



(10) 授权公告号 CN 107779506 B

(45) 授权公告日 2022.07.15

(21) 申请号 201711070698.4

(22) 申请日 2013.06.14

(65) 同一申请的已公布的文献号
申请公布号 CN 107779506 A

(43) 申请公布日 2018.03.09

(30) 优先权数据

61/662,878 2012.06.21 US

61/682,725 2012.08.13 US

61/695,795 2012.08.31 US

61/711,172 2012.10.08 US

13/801,748 2013.03.13 US

(62) 分案原申请数据

201380042981.X 2013.06.14

(73) 专利权人 香港中文大学

地址 中国香港新界

(72) 发明人 赵慧君 卢煜明 陈君赐 江培勇

(74) 专利代理机构 北京英赛嘉华知识产权代理
有限责任公司 11204

专利代理师 王达佐 洪欣

(51) Int.Cl.

C12Q 1/6886 (2018.01)

C12Q 1/6827 (2018.01)

(56) 对比文件

WO 2012071621 A1, 2012.06.07

CN 101849236 A, 2010.09.29

WO 2011090557 A1, 2011.07.28

Forshe, Tim; Murtaza, Muhammed;
Parkinson, Christine; 等. Noninvasive
identification and monitoring of cancer
mutations by targeted deep sequencing of
plasma DNA. 《SCIENCE TRANSLATIONAL
MEDICINE》. 2012, 第4卷 (第136期),

审查员 姜紫耀

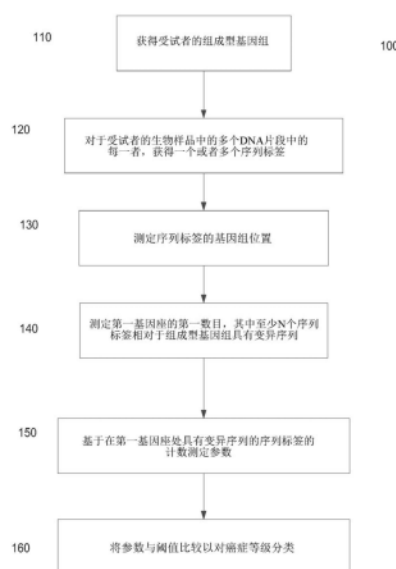
权利要求书9页 说明书41页 附图30页

(54) 发明名称

用于癌症检测的血浆DNA突变分析

(57) 摘要

可以将经历针对癌症的筛选或监测的受试者的生物样品(例如血浆或血清)中的体细胞突变频率与同一受试者的组成型DNA中的体细胞突变频率比较。参数可以来源于这些频率并且将其用于确定癌症等级的分类。可以通过要求任何变异基因座都具有至少规定数目的变异序列读取(标签)来将假阳性过滤掉, 由此提供更准确的参数。可以分析不同变异基因座的相对频率以测定患者中肿瘤的异质性等级。



1. 确定受试者中癌症等级的分类的系统,所述系统包括:
用于获得所述受试者的基因组的一致性序列的装置;
用于接收所述受试者的生物样品中的多个DNA片段中每一个的一或多个序列标签的装置,所述生物样品包括无细胞的DNA;
用于确定所述序列标签的基因组位置的装置;
用于将所述序列标签与所述一致性序列比较以确定第一基因座的第一数目的装置,其中:
在每个第一基因座处,相对于所述一致性序列具有单核苷酸变异体的序列标签的数目高于截止值,所述截止值大于一;
用于基于在所述第一基因座处具有单核苷酸变异体的序列标签的计数来确定参数的装置;以及
用于将所述参数与阈值比较来确定所述受试者中癌症等级的分类的装置。
2. 如权利要求1所述的系统,其中由来自一或多名其它受试者的一或多种样品确定所述阈值。
3. 如权利要求1所述的系统,其中基因座的所述截止值取决于在所述基因座上具有基因组位置的序列标签的总数目。
4. 如权利要求1所述的系统,其中不同截止值用于所述第一基因座中的至少两个。
5. 如权利要求4所述的系统,其进一步包含:
用于动态地确定所述第一基因座之一的第一截止值的装置,所述第一基因座之一驻留于第一区域内。
6. 如权利要求5所述的系统,其中基于所述第一基因座之一的测序深度确定所述第一截止值。
7. 如权利要求5所述的系统,其中基于假阳性率确定所述第一截止值,所述假阳性率取决于测序错误率、所述第一区域的测序深度和所述第一区域中的核苷酸位置的数目。
8. 如权利要求7所述的系统,其中基于所述第一区域中的真实阳性的数目确定所述第一截止值。
9. 如权利要求8所述的系统,其进一步包含:
用于基于所述生物样品中的所述第一区域的测序深度D和源自于肿瘤的DNA百分比浓度f来计算关于所述第一截止值真实阳性的数目的装置。
10. 如权利要求9所述的系统,其中使用根据下式的泊松分布概率计算真实阳性的数目: $Pb = 1 - \sum_{i=0}^{r-1} P(i, M_p)$, 其中Pb是检测真实阳性的概率,并且r是所述第一截止值,并且 $M_p = D \times f / 2$ 。
11. 如权利要求5所述的系统,其中使用以下准则中的任一者确定所述第一截止值:
如果所述测序深度小于50,那么所述第一截止值是5,
如果所述测序深度是50-110,那么所述第一截止值是6,
如果所述测序深度是111-200,那么所述第一截止值是7,
如果所述测序深度是201-310,那么所述第一截止值是8,
如果所述测序深度是311-450,那么所述第一截止值是9,

如果所述测序深度是451-620,那么所述第一截止值是10,和

如果所述测序深度是621-800,那么所述第一截止值是11。

12.如权利要求1所述的系统,其中所述参数是第一基因座的所述第一数目的加权和,其中每个第一基因座的贡献值是基于分配给各个第一基因座的权重值来求得。

13.如权利要求1所述的系统,其中所述参数包括在所述第一数目的第一基因座处含有单核苷酸变异体的所述序列标签的总和。

14.如权利要求13所述的系统,其中所述总和是加权和,并且其中所述第一个第所述一基因座中的第一权重不同于第二个所述第一基因座中的第二权重。

15.如权利要求14所述的系统,其中所述第一权重大于所述第二权重,并且其中第一个所述第一基因座是与癌症相关,并且第二个所述第一基因座是与癌症不相关。

16.如权利要求1所述的系统,其中所述参数是第一基因座的所述第一数目。

17.如权利要求1所述的系统,其中用于确定序列标签的基因组位置的装置包括:

用于将至少一部分所述序列标签与参考基因组比对的装置,其中序列标签的所述比对允许所述序列标签与所述参考基因组之间有一或多个错配。

18.如权利要求17所述的系统,其中用于将所述序列标签与所述一致性序列比较的装置包括:

用于将所述一致性序列与所述参考基因组比较以确定相对于所述参考基因组具有变异序列的第二基因座的第二数目的装置;

用于基于所述比对,确定第三基因座的第三数目的装置,其中:

在每个第三基因座处,相对于所述参考基因组具有单核苷酸变异体的所述序列标签的数目高于截止值;和

用于获取所述第三数目与所述第二数目的差值以获得第一基因座的所述第一数目的装置。

19.如权利要求18所述的系统,其中获取所述第三数目与所述第二数目的所述差值鉴定所述第一基因座。

20.如权利要求19所述的系统,其中用于确定所述参数的装置包括:

对于所述具有第一数目的第一基因座中的每个基因座:

用于对与所述基因座比对上并且在所述基因座处具有单核苷酸变异体的序列标签的数目进行计数的装置;和

用于基于各个计数来确定所述参数的装置。

21.如权利要求1所述的系统,其中所述一致性序列来源于所述受试者的含有多于50%组成型DNA的组成型样品。

22.如权利要求1所述的系统,其中用于确定序列标签的基因组位置的装置包括:

用于将至少一部分所述序列标签与所述一致性序列比对的装置,其中序列标签的所述比对允许所述序列标签与所述一致性序列之间有一或多个错配。

23.如权利要求22所述的系统,其中用于将所述序列标签与所述一致性序列比较的装置包括:

用于基于所述比对,鉴别相对于所述受试者的所述一致性序列在基因组位置处具有单核苷酸变异体的序列标签的装置;

对于展现单核苷酸变异体的每个基因组位置：

用于计算比对到所述基因组位置并且在所述基因组位置处具有单核苷酸变异体的序列标签的各个数目的装置；

用于基于所述各个数目确定参数的装置。

24. 如权利要求23所述的系统，其中用于基于所述各个数目确定参数的装置包括：

用于对所述各个数目求和以获得第一总和的装置；和

用于使用所述第一总和以确定所述参数的装置。

25. 如权利要求24所述的系统，其中用于使用所述第一总和以确定所述参数的装置包括：

用于由所述第一总和减去展现单核苷酸变异体的基因组位置的数目的装置。

26. 如权利要求24所述的系统，其中用于使用所述第一总和以确定所述参数的装置包括：

用于基于比对上的序列标签的某一数量标准化所述第一总和的装置。

27. 如权利要求1所述的系统，其进一步包含：

用于获得多个DNA片段中的每一者的一个或多个第二序列标签的装置，其中所述一个或多个第二序列标签通过对来自含有多于90%组成型DNA的组成型样品的DNA片段的随机测序获得；

用于将至少一部分所述第二序列标签与参考基因组比对的装置，其中第二序列标签的所述比对允许所述第二序列标签与所述参考基因组之间在M个基因组位置处有错配，其中M是等于或大于一的整数；和

用于基于所述第二序列标签和所述比对构建所述一致性序列的装置。

28. 如权利要求27所述的系统，其中所述组成型样品是所述生物样品，并且其中用于构建所述一致性序列的装置包括：

用于确定纯合子基因座或具有两种等位基因的杂合子基因座的装置。

29. 如权利要求27所述的系统，其中：

其中所述生物样品是获自血液样品的血浆或血清，以及

所述组成型样品是获自所述血液样品的白血球。

30. 如权利要求1所述的系统，其中由对所述生物样品中的DNA片段进行随机测序产生所述一或多个序列标签。

31. 如权利要求1所述的系统，其中所述参数是肿瘤来源的DNA的百分比浓度。

32. 如权利要求1所述的系统，其中所述序列标签提供全基因组水平的单核苷酸变异体检测。

33. 一种分析受试者的一或多个肿瘤的异质性的系统，所述系统包含：

用于获得所述受试者的一致性序列的装置；

用于接收所述受试者的生物样品中的多个DNA片段中每一个的一或多个序列标签的装置，所述生物样品包括无细胞的DNA分子；

用于确定所述序列标签的基因组位置的装置；

用于将所述序列标签与所述一致性序列比较以确定第一基因座的数目的装置，其中：

在每个第一基因座处，相对于所述一致性序列具有单核苷酸变异体的序列标签的第一

数目高于截止值,所述截止值大于一;和

用于基于所述第一基因座的序列标签的各自第一数目计算所述一或多个肿瘤的异质性度量的装置。

34.如权利要求33所述的系统,其进一步包含:

用于将所述异质性度量与一或多个阈值比较以确定异质性等级的分类的装置。

35.如权利要求34所述的系统,其中由一或多名其它受试者确定所述一个或多个阈值,并且其中使用一名或多名其它受试者的包括无细胞的DNA的生物样品的异质性的度量以确定阈值,所述一或多名其它受试者的肿瘤已经被活检和分析以确定经活检得到的肿瘤中的突变以确定肿瘤异质性等级。

36.如权利要求35所述的系统,其中用于将所述异质性度量与一或多个阈值比较的装置包括:

用于将所述异质性度量输入到校准函数的装置,所述校准函数基于所述异质性度量输出异质性等级。

37.如权利要求33所述的系统,其中所述异质性度量包括第一基因座的总数目,在所述第一基因座处,多于一个DNA片段比对到一致性序列同时在该基因座处具有单核苷酸变异体。

38.如权利要求33所述的系统,其中计算多个异质性度量,其中用于计算所述异质性度量的装置包括:

对于第一基因座中的每一个,用于计算具有变异序列的序列标签的比例的装置;

用于产生一个关于多个第一基因座的百分比数值的直方图的装置;和

用于鉴别所述直方图中的多个峰值的装置。

39.如权利要求38所述的系统,其中所述多个异质性度量之一对应于多个鉴别到的峰值。

40.如权利要求38所述的系统,其中所述多个异质性度量之一包括所述多个峰值中两个峰值的高度的比率。

41.如权利要求38所述的系统,其中所述比例各自代表在特定第一基因座处测量的肿瘤DNA的百分比浓度。

42.如权利要求33所述的系统,其中所述异质性度量对应于在第一基因座具有变异序列的序列标签的第一特定数量的第一比例与在第一基因座具有变异序列的序列标签的第二特定数量的第二比例的比率。

43.如权利要求42所述的系统,其中所述第一特定数量小于所述第二特定数量。

44.如权利要求43所述的系统,其中所述第一特定数量是第一范围并且所述第二特定数量是第二范围,所述第一范围低于所述第二范围。

45.如权利要求42所述的系统,其中所述第一特定数量和所述第二特定数量对应于具有变异序列的序列标签的百分比浓度或绝对数。

46.如权利要求33所述的系统,其中所述一或多个肿瘤包括单一对象中的多个克隆,并且其中所述异质性包括肿瘤内异质性。

47.如权利要求33所述的系统,其中所述一或多个肿瘤是多个肿瘤,并且其中所述异质性包括肿瘤间异质性。

48. 根据权利要求33所述的系统,其中由对应于各个所述第一基因座的第一数目的直方图确定所述异质性度量。

49. 如权利要求48所述的系统,其中计算多个异质性度量,其中第一基因座包括第一子集和第二子集,其中多个异质性度量包括对应于所述第一子集的各个所述第一数目的第一直方图和对应该于所述第二子集的各个所述第二数目的第二直方图。

50. 如权利要求33所述的系统,其中所述异质性度量包括具有高于规定数值的对应的第一数目的第一基因座的比例。

51. 计算机可读介质,其存储用于控制计算机系统以执行确定受试者中癌症等级的分类的操作的多个指令,所述操作包括:

获得所述受试者的一致性序列;

接收所述受试者的生物样品中的多个DNA片段中每一个的一或多个序列标签,所述生物样品包括无细胞的DNA;

确定所述序列标签的基因组位置;

将所述序列标签与所述一致性序列比较以确定第一基因座的第一数目,其中:

在每个第一基因座处,相对于所述一致性序列具有单核苷酸变异体的序列标签的数目高于截止值,所述截止值大于一;

基于在所述第一基因座处具有单核苷酸变异体的序列标签的计数来确定参数;以及

将所述参数与阈值比较来确定所述受试者中癌症等级的分类。

52. 如权利要求51所述的计算机可读介质,其中由来自一或多名其它受试者的一或多种样品确定所述阈值。

53. 如权利要求51所述的计算机可读介质,其中基因座的所述截止值取决于在所述基因座上具有基因组位置的序列标签的总数目。

54. 如权利要求51所述的计算机可读介质,其中不同截止值用于所述第一基因座中的至少两个。

55. 如权利要求54所述的计算机可读介质,所述操作进一步包括:

动态地确定所述第一基因座之一的第一截止值,所述第一基因座之一驻留于第一区域内。

56. 如权利要求55所述的计算机可读介质,其中基于所述第一基因座之一的测序深度确定所述第一截止值。

57. 如权利要求55所述的计算机可读介质,其中基于假阳性率确定所述第一截止值,所述假阳性率取决于测序错误率、所述第一区域的测序深度和所述第一区域中的核苷酸位置的数目。

58. 如权利要求57所述的计算机可读介质,其中基于所述第一区域中的真实阳性的数目确定所述第一截止值。

59. 如权利要求58所述的计算机可读介质,所述操作进一步包括:

基于所述生物样品中的所述第一区域的测序深度 D 和源自于肿瘤的DNA百分比浓度 f 来计算关于所述第一截止值真实阳性的数目。

60. 如权利要求59所述的计算机可读介质,其中使用根据下式的泊松分布概率计算真

实阳性的数目： $Pb = 1 - \sum_{i=0}^{r-1} P(i, M_p)$ ，其中 Pb 是检测真实阳性的概率，并且 r 是所述第一截止值，并且 $M_p = D \times f/2$ 。

61. 如权利要求55所述的计算机可读介质，其中使用以下准则中的任一者确定所述第一截止值：

如果所述测序深度小于50，那么所述第一截止值是5，
如果所述测序深度是50-110，那么所述第一截止值是6，
如果所述测序深度是111-200，那么所述第一截止值是7，
如果所述测序深度是201-310，那么所述第一截止值是8，
如果所述测序深度是311-450，那么所述第一截止值是9，
如果所述测序深度是451-620，那么所述第一截止值是10，和
如果所述测序深度是621-800，那么所述第一截止值是11。

62. 如权利要求51所述的计算机可读介质，其中所述参数是第一基因座的所述第一数目的加权和，其中每个第一基因座的贡献值是基于分配给各个第一基因座的权重值来求得。

63. 如权利要求51所述的计算机可读介质，其中所述参数包括在所述第一数目的第一基因座处含有单核苷酸变异体的所述序列标签的总和。

64. 如权利要求63所述的计算机可读介质，其中所述总和是加权和，并且其中所述第一个所述第一基因座中的第一权重不同于第二个所述第一基因座中的第二权重。

65. 如权利要求64所述的计算机可读介质，其中所述第一权重大于所述第二权重，并且其中第一个所述第一基因座是与癌症相关，并且第二个所述第一基因座是与癌症不相关。

66. 如权利要求51所述的计算机可读介质，其中所述参数是第一基因座的所述第一数目。

67. 如权利要求51所述的计算机可读介质，其中确定序列标签的基因组位置包括：

将至少一部分所述序列标签与参考基因组比对，其中序列标签的所述比对允许所述序列标签与所述参考基因组之间有一或多个错配。

68. 如权利要求67所述的计算机可读介质，其中将所述序列标签与所述一致性序列比较包括：

将所述一致性序列与所述参考基因组比较以确定相对于所述参考基因组具有变异序列的第二基因座的第二数目；

基于所述比对，确定第三基因座的第三数目，其中：

在每个第三基因座处，相对于所述参考基因组具有单核苷酸变异体的所述序列标签的数目高于截止值；和

获取所述第三数目与所述第二数目的差值以获得第一基因座的所述第一数目。

69. 如权利要求68所述的计算机可读介质，其中获取所述第三数目与所述第二数目的所述差值鉴定所述第一基因座。

70. 如权利要求69所述的计算机可读介质，其中确定所述参数包括：

对于所述具有第一数目的第一基因座中的每个基因座：

对与所述基因座比对上并且在所述基因座处具有单核苷酸变异体的序列标签的数目

进行计数;和

基于各个计数来确定所述参数。

71. 如权利要求51所述的计算机可读介质,其中所述一致性序列来源于所述受试者的含有多于50%组成型DNA的组成型样品。

72. 如权利要求51所述的计算机可读介质,其中确定序列标签的基因组位置包括:

将至少一部分所述序列标签与所述一致性序列比对,其中序列标签的所述比对允许所述序列标签与所述一致性序列之间有一或多个错配。

73. 如权利要求72所述的计算机可读介质,其中将所述序列标签与所述一致性序列比较包括:

基于所述比对,鉴别相对于所述受试者的所述一致性序列在基因组位置处具有单核苷酸变异体的序列标签;

对于展现单核苷酸变异体的每个基因组位置:

计算比对到所述基因组位置并且在所述基因组位置处具有单核苷酸变异体的序列标签的各个数目;

基于所述各个数目确定参数。

74. 如权利要求73所述的计算机可读介质,其中基于所述各个数目确定参数包括:

对所述各个数目求和以获得第一总和;和

使用所述第一总和以确定所述参数。

75. 如权利要求74所述的计算机可读介质,其中使用所述第一总和以确定所述参数包括:

由所述第一总和减去展现单核苷酸变异体的基因组位置的数目。

76. 如权利要求74所述的计算机可读介质,其中使用所述第一总和以确定所述参数包括:

基于比对上的序列标签的数量标准化所述第一总和。

77. 如权利要求51所述的计算机可读介质,所述操作进一步包括:

获得多个DNA片段中的每一者的一个或多个第二序列标签,其中所述一个或多个第二序列标签通过对来自含有多于90%组成型DNA的组成型样品的DNA片段的随机测序获得;

将至少一部分所述第二序列标签与参考基因组比对,其中第二序列标签的所述比对允许所述第二序列标签与所述参考基因组之间在M个基因组位置处有错配,其中M是等于或大于一的整数;和

基于所述第二序列标签和所述比对构建所述一致性序列。

78. 如权利要求77所述的计算机可读介质,其中所述组成型样品是所述生物样品,并且其中构建所述一致性序列包括:

确定纯合子基因座或具有两种等位基因的杂合子基因座。

79. 如权利要求77所述的系统,其中:

其中所述生物样品是获自血液样品的血浆或血清,以及

所述组成型样品是获自所述血液样品的白血球。

80. 如权利要求51所述的计算机可读介质,其中由对所述生物样品中的DNA片段进行随机测序产生所述一或多个序列标签。

81. 如权利要求51所述的计算机可读介质,其中所述参数是肿瘤来源的DNA的百分比浓度。

82. 如权利要求51所述的系统,其中所述序列标签提供全基因组水平的单核苷酸变异体检测。

83. 计算机可读介质,其存储用于控制计算机系统以执行分析受试者的一或多个肿瘤的异质性的操作的多个指令,所述操作包括:

获得所述受试者的一致性序列;

接收所述受试者的生物样品中的多个DNA片段中每一个的一或多个序列标签,所述生物样品包括无细胞的DNA分子;

确定所述序列标签的基因组位置;

将所述序列标签与所述一致性序列比较以确定第一基因座的数目,其中:

在每个第一基因座处,相对于所述一致性序列具有单核苷酸变异体的序列标签的第一数目高于截止值,所述截止值大于一;和

基于所述第一基因座的序列标签的各自第一数目计算所述一或多个肿瘤的异质性度量。

84. 如权利要求83所述的计算机可读介质,所述操作进一步包括:

将所述异质性度量与一或多个阈值比较以确定异质性等级的分类。

85. 如权利要求84所述的计算机可读介质,其中由一或多名其它受试者确定所述一个或多个阈值,并且其中使用一名或多名其它受试者的包括无细胞的DNA的生物样品的异质性的度量以确定阈值,所述一或多名其它受试者的肿瘤已经被活检和分析以确定经活检得到的肿瘤中的突变以确定肿瘤异质性等级。

86. 如权利要求85所述的计算机可读介质,其中将所述异质性度量与一或多个阈值比较包括:

将所述异质性度量输入到校准函数,所述校准函数基于所述异质性度量输出异质性等级。

87. 如权利要求83所述的计算机可读介质,其中所述异质性度量包括第一基因座的总数目,在所述第一基因座处,多于一个DNA片段比对到一致性序列同时在该基因座处具有单核苷酸变异体。

88. 如权利要求83所述的计算机可读介质,其中计算多个异质性度量,其中计算所述异质性度量包括:

对于第一基因座中的每一个,计算具有变异序列的序列标签的比例;

产生一个关于多个第一基因座的百分比数值的直方图;和

鉴别所述直方图中的多个峰值。

89. 如权利要求88所述的计算机可读介质,其中所述多个异质性度量之一对应于多个鉴别到的峰值。

90. 如权利要求88所述的计算机可读介质,其中所述多个异质性度量之一包括所述多个峰值中两个峰值的高度的比率。

91. 如权利要求88所述的计算机可读介质,其中所述比例各自代表在特定第一基因座处测量的肿瘤DNA的百分比浓度。

92. 如权利要求83所述的计算机可读介质,其中所述异质性度量对应于在第一基因座具有变异序列的序列标签的第一特定数量的第一比例与在第一基因座具有变异序列的序列标签的第二特定数量的第二比例的比率。

93. 如权利要求92所述的计算机可读介质,其中所述第一特定数量小于所述第二特定数量。

94. 如权利要求93所述的计算机可读介质,其中所述第一特定数量是第一范围并且所述第二特定数量是第二范围,所述第一范围低于所述第二范围。

95. 如权利要求92所述的计算机可读介质,其中所述第一特定数量和所述第二特定数量对应于具有变异序列的序列标签的百分比浓度或绝对数。

96. 如权利要求83所述的系统,其中所述一或多个肿瘤包括单一对象中的多个克隆,并且其中所述异质性包括肿瘤内异质性。

97. 如权利要求83所述的系统,其中所述一或多个肿瘤是多个肿瘤,并且其中所述异质性包括肿瘤间异质性。

98. 如权利要求83所述的计算机可读介质,其中由对应于各个所述第一基因座的第一数目的直方图确定所述异质性度量。

99. 如权利要求98所述的计算机可读介质,其中计算多个异质性度量,其中第一基因座包括第一子集和第二子集,其中多个异质性度量包括对应于所述第一子集的各个所述第一数目的第一直方图 and 对应于所述第二子集的各个所述第二数目的第二直方图。

100. 如权利要求83所述的计算机可读介质,其中所述异质性度量包括具有高于规定数值的对应的第一数目的第一基因座的比例。

用于癌症检测的血浆DNA突变分析

[0001] 相关申请的交叉引用

[0002] 本申请是2012年6月21日提交的标题是“用于癌症检测的血浆DNA突变分析 (MUTATIONAL ANALYSIS OF PLASMA DNA FOR CANCER DETECTION)”的美国临时专利申请第61/662,878号、2012年8月13日提交的标题是“用于癌症检测的血浆DNA突变分析”的美国临时专利申请第61/682,725号、2012年8月31日提交的标题是“用于癌症检测的血浆DNA突变分析”的美国临时专利申请第61/695,795号、和2012年10月8日提交的标题是“用于癌症检测的血浆DNA突变分析”的美国临时专利申请第61/711,172号的非临时申请,并且要求所述临时专利申请的权益,所述临时专利申请以全文引用的方式并入本文中用于所有目的。

背景技术

[0003] 已经显示,肿瘤来源的DNA存在于癌症患者的无细胞的血浆/血清中(陈XQ (Chen XQ) 等人自然医学 (Nat Med) 1996;2:1033-1035)。大多数现行方法基于对已知与癌症相关的突变进行直接分析(迪尔F (Diehl F) 等人美国国家科学院院刊 (Proc Natl Acad Sci) 2005;102:16368-16373;弗休T (Forsheew T) 等人科学·转化医学 (Sci Transl Med) 2012;4:136ra68)。另一种方法已经研究了通过对血浆DNA进行随机测序而检测的癌症相关基因组拷贝数变异(洛 (Lo) 等人的美国专利公开2013/0040824)。

[0004] 已知随着时间,多于一种癌细胞将获得生长优势并且产生多个子代细胞克隆。最终,肿瘤的生长和/或其转移灶将含有多种克隆癌细胞群的聚结物。此现象典型地称为肿瘤异质性(格林杰M (Gerlinger M) 等人新英格兰医学杂志 (N Engl J Med) 2012;366:883-892;叶TA (Yap TA) 等人科学·转化医学2012;4:127ps10)。

[0005] 已知癌症是高度异质的,即相同组织类型的癌症的突变分布可以大幅变化。因此,对特定突变进行直接分析典型地仅可以检测已知与那些特定突变相关的特定癌症类型内的病例的子集。另外,肿瘤来源的DNA通常是人类血浆中DNA的微量物质;血浆中DNA的绝对浓度低。因此,即使在患有已知具有标靶突变的癌症的患者中,直接检测血浆或血清中某一或某一小批癌症相关的突变也可能会获得低分析敏感性。此外,已经显示,即使在单一肿瘤内就突变来说也存在显著瘤内异质性。突变可以仅见于肿瘤细胞的亚群中。原发性肿瘤与转移性病变之间的突变分布差异甚至更大。一个关于瘤内和原发性-转移性间的异质性的实例包括罹患结肠直肠癌的患者中的KRAS、BRAF和PIK3CA基因(巴尔杜斯 (Baldus) 等人临床癌症研究 (Clin Cancer Research) 2010.16:790-9.)。

[0006] 在其中患者具有原发性肿瘤(携带KRAS突变但无PIK3CA突变)和隐蔽转移性病变(携带PIK3CA突变但无KRAS突变)的情形下,如果集中于检测原发性肿瘤中的KRAS突变,那么无法检测隐蔽转移性病变。然而,如果在分析中包括两种突变,那么可以检测原发性肿瘤和隐蔽转移性病变两者。因此,包括两种突变的测试在残余肿瘤组织的检测中将具有更高敏感性。当针对癌症进行筛选时,当拥有极少或不拥有关于可能出现的突变类型的信息,这种简单实例变得更复杂。

[0007] 因此需要提供新技术,以便针对癌症执行广泛筛选、检测或评估。

发明内容

[0008] 实施方案可以观察经历针对癌症的筛选或监测的受试者的生物样品(例如血浆或血清)中的体细胞突变频率,当与同一受试者的组成型DNA中的体细胞突变频率相比时。可以使用随机测序来测定这些频率。参数可以来源于这些频率并且用以确定癌症等级的分类。可以通过要求任何变异基因座都具有至少规定数目的变异序列读取(标签)来将假阳性过滤掉,由此提供更准确的参数。可以分析不同变异基因座的相对频率以测定患者中肿瘤的异质性等级。

[0009] 在一个实施方案中,可以将参数与来源于一组不患有癌症或具有低癌症风险的受试者的相同参数比较。获自测试受试者的参数与来自不患有癌症或具有低癌症风险的受试者群组的参数的显著差异可以表明,测试受试者患有癌症或癌变前病况或在将来将患上癌症的风险增加。因此,在一个实施方案中,可以在无肿瘤的先前基因组信息的情况下进行血浆DNA分析。这种实施方案因此尤其适用于筛选癌症。

[0010] 在另一个实施方案中,实施方案还可以用于监测治疗后的癌症患者并且查看是否存在残余肿瘤或肿瘤是否已经复发。举例来说,具有残余肿瘤或肿瘤已经复发的患者将具有比不存在残余肿瘤或未观察到肿瘤复发的患者中更高的体细胞突变频率。所述监测可以包括在治疗后的多个时间点从癌症患者获得样品,以便确定体液或具有无细胞的核酸的其它样品(例如血浆或血清)中肿瘤相关的遗传畸变的时间变化。

[0011] 根据一个实施方案,一种方法检测受试者中的癌症或癌变前变化。获得受试者的组成型基因组。接收受试者的生物样品中的多个DNA片段中每一个的一个或多个序列标签,其中生物样品包括无细胞的DNA。测定序列标签的基因组位置。将序列标签与组成型基因组比较以测定某第一基因座的第一数目。在每个第一基因座处,相对于组成型基因组具有变异序列的序列标签的数目高于某一截止值,其中截止值大于一。基于在第一基因座处具有变异序列的序列标签的数目测定参数。将参数与阈值比较以确定受试者中癌症等级的分类。

[0012] 根据另一个实施方案,一种方法分析受试者的一或多个肿瘤的异质性。获得受试者的组成型基因组。接收受试者的生物样品中的多个DNA片段中每一个的一个或多个序列标签,其中生物样品包括无细胞的DNA。测定序列标签的基因组位置。将序列标签与组成型基因组比较以测定第一基因座的第一数目。在每个第一基因座处,相对于组成型基因组具有变异序列的序列标签的数目高于某一截止值,其中截止值大于一。基于第一基因组位置集的各自第一数目计算一或多个肿瘤的异质性的度量。

[0013] 根据另一个实施方案,一种方法测定包括无细胞的DNA的生物样品中的肿瘤DNA的百分比浓度。接收生物样品中的多个DNA片段中的每一个的一个或多个序列标签。测定序列标签的基因组位置。对于多个基因组区域中的每一者,由在基因组区域内具有基因组位置的序列标签测定基因组区域内DNA片段的各自的量。将各自的量标准化以获得各自的密度。将各自的密度与参考密度比较以鉴别基因组区域是展现出1拷贝损失还是1拷贝增加。由鉴别为展现1拷贝损失的各自的密度或由鉴别为展现1拷贝增加的各自的密度计算第一密度。通过将第一密度与另一个密度比较以获得差别来计算百分比浓度,其中将差别用参考密度标准化。

[0014] 其它实施方案涉及与本文中所描述的方法相关的系统和计算机可读介质。

[0015] 可以参考以下具体实施方式和附图获得对本发明的本质和优势的更佳理解。

附图说明

[0016] 图1是根据本发明的实施方案检测受试者中的癌症或癌变前变化的方法100的流程图。

[0017] 图2展示了根据本发明的实施方案将样品基因组 (SG) 直接与组成型基因组 (CG) 比较的方法的流程图。

[0018] 图3展示了根据本发明的实施方案使用参考基因组 (RG) 将样品基因组 (SG) 与组成型基因组 (CG) 比较的方法300的流程图。

[0019] 图4是表400,其展示当假定样品中的肿瘤来源的DNA的百分比浓度是10%时,根据本发明的实施方案使用不同出现数目作为对样品中所存在的突变分类的标准正确鉴别的癌症相关的单核苷酸突变的数目。

[0020] 图5是表,其展示当假定样品中的肿瘤来源的DNA的百分比浓度是5%时鉴别的假阳性基因座的预期数目和突变的预期数目。

[0021] 图6A是图表600,其展示肿瘤来源的DNA的血浆百分比浓度是10%和20%的血浆中的癌症相关的突变的检测率并且使用四次和六次出现 (r) 作为呼叫潜在癌症相关的突变的准则。图6B是图表650,其展示相对于测序不同测序深度的错误分类,其分别使用突变核酸出现次数 (r) 4、5、6和7作为识别突变位点的标准。

[0022] 图7A是图表700,其展示当假定样品中的肿瘤来源的DNA的百分比浓度是5%时真实癌症相关的突变位点和假阳性位点的数目相对于不同测序深度的变化。图7B是图表750,其展示假阳性位点的预测数目,包括分析全基因组 (WG) 和所有外显子。

[0023] 图8是表800,其展示根据本发明的实施方案4名HCC患者在治疗前后的结果,包括血浆中的肿瘤来源的DNA的百分比浓度。

[0024] 图9是表900,其展示根据本发明的实施方案检测16名健康对照受试者中的HCC相关的SNV。

[0025] 图10A展示了根据本发明的实施方案HCC患者的肿瘤样品的序列读取密度的分布图。图10B展示了根据本发明的实施方案HCC患者的血浆中的所有基因组区段的z分数的分布图1050。

[0026] 图11展示了根据本发明的实施方案HCC患者的血浆的z分数的分布图1100。

[0027] 图12是根据本发明的实施方案测定包括无细胞的DNA的生物样品中的肿瘤DNA的百分比浓度的方法1200的流程图。

[0028] 图13A展示了根据本发明的实施方案在诊断时分析患有卵巢癌和乳癌的患者血浆中的突变的表1300。

[0029] 图13B展示了根据本发明的实施方案关于患有双侧卵巢癌和乳癌患者,分析在肿瘤切除之后的血浆中的突变的表1350。

[0030] 图14A是展示检测HCC1的血浆DNA中的单核苷酸变异的表1400。图14B是展示检测HCC2的血浆DNA中的单核苷酸变异的表1450。

[0031] 图15A是展示检测HCC3的血浆DNA中的单核苷酸变异的表1500。图15B是展示检测HCC4的血浆DNA中的单核苷酸变异的表1550。

- [0032] 图16是展示检测患有卵巢癌(和乳癌)的患者的血浆DNA中的单核苷酸变异的表1600。
- [0033] 图17是展示突变出现频率的不同要求和不同测序深度所对应的预测敏感性的表1700。
- [0034] 图18是展示针对不同截止值和不同测序深度所对应的假阳性基因座的预测数目的表1800。
- [0035] 图19展示了说明不同肿瘤位点检测到的突变数目的树图。
- [0036] 图20是表2000,其展示治疗前和治疗后血浆样品中携带肿瘤来源的突变的片段的数目。
- [0037] 图21是图表2100,其展示血浆中单一肿瘤位点检测到的突变和所有四个肿瘤位点都检测到的突变的出现频率分布。
- [0038] 图22是图表2200,其展示血浆中来自异质肿瘤的突变的预测出现频率分布。
- [0039] 图23展现了本发明的实施方案在16名招募的健康对照受试者中特异性。
- [0040] 图24是根据本发明的实施方案分析受试者的一或多个肿瘤的异质性的方法2400的流程图。
- [0041] 图25展示了可与根据本发明的实施方案的系统和方法一起使用的例示性计算机系统2500的框图。

具体实施方式

[0042] 定义

[0043] 如本文中所用,术语“基因座(locus)”或其复数形式“基因座(loci)”是任何长度的核苷酸(或碱基对)的位置或地址,其可以在基因组中具有变化。“区间(bin)”是基因组中的预定长度的区域。多个区间可以具有相同第一长度(分辨率),而不同多个可以具有相同第二长度。在一个实施方案中,区间不彼此重叠。

[0044] 如本文中所用,术语“随机测序”是指其中在测序程序之前所测序的核酸片段尚未具体鉴别或预定的测序。不需要序列特异性引物来靶向特定基因座。术语“通用测序”是指其中测序可以起始于任何片段的测序。在一个实施方案中,添加衔接子到片段的末端,并且使用于测序的引物连接到衔接子。因此,任何片段都可以用相同引物测序,并且因此测序可以是随机的。

[0045] 如本文中所用,术语“序列标签”(也称为序列读取)是指由核酸分子的任何部分或全部测序的核苷酸串。举例来说,序列标签可以是由核酸片段测序的短核苷酸串(例如约30个)、在核酸片段的两个末端的短核苷酸串或测序生物样品中存在的整个核酸片段。核酸片段是较大核酸分子的任何部分。片段(例如基因)可以与较大核酸分子的其它部分分开地存在(即不连接)。

[0046] 术语“组成型基因组”(也称为CG)由基因组内基因座处的共有核苷酸组成,并且因此可以视为一致性(共有)序列。CG可以覆盖受试者的整个基因组(例如人类基因组),或仅覆盖部分基因组。组成基因组(CG)可以获自细胞的DNA以及无细胞的DNA(例如如可以见于血浆中)。理想地,共有核苷酸应指示,基因座对于一种等位基因是纯合子或对于两种等位基因是杂合子。杂合子的基因座典型地含有两种等位基因,其是遗传多态性的成员。作为一

个实例,确定基因座是否是异型接合的准则可以是两种等位基因的阈值,所述等位基因各自以至少含有预定百分比(例如30%或40%)的读取比对到对应基因座上。如果一种核苷酸以足够大的百分比(例如70%或更大)呈现,那么可以将基因座测定为在CG中为纯合子纯合子。尽管一种健康细胞的基因组可以由于在细胞分裂期间自发出现的随机突变而不同于另一种健康细胞的基因组,但当使用这种共有核苷酸时CG不应不同。一些细胞可以具有具备基因组重排的基因组,例如B和T淋巴细胞,例如包括抗体和T细胞受体基因。所述大规模差异仍将是血液中的总体有核细胞群体的相对较小群体,并且因此所述重排将不影响以足够取样(例如测序深度)来测定血细胞的组成基因组。其它细胞类型(包括口腔细胞、皮肤细胞)、毛囊或各种正常身体组织的活检体也可以充当CG的来源。

[0047] 术语“组成型DNA”是指反映受试者出生时所具有的遗传组成的任何来源的DNA。对于受试者,可以获得组成型DNA的“组成样品”的实例包括健康血细胞DNA、口腔细胞DNA和毛根DNA。来自这些健康细胞的DNA界定了受试者的CG。可以按多种方式,例如当已知一个人不患有癌症或样品可以获自不大可能含有癌性或癌变前细胞的组织(例如当怀疑肝癌时,是毛根DNA)时,将细胞鉴别为健康的。作为另一个实例,当患者无癌时,可以获得血浆样品,并且将所测定的组成DNA针对后续血浆样品(例如一年或更久后)的结果进行比较。在另一个实施方案中,含有<50%的肿瘤DNA的单一生物样品可以用于推断组成型基因组和肿瘤相关的遗传变化。在这种样品中,肿瘤相关的单核苷酸突变的浓度将比CG中的杂合子SNP的每种等位基因的浓度低。这种样品可以与下文描述的用以测定样品基因组的生物样品相同。

[0048] 如本文中所用,术语“生物样品”是指获自受试者(例如人类、患有癌症的个人、疑似患有癌症的个人或其它生物体)并且含有一或多种所关注的无细胞的核酸分子的任何样品。生物样品可以包括无细胞的DNA,其中的一些可以来源于健康细胞并且一些来源于肿瘤细胞。举例来说,肿瘤DNA可以见于血液;或其它体液,例如尿液、胸膜液、腹水液、腹膜液、唾液、泪液或脑脊髓液。非体液实例是大便样品,其可以与腹泻液混合。对于一些所述样品,生物样品可以非侵入性地获得。在一些实施方案中,生物样品可以用作组成样品。

[0049] 术语“样品基因组”(也称为SG)是已经与基因组(例如人类基因组)的位置比对的一批序列读取。样品基因组(SG)不是共有序列,但包括可以仅以足够读取数目(例如至少2或3,或更高截止值)呈现的核苷酸。如果等位基因呈现足够次数并且其不是CG的一部分(即不是共有序列的一部分),那么等位基因可以指示“单核苷酸突变”(也称为SNM)。还可以使用本发明检测其它类型的突变,例如包括两种或更多种核苷酸的突变(例如影响微卫星或简单串联重复多态性中的串联重复单元的数目)、染色体易位(其可以是染色体内部或染色体间的)和序列反转。

[0050] 术语“参考基因组”(也称为RG)是指单倍或二倍基因组,可以将来自生物样品和组成型样品的序列读取与其比对并且比较。对于单倍基因组,每个基因座处仅存在一种核苷酸。对于二倍基因组,可以鉴别杂合子型基因座,这种基因座具有两种等位基因,其中任一等位基因可以允许与基因座匹配的比对。

[0051] 术语“癌症等级”可以指癌症是否存在、癌症的分期、肿瘤的大小和/或癌症的严重程度的其它度量。癌症等级可以是数字或其它特征。等级可以是零。癌症等级还包括与突变或突变的数目相关的癌变前或癌前期病况(状态)。癌症等级可以按各种方式使用。举例来说,筛选可以检查已知先前不患有癌症的某人是否存在癌症。评估可以研究已经诊断为患

有癌症的某人。检测可以意指‘筛选’或可以意指检查具有癌症的暗示特征(例如症状或其它阳性测试)的某人是否患有癌症。

[0052] 实施方式

[0053] 提供检测癌症的实施方案,其通过分析不由肿瘤直接获取并且包括无细胞的核酸的生物样品(例如血浆/血清样品)来进行。可以对于遍及身体的各种类型的组织产生无细胞的核酸。以此方式,可以执行检测各种癌症的广泛分析。

[0054] 遗传畸变(包括单核苷酸突变、缺失、扩增和重排)在患癌症期间积聚于肿瘤细胞中。在实施方案中,大规模平行测序可以用以检测和定量体液(例如血浆、血清、唾液、腹水、胸膜液和脑脊髓液)中的单核苷酸突变(SNM),也称为单核苷酸变异(SNV),以便检测和监测癌症。对SNM(或其它类型的突变)的数目进行定量可以提供作为筛选测试的一部分鉴别癌症早期的机制。在各种实施方案中,小心地区别测序错误和区别健康细胞中出现的自发突变(例如通过要求在特定基因座处鉴别多个SNM,例如至少3、4或5个)。

[0055] 一些实施方案还提供分析肿瘤异质性的非侵入性方法,其可以包括体内的同一肿瘤内的细胞(即瘤内异质性)或不同肿瘤的细胞(来自同一位点或来自不同位点)。举例来说,可以非侵入性地分析所述肿瘤异质性的克隆结构,包括评估含有每种突变的相对肿瘤细胞质量。以较高相对浓度存在的突变存在于体内的较大数目的恶性细胞中,例如相对于体内仍存在的其它恶性细胞在致瘤过程期间较早地出现的细胞(韦尔奇JS(Welch JS)等人细胞(Cell)2012;150:264-278)。由于所述突变具有较高相对丰度,预期其对于检测癌症DNA展现比具有较低相对丰度的突变更高的诊断敏感性。连续监测突变的相对丰度的变化将允许非侵入性地监测肿瘤的克隆架构随着疾病进展自发的或响应于治疗的变化。所述信息将适用于评估预后或早期检测肿瘤对治疗的抗性。

[0056] I. 引言

[0057] 突变可能在细胞分裂期间由于DNA复制和/或DNA修复的错误而出现。一种类型的所述突变包括单核苷酸的变化,其可以包括来自基因组不同部分的多种序列。癌症一般来说被认为是由于已经获得了生长优势的单癌细胞的克隆扩充。此克隆扩充将导致来源于上代癌细胞的所有癌细胞中突变(例如单核苷酸突变)积聚。这些子代肿瘤细胞将共享一套突变(例如单核苷酸突变)集。如本文中所描述,癌症相关的单核苷酸突变在癌症患者的血浆/血清中是可检测的。

[0058] 一些实施方案可以针对生物样品(例如血浆或血清)中的所有突变有效地筛选。因为突变的数目不固定(可以检测到来自不同肿瘤细胞亚群的数百、数千或数百万的癌症相关的突变),所以实施方案可以提供比检测特定突变的技术更佳的敏感性。突变的数目可以用以检测癌症。

[0059] 为了对许多或所有突变提供这种筛选,实施方案可以对生物样品(例如体液,包括血浆和血清)中的遗传变异执行检索(例如随机检索),所述生物样品可以含有肿瘤来源的DNA。使用样品(例如血浆)消除了对肿瘤或癌症执行侵入性活检的需要。此外,因为筛选可以覆盖基因组的所有或大区域,所以筛选不限于任何可列举并且已知的突变,但可以使用存在的任何突变。此外,因为累加所有跨越基因组的或大区域的突变的数目,所以可以获得较高敏感性。

[0060] 然而,在人类基因组中存在多态位点,包括单核苷酸多态性(SNP),其不应计数于

突变中。实施方案可以确定已经检测的遗传变异很可能是癌症相关的突变还是基因组中的多态性。举例来说,作为在癌症相关的突变与基因组中的多态性之间确定的一部分,实施方案可以测定组成型基因组,其可以包括多态性。组成基因组(CG)的多态性可以限制于在测序数据中以足够高的百分比(例如30-40%)展现的多态性。

[0061] 然后将获自生物样品的序列与组成型基因组比对,并且鉴别是单核苷酸突变(SNM)或其它类型的突变的变异。这些SNM将是不包括于已知多态性中的变异,并且因此可以标记为癌症相关的,并且不是组成型基因组的一部分。健康个人可以由于健康细胞之中的随机突变而具有一定数目的SNM,例如在细胞分裂期间产生,但患有癌症的个人将具有更多。

[0062] 举例来说,对于患有癌症的个人,体液中可检测的SNM的数目将高于同一个人的组成基因组中所存在的多态性。可以在含有肿瘤来源的DNA的体液样品与含有多半组成型DNA的DNA样品中检测到的变异的量之间进行比较。在一个实施方案中,术语‘多半’将意指多于90%。在另一个优选实施方案中,术语‘多半’将意指多于95%、97%、98%或99%。当体液中的变异的量超过具有多半组成DNA的样品的变异量时,体液可能含有肿瘤来源的DNA的可能性增加。

[0063] 一种可以用以针对DNA样品中的变异随机检索的方法是随机或鸟枪测序(例如使用大规模平行测序)。可以使用任何大规模平行测序平台,包括连接测序平台(例如生命技术(Life Technologies) SOLiD平台)、离子激流/离子质子(Ion Torrent/Ion Proton)、半导体测序、罗氏(Roche) 454、单分子测序平台(例如赫利克斯(Helicos)、太平洋生物科学(Pacific Biosciences)和纳米孔)。然而,已知测序错误可能会出现并且可能会曲解为组成DNA中的变异或来源于肿瘤DNA的突变。因此,为改进我们提出的方法的特异性,可以考虑造成测序错误或分析错误的其它组分的概率,例如通过使用在某个基因座上具有适当测序深度以及满足至少规定数目(例如2或3个)要求的等位基因数目来确定SNM。

[0064] 如本文中所描述,当样品中所存在的随机检测的遗传变异的量超过可能由于分析错误(例如测序错误)无意检测的组成型DNA的变异的预期量时,实施方案可以提供生物样品(例如体液)中的存在肿瘤来源的DNA的证据。所述信息可以用于筛选、诊断、预测和监测癌症。在以下部分中,我们描述可以用于检测血浆/血清或其它样品(例如体液)中的单核苷酸突变的分析步骤。体液可以包括血浆、血清、脑脊髓液、胸膜液、腹水液、乳头溢液、唾液、支气管肺泡灌洗液、痰液、泪液、汗液和尿液。除了体液之外,所述技术还可以应用于大便样品,因为后者已经有报道显示含有来自结肠直肠癌的肿瘤DNA(伯杰BM(Berger BM),阿尔奎斯特DA(Ahlquist DA).病理学(Pathology) 2012;44:80-88)。

[0065] II. 通用筛选方法

[0066] 图1是根据本发明的实施方案检测受试者中的癌症或癌变前变化的方法100的流程图。实施方案可以分析受试者的生物样品中的无细胞的DNA以检测很可能由肿瘤产生的无细胞的DNA的变异。分析可以使用受试者的组成型基因组可以解释是健康细胞的一部分的多态性,并且解释测序错误。方法100和本文中所描述的任何方法可以完全或部分用计算机系统执行,所述计算机系统包括一或多个处理器。

[0067] 在步骤110中,获得受试者的组成型基因组。可以由所测试的受试者的组成型DNA测定组成型基因组(CG)。在各种实施方案中,CG可以由存储器读取或例如通过分析组成型

DNA的序列读取而有效测定,所述组成型DNA可以在来自包括无细胞的DNA的样品的细胞中。举例来说,当怀疑非血液恶性病时,可以分析血细胞以测定受试者的组成DNA。

[0068] 在各种实施方案中,对组成型DNA的分析可以使用大规模平行测序、基于阵列的杂交、基于探针的溶液中杂交、基于连接的分析、引物延伸反应分析和质谱分析来执行。在一个实施方案中,CG可以在受试者生命的一个时间点、例如在出生时或甚至产前时期中测定(其可以使用胎儿细胞或经由无细胞的DNA片段进行,参看美国公开2011/0105353),并且然后当在受试者生命的其它时间获得体液或其它样品时被参考。因此,CG可以简单地由计算机存储器读取。组成型基因组可以读出为基因座列表,其中组成型基因组不同于参考基因组。

[0069] 在步骤120中,对于受试者的生物样品中的多个DNA片段中的每一者,获得一或多种序列标签,其中生物样品包括无细胞的DNA。在一个实施方案中,由对生物样品中的DNA片段进行随机测序产生一个或多个序列标签。当执行双末端测序时,可以获得多于一种序列标签。一种标签将对应于DNA片段的每个末端。

[0070] 可以分析样品(例如血浆、血清或其它体液)中的无细胞的DNA以对遗传变异检索。可以使用与已经用于分析组成DNA的平台相同的分析平台分析无细胞的DNA。或者,可以使用不同分析平台。举例来说,可以使用大规模平行测序对无细胞的DNA样品测序,或可以在大规模平行测序之前捕获基因组的部分或使其富集。如果使用富集,那么可以例如使用基因组的所选部分的液相或固相捕获。然后,可以对所捕获的DNA进行大规模平行测序。

[0071] 在步骤130中,测定序列标签的基因组位置。在一个实施方案中,将序列标签与参考基因组比对,所述参考基因组获自一或多名其它受试者。在另一个实施方案中,将基因组序列标签与所测试的受试者的组成基因组比对。可以使用本领域技术人员已知的技术、例如使用局部序列基本搜索工具(BLAST)执行比对。

[0072] 在步骤140中,测定基因座的第一数目,其中至少N个序列标签相对于组成基因组(CG)具有变异序列。N等于或大于二。如下文更详细论述,可以通过使N等于2、3、4、5或更高来去除测序错误以及细胞中随机出现的体细胞突变(例如由于细胞分裂)。满足一或多个规定准则的基因座可以鉴别为突变(变异体)或突变基因座(变异基因座),然而具有变异体但不满足一个或多个准则的基因座(例如如仅一个变异序列标签)称为潜在或推定的突变。序列变异体可以针对仅一个核苷酸或多个核苷酸。

[0073] N可以测定为基因座的总标签的百分比,与绝对值相对。举例来说,当由变异体读取推断的肿瘤DNA的相对百分比浓度测定为等于或大于10%(或某一其它百分比)时,可以鉴别变异基因座。换句话说,当基因座由200个序列读取覆盖时,可能需要至少10个序列读取展示变异等位基因的标准来将变异体界定为突变。变异等位基因的10个序列读取和野生型等位基因的190个读取将得到10% ($2 \times 10 / (10 + 190)$) 的肿瘤DNA百分比浓度。

[0074] 在一个实施方案中,可以将序列标签(统称为样品基因组)直接与CG比较以测定变异体。在另一个实施方案中,通过参考基因组(RG)将样品基因组(SG)与CG比较以测定变异体。举例来说,可以将CG和SG两者与RG比较以测定展现变异体的基因座的分别的数目(例如集),并且然后可以取得差值以获得基因座的第一数目。第一数目可以简单地以数目形式获得或可以对应于一套特定基因座的集合,其然后可以经进一步分析以由第一基因座处的序列标签测定参数。

[0075] 在一个实施方案中,将组成型DNA和血浆DNA的测序结果比较以确定单核苷酸突变是否存在于血浆DNA中。可以分析组成型DNA纯合子的区域。出于说明的目的,假定特定基因座的基因型在组成型DNA中是纯合子并且是AA。然后在血浆中,不是A的等位基因的存在将指示在特定基因座处可能存在单核苷酸突变(SNM)。指示可能存在SNM的基因座可以形成步骤140中基因座的第一数目。

[0076] 在一个实施方案中,可以适用的是靶向基因组的已知在特定癌症类型中或在所述群体的特定子集中特别易于遭受突变的部分。与后一方面相关地,实施方案可以寻找特定群体组中特别普遍的突变类型,例如是B型肝炎病毒(对于肝癌)或人乳头瘤病毒(对于宫颈癌)的携带者的或具有倾向于体细胞突变的遗传因子的受试者或者在DNA错配修复基因中具有种系突变的受试者中尤其常见的突变。所述技术还将适用于针对具有BRCA1或BRCA2突变的受试者的卵巢癌和乳腺癌中的突变进行筛选。所述技术类似地将适用于针对具有APC突变的受试者的结肠直肠癌中的突变进行筛选。

[0077] 在步骤150中,基于计算在第一基因座处具有变异序列的序列标签的数目测定参数。在一个实例中,参数是基因座的第一数目,其中至少N个DNA片段在基因座处相对于组成基因组具有变异序列。因此,计数可以简单地用以确保基因座具有多于N个拷贝的特定变异体被鉴别,随后包括于第一数目中。在另一个实施方案中,参数可以是或包括在第一基因座处相对于组成基因组具有序列变异体的序列标签的总数目。

[0078] 在步骤160中,将受试者的参数与阈值(例如来源于一或多名其它受试者)比较以确定受试者中癌症等级的分类。癌症等级的实例包括受试者是否具有癌症还是癌变前病况,或增加的患癌症可能性。在一个实施方案中,可以由来自受试者的先前获得的样品测定阈值。

[0079] 在另一个实施方案中,一或多名其它受试者可以测定为不患有癌症或具有低癌症风险。因此,阈值可以是正常值、正常范围,或指示与正常值或范围有统计性显著偏差。举例来说,不患有癌症或具有低癌症风险的受试者的血浆中相对于特定受试者的CG的可检测的突变数目可以用作正常范围以确定所测试的受试者中检测到的突变数目是否正常。在另一个实施方案中,其它受试者可以已知患有癌症,并且因此类似突变数目可以指示癌症。

[0080] 在一个实施方案中,其它受试者可以经选择以使临床特征与测试受试者的临床特征(例如性别、年龄、饮食、吸烟习惯、药物史、先前疾病、家族史、所选基因组基因座的基因型、病毒感染(例如B或C型肝炎病毒或人乳头瘤病毒或人类免疫缺陷病毒或艾伯斯坦-巴尔病毒(Epstein-Barr virus)感染)或其它传染因子(例如细菌(例如幽门螺旋杆菌(*Helicobacter pylori*)) 和寄生虫(例如华支睾吸虫(*Clonorchis sinensis*)))感染的状态等)匹配。举例来说,是B或C型肝炎病毒的携带者的受试者具有增加的患上肝细胞癌的风险。因此,具有与B或C型肝炎的携带者类似数目或模式的突变的测试受试者可以视为具有增加的患上肝细胞癌的风险。另一个方面,如果使用适当基线(即相对于另一名肝炎患者),比另一名肝炎患者展现更多突变的B或C型肝炎患者可以适当地鉴别为具有更高癌症等级分类。类似地,是人乳头瘤病毒感染的携带者的受试者具有增加的宫颈癌和头颈癌风险。艾伯斯坦-巴尔病毒感染已经与鼻咽癌、胃癌、霍奇金淋巴瘤(Hodgkin's lymphoma)和非霍奇金淋巴瘤(non-Hodgkin's lymphoma)相关。幽门螺旋杆菌感染已经与胃癌相关。华支睾吸虫感染已经与胆管癌相关。

[0081] 监测不同时间点突变数目的变化可以用于监测癌症进展和治疗反应。所述监测还可以用以记录癌变前病况的进展或受试者将患上癌症的风险的变化。

[0082] 展示变异的序列标签的量也可以用以监测。举例来说,可以使用基因座处变异体读取的百分比浓度。在一个实施方案中,在连续监测期间样品中肿瘤相关的遗传畸变的百分比浓度的增加可以表示疾病进展或即将发生复发。类似地,在连续监测期间样品中肿瘤相关的遗传畸变的百分比浓度的减小可以表示治疗反应和/或缓解和/或良好预后。

[0083] III. 测定基因组

[0084] 下文更详细说明上文论述的各种基因组。举例来说,论述参考基因组、组成型基因组和样品基因组。

[0085] A. 参考基因组

[0086] 参考基因组 (RG) 是指受试者的或群体共有的单倍或二倍基因组。参考基因组是已知的并且因此可以用以比较来自新患者的测序读取。可以将来自患者的样品的序列读取与 RG 比对并且比较以鉴别读取的变异。对于单倍基因组,每个基因座处仅存在一种核苷酸,并且因此每个基因座可以视为半合子的。对于二倍基因组,可以鉴别杂合子基因座,这种基因座具有两种等位基因,其中任一等位基因可以允许与基因座的某一比对相匹配。

[0087] 参考基因组在受试者群体之中可以是相同的。此相同参考基因组可以用于健康受试者以测定用于对患者分类(例如患有或不患有癌症)的适当阈值。然而,不同参考基因组可以用于不同群体,例如用于不同种族或甚至用于不同家族。

[0088] B. 组成基因组

[0089] 受试者(例如人或其它二倍生物体)的组成型基因组 (CG) 是指受试者的二倍基因组。CG 可以规定杂合性基因座,其中第一等位基因来自第一单倍型并且不同的第二等位基因来自第二单倍型。注意,覆盖该对杂合子基因座的两种单倍型的结构不需要已知,即哪一个杂合子基因座上的等位基因与另一个杂合子基因座的等位基因是相同单倍型的结构即不需要已知。仅知道在每个杂合子基因座处存在两种等位基因就足够的。

[0090] CG 可以由于多态性而不同于 RG。举例来说, RG 上的基因座对于 T 可以是纯合的,但 CG 对于 T/A 是杂合的。因此, CG 将在此基因座处展现变异。CG 还可以由于遗传的突变(例如其在家族中蔓延)或新生突变(其出现于胎儿中,但其不存在于其父母中)而不同于 RG。遗传的突变典型地称为‘种系突变’。一些所述突变与癌症倾向性相关,例如在家族中遗传的 BRCA1 突变。所述突变不同于可以在人一生中由于细胞分裂而出现的‘体细胞突变’并且可以推动细胞和其子代即将变为癌症。

[0091] 测定 CG 的目标是从样品基因组 (SG) 的突变中去除所述种系突变和新生突变以便鉴别体细胞突变。SG 中体细胞突变的量因此可以用以评估受试者中癌症的可能性。可以经进一步对这些体细胞突变过滤以去除测序错误,并且可能去除很少出现的体细胞突变(例如展示变异体的仅一种读取),因为这些该去除的所述体细胞突变不大可能与癌症相关。

[0092] 在一个实施方案中,可以使用细胞(白血球 DNA)测定 CG。然而,同样还可以由无细胞的 DNA(例如血浆或血清)测定 CG。对于其中大部分细胞是非恶性的样品类型,例如来自健康受试者的白血球,则大多数或一致性的基因组是 CG。对于 CG,每个基因组中的基因座由所取样的组织中的大部分细胞所具有的 DNA 序列组成。测序深度应足以阐明组成型基因组内的杂合位点。

[0093] 作为另一个实例,血浆可以用作组成型样品以测定CG。举例来说,对于其中血浆中的肿瘤DNA少于50%并且SNM处于杂合子状态的情况,例如突变是新增加的等位基因,则新等位基因可以具有小于25%的浓度。然而,CG中SNP的杂合等位基因的浓度应约为50%。因此,可以在体细胞突变与CG的多态性之间加以区别。在一个实施方案中,当使用血浆或具有显著肿瘤浓度的其它混合物时,可以在30-40%之间使用合适截止值加以区分体细胞突变和多态性位点。肿瘤DNA浓度的测量可以适用于确保血浆中的肿瘤DNA少于50%。测定肿瘤DNA浓度的实例描述于本文中。

[0094] C. 样品基因组

[0095] 样品基因组(SG)不简单地如同于RG和CG情况下的单倍或二倍基因组。SG是一批来自样品的读取,并且可以包括:来自组成型DNA的对应于CG的读取、来自肿瘤DNA的读取、来自健康细胞的相对于CG展示随机突变(例如由于由细胞分裂产生的突变)的读取和测序错误。各种参数可以用以确切控制哪些读取包括于SG中。举例来说,要求等位基因在至少5个读取从而减少测序错误在SG中出现,以及减小源自于随机突变的读取。

[0096] 作为一个实例,假定受试者是健康的,即并不患有癌症。出于说明的目的,来自1000个细胞的DNA处于获自此受试者的1ml血浆(即1000基因组当量的DNA)中。血浆DNA典型地由约150bp的DNA片段组成。因为人类基因组是 3×10^9 bp,所以每单倍基因组将存在约 2×10^7 个DNA片段。因为人类基因组是二倍的,所以每ml血浆将存在约 4×10^7 个DNA片段。

[0097] 因为每单位时间在血浆中数百万到数十亿细胞释放其DNA并且来自这些细胞的片段在循环期间将混合在一起,所以 4×10^7 个DNA片段可以来自 4×10^7 个不同细胞。如果这些细胞彼此不带有相近的(与相远相对,例如原始受精卵)克隆关系(即其不共有最近上代细胞),那么统计上可能的是在这些片段之中将可见突变不多于一次。

[0098] 另一个方面,如果在每ml 1000基因组当量的血浆DNA之中,存在一定百分比的细胞共有相近的上代细胞(即其彼此共享相关的克隆关系),那么可以看到来自此克隆的突变优先呈现于血浆DNA中(例如在血浆中展现克隆突变分布)。所述克隆地相关的细胞可以是癌细胞,或正要变为癌症但尚未是癌症的(即肿瘤前的)细胞。因此,要求突变出现多于一次可以去除样品中鉴别到“突变”中的此中自然的差异,其可以留下与癌细胞或肿瘤前细胞相关的更多突变,由此允许对癌症或癌前期病况进行检测、尤其早期检测。

[0099] 在一个近似情况中,已经陈述,平均一个突变将在每次细胞分裂之后积聚于基因组中。先前研究已经显示,大部分血浆DNA来自造血细胞(磊YY(Lui YY)等人临床化学(Clin Chem) 2002;48:421-427)。据估计,造血干细胞每25-50周复制一次(卡特林SN(Catlin SN)等人血液(Blood) 2011;117:4460-4466)。因此,作为简单化的近似情况,40岁的健康受试者将每造血干细胞积聚大约40到80个突变。

[0100] 如果每ml此人的血浆中存在1000基因组当量,并且如果这些细胞中的每一者来源于不同造血干细胞,那么在 4×10^{10} 个DNA片段之中可以预期40,000到80,000个突变(即每基因组的 4×10^7 个DNA片段和每ml血浆1000基因组当量)。然而,因为每种突变将可见一次,所以每种突变仍可能低于检测极限(例如在截止值N大于1时),并且因此这些突变可以被过滤掉,由此使得分析可集中于更可能由癌性病况产生的突变。截止值可以是大于1的任何值(整数或非整数),并且对于不同基因座和区域可以是动态的。肿瘤DNA的测序深度和分数浓度还可以影响由癌细胞或肿瘤前细胞检测突变(例如可检测的突变的百分比)的敏感性。

[0101] IV. 将SG直接与CG比较

[0102] 一些实施方案可以鉴别CG是纯合的核苷酸位置,但其中SG中的少数物质(即肿瘤DNA)是杂合的。当以高深度(例超过50倍覆盖)对一个位置测序时,可以检测在健康与癌细胞的DNA混合物中在所述位置处是否存在一或两种等位基因。当检测到两种等位基因时,(1)CG是杂合的,或(2)CG是纯合的但SG是杂合的。这两种情形可以通过查看主要和次要等位基因的相对计数来区分。在上一情形下,两种等位基因将具有类似计数数目;但在下一情形下,其计数数目将存在大差异。测试样品的读取在等位基因上此相对计数的比较是将序列标签与组成基因组比较的一个实施方案。方法100的第一基因座可以由其中等位基因的数目低于上限阈值(对应于CG中的多态性的阈值)并且高于下限阈值(对应于与癌性病况不相关的且以足够低的比率出现的错误和体细胞突变的阈值)的基因座。因此,可以同时测定组成基因组和第一基因座。

[0103] 在另一个实施方案中,鉴别突变的方法可以首先测定CG,并且然后测定相对于CG具有足够突变数目的基因座。可以由不同于测试样品的组成型样品测定CG。

[0104] 图2展示了根据本发明的实施方案将样品基因组(SG)直接与组成型基因组(CG)比较的方法200的流程图。在框210处,获得受试者的组成型基因组。可以例如由先前适时取得的样品或临在实施方法200之前获得并且分析的组成型样品获得组成基因组。

[0105] 在框220处,对于受试者的生物样品中的多个DNA片段中的每一者,获得一个或多个序列标签。可以使用如本文中提及的各种技术执行测序。序列标签是对片段的序列内容为何物的测量。但序列标签的一个或多个碱基可以是错误的。

[0106] 在框230处,将至少一部分序列标签与组成基因组比对。比对可以涉及CG在许多基因座处是杂合的。比对将不需要确切匹配以使得可以检测变异体。

[0107] 在框240处,鉴别在基因座处相对于组成型基因组具有变异序列的序列标签。有可能的是序列标签可以具有多于一个变异体。可以针对每个基因座和每种序列标签追踪变异体。变异体可以是不在CG中的任何等位基因。举例来说,CG相对于A/T而言是杂合性,并且变异体可以是G或C。

[0108] 在框250处,对于具有变异体的每个基因座,计算机系统可以对与基因座比对并且在基因座处具有变异序列的序列标签各自对应的第一数目计数。因此,每个基因座可以具有相关计数的数目的变异体分布在其基因座处。典型地,与对应于CG的序列标签相比,在基因座处将可见较少变异体,例如由于肿瘤DNA浓度小于50%。然而,一些样品的肿瘤DNA浓度可以大于50%。

[0109] 在框260处,基于各自第一数目测定参数。在一个实施方案中,如果各自的数目大于截止值(例如大于二),那么各自数目可以加到总和,其即是参数或用以测定参数。在另一个实施方案中,各自数目大于截止值的基因座的数目用作参数。

[0110] 在框270处,将参数与阈值比较以对癌症等级分类。如上文所述,可以通过分析来自其它受试者的样品测定阈值。取决于这些其它受试者的健康或癌症状态,可以确定分类。举例来说,如果其它受试者患有4期癌症,那么如果当前参数接近于(例如在特定范围内)获自其它受试者的参数的值,则当前受试者可以分类为患有4期癌症。然而,如果参数超过阈值(即大于或小于,取决于如何界定参数),那么分类可以鉴别为不到4期。当其它受试者不患有癌症时,可以进行类似分析。

[0111] 可以使用多个阈值来确定分类,其中由不同组受试者集合测定每个阈值。每个受试者集合可以具有常见癌症等级。因此,可以将当前参数与每组受试者集合的值比较,其可以提供与所述集合中的一者的匹配或提供范围。举例来说,参数可以约等于关于癌症前期的或在2期的受试者获得的参数。作为另一个实例,当前参数可以处于可能与数种不同癌症等级匹配的范围。因此,分类可以包括多于一种癌症等级。

[0112] V. 使用参考基因组

[0113] 可以将组成型DNA和来自生物样品的DNA两者的基因组序列与人类参考基因组比较。当与参考基因组相比,如果血浆样品中比组成型DNA存在更多变化时,则存在更高癌症概率。在一个实施方案中,研究参考基因组中的纯合基因座。比较组成DNA和来自生物样品的DNA两者中的杂合基因座的量。当由生物样品的DNA检测的杂合位点的量超过组成DNA的杂合位点量时,存在更高癌症概率。

[0114] 分析还可以限于CG中纯合的基因座。对于杂合基因座同样可以界定SNM,但这一般来说将需要产生第三种变异体。换句话说,如果杂合基因座是A/T,那么新变异体将是C或G。鉴别纯合基因座的SNM一般来说更容易。

[0115] 当与健康受试者中可见的变化率相比时,生物样品DNA中的杂合基因座的量相对于组成型DNA的增加的程度可以暗示癌症或癌变前状态。举例来说,如果所述位点的增加程度超过与健康受试者中所观察到的程度对应的某一阈值,那么可以将所述数据视为暗示癌症或癌变前状态。在一个实施方案中,确定不患有癌症的受试者中的突变分布,并且可以将阈值视为一定数目的标准偏差(例如2或3个标准偏差)。

[0116] 一个实施方案可以要求在基因座处有至少规定数目的变异体,随后才对所述基因座计数。另一个实施方案提供测试,甚至用于基于看到只有一次变化的数据。举例来说,当血浆中可见的变异的总数目(错误+真突变或多态性)统计上显著高于组成型DNA中的总数目时,则为存在癌症的证据。

[0117] 图3展示了根据本发明的实施方案使用参考基因组(RG)将样品基因组(SG)与组成型基因组(CG)比较的方法300的流程图。方法300假定已经获得RG,并且已经获得生物样品的序列标签。

[0118] 在框310处,将至少一部分序列标签与参考基因组比对。比对可以允许错配当变异要被检测。参考基因组可以来自与受试者类似的群体。比对的序列标签有效地包含样品基因组(样品基因组)。

[0119] 在框320处,鉴别第一数目(A)的潜在变异体,例如单核苷酸突变(SNM)。潜在SNM是其中SG的序列标签展示不同于RG的核苷酸的基因座。可以使用其它准则,例如展示变异的序列标签的数目必须大于截止值和基因座在RG中是否是纯合的。当特定基因座被鉴别并且通过将基因座存储于存储器中而追踪时,潜在SNM集合可以集合A。可以测定特定基因座或可以直接测定若干所述SNM。

[0120] 在框330处,通过将通过对来自组成型样品的DNA片段测序而获得的序列标签与参考基因组比对来测定组成基因组。可以在任何先前时间并且使用在任何先前时间获得的组成型样品来执行此步骤。CG可以简单地由存储器读取,其中事先进行比对。在一个实施方案中,组成型样品可以是血细胞。

[0121] 在框340处,鉴别第二数目(B)的基因座,其中CG的比对的序列标签在基因座处相

对于参考基因组具有变异体(例如SNM)。如果具体来说追踪基因座集合,那么B可以代表所述集,与仅一数目相对。

[0122] 在框350处,由集合A减去集合B以鉴别存在于样品基因组中但不存在于CG中的变异体(SNM)。在一个实施方案中,SNM集可以限于CG是纯合的核苷酸位置。为实现此过滤,可以在集C中鉴别其中CG是纯合的特定基因座。在另一个实施方案中,如果在基因座处CG不是纯合的,那么该基因座不计数于第一数目A或第二数目B中。在另一个实施方案中,可以将任何已知多态性(例如凭借其于SNP数据库中的存在)过滤掉。

[0123] 在一个实施方案中,框350中的减法可以仅是数值的减法,并且因此不去除特定潜在SNM,但仅减去值。在另一个实施方案中,减法取得集合A与集合B之间的差值(例如在集合B是集合A的子集时)以鉴别不在集合B中的特定SNM。在逻辑值中,此可以表示为[A AND NOT (B)]。所得鉴别的变异体集合可以标记为C。参数可以测定为数目C或由集合C测定。

[0124] 在一些实施方案中,可以考虑突变的本质,并且不同权重对应于不同突变类别。举例来说,通常与癌症相关的突变可以赋予较高权重(当涉及基因座的相对权重时,也称为重要值)。所述突变可以见于肿瘤相关的突变的数据库中,例如癌症中的体细胞突变的目录(Catalogue of Somatic Mutations in Cancer, COSMIC) (www.sanger.ac.uk/genetics/CGP/cosmic/)。作为另一个实例,与非同义变化相关的突变可以赋予较高权重。

[0125] 因此,第一数目A可以测定为加权和,其中在一个基因座处展示变异体的标签的计数可以具有与在另一个基因座处的标签的计数不同的权重。第一数目A可以反映此加权和。因此,可以对B和数目C执行类似计算,并且参数可以反映此权重。在另一个实施方案中,当测定特定基因座集C时,考虑整合此权重。举例来说,可以针对集C的基因座的计数测定加权和。所述权重可以用于本文中所描述的其它方法。

[0126] 因此,与阈值比较以确定癌症等级的分类的参数可以是相对于RG展现SG和CG的变异的基因座的数目。在其它实施方案中,可以计数对展示变异的DNA片段(如经由序列标签计数)的总数目。在其它实施方案中,可以在另一个式中使用所述数目以获得参数。

[0127] 在一个实施方案中,每个基因座处变异体的浓度可以是参数并且与阈值进行比较。此阈值可以用以确定基因座是否是潜在变异基因座(包括展示变异体的读取的特定数目的截止值),并且然后对基因座进行计数。浓度还可以在SNM的总和中用作加权因子。

[0128] VI. 使用截止值减少假阳性

[0129] 如上文所提及,可以在多个无细胞的DNA片段(例如血浆中的循环DNA)中针对大基因组区域(例如整个基因组)或多个基因组区域勘察单核苷酸突变来改进方法的敏感性。然而,分析错误(例如测序错误)可以影响此方法的可行性、准确性和特异性。此处,我们使用大规模平行测序平台作为一个实例来说明测序错误的重要性。伊路米那(Illumina)合成测序平台的测序错误率是每测序的核苷酸约0.1%到0.3%(米诺切(Minoche)等人基因组生物学(Genome Biol) 2011, 12:R112)。可以使用任何大规模平行测序平台,包括连接测序平台(例如生命技术SOLiD平台)、离子激流/离子质子、半导体测序、罗氏454、单分子测序平台(例如赫利克斯、太平洋生物科学和纳米孔)。

[0130] 在先前关于肝细胞癌的研究中,已经显示,对于癌症全基因组存在约3,000个单核苷酸突变(陶Y(Tao Y)等人2011美国国家科学院院刊(Proc Natl Acad Sci USA); 108: 12042-12047)。假定循环中仅10%的总DNA来源于肿瘤细胞,并且我们以单倍型基因组覆盖

度为1的平均测序深度对血浆DNA测序,我们将由于测序错误而遇到9百万($3 \times 10^9 \times 0.3\%$)个单核苷酸变异(SNV)。然而,预期大部分单核苷酸突变仅出现于两种同源染色体中的一者上。在测序深度在具有100%肿瘤DNA的样品中为单倍型基因组覆盖度为1的情况下,我们期望仅检测到3,000个突变的一半,即1,500个突变。当我们以单倍型基因组覆盖为1对含有10%肿瘤来源的DNA的血浆样品测序时,我们期望仅检测到150($1,500 \times 10\%$)个癌症相关的单核苷酸突变。因此,癌症相关的突变的检测的信噪比是1比60,000。此极低信噪比表明,如果我们简单地使用生物样品(例如血浆)中的所有单核苷酸变化作为参数,那么使用此方法区分正常和癌症情况的准确性将极低。

[0131] 预期随着测序技术的进展,测序错误率将不断减小。还可以使用多于一个测序平台来分析相同样品,并且通过比较跨平台测序结果来指出可能受测序错误影响的读取。另一种方法是分析在不同时间由同一受试者获取的两种样品。然而,所述方法是耗时的。

[0132] 在一个实施方案中,在检测癌症患者的血浆中的单核苷酸突变中增强信噪比的一种方法是仅当样品中多次出现相同突变时才对突变计数。在所选测序平台中,包括特定核苷酸取代性的测序错误可能更常见并且将影响测试受试者和对照受试者两者的测试样品和组成型DNA样品的测序结果。然而,一般来说,测序错误随机出现。

[0133] 当在多个DNA片段中的相同核苷酸位置处观察到相同变化时,具有测序错误的机率将指数性较低。另一个方面,检测到样品中的真癌症相关的突变变化的机率受样品中的肿瘤DNA的测序深度和百分比浓度影响。在多个DNA片段中观察到突变的机率将随肿瘤DNA的测序深度和百分比浓度而增加。在使用具有无细胞的肿瘤DNA(例如血浆中)的样品的各种实施方案中,分数浓度可以是5%、10%、20%和30%。在一个实施方案中,分数浓度小于50%。

[0134] 图4是表400,其展示根据本发明的实施方案使用不同出现频数作为对样品中所存在的突变分类的准则以正确鉴别的癌症相关的单核苷酸突变的数目。还展示基于相同分类准则由于测序错误而错误地鉴别为具有突变的核苷酸位置的数目。假定测序错误率是0.1%(米诺切等人基因组生物学2011,12:R112)。假定样品中肿瘤来源的DNA的分数浓度是10%。

[0135] 图4显示,当假定样品中的肿瘤来源的DNA的分数浓度是10%时,血浆中检测到的癌症相关的突变的数目与假阳性的数目之间的比率将随在样本中用于界定突变的相同变化的次数增加而指数增加。换句话说,癌症突变检测的敏感性和特异性两者都将改进。另外,检测癌症相关的突变的敏感性受测序深度影响。在测序的单倍型基因组覆盖度100次的情况下,即使使用在样品中的至少4个DNA片段中出现特定突变的准则也可以检测到3,000个突变中的2,205个(73.5%)。可以使用此类片段的最小数目的其它值,例如3、5、8、10个和大于10个。

[0136] 图5是表500,其展示当假定样品中的肿瘤来源的DNA的百分比浓度是5%时鉴别的假阳性基因座的预期数目和突变的预期数目。在样品中的肿瘤来源的DNA的百分比浓度较低的情况下,将需要较高测序深度来实现检测癌症相关的突变的相同敏感性。还将需要更严格的准则来维持特异性。举例来说,将需要使用在至少5个DNA片段中出现特定突变的准则,而不是在10%肿瘤DNA浓度的情况下在样品中出现至少4次的准则。表400和500提供在给定倍数测序覆盖度和肿瘤DNA浓度情况下使用截止值的指导,其可以如本文中所描述来

假定或测量。

[0137] 使用检测到单核苷酸变化多于一次来界定突变的准则的另一种优势是,预期此可将由于非恶性组织中的单核苷酸变化的假阳性检测减到最少。因为核苷酸变化可以在正常细胞的有丝分裂期间出现,所以体内的每个健康细胞可以具有多个单核苷酸变化。这些变化可能会产生假阳性结果。然而,当细胞死亡时,细胞的变化将存在于血浆/血清中。虽然预期不同正常细胞会携带不同突变集,但一种细胞中出现的突变不大可能存在于血浆/血清中的众多拷贝中。这与肿瘤细胞内的突变形成对比,其中预期多个拷贝可见于血浆/血清中,因为肿瘤生长实质上是克隆的。因此,来自克隆的多个细胞将死亡并且释放代表克隆的标志性突变。

[0138] 在一个实施方案中,可以在测序之前,执行特定基因组区域的标靶富集。此标靶富集步骤可以在执行相同总量的测序的情况下增加所关注的区域的测序深度。在另一个实施方案中,可以首先在第一回合执行以相对低测序深度的测序。然后可以使展示至少一个单核苷酸变化的区域富集,以进行第二回合的测序,其具有较高倍数覆盖。然后,可以针对具有标靶富集的测序结果,应用多次出现的准则来界定突变。

[0139] VII. 动态截止

[0140] 如上文所述,支持变异体(潜在突变)的读取的数目的截止值N可以用以确定基因座是否够资格作为待计数的突变(例如SNM)。使用这种截止值可以减少假阳性。以下论述提供针对不同基因座选择截止的方法。在以下实施方案中,我们假定存在单一主要癌症克隆。可以对包括释放不同量的肿瘤DNA到血浆中的多个癌细胞克隆的情形进行类似分析。

[0141] A. 血浆中检测到的癌症相关的突变的数目

[0142] 血浆中可检测的癌症相关的突变的数目可以受以下多个参数影响,例如:(1)肿瘤组织中突变的数目(N_T),肿瘤组织中所存在的突变的总数目是患者血浆中可检测的肿瘤相关的突变的最大数目;(2)血浆中的肿瘤来源的DNA的百分比浓度(f),血浆中的肿瘤来源的DNA的百分比浓度越高,在血浆中检测到肿瘤相关的突变的机率越高;(3)测序深度(D),测序深度是指所测序的区域由序列读取覆盖的次数。举例来说,10倍的平均测序深度意指,所测序的区域内的每种核苷酸平均由10个序列读取覆盖。当测序深度增加时,检测到癌症相关的突变的机率将增加;和(4)在血浆中将核苷酸变化界定为潜在癌症相关的突变所需要的检测到该变化的最小次数(r),其是用以将测序错误与真实癌症相关的突变区别开的截止值。

[0143] 在一个实施方案中,泊松分布用以预测血浆中可以检测到的癌症相关的突变的数目。假定突变以测序深度 D 存在于两种同源染色体中的一者上的核苷酸位置中,突变存在于血浆中的预期次数(M_p)如下计算: $M_p = D \times f / 2$ 。

[0144] 在特定突变位点处在血浆中检测到突变的概率(P_b)如下计算:

$$[0145] \quad P_b = 1 - \sum_{i=0}^{r-1} \frac{e^{-M_p} M_p^i}{i!}$$

[0146] 其中 r (截止值)是为了将核苷酸变化界定为潜在肿瘤相关的突变所需要在血浆中可见改变化的次数;泊松(i, M_p)是具有 i 次出现、平均数目是 M_p 的泊松分布概率。

[0147] 预期在血浆中检测到的癌症相关的突变的总数目(N_p)可以如下计算: $N_p = N_T \times P_b$,

其中 N_T 是肿瘤组织中所存在的突变的数目。以下图表展示使用检测潜在突变的不同出现频数(r)准则和不同测序深度时,预期在血浆中检测到的肿瘤相关的突变的百分比。

[0148] 图6A是图表600,其展示当肿瘤来源的DNA的血浆分数浓度是10%和20%和使用出现频数(r)为四次和六次作为识别潜在癌症相关的突变的准则是,在血浆中的癌症相关的突变的检测率的变化。在 r 相同的情况下,血浆中的肿瘤来源的DNA的百分比浓度越高,将导致血浆中可检测到的癌症相关的突变的数目越高。在血浆中的肿瘤来源的DNA的百分比浓度相同的情况下, r 越高,将导致检测到的突变的数目越少。

[0149] B.由于错误导致检测到的假阳性的数目

[0150] 血浆DNA测序数据中的单核苷酸变化可以由于测序和比对错误而出现。可以基于数学上的二项分布模型预测具有假阳性单核苷酸变化的核苷酸位置的数目。影响假阳性位点的数目(N_{FP})的参数可以包括:(1)测序错误率(E),测序错误率定义为测序的核苷酸不正确的比例;(2)测序深度(D),测序深度越高,展示测序错误的核苷酸位置的数目将增加;(3)界定潜在癌症相关的突变所需要的相同核苷酸变化的最少出现数目(r);和(4)所关注的区域内的核苷酸位置的总数目(N_I)。

[0151] 突变的出现一般来说可以视为随机过程。因此,随着界定潜在突变的出现的准则增加,假阳性核苷酸位置的数目将随 r 指数性地减少。在一些现有测序平台中,某些序列上下文更易于具有测序错误。所述测序上下文的实例包括GGC基序、均聚物(例如AAAAAAA)和简单重复(例如ATATATATAT)。这些序列上下文将实质上增加单核苷酸变化或插入/缺失假象(中村K(Nakamura K)等人核酸研究(Nucleic Acids Res)2011;39,e90和米诺切AE等人基因组生物学2011;12,R112)。另外,重复序列(例如均聚物和简单重复)将计算上引入比对中的不明确性,并且因此引起单核苷酸变异的假阳性结果。

[0152] 所关注的区域越大,将观察到的假阳性核苷酸位置的数目越高。如果在全基因组中查找突变,那么所关注的区域将是全基因组并且涉及的核苷酸的数目将是30亿。另一个方面,如果集中于外显子,那么编码外显子的核苷酸的数目(即约4500万)将构成所关注的区域。

[0153] 可以基于以下计算测定与测序错误相关的假阳性核苷酸位置的数目。由于测序错误而导致在相同位置处具有相同核苷酸变化的概率(P_{Er})可以如下计算:

[0154] 其中 $C(D, r)$ 是由总共 D 个元素选择 r 个元素的可能的组合数目; r 是界定潜在突变所需要的出现数目; D 是测序深度;并且 E 是测序错误率。 $C(D, r)$ 可以如下计算:

[0155] 对于突变假阳性的核苷酸位置的数目(N_{FP})可以如下计算:

$$[0156] \quad N_{FP} = N_I P_{Er}$$

[0157] 其中 N_I 是所关注的区域中的核苷酸位置的总数目。

[0158] 图6B是图表650,其展示在使用4、5、6和7的出现频数(r)的准则来界定核苷酸变化时,其错误分类的预期数目相对于测序深度的变化情况。在此计算中假定所关注的区域是全基因组(30亿个核苷酸位置)。假定测序错误率是0.3%的测序的核苷酸。如可以看到的, r 的值对假阳性具有显著影响。但,由图6A可以看出,较高 r 值还减少检测的突变的数目,至少直到使用显著较高测序深度。

[0159] C.选择最少出现(r)

[0160] 如上文所论述,真实癌症相关的突变位点和归因于测序错误的假阳性位点的数目

将随测序深度而增加。然而,其增加速率将不同。因此,有可能利用测序深度和r值的选择来实现真实癌症相关的突变的检测达到最大化同时保持假阳性位点的数目处于低水平。

[0161] 图7A是图表700,其展示不同测序深度下真实癌症相关的突变位点和假阳性位点的数目。假定肿瘤组织中的癌症相关的突变的总数目是3,000,并且假定血浆中的肿瘤来源的DNA的百分比浓度是10%。假定测序错误率是0.3%。在图例中,TP表示相应突变存在于肿瘤组织中的真实阳性位点,并且FP表示相应突变不存在于肿瘤组织中并且测序数据中所存在的假阳性位点,其核苷酸变化归因于测序错误。

[0162] 根据图表700,如果我们使用出现频数最少6次($r=6$)作为准则来界定血浆中的潜在突变位点,那么在110倍的测序深度下,将检测到约1,410个真实癌症相关的突变。使用此准则,将检测到仅约20个假阳性位点。如果我们使用出现频数最少7次($r=7$)作为准则来界定潜在突变,那么可以检测到的癌症相关的突变的数目将减少470,即约为940。因此, $r=6$ 的准则将使得血浆中的癌症相关的突变的检测更敏感。

[0163] 另一个方面,如果我们分别使用出现频数(r)最少6和7次的准则来界定潜在突变,那么在200倍的测序深度下,所检测的真实癌症相关的突变的数目将是约2,800和2,600。使用这两个 r 值,假阳性位点的数目将分别是约740和20。因此,在200倍的测序深度下,使用 $r=7$ 的更严格准则来界定潜在突变可以大大减少假阳性位点的数目而不会显著不利地影响检测真实癌症相关的突变的敏感性。

[0164] D. 界定血浆中的潜在突变的测序数据的动态计算机可读介质

[0165] 所关注的区域内的每种核苷酸的测序深度将不同。如果我们应用核苷酸变化出现的固定截止值来界定血浆中的潜在突变,那么与具有较低测序深度的核苷酸相比,由较多序列读取覆盖(即较高测序深度)的核苷酸将具有较高的概率将在肿瘤组织中不存在这种变化错误地记为具有该核苷酸变异。解决此问题的一个实施方案是根据特定核苷酸位置的实际测序深度和根据识别假阳性变异的概率的所要球求的上限,将 r 对应的动态截止值应用于不同核苷酸位置。

[0166] 在一个实施方案中,最大容许假阳性率可以固定于1比 1.5×10^8 个核苷酸位置。在此最大容许假阳性率下,全基因组中鉴别的假阳性位点的总数目将小于20。针对不同测序深度的 r 值可以根据图6B中所示的曲线来测定并且这些截止值展示于表1中。在其它实施方案中,可以使用其它不同最大容许假阳性率,例如1比 3×10^8 、1比 10^8 或1比 6×10^7 。假阳性位点的相应总数目将分别小于10、30和50。

[0167]	特定核苷酸位置的测序深度	用以界定潜在突变所需要在血浆 DNA 测序数据中所存在的核苷酸变化的最少出现数目 (r)
	<50	5
	50 - 110	6
	111 - 200	7
	201 - 310	8
	311 - 450	9
	451 - 620	10
	621 - 800	11

[0168] 表1. 相对于特定核苷酸位置的不同测序深度,用以界定潜在突变所需要在血浆中

所存在的核苷酸变化的最少出现数目(r)。最大假阳性率固定于1比 1.5×10^8 个核苷酸。

[0169] E. 标靶富集测序

[0170] 如图7A中所示,较高测序深度可以导致检测癌症相关的突变的敏感性较佳同时通过允许使用较高r值保持较低的假阳性位点数目。举例来说,在110倍的测序深度下,使用r值6可以在血浆中检测到1,410个真实癌症相关的突变,而当测序深度增加到200倍并且应用r值7时,所检测的真实癌症相关的突变的数目将是2,600。两个数据集预期将得到约20个假阳性位点。

[0171] 虽然目前以200倍的深度对全基因组测序相对昂贵,但实现这种测序深度的一种可能的方式将是集中于较小的所关注的区域。可以例如通过但不限于使用DNA或RNA诱饵通过杂交来捕获所关注的基因组区域,从而实现标靶区域的分析。然后将所捕获的区域例如通过磁性手段拉下,并且然后使其进行测序。可以例如使用安捷伦(Agilent) SureSelect标靶富集系统、罗氏Nimblegen标靶富集系统和伊路米那(Illumina)标靶重测序系统执行所述标靶捕获。另一种方法是对标靶区域执行PCR扩增并且然后执行测序。在一个实施方案中,所关注的区域是外显子组。在这种实施方案中,可以对血浆DNA执行所有外显子的标靶捕获,并且然后可以对外显子区域富集的血浆DNA测序。

[0172] 除了具有较高测序深度之外,集中于特定区域而不是分析全基因组将显著减少检索空间中的核苷酸位置的数目,并且将导致给定相同测序错误率下假阳性位点的数目减少。

[0173] 图7B是图表750,其展示假阳性位点的预测数目,包括分析全基因组(WG)和所有外显子。对于每种类型的分析,使用两个不同r值,5和6。在200倍的测序深度下,如果将 $r=5$ 用以界定血浆中的突变,那么对于全基因组和所有外显子,假阳性位点的预测数目分别是约23,000和230。如果将 $r=6$ 用以界定血浆中的突变,那么假阳性位点的预测数目分别是750和7。因此,所关注的区域中的核苷酸的数目的限制可以显著减少血浆突变分析中的假阳性的数目。

[0174] 在外显子捕获或甚至外显子组捕获测序中,搜索空间中的核苷酸的数目减少。因此,即使我们允许癌症相关的突变的检测有较高假阳性率,假阳性位点的绝对数也可以保持为相对低等级。允许较高假阳性率将允许待使用的界定血浆中的单核苷酸变异的最少出现(r)的准则不太严格。这将导致检测真实癌症相关的突变的敏感性较高。

[0175] 在一个实施方案中,我们可以使用 1.5×10^6 的最大容许假阳性率。在此假阳性率下,标靶外显子内的假阳性位点的总数目将仅是20。使用 1.5×10^6 的最大容许假阳性率针对不同测序深度的r值展示于表2中。在其它实施方案中,可以使用其它不同最大容许假阳性率,例如1比 3×10^6 、1比 10^6 或1比 6×10^5 。假阳性位点的相应总数目将分别小于10、30和50。在一个实施方案中,如上文所述,不同类别的突变可以带来不同权重。

[0176]	特定核苷酸位置的测序深度	用以界定潜在突变所需要在血浆DNA 测序数据中所存在的核苷酸变化的最少出现数目 (r)
	<50	4
	50 - 125	5
	126 - 235	6
	236 - 380	7
	381 - 560	8
	561 - 760	9

[0177] 表2. 相对于特定核苷酸位置的不同测序深度, 用以界定潜在突变所需要在血浆中所存在的核苷酸变化的最少出现数目 (r)。最大假阳性率固定于1比 1.5×10^6 个核苷酸。

[0178] VIII. 癌症检测

[0179] 如上文所提及, 可以按各种方式使用变异基因座处的序列标签的计数来测定参数, 将所述参数与阈值比较以对癌症等级分类。一个基因座或许多基因座处的变异体读取相对于所有读取的百分比浓度是可以使用的另一个参数。以下是计算参数和阈值的一些实例。

[0180] A. 测定参数

[0181] 如果CG在特定基因座处对于第一等位基因是纯合的并且变异等位基因可见于生物样品 (例如血浆) 中, 那么百分比浓度可以如下计算: $2p/(p+q)$, 其中p是具有变异等位基因的序列标签的数目, 并且q是具有CG的第一等位基因的序列标签的数目。此式假定肿瘤的单倍型中的仅一者具有变异体, 典型地将如此。因此, 对于每个纯合基因座, 可以计算百分比浓度。可以将百分比浓度平均。在另一个实施方案中, 计数p可以包括所有基因座的序列标签的数目, 并且对于计数q类似地, 以测定百分比浓度。现描述一个实例。

[0182] 以全基因组水平研究和检测4名HCC患者的血浆中的肿瘤来源的单核苷酸变异体 (SNV)。我们分别以单倍型基因组平均覆盖深度分别为29.5倍 (范围, 27倍到33倍) 和43倍 (范围, 39倍到46倍) 对肿瘤DNA和白血球DNA测序。比较来自4名HCC患者中的每一者的肿瘤DNA与白血球DNA的MPS数据, 并且以严格生物信息学算法寻求存在于肿瘤DNA中但不存在于白血球DNA中的SNV。此算法需要使候选的SNV对应的测序得到的肿瘤DNA片段 (即在相应序列标签) 至少超过某一阈值数目, 随后才将其分类为真实SNV。例如如本文中所描述, 通过考虑特定核苷酸的测序深度和测序错误率测定阈值数目。

[0183] 图8是表800, 其展示根据本发明的实施方案4名HCC患者在治疗前后的结果, 包括血浆中的肿瘤来源的DNA的百分比浓度。在4个HCC病例中, 肿瘤相关的SNV的数目在1,334到3,171范围内。在治疗前后列出血浆中可检测的所述SNV的比例。在治疗之前, 在血浆中检测出15%-94%的肿瘤相关的SNV。在治疗之后, 检测出的百分比在1.5%-5.5%之间。因此, 检测的SNV的数目的确与癌症等级相关。这显示, SNV的数目可以用作参数来对癌症等级分类。

[0184] 血浆中的肿瘤来源的DNA的百分比浓度通过突变体相对于总 (即突变体加野生型) 序列的分数计数来测定。公式是 $2p/(p+q)$, 其中2考虑了肿瘤上突变的仅发生在一种单倍型中。这些百分比浓度充分与由全基因组聚集等位基因丢失 (GAAL) 分析测定的那些相关 (陈 KC (Chan KC) 等人临床化学2013;59:211-24) 并且在手术之后降低。因此, 表明百分比浓度也是可用于测定癌症等级的参数。

[0185] 来自SNV分析的百分比浓度可以反应肿瘤负荷。具有较高肿瘤负荷(例如较高推断得到的百分比浓度)的癌症患者将具有比具有较低肿瘤负荷的癌症患者更高的体细胞突变频率。因此,实施方案还可以用于预后。一般来说,具有较高肿瘤负荷的癌症患者具有比具有较低肿瘤负荷的癌症患者更差的预后。前一组因此将具有更高的因所述疾病而死亡的机率。在一些实施方案中,如果可以测定(例如使用实时PCR或荧光测定法)生物样品(例如血浆)中的DNA的绝对浓度,那么可以测定肿瘤相关的遗传畸变的绝对浓度并且将其用于临床检测和/或监测和/或预后。

[0186] B. 测定阈值

[0187] 表800可以用以测定阈值。如上文所提及,通过SNV分析测定的SNV数目和百分比浓度与癌症等级相关。阈值可以根据个人为基础测定。举例来说,治疗前的参数值可以用以测定阈值。在各种实施方案中,阈值可以是相对于治疗前绝对值的相对变化。合适的阈值可以是SNV数目或百分比浓度的减少50%对应的数值。这种阈值将向表800中的每一病例提供较低癌症等级的分类。注意,所述阈值可以取决于测序深度。

[0188] 在一个实施方案中,阈值可以跨越样品而使用,并且可以或可以不顾及参数的治疗前值。举例来说,100个SNV的阈值可以用以将受试者分类为不患有癌症或具有低癌症等级。在表800中,四个病例中的每一者都满足此100个SNV的阈值。如果将百分比浓度用作参数,那么1.0%的阈值将HCC1-HCC3分类为几乎零等级癌症,并且1.5%的第二阈值将HCC4分类为低癌症等级。因此,多于一个阈值可以用以获得多于两个分类。

[0189] 为说明其它可能的阈值,我们分析健康对照的血浆的肿瘤相关的SNV。可以对健康受试者进行众多测量以测定预期生物样品相对于组成型基因组的变异多少的范围。

[0190] 图9是表900,其展示根据本发明的实施方案检测16名健康对照受试者中的HCC相关的SNV。表900可以用以估计SNV分析方法的特异性。16名健康对照列于不同行中。每一列是针对特定HCC患者检测的SNV,并且展示在变异基因座处具有变异等位基因的序列读取的数目和具有野生型等位基因(即来自CG的等位基因)的序列读取的数目。举例来说,对于HCC1,对照C01在所述变异基因座处具有40个变异体读取,但具有31,261个野生型等位基因读取。最后一列展示跨越HCC患者的所有SNV的总百分比浓度。因为HCC相关的SNV特异于HCC患者,所以HCC相关的SNV的存在代表假阳性。如本文中所描述,如果截止值应用于这些表观序列变异体,那么所有这些假阳性将被过滤掉。

[0191] 16名健康对照的血浆中的这些微小数量的待定的肿瘤相关突变代表了此方法的对应“随机噪声”并且可能是由于测序错误所导致。由所述噪声估计的平均百分比浓度是0.38%。这些值显示了健康受试者的范围。因此,HCC的零癌症等级的分类的阈值可以是约0.5%,因为最高百分比浓度是0.43%。因此,如果由HCC患者去除所有癌细胞,那么将预期其会表现出类似于这样的低百分比浓度。

[0192] 返回参看表800,如果将0.5%用作零癌症等级的阈值,那么基于SNV分析,HCC1和HCC3的治疗后血浆数据将测定为具有零等级。HCC2可以分类为从零向上的一个等级。HCC4也可以分类为从零向上的一个等级,或某一更高等级,但与治疗前样品相比仍是相对低等级。

[0193] 在其中参数对应于变异基因座数目的一个实施方案中,阈值可以是零(即一个变异基因座可以指示非零癌症等级)。然而,在许多设定(例如测序深度)下,阈值将更高,例如

绝对值是5或10。在其中在治疗之后监测一个人的一个实施方案中,阈值可以是样品中出现的SNV (通过直接分析肿瘤而鉴别) 的一定百分比。如果在基因座处需要的变异体读取的数目的截止值足够大,那么仅具有一个变异基因座就可以指示非零癌症等级。

[0194] 因此,生物样品(例如血浆)的DNA中的变异(例如单核苷酸变异)的定量分析可以用于诊断、监测和预测癌症。为了检测癌症,可以将测试受试者的血浆中检测到的单核苷酸变异数目与一组健康受试者的单核苷酸变异数目比较。在健康受试者中,血浆中的表观单核苷酸变异可以是源自于测序错误,来自于血细胞和其它器官的非克隆突变。已经显示,正常健康受试者中的细胞可以携带小数目的突变(康拉德DF (Conrad DF) 等人自然遗传 (Nat Genet) 2011;43:712-4), 如表900中所示。因此,一组表观健康受试者的血浆中的表观单核苷酸变异的总数目可以用作参考范围来确定在所测试的患者血浆中具有异常高数目的单核苷酸变异是否对应于非零癌症等级。

[0195] 可以使用于确定参考范围的健康受试者与所测试的受试者在年龄和性别方面匹配。在先前研究中,已经显示,体细胞中的突变的数目将随着年龄而增加(张NK (Cheung NK) 等人,美国医学会杂志 (JAMA) 2012;307:1062-71)。因此,随着我们变得年老,则积累细胞克隆将是‘正常的’,纵使所述细胞克隆大部分时间是相对良性的,或将花费极长时间来变得临床上有影响。在一个实施方案中,可以针对不同受试者群组,例如不同年龄、性别、种族和其它参数(例如吸烟状态、肝炎状态、酒精、药物史),产生参考水平。

[0196] 参考范围可以基于所用截止值(即在基因座处需要的变异序列标签的数目)以及假设的假阳性率和其它变量(例如年龄)而变化。因此,可以关于一或多个准则来决定来测定参考范围,并且将相同准则用以测定样品的参数。然后,可以将参数与参考范围比较,因为两者是使用相同准则测定的。

[0197] 如上文所提及,实施方案可以使用多个阈值来测定癌症等级。举例来说,第一等级可以确定低于阈值的参数为无癌症的征象;和至少第一癌症等级,其可以是肿瘤前的等级。其它等级可以对应于不同癌症阶段。

[0198] C. 对于实验变量的依赖性

[0199] 测序深度对于确定少数(例如肿瘤)基因组的最低检测阈值可以是重要的。举例来说,如果使用单倍型基因组的测序深度为10,那么即使用无任何错误的测序技术可以检测的最低肿瘤DNA浓度是1/5,即20%。另一个方面,如果使用单倍型基因组的测序深度为100,那么最低检出浓度可以下降到2%。此分析涉及仅分析一个突变基因座的情形。然而,当分析更多个突变基因座时,最低肿瘤DNA浓度可以更低并且由二项概率函数决定。举例来说,如果测序深度是10倍并且肿瘤DNA的百分比浓度是20%,那么检测到突变的机率是10%。然而,如果我们具有10个突变,那么检测到至少一个突变的机率将是 $1 - (1 - 10\%)^{10} = 65\%$ 。

[0200] 增加测序深度存在数种效应。测序深度越高,将可遇见越多测序错误,参看图4和5。然而,在较高测序深度下,可以将更容易区分测序错误和细胞(例如癌细胞)亚群的克隆扩增的突变,因为测序错误是随机出现于基因组中,但既定细胞群体对应的突变是出现在相同位置处。

[0201] 测序深度越高,将由“健康细胞”鉴别出越多突变。然而,当不存在这些健康细胞的克隆扩充并且其突变分布不同时,则这些健康细胞中的突变可以通过其在血浆中的出现频率与癌症相关的突变区分开来(例如对于展现突变的读取的所需数目使用截止值N,例如N

等于2、3、4、5或更大)。

[0202] 如上文所提及,阈值可以取决于健康细胞中的突变的量,所述健康细胞将克隆地扩增,并且因此通过其它机制可能不被过滤掉。预计此差异可以通过分析健康受试者而获得。随着克隆扩增随时间而发生,患者的年龄可以影响健康受试者中存在的变异,并且因此阈值对于年龄可能有一定依赖。

[0203] D. 与标靶方法的组合

[0204] 在一些实施方案中,随机测序可以与标靶方法组合使用。举例来说,可以在患者癌症发生后对血浆样品执行随机测序。可以针对拷贝数畸变和SNV分析血浆DNA的测序数据。可以靶向展示畸变(例如扩增/缺失或高密度的SNV)的区域以用于连续监测目的。监测可以按单一程序形式有效地,在一段时间内进行,或在随机测序之后立即进行。为了标靶分析,基于液相杂交的捕获方法已经成功地用以使血浆DNA富集以用于非侵入性产前诊断(廖GJ (Liao GJ) 等人临床化学2011;57:92-101)。所述技术如上文所提及。因此,标靶和随机方法可以组合使用以用于癌症检测和监测。

[0205] 因此,可以对上文所提及的非标靶全基因组方法发现的可能突变的基因座执行标靶测序。所述标靶测序可以使用液相或固相杂交技术(例如使用安捷伦SureSelect、NimbleGen序列捕获或伊路米那标靶重测序系统)接着是大规模平行测序来执行。另一种方法是基于扩增(例如基于PCR的)系统来实现标靶测序(弗休T (Forshew T) 等人科学·转化医学2012;4:135ra68)。

[0206] IX. 百分比浓度

[0207] 肿瘤DNA的百分比浓度可以用以决定基因座处所需的变异数目的截止值,随后再鉴定基因座上的突变。举例来说,如果已知百分比浓度相对高,那么高截止值可以用以过滤掉更多假阳性,因为知晓对于真实SNV而言,应存在相对高数目的变异体读取。另一个方面,如果百分比浓度低,那么可能需要较低截止值以便不会遗漏一些SNV。在此情况下,百分比浓度将通过与SNV分析不同的方法来测定,以其用作参数。

[0208] 各种技术可以用于测定百分比浓度,其中的一些描述于本文中。这些技术可以用以测定混合物中的肿瘤来源的DNA的百分比浓度,所述混合物例如含有肿瘤细胞和非恶性细胞的混合物的活检样品或来自癌症患者的含有由肿瘤细胞释放的DNA和由非恶性细胞释放的DNA的血浆样品。

[0209] A. GAAL

[0210] 全基因组聚集等位基因丢失(GAAL)分析关注于那些已经丧失了杂合性的基因座(陈KC (Chan KC) 等人临床化学2013;59:211-24)。对于在组成型基因组CG中的杂合性位点,肿瘤通常涉及等位基因中的一者缺失的基因座。因此,这种基因座的序列读取将展示比另一者多于一个的等位基因,其中差异与样品中的肿瘤DNA的百分比浓度成正比。这种计算的一个实例如下。

[0211] 用昂飞(Affymetrix)全基因组人SNP阵列6.0系统对由HCC患者的白血球和肿瘤组织提取的DNA基因分型。用昂飞基因分型控制台4.1版处理微阵列数据。用Birdseedv2算法执行基因分型分析和单核苷酸多态性(SNP)呼叫。将白血球和肿瘤组织的基因分型数据用于鉴别杂合性丢失(LOH)区域并且用于执行拷贝数分析。用来自昂飞的基因分型控制台以默认参数、并且以最小基因组片段为100bp和片段内最少遗传标记为5来执行拷贝数分析。

[0212] 具有LOH的区域鉴别为在肿瘤组织中具有1个拷贝并且在白血球细胞中具有2个拷贝的区域,这些区域内的SNP在白血球细胞中是杂合的但在肿瘤组织中是纯合。对于肿瘤组织中展现LOH的基因组区域,存在于白血球细胞中但肿瘤组织中不存在或强度降低的SNP等位基因被视为含有片段缺失的染色体区域上的等位基因。存在于白血球和肿瘤组织两者中的等位基因被认为已经是来源于染色体区域的非缺失片段。对于肿瘤中的具有单一拷贝损失的所有染色体区域,计算对携带缺失等位基因和非缺失等位基因的序列读取的总数目。这两个值的差值用以使用以下方程式来推断样品中的肿瘤来源的DNA的百分比浓度(F_{GAAL}):

$$[0213] \quad F_{GAAL} = \frac{N_{non-del} - N_{del}}{N_{non-del}}$$

[0214] 其中 $N_{non-del}$ 代表携带非缺失等位基因的序列读取的总数目,并且 N_{del} 代表携带缺失等位基因的序列读取的总数目。

[0215] B. 基于基因组呈现量的评估

[0216] GAAL技术的一个问题是,鉴别出特定基因座(即展现LOH的基因座),并且仅使用与所述基因座的序列读取比对。所述需求可能会增加额外步骤并且因此增加成本。现描述一个实施方案,其仅使用拷贝数,例如序列读取密度。

[0217] 在癌症基因组中经常观察到染色体畸变,例如扩增和缺失。癌症组织中观察到的染色体畸变通常涉及亚染色体区域,并且这些畸变可以短于1Mb。并且,癌症相关的染色体畸变在不同患者中是异质的,并且因此在不同患者中不同区域可以受影响。数十、数百或甚至数千拷贝数的畸变见于癌症基因组中也并非不寻常的。所有这些因素使得难以测定肿瘤DNA浓度。

[0218] 实施方案包括分析由肿瘤相关的染色体畸变产生的量变。在一个实施方案中,使用大规模平行测序、例如通过伊路米那HiSeq2000测序平台对含有来源于癌细胞和正常细胞的DNA的DNA样品测序。来源DNA可以是血浆或其它合适生物样品中的无细胞的DNA。

[0219] 肿瘤组织中扩增的染色体区域将具有增加的被测序的概率,并且肿瘤组织中缺失的区域将具有降低的被测序的概率。因此,比对到扩增区域的序列标签对应的读取密度将增加,并且比对到缺失区域的序列标签对应的读取密度将减少。变异的程度与DNA混合物中的肿瘤来源的DNA的百分比浓度成正比。来自肿瘤组织的DNA的比例越高,将由染色体畸变引起的变化越大。

[0220] 1. 具有高肿瘤浓度的样品中的评估

[0221] 由四名肝细胞癌患者的肿瘤组织提取DNA。将DNA使用科瓦里亚(Covaria)DNA超声处理系统片段化并且使用如所描述的伊路米那HiSeq2000平台测序进行测序(陈KC等人临床化学2013;59:211-24)。将序列读取与人参考基因组(hg18)比对。然后将基因组分成1Mb区间(区域),并且在如所描述方法调节GC引起的偏差之后计算每个区间的序列读取密度(陈EZ等人公共科学图书馆·综合(PLoS One).2011;6:e21791)。

[0222] 在将序列读取与参考基因组比对之后,可以计算各个区域的序列读取密度。在一个实施方案中,序列读取密度是定位到特定区间(例如1Mb区域)的读取的数目除以可以与参考基因组(例如与参考基因组中的独特位置)比对的总序列读取的比例。预期与肿瘤组织中扩增的染色体区域重叠的区间比不具有所述重叠的区间具有更高的序列读取密度。另一

个方面,预期与缺失的染色体区域重叠的区间比不具有所述重叠的区间具有更低的序列读取密度。具有与不具有染色体畸变的区域之间的序列读取密度的差值的量级主要受样品中的肿瘤来源的DNA的比例和肿瘤细胞中的扩增/缺失的程度影响。

[0223] 各种统计模型可以用以鉴别具有对应于不同类型的染色体畸变的序列读取密度的区间。在一个实施方案中,可以使用正态分布混合物模型(麦克拉伦G (McLachlan G) 和皮尔D (Peel D). 多变量正态分布混合物 (Multivariate normal mixtures). 有限分布混合物模型 (Finite mixture models) 2004: 第81-116页. 约翰·威利父子出版社 (John Wiley & Sons Press)). 还可以使用其它统计模型, 例如二项分布混合物模型和泊松回归模型(麦克拉伦G和皮尔D. 具有非正常组分的混合物, 有限混合物模型2004: 第135-174页. 约翰·威利父子出版社)。

[0224] 区间的序列读取密度可以使用在相同区间处由对白血球DNA测序所测定的序列读取密度来标准化。不同区间的序列读取密度可以受特定染色体区域的序列所处的上下文环境所影响, 并且因此标准化可以帮助更准确地鉴别展示畸变的区域。举例来说, 不同染色体区域的可比对性(其是指将序列比对回其原始位置的概率)可以不同。另外, 拷贝数的多态性(即拷贝数变异)也将影响区间的序列读取密度。因此, 用白血球DNA标准化可能可以使与不同染色体区域之间的序列上下文的差异相关的变异减到最少。

[0225] 图10A展示了根据本发明的实施方案HCC患者的肿瘤样品的序列读取密度的分布图1000。肿瘤组织在由HCC患者手术切除之后获得。x轴代表患者的肿瘤组织与白血球组织之间的序列读取密度的比率(R)的以二为底的对数值。y轴代表对应区间的数目。

[0226] 可以使用正态分布混合物模型将峰值拟合到分布曲线以代表具有缺失、扩增和不具有染色体畸变的区域。在一个实施方案中, 峰值的数目可以通过覆盖不同可能值所对应的赤池信息准则 (Akaike's information criterion, AIC) 来测定。 $\log_2 R = 0$ (即 $R = 1$) 的中心峰代表不具有任何染色体畸变的区域。左峰(相对于中心峰)代表具有一拷贝损失的区域。右峰(相对于中心峰)代表具有一拷贝扩增的区域。

[0227] 肿瘤来源的DNA的百分比浓度可以由代表扩增与缺失区域的峰值之间的距离来反映。距离越大, 样品中的肿瘤来源的DNA的百分比浓度将越高。样品中的肿瘤来源的DNA的百分比浓度可以使用以下方程式通过此基因组呈现量的方法来测定, 表示为 $F_{GR} : F_{GR} = R_{right} - R_{left}$, 其中 R_{right} 是右峰的R值并且 R_{left} 是左峰的R值。最大差值将是1, 对应于100%。估计获自HCC1患者的肿瘤样品中的肿瘤来源的DNA的百分比浓度是66%, 其中 R_{right} 和 R_{left} 的值分别是1.376和0.712。

[0228] 为证实此结果, 独立地使用全基因组聚集等位基因损失 (GAAL) 分析的方法来测定肿瘤DNA的百分比浓度(陈KC等人临床化学2013; 59: 211-24)。表3展示使用基因组呈现量 (F_{GR}) 和GAAL (F_{GAAL}) 方法得到的, 四名HCC患者的肿瘤组织中的肿瘤来源的DNA的百分比浓度。通过这两种不同方法测定的值彼此充分一致。

[0229]

HCC肿瘤	F_{GAAL}	F_{GR}
1	60.0%	66.5%
2	60.0%	61.4%
3	58.0%	58.9%
4	45.7%	42.2%

[0230] 表3展示通过GAAL和基因组呈现量 (GR) 测定的分数浓度。

[0231] 2. 具有低肿瘤浓度的样品中的评估

[0232] 以上分析已经显示, 我们的基因组呈现量的方法可以用在肿瘤来源DNA多于50%的样品DNA中, 即在肿瘤DNA是大多数比例时的情况下测量肿瘤DNA的百分比浓度。在先前分析中, 我们已经展示, 此方法还可以应用于其中肿瘤来源的DNA仅有微量比例 (即低于50%) 的样品。可以含有微量比例的肿瘤DNA的样品包括 (但不限于) 癌症患者的血液、血浆、血清、尿液、胸膜液、脑脊髓液、泪液、唾液、腹水液和粪便。在一些样品中, 肿瘤来源的DNA的百分比浓度可以是49%、40%、30%、20%、10%、5%、2%、1%、0.5%、0.1%或更低。

[0233] 对于所述样品, 代表具有扩增和缺失的区域的序列读取密度的峰值可能不比如以上所说明在含有相对高浓度的肿瘤来源的DNA的样品中明显。在一个实施方案中, 癌细胞中的具有染色体畸变的区域可以通过与已知不含有癌症DNA的参考样品进行比较而鉴别出。举例来说, 不患有癌症的受试者的血浆可以用作参考来测定染色体区域的序列读取密度的正常范围。可以将所测试的受试者的序列读取密度与参考组别对应的值比较。在一个实施方案中, 可以测定序列读取密度的平均值和标准偏差 (SD)。对于每个区间, 使用下式将所测试的受试者的序列读取密度与参考组别的平均值比较以测定z分数:

$$[0234] \quad z \text{ 分数} = \frac{(GR_{\text{test}} - \overline{GR}_{\text{ref}})}{SD_{\text{ref}}}, \text{ 其中 } GR_{\text{test}} \text{ 代表癌症患者的序列读取密度; } \overline{GR}_{\text{ref}} \text{ 代表参}$$

考受试者的平均序列读取密度, 并且 SD_{ref} 代表参考受试者的序列读取密度的SD。

[0235] z分数<-3的区域表示癌症患者中的特定区间的序列读取密度的显著呈现不足, 表明肿瘤组织中存在缺失。z分数>3的区域表示癌症患者中的特定区间的序列读取密度的显著呈现过度, 表明肿瘤组织中存在扩增。

[0236] 然后, 可以对所有区间的z分数的分布进行构筑以鉴别具有不同数目的拷贝增加和损失的区域, 例如缺失1或2个拷贝的染色体; 和扩增, 产生1、2、3和4个额外拷贝的染色体。在一些情况下, 可以涉及多于一个染色体或多于一个染色体区域。

[0237] 图10B展示了根据本发明的实施方案HCC患者的血浆中的所有区间的z分数的分布图1050。将代表1拷贝损失、无拷贝变化、1拷贝增加和2拷贝增加的峰值 (从左到右) 拟合到z分数分布。然后可以例如使用如上文所述的正态分布混合物模型鉴别出具有不同类型的染色体畸变的区域。

[0238] 然后可以由展现一拷贝增加或一拷贝损失的区域的序列读取密度推断样品中的癌症DNA的百分比浓度 (F)。关于特定位元子测定的百分比浓度可以如下计算:

$$F = \frac{|GR_{\text{test}} - \overline{GR}_{\text{ref}}| \times 2}{\overline{GR}_{\text{ref}}} \times 100\%。 \text{ 此还可以表示为: } F = \frac{|z \text{ 分数} \times SD_{\text{ref}}|}{\overline{GR}_{\text{ref}}} \times 2, \text{ 其可以重写为: } F =$$

$|z \text{ 分数}| \times CV \times 2$, 其中CV是测量参考受试者的序列读取密度的变异系数; 并且

$$CV = \frac{SD_{\text{ref}}}{\overline{GR}_{\text{ref}}}。$$

[0239] 在一个实施方案中, 将各区间的结果组合。举例来说, 可以将展示1拷贝增加的位元子的z分数平均或将所得F值平均。在另一个实施方案中, 用于推断F的z分数的值通过统计模型测定并且由图10B和图11中所示的峰值表示。举例来说, 右峰的z分数可以用以测定

展现1拷贝增加的区域的分数的浓度。

[0240] 在另一个实施方案中，z分数<-3和z分数>3的所有区间都可以分别归于具有单一拷贝损失和单一拷贝增加的区域，因为这两种类型的染色体畸变是最常见的。此近似在具有染色体畸变的区间的数目相对小并且正态分布的拟合可能不准确时最适用。

[0241] 图11展示了根据本发明的实施方案HCC患者的血浆的z分数的分布图1100。虽然与染色体畸变重叠的区间的数目相对小，但将z分数<-3和z分数>3的所有区间分别拟合到单一拷贝损失和单一拷贝增加的正态分布。

[0242] 使用GAAL分析和此基于GR的方法测定四名HCC患者的血浆中的肿瘤来源的DNA的百分比浓度。结果展示于表4中。可以看出，推断分数代表在GAAL分析与GR分析之间充分相关。

样品	血浆中的肿瘤来源的 DNA 的分数浓度	
	GAAL 分析	GR 分析
病例 11	4.3%	4.5%
病例 13	5%	5.5%
病例 23	52%	62%
病例 27	7.6%	6.1%

[0244] 表4. 通过染色体畸变的分析推断的血浆中的肿瘤来源的DNA的百分比浓度。

[0245] C. 测定百分比浓度的方法

[0246] 图12是根据本发明的实施方案测定包括无细胞的DNA的生物样品中的肿瘤DNA的百分比浓度的方法1200的流程图。方法1200可以经由各种实施方案，包括上文所描述的实施方案执行。

[0247] 在框1210处，对于生物样品中的多个DNA片段中的每一者，获得一个或多个序列标签。框1210可以如本文中关于其它方法所描述来执行。举例来说，可以由血浆样品对DNA片段的一个末端测序。在另一个实施方案中，可以对DNA片段的两个末端测序，由此使得可估计片段的长度。

[0248] 在框1220处，测定基因组位置的序列标签。可以例如如本文中所描述通过将序列标签与参考基因组比对来测定基因组位置。如果对片段的两个末端测序，那么可以以成对的方式将配对的标签分别比对到参考基因组，并且要求在两个标签之间的距离小于规定距离（例如500或1,000个碱基）。

[0249] 在框1230处，对于多个基因组区域中的每一者，由在基因组区域内具有基因组位置的序列标签测定各自对应的基因组区域内DNA片段的量。基因组区域可以是参考基因组中的相等长度的非重叠区域。在一个实施方案中，可以对与区间比对的多个标签计数。因此，每个区间可以具有相应的比对标签的数目。可以计算直方图，说明区间具有一定数目的比对标签的频率。可以对各自具有相同长度（例如1Mb区间）的基因组区域执行方法1200，其中所述区域是非重叠的。在其它实施方案中，可以考虑使用不同的长度并且区域可能重叠。

[0250] 在框1240处，标准化各自的量以获得各自的密度。在一个实施方案中，标准化各自的量以获得各自的密度包括使用与比对得到的参考标签总数相同的数目以测定各自的密度和参考密度。在另一个实施方案中，各自的量可以除以比对得到的参考标签的总数目。

[0251] 在框1250处，将各自的密度与参考密度比较以鉴别基因组区域展现出1拷贝损失

还是1拷贝增加。在一个实施方案中,计算各自的密度与参考密度之间的差值(例如作为测定z分数的一部分),并且将其与截止值比较。在各种实施方案中,参考密度可以获自健康细胞的样品(例如来自白血球)或获自各自量的本身(例如通过在大多数区域不展现损失或增加的假设下,获取中值或平均值)。

[0252] 在框1260处,由鉴别为展现1拷贝损失的一个或多个各自的密度或由鉴别为展现1拷贝增加的一或多个各自的密度计算第一密度。第一密度可以对应于仅一个基因组区域,或可以由多个基因组区域的密度测定。举例来说,可以第一密度由具有1拷贝损失的各自密度计算。各自密度提供了给定肿瘤浓度下由肿瘤中的缺失区域产生的密度差值的度量。类似地,如果第一密度来自具有1拷贝增加的各自密度,那么可以获得由肿瘤中的扩增区域产生的密度差值的度量。以上部分描述了关于如何用多个区域的密度来测定平均密度来确定第一密度的各种实例。

[0253] 在框1270处,通过以下方式计算百分比浓度:将第一密度与另一个密度比较以获得差别。将差别用参考密度标准化,其可以在框1270中进行。举例来说,可以通过将差别除以参考密度来将差别用参考密度标准化。在另一个实施方案中,可以在前面的框中标准化差别。

[0254] 在一个实施方案中,另一个密度是参考密度,例如如以上部分2中。因此,计算百分比浓度可以包括将差别乘以二。在另一个实施方案中,另一个密度是由鉴别为展现1拷贝损失(其中使用鉴别为展现1拷贝增加的各自密度计算第一密度)的各自密度计算的第二密度,例如如以上部分1中所描述。在此情况下,可以通过计算第一密度与参考密度的第一比率(例如 R_{right})并且计算第二密度与参考密度的第二比率(R_{left})来测定标准化的差别,其中差别在第一比率与第二比率之间。如上文所述,可以通过将峰值拟合到各自的密度的直方图的分布曲线来执行展现1拷贝损失或1拷贝增加的基因组区域的鉴别。

[0255] 总之,实施方案可以分析不同染色体区域中的血浆DNA的基因组呈现量,以同时确定染色体区域在肿瘤组织中是否存在扩增或缺失,并且如果所述区域存在扩增或缺失,以将其基因组呈现量用以推断肿瘤来源的DNA的百分比浓度。一些实施方案使用正态分布混合物模型来分析不同区域的基因组呈现量的总分布以便测定与不同类型的畸变相关的基因组呈现量,即增加1、2、3或4个拷贝和损失1或2个拷贝。

[0256] 实施方案具有数种优于其它方法的优势,所述其它方法例如全基因组聚集等位基因丢失(GAAL)方法(美国专利申请13/308,473;陈KC等人临床化学2013;59:211-24)和肿瘤相关的单核苷酸突变的分析(弗休T等人科学·转化医学2012;4:136ra68)。定位到具有染色体畸变的区域的所有序列读取都可以用以测定区域的序列读取密度,并且因此关于肿瘤DNA的分数浓度是有信息量的。另一个方面,在GAAL分析中,仅覆盖个体中的杂合子并且定位于具有染色体增加或缺失的染色体区域内的单核苷酸的序列读取将是有信息量的。类似地,对于癌症相关的突变的分析,仅覆盖突变的序列读取将适用于推断肿瘤DNA浓度。因此,实施方案可以允许更成本有效地使用测序数据,因为当与其它方法相比时,可能需要相对更少的测序读取来在肿瘤来源的DNA的百分比浓度的评估中实现相同准确性程度。

[0257] X. 替代性方法

[0258] 除使用特定突变可见于序列标签上的次数作为将基因座鉴别为真实突变的准则(由此调节阳性预测值)之外,可以代替或加之于使用截止值的方法,采用其它技术以在鉴

别癌症相关突变中提供更大预测作用。举例来说,当处理测序数据时,可以例如通过考虑测序的核苷酸的质量分数,使用不同严格性的生物信息学过滤手段。在一个实施方案中,可以使用具有不同测序错误分布的DNA测序仪和测序化学物质。具有较低测序错误率的测序仪和化学物质将得到较高阳性预测值。还可以使用对相同DNA片段的重复测序以增加测序准确性。一种可能的策略是太平洋生物科学提供的循环共有测序策略。

[0259] 在另一个实施方案中,可以将所测序的片段的大小信息并入到数据的解释中。因为肿瘤来源的DNA比血浆中的非肿瘤来源的DNA短(参看美国专利申请第13/308,473号),所以含有潜在肿瘤来源的突变的较短血浆DNA片段的阳性预测值将高于较长血浆DNA片段的阳性预测值。如果对血浆DNA执行双末端测序,那么大小数据将可容易获得。作为一个替代方案,可以使用具有长读取长度的DNA测序仪,因此产生血浆DNA片段的完整长度。还可以在DNA测序之前对血浆DNA样品执行大小分级。可以用于大小分级的方法的实例包括凝胶电泳、使用微流体方法(例如卡立珀(Caliper) LabChip XT系统)和尺寸排阻层析。

[0260] 在另一个实施方案中,如果集中于血浆中的较短DNA片段,那么预期患有非血液癌症的患者的血浆中的肿瘤相关的突变的百分比浓度将增加。在一个实施方案中,可以比较血浆中的肿瘤相关的突变在两个或更多个不同大小分布的DNA片段中的百分比浓度。当与较大片段相比时,患有非血液癌症的患者在较短片段中将具有较高的肿瘤相关的突变的百分比浓度。

[0261] 在一些实施方案中,可以将来自相同血液样品的两个或更多个等分试样或来自在相同场合或不同场合获取的两个或更多个血液样品的测序结果组合。可见于多于一个等分试样或样品中的潜在突变将具有肿瘤相关的突变的更高阳性预测值。阳性预测值将随展示这种突变的样品的数目增加而增加。存在于在不同时间点获取的血浆样品中的潜在突变可以视为潜在突变。

[0262] XI. 实例

[0263] 以下是实例技术和数据,本发明的实施方案不应视为限制于此实例技术和数据。

[0264] A. 材料和方法

[0265] 关于样品收集,招募肝细胞癌(HCC)患者、慢性B型肝炎的携带者以及同时患有乳腺癌和卵巢癌的患者。所有HCC患者都具有巴塞罗纳临床肝癌(Barcelona Clinic Liver Cancer) A1期疾病。将来自所有参与者的外周血液样品收集到含EDTA的管子中。在HCC患者的癌症切除手术期间获得其肿瘤组织。

[0266] 将外周血液样品在4℃下在1,600g下离心10分钟。将血浆部分在4℃下在16,000g下再离心10分钟,并且然后在80℃下储存。根据QIAamp DSP DNA血液微型试剂盒(凯杰(Qiagen))的血液和体液方案提取来自4.8mL血浆的无细胞的DNA分子。对于每一病例,将血浆DNA用SpeedVac浓缩器(Savant DNA120;赛默科技(Thermo Scientific))浓缩到40μl最终体积,以用于后续制备DNA测序库。

[0267] 根据QIAamp DSP DNA血液微型试剂盒的血液和体液方案由患者的血沉棕黄层样品提取基因组DNA。用QIAamp DNA微型试剂盒(凯杰)由肿瘤组织提取DNA。

[0268] 用双末端样品制备试剂盒(伊路米那)根据制造商说明构筑基因组DNA样品的测序库。简单来说,首先将1-5微克基因组DNA用科瓦里斯(Covaris) S220聚焦超声发生器碎成为200bp片段。然后,将DNA分子用T4DNA聚合酶和克列诺(Klenow)聚合酶末端修复;然后将T4

多核苷酸激酶用以使5'末端磷酸化。通过缺乏3'-5'核酸外切酶的克列诺片段产生3'悬垂物。将伊路米那衔接子寡核苷酸连接到粘性末端。用12循环的PCR使衔接子连接的DNA富集。因为血浆DNA分子是短片段并且血浆样品中的总DNA的量相对小,所以当构筑来自血浆样品的DNA库时,我们省略片段化步骤并且使用15循环的PCR。

[0269] 将安捷伦2100生物分析仪(安捷伦技术)用以检查衔接子连接的DNA库的品质和大小。然后通过卡帕(KAPA)库定量试剂盒(卡帕生物系统(Kapa Biosystems))根据制造商说明测量DNA库。将DNA库稀释并且与双末端测序流动细胞杂交。在cBot簇产生系统(伊路米那)上用TruSeq PE簇产生试剂盒v2(伊路米那)产生DNA簇,接着在HiSeq 2000系统(伊路米那)上用TruSeq SBS试剂盒v2(伊路米那)进行51x2循环或76x2个循环的测序。

[0270] 借助于短寡核苷酸比对程序2(SOAP2)以双末端模式分析双末端测序数据。对于每个双末端读取,将来自每个末端的50bp或75bp与参考人类基因组(hg18)比对。对于每个末端的比对,允许至多2个核苷酸错配。然后分析2个末端的这些潜在比对的基因组坐标以确定任何组合是否将满足2个末端以正确定向且与相同染色体比对,横跨插入区域小于或等于600bp,并且定位到参考人类基因组中的单一位置。重复读取定义为其中插入的DNA分子在人类基因组中展示相同起始和末端位置的双末端读取;如先前所述去除重复读取(洛等人科学·转化医学2010;2:61ra91)。

[0271] 在一些实施方案中,对于配对的肿瘤和组成型DNA样品测序以鉴别肿瘤相关的单核苷酸变异体(SNV)。在一些实施方案中,我们集中于组成型DNA(在此实例中,是血沉棕黄层DNA)中的纯合位点处出现的SNV。原则上,肿瘤组织的测序数据中检测到但组成型DNA中不存在的任何核苷酸变异都可以是潜在突变(即SNV)。然而,如果将肿瘤组织的测序数据中任何核苷酸变化的单一出现视为肿瘤相关的SNV,那么由于测序错误(占0.1%-0.3%测序得到核苷酸),将在基因组中鉴别出数百万假阳性。一种减少假阳性数目的方法将是建立如下准则:观察到肿瘤组织的测序数据中要求相同核苷酸变化多次出现,随后将检测肿瘤相关的SNV。

[0272] 因为测序错误的出现是随机过程,所以由于测序错误导致的假阳性的数目将随观察的SNV够资格作为肿瘤相关的SNV所需的出现数目的增加而呈指数性减少。另一个方面,假阳性的数目将随测序深度增加而增加。这些关系可以用泊松和二项分布函数来预测。实施方案可以测定观察的SNV够资格作为肿瘤相关的出现的动态截止值。实施方案可以考虑肿瘤测序数据中的特定核苷酸的实际覆盖、测序错误率、所允许的最大假阳性率和突变检测的所要敏感性。

[0273] 在一些实例中,我们设定非常严格的准则来减少假阳性。举例来说,可能需要突变完全不存在于组成型DNA测序中,并且特定核苷酸位置的测序深度必须是20倍。在一些实施方案中,实现假阳性检测率小于 10^{-7} 所对应的突变序列标签出现频率的截止值的。在一些实例中,我们还将着丝粒、端粒和低复杂区域内的SNV过滤掉以使由于错误比对导致的假阳性减到最少。另外,还去除在特定SNV中定位到dbSNP构建135数据库中的已知SNP。

[0274] B. 切除前后

[0275] 图13A展示了根据本发明的实施方案在诊断时分析患有卵巢癌和乳癌的患者血浆中的突变的表1300。此处,我们证实患有双侧卵巢癌和乳癌的一个实例。将血浆的测序数据与患者的组成型DNA(白血球)的测序结果比较。将存在于血浆中但不存在于组成

型DNA中的单核苷酸变化视为潜在突变。各自在两个位点对患者的右和左侧上的卵巢癌取样,产生总共四个肿瘤样品。肿瘤突变是在所有四个来自不同位点的卵巢肿瘤组织中都检测到的突变。

[0276] 通过测序在血浆中检测到超过360万个单核苷酸变化至少一次。在这些变化之中,仅2,064个也在肿瘤组织中检测到,得到0.06%的阳性预测值。使用在血浆中检测到至少两次的准则,潜在突变的数目显著减少99.5%,变为18,885。肿瘤突变的数目仅减少3%,变为2,003,并且阳性预测值增加到11%。

[0277] 使用在血浆中检测到至少五次的准则,仅2,572个潜在突变被检测到,并且在其中,1,814个是在所有肿瘤组织中都检测到的突变,因此得到71%的阳性预测值。取决于所需要的敏感性和阳性预测值,出现数目(例如2、3、4、6、7、8、9、10等)的其它准则可以用于界定潜在突变。用作准则的出现数目越高,得到越高阳性预测值但是会相应减少敏感性。

[0278] 图13B展示了根据本发明的实施方案在肿瘤切除之后分析患有双侧卵巢癌和乳癌的患者血浆中的突变的表1350。对患者执行手术切除。在切除卵巢癌和乳癌之后一天获取血液样品。然后对血浆DNA测序。对于此实例,仅分析来自卵巢癌的突变。在血浆样品中检测到超过300万个潜在突变至少一次。然而,使用有至少五次出现的准则,潜在突变的数目减少到238。当与在诊断时获取的样品的、使用相同五个突变的准则所得的潜在突变的数目相比时,观察到显著减少。

[0279] 在一个实施方案中,血浆中检测的单核苷酸变化的数目可以用作参数来对癌症患者进行检测、监测和预测。不同出现数目可以用作准则来实现所要敏感性和特异性。预期具有较高肿瘤负荷并且因此具有较差预后的患者将具有可见于血浆中的较高突变负荷。

[0280] 对于所述分析,可以针对不同类型的癌症建立突变负荷分布。出于监测的目的,将看到,对治疗起反应的患者的血浆中的突变负荷将减少。如果肿瘤已经复发,例如在复发期间,那么预期突变负荷将增加。所述监测将允许监测所选治疗模式对患者的功效并且检测对特定治疗的抗性的出现。

[0281] 通过分析可以见于血浆DNA测序结果中的特定突变,还可以鉴别和预测对特定靶治疗的敏感性(例如表皮生长因子受体基因中的突变,并且响应于酪氨酸激酶抑制剂治疗)和抗性(例如结肠直肠癌中的KRAS突变,并且对通过帕尼单抗和西妥昔单抗的治疗具有抗性),并且可以引导治疗方案的计划。

[0282] 以上实例是针对双侧卵巢癌。还可以对乳癌的突变执行相同分析,并且然后将能够追踪血浆中的这些癌症类型两者的突变。还可以使用类似策略来追踪原发性癌症和其转移性癌症。

[0283] 实施方案将适用于筛选表观健康受试者或具有特定风险因素的受试者(例如吸烟状态、病毒状态(例如肝炎病毒携带者、人乳头瘤病毒感染受试者))中的癌症。可以见于所述受试者的血浆中的突变负荷将导致受试者将在特定时间范围内患上症状性癌症的风险。因此,预期血浆中具有较高突变负荷的受试者将比具有较低突变负荷的受试者表现出更高的风险。此外,血浆中的所述突变负荷的时间分布也将是强有力的风险指示。举例来说,如果受试者每年执行一次血浆突变负荷分析并且如果突变负荷逐渐地增加,那么此受试者应转移到癌症的其它筛选模式,例如使用胸部X射线、超声波、计算机断层扫描、磁共振成像或正电子发射断层扫描。

[0284] C.用以由测序血浆推断突变的动态截止值

[0285] 招募四名患有肝细胞癌 (HCC) 的患者和一名患有卵巢癌和乳癌的患者用于此研究。对于后一患者,我们集中于分析卵巢癌。在肿瘤手术切除前后由每名患者收集血液样品。还收集切除的肿瘤组织。使用HiSeq2000测序系统(伊路米那)对由肿瘤组织、手术前血液样品的白血细胞以及手术前后血浆样品提取的DNA测序。使用短寡核苷酸分析包2 (SOAP2) 将测序数据与参考人类基因组序列 (hg18) 比对 (李R (Li R) 等人生物信息学 (Bioinformatics) 2009;25:1966-1967)。将白血细胞的DNA序列视为每名研究受试者的组成型DNA序列。

[0286] 在此实例中,首先在不参考肿瘤组织的情况下由血浆DNA测序数据和CG推断肿瘤相关的SNM。然后,将来自血浆的推断结果与由肿瘤组织产生的测序数据(作为标准)比较以确定推断结果的准确性。在此方面,通过将来自肿瘤组织的测序数据与组成型序列比较以研究出肿瘤组织中的突变来制定标准。在此分析中,我们集中于所研究的受试者的组成DNA是纯合的核苷酸位置。

[0287] 1. 非标靶全基因组分析

[0288] 每名患者的白血细胞、肿瘤组织和血浆DNA的测序深度展示于表5中。

病例	中值测序深度（倍数）				
	白血细胞	肿瘤组织	手术前血浆	手术后血浆	
[0289]	HCC1	39	29	23	24
	HCC2	39	29	25	28
[0290]	HCC3	46	33	18	21
	HCC4	46	27	20	23
	卵巢癌患者	44	53	37	28

[0291] 表5. 四个HCC病例的不同样品的中值测序深度。

[0292] 将如表1中所示的用于界定血浆突变的最少出现频数 (r) 的动态截止值,将其用于鉴别每名患者的血浆中的突变。因为每个基因座的测序深度可以变化,所以截止值可以变化,其有效地提供了截止值对于基因座的读取的总数目的依赖性。举例来说,尽管中值深度小于50 (表5),但个别基因座的测序深度可以变化非常大并且被覆盖>100次。

[0293] 除了测序错误之外,另一个错误来源将是比对错误。为将此类型的错误减到最少,使用鲍泰伊 (Bowtie) 比对程序将携带突变的序列读取与参考基因组再比对 (兰米德B (Langmead B) 等人基因组生物学2009,10:R25)。仅将可以通过SOAP2和鲍泰伊与参考基因组的独特位置比对的读取用于血浆突变的下游分析。也可以使用基于不同算法的比对软件包的其它组合。

[0294] 为了进一步使实际测序数据中的测序和比对错误减到最少,我们应用两个额外过滤算法来检查展示序列读取中的单核苷酸变异所对应的核苷酸位置:(1) 可以使用鲍泰伊以比对质量 $\geq Q20$ (即错误比对概率 $<1\%$) 将 $\geq 70\%$ 的携带突变的序列读取与相同基因组坐标再比对;(2) $\geq 70\%$ 的携带突变的序列读取不在序列读取的两个末端 (即5'和3'末端) 的5bp内。建立此过滤规则,因为测序错误在序列读取的两个末端处更普遍。

[0295] 我们还研究影响在无肿瘤基因组的先验知识的情况下推断肿瘤的因素。一个所述参数是血浆中的肿瘤来源的DNA的百分比浓度。此参数可以视为另一个标准参数,并且在肿

瘤基因组的先验知识下使用GAAL推断以用于参考目的。

[0296] 表6展示治疗前和血浆中检测到的核苷酸变异。对于HCC1,在无肿瘤基因组的先验知识的情况下,检测到总共961个单核苷酸变异。在血浆中检测到的这些核苷酸变异之中,828个是癌症相关的突变。在手术切除HCC之后,核苷酸变异的总数目减少到43并且其无一者是癌症相关的突变。

[0297] 出于参考目的,手术前血浆样品中的肿瘤来源的DNA的百分比浓度是53%,并且在有肿瘤基因组的先验知识下推断。对于HCC2、HCC3和HCC4,在无肿瘤基因组的先验知识的情况下,血浆中的单核苷酸变异的数目对于手术前血浆样品推断为在27到32范围内。这些结果与以下数学预测相一致,在约20倍的测序深度下,在血浆中可以检测到极低百分比的癌症相关的突变并且血浆中检测到的大多数序列变异归因于测序错误。在肿瘤切除之后,所检测到的序列变异的数目不存在显著变化。出于参考目的,血浆中的肿瘤来源的DNA的百分比浓度推断为在2.1%到5%范围内,并且在有肿瘤基因组的先验知识下推断。

[0298]		手术前血浆			手术后血浆		
		肿瘤来源的 DNA 的百分 比浓度	单核苷酸 变异的总 数目	所鉴别的 癌症相关 的突变的 数目	肿瘤来源的 DNA 的百分 比浓度	单核苷酸 变异的总 数目	所鉴别的 癌症相关 的突变的 数目
	HCC1	53%	961	828	0.4%	43	0
	HCC2	5%	32	0	0.6%	49	0
	HCC3	2.1%	29	0	0.2%	32	0
	HCC4	2.6%	27	0	1.3%	35	1
	卵巢癌(和乳 癌) 患者	46%	1718	1502	0.2%	2	0

[0299] 表6. 血浆中检测到的核苷酸变异。

[0300] 2. 外显子的标靶富集

[0301] 如上文所论述,增加所关注的区域的测序深度可以增加鉴别血浆中的癌症相关的突变的敏感性和特异性两者,并且因此增加癌症患者与非癌症受试者之间的区别能力。而增加全基因组的测序深度仍然是成本极高的,一种替代方案是使某些区域富集以用于测序。在一个实施方案中,可以使所选外显子或实际上整个外显子组标靶富集以用于测序。此方法可以显著增加标靶区域的测序深度而不增加序列读取的总量。

[0302] 使用用于使外显子组标靶富集的安捷伦SureSelect所有外显子试剂盒捕获HCC患者和患有卵巢癌(和乳癌)的患者的血浆DNA的测序库。然后使用HiSeq 2000测序系统对外显子富集的测序库测序。将序列读取与人参考基因组(hg18)比对。在比对之后,分析独特地定位到外显子的序列读取的单核苷酸变异。为了在外显子组捕获分析中鉴别血浆中的单核苷酸变异,使用表2中所示的动态截止值。

[0303] 图14A是展示检测HCC1的血浆DNA中的单核苷酸变异的表1400。在无肿瘤基因组的先验知识的情况下,我们由标靶测序数据推断血浆中的总共57个单核苷酸变异。在后续由获自肿瘤组织的测序数据的证实中,发现55个是真实肿瘤相关的突变。如之前所论述,手术前血浆中的肿瘤来源的DNA的百分比浓度是53%。在肿瘤切除之后,在获自血浆的标靶测序数据中未检测到单核苷酸变异。这些结果表明,血浆中的单核苷酸变异的数目的定量分析可以用于监测癌症患者的疾病进展。

[0304] 图14B是展示检测HCC2的血浆DNA中的单核苷酸变异的表1450。在无肿瘤基因组的先验知识的情况下,我们由标靶测序数据推断血浆中的总共18个单核苷酸变异。所有这些突变都见于肿瘤组织中。如之前所论述,手术前血浆中的肿瘤来源的DNA的百分比浓度是5%。在肿瘤切除之后,在血浆中未检测到单核苷酸变异。与血浆中的肿瘤来源的DNA的百分比浓度较高的HCC1相比,在HCC2的病例的血浆中检测到较少单核苷酸变异。这些结果表明,血浆中的单核苷酸变异的数目可以用作参数来反映血浆中的肿瘤来源的DNA的百分比浓度和患者的肿瘤负荷,因为已经显示血浆中的肿瘤来源的DNA的浓度与肿瘤负荷正相关(陈KC等人临床化学2005;51:2192-5)。

[0305] 图15A是展示检测HCC3的血浆DNA中的单核苷酸变异的表1500。在无肿瘤基因组的先验知识的情况下,我们由标靶测序数据在切除前和切除后血浆样品两者中均未观察到任何单核苷酸变异。这可能是由于此患者的血浆中的肿瘤来源的DNA的相对低百分比浓度(2.1%)。预测测序深度的进一步增加可改进在肿瘤来源的DNA的百分比浓度低的情况下检测癌症相关的突变的敏感性。

[0306] 图15B是展示检测HCC4的血浆DNA中的单核苷酸变异的表1550。在无肿瘤基因组的先验知识的情况下,我们由标靶测序数据推断血浆中的总共3个单核苷酸变异。所有这些突变都见于肿瘤组织中。与血浆中的肿瘤来源的DNA的分数浓度较高的HCC1和HCC2相比,在血浆中的分数肿瘤DNA是2.6%的病例HCC4的血浆中检测到较少单核苷酸变异。这些结果表明,血浆中的单核苷酸变异的数目可以用作参数来反映血浆中的肿瘤来源的DNA的百分比浓度和患者的肿瘤负荷。

[0307] 图16是展示检测患有卵巢癌(和乳癌)的患者的血浆DNA中的单核苷酸变异的表1600。在无肿瘤基因组的先验知识的情况下,我们由标靶测序数据推断血浆中的总共64个单核苷酸变异。在其之中,发现59个在卵巢肿瘤组织中。血浆中的卵巢肿瘤来源的DNA的估计百分比浓度是46%。在切除卵巢癌之后,检测到血浆中单核苷酸变异的总数目显著减少。

[0308] 除了使用SureSelect标靶富集系统(安捷伦)之外,我们还使用Nimblegen SeqCap EZ外显子组+UTR标靶富集系统(罗氏)使来自外显子的序列富集以用于测序。Nimblegen SeqCap系统覆盖基因组的外显子区域以及5'和3'非翻译区域。分析四名HCC患者、两名健康对照受试者和两名不患有癌症的慢性B型肝炎携带者的治疗前血浆样品(表7)。在其它实施方案中,可以使用其它标靶富集系统,包括(但不限于)使用基于液相或固相杂交的那些富集系统。

[0309]		治疗前血浆			治疗后血浆	
		通过 GAAL 分	血浆中检测	与相应肿瘤组	血浆中检测	与相应肿瘤组
[0310]		析所得的血浆中的肿瘤来源的 DNA 的百分比浓度	到的序列变异满足动态截止值的数目	织中检测到的突变重叠的序列变异的数目	到的序列变异满足动态截止值的数目	织中检测到的突变重叠的序列变异的数目
	HCC1	53%	69	64	1	1
	HCC2	5%	51	47	3	0
	HCC3	2.1%	0	0	1	0
	HCC4	2.6%	8	7	0	0

[0311] 表7. 四名HCC患者 (HCC1-4) 使用用于序列捕获的Nimblegen SeqCap EZ外显子组+

UTR标靶富集系统所得的外显子组测序结果。HCC3的治疗前血浆的测序分析由于PCR复制读取的较高百分比而是次优的。

[0312] 在两名慢性B型肝炎携带者和两名健康对照受试者中，检测到一个或更少单核苷酸变异满足动态截止值准则(表8)。在四名HCC患者中的三者中，血浆中检测到的序列变异满足动态截止值需求的数目是至少8。在HCC3中，未检测到满足动态截止值的SNV。在此样品中，在序列读取中存在高比例PCR复制读取，导致非复制序列读取的数目较低。在手术切除肿瘤之后，观察到血浆中可检测的SNV显著减少。

[0313]		血浆中检测到的序列变异满足动态截止的数目
	HBV1	0
	HBV2	1
	Ctrl1	1
	Ctrl2	1

[0314] 表8.2名慢性B型肝炎携带者(HBV1和HBV2)和2名健康对照受试者(Ctrl1和Ctrl2)使用用于序列捕获的Nimblegen SeqCap EZ外显子组+UTR标靶富集系统所得的外显子组测序结果。

[0315] XII. 肿瘤异质性

[0316] 生物样品(例如血浆/血清)中的单核苷酸突变的定量还适用于分析肿瘤异质性，肿瘤内和肿瘤间异质性。肿瘤内异质性涉及相同肿瘤内肿瘤细胞的多种克隆的存在。肿瘤间异质性涉及相同组织学类型的两个或更多个肿瘤的肿瘤细胞的多种克隆的存在，但存在于不同位点(同一器官或不同器官中)。在某些类型的肿瘤中，肿瘤异质性的存在是不良的预后指示(尹HH(Yoon HH)等人临床肿瘤学杂志(J Clin Oncol)2012;30:3932-3938;梅洛LMF(Merlo LMF)等人癌症预防研究(Cancer Prev Res)2010;3:1388-1397)。在某些类型的肿瘤中，肿瘤异质性的程度越高，在标靶治疗之后肿瘤进展或出现抗性克隆的机率将越高。

[0317] 尽管认为癌症由一种肿瘤细胞的克隆扩充产生，但癌症的生长和进化将导致新的并且不同的突变积聚于癌症的不同部分中。举例来说，当癌症患者出现转移时，定位于原始器官处的肿瘤和转移性肿瘤将共有多个突变。然而，两个位点的癌细胞还将携带不存在于另一个肿瘤位点中的独特突变集。预期由两个位点共有的突变以比仅在一个肿瘤位点中观察到的那些突变有更高的浓度存在。

[0318] A. 实例

[0319] 我们分析患有双侧卵巢癌和乳癌的患者血浆。两个卵巢肿瘤都是浆液性腺癌。在最长尺寸方面，左肿瘤测量为6cm并且右肿瘤测量为12cm。还在结肠和网膜处存在多个转移性病变。使用来自伊路米那的边合成边测序平台以平均单倍型基因组覆盖度为44倍对由白血球提取的DNA测序。针对血浆中的单核苷酸突变，进一步分析仅展示一种等位基因、即纯合的核苷酸位置。

[0320] 由左和右肿瘤的四个不同位点提取DNA，并且使用伊路米那测序平台对其测序。两个位点(位点A和B)来自右肿瘤，并且另两个位点(位点C和D)来自左肿瘤。位点A和B相隔约4cm。位点C与D之间的距离也是约4cm。在手术切除卵巢肿瘤前后由患者收集血浆样品。然后由患者的血浆提取DNA。位点A、B、C和D的肿瘤以及血浆样品的测序深度展示于表9中。

[0321]

样品	原始测序读取的数目	比对的读取的数目	单倍型基因组覆盖的倍数
来自血沉棕黄层的组成型 DNA	1,091,250,072	876,269,922	43.81
右卵巢肿瘤 (位点 A)	1,374,495,256	1,067,277,229	53.36
右卵巢肿瘤 (位点 B)	934,518,588	803,007,464	40.15
左卵巢肿瘤 (位点 C)	1,313,051,122	1,036,643,946	51.83
左卵巢肿瘤 (位点 D)	1,159,091,833	974,823,207	48.74
在手术之前收集的血浆样品	988,697,457	741,982,535	37.10
在手术之后收集的血浆样品	957,295,879	564,623,127	28.23

[0322] 表9. 位点A、B、C和D的肿瘤的测序深度。

[0323] 在当前实例中,为了界定单一肿瘤相关的单核苷酸突变,对核苷酸位置测序至少20次(在肿瘤组织中)和30次(在组成型DNA中)。在其它实施方案中,可以使用其它测序深度,例如35、40、45、50、60、70、80、90、100和>100倍。测序成本的降低将使得增加测序深度要容易得多。核苷酸位置在组成型DNA中是纯合的,而在肿瘤组织中观察到核苷酸变化。肿瘤组织中的核苷酸变化的出现的准则取决于肿瘤组织中的特定核苷酸位置的总测序深度。对于20到30倍的核苷酸覆盖,核苷酸变化的出现频数(截止值)是至少五次。对于31到50倍的覆盖,核苷酸变化的出现频数是至少六次。对于51到70倍的覆盖,出现频数需求是至少七次。这些准则来源于对使用泊松分布检测真实突变的敏感性和假阳性基因座的数目的预期值的预测。

[0324] 图17是展示不同出现频数需求和测序深度的预测得到的敏感性的表1700。敏感性将对应于使用特定截止值在特定倍数深度下检测到的真实突变的数目。测序深度越高,关于既定截止值越可能检测到突变,因为将获得越多突变序列读取。对于较高截止值,将不大可能检测到突变体,因为准则更严格。

[0325] 图18是展示针对不同截止值和不同测序深度的假阳性基因座的预测数目的表1800。假阳性的数目随测序深度增加而增加,因为获得更多序列读取。然而,对于五或更大的截止值,即使高达70的测序深度,也无假阳性。在其它实施方案中,可以使用不同出现准则以便实现所要的敏感性和特异性。

[0326] 图19展示了说明不同肿瘤位点检测到的突变数目的树图。通过直接对肿瘤测序来测定突变。位点A具有71个对于所述肿瘤特异性的突变,并且位点B具有122个位点特异性突变,纵使其仅相隔4cm。在位点A和B两者中可见10个突变。位点C具有168个对于所述肿瘤特异性的突变,并且位点D具有248个位点特异性突变,纵使其仅相隔4cm。在位点C和D两者中可见12个突变。不同肿瘤位点的突变分布存在显著异质性。举例来说,248个突变仅在位点D肿瘤中检测到,但在另三个肿瘤位点中未检测到。跨越所有位点可见总共2,129个突变。因此,许多突变在不同肿瘤之中是共有的。因此,存在七个SNV类别。在这四个区域之中在拷贝数畸变方面不存在可观察的差异。

[0327] 图20是表2000,其展示治疗前和治疗后血浆样品中携带肿瘤来源的突变片段的数目。还展示携带各自突变的肿瘤来源的DNA的推断得到的百分比浓度。突变的类别是指检测到突变的肿瘤位点。举例来说,类别A突变是指仅存在于位点A的突变,而类别ABCD突变是指

存在于所有四个肿瘤位点的突变。

[0328] 对于2,129个存在于所有四个肿瘤位点处的突变,2,105个(98.9%) 在至少一个血浆DNA片段中可检测到。另一个方面,对于609个存在于四个肿瘤位点中的仅一者中的突变,仅77个(12.6%) 在至少一个血浆DNA片段中可检测到。因此,血浆中的单核苷酸突变的定量可以用于反映肿瘤组织中的这些突变的相对丰度。此信息将适用于研究癌症异质性。在此实例中,潜在突变只要求在测序数据中出现至少一次。

[0329] 测定每个SNV类别对应的循环肿瘤DNA的百分比浓度。如通过由所有4个区域共有的SNV(即群组ABCD)测定在手术之前和在手术之后血浆中的肿瘤DNA的百分比浓度,其分别是46%和0.18%。这些较后百分比与GAAL分析中获得的百分比46%和0.66%充分相关。由所有4个区域共有的突变(即群组ABCD)向血浆提供肿瘤来源的DNA是最高分数的贡献。

[0330] 关于群组AB和CD的SNV所测定的手术前血浆中的肿瘤来源的DNA的百分比浓度分别是9.5%和1.1%。这些浓度与右和左卵巢肿瘤的相对大小一致。关于区域独特SNV(即群组A、B、C和D中的那些)所测定的肿瘤来源的DNA的百分比浓度一般来说很低。这些数据表明,对于癌症患者中的总肿瘤负荷的准确测量,与更传统的靶向特定肿瘤相关突变的方法相比,使用全基因组鸟枪方法可以提供更具代表性的图像。对于后一种方法,如果仅肿瘤细胞的子集具有标靶突变,那么可能会错漏关于由不具有标靶突变的肿瘤细胞引起的即将发生的复发或疾病进展的重要信息,或可能会错过发现治疗抗性克隆的出现。

[0331] 图21是图表2100,其展示血浆中单一肿瘤位点检测到的突变和所有四个肿瘤位点检测到的突变的出现分布。条形图2100展示两种类型的突变的数据:(1) 仅在一个位点中检测到的突变,和(2) 在所有四个肿瘤位点中检测到的突变。横轴是血浆中检测到突变的次数。竖轴展示突变的对应于横轴上的特定值的百分比。举例来说,约88%的类型(1) 突变在血浆中仅出现一次。如可以看到,在一个位点中出现的突变大部分被检测到一次,并且不多于四次。与存在于所有四个肿瘤位点中的突变相比,存在于单一肿瘤位点中的突变在血浆中被检测到的频率要低得多。

[0332] 此技术的一个应用将是使得临床医师可估计携带不同类别的突变的肿瘤细胞的负荷。这些突变的比例将可能可用标靶药剂治疗。预期靶向由较高比例的肿瘤细胞携带的突变的药剂将具有较为突出的治疗效果。

[0333] 图22是图表2200,其展示血浆中来自异质肿瘤的突变的预测出现分布。肿瘤含有两组突变。一组突变存在于所有肿瘤细胞中,并且另一组突变仅存在于1/4的肿瘤细胞中,基于两个位点代表每个卵巢肿瘤的近似。假定血浆中的肿瘤来源的DNA的总百分比浓度是40%。假定以每核苷酸位置50倍的平均深度对血浆样品测序。根据血浆中的此出现频数的预测分布,可以通过其在血浆中的出现频数将存在于所有肿瘤组织中的突变与仅存在于1/4肿瘤细胞中的突变区分。举例来说,6次出现可以用作截止值。对于存在于所有肿瘤细胞中的突变,92.3%的突变将存在于血浆中至少6次。相比之下,对于存在于1/4肿瘤细胞中的突变,仅12.4%的突变将存在于血浆中至少6次。

[0334] 图23是表2300,其本发明的实施方案阐述在16名健康对照受试者中的特异性。以30倍的中值覆盖对其血浆DNA样品测序。在这些健康受试者的血浆样品中对存在于以上卵巢癌患者的血浆中的突变执行检测。存在于卵巢癌患者的肿瘤中的突变极罕见地在健康对照受试者的血浆的测序数据中检测到,并且所述类别的突变中无一者的表观百分比浓度>

1%。这些结果显示,此检测方法是高度特异性的。

[0335] B. 方法

[0336] 图24是根据本发明的实施方案分析受试者的一个或多个肿瘤的异质性的方法2400的流程图。可能如本文中所描述执行方法2400的某些步骤。

[0337] 在框2410处,获得受试者的组成型基因组。在框2420处,对于受试者的生物样品中的多个DNA片段中的每一者,获得一或多种序列标签,其中生物样品包括无细胞的DNA。在框2430处,测定序列标签的基因组位置。在框2440处,将序列标签与组成型基因组比较以测定第一基因座的第一数目。在每个第一基因座处,相对于组成型基因组具有变异序列的序列标签的数目高于截止值,其中截止值大于一。

[0338] 在框2450处,基于第一基因组位置集合的各个第一数目计算一个或多个肿瘤的异质性的度量。在一个方面,度量可以提供代表由肿瘤共用的突变的数目相对于不由肿瘤共用的突变的数目的值。此处,多个肿瘤可以在单一对象中存在,所述对象内的不同肿瘤可以代表通常称为肿瘤内异质性者。度量还可以指与许多或大多数肿瘤中的突变相比,是否有一些突变在一个还是少数几个肿瘤中。可以计算多于一个异质性度量。

[0339] 在框2460处,可以将异质性度量与阈值比较以确定异质性等级的分类。可以按各种方式使用一或多个度量。举例来说,一或多个异质性度量可以用以预测肿瘤进展的机率。在一些肿瘤中,异质性越大,在治疗(例如标靶治疗)之后,癌变发生机率越高,并且抗性克隆出现的机率越高。

[0340] C. 肿瘤异质性度量

[0341] 异质性度量的一个实例是血浆中的不同组别的突变的‘浓度带’的数目。举例来说,如果患者中存在两种主要肿瘤克隆,并且如果这些克隆以不同浓度存在,那么我们将预期在血浆中看到不同浓度的两种不同突变。可以通过测定不同突变集合的百分比浓度来计算这些不同值,其中每个集合对应于肿瘤中的一者。

[0342] 这些浓度中的每一者可以称为‘浓度带’或‘浓度类’。如果患者具有更多个克隆,那么将观察到更多浓度带/类。因此,带越多,异质性越大。可以通过绘制各种突变的百分比浓度看到浓度带的数目。可以对各个浓度制直方图,其中不同峰值对应于不同肿瘤(或一个肿瘤的不同克隆)。大峰值将可能是由所有或一些肿瘤(或肿瘤的克隆)共有的突变。可以分析这些峰值以确定哪些较小峰值组合确定了较大峰值。可以使用拟合程序,例如类似于图10B和11的拟合程序。

[0343] 在一个实施方案中,直方图是y轴是基因座的量(例如数目或比例)并且x轴是百分比浓度的图。由所有或一些肿瘤共有的突变将产生较高百分比浓度。峰值大小将代表产生特定百分比浓度的基因座的量。峰值在低和高浓度下的相对大小将反映肿瘤(或肿瘤的克隆)的异质性的程度。高浓度下的较大峰值反映,大多数突变由大多数或所有肿瘤(或肿瘤的克隆)共有并且指示较低肿瘤异质性程度。如果低浓度下的峰值较大,那么大多数突变由几个肿瘤(或肿瘤的几个克隆)共有。这将指示较高肿瘤异质性程度。

[0344] 存在的峰值越多,存在的位点特异性突变越多。每个峰值可以对应于不同突变集合,其中突变集来自肿瘤的子集(例如仅一或两个肿瘤,如以上所说明)。对于图19的实例,可以存在总共7个峰值,4个可能具有最小浓度(取决于肿瘤的相对大小)的位点特异的峰值、两个关于AB位点和CD位点的峰值和一个关于由所有位点共有的突变的峰值。

[0345] 峰值的位置还可以提供肿瘤的相对大小。较大浓度将与较大肿瘤相关,因为较大肿瘤将释放较多肿瘤DNA到样品中,例如到血浆中。因此,可以估计携带不同类别的突变的肿瘤细胞的负荷。

[0346] 异质性度量的另一个实例是与变异体读取相对高(例如9-13)的突变读取的比例相比变异体读取相对少(例如4、5或6)的突变位点的比例。返回参看图22,可以看到,位点特异性突变具有较少变异体读取(其也产生较小百分比浓度)。共有突变具有较多变异体读取(其也产生较大百分比浓度)。在6(较小计数)下的第一比例除以在10(较大计数)下的第二比例的比率产生了异质性度量。如果比率小,那么极少突变是位点特异性的,并且因此异质性等级低。如果比率大(或至少大于由已知样品校准的值),那么异质性等级较大。

[0347] D. 测定阈值

[0348] 可以由肿瘤被活检(例如如上文所述)的受试者测定阈值以直接确定异质性等级。等级可以按各种方式界定,例如位点特异性突变与共有突变的比率。然后可以分析生物样品(例如血浆样品)以测定异质性度量,其中生物样品的异质性度量可以与通过直接分析肿瘤的细胞测定的异质性等级相关。

[0349] 这种程序可以提供阈值相对于异质性等级的校准。如果测试异质性度量处于两个阈值之间,那么异质性等级可以估计为在对应于阈值的等级之间。

[0350] 在一个实施方案中,可以在由活检体测定的异质性等级与由血浆样品(或其它样品)测定的相应异质性度量之间计算校准曲线。在这种实例中,异质性等级是数值,其中这些数值等级可以对应于不同分类。不同数值等级范围可以对应于不同诊断,例如不同癌症阶段。

[0351] E. 使用来自基因组呈现量的百分比浓度的方法

[0352] 还可以使用例如如使用方法1200的实施方案测定的百分比浓度分析肿瘤异质性。展现一拷贝损失的基因组区域可以来自不同肿瘤。因此,各个基因组区域的百分比浓度可以不同,其取决于取决于扩增(或1拷贝缺失)是否存在于仅一个肿瘤还是多个肿瘤。因此,可以使用关于经由方法1200的实施方案测定的百分比浓度来度量的肿瘤异质性。

[0353] 举例来说,一个基因组区域可以鉴别为对应于1拷贝损失,并且分数浓度可以仅由所述基因组区域处的各自密度测定(各自密度可以用作百分比浓度)。可以通过计算具有各种密度的区域的数目测定直方图。如果仅一个肿瘤或一个肿瘤克隆或一个肿瘤种植在特定区域中有所增加,那么所述区域的密度将小于有多个肿瘤或多个肿瘤克隆或多个肿瘤种植增加的区域的密度(即共有区域中的肿瘤DNA的百分比浓度将大于位点特异性区域)。上文所描述的异质性度量因此可以应用于使用各种区域中的拷贝数增加或损失而鉴别的峰值,正如不同位点的百分比浓度展示百分比浓度的分布。

[0354] 在一个实施方案中,如果各自密度用于直方图,可以将使增加和损失的区域分开来分析。可以通过仅对于增加的区域对应的各自密度产生直方图,并且可以仅对于损失的区域对应的各自密度产生的直方图。如果使用百分比浓度,那么可以将损失和增加所对应的峰值一起分析。举例来说,使用百分比浓度与参考密度的差值(例如以绝对值形式),并且因此增加和损失的百分比浓度可以对同个峰值有贡献。

[0355] XIII. 计算机系统

[0356] 本文中提及的任何计算机系统都可以利用任何合适数目的子系统。所述子系统的

实例展示于图25中计算机设备2500。在一些实施方案中,计算机系统包括单一计算机设备,其中子系统可以是计算机设备的组件。在其它实施方案中,计算机系统可以包括多个具有内部组件的计算机设备,其各自是子系统。

[0357] 图25中所示的子系统经由系统总线2575互连。展示其它子系统,例如印刷机2574、键盘2578、固定磁盘2579、与显示适配器2582耦合的监视器2576等。与I/O控制器2571耦合的外围装置和输入/输出(I/O)装置可以通过本领域中已知的许多构件(例如串行端口2577)与计算机系统连接。举例来说,串行端口2577或外部接口2581(例如以太网、Wi-Fi等)可以用以将计算机系统2500连接到广域网(例如因特网)、鼠标输入装置或扫描仪。经由系统总线2575的互连使得中央处理器2573与每个子系统连通并且控制来自系统存储器2572或固定磁盘2579的指令的执行以及子系统之间的信息的交换。系统存储器2572和/或固定磁盘2579可以体现计算机可读介质。本文中提及的任何值可以由一个组件向另一个组件输出并且可以向用户输出。

[0358] 计算机系统可以包括例如通过外部接口2581或通过内部接口连接在一起的多个相同组件或子系统。在一些实施方案中,计算机系统、子系统或设备可以经网络连通。在所述情况下,一个计算机可以视为客户端并且另一个计算机视为服务器,其中每一者可以是同一计算机系统的一部分。客户端和服务器可以各自包括多个系统、子系统或组件。

[0359] 应理解,本发明的任何实施方案可以按控制逻辑形式以模块化或集成方式使用硬件(例如专用集成电路或现场可编程门阵列)和/或使用通用可编程处理器的计算机软件来实施。如本文中所用,处理器包括同一集成芯片上的多核处理器,或单一电路板或网络连接的电路板上的多个处理单元。基于本发明和本文中所提供的传授内容,本领域的普通技术人员将知道并且理解使用硬件和硬件与软件的组合来实施本发明的实施方案的其它方式和/或方法。

[0360] 本申请中描述的任何软件组件或函数可以实施为由处理器使用任何合适的计算机语言(例如Java、C++或Perl)、使用例如常规或面向对象的技术执行的软件代码。软件代码可以存储为用于存储和/或传输的计算机可读介质上的一系列指令或命令,合适的媒体包括随机存取存储器(RAM)、只读存储器(ROM)、磁性媒体(例如硬盘驱动器或软性磁盘)或光学媒体(例如光盘(CD)或DVD(数字通用光盘))、快闪存储器等。计算机可读介质可以是所述存储或传输装置的任何组合。

[0361] 所述程序还可以使用适合于经由符合多种方案的有线、光学和/或无线网络(包括因特网)传输的载波信号来编码和传输。因此,根据本发明的一个实施方案的计算机可读介质可以使用以所述程序编码的数据信号产生。以程序代码编码的计算机可读介质可以与相容装置一起封装或与其它装置分开地提供(例如经由因特网下载)。任何所述计算机可读介质可以存在于单一计算机程序产品(例如硬盘驱动器、CD或整个计算机系统)之上或之内,并且可以存在于系统或网络内的不同计算机程序产品之上或之内。计算机系统可以包括用于向用户提供本文中提及的任何结果的监视器、印刷机或其它合适显示器。

[0362] 本文中所描述的任何方法可以完全或部分用计算机系统执行,所述计算机系统包括一或多个处理器,所述处理器可以经配置以执行所述步骤。因此,实施方案可以涉及经配置以执行本文中所描述的任何方法的步骤的计算机系统,可能用不同组件执行分别的步骤或分别的步骤群组。尽管本文中方法的步骤以经编号步骤的形式呈现,但其可以同时或以

不同顺序执行。另外,这些步骤的部分可以与其它方法的其它步骤的部分一起使用。此外,步骤的全部或部分可以是任选的。另外,任何方法的任何步骤都可以用执行这些步骤的模块、电路或其它构件来执行。

[0363] 可以在不脱离本发明的实施方案的精神和范围的情况下以任何合适方式组合特定实施方案的特定细节。然而,本发明的其它实施方案可以涉及与每个个别方面或这些个别方面的特定组合相关的特定实施方案。

[0364] 出于说明和描述的目的呈现本发明的例示性实施方案的以上描述。其并不打算是穷尽性的或将本发明限制于所描述的精确形式,并且鉴于以上传授内容许多修改和变化是可能的。所述实施方案经选择和描述以便最佳地解释本发明的原理和其实际应用,由此使得本领域的其它技术人员能够在各种实施方案中并且以适于所预期的特定用途的各种修改最佳地利用本发明。

[0365] 除非具体地指示相反的意思,否则“一(a/an)”或“所述(the)”的叙述打算意指“一或多个”。

[0366] 此处提及的所有专利、专利申请、公开和描述都出于所有目的以全文引用的方式并入。不承认任一者是现有技术。

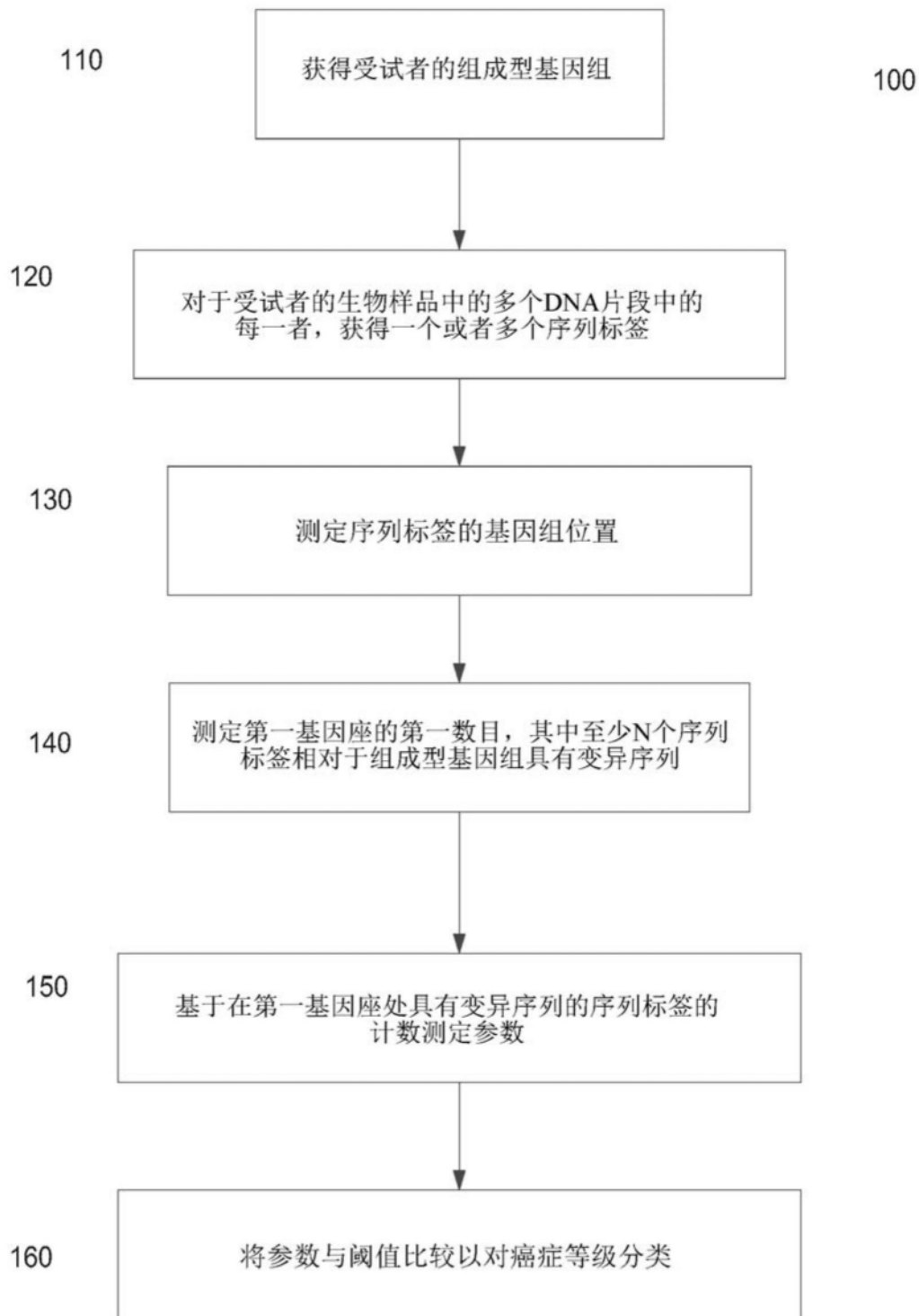


图1

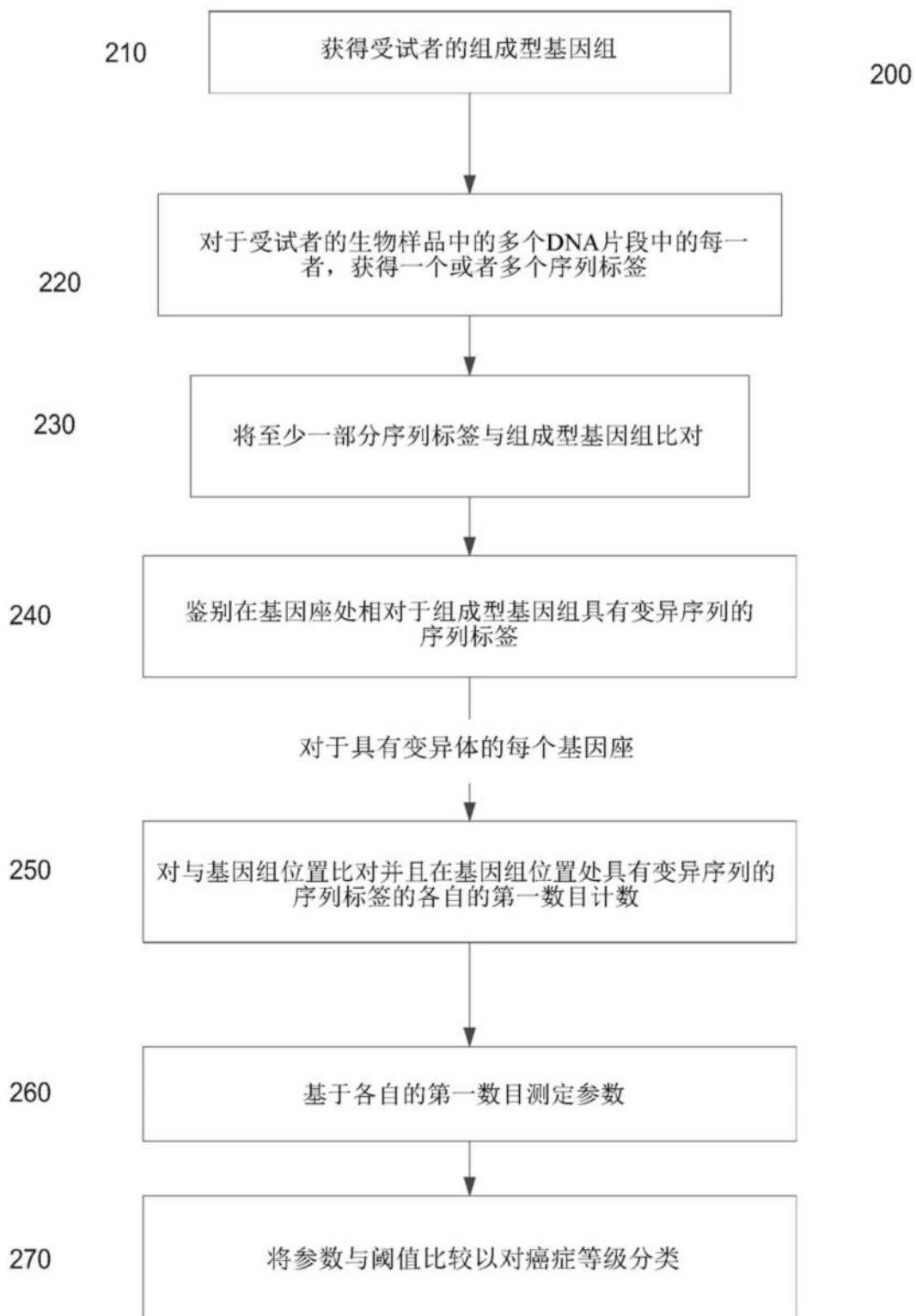


图2

300

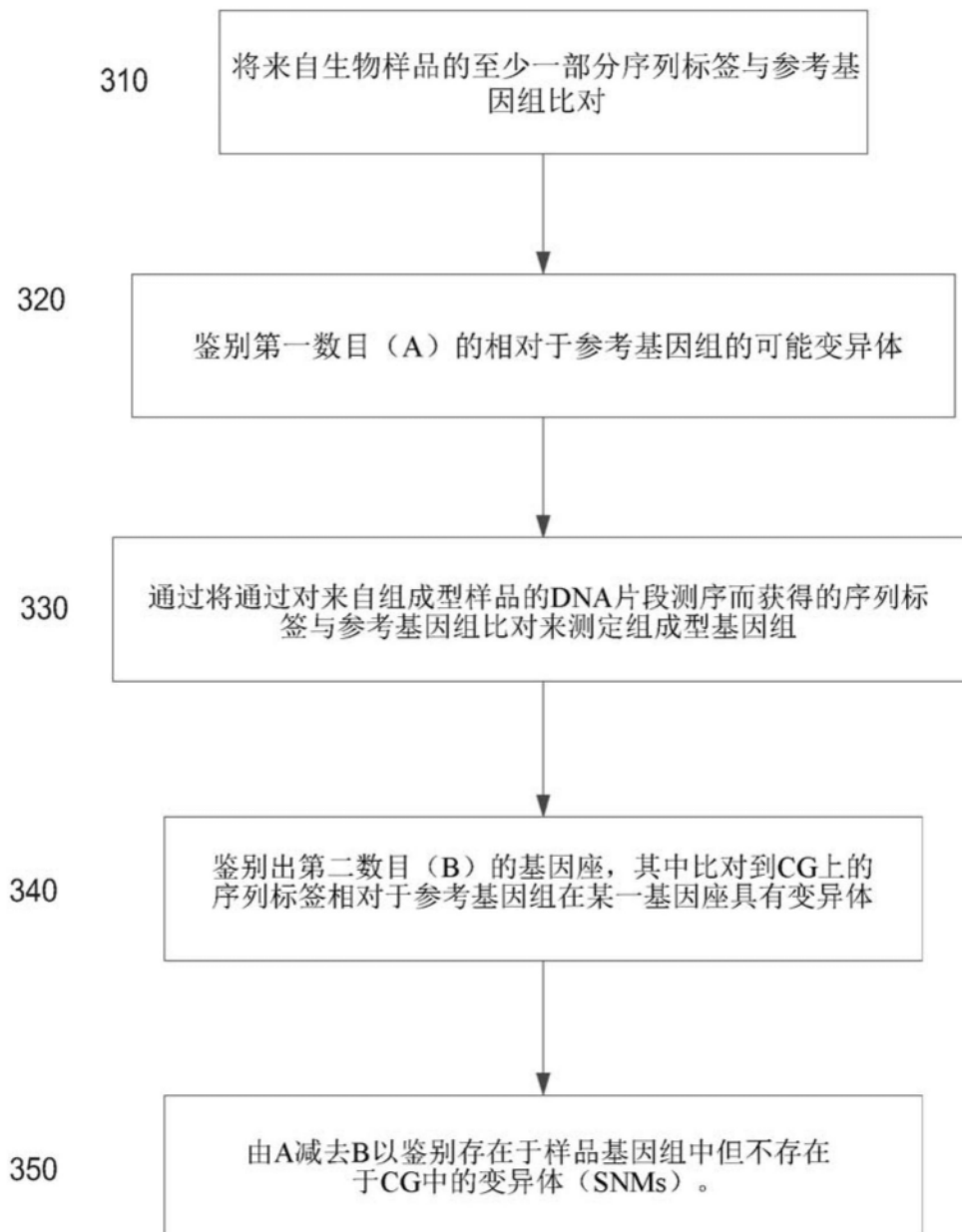


图3

被看到的 测序深度	1x		2x		3x		4x		5x		6x	
	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误的预期数目	鉴别的突变的预期数目
1	2,998,500	146	1,499	4	0	0	0	0	0	0	0	0
10	29,850,499	1,180	45,000	271	40	43	0	5	0	1	0	0
20	59,403,980	1,896	190,000	793	380	241	1	57	0	11	0	2
40	117,631,683	2,594	780,000	1,782	3,293	970	10	429	0	158	0	50
60	174,706,399	2,851	1,770,000	2,403	11,407	1,730	54	1,058	0	554	0	252
80	230,650,961	2,945	3,160,000	2,725	27,387	2,286	176	1,700	1	1,113	0	645
100	285,487,746	2,980	4,950,000	2,879	53,900	2,626	436	2,205	3	1,679	0	1,152
120	339,238,690	2,993	7,140,000	2,948	93,613	2,814	913	2,546	7	2,145	0	1,663
140	391,925,294	2,997	9,730,000	2,978	149,193	2,911	1,703	2,755	15	2,481	0	2,098
160	443,568,633	2,999	12,720,000	2,991	223,307	2,959	2,922	2,873	30	2,701	0	2,426
180	494,189,366	3,000	16,110,000	2,996	318,620	2,981	4,700	2,936	55	2,835	1	2,653
200	543,807,741	3,000	19,900,000	2,999	437,800	2,992	7,187	2,969	94	2,912	1	2,799

400

图4

被看到的 测序深度	1x		2x		3x		4x		5x		6x	
	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目	假阳性基因座的针对测序错误数 期数目	鉴别的突变的预期数目
1	2,998,500	74	1,499	1	0	0	0	0	0	0	0	0
10	29,850,499	664	45,000	79	40	6	0	0	0	0	0	0
20	59,403,980	1,180	190,000	271	380	43	1	5	0	1	0	0
40	117,631,683	1,896	780,000	793	3,293	241	10	57	0	11	0	2
60	174,706,399	2,331	1,770,000	1,327	11,407	573	54	197	0	56	0	13
80	230,650,961	2,594	3,160,000	1,782	27,387	970	176	429	1	158	0	50
100	285,487,746	2,754	4,950,000	2,138	53,900	1,369	436	727	3	326	0	126
120	339,238,690	2,851	7,140,000	2,403	93,613	1,730	913	1,058	7	554	0	252
140	391,925,294	2,909	9,730,000	2,592	149,193	2,037	1,703	1,390	15	824	0	427
160	443,568,633	2,945	12,720,000	2,725	223,307	2,286	2,922	1,700	30	1,113	0	645
180	494,189,366	2,967	16,110,000	2,817	318,620	2,479	4,700	1,973	55	1,404	1	891
200	543,807,741	2,980	19,900,000	2,879	437,800	2,626	7,187	2,205	94	1,679	1	1,152

500

图5

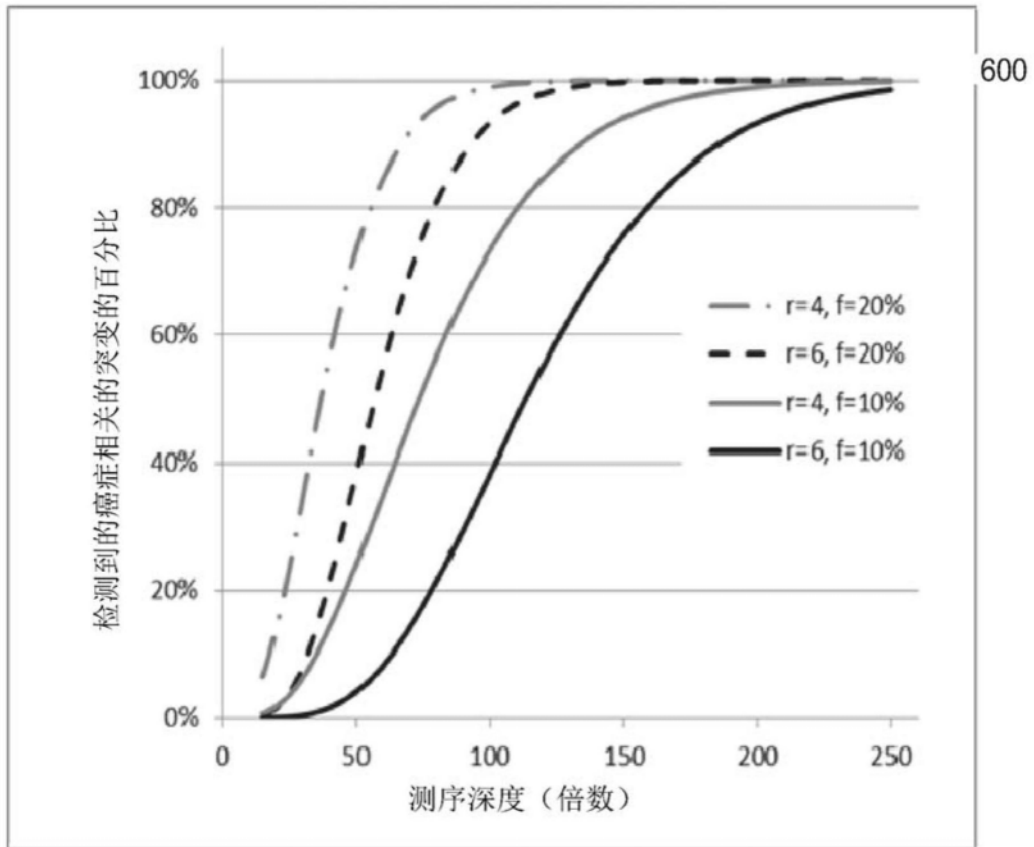


图6A

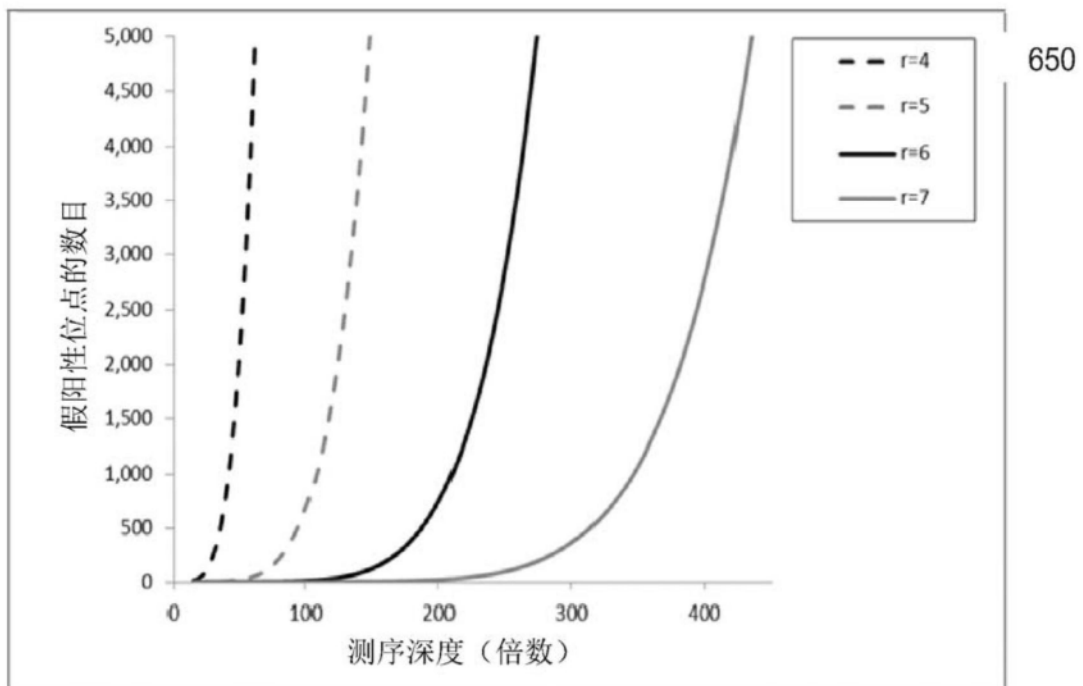
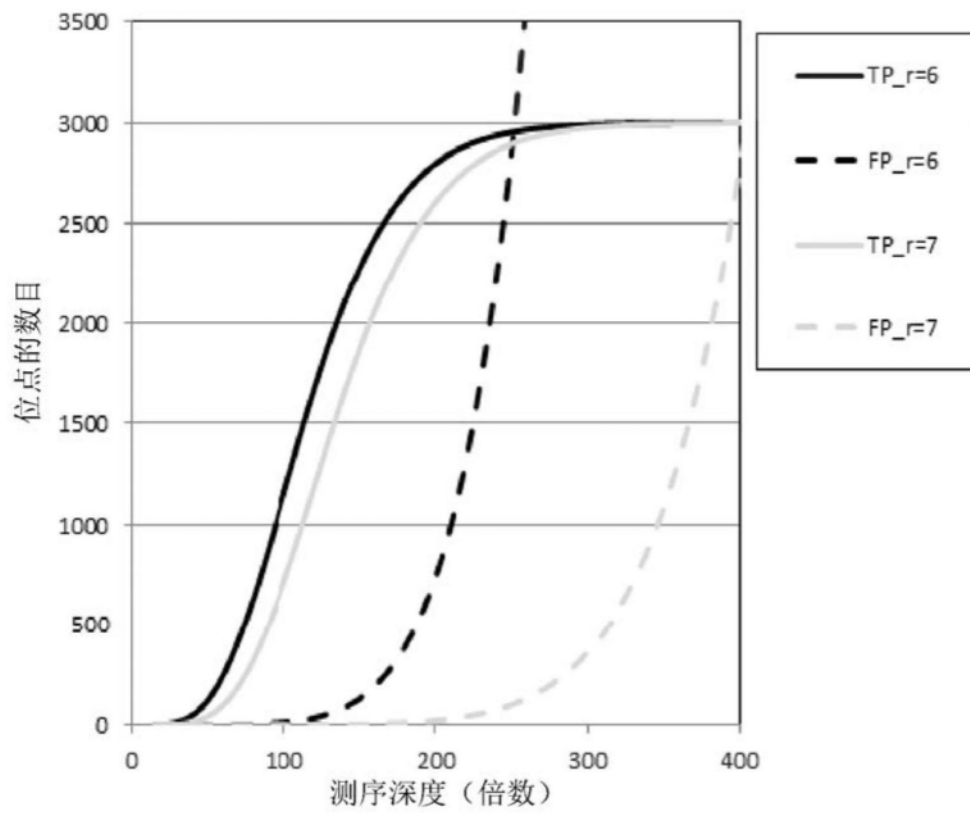


图6B



700

图7A

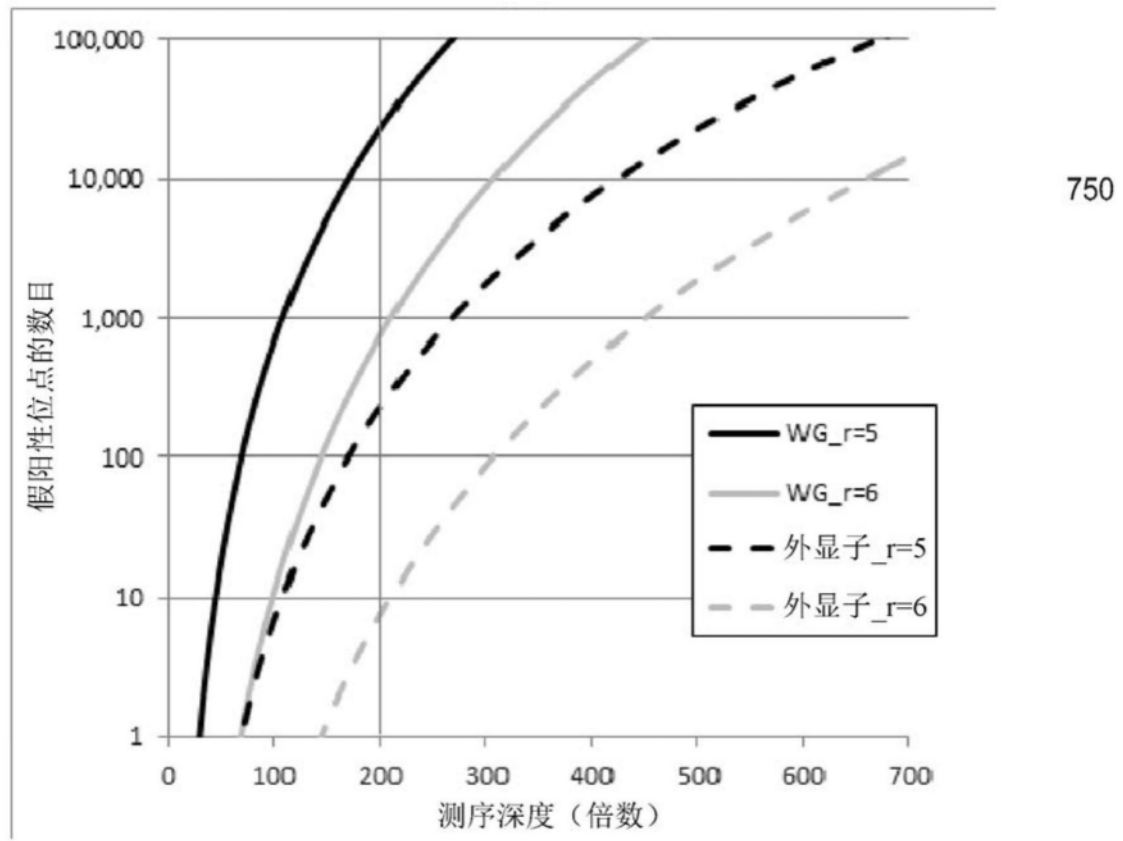


图7B

病例	肿瘤组织中检测到的SNV的数目	时间点	由血浆测序的肿瘤相关的SNV的数目 (肿瘤组织中所见的SNV的%)	展示SNV的血浆DNA序列读取的数目 (p)	展示野生型序列的血浆DNA序列读取的数目 (q)	通过SNV分析所得的肿瘤来源的DNA的推论百分比浓度 $(\frac{2p}{p+q})$	通过GAAL分析所得的肿瘤来源的DNA的推论百分比浓度
HCC1	2,840	治疗前	2,569 (94%)	11,389	31,602	53%	52%
		治疗后	44 (1.5%)	91	46,898	0.4%	1.4%
HCC2	3,105	治疗前	1,097 (35%)	1,490	57,865	5.0%	4.3%
		治疗后	72 (2.3%)	206	66,692	0.6%	0.9%
HCC3	3,171	治疗前	461 (15%)	525	48,886	2.1%	5.6%
		治疗后	31 (1%)	67	58,862	0.2%	0.9%
HCC4	1,334	治疗前	201 (15%)	248	18,527	2.6%	7.6%
		治疗后	74 (5.5%)	149	22,144	1.3%	2.7%

800

图8

60

	HCC1 SNV		HCC2 SNV		HCC3 SNV		HCC4 SNV		所有SNV		
	展示肿瘤相关的SNV的血浆DNA序列读取	展示野生型序列的血浆DNA序列读取	展示肿瘤相关的SNV的血浆DNA序列读取	展示野生型序列的血浆DNA序列读取	展示肿瘤相关的SNV的血浆DNA序列读取	展示野生型序列的血浆DNA序列读取	展示肿瘤相关的SNV的血浆DNA序列读取	展示野生型序列的血浆DNA序列读取	展示肿瘤相关的SNV的血浆DNA序列读取 (p)	展示野生型序列的血浆DNA序列读取 (q)	血浆中的所有肿瘤来源的SNV的百分比浓度 $(\frac{2p}{p+q})$
对照											
C01	40	31,261	66	38,172	30	33,823	69	14,389	205	117,645	0.35%
C02	60	34,368	112	42,660	37	37,462	58	16,097	267	130,587	0.41%
C03	55	37,931	100	46,297	50	41,181	74	17,039	279	142,448	0.39%
C04	41	40,277	87	48,441	39	43,541	87	17,358	254	149,617	0.34%
C05	42	31,655	71	38,787	46	34,566	70	14,821	229	119,829	0.38%
C06	41	34,604	97	43,999	26	37,477	79	16,762	243	132,842	0.37%
C07	52	38,664	118	47,200	61	41,840	72	17,601	303	145,305	0.42%
C08	57	38,514	131	46,464	42	41,662	81	17,275	311	143,915	0.43%
C09	41	34,945	115	43,760	33	38,039	81	16,503	270	133,247	0.40%
C10	44	35,197	89	43,622	51	38,370	65	16,009	249	133,198	0.37%
C11	41	36,693	95	45,108	39	39,732	72	16,801	247	138,334	0.36%
C12	45	39,506	104	50,085	46	42,536	91	18,598	286	150,725	0.38%
C13	36	33,799	71	40,309	39	36,241	70	14,450	216	124,799	0.35%
C14	42	38,199	104	47,832	42	41,245	91	17,807	279	145,083	0.38%
C15	49	36,106	90	43,129	41	38,432	73	16,154	253	133,821	0.38%
C16	38	33,633	66	40,023	41	36,431	52	14,502	197	124,589	0.32%
										平均值	0.38%

900

图9

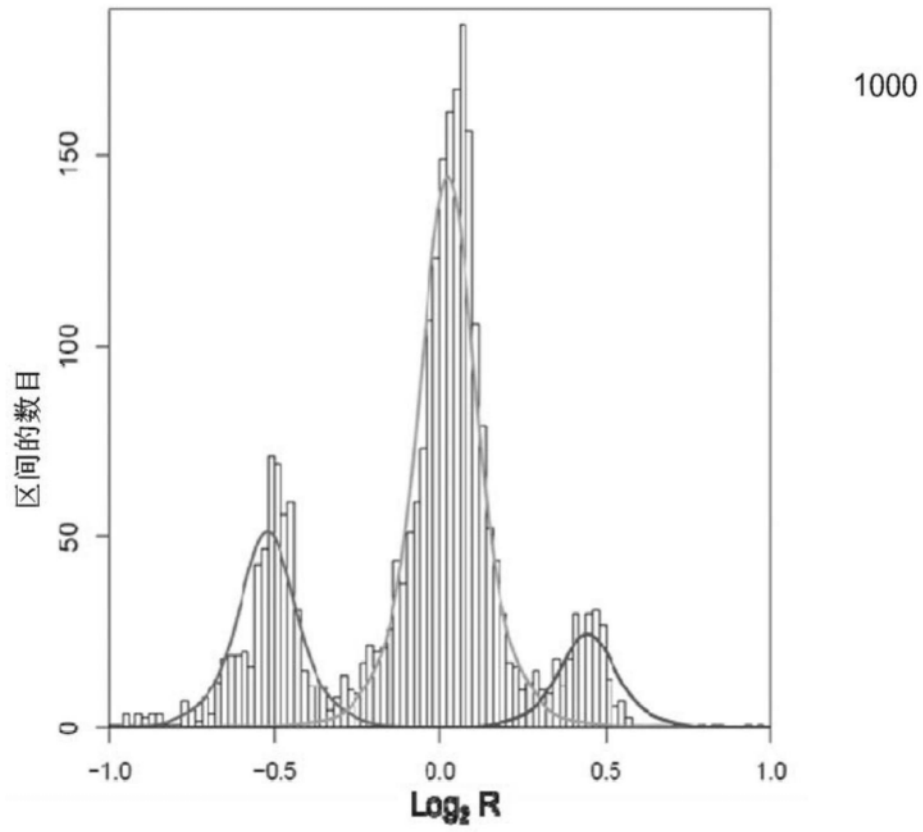


图10A

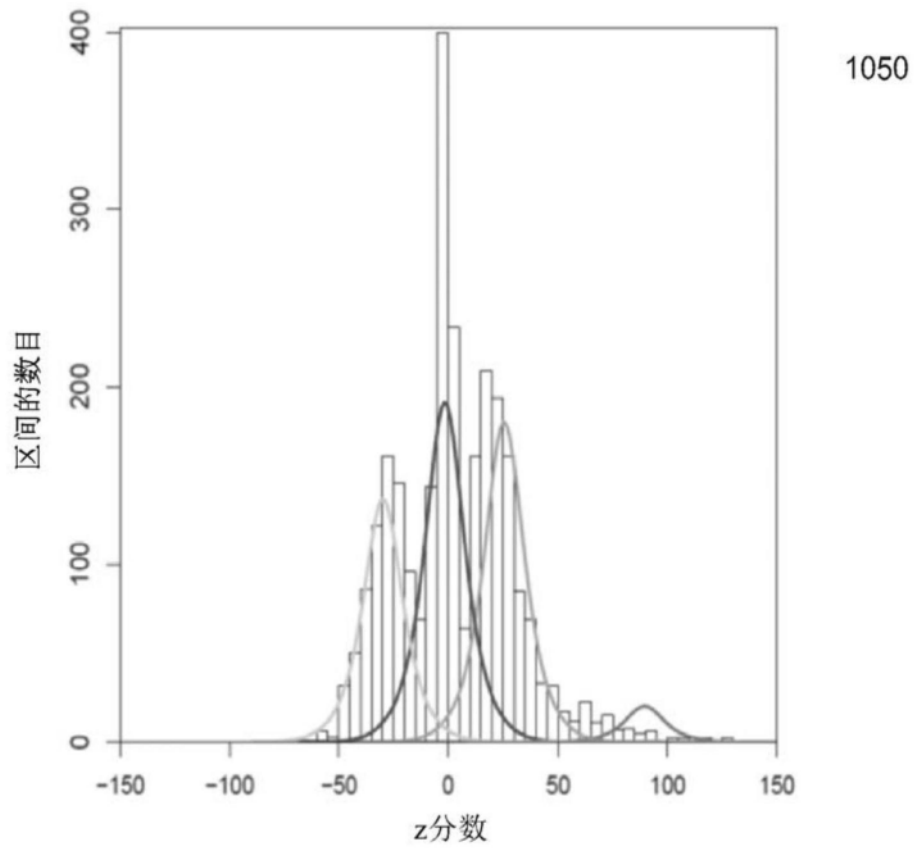


图10B

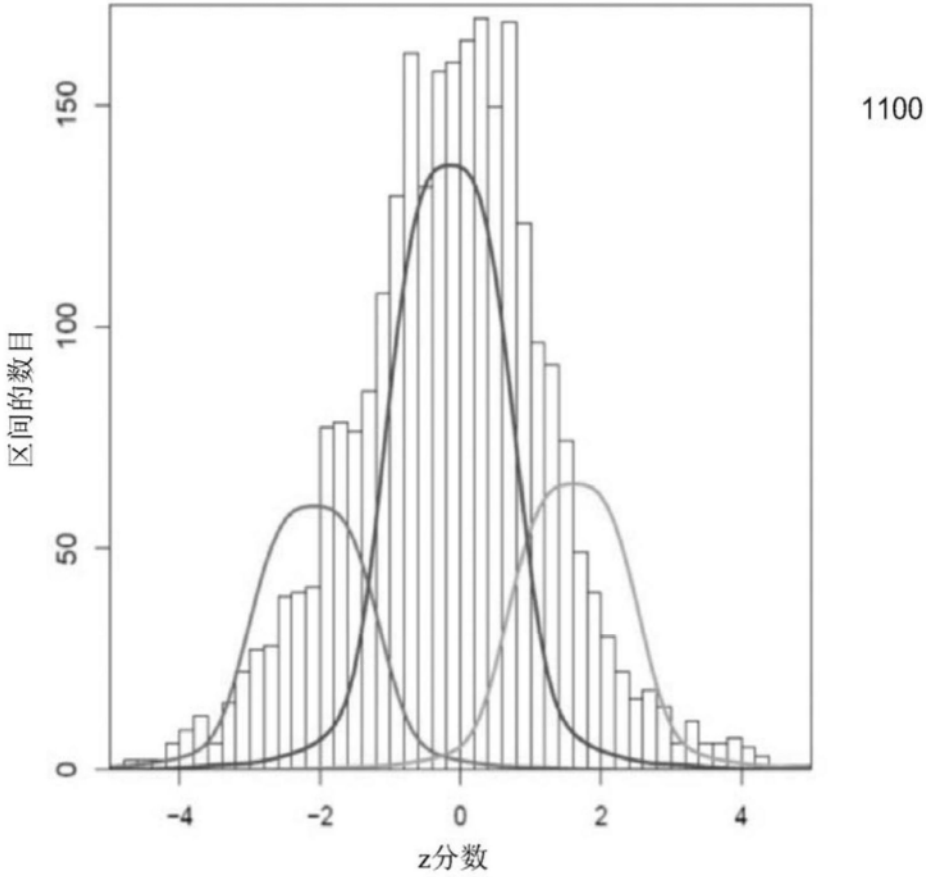


图11

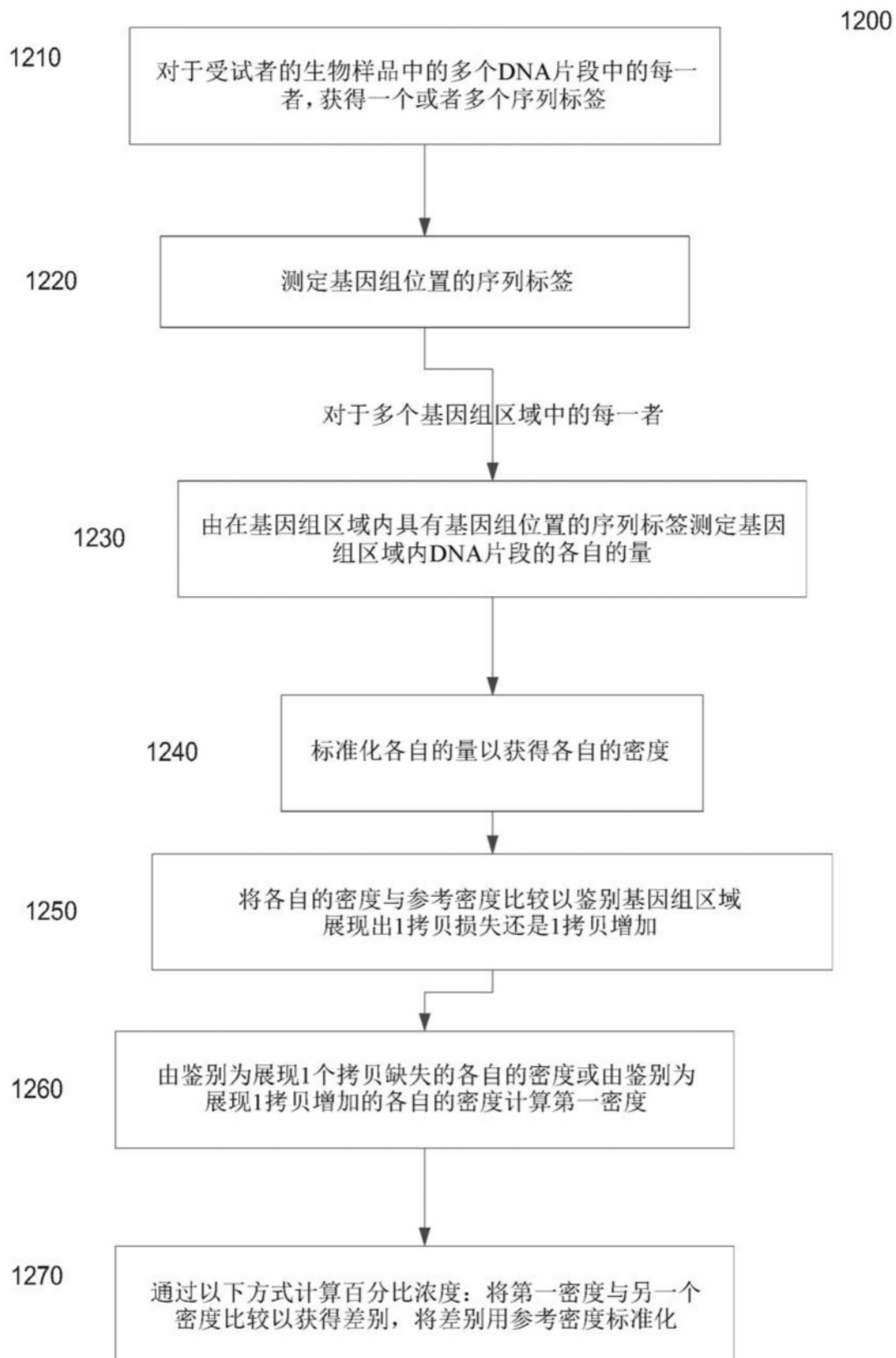


图12

血浆中检测到的次数 (N)	血浆中检测到N次的潜在突变的数目	血浆中检测到 $\geq N$ 次的潜在突变的数目	血浆中检测到N次的肿瘤突变	血浆中检测到 $\geq N$ 次的肿瘤突变	检测到的突变与N=1相比时的百分比	阳性预测值
1	3,605,085	3,623,970	61	2,064	100%	0.06%
2	13,518	18,885	35	2,003	97%	11%
3	2,029	5,367	71	1,968	95%	37%
4	766	3,338	83	1,897	92%	57%
5	437	2,572	140	1,814	88%	71%
6	313	2,135	161	1,674	81%	78%
7	265	1,822	165	1,513	73%	83%
8	261	1,557	192	1,348	65%	87%
9	231	1,296	185	1,156	56%	89%
10	191	1,065	160	971	47%	91%
11	179	874	165	811	39%	93%
12	148	695	137	646	31%	93%
13	122	547	109	509	25%	93%
14	99	425	92	400	19%	94%
15	74	326	71	308	15%	94%
16	54	252	49	237	11%	94%
17	39	198	39	188	9%	95%
18	35	159	33	149	7%	94%
19	24	124	23	116	6%	94%
20	20	100	20	93	5%	93%
21	18	80	17	73	4%	91%
22	11	62	10	56	3%	90%
23	9	51	9	46	2%	90%
24	8	42	5	37	2%	88%
25	6	34	6	32	2%	94%
26	7	28	5	26	1%	93%
27	8	21	8	21	1%	100%
28	1	13	1	13	1%	100%
29	4	12	4	12	1%	100%
30	8	8	8	8	0%	100%

1300

图13A

血浆中检测到的次数 (N)	血浆中检测到N次的潜在突变的数目	血浆中检测到 $\geq N$ 次的潜在突变的数目	血浆中检测到N次的肿瘤突变	血浆中检测到 $\geq N$ 次的肿瘤突变	检测到的突变与N=1相比时的百分比	阳性预测值
1	3,155,634	3,167,144	51	54	100%	<0.01%
2	9,835	11,510	2	3	6%	0.03%
3	1,123	1,675	1	1	2%	0.06%
4	314	552	-	-	-	-
5	132	238	-	-	-	-
6	49	106	-	-	-	-
7	22	57	-	-	-	-
8	11	35	-	-	-	-
9	6	24	-	-	-	-
10	5	18	-	-	-	-
11	4	13	-	-	-	-
12	2	9	-	-	-	-
13	2	7	-	-	-	-
14	3	5	-	-	-	-
15	-	2	-	-	-	-
16	-	2	-	-	-	-
17	1	2	-	-	-	-
18	-	1	-	-	-	-
19	1	1	-	-	-	-
20	-	-	-	-	-	-
21	-	-	-	-	-	-
22	-	-	-	-	-	-
23	-	-	-	-	-	-
24	-	-	-	-	-	-
25	-	-	-	-	-	-
26	-	-	-	-	-	-
27	-	-	-	-	-	-
28	-	-	-	-	-	-
29	-	-	-	-	-	-
30	-	-	-	-	-	-

1300

图13B

测序深度	治疗前血浆			治疗后血浆		
	规定测序深度下 核苷酸的百分比	血浆中检测到单核 苷酸变异的核苷酸 位置的总数目	肿瘤中具有相 应变化的位点 的数目	规定测序深度下 核苷酸的百分比	血浆中检测到单核 苷酸变异的核苷酸 位置的总数目	肿瘤中具有相 应变化的位点 的数目
<50	36.3%	9	8	30.5%	0	0
50 – 125	29.9%	15	15	31.2%	0	0
126 – 235	18.6%	21	21	20.8%	0	0
236 – 380	9.4%	7	6	10.9%	0	0
381 – 560	4.2%	2	2	4.8%	0	0
561 – 760	1.6%	3	3	1.7%	0	0
总计		57	55		0	0

图14A

测序深度	治疗前血浆			治疗后血浆		
	规定测序深度下核苷酸的百分比	血浆中检测到单核苷酸变异位置的总数目	肿瘤中具有相应变化的位点的数目	规定测序深度下核苷酸的百分比	血浆中检测到单核苷酸变异位置的总数目	肿瘤中具有相应变化的位点的数目
<50	30.0%	0	0	27.1%	0	0
50 – 125	33.6%	2	2	32.5%	0	0
126 – 235	21.6%	6	6	22.2%	0	0
236 – 380	10.2%	5	5	11.5%	0	0
381 – 560	3.7%	4	4	4.9%	0	0
561 – 760	0.9%	1	1	1.7%	0	0
总计		18	18		0	0

1450

图14B

测序深度	治疗前血浆			治疗后血浆		
	规定测序深度下 核苷酸的百分比	血浆中检测到单核苷酸 变异的核苷酸位置 的数目	肿瘤中具有相 应变化的位点 的数目	规定测序深度下 核苷酸的百分比	血浆中检测到单核苷酸 变异的核苷酸位置 的数目	肿瘤中具有相 应变化的位点 的数目
<50	38.8%	0	0	18.5%	0	0
50 – 125	45.7%	0	0	28.6%	0	0
126 – 235	15.2%	0	0	25.1%	0	0
236 – 380	0.2%	0	0	16.2%	0	0
381 – 560	<0.1%	0	0	8.3%	0	0
561 – 760	<0.1%	0	0	3.4%	0	0
总计		0	0		0	0

1500

图15A

测序深度	治疗前血浆			治疗后血浆		
	规定测序深度下 核苷酸的百分比	血浆中检测到单核 苷酸变异的核苷酸 位置的总数目	肿瘤中具有相 应变化的位点 的数目	规定测序深度下 核苷酸的百分比	血浆中检测到 单核苷酸变异 的核苷酸位置 的总数目	肿瘤中具有相 应变化的位点 的数目
<50	26.7%	0	0	24.4%	0	0
50 – 125	35.9%	1	1	31.2%	1	1
126 – 235	24.7%	2	2	23.1%	0	0
236 – 380	10.4%	0	0	13.1%	0	0
381 – 560	2.2%	0	0	6.0%	0	0
561 – 760	0.1%	0	0	2.2%	0	0
总计		3	3		1	1

1550

图15B

测序深度	治疗前血浆			治疗后血浆		
	规定测序深度下核苷酸的百分比	血浆中检测到单核苷酸变异的核苷酸位置的总数目	肿瘤中具有相应变化的位点的数目	规定测序深度下核苷酸的百分比	血浆中检测到单核苷酸变异的核苷酸位置的总数目	肿瘤中具有相应变化的位点的数目
<50	31.1%	7	7	30.8%	0	0
50 – 125	30.3%	14	14	33.2%	1	0
126 – 235	20.5%	19	19	21.6%	0	0
236 – 380	11.3%	13	13	10.3%	0	0
381 – 560	5.2%	7	4	3.5%	0	0
561 – 760	1.8%	4	2	0.7%	0	0
总计		64	59		1	0
1600						

图16

测序深度 (倍数)	当突变体界定为见于肿瘤组织测序数据中至少n次时的敏感性, 其中n=					
	4	5	6	7	8	9
20	73%	56%	38%	24%	13%	7%
21	77%	60%	43%	28%	16%	9%
22	80%	64%	47%	31%	19%	11%
23	83%	68%	51%	35%	22%	13%
24	85%	71%	55%	39%	26%	15%
25	87%	75%	59%	43%	29%	18%
26	89%	78%	63%	47%	33%	21%
27	90%	80%	67%	51%	36%	24%
28	92%	83%	70%	55%	40%	27%
29	93%	85%	73%	59%	44%	30%
30	94%	87%	76%	62%	48%	34%
31	95%	89%	78%	66%	51%	37%
32	96%	90%	81%	69%	55%	41%
33	96%	91%	83%	72%	58%	44%
34	97%	93%	85%	74%	61%	48%
35	97%	94%	87%	77%	65%	51%
36	98%	95%	88%	79%	68%	54%
37	98%	95%	90%	82%	70%	58%
38	99%	96%	91%	84%	73%	61%
39	99%	97%	92%	85%	76%	64%
40	99%	97%	93%	87%	78%	67%
41	99%	98%	94%	88%	80%	69%
42	99%	98%	95%	90%	82%	72%
43	99%	98%	96%	91%	84%	75%
44	100%	98%	96%	92%	86%	77%
45	100%	99%	97%	93%	87%	79%
46	100%	99%	97%	94%	89%	81%
47	100%	99%	98%	95%	90%	83%
48	100%	99%	98%	95%	91%	84%
49	100%	99%	98%	96%	92%	86%
50	100%	99%	99%	97%	93%	88%
51	100%	100%	99%	97%	94%	89%
52	100%	100%	99%	97%	95%	90%
53	100%	100%	99%	98%	95%	91%
54	100%	100%	99%	98%	96%	92%
55	100%	100%	99%	98%	96%	93%
56	100%	100%	99%	99%	97%	94%
57	100%	100%	100%	99%	97%	95%
58	100%	100%	100%	99%	98%	95%
59	100%	100%	100%	99%	98%	96%
60	100%	100%	100%	99%	98%	96%
61	100%	100%	100%	99%	98%	97%
62	100%	100%	100%	99%	99%	97%
63	100%	100%	100%	100%	99%	97%
64	100%	100%	100%	100%	99%	98%
65	100%	100%	100%	100%	99%	98%
66	100%	100%	100%	100%	99%	98%
67	100%	100%	100%	100%	99%	99%
68	100%	100%	100%	100%	99%	99%
69	100%	100%	100%	100%	100%	99%
70	100%	100%	100%	100%	100%	99%

1700

图17

测序深度 (倍数)	当突变体界定为见于肿瘤测序数据中至少n次时的全基因组中的假阳性的数目, 其中n=						
	4	5	6	7	8	9	
20	1	0	0	0	0	0	1800
21	1	0	0	0	0	0	
22	1	0	0	0	0	0	
23	1	0	0	0	0	0	
24	1	0	0	0	0	0	
25	1	0	0	0	0	0	
26	2	0	0	0	0	0	
27	2	0	0	0	0	0	
28	2	0	0	0	0	0	
29	3	0	0	0	0	0	
30	3	0	0	0	0	0	
31	3	0	0	0	0	0	
32	4	0	0	0	0	0	
33	5	0	0	0	0	0	
34	5	0	0	0	0	0	
35	6	0	0	0	0	0	
36	7	0	0	0	0	0	
37	7	0	0	0	0	0	
38	8	0	0	0	0	0	
39	9	0	0	0	0	0	
40	10	0	0	0	0	0	
41	11	0	0	0	0	0	
42	12	0	0	0	0	0	
43	14	0	0	0	0	0	
44	15	0	0	0	0	0	
45	17	0	0	0	0	0	
46	18	0	0	0	0	0	
47	20	0	0	0	0	0	
48	22	0	0	0	0	0	
49	24	0	0	0	0	0	
50	26	0	0	0	0	0	
51	28	0	0	0	0	0	
52	30	0	0	0	0	0	
53	33	0	0	0	0	0	
54	35	0	0	0	0	0	
55	38	0	0	0	0	0	
56	41	0	0	0	0	0	
57	44	0	0	0	0	0	
58	47	0	0	0	0	0	
59	51	0	0	0	0	0	
60	54	0	0	0	0	0	
61	58	0	0	0	0	0	
62	62	0	0	0	0	0	
63	66	0	0	0	0	0	
64	71	0	0	0	0	0	
65	75	0	0	0	0	0	
66	80	0	0	0	0	0	
67	85	0	0	0	0	0	
68	90	0	0	0	0	0	
69	96	0	0	0	0	0	
70	102	0	0	0	0	0	

图18

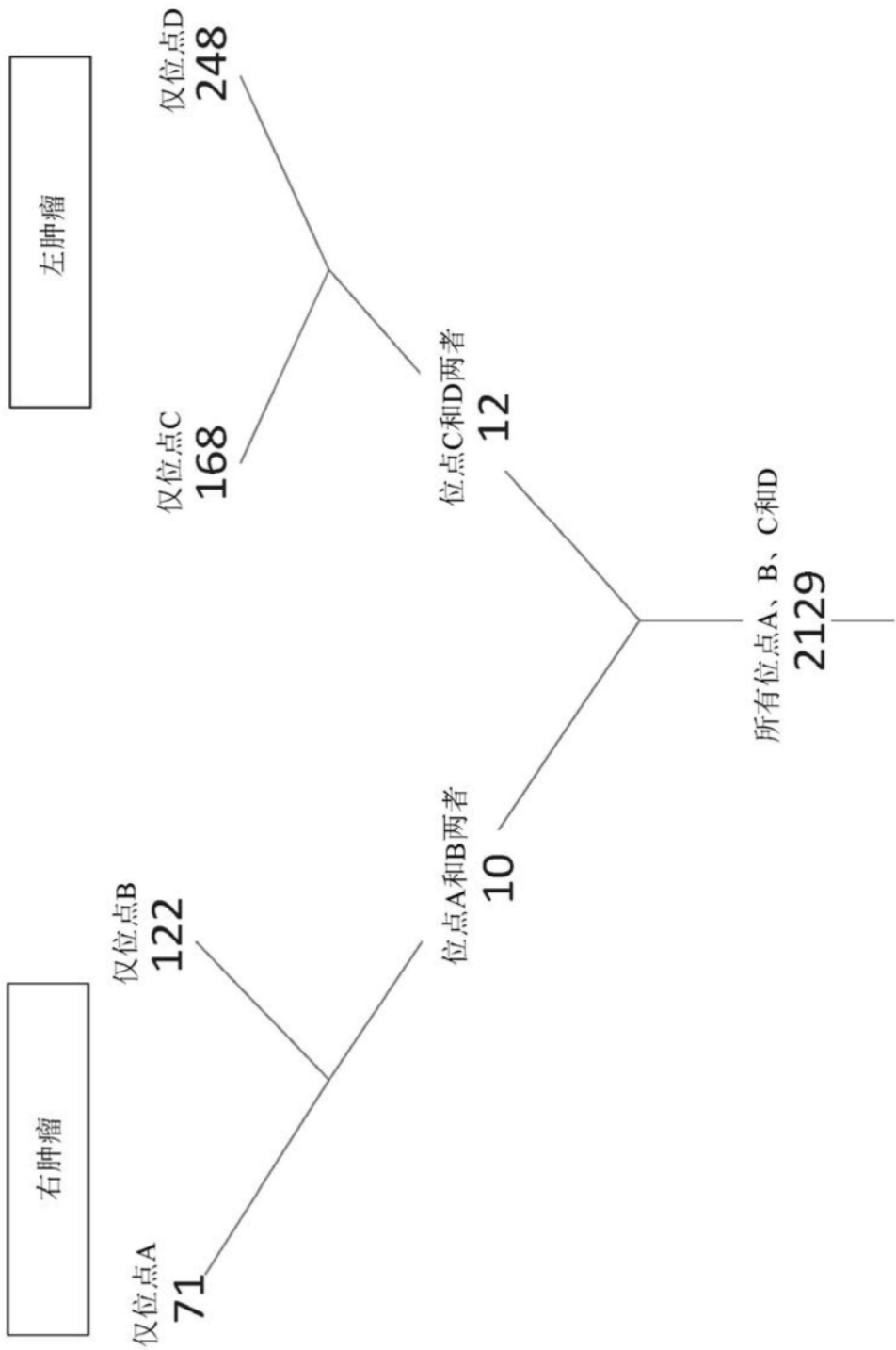


图19

类别	具有突变的基因座的数目	治疗前血浆			治疗后血浆		
		携带野生型等位基因的DNA片段的数目	携带突变等位基因的DNA片段的数目	肿瘤来源的DNA的百分比浓度	携带野生型等位基因的DNA片段的数目	携带突变等位基因的DNA片段的数目	肿瘤来源的DNA的百分比浓度
A	71	3,321	37	2.20%	2,149	5	0.46%
B	122	5,633	8	0.28%	3,516	1	0.06%
C	168	8,507	51	1.19%	5,438	2	0.07%
D	248	10,880	26	0.48%	7,060	2	0.06%
AB	10	423	21	9.46%	297	0	0.00%
CD	12	569	3	1.05%	333	0	0.00%
ABCD	2129	70,940	21,344	46.26%	61,417	54	0.18%

2000

图20



图21

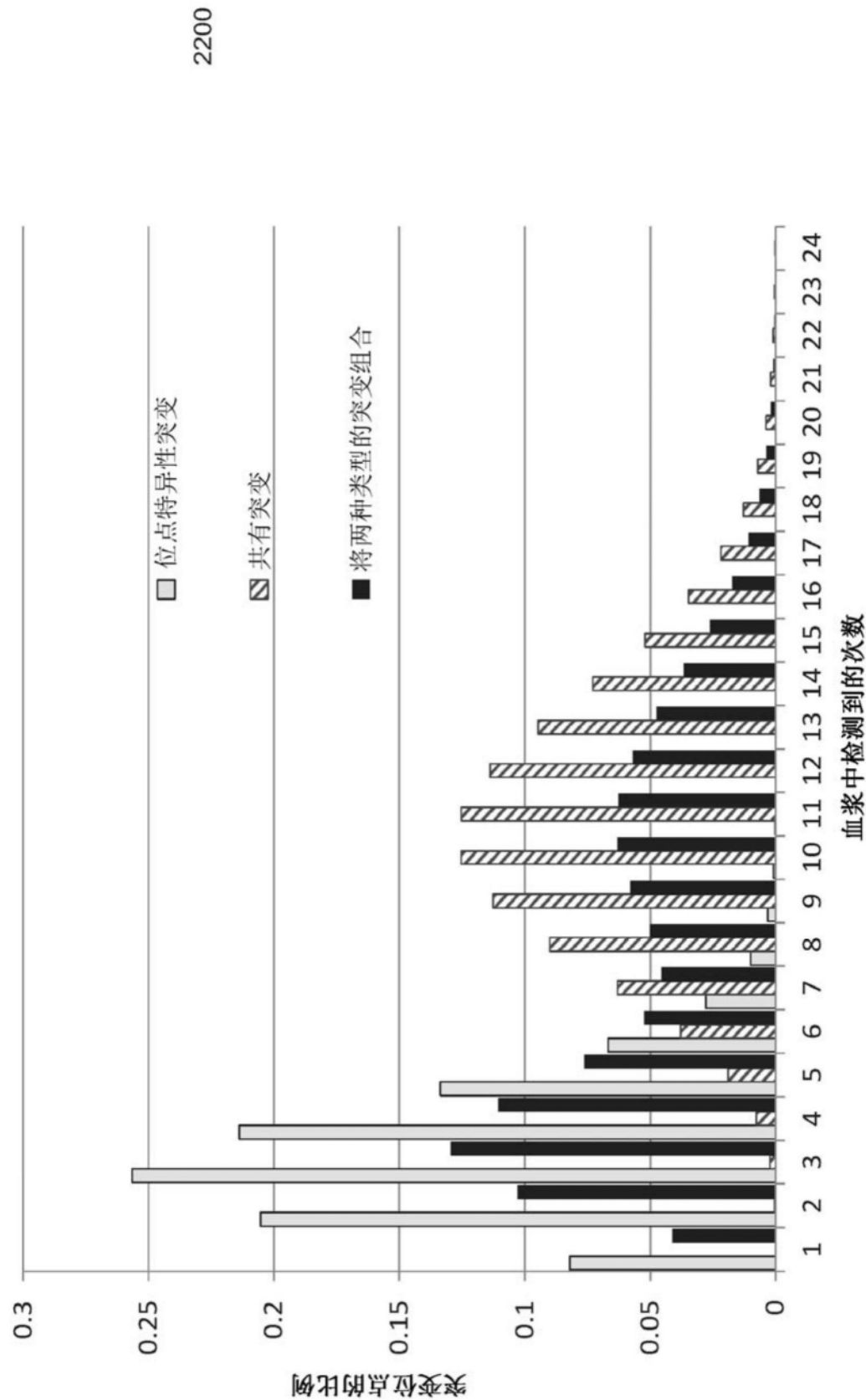


图22

类别	基因座的数目	健康对照受试者															
		TC1	TC2	TC3	TC4	TC5	TC ₆	TC7	TC8	TC9	TC10	TC11	TC12	TC13	TC14	TC15	TC16
		血浆中的突变等位基因的数目（百分比浓度）															
A	71	0	$\frac{1}{(0.17\%)}$	$\frac{1}{(0.18\%)}$	0	$\frac{1}{(0.17\%)}$	$\frac{1}{(0.18\%)}$	$\frac{1}{(0.07\%)}$	$\frac{1}{(0.20\%)}$	0	$\frac{6}{(0.17\%)}$	0	0	$\frac{1}{(0.17\%)}$	$\frac{2}{(0.38\%)}$	0	$\frac{2}{(0.38\%)}$
B	122	0	0	0	0	0	0	$\frac{1}{(0.04\%)}$	0	$\frac{1}{(0.11\%)}$	$\frac{3}{(0.05\%)}$	$\frac{1}{(0.10\%)}$	$\frac{1}{(0.12\%)}$	0	0	0	$\frac{1}{(0.11\%)}$
C	168	0	0	0	0	0	0	$\frac{2}{(0.06\%)}$	0	$\frac{2}{(0.16\%)}$	0	0	0	$\frac{1}{(0.07\%)}$	0	0	0
D	248	0	0	$\frac{1}{(0.06\%)}$	0	0	0	0	0	0	$\frac{1}{(0.01\%)}$	0	0	$\frac{1}{(0.05\%)}$	0	0	0
AB	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CD	12	0	0	0	0	0	0	0	0	0	$\frac{1}{(0.16\%)}$	0	0	$\frac{1}{(0.92\%)}$	0	0	0
ABCD	$\frac{212}{9}$	$\frac{11}{(0.07\%)}$	$\frac{3}{(0.02\%)}$	$\frac{2}{(0.01\%)}$	$\frac{8}{(0.06\%)}$	$\frac{5}{(0.03\%)}$	$\frac{5}{(0.03\%)}$	$\frac{10}{(0.02\%)}$	$\frac{4}{(0.03\%)}$	$\frac{4}{(0.03\%)}$	$\frac{19}{(0.02\%)}$	$\frac{3}{(0.02\%)}$	0	$\frac{1}{(0.01\%)}$	$\frac{5}{(0.03\%)}$	$\frac{2}{(0.01\%)}$	$\frac{4}{(0.03\%)}$

2300

图23

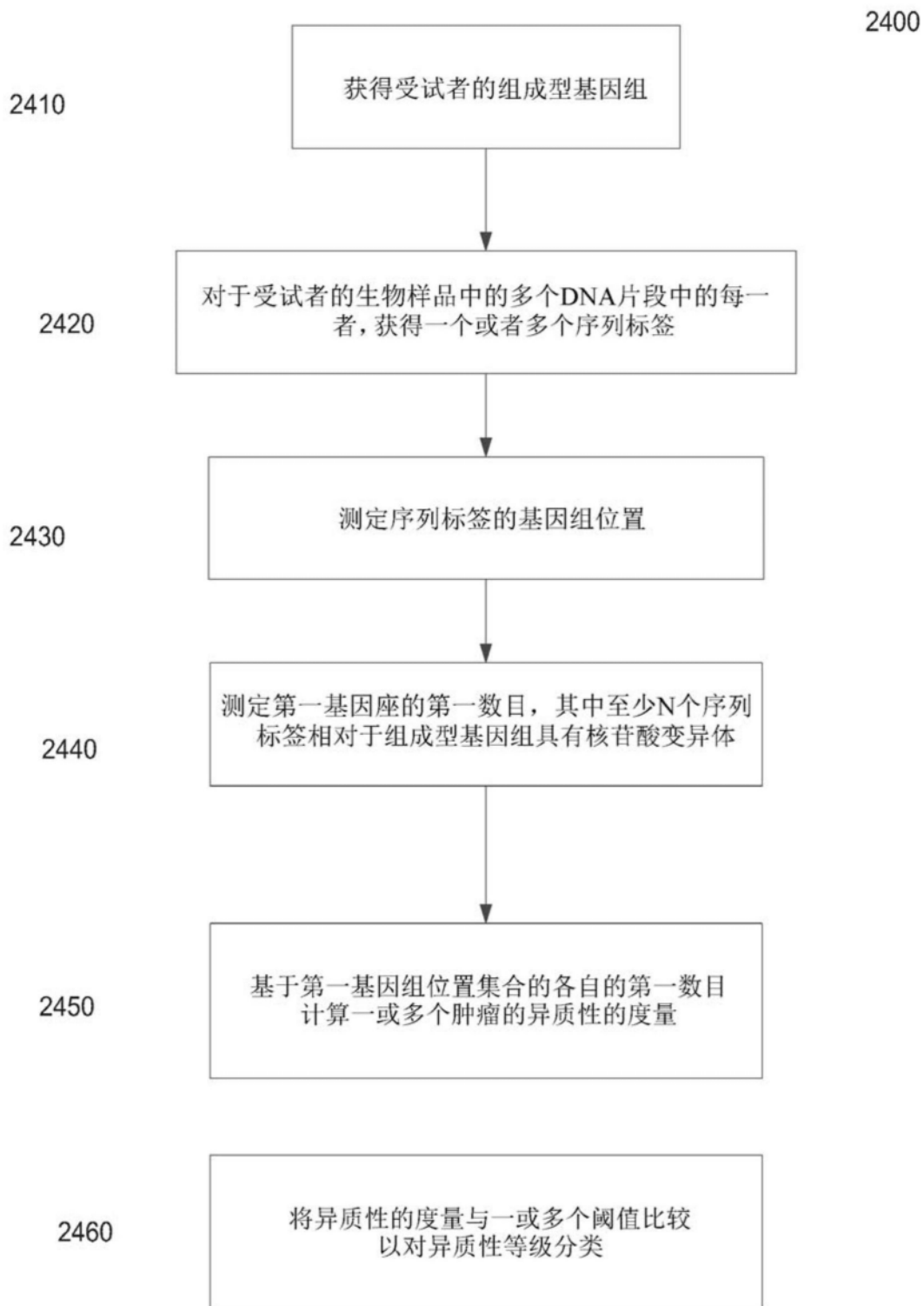


图24

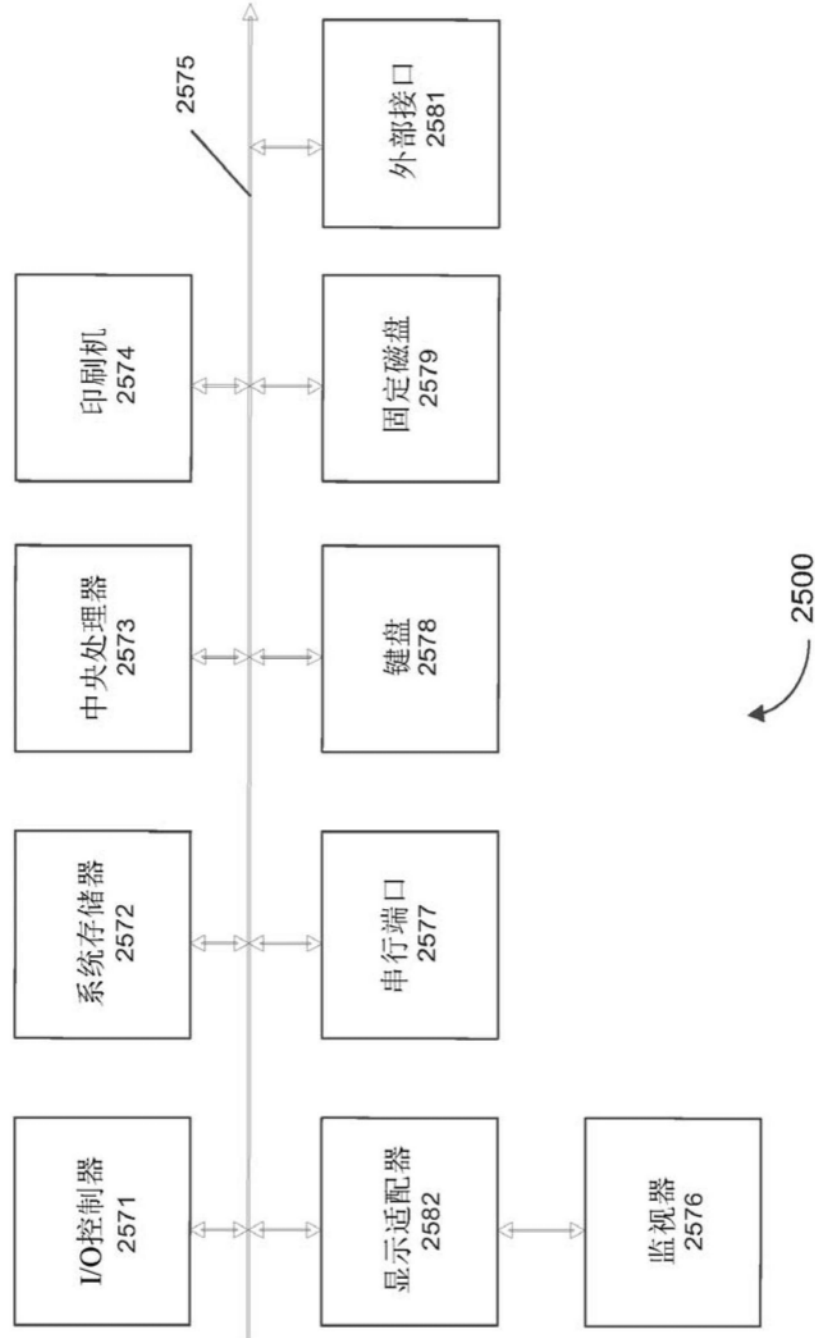


图25