



(12) 发明专利申请

(10) 申请公布号 CN 113710818 A

(43) 申请公布日 2021. 11. 26

(21) 申请号 202080027120.4

(22) 申请日 2020.04.01

(30) 优先权数据

62/828,224 2019.04.02 US

62/961,517 2020.01.15 US

(85) PCT国际申请进入国家阶段日

2021.10.08

(86) PCT国际申请的申请数据

PCT/US2020/026269 2020.04.01

(87) PCT国际申请的公布数据

W02020/206041 EN 2020.10.08

(71) 申请人 格里尔公司

地址 美国加利福尼亚州门洛帕克市

(72) 发明人 卢煜明 赵慧君 陈君赐 江培勇

林伟棋 吉璐

(74) 专利代理机构 上海翼胜专利商标事务所

(普通合伙) 31218

代理人 翟羽

(51) Int.Cl.

C12Q 1/6886 (2018.01)

C12Q 1/6883 (2018.01)

C12Q 1/70 (2006.01)

G16H 50/30 (2018.01)

权利要求书7页 说明书55页 附图22页

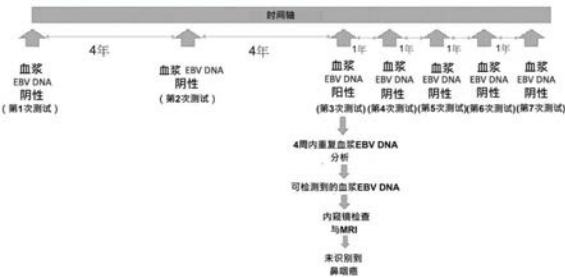
(54) 发明名称

病毒相关联的癌症风险分层

(57) 摘要

本文提供了基于对来自一对象的一生物样本的多个无细胞核酸分子的分析来对所述对象发展出一病原体相关联的疾病的风险进行分层的多种方法和多种系统。在各种示例中,筛选频率是基于风险分析来确定。本文还提供了用于分析多个无细胞核酸分子中一病原体基因组的多个变异模式的多种方法和多种系统。

基于血浆EBV DNA状态的鼻咽癌筛选的可变间隔的说明性示例



1. 一种筛选一对象中一病原体相关联的疾病的方法,其特征在于,所述方法包含:

接收来自一第一时间点执行的一第一分析的数据,所述第一分析包含确定来自所述对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征,其中来自所述病原体的所述多个无细胞核酸分子的所述特征包含与来自所述生物样本中的所述对象的所述多个无细胞核酸分子相比的一数量、一甲基化状态、一变异模式、一片段大小、或一相对丰度,及其中所述特征表明所述对象发展出所述病原体相关联的疾病的一风险;以及

基于所述特征确定在一第二时间点处被执行的一第二分析,以筛选所述对象中的所述病原体相关联的疾病,其中所述第一时间点和所述第二时间点之间的一间隔与所述风险为负相关。

2. 一种预测一对象中一病原体相关联的疾病的方法,其特征在于,所述方法包含:

接收来自一第一分析的数据,所述第一分析包含确定来自所述对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征,其中来自所述病原体的所述多个无细胞核酸分子的所述特征包含与来自所述生物样本中的所述对象的所述多个无细胞核酸分子相比的一数量、一甲基化状态、一变异模式、一片段大小、或一相对丰度;以及

基于来自所述病原体的所述多个无细胞核酸分子的所述特征以及下述其中一个或多个因素:所述对象的年龄、所述对象的吸烟习惯、所述对象的病原体相关联的疾病的家族史、所述对象的多个基因型因素、所述对象的种族或所述对象的饮食史,来生成表明所述对象发展出病原体相关联的疾病的一风险的一报道。

3. 如权利要求1所述的方法,其特征在于,所述第一分析的结果不会导致针对所述病原体相关联的疾病对所述对象进行一医学治疗。

4. 如权利要求3所述的方法,其特征在于,所述医学治疗包含使用多个治疗剂治疗、放射治疗或外科治疗。

5. 如权利要求1、3或4所述的方法,其特征在于,在通过假阳性率低于1%的一临床诊断检查来确定一第二时间点之前,将所述对象诊断为没有患有所述病原体相关联的疾病。

6. 如权利要求5所述的方法,其特征在于,所述临床诊断检查包含物理检查、侵入性活检、内窥镜检查、磁共振成像、正电子发射断层扫描、计算机断层扫描或x射线成像。

7. 如权利要求5所述的方法,其特征在于,所述临床诊断检查包含一侵入性活检,所述侵入性活检包含组织学分析、细胞学分析或细胞核酸分析。

8. 如权利要求1或3至7任一项所述的方法,其特征在于,所述间隔为至少约为2个月、4个月、6个月、8个月、10个月或12个月。

9. 如权利要求8所述的方法,其特征在于,所述间隔为至少约为12个月。

10. 如权利要求1至9任一项所述的方法,其特征在于,所述方法还包含执行所述第一分析。

11. 如权利要求10所述的方法,其特征在于,所述执行所述第一分析的步骤包含:

(i) 从所述对象获得一第一生物样本;以及

(ii) 在所述第一生物样本中测量来自所述病原体的多个无细胞核酸分子的一第一数量。

12. 如权利要求11所述的方法,其特征在于,所述测量所述第一数量的步骤包含:测量所述第一生物样本中来自所述病原体的所述多个无细胞核酸分子的一拷贝数。

13. 如权利要求11或12所述的方法,其特征在于,所述测量包含聚合酶链反应(PCR)。

14. 如权利要求11或12所述的方法,其特征在于,所述测量包含定量聚合酶链反应(qPCR)。

15. 如权利要求11所述的方法,其特征在于,所述第一数量包含:测量所述第一生物样本中来自所述病原体的所述多个无细胞核酸分子的第一百分比。

16. 如权利要求11至15任一项所述的方法,其特征在于,所述第一分析还包含步骤:

(iii) 如果所述第一数量高于一阈值,则从所述对象获得一第二生物样本,及在所述第二生物样本中从所述病原体中测量多个无细胞核酸分子的第二数量。

17. 如权利要求16所述的方法,其特征在于,所述第二生物样本在所述第一生物样本之后约4周获得。

18. 如权利要求16或17所述的方法,其特征在于,与如果所述第二数量低于所述阈值的一间隔相比,如果所述第一数量和第二拷贝数都高于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔更短。

19. 如权利要求16至18任一项所述的方法,其特征在于,与如果所述第一数量高于所述阈值的一间隔相比,如果所述第一数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔更长。

20. 如权利要求16至19任一项所述的方法,其特征在于,如果所述第一数量和所述第二数量均高于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约1年。

21. 如权利要求16至20任一项所述的方法,其特征在于,如果所述第二数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约2年。

22. 如权利要求16至21任一项所述的方法,其特征在于,如果所述第一数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约4年。

23. 如权利要求10所述的方法,其特征在于,所述第一分析包含步骤:

在所述生物样本中确定来自所述病原体的多个无细胞核酸分子的一甲基化状态。

24. 如权利要求23所述的方法,其特征在于,所述确定所述甲基化状态的步骤包含:使用一甲基化敏感限制酶或亚硫酸氢盐处理所述生物样本中的所述多个无细胞核酸分子。

25. 如权利要求23所述的方法,其特征在于,所述确定所述甲基化状态的步骤包含:对所述对象的所述生物样本中的所述多个无细胞核酸执行一可识别甲基化测序。

26. 如权利要求25所述的方法,其特征在于,所述可识别甲基化测序包含未甲基化胞嘧啶转化到尿嘧啶的亚硫酸氢盐转化。

27. 如权利要求25所述的方法,其特征在于,所述可识别甲基化测序包含:使用一甲基化敏感限制酶进行处理。

28. 如权利要求10所述的方法,其特征在于,所述第一分析包含步骤:

在所述生物样本中确定来自所述病原体的多个无细胞核酸分子的一片段大小分布。

29. 如权利要求28所述的方法,其特征在于,所述确定所述片段大小分布的步骤包含:对所述生物样本中的多个无细胞核酸分子进行测序,以及基于映射到所述病原体的一参考基因组的多个序列读数确定来自所述生物样本中所述病原体的所述多个无细胞核酸分子的一片段大小。

30. 如权利要求10所述的方法,其特征在于,所述第一分析包含步骤:

在所述生物样本中确定来自所述病原体的所述多个无细胞核酸分子的一变异模式。

31. 如权利要求30所述的方法, 其特征在于, 所述确定所述变异模式的步骤包含: 对所述生物样本中的多个无细胞核酸分子进行测序, 以及基于映射到所述病原体的一参考基因组的多个序列读数以确定来自所述生物样本中所述病原体的所述多个无细胞核酸分子的所述变异模式。

32. 如权利要求30或31所述的方法, 其特征在于, 来自所述病原体的所述多个无细胞核酸分子的所述变异模式包含多个单核苷酸变异。

33. 如权利要求32所述的方法, 其特征在于, 所述识别所述变异模式的步骤包含:

确定介于映射到所述病原体的所述参考基因组的多个序列读数与所述病原体的一疾病相关的参考基因组之间的一相似性水平。

34. 如权利要求33所述的方法, 其特征在于, 所述病原体的疾病相关参考基因组包含在一患病组织中被识别的所述病原体的一基因组。

35. 如权利要求33或34所述的方法, 其特征在于, 所述确定相似性水平的步骤包含:

将所述病原体的所述参考基因组分离到多个箱中; 以及

确定对于所述多个箱中的每个箱相对于所述病原体的所述疾病相关的参考基因组的一相似性指数, 其中所述相似性指数与相应箱内的多个变异位点的一比例相关, 所述相应箱中映射到所述病原体的所述参考基因组的所述多个序列读数中的至少一个与所述病原体的所述疾病相关的参考基因组具有一相同的核苷酸变异。

36. 如权利要求35所述的方法, 其特征在于, 所述病原体的所述疾病相关的参考基因组包含所述病原体的多个疾病相关的参考基因组, 及其中所述确定相似性水平的步骤包含:

确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数; 及

基于所述多个疾病相关的参考基因组的一比例来确定所述多个箱中的每一个箱的一箱分数, 相对于所述箱分数, 所述多个箱内的所述相应相似性指数是高于一截止值。

37. 如权利要求35或36所述的方法, 其特征在于, 所述多个箱中的每一个具有一长度为约100、200、300、400、500、600、700、800、900或1000个碱基对。

38. 如权利要求10至37任一项所述的方法, 所述第一分析包含步骤包含: 确定所述生物样本中来自所述病原体的所述多个无细胞核酸分子的所述甲基化状态、所述片段大小分布或所述变异模式。

39. 如先前的权利要求任一项所述的方法, 所述方法进一步包含: 使用应用于一数据输入的一分类器来计算对于所述对象发展出所述病原体相关联的疾病的一风险分数, 所述数据输入包含所述生物样本中来自所述病原体的所述多个无细胞核酸分子的所述特征, 其中所述分类器被配置为对所述数据输入应用一函数以生成一输出, 所述数据输入包含来自所述生物样本中所述病原体的所述多个无细胞核酸分子的所述特征, 所述输出包含所述风险分数, 所述风险分数评估所述对象发展出疾病的风险。

40. 如权利要求39所述的方法, 其特征在于, 所述分类器使用一标记的数据集进行训练。

41. 如权利要求1所述的方法, 其特征在于, 所述方法进一步包含在所述第二时间点执行所述第二分析。

42. 如权利要求41所述的方法,其特征在于,所述第二分析与所述第一分析相同。

43. 如权利要求41所述的方法,其特征在于,所述第二分析包含来自所述对象的所述多个无细胞核酸分子的一分析、所述对象的侵入性活检、所述对象的内窥镜检查,或所述对象的一磁共振成像检查。

44. 一种分析来自一对象的一生物样本的多个核酸分子的方法,其特征在于,所述方法包含:

在一计算机系统中,从所述对象的所述生物样本获得多个无细胞核酸分子的多个序列读数,其中所述生物样本包含来自所述对象和潜在地来自一病原体的多个无细胞核酸分子;

在所述计算机系统中,将所述多个无细胞核酸分子的所述多个序列读数与所述病原体的一参考基因组比对;及

在所述计算机系统中,识别来自所述病原体的所述多个无细胞核酸分子的一变异模式,其中所述变异模式表征在所述病原体的所述参考基因组上的多个变异位点中的每一个处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,其中所述多个变异位点包含横跨所述病原体的所述参考基因组的至少30个位点,及所述变异模式表明所述对象中所述病原体相关联的疾病的一状态或一风险。

45. 如权利要求44所述的方法,其特征在于,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少40个、至少50个、至少60个、至少70个、至少80个、至少90个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1000个、至少1100个,或至少1200个位点。

46. 如权利要求44所述的方法,其特征在于,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少600个位点。

47. 如权利要求44所述的方法,其特征在于,所述多个变异位点包含横跨所述病原体的所述参考基因组的约660个位点。

48. 如权利要求44所述的方法,其特征在于,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少1000个位点。

49. 如权利要求44所述的方法,其特征在于,所述多个变异位点包含横跨所述病原体的所述参考基因组的约1100个位点。

50. 如权利要求44所述的方法,其特征在于,所述多个变异位点由所有位点组成,在所述所有位点处,映射到所述病原体的所述参考基因组的所述多个序列读数具有与所述病原体的所述参考基因组的一不同的核苷酸变异。

51. 如权利要求44至50任一项所述的方法,其特征在于,所述比对所述多个序列读数的步骤被配置为允许映射到所述病原体的所述参考基因组的多个序列读数和所述病原体的所述参考基因组之间存在10、9、8、7、6、5、4、3、2或1个碱基的最大不匹配。

52. 如权利要求44至50任一项所述的方法,其特征在于,所述比对所述多个序列读数的步骤被配置为允许映射到所述病原体的所述参考基因组和所述病原体的所述参考基因组之间存在2个碱基的最大不匹配。

53. 如权利要求44至52任一项所述的方法,其特征在于,所述方法进一步包含:

基于映射到所述病原体的所述参考基因组的所述多个序列读数的所述变异模式来诊

断、预测或监测所述对象的所述病原体相关联的疾病。

54. 如权利要求44至53任一项所述的方法,其特征在于,来自所述病原体的所述多个无细胞核酸分子的所述变异模式包含多个单核苷酸变异。

55. 如权利要求44至54任一项所述的方法,其特征在于,所述识别所述变异模式的步骤包含:

确定介于映射到所述病原体的所述参考基因组的多个序列读数与所述病原体的一疾病相关的参考基因组之间的一相似性水平。

56. 如权利要求55所述的方法,其特征在于,所述病原体的疾病相关参考基因组包含在一患病组织中被识别的所述病原体的一基因组。

57. 如权利要求55或56所述的方法,其特征在于,所述确定相似性水平的步骤包含:

将所述病原体的所述参考基因组分离到多个箱中;以及

确定对于所述多个箱中的每个箱相对于所述病原体的所述疾病相关的参考基因组的一相似性指数,其中所述相似性指数与相应箱内的多个变异位点的一比例相关,所述相应箱中映射到所述病原体的所述参考基因组的所述多个序列读数中的至少一个与所述病原体的所述疾病相关的参考基因组具有一相同的核苷酸变异。

58. 如权利要求57所述的方法,其特征在于,所述病原体的所述疾病相关的参考基因组包含所述病原体的多个疾病相关的参考基因组,及其中所述确定相似性水平的步骤包含:

确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数;及

基于所述多个疾病相关的参考基因组的一比例来确定所述多个箱中的每一个箱的一箱分数,相对于所述箱分数,所述多个箱内的所述相应相似性指数是高于一截止值。

59. 如权利要求58所述的方法,其特征在于,其中所述截止值为约0.9。

60. 如权利要求57至59任一项所述的方法,其特征在于,所述多个箱中的每一个具有一长度为约100、200、300、400、500、600、700、800、900或1000个碱基对。

61. 如权利要求44至60任一项所述的方法,其特征在于,所述方法进一步包含:使用应用于一数据输入的一分类器来计算对于所述对象发展出所述病原体相关联的疾病的一风险分数,所述数据输入包含来自所述病原体的所述多个无细胞核酸分子的所述变异模式,其中所述分类器被配置为对所述数据输入应用一函数以生成一输出,所述数据输入包含来自所述病原体的所述多个无细胞核酸分子的所述变异模式,所述输出包含所述风险分数,所述风险分数评估所述对象发展出疾病的风险。

62. 如权利要求61所述的方法,其特征在于,所述分类器使用一标记的数据集进行训练。

63. 如权利要求61或62所述的方法,其特征在于,所述分类器包含一数学模型,所述数学模型使用朴素贝叶斯模型、逻辑回归、随机森林、决策树、梯度提升树、神经网络、深度学习、线性/核支持向量机(SVM)、线性/非线性回归,或线性判别分析。

64. 如权利要求44至63任一项所述的方法,其特征在于,其中所述病原体是一病毒。

65. 如权利要求64所述的方法,其特征在于,其中所述病毒为爱泼斯坦-巴尔病毒(EBV)。

66. 如权利要求65所述的方法,其特征在于,所述病原体相关联的疾病包含鼻咽癌、NK

细胞淋巴瘤、伯基特淋巴瘤、移植后淋巴增生性疾病、或霍奇金淋巴瘤。

67. 如权利要求65或66所述的方法,其特征在于,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点包含从表6所列的多个基因组位点中相对于一EBV参考基因组(AJ507799.2)选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。

68. 如权利要求67所述的方法,其特征在于,所述多个变异位点包含如表6所列相对于一EBV参考基因组(AJ507799.2)的一基因组位点。

69. 如权利要求65或66所述的方法,其特征在于,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点是从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)随机选择的。

70. 如权利要求65或66所述的方法,其特征在于,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点包含从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)随机选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。

71. 如权利要求64所述的方法,其特征在于,其中所述病毒为一人类乳头瘤病毒(HPV)。

72. 如权利要求71所述的方法,其特征在于,所述病原体相关联的疾病包含宫颈癌、口咽癌或头颈癌。

73. 如权利要求64所述的方法,其特征在于,其中所述病毒为一乙型肝炎病毒(HBV)。

74. 如权利要求73所述的方法,其特征在于,所述病原体相关联的疾病包含一肝硬化或一肝细胞癌(HCC)。

75. 如权利要求44至74任一项所述的方法,其特征在于,所述变异模式表明所述对象中所述病原体相关联的疾病的一状态,其中所述病原体相关联的疾病的所述状态包含所述对象中所述病原体相关联的疾病的一存在、所述对象中一肿瘤组织的一数量、所述对象中一肿瘤组织的一大小、所述对象中一肿瘤的一阶段、所述对象中的一肿瘤负荷、或所述对象中肿瘤转移的一存在。

76. 如权利要求44至74任一项所述的方法,其特征在于,所述生物样本选自由以下组成的群组:全血、血浆、血清、尿液、脑脊液、血沉棕黄层、阴道液、阴道冲洗液、唾液、口腔冲洗液、鼻腔冲洗液、一鼻腔刷样本及其组合。

77. 一种非暂时性计算机可读介质,其特征在于,所述非暂时性计算机可读介质包含一机器可执行代码,所述机器可执行代码通过一个或多个计算机处理器执行时,实行如权利要求1至76任一项所述的方法。

78. 一种计算机产品,其特征在于,所述计算机产品包含一计算机可读介质,所述计算机可读介质存储多个指令,所述多个指令用于控制一计算机系统以执行如权利要求1至76任一项所述的方法的操作。

79. 一种系统,其特征在于,所述系统包含:

如权利要求78所述的计算机产品;及

一个或多个处理器,用于执行存储在所述计算机可读介质上的多个指令。

## 病毒相关联的癌症风险分层

[0001] 相关申请案

[0002] 本申请主张2020年1月15日提交的美国临时申请案第62/961,517号和2019年4月2日提交的美国临时申请案第62/828,224号的权益,其中通过引用并入本文。

### 背景技术

[0003] 许多疾病和症状可能与多种病原体(例如,病毒)的感染有关联。鼻咽癌(NPC)是中国南部和东南亚地区最常见的癌症之一,且鼻咽癌的发病机制与爱泼斯坦-巴尔病毒(Epstein-Barr virus,EBV)感染密切相关。在鼻咽癌高发区域,几乎所有鼻咽癌肿瘤都含有EBV基因组。基于EBV与鼻咽癌之间的密切关系,血浆EBV-DNA已发展成为鼻咽癌的生物标志物。使用实时聚合酶链反应(polymerase chain reaction,PCR)分析,血浆EBV DNA检测对检测鼻咽癌的敏感性为95%,及特异性为93%(Lo等人,癌症研究期刊,1999;59:1188-91)。基于对生物样本中病原体的多个无细胞核酸分子的分析,开发非侵入性或微创诊断分析方法对这些病原体相关联的疾病的风险进行分层可能具有重大的临床益处。

### 发明内容

[0004] 在一些方面中,本文提供了一种筛选一对象中一病原体相关联的疾病的方法,所述方法包含:接收来自在第一时间点执行的一第一分析的数据,所述第一分析包含确定来自所述对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征,其中来自所述病原体的所述多个无细胞核酸分子的所述特征包含与来自所述生物样本中的所述对象的所述多个无细胞核酸分子相比的一数量、一甲基化状态、一变异模式、一片段大小、或一相对丰度,其中所述特征表明所述对象发展出所述病原体相关联的疾病的一风险;及基于所述特征确定在一第二时间点处被执行的一第二分析,以筛选所述对象中的所述病原体相关联的疾病,其中所述第一时间点和所述第二时间点之间的一间隔与所述风险为负相关。

[0005] 在一些方面,本文提供了一种预测一对象中一病原体相关联的疾病的方法,所述方法包含:接收来自一第一分析的数据,所述第一分析包含确定来自所述对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征,其中来自所述病原体的所述多个无细胞核酸分子的所述特征包含与来自所述生物样本中的所述对象的所述多个无细胞核酸分子相比的一数量、一甲基化状态、一变异模式、一片段大小、或一相对丰度;以及基于来自所述病原体的所述多个无细胞核酸分子的所述特征以及下述其中一个或多个因素:所述对象的年龄、所述对象的吸烟习惯、所述对象的病原体相关联的疾病的家族史、所述对象的多个基因型因素、所述对象的种族或所述对象的饮食史,来生成表明所述对象发展出病原体相关联的疾病的一风险的一报道。

[0006] 在某些情况下,所述第一分析的结果不会导致针对所述病原体相关联的疾病对所述对象进行一医学治疗。在某些情况下,所述医学治疗包含使用多个治疗剂治疗、放射治疗或外科治疗。在某些情况下,在通过假阳性率低于1%的一临床诊断检查来确定一第二时间点之前,将所述对象诊断为没有患有所述病原体相关联的疾病。在某些情况下,所述临床诊

断检查包含物理检查、侵入性活检、内窥镜检查、磁共振成像、正电子发射断层扫描、计算机断层扫描或x射线成像。在某些情况下,所述临床诊断检查包含一侵入性活检,所述侵入性活检包含组织学分析、细胞学分析或细胞核酸分析。在某些情况下,所述间隔为至少约为2个月、4个月、6个月、8个月、10个月或12个月。在某些情况下,所述间隔为至少约为12个月。

[0007] 在某些情况下,所述方法还包含执行所述第一分析。在某些情况下,所述执行所述第一分析的步骤包含:(i)从所述对象获得一第一生物样本;以及(ii)在所述第一生物样本中测量来自所述病原体的多个无细胞核酸分子的一第一数量。在某些情况下,所述测量所述第一数量的步骤包含:测量所述第一生物样本中来自所述病原体的所述多个无细胞核酸分子的一拷贝数。在某些情况下,所述测量包含聚合酶链反应(PCR)。在某些情况下,所述测量包含定量聚合酶链反应(qPCR)。在某些情况下,所述第一数量包含:测量所述第一生物样本中来自所述病原体的所述多个无细胞核酸分子的一第一百分比。在某些情况下,所述第一分析还包含:(iii)如果所述第一数量高于一阈值,则从所述对象获得一第二生物样本,及在所述第二生物样本中从所述病原体中测量多个无细胞核酸分子的一第二数量。在某些情况下,所述第二生物样本在所述第一生物样本之后约4周获得。在某些情况下,与如果所述第二数量低于所述阈值的一间隔相比,如果所述第一数量和第二拷贝数都高于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔更短。在某些情况下,与如果所述第一数量高于所述阈值的一间隔相比,如果所述第一数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔更长。在某些情况下,如果所述第一数量和所述第二数量均高于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约1年。在某些情况下,如果所述第二数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约2年。在某些情况下,如果所述第一数量低于所述阈值,则所述第一时间点和所述第二时间点之间的所述间隔为约4年。在某些情况下,所述第一分析包含:在所述生物样本中确定来自所述病原体的多个无细胞核酸分子的一甲基化状态。在某些情况下,所述确定所述甲基化状态的步骤包含:使用一甲基化敏感限制酶或亚硫酸氢盐处理所述生物样本中的所述多个无细胞核酸分子。在某些情况下,所述确定所述甲基化状态的步骤包含:对所述对象的所述生物样本中的所述多个无细胞核酸执行一可识别甲基化测序。在某些情况下,所述可识别甲基化测序包含未甲基化胞嘧啶转化到尿嘧啶的亚硫酸氢盐转化。在某些情况下,所述可识别甲基化测序包含:使用一甲基化敏感限制酶进行处理。在某些情况下,所述第一分析包含:在所述生物样本中确定来自所述病原体的多个无细胞核酸分子的一片段大小分布。在某些情况下,所述确定所述片段大小分布的步骤包含:对所述生物样本中的多个无细胞核酸分子进行测序,以及基于映射到所述病原体的一参考基因组的多个序列读数确定来自所述生物样本中所述病原体的所述多个无细胞核酸分子的一片段大小。

[0008] 在某些情况下,所述第一分析包含:在所述生物样本中确定来自所述病原体的所述多个无细胞核酸分子的一变异模式。在某些情况下,所述确定所述变异模式的步骤包含:对所述生物样本中的多个无细胞核酸分子进行测序,以及基于映射到所述病原体的一参考基因组的多个序列读数以确定来自所述生物样本中所述病原体的所述多个无细胞核酸分子的所述变异模式。在某些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式包含多个单核苷酸变异。在某些情况下,所述识别所述变异模式的步骤包含:确定介于映射到所述病原体的所述参考基因组的多个序列读数与所述病原体的一疾病相关的参

考基因组之间的一相似性水平。在某些情况下,所述病原体的疾病相关参考基因组包含在一患病组织中被识别的所述病原体的一基因组。在某些情况下,所述确定相似性水平的步骤包含:将所述病原体的所述参考基因组分离到多个箱中;以及确定对于所述多个箱中的每个箱相对于所述病原体的所述疾病相关的参考基因组的一相似性指数,其中所述相似性指数与相应箱内的多个变异位点的一比例相关,所述相应箱中映射到所述病原体的所述参考基因组的所述多个序列读数中的至少一个与所述病原体的所述疾病相关的参考基因组具有一相同的核苷酸变异。在某些情况下,所述病原体的所述疾病相关的参考基因组包含所述病原体的多个疾病相关的参考基因组,及所述确定相似性水平的步骤包含:确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数;及基于所述多个疾病相关的参考基因组的一比例来确定所述多个箱中的每一个箱的一箱分数,相对于所述箱分数,所述多个箱内的所述相应相似性指数是高于一截止值。在某些情况下,所述多个箱中的每一个具有一长度为约100、200、300、400、500、600、700、800、900或1000个碱基对。在某些情况下,所述第一分析包含步骤包含:确定所述生物样本中来自所述病原体的所述多个无细胞核酸分子的所述甲基化状态、所述片段大小分布或所述变异模式。

[0009] 在某些情况下,所述方法进一步包含:使用应用于一数据输入的一分类器来计算对于所述对象发展出所述病原体相关联的疾病的一风险分数,所述数据输入包含所述生物样本中来自所述病原体的所述多个无细胞核酸分子的所述特征,其中所述分类器被配置为对所述数据输入应用一函数以生成一输出,所述数据输入包含来自所述生物样本中所述病原体的所述多个无细胞核酸分子的所述特征,所述输出包含所述风险分数,所述风险分数评估所述对象发展出疾病的风险。在某些情况下,所述分类器使用一标记的数据集进行训练。

[0010] 在某些情况下,所述方法进一步包含在所述第二时间点执行所述第二分析。在某些情况下,所述第二分析与所述第一分析相同。在某些情况下,所述第二分析包含来自所述对象的所述多个无细胞核酸分子的一分析、所述对象的侵入性活检、所述对象的内窥镜检查,或所述对象的一磁共振成像检查。

[0011] 在一些方面,本文提供了一种分析来自一对象的一生物样本的多个核酸分子的方法,所述方法包含:在一计算机系统中,从所述对象的所述生物样本获得多个无细胞核酸分子的多个序列读数,其中所述生物样本包含来自所述对象和潜在地来自一病原体的多个无细胞核酸分子;在所述计算机系统中,将所述多个无细胞核酸分子的所述多个序列读数与所述病原体的一参考基因组比对;及在所述计算机系统中,识别来自所述病原体的所述多个无细胞核酸分子的一变异模式,所述变异模式表征在所述病原体的所述参考基因组上的多个变异位点中的每一个处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,其中所述多个变异位点包含横跨所述病原体的所述参考基因组的至少30个位点,及所述变异模式表明所述对象中所述病原体相关联的疾病的一状态或一风险。

[0012] 在某些情况下,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少40个、至少50个、至少60个、至少70个、至少80个、至少90个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1000个、至少1100个,或至少1200个位点。在某些情况下,所述多个变异位点包含横跨所述病原体的所

述参考基因组的至少600个位点。在某些情况下,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少660个位点。在某些情况下,所述多个变异位点包含横跨所述病原体的所述参考基因组的至少1000个位点。在某些情况下,所述多个变异位点包含横跨所述病原体的所述参考基因组的约1100个位点。在某些情况下,所述多个变异位点由所有位点组成,在所述所有位点处,映射到所述病原体的所述参考基因组的所述多个序列读数具有与所述病原体的所述参考基因组的一不同的核苷酸变异。在某些情况下,所述比对所述多个序列读数的步骤被配置为允许映射到所述病原体的所述参考基因组的多个序列读数和所述病原体的所述参考基因组之间存在10、9、8、7、6、5、4、3、2或1个碱基的最大不匹配。在某些情况下,所述比对所述多个序列读数的步骤被配置为允许映射到所述病原体的所述参考基因组的多个序列读数和所述病原体的所述参考基因组之间存在2个碱基的最大不匹配。在某些情况下,所述方法还包含:基于映射到所述病原体的所述参考基因组的所述多个序列读数的所述变异模式来诊断、预测或监测所述对象的所述病原体相关联的疾病。在某些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式包含多个单核苷酸变异。在某些情况下,所述识别所述变异模式的步骤包含:确定介于映射到所述病原体的所述参考基因组的多个序列读数与所述病原体的一疾病相关的参考基因组之间的一相似性水平。在某些情况下,所述病原体的疾病相关参考基因组包含在一患病组织中被识别的所述病原体的一基因组。在某些情况下,所述确定相似性水平的步骤包含:确定对于所述多个箱中的每个箱相对于所述病原体的所述疾病相关的参考基因组的一相似性指数,其中所述相似性指数与相应箱内的多个变异位点的一比例相关,所述相应箱中映射到所述病原体的所述参考基因组的所述多个序列读数中的至少一个与所述病原体的所述疾病相关的参考基因组具有一相同的核苷酸变异。在某些情况下,所述病原体的所述疾病相关的参考基因组包含所述病原体的多个疾病相关的参考基因组,及所述确定相似性水平的步骤包含:确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数;及基于所述多个疾病相关的参考基因组的一比例来确定所述多个箱中的每一个箱的一箱分数,相对于所述箱分数,所述多个箱内的所述相应相似性指数是高于一截止值。在某些情况下,所述截止值为约0.9。在某些情况下,所述多个箱中的每一个具有一长度为约100、200、300、400、500、600、700、800、900或1000个碱基对。在某些情况下,所述方法还包含:使用应用于一数据输入的一分类器来计算对于所述对象发展出所述病原体相关联的疾病的一风险分数,所述数据输入包含来自所述病原体的所述多个无细胞核酸分子的所述变异模式,其中所述分类器被配置为对所述数据输入应用一函数以生成一输出,所述数据输入包含来自所述病原体的所述多个无细胞核酸分子的所述变异模式,所述输出包含所述风险分数,所述风险分数评估所述对象发展出疾病的风险。在某些情况下,所述分类器使用一标记的数据集进行训练。在某些情况下,所述分类器包含一数学模型,所述数学模型使用朴素贝叶斯模型、逻辑回归、随机森林、决策树、梯度提升树、神经网络、深度学习、线性/核支持向量机(SVM)、线性/非线性回归,或线性判别分析。

[0013] 在某些情况下,所述病原体是一病毒。在某些情况下,所述病毒为爱泼斯坦-巴尔病毒(EBV)。在某些情况下,所述病原体相关联的疾病包含鼻咽癌、NK细胞淋巴瘤、伯基特淋巴瘤、移植后淋巴增生性疾病、或霍奇金淋巴瘤。在某些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在多个变异位点中的每个位点处映射到所述病原

体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点包含从表6所列的多个基因组位点中相对于一EBV参考基因组(AJ507799.2)选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。在某些情况下,所述多个变异位点包含如表6所列相对于一EBV参考基因组(AJ507799.2)的一基因组位点。在某些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点是从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)随机选择的。在某些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点包含从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)随机选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。

[0014] 在某些情况下,所述病毒为一人类乳头瘤病毒(HPV)。在某些情况下,所述病原体相关联的疾病包含宫颈癌、口咽癌或头颈癌。在某些情况下,所述病毒为一乙型肝炎病毒(HBV)。在某些情况下,所述病原体相关联的疾病包含一肝硬化或一肝细胞癌(HCC)。在某些情况下,所述变异模式表明所述对象中所述病原体相关联的疾病的一状态,所述病原体相关联的疾病的所述状态包含所述对象中所述病原体相关联的疾病的一存在、所述对象中一肿瘤组织的一数量、所述对象中一肿瘤组织的一大小、所述对象中一肿瘤的一阶段、所述对象中的一肿瘤负荷、或所述对象中肿瘤转移的一存在。在某些情况下,所述生物样本选自自由以下组成的群组:全血、血浆、血清、尿液、脑脊液、血沉棕黄层、阴道液、阴道冲洗液、唾液、口腔冲洗液、鼻腔冲洗液、一鼻腔刷样本及其组合。

[0015] 在一些方面,本文提供了一种非暂时性计算机可读介质,所述非暂时性计算机可读介质包含一机器可执行代码,所述机器可执行代码通过一个或多个计算机处理器执行时,实行上述多种方法中的任一种方法。

[0016] 在一些方面,本文提供一种计算机产品,所述计算机产品包含一计算机可读介质,所述计算机可读介质存储多个指令,所述多个指令用于控制一计算机系统以执行上述多种方法中的任一种方法的操作。

[0017] 在一些方面,本文提供了一种系统,所述系统包含:如本文所述的计算机产品;以及一个或多个处理器,用于执行存储在所述计算机可读介质上的多个指令。

[0018] 在一些方面,本文提供了一种系统,所述系统包含:用于执行上述多种方法的任一种方法的装置。

[0019] 在一些方面,本文提供了一种系统,所述系统被配置为执行上述多种方法的任一种方法。

[0020] 在一些方面,本文提供了一种系统,所述系统包含多个模块,所述多个模块分别执行上述多种方法的任一种方法的多个步骤。

[0021] 援引参照:

[0022] 本说明书中提及的所有出版物、专利和专利申请均以引用的方式整体并入本文,其程度与每个单独的出版物、专利或专利申请被具体地和单独地表示以引用方式并入的程度相同。

## 附图说明

[0023] 本文所述的新颖特征在所附的权利要求中得到了详细阐述。通过参考以下阐述说明性实施例的详细描述,其中利用了本文所述的原理,以及附图,可以更好地理解本文所描述的特征和优点,其中所述附图:

[0024] 图1是对超过20000名对象的一群组进行鼻咽癌(NPC)筛选研究的设计示意图。

[0025] 图2示出了根据本公开的鼻咽癌筛选方案的示例性示意图。

[0026] 图3总结了基于来自鼻咽癌患者和非鼻咽癌对象样本的EBV变异基因谱的系统发育树分析。

[0027] 图4总结了基于来自鼻咽癌患者和非鼻咽癌对象样本(不包含29个报道的变异)的EBV变异基因谱的系统发育树分析。

[0028] 图5总结了基于来自鼻咽癌患者、非鼻咽癌对象和鼻咽癌前对象样本的EBV变异基因谱的系统发育树分析。

[0029] 图6总结了基于鼻咽癌患者、非鼻咽癌对象和鼻咽癌前对象(不包含29个报道的变异)样本的EBV变异基因谱的系统发育树分析。

[0030] 图7示出了以区块为基础的(block-based)变异模式分析的原理。

[0031] 图8总结了13例鼻咽癌、16例非鼻咽癌和4例鼻咽癌前样本的EBV DNA变异模式的以区块为基础的分析。

[0032] 图9总结了13个鼻咽癌、16个非鼻咽癌和4个鼻咽癌前(排除了29个报道的变异)样本的EBV DNA变异模式的以区块为基础的分析。

[0033] 图10A显示了基于使用以区块为基础的变异分析的所有EBV变异的分析,使用经训练的分类器计算的鼻咽癌风险分数。图10B显示了基于对29个报道的EBV变异的分析使用经训练的分类器计算的鼻咽癌风险分数。图10C显示了基于使用以区块为基础的变异分析(但排除了29个报道的变异)的所有EBV变异的分析,使用经训练的分类器计算的鼻咽癌风险分数。

[0034] 图11总结了具有EBV DNA短暂阳性或持续阳性的鼻咽癌患者和非鼻咽癌对象的甲基化水平。

[0035] 图12是示出了通过甲基化敏感酶消化诱导的具有血浆EBV DNA阳性的非癌症对象的血浆DNA的大小变化的示意图。填充和未填充的棒棒糖图(lollipop)分别代表甲基化和未甲基化的CpG位点。黄色水平长条表示血浆EBV DNA分子。随着酶消化,大小分布向左侧移动。

[0036] 图13是示出了通过甲基化敏感酶消化诱导的具有EBV DNA阳性的鼻咽癌患者的血浆DNA的大小变化的示意图。填充和未填充的棒棒糖图(lollipop)分别代表甲基化和未甲基化的CpG位点。黄色水平长条表示血浆EBV DNA分子。随着酶消化,大小分布向左侧移动。

[0037] 图14显示了使用甲基化敏感限制酶HpaII的以电脑模拟(in-silico)或不以电脑模拟消化的血浆EBV DNA的尺寸基因谱。

[0038] 图15显示了鼻咽癌患者和非鼻咽癌对象以甲基化敏感限制酶消化或不以甲基化敏感限制酶消化的血浆EBV DNA的累积大小分布。

[0039] 图16A是展示了EBV基因组中661个SNV位点的训练集中与鼻咽癌相关联的三个假设位点a、B和C的示意图。测试样本的鼻咽癌风险分数由多个血浆EBV DNA读数(例如,具有

可用基因型信息)覆盖的661个SNV位点子集的多个基因型模式来确定。从测试样本的血浆测序数据来看,基因型信息仅适用于位点A和C,而不适用于位点B,因为位点B未被任何测序EBV DNA读数覆盖。图16B是通过分析训练集中所有63个鼻咽癌样本和88个非鼻咽癌样本在2个位点上的基因型来展示位点A和C的基因型权重的示意图。建立了逻辑回归模型,以提供A和C位点高风险基因型的权重的信息。图16C是展示了基于A和C位点的基因型,通过从训练模型推导出的相应系数进行加权,得出测试样本的NPC风险分数的过程的示意图。图16D显示了训练集中鼻咽癌和非鼻咽癌样本中5678个SNV在EBV基因组中的分布(显示了在EBV基因组中1000个核苷酸的滑动窗口中的变异总数)。

[0040] 图17A和17B是使用略去一项方法总结训练集中鼻咽癌风险分数的图表。图17A显示训练集中鼻咽癌和非鼻咽癌血浆样本的鼻咽癌风险分数。图17B显示了通过鼻咽癌风险分数分析鼻咽癌和非鼻咽癌样本差异的ROC曲线分析。

[0041] 图18A和18B是总结测试集中鼻咽癌风险分数的图表。图18A显示测试集中鼻咽癌和非鼻咽癌血浆样本的鼻咽癌风险分数。图18B显示了通过鼻咽癌风险分数分析鼻咽癌和非鼻咽癌样本差异的ROC曲线分析。

[0042] 图19A和19B是通过分析EBER区域的多个基因型模式总结鼻咽癌风险分析的图表。图19A通过分析EBER区域的多个基因型模式,显示了测试集中鼻咽癌和非鼻咽癌血浆样本的鼻咽癌风险分数。图19B显示了基于EBER区域鼻咽癌风险分数分析的鼻咽癌和非鼻咽癌样本差异的ROC曲线分析。

[0043] 图20A和20B是通过分析BALF2区域的多个基因型模式总结鼻咽癌风险的图表。图20A通过分析BALF2区域的多个基因型模式,显示了测试集中鼻咽癌和非鼻咽癌血浆样本的鼻咽癌风险分数。图20B显示了基于BALF2区域鼻咽癌风险分数分析的鼻咽癌和非鼻咽癌样本差异的ROC曲线分析。

[0044] 图21显示了一种计算机控制系统,所述计算机控制系统可被编程或以其他方式配置以实行本文提供的多种方法。

[0045] 图22示出了本文公开的多种方法和多种系统的示意图。

## 具体实施方式

### [0046] 概述

[0047] 在多个方面中,本文提供了筛选一对象中一病原体相关联的疾病的多种方法及多种系统。所述多种方法和多种系统可基于来自所述对象的生物样本中来自所述病原体的多个无细胞核酸分子的特征,提供对所述对象发展出所述病原体相关联的疾病的风险的评估。在其他多种方法和多种系统中,风险预测可以确定适当的筛选频率。适当和及时的跟进筛选不仅可以节省所述对象的费用,而且可以早期发现疾病。例如,EBV-鼻咽癌(EBV-NPC)的分期分布转移到较早期可以显着改善鼻咽癌患者的无进展(progression-free)生存率。

[0048] 对象发展出病原体相关联的疾病的疾病的风险可指所述对象倾向于发展出病原体相关联的疾病的疾病的可能性。在某些情况下,如本文所述的风险是指对象中病原体相关联的疾病发展为可在未来时间点进行临床检测的状态(“临床可检测疾病”)的可能性。在某些情况下,对象在第一时间点通过筛选分析进行筛选,所述筛选分析测试从所述对象的生物样本中的病原体的多个无细胞核酸分子,并且当所述对象在第一时间点被诊断为不患有临床可检测

的病原体相关联的疾病时,所述对象的所述生物样本中所述病原体的所述多个无细胞核酸分子的特征可能表明所述对象在未来某个时间点存在临床可检测疾病的风险。

[0049] 临床可检测疾病是指表现出多个病症症状的疾病,可通过一项或多项成熟的临床诊断检查进行检测。在某些情况下,成熟的临床诊断检查包含对病原体相关联的疾病的低假阳性检出率的医学试验/分析,例如低于30%、20%、10%、8%、7%、6%、5%、4%、3%、2.5%、2%、1%、0.8%、0.5%、0.25%、0.15%、0.1%、0.08%、0.05%、0.02%、0.01%、0.005%、0.002%、0.001%甚至更低。成熟的临床诊断检查包含多种医学测试/分析,也可以具有检测病原体相关联的疾病的高敏感性,例如,至少30%、40%、50%、60%、70%、80%、85%、90%、92%、94%、95%、96%、97%、98%、99%、99.5%或100%。在某些情况下,病原体相关联的疾病是病原体相关联的增殖性疾病,例如癌症,通过一种或多种侵入性活检,然后通过对活检组织进行组织学或其他检查((如组织分析、细胞检查,如细胞DNA或蛋白质分析)、成像检查,如X射线、磁共振成像(MRI)、正电子发射断层扫描(PET)或计算机断层扫描(CT)或正电子发射计算机断层扫描(PET-CT)、实验室检查(如血液或尿液检查)或体检),可以高置信度和低假阳性率对癌症进行临床诊断。病原体相关联的疾病的诊断可由认证医生根据上述或其他成熟的临床检查结果进行。在某些情况下,第一次筛选分析的结果不会导致对象对病原体相关联的疾病进行医学治疗,因为所述对象通过成熟的临床诊断检查被诊断为不患有所述疾病。

[0050] 基于评估的风险,在某些情况下,这些方法包含确定对象中病原体相关联的多个筛选分析频率。筛选分析的频率可与风险相关,并且两次筛选分析(例如,本文所述的筛选分析和后续随后的筛选分析)之间的间隔可与风险反相关。在某些情况下,所述方法包含从在第一时间点执行的第一筛选分析接收数据。第一筛选分析可包含确定来自对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征。例如,第一筛选分析包含从对象获得的生物样本,并且所述生物样本包含来自所述对象(及可能来自所述病原体)的多个无细胞核酸分子(例如,无细胞DNA)。第一筛选分析还可包含确定生物样本中病原体的多个无细胞核酸分子的特征。本文提供的多种方法和多种系统中使用的来自病原体的多个无细胞核酸分子的非限制性特征包含与生物样本中对象的多个无细胞核酸分子相比的数量(例如,拷贝数或百分比)、甲基化状态、片段大小、变异模式、相对丰度。如本文所述,关于来自对象的生物样本或在对象执行检查或分析的时间点可以是指对象接受检查的时间点或从对象获得生物样本的时间点,而不是对生物样本执行实际分析的时间点。

[0051] 在某些情况下,本文提供的多种方法包含(a)接收来自在第一时间点执行的第一分析的数据,所述第一分析包含确定来自所述对象的一生物样本中的一病原体的多个无细胞核酸分子的一特征,其中来自所述病原体的所述多个无细胞核酸分子的所述特征包含与来自所述生物样本中的所述对象的所述多个无细胞核酸分子相比的一数量(例如,拷贝数或百分比)、一甲基化状态、一变异模式、一片段大小、或一相对丰度,其中所述特征表明所述对象发展出所述病原体相关联的疾病的一风险;及(b)基于所述特征确定在一第二时间点处被执行的一第二分析,以筛选所述对象中的所述病原体相关联的疾病,其中所述第一时间点和所述第二时间点之间的一间隔与所述风险为负相关。

[0052] 在某些情况下,如本文所述的对象的生物样本中的多个无细胞核酸分子的一个或多个特征使得能够采用非侵入性方法来评估在对象中病原体相关联的疾病(例如,癌症)的

状态或所述对象在未来发展为病原体相关联的疾病的风险。在不希望受到某些理论约束的情况下,可以存在至少两种可能的场景,所述两种可能场景是可用于多种方法和多种系统中的多个无细胞核酸分子的一个或多个特征与对象发展出病原体相关联的疾病的风险之间的关联的基础。在一种可能的情况下,遭受病原体相关联的疾病(例如,病原体相关肿瘤)的病变组织可能在初始筛选(例如,第一筛选分析)时已经存在。然而,病变组织(例如肿瘤)的大小可能太小,无法通过其他传统的医学检查方法(例如检测病原体相关联的疾病的假阳性率低于10%、5%、2%、1%、0.5%、0.1%或0.05%的方法,如内窥镜检查 and 磁共振成像(MRI))检测到。例如,随着疾病的发展,疾病组织(例如肿瘤,在大小上)的生长可以在后续筛选(第二筛选分析)中检测到(更晚期的疾病组织,例如扩大的组织(例如扩大的肿瘤))。另一种可能的情况是:病原体的多个核酸分子,例如EBV DNA,可以由处于初始病变状态的细胞释放,例如,癌前细胞,这些细胞随后可能发展为病变细胞,例如癌细胞。无论关联背后的确切情况如何,本文描述的对象可用于对多个对象进行分层,对所述多个对象随后出现临床可检测鼻咽癌的风险进行分层。

[0053] 在某些情况下,本文所述的多个具体筛选计划所用的多个实际时间间隔根据健康经济考量(例如,筛选成本)、对象偏好(例如,更频繁的筛选间隔可能对某些对象的生活方式更具破坏性)和其他临床参数调整(例如,个体的多个基因型(例如HLA状态)(Bei等人, Nat Genet. 期刊, 2010年; 42:599-603; Hildesheim等人, J Natl Cancer Inst. 期刊, 2002年; 94:1780-9。)、鼻咽癌家族史、饮食史、民族血统(如广东人)。

[0054] 在一些情况下,本文提供的多种方法包含:接收来自第一分析的数据,所述第一分析包括确定对象的生物样本中病原体的多个无细胞核酸分子的特征,其中来自病原体的多个无细胞核酸分子的特征包括与生物样品中来自对象的多个无细胞核酸分子相比的数量(例如,拷贝数或百分比)、甲基化状态、变异模式、片段大小、多个片段端部的坐标、多个片段端部的序列基序或相对丰度;以及基于来自所述病原体的所述多个无细胞核酸分子的所述特征以及下述其中一个或多个因素:所述对象的年龄、所述对象的吸烟习惯、所述对象的病原体相关联的疾病的家族史、所述对象的多个基因型因素或所述对象的饮食史来生成表明所述对象发展出病原体相关联的疾病的一风险的一报道。

[0055] 在多个方面中,本文提供了用于分析来自对象的生物样本中的多个核酸分子的多种方法和多种系统。所述多种方法和多种系统的示例可涉及分析生物样本中来自病原体的多个核酸分子的变异模式。在某些情况下,生物样本中病原体的多个核酸分子包含多个无细胞核酸分子。变异模式分析可以涉及生物样本中多个核酸分子的序列的比较,所述生物样本中所述多个核酸分子的所述序列被识别为源自具有一个或多个参考基因组的病原体,并随后在所述生物样本中确定来自所述病原体的所述多个核酸分子中的核苷酸变异模式。

[0056] 在一些情况下,本文提供的多种方法和多种系统包含基于生物样本中来自病原体的多个核酸分子中的变异模式来确定对象中病原体相关联的疾病的状态或风险。例如,在血浆中检测到的EBV基因组的遗传变异可用于预测未来鼻咽癌发展的风险。先前已被报道,EBV相关肿瘤和对照样本中存在EBV品系(Palser等人, J Virol 期刊, 2015年; 89:5222-37)可能有所不同,本研究中的肿瘤和对照样本来自不同的地理位置。鉴于EBV变异的地理变异,因此很难断定肿瘤样本中识别的多个变异是地理相关的还是疾病相关的。

[0057] 在某些情况下,本文所述的变异模式分析涉及生物样本中来自病原体的多个核酸

分子与病原体的一个或多个参考基因组之间的全基因组 (genomewide) 的比较。全基因组的比较可能涉及横跨病原体整个基因组的序列比对以及随后核苷酸变异模式的聚类分析。在某些情况下,全基因组的比较涉及横跨病原体的参考基因组中大量位点处的核苷酸变异的分析。这些位点可以包含横跨病原体的整个基因组中的所有位点。替代地,横跨病原体的参考基因组的这些位点或变异位点可包含至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少1100、至少1200个、至少1300个、至少1400个、至少1500个、至少1600个、至少1700个、至少1800个、至少1900个、至少2000个、至少3000个、至少4000个或至少5000个通常可以被发现的核苷酸变异的位点。本文所述的核苷酸变异可包含多种单核苷酸变异 (SNVs)。本文提供的用于变异模式分析的多个变异位点可包含在病原体的基因组中识别的典型单核苷酸变异 (SNV)。在某些情况下,多个变异位点可能包含多种插入、多种删除和多种融合。

[0058] 本文提供的全基因组变异模式分析可能优于个别的多个单核苷酸多态性 (SNPs) 分析。在示例性情况下,虽然固定数量的位点上的SNPs可与可导致对象的病状的病原体的特定(多种)品系或(多种)亚型相关联,但基于这些个别的SNPs的分析的风险评估可限于所述病原体的所述特定(多种)品系或所述(多种)亚型,及如果存在病原体的其他致病的(多种)品系或(多种)亚型,则无法提供准确的风险评估。在另一示例性情况下,当生物样本中的病原体核酸分子稀少时,例如,当分析例如血浆等生物样本中的多个无细胞核酸分子时,本文提供的全基因组变异模式分析是有益的。生物样本中可用的病原体核酸分子可能没有病原体基因组覆盖的显著数量。因此,涉及横跨病原体的整个基因组中大量的变异位点的全基因组变异模式分析可以相对更全面地读出生物样本中病原体的多个无细胞核酸分子的基因型特征,然而,涉及固定数量的个别的多态性的分析仅限于基因组的相对小的区域或一些小的区域,因此可以提供生物样本中病原体的多个无细胞核酸分子的基因型特征相对有限的读出。

[0059] 在一些情况下,本文提供的变异模式分析包含以区块为基础的模式分析,其涉及将病原体的参考基因组分进行分层到多个箱(bin)中,并分析相对于所述多个箱中的每个箱的多个序列读数。在一些情况下,所述方法包含:确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数。所述相似性指数与相应箱内的多个变异位点的一比例相关,所述相应箱中映射到所述病原体的所述参考基因组的所述多个序列读数中的至少一个与所述病原体的所述疾病相关的参考基因组具有一相同的核苷酸变异。在某些情况下,所述病原体的所述疾病相关的参考基因组包含所述病原体的多个疾病相关的参考基因组,所述方法包含确定对于所述多个箱中的每一个箱相对于所述病原体的所述多个疾病相关的参考基因组中的每一个的一相应相似性指数;及基于所述多个疾病相关的参考基因组的一比例来确定所述多个箱中的每一个箱的一箱分数,相对于所述箱分数,所述多个箱内的所述相应相似性指数是高于一截止值。

[0060] 无细胞核酸分子的分析

[0061] 从对象的生物样本中筛选多个无细胞核酸分子的分析可以是任何适当的核酸分析。例如,多种测序方法可用于分析多个无细胞核酸分子的数量(例如,拷贝数或百分比)、甲基化状态、片段大小或相对丰度。替代地或另外地,也可以使用以扩增或杂交为基础的方

法,例如各种聚合酶链反应(PCR)方法或以微阵列为基础的方法。在某些情况下,多种免疫沉淀法被用来分析多个核酸分子的甲基化状态。

[0062] 在本公开的一些示例中,用于检测多个无细胞病原体核酸分子(例如,无细胞EBV DNA)的筛选分析包含在不同时间点进行的多于一次的测试,及在多次测试中,所述多个无细胞病原体核酸分子的可检测性可表明对象发展出病原体相关联的疾病的风险。例如,所述分析可包含两步骤分析,或包含3、4、5、6、7、8、9、10或甚至更多测试的分析方案。所述多个测试中的一些可在同一时间点执行,而其他测试可在(多个)不同时间点执行,替代地,所有测试可在不同时间点执行。

[0063] 不同筛选分析的时间或筛选频率可通过本文提供的多种方法和多种系统确定。第一筛选分析和第二筛选分析之间的间隔可以为至少为约2个月、4个月、6个月、8个月、10个月或12个月。在某些情况下,所述间隔至少为约12个月。第一筛选分析和第二次筛选分析之间的间隔可为约1年、1.5年、2年、2.5年、3年、3.5年、4年、4.5年、5年、6年、7年、8年、9年、10年或更长。只要对象通过成熟的临床诊断方法正常诊断为未患有病原体相关联的疾病(例如,未患有临床可检测的病原体相关联的疾病),间隔可以很长,即使第一筛选分析可以给出阳性结果,表明存在病原体相关联的疾病。本文提供的多种方法和多种系统能够预测对象在未来(例如,在6个月、12个月、2年、3年、5年或10年内)发展出病原体相关联的疾病的风险。基于评估的风险,可以确定适当的跟进时间点。

[0064] 可优化获得样本和执行分析之间的时间,以提高分析或方法的敏感性和/或特异性。在一些实施例中,可在执行分析之前立即获得样本(例如,在执行第一分析之前获得第一样本,并且在执行第一分析之后但在执行第二分析之前获得第二样本)。在一些实施例中,可在执行分析之前获得样本并存储一段时间(例如,几个小时、几天或几周)。在一些实施例中,可在从对象获得样本后的1天、2天、3天、4天、5天、6天、1周、2周、3周、4周、5周、6周、7周、8周、3个月、4个月、5个月、6个月、1年或1年以上对样本进行分析。

[0065] 执行分析(例如,第一分析或第二分析)与确定样本是否包含表明疾病(例如,肿瘤)的标记物或标记物集之间的时间可以变化。在某些情况下,可以优化时间以提高分析或方法的敏感性和/或特异性。在一些实施例中,确定样本是否包含表明肿瘤的标记物或标记物集可在执行分析的至多0.1小时、0.5小时、1小时、2小时、4小时、8小时、12小时、24小时、2天、3天、4天、5天、6天、1周、2周、3周或1个月内发生。

[0066] 如本文所述的生物样本的测序分析可用于分析来自病原体的多个无细胞核酸分子的一个或多个特征。本文提供的多种方法可包含对来自生物样本的多个核酸分子(例如,多个无细胞核酸分子、多个细胞核酸分子或两者)进行测序。在一些示例中,本文提供的多种方法包含分析来自生物样本的多个核酸分子的多个测序结果,例如多个测序读数。本文提供的多种方法和多种系统可涉及或不涉及测序的主动步骤。多种方法和多种系统可以包含或提供用于从测序器接收和处理测序数据的方法。多种方法和多种系统还可以包含或提供向测序器提供命令以调整测序程序的(多个)参数的方法,例如,基于多个测序结果分析的多个命令。

[0067] 商用测序设备可用于本公开中提供的多种方法,如Illumina测序平台和454/Roche平台。可使用本领域已知的任何方法对核酸进行测序。例如,测序可以包含下一代测序(next generation sequencing)。在某些情况下,可使用链终止测序、杂交测序、

Illumina测序(例如,使用多种可逆终止剂染料)、ion torrent半导体测序、质量分光光度法测序、大规模平行签名测序(MPSS)、马克萨姆-吉尔伯特测序(Maxam-Gilbert sequencing)、纳米孔测序、聚合酶克隆测序、焦磷酸测序、霰弹枪测序、单分子实时(SMRT)测序、SOLiD测序(使用四个荧光标记的双碱基探针杂交)、通用测序或其任何组合。

[0068] 可在本文所提供的多种方法中使用的一种测序方法可涉及双端测序,例如,使用Illumina“用对读测序模块”及其基因组分析仪。使用此模块,在基因组分析仪完成第一测序读数后,用对读测序模块可以指导原始模板的重新合成和第二轮簇的生成。通过在本文提供的多种方法中使用双端读数,可以从核酸分子的两端获得序列信息,并将两端映射到参考基因组,例如,病原体的基因组或宿主生物体的基因组。在对两端进行映射之后,人们可以根据如本文所提供的多种方法的一些实施例确定病原体整合轮廓。

[0069] 在双端测序期间,从核酸分子的第一端部序列读数可包含至少20、至少25、至少30、至少35、至少40、至少45、至少50、至少55、至少60、至少65、至少70、至少75、至少80、至少85、至少90、至少95、至少100、至少105、至少110、至少105、至少120、至少125、至少130、至少135、至少140、至少145、至少150、至少155、至少160、至少165、至少170、至少175或至少180个连续的核苷酸。从核酸分子的第一端部序列读数可包含最多24、最多28、最多32、最多38、最多42、最多48、最多52、最多58、最多62、最多68、最多72、最多78、最多82、最多88、最多92、最多98、最多102、最多108、最多122、最多128、最多132、最多138、,最多142个、最多148个、最多152个、最多158个、最多162个、最多168个、最多172个或最多180个连续的核苷酸。从核酸分子的第一端部序列读数可包含约20、约25、约30、约35、约40、约45、约50、约55、约60、约65、约70、约75、约80、约85、约90、约95、约100、约105、约110、约105、约120、约125、约130、约135、约140、约145、约150、约155、约160、约165、约170、约175或约180个连续的核苷酸。从核酸分子的第二端部序列读数可包含至少20、至少25、至少30、至少35、至少40、至少45、至少50、至少55、至少60、至少65、至少70、至少75、至少80、至少85、至少90、至少95、至少100、至少105、至少110、至少105、至少120、至少125、至少130、至少135、至少140、至少145、至少150、至少155、至少160、至少165、至少170、至少175或至少180个连续的核苷酸。从核酸分子的第二端部序列读数可包含最多24、最多28、最多32、最多38、最多42、最多48、最多52、最多58、最多62、最多68、最多72、最多78、最多82、最多88、最多92、最多98、最多102、最多108、最多122、最多128、最多132、最多138、,最多142个、最多148个、最多152个、最多158个、最多162个、最多168个、最多172个或最多180个连续的核苷酸。从核酸分子的第二端部序列读数可包含约20、约25、约30、约35、约40、约45、约50、约55、约60、约65、约70、约75、约80、约85、约90、约95、约100、约105、约110、约105、约120、约125、约130、约135、约140、约145、约150、约155、约160、约165、约170、约175或约180个连续的核苷酸。在某些情况下,从核酸分子的第一端部序列读数可包含至少75个连续的核苷酸。在某些情况下,从核酸分子的第二端部序列读数可包含至少75个连续的核苷酸。从核酸分子的第一端和第二端部序列读数可以具有相同的长度或不同的长度。从生物样本中读取的多个核酸分子的序列可以具有相同的长度或不同的长度。

[0070] 本文提供的多种方法中的测序可在不同的测序深度下执行。测序深度是指基因座被与所述基因座比对的序列读数覆盖的次数。基因座可以小到核苷酸,或大到染色体臂,或大到整个基因组。本文提供的多种方法中的测序深度可以是50倍(50x)、100倍(100x)等,其

中“倍(x)”之前的数字是指一个位点被序列读数覆盖的次数。测序深度也可应用于多个位点或整个基因组,在这种情况下,倍(x)可分别指基因座或单倍体基因组或整个基因组被测序的平均次数。在某些情况下,超深测序在本文所述的多种方法中执行,其可指执行至少100倍测序深度。

[0071] 在测序过程中读取核酸内特定核苷酸的次数或平均次数(例如,测序深度)可比正在测序的核酸长度大数倍。在某些情况下,当测序深度显著大(例如,至少5倍)于核酸长度时,测序可称为深度测序。在一些示例中,测序深度平均可比正在测序的核酸长度大至少约5倍、大至少约10倍、大至少约20倍、大至少约30倍、大至少约40倍、大至少约50倍、大至少约60倍、大至少约70倍、大至少约80倍、大至少约90倍、大至少约100倍。在某些情况下,可针对特定分析物(例如,核酸片段或癌症特异性核酸片段)富集样本。

[0072] 在本文提供的多种方法中生成的序列读数(或多个测序读数)可指从核酸分子的任何部分或全部测序的核苷酸串。例如,序列读数可以是与核酸片段互补的核苷酸短串(例如,20-150),与核酸片段端部互补的核苷酸串,或与生物样本中存在的整个核酸片段互补的核苷酸串。序列读数可以通过多种方式获得,例如,使用多种测序技术

[0073] 数量/可检测性

[0074] 可用于多种方法和多种系统中的多个无细胞核酸分子的特征之一是来自病原体的多个无细胞核酸分子的数量(例如,拷贝数或百分比)。本公开的一些方面涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的数量(例如,拷贝数或百分比)的评估所述对象发展出病原体相关联的疾病的风险分层。

[0075] 生物样本中核酸分子的拷贝数与核酸分子的可检测性有关。给定特定的分析方法,核酸模板的可检测性可与模板分子的拷贝数相关,例如,低于分析方法检测较低的拷贝数可以是不可检测,而等于或高于分析方法检测下限的拷贝数可称为“可检测”例如,定量聚合酶链反应(qPCR)方法通常有一个检测下限,在此下限下,模板分子的信号无法与背景噪声区分。因此,在某些情况下,本文提供的多种方法和多种系统直接依赖于生物样本中多个无细胞核酸分子的可检测性,其可与生物样本中的其拷贝数相关。在某些情况下,直接测量生物样本中多个无细胞核酸分子的拷贝数。在其他情况下,通过检测多个无细胞核酸分子本身,隐式测量或推断拷贝数。

[0076] 可以执行多种检测分析,例如聚合酶链反应(PCR)或定量聚合酶链反应(qPCR),以评估生物样本中病原体的多个无细胞核酸分子的存在或不存在或拷贝数。多种探针可设计为针对病原体特异性基因组区域,例如,EBV特异性基因组DNA序列、人类乳头瘤病毒(HPV)特异性基因组DNA序列或乙型肝炎病毒(HBV)特异性基因组DNA序列。

[0077] 虽然本文提供了多种示例和实施例,但与例如拷贝数和鼻咽癌相关的附加技术和实施例可在2011年11月30日提交的PCT国际专利申请案第PCT AU/2011/001562号中找到,其全部通过引用并入本文。鼻咽癌可能与爱泼斯坦-巴尔病毒(EBV)感染密切相关。在中国南方,几乎所有鼻咽癌患者的肿瘤组织中都可以发现EBV基因组。来自鼻咽癌组织的血浆EBV DNA已被开发为鼻咽癌的肿瘤标记物(Lo等人,Cancer Res期刊1999年;59:1188-1191)。特别是,实时qPCR分析可用于针对EBV基因组的BamHI-W片段的血浆EBV DNA分析。每个EBV基因组5中可能有约6到12个BamHI-W片段重复,每个鼻咽癌肿瘤细胞中可能有大约50个EBV基因组(Longnecer等人,Fields Virology,第5版,第61章爱泼斯坦-巴尔病毒;

Tierney等人, J Virol. 期刊2011年;85:12362-12375)。换句话说,每个鼻咽癌肿瘤细胞中可能有300-600(例如,约500)个PCR靶拷贝。这样每个肿瘤细胞的高靶数量可以解释为什么血浆EBV DNA是检测早期鼻咽癌的高度敏感标记物。鼻咽癌细胞能够沉积EBV DNA片段到对象的血液中。此肿瘤标记物可用于监测(Lo等人, Cancer Res 期刊1999年;59:5452-5455)和预测(Lo等人, Cancer Res 期刊2000年;60:6878-6881)。

[0078] qPCR分析也可以类似于本文所述的EBV的方式用于测量样本中HPV、HBV或任何其他病毒DNA的量。这种分析对于筛选宫颈癌(CC)、头颈部鳞状细胞癌(HNSCC)、肝硬化或肝癌(HCC)尤其有用。在一个示例中,qPCR分析靶HPV基因组的多态性L1区域内的区域(例如,200个核苷酸)。更具体地说,本文所设想的是qPCR引物的使用,其选择性地杂交到编码L1区域中的一个或多个高变表面环的多个序列。

[0079] 或者,可以使用多种测序技术检测和量化来自病原体的多个无细胞核酸分子。例如,cfDNA片段可以被测序并与HPV参考基因组比对和量化。或者在其他示例中,多个cfDNA片段的多个序列读数与EBV或HBV的参考基因组比对和量化。

[0080] 通过本文提供的分析测量的来自病原体的多个无细胞核酸分子的可检测性或拷贝数可表明对象发展出病原体相关联的疾病的风险。在一些示例中,来自病原体的多个无细胞核酸分子的拷贝数越高,对象发展出病原体相关联的疾病的风险越高。在某些情况下,来自病原体的多个无细胞核酸分子在一个特定时间点或多个时间点的分析中的可检测性表明对象发展出病原体相关联的疾病的风险。当来自对象的生物样本中的来自病原体的多个无细胞核酸分子可检测时,与通过本文提供的分析无法检测到所述多个分子时相比,对象可被置于更高的病原体相关联的疾病的风险中。可在如上所述的时间执行多步骤检测分析。

[0081] 在本公开的一些示例中,执行两步骤分析以检测生物样本中的多个无细胞病原体核酸分子。在某些情况下,取决于第一时间点的分析结果,执行两步骤分析的第一试验,随后执行或不执行两步骤分析的第二试验。例如,如果第一试验提供阳性结果,例如,在第一生物样本中检测到无细胞病原体核酸分子,则可以执行两步骤检测分析的第二试验;如果从第一试验获得阴性结果,则可以不执行第二试验。在其他情况下,无论第一次测试如何,都会执行第二次测试。在一些示例中,两步骤检测分析的两个试验都有阳性结果的情况称为永久阳性,而只有第一或第二试验有阳性结果的情况称为暂时阳性。在一个说明性示例中,“阴性”分析结果相比,“阳性”分析结果表明对象发展出病原体相关联的疾病(如EBV相关联的鼻咽癌)的风险更高,而与“短暂阳性”分析结果相比,“永久阳性”分析结果表明更高的风险。在一些说明性示例中,与获得短暂阳性结果时相比,当从第一时间点执行的两步骤检测分析中获得永久阳性结果时,可以在第一时间点和第二时间点之间设置更长的间隔。例如,在EBV相关联的鼻咽癌筛选中,如果从两步骤检测分析中的第一检测分析获得永久阳性结果,则可建议在第一检测分析后约一年内进行后续第二筛选分析。相反,如果从两步骤检测分析中的第一检测分析获得短暂阳性结果,则可在第一检测分析的大约两年内执行后续第二筛选分析。如果获得阴性结果,后续筛选试验的间隔时间可以为四年甚至更长。在某些情况下,表明较高风险的先前阳性结果可以推翻区间选择,区间选择将通过表明较低风险的后续结果设置。例如,在第1年获得了永久阳性结果,则对象将在接下来的4年内每年进行跟进,无论后续4年内进行的跟进分析结果如何。图2中给出了说明性示例,并在示例

2中进行了更详细的描述。与检测分析类似,基于病原体的多个无细胞核酸分子的其他特征的风险评估也可以遵循此示例性或类似的筛选方案。

[0082] 可在第一试验后的数小时、数天或数周内进行第二试验。在一个示例中,可在第一分析之后立即执行第二分析。在其他情况下,可在第一分析后1天、2天、3天、4天、5天、6天、1周、2周、3周、4周、5周、6周、7周、8周、3个月、4个月、5个月、6个月、1年或1年以上进行第二分析。在特定示例中,可在第一样本的2周内执行第二分析。通常,第二试验可用于提高特异性,从而在患者中检测到病原体相关联的疾病,例如肿瘤。执行第一试验和第二试验之间的时间可通过实验来确定。在一些实施例,所述方法可包含2个或多个试验,且两个试验使用相同的样本(例如,在执行第一分析之前从对象(例如,患者处获得单个样本),并保存一段时间直到执行第二分析)。例如,可以同时从一对象身上获得两管血液。第一管可用于第一试验。只有当对象的第一试验结果为阳性时,才可以使用第二管。可使用本领域技术人员已知的任何方法(例如,低温法)保存样本。这种保存在某些情况下可能是有益的,例如,对象可以收到阳性检测结果(例如,第一分析表明有癌症),患者可以不等到进行第二次检测,而是选择寻求第二意见。

[0083] 甲基化状态

[0084] 本公开的一些方面涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的甲基化状态的评估,对于对象发展出病原体相关联的疾病的风险的分层。

[0085] 多个无细胞病原体核酸分子的甲基化可以区分患有病原体相关联的疾病的患者(如EBV相关联的鼻咽癌或HPV相关联的宫颈癌)和没有患有所述疾病的对象(如非鼻咽癌对象)的样本的差异。例如,如美国专利申请案第16/046,795号所示,与鼻咽癌相关联的血浆EBV DNA的甲基化状态可以不同于在非鼻咽癌对象中检测到的血浆EBV DNA的甲基化状态,其通过引用全部并入本文。当通过亚硫酸氢盐测序分析时,鼻咽癌患者和非鼻咽癌对象的血浆DNA之间可能存在差异甲基化的多个区域,和可检测到EBV DNA。因此,分析这些差异甲基化区域的甲基化状态可以区分鼻咽癌和非鼻咽癌对象。如本文所述,鼻咽癌相关联的EBV DNA甲基化状态也可预测鼻咽癌发展的风险,并可用于调整鼻咽癌筛选间隔。例如,与没有鼻咽癌相关联的EBV DNA甲基化模式的对象相比,具有鼻咽癌相关联的EBV DNA甲基化模式的对象可以更频繁地进行筛选。在某些情况下,可以使用其他类型的可识别甲基化测序来代替亚硫酸氢盐测序,例如,使用单分子测序系统,如Pacific Biosciences公司的测序系统(Kelleher等人,Methods Mol Biol.期刊2018年;1681:127-137;Powers等人,BMC Genomics.期刊2013年;14:675)和牛津纳米孔(Simpson等人,Nat Methods.期刊2017年;14:407-10),以及测序前使用甲基化敏感限制酶处理。在另一种情况下,可以使用可识别甲基化且不是以测序为基础的分子方法,例如甲基化特异性PCR(Herman等人,美国科学院院报1996年;93:9821-6),基于甲基化敏感酶(如限制酶)的检测系统,和亚硫酸氢盐转化然后进行质谱分析(van den Boom等人,Methods Mol Biol.期刊2009年;507:207-27;Nygren等人,Clin Chem.期刊2010年;56:1627-35),以及基于DNA分子的甲基化状态的差异沉析的方法(例如,使用抗甲基化胞嘧啶抗体(Shen等人,Nature.期刊2018年;563:579-83;Zhou等人,PLoS One.期刊2018年;13:e0201586)或甲基化结合蛋白(Zhang等人,Nat Commun.期刊2013年;4:1517)。

[0086] 在某些情况下,多个无细胞病原体核酸分子(例如血浆EBV DNA)的甲基化模式可

用于检测多种病原体相关联的疾病(例如,病原体相关联的癌症,例如鼻咽癌),或预测具有临床可检测疾病的未来风险。如上所述,一种方法是使用亚硫酸氢盐处理多个核酸分子以将未甲基化胞嘧啶转化为尿嘧啶。甲基化的胞嘧啶不会被亚硫酸氢盐改变,而是保持为胞嘧啶。亚硫酸氢盐处理过的多个核酸分子的后续检查,如测序,可用于检测生物样本中多个核酸分子的甲基化状态。

[0087] 在一个示例中,使用甲基化敏感限制酶分析确定血浆EBV DNA甲基化水平的差异。甲基化敏感限制酶的一个非限制性示例是HpaII,它可以切割携带未甲基化“CCGG”基序的分子,但留下没有“CCGG”或甲基化“CCGG”的分子未改变。或者,也可以使用其他甲基化敏感限制酶。在一个示例中,由于非癌症对象血浆EBV DNA的甲基化水平较低,非癌症对象血浆EBV DNA可能更容易被甲基化敏感限制酶的切割。可以确定酶消化的敏感性,例如但不限于大规模平行测序、凝胶电泳、毛细管电泳、聚合酶链反应(PCR)和实时PCR。

[0088] 在使用测序(例如大规模平行测序)分析甲基化敏感限制酶的消化程度的情况下,可使用病原体的多个无细胞核酸分子(例如血浆EBV DNA)的大小分布(以酶消化和不以酶消化)来反映消化程度。如图12和13所示,大小分布曲线向左移动可表明血浆EBV DNA的大小分布变短。曲线越向左移动,反映酶消化程度越高,意味着DNA甲基化水平越低。

[0089] 如本文所述的多个无细胞病原体核酸分子的甲基化状态可包含个别的甲基化位点的甲基化密度、在病原体基因组上相邻区域上甲基化/未甲基化位点的分布,病原体基因组上一个或多个特定区域内或整个病原体基因组中每个个别的甲基化位点的甲基化模式或水平,以及非CpG甲基化。在某些情况下,甲基化状态包含个别的差异甲基化位点的甲基化水平(或甲基化密度),这些甲基化位点可在例如具有病原体相关联的疾病(例如,EBV相关联的鼻咽癌或HPV相关联的宫颈癌)的患者和无所谓疾病的对象(例如,非鼻咽癌对象)的样本之间被识别。对于给定的甲基化位点,甲基化密度可指在给定的甲基化位点上甲基化的核酸分子的部分超过包含这样的甲基化位点的感兴趣的核酸分子的总数。例如,肝组织中第一甲基化位点的甲基化密度可以是指第一位点处甲基化的多个肝DNA分子在整个肝DNA分子中的一部分。在某些情况下,甲基化状态包含个别的甲基化位点之间甲基化/未甲基化状态的一致性(例如,模式或单倍型(haplotype))。

[0090] 在某些情况下,如本文所述的筛选分析(例如,第一分析或第二分析)可包含通过任何可用技术确定多个无细胞核酸分子的甲基化状态,例如但不限于执行可识别甲基化测序(methylation-aware sequencing)、甲基化敏感性扩增(methylation-sensitive amplification)或甲基化敏感性沉淀析(methylation-sensitive precipitation)。虽然本文提供了示例和实施例,但与例如确定甲基化状态有关的附加技术和实施例可在2013年9月20日提交的PCT专利申请案第PCT AU/2013/001088号中找到,其通过引用完全并入本文。

[0091] 片段大小

[0092] 本公开的一些方面涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的片段大小的评估,对于对象发展出病原体相关联的疾病的风险的分层。

[0093] 多个无细胞病原体核酸分子的片段大小分布和/或相对丰度可区分患有病原体相关联的疾病的多个患者(例如,EBV相关联的鼻咽癌或HPV相关联的宫颈癌)和没有所述疾病的对象(例如,非鼻咽癌对象)的样本。例如,血浆EBV DNA分子的大小分布以及循环DNA分子映射到EBV基因组和人类基因组的比率有助于区分鼻咽癌患者和可检测到血浆EBV DNA的

非鼻咽癌对象,如大规模平行测序所示(Lam等人,美国科学院院报2018年;115:E5115-E5124),其全部通过引用并入本文。根据本公开的一些示例,与鼻咽癌相关联的大小分布和映射到EBV和人类基因组的循环DNA的相对丰度也可用于预测未来临床可检测到鼻咽癌的发展风险。在一个实施方案中,在血浆DNA测序中具有这些鼻咽癌相关联的特征但未检测到鼻咽癌的对象可以比具有可检测到血浆EBV DNA但没有这些鼻咽癌相关联的特征的对象更频繁地跟进。对于使用上述的两步骤分析,使用这种基于测序的分析对鼻咽癌的风险进行分层有一个潜在的实际优势,那就是可以省去从病人那里收集另一个血样。

[0094] 在某些情况下,分析(例如,第一分析或第二分析)可包含执行分析(例如,下一代测序分析)以分析核酸片段大小(例如,血浆EBV DNA的片段大小)。在某些情况下,测序用于评估样本中无细胞病毒核酸的大小。例如,每个已测序的血浆DNA分子的大小可以从序列的起始坐标和结束坐标得到,其中所述坐标可以通过将多个序列读数映射(比对)到病毒基因组来确定。在各种示例中,DNA分子的起始和结束坐标可以通过两个双端读数或覆盖两端的单个读数(这在单分子测序中可以实行)来确定。在某些情况下,基于扩增或杂交的方法也可用于片段大小的分析。例如,探针可设计成标靶不同长度的基因组区域,扩增(例如,PCR或qPCR)或杂交信号可表明标靶基因组区域(同时长度等于或大所述标靶区域)的无细胞核酸片段的数量。由此可以推断片段大小的分布。用于片段大小分析(assay及analyses)的多种方法可包含美国专利申请案公开第US20180208999A1号中所述的方法,其通过引用全部并入本文。

[0095] 片段大小分布可以显示为水平轴上核酸片段大小的直方图。可以确定每种大小(例如,在1碱基对(bp)分辨率内)的核酸片段的数量并绘制在垂直轴上(例如,作为原始数量或频率百分比)。大小分辨率可以超过1bp(例如,2、3、4或5bp分辨率)。以下对大小分布的分析(也称为大小分布图)示出来自鼻咽癌对象的无细胞混合物中的病毒DNA片段在统计学上比无明显病症的对象更长。在一个说明性示例中,在从血浆EBV DNA分析获得的片段大小分布曲线中,鼻咽癌患者的血浆EBV DNA大小曲线中可能存在表示166bp峰值的特征(核小体模式),而非癌症对象的血浆EBV DNA不显示典型的核小体模式。

[0096] 在某些情况下,计算来自病原体的多个无细胞核酸分子相对于来自对象的多个无细胞核酸分子的相对丰度,以评估风险。在某些情况下,相对丰度是根据大小比例进行分析的。在各种示例中,病原体片段与来自对象的无细胞片段的大小比例是指来自病原体的无细胞核酸片段与来自对象的无细胞核

$$\text{大小80-110bp比例} = \frac{\text{80-110bp内的EBV DNA片段的比例}}{\text{80-110bp内的常染色体DNA片段的比例}}$$

[0097] 酸片段之间的数量比率。例如,80到110碱基对之间的EBV DNA片段的大小比例(size ratio)可以是:

[0098] 在各种情况下,截止值或阈值为评估所设置。例如,可以存在尺寸阈值(size threshold)用于确定病原体片段和对象常染色体(autosomal)片段之间的大小比例。或者在某些情况下,设置尺寸阈值,以便将大小低于或高于阈值的多个片段视为表明对象发展出病原体相关联的疾病的风险。应该理解,尺寸阈值可以是任何值。尺寸阈值可至少为约10bp、20bp、25bp、30bp、35bp、40bp、45bp、50bp、55bp、60bp、65bp、70bp、75bp、80bp、85bp、

90bp、95bp、100bp、105bp、110bp、115bp、120bp、125bp、130bp、135bp、140bp、145bp、150bp、155bp、160bp、165bp、170bp、175bp、180bp、185bp、190bp、195bp、200bp、210bp、230bp、240bp、250bp,或大于250bp。例如,尺寸阈值可以是150bp。在另一个示例中,尺寸阈值可以是180bp。在一些实施例中,可以使用较高和较低的尺寸阈值(例如,一范围的多个值)。在一些实施例中,较高和较低的尺寸阈值可用于选择具有长度在较高和较低截止值之间的核酸片段。在一些实施例中,较高和较低的截止值可用于选择具有长度大于较高截止值且小于较低尺寸阈值的核酸片段。在某些情况下,大小比例的截止值用于确定对象是否有风险,或对象发展出病原体相关联的疾病(如鼻咽癌)的风险有多大。例如,患有鼻咽癌的对象较血浆EBV DNA结果呈假阳性的对象具有较低的大小比例,在80-110bp的大小范围内。在一些情况下,大小比例的截止值可以是约0.1、约0.5、约1、约2、约3、约4、约5、约6、约7、约8、约9、约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约25、约50、约100,或大于约100。在一些情况下,大小指标的截止值可以是约或至少10、约或至少2、约或至少1、约或至少0.5、约或至少0.333、约或至少0.25、约或至少0.2、约或至少0.167、约或至少0.143、约或至少0.125、约或至少0.111、约或至少0.1、约或至少0.091、约或至少0.083、约或至少0.077、约或至少0.071、约或至少0.067、约或至少0.063、约或至少0.059、约或至少0.056、约或至少0.053、约或至少0.05、约或至少0.04、约或至少0.02、约或至少0.001,或小于约0.001。

[0099] 可以确定核酸片段的大小分布的各种统计值。例如,可以使用大小分布的平均值、众数(mode)、中值或平均值。可以使用其他统计值,例如,给定大小的累积频率或不同大小的核酸片段数量的各种比率。累积频率可对应于给定大小或小于或大于给定大小的DNA片段的比例(例如百分比)。多种统计值提供有关核酸片段大小分布的信息,以便相对于一个或多个截止值比较,用于确定病原体引起的病症水平。可使用健康对象、已知具有一种或多种病症的对象、与病原体相关联的病症为假阳性的对象以及本文提及的其他对象的群组来确定截止值。本领域技术人员将知道如何基于本文的描述来确定这样的截止值。

[0100] 在一些示例中,病原体片段大小的第一统计值可以与人类基因组大小的参考统计值进行比较。例如,可以在第一统计值和参考统计值之间确定分离值(separation value)(例如,差值或比率),参考统计值例如从病原体参考基因组中的其他区域确定或从人类核酸确定。分离值也可以从其他值中确定。例如,可以根据多个区域的统计值来确定参考值。可将分离值与尺寸阈值进行比较以获得尺寸分类(例如,DNA片段是否较短、较长或与正常区域相同)。

[0101] 一些示例可计算一参数(分离值),所述参数可定义为参考病原体基因组和参考人类基因组之间短DNA片段比例的差异,使用以下等式:

$$[0102] \quad \Delta F = P(\leq 150\text{bp})_{\text{测试}} - P(\leq 150\text{bp})_{\text{参考}}$$

[0103] 其中 $P(\leq 150\text{bp})_{\text{测试}}$ 表示来自测试区域具有大小为 $\leq 150\text{bp}$ 的测序片段的比例, $P(\leq 150\text{bp})_{\text{参考}}$ 表示来自参考区域具有大小为 $\leq 150\text{bp}$ 的测序片段的比例。在其他实施例中,可以使用其他尺寸阈值,例如但不限于100bp、110bp、120bp、130bp、140bp、160bp和166bp。在其他实施例中,尺寸阈值可以用碱基、核苷酸或其他单位表示。

[0104] 可使用对照对象的平均值和SD值计算基于尺寸的z分数。

$$[0105] \quad \text{基于尺寸的 z 分数} = \frac{\Delta F_{\text{样本}} - \text{平均}\Delta F_{\text{对照}}}{SD\Delta F_{\text{对照}}}$$

[0106] 在一些实施例中,基于尺寸的z分数>3表明病原体的短片段比例增加,而基于尺寸的z分数<-3表明病原体的短片段比例减少。可以使用其他尺寸阈值。有关基于尺寸的方法的更多详细信息,请参见美国专利第8,620,593号和第8,741,811号以及美国专利申请案公开第2013/0237431号,其中每一项均以引用方式全部并入。

[0107] 为了确定核酸片段的大小,本公开的至少一些示例可以与任何单分子分析平台一起工作,在所述平台中可以分析染色体起源和分子长度,例如电泳、光学方法(例如,光学映射及其变异, [en.wikipedia.org/wiki/optical#umapping#cite#note-Nanocoding-3](http://en.wikipedia.org/wiki/optical#umapping#cite#note-Nanocoding-3),及Jo等人,美国科学院院报2007年;104:2673-2678),基于荧光的方法,基于探针的方法,数字PCR(基于微流体或基于乳状液,例如,BEAMing技术(Dressman等人,美国科学院院报2003年;100:8817-8822)、RainDance([www.raindancetech.com/technology/pcrgenomics\\_research.asp](http://www.raindancetech.com/technology/pcrgenomics_research.asp)))、滚环扩增、质谱、熔解分析(或熔解曲线分析(meltingcurve analysis))、分子筛等。作为质谱的一个示例,较长的分子将具有较大的质量(尺寸值的一个示例)。

[0108] 在一个示例中,可以使用双端测序协议对核酸分子进行随机测序。两端的两个读数可以映射(比对)到参考基因组,参考基因组可以被重复注释(例如,当与人类基因组比对时)。DNA分子的大小可以通过两个读数所对应的基因组位置之间的距离来确定。

[0109] 变异模式分析

[0110] 本公开的一些方面涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的变异模式的评估,对于对象发展出病原体相关联的疾病的风险的分层。在生物样本中检测到的病原体基因组的遗传变异可用于预测病原体相关联的疾病未来发展的风险。

[0111] 与来自无病原体相关联的疾病的对象的样本相比,患有病原体相关联的疾病(例如,病原体相关联的恶性肿瘤)的患者的病变组织中病原体核酸分子的变异模式可能不同。已被报道EBV相关联的肿瘤和对照样本中存在EBV品系(Palser等人, *J Virol.* 期刊2015年;89:5222-37)可能会有所不同。然而,在此先前的研究中,肿瘤和对照样本来自不同的地理位置。鉴于EBV变异的潜在地理变异,很难断定肿瘤样本中识别的变异是地理相关联的还是疾病相关联的。以前曾尝试通过分析鼻咽癌肿瘤样本来确定鼻咽癌相关联的EBV变异。在一项全基因组关联研究(GWAS)中(Hui等人, *Int J Cancer* 期刊2019年, [doi.org/10.1002/ijc.32049](https://doi.org/10.1002/ijc.32049))分析了来自同一地理区域无EBV相关联的疾病的个体的鼻咽癌肿瘤和唾液样本,有29个多态性(单核苷酸多态性(SNP)或插入或者缺失(indels))被识别为低于假发现率,调整后的P为0.05。这29个鼻咽癌相关联的EBV变异在90%以上的鼻咽癌病例中存在,但在对照病例中仅为40-50%。

[0112] 与发展为鼻咽癌的个别的EBV多态性分析相反(Hui等人, *Int J Cancer* 期刊2019年, [doi.org/10.1002/ijc.32049](https://doi.org/10.1002/ijc.32049);Feng等人, *Chin J Cancer* 期刊2015年;34:61),本公开的各个方面提供了以全基因组方式分析病原体核酸分子的变异模式的多种方法和多种系统。此外,不是通过分析肿瘤和细胞系样本来识别与疾病相关联的EBV变异(Palser等人, *J Virol.* 期刊2015年;89:5222-37,Correia等人, *J Virol.* 期刊2018年;92:e01132-18,Hui等

人,Int J Cancer期刊2019年,doi.org/10.1002/ijc.32049),与肿瘤的侵入性活检相比,本公开的各个方面提供了通过分析例如血液(例如血浆或血清)、鼻腔冲洗液、鼻腔刷子样本或通过非侵入性或微创程序获得的其他体液中的无细胞病原体核酸分子来分析病原体变异模式的多种方法和多种系统。在一个示例中,血液中EBV DNA分子的低丰度和片段性质可能对分析造成技术挑战。以非侵入性方式分析无细胞病毒DNA分子的变异模式可以改善临床应用(包含筛选、预测医学、风险分层、监测和预测)。在一个示例中,分析可用于区分具有不同病毒相关联的症状的对象,例如,在筛选背景下,具有可检测血浆EBV DNA的鼻咽癌患者和非鼻咽癌对象。在另一个示例中,它可以用于疾病或癌症的风险预测。

[0113] 可以使用不同的方法来获得不同的模式。非限制性的分析方法可包含大规模平行测序(MPS)、Sanger测序(如Lorenzetti等人,J Clin Microbiol.期刊2012年;50:609-18),以及基于微阵列的SNP分析(如Wang等人,美国科学院院报2002年;99:15687-92),杂交分析和质谱分析。在一个说明性示例中,使用例如具有捕获富集的靶向测序、MPS或Sanger测序的测序方法,并且基于每个核苷酸参考病原体的参考基因组(例如,EBV参考基因组)分析多个序列读数。所述方法可包含从对象的生物样本获得多个无细胞核酸分子的多个序列读数。所述方法还可包含将多个序列读数与病原体的参考基因组比对。所述方法还可包含通过分析病原体的参考基因组和映射到病原体的参考基因组的多个序列读数之间的核苷酸变异来分析横跨病原体的参考基因组的核苷酸变异模式。本文提供的所述变异模式表征在所述病原体的所述参考基因组上的多个变异位点中的每一个处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异。所述多个变异位点可以包含横跨所述病原体的所述参考基因组的至少30个、至少40个、至少50个、至少60个、至少70个、至少80个、至少90个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1000个、至少1100个,或至少1200个位点。所述多个变异位点包含横跨所述病原体的所述参考基因组的至少1000个位点。所述多个变异位点包含横跨所述病原体的所述参考基因组的约1100个位点。所述多个变异位点包含横跨所述病原体的所述参考基因组的至少600个位点。所述多个变异位点包含横跨所述病原体的所述参考基因组的约660个位点。在一些情况下,所述多个变异位点包含从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。在一些情况下,所述多个变异位点包含如表6所列相对于一EBV参考基因组(AJ507799.2)的一基因组位点。

[0114] 在一些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点是从表6所列的多个基因组位点中相对于EBV参考基因组(AJ507799.2)随机选择的。在一些情况下,本文提供的多种方法包含从表6所列的多个基因组位点中相对于一EBV参考基因组(AJ507799.2)随机选择的一步骤。所述方法还可包含通过分析病原体的参考基因组和映射到病原体的参考基因组的序列读数之间的核苷酸变异来分析横跨病原体的参考基因组的核苷酸变异模式。

[0115] 在一些情况下,来自所述病原体的所述多个无细胞核酸分子的所述变异模式表征在所述多个变异位点中的每个位点处映射到所述病原体的所述参考基因组的所述多个序列读数的一核苷酸变异,所述多个变异位点包含从表6所列的多个基因组位点中相对于EBV

参考基因组 (AJ507799.2) 随机选择的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600个位点。

[0116] 在一些情况下,所述多个变异位点由所有位点组成,在所述所有位点处,映射到所述病原体的所述参考基因组的所述多个序列读数具有与所述病原体的所述参考基因组的一不同的核苷酸变异。

[0117] 在一些情况下,使用野生型病原体基因组作为参考基因组。例如,一野生型 (wide type) EBV基因组 (GenBank:AJ507799.2) 可以用作参考EBV基因组。在其他情况下,使用其他病原体基因组作为参考基因组。在另一个示例中,使用多个病原体基因组 (例如EBV基因组) 作为参考。在另一个示例中,使用共有序列 (consensus sequence) 作为参考。可通过组合不同病原体基因组序列的变异来建立共有,例如,所述EBV基因组共有序列,例如,de Jesus等人, J Gen Virol期刊2003年;84:1443-50中描述的EBV基因组的共有序列。

[0118] 本文所提供的多种方法和多种系统中使用的序列比对 (例如,用于分析拷贝数、甲基化状态、片段大小、相对丰度或变异模式) 可通过任何适当的生物信息学算法、程序、工具包或软件包 (package) 来执行。例如,可以使用短寡核苷酸分析包 (short oligonucleotide analysis package, SOAP) 作为本文所提供的多种方法和多种系统的多种应用的比对工具。可在本文提供的多种方法和多种系统中使用的短序列读数分析工具的示例包含Arioc、BarraCUDA、BBMap、BFAST、BigBWA、BLASTN、BLAT、Bowtie、Bowtie2、BWA、BWA-PSSM、CASHX、Cloudburst、CUDA-EC、CUSHAW、CUSHAW2、CUSHAW2-GPU、CUSHAW3、drFAST、ELAND、ERNE、GASSST、GEM、Genalice MAP、Geneious Assembler、GensearchNGS、GMAP及GSNAP、GNU MAP、HIVE-hexagon、Isaac、LAST、MAQ、mrFAST、mrsFAST、MOM、MOSAIC、MPscan、Novoalign& NovoalignCS、NextGENe、NextGenMap、Omixon Variant Toolkit、PALMapper、Partek Flow、PASS、PerM、PRIMEX、QPalma、RazerS、REAL、cREAL、RMAP、rNA、RTG Investigator、Segemehl、SeqMap、Shrec、SHRiMP、SLIDER、SOAP、SOAP2、SOAP3、SOAP3-dp、SOCS、SparkBWA、SSAHA、SSAHA2、Stampy、SToRM、Subread、Subjunc、Taipan、UGENE、VelociMapper、XpressAlign、及 ZOOM。

[0119] 序列读数中的一数量的连续的核苷酸 (“序列段 (sequence stretch)”) 可用于与参考基因组比对,以发出关于比对的调用。例如,所述比对可包含比对参考基因组 (例如病原体的参考基因组或宿主生物的参考基因组) 的序列读数的至少4、至少6、至少8、至少10、至少12、至少14、至少16、至少18、至少20、至少22、至少24、至少25、至少26、至少28、至少30、至少32、至少34、至少35、至少36、至少38、至少40、至少42、至少44、至少45、至少46、至少48、至少50、至少52、至少54、至少55、至少56、至少58、至少60、至少62、至少64、至少65、至少66、至少67、至少68、至少69、至少70、至少71、至少72、至少73、至少74、至少75、至少76、至少78、至少80、至少82、至少84、至少85、至少86、至少88、至少90、至少92、至少94、至少95、至少96、至少98、至少100、至少102、至少104、至少106、至少108、至少110、至少112、至少114、至少116、至少118、至少120、至少122、至少124、至少126、至少128、至少130、至少132、至少134、至少136、至少138、至少140、至少142、至少145、至少146、至少148或至少150个连续的核苷酸。在一些情况下,如本文所述的比对可以包含比对参考基因组 (例如病原体的参考基因组或宿主生物的参考基因组) 的序列读数的至多5、至多7、至多9、至多11、至多13、至多15、至多17、至多19、至多21、至多23、至多25、至多27、至多29、至多31、至多33、至多35、至多37、至

多39、至多41、至多43、至多45、至多47、至多49、至多51,最多53、最多55、最多57、最多59、最多61、最多63、最多65、最多67、最多68、最多69、最多70、最多71、最多72、最多73、最多74、最多75、最多76、最多78、最多80、最多81、最多83、最多85、最多87、最多89、最多91、最多93、最多95、最多97、最多99,最多101、最多103、最多105、最多107、最多109、最多111、最多113、最多115、最多117、最多119、最多121、最多123、最多125、最多127、最多129、最多131、最多133、最多135、最多137、最多139、最多141、最多143、最多145、最多147、最多149,或最多151个连续的核苷酸。在一些示例中,如本文所述的比对包含比对参考基因组(例如病原体的参考基因组或宿主生物的参考基因组)的序列读数的约20、约22、约24、约25、约26、约28、约30、约32、约34、约35、约36、约38、约40、约42、约44、约45、约46、约48、约50、约52、约54、约55、约56、约58、约60、约62、约64、约65,约66、约67、约68、约69、约70、约71、约72、约73、约74、约75、约76、约78、约80、约82、约84、约85、约86、约88、约90、约92、约94、约95、约96、约98、约100、约102、约104、约106、约108、约110、约112、约114、约116、约118,约120、约122、约124、约126、约128、约130、约132、约134、约136、约138、约140、约142、约145、约146、约148、约150、约152、约154、约155、约156、约158、约160、约162、约164、约165、约166、约168、约170、约172、约174、约175,约176、约178、约180、约185、约190、约195或约200个连续的核苷酸。

[0120] 在某些情况下,当序列段在整个序列读数上与参考基因组(例如,人类参考基因组)的特定区域具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或100%的序列同一性或互补性时,进行比对调用。在某些情况下,当序列段在整个序列读数中与参考基因组(例如,人类参考基因组)的特定区域具有至少80%的序列同一性或互补性时,进行比对调用。在某些情况下,当序列段与参考基因组(例如,人类参考基因组)的特定区域相同或互补,具有不超过20、15、10、9、8、7、6、5、4、3、2或1碱基的不匹配(mismatch),或具有零个不匹配时,进行比对调用。在某些情况下,当序列段与参考基因组(例如人类参考基因组)的特定区域相同或互补,且不超过2个碱基的不匹配时,进行比对调用。最大的不匹配数或百分比,或最小相似数或百分比可以随选择标准而变化,这取决于本文提供的多种方法和多种系统的应用目的和背景。

[0121] 在某些情况下,序列读数与病原体的参考基因组的比对允许最大不匹配不超过20、15、10、9、8、7、6、5、4、3、2或1个碱基。映射的序列读数与病原体的参考基因组之间的不匹配可表明存在生物样本中病原体基因组序列中的核苷酸变异,在其他情况下,也可表明测序错误。不希望受到某一理论的约束,在一个生物样本中,在给定的基因组位点处识别出的一个以上的核苷酸变异可能是源自因为无细胞病原体核酸分子来源的病变细胞的测序错误或异质性。在某些情况下,如果在给定的生物样本中识别出超过1、2或3个核苷酸变异,则在一基因组位点处多个核苷酸变异被排除在分析之外。

[0122] 在一个说明性示例中,使用捕获富集的靶向测序来分析鼻咽癌对象和非鼻咽癌对象循环中具有可检测的血浆EBV DNA的无细胞病毒DNA分子。多种捕获探针可以设计成覆盖整个EBV基因组。在其他情况下,可以只分析EBV基因组的一部分,多种捕获探针被设计成只能覆盖EBV基因组的一部分。在相同的分析中,多种捕获探针也可以包含在人类基因组的标靶基因组区域。例如,多种探针可以包含标靶人类常见的单核苷酸多态性(SNP)位点和多种人类白细胞抗原(HLA)的SNP。在一个实施例中,更多探针可设计成与其他病毒基因组序列

杂交,例如,HPV或HBV基因组。

[0123] 在某些情况下,通过直接比较映射到参考基因组的序列读数和参考基因组来分析病原体基因组的变异模式。可以以任何适当的方式进一步处理比较结果,例如,用于聚类分析或系统发育树分析。用于这些分析的可用生物信息学工具包含MEGA4、MEGA5、CLUSTALW、Phylip、RAxML、BEAST、PhyML、TreeView、MAFFT、MrBayes、BIONJ、MLTreeMap、Newick实用程序(Newick Utilities)、Phylo.io、Phylogeny.fr、REALPHY、SuperTree和系统发育网络重复器(The PhyloGenetic Web Repeater,POWER)。聚类分析或系统发育树分析将映射到病原体参考基因组的序列读数与一个或多个从患病组织或健康对象获得的,或表明为能够或不能引起病原体相关联的疾病的,或表明为对引起病原体相关联的疾病有效或无效的病原体基因组进行比较。

[0124] 在说明性示例中,本文提供的多种方法和多种系统包含以区块为基础的变异模式分析。以区块为基础的变异模式分析可包含将病原体的参考基因组进行分层到多个箱(“区块”)中。将映射到病原体参考基因组的多个序列读数与相对于多个箱中的每个箱内的疾病相关联的病原体基因组进行比较。在某些情况下,存在多个(例如,至少2、3、4、5、6、7、8、9、10、12、14、16、18、20、22、24、26、28、30、40、50、60、70、80、90、100、120、140、160、180、200、300、400、500、600、700、800、900或1000个)不同的病原体基因组与以区块为基础的分析进行比较,所述以区块为基础的分析包含与疾病相关联的病原体基因组,以及可选地,已知或表明为不能导致病原体相关联的疾病或对导致病原体相关联的疾病不起作用的病原体基因组(与疾病无关的病原体基因组)。在以区块为基础的分析中,在多个箱中的每个箱中,基于映射到病原体参考基因组的序列读数与多个疾病相关联的病原体基因组中的每个或与疾病无关的多个病原体基因组之间的共享核苷酸变异来计算相似性指数。相似性指数可取决于多个变异位点的比例,其中映射到多个病原体参考基因组的多个序列读数中的至少一个处具有与疾病相关联的或疾病无关的病原体基因组相同的核苷酸变异。基于相对于多个序列读数比较的多个病原体基因组中的每一个的相似性指数,可以基于例如通过相似性指数所反映的相似性水平来计算箱分数。在一个示例中,箱分数可以取决于高于预定截止值的多个相似性指数的比例。相似性指数可以有一个截止值集,例如,约0.6、0.7、0.75、0.8、0.85、0.9或0.95。高于截止值的相似性指数可以表明序列读数与跟所述序列读数相比的病原体基因组的“相似”。基于上述分析,然后可以使用计算出的多个相似性指数或多个箱分数在横跨病原体基因组或部分病原体基因组的较大尺度上执行模式分析。类似于上述的聚类分析或系统发育分析可以遵循以区块为基础的分析来预测病原体相关联的疾病(如EBV相关联的鼻咽癌)的发展风险。

[0125] 风险分数

[0126] 本公开的一些方面涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的一个或多个特征的组合考量,对于对象发展出病原体相关联的疾病的的分层。在某些情况下,会生成一风险分数,表明对象发展出病原体相关联的疾病(例如EBV相关联的鼻咽癌)的风险。

[0127] 在某些情况下,本公开涉及基于对来自对象的生物样本中来自病原体的多个无细胞核酸分子的一个或多个特征的组合考量,以及一个或多个因素:对象的年龄、对象的吸烟习惯、对象的鼻咽癌家族史、对象的多个基因型因素、饮食史或对象的种族,对于对象发展

出病原体相关联的疾病的风险的分层。在没有临床检测到鼻咽癌的对象中,血浆EBV DNA检测的阳性率与对象的年龄之间可能存在正相关。对象的吸烟习惯会增加对象患鼻咽癌的风险。有鼻咽癌家族史的对象自身发生鼻咽癌的风险较高。基因型因素也可能与鼻咽癌的风险相关,如HLA状态,如Bei等人,Nat Genet.期刊,2010年;42:599-603和Hildesheim等人,J Natl Cancer Inst.期刊,2002年;94:1780-9,其中每一个都完整地并入在本文中。此外,饮食史可能与患鼻咽癌的风险相关,例如,食用大量咸鱼的对象患鼻咽癌的风险相对较高。某些种族,如广东人,也可能与发展为鼻咽癌的高风险相关。

[0128] 在一些情况下,所述多种方法和多种系统还包含生成表明对象发展出病原体相关联的疾病的风险的报道。此类报道可以具有数字风险分数值或明确的风险评估。在某些情况下,报道包含筛选频率或后续筛选分析的未来时间点的建议。报道可提供给对象、为对象提供服务的医疗机构或医疗专业人员,或任何相关第三方,如医疗保险公司。在报道发布之前或之后,可由认证医生对报道进行审查、评估或编辑。在某些情况下,认证医生会根据他/她的医学意见或多种独立的检查,对风险评估提供额外的意见或参与最终风险评估。

[0129] 在一些情况下,本公开提供了通过使用一分类器对发展出一病原体相关联的疾病的风险进行分层的多种方法,例如病原体相关联的增殖性疾病,例如EBV相关联的鼻咽癌。此类分类器可将本文所述的一个或多个因素作为数据输入,并提供包含风险分数的输出,其可表明对象发展出病原体相关联的疾病的风险。可输入分类器的一个或多个因素可包含多个无细胞病原体核酸分子的一个或多个特征、来自对象的生物样本中来自病原体的多个无细胞核酸分子的一个或多个特征,以及一个或多个因素:对象的年龄、对象的吸烟习惯、对象的鼻咽癌家族史、对象的基因型因素、饮食史和对象的种族。作为分类器的输出的风险分数可以表明对象当前患有或将来发展为病原体相关联的疾病的风险。在某些情况下,风险分数表明对象目前可能患有病原体相关联的疾病。在某些情况下,风险分数表示对象在未来一段时间内(例如但不限于1年、2年、3年、4年、5年、10年或15年)发展出病原体相关联的疾病的可能性。在某些情况下,分类器提供的输出包含推荐的筛选频率或后续筛选分析的未来时间点。这样的输出可以是临床建议的形式,也可以在如上所述的报道中提供给对象、医疗机构或医疗专业人员,或任何第三方,如医疗保险公司。

[0130] 如本文所述,分类器可以指实行分类的任何算法。在本公开中,分类器可以是基于用于预测病原体相关联的疾病未来发展风险的任何适当算法构建的分类模型。适当的算法可以包含机器学习算法和其他数学/统计模型,例如但不限于支持向量机(SVM)、朴素贝叶斯(Naïve Bayes)、逻辑回归、随机森林、决策树、梯度提升树、神经网络、深度学习、线性/核支持向量机(SVM)、线性/非线性回归、线性判别分析等。在某些情况下,分类器是使用包含多个输入-输出对的标记的数据集进行训练。例如,从已经被诊断为无鼻咽癌或患有鼻咽癌的一数量的对象的样本的分析结果生成的数据集。在这些情况下,数据集可包含具有来自这些对象的血浆EBV DNA特征(例如,变异模式、甲基化状态、可检测性/拷贝数或片段大小)、年龄、家族史、吸烟习惯、种族或饮食史的一个或多个因素的输入,以及相应的输出,表明相应的对象是否具有鼻咽癌。在说明性示例中,可以使用包含大量输入-输出对(例如至少10、20、50、100、200、500、1000、2000、5000、10000或20000对)的标记的数据集来训练分类器。

[0131] 在一个示例中,提供了一个分类模型,以使用变异模式的分析来预测具有可检测

血浆EBV DNA的对象未来鼻咽癌发展的风险。分类模型可以是使用支持向量机(SVM)算法如下构造的分类器:

[0132] 给定包含n个样本的训练数据集:

[0133]  $(M_1, Y_1), \dots, (M_n, Y_n)$

[0134] 其中 $Y_i$ 表示样本i的鼻咽癌状态。 $Y_i$ 对于来自鼻咽癌患者的样本为1)或对于来自无鼻咽癌的对象样本为-1; $M_i$ 是一p维向量,包含样本i的病毒变异模式。例如, $M_i$ 可以是一系列的多个变异位点(例如,表6中所列的与鼻咽癌相关联的29个变异位点或与鼻咽癌相关联的661个变异位点)。替代地, $M_i$ 可以是关于已知患有鼻咽癌的对象中存在的多个参考EBV变异的一系列以区块为基础的变异相似性分数(例如,500bp的非重叠窗口)。

[0135] 可以识别一“超平面”,所述“超平面”在训练数据集中尽可能准确地将非鼻咽癌组和鼻咽癌组分开,通过寻找一系数集(W和p维向量)满足以下条件:

[0136] 准则1:

[0137]  $W \cdot M_i - b \geq 1$  (对于鼻咽癌群组中的任何对象)

[0138] 及

[0139] 准则2

[0140]  $W \cdot M_i - b \leq -1$  (对于非鼻咽癌群组中的任何对象)

[0141] 其中W是确定超平面的多个系数的p维向量; $M$ 是一矩阵(p x n维),具有p个变量(或多个以区块为基础的相似性分数)和n个样本;b是截距。

[0142] 这两个准则(即准则1和2)也可以写成:

[0143]  $Y_i (W \cdot M_i - b) \geq 1$  ( $\geq$  准则3)

[0144] 其中 $Y_i$ 为-1(非鼻咽癌)或1(鼻咽癌)。

[0145] 准则1和2之间的边界距离(D)为:  $\frac{2}{\|W\|}$

[0146] 其中 $\|W\|$ 是使用点到平面方程的距离计算。

[0147] 根据准则3,通过最小化 $\|W\|$ 使D最大化。

[0148] 基于这一原理,可以确定分类器的参数(W和b)。因此,使用经训练的参数(W和b)实行的经训练的分类器可用于计算多个测试样本的鼻咽癌风险分数。

[0149] 在一个示例中,鼻咽癌风险分数被计算为横跨病毒基因组中SNV位点的固定集处的多个EBV基因型的加权总和(作为二元逻辑回归模型中的解释变量)。在所述示例中,通过分析训练集中来自鼻咽癌和非鼻咽癌样本的多个EBV SNV基因谱(profile)中的差异来识别鼻咽癌相关联的SNV集。可以分析EBV基因组中每个变异与鼻咽癌病例的关联,例如使用Fisher精确检验(Fisher's exact test)。然后可以获得的显著的SNV的固定集,例如,具有假发现率(false discovery rate, FDR)控制在5%。测试样本的鼻咽癌风险分数可通过其多个EBV基因型在从训练集识别的多个显著的SNV位点(significant SNV sites)的特定集上来确定,所述训练集包含从已知鼻咽癌和非鼻咽癌对象的血浆DNA样本的测序数据。在某些情况下,血浆中EBV DNA分子的浓度可能较低,因此通过多个测序的EBV DNA读数可以不完全覆盖整个EBV基因组。所述分数可通过多个血浆EBV DNA读数(例如,具有可用的基因型信息)所覆盖的多个SNV位点的基因型模式来确定。为了得出鼻咽癌风险分数,可以首先识别样本中多个血浆EBV DNA读数所覆盖的显著的SNV位点子集,然后在所述显著的SNV位点

子集内确定每个位点的多个基因型的权重(效应大小)。可构建如下的逻辑回归模型,以提供鼻咽癌每个SNV位点处多个风险基因型的影响大小的信息:

$$[0150] \quad P = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^n \beta_k X_k)}}$$

[0151] 其可以被重写为:

$$[0152] \quad \text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{k=1}^n \beta_k X_k,$$

[0153] 其中n是显着的SNV位点的数量 $\beta_0$ ;及 $\beta_k$ 是可以通过最大似然性估计器确定的多个系数;P为EBV阳性患者发生鼻咽癌的概率;此变量 $X_k$ 代表基因组位置k处的SNV位点。如果样本中存在与EBV参考基因组相同的变异,则 $X_k$ 可编码为-1。如果样本中存在替代变量,则 $X_k$ 可编码为1。如果样本中未包含分析的变异位点,则 $X_k$ 可将其编码为0。因此,系数 $\beta_0$ 和 $\beta_k$ 可以被估计,例如,使用python中的‘逻辑回归(Logistic Regression)’函数。这可以通过分析训练数据集中鼻咽癌和非鼻咽癌样本中每个位点的基因型模式来实行。因此,测试样本的鼻咽癌风险分数可以根据多个SNV位点处的其拥有的基因型得出,并通过从训练模型中推导出相应的系数 $\beta_0$ 和 $\beta_k$ 进行加权。

[0154] 生物样本

[0155] 在本文所提供的多种方法中使用的生物样本可包含从活的或死的对象获得的任何组织或材料。生物样本可以是无细胞样本。生物样本可包含核酸(例如,DNA或RNA)或其片段。样本中的核酸可以是无细胞核酸。样本可以是液体样本或固体样本(例如,细胞或组织样本)。生物样本可以是体液,例如血液、血浆、血清、尿液、口腔冲洗液、鼻腔冲洗液、鼻刷样本、阴道液、鞘膜积液(例如睾丸鞘膜积液)、阴道冲洗液、胸膜液、腹水、脑脊液、唾液、汗液、泪液、痰、支气管肺泡灌洗液、乳头排出液,从身体不同部位(如甲状腺、乳房)的吸取液等。也可以使用粪便样本。在各种示例中,已被富集的非细胞DNA的生物样本中的大部分DNA(例如,通过离心方案获得的血浆样本)可以是无细胞的(例如,大于50%、60%、70%、80%、90%、95%或99%的DNA可以是无细胞的)。生物样本可经处理以物理破坏组织或细胞结构(例如,离心和/或细胞裂解),从而将细胞内组分释放到溶液中,所述溶液可进一步包含用于制备样本以供分析的酶、缓冲液、盐、洗涤剂等。

[0156] 本文提供的多种方法和多种系统可用于分析生物样本中的核酸分子。所述核酸分子可以是多个细胞核酸分子、多个无细胞核酸分子或两者。通过本文所提供的多种方法使用的无细胞核酸可以是生物样本中细胞外的核酸分子。无细胞核酸分子可以存在于各种体液中,例如血液、唾液、精液和尿液。无细胞DNA分子可以由于健康状况和/或疾病(如病毒感染和肿瘤生长)引起的各种组织中的细胞死亡来产生。多个无细胞核酸分子可以包含由病原体整合事件产生的序列。

[0157] 本文所述方法中使用的多个无细胞核酸分子,例如无细胞DNA,可存在于血浆、尿液、唾液或血清中。无细胞DNA可以以短片段的形式自然产生。无细胞DNA片段化是指当产生或释放无细胞DNA分子时,高分子量DNA(如细胞核中的DNA)被切割、断裂或消化成短片段的过程。本文提供的多种方法和多种系统可用于在某些情况下分析细胞核酸分子,例如,来自肿瘤组织的细胞DNA,或者当病人患有白血病、淋巴瘤或骨髓瘤时来自白细胞的细胞DNA。根

据本公开的一些示例,可以对取自肿瘤组织的样本进行多种分析(assays及analyses)。

[0158] 对象

[0159] 本文提供的多种方法和多种系统可用于分析来自对象(例如生物体,例如宿主生物体)的样本。对象可以是任何人类患者,例如癌症患者、有癌症风险的患者或有癌症家族或个人疾病史的患者。在某些情况下,对象处于癌症治疗的特定阶段。在某些情况下,对象可能患有或怀疑患有癌症。在某些情况下,对象是否患有癌症未知。

[0160] 在某些情况下,根据本文提供的筛选分析结果,对象接受或不接受病原体相关联的疾病的的治疗。在一个示例中,虽然第一筛选分析显示阳性结果,表明对象发展为病原体相关联的疾病的高风险,但通过后续诊断检查,对象被诊断为没有病原体相关联的疾病(例如,EBV相关联的鼻咽癌)。在这种情况下,对象没有接受医疗治疗,例如但不限于使用治疗剂(例如,化疗)、放疗、手术或其任何组合进行治疗。在另一个示例中,对象被筛选为具有发展出病原体相关联的疾病(例如,HPV相关联的宫颈癌)的高风险,并进一步被诊断为患有所述疾病。结果,对象可接受所述疾病的医疗治疗,例如但不限于手术、化疗、放疗、靶向治疗、免疫治疗或其任何组合。

[0161] 本文提供的多种方法和多种系统可适用的病原体相关联的疾病可包含增殖性疾病,例如癌症。这些疾病可能与病毒、细菌或真菌等病原体有关联或由其引起。可与本文所述疾病相关联的病毒可包含EBV、卡波西肉瘤相关联的疱疹病毒(KSHV)、HPV(例如但不限于HPV 16、18、31、33、34、35、39、45、51、52、56、58、59、66、68和70)(Burd等人,Clin Microbiol Rev期刊2003年:16:1-17)、默克尔细胞多瘤病毒(MCPV)、HBV、HCV和人类T淋巴细胞营养病毒-1(HTLV1)。适用的病原体相关联的癌症包含Burkitts淋巴瘤、Hodgkins淋巴瘤、免疫抑制相关淋巴瘤、T细胞和NK细胞淋巴瘤;鼻咽癌或胃癌,可能与EB病毒有关联。适用的病原体相关联的癌症可包含原发性渗出性淋巴瘤或卡波西肉瘤,其可与KSHV相关联。适用的病原体相关联的癌症可包含宫颈癌、头颈癌或肛门生殖道癌,其可与HPV相关联。适用的病原体相关联的癌症可包含与MCPV相关联的默克尔细胞癌。适用的病原体相关联的癌症可包含与HBV或丙型肝炎病毒(HCV)相关联的HCC。适用的病原体相关联的癌症可包含与HTLV1相关联的成人T细胞白血病/淋巴瘤。

[0162] 对象可能患有任何类型的癌症或肿瘤,或有患任何类型癌症或肿瘤的风险。例如,对象可能患有鼻咽癌或鼻腔的癌症。在另一个示例中,对象可能患有口咽癌或口腔的癌症。癌症的非限制性示例可包含但不限于肾上腺癌、肛门癌、基底细胞癌、胆管癌、膀胱癌、血癌、骨癌、脑瘤、乳腺癌、支气管癌、心血管系统癌、宫颈癌、结肠癌、结直肠癌、消化系统癌症、内分泌系统癌症、子宫内膜癌、食管癌、眼癌、胆囊癌、胃肠道肿瘤、肝细胞癌、肾癌、造血系统恶性肿瘤、喉癌、白血病、肝癌、肺癌、淋巴瘤、黑色素瘤、间皮瘤、肌肉系统癌症、骨髓增生异常综合征(MDS)、骨髓瘤、鼻腔癌、鼻咽癌、神经系统癌、淋巴系统癌、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、阴茎癌、垂体瘤、前列腺癌、直肠癌、肾盂癌、生殖系统癌、呼吸系统癌、肉瘤、涎腺癌、骨骼系统癌、皮肤癌、小肠癌、胃癌、睾丸癌、喉癌、胸腺癌、甲状腺癌、甲状腺癌、泌尿系统癌、子宫癌、阴道癌或外阴癌。淋巴瘤可以是任何类型的淋巴瘤,包含B细胞淋巴瘤(例如,弥漫性大B细胞淋巴瘤、滤泡性淋巴瘤、小淋巴细胞淋巴瘤、套细胞淋巴瘤、边缘区B细胞淋巴瘤、伯基特淋巴瘤、淋巴浆细胞淋巴瘤、毛细胞白血病或原发性中枢神经系统淋巴瘤)或T细胞淋巴瘤(例如,前体T淋巴细胞淋巴瘤或外周T细胞淋巴瘤)。白血病可以

是任何类型的白血病,包含急性白血病或慢性白血病。白血病的类型包含急性髓系白血病、慢性髓系白血病、急性淋巴细胞白血病、急性未分化白血病或慢性淋巴细胞白血病。在某些情况下,癌症患者没有患有特定类型的癌症。例如,在某些情况下,患者可能患有不是乳腺癌的癌症。

[0163] 癌症的示例包含引起实体肿瘤的癌症和不引起实体肿瘤的癌症。此外,本文提及的任何癌症可以是原发性癌症(例如,以其最初开始生长的身体部分命名的癌症)或继发性或转移性癌症(例如,起源于身体另一部分的癌症)。

[0164] 通过本文所述的任何方法诊断的对象可以是任何年龄的,并且可以是成年人、婴儿或儿童。在某些情况下,对象是0、1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、68、69、70、71、78、78、78、78、78、80、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98或99岁,或在一年龄范围内(例如,2岁至20岁、20岁至40岁或40岁至90岁)。可以受益的一特定类别的病人可以是40岁以上的病人。可以受益的另一特定类别的病人可以是儿科患者。此外,通过本文所述的任何方法或组合诊断的对象可以是男性或女性。

[0165] 在一些实施例中,本公开的方法可检测对象中的肿瘤或癌症,其中所述肿瘤或癌症具有疾病的地理模式。在一个示例中,对象可以有EBV相关联的癌症(例如鼻咽癌),这在中国南方地区(例如香港特区)是普遍存在的。在另一个示例中,对象可能患有HPV相关的癌症(例如口咽癌),这种癌症可能在美国和西欧流行。在另一个示例中,对象可能患有HTLV-1相关癌症(例如,成人T细胞白血病/淋巴瘤),这种癌症可能在日本南部、加勒比、中非、南美洲部分地区以及美国东南部的一些移民群组中流行。

[0166] 本文公开的任何方法也可对非人类对象(例如实验室或农场动物)或源自本文公开的生物体的细胞样本执行。所述非人类对象的非限制性示例包含狗、山羊、豚鼠、仓鼠、老鼠、猪、非人类灵长类动物(例如大猩猩、猿、猩猩、狐猴或狒狒)、老鼠、绵羊、奶牛或斑马鱼。

[0167] 计算机系统

[0168] 本文公开的任何方法可由一个或多个计算机系统执行和/或控制。在一些示例中,本文公开的多种方法的任何步骤都可以由一个或多个计算机系统完全、单独或顺序地执行和/或控制。本文提到的任何计算机系统都可以利用任何适当数量的子系统。在一些实施例中,计算机系统包含单个计算机设备,其中子系统可以是计算机设备的多个组件。在其他实施例中,计算机系统可以包含多个计算机设备,每个计算机设备都是具有多个内部组件的子系统。计算机系统可以包含台式机和笔记本电脑、平板电脑、移动电话和其他移动设备。

[0169] 子系统可通过系统总线互连。其他子系统包含打印机、键盘、存储设备和耦合到显示适配器的显示器。耦合到I/O控制器的外围装置和输入/输出(I/O)设备可以通过本领域已知的任意数量的连接连接到计算机系统,例如输入/输出(I/O)端口(例如USB、FireWire®)。例如,I/O端口或外部接口(如以太网、Wi-Fi等)可用于将计算机系统连接到广域网,如互联网、鼠标输入设备或扫描仪。经由系统总线的互连允许中央处理器与每个子系统通信,并控制来自系统存储器或存储设备(例如,固定磁盘(例如硬盘驱动器或光盘))的多条指令的执行,以及多个子系统之间的信息交换。系统存储器和/或存储设备可以实施为计算机可读介质。另一子系统是数据收集设备,例如摄像机、麦克风、加速度计等。本文提

到的任何数据都可以从一个组件输出到另一个组件,并且可以输出到用户。

[0170] 计算机系统可以包含多个相同的组件或子系统,例如,通过外部接口或内部接口连接在一起。在一些实施例中,计算机系统、子系统或设备可以通过网络进行通信。在这种情况下,一台计算机可以被视为客户端,另一台计算机可以被视为服务器,其中每台计算机都可以是同一计算机系统的一部分。客户机和服务器可以分别包含多个系统、子系统或组件。

[0171] 本公开提供了计算机控制系统,其被编程以实行本公开的用于对病原体相关联的疾病的风险进行分层的方法。图21示出了计算机系统1101,其被编程或以其他方式配置为分析多个无细胞核酸分子或其序列读数,分析与疾病风险相关联的其他因素,评估风险,或生成表明本文所述风险的报道。计算机系统1101可以实行和/或调节本公开中提供的方法的各个方面,例如,控制来自生物样本的核酸分子的测序,执行如本文所述的测序数据的生物信息学分析的各个步骤,集成数据收集,分析和结果报道,以及数据管理。计算机系统1101可以是用户的电子装置或相对于所述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。

[0172] 计算机系统1101包含中央处理单元(CPU,本文也称为“处理器”和“计算机处理器”)1105,其可以是单核或多核处理器,或者用于并行处理的多个处理器。计算机系统1101还包含存储器或存储器位置1110(例如,随机存取存储器、只读存储器、闪存)、电子存储单元1115(例如,硬盘)、用于与一个或多个其他系统通信的通信接口1120(例如,网络适配器)以及外围装置1125,例如高速缓存、其他存储器、数据存储器 and/或电子显示适配器。存储器1110、存储单元1115、接口1120和外围装置1125通过例如主板的通信总线(实体线路)与CPU 1105相通信。存储单元1115可以是用于存储数据的数据存储单元(或数据存储库)。计算机系统1101可以借助通信接口1120操作地耦合到计算机网络(“网络”)1130。网络1130可以是因特网、互联网和/或外联网,或者与因特网通信的内联网和/或外联网。在某些情况下,网络1130是电信和/或数据网络。网络1130可以包含一个或多个计算机服务器,其可以实行分布式计算,例如云计算。在某些情况下,借助计算机系统1101、网络1130可以实行对等网络,其可以使耦合到计算机系统1101的多个装置充当客户端或服务器。

[0173] CPU 1105可执行机器可读指令序列,其可实施在程序或软件中。指令可以存储在存储器位置中,例如存储器1110。这些指令可引导到CPU 1105,其随后可编程或以其他方式配置CPU 1105以实行本公开的方法。CPU 1105执行的操作的示例可以包含获取、解码、执行和写回。

[0174] CPU 1105可以是电路的一部分,例如集成电路。系统1101的一个或多个其他组件可以包含在电路中。在某些情况下,所述电路是专用集成电路(ASIC)。

[0175] 存储单元1115可以存储文件,例如驱动程序、文库和保存的程序。存储单元1115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1101可以包含计算机系统1101外部的一个或多个附加数据存储单元,例如位于通过内部网或因特网与计算机系统1101通信的远程服务器上。

[0176] 计算机系统1101可以通过网络1130与一个或多个远程计算机系统通信。例如,计算机系统1101可以与用户的远程计算机系统(例如,安装有接收和显示从计算机系统1101发送的样本分析结果的应用程序的智能电话)通信。远程计算机系统的示例包含多种个人

计算机(如便携式PC)、平板电脑(slate PC)或平板电脑(tablet PC)(如 Apple®iPad、Samsung®GalaxyTab)、电话、智能手机(如 Apple®iPhone、安卓装置、Blackberry®)或个人数字助理。用户可以经由网络1130访问计算机系统1101。

[0177] 本文所述的多种方法可以通过存储在计算机系统1101的电子存储位置(例如,存储器1110或电子存储单元1115)上的机器(例如,计算机处理器)可执行代码来实行。机器可执行或机器可读代码可以以软件的形式提供。在使用期间,所述代码可由处理器1105执行。在一些情况下,可以从存储单元1115检索代码并将其存储在存储器1110上,以供处理器1105随时访问。在一些情况下,可以排除电子存储单元1115,并且机器可执行指令存储在存储器1110上。

[0178] 代码可以预编译并配置为与具有适于执行代码的处理器机器一起使用,或者可以在运行时编译。代码可以用编程语言提供,可以选择编程语言使代码以预编译或编译方式执行。

[0179] 本文提供的系统和方法的多个方面,例如计算机系统1101,可以在编程中实行。技术的各个方面可以被认为典型地以机器(或处理器)可执行代码和/或在一种机器可读介质中承载或实施的相关数据的形式“产品”或“制品”。机器可执行代码可存储在电子存储单元上,例如存储器(例如,只读存储器、随机存取存储器、闪存)或硬盘。“存储器”类型的介质可包含计算机、处理器等的任何或所有有形的存储器,或其相关联的模块,例如各种半导体存储器、磁带机、磁盘机等,其可随时为软件编程提供非暂时性存储。软件的全部或部分有时可通过因特网或各种其他电信网络进行通信。例如,这种通信可以使得能够将软件从一台计算机或处理器加载到另一台计算机或处理器,例如从管理服务器或主机加载到应用服务器的计算机平台。因此,可承载软件元件的另一种类型的介质包含光波、电波和电磁波,例如通过有线和光纤固网网络以及各种空中链结多多个本地装置之间的物理接口使用。携带这种波的物理元件,例如有线或无线链结、光链结等,也可以被视为承载软件的介质。如本文所使用的,除非限于非暂时性、有形的“存储”介质,否则例如计算机或机器的“可读介质”的术语是指参与向处理器提供指令以供执行的任何介质。

[0180] 因此,例如计算机可执行代码的机器可读介质可以采取多种形式,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含例如光盘或磁盘,例如任何计算机中的任何存储装置等,例如可用于实行附图中所示的数据库等。易失性存储介质包含动态存储器,例如计算机平台的主存储器。有形传输介质包含同轴电缆;铜线和光纤,包含计算机系统总线中的总线导线。载波传输介质的形式可以是电信号或电磁信号,或声波或光波,例如在射频(RF)和红外(IR)数据通信期间产生的那些。因此,计算机可读介质的常见形式包含例如:软盘(floppy disk、flexible disk)、硬盘、磁带、任何其他磁性介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、穿孔卡纸带、任何其他具有孔图案的物理存储介质、RAM、ROM、PROM和EPROM、闪存-EPROM、,任何其他存储芯片或盒带、传输数据或指令的载波、传输此类载波的电缆或链结,或计算机可从中读取编程代码和/或数据的任何其他介质。许多这些形式的计算机可读介质中可能涉及将一个或多个指令的一个或多个序列携带到处理器以执行。

[0181] 计算机系统1101可包含电子显示器1135,或与电子显示器1135通信,电子显示器

1135包含用户界面(UI) 1140,用于提供例如样本分析结果,例如但不限于病原体整合简档的图形显示、病原体整合断点的基因组位置、病症分类(例如,疾病或癌症类型和癌症级别),以及基于病症分类的治疗建议或预防措施建议。UI的示例包含但不限于图形用户界面(GUI)和基于web的用户界面。

[0182] 本公开的多种方法和多种系统可以通过一个或多个算法来实行。算法可在由中央处理单元1105执行时通过软件实行。例如,所述算法可以控制来自样本的核酸分子的测序、直接收集测序数据、分析测序数据、执行以区块为基础的变异模式分析、评估风险或生成指示风险的报道。

[0183] 在一些情况下,如图22所示,可以从对象1201(例如人类对象)获得样本1202。样本1202可经受如本文所述的一种或多种方法,例如执行分析。在某些情况下,分析可能包含杂交、扩增、测序、标记、表观遗传学修饰碱基或其任何组合。来自方法的一个或多个结果可以输入到处理器1204中。可以将一个或多个输入参数(例如样本标识、对象标识、样本类型、参考或其他信息)输入到处理器1204中。来自分析的一个或多个度量可以输入处理器1204,使得处理器可以产生结果,例如病症分类(例如,诊断)或治疗建议。处理器可以将结果、输入参数、度量、参考或其任何组合发送到显示器1205,例如视觉显示器或图形用户界面。处理器1204可以:(i)向服务器1207发送结果、输入参数、度量或其任何组合,(ii)从服务器1207接收结果、输入参数、度量或其任何组合,(iii)或其组合。

[0184] 可以使用硬件(例如,专用集成电路或现场可编程门阵列)和/或以模块化或集成方式使用具有一般可编程处理器的计算机软件,以控制逻辑的形式实行本公开的各个方面。如本文所使用的,处理器包含单核处理器、同一集成芯片上的多核处理器、或单个电路板或网络上的多个处理单元。基于本文提供的公开和教导,本领域普通技术人员将知道并理解使用硬件以及硬件和软件的组合来实行本文描述的多个实施例的其他方式和/或方法。

[0185] 应用程序中描述的任何软件组件或功能可以被实行为软件处理器,使用诸如java、C++、C++、C语言、ObjaveC、Swift,或脚本语言(如Perl或Python)使用的任何适当的计算机语言,例如,使用常规或对象导向的技术。软件代码可以作为一系列指令或命令存储在计算机可读介质上,用于存储和/或传输。合适的非暂时性计算机可读介质可以包含随机存取存储器(RAM)、只读存储器(ROM)、诸如硬盘驱动器或软盘的磁性介质,或者诸如光盘(CD)或DVD(数字多功能盘)等光学介质、闪存等。计算机可读介质可以是这样的存储器或传输装置的任意组合。

[0186] 还可以使用适于经由符合各种协议(包含因特网)的有线、光和/或无线网络传输的载波信号来编码和传输此类程序。因此,可以使用以这样的程序编码的数据信号来创建计算机可读介质。用程序代码编码的计算机可读介质可与兼容装置打包或与其他装置分开提供(例如,通过互联网下载)。任何这样的计算机可读介质可以驻留在单个计算机产品(例如,硬盘机、CD或整个计算机系统)上或在单个计算机产品内,并且可以存在于系统或网络内的不同计算机产品上或在不同计算机产品内。计算机系统可以包含显示器、打印机或其他合适的显示器,用于向用户提供本文所述的任何结果。

[0187] 本文描述的任何方法都可以使用包含一个或多个处理器的计算机系统全部或部分地执行,所述计算机系统可以被配置为执行所述多个步骤。因此,可以将实施例定向到配

置为执行本文所述的任何方法的步骤的计算机系统,其中不同的组件执行相应的步骤或相应的步骤组。尽管呈现为编号的步骤,但本文中的方法步骤可以同时或以不同顺序执行。此外,这些步骤的一部分可以与来自其他方法的其他步骤的一部分一起使用。此外,步骤的全部或部分可以是可选的。另外地,可以使用模块、单元、电路或用于执行这些步骤的其他方法来执行任何方法的任何步骤。

[0188] 其他实施例

[0189] 本文使用的章节标题仅用于组织的目的,不得解释为限制主要内容。

[0190] 应当理解,本文所描述的方法不限于本文所描述的特定方法、方案、对象和测序技术,因此可以变化。还应理解,本文使用的术语仅用于描述特定实施例的目的,并不旨在限制本文所述的多种方法和组合的范围,其将仅受所附权利要求书的限制。虽然本公开的一些实施例已在本文中示出和描述,但对于本领域技术人员来说,这些实施例仅通过示例的方式提供是显而易见的。在不脱离本公开的情况下,本领域技术人员现在将想到许多变异、改变和替换。应当理解,在实施本公开时可以采用在此所述的本公开的实施例的各种替代方案。以下权利要求旨在定义本公开的范围,并涵盖在这些权利要求范围内的方法和结构及其等价物。

[0191] 为了说明,参考示例的应用来描述几个方面。除非另有表明,否则任何实施例都可以与任何其他实施例组合。应当理解,阐述了许多具体细节、关系和方法,以提供对本文所述的特征的全面理解。然而,本领域技术人员将容易认识到,本文所描述的特征可以在没有—个或多个具体细节或与其他方法一起的情况下实施。本文描述的特征不受所示的动作或事件的顺序的限制,因为一些动作可以以不同的顺序发生和/或与其他动作或事件同时发生。此外,并非所有示出的动作或事件都需要实行为根据本文描述的特征的一个方法。

[0192] 示例

[0193] 提供以下示例是为了进一步说明本公开的一些实施例,但并不旨在限制本公开的范围;将通过其示例性可以理解,本领域技术人员已知的其他程序、方法或技术可以替代地被使用。

[0194] 示例1 4年内对20000多名对象进行鼻咽癌筛选

[0195] 本示例描述了一项大规模筛选研究,所述研究在大约4年的时间里对20000多名对象进行了群组(cohort)研究。图1示出了本研究的设计图。在首轮筛选中,使用血浆EBV DNA分析对20000多名年龄在40至62岁之间的男性进行了鼻咽癌筛选。可检测到血浆EBV DNA的对象在中值为4周后用第二组血样重新测试。这种安排的目的是将鼻咽癌患者与那些无鼻咽癌但可检测到血浆EBV DNA的患者区分开来。在先前的一项研究中,研究表明,在无鼻咽癌的对象中,血浆EBV DNA的存在通常是短暂现象。在这些个体的三分之二的人中,血浆EBV DNA在两周后变成检测不到。使用鼻内窥镜和鼻咽磁共振成像(MRI)对血浆EBV DNA结果持续阳性的对象进行进一步研究,以确认或排除鼻咽癌的存在。根据这一安排,确定了34例鼻咽癌。

[0196] 随后,在群组进行了另一轮(第二轮)的鼻咽癌筛选,中值为首轮筛选后4年。在第二轮鼻咽癌筛选中,与第一轮筛选一样,检测结果呈阳性的对象将在大约4周后再次接受检测。4周连续两次检测结果呈阳性的对象将接受鼻内窥镜和MRI进一步检查。第二轮筛选于2017年开始。截至2018年9月15日,共有8335名对象完成了第二轮筛选。784名(9.4%)对象

血浆EBV DNA呈阳性。在4周的重新测试中,230名(2.7%)对象仍然可以检测到血浆EBV DNA。表1总结了两轮鼻咽癌筛选的检测结果。

[0197] 表1第一轮和第二轮鼻咽癌筛选中血浆EBV DNA的状况

| [0198] | 第一轮筛选中的<br>血浆 EBV DNA<br>状况 | 数量   | 第二轮筛选中的血浆 EBV DNA 状况 |          |          |
|--------|-----------------------------|------|----------------------|----------|----------|
|        |                             |      | 阴性                   | 短暂阳性     | 持续阳性     |
|        | 阴性                          | 7907 | 7267 (92%)           | 479 (6%) | 161 (2%) |
|        | 短暂<br>阳性                    | 276  | 218 (79%)            | 30 (11%) | 28 (10%) |
|        | 持续阳性                        | 152  | 66 (43%)             | 48 (32%) | 38 (25%) |

[0199] 如表1所示,在第二轮鼻咽癌筛选中可检测到血浆EBV DNA的概率与第一轮筛选中血浆EBV DNA的状态相关。在第一轮筛选中,血浆EBV DNA呈阴性、短暂阳性和持续阳性的对象在第二轮筛选的初始分析中有8%、21%和57%的概率可检测到血浆EBV DNA。此外,在4周后,具有血浆EBV DNA持续阳性的几率在三组中从2%逐渐增加到25%。

[0200] 通过本文所述的筛选确定的鼻咽癌患者的分期分布比未接受鼻咽癌筛选的历史群组中的患者早得多。早期疾病(I期和II期)的百分比分别为70%和20%。分期分布的这种变化使患者的无进展生存期得到显著改善,危险比为0.1。表2总结了第一轮和第二轮筛选中鼻咽癌病例的分期分布。在第二轮筛选8335名对象后,发现了13例新的鼻咽癌病例。在第一轮和第二轮筛选中,早期疾病患者的百分比分别为71%和69%。早期疾病患者的百分比没有显着差异(P=0.93,卡方检验)。

[0201] 表2两轮的筛选中鼻咽癌患者的分期分布

| [0202] | 分期  | 第一轮筛选    | 第二轮筛选   |
|--------|-----|----------|---------|
|        | I   | 16 (47%) | 4 (31%) |
|        | II  | 8 (24%)  | 5 (38%) |
|        | III | 8 (24%)  | 4 (31%) |
|        | IV  | 2 (6%)   | 0 (0%)  |

[0203] 如表3所总结,与在第一轮筛选中未检测到血浆EBV DNA的对象相比,在第一轮筛选后4年进行的第二轮筛选中,在第一轮筛选中短暂和持续可检测到血浆EBV DNA的对象在第二轮筛选中检测到患有鼻咽癌的风险更高。这两组的相对风险值分别为7.2和19.7。

[0204] 表3第二轮筛选中被识别的鼻咽癌病例的数量(按第一轮血浆EBV DNA状态分类)

| [0205] | 第一轮筛选中的血浆 EBV DNA 状况 | 数量   | 第二轮中检测到鼻咽癌的数量<br>(具有相同血浆 EBV DNA 状态的对象的%) | 相对于第一轮中检测不到血浆 EBV DNA 的对象的鼻咽癌相对风险 |
|--------|----------------------|------|-------------------------------------------|-----------------------------------|
|        | 阴性                   | 7907 | 8 (0.10%)                                 | 1                                 |
|        | 短暂阳性                 | 276  | 2 (0.72%)                                 | 7.2                               |
|        | 持续阳性                 | 152  | 3 (1.97%)                                 | 19.7                              |

[0206] 这些结果表明,血浆EBV DNA分析不仅有助于筛选鼻咽癌的当前状态,而且有助于预测未来具有临床可观察到的鼻咽癌的风险。此发现的一个实际应用是根据早期筛选对象的血浆EBV DNA状态定制重复筛选的时间间隔。例如,与检测不到血浆EBV DNA的对象相比,在基线具有可检测的血浆EBV DNA但没有可识别鼻咽癌的对象可以在更短的时间间隔后重新筛选。同样作为说明,对于检测不到血浆EBV DNA、短暂可检测到和持续可检测到的对象,重复筛选的间隔分别为4年、2年和1年。

[0207] 示例2基于血浆EBV-DNA可检测性的鼻咽癌筛选

[0208] 本示例描述了基于对象血浆中EBV DNA的可检测性为所述对象设计的鼻咽癌筛选方案。图2示出了如本文所述的方案的示意图。

[0209] 根据此方案,早期筛选示例中检测不到血浆EBV DNA的对象在4年后重新筛选,因为未来4年内检测不到EBV DNA的对象患鼻咽癌的风险相对较低。如果后续筛选对血浆EBV DNA呈阴性,则后续筛选的间隔为4年。然而,当对象在一次筛选中检测到EBV DNA,但未检测到鼻咽癌时,下一次筛选安排在一年后。当血浆EBV DNA持续阴性4年时,筛选间隔恢复到4年。具体筛选计划使用的实际时间间隔也根据健康经济因素(例如筛选成本)、对象偏好(例如,更频繁的筛选间隔可能对某些对象的生活方式更具破坏性)和其他临床参数进行调整(例如个人基因型、鼻咽癌家族史、饮食史、种族血统(例如广东人))。

[0210] 示例3无细胞EBV DNA分子的变异模式分析

[0211] 在本示例中,使用捕获富集的靶向测序来分析鼻咽癌对象、非鼻咽癌对象具有可检测的血浆EBV DNA和鼻咽癌前对象的循环中的无细胞病毒DNA分子(详见下一节)。多种捕获探针可以设计成覆盖整个EBV基因组。在同一分析中,还包含标靶约3000个人类常见单核苷酸多态性(SNP)位点和人类白细胞抗原(HLA)SNP的多种探针。

[0212] 在本示例中,分析了13例鼻咽癌患者和16例可检测到血浆EBV DNA的非鼻咽癌对象的血浆EBV DNA。所述13例鼻咽癌患者出现症状,及是从威尔斯亲王医院的临床肿瘤科或耳鼻咽喉科招募。如示例1所述,16名非鼻咽癌对象来自20000多名对象鼻咽癌筛选群组。

[0213] 在此分析中,使用了通过专门设计的捕获探针进行捕获富集的靶向测序。对于分析的每个血浆样本,使用QIAamp循环核酸试剂盒从4毫升的血浆中提取DNA。对于每个病例,使用TruSeq纳米DNA文库制备试剂盒(Illumina公司)将所有提取的DNA用于测序文库的制备。条形码是使用结合唯一分子标识符(unique molecular identifier,UMI)序列的双标签(dual-indexing)系统(xGen双标签UMI适配器,集成DNA技术)进行的。使用TruSeq纳米试剂盒(Illumina公司)对适配器连接的样本进行八次循环的PCR扩增。然后使用myBait定制

捕获板系统(myBait custom capture panel system) (Arbor Biosciences公司),使用覆盖上述病毒和人类基因组区域的定制设计探针捕获扩增产物。目标捕获后,通过14次循环的PCR对捕获的产物进行富集以生成DNA文库。DNA文库在NextSeq平台(Illumina公司)上测序。对于每次测序运行,使用双端模式对具有唯一样本条形码的十个样本进行测序。每个DNA片段将从两端中的每一端测序71个核苷酸。测序后,序列读取将被映射到一人工组合的参考序列,所述序列由整个人类基因组(hg19)、整个EBV基因组(GenBank:AJ507799.2)、整个HBV基因组和整个HPV基因组组成。使用SOAP2(生物信息学2009年;25:1966-7)进行比对,允许在插入大小不超过600bp的情况下,在正确方向上每个读数最多2个不匹配。映射到组合的基因组序列中的多个独特位置的多个测序读数将用于下游分析。所有具有相同唯一分子标识符的重复片段都将被过滤。

[0214] 基于比对结果,识别序列读数和EBV参考基因组(GenBank:AJ507799.2)之间的核苷酸差异,包括但不限于单核苷酸变异(SNV)。在来自13名鼻咽癌对象的44份样本中,16名可检测到血浆EBV DNA的非鼻咽癌对象和4名鼻咽癌前对象识别的中位数为1116个SNV(四分位距(IQR):902-1216)。在这些血浆样本中,在EBV基因组的某些核苷酸位置观察到两种不同的等位基因。这一观察结果可能是由于测序错误或肿瘤异质性的存在。只有26个位置(IQR:20-35)的中值在血浆EBV DNA中有一个以上的等位基因。

[0215] 在图3所示的系统发育树分析中,鼻咽癌对象聚类在一起,并与非鼻咽癌对象分离。这些结果表明鼻咽癌和非鼻咽癌对象之间存在不同的EBV变异基因谱。因此,血浆EBV DNA的EBV变异基因谱分析可用于在筛选中区分鼻咽癌和非鼻咽癌对象。三名非鼻咽癌对象(AC106、AP080和FF159)有两个连续采集的样本,分别在4周时采集。来自同一个体的两个样本聚类在一起,表明它们具有非常相似的变异。

[0216] 系统发育树分析也基于EBV变异进行,但不包含Hui等人(Hui等人,Int J Cancer期刊2019年,doi.org/10.1002/ijc.32049)对同一群组13例鼻咽癌患者和16例可检测到血浆EBV DNA的非鼻咽癌对象的研究中报道的29个变异。如图4所示,鼻咽癌对象也被聚类在一起,并与非鼻咽癌对象分开。

[0217] 在第一轮筛选中(如示例1所述)血浆EBV DNA持续阳性但在内窥镜检查 and MRI上未检测到鼻咽癌的4名对象随后被诊断患有鼻咽癌。所有4名对象(BB096、DN054、FK015和HB121)在第一轮筛选后3年被诊断为鼻咽癌。所有4名对象在耳鼻喉科门诊跟进期间,在第一轮筛选后1年采集了一份额外的血浆样本。对于这四名对象中的每一位,在第一轮筛选和1年后收集的两个样本进行了EBV变异分析。如图5所示,来自鼻咽癌前对象的样本与鼻咽癌样本聚类在一起,表明与鼻咽癌相关联的EBV变异在癌症实际发生之前就存在。这表明,具有鼻咽癌相关联的EBV变异的个体将来患鼻咽癌的风险更高。系统发育树分析也基于EBV变异进行,但不包含Hui等人(Hui等人,Int J Cancer期刊2019年,doi.org/10.1002/ijc.32049)对同一群组鼻咽癌、非鼻咽癌及鼻咽癌前的对象的研究中报道的29个变异。如图6所示,鼻咽癌前对象的样本仍然与鼻咽癌样本聚类在一起,进一步表明EBV变异分析能够预测未来鼻咽癌的风险。

[0218] 示例4以区块为基础的变异模式分析

[0219] 本示例描述了示例性以区块为基础的变异模式分析方法的工作原理及其在示例3中所述样本的EBV变异模式分析中的应用。

[0220] 图7示出了以区块为基础的 (block-based) 变异模式分析的原理。以区块为基础的分析用于评估不同样本的血浆EBV DNA测序得出的EBV DNA变异模式与参考基因组的相似性, 本文使用公共数据库中可用的NPC测序数据 (Kwok等人, J Virol期刊2014年; 88:10662-72, Li等人, Nat Comm期刊2017年; 8:14121) 用作为参考。在以区块为基础的分析中, EBV基因组被划分为大小为500bp的箱 (共344个箱), 并将每个箱的变异模式与参考集中的24个鼻咽癌样本的相似性进行比较。作为一示例, 如果一个特定箱内有8个变异位点, 则分析测试样本的所述箱内的这些位点上的等位基因, 并与24个参考样本的相同位点上的等位基因进行比较。基于与参考样本具有完全相同的等位基因的比例推导出相似性指数。例如, 如果测试样本与一个参考样本在8个变异位点中的7个位点上具有完全相同的等位基因, 则所述箱与所述参考样本的相似性指数为7/8。与24个参考样本相比, 测试样本的所述箱将有24个相似性指数。基于所述箱的24个相似性指数计算箱分数, 所述箱分数表示变异模式与参考样本的总体相似性。例如, 如果相似性指数的截止值设置为0.9, 则箱分数统计所述多个指数高于截止值的箱的比例。因此, 如果24个相似性指数中只有两个高于0.9, 则箱分数为2/24。箱分数越高, 测试样本的变异模式与参考样本集越相似。

[0221] 图8示出了对13例鼻咽癌、16例非鼻咽癌和4例鼻咽癌前样本的EBV DNA变异模式的以区块为基础的分析。对于4名鼻咽癌前对象中的每一名, 分析了两个时间点的样本, 因此总共有8名对象。EBV基因组的344个箱的箱分数是根据这些样本得出的。基于这些样本的箱分数, 进行无监督聚类分析。鼻咽癌样本 (黑色) 聚类在一起, 非鼻咽癌样本 (用点标记) 聚类在一起。鼻咽癌前对象的EBV变异基因谱与鼻咽癌对象的EBV变异基因谱聚类在一起。值得注意的是, 这4名鼻咽癌前对象的变异特征是通过分析其基线样本获得的, 这些基线样本是在发展出鼻咽癌前几年收集的。

[0222] 图9示出了基于EBV变异的EBV DNA变异的以区块为基础的分析, 不包含Hui等人 (Hui等人, Int J Cancer期刊2019年, doi.org/10.1002/ijc.32049) 在同一组13名鼻咽癌的对象、16名非鼻咽癌的对象和4名鼻咽癌前的对象的研究中报道的29种变异。同样, 观察到鼻咽癌样本 (黑色) 的聚类。此外, 鼻咽癌前对象的EBV变异基因谱与鼻咽癌对象的EBV变异基因谱聚类在一起。鼻咽癌前和鼻咽癌样本的聚类分析表明, 变异分析可以预测鼻咽癌的未来发展。总而言之, 示例3和示例4中的数据显示, 招募时未患鼻咽癌但后来发展出癌症的对象在基线血液样本中具有与其他鼻咽癌患者相似的EBV变异模式。

[0223] 示例5使用数学模型的鼻咽癌风险预测

[0224] 本示例描述了一个分类模型的构建, 所述模型使用变异模式的分析以及使用分类模型的测试结果来预测具有可检测到的血浆EBV DNA的对象未来发展出鼻咽癌的风险。

[0225] 使用支持向量机 (SVM) 算法, 使用训练数据集构建分类器, 所述训练数据集包含18名无鼻咽癌的对象和8名鼻咽癌患者, 如示例4所述。测试数据集包含5名鼻咽癌患者、5名无鼻咽癌的对象以及从4名在样本采集时通过内窥镜检查 and MRI 没有检测到鼻咽癌, 但随后被诊断为患有鼻咽癌 (标记为鼻咽癌前) 对象中采集的8份样, 如示例4所述。

[0226] SVM分析的方法描述如下:

[0227] 给定包含n个样本的训练数据集:

[0228]  $(M_1, Y_1), \dots, (M_n, Y_n)$

[0229] 其中,  $Y_i$  表明样本i的鼻咽癌状态。 $Y_i$  为1 (鼻咽癌患者样本) 或-1 (非鼻咽癌对象样

本);  $M_i$  是一  $p$  维向量, 包含样本  $i$  的病毒变异模式。例如,  $M_i$  可以是一系列的变异位点, 例如与鼻咽癌相关联的 29 个变异。替代地,  $M_i$  可以是关于已知患有鼻咽癌的对象中存在的多个参考 EBV 变异的一系列以区块为基础的变异相似性分数 (例如, 500bp 的非重叠窗口)。

[0230] 识别一“超平面”, 所述“超平面”在训练数据集中尽可能准确地将非鼻咽癌组和鼻咽癌组分开, 通过寻找一系数集 ( $W$  和  $p$  维向量) 满足以下条件:

[0231] 准则1:

[0232]  $W \cdot M_i - b \geq 1$  (对于鼻咽癌群组中的任何对象)

[0233] 及

[0234] 准则2

[0235]  $W \cdot M_i - b \leq -1$  (对于非鼻咽群组中的任何对象)

[0236] 其中  $W$  是确定超平面的多个系数的  $p$  维向量;  $M$  是一矩阵 ( $p \times n$  维), 具有  $p$  个变量 (或多个以区块为基础的相似性分数) 和  $n$  个样本;  $b$  是截距。

[0237] 这两个准则 (即准则1和2) 也可以写成:

[0238]  $Y_i (W \cdot M_i - b) \geq 1$  ( $\geq$  准则3)

[0239] 其中  $Y_i$  为 -1 (非鼻咽癌) 或 1 (鼻咽癌)。

[0240] 准则1和2之间的边界距离 ( $D$ ) 为:  $\frac{2}{\|W\|}$ ,

[0241] 其中  $\|W\|$  是使用点到平面方程的距离计算。

[0242] 根据准则3, 通过最小化  $\|W\|$  使  $D$  最大化。

[0243] 基于这一原理, 确定分类器的参数 ( $W$  和  $b$ )。然后通过使用训练参数 ( $W$  和  $b$ ) 计算每个测试样本的鼻咽癌风险分数。

[0244] 图10A显示了基于使用以区块为基础的变异分析的所有 EBV 变异的分析, 使用经训练的分类器计算的鼻咽癌风险分数。对于此分析, 如示例4所述, 将 EBV 基因组划分为 344 个 500bp 的片段, 用于计算箱分数。箱分数被认为是机器学习的一个特征。鼻咽癌样本的鼻咽癌风险分数显著高于从非鼻咽癌对象中采集的样本 (平均鼻咽癌风险分数: 0.15 对 0.53,  $p$  值  $< 0.01$ , 学生  $t$  检验)。类似地, 与无鼻咽癌的对象相比, 从鼻咽癌前对象身上采集的样本的鼻咽癌风险分数显著较高 (平均风险分数: 0.58 对 0.15,  $p$  值  $< 0.01$ , 学生  $t$  检验)。使用 0.32 的截止值, 来自鼻咽癌患者和鼻咽癌前对象的样本可以与无鼻咽癌的样本进行区分, 具有 100% 的敏感性和 100% 的特异性。

[0245] 图10B示出了基于 Hui 等人在研究中报道的 29 个 EBV 变异的分析, 使用经过训练的分类器计算出的鼻咽癌风险分数 (Hui 等人, Int J Cancer 期刊 2019 年, doi.org/10.1002/ijc.32049)。鼻咽癌样本的鼻咽癌风险分数显著高于从非鼻咽癌对象中采集的样本 (平均鼻咽癌风险分数: 0.89 对 0.18,  $p$  值  $< 0.01$ , 学生  $t$  检验)。类似地, 与无鼻咽癌的对象相比, 从鼻咽癌前对象身上采集的样本的鼻咽癌风险分数显著较高 (平均风险分数: 0.57 对 0.18,  $p$  值 = 0.02, 学生  $t$  检验)。使用 0.6 的截止值, 鼻咽癌患者和鼻咽癌前对象的样本可以与无鼻咽癌的样本进行区分, 具有 74% 的敏感性和 100% 的特异性。

[0246] 图10C示出了使用经训练分类器计算的鼻咽癌风险分数, 所述分类器基于使用以区块为基础的变异分析对所有 EBV 变异的分析, 但不包含 Hui 等人先前报道的与鼻咽癌相关联的 29 个变异 (Hui 等人, 国际癌症期刊 2019 年, doi:10.1002/ijc.32049)。鼻咽癌样本的鼻

咽癌风险分数显著高于从非鼻咽癌对象中采集的样本(平均鼻咽癌风险分数:0.58对0.15,  $p$ 值 $<0.01$ , 学生 $t$ 检验)。类似地, 与无鼻咽癌的对象相比, 从鼻咽癌前对象身上采集的样本的鼻咽癌风险分数显著较高(平均风险分数:0.53对0.15,  $p$ 值 $<0.01$ , 学生 $t$ 检验)。使用0.31的截止值, 从鼻咽癌患者和随后发生鼻咽癌的患者身上采集的样本可以与没有鼻咽癌的样本进行区分, 具有100%敏感性和100%特异性。这些结果表明, 从分析中排除先前报道的29种EBV变异不会对所述分析的准确性产生不利影响。

[0247] 示例6通过亚硫酸氢盐测序分析血浆EBV-DNA甲基化状态

[0248] 本示例说明了使用亚硫酸氢盐测序基于血浆EBV DNA的甲基化状态来区分鼻咽癌患者和非鼻咽癌但具有可检测到的血浆EBV DNA的对象。

[0249] 采用亚硫酸氢盐测序法确定鼻咽癌患者和无鼻咽癌对象血浆中EBV DNA的甲基化水平。亚硫酸氢盐转化可将未甲基化的胞嘧啶转化为尿嘧啶。甲基化胞嘧啶不能被亚硫酸氢盐改变, 可以保持为胞嘧啶。在测序过程中, 尿嘧啶可以被确定为胸腺嘧啶。测序后, 任何CpG二核苷酸背景下的胞嘧啶甲基化状态都可以通过检查胞嘧啶是否已改变为胸腺嘧啶来确定。

[0250] 确定了10名鼻咽癌患者和40名未患癌症但血浆中可检测到EBV DNA的对象(非鼻咽癌对象)的血浆EBV DNA甲基化水平。对于40名非鼻咽癌对象, 在4周后分别采集另一份血样。当中的20例血浆EBV DNA呈阴性, 并被标记为具有短暂阳性的血浆EBV DNA。当中的20人的血浆EBV DNA仍然呈阳性, 并被标记为具有持续阳性的血浆EBV DNA。

[0251] 如图11所示, 鼻咽癌患者的EBV DNA甲基化水平显著高于具有血浆EBV DNA短暂阳性的无癌症对象( $P$ 值 $<0.01$ , 学生 $t$ 检验)和具有血浆EBV DNA持续阳性的无癌症对象( $P$ 值 $<0.01$ , 学生 $t$ 检验)。这些结果表明, 对血浆EBV DNA甲基化的分析有助于区分鼻咽癌患者和无鼻咽癌但可检测到血浆EBV DNA的对象。

[0252] 示例7使用甲基化敏感限制酶分析血浆EBV-DNA甲基化状态

[0253] 本示例描述了一个以电脑模拟的模拟实验, 演示了使用血浆EBV DNA的甲基化敏感限制酶分析来区分鼻咽癌患者和无鼻咽癌但可检测到血浆EBV DNA的对象。

[0254] 将来自一非鼻咽癌对象和一鼻咽癌患者的样本进行血浆DNA亚硫酸氢盐测序。两名对象的血浆DNA中分别获得347,516和627,1012个EBV DNA片段。血浆EBV DNA甲基化水平分别为48.9%和86.3%。确定大约一半的血浆EBV DNA分子含有至少一个“CCGG”基序。

[0255] 为了模拟限制酶对血浆EBV DNA的消化, 以电脑模拟的血浆EBV DNA分子的消化取决于从亚硫酸盐测序结果推断出的在“CCGG”序列背景下的甲基化状态来执行。因此, 获得了使用甲基化敏感限制酶HpaII以电脑模拟和不以电脑模拟消化的血浆EBV DNA的模拟尺寸基因谱, 如图14所示。未经酶消化, 非鼻咽癌对象血浆EBV DNA的大小分布位于鼻咽癌对象的左侧, 表明非鼻咽癌对象的大小分布较短。这种片段大小的差异也在酶消化后的大小分布基因谱中观察到, 在非鼻咽癌对象中, 酶消化后50bp以下的短DNA的丰度比未经酶消化的显著增加。对于鼻咽癌患者, 有酶消化和无酶消化的样本中 $<50$ bp的DNA分子比例分别为5.87%和0.84%。然而, 对于非鼻咽癌对象, 有酶消化和无酶消化的样本中 $<50$ bp的DNA分子比例分别为22.24%和4.99%。对于鼻咽癌患者和非鼻咽癌对象, 酶消化后 $<50$ bp的DNA比例分别增加17.2%和5.0%。图15示出了鼻咽癌患者和非鼻咽癌对象在甲基化敏感限制酶消化前后血浆EBV DNA的累积大小分布。酶消化程度的差异可以通过累积频率曲线相对于大

小的关系更容易理解。有酶消化和无酶消化的两条曲线之间的间隙反映了消化程度。间隙越大,酶对血浆EBV DNA的消化程度越大,因此表明血浆EBV DNA中的甲基化水平较低。如图所示,与鼻咽癌患者相比,非鼻咽癌对象的间隙更大。鼻咽癌患者和非鼻咽癌对象无酶消化和有酶消化曲线之间的最大距离分别为8.1和18.3;鼻咽癌患者和非鼻咽癌对象的两条曲线之间的面积分别为2395和942.9。

[0256] 示例8无细胞EBV-DNA分子的SNV基因谱分析

[0257] 在包含63名鼻咽癌对象和88名非鼻咽癌对象的血浆DNA测序数据的训练数据集中,分析了两组之间的EBV SNV基因谱的差异。横跨EBV基因组中的多个差异SNV被识别出。建议从这些SNV位点的基因型模式得出鼻咽癌风险分数,随后在31个鼻咽癌样本和40个非鼻咽癌样本的测试集中进行分析。在本示例中,从训练集中总共识别了661个EBV基因组中的显着的SNV(图16D)。在测试集中,鼻咽癌血浆样本显示有较高的鼻咽癌风险分数;可以有鼻咽癌相关联的EBV SNV基因谱。在非鼻咽癌样本中,鼻咽癌风险分数范围广泛。非鼻咽癌对象可能具有不同的EBV SNV特征。

[0258] 材料和方法。

[0259] 研究参与者和设计。

[0260] 本研究涉及对先前在Lam等人,美国科学院院报2018年;115:E5115-E5124中报道的鼻咽癌和非鼻咽癌血浆样本测序数据集的子集(作为训练集),以及来自鼻咽癌和非鼻咽癌对象的新测序的血浆DNA样本(作为测试集)进行分析。

[0261] 训练数据集包含在Lam等人,美国科学院院报2018年;115:E5115-E5124中描述的先前前瞻性鼻咽癌筛选研究中筛选出的鼻咽癌患者和非鼻咽癌对象的血浆样本。这些非鼻咽癌对象通过实时PCR检测血浆EBV DNA水平。数据集还包含来自独立群组的有症状的鼻咽癌患者的样本。研究来自所有样本的EBV分离的EBV基因型信息,以建立鼻咽癌风险分数预测的训练模型。在这项研究中,对另外31名有症状的鼻咽癌患者和40名非鼻咽癌对象的血浆样本进行靶向捕获测序,以作为测试集。这31个有症状的鼻咽癌患者来自香港威尔斯亲王医院的临床肿瘤科。非鼻咽癌对象也来自前面提到的鼻咽癌筛选群组(包含20000多名的对象),并从中随机选择。对这些鼻咽癌和非鼻咽癌样本的EBV基因型变异进行分析,并基于训练模型得出其鼻咽癌风险分数。训练和测试集中的所有鼻咽癌和非鼻咽癌样本没有重叠。

[0262] 靶向捕获测序。

[0263] 通过捕获探针系统(myBaits Custom capture Panel, Arbor Biosciences公司)富集血浆DNA文库中的EBV DNA分子,对血浆样本进行靶向捕获测序。EBV捕获探针被设计用于覆盖整个病毒基因组。还包含标靶3000个人类单核苷酸多态性(SNP)位点的探针,以供参考。在每次捕获反应中使用含有EBV探针与常染色体DNA探针的摩尔比为100:1的探针混合物。来自10个血浆样本的DNA文库在一次捕获反应中被多路复用,使用来自每个样本的等量DNA文库。表4A和4B中列出了所有病例(包含以前报道用作当前训练集的病例)的序列统计数据。

[0264] 表4A训练集中所有鼻咽癌和非鼻咽癌病例的测序统计

[0265]

| 训练集   |      |           |           |        |            |
|-------|------|-----------|-----------|--------|------------|
| 样本    | 群组** | 原始片段数量    | 映射的片段数量   | 映射率(%) | PCR 重复率(%) |
|       |      |           |           |        |            |
| GG017 | 0    | 32715321  | 30223262  | 92.4   | 43.1       |
| HL059 | 0    | 144554902 | 126762070 | 87.7   | 68.4       |
| DN045 | 0    | 78914933  | 68428310  | 86.7   | 66.9       |

[0266]

| 训练集   |      |           |           |        |            |
|-------|------|-----------|-----------|--------|------------|
| 样本    | 群组** | 原始片段数量    | 映射的片段数量   | 映射率(%) | PCR 重复率(%) |
| BP015 | 0    | 94168529  | 86145241  | 91.5   | 51.4       |
| AB126 | 0    | 56541949  | 54346856  | 96.1   | 24         |
| AC166 | 0    | 64450578  | 60439270  | 93.8   | 17.4       |
| AD092 | 0    | 71510547  | 69046150  | 96.5   | 16.1       |
| AE058 | 0    | 79728136  | 76825948  | 96.4   | 21.3       |
| AQ104 | 0    | 96938063  | 84743586  | 87.4   | 16.4       |
| BX011 | 0    | 72498952  | 70129591  | 96.7   | 14.9       |
| CA062 | 0    | 72180027  | 69744659  | 96.6   | 15.3       |
| CH131 | 0    | 71459860  | 68990753  | 96.5   | 22.2       |
| DC078 | 0    | 76239599  | 73238855  | 96.1   | 28.2       |
| DF038 | 0    | 100612788 | 97254251  | 96.7   | 26.1       |
| AG067 | 0    | 94932887  | 85387366  | 89.9   | 77.4       |
| AR027 | 0    | 61611288  | 59001573  | 95.8   | 15.1       |
| BL058 | 0    | 69559074  | 66513711  | 95.6   | 14.4       |
| AF118 | 0    | 64803996  | 61659065  | 95.2   | 14.4       |
| AF121 | 0    | 47656000  | 45104454  | 94.7   | 16         |
| AO097 | 0    | 64803246  | 62335332  | 96.2   | 14         |
| GV094 | 0    | 55594689  | 53398818  | 96     | 13.2       |
| AL092 | 0    | 88202778  | 84617253  | 95.9   | 20.7       |
| AM164 | 0    | 92235133  | 88753051  | 96.2   | 21.5       |
| EI030 | 0    | 67332747  | 64898723  | 96.4   | 13.7       |
| ER057 | 0    | 75611966  | 72851241  | 96.3   | 15.6       |
| FF077 | 0    | 88728791  | 84934257  | 95.7   | 18.3       |
| FF094 | 0    | 67950009  | 65456835  | 96.3   | 16.5       |
| AO100 | 0    | 74073437  | 71534001  | 96.6   | 14.4       |
| HE119 | 0    | 75939094  | 70594529  | 93     | 46.3       |
| GC110 | 0    | 109911126 | 101627813 | 92.5   | 30         |
| GT107 | 0    | 73134341  | 66124665  | 90.4   | 36.9       |
| GZ039 | 0    | 58128740  | 54517308  | 93.8   | 26.1       |
| AE151 | 0    | 118973652 | 109516490 | 92     | 21         |
| AH116 | 0    | 97765995  | 88477724  | 90.5   | 28         |
| AM095 | 0    | 87643692  | 80164284  | 91.5   | 19.6       |
| BP065 | 0    | 84740540  | 80067572  | 94.5   | 37.4       |
| EN086 | 0    | 32884093  | 31068440  | 94.5   | 38.3       |
| GC038 | 0    | 52719658  | 49985247  | 94.8   | 38.1       |
| AC106 | 0    | 46473277  | 43990963  | 94.7   | 82.5       |
| AP080 | 0    | 38659615  | 36293332  | 93.9   | 60         |
| GT123 | 0    | 90634113  | 82011875  | 90.5   | 65.1       |
| AE011 | 0    | 64587311  | 59269827  | 91.8   | 49.2       |
| BV159 | 0    | 108366362 | 97270043  | 89.8   | 73.8       |
| CZ031 | 0    | 104890395 | 93619970  | 89.3   | 73.4       |
| AL071 | 0    | 35231149  | 32775649  | 93     | 74.6       |
| AL122 | 0    | 132811199 | 123757690 | 93.2   | 76.6       |
| AS079 | 0    | 33454154  | 31094045  | 93     | 74.3       |
| AX070 | 0    | 82769034  | 77118993  | 93.2   | 75.8       |

[0267]

| 训练集   |      |           |           |        |            |
|-------|------|-----------|-----------|--------|------------|
| 样本    | 群组** | 原始片段数量    | 映射的片段数量   | 映射率(%) | PCR 重复率(%) |
| DC125 | 0    | 82353895  | 76845022  | 93.3   | 64.2       |
| DO041 | 0    | 98527392  | 91944421  | 93.3   | 63         |
| DN037 | 0    | 73898976  | 66401716  | 89.8   | 69.3       |
| DN131 | 0    | 85896965  | 77109501  | 89.8   | 68.8       |
| DS050 | 0    | 97058938  | 87190650  | 89.8   | 68         |
| DZ071 | 0    | 130632583 | 117555933 | 90     | 67.8       |
| EH050 | 0    | 144211569 | 131747254 | 91.4   | 67.5       |
| DZ026 | 0    | 63577798  | 60575778  | 95.3   | 24.9       |
| HM142 | 0    | 74460599  | 71830670  | 96.5   | 28.9       |
| HN068 | 0    | 58569268  | 56499964  | 96.5   | 27.6       |
| HR120 | 0    | 78697168  | 75901684  | 96.5   | 28.7       |
| CD005 | 0    | 67185044  | 64398576  | 95.8   | 18.9       |
| DC146 | 0    | 67286289  | 64869690  | 96.4   | 20.4       |
| DD090 | 0    | 72863832  | 69973561  | 96     | 18.9       |
| DE103 | 0    | 74532024  | 71748839  | 96.3   | 20.1       |
| DF112 | 0    | 80285807  | 77313233  | 96.3   | 16.6       |
| DH045 | 0    | 73283371  | 70644621  | 96.4   | 21         |
| DK016 | 0    | 98640353  | 95198449  | 96.5   | 22.8       |
| DK057 | 0    | 65024042  | 62488386  | 96.1   | 19.8       |
| DL055 | 0    | 64127942  | 61316770  | 95.6   | 18.9       |
| CE144 | 0    | 55972062  | 53546313  | 95.7   | 15.4       |
| CP042 | 0    | 67609649  | 64706108  | 95.7   | 15.2       |
| CZ046 | 0    | 55236628  | 52985764  | 95.9   | 13.5       |
| AP047 | 0    | 73544542  | 70437730  | 95.8   | 19.9       |
| AS108 | 0    | 74546824  | 71474684  | 95.9   | 22.1       |
| BF137 | 0    | 87739825  | 83608642  | 95.3   | 19.2       |
| AG020 | 0    | 67573799  | 63087296  | 93.4   | 17.6       |
| AE055 | 0    | 62308055  | 59551554  | 95.6   | 11.4       |
| AE105 | 0    | 59317164  | 56861140  | 95.9   | 10.2       |
| AE107 | 0    | 69376388  | 66837992  | 96.3   | 13.3       |
| AB004 | 0    | 69373853  | 66823399  | 96.3   | 12.4       |
| AC153 | 0    | 83546018  | 80433313  | 96.3   | 13.4       |
| AE026 | 0    | 80236204  | 77227885  | 96.2   | 13.8       |
| AF091 | 0    | 79865448  | 76665569  | 96     | 12.4       |
| HF020 | 0    | 73890276  | 69898875  | 94.6   | 11.9       |
| BO049 | 0    | 54341974  | 49518640  | 91.1   | 12.2       |
| CV094 | 0    | 69353920  | 62090890  | 89.5   | 11.9       |
| DM146 | 0    | 86198122  | 83306628  | 96.7   | 13.7       |
| DN054 | 0    | 57906125  | 55516552  | 95.9   | 21.6       |
| DN092 | 0    | 65436665  | 62867803  | 96.1   | 16.7       |
| AC173 | 1    | 77221448  | 69636427  | 90.2   | 53.5       |
| AO050 | 1    | 94201867  | 84771216  | 90     | 51.9       |
| AQ014 | 1    | 64826863  | 58371226  | 90     | 47.2       |
| AZ118 | 1    | 75307129  | 67827313  | 90.1   | 47.7       |
| AC088 | 1    | 76597786  | 55250665  | 72.1   | 47.2       |

[0268]

| 训练集     |      |           |           |        |            |
|---------|------|-----------|-----------|--------|------------|
| 样本      | 群组** | 原始片段数量    | 映射的片段数量   | 映射率(%) | PCR 重复率(%) |
| AL038   | 1    | 76499430  | 55322894  | 72.3   | 45.7       |
| AM086   | 1    | 84280496  | 61284379  | 72.7   | 43.4       |
| AT038   | 1    | 64157394  | 46063166  | 71.8   | 45.8       |
| BK041   | 1    | 61505610  | 44247376  | 71.9   | 44.8       |
| CF028   | 1    | 97748094  | 88104244  | 90.1   | 59.1       |
| CH047   | 1    | 123975141 | 112556783 | 90.8   | 56.6       |
| CL037   | 1    | 106862473 | 96469537  | 90.3   | 60.7       |
| CP006   | 1    | 61469649  | 54366171  | 88.4   | 59.4       |
| CD007   | 1    | 103710165 | 93643893  | 90.3   | 61.9       |
| DF120   | 1    | 96451355  | 89089726  | 92.4   | 51.6       |
| DH101   | 1    | 73023724  | 67311149  | 92.2   | 60.3       |
| EG016   | 1    | 83087673  | 77307393  | 93     | 24.2       |
| EN070   | 1    | 35732253  | 32582501  | 91.2   | 52.5       |
| EV013   | 1    | 70202729  | 64881793  | 92.4   | 35.8       |
| FD089   | 1    | 106149891 | 88230410  | 83.1   | 51.9       |
| FG092   | 1    | 58840935  | 54320095  | 92.3   | 36.8       |
| FM073   | 1    | 65062459  | 60232085  | 92.6   | 39.3       |
| FZ037   | 1    | 46211337  | 42733248  | 92.5   | 37.6       |
| GC137   | 1    | 73772882  | 68339539  | 92.6   | 62.9       |
| GS059   | 1    | 103768139 | 95756898  | 92.3   | 64.4       |
| GX170   | 1    | 112376826 | 104300963 | 92.8   | 60.7       |
| HD083   | 1    | 80146546  | 74256782  | 92.7   | 59.8       |
| HM169   | 1    | 69203940  | 64144652  | 92.7   | 59.7       |
| AG006   | 1    | 73346449  | 68476847  | 93.4   | 22.9       |
| FD163   | 1    | 62554476  | 58856976  | 94.1   | 27.7       |
| CX027   | 1    | 88012245  | 80202542  | 91.1   | 67.7       |
| CV009   | 1    | 60922871  | 56232165  | 92.3   | 45.6       |
| TBR1433 | 2    | 77708246  | 70039392  | 90.1   | 30.2       |
| TBR1470 | 2    | 73941394  | 67495510  | 91.3   | 21.6       |
| TBR1572 | 2    | 71106989  | 64814893  | 91.2   | 23.6       |
| TBR1605 | 2    | 115061297 | 94605333  | 82.2   | 47.8       |
| TBR1606 | 2    | 60654197  | 55309308  | 91.2   | 32         |
| TBR1607 | 2    | 75439582  | 69608132  | 92.3   | 28.1       |
| TBR1650 | 2    | 83518964  | 76881089  | 92     | 21.8       |
| TBR1665 | 2    | 73581524  | 68005926  | 92.4   | 26.7       |
| TBR1685 | 2    | 64858923  | 59295059  | 91.4   | 28.4       |
| TBR1794 | 2    | 77616481  | 72400504  | 93.3   | 31.9       |
| TBR1795 | 2    | 84087680  | 78757703  | 93.7   | 25.2       |
| TBR1821 | 2    | 89364373  | 83561953  | 93.5   | 25.2       |
| TBR1822 | 2    | 74207438  | 69089332  | 93.1   | 32.3       |
| TBR1841 | 2    | 76709226  | 71246483  | 92.9   | 27.6       |
| TBR1857 | 2    | 93499651  | 85084161  | 91     | 29.1       |
| TBR1911 | 2    | 102778437 | 93039420  | 90.5   | 28.3       |
| TBR1937 | 2    | 108092562 | 98448107  | 91.1   | 31.5       |
| TBR1950 | 2    | 100931791 | 92237772  | 91.4   | 31.7       |

[0269]

| 训练集     |      |           |           |        |            |
|---------|------|-----------|-----------|--------|------------|
| 样本      | 群组** | 原始片段数量    | 映射的片段数量   | 映射率(%) | PCR 重复率(%) |
| TBR1961 | 2    | 120837880 | 110269912 | 91.2   | 23.3       |
| TBR2032 | 2    | 74713097  | 70057803  | 93.8   | 27.1       |
| TBR2044 | 2    | 74572414  | 69808426  | 93.6   | 21.7       |
| TBR2059 | 2    | 68180154  | 63969165  | 93.8   | 22.8       |
| TBR2066 | 2    | 71590556  | 67039888  | 93.6   | 24.7       |
| TBR2129 | 2    | 67520639  | 63360453  | 93.8   | 22.9       |
| TBR1344 | 2    | 89830107  | 79295024  | 88.3   | 35.2       |
| TBR1358 | 2    | 37407353  | 35051007  | 93.7   | 41.9       |
| TBR1360 | 2    | 73282234  | 61715512  | 84.2   | 49.8       |
| TBR1378 | 2    | 54841088  | 50538475  | 92.2   | 34.5       |
| TBR1379 | 2    | 61335101  | 51046779  | 83.2   | 48.6       |
| TBR1390 | 2    | 50153930  | 44313840  | 88.4   | 45         |
| TBR1557 | 2    | 35803478  | 32801152  | 91.6   | 43.1       |

[0270] \*\*:第0组=非鼻咽癌对象,第1组=鼻咽癌对象(筛选群组),第2组=鼻咽癌(外部群组)。

[0271] 表4B测试集中所有鼻咽癌和非鼻咽癌病例的测序统计

[0272]

| 测试集   |      |          |          |         |            |         |
|-------|------|----------|----------|---------|------------|---------|
| 样本    | 群组## | 原始片段数量   | 映射的片段数量  | 映射率(%)  | PCR 重复率(%) | 鼻咽癌风险分数 |
| AB069 | 0    | 62333414 | 56996119 | 91.4375 | 67.0529    | 0.25    |
| AG102 | 0    | 50527076 | 47272142 | 93.558  | 79.7162    | 1.00    |
| BF034 | 0    | 30900262 | 29069989 | 94.0768 | 79.9262    | 0.06    |
| BH035 | 0    | 27968166 | 25683364 | 91.8307 | 78.2321    | 1.00    |
| BM060 | 0    | 44571256 | 41656811 | 93.4612 | 82.7252    | 1.00    |
| BN052 | 0    | 32654549 | 30177844 | 92.4154 | 77.7825    | 0.00    |
| BO115 | 0    | 20605498 | 18891596 | 91.6823 | 76.3716    | 0.00    |
| BR067 | 0    | 35222869 | 31942475 | 90.6867 | 10.9972    | 1.00    |
| BS030 | 0    | 29488585 | 26961246 | 91.4294 | 66.5338    | 0.99    |
| CB025 | 0    | 35335207 | 32498897 | 91.9731 | 81.8117    | 1.00    |
| CI095 | 0    | 44920271 | 41857137 | 93.181  | 64.8167    | 0.00    |
| CO003 | 0    | 22618823 | 20545705 | 90.8345 | 66.4679    | 1.00    |
| DK129 | 0    | 26650610 | 24552495 | 92.1273 | 66.7223    | 1.00    |
| DM162 | 0    | 46869923 | 42223785 | 90.0872 | 65.1806    | 0.99    |
| DO001 | 0    | 35030693 | 32412652 | 92.5264 | 64.0082    | 1.00    |
| DR058 | 0    | 33151251 | 30641021 | 92.4279 | 77.5861    | 0.41    |
| DX145 | 0    | 30538948 | 28353858 | 92.8449 | 64.0698    | 0.00    |
| DZ091 | 0    | 48775427 | 45509608 | 93.3044 | 79.647     | 0.00    |

[0273]

| 测试集     |      |           |          |         |            |         |
|---------|------|-----------|----------|---------|------------|---------|
| 样本      | 群组## | 原始片段数量    | 映射的片段数量  | 映射率(%)  | PCR 重复率(%) | 鼻咽癌风险分数 |
| EB064   | 0    | 15486333  | 14294637 | 92.3049 | 77.2137    | 0.52    |
| EC056   | 0    | 44264275  | 41421171 | 93.577  | 64.8678    | 0.28    |
| EI052   | 0    | 30414618  | 28373013 | 93.2874 | 79.4382    | 0.98    |
| ER022   | 0    | 29318005  | 25814308 | 88.0493 | 64.2827    | 0.00    |
| ET022   | 0    | 28303377  | 26549950 | 93.8049 | 79.5254    | 0.97    |
| EZ015   | 0    | 34114519  | 31826767 | 93.2939 | 79.4083    | 0.65    |
| FF159   | 0    | 27631827  | 25177560 | 91.118  | 66.2635    | 0.00    |
| FH039   | 0    | 25047700  | 23182787 | 92.5546 | 73.199     | 1.00    |
| FV078   | 0    | 59919758  | 55955981 | 93.3849 | 82.1063    | 1.00    |
| GC157   | 0    | 22988959  | 21147818 | 91.9912 | 72.2857    | 0.00    |
| GG040   | 0    | 58823944  | 53857823 | 91.5577 | 10.9781    | 0.14    |
| GK072   | 0    | 28087271  | 26012505 | 92.6131 | 72.1235    | 0.99    |
| GV071   | 0    | 30298816  | 27995522 | 92.3981 | 81.7554    | 1.00    |
| GX058   | 0    | 52901878  | 47527912 | 89.8416 | 72.5617    | 0.00    |
| GZ082   | 0    | 33025312  | 30743443 | 93.0905 | 76.508     | 0.00    |
| HB042   | 0    | 39832106  | 37486823 | 94.1121 | 79.7558    | 0.59    |
| HC056   | 0    | 27801939  | 25722722 | 92.5213 | 77.5543    | 0.80    |
| HE176   | 0    | 26672711  | 24740453 | 92.7557 | 65.5094    | 0.00    |
| HE181   | 0    | 20151536  | 18596587 | 92.2837 | 77.1676    | 0.00    |
| HF010   | 0    | 36767150  | 34443572 | 93.6803 | 83.3378    | 0.99    |
| HK068   | 0    | 24744347  | 22950199 | 92.7493 | 66.3875    | 0.02    |
| HN102   | 0    | 18847144  | 17418641 | 92.4206 | 66.0707    | 0.00    |
| p003704 | 1    | 24089077  | 22256290 | 92.3916 | 75.6729    | 1.00    |
| p100405 | 1    | 27917819  | 25958361 | 92.9813 | 76.6278    | 1.00    |
| p100742 | 1    | 33868828  | 31121633 | 91.8887 | 77.043     | 1.00    |
| p101161 | 1    | 22077183  | 20555644 | 93.1081 | 76.2116    | 1.00    |
| TBR2003 | 1    | 89502393  | 78014093 | 87.1643 | 67.8335    | 1.00    |
| TBR2197 | 1    | 49274726  | 46072820 | 93.5019 | 79.8709    | 1.00    |
| TBR2230 | 1    | 19463878  | 17991477 | 92.4352 | 77.7681    | 1.00    |
| TBR2239 | 1    | 40477218  | 37931905 | 93.7117 | 79.5694    | 1.00    |
| TBR2269 | 1    | 36732370  | 33345425 | 90.7794 | 10.8014    | 0.85    |
| TBR2329 | 1    | 102625376 | 87445869 | 85.2088 | 79.1855    | 0.99    |
| TBR2343 | 1    | 47646593  | 41027985 | 86.109  | 80.656     | 1.00    |
| TBR2330 | 1    | 36942083  | 33822640 | 91.5559 | 11.0708    | 0.00    |
| TBR2385 | 1    | 42000104  | 39181234 | 93.2884 | 81.8537    | 1.00    |
| TBR2406 | 1    | 66799222  | 60524426 | 90.6065 | 83.3811    | 0.00    |
| TBR2430 | 1    | 19062836  | 17515880 | 91.885  | 77.2878    | 1.00    |
| TBR2466 | 1    | 39167493  | 35820959 | 91.4558 | 66.6063    | 1.00    |
| TBR2553 | 1    | 20976134  | 19085605 | 90.9872 | 78.5291    | 1.00    |
| TBR2605 | 1    | 28691106  | 26101695 | 90.9749 | 65.7645    | 1.00    |
| TBR2615 | 1    | 33489016  | 29864524 | 89.1771 | 68.4423    | 1.00    |

[0274]

| 测试集     |      |           |          |         |            |         |
|---------|------|-----------|----------|---------|------------|---------|
| 样本      | 群组## | 原始片段数量    | 映射的片段数量  | 映射率(%)  | PCR 重复率(%) | 鼻咽癌风险分数 |
| TBR2641 | 1    | 113077610 | 94235991 | 83.3374 | 54.0705    | 0.98    |
| TBR2647 | 1    | 52926587  | 46699098 | 88.2337 | 68.1284    | 1.00    |
| TBR2655 | 1    | 44805097  | 41374955 | 92.3443 | 65.3989    | 1.00    |
| TBR2669 | 1    | 43399057  | 39819658 | 91.7524 | 65.4329    | 1.00    |
| TBR2682 | 1    | 35617499  | 32625124 | 91.5986 | 77.4284    | 1.00    |
| TBR2699 | 1    | 78986032  | 67322508 | 85.2334 | 80.332     | 1.00    |
| TBR2709 | 1    | 60912602  | 54630334 | 89.6864 | 78.8851    | 0.97    |
| TBR2847 | 1    | 19610868  | 17657654 | 90.0401 | 52.1991    | 1.00    |
| TBR2849 | 1    | 15220276  | 14043817 | 92.2704 | 51.0899    | 1.00    |
| TBR2868 | 1    | 21065832  | 18609241 | 88.3385 | 53.7439    | 1.00    |
| TBR2892 | 1    | 17905000  | 16600383 | 92.7137 | 51.5529    | 1.00    |
| TBR2906 | 1    | 29385280  | 26298916 | 89.4969 | 53.0486    | 1.00    |

[0275] 第0组 = 非鼻咽癌对象, 第1组 = 鼻咽癌对象

[0276] EBV变异调用

[0277] 使用BWA比对器, 将测序读数与人类 (hg19) 和EBV参考基因组 (AJ507799.2) 比对, 其在Li H等人所述, 生物信息学2010年; 26:589-95, 其全部通过引用并入本文。当检测到EBV基因组位点上与参考病毒基因组不同的替代等位基因时, 使用Samtools识别了EBV单核苷酸变异 (SNV), 如Li H等人所述, 生物信息学2009年; 25:2078-9, 其全部通过引用并入本文。在随后的鼻咽癌风险分数分析中, 滤除检测到1种类型以上的等位基因的SNV位点 (次要等位基因频率截止值设置为5%)。

[0278] 鼻咽癌风险分数

[0279] 在此示例中, 鼻咽癌风险分数是横跨病毒基因组中SNV位点的固定集处的多个EBV基因型的加权总和 (作为二元逻辑回归模型中的解释变量)。通过分析训练集中来自鼻咽癌和非鼻咽癌样本的多个EBV SNV基因谱 (profile) 中的差异来首先识别鼻咽癌相关联的SNV集。使用Fisher精确检验分析EBV基因组中每个变异与鼻咽癌病例的关联。然后, 获得的显著的SNV的固定集具有假发现率 (FDR) 控制在5%。

[0280] 测试样本的鼻咽癌风险分数可通过其多个EBV基因型在从训练集识别的多个显著的SNV位点的特定集上来确定。如前所述, 由于血浆EBV DNA分子浓度较低, EBV DNA测序可能无法完全覆盖整个EBV基因组。因此, 所述分数由血浆EBV DNA读数 (例如, 具有可用的基因型信息) 所覆盖的SNV位点的基因型模式来确定 (图16A、16B和16C)。为了得出鼻咽癌风险分数, 首先识别显著的SNV位点的子集, 这些位点被测试样本中的血浆EBV DNA读数所覆盖。然后, 在显著的SNV位点子集内确定每个位点的基因型权重 (效应大小)。这是通过分析训练数据集中鼻咽癌和非鼻咽癌样本中每个位点的基因型模式来完成的 (图16B)。在此基础上, 构建了一个逻辑回归模型, 以提供每个SNV位点的风险基因型对鼻咽癌的影响大小。逻辑模型编写如下:

$$[0281] \quad P = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^n \beta_k X_k)}}$$

[0282] 可以改写为:

$$[0283] \quad \text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{k=1}^n \beta_k X_k,$$

[0284] 其中n是显着的SNV位点的数量; $\beta_0$ 和 $\beta_k$ 是可以通过最大似然性估计器确定的多个系数;P为EBV阳性患者发生鼻咽癌的概率;此变量 $X_k$ 代表基因组位置k处的SNV位点。如果样本中存在与EBV参考基因组相同的变异,则 $X_k$ 编码为-1。如果样本中存在替代变异,则 $X_k$ 编码为1。如果样本中未包含分析的变异位点,则 $X_k$ 编码为0。Python中使用了‘逻辑回归函数’(惩罚(penalty) = ‘l2’, C=1, 解算器(solver) = saga, 最大迭代次数(max\_iter) = 5000, 及随机状态(random\_state) = 0)来估计系数 $\beta_0$ 和 $\beta_k$ 。这是通过分析训练数据集中鼻咽癌和非鼻咽癌样本中每个位点的基因型模式来完成的。将矩阵(c+d) × n输入到python,其中c是鼻咽癌样本数,d是训练集中非鼻咽癌样本数,n是基因型变异数。每行代表一个样本(无鼻咽癌的患者为0;有鼻咽癌的患者为1),每列代表一个变量。然后可以推导出系数( $\beta_0$ 和 $\beta_k$ )。然后,测试样本的鼻咽癌风险分数根据多个SNV位点处的其拥有的基因型得出,并通过从训练模型中推导得出相应的系数 $\beta_0$ 和 $\beta_k$ 进行加权。(图16C)。

[0285] 结果

[0286] 构建鼻咽癌风险分数训练模型。

[0287] 如上所述,先前报道的鼻咽癌和非鼻咽癌样本的血浆EBV DNA测序数据用于鼻咽癌风险分数训练模型的开发。靶向捕获测序已经被进行以富集血浆样本中的EBV DNA。本文研究了鼻咽癌和非鼻咽癌样本中EBV分离的病毒SNV基因谱。从此数据集中,选择了通过测序的EBV DNA读数在EBV基因组上至少覆盖30%的鼻咽癌和非鼻咽癌病例。此截止值被选择是因为训练数据集中95%以上的鼻咽癌样本具有病毒基因组覆盖率大于截止值(表4A和4B)。表5中详细列出了这些选定的鼻咽癌和非鼻咽癌对象的人口统计数据,包含年龄和性别,以及鼻咽癌患者的癌症分期信息(第8版AJCC)。这些选定鼻咽癌和非鼻咽癌样本的测序统计数据陈述在(表4A和4B)中。

[0288] 表5训练集中所有鼻咽癌和非鼻咽癌案例的对象特征

|               | 鼻咽癌患者            | 非鼻咽癌对象       |
|---------------|------------------|--------------|
| 数量            | 63               | 88           |
| 性别            |                  |              |
| 男性            | 56               | 88           |
| 女性            | 7                | 0            |
| 年龄中值, 年 (IQR) | 53 (47.5 – 57.5) | 54 (48 – 59) |
| 肿瘤分期          |                  |              |
| I             | 17               | NA (不适用)     |

[0290]

|     | 鼻咽癌患者 | 非鼻咽癌对象 |
|-----|-------|--------|
| II  | 11    | NA     |
| III | 26    | NA     |
| IV  | 9     | NA     |

[0291] 对63例鼻咽癌和88例非鼻咽癌样本的EBV-SNV基因谱进行了分析。所有样本的EBV基因组的中值测序深度为2倍 (2x) (四分位距 (IQR), 1.0倍-9.2倍)。从鼻咽癌样本中识别出的EBV SNV的平均数量为800 (IQR, 662-958), 而非鼻咽癌样本中的SNV的平均数量为539 (范围为363-656)。总的来说, 所有样本共识别出5678种不同的SNV。这些SNV在EBV基因组中的分布如图16D所示。

[0292] 在训练集中, 每种病毒的SNV与鼻咽癌样本的相关性也通过Fisher精确试验进行了研究。通过将假发现率 (FDR) 控制在0.05, 总共识别出661个与鼻咽癌相关联的显著的SNV, 使用调整的p值。表6列出了这些661个SNV的基因组位置。随后, 基于这些661个SNV位点的基因型模式, 得出鼻咽癌和非鼻咽癌对象血浆样本测试集的鼻咽癌风险分数。

[0293] 表6 661个示例的SNV的EBV基因组位置 (相对于AJ507799.2)

[0294]

| EBV 基因组位置                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 46, 156, 158, 206, 212, 246, 390, 409, 475, 505, 536, 570, 612, 628, 631, 866, 1067, 1072, 1074, 1133, 1137, 1176, 1194, 1195, 1322, 1349, 1373, 1384, 1391, 1534, 1875, 1992, 2709, 2772, 3223, 3379, 3820, 3941, 4863, 5398, 5745, 5802, 5849, 6066, 6108, 6209, 6287, 6379, 6483, 6555, 6583, 6865, 6883, 6885, 6910, 6943, 6998, 7000, 7015, 7047, 7133, 7188, 7208, 7212, 7232, 7246, 7261, 7296, 7326, 7356, 7385, 8233, 8344, 8455, 8567, 8872, 10623, 11323, 11694, 35308, 35492, 35526, 35550, 35583, 35615, 35637, 35678, 35856, 35869, 35974, 36067, 36166, 36577, 36667, 36694, 36768, 36798, 36847, 36948, 36950, 37051, 37053, 37284, 37465, 37624, 37641, 37671, 37682, 37701, 37739, 37834, 37954, 40549, 40555, 40835, 41153, 41402, 42209, 42321, 42422, 42712, 42948, 42992, 43088, 43235, 43280, 43312, 43396, 43419, 43611, 43806, 43819, 44122, |

[0295]

## EBV 基因组位置

44530, 44650, 45100, 45616, 45691, 45694, 45823, 46105, 46133, 46610, 46895, 47904, 48633, 48730, 48997, 50133, 50754, 50764, 50881, 50946, 51080, 51151, 51152, 51227, 51269, 51379, 51435, 51514, 51517, 51588, 51847, 52549, 53683, 57411, 58192, 58207, 59205, 59334, 59390, 59435, 59489, 59588, 60005, 60239, 60453, 60887, 60893, 61256, 62141, 62456, 62499, 62509, 62741, 62819, 63302, 63911, 64131, 64171, 64216, 64234, 64882, 64921, 65465, 66364, 66434, 66718, 66749, 66961, 67054, 67621, 67721, 67745, 67867, 68260, 68303, 68304, 68509, 68885, 69483, 75030, 75287, 75326, 76761, 76917, 77195, 77815, 77816, 78662, 79264, 79318, 79649, 79739, 80313, 80349, 80609, 80626, 80635, 80840, 80919, 80978, 81110, 81212, 81682, 81722, 82332, 82369, 83062, 83639, 84127, 84257, 84345, 84390, 84413, 84524, 84739, 84766, 84799, 84883, 84887, 84917, 84970, 85076, 85125, 85128, 85224, 85227, 85228, 85801, 85840, 86113, 86779, 86794, 87397, 87556, 88012, 88121, 88223, 88303, 88464, 88500, 88552, 88597, 88636, 88837, 88900, 89630, 89819, 89850, 89920, 90477, 90553, 90585, 90641, 91005, 91011, 91046, 91179, 91429, 91430, 91437, 91765, 93097, 93367, 93468, 94793, 95291, 95379, 95458, 95509, 95631, 98147, 98243, 98261, 98376, 98489, 98841, 98984, 98985, 99057, 99069, 99329, 99350, 99355, 99736, 99760, 99805, 100552, 101509, 101691, 101920, 101986, 102922, 103333, 103824, 104286, 104432, 104549, 104554, 104672, 104804, 105670, 106006, 106374, 106468, 107457, 107592, 108012, 108332, 108351, 108355, 108419, 109234, 109507, 109576, 109775, 109939, 110032, 110477, 110687, 110773, 110873, 110939, 111026, 111694, 112486, 112980, 113691, 113718, 114468, 114762, 114811, 115371, 115462, 115574, 115639, 115711, 115726, 116058, 116310, 116393, 116394, 116501, 116583, 116807, 117030, 117291, 117456, 117564, 117994, 118097, 118210, 118349, 118432, 118460, 118505, 118955, 119031, 119295, 119381, 119417, 119786, 119804, 120294, 120318, 120360, 120672, 120866, 121160, 121164, 121230, 121383, 121473, 121689, 121719, 121737, 121776, 121893, 122140, 122208, 122340, 122343, 122361, 122443, 122481, 122490, 122607, 122610, 122820, 123174, 123312, 124938, 125271, 126135, 126225, 126442, 126601, 126681, 127197, 127408, 127465, 127597, 127615, 127840, 127991, 128036, 128268, 129730, 129835, 129904, 130450, 130453, 130687, 132047, 132182, 132224, 133635, 133648, 133779, 133947, 134155, 134157, 134199, 134349, 134371, 134385, 134718, 134729, 134760, 134766, 134788, 134874, 135060, 135078, 135102, 135108, 135117, 135354, 135606, 135866, 135949, 136053, 136077, 136185, 136554, 136645, 136914, 136932, 136974, 137080, 137142, 137315, 137346, 137480, 138869, 139209, 139440, 139495, 139683, 139945, 140001, 140059,

|        | EBV 基因组位置                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [0296] | 140227, 140254, 140256, 140305, 140492, 140569, 140600, 140688, 140744, 143451,<br>144072, 144086, 144354, 144564, 144684, 145144, 145245, 145538, 145736, 145918,<br>146158, 146237, 146241, 146242, 146249, 146270, 146557, 146627, 146690, 146744,<br>146756, 146764, 146887, 147059, 147060, 147068, 147088, 147102, 147310, 147426,<br>147478, 147492, 147607, 147651, 147663, 147681, 147698, 147708, 147731, 147773,<br>147783, 147849, 147882, 147899, 148050, 148230, 148283, 148488, 148627, 148636,<br>148930, 148971, 149130, 149318, 149354, 149643, 149835, 149925, 150021, 150027,<br>150171, 150356, 150470, 150749, 150777, 151139, 151146, 151202, 151255, 151337,<br>151352, 151370, 151643, 151821, 151876, 151942, 152023, 152086, 152244, 152611,<br>152945, 152946, 153011, 154386, 154614, 154971, 155084, 155388, 155390, 155608,<br>155919, 155988, 156012, 156132, 156138, 156153, 156183, 156282, 156636, 156695,<br>156797, 156809, 156818, 157052, 157124, 157229, 157427, 157466, 157805, 157823,<br>158015, 158142, 158407, 158429, 158480, 158777, 159219, 160803, 160826, 160970,<br>161035, 162116, 162146, 162194, 162214, 162236, 162463, 162475, 162506, 162851,<br>163106, 163286, 163292, 163363, 163403, 163421, 163463, 163610, 163628, 163685,<br>163925, 163994, 164723, 165086, 165850, 167201, 168172, 168176, 168411, 168432,<br>168466, 168559, 168593, 168596, 168659, 169008, 169428 |

[0297] 鼻咽癌风险分数训练模型的评估

[0298] 对训练模型进行评估,以使用遗漏法(leave one-out approach)分析训练集中样本的鼻咽癌风险分数。在遗漏法中,建立训练模型和推导鼻咽癌风险分数的原则与上述方法中描述的相同。除训练集中的一个样本外,所有样本均用于构建训练模型,遗漏的一个样本可用于分析其鼻咽癌风险分数。在遗漏分析中,鼻咽癌群组的鼻咽癌风险分数中值为0.99(IQR,0.98-1.0),非鼻咽癌群组为0.01(IQR,0.00-0.89)(图17A)。接收者操作特征(ROC)曲线分析用于通过鼻咽癌风险分数评估鼻咽癌和非鼻咽癌样本的差异。曲线下的面积值为0.91(图17B)。

[0299] 测试集中的鼻咽癌风险分数分析

[0300] 对另外31名鼻咽癌患者和45名非鼻咽癌对象的血浆样本进行靶向捕获测序。其中所有31例鼻咽癌样本和40例非鼻咽癌样本通过测序的EBV DNA读数具有至少30%或以上的EBV基因组覆盖率。表7总结了这些鼻咽癌和非鼻咽癌对象的临床特征。表4A和4B中也说明了样本的此测试集的测序统计。

[0301] 表7测试集中所有鼻咽癌和非鼻咽癌病例的对象特征

|        |               |                |              |
|--------|---------------|----------------|--------------|
| [0302] |               | 鼻咽癌患者          | 非鼻咽癌对象       |
|        | 数量            | 31             | 40           |
|        | 性别            |                |              |
|        | 男性            | 26             | 40           |
|        | 女性            | 5              | 0            |
|        | 年龄中值, 年 (IQR) | 53 (47 – 61.5) | 53 (50 – 57) |
|        | 肿瘤分期          |                |              |
|        | I             | 6              | NA (不适用)     |
|        | II            | 2              | NA           |
|        | III           | 12             | NA           |
|        | IV            | 11             | NA           |

[0303] 基于所开发的训练模型,对31个鼻咽癌样本和40个非鼻咽癌样本的测试集的鼻咽癌风险分数进行分析。样本的鼻咽癌风险分数可以通过其在训练集中确定的661个显著的SNV位置上的变异模式来确定。由于EBV基因组可能不完全覆盖,只有测序的EBV DNA读数覆盖的SNV位点和具有相应等位基因信息的SNV位点可以包含在鼻咽癌风险分数分析中(图16A、16B和16C)。

[0304] 鼻咽癌组的鼻咽癌风险分数中值为0.999 (IQR, 0.996-0.999), 非鼻咽癌组的鼻咽癌风险分数中值为0.557 (IQR, 0.000-0.996) (图18A)。同样,在这些31个鼻咽癌样本中也发现了高鼻咽癌风险分数。测试集中的鼻咽癌样本可以与训练集中的鼻咽癌样本共享相似的EBV SNV基因谱。通过ROC曲线分析评估鼻咽癌风险分数对鼻咽癌和非鼻咽癌样本的区分。曲线下的面积值为0.83 (图18B)。

[0305] 测试集中高风险变异位点的基因型模式分析。

[0306] 在EBER (EBV编码的小RNA (EBV-encoded small RNA)) 区域存在高风险的鼻咽癌相关联的EBV变异。在EBER区域,Hui等人报道了23个重要的SNV。在31个鼻咽癌和40个非鼻咽癌样本的测试集中采用了类似的鼻咽癌风险预测方法,但仅基于EBER区域中23个已报道的SNV的基因型模式进行分析。

[0307] 在测试集中,71个鼻咽癌和非鼻咽癌样本中的31个(44%)具有EBV DNA读数覆盖了所有23个SNV位点。如表8所示,对于这些23个SNV位点中的每一个,只有一部分样本具有可用的基因型信息,具有多个读数覆盖多个SNV位点(即,并非所有23个SNV位点都被样本中的血浆EBV DNA读数覆盖)。鼻咽癌样本中23个SNV位点中每个位点处的高风险基因型百分比在86%到97%之间。非鼻咽癌样本中高风险基因型的百分比在35%到52%之间。分析的鼻咽癌和非鼻咽癌样本数量是指具有可用的基因型信息的样本(例如,具有多个EBV DNA读数覆盖多个SNV位点)。在测试集(31个鼻咽癌样本和40个非鼻咽癌样本)中,只有一比例的样本具有覆盖SNV位点的读数和相应位点的可用基因型信息。通过ROC曲线分析,通过仅分

析EBER区域的23个SNV的基因型模式,来评估鼻咽癌和非鼻咽癌样本的差异。曲线值下的面积为0.72(图19A和19B)。此值低于从整个EBV基因组的基因型模式分析得出的值(0.83)。与固定的病毒基因组区域相比,分析全EBV基因组的基因型模式可以更好地区分鼻咽癌和非鼻咽癌样本。

[0308] 表8测试集中在EBER基因的23个SNV位点处鼻咽癌和非鼻咽癌病例的基因型模式

[0309]

| SNV 位置 | 风险等位基因 | 分析的鼻咽癌样本数量 | 分析的非鼻咽癌样本数量 | 具有风险等位基因的鼻咽癌样本数量(百分比) | 具有风险等位基因的非鼻咽癌样本数量(百分比) |
|--------|--------|------------|-------------|-----------------------|------------------------|
| 5398   | A      | 29         | 31          | 25 (86%)              | 12 (39%)               |
| 5849   | T      | 28         | 27          | 24 (86%)              | 11 (41%)               |
| 6483   | T      | 29         | 19          | 25 (86%)              | 9 (47%)                |
| 6583   | G      | 29         | 16          | 25 (86%)              | 7 (44%)                |
| 6865   | A      | 29         | 25          | 26 (90%)              | 9 (36%)                |
| 6883   | G      | 29         | 25          | 27 (93%)              | 11 (44%)               |
| 6885   | T      | 29         | 23          | 26 (90%)              | 10 (43%)               |
| 6910   | A      | 29         | 23          | 26 (90%)              | 8 (35%)                |
| 6943   | G      | 29         | 23          | 28 (97%)              | 11 (48%)               |
| 6998   | G      | 30         | 26          | 29 (97%)              | 11 (42%)               |
| 7000   | T      | 30         | 25          | 29 (97%)              | 10 (40%)               |
| 7011   | G      | 30         | 26          | 29 (97%)              | 11 (42%)               |
| 7015   | T      | 30         | 25          | 29 (97%)              | 11 (44%)               |
| 7047   | C      | 30         | 29          | 29 (97%)              | 14 (48%)               |
| 7124   | G      | 29         | 28          | 28 (97%)              | 11 (39%)               |
| 7133   | C      | 29         | 28          | 28 (97%)              | 12 (43%)               |
| 7197   | T      | 28         | 26          | 27 (96%)              | 10 (38%)               |
| 7205   | A      | 28         | 26          | 27 (96%)              | 11 (42%)               |
| 7212   | C      | 28         | 27          | 27 (96%)              | 11 (41%)               |
| 7232   | A      | 29         | 28          | 25 (86%)              | 11 (39%)               |
| 7261   | A      | 29         | 27          | 28 (97%)              | 14 (52%)               |
| 7296   | T      | 28         | 26          | 27 (96%)              | 13 (50%)               |
| 7326   | C      | 28         | 26          | 27 (96%)              | 12 (46%)               |

[0310] 同样,BALF2 (BamHI A左阅读框-2 (BamHI A left frame-2)) 基因上的3个高风险SNV也已被报道 (Xu等人,Nat Genet期刊,2019年;51:1131-6)。在测试集中,71个样本中有55个 (78%) 的EBV DNA读数覆盖了所有3个SNV。对于这3个SNV位点中的每一个,测试集中只有一部分样本的读数覆盖了SNV位点和具有可用的基因型信息 (表9)。鼻咽癌样本中3个SNV位点中每个位点处的高风险基因型百分比在86%到93%之间。非鼻咽癌样本中高风险基因型的百分比在47%到65%之间。有4例没有EBV DNA读数覆盖BALF2基因上3个已报道的SNV中的任何一个 (1个鼻咽癌和3个非鼻咽癌样本), 这些病例无法分析。在测试集中剩余的30个鼻咽癌和37个非鼻咽癌样本中采用了类似的鼻咽癌风险预测方法,仅分析了BALF2区域报道的3个SNV的基因型模式。通过ROC曲线分析评估鼻咽癌和非鼻咽癌样本的差异。曲线值下的面积为0.77 (图20A和20B)。此值低于从整个EBV基因组的基因型模式分析得出的值 (0.83)。与固定的病毒基因组区域相比,分析全EBV基因组的基因型模式可以更好地区分鼻咽癌和非鼻咽癌样本。

[0311] 表9测试集中在BALF2基因的3个SNV位点处鼻咽癌和非鼻咽癌病例的基因型模式

[0312]

| SNV 位置 | 风险等位基因 | 分析的鼻咽癌样本数量 | 分析的非鼻咽癌样本数量 | 具有风险等位基因的鼻咽癌样本数量<br>(百分比) | 具有风险等位基因的非鼻咽癌样本数量<br>(百分比) |
|--------|--------|------------|-------------|---------------------------|----------------------------|
| 162214 | C      | 30         | 31          | 28 (93%)                  | 20 (65%)                   |
| 162475 | C      | 30         | 32          | 27 (90%)                  | 17 (53%)                   |
| 163363 | T      | 29         | 32          | 25 (86%)                  | 15 (47%)                   |

[0313] 本示例中描述的鼻咽癌风险分数分析允许基于EBV基因组661个显着的SNV集中随机选择SNV的浮动数量的基因型模式进行鼻咽癌风险预测 (表6)。用于鼻咽癌风险分数分析的SNV位点的浮动数量可以通过测序的EBV DNA读数是否覆盖SNV位点以及是否具有相应的等位基因信息来确定。对661个显着的SNV集进行了下采样,并在已被分析的测试集中使用与下采样的SNV集内的SNV的浮动数量相同的方法分析样本的鼻咽癌预测性能。对于下采样分析,从661个显着的SNV中随机选择一定数量的SNV (例如、23、25、100、200或500个)。然后,对于测试样本,识别EBV DNA序列读数所覆盖的下采样SNV集中的SNV位点。然后,通过在覆盖的、向下采样的SNV位点在训练集中使用鼻咽癌和非鼻咽癌样本的基因型模式训练模型,来获得鼻咽癌风险分数训练模型。通过训练,为训练模型确定每个位点的基因型权重。然后,通过在这些覆盖的、向下采样的SNV位点上应用其自身的基因型模式于鼻咽癌风险分数训练模型,所述鼻咽癌风险分数训练模型在相同向下采样的SNV位点上加权,从而得出测试样本的鼻咽癌风险分数。表10总结了不同数量的SNV位点的鼻咽癌风险分数训练模型的预测性能。对于给定数量的SNV位点,随机选择SNV进行10次下采样,表10中曲线下的面积值是10次随机下采样的平均结果。全EBV基因组中的SNV集被下采样至23个,这与EBER区域中报道的SNV数量相同。通过ROC曲线分析评估鼻咽癌和非鼻咽癌样本的差异。曲线下的面积值为0.78。此值高于以EBER区域的23个已报道的SNV基因型模式的分析 (0.72)。

[0314] 表10基于不同数量的SNV的鼻咽癌预测性能

|        |             |               |
|--------|-------------|---------------|
| [0315] | 下采样的SNV数量   | 曲线下面积 (AUC) 值 |
|        | 23          | 0.78          |
|        | 25          | 0.78          |
|        | 100         | 0.77          |
|        | 200         | 0.83          |
|        | 500         | 0.79          |
|        | 661 (所有SNV) | 0.83          |

[0316] 本研究报告通过血浆DNA测序分析EBV基因型信息。通过双端测序,在鼻咽癌对象和携带血浆EBV DNA的非鼻咽癌对象之间识别(包含数量和大小)血浆EBV DNA分子的区别分子特征。将这种基于计数和大小的血浆EBV DNA分析纳入其中,几乎可以将目前基于PCR的方案阳性预测值提高一倍,这可以构成第二代基于测序的筛选试验的基础。对鼻咽癌和非鼻咽癌对象的血浆样本进行测序可以另外产生EBV基因型信息,并可以增进其潜在的临床应用价值。

[0317] 鼻咽癌风险分数可由病毒全基因组标记而不是单个基因标记确定。本文的风险分数是根据EBV基因组中SNV位点的变异模式得出的。用于EBV基因型信息的血浆测序可能涉及对低浓度EBV DNA分子的血浆样本进行测序,因此导致EBV基因组的不完全覆盖。在某些情况下,提供有用信息的SNV位点可能不会被任何EBV DNA读数所覆盖,并且在某些情况下,无法判断个体是否携带高风险EBV品系类型。这得到了以下结果的支持:对于EBER基因上23个已报道的SNV位点中的每一个,在测试集的71个分析样本中,只有一些样本的读数覆盖了这些位点。测试集中的鼻咽癌样本显示有较高的鼻咽癌风险分数,这可以表明存在鼻咽癌相关联的EBV SNV基因谱。本文采用捕获探针法富集血浆样本中的EBV DNA分子。扩增子测序方法也可用于富集EBV DNA片段,所述片段可针对高风险变异区域获得基因型信息。

[0318] 本文分析了最近报道的EBER基因和BALF2基因高风险变异位点在测试集中鼻咽癌和非鼻咽癌样本的基因型模式。鼻咽癌和非鼻咽癌样本中高风险基因型的分布与分析的细胞样本(即鼻咽癌肿瘤组织和正常对照对象的唾液样本)的两项研究结果一致。由于包含本研究在内的三项研究都是在中国南部地区的相同或邻近地区进行的,因此正常对照组中EBV基因型的分布可能是相似的。这为通过对血浆样本测序进行EBV基因型分析的可行性提供了证据。

[0319] 在筛选的背景下,从血浆样本中分析EBV SNV可以具有临床实用性。如前所述,大约5%的筛选人群血浆中含有EBV DNA,但没有鼻咽癌(假阳性组)。本文的数据显示,这些非鼻咽癌对象的鼻咽癌风险分数不同,可能涉及不同的EBV SNV基因谱。可能存在异质性个体群组,他们在未来患有鼻咽癌的风险不同。其中一些携带高风险EBV品系的人将来患鼻咽癌的风险更高。鼻咽癌风险分数可用于根据病毒全基因组的SNV基因谱将非鼻咽癌对象分为不同的风险组。在一个示例中,对于鼻咽癌风险分数高的对象,可以保证进行更频繁的筛选。

[0320] 通过对鼻咽癌患者和非鼻咽癌对象的血浆样本进行测序分析,分析其EBV基因型信息。虽然先前的研究侧重于在人群水平上识别与鼻咽癌相关联的高风险变异,但本研究提供了病毒基因型分析的临床应用见解。这样的分析可以通过对EBV基因型的表征,在个体

基础上告知癌症风险。

[0321] 虽然本公开的较佳实施例已在本文中示出和描述,但对于本领域技术人员来说,这些实施例仅通过示例的方式提供是显而易见的。在不脱离本公开的情况下,本领域技术人员现在将想到许多变异、改变和替换。应当理解,在实施本公开时可以采用在此所述的本公开的实施例的各种替代方案。以下权利要求旨在定义本公开的范围,并涵盖在这些权利要求范围内的方法和结构及其等价物。

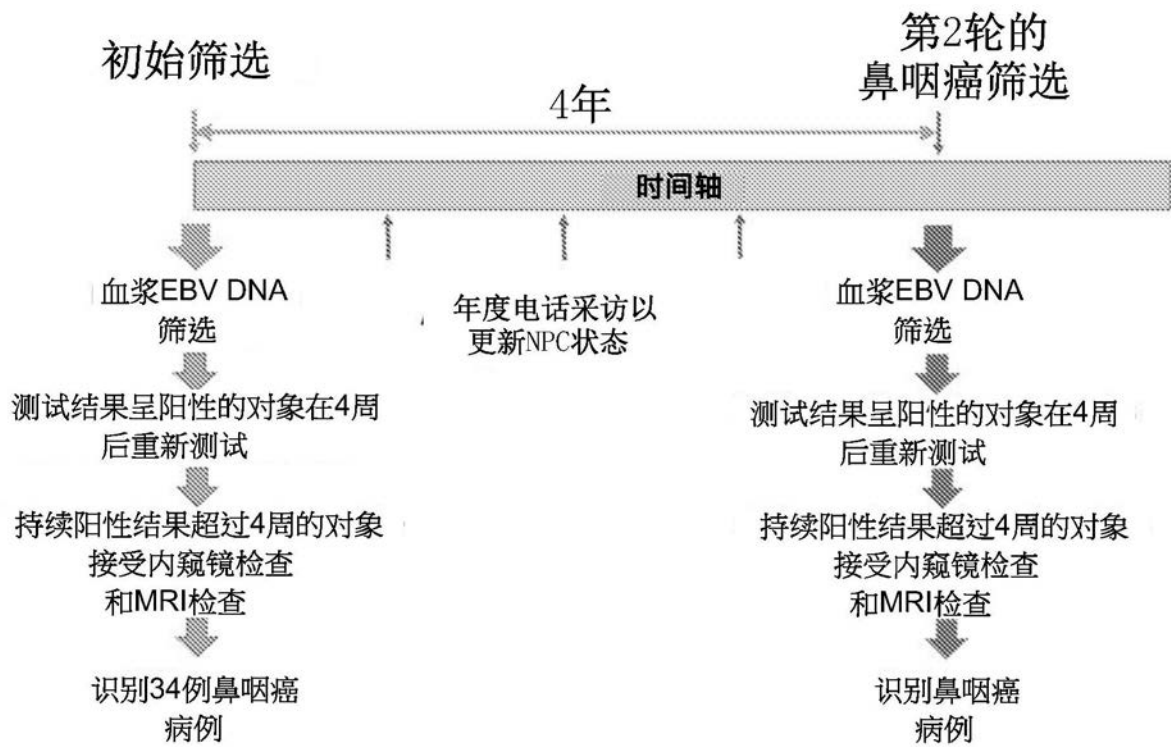


图1

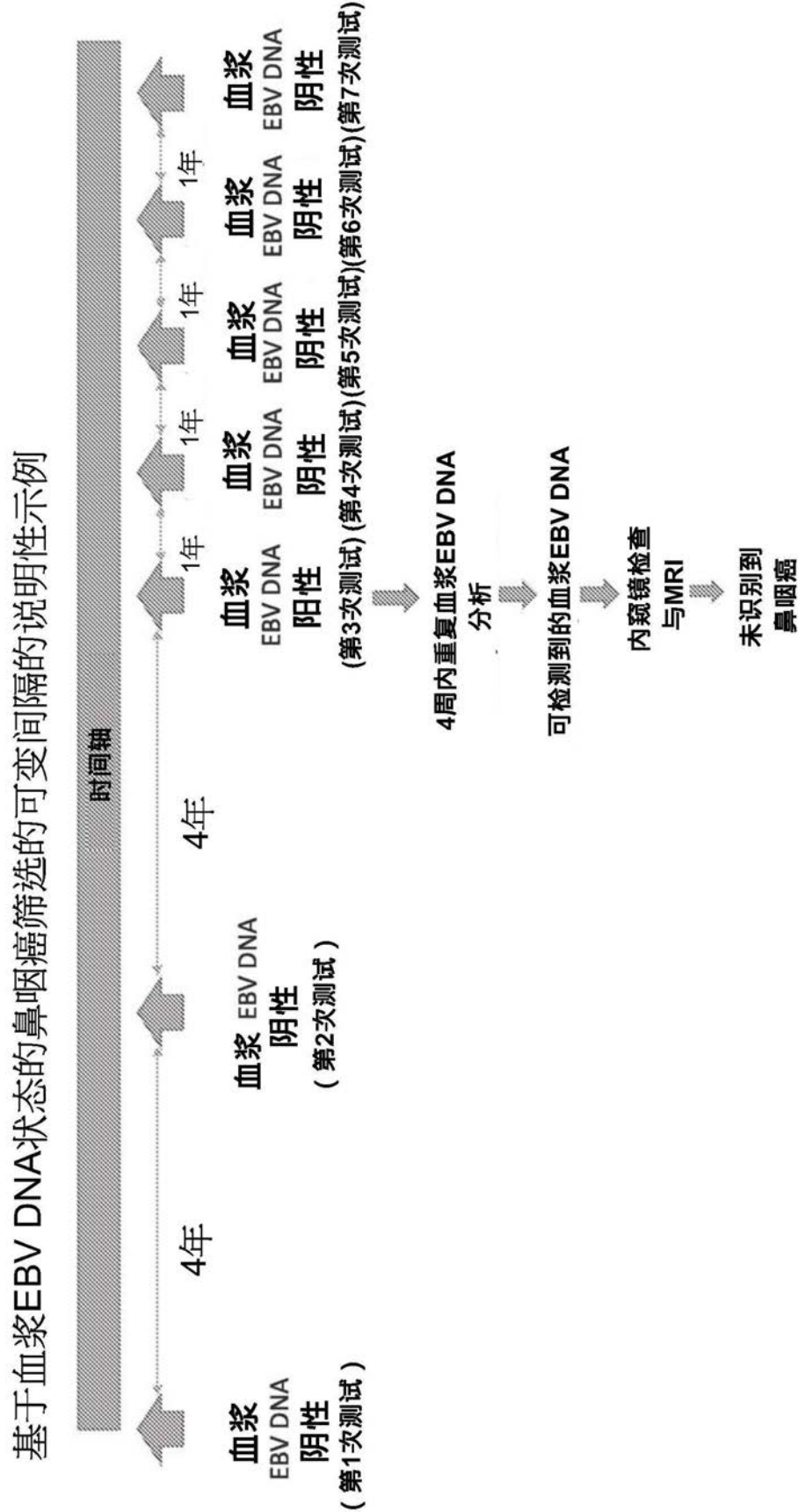


图2

65

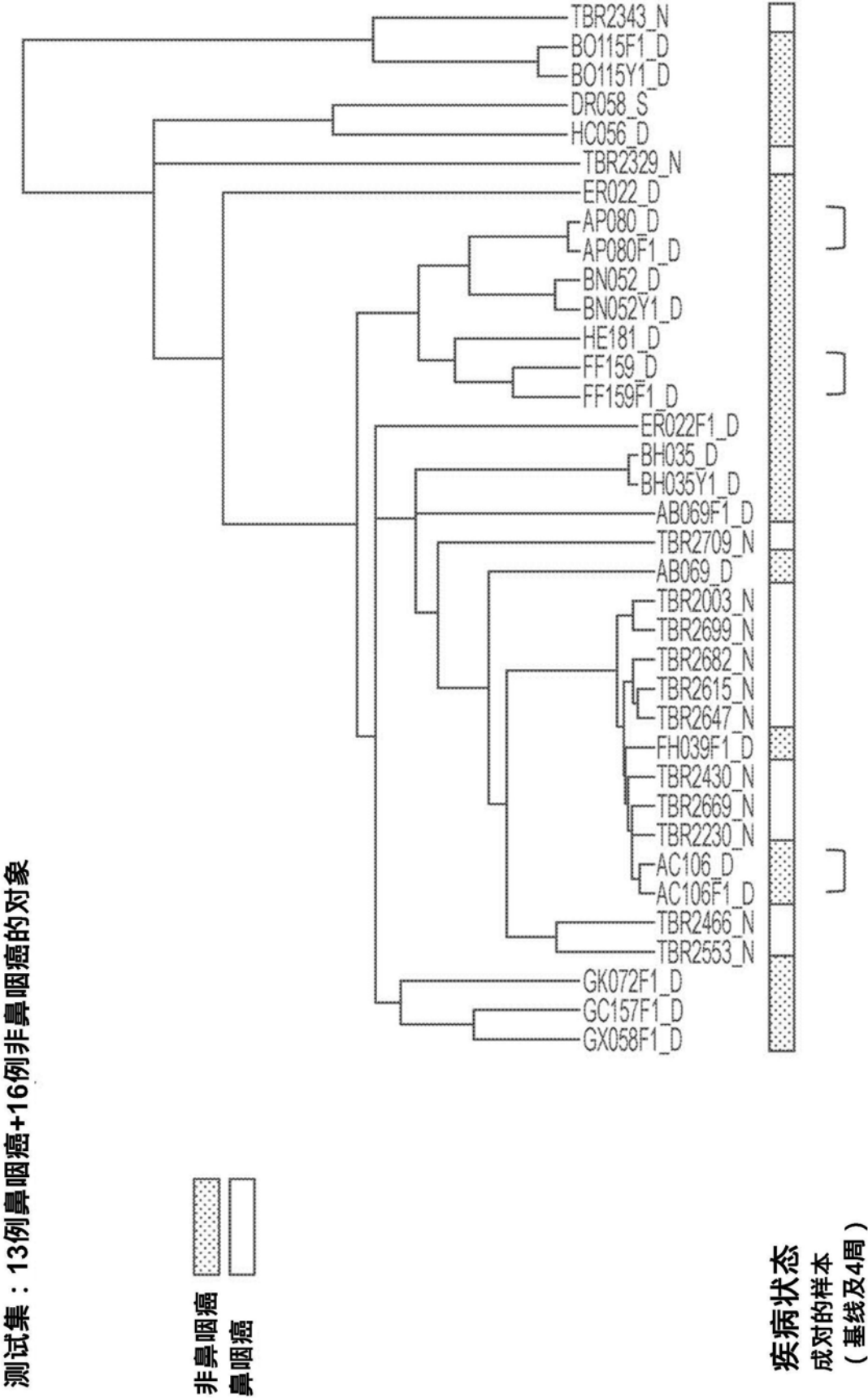


图3

测试集：13例鼻咽癌+16例非鼻咽癌的对象

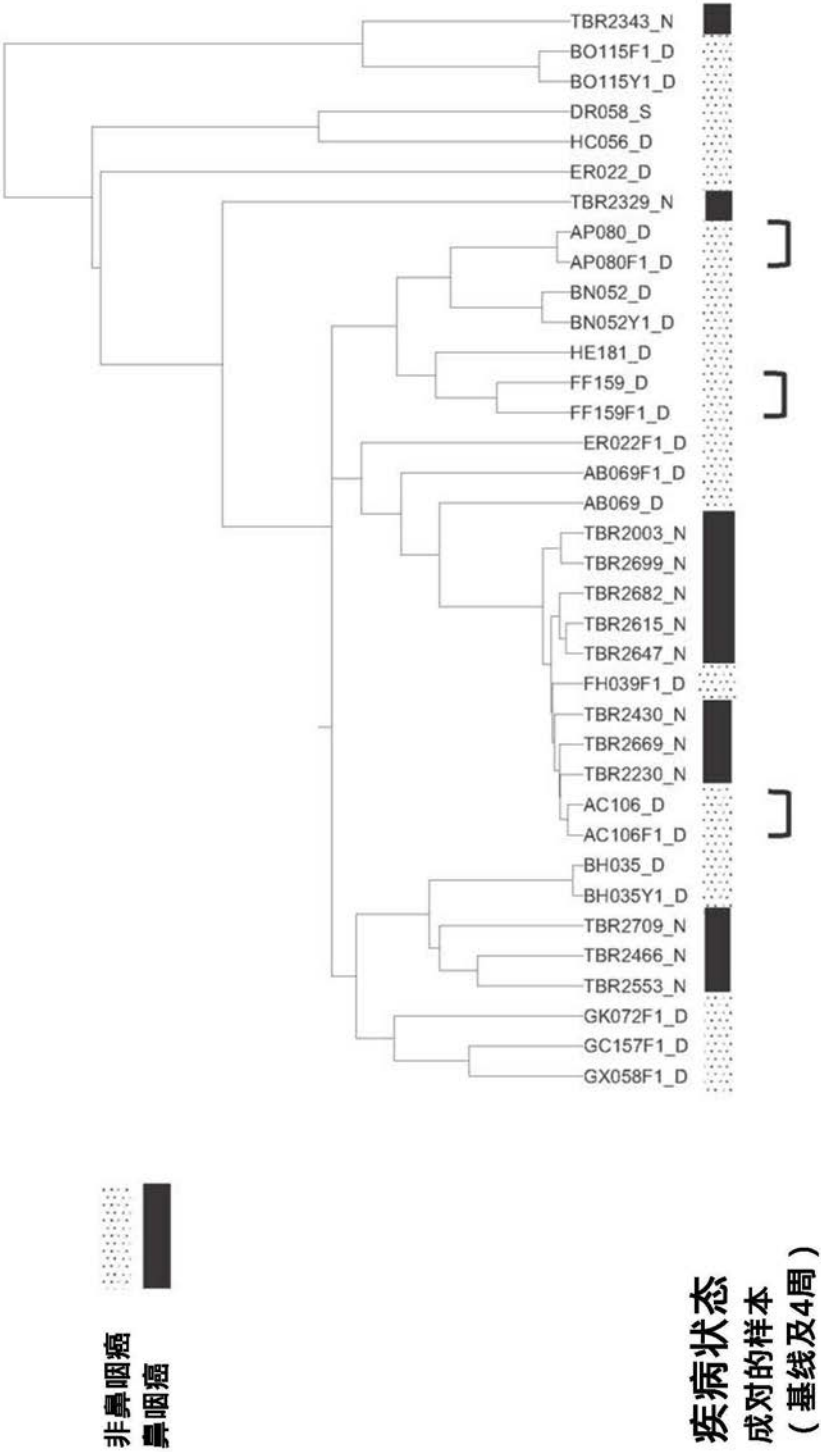


图4

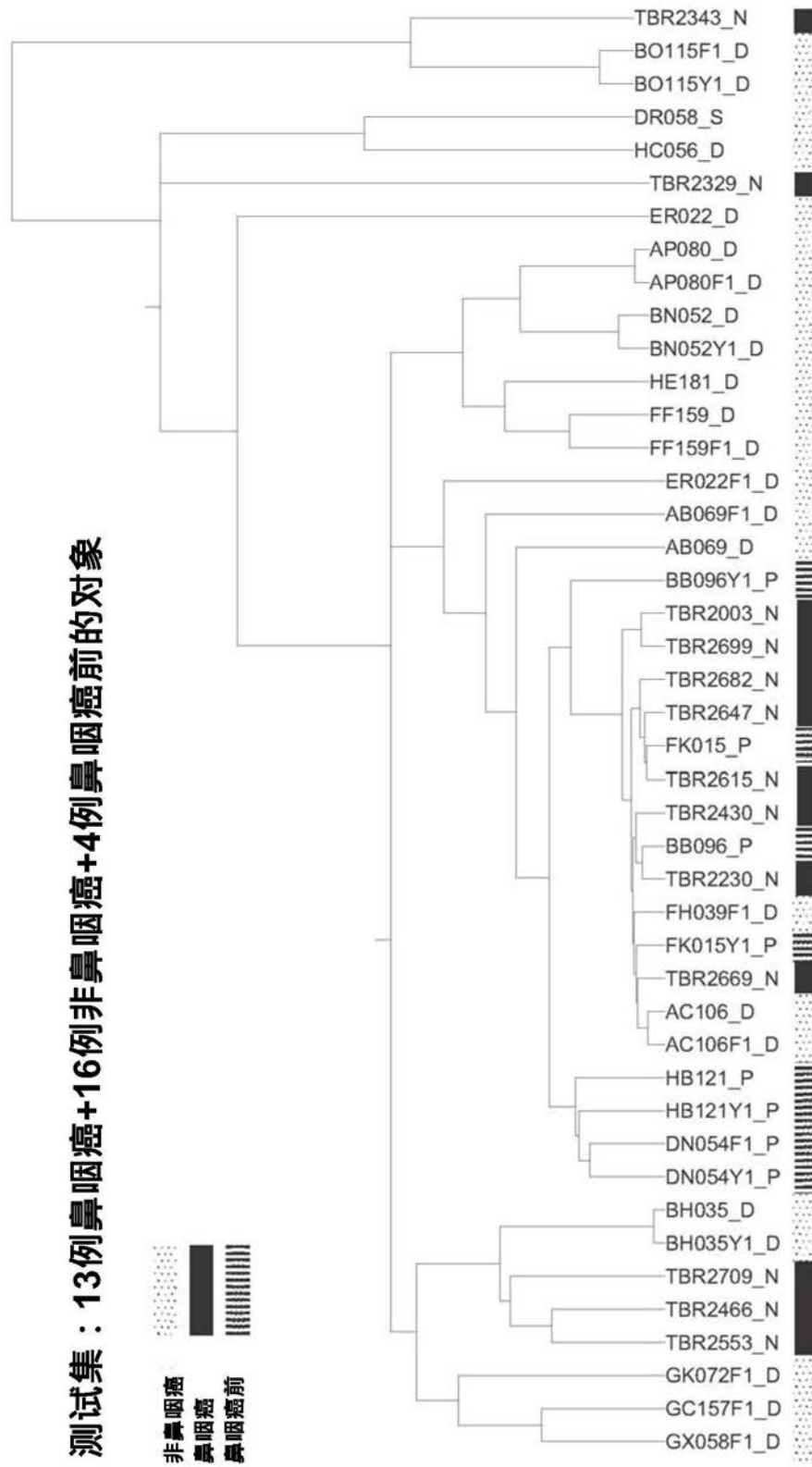


图5

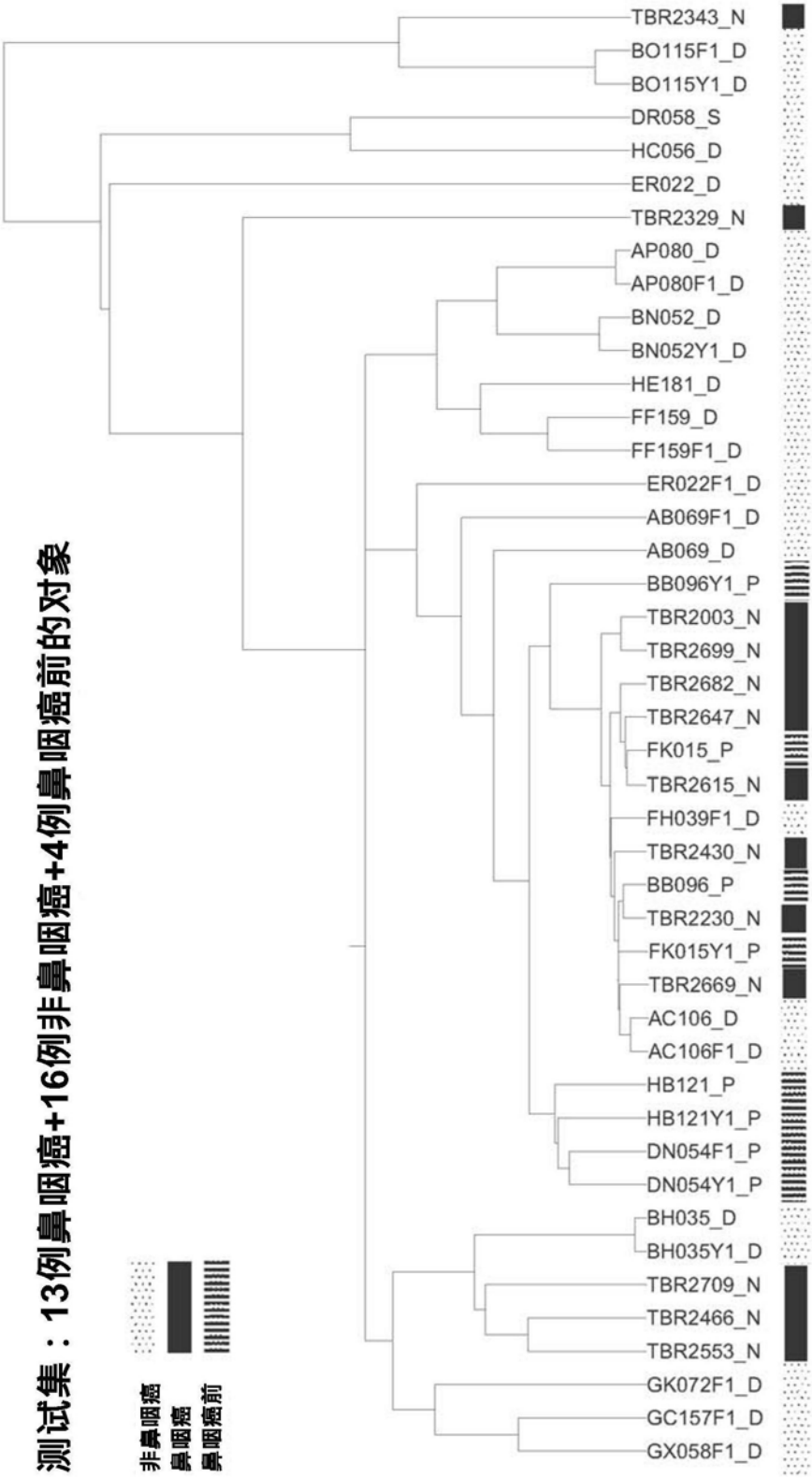


图6

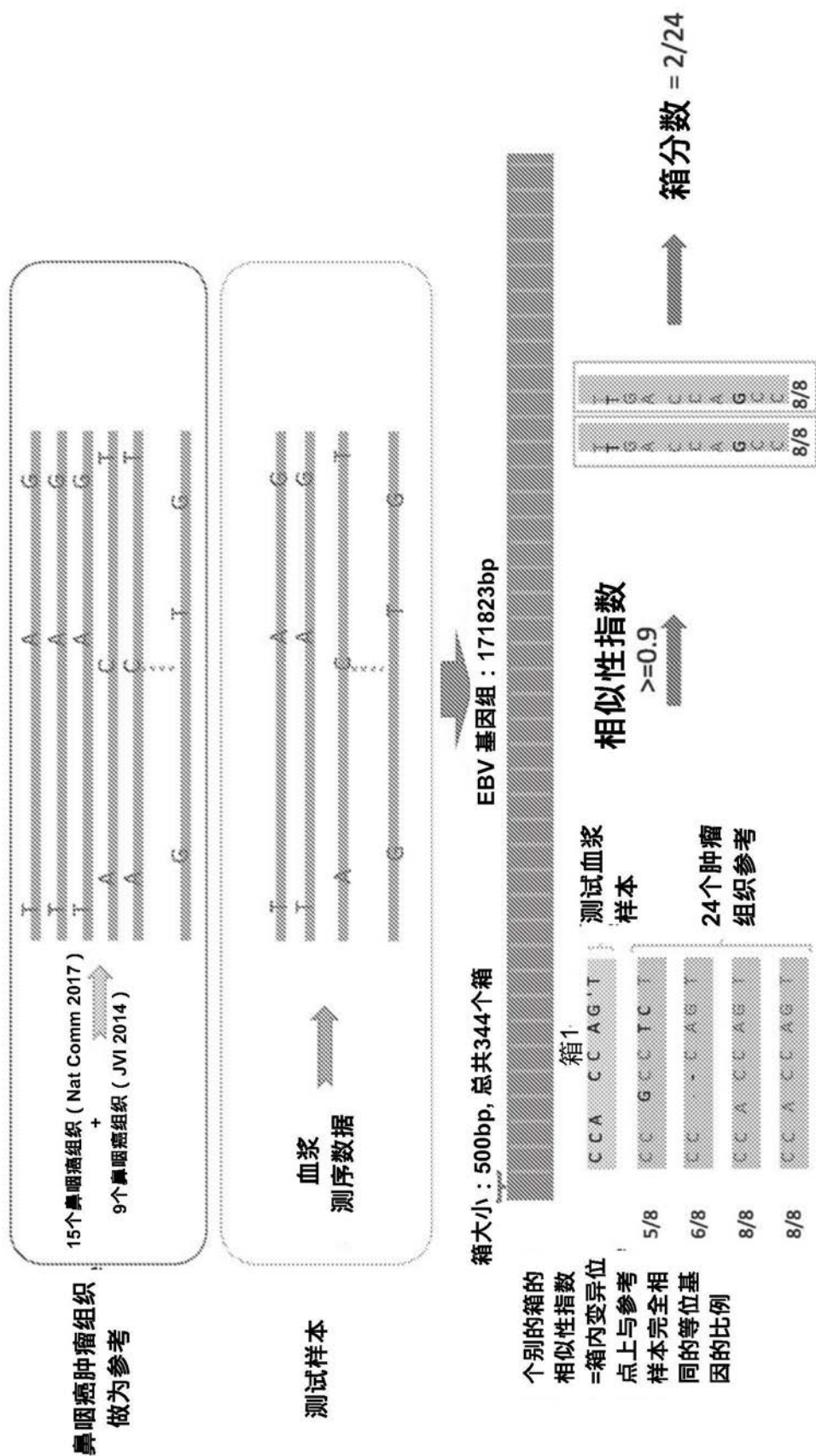


图7

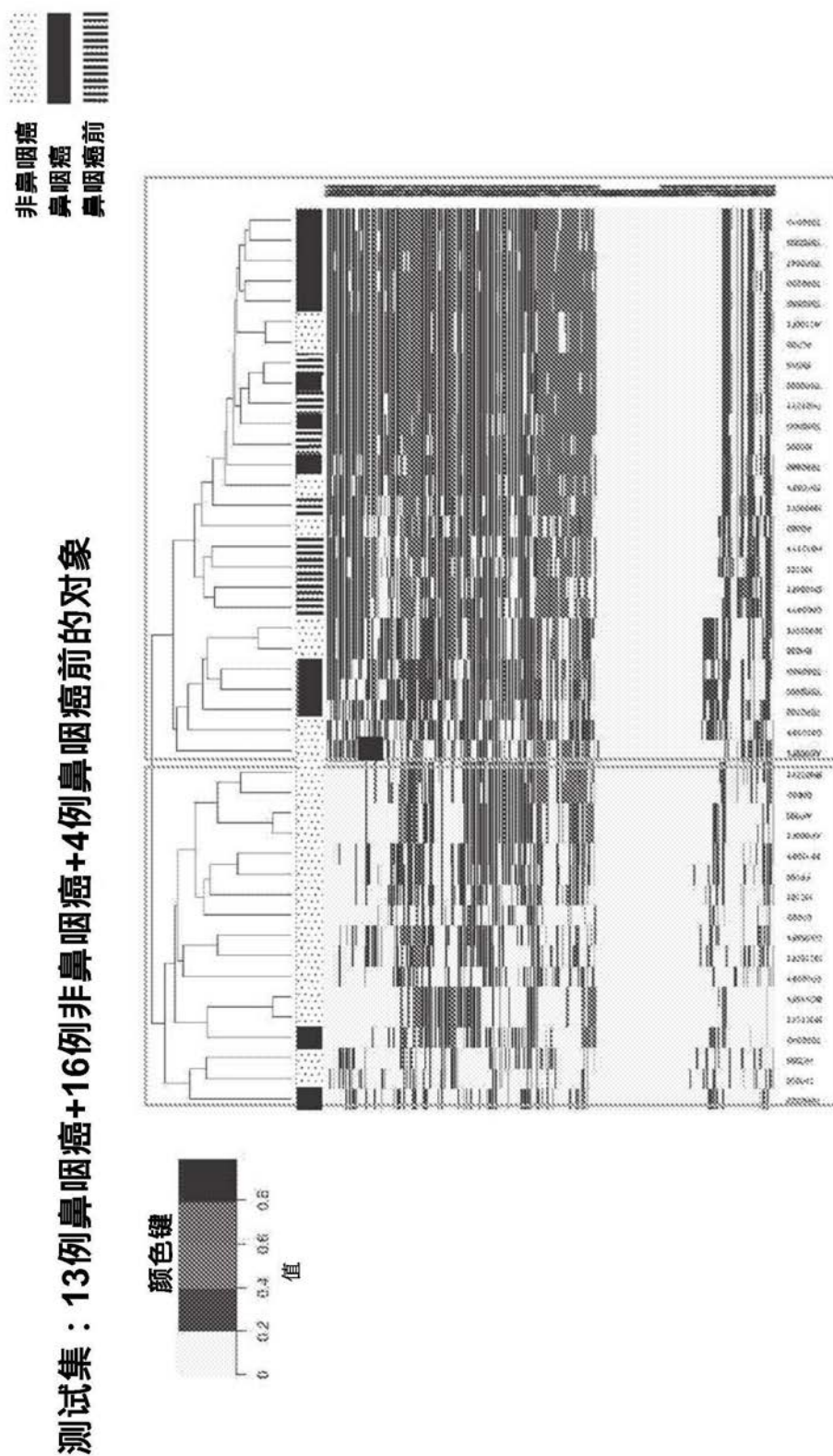


图8

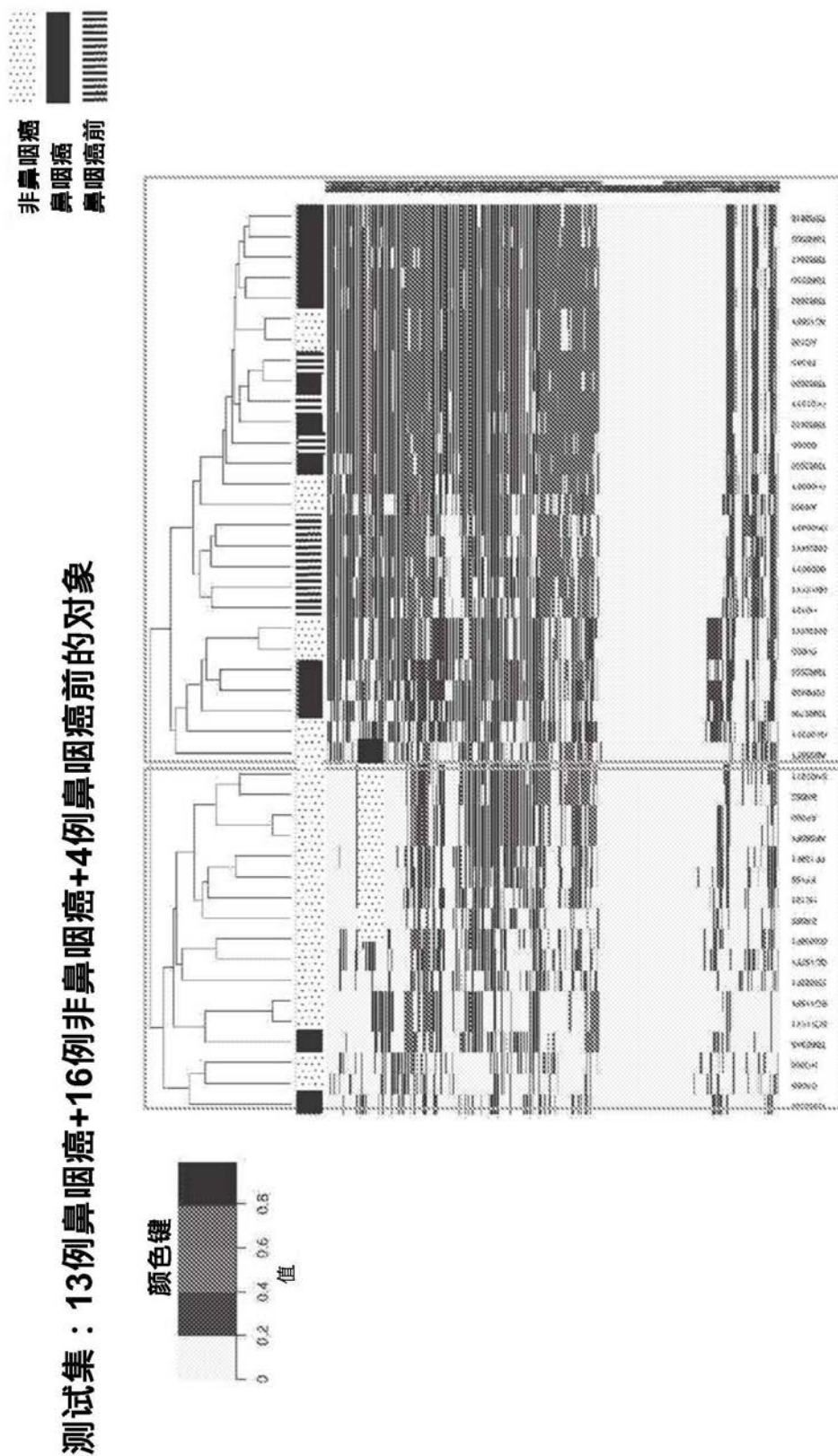


图9

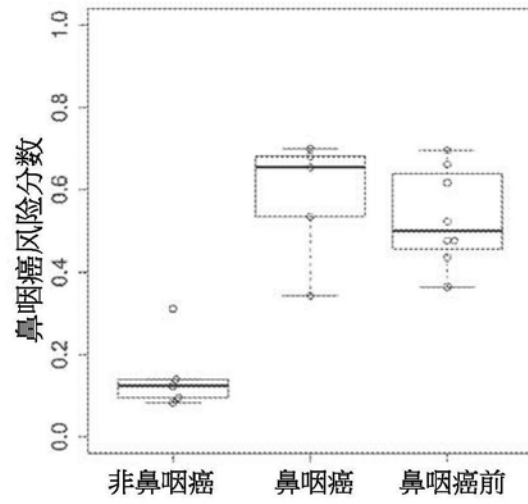


图10A

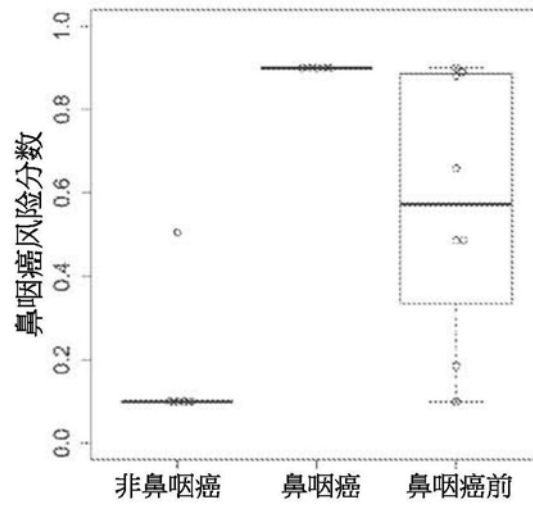


图10B

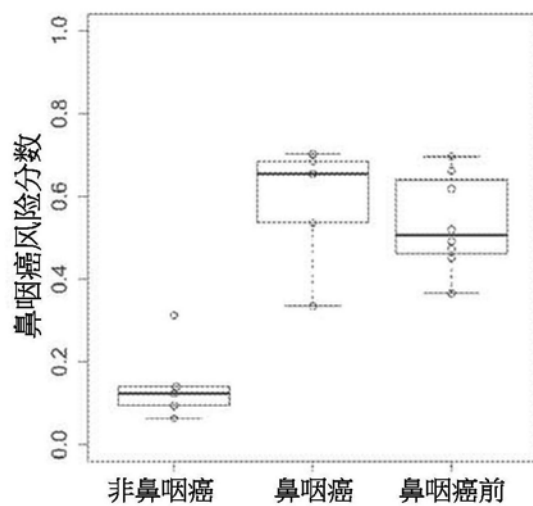


图10C

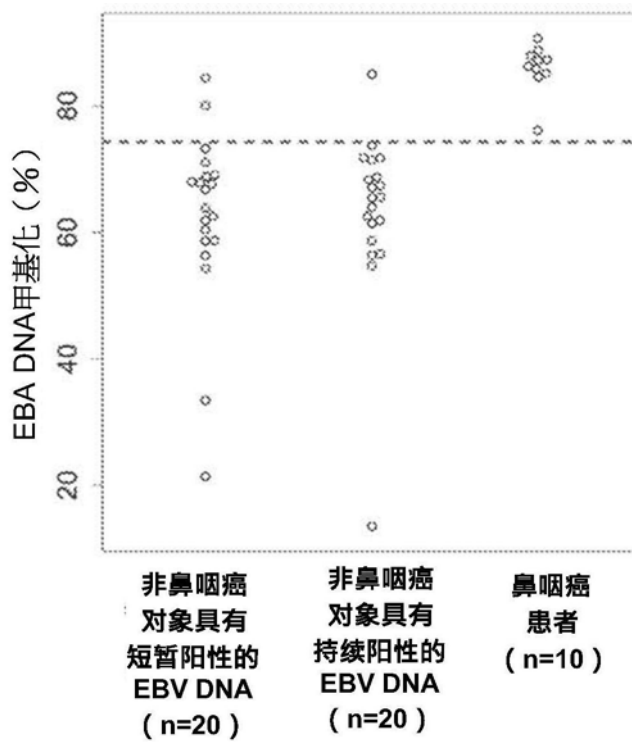


图11

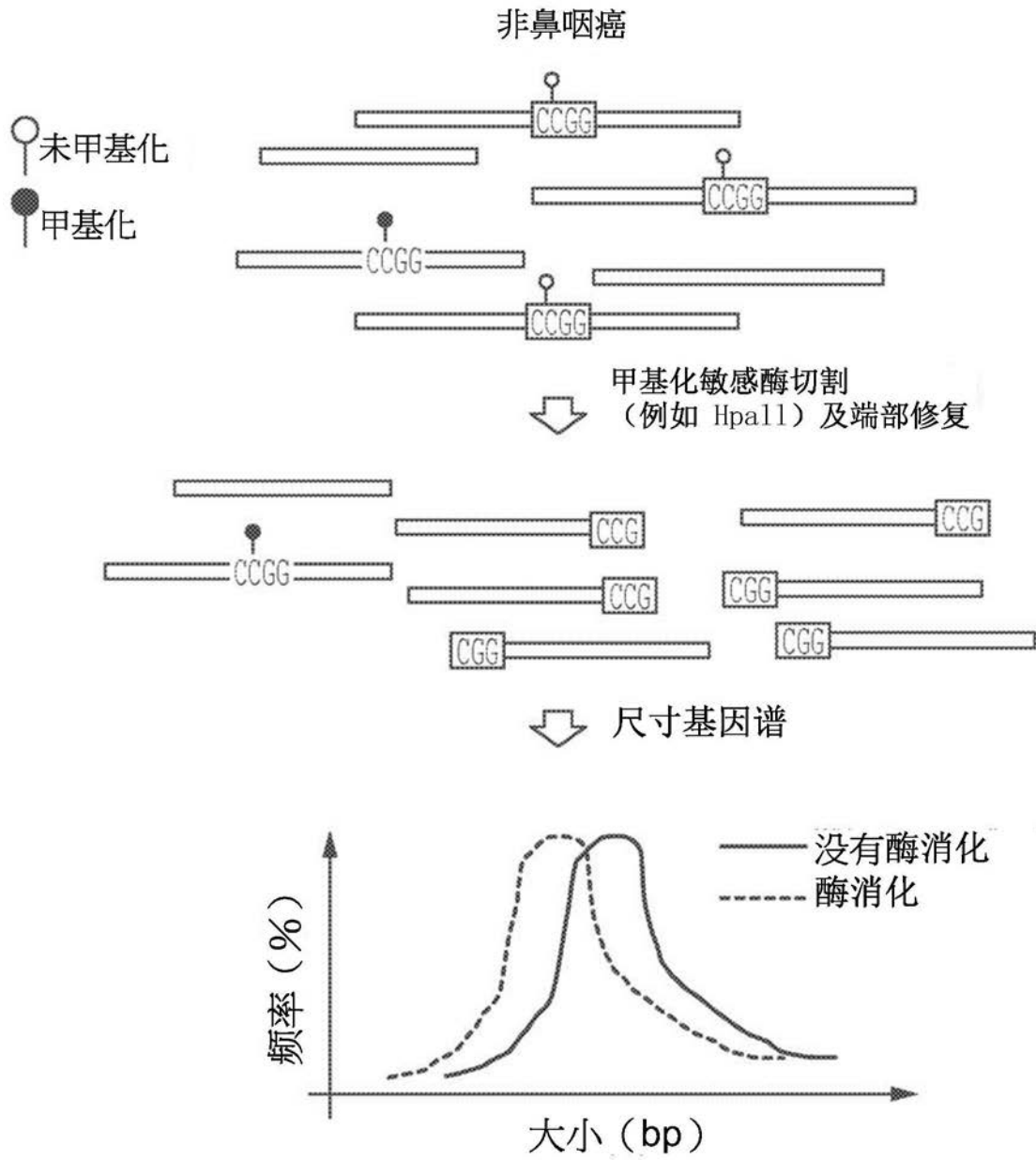


图12

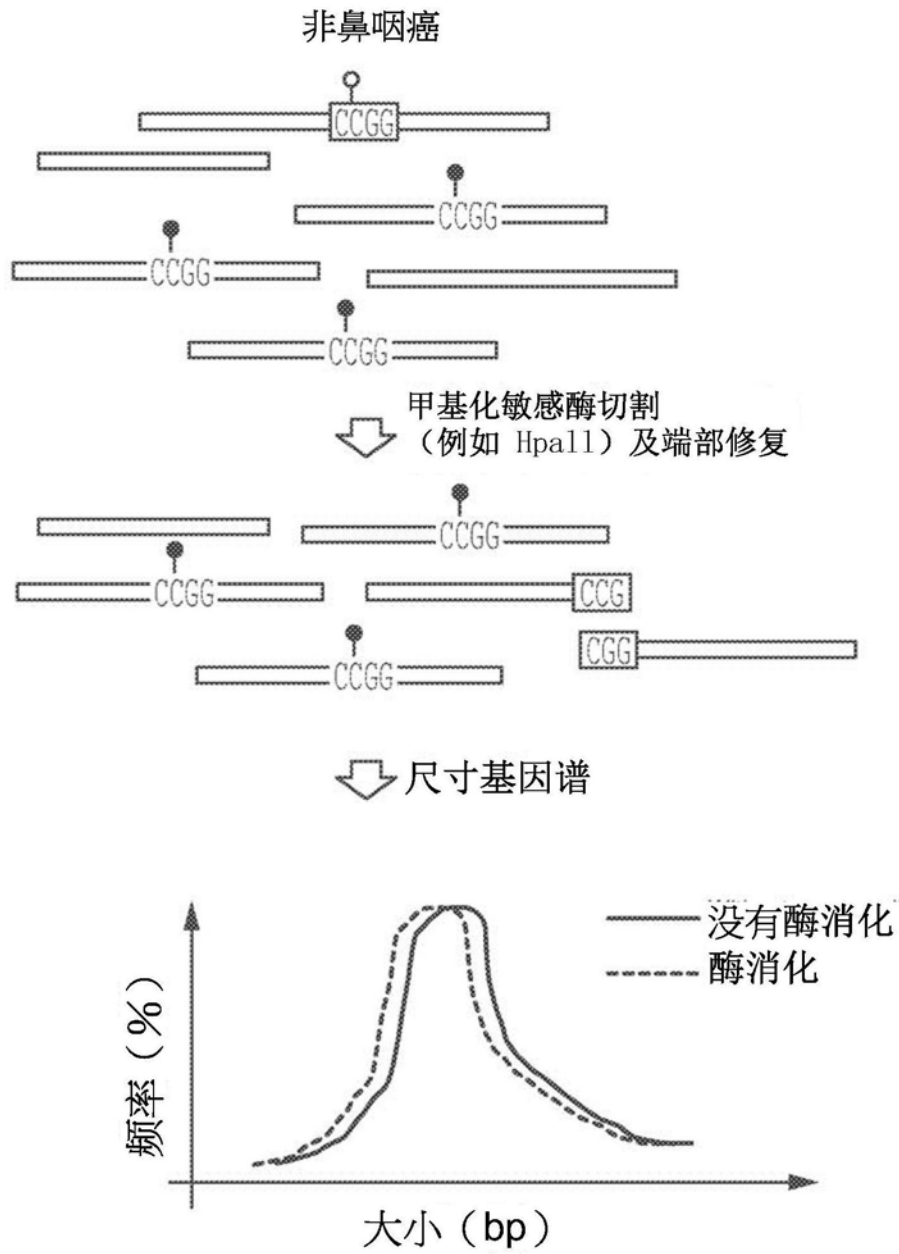


图13

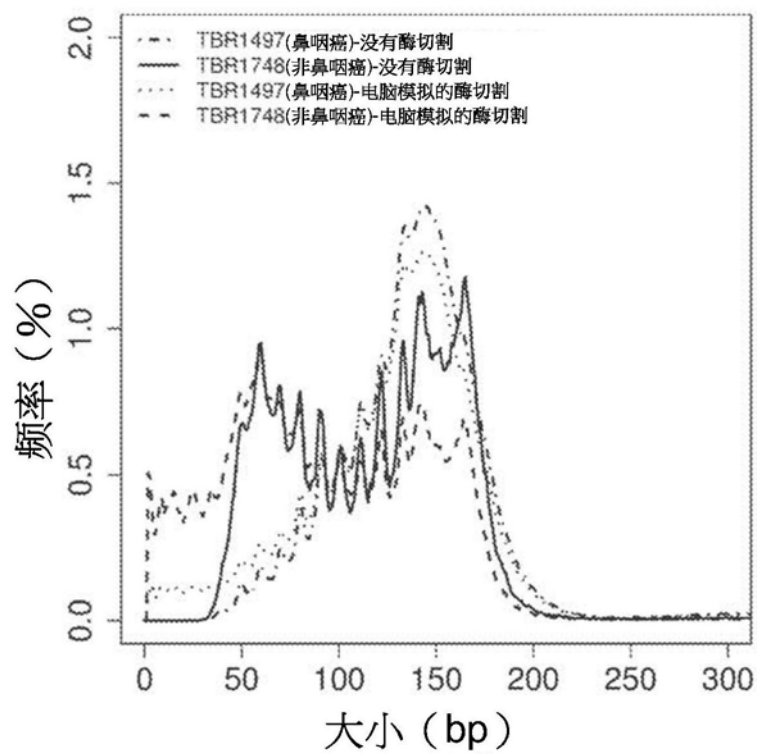


图14

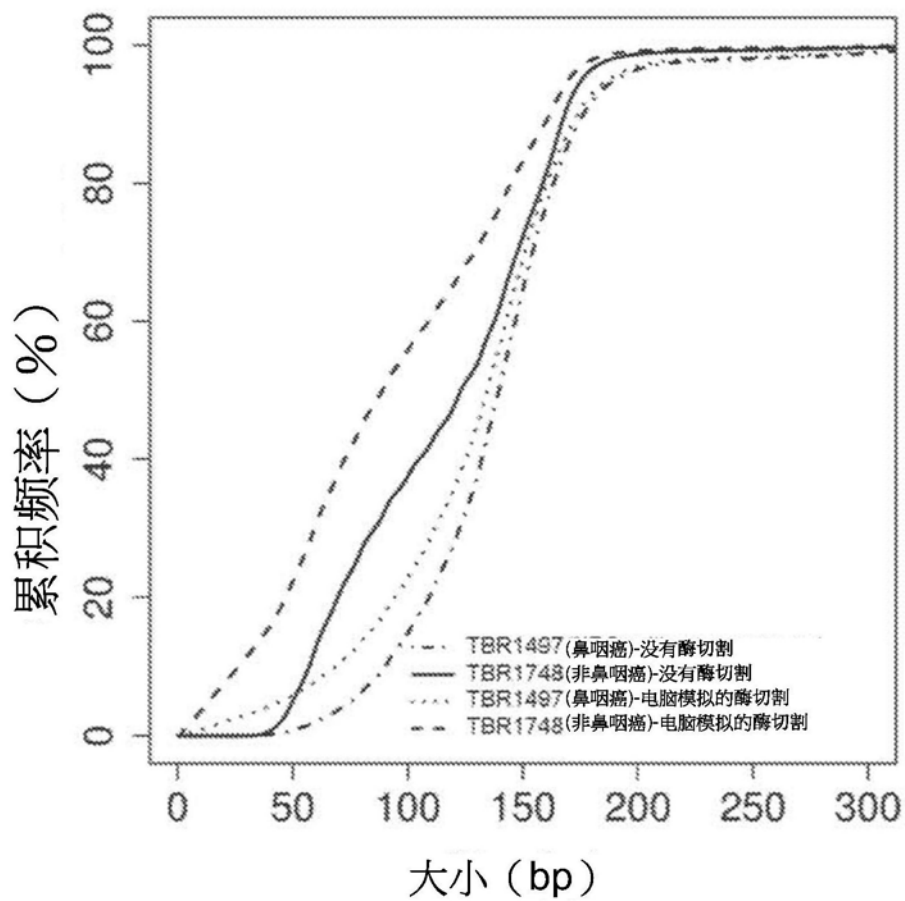


图15

来自测试样本的血浆EBV DNA  
读数与EBV基因组比对

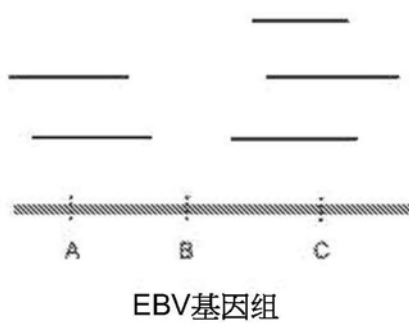


图16A

## 训练模型

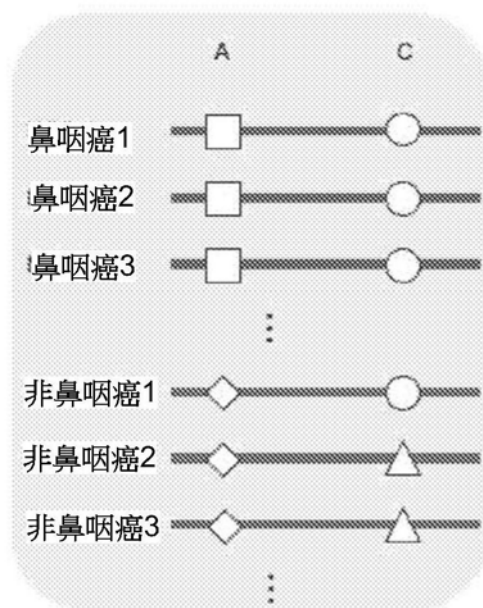


图16B

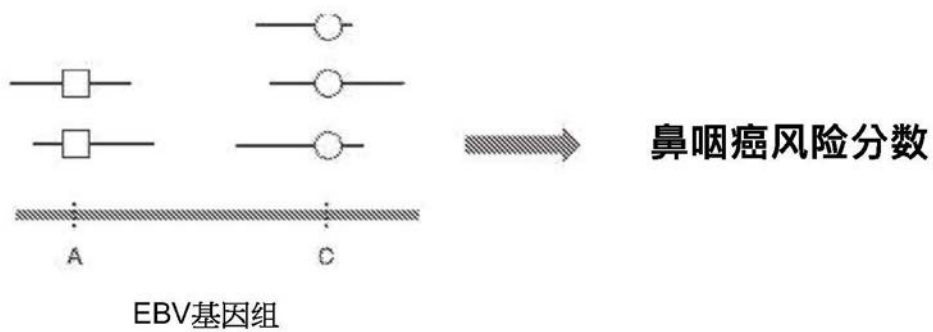


图16C

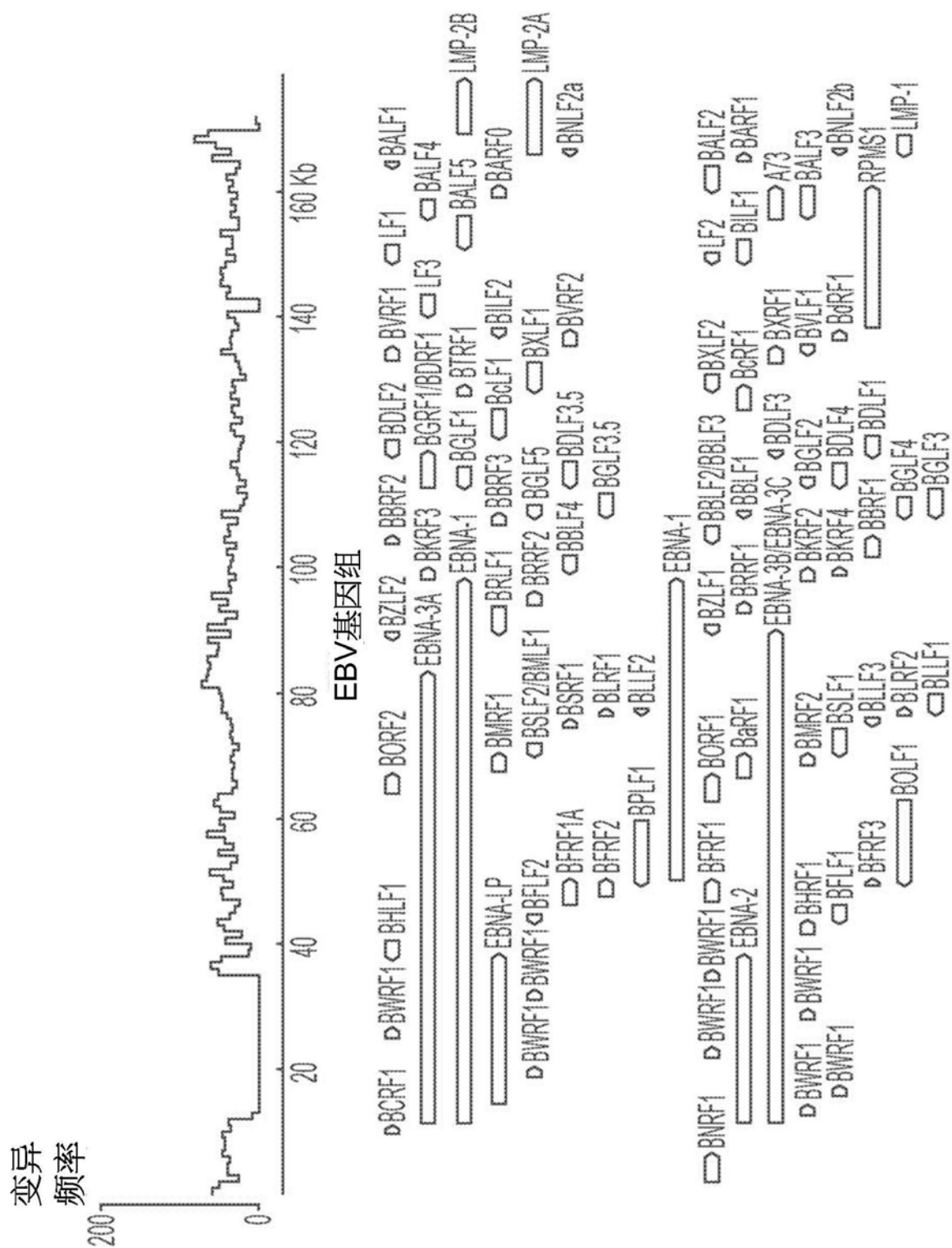


图16D

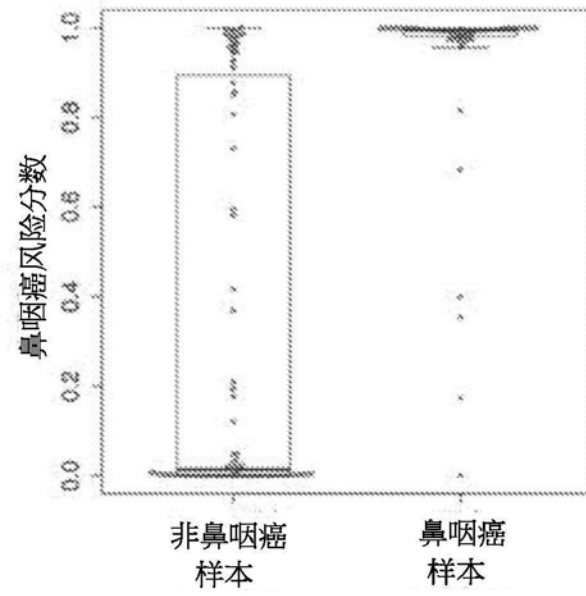


图17A

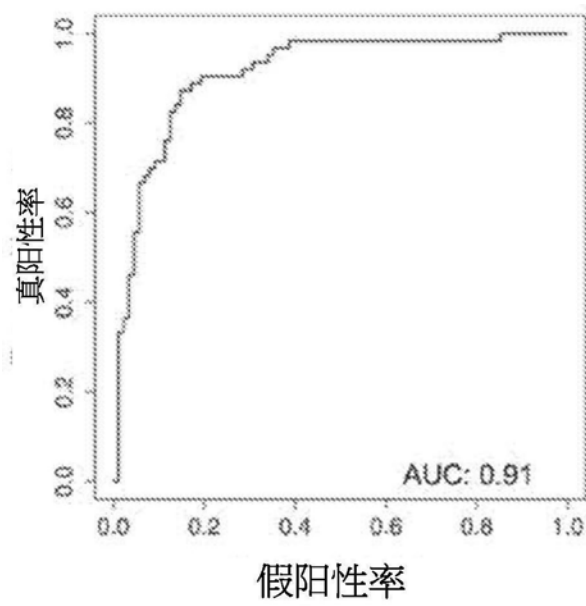


图17B

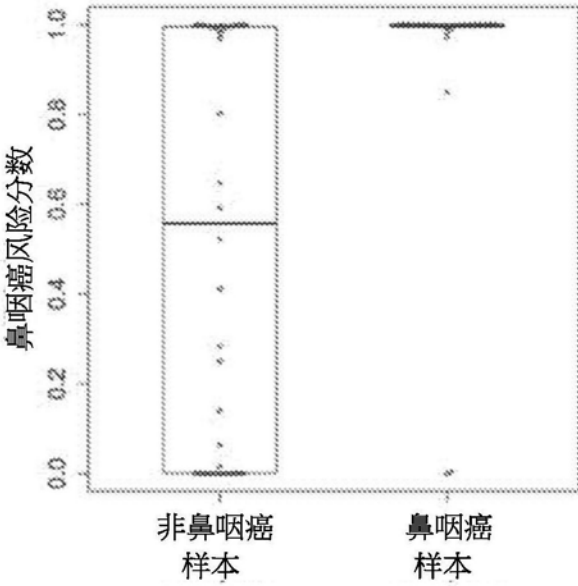


图18A

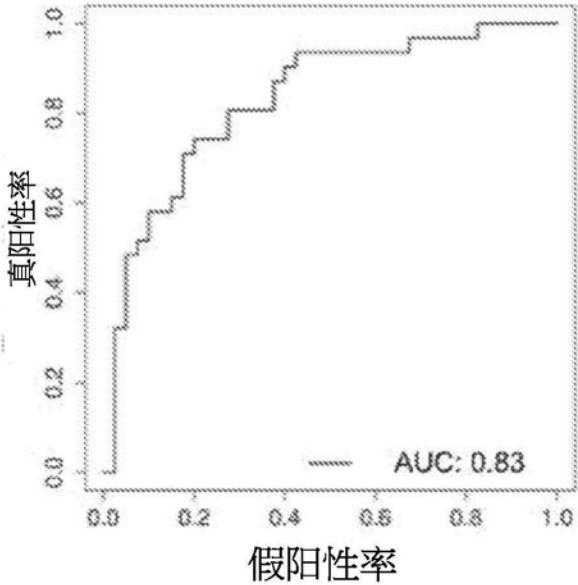


图18B

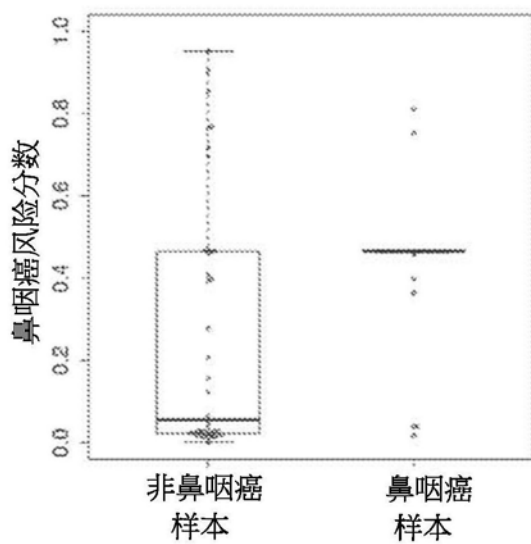


图19A

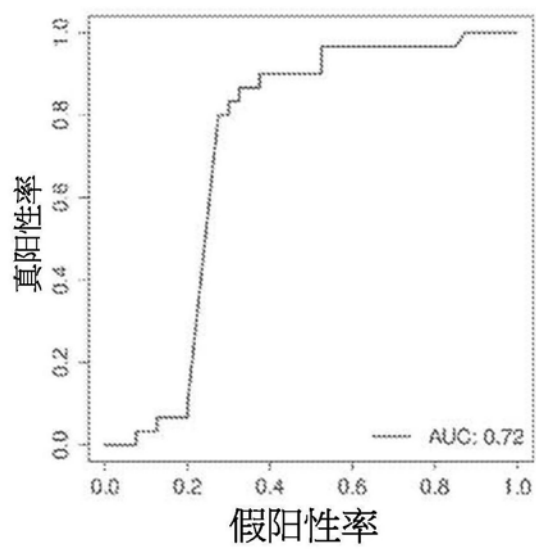


图19B

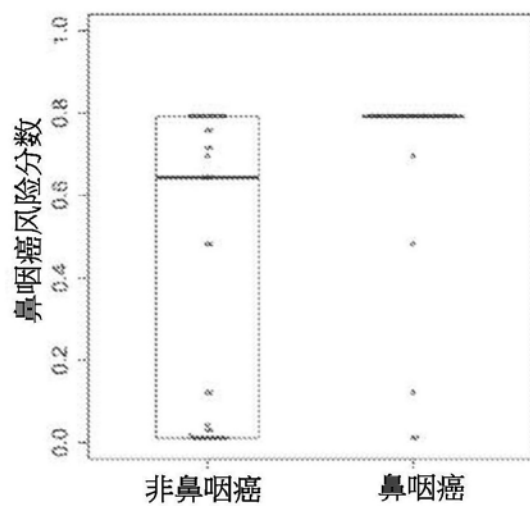


图20A

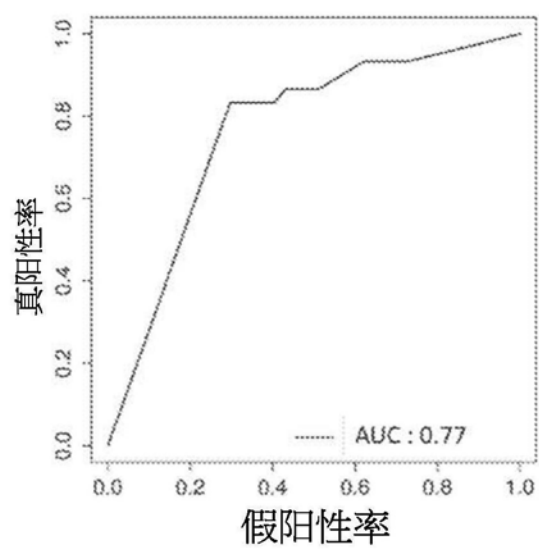


图20B

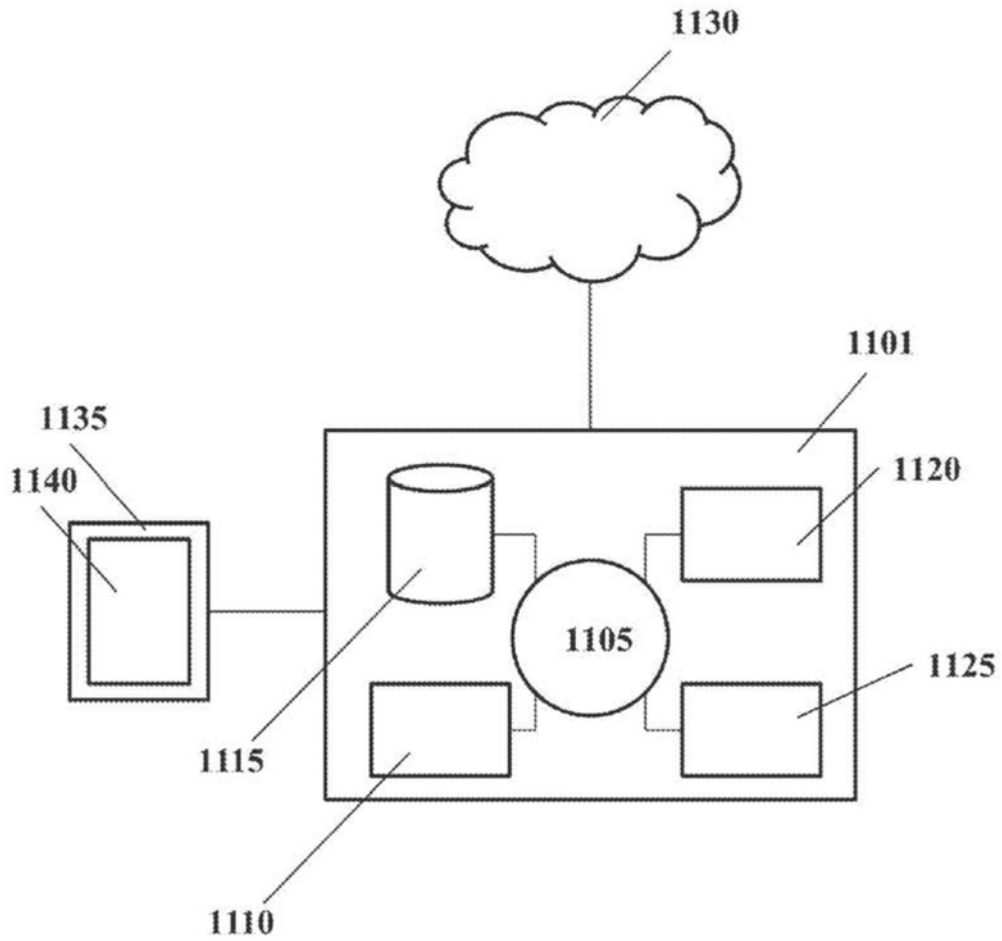


图21

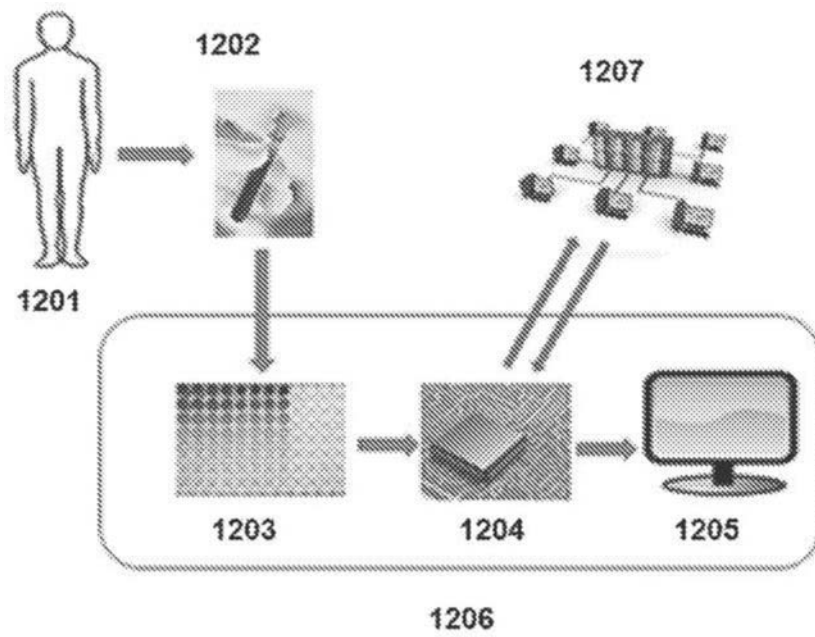


图22