(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117174167 A
(43)申请公布日 2023.12.05

(21)申请号 202311131468.X
(22)申请日 2016.12.16
(30)优先权数据
62/269,051 2015.12.17 US
(62)分案原申请数据
201680081723.6 2016.12.16
(71)申请人 夸登特健康公司
地址 美国加利福尼亚州
(72)发明人 埃尔米·埃尔图凯
阿米尔阿里·塔拉萨兹
达里娅·丘多瓦
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 韦昌金 郑霞
(51)Int.Cl.
G16B 20/10(2019.01)
G16B 30/00(2019.01)
G16B 30/10(2019.01)
G16B 99/00(2019.01)
C12Q 1/6886(2018.01)

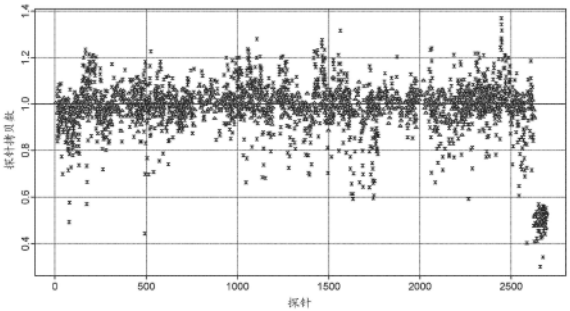
权利要求书3页 说明书22页 附图17页

(54)发明名称

通过分析无细胞DNA确定肿瘤基因拷贝数的方法

(57)摘要

本申请涉及通过分析无细胞DNA确定肿瘤基因拷贝数的方法。本文提供了方法以改进核酸样品中拷贝数变异的自动检测。这些方法提供了用于确定样品内遗传基因座的基线拷贝数的改进方法,减少了由于遗传基因座的特征、样品制备和探针耗尽导致的变异。



CN 117174167 A

1.一种系统,所述系统包括:

i)网络;

ii)数据库,所述数据库被连接到所述网络,包括配置成存储核酸序列数据的计算机存储器;

iii)生物信息学计算机,其包括计算机存储器和一个或更多个计算机处理器,所述计算机被连接到所述网络;

其中所述计算机还包括机器可执行代码,当由所述一个或更多个计算机处理器执行时,所述机器可执行代码复制存储在所述数据库中的所述核酸序列数据,将复制的数据写入所述生物信息学计算机中的存储器,并进行包括以下的步骤:

(a)从所述核酸序列数据生成第一数据集,其包含对于多个遗传基因座中的每个遗传基因座,(i)测序读段覆盖("读段覆盖")相关的定量量度,其中测序读段覆盖相关的定量量度包括所述多个遗传基因座中的每个遗传基因座处的独特分子计数(UMC);和(ii)所述遗传基因座的鸟嘌呤-胞嘧啶含量("GC含量")相关的定量量度,其中所述核酸序列数据从DNA样品获得;

(b)从所述第一数据集去除高变异遗传基因座,其中去除包括:

(i)拟合涉及鸟嘌呤-胞嘧啶含量相关的定量量度和所述遗传基因座的测序读段覆盖的定量量度的模型;和

(ii)从所述第一数据集去除至少10%的遗传基因座,其中去除遗传基因座包括去除与所述模型差异最大的遗传基因座的至少10%,由此提供基线定线遗传基因座的第一数据集;

(c)通过进行饱和平衡校正和探针效率校正来校正所述第一数据集,其中进行饱和平衡校正包括将基线定线遗传基因座的所述第一数据集转换为饱和校正数据集,并且其中进行探针效率校正包括转换所述饱和校正数据集为探针效率校正数据集;

(d)确定对于所述第一数据集的基线测序读段覆盖,其中所述基线测序读段覆盖包含预期测序读段覆盖并且是依赖于所述饱和平衡和所述探针效率的函数;并且

(e)确定所述多个遗传基因座中每个遗传基因座相对于所述基线读段覆盖的拷贝数状态。

2.根据权利要求1所述的系统,其中进行饱和平衡校正包括通过以下将基线定线遗传基因座的所述第一数据集转换为饱和校正数据集:

(a)对于每个遗传基因座,从基线定线遗传基因座的第一数据集确定来源于所述遗传基因座的来自样品的DNA分子链在测序读段内被代表的概率相关的定量量度;

(b)通过将基线定线遗传基因座的所述第一数据集中的读段覆盖与基线定线遗传基因座的所述第一数据集的GC含量和与基线定线遗传基因座的所述第一数据集中的来源于每个基因座的DNA链在所述测序读段内被代表的概率有关的定量量度相关联,确定所述读段覆盖的第一转换;和

(c)将所述第一转换应用于来自基线定线遗传基因座的第一数据集的每个遗传基因座的读段覆盖以提供所述饱和校正数据集,其中所述饱和校正数据集包括基线定线遗传基因座的第一数据集的转换的读段覆盖的第一集。

3.根据权利要求1所述的系统,其中进行探针效率校正包括通过以下转换所述饱和校

正数据集为探针效率校正数据集：

(a)从所述饱和校正数据集去除相对于转换的读段覆盖的所述第一集为高变异遗传基因座的遗传基因座,从而提供基线定线遗传基因座的第二数据集；

(b)确定与基线定线遗传基因座的所述第二数据集的探针效率相关的转换的读段覆盖的第一集的第二转换,其中所述探针效率通过对一个或更多个参考样品进行所述饱和平衡校正来确定,其中所述探针效率是通过进行所述饱和平衡校正获得的所述转换的读段覆盖;和

(c)利用所述第二转换将基线定线遗传基因座的所述第二数据集的转换的读段覆盖的所述第一集进行转换,从而提供探针效率校正数据集,其中所述探针效率校正数据集包括基线定线遗传基因座的所述第二数据集的转换的读段覆盖的第二集。

4.根据权利要求3所述的系统,其中从所述饱和校正数据集去除为高变异遗传基因座的遗传基因座包括：

(i)拟合涉及所述GC含量和所述饱和校正数据集的转换的读段覆盖的第一集的模型;和

(ii)从饱和校正数据集去除遗传基因座的子集,其中去除所述子集包括去除与所述模型差异最大的至少5%的遗传基因座,由此提供基线定线遗传基因座的第二数据集。

5.根据权利要求1所述的系统,其中确定来源于遗传基因座的DNA链在所述测序读段内被代表的概率相关的定量量度包括将测序读段分选为配对读段和非配对读段,其中(i)所述配对读段中每个配对读段对应于所述DNA样品中由来源于双链DNA分子的第一加标签链和第二差异加标签互补链生成的序列读段,并且(ii)所述非配对读段中每个非配对读段代表第一加标签链,来源于双链DNA分子的第二差异加标签互补链在序列读段中的所述序列读段之中没有被代表。

6.根据权利要求5所述的系统,其中映射到一个或更多个遗传基因座中的每一个的(i)所述配对读段和(ii)所述非配对读段的定量量度被确定,以基于映射到所述一个或更多个遗传基因座中的每个遗传基因座的配对读段和非配对读段的定量量度产生无细胞体液样品中映射到所述一个或更多个遗传基因座中的每一个的全部双链DNA分子相关的定量量度。

7.根据权利要求1所述的系统,其中所述一个或更多个遗传基因座的测序读段覆盖是与所述一个或更多个遗传基因座的区域的测序读段覆盖的中心趋势相关的量度。

8.根据权利要求1所述的系统,其中与测序读段覆盖相关的定量量度包括通过与文库大小相关的度量将UMC归一化以提供归一化UMC。

9.根据权利要求8所述的系统,其中将UMC归一化包括将遗传基因座的UMC除以所有UMC的总和,或者将遗传基因座的UMC除以所有常染色体UMC的总和。

10.根据权利要求8所述的系统,其中归一化UMC进一步归一化,并且进一步归一化包括:(i)从来源于训练样本的测序读段,确定相应的遗传基因座的归一化UMC;(ii)对于每个遗传基因座,样本的归一化UMC通过训练样本在相应基因座处的归一化UMC的中值进一步归一化,由此提供遗传基因座的相对丰度(RA)。

11.根据权利要求10所述的系统,其中所述一个或更多个遗传基因座的测序读段覆盖相关的定量量度包括所述探针的RA的中位数。

12.根据权利要求1所述的系统,其中所述测序读段覆盖的定量量度是对基因组的特定子区域确定的。

13.根据权利要求11所述的系统,其中所述区域能够是箱、感兴趣的基因、外显子、对应于序列探针的区域、对应于引物扩增产物的区域,或者对应于引物结合位点的区域。

14.根据权利要求11所述的系统,其中所述基因组的子区域是对应于寡核苷酸探针的区域。

15.根据权利要求1所述的系统,其中所述多个遗传基因座中的一个或更多个遗传基因座的GC含量相关的定量量度是与一种或更多种寡核苷酸探针的GC含量的中心趋势相关的量度。

16.根据权利要求15所述的系统,其中与GC含量的中心趋势相关的量度包括(i)平均值、(ii)中位数或(iii)众数。

17.根据权利要求1所述的系统,其中与GC含量相关的定量量度是通过将鸟苷和胞嘧啶碱基的数量除以区域上的碱基总数获得的量度。

18.根据权利要求1所述的系统,其中所述DNA样品是从受试者的组织获得的基因组DNA样品。

19.根据权利要求1所述的系统,其中所述DNA样品是从受试者的体液获得的无细胞DNA样品,其中所述体液选自由血清、血浆、尿和脑脊液组成的组。

20.根据权利要求1所述的系统,其中所述读段覆盖通过将所述测序读段映射于参考基因组来确定。

通过分析无细胞DNA确定肿瘤基因拷贝数的方法

[0001] 本申请是申请日为2016年12月16日,申请号为201680081723.6,发明名称为“通过分析无细胞DNA确定肿瘤基因拷贝数的方法”的申请的分案申请。

[0002] 交叉引用

[0003] 本申请要求于2015年12月17日提交的美国临时申请号62/269,051的优先权,其在此通过引用以其整体并入。

[0004] 背景

[0005] 癌症是由个体正常细胞内突变的积累引起的,其中至少一些导致细胞分裂调节不当。这种突变通常包括拷贝数变异,其中肿瘤基因组内基因的拷贝数相对于受试者的非癌性细胞增加或减少。

[0006] 利用检测和表征肿瘤细胞中的拷贝数变异来监测肿瘤的进展,预测患者的结果,和改善治疗选择。然而,常规方法是对细胞样品进行的,细胞样品通常通过痛苦和时间密集的活检获得。这种活检通常也只能检查受试者体内的一小部分肿瘤细胞,因此并不总是代表肿瘤细胞的群体。对于不需要细胞活检、荧光原位杂交(FISH)、比较基因组杂交阵列或定量荧光聚合酶链式反应(PCR)测定的用于肿瘤中拷贝数变异的更简单、更快速的检测存在需求。

[0007] 使用测序数据确定拷贝数变异的一个特殊挑战是,出于与真实拷贝数无关的原因,遗传基因座的覆盖深度会出现变异。例如,即使对于样品中存在的在同样拷贝数的单独遗传基因座,扩增效率、PCR效率和鸟嘌呤-胞嘧啶含量也可导致不同的覆盖深度。为了改善拷贝数检测,需要消除由于这种效应的偏倚的改进方法。

[0008] 概述

[0009] 对在来源于无细胞体液的样品的肿瘤细胞中检测拷贝数变异的改进方法存在相当大的需求。本方面解决了这种需求并且提供了另外的优点。在一方面,本公开内容提供了一种方法,包括:(a)获得受试者的无细胞体液样品的脱氧核糖核酸(DNA)分子的测序读段;(b)从序列读段生成第一数据集,对于多个遗传基因座中的每个遗传基因座,所述第一数据集包含与测序读段覆盖(“读段覆盖”)有关的定量量度;(c)通过进行饱和平衡校正和探针效率校正来校正第一数据集;(d)确定第一数据集的基线读段覆盖,其中基线读段覆盖涉及饱和平衡和探针效率;和(e)确定多个遗传基因座中每个遗传基因座相对于基线读段覆盖的拷贝数状态。在一些实施方案中,对于多个遗传基因座中的每个遗传基因座,第一数据集包含(i)遗传基因座的鸟嘌呤-胞嘧啶含量(“GC含量”)相关的定量量度。在一些实施方案中,该方法包括在(c)之前从第一数据集去除为高变异遗传基因座的遗传基因座,其中去除包括:(i)拟合涉及鸟嘌呤-胞嘧啶含量相关的定量量度和遗传基因座的测序读段覆盖的定量量度的模型;和(ii)从遗传基因座去除至少10%的遗传基因座,其中去除遗传基因座包括去除与该模型最不同的遗传基因座,由此提供基线定线(baselining)遗传基因座的第一数据集。在一些实施方案中,该方法包括去除至少45%的遗传基因座。

[0010] 在一些实施方案中,进行饱和平衡校正包括通过以下将基线定线数据遗传基因座的第一数据集转换为饱和校正数据集:(i)对于来自基线定线遗传基因座的第一数据集每

个遗传基因座,确定来源于遗传基因座的DNA分子链在测序读段内被代表的概率有关的定量量度;(ii)通过将基线定线遗传基因座的第一数据集的读段覆盖与基线定线遗传基因座的第一数据集的GC含量和与来源于基线定线遗传基因座的第一数据集中的每个基因座的DNA链在测序读段内被代表的概率有关的定量量度相关联来确定读段覆盖的第一转换;和(iii)将所述第一转换应用于来自基线定线遗传基因座的第一数据集的每个遗传基因座的读段覆盖以提供饱和校正数据集,其中所述饱和校正数据集包括基线定线遗传基因座的第一数据集的转换的读段覆盖的第一集。

[0011] 在一些实施方案中,确定第一转换包括(i)确定与基线定线遗传基因座的第一数据集的读段覆盖的中心趋势有关的量度;(ii)基于遗传基因座的GC含量和与来源于遗传基因座的DNA链被代表在测序读段内的概率相关的定量量度,确定拟合与基线定线遗传基因座的第一数据集的读段覆盖的中心趋势有关的量度的函数;和(iii)对于基线定线遗传基因座的第一数据集的每个遗传基因座,确定由所述函数预测的读段覆盖与所述读段覆盖的差异,其中所述差异为转换的读段覆盖。在一些实施方案中,该函数是表面近似。在本文提供的一些实施方案中,表面近似是二维二次多项式。

[0012] 在一些实施方案中,进行探针效率校正包括通过以下转换饱和校正数据集为探针效率校正数据集:(i)从饱和校正数据集去除相对于转换的读段覆盖的第一集为高变异遗传基因座的遗传基因座,从而提供基线定线遗传基因座的第二数据集;(ii)确定与基线定线遗传基因座的第二数据集的探针效率相关的转换的读段覆盖的第一集的第二转换;和(iii)利用第二转换将基线定线遗传基因座的第二数据集的转换的读段覆盖的第一数据集进行转换,从而提供探针效率校正数据集,其中探针效率校正数据集包括基线定线遗传基因座的第二数据集的转换的读段覆盖的第二集。在一些实施方案中,从第一数据集去除为高变异遗传基因座的遗传基因座包括:(i)拟合涉及GC含量和饱和校正数据集的转换的读段覆盖的第一集的模型;和(ii)从饱和校正数据集去除至少10%的遗传基因座,其中去除遗传基因座包括去除与该模型最不同的遗传基因座,由此提供基线定线遗传基因座的第二数据集。在本文提供的一些实施方案中,去除是至少45%的遗传基因座。

[0013] 在一些实施方案中,探针效率通过对一种或更多种参考样品进行饱和平衡校正来确定,其中探针效率是通过进行饱和平衡校正获得的转换的读段覆盖。在一些实施方案中,一种或更多种参考样品是来自无癌症的受试者的无细胞体液样品。在本文提供的一些实施方案中,一种或更多种参考样品是来自具有癌症的受试者的无细胞体液样品,其中对应的遗传基因座未经历拷贝数改变。

[0014] 在一些实施方案中,确定第二转换包括(i)将对来自一种或更多种参考样品的遗传基因座确定的探针效率拟合到来自基线定线遗传基因座的第二数据集的读段覆盖的第一集;(ii)将基线定线遗传基因座的第二数据集的每个遗传基因座的转换的读段覆盖除以基于(i)的拟合的预测的探针效率。在一些实施方案中,该方法还包括:(f)通过将基线定线遗传基因座的第二数据集的转换的读段覆盖与基线定线遗传基因座的第二数据集的GC含量和与来源于基线定线遗传基因座的第二数据集中的每个基因座的DNA链在测序读段内被代表的概率有关的定量量度相关联,确定转换的读段覆盖的第二集的第三转换;和(g)将第三转换应用于转换的读段覆盖的第二集以提供第四数据集,其中第四数据集包括转换的定量读段覆盖的第三集。

[0015] 在一些实施方案中,无细胞体液样品的DNA使用与来自遗传基因座的集的遗传基因座的至少一部分互补的一种或更多种寡核苷酸探针富集遗传基因座的集。在一些实施方案中,来自遗传基因座的集的每个遗传基因座的GC含量是与来自遗传基因座的集的遗传基因座的至少一部分互补的一种或更多种寡核苷酸探针的鸟嘌呤-胞嘧啶含量的中心趋势相关的量度。在一些实施方案中,遗传基因座的读段覆盖是与对应于一种或更多种寡核苷酸探针的遗传基因座区域的读段覆盖的中心趋势相关的量度。在一些实施方案中,进行饱和平衡校正和进行探针效率校正包括拟合Langmuir模型,其中Langmuir模型包含探针效率(K)和饱和平衡常数(Isat)。在一些实施方案中,K和Isat对于一种或更多种寡核苷酸探针中的每种寡核苷酸探针根据经验确定。在一些实施方案中,进行饱和平衡校正和进行探针校正包括将遗传基因座的读段覆盖拟合于Langmuir模型,假设遗传基因座以相同的拷贝数状态存在,从而提供基线读段覆盖。在一些实施方案中,相同的拷贝数状态是二倍体。在一些实施方案中,基线读段覆盖是依赖于探针效率和饱和平衡的函数。

[0016] 在一些实施方案中,确定拷贝数状态包括比较遗传基因座的读段覆盖与基线读段覆盖。在一些实施方案中,无细胞体液选自由以下组成的组:血清、血浆、尿和脑脊液。在一些实施方案中,读段覆盖通过将测序读段映射到参考基因组来确定。在一些实施方案中,获得测序读段包括将衔接子与来自受试者的无细胞体液的DNA分子连接。在一些实施方案中,DNA分子是双链DNA分子并且衔接子与双链DNA分子连接,使得每个衔接子将DNA分子的互补链不同地加标签以提供加标签链。在一些实施方案中,确定来源于遗传基因座的DNA链在所述测序读段内被代表的概率相关的定量量度包括将测序读段分选为配对读段和非配对读段,其中(i)每个配对读段对应于所述集中由来源于双链多核苷酸分子的第一加标签链和第二差异加标签互补链生成的序列读段,并且(ii)每个非配对读段代表第一加标签链,并且来源于双链多核苷酸分子的第二差异加标签互补链在所述序列读段集中的所述序列读段之中没有被代表。在一些实施方案中,该方法还包括确定映射到一个或更多个遗传基因座中的每一个的(i)所述配对读段和(ii)所述非配对读段的定量量度,以基于映射到每个基因座的配对读段和非配对读段相关的所述定量量度确定所述样品中映射到所述一个或更多个遗传基因座中的每一个的全部双链DNA分子相关的定量量度。在一些实施方案中,衔接子包括条形码序列。

[0017] 在一些实施方案中,确定读段覆盖包括基于测序读段映射到参考基因组的位置和条形码序列叠并测序读段。在一些实施方案中,遗传基因座包括一种或更多种致癌基因。在一些实施方案中,一种方法包括通过确定受试者的种系基因组对其为杂合的基线定线遗传基因座内的变体的相对量,确定受试者的肿瘤细胞中基线定线遗传基因座的至少子集已经经历拷贝数改变。在一些实施方案中,变体的相对量不是近似相等的。在一些实施方案中,将变体的相对量对其不是近似相等的基线定线遗传基因座从基线定线遗传基因座去除,从而提供等位基因频率校正的基线定线遗传基因座。在一些实施方案中,等位基因频率校正的基线定线遗传基因座在本文描述的方法中用作基线定线基因座。

[0018] 在另一方面,本公开内容提供了包括以下的方法:接收受试者的无细胞体液样品的脱氧核糖核酸(DNA)分子的测序读段到存储器中;用计算机处理器执行代码以进行以下步骤:从序列读段生成对于多个遗传基因座中的每个遗传基因座包含与测序读段覆盖(“读段覆盖”)有关的定量量度的第一数据集;通过进行饱和平衡校正和探针效率校正来校正第

一数据集;确定第一数据集的基线读段覆盖,其中基线读段覆盖涉及饱和平衡和探针效率;和确定多个遗传基因座中每个遗传基因座相对于基线读段覆盖的拷贝数状态。

[0019] 在另一方面,本公开内容提供了一种系统,包括:网络;数据库,所述数据库连接到网络,包括配置成存储核酸(例如,DNA)序列数据的计算机存储器;生物信息学计算机,其包括计算机存储器和一个或更多个计算机处理器,所述计算机连接到所述网络;其中计算机还包括机器可执行代码,当由一个或更多个计算机处理器执行时,该机器可执行代码复制存储在数据库上的核酸(例如,DNA)序列数据,将复制的数据写入生物信息学计算机中的存储器,并进行包括以下的步骤:从核酸(例如,DNA)序列数据生成第一数据集,其包含对于多个遗传基因座中的每个遗传基因座,与测序读段覆盖(“读段覆盖”)相关的定量量度;通过进行饱和平衡校正和探针效率校正来校正第一数据集;确定对于第一数据集的基线读段覆盖,其中基线读段覆盖与饱和平衡和探针效率有关;并确定多个遗传基因座中每个遗传基因座相对于基线读段覆盖的拷贝数状态。在一些实施方案中,数据库连接于DNA测序仪。

[0020] 通过引用并入

[0021] 本说明书中提及的所有出版物、专利和专利申请通过引用并入本文,其程度如同每一个单独的出版物、专利或专利申请被具体和单独地指明通过引用并入的相同程度。

[0022] 附图简述

[0023] 本发明的新颖的特征在所附权利要求书中具体陈述。本发明的特征和优势的更好的理解将通过参考以下详细说明和附图来获得,详细说明陈述了其中利用本发明的原理的说明性实施方案,在附图中:

[0024] 图1说明示例性致癌基因和用于序列捕获探针的靶。

[0025] 图2说明基因水平信号相对于跨三个加标(spike-in)的理论拷贝数和跨加标基因的探针水平信号变异。

[0026] 图3说明将诱饵量与独特分子计数相关联的诱饵优化实验。

[0027] 图4A和图4B说明p(图4A)和GC含量(图4B)对独特分子计数的非线性影响。

[0028] 图5说明没有进行饱和或探针效率校正的每个探针的独特分子计数。

[0029] 图6说明饱和校正后每个探针的独特分子计数。

[0030] 图7说明饱和校正后和探针效率校正后每个探针的独特分子计数。

[0031] 图8说明真实拷贝数与探针饱和和探针效率相关的独特分子计数之间的相互作用的提议的Langmuir模型。

[0032] 图9说明典型临床样品中饱和校正、探针效率校正和第二轮探针效率校正后用于基线定线遗传基因座的探针信号降噪。

[0033] 图10A和图10B说明将饱和后校正的UMC针对参考样品中确定的探针效率作图,以便进行探针效率校正。图10A是来自肿瘤细胞中无拷贝数改变的受试者。图10B是来自肿瘤细胞中具有拷贝数改变的受试者。

[0034] 图11说明患者样品中饱和和探针效率校正的拷贝数变异检测的最终报告。样品上方的星号表示基于校正的信号和次等位基因频率校正的基线优化检测的基因扩增。

[0035] 图12示出了被编程或以其他方式配置为实现本公开内容的方法的计算机系统1201。

[0036] 图13说明使用本公开内容的方法测量的基因ERBB2的观察到的拷贝数(CN)vs.理

论CN。实心圆点表示观察到的~2拷贝数(二倍体样品),空心圆点表示检测到的扩增事件,且粗的水平虚线表示平均基因CN截断值。

[0037] 图14说明与对照方法(方形)相比使用本公开内容的方法(圆点)测量的基因ERBB2的观察到的拷贝数(CN)vs.理论CN。实心圆点表示观察到的~2拷贝数(二倍体样品),空心圆点表示检测到的扩增事件,且粗的水平虚线表示平均基因CN截断值。

[0038] 图15说明将探针拷贝数针对用于本公开内容的方法(三角形)vs.对照方法(X)的验证研究中使用的探针作图。

[0039] 详细说明

[0040] 定义

[0041] 如本文所用的术语"遗传变异"通常是指受试者的核酸样品或基因组中的改变、变异或多态性。这样的改变、变异或多态性可以是相对于参考基因组的,该参考基因组可以是所述受试者或其他个体的参考基因组。单核苷酸多态性(SNP)是多态性的一种形式。在一些实例中,一个或更多个多态性包括一个或更多个单核苷酸变异(SNV)、插入、缺失、重复、小的插入、小缺失、小重复、结构变体接合、可变长度串联重复和/或侧翼序列。拷贝数变异(CNV)、颠换和其他重排也是遗传变异的形式。基因组改变可以是碱基变化、插入、缺失、重复、拷贝数变异或颠换。

[0042] 如本文使用的,术语"多核苷酸"通常指包含一个或更多个核酸亚单位的分子。多核苷酸可包含选自腺苷(A)、胞嘧啶(C)、鸟嘌呤(G)、胸腺嘧啶(T)和尿嘧啶(U)或其变体的一个或更多个亚单位。核苷酸可包含A、C、G、T或U或其变体。核苷酸可包括可被掺入到增长的核酸链中的任何亚单位。此类亚单位可以是A、C、G、T或U,或为一个或更多个互补A、C、G、T或U特异性的或与嘌呤(即,A或G或其变体)或嘧啶(即,C、T或U或其变体)互补的任何其他亚单位。亚单位能够使得个体核酸碱基或碱基组(例如,AA、TA、AT、GC、CG、CT、TC、GT、TG、AC、CA或尿嘧啶-其对应物)得以解析。在一些实例中,多核苷酸为脱氧核糖核酸(DNA)或核糖核酸(RNA)或其衍生物。多核苷酸可以是单链或双链的。

[0043] 如本文所用的术语"受试者"通常是指动物如哺乳动物物种(例如,人),或禽类(例如,鸟)物种,或其他生物体如植物。更具体地,所述受试者可以是脊椎动物、哺乳动物、小鼠、灵长类动物、猿或人。动物包括,但不限于,农场动物、运动动物和宠物。受试者可以是健康的个体,患有或疑似患有疾病或有患病倾向的个体,或需要治疗或疑似需要治疗的个体。受试者可以是患者。

[0044] 术语"基因组"通常指生物体的遗传信息的整体。基因组可以以DNA或以RNA来编码。基因组可包含编码蛋白的编码区域和非编码区域。基因组可包含生物体中所有染色体的序列的集合。例如,人类基因组具有总计46条染色体。所有这些的序列共同构成人类基因组。

[0045] 术语"衔接子(adaptor)"、"衔接子(adaptor)"和"标签"在本说明书全文中作为同义词使用。可通过任何方法,包括连接、杂交或其他方法,使衔接子或标签与待"加标签"的多核苷酸序列偶联。

[0046] 如本文所用的术语"文库衔接子(library adaptor)"或"文库衔接子(library adaptor)"通常是指其身份(例如,序列)可用来区分生物样品(在本文中也称为"样品")中的多核苷酸的分子(例如,多核苷酸)。

[0047] 如本文所用的术语“测序衔接子”通常是指适合于允许测序仪器对靶多核苷酸进行测序(例如通过与靶多核苷酸相互作用以实现测序)的分子(例如,多核苷酸)。测序衔接子允许测序仪器对靶多核苷酸进行测序。在一个实例中,测序衔接子包含与附接至测序系统的固体支持物如流通池的捕获多核苷酸杂交或结合的核苷酸序列。在另一个实例中,测序衔接子包含与多核苷酸杂交或结合以生成发夹环的核苷酸序列,该发夹环允许测序系统对靶多核苷酸进行测序。测序衔接子可包含测序仪基序,该测序仪基序可以是与其他分子(例如,多核苷酸)的流通池序列互补并可被测序系统使用以对靶多核苷酸进行测序的核苷酸序列。该测序仪基序还可包括用于在测序如合成测序中使用的引物序列。该测序仪基序可包括使文库衔接子与测序系统偶联和对靶多核苷酸进行测序所需的序列。

[0048] 除非另外确定,当如本文所用的术语“至少”、“至多”或“约”在系列之前时是指所述系列的每一个成员。

[0049] 与参考数值相关的术语“约”及其语法上的等同表述可包括从该值加上或减去10%的值的范围。例如,“约10”的量可包括9至11的量。在其他实施方案中,与参考数值相关的术语“约”可包括从该值加上或减去10%、9%、8%、7%、6%、5%、4%、3%、2%或1%的值的范围。

[0050] 与参考数值相关的术语“至少”及其语法上的等同表述可包括该参考数值和大于该值的值。例如,“至少10”的量可包括值10和大于10的任何数值,如11、100和1,000。

[0051] 与参考数值相关的术语“至多”及其语法上的等同表述可包括该参考数值和小于该值的值。例如,“至多10”的量可包括值10和10以下的任何数值,如9、8、5、1、0.5和0.1。

[0052] 术语“定量量度”是指数量的任何量度,包括绝对和相对量度。定量量度可以是,例如,数目(例如计数)、百分比、程度或阈值。

[0053] 术语“读段覆盖”是指被原始序列读段或由处理序列读段的覆盖,例如从原始序列读段推断的独特分子计数。

[0054] 术语“基线读段覆盖”是指基于给定探针参数例如GC含量、探针效率、连接效率或下拉效率,包含二倍体基因组环境的样品中探针的预期读段覆盖。

[0055] 本文所用的“探针”是指包含官能性的多核苷酸。官能性可以是可检测标记(荧光)、结合部分(生物素)、或固体支持物(磁力可吸引的颗粒或芯片)。

[0056] “互补性”是指核酸与另一核酸序列通过传统Watson-Crick类型或其他非传统类型形成氢键的能力。互补性百分比指示核酸分子中可与第二核酸序列形成氢键(Watson-Crick碱基配对)的残基的百分比(10个中的5、6、7、8、9、10个分别为50%、60%、70%、80%、90%和100%互补)。“完全互补”是指核酸序列中所有的连续残基将与第二核酸序列中相同数目的连续残基氢键键合。

[0057] 本文所用的“大体上互补”是指跨越8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、30、35、40、45、50或更多个核苷酸的区域至少60%、65%、70%、75%、80%、85%、90%、95%、97%、98%、99%或100%的互补性程度,或是指在严格条件下杂交的两个核酸。序列同一性,诸如为了评估互补性百分比的目的,可通过任何合适的比对算法测量,所述比对算法包括但不限于Needleman-Wunsch算法(参见例如在万维网网页:ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html可得的EMBOSS Needle aligner,任选地以默认设置)、BLAST算法(参见例如在blast.ncbi.nlm.nih.gov/Blast.cgi可得的

BLAST比对工具,任选地以默认设置)、或Smith-Waterman算法(参见例如在万维网网页:ebi.ac.uk/Tools/psa/emboss water/nucleotide.html可得的EMBOSS Water aligner,任选地以默认设置)。最佳比对可使用选择的算法的任何合适的参数(包括默认参数)评估。

[0058] “杂交”是指其中一个或更多个多核苷酸反应形成经由核苷酸残基的碱基之间的氢键被稳定化的复合体的反应。氢键键合可通过Watson Crick碱基配对、Hoogstein结合、或以符合碱基互补性的任何其他序列特异性方式发生。复合体可包括形成双链体结构的两条链、形成多链复合体的三条或更多条链、单个自杂交链、或这些的任何组合。杂交反应可构成更广泛的过程诸如PCR的起始、或内切核酸酶对多核苷酸的酶促裂解中的一个步骤。与第一序列互补的第二序列被称为第一序列的“互补物”。应用于多核苷酸的术语“可杂交的”是指杂交反应中多核苷酸形成经由核苷酸残基的碱基之间的氢键键合被稳定化的复合体的能力。

[0059] 术语“严格杂交条件”指在其下多核苷酸将优先与其靶子序列杂交,并且在较小程度上与其它序列杂交或根本不与其他序列杂交的条件。在核酸杂交实验的上下文中的“严格杂交”是序列依赖性的,并在不同环境参数下是不同的。核酸杂交的广泛指南见于Tijssen(1993)Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes第I部分第2章“Overview of principles of hybridization and the strategy of nucleic acid probe assays”,Elsevier,New York。

[0060] 通常,高度严格的杂交和洗涤条件被选择为比在定义的离子强度和pH下特定序列的热解链温度(Tm)低约5℃。Tm为50%的靶序列与完全匹配的探针杂交的温度(在定义的离子强度和pH下)。非常严格的条件被选择为等于特定探针的Tm。

[0061] 严格杂交条件包括包含水、缓冲剂(在pH 6-9或pH 7-8的磷酸盐、tris、SSPE或SSC缓冲液)、盐(钠或钾)、和变性剂(SDS、甲酰胺或吐温)的缓冲液和37℃-70℃、60℃-65℃的温度。

[0062] 在DNA印迹或RNA印迹中的滤器上用于具有多于100个互补残基的互补核酸的杂交的严格杂交条件的实例为在42℃,50%福尔马林与1mg肝素,杂交被进行过夜。高度严格的洗涤条件的实例为在72℃,0.15M NaCl,持续约15分钟。严格洗涤条件的实例为在65℃,0.2X SSC洗涤,持续15分钟(参见,Sambrook等人对SSC缓冲液的描述)。通常,高严格洗涤之前进行低严格洗涤以去除背景探针信号。对于多于100个核苷酸的双链体的示例性中度严格洗涤为在45℃,1x SSC,持续15分钟。对于,例如,多于100个核苷酸的双链体的示例性低度严格洗涤为在40℃,4-6x SSC,持续15分钟。通常,比起在特定杂交测定中对于不相关探针观察到的2倍(或更高)的信噪比指示检测到特异性杂交。

[0063] 在一方面,本公开内容提供了一种方法,包括:(a)获得来源于受试者的无细胞体液样品的脱氧核糖核酸(DNA)分子的测序读段;(b)从测序读段生成第一数据集,所述第一数据集包含对于多个遗传基因座中的每个遗传基因座,与以下相关的定量量度:(i)遗传基因座的鸟嘌呤-胞嘧啶含量和(ii)遗传基因座的测序读段覆盖相关的定量量度;(c)通过以下转换第一数据集为第二数据集:(i)从所述第一数据集去除相对于测序读段覆盖相关的定量量度为高变异遗传基因座的遗传基因座;(ii)对于来自剩余遗传基因座的第一集的每个遗传基因座,确定来源于遗传基因座的样品的DNA链在测序读段内被代表的概率有关的

定量量度;(iii)通过将剩余遗传基因座的第一集的测序读段覆盖相关的定量量度与剩余遗传基因座的第一集的GC含量相关的定量量度和与来源于剩余遗传基因座的第一集中的每个基因座的DNA链在测序读段内被代表的概率有关的定量量度相关联来确定测序读段覆盖相关的定量量度的第一转换;和(iv)将所述第一转换应用于来自剩余遗传基因座的第一集的每个遗传基因座的序列读段覆盖以提供第二数据集,其中第二数据集包括剩余遗传基因座的第一集的测序读段覆盖的转换的定量量度的第一集。

[0064] 在一些实施方案中,该方法还包括通过以下将第二数据集转换为第三数据集:(d)从第二数据集去除相对于测序读段覆盖的转换的定量量度的第一集为高变异遗传基因座的遗传基因座,从而提供剩余遗传基因座的第二集;(e)确定与剩余遗传基因座的第二集的效率相关的测序读段覆盖的转换的定量量度的第一集的第二转换和(f)利用第二转换将剩余遗传基因座的第二集的测序读段覆盖的转换的定量量度的第一集进行转换,从而提供第三数据集,其中第三数据集包括(d,i)的剩余遗传基因座的第二集的测序读段覆盖相关的转换的定量量度的第二集;

[0065] 获得来自受试者的无细胞体液的DNA分子的测序读段

[0066] 获得来自受试者的无细胞体液的DNA分子的测序读段可包括获得无细胞体液。示例性的无细胞体液是或可来源于血清、血浆、血液、唾液、尿、滑液、全血、淋巴液、腹水、间质液或细胞外液、细胞之间的间隙中的流体(包括龈沟液)、骨髓、脑脊液、唾液、粘液、痰液、精液、汗液、尿液或任何其他体液。无细胞体液可选自由以下组成的组:血浆、尿或脑脊液。无细胞体液可以是血浆。无细胞体液可以是尿。无细胞体液可以是脑脊液。

[0067] 核酸分子,包括DNA分子,可从无细胞体液提取。DNA分子可以是基因组DNA。DNA分子可来自受试者的健康组织的细胞。DNA分子可来自已经历体细胞突变的非癌性细胞。DNA分子可来自母体样品中的胎儿。本领域技术人员将理解,在其中DNA分子来自母体样品中的胎儿的实施方案中,受试者可以是指胎儿,尽管样品是母体样品。DNA分子可来自受试者的癌变前细胞。DNA分子可来自受试者的癌性细胞。DNA分子可来自受试者的原代肿瘤内的细胞。DNA分子可来自受试者的继发肿瘤。DNA分子可以是循环DNA。循环DNA可包括循环肿瘤DNA(ctDNA)。DNA分子可以是双链或单链的。可选地,DNA分子可包括双链部分和单链部分的组合。DNA分子不必是无细胞的。在一些情形中,DNA分子可从样品分离。例如,DNA分子可以是从体液例如血清或血浆分离的无细胞DNA。

[0068] 样品可包含不同量的核酸分子的基因组等同物。例如,约30ng DNA的样品可含有约10,000个单倍体人类基因组等同物,而在cfDNA的情况下,可含有约2000亿个单独的多核苷酸分子。类似地,约100ng DNA的样品可含有约30,000个单倍体人类基因组等同物,而在cfDNA的情况下,可含有约6000亿个单独的分子。

[0069] 无细胞DNA分子可使用本领域已知的多种技术从体液分离和提取。在一些情况下,可使用商购可得的试剂盒诸如Qiagen Qiamp® Circulating Nucleic Acid试剂盒方案分离、提取和制备无细胞核酸。在其他实例中,可使用Qiagen Qubit™ dsDNA HS测定试剂盒方案、Agilent™ DNA 1000试剂盒或TruSeq™ Sequencing Library Preparation;Low-Throughput(LT)方案来定量核酸。无细胞核酸可以是胎儿来源的(经由从妊娠受试者取得的流体),或可源自受试者自身的组织。无细胞核酸可来源于赘生物(例如肿瘤或腺瘤)。

[0070] 通常,通过分配步骤(partitioning step)从体液中提取和分离无细胞核酸,在该

分配步骤中,如在溶液中发现的无细胞核酸与细胞和体液的其他不可溶性组分分开。分配可包括但不限于技术诸如离心或过滤。在其他情况下,细胞并非首先与无细胞核酸分配,而是经裂解。在一个实例中,完整细胞的基因组DNA通过选择性沉淀来分配。无细胞核酸(包括DNA)可保持可溶性并可与不可溶性基因组DNA分开并被提取。通常,在不同试剂盒特定的添加缓冲液和其他洗涤步骤之后,可使用异丙醇沉淀来沉淀核酸。可使用进一步的清洁步骤诸如基于二氧化硅的柱以去除污染物或盐。可针对特定应用优化一般步骤。例如,可在整个反应中添加非特异性批量(bulk)载体核酸以优化该程序的某些方面诸如收率。

[0071] 无细胞DNA分子可以是至多500个核苷酸的长度、至多400个核苷酸的长度、至多300个核苷酸的长度、至多250个核苷酸的长度、至多225个核苷酸的长度、至多200个核苷酸的长度、至多190个核苷酸的长度、至多180个核苷酸的长度、至多170个核苷酸的长度、至多160个核苷酸的长度、至多150个核苷酸的长度、至多140个核苷酸的长度、至多130个核苷酸的长度、至多120个核苷酸的长度、至多110个核苷酸的长度、或至多100个核苷酸的长度。

[0072] 无细胞DNA分子可以是至少500个核苷酸的长度、至少400个核苷酸的长度、至少300个核苷酸的长度、至少250个核苷酸的长度、至少225个核苷酸的长度、至少200个核苷酸的长度、至少190个核苷酸的长度、至少180个核苷酸的长度、至少170个核苷酸的长度、至少160个核苷酸的长度、至少150个核苷酸的长度、至少140个核苷酸的长度、至少130个核苷酸的长度、至少120个核苷酸的长度、至少110个核苷酸的长度、或至少100个核苷酸的长度。尤其地,无细胞核酸可以是在140和180个核苷酸之间的长度。

[0073] 无细胞DNA可包含来自健康组织和肿瘤的不同量的DNA分子。肿瘤来源的无细胞DNA可以是样品中无细胞DNA总量的至少0.1%、样品中无细胞DNA总量的至少0.2%、样品中无细胞DNA总量的至少0.5%、样品中无细胞DNA总量的至少0.7%、样品中无细胞DNA总量的至少1%、样品中无细胞DNA总量的至少2%、样品中无细胞DNA总量的至少3%、样品中无细胞DNA总量的至少4%、样品中无细胞DNA总量的至少5%、样品中无细胞DNA总量的至少10%、样品中无细胞DNA总量的至少15%、样品中无细胞DNA总量的至少20%、样品中无细胞DNA总量的至少25%、或样品中无细胞DNA总量的至少30%、或更多。

[0074] 在一些情况中,DNA分子可在提取过程中被修剪并包括在100和400个核苷酸之间的长度的片段。在一些情况中,核酸可在提取后被修剪并可包括在100和400个核苷酸之间的长度的核苷酸。在某些情况下,DNA分子的长度已经在100到400个核苷酸之间,且另外的修剪并非有目的地实施。

[0075] 受试者可以是动物。受试者可以是哺乳动物,诸如犬、马、猫、小鼠、大鼠或人类。受试者可以是人类。受试者可疑似患有癌症。受试者可此前接受过癌症诊断。受试者的癌症状态可以是未知的。受试者可以是雄性或雌性的。受试者可以是至少20岁、至少30岁、至少40岁、至少50岁、至少60岁或至少70岁。

[0076] 测序可通过本领域已知的任何方法。例如,测序技术包括经典技术(例如,使用标记的终止子或引物的双脱氧测序反应(Sanger方法)以及平板或毛细管中的凝胶分离)和下一代技术。示例性技术包括使用可逆终止的标记的核苷酸的合成测序、焦磷酸测序、454测序、Illumina/Solexa测序、与标记的寡核苷酸探针文库的等位基因特异性杂交、使用等位基因特异性杂交至标记的克隆文库,随后连接的合成测序、在聚合步骤期间实时监测标记的核苷酸的掺入、聚合酶克隆测序(polony sequencing)、SOLiD测序靶向测序、单分子实时

测序、外显子测序、基于电子显微术的测序、组测序(panel sequencing)、晶体管介导的测序、直接测序、随机鸟枪法测序、全基因组测序、杂交测序、毛细管电泳、凝胶电泳、双链体测序、循环测序、单碱基延伸测序、固相测序、高通量测序、大规模平行标志测序、乳液PCR、在较低变性温度共扩增-PCR(COLD-PCR)、多重PCR、可逆染料终止子测序、双末端测序(paired-end sequencing)、近期测序(near-term sequencing)、核酸外切酶测序、连接测序、短读段测序、单分子测序、实时测序、反向终止子测序、纳米孔测序、MS-PET测序及其组合。在一些实施方案中,测序方法是大规模平行测序,即,对至少100、1000、10,000、100,000、100万、1000万、1亿、或10亿个多核苷酸分子的任一种同时(或快速连续)测序。在一些实施方案中,测序可由基因分析仪,诸如例如可从Illumina或Applied Biosystems商业获得的基因分析仪进行。最近通过利用聚合酶或连接酶的连续或单一延伸反应以及通过与探针文库的单一或连续差示杂交已经证明了单独分子的测序。可由DNA测序仪(例如,设计用于进行测序反应的机器)进行测序。在一些实施方案中,DNA测序仪可包含数据库或连接到数据库,例如,包含DNA序列数据的数据库。

[0077] 可使用的测序技术包括,例如,使用合成测序系统。在第一步骤中,将DNA剪切成约300-800个碱基对的片段,并将片段平端化。寡核苷酸衔接子然后连接到片段的末端。衔接子充当用于片段扩增和测序的引物。可使用例如含有5'-生物素标签的衔接子B将片段附接至DNA捕获珠,例如链霉抗生物素蛋白包被的珠。附着于珠的片段在油-水乳液的液滴内PCR扩增。结果是每个珠上克隆扩增的DNA片段的多个拷贝。在第二步骤中,珠被捕获在孔中(皮升大小)。焦磷酸测序对每个DNA片段平行进行。加入一个或更多个核苷酸生成光信号,其由测序仪器中的CCD相机记录。信号强度与掺入的核苷酸数量成正比。焦磷酸测序利用核苷酸添加时释放的焦磷酸(PPi)。在腺苷5'磷酸硫酸盐存在下,通过ATP硫酸化酶将PPi转化为ATP。萤光素酶使用ATP将萤光素转化为氧化萤光素,并且该反应生成被检测和分析的光。

[0078] 可使用的DNA测序技术的另一个实例是来自Life Technologies Corporation(Carlsbad,Calif.)的Applied Biosystems的SOLiD技术。在SOLiD测序中,将基因组DNA剪切成片段,并将衔接子附接到片段的5'和3'末端以生成文库片段。可选地,可通过以下引入内部衔接子:连接衔接子到片段的5'和3'末端,环化片段,消化环化片段以生成内部衔接子,并将衔接子附接到所得片段的5'和3'末端以生成配偶-配对文库。接下来,在含有珠、引物、模板和PCR组分的微反应器中制备克隆珠群体。PCR后,模板被变性并富集珠以分离具有延伸的模板的珠。在选定的珠上的模板经历允许与载玻片键合的3'修饰。序列可通过连续杂交和连接具有由特定荧光团识别的中心确定的碱基(或碱基对)的部分随机的寡核苷酸来确定。记录颜色后,去除连接的寡核苷酸,且然后重复该过程。

[0079] 可使用的DNA测序技术的另一个实例是离子半导体测序,例如使用Life Technologies(South San Francisco,Calif.)的Ion Torrent以商标ION TORRENT出售的系统。离子半导体测序描述于,例如Rothberg,等人,An integrated semiconductor device enabling non-optical genome sequencing,Nature 475:348-352(2011);美国公布号2010/0304982;美国公布号2010/0301398;美国公布号2010/0300895;美国公布号2010/0300559;和美国公布号2009/0026082中,其每一个的内容通过引用以其整体并入。

[0080] 可使用的测序技术的另一个实例是Illumina测序。Illumina测序是基于使用折叠PCR(fold-back PCR)和锚定引物在固体表面上扩增DNA。基因组DNA被片段化,且衔接子被

添加到片段的5'和3'末端。附着于流通池通道表面的DNA片段被延伸并被桥式扩增。片段变成双链,且双链分子变性。多个循环的固相扩增、随后变性可在流通池的每个通道中产生相同模板的单链DNA分子的约1,000个拷贝的几百万个簇。引物、DNA聚合酶和四种荧光团标记的、可逆终止的核苷酸用于进行顺序测序。核苷酸掺入后,使用激光激发荧光团,拍摄图像并记录第一碱基的身份。来自每个掺入的碱基的3'终止子和荧光团被除去,并重复掺入、检测和鉴定步骤。按照这种技术的测序描述在美国专利第7,960,120号;美国专利第7,835,871号;美国专利第7,232,656号;美国专利第7,598,035号;美国专利第6,911,345号;美国专利第6,833,246号;美国专利第6,828,100号;美国专利第6,306,597号;美国专利第6,210,891号;美国公布号2011/0009278;美国公布号2007/0114362;美国公布号2006/0292611;和美国公布号2006/0024681,其每一个通过引用以其整体并入。

[0081] 可使用的测序技术的另一个实例包括Pacific Biosciences(Menlo Park,Calif.)的单分子、实时(SMRT)技术。在SMRT中,四种DNA碱基中的每一种与四种不同荧光染料中的一种相连。这些染料是磷酸连接的(phospholinked)。单个DNA聚合酶与模板单链DNA的单个分子一起固定在零模波导(ZMW)的底部。将一个核苷酸掺入增长链需要几毫秒。在此期间,荧光标记物被激发并产生荧光信号,且荧光标签被切下。检测染料的相应荧光表明哪个碱基被掺入。重复该过程。

[0082] 可使用的测序技术的另一个实例是纳米孔测序(Soni&Meller,2007,Progress toward ultrafast DNA sequence using solid-state nanopores,Clin Chem 53(11):1996-2001)。纳米孔是直径为1纳米级别的小孔。将纳米孔浸入导电流体中并跨其施加电势导致由于离子通过纳米孔的传导而产生轻微的电流。流动的电流量对纳米孔的尺寸敏感。当DNA分子通过纳米孔时,DNA分子上的每个核苷酸以不同程度阻挡纳米孔。因此,当DNA分子通过纳米孔时,通过纳米孔的电流的变化代表DNA序列的读取。

[0083] 可使用的测序技术的另一个实例涉及使用化学品敏感的场效应晶体管(chemFET)阵列来对DNA测序(例如,如美国公布号2009/0026082中描述的)。在该技术的一个实例中,可将DNA分子置于反应室中,并且模板分子可与结合聚合酶的测序引物杂交。在测序引物的3'末端将一个或更多个三磷酸掺入到新核酸链中可通过chemFET的电流变化来检测。阵列可有多个chemFET传感器。在另一个实例中,可将单个核酸与珠附接,并且可在珠上扩增核酸,并且可将单个珠转移到chemFET阵列上的单个反应室中,每个室具有chemFET传感器,并且核酸可被测序。

[0084] 可使用的测序技术的另一个实例涉及使用电子显微镜,例如,由Moudrianakis,E.N.和Beer M.,在Base sequence determination in nucleic acids with the electron microscope,III.Chemistry and microscopy of guanine-labeled DNA,PNAS 53:564-71(1965)中所描述的。在该技术的一个实例中,使用可使用电子显微镜区分的金属标记物来标记单个DNA分子。然后将这些分子在平坦的表面上展开并使用电子显微镜成像以测量序列。

[0085] 在测序之前,衔接子序列可附接到核酸分子,并且可以为感兴趣的特定序列富集核酸。序列富集可在衔接子序列附接之前或之后发生。

[0086] 核酸分子或富集的核酸分子可附接到适用于本文公开的任何测序平台的任何测序衔接子。例如,序列衔接子可包括流通池序列、样品条形码或两者。在另一个实例中,序列

衔接子可以是发夹形衔接子、Y形衔接子、叉状衔接子,和/或包含样品条形码。在某些情况下,衔接子不包含测序引物区域。在某些情况下,衔接子附着的DNA分子被扩增,并且扩增产物如本文所述富集特定序列。在某些情况下,在准备测序文库后,DNA分子富集了特定序列。衔接子可包含条形码序列。不同的条形码可以是至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25或更多个(或如遍及全文描述的任何长度)核酸碱基,例如7个碱基。条形码可以是随机序列、简并序列、半简并序列或定义的序列。在某些情况下,存在足够的条形码多样性,使得大体上(例如,至少70%、至少80%、至少90%或至少99%的)每种核酸分子以不同条形码序列加标签。在某些情况下,存在足够的条形码多样性,使得来自特定遗传基因座的大体上(例如,至少70%、至少80%、至少90%或至少99%的)每种核酸分子以不同条形码序列加标签。

[0087] 测序衔接子可包含能够与一种或更多种测序引物杂交的序列。测序衔接子还可包含与固体支持物杂交的序列,例如流通池序列。例如,测序衔接子可以是流通池衔接子。该测序衔接子可附接至多核苷酸片段的一端或两端。在另一个实例中,测序衔接子可以是发夹形。例如,发夹形衔接子可包含互补双链部分和环部分,其中该双链部分可附接(例如,连接)至双链多核苷酸。发夹形测序衔接子可附接至多核苷酸片段的两端以生成环形分子,可对该环形分子多次测序。

[0088] 在一些情况下,所述文库衔接子均不含有样品标识基序(或样品分子条形码)。可经由测序衔接子提供这样的样品标识基序。样品标识基序可包含至少4、5、6、7、8、9、10、20、30或40个核苷酸碱基的测序仪序列,该测序仪序列允许从来自其他样品的多核苷酸分子中鉴定来自给定样品的多核苷酸分子。例如,这可允许在相同池中对来自两个受试者的多核苷酸分子进行测序并随后鉴定所述受试者的序列读段。

[0089] 测序仪基序包括使文库衔接子与测序系统偶联并对与该文库衔接子偶联的靶多核苷酸进行测序所需的核苷酸序列。该测序仪基序可包括与流通池序列互补的序列和可与用于在测序中使用的引物(或引发序列)选择性地杂交的序列(测序起始序列)。例如,这样的测序起始序列可与在合成测序(例如,Illumina)中使用的引物互补。这样的引物可包含在测序衔接子中。测序起始序列可以是引物杂交位点。

[0090] 在一些情况下,所述文库衔接子均不含有完整的测序仪基序。所述文库衔接子可不含有或含有部分测序仪基序。在一些情况下,所述文库衔接子包含测序起始序列。所述文库衔接子可包含测序起始序列但不包含流通池序列。该测序起始序列可与用于测序的引物互补。该引物可以是序列特异性引物或通用引物。这样的测序起始序列可位于所述文库衔接子的单链部分上。作为替代方案,这样的测序起始序列可以是引发位点(例如,扭结(kink)或切口)以允许聚合酶在测序过程中与文库衔接子偶联。

[0091] 衔接子可通过连接附着于DNA分子。在某些情况下,衔接子与双链DNA分子连接,使得每个衔接子将DNA分子的互补链不同地加标签。在一些情况下,衔接子序列可通过PCR附着,其中单链DNA的第一部分与靶序列互补,且第二部分包含衔接子序列。

[0092] 可通过序列捕获方法进行对感兴趣的特定序列的富集。序列捕获可使用与感兴趣的靶杂交的固定的探针进行。序列捕获可使用附接于官能团例如生物素的探针进行,其允许探针与特定序列杂交以通过下拉(pulldown)从样品富集所述特定序列。在一些情况中,在与官能化的探针杂交之前,来自文库片段的特定序列诸如衔接子序列可通过使互补的、

未官能化的多核苷酸序列与片段退火来掩蔽,以减少非特异性或脱靶结合。序列探针可靶向特定基因。序列捕获探针可靶向特定遗传基因座或基因。这样的基因可以是致癌基因。被捕获探针靶向的示例性基因包括图1所示的那些。带有点突变(SNV)的示例性基因包括但不限于,AKT1、ATM、CCNE1、CTNNB1、FGFR1、GNAS、JAK3、MLH1、NPM1、PTPN11、RIT1、TERT、ALK、BRAF、CDH1、EGFR、FGFR2、HNF1A、KIT MPL、NRAS、RAF1、ROS1、TP53、APC、BRCA1、CDK4、ERBB2、FGFR3、HRAS、KRAS、MYC、NTRK1、RB1、SMAD4、TSC1、AR、BRCA2、CDK6、ESR1、GATA3、IDH2、MAP2K2、NFE2L2、PIK3CA、RHEB、SRC、ARID1A、CCND2、CDKN2B、FBXW7、GNAQ、JAK2、MET、NOTCH1、PTEN、RHOA和STK11。带有拷贝数变异的示例性基因包括但不限于,AR、CCNE1、CDK6、ERBB2、FGFR2、KRAS、MYC、PIK3CA、BRAF、CDK4、EGFR、FGFR1、KIT、MET、PDGFRA和RAF1。具有基因融合的示例性基因包括但不限于:ALK、FGFR2、FGFR3、NTRK1、RET和ROS1。具有插入缺失的示例性基因包括但不限于:EGFR(例如,在外显子19和20处)、ERBB2(例如,在外显子19和20处)、和MET(例如,跳读外显子14)。示例性靶可包括CCND1和CCND2。序列捕获探针可平铺在一个基因上(例如,探针可靶向重叠区域)。序列探针可靶向非重叠区域。序列探针可对长度、解链温度和二级结构优化。

[0093] 鸟嘌呤-胞嘧啶(GC)含量的定量量度

[0094] 鸟嘌呤-胞嘧啶含量是DNA分子中为鸟嘌呤或胞嘧啶的含氮碱基的百分比。与遗传基因座的GC含量相关的定量量度可以是整个遗传基因座的GC含量。与遗传基因座的GC含量相关的定量量度可以是基因的外显子区域的GC含量。与遗传基因座的GC含量相关的定量量度可以是被映射到遗传基因座的读段覆盖的区域的GC含量。与GC含量相关的定量量度可以是对应于遗传基因座的序列捕获探针的GC含量。与遗传基因座的GC含量相关的定量量度可以是对应于遗传基因座的序列捕获探针的GC含量的中心趋势相关的量度。与中心趋势相关的量度可以是中心趋势的任何量度,如平均值、中位数或众数。与中心趋势相关的量度可以是中位数。给定区域的GC含量可通过将鸟苷和胞嘧啶碱基的数量除以该区域上的碱基总数来测量。

[0095] 测序读段覆盖的定量量度

[0096] 与测序读段覆盖相关的定量量度是指示来源于对应于遗传基因座(例如来自参考基因组的特定位置、碱基、区域、基因或染色体)的DNA分子的读段的数量的量度。为了将读段与遗传基因座相关联,读段可被映射于参考或与参考比对。进行映射或比对的软件(例如,Bowtie、BWA、mrsFAST、BLAST、BLAT)可将测序读段与遗传基因座相关联。在映射过程期间,可优化特定的参数。优化映射过程的非限制性示例可包括掩蔽重复区域;采用映射质量(例如,MAPQ)分数截止;使用不同的种子长度来生成比对;并限制基因组位置之间的编辑距离。

[0097] 与测序读段覆盖相关的定量量度可包括与遗传基因座相关的读段的计数。在某些情况下,将计数转换为新的度量,以减轻不同测序深度、文库复杂度或遗传基因座大小的影响。示例性度量为每百万的每千碱基读段(Read Per Kilobase per Million,RPKM)、每百万的每千碱基片段(Fragments Per Kilobase per Million,FPKM)、M值的修剪平均值(TMM)、方差稳定原始计数和对数转换原始计数。可用于特定应用的其他转换对于本领域技术人员也是已知的。

[0098] 可使用叠并读段来确定定量量度,其中每个叠并读段对应于初始模板DNA分子。叠

并和定量读段家族的方法见于PCT/US2013/058061和PCT/US2014/000048中,其中的每一个通过引用整体并入本文。具体而言,可使用以下叠并方法,所述叠并方法使用条形码和来自测序读段的信息以将读段叠并为家族,使得每个家族共享条形码序列和至少一部分的测序读段序列。于是,对于大多数的家族,每个家族来源于单个初始模板DNA分子。从来自于家族的映射序列得到的计数可称为“独特分子计数”(UMC)。在某些情况下,确定与测序读段覆盖相关的定量量度包括通过与文库大小相关的度量将UMC归一化以提供归一化的UMC(“归一化的UMC”)。示例性方法是将遗传基因座的UMC除以所有UMC的总和;将遗传基因座的UMC除以所有常染色体UMC的总和。当比较多个测序读段数据集时,UMC可例如被两个测序读段数据集的遗传基因座的中值UMC归一化。在某些情况下,与测序读段覆盖相关的定量量度可以是归一化UMC,其如下进一步归一化:(i)从来源于训练样本的测序读段,确定相应的遗传基因座的归一化UMC;(ii)对于每个遗传基因座,样本的归一化UMC通过训练样本在相应基因座处的归一化UMC的中值归一化,由此提供遗传基因座的相对丰度(RA)。

[0099] 共有序列可基于它们的序列来鉴定,例如通过基于前5个、10个、15个、20个或25个碱基内的相同序列叠并测序读段。在某些情况下,叠并允许另外相同的读段中的1个差异、2个差异、3个差异、4个差异或5个差异。在某些情况下,叠并使用读段的映射位置,例如测序读段的初始碱基的映射位置。在某些情况下,叠并使用条形码,且共享条形码序列的测序读段被叠并成共有序列。在某些情况下,叠并使用条形码和初始模板分子的序列二者。例如,共享条形码和映射到参考基因组中相同位置的所有读段可被叠并。在另一个实例中,可叠并共享条形码和初始模板分子的序列(或与初始模板分子的序列的同一性百分比)的所有读段。

[0100] 在某些情况下,测序读段覆盖的定量量度是对基因组的特定子区域确定的。区域可以是箱(bin)、感兴趣的基因、外显子、对应于序列探针的区域、对应于引物扩增产物的区域、或对应于引物结合位点的区域。在某些情况下,基因组的子区域是对应于序列捕获探针的区域。如果读段的至少一部分映射到对应于序列捕获探针的区域的至少一部分,则读段可映射到对应于序列捕获探针的区域。如果读段的至少一部分映射到对应于序列捕获探针的区域的大部分,则读段可映射到对应于序列捕获探针的区域。如果读段的至少一部分映射跨对应于序列捕获探针的区域的中心点,则读段可映射到对应于序列捕获探针的区域。在某些情况下,与遗传基因座的测序读段覆盖相关的定量量度是遗传基因座内对应于基因组位置的探针的RA的中位数。例如,如果KRAS被三个探针覆盖,其具有2、3和5的RA,则遗传基因座的RA将是3。

[0101] “饱和平衡”校正

[0102] 通常,本文描述的方法可用于增加核酸样品中变体调用(例如检测拷贝数变体)的特异性和灵敏度。例如,方法可减少数据样本中的噪声或失真的数量,减少检测到的假阳性变体的数量。当噪声和/或失真降低时,特异性和灵敏度增加。噪声可被认为是信号的不希望的随机增加。失真可被认为是信号或信号的一部分的幅值改变。

[0103] 噪声可通过在拷贝和/或读取多核苷酸中的错误而引入。例如,在测序过程中,单个多核苷酸可首先经历扩增。扩增可能引入错误,致使扩增的多核苷酸的亚组可能在特定的基因座处包含与在该基因座处的原始碱基不同的碱基。此外,在读取过程中,在任何特定基因座处的碱基可能被不正确地读取。因此,序列读段的集合可能在基因座处包含一定百

分比的与原始碱基不同的碱基判定。在通常的测序技术中,这种错误率可以是个位数,例如,2%-3%。当对全部被假定为具有相同序列的分子的集合进行测序时,这样的噪声是足够小的,使得人们可以以高可靠性鉴定原始碱基。

[0104] 然而,如果亲本多核苷酸的集合包括具有在特定基因座处的序列变体的多核苷酸亚组,则噪声可能是一个显著的问题。例如,当无细胞DNA不仅包括种系DNA还包括来自另一个来源的DNA诸如胎儿DNA或来自癌细胞的DNA时,可能是这样的情况。在这种情况下,如果具有序列变体的分子的频率与通过测序过程引入的错误的频率在相同的范围内,则真序列变体可能无法与噪声区别。这可能会干扰,例如,样品中的序列变体的检测。

[0105] 失真在测序过程中可表现为由亲本群体中相同频率的分子产生的信号强度的差异,例如,序列读段的总数。例如,可通过扩增偏倚、GC偏倚或测序偏倚引入失真。这可能会干扰样品中的拷贝数变异的检测。GC偏倚导致了在序列读取中GC含量丰富或贫乏区域的不均匀呈现。

[0106] 本文公开的方法包括通过以下确定用于确定基线的遗传基因座的初始集:从数据集去除对其的测序读段覆盖相关的定量量度或测序读段覆盖相关的转化的定量量度与预测模型差异最大的那些遗传基因座(其在本文中可被称为去除高变异遗传基因座),从而提供剩余遗传基因座的第一集。在一些情况下,去除这些遗传基因座包括拟合将测序读段覆盖相关的定量量度与遗传基因座的GC含量相关的定量量度相关联的模型。例如,预测模型可将遗传基因座的RA与基因座的GC含量相关联。在某些情况下,预测模型是回归模型,包括非参数回归模型,例如LOESS和LOWESS回归模型。在某些情况下,基线定线通过去除与预测模型偏离最大的遗传基因座的5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%或70%来进行。在某些情况下,基线定线通过去除与预测模型偏离最大的遗传基因座的至少5%、至少10%、至少15%、至少20%、至少25%、至少30%、至少35%、至少40%、至少45%、至少50%、至少55%、至少60%、至少65%或至少70%来进行。在某些情况下,通过测量遗传基因座相对于模型的残差来确定偏离。可以选择确切的截断值以提供排除从剩余遗传基因座的特定数量的变异。

[0107] PCT/US2014/072383中公开了确定来源于遗传基因座的来自样品的DNA链在测序读段内被代表的概率相关的定量量度的方法,其全部内容通过引用特此并入。确定定量量度可包括估计样品中存在的来源于基因座的初始模板DNA分子的数量。双链多核苷酸不生成序列读段的概率可以基于代表初始模板DNA分子的两条链的读段和仅代表初始模板DNA分子的单链的读段的相对数量来确定。

[0108] 样品中未检测到的初始模板DNA分子的数量可以基于代表初始模板DNA分子的两条链的读段和仅代表初始模板DNA分子的单链的读段的相对数量来估计。作为实例,记录了特定遗传基因座即基因座A的计数,其中1000个分子是配对的(例如,两个链都被检测到)并且1000个分子是未配对的(例如,仅检测到单个链)。应注意,这里使用的术语“配对”和“未配对”与有时将其应用于测序读段以指示分子的两端还是单端被测序的这些术语不同。对于单个Watson或Crick链,假设使其通过转化后的过程具有统一的概率p,人们可以如下计算未能使其通过该过程的(未发现的)分子的比例:R,配对与未配对分子之比=1000/1000=1,因此$R=1=p^2/(2p(1-p))$。这意味着p=2/3,并且丢失的分子的量等于$(1-p)^2=1/9$。因此,在该实例中,大约11%的转化分子丢失并且从未被检测到。除了采用二项分布外,估算

未发现的分子的数目的其他方法还包括基于观察到的序列读段的冗余的指数、β、γ或经验分布。在后一种情况下,配对和未配对分子的读出计数的分布可来源于这样的冗余,从而推断特定基因座处原始多核苷酸分子的潜在分布。这通常可导致更好地估算未发现的分子的数目。在某些情况下,p是来源于遗传基因座的来自样品的DNA链在测序读段内被代表的概率相关的定量量度。在一些情况下,p被类似地推导,但是使用不同的读段分布模型(例如,二项式、泊松、β、γ和负二项分布)。

[0109] 对于测序读段覆盖的定量量度的转换可通过以下来确定:将来自去除了高变异遗传基因座的遗传基因座的集的定量量度或转换的测序读段覆盖与GC含量相关的定量量度和来源于遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度相关联。在某些情况下,剩余的遗传基因座被假定为二倍体和/或以相同的拷贝数存在。在某些情况下,转换通过以下来确定:将剩余遗传基因座的测序读段覆盖相关的定量量度的中心趋势相关的量度与GC含量相关的定量量度和来源于遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度拟合。转换可以,例如,(i)将去除高变异遗传基因座后剩余遗传基因座的测序读段覆盖的定量量度的中心趋势与GC含量相关的定量量度和来源于遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度二者拟合。在某些情况下,剩余基因座的测序读段覆盖的定量量度的中心趋势相关的量度是剩余遗传基因座的UMC的中心趋势。在某些情况下,表面近似用于将剩余遗传基因座的UMC的表面或剩余遗传基因座的UMC的中心趋势与以下拟合:(i)GC含量相关的定量量度和(ii)来源于遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度。例如,表面近似可以是初始模板DNA分子相关的量度(例如,UMC)与GC含量和p的定量量度的二维二次多项式表面拟合。在某些情况下,测序覆盖相关的转换的定量量度是从以下计算的基于以上确定的转换预期的值:(i)GC含量相关的定量量度和(ii)来源于遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度。在某些情况下,测序覆盖相关的转换的定量量度是每个遗传基因座的残差(例如,基于表面近似的基因座的测序读段覆盖相关的预期定量量度与样品中遗传基因座的测序读段覆盖相关的观察到的定量量度的差异或商)。任选地,在确定了测序覆盖相关的转换的定量量度之后,可以基于测序读段覆盖的新转换的定量量度再次如上所述去除高变异遗传基因座。

[0110] “探针效率”校正

[0111] 本文公开了使用参考样品确定和去除遗传基因座的偏倚的方法。在某些情况下,参考样品是来自无癌症的受试者的无细胞DNA的测序读段。在某些情况下,参考样品是来自具有在感兴趣的遗传基因座中基本上没有拷贝数变异的癌症细胞的受试者的无细胞DNA的测序读段。在某些情况下,参考样品是来自患有癌症的受试者的无细胞DNA的测序读段,其中怀疑已经历拷贝数变异的区域被排除在分析之外。在某些情况下,参考样品是来自无癌症的受试者的血浆样品。在某些情况下,参考样品是来自患有癌症的受试者的血浆样品。

[0112] 参考样品的每个遗传基因座可以如以上“饱和平衡校正”中描述的处理,以提供测序读段覆盖的转换的定量量度。在某些情况下,测序覆盖相关的转换的定量量度是从以下计算的基于以上确定的转换预期的值:(i)GC含量相关的定量量度和(ii)来源于来自参考遗传基因座的遗传基因座的DNA链在测序读段内被代表的概率相关的定量量度。在某些情况下,测序覆盖相关的转换的定量量度是每个参考遗传基因座的残差(例如,基于表面近似的基因座的测序读段覆盖相关的预期定量量度与参考样品中遗传基因座的测序读段覆盖

相关的观察到的定量量度的差异或商)。参考样品中遗传基因座的测序读段覆盖相关的转换的定量量度可被认为是遗传基因座的“效率”。例如,低效扩增的遗传基因座比非常有效地扩增的遗传基因座(在样品中以相同拷贝数存在)具有更低的UMC。

[0113] 可以基于来自参考样品的遗传基因座确定的效率来校正与样品的测序读段覆盖相关的转换的定量量度。这种校正可以减少由从样品产生测序读段的过程引入到样品中的变异,这可以与连接效率、下拉效率、PCR效率、流通池聚类损失、解复用损失、叠并损失和比对损失相关。在一个实施方案中,校正包括将样品的测序覆盖的饱和后的转换的定量量度除以或减去与测序覆盖相关的预测的饱和后转换的定量量度。在某些情况下,与遗传基因座的测序覆盖有关的预测的饱和后转换的定量量度通过拟合来自样品的遗传基因座的测序覆盖相关的饱和后转换的定量量度与参考的测序读段覆盖相关的饱和后转换的定量量度之间的关系来确定。在一些情况下,拟合包括对参考的测序读段覆盖相关的饱和后转换的定量量度进行来自样品的遗传基因座的测序覆盖相关的饱和后转换的定量量度的局部回归(例如LOESS或LOWESS)或稳健线性回归。在某些情况下,拟合可以是线性回归、非线性回归或非参数回归。

[0114] 任选地,来自探针效率校正的转换的定量量度可以是输入到“饱和平衡校正”转换中的输入,以产生变异减小的与测序读段覆盖相关的第三、进一步转换的定量量度。一般而言,为了进一步减少在测序读段覆盖的转换的定量量度内的变异,可以使用本文公开的任何方法来转换测序覆盖的转换的定量量度另外的次数。

[0115] 基因水平概要

[0116] 推断的拷贝数的基因水平概要可以基于如本文所公开的确定的测序读段覆盖的转换的定量量度来确定。通过丢弃高变异遗传基因座,可以推断相对于上述操作中选择的基线的拷贝数。例如,如果样品中剩余的遗传基因座被推断为二倍体,那么测序覆盖相关的转换的定量量度与基线不同的遗传基因座可以被推断为在肿瘤细胞中经历了拷贝数改变。在某些情况下,使用观察到的探针信号的基因水平中位数和使用基因中观察到的探针水平标准差计算的估计的标准差以及全基因组正常二倍体探针信号标准差,计算基因水平z-分数。

[0117] 次等位基因频率基线优化

[0118] 本文提供了使用测序读段中的变体的次等位基因频率来检测本文所述的拷贝数的基因水平概要中的错误和校正错误的方法。10%和90%之间、20%和80%之间、30%和70%之间、40%和60%之间、或约50%的来自无细胞体液的核酸的测序读段中存在的序列变体可以是受试者的种系序列中存在的杂合变体。在一些情况下,遗传基因座已被确定经历了如上所述的扩增。将变体的数量与推断的拷贝数进行比较以确定变体频率是否与推断的拷贝数不一致。在一个实例中,杂合遗传基因座可以在用于确定基线拷贝数的遗传基因座(例如,排除高变异遗传基因座后剩余的遗传基因座)中检查。在某些情况下,样品中大量的遗传基因座已被扩增,并且该基线可能被错误鉴定。在这种情况下,杂合性可能偏离1:1的比率,且检测并纠正不准确的基线定线。在第二个实例中,例如,基于与测序读段覆盖相关的转换的定量量度,可以推断遗传基因座以三倍体拷贝数存在。如果受试者的种系基因组具有带有遗传基因座的第一等位基因的一个染色体,并且第二染色体具有第二等位基因,则第一或第二等位基因可在癌细胞中已经重复。

[0119] Langmuir样饱和模型

[0120] 不受理论束缚,本文公开了假设为基于探索历史临床数据以及涉及合成的加标模型系统的靶向实验的诱饵-cfDNA相互作用的调控机制的Langmuir样饱和模型。因此,在干扰测定效应(例如连接效率、PCR扩增偏差、测序假象等)的不存在下,诱饵下拉过程可被描述为

[0121] $$\text{独特分子计数} = I_{sat}\frac{K \bullet \text{拷贝数}}{1 + K \bullet \text{拷贝数}}$$

[0122] 本描述中的K是诱饵效率,其依赖于诱饵序列特征以及其与基因组中邻近靶向的诱饵位置的DNA片段的相互作用。I_{sat}是由下拉反应中有限的初始诱饵计数驱动的饱和参数,其是总诱饵池浓度以及复制计数的函数。本文使用的复制计数是指存在的序列捕获探针的相对或绝对量。例如,序列捕获阵列可以提供阵列上不同摩尔量的探针,以考虑不同的探针效率。图8说明了基于诱饵效率K和饱和参数I_{sat}使真实拷贝数和独特分子计数相关联的模型。

[0123] 诱饵效率K主要由GC含量驱动,而I_{sat}由更复杂的诱饵耗尽机制和RNA二级结构相互作用驱动,可通过研究独特分子计数与总读段计数相互作用来粗略检查。除了非线性下拉反应外,探针信号还可以通过包含以下假设的乘法模型进一步建模:在幼稚模型下,cfDNA片段按基因组位置均匀分布,且随机采样过程是造成覆盖变异的支配因素。然后,可以通过将观察到的UMC与样品中的真实分子计数相关联来对拷贝数信号(例如,UMC)建模,考虑到潜在位置cfDNA特征(profile)、连接效率、下拉效率、PCR效率、流通池聚类损失、解复用损失、叠并损失和比对损失。

[0124] 除了非线性下拉反应外,探针信号还可以通过简单乘法模型进一步建模,包括以下假设。在幼稚模型下,cfDNA片段按基因组位置均匀分布,且随机采样过程是造成覆盖变异的支配因素。然后,拷贝数信号,即与给定探针相关的读段计数可以被建模为:

[0125] 观察到的UMC=真实UMC x潜在位置cfDNA特征(诱饵,cfDNA片段)x连接效率(位置、大小、cfDNA片段)x下拉效率(探针,cfDNA片段)xPCR效率(DNA片段)x流通池聚类损失x解复用损失和叠并损失x比对损失(cfDNA片段序列)。

[0126] 这个模型假设了上述模型的乘性。可以通过一系列步骤,例如本文公开的基线确定方法,从观察到的UMC(例如,给定序列捕获探针的UMC)关于已建立的基线推断潜在的诱饵特异性拷贝数信号。

[0127] 本文公开的方法提供了用于从样本集和训练集估算探针效率和诱饵饱和的方法。可选地,可以通过进行一组诱饵滴定实验来推断这样的参数,其中对每种探针观察到改变靶序列浓度对UMC的影响。如果K、I_{sat}和UMC是已知的,则确定对应于未经历拷贝数变异的肿瘤细胞的UMC值或范围可以是可能的。例如,假设大部分遗传基因座没有经过拷贝数改变,观察到的UMC将主要来源于二倍体样品。已经经历拷贝数变异的样品将是其UMC落在探针的相应的K和I_{sat}值的期望范围之外的那些遗传基因座。在某些情况下,例如,UMC值或范围将是依赖于每种探针的K和I_{sat}的函数。例如,对应于二倍体拷贝数的UMC在两种探针之间可以是不同的。

[0128] 计算机控制系统

[0129] 本公开内容提供了被编程为实现本公开内容的方法的计算机控制系统。图12示出

了被编程或以其他方式配置为实现本公开内容的方法的计算机系统1201。计算机系统1201包括中央处理单元(CPU,本文也被称为“处理器”和“计算机处理器”)1205,其可以是单核或多核处理器或用于并行处理的多于一个处理器。计算机系统1201还包括存储器或存储器位置1210(例如,随机存取存储器、只读存储器、闪速存储器)、电子存储单元1215(例如,硬盘)、用于与一个或更多个其他系统进行通信的通信界面1220(例如,网络适配器)和外围设备1225,诸如高速缓冲存储器、其他存储器、数据存储和/或电子显示适配器。存储器1210、储存单元1215、界面1220和外围设备1225与CPU 1205通过通信总线(实线),诸如主板(motherboard)通信。存储单元1215可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统1201可借助于通信界面1220被可操作地耦合至计算机网络(“网络”)1230。网络1230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1230为电信和/或数据网络。网络1230可包括局部区域网络。网络1230可包括一个或更多个计算机服务器,这可支持分布式计算,诸如云计算。在一些情况下,借助于计算机系统1201,网络1230可实现对等网络(peer-to-peer network),其可使耦合至计算机系统1201的设备能够作为客户端或服务器运行。

[0130] CPU 1205可执行一系列的机器可读指令,该机器可读指令可以以程序或软件来体现。指令可被存储于存储器位置,诸如存储器1210中。指令可被导向CPU 1205,其可随后编程或以其他方式配置CPU 1205,以实现本公开内容的方法。由CPU 1205进行的操作的实例可包括读取、解码、执行和写回。

[0131] CPU 1205可以是电路诸如集成电路的一部分。系统1201的一个或更多个其他组件可被包含在该电路中。在一些情况下,电路为专用集成电路(ASIC)。

[0132] 存储单元1215可存储文件,诸如驱动程序、库和保存的程序。存储单元1215可存储用户数据,例如,用户偏好和用户程序。在一些情况下,计算机系统1201可包括一个或更多个另外的数据存储单元,该数据存储单元在计算机系统1201的外部,诸如位于通过内联网或因特网而与计算机系统1201通信的远程服务器上。

[0133] 计算机系统1201可与一个或更多个远程计算机系统通过网络1230进行通信。例如,计算机系统1201可与用户的远程计算机系统进行通信。远程计算机系统的实例包括个人计算机(例如便携式PC)、板型或平板PC(例如Apple® iPad、Samsung® Galaxy Tab)、电话、智能电话(例如Apple® iPhone、Android支持的设备、Blackberry®)或个人数字助理。用户可经由网络1230访问计算机系统1201。

[0134] 如本文描述的方法可通过机器(例如,计算机处理器)可执行代码的方式实现,该机器可执行代码被存储在计算机系统1201的电子存储位置,诸如,例如存储器1210或电子存储单元1215上。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,代码可由处理器1205执行。在一些情况下,代码可从存储单元1215检索并存储在存储器1210上,以用于由处理器1205即时访问。在一些情况下,可排除电子存储单元1215,而将机器可执行指令存储于存储器1210中。

[0135] 代码可被预编译并配置为用于与具有适用于执行该代码的处理器的机器一起使用,或者可在运行时间期间被编译。代码可以以编程语言的形式提供,该编程语言可被选择以便使得代码能够以预编译的或按编译原样(as-compiled)的方式被执行。

[0136] 本文所提供的系统和方法的各方面,诸如计算机系统1201,可以以编程来体现。技术的多个方面可被认为是通常呈一种机器可读介质进行或体现的机器(或处理器)可执行代码和/或相关数据的形式的“产品”或“制品(articles of manufacture)”。机器可执行代码可被存储于电子存储单元诸如存储器(例如,只读存储器、随机存取存储器、闪速存储器)或硬盘上。“存储”型介质可包括计算机、处理器等的任何或所有有形存储器,或其相关模块,诸如多种半导体存储器、磁带驱动器、磁盘驱动器等,其可在任何时间为软件编程提供非暂时性存储。该软件的所有或部分有时可通过因特网或多种其他电信网络进行通信。例如,此类通信可使得将软件从一个计算机或处理器加载到另一个计算机或处理器,例如,从管理服务器或主机加载到应用服务器的计算机平台。因此,能够携带软件元件的另一类型的介质包括诸如在本地设备之间的物理界面、通过有线和光纤陆线网络以及在多种空中链路(air-links)上使用的光波、电波和电磁波。携带此类波的物理元件,诸如有线或无线链路、光链路等,也可被认为是携带软件的介质。如本文使用的,除非被限制为非暂时性的、有形的“储存”介质,否则术语诸如计算机或机器“可读介质”指参与将指令提供至处理器用于执行的任何介质。

[0137] 因此,机器可读介质,诸如计算机可执行代码,可采取多种形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括,例如光盘或磁盘,诸如在任何计算机等中的任何存储设备,诸如可用于实现如附图中示出的数据库等。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。有形的传输介质包括同轴电缆;铜线和光纤,包括构成计算机系统内的总线的导线。载波传输介质可采取电信号或电磁信号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,计算机可读介质的常见形式包括,例如:软盘(floppy disk)、软性磁盘(flexibledisk)、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、穿孔卡片纸带、具有打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,或者计算机可从其读取编程代码和/或数据的任何其他介质。这些计算机可读介质的形式中的许多形式可参与向处理器传送一个或更多个指令的一个或更多个序列以用于执行。

[0138] 计算机系统1201可包括电子显示器1235或与之通信,该电子显示器1235包括用户界面(UI)1240,用于提供例如报告。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。

[0139] 本公开内容的方法和系统可通过一个或更多个算法来实现。算法可在由中央处理器1205执行后通过软件来实现。

实施例

[0140] 实施例1

[0141] 检查先前生成的拷贝数变异加标数据揭示了在原始读段计数和UMC二者中的显著探针间信号变异,以及探针/基因水平拷贝数信号对潜在拷贝数变化的响应。参见图2。图3说明了三种基因(CCND1、CCND2和ERBB2)的推断拷贝数相比于理论拷贝数,证明了归一化的覆盖对样品中诱饵量的非线性响应。这些结果表明在下拉期间的诱饵耗竭,这通过在相同基因内在相邻探针中追踪诱饵滴定效应而证实,且独特分子计数有相当大的差异(从而对

于具有高初始UMC的探针观察到更快的独特分子计数饱和)。

[0142] 图4A说明,与每种探针相关的UMC关于探针p具有非线性响应。图4B说明,与每种探针相关的UMC关于探针GC含量具有非线性响应。

[0143] 图5说明未进行饱和或探针效率校正的探针的UMC。图6显示饱和校正后的相同样品。图7显示探针效率校正后的相同样品。在每个阶段,基因组位置内的变异减少,导致肿瘤细胞潜在拷贝数变异的更清晰图像出现。图7中中位数探针探针效率校正后信号高于1.2的基因被称为已经发生了拷贝数变异。不同水平的探针效率校正后信号可能是由于肿瘤异质性或继发性肿瘤。

[0144] 图9显示了饱和校正和探针效率校正后基线定线遗传基因座探针信号降噪的典型进程。

[0145] 图10A示出了x轴上的参考样品的探针效率和来自肿瘤细胞中无拷贝数变异的受试者的样品的饱和校正后信号的图。关系是近似线性的。图10B说明来自肿瘤细胞中带有拷贝数变异的受试者的类似图。响应不像图10A那样线性。由通过确定来自参考样品的探针效率与基线定线遗传基因座(以黑色表示)的饱和校正后的UMC之间的关系来推断的预测的效率校正,将减少推定已经在肿瘤细胞中经历拷贝数扩增的遗传基因座(灰色圆点)中由于不同探针效率的变异。图11示出了基于饱和和探针效率校正后的UMC和MAF优化的基线定线,来自患者样品的拷贝数变异的示例性报告。星号表示指示属于在受试者的肿瘤细胞中已经经历拷贝数变异的遗传基因座的点。

[0146] 实施例2

[0147] 无细胞DNA从患有癌症的受试者获得,制备了条形码化的测序文库,通过用探针集序列捕获来富集一组致癌基因,并对条形码化的测序文库进行测序。将测序读段映射到参考基因组,并基于条形码序列和映射位置叠并成家族。对于对应于来自探针组的探针的中点的每个基因组坐标,计数跨越该中点的读段家族的数目以获得每探针UMC。对每种基因确定中位数每探针UMC。为了进行“饱和平衡校正”,将基因按其中位数每探针GC含量分组。去除其中位数每探针UMC与具有相似的中位数每探针GC含量的那些基因显著不同的基因。

[0148] 对于每种探针,如本文所述确定p和GC含量。来自上一步的剩余的基因用于进行中位数基因水平UMC的二维二次多项式表面拟合以探测p和GC含量。将p和GC含量与预期的UMC相关联的函数用于确定预期的每探针UMC。通过将观察到的每探针UMC除以期望的每探针UMC来确定数据集的残差。每种探针的残留UMC是测序覆盖的转换的定量量度。

[0149] 将基因再次按其中位数每探针GC含量分组,并且去除了其中位数每探针残留UMC与具有相似中位数每探针GC含量的基因显著不同的基因。“探针效率”校正然后如前述段落中所述的通过获得参考样品的残留UMC进行。然后将来自样品的每种探针的残留UMC除以来自参考的每种相应探针的残留UMC以获得探针效率校正后的UMC。

[0150] 类似于以上的饱和平衡校正,剩余的基因用于进行探针效率校正后的UMC的二维二次多项式表面拟合以探测p和GC含量。将p和GC含量与预期的探针效率校正后的UMC相关联的函数用于确定预期的每探针探针效率校正后的UMC。通过将观察到的每探针探针效率校正后的UMC除以期望的每探针探针效率校正后的UMC来确定数据集的残差。每种探针的残留探针效率校正后的UMC是探针GC校正后的信号。

[0151] 将剩余基因按其中位数每探针GC含量分组,并且去除了其中位数每探针GC校正后

的信号与具有相似中位数每探针GC含量的那些基因显著不同的基因。

[0152] 重复该实施例的过程,以探针GC校正后的信号作为起始输入代替初始UMC。

[0153] 对于每种基因,使用探针GC校正后信号的中位数来总结每种基因。其中位数探针GC校正后信号与其他基因显著不同的基因被认为是肿瘤细胞中经历基因扩增或缺失的候选者。

[0154] 对于每种基因,确定种系杂合等位基因并定量每种等位基因的相对频率。发现用于基线定线的遗传基因座具有的等位基因比率为约1:1,验证了基线定线遗传基因座的选择。

[0155] 基于基因水平中位数探针GC校正后的信号和估计的从全基因组正常二倍体探针信号的标准差确定每种基因的Z-分数。Z-分数高于截止值的基因据报道在肿瘤细胞中已经经历了基因扩增。

[0156] 实施例3

[0157] 本文所述的方法通过在本公开内容的方法对比于对照方法中测量ERBB2拷贝数来验证。本公开内容的方法产生了观察到的拷贝数(CN)对比于理论拷贝数的线性响应,没有观察到正常(健康)队列中的假阳性CNV结果。参见图13,其显示推断的基因拷贝数相比于理论拷贝数估计值,其中实心圆点表示观察到的拷贝数~2(二倍体样品),开放圆点表示检测到的扩增事件且粗的水平虚线标记平均基因CN截止值。还参见图14,其描绘了图13的数据,其中对照数据由正方形表示。所有的CNV遵循预期的滴定趋势,降至2.15个拷贝。此外,由于变异减小,本公开内容的方法减少了数据中观察到的“噪声”,与对照方法相比,允许容易地区分CNV。参见图15的最右侧;三角形表示本公开内容的方法,而X表示对照方法。

[0158] 虽然本文已经显示和描述了本发明的优选实施方案,但对于本领域技术人员将明显的是,此类实施方案仅通过示例的方式提供。并非意图将本发明限制于本说明书中提供的具体实例。虽然已参考以上提及的说明书描述了本发明,但本文实施方案的描述和说明并不意图以限制性的意义来解释。在不偏离本发明的情况下,本领域技术人员现在将想到许多变化、改变和替换。此外,应当理解,本发明的所有方面并不限于本文阐述的取决于多种条件和变量的具体描写、配置或相对比例。应当理解,在实践本发明时可采用本文描述的本发明的实施方案的各种替代选择。因此可预期,本发明还应涵盖任何此类的替代选择、修改、变化或等同物。以下权利要求意图界定本发明的范围,并且从而涵盖在这些权利要求范围内的方法和结构及其等同物。

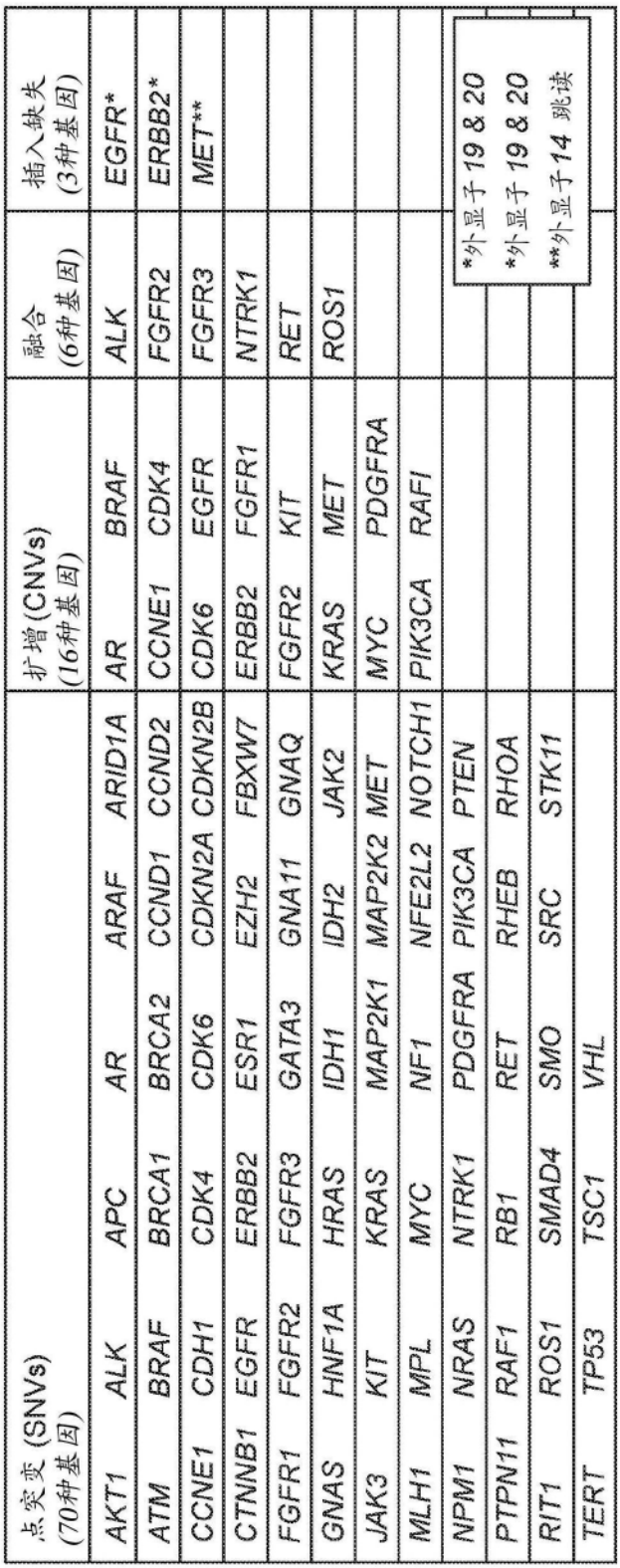

点突变 (SNVs)
(70种基因)
AKT1 ALK APC AR ARAF ARID1A
ATM BRAF BRCA1 BRCA2 CCND1 CCND2
CCNE1 CDH1 CDK4 CDK6 CDKN2A CDKN2B
CTNNB1 EGFR ERBB2 ESR1 EZH2 FBXW7
FGFR1 FGFR2 FGFR3 GATA3 GNA11 GNAQ
GNAS HNF1A HRAS IDH1 IDH2 JAK2
JAK3 KIT KRAS MAP2K1 MAP2K2 MET
MLH1 MPL MYC NF1 NFE2L2 NOTCH1
NPM1 NRAS NTRK1 PDGFRA PIK3CA PTEN
PTPN11 RAF1 RB1 RET RHEB RHOA
RIT1 ROS1 SMAD4 SMO SRC STK11
TERT TP53 TSC1 VHL
扩增(CNVs)
(16种基因)
AR BRAF CCNE1 CDK4 CDK6 EGFR
ERBB2 FGFR1 FGFR2 KIT KRAS MET
MYC PDGFRA PIK3CA RAFI
融合
(6种基因)
ALK FGFR2 FGFR3 NTRK1 RET ROS1
插入缺失
(3种基因)
EGFR*
ERBB2*
MET**
*外显子19 & 20
*外显子19 & 20
**外显子14 跳读

图1

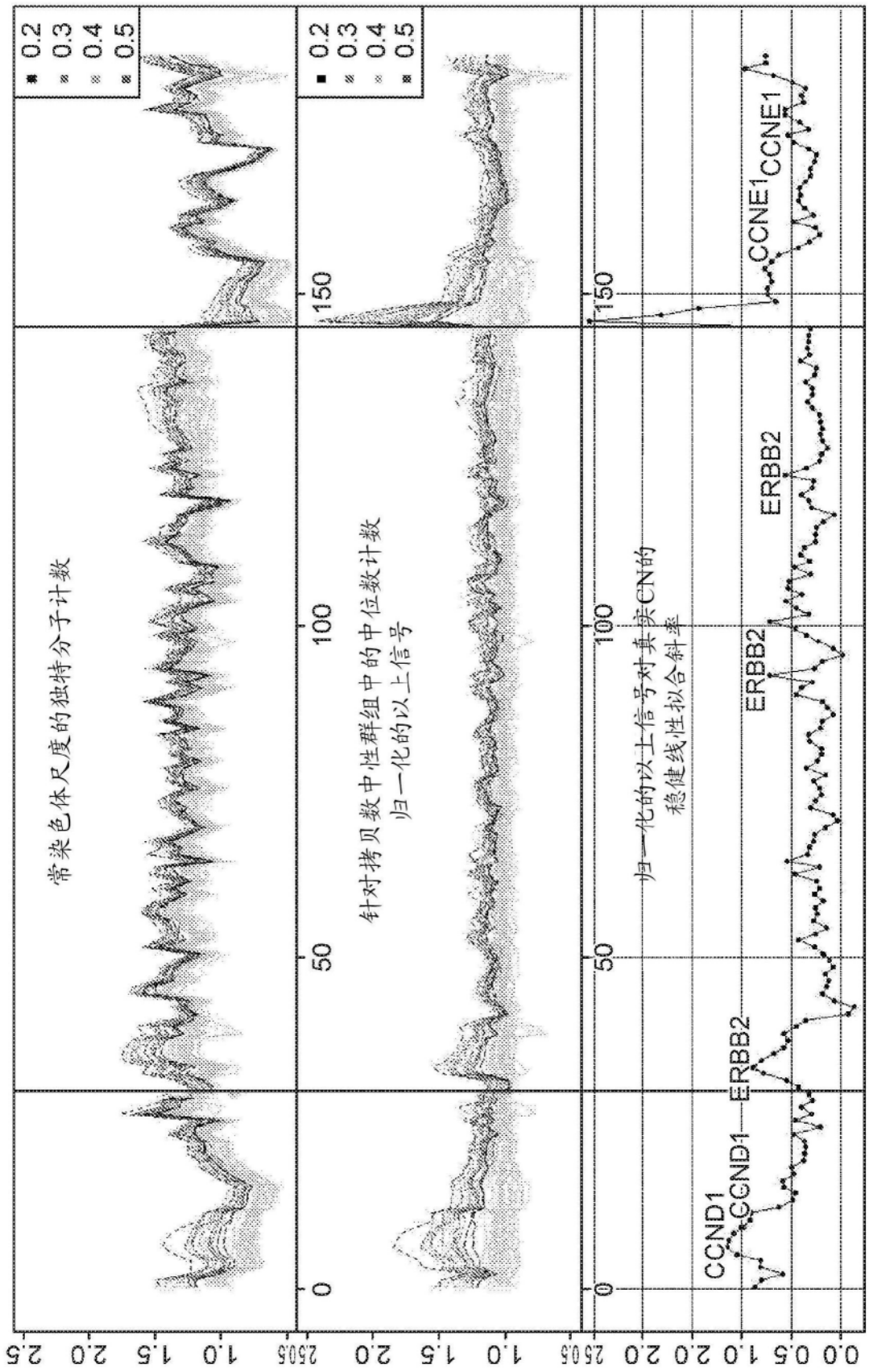

常染色体尺度的独特分子计数
针对拷贝数中性群组中的中位数计数归一化的以上信号
归一化的以上信号对真实CN的稳健线性拟合斜率
0.2
0.3
0.4
0.5
CCND1
ERBB2
ERBB2
ERBB2
CCNE1
CCNE1

图2

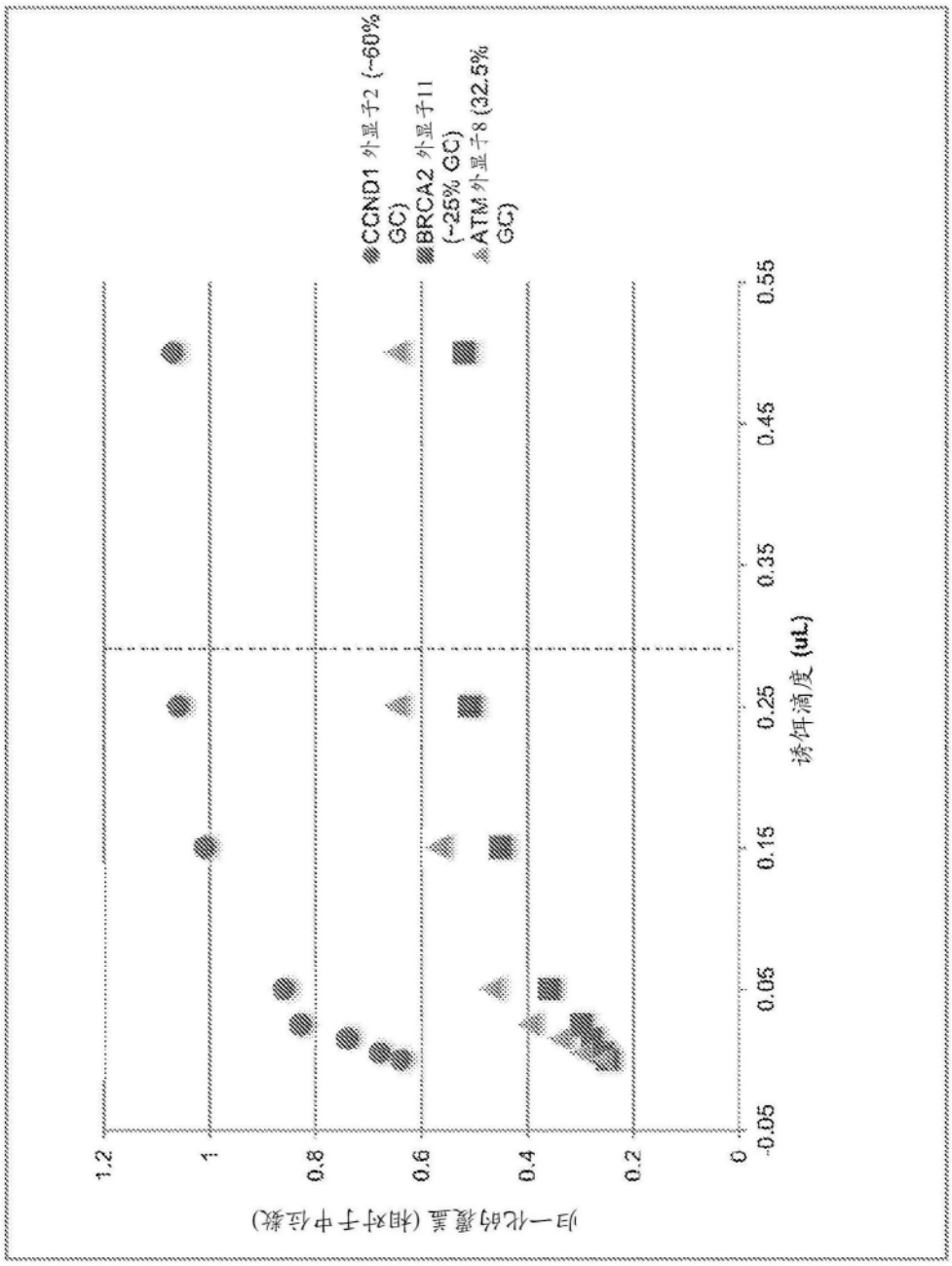

CCND1 外显子2 (~60% GC)
BRCA2 外显子11 (~25% GC)
ATM 外显子8 (32.5% GC)
归一化的覆盖 (相对于中位数)
诱饵滴度 (uL)
1.2
1
0.8
0.6
0.4
0.2
0
-0.05
0.05
0.15
0.25
0.35
0.45
0.55

图3

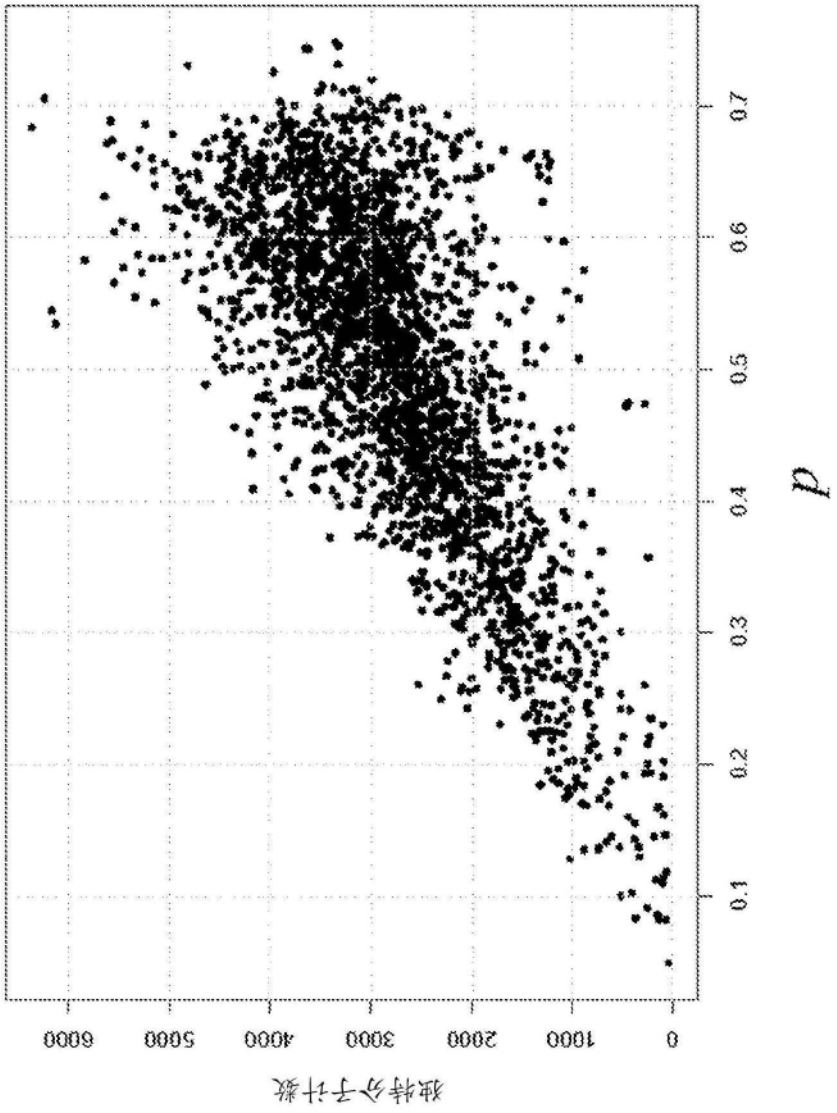


图4A

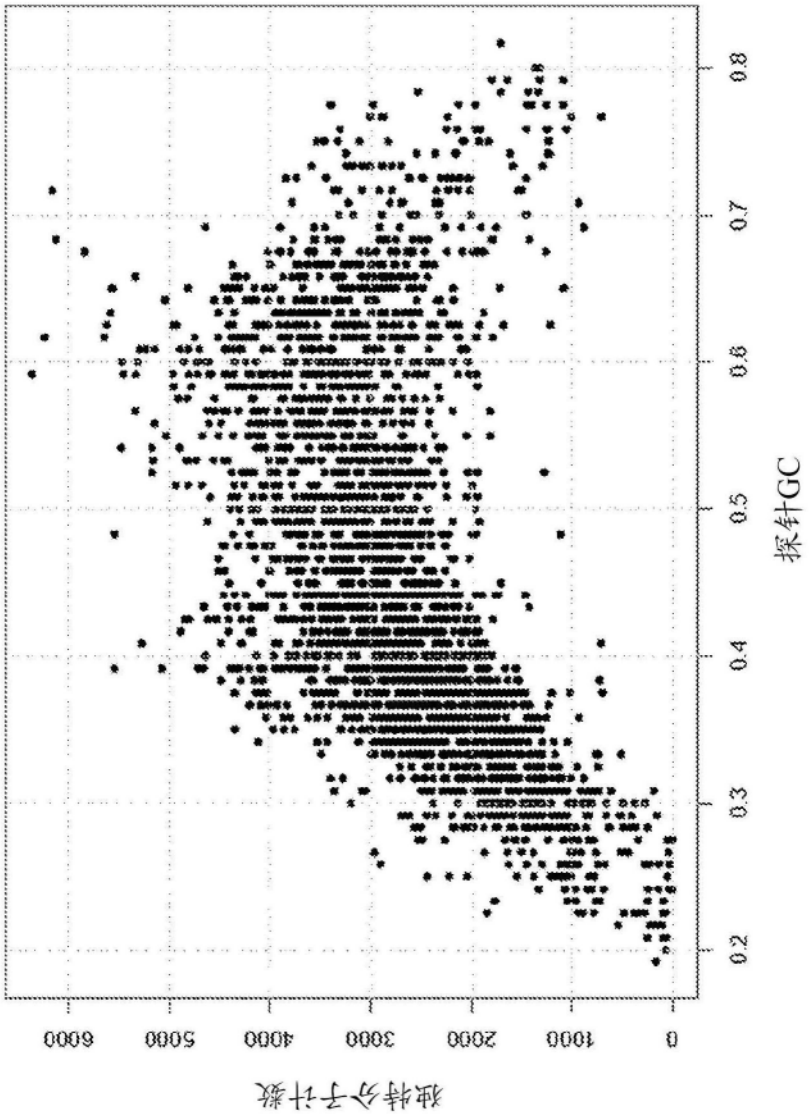


图4B

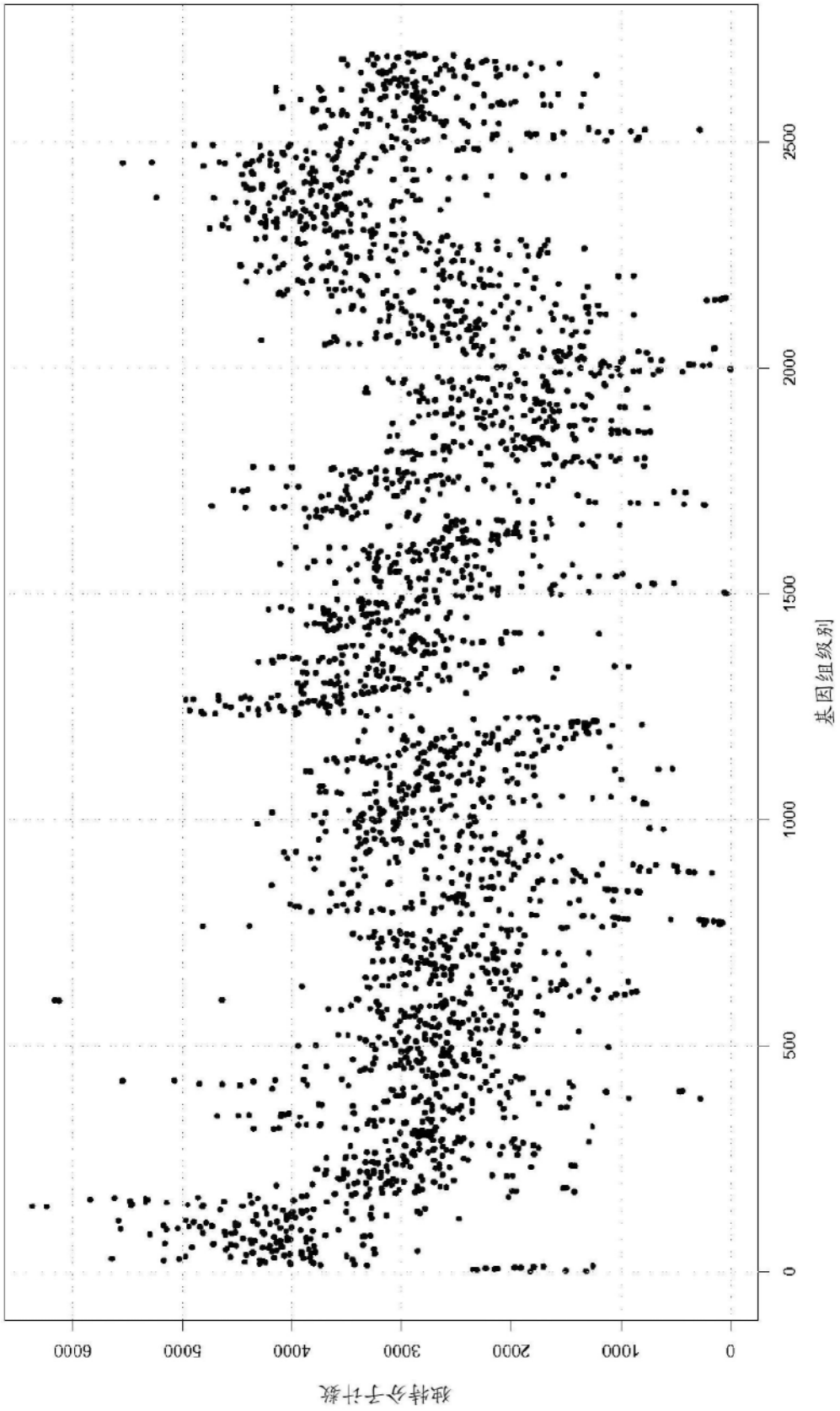

基因组级别
0
500
1000
1500
2000
2500
独特分子计数
0
1000
2000
3000
4000
5000
6000

图5

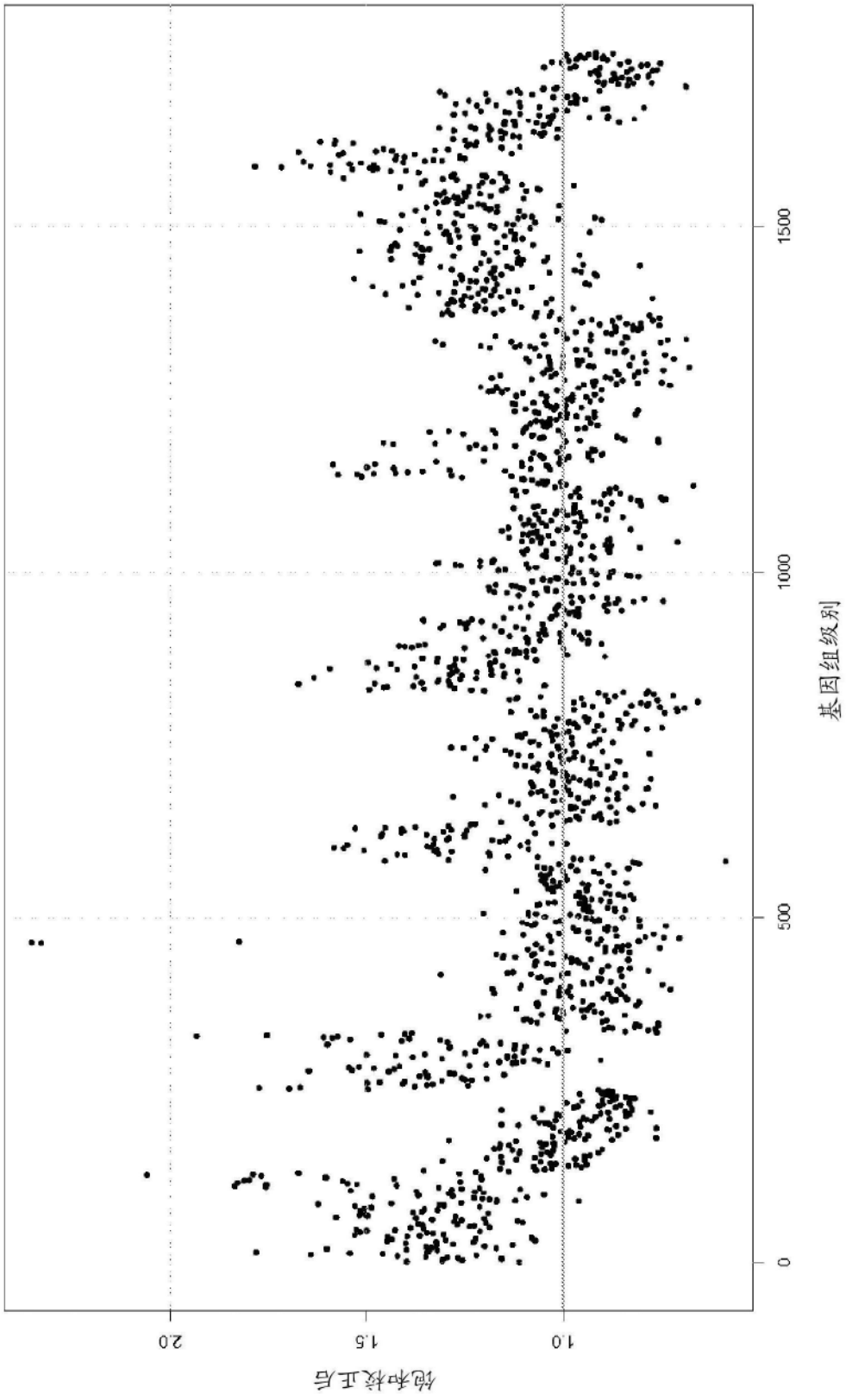


图6

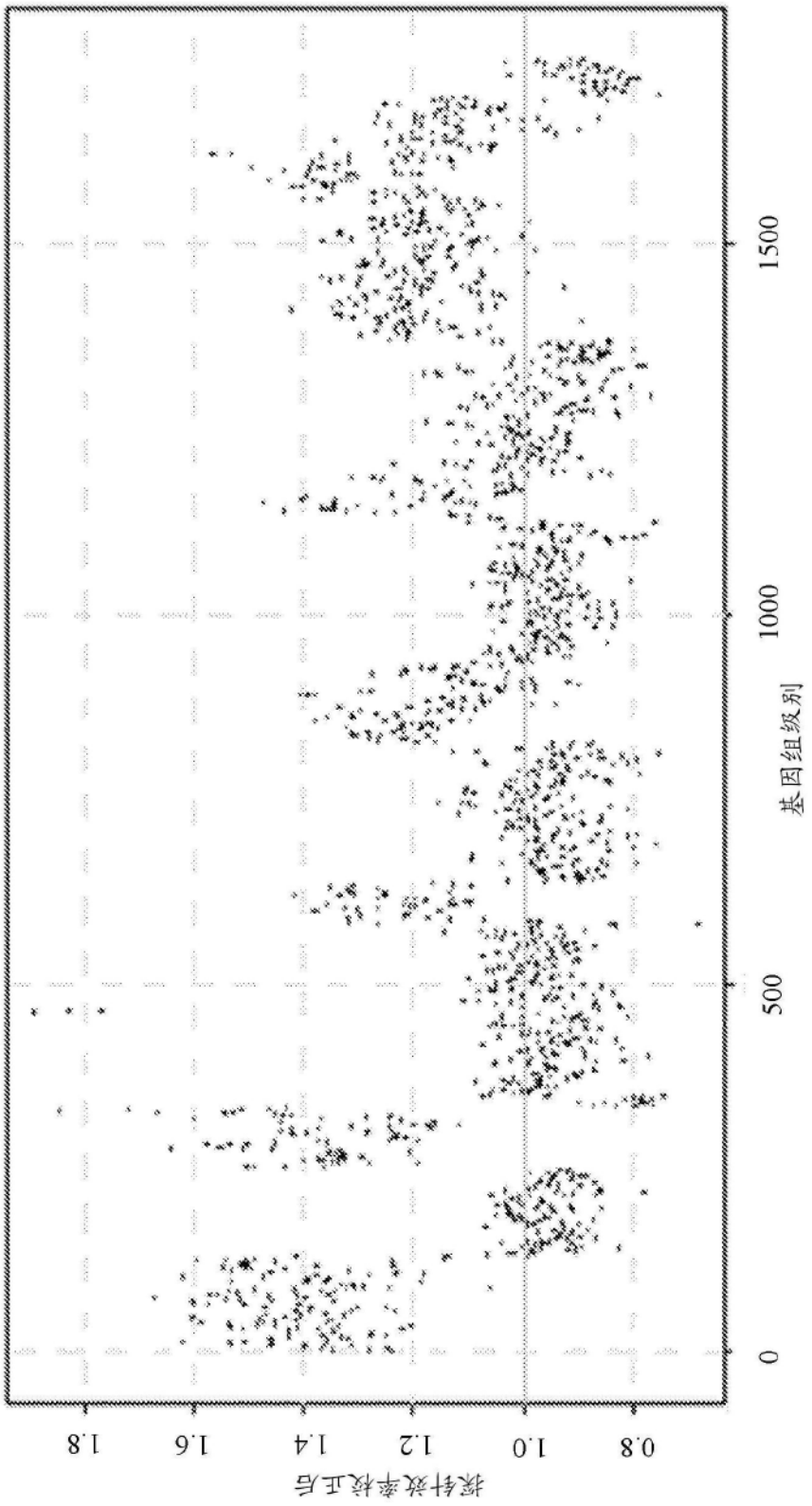


图7

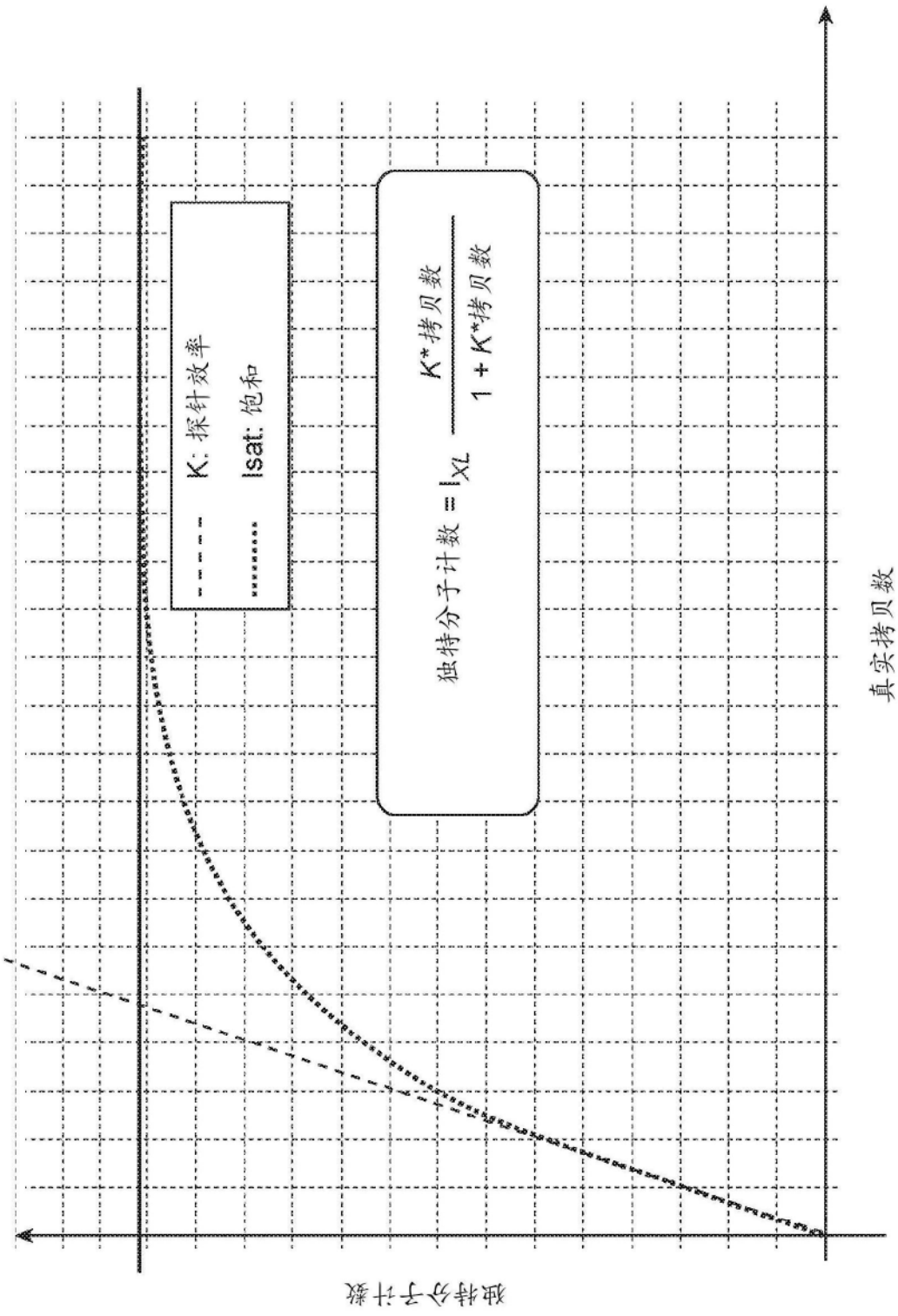


图8

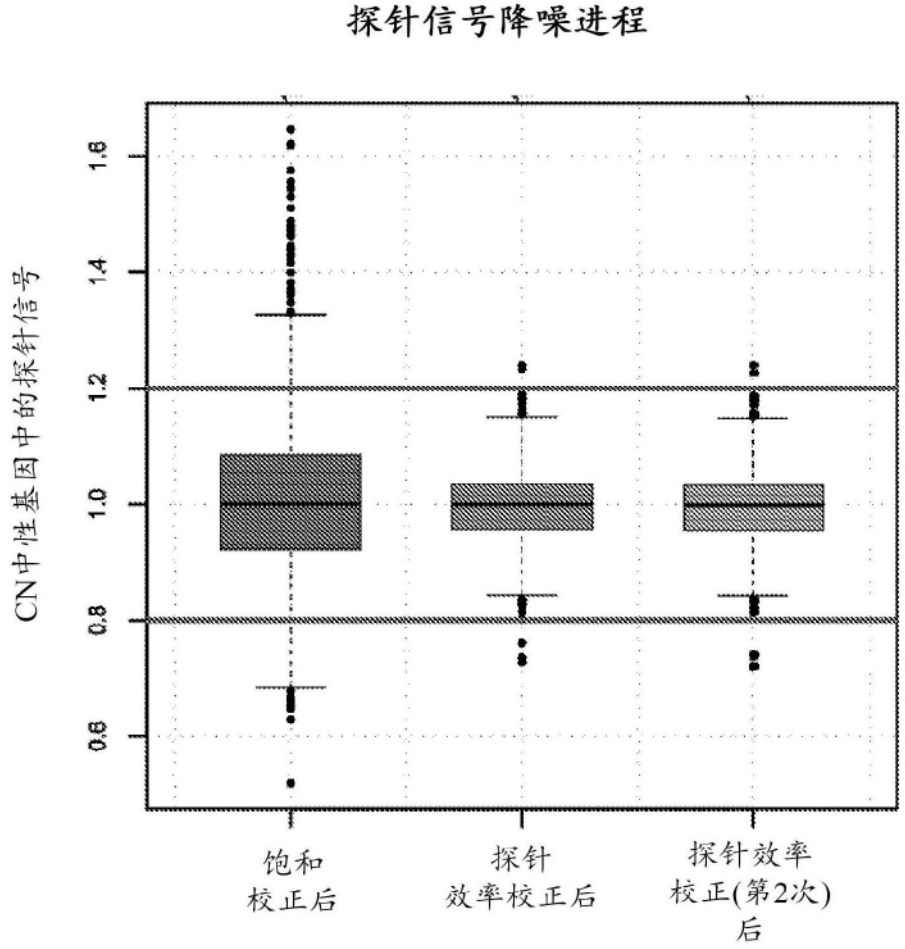

探针信号降噪进程
CN中性基因中的探针信号
0.6
0.8
1.0
1.2
1.4
1.6
饱和
校正后
探针
效率校正后
探针效率
校正(第2次)
后

图9

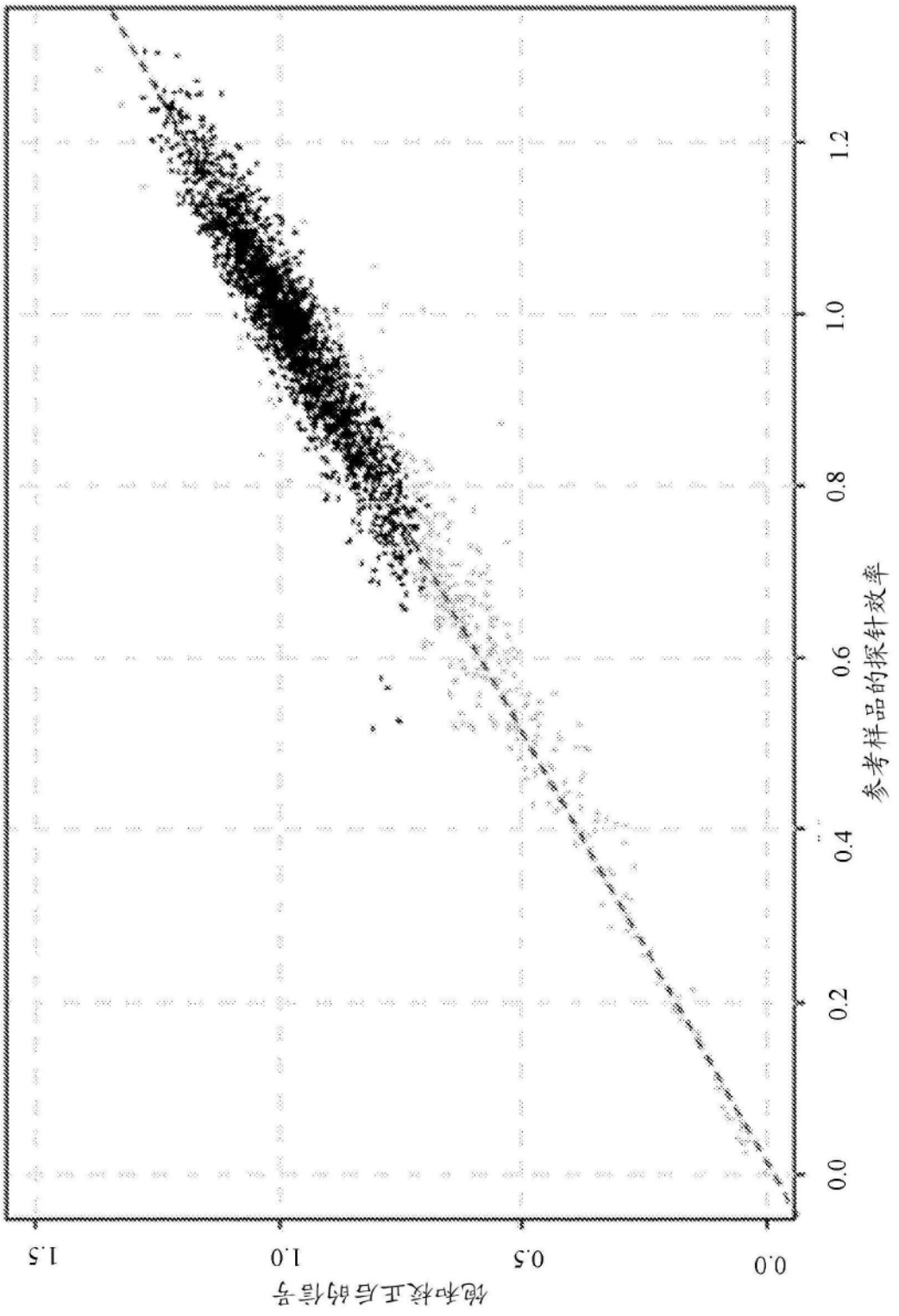


图10A

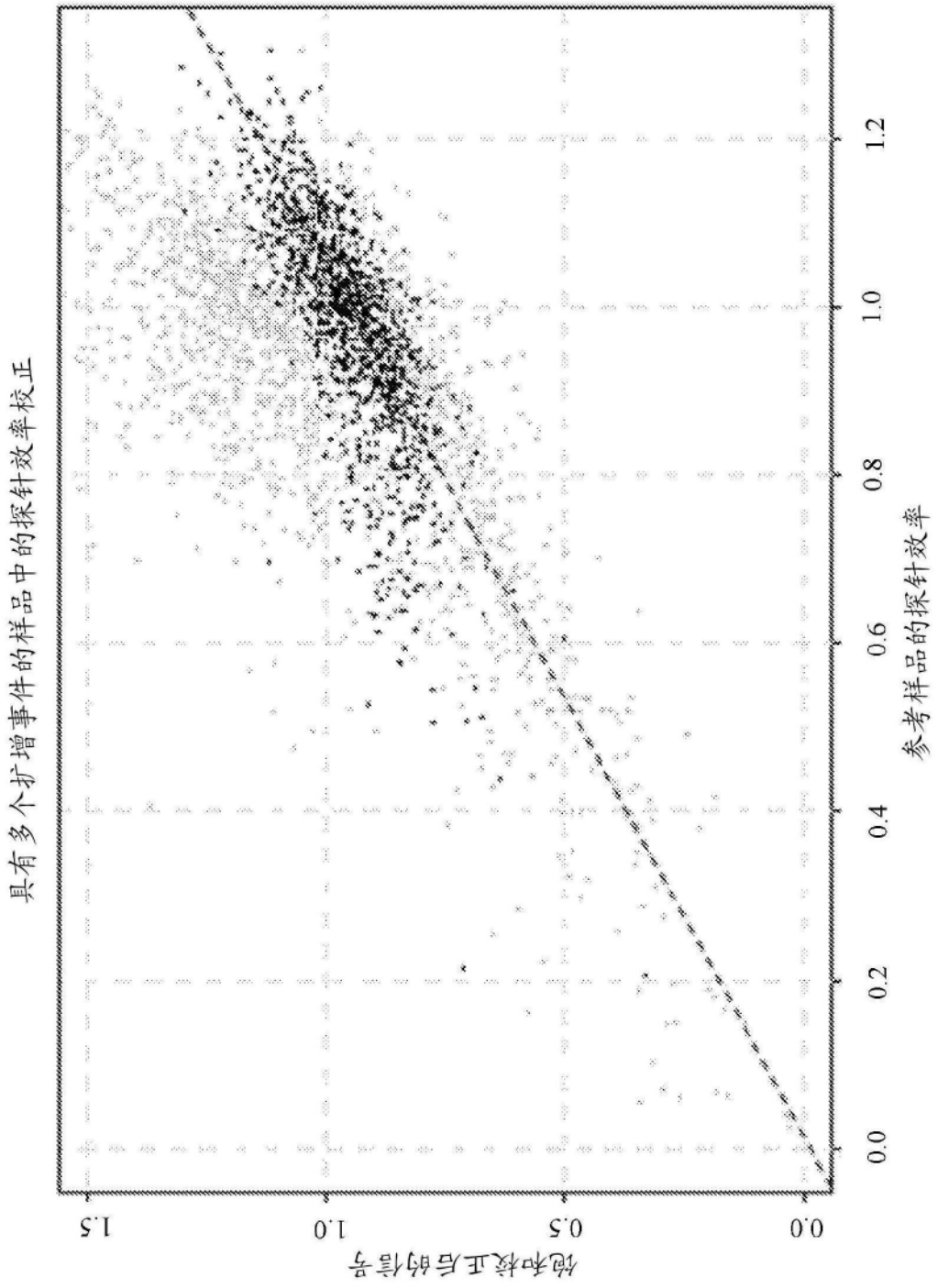

具有多个扩增事件的样品中的探针效率校正
饱和校正后的信号
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
参考样品的探针效率

图10B

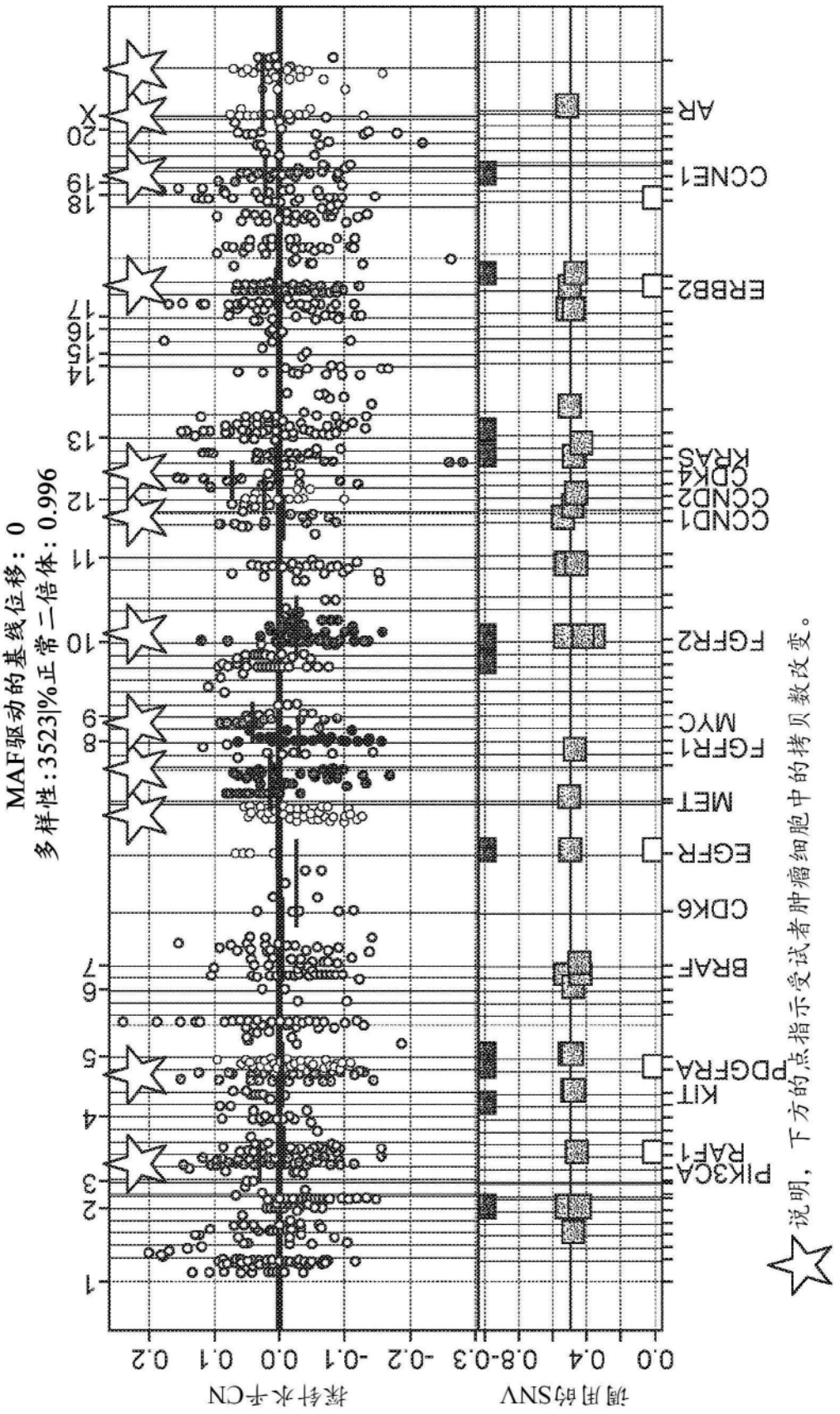

MAF驱动的基线位移：0
多样性:3523|%正常二倍体:0.996
调用的SNV
探针水平CN
0.0 0.4 0.8 0.3 0.2 0.1 0.0 0.1 0.2
AR
CCNE1
ERBB2
CCND2
CCND1
FGFR2
MYC
FGFR1
MET
EGFR
CDK6
BRAF
PDGFRA
KIT
RAF1
PIK3CA
说明，下方的点指示受试者肿瘤细胞中的拷贝数改变。

图11

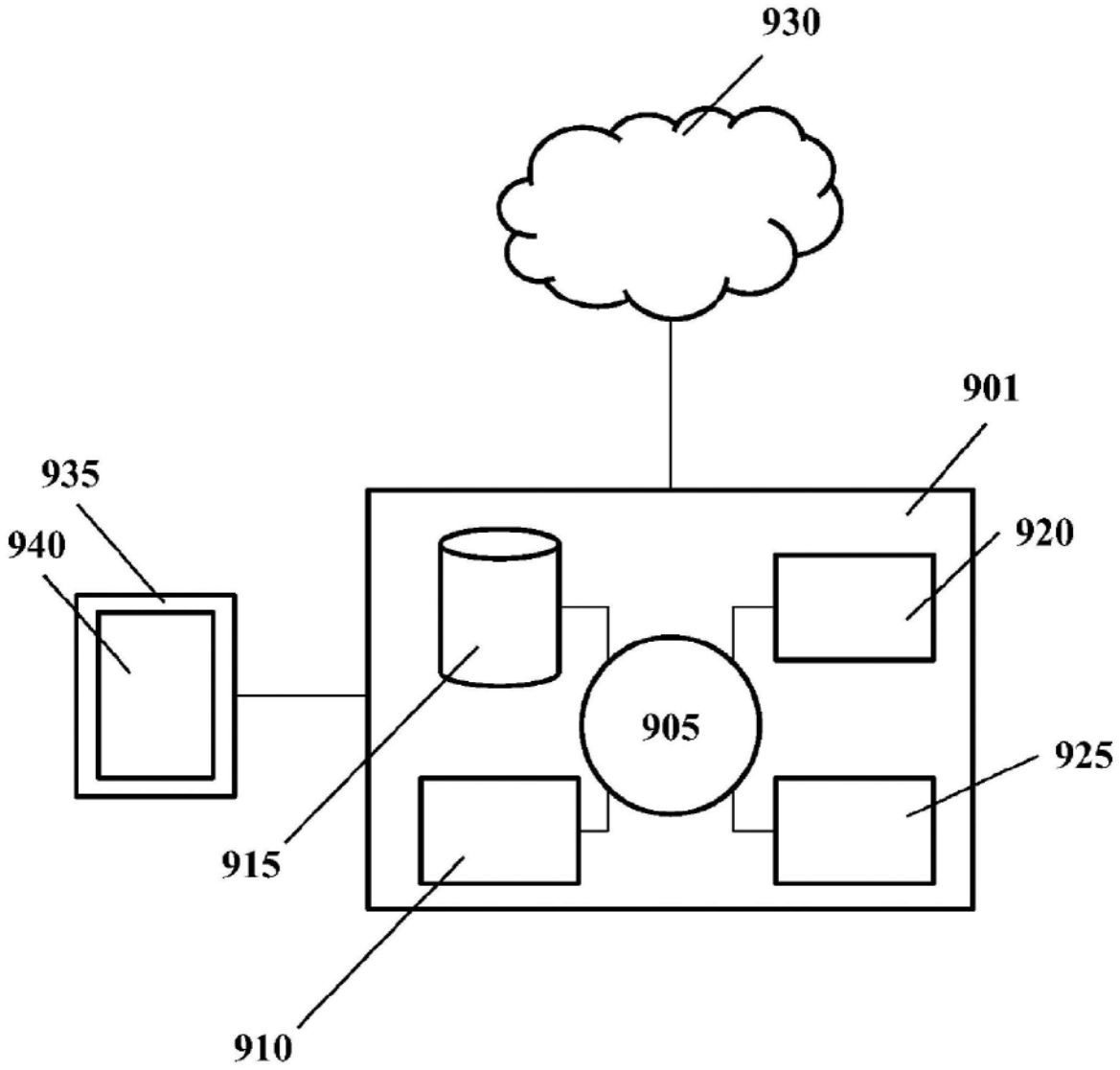


图12

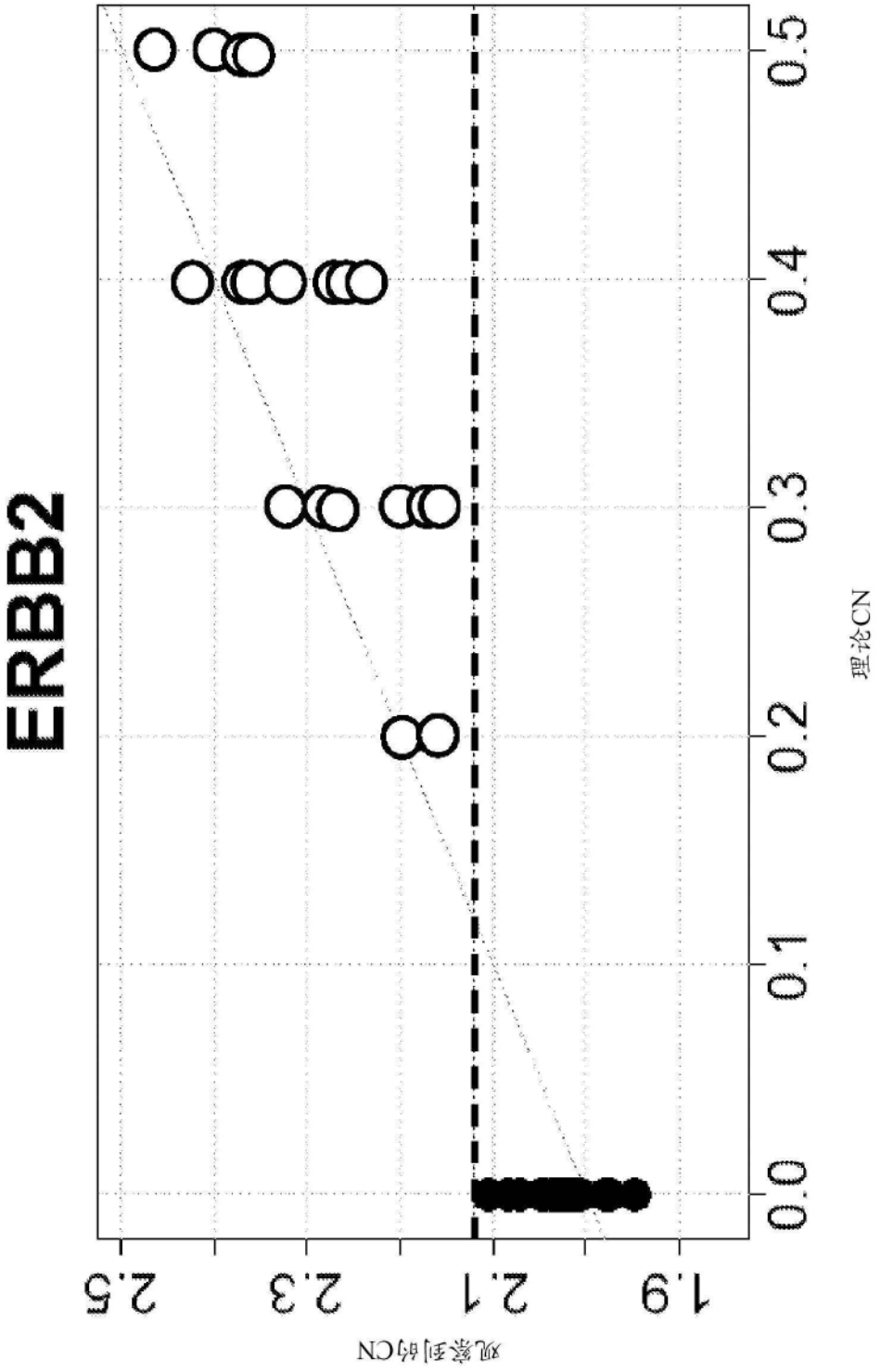


图13



图14

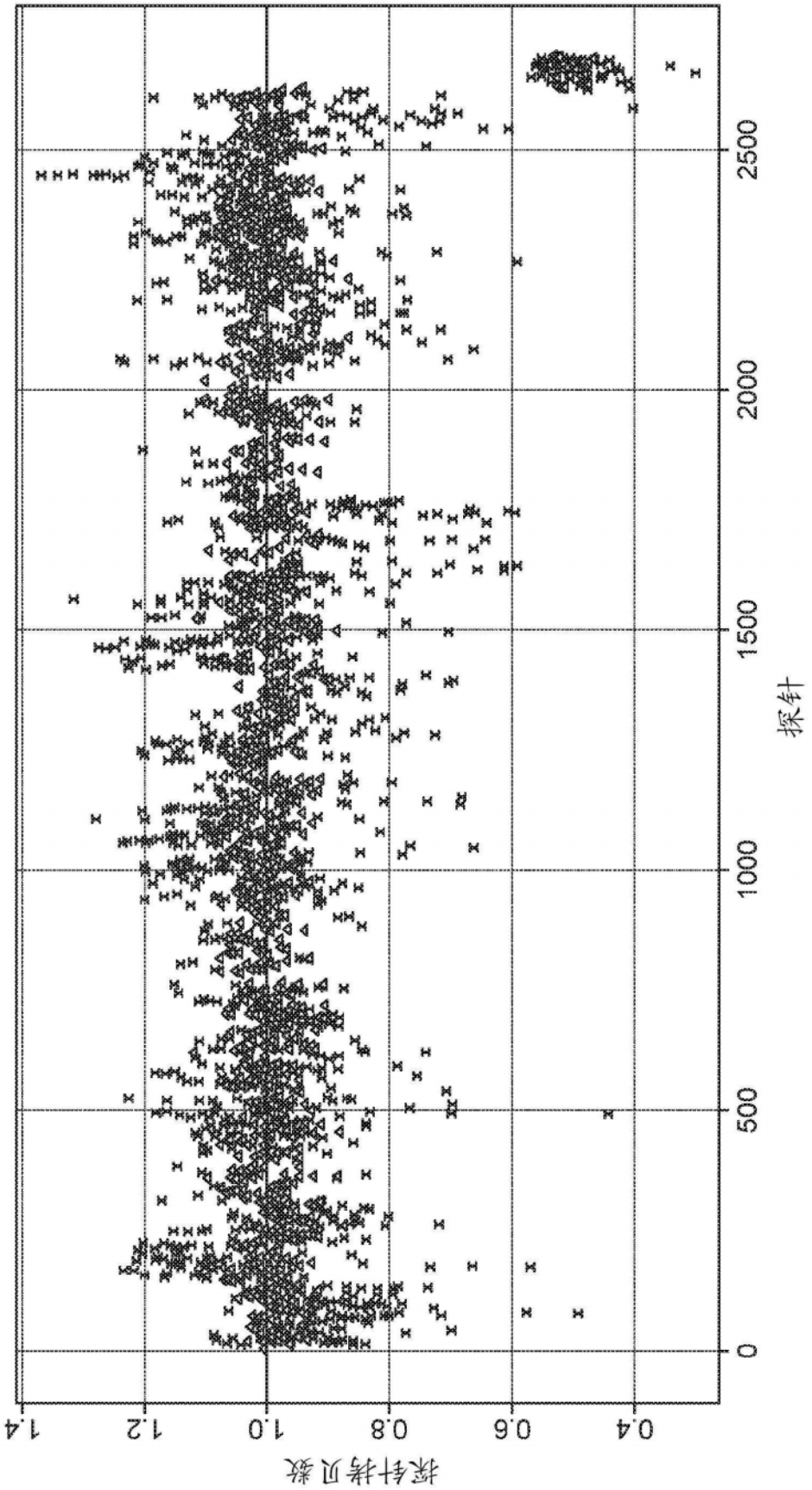


图15