(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118711654 A
(43)申请公布日 2024.09.27

(21)申请号 202410873100.9

(22)申请日 2018.05.16

(30)优先权数据
62/507,127 2017.05.16 US

(62)分案原申请数据
201880047621.1 2018.05.16

(71)申请人 夸登特健康公司
地址 美国加利福尼亚州
申请人 丹娜—法伯癌症研究所

(72)发明人 理查德·B·兰曼
杰弗里·R·奥克斯纳德

(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 王玮玮 郑霞

(51)Int.Cl.
G16B 20/10(2019.01)
G16B 20/20(2019.01)
G16B 30/10(2019.01)
G16B 30/20(2019.01)
G06N 3/123(2023.01)
C12Q 1/6886(2018.01)

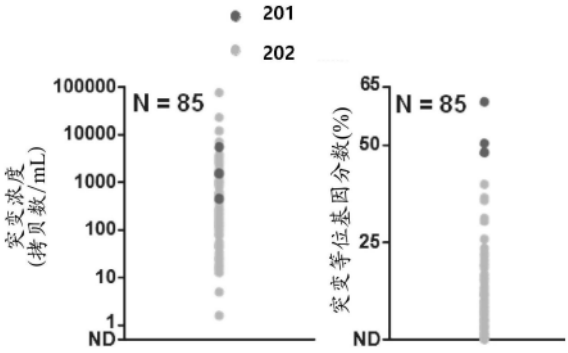
权利要求书3页 说明书64页 附图14页

(54)发明名称

无细胞DNA的体细胞来源或种系来源的鉴定

(57)摘要

本申请涉及无细胞DNA的体细胞来源或种系来源的鉴定。本公开内容提供了从无细胞DNA(cfDNA)中检测体细胞变异或种系变异的系统和方法。通常,这些系统和方法包括接收来自受试者的cfDNA的测序信息,该测序信息包括来自多于一个基因组基因座的cfDNA测序读段;基于该cfDNA测序读段来确定该多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;确定每一个所述AF量度的标准差(STDEV);提供STDEV阈值和AF阈值;确定每一个所述AF量度具有的STDEV高于还是低于该STDEV阈值;确定每一个所述AF量度高于还是低于该AF阈值;以及将STDEV低于该STDEV阈值且AF量度低于该AF阈值的每一个基因座分类为体细胞来源的,或者将STDEV低于该STDEV阈值且AF量度高于该AF阈值的每一个基因座分类为种系的。



CN 118711654 A

1.一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:

接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的cfDNA测序读段;

基于所述cfDNA测序读段来确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;

确定每一个所述AF量度的标准差(STDEV);

提供STDEV阈值和AF阈值;

确定每一个所述AF量度具有的STDEV高于还是低于所述STDEV阈值;

确定每一个所述AF量度高于还是低于所述AF阈值;以及

将STDEV低于所述STDEV阈值且AF量度低于所述AF阈值的每一个基因座分类为体细胞来源的。

2.一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:

接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的cfDNA测序读段;

基于所述cfDNA测序读段来确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;

确定每一个所述AF量度的标准差(STDEV);

提供STDEV阈值和AF阈值;

确定每一个所述AF量度具有的STDEV高于还是低于所述STDEV阈值;

确定每一个所述AF量度高于还是低于所述AF阈值;以及

将STDEV低于所述STDEV阈值且AF量度高于所述AF阈值的每一个基因座分类为种系来源的。

3.如权利要求1或2所述的方法,其中基因组基因座的AF量度低于所述STDEV阈值指示所述基因组基因座的低拷贝数变异(CNV)。

4.如权利要求1或2所述的方法,其中基因组基因座的AF量度高于所述STDEV阈值指示相关基因组基因座的高拷贝数变异(CNV)。

5.如权利要求1或2所述的方法,其中所述AF阈值根据经验确定。

6.一种用于在来自患有癌症的受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:

接收来自所述受试者在用癌症治疗剂治疗之前的第一时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第一组cfDNA测序读段;

接收来自所述受试者在用癌症治疗剂治疗之后的第二时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第二组cfDNA测序读段;

基于所述第一时间点的所述cfDNA测序读段并基于所述第二时间点的所述cfDNA测序读段,确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;

比较来自所述第一时间点和来自所述第二时间点的所述AF量度;

其中所述癌症对所述癌症治疗剂有响应;并且

如果来自基因组基因座的AF量度在所述第一时间点和所述第二时间点之间降低,则将所述基因组基因座鉴定为体细胞来源的。

7.一种用于在来自患有癌症的受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:

接收来自所述受试者在用癌症治疗剂治疗之前的第一时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第一组cfDNA测序读段;

接收来自所述受试者在用癌症治疗剂治疗之后的第二时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第二组cfDNA测序读段;

基于所述第一时间点的所述cfDNA测序读段并基于所述第二时间点的所述cfDNA测序读段,确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;

比较来自所述第一时间点和来自所述第二时间点的所述AF量度;

其中所述癌症对所述癌症治疗剂有响应;并且

如果来自基因组基因座的AF量度在所述第一时间点和所述第二时间点之间未降低,则将所述基因组基因座识别为种系来源的。

8.一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源或种系来源的方法,所述方法包括:

接收在第一时间点收集的来自所述受试者的所述cfDNA的测序信息,所述测序信息包括第一cfDNA测序读段;

提供来自所述多于一个基因组基因座的序列信息;

对所述多于一个基因组基因座中的每一个进行分箱,其中所述分箱包括为所述多于一个基因组基因座中的每一个基因组基因座分配初始分类,所述初始分类选自由以下组成的组:

a)假定体细胞来源;

b)假定种系来源;或者

c)不确定来源;

从而生成包含假定体细胞来源的基因组基因座的第一箱、包含假定种系来源的基因组基因座的第二箱和包含不确定来源的基因组基因座的第三箱;

对于所述第一箱、所述第二箱和所述第三箱中的所述基因组区域中的每一个,基于所述第一cfDNA测序读段来确定定量等位基因分数(AF)量度,以分别生成第一AF组、第二AF组和第三AF组;

基于所述第一AF组生成第一频率分布,并且基于所述第二AF组生成第二频率分布,其中在所述第一频率分布和所述第二频率分布之间不存在重叠;

基于所述第一频率分布和所述第二频率分布来鉴定AF阈值,该AF阈值(i)不小于所述第一AF组中最大的定量AF量度,并且(ii)不大于所述第二AF组中最小的定量AF量度;以及

为所述第三箱的基因组基因座中的每一个分配最终分类,(A)如果所述基因组基因座具有不大于所述AF阈值的定量AF量度,则所述最终分类被假定为体细胞来源,或者(B)如果所述基因组区域具有不小于所述AF阈值的定量AF量度,则所述最终分类被假定为种系来源。

9.一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每

一个的体细胞来源或种系来源的方法,所述方法包括:

接收在第一时间点收集的来自所述受试者的所述cfDNA的测序信息,所述测序信息包括第一cfDNA测序读段;

提供来自所述多于一个基因组基因座的序列信息;

对所述多于一个基因组基因座的每一个进行分箱,其中所述分箱包括为所述多于一个基因组基因座中的每一个基因组基因座分配初始分类,所述初始分类选自由以下组成的组:

a)假定体细胞来源;

b)假定种系来源;或者

c)不确定来源;

从而生成包含假定体细胞来源的基因组基因座的第一箱、包含假定种系来源的基因组基因座的第二箱和包含不确定来源的基因组基因座的第三箱;

对于所述第一箱、所述第二箱和所述第三箱中的所述基因组区域中的每一个,基于所述第一cfDNA测序读段来确定定量等位基因分数(AF)量度,以分别生成第一AF组、第二AF组和第三AF组;

基于所述第一AF组生成第一频率分布,并且基于所述第二AF组生成第二频率分布,其中在所述第一频率分布和所述第二频率分布之间存在重叠;

基于所述第一频率分布和第二频率分布来识别第一AF阈值,该第一AF阈值为所述第一AF组中最大的定量AF量度;

基于所述第一频率分布和第二频率分布来识别第二AF阈值,该第二AF阈值为所述第二AF组中最小的定量AF量度;以及

为所述多于一个基因组基因座中的每一个分配最终分类,其中(A)如果所述基因组区域具有不大于所述第一AF阈值的定量AF量度,则所述最终分类被假定为体细胞来源,(B)如果所述基因组区域具有不小于所述第二AF阈值的定量AF量度,则所述最终分类被假定为种系来源,或者(C)如果所述基因组区域具有大于所述第一AF阈值且小于所述第二AF阈值的定量AF量度,则所述最终分类是不确定的。

10.一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:

接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的一组cfDNA测序读段;

基于所述cfDNA测序读段确定第一组定量等位基因分数(AF)量度,所述第一组AF量度包括对所述多于一个基因组基因座中的每一个AF量度;

提供第二组AF量度,所述第二组AF量度包括对一种或更多种已知体细胞变异中的每一种的AF量度;

比较来自所述第一组AF量度的基因组基因座的AF量度与来自所述第二组AF量度的AF量度;

如果来自基因组基因座的所述第一组AF量度的所述AF量度和来自所述第二组AF量度的所述AF量度之间存在10%或更小的差异,则将基因组基因座鉴定为体细胞来源。

无细胞DNA的体细胞来源或种系来源的鉴定

[0001] 本申请是申请日为2018年05月16日,申请号为201880047621.1,发明名称为“无细胞DNA的体细胞来源或种系来源的鉴定”的申请的分案申请。

[0002] 交叉引用

[0003] 本申请要求于2017年5月16日提交的美国临时申请第62/507,127号的权益,该美国临时申请通过引用以其整体并入本文。

背景

[0005] 受试者基因组和参考基因组(例如,GRCh38.p4)的比较通常会显示出约0.01%的碱基的差异(遗传变异)。种系中的遗传变异可以呈现为通过正常遗传或通过种系突变传递的SNP。变异可以纯合或杂合的形式存在。

[0006] 某些病理学状态,诸如癌症的特征在于病变细胞基因组与种系基因组相比的遗传变异。这些变异是由体细胞中的突变引起的,并被称为体细胞突变。

[0007] 携带体细胞突变的多核苷酸可以在无细胞DNA(cfDNA)中被检测到,在无细胞DNA中,它们与来自具有种系基因组的细胞的DNA混合在一起。在cfDNA中存在大的背景(种系)的情况下,没有计算机实施的方法能够自动地区分种系变异与体细胞突变。替代地,常规系统依赖于个体人类专家或专家团队(两种情况下都称为肿瘤讨论会(Tumor Board))的专业知识来区分体细胞突变与种系突变。

[0008] 如果不存在噪声和偏倚,种系变异将是等位基因分数为50%(在杂合(het)基因座的情况下)或100%(在纯合(homo)基因座的情况下)的变异。然而,在实践中,系统中噪声和偏倚的存在使这些明确的数字变得模糊。换言之,het或homo基因座无法被精确地检测为50%或100%,而是替代地处于het和homo类别的每一个的置信下限和置信上限范围之间。例如,het基因座可能处于40%至60%的范围内,而homo基因座可能处于98%至100%的范围内。

[0009] 概述

[0010] 从以下详述的描述,本公开内容的另外的方面和优势对本领域技术人员而言将变得明显,详细描述中仅示出和描述了本公开内容的说明性实施方案。如将会意识到的,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具有多个明显方面的修改,所有这些都不偏离本公开内容。相应地,附图和描述应被认为在本质上是说明性而而非限制性的。

[0011] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的cfDNA测序读段;基于所述cfDNA测序读段来确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;确定每一个所述AF量度的标准差(STDEV);提供STDEV阈值和AF阈值;确定每一个所述AF量度具有的STDEV高于还是低于所述STDEV阈值;确定每一个所述AF量度高于还是低于所述AF阈值;以及将STDEV低于所述STDEV阈值且AF量度低于所述AF

阈值的每一个基因座分类为体细胞来源。

[0012] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的cfDNA测序读段;基于所述cfDNA测序读段来确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;确定每一个所述AF量度的标准差(STDEV);提供STDEV阈值和AF阈值;确定每一个所述AF量度具有的STDEV高于还是低于所述STDEV阈值;确定每一个所述AF量度高于还是低于所述AF阈值;以及将STDEV低于所述STDEV阈值且AF量度高于所述AF阈值的每一个基因座分类为种系来源。

[0013] 在一些实施方案中,基因组基因座的AF量度低于所述STDEV阈值指示所述基因组基因座的低拷贝数变异(CNV)。

[0014] 在一些实施方案中,基因组基因座的AF量度高于所述STDEV阈值指示相关基因组基因座的高拷贝数变异(CNV)。

[0015] 在一些实施方案中,AF阈值根据经验确定。

[0016] 在一方面中,本公开内容提供了一种用于在来自患有癌症的受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:接收来自所述受试者在用癌症治疗剂治疗之前的第一时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第一组cfDNA测序读段;接收来自所述受试者在用癌症治疗剂治疗之后的第二时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第二组cfDNA测序读段;基于所述第一时间点的所述cfDNA测序读段并基于所述第二时间点的所述cfDNA测序读段,确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;比较来自所述第一时间点和来自所述第二时间点的所述AF量度;其中所述癌症对所述癌症治疗剂有响应;并且如果来自基因组基因座的AF量度在所述第一时间点和所述第二时间点之间降低,则将所述基因组基因座鉴定为体细胞来源。

[0017] 在一方面中,本公开内容提供了一种用于在来自患有癌症的受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:接收来自所述受试者在用癌症治疗剂治疗之前的第一时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第一组cfDNA测序读段;接收来自所述受试者在用癌症治疗剂治疗之后的第二时间点的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的第二组cfDNA测序读段;基于所述第一时间点的所述cfDNA测序读段并基于所述第二时间点的所述cfDNA测序读段,确定所述多于一个基因组基因座中的每一个的定量等位基因分数(AF)量度;比较来自所述第一时间点和来自所述第二时间点的所述AF量度;其中所述癌症对所述癌症治疗剂有响应;并且如果来自基因组基因座的AF量度在所述第一时间点和所述第二时间点之间未降低,则将所述基因组基因座鉴定为种系细胞来源。

[0018] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源或种系来源的方法,所述方法包括:接收在第一时间点收集的来自所述受试者的所述cfDNA的测序信息,所述测序信息包括第一

cfDNA测序读段;提供来自所述多于一个基因组基因座的序列信息;对所述多于一个基因组基因座中的每一个进行分箱,其中所述分箱包括为所述多于一个基因组基因座中的每一个基因组基因座分配初始分类,所述初始分类选自由以下组成的组:a)假定体细胞来源;b)假定种系来源;或者c)不确定来源;从而生成包含假定体细胞来源的基因组基因座的第一箱、包含假定种系来源的基因组基因座的第二箱和包含不确定来源的基因组基因座的第三箱;对于所述第一箱、所述第二箱和所述第三箱中的所述基因组区域中的每一个,基于所述第一cfDNA测序读段来确定定量等位基因分数(AF)量度,以分别生成第一AF组、第二AF组和第三AF组;基于所述第一AF组生成第一频率分布,并且基于所述第二AF组生成第二频率分布,其中在所述第一频率分布和所述第二频率分布之间不存在重叠;基于所述第一频率分布和第二频率分布来鉴定AF阈值,该AF阈值(i)不小于所述第一AF组中最大的定量AF量度,并且(ii)不大于所述第二AF组中最小的定量AF量度;以及为所述第三箱的基因组基因座中的每一个分配最终分类,(A)如果所述基因组基因座具有不大于所述AF阈值的定量AF量度,则该最终分类为假定体细胞来源,或者(B)如果所述基因组区域具有不小于所述AF阈值的定量AF量度,则为假定种系来源。

[0019] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源或种系来源的方法,所述方法包括:接收在第一时间点收集的来自所述受试者的所述cfDNA的测序信息,所述测序信息包括第一cfDNA测序读段;提供来自所述多于一个基因组基因座的序列信息;对所述多于一个基因组基因座中的每一个进行分箱,其中所述分箱包括为所述多于一个基因组基因座中的每一个基因组基因座分配初始分类,所述初始分类选自由以下组成的组:a)假定体细胞来源;b)假定种系来源;或c)不确定来源;从而生成包含假定体细胞来源的基因组基因座的第一箱、包含假定种系来源的基因组基因座的第二箱和包含不确定来源的基因组基因座的第三箱;对于所述第一箱、所述第二箱和所述第三箱中的所述基因组区域中的每一个,基于所述第一cfDNA测序读段来确定定量等位基因分数(AF)量度,以分别生成第一AF组、第二AF组和第三AF组;基于所述第一AF组生成第一频率分布,并且基于所述第二AF组生成第二频率分布,其中在所述第一频率分布和所述第二频率分布之间存在重叠;基于所述第一频率分布和第二频率分布来鉴定第一AF阈值,该第一AF阈值为所述第一AF组中最大的定量AF量度;基于所述第一频率分布和第二频率分布来鉴定第二AF阈值,该第二AF阈值为所述第二AF组中最小的定量AF量度;以及为所述多于一个基因组基因座中的每一个分配最终分类,其中(A)如果所述基因组区域具有不大于所述第一AF阈值的定量AF量度,则所述最终分类被假定为体细胞来源,(B)如果所述基因组区域具有不小于所述第二AF阈值的定量AF量度,则所述最终分类被假定为种系来源,或者(C)如果所述基因组区域具有大于所述第一AF阈值且小于所述第二AF阈值的定量AF量度,则所述最终分类为不确定的。

[0020] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的一组cfDNA测序读段;确定第一组定量等位基因分数(AF)量度,所述第一组AF量度包括基于所述cfDNA测序读段对所述多于一个基因组基因座中的每一个的AF量度;提供第二组AF量度,所述第二组AF量度包括一种或更多种已知体细胞变异中的每一种的AF量度;比较来自所述

第一组AF量度的基因组基因座的AF量度与来自所述第二组AF量度的AF量度;如果来自基因组基因座的所述第一组AF量度的所述AF量度和来自所述第二组AF量度的所述AF量度之间存在10%或更小的差异,则将基因组基因座鉴定为体细胞来源。

[0021] 在一些实施方案中,所述第二组AF量度包括来自基于所述cfDNA测序读段对第二多于一个基因组基因座的AF量度。

[0022] 在一些实施方案中,所述第二组AF量度包括对来自多于一个对照受试者的cfDNA的多于一个基因组基因座的AF量度。

[0023] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的一组cfDNA测序读段;确定第一组定量等位基因分数(AF)量度,所述第一组AF量度包括基于所述cfDNA测序读段对所述多于一个基因组基因座中的每一个的AF量度;提供第二组AF量度,所述第二组AF量度包括对一种或更多种已知体细胞变异中的每一种的AF量度;比较来自所述第一组AF量度的基因组基因座的AF量度与来自所述第二组AF量度的AF量度;如果来自基因组基因座的所述第一组AF量度的所述AF量度和来自所述第二组AF量度的所述AF量度之间存在大于10%的差异,则将基因组基因座鉴定为种系来源。

[0024] 在一些实施方案中,所述第二组AF量度包括基于所述cfDNA测序读段对来自第二多于一个基因组基因座的AF量度。

[0025] 在一些实施方案中,所述第二组AF量度包括对来自多于一个对照受试者的cfDNA的多于一个基因组基因座的AF量度。

[0026] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的种系来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的一组cfDNA测序读段;确定第一组定量等位基因分数(AF)量度,所述第一组AF量度包括基于所述cfDNA测序读段对所述多于一个基因组基因座中的每一个的AF量度;提供第二组AF量度,所述第二组AF量度包括对一种或更多种已知种系变异中的每一种的AF量度;比较来自所述第一组AF量度的基因组基因座的AF量度与来自所述第二组AF量度的AF量度;如果来自基因组基因座的所述第一组AF量度的所述AF量度和来自所述第二组AF量度的所述AF量度之间存在10%或更小的差异,则将基因组基因座鉴定为种系来源。

[0027] 在一些实施方案中,所述第二组AF量度包括基于所述cfDNA测序读段对来自第二多于一个基因组基因座的AF量度。

[0028] 在一些实施方案中,所述第二组AF量度包括对来自多于一个对照受试者的cfDNA的多于一个基因组基因座的AF量度。

[0029] 在一方面中,本公开内容提供了一种用于在来自受试者的无细胞DNA(cfDNA)中鉴定多于一个基因组基因座中的每一个的体细胞来源的方法,所述方法包括:接收来自所述受试者的所述cfDNA的测序信息,所述测序信息包括来自所述多于一个基因组基因座的一组cfDNA测序读段;确定第一组定量等位基因分数(AF)量度,所述第一组AF量度包括基于所述cfDNA测序读段对所述多于一个基因组基因座中的每一个的AF量度;提供第二组AF量度,所述第二组AF量度包括对一种或更多种已知种系变异中的每一种的AF量度;比较来自所述

第一组AF量度的基因组基因座的AF量度与来自所述第二组AF量度的AF量度;如果来自基因组基因座的所述第一组AF量度的所述AF量度和来自所述第二组AF量度的所述AF量度之间存在大于10%的差异,则将基因组基因座鉴定为体细胞来源。

[0030] 在一些实施方案中,所述第二组AF量度包括基于所述cfDNA测序读段对来自第二多于一个基因组基因座的AF量度。

[0031] 在一些实施方案中,所述第二组AF量度包括对来自多于一个对照受试者的cfDNA的多于一个基因组基因座的AF量度。

[0032] 在一些实施方案中,一个或更多个所述基因组基因座是BRCA基因中的基因座。

[0033] 在一方面中,本公开内容提供了一种方法,所述方法包括:a)提供一组cfDNA分子的序列读段,其中所述序列读段映射至参考基因组的选择的基因组区域(例如,基因、外显子、内含子、基因的一部分(例如,至少100个核苷酸、至少500个核苷酸或至少1000个核苷酸));b)确定所述基因组区域中包含多于一个遗传变异(例如,不同于参考序列的核苷酸)的组的等位基因频率,其中所述组包括感兴趣的变异;c)确定所述组中所述遗传变异的所述等位基因频率的变异性量度(例如,标准差或方差);d)提供变异性量度阈值和等位基因频率阈值;e)确定变异性量度是否低于变异性阈值;以及f)如果所述变异性量度低于所述变异性阈值:(i)如果感兴趣的变异的等位基因频率高于所述等位基因频率阈值,则判定所述感兴趣的变异具有种系来源,并且(ii)如果感兴趣的变异的等位基因频率低于所述等位基因频率阈值,则判定所述感兴趣的变异具有体细胞来源。

[0034] 通过引用并入

[0035] 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,其程度如同每一个单独的出版物、专利或专利申请被具体且单独地指明通过引用并入的相同程度。在通过引用并入的出版物和专利或专利申请与本说明书中包含的公开内容相矛盾的程度上,意图本说明取代和/或优先于任何这种矛盾的材料。

附图简述

[0037] 本公开内容的新的特征特别地在所附权利要求中阐述。通过参考以下详述及其附图将会获得对本公开内容的特征和优势的更好的理解,详述阐述了利用了本公开内容的原理的说明性实施方案,在附图中:

[0038] 图1示出了被编程或以其他方式配置为实现本文提供的方法的计算机系统。

[0039] 图2A示出了种系T790M突变(201-黑色点)存在的浓度虽然具有与体细胞T790M突变(202-灰色点)相似的浓度,但等位基因分数(AF)较高。

[0040] 图2B示出了在四名治疗中(on treatment)的患者中,体细胞EGFR突变的浓度降低,而种系EGFR T790M的浓度保持恒定(203代表EGFR驱动突变,而204代表EGFR T790M)。

[0041] 图2C示出了950个病例的血浆NGS结果中EGFR T790M的AF分布(下图)包括对于EGFR驱动突变也可见的体细胞峰(上图),以及还有对于普通SNP(EGFR Q787,中图)更清楚地可见的杂合峰(箭头)。

[0042] 图3A示出了来自三个初始病例的治疗前和治疗中的血浆样本的血浆NGS,所有这些样品对种系EGFR T790M都呈阳性。在检测出的所有编码和非编码变异中,对应于纯合、杂合和肿瘤来源变异的预期AF的三组变异是明显的。肿瘤来源组中的变异对治疗有响应,而

纯合组和杂合组中的变异保持相对恒定的AF。

[0043] 图3B示出了来自总计105个病例中另外102个病例的血浆NGS结果。对于在105个病例中检测出的所有编码和非编码变异观察到三峰分布,这些峰接近0%(可能肿瘤来源)、49%(可能杂合)和100%(可能纯合)。

[0044] 图3C示出了对于错义变异和无义变异,存在低AF处的富集(箭头),其中将预期发现肿瘤来源的变异。相比之下,同义变异,可能反映良性种系多态性,在富集在50%和100%AF附近。

[0045] 图4A示出了在105个EGFR突变阳性的病例的血浆NGS上发现的所有变异的AF,EGFR驱动突变AF(401-黑色点)呈递增顺序,并且示出了常见的EGFR SNP(402-较大的灰色点)。

[0046] 图4B示出了对于25%和75%之间的变异AF,病例平均值和群体平均值之间的标准差和绝对差随着EGFR驱动突变AF的增加而增加。

[0047] 图5示出了在具有低拷贝数变异的病例中,杂合和肿瘤来源的编码变异之间的区别。图5还示出了当存在较低的拷贝数变异时,可以目视区分具有种系EGFR T790M(501)的哪些病例可能是种系的。

[0048] 图6A示出了NGS结果,该NGS结果显示来自31,414名独特癌症患者数据库的48名(0.15%)癌症患者被发现携带种系EGFR T790M突变,并且在这些患者中主要诊断到非鳞状非小细胞肺癌(NSCLC)。

[0049] 图6B示出了与参考群组中种系EGFR T790M的群体患病率(0.008%)相比,患有非鳞状NSCLC的受试者中存在更高的患病率(0.34%),但患有其他癌症的受试者中的患病率不高(0.03%,p=0.06),这表明种系EGFR T790M是肺癌的风险变异。

[0050] 图7示出了在三个时间点检测到的编码变异和非编码变异的AF图(时间2的TP53突变AF被估算为0%)。在包括常见SNP(EGFR Q787Q)的变异带中观察到EGFR T790M突变,怀疑是偶然检测到的种系EGFR T790M。

[0051] 图8A示出了即使存在异常值,也可以拟合对于标准差的曲线。

[0052] 图8B示出了即使存在异常值,也可以拟合对于平均值的曲线。

[0053] 图9示出了在被指定为具有低拷贝数变异和高的EGFR T790M的AF的11个血浆NGS病例(群组A)中(左图),所有11个均被确认为种系(100%阳性预测值)。在具有高拷贝数变异和高的EGFR T790M的AF的10个病例(群组B)中,1个病例为种系T790M突变阳性。

[0054] 详述

[0055] 本文使用的章节标题仅用于组织目的,并且不应被解释为以任何方式限制所描述的主题。

[0056] 在这些对多种实施方案的详细描述中,出于说明的目的,阐述了许多具体细节以提供对所公开的实施方案的透彻理解。然而,本领域技术人员将理解,可以在具有或没有这些具体细节的情况下实践这些不同的实施方案。在其他情况下,结构和装置以框图形式示出。此外,本领域技术人员可以容易地理解,方法被提供及进行的具体顺序是说明性的,并且设想顺序可以不同并且仍然保持在本文公开的多种实施方案的精神和范围内。

[0057] 出于任何目的,本申请中引用的所有文献和类似材料,包括但不限于专利、专利申请、文章、书籍、论文和互联网网页,均明确地通过引用以其整体并入。除非另外描述,否则本文使用的所有技术术语和科学术语具有如本文描述的多种实施方案所属领域的普通技

术人员通常理解的含义。

[0058] 应当理解的是,在本教导中讨论的温度、浓度、时间、碱基数、覆盖度等之前存在暗示的“约”,使得轻微和非实质性不同的的等效值在本教导的范围内。在本申请中,除非另外具体说明,单数的使用包括复数。此外,“包含(comprise)”、“包含(comprises)”、“包含(comprising)”、“含有(contain)”、“含有(contains)”、“含有(containing)”、“包括(include)”、“包括(includes)”、“包括(including)”的使用意图是非性限制性的。应理解,前述一般描述和以下详细描述两者仅是示例性和说明性的,而不限制本教导。

[0059] 如本文使用的,“一(a)”或“一(an)”也可以指“至少一”或“一或更多”。此外,“或”的使用是包含性的,使得措辞“A或B”当“A”为真、“B”为真、或“A”和“B”两者都为真时为真。

[0060] 此外,除非上下文另外要求,否则单数术语应包括复数,并且复数术语应包括单数。通常,本文描述的与细胞和组织培养、分子生物学和蛋白及寡核苷酸或多核苷酸化学以及杂交关联使用的命名和技术是本领域熟知且常用的那些。例如,标准技术被用于核酸纯化和制备、化学分析、重组核酸及寡核苷酸合成。酶促反应和纯化技术根据生产商的说明或如本领域通常完成的或如本文描述的来进行。本文描述的技术和程序通常根据本领域熟知的常规方法以及如多种一般和更具体的参考文献中描述的进行,所述参考文献在本说明书全文中被引用及讨论。参见例如,Sambrook等人,Molecular Cloning:ALaboratory Manual(第3版,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.2000)。结合本文描述的实验室程序和技术使用的术语是本领域熟知且常用的那些。

[0061] “系统”阐述了一组构成整体的真实或抽象的组成部分,其中每个组成部分与整体中的至少一个其他组成部分相互作用或相关。

[0062] “生物分子”可以指由生物有机体产生的任何分子,包括大的聚合分子,诸如蛋白、多糖、脂质和核酸(DNA和RNA),以及小分子,诸如初级代谢物、次级代谢物和其他天然产物。

[0063] 如本文使用的,术语“测序”是指用于确定生物分子,例如核酸,诸如DNA或RNA的序列的若干种技术中的任一种。示例性测序方法包括但不限于靶向测序、单分子实时测序、外显子测序、基于电子显微术的测序、panel测序、晶体管介导的测序、直接测序、随机鸟枪法测序、Sanger双脱氧终止测序、全基因组测序、杂交测序、焦磷酸测序、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、大规模并行信号测序(massively parallel signature sequencing)、乳液PCR、低变性温度共扩增PCR(COLD-PCR)、多重PCR、可逆染料终止子测序、配对末端测序、近末端测序(near-term sequencing)、外切核酸酶测序、连接测序、短读段测序、单分子测序、合成测序、实时测序、反向终止子测序、纳米孔测序、454测序、Solexa基因组分析仪测序、SOLiD™测序、MS-PET测序及其组合。在一些实施方案中,测序可以通过基因分析仪进行,诸如,例如可从Illumina或Applied Biosystems商业上获得的基因分析仪。

[0064] 措辞“下一代测序”或NGS是指与传统的基于Sanger和毛细管电泳的方法相比具有增加的通量的测序技术,例如,具有一次产生数十万个相对较小的序列读段的能力。下一代测序技术的一些实例包括但不限于合成测序、连接测序和杂交测序。

[0065] 措辞“测序运行”是指为确定与至少一种生物分子(例如,核酸分子,诸如DNA或RNA)相关的一些信息而进行的测序实验的任何步骤或部分。

[0066] DNA(脱氧核糖核酸)是由四种类型的核苷酸:腺嘌呤(A)、胸腺嘧啶(T)、胞嘧啶(C)

和鸟嘌呤(G)构成的核苷酸链。RNA(核糖核酸)是由四种类型的核苷酸:A、尿嘧啶(U)、G和C构成的核苷酸链。特定核苷酸对以互补方式彼此特异性结合(称为互补碱基配对)。在DNA中,腺嘌呤(A)与胸腺嘧啶(T)配对并且胞嘧啶(C)与鸟嘌呤(G)配对。在RNA中,腺嘌呤(A)与尿嘧啶(U)配对并且胞嘧啶(C)与鸟嘌呤(G)配对。当第一核酸链与由与该第一链中的核苷酸互补的核苷酸构成的第二核酸链结合时,两条链结合形成双链。如本文使用的,"核酸测序数据"、"核酸测序信息"、"核酸序列"、"核苷酸序列"、"基因组序列"、"遗传序列"或"片段序列"或"核酸测序读段"表示指示核酸诸如DNA或RNA的分子(例如,全基因组、全转录组、外显子组、寡核苷酸、多核苷酸或片段)中核苷酸碱基(例如,腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶或尿嘧啶)顺序的任何信息或数据。应当理解,本教导设想了使用所有可用的各种技术、平台或科术(technologies)获得的序列信息,包括但不限于:毛细管电泳、微阵列、基于连接的系统、基于聚合酶的系统、基于杂交的系统、直接或间接核苷酸鉴定系统、焦磷酸测序、基于离子或pH的检测系统以及基于电子信号的系统。

[0067] "多核苷酸"、"核酸"或"寡核苷酸"是指核苷(包括脱氧核糖核苷、核糖核苷或其类似物)通过核苷间键连接的线性聚合物。通常,多核苷酸包含至少三个核苷。寡核酸的尺寸范围通常从几个单体单元例如3-4个到几百个单体单元。除非另外注明,否则每当多核苷酸以字母序列诸如"ATGCCTG"表示时,应该理解,该核苷酸从左到右是5'→3'的顺序,并且"A"表示脱氧腺苷,"C"表示脱氧胞苷,"G"表示脱氧鸟苷,并且"T"表示胸苷。字母A、C、G和T可以用于指碱基本身、包含碱基的核苷或核苷酸,这是本领域的标准。

[0068] 术语"衔接子(adaptor)"、"衔接子(adapter)"和"标签"在本说明书全文中作为同义词使用。可以通过任何方法,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。

[0069] 如本文使用的,常见变异具有至少5%的GMAF(次要等位基因总频率),而低频变异具有约0.1%-5%的GMAF,并且罕见变异具有0.5%或更少的GMAF,其中GMAF是给定群体中最不常见的等位基因出现的频率。

[0070] 如本文使用的,"基因型"是指一条或更多条种系染色体上的遗传基因座处的等位基因同一性。这包括完全基因型(所有染色体上的等位基因相同性)、部分基因型(至少一条染色体上的等位基因相同性)和空白基因型(null genotype)(在一条或更多条或所有染色体上不存在等位基因),包括确定基因座处的纯合性或杂合性("等位基因分配(allelic designation)")。

[0071] 如本文使用的,体细胞变异表示来源是癌性组织。如本文使用的,遗传变异的体细胞来源是指先在体细胞中而不是种系中发生的遗传变异。这种体细胞变异与种系变异相对,种系变异以正常细胞为来源。体细胞变异可以通过有丝分裂传递给子细胞。这可以导致生物体的一组细胞与其余细胞具有遗传差异。另外,由于该变异不发在种系细胞中,因此该变异不会被后代生物体遗传。

[0072] 通常在种系变异的情况下,SNP可以指单核苷酸多态性或群体变异(variation in the population),而SNV可以指单核苷酸变异,而SSNV可以指体细胞单核苷酸变异(通常在癌症相关变异的情况下使用)。对于个体,对在体细胞(癌性)cfDNA和种系(正常)cfDNA两者中检测到的变异使用术语SNV。

[0073] CNV可以指拷贝数变异(基因水平的拷贝数突变,通常由复制事件引起)。

[0074] 来自患有癌症的受试者的无细胞DNA(cfDNA)包括来自携带种系基因组的细胞(例如,来自“健康细胞”)的DNA(“种系DNA”)和来自通常携带体细胞突变的癌细胞的DNA(“癌症DNA”)两者。无细胞DNA样品中种系DNA和癌症DNA的相对量取决于癌症的发展程度。在早期时期,只有少量的DNA是癌症DNA。例如,这可能占总DNA的约1%-5%。因此,检测到少量(例如,样品中cfDNA的约1%-5%)的遗传变异,诸如次要等位基因,可以指示体细胞突变,并因此指示癌症DNA的存在。然而,随着疾病进展和肿瘤扩大,无细胞DNA样品中的癌症DNA的量可能显著增加,例如增加到多于总无细胞DNA的25%。当携带遗传变异的DNA分子的百分比达到高水平时,变异是代表源自癌细胞的体细胞突变,还是代表种系DNA中的杂合性可能变得不明确。

[0075] 血浆cfDNA的基因组分析可以作为用于基因组发现及用于辅助精准癌症药物递送的工具,但是癌症来源的DNA向血浆中的脱落可以是高度可变的,并且取决于癌症阶段、转移扩散的程度以及癌症是在响应还是在进展。另外,体细胞基因组改变的血浆水平可以响应疗法而高度动态化,有时变得在两周内检测不到。因此,在许多患者中,血浆cfDNA的大部分是种系DNA,主要从良性造血细胞或内皮细胞脱落。本公开内容提供了一种方法,该方法可以在cfDNA下一代测序(NGS)谱中区分种系变异与癌症来源的体细胞变异,因此提供用于疗法选择的肿瘤基因分型以及用于通过单一测定评估遗传风险的种系表征。

[0076] 本公开内容提供的方法具有许多应用。血浆NGS有时会鉴定到癌症患者中的偶然性种系突变,这对患者及其家人具有潜在的临床意义。本文描述的某些种系EGFR突变是被认为与遗传性癌症风险相关的罕见风险等位基因,并且本文描述的方法可以应用于其他种系突变。其他癌症相关基因(例如TP53或BRCA1/2和错配修复基因)也可以通过血浆NGS来测序,并且这些基因中的种系突变可能具有深远的临床意义。本公开内容描述了使用生物信息学算法预测疑似种系突变的存在,确认的种系突变具有高阳性预测值,这是测试中规则(rule-in test)的重要诊断性能特征。

[0077] 分辨血浆cfDNA中的种系变异和体细胞变异还可能影响对癌症生物学的理解。通过肿瘤NGS,可能难以确定癌基因中未知意义的变异是否代表潜在的驱动突变或种系多态性。在没有高拷贝数变异的血浆NGS的情况下,本公开内容允许用单一(血液)样品区分这两种类型的基因组改变,从而减少种系多态性在治疗上被错误靶向的风险。另外,在使用连续血浆基因分型(serial plasma genotyping)随时间推移监测对治疗的响应和耐药性的情况下,区分血浆cfDNA中的种系变异和体细胞变异的能力可以使准确跟踪肿瘤DNA水平变得更容易。

[0078] 肿瘤突变负荷(TMB)是新兴的生物标志物,用于了解对免疫检查点抑制剂的敏感性和耐药性。癌症中更多的突变可以导致更多的刺激免疫的细胞表面新抗原。然而,使用肿瘤NGS可能难以计算突变负荷,因为种系多态性可能被误认为潜在抗原性体细胞突变。本公开内容提供了克服这一挑战的能力,允许在生物信息学上区分种系变异和体细胞变异,并且更明确地鉴定抗原性体细胞变异,从而减少种系多态性被误认为潜在抗原性体细胞突变的几率。

[0079] 受试者的种系DNA在任何遗传基因座处都可以是纯合的或杂合的。基因座处的测量可以采取等位基因分数(AF)的形式,它测量在样品中观察到等位基因的频率。出于多种原因(包括,例如,在DNA测序中的错误),在由非癌症受试者的cfDNA产生的一组序列读段

中,映射至受试者在一个遗传基因座处对于其为纯合的等位基因形式(例如,SNV)的读段计数可能不严格地是100%。类似地,映射至受试者在一个遗传基因座处对于其为杂合的等位基因形式的读段计数可能不严格地是正好50%。如果个体的种系中的遗传变异是纯合的(与参考基因组中的等位基因不匹配),则携带该遗传变异的碱基判定的百分比通常将接近100%判定,但不总是与等同100%。类似地,如果个体的种系中的遗传变异是杂合的,则携带该遗传变异的碱基判定的百分比通常将接近50%,但是可以例如从30%至70%变化。在该范围内的测量与在该基因座处的杂合性一致。然而,测量可能做出不明确的确定。在这种情况下,人们可以用一定水平的置信度或概率将一个基因座处的基因型判定为杂合的或纯合的。

[0080] 因此,如果受试者患有癌症并且在一个基因座处的遗传变异被测量为在与杂合性一致的范围内,则该变异是由体细胞突变导致的置信度与在纯合性和杂合性之间的范围内的测量值相比可能降低。例如,在5%至20%范围内的测量值可以指示基因座包含的遗传变异的量过高而不能用纯合性来解释,并且过低而不能用杂合性来解释。因此,该测量可能是体细胞突变的结果。相比之下,约40%的测量可以指示杂合性,或者它可以指示大量DNA包含体细胞突变(例如,如果该体细胞突变已经导致肿瘤,该肿瘤已经对样品贡献了相对大量的DNA)。

[0081] 本公开内容尤其提供了确定在无细胞DNA样品中检测到的遗传变异的来源更可能是种系(例如,代表种系的杂合性)还是体细胞(例如,来自癌症)的方法。具体地,本公开内容提供了利用AF来进行这种确定的方法。

[0082] 在一些实施方案中,本公开内容提供了用于在来自受试者的无细胞DNA(cfDNA)中用一个或更多个阈值鉴定多于一个基因组基基因座的每一个的种系来源或体细胞来源的方法,该阈值可以用于确定基因座处的变异是种系来源还是体细胞来源。可以使用的一个示例性阈值是标准差(STDEV)阈值。例如,技术人员在确定基因组基因座的定量等位基因分数(AF)后,可以确定AF量度的标准差。随着拷贝数变异(CNV)的增加,STDEV预期也会增加。因此,可以假设低STDEV具有低CNV,使得这些数据更容易处理。STDEV阈值可以用于将高CNV与低CNV分开,增加了该方法的预测能力。AF的第二阈值可以与CNV阈值组合使用或作为其替代。因为种系来源的变异中的AF量度被预期高于体细胞变异中的AF量度,高于AF阈值的AF量度可以被分类为种系来源的,而低于AF阈值的AF量度可以分类为体细胞来源的。示例性AF阈值包括但不限于约10%、约11%、约12%、约13%、约14%、约15%、约16%、约17%、约18%、约19%、约20%、约21%、约22%、约23%、约24%、约25%、约26%、约27%、约28%、约29%、约30%、约31%、约32%、约33%、约34%及约35%。在一些实施方案中,根据经验确定AF阈值。

[0083] 本文描述的方法也可以用来基于对治疗的响应确定来自cfDNA的基因座是种系来源还是体细胞来源。例如,可以从患有癌症的受试者获得用癌症治疗剂治疗之前和之后的序列信息。如果癌症对癌症治疗剂有响应,并且基因座处的癌症相关变异是体细胞来源,则其AF将降低。因此,可以测量治疗之前和之后的AF,并将这些值进行比较,以确定体细胞来源或种系来源。如果AF量度降低,则变异可以被鉴定为是体细胞来源的。如果AF量度没有降低(即,它保持不变或增加),则变异可以被鉴定为是种系来源的。

[0084] 在一些实施方案中,本文描述的方法可以用于通过根据假定体细胞来源、假定种

系来源或不确定来源的初始分类将基因座分箱来确定来自cfDNA的基因座是种系来源还是体细胞来源。然后可以确定每个箱(bin)的基因座的定量AF量度,以生成AF组,随后使用AF组来生成假定体细胞来源或假定种系来源的基因座的频率分布。该分布可以用于设置AF阈值,例如,阈值不小于“假定体细胞”AF组中最大的定量AF量度,并且不大于“假定种系”AF组中最小的定量AF量度。因此,可以基于基因座的AF是高于AF阈值(并因此是种系的)还是低于AF阈值(并因此是体细胞的),将“不确定来源”的基因座分类为种系的或体细胞的。可选地,当“假定体细胞”AF量度的频率分布和“假定种系”AF量度的频率分布之间存在重叠时,可以确定两个阈值,使得第一AF阈值是“假定体细胞”AF组中最大的定量AF量度,而第二AF阈值是“假定种系”AF组中最小的定量AF量度。在这样的实施方案中,具有低于“假定体细胞”阈值的定量AF量度的基因座被分类为体细胞的,具有高于“假定种系”阈值的定量AF量度的基因座被分类为种系的,并且具有两个阈值之间的定量AF量度的基因座被分类为不明确的。然后,这些不明确的基因座例如,基于它们的AF量度在频率分布中的位置可以被指定它们是种系来源还是体细胞来源的概率。

[0085] 在一些实施方案中,本公开内容提供了用于通过将来自样品中基因组基因座的AF量度与来自已知体细胞变异或种系变异的一个或更多个AF量度进行比较来鉴定cfDNA中基因组基因座的体细胞来源或种系来源的方法。例如,如果使用来自已知体细胞变异的AF量度,具有相似AF量度(例如,在30%以内、在25%以内、在20%以内、在15%以内、在10%以内、在9%以内、在8%以内、在7%以内、在6%以内、在5%以内、在4%以内、在3%以内、在2%以内、在1%以内、或在0.1%以内)的基因组基因座可以被分类为体细胞来源,而具有不相似AF量度(例如,不在30%以内、不在25%以内、不在20%以内、不在15%以内、不在10%以内、不在9%以内、不在8%以内、不在7%以内、不在6%以内、不在5%以内、不在4%以内、不在3%以内、不在2%以内、不在1%以内、或不在0.1%以内)的基因组基因座可以被分类为种系来源。同样,如果使用来自已知种系变异的量度,具有相似AF量度(例如,在30%以内、在25%以内、在20%以内、在15%以内、在10%以内、在9%以内、在8%以内、在7%以内、在6%以内、在5%以内、在4%以内、在3%以内、在2%以内、在1%以内、或在0.1%以内)的基因组基因座可以被分类为种系来源,而具有不相似AF量度(例如,不在30%以内、不在25%以内、不在20%以内、不在15%以内、不在10%以内、不在9%以内、不在8%以内、不在7%以内、不在6%以内、不在5%以内、不在4%以内、不在3%以内、不在2%以内、不在1%以内、或不在0.1%以内)的基因组基因座可以被分类为体细胞来源。来自已知体细胞变异或种系变异的AF量度可以来自被测试的受试者或多于一个对照受试者的cfDNA测序读段。

[0086] 在一些实施方案中,对来自受试者的无细胞DNA进行测序,并对一种或更多种遗传变异进行检测及定量。例如,确定映射至包含变异的基因座的总读段的相对量(读段计数的数目)。如果相对量与纯合性一致,则人们可以具有变异存在于种系中的高置信度。这样的量可以是,例如,高于95%、高于96%、高于97%、高于98%、高于99%或100%。可以将该判定与确定的基因型比较以进行确认。

[0087] 如果相对量与基因座处的纯合基因型或杂合基因型不一致,则人们可以具有高置信度变异是体细胞突变的结果并且不存在于种系中。这样的量可以是,例如,低于30%、低于25%、低于20%、低于15%、低于10%、低于9%、低于8%、低于7%、低于6%、低于5%、低于4%、低于3%、低于2%、或低于1%。同样,可以将该判定与确定的基因型比较以进行确

认。

[0088] 可选地,相对量可以与基因座处的杂合性一致。这样的量可以是,例如,在30%和70%之间,例如,在40%和60%之间、在45%和55%之间、在46%和54%之间、在47%和53%之间、在48%和52%之间、或者在49%和51%之间。在一些实施方案中,受试者基因座处的可能的种系基因型(例如,如从gDNA获得的)被确定。在一些实施方案中,将基因型与在无细胞DNA中发现的变异的身份进行比较。在某些实施方案中,如果基因型是纯合的,则人们可以以高置信度推断变异代表体细胞突变,并且最可能是大量的。如果基因型被确定为杂合的,并且变异与杂合等位基因中的一种一致,则人们可以推断变异不是体细胞突变,而是代表种系基因型中的杂合性。

[0089] 在一些实施方案中,纯合基因型可以被以高置信度排除,但是杂合基因型不能被以高置信度确定,导致可能的不明确的基因型。例如,可以在基因组DNA上测量到变异处于范围的最末端,例如在30%处。在这种情况下,人们可能不能以高置信度确定在cfDNA中检测到的变异的量是否更有可能代表体细胞突变或种系杂合性。因此当样品中由于例如肿瘤细胞的快速生长而存在大量包含体细胞突变的DNA时,可能会出现这样的测量。应当注意,在任何测量的水平上,在基因组DNA中检测到的变异都有一定概率不代表杂合性。然而,种系中30%和70%之间的变异检测最有可能代表杂合性,并且在cfDNA中检测到的变异可以据此被测量。

[0090] 在这种情况下,可以以贝叶斯方式使用其他信息来增加或减少cfDNA中的变异代表体细胞突变或种系中的杂合性的概率。例如,人口研究可以例如,基于遗传祖先指示不同群体的种系中变异的流行程度。因此,例如,如果个体中的杂合基因型的判定具有低置信度,并且在与受试者共有遗传祖先的人中发现高发生率的变异,则人们可以以较高置信度确定该人确实是杂合的,并且cfDNA中的变异不代表体细胞突变。相反,如果在与受试者共有遗传祖先的人中仅发现非常低发生率的变异,则人们可以以较高置信度确定该人不是杂合的,并且cfDNA中的变异代表体细胞突变。

[0091] 本公开内容设想了确定量(例如,读段计数)是与杂合基因型一致还是不一致的若干方法。在一些实施方案中,使用截止值。例如,对于基因座处的特定遗传变异可以设置总读段计数的30%的截止值。在一些实施方案中,低于截止量的值被假定代表体细胞突变。在一些实施方案中,高于截止量的值,并且通常低于纯合性的截止值,可以被假定与杂合性一致,并且从而在将变异判定为体细胞突变之前需要进一步分析。

[0092] 在一些实施方案中,使用概率函数(例如,贝叶斯函数)计算量代表杂合性的概率。高于特定水平的概率可以触发基因型比较。

[0093] 在一些实施方案中,基因型的确定作为分析的常规部分来进行。在一些实施方案中,仅当变异的丰度与杂合性的解释一致时,基因型的确定才被确定。

[0094] 在一些实施方案中,本公开内容的方法减少了可能比可靠地检测出与癌症相关的从头基因组改变所需要的数量级高几个数量级的错误率和偏倚。在一些实施方案中,该方法首先通过收集体液样品作为遗传物质的来源(血液、唾液、汗液以及其他),随后对物质进行测序,来捕获遗传信息。例如,样品中的多核苷酸可以被测序,产生多于一个序列读段。包含多核苷酸的样品中的肿瘤负荷可以被估计为携带变异的序列读段针对于从该样品生成的序列读段的总数目的相对数目。此外,在拷贝数变异的情况下,肿瘤负荷可以被估计为测

试基因座与对照基因座处的序列读段总数目的相对过量(在基因重复的情况下)或相对缺乏(在基因消除的情况下)。因此,例如,运行可以产生映射至癌基因基因座的1000个读段,其中900个对应于野生型,并且100个对应于癌症突变,指示该基因处的拷贝数变异。接下来,处理遗传信息并鉴定遗传变异。遗传变异包括序列变异、拷贝数变异和核苷酸修饰变异。序列变异是遗传核苷酸序列中的变异。拷贝数变异是基因组的一部分的拷贝数与野生型的偏差。遗传变异包括例如单核苷酸变异(SNP)、插入、缺失、倒位、颠换、易位、基因融合、染色体融合、基因截短、拷贝数变异(例如非整倍性、部分非整倍性、多倍性、基因扩增)、核酸化学修饰的异常改变、表观遗传模式的异常改变和核酸甲基化的异常改变。然后该过程确定包含遗传物质的样品中的遗传变异的频率。由于该过程为有噪声的,该过程将信息与噪声分开。

[0095] 测序方法具有错误率。例如,Illumina的mySeq系统可以产生低个位数的错误率百分比。因此,对于映射至基因座的1000个序列读段,人们可以预期约50个读段(约5%)包括错误。某些方法,诸如在WO 2014/149134(Talasaz和Eltoukhy)中描述的那些可以显著减少错误率。错误产生噪声,噪声可以使样品中以低水平存在的来自癌症的信号模糊。因此,如果样品具有在测序系统错误率附近的水平(例如,约0.1%-5%)的肿瘤负荷,则可能难以将对应于由于癌症引起的遗传变异的信号与由噪声引起的信号区分开。

[0096] 甚至在噪声的存在下,也可以通过分析遗传变异做出癌症诊断。分析可以基于序列变异的频率或CNV的水平,并且可以建立用于检测在噪声范围内的遗传变异的诊断置信度指示或水平。接下来,该过程增加了诊断置信度。这可以通过以下做出:使用多于一个测量来增加诊断的置信度,或者可选地,使用在多于一个时间点的测量来确定癌症是否进展、缓解或稳定。诊断置信度可以用来鉴定疾病状态。例如,从受试者取得的无细胞多核苷酸可以包括源自正常细胞的多核苷酸,以及源自病变细胞诸如癌细胞的多核苷酸。来自癌细胞的多核苷酸可以携带遗传变异,诸如体细胞突变和拷贝数变异。当来自受试者的样品的无细胞多核苷酸被测序时,这些癌症多核苷酸被检测为序列变异或被检测为拷贝数变异。无细胞多核苷酸样品中肿瘤多核苷酸的相对量被称为“肿瘤负荷”。

[0097] 对参数的测量,无论它们是否在噪声范围内,均可以提供置信区间。随时间推移进行测试,人们可以通过比较随时间推移的置信区间来确定癌症是否进展、稳定或缓解。在置信区间不重叠的情况下,这指示疾病的方向。

[0098] 接下来,该过程生成遗传报告/诊断。该过程接收种系SNP和体细胞癌症突变,并标记体细胞癌症突变及生成与人类肿瘤讨论会分析类似的注释体细胞突变的报告,并提供待由实验室主管审查及批准的治疗选项。

[0099] 现在转到生成肿瘤讨论会建议的过程,在一些实施方案中,该系统使用来自cBioPortal SNV的GH2.7中的68个基因的数据,其中GH2.7是Guardant Health的组和2015年2月发布的相关测试过程(Guardant360测试)的组。用于肿瘤基因组学的cBioPortal(http://cbioportal.org)提供探索、可视化及分析多维癌症基因组学数据的网络资源。该门户网站将来自癌症组织和细胞系的分子谱数据简化为易于理解的遗传学、表观遗传学、基因表达和蛋白质组学事件。查询界面与定制数据存储组合,使研究人员能够交互式地探索样品、基因和途径之间的遗传改变,并在基础数据可用时将这些与临床结果相联系。该门户网站提供了来自多个平台的基因水平数据的图形概述、网络可视化和分析、存活分析、以患者为中

心的查询和软件编程访问。该系统提供了变异水平判定以及样品水平判定,以确定主管是否应当深入审查测试。

[0100] 许多癌症可以使用本文描述的方法和系统来检测。癌细胞,如大部分细胞一样,可以通过更新率表征,其中旧细胞死亡并被较新的细胞代替。通常,与给定受试者的脉管系统相接触的死亡的细胞可以将DNA或DNA片段释放到血流中。在疾病不同时期中的癌细胞也是如此。根据疾病的时期,癌细胞还可以通过多种遗传畸变诸如拷贝数变异以及突变来表征。这种现象可以用于使用本文描述的方法和系统检测个体中癌症的存在或不存在。

[0101] 在一些实施方案中,本公开内容的方法可以用于诊断疾病或状况,诸如癌症或炎性状况。如本文使用的术语“诊断”是指技术人员能够估计和/或确定患者是否罹患特定疾病或状况的方法。在一些实施方案中,本公开内容的方法可以用于疾病或状况诸如癌症或炎性状况的预后。如本文使用的术语“预后”是指疾病或状况进展的可能性,包括疾病或状况的复发。在一些实施方案中,本公开内容的方法可以用于评估发展疾病或状况诸如癌症或炎性状况的风险。在一些实施方案中,本公开内容的方法可以用于评估治疗疾病或状况诸如癌症或炎性状况的功效。例如,本公开内容的方法可以在治疗患有疾病或状况的患者之前和之后(例如,在施用诸如化学治疗剂的药物之前和之后)使用。在一些实施方案中,本公开内容的方法可以用于监测疾病或状况诸如癌症或炎性状况的进展或减退。例如,本公开内容的方法可以在不同的时间点进行,以监测进展或减退。在一些实施方案中,本公开内容的方法可以用于鉴定用于改善或治疗疾病或状况诸如癌症或炎性状况的化合物。例如,本公开内容的方法可以在施用化合物之前和之后使用,以确定化合物是否改善或治疗了疾病。

[0102] 如本文使用的,“治疗”疾病或状况是指采取步骤以获得有益或期望的结果,包括临床结果。有益或期望的临床结果包括但不限于缓解或改善与疾病或状况相关的一种或更多种症状。如本文使用的,向受试者“施用(administering)”或“施用(administration of)”化合物或剂可以使用本领域技术人员已知的多种方法中的一种来进行。例如,化合物或剂可以静脉内、动脉内、皮内、肌肉内、腹膜内、静脉内、皮下、眼部、舌下、口服(通过摄入)、鼻内(通过吸入)、脊椎内、脑内和透皮(通过吸收,例如通过皮肤导管)施用。化合物或剂还可以通过可再充装或生物可降解的聚合物装置或其他装置(例如,贴剂和泵)或制剂适当地引入,所述装置或制剂提供化合物或剂的延长、缓慢或控制释放。施用还可以进行例如一次、多于一次和/或在一个或更多个延长的时间段内进行。在一些方面中,施用包括直接施用(包括自我施用)和间接施用(包括给药物开处方的行为)。例如,如本文使用的,指导患者自我施用药物或者由另一个人施用药物和/或向患者提供药物处方的医师正在向患者施用药物。在一些实施方案中,化合物或剂通过口服例如通过摄入施用至受试者,或者通过静脉内例如通过注射施用至受试者。在一些实施方案中,口服施用的化合物或剂处于延长释放或缓慢释放的制剂中,或者使用用于这种缓慢或延长释放的装置施用。

[0103] 在一些实施方案中,可以从处于患癌风险的受试者抽取血液并如本文描述地准备以生成无细胞多核苷酸群体。在一个实例中,这可以是无细胞DNA。本公开内容的系统和方法可以用于检测可存在于某些现有癌症中的突变或拷贝数变异。该方法可以帮助检测身体中癌性细胞的存在,即使不存在疾病的症状或其他标志(hallmarks)。

[0104] 如本文使用的,术语“癌症”包括但不限于多种类型的恶性赘生物,其中大多数可

以侵入周围组织,并可以转移至不同部位(参见例如,PDR Medical Dictionary,第1版(1995),出于所有目的通过引用以其整体并入本文)。术语“赘生物(neoplasm)”和“肿瘤(tumor)”是指通过细胞增殖比正常组织更迅速地生长并且在去除引发增殖的刺激之后继续生长的异常组织。这种异常组织可以是良性的(诸如良性肿瘤)或恶性的(诸如恶性肿瘤),显示出部分或完全缺乏有结构的组织和与正常组织的功能协调。癌症的一般类别的实例包括但不限于上皮癌(carcinoma)(源自上皮细胞的恶性肿瘤,诸如,例如,乳腺癌、前列腺癌、肺癌和结肠癌的常见形式)、肉瘤(源自结缔组织或间充质细胞的恶性肿瘤)、淋巴瘤(源自造血细胞的恶性肿瘤)、白血病(源自造血细胞的恶性肿瘤)和生殖细胞肿瘤(源自全能性细胞的肿瘤,在成人中最常见于睾丸或卵巢中;在胎儿、婴儿和年幼儿童中,最常见于身体中线,特别是在尾骨尖处)、母细胞肿瘤(blastic tumor)(类似未成熟组织或胚胎组织的典型恶性肿瘤)等。意图被本公开内容包括的赘生物类型的实例包括但不限于与神经组织、血液形成组织、乳房、皮肤、骨骼、前列腺、卵巢、子宫、子宫颈、肝、肺、脑、喉、胆囊、胰腺、直肠、甲状旁腺、甲状腺、肾上腺、免疫系统、头和颈、结肠、胃、支气管和/或肾的癌症相关的那些赘生物。在特定实施方案中,可以被检测的癌症的类型和数目包括但不限于血癌、脑癌、肺癌、皮肤癌、鼻癌、喉癌、肝癌、骨癌、淋巴瘤、胰腺癌、皮肤癌、肠癌、直肠癌、甲状腺癌、膀胱癌、肾癌、口腔癌、胃癌、实体瘤(solid state tumors)、异质肿瘤、均质肿瘤等。

[0105] 在一些实施方案中,该系统和方法可以用于检测可能导致或起因于癌症的任何数目的遗传畸变。这些可以包括但不限于突变、突变、插入/缺失、拷贝数变异、颠换、易位、倒位、缺失、非整倍性、部分非整倍性、多倍性、染色体不稳定性、染色体结构改变、基因融合、染色体融合、基因截短、基因扩增、基因重复、染色体损伤、DNA损伤、核酸化学修饰的异常改变、表观遗传模式的异常改变、核酸甲基化的异常改变、感染及癌症。

[0106] 另外,本文描述的系统和方法还可以用于帮助表征某些癌症。由本公开内容的系统和方法产生的遗传数据可以允许帮助从业者更好地表征具体形式的癌症。许多时候,癌症在组成和分期两个方面是异质的。遗传谱数据可以允许表征癌症的具体亚型,该表征在该具体亚型的诊断或治疗中可能是重要的。该信息还可以为受试者或从业者提供关于癌症具体类型的预后的线索。

[0107] 在一些实施方案中,本文提供的系统和方法用于监测特定受试者中已经知晓的癌症或其他疾病。这可以允许受试者或从业者根据疾病的进展调整治疗选项。在该实例中,本文描述的系统和方法可以用于构建疾病进程中特定受试者的遗传谱。在一些情况下,癌症可以进展,变成更具侵袭性和遗传上不稳定性。在其他实例中,癌症可以保持为良性的、非活动的、或休眠的。本公开内容的系统和方法可以用于确定疾病进展。

[0108] 此外,本文描述的系统和方法可以用于确定特定治疗选项的功效。在一些实施方案中,如果治疗成功,则成功的治疗选项可以实际上增加在受试者血液中检测到的拷贝数变异或突变的量,因为更多的癌症可能死亡并使DNA脱落。在其他实施方案中,这可能不发生。在一些实施方案中,某些治疗选项与癌症随时间推移的遗传谱相关。这种相关性可以用于选择疗法。另外,如果观察到癌症在治疗之后正在缓解,则本文描述的系统和方法可以用于监测剩余的疾病或疾病的复发。

[0109] 本文描述的方法和系统不限于仅与癌症相关的突变和拷贝数变异的检测。多种其他疾病和感染可能导致其他类型的状况,这可以适用于早期检测和监测。例如,在某些情况

下,遗传紊乱或传染性疾病可以在受试者中引起某些遗传镶嵌现象(genetic mosaicism)。这种遗传镶嵌现象可以引起可观察到的拷贝数变异和突变。在一些实施方案中,本公开内容的系统和方法也可以用于监测身体内免疫细胞的基因组。当存在某些疾病后,免疫细胞,诸如B细胞,可以经历快速克隆扩增。克隆扩增可以使用拷贝数变异检测来监测并且可以监测某些免疫状态。在该实例中,拷贝数变异分析可以随时间推移而进行,以产生特定疾病可能如何进展的谱。

[0110] 在一些实施方案中,本公开内容的方法适用于自身免疫性或免疫相关的疾病或状况。如本文使用的,“自身免疫性或免疫相关的疾病或状况”可以指影响免疫系统或与免疫系统相关的任何疾病、紊乱或状况。自身免疫性或免疫相关的疾病或状况的实例包括但不限于,炎症、抗磷脂综合征、系统性红斑狼疮、类风湿性关节炎、自身免疫性血管炎、乳糜泻、自身免疫性甲状腺炎、输血后免疫、母体胎儿不相容(maternal-fetal incompatibility)、输血反应、免疫缺陷诸如IgA缺陷、常见变异型免疫缺陷、药物诱导性狼疮、糖尿病、I型糖尿病、II型糖尿病、幼年型糖尿病、幼年型类风湿性关节炎、银屑病性关节炎、多发性硬化、免疫缺陷、过敏、哮喘、银屑病、特应性皮炎、过敏性接触性皮炎、慢性皮肤病、肌萎缩侧索硬化、化疗所致损伤、移植物抗宿主病(graft-vs-host diseases)、骨髓移植排斥、强直性脊柱炎、特应性湿疹、天疱疮、白塞病、慢性疲劳综合症纤维肌痛、化疗所致损伤、重症肌无力、肾小球肾炎、过敏性视网膜炎、系统性硬化、亚急性皮肤型红斑狼疮、包括冻疮样红斑狼疮的皮肤型红斑狼疮、干燥综合征、自身免疫性肾炎、自身免疫性血管炎、自身免疫性肝炎、自身免疫性心脏炎、自身免疫性脑炎、自身免疫介导的血液病、lc-SSc(局限性皮肤型硬皮病)、dc-SSc(弥漫性皮肤型硬皮病)、自身免疫性甲状腺炎(AT)、格雷夫斯病(GD)、重症肌无力、多发性硬化(MS)、强直性脊柱炎、移植排斥、免疫衰老、风湿性/自身免疫性疾病、混合型结缔组织病、脊柱关节病、银屑病、银屑病性关节炎、肌炎、硬皮病、皮肌炎、自身免疫性血管炎、混合型结缔组织病、特发性血小板减少性紫癜、克罗恩病、人类佐剂病、骨性关节炎、幼年型慢性关节炎、脊柱关节病、特发性炎性肌病、系统性血管炎、结节病、自身免疫性溶血性贫血、自身免疫性血小板减少症、甲状腺炎、免疫介导的肾病、中枢或外周神经系统脱髓鞘病、特发性脱髓鞘多神经病、格林-巴利综合征、慢性炎性脱髓鞘多神经病、肝胆疾病、传染性或自身免疫性慢性活动性肝炎、原发性胆汁性肝硬化、肉芽肿性肝炎、硬化性胆管炎、炎性肠病、谷蛋白敏感性肠病、惠普尔病、自身免疫性或免疫介导的皮肤病、大疱性皮肤病、多形性红斑、过敏性鼻炎、特应性皮炎、食物过敏、荨麻疹、肺部免疫性疾病、嗜酸性肺炎、特发性肺纤维化、过敏性肺炎、移植相关疾病、移植物排斥或移植物抗宿主病、银屑病性关节炎、银屑病、皮炎、多肌炎/皮肌炎、中毒性表皮坏死松解症、系统性硬皮病和硬化、与炎性肠病相关的应答、克罗恩病、溃疡性结肠炎、呼吸窘迫综合征、成人型呼吸窘迫综合征(ARDS)、脑膜炎、脑炎、葡萄膜炎、结肠炎、肾小球肾炎、过敏性状况、湿疹、哮喘、涉及T细胞浸润和慢性炎性应答的状况、动脉粥样硬化、自身免疫性心肌炎、白细胞黏附缺陷症、过敏性脑脊髓炎、与由细胞因子和T淋巴细胞介导的急性和迟发性过敏相关的免疫应答、结核病、结节病、包括韦格纳肉芽肿病的肉芽肿病、粒细胞缺乏症、血管炎(包括ANCA)、再生障碍性贫血、Diamond Blackfan贫血、包括自身免疫性溶血性贫血(AIHA)的免疫性溶血性贫血、恶性贫血、纯红细胞再生障碍(PRCA)、因子VIII缺乏症、血友病A、自身免疫性中性粒细胞减少症、全血细胞减少症、白细胞减少症、涉及白细胞渗出的疾病、中枢神经系统(CNS)炎性紊乱、多

器官损伤综合征、重症肌无力、抗原-抗体复合物介导的疾病、抗肾小球基膜病、抗磷脂抗体综合征、过敏性神经炎、白塞病、Castleman综合征、Goodpasture综合征、兰伯特-伊顿肌无力综合征、雷诺综合征、干燥综合征、斯-约综合征、大疱性类天疱疮、天疱疮、自身免疫性多内分泌腺疾病、Reiter病、僵人综合征、巨细胞性动脉炎、免疫复合物性肾炎、IgA肾病、IgM多神经病或IgM介导的神经病、特发性血小板减少性紫癜(ITP)、血栓性血小板减少性紫癜(TTP)、自身免疫性血小板减少症、包括自身免疫性睾丸炎和卵巢炎的睾丸和卵巢自身免疫性疾病、原发性甲状腺功能减退症、包括自身免疫性甲状腺炎的自身免疫性内分泌疾病、慢性甲状腺炎(桥本甲状腺炎)、亚急性甲状腺炎、特发性甲状腺功能减退症、爱迪生氏病、格雷夫斯病、自身免疫性多内分泌腺综合征(或多腺性内分泌病综合征)、席汉氏综合征、自身免疫性肝炎、淋巴细胞性间质性肺炎、HIV、闭塞性细支气管炎(非移植)对NSIP、格林-巴利综合征、大血管血管炎(包括风湿性多肌痛和巨细胞性(高安)动脉炎)、中血管血管炎(包括川崎病和结节性多动脉炎)、强直性脊柱炎、Berger病(IgA肾病)、快速进行肾小球肾炎、原发性胆汁性肝硬化、口炎性腹泻(Celiac sprue)(谷蛋白肠病)、冷球蛋白血症、和肌萎缩侧索硬化(ALS)。在某些实施方案中,本公开内容的方法适用于炎性状况,包括但不限于哮喘、多发性硬化(例如,复发缓解型多发性硬化和继发性进行性多发性硬化)、关节炎(例如,类风湿性关节炎、骨关节炎和银屑病性关节炎)、红斑狼疮和银屑病。

[0111] 在一些实施方案中,本公开内容的系统和方法可以用于监测全身性感染本身,其可以由病原体诸如细菌或病毒引起。对拷贝数变异或甚至突变的检测可以用于确定病原体群体在感染过程期间是如何改变的。这在慢性感染诸如HIV/AIDS或肝炎感染期间可能特别重要,其中病毒可以在感染过程期间改变生命周期状态和/或突变为毒力更强的形式。

[0112] 在一些实施方案中,本公开内容的系统和方法可以用于监测移植受试者。通常,移植组织在移植后经历一定程度的身体排斥。本公开内容的方法可以用于确定或谱分析免疫细胞试图破坏移植组织时的宿主体的排斥活动。这可以用于监测移植组织的状态以及改变治疗过程或预防排斥。

[0113] 此外,在一些实施方案中,本公开内容的方法可以用于表征受试者中的异常状况的异质性,该方法包括生成受试者中的细胞外多核苷酸的遗传谱,其中该遗传谱包含由拷贝数变异和突变分析得到的多于一个数据。在一些情况下,包括但不限于癌症,疾病可以是异质的。疾病细胞可以不相同。在癌症的实例中,已知一些肿瘤包含不同类型的肿瘤细胞,一些细胞处于癌症的不同时期。在一些实施方案中,异质性包括疾病的多个病灶。同样,在癌症的实例中,可以存在多个肿瘤病灶,或许其中一个或更多个病灶为已从原发部位扩散的转移的结果。

[0114] 本公开内容的方法可以用于生成为由异质性疾病中的不同细胞得到的遗传信息的总和的谱、指纹图谱或数据集。该数据集可以包含单独的或组合的拷贝数变异和突变分析。

[0115] 另外,本公开内容的系统和方法可以用于诊断、预后、监测或观察胎儿来源的癌症或其他疾病。即,这些方法可以用于妊娠的受试者,以诊断、预后、监测或观察未出生受试者的癌症或其他疾病,未出生受试者的DNA和其他多核苷酸可以与母体分子共循环。在一些实施方案中,该系统和方法可用于诊断、预后、监测或观察产前或妊娠相关的疾病或状况。如本文使用的,术语“产前或妊娠相关的疾病或状况”是指影响妊娠女性、胚胎或胎儿的任何

疾病、紊乱或状况。产前或妊娠相关的状况也可以指与妊娠相关或由妊娠结果直接或间接引起的任何疾病、紊乱或状况。这些疾病或状况可以包括任何及所有出生缺陷、先天性状况或遗传性疾病或状况。产前或妊娠相关的疾病的实例包括但不限于,Rhesus病、新生儿溶血病、β地中海贫血、性别决定、妊娠确定、遗传性孟德尔遗传紊乱、染色体畸变、胎儿染色体非整倍性、胎儿染色体三体、胎儿染色体单体、8三体(trisomy8)、13三体(Patau综合征)、16三体、18三体(Edwards综合征)、21三体(唐氏综合征)、X染色体连锁紊乱、X三体(XXX综合征)、X单体(Turner综合征)、XXY综合征、XYY综合征、XYY综合征、XXXY综合征、XXYY综合征、XYYY综合征、XXXXX综合征、XXXXY综合征、XXXYY综合征、XXYYY综合征、脆性X综合征、胎儿生长受限、囊性纤维化、血红蛋白病、胎儿死亡、胎儿酒精综合征、镰状细胞贫血、血友病、克兰费尔特综合征、dup(17)(p11.2p1.2)综合征、子宫内膜异位症、佩利措伊斯-梅茨巴赫病、dup(22)(q11.2q11.2)综合征、猫眼综合征、猫叫综合征、Wolf-Hirschhorn综合征、Williams-Beuren综合征、夏科-马里-图思病、压力易感性周围神经病(neuropathy with liability to pressure palsies)、史密斯-马吉利综合症、神经纤维瘤病、Alagille综合征、Velocardiofacial综合征、DiGeorge综合征、类固醇硫酸酯酶缺乏症、普拉德-威利综合征、卡尔曼综合征、小眼畸形合并线性皮肤缺损(microphthalmia with linear skin defects)、肾上腺发育不良、甘油激酶缺乏症、佩利措伊斯-梅茨巴赫病、Y上的睾丸决定因子、无精子症(因子a)、无精子症(因子b)、无精子症(因子c)、1p36缺失、苯丙酮尿症、Tay-Sachs病、肾上腺增生、范科尼贫血、脊髓性肌萎缩、杜氏肌营养不良、亨廷顿舞蹈症、强直性肌营养不良、罗伯逊易位、Angelman综合征、结节性硬化、共济失调毛细血管扩张症(ataxia telangieltasia)、开放性脊柱裂、神经管缺陷、腹壁缺陷(ventral wall defects)、小于胎龄儿(small-for-gestational-age)、先天性巨细胞病毒、软骨发育不全、马凡综合征、先天性甲状腺功能减退症、先天性弓形体病、生物素酰胺酶缺乏症、半乳糖血症、枫糖尿症、同型胱氨酸尿症、中链脂酰基Co-A脱氢酶缺乏症、结构性出生缺陷、心脏缺陷、肢体异常(abnormal limbs)、畸形足、无脑畸形、无嗅脑畸形/全前脑畸形、脑积水、无眼畸形/小眼畸形、无耳畸形/小耳畸形、大血管错位、法洛四联症、左心发育不全综合征、主动脉缩窄、腭裂合并唇裂(cleft palate without cleft lip)、唇裂合并腭裂或唇裂没有腭裂、食管闭锁/狭窄合并或没有瘘、小肠闭锁/狭窄、肛门直肠闭锁/狭窄、尿道下裂、性别不定(性别不定)、肾不发育(renal agenesis)、囊性肾、轴前多指(preaxial polydactyly)、肢体缩小缺陷、膈疝、盲(blindness)、白内障、视觉问题、听觉损失、耳聋、X连锁肾上腺脑白质营养不良、Rett综合征、溶酶体紊乱、大脑性麻痹、孤独症、无舌畸形、白化病、眼白化病、眼皮肤白化病、妊娠糖尿病、阿诺德-基亚里畸形、CHARGE综合征、先天性膈疝、短指畸形、无虹膜、足裂和手裂、虹膜异色症、Darwinian耳、埃勒斯-当洛斯综合征、大疱性表皮松解症、Gorham病、桥本综合征、胎儿水肿、张力过低、克利佩尔-费尔综合征、肌营养不良、成骨不全、早老症、Smith Lemli Opitz综合征、色盲(chromatelopsia)、X连锁淋巴增生性疾病、脐膨出、腹裂、先兆子痫、子痫、未足月产、早产、流产、宫内发育迟缓、异位妊娠、妊娠剧吐、早孕反应或成功引产的可能性。

[0116] 此外,在一些实施方案中,报告经由互联网以电子方式进行提交及访问。在某些实施方案中,在受试者的位置以外的地点进行序列数据的分析。生成报告并发送至受试者的位置。经由支持互联网的计算机(internet enabled computer),受试者访问反映其肿瘤负

荷的报告。

[0117] 注释的信息可以由健康护理提供者使用以选择其他药物治疗选项和/或向保险公司提供关于药物治疗选项的信息。该方法可以包括以例如NCCN肿瘤学临床实践指南(the NCCN Clinical Practice Guidelines in Oncology)或美国临床肿瘤学会(the American Society of Clinical Oncology)(ASCO)临床实践指南为状况注释药物治疗选项。

[0118] 在报告中被分级的药物治疗选项可以通过列出另外的药物治疗选项在报告中进行注释。另外的药物治疗可以是用于FDA批准的非正式批准用途的药物。1993年综合预算调节法案(Omnibus Budget Reconciliation Act)(OBRA)中的一项规定要求医疗保险(Medicare)覆盖标准医学纲要(standard medical compendia)中包含的抗癌药物的非正式批准用途。用于注释列表的药物可以见于CMS批准的纲要,包括the National Comprehensive Cancer Network(NCCN)Drugs and Biologics Compendium、Thomson Micromedex DrugDex®、Elsevier Gold Standard's Clinical Pharmacology compendium、和American Hospital Formulary Service—Drug Information Compendium®。

[0119] 药物治疗选项可以通过列出可以用于治疗具有特定状态的一种或更多种分子标志物的癌症的实验药物来注释。实验药物可以是可获得其体外数据、体内数据、动物模型数据、临床前试验数据或临床试验数据的药物。数据可以被公布于同行评议的医学文献中,所述同行评议的医学文献见于CMS医疗保险福利政策手册(CMS Medicare Benefit Policy Manual)中列出的期刊,包括,例如,American Journal of Medicine、Annals of Internal Medicine、Annals of Oncology、Annals of Surgical Oncology、Biology of Blood and Marrow Transplantation、Blood、Bone Marrow Transplantation、British Journal of Cancer、British Journal of Hematology、British Medical Journal、Cancer、Clinical Cancer Research、Drugs、European Journal of Cancer(原先是European Journal of Cancer and Clinical Oncology)、Gynecologic Oncology、International Journal of Radiation,Oncology,Biology,and Physics、The Journal of the American Medical Association、Journal of Clinical Oncology、Journal of the National Cancer Institute、Journal of the National Comprehensive Cancer Network(NCCN)、Journal of Urology、Lancet、Lancet Oncology、Leukemia、The New England Journal of Medicine、和Radiation Oncology。

[0120] 药物治疗选项可以通过提供基于电子报告的链接来连接列出的药物与关于该药物的科学信息来注释。例如,可以提供关于药物临床试验信息的链接(clinicaltrials.gov)。如果报告经由计算机或计算机网站来提供,则链接可以是脚注、网站超链接、带有信息的弹出框或悬浮框等。报告和注释信息可以以打印形式提供,并且注释可以是例如对参考的脚注。

[0121] 在报告中,用于注释一个或更多个药物治疗选项的信息可以由存储科学信息的商业实体提供。健康护理提供者可以用注释信息中列出的实验药物来治疗受试者诸如癌症患者,并且健康护理提供者可以访问注释的药物治疗选项,检索科学信息(例如,打印医学期刊文章)并将科学信息(例如,打印的医学期刊文章)连同用于提供药物治疗的报销要求提

交至保险公司。医师可以使用多种诊断相关组(Diagnosis-related group)(DRG)代码中的任何一种来实现报销。

[0122] 报告中的药物治疗选项也可以用关于药物影响的途径中的其他分子组分的信息(例如,关于靶向在作为药物靶的细胞表面受体下游的激酶的药物的信息)来注释。药物治疗选项可以用关于靶向一种或更多种其他分子途径组分的药物的信息来注释。与途径相关的信息的鉴定和/或注释可以外包或分包给另一公司。

[0123] 注释的信息可以是,例如,药物名称(例如,用于FDA批准的非正式批准用途的药物;见于CMS批准的纲要中的药物、和/或科学(医学)期刊文章中描述的药物)、关于一个或更多个药物治疗选项的科学信息、关于一种或更多种药物的科学信息的一个或更多个链接、关于一种或更多种药物的临床试验信息(例如,来自clinicaltrials.gov/的信息)、关于药物的科学信息的引用的一个或更多个连接等。

[0124] 注释的信息可以被插入到报告中的任何位置。注释的信息可以被插入到报告中的多个位置。注释的信息可以被插入到报告中在关于分级药物治疗选项的部分附近。注释的信息可以被插入到报告中在与分级药物治疗选项分开的页面上。不包含分级药物治疗选项的报告可以用信息来注释。

[0125] 系统还可以包括药物对从受试者(例如癌症患者)分离的样品(例如肿瘤细胞)的影响的报告。使用来自癌症患者的肿瘤的体外培养可以使用多种技术来建立。系统还可以包括使用所述体外培养和/或异种移植模型来高通量筛选FDA批准的非正式批准用途的药物或实验药物。系统还可以包括监测肿瘤抗原以用于复发检测。

[0126] 系统可以提供支持互联网访问的患有癌症的受试者的报告。系统可以使用手持式DNA测序仪或台式DNA测序仪。DNA测序仪为用于自动化DNA测序过程的科学仪器。对给定DNA样品,DNA测序仪用于确定四种碱基的顺序:腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶。DNA碱基的顺序被报告为文本字符串,称为读段。一些DNA测序仪也可以被认为是光学仪器,因为它们分析源于与核苷酸附接的荧光染料的光信号。

[0127] DNA测序仪可以应用基于DNA的化学修饰随后在特定碱基处裂解的Gilbert测序方法,或者DNA测序仪可以应用基于双脱氧核苷酸链终止的Sanger技术。Sanger方法由于其增加的效率和低放射性而变得流行。DNA测序仪可以使用不需要DNA扩增(聚合酶链式反应—PCR)的技术,这加快了测序前的样品制备并减少了错误。另外,从互补链中的核苷酸的实时添加引起的反应收集测序数据。例如,DNA测序仪可以利用被称为单分子实时(single-molecule real-time)(SMRT)的方法,其中测序数据通过当核苷酸由包含荧光染料的酶添加至互补链时发射的光(由相机捕获)来产生。可选地,DNA测序仪可以使用基于纳米孔感测技术的电子系统。

[0128] 数据由DNA测序仪通过直接连接或通过互联网发送至计算机进行处理。系统的数据处理方面可以以数字电子电路或以计算机硬件、固件、软件或其组合来实现。本公开内容的数据处理设备可以有形地体现在机器可读存储装置中以用于通过可编程处理器执行的计算机程序产品来实施;并且本公开内容的数据处理方法步骤可以由执行指令程序的可编程处理器进行,以通过操作输入数据并生成输出来进行本公开内容的功能。本公开内容的数据处理方面可以有利地在一个或更多个计算机程序中实施,所述一个或更多个计算机程序可在可编程系统中执行,所述可编程系统包括耦合以从数据存储系统接收数据和指令并

向数据存储系统传输数据和指令的至少一个可编程处理器、至少一个输入装置以及至少一个输出装置。如果需要,每一个计算机程序可以以高级程序或面向对象编程语言或者汇编或机器语言来实施;并且,在任何情况下,语言可以是编译或解译语言。合适的处理器包括,例如,通用和专用的微处理器两者。通常,处理器将从只读存储器和/或随机存取存储器接收指令和数据。适用于有形地体现计算机程序指令和数据的存储装置包括非易失性存储器的所有形式,包括例如半导体存储器装置,诸如EPROM、EEPROM和闪存装置;磁盘,诸如内置硬盘和可移动磁盘;磁光盘;和CD-ROM盘。前述中的任一项可以由ASIC(专用集成电路)补充或并入ASIC中。

[0129] 为了提供与用户的交互,本公开内容可以使用具有显示装置和输入装置的计算机系统来实现,所述显示装置诸如监视器或LCD(液晶显示器)屏幕用于向用户显示信息,用户可以通过所述输入装置将输入提供至计算机系统,所述输入装置诸如键盘、二维点击装置诸如鼠标或轨迹球、或者三维点击装置诸如数据手套或陀螺仪鼠标。计算机系统可以被编程为提供图形用户界面,计算机程序通过该图形用户界面与用户交互。计算机系统可以被编程为提供虚拟现实的三维显示界面。

[0130] 测试样品

[0131] 本文公开的方法可以包括分离一种或更多种多核苷酸。

[0132] 多核苷酸可以包括任何类型的核酸,诸如DNA和/或RNA。例如,如果多核苷酸是DNA,它可以是基因组DNA、互补DNA(cDNA)或任何其他脱氧核糖核酸。多核苷酸也可以是无细胞核酸,诸如无细胞DNA(cfDNA)。例如,多核苷酸可以是循环cfDNA。循环cfDNA可以包括经由凋亡或坏死从身体细胞脱落的DNA。经由凋亡或坏死脱落的cfDNA可以来源于正常的身体细胞。在存在异常组织生长的情况下,诸如癌症,肿瘤DNA可能会脱落。循环cfDNA可以包括循环肿瘤DNA(ctDNA)。如本文描述的,本公开的方法允许技术人员从cfDNA中确定遗传基因座(例如,遗传基因座处的变异)的来源是种系还是体细胞,而不需要从基因组DNA中分离序列信息。

[0133] 多核苷酸可以是双链的或单链的。可选地,多核苷酸可以包括双链部分和单链部分的组合。

[0134] 样品可以是从受试者分离的任何生物样品。例如,样品可以包括但不限于体液、全血、血小板、血清、血浆、粪便、红细胞、白细胞(white blood cell)或白细胞(leukocyte)、内皮细胞、组织活组织检查、滑液、淋巴液、腹水、间质或细胞外液、细胞间空间的液体,包括龈沟液、骨髓、脑脊液、唾液、粘液、痰、精液、汗液、尿液、鼻刷液、巴氏涂片液或任何其他体液。体液可以包括唾液、血液或血清。例如,多核苷酸可以是从体液例如血液或血清分离的无细胞DNA。样品也可以是肿瘤样品,肿瘤样品可以通过各种方法从受试者获得,所述方法包括但不限于静脉穿刺、排泄、射精、按摩、活组织检查、针抽吸、灌洗、刮擦、手术切口或介入或其他方法。样品可以是无细胞样品(例如,不包含任何细胞)。

[0135] 样品可以包含一定体积的含有无细胞DNA分子的血浆。样品可以包含足以实现给定读段深度的体积的血浆。取样血浆的体积可以是至少0.5毫升(mL)、1mL、5mL、10mL、20mL、30mL或40mL。取样血浆的体积为至多0.5mL、1mL、5mL、10mL、20mL、30mL或40mL。取样血浆的体积可以是5mL至20mL。取样血浆的体积可以是10mL至20mL。

[0136] 样品可以包含不同量的包含基因组当量的核酸。例如,约30ng DNA的样品可以包

含约10,000(1×10^4)个单倍体人类基因组当量,而在cfDNA的情况下,可以包含约2000亿(2×10^{11})个单独的多核苷酸分子。类似地,约100ng DNA的样品可以包含约30,000个单倍体人类基因组当量,而在cfDNA的情况下,可以包含约6000亿个单独的分子。

[0137] 样品可以包含来自不同来源的核酸。例如,样品可以包含种系DNA或体细胞DNA。样品可以包含携带突变的核酸。例如,样品可以包含携带种系突变和/或体细胞突变的DNA。样品还可以包含携带癌症相关突变(例如,癌症相关的体细胞突变)的DNA。在一些实施方案中,样品包含以下的一种或更多种:单碱基取代、拷贝数变异、插入/缺失、基因融合、颠换、易位、倒位、缺失、非整倍性、部分非整倍性、多倍性、染色体不稳定性、染色体结构改变、染色体融合、基因截短、基因扩增、基因重复、染色体损伤、DNA损伤、核酸化学修饰的异常改变、表观遗传模式的异常改变、基因组区域内核酸(例如,cfDNA)片段分布的异常改变、核酸(例如,cfDNA)片段长度分布的异常改变、和核酸甲基化的异常改变。

[0138] 本文的方法可以包括从样品获得一定量的核酸分子,例如,无细胞核酸分子。例如,该方法可以包括从样品获得多达约600ng、多达约500ng、多达约400ng、多达约300ng、多达约200ng、多达约100ng、多达约50ng、或多达约20ng的无细胞核酸分子。该方法可以包括获得至少1飞克(fg)、至少10fg、至少100fg、至少1皮克(pg)、至少10pg、至少100pg、至少1ng、至少10ng、至少100ng、至少150ng、或至少200ng的无细胞核酸分子。该方法可以包括获得至多1飞克(fg)、至多10fg、至多100fg、至多1皮克(pg)、至多10pg、至多100pg、至多1ng、至多10ng、至多100ng、至多150ng、或至多200ng的无细胞核酸分子。该方法可以包括获得1飞克(fg)至200ng、1皮克(pg)至200ng、1ng至100ng、10ng至150ng、10ng至200ng、10ng至300ng、10ng至400ng、10ng至500ng、10ng至600ng、10ng至700ng、10ng至800ng、10ng至900ng、或10ng至1000ng的无细胞核酸分子。无细胞核酸分子的量可以相当于单倍体基因组拷贝的数目。因为单倍体基因组拷贝具有约3.3皮克(pg)的质量,所以每纳克(ng)无细胞核酸分子可以相当于约300个单倍体基因组拷贝。例如,5ng无细胞核酸分子可以相当于1,500个基因组拷贝。

[0139] 无细胞核酸可以是不附着于细胞的任何细胞外核酸。无细胞核酸可以是在血液中循环的核酸。可选地,无细胞核酸可以是本文公开的其他体液例如尿液中的核酸。无细胞核酸可以是脱氧核糖核酸(“DNA”),例如,基因组DNA、线粒体DNA或其片段。无细胞核酸可以是核糖核酸(“RNA”),例如,mRNA、短干扰RNA(siRNA)、微RNA(miRNA)、循环RNA(cRNA)、转移RNA(tRNA)、核糖体RNA(rRNA)、小核仁RNA(snoRNA)、Piwi相互作用RNA(piRNA)、长非编码RNA(长ncRNA)或其片段。在一些情况下,无细胞核酸是DNA/RNA杂交体。无细胞核酸可以是双链的、单链的或其杂交体。无细胞核酸可以通过分泌或细胞死亡过程例如细胞坏死和凋亡释放到体液中。

[0140] 无细胞核酸可以包含一种或更多种表观遗传学修饰。例如,无细胞核酸可以被乙酰化、甲基化、泛素化、磷酸化、sumo化(sumoylated)、核糖基化和/或瓜氨酸化。例如,无细胞核酸可以是甲基化的无细胞DNA。

[0141] 无细胞DNA通常具有约110个至约230个核苷酸的尺寸分布,具有约168个核苷酸的模式。在定量无细胞核酸分子长度的测定中检测到的第二个次峰具有240个至440个核苷酸之间的范围。另外的更高量级的核苷酸峰也以更长的长度存在。

[0142] 在本公开内容的一些实施方案中,无细胞核酸的长度可以为至多1,000个核苷酸

(nt)、长度为至多500个核苷酸、长度为至多400个核苷酸、长度为至多300个核苷酸、长度为至多250个核苷酸、长度为至多225个核苷酸、长度为至多200个核苷酸、长度为至多190个核苷酸、长度为至多180个核苷酸、长度为至多170个核苷酸、长度为至多160个核苷酸、长度为至多150个核苷酸、长度为至多140个核苷酸、长度为至多130个核苷酸、长度为至多120个核苷酸、长度为至多110个核苷酸、或长度为至多100个核苷酸。

[0143] 在本公开内容的一些实施方案中,无细胞核酸的长度可以为至少1,000个核苷酸、长度为至少500个核苷酸、长度为至少400个核苷酸、长度为至少300个核苷酸、长度为至少250个核苷酸、长度为至少225个核苷酸、长度为至少200个核苷酸、长度为至少190个核苷酸、长度为至少180个核苷酸、长度为至少170个核苷酸、长度为至少160个核苷酸、长度为至少150个核苷酸、长度为至少140个核苷酸、长度为至少130个核苷酸、长度为至少120个核苷酸、长度为至少110个核苷酸、或长度为至少100个核苷酸。无细胞核酸的长度可以为从140个至180个核苷酸。

[0144] 在本公开内容的一些实施方案中,受试者中的无细胞核酸可以源自肿瘤。例如,从受试者分离的无细胞DNA可以包括循环肿瘤DNA(ctDNA)。下一代测序允许检测及测量罕见突变。在一部分无细胞DNA中检测到相对于种系序列的突变可以指示ctDNA的存在,从而指示肿瘤的存在。对无细胞DNA进行测序可以允许检测已知指示癌症存在的遗传变异。例如,对无细胞DNA进行测序可以允许检测癌症相关基因的突变。

[0145] 分离和提取

[0146] 无细胞多核苷酸可以是胎儿来源的(经由从妊娠受试者取得的液体),或者可以源自受试者自身的组织。无细胞多核苷酸可以源自健康组织、源自疾病组织诸如肿瘤组织、或源自移植器官。

[0147] 在一些实施方案中,无细胞多核苷酸源自血液样品或其级分。例如,血液样品(例如,约10ml至约30ml)可取自受试者,离心去除细胞,并将所得血浆用于cfDNA提取。

[0148] 多核苷酸的分离和提取可以通过使用多种技术收集体液来进行。在一些情况下,收集可以包括使用注射器从受试者抽吸体液。在其他情况下,收集可以包括将液体抽吸或直接收集到收集容器中。

[0149] 在收集体液之后,多核苷酸可以使用本领域利用的多种技术来分离和提取。在一些情况下,无细胞DNA可以使用商购可得的试剂盒诸如Qiagen Qiamp® Circulating Nucleic Acid试剂盒方案来分离、提取和制备。在其他实例中,可以使用Qiagen Qubit™ dsDNAHS测定试剂盒方案、Agilent™ DNA1000试剂盒、或TruSeq™ Sequencing Library Preparation;Low-Throughput(LT)方案。

[0150] 通常,无细胞多核苷酸可以通过分区步骤(partitioning step)从体液中提取和分离,在该分区步骤中,如在溶液中存在的无细胞DNA被与细胞和体液的其他不可溶性组分分开。分区可以包括但不限于诸如离心或过滤的技术。在其他情况下,可以先不将细胞与无细胞DNA分离,而是裂解。例如,完整细胞的基因组DNA可以通过选择性沉淀来分离。样品分离可以与用标识符(诸如包括条形码的标识符)对核酸加标签相结合,或者可以在不使用标识符的方法中进行。样品可以被分成分区,使得每个分区可以被独立地加条形码(例如,每个分区具有一个独特的条形码),并且来自分区的测序数据可以在以后被重新组合。样品可以被分成分区,并且核酸分子在分区内或分区之间相对于彼此被非独特地加标签。在一些

实施方案中,样品可以被分成分区,而不使用标识符。在一个实例中,cfDNA样品被分成4个或更多个分区,其中每个分区为空间上可寻址的位置。样品制备和测序在每个空间上可寻址的分区上进行,并且生物信息学可以利用可寻址的位置来进一步鉴定独特的分子。在一个实例中,核酸分子可以被分成多个分区,例如,包含不同类型的核酸分子(例如,双链核酸诸如DNA、和/或单链核酸诸如RNA和/或单链DNA)。无细胞多核苷酸(包括DNA)可以保持可溶性并可以与不可溶性基因组DNA分离并被提取。通常,在添加不同试剂盒特定的缓冲液和其他洗涤步骤之后,可以使用异丙醇沉淀来沉淀DNA。可以使用进一步的清洁(clean up)步骤诸如基于二氧化硅的柱或珠(诸如磁珠)来去除污染物或盐。可以针对具体应用来优化一般步骤。例如,可以在整个反应中添加非特异性批量(bulk)载体多核苷酸以优化该程序的某些方面诸如收率。

[0151] 在一些实施方案中,处理血浆样品以使蛋白酶K降解,并将DNA用异丙醇来沉淀并且随后在Qiagen柱上捕获。然后可以将DNA洗脱(例如,使用100微升(μl)洗脱液,诸如水或Tris-EDTA(TE)洗脱缓冲液)。在一些实施方案中,一部分DNA可以基于尺寸(例如,长度为500个核苷酸或更少的DNA),例如,使用固相可逆固定化(SPRI)珠,诸如AgenCourt®AMPure®珠来进行选择。在一些实施方案中,可以将DNA重悬在较小的体积诸如30μl水中,并检查DNA的尺寸分布(例如,检查166个核苷酸处的主峰和330个核苷酸处的次峰)。约5ng的DNA可能相当于约1500个单倍体基因组当量("HGE")。

[0152] 在提取之后,样品可以产生多达1微克(克)的DNA、多达800ng的DNA、多达500ng的DNA、多达300ng的DNA、多达250ng的DNA、多达200ng的DNA、多达180ng的DNA、多达160ng的DNA、多达140ng的DNA、多达120ng的DNA、多达100ng的DNA、多达90ng的DNA、多达80ng的DNA、多达70ng的DNA、多达60ng的DNA、多达50ng的DNA、多达40ng的DNA、多达30ng的DNA、多达20ng的DNA、多达10ng的DNA、多达9ng的DNA、多达8ng的DNA、多达7ng的DNA、多达6ng的DNA、多达5ng的DNA、多达4ng的DNA、多达3ng的DNA、多达2ng的DNA、或多达1ng的DNA。

[0153] 在提取之后,样品可以产生至少1ng的DNA、至少3ng的DNA、至少5ng的DNA、至少7ng的DNA、至少10ng的DNA、至少20ng的DNA、至少30ng的DNA、至少40ng的DNA、至少50ng的DNA、至少70ng的DNA、至少100ng的DNA、至少150ng的DNA、至少200ng的DNA、至少250ng的DNA、至少300ng的DNA、至少400ng的DNA、至少500ng的DNA、或至少700ng的DNA。

[0154] 可以将一种或更多种无细胞核酸与样品中的细胞碎片分离。在一些情况下,将一种或更多种无细胞核酸与膜、细胞器、核小体、外来体或细胞核、线粒体、糙面内质网、核糖体、光面内质网、叶绿体、高尔基体(Golgi apparatus)、高尔基体(Golgi body)、糖蛋白、糖脂、池(cisternae)、脂质体、过氧化物酶体、乙醛酸循环体、中心粒、细胞骨架、溶酶体、纤毛、鞭毛、收缩液泡、囊泡(vesicle)、核被膜、液泡、微管、核仁、质膜、内体(endosomes)、染色质或其组合分离。可以将一种或更多种无细胞核酸与一种或更多种外来体分离。在一些情况下,将一种或更多种无细胞核酸与一种或更多种细胞表面结合的核酸分离。

[0155] 无细胞DNA的纯化可以使用任何方法来完成,所述方法包括但不限于使用由诸如Sigma Aldrich、Life Technologies、Promega、Affymetrix、IBI等公司提供的商业试剂盒和方案。试剂盒和方案还可以是非商购可得的。

[0156] 在分离之后,在一些情况下,无细胞多核苷酸可以在测序前与一种或更多种另外的物质诸如一种或更多种试剂(例如,连接酶、蛋白酶、聚合酶)预混合。

[0157] 无细胞DNA可以被以足以检测样品中频率低至0.0005％的遗传变异的读段深度测序。无细胞DNA可以被以足以检测样品中频率低至0.001％的遗传变异的读段深度测序。无细胞DNA可以被以足以检测样品中频率低至1.0％、0.75％、0.5％、0.25％、0.1％、0.075％、0.05％、0.025％、0.01％或0.005％的遗传变异的读段深度测序。因此,对无细胞DNA测序允许非常灵敏地检测受试者中的癌症。

[0158] 本文的方法可以用于检测受试者中的癌症。可以对未知患有癌症或怀疑患有癌症的受试者的无细胞DNA测序,以诊断癌症的存在或不存在。对无细胞DNA测序提供了用于癌症的早期检测或已知癌症的“活组织检查”的无创方法。可以对被诊断为患有癌症的受试者中的无细胞DNA测序,以提供关于癌症的信息。可以对受试者的癌症治疗之前和之后的无细胞DNA测序,以确定治疗的功效。

[0159] 受试者可能被怀疑患有癌症或可能不被怀疑患有癌症。受试者可能已经经历了与癌症诊断一致的症状。受试者可能未经历任何症状或可能已经表现出与癌症不一致的症状。受试者可能已经被基于生物成像方法被诊断为患有癌症。受试者可能不具有可通过成像方法检测的癌症。成像方法可以是正电子发射断层扫描、磁共振成像、X射线、计算机轴向断层扫描、超声波或其组合。

[0160] 受试者可能表现出癌症。可选地,受试者可能不表现出可检测的癌症。在一些情况下,不表现出可检测的癌症的受试者可以患有癌症,但没有可检测的症状。未知患有癌症或怀疑患有癌症的受试者可以患有使用多种癌症筛查方法无法检测到的癌症。使用多种成像方法均不能检测到癌症。成像方法可以包括,例如,正电子发射断层扫描、磁共振成像、X射线、计算机轴向断层扫描、内窥镜检查、超声波或其组合。对于未知患有癌症或怀疑患有癌症的受试者,测试诸如组织活组织检查、骨髓抽吸、pap测试、粪便潜血测试、蛋白生物标志物检测例如前列腺特异性抗原测试、甲胎蛋白血液测试或CA-125测试、或其组合可以指示受试者未患有癌症,例如,对受试者未检测到癌症。在其他情况下,不表现出可检测的癌症的受试者可以没有任何癌症。

[0161] 受试者患癌症的风险可能比一般人群高。受试者可能具有癌症家族史。受试者可能具有已知的癌症风险的遗传来源。受试者可能已经暴露于已知会增加或引起癌症风险的环境条件。受试者可以是癌症的唯一风险因素为年龄和/或性别的患者。受试者可能不具有已知的癌症危险因素。

[0162] 受试者可能已经被诊断患有癌症。癌症可能为早期或晚期。癌症可能是转移性的或可能不是转移性的。受试者可能已经被诊断为患有的癌症类型包括但不限于:癌、肉瘤、淋巴瘤、白血病、生殖细胞肿瘤和母细胞瘤。受试者可能已经被诊断为患有的癌症类型包括但不限于:急性成淋巴细胞性白血病(ALL)、急性髓性白血病、肾上腺皮质癌、成人急性髓性白血病、成人原发部位不明癌、成人恶性间皮瘤、艾滋病相关癌症、艾滋病相关淋巴瘤、肛门癌、阑尾癌、星形细胞瘤、儿童小脑或大脑癌、基底细胞癌、胆管癌、膀胱癌、骨肿瘤、骨肉瘤/恶性纤维组织细胞瘤、脑癌、脑干胶质瘤、乳腺癌、支气管腺瘤/类癌、伯基特淋巴瘤、类癌瘤、原发性不明的癌、中枢神经系统淋巴瘤、小脑星形细胞瘤、大脑星形细胞瘤/恶性神经胶质瘤、宫颈癌、儿童急性髓性白血病、儿童原发部位不明的癌症、儿童癌症、儿童大脑星形细胞瘤、儿童间皮瘤、软骨肉瘤、慢性淋巴细胞白血病、慢性髓细胞性白血病、慢性骨髓增生性紊乱、结肠癌、皮肤T细胞淋巴瘤、促结缔组织增生性小圆细胞肿瘤、子宫内膜癌、子宫内膜

子宫癌、室管膜瘤、上皮样血管内皮瘤(EHE)、食管癌、尤因肿瘤肉瘤家族、尤因肿瘤家族中的尤因肉瘤、颅外生殖细胞肿瘤、性腺外生殖细胞肿瘤、肝外胆管癌、眼癌、眼内黑素瘤、胆囊癌、胃(gastric)(胃(stomach))癌、胃类癌、胃肠道类癌肿瘤、胃肠道间质瘤(GIST)、妊娠性滋养层细胞瘤、脑干胶质瘤、胶质瘤、毛细胞白血病、头颈癌、心脏癌、肝细胞(肝)癌、霍奇金淋巴瘤、下咽癌、下丘脑和视觉途径胶质瘤、胰岛细胞癌(内分泌胰腺)、卡波西肉瘤、肾癌(肾细胞癌)、喉癌、急性成淋巴细胞性白血病(也称为急性淋巴细胞白血病)、急性髓性白血病(也称为急性髓细胞性白血病)、慢性淋巴细胞性白血病(也称为慢性淋巴细胞白血病)、白血病(leukaemia)、慢性髓细胞性白血病(也称为慢性髓性白血病)、毛细胞白血病(leukemia)、唇及口腔癌、脂肪肉瘤、肝癌(原发性)、非小细胞肺癌、小细胞肺癌、淋巴瘤(艾滋病相关)、淋巴瘤、Waldenström巨球蛋白血症、男性乳腺癌、骨恶性纤维组织细胞瘤/骨肉瘤、髓母细胞瘤、黑素瘤、梅克尔细胞癌、原发灶隐匿转移性颈部鳞状癌、口癌、多发性内分泌肿瘤综合征、儿童多发性骨髓瘤(骨髓癌)、多发性骨髓瘤/浆细胞赘生物、蕈样肉芽肿、骨髓增生异常综合征、骨髓增生异常/骨髓增生性疾病、慢性髓细胞性白血病、粘液瘤、鼻腔和副鼻窦癌、鼻咽癌、神经母细胞瘤、非霍奇金淋巴瘤、非小细胞肺癌、少突神经胶质瘤、口腔癌、口咽癌、骨肉瘤/骨恶性纤维组织细胞瘤、卵巢癌、卵巢上皮癌(表面上皮间质肿瘤)、卵巢生殖细胞瘤、卵巢低恶性潜能肿瘤、胰腺癌、胰岛细胞癌、副鼻窦和鼻腔癌、甲状旁腺癌、阴茎癌、咽癌、嗜铬细胞瘤、松果体星形细胞瘤、松果体生殖细胞瘤、松果体母细胞瘤和幕上原始神经外胚层肿瘤、垂体腺瘤、浆细胞赘生物/多发性骨髓瘤、胸膜肺母细胞瘤、原发性中枢神经系统淋巴瘤、前列腺癌、直肠癌、肾细胞癌(肾癌)、肾盂和输尿管移行细胞癌、视网膜母细胞瘤、横纹肌肉瘤、唾液腺癌、Sézary综合征、皮肤癌(黑素瘤)、皮肤癌(非黑素瘤)、梅克尔细胞皮肤癌、小细胞肺癌、小肠癌、软组织肉瘤、鳞状细胞癌、原发灶隐匿转移性颈部鳞状癌、胃癌、幕上原始神经外胚层肿瘤、皮肤T细胞淋巴瘤、睾丸癌、喉癌、胸腺瘤和胸腺癌、胸腺瘤、甲状腺癌、肾盂和输尿管移行细胞癌、输尿管和肾盂移行细胞癌、尿道癌、子宫肉瘤、阴道癌、视觉途径和下丘脑神经胶质瘤、儿童视觉途径和下丘脑神经胶质瘤、外阴癌、Waldenström巨球蛋白血症和肾母细胞瘤(肾癌)。

[0163] 受试者可能先前接受过癌症治疗。受试者可能已经接受手术治疗、放射治疗、化学治疗、靶向癌症治疗剂或癌症免疫治疗。受试者可能已经被用癌症疫苗治疗。受试者可能已经被用实验性癌症治疗治疗。受试者可能尚未接受过癌症治疗。受试者可能从癌症缓解。受试者可能先前接受过癌症治疗,并且不表现出任何可检测的症状。

[0164] 遗传分析

[0165] 某些DNA测序方法使用序列捕获来富集感兴趣的序列。序列捕获通常包括使用与感兴趣的序列杂交的寡核苷酸探针。探针集策略可以包括将探针平铺在感兴趣的区域内。这样的探针可以是,例如,长度约60个至120个碱基。探针集可以具有约2x、3x、4x、5x、6x、8x、9x、10x、15x、20x、50x或更大的深度。序列捕获的有效性部分地取决于靶分子中与探针序列互补(或几乎互补)的序列的长度。富集的核酸分子可以代表多于5,000个人类基因组碱基、多于10,000个人类基因组碱基、多于15,000个人类基因组碱基、多于20,000个人类基因组碱基、多于25,000个人类基因组碱基、多于30,000个人类基因组碱基、多于35,000个人类基因组碱基、多于40,000个人类基因组碱基、多于45,000个人类基因组碱基、多于50,000个人类基因组碱基、多于55,000个人类基因组碱基、多于60,000个人类基因组碱基、多于

65,000个人类基因组碱基、多于70,000个人类基因组碱基、多于75,000个人类基因组碱基、多于80,000个人类基因组碱基、多于85,000个人类基因组碱基、多于90,000个人类基因组碱基、多于95,000个人类基因组碱基、或多于100,000个人类基因组碱基。富集的核酸分子可以代表不多于5,000个人类基因组碱基、不多于10,000个人类基因组碱基、不多于15,000个人类基因组碱基、不多于20,000个人类基因组碱基、不多于25,000个人类基因组碱基、不多于30,000个人类基因组碱基、不多于35,000个人类基因组碱基、不多于40,000个人类基因组碱基、不多于45,000个人类基因组碱基、不多于50,000个人类基因组碱基、不多于55,000个人类基因组碱基、不多于60,000个人类基因组碱基、不多于65,000个人类基因组碱基、不多于70,000个人类基因组碱基、不多于75,000个人类基因组碱基、不多于80,000个人类基因组碱基、不多于85,000个人类基因组碱基、不多于90,000个人类基因组碱基、不多于95,000个人类基因组碱基、或不多于100,000个人类基因组碱基。富集的核酸分子可以代表5,000-100,000个人类基因组碱基、5,000-50,000个人类基因组碱基、5,000-30,000个人类基因组碱基、10,000-100,000个人类基因组碱基、10,000-50,000个人类基因组碱基、或10,000-30,000个人类基因组碱基。富集的核酸分子可以代表多种核酸特征,包括遗传变异,诸如核苷酸变异(SNV)、拷贝数变异(CNV)、插入或缺失(例如,插入/缺失)、与癌症相关的核小体区域、基因融合和倒位。

[0166] 通常,本文提供的方法和系统可用于制备无细胞多核苷酸序列以用于下游应用测序反应。测序方法可以是大规模并行测序,即,同时(或以快速相继)对至少100个、1000个、10,000个、100,000个、1百万个、1千万个、1亿个、10亿个或100亿个多核苷酸分子中的任一个测序。测序方法可以包括但不限于:高通量测序、焦磷酸测序、边合成边测序、单分子测序、纳米孔测序、半导体测序、边连接边测序、边杂交边测序、RNA-Seq(Illumina)、数字基因表达(Helicos)、下一代测序、单分子边合成边测序(SMSS)(Helicos)、大规模并行测序、克隆单分子阵列(Solexa)、鸟枪法测序、Maxim-Gilbert测序或Sanger测序、引物步移、使用PacBio、SOLiD、Ion Torrent、或Nanopore平台测序、及本领域中已知的任何其他测序方法。

[0167] 基因组核酸样品(例如,基因组DNA样品)中的单独的多核苷酸片段可以通过用非独特标识符加标签,例如,对单独的多核苷酸片段非独特地加标签,而独特地鉴定。

[0168] 测序组

[0169] 为了提高检测指示肿瘤的突变的可能性,测序的DNA区域可以包含一组基因或基因组区域。选择有限的测序区域(例如,有限的组)可以减少所需的总测序(例如,测序的核苷酸总量)。测序组可以靶向多个不同的基因或区域来检测单个癌症、一组癌症或所有癌症。

[0170] 在一些方面中,选择靶向多种不同基因或基因组区域的组,使得确定比例的患有癌症的受试者在该组中的一个或更多个不同基因或基因组区域中显示遗传变异或肿瘤标志物。可以选择组以将测序区域限制在固定数目的碱基对。可以选择组以对期望的量的DNA测序。还可以选择组以实现期望的序列读段深度。可以选择组,以针对一定量的测序碱基对实现期望的序列读段深度或序列读段覆盖度。可以选择组以实现检测样品中的一种或更多种遗传变异的理论灵敏度、理论特异性和/或理论准确度。

[0171] 用于检测区域的组的探针可以包括用于检测热点区域的探针以及核小体感知(nucleosome-aware)探针(例如,KRAS密码子12和13),并且可以被设计成基于对受核小体

结合模式和GC序列组成影响的cfDNA覆盖度和片段尺寸变化的分析来优化捕获。本文使用的区域还可以包括基于核小体位置和GC模型来优化的非热点区域。该组可以包括多于一个子组,包括用于鉴定来源组织的子组(例如,使用已公布的文献来定义50-100个诱饵,这些诱饵代表跨组织具有最多样的转录谱(不一定为启动子)的基因)、全基因组骨架(例如,用于鉴定超保守基因组含量,及为了拷贝数碱基排列的目的用少数探针跨染色体稀疏平铺)、转录起始位点(TSS)/CpG岛(例如,用于捕获在例如肿瘤抑制基因(例如,结肠直肠癌中的SEPT9/VIM)启动子中的差异甲基化区域(例如,差异性甲基化区域(DMR))。在一些实施方案中,来源组织的标志物是组织特异性的表观遗传标志物。

[0172] 感兴趣的基因组位置的示例性列表可以见于表1和表2中。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表1中的至少5个、至少10个、至少15个、至少20个、至少25个、至少30个、至少35个、至少40个、至少45个、至少50个、至少55个、至少60个、至少65个、至少70个、至少75个、至少80个、至少85个、至少90个、至少95个或97个基因中的至少一部分。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表1中的至少5个、至少10个、至少15个、至少20个、至少25个、至少30个、至少35个、至少40个、至少45个、至少50个、至少55个、至少60个、至少65个或70个SNV。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表1中的至少1个、至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少11个、至少12个、至少13个、至少14个、至少15个、至少16个、至少17个或18个CNV。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表1中的至少1个、至少2个、至少3个、至少4个、至少5个或6个融合。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表1中的至少1个、至少2个或3个插入/缺失中的至少一部分。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表2中的至少5个、至少10个、至少15个、至少20个、至少25个、至少30个、至少35个、至少40个、至少45个、至少50个、至少55个、至少60个、至少65个、至少70个、至少75个、至少80个、至少85个、至少90个、至少95个、至少100个、至少105个、至少110个或115个基因中的至少一部分。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表2中的至少5个、至少10个、至少15个、至少20个、至少25个、至少30个、至少35个、至少40个、至少45个、至少50个、至少55个、至少60个、至少65个、至少70个或73个SNV。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表2中的至少1个、至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少11个、至少12个、至少13个、至少14个、至少15个、至少16个、至少17个或18个CNV。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表2中的至少1个、至少2个、至少3个、至少4个、至少5个或6个融合。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表2中的至少1个、至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少11个、至少12个、至少13个、至少14个、至少15个、至少16个、至少17个或18个插入/缺失中的至少一部分。这些感兴趣的基因组位置中的每一个都可以被鉴定为给定诱饵集组的骨架区域或热点区域。感兴趣的热点基因组位置的示例性列表可以见于表3中。在一些实施方案中,本公开内容的方法中使用的基因组区域包括表3中的至少1个、至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少11个、至少12个、至少13个、至少14个、至少15个、至少16个、至少17个、至少18个、至少19个、或至少20个基因中的至少一部分。列出了每个热点基

因组区域的若干特征,包括相关基因、其所在的染色体、代表基因基因座的基因组的起始和终止位置、基因基因座以碱基对的长度、基因覆盖的外显子、以及可能寻求捕获的感兴趣的给定基因组区域的关键特征(例如,突变类型)。

[0173] 表1

[0174]

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和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和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						

[0175] 表2

[0176]

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和20)
ATM	BRAF	BRCA1	BRCA2	CCND1	CCND2	CCND1	CCND2	FGFR2	
CCNE1	CDH1	CDK4	CDK6	CDKN2A	DDR2	CCNE1	CDK4	FGFR3	
CTNNB1	EGFR	ERBB2	ESR1	EZH2	FBXW7	CDK6	EGFR	NTRK1	ERBB2 (外显子19和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				ATM
RIT1	ROS1	SMAD4	SMO	MAPK1	STK11				
TERT	TP53	TSC1	VHL	MAPK3	MTOR				
NTRK3									APC
									ARID1A
									BRCA1
									BRCA2

[0177]

				CDH1
				CDKN2A
				GATA3
				KIT
				MLH1
				MTOR
				NF1
				PDGFRA
				PTEN
				RB1
				SMAD4
				STK11
				TP53
				TSC1
				VHL

[0178] 表3

[0179]

基因	染色体	起始位置	终止位置	长度 (bp)	覆盖的外显子	关键特征
ALK	chr2	29446405	29446655	250	内含子19	融合
ALK	chr2	29446062	29446197	135	内含子20	融合
ALK	chr2	29446198	29446404	206	20	融合
ALK	chr2	29447353	29447473	120	内含子19	融合
ALK	chr2	29447614	29448316	702	内含子19	融合
ALK	chr2	29448317	29448441	124	19	融合
ALK	chr2	29449366	29449777	411	内含子18	融合
ALK	chr2	29449778	29449950	172	18	融合
BRAF	chr7	140453064	140453203	139	15	BRAF V600
CTNNB1	chr3	41266007	41266254	247	3	S37
EGFR	chr7	55240528	55240827	299	18 和 19	G719 和缺失
EGFR	chr7	55241603	55241746	143	20	插入/T790M

[0180]

EGFR	chr7	55242404	55242523	119	21	L858R
ERBB2	chr17	37880952	37881174	222	20	插入
ESR1	chr6	152419857	152420111	254	10	V534, P535, L536, Y537, D538
FGFR2	chr10	123279482	123279693	211	6	S252
GATA3	chr10	8111426	8111571	145	5	SS / 插入/缺失
GATA3	chr10	8115692	8116002	310	6	SS / 插入/缺失
GNAS	chr20	57484395	57484488	93	8	R844
IDH1	chr2	209113083	209113394	311	4	R132
IDH2	chr15	90631809	90631989	180	4	R140, R172
KIT	chr4	55524171	55524258	87	1	
KIT	chr4	55561667	55561957	290	2	
KIT	chr4	55564439	55564741	302	3	
KIT	chr4	55565785	55565942	157	4	
KIT	chr4	55569879	55570068	189	5	
KIT	chr4	55573253	55573463	210	6	
KIT	chr4	55575579	55575719	140	7	
KIT	chr4	55589739	55589874	135	8	
KIT	chr4	55592012	55592226	214	9	
KIT	chr4	55593373	55593718	345	10 和11	557, 559, 560, 576
KIT	chr4	55593978	55594297	319	12 和13	V654
KIT	chr4	55595490	55595661	171	14	T670, S709
KIT	chr4	55597483	55597595	112	15	D716
KIT	chr4	55598026	55598174	148	16	L783
KIT	chr4	55599225	55599368	143	17	C809, R815, D816, L818, D820, S821F, N822, Y823

[0181]

KIT	chr4	55602653	55602785	132	18	A829P
KIT	chr4	55602876	55602996	120	19	
KIT	chr4	55603330	55603456	126	20	
KIT	chr4	55604584	55604733	149	21	
KRAS	chr12	25378537	25378717	180	4	A146
KRAS	chr12	25380157	25380356	199	3	Q61
KRAS	chr12	25398197	25398328	131	2	G12/G13
MET	chr7	116411535	116412255	720	13, 14, 内含子13, 内含子14	MET 外显子 14 SS
NRAS	chr1	115256410	115256609	199	3	Q61
NRAS	chr1	115258660	115258791	131	2	G12/G13
PIK3CA	chr3	178935987	178936132	145	10	E545K
PIK3CA	chr3	178951871	178952162	291	21	H1047R
PTEN	chr10	89692759	89693018	259	5	R130
SMAD4	chr18	48604616	48604849	233	12	D537
TERT	chr5	1294841	1295512	671	启动子	chr5:1295228
TP53	chr17	7573916	7574043	127	11	Q331, R337, R342
TP53	chr17	7577008	7577165	157	8	R273
TP53	chr17	7577488	7577618	130	7	R248
TP53	chr17	7578127	7578299	172	6	R213/Y220
TP53	chr17	7578360	7578564	204	5	R175 / 缺失
TP53	chr17	7579301	7579600	299	4	
				12574 (靶区域总数)		
				16330 (总探针覆盖度)		

[0182] 在一些实施方案中,组中的一个或更多个区域包括来自一个或更多个基因的一个或更多个基因座,用于检测手术之后残留的癌症。这种检测可以比现有癌症检测方法可能的检测更早。在一些实施方案中,组中的一个或更多个区域包括来自一个或更多个基因的一个或更多个基因座,用于检测高风险患者群体的癌症。例如,吸烟者患肺癌的比率比一般群体高得多。此外,吸烟者还可能发展其他肺部状况,这使癌症检测更加困难,诸如肺中不

规则结节的发展。在一些实施方案中,本文描述的方法比现有的癌症检测方法可能的检测更早地检测高风险患者的癌症。

[0183] 可以选择待包括在测序组中的区域,所述选择基于在该基因或区域中具有肿瘤标志物的患有癌症的受试者的数目。可以基于患有癌症的受试者的患病率和该基因中存在的肿瘤标志物来选择待包括在测序组中的区域。区域中肿瘤标志物的存在可以指示受试者患有癌症。

[0184] 在一些情况下,可以使用来自一个或更多个数据库的信息来选择组。关于癌症的信息可以从癌症肿瘤活组织检查或cfDNA测定获得。数据库可以包括描述被测序的肿瘤样品群体的信息。数据库可以包括关于肿瘤样品中mRNA表达的信息。数据库可以包括关于肿瘤样品中调控元件的信息。与被测序的肿瘤样品相关的信息可以包括各种遗传变异的频率,并描述遗传变异发生的基因或区域。遗传变异可以是肿瘤标志物。这种数据库的非限制性实例是COSMIC。COSMIC是在多种癌症中发现的体细胞突变的目录。对于特定的癌症,COSMIC基于突变频率将基因排名。可以通过在给定基因中具有高突变频率来选择基因以包括在组中。例如,COSMIC表明,33%的被测序的乳腺癌样品群体在TP53中具有突变,并且22%的被取样的乳腺癌群体在KRAS中具突变。其他被排名的基因,包括APC,具有仅见于约4%的被测序乳腺癌样品群体中的突变。基于在被取样的乳腺癌中具有相对高的频率(例如,与以约4%的频率存在的APC相比),TP53和KRAS可以被包括在测序组中。COSMIC被作为非限制性实例提供,然而,可以使用将癌症与位于基因或遗传区域中的肿瘤标志物相关联的任何数据库或信息集。在另一个实例中,如由COSMIC提供的,在1156个胆管癌样品中,380个样品(33%)在TP53中携带突变。其他若干基因,诸如APC,在所有样品中的4%-8%中具有突变。因此,可以基于在胆管癌样品群体中相对高的频率来选择TP53以包括在组中。

[0185] 可以为组选择肿瘤标志物在被取样的肿瘤组织或循环肿瘤DNA中的频率显著高于在给定背景群体中发现的频率的基因或区域。可以选择将多个区域的组合包括在组中,使得至少大多数患有癌症的受试者将在该组中的至少一个区域或基因中存在肿瘤标志物。区域的组合可以基于对于特定的癌症或一组癌症指示大多数受试者在所选择的区域的一个或更多个中具有一种或更多种肿瘤标志物的数据来选择。例如,为了检测癌症1,可以基于指示90%患有癌症1的受试者在组的区域A、B、C和/或D中具有肿瘤标志物的数据来选择包括区域A、B、C和/或D的组。可选地,可能显示肿瘤标志物独立地存在于患有癌症的受试者的两个或更多个区域中,使得两个或更多个区域组合中的肿瘤标志物存在于大多数患有癌症的受试者群体中。例如,为了检测癌症2,可以基于指示90%的受试者在一个或更多个区域中具有肿瘤标志物,并且在30%的这样的受试者中仅在区域X中检测到肿瘤标志物,而对于被检测出肿瘤标志物的其余受试者仅在区域Y和/或Z中检测出肿瘤标志物的数据来选择包括区域X、Y和Z的组。如果50%或更多时间在这些区域的一个或更多个中检测出肿瘤标志物,则在一个或更多个区域中存在的先前显示与一种或更多种癌症相关的肿瘤标志物可以指示或预测受试者患有癌症。在对于一个或更多个区域中的一组肿瘤标志物已知癌症频率的情况下,可以使用计算方法,诸如采用检测癌症的条件概率的模型,来预测哪些区域单独或组合可以预测癌症。用于组选择的其他方法包括使用数据库,所述数据库描述来自采用以大组和/或全基因组测序(WGS、RNA-Seq、芯片-Seq、亚硫酸氢盐测序、ATAC序列及其他)进行的肿瘤综合基因组谱分析的研究的信息。从文献收集的信息也可以描述某些癌症中通常

受影响及突变的途径。通过使用描述遗传信息的知识本体,可以进一步为组选择提供信息。

[0186] 被包括在测序组中的基因可以包括完全转录区、启动子区、增强子区、调控元件和/或下游序列。为了进一步增加检测指示肿瘤的突变的可能性,可以仅将外显子包括在组中。组可以包括选择的基因的所有外显子,或者仅包括选择的基因的一个或更多个外显子。组可以包括来自多于一个不同基因中每一个的外显子。组可以包括来自多于一个不同基因中每一个的至少一个外显子。

[0187] 在一些方面中,选择来自多于一个不同基因中每一个的外显子的组,使得确定比例的患有癌症的受试者在外显子的组中的至少一个外显子中显示出遗传变异。

[0188] 可以对来自基因组中每一个不同基因中的至少一种完整外显子进行测序。被测序的组可以包括来自多于一个基因的外显子。组可以包括来自2个至100个不同基因、来自2个至70个基因、来自2个至50个基因、来自2个至30个基因、来自2个至15个基因、或来自2个至10个基因的外显子。

[0189] 选择的组可以包括不同数目的外显子。组可以包括从2个至3000个外显子。组可以包括从2个至1000个外显子。组可以包括从2个至500个外显子。组可以包括从2个至100个外显子。组可以包括从2个至50个外显子。组可以包括不多于300个外显子。组可以包括不多于200个外显子。组可以包括不多于100个外显子。组可以包括不多于50个外显子。组可以包括不多于40个外显子。组可以包括不多于30个外显子。组可以包括不多于25个外显子。组可以包括不多于20个外显子。组可以包括不多于15个外显子。组可以包括不多于10个外显子。组可以包括不多于9个外显子。组可以包括不多于8个外显子。组可以包括不多于7个外显子。

[0190] 组可以包括来自多于一个不同基因的一个或更多个外显子。组可以包括来自多于一个不同基因的一部分中的每一个的一个或更多个外显子。组可以包括来自不同基因的每一个的至少25%、50%、75%或90%中的至少两个外显子。组可以包括来自不同基因的每一个的至少25%、50%、75%或90%中的至少三个外显子。组可以包括来自不同基因的每一个的至少25%、50%、75%或90%中的至少四个外显子。

[0191] 测序组的尺寸可能有所不同。测序组可以根据几个因素而变得更大或更小(就核苷酸尺寸而言),这些因素包括,例如,被测序的核苷酸总量或针对组中特定区域测序的独特分子的数目。测序组的尺寸可以是5kb至50kb。测序组的尺寸可以是10kb至30kb。测序组的尺寸可以是12kb至20kb。测序组的尺寸可以是12kb至60kb。测序组的尺寸可以是至少10kb、12kb、15kb、20kb、25kb、30kb、35kb、40kb、45kb、50kb、60kb、70kb、80kb、90kb、100kb、110kb、120kb、130kb、140kb或150kb。测序组的尺寸可以小于100kb、90kb、80kb、70kb、60kb或50kb。

[0192] 选择用于测序的组可以包括至少1个、5个、10个、15个、20个、25个、30个、40个、50个、60个、80个或100个区域。在一些情况下,选择组中区域的尺寸相对较小的区域。在一些情况下,组中的区域具有的尺寸为约10kb或更小、约8kb或更小、约6kb或更小、约5kb或更小、约4kb或更小、约3kb或更小、约2.5kb或更小、约2kb或更小、约1.5kb或更小、或约1kb或更小或者更小。在一些情况下,组中的区域具有的尺寸为从约0.5kb至约10kb、从约0.5kb至约6kb、从约1kb至约11kb、从约1kb至约15kb、从约1kb至约20kb、从约0.1kb至约10kb或从约0.2kb至约1kb。例如,组中的区域可以具有的尺寸为从约0.1kb至约5kb。

[0193] 本文选择的组可以允许足以(例如,在从样品获得的无细胞核酸分子中)检测低频

率遗传变异的深度测序。样品中遗传变异的量可以依据给定遗传变异的次要等位基因频率来表示。次要等位基因频率可以指在给定核酸群体诸如样品中次要等位基因(例如,不是最常见的等位基因)出现的频率。以低次要等位基因频率的遗传变异在样品中存在的频率可能相对较低。在一些情况下,组允许检测以至少0.0001%、0.001%、0.005%、0.01%、0.05%、0.1%或0.5%的次要等位基因频率的遗传变异。组可以允许检测以0.001%或更高的次要等位基因频率的遗传变异。组可以允许检测以0.01%或更高的次要等位基因频率的遗传变异。组可以允许检测样品中以低至0.0001%、0.001%、0.005%、0.01%、0.025%、0.05%、0.075%、0.1%、0.25%、0.5%、0.75%或1.0%的频率存在的遗传变异。组可以允许检测样品中以至少0.0001%、0.001%、0.005%、0.01%、0.025%、0.05%、0.075%、0.1%、0.25%、0.5%、0.75%或1.0%的频率存在的肿瘤标志物。组可以允许检测样品中以低至1.0%的频率的肿瘤标志物。组可以允许检测样品中以低至0.75%的频率的肿瘤标志物。组可以允许检测样品中以低至0.5%的频率的肿瘤标志物。组可以允许检测样品中以低至0.25%的频率的肿瘤标志物。组可以允许检测样品中以低至0.1%的频率的肿瘤标志物。组可以允许检测样品中以低至0.075%的频率的肿瘤标志物。组可以允许检测样品中以低至0.05%的频率的肿瘤标志物。组可以允许检测样品中以低至0.025%的频率的肿瘤标志物。组可以允许检测样品中以低至0.01%的频率的肿瘤标志物。组可以允许检测样品中以低至0.005%的频率的肿瘤标志物。组可以允许检测样品中以低至0.001%的频率的肿瘤标志物。组可以允许检测样品中以低至0.0001%的频率的肿瘤标志物。组可以允许在被测序的cfDNA中检测样品中以低至1.0%至0.0001%的频率的肿瘤标志物。组可以允许在被测序的cfDNA中检测样品中以低至0.01%至0.0001%的频率的肿瘤标志物。

[0194] 遗传变异可以以患有疾病(例如,癌症)的受试者群体的百分比来显示。在一些情况下,患有癌症的群体的至少1%、2%、3%、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或99%在组中的至少一个区域中显示出一种或更多种遗传变异。例如,患有癌症的群体的至少80%可以在组中的至少一个区域中显示出一种或更多种遗传变异。

[0195] 组可以包括来自一个或更多个基因中的每一个的一个或更多个区域。在一些情况下,组可以包括来自至少1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、15个、20个、25个、30个、40个、50个或80个基因中的每一个的一个或更多个区域。在一些情况下,组可以包括来自至多1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、15个、20个、25个、30个、40个、50个或80个基因中的每一个的一个或更多个区域。在一些情况下,组可以包括来自从约1个至约80个、从1个至约50个、从约3个至约40个、从5个至约30个、从10个至约20个不同基因中的每一个的一个或更多个区域。

[0196] 可以选择组中的多个区域以检测到一个或更多个表观遗传修饰区域。一个或更多个表观遗传修饰的区域可以被乙酰化、甲基化、泛素化、磷酸化、sumo化、核糖基化和/或瓜氨酸化。例如,可以选择组中的多个区域以检测到一个或更多个甲基化区域。

[0197] 可以选择组中它们包括跨一种或更多种组织差异性转录的序列的区域。在一些情况下,这些区域可以包括在某些组织中相比于其他组织以较高水平转录的序列。例如,这些区域可以包括在某些组织中转录但在其他组织中不转录的序列。

[0198] 组中的区域可以包括编码序列和/或非编码序列。例如,组中的区域可以包括外显子、内含子、启动子、3'非翻译区、5'非翻译区、调控元件、转录起始位点和/或剪接位点中的

一个或更多个序列。在一些情况下,组中的区域可以包括其他非编码序列,所述其他非编码序列包括假基因、重复序列、转座子、病毒元件和端粒。在一些情况下,组中的区域可以包括非编码RNA中的序列,所述非编码RNA例如,核糖体RNA、转移RNA、Piwi相互作用RNA和微RNA。

[0199] 可以选择组中的多个区域以以期望的灵敏度水平检测(诊断)癌症(例如,通过检测一种或更多种遗传变异)。例如,可以选择组中的多个区域,以以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的灵敏度检测癌症(例如,通过检测一种或更多种遗传变异)的区域。可以选择组中的多个区域以以100%的灵敏度检测癌症。

[0200] 为了以期望的特异性水平检测(诊断)癌症(例如,通过检测一种或更多种遗传变异)可以选择组中的多个区域。例如,可以选择组中的多个区域以以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的特异性检测癌症(例如,通过检测一种或更多种遗传变异)。可以选择组中的多个区域以100%的特异性检测一种或更多种遗传变异。

[0201] 可以选择组中的多个区域以以期望的阳性预测值检测(诊断)癌症。阳性预测值可以通过增加灵敏度(例如,检测到实际阳性的几率)和/或特异性(例如,不将实际阴性误认为阳性的几率)来增加。作为非限制性实例,可以选择组中的多个区域以以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的阳性预测值检测一种或更多种遗传变异。可以选择组中的多个区域以以100%的阳性预测值检测一种或更多种遗传变异。

[0202] 可以选择组中的多个区域以以期望的准确度检测(诊断)癌症。如本文使用的,术语“准确度”可以指测试区分疾病状况(例如,癌症)和健康的能力。准确度可以使用量度诸如灵敏度和特异性、预测值、似然比、ROC曲线下面积、约登指数和/或诊断比值比来定量。

[0203] 准确度可以表示为百分比,是指给出正确结果的测试数和进行的测试总数之间的比率。可以选择组中以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的准确度检测癌症的区域。可以选择组中以100%的准确度检测癌症的区域。

[0204] 可以选择高度灵敏并检测低频率遗传变异的组。例如,可以选择可以以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的灵敏度检测样品中以低至0.01%、0.05%或0.001%的频率存在的遗传变异或肿瘤标志物的组。可以选择组中以70%或更高的灵敏度检测样品中以1%或更低的频率存在的肿瘤标志物的区域。可以选择以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的灵敏度检测样品中以低至0.1%的频率的肿瘤标志物的组。可以选择以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的灵敏度检测样品中以低至0.01%的频率的肿瘤标志物的组。可以选择以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%或99.9%的灵敏度检测样品中以低至0.001%的频率的肿瘤标志物的组。

[0205] 可以选择高度特异性并检测低频率遗传变异的组。例如,可以选择可以以至少50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、96%、97%、98%、99%、99.5%

或99.9％的特异性检测样品中以低至0.01％、0.05％或0.001％的频率存在的遗传变异或肿瘤标志物的组。可以选择组中以70％或更高的特异性检测样品中以1％或更低的频率存在的肿瘤标志物的区域。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的特异性检测样品中以低至0.1％的频率的肿瘤标志物的组。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的特异性检测样品中以低至0.01％的频率的肿瘤标志物的组。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的特异性检测样品中以低至0.001％的频率的肿瘤标志物的组。

[0206] 可以选择高度准确度并检测低频率遗传变异的组。可以选择可以以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的准确度检测样品中以低至0.01％、0.05％或0.001％的频率存在的遗传变异或肿瘤标志物的组。可以选择组中以70％或更高的准确度检测样品中以1％或更低的频率存在的肿瘤标志物的区域。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的准确度检测样品中以低至0.1％的频率的肿瘤标志物的组。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的准确度检测样品中以低至0.01％的频率的肿瘤标志物的组。可以选择以至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的准确度检测样品中以低至0.001％的频率的肿瘤标志物的组。

[0207] 可以选择高度预测并检测低频率遗传变异的组。可以选择样品中以低至0.01％、0.05％或0.001％的频率存在的遗传变异或肿瘤标志物可以具有至少70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.5％或99.9％的阳性预测值的组。

[0208] 组中使用的探针或诱饵的浓度可以增加(2ng/μL至6ng/μL),以捕获样品中更多的核酸分子。组中使用的探针或诱饵的浓度可以为至少2ng/μL、3ng/μL、4ng/μL、5ng/μL、6ng/μL或更高。探针的浓度可以为约2ng/μL至约3ng/μL、约2ng/μL至约4ng/μL、约2ng/μL至约5ng/μL、约2ng/μL至约6ng/μL。组中使用的探针或诱饵的浓度可以为2ng/μL或更高至6ng/μL或更低。在一些情况下,这可以允许生物样品中的更多分子被分析,从而能够使较低频率的等位基因被检测到。

[0209] 测序深度

[0210] 从cfDNA分子样品富集的DNA可以被以多种读段深度进行测序,以检测样品中的低频率遗传变异。对于给定位置,读段深度可以指来自样品映射至位置的所有分子包括原始分子和通过扩增原始分子产生的分子的所有读段的数目。因此,例如,50,000个读段的读段深度可以指来自5,000个分子的读段数目,其中每个分子10个读段。映射至位置的原始分子可以是独特且非冗余的(例如,非扩增的样品cfDNA)。

[0211] 为了评估样品分子在给定位置处的读段深度,可以跟踪样品分子。分子跟踪技术可以包括用于标记DNA分子的各种技术,诸如加条形码标签以独特地鉴定样品中的DNA分子的技术。例如,可以将一个或更多个独特的条形码序列附接至样品cfDNA分子的一个或更多个末端。在确定给定位置处的读段深度时,映射至该位置的加不同条形码标签的cfDNA分子的数目可以指示该位置的读段深度。在另一个实例中,样品cfDNA分子的两个末端可以用八种条形码序列中的一种来加标签。给定位置处的读段深度可以通过定量给定位置处的原始

cfDNA分子的数目来确定,例如,通过叠并(cllapse)来自扩增的冗余读段并基于条形码标签和内源序列信息来鉴定独特的分子。

[0212] DNA可以被测序至以下读段深度:每个碱基至少3,000个读段、每个碱基至少4,000个读段、每个碱基至少5,000个读段、每个碱基至少6,000个读段、每个碱基至少7,000个读段、每个碱基至少8,000个读段、每个碱基至少9,000个读段、每个碱基至少10,000个读段、每个碱基至少15,000个读段、每个碱基至少20,000个读段、每个碱基至少25,000个读段、每个碱基至少30,000个读段、每个碱基至少40,000个读段、每个碱基至少50,000个读段、每个碱基至少60,000个读段、每个碱基至少70,000个读段、每个碱基至少80,000个读段、每个碱基至少90,000个读段、每个碱基至少100,000个读段、每个碱基至少110,000个读段、每个碱基至少120,000个读段、每个碱基至少130,000个读段、每个碱基至少140,000个读段、每个碱基至少150,000个读段、每个碱基至少160,000个读段、每个碱基至少170,000个读段、每个碱基至少180,000个读段、每个碱基至少190,000个读段、每个碱基至少200,000个读段、每个碱基至少250,000个读段、每个碱基至少500,000个读段、每个碱基至少1,000,000个读段、或每个碱基至少2,000,000个读段。DNA可以被测序至以下读段深度:每个碱基约3,000个读段、每个碱基约4,000个读段、每个碱基约5,000个读段、每个碱基约6,000个读段、每个碱基约7,000个读段、每个碱基约8,000个读段、每个碱基约9,000个读段、每个碱基约10,000个读段、每个碱基约15,000个读段、每个碱基约20,000个读段、每个碱基约25,000个读段、每个碱基约30,000个读段、每个碱基约40,000个读段、每个碱基约50,000个读段、每个碱基约60,000个读段、每个碱基约70,000个读段、每个碱基约80,000个读段、每个碱基约90,000个读段、每个碱基约100,000个读段、每个碱基约110,000个读段、每个碱基约120,000个读段、每个碱基约130,000个读段、每个碱基约140,000个读段、每个碱基约150,000个读段、每个碱基约160,000个读段、每个碱基约170,000个读段、每个碱基约180,000个读段、每个碱基约190,000个读段、每个碱基约200,000个读段、每个碱基约250,000个读段、每个碱基约500,000个读段、每个碱基约1,000,000个读段、或每个碱基约2,000,000个读段。DNA可以被测序至以下读段深度:每个碱基从约10,000个至约30,000个读段、每个碱基10,000个至约50,000个读段、每个碱基10,000个至约5,000,000个读段、每个碱基50,000个至约3,000,000个读段、每个碱基100,000个至约2,000,000个读段、或每个碱基约500,000个至约1,000,000个读段。在一些实施方案中,DNA可以针对选自以下的组尺寸被测序至任何上文读段深度:小于70,000个碱基、小于65,000个碱基、小于60,000个碱基、小于55,000个碱基、小于50,000个碱基、小于45,000个碱基、小于40,000个碱基、小于35,000个碱基、小于30,000个碱基、小于25,000个碱基、小于20,000个碱基、小于15,000个碱基、小于10,000个碱基、小于5,000个碱基、以及小于1,000个碱基。例如,组的读段总数可以低至600,000个(对于1,000个碱基,每个碱基3,000个读段),并且高至$1.4x\ 10^{11}$个(对于70,000个碱基,每个碱基2,000,000个读段)。在一些实施方案中,DNA可以针对选自以下的组尺寸被测序至任何上文读段深度:5,000个碱基至70,000个碱基,5,000个碱基至60,000个碱基,10,000个碱基至70,000个碱基,或10,000个碱基至60,000个碱基。

[0213] 读段覆盖度可以包括来自核酸分子的一条链或两条链的读段。例如,读段覆盖度可以包括来自样品映射至组中每一个核苷酸的至少5,000个、至少10,000个、至少15,000个、至少20,000个、至少25,000个、至少30,000个、至少35,000个、至少40,000个、至少45,

000个或至少50,000个DNA分子的两条链的读段。

[0214] 在碱基读段量固定的情况下,可以选择一个组以针对期望的读段深度进行优化。

[0215] 加标签

[0216] 在本公开的一些实施方案中,在测序之前制备核酸文库。例如,基因组核酸样品(例如,基因组DNA样品)中的单独的多核苷酸片段可以通过用非独特标识符加标签,例如,对单独的多核苷酸片段非独特地加标签,而独特地鉴定。在一些实施方案中,核酸分子相对于彼此被非独特地加标签。

[0217] 本文公开的多核苷酸可以被加标签。例如,双链多核苷酸可以用双链体标签来加标签,这些标签差异性标记双链分子的互补链(即"沃森"链和"克里克"链)。在一些情况下,双链体标签是具有互补和非互补部分的多核苷酸。

[0218] 标签可以是附接至多核苷酸的任何类型的分子,包括但不限于核酸、化学化合物、荧光探针或放射性探针。标签也可以是寡核苷酸(例如,DNA或RNA)。标签可以包括已知序列、未知序列或两者。标签可以包括随机序列、预定序列或两者。标签可以是双链的或单链的。双链标签可以是双链体标签。双链标签可以包含两条互补链。可选地,双链标签可以包含杂交部分和非杂交部分。双链标签可以是Y形的,例如,杂交部分处于标签的一个末端处,而非杂交部分处于标签的相对末端处。一个这样的实例是Illumina测序中使用的"Y衔接子"。其他实例包括发夹形衔接子或泡形衔接子。泡形衔接子具有两侧上都侧接互补序列的非互补序列。在一些实施方案中,Y形衔接子包含长度为2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个或32个核苷酸的条形码。在一些组合中,这可以与平端修复及连接相结合。

[0219] 不同标签的数目可以大于样品中的分子的估计或预定数目。例如,对于独特的加标签,可以使用样品中分子的估计或预定数目的至少两倍的不同标签。

[0220] 用于对集合中的分子加标签的不同鉴定标签的数目的范围可以,例如,在该范围的下限处的2、3、4、5、6、7、8、9、10、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48或49中的任一个和在范围的上限处的50、100、500、1000、5000和10,000中的任一个之间。用于对集合中的分子加标签的鉴定标签的数目可以是至少2、3、4、5、6、7、8、9、10、15、20、25、30、35、40、45、50、55、60或更多。因此,例如,从1千亿个至1万亿个分子的集合可以用从4个至100个不同的鉴定标签来加标签。从1千亿个至1万亿个分子的集合可以用从8个至10,000个不同的鉴定标签来加标签。从1千亿个至1万亿个分子的集合可以用从16个至10,000个不同的鉴定标签来加标签。从1千亿个至1万亿个分子的集合可以用从16个至5,000个不同的鉴定标签来加标签。从1千亿个至1万亿个分子的集合可以用从16个至1,000个不同的鉴定标签来加标签。

[0221] 如果分子的集合中存在的分子多于标签,则可以认为该分子集合是"非独特地加标签的"。如果分子的集合中的分子的至少1%、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%或至少50%或约50%中的每一个都携带被该集合中的至少一个其他分子所共享的识别标签("非独特标签"或"非独特标识符"),则可以认为该集合是"非独特地加标签的"。标识符可以包含一个条形码或两个条形码。通过用少于群体中核酸分子总数的标签对核酸分子加标签,可以对核酸分子群体非独特地加标

签。对于非独特地加标签的群体,不多于1%、5%、10%、15%、20%、25%、30%、35%、40%、45%或50%的分子可以被独特地加标签。在一些实施方案中,核酸分子通过非独特标签与来自序列读段的起始位置和终止位置或序列的组合来鉴定。在一些实施方案中,被测序的核酸分子的数目小于或等于标识符与起始位置和终止位置或序列的组合的数目。

[0222] 在一些情况下,本文的标签包含分子条形码。这种分子条形码可以用于区分样品中的多核苷酸。分子条形码可以彼此不同。例如,分子条形码其之间可以具有差异,所述差异可以表征为预定的编辑距离或汉明距离。在一些情况下,本文的分子条形码具有1、2、3、4、5、6、7、8、9或10的最小编辑距离。为了进一步提高未加标签分子向加标签分子的转化(例如,加标签)效率,人们利用短标签。例如,文库衔接子标签的长度可以为多达65个、60个、55个、50个、45个、40个或35个核苷酸碱基。这种短文库条形码的集合可以包括一定数目的不同的分子条形码,例如,至少2个、4个、6个、8个、10个、12个、14个、16个、18个或20个不同的条形码,具有1、2、3或更大的最小编辑距离。

[0223] 因此,分子的集合可以包括一个或更多个标签。在一些情况下,集合中的一些分子可以包含识别标签(“标识符”),诸如不被该集合中的任何其他分子所共享的分子条形码。例如,在分子的集合的一些情况下,该集合中的分子的100%或至少50%、60%、70%、80%、90%、95%、97%、98%或99%可以包含不被该集合中的任何其他分子所共享的标识符或分子条形码。如本文使用的,如果分子的集合中的分子的至少95%中的每一个都携带不被该集合中的任何其他分子所共享的标识符(“独特标签”或“独特标识符”),则认为该集合是“独特地加标签的”。在一些实施方案中,核酸分子相对于彼此被独特地加标签。如果分子的集合中的分子的至少1%、5%、10%、15%、20%、25%、30%、35%、40%、45%或50%中的每一个都携带被该集合中的至少一个其他分子所共享的识别标签或分子条形码(“非独特标签”或“非独特标识符”),则认为该集合是“非独特地加标签的”。在一些实施方案中,核酸分子相对于彼此被非独特地加标签。相应地,在非独特地加标签的群体中,不多于1%的分子被独特地加标签。例如,在非独特地加标签的群体中,不多于1%、5%、10%、15%、20%、25%、30%、35%、40%、45%或50%的分子可以被独特地加标签。

[0224] 基于样品中分子的估计数目,可以使用一定数目的不同标签。在一些加标签方法中,不同标签的数目可以与样品中分子的估计数目至少相同。在其他加标签方法中,不同标签的数目可以是样品中分子的估计数目的至少二倍、三倍、四倍、五倍、六倍、七倍、八倍、九倍、十倍、一百倍或一千倍。在独特的加标签中,可以使用样品中分子的估计数目的至少两倍(或更多)的不同标签。

[0225] 多核苷酸片段(在加标签之前)可以包含任何长度的序列。例如,多核苷酸片段(在加标签之前)可以包含至少50个、55个、60个、65个、70个、75个、80个、85个、90个、95个、100个、105个、110个、115个、120个、125个、130个、135个、140个、145个、150个、155个、160个、165个、170个、175个、180个、185个、190个、195个、200个、205个、210个、215个、220个、225个、230个、235个、240个、245个、250个、255个、260个、265个、270个、275个、280个、285个、290个、295个、300个、400个、500个、600个、700个、800个、900个、1000个、1100个、1200个、1300个、1400个、1500个、1600个、1700个、1800个、1900个、2000个或更多个核苷酸的长度。多核苷酸片段可以约为无细胞DNA的平均长度。例如,多核苷酸片段可以包含约160个碱基的长度。多核苷酸片段也可以从较大片段被片段化成长度为约160个碱基的较小片段。

[0226] 可以实现对测序的改进,只要复制物或同源物多核苷酸中的至少一些相对于彼此携带独特的标识符,即携带不同的标签即可。然而,在某些实施方案中,选择所使用的标签的数目,使得在任一位置处起始的所有复制物分子携带独特标识符的机会至少为95%。例如,在包含约10,000个单倍体人类基因组当量的片段化基因组DNA例如cfDNA的样品中,预期z在2和8之间。这样的群体可以用约10个和100个之间不同的标识符,例如,约2个标识符、约4个标识符、约9个标识符、约16个标识符、约25个标识符、约36个不同的标识符、约49个不同的标识符、约64个不同的标识符、约81个不同的标识符、或约100个不同的标识符来加标签。

[0227] 具有可识别序列的核酸条形码,包括分子条形码,可以用于加标签。例如,多于一个DNA条形码可以包含多个数目的核苷酸序列。可以使用具有2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个或更多个可识别核苷酸序列的多于一个DNA条形码。当仅附接至多核苷酸的一个末端时,多于一个DNA条形码可以产生2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个或更多个不同的标识符。可选地,当附接至多核苷酸的两个末端时,个多于一个DNA条形码可以产生4个、9个、16个、25个、36个、49个、64个、81个、100个、121个、144个、169个、196个、225个、256个、289个、324个、361个、400个或更多个不同的标识符(当DNA条形码仅附接至多核苷酸的一个末端时为^2)。在一个实例中,可以使用具有6个、7个、8个、9个或10个可识别核苷酸序列的个多于一个DNA条形码。当附接至多核苷酸的两个末端时,它们分别产生36个、49个、64个、81个或100个可能的不同标识符。在一个特定实例中,多于一个DNA条形码可以包含8个可识别核苷酸序列。当仅附接至多核苷酸的一个末端时,多于一个DNA条形码可以产生8个不同的标识符。可选地,当附接至多核苷酸的两个末端时,多于一个DNA条形码可以产生64个不同的标识符。以这个方式加标签的样品可以是具有约10ng至约200ng、约1μg、约10μg中的任一个的范围的片段化多核苷酸,例如基因组DNA,例如cfDNA的样品。

[0228] 多核苷酸可以以多种方式来独特地鉴定。多核苷酸可以通过独特的条形码来独特地鉴定。例如,样品中任何两种多核苷酸都附接两个不同的条形码。条形码可以是DNA条形码或RNA条形码。例如,条形码可以是DNA条形码。

[0229] 可选地,多核苷酸可以通过条形码和多核苷酸的一个或更多个内源序列的组合来独特地鉴定。条形码可以是非独特标签或独特标签。在一些情况下,条形码是非独特标签。例如,样品中任何两种多核苷酸可以附接至包含相同条形码的条形码,但是这两种多核苷酸仍然可以通过不同的内源序列来鉴定。这两种多核苷酸可以通过不同的内源序列中的信息来鉴定。这些信息包括内源序列或其一部分的序列、内源序列的长度、内源序列的位置、内源序列的一种或更多种表观遗传修饰或内源序列的任何其他特征。在一些实施方案中,多核苷酸可以通过标识符(包含一个条形码或包含两个条形码)与来自序列读段的起始序列和终止序列组合来鉴定。

[0230] 非独特标签和内源序列信息的组合可以用于明确检测核酸分子。例如,来自样品的非独特地加标签的核酸分子(“亲本多核苷酸”)可以被扩增以产生子代多核苷酸。然后可以对亲本多核苷酸和子代多核苷酸测序以产生序列读段。为了减少误差,可以叠并序列读

段以生成一组共有序列。为了生成共有序列,可以基于非独特标签中的序列信息和内源序列信息来叠并序列读段,所述内源序列信息包括序列读段开始区域处的序列信息、序列读段结束区域处的序列信息和序列读段的长度。在一些实施方案中,共有序列通过环测序来产生,其中相同的核酸链在滚环中被多次测序以获得共有序列。共有序列可以在逐个分子的基础上来确定(其中共有序列在一段碱基上来确定),或在逐个碱基的基础上来确定(其中共有核苷酸针对给定位置处的碱基来确定)。在一些实施方案中,概率模型被构建来对扩增和测序误差谱进行建模,并用于估计分子的每一个位置中的真实核苷酸的概率。在一些实施方案中,概率模型参数的估计基于被一起处理的单独样品或一批样品或参考样品组中观察到的误差谱来更新。在一些实施方案中,使用对来自受试者的单独cfNA(例如,cfDNA)分子加标签的条形码来确定共有序列。

[0231] 内源序列可以处于多核苷酸的末端上。例如,内源序列可以与附接的条形码相邻(例如,之间有碱基)。在一些情况下,内源序列的长度可以是至少2个、4个、6个、8个、10个、20个、30个、40个、50个、60个、70个、80个、90个或100个碱基。内源序列可以是待分析片段/多核苷酸的末端序列。内源序列可以是一定长度的序列。例如,包括8个不同条形码的多于一个条形码可以附接至样品中每一个多核苷酸的两个末端。样品中的每一个多核苷酸可以通过条形码和多核苷酸末端上的约10个碱基对内源序列的组合来鉴定。不囿于理论,多核苷酸的内源序列也可以是整个多核苷酸序列。

[0232] 本文还公开了加标签的多核苷酸的组合物。加标签的多核苷酸可以是单链的。可选地,加标签的多核苷酸可以是双链的(例如,加标签的多核苷酸双链体)。相应地,本公开内容还提供了加标签的多核苷酸双链体的组合物。多核苷酸可以包括任何类型的核酸(DNA和/或RNA)。多核苷酸包括本文公开的任何类型的DNA。例如,多核苷酸可以包括DNA,例如,片段化的DNA或cfDNA。组合物中映射至基因组中的可映射的碱基位置的一组多核苷酸可以被非独特地加标签,即,不同标识符的数目可以是至少2并且少于映射至可映射的碱基位置的多核苷酸的数目。不同标识符的数目也可以是至少3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25,并且少于映射至可映射的碱基位置的多核苷酸的数目。

[0233] 在一些情况下,当组合物从约1ng至约10μg或更高时,可以使用更大组的不同的分子条形码。例如,5个和100个之间的不同的文库衔接子可以用于对cfDNA样品中的多核苷酸加标签。

[0234] 可以将分子条形码分配至本公开内容中公开的任何类型的多核苷酸。例如,可以将分子条形码分配至无细胞多核苷酸(例如,cfDNA)。通常,本文公开的标识符可以是用于对多核苷酸加标签的条形码寡核苷酸。条形码标识符可以是核酸寡核苷酸(例如,DNA寡核苷酸)。条形码标识符可以是单链的。可选地,条形码标识符可以是双链的。条形码标识符可以使用本文公开的任何方法被附接至多核苷酸。例如,条形码标识符可以通过使用酶连接被附接至多核苷酸。条形码标识符也可以通过PCR被掺入多核苷酸中。在其他情况下,反应可以包括向分析物直接添加金属同位素或通过用同位素标记的探针添加。通常,在本公开内容的反应物中,独特或非独特标识符或分子条形码的分配可以遵循由例如以下描述的方法和系统:美国专利申请2001/0053519、2003/0152490、2011/0160078和美国专利第6,582,908号,其各自通过引用整体并入本文。

[0235] 本文使用的标识符或分子条形码可以是完全内源的,由此可以进行单独片段的环

连接,随后进行随机剪切或靶向扩增。在这种情况下,分子的新起始点和终止点与原始分子内连接点的组合可以形成特定的标识符。

[0236] 本文使用的标识符或分子条形码可以包括任何类型的寡核苷酸。在一些情况下,标识符可以是预定、随机或半随机序列的寡核苷酸。标识符可以是条形码。例如,可以使用多于一个条形码,使得条形码在所述多于一个条形码中相对于彼此不必是独特的。可选地,可以使用多于一个条形码,使得每一个条形码在所述多于一个条形码中相对任何其他条形码都是独特的。条形码可以包括可以被单独跟踪的特定序列(例如,预定序列)。此外,条形码可以附接(例如,通过连接)至单独的分子,使得条形码与其可以附接的序列的组合产生可以被单独跟踪的特定序列。如本文描述的,条形码的检测与在序列读段的开始(起始)和/或结束(终止)部分的序列数据组合可以允许将独特的身份分配至特定分子。单独序列读段的长度或碱基对数目也可以用于将独特身份分配至此类分子。如本文描述的,来自已经被分配了独特身份的核酸的单链的片段可以从而允许随后鉴定来自该亲本链的片段。以此方式,样品中的多核苷酸可以被独特地或基本上独特地加标签。双链体标签可以包括简并或半简并核苷酸序列,例如,随机简并序列。核苷酸序列可以包含任何数目的核苷酸。例如,核苷酸序列可以包含1个(如果使用非天然核苷酸)、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个或更多个核苷酸。在一个特定实例中,序列可以包含7个核苷酸。在另一个实例中,序列可以包含8个核苷酸。序列也可以包含9个核苷酸。序列可以包含10个核苷酸。

[0237] 条形码可以包含连续性序列或非连续性序列。包含至少1个、2个、3个、4个、5个或更多个核苷酸的条形码是连续性序列或非连续性序列。包含4个核苷酸的条形码是连续的,如果这4个核苷酸没有被任何其他核苷酸打断)。例如,如果条形码包含序列TTGC,那么如果条形码是TTGC,则该条形码是连续性的。另一方面,如果条形码是TTXGC,则该条形码是非连续性的,其中X是核酸碱基。

[0238] 标识符或分子条形码可以具有n-mer序列,所述n-mer序列的长度可以是2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个或更多个核苷酸。本文的标签可以包含任何长度范围的核苷酸。例如,序列的长度可以是在2个至100个、10个至90个、20个至80个、30个至70个、40个至60个之间或约50个核苷酸。条形码群体可以包括相同长度或不同长度的条形码。

[0239] 标签可以包括在标识符或分子条形码下游的双链固定参考序列。可选地,标签可以包括在标识符或分子条形码上游或下游的双链固定参考序列。双链固定参考序列的每一条链的长度可以是,例如,3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个、20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个核苷酸。

[0240] 本文公开的加标签可以使用任何方法来进行。多核苷酸可以通过杂交用衔接子来

加标签。例如,衔接子可以具有与多核苷酸的序列的至少一部分互补的核苷酸序列。作为替代,多核苷酸可以通过连接用衔接子来加标签。

[0241] 条形码或标签可以使用多种技术来附接。附接可以通过包括例如连接(平端或粘端)或退火优化的分子倒置探针的方法来进行。例如,加标签可以包括使用一种或更多种酶。酶可以是连接酶。连接酶可以是DNA连接酶。例如,DNA连接酶可以是T4 DNA连接酶、大肠杆菌(E.coli)DNA连接酶和/或哺乳动物连接酶。哺乳动物连接酶可以是DNA连接酶I、DNA连接酶III或DNA连接酶IV。连接酶也可以是热稳定的连接酶。标签可以被连接至多核苷酸的平端(平端连接)。可选地,标签可以被连接至多核苷酸的粘端(粘端连接)。连接效率可以通过优化多种条件来增加。连接效率可以通过优化连接反应时间来增加。例如,连接反应时间可以少于1小时、2小时、3小时、4小时、5小时、6小时、7小时、8小时、9小时、10小时、11小时、12小时、13小时、14小时、15小时、16小时、17小时、18小时、19小时或20小时。在特定实例中,连接反应时间少于20小时。连接效率可以通过优化反应中的连接酶浓度来增加。例如,连接酶浓度可以为至少10单位/微升、50单位/微升、100单位/微升、150单位/微升、200单位/微升、250单位/微升、300单位/微升、400单位/微升、500单位/微升、或600单位/微升。效率也可以通过添加适于连接的酶、酶辅因子或其他添加剂或改变适于连接的酶、酶辅因子或其他添加剂的浓度和/或优化具有酶的溶液的温度来优化。效率还可以通过改变反应的多种组分的添加顺序来优化。标签序列的末端可以包含二核苷酸来增加连接效率。当标签包含非互补部分(例如,Y形衔接子)时,标签衔接子的互补部分上的序列可以包含一个或更多个增加连接效率的选择的序列。这种序列位于标签的末端处。这种序列可以包含1个、2个、3个、4个、5个或6个末端碱基。也可以使用具有高粘度(例如,低雷诺数)的反应溶液来增加连接效率。例如,溶液可以具有的雷诺数小于3000、2000、1000、900、800、700、600、500、400、300、200、100、50、25或10。还设想了可以使用大致统一分布的片段(例如,严格的标准差)来增加连接效率。例如,片段尺寸的变化可以相差(vary by)小于20%、15%、10%、5%或1%。加标签也可以包括引物延伸,例如通过聚合酶链式反应(PCR)。加标签也可以包括基于连接的PCR、多重PCR、单链连接或单链环化中的任何一种。加标签(例如,通过连接)的效率可以增加至对至少20%、至少30%、至少40%、至少50%、至少60%、至少70%、至少80%、至少90%、至少95%或至少98%的分子加标签的效率(转化效率)。

[0242] 可以进行连接反应,在该连接反应中,样品中的亲本多核苷酸与包含y个不同条形码寡核苷酸的反应混合物混合,其中y=n的平方根。该连接可以导致条形码寡核苷酸与样品中的亲本多核苷酸的随机附接。然后可以将反应混合物在足以实现条形码寡核苷酸与样品的亲本多核苷酸连接的连接条件下孵育。在一些实施方案中,将选自y个不同条形码寡核苷酸的随机条形码与亲本多核苷酸的两个末端连接。y个条形码与亲本多核苷酸的一个或两个末端的随机连接可以导致产生y^2个独特标识符。例如,包含约10,000个单倍体人类基因组当量的cfDNA的样品可以用约36个独特标识符来加标签。独特标识符可以包含6个独特DNA条形码。6个独特条形码与多核苷酸的两个末端连接可以导致产生36个可能的独特标识符。

[0243] 在一些实施方案中,包含约10,000个单倍体人类基因组当量的DNA的样品用一定数目的独特标识符来加标签,所述一定数目的独特标识符通过将一组独特条形码与亲本多核苷酸的两个末端连接来产生。例如,通过将8个独特条形码与亲本多核苷酸的两个末端连

接,可以产生64个独特标识符。同样,通过将10个独特条形码与亲本多核苷酸的两个末端连接可以产生100个独特标识符,通过将15个独特条形码与亲本多核苷酸的两个末端连接可以产生225个独特标识符,通过将20个独特条形码与亲本多核苷酸的两个末端连接可以产生400个独特标识符,通过将25个独特条形码与亲本多核苷酸的两个末端连接可以产生625个独特标识符,通过将30个独特条形码与亲本多核苷酸的两个末端连接可以产生900个独特标识符,通过将35个独特条形码与亲本多核苷酸的两个末端连接可以产生1225个独特标识符,通过将40个独特条形码与亲本多核苷酸的两个末端连接可以产生1600个独特标识符,通过将45个独特条形码与亲本多核苷酸的两个末端连接可以产生2025个独特标识符,并且通过将50个独特条形码与亲本多核苷酸的两个末端连接可以产生2500个独特标识符。反应的连接效率可以超过10%、超过20%、超过30%、超过40%、超过50%、超过60%、超过70%、超过80%或超过90%。连接条件可以包括使用能够结合片段的任一末端并且仍可扩增的双向衔接子。连接条件可以包括粘端连接衔接子,每一个粘端连接衔接子具有至少一个核苷酸碱基的突出端。在一些情况下,连接条件可以包括具有不同碱基突出端的衔接子,以增加连接效率。作为非限制性实例,连接条件可以包括的衔接子具有单碱基胞嘧啶(C)突出端(即,C-加尾的衔接子)、单碱基胸腺嘧啶(T)突出端(T-加尾的衔接子)、单碱基腺嘌呤(A)突出端(A-加尾的衔接子)和/或单碱基鸟嘌呤(G)突出端(G-加尾的衔接子)。连接条件可以包括平端连接,这不同于加尾。连接条件可以包括仔细滴定衔接子和/或条形码寡核苷酸的量。连接条件可以包括使用与反应混合物中的亲本多核苷酸片段的量相比超过2X、超过5X、超过10X、超过20X、超过40X、超过60X、超过80X(例如,~100X)摩尔过量的衔接子和/或条形码寡核苷酸。连接条件可以包括使用T4 DNA连接酶(例如,NEBNExt Ultra Ligation Module)。在一个实例中,将18微升连接酶主混合物与90微升的连接物(90份中的18份)和连接增强剂一起使用。相应地,用n个独特标识符对亲本多核苷酸加标签可以包括使用数目为y的不同条形码,其中y=n的平方根。以此方式加标签的样品可以是这样的样品:其具有范围为约10ng至约100ng、约200ng、约300ng、约400ng、约500ng、约1μg、约10μg中的任一种的片段化多核苷酸,例如,基因组DNA,例如,cfDNA。用于鉴定样品中的亲本多核苷酸的条形码的数目y可以取决于样品中的核酸的量。

[0244] 一种增加转化效率的方法涉及使用针对在单链DNA上的最佳反应性而工程化的连接酶,诸如ThermoPhage单链DNA(ssDNA)连接酶衍生物。此类连接酶绕过文库制备中末端修复和A加尾的传统步骤,该步骤由于中间清洁步骤而可能具有较差的效率和/或累积的损失,并允许有义或反义起始多核苷酸被转化为适当地加标签的多核苷酸的概率加倍。它还转化可能具有突出端的双链多核苷酸,该突出端可能无法通过通常的末端修复反应充分地平端化。用于该ssDNA反应的最佳反应条件为:1x反应缓冲液(50毫摩尔(mM)MOPS(pH 7.5)、1mM DTT、5mM $MgCl_2$、10mM KCl)。在65℃,将200pmol 85nt ssDNA寡聚体和5U ssDNA连接酶与50mM ATP、25mg/ml BSA、2.5mM $MnCl_2$孵育1小时。随后使用PCR扩增可以进一步将加标签的单链文库转化为双链文库并产生远高于20%的总转化效率。将转化率增加至例如高于10%的其他方法包括,例如,单独或组合的以下中的任一种:退火优化的分子倒置探针、具有良好控制的多核苷酸尺寸范围的平端连接、选择高效聚合酶、粘端连接或者使用或不使用融合引物的预先多重扩增步骤、优化靶序列中的末端碱基、优化反应条件(包括反应时间)、及在连接期间引入一个或更多个步骤来清洁反应物(例如,不期望的核酸片段),以及

优化温度或缓冲条件。粘端连接可以使用多个核苷酸突出端来进行。粘端连接可以使用包含A、T、C或G碱基的单个核苷酸突出端进行。

[0245] 本公开内容还提供了加标签的多核苷酸的组合物。多核苷酸可以包含片段化DNA，例如cfDNA。组合物中映射至基因组中的可映射的碱基位置的一组多核苷酸可以被非独特地加标签，即，不同标识符的数目可以是至少2且少于映射至可映射的碱基位置的多核苷酸的数目。约10ng至约10μg之间(例如，约10ng-1μg、约10ng-100ng、约100ng-10μg、约100ng-1μg、约1μg-10μg中的任一种)的组合物可以携带2个、5个、10个、50个或100个中的任一个至100个、1000个、10,000个或100,000个中的任一个之间的不同标识符。例如，5个和100个之间的不同标识符可以用于对此类组合物中的多核苷酸加标签。

[0246] 测序

[0247] 可以对加标签的多核苷酸测序以生成序列读段。例如，可以对加标签的双链体多核苷酸测序。序列读段只能从加标签的双链体多核苷酸的一条链生成。可选地，加标签的双链体多核苷酸的两条链都可以生成序列读段。加标签的双链体多核苷酸的两条链可以包含相同的标签。可选地，加标签的双链体多核苷酸的两条链可以包含不同的标签。当加标签的双链体多核苷酸的两条链被加不同标签时，可以将从一条链(例如，沃森链)生成的序列读段与从另一条链(例如，克里克链)生成的序列读段区分开。测序可以包括对每个分子生成多个序列读段。例如，这在测序过程期间，作为例如通过PCR扩增单独多核苷酸链的结果发生。

[0248] 本文公开的方法可以包括扩增多核苷酸。扩增可以在加标签之前、加标签之后或两者进行。多核苷酸扩增可以导致核苷酸掺入核酸分子或引物中，从而形成与模板核酸互补的新核酸分子。新形成的多核苷酸分子及其模板可以被用作合成另外的多核苷酸的模板。被扩增的多核苷酸可以是任何核酸，例如脱氧核糖核酸，包括基因组DNA、cDNA(互补DNA)、cfDNA和循环肿瘤DNA(ctDNA)。被扩增的多核苷酸也可以是RNA。如本文使用的，一个扩增反应可以包括许多轮的DNA复制。DNA扩增反应可以包括，例如，聚合酶链式反应(PCR)。一个PCR反应可以包括2个-100个变性、退火及合成DNA分子的“循环”。例如，在扩增步骤期间可以进行2个-7个、5个-10个、6个-11个、7个-12个、8个-13个、9个-14个、10个-15个、11个-16个、12个-17个、13个-18个、14个-19个、或15个-20个循环。PCR的条件可以基于包括引物在内的序列的GC含量来优化。可以选择扩增引物以针对感兴趣的靶序列进行选择。可以设计引物来优化转化效率或使转化效率最大化。在一些实施方案中，引物包含在引物之间的短序列，以便拉出感兴趣的小区域。在一些实施方案中，引物靶向核小体区域，使得引物与存在核小体的区域杂交，而不是核小体之间的区域杂交，因为核小体间区域被更高度地裂解而因此不太可能作为靶存在。

[0249] 在一些实施方案中，基因组中被核小体和癌细胞、肿瘤微环境或免疫系统组分(粒细胞、肿瘤浸润淋巴细胞等)中的其他调控机制差异性保护的区域被靶向。在一些实施方案中，在肿瘤细胞中稳定和/或非差异性调控的其他区域被靶向。在这些区域内，覆盖度、裂解位点、片段长度、序列含量、片段终点处的序列含量或附近基因组环境的序列含量的差异可以用于推断某些癌细胞分类(例如，EGFR突变、KRAS突变、ERBb2扩增或PD-1表达癌症)或癌症类型(例如，肺腺癌、乳腺癌或结肠直肠癌)的存在或不存在。这种靶向还可以通过增强某些位点处的覆盖度或捕获概率来增强测定的灵敏度和/或特异性。这些原理适用于靶向方

法,靶向方法包括但不限于:基于连接加杂交捕获的富集、基于扩增的富集、用序列/基因组位置特异性起始引物基于滚环的富集、以及其他方法。可以用这些方法靶向及随后分析的区域包括但不限于:内含子区域、外显子区域、启动子区域、TSS区域、远端调控元件、增强子区域和超增强子区域和/或前述区域的连接点。这些方法也可以与本文描述的用于确定样品中包含的变异(例如,种系变异或体细胞变异)的其他技术组合,用于推断肿瘤的来源组织和/或肿瘤负荷的量度。例如,种系变异可以确定某些癌症类型的易感性,而体细胞变异可以基于受影响的基因、途径和变异百分比而与某些癌症类型相关联。然后,该信息可以和与调控机制和/或化学修饰(诸如,例如,甲基化、羟甲基化、乙酰化)和/或RNA相关的表观遗传特征组合使用。核酸文库可以包括对DNA、DNA修饰和RNA的组合分析,以增强检测癌症、癌症类型、特定疾病中活化的分子途径、来源组织以及对应于肿瘤负荷的量度的灵敏度和特异性。用于分析上文中的每一个的方法已经在别处概述,并且可以组合用于分析来自同一患者的单个或多个样品,其中样品可以来自多种身体样本。

[0250] 核酸扩增技术可以与本文描述的测定一起使用。一些扩增技术是PCR方法,PCR方法可以包括但不限于溶液PCR和原位PCR。例如,扩增可以包括基于PCR的扩增。可选地,扩增可以包括非基于PCR的扩增。模板核酸的扩增可以包括使用一种或更多种聚合酶。例如,聚合酶可以是DNA聚合酶或RNA聚合酶。在一些情况下,高保真度扩增使用诸如高保真度聚合酶(例如,Phusion RTM高保真度DNA聚合酶)或PCR方案来进行。在一些情况下,聚合酶可以是高保真度聚合酶。例如,聚合酶可以是KAPA HiFi DNA聚合酶。聚合酶也可以是Phusion DNA聚合酶或Ultra II聚合酶。聚合酶可以在使由于例如片段长度和/或GC含量引起的扩增偏倚减少或最小化的反应条件下使用。

[0251] 通过PCR扩增多核苷酸的单链将生成该链及其互补序列两者的拷贝。在测序期间,该链及其互补序列两者都将生成序列读段。然而,从例如沃森链的互补序列生成的序列读段可以被这样鉴定,因为它们携带对原始沃森链加标签的双链体标签部分的互补序列。相比之下,从克里克链或其扩增产物生成的序列读段将携带对原始克里克链加标签的双链体标签部分。以此方式,可以将从沃森链的互补序列的扩增产物生成的序列读段与从原始分子的克里克链的扩增产物生成的互补序列读段区分开。

[0252] 扩增,诸如PCR扩增,通常以若干轮来进行。扩增的示例性轮数包括1轮、2轮、3轮、4轮、5轮、6轮、7轮、8轮、9轮、10轮、11轮、12轮、13轮、14轮、15轮、16轮、17轮、18轮、19轮、20轮、25轮、30轮或更多轮的扩增。扩增条件可以例如针对缓冲条件和聚合酶类型及条件来优化。扩增也可以被修改,以减少样品处理中的偏倚,例如,通过减少非特异性扩增偏倚、GC含量偏倚和尺寸偏倚。

[0253] 在一些实施方案中,序列可以在测序之前被富集。富集可以针对特定靶区域来进行或非特异性地进行。在一些实施方案中,感兴趣的靶向基因组区域可以用针对一个或更多个诱饵集组选择的捕获探针(“诱饵”)使用差异性平铺和捕获方案来富集。差异性平铺和捕获方案使用不同相对浓度的诱饵集在与诱饵相关的基因组区域中差异性平铺(例如,以不同的“分辨率”),受一组限制(例如,测序仪限制,诸如测序载量、每种诱饵的效用等),并以下游测序所需的水平捕获它们。这些感兴趣的靶向基因组区域可以包括单核苷酸变异(SNV)和插入/缺失(即,插入或缺失)。感兴趣的靶向基因组区域可以包括感兴趣的骨架基因组区域(“骨架区域”)或感兴趣的热点基因组区域(“热点(hot-spot)区域”或“热点

(hotspot)区域”或者“热点(hot-spot)”或“热点(hotspot)”)。虽然“热点”可以指与序列变异相关的特定基因座,“骨架”区域可以指较大的基因组区域,但是它们每一个都可以具有一个或更多个潜在的序列变异。例如,骨架区域可以是包含一个或更多个癌症相关突变的区域,而热点可以是具有与癌症复发相关的特定突变的基因座或具有与癌症相关的特定复发突变的基因座。感兴趣的骨架和热点基因组区域两者都可以包括通常在液体活组织检查测定中包括的肿瘤相关标志物基因(例如,BRAF、BRCA1/2、EGFR、KRAS、PIK3CA、ROS1、TP53等),对于这些标志物基因,可以预期在患有癌症的受试者中观察到一个或更多个变异。在一些实施方案中,具有针对一个或更多个感兴趣区域的探针的生物素标记的珠可以用于捕获靶序列,任选地随后扩增这些区域,以富集感兴趣区域。

[0254] 可以从样品获得的测序数据的量是有限的,并且受限于诸如核酸模板的质量、靶序列的数目、特定序列的稀缺性、测序技术的限制等因素以及诸如时间和费用的实际考虑。因此,“读段预算”是一种将可以从样品提取的遗传信息的量概念化的方法。每个样品的读段预算可以被选择为其在测序实验中鉴定到待被分配至包含预定量的DNA的测试样品的碱基读段的总数。读段预算可以基于产生的读段总数,例如,包括通过扩增产生的冗余读段。可选地,它可以基于样品中检测到的独特分子的数目。在某些实施方案中,读段预算可以反映支持在一个基因座处的判定的双链的量。即,从DNA分子的两条链检测到读段的基因座百分比。

[0255] 读段预算的因素包括读段深度和组长度。例如,可以将3,000,000,000个读段的读段预算分配为平均读段深度为20,000个读段/碱基的150,000个碱基。读段深度可以指在基因座处产生读段的分子的数目。在本公开内容中,每个碱基处的读段可以被分配为第一平均读段深度的组中的骨架区域中的碱基和更深读段深度的组中的热点区域中的碱基之间。在一些实施方案中,样品被测序至由样品中存在的核酸量确定的读段深度。在一些实施方案中,样品被测序至设置的读段深度,使得包含不同量核酸的样品被测序至相同的读段深度。例如,包含300ng核酸的样品被测序至的读段深度可以是包含30ng核酸的样品的读段深度的1/10。在一些实施方案中,来自两个或更多个不同受试者的核酸可以按基于从每一个受试者获得的核酸量的比被添加在一起。

[0256] 通过非限制性实例的方式,如果读段预算由给定样品的100,000个读段计数组成,则这100,000个读段计数将在骨架区域的读段和热点区域的读段之间来划分。将大量读段(例如,90,000个读段)分配至骨架区域将导致少量读段(例如,剩余的10,000个读段)被分配至热点区域。相反,将大量读段(例如,90,000个读段)分配至热点区域将导致少量读段(例如,剩余的10,000个读段)被分配至骨架区域。因此,技术人员可以分配读段预算来提供期望的灵敏度和特异性水平。在某些实施方案中,读段预算可以在100,000,000个读段和100,000,000,000个读段之间,例如,在500,000,000个读段和50,000,000,000个读段之间,或者在约1,000,000,000个读段和5,000,000,000个读段之间,跨例如20,000个碱基至100,000个碱基。

[0257] 所有多核苷酸(例如,扩增的多核苷酸)可以被提交至测序装置进行测序。可选地,将所有扩增的多核苷酸的取样或子集提交至测序装置进行测序。对于任何原始双链多核苷酸,对于测序可以存在三种结果。第一,序列读段可以从原始分子的两条互补链(即,从沃森链和从克里克链两者)生成。第二,序列读段仅能从两条互补链中的一条生成(即,从沃森链

或从克里克链生成,但不能从两者都生成)。第三,从两条互补链中的任一条都不能生成序列读段。因此,对映射至遗传基因座的独特序列读段计数将低估原始样品中映射至基因座的双链多核苷酸的数目。本文描述了估计未发现和未计数的多核苷酸的方法。

[0258] 测序方法可以是大规模并行测序,即,同时(或以快速相继)对至少100个、1000个、10,000个、100,000个、1百万个、1千万个、1亿个或10亿个多核苷酸分子中的任一个测序。

[0259] 测序方法可以包括但不限于:高通量测序、焦磷酸测序、边合成边测序、单分子测序、纳米孔测序、半导体测序、边连接边测序、边杂交边测序、RNA-Seq(Illumina)、数字基因表达(Helicos)、下一代测序、单分子边合成边测序(SMSS)(Helicos)、大规模并行测序、克隆单分子阵列(Solexa)、鸟枪法测序、Maxim-Gilbert测序或Sanger测序、引物步移、使用PacBio、SOLiD、Ion Torrent、或Nanopore平台测序、及本领域中已知的任何其他测序方法。

[0260] 所述方法可以包括对至少1百万、1千万、1亿、5亿、10亿、11亿、12亿、15亿、20亿、25亿、30亿、35亿、40亿、45亿、50亿、55亿、60亿、65亿、70亿、80亿、90亿或100亿碱基对测序。在一些情况下,所述方法可以包括对从约10亿至约70亿、从约11亿至约68亿、从约12亿至约65亿、从约11至到约64亿、从约15亿至约70亿、从约20亿至约60亿、从约25亿至约55亿、从约30亿至约50亿碱基对测序。例如,所述方法可以包括对从约12亿至约65亿碱基对测序。

[0261] 肿瘤标志物

[0262] 肿瘤标志物是与一种或更多种癌症相关的遗传变异。肿瘤标志物可以使用若干资源或方法中的任一种来确定。肿瘤标志物可以是先前已经发现的,或者可以是使用实验技术或流行病学技术从头发现的。当肿瘤标志物与癌症高度相关时,肿瘤标志物的检测可以指示癌症。当区域或基因中的肿瘤标志物出现的频率大于给定背景群体或数据集的频率时,肿瘤标志物的检测可以指示癌症。

[0263] 公开可得的资源,诸如科学文献和数据库可以详细描述被发现与癌症相关的遗传变异。科学文献可以描述将一个或更多个遗传变异与癌症相关联的实验或全基因组关联研究(GWAS)。数据库可以聚集从诸如科学文献的来源收集的信息,以提供用于确定一种或更多种肿瘤标志物的更全面的资源。数据库的非限制性实例包括FANTOM、GTex、GEO、Body Atlas、INSiGHT、OMIM(在线人类孟德尔遗传(Online Mendelian Inheritance in Man),omim.org)、cBioPortal(cbioportal.org)、CIViC(癌症变异的临床解释,civic.genome.wustl.edu)、DOCM(生物处理突变数据库(Database of Curated Mutations)、docm.genome.wustl.edu)和ICGC数据门户(dcc.icgc.org)。在另外的实例中,COSMIC(癌症体细胞突变目录)数据库允许按癌症、基因或突变类型检索肿瘤标志物。肿瘤标志物也可以通过进行实验,诸如病例对照或关联研究(例如,全基因组关联研究)来从头确定。

[0264] 一种或更多种肿瘤标志物可以在测序组中被检测到。肿瘤标志物可以是与癌症相关的一个或更多个遗传变异。肿瘤标志物可以选自单核苷酸变异(SNV)、拷贝数变异(CNV)、插入或缺失(例如,插入/缺失)、基因融合和倒位。肿瘤标志物可以影响蛋白水平。肿瘤标志物可以位于启动子或增强子中,并且可以改变基因的转录。肿瘤标志物可以影响基因的转录和/或翻译效率。肿瘤标志物可以影响转录的mRNA的稳定性。肿瘤标志物可以导致翻译蛋白的氨基酸序列改变。肿瘤标志物可以影响剪接,可以改变由特定密码子编码的氨基酸,可以导致移码,或可以导致过早终止密码子。肿瘤标志物可以导致氨基酸的保守取代。一种或

更多种肿瘤标志物可导致氨基酸的保守取代。一种或更多种肿瘤标志物可以导致氨基酸的非保守取代。

[0265] 一种或更多种肿瘤标志物可以是驱动突变。驱动突变是这样的突变,它通过增加肿瘤细胞的存活或繁殖,赋予肿瘤细胞在其微环境中的选择优势。肿瘤标志物可以全都不是驱动突变。一种或更多种肿瘤标志物可以是乘客突变(passenger mutation)。乘客突变是对肿瘤细胞的适应性没有影响但可能与克隆扩增相关的突变,因为它与驱动突变发生在同一基因组中。

[0266] 肿瘤标志物的频率可以低至0.001%。肿瘤标志物的频率可以低至0.005%。肿瘤标志物的频率可以低至0.01%。肿瘤标志物的频率可以低至0.02%。肿瘤标志物的频率可以低至0.03%。肿瘤标志物的频率可以低至0.05%。肿瘤标志物的频率可以低至0.1%。肿瘤标志物的频率可以低至1%。

[0267] 多于50%的患有癌症的受试者中可以不存在单一肿瘤标志物。多于40%的患有癌症的受试者中可以不存在单一肿瘤标志物。多于30%的患有癌症的受试者中可以不存在单一肿瘤标志物。多于20%的患有癌症的受试者中可以不存在单一肿瘤标志物。多于10%的患有癌症的受试者中可以不存在单一肿瘤标志物。多于5%的患有癌症的受试者中可以不存在单一肿瘤标志物。0.001%至50%的患有癌症的受试者中可以存在单一肿瘤标志物。0.01%至50%的患有癌症的受试者中可以存在单一肿瘤标志物。0.01%至30%的患有癌症的受试者中可以存在单一肿瘤标志物。0.01%至20%的患有癌症的受试者中可以存在单一肿瘤标志物。0.01%至10%的患有癌症的受试者中可以存在单一肿瘤标志物。0.1%至10%的患有癌症的受试者中可以存在单一肿瘤标志物。0.1%至5%的患有癌症的受试者中可以存在单一肿瘤标志物。

[0268] 肿瘤标志物的检测可以指示一种或更多种癌症的存在。检测可以指示选自包括以下的组的癌症的存在:卵巢癌、胰腺癌、乳腺癌、结肠直肠癌、非小细胞肺癌(例如,鳞状细胞癌或腺癌)或任何其他癌症。检测可以指示选自包括以下的组的任何癌症的存在:卵巢癌、胰腺癌、乳腺癌、结肠直肠癌、非小细胞肺癌(例如,鳞状细胞癌或腺癌)或任何其他癌症。检测可以指示选自包括以下的组的多于一种癌症中的任一种的存在:卵巢癌、胰腺癌、乳腺癌、结肠直肠癌和非小细胞肺癌(例如,鳞状细胞癌或腺癌)或任何其他癌症。检测可以指示本申请中提及的任何癌症中的一种或更多种的存在。

[0269] 一种或更多种癌症可以在组中的至少一个外显子中显示出肿瘤标志物。选自包括卵巢癌、胰腺癌、乳腺癌、结肠直肠癌、非小细胞肺癌(鳞状细胞或腺癌)或任何其他癌症的组的一种或更多种癌症各自在组中的至少一个外显子中显示出肿瘤标志物。至少3种癌症中的每一种都可以在组中的至少一个外显子中显示出肿瘤标志物。至少4种癌症中的每一种都可以在组中的至少一个外显子中显示出肿瘤标志物。至少5种癌症中的每一种都可以在组中的至少一个外显子中显示出肿瘤标志物。至少8种癌症中的每一种都可以在组中的至少一个外显子中显示出肿瘤标志物。至少10种癌症中的每一种都可以在组中的至少一个外显子中显示出肿瘤标志物。所有癌症都可以在组中的至少一个外显子中显示出肿瘤标志物。

[0270] 如果受试者患有癌症,则受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少85%的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿

瘤标志物。至少90％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少92％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少95％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少96％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少97％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少98％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少99％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。至少99.5％的患有癌症的受试者可以在组中的至少一个外显子或基因中显示出肿瘤标志物。

[0271] 如果受试者患有癌症,则受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少85％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少90％的癌症受试者可以在组中的至少一个区域中显示肿瘤标志物。至少92％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少95％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少96％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少97％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少98％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少99％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。至少99.5％的患有癌症的受试者可以在组中的至少一个区域中显示出肿瘤标志物。

[0272] 检测可以以高灵敏度和/或高特异性来进行。灵敏度可以指被正确鉴定为阳性的阳性的比例的量度。在一些情况下,灵敏度是指所有被检测到的现有肿瘤标志物的百分比。在一些情况下,灵敏度是指被正确鉴定为患有某种疾病的病人的百分比。特异性可以指被正确鉴定为阴性的阴性的比例的量度。在一些情况下,特异性是指被正确鉴定的未改变碱基的比例。在一些情况下,特异性是指被正确鉴定为没有某些疾病的健康人的百分比。先前描述的非独特加标签方法通过减少由扩增和测序误差生成的噪声显著增加检测的特异性,这减少了假阳性的频率。检测可以以至少95％、97％、98％、99％、99.5％或99.9％的灵敏度和/或至少80％、90％、95％、97％、98％或99％的特异性来进行。检测可以以至少90％、95％、97％、98％、99％、99.5％、99.6％、99.98％、99.9％或99.95％的灵敏度来进行。检测可以以至少90％、95％、97％、98％、99％、99.5％、99.6％、99.98％、99.9％或99.95％的特异性来进行。检测可以以至少70％的特异性和至少70％的灵敏度、至少75％的特异性和至少75％的灵敏度、至少80％的特异性和至少80％的灵敏度、至少85％的特异性和至少85％的灵敏度、至少90％的特异性和至少90％的灵敏度、至少95％的特异性和至少95％的灵敏度、至少96％的特异性和至少96％的灵敏度、至少97％的特异性和至少97％的灵敏度、至少98％的特异性和至少98％的灵敏度、至少99％的特异性和至少99％的灵敏度、或100％的特异性100％的灵敏度来进行。在一些情况下,所述方法可以以约80％或更高的灵敏度检测肿瘤标志物。在一些情况下,所述方法可以以约95％或更高的灵敏度检测肿瘤标志物。在一些情况下,所述方法可以以约80％或更高的灵敏度和约95％或更高的灵敏度检测肿瘤标志物。

[0273] 检测可以高度准确。准确度可以适用于无细胞DNA中肿瘤标志物的鉴定和/或癌症

的诊断。统计工具,诸如上文描述的协变量分析,可以用于增加和/或测量准确度。所述方法可以以至少80%、90%、95%、97%、98%或99%、99.5%、99.6%、99.98%、99.9%或99.95%的准确度检测肿瘤标志物。在一些情况下,所述方法可以以至少95%或更高的准确度检测肿瘤标志物。

[0274] 检测限值/噪声范围

[0275] 噪声可能通过拷贝和/或读取多核苷酸中的错误而引入。例如,在测序过程中,单个多核苷酸可以首先经历扩增。扩增可能引入错误,使得扩增的多核苷酸的子集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%。在一些情况下,错误率可以多达约10%、多达约9%、多达约8%、多达约7%、多达约6%、多达约5%、多达约4%、多达约3%、多达约2%、或多达约1%。当对所有被假定为具有相同序列的分子的集合进行测序时,这样的噪声可以足够小,使得人们可以以高可靠性鉴定原始碱基。

[0276] 然而,如果亲本多核苷酸的集合包括在特定基因座处变化的多核苷酸的子集,则噪声可能是一个重大问题。例如,当无细胞DNA不仅包括种系DNA,还包括来自另一来源的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,如果具有序列变异的分子的频率可能与通过测序过程引入的错误的频率在相同的范围内,则可能无法将真序列变异与噪声区分开。这可能会干扰,例如,样品中的序列变异的检测。例如,序列可以具有0.5%-1%的每碱基错误率。扩增偏倚和测序错误将噪声引入到最终测序产物中。这种噪声可以降低检测的灵敏度。作为非限制性实例,频率低于测序错误率的序列变异可被误认为噪声。

[0277] 噪声范围或检测限值是指其中具有序列变异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的情况。“检测限值”也可以指其中对于待被检测的变异,太少携带该变异的分子被测序的情况。具有序列变异的分子的频率可能由于核酸分子的量少而与错误的频率在相同的范围内。作为非限制性实例,采样量,例如,100ng的核酸,可能包含相对少量的无细胞核酸分子,例如,循环肿瘤DNA分子,使得序列变异的频率可能是低的,尽管该变异可能存在于大多数循环肿瘤DNA分子中。可选地,序列变异可能是罕见的,或者仅发生在非常少量的采样核酸中,使得检测到的变异不能与噪声和/或测序错误区分开。作为非限制性实例,在特定基因座处,肿瘤标志物可能仅在该基因座处的全部读段的0.1%至5%中被检测到。

[0278] 失真在测序过程中可以表现为由亲本群体中相同频率的分子产生的信号强度例如序列读段的总数的差异。例如,失真可以通过扩增偏倚、GC偏倚或测序偏倚被引入。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段中GC含量丰富或贫乏区域的不均匀呈现。此外,通过以比它们在群体中的实际数目更大或更小的量提供序列读段,扩增偏倚可以使拷贝数变异的测量失真。

[0279] 减少来自单个个体分子或来自分子整体的噪声和/或失真的一种方式为,将序列读段分组为源自原始个体分子的家族,以减少来自单个个体分子或来自分子整体的噪声和/或失真。将初始遗传物质的样品中的个体多核苷酸有效转化为测序就绪的加标签的亲

本多核苷酸,可以增加初始遗传物质的样品中的个体多核苷酸将在测序就绪的样品中呈现的概率。这可以产生关于初始样品中的更多多核苷酸的序列信息。另外,通过从加标签的亲本多核苷酸扩增的子代多核苷酸的高比率采样,以及将生成的序列读段叠并为呈现加标签的亲本多核苷酸的序列的共有序列,来高产量地生成加标签的亲本多核苷酸的共有序列,可以减少由扩增偏倚和/或测序错误引入的噪声,并且可以增加检测的灵敏度。将序列读段叠并为共有序列是减少从一个分子接收到的信息中的噪声的一种方式。使用概率函数是减少噪声和/或失真的另一种方法,概率函数将接收到的频率转换为每个可能的真核苷酸的似然性或后估计值(使用扩增和测序误差曲线的特定估计值)。关于分子整体,将读段分组为家族并确定家族的定量量度减少了例如在多于一个不同基因座中的每一个基因座处的分子的量的失真。同样,将不同家族的序列读段叠并为共有序列消除了由扩增和/或测序错误引入的错误。此外,基于由家族信息得到的概率来确定碱基判定的频率也减少了从分子整体接收到的信息中的噪声。频率报告或肿瘤标志物判定也可以使用多于一个参考序列和覆盖度观察来进行,从这些参考序列和覆盖度观察将确定在位置处观察肿瘤标志物的频率。参考序列可以包括来自健康个体或来自患有疾病或状况诸如癌症的个体的序列或标志物谱。来自“已知”参考样品的频率可以用于设置进行标志物检测判定的阈值频率。例如,对于在某一位置处具有“A”的核苷酸,0.1%的频率可以被用作用于确定是否在测试受试者中判定该位置处的碱基为“A”的阈值。例如,可以使用至少20个、至少50个、至少100个、至少500个、至少1,000个、至少2,000个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000个、至少10,000个、至少11,000个、至少12,000个、至少13,000个、至少14,000个、至少15,000个、至少16,000个、至少17,000个、至少18,000个、至少19,000个、至少20,000个、至少30,000个、至少40,000个、至少50,000个、至少60,000个、至少70,000个、至少80,000个、至少90,000个、或至少100,000个参考序列。

[0280] 通过将加标签的分子和位置信息与被处理的样品内或跨多批样品的观察分子集合进行比较,以鉴定来自其他处理样品的污染分子,可以进一步减少噪声和/或失真。噪声和/或失真可以通过将序列读段中的遗传变异与其他序列读段中的遗传变异进行比较来进一步减少。在一个序列读段中以及再次在其他序列读段中观察到的遗传变异增加了检测到的变异实际上为肿瘤标志物而不仅仅为测序错误或噪声的概率。作为非限制性实例,如果遗传变异在第一序列读段中被观察到并且也在第二序列读段中被观察到,则可以关于该变异是否实际上是遗传变异而不是测序错误进行贝叶斯推断。

[0281] 重复检测变异可以增加变异被准确地检测到的概率、似然性和/或置信度。可以通过比较两组或更多组遗传数据或遗传变异来重复地检测变异。两组或更多组遗传变异可以在多个时间点的样品和在相同时间点的不同样品(例如重新分析的血液样品)两者中检测。当检测到变异处于噪声范围内或低于噪声阈值时,重新采样或重复检测低频率变异使得该变异更可能实际上是变异而不是测序错误。重新采样可以来自相同的样品,诸如重新分析或重新运行的样品,或来自在不同时间点的样品。

[0282] 共变量检测可以增加变异被准确地检测到的概率、似然性和/或置信度。对于共变量肿瘤标志物,一种肿瘤标志物的存在与一种或更多种其他肿瘤标志物的存在相关。基于对共变量遗传变异的检测,甚至在相关遗传变异以低于检测限值存在的情况下,推断相关共变量遗传变异的存在也是可能的。可选地,基于对共变量遗传变异的检测,相关遗传变异

的诊断置信度指示可以被增加。此外,在共变量变异被检测到的一些情况下,低于检测限值检测到的共变量变异的检测阈值可以被降低。共变量变异或基因的非限制性实例包括:驱动突变和耐受性突变、驱动突变和乘客突变。作为共变异或基因的具体实例为在肺癌中发现的EGFR L858R活化突变和EGFR T790M耐受性突变。许多其他共变量变异和基因与多种耐受性突变相关,并且将被本领域技术人员鉴定。

[0283] 在一个实施方式中,使用来自基本上同时或在多于一个时间点内收集的多于一个样品的测量,可以调整每一个变异的诊断置信度指示以指示预测拷贝数变异(CNV)或突变或肿瘤标志物的观察的置信度。可以通过使用在多于一个时间点的测量来增加置信度,以确定癌症是否进展、缓解或稳定。诊断置信度指示可以通过许多统计方法中的任何一种来指定,并且可以至少部分地基于在一段时间段内观察到的测量的频率。例如,可以做出当前和先前结果的统计相关性。可选地,对于每一个诊断,可以建立隐马尔可夫模型(hidden Markov model),使得可以基于来自多于一个测量或时间点的特定测试事件的发生频率来做出最大似然性或最大后验决定。作为该模型的一部分,也可以输出特定决定的误差概率和所得的诊断置信度指示。以这种方式,参数的测量,无论它们是否在噪声范围内,均可以被提供置信区间。随时间推移进行测试,人们可以通过比较随时间推移的置信区间来增加癌症是否进展、稳定或缓解的预测置信度。两个采样时间点可以被分开至少约1微秒、1毫秒、1秒、10秒、30秒、1分钟、10分钟、30分钟、1小时、12小时、1天、1周、2周、3周、1个月或1年。两个时间点可以被分开约一个月至约一年、约一年至约5年、或不多于约三个月、两个月、一个月、三周、两周、一周、一天或12小时。在一些实施方案中,两个时间点可以被治疗事件诸如进行治疗施用或进行外科手术分开。当两个时间点被治疗事件分开时,可以比较在事件之前和之后检测到的CNV或突变。

[0284] 在收集无细胞多核苷酸序列的测序数据之后,可以对序列数据应用一个或更多个生物信息学处理,以检测遗传特征或变异,诸如调控元件处的cfDNA特征、核小体间距/核小体结合模式、核酸的化学修饰、拷贝数变异、及突变或包括但不限于甲基化谱的表观遗传标志物的改变、以及遗传变异诸如SNV、CNV、插入/缺失和/或融合。在需要拷贝数变化分析的一些情况下,序列数据可以:1)与参考基因组比对并映射至单独的分子;2)过滤;4)分到序列窗或箱中;5)对每一个窗的覆盖度读段计数;6)然后可以使用统计建模算法对覆盖度分子进行归一化;以及7)可以生成输出文件,其反映在基因组中的多个位置处的离散的拷贝数状态。在一些情况下,计数与参考基因组的特定基因座比对的覆盖度读段/分子或归一化的覆盖度读段的数目。在需要突变分析的其他情况下,序列数据可以1)与参考基因组比对并映射至单独的分子;2)过滤;4)基于该特定碱基的覆盖度读段来计算变异碱基的频率;5)使用随机、统计或概率建模算法来对变异碱基频率进行归一化;以及6)可以生成输出文件,其反映在基因组中的多个位置处的突变状态。用于映射的参考基因组可以包括任何感兴趣物种的基因组。可用作参考的人类基因组序列可以包括hg19组装、GRCh38.p4或任何先前或可得的hg组装。这些序列可以使用在genome.ucsc.edu/index.html可得的基因组浏览器来查询。其他物种基因组包括,例如PanTro2(黑猩猩)和mm9(小鼠)。

[0285] 在一些情况下,标识符(诸如包括条形码的标识符)可以用于在突变分析期间对序列读段进行分组。在一些情况下,序列读段被分组到家族中,例如,通过使用标识符或标识符和起始/终止位置或序列的组合。在一些情况下,碱基判定可以如下进行:通过将一个或

更多个家族中的核苷酸与参考序列进行比较,并且确定特定碱基1)在每个家族中,以及2)在家族和参考序列之间的频率。核苷酸碱基判定可以基于标准诸如在位置处具有碱基的家族的百分比来进行。在一些情况下,如果碱基判定频率大于如通过多于一个参考序列(例如,来自健康个体的序列)中的频率确定的噪声阈值,则报告碱基判定。来自患者或受试者的当前和先前分析的时间信息用于增强分析和确定。在一些实施方案中,将来自患者或受试者的序列信息与从健康个体的群组、癌症患者的群组或来自患者或受试者的种系DNA获得的序列信息进行比较。种系DNA可以从以下但不限于以下获得:体液、全血、血小板、血清、血浆、粪便、红细胞、白细胞(white blood cell)或白细胞(leukocyte)、内皮细胞、组织活组织检查、滑液、淋巴液、腹水、间质或细胞外液、细胞间空间的液体,包括龈沟液、骨髓、脑脊液、唾液、粘液、痰、精液、汗液、尿液或任何其他体液。癌症患者群组可以与患者或受试者患有相同类型的癌症、与患者或受试者患有相同时期的癌症,两者或都不具有。在一些实施方案中,癌症患者的群组、健康个体的群组或来自受试者的种系DNA被用于提供位置处的碱基的基线频率,并且基线频率被用于在受试者中进行碱基判定。不受限制地,可以将健康个体的群组或来自受试者的种系DNA中的位置处的碱基频率与在来自受试者的序列读段中检测到的碱基频率进行比较。

[0286] 在一些实施方案中,本公开内容的方法和系统可以用于检测0.025%或更低、0.05%或更低、0.075%或更低、或者0.1%或更低的次要等位基因频率(MAF)。拷贝数变异可以按(1)测试样品中基因的独特分子计数(UMC)与(2)参考样品(例如,对照样品)中该基因的UMC的比率来测量。在一些实施方案中,本公开内容的方法和系统可以用于检测为拷贝数扩增(CNA)的拷贝数变异。在一些实施方案中,本公开内容的方法和系统可以用于检测至少1.5、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、55、60或更多的CNA。在一些实施方案中,本公开内容的方法和系统可以用于检测为拷贝数损失(CNL)的拷贝数变异。在一些实施方案中,本公开内容的方法和系统可以用于检测少于0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1或0.05的CNL。

[0287] 多种不同的反应和/或操作可以在本文公开的系统和方法中发生,包括但不限于:核酸测序、核酸定量、测序优化、检测基因表达、定量基因表达、基因组谱分析、癌症谱分析或表达的标志物的分析。此外,所述系统和方法具有许多医学应用。例如,它可以用于多种遗传和非遗传疾病和紊乱(包括癌症)的鉴定、检测、诊断、治疗、监测、分期或风险预测。它可以用于评估受试者对遗传和非遗传疾病的不同治疗的响应,或提供关于疾病进展和预后的信息。

[0288] 计算机控制系统

[0289] 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1示出了被编程或以其他方式配置为分析序列数据、检测肿瘤标志物及确定癌症状态的计算机系统101。计算机系统101可以调控本公开内容的序列分析的各个方面,诸如,例如将数据针对已知序列和变异进行匹配。计算机系统101可以是用户的电子装置或相对于该电子装置远程定位的计算机系统。电子装置可以是移动电子装置。

[0290] 计算机系统101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”)105,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101还包括存储器或存储器位置110(例如,随机存取存储器、只读存储器、闪速存储器)、电子存

储单元115(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120(例如,网络适配器)和外围装置125,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110、储存单元115、界面120和外围装置125与CPU 105通过通信总线(实线),诸如主板(motherboard)通信。存储单元115可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统101可以借助于通信界面120被可操作地耦合至计算机网络(“网络”)130。网络130可以是互联网(Internet)、内联网和/或外联网、或与互联网通信的内联网和/或外联网。在一些情况下,网络130为通信和/或数据网络。网络130可以包括一个或更多个计算机服务器,这可以支持分布式计算,诸如云计算。在一些情况下,借助于计算机系统101,网络130可以实现对等网络(peer-to-peer network),其可以使耦合至计算机系统101的装置能够作为客户端或服务器运行。

[0291] CPU 105可以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。指令可以被存储于存储器位置,诸如存储器110中。指令可以被导向CPU 105,该指令可以随后编程或以其他方式配置CPU 105以实现本公开内容的方法。由CPU 105进行的操作的实例可以包括读取、解码、执行和写回。

[0292] CPU 105可以是电路诸如集成电路的一部分。系统101的一个或更多个其他组件可以被包含在电路中。在一些情况下,电路为专用集成电路(ASIC)。

[0293] 存储单元115可以存储文件,诸如驱动程序、库和保存的程序。存储单元115可以存储用户数据,例如,用户偏好和用户程序。在一些情况下,计算机系统101可以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101的外部,诸如位于通过内联网或互联网而与计算机系统101通信的远程服务器上。

[0294] 计算机系统101可以与一个或更多个远程计算机系统通过网络130进行通信。例如,计算机系统101可以与用户(例如,医师)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930访问计算机系统101。

[0295] 如本文描述的方法可以通过机器(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101的电子存储位置,诸如,例如存储器110或电子存储单元115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,代码可以由处理器105执行。在一些情况下,代码可以从存储单元115检索并存储在存储器110上,以用于由处理器105即时访问。在一些情况下,可以排除电子存储单元115,而将机器可执行指令存储于存储器110中。

[0296] 代码可以被预编译并配置为用于与具有适于执行该代码的处理器的机器一起使用,或者可以在运行时间期间被编译。代码可以以编程语言的形式提供,该编程语言可以被选择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。

[0297] 本文提供的系统和方法的各方面,诸如计算机系统101,可以以编程来体现。技术的多个方面可以被认为是通常呈一种机器可读介质执行或体现的机器(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)”。机器可执行代码可以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”型介质可以包括计算机、处理器等的任何或所有有形存储器,或其相关模块,诸如多种半导体存储器、磁带驱动器、磁盘驱动器等,其可以在任何时间为软件编程提供非暂时性存储。软件的所有或部分有时可以通过互联网或多种其他通信网络进行通信。例如,此类通信可以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中,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中。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装置之间的物理界面、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“存储”介质,否则术语诸如计算机或机器“可读介质”是指参与将指令提供至处理器用于执行的任何介质。

[0298] 因此,机器可读介质,诸如计算机可执行代码,可以采取多种形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括,例如光盘或磁盘,诸如在任何计算机等中的任何存储设备,诸如可以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。载波传输介质可以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,或者计算机可以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以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。

[0299] 计算机系统101可以包括电子显示器135或与之通信,该电子显示器135包括用户界面(UI)140,用于提供例如关于癌症诊断的信息。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。

[0300] 在一个方面,本文提供了包括计算机的系统,计算机包括处理器和计算机存储器,其中计算机与通信网络通信,并且其中计算机存储器包括代码,当代码由处理器执行时,(1)通过通信网络将序列数据接收到计算机存储器中;(2)使用本文所述的方法,确定序列数据中的遗传变异代表种系突变还是体细胞突变;以及(3)通过通信网络报告出该确定。

[0301] 通信网络可以是连接到互联网的任何可用网络。通信网络可以利用例如高速传输网络,包括但不限于电力线宽带(BPL)、电缆调制解调器、数字用户线路(DSL)、光纤、卫星和无线电。

[0302] 在一个方面,本文提供了一种系统,该系统包括:局域网;一个或更多个DNA测序仪,包括被配置为存储DNA序列数据的连接到局域网的计算机存储器;生物信息学计算机,包括计算机存储器和处理器,该计算机连接到局域网;其中所述计算机还包括代码,当所述代码被执行时,复制存储在DNA测序仪上的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、并进行如本文描述的步骤。

[0303] 本公开内容的方法和系统可以通过一个或更多个算法来实现。算法可以在由中央处理单元105执行后通过软件来实现。算法可以例如确定癌症是否存在和/或正在进展。

[0304] 本说明书中提及的所有出版物和专利申请均通过引用并入本文,其程度如同每个单独出版物或专利申请被具体及单独地指明通过引用并入的相同程度。

[0305] 从前述将理解,尽管为了说明的目的,本文已经描述了本公开内容的具体实施方案,但是可以进行各种修改而不偏离本公开内容的精神和范围。因此,本公开内容除了由所附权利要求限制以外,不受限制。

[0306] 虽然本文已经显示出及描述了本公开内容的优选实施方案,但是对于本领域技术人员明显的是,这些实施方案仅被作为实例提供。并不意图本公开内容受限于本说明书中提供的具体实例。虽然已经参考以上提及的说明书描述了本公开内容,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。本领域技术人员现在将想到不偏离本公开内容的许多变化、改变和替换。此外,应当理解,本公开内容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,本文描述的本公开内容的实施方案的各种替代方案均可以用于实践本公开内容。因此设想,本公开内容还应当涵盖任何这样的替代、修改、变化或等同物。意图所附权利要求限定本公开内容的范围,并且从而涵盖在这些权利要求的范围内的方法和结构及其等同物。

实施例

[0307] 实施例1:用于检测早期癌症患者中的ctDNA的下一代测序测定

[0308] 本研究中包括来自10288名经历临床循环肿瘤DNA检测的晚期癌症患者(pt)的去鉴定的(deidentified)cfDNA测序数据(表2中的73个基因)。从血浆提取cfDNA并定量。制备了DNA文库,并测序至15,000X平均读段深度。使用Ingenuity Variant Analysis,将怀疑为种系来源(等位基因分数40%-60%)的点突变和小的插入/缺失按照美国医学遗传学和基因组学学院(American College of Medical Genetics and Genomics)的指南进行分类。研究了多于50种癌症类型,包括肺癌(40%)、乳腺癌(20%)、结肠直肠癌(CRC)(8%)、前列腺癌(6%)和胰腺癌(3%)。受试者的平均年龄为63.6岁(范围:18岁-95岁),并且42%为男性。在已鉴定的34,873个推定种系变异中,520个(1.5%)为致病性的或可能致病性的(PV),16,939种(49%)为意义不明,并且17,414种(50%)为良性或可能良性的。在具有遗传性癌症综合征基因PV的250名受试者(2.4%)中,83名由于高水平的体细胞肿瘤负荷被排除,剩余的167名(1.6%)具有假定的种系PV;总体上,年龄在50岁以下的患者的比率大于那些50岁及更年长的患者(3.3%相比于1.4%,p=0.02),并且乳腺癌患者的比率也大于50岁及更年长的患者(4.3%相比于1.5%,p=0.03)。结果在表4中示出。

[0309] 表4:按癌症类型的PV(N)

[0310]

基因	总数	卵巢	胰腺	前列腺	乳腺	肺	CRC[1]	其他
BRCA2	**78**	3	7	18	26	17	1	6
BRCA1	**38**	11	1		9	12	1	4
TP53	**16**				4	8	2	2
CDKN2A	**10**		1		3	2		4
ATM	**5**		1		2	2		
KIT	**4**				1	2	1	
NF1	**4**			1		1		2
RET	**4**		1			2		1
APC	**4**						1	3
RB1	**2**							2
MLH1	**1**						1	
SMAD4	**1**						1	
总计	**167**	**14**	**11**	**19**	**45**	**46**	**8**	**24**
患者数	10288	205	328	577	2047	4136	830	2165
患者的%	**1.6%**	**6.8%**	**3.4%**	**3.3%**	**2.2%**	**1.1%**	**1.0%**	**1.1%**

[0311] [1]未对除MLH1外显子12外的林奇基因测序

[0312] 偶然鉴定的假定种系PV的观察频率低于真实种系率,但这些发现说明由cfDNA检测在临床上是可行的。重要的是,偶然的种系发现可以影响肿瘤治疗计划(例如用于BRCA1/2突变的PARP抑制剂),并且可以经由增加监测/一级预防使家庭受益。

[0313] 实施例2:辨别无细胞DNA中的种系EGFR T790M突变

[0314] 为了解决血浆cfDNA的基因组分析是否能够允许同时进行肿瘤和种系基因分型及准确分辨肿瘤来源变异和种系变异,研究了EGFR基因中的体细胞变异和种系变异,包括已知的种系变异和10%-20%的NSCLC患者中存在的一组致癌突变。一种EGFR突变,T790M,很少能被作为种系变异检测到,其中已将T790M的存在与家族性肺癌关联。EGFR T790M更通常被视为NSCLC患者对EGFR酪氨酸激酶抑制剂(TKI)产生耐药性后的获得性体细胞突变。在初始治疗之后携带T790M介导的耐药性的肺癌显示出对第三代EGFR TKI奥希替尼的敏感性。

[0315] 一名49周岁的从未吸烟且有肺癌家族史的患者出现转移性肺腺癌,针对第二代EGFR酪氨酸激酶抑制剂(TKI)阿法替尼具有主要进展。初始的组织基因分型已经显示出EGFR L858R和T790M突变,以及CDKN2A、TP53和CTNNB1中的其他体细胞改变。由于EGFR中的L858R突变,开始一线阿法替尼治疗。然而,该患者在治疗仅两个月后就发生了进展性脑转移。在转诊时,血浆下一代测序(NGS)显示出先前观察到的EGFR L858R、TP53和CTNNB1变异的等位基因分数(AF)为1.4%-5.3%,而检测到EGFR T790M等位基因的AF为50.9%,如表5中可见。

[0316] 表5表5

[0317]

改变	时间1(AF)	时间2(AF)	时间3(AF)
EGFR L858R	53%	0.6%	18.1%
EGFR T790M	50.9%	49.2%	54.4%
EGFR C797S	ND	ND	1.3%

EGFR Q787Q	51.5%	48.7%	54.8%
TP53P278R	3.8%	ND	19.7%

[0318] 该患者开始服用EGFR酪氨酸激酶抑制剂(TKI)奥希替尼,这种抑制剂在EGFR T790M介导的对初始EGFR TKI耐受的环境中有活性,并具有持续九个月的临床益处,此时扫描显示出肺中的早期进展。重复血浆NGS显示出EGFR L858R变异为0.6% AF,但T790M相对稳定在49.2% AF(图7,其中701是EGFR L858R;702是EGFR T790M;703是EGFR Q787,并且704是TP53 P278R)。然后患者接受了临床试验的研究性治疗,并发展了进一步的疾病进展。此时重复血浆NGS显示出EGFR L858R的水平增加到18% AF、T790M的水平增加到54% AF、以及介导对奥希替尼的获得性耐药性的第三种EGFR突变C797S的水平增加到1.3% AF。该突变可能介导了对奥希替尼的获得性耐受性,并且在肺癌初始诊断时EGFR T790M突变的存在、连同其在cfDNA分析中的高AF以及肺癌家族史,引起了一种怀疑,即EGFR T790M突变可能代表种系风险等位基因。

[0319] 液滴数字PCR

[0320] 将血液(6mL-10mL)收集到含EDTA的淡紫色(lavender)盖子的vacutainer管中,并以1200g离心10min。使血浆上清液通过3000g离心10min进一步澄清。将第二上清液在-80℃储存在低温恒温管中,直至使用。使用QIAmp循环核酸试剂盒(Qiagen)分离无细胞DNA,并进行液滴数字PCR(ddPCR)。简言之,由2x ddPCR主混合物(Bio-Rad)和被制备用于每个测定的40x TaqMan探针/引物组装TaqMan PCR反应混合物。液滴使用自动化液滴生成器(Bio-RAD)来生成。将PCR进行至终点。在PCR之后,将液滴在QX100或QX200液滴读取器(Bio-Rad)上进行读取。ddPCR数据的分析用QuantaSoft分析软件(Bio-Rad)来进行。所有ddPCR试剂均从Bio-Rad订购。所有引物和探针均从Life Technologies定制。引物和条件如下:

[0321] EGFR L858R正向引物,5'-GCAGCATGTCAAGATCACAGATT-3'(SEQ ID NO.1);反向引物,5'-CCTCCTTCTGCATGGTATTCTTTCT-3'(SEQ ID NO.2);探针序列:5'-VIC-AGTTTGGCCAGCCCAA-MGB-NFQ-3'(SEQ ID NO.3),5'-FAM-AGTTTGGCCCGCCCAA-MGB-NFQ-3'(SEQ ID NO.4)。循环条件:95℃x 10min(1个循环),94℃x 30s和58℃x 1min的40个循环,并保持10℃。

[0322] EGFR del 19正向引物,5'-GTGAGAAAGTTAAAATTCCCGTC-3'(SEQ ID NO.5);反向引物,5'-CACACAGCAAAGCAGAAAC-3'(SEQ ID NO.6);探针序列:5'-VIC-ATCGAGGATTTCCTTGTTG-MGB-NFQ-3'(SEQ ID NO.6),5'-FAM-AGGAATTAAGAGAAGCAACATC-MGB-NFQ-3'(SEQ ID NO.7)。循环条件:95℃x 10min(1个循环),94℃x 30s和55℃x 1min的40个循环,随后保持10℃。

[0323] EGFR T790M,正向引物,5'-GCCTGCTGGGCATCTG-3'(SEQ ID NO.8),反向引物,5'-TCTTTGTGTTCCCGGACATAGTC-3'(SEQ ID NO.9);探针序列:5'-VIC-ATGAGCTGCGTGATGAG-MGB-NFQ-3'(SEQ ID NO.10),5'-FAM-ATGAGCTGCATGATGAG-MGB-NFQ-3'(SEQ ID NO.11)。循环条件:95℃x 10min(1个循环),94℃x 30s和58℃x 1min的40个循环,随后保持10℃。

[0324] 血浆下一代测序

[0325] 从抽取在无细胞DNA管中的10mL全血中分离cfDNA,通过靶向70个基因的外显子和6个基因的关键内含子的杂交捕获来富集,并在Illumina NextSeq500测序仪上测序至~15,000X的平均深度。

[0326] 种系测序

[0327] 对于选择的病例,提供了去鉴定的血沉棕黄层样本,并提取基因组DNA用于对EGFR进行Sanger测序。

[0328] 统计分析

[0329] 使用线性回归分析杂合组变异中的EGFR驱动突变的AF和拷贝数变异的量度之间的关系。使用高斯近似估计个体病例的标准差和平均AF的分布的概率密度函数,并使用Tukey方法鉴定异常值。在每个感兴趣的诊断中,对EGFR T790M患病率确定95%的置信区间。使用双尾Fisher精确检验比较不同诊断中的患病率。

[0330] 结果

[0331] 在具有EGFR T790M突变的85名晚期NSCLC患者中,基于先前的种系测序,已知有3名患者携带种系EGFR T790M突变,而其余患者在TKI治疗之后获得了EGFR T790M。研究T790M等位基因的绝对浓度,以拷贝数/mL血浆计,一些具有体细胞T790M的病例的血浆中携带的突变T790M等位基因的浓度甚至高于三个具有种系EGFR T790M的病例(图2A)。相比之下,对于T790M的AF,以突变T790M的拷贝数占该基因座处的所有突变或野生型变异的比例来计算,三个种系病例的AF徘徊在50%附近,高于体细胞T790M病例的AF(图2A)。然后研究了用第三代EGFR TKI诸如奥希替尼治疗后血浆cfDNA中体细胞EGFR突变水平相比于种系EGFR突变水平的变化。在第一代TKI耐药之后携带获得性EGFR T790M的患者中,EGFR T790M突变和驱动突变(例如L858R或外显子19缺失)两者的浓度响应于治疗而显著降低(图2B)。相比之下,在具有种系EGFR T790M突变的患者中,EGFR驱动突变对治疗有响应,但EGFR T790M水平保持相对稳定(图2B)。这些数据提供了概念验证,即血浆cfDNA中变异水平的定量可以用于区分肿瘤相关突变的体细胞来源和种系来源。

[0332] 下一代测序(NGS)具有捕获跨越许多癌症相关基因的大量变异的潜力。为了进一步研究血浆cfDNA中种系EGFR突变和体细胞EGFR突变的表现,对70个癌基因和肿瘤抑制基因的外显子区域和6个发生致癌性重排的基因的内含子区域进行了测序。查询了临床血浆NGS结果数据库以研究体细胞EGFR突变和种系EGFR突变的分布,结果是对950个连续的NSCLC样品的测试组鉴定了以下的每一个:已知的体细胞突变(L858R和外显子19缺失)、EGFR酪氨酸激酶结构域中常见的种系单核苷酸多态性(SNP)(Q787Q)(17)和T790M,并绘制了每一个的AF分布(图2C)。已知SNP的分布包括两个离散的以50%和100%的AF为中心的正态分布概率分布,这与Q787Q等位基因的杂合性和纯合性相容。相比之下,已知的体细胞改变(L858R和外显子19缺失)的分布显示出指数衰减分布(从测定的检测限值开始,拖着延伸至AF高于90%的长尾巴),这与变化很大,但通常较低(<5%)的体细胞AF相容。T790M的分布显著符合该相同的体细胞分布。然而,存在较小但离散的、以50% AF为中心的正态分布的子群体(图2C)。该模式印证了,研究cfDNA中的变异AF是用于对变异如种系或体细胞来源的EGFR T790M进行分类的一种方法。

[0333] 通过对来自具有种系EGFR T790M的三个病例的治疗前和治疗中的血浆样品进行血浆NGS,进一步研究了AF分布,已知所述三个病例在他们的癌症中携带EGFR驱动突变(两个具有L858R,一个具有L861Q)。研究在血浆NGS上鉴定的所有编码和非编码变异的AF,清楚地看到三组变异(图3A,其中301是EGFR T790M;302是EGFR驱动突变;303是TP53突变;304是其他改变;305是纯合带;306是杂合子带,并且307是肿瘤带)。最低AF组的变异包括EGFR驱

动突变和TP53突变,代表癌症来源的变异。最高AF组的变异以100％ AF附近为中心,代表纯合种系变异。最后,中间组的变异以50％附近为中心,包括已知的种系EGFR T790M突变,并代表杂合种系变异。在用第三代EGFR TKI(两个用奥希替尼,一个用ASP7283)治疗时,低AF的癌症来源的变异降低或变得检测不到(24％→0.2％,3.7％→ND,1.1％→ND),低AF的癌症来源的变异降低或变得检测不到。相比之下,中间组的杂合种系变异仅较小变化,并保持以50％AF为中心(56％→49％,52％→49％,49％→50％)。有趣的是,随着癌症对治疗的响应,这些杂合变异中的一些表现为具有AF的增加,而其他则具有AF的降低。杂合组在治疗时的这些变化可能代表肿瘤来源的拷贝数变异的减少,导致cfDNA中变异AF的变化。

[0334] 然后研究了来自上述初始病例的血浆NGS的所有编码和非编码变异(表5)。这揭示了与所研究的种系EGFR T790M病例相似的模式,其中患者的EGFR T790M突变落入杂合组的变异中,并且在治疗时的AF与EGFR L858R突变相比最低程度地变化(图7)。

[0335] 为了进一步研究cfDNA中肿瘤含量和杂合拷贝数变异之间的关系,针对另外63个EGFR T790M阳性的血浆NGS病例和39个EGFR驱动突变阳性而无T790M的血浆NGS病例,查询了数据库。这105个病例的每一个都被检测到编码变异和非编码变异的中位数为107。观察总计所有10,702个变异的AF分布(图3B),清楚地看到三峰分布,三个AF峰分别在～0％、49％和100％。与非编码外显子变异和内含子变异相比,编码错义变异和无义变异富集在低AF变异组中(图3C),与这代表癌症来源的变异的组一致。

[0336] 为了研究潜在种系变异和体细胞变异之间的关系,每个血浆NGS病例按照EGFR驱动突变从低AF到高AF的顺序被单独地绘制在图上(图4A,其中401(黑色点)是EGFR驱动突变;402(较大的灰色点)是EGFR Q787Q(已知的SNP);403是杂合子带的平均值;404(中等尺寸的灰色点)是其他编码改变,且405(较小的灰色点)是非编码改变)。虽然驱动突变的AF并不是对cfDNA中肿瘤含量的完美量度(由于在一些病例中存在或不存在EGFR基因扩增),但它能够用作对整个群组的cfDNA中肿瘤含量的估计。研究杂合组中变异AF的分布时,此处指定所有变异具有在25％和75％之间的AF,分布随着EGFR驱动突变AF的增加而变化。EGFR驱动突变AF的增加与杂合组标准差的增加(图4B)以及病例和群体平均值之间的绝对差的增加相关,这两者都表明存在癌症来源的拷贝数变异。研究杂合组中变异AF的标准差时,正态分布适于94个病例,而11个病例具有异常特征(图8A)。类似地,研究杂合组中变异的AF中位数时,正态分布适于94个病例,而11个病例具有异常特征(图8B)。因为这些异常群体重叠,总计16个病例显示出这两种异常特征中的一种,证实了cfDNA中的高拷贝数变异,这两种异常特征可能是由于高水平的肿瘤DNA导致种系变异的AF的可变性造成的。

[0337] 因为高拷贝数变异会导致种系变异的AF与预期的50％有很大偏差,所以在这些异常病例中的种系-体细胞的辨别会被削弱。因此,这16个异常病例与没有异常特征的89个病例分离(图5)。用目视检查异常病例的编码变异来区分种系杂合变异和体细胞癌症来源的变异之间的明显分离可能是有挑战性的,但相比之下,目视检查没有这些高拷贝数变异特征的病例的编码变异(图5,其中501(较大的灰色点)是EGFR T790M;502(黑色点)是EGFR驱动突变,并且503(较小的灰色点)是其他编码改变)允许清楚地区分AF在35％至60％范围内的杂合变异组,它们与AF低于30％的癌症来源的变异组不重叠。因此,通过排除具有高拷贝数变异(并因此肿瘤含量高)的血浆NGS病例,血浆NGS结果可以在癌症来源的组中准确地区分体细胞变异,并偶然地在杂合组中鉴定到种系风险等位基因。

[0338] 遵循这些概念验证研究的逻辑,开发并评估了一种整合的生物信息学算法来分离使用血浆NGS测定的70个基因中的种系改变和体细胞改变。该算法首先使用先验知识,包括已知的种系变异和体细胞变异(致病性的和良性的)的内部和外部数据库,为变异分配假定的种系或体细胞来源。例如,EGFR Q787Q改变是一种良性多态性,存在于ExAC数据库(http://exac.broadinstitute.org/)中~52%的种系外显子组中,允许将其指定为假定的种系来源,而不管等位基因分数怎样。相反,EGFR L858R改变虽是NSCLC中一种相对常见的致癌突变,但并未出现在种系数据库中,允许将其指定为推测的体细胞来源。这样的先验分箱通常导致每个病例中位数为78的变异被指定为种系的,这允许如以上研究中描述的通过变异AF构建杂合概率分布。如果所有假定的体细胞突变(通常数量较少)的存在都低于这种杂合种系分布的下限,则剩余未分配的变异的种系-体细胞辨别根据它们相对于由先验变异分类描述的种系分布的AF来进行。然而,如果假定的体细胞变异的AF高于杂合种系分布下限的AF,或者如果检测到极端的染色体不稳定性(如通过基因组的表观二倍体分数评估的),则认为保留在该重叠区域内的变异的种系/体细胞区分是不确定的,并且变异被假定为体细胞来源并如此报告。这种方法允许鉴定具有高阳性预测值的可疑种系变异,理解在高肿瘤DNA含量的环境中种系来源的变异的灵敏度将降低。

[0339] 然后将该算法应用于21个预先收集的临床样品,这些样品在血浆NGS上被检测到具有高AF(30%-75%)的EGFR T790M突变(图9,其中901(较大的灰色点)是EGFR T790M;902(黑色点)是EGFR驱动突变,并且903(较小的灰色点)是其他编码改变)。基于上述EGFR T790M种系-体细胞分离,将病例分为两个群组。群组A包括11个病例,其中体细胞来源的变异相对于种系来源的变异的分布导致对种系T790M突变存在的预测。群组B包括10个病例,其中由于高拷贝数变异和广的杂合组使种系与体细胞的确定复杂化。然后,将每个样品中含有基因组DNA的细胞级分被不可逆地去鉴定,并以双盲的方式提交至CLIA认证的临床实验室进行EGFR测序,使得种系结果不能追溯到任何个体患者。群组A中的所有11例病例均被确认携带种系EGFR T790M(阳性预测值为100%,11/11)。在群组B中的10个病例中,发现1个是种系的,导致灵敏度为92%(11/12),并且总体准确度为95%(20/21)。怀疑群组B中存在非人一个种系样品是具有高肿瘤含量的病例,使得假定的体细胞突变的AF与杂合种系分布重叠,使得难以确定地区分种系变异。

[0340] 已经验证了一种用于鉴定携带种系EGFR T790M的血浆NGS病例的方法后,现有的血浆NGS数据被用于了解种系变异与特定癌症类型的关联。查询了代表大量成人实体瘤类型的31,414名连续发病的独特患者(consecutive unique patient)的临床测试数据库,鉴定到911个EGFR T790M阳性的病例,其中48个通过上文的方法被判定为种系来源。虽然整个患者群组中少数患者的癌症诊断为非鳞状NSCLC(41%),但是具有种系EGFR T790M的48名患者中的43名患者的癌症诊断为非鳞状NSCLC(90%,图6A)。此外,在其余5名具有种系EGFR T790M的患者中,3名有相关诊断(鳞状NSCLC、小细胞肺癌、原发性不明癌)。非鳞状NSCLC患者中种系EGFR T790M的群体频率(43/12,774,0.34%)显著高于具有另一种癌症诊断的患者中可见的群体频率(5/18,640,0.03%,图6B),后者仅略高于一般群体测序工作所报告的群体频率(例如ExAC的等位基因频率中位值为0.0082%)。这些观察结果与具有种系T790M的患者处于增加的特别是NSCLC风险的概念一致,并且它们表明该等位基因不提供除肺癌以外的其他癌症的显著增加的风险。

[0341] 以上分析证明了cfDNA基因组学作为研究种系癌症风险等位基因的工具的能力。使用现有的数据和来自正在进行的临床研究工作的样品,开发并验证了一种用于在cfDNA NGS谱中区分种系变异与癌症来源的体细胞变异的生物信息学算法,所述生物信息学算法提供了一种测定,该测定提供了用于治疗选择的对肿瘤基因型的洞察以及对遗传性风险等位基因的筛选。查询了临床测试数据库以探索罕见种系等位基因EGFR T790M,并在非鳞状NSCLC患者中观察到该突变的富集。以上数据突出了当前用于常规临床护理的血浆基因分型检测种系变异并在某些情况下将种系变异与体细胞变异区分开的能力。

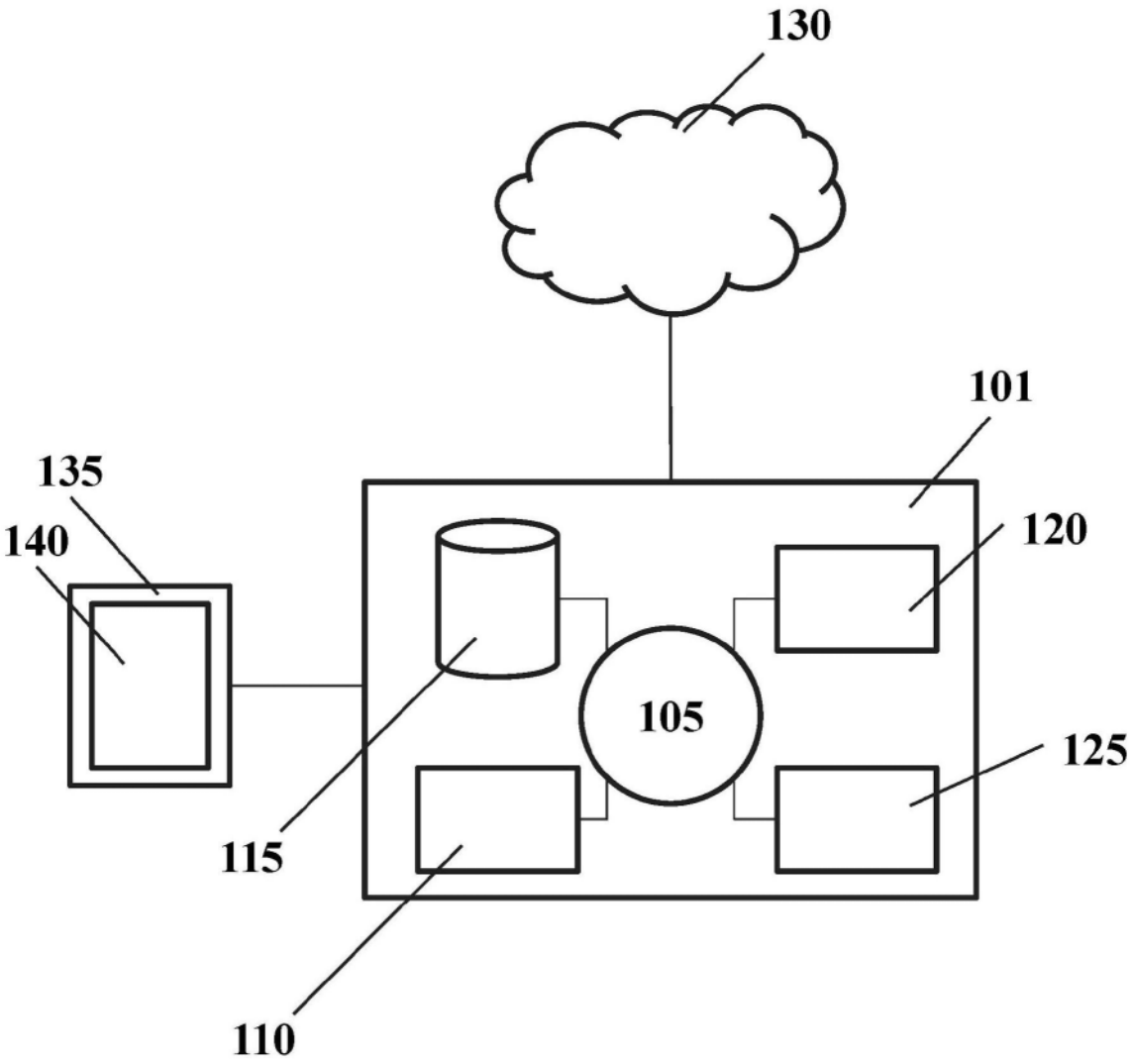


图1

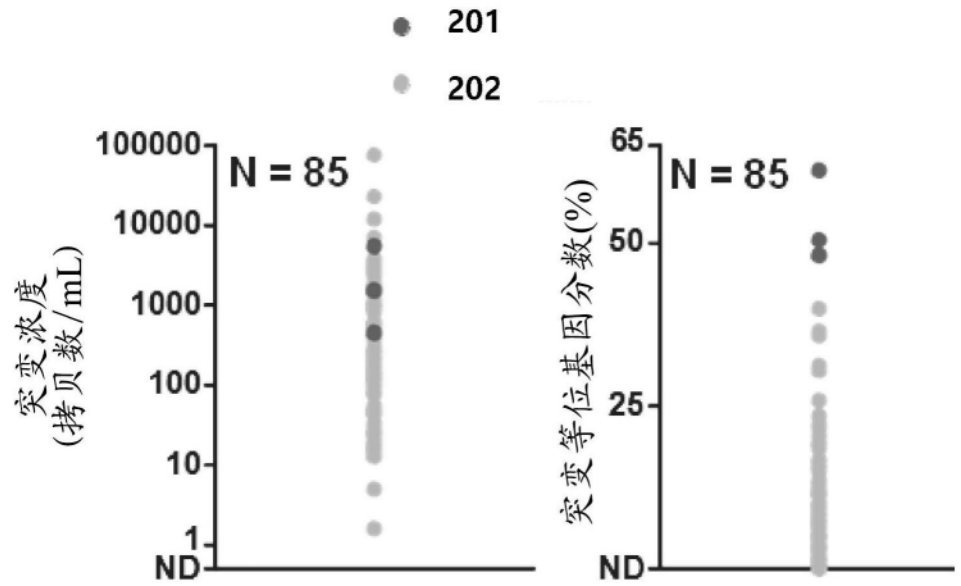

201
202
100000
10000
1000
100
10
1
ND
N = 85
突变浓度
(拷贝数/mL)
65
50
25
ND
N = 85
突变等位基因分数(%)

图2A

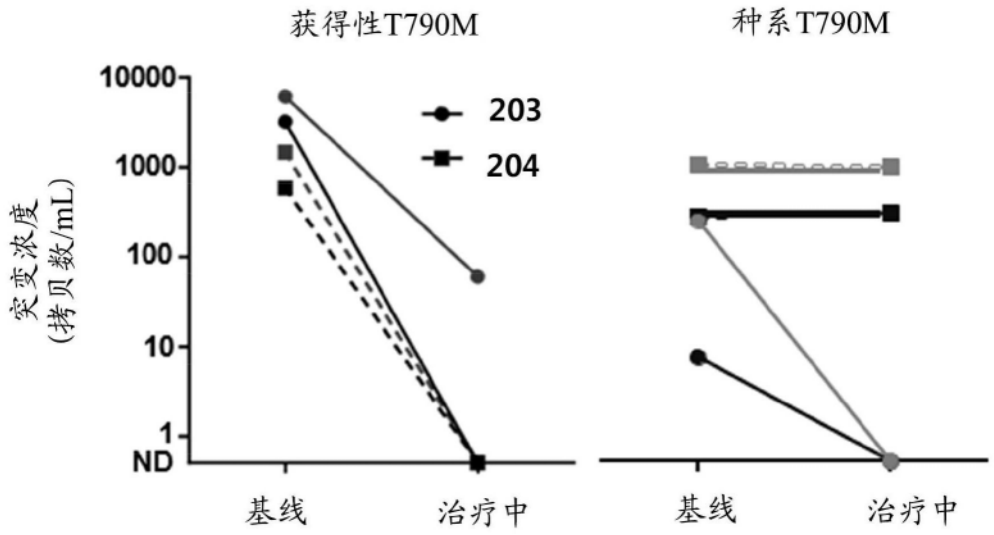

获得性T790M
种系T790M
10000
1000
100
10
1
ND
203
204
突变浓度
(拷贝数/mL)
基线
治疗中
基线
治疗中

图2B

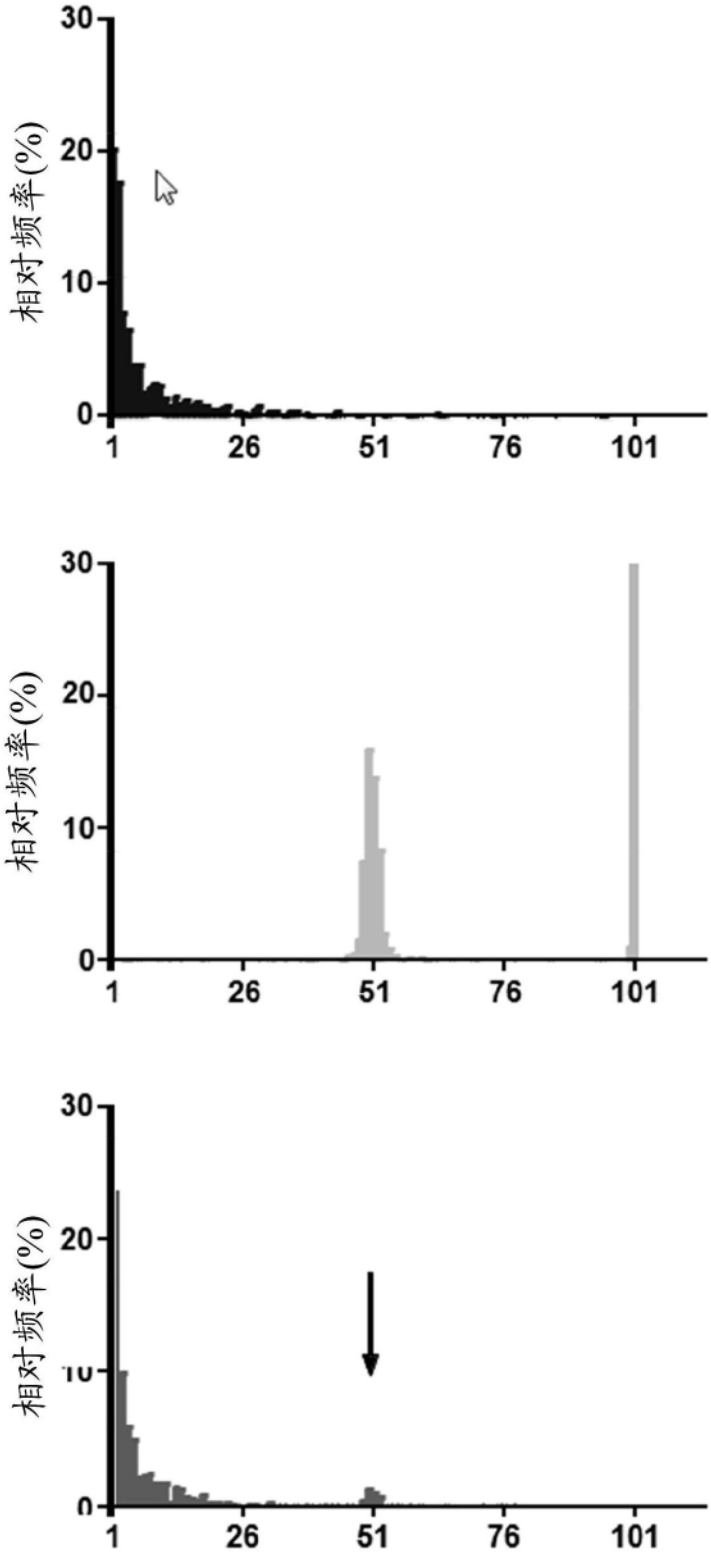


图2C

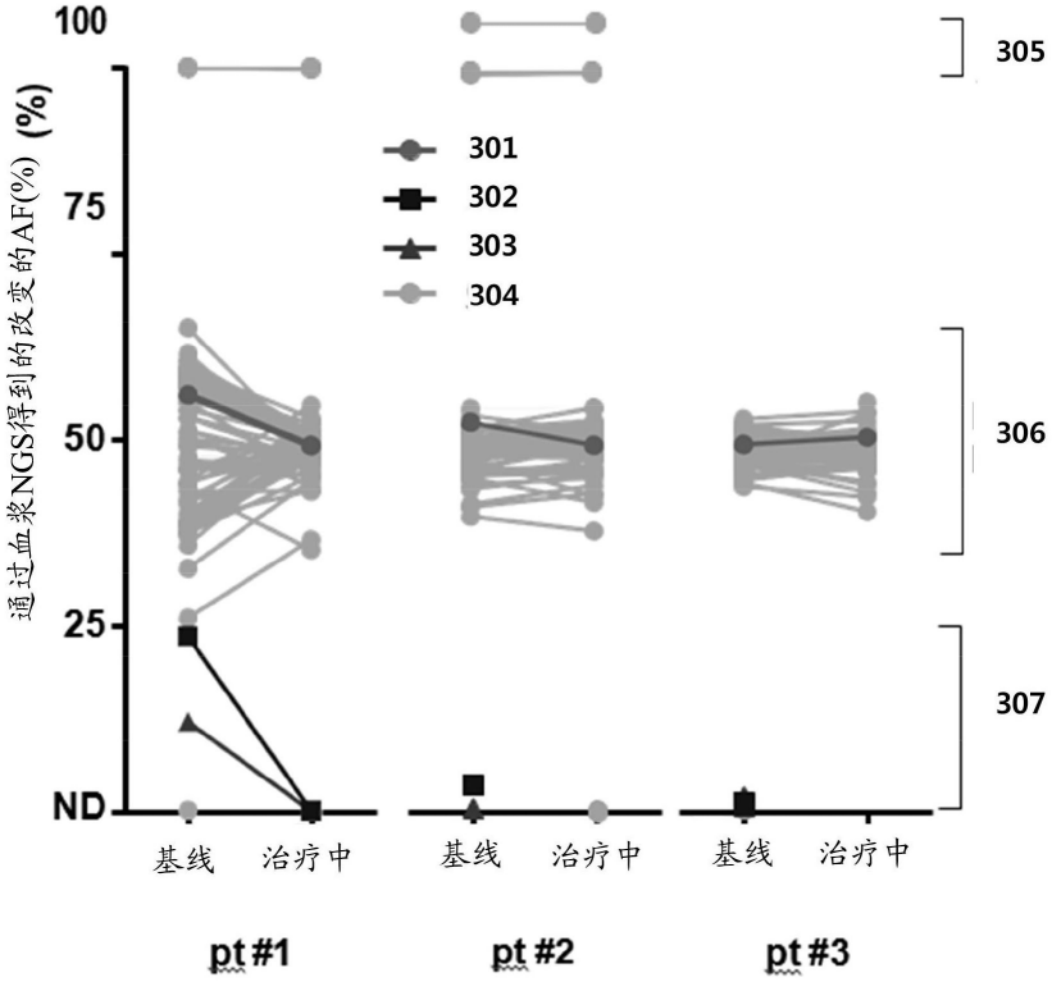

100
75
50
25
ND
通过血浆NGS得到的改变的AF(%) (%)
301
302
303
304
305
306
307
基线
治疗中
基线
治疗中
基线
治疗中
pt #1
pt #2
pt #3

图3A

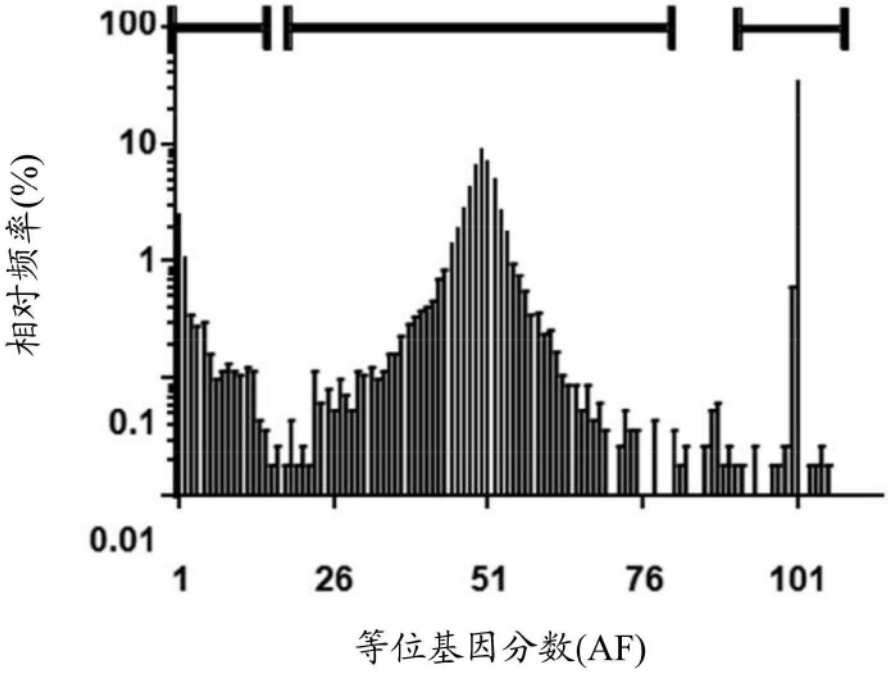


图3B

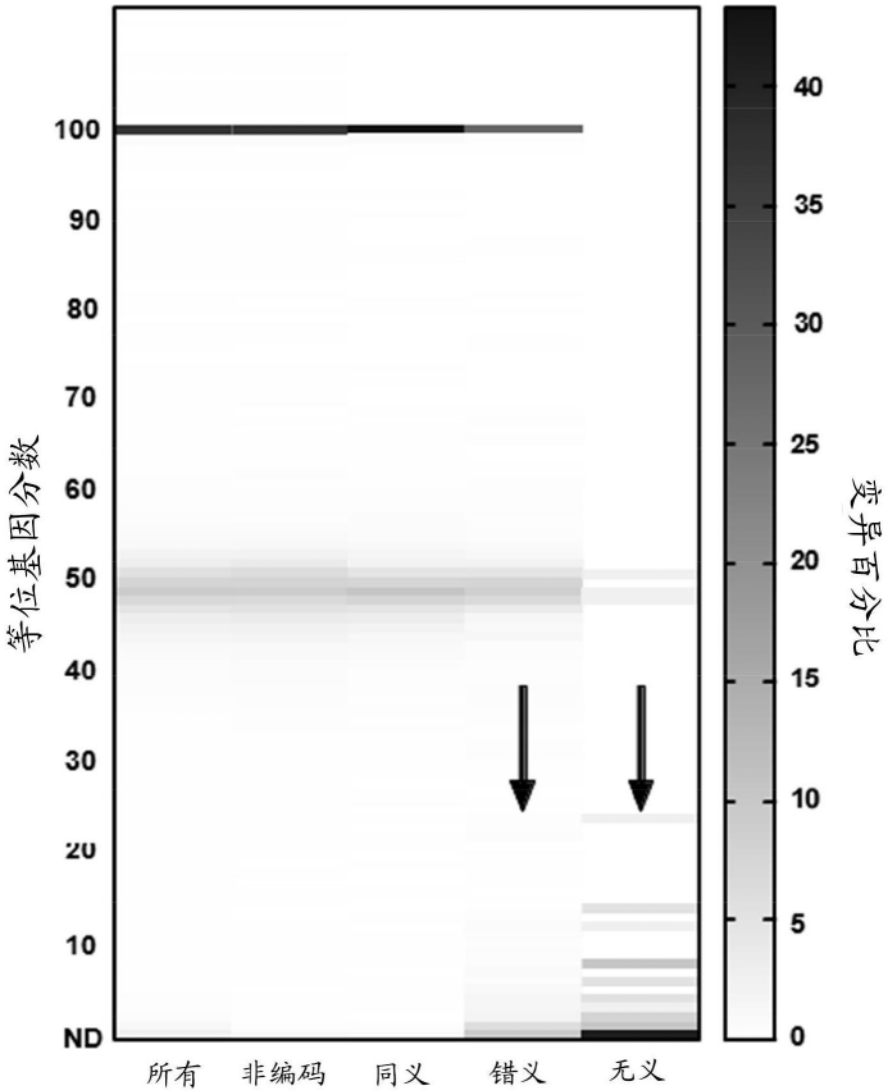

等位基因分数
100
90
80
70
60
50
40
30
20
10
ND
所有
非编码
同义
错义
无义
40
35
30
25
20
15
10
5
0
变异百分比

图3C

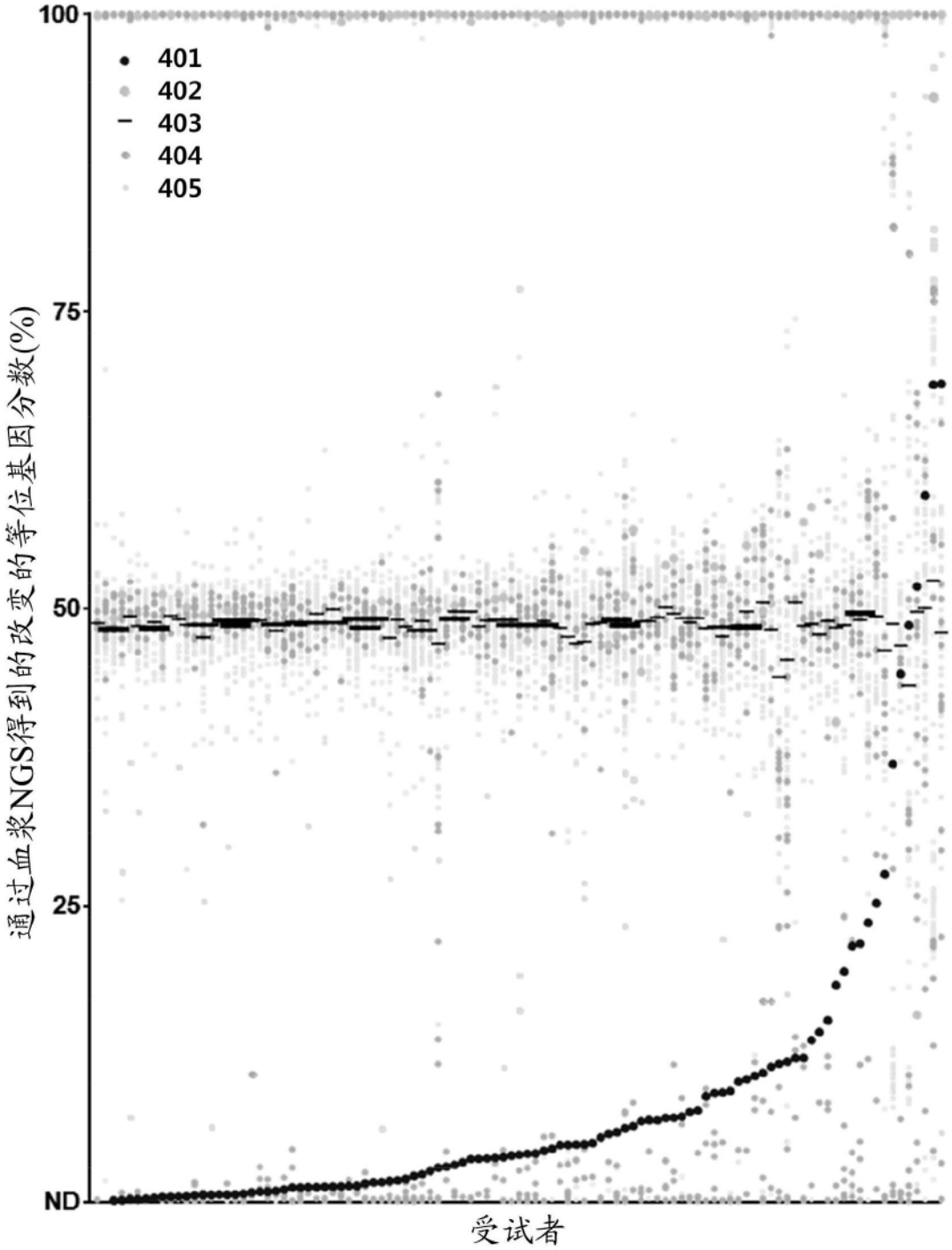


图4A

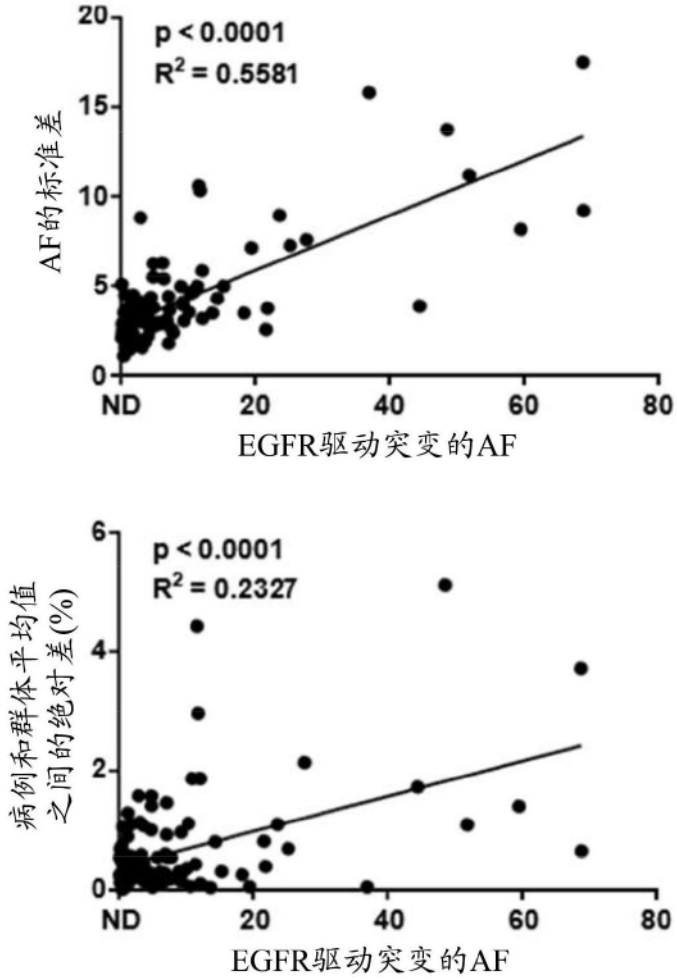

p < 0.0001
R2 = 0.5581
AF的标准差
EGFR驱动突变的AF
p < 0.0001
R2 = 0.2327
病例和群体平均值之间的绝对差(%)
EGFR驱动突变的AF

图4B

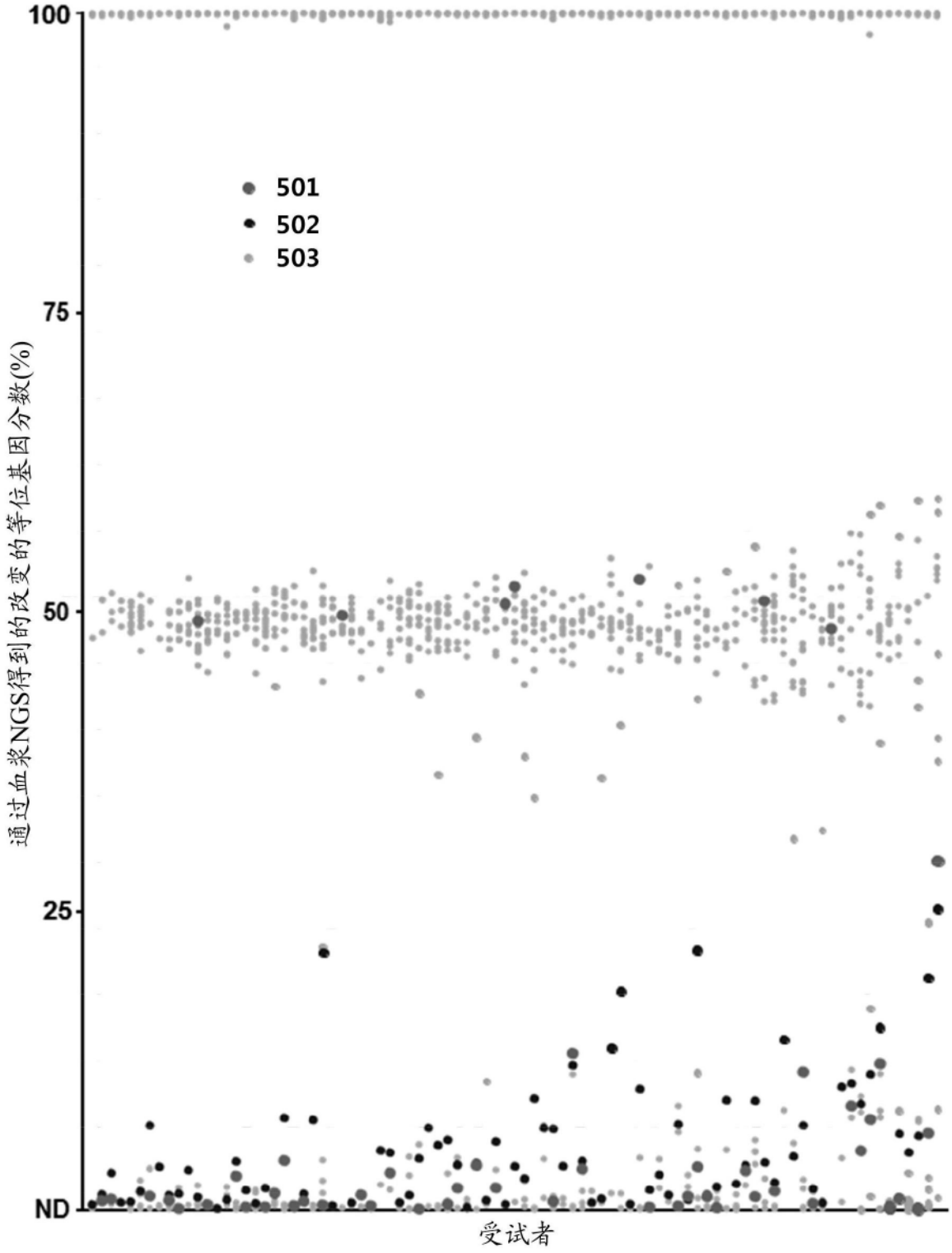


图5

	检测到的 种系T790M	未检测到 种系T790M
总计	48	31,366
非鳞状NSCLC	43 (90%)	12,731 (41%)
其他癌症	5 (10%)	18,635 (59%)

图6A

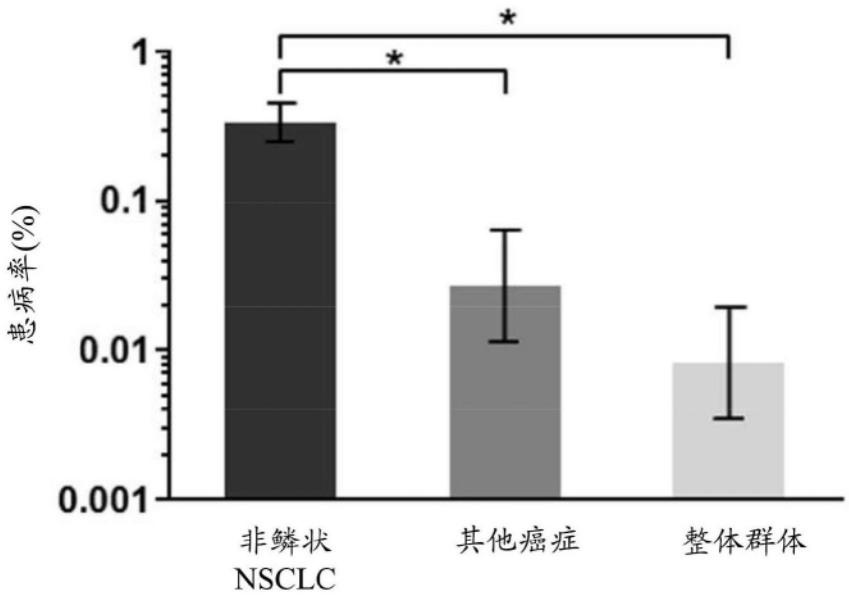


图6B

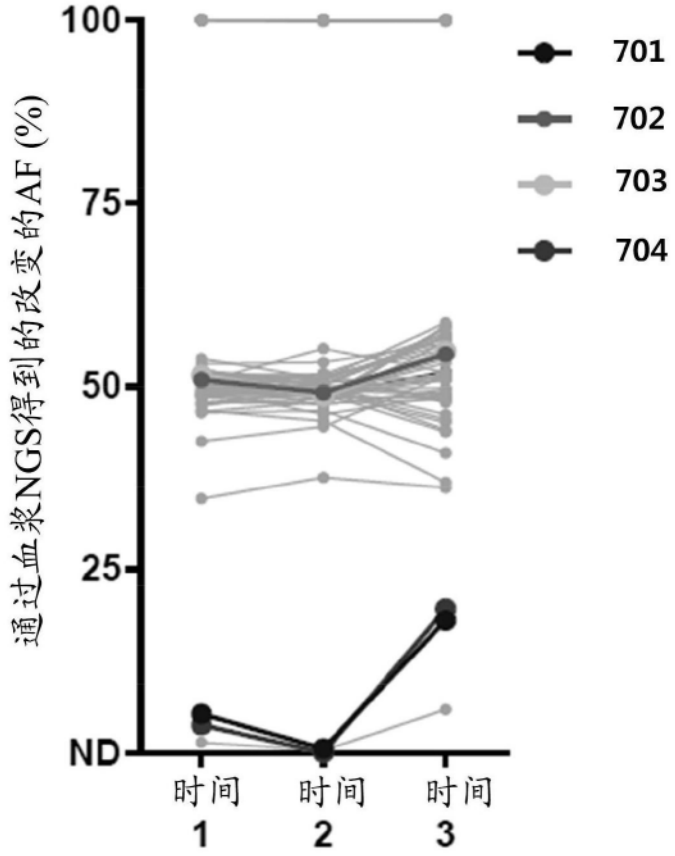


图7

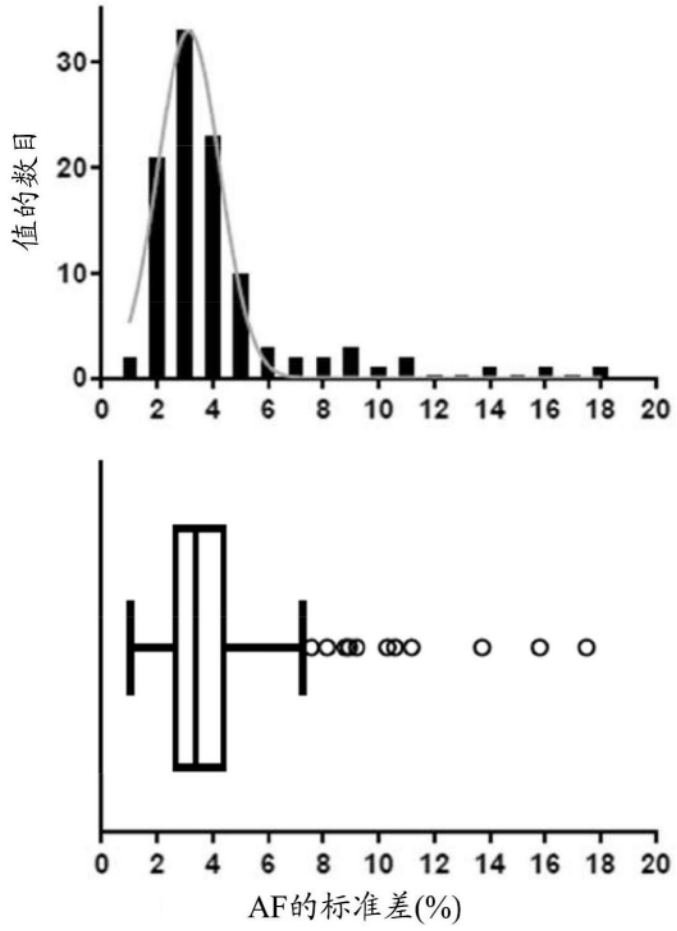

值的数目
30
20
10
0
0 2 4 6 8 10 12 14 16 18 20
0 2 4 6 8 10 12 14 16 18 20
AF的标准差(%)

图8A

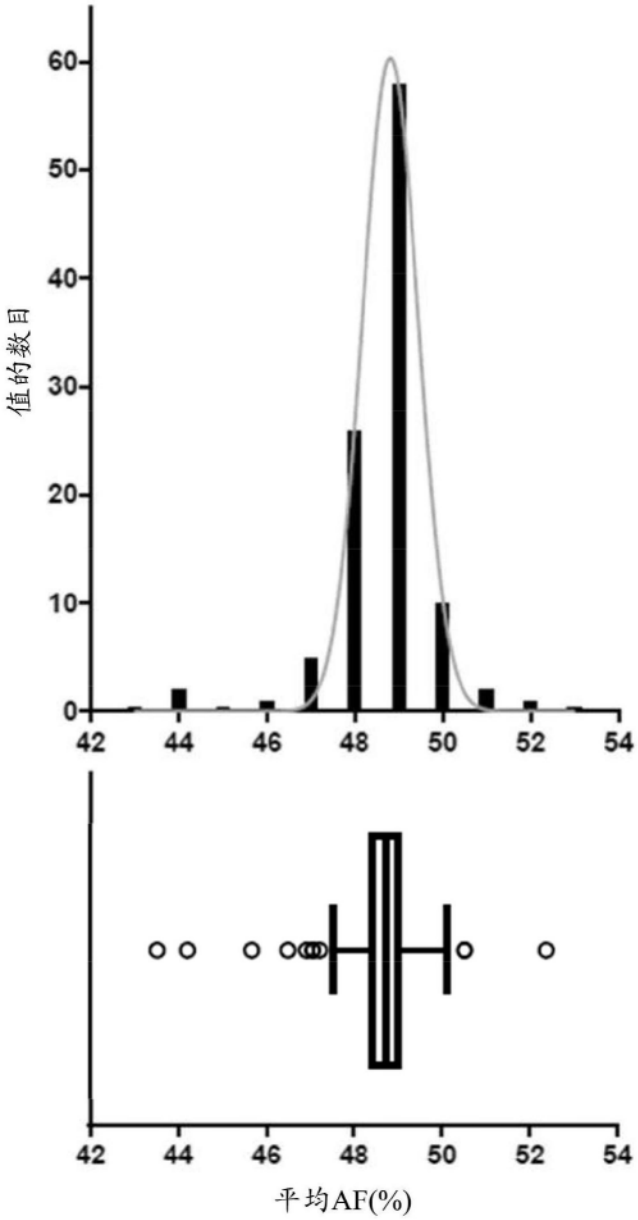


图8B

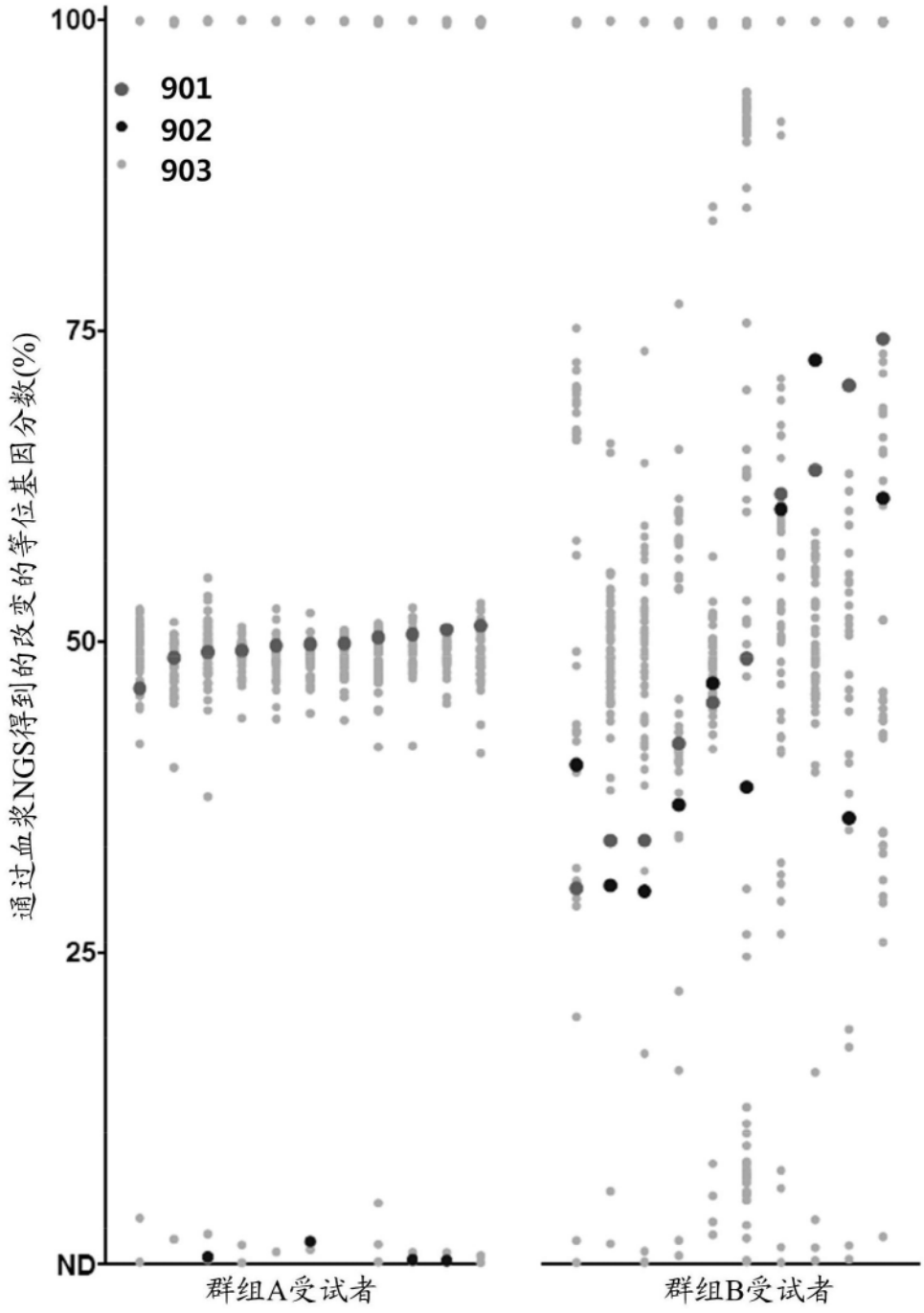

100
75
50
25
ND
901
902
903
通过血浆NGS得到的改变的等位基因分数(%)
群组A受试者
群组B受试者

图9