(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113228194 A
(43)申请公布日 2021.08.06

(21)申请号 201980080958.7

(22)申请日 2019.10.14

(30)优先权数据
62/745,150 2018.10.12 US

(85)PCT国际申请进入国家阶段日
2021.06.07

(86)PCT国际申请的申请数据
PCT/US2019/056166 2019.10.14

(87)PCT国际申请的公布数据
WO2020/077352 EN 2020.04.16

(71)申请人 人类长寿公司
地址 美国加利福尼亚州

(72)发明人 A·哈雷 E·辛布洛特 C·劳

(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅

(51)Int.Cl.
G16B 40/20(2006.01)
G16H 50/20(2006.01)

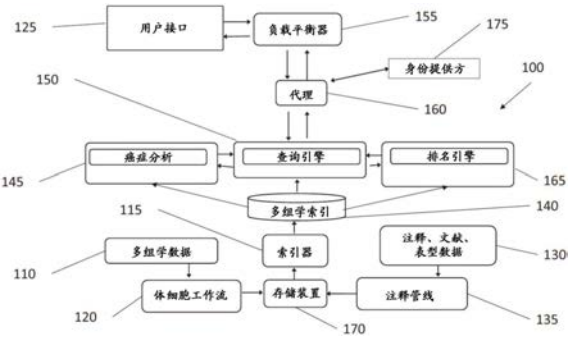
权利要求书8页 说明书37页 附图20页

(54)发明名称

用于癌症基因组和临床数据综合分析的多组学搜索引擎

(57)摘要

提供了利用多组学数据索引来进行肿瘤概况分析的方法。方法可以包括存储多个多组学数据索引,其中多个多组学数据索引中的每一个包括癌症特定的记号化数据;摄取附加的多组学数据以及与附加的多组学数据相关联的任何注释,附加的多组学数据与一个或多个索引有关;对所摄取的附加多组学数据和注释进行索引,同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射,以产生经记号化的所摄取的附加多组学数据;接收用户查询;基于用户查询来选择一个或多个相关多组学数据索引;基于临床可操作性、致病性、特征权重或频率中的至少一项,将所选择的一个或多个多组学数据索引排名;以及向用户返回经排名的一个或多个多组学数据索引。



CN 113228194 A

1.一种用于利用多组学数据索引来进行肿瘤概况分析的方法,所述方法包括:

存储多个多组学数据索引,其中所述多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据;

摄取附加多组学数据以及与所述附加多组学数据相关联的任何注释,所述附加多组学数据与一个或多个索引有关;

对所摄取的所述附加多组学数据和所述注释进行索引,同时在特定的所述索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射,以产生记号化的所摄取的附加多组学数据;

接收用户查询;

基于所述用户查询来选择相关的一个或多个多组学数据索引;

基于临床可操作性、致病性、特征权重或频率中的至少一项,将所选择的所述一个或多个多组学数据索引排名,以及

向所述用户返回经排名的所述一个或多个多组学数据索引。

2.根据权利要求1所述的方法,其中所述多组学数据选自包括以下项的组:基因组的、转录组的、表观遗传的、染色质可及性、微生物组的、蛋白质组的、表型的、图像、相关文献、集成多组学数据、以及前述项的组合。

3.根据权利要求1所述的方法,其中所述多个多组学数据索引还包括体细胞基因组改变、正常基因组改变和癌症注释源。

4.根据权利要求1所述的方法,还包括针对所选择的所述一个或多个多组学数据索引得出癌症分析,其中所述癌症分析包括肿瘤特性,所述肿瘤特性选自包括以下项的组:质量控制、肿瘤突变负荷、基因组突变签名、微卫星不稳定状态、新抗原、HLA等位基因分型、RNA确认的变体、拷贝数变体、结构性变体、非编码调控变体、基因融合、通路富集、癌症驱动因素标识、突变概要、差异基因表达、免疫签名、关于类似患者的治疗结果的匹配信息以及前述项的组合。

5.根据权利要求4所述的方法,其中所述癌症分析针对个体样本或样本同类群组而被得出。

6.根据权利要求4所述的方法,其中所述癌症分析包括机器学习预测和经排名的特征。

7.根据权利要求6所述的方法,其中所述机器学习预测选自包括以下项的组:主要原发部位分类器、未来转移部位预测分类器、微卫星不稳定状态预测、新抗原结合亲和力预测、疾病状态分层、确定癌症谱系、以及前述项的组合。

8.根据权利要求1所述的方法,还包括将注释从基因组层级的更高级别向基因组层级的更低级别传播。

9.根据权利要求1所述的方法,还包括将所选择的所述一个或多个多组学数据索引从基因组层级的更高级别到基因组层级的更低级别排名。

10.根据权利要求1所述的方法,其中所述排名包括针对癌症变体和基因的临床排名和致病性的排名。

11.根据权利要求1所述的方法,其中所述排名包括通过并入针对癌症数据的潜在空间表示来将同类群组分层。

12.根据权利要求11所述的方法,其中所述同类群组被分层为反应者和非反应者。

13.根据权利要求11所述的方法，其中所述同类群组被分层为长期无进展生存时间和短期无进展生存时间。

14.根据权利要求11所述的方法，其中所述同类群组被分层为癌症的不同亚型。

15.根据权利要求11所述的方法，其中所述潜在空间表示由神经网络来执行。

16.根据权利要求11所述的方法，其中所述潜在空间表示由降维技术来执行。

17.根据权利要求16所述的方法，其中所述神经网络选自包括以下项的组：自动编码器、变分自动编码器、深度置信网络、受限玻耳兹曼机、前馈、卷积、递归、门控递归、长短期记忆、残差、以及生成对抗网络。

18.根据权利要求1所述的方法，其中所述排名还包括用于学习排名的模型，所述用于学习排名的模型选自包括以下项的组：支持向量机、提升决策树、回归方法、神经网络、以及前述项的组合。

19.根据权利要求1所述的方法，其中所述排名还包括深度学习排名。

20.根据权利要求19所述的方法，其中所述深度学习排名从深度学习模型中得出，所述深度学习模型选自包括以下项的组：深度语义相似度模型、卷积深度语义相似度模型、递归深度语义相似度模型、深度相关性匹配模型、深度和广度模型、深度语言模型、变换器网络、长短期记忆网络、学习到的深度学习文本嵌入、学习到的命名实体识别、孪生神经网络、交互孪生网络、词法和语义匹配网络、以及前述项的组合。

21.根据权利要求1的方法，其中所述多组学数据选自包括以下项组的组：从全基因组序列数据的体细胞调用、从全外显子组序列数据的体细胞调用、从新鲜冷冻组织的体细胞组套测序、从福尔马林固定石蜡包埋组织的体细胞组套测序、从液体活检的体细胞组套测序、肿瘤和正常变体调用、被索引为在RNA或基因表达水平上被确认为变体的肿瘤/正常转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、单个细胞测序数据、以及前述项的组合。

22.根据权利要求1所述的方法，其中所述多组学数据索引还包括被提取的表型数据。

23.根据权利要求22所述的方法，其中所述表型数据选自包括以下项的组：电子健康记录、临床数据、功能数据、以及前述项的组合。

24.根据权利要求1所述的方法，其中所述多组学数据索引还包括特征化的影像学数据。

25.根据权利要求24所述的方法，其中所述特征化的影像学数据选自包括以下项的组：组织学投影片、MRI图像、X射线、乳房X线照片、超声、PET图像、CT扫描、以及前述项的组合。

26.根据权利要求4所述的方法，其中所述癌症分析在所述用户查询的接收之后被动态计算。

27.根据权利要求1所述的方法，其中所述对所摄取的附加多组学数据和所述注释进行索引还包括对如下得出的数据进行索引，所述得出的数据选自包括以下项的组：癌症分析、注释、从影像学数据提取的特征、表型、医学文献数据及其嵌入、以及前述各项的组合。

28.根据权利要求1所述的方法，其中所述排名还包括将样本改变与已建立的药物靶标标签和可用的临床试验相匹配。

29.根据权利要求1所述的方法，其中所述排名还包括通过检测基于感兴趣的临床变量和/或统计显著性将同类群组分层的潜在生物标志物来进行所述同类群组中的癌症药物靶

标的标识，并且其中向所述用户返回经排名的所述一个或多个多组学数据索引包括分层可视化。

30.根据权利要求1所述的方法，其中所述向所述用户返回经排名的所述一个或多个多组学数据索引还包括针对个体患者和/或同类群组来动态创建超链接报告，所述超链接报告提供肿瘤的全面概况分析。

31.根据权利要求1所述的方法，其中所述用户查询能够包括用户上传的数据，所述用户上传的数据选自包括以下项的组：一组套的变体、基因、通路、疾病状态病症、感兴趣的表型，并且其中所述选择包括查询由所述上传的数据子选择的个体样本或同类群组数据。

32.根据权利要求1所述的方法，其中所述用户查询能够经由用户接口而被提供并且能够包括上传用于进行索引的数据，所述用于进行索引的数据选自包括以下项的组：基因组数据、转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、表型数据、注释数据、以及前述项的组合。

33.根据权利要求1所述的方法，还包括：规范化和/或扩大所述用户查询、将所述查询的意图分类、汇总检索到的文档、以及使用深度学习方法基于所述查询与潜在空间中的文档之间的相似度来执行文档检索。

34.根据权利要求1所述的方法，其中所述进行索引、所述选择和所述排名中的至少一项包括利用深度神经网络。

35.根据权利要求4所述的方法，其中得出所述癌症分析包括利用深度神经网络。

36.根据权利要求1所述的方法，其中所述向所述用户返回经排名的所述一个或多个多组学数据索引还包括：返回所返回的结果的概要可视化以及经排名的所述结果的列表。

37.一种其中存储有程序的非瞬态计算机可读介质，所述程序用于使计算机执行用于利用多组学数据索引来进行肿瘤概况分析的方法，所述方法包括：

存储多个多组学数据索引，其中所述多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据；

摄取附加多组学数据以及与所述附加多组学数据相关联的任何注释，所述附加多组学数据与一个或多个索引有关；

对所摄取的所述附加多组学数据和所述注释进行索引，同时在特定的所述索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生记号化的所摄取的附加多组学数据；

接收用户查询；

基于所述用户查询来选择相关的一个或多个多组学数据索引；

基于临床可操作性中的至少一项，将所选择的所述一个或多个多组学数据索引排名，以及

向所述用户返回经排名的所述一个或多个多组学数据索引。

38.根据权利要求37所述的方法，其中所述多组学数据选自包括以下项的组：基因组的、转录组的、表观遗传的、染色质可及性、微生物组的、蛋白质组的、表型的、图像、相关文献、集成多组学数据、以及前述项的组合。

39.根据权利要求37所述的方法，其中所述多个多组学数据索引还包括体细胞基因组改变、正常基因组改变和癌症注释源。

40.根据权利要求37所述的方法，还包括针对所选择的所述一个或多个多组学数据索引得出癌症分析，其中所述癌症分析包括癌症特性，所述癌症特性选自包括以下项的组：质量控制、肿瘤突变负荷、基因组突变签名、微卫星不稳定状态、新抗原、HLA等位基因分型、RNA确认的变体、拷贝数变体、结构性变体、非编码调控变体、基因融合、通路富集、癌症驱动因素标识、突变概要、差异基因表达、免疫签名、关于类似患者的治疗结果的匹配信息以及前述项的组合。

41.根据权利要求40所述的方法，其中所述癌症分析针对个体样本或样本同类群组而被得出。

42.根据权利要求40所述的方法，其中所述癌症分析包括机器学习预测和经排名的特征。

43.根据权利要求42所述的方法，其中所述机器学习预测选自包括以下项的组：主要原发部位分类器、未来转移部位预测分类器、微卫星不稳定状态预测、新抗原结合亲和力预测、疾病状态分层、确定癌症谱系、以及前述项的组合。

44.根据权利要求37所述的方法，还包括将注释从基因组层级的更高级别向基因组层级的更低级别传播。

45.根据权利要求37所述的方法，还包括将所选择的所述一个或多个多组学数据索引从基因组层级的更高级别到基因组层级的更低级别排名。

46.根据权利要求37所述的方法，其中所述排名包括针对癌症变体和基因的临床排名。

47.根据权利要求3375所述的方法，其中所述排名包括通过并入针对癌症数据的潜在空间表示来将同类群组分层。

48.根据权利要求47所述的方法，其中所述同类群组被分层为反应者和非反应者。

49.根据权利要求47所述的方法，其中所述同类群组被分层为长期无进展生存时间和短期无进展生存时间。

50.根据权利要求47所述的方法，其中所述潜在空间表示由神经网络来执行。

51.根据权利要求50所述的方法，其中所述神经网络选自包括以下项的组：自动编码器、变分自动编码器、深度置信网络、受限玻耳兹曼机、前馈网络、卷积网络、递归网络、长短期记忆网络、以及生成对抗网络。

52.根据权利要求37所述的方法，其中所述排名还包括用于学习排名的模型，所述用于学习排名的模型选自包括以下项的组：支持向量机、提升决策树、回归模型、神经网络、以及前述项的组合。

53.根据权利要求37所述的方法，其中所述排名还包括深度学习排名。

54.根据权利要求53所述的方法，其中所述深度学习排名从深度学习模型被得出，所述深度学习模型选自包括以下项的组：深度语义相似度模型、深度和广度模型、深度语言模型、学习到的深度学习文本嵌入、学习到的命名实体识别、孪生神经网络、以及前述项的组合。

55.根据权利要求37所述的方法，其中所述多组学数据选自包括以下项组的组：从全基因组序列数据的体细胞调用、从全外显子组序列数据的体细胞调用、从新鲜冷冻组织的体细胞组套测序、从福尔马林固定石蜡包埋组织的体细胞组套测序、从液体活检的体细胞组套测序、肿瘤和正常变体调用、被索引为在RNA或基因表达水平上被确认为变体的肿瘤/正

常转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、单个细胞测序数据、以及前述项的组合。

56.根据权利要求37所述的方法，其中所述多组学数据索引还包括被提取的表型数据。

57.根据权利要求56所述的方法，其中所述表型数据选自包括以下项的组：电子健康记录、临床数据、功能数据、以及前述项的组合。

58.根据权利要求37所述的方法，其中所述多组学数据索引还包括特征化的影像学数据。

59.根据权利要求58所述的方法，其中所述特征化的影像学数据选自包括以下项的组：组织学投影片、MRI图像、X射线、乳房X线照片、超声、PET图像、CT扫描、以及前述项的组合。

60.根据权利要求40所述的方法，其中所述癌症分析在所述用户查询的接收之后被动态计算。

61.根据权利要求40所述的方法，其中所述对所摄取的附加多组学数据和所述注释进行索引还包括对如下得出的数据进行索引，所述得出的数据选自包括以下项的组：癌症分析、注释、从影像学数据提取的特征、表型、医学文献数据及其嵌入、以及前述各项的组合。

62.根据权利要求37所述的方法，其中所述排名还包括将样本改变与已建立的药物靶标标签和可用的临床试验相匹配。

63.根据权利要求37所述的方法，其中所述排名还包括通过检测基于感兴趣的临床变量和/或统计显著性将同类群组分层的潜在生物标志物来进行所述同类群组中的癌症药物靶标的标识，并且其中向所述用户返回经排名的所述一个或多个多组学数据索引包括分层可视化。

64.根据权利要求37所述的方法，其中所述向所述用户返回经排名的所述一个或多个多组学数据索引还包括针对个体患者和/或同类群组来动态创建超链接报告，所述超链接报告提供肿瘤的全面概况分析。

65.根据权利要求37所述的方法，其中所述用户查询能够包括用户上传的数据，所述用户上传的数据选自包括以下项的组：一组套的变体、基因、通路、疾病状态病症、感兴趣的表型，并且其中所述选择包括查询由所述上传的数据子选择的个体样本或同类群组数据。

66.根据权利要求37所述的方法，其中所述用户查询能够经由用户接口而被提供并且能够包括上传用于进行索引的数据，所述用于进行索引的数据选自包括以下项的组：基因组数据、转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、表型数据、注释数据、以及前述项的组合。

67.根据权利要求37所述的方法，还包括：规范化和/或扩大所述用户查询、将所述查询的意图分类、汇总检索到的文档、以及使用深度学习方法基于所述查询与潜在空间中的文档之间的相似度来执行文档检索。

68.根据权利要求37所述的方法，其中所述进行索引、所述选择和所述排名中的至少一个包括利用深度神经网络。

69.根据权利要求40所述的方法，其中得出所述癌症分析包括利用深度神经网络。

70.根据权利要求37所述的方法，其中所述向所述用户返回经排名的所述一个或多个多组学数据索引还包括：返回所返回的结果的概要可视化以及经排名的所述结果的列表。

71.一种用于利用多组学数据索引来进行肿瘤概况分析的系统，所述系统包括：

索引单元，包括：

存储元件，被配置为存储多个多组学数据索引，其中所述多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据，以及

索引引擎，被配置为

摄取附加多组学数据以及与所述附加多组学数据相关联的任何注释，所述附加多组学数据与一个或多个索引相关，以及

对所摄取的所述附加多组学数据和注释进行索引，同时在特定的所述索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生记号化的所摄取的附加多组学数据；

用户接口，被配置为接收用户查询；

查询引擎，被配置为基于所述用户查询来从所述索引单元选择相关的一个或多个多组学数据索引；以及

排名引擎，被配置为接收所选择的相关的所述一个或多个多组学数据索引，基于临床可操作性、致病性、特征权重或频率中的至少一项来将所选择的所述一个或多个多组学数据索引排名，并且经由所述用户接口，向所述用户返回经排名的所述一个或多个多组学数据索引。

72.根据权利要求71所述的系统，其中所述多组学数据选自包括以下项的组：基因组的、转录组的、表观遗传的、染色质可及性、微生物组的、蛋白质组的、表型的、图像、相关文献、集成多组学数据、以及前述项的组合。

73.根据权利要求71所述的系统，其中所述多个多组学数据索引还包括体细胞基因组改变、正常基因组改变和癌症注释源。

74.根据权利要求71所述的系统还包括癌症分析引擎，所述癌症分析引擎被配置为针对所选择的所述一个或多个多组学数据索引得出癌症分析，其中所述癌症分析包括癌症特性，所述癌症特性选自包括以下项的组：质量控制、肿瘤突变负荷、基因组突变签名、微卫星不稳定状态、新抗原、HLA等位基因分型、RNA确认的变体、拷贝数变体、结构性变体、非编码调控变体、基因融合、通路富集、癌症驱动因素标识、突变概要、差异基因表达、免疫签名、关于类似患者的治疗结果的匹配信息及前述项的组合。

75.根据权利要求74所述的系统，其中所述癌症分析针对个体样本或样本同类群组而被得出。

76.根据权利要求74所述的系统，其中所述癌症分析包括机器学习预测和经排名的特征。

77.根据权利要求76所述的系统，其中所述机器学习预测选自包括以下项的组：主要原发部位分类器、未来转移部位预测分类器、微卫星不稳定状态预测、新抗原结合亲和力预测、疾病状态分层、确定癌症谱系、以及前述项的组合。

78.根据权利要求71所述的系统，其中所述索引引擎被配置为将注释从基因组层级的更高级别向基因组层级的更低级别传播。

79.根据权利要求71所述的系统，其中所述排名引擎被配置为将所选择的所述一个或多个多组学数据索引从基因组层级的更高级别到基因组层级的更低级别排名。

80.根据权利要求71所述的系统，其中所述排名包括针对癌症变体和基因的临床排名。

81.根据权利要求71所述的系统，其中所述排名包括通过并入针对癌症数据的潜在空间表示来将同类群组分层。

82.根据权利要求81所述的系统，其中所述同类群组被分层为反应者和非反应者。

83.根据权利要求81所述的系统，其中所述同类群组被分层为长期无进展生存时间和短期无进展生存时间。

84.根据权利要求79所述的系统，其中所述同类群组被分层为不同的癌症亚型。

85.根据权利要求81所述的系统，其中所述潜在空间表示由神经网络来执行。

86.根据权利要求85所述的系统，其中所述神经网络选自包括以下项的组：自动编码器、变分自动编码器、深度置信网络、受限玻耳兹曼机、前馈、卷积、递归、门控递归、长短期记忆、残差、以及生成对抗网络。

87.根据权利要求71所述的系统，其中所述排名引擎还包括用于学习排名的模型，所述用于学习排名的模型选自包括以下项的组：支持向量机、提升决策树、回归模型、神经网络、以及前述项的组合。

88.根据权利要求71所述的系统，其中所述排名还包括深度学习排名。

89.根据权利要求88所述的系统，其中所述深度学习排名从选自包括以下项的组的深度学习模型被得出：深度语义相似度模型、深度和广度模型、深度语言模型、学习到的深度学习文本嵌入、学习到的命名实体识别、孪生神经网络、以及前述项的组合。

90.根据权利要求71所述的系统，其中所述多组学数据选自包括以下项组的组：从全基因组序列数据的体细胞调用、从全外显子组序列数据的体细胞调用、从新鲜冷冻组织的体细胞组套测序、从福尔马林固定石蜡包埋组织的体细胞组套测序、从液体活检的体细胞组套测序、肿瘤和正常变体调用、被索引为在RNA或基因表达水平上被确认为变体的肿瘤/正常转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、单个细胞测序数据、以及前述项的组合。

91.根据权利要求71所述的系统，其中所述多组学数据索引还包括被提取的表型数据。

92.根据权利要求91所述的系统，其中所述表型数据选自包括以下项的组：电子健康记录、临床数据、功能数据、以及前述项的组合。

93.根据权利要求71所述的系统，其中所述多组学数据索引还包括特征化的影像学数据。

94.根据权利要求93所述的系统，其中所述特征化的影像学数据选自包括以下项的组：组织学投影片、MRI图像、X射线、乳房X线照片、超声、PET图像、CT扫描、以及前述项的组合。

95.根据权利要求74所述的系统，其中所述癌症分析在所述用户查询的接收之后被动态计算。

96.根据权利要求71所述的系统，其中所述索引引擎还被配置对如下得出的数据进行索引，所述得出的数据选自包括以下项的组：癌症分析、注释、从影像学数据所提取的特征、表型、医学文献数据及其嵌入、以及前述项的组合。

97.根据权利要求71所述的系统，其中所述排名引擎还被配置为将样本改变与已建立的药物靶标标签和可用的临床试验相匹配。

98.根据权利要求71所述的系统，其中所述排名引擎还被配置为通过检测基于感兴趣的临床变量和/或统计显著性将同类群组分层的潜在生物标志物来进行所述同类群组中的

癌症药物靶标的标识，并且所述排名引擎还被配置为经由分层可视化向所述用户返回经排名的所述一个或多个多组学数据索引。

99.根据权利要求71所述的系统，其中所述排名引擎被配置为经由针对个体患者和/或同类群组动态创建超链接报告来向所述用户返回将经排名的所述一个或多个多组学数据索，所述超链接报告提供肿瘤的全面的概况分析。

100.根据权利要求71所述的系统，其中所述用户查询包括用户上传的数据，所述用户上传的数据选自包括以下项的组：一组套的变体、基因、通路、疾病状态病症、感兴趣的表型，并且其中所述选择包括查询由所述所上传的数据子选择的个体样本或同类群组数据。

101.根据权利要求71所述的系统，其中所述用户接口被配置为接收用户查询，所述用户查询包括用于进行索引的被上传的数据，所述数据选自包括以下项的组：基因组数据、转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、表型数据、注释数据、以及前述项的组合。

102.根据权利要求71所述的系统，其中所述查询引擎还被配置为规范化和/或扩大所述用户查询、将所述查询的意图分类、汇总检索到的文档、以及使用深度学习方法基于所述查询与潜在空间中的文档之间的相似度来执行文档检索。

103.根据权利要求71所述的系统，其中所述索引引擎、所述查询引擎和所述排名引擎中的至少一个被配置为利用深度神经网络。

104.根据权利要求74所述的系统，其中所述癌症分析引擎被配置为利用深度神经网络来得出所述癌症分析。

105.根据权利要求71所述的系统，其中所述排名引擎还被配置为：还通过返回所返回的结果的概要可视化以及经排名的结果列表，来向所述用户返回经排名的所述一个或多个多组学数据索引。

106.一种用于利用多组学数据索引来进行肿瘤概况分析的系统，所述系统包括：

索引单元，包括：

存储元件，被配置为存储多个多组学数据索引，其中所述多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据，以及

索引引擎，被配置为

摄取附加多组学数据以及与所述附加多组学数据相关联的任何注释，所述附加多组学数据与一个或多个索引相关，以及

对所摄取的所述附加多组学数据和注释进行索引，同时在所述特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生记号化的所摄取的附加多组学数据；

用户接口，被配置为接收用户查询；以及

查询引擎，被配置为基于所述用户查询来从所述索引单元选择相关的一个或多个多组学数据索引，以基于临床可操作性、致病性、特征权重或频率中的至少一项来将所选择的所述一个或多个多组学数据索引排名，并且经由所述用户接口，向所述用户返回经排名的所述一个或多个多组学数据索引。

用于癌症基因组和临床数据综合分析的多组学搜索引擎

背景技术

[0001] 随着癌症基因组测序的重要性日益提高，数以千计的癌症基因组、外显子组、转录组、蛋白质组和其他癌症数据已经由私人和公共机构(例如，癌症基因组图谱[TCGA]、国际癌症基因组联盟[ICGC])两者进行测序。肿瘤和正常测序数据的解释和分析取决于对私人和公共基因组数据和数据库两者的综合分析相关。

[0002] 工业界、生物制药公司、研究机构和国际癌症协会面临障碍，诸如例如(1)提供对任何样本或样本子集的立即访问；(2)集成多组学数据集来形成肿瘤生物学的完整图景；(3)将预后、诊断和治疗信息与所有可用数据(例如，基因组的、转录组的、蛋白质组的、功能的、医学的、影像学的、文献数据)有效地相关联，以提供针对个体癌症患者的临床见解和可操作性、以及根据潜在的多组学预后、诊断或(多个)治疗生物标志物的患者同类群组(cohort)分层。

[0003] 目前，公开可用的数据分散在出版物、指南和基于web的资源中。最终，解决上述三个问题的解决方案将使得癌症基因组分析广泛应用于临床。

[0004] 数据集成和整合在癌症测序中提出了特别严峻的挑战，即，允许用户合并数据的多个源并且标识临床和生物学相关信息的标准化和集成。附加地，与种系序列分析相比较，癌症的基因组分析需要广泛的生物信息学管线，并且需要针对同一样本产生数据的多组学流。例如，对于典型的癌症活检和血常规，针对肿瘤DNA、正常DNA、肿瘤RNA、有时正常RNA的二进制碱基调用(BCL)必须经由与参考基因组对准、去重复、重新对准和变体重新校准而被转换为变体调用格式(VCF)。此外，运行多个体细胞变体调用器来推导一组共识的体细胞单核苷酸变体(SNV)和小的插入和缺失(indels)通常是行业标准。例如，进一步令人感兴趣的是肿瘤的拷贝数变体(CNV)的检测、肿瘤与正常RNA-Seq复制之间的差异基因表达、用于确认在体细胞(肿瘤)DNA中检测到的变体也在RNA中表达的数据处理、以及检测基因融合的管线。进一步令人感兴趣的是使用调用大型结构性变体的工具以及执行高级生物信息学的工具来注释癌症改变并且计算肿瘤的相关性质(例如，肿瘤突变负荷、基因组突变签名、微卫星状态、被表达的新抗原、正常基因组的HLA分型)，以及标识与临床相关的肿瘤改变。

[0005] 现代癌症剖析技术可以容易地每样本生成25吉字节的多组学数据，这意味着进行中等大小的癌症生物标志物发现研究的研究人员容易面临太字节的原始数据。因此，标识相关的生物标志物类似于“大海捞针”。而且，一旦分析管线完成运行，实际上就无法与结果交互来形成新的假设。

[0006] 解决当前癌症数据的可访问性、多集成性和可操作性问题的最常用方法是设计门户来显示经预先过滤的数据表以及基于先前所编策的文件和被预先计算的工作流程的分析。门户的示例包括Illumina BaseSpace Correlation Engine and Cohort Analyzer、WuXI nextCODE TCGA门户、cBioPortal、IntOGen、Tumorscape、Tumorportal、Xena、ICGC Data Portal、St.Jude PeCan和Qiagen OmicSoft。但是，这些门户通常限制了可以被解决的问题类型以及可以被执行的附加分析。此外，在生物信息学管线的许多级别处，数据通常不可访问以供查询。门户中的数据通常被预先过滤，未被集成并且通常未被排名。附加地，

大多数门户不托管个体用户数据。允许用户上传用户自己的数据的少数几个门户通常不提供将用户的数据与门户数据集成、或者推导高级癌症分析并且使得这些数据可访问并且按临床可操作性、致病性、特征权重或频率而被排名的方法。

[0007] 因此，需要提供有效并且有效率地提供对任何样本或样本子集的即时访问的系统和方法。还需要提供有效并且由效率集成多组学数据集来形成肿瘤生物学的完整图景的系统和方法。还需要提供将预后、诊断和治疗信息与所有可用数据(例如，基因组的、转录组的、蛋白质组的、功能的、医学的、影像学的、文献数据)有效并且有效率地相关联来提供针对个体癌症患者的临床见解和可操作性、并且根据潜在的多组学预后或者(多个)治疗生物标志物将患者的同类群组分层的系统和方法。

发明内容

[0008] 概况分析。一种方法可以包括存储多个多组学数据索引，其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化(tokenized)数据。该方法还可以包括摄取附加多组学数据以及与附加多组学数据相关联的任何注释，附加多组学数据与一个或多个索引有关。该方法还可以包括对所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。该方法还可以包括接收用户查询。该方法还可以包括基于用户查询来选择相关的一个或多个多组学数据索引。该方法还可以包括基于临床可操作性、致病性、特征权重或频率中的至少一项，将所选择的一个或多个多组学数据索引排名。该方法还可以包括向用户返回经排名的一个或多个多组学数据索引。

[0009] 根据各种实施例，提供了一种其中存储有程序的非瞬态计算机可读介质，该程序用于使计算机执行用于利用多组学数据索引以供肿瘤概况分析(tumor profiling)的方法。该方法可以包括存储多个多组学数据索引，其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据。该方法还可以包括摄取附加多组学数据以及与附加的多组学数据相关联的注释，附加多组学数据与一个或多个索引有关。方法还可以包括对所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。该方法还可以包括接收用户查询。该方法还可以包括基于用户查询来选择相关的一个或多个多组学数据索引。方法还可以包括基于临床可操作性、致病性、特征权重或频率中的至少一项，将所选择的一个或多个多组学数据索引排名。该方法还可以包括向用户返回经排名的一个或多个多组学数据索引。

[0010] 根据各种实施例，提供了用于利用多组学数据索引进行肿瘤概况分析的系统。该系统可以包括索引单元。索引单元可以包括存储元件，存储元件被配置为存储多个多组学数据索引，其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据。索引单元还可以包括索引引擎。索引单元可以被配置为摄取附加多组学数据以及与附加多组学数据相关联的注释，附加多组学数据与一个或多个索引有关。索引单元还可以被配置为对所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。该系统还可以包括被配置为接收用户查询的用户接口。该系统还可以包

括查询引擎，查询引擎被配置为基于用户查询来从索引单元中选择相关的一个或多个多组学数据索引。该系统还可以包括排名引擎，排名引擎被配置为接收所选择的相关的一个或多个多组学数据索引，并且基于临床可操作性、致病性、特征权重、或频率中的至少一项，将所选择的一个或多个多组学数据索引排名。排名引擎还可以被配置为经由用户接口向用户返回经排名的一个或多个多组学数据索引。

[0011] 根据各种实施例，提供了用于利用多组学数据索引进行肿瘤概况分析的系统。该系统可以包括索引单元。索引单元可以包括存储元件，存储元件被配置为存储多个多组学数据索引，其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据。索引单元还可以包括索引引擎。索引单元可以被配置为摄取附加多组学数据以及与附加多组学数据相关联的注释，附加的多组学数据与一个或多个索引有关。索引单元还可以被配置为对所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。该系统还可以包括被配置为接收用户查询的用户接口。系统还可以包括查询引擎，查询引擎被配置为基于用户查询来从索引单元中选择相关的一个或多个多组学数据索引。查询引擎还可以被配置为基于临床可操作性、致病性、特征权重或频率中的至少一项来将所选择的一个或多个多组学数据索引排名。查询引擎还可以被配置为经由用户接口，向用户返回经排名的一个或多个多组学数据索引。

[0012] 根据各种实施例，提供了用于肿瘤概况分析的多组学癌症搜索引擎系统。该系统可以包括：存储元件，被配置为存储多个经集成的多组学索引；高级癌症分析软件模块；多组学索引管线；反映多组学癌症改变的临床效用的排名引擎；查询引擎，该查询引擎选择并且组合相关的多组学索引并且返回针对个体样本和样本的同类群组的经排名的多组学改变；以及用户接口，被配置为接收用户查询并且执行对癌症数据的搜索。

[0013] 根据以下的具体实施方式以及所附权利要求书和附图，附加方面将变得明显。

附图说明

[0014] 各个方面和实现的上述例示性示例提供了概述或框架，用于理解所要求保护的各方面和实现的性质和特征：

[0015] 图1图示了根据各种实施例的用于多组学癌症搜索引擎的系统架构的示例。

[0016] 图2a图示了根据各种实施例的多组学索引组织的示例。图2b图示了根据各种实施例的注释的分层传播和变体的排名的示例。

[0017] 图3图示了根据各种实施例的针对个体样本和同类群组而动态地被预计算和计算的一组癌症分析的示例。

[0018] 图4a图示了根据各种实施例的用于学习变体排名的广度和深度模型的示例。图4b图示了根据各种实施例的依赖于针对生物医学数据的深度语义相似度模型(DSSM)的学习排名引擎的示例。

[0019] 图5a和图5b一起图示了根据各种实施例的针对查询引擎的操作的工作流的示例。

[0020] 图6图示了根据各种实施例的用户接口的示例。如图所示，例如，单个搜索框允许用户录入不同的查询并且接收经排名的结果。

[0021] 图7图示了根据各种实施例的利用特定语法所获得的搜索结果的示例。

[0022] 图8a和图8b图示了根据各种实施例的利用特定语法所获得的搜索结果的示例。

[0023] 图9图示了根据各种实施例的从用户查询所返回的搜索结果的示例。

[0024] 图10图示了根据各种实施例的从用户查询所返回的搜索结果的示例。

[0025] 图11图示了根据各种实施例的从用户查询所返回的搜索结果的示例。

[0026] 图12图示了根据各种实施例的从用户查询所返回的搜索结果的示例。

[0027] 图13图示了根据各种实施例的计算机系统的框图。

[0028] 图14图示了根据各种实施例的用于利用多组学数据索引进行肿瘤概况分析的方法的流程图。

[0029] 图15图示了根据各种实施例的用于利用多组学数据索引进行肿瘤概况分析的系统。

[0030] 图16图示了根据各种实施例的用于利用多组学数据索引进行肿瘤概况分析的系统。

[0031] 应理解，附图不一定按比例绘制，附图中的对象也不必以其彼此之间的关系而按比例绘制。附图是为了使得本文所公开的装置、系统和方法的各种实施例更清楚且易于理解。在所有附图中，将尽可能使用相同的附图标记指代相同或相似的部分。此外，应当理解，附图无意以任何方式限制本教导的范围。

具体实施方式

[0032] 本说明书描述了用于癌症基因组和临床数据的综合分析的多组学搜索引擎的各种示例性实施例，以及与之相关联的系统和方法。然而，本公开不限于这些示例性实施例和应用，也不限于本文中示例性实施例和应用操作或者被描述的方式。

[0033] 除非另有定义，否则本文所使用的所有技术术语具有与本文所公开的实施例所属领域的普通技术人员通常所理解的含义相同的含义。除非上下文另外明确指出，否则如本说明书和所附权利要求书中所使用的，单数形式的“一(a)”、“一个(an)”和“所述(the)”包括复数引用。除非另有说明，否则本文中对“或”的任何引用旨在涵盖“和/或”。

[0034] 本公开描述了用于操作多组学搜索引擎来对癌症基因组和临床数据进行综合分析的系统和方法，并且在本文中可以由简写“癌症搜索”(Cancer Search或者cancer search)来指代。

[0035] 除非另有定义，否则与本文所述的本教导结合使用的科学术语和技术术语应具有本领域普通技术人员通常所理解的含义。此外，除非上下文另外要求，否则单数术语应包括复数，并且复数术语应包括单数。通常，本文所描述的与细胞和组织培养、分子生物学、以及蛋白质和寡核苷酸或多核苷酸化学和杂交结合所利用的术语和这些内容的技术是本领域众所周知的和被常用的。标准技术被用于例如核酸纯化和制备、化学分析、重组核酸和寡核苷酸合成。酶促反应和纯化技术根据制造方的说明书或如本领域通常所达成的的或者如本文所述的来执行。本文所描述的技术和流程通常根据本领域众所周知的、并且如在本说明书全文中所引用和讨论的各种一般性和更具体的参考文献中所描述的常规方法来执行。参见例如，Sambrook等人的Molecular Cloning:A Laboratory Manual(Third ed.,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.2000)。本文所描述的与之结合使用的命名法、以及实验室流程和技术是本领域众所周知的和常用的。

[0036] 如本文所使用的，“DNA”(脱氧核糖核酸)是指由4个类型的核苷酸组成的核苷酸链:A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)和G(鸟嘌呤)，并且RNA(核糖核酸)由4个类型的核苷酸组成:A、U(尿嘧啶)、G和C。某些核苷酸对以互补方式彼此特异性结合(被称为互补碱基配对)。即，腺嘌呤(A)与胸腺嘧啶(T)配对(但在RNA的情况下，腺嘌呤(A)与尿嘧啶(U)配对)，并且胞嘧啶(C)与鸟嘌呤(G)配对。当第一核酸链结合到由与第一链中的核苷酸互补的核苷酸组成的第二核酸链时，两条链结合以形成双链。如本文所使用的，“核酸测序数据”、“核酸测序信息”、“核酸序列”、“基因组序列”、“遗传序列”或“片段序列”或“核酸测序读段”表示指示核苷酸碱基(例如，腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶/尿嘧啶)在DNA或RNA分子(例如，全基因组、全转录组、外显子组、寡核苷酸、多核苷酸、片段等)中的顺序的任何信息或数据。

[0037] 应理解，本教导考虑了使用所有可用的技术(technique)、平台或技术(technology)所获得的序列信息，包括但不限于:毛细管电泳、微阵列、基于连接反应的系统、基于聚合酶的系统、基于杂交的系统、直接或间接核苷酸标识系统、焦磷酸测序、基于离子或pH的检测系统、基于电子签名的系统等。“多核苷酸”、“核酸”或“寡核苷酸”是指由核苷酸间键合连结的核苷(包括脱氧核糖核苷、核糖核或其类似物)的线性聚合物。通常，多核苷酸包含至少三个核苷。通常，寡核苷酸的大小范围为几个单体单元到(例如，3-4个)几百个单体单元。除非另有说明，否则每当多核苷酸(诸如寡核苷酸)由字母序列(诸如“ATGCCTG”)表示时，应理解，核苷酸从左到右的顺序为5’->3’，并且“A”表示脱氧腺苷，“C”表示脱氧胞苷、“G”表示脱氧鸟苷，并且“T”表示胸苷。如本领域中的标准，字母A、C、G和T可以被用于指代碱基本身、指代核苷或者指代包括碱基的核苷酸。

[0038] 短语“下一代测序”(NGS)是指与传统的基于桑格和毛细管电泳的方法相比具有提高的吞吐量的测序技术，例如，具有一次生成数十万个相对小的序列读段的能力。下一代测序技术的一些示例包括但不限于通过合成测序、通过连接反应测序以及通过杂交测序。更具体地，Illumina的MISEQ、HISEQ和NEXTSEQ系统以及Life Technologies Corp.的个人基因组机器(PGM)和SOLiD测序系统提供对全基因组或目标基因组的大规模并行测序。在国际申请日为2006年2月1日的题为“Reagents,Methods,and Libraries for Bead-Based Sequencing”的PCT公布号WO 2006/084132、于2010年8月31日提交的题为“Low-Volume Sequencing System and Method of Use”的美国专利申请系列号12/873,190以及于2010年8月31日提交的题为“Fast-Indexing Filter Wheel and Method of Use”的美国专利申请系列号12/873,132中更详细描述了SOLiD系统及相关联的工作流、方案、化学等，这些申请中的每个申请的全部内容通过引用并入本文。

[0039] 短语“测序运行”指代为了确定与至少一个生物分子(例如，核酸分子)有关的某些信息而被执行的测序实验的任何步骤或部分。

[0040] 如本文所使用的，短语“基因组特征”可以指代具有某些经注释功能的基因组区域(例如，基因、蛋白质编码序列、mRNA、tRNA、rRNA、重复序列、反向重复、miRNA、siRNA等)或遗传/基因组变体(例如，单核苷酸多态性/变体、插入/缺失序列、拷贝数变异、倒位等)，该遗传/基因组变体表示由于突变、重组/交叉或遗传漂移而参照特定物种或特定物种内的亚种群已经经历变化的(DNA或RNA中)单个基因或基因的分组。

[0041] 如本文所使用的，术语“生物标志物”指代生物学状态的客观可测量指示物。

[0042] 如本文所使用的,术语“致病性”指代增加个体对某种疾病或功能障碍的易感性或易患性的遗传改变变体性质。也被称为易患突变、有害突变和致病突变。

[0043] 如本文所使用的,术语“种系”指代被并入后代体内每个细胞的DNA中的衍生自生殖细胞(卵或精子)的组织,其。种系突变可以从亲代传给后代。

[0044] 如本文所使用的,术语“体细胞(的)”指代细胞在细胞分裂过程中所获得的基因改变。体细胞突变不同于种系突变,后者是发生在生殖细胞中的遗传性的基因改变。

[0045] 如本文所使用的,术语“密码子”指代对应于特定氨基酸的DNA或RNA的三核苷酸序列。

[0046] 如本文所使用的,术语“UI”是用户接口的首字母缩写。

[0047] 如本文所使用的,术语“查询时间”指待用户提交查询的时间点。

[0048] 如本文所使用的,术语“学习排名”或“排名引擎”或“相关性学习引擎”指代在针对信息检索系统的排名模型的构造中对机器学习的应用,通常是监督学习、半监督学习或强化学习。训练数据由项的列表组成,并且在每个列表中的项之间指定了一些偏序。该顺序通常通过针对每个项给出数字或序数得分或者二元判断(例如,“相关”或“不相关”)而被引发。排名模型的目的是排名,即,以在某种意义上与训练数据中的排名“类似”的方式,在新的、看不见的列表中产生项的排列。

[0049] 如本文所使用的,术语“潜在空间”或“隐藏空间”是指特征所在的空间。

[0050] 如本文所使用的,术语“嵌入”指代将文档(例如,文本、图像、结构化数据)映射到更低维度的潜在空间,保留对象主要特性。

[0051] 如本文中所使用的,术语“深度和广度模型”指代与深度神经网络(例如,用于泛化)联合训练广度线性模型(例如,用于记忆)的深度学习模型。

[0052] 如本文中所使用的,术语“语言模型”指代关于词语序列的概率分布。

[0053] 如本文中所使用的,术语“变换器模型”指代具有核心思想自注意力的深度学习模型,自注意力即注意输入序列的不同位置来计算该序列的表示的能力。

[0054] 如本文中所使用的,术语“BM25”指代信息检索中广泛的一族统计功能,这些统计功能考虑了文档或一组文档中每个查询术语的出现数目,即,术语频率(TF)以及对应的(多个)反向文档,并且在不考虑它们在文档中的接近性的情况下,基于每个文档中出现的查询术语而将一组文档排名。

[0055] 如本文中所使用的,术语“RM3”指代对于相关性反馈和伪相关性反馈两者都有用的信息检索模型。

[0056] 如本文中所使用的,术语“DSSM”是代表深度语义相似度模型的首字母缩写。

[0057] 如本文中所使用的,术语“孪生网络”指代在两个不同的输入向量上协同工作时使用相同权重来计算可比较的输出向量的人工神经网络。

[0058] 如本文中所使用的,术语“FDA”是美国食品和药物管理局的首字母缩写。

[0059] 如本文中所使用的,术语“NCCN”是国家综合癌症网络的首字母缩写。

[0060] 如本文中所使用的,术语“COSMIC”是癌症中体细胞突变目录的首字母缩写。

[0061] 如本文中所使用的,术语“TCGA”是癌症基因组图谱的首字母缩写。

[0062] 如本文中所使用的,术语“CPRA”是染色体、位置、参考和替代的首字母缩写。

[0063] 如本文中所使用的,术语“SNV”是单核苷酸变体的首字母缩写。

[0064] 如本文中所使用的,术语“CNV”是拷贝数变体的首字母缩写。

[0065] 如本文中所使用的,术语“BCL”是二进制碱基调用的首字母缩写。

[0066] 如本文中所使用的,术语“FASTQ”指代用于存储生物学序列(通常是核苷酸序列)以及其对应质量得分两者的基于文本的格式。为简洁起见,序列字母和质量得分各自都以单个ASCII字符而被编码。

[0067] 如本文中所使用的,术语“BAM”指代用于存储序列数据的二进制格式。

[0068] 如本文中所使用的,术语“VCF”是代表变体调用格式的首字母缩写,并且指代在生物信息学中被用于存储基因序列变异的文本文件的格式。

[0069] 如本文中所使用的,术语“EHR”是代表电子健康记录的首字母缩写。

[0070] 如本文中所使用的,术语“ASCO”是代表美国临床肿瘤学会的首字母缩写。

[0071] 本公开描述了用于癌症基因组和临床数据的综合分析的多组学搜索引擎的各种实施例,本文中被简称为“癌症搜索”。癌症搜索是于2017年3月21日提交的题为“Genomic Metabolic and Microbiombic Search Engine”的美国专利申请号15/465,454中提出的工作的扩展,其全部内容通过引用并入本文。

[0072] 根据各种实施例,提供了通用搜索引擎架构,该通用搜索引擎架构可以被配置为适配癌症多组学数据的特定需求。以下参考图1更详细地所讨论的整体架构1可以包括各种组件。例如,通用架构可以包括基于Web的用户接口、查询引擎、可以利用所有注释来对癌症多组学数据进行索引的索引管线、癌症分析软件模块以及排名引擎。查询引擎可以被配置为响应于请求来搜索可用于个体样本或同类群组的多组学数据流的任何组合。癌症分析(例如,在软件模块或引擎中)可以被配置为通过预先计算一些特性并且在查询时间动态地计算其他特性来得出重要的肿瘤特性。排名引擎可以被配置,使得在索引时间其将预加载默认的临床可操作的或者与致病性相关的排名,并且在查询服务时间其将基于所检测到的查询意图而进一步增强排名。以下将提供与各种数据类型、管线、引擎、模块和分析有关的更多详细信息。

[0073] 用户接口(UI)的总体功能可以被配置为呈现统一并且高度响应的方式用于查询和导航多组学癌症搜索结果。UI可以主动维护用户搜索会话的状态。UI可以被配置为接受用户查询,可以将用户查询中继到查询引擎,可以在可用的情况下呈现结果的集成多组学排名结果以及该结果的其概要可视化,并且可以允许用户与搜索结果交互。用户可以经由UI以各种方式与搜索结果交互,包括例如通过提供相关性反馈(例如,对结果有多好地回答用户信息需求的提升/降级/固定/删除类型评估)、通过对由搜索结果呈现的信息的准确性的评论(例如,特定注释源/公布过时或者不一致)、以及通过标记要被包括在动态个体患者或同类群组报告中的特定结果。与UI相关的更多详细信息将在下文提供。

[0074] 图1表示多组学癌症搜索系统100的通用架构的非限制性示例。针对(多个)样本(例如,肿瘤和/或正常样本)的一组多组学数据110(例如,基因组的、转录组的等)可以从体细胞工作流120被添加到索引管线或索引器115,或者经由用户接口125被上传。上传格式的非限制性示例可以包括FASTQ、BAM、针对肿瘤、正常、体细胞VCF的VCF、RNA-Seq变体确认VCF、表格形式的RNA-Seq差异基因表达、CNV VCF、结构性变体VCF、融合调用VCF或者它们的任意组合。多组学数据110可以是癌症多组学数据,该癌症多组学数据包括BCL、FASTQ、BAM、VCF、表格癌症数据、文本癌症数据、图像癌症数据。一组注释、文献和表型数据130可以经由

注释管线135而被添加到索引器115。数据可以或者驻留在存储单元170(例如,云存储、内部计算机存储)上,或者由用户经由专用搜索上传接口来上传。由索引管线115添加的数据可以被存储在一个或多个索引140中。系统架构还可以包括癌症分析引擎或模块145,癌症分析引擎或模块145可以被配置为在索引和服务时间得出肿瘤的重要特性。癌症分析引擎145可以得出所述重要特性,无论分析是针对单个样本还是同类群组。用户接口125可以允许用户录入查询并且接收由查询引擎150提供的结果。查询引擎150可以被配置为接受用户查询;选择、预连结、聚合和汇总相关的多组学索引;并且返回经排名的多组学数据或特征。根据各种实施例,系统架构还可以包括负载平衡器155,以针对大量用户提供UI 125和查询引擎150之间的双向数据传送。根据各种实施例,系统架构还可以包括认证代理160,并且包括身份提供方175(例如,第三方提供方)。从索引器115检索到的结果可以由排名引擎165(例如,学习排名引擎)排名,排名引擎165可以被配置为得出例如针对变体、基因、通路、表型、文本数据和图像的排名模型。从索引中检索到的结果可以由排名引擎排名,并且以经排名的顺序被呈现给用户。如本文将详细讨论的,无论其是基因组的、转录组的、表观遗传的、染色质可及性数据、微生物组的、蛋白质组的、医学文献、表型数据、文本数据、影像学数据、注释源、癌症分析、预测模型、有助于模型准确性的特征等,可以被查询、分析和排名的数据类型都是巨大的。以下将呈现关于与通用架构的该示例有关的各种方法和系统实施例的更多细节。

[0075] 现在参考图14,并且根据各种实施例,提供了方法1400用于利用多组学数据索引来进行肿瘤概况分析。该方法可以包括,在步骤1410处,存储多个多组学数据索引,其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据。贯穿本公开提供了与例如存储特征、多组学数据索引和癌症特定的数据有关的进一步讨论,并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0076] 该方法还可以包括,在步骤1420处,摄取附加多组学数据以及与附加多组学数据相关联的注释,附加多组学数据与一个或多个索引有关。贯穿本公开提供了与例如注释和摄取特征有关的进一步讨论,并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0077] 该方法还可以包括,在步骤1430处,对所摄取的附加多组学数据和注释进行索引,同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射,以产生经记号化的所摄取的附加多组学数据。贯穿本公开提供了与例如索引、基因名称、基因变体名称和多组学映射有关的进一步讨论,并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0078] 该方法还可以包括,在步骤1440处,接收用户查询。贯穿本公开提供了与例如接收特征和用户查询有关的进一步讨论,并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0079] 该方法还可以包括,在步骤1450处,基于用户查询来选择相关的一个或多个多组学数据索引。贯穿本公开提供了与例如选择特征、多组学索引的预连结以及相关性确定有关的进一步讨论,并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0080] 该方法还可以包括,在步骤1460处,基于临床可操作性、致病性、特征权重和频率中的至少一项,将所选择的一个或多个多组学数据索引排名。也可以包括其他排名因素,例

如与查询意图有关的因素。贯穿本公开提供了与排名有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0081] 该方法还可以包括，在步骤1470处，向用户返回经排名的一个或多个多组学数据索引。贯穿本公开提供了与例如返回特征、显示和报告有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0082] 根据各种实施例，一种非瞬态计算机可读介质存储有用于使计算机执行用于利用多组学数据索引进行肿瘤概况分析的方法的程序。该方法内的步骤可以类似于以上所提供的步骤，或者可以如所需要的而变化。

[0083] 该方法可以包括存储多个多组学数据索引，其中多个多组学数据索引中的每多组学数据索引包括癌症特定的记号化数据。贯穿本公开提供了与例如存储特征、多组学数据索引和癌症特定的数据有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0084] 该方法还可以包括摄取附加多组学数据以及与附加多组学数据相关联的注释，附加多组学数据与一个或多个索引有关。贯穿本公开提供了与例如注释和摄取特征有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0085] 该方法还可以包括对所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。贯穿本公开提供了与例如索引、基因名称、基因变体名称和多组学映射有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0086] 该方法还可以包括接收用户查询。贯穿本公开提供了与例如接收特征和用户查询有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0087] 方法还可以包括基于用户查询来选择相关的一个或多个多组学数据索引。贯穿本公开提供了与例如选择特征和相关性确定有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0088] 该方法还可以包括基于临床可操作性、致病性、特征权重和频率中的至少一项，将所选择的一个或多个多组学数据索引排名。应注意，排名可以由查询的意图来进一步更改(例如，按照反向频率的顺序排名、按照对模型的特定预测的特征贡献的顺序排名、按照其权重的倒序将突变签名排名等)。像这样，如果其他排名未被请求并且其他意图不容易(或者无法)被推断，则临床可操作性可以用作默认排名。贯穿本公开提供了与将特征排名和确定有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0089] 该方法还可以包括，向用户返回经排名的一个或多个多组学数据索引。贯穿本公开提供了与例如返回特征有关的进一步讨论，并且这些讨论将适用于本文所讨论或考虑的该实施例和所有实施例。

[0090] 根据各种实施例，多组学数据可以选自包括以下项的组：基因组的、转录组的、表观遗传的、染色质可及性数据、微生物的、蛋白质组的、表型的、图像、相关文献、集成多组学数据、以及前述项的组合。根据各种实施例，多个多组学数据索引还可以包括肿瘤(体细胞)基因组改变、正常(种系)基因组改变和癌症注释源。

[0091] 根据各种实施例，本文所讨论或考虑的方法还可以包括针对所选择的一个或多个多组学数据索引得出癌症分析。癌症分析可以包括选自包括以下项的组的肿瘤特性：质量控制、肿瘤突变负荷、基因组突变签名、微卫星不稳定性状态、新抗原及其结合亲和力、HLA等位基因分型、RNA确认的变体、拷贝数变体、结构性变体、非编码调控变体、基因融合、通路富集、癌症驱动因素标识、突变概要、差异基因表达、免疫签名、以及前述项的组合。根据各种实施例，癌症分析可以针对单个样本或样本同类群组来得出。此外，癌症分析可以包括关于类似患者的治疗结果的匹配信息。根据各种实施例，癌症分析可以包括机器学习预测和经排名的特征。根据各种实施例，癌症分析可以包括机器学习预测、以及按照其与特定预测的相关性的顺序而被排名机器学习模型特征。机器学习预测可以选自包括以下项的组：主要原发部位分类器、未来转移部位预测分类器、微卫星不稳定状态预测、新抗原结合亲和力预测、疾病状态分层、确定癌症谱系、以及前述项的组合。癌症分析可以在接收到用户查询之后被动态计算。癌症分析的得出可以包括利用深度神经网络和其他机器学习方法(例如，支持向量分类器、树方法、集合方法(Ensemble Methods))。模型特征重要性的得出可以包括梯度归因方法或者其他特征重要性方法

[0092] 根据各种实施例，本文所讨论或考虑的方法还可以包括将注释从更高级别的基因组层级向更低级别的基因组层级传播。

[0093] 根据各种实施例，本文所讨论或考虑的方法还可以包括对所选择的一个或多个多组学数据索引从更高级别的基因组层级到更低级别的基因组层级的排名的传播。排名可以包括针对癌症变体和基因的临床的排名。排名可以包括基因的富集属于特定通路的概率。排名可以包括针对机器学习模型的特征所确定的重要性权重。排名可以包括通过并入癌症数据的潜在空间表示以及子选择表示来将同类群组分层，该分层引起反应者与非反应者、短期与长期无进展生存、一种癌症亚型与另一种癌症亚型等之间的最大解纠缠。同类群组可以被分层为反应者和非反应者。同类群组可以被分层为长期无进展生存时间和短期无进展生存时间。同类群组可以被分层为不同的癌症亚型。潜在空间表示可以由神经网络或者任何其他降维方法(例如，主成分分析、个体成分分析、流形学习)来执行。神经网络可以选自包括以下项的组：自动编码器、变分自动编码器、深度置信网络、受限玻耳兹曼机、前馈、卷积、递归、门控递归、长短期记忆、残差、以及生成对抗网络。

[0094] 根据各种实施例，包括本文所讨论或考虑的方法，排名还可以包括用于学习排名的模型，该用于学习排名的模型选自包括以下项的组：支持向量机、提升决策树、回归方法、神经网络、以及前述项的组合。用于学习排名的模型还可以包括其他机器学习模型或者深度神经网络。排名还可以包括深度学习排名。排名还可以包括查询的嵌入与在经由深度学习方法所学习到的联合嵌入空间中的经索引文档之间的相似度。深度学习排名可以从选自包括以下项的组的深度学习模型得出：深度语义相似度模型、深度和广度模型、深度语言模型、学习到的深度学习文本嵌入、学习到的命名实体识别、孪生神经网络、以及前述项的组合。

[0095] 根据各种实施例，包括本文所讨论或考虑的方法，多组学数据可选自包括以下项的组：从全基因组序列数据的体细胞(和种系)调用、从全外显子组序列数据的体细胞(和种系)调用、从新鲜冷冻组织的体细胞(和种系)组套(panel)测序、从福尔马林固定石蜡包埋组织的体细胞(和种系)组套测序、从液体活检的体细胞(和种系)组套测序、肿瘤和正常变

体调用、被索引为在RNA或基因表达水平上确认变体的肿瘤/正常转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、单个细胞测序数据、以及前述项的组合。在各种实施例中，被索引的多组学数据可以或者来自内部的体细胞调用和免疫管线，或者可以以FASTQ、BAM、VCF和其他表格格式的形式而从任何外部合作伙伴实时被提供或上传。

[0096] 根据各种实施例，包括本文所讨论或考虑的方法，多组学数据索引还可以包括被提取的表型数据。表型数据可以选自包括以下项的组：电子健康记录、临床数据、功能数据、以及前述项的组合。

[0097] 根据各种实施例，包括本文所讨论或考虑的方法，多组学数据索引还可以包括特征化/嵌入式影像学数据。特征化的影像学数据可以选自包括以下项的组：组织学投影片、MRI图像、X射线、乳房X线照片、超声、PET图像、CT扫描、以及前述项的组合。

[0098] 根据各种实施例，包括本文所讨论或考虑的方法，对所摄取的附加多组学数据和注释的进行索引还可以包括对所得出的数据进行索引，所得出的数据选自包括以下项的组：癌症分析、注释、从影像学数据中所提取的特征、数据、表型、医学文献数据、数据嵌入、以及前述项的组合。

[0099] 根据各种实施例，包括本文所讨论或考虑的方法，排名还可以包括将样本改变与已建立的药物靶标标签和可用的临床试验相匹配。排名还可以包括通过检测基于感兴趣的临床变量和/或统计显著性将同类群组分层的潜在生物标志物来进行所述同类群组中的癌症药物靶标的标识，并且其中向用户返回经排名的一个或多个多组学数据索引包括分层可视化。

[0100] 根据各种实施例，包括本文所讨论或考虑的方法，向用户返回经排名的一个或多个多组学数据索引还可以包括针对个体患者和/或同类群组动态创建超链接报告(例如，包含经排名的改变，其中每个条目被超链接到搜索查询)，超链接报告提供对肿瘤或癌症的全面概况分析。向用户返回经排名的一个或多个多组学数据索引还可以进包括返回所返回的结果的概要可视化以及经排名的结果的列表。

[0101] 根据各种实施例，包括本文所讨论或考虑的方法，用户查询可以包括用户上传的数据，用户上传的数据选自包括以下项的组：一组套的变体、基因、通路、疾病状态病情、感兴趣的表型，并且其中选择包括查询由所上传的数据子选择的个体样本或同类群组数据。用户查询可以经由用户接口而被提供，并且可以包括上传用于进行索引的数据，该用于进行索引的数据选自包括以下项的组：基因组数据、转录组数据、表观遗传数据、染色质可及性数据、微生物组数据、蛋白质组数据、表型数据、注释数据、以及前述项的组合。

[0102] 根据各种实施例，本文所讨论或考虑的方法还可以包括规范化和/或扩大用户查询、将查询意图分类、汇总检索到的文档、以及使用深度学习方法基于查询和潜在空间中的文档之间的相似度来执行文档检索。

[0103] 根据各种实施例，包括本文所讨论或考虑的方法，进行索引、选择和排名中的至少一项包括利用深度神经网络。

[0104] 根据各种实施例，本文所讨论或考虑的方法(和系统)可以操作以集中大量癌症多组学数据，来为肿瘤学家、医学从业者、研究科学家和其他非程序员提供以任何细节级别询问癌症生物信息学管线、以及获得对癌症生物学和潜在癌症临床治疗方法的临床和生物学

见解的平台。数据类型可以包括例如基因组(单核苷酸变异、肿瘤和正常组织的插入和缺失、结构性重排、拷贝数变异、基因融合和肿瘤基因组的被表达的变体)、转录组的、表观遗传的、染色质可及性、微生物组的、蛋白质组丰度和定位、医学文献数据(出版物、治疗指南、临床试验纳入/排除标准)、表型数据(功能、临床、电子病历、组织病理学和放射学报告)、影像学数据(组织病理学投影片、MRI扫描、X射线、乳房X线照片、超声、PET图像、CT扫描)、癌症注释源(变体、基因、通路、药物)、所得出的癌症分析(肿瘤突变负荷、突变签名、微卫星不稳定状态、RNA序列确认的变体、差异表达的基因、空间多组学谱系表示、MHC I类和II类分子的新抗原结合亲和力)。

[0105] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)包括癌症分析(例如,作为步骤、特征、引擎、模块或软件模块)。癌症分析允许用户可以访问肿瘤的重要特性,包括例如肿瘤突变负荷、突变签名、空间多组学谱系表示、MHC I类和II类分子的新抗原结合亲和力、RNA序列确认的变体、差异表达的基因、通路富集、微卫星不稳定性状态和微卫星重复位点以及从影像学和临床数据中提取的特征。根据各种实施例,该数据可以针对个体样本被预先计算或者针对同类群组样本被动态计算。根据各种实施例,癌症分析可以提供对来自机器学习模型的预测及其按特征对特定分类的贡献被排名的特征的集成。特定分类可以包括例如主要原发部位、未来转移部位的预测、将变体分类为真阳性或假阳性、关于类似患者的治疗结果的信息、测序质量的异常值检测以及使用潜在和实际表示的针对类群组的疾病状态预测。返回按特征对特定分类的贡献被排名的特征的优点在于,模型预测对用户而言更加可解释。

[0106] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)包括多模式排名(例如,作为步骤、特征、引擎、模块或软件模块)。多模式排名可以提供相关性学习引擎,以集成多组学遗传数据、注释源、文献数据、临床试验结果以及良好表征的同类群组中显著突变的基因来了解癌症数据的临床可操作性排名。在各种实施例中,机器学习模型可以被用于权衡来自多组学数据的注释的贡献。在各种实施例中,深度学习和机器学习降维技术可以被用于得出样本同类群组的潜在空间表示。在各种实施例中,学习到的嵌入可以被用于将基因组的、文本和影像学数据进行排名。

[0107] 如上所述,并且如将在下文进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)还包括用于将多个癌症注释源集成和排名的机制(例如,作为步骤、特征、引擎、模块或软件模块)。这些源可以包括例如FDA标签、NCCN指南、临床试验、CIViC、DoCM、OncoKB、Mycancergenome、癌症药物基因组生物标志物数据库、TCGA、ICGC、COSMIC、NCI60、CCLE、Drugbank、ClinVar、HGMD、PGMD、PharmGKB、dbSNP、dbNSFP、1000Genomes、EXEC、CPDB、KEGG、BioCarta、BioCyc、Reactome、GenMAPP、MsigDB、Brenda、CTD、HPRD、GXD、BIND。在各种实施例中,注释和排名可以从更高级别的表示传播更低级别(例如,从通路到基因到变体、或者从基因到变体密码子到完整的变体规范-染色体、位置、参考、替代)。

[0108] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,根据各种实施例,本文所描述和考虑的各种方法(和系统)还包括用于集成若干深度学习模型的机制(例如,作为步骤、特征、引擎、模块或软件模块)。集成可以起到提供神经数据索引的作用(例如,单独和一起嵌入多组学数据集,以将它们针对DNA和RNA肿瘤改变的相应潜在空间规则化;嵌入来自电子健康记录、临床笔记、文献、注释的文本数据;深度变换器模型,用于命名

实体识别以及文本和注释数据的汇总;嵌入影像学数据)。集成还可以提供神经学习排名模型(例如,深度语义相似度模型、卷积深度语义相似度模型、循环深度语义相似度模型、深度相关度匹配模型、交互孪生网络、词法和语义匹配网络、DeepRank),该神经学习排名模型可以被用于解决学习排名的特征工程问题。集成可以提供神经查询模型(例如,用于查询规范化、同义词展开、缩写展开、术语去歧义化、备选建议的深度学习变换器模型)。集成可以起到提供用于高级癌症分析的神经模型(例如,原发部位分类、未来转移部位的预测、新抗原结合亲和力预测、将变体分类为真或假阳性、药物和试验匹配、使用来自经索引的相似病例的信息的治疗推荐系统、比较等位基因分数的减少、增加、维持的模型、拷贝数变异、系列活检在每个位置处的RNA表达、以及用于同类群组分析和分层的深度学习自动编码器方法和其他降维技术)的作用。

[0109] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文中所描述和考虑的各种方法(和系统)还可以包括用于标识(多个)诊断、预后或预测性生物标志物的统计、机器学习和深度学习方法(例如,作为步骤、特征、引擎、模块或软件模块)。在各种实施例中,当用户(例如,学术或行业研究人员)录入关于样本同类群组的表型查询时,经排名的生物标志物被返回,该经排名的生物标志物可以将同类群组、其统计显著性以及其概要可视化分层。在各种实施例中,验证查询可以由搜索引擎建议以执行鲁棒的算法和统计验证。在各种实施例中,系统和方法可以经由所建议的查询细化来自动建议迭代假设细化。根据各种实施例,针对癌症同类群组查询得出的统计学可视化和分析可以包括例如提供了统计显著性的Kaplan-Meier生存分析可视化、对数秩检验结果可视化、Cox比例风险回归分析可视化、树状生存模型可视化、热图、散点图、箱形图和条形图。

[0110] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)还可以包括使用和/或接收交互式概要可视化和/或经排名的变体、基因、通路、所得出的癌症分析、集成机器学习模型的输出(例如,癌症类型分类、最可能的复发部位)(例如,作为步骤、特征、引擎、模块或软件模块)。这可以经由(以下将更详细地讨论的)查询引擎而被提供。在各种实施例中,概要可视化可以是动态的,并且每个数据点可以被链接到所返回的特定结果。

[0111] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)还可以在10000、5000、4000、3000、2000、1000、900、800、700、500、400、300、200、100毫秒或更短的访问内,或者在上述值之间的任何访问范围内,提供对按临床可操作性、致病性、特征权重或频率而被排名的多组学癌症数据的交互式和快速的访问。

[0112] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以提供通用搜索接口(如与许多不同的入口点相对)。在各种实施例中,所有知识,例如,多组学癌症数据、样本、变体、基因、药物、通路、表型、医学文献、图像数据、所得出的癌症分析、用于预测肿瘤特性以及其特征的机器学习模型、用户数据的上传等,可以通过相同的简单搜索接口可访问。

[0113] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考虑的各种方法(和系统)还可以提供(比较顺序的活检样本并且提供新旧癌症驱动因素、变体等位基因分数变化、拷贝数变化以及癌症改变的RNA确认状态变化之间的差异(增加、减少、维持)的能力(例如,作为步骤、特征、引擎、模块或软件模块)。

[0114] 如上所述,并且如将在以下进一步详细讨论的,根据各种实施例,本文所描述和考

虑的各种方法(和系统)还可以提供各种比较方案(例如,作为步骤、特征、引擎、模块或软件模块)。这些方案可以包括例如,(1)样本与样本的比较,同一患者内多组学数据流的任意组合的比较,(2)样本与同类群组的比较(例如,将个体样本与TCGA中的相同癌症亚型比较),以及(3)成对的同类群组比较(例如,将同类群组与具有相同癌症类型的良好表征的TCGA同类群组比较)。

[0115] 根据各种实施例,本文所描述和考虑的各种方法(和系统)可以提供对来自用户机构的变体/基因药物靶标组套(或者当前在实践中所使用的组套)的动态上传(例如,作为步骤、特征、引擎、模块或软件模块)。后续的查询可以指示使用所上传的组套和针对(多个)样本所存储的多组学数据的交集。

[0116] 在公共领域中,并且如本文已经讨论的,已经提出了通用基因组搜索来解决对种系基因组数据的立即访问的问题。G该搜索表示种系基因组概况分析的显著不同的问题,这些问题聚焦于孟德尔稀有变体、GWAS命中率、常见疾病的负荷测试和多基因风险、以及遗传性风险。为了有效解决以上和本文所讨论的全面的癌症表征中的所有三个主要问题,根据所提供和考虑的各种实施例,本文所述的系统和方法还可以包括针对个体样本和同类群组的高级癌症分析,以及(以上和本文所详细讨论的)排名引擎。根据本文所提供的各种实施例,本文所描述的系统和方法可以扩充现有通用种系搜索系统的所有部分,以在进行索引和服务时间期间集成多组学数据、由于癌症改变的临床相关性和致病性而将癌症改变排名、并且使得搜索引擎范式对针对个体样本和同类群组的全面癌症概况分析有用。附加地,根据本文所提供的各种实施例,本文所描述的系统和方法可以包括构建在癌症搜索引擎之上的癌症同类群组分层分析,该分层分析整体是先前工作所缺少的。

[0117] 根据各种实施例,图15图示了被提供用于利用多组学数据索引来进行肿瘤概况分析的系统1500。系统1500可以包括索引单元1510。索引单元可以包括被配置为存储多个多组学数据索引的存储元件1520,其中多个多组学数据索引中的每个多组学数据搜索引擎包括癌症特定的记号化数据。索引单元1510还可以包括索引引擎1530。索引单元1510可以被配置为经由数据源1540来摄取附加多组学数据以及与附加多组学数据相关联的注释,附加多组学数据与一个或多个索引有关。索引单元1510还可以被配置为对来自数据源1540的所摄取的附加多组学数据和注释进行索引,同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射,以产生经记号化的所摄取的附加多组学数据。

[0118] 系统1500还可以包括被配置为接收用户查询1560的用户接口1550。

[0119] 系统1500还可以包括查询引擎1570,查询引擎1570被配置为基于用户查询1560来从索引单元1510选择相关的一个或多个多组学数据索引。

[0120] 系统1500还可以包括排名引擎1580,排名引擎1580被配置为(例如,从查询引擎1570)接收所选择的相关的一个或多个多组学数据索引,以将所选择的一个或多个多组学数据索引排名,并且经由用户接口1550来向用户返回经排名的一个或多个多组学数据索引。

[0121] 根据各种实施例,图16图示了被提供用于利用多组学数据索引来进行肿瘤概况分析的系统1600。系统1600可以包括索引单元1610。索引单元可以包括存储元件被配置为存储多个多组学数据索引的存储元件1620,其中多个多组学数据索引中的每个多组学数据索

引包括癌症特定的记号化数据。索引单元1610还可以包括索引引擎1630。索引单元1610可以被配置为经由数据源1640来摄取附加多组学数据以及与附加多组学数据相关联的注释，附加多组学数据与一个或多个索引有关。索引单元1610还可以被配置为对来自数据源1640的所摄取的附加多组学数据和注释进行索引，同时在特定索引中保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多组学映射，以产生经记号化的所摄取的附加多组学数据。

[0122] 系统1600还可以包括被配置为接收用户查询1660的用户接口1650。

[0123] 系统1600还可以包括查询引擎1670，查询引擎1670被配置为基于用户查询1660来从索引单元1610选择相关的一个或多个多组学数据索引。查询引擎1670还可以被配置为基于临床可操作性、致病性、特征权重或频率而将所选择的一个或多个多组学数据索引排名。查询引擎还可以被配置为经由用户接口1650来向用户返回经排名的一个或多个多组学数据索引。

[0124] 注意根据各种实施例，所有附加特征的所有先前讨论，特别是关于先前所描述的方法和非瞬态计算机可读介质的，适用于本文所描述和考虑的各种系统实施例的特征。

[0125] 根据各种实施例，提供了用于利用多组学数据索引来进行肿瘤概况分析的计算机实现的系统。该系统可以包括计算机存储装置、包括至少一个处理器的数字处理设备、被配置为执行可执行指令的操作系统、存储器以及包括指令的计算机程序，该指令由数字处理设备可执行来创建多组学癌症搜索引擎应用。多组学癌症搜索引擎应用可以包括在计算机存储装置中被记录的多个集成的多组学索引、以及提供高级癌症分析的软件模块。多组学癌症搜索引擎应用可以包括如下软件模块：该软件模块提供多组学索引管线，多组学索引管线摄取与多组学基因组和影像学数据相关联的多组学癌症数据、注释、医学和临床数据，将数据记号化，同时保留变体命名法、基因名称和药物名称，并且利用经记号化的数据来更新索引。多组学癌症搜索引擎应用还可以包括如下软件模块：该软件模块负责将反映癌症改变的临床效用的集成的多组学数据排名。多组学癌症搜索引擎应用可以包括查询引擎，该查询引擎选择并且组合相关的多组学索引，并且返回针对个体样本和样本同类群组的经排名的多组学改变。多组学癌症搜索引擎应用可以包括呈现用户接口的软件模块，该用户接口允许用户录入用户查询并且对多组学数据执行分面搜索。

[0126] 根据各种实施例，提供了被编码为具有计算机程序的非瞬态计算机可读存储介质，该计算机程序包括由处理器可执行来创建多组学癌症搜索引擎应用的指令。多组学癌症搜索引擎应用可以包括在计算机存储装置中被记录的多个集成的多组学索引、以及提供高级癌症分析的软件模块。多组学癌症搜索引擎应用可以包括如下软件模块：该软件模块提供多组学索引管线，以摄取与多组学基因组和影像学数据相关联的多组学癌症数据、注释、医学和临床数据，将数据记号化，同时保留变体命名法、基因名称和药物名称，并且利用经记号化的数据来更新索引。多组学癌症搜索引擎应用还可以包括如下软件模块：该软件模块负责将反映所返回结果的临床效用、致病性、频率、特征权重的集成的多组学数据排名。多组学癌症搜索引擎应用可以包括查询引擎，该查询引擎选择并且组合相关的多组学索引、并且返回针对个体样本和样本同类群组的经排名的多组学改变。多组学癌症搜索引擎应用可以包括呈现用户接口的软件模块，该用户接口允许用户录入用户查询并且对多组学数据执行分面搜索。

[0127] 根据各种实施例，提供了提供多组学癌症搜索引擎应用的计算机实现的方法。多组学癌症搜索引擎应用可以包括在计算机存储装置中被记录的多个集成的多组学索引、以及提供高级癌症分析的软件模块。多组学癌症搜索引擎应用可以包括如下软件模块：该软件模块提供多组学索引管线，多组学索引管线摄取与多组学基因组和影像学数据相关联的多组学癌症数据、注释、医学和临床数据，将数据记号化，同时留变体命名法、基因名称和药物名称，并且利用经记号化的数据来更新索引。多组学癌症搜索引擎应用可以包括如下软件模块：该软件模块负责对反映所返回结果的癌症改变临床效用、致病性、频率、特征权重的集成的多组学数据进行排名。多组学癌症搜索引擎应用可以包括查询引擎，查询引擎选择并且组合相关的多组学索引、并且返回针对个体样本和样本同类群组的经排名的多组学改变。多组学癌症搜索引擎应用可以包括呈现用户接口的软件模块，该用户接口允许用户录入用户查询并且对多组学数据执行多面搜索。在各种实施例中，索引被最佳地格式化为部分预连结的配置，并且临床排名被预加载，使得搜索速度被提高，并且搜索和结果之间的滞后时间被减少。在各种实施例中，多组学索引的预连结发生在用户录入查询之前。

[0128] 注意根据各种实施例，附加特征的所有先前讨论，特别是关于先前所描述的计算机实现的方法、计算机实现的系统和非瞬态计算机可读介质的，适用于本文所描述和考虑的各种系统实施例的特征。

[0129] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以集中包括的大量癌症多组学数据。该数据可以包括例如基因组的(例如，单核苷酸变异、肿瘤和正常组织中的插入和缺失、结构性重排、拷贝数变异、基因融合、以及肿瘤基因组的被表达的变体)、转录组的(例如，RNA-Seq变体确认和差异基因表达)、表观遗传的、染色质可及性、微生物组的、蛋白质组丰度和定位、医学文献数据(例如，出版物、治疗指南、临床试验纳入/排除标准)、表型数据(例如，功能、临床、EHR)、影像学数据(例如，组织学、MRI、X射线、乳房X线照片、超声、PET图像、CT扫描)、癌症注释源(例如，变体、基因、通路、药物)、所得出的癌症分析(例如，肿瘤突变负荷、突变签名、微卫星不稳定状态、空间多组学谱系表示、MHC I类和II类分子的新抗原结合亲和力)、来自机器学习模型的预测及其特征(例如，主要原发部位、微卫星不稳定性、潜在的未来转移部位、药物和试验匹配项)。根据各种实施例，基因组数据可以是全外显子组、全基因组、基因组套数据、SNP阵列的形式。根据各种实施例，出于监测疾病进展、耐药性的发展和复发监测的目的，顺序的活检多组学数据可以被索引。

[0130] 根据各种实施例，经索引的数据可以是例如但不限于针对肿瘤和正常两者、或者仅针对肿瘤的变体调用格式(VCF)、BAM和FASTQ的形式。根据各种实施例，表型数据可以以表格格式或者原始格式(例如，HER、临床记录、pdf报告)而被提供。

[0131] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以包括注释源。注释源的示例可以包括但不限于：FDA标签、NCCN指南、临床试验、CIViC、DoCM、OncoKB、Mycancergenome、癌症药物基因组生物标志物数据库、TCGA、ICGC、COSMIC、NCI60、CCLE、Drugbank、ClinVar、HGMD、PGMD、PharmGKB、dbSNP、dbNSFP、1000Genomes、EXEC、CPDB、CADD、PolyPhen、dbNSFP等。

[0132] 根据各种实施例，本文所述的系统和方法还可以包括药物靶标信息，药物靶标信息可以从多个源被得出并且被集成。这些源包括但不限于FDA标签、NCCN药品和生物制品纲要、Thomson Micromedex DrugDex、Elsevier Gold Standard的临床药理纲要、美国医院处

方药品信息纲要、ESMO指南、ASCO指南、NCCN指南以及在其他癌症知识数据库(诸如例如OncoKB、CIViC、DoCM、COSMIC)中所注释的突变。根据各种实施例,药物靶标可以在变体、基因和通路级别被索引。根据各种实施例,药物适应症、证据、癌症类型、所报告的不良反应和附加信息可以被存储在搜索索引中。

[0133] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以包括癌症分析(或者高级癌症分析),或者提供高级癌症分析的软件模块,或者对前述项的使用。软件模块可以提供被预先计算的(例如,在索引时间被计算)和动态的(例如,在查询时间计算的)被得出的癌症分析两者。根据各种实施例,高级分析还可以在查询时间被可视化。图3图示了针对个体样本和同类群组被预先计算和被动态计算的癌症分析的示例。高级分析模块可以集成来自机器学习和深度学习模型的预测,用于预测肿瘤生物学的重要特性。

[0134] 根据各种实施例,针对个体样本的被预先计算的得出的癌症分析可以包括例如但不限于肿瘤突变负荷(用于疗法(例如免疫疗法)的重要生物标志物)、微卫星不稳定性状态(重要的癌症状态,其中错配修复蛋白失去作用)、基因组突变签名(癌症的潜在病因和机制基础)、所检测到的neoORF(可能导致新的氨基酸序列的移码突变,可以对癌症疫苗有用)、所检测到的新抗原、MHC I类和II类分子的新抗原结合亲和力、HLA等位基因分型(癌症疫苗设计的重要变量)、被表达的免疫基因(例如,对免疫治疗产生反应的基因)、RNA序列确认的变体以及差异表达的基因。

[0135] 根据各种实施例,针对个体样本的动态高级癌症分析可以包括,例如但不限于,针对特定类型的变体(基于查询,例如,非沉默变体)的通路富集分析和空间多组学谱系表示。根据各种实施例,针对样本同类群组的动态高级癌症分析可以包括但不限于:同类群组突变签名;通过折叠相同基因中的复发性体细胞改变并且在纠正非沉默与沉默变体的比、基因复制时间和癌症生物学的其他性质之后,检测显著突变的基因和癌症驱动因素;疾病状态分层;空间多组学谱系表示;以及针对变体子集(例如,非沉默突变)的通路富集分析。

[0136] 根据各种实施例,癌症分析可以经由高级分析模块而被提供,高级分析模块可以被配置为集成来自例如机器学习和深度学习模型的预测,用于预测肿瘤生物学的重要特性(例如,针对微卫星不稳定状态的仅肿瘤和肿瘤-正常分类器;针对未知原发的转移性肿瘤的肿瘤原发分类;用于预测特定患者的最可能复发部位的模型;用于仅肿瘤变体调用的深度学习和机器学习方法;新抗原结合预测;用于不同癌症类型的遗传性癌症风险预测的机器学习模型;用于免疫疗法结果预测的机器学习模型;将变体分类为真阳性或假阳性;针对变体、基因、药物和疾病的深度学习方法;用于处理文献、EHR和临床试验数据的命名实体识别;用于标识感兴趣的区域并且从非结构化的组织学和放射学投影片以及其他影像学数据中提取特征的深度学习方法;用于学习癌症多组学疾病状态的潜在嵌入的深度学习模型;用于药物和试验匹配的深度学习方法;用于标识类似患者的机器学习模型;基于治疗类似患者的结果的癌症治疗推荐器系统;以及针对(多个)同类群组生物标志物分层和同类群组疾病状态标识的机器学习和深度学习方法)。

[0137] 根据各种实施例,本文所述的系统和方法可以包括例如(例如,从电子健康记录、临床和功能记录所学习到的)表型数据、注释源、医学文献或影像学数据(例如,组织学投影片、MRI、X射线、乳房X线照片、超声、PET图像、CT扫描)的深度学习嵌入。

[0138] 根据各种实施例,本文所述的系统和方法可以包括高级癌症分析模块,高级癌症

分析模块设置关于质量控制的统计阈值，针对经索引的测序质量指标来标识的异常值。感兴趣的质量控制指标的一些非限制性示例可以包括针对肿瘤-正常匹配的质量控制(例如，血缘关系和同一性值)；肿瘤和正常测序指标(例如，反映潜在肿瘤/正常污染的Freemix/Conpair指标，包括但不限于平均总覆盖率、对齐的百分比读段、重复百分比和Y/X比的测序指标)；以及体细胞测序质量控制指标，包括但不限于dbSNP中的变体数目、dbSNP富集、dbSNP插入缺失比、dbSNP转换/颠换比和异质/均质变体比率(杂合/纯合变体比)。

[0139] 根据各种实施例，高级癌症分析(或其相关联的模块)可以提供例如动态算法，该动态算法用于基于样本同类群组中的可疑(多组学)生物标志物的突变概要、癌症驱动因素标识、多个活组织检查的比较以及同类群组分层。在各种实施例中，可以实现样本与样本同类群组的比较，以及多个同类群组的比较。

[0140] 根据各种实施例，本文所述的系统和方法可以包括索引和集中大量癌症多组学数据。如上文详细地所讨论的，数据可以包括例如但不限于基因组数据(例如，单核苷酸变异、肿瘤和正常的插入和缺失、结构性重排、拷贝数变异、基因融合和肿瘤基因组的被表达的变体)、转录组数据、表观遗传数据、染色质可及性数据、微生物学数据、蛋白质组丰度和定位数据、医学文献数据(例如，出版物、治疗指南、临床试验纳入/排除标准)、表型数据(例如，功能、临床、EHR)、影像学数据(例如，组织学投影片、MRI、X射线、乳房X线照片、超声、PET图像、CT扫描)、癌症注释源(例如，变体、基因、通路、药物)、所得出的癌症分析(例如，肿瘤突变负荷、突变签名、差异表达的基因、空间多组学谱系表示、来自机器学习模型主要原发部位、未来转移部位、微卫星不稳定状态、MHC I类和II类分子的新抗原结合亲和力的预测和特征)。

[0141] 申请人已有利地发现，通过对原始数据和所得出的分析进行索引，来自机器学习和深度学习模型的预测以及其(得出的)特征和嵌入可以包括更好的机器学习可解释性、迭代假设生成以及对用户的后续查询的细化，从而可以更好地表征和了解肿瘤生物学。

[0142] 根据各种实施例并且如上所讨论，本文所公开的系统和方法可以包括软件模块，软件模块用于对癌症数据、与基因组和影像学数据相关联的注释、医学和临床数据进行多组学索引，将数据记号化，同时保留变体命名法、基因名称和药物名称，并且利用经记号化的数据来更新索引。根据各种实施例，多组学索引的步骤可以包括在变体、基因、通路、癌症亚型或样本的级别上集成和预连结多组学索引。

[0143] 特定于癌症注释数据，根据各种实施例，本文所述的系统和方法可以包括索引步骤(参见上文)，或者提供针对癌症注释数据进行多组学索引的软件模块。癌症注释数据可以包括但不限于FDA标签和NCCN指南、临床试验、公共癌症数据库(CIViC、DoCM、OncoKB、Mycancergenome、COSMIC、癌症药物基因组生物标志物数据库、ICGC、TCGA)、公共基因组数据库(ClinVar、dbNSFP、dbSNP)、商业数据源(HGMD、PGMD、PharmGKB、CPDB)。在另一方面，多组学索引软件模块也索引非聚焦癌症的注释源：ClinVar、dbNSFP、dbSNP、CPDB、HGMD、PGMD。根据各种实施例，用于多组学索引的软件模块可以被配置为在变体、基因密码子编号、基因、通路、癌症亚型或样本的级别上集成和预连结多组学注释数据。

[0144] 根据各种实施例，索引还可以包括利用所得出的内容嵌入来对复杂的表型、文献数据、组织病理学、MRI、X射线、乳房X线照片、超声、PET图像、CT扫描图像进行索引。

[0145] 根据各种实施例，本文所述的系统和方法还可以包括索引流程，在索引流程中，在

进行索引期间的多组学数据集成首先在样本级别上发生，然后在如图2a和图2b所描绘的变体、基因密码子编号、基因或通路级别或者其任何组合上发生。在图2a中所示的多组学索引集成的非限制性示例中，所摄入的多组学癌症数据选自包括以下项的组：单核苷酸变体(SNV)和小的插入和缺失(indel)(被表示为染色体编号、染色体位置、参考、替代等位基因-CPRA)、拷贝数变体(CNV)和RNA中确认的变体。SNV可以从包含体细胞VCF的SNV和小的插入和缺失被索引。在染色体区域上被调用的拷贝数变体(CNV)(例如，也使用高级癌症分析模块在基因级别上被映射)可以从拷贝数调用VCF被索引(CNV也在基因级别上被映射)。RNA-Seq确认的变体可以从(从高级癌症分析模块所得出的)RNA-Seq分析被获得。多组学索引可以被连结以回答复杂的查询(例如，获取SNV和小的插入和缺失相重叠的CNV增加和丢失，其表达于针对样本组的RNA中)。差异表达的基因可以从例如高级分析软件模块得出。

[0146] 根据各种实施例，被连结的多组学索引可以经由所选择的索引方法而被产生，索引方法诸如例如但不限于用于对拷贝数变体和确认的RNA变体(再次参见图2a)进行索引的KEYSxCPRA、KEYSxCNV、KEYSxCNV_RANGE、KEYSxCNV_GENE、KEYSxCPRA_RNA和KEYSxGENE_RNA。申请人已有利地发现，对多个信息流的交叉索引提供了例如查询多组学数据流的任意组合或个体流自身的能力、以及执行变体、基因密码子编号、基因、通路以及其他级别的实体链接的能力。

[0147] 参考图2a所示的示例，第一索引表210按照在具有KEYS样本ID 222的样本中出现的其CPRA 212(染色体214、位置216、参考218、替代等位基因220)描述了DNA中的单核苷酸多态性和小的插入和缺失。第二索引表230按照在具有KEYS样本ID 242的样本中出现的范围232(染色体234、起始236、结束238)描述了拷贝数变体(CNV)。第三索引表250按照在具有KEYS样本ID 262的样本中出现的RNA-Seq描述了DNA中的变体(CPRA)252变体(参见第一索引表210)。第四索引表270以其范围相对于DNA(CPRA)274中的单核苷酸多态性和小的插入和缺失描述了拷贝数变体CNV 272。

[0148] 参考图2b所示的示例，其提供了CPRAxTERM排名300，CPRAxTERM排序排名300由针对CPRA级别310、GENE_CODON级别312和GENE级别314上被聚合的注释(术语)的排名组成。公式320提供关于如何在GENE_CODON级别上针对CPRA来计算排名的示例。公式322提供关于如何在GENE级别上针对CPRA来计算排名的示例。第五索引表330提供通过GENE_CODON映射索引表的CPRA的示例。第六索引表340提供GENE_CODON级别注释索引表的示例。第七索引表350提供CPRA级别注释索引表的示例。

[0149] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以提供将所选择的一个或多个多组学数据索引的排名。在各种实施例中，排名可以在没有对可用癌症多组学数据相关联的过滤的情况下发生。如以上所讨论，可访问的数据可以包括例如变体、基因、通路、RNA序列确认的变体、差异表达的基因、高/低甲基化区域、被表达的蛋白质、拷贝数变体、结构性变体、基因融合、表型、家族史、注释、药物、临床试验纳入/排除标准、所得出的分析(例如，突变签名权重、微卫星重复位点、从影像学数据和图像本身所提取的特征和文献数据以及其嵌入)、以及机器学习模型预测及其特征(例如，微卫星不稳定状态和微卫星不稳定位点、所预测的主要原发部位以及按照其相对重要性的被标识为该模型的关键特征的改变、所预测的转移部位和模型关键特征、以及所预测的MHC I类和II类分子的新抗原结合亲和力)。在各种实施例中，不同的多组学流的任何组合或者个体的数据流可以基于用户查

询而被返回。

[0150] 图2b例如图示了注释的分层传播和由变体级CPRA x cpraTERM、密码子级CPRAxcodonTERM和基因级CPRA xgeneTERM注释的加权排名累积的变体(CPRA)的排名的示例。

[0151] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以提供对多个癌症注释源的集成和排名。这些多个癌症注释源可以包括例如FDA标签、NCCN指南、NCCN纲要生物标志物、临床试验、CIViC、DoCM、OncoKB、Mycancergenome、癌症药物基因组生物标志物数据库、TCGA、ICGC、COSMIC、NCI60、CCLE、DrugBank、ClinVar、HGMD、PGMD、PharmGKB、dbSNP、dbNSFP、1000Genomes、EXAC、CPDB、KEGG、BioCarta、BioCyc、Reactome、GenMAPP、MSigDB、Brenda、CTD、HPRD、GXD和BIND。

[0152] 根据各种实施例,多模式排名引擎(或模块)还可以进一步相关性学习引擎来集成例如注释源、文献数据、临床试验结果和良好表征的同类群组中的显著突变的基因(诸如TCGA),以学习个体患者和同类群组两者查询用例设置中的多组学数据的临床可行排名。在其他实施例中,学习到的排名可以基于具有未知临床显著性的改变的预测的致病性。

[0153] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以提供按照癌症基因组改变的临床可操作性、致病性、特征权重或频率方面的对癌症基因组改变的排名。根据各种实施例,排名模型可以通过训练监督学习模型而被得出,训练监督学习模型是通过学习对针对多组学癌症数据所提取的特征进行加权。对于变体(例如,在精确位置和特定密码子处)或基因(例如,考虑突变类型),这可以包括例如在FDA标签、NCCN指南、NCCN生物标志物纲要、ASCO指南、ESMO指南或其他顶级癌症指南中是否已经暗示基因中的变体/或改变类型的指示物,以及是否有特定药物的适应症/禁忌症;从其他癌症注释源(诸如例如,临床试验、OncoKB、Mycancergenome、CIViC、DoCM和癌症药物基因组生物标志物数据库)所提取的基因变体或改变类型的特征;从其他相关注释源中所提取的特征,诸如例如TCGA、TCGA显著突变的基因、COSMIC癌症基因普查、COSMIC、ICGC、Drugbank、Swissprot、dbNSFP、HGMD、PGMD、PharmGKB和ClinVar;来自HLI、HLI癌症、TCGA、COSMIC、ICGC、1000Genomes、EXAS、Gnomad的种群等位基因频率数据;来自从相关临床试验、PubMed、Medline、OMIM文章和其他医学文献中所取的文本中的嵌入;以及从医学文献中所提取的命名实体的嵌入。

[0154] 根据各种实施例,排名可以基于支持向量回归、提升树、以及对来自注释源的信息进行加权的其他机器学习模型,注释源诸如例如有FDA、NCCN指南、NCCN生物标志物纲要、编策的癌症基因、COSMIC、TCGA显著突变的基因、已知的热点、临床试验以及计算机预测的功能获得型/功能缺失型(loss/gain of function)得分(例如,CADD、FATHMM、SIFT、Polyphen)。

[0155] 根据各种实施例,三种学习排名方法被用于得出排名。这些方法包括逐点方法(例如,逻辑回归)、成对方法(例如,RankSVM、RankBoost)和基于列表方法(LambdaMart)。

[0156] 根据各种实施例,与其他文件(例如,医学文献)的排名相比,对变体和基因的排名可以被单独学习,其中单独的学习排名模型被训练为使用可以包括例如BM25、PageRank、RM3和针对文本文档的其他排名模型的加权变换特征集。

[0157] 根据各种实施例,真对变体和基因的排名可以单独被学习,或者作为深度和广度模式的一部分与针对其他文档类型的排名一起被学习。在一些实施例中,针对文本文档的

排名利用深度学习语言建模(LM)，通过给定查询的文档的概率来将项排名。根据各种实施例，深度学习语言模型可以是关于相关数据上被微调的变换器模型(例如，BERT、RoBERTa、XLnet、Albert)。这样的模型可以是大规模的、经预训练的语言模型嵌入。根据各个实施例，文档相关性可以使用文档的文本和时间部分例如通过得出特征的多个类而被生成，多个特征类包括例如从一组注释得出的实体特征和时间特征，命名实体识别(NER)和时间标注。

[0158] 根据各种实施例，为了提供附加的语义理解，深度学习方法(例如，深度语义相似度模型、卷积深度语义相似度模型、递归深度语义相似度模型、深度相关性匹配模型、交互孪生网络、词法和语义匹配网络、长短期记忆网络、变换器网络、词语嵌入方法、DeepRank)可以被用于通过首要使用从查询的原始文本和文档所自动学习到的特征来解决学习排名的特征工程任务。像这样，深度学习方法可以使用不同类型的神经网络，无论其是例如，卷积的还是递归的。

[0159] 如以上所讨论，根据各种实施例，排名可以包括针对癌症变体和基因的临床的排名。排名可以包括深度学习排名，其中深度学习排名可以从深度学习模型得出，该深度学习模型选自包括以下项的组：深度语义相似度模型、深度和广度模型、深度语言模型、学习到的深度学习文本嵌入、学习到的命名实体识别、孪生神经网络、以及前述项的组合。

[0160] 图4a图示了用于学习变体排名的广度和深度模型的示例。广度部分可以使用来自不同注释源的交叉积特征变换来有效地记忆稀疏特征以及其交互，而深度部分可以概括为先前看不见的特征交互和文献嵌入。

[0161] 图4b图示了依赖于针对生物医学数据的深度语义相似度模型(参见以上讨论)的学习排名引擎的示例。在图4所示的特定示例中，孪生网络被用于允许通过学习联合查询和文档嵌入来学习查询(Q)与相关文档(D^+)之间的语义相似度。相关性可以通过查询与文档嵌入之间的余弦相似度R(Q,D)而被估计。网络可以将对随机地被采样的负面文档D^-的交叉熵损失最小化：

[0162] $$\mathcal{L}(q,d^+,D^-) = -log\left(\frac{e^{\gamma\cdot\cos(\vec{q},\vec{d}^+)}}{\sum_{d\epsilon D} e^{\gamma\cdot\cos(\vec{q},\vec{d})}}\right)$$，其中$D=\{d^+\}\cup D^-$

[0163] 在排名模型被训练之后，文档嵌入可以被预先计算(例如，作为文档中词语的所有单位向量的质心)。在查询时间，在评估查询与联合潜在空间中的文档表示之间的相似度之前，查询向量嵌入可以被生成。注意，在图4b中所引用的特定查询和文档仅是示例性的，并且不以任何方式限制所提交的查询和所分析的文档的类型。

[0164] 根据各种实施例，全局排名可以针对临床可操作性(或当临床效用未知时的致病性)而被优化，并且被预先加载到索引中，从而(例如，历经top-K算法的)结果可以被重新排名，以进一步满足特定的信息需求。根据各种实施例，重新排名可以涉及使用语言建模或者来自标准信息检索模型(例如，PageRank、BM25、RM3)的加权的经变换特征。

[0165] 根据各种实施例，针对样本同类群组中的潜在生物标志物的排名可以通过首先学习多组学数据流(例如，如本文所讨论的DNA和RNA以及其他)的潜在空间表示、并且然后将表示聚类并且标识一组特征(例如，生物标志物)而被达成，该一组特征负责感兴趣的子同类群组之间的最大解纠缠。根据各种实施例，多组学无监督深度学习方法(例如，变分自动编码器)可以出于目的而被构造。根据各种实施例，深度生成对抗网络可以利用多个数据流

之间的循环损失而被构造。根据各种实施例，标准降维技术(例如，主成分分析、个体成分分析、流形学习)可以被用于将稀疏的、广泛的多组学数据变换为有意义的潜在空间。这些方法有利地可以增加用于检测多组学生物标志物的能力。

[0166] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以传播从更高级别的生物层级所学习到的排名，以通知更低级别的生物层级。例如，基因级别排名可以通知变体级别排名，在变体级别排名种关于各种癌症注释源中变体的发生的信息可能不可用。

[0167] 根据各种实施例，针对缺失注释的变体的排名可以被构造为针对基因和突变类型的排名的聚合。例如，聚合函数被学习，该聚合函数在给定这些方面的情况下预测总体相关性，在这之后常规的学习排名算法可以被应用于学习排名。

[0168] 根据各种实施例，临床上可操作的和致病性排名可以被预先加载到索引中，以提高检索速度。根据各种实施例，针对多组学流的特定组合而被学习的排名公式可以在索引检索时间被应用。

[0169] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以包括对针对特定用户查询所返回的结果的排名，返回结果可以取决于所查询的多组学数据流的组合，并且可以考虑到个体的和经组合的多组学数据流的临床相关性，响应于用户用户查询而基于用户偏好来变化。。

[0170] 根据各种实施例，排名可以由用户改变(例如，所返回的结果可以被提升或降级)。根据各种实施例，排名可以由来自用户的间接反馈(诸如例如，对特定的被返回结果的点击率和停留时间)来更改。

[0171] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以提供用于借助web交互性来收集用户反馈以改进结果的多组学排名。例如，变体、基因、通路、所得出的分析可以基于用户反馈而在所返回结果的列表中被提升或降级。根据各种实施例，附加编策信息可以被提供并且被保存在索引中。

[0172] 在各种实施例中，本文所描述的系统和方法可以提供接口(或者与接口的交互)以收集关于所返回结果的相关性的明确用户反馈(例如，用户点赞(give thumbs up)/提升/保存/保存以供报告/固定/导出特定结果，或者用户点灭(give thumbs down)/降级从所返回结果的列表删除结果)。

[0173] 在各种实施例中，本文所描述的系统和方法可以促进从搜索日志收集和分析隐式用户反馈(例如，分析点击、停留时间、查询序列、所返回结果的数目)。

[0174] 在各种实施例中，可以提供协作式搜索用户接口(或与之交互)，以允许多个用户协作地完善对多组学癌症改变的排名的质量(例如，在虚拟肿瘤板设置中)。

[0175] 如以上所讨论，根据各种实施例，本文所述的系统可以包括查询引擎，该查询引擎可以被配置为执行以下中的至少一项：接受用户查询，选择、聚合和汇总相关的多组学索引，以及返回个体样本和/或癌症样本同类群组的经排名的多组学改变。

[0176] 在各种实施例中，查询引擎可以是无状态服务器，该无状态服务器基于被预先计算和预连结的多组学索引文件的集合来接受用户查询(例如，作为HTTP POST请求)并且以结果的经排名列表来响应(例如，作为异步JSON)。在各种实施例中，查询引擎可以执行以下功能中的至少一个功能：(a)解析查询并且将用户意图分类(例如，用户是否想要变体、基因、通路、样本、单个样本数据、同类群组样本数据、样本与同类群组比较、同类群组与同类

群组比较、出版物、图像)；(b)提供查询自动纠正(例如，使用关于日志上被微调的自动纠正深度学习模型)，提供选择性的同义词展开和缩写展开，生成备选查询(例如，使用深度学习的经微调变换器模型)并且提供基于内容的建议(例如，对连续查询使用经微调的语言模型、使用利用经索引数据的模型)；(c)决定要使用的适当多组学索引的组合；(e)按结果与所预测的查询意图(例如，临床相关性和致病性—默认排名、某些查询的频率、其他查询的互信息量、特征权重等)的相关性，将结果排名；(f)汇总注释文档和医学文献(例如，使用深度学习汇总技术)；以及(g)处理来自UI的交互/反馈信号。在各种实施例中，查询引擎可以允许每个查询的亚秒级时延以及对数十万个并发用户的可缩放性。

[0177] 在图5a至图5b的示例工作流中图示了这些功能中的至少一些功能，图5a和图5b图示了查询引擎工作流，该查询引擎工作流起以下功能(1)产生同义词和缩写展开，(2)产生备选(类似)查询，(3)产生基于内容的建议并且提供查询自动完成和自动纠正功能，(4)将用户查询意图分类(例如，用户是否想要变体、基因、通路、样本、单个样本数据、同类群组样本数据、样本与同类群组比较、同类群组与同类群组比较、出版物、图像?)，(5)执行神经信息检索(例如，基于查询和经索引文档的联合嵌入)以及(6)提供文档的汇总(例如，多个源文本汇总)，这些汇总可以经由系统UI而被传递回给用户。根据各种实施例，主题特定的项嵌入可以被用于查询扩大，特别是以上(2)中的查询扩大。根据各种实施例，对于文本数据，神经信息检索模型可以考虑术语空间中的匹配以及潜在空间中的匹配两者。此外，针对例如变体、基因、通路、药物和癌症类型的命名实体识别模型也可以被集成来改进召回率。注意，在图5a和图5b中所引用的特定查询、数据和概要仅是示例性的，并且不以任何方式限制所提交的查询的类型、所分析的文档以及所产生的概要。例如，在图5a至图5b所示的特定示例工作流的情况下，给定该查询的特定参数，查询引擎可以得出结论，尽管TP53中的功能丧失事件在癌症中非常普遍，但是R248变体似乎不仅导致肿瘤抑制的丧失，而且还可以作为可以在小鼠模型中促进肿瘤发生(请参见注释源CIViC和癌症药物基因组生物标志物数据库[GDKB])的功能获得突变。

[0178] 如以上所讨论，根据各种实施例，本文所述的系统和方法可以使用在可用的生物医学文献和医学本体(例如，GO、UMLS、DO、MeSH、eVOC、HPO、MPO)上所训练的深度学习模型来促进查询术语扩大的集成。

[0179] 如以上所讨论，根据各种实施例，本文所述的系统可以促进神经信息检索模型的集成，旨在提供更好的语义理解能力，以供将文献、图像和注释进行排名。在各种实施例中，词语的分布式表示(例如，由word2vec生成的词语)可以被组合以生成针对查询和文档的嵌入，并且平均嵌入可以被用来生成有效的文档相似度检索。

[0180] 进行查询的排名的有效方法的示例是针对每个查询独立构建排名方案。但是，分别针对每个查询训练模型遭受缺少看不见的查询的有标签数据。然而，根据各种实施例，癌症基因组改变搜索引擎可以允许将查询类型分组并且具有关键临床重要性的特定查询子集进行微调排名(例如，按癌症改变的临床可操作性和致病性顺序返回癌症改变的查询、按基因的临床可操作性顺序返回基因的查询)。为了得出变体和基因的临床可操作性，可以使用查询和文档对的经手工加标签的语料库。在各种实施例中，结果的精度和召回率可以被测量。

[0181] 在各种实施例中，训练语料集可以包括由癌症分析员手动检查的全面的癌症病

例。

[0182] 在各种实施例中,手动训练语料库可以由例如癌症分析员/编策员构造。分析员/编策员可以检查例如,(1)在相同癌症类型的良好表征的同类群组(例如,TCGA、ICGC、内部同类群组)内显著突变的基因中的改变(>0.02来自MutSigCV的p或q值);(2)显著突变的基因的排名;(3)所检测到的突变是否与良好表征的同类群体相同(例如,错义(missense)、插入和缺失、无意义);(4)如果突变是错义的,则其是否发生在热点处;(5)具有该突变的来自良好表征的同类群组的具有该突变的患者数目;以及(6)在一些情况下,对具有突变的患者的突变、位置、结构和癌症类型进行的进一步检查。

[0183] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以提供通用搜索接口(如与许多不同的入口点相对)。在各种实施例中,所有知识(无论其是例如多组学癌症数据、样本、变体、基因、药物、通路、表型、医学文献、图像数据、所得出的癌症分析、用于预测肿瘤特性的机器学习模型及其特征、用户数据的上传等)可以通过相同的简单搜索接口可访问。

[0184] 如以上所讨论,根据各种实施例,本文所述的系统和方法可以为利用个体样本或者样本同类群组来工作的临床医生或研究人员提供关键可操作和重要的癌症改变、所得出的癌症分析以及质量控制指标的清单/终端。

[0185] 根据各种实施例,本文所述的系统和方法可以提供根据ACMG指南所报告的重要的癌症和遗传性癌症变体。

[0186] 根据各种实施例,本文所述的系统和方法可以提供动态被超链接的个体患者和同类群组报告,其中报告上的项中的至少一些项被超链接到多模式癌症搜索查询,癌症改变被排名。在各种实施例中,经超链接的报告内容可以基于用户出于报告目的而进行并且保存的查询而动态地被生成。

[0187] 根据各种实施例,本文所述的系统和方法允许将以下至少一项包括在由所保存的用于报告的用户查询生成的动态报告中:集成的多组学结果、可视化、图像、医学文献、高级癌症分析和来自癌症生物信息学管线的任何级别的数据(例如,测序覆盖率、碱基对变化类型的百分比、支持个体变体的测序读段的可视化)。

[0188] 根据各种实施例,本文所述的系统和方法可以作为具有两个因素认证和访问控制层的web服务来运行,以帮助确保每个客户端只能访问他们被授权访问的样本,并且没有分析跨对其的访问由不同实体控制的独立数据集地被执行。

[0189] 在各种实施例中,查询可以包括与特殊运算符组合的(在概念上可以是任意的)自然语言术语。在各种实施例中,查询可以包括语音到文本的模型。在各种实施例中,特殊运算符可以使得用户能够无歧义地引用某些信息(例如,特定客户端)或施加某些约束(例如,仅提供基因或通路作为结果)。在各种实施例中,运算符可以包括例如加号、减号、等号、和号、星号、引号、大括号、括号、花括号、反斜杠、正斜杠、冒号、分号、井号(#)、@符号(@)、波浪号(~)、等于号(=)、大于号(>)、小于号(<)、以及词语和、或、非、除了。在各种实施例中,查询由与特殊运算符组合的自然语言术语组成。在各种实施例中,特殊运算符可以使得用户能够无歧义地引用某些信息。

[0190] 图6图示了具有单个搜索框610的用户接口600的示例,搜索框610允许用户录入不同的查询并且接收经排名的结果。每个变体可以利用丰富的数据、,变体变体以及使用集成

的基因组变体浏览器(IGV)来查看突变和周围测序读段并且在UCSC基因组浏览器中探索变体的能力而被显示,这些丰富的数据包括例如变体质量控制、变体指标、与种群数据库相比的等位基因频率、治疗性药物注释、与癌症数据库和注释源的比较。

[0191] UI 600的部分620允许用户检查变体调用的位置和质量。染色体、位置和变体可以利用以不同于参考的颜色来突出显示的突变碱基而被列出。UCSC链接允许用户在基因组浏览器中查看变体(允许对变体的深入调查)。实际的测序读段可以使用IGV链接而被可视化,这将允许用户例如确定变体调用的可靠性、查看变体是否出现在混乱区域中或者调用是否由于测序伪像而不可靠。

[0192] UI 600的部分630列出了基因级别的信息。基因名称被列出,并且在被单击时,可以进行到有关变体的深入信息,包括基因概要、该变体在TCGA数据中的频率。像这样,用户可以调查变体是否被找到,以及变体在相同以及其他肿瘤类型中以什么频率被找到。针对该变体的临床试验以及其他相关的临床信息可以被显示。HGVS选项卡显示了蛋白质级别的变体。Ensembl选项卡显示了被用于映射蛋白质的转录本,并且dbSNP rsID也被列出。变体可以与健康种群中所发现的频率进行比较(参见图6中的"HLI健康等位基因频率")。PubMed选项卡链接到来自PubMed的科学文献中与该变体有关的相关论文。

[0193] UI 600的部分640可以允许用户执行变体调用的质量控制。如果RNA-Seq也被执行,则RNA-Seq等位基因分数被显示。肿瘤和正常等位基因分数和读段深度允许用户确定调用质量,以及正常血液中是否存在变体的任何证据。

[0194] UI 600的框650提供临床信息(如果可用)。

[0195] 在各种实施例中,本文描述的系统可以包括允许用户录入用户查询或使用用户查询的接口。在各种实施例中,本文描述的方法可以提供经由接口来录入用户查询或使用用户查询。如以上所讨论,在各种实施例中,用户查询可以是通过语音的。在各种实施例中,用户查询可以包括例如患者/个体ID号、同类群组名称/ID号、某个基因名称或基因符号、特定注释源、变体和/或表型。在各种实施例中,输入可以是复选框或可点击按钮,该复选框或可点击按钮将输出限制或过滤为序列,例如,变体、基因、表型数据、多组学数据流的特定组合以及统计上显著的变体、基因、通路。在各种实施例中,结果可以是可排序的、在适当时被指定为收藏夹或者被导出到另一程序或者被导出到被动态生成的报告。在各种实施例中,个体搜索项可以是可组合的。在各种实施例中,个体(或用户)可以使用附加的用户查询或过滤来在某个结果集内搜索附加信息。表1举例说明了期望信息的示例、示例用户输入和示例输出的非穷举列表。表1不是可由用户部署的查询的排他性或穷举性列表。

[0196]

用户所期望的信息类型	示例用户输入	示例输出
患者信息(医师的单独或所有患者)	@patientSeqID person	患者的癌症类型、测序深度、所选择的测序质量指标
具有 FDA 批准的疗法的体细胞突变	@PatientSeqlD fda	具有来自 FDA 标签的注释的基因和/或变体的经排名列表; 链接到 FDA 和其他注释源

具有 FDA 批准或专业指南疗法的体细胞突变，图 7	fda+nccn@PatientSeqID	具有来自 FDA 和 NCCN 的注释的基因和/或变体的经排名列表，链接到注释源
与临床试验匹配的体细胞突变	@PatientSeqID nonsilent genes	具有临床试验信息的基因和/或变体排名列表；链接到 clinicaltrials.gov 和其他注释源
特定癌症基因的体细胞突变	@PatientSeqID TP53	列出 TP53 基因中的所有体细胞变体
肿瘤突变负荷（TMB），图 8	@PatientSeqID afrac>0.05 tmb	表示外显子组中的非沉默突变除以外显子组大小（突变/MB）的数字，叠加在各种癌症类型的癌症基因组图谱（TCGA）同类群组上
突变签名 图 9	@PatientSeqID mutsig	将导致体细胞突变的底层突变过程分类的模式
微卫星不稳定性（MSI）状态	@PatientSeqID msi	评分系统，用于将位于被称为基因组微卫星的重复 DNA 序列中的体细胞插入和缺失分类
所有相关体细胞突变	@PatientSeqID nonsilent panel:reportable afrac>0.05	在被定义的癌症基因集中，肿瘤等位基因分数等于或高于 5％的经排名的非沉默体细胞突变的列表
遗传性的癌症风险变体	@PatientSeqID g nonsilent panel:hli-inh af<0.02	在被定义的基因集中，与遗传性癌症风险相关联并且在参考种群中变体频率低于 2％的非沉默种系变体列表
基于肿瘤 RNA 序列	@PatientSeqID	免疫疗法考虑的属性，例如，

[0197]

[0198]

的免疫概况	immuno	TMB、非沉默突变列表、新开放读段框架、HLA 分型
来自同一患者或具有相同肿瘤类型的不同患者的两个不同肿瘤样本中的独特体细胞变体	@PatientSeqlDl-@PatientSeqlD2	肿瘤 1 中存在但不在肿瘤 2 中的体细胞变体的经排名列表
两个或更多个不同肿瘤样本中的共同体细胞变体	@PatientSeqlDl @PatientSeqlD2	肿瘤 1 和肿瘤 2 中都存在的体细胞变体的经排名列表
将患者的体细胞变体与相同肿瘤类型的公共数据比较	@PatientSeqlDl vs cohort:tcga_paad	与相同肿瘤类型的 TCGA 同类群组中的每个患者相比较的患者的体细胞突变的图形表示
了解基于体细胞突变的受影响生物通路	@PatientSeqlD nons1lent pathways	受所标识的变体影响的生物通路的通路富集分析
比较同类群组中的每个患者的肿瘤突变负荷（TMB），图 10	@cohort:cohortlD tmb	显示同类群组中每个患者的 TMB（外显子组中的非沉默突变除以外显子组大小，以突变/MB 表达）
基因级别的变体概况以及比较同类群组的临床信息，图 11	@cohort:cohortlD panel:cgc nonsilent	基于同类群组中最频繁突变的基因或经定义的基因集，显示每个患者的体细胞突变，显示每个患者中的体细胞突变的数目和类型，并且与临床数据（如果可用）对准
基于体细胞突变和临床信息，将同类群组	@cohort:responders	每个患者在 EGFR 基因中的体细胞突变的基因和蛋白质

[0199]

进行分层，图 12	cohort:nonresponders egfr	预测，与患者对特定临床治疗的反应对准

[0200] 注意，表1中对附图的所有参考仅用于指导，并不意味着相对于用户所期望的信息类型，限制了相对用户输入和示例输出。例如，图7图示了根据各种实施例的利用特定语法（“fda+nccn@PatientSeqID”）获得的搜索结果的示例。

[0201] 此外，例如，图8a和图8b图示了根据各种实施例的以特定语法(“@PatientSeqID afrac>0.05tmb”)所获得的搜索结果的示例。具体地，图8b图示了在该特定示例中，导致肿瘤的总体肿瘤突变负荷的非沉默突变之一的显示。更详细地，图8a和图8b示出了利用以上引用的特定语法所获得的搜索结果的一个示例，其中用户希望仅计入等位基因分数大于5％的突变的肿瘤突变负荷值。癌症突变负荷然后可以被显示在按同类群组分组的癌症基因组图谱肿瘤突变值上的背景上。在肿瘤样本中发现的非沉默突变的类型数目也可以被显示在所图示的饼图中(参见图8b)。该显示允许用户快速评估潜在的癌症亚型、潜在的测序问题、以及对肿瘤突变负荷值背后的因素的总体评估。所示饼图的中心区域显示了非沉默突变的总计数。通过在饼图的中心区域之外的参考(饼形图旁边所提供的图例)，非沉默突变的总数再次被进一步细分为已经被标识的非沉默突变类型。在许多癌症中(如本示例所见)，错义突变可能是最频繁的。如果微卫星不稳定的移码突变构成了很大一部分突变，则饼形图显示功能允许快速检查该参数。各种测序伪像也可能导致通常在该癌症中不常见的突变类型的高百分比。饼图显示功能还可以被用于确定肿瘤突变负荷的临床相关性。一些免疫治疗剂对主要由移码突变或其他特定突变类型组成的肿瘤作用最佳。像这样，饼图显示功能将允许用户快速评估那些可能性。在图表下方，接口产生所有非沉默变体的经排名列表，其中大于5％的等位基因分数被显示(图8b由于空间不足而显示单个命中)。

[0202] 进一步地，例如，图9图示了根据各种实施例的从用户查询所返回的搜索结果的示例。具体地，图9图示了利用特定语法“@PatientSeqID mutsig”所获得的搜索结果的非限制性示例。突变签名是肿瘤中所有基因中发生的碱基对变化的总体模式。突变签名可以通过计算上下文中所有碱基对的变化得出整体突变发生模式来得出。易于使用的突变签名定义可以在https://cancer.sanger.ac.uk/cosmic/signatures处找到。突变签名的标识可以指导治疗、可以帮助解释肿瘤的底层原因并且可以帮助解决显著性未知的变体。因此，突变签名对于分析肿瘤的整体特性很重要。

[0203] 图9的部分A在围绕突变的碱基对的上下文中显示了碱基对替换类型的类型(即，C>A、C>G、C>T、T>A、T>C、T>G)的X-Y图(3bp，被显示在X轴上)。每个突变类型的频率被绘制在Y轴上。在该示例情况下，图与COSMIC所标识的签名进行比较，以得出肿瘤的整体突变签名。

[0204] 部分B在饼形图上显示了在肿瘤中发现的总体突变签名的百分比。该显示可以允许用户确定肿瘤中的主要签名以及所标识的任何次要签名。在该示例中，来自黑色素瘤肿瘤，所显示的主要签名是S7，这与文献一致。如果所显示的突变签名不是针对该癌症类型所预期的，则用户可以进行进一步的调查。

[0205] 突变签名还可以帮助指导临床决策。例如，考虑乳腺癌和卵巢癌中的BRCA1/2突变。PARP抑制剂可以被用于BRCA1/2突变的乳腺癌和卵巢癌患者。COSMIC签名3的特征可以在于BRCA或通路基因的缺陷，从而即使在缺少被标识的突变的情况下，在肿瘤中标识签名3仍然指示BRCA突变过程。如果肿瘤包含未知显著性BRCA突变，则测定签名3的存在可以帮助确定突变是否起作用。在这两种情况下，PARP抑制剂的潜在益处可以被探索。

[0206] 此处可访问的另一功能是96个三元组中的每个三元组的重构权重(未示出)。

[0207] 另外，例如，图10图示了根据各种实施例的从用户查询所返回的搜索结果的示例。具体地，图10图示了利用特定语法“cohort:CohortID tmb”所获得的搜索结果的非限制性示例。针对该情况的查询可以是标识同类群组中的肿瘤突变负荷。同类群组中的每个肿瘤

的肿瘤突变负荷(TMB,突变/mb)(具有与其相关联的数字TMB值的圆圈)可以通过癌症基因组图谱(该图上其余的和大多数圆圈,未参考任何相关联的TMB值)而与来自相同癌症类型(在该情况下,胰腺癌-PAAD)的肿瘤的TMB比较。TMB被表示在Y轴上,这允许用户查看同类群组中所标识的TMB是否与有关该癌症的先验知识一致。针对PAAD的TCGA中位数被示出为框中间的水平线。使用箱线图和须状图的表示,允许用户查看同类群组样本图在TCGA中发现的平均范围还是异常范围内。

[0208] 参考图10,提供了同类群组TMB图表500,其中TMB 510被表示在Y轴512上。同类群组中每个肿瘤的肿瘤突变负荷(TMB,突变/mb)是第一点520,第一点520具有与之关联的相关数字TMB值522。通过由第二点530表示的癌症基因组图谱,那些值与来自相同癌症类型(在该情况下,胰腺癌-PAAD)的肿瘤的TMB进行比较,第二点530不具有与其相关联的TMB值,并且在本示例中,形成大部分捕获点。

[0209] 此外,例如,图11图示了根据各种实施例的从用户查询返回的搜索结果的示例。具体地,图11显示了响应于要求汇总特定同类群组中的癌症基因普查组套的非沉默突变的用户查询“cohort:CohortID panel:cgc nonsilent”而在样本同类群组中被检查的多个基因组改变和临床信息的集成概要。有效地,针对这种情况的查询可以是标识给定同类群组中的样本是否具有相同数目和类型的突变。每个肿瘤样本可以被显示在列中,每个基因被显示在行中,并且可用的临床信息可以被添加到表中。该图可以由所显示的临床参数中的任何临床参数分层。该图可以按照同类群组中最频繁突变的癌症基因初步地被排序(如图所示),并且显示基因级别的频率。突变类型(例如,错义、无意义、移码)可以按变体的类型使用不同的框颜色变体来标识(参见图11的部分B)。在所示的示例中,驱动基因(NRAS)是如预期的错义突变。还可以显示每个样本的总突变计数,用户可以使用其信息来将图排序。该显示特征允许用户对同类群组执行深入分析,以及标识任何个体样本的特定改变。在该图中可以看到突变的共现或互斥。单独突变可以在图表下方列出(未示出)。

[0210] 在图11所图示的情况下,部分A图示了最左端的样本具有最高量的突变。在该同类群组之中,突变类型相当一致。在一些情况下,可能会观察到具有极高突变计数和高移码类型突变的样本。该观察结果可能需要更多的探索,以确定样本是否是微卫星不稳定的或存在伪像的。此外,左起第三个的样本不具有其余样本所具有的NRAS突变。但是,突变的数目和类型与该同类群组的其余不同。该观察可能需要更彻底的探索,以确定该差异是伪像还是生物学的。部分C图示了可以使用临床数据被排序的突变表图。

[0211] 进一步地,例如,图12图示了根据各种实施例的从用户查询返回的搜索结果的示例。具体地,图12示出了利用特定语法“cohort:responders cohort:nonresponders egfr”所获得的搜索结果的非限制性示例,其中用户希望比较两个子同类群组中的基因EGFR突变:反应者和非反应者。经排名的个体突变可在下方列出(该图中未示出)。在该示例中,部分A提供了两个同类群组(同类群组反应者与同类群组非反应者)中的种系/体细胞突变的EGFR基因级别的示意图。部分B提供了3D蛋白结构,突出显示了受聚集在两个同类群组的药物(吉非替尼(gefitinib))结合部位附近的热点突变影响的位置。

[0212] 图13是图示了在其上可以实现本教导的实施例或实施例的部分的计算机系统1000的框图。在本教导的各种实施例中,计算机系统1000可以包括用于传达信息的总线1002或其他通信机制,以及与总线1002耦合用于处理信息的处理器1004。在各种实施例中,

计算机系统1000还可以包括存储器1006，存储器1006可以是随机存取存储器(RAM)或其他动态存储设备，存储器1006耦合到总线1002用于确定要由处理器1004执行的指令。存储器1006还可以被用于在执行要由处理器1004执行的指令期间，存储临时变量或其他中间信息。在各种实施例中，计算机系统1000还可以包括耦合到总线1002的只读存储器(ROM)1008或其他静态存储设备，用于存储针对处理器1004的静态信息和指令。诸如磁盘或光盘的存储设备1010可以被提供并且耦合到总线1002用于存储信息和指令。

[0213] 在各种实施例中，计算机系统1000可以经由总线1002耦合到显示器1012(诸如阴极射线管(CRT)或液晶显示器(LCD))用于向计算机用户显示信息。包括字母数字和其他键的输入设备1014可以耦合到总线1002，用于向处理器1004传达信息和命令选择。另一类型的用户输入设备是光标控件1016(诸如鼠标、轨迹球或光标方向键)，用于向处理器1004传达方向信息和命令选择以及用于控制显示器1012上的光标移动。该输入设备1014通常具有在两个轴(第一轴(即，x)和第二轴(即，y))上的两个自由度，这两个自由度允许设备平面中指定位置。但是，应理解，本文还考虑了允许3维(x、y和z)光标移动的输入设备1014。本文中更详细地讨论了关于超出在此讨论的能力的显示和输入设备(或者本文中也使用的接口)。

[0214] 与本教导的某些实现一致，响应于处理器1004执行存储器1006中所包含的一个或多个指令的一个或多个序列，结果可以由计算机系统1000提供。这样的指令可以从另一计算机可读介质或计算机可读存储介质(诸如存储设备1010)被读取到存储器1006中。执行存储器1006中所包含的指令的序列可以使处理器1004执行本文所述的过程。备选地，可以使用硬连线电路来代替软件指令或者与软件指令结合使用来实现本教导。因此，本教导的实现方式不限于硬件电路装置和软件的任何特定组合。

[0215] 如本文中使用的并且将在以下更详细地讨论的术语“计算机可读介质”(例如，数据存储库、数据存储装置等)或“计算机可读存储介质”指代参与向处理器1004提供指令以供执行的任何介质。这样的介质可以采取许多形式，包括但不限于非易失性介质、易失性介质和传输介质。非易失性介质的示例可以包括但不限于光学、固态、磁性盘，诸如存储设备1010。易失性介质的示例可以包括但不限于动态存储器，诸如存储器1006。传输介质的示例可以包括但不限于同轴电缆、铜线和光纤，包括构成总线1002的线。

[0216] 计算机可读介质的常见形式包括例如软盘、软性盘、硬盘、磁带或任何其他磁性介质、CD-ROM、任何其他光学介质、打孔卡、纸带、具有孔模式的任何其他物理介质、RAM、PROM和EPROM、FLASH-EPROM、任何其他存储器芯片或盒式磁带、或者计算机可以从其读取的任何其他有形介质。以下提供了有关介质的进一步讨论。

[0217] 除了计算机可读介质之外，指令或数据还可以被提供为通信装置或系统中所包括的传输介质上的信号，以向计算机系统1000的处理器1004提供一个或多个指令的序列以供执行。例如，通信装置可以包括具有指示指令和数据的信号的收发器。该指令和数据被配置为使一个或多个处理器实现本文的公开中概述的功能。数据通信传输连接的代表性示例可以包括但不限于电话调制解调器连接、广域网(WAN)、局域网(LAN)、红外数据连接、NFC连接等。以下提供了有关数据通信的进一步讨论。

[0218] 应当理解，本文所述的包括流程图、图解和所附公开的方法可以使用作为独立设备的计算机系统1000而被实现或者在共享计算机处理资源的分布式网络(例如云计算网络)上被实现。

[0219] 还应当理解，在某些实施例中，机器可读存储设备被提供用于存储非瞬态机器可读指令，该非瞬态机器可读指令用于执行(execute)或执行(carry out)本文描述的方法。机器可读指令可以控制本文所描述的系统和方法的所有方面。此外，机器可读指令可以最初被加载到存储器模块中，或者经由云或经由API而被访问。

[0220] 在各种实施例中，本文所描述的系统和方法可以包括数字处理设备或者使用数字处理设备。在各种实施例中，数字处理设备可以包括执行设备功能的一个或多个硬件中央处理单元(CPU)或者通用图形处理单元(GPGPU)。在各种实施例中，数字处理设备还包括被配置为执行可执行指令的操作系统。在各种实施例中，数字处理设备可以可选地连接到计算机网络。在各种实施例中，数字处理设备可以可选地连接到互联网，使得其访问万维网。在各种实施例中，数字处理设备可以可选地连接到云计算基础设施。在各种实施例中，数字处理设备可以可选地连接到内联网。在各种实施例中，数字处理设备可以可选地连接到数据存储设备。

[0221] 根据各种实施例，作为非限制性示例，合适的数字处理设备可以包括服务器计算机、台式计算机、膝上型计算机、笔记本计算机、子笔记本计算机、上网本计算机、上网本平板计算机、手持式计算机、互联网设备、移动智能电话、平板计算机和个人数字助理。本领域普通技术人员将认识到，许多智能电话适合于在本文所述的系统中使用。本领域普通技术人员还将认识到，具有可选的计算机网络连接性的精选电视、视频播放器和数字音乐播放器适合于在本文所述的系统中使用。合适的平板计算机包括本领域普通技术人员已知的具有书册、板和可转换配置的平板计算机。

[0222] 在各种实施例中，数字处理设备包括被配置为执行可执行指令的操作系统。操作系统可以是例如包括程序和数据的软件，其管理设备的硬件并且提供用于执行应用的服务。本领域普通技术人员将认识到，作为非限制性示例，合适的服务器操作系统包括FreeBSD、OpenBSD、Net BSD、Linux、Apple®Mac OSXServer®、Oracle® Solaris®、WindowsServer®和Novell® NetWare®。本领域普通技术人员将认识到，作为非限制性示例，合适的个人计算机操作系统包括Microsoft® Windows®、Apple®Mac OSX®、UNIX®和类似UNIX的操作系统，诸如GNU/Linux®。在各种实施例中，操作系统由云计算提供。本领域普通技术人员还将认识到，作为非限制性示例，合适的移动智能电话操作系统包括Nokia® Symbian® OS、Apple® iOS®、Research InMotion®Black BerryOS®、Google® Android®、Microsoft®Windows Phone®OS、Microsoft®WindowsMobile®OS、Linux®和Palm® WebOS®。

[0223] 在各种实施例中，设备包括存储和/或存储器设备。存储和/或存储设备是用于临时或永久地存储数据或程序的一个或多个物理装置。在各种实施例中，设备是易失性存储器，并且要求电力来维持所存储的信息。在各种实施例中，设备是非易失性存储器，并且在数字处理设备不通电时保留所存储的信息。在各种实施例中，非易失性存储器包括闪存。在一些实施例中，非易失性存储器包括动态随机存取存储器(DRAM)。在各种实施例中，非易失性存储器包括铁电型随机存取存储器(FRAM)。在各种实施例中，非易失性存储器包括相变

型随机存取存储器(PRAM)。在各种实施例中,设备是存储设备,作为非限制性示例,包括CD-ROM、DVD、闪存设备、磁盘驱动器、磁带驱动器、光盘驱动器和基于云计算的存储装置。在各种实施例中,存储和/或存储器设备是诸如本文所公开的那些设备的组合。

[0224] 在各种实施例中,数字处理设备包括用于向用户发送视觉信息的显示器。在各种实施例中,显示器是阴极射线管(CRT)。在各种实施例中,显示器是液晶显示器(LCD)。在各种实施例中,显示器是薄膜晶体管液晶显示器(TFT-LCD)。在各种实施例中,显示器是有机发光二极管(OLED)显示器。在各种实施例中,在OLED显示器上是无源矩阵OLED(PMOLED)或有源矩阵OLED(AMOLED)显示器。在各种实施例中,显示器是等离子显示器。在各种实施例中,显示器是视频投影仪。在各种实施例中,显示器是诸如本文所公开的那些设备的组合。

[0225] 在各种实施例中,数字处理设备包括用于从用户接收信息的输入设备。在各种实施例中,输入设备是键盘。在各种实施例中,输入设备是指点设备,作为非限制性示例,包括鼠标、轨迹球、轨迹板、操纵杆、游戏控制器或手写笔。在各种实施例中,输入设备是触摸屏或多点触摸屏。在各种实施例中,输入设备是麦克风,用于捕获语音或其他声音输入。在各种实施例中,输入设备是视频相机或其他传感器,用于捕获运动或视觉输入。在各种实施例中,输入设备是Kinect、Leap Motion等。在各种实施例中,输入设备是诸如本文所公开的那些设备的组合。

[0226] 在各种实施例中,本文所公开的系统可以包括一个或多个非瞬态计算机可读存储介质,并且本文的方法可以在一个或多个非瞬态计算机可读存储介质上运行,该非瞬态计算机可读存储介质利用程序被编码,该程序包括由可选地被联网的数字处理设备的操作系统可执行的指令。在各种实施例中,计算机可读存储介质是数字处理设备的有形组件。在各种实施例中,计算机可读存储介质从数字处理设备可选地可移除。在各种实施例中,作为非限制性示例,计算机可读存储介质包括CD-ROM、DVD、闪存设备、固态存储器、磁盘驱动器、磁带驱动器、光盘驱动器、云计算系统和服务等。在各种实施例中,程序和指令被永久地、基本上永久地、半永久地或者非临时地编码在介质上。

[0227] 在各种实施例中,本文所公开的系统和方法可以包括至少一个计算机程序或者使用至少一个计算机程序。计算机程序包括在数字处理设备的CPU中可执行的、被编写为执行指定任务的指令序列。计算机可读指令可以被实现为执行特定任务或实现特定抽象数据类型的程序模块,诸如功能、对象、应用编程接口(APi)、数据结构等。本领域普通技术人员将认识到,可以以各种语言的各种版本来编写计算机程序。

[0228] 计算机可读指令的功能可以根据需要在各种环境中被组合或分布。在各种实施例中,计算机程序包括一个指令序列。在各种实施例中,计算机程序包括多个指令序列。在各种实施例中,计算机程序从一个位置被提供。在各种实施例中,计算机程序从多个位置被提供。在各种实施例中,计算机程序包括一个或多个软件模块。在各种实施例中,计算机程序部分或整体包括一个或多个web应用、一个或多个移动应用、一个或多个独立应用、一个或多个web浏览器插件、扩展、加载项、或附加项、或者其组合。

[0229] 在各种实施例中,计算机程序包括web应用。本领域普通技术人员将认识到,在各种实施例中,web应用利用一个或多个软件框架和一个或多个数据库系统。在各种实施例中,web应用基于诸如Microsoft®.NET或Ruby on Rails(RoR)的软件框架而被创建。在各种实施例中,web应用利用一个或多个数据库系统,作为非限制性示例,数据库系统包括关

系型、非关系型、面向对象、关联和XML数据库系统。在各种实施例中，作为非限制性示例，合适的关系数据库系统包括Microsoft®SQL Server、mySQL™和Oracle®。本领域普通技术人员还将认识到，在各种实施例中，web应用以一个或多个语言的一个或多个版本被编写。web应用可以以一个或多个标记语言、表示定义语言、客户端脚本语言、服务器端编码语言、数据库查询语言或其组合被编写。在各种实施例中，web应用在某种程度上以诸如超文本标记语言(HTML)、可扩展超文本标记语言(XHTML)或可扩展标记语言(XML)的标记语言被编写。在各种实施例中，web应用在某种程度上以诸如级联样式表(CSS)的表示定义语言来编写。在各种实施例中，web应用在某种程度上以诸如异步Javascript和XML(AJAX)、Flash®ActionScript、Javascript或Silverlight®的客户端脚本语言被编写。在各种实施例中，web应用在某种程度上以服务器端编码语言(诸如，主动服务器页面(ASP)、Perl、Java™、Java服务器页面(JSP)、超文本预处理器(PHP)、Python™、Ruby、Tel、Smalltalk、WebDNA®或Groovy)被编写。在各种实施例中，web应用在某种程度上以诸如结构化查询语言(SQL)的数据库查询语言来编写。在各个实施例中，web应用集成了企业服务器产品，诸如IBM® LotusDomino®。在各种实施例中，web应用包括媒体播放器元素。在各种实施例中，媒体播放器元素利用许多合适的多媒体技术中的一个或多个中的一个或多个合适的多媒体计数，作为非限制性示例，多媒体技术包括Adobe® Flash®、HTML 5、Apple® QuickTime®、Microsoft® Silverlight®、Java™和Unity®。

[0230] 在各种实施例中，计算机程序包括被提供给移动数字处理设备的移动应用。在各种实施例中，移动应用在制造时被提供给移动数字处理设备。在各种实施例中，移动应用经由本文描述的计算机网络而被提供给移动数字处理设备。

[0231] 移动应序可以使用本领域已知的硬件、语言和开发环境，通过本领域普通技术人员已知的技术被创建。本领域普通技术人员将认识到，移动应用可以以若干语言被编写。作为非限制性示例，合适的编程语言包括C、C++、C#、Objective-C、Java™、Javascript、Pascal、Object Pascal、Python™、Ruby、VB.NET、WML以及具有或没有CSS的XHTML/HTML、或者其组合。

[0232] 合适的移动应用开发环境若干源可得。作为非限制性示例，可商购的开发环境包括AirplaySDK、alcheMo、Appcelerator®、Celsius、Bedrock、Flash Lite、.NET Compact Frame-work、Rhomobile和WorkLight Mobile Platform。其他开发环境免费可用，作为非限制性示例，包括Lazarus、MobiFlex、MoSync和Phonegap。此外，移动设备制造商还分发软件开发人员工具包，作为非限制性示例，包括iPhone和iPad(iOS)SDK、Android™SDK、BlackBerry®SDK、BREW SDK、Palm®OS SDK、Symbian SDK、webOS SDK和Windows® Mobile SDK。

[0233] 本领域普通技术人员将认识到，若干商业论坛可用于分发移动应用，作为非限制性示例，包括Apple®App Store、Google®Play、Chrome WebStore、BlackBerry® App World、用于Palm设备的App Store、用于webOS的App Catalog、用于Mobile的Windows®Marketplace、用于Nokia®设备的Ovi Store、Samsung®Apps和

Nintendo DSi Shop。

[0234] 在各种实施例中，计算机程序包括独立应用，独立应用是作为独立计算机进程而不是现有进程的加载项(例如，不是插件)而运行的程序。本领域普通技术人员将认识到，独立的应用经常是经编译的。编译器是(多个)计算机程序，该(多个)计算机程序将以编程语言所编写的源代码转换为二进制目标代码，诸如汇编语言或机器代码。作为非限制性示例，合适的编译编程语言包括但C、C++、Objective-C、COBOL、Delphi、Eiffel、Java™、Lisp、Python™、Visual Basic和VB.NET或者其组合。编译通常至少部分地被执行，以创建可执行程序。在各种实施例中，计算机程序包括一个或多个可执行的经编译应用。

[0235] 在各个实施例中，计算机程序包括web浏览器插件(例如，扩展等)。在计算中，插件是向较大的软件应用添加特定功能性的一个或多个软件组件。软件应用的制作方支持插件，以使得第三方开发人员能够创建扩展应用的能力，以支持容易地添加新功能并且减少应用的大小。当受支持时，插件使得自定义软件应用的功能性成为可能。例如，插件在web浏览器中通常被用于播放视频、生成交互性、扫描病毒以及显示特定文件类型。本领域普通技术人员将熟悉若干web浏览器插件，包括Adobe® Flash® Player、Microsoft® Silverlight®和Apple® QuickTime®。在各种实施例中，工具栏包括一个或多个web浏览器扩展、加载项或附加项。在各种实施例中，工具栏包括一个或多个资源管理器栏、工具栏或桌面栏。

[0236] 本领域普通技术人员将认识到，若干插件框架可用，使得以各种编程语言开发插件成为可能，编程语言包括但不限于C++、Delphi、Java™、PHP、Python™和VB.NET或者其组合。

[0237] Web浏览器(也被称为Internet浏览器)是软件应用，被设计用于与联网的数字处理设备一起使用来检索、呈现和遍历万维网上的信息资源。作为非限制性示例，合适的web浏览器包括Microsoft®Internet Explorer®、Mozilla® Firefox®、Google® Chrome、Apple® Safari®、OperaSoftware® Opera®和KDE Konqueror。在各种实施例中，web浏览器是移动web浏览器。移动web浏览器(也被称为微浏览器、微型浏览器和无线浏览器)被设计用于在移动数字处理设备上使用，作为非限制性示例，移动数字处理设备包括手持式计算机、平板计算机、上网本计算机、子笔记本计算机、智能电话和个人数字助理(PDA)。作为非限制性示例，合适的移动web浏览器包括Google® Android®浏览器、RIMBlackBerry®Browser、Apple®Safari®、Palm®Blazer、Palm® WebOS®Browser、用于移动设备的Mozilla® Firefox®、Microsoft®Internet Explorer®Mobile、Amazon® Kindle®Basic Web、Nokia®Browser、OperaSoftware® Opera® Mobile和Sony PSP™浏览器。

[0238] 在各种实施例中，本文所公开的系统和方法包括软件、服务器和/或数据库模块，或且将其使用并入根据本文所公开的各种实施例的方法中。软件模块可以通过本领域普通技术人员已知的技术、使用本领域已知的机器、软件和语言而被创建。本文所公开的软件模

块以多种方式被实现。在各种实施例中，软件模块包括文件、代码段、编程对象、编程结构、或者其组合。在进一步的各种实施例中，软件模块包括多个文件、多个代码段、多个编程对象、多个编程结构、或者其组合。在各种实施例中，作为非限制性示例，一个或多个软件模块包括web应用、移动应用和独立应用。在各种实施例中，软件模块在一个计算机程序或应用中。在各种实施例中，软件模块在多于一个计算机程序或应用中。在各种实施例中，软件模块被托管在一个机器上。在各种实施例中，软件模块被托管在多于一个机器上。在各个实施例中，软件模块被托管在云计算平台上。在各种实施例中，软件模块被托管在一个位置中的一个或多个机器上。在各种实施例中，软件模块被托管在多于一个位置中的一个或多个机器上。

[0239] 在各种实施例中，本文所公开的系统和方法包括一个或多个数据库，或者将其使用并入根据本文所公开的各种实施例的方法中。本领域普通技术人员将认识到，许多数据库适合用于存储和检索用户、查询、记号和结果信息。在各种实施例中，作为非限制性示例，合适的数据库包括关系型数据库、非关系型数据库、面向对象的数据库、对象数据库、实体关系模型数据库、关联数据库和XML数据库。其他非限制性示例包括SQL、PostgreSQL、MySQL、Oracle、DB2和Sybase。在各种实施例中，数据库是基于互联网的。在进一步的Web.中，作为非限制性示例，合适的web浏览器包括Microsoft® Internet Explorer®、Mozilla® Firefox®、Google® Chrome、Apple® Safari®、Opera Software® Opera®和KDE Konqueror。在各种实施例中，web浏览器是移动web浏览器。移动web浏览器(也被称为微浏览器、微型浏览器和无线浏览器)被设计用于在移动数字处理设备上使用，作为非限制性示例，移动数字处理设备包括手持式计算机、平板计算机、上网本计算机、子笔记本计算机、智能电话和个人数字助理(PDA)。作为非限制性示例，合适的移动web浏览器包括Google® Android®浏览器、RIM BlackBerry® Browser、Apple® Safari®、Palm® Blazer、Palm® WebOS® Browser、用于移动设备的Mozilla®Firefox®、Microsoft® Internet Explorer® Mobile、Amazon® Kindle® Basic Web、Nokia® Browser、OperaSoftware® Opera®Mobile和Sony PSP™浏览器。

[0240] 在各种实施例中，数据库是基于web的。在各种实施例中，数据库是基于云计算的。在其他实施例中，数据库基于一个或多个本地计算机存储设备。

[0241] 在各种实施例中，本文所公开的系统和方法包括用于防止未经授权的访问的一个或多个特征。安全性措施可以例如保护用户的数据。在各种实施例中，数据被加密。在各种实施例中，对系统的访问需要多因素认证和访问控制层。在各种实施例中，对系统的访问需要两步认证(例如，基于web的接口)。在各种实施例中，除了用户名和密码之外，两步认证还要求用户输入发送到用户的电子邮件或手机的访问码。在一些实例中，用户在输入正确的用户名和密码失败之后，被锁定在帐户之外。在各种实施例中，本文所公开的系统和方法还包括用于保护用户的基因组及其在任何基因组中的搜索的匿名性的机制。

[0242] 在各个实施例中，本文所述的系统和方法可以通过以下来协助肿瘤学家在病例审查期间或者在虚拟肿瘤讨论会期间的协作式环境中得出临床见解：允许在癌症生物信息学管线的任何级别处探查患者或患者集合的数据、验证哪些癌症改变是真实的并且不表示测序伪像、报告质量控制值、集成多组学数据流和高级分析来提供关键仪表板或者癌症特性

和发现的“不容错过”清单，以及为所返回的每个经排名结果提供临床、预后、诊断和治疗信息。在各种实施例中，本文描述的多组学癌症搜索向医师提供“增强的智能”来帮助临床决策。

[0243] 根据各种实施例，本文描述的系统和方法的使用可以包括临床医生作为用户。这些用户可以使用本文所述的系统和方法来执行对肿瘤(和正常)基因组中的药物靶标和关键改变的全面报告。

[0244] 根据各种实施例，本文所述的系统和方法可以被用于虚拟肿瘤讨论会。根据各种实施例，本文所述的系统和方法可以由单独临床医生用作不容错过的重要肿瘤性质的核对清单，以及对肿瘤科医生机构内或全球可用的临床试验的检查。根据各种实施例，本文所述的系统和方法可以由肿瘤科医生在患者-肿瘤科医生访问对话期间使用。在各种实施例中，多个临床医生可以使用将临床上可操作的和致病性的癌症变化查询、可视化、重新排名的协作式功能，帮助在虚拟分子肿瘤讨论会期间来导航可用的表型的、影像学和文献数据，以确定最佳的诊断和治疗。本文所述的系统和方法可以解决的问题的一些非限制性示例可以包括，临床上相关的癌症变体是什么?存在潜在的疗法(经FDA批准的、NCNN、临床试验)吗?肿瘤中标识的突变是真实的吗?其由高质量的测序读段支持吗?该突变处于难以测序的区域吗?它仅存在于肿瘤中而不存在于正常组织中吗?它在RNA中被表达吗?该突变起作用吗?全局肿瘤性质、肿瘤突变负荷或微卫星不稳定性是什么?系统可以显示多个指标，该多个指标可以被用于确定总体质量和单个变体的质量。根据各种实施例的系统和方法可以提供用于将患者的突变与先前在诸如癌症基因组图谱(TCGA)的公共数据集中已经被描述的突变比较。根据各种实施例的系统和方法可以提供用于比较同一患者的多个活检。

[0245] 在各种实施例中，本文描述的系统和方法的用户可以包括生物制药或学术研究人员，生物制药或学术研究人员然后可以执行例如同类群组肿瘤概况分析来表征具有良好/不良预后的患者、反应者/无反应者的遗传概况，执行质量控制检查、药物靶标标识，关于潜在药物反应生物标志物将同类群组分层、以及在对附加验证或测试同类群组的更广泛的分析之前进行快速且迭代的假设生成。在各种实施例中，可以将同类群组分层的经排名的生物标志物、该生物标志物的统计显著性以及概要可视化可以由系统返回。在各种实施例中，验证查询可以由搜索引擎建议来执行鲁棒的算法和统计验证。在各种实施例中，系统可以经由所提议的查询精化来自动建议迭代假设精化。

[0246] 在各种实施例中，本文所述的系统和方法可以例如：标识与存活、抗性、反应相关的蛋白质、通路、突变过程；深入研究一个组中所找到的任何差异；与其他数据集比较；基于质量控制参数中的一个质量控制参数，检查同类群组质量控制来确保同类群组分析可靠并且不偏斜；调查任何异常结果，以确保它们不是由于系统性问题；向下钻取单个样本、离群值或异常结果，以确保它是真实结果；进一步探索并且快速获得分析的统计显著性；执行多目标数据探索；以及针对潜在疗法搜索文献和注释源。标准的生物信息学分析通常不给予使用领域知识来交互式地查询数据和精化假设的能力。内部系统经常基于数据库系统，而不基于能够提供相关性排名、执行多个信息流(例如，基因组的、转录组的、注释、文献)的集成、并且包括相关的内置机器学习模型搜索索引(诸如本文所讨论的索引)。

[0247] 如上所讨论，根据各种实施例，本文所述的系统和方法可以被配置为提供被动态超链接的个体患者和同类群组的变体报告，其中报告上的所有项被超链接到多模式癌症搜

索查询。在各种实施例中，被超链接的报告内容基于用户进行和突出显示、并且出于报告目的保存的查询而动态地被生成。

[0248] 如上所讨论，根据变体实施例，本文所述的系统和方法可以被配置为拥有专家审阅能力，从而给予用户选择哪些查询结果用于被超链接的实时报告生成的能力。

[0249] 在各种实施例中，动态报告从不会过时，并且基于新被索引的信息而被更新。此外，用户可以被通知任何可用的新的注释、药物、临床试验。

[0250] 在各种实施例中，本文提供的系统和方法可以允许分析扩展超越静态临床报告和被预先计算的癌症门户分析两者，以提供针对个体患者或同类群组的被超链接的报告的动态生成。这样的报告的示例包括但不限于肿瘤概况分析、药物和试验匹配、以及个体样本的免疫报告、以及样本同类群组的同类群组概况分析报告。报告可以基于用户查询而被定制，并且在各种实施例中，报告包含由多组学癌症搜索返回的经用户预先选择的结果。

[0251] 申请人已有利地发现，基于多组学癌症搜索系统的动态报告范例可以提供(1)用户与数据的交互，该用户与数据的交互超越了在广泛的生物信息学管线已经被运行之后无法被修改或更新的标准静态PDF报告的能力；(2)按照其临床可操作性、致病性、特征权重或频率对所有多组学癌症变化进行排名；(3)用户询问从BAM到VCF到输出的任何级别的管线输出，以进行更复杂的分析；(4)用户不仅查看机器学习模型预测，还查看指导特定预测的经排名特征的列表。

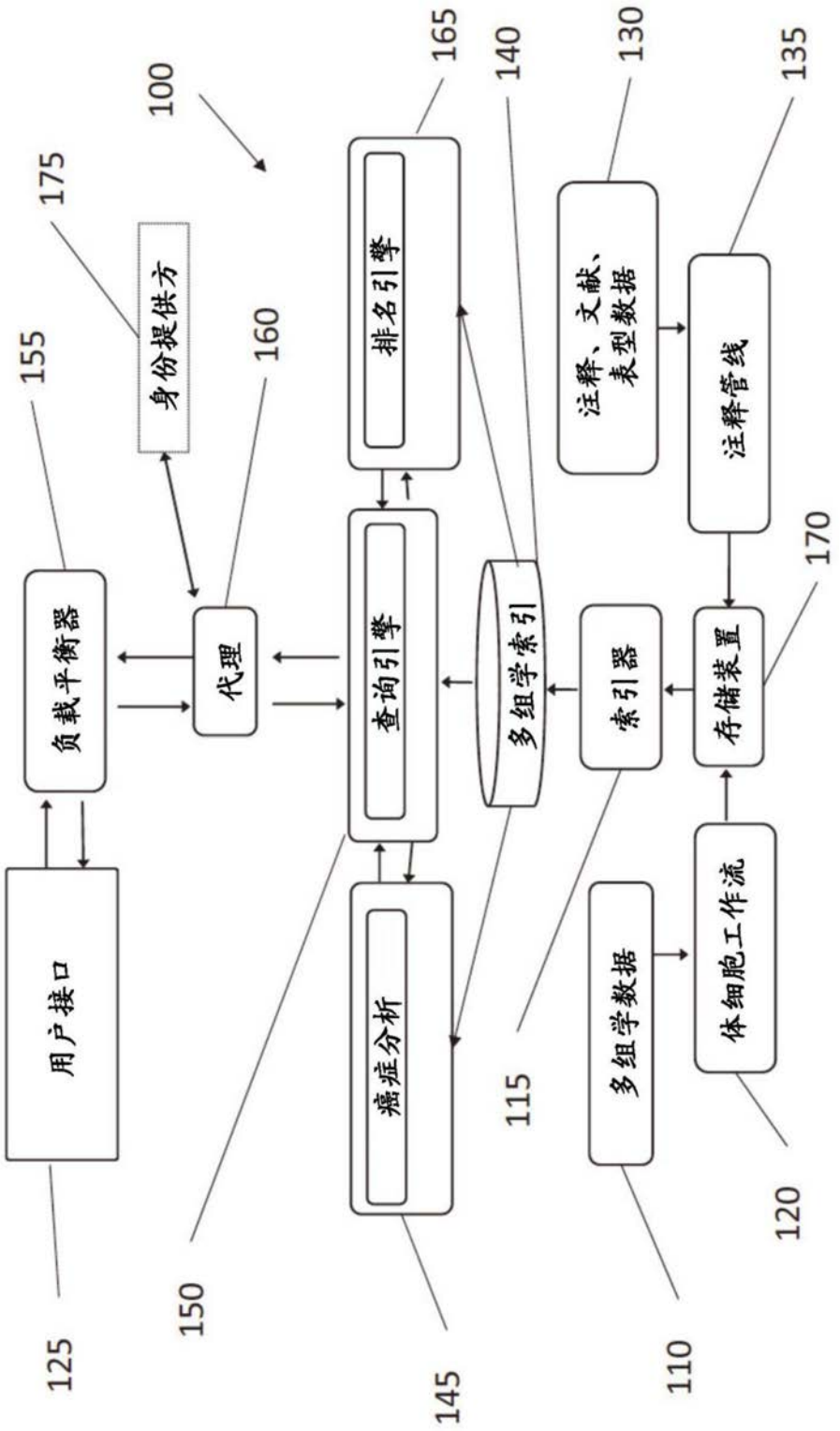

100
165
140
130
135
175
155
160
排名引擎
注释、文献、表型数据
注释管线
身份提供方
负载平衡器
代理
查询引擎
多组学索引
索引器
存储装置
170
用户接口
癌症分析
体细胞工作流
115
多组学数据
150
145
110
120
125

图1

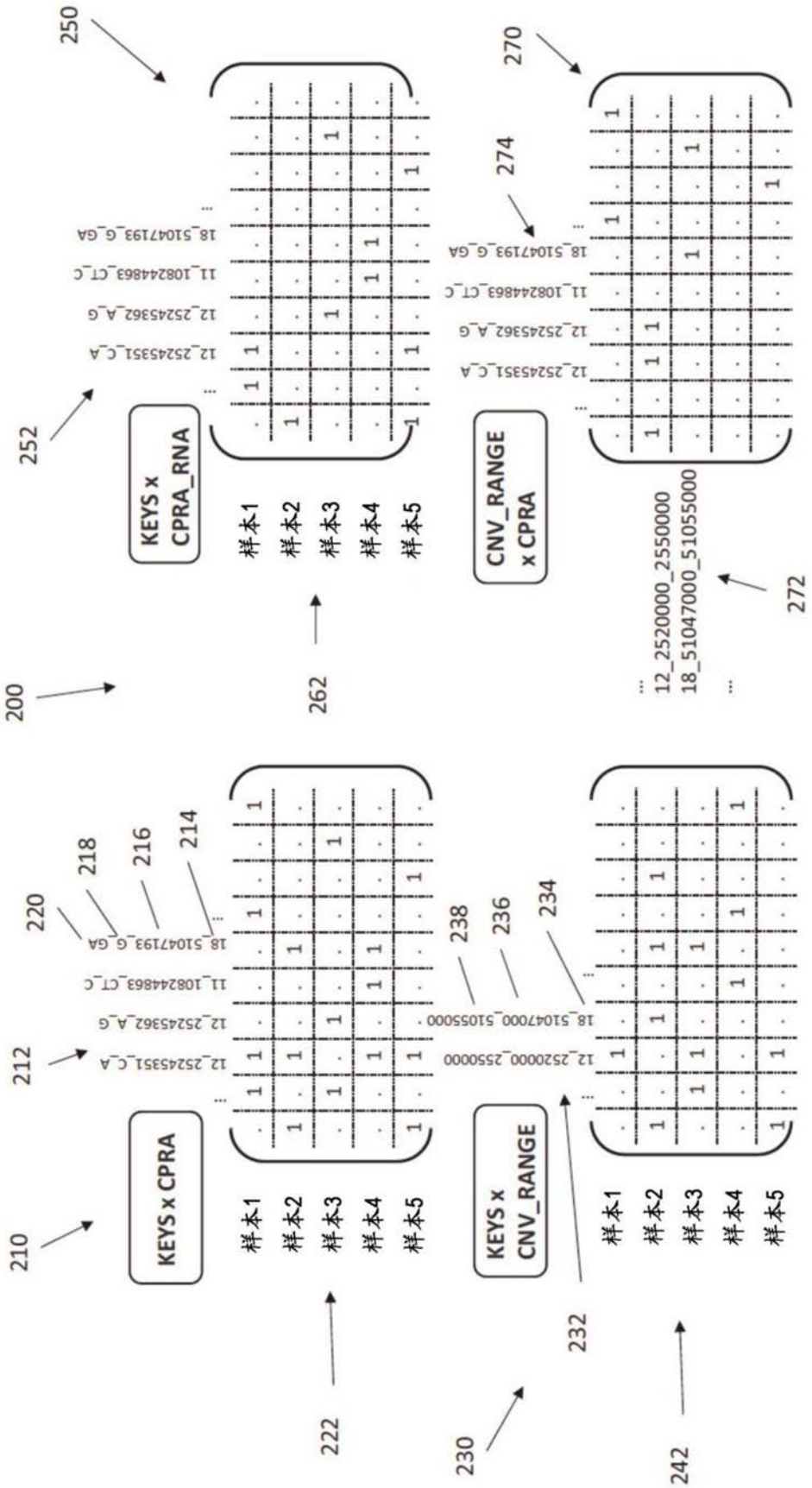

250
270
274
252
KEYS x CPRA_RNA
CNV_RANGE x CPRA
样本1
样本2
样本3
样本4
样本5
12_25245351_C_A
12_25245362_A_G
11_108244863_CT_C
18_51047193_G_GA
200
262
272
12_2520000_2550000
18_51047000_51055000
218
216
214
220
238
236
234
212
210
KEYS x CPRA
KEYS x CNV_RANGE
222
230
232
242

图2a

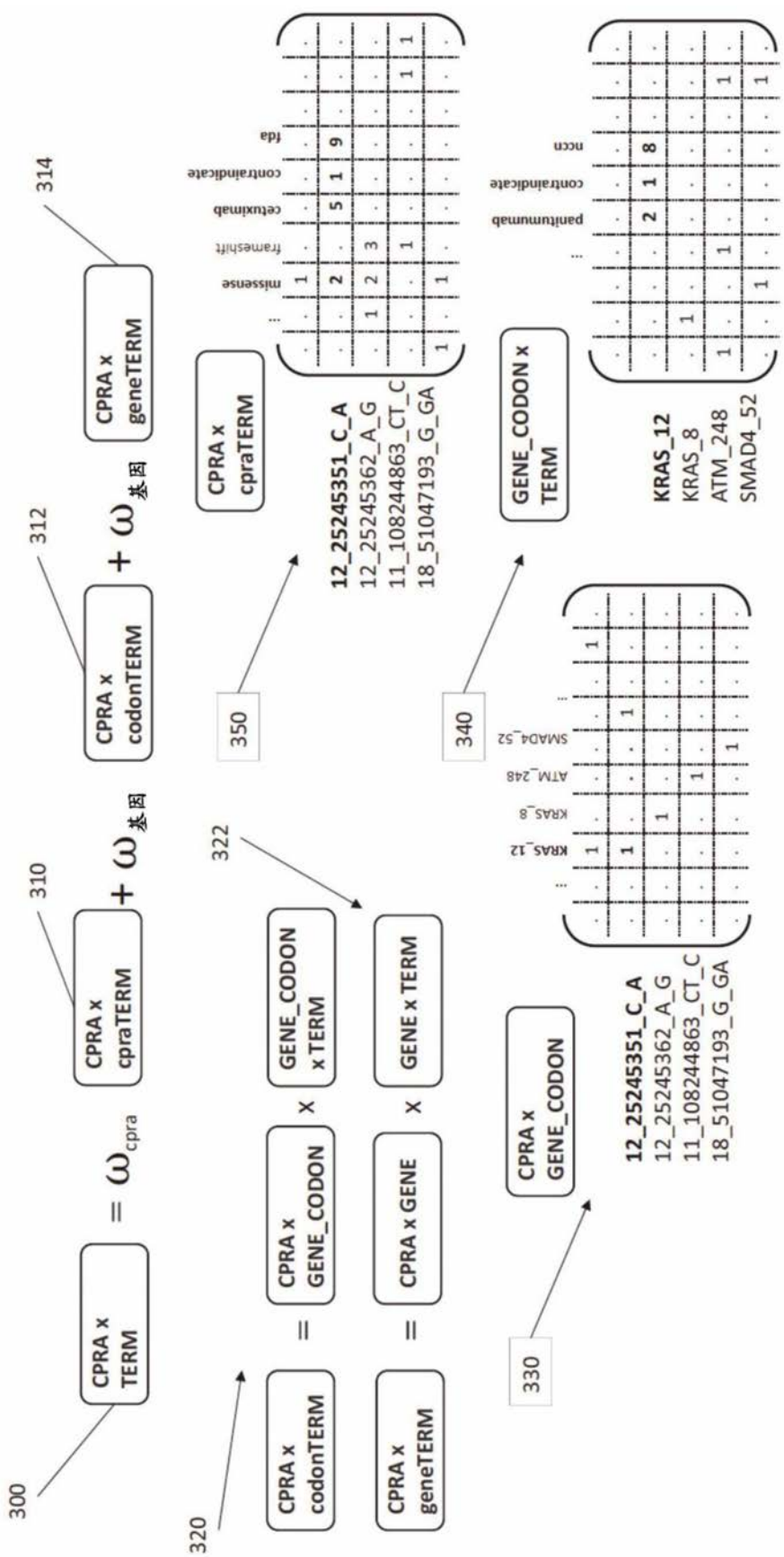

300
CPRA x TERM
= ω cpra
CPRA x cpraTERM
+ ω 基因
CPRA x codonTERM
+ ω 基因
CPRA x geneTERM
310
312
314
320
CPRA x codonTERM
CPRA x geneTERM
=
CPRA x GENE_CODON
CPRA x GENE
X
GENE_CODON x TERM
GENE x TERM
322
330
CPRA x GENE_CODON
350
CPRA x cpraTERM
340
GENE_CODON x TERM
12_25245351_C_A
12_25245362_A_G
11_108244863_CT_C
18_51047193_G_GA
KRAS_12
KRAS_8
ATM_248
SMAD4_52
missense
frameshift
cetuximab
contraindicate
fda
panitumumab
nccn

图2b

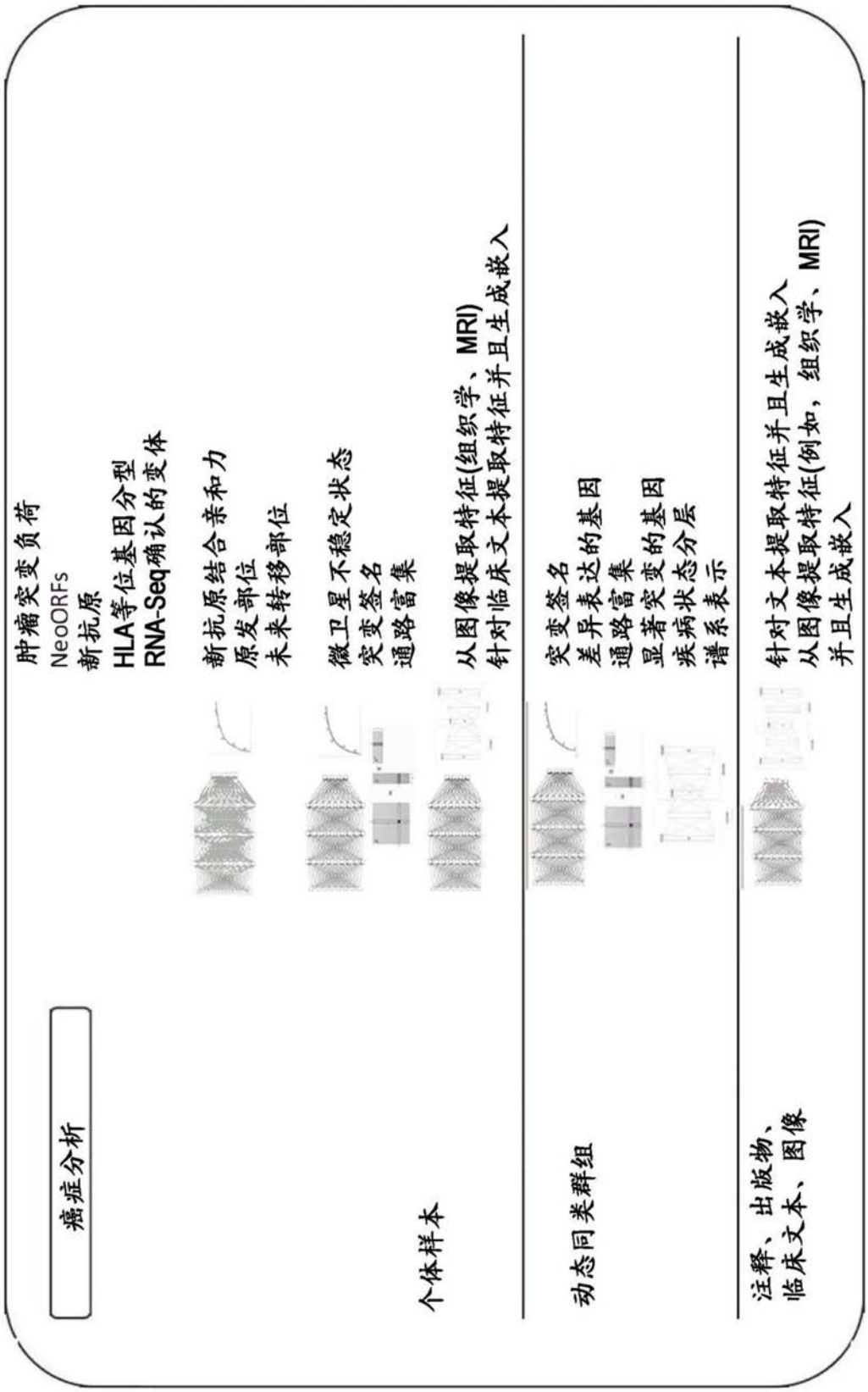

癌症分析
肿瘤突变负荷
NeoORFs
新抗原
HLA等位基因分型
RNA-Seq确认的变体
新抗原结合亲和力
原发部位
未来转移部位
微卫星不稳定状态
突变签名
通路富集
个体样本
从图像提取特征(组织学、MRI)
针对临床文本提取特征并且生成嵌入
突变签名
差异表达的基因
通路富集
显著突变的基因
疾病状态分层
谱系表示
动态同类群组
注释、出版物、
临床文本、图像
针对文本提取特征并且生成嵌入
从图像提取特征(例如，组织学、MRI)
并且生成嵌入

图3

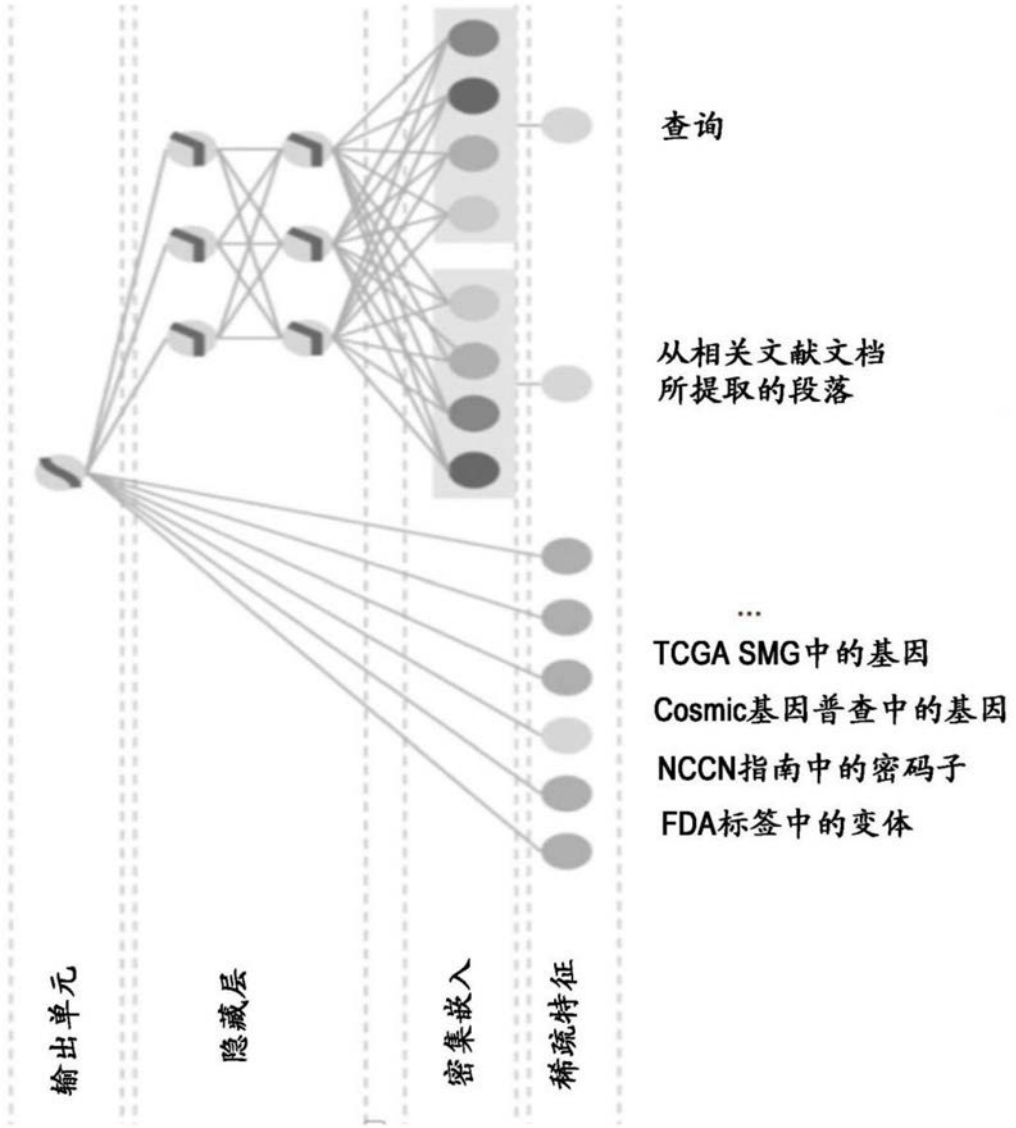

查询
从相关文献文档
所提取的段落
...
TCGA SMG中的基因
Cosmic基因普查中的基因
NCCN指南中的密码子
FDA标签中的变体
输出单元
隐藏层
密集嵌入
稀疏特征

图4a

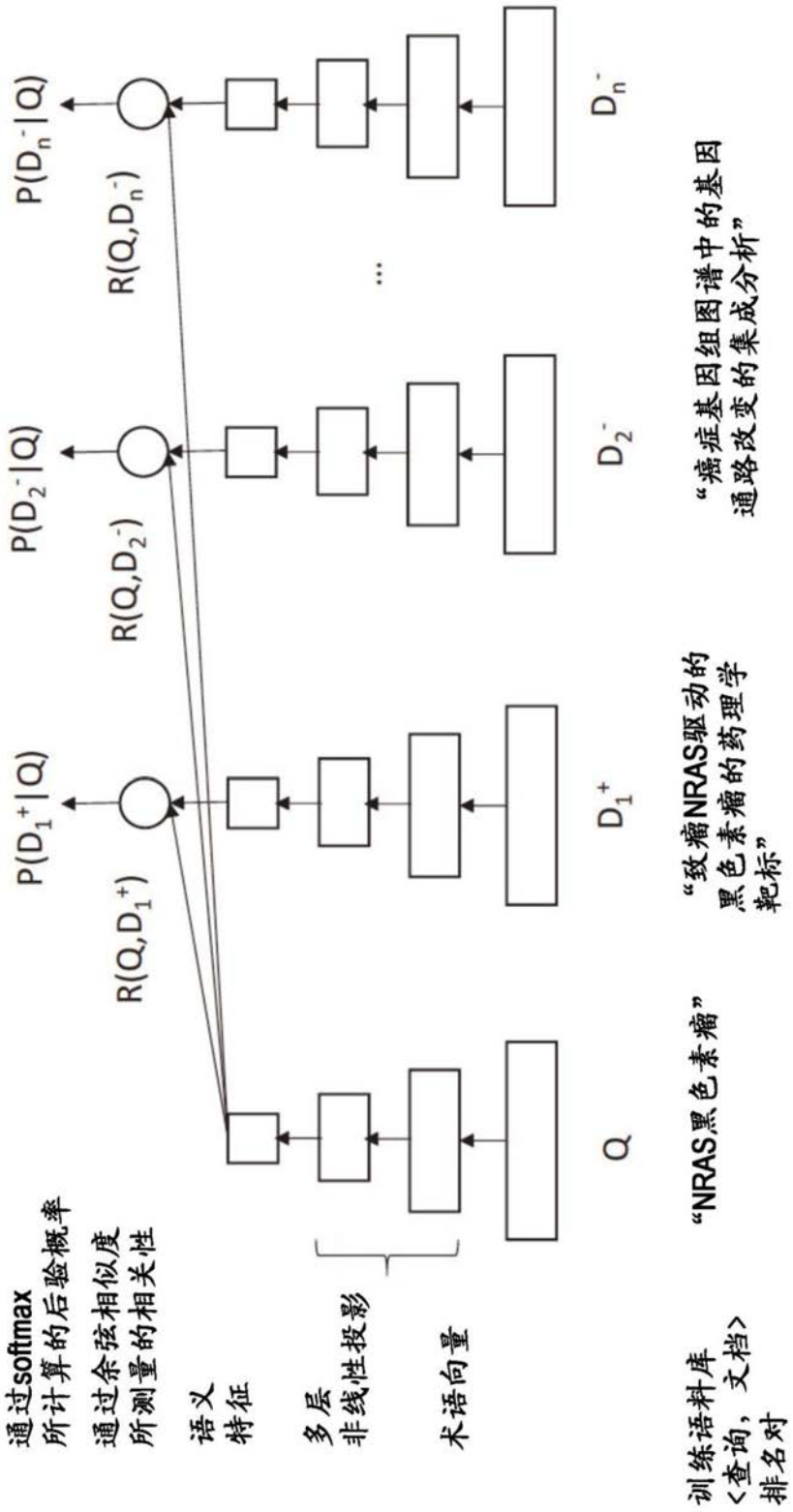

通过softmax
所计算的后验概率
通过余弦相似度
所测量的相关性
语义
特征
多层
非线性投影
术语向量
训练语料库
<查询，文档>
排名对
Q
D1+
D2-
Dn-
"NRAS黑色素瘤"
"致癌NRAS驱动的
黑色素瘤的药理学
靶标"
"癌症基因组图谱中的基因
通路改变的集成分析"
R(Q,D1+)
R(Q,D2-)
R(Q,Dn-)
P(D1+|Q)
P(D2-|Q)
P(Dn-|Q)

图4b

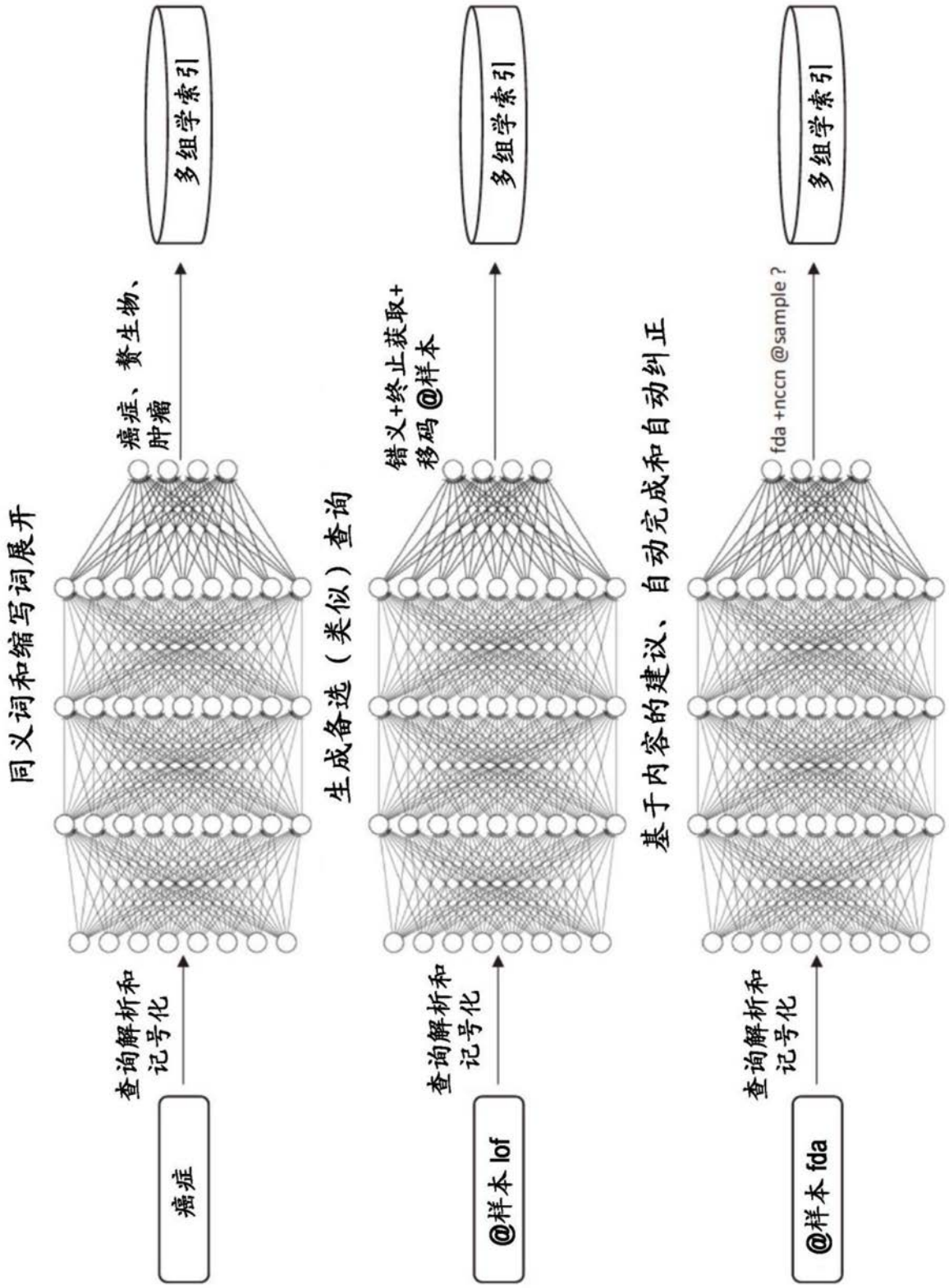

癌症
查询解析和记号化
同义词和缩写词展开
癌症、赘生物、肿瘤
多组学索引
@样本 lof
查询解析和记号化
生成备选（类似）查询
错义+终止获取+移码 @样本
多组学索引
@样本 fda
查询解析和记号化
基于内容的建议、自动完成和自动纠正
fda +nccn @sample ?
多组学索引

图5a

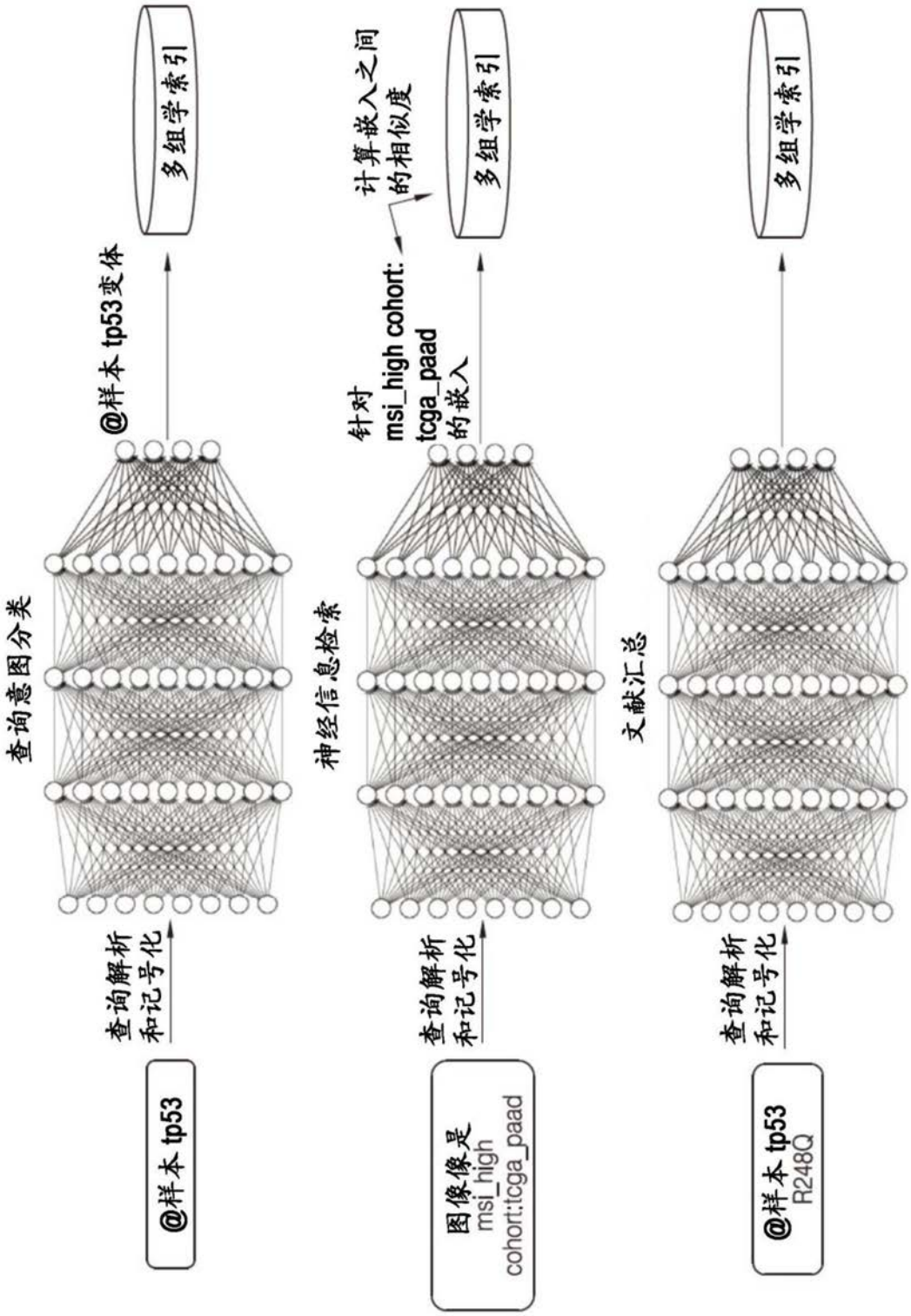

@样本 tp53
查询解析和记号化
查询意图分类
@样本 tp53变体
多组学索引
图像像是 msi_high cohort:tcga_paad
查询解析和记号化
神经信息检索
针对 msi_high cohort: tcga_paad 的嵌入
计算嵌入之间的相似度
多组学索引
@样本 tp53 R248Q
查询解析和记号化
文献汇总
多组学索引

图5b

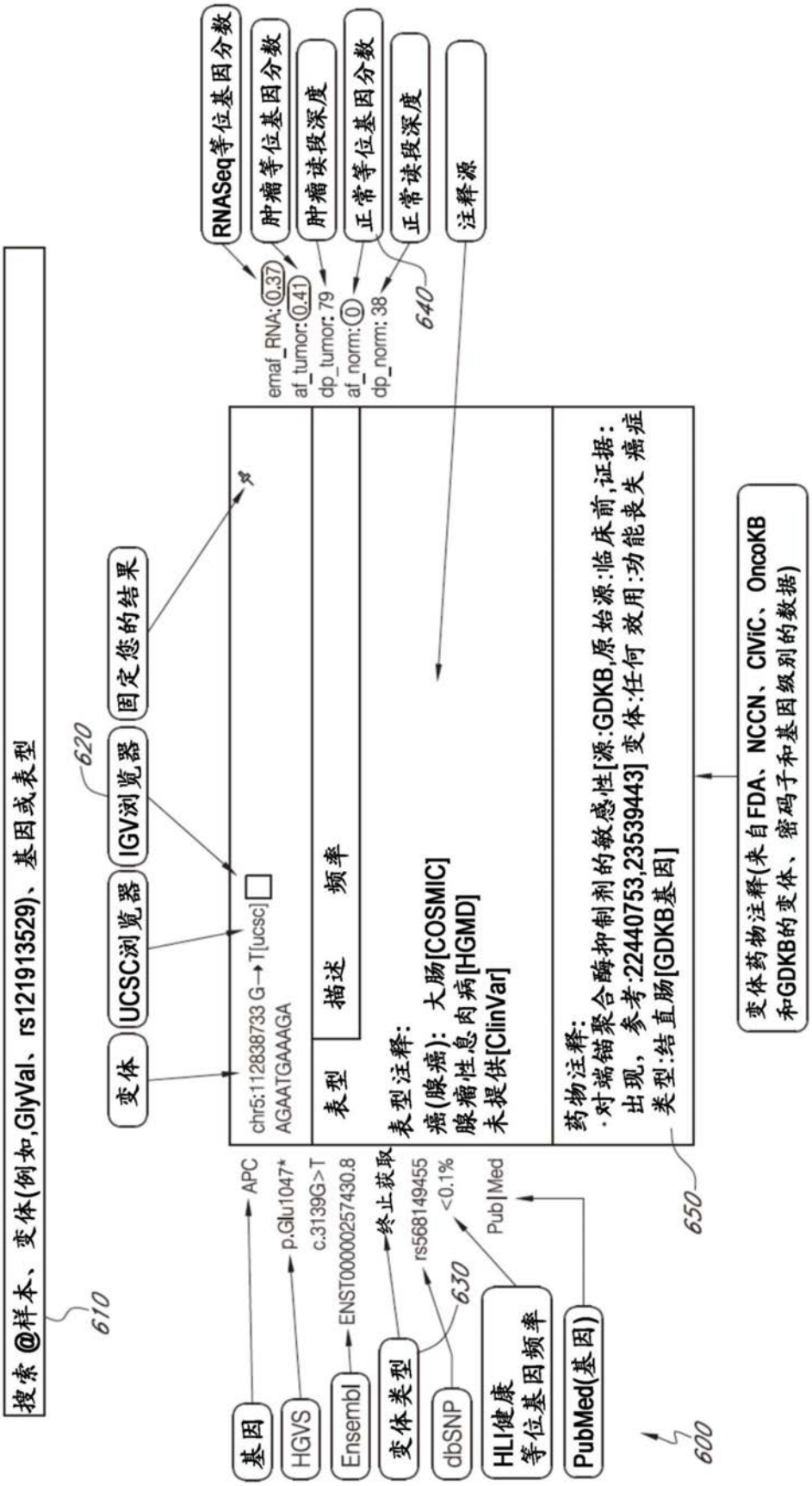

搜索 @样本、变体(例如,GlyVal、rs121913529)、基因或表型
610
变体
UCSC浏览器
IGV浏览器
固定您的结果
620
chr5:112838733 G→T[ucsc]
AGAATGAAAGA
表型
描述
频率
表型注释:
癌(腺癌): 大肠[COSMIC]
腺瘤性息肉病[HGMD]
未提供[ClinVar]
药物注释:
·对端锚聚合酶抑制剂的敏感性[源:GDKB,原始源:临床前,证据:出现, 参考:22440753,23539443] 变体:任何 效用:功能丧失 癌症类型:结直肠[GDKB基因]
650
emaf_RNA: 0.37
af_tumor: 0.41
dp_tumor: 79
af_norm: ①
dp_norm: 38
640
RNASeq等位基因分数
肿瘤等位基因分数
肿瘤读段深度
正常等位基因分数
正常读段深度
注释源
变体药物注释(来自FDA、NCCN、CIViC、OncoKB和GDKB的变体、密码子和基因级别的数据)
基因
HGVS
Ensembl
变体类型
dbSNP
HLI健康等位基因频率
PubMed(基因)
APC
p.Glu1047*
c.3139G>T
ENST00000257430.8
终止获取
rs568149455
<0.1%
PubMed
630
600

图6

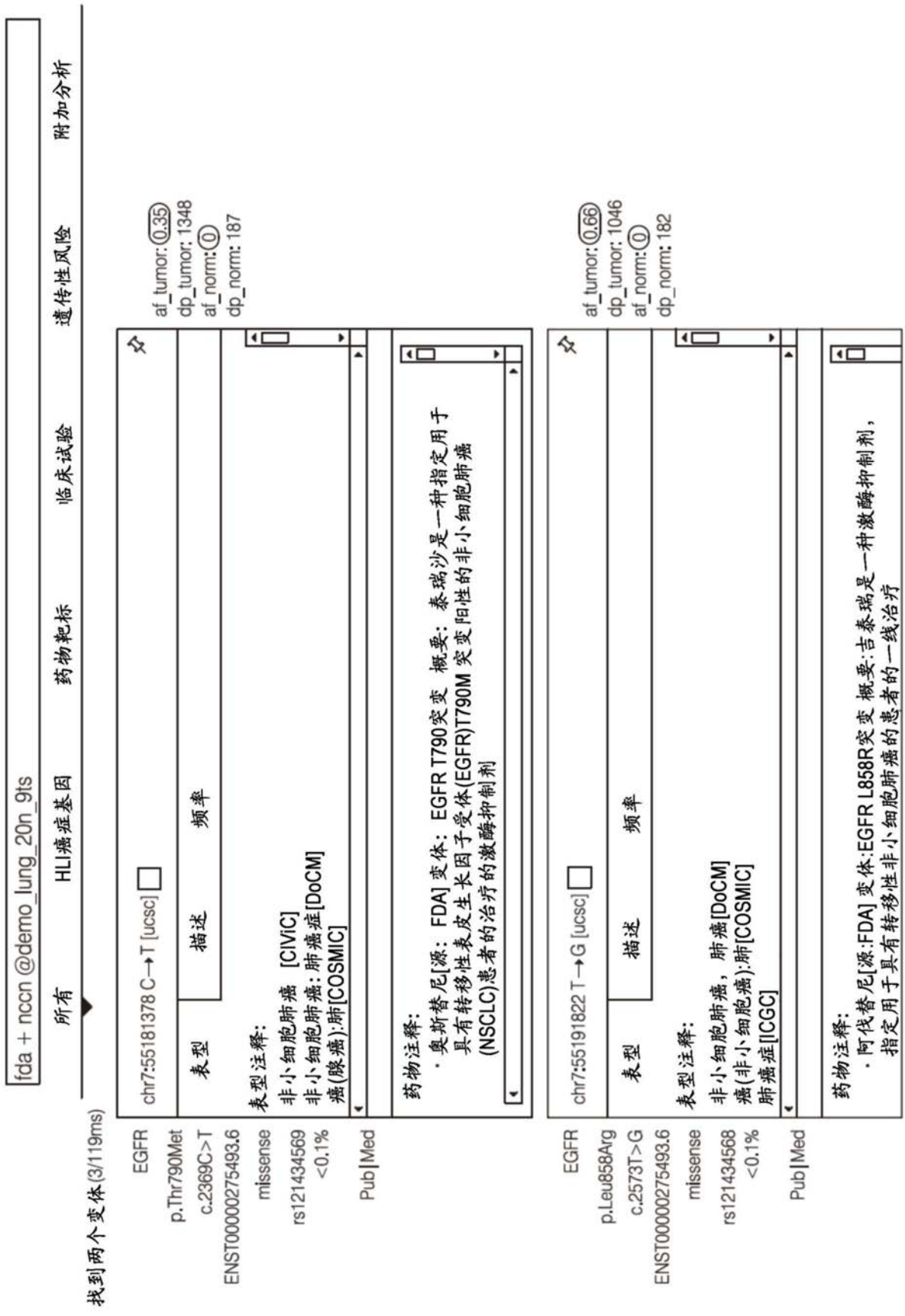

fda + nccn @demo_lung_20n_9ts
所有
HLI癌症基因
药物靶标
临床试验
遗传性风险
附加分析
找到两个变体(3/119ms)
EGFR
p.Thr790Met
c.2369C>T
ENST00000275493.6
missense
rs121434569
<0.1%
Pub|Med
chr7:55181378 C→T [ucsc]
af_tumor: 0.35
dp_tumor: 1348
af_norm: 0
dp_norm: 187
表型
描述
频率
表型注释:
非小细胞肺癌 [CIViC]
非小细胞肺癌:肺癌症[DoCM]
癌(腺癌):肺[COSMIC]
药物注释:
·奥斯替尼[源: FDA] 变体: EGFR T790突变 概要: 泰瑞沙是一种指定用于具有转移性表皮生长因子受体(EGFR)T790M 突变阳性的非小细胞肺癌(NSCLC),患者的治疗的激酶抑制剂
EGFR
p.Leu858Arg
c.2573T>G
ENST00000275493.6
missense
rs121434568
<0.1%
Pub|Med
chr7:55191822 T→G [ucsc]
af_tumor: 0.66
dp_tumor: 1046
af_norm: 0
dp_norm: 182
表型
描述
频率
表型注释:
非小细胞肺癌, 肺癌[DoCM]
癌(非小细胞癌):肺[COSMIC]
肺癌症[ICGC]
药物注释:
·阿伐替尼[源:FDA] 变体:EGFR L858R突变 概要:吉泰瑞是一种激酶抑制剂,指定用于具有转移性非小细胞肺癌的患者的一线治疗

图7

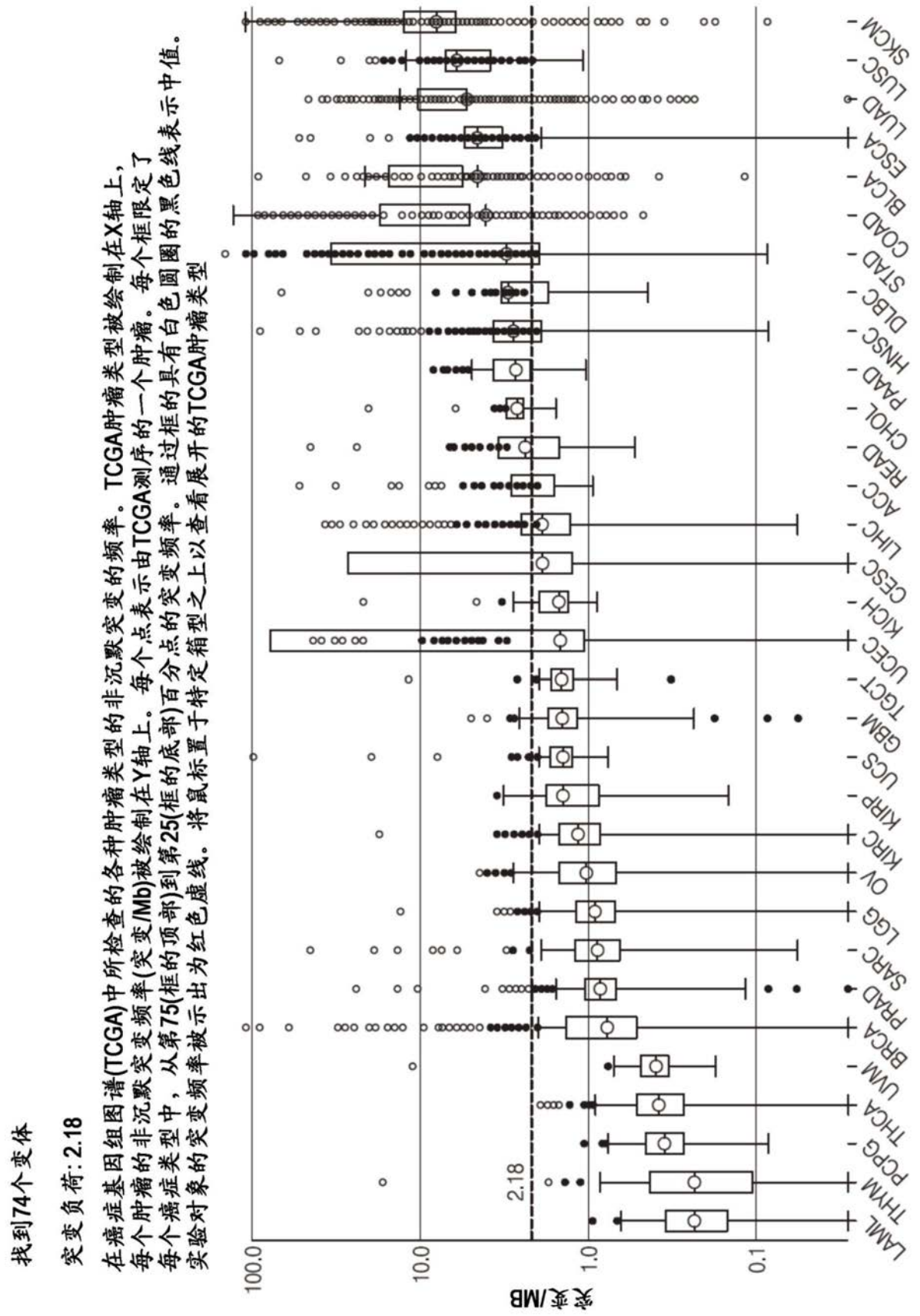

找到74个变体
突变负荷: 2.18
在癌症基因组图谱(TCGA)中所检查的各种肿瘤类型的非沉默突变的频率。TCGA肿瘤类型被绘制在X轴上，
每个肿瘤的非沉默突变频率(突变/Mb)被绘制在Y轴上。每个点表示由TCGA测序的一个肿瘤。每个框限定了
每个癌症类型中，从第75(框的顶部)到第25(框的底部)百分点的突变频率。通过框的具有白色圆圈的黑色线表示中值。
实验对象的突变频率被示出为红色虚线。将鼠标置于特定箱型之上以查看展开的TCGA肿瘤类型
突变/Mb
100.0
10.0
1.0
0.1
2.18
LAML
THYM
PCPG
THCA
UVM
BRCA
PRAD
SARC
LGG
OV
KIRC
KIRP
UCS
GBM
TGCT
UCEC
KICH
CESC
LIHC
ACC
READ
CHOL
PAAD
HNSC
DLBC
STAD
COAD
BLCA
ESCA
LUAD
LUSC
SKCM

图8a

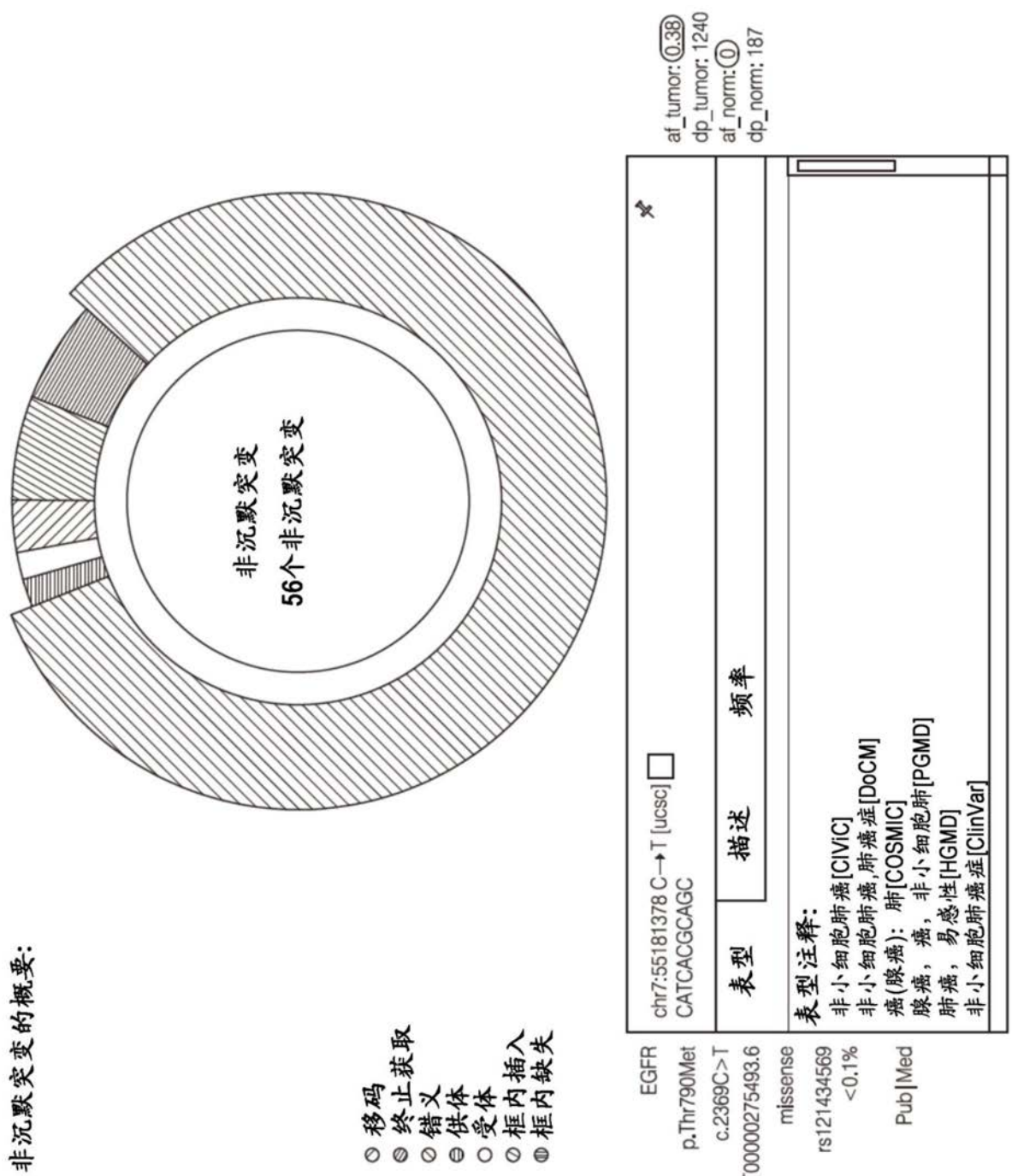

非沉默突变的概要:
非沉默突变
56个非沉默突变
移码
终止获取
错义
供体
受体
框内插入
框内缺失
EGFR
p.Thr790Met
c.2369C>T
ENST00000275493.6
missense
rs121434569
<0.1%
Pub|Med
chr7:55181378 C→T [ucsc]
CATCACGCAGC
af_tumor: 0.38
dp_tumor: 1240
af_norm: 0
dp_norm: 187
表型
描述
频率
表型注释:
非小细胞肺癌[CIViC]
非小细胞肺癌,肺癌症[DoCM]
癌(腺癌): 肺[COSMIC]
腺癌, 癌, 非小细胞肺[PGMD]
肺癌, 易感性[HGMD]
非小细胞肺癌症[ClinVar]

图8b

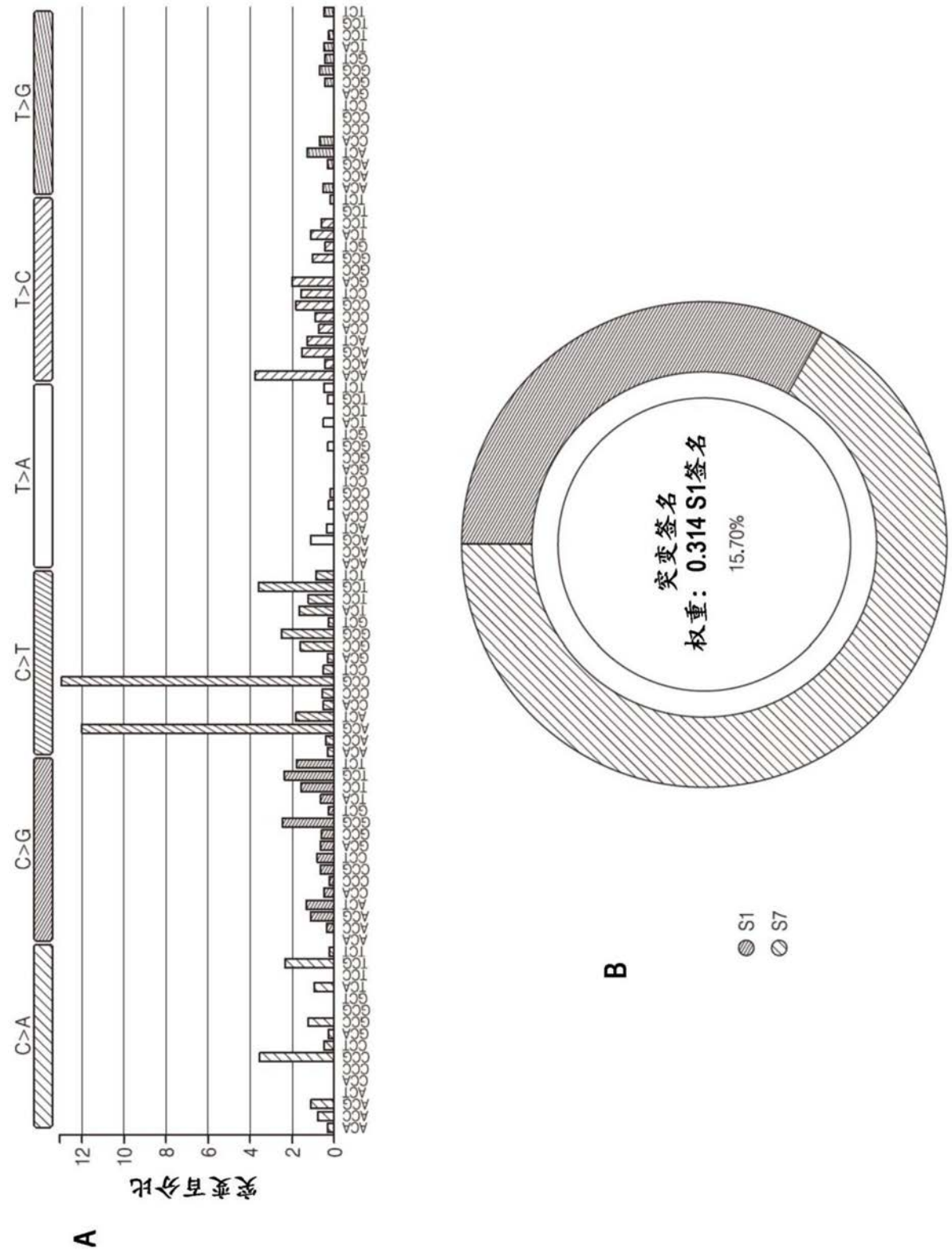

A
突变百分比
0
2
4
6
8
10
12
C>A
C>G
C>T
T>A
T>C
T>G
B
突变签名
权重：0.314 S1签名
15.70%
S1
S7

图9

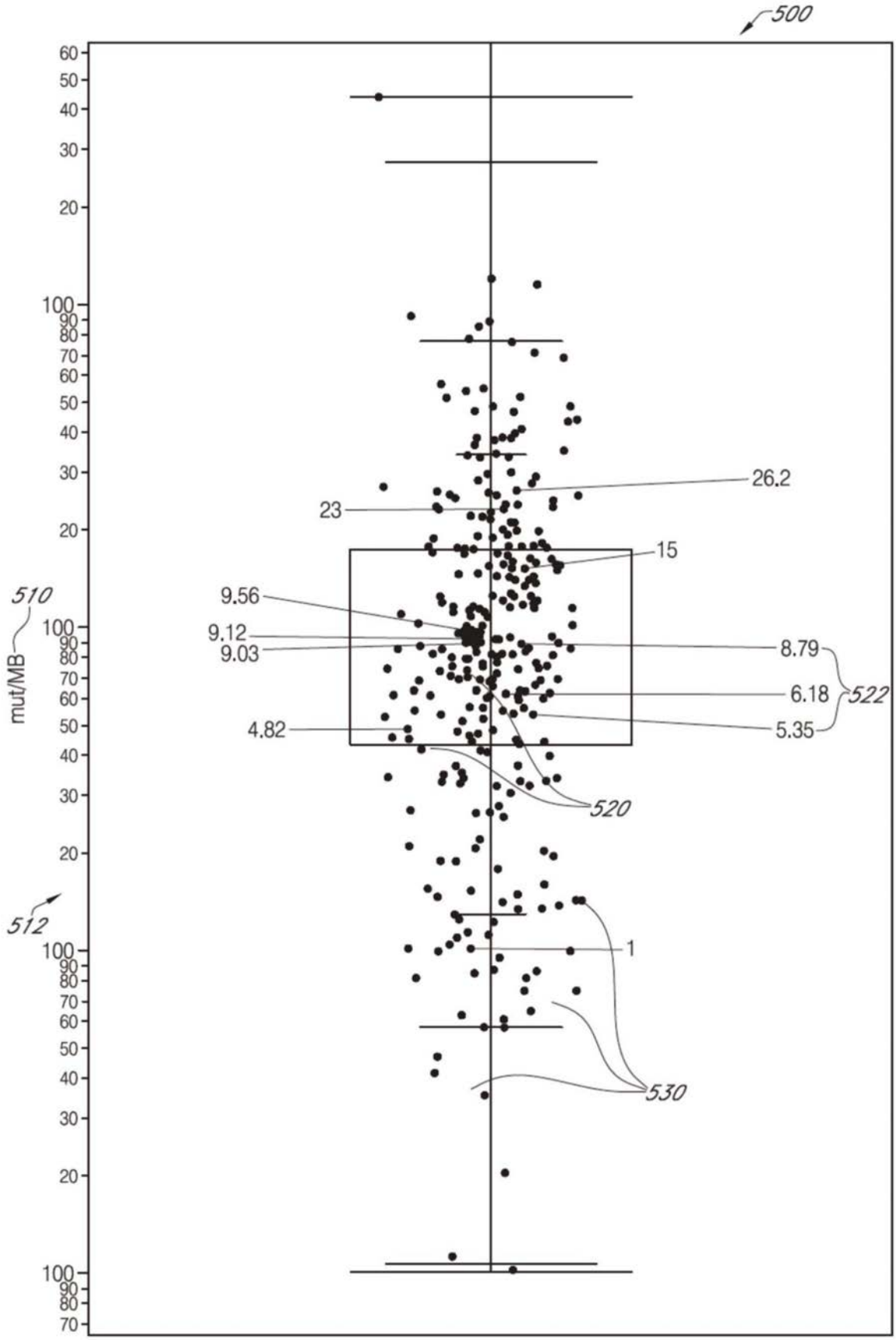

500
510
512
mut/MB
60
50
40
30
20
100
90
80
70
60
50
40
30
20
100
90
80
70
60
50
40
30
20
100
90
80
70
26.2
23
15
9.56
9.12
9.03
8.79
6.18
5.35
522
4.82
520
1
530

图10

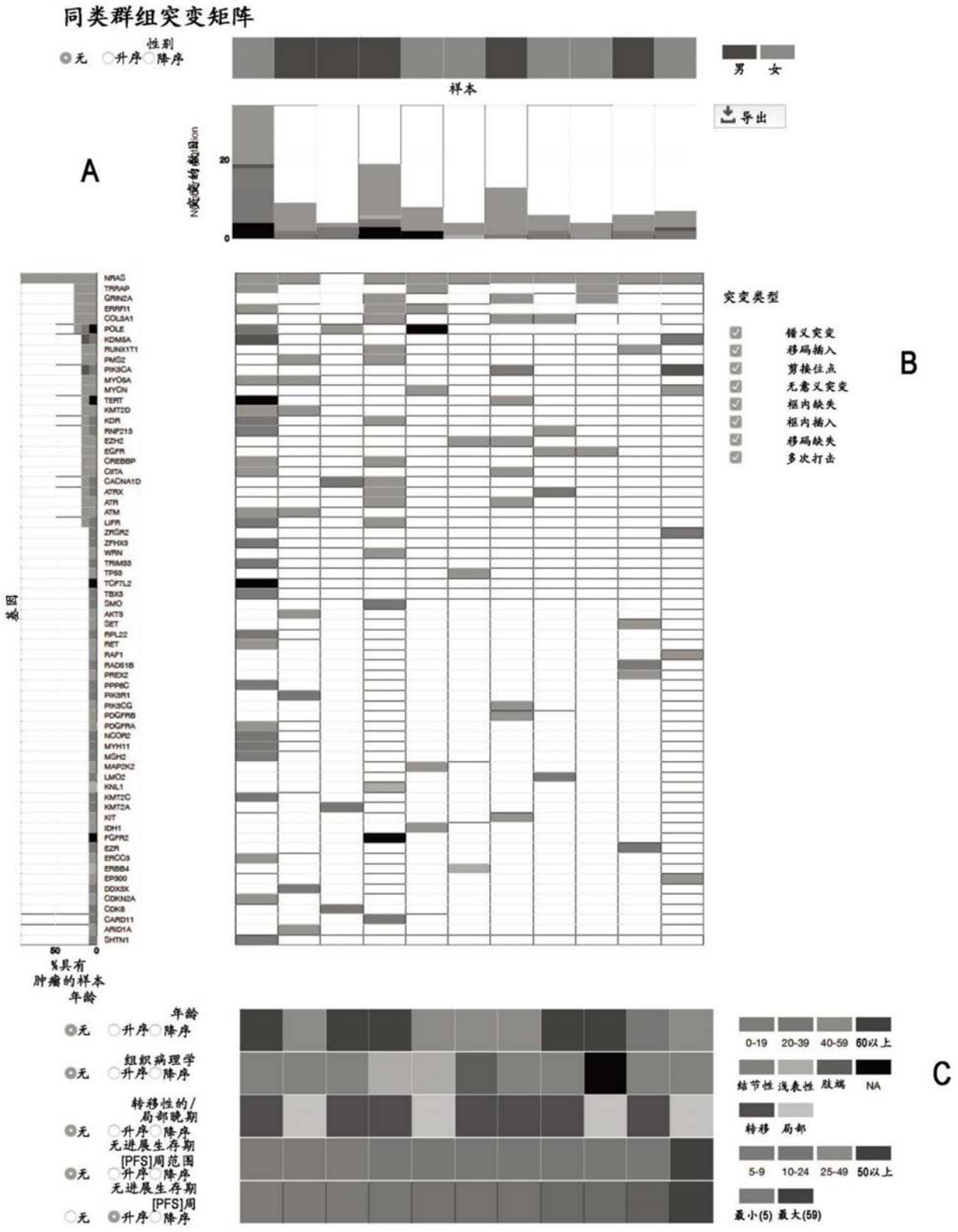

同类群组突变矩阵
性别
无 升序 降序
男 女
样本
导出
A
20
0
B
突变类型
错义突变
移码插入
剪接位点
无意义突变
框内缺失
框内插入
移码缺失
多次打击
基因
NRAS
TRRAP
GRIN2A
ERRFI1
COL3A1
POLE
KDM5A
RUNX1T1
PMS2
PIK3CA
MYO5A
MYCN
TERT
KMT2D
KDR
RNF213
EZH2
EGFR
CREBBP
CIITA
CACNA1D
ATRX
ATR
ATM
LIFR
ZRSR2
ZFHX3
WRN
TRIM33
TP53
TCF7L2
TBX3
SMO
AKT3
SET
RPL22
RET
RAF1
RAD51B
PREX2
PPP6C
PIK3R1
PIK3CG
PDGFRB
PDGFRA
NCOR2
MYH11
MGH2
MAP2K2
LMO2
KNL1
KMT2C
KMT2A
KIT
IDH1
FGFR2
EZR
ERCC3
ERBB4
EP300
DDX3X
CDKN2A
CDK8
CARD11
ARID1A
SHTN1
60
0
%具有肿瘤的样本
年龄
年龄
无 升序 降序
组织病理学
无 升序 降序
转移性的/局部晚期
无 升序 降序
无进展生存期[PFS]周范围
无 升序 降序
无进展生存期[PFS]周
无 升序 降序
0-19 20-39 40-59 60以上
结节性 浅表性 肢端 NA
转移 局部
5-9 10-24 25-49 50以上
最小(5) 最大(59)
C

图11

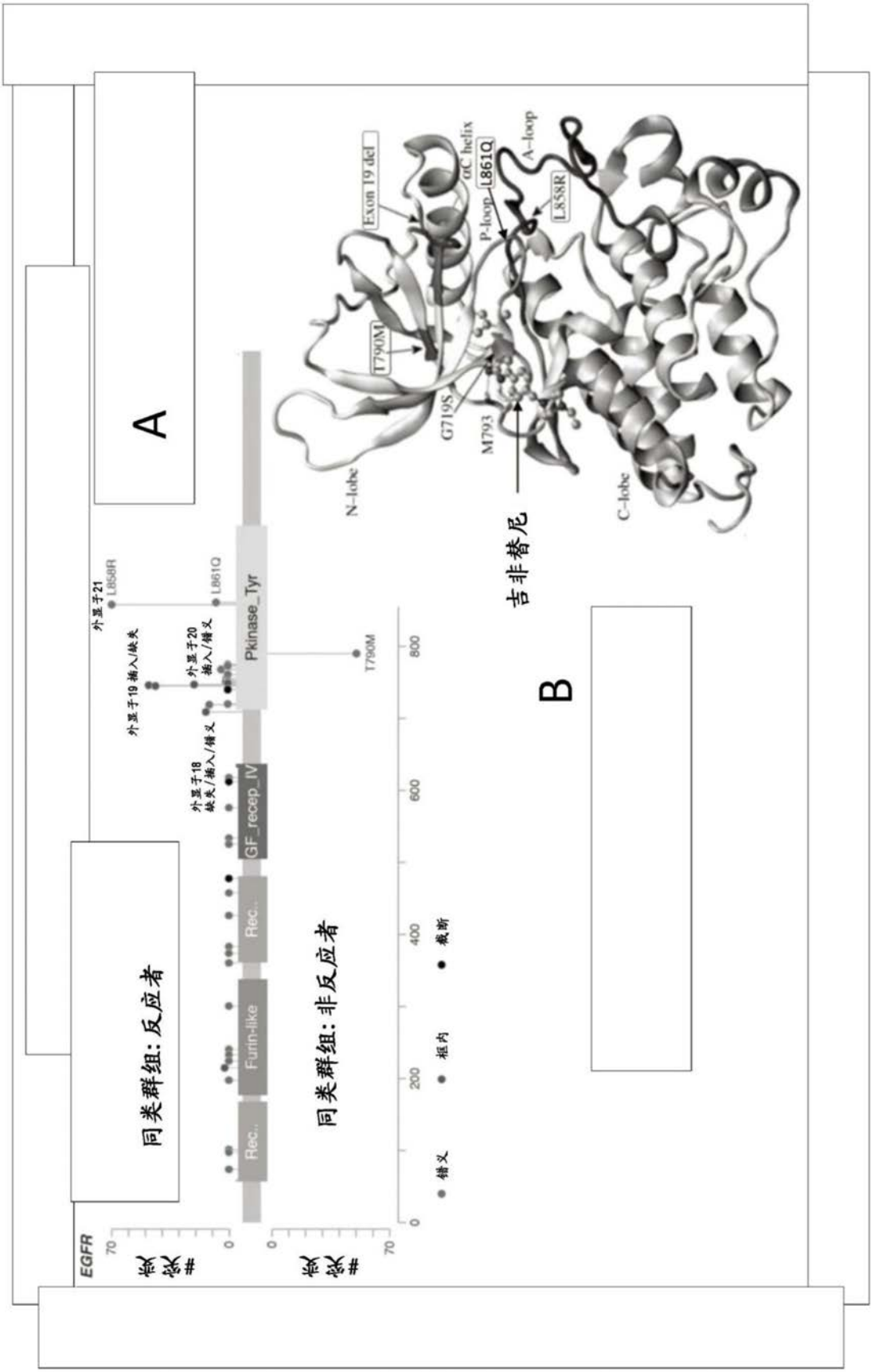

A
B
同类群组: 反应者
同类群组: 非反应者
EGFR
Rec..
Furin-like
Rec..
GF_recep_IV
Pkinase_Tyr
外显子21
L858R
外显子19 插入/缺失
外显子18
缺失/插入/错义
外显子20
插入/错义
L861Q
T790M
错义
框内
截断
N-lobe
C-lobe
Exon 19 del
T790M
G719S
M793
αC helix
L861Q
P-loop
A-loop
L858R
吉非替尼

图12

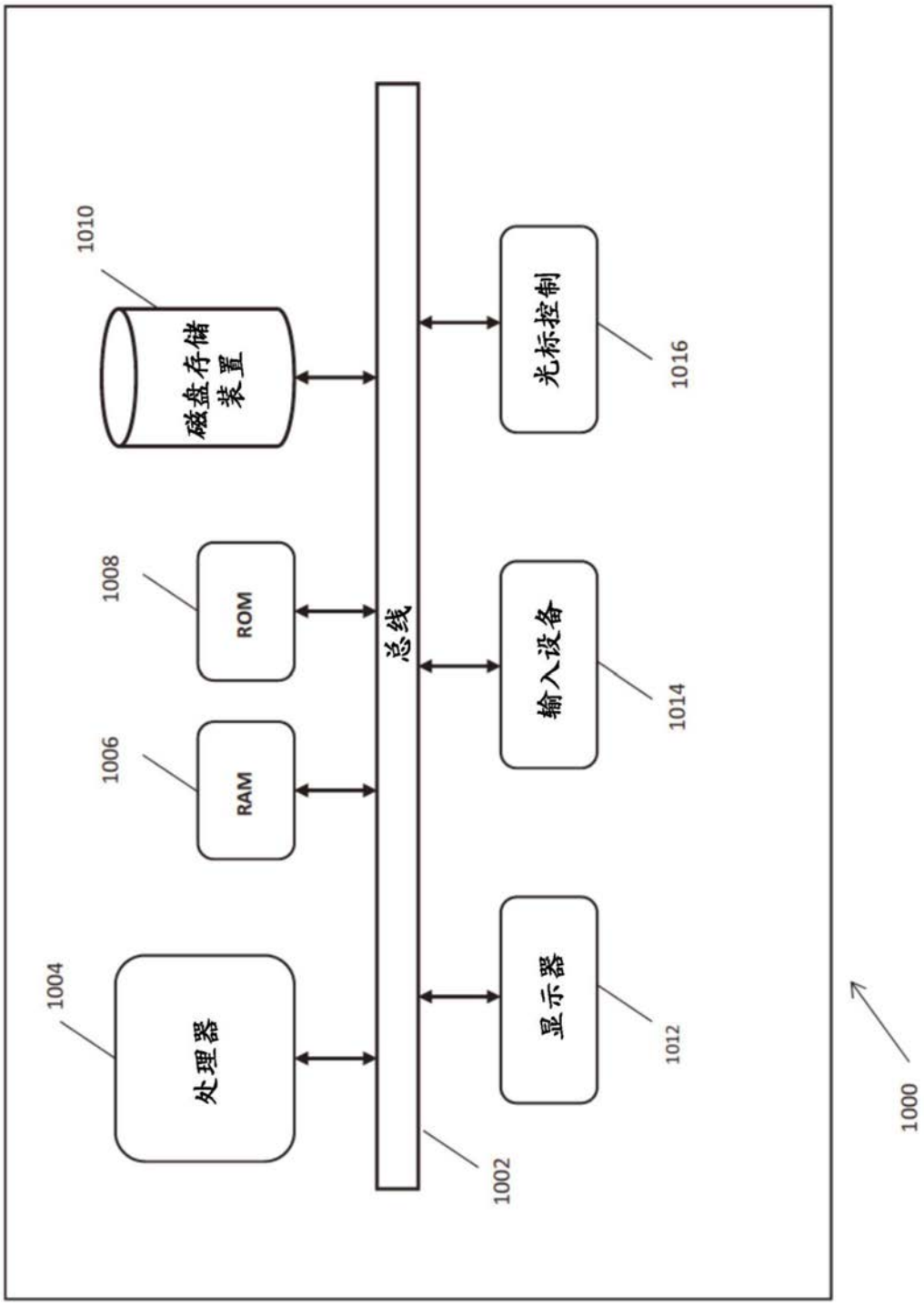

1010
磁盘存储装置
光标控制
1016
1008
ROM
总线
1006
RAM
输入设备
1014
1004
处理器
显示器
1012
1002
1000

图13

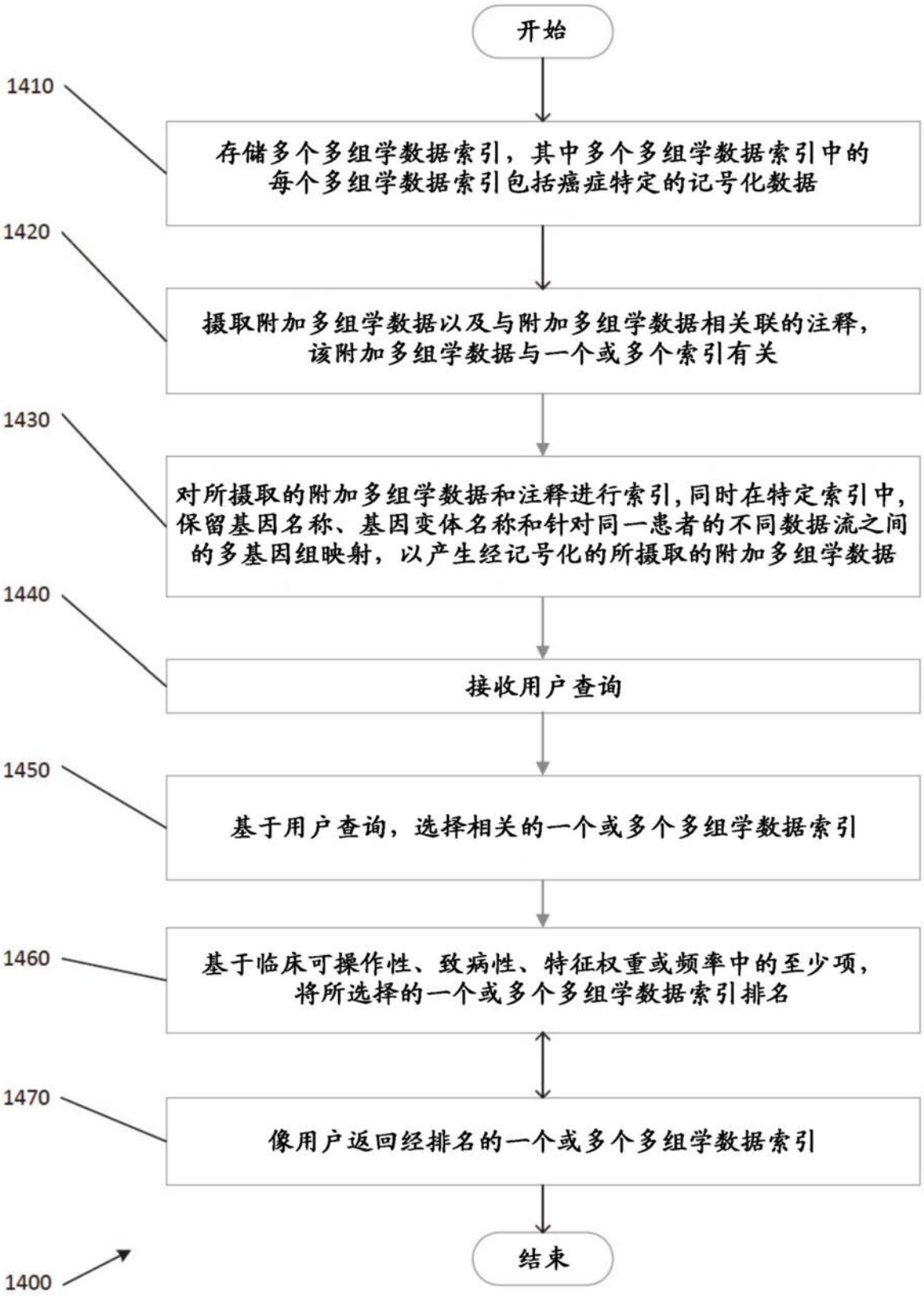

开始
1410
存储多个多组学数据索引，其中多个多组学数据索引中的每个多组学数据索引包括癌症特定的记号化数据
1420
摄取附加多组学数据以及与附加多组学数据相关联的注释，该附加多组学数据与一个或多个索引有关
1430
对所摄取的附加多组学数据和注释进行索引，同时在特定索引中，保留基因名称、基因变体名称和针对同一患者的不同数据流之间的多基因组映射，以产生经记号化的所摄取的附加多组学数据
1440
接收用户查询
1450
基于用户查询，选择相关的一个或多个多组学数据索引
1460
基于临床可操作性、致病性、特征权重或频率中的至少项，将所选择的一个或多个多组学数据索引排名
1470
像用户返回经排名的一个或多个多组学数据索引
结束
1400

图14

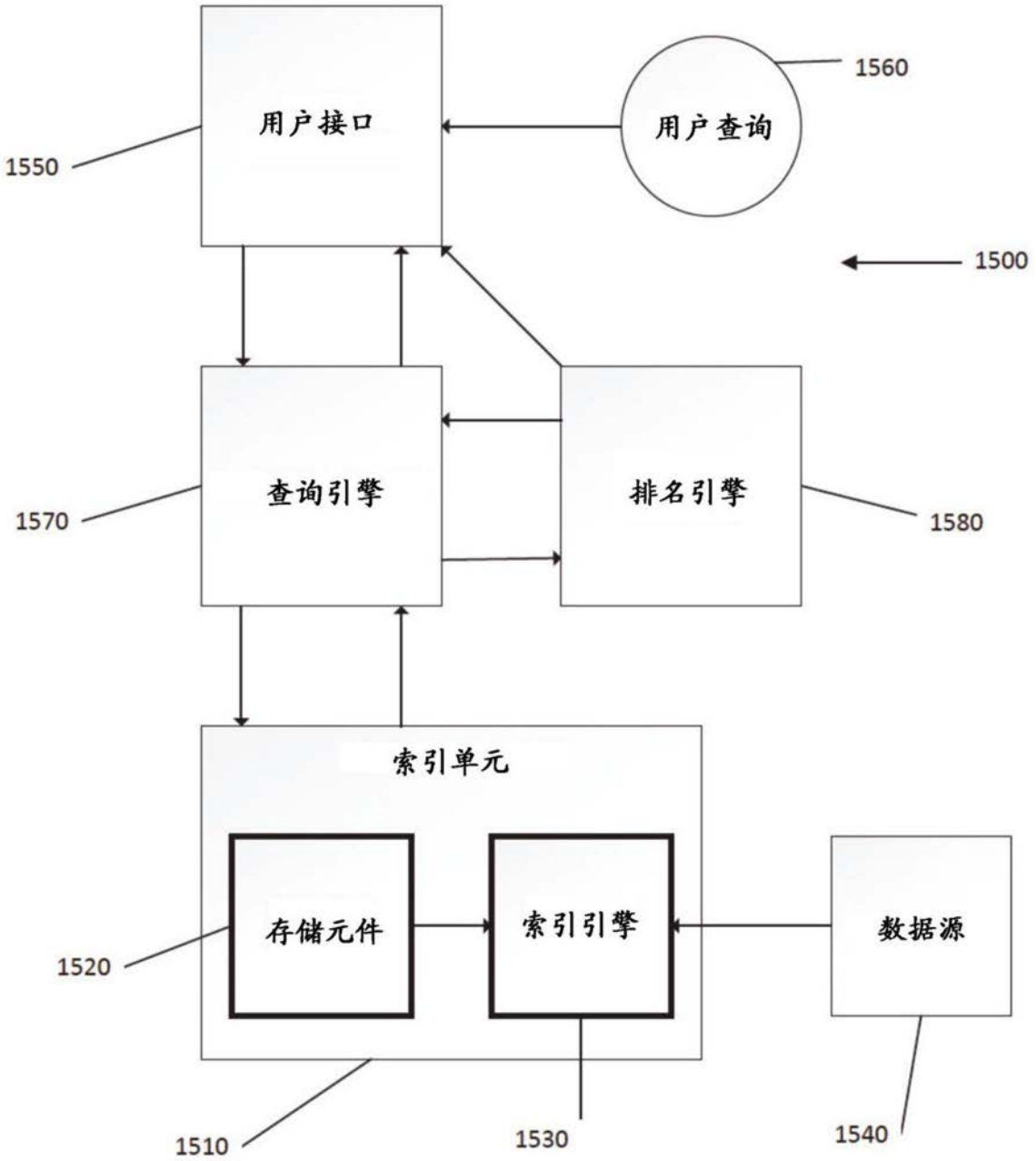

用户接口
用户查询
1560
1550
1500
查询引擎
排名引擎
1570
1580
索引单元
存储元件
索引引擎
数据源
1520
1510
1530
1540

图15

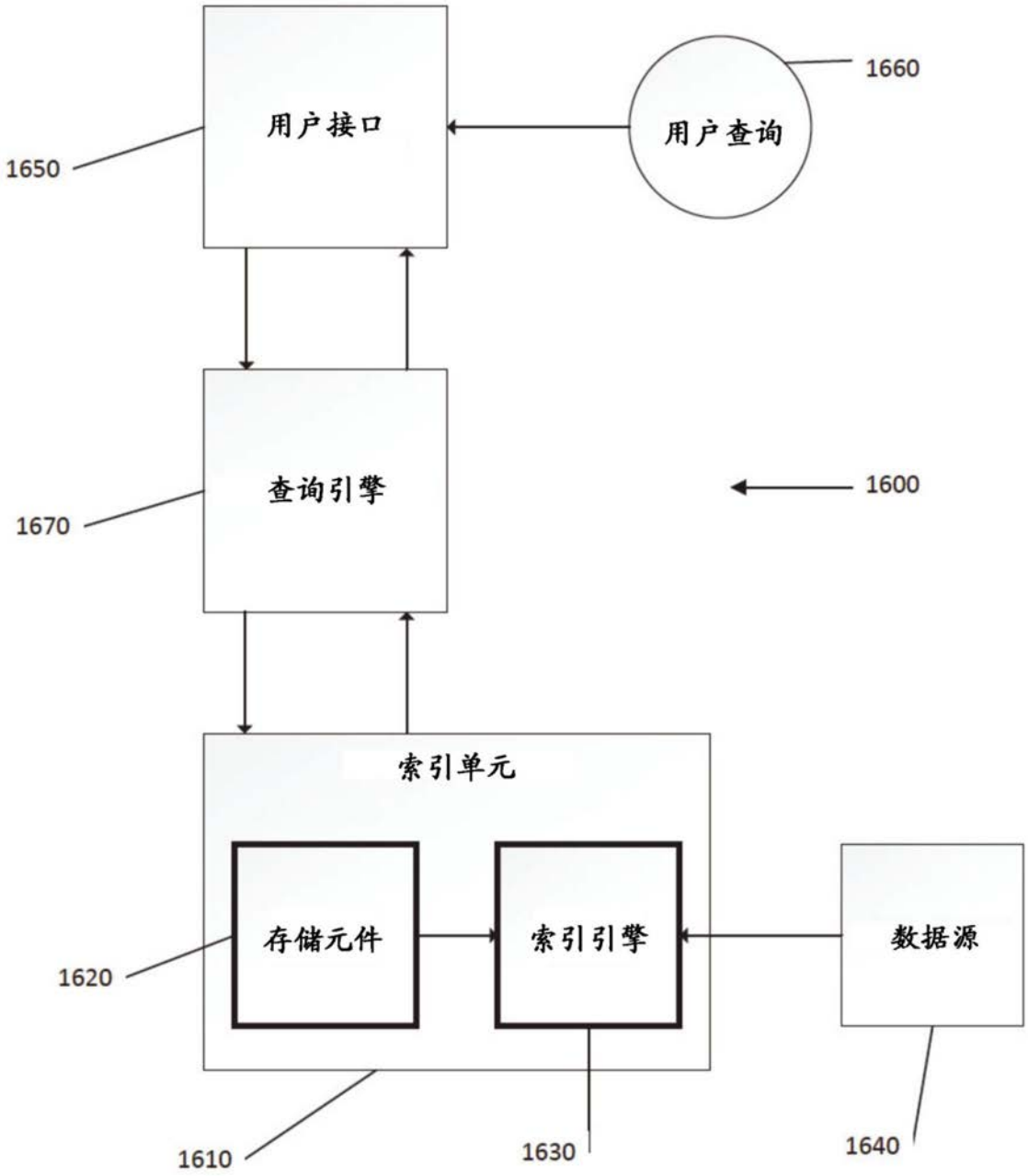

用户接口
1650
用户查询
1660
查询引擎
1670
1600
索引单元
存储元件
1620
索引引擎
数据源
1610
1630
1640

图16