



(12) 发明专利申请

(10) 申请公布号 CN 119013418 A

(43) 申请公布日 2024. 11. 22

(21) 申请号 202380033859.X

(22) 申请日 2023.03.09

(30) 优先权数据

63/318,479 2022.03.10 US

(85) PCT国际申请进入国家阶段日

2024.10.12

(86) PCT国际申请的申请数据

PCT/US2023/064065 2023.03.09

(87) PCT国际申请的公布数据

W02023/173034 EN 2023.09.14

(71) 申请人 麦克诺玛公司

地址 美国加利福尼亚州

(72) 发明人 塞丽娜·弗拉拉西奥

斯蒂芬·万德罗 埃迪·亚当斯

格雷戈里·D·普尔

(74) 专利代理机构 北京安信方达知识产权代理有限公司 11262

专利代理师 刘晓杰 郑霞

(51) Int.Cl.

C12Q 1/6886 (2006.01)

C12Q 1/6806 (2006.01)

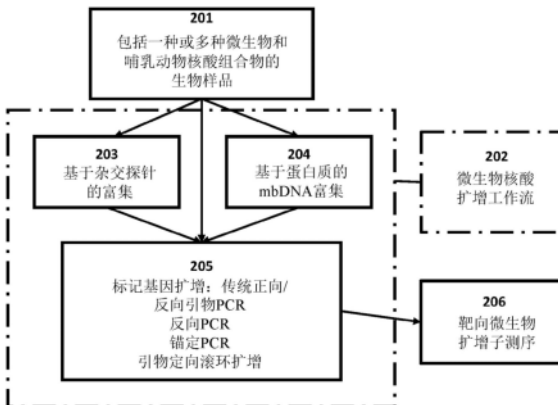
权利要求书13页 说明书33页 附图12页

(54) 发明名称

通过靶向微生物扩增子测序获得的疾病分类器

(57) 摘要

本文提供了诊断一种或多种如本文其它部分所述的疾病的多模态方法和系统。



1. 一种生成用于区分一个或多个受试者的癌症和非癌症健康状况的基因组特征集的方法,所述方法包括:

(a) 提供一个或多个受试者的一种或多种核酸分子和对应的健康状况;

(b) 扩增所述一种或多种核酸分子的一种或多种非哺乳动物核酸分子的一个或多个基因组特征,由此生成经扩增的一个或多个基因组特征;

(c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及

(d) 通过将所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度与所述一个或多个受试者的所述健康状况组合来生成特征集,所述特征集被配置成区分癌症和非癌症健康状况。

2. 根据权利要求1所述的方法,其中所述基因组特征包括微生物系统发育标记基因或其标记基因片段。

3. 根据权利要求2所述的方法,其中所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。

4. 根据权利要求2所述的方法,其中所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。

5. 根据权利要求3所述的方法,其中所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。

6. 根据权利要求4所述的方法,其中所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。

7. 根据权利要求2所述的方法,其中微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。

8. 根据权利要求1所述的方法,其中扩增包括进行聚合酶链式反应或其衍生物。

9. 根据权利要求8所述的方法,其中所述其衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。

10. 根据权利要求8所述的方法,其中所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。

11. 根据权利要求10所述的方法,其中所述一个或多个基因组特征包括线粒体DNA基因组特征。

12. 根据权利要求1所述的方法,其包括富集所述一种或多种核酸分子。

13. 根据权利要求12所述的方法,其中所述一种或多种核酸分子包括哺乳动物核酸分子、非哺乳动物核酸分子或其任何组合。

14. 根据权利要求13所述的方法,其中所述一种或多种哺乳动物核酸分子包括DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA或其任何组合。

15. 根据权利要求13所述的方法,其中所述一种或多种非哺乳动物核酸分子包括微生物DNA、微生物RNA、微生物游离DNA、微生物游离RNA或其任何组合。

16. 根据权利要求13所述的方法,其中所述富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述非哺乳动物核酸分子与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;

(b) 在促进所述一种或多种非哺乳动物核酸分子的靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针、所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸;

(c) 使未经结合的杂交探针和与所述一种或多种非哺乳动物核酸分子结合的经杂交探针分离;以及

(d) 洗涤与所述一种或多种非哺乳动物核酸分子结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸分子。

17. 根据权利要求16所述的方法,其中洗涤被配置成去除非特异性相关核酸和其它反应组分。

18. 根据权利要求13所述的方法,其中所述一种或多种核酸分子的富集包括非哺乳动物DNA富集。

19. 根据权利要求18所述的方法,其中所述非哺乳动物DNA富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;

(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物核酸分子或所述一种或多种非哺乳动物核酸分子的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;

(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白;以及

(d) 洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸以供扩增。

20. 根据权利要求19所述的方法,其中洗涤被配置成去除非特异性相关核酸分子和所述蛋白质-DNA结合反应组分的剩余部分。

21. 根据权利要求1所述的方法,其中所述一种或多种核酸分子源自所述一个或多个受试者的一个或多个生物样品。

22. 根据权利要求21所述的方法,其中所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。

23. 根据权利要求1所述的方法,其中所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。

24. 根据权利要求1所述的方法,其包括过滤所述一个或多个非哺乳动物测序读段。

25. 根据权利要求24所述的方法,其中过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。

26. 根据权利要求25所述的方法,其中过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。

27. 根据权利要求26所述的方法,其包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。

28. 根据权利要求27所述的方法,其中去污包括计算机模拟去污。
29. 根据权利要求27所述的方法,其中去污被配置成去除非内源性微生物测序读段,由此生成经去污的微生物分类学分配和相关数量的测序读段。
30. 根据权利要求26所述的方法,其中映射用QIIME2或其支持的其它版本进行。
31. 根据权利要求26所述的方法,其中所述一个或多个微生物参考数据库包括细菌16S rRNA数据库Greengenes;细菌、真菌和古菌rRNA数据库SILVA;真核核糖体ITS区数据库UNITE;源自可公开获得且完整的微生物基因组序列的定制数据库;或其任何组合。
32. 根据权利要求1所述的方法,其中所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度包括微生物功能性基因丰度、生化通路丰度或其任何组合。
33. 根据权利要求29所述的方法,其包括预测所述经去污的微生物分类学分配的宏基因组功能性含量,由此产生一个或多个功能性丰度。
34. 根据权利要求33所述的方法,其中预测所述宏基因组功能性含量通过PICRUST2进行。
35. 根据权利要求22所述的方法,其中所述液体活检样品包括:血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。
36. 根据权利要求1所述的方法,其中所述癌症包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺腺癌、前列腺癌、皮肤癌或其任何组合。
37. 根据权利要求36所述的方法,其中所述肺癌包括非小细胞肺癌。
38. 根据权利要求1所述的方法,其中所述非癌症健康状况包括健康、疾病或其任何组合。
39. 根据权利要求38所述的方法,其中所述疾病状态包括肺病,其中所述肺病包括:类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。
40. 根据权利要求1所述的方法,其中所述癌症包括I期、II期或III期癌症。
41. 根据权利要求1所述的方法,其包括生成经训练的预测模型,其中所述经训练的预测模型是用所述一个或多个受试者的所述特征集和所述健康状况训练的。
42. 根据权利要求41所述的方法,其中所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。
43. 根据权利要求41所述的方法,其中所述经训练的预测模型包括正规化机器学习模型。
44. 根据权利要求42所述的方法,其中所述机器学习模型包括机器学习分类器。
45. 根据权利要求42所述的方法,其中所述机器学习模型包括梯度提升机、神经网络、支持向量机、k均值、分类树、随机森林、回归或其任何组合。
46. 一种诊断一个或多个受试者的癌症的方法,所述方法包括:
- (a) 提供一个或多个受试者的一种或多种核酸分子;
- (b) 扩增所述一种或多种核酸分子的一种或多种非哺乳动物核酸分子的一个或多个基因组特征,由此生成经扩增的一个或多个基因组特征;
- (c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及

(d) 至少作为将所述一个或多个基因组特征作为输入提供给经训练的预测模型的结果,输出对所述一个或多个受试者的癌症或非癌症健康状况的诊断。

47. 根据权利要求46所述的方法,其中所述非哺乳动物核酸分子包括微生物核酸。

48. 根据权利要求46所述的方法,其中所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。

49. 根据权利要求46所述的方法,其中所述一种或多种核酸分子包括DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA、游离微生物DNA、游离微生物RNA或其任何组合的总群体。

50. 根据权利要求46所述的方法,其中所述一个或多个基因组特征包括微生物系统发育标记基因或其标记基因片段。

51. 根据权利要求50所述的方法,其中所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。

52. 根据权利要求50所述的方法,其中所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。

53. 根据权利要求51所述的方法,其中所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。

54. 根据权利要求52所述的方法,其中所述真菌标记基因可包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。

55. 根据权利要求50所述的方法,其中所述微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。

56. 根据权利要求46所述的方法,其中扩增包括进行聚合酶链式反应或其衍生物。

57. 根据权利要求56所述的方法,其中所述其衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。

58. 根据权利要求56所述的方法,其中所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。

59. 根据权利要求58所述的方法,其中所述一个或多个基因组特征包括线粒体DNA基因组特征。

60. 根据权利要求46所述的方法,其包括富集所述一种或多种核酸分子。

61. 根据权利要求60所述的方法,其中所述一种或多种核酸分子包括一种或多种哺乳动物核酸分子、非哺乳动物核酸分子或其任何组合。

62. 根据权利要求61所述的方法,其中富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;

(b) 在促进所述一种或多种非哺乳动物核酸分子的靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物核酸分子以及所述一种或多种非哺乳动物核酸分子;

(c) 使未经结合的杂交探针和与所述一种或多种非哺乳动物核酸分子结合的经杂交探

针分离;以及

(d) 洗涤与所述一种或多种非哺乳动物核酸分子结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸分子。

63. 根据权利要求62所述的方法,其中洗涤被配置成去除非特异性相关核酸分子和其它反应组分。

64. 根据权利要求61所述的方法,其中所述一种或多种核酸分子的富集包括非哺乳动物DNA富集。

65. 根据权利要求64所述的方法,其中所述非哺乳动物DNA富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;

(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物核酸分子或所述一种或多种非哺乳动物核酸的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;

(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸分子片段结合的重组CXXC结构域蛋白;

(d) 洗涤与所述非甲基化CpG核酸分子片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸以供扩增。

66. 根据权利要求65所述的方法,其中所述重组CXXC结构域蛋白包括:重组的含有锌指CXXC结构域的蛋白KDM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、源自其的重组CXXC结构域或其任何组合。

67. 根据权利要求65所述的方法,其中洗涤被配置成去除非特异性相关核酸分子和所述蛋白质-DNA结合反应组分的所述剩余部分。

68. 根据权利要求46所述的方法,其中所述一种或多种核酸分子源自所述一个或多个受试者的一个或多个生物样品。

69. 根据权利要求68所述的方法,其中所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。

70. 根据权利要求61所述的方法,其中所述一种或多种哺乳动物核酸分子包括DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA或其任何组合。

71. 根据权利要求61所述的方法,其中所述一种或多种非哺乳动物核酸分子包括微生物DNA、微生物RNA、微生物游离DNA、微生物游离RNA或其任何组合。

72. 根据权利要求61所述的方法,其包括过滤所述一个或多个非哺乳动物测序读段。

73. 根据权利要求72所述的方法,其中过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。

74. 根据权利要求73所述的方法,其中过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。

75. 根据权利要求74所述的方法,其包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。

76. 根据权利要求75所述的方法,其中去污包括计算机模拟去污。

77. 根据权利要求75所述的方法, 其中去污被配置成去除非内源性微生物测序读段, 由此生成经去污的微生物分类学分配和相关数量的测序读段。

78. 根据权利要求74所述的方法, 其中映射用QIIME2或其支持的其它版本进行。

79. 根据权利要求74所述的方法, 其中所述一个或多个微生物参考数据库包括细菌16S rRNA数据库Greengenes; 细菌、真菌和古菌rRNA数据库SILVA; 真核核糖体ITS区数据库UNITE; 源自可公开获得且完整的微生物基因组序列的定制数据库; 或其任何组合。

80. 根据权利要求46所述的方法, 其中所述一个或多个基因组特征包括所述一个或多个非哺乳动物测序读段的微生物功能性基因、生化通路的丰度或其任何组合丰度。

81. 根据权利要求77所述的方法, 其包括预测所述经去污的微生物分类学分配的宏基因组功能性含量, 由此产生一个或多个功能性丰度。

82. 根据权利要求81所述的方法, 其中预测所述宏基因组功能性含量通过PICRUSt2进行。

83. 根据权利要求69所述的方法, 其中所述液体活检样品包括: 血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。

84. 根据权利要求46所述的方法, 其中所述癌症健康状况包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺腺癌、前列腺癌、皮肤癌或其任何组合。

85. 根据权利要求84所述的方法, 其中所述肺癌包括非小细胞肺癌。

86. 根据权利要求46所述的方法, 其中所述非癌症状态包括健康非癌症状态、疾病非癌症状态或其任何组合。

87. 根据权利要求86所述的方法, 其中所述疾病包括肺病, 其中所述肺病包括: 类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎或其任何组合。

88. 根据权利要求46所述的方法, 其中所述癌症健康状况包括I期、II期或III期癌症。

89. 根据权利要求46所述的方法, 其中所述经训练的预测模型是用一个或多个受试者的特征集和健康状况训练的。

90. 根据权利要求46所述的方法, 其中所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。

91. 根据权利要求46所述的方法, 其中所述经训练的预测模型包括正规化机器学习模型。

92. 根据权利要求90所述的方法, 其中所述机器学习模型包括机器学习分类器。

93. 根据权利要求90所述的方法, 其中所述机器学习模型包括梯度提升机机器学习模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组合。

94. 一种用于诊断一个或多个受试者的癌或非癌健康状况的系统, 所述系统包括:

(a) 一个或多个处理器; 以及

(b) 包含软件的非暂时性计算机可读存储介质, 所述软件被配置成使所述一个或多个处理器:

(i) 接收一个或多个受试者的生物样品的所述一个或多个受试者的一个或多个核酸分子测序读段, 其中所述一个或多个核酸分子测序读段包括一种或多种非哺乳动物核酸分子的经扩增的一个或多个基因组特征的序列; 以及

(ii) 至少作为将所述一种或多种非哺乳动物核酸测序读段的一个或多个基因组特征作为输入提供给经训练的预测模型的结果,输出对所述一个或多个受试者的癌或非癌健康状况的诊断。

95. 根据权利要求94所述的系统,其中所述非哺乳动物核酸分子包括微生物核酸。

96. 根据权利要求94所述的系统,其中所述生物样品包括组织活检样品、液体活检样品或其任何组合。

97. 根据权利要求94所述的系统,其中所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。

98. 根据权利要求94所述的系统,其中所述一个或多个核酸分子测序读段包括DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA、游离微生物DNA、游离微生物RNA或其任何组合的测序读段。

99. 根据权利要求94所述的系统,其中所述一个或多个基因组特征包括微生物系统发育标记基因或其标记基因片段。

100. 根据权利要求99所述的系统,其中所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。

101. 根据权利要求99所述的系统,其中所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。

102. 根据权利要求100所述的系统,其中所述细菌标记基因包括核糖体RNA基因。

103. 根据权利要求102所述的系统,其中所述核糖体RNA基因包括5S核糖体RNA基因、16S核糖体RNA基因、23S核糖体RNA基因或其任何组合。

104. 根据权利要求100所述的系统,其中所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。

105. 根据权利要求101所述的系统,其中所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。

106. 根据权利要求94所述的系统,其中所述一种或多种非哺乳动物核酸分子的所述经扩增的一个或多个基因组特征是通过聚合酶链式反应或其衍生物扩增的。

107. 根据权利要求106所述的系统,其中所述其衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。

108. 根据权利要求106所述的系统,其中所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。

109. 根据权利要求108所述的系统,其中所述一个或多个基因组特征包括线粒体DNA基因组特征。

110. 根据权利要求94所述的系统,其中所述一个或多个核酸分子测序读段包括一种或多种经富集的核酸分子的测序读段。

111. 根据权利要求110所述的系统,其中所述一个或多个核酸分子测序读段的所述一种或多种核酸分子包括哺乳动物核酸分子测序读段、非哺乳动物核酸分子测序读段或其任何组合。

112. 根据权利要求110所述的系统,其中所述一种或多种经富集的核酸分子是通过以下产生的:

(a) 将所述一个或多个核酸分子测序读段的一种或多种核酸分子与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;

(b) 在促进所述一种或多种核酸分子的靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种核酸分子;

(c) 使未经结合的杂交探针和与一种或多种非哺乳动物核酸分子结合的经杂交探针分离;以及

(d) 洗涤与所述一种或多种非哺乳动物核酸分子结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸分子。

113. 根据权利要求112所述的系统,其中洗涤被配置成去除非特异性相关核酸和其它反应组分。

114. 根据权利要求110所述的系统,其中所述一种或多种经富集的核酸分子是通过非哺乳动物DNA富集产生的。

115. 根据权利要求114所述的系统,其中所述非哺乳动物DNA富集包括:

(a) 将所述一种或多种核酸分子与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;

(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种核酸分子的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;

(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸分子片段结合的重组CXXC结构域蛋白;

(d) 洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸分子以供扩增。

116. 根据权利要求115所述的系统,其中洗涤被配置成去除非特异性相关核酸分子和所述蛋白质-DNA结合反应组分的所述剩余部分。

117. 根据权利要求94所述的系统,其中所述软件将所述一个或多个处理器配置成过滤所述一个或多个核酸分子测序读段。

118. 根据权利要求117所述的系统,其中所述过滤包括过滤所述一个或多个核酸分子测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。

119. 根据权利要求118所述的系统,其中过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。

120. 根据权利要求119所述的系统,其中所述软件将所述一个或多个处理器配置成对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。

121. 根据权利要求120所述的系统,其中去污包括计算机模拟去污。

122. 根据权利要求120所述的系统,其中所述去污被配置成去除非内源性微生物测序读段,由此生成经去污的微生物分类学分配和相关数量的测序读段。

123. 根据权利要求119所述的系统,其中映射用QIIME2或其支持的其它版本进行。

124. 根据权利要求119所述的系统,其中所述一个或多个微生物参考数据库包括细菌

16S rRNA数据库Greengenes;细菌、真菌和古菌rRNA数据库SILVA;真核核糖体ITS区数据库UNITE;源自可公开获得且完整的微生物基因组序列的定制数据库;或其任何组合。

125. 根据权利要求111所述的系统,其中所述经扩增的一个或多个基因组特征包括一个或多个非哺乳动物测序读段的微生物功能性基因、生化通路的丰度或其任何组合丰度。

126. 根据权利要求122所述的系统,其中所述软件将所述一个或多个处理器配置成预测所述经去污的微生物分类学分配的宏基因组功能性含量,由此产生一个或多个功能性丰度。

127. 根据权利要求126所述的系统,其中预测所述宏基因组功能性含量通过PICRUST2进行。

128. 根据权利要求94所述的系统,其中所述生物样品包括组织活检样品、液体活检样品或其任何组合。

129. 根据权利要求128所述的系统,其中所述液体活检样品包括:血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。

130. 根据权利要求94所述的系统,其中所述癌健康状况包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺腺癌、前列腺癌、皮肤癌或其任何组合。

131. 根据权利要求130所述的系统,其中所述肺癌包括非小细胞肺癌。

132. 根据权利要求94所述的系统,其中所述非癌健康状况包括健康非癌症状态、疾病非癌症状态或其任何组合。

133. 根据权利要求132所述的系统,其中所述疾病状态包括肺病,其中所述肺病包括:类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。

134. 根据权利要求94所述的系统,其中所述癌健康状况包括I期、II期或III期癌症。

135. 根据权利要求94所述的系统,其中所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。

136. 根据权利要求94所述的系统,其中所述经训练的预测模型包括正规化机器学习模型。

137. 根据权利要求135所述的系统,其中所述机器学习模型包括机器学习分类器。

138. 根据权利要求135所述的系统,其中所述机器学习模型包括梯度提升机器学习模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组合。

139. 根据权利要求94所述的系统,其中所述癌健康状况或所述非癌健康状况包括癌症或疾病的类别、组织特异性位置或其任何组合。

140. 根据权利要求94所述的系统,其中所述癌健康状况包括一种或多种癌症类型、一种或多种癌症亚型、癌症分期、癌症预后或其任何组合。

141. 根据权利要求94所述的系统,其中所述经训练的预测模型用于预测所述一个或多个受试者的癌症疗法应答。

142. 根据权利要求94所述的系统,其中所述经训练的预测模型用于选择针对所述一个或多个受试者的疾病或癌症的最佳疗法。

143. 根据权利要求94所述的系统,其中所述经训练的预测模型用于对所述一个或多个

受试者的一种或多种癌症或一种或多种疾病对疗法的应答过程进行纵向建模并且然后调整治疗方案。

144. 根据权利要求94所述的系统,其中所述癌健康状况包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、肺腺癌、肺鳞状细胞癌、淋巴样增生性弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合癌症。

145. 根据权利要求94所述的系统,其中所述经训练的预测模型被配置成从所述一个或多个核酸分子测序读段中去除受污染的非哺乳动物特征,同时选择性地保留所述一种或多种非哺乳动物核酸分子的未受污染的非哺乳动物特征的序列。

146. 一种生成用于区分一个或多个受试者的癌症类型的特征集的方法,所述方法包括:

(a) 提供一个或多个受试者的一种或多种核酸分子和对应的健康状况;

(b) 扩增所述一种或多种核酸分子的一种或多种非哺乳动物核酸分子的一个或多个基因组特征,由此生成经扩增的一个或多个基因组特征;

(c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及

(d) 通过将所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征的丰度与所述一个或多个受试者的所述健康状况组合来生成特征集,所述特征集被配置成区分癌症类型。

147. 根据权利要求146所述的方法,其中所述一个或多个基因组特征包括微生物系统发育标记基因或其标记基因片段。

148. 根据权利要求147所述的方法,其中所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。

149. 根据权利要求147所述的方法,其中所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。

150. 根据权利要求148所述的方法,其中所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。

151. 根据权利要求149所述的方法,其中所述真菌标记基因可包括以下中的一者或多者:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S以及内部转录间隔区1和2。

152. 根据权利要求147所述的方法,其中微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。

153. 根据权利要求146所述的方法,其中扩增包括进行聚合酶链式反应或其衍生物。

154. 根据权利要求153所述的方法,其中所述其衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。

155. 根据权利要求153所述的方法, 其中所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。

156. 根据权利要求155所述的方法, 其中所述一个或多个基因组特征包括线粒体DNA基因组特征。

157. 根据权利要求146所述的方法, 其包括富集所述一种或多种核酸分子。

158. 根据权利要求157所述的方法, 其中所述一种或多种核酸分子包括哺乳动物核酸分子、非哺乳动物核酸分子或其任何组合。

159. 根据权利要求158所述的方法, 其中富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与杂交探针组合, 其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;

(b) 在促进所述一种或多种非哺乳动物核酸分子的靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物核酸分子以及所述一种或多种非哺乳动物核酸分子;

(c) 使未经结合的杂交探针和与所述一种或多种非哺乳动物核酸分子结合的经杂交探针分离; 以及

(d) 洗涤与所述一种或多种非哺乳动物核酸分子结合的所述经杂交探针, 由此产生一种或多种经富集的非哺乳动物核酸分子。

160. 根据权利要求159所述的方法, 其中洗涤被配置成去除非特异性相关核酸分子和其它反应组分。

161. 根据权利要求158所述的方法, 其中所述一种或多种核酸分子的富集包括非哺乳动物DNA富集。

162. 根据权利要求161所述的方法, 其中所述非哺乳动物DNA富集包括:

(a) 将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;

(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物核酸分子或所述一种或多种非哺乳动物核酸分子的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;

(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸分子片段结合的重组CXXC结构域蛋白;

(d) 洗涤与所述非甲基化CpG核酸分子片段结合的所述重组CXXC结构域蛋白, 由此产生一种或多种经富集的核酸分子以供扩增。

163. 根据权利要求162所述的方法, 其中洗涤被配置成去除非特异性相关核酸分子和所述蛋白质-DNA结合反应组分的剩余部分。

164. 根据权利要求146所述的方法, 其中所述一种或多种核酸源自所述一个或多个受试者的一个或多个生物样品。

165. 根据权利要求164所述的方法, 其中所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。

166. 根据权利要求146所述的方法, 其中所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。

167. 根据权利要求158所述的方法, 其中所述一种或多种哺乳动物核酸分子包括DNA核酸、RNA核酸、游离DNA核酸、游离RNA核酸、外泌体DNA核酸、外泌体RNA核酸或其任何组合。

168. 根据权利要求158所述的方法, 其中所述一种或多种非哺乳动物核酸分子包括微生物DNA、微生物RNA、微生物游离DNA、微生物游离RNA或其任何组合。

169. 根据权利要求146所述的方法, 其包括过滤所述一个或多个非哺乳动物测序读段。

170. 根据权利要求169所述的方法, 其中过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。

171. 根据权利要求170所述的方法, 其中过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。

172. 根据权利要求171所述的方法, 其包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。

173. 根据权利要求172所述的方法, 其中去污包括计算机模拟去污。

174. 根据权利要求172所述的方法, 其中去污被配置成去除非内源性微生物测序读段, 由此生成经去污的微生物分类学分配和相关数量的测序读段。

175. 根据权利要求171所述的方法, 其中映射用QIIME2或其支持的其它版本进行。

176. 根据权利要求171所述的方法, 其中所述一个或多个微生物参考数据库包括细菌16S rRNA数据库Greengenes; 细菌、真菌和古菌rRNA数据库SILVA; 真核核糖体ITS区数据库UNITE; 源自可公开获得且完整的微生物基因组序列的定制数据库; 或其任何组合。

177. 根据权利要求146所述的方法, 其中所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度包括微生物功能性基因丰度、生化通路丰度或其任何组合。

178. 根据权利要求174所述的方法, 其包括预测所述经去污的微生物分类学分配的宏基因组功能性含量, 由此产生一个或多个功能性丰度。

179. 根据权利要求178所述的方法, 其中预测所述宏基因组功能性含量通过PICRUST2进行。

180. 根据权利要求165所述的方法, 其中所述液体活检样品包括: 血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。

181. 根据权利要求146所述的方法, 其中所述癌症类型包括肺癌、乳腺癌、卵巢癌、胃肠道癌、头颈癌、肝癌、胰腺腺癌、前列腺癌、皮肤癌或其任何组合。

182. 根据权利要求181所述的方法, 其中所述肺癌包括非小细胞肺癌。

183. 根据权利要求146所述的方法, 其中所述癌症类型包括I期、II期或III期癌症。

184. 根据权利要求146所述的方法, 其包括生成经训练的预测模型, 其中所述经训练的预测模型是用所述一个或多个受试者的所述特征集和所述健康状况训练的。

185. 根据权利要求184所述的方法, 其中所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。

186. 根据权利要求184所述的方法, 其中所述经训练的预测模型包括正规化机器学习模型。

187. 根据权利要求185所述的方法, 其中所述机器学习模型包括机器学习分类器。

188. 根据权利要求185所述的方法, 其中所述机器学习模型包括梯度提升机机器学习

模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组合。

通过靶向微生物扩增子测序获得的疾病分类器

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年3月10日提交的美国临时申请第63/318,479号的权益,所述美国临时申请出于所有目的通过引用特此并入。

发明内容

[0003] 本公开的各方面提供了一种生成用于区分一个或多个受试者的癌症和非癌症健康状况的特征集的方法。在一些实施例中,所述方法基于一种或多种微生物基因组特征的靶向扩增子测序。在一些实施例中,所述方法包括以下步骤:(a) 提供一个或多个受试者的一种或多种核酸和对应的健康状况;(b) 扩增所述一种或多种核酸的一种或多种非哺乳动物核酸的一个或多个基因组特征,由此生成经扩增的一个或多个基因组特征;(c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及(d) 通过将所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度与所述一个或多个受试者的所述健康状况组合来生成特征集,所述特征集被配置成区分癌症和非癌症健康状况。在一些实施例中,所述基因组特征包括微生物系统发育标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。在一些实施例中,所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。在一些实施例中,所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。在一些实施例中,所述微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。在一些实施例中,扩增包括进行聚合酶链式反应或其衍生物。在一些实施例中,聚合酶链式反应衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。在一些实施例中,所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。在一些实施例中,所述一个或多个基因组特征包括线粒体DNA基因组特征。在一些实施例中,所述阻断引物抑制线粒体DNA基因组特征的扩增。在一些实施例中,所述方法进一步包括富集所述一种或多种核酸。在一些实施例中,所述一种或多种核酸包括哺乳动物核酸、非哺乳动物核酸或其任何组合。在一些实施例中,核酸富集包括以下步骤:(a) 将所述一种或多种哺乳动物和非哺乳动物核酸与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;(b) 在促进靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物和非哺乳动物核酸;(c) 使未经结合的杂交探针和与非哺乳动物核酸结合的经杂交探针分离;以及(d) 洗涤与非哺乳动物核酸结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸。在一些实施例中,洗涤被配置成去除非特异性相关核酸和其它反应组分。在一些实施例中,所述一种或多种核酸的富集包括非哺乳动物DNA富集。在一些实施例中,非哺乳

动物DNA富集包括以下步骤：(a) 将所述一种或多种哺乳动物和非哺乳动物核酸与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应；(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物或非哺乳动物核酸的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应；(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白；以及(d) 洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白，由此产生一种或多种经富集的核酸以供扩增。在一些实施例中，洗涤被配置成去除非特异性相关核酸和蛋白质-DNA结合反应组分的剩余部分。在一些实施例中，所述一种或多种核酸分子源自所述一个或多个受试者的一个或多个生物样品。在一些实施例中，所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。在一些实施例中，所述液体活检样品包括：血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中，所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。在一些实施例中，所述哺乳动物和非哺乳动物核酸包括：DNA核酸、RNA核酸、微生物游离DNA核酸、微生物游离RNA核酸、游离DNA核酸、游离RNA核酸、外泌体DNA核酸、外泌体RNA核酸或其任何组合。在一些实施例中，所述方法包括过滤所述一个或多个非哺乳动物测序读段。在一些实施例中，过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。在一些实施例中，过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。在一些实施例中，所述方法包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。在一些实施例中，去污包括计算机模拟去污。在一些实施例中，去污被配置成去除非内源性微生物测序读段，由此生成经去污的微生物分类学分配和相关数量的测序读段。在一些实施例中，非哺乳动物测序读段映射用QIIME2或其支持的其它版本进行。在一些实施例中，所述一个或多个微生物参考数据库包括细菌16S rRNA数据库Greengenes；细菌、真菌和古菌rRNA数据库SILVA；真核核糖体ITS区数据库UNITE；源自可公开获得且完整的微生物基因组序列的定制数据库；或其任何组合。在一些实施例中，所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度包括微生物功能性基因丰度、生化通路丰度或其任何组合。在一些实施例中，所述方法包括预测所述经去污的微生物分类学分配的宏基因组功能性含量，由此产生一个或多个功能性丰度。在一些实施例中，预测所述宏基因组功能性含量通过PICRUSt2进行。在一些实施例中，所述癌症包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺癌、前列腺癌、皮肤癌或其任何组合。在一些实施例中，肺癌包括非小细胞肺癌。在一些实施例中，所述癌症包括I期、II期或III期癌症。在一些实施例中，所述非癌症状态包括健康非癌症状态、疾病非癌症状态或其任何组合。在一些实施例中，所述疾病状态包括肺病，其中所述肺病包括：类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。在一些实施例中，所述方法包括生成经训练的预测模型，其中所述经训练的预测模型是用所述一个或多个受试者的所述特征集和所述健康状况训练的。在一些实施例中，所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。在一些实施例中，所述经训练的预测模型包括正规化的机器学习模型。在一些实施例中，机器学习模型包括机器学习分类器。在一些实施

例中,所述机器学习模型包括梯度提升机机器学习模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组合。

[0004] 本文所公开的各方面提供了一种使用经训练的预测模型的输出来诊断一个或多个受试者的癌症或非癌症健康状况的方法。在一些实施例中,所述方法包括以下步骤:(a) 提供一个或多个受试者的一种或多种核酸;(b) 扩增一种或多种非哺乳动物核酸的一个或多个基因组特征,由此产生经扩增的一个或多个基因组特征;(c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及(d) 至少作为将所述一个或多个基因组特征作为输入提供给经训练的预测模型的结果,输出对所述一个或多个受试者的癌症或非癌症健康状况的诊断。在一些实施例中,所述非哺乳动物核酸包括微生物核酸。在一些实施例中,所述一种或多种核酸源自所述一个或多个受试者的一种或多种生物样品。在一些实施例中,所述一个或多个生物样品包括:组织活检样品、液体活检样品或其任何组合。在一些情况下,所述液体活检样品包括:血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。在一些实施例中,所述一种或多种核酸包括DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA、游离微生物DNA、游离微生物RNA或其任何组合的总群体。在一些实施例中,所述一个或多个基因组特征包括微生物系统发育标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因可以包括细菌标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。在一些实施例中,所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。在一些实施例中,所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。在一些实施例中,所述微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。在一些实施例中,扩增包括进行聚合酶链式反应或其衍生物。在一些实施例中,聚合酶链式反应衍生物包括:反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。在一些实施例中,所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。在一些实施例中,所述一个或多个基因组特征包括线粒体DNA基因组特征。在一些实施例中,所述阻断引物抑制线粒体DNA基因组特征的扩增。在一些实施例中,所述方法包括富集所述一种或多种核酸。在一些实施例中,所述一种或多种核酸包括哺乳动物核酸、非哺乳动物核酸或其任何组合。在一些实施例中,所述哺乳动物和非哺乳动物核酸包括DNA核酸、RNA核酸、微生物游离DNA核酸、微生物游离RNA核酸、游离DNA核酸、游离RNA核酸、外泌体DNA核酸、外泌体RNA核酸或其任何组合。在一些实施例中,核酸富集包括以下步骤:(a) 将所述一种或多种哺乳动物和非哺乳动物核酸与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;(b) 在促进靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物和非哺乳动物核酸;(c) 使未经结合的杂交探针和与非哺乳动物核酸结合的经杂交探针分离;以及(d) 洗涤与非哺乳动物核酸结合的所述经杂交探针,由此产

生一种或多种经富集的非哺乳动物核酸。在一些实施例中,洗涤被配置成去除非特异性相关核酸和其它反应组分。在一些实施例中,所述一种或多种核酸的富集包括非哺乳动物DNA富集。在一些实施例中,非哺乳动物DNA富集包括以下步骤:(a)将所述一种或多种哺乳动物和非哺乳动物核酸与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;(b)在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物或非哺乳动物核酸的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;(c)从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白;以及(d)洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸以供扩增。在一些实施例中,洗涤被配置成去除非特异性相关核酸分子和所述蛋白质-DNA结合反应组分的剩余部分。在一些实施例中,所述重组CXXC结构域蛋白包括:重组的含有锌指CXXC结构域的蛋白KDM2A、KDM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、源自其的重组CXXC结构域或其任何组合。在一些实施例中,所述方法包括过滤所述一种或多种非哺乳动物测序。在一些实施例中,过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。在一些实施例中,过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。在一些实施例中,所述方法包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。在一些实施例中,所述去污包括计算机模拟去污。在一些实施例中,去污被配置成去除非内源性微生物测序读段,由此生成经去污的微生物分类学分配和相关数量的测序读段。在一些实施例中,非哺乳动物测序读段映射用QIIME2或其支持的其它版本进行。在一些实施例中,所述一个或多个微生物参考数据库包括:细菌16S rRNA数据库Greengenes;细菌、真菌和古菌rRNA数据库SILVA;真核核糖体ITS区数据库UNITE;源自可公开获得且完整的微生物基因组序列的定制数据库;或其任何组合。在一些实施例中,所述一个或多个基因组特征包括所述一个或多个非哺乳动物测序读段的微生物功能性基因、生化通路的丰度或其任何组合丰度。在一些实施例中,所述方法包括预测所述经去污的微生物分类学分配的所述宏基因组功能性含量,由此产生一个或多个功能性丰度。在一些实施例中,宏基因组功能性含量通过PICRUSt2进行。在一些实施例中,所述癌症健康状况包括:肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺癌、前列腺癌、皮肤癌或其任何组合。在一些实施例中,肺癌包括非小细胞肺癌。在一些实施例中,所述癌症包括I期、II期或III期癌症。在一些实施例中,所述非癌症状态包括健康非癌症状态、疾病非癌症状态或其任何组合。在一些实施例中,所述疾病状态包括肺病,其中所述肺病包括:类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。在一些实施例中,所述经训练的预测模型是用一个或多个受试者的特征集和健康状况训练的。在一些实施例中,所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。在一些实施例中,所述经训练的预测模型包括正规化的机器学习模型。在一些实施例中,机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括梯度提升机机器学习模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组

合。

[0005] 本文公开的各方面提供了一种用于诊断一个或多个受试者的癌或非癌健康状况的系统。在一些实施例中,所述系统包括:(a)处理器;以及(b)包含软件的非暂时性计算机可读存储介质,所述软件被配置成使所述处理器:(i)接收一个或多个受试者的生物样品的所述一个或多个受试者的一个或多个核酸测序读段,其中所述一个或多个核酸测序读段包括一种或多种非哺乳动物核酸的经扩增的一个或多个基因组特征;以及(ii)至少作为将所述一个或多个非哺乳动物核酸测序读段的一个或多个基因组特征作为输入提供给经训练的预测模型的结果,输出对所述一个或多个受试者的癌或非癌健康状况的诊断。在一些实施例中,所述非哺乳动物核酸可以包括微生物核酸。在一些实施例中,所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。在一些实施例中,所述一个或多个受试者可以包括人受试者、非人哺乳动物受试者或其任何组合。在一些实施例中,所述液体活检样品包括:血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述一种或多种核酸包括:DNA、RNA、游离DNA、游离RNA、外泌体DNA、外泌体RNA、游离微生物DNA、游离微生物RNA或其任何组合。在一些实施例中,所述基因组特征可以包括微生物系统发育标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因可以包括细菌标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因可以包括真菌标记基因或其标记基因片段。在一些实施例中,所述细菌标记基因包括核糖体RNA基因。在一些实施例中,所述核糖体RNA基因包括5S核糖体RNA基因、16S核糖体RNA基因、23S核糖体RNA基因或其任何组合。在一些实施例中,所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。在一些实施例中,所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。在一些实施例中,所述微生物系统发育标记基因可以包括细菌标记基因、真菌标记基因或其任何组合。在一些实施例中,所述一种或多种非哺乳动物核酸的所述经扩增的一个或多个基因组特征是通过聚合酶链式反应或其衍生物扩增的。在一些实施例中,聚合酶链式反应衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。在一些实施例中,所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。在一些实施例中,所述一个或多个基因组特征包括线粒体DNA基因组特征。在一些实施例中,所述阻断引物抑制线粒体DNA基因组特征的扩增。在一些实施例中,所述一个或多个核酸测序读段包括一种或多种经富集的核酸的测序读段。在一些实施例中,所述一种或多种核酸可以包括哺乳动物核酸、非哺乳动物核酸或其任何组合。在一些实施例中,所述一种或多种经富集的核酸分子是通过以下产生的:(a)将所述一种或多种哺乳动物和非哺乳动物核酸与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补的核酸序列;(b)在促进靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物和非哺乳动物核酸;(c)使未经结合的杂交探针和与非哺乳动物核酸结合的经杂交探针分离;以及(d)洗涤与非哺乳动物核酸结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸。在一些实施例中,洗涤被配置成去除非特异性相关核酸和其它反应组

分。在一些实施例中,所述一种或多种经富集的核酸分子是通过非哺乳动物DNA富集产生的。在一些实施例中,所述非哺乳动物富集包括:(a)将所述一种或多种哺乳动物和非哺乳动物核酸与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;(b)在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物或非哺乳动物核酸的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;(c)从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白;以及(d)洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸以供扩增。在一些实施例中,洗涤被配置成去除非特异性相关核酸和蛋白质-DNA结合反应组分的剩余部分。在一些实施例中,所述重组CXXC结构域蛋白包括:重组的含有锌指CXXC结构域的蛋白KDM2A、KDM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、源自其的重组CXXC结构域或其任何组合。在一些实施例中,所述软件配置所述处理器以过滤所述一个或多个核酸测序读段。在一些实施例中,过滤包括过滤所述一个或多个测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。在一些实施例中,过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。在一些实施例中,所述软件配置所述处理器以对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。在一些实施例中,所述去污包括计算机模拟去污。在一些实施例中,去污被配置成去除非内源性微生物测序读段,由此生成经去污的微生物分类学分配和相关数量的测序读段。在一些实施例中,映射用QIIME2或其支持的其它版本进行。在一些实施例中,所述一个或多个微生物参考数据库包括:细菌16S rRNA数据库Greengenes;细菌、真菌和古菌rRNA数据库SILVA;真核核糖体ITS区数据库UNITE;源自可公开获得且完整的微生物基因组序列的定制数据库;或其任何组合。在一些实施例中,所述经扩增的一个或多个基因组特征包括一个或多个非哺乳动物测序读段的微生物功能性基因、生化通路的丰度或其任何组合丰度。在一些实施例中,预测宏基因组功能性含量在经去污的微生物分类学分配上进行,由此产生一个或多个功能性丰度。在一些实施例中,所述软件配置所述处理器以预测所述经去污的微生物分类学分配的宏基因组功能性含量,由此产生一个或多个功能性丰度。在一些实施例中,预测所述宏基因组功能性含量通过PICRUST2进行。在一些实施例中,所述癌健康状况包括:肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺癌、前列腺癌、皮肤癌或其任何组合。在一些实施例中,肺癌包括非小细胞肺癌。在一些实施例中,所述癌健康状况包括I期、II期或III期癌症。在一些实施例中,所述非癌健康状况包括健康、疾病或其任何组合非癌状况。在一些实施例中,所述疾病状态可以包括肺病,其中肺病包括:类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。在一些实施例中,所述经训练的预测模型是用所述一个或多个受试者的一个或多个基因组特征集和所述健康状况训练的。在一些实施例中,所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。在一些实施例中,所述经训练的预测模型包括正规化的机器学习模型。在一些实施例中,机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括梯度提升机机器学习模型、神经网络机器学习模型、支持向量机机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习

模型、回归机器学习模型或其任何组合。在一些实施例中,所述癌健康状况包括一种或多种类型的癌症、一种或多种癌症亚型、癌症分期、癌症预后或其任何组合。在一些实施例中,所述癌或非癌健康状况包括癌症或疾病的类别、组织特异性位置或其任何组合。在一些实施例中,所述经训练的预测模型用于预测所述一个或多个受试者的癌症疗法应答。在一些实施例中,所述经训练的预测模型用于选择针对所述一个或多个受试者的最佳疗法。在一些实施例中,所述经训练的预测模型用于对一个或多个受试者的一种或多种癌症对疗法的应答过程进行纵向建模并且然后调整治疗方案。在一些实施例中,所述癌健康状况可以包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、肺腺癌、肺鳞状细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合癌症。在一些实施例中,所述经训练的预测模型可以被配置成去除受污染的非哺乳动物特征,同时选择性地保留其它未受污染的非哺乳动物特征。

[0006] 本文所公开的各方面提供了一种生成用于区分一个或多个受试者的癌症类型的特征集的方法,所述方法包括:(a) 提供一个或多个受试者的一种或多种核酸和对应的健康状况;(b) 扩增所述一种或多种核酸的一种或多种非哺乳动物核酸的一个或多个基因组特征,由此生成经扩增的一个或多个基因组特征;(c) 对所述经扩增的一个或多个基因组特征进行测序以生成一个或多个非哺乳动物测序读段;以及(d) 通过将所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度与所述一个或多个受试者的所述健康状况组合来生成特征集,所述特征集被配置成区分癌症类型。在一些实施例中,所述基因组特征包括微生物系统发育标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因包括细菌标记基因或其标记基因片段。在一些实施例中,所述微生物系统发育标记基因包括真菌标记基因或其标记基因片段。在一些实施例中,所述细菌标记基因包括:核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。在一些实施例中,所述真菌标记基因包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。在一些实施例中,所述微生物系统发育标记基因包括细菌标记基因、真菌标记基因或其任何组合。在一些实施例中,扩增包括进行聚合酶链式反应或其衍生物。在一些实施例中,聚合酶链式反应衍生物包括反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。在一些实施例中,所述聚合酶链式反应包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。在一些实施例中,所述一个或多个基因组特征包括线粒体DNA基因组特征。在一些实施例中,所述阻断引物抑制线粒体DNA基因组特征的扩增。在一些实施例中,所述方法进一步包括富集所述一种或多种核酸。在一些实施例中,所述一种或多种核酸包括哺乳动物核酸、非哺乳动物核酸或其任何组合。在一些实施例中,核酸富集包括以下步骤:(a) 将所述一种或多种哺乳动物和非哺乳动物核酸与杂交探针组合,其中所述杂交探针包括与非哺乳动物基因组特征互补

的核酸序列；(b) 在促进靶核酸特征与所述杂交探针之间的核酸碱基配对的条件下温育所述杂交探针和所述一种或多种哺乳动物和非哺乳动物核酸；(c) 使未经结合的杂交探针和与非哺乳动物核酸结合的经杂交探针分离；以及(d) 洗涤与非哺乳动物核酸结合的所述经杂交探针，由此产生一种或多种经富集的非哺乳动物核酸。在一些实施例中，洗涤被配置成去除非特异性相关核酸和其它反应组分。在一些实施例中，所述一种或多种核酸的富集包括非哺乳动物DNA富集。在一些实施例中，非哺乳动物DNA富集包括以下步骤：(a) 将所述一种或多种哺乳动物和非哺乳动物核酸与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应；(b) 在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物或非哺乳动物核酸的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应；(c) 从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白；以及(d) 洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白，由此产生一种或多种经富集的核酸以供扩增。在一些实施例中，洗涤被配置成去除非特异性相关核酸和蛋白质-DNA结合反应组分的剩余部分。在一些实施例中，所述一种或多种核酸分子源自所述一个或多个受试者的一个或多个生物样品。在一些实施例中，所述一个或多个生物样品包括组织活检样品、液体活检样品或其任何组合。在一些实施例中，所述液体活检样品包括：血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中，所述一个或多个受试者包括人受试者、非人哺乳动物受试者或其任何组合。在一些实施例中，所述哺乳动物和非哺乳动物核酸包括：DNA核酸、RNA核酸、微生物游离DNA核酸、微生物游离RNA核酸、游离DNA核酸、游离RNA核酸、外泌体DNA核酸、外泌体RNA核酸或其任何组合。在一些实施例中，所述方法包括过滤所述一个或多个非哺乳动物测序读段。在一些实施例中，过滤包括过滤所述一个或多个非哺乳动物测序读段以产生一个或多个线粒体DNA耗竭的非哺乳动物测序读段。在一些实施例中，过滤包括将所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段相对于一个或多个微生物参考数据库进行映射以确定所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。在一些实施例中，所述方法包括对所述一个或多个线粒体DNA耗竭的非哺乳动物测序读段进行去污。在一些实施例中，去污包括计算机模拟去污。在一些实施例中，去污被配置成去除非内源性微生物测序读段，由此生成经去污的微生物分类学分配和相关数量的测序读段。在一些实施例中，非哺乳动物测序读段映射用QIIME2或其支持的其它版本进行。在一些实施例中，所述一个或多个微生物参考数据库包括细菌16S rRNA数据库Greengenes；细菌、真菌和古菌rRNA数据库SILVA；真核核糖体ITS区数据库UNITE；源自可公开获得且完整的微生物基因组序列的定制数据库；或其任何组合。在一些实施例中，所述一个或多个非哺乳动物测序读段的所述一个或多个基因组特征丰度包括微生物功能性基因丰度、生化通路丰度或其任何组合。在一些实施例中，所述方法包括预测所述经去污的微生物分类学分配的宏基因组功能性含量，由此产生一个或多个功能性丰度。在一些实施例中，预测所述宏基因组功能性含量通过PICRUSt2进行。在一些实施例中，所述癌症包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺癌、前列腺癌、皮肤癌或其任何组合。在一些实施例中，肺癌包括非小细胞肺癌。在一些实施例中，所述癌症包括I期、II期或III期癌症。在一些实施例中，所述方法包括生成经训练的预测模型，其中所述经训练的预测模型是用所述一个或多个受试者的所述特征集和所述健康状况训练的。在一些实施例

中,所述经训练的预测模型包括机器学习模型、一种或多种机器学习模型、机器学习模型的集合或其任何组合。在一些实施例中,所述经训练的预测模型包括正规化的机器学习模型。在一些实施例中,机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括梯度提升机器学习模型、神经网络机器学习模型、支持向量机器学习模型、k均值机器学习模型、分类树机器学习模型、随机森林机器学习模型、回归机器学习模型或其任何组合。

[0007] 本文所提供的公开内容的各方面描述了一种确定受试者的疾病的方法,所述方法包括:接收受试者的生物样品、电子病历信息和一个或多个放射性图像;对从所述生物样品中分离出的一种或多种核酸分子进行测序,由此生成一个或多个核酸分子测序读段;以及在预测模型提供所述受试者的一个或多个核酸分子测序读段、电子病历信息和源自一个或多个放射性图像的数据作为输入时,确定所述受试者的疾病作为预测模型的输出。在一些实施例中,所述方法进一步包括鉴定来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述预测模型提供来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFRII、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII或其组合。在一些实施例中,所述疾病包括癌症或非癌疾病。在一些实施例中,所述生物样品包括液体活检、组织活检或其组合。在一些实施例中,所述一个或多个放射性图像包括x光图像、计算机断层扫描(CT)图像、低剂量计算机断层扫描图像、磁共振成像(MRI)图像、超声波图像、正电子发射断层扫描图像、透视图像、血管造影图像或其任何组合。在一些实施例中,所述癌症包括直径小于3厘米的肿瘤块。在一些实施例中,测序包括基于扩增子的16S rRNA测序。在一些实施例中,所述基于扩增子的16S rRNA测序对所述一种或多种核酸分子的V6区进行测序。在一些实施例中,所述一种或多种核酸分子包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合。在一些实施例中,所述液体活检包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述方法进一步包括计算所述一个或多个放射性图像的一个或多个特征,其中所述一个或多个放射性图像的一个或多个特征作为输入向所述预测模型提供。在一些实施例中,所述一个或多个特征包括布洛克癌症概率评分、病变直径、病变棘突、病变坚实度或其任何组合。在一些实施例中,所述方法进一步包括将所述一个或多个核酸测序读段相对于

基因组数据库进行映射或比对以确定所述一个或多个核酸测序读段的一个或多个特征。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期、III期或IV期。在一些实施例中,所述方法进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,测序包括鸟枪法宏基因组测序、下一代测序、长读段测序或其任何组合。在一些实施例中,所述方法进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、Bowtie2、Kraken或其任何组合完成。

[0008] 本文所提供的公开内容的另一方面描述了一种方法,其包括:接收一个或多个受试者的生物样品、电子病历信息、源自一个或多个放射性图像的数据以及对应疾病;对从所述生物样品中分离出的一种或多种核酸分子进行测序,由此生成一个或多个核酸分子测序读段;以及鉴定对应于所述一个或多个受试者的所述疾病的所述一个或多个核酸分子测序读段的一个或多个特征、电子病历信息以及源自一个或多个放射性图像的数据。在一些实施例中,鉴定包括将所述一个或多个测序读段与基因组数据库进行比对。在一些实施例中,所述方法进一步包括用所述核酸分子测序读段的所述一个或多个特征、电子病历信息和源自所述一个或多个放射性图像的所述数据以及所述一个或多个受试者的对应疾病来训练预测模型。在一些实施例中,所述疾病包括癌症或非癌疾病。在一些实施例中,所述方法进一步包括鉴定所述受试者的所述生物样品的一种或多种蛋白质生物标记物的一种或多种特征。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFR1I、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFR1I或其组合。在一些实施例中,所述生物样品包括液体活检、组织活检或其组合。在一些实施例中,所述一个或多个放射性图像包括x光图像、计算机断层扫描(CT)图像、低剂量计算机断层扫描图像、磁共振成像(MRI)图像、超声波图像、正电子发射断层扫描图像、透视图像、血管造影图像或其任何组合。在一些实施例中,所述癌症包括直径小于3厘米的肿瘤块。在一些实施例中,测序包括基于扩增子的16S rRNA测序。在一些实施例中,所述基于扩增子的16S rRNA测序对所

述一种或多种核酸分子的V6区进行测序。在一些实施例中,所述一种或多种核酸分子包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合。在一些实施例中,所述液体活检包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述一个或多个放射性图像特征包括布洛克癌症概率评分、病变直径、病变棘突、病变坚实度或其任何组合。在一些实施例中,所述方法进一步包括将所述一个或多个核酸测序读段相对于基因组数据库进行映射或比对以确定所述一个或多个核酸测序读段的一个或多个人、非人或其组合特征。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期或III期或IV期。在一些实施例中,所述方法进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,测序包括鸟枪法测序、下一代测序、长读段测序或其任何组合。在一些实施例中,所述方法进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、Bowtie2、Kraken或其任何组合完成。

[0009] 本文所提供的公开内容的另一方面描述了一种被配置成确定受试者的疾病的计算机系统,所述计算机系统包括:(a) 一个或多个处理器;以及(b) 包含软件的非暂时性计算机可读存储介质,其中所述软件包括可执行指令,所述可执行指令作为执行的结果使所述计算机系统的所述一个或多个处理器:(i) 接收受试者的生物样品的一个或多个测序读段、电子病历信息和一个或多个图像;以及(ii) 在预测模型提供所述受试者的一个或多个核酸

分子测序读段、电子病历信息和源自一个或多个放射性图像的数据作为输入时,确定所述受试者的疾病作为预测模型的输出。在一些实施例中,所述疾病包括癌症或非癌疾病。在一些实施例中,所述生物样品包括组织活检、液体活检或其组合。在一些实施例中,所述可执行指令包括接收来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述预测模型提供来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白或其组合。在一些实施例中,所述预测模型是用所述核酸分子测序读段的所述一个或多个特征、电子病历信息和源自所述一个或多个放射性图像的所述数据以及所述一个或多个受试者的对应疾病进行训练的。在一些实施例中,所述可执行指令包括鉴定所述受试者的所述生物样品的一种或多种蛋白质生物标记物的一个或多个特征。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFR1I、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFR1I或其组合。在一些实施例中,所述一个或多个放射性图像包括x光图像、计算机断层扫描(CT)图像、低剂量计算机断层扫描图像、磁共振成像(MRI)图像、超声波图像、正电子发射断层扫描图像、透视图像、血管造影图像或其任何组合。在一些实施例中,所述癌症包括直径小于3厘米的肿瘤块。在一些实施例中,所述一个或多个核酸分子测序读段包括一个或多个基于扩增子的16S rRNA测序读段。在一些实施例中,所述基于扩增子的16S rRNA测序读段包括所述一种或多种核酸分子的V6区的测序读段。在一些实施例中,所述一种或多种核酸分子测序读段包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合的测序读段。在一些实施例中,所述液体活检包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述一个或多个放射性图像特征包括布洛克癌症概率评分、病变直径、病变棘突、病变坚实度或其任何组合。在一些实施例中,所述可执行指令进一步包括将所述一个或多个核酸测序读段相对于基因组数据库进行映射或比对以确定所述一个或多个核酸测序读段的一个或多个个人、非人或其组合特征。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预

测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期、III期或IV期。在一些实施例中,所述可执行指令进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,所述一个或多个测序读段是通过鸟枪法测序、下一代测序、长读段测序或其任何组合产生的。在一些实施例中,所述可执行指令进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、Bowtie2、Kraken或其任何组合完成。

[0010] 本文所提供的公开内容的另一方面描述了一种确定受试者疾病的方法,所述方法包括:接收来自受试者的生物样品;对所述生物样品的一种或多种核酸分子进行测序,由此生成一个或多个核酸分子测序读段;以及当预测模型提供所述受试者的一个或多个核酸分子测序读段时,确定所述受试者的疾病作为预测模型的输出,其中所述预测模型是用一个或多个液体生物样品和一个或多个组织生物样品的一个或多个核酸分子测序读段以及一个或多个受试者的对应疾病来训练的。在一些实施例中,所述疾病包括癌症、非癌疾病或其组合。在一些实施例中,所述方法进一步包括鉴定来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述预测模型提供来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFR1I、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFR1I或其组合。在一些实施例中,所述癌症包括直径小于3厘米毫米的肿瘤块。在一些实施例中,所述测序包括基于扩增子的16S rRNA测序。在一些实施例中,所述基于扩增子的16SrRNA测序对所述一种或多种核酸分子的V6区进行测序。在一些实施例中,所述一种或多种核酸分子包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合。在一些实施例中,所述液体活检包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的

皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述方法进一步包括将所述一个或多个核酸测序读段相对于基因组数据库进行映射或比对以确定所述一个或多个核酸测序读段的一个或多个人、非人或其组合特征,所述特征作为输入向所述预测模型提供。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期、III期或IV期。在一些实施例中,所述方法进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段,其中所述一个或多个经去污的核酸分子作为输入向所述预测模型提供。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,测序包括鸟枪法测序、下一代测序、长读段测序或其任何组合。在一些实施例中,所述方法进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、PICRUSt2、Bowtie2、Kraken或其任何组合完成。

[0011] 本文所提供的公开内容的另一方面描述了一种鉴定一个或多个非人类基因组特征的方法,所述方法包括:接收一个或多个液体生物样品、一个或多个组织生物样品和一个或多个受试者的对应疾病;对所述一个或多个液体生物样品和所述一个或多个组织生物样品的一个或多个核酸分子进行测序,由此生成一个或多个测序读段;以及鉴定来自一个或多个测序读段的对应于所述一个或多个受试者的所述疾病的一个或多个非人基因组特征。在一些实施例中,鉴定包括将所述一个或多个测序读段相对于基因组数据库进行比对或映射以确定所述一个或多个核酸测序读段的一个或多个人、非人或其组合特征。在一些实施例中,所述方法进一步包括用所述一个或多个非人基因组特征和所述一个或多个受试者的所述对应疾病来训练预测模型。在一些实施例中,所述疾病包括癌症或非癌疾病。在一些实施例中,所述方法进一步包括鉴定所述一种或多种液体生物样品、所述一种或多种组织生物样品或其组合的一种或多种蛋白质生物标记物的一种或多种特征。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFRII、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII或其组合。在一些实施例中,所述癌症包括直径小于3厘米的肿瘤块。在一些实施例中,所述测序包括基于扩增子的16S rRNA测序。在一些实施例

中,所述基于扩增子的16S rRNA测序对所述一种或多种核酸分子的V6区进行测序。在一些实施例中,所述一种或多种核酸分子包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合。在一些实施例中,所述液体生物样品包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样赘生物弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期、III期或IV期。在一些实施例中,所述方法进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,测序包括鸟枪法测序、下一代测序、长读段测序或其任何组合。在一些实施例中,所述方法进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、PICRUSt2、Bowtie2、Kraken或其任何组合完成。

[0012] 本文所提供的公开内容的另一方面描述了一种被配置成确定受试者的疾病的计算机系统,所述计算机系统包括:(a) 一个或多个处理器;以及(b) 包含软件的非暂时性计算机可读存储介质,其中所述软件包括可执行指令,所述可执行指令作为执行的结果使所述计算机系统的所述一个或多个处理器:(i) 接收受试者的生物样品的一个或多个测序读段;以及(ii) 当预测模型提供所述受试者的一个或多个核酸分子测序读段时,确定所述受试者的疾病作为预测模型的输出,其中所述预测模型是用一个或多个液体生物样品和一个或多个组织生物样品的一个或多个核酸分子测序读段以及一个或多个受试者的对应疾病来训练的。在一些实施例中,所述疾病包括癌症或非癌疾病。在一些实施例中,所述疾病包括癌

症或非癌疾病。在一些实施例中,所述可执行指令包括接收来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述预测模型提供来自所述受试者的所述生物样品的一种或多种蛋白质生物标记物。在一些实施例中,所述可执行指令包括鉴定所述受试者的所述生物样品的一种或多种蛋白质生物标记物的一个或多个特征。在一些实施例中,所述一种或多种蛋白质生物标记物包括癌胚抗原、骨桥蛋白、癌症抗原15-3、癌症抗原19-9、癌症抗原125、白细胞介素-8、催乳激素、细胞角蛋白19片段(CYFRA 21-1)、MMP-9、sTNFR1I、MMP-7、抵抗素、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFR1I或其组合。在一些实施例中,所述癌症包括直径小于3厘米的肿瘤块。在一些实施例中,所述一个或多个核酸分子测序读段包括一个或多个基于扩增子的16S rRNA测序读段。在一些实施例中,所述基于扩增子的16S rRNA测序读段包括所述一种或多种核酸分子的V6区的测序读段。在一些实施例中,所述一种或多种核酸分子测序读段包括哺乳动物RNA、哺乳动物DNA、哺乳动物游离DNA、哺乳动物游离RNA、哺乳动物外泌体DNA、哺乳动物外泌体RNA、非人RNA、非人DNA、非人游离DNA、非人游离RNA、非人外泌体DNA、非人外泌体RNA、循环肿瘤DNA、循环肿瘤RNA或其任何组合的测序读段。在一些实施例中,所述液体生物样品包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些实施例中,所述癌症包括肺腺癌(LUAD)、肺鳞状细胞癌(LUSC)、小细胞肺癌(SCLC)或其任何组合。在一些实施例中,所述癌症包括:急性骨髓性白血病、肾上腺皮质癌、膀胱尿路上皮癌、脑低级神经胶质瘤、乳腺浸润性癌、宫颈鳞状细胞癌和宫颈腺癌、胆管癌、结肠腺癌、食管癌、多形性胶质母细胞瘤、头颈部鳞状细胞癌、肾脏嫌色细胞、肾脏肾透明细胞癌、肾脏肾乳头状细胞癌、肝细胞癌、淋巴样增生性弥漫性大B细胞淋巴瘤、间皮瘤、卵巢浆液性囊腺癌、胰腺腺癌、嗜铬细胞瘤和副神经节瘤、前列腺腺癌、直肠腺癌、肉瘤、皮肤的皮肤黑色素瘤、胃腺癌、睾丸生殖细胞肿瘤、胸腺瘤、甲状腺癌、子宫癌肉瘤、子宫体子宫内膜癌、葡萄膜黑色素瘤或其任何组合。在一些实施例中,所述可执行指令进一步包括将所述一个或多个核酸测序读段相对于基因组数据库进行映射或比对以确定所述一个或多个核酸测序读段的一个或多个个人、非人或其组合特征。在一些实施例中,所述基因组数据库包括人基因组数据库。在一些实施例中,所述预测模型包括机器学习模型。在一些实施例中,所述预测模型包括神经网络、卷积神经网络、逻辑回归、随机森林、支持向量机或其任何组合。在一些实施例中,所述机器学习模型包括机器学习分类器。在一些实施例中,所述机器学习模型包括堆叠机器学习模型、一种或多种机器学习模型、集合机器学习模型或其组合。在一些实施例中,所述预测模型使用留一法验证进行训练。在一些实施例中,所述预测模型被配置成确定癌症的阶段、癌症的解剖起源或其组合。在一些实施例中,所述癌症的阶段是I期、II期、III期或IV期。在一些实施例中,所述可执行指令进一步包括对所述一个或多个核酸分子测序读段进行去污以产生一个或多个经去污的核酸分子测序读段。在一些实施例中,去污包括计算机模拟去污、实验对照去污或其组合。在一些实施例中,所述预测模型以至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约91%、至少约92%、至少约93%、至少约94%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的准确性确定所述疾病。在一些实施例中,所述一个或多个测序读段是通过鸟枪法测序、下一代测序、长读段测序或其任何组合产生的。在一些实施例中,所述可执行指令进一步包括确定所述一个或多个核酸分子测序读段的一个或多个特

征。在一些实施例中,所述一种或多种核酸分子的一个或多个特征包括非微生物分类学丰度、哺乳动物基因组坐标、注释的基因组位点、哺乳动物功能性基因和/或生化通路丰度或其任何组合特征,以及与所述一个或多个特征相关的多个测序读段。在一些实施例中,所述预测模型被配置成区分所述受试者的癌症和非癌疾病。在一些实施例中,映射或比对用Deblur、PICRUSt2、Bowtie2、Kraken或其任何组合完成。

[0013] 通过引用并入

[0014] 本说明书中所提到的所有公开、专利和专利申请均通过相同的程度引用并入本文,如同特定且单独地指示每个单独的公开、专利或专利申请是通过引用并入的。

附图说明

[0015] 本发明的新颖特征在所附权利要求书中具体阐述。通过参考阐述了说明性实施例的以下具体实施方式,将获得对本发明的特征和优点的更好理解,在所述实施例中利用了本发明的原理,并且在其附图中:

[0016] 图1显示了如本文中的一些实施例中所述的微生物核酸扩增和/或富集方法的流程图。

[0017] 图2显示了如本文中的一些实施例中所述的微生物分类学计算方法的流程图。

[0018] 图3显示了如本文中的一些实施例中所述的微生物功能注释计算方法的流程图。

[0019] 图4显示了用于从健康、癌症和/或非癌非健康受试者的核酸样品中生成一个或多个基于微生物分类学的预测模型分类器的方法的流程图。

[0020] 图5显示了用于从健康、癌症和/或非癌非健康受试者的核酸样品中生成一个或多个微生物功能性注释预测模型分类器的方法的流程图。

[0021] 图6显示了如本文中的一些实施例中所述的被配置成执行本文所提供的公开内容的方法的系统。

[0022] 图7A-7B显示了如本文中的一些实施例中所述的用于扩增系统发育上不同的细菌的16S区的16S核糖体RNA高变区和对应的16S引物。

[0023] 图8显示了如本文中的一些实施例中所述的具有内部转录(ITS)区的真菌核糖体RNA基因簇的示意图。

[0024] 图9显示了如本文中的一些实施例中所述的用V6引物对和用所述V6引物对扩增的微生物DNA标准品组合物进行的16S核糖体DNA扩增的实验数据。

[0025] 图10显示了在存在和不存在人基因组DNA的情况下用V6引物对进行的微生物16S核糖体DNA扩增的实验数据。

[0026] 图11显示了如本文中的一些实施例中所述的用V6引物对进行的16S核糖体DNA扩增的特异性的实验数据。

[0027] 图12A-B显示了如本文中的一些实施例中所述的用于16S测序文库制备(图12A)和西部验证(图12B)的流程图。

[0028] 图13显示了如本文中的一些实施例中所述的用于16S测序处理的流程图。

[0029] 图14A-14B显示了如本文中的一些实施例中所述的在16S测序处理过程中的各点处的实验数据测序读段计数。

[0030] 图15A-15B显示了如本文其它部分所述的经训练的预测模型用于区分来自一个或

多个受试者的非小细胞肺癌和非癌症核酸样品的接收者操作特征曲线。

具体实施方式

[0031] 本文所提供的公开内容描述了确定、鉴定、分类和/或生成一个或多个受试者的一个或多个核酸分子特征的方法和系统,所述方法和系统可以区分、分类和/或诊断一个或多个受试者和/或一组或多组受试者的健康状况。在一些情况下,一个或多个核酸分子特征可以从受试者和/或多个受试者的一个或多个生物样品的一种或多种核酸分子中衍生、获得、接收和/或确定。在一些情况下,一种或多种核酸分子可以包括一种或多种哺乳动物核酸分子、一种或多种非哺乳动物核酸分子或其组合。在一些情况下,一种或多种非哺乳动物核酸分子可以包括一种或多种来自细菌、真菌或其组合的核酸分子。在一些情况下,如本文其它部分所述,一个或多个受试者的健康状况可以包括癌健康状况、非癌疾病健康状况、健康的健康状况或其组合。在一些情况下,癌健康状况可以包括患有癌症的个体。在一些情况下,所述癌症可以包括肺癌、乳腺癌、卵巢癌、胃肠癌、头颈癌、肝癌、胰腺癌、前列腺癌、皮肤癌或其任何组合。在一些情况下,肺癌可以包括非小细胞肺癌。在一些情况下,癌健康状况可以包括癌症的阶段(例如,I期、II期、III期等)的诊断。在一些情况下,健康状况可以包括癌症和/或疾病在受试者或多个受试者的体内的空间位置(即解剖位置)。在一些情况下,生物样品可以包括液体生物样品、组织生物样品或其组合。在一些情况下,非癌疾病健康状况可以包括肺病。在一些情况下,肺病可以包括:类癌、错构瘤、肉芽肿、间质纤维化、肺气肿、支气管炎、慢性阻塞性肺部疾病、肺炎、结节病或其任何组合。在一些情况下,液体生物样品可以包括液体活检。在一些情况下,液体活检可以包括血浆、血清、全血、尿液、脑脊液、唾液、汗液、泪液、呼出气冷凝液或其任何组合。在一些情况下,组织生物样品可以包括受试者的一个或多个区、器官和/或解剖位置(例如,肺、皮肤、肝脏、胰腺、大脑等)的组织活检。

[0032] 在一些情况下,在对应的一个或多个经富集、经过滤和/或经去污的核酸分子特征以至少约80%、至少约85%、至少约86%、至少约87%、至少约88%、至少约89%、至少约90%、至少约92%、至少约94%、至少约96%、至少约98%或至少约99%的准确率确定、分类、鉴定和/或诊断一个或多个受试者的健康状况时,将一种或多种核酸分子和/或一种或多种核酸分子的一个或多个测序读段扩增、富集、过滤和/或去污可以提供比预期更好的结果。

[0033] 在一些情况下,一个或多个受试者的生物样品可以包括一种或多种微生物核酸分子组合物、一种或多种哺乳动物核酸分子组合物或其组合201,如图1中所示。在一些情况下,一个或多个微生物核酸分子、一种或多种哺乳动物核酸分子组合物或其组合可以通过微生物核酸富集和扩增工作流202来富集。在一些情况下,包括一个或多个微生物核酸分子和一个或多个哺乳动物核酸分子的生物样品可以通过杂交探针富集203和/或通过基于蛋白质的微生物DNA富集204来富集。在一些情况下,基于杂交的富集可以包括:将一种或多种哺乳动物核酸分子和一种或多种非哺乳动物核酸分子与杂交探针组合,其中杂交探针可以包括与非哺乳动物基因组特征互补的核酸序列;在促进所述一种或多种非哺乳动物核酸分子的靶核酸特征与所述杂交探针之间的核酸分子碱基配对的条件下温育所述杂交探针、所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子;使未经结合的杂交探针和与所述一种或多种非哺乳动物核酸分子结合的经杂交探针分离;以及洗涤与所述

一种或多种非哺乳动物核酸分子结合的所述经杂交探针,由此产生一种或多种经富集的非哺乳动物核酸分子。

[0034] 方法

[0035] 在一些情况下,本文所提供的公开内容描述了一种富集一种或多种非哺乳动物核酸分子(例如,非哺乳动物DNA)的方法。在一些情况下,富集一种或多种非哺乳动物核酸分子可以通过基于蛋白质的非哺乳动物(例如,微生物)核酸分子富集204来富集。在一些情况下,非哺乳动物DNA富集可以包括:将所述一种或多种哺乳动物核酸分子和所述一种或多种非哺乳动物核酸分子与一种或多种重组CXXC结构域蛋白组合以形成蛋白质-DNA结合反应;在促进所述重组CXXC结构域蛋白与所述一种或多种哺乳动物核酸分子或所述一种或多种非哺乳动物核酸分子的非甲基化CpG基序之间的相互作用的条件下温育所述蛋白质-DNA结合反应;从所述蛋白质-DNA结合反应的剩余部分中分离未经结合的重组CXXC结构域蛋白和与所述非甲基化CpG核酸片段结合的重组CXXC结构域蛋白;以及洗涤与所述非甲基化CpG核酸片段结合的所述重组CXXC结构域蛋白,由此产生一种或多种经富集的核酸分子以供扩增。

[0036] 在一些情况下,扩增可以包括标记基因扩增205。在一些情况下,标记基因可以包括微生物系统发育标记基因或其标记基因片段。在一些情况下,微生物系统发育标记基因可以包括细菌标记基因、真菌标记基因或其任何组合。在一些情况下,微生物系统发育标记基因可以包括细菌标记基因或其标记基因片段。在一些情况下,微生物系统发育标记基因可以包括真菌标记基因或其标记基因片段。在一些情况下,细菌标记基因可以包括核糖体RNA基因5S、核糖体RNA基因16S、核糖体RNA基因23S、细菌管家基因dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf或其任何组合。在一些情况下,细菌核糖体RNA基因,如图7A中所示,可以包括高变区(V1-V9),所述高变区可以用于区分和/或分类微生物分类学。在一些情况下,一个或多个正向和/或反向引物,如图7B中所示,可以用于扩增细菌核糖体RNA基因的16S区以区分系统发育的各种细菌集,所述系统发育的各种细菌集可以用作区分、确定和/或诊断受试者和/或一组受试者的健康状况的特征。

[0037] 在一些情况下,所述真菌标记基因可以包括:核糖体RNA基因18S、核糖体RNA基因5.8S、核糖体RNA基因28S、内部转录间隔区1和2或其任何组合。在一些情况下,内部转录间隔区1和2(分别为ITS1和ITS2)位于小和大核糖体RNA(rRNA)提交18S rRNA、5.8S rRNA和28S RNA之间,如图8中所示。在一些情况下,如本文其它部分中所述,ITS1和/或ITS2区的扩增和测序提供基因组特征和/或标记以检测和/或确定受试者和/或一组受试者的生物样品中是否存在一种或多种真菌。在一些情况下,一种或多种真菌可以提供可以区分、分类和/或诊断受试者和/或多个受试者的健康状况的分类学特征。在一些情况下,ITS1和/或ITS2区。在一些情况下,扩增可以通过进行聚合酶链式反应(PCR)或其衍生物实现和/或完成。在一些情况下,聚合酶链式反应的衍生物可以包括反向引物PCR、反向PCR、锚定PCR、引物定向滚环扩增或其任何组合。在一些情况下,所述聚合酶链式反应扩增可以包括被配置成防止一个或多个基因组特征扩增的阻断引物、标记基因引物或其任何组合。在一些情况下,所述一个或多个基因组特征可以包括线粒体DNA基因组特征。

[0038] 在一些情况下,所述经富集和/或经扩增的一种或多种核酸分子可以通过序列文库制备300被制备以用于测序,如图12A中所示。在一些情况下,所述方法可以包括:提供经扩增和/或经富集的一种或多种核酸302;偶联条形码索引序列304;以及将一种或多种适配器序列与条形码索引序列相偶联306。在一些情况下,核酸分子的文库可以包括约256bp的碱基对长度。在一些情况下,经扩增和/或经富集的一种或多种核酸分子可以包括约90bp的长度。在一些情况下,经扩增的一种或多种核酸分子可以包括用V6引物扩增的血浆生物样品的游离DNA。如图12B所示的示例实验的凝胶电泳结果所示,经富集和/或经扩增的核酸分子的各种长度从文库制备之前308到文库制备之后310都在增加。

[0039] 然后将哺乳动物和/或非哺乳动物核酸分子的一种或多种经富集和/或经扩增的核酸分子组合物的所得制备文库通过靶向扩增子测序206方法进行测序,例如,靶向微生物扩增子测序可以被用于微生物分类学特征方法213和/或微生物功能性特征方法216中,分别如图2和3中所示的。在一些情况下,靶向微生物扩增子测序可以包括微生物16S扩增子测序。在一些情况下,通过对一个或多个经富集和/或经扩增的核酸分子组合物进行测序所生成的一个或多个测序读段可以被预处理,如图13中所示。在一些情况下,所述预处理可以包括:通过fastp处理经富集和/或经扩增的核酸组合物的一个或多个测序读段以去除适配器序列并执行质量控制以生成一个或多个经处理的测序读段312;从经处理的一个或多个测序读段生成子操作分类学单元以执行质量314;以及对照基因组数据库查询子操作分类学单元以分配一个或多个子操作单元分类学316。在一些情况下,质量控制可以包括约30或至少约30的平均读取质量。在一些情况下,基因组数据库可以包括16S GreenGenes 13.8。在一些情况下,Qiime2的sklearn分类器可用于分配子操作单元分类学。在一些情况下,子操作分类学单元可以使用基于质量分数、预期错误率、观察到的每个独特序列的频率或其组合对序列的误差轮廓进行建模的去噪工具Deblur生成。在一些情况下,靶向微生物扩增子测序206可以包括鸟枪法测序、下一代测序207、通过合成进行的测序或其组合。在一些情况下,微生物分类学特征213和/或微生物功能性特征可以是一组一个或多个核酸分子特征的一部分,如本文其它部分所述。在一些情况下,微生物分类学未来方法213可以确定经富集和/或经扩增的核酸分子的一种或多种微生物分类学分配和相关微生物丰度。在一些情况下,微生物功能性特征方法可以确定经富集和/或经扩增的核酸分子的一种或多种微生物功能性途径。

[0040] 在一些情况下,微生物功能性特征方法216(图3)可以包括:对经富集和/或经扩增的核酸分子文库进行测序,例如,使用下一代测序,以生成一组测序读段207;从所述组测序读段中过滤一种或多种核酸分子序列(例如,线粒体DNA)208,由此生成一个或多个线粒体DNA耗竭的测序读段209;鉴定一个或多个线粒体DNA耗竭的测序读段的一个或多个微生物分类学分配210;对一个或多个微生物分类学分配进行去污211;注释和/或鉴定一个或多个经去污的微生物分类学测序读段的一个或多个微生物功能性特征214;以及输出一个或多个经鉴定和/或经注释的微生物功能性特征的特征集215。在一些情况下,一个或多个微生物功能性特征215可以与受试者的已知健康状况(217、218、219)结合使用以训练预测模型(例如,机器学习分类器),如图5中所示,其它部分中所述。在一些情况下,微生物功能性特征可以包括宏基因组功能性特征。在一些情况下,PICRUSt2可以确定和/或鉴定一个或多个经去污的微生物分类学测序读段的一个或多个宏基因组功能性特征。在一些情况下,微生

物分类学工作流程可以包括映射以确定来自线粒体DNA耗竭的测序读段的微生物分类学分配210。在一些情况下,去污可以包括计算机模拟去污。在一些情况下,去污可以去除一个或多个非内源性微生物测序读段,由此从线粒体DNA耗竭的非哺乳动物测序读段的一个或多个微生物分类学身份生成一个或多个经去污的微生物分类学分配和相关数量的测序读段。

[0041] 在一些情况下,微生物功能性方法213可以包括:对经富集和/或经扩增的核酸分子文库进行测序,例如,使用下一代测序,以生成一组测序读段207;从所述组测序读段中过滤一种或多种核酸分子序列(例如,线粒体DNA)208,由此生成一个或多个线粒体DNA耗竭的测序读段209;鉴定一个或多个线粒体DNA耗竭的测序读段的一个或多个微生物分类学分配210;对一个或多个微生物分类学分配进行去污211;以及输出经富集和/或经扩增的核酸分子文库的一个或多个经去污的微生物分类学特征212。在一些情况下,一个或多个微生物分类学特征可以与受试者的已知健康状况(217、218、219)结合使用以训练预测模型(例如,机器学习分类器),如图5中所示,其它部分中所述。在一些情况下,微生物分类学工作流程可以包括映射以确定来自线粒体DNA耗竭的测序读段的微生物分类学分配210。在一些情况下,映射可以包括将一个或多个线粒体DNA耗竭的核酸分子测序读段相对于一个或多个微生物参考数据库进行映射以确定线粒体DNA耗竭的非哺乳动物测序读段的微生物分类学身份。在一些情况下,映射可以由QIME2或其支持的其它版本执行。在一些情况下,一个或多个微生物参考数据库可以包括:细菌16S rRNA数据库Greengenes;细菌、真菌和古菌rRNA数据库SILVA;真核核糖体ITS区数据库UNITE;源自可公开获得且完整的微生物基因组序列的定制数据库;或其任何组合。在一些情况下,去污可以包括计算机模拟去污。在一些情况下,去污可以去除一个或多个非内源性微生物测序读段,由此从线粒体DNA耗竭的非哺乳动物测序读段的一个或多个微生物分类学身份生成一个或多个经去污的微生物分类学分配和相关数量的测序读段。

[0042] 在一些情况下,一个或多个核酸分子特征可以包括一种或多种核酸分子的基因组特征。在一些情况下,基因组特征可以包括微生物系统发育标记基因或其标记基因片段。在一些情况下,微生物系统发育标记基因可以包括细菌标记基因或其标记基因片段。在一些情况下,微生物系统发育标记基因可以包括真菌标记基因或其标记基因片段。在一些情况下,一个或多个核酸分子特征可以包括一种或多种非哺乳动物核酸分子(例如,微生物核酸分子)的特征、特征集和/或特征组,如本文其它部分所述。

[0043] 在一些情况下,一个或多个核酸分子特征可用于训练220一个或多个预测模型221(例如,机器学习分类器),如图4和5所示,如本文其它部分所述。在一些情况下,一个或多个核酸分子特征(213、216)可以包括微生物、细菌、真菌或其组合分类学和/或功能性分类和/或受试者或多个受试者的生物样品的一个或多个核酸分子的特征,如本文其它部分所述。

[0044] 在一些情况下,预测模型(例如,机器学习分类器)可以用具有以下已知健康状况的受试者的生物样品的一个或多个核酸分子的一个或多个核酸分子特征进行训练:健康217、非癌疾病219或癌218。在一些情况下,预测模型可以用一个或多个受试者的一个或多个微生物分类学特征213以及相关健康状况进行训练220,如图4中所示。在一些情况下,预测模型可以用一个或多个受试者的一个或多个微生物功能性特征216以及相关健康状况进行训练,如图5中所示。在一些情况下,经训练的预测模型221可以包括一个或多个分类器(222、223、224),所述一个或多个分类器可以区分、分类和/或诊断未纳入预测模型的训练

的一个或多个受试者的健康状况。在一些情况下,一个或多个分类器可以包括健康相对于癌症健康状况分类器222、癌相对于非癌疾病健康分类器223、非癌疾病相对于健康分类器或其任何组合。

[0045] 预测模型

[0046] 本公开的方法和系统可以利用或访问基于一个或多个核酸分子特征训练的可以用于分类、诊断和/或表征受试者、多个受试者和/或一组或多组受试者的健康状况的人工智能、预测模型和/或机器学习的外部能力。在一些情况下,如本文其它部分所述,一个或多个核酸分子特征(例如,微生物功能性特征、微生物分类学特征等)可以预测、分类和/或鉴定一个或多个受试者的癌症和/或非癌疾病。在一些情况下,本文其它部分所述的一个或多个核酸分子特征可以用于训练一种或多种预测模型。这些特征可用于准确预测、分类和/或表征受试者、多个受试者和/或一组或多组受试者的健康状况,例如,癌症、非癌疾病、病症或其任何组合。使用此类预测能力,医疗保健提供者(例如,医师)可以能够做出明智的、准确的基于风险的决策,由此提高向受试者提供的关于癌症、非癌疾病、病症或其任何组合的护理和监测的质量。

[0047] 本公开的方法和系统可以分析微生物的存在和/或丰度(例如,特定属、分类学、微生物功能性途径的微生物丰度)。然后微生物的存在和/或丰度可用于确定一个或多个核酸分子特征,例如,可以预测一个或多个受试者的癌症和/或非癌疾病的非哺乳动物核酸分子特征。在一些情况下,本文其它部分描述的方法和/或系统可以用指示受试者的健康状况(例如,癌症和/或非癌疾病)的一个或多个核酸分子特征来训练预测模型。在一些情况下,然后经训练的预测模型可以用于生成与用于训练预测模型的一个或多个受试者不同的一个或多个受试者的癌症和/或非癌疾病的可能性(例如,预测)。经训练的预测模型可以包括基于人工智能的模型,如基于机器学习的分类器,所述分类器被配置成处理来自一种或多种核酸分子和/或经富集、经过滤和/或经扩增的一种或多种核酸分子的一个或多个核酸分子特征,以生成受试者患有癌症、非癌疾病或病症的可能性。所述模型可以使用来自一组或多组受试者,例如,癌症受试者、患有非癌疾病的受试者、无疾病和无癌症的受试者、接受癌症治疗的癌症受试者、接受非癌疾病治疗的受试者或其任何组合的微生物分类学特征或微生物功能性途径的丰度进行训练。在一些情况下,预测模型可以被训练以提供治疗预测以治疗并非预测模型的训练数据集的一部分的一个或多个受试者的癌症。当提供患者的经杂交富集的生物样品的一种或多种微生物的存在和丰度的输入时,此类预测模型可以向并非训练数据集的一部分的一个或多个受试者输出治疗建议。

[0048] 预测模型可以包括一种或多种预测模型。模型可以包括一种或多种机器学习算法。机器学习算法的实例可以包含支持向量机(SVM)、朴素贝叶斯分类、随机森林、神经网络(如深度神经网络(DNN))、循环神经网络(RNN)、深度RNN、长短期记忆(LSTM)循环神经网络(RNN)、门控循环单元(GRU)、梯度提升机、随机森林或其它监督学习算法或无监督机器学习、统计、线性回归、k最近邻、k均值、决策树、逻辑回归或其任何组合。模型可以用于分类或回归。模型同样可以涉及包含多个预测模型的集成模型的估计,并利用如梯度提升等技术,例如在梯度提升决策树的构造中。模型可以使用一个或多个训练数据集进行训练,所述一个或多个训练数据集包括一个或多个核酸分子特征、受试者数据(例如,受试者病史、受试者家族病史)、受试者生命体征(例如,血压、脉搏、体温、血氧饱和度)、受试者的已知健康状

况或其任何组合。

[0049] 预测模型可以包括任何数量的机器学习算法。在一些实施例中,随机森林机器学习算法可以是袋装决策树的集合。集合可以是至少约1、2、3、4、5、10、20、30、40、50、60、70、80、90、100、120、140、160、180、200、250、500、1000棵或更多棵袋装决策树。集合可以是至多约1000、500、250、200、180、160、140、120、100、90、80、70、60、50、40、30、20、10、5、4、3、2棵或更少棵袋装决策树。集合可以是约1至1000、1至500、1至200、1至100或1至10棵袋装决策树。

[0050] 在一些实施例中,机器学习算法可以有各种参数。各种参数可以是例如,学习率、小批量大小、训练的次数、动量、学习权重衰减或神经网络层数等。

[0051] 在一些实施例中,学习率可以为约0.00001至0.1。

[0052] 在一些实施例中,小批量大小可以为约16至128。

[0053] 在一些实施例中,神经网络可以包括神经网络层。神经网络可以具有至少约2至1000个或更多个神经网络层。

[0054] 在一些实施例中,训练的次数可以为至少约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、150、200、250、500、1000、10000次或更多次。

[0055] 在一些实施例中,动量可以为至少约0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更大。在一些实施例中,动量可以为至多约0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1或更小。

[0056] 在一些实施例中,学习权重衰减可以为至少约0.00001、0.0001、0.001、0.002、0.003、0.004、0.005、0.006、0.007、0.008、0.009、0.01、0.02、0.03、0.04、0.05、0.06、0.07、0.08、0.09、0.1或更大。在一些实施例中,学习权重衰减可以为至多约0.1、0.09、0.08、0.07、0.06、0.05、0.04、0.03、0.02、0.01、0.009、0.008、0.007、0.006、0.005、0.004、0.003、0.002、0.001、0.0001、0.00001或更小。

[0057] 在一些实施例中,机器学习算法可以使用损失函数。损失函数可以是例如,回归损失、平均绝对误差、平均偏置误差、铰链损失、亚当优化器和/或交叉熵等。

[0058] 在一些实施例中,机器学习算法的参数可以借助人和/或计算机系统进行调整。

[0059] 在一些实施例中,机器学习算法可以优先考虑某些特征。机器学习算法可以优先考虑可能与检测癌症、非癌疾病、病症或其任何组合更相关的特征。如果在确定癌症、非癌疾病和/或病症时,所述特征比另一特征更常被分类,则所述特征可能与检测癌症、非癌疾病和/或病症更相关。在一些情况下,特征可以使用加权系统进行优先排序。在一些情况下,特征可以基于特征出现的频率和/或数量按照概率统计学进行优先排序。机器学习算法可以借助人和/或计算机系统优先考虑某些特征。

[0060] 在一些情况下,机器学习算法可以优先考虑某些特征以降低计算成本、节省处理能力、节省处理时间、提高可靠性或减少随机存取存储器的使用等。

[0061] 训练数据集可以例如从患有常见癌症、非癌疾病或病症诊断的一组或多组受试者生成。训练数据集可以包括呈一个或多个受试者的生物样品中存在的微生物的丰度分类学分配特征和/或生物样品中存在的微生物的微生物功能性途径特征的一个或多个核酸分子特征。特征可以包括一个或多个受试者与微生物特征相对应的癌症诊断。在一些情况下,特征可以包括患者信息,如患者年龄、患者病史、其它医疗状况、当前或过去药物、临床风险评

分以及自上次观察以来的时间。例如,在给定时间点从给定患者收集的特征集可以共同用作签名,所述签名可以指示患者在给定时间点的健康状况或状态。

[0062] 标记可以包括临床结果,如受试者(例如,患者)的癌症、非癌疾病、病症或其组合的存在、不存在、诊断和/或预后。临床结果可以包括治疗功效(例如,受试者是否对基于癌症和/或疾病的治疗积极或消极响应)。

[0063] 输入特征可以通过将数据聚集到箱中或者可替代地使用独热编码来构造。输入还可以包含源自前面提到的输入的特征值或向量,如交叉相关。

[0064] 训练数据集可以从由指示癌症、非癌疾病、病症或其任何组合的生物样品的经富集和/或经扩增的核酸分子中鉴定和/或分类的例如一个或多个微生物分类学特征、一个或多个微生物功能性通路或其组合的一个或多个核酸摩尔特征的存在和/或丰度进行构建。

[0065] 模型可以处理输入特征以生成包括一种或多种分类、一种或多种预测或其组合的输出值。例如,此类分类或预测可以包含癌症或无癌症存在的二元分类;非癌疾病的存在;病症的存在;或受试者的其任何组合分类。在一些情况下,一个或多个预测模型和/或机器学习算法可以介于以下对受试者进行分类:一组分类标记(例如,‘无癌症、非癌症疾病和/或病症’、‘明显癌症、非癌症疾病和/或病症’和‘可能癌症、非癌症疾病和/或病症’);发生特定癌症、非癌疾病和/或病症的可能性(例如,相对可能性或概率);指示存在癌症、非癌疾病和/或病症的评分;患者死亡可能性的‘风险因素’以及任何数值预测的置信区间。各种机器学习技术可以级联,使得机器学习技术的输出也可以用作模型的后续层或分段的输入特征。

[0066] 为了训练模型(例如,通过确定模型的权重和相关性)以生成实时分类或预测,模型可以使用本文其它部分描述的训练数据集和/或一种或多种训练特征进行训练。此类数据集和/或特征可能足够大以生成统计上显著的分类或预测。例如,数据集可以包括源自来自一个或多个受试者的生物样品中的真菌、病毒、古菌、细菌或其任何组合微生物存在和/或丰度的测序数据的一个或多个核酸分子特征。

[0067] 数据集可以被分成子集(例如,离散的或重叠的),如训练数据集、开发数据集和测试数据集。例如,数据集可以被分成包括80%的数据集的训练数据集、包括10%的数据集的开发数据集和包括10%的数据集的测试数据集。训练数据集可以包括数据集的约10%、约20%、约30%、约40%、约50%、约60%、约70%、约80%或约90%。开发数据集可以包括数据集的约10%、约20%、约30%、约40%、约50%、约60%、约70%、约80%或约90%。测试数据集可以包括数据集的约10%、约20%、约30%、约40%、约50%、约60%、约70%、约80%或约90%。在一些实施例中,可以采用留一法交叉验证。训练集(例如,训练数据集)可以通过将对应于一个或多个患者队列的数据集随机采样来选择以确保采样的独立性。可替代地,训练集(例如,训练数据集)可以通过将对应于一个或多个患者队列的数据集按比例采样来选择以确保采样的独立性。

[0068] 为了提高模型预测的准确度并减少模型的过度拟合,数据集可以被扩充以增加训练集内的样品数量。例如,数据扩充可以包括将训练记录中的观察顺序重新排列。为了适应具有耗竭观察的数据集,可以使用推断耗竭数据的方法,如前向填充、后向填充、线性插值和多任务高斯过程。数据集可以被过滤或批量校正以消除或减轻混杂因素。例如,在数据库内,可以排除受试者的子集。

[0069] 模型可以包括一种或多种神经网络,如神经网络、卷积神经网络(CNN)、深度神经网络(DNN)、循环神经网络(RNN)或深度RNN。循环神经网络可以包括可以是长短期记忆(LSTM)单元或门控循环单元(GRU)的单元。例如,所述模型可以包括算法架构,所述算法架构包括具有输入特征集的神经网络,如本文其它部分所述,例如,一个或多个核酸分子特征、生命测量结果、受试者病史、受试者人口统计学特征或其任何组合。在训练模型期间可以使用神经网络技术,如丢弃或正则化,以防止过度拟合。神经网络可以包括多个子网络,所述多个子网络中的每一者被配置成生成不同类型的输出信息的分类或预测,所述分类或预测可以被组合以形成神经网络的总体输出。机器学习模型可以可替代地利用统计或相关算法,包含随机森林、分类和回归树、支持向量机、判别分析、回归技术以及其集成和梯度提升变体。

[0070] 当模型生成癌症、非癌疾病、病症或其组合的分类或预测时,可以生成通知(例如,警报或报警)并将其传输给医疗保健提供者,如医院内的医生、护士或受试者的治疗团队的其它成员。通知可以经由自动电话呼叫、短信服务(SMS)或多媒体消息服务(MMS)消息、电子邮件和/或仪表板内的警报来传输。通知可以包括输出信息,如癌症、非癌疾病和/或病症的预测;预测的癌症、非癌疾病和/或病症的可能性;癌症、非癌疾病和/或病症预计发病前的时间;癌症、非癌疾病和/或病症的可能性或时间的置信区间、推荐治疗方案或其任何组合信息。

[0071] 为了验证模型的性能,可以生成不同的性能指标。例如,接受者操作特征曲线下面积(AUROC)可用于确定模型的诊断、预后、筛查或其任何组合能力。例如,模型可以使用可调节的分类学阈值,使得特异性和灵敏度是可调的,并且接受者操作特征曲线(ROC)可以用于鉴定与不同的特异性和灵敏度值相对应的不同操作点。

[0072] 在一些情况下,如当数据集不够大时,可以进行交叉验证以评估模型在不同训练和测试数据集集中的稳健性。

[0073] 为了计算如灵敏度、特异性、准确度、阳性预测值(PPV)、阴性预测值(NPV)、精确回忆曲线下面积(AUPR)、AUROC或类似物的性能指标,可以使用以下定义。“假阳性”可以指错误地或过早地生成阳性结果或后果的结果(例如,在癌症、非癌疾病和/或病症实际发作之前或没有任何发作)。“真阳性”可以指当患者患有癌症、非癌疾病和/或病症(例如,患者显示出癌症、非癌疾病和/或病症的症状,或者患者的记录表明癌症、非癌疾病和/或病症)时已正确生成阳性结果或后果的结果。“假阴性”可以指其中已生成阴性结果或后果,但患者患有癌症、非癌疾病和/或病症的结果(例如,患者显示出癌症、非癌疾病和/或病症的症状,或者患者的记录表明癌症、非癌疾病和/或病症)。“真阴性”可以指已生成阴性结果或后果的结果(例如,在癌症、非癌疾病和/或病症实际发作之前或没有任何发作)。

[0074] 可以训练模型直到满足准确度或性能的某些预定条件,如具有与诊断准确度测量相对应的最小期望值。例如,诊断准确度测量可以对应于对受试者的癌症、非癌疾病和/或病症发生的可能性的预测。作为另一个实例,诊断准确度测量可以对应于先前已接受治疗的受试者的癌症、非癌疾病和/或病症恶化或复发的可能性的预测。诊断准确度测量的实例可以包含对应于检测或预测癌症、非癌疾病和/或病症的诊断准确度的灵敏度、特异性、阳性预测值(PPV)、阴性预测值(NPV)、准确度、AUPR和AUROC。

[0075] 例如,此类预定条件可以是预测癌症、非癌疾病和/或病症的灵敏度包括例如至少

约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的值。

[0076] 作为另一个实例,此类预定条件可以是预测癌症、非癌疾病和/或病症的特异性包括例如至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的值。

[0077] 作为另一个实例,此类预定条件可以是预测癌症、非癌疾病和/或病症的阳性预测值 (PPV) 包括例如至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的值。

[0078] 作为另一个实例,此类预定条件可以是预测癌症、非癌疾病和/或病症的阴性预测值 (NPV) 包括例如至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的值。

[0079] 作为另一个实例,此类预定条件可以是预测癌症、非癌疾病和/或病症的接受者操作特征 (ROC) 曲线的曲线下面积 (AUC) (AUROC) 包括至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.85、至少约0.90、至少约0.95、至少约0.96、至少约0.97、至少约0.98或至少约0.99的值。

[0080] 作为另一个实例,此类预定条件可以是预测癌症、非癌疾病和/或病症的精确回忆曲线下面积 (AUPR) 包括至少约0.10、至少约0.15、至少约0.20、至少约0.25、至少约0.30、至少约0.35、至少约0.40、至少约0.45、至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.85、至少约0.90、至少约0.95、至少约0.96、至少约0.97、至少约0.98或至少约0.99的值。

[0081] 在一些实施例中,经训练的模型可以被训练或配置成以至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的灵敏度来预测癌症、非癌疾病和/或病症。

[0082] 在一些实施例中,经训练的模型可以被训练或配置成以至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的特异性来预测癌症、非癌疾病和/或病症。

[0083] 在一些实施例中,经训练的模型可以被训练或配置成以至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的阳性预测值 (PPV) 来预测癌症、非癌疾病和/或病症。

[0084] 在一些实施例中,经训练的模型可以被训练或配置成以至少约50%、至少约55%、至少约60%、至少约65%、至少约70%、至少约75%、至少约80%、至少约85%、至少约90%、至少约95%、至少约96%、至少约97%、至少约98%或至少约99%的阴性预测值 (NPV) 来预

测癌症、非癌疾病和/或病症。

[0085] 在一些实施例中,经训练的模型可以被训练或配置成以至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.85、至少约0.90、至少约0.95、至少约0.96、至少约0.97、至少约0.98或至少约0.99的接受者操作特征(ROC)曲线的曲线下面积(AUC)(AUROC)来预测癌症、非癌疾病和/或病症。

[0086] 在一些实施例中,经训练的模型可以被训练或配置成以至少约0.10、至少约0.15、至少约0.20、至少约0.25、至少约0.30、至少约0.35、至少约0.40、至少约0.45、至少约0.50、至少约0.55、至少约0.60、至少约0.65、至少约0.70、至少约0.75、至少约0.80、至少约0.85、至少约0.90、至少约0.95、至少约0.96、至少约0.97、至少约0.98或至少约0.99的精确回忆曲线下面积(AUPR)来预测癌症、非癌疾病和/或病症。

[0087] 训练数据集可以从训练受试者(例如,人)收集。每次训练都具有诊断状态,表明所述受试者已被诊断出患有生物病状或尚未被诊断出患有癌症、非癌疾病和/或病症。

[0088] 在一些实施例中,模型是神经网络或卷积神经网络。参见Vincent等人,2010,“堆叠式去噪自动编码器:使用局部降噪标准在深度网络中学习有用的表示(Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion)”,《机器学习研究杂志(J Mach Learn Res)》11,第3371-3408页;Larochelle等人,2009,“探索训练深度神经网络的策略(Exploring strategies for training deep neural networks)”,《机器学习研究杂志》10,第1-40页;以及Hassoun,1995,《人工神经网络基础知识(Fundamentals of Artificial Neural Networks)》,麻省理工学院(Massachusetts Institute of Technology),所述参考文献中的每一者通过引用特此并入。

[0089] 在一些实施例中,独立成分分析(ICA)被用于对数据进行去维处理,如Lee,T.-W.(1998):《独立成分分析:理论与应用(Independent component analysis:Theory and applications)》,波士顿马萨诸塞州:克鲁维尔学术出版社(Boston,Mass:Kluwer Academic Publishers),ISBN 0-7923-8261-7;以及Hyvärinen,A.;Karhunen,J.;Oja,E.(2001):《独立成分分析(Independent Component Analysis)》,纽约:威利出版社(New York:Wiley),ISBN 978-0-471-40540-5中所述,所述参考文献中的每一者通过引用以其整体特此并入。

[0090] 在一些实施例中,主要成分分析(PCA)被用于对数据进行去维处理,如Jolliffe,I.T.(2002).主要成分分析(Principal Component Analysis).《斯普林格统计学系列(Springer Series in Statistics)》.纽约:施普林格出版社(Springer-Verlag).doi:10.1007/b98835.ISBN 978-0-387-95442-4中所述,所述参考文献通过引用以其整体特此并入。

[0091] SVM描述于Cristianini和Shawe-Taylor,2000,“《支持向量机入门(An Introduction to Support Vector Machines)》”,剑桥的剑桥大学出版社(Cambridge University Press,Cambridge);Boser等人,1992,“最优边际分类器的训练算法(A training algorithm for optimal margin classifiers)”于《第五届ACM年度计算学习理论研讨会论文集(5th Annual ACM Workshop on Computational Learning Theory)》中,宾夕法尼亚州匹兹堡的ACM出版社(ACM Press,Pittsburgh,Pa.),第142-152页;Vapnik,

1998,《统计学习理论(Statistical Learning Theory)》,纽约的威利出版社;Mount,2001,《生物信息学:序列和基因组分析(Bioinformatics:sequence and genome analysis)》,纽约冷泉港的冷泉港实验室出版社(Cold Spring Harbor Laboratory Press,Cold Spring Harbor);Duda,《模式分类(Pattern Classification)》,第二版,2001,约翰·威利父子出版公司(John Wiley&Sons,Inc.),第259页,262-265;以及Hastie,2001,《统计学习的要素(The Elements of Statistical Learning)》,纽约的施普林格出版社;以及Furey等人,2000,《生物信息学(Bioinformatics)》16,906-914,所述参考文献中的每一者通过引用以其整体特此并入。当用于分类时,SVM将给定的经二进制标记的数据集与最大程度地远离所标记的数据的超平面分离。对于可能没有线性分离的情况,SVM可以结合自动实现对特征空间的非线性映射的“核”技术运行。由SVM在特征空间中发现的超平面对应于输入空间中的非线性决策边界。

[0092] 决策树通过以下总体描述:Duda,2001,《模式分类》,纽约的约翰·威利父子出版公司,第395-396页,所述参考文献通过引用特此并入。基于树的方法将特征空间划分成一组矩形,并且然后在每个矩形中拟合模型(比如常量)。在一些实施例中,决策树是随机森林回归。可以使用的一个特定算法是分类和回归树(CART)。其它特定决策树算法包含但不限于ID3、C4.5、MART和随机森林。CART、ID3和C4.5描述于Duda,2001,《模式分类》,纽约的约翰·威利父子出版公司,第396-408页和第411-412页,所述参考文献通过引用特此并入。CART、MART和C4.5描述于Hastie等人,2001,《统计学习的要素》,纽约的施普林格出版社,第9章,所述参考文献通过引用以其整体特此并入。随机森林描述于Breiman,1999,“随机森林—随机特征(Random Forests—Random Features)”,《技术报告(Technical Report)》567,伯克利的加州大学统计系(Statistics Department,U.C.Berkeley),1999年9月,所述文献通过引用以其整体特此并入。

[0093] 聚类(例如,无监督聚类模型算法和监督聚类模型算法)描述于Duda和Hart,《模式分类和场景分析(Pattern Classification and Scene Analysis)》,1973,纽约的约翰·威利父子出版公司的第211-256页中(下文中称为“Duda 1973”),所述文献通过引用以其整体特此并入。如Duda 1973的第6.7节中所描述的,聚类问题被描述为在数据集中寻找自然分组的问题。为了鉴定自然分组,解决了两个问题。第一,确定测量两个样品之间的类似性(或相异度)的方式。此度量(类似性量度)被用于以确保一个集群中的样品比其它集群中的样品彼此更类似。第二,确定用于使用类似性量度将数据划分成集群的机制。Duda 1973的6.7节中讨论了类似性量度,其中称开始聚类调查的一种方式定义距离函数并且计算训练集中所有对样品之间的距离的矩阵。如果距离是良好类似性量度,则相同集群中的参考实体之间的距离将显著小于不同集群中的参考实体之间的距离。然而,如Duda 1973的第215页所述,聚类不需要使用距离量度。例如,非度量类似性函数 $s(x, x')$ 可用于比较两个向量 x 和 x' 。通常,当 x 和 x' 在某种程度上“类似”时, $s(x, x')$ 为值较大的对称函数。Duda 1973的第218页上提供了非对称类似性函数 $s(x, x')$ 的实例。一旦已经选择了用于测量数据集中的点之间的“类似性”或“相异度”的方法,聚类就需要测量数据的任何分区的聚类质量的准则函数。将准则函数极值化的数据集的分区被用于对数据进行聚类。参见Duda 1973的第217页。准则函数讨论于Duda1973的第6.8节。最近,已出版了Duda等人,《模式分类》,第2版,纽约的约翰·威利父子出版公司第537-563页详细地描述了聚类。关于聚类技术的更多信

息可见于Kaufman和Rousseeuw,1990,《数据调查组:聚类分析导论 (Finding Groups in Data:An Introduction to Cluster Analysis)》,纽约州纽约市的威利出版社;Everitt,1993,《聚类分析 (Cluster analysis)》(第3版),纽约州纽约市的威利出版社;以及Backer,1995,《聚类分析中的计算机辅助推理 (Computer-Assisted Reasoning in Cluster Analysis)》,新泽西州上鞍河的普伦蒂斯·霍尔出版社 (Prentice Hall,Upper Saddle River,New Jersey),所述参考文献中的每一者通过引用特此并入。可用于本公开的特定示例性聚类技术包含但不限于层次聚类(使用最近邻域算法、最远邻域算法、平均连接算法、质心算法或平方和算法的聚集聚类)、k均值聚类、模糊k均值聚类算法和帕特里克贾维斯聚类(Jarvis-Patrick clustering)。在一些实施例中,聚类包括无监督聚类,其中当对训练集进行聚类时没有强加应形成什么聚类的预想概念。

[0094] 回归模型,如多分类学逻辑模型的回归模型描述于Agresti,《分类数据分析简介 (An Introduction to Categorical Data Analysis)》,1996,纽约的约翰·威利父子出版公司,第8章,所述参考文献通过引用以其整体特此并入。在一些实施例中,模型利用Hastie等人,2001,《统计学习的要素》,纽约的施普林格出版社中公开的回归模型,所述参考文献通过引用以其整体特此并入。在一些实施例中,梯度提升模型用于例如本文描述的分类学算法;这些梯度提升模型描述于Boehmke,Bradley;Greenwell,Brandon (2019)。“梯度提升 (Gradient Boosting)”。《动手学机器学习与R (Hands-On Machine Learning with R)》。查普曼和霍尔 (Chapman&Hall)。第221-245页 ISBN 978-1-138-49568-5.,所述参考文献通过引用以其整体特此并入。在一些实施例中,使用集成建模技术;这些集成建模技术描述于本文中的分类模型的实施方案,并且描述于Zhou Zhihua (2012)《组合方法:基础和算法 (Ensemble Methods:Foundations and Algorithms)》。查普曼和霍尔/CRC. ISBN 978-1-439-83003-1,所述参考文献通过引用以其整体特此并入。

[0095] 在一些实施例中,机器学习分析由执行包含进行数据分析的指令的一个或多个程序(例如,存储在非持久性存储器或持久性存储器中的一个或多个程序)的装置来进行。在一些实施例中,数据分析由包括至少一个处理器(例如,处理核心)和包括进行数据分析的指令的存储器(例如,存储在非持久性存储器或持久性存储器中的一个或多个程序)的系统来进行。

[0096] 计算机系统

[0097] 本公开提供了被编程以实现本公开的方法的计算机系统。图6显示了计算机系统600,所述计算机系统经过编程或以其它方式被配置成预测一个或多个受试者的癌症、非癌疾病或其任何组合的健康状况;训练本文其它部分所述的预测模型;生成推荐疗法;或本文其它部分所述的其任何组合方法。计算机系统600可以是用户的电子装置或相对于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。

[0098] 计算机系统600包含中央处理单元(CPU,本文中也称为“处理器”和“计算机处理器”)606,所述中央处理单元可以是单核或多核处理器或用于并行处理的多个处理器。计算机系统600还包含存储器或存储器位置604(例如,随机存取存储器、只读存储器、闪存)、电子存储单元602(例如,硬盘)、用于与一个或多个其它系统通信的通信接口608(例如,网络适配器)以及外围装置610,如高速缓存、其它存储器、数据存储设备和/或电子显示适配器。存储器604、存储单元602、接口608和外围装置610通过如母线等通信总线(实线)与CPU 606

通信。存储单元602可以是用于存储数据的数据存储单元(或数据储存库)。计算机系统600可以借助于通信接口608可操作地耦接到计算机网络(“网络”)612。网络612可以是英特网、互联网和/或外联网或与英特网通信的内联网和/或外联网。在一些情况下,网络612是电信网络和/或数据网络。网络612可以包含一个或多个计算机服务器,所述一个或多个计算机服务器可以实现如云计算等分布式计算。在一些情况下,网络612借助于计算机系统600可以实施对等网络,所述对等网络可以实现将装置耦接到计算机系统600以充当客户端或服务器。

[0099] CPU 606可以执行一系列机器可读指令,所述一系列机器可读指令可以体现在程序或软件中。指令可以存储在如存储器604等存储器位置中。指令可以被引导至CPU 606,其可以随后对CPU 606进行编程或以其它方式配置以实现本文其它部分描述的本公开的方法。由CPU 606进行的操作的实例可以包含提取、解码、执行和写回。

[0100] CPU 606可以是如集成电路等电路的一部分。系统600的一个或多个其它组件可以包含在电路中。在一些情况下,所述电路是专用集成电路(ASIC)。

[0101] 存储单元602可以存储文件,如驱动程序、文库和保存的程序。存储单元602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600可以包含一个或多个另外的数据存储单元,所述一个或多个另外的数据存储单元位于计算机系统600外部,如定位在通过内联网或因特网与计算机系统600通信的远程服务器上。

[0102] 计算机系统600可以通过网络612与一个或多个远程计算机系统通信。例如,计算机系统600可以与用户的远程计算机系统通信。远程计算机系统的实例可以包含个人计算机(例如,便携式PC)、平板计算机或平板PC(例如, **Apple® iPad**、**Samsung® Galaxy Tab**)、电话、智能手机(例如, **Apple® iPhone**、支持Android的装置、**Blackberry®**)或个人数字助理。用户可以通过网络612访问计算机系统600。

[0103] 可以通过存储在计算机系统600的电子存储位置上(如存储器604或电子存储单元602上)的机器(例如,计算机处理器)可执行代码实施如本文所描述的方法。机器可执行代码或机器可读代码可以软件的形式提供。在使用期间,代码可以由处理器606执行。在一些情况下,代码可以从存储单元602中检索并且存储在存储器604上以供由处理器606随时访问。在一些情形下,可以排除电子存储单元602,并且在存储器604上存储机器可执行指令。

[0104] 代码可以被预先编译且被配置成用于与具有被适配成执行所述代码的处理器器的机器一起使用或可以在运行期间进行编译。代码可以编程语言供应,可以选择所述编程语言以实现以预先编译或类编译(as-compiled)的方式执行代码。

[0105] 在一些实施例中,如本文其它部分所述的,系统可以包括一种用于诊断一个或多个受试者的癌或非癌健康状况的系统。在一些情况下,所述系统可以包括:(a)一个或多个处理器;以及(b)包含软件的非暂时性计算机可读存储介质,所述软件被配置成使所述一个或多个处理器:(i)接收一个或多个受试者的生物样品的所述一个或多个受试者的一个或多个核酸分子测序读段,其中所述一个或多个核酸分子测序读段包括一种或多种非哺乳动物核酸分子的经扩增的一个或多个基因组特征的序列;以及(ii)至少作为将所述一个或多个非哺乳动物核酸测序读段的一个或多个基因组特征作为输入提供给经训练的预测模型的结果,输出对所述一个或多个受试者的癌或非癌健康状况的诊断。

[0106] 本文提供的系统和方法的各个方面,如计算机系统600,可以体现在编程中。所述

技术的各方面可以被认为是通常以机器(或处理器)可执行代码和/或相关联的数据的形式“产品”或“制品”,所述机器可执行代码和/或相关联的数据被承载在或体现在一种类型的机器可读介质中。机器可执行代码可以存储在如存储器(例如,只读存储器、随机存取存储器、闪存)或硬盘等电子存储单元上。“存储”类型介质可以包含计算机、处理器等的任何或全部有形存储器或其相关联的模块,如可以在任何时间为软件编程提供非暂时性存储的各种半导体存储器、磁带驱动、硬盘驱动等。软件的全部或部分有时可以通过因特网或各种其它电信网络通信。此类通信,例如,可以实现将软件从一个计算机或处理器加载到另一个计算机或处理器中,例如从管理服务器或主机计算机加载到应用服务器的计算机平台中。因此,可以承载软件元素的另一种类型的介质包含光波、电波和电磁波,如通过有线和光陆线网络以及通过各种空中链路在本地装置之间的物理接口上使用。承载此类波的物理元件,如有线或无线链路、光链路等,也可以被认为是承载软件的介质。如本文所用的,除非限定为非暂时性有形“存储”介质,否则如计算机或机器“可读介质”之类的术语是指参与向处理器提供指令以供执行的任何介质。

[0107] 因此,如计算机可执行代码等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,如任何计算机等中任何存储装置,如可以用于实现在附图中示出的数据库等。易失性存储介质包含动态存储器,如此类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,包含包括计算机系统内的总线的线。载波传输介质可以采取电信号或电磁信号,或声波或光波的形式,如在射频(RF)和红外(IR)数据通信期间产生的信号形式。因此,常见形式的计算机可读介质包含例如:软盘、可折叠磁盘、硬盘、磁带、任何其它磁介质、CD-ROM、DVD或DVD-ROM、任何其它光学介质、穿孔卡纸磁带、具有孔图案的任何其它物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其它存储芯片或盒式磁带、载波传输数据或指令、传输此类载波的电缆或链路或计算机可以从中读取编程代码和/或数据的任何其它介质。多种这些形式的计算机可读介质可涉及将一个或多个指令的一个或多个序列传送到处理器以便执行。

[0108] 计算机系统600可以包含电子显示器616或与所述电子显示器通信,所述电子显示器包括用于提供例如用于预测结果可视化的显示器或用于训练预测模型的界面的用户界面(UI)614。UI的实例包含但不限于图形用户界面(GUI)和基于网络的用户界面。

[0109] 虽然本文已经示出和描述本发明的优选实施例,但对于本领域技术人员明显的是此类实施例仅以实例的方式提供。在不脱离本发明的情况下,本领域技术人员现在将想到许多变化、改变和取代。应当理解,本文描述的本发明实施例的各种替代方案可用于实践本发明。预期的是以下权利要求书限定了本发明的范围以及由此覆盖在这些权利要求及其等效物的范围内的方法和结构。

[0110] 尽管如所述和本文权利要求的方法的步骤显示了根据实施例的所述方法或所述组中的每一者操作,但本领域的普通技术人员会根据本文所述的教学认识到许多变化。所述步骤可能会以不同的顺序完成。可以添加或省略步骤。一些步骤可以包括子步骤。许多步骤可以重复多次,只要有益即可。

[0111] 所述方法或所述组中的每一者操作的一个或多个步骤都可以用如本文所述的电路,例如,处理器或逻辑电路中,如现场可编程门阵列的可编程阵列逻辑中的一个或多个来

执行。例如,电路可以被编程以提供所述方法或所述组中的每一者操作的一个或多个步骤,并且程序可以包括存储在计算机可读存储器上的程序指令或逻辑电路(如可编程阵列逻辑或现场可编程门阵列)的编程步骤。

[0112] 实例1:16S rDNA V6引物扩增效率和特异性

[0113] 对16S rDNA V6扩增引物(例如,967F和1064R)的效率进行了评估,并对人类基因组和微生物标准品稀释系列进行了比较,如图9中可见。图9显示了各种组合物的针对所观察到的PCR反应产物产生的信号强度所绘制的聚合酶链式反应(PCR)周期:人基因组DNA 318、微生物标准品稀释系列322和反应中不含有DNA的阴性对照320。微生物标准品稀释系列包含具有790、7,896、78,955、789,554和7,895,540个16S拷贝数的5个标准品稀释。从PCR循环图中可以看出,当只有人基因组DNA存在时318,在与微生物标准品稀释322相比时,V6引物扩增人gDNA的效率较低。这些结果表明V6引物选择性地扩增微生物核酸分子,由此减少通过扩增宿主哺乳动物线粒体DNA产生的噪音或杂质扩增子。

[0114] 图10显示了在存在人gDNA的情况下,评估16S rDNA V6扩增引物(例如,967F、1064R)扩增效率的实验。通过V6引物制备了四组用于PCR扩增:掺入到来自白细胞的人基因组DNA中的微生物DNA标准品(V6/wbDNA)(324,326,328,330,332)、微生物DNA标准品(V6/mbDNA)(338,340,342,344,346)、仅人gDNA(334)和无模板对照(即阴性对照)(336)。反应中使用的人gDNA含量固定为5ng,对应于大约1,396个基因组当量。按质量计算,每个实验组中的人gDNA以大于或等于微生物DNA标准品的量的1000倍存在。五个微生物标准品组包括790(324,338)、7,896(326,340)、78,955(328,342)、789,554(330,344)和7,895,540(332,346)的16S拷贝数。从图10中显示的结果可以看出,尽管在PCR反应中在只存在人gDNA时观察到了扩增,但当两种DNA源混合时,人gDNA引物杂交事件并不妨碍微生物DNA的特异性扩增。

[0115] 图11显示了在存在人DNA的情况下进行以评估16S rDNA V6引物(例如,967F、1064R)扩增微生物DNA的特异性的实验。制备具有不同量的人基因组DNA的六个实验性小组:3ng(410)、0.3ng(402)、0.03ng(408)、0.003ng(414)、0.0003ng(406)和0ng(400),并在存在5pg微生物DNA标准品(7,895个基因组当量)的情况下扩增。此外,还使用了无模板对照组(412)作为阴性对照。从图37中所示的PCR循环图中可以看出,在所提及的所有水平的人gDNA上,都优先扩增微生物gDNA。在不存在微生物DNA的情况下,人DNA唯一对照“hDNA 3ng”曲线在效率降低时扩增,产生31的Cp。这些结果表明,V6引物的特异性很高,而且由于PCR扩增曲线是重叠的,因此不会受到大量人类DNA前景的阻碍。

[0116] 实例2:扩增子测序空白过滤

[0117] 在本文所述工作流和/或方法的不同步骤中,扩增子测序读段计数如图14A-14B中所示。图14B显示了在图14A中所示的“OTU”和“OTU_过滤”测序读段/样品的放大视图。实验组包含血浆(500)、空白阴性对照(502)和按规定浓度混合的工业Zymo商购可得DNA微生物样品(504)。每个样品的读段数图显示了如本文其它部分所述的核酸分子测序读段的不同点的读段数/样品:“原始_读段”是质量过滤或分类学分配前的总读段/样品;“qf_读段”是去除PCR重复序列的质量过滤步骤后剩余的测序读段;“OTU”对应于通过Deblur处理鉴定的次级操作分类学单元的每个样品的读段数,并且“OTU_过滤”对应于减去DNA提取空白对照(即“空白”)中存在的sOTU后剩余的sOTU。具有整个数据集中的至少10的丰度的特征被保留下来以便进一步进行下游处理。

[0118] 从图14A-14B中所示的结果可知,从空白对照中减去测序读段后,DNA提取过程本身存在的污染已被考虑在内,有数千个经去污的读段待进一步处理以确定一个或多个核酸特征以用于分类和/或鉴定一个或多个受试者的健康状况,如本文其它部分所述。经过空白过滤后,血浆样品中总丰度的75.75%被去除,即从总共30,689个微生物中去除了12,509个微生物(去除了微生物的40.76%)。因此,通过本文其它部分所述的方法和系统检测和分析的不是DNA污染,而是受试者的生物样品中存在的核酸分子信号。

[0119] 实例3:机器学习分类器性能

[0120] 用一个或多个受试者的16S扩增子序列(例如,V6超变区)和已知的健康状况标记(即特定的非癌疾病标记和/或非小细胞肺癌阶段)训练机器学习分类器。各种经标记的健康状况的受试者和/或样品的数量分布示出于图15A中。如本文其它部分所述,从每组所有受试者中获取血浆,并用V6 16S引物扩增,然后进行下一代测序。然后对测序读段进行去污处理以鉴定一种或多种微生物分类学特征。微生物分类学特征包含已鉴定的微生物分类学丰度,并折算到属一级。然后微生物分类学特征的相关读段被用于训练三个随机森林机器学习模型,并进行5倍交叉验证。这三种随机森林机器学习模型包含分类器以分类和/或表征I期、II期和III期的癌症健康状况。分类器的性能接收器操作特征曲线和相关曲线下面积(AUC),即I期的0.891、II期的0.71和III期癌症的0.88示出于图15B中。

[0121] 使用鸟枪法测序和测序读段的生命之网(WoL)属相对丰度生成用于I期(506)、II期(510)和III期(508)癌症健康状况分类和/或鉴定的三个单独的机器学习分类器。用鸟枪法测序和WoL丰度进行训练的机器学习分类器的相关AUC为I期癌症的0.837,II期癌症的0.66和III期癌症的0.871。

[0122] 从这些结果中可以看出,V6 16S扩增引物提供了一种或多种经富集和/或经扩增的核酸分子,所述一种或多种经富集和/或经扩增的核酸分子可用于开发在区分I期、II期和III期癌症方面具有高准确性的一种或多种微生物分类学特征。

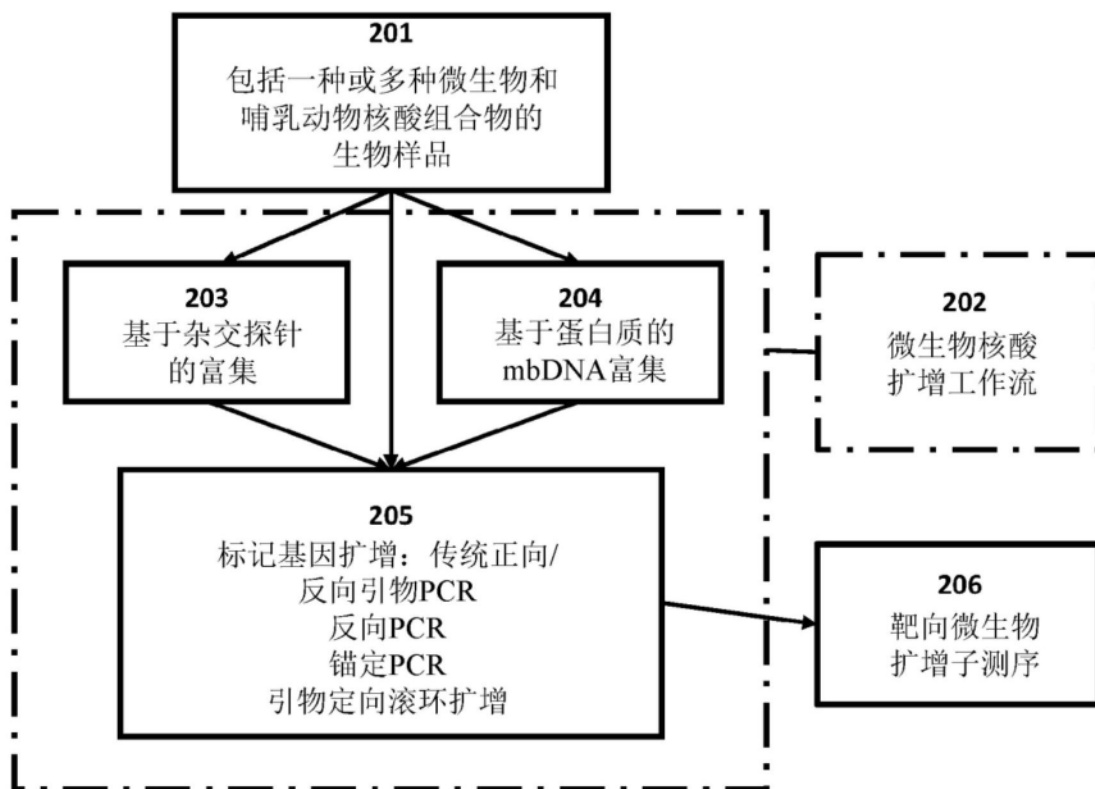


图1

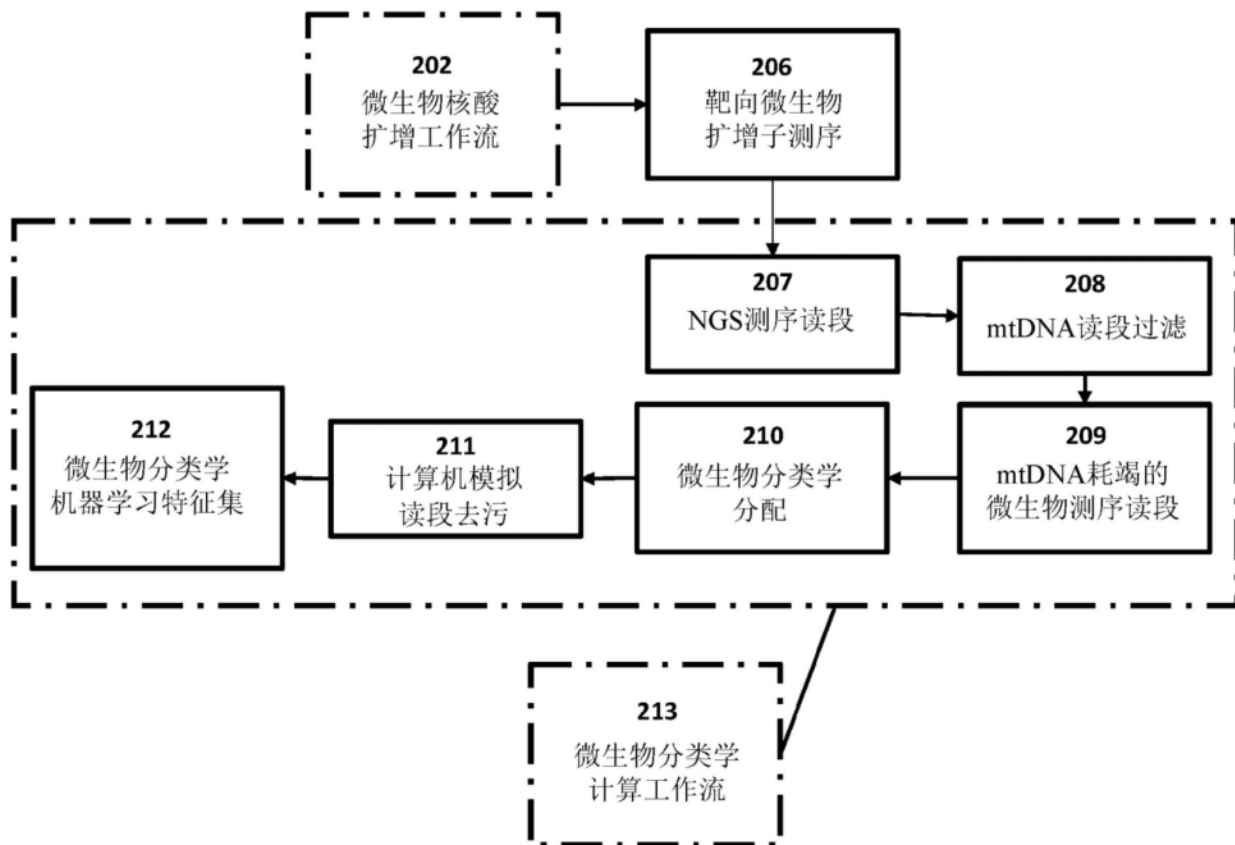


图2

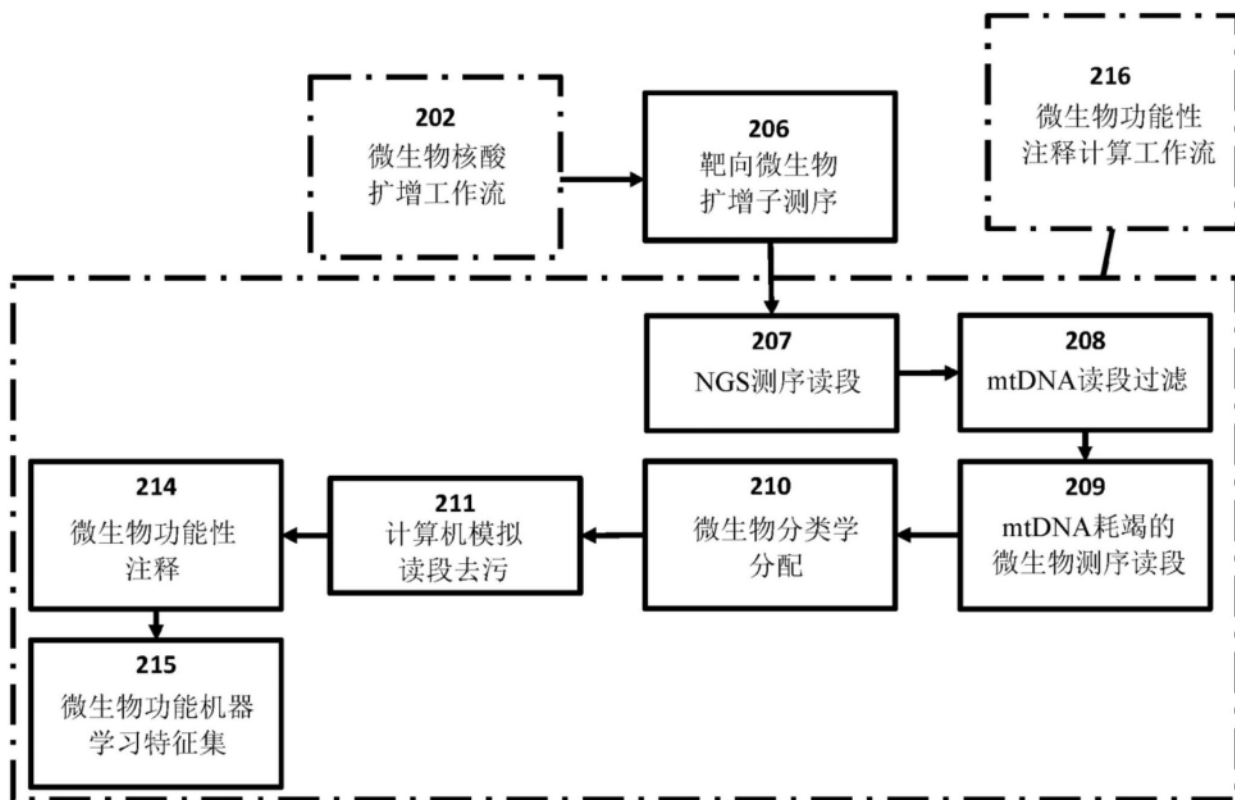


图3

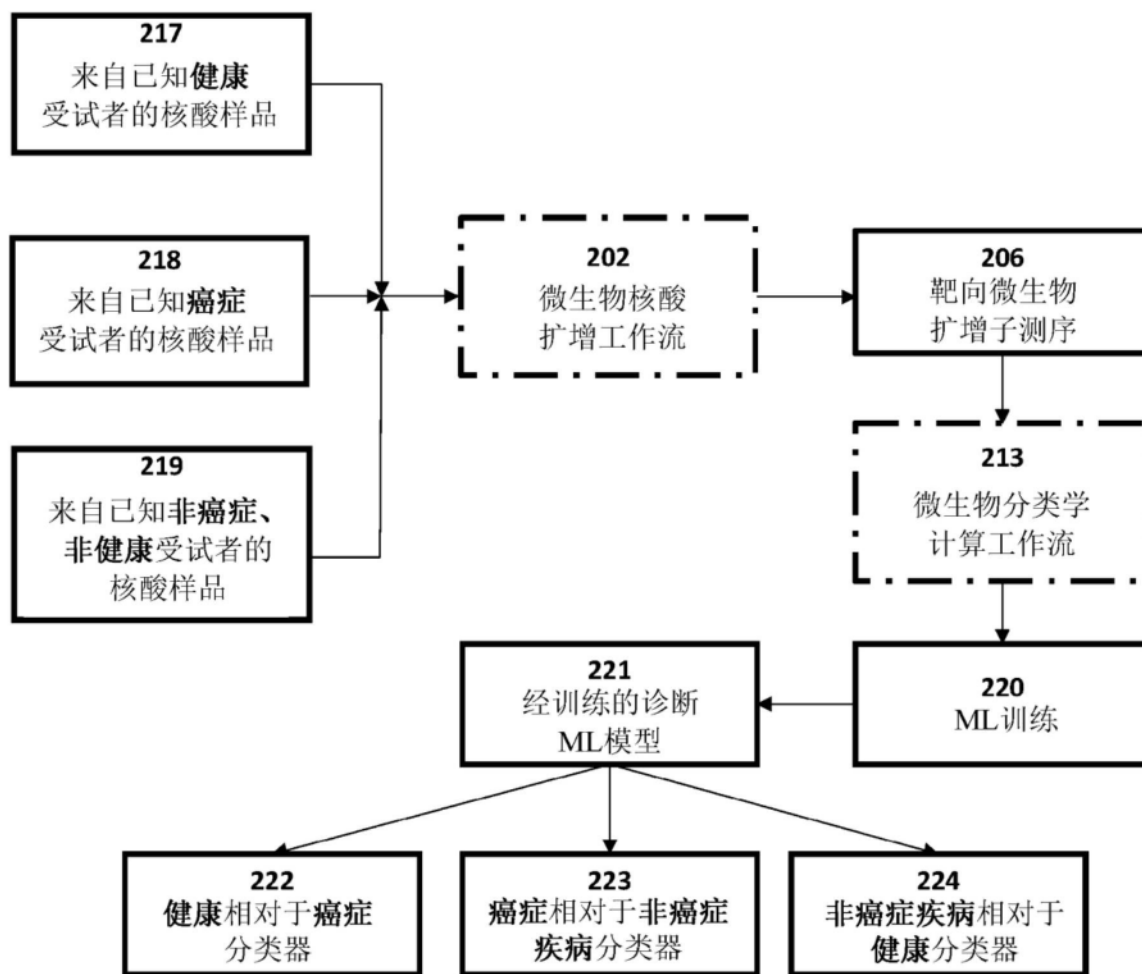


图4

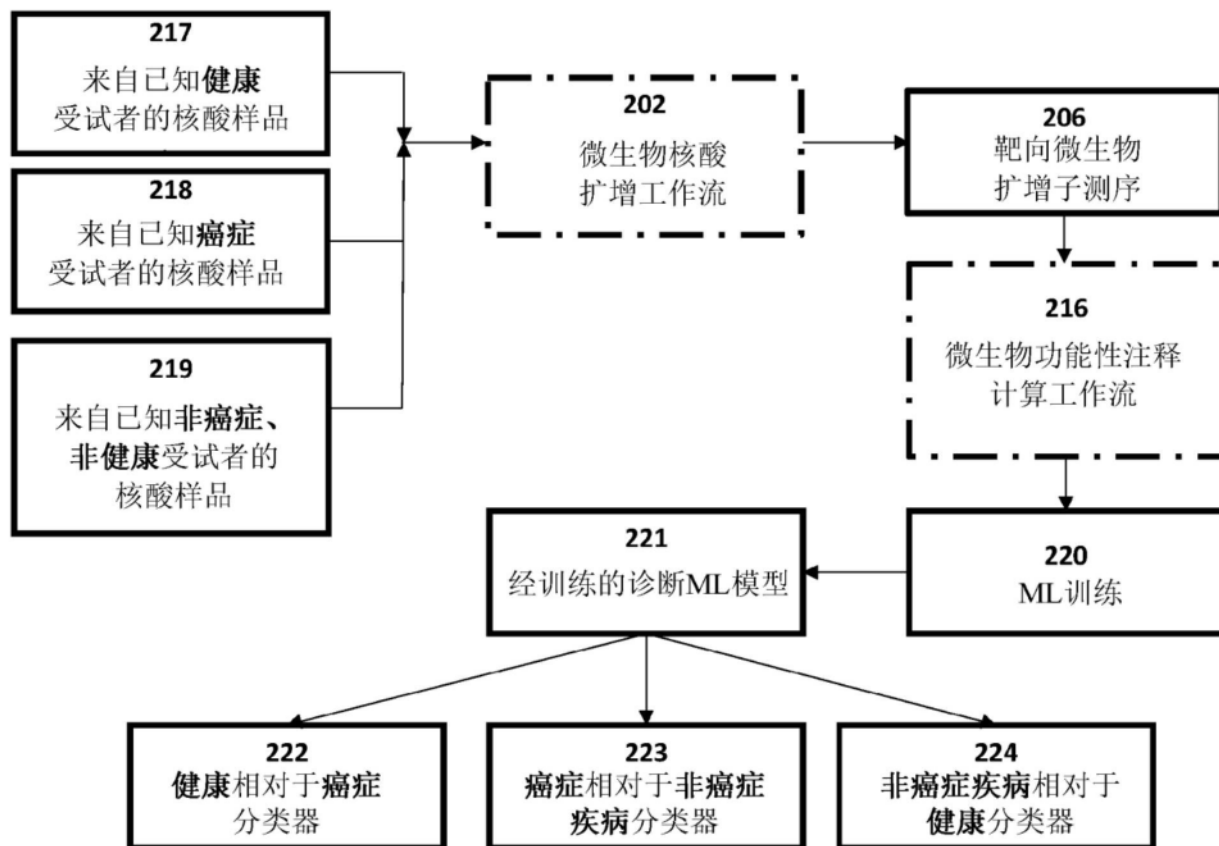


图5

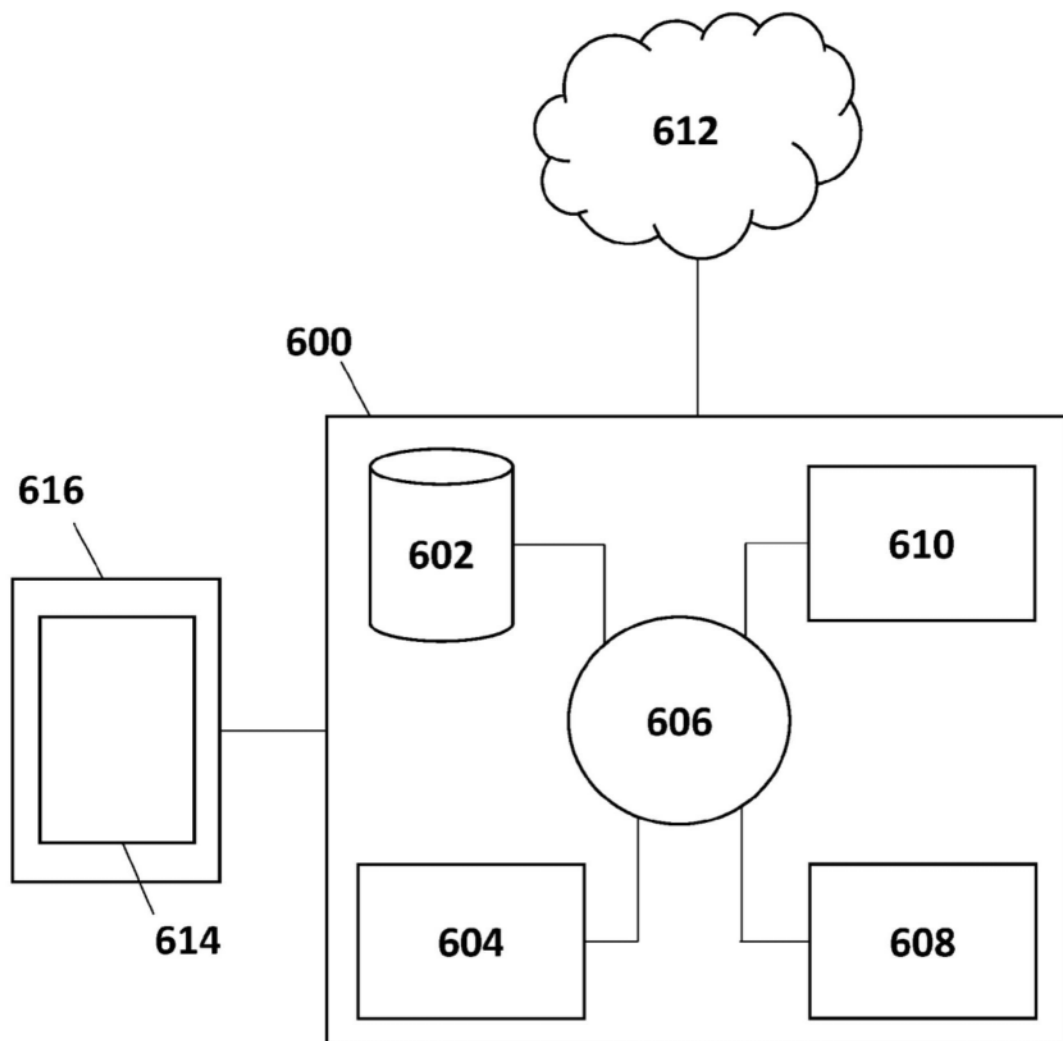


图6

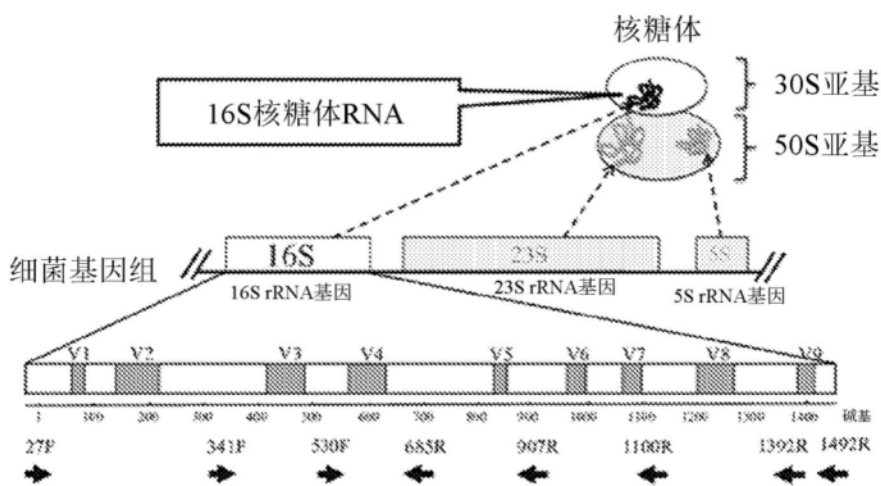


图7A

16S引物名称	序列 (5'-3')
8F	AGA GTT TGA TCC TGG CTC AG
27F	AGA GTT TGA TCM TGG CTC AG
U1492R	GGT TAC CTT GTT ACG ACT T
928F	TAA AAC TYA AAK GAA TTG ACG GG
336R	ACT GCT GCS YCC CGT AGG AGT CT
1100F	YAA CGA GCG CAA CCC
1100R	GGG TTG CGC TCG TTG
337F	GAC TCC TAC GGG AGG CWG CAG
907R	CCG TCA ATT CCT TTR AGT TT
785F	GGA TTA GAT ACC CTG GTA
806R	GGA CTA CVS GGG TAT CTA AT
533F	GTG CCA GCM GCC GCG GTA A
518R	GTA TTA CCG CGG CTG CTG G
1492R	CGG TTA CCT TGT TAC GAC TT
967F	CAA CGC GAA GAA CCT TAC C
1064R	CGA CGA CCA TGC ANC ACC T

图7B

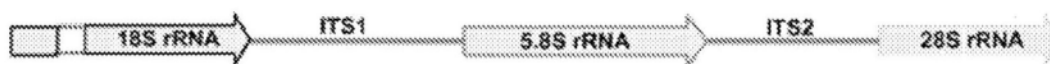


图8

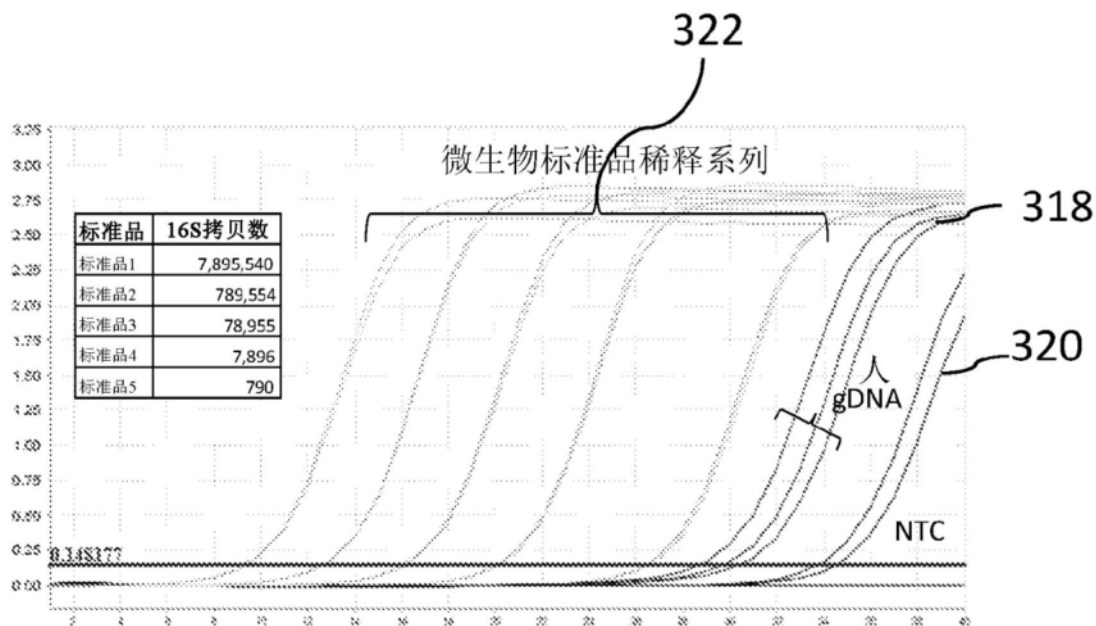


图9

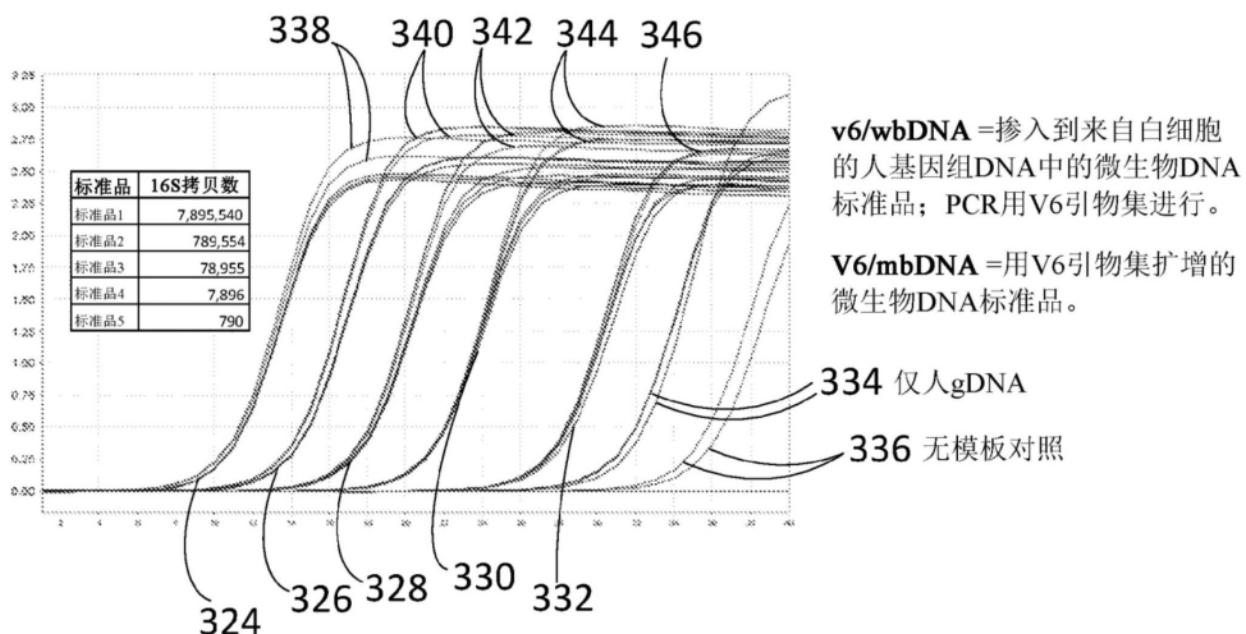


图10

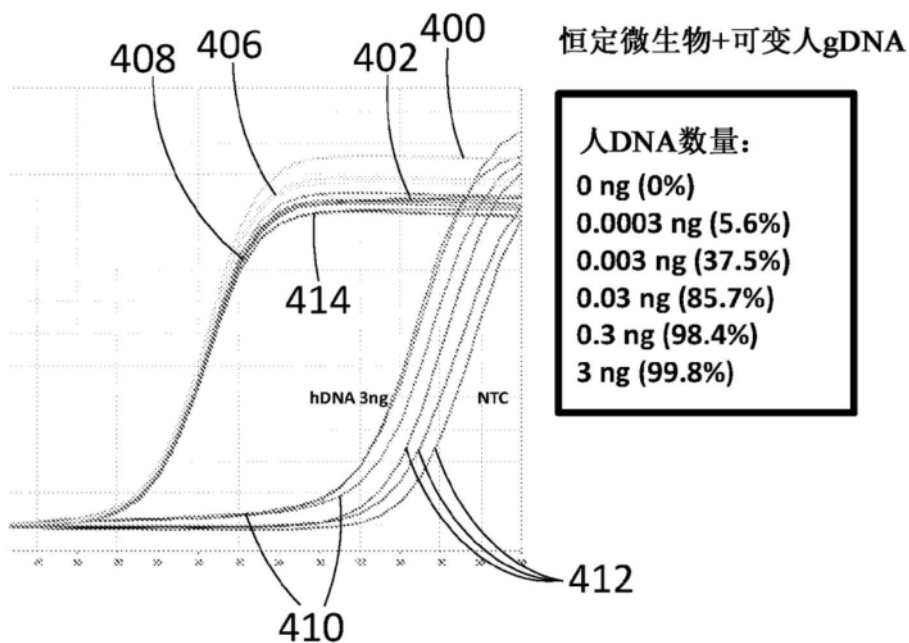


图11

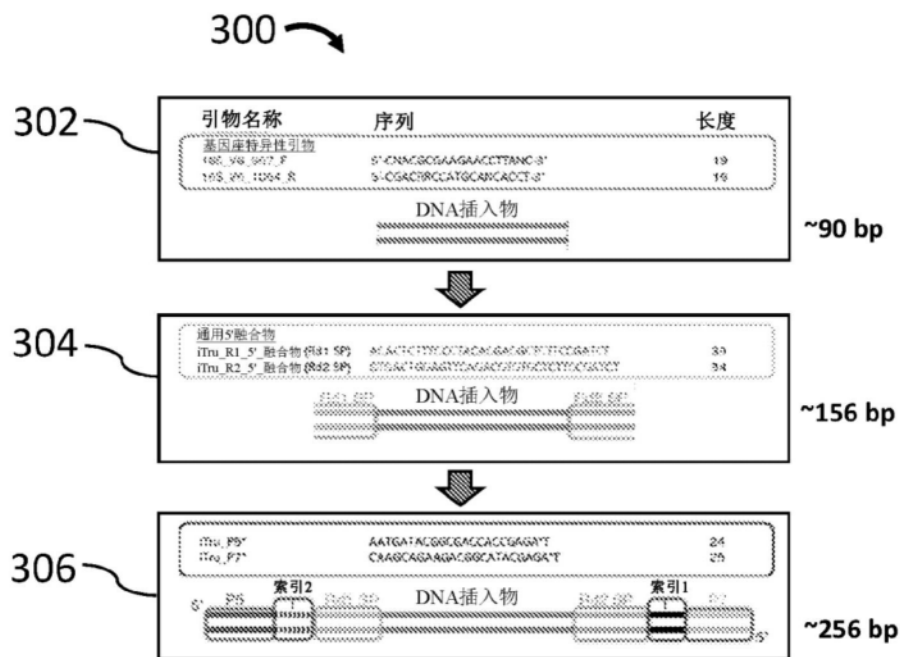


图12A

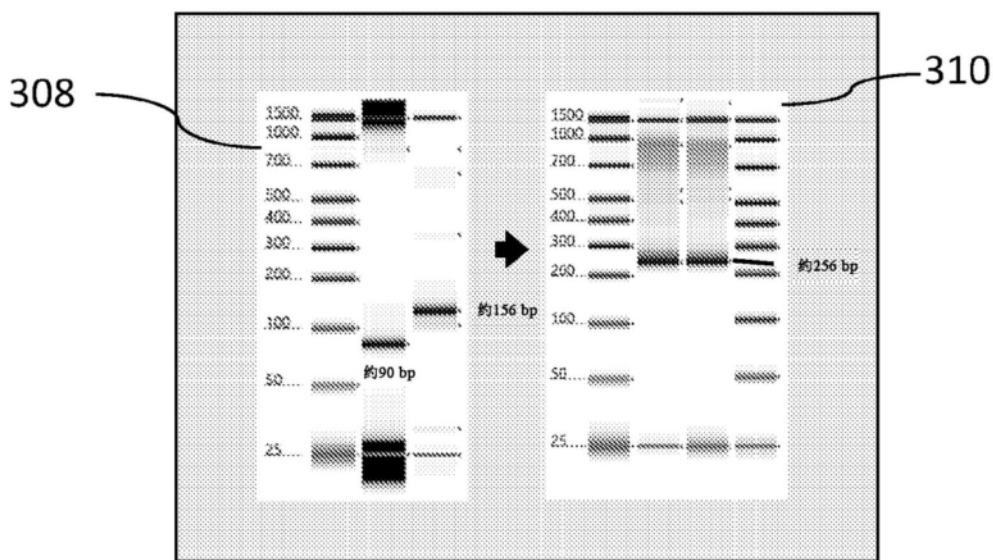


图12B

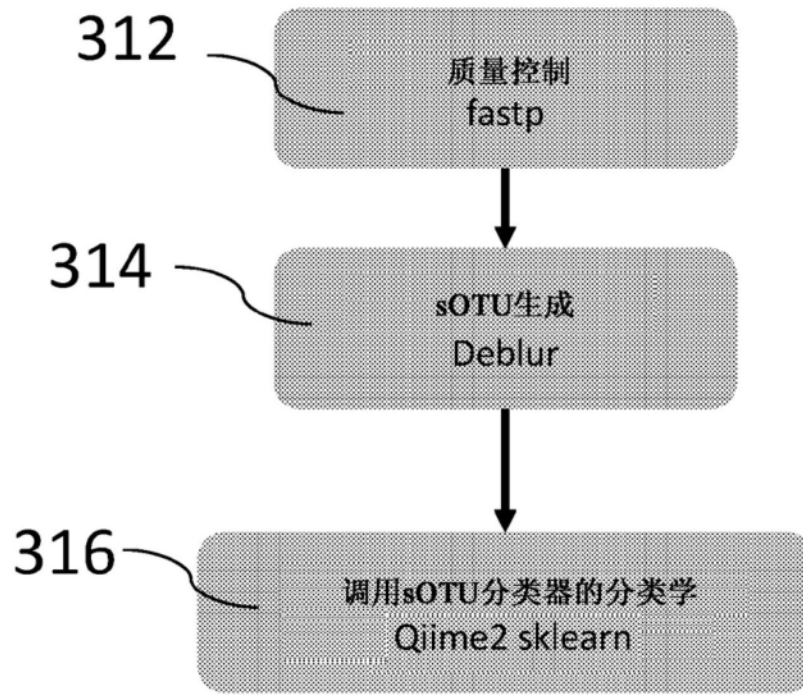


图13

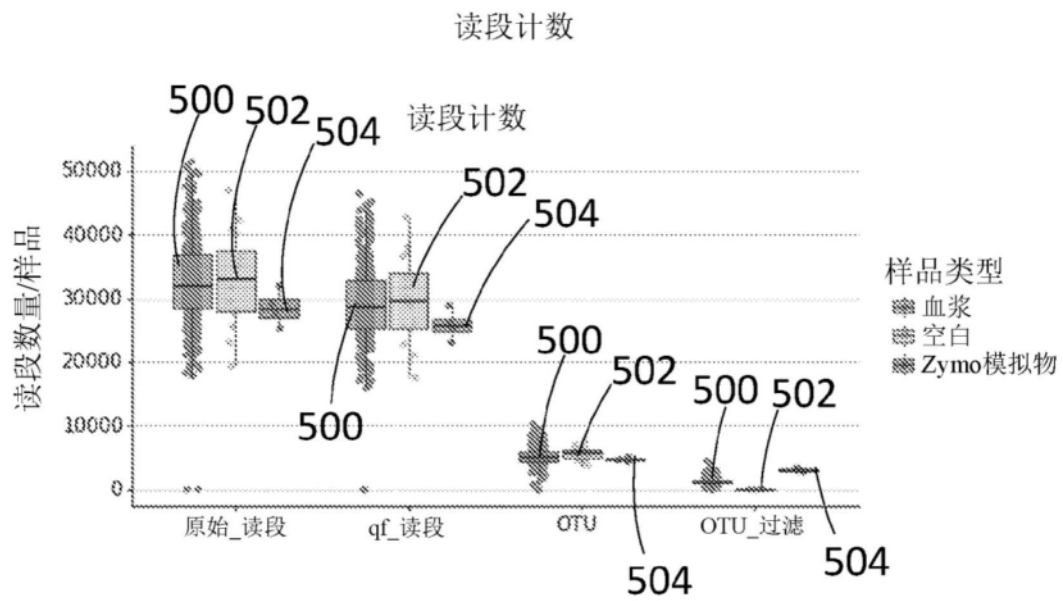


图14A

空白过滤

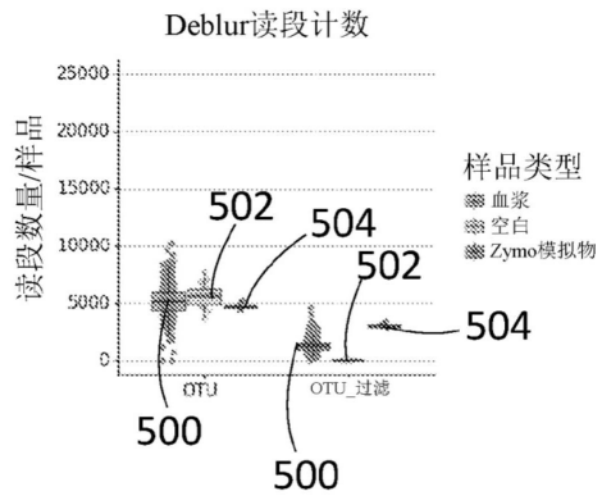


图14B

	诊断	样品数量
非癌症	类癌	50
	错构瘤	20
	肉芽肿	16
	间质纤维化	12
	机化性肺炎	6
	支气管扩张	5
非小细胞肺癌	诊断	
	I期	51
	II期	57
	III期	19

图15A

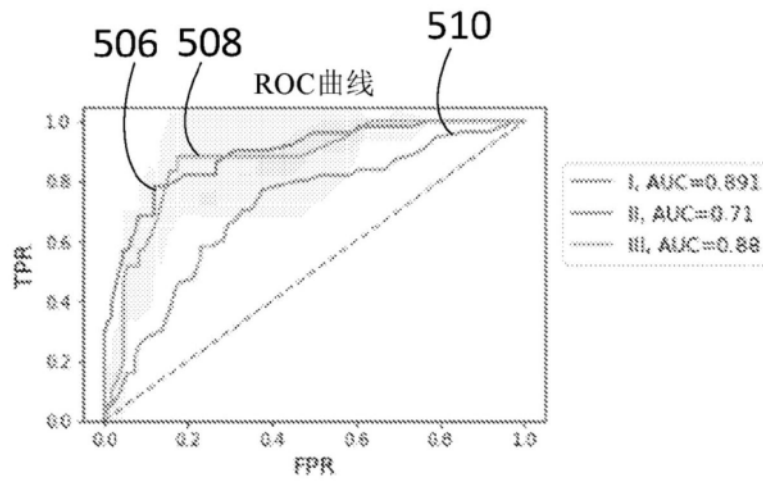


图15B