护的三个峰(例如,对应于约320至约340个碱基位置(bp)范围内的片段大小)。这些峰中的每一个可以包括位置(例如,沿着横轴在峰的中心处)、片段大小(例如,沿着纵轴在峰的中心处)和峰宽度(例如,沿着轴中的一个)。

[0373] 在20个ERBB2阴性和ERBB2阳性晚期乳腺癌患者的群组中,通过全基因组分析检查两种调节元件(例如,与ERBB2基因相关的启动子和增强子区域)。此类研究揭示了在ERBB2阳性病例中,核小体清除的预期染色质结构具有足够的片段覆盖率,并且存在与表达相关的双核小体簇,如图27B和图27C所示。

[0374] 图27B显示了20个样品的四个聚合的晚期乳腺癌群组中的ERBB2启动子区域的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(ii)包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(iii)包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组,以及(iv)包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组。

[0375] 包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组代表可能在肿瘤中的ERBB2基因中具有低肿瘤负荷和低CNV的受试者。包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组代表可能在肿瘤中的ERBB2基因中具有高肿瘤负荷但低CNV的受试者。如在图27B的顶部两行中的热图谱中所见,肿瘤中ERBB2基因中具有低CNV的受试者在低突变负荷和高突变负荷情况二者表现出相似的片段组谱。

[0376] 包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组代表可能在肿瘤中的ERBB2基因中具有低肿瘤负荷但具有高CNV的受试者。包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组代表可能在肿瘤中的ERBB2基因中具有高肿瘤负荷并且具有高CNV的受试者。如在图27B的底部两行中的热图谱中所见,肿瘤中ERBB2基因中具有高CNV的受试者在低突变负荷和高突变负荷情况二者表现出相似的片段组谱。此外,ERBB2基因中具有高CNV的受试者表现出片段组谱,所示片段组谱具有(i)更多双核小体峰(位于沿着对应于片段长度的纵轴的每行热图谱的上部)的外观以及(ii)两个峰之间的更大距离以及其他峰的“拖尾”(例如,不太明显的峰,其具有更大的宽度并因此开始合并在一起)。

[0377] 图27C显示了20个样品的四个聚合的晚期乳腺癌群组中的ERBB2增强子区域的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(ii)包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组,(iii)包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组,以及(iv)包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组。

[0378] 包含低突变负荷和近二倍体ERBB2拷贝数(CN)的群组代表可能在肿瘤中的ERBB2基因中具有低肿瘤负荷和低CNV的受试者。包含高突变负荷和近二倍体ERBB2拷贝数(CN)的群组代表可能在肿瘤中的ERBB2基因中具有高肿瘤负荷但具有低CNV的受试者。如在图27C的顶部两行中的热图谱中所见,肿瘤中ERBB2基因中具有低CNV的受试者在低突变负荷和高

突变负荷情况二者表现出相似的片段组谱。

[0379] 包含低突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组代表可能在肿瘤中的ERBB2基因中具有低肿瘤负荷但具有高CNV的受试者。包含高突变负荷和高ERBB2拷贝数(CN)(例如,大于约4)的群组代表可能在肿瘤中的ERBB2基因中具有高肿瘤负荷并且具有高CNV的受试者。如在图27C的底部两行中的热图谱中所见,肿瘤中ERBB2基因中具有高CNV的受试者在低突变负荷和高突变负荷情况二者表现出相似的片段组谱。此外,ERBB2基因中具有高CNV的受试者表现出片段组谱,所述片段组谱具有更多的双核小体峰(位于沿着对应于片段长度的纵轴的每行热图谱的上部)的外观。

[0380] 个体受试者样品的片段组分析证实了使用靶向测定诸如液体活检测定法进行染色质结构检测的可行性,如图28A和图28B所示。

[0381] 图28A显示了对齐的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(右侧;如从上到下所示):(i)从单个样品(来自ERBB2阳性受试者)生成的ERBB2增强子区域(右上)的热图谱,(ii)从多于一个健康对照生成的聚合群组热图谱,以及(iii)从多于一个高ERBB2CN/低突变负荷受试者生成的聚合群组热图谱。此外,示出了在4个不同的基因组区域(例如,对应于TP53、NF1、ERBB2和BRCA1基因)处的单核小体和双核小体计数(例如,测试样品中计数的在该基因组位置处起始的片段的数量)的覆盖图(左侧)。与健康对照群组相比,测试样品表现出更类似于高ERBB2 CN和低突变负荷群组(例如,具有双核小体片段的峰的外观或"双核小体峰")的片段组谱(右侧)。此外,与其他3个基因(TP53、NF1和BRCA1)相比,测试样品表现出单核小体和双核小体计数的覆盖图(左侧),这两者在ERBB2基因区域都显著升高(例如,几倍)。因此,测试样品的片段组谱和覆盖图均指示并证实测试受试者可能是ERBB2阳性。通过进行片段组谱分析,在不考虑ERBB2基因的基因座中每个碱基位置的碱基身份的情况下,测量并获得了ERBB2基因中CN遗传畸变的存在。

[0382] 图28B显示了对齐的2D片段起始位置(x-轴)和片段长度(y-轴)密度热图谱(如从上到下所示):(i)从单个样品(来自ERBB2阴性受试者)生成的ERBB2增强子区域(右上)的热图谱,(ii)从多于一个健康对照生成的聚合群组热图谱,以及(iii)从多于一个高ERBB2CN/低突变负荷受试者生成的聚合群组热图谱。此外,示出了在4个不同的基因组区域(例如,对应于TP53、NF1、ERBB2和BRCA1基因)处的单核小体和双核小体计数(例如,测试样品中计数的在该基因组位置处起始的片段的数量)的覆盖图。与高ERBB2CN和低突变负荷群组相比,测试样品表现出更类似于健康对照群组(例如,缺乏双核小体片段的峰或"双核小体峰")的片段组谱(右侧)。此外,与其他3个基因(TP53、NF1和BRCA1)相比,测试样品表现出单核小体和双核小体计数的覆盖图(左侧),其在ERBB2基因区域中未升高。因此,测试样品的片段组谱和覆盖图均指示并证实测试受试者可能是ERBB2阴性。通过进行片段组谱分析,在不考虑ERBB2基因的基因座中每个碱基位置的碱基身份的情况下,测量并获得了ERBB2基因中CN遗传畸变的不存在。

[0383] 在一个方面,本文公开了一种用于生成指示来自从受试者获得的无细胞样品(或无细胞DNA)的脱氧核糖核酸(DNA)片段中遗传畸变的存在或不存在的输出的方法。该方法可以包括从片段组谱(例如,2D热图谱图)鉴定一个或更多个峰。此类鉴定可以包括构建来自无细胞样品(或无细胞DNA)的DNA片段跨基因组中多于一个碱基位置的分布。接下来,可以在DNA片段的分布中鉴定多于一个碱基位置的一个或更多个碱基位置处的一个或更多个

峰。每个此类峰可以包括峰值和峰分布宽度。接下来,可以确定受试者中遗传畸变的存在或不存在。此类确定可以至少基于(i)一个或更多个碱基位置,(ii)峰值,和/或(iii)峰分布宽度。在一些实施方案中,一个或更多个峰包含双核小体峰和/或单核小体峰。

[0384] 在一些实施方案中,至少基于指示与双核小体峰相关的第一峰值和与单核小体峰相关的第二峰值的比率的定量量度来确定指示遗传畸变的存在或不存在的所述输出,或反之亦然。例如,可以使用双核小体峰值(和/或峰分布宽度("峰宽度"))与单核小体峰值(和/或峰宽度)的比率来指示测试样品的片段组谱是否可以模式匹配到一个或多个健康对照受试者(或群组)和/或一个或更多个患病受试者(或群组)的片段组谱(具有相似的峰位置、峰值和/或峰宽度)。

[0385] 一旦生成多参数分布(例如,2D密度图或热图谱),就可以估计多模态密度;然而,即使在一个维度上,此类估计也可能具有挑战性。对于单模态模型,密度形状可以通过可以使用众所周知的多元分布分析方法生成的参数(例如,偏度和峰度)来描述。对于多模态模型,可以执行多模态密度分析(例如,诸如片段起始位置("片段起始")的参数的多模态密度分析)以确定模式的数量和每个此类模式的位置,因为模式是模拟染色质标志物的表观遗传帽分析基因表达(CAGE)峰的主要特征,并且可能是潜在的染色质组织的迹象。

[0386] 多模态密度分析可以包括使用混合物模型,该混合物模型以与多模态密度配置一致的方式将采样群体分解成一组均匀组分。可以使用各种方法(methods)和方法(approaches)来确定多元正态混合物的模态行为,例如机器学习算法。作为实例,可以对多参数分布(例如,片段组2D密度)执行图像处理和图像分割算法,诸如适合于拓扑图的分水岭变换。此类分水岭变换方法可以代表片段组谱,使得每个点的亮度代表其高度,因此多模态密度分析可以包括确定沿着此类分水岭图的脊的顶部延伸的一条或多条线。使用此类变换方法,分析片段组谱以经由二元正态混合物的拓扑建模来映射规范核小体架构,如图29A所示。

[0387] 图29A显示ERBB2和NF1外显子结构域的2D核小体映射(无扩增)。例如,可以通过在染色体17上进行与ERBB2启动子区域和相邻基因NF1相关的片段组谱的脊线重建来获得此类核小体映射。在该过程中,将核小体掩模拟合至片段组谱。

[0388] 这里,信号代表核小体边界的轮廓以及此类轮廓上的密度的变化。在该图的底部,显示了2D密度估计和图像处理。在该图的顶部,跨30个近二倍体ERBB2临床病例(例如,其液体活检测定报告MAF值指示低CNV或无CNV的受试者)中观察到的规范结构域的核小体掩模。对健康受试者进行检查并进行片段组谱分析,并且确定预期存在核小体的轮廓。此类分析包括使用差量信号,其中每个差量信号包括DNA片段(例如,测试样品的DNA片段)的分布和参考分布(例如健康对照的规范分布)之间的差异。基于健康对照构建掩模,并将该掩模应用于测试样品。得到的图指示,该测试样品具有与健康对照群组非常相似的片段组谱。

[0389] 然后将该核小体掩蔽方法应用于染色体17(chr17)的整个靶向结构域,并扩展至其通过液体活检测定法测定的7,000个样品的较大临床群组,该样品代表跨4种组织类型(前列腺、结肠、乳房和肺)的晚期癌症患者。对片段组信号进行去卷积以产生chr17靶向结构域的规范核小体掩模,该chr17靶向结构域包括4种基因ERBB2、NF1、BRCA1和TP53。

[0390] 接下来,通过将被测定用于肿瘤相关的次要等位基因频率(MAF)的跨811个晚期乳腺癌样品中ERBB2基因的残差掩模与相邻基因的残差掩膜进行对比,使用源自泛癌近二倍

体ERBB2拷贝数训练集的核小体特异性特征来估计ERBB2表达组分和染色体17肿瘤负荷。具体地,将肿瘤负荷评估为跨非ERBB2结构域的迭代残差测量值,针对局灶扩增事件稳健化(robustified)(如图30中所示),并且跨811个乳腺癌样品,将ERBB2表达量度计算为ERBB2双核小体对比单核小体通道中的残差密度估计值,用于ERBB2表达对比拷贝数估计值(如图31A所示)。将ERBB2拷贝数确定为ERBB2单核小体中的残差密度,针对突变负荷进行校正,并且在ERBB2边界之外进行评估。

[0391] 图29B显示ERBB2和NF1外显子结构域的2D核小体映射(无扩增)。在该图的底部,显示了2D密度估计和图像处理。在该图的顶部,显示了跨30个ERBB2临床病例观察到的规范结构域的核小体掩模。在该过程中,使用测试样品和规范健康谱之间的比较(例如,通过执行信号去卷积和对去卷积的信号的模式识别)来执行模式匹配。可以使用多种方法进行比较以观察差异。例如,可以计算对数似然性以测量观察到的信号与以下之间的距离(或差量信号):(i)一个或更多个规范掩模(例如,来自健康对照),(ii)一个或更多个阳性异常谱,或(iii)两者的组合。作为另一实例,可以执行图像处理算法以用于片段组谱比较。然后可以比较此类距离或差量信号,以确定给定测试样品是否具有指示受试者更可能处于健康或患病状态的片段组谱。与多于一个参考分布(例如,一个或更多个健康和一个或更多个患病)的比较可以并入单个比较中。

[0392] 图30显示了跨4个不同群组推断的染色体17肿瘤负荷图,这些群组先前已通过液体活检测定法测定最大MAF:(i)最大MAF范围为(0,0.5]的群组,(ii)最大MAF范围为(0.5,5]的群组,(iii)最大MAF范围为(5,20]的群组,以及(iv)最大MAF范围为(20,100]的群组。肿瘤的细胞清除(例如,肿瘤脱落细胞和无细胞DNA进入循环的趋势)可以通过计算NF1基因或其他非癌症标志物的定量量度来测量。例如,此类定量量度可以是具有双核小体保护的测量片段的数量与具有单核小体保护的测量片段的数量的比率。来自从受试者获得的无细胞样品(或无细胞DNA)的DNA片段的分布(例如,多参数分布或单参数分布)可以在遗传基因座处去卷积成一种或更多种组分。此类组分可包含拷贝数(CN)、细胞清除和基因表达中的一个、两个、三个。去卷积可以包括构建来自无细胞样品(或无细胞DNA)的DNA片段跨基因组中多于一个碱基位置的覆盖率的分布。接下来,对于一个或更多个遗传基因座中的每一个,去卷积可以包括对覆盖率的分布进行去卷积,从而生成与拷贝数(CN)组分、细胞清除组分和/或基因表达组分相关的分数贡献。

[0393] 图31A显示ERBB2表达组分对比ERBB2拷贝数的图。在此,将ERBB2表达测量值(y-轴)计算为跨811个乳腺癌样品ERBB2双核小体对比单核小体通道的残差密度估计值。检查ERBB2启动子区域以观察与拷贝数变化相关的染色质重组事件。由于拷贝数变化与表达有关,因此可以从片段组信号估计表达。对于先前经由FISH和/或免疫组织化学(IHC)证实为HER2阳性的ERBB2状态的受试者群组,在该群组中的ERBB2启动子区域中检查片段组谱,并鉴定ERBB2阳性表达的掩模。类似地,生成ERBB2阴性群组的掩模(再次,通过FISH和/或IHC临床验证)以鉴定ERBB2阴性表达的掩模。因此,对于给定的测试样品,分析相关的片段组谱(例如,作为ERBB阳性谱和ERBB2阴性谱的混合物)可以揭示匹配ERBB2阳性或ERBB2阴性片段组谱的似然(例如,与模式匹配相关的对数似然)。对于群组中的每个受试者,从相关的片段组谱的覆盖数量测量ERBB2拷贝数。

[0394] 图31B示出了使用ERBB2阴性训练集的2D阈值化的图,其经由构建方差-协方差矩

阵、对方差-协方差矩阵求逆以及生成椭圆判别函数来执行。ERBB2表达和拷贝数的多元正态分布用平均向量μ和协方差矩阵Σ参数化，并用于产生判别评分。该程序用于测试测试样品是否被包含在由ERBB2阴性训练数据的二元正态近似产生的椭圆内。椭圆(如图31B所示)由数据的一阶矩和二阶矩确定。ERBB2表达和拷贝数的多元正态分布的方差-协方差矩阵求逆产生了判别评分。将该判别评分计算为二元正态密度的负对数。

[0395] 表2

[0396]

		FISH|IHC				FISH|IHC		
		阴性	阳性			阴性	阳性	
常规CNU	检测到的	4	17	21	片段组学	2	21	23
	未检测到的	26	11	37		28	7	35
	总计	30	28	58		30	28	58
		估计的	95%置信区间			估计的	95%置信区间	
		值	下限	上限		值	下限	上限
	灵敏度	0.61	0.41	0.78		0.75	0.55	0.89
	特异性	0.87	0.68	0.96		0.93	0.76	0.99

[0397] 表2显示了具有已知HER2免疫组织化学状态的58个样品中的扩增检测概要结果。这些结果包括ERBB2阳性和ERBB2阴性乳腺癌病例的独立测试集的灵敏度和特异性总结，这些结果通过免疫组织化学(IHC)和荧光原位杂交(FISH)得到验证。这些结果指示，与传统的CNV检测方法相比，片段组学(片段组谱的分析)使ERBB2阳性和ERBB2阴性乳腺癌病例的扩增检测能够具有更高的灵敏度和特异性。此类片段组学方法可以与传统CNV检测方法(例如，考虑一个或更多个遗传基因座中碱基位置的碱基身份的方法)并行进行，以更高灵敏度和更高特异性检测CNV。可选地，此类片段组学方法可以与传统的CNV检测方法(例如，考虑一个或更多个遗传基因座中碱基位置的碱基身份的方法)组合进行，以比单独的任一种方法更高的灵敏度和更高的特异性检测CNV。

[0398] 实施例5:无细胞DNA片段化模式(片段组谱分析或“片段组学”分析)揭示了指示与癌症相关的免疫细胞类型存在的变化

[0399] 一组片段组谱，该片段组谱包含由chr1:43814893-43815072的单个连续段代表的MPL基因(MPL原癌基因、血小板生成素受体)基因座的片段起始分布，通过以下检查:(i)跨越至少6个不同组织的一组2,360个晚期恶性病例，以及(ii)43个健康生物库对照受试者。对于每个片段组谱，在滑动的30bp窗口中计算双核小体比率，该双核小体比率定义为观察到的双核小体片段(长度在～240bp到～360bp范围内)的数量除以单核小体片段(长度小于240bp)的数量。接下来，通过减去跨健康对照受试者的中值谱，对每个片段组谱获得此类双核小体比率的残差。如图32A所示，生成如由热图谱所代表的残差图，其中行对应于样品并且列对应于跨越180bp的MPL靶向结构域的单个窗口，并且其中y-轴通过液体活检测定中观察到的递增最大突变等位基因频率(MAF)来排序。

[0400] 高MAF样品(大于约30％)(即来自具有最高肿瘤负荷并因此代表相对晚期转移性疾病的受试者的那些)表现出双核小体残差的富集，其指示与健康对照受试者相比在高肿瘤负荷癌症中的短程(亚核小体，小于～180bp)差异性染色质架构。检查靶向MPL结构域的ENSEMBL转录结构揭示了残差双核小体比率信号中的断点(如图32B和图32C所示)，这与随着高肿瘤负荷癌症样品中片段的富集而导致的转录物结构变异相关，这与MPL的可变转录物中截短外显子使用一致。此类断点指示MPL基因中的替代剪接事件，并且代表跨越两个不同转录物的亚核小体片段信号，其中一个转录物是另一个转录物的截短形式。截短形式的

转录物(规范形式)显示在顶部,而非规范形式的转录物显示在底部。

[0401] 对断点与组织特异性替代外显子使用的关联的进一步检查(如图32C所示)揭示了定义跨膜Mpl变体、MPLK(完整)和MPLP(截短)的鉴定。在单核细胞、B淋巴细胞和T细胞群体中检测到MPLP变体,而在单核细胞、B细胞和T细胞中MPLK mRNA表达较低。我们观察到与较短转录物边缘相关的断点,而与较长转录物相关的小的分数(即较低信号)。在免疫细胞类型群体中观察到较长的转录物,并且可以指示癌症的存在和/或侵袭性。这些结果指示,相对于健康的正常对照受试者,具有高肿瘤负荷的受试者携带另外的无细胞DNA负荷,该另外的无细胞DNA负荷富含MPLP特征。此类特征指示与癌症存在和侵袭性相关的免疫细胞类型存在(例如,如[Different mutations of the human c-mpl gene indicate distinct hematopoietic diseases,Xin He等人,Journal of Hematology&Oncology20136:11]中所述)。因此,这些结果指示,片段组学(片段组谱的分析)使得能够检测和鉴定其存在与癌症相关的免疫细胞类型的存在或相对增加的量。

$$\mathrm{cfDNA} = \sum_{\substack{\text{组织}\\(\text{包括血液})}} \begin{bmatrix} \text{良性全身响应} \\ \text{肿瘤全身响应} \\ \text{肿瘤微环境} \\ \text{肿瘤} \end{bmatrix}$$

图1A

$$\mathrm{cfDNA} = \sum_{\substack{\text{组织}\\(\text{包括血液})}} \begin{bmatrix} \text{良性全身响应} \times \text{良性清除} \\ \text{肿瘤全身响应} \times \text{炎症清除} \\ \text{肿瘤微环境} \times \text{TM清除} \\ \text{肿瘤} \times \text{肿瘤清除} \end{bmatrix}$$

图1B

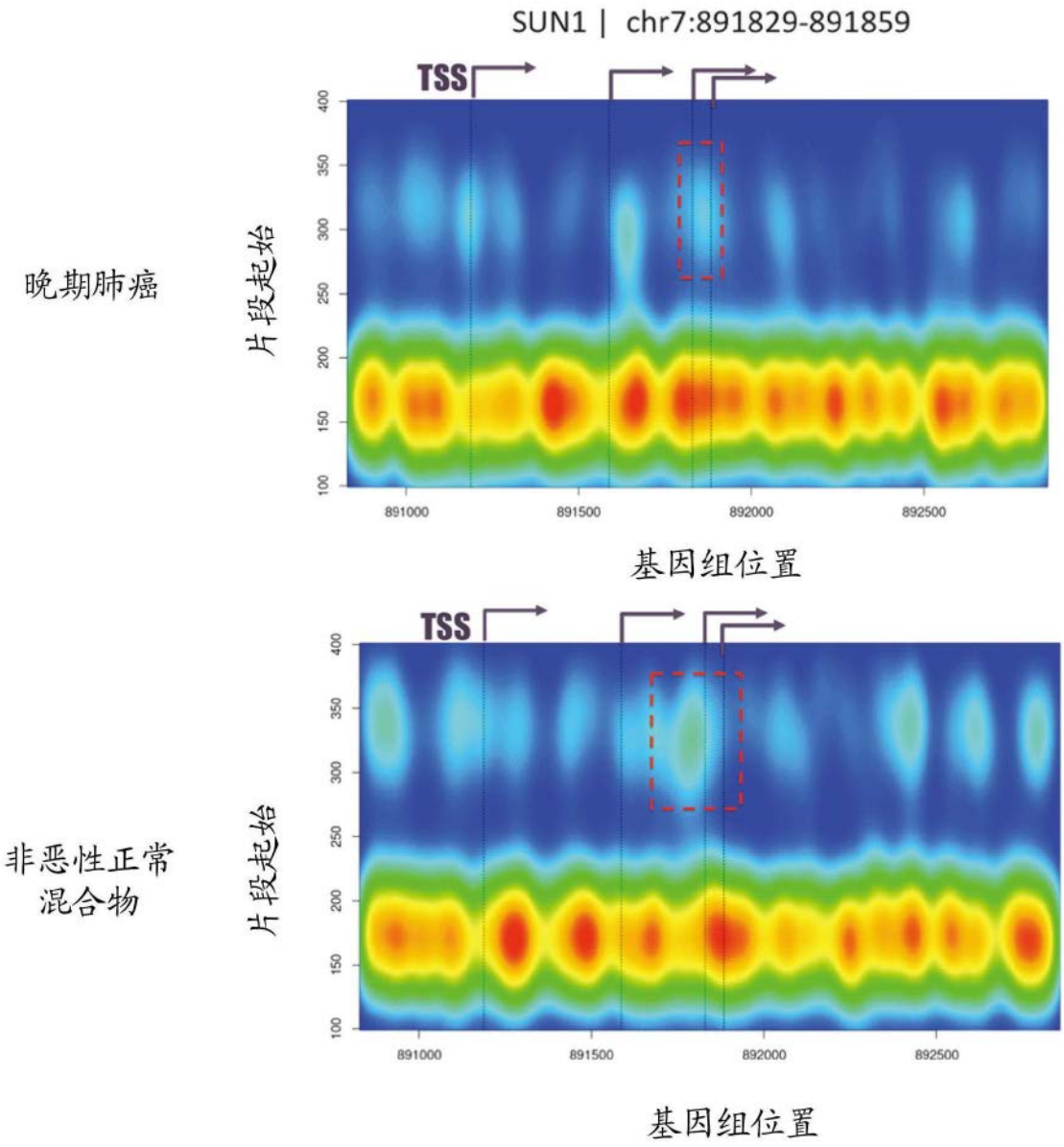

SUN1 | chr7:891829-891859
晚期肺癌
TSS
片段起始
400
350
300
250
200
150
100
891000
891500
892000
892500
基因组位置
非恶性正常
混合物
TSS
片段起始
400
350
300
250
200
150
100
891000
891500
892000
892500
基因组位置

图1C

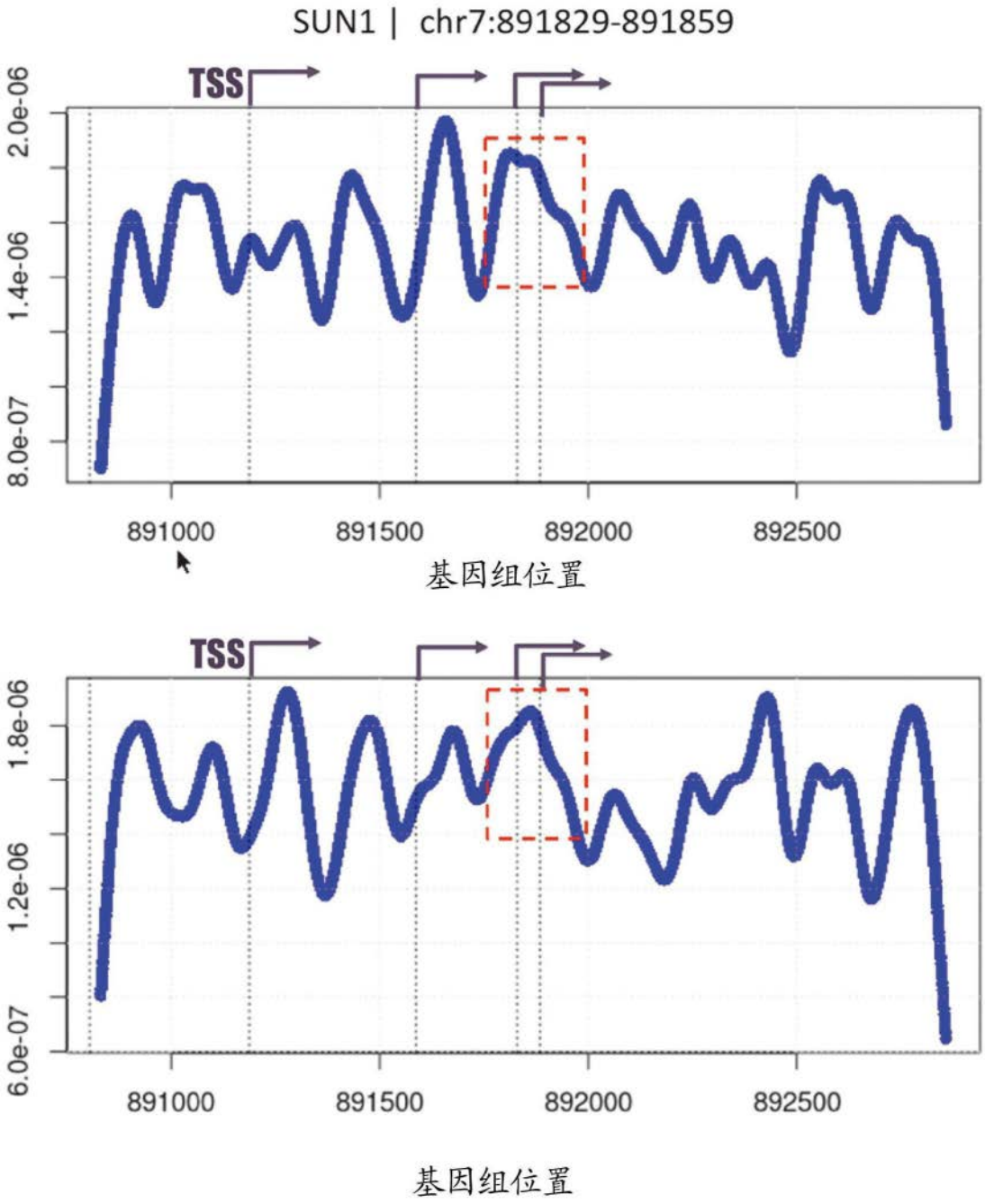

SUN1 | chr7:891829-891859
TSS
2.0e-06
1.4e-06
8.0e-07
891000
891500
892000
892500
基因组位置
TSS
1.8e-06
1.2e-06
6.0e-07
891000
891500
892000
892500
基因组位置

图1D

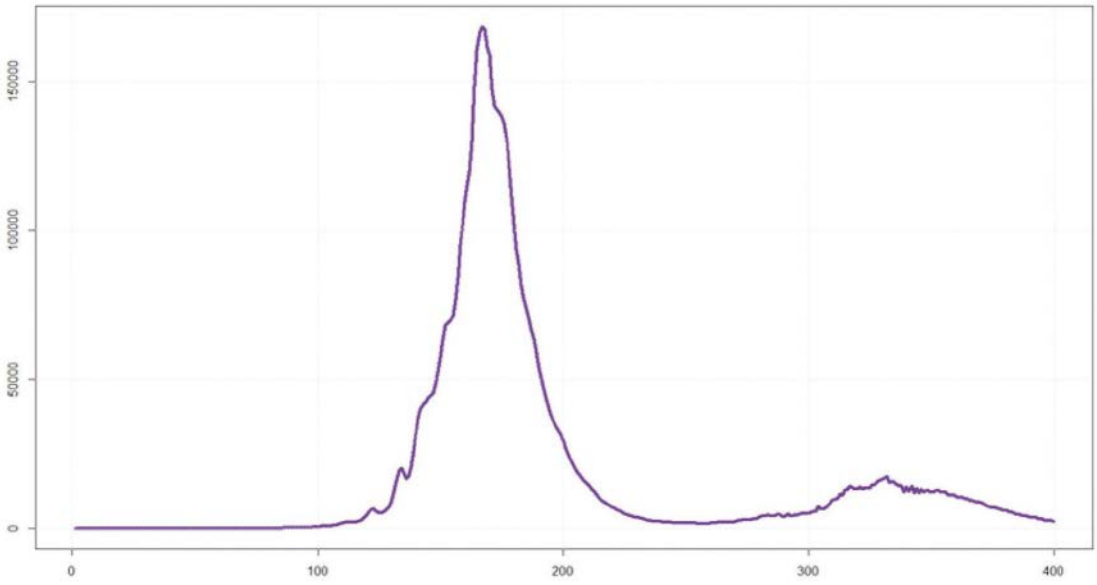


图1E

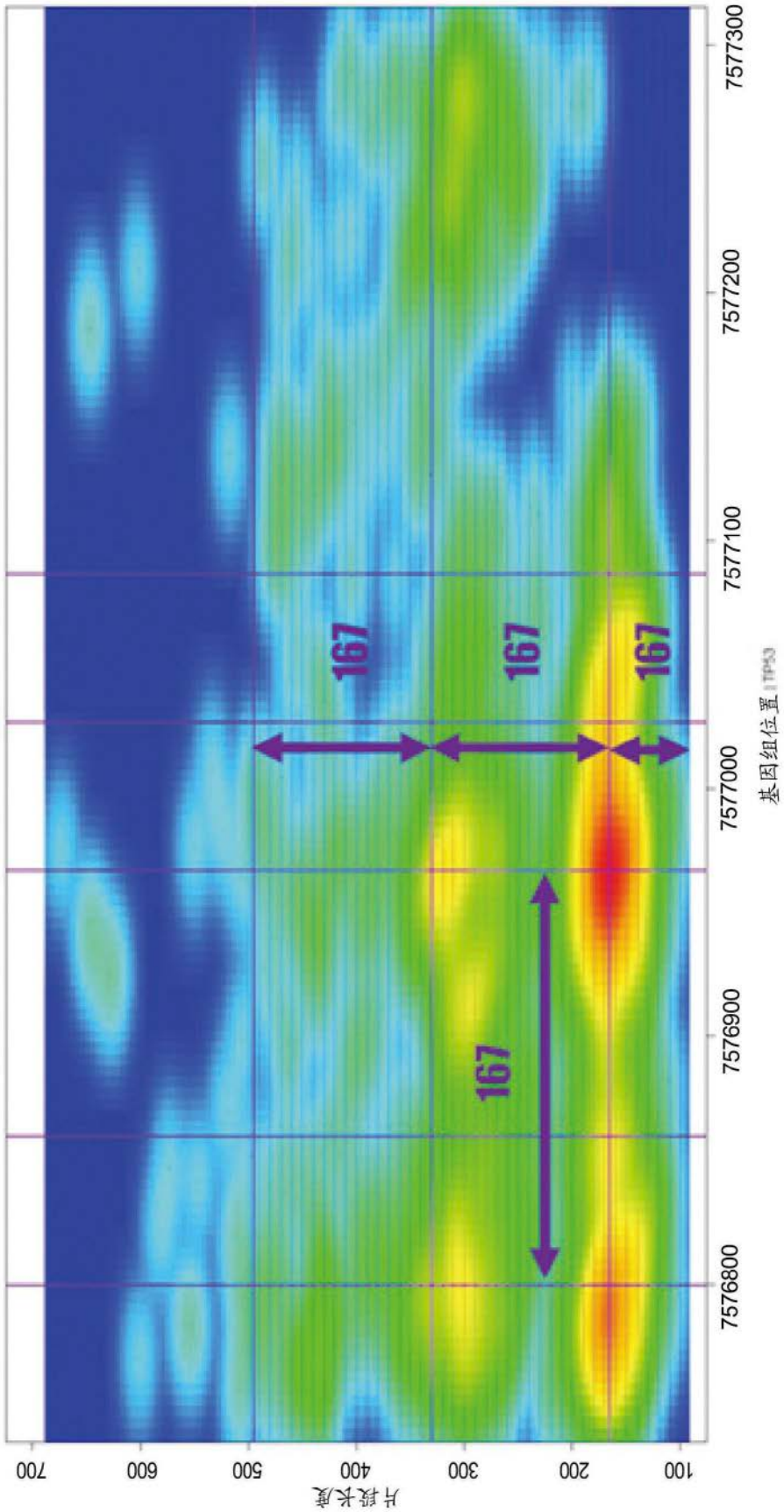

7577300
7577200
7577100
7577000
7576900
7576800
基因组位置
167
167
167
167
100
200
300
400
500
600
700
片段长度

图2

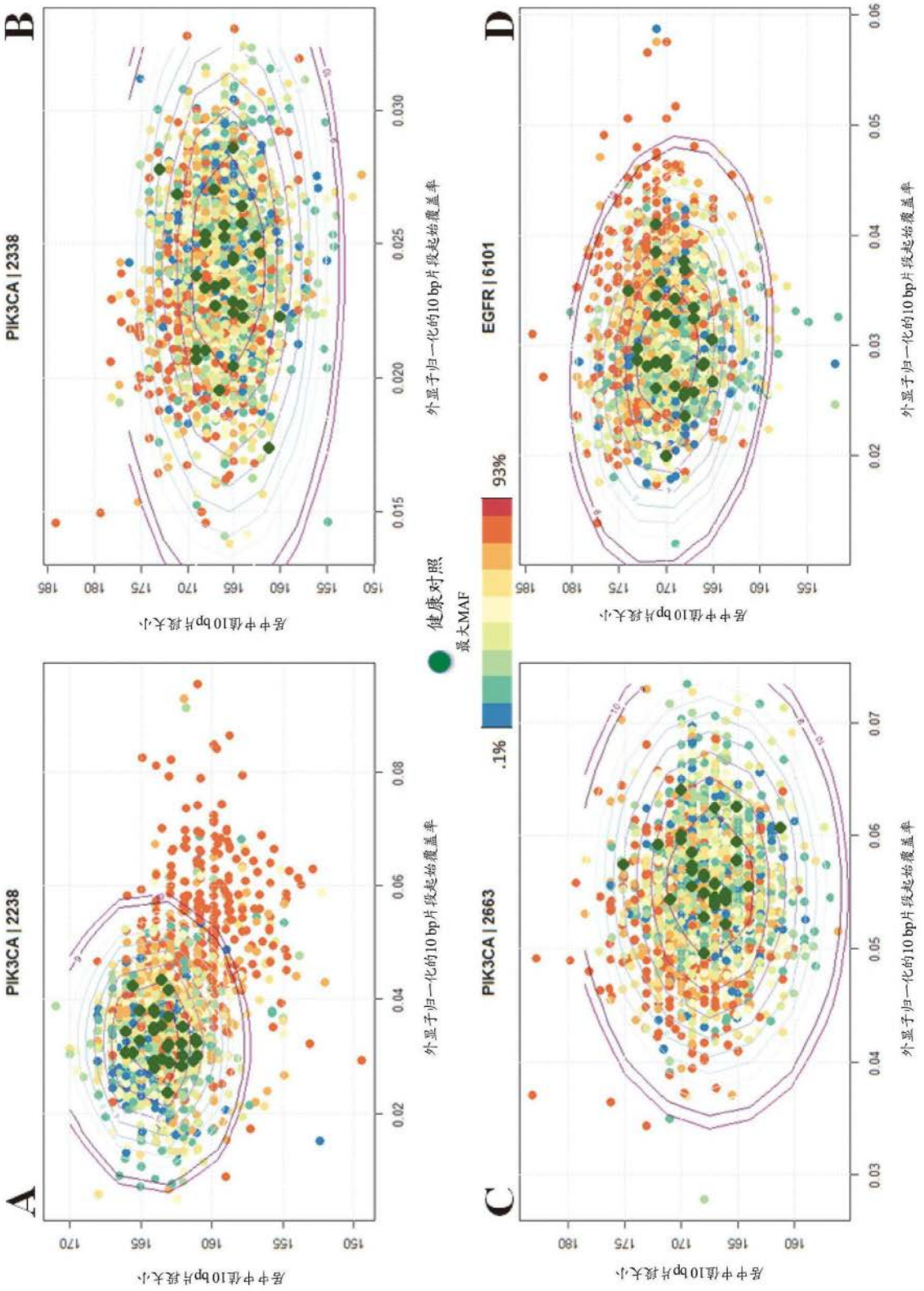

A
PIK3CA | 2238
B
PIK3CA | 2338
C
PIK3CA | 2663
D
EGFR | 6101
居中中值10 bp片段大小
外显子归一化的10 bp片段起始覆盖率
健康对照
最大MAF
93%
.1%

图3

	样品id	基因	改变	百分比	拷贝数	类型	突变类型	癌症类型
59791	A2570201_1	RB1	E209*	63.88	NA	SNV	无意义	小细胞肺癌
59792	A2570201_1	PIK3CA	AMP	0.00	3.821532	CNV	amp	小细胞肺癌
59794	A2570201_1	CCND2	AMP	0.00	2.670914	CNV	amp	小细胞肺癌
59795	A2570201_1	MYC	AMP	0.00	2.586206	CNV	amp	小细胞肺癌
59797	A2570201_1	CCND1	AMP	0.00	2.585685	CNV	amp	小细胞肺癌
59798	A2570201_1	CDK4	AMP	0.00	2.486675	CNV	amp	小细胞肺癌
59793	A2570201_1	ERBB2	AMP	0.00	2.425024	CNV	amp	小细胞肺癌
59796	A2570201_1	KRAS	AMP	0.00	2.384474	CNV	amp	小细胞肺癌
59799	A2570201_1	CDKN2A	G23G	0.28	NA	SNV	同义的	小细胞肺癌

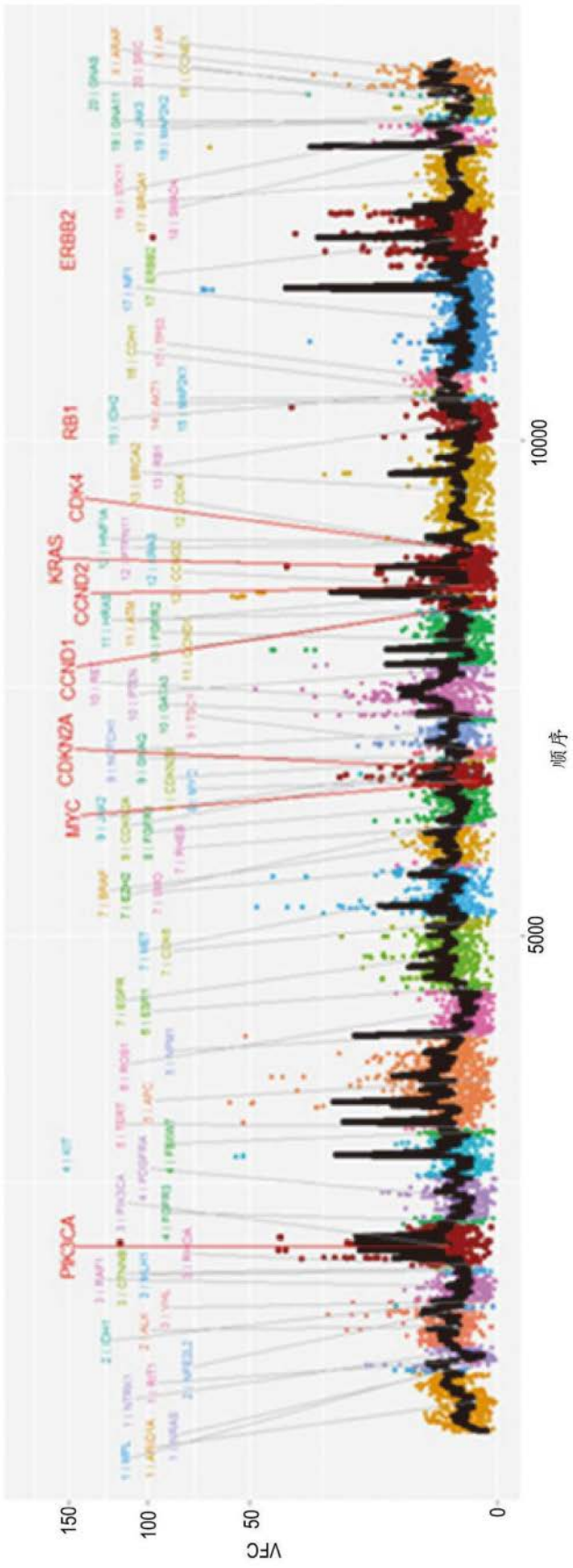


图4

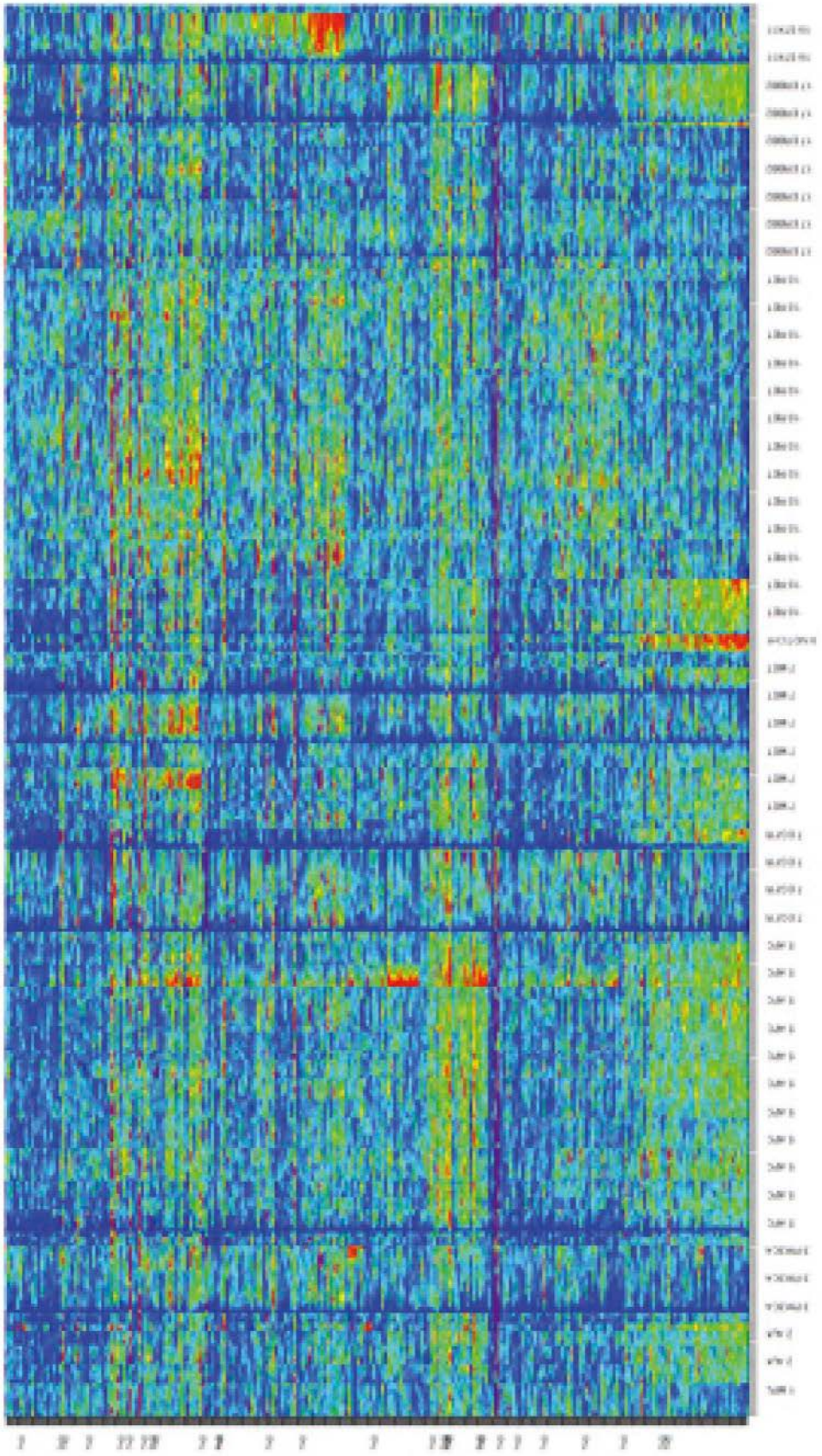


图5

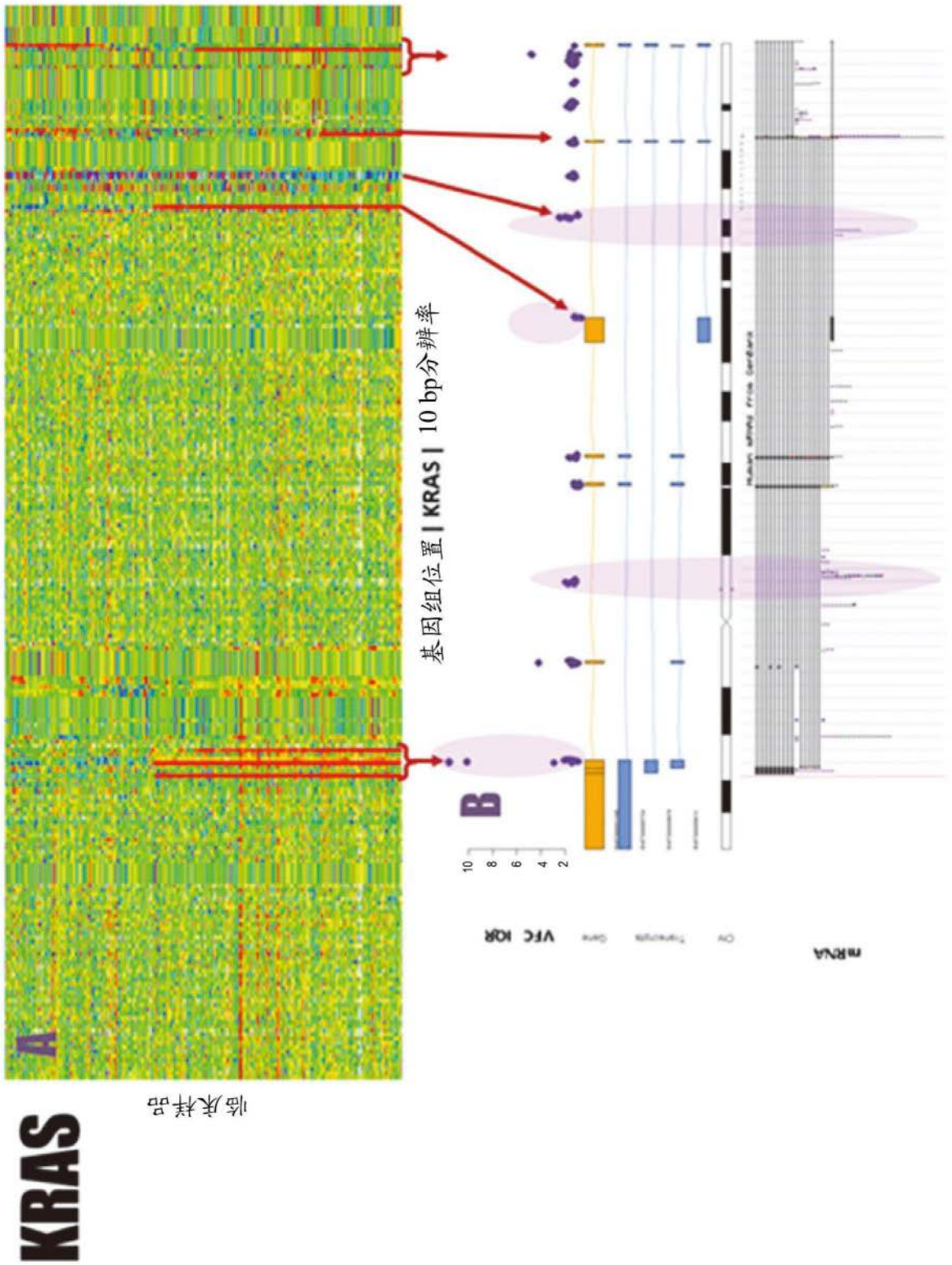


图6

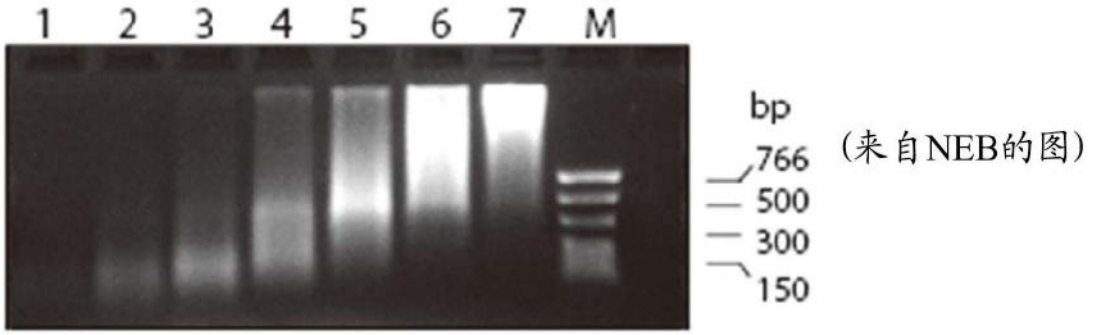

1
2
3
4
5
6
7
M
bp
766
500
300
150
(来自NEB的图)

图7

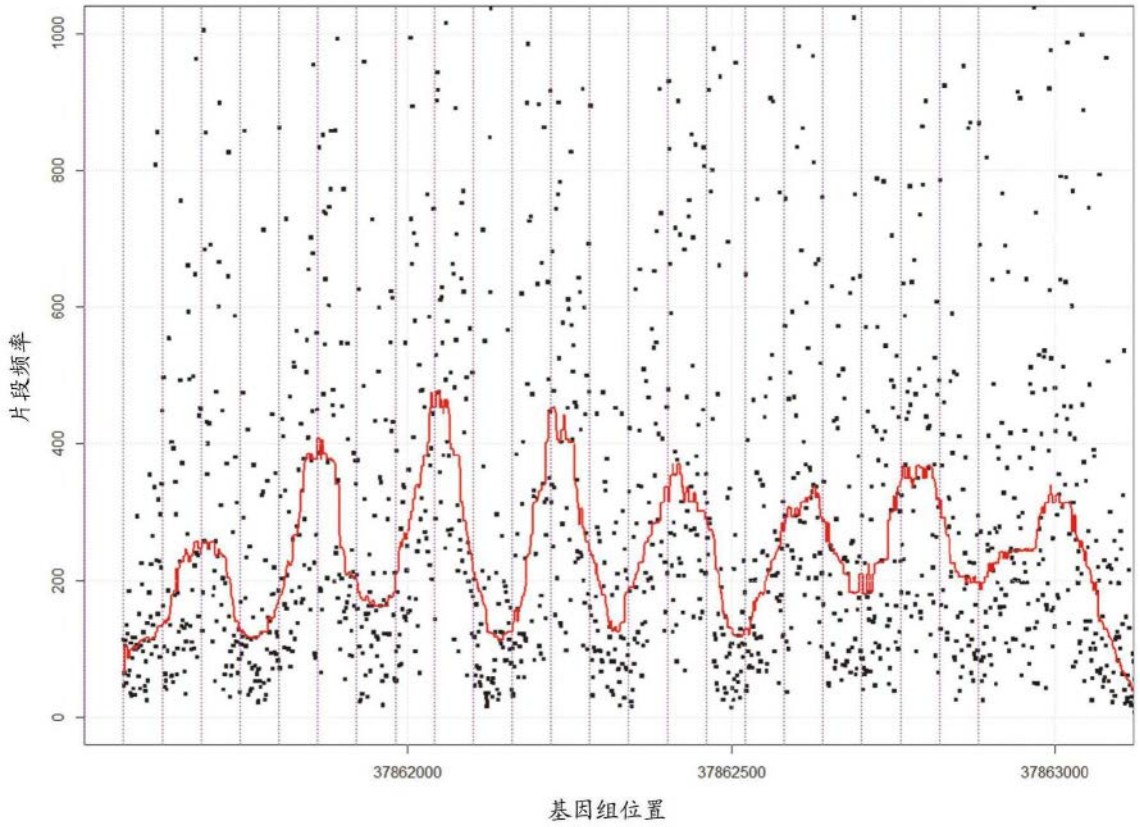

片段频率
0
200
400
600
800
1000
37862000
37862500
37863000
基因组位置

图8

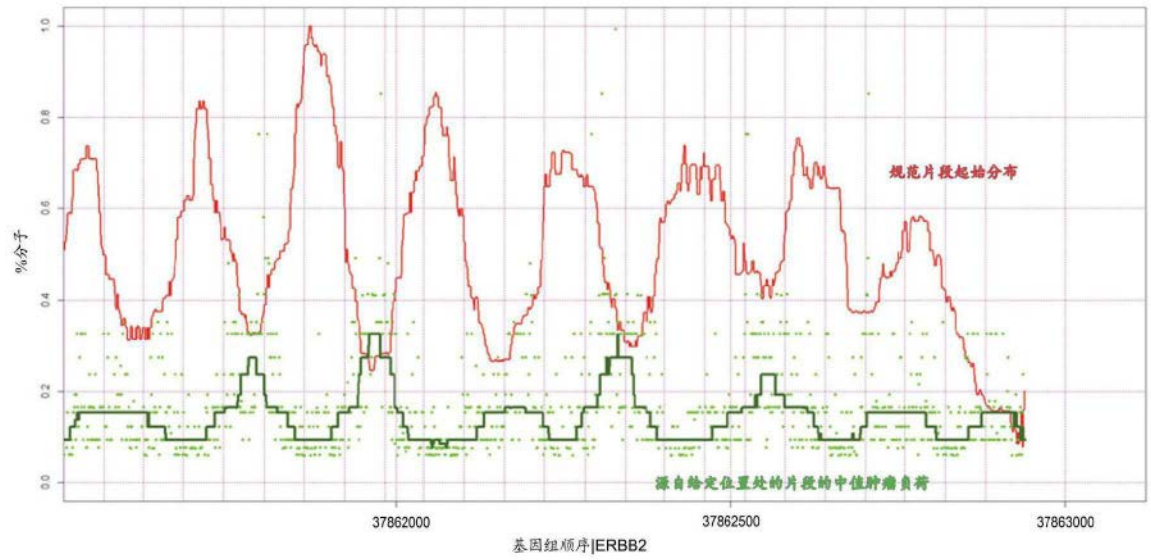

规范片段起始分布
源自给定位置处的片段的中值肿瘤负荷
%分子
37862000
37862500
37863000
基因组顺序|ERBB2

图9

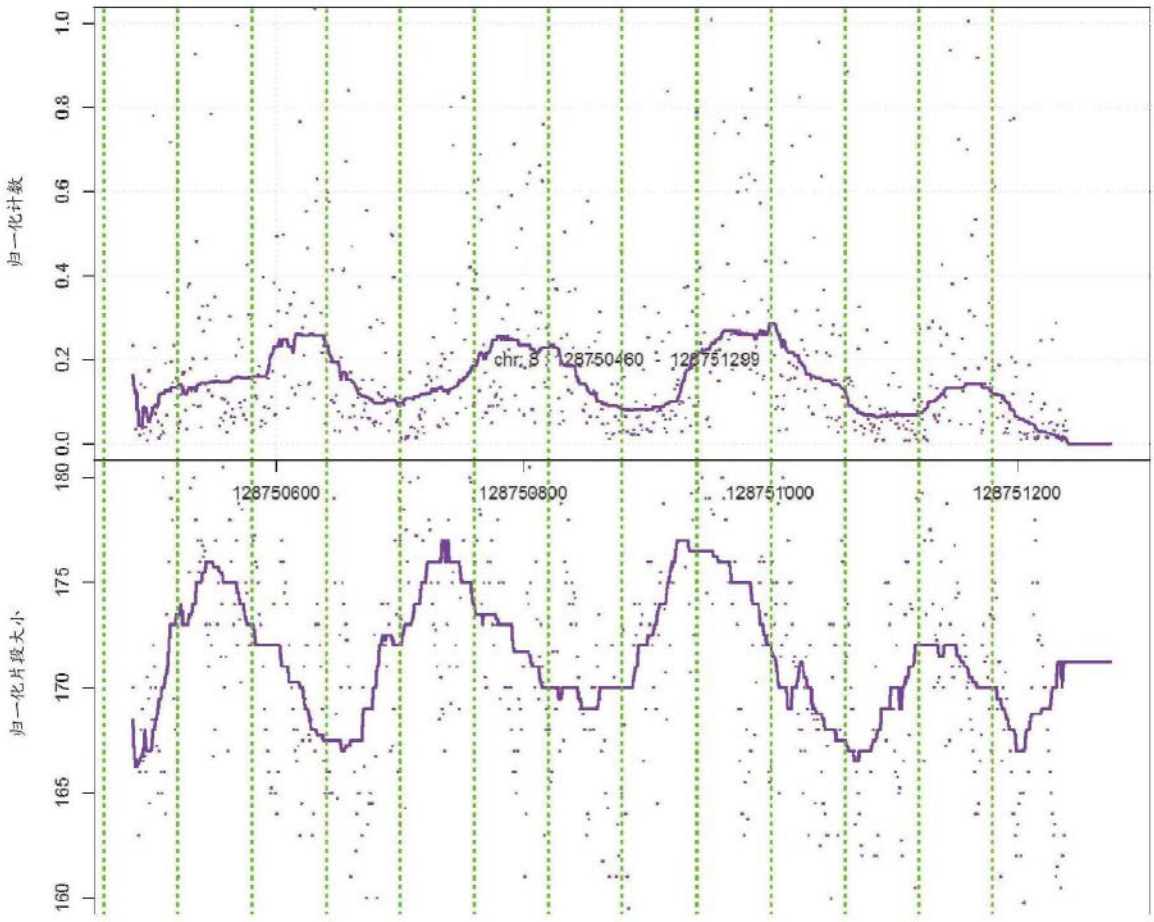

归一化计数
chr: 8 128750460 - 128751299
128750600
128750800
128751000
128751200
归一化片段大小

图10

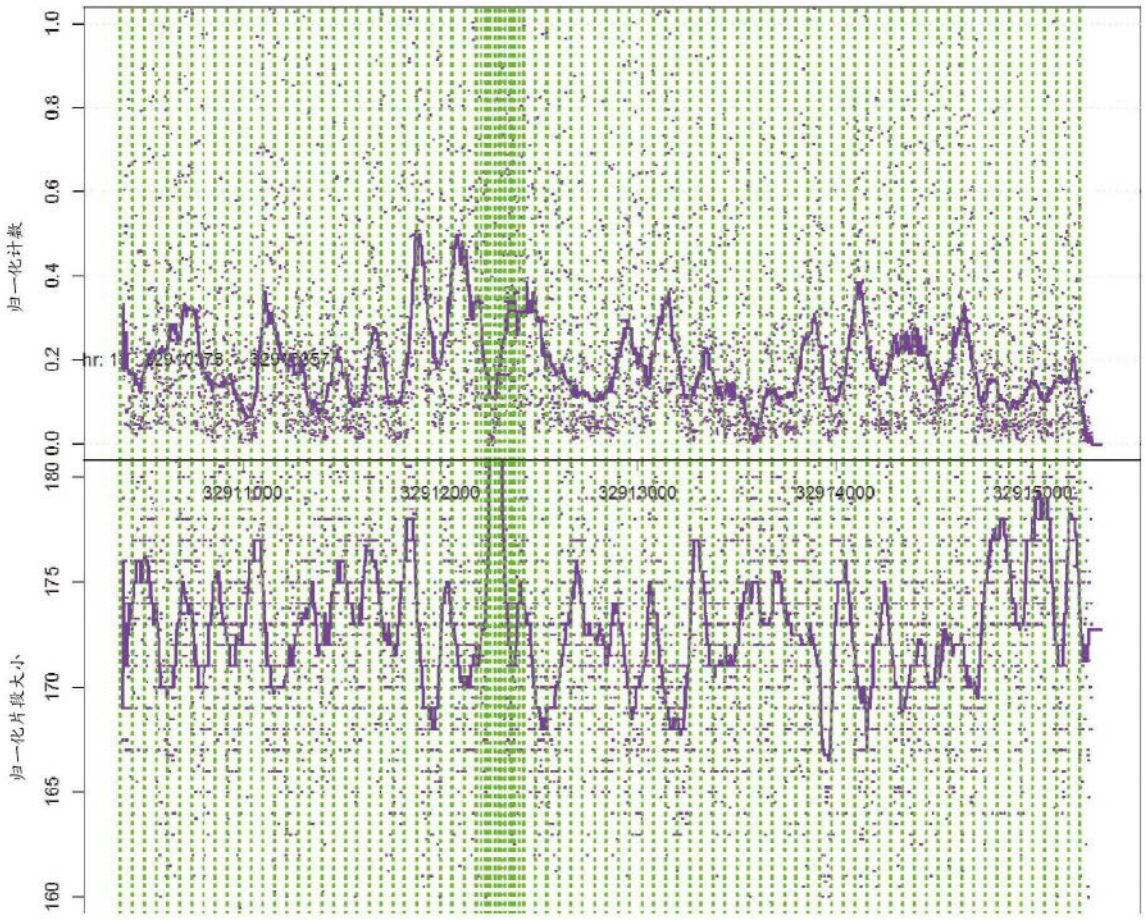

归一化计数
1.0
0.8
0.6
0.4
0.2
0.0
32911000
32912000
32913000
32914000
32915000
归一化片段大小
180
175
170
165
160

图11

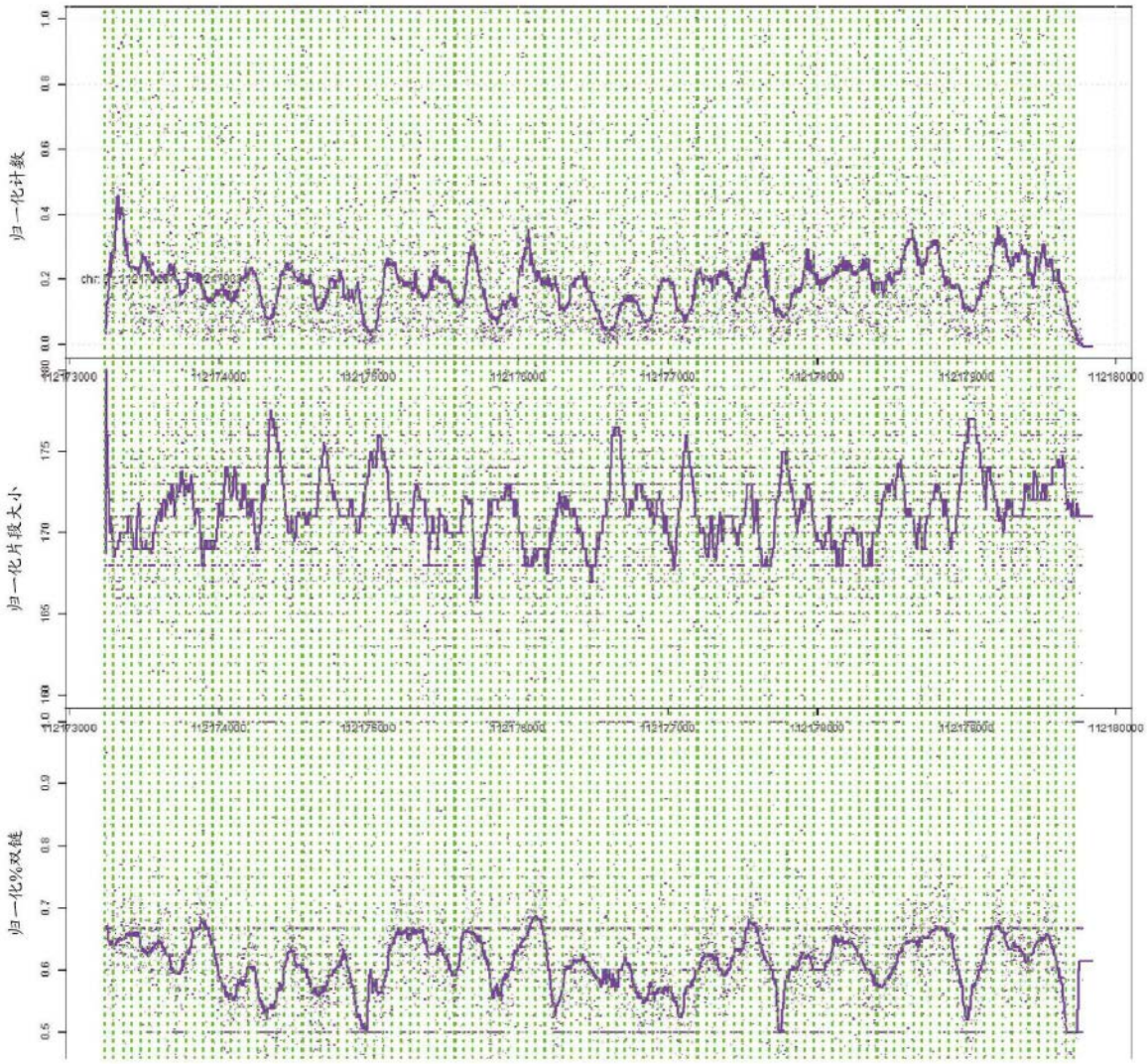


图12

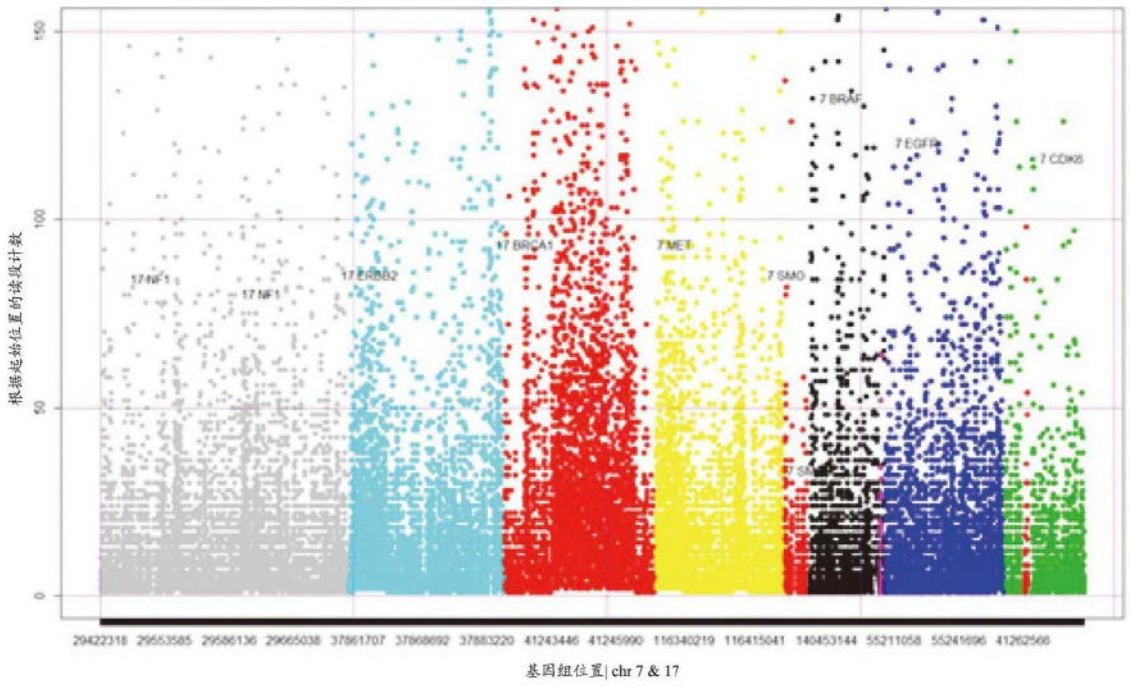

根据起始位置的读段计数
17 NF1
17 NF1
17 ERBB2
17 BRCA1
7 MET
7 SMO
7 BRAF
7 EGFR
7 CDK6
29422318 29553585 29586136 29665038 37861707 37868692 37883220 41243446 41245990 116340219 116415041 140453144 55211058 55241696 41262566
基因组位置| chr 7 & 17

图13

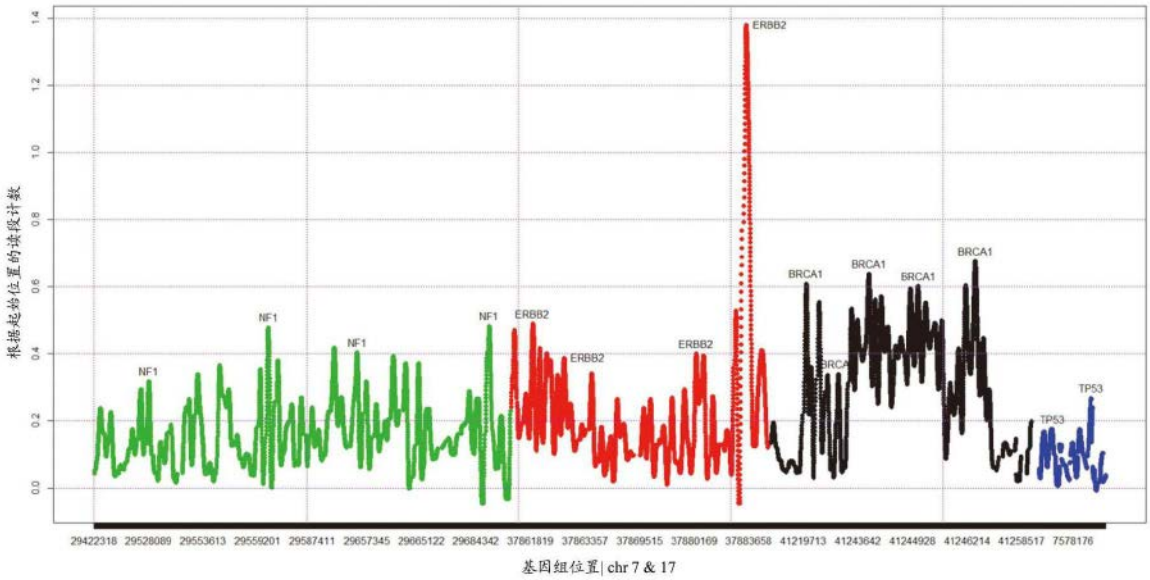

根据起始位置的读段计数
NF1
NF1
NF1
NF1
ERBB2
ERBB2
ERBB2
ERBB2
BRCA1
BRCA1
BRCA1
BRCA1
TP53
TP53
29422318 29528089 29553613 29559201 29587411 29657345 29665122 29684342 37861819 37863357 37869515 37880169 37883658 41219713 41243642 41244928 41246214 41258517 7578176
基因组位置| chr 7 & 17

图14

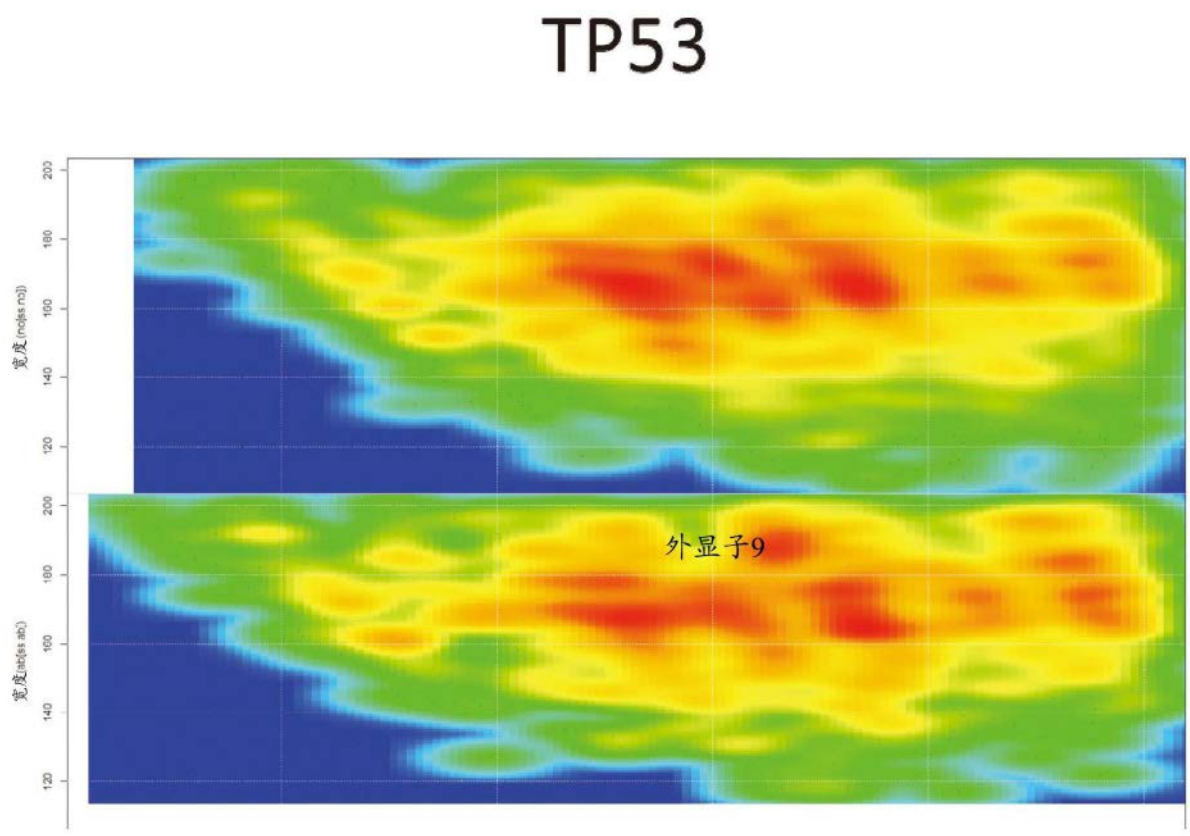


图15

NF1

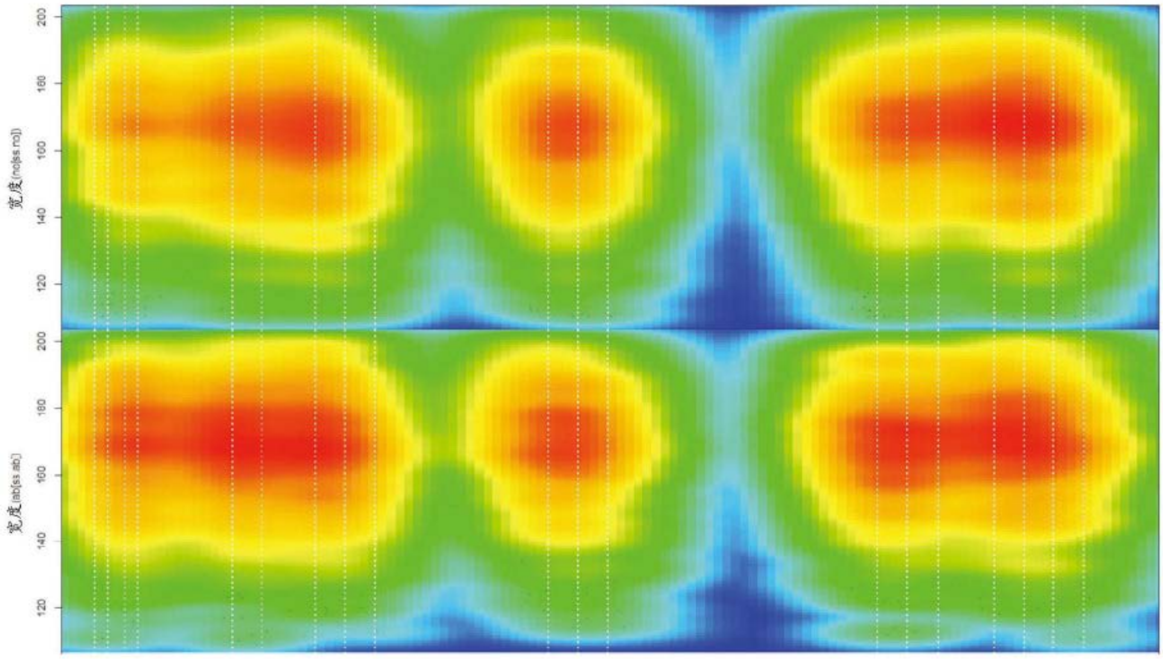


图16

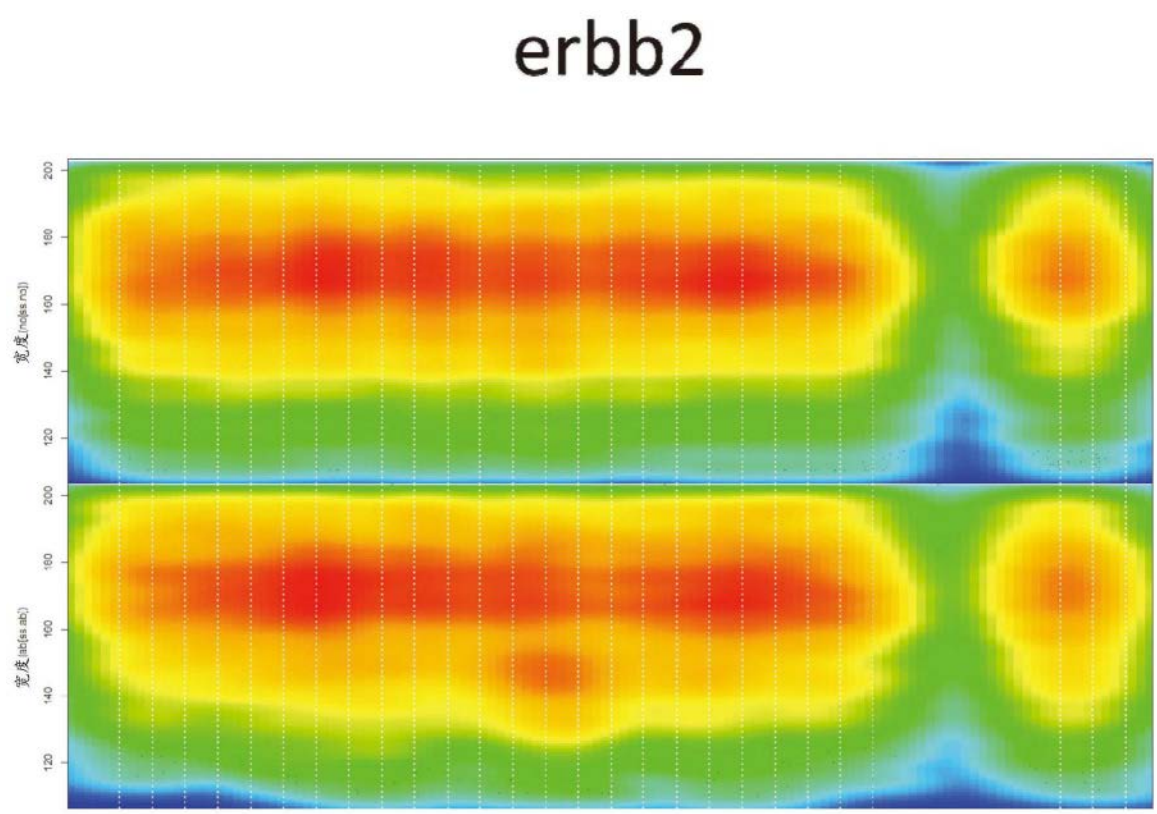

erbb2
宽度(nobs no))
宽度(nobs ab)
120
140
160
180
200

图17

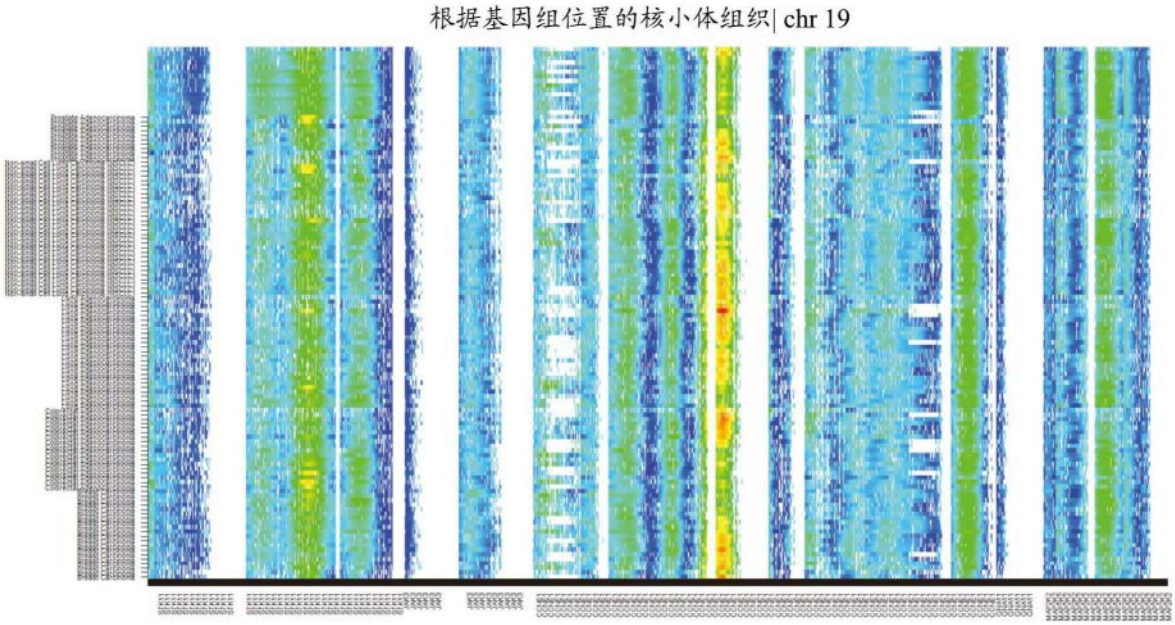

根据基因组位置的核小体组织| chr 19

图18

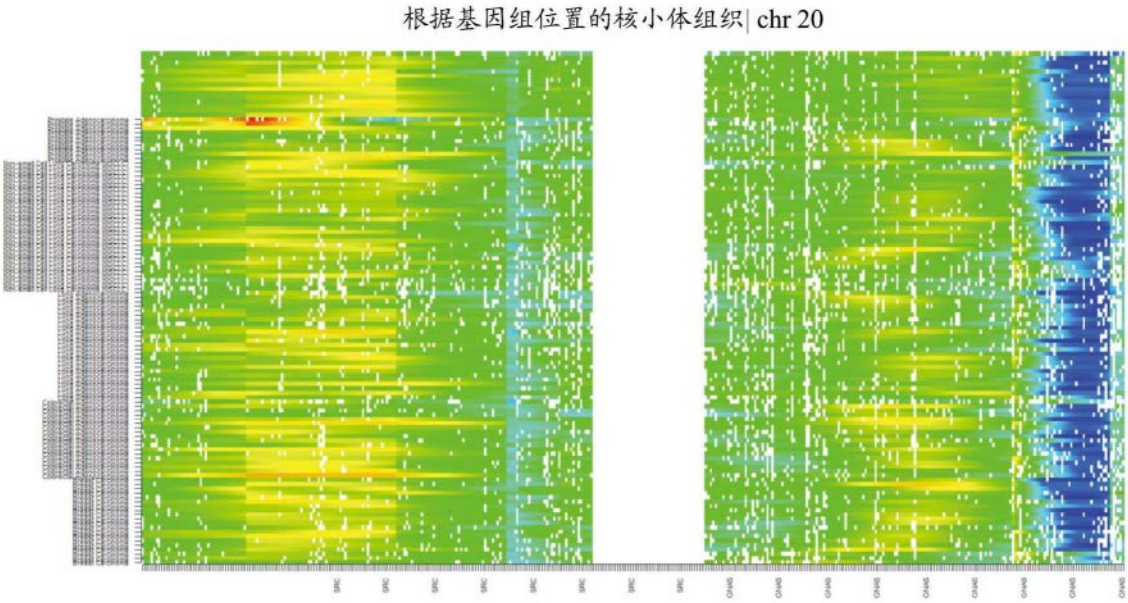
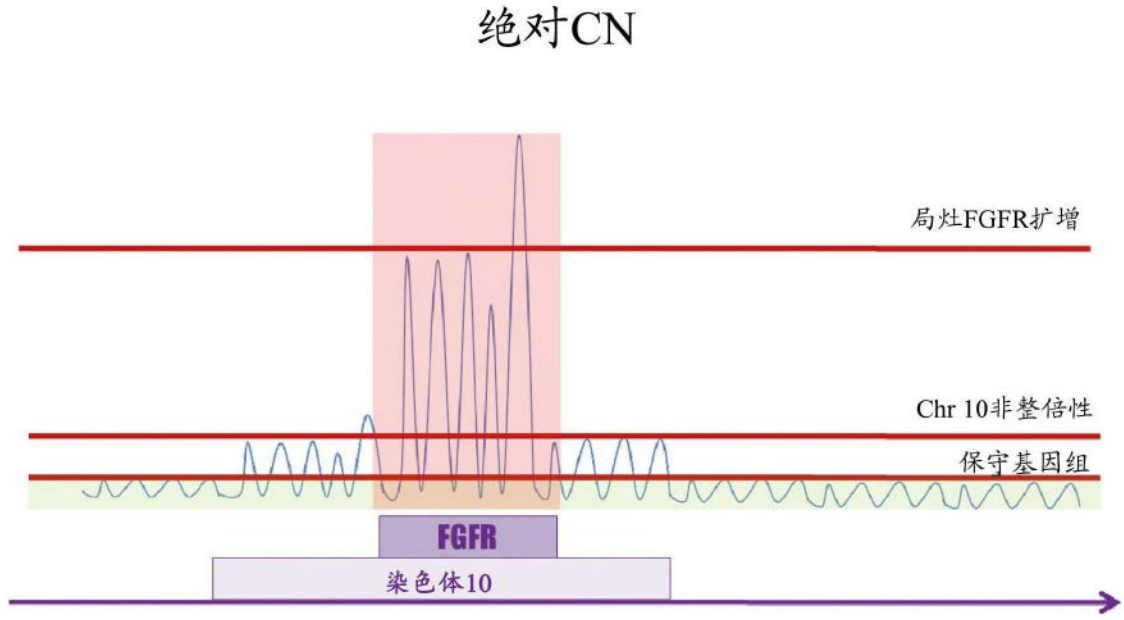


图19



定位核小体位置并将它们与正常群组中预期的匹配
对于FGFR中的每个核小体窗口
确定超保守的非chr10核小体位点的集合(基因组UC)
确定超保守的chr10核小体位点的集合（ch10 UC）
跨FGFR核小体位点的位置与插入物大小密度进行整合-（UC）x非整倍性因子（ch10 UC /基因组UC）

图20

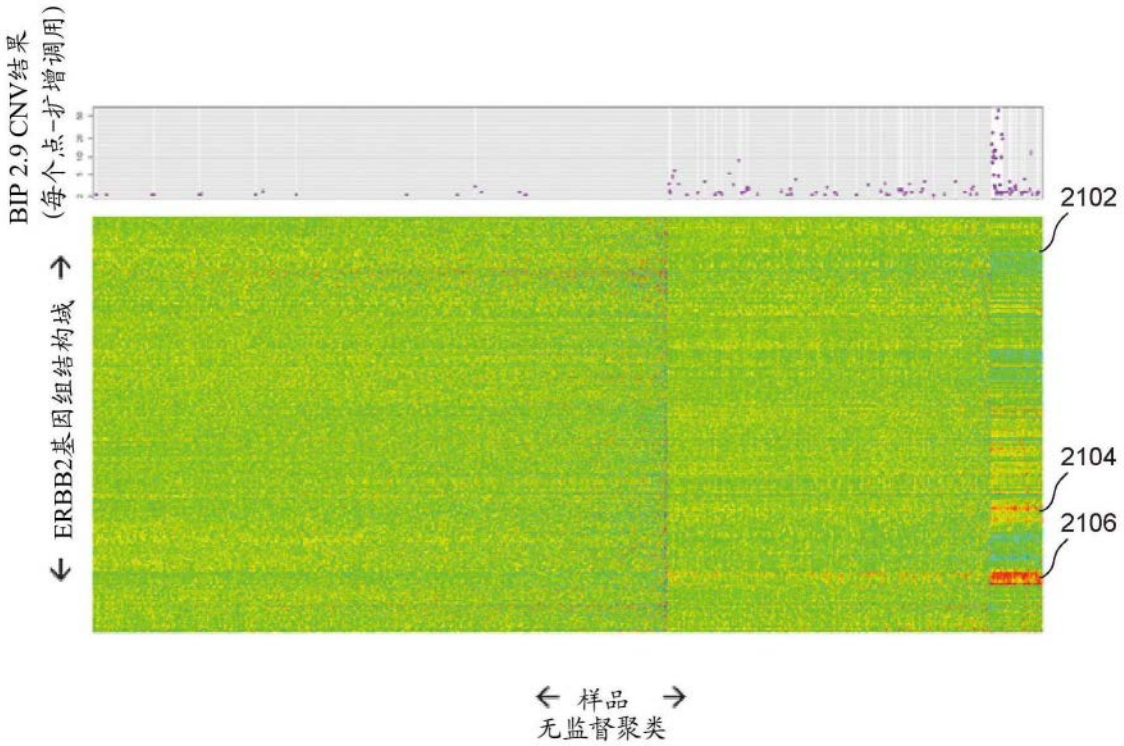

BIP 2.9 CNV结果
(每个点-扩增调用)
ERBB2基因组结构域
2102
2104
2106
样品
无监督聚类

图21A

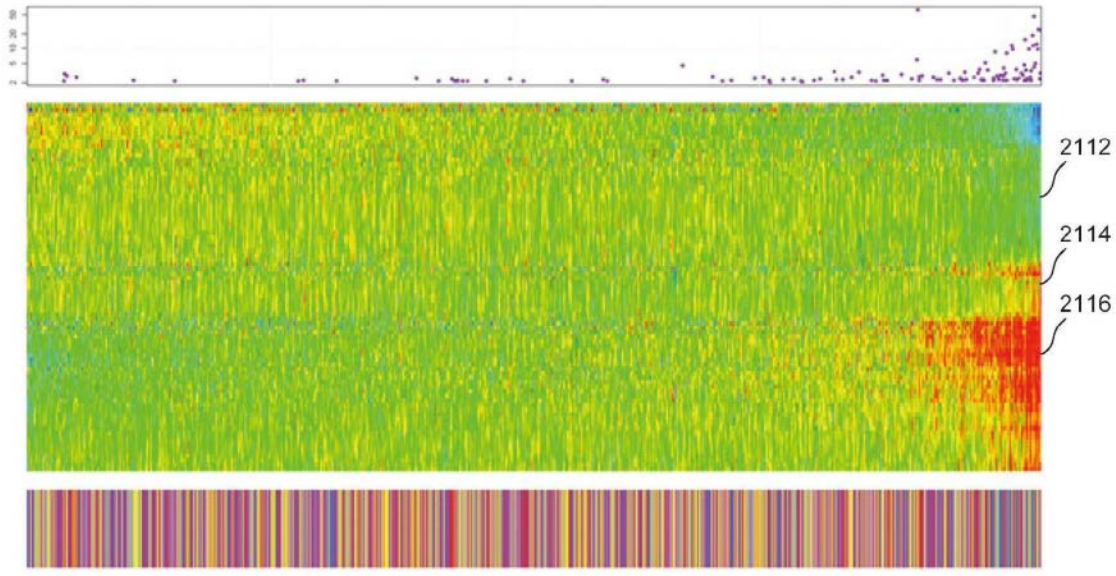

2112
2114
2116

图21B

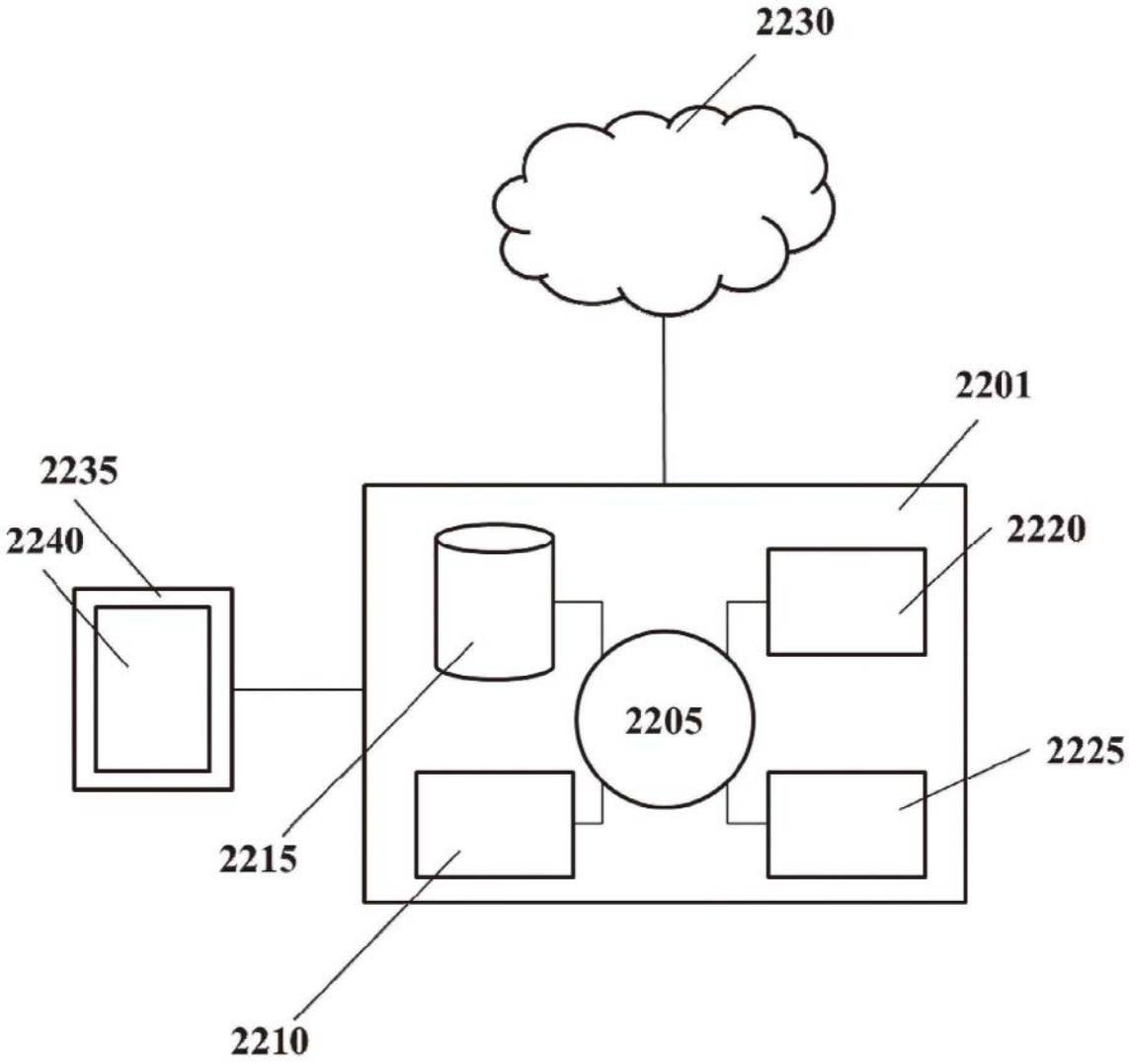


图22

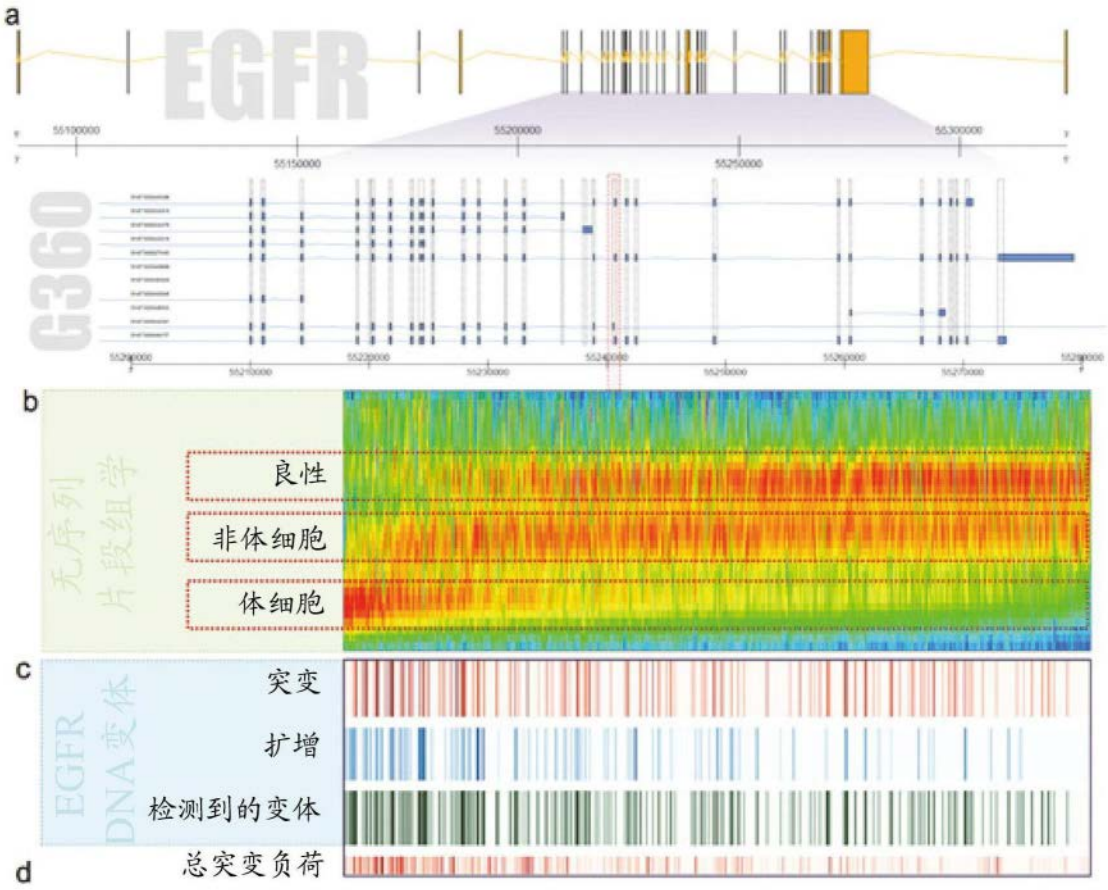


图23

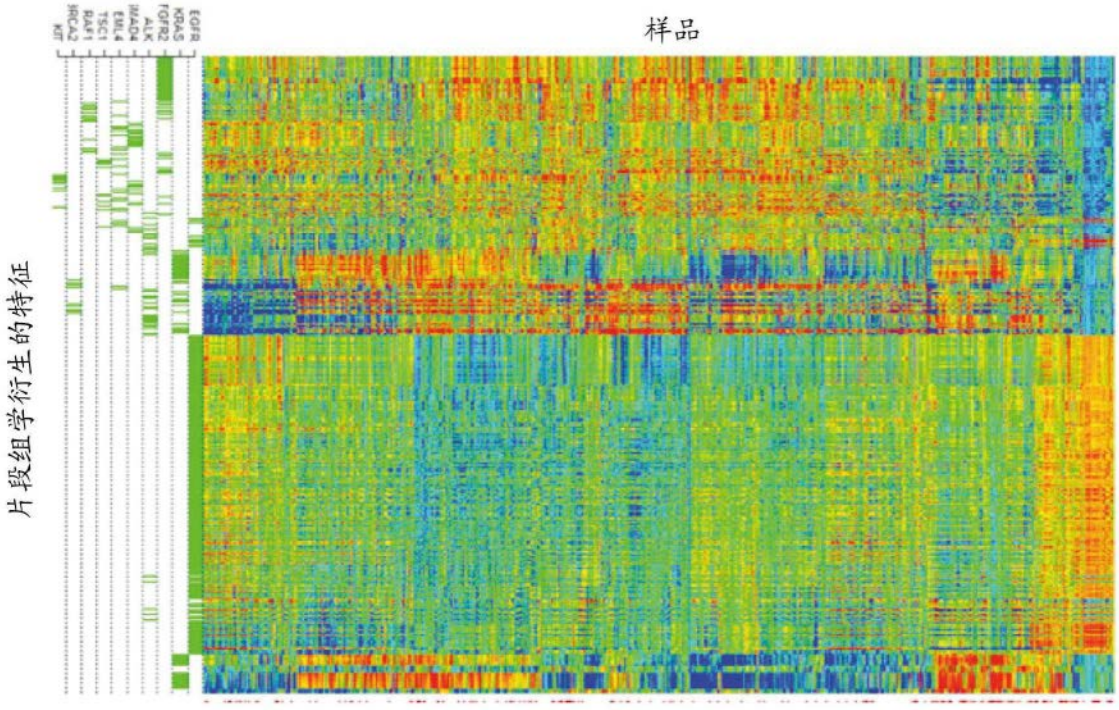


图24

$$p(\mathbf{X}=\mathbf{x}\mid\theta)=\sum_{k=1}^{K}\pi_k\, p(\mathbf{X}=\mathbf{x}\mid C=c_k,\theta_k)$$

其中：

X为代表个体DNA片段单元的多元随机变量

C为分类变量，其值来自$\{c_1 \dots c_k\}$，其中一个c为可能癌症

π_k为第k类的边际概率（权重）

Θ_k为针对每列/变量的含有概率分布的第k类的参数估计值

图25

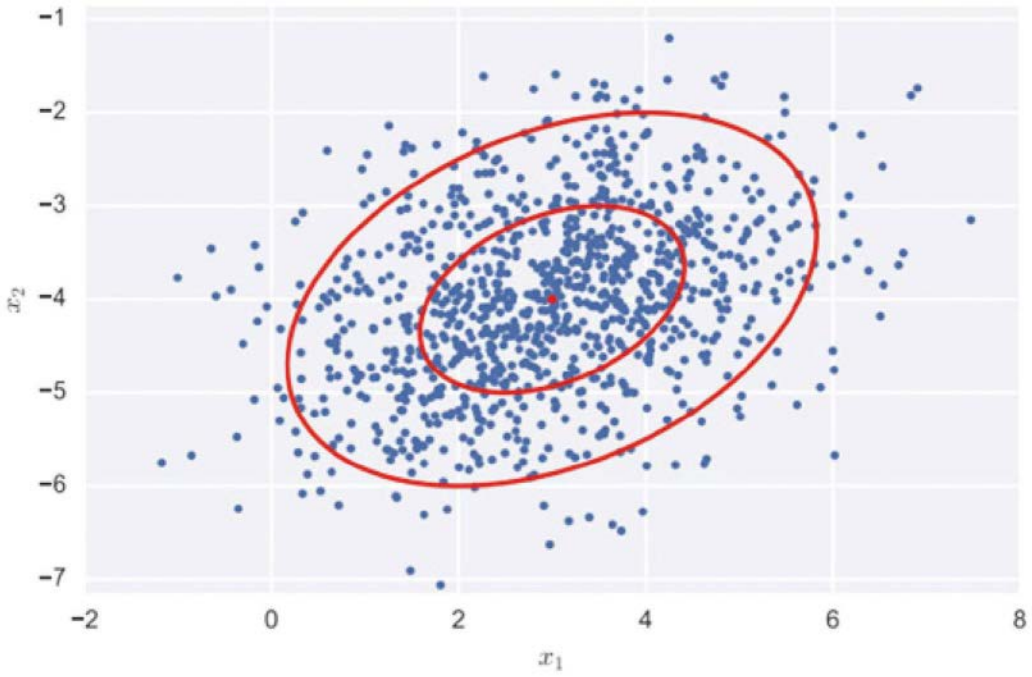

-1
-2
-3
-4
-5
-6
-7
x_2
-2
0
2
4
6
8
x_1

图26A

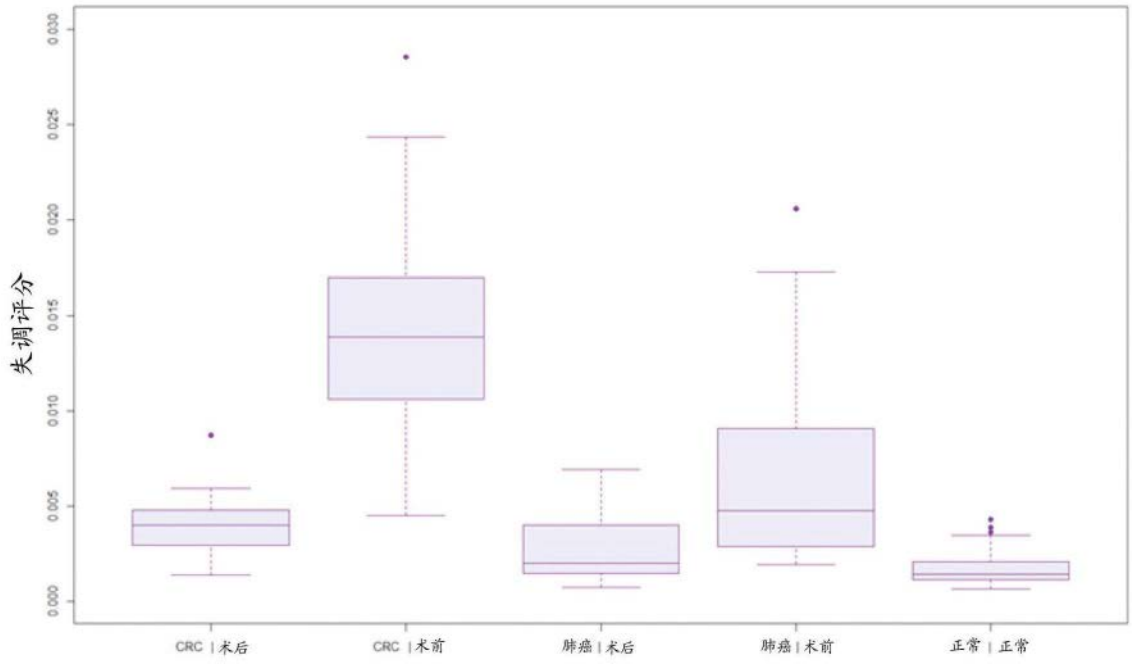

失调评分
0.000
0.005
0.010
0.015
0.020
0.025
0.030
CRC | 术后
CRC | 术前
肺癌 | 术后
肺癌 | 术前
正常 | 正常

图26B

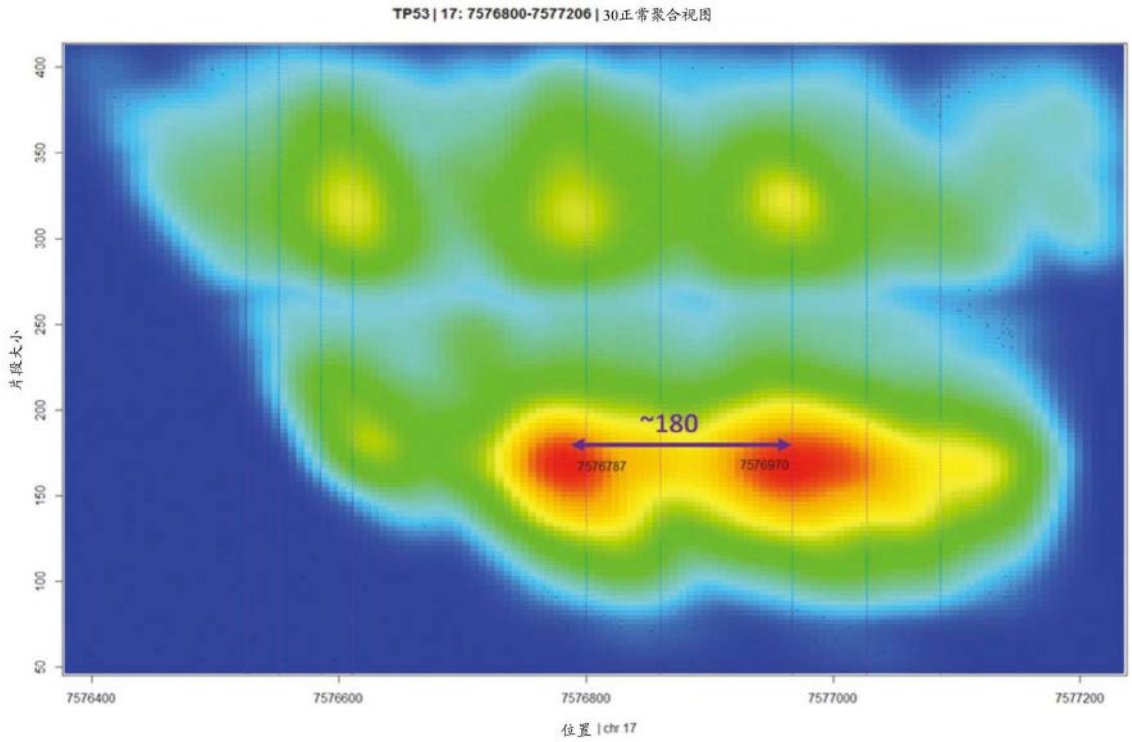

TP53 | 17: 7576800-7577206 | 30正常聚合视图
片段大小
400
350
300
250
200
150
100
50
~180
7576787
7576970
7576400
7576600
7576800
7577000
7577200
位置 | chr 17

图27A

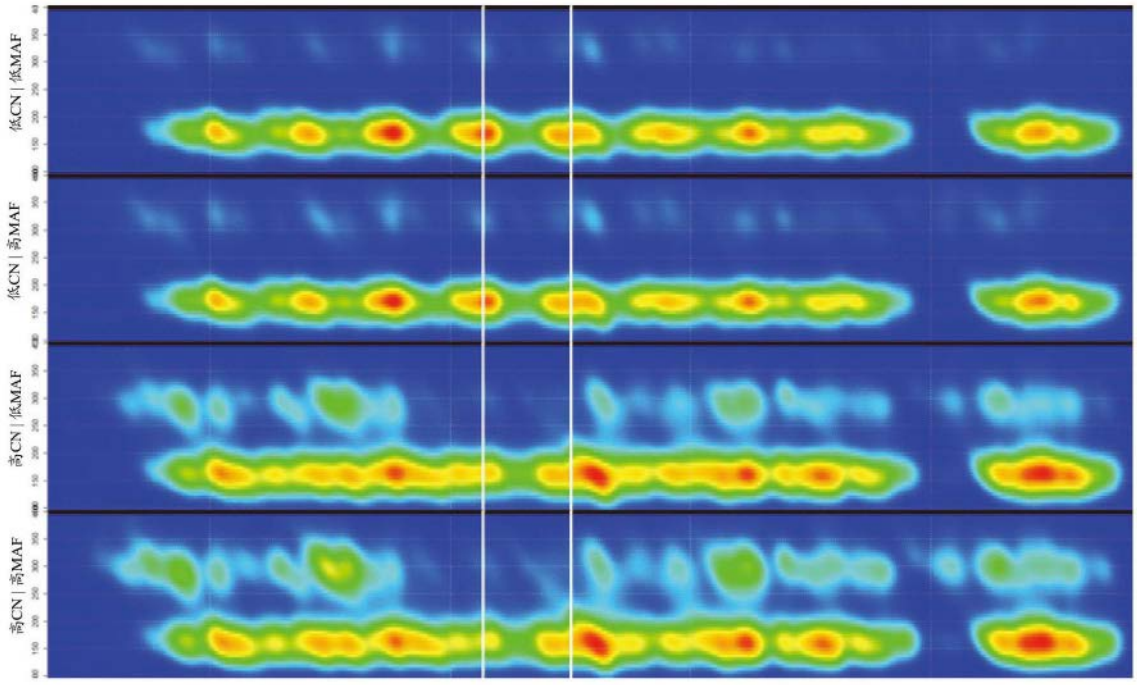

低CN | 低MAF
低CN | 高MAF
高CN | 低MAF
高CN | 高MAF

图27B

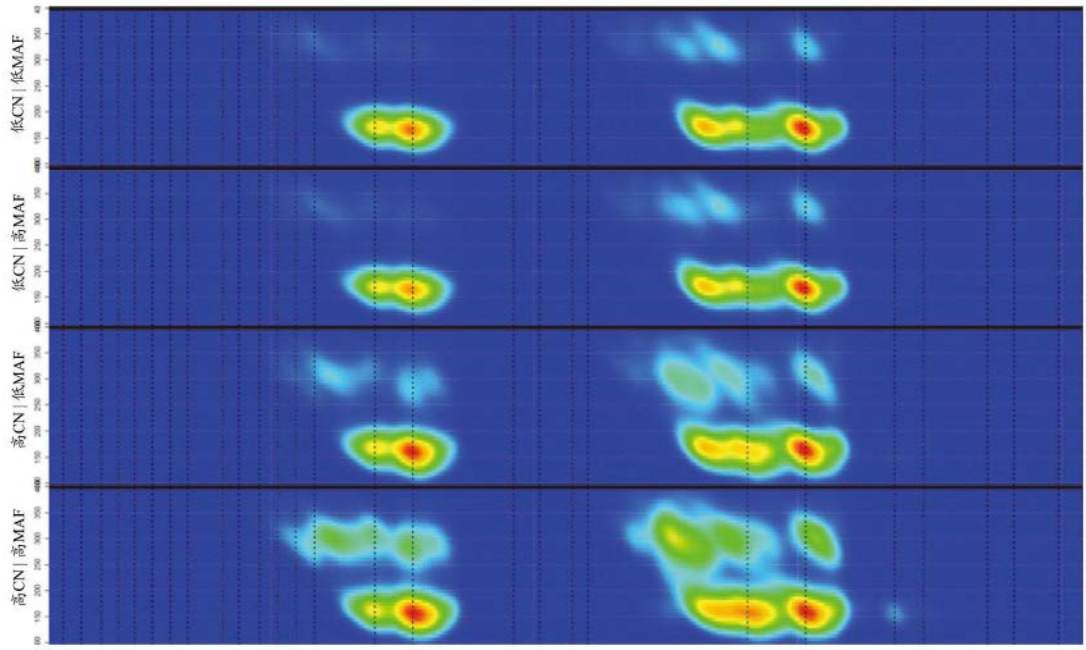

低CN | 低MAF
低CN | 高MAF
高CN | 低MAF
高CN | 高MAF

图27C

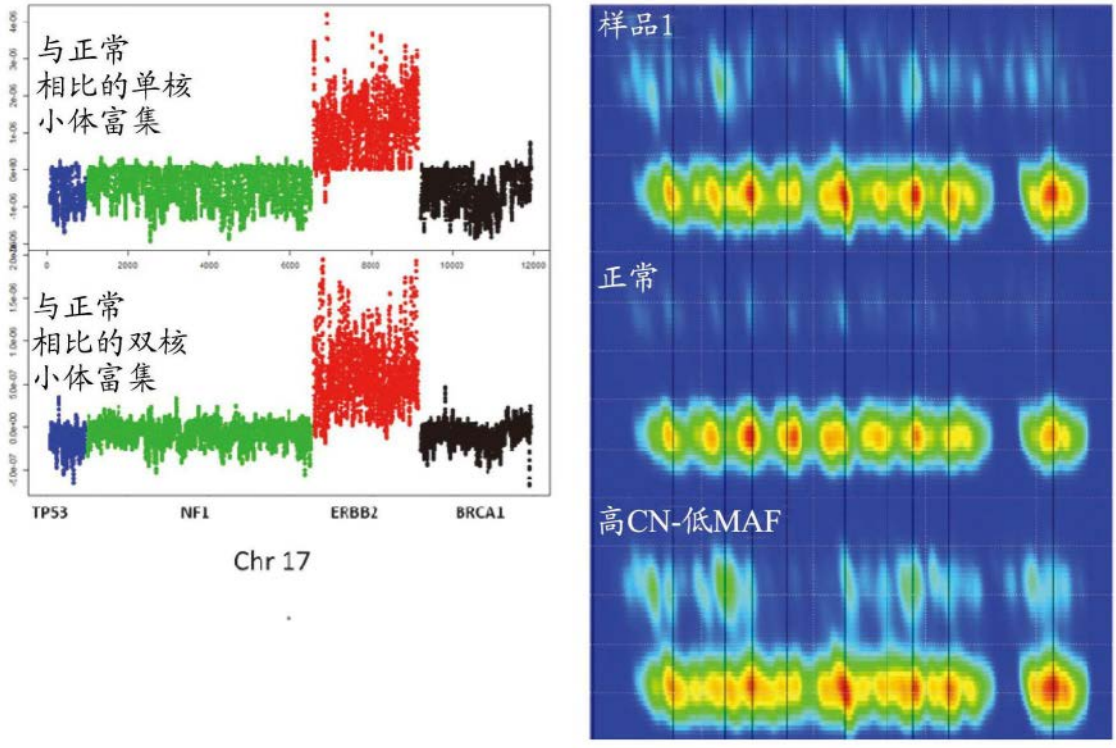

与正常
相比的单核
小体富集
与正常
相比的双核
小体富集
TP53
NF1
ERBB2
BRCA1
Chr 17
样品1
正常
高CN-低MAF

图28A

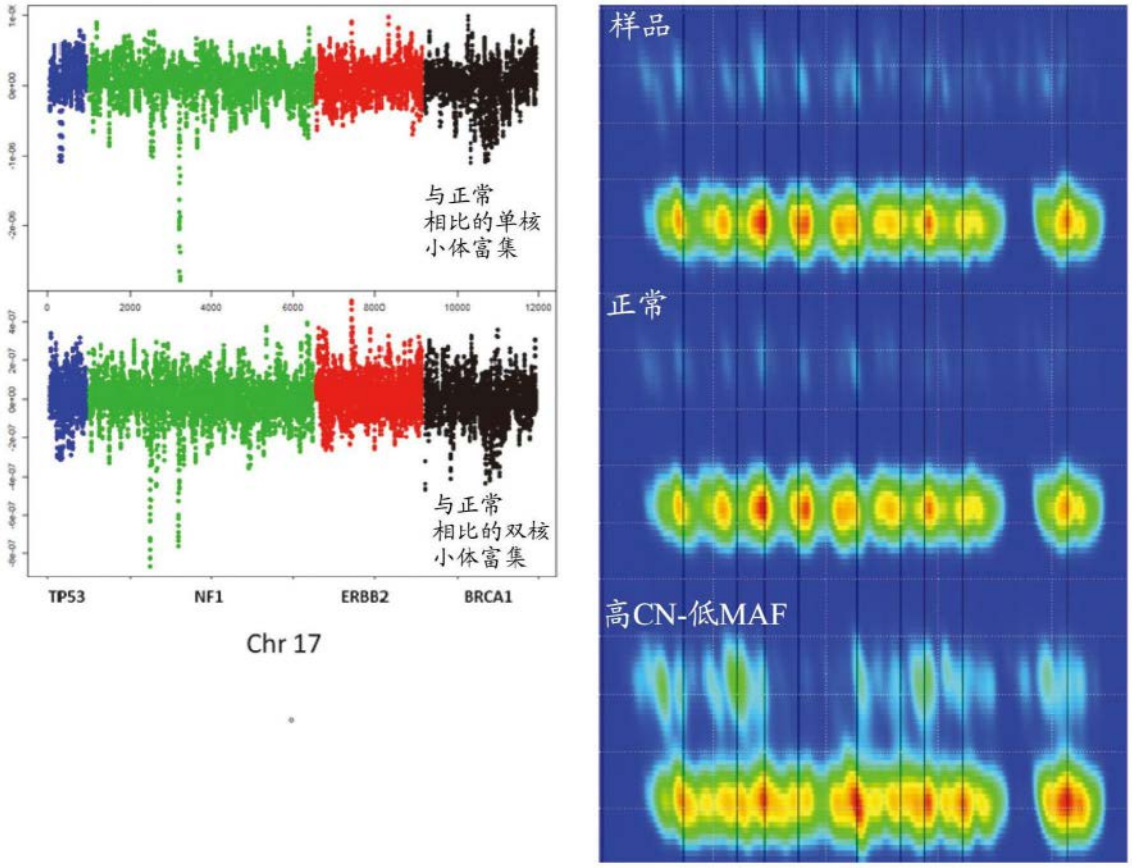

与正常
相比的单核
小体富集
与正常
相比的双核
小体富集
TP53
NF1
ERBB2
BRCA1
Chr 17
样品
正常
高CN-低MAF

图28B

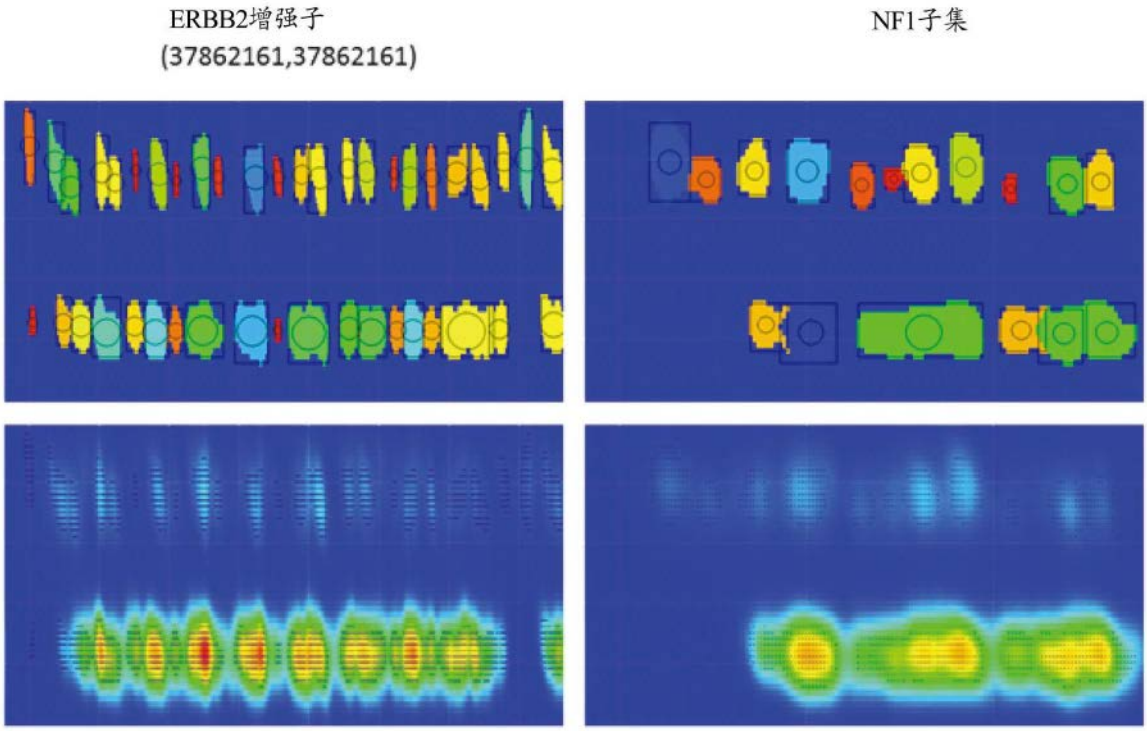


图29A

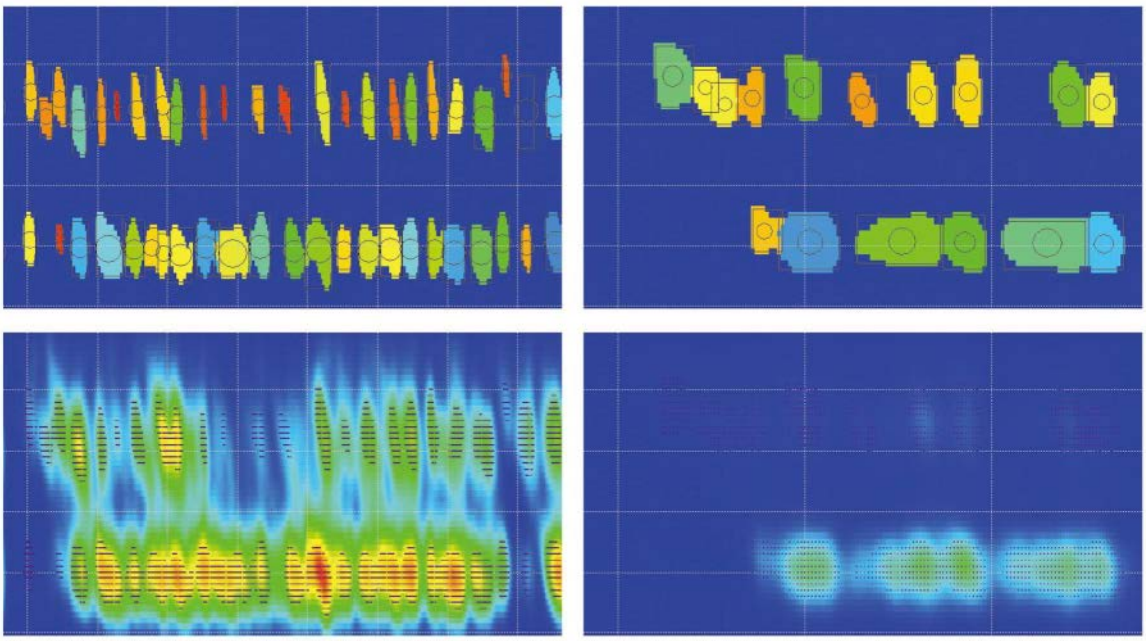


图29B

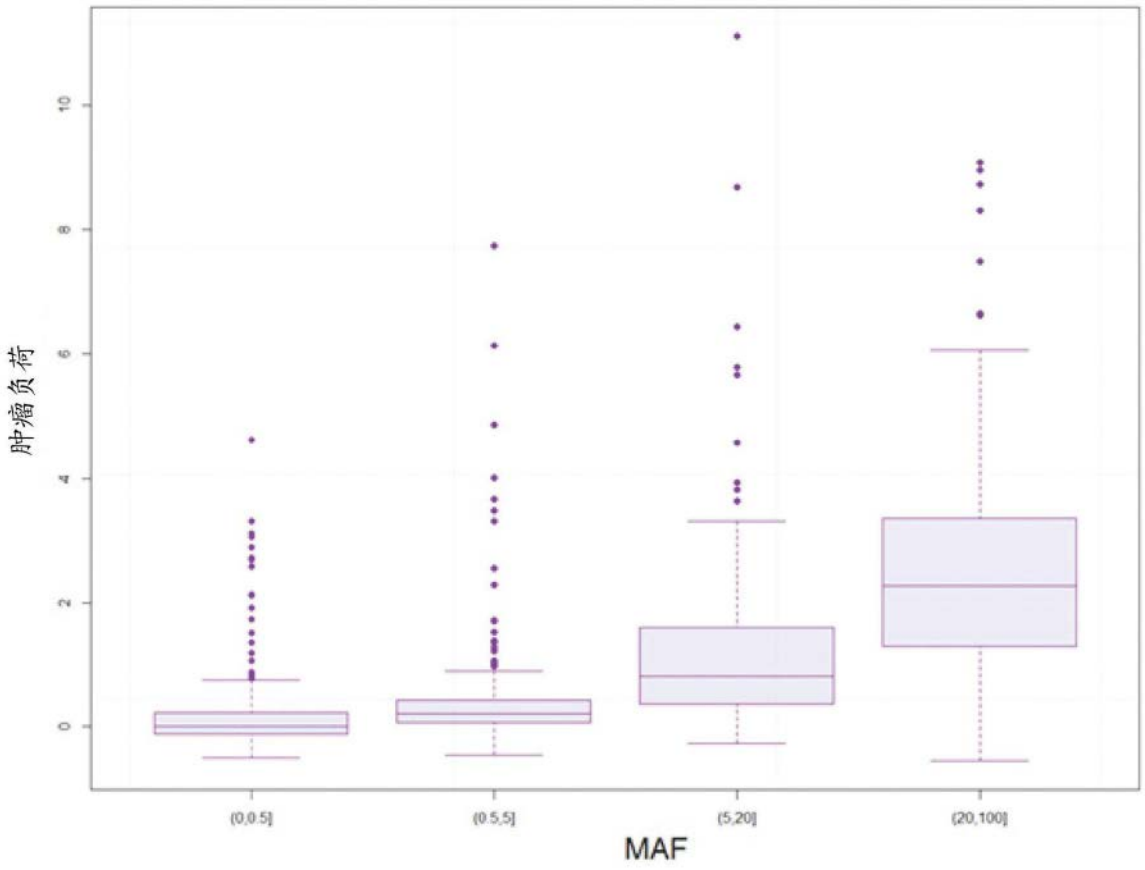


图30

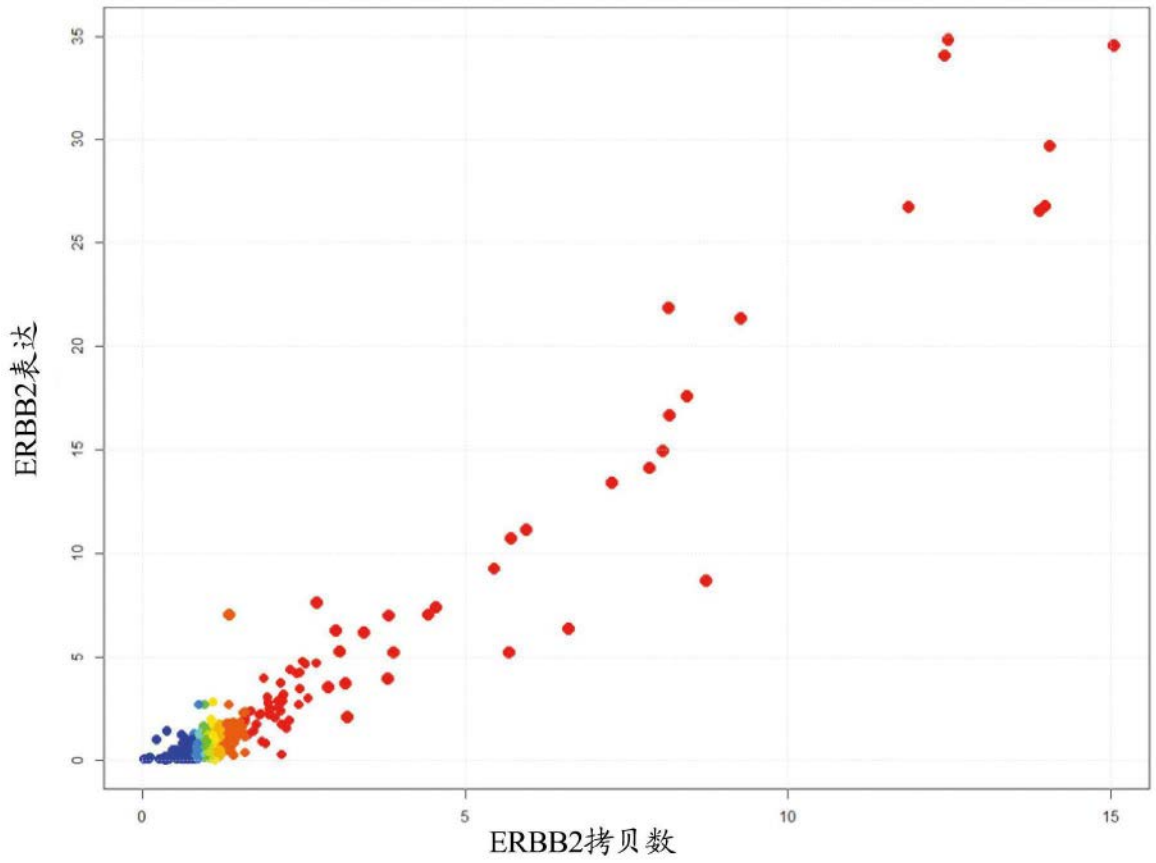


图31A

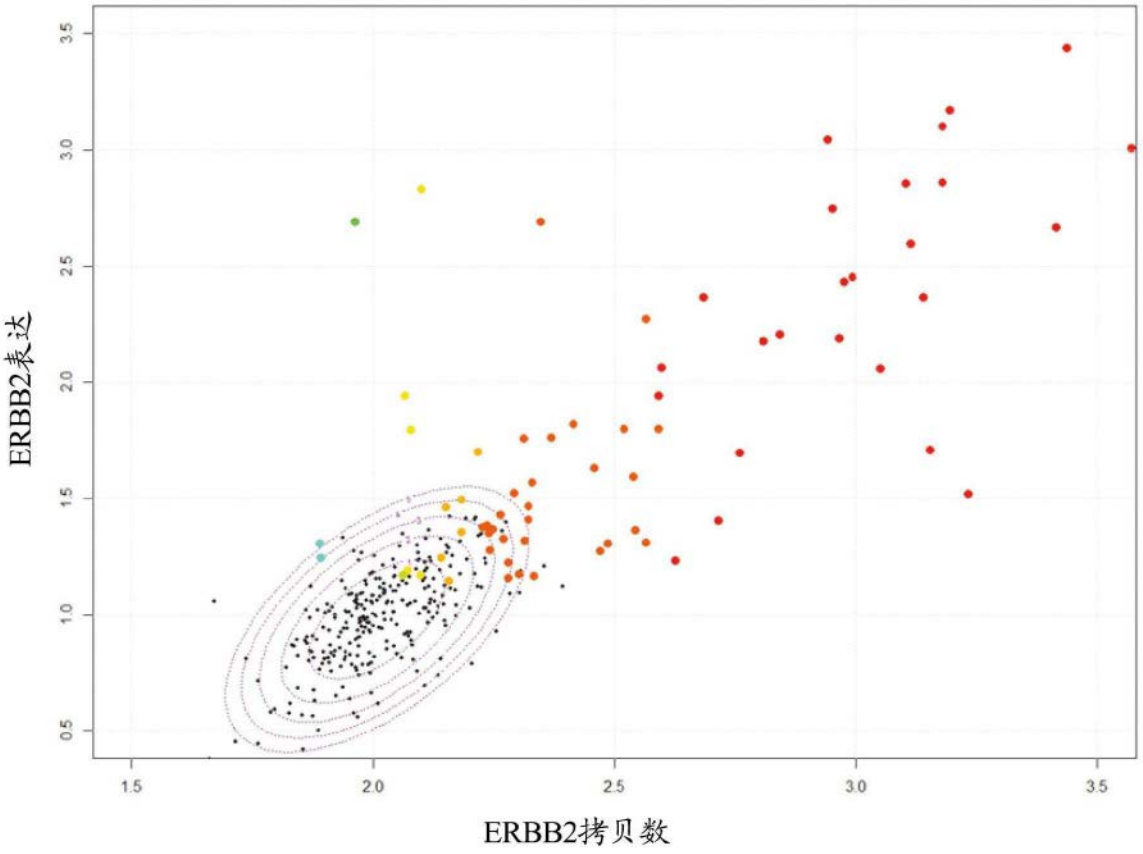

ERBB2表达
ERBB2拷贝数
0.5
1.0
1.5
2.0
2.5
3.0
3.5

图31B

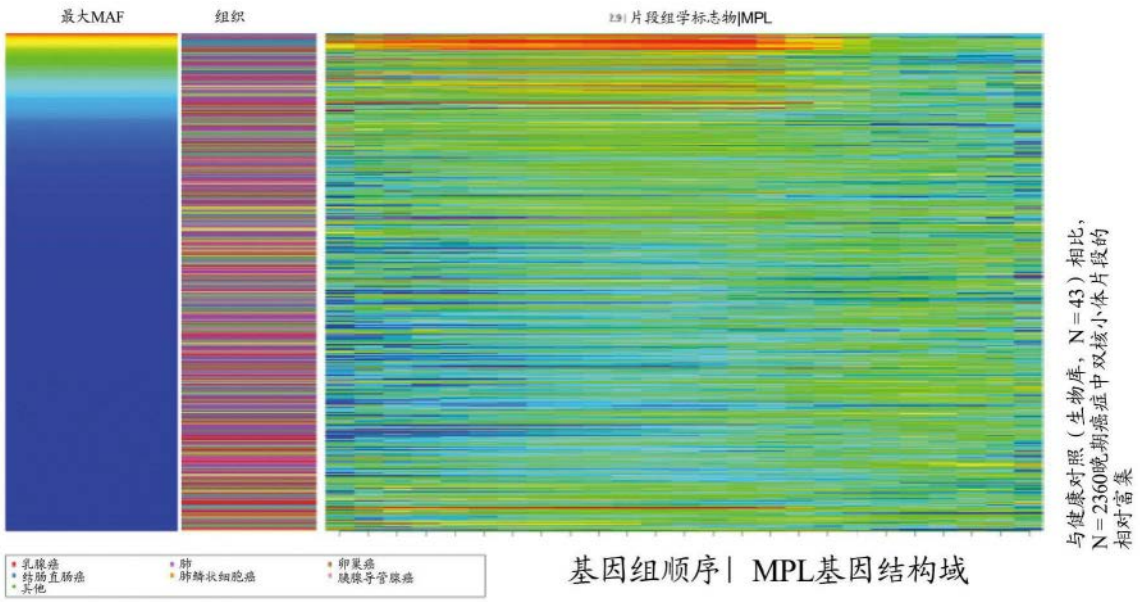

最大MAF
组织
片段组学标志物|MPL
乳腺癌
结肠直肠癌
其他
肺
肺鳞状细胞癌
卵巢癌
胰腺导管腺癌
基因组顺序| MPL基因结构域
与健康对照（生物库，N=43）相比，
N=2360晚期癌症中双核小体片段的
相对富集

图32A

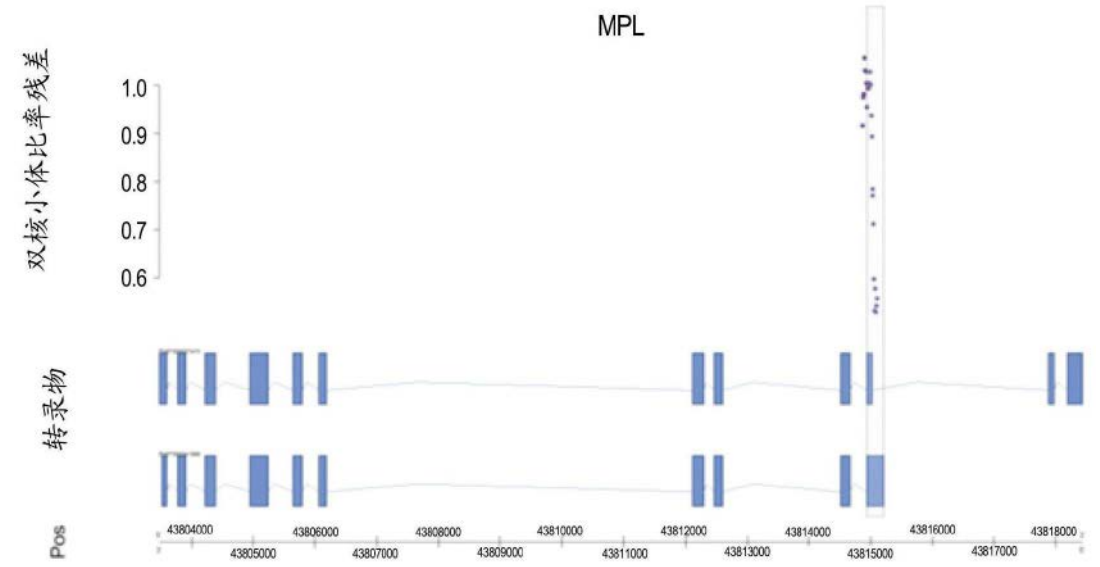

MPL
双核小体比率残差
1.0
0.9
0.8
0.7
0.6
转录物
Pos
43804000
43805000
43806000
43807000
43808000
43809000
43810000
43811000
43812000
43813000
43814000
43815000
43816000
43817000
43818000

图32B

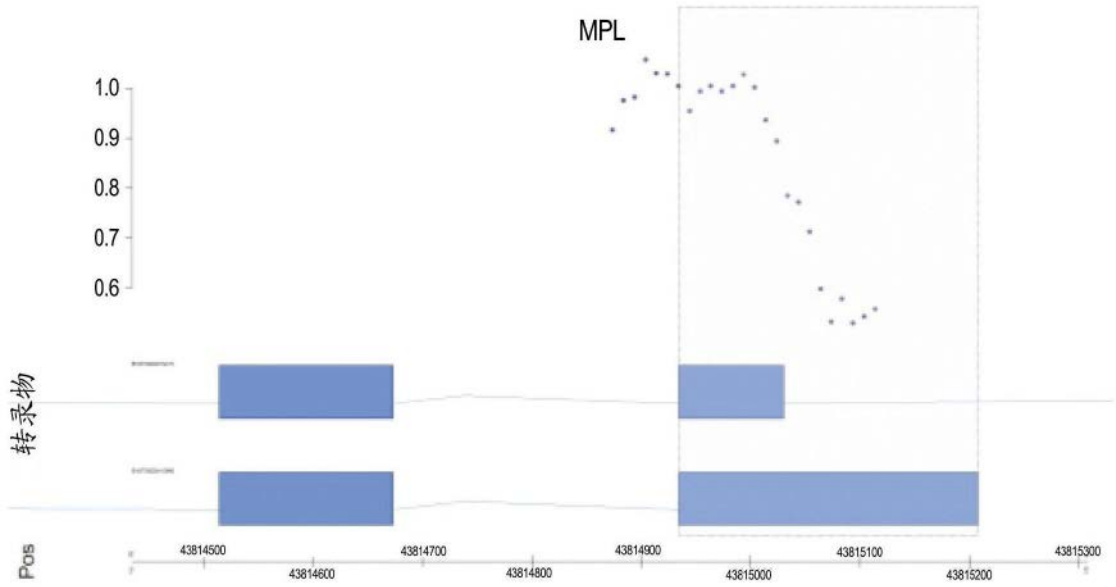

MPL
1.0
0.9
0.8
0.7
0.6
转录物
Pos
43814500
43814600
43814700
43814800
43814900
43815000
43815100
43815200
43815300

图32C