



(12) 发明专利申请

(10) 申请公布号 CN 119851752 A

(43) 申请公布日 2025. 04. 18

(21) 申请号 202411604031.8

(22) 申请日 2019.02.27

(30) 优先权数据

62/636,061 2018.02.27 US

(62) 分案原申请数据

201980019430.9 2019.02.27

(71) 申请人 磨石生物公司

地址 美国加利福尼亚州

(72) 发明人 T·F·鲍彻 B·布里克-沙利文

J·巴斯比 M·斯科伯恩

R·耶冷斯凯

(74) 专利代理机构 北京市金杜律师事务所

11256

专利代理师 陈文平 王北南

(51) Int.Cl.

G16B 20/20 (2019.01)

G12Q 1/6869 (2018.01)

G01N 33/68 (2006.01)

G16B 40/00 (2019.01)

权利要求书5页 说明书152页

序列表(电子公布) 附图59页

(54) 发明名称

利用泛等位基因模型进行的新抗原鉴别

(57) 摘要

一种用于鉴别可能被MHC等位基因呈递在受试者的肿瘤细胞表面上的新抗原的方法。通过对所述受试者的肿瘤细胞进行测序来获得肿瘤新抗原的肽序列和MHC等位基因的肽序列。将所述肿瘤新抗原的肽序列和所述MHC等位基因的肽序列输入机器学习呈递模型中以产生所述肿瘤新抗原的呈递可能性,每个呈递可能性代表了新抗原被所述MHC等位基因中的至少一个呈递在所述受试者的肿瘤细胞表面上的可能性。基于所述呈递可能性选择所述新抗原的子集。



1. 一种用于鉴别来自受试者的一个或多个肿瘤细胞的可能被一个或多个MHC等位基因呈递在所述肿瘤细胞表面上的至少一种新抗原的方法,所述方法包括以下步骤:

从所述受试者的所述肿瘤细胞和正常细胞获得外显子组、转录组或全基因组核苷酸测序数据中的至少一种,其中所述核苷酸测序数据被用于获得代表通过比较来自所述肿瘤细胞的核苷酸测序数据和来自所述正常细胞的核苷酸测序数据鉴别的新抗原集合中每一种新抗原的肽序列的数据,其中每种新抗原的肽序列包含至少一个使其不同于从所述受试者的正常细胞鉴别的相应野生型肽序列的变化;

将所述新抗原中的每一种的肽序列编码成相应的数字矢量,每个数字矢量包含有关构成所述肽序列的多个氨基酸和所述氨基酸在所述肽序列中的位置的集合的信息;

从所述受试者的所述肿瘤细胞获得外显子组、转录组或全基因组核苷酸测序数据中的至少一种,其中所述核苷酸测序数据被用于获得代表所述受试者的所述一个或多个MHC等位基因中的每一个的肽序列的数据;

将所述受试者的所述一个或多个MHC等位基因中的每一个的肽序列编码成相应的数字矢量,每个数字矢量包含有关构成所述肽序列的多个氨基酸和所述氨基酸在所述肽序列中的位置的集合的信息;

使用计算机处理器将编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型中,以产生对于所述新抗原集合的呈递可能性集合,所述集合中的每个呈递可能性代表了相应的新抗原被所述一个或多个MHC等位基因呈递在所述受试者的肿瘤细胞表面上的可能性,所述机器学习呈递模型包含:

至少基于训练数据集鉴别的多个参数,所述训练数据集包含:

对于多个样品中的每个样品,通过质谱测量与被鉴别为存在于所述样品中的MHC等位基因集合中的至少一个MHC等位基因结合的肽的存在而获得的标记;

对于每个样品,编码为包含有关构成所述肽的多个氨基酸和所述氨基酸在所述肽中的位置的集合的信息的数字矢量的训练肽序列;以及

对于每个样品,编码为包含有关构成与所述样品的所述肽结合的至少一个MHC等位基因的多个氨基酸以及所述氨基酸在所述至少一个MHC等位基因中的位置的集合的信息的数字矢量的训练肽序列;

代表作为输入接收的编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量与基于所述数字矢量和所述参数作为输出生成的呈递可能性之间的关系的函数;

基于所述呈递可能性集合选择所述新抗原集合的子集,以产生选定的新抗原的集合;以及

回收所述选定的新抗原的集合。

2. 根据权利要求1所述的方法,其中将编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入所述机器学习呈递模型中包括:

将所述机器学习呈递模型应用于所述新抗原的所述肽序列并应用于所述一个或多个MHC等位基因的所述肽序列,以生成所述一个或多个MHC等位基因中的每一个的依赖性分

数,所述依赖性分数指示所述MHC等位基因是否会基于所述肽序列的特定位置处的特定氨基酸呈递所述新抗原。

3.根据权利要求2所述的方法,其中将编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入所述机器学习呈递模型中进一步包括:

变换所述依赖性分数以产生每一MHC等位基因的相应独立等位基因可能性,由此指示所述相应MHC等位基因会呈递所述相应新抗原的可能性;及

将所述独立等位基因可能性组合以产生所述新抗原的呈递可能性。

4.根据权利要求3所述的方法,其中变换所述依赖性分数将所述新抗原的呈递建模为在所述一个或多个MHC等位基因之间相互排斥。

5.根据权利要求2所述的方法,其中将编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入所述机器学习呈递模型中进一步包括:

变换所述依赖性分数的组合以产生呈递可能性,其中变换所述依赖性分数的组合将所述新抗原的呈递建模为在所述一个或多个MHC等位基因之间存在干扰。

6.根据权利要求2-5中任一项所述的方法,其中所述呈递可能性集合通过至少一个或多个等位基因非相互作用特征进一步鉴别,并且进一步包括:

将所述机器学习呈递模型应用于所述等位基因非相互作用特征,以产生所述等位基因非相互作用特征的依赖性分数,所述依赖性分数指示所述相应新抗原的肽序列是否将基于所述等位基因非相互作用特征而被呈递。

7.根据权利要求6所述的方法,其进一步包括:

将所述一个或多个MHC等位基因的每个MHC等位基因的依赖性分数与所述等位基因非相互作用特征的依赖性分数组合;

变换每个MHC等位基因的所述组合的依赖性分数以产生每个MHC等位基因的独立等位基因可能性,由此指示所述相应MHC等位基因将呈递所述相应新抗原的可能性;及

组合所述独立等位基因可能性以产生所述呈递可能性。

8.根据权利要求6所述的方法,其进一步包括:

组合所述MHC等位基因中每一个的所述依赖性分数与所述等位基因非相互作用特征的所述依赖性分数;及

变换所述组合的依赖性分数以产生所述呈递可能性。

9.根据权利要求1-8中任一项所述的方法,其中所述一个或多个MHC等位基因包括两个或更多个不同的MHC等位基因。

10.根据权利要求1-9中任一项所述的方法,其中所述肽序列包括长度不为9个氨基酸的肽序列。

11.根据权利要求1-10中任一项所述的方法,其中编码肽序列包括使用独热编码方案编码所述肽序列。

12.根据权利要求1-11中任一项所述的方法,其中所述多个样品包括以下中的至少一种:

(a) 被工程改造成表达单个MHC等位基因的一种或多种细胞系;

- (b) 被工程改造成表达多个MHC等位基因的一种或多种细胞系;
- (c) 从多个患者获得或得到的一种或多种人细胞系;
- (d) 从多个患者获得的新鲜或冷冻的肿瘤样品;以及
- (e) 从多个患者获得的新鲜或冷冻的组织样品。

13. 根据权利要求1-12中任一项所述的方法, 其中所述训练数据集还包含以下中的至少一种:

- (a) 与所述肽中的至少一种的肽-MHC结合亲和力测量值相关的数据; 及
- (b) 与所述肽中的至少一种的肽-MHC结合稳定性测量值相关的数据。

14. 根据权利要求1-13中任一项所述的方法, 其中所述呈递可能性集合进一步通过至少由RNA-seq或质谱法测量的所述受试者中所述一个或多个MHC等位基因的表达水平鉴别。

15. 根据权利要求1-14中任一项所述的方法, 其中所述呈递可能性集合通过特征进一步鉴别, 所述特征包括以下中的至少一种:

- (a) 预测的所述新抗原集合中的新抗原与所述一个或多个MHC等位基因之间的亲和力; 及
- (b) 预测的所述新抗原编码的肽-MHC复合物的稳定性。

16. 根据权利要求1-15中任一项所述的方法, 其中所述数字可能性集合通过特征进一步鉴别, 所述特征包括以下中的至少一种:

- (a) 在源蛋白质序列内侧接所述新抗原编码肽序列的C末端序列; 及
- (b) 在源蛋白质序列内侧接所述新抗原编码肽序列的N末端序列。

17. 根据权利要求1-16中任一项所述的方法, 其中选择所述选定的新抗原的集合包括基于所述机器学习呈递模型, 选择被呈递在所述肿瘤细胞表面上的可能性相对于未选择的新抗原有所增加的新抗原。

18. 根据权利要求1-17中任一项所述的方法, 其中选择所述选定的新抗原的集合包括基于所述机器学习呈递模型, 选择能够在所述受试者体内诱导肿瘤特异性免疫应答的可能性相对于未选择的新抗原有所增加的新抗原。

19. 根据权利要求1-18中任一项所述的方法, 其中选择所述选定的新抗原的集合包括基于所述呈递模型, 选择能够被专职抗原呈递细胞 (APC) 呈递至天然T细胞的可能性相对于未选择的新抗原有所增加的新抗原, 任选地其中所述APC是树突状细胞 (DC)。

20. 根据权利要求1-19中任一项所述的方法, 其中选择所述选定的新抗原的集合包括基于所述机器学习呈递模型, 选择经历中枢或外周耐受性抑制的可能性相对于未选择的新抗原有所降低的新抗原。

21. 根据权利要求1-20中任一项所述的方法, 其中选择所述选定的新抗原的集合包括基于所述机器学习呈递模型, 选择能够在所述受试者体内诱导针对正常组织的自体免疫应答的可能性相对于未选择的新抗原有所降低的新抗原。

22. 根据权利要求1-21中任一项所述的方法, 其中所述一种或多种肿瘤细胞选自自由以下组成的组: 肺癌、黑素瘤、乳癌、卵巢癌、前列腺癌、肾癌、胃癌、结肠癌、睾丸癌、头颈癌、胰腺癌、脑癌、B细胞淋巴瘤、急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病和T细胞淋巴细胞性白血病、非小细胞肺癌和小细胞肺癌。

23. 根据权利要求1-22中任一项所述的方法, 其进一步包括从所述选定的新抗原的集

合产生用于构建个性化癌症疫苗的输出。

24. 根据权利要求23所述的方法, 其中所述用于所述个性化癌症疫苗的输出包括编码所述选定的新抗原的集合的至少一个肽序列或至少一个核苷酸序列。

25. 根据权利要求1-24中任一项所述的方法, 其中所述机器学习呈递模型是神经网络模型。

26. 根据权利要求25所述的方法, 其中所述神经网络模型包括包含布置在一个或多个层中的一系列节点的单个神经网络模型, 所述单个神经网络模型被配置为接收编码多个不同MHC等位基因的肽序列的数字矢量。

27. 根据权利要求26所述的方法, 其中所述神经网络模型通过更新所述神经网络模型的参数来进行训练。

28. 根据权利要求25-27中任一项所述的方法, 其中所述机器学习呈递模型是包括一个或多个节点层的深度学习模型。

29. 根据权利要求1-28中任一项所述的方法,
其中被编码为包含有关构成与所述样品的所述肽结合的至少一个MHC等位基因的多个氨基酸以及所述氨基酸在所述至少一个MHC等位基因中的位置的集合的信息的数字矢量的训练肽序列不包括: 被输入所述机器学习呈递模型中以生成所述新抗原集合的呈递可能性集合的受试者的MHC等位基因的肽序列。

30. 根据权利要求1-29中任一项所述的方法, 其中与所述训练数据集的所述多个样品中的每个样品的肽结合的至少一个MHC等位基因属于所述受试者的一个或多个MHC等位基因所属于的基因家族。

31. 根据权利要求1-30中任一项所述的方法, 其中与所述训练数据集的所述多个样品中的每个样品的肽结合的至少一个MHC等位基因包括一个MHC等位基因。

32. 根据权利要求1-30中任一项所述的方法, 其中与所述训练数据集的所述多个样品中的每个样品的肽结合的至少一个MHC等位基因包括多于一个MHC等位基因。

33. 根据权利要求1-32中任一项所述的方法, 其中所述一个或多个MHC等位基因是I类MHC等位基因。

34. 一种计算机系统, 其包括:

计算机处理器;

内存, 其用于存储计算机程序指令, 所述指令在由所述计算机处理器执行时会导致所述计算机处理器执行以下操作:

从受试者的肿瘤细胞和正常细胞获得外显子组、转录组或全基因组核苷酸测序数据中的至少一种, 其中所述核苷酸测序数据被用于获得代表通过比较来自所述肿瘤细胞的核苷酸测序数据和来自所述正常细胞的核苷酸测序数据鉴别的新抗原集合中每一种新抗原的肽序列的数据, 其中每种新抗原的肽序列包含至少一个使其不同于从所述受试者的正常细胞鉴别的相应野生型肽序列的变化;

将所述新抗原中的每一种的肽序列编码为相应的数字矢量, 每个数字矢量包含有关构成所述肽序列的多个氨基酸和所述氨基酸在所述肽序列中的位置的集合的信息;

从所述受试者的一个或多个MHC等位基因中的每一个获得外显子组、转录组或全基因组核苷酸测序数据中的至少一种, 其中所述核苷酸测序数据被用于获得代表所述受试者的

一个或多个MHC等位基因中的每一个的肽序列的数据；

将所述受试者的一个或多个MHC等位基因中的每一个的肽序列编码为相应的数字矢量,每个数字矢量包含有关构成所述肽序列的多个氨基酸和所述氨基酸在所述肽序列中的位置的集合的信息；

使用计算机处理器将编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型中,以产生对于所述新抗原集合的呈递可能性集合,所述集合中的每个呈递可能性代表了相应的新抗原被所述一个或多个MHC等位基因呈递在所述受试者的肿瘤细胞表面上的可能性,所述机器学习呈递模型包含：

至少基于训练数据集鉴别的多个参数,所述训练数据集包含：

对于多个样品中的每个样品,通过质谱测量与被鉴别为存在于所述样品中的MHC等位基因集合中的至少一个MHC等位基因结合的肽的存在而获得的标记；

对于每个样品,编码为包含有关构成所述肽的多个氨基酸和所述氨基酸在所述肽中的位置的集合的信息的数字矢量的训练肽序列；以及

对于每个样品,编码为包含有关构成与所述样品的所述肽结合的至少一个MHC等位基因的多个氨基酸以及所述氨基酸在所述至少一个MHC等位基因中的位置的集合的信息的数字矢量的训练肽序列；

代表作为输入接收的编码所述新抗原中的每一种的肽序列的数字矢量和编码所述一个或多个MHC等位基因中的每一个的肽序列的数字矢量与基于所述数字矢量和所述参数作为输出生成的呈递可能性之间的关系的函数；

基于所述呈递可能性集合选择所述新抗原集合的子集,以产生选定的新抗原的集合；以及

回收所述选定的新抗原的集合。

利用泛等位基因模型进行的新抗原鉴别

[0001] 本申请是申请号为201980019430.9,发明名称为“利用泛等位基因模型进行的新抗原鉴别”的专利申请的分案申请。

[0002] 相关申请的交叉引用

[0003] 本申请要求于2018年2月27日提交的美国临时申请号62/636,061的权益和优先权。上面引用的申请的内容通过引用整体并入。

背景技术

[0004] 基于肿瘤特异性新抗原的治疗性疫苗和T细胞疗法作为下一代个性化癌症免疫疗法具有广阔的前景。¹⁻³鉴于产生新抗原的可能性相对较高,具有高突变负荷的癌症,如非小细胞肺癌(NSCLC)和黑素瘤成为此类疗法的特别值得关注的靶标。^{4,5}早期有证据显示,基于新抗原的疫苗接种能够引起T细胞应答⁶并且靶向新抗原的T细胞疗法在某些情况下能够在选择的患者中引起肿瘤消退。⁷I类MHC和II类MHC都对T细胞响应具有影响⁷⁰⁻⁷¹。

[0005] 但是,鉴别新抗原和识别新抗原的T细胞已成为评估肿瘤响应^{77,110}、检查肿瘤进展¹¹¹和设计下一代个性化疗法¹¹²的主要挑战。目前的新抗原鉴别技术是费时和费力的^{84,96},或者不够精确^{87,91-93}。尽管最近已证明识别新抗原的T细胞是TIL的主要成分^{84,96,113,114}并且在癌症患者的外周血中循环¹⁰⁷,但目前用于鉴别新抗原反应性T细胞的方法具有以下三个局限性的组合:(1)其依赖于难以获得的临床试样,例如TIL^{97,98}或白细胞分离术(leukaphereses)¹⁰⁷(2)其需要筛选不切实际的大肽库⁹⁵或(3)其依赖于MHC多聚体,这实际上只对很小数量MHC等位基因数可用。

[0006] 另外,提出的初步方法并入了使用下一代测序的基于突变分析、RNA基因表达及候选新抗原肽的MHC结合亲和力预测⁸。然而,提出的这些方法都无法模拟整个表位产生过程,该过程除含有基因表达和MHC结合外,还含有许多步骤(例如TAP转运、蛋白酶体裂解、MHC结合、将肽-MHC复合物转运至细胞表面和/或TCR对MHC-I的识别;内吞或自噬、通过细胞外或溶酶体蛋白酶(例如组织蛋白酶)裂解、与CLIP肽竞争HLA-DM催化的HLA结合、将肽-MHC复合物转运至细胞表面和/或TCR对MHC-II的识别)⁹。因此,现有的方法可能会有低阳性预测值(PPV)降低的问题。(图1A)

[0007] 事实上,多个研究团队所进行的关于由肿瘤细胞呈递的肽的分析显示,预计使用基因表达和MHC结合亲和力呈递的肽中不到5%可以在肿瘤表面MHC上发现^{10,11}(图1B)。近期观察到的仅针对突变数量的检查点抑制剂反应无法提高对结合受限的新抗原的预测准确性进一步支持了结合预测与MHC呈递之间的这一低相关性。¹²

[0008] 现有的呈递预测方法的这一低阳性预测值(PPV)提出了有关基于新抗原的疫苗设计和基于新抗原的T细胞疗法的问题。如果使用PPV低的预测方法来设计疫苗,则大多数患者不太可能接受治疗性新抗原,且少数患者可能要接受一种以上新抗原(即使假设所有呈递的肽都具有免疫原性)。同样,如果治疗性T细胞是基于低PPV的预测设计的,则大多数患者不太可能接受对肿瘤新抗原具有反应性的T细胞,以及使用下游实验室技术在预测后鉴定预测性新抗原的时间和物理资源成本可能过高。因此,用当前方法进行新抗原疫苗接种

和T细胞疗法不太可能在众多具有肿瘤的受试者中取得成功。(图1C)

[0009] 此外,先前的方法仅使用顺式作用突变来产生候选新抗原,而在很大程度上忽视了考虑neo-ORF的其它来源,包括在多种肿瘤类型中出现且导致许多基因异常剪接的剪接因子突变¹³,及产生或移除蛋白酶裂解位点的突变。

[0010] 最后,由于文库构建、外显子组和转录组捕捉、测序或数据分析的条件并非最佳条件,故肿瘤基因组和转录组分析的标准方法可能会遗漏产生候选新抗原的体细胞突变。同样,标准肿瘤分析方法可能会无意中促成序列伪影或生殖系多态现象作为新抗原,而分别导致疫苗能力的低效使用或自身免疫的风险。

发明内容

[0011] 本文公开了一种鉴别和选择用于个性化癌症疫苗、用于T细胞疗法或其二者的新抗原的优化方法。首先,提出了使用下一代测序(NGS)鉴别新抗原候选物的优化的肿瘤外显子组和转录组分析方法。这些方法建立在标准NGS肿瘤分析方法的基础之上,以确保在所有类别的基因组变化内推进最高敏感性和特异性的新抗原候选物。其次,提出了选择高PPV新抗原的新颖方法来克服特异性问题并确保打算包括在疫苗中和/或作为T细胞疗法的靶标的新抗原较大可能地引发抗肿瘤免疫。取决于实施方案,这些方法包括训练的统计回归或非线性深度学习模型,该模型被配置成预测多种长度的肽的呈递,从而在泛等位基因(pan-allele)的基础上共享跨越不同长度的肽的统计强度。该模型能够预测肽将被任何MHC等位基因(包括该模型先前在训练过程中未遇到的未知MHC等位基因)呈递的可能性。非线性深度学习模型可以被特别地设计和训练成将同一细胞中的不同MHC等位基因视为独立的,由此解决了线性模型所具有的不同MHC等位基因会相互干扰的问题。最后,解决了基于新抗原的个性化疫苗设计和制造以及用于T细胞疗法的个性化新抗原特异性T细胞生产的其它需要考虑的问题。

[0012] 本文公开的模型优于在结合亲和力上训练的最新预测器和基于MS肽数据的早期预测器多达一个数量级。通过更可靠地预测肽的呈递,该模型可以使用临床实践方法以更具时间和成本效益的方式鉴别用于个性化疗法的新抗原特异性或肿瘤抗原特异性T细胞,其使用有限量的患者外周血,每位患者筛查少量的肽,并且不一定依赖于MHC多聚体。然而,在另一个实施方案中,本文公开的模型可通过减少为了鉴别新抗原或肿瘤抗原特异性T细胞而需要筛选的与MHC多聚体结合的肽的数目,来使用MHC多聚体以更具时间和成本效益的方式鉴别肿瘤抗原特异性T细胞。

[0013] 本文公开的模型在TIL新表位数据集上的预测性能和前瞻性新抗原反应性T细胞鉴别任务证明,现在有可能通过对HLA处理和呈递进行建模来获得治疗上有用的新表位预测。总而言之,这项作为抗原靶向免疫治疗提供了实用的计算机模拟抗原鉴别,从而加快了治愈患者的进程。

附图说明

[0014] 参照以下描述和附图将更好地理解本发明的这些和其它特征、方面及优势,在附图中:

[0015] 图1A显示当前用于鉴别新抗原的临床方法。

- [0016] 图1B显示<5%的预测结合肽被呈递在肿瘤细胞上。
- [0017] 图1C显示新抗原预测特异性问题的影响。
- [0018] 图1D显示结合预测不足以进行新抗原鉴别。
- [0019] 图1E显示MHC-I呈递的概率随肽长度的变化。
- [0020] 图1F显示由Promega动态范围标准(dynamic range standard)生成的示例性肽谱。
- [0021] 图1G显示添加特征如何增加模型阳性预测值。
- [0022] 图2A是根据一个实施方案,用于鉴别患者体内肽呈递的可能性的环境的概述。
- [0023] 图2B和2C示出了根据一个实施方案的获得呈递信息的方法。
- [0024] 图3是一个高级框图,示出了根据一个实施方案的呈递鉴别系统的计算机逻辑组件。
- [0025] 图4示出了根据一个实施方案的一组示例训练数据。
- [0026] 图5示出了与MHC等位基因相关联的示例网络模型。
- [0027] 图6示出了根据一个实施方案的由MHC等位基因共享的示例网络模型 $NN_H(\bullet)$ 。
- [0028] 图7示出了使用示例网络模型生成与一个MHC等位基因相关联的肽的呈递可能性。
- [0029] 图8示出了使用示例网络模型生成与一个MHC等位基因相关联的肽的呈递可能性。
- [0030] 图9示出了使用示例网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。
- [0031] 图10示出了使用示例网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。
- [0032] 图11示出了使用示例网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。
- [0033] 图12示出了使用示例网络模型生成与多个MHC等位基因相关联的肽的呈递可能性。
- [0034] 图13示出了根据一个实施方案的由MHC等位基因共享的示例网络模型 $NN_H(\bullet)$ 。
- [0035] 图14示出了与MHC等位基因不相关联的示例网络模型。
- [0036] 图15示出了使用由MHC等位基因共享的示例网络模型来产生与MHC等位基因相关联的肽的呈递可能性。
- [0037] 图16示出了针对第一测试样品的由包含神经网络并且在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线,以及针对第一测试样品的由包含神经网络但没有在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线。
- [0038] 图17示出了针对第二测试样品的由包含神经网络并且在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线,以及针对第二测试样品的由包含神经网络但没有在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线。
- [0039] 图18示出了针对第三测试样品的由包含神经网络并且在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线,以及针对第三测试样品的由包含神经网络但没有在包含测试的HLA等位基因的样品上加以训练的泛等位基因模型输出的精确率/召回率曲线。

[0040] 图19示出了由包含神经网络的泛等位基因模型、随机森林模型(random forest model)、二次判别模型(quadratic discriminant model),以及在包含测试的HLA等位基因的样品上加以训练的MHCFlurry模型输出的精确率/召回率曲线。

[0041] 图20示出了第一测试样品的由包含神经网络的泛等位基因模型、随机森林模型、二次判别模型,以及在包含测试的HLA等位基因的样品上加以训练的MHCFlurry模型输出的精确率/召回率曲线。

[0042] 图21示出了第二测试样品的由包含神经网络的泛等位基因模型、随机森林模型、二次判别模型,以及在包含测试的HLA等位基因的样品上加以训练的MHCFlurry模型输出的精确率/召回率曲线。

[0043] 图22示出了第三测试样品的由包含神经网络的泛等位基因模型、随机森林模型、二次判别模型,以及在包含测试的HLA等位基因的样品上加以训练的MHCFlurry模型输出的精确率/召回率曲线。

[0044] 图23A示出了NSCLC患者中突变负荷的样品频率分布。

[0045] 图23B示出了根据一个实施方案的基于患者是否满足最小突变负荷的纳入标准选择的患者的模拟疫苗中呈递的新抗原的数量。

[0046] 图23C根据一个实施方案比较了与包含基于呈递模型鉴别的治疗子集的疫苗相关的所选患者和与包含通过现有技术水平模型鉴别的治疗子集的疫苗相关的所选患者之间的模拟疫苗中呈递的新抗原的数量。

[0047] 图23D比较了与包含基于HLA-A*02:01的单独等位基因呈递模型鉴别的治疗子集的疫苗相关的所选患者和与包含基于HLA-A*02:01和HLA-B*07:02的双独立等位基因呈递模型鉴别的治疗子集的疫苗相关的所选患者之间的模拟疫苗中呈递的新抗原的数量。根据一个实施方案,疫苗容量设置为 $v=20$ 个表位。

[0048] 图23E根据一个实施方案比较了基于突变负荷选择的患者和通过期望效用分数选择的患者之间的模拟疫苗中呈递的新抗原的数量。

[0049] 图24比较了当在五个留出测试样品上测试使用呈递热点参数的泛等位基因呈递模型和不使用呈递热点参数的泛等位基因呈递模型时该泛等位基因模型在40%召回率下的阳性预测值(PPV)。

[0050] 图25A比较了对于包含12个不同测试样品(每个测试样品均取自具有至少一种预先存在的T细胞应答的患者)的测试集而言,在排名前5位、前10位和前20位的体细胞突变中由T细胞(例如,预先存在的T细胞应答)识别的体细胞突变所占的比例,所述排名前5位、前10位和前20位的体细胞突变是使用标准HLA结合亲和力预测以 >2 TPM阈值对如通过RNA-seq、等位基因特异性神经网络模型和泛等位基因神经网络模型测定的基因表达鉴别出的。

[0051] 图25B比较了对于包含12个不同测试样品(每个测试样品均取自具有至少一种预先存在的T细胞应答的患者)的测试集而言,在排名前5位、前10位和前20位的最小新表位中由T细胞(例如,预先存在的T细胞应答)识别的最小新表位所占的比例,所述排名前5位、前10位和前20位的最小新表位是使用标准HLA结合亲和力预测以 >2 TPM阈值对如通过RNA-seq、等位基因特异性神经网络模型和泛等位基因神经网络模型测定的基因表达鉴别出的。

[0052] 图26A描绘了对于9位患者的患者特异性新抗原肽库的T细胞应答的检测。

[0053] 图26B描绘了对于4位患者的个体患者特异性新抗原肽的T细胞应答的检测。

- [0054] 图26C描绘了患者CU04的ELISpot孔的示例图像。
- [0055] 图27A描绘了来自在HLA匹配的健康供体中用新抗原进行的对照实验的结果。
- [0056] 图27B描绘了来自在HLA匹配的健康供体中用新抗原进行的对照实验的结果。
- [0057] 图28描绘了对于图26A中描绘的每个供体和每个体外扩增,对PHA阳性对照的T细胞应答的检测。
- [0058] 图29A描绘了针对患者CU04对库#2中的每个个体患者特异性新抗原肽的T细胞应答的检测。
- [0059] 图29B描绘了对于患者CU04的三次访问中的每一次访问以及对于患者1-024-002的两次访问中的每一次访问,针对对个体患者特异性新抗原肽的T细胞应答的检测,每次访问发生在不同的时间点。
- [0060] 图29C描绘了对于患者CU04的两次访问中的每一次访问以及对于患者1-024-002的两次访问中的每一次访问,针对对个体患者特异性新抗原肽和患者特异性新抗原肽库的T细胞应答的检测,每次访问发生在不同的时间点。
- [0061] 图30描绘了对于图26A的患者,对两个患者特异性新抗原肽库和DMSO阴性对照的T细胞应答的检测。
- [0062] 图31A描绘了泛等位基因模型和等位基因特异性模型的针对每个包含II类MHC等位基因的测试样品0的精确率-召回率曲线。
- [0063] 图31B描绘了泛等位基因模型和等位基因特异性模型的针对每个包含II类MHC等位基因的测试样品1的精确率-召回率曲线。
- [0064] 图31C描绘了泛等位基因模型和等位基因特异性模型的针对每个包含II类MHC等位基因的测试样品2的精确率-召回率曲线。
- [0065] 图31D描绘了泛等位基因模型和等位基因特异性模型的针对每个包含II类MHC等位基因的测试样品4的精确率-召回率曲线。
- [0066] 图32描绘了对来自NSCLC患者的外周血的新抗原特异性记忆T细胞的TCR进行测序的方法。
- [0067] 图33描绘了用于将TCR引入到受体细胞中的TCR构建体的示例性实施方案。
- [0068] 图34描绘了用于将TCR克隆到表达系统中以进行疗法开发的示例性P526构建体骨架核苷酸序列。
- [0069] 图35描绘了用于将患者新抗原特异性TCR,克隆型1TCR克隆到表达系统中以进行疗法开发的示例性构建体序列。
- [0070] 图36描绘了用于将患者新抗原特异性TCR,克隆型3克隆到表达系统中以进行疗法开发的示例性构建体序列。
- [0071] 图37是根据实施方案的用于向患者提供定制的新抗原特异性治疗的方法的流程图。
- [0072] 图38示出了用于实施图1和3中所示实体的示例计算机。

具体实施方式

[0073] I. 定义

[0074] 一般说来,权利要求书和说明书中使用的术语意图解释为具有与本领域普通技术

人员所理解的普通含义。为清楚起见,以下定义某些术语。如果普通含义与所提供的定义之间存在矛盾,应使用所提供的定义。

[0075] 如本文所使用,术语“抗原”是诱导免疫应答的物质。

[0076] 如本文所使用,术语“新抗原”是具有至少一个使其不同于相应野生型亲本抗原的变化的抗原,例如,该变化是肿瘤细胞突变或肿瘤细胞特异性翻译后修饰。新抗原可以包括多肽序列或核苷酸序列。突变可以包括移码或非移码插入缺失、错义或无义取代、剪接位点变化、基因组重排或基因融合,或产生neoORF的任何基因组或表达变化。突变还可以包括剪接变体。肿瘤细胞特异性翻译后修饰可以包括异常磷酸化。肿瘤细胞特异性翻译后修饰还可以包括蛋白酶体产生的剪接抗原。参见Liepe等人,A large fraction of HLA class II ligands are proteasome-generated spliced peptides;Science.2016 Oct 21;354(6310):354-358。

[0077] 如本文所使用,术语“肿瘤新抗原”是存在于受试者的肿瘤细胞或组织中但不存在于受试者的相应正常细胞或组织中的新抗原。

[0078] 如本文所使用,术语“基于新抗原的疫苗”是基于一个或多个新抗原,例如多种新抗原的疫苗构建体。

[0079] 如本文所使用,术语“候选新抗原”是产生可以代表新抗原的新序列的突变或其它异常。

[0080] 如本文所使用,术语“编码区”是基因中编码蛋白质的部分。

[0081] 如本文所使用,术语“编码突变”是在编码区中存在的突变。

[0082] 如本文所使用,术语“ORF”是指开放阅读框。

[0083] 如本文所使用,术语“NEO-ORF”是由突变或其它异常如剪接而产生的肿瘤特异性ORF。

[0084] 如本文所使用,术语“错义突变”是导致一个氨基酸被另一个氨基酸取代的突变。

[0085] 如本文所使用,术语“无义突变”是导致一个氨基酸被终止密码子取代的突变。

[0086] 如本文所使用,术语“移码突变”是导致蛋白质框架改变的突变。

[0087] 如本文所使用,术语“插入缺失”是一个或多个核酸的插入或缺失。

[0088] 如本文在两个或更多个核酸或多肽序列的情况下使用的术语“同一性”百分比是指当比较并对准达到最大对应性时,如使用以下描述的序列比较算法(例如BLASTP和BLASTN,或技术人员可用的其它算法)之一测量或通过目测检查得到的两个或更多个序列或子序列具有指定百分比的核苷酸或氨基酸残基是相同的。取决于应用,“同一性”百分比可以存在于所比较的序列的某一区域内,例如在功能结构域内,或者存在于待比较的两个序列的全长内。

[0089] 为进行序列比较,通常,一个序列充当参考序列,以与测试序列相比较。当使用序列比较算法时,将测试序列和参考序列输入计算机,必要时指定子序列座标,并且指定序列算法程序参数。然后,序列比较算法基于指定的程序参数计算测试序列相对于参考序列的序列同一性百分比。或者,可以通过组合在所选序列位置(例如序列基元)处特定核苷酸,或对于翻译的序列来说特定氨基酸的存在或不存在来确定序列相似性或不相似性。

[0090] 用于比较的最佳序列比对可以例如通过Smith和Waterman,Adv. Appl. Math. 2:482(1981)的局部同源性算法;Needleman和Wunsch,J. Mol. Biol. 48:443(1970)的同源性比对

算法;Pearson和Lipman,Proc.Nat'l.Acad.Sci.USA 85:2444(1988)的相似性搜索方法;这些算法的计算机化实施(Wisconsin Genetics软件包中的GAP、BESTFIT、FASTA和TFASTA;Genetics Computer Group,575 Science Dr.,Madison,Wis.);或通过目测检查(一般参见Ausubel等人,见下文)来进行。

[0091] 适于测定序列同一性和序列相似性百分比的算法的一个实例是Altschul等人,J.Mol.Biol.215:403-410(1990)中描述的BLAST算法。执行BLAST分析的软件通过National Center for Biotechnology Information公开可用。

[0092] 如本文所使用,术语“无终止或通读”是导致天然终止密码子移除的突变。

[0093] 如本文所使用,术语“表位”是抗原中通常由抗体或T细胞受体结合的特定部分。

[0094] 如本文所使用,术语“免疫原性”是例如通过T细胞、B细胞或两者引发免疫应答的能力。

[0095] 如本文所使用,术语“HLA结合亲和力”、“MHC结合亲和力”意思指特定抗原与特定MHC等位基因之间的结合亲和力。

[0096] 如本文所使用,术语“诱饵(bait)”是用于自样品富集特定DNA或RNA序列的核酸探针。

[0097] 如本文所使用,术语“变体”是受试者的核酸与用作对照的参考人基因组之间的差异。

[0098] 如本文所使用,术语“变体识别(variant call)”是对通常由测序确定的变体存在的算法确定。

[0099] 如本文所使用,术语“多态现象”是生殖系变体,即,在个体的所有带有DNA的细胞中所发现的变体。

[0100] 如本文所使用,术语“体细胞变体”是在个体的非生殖系细胞中产生的变体。

[0101] 如本文所使用,术语“等位基因”是基因的一种形式,或是基因序列的一种形式,或是蛋白质的一种形式。

[0102] 如本文所使用,术语“HLA型”是HLA基因等位基因的互补序列。

[0103] 如本文所使用,术语“无义介导的衰变”或“NMD”是由过早终止密码子引起的细胞对mRNA的降解。

[0104] 如本文所使用,术语“躯干突变”是起源于肿瘤发展早期且存在于大多数肿瘤细胞中的突变。

[0105] 如本文所使用,术语“亚克隆突变”是起源于肿瘤发展后期且仅存在于一小部分肿瘤细胞中的突变。

[0106] 如本文所使用,术语“外显子组”是编码蛋白质的基因组的子组。外显子组可以是基因组的全体外显子。

[0107] 如本文所使用,术语“逻辑回归”是由统计得到的二进制数据的回归模型,其中因变量等于1的概率的分对数被建模为因变量的线性函数。

[0108] 如本文所使用,术语“神经网络”是用于分类或回归的机器学习模型,由多层线性变换,继之以通常通过随机梯度下降和反向传播训练的逐元素非线性组成。

[0109] 如本文所使用,术语“蛋白质组”是由细胞、细胞群或个体表达和/或翻译的所有蛋白质的集合。

[0110] 如本文所使用,术语“肽组”是由MHC-I或MHC-II呈递于细胞表面上的所有肽的集合。肽组可以指一个细胞或一组细胞(例如肿瘤肽组,意思指构成肿瘤的所有细胞的肽组的联合)的特性。

[0111] 如本文所使用,术语“ELISPOT”意思指酶联免疫吸附斑点测定,这是一种用于监测人和动物的免疫应答的常用方法。

[0112] 如本文所使用,术语“dextramer”是在流式细胞术中用于抗原特异性T细胞染色的基于葡聚糖的肽-MHC多聚体。

[0113] 如本文所用,术语“MHC多聚体”是包含多个肽-MHC单体单元的肽-MHC复合物。

[0114] 如本文所用,术语“MHC四聚体”是包含四个肽-MHC单体单元的肽-MHC复合物。

[0115] 如本文所使用,术语“耐受性或免疫耐受性”是对一种或多种抗原,例如自身抗原免疫无反应性的状态。

[0116] 如本文所使用,术语“中枢耐受性”是通过缺失自身反应性T细胞克隆或通过促进自身反应性T细胞克隆分化成免疫抑制性调控性T细胞(Treg)而在胸腺中经历的耐受性。

[0117] 如本文所使用,术语“外周耐受性”是通过使经历中枢耐受性而存活的自身反应性T细胞下调或无反应性(anergizing),或通过促进这些T细胞分化成Treg而在外周经历的耐受性。

[0118] 术语“样品”可以包括借助于包括静脉穿刺、排泄、射精、按摩、活组织检查、针抽吸、灌洗样品、刮取、手术切口或干预在内的手段,或本领域中已知的其它手段从受试者获取单个细胞或多个细胞,或细胞碎片,或体液等分试样。

[0119] 术语“受试者”涵盖细胞、组织或生物体、人或非人,无论是体内、离体还是体外,雄性还是雌性的。术语受试者包括含人在内的哺乳动物。

[0120] 术语“哺乳动物”涵盖人和非人两种,并且包括但不限于人、非人灵长类动物、犬科动物、猫科动物、鼠科动物、牛科动物、马科动物及猪科动物。

[0121] 术语“临床因素”是指受试者状况,例如疾病活动性或严重程度的量度。“临床因素”涵盖受试者健康状况的所有标志物,包括非样品标志物,和/或受试者的其它特征,如但不限于年龄和性别。临床因素可以是能通过在确定条件下评价来自受试者的一个样品(或样品群)或受试者而获得的分数、一个值或一组值。临床因素也可以由标志物和/或如基因表达替代物之类其它参数进行预测。临床因素可以包括肿瘤类型、肿瘤亚型和吸烟史。

[0122] 缩写:MHC:主要组织相容性复合物;HLA:人白细胞抗原或人MHC基因座;NGS:下一代测序;PPV:阳性预测值;TSNA:肿瘤特异性新抗原;FFPE:福尔马林固定、石蜡包埋;NMD:无义介导的衰变;NSCLC:非小细胞肺癌;DC:树突状细胞。

[0123] 除非上下文另外清楚地规定,否则如本说明书和所附权利要求中所使用,单数形式“一个(种)(a/an)”和“所述”包括多个参照物。

[0124] 本文中未直接定义的任何术语应理解为具有与本发明领域内所理解的通常与之相关的含义。本文论述的某些术语是为了向从业人员描述本发明各方面的组合物、装置、方法等以及其制备或使用提供额外的指导。应了解,相同的事物可以按超过一种方式表示。因此,替代性措辞和同义词可以用于本文所论述的任一个或多个术语。无论本文中是否阐述或论述术语都无关紧要。提供了一些同义词或可取代的方法、材料等。除非明确陈述,否则对一个或数个同义词或等效物的叙述不排除其它同义词或等效物的使用。实例,包括术语

实例的使用只是出于说明的目的,且并非在本文中限制本发明各方面的范围和含义。

[0125] 说明书正文内引用的所有参考文献、颁布的专利和专利申请都是以引用的方式整体并入本文中用于所有目的。

[0126] II. 鉴别新抗原的方法

[0127] 本文公开了用于鉴别来自受试者的一个或多个肿瘤细胞的可能被一个或多个MHC等位基因呈递在所述肿瘤细胞表面上的至少一种新抗原的方法。所述方法包括从受试者的肿瘤细胞以及正常细胞获得外显子组、转录组和/或全基因组核苷酸测序数据。该核苷酸测序数据被用于获得新抗原集合中的每种新抗原的肽序列。通过比较来自肿瘤细胞的核苷酸测序数据和来自正常细胞的核苷酸测序数据来鉴别新抗原集合。具体地,新抗原集合中的每种新抗原的肽序列包含至少一个使其不同于从受试者的正常细胞鉴别的相应野生型肽序列的变化。所述方法进一步包括将新抗原集合中每种新抗原的肽序列编码成相应的数字矢量。每个数字矢量包含描述构成肽序列的氨基酸和该氨基酸在肽序列中的位置的信息。所述方法进一步包括从受试者肿瘤细胞获得外显子组、转录组和/或全基因组核苷酸测序数据。这种核苷酸测序数据被用于获得受试者的一个或多个MHC等位基因中的每一个的肽序列。将受试者的一个或多个MHC等位基因中的每一个的肽序列编码为相应的数字矢量。每个数字矢量包含描述构成MHC等位基因肽序列的氨基酸以及这些氨基酸在MHC等位基因的肽序列中的位置的信息。所述方法进一步包括将编码新抗原中的每一种的肽序列的数字矢量以及编码一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型中,以产生对于新抗原集合中每种新抗原的呈递可能性。每个呈递可能性代表了相应的新抗原被一个或多个MHC等位基因呈递在受试者的肿瘤细胞表面上的可能性。机器学习呈递模型包含多个参数和函数。所述多个参数基于训练数据集鉴别。所述训练数据集包含:对于多个样品中的每个样品,通过质谱测量与被鉴别为存在于所述样品中的MHC等位基因集合中的至少一个MHC等位基因结合的肽的存在而获得的标记、编码为包含描述构成肽的氨基酸和该氨基酸在肽中的位置的信息的数字矢量的训练肽序列,以及编码为包含描述构成与样品的肽结合的至少一个MHC等位基因的氨基酸和该氨基酸在MHC等位基因肽中的位置的信息的数字矢量的训练肽序列。所述函数代表由机器学习呈递模型作为输入接收的所述数字矢量与由机器学习呈递模型基于所述数字矢量和所述多个参数作为输出生成的呈递可能性之间的关系。所述方法进一步包括基于所述呈递可能性选择所述新抗原集合的子集,以产生选定的新抗原的集合;以及回收所选定新抗原的集合。

[0128] 在一些实施方案中,将编码新抗原中的每一种的肽序列的数字矢量和编码一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型包括:将机器学习呈递模型应用于新抗原的肽序列并应用于一个或多个MHC等位基因的肽序列,以生成一个或多个MHC等位基因中的每一个的依赖性分数。MHC等位基因的依赖性分数指示MHC等位基因是否会基于肽序列的特定位置处的特定氨基酸呈递新抗原。在进一步的实施方案中,将编码新抗原中的每一种的肽序列的数字矢量和编码一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型进一步包括:变换依赖性分数以得到每一MHC等位基因的相应独立等位基因可能性,由此指示相应MHC等位基因会呈递相应新抗原的可能性;及将独立等位基因可能性组合以产生新抗原的呈递可能性。在一些实施方案中,变换依赖性分数将新抗原的呈递建模为在一个或多个MHC等位基因之间相互排斥。在可替代实

施方案中,将编码新抗原中的每一种的肽序列的数字矢量和编码一个或多个MHC等位基因中的每一个的肽序列的数字矢量输入机器学习呈递模型进一步包括:变换依赖性分数的组合以产生呈递可能性。在这样的实施方案中,变换依赖性分数的组合将新抗原的呈递建模为在一个或多个MHC等位基因之间进行干扰。

[0129] 在一些实施方案中,呈递可能性的集合通过至少一个或多个等位基因非相互作用特征进一步鉴别。在这样的实施方案中,所述方法进一步包括将机器学习呈递模型应用于等位基因非相互作用特征,以产生所述等位基因非相互作用特征的依赖性分数。依赖性分数指示相应新抗原的肽序列是否将基于所述等位基因非相互作用特征而被呈递。在一些实施方案中,所述方法进一步包括将一个或多个MHC等位基因中的每个MHC等位基因的依赖性分数与等位基因非相互作用特征的依赖性分数组合,变换每个MHC等位基因的组的依赖性分数以产生每个MHC等位基因的独立等位基因可能性,及组合独立等位基因可能性以产生呈递可能性。MHC等位基因的独立等位基因可能性指示MHC等位基因将呈递相应新抗原的可能性。在替代实施方案中,所述方法进一步包括组合每个MHC等位基因的依赖性分数与所述等位基因非相互作用特征的依赖性分数;及变换组合的依赖性分数以产生呈递可能性。

[0130] 在一些实施方案中,所述一个或多个MHC等位基因包括两个或更多个不同的MHC等位基因。

[0131] 在一些实施方案中,肽序列包含长度不为9个氨基酸的肽序列。

[0132] 在一些实施方案中,编码肽序列包括使用独热编码方案编码所述肽序列。

[0133] 在一些实施方案中,所述多个样品包括以下至少一种:被工程改造成表达单个MHC等位基因的细胞系;被工程改造成表达多个MHC等位基因的细胞系;从多个患者获得或得到的人细胞系;从多个患者获得的新鲜或冷冻的肿瘤样品;以及从多个患者获得的新鲜或冷冻的组织样品。

[0134] 在一些实施方案中,所述训练数据集还包含以下至少一种:与所述肽中的至少一个的肽-MHC结合亲和力测量值相关的数据;及与所述肽中的至少一个的肽-MHC结合稳定性测量值相关的数据。

[0135] 在一些实施方案中,呈递可能性集合进一步通过如由RNA-seq或质谱法测量的所述受试者中一个或多个MHC等位基因的表达水平鉴别。

[0136] 在一些实施方案中,呈递可能性集合通过特征进一步鉴别,所述特征包括以下中的至少一种:预测的所述新抗原集合中的新抗原与一个或多个MHC等位基因之间的亲和力;及预测的新抗原编码的肽-MHC复合物的稳定性。

[0137] 在一些实施方案中,所述数字可能性集合通过特征进一步鉴别,所述特征包括以下中的至少一种:在其源蛋白质序列内侧接所述新抗原编码肽序列的C末端序列;及在其源蛋白质序列内侧接所述新抗原编码肽序列的N末端序列。

[0138] 在一些实施方案中,选择所述选定的新抗原的集合包括基于机器学习呈递模型,选择被呈递在所述肿瘤细胞表面上的可能性相对于未选择的新抗原有所增加的新抗原。

[0139] 在一些实施方案中,选择所述选定的新抗原的集合包括基于机器学习呈递模型,选择能够在受试者体内诱导肿瘤特异性免疫应答的可能性相对于未选择的新抗原有所增加的新抗原。

[0140] 在一些实施方案中,选择所述选定的新抗原的集合包括基于呈递模型,选择能够

被专职抗原呈递细胞 (APC) 呈递至天然T细胞的可能性相对于未选择的新抗原有所增加的新抗原。在这样的实施方案中,APC任选地是树突状细胞 (DC)。

[0141] 在一些实施方案中,选择所述选定的新抗原的集合包括基于机器学习呈递模型,选择经历中枢或外周耐受性抑制的可能性相对于未选择的新抗原有所降低的新抗原。

[0142] 在一些实施方案中,选择所述选定的新抗原的集合包括基于机器学习呈递模型,选择能够在所述受试者体内诱导针对正常组织的自体免疫应答的可能性相对于未选择的新抗原有所降低的新抗原。

[0143] 在一些实施方案中,所述一种或多种肿瘤细胞选自以下组成的组:肺癌、黑素瘤、乳癌、卵巢癌、前列腺癌、肾癌、胃癌、结肠癌、睾丸癌、头颈癌、胰腺癌、脑癌、B细胞淋巴瘤、急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病和T细胞淋巴细胞性白血病、非小细胞肺癌和小细胞肺癌。

[0144] 在一些实施方案中,该方法进一步包括从所述选定的新抗原的集合产生用于构建个性化癌症疫苗的输出。在这样的实施方案中,个性化癌症疫苗的输出可包括编码所述选定的新抗原的集合的至少一个肽序列或至少一个核苷酸序列。

[0145] 在一些实施方案中,机器学习呈递模型是神经网络模型。在这样的实施方案中,神经网络模型可以是包括布置在一个或多个层中的一系列节点的单个神经网络模型。该单个神经网络模型可以被配置为接收编码多个不同MHC等位基因的肽序列的数字矢量。在这样的实施方案中,神经网络模型可通过更新神经网络模型的参数来进行训练。在一些实施方案中,机器学习呈递模型可以是包括一个或多个节点层的深度学习模型。

[0146] 在一些实施方案中,被编码为包括有关构成与样品的肽结合的至少一个MHC等位基因的多个氨基酸以及该氨基酸在该至少一个MHC等位基因中的位置的集合的信息的数字矢量的训练肽序列不包括:被输入机器学习呈递模型中以产生新抗原集合的呈递可能性集合的受试者的MHC等位基因的肽序列。

[0147] 在本文公开的某些方面,与训练数据集的多个样品中的每个样品的肽结合的至少一个MHC等位基因属于受试者的一个或多个MHC等位基因所属于的基因家族。

[0148] 在一些实施方案中,与训练数据集的多个样品中的每个样品的肽结合的至少一个MHC等位基因包括一个MHC等位基因。在替代实施方案中,与训练数据集的多个样品中的每个样品的肽结合的至少一个MHC等位基因包括多于一个MHC等位基因。

[0149] 在一些实施方案中,该一个或多个MHC等位基因是I类MHC等位基因。

[0150] 本文还公开了计算机系统,其包括计算机处理器和存储计算机程序指令的内存,该计算机程序指令在被计算机处理器执行时使计算机处理器执行上述方法的实施方案。

[0151] III. 鉴别新抗原中的肿瘤特异性突变

[0152] 本文还公开了用于鉴别某些突变(例如癌细胞中存在的变体或等位基因)的方法。确切地说,这些突变可以存在于患有癌症的受试者的癌细胞的基因组、转录组、蛋白质组或外显子组中,但不存在于受试者的正常组织中。

[0153] 若肿瘤中的基因突变仅导致肿瘤中蛋白质的氨基酸序列改变,则认为这些突变可用于免疫靶向肿瘤。有用的突变包括:(1) 导致蛋白质中的氨基酸不同的非同义突变;(2) 通读突变,其中终止密码子被修饰或缺失,导致翻译得到在C末端具有新肿瘤特异性序列的较长蛋白质;(3) 导致在成熟mRNA中包括内含子且由此产生独特肿瘤特异性蛋白质序列的剪

接位点突变；(4) 产生在2种蛋白质的接合处具有肿瘤特异性序列的嵌合蛋白的染色体重排(即, 基因融合)；(5) 产生具有新肿瘤特异性蛋白质序列的新开放阅读框的移码突变或缺失。突变还可以包括非移码插入缺失、错义或无义取代、剪接位点变化、基因组重排或基因融合, 或产生neoORF的任何基因组或表达变化中的一种或多种。

[0154] 在肿瘤细胞中具有突变的肽或由例如剪接位点突变、移码突变、通读突变或基因融合突变产生的突变多肽可以通过对肿瘤和正常细胞中的DNA、RNA或蛋白质进行测序来鉴别。

[0155] 突变还可以包括先前鉴别的肿瘤特异性突变。已知的肿瘤突变可以见于癌症体细胞突变目录(Catalogue of Somatic Mutations in Cancer, COSMIC) 数据库。

[0156] 多种方法可用于检测个体的DNA或RNA中特定突变或等位基因的存在。本领域中的改进之处在于提供准确、容易且便宜的大规模SNP基因分型。举例来说, 已描述若干技术, 包括动态等位基因特异性杂交(DASH)、微板阵列对角线凝胶电泳(microplate array diagonal gel electrophoresis, MADGE)、焦磷酸测序、寡核苷酸特异性连接、TaqMan系统以及各种DNA“芯片”技术, 如Affymetrix SNP芯片。这些方法通常通过PCR扩增靶基因区。一些其它的方法基于通过侵袭式裂解产生小信号分子, 随后进行质谱法或固定化挂锁探针(padlock probe)和滚环扩增。本领域中已知用于检测特定突变的若干方法概述于下。

[0157] 基于PCR的检测手段可以包括同时多重扩增多个标志物。举例来说, 本领域中众所周知, 选择PCR引物产生尺寸不重叠且可以同时分析的PCR产物。或者, 可用以不同方式标记且由此可以通过不同方式检测的引物扩增不同标志物。当然, 基于杂交的检测手段能够以不同方式检测样品中的多个PCR产物。本领域中已知能够多重分析多个标志物的其它技术。

[0158] 已经开发出数种方法来促进基因组DNA或细胞RNA中单核苷酸多态性的分析。举例来说, 可以通过使用专用的核酸外切酶抗性核苷酸检测单碱基多态性, 如例如Mundy, C.R. (美国专利第4,656,127号)中所公开的。根据该方法, 与紧靠多态性位点3'端的等位基因序列互补的引物能够与从特定动物或人获得的靶分子杂交。如果靶分子上的多态性位点含有与存在的特定核酸外切酶抗性核苷酸衍生物互补的核苷酸, 则该衍生物将被合并至杂交引物的末端上。此类合并使得引物对核酸外切酶具有抗性, 并由此允许其检测。由于样品的核酸外切酶抗性衍生物的身份是已知的, 故引物对核酸外切酶产生抗性的发现披露, 靶分子多态性位点中存在的核苷酸与反应中使用的核苷酸衍生物互补。该方法的优势在于, 它不需要测定大量无关的序列数据。

[0159] 可以使用基于溶液的方法来确定多态性位点的核苷酸的身份。Cohen, D. 等人(法国专利2,650,840; PCT申请第W091/02087号)。如在美国专利第4,656,127号的Mundy方法中所述, 采用与紧靠多态性位点3'端的等位基因序列互补的引物。该方法使用标记过的双脱氧核苷酸衍生物来确定该位点的核苷酸的身份, 如果与多态性位点的核苷酸互补, 则该核苷酸将被合并至引物末端上。Goelet, P. 等人(PCT申请第92/15712号)描述了一种替代性方法, 称为遗传位点分析(Genetic Bit Analysis)或GBA。Goelet, P. 等人的方法使用了标记过的终止子和与在多态性位点3'端的序列互补的引物的混合物。由此通过存在于所评价靶分子的多态性位点中的核苷酸来确定合并的标记过的终止子并且该终止子与存在于所评价靶分子的多态性位点中的核苷酸互补。与Cohen等人(法国专利2,650,840; PCT申请第W091/02087号)的方法相比, Goelet, P. 等人的方法可以是非均相测定, 其中引物或靶分子

被固定于固相。

[0160] 已描述数种引物引导的用于测定DNA中的多态性位点的核苷酸并入程序(Komher, J.S.等人, Nucl. Acids. Res. 17:7779-7784 (1989); Sokolov, B.P., Nucl. Acids Res. 18:3671 (1990); Syvanen, A.-C.等人, Genomics 8:684-692 (1990); Kuppaswamy, M.N.等人, Proc. Natl. Acad. Sci. (U.S.A.) 88:1143-1147 (1991); Prezant, T.R.等人, Hum. Mutat. 1:159-164 (1992); Ugozzoli, L.等人, GATA 9:107-112 (1992); Nyren, P.等人, Anal. Biochem. 208:171-175 (1993))。这些方法与GBA的不同之处在于,它们利用并入经过标记的脱氧核苷酸来区别多态性位点处的碱基。在此类形式中,由于信号与并入的脱氧核苷酸的数量成比例,故在同一核苷酸的操作中出现的多态现象可以产生与该操作的长度成比例的信号(Syvanen, A.-C.等人, Amer. J. Hum. Genet. 52:46-59 (1993))。

[0161] 许多方案直接从数百万个独立DNA或RNA分子中并行获得序列信息。实时单分子边合成边测序技术依赖于荧光核苷酸的检测,因为这些核苷酸被并入与测序模板互补的新生DNA链中。在一种方法中,将30-50个碱基长度的寡核苷酸以5'端共价锚定至玻璃盖玻片上。这些锚定链执行两种功能。首先,如果模板被配置成具有与表面结合的寡核苷酸互补的捕捉尾部,则其充当靶模板链的捕捉位点。这些锚定链还充当模板引导的引物延伸的引物,形成序列读取的基础。捕捉引物用作固定位点以便使用多个合成、检测以及染料-连接子化学裂解以移除染料的循环进行序列测定。每个循环由添加聚合酶/标记过得核苷酸混合物、冲洗、成像及染料裂解组成。在一种替代方法中,聚合酶被修饰成具有荧光供体分子并且被固定于玻璃载片上,而各核苷酸用附接至 γ -磷酸的受体荧光部分进行颜色编码。当核苷酸被并入从头合成的链中时,该系统检测荧光标记的聚合酶与荧光修饰的核苷酸之间的相互作用。还存在其它边合成边测序技术。

[0162] 任何适合的边合成边测序平台都可以用于鉴别突变。如上文所描述,目前有四个主要的边合成边测序平台:来自Roche/454 Life Sciences的基因组测序仪、来自Illumina/Solexa的1G分析仪、来自Applied BioSystems的SOLiD系统以及来自Helicos Biosciences的Heliscope系统。Pacific BioSciences和VisiGen Biotechnologies也描述过边合成边测序平台。在一些实施方案中,使所测序的多个核酸分子结合至支撑物(例如固体支撑物)上。为了将核酸固定于支撑物上,可以在模板的3'和/或5'端添加捕捉序列/通用引发位点。可以通过使捕捉序列与共价附接至支撑物的互补序列杂交而使核酸结合至支撑物。捕捉序列(又称为通用捕捉序列)是与附接至支撑物的序列互补的核酸序列,该序列还可以充当通用引物。

[0163] 作为捕捉序列的替代,可以将偶合对(如抗体/抗原、受体/配体,或抗生物素-生物素对,如例如美国专利申请第2006/0252077号中所述)的一个成员连接至各片段以将其捕捉在涂有该偶合对的相应第二成员的表面上。

[0164] 在捕捉后,可以例如实施例和美国专利第7,283,337号中所描述,通过例如单分子检测/测序,包括模板依赖性边合成边测序对该序列进行分析。在边合成边测序时,使表面结合的分子在聚合酶存在下暴露于多个标记过得核苷酸三磷酸。模板序列由并入正在生长的链的3'端的标记过的核苷酸的顺序决定。这可以实时进行或者可以按分步重复模式进行。对于实时分析,可以将不同光学标记并入各核苷酸并且可以利用多种激光器刺激并入的核苷酸。

[0165] 测序还可以包括其它大规模平行测序或下一代测序 (NGS) 技术和平台。大规模平行测序技术和平台的其它实例有IlluminaHiSeq或MiSeq、Thermo PGM或Proton、Pac Bio RS II或Sequel、Qiagen公司的Gene Reader及Oxford Nanopore MinION。可以使用当前其它类似的大规模平行测序技术,以及这些技术的改进形式。

[0166] 任何细胞类型或组织都可以用于获得用于本文所描述的方法中的核酸样品。举例来说,DNA或RNA样品可以从肿瘤或体液,例如利用已知技术(例如静脉穿刺)获得的血液,或唾液获得。或者,可以对干燥样品(例如毛发或皮肤)进行核酸测试。此外,可以从肿瘤获得一份测序样品,并且可以从正常组织获得另一份测序样品,其中正常组织与肿瘤同属相同组织类型。可以从肿瘤获得一份测序样品,并且可以从正常组织获得另一份测序样品,其中正常组织与肿瘤属于不同组织类型。

[0167] 肿瘤可以包括以下一种或多种:肺癌、黑素瘤、乳癌、卵巢癌、前列腺癌、肾癌、胃癌、结肠癌、睾丸癌、头颈癌、胰腺癌、脑癌、B细胞淋巴瘤、急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病和T细胞淋巴细胞性白血病、非小细胞肺癌和小细胞肺癌。

[0168] 或者,可以使用蛋白质质谱法鉴别或验证结合至肿瘤细胞上的MHC蛋白质的突变肽的存在。肽可以用酸从肿瘤细胞或从自肿瘤免疫沉淀的HLA分子洗脱,并且接着使用质谱法鉴别。

[0169] IV. 新抗原

[0170] 新抗原可以包括核苷酸或多肽。举例来说,新抗原可以是编码多肽序列的RNA序列。因此,可用于疫苗中的新抗原包括核苷酸序列或多肽序列。

[0171] 本文公开了包含通过本文所公开的方法鉴别的肿瘤特异性突变的分离的肽、包含已知肿瘤特异性突变的肽,以及通过本文所公开的方法鉴别的突变多肽或其片段。新抗原肽可以在其编码序列背景下描述,其中新抗原包括编码相关多肽序列的核苷酸序列(例如DNA或RNA)。

[0172] 由新抗原核苷酸序列编码的一个或多个多肽可以包含以下至少一种:以低于1000nM的IC50值的与MHC的结合亲和力;对于长度是8-15个,即8、9、10、11、12、13、14或15个氨基酸的I类MHC肽,在该肽内或附近存在促进蛋白酶体裂解的序列基元;及存在促进TAP转运的序列基元。对于长度是6-30个,即6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个氨基酸的II类MHC肽,在该肽内或附近存在促进通过细胞外或溶酶体蛋白酶(组织蛋白酶)的切割或HLA-DM催化的HLA结合的序列基元。

[0173] 一种或多种新抗原可以被呈递在肿瘤表面上。

[0174] 一种或多种新抗原可以在患肿瘤的受试者中具有免疫原性,例如能够在该受试者体内引起T细胞应答或B细胞应答。

[0175] 在产生用于患肿瘤的受试者的疫苗的情况下,可以考虑排除在受试者体内诱导自体免疫应答的一种或多种新抗原。

[0176] 至少一个新抗原肽分子的尺寸可以包括但不限于约5个、约6个、约7个、约8个、约9个、约10个、约11个、约12个、约13个、约14个、约15个、约16个、约17个、约18个、约19个、约20个、约21个、约22个、约23个、约24个、约25个、约26个、约27个、约28个、约29个、约30个、约31个、约32个、约33个、约34个、约35个、约36个、约37个、约38个、约39个、约40个、约41个、约42个、约43个、约44个、约45个、约46个、约47个、约48个、约49个、约50个、约60个、约70个、约80

个、约90个、约100个、约110个、约120个或更多个氨基分子残基,以及由其中可衍生的任何范围。在特定实施方案中,新抗原肽分子等于或少于50个氨基酸。

[0177] 新抗原肽和多肽可以:对于I类MHC是15个或更少残基长度并且通常由介于约8个与约11个之间的残基,特别是9个或10个残基组成;对于II类MHC是6-30个残基(包括端点在内)。

[0178] 必要时,可以通过若干方式设计出更长的肽。在一种情况下,当预测出或已知肽在HLA等位基因上呈递的可能性时,较长的肽可以由以下任一种组成:(1)朝各相应基因产物的N末端和C末端延伸2-5个氨基酸的个别呈递的肽;(2)一些或全部呈递肽与各自的延伸序列的串接。在另一情况下,当测序披露在肿瘤中存在较长的(>10个残基)新表位序列(例如由产生新颖肽序列的移码、通读或包括内含子引起)时,较长的肽将由以下组成:(3)由新颖肿瘤特异性氨基酸组成的整个延伸段,由此绕过了对基于计算或体外测试来选择HLA呈递最强的较短肽的需求。在两种情况下,较长链的使用使患者细胞能够进行内源性加工并且可以产生更有效的抗原呈递和T细胞应答的诱导作用。

[0179] 新抗原肽和多肽可以被呈递于HLA蛋白质上。在一些方面,新抗原肽和多肽是以高于野生型肽的亲和力呈递于HLA蛋白质上。在一些方面,新抗原肽或多肽的IC₅₀值可以是至少低于5000nM、至少低于1000nM、至少低于500nM、至少低于250nM、至少低于200nM、至少低于150nM、至少低于100nM、至少低于50nM或更低。

[0180] 在一些方面,新抗原肽和多肽当施用给受试者时不会诱导自体免疫应答和/或激发免疫耐受性。

[0181] 还提供了包含至少两个或更多个新抗原肽的组合物。在一些实施方案中,该组合物含有至少两个不同的肽。至少两个不同的肽可以来源于同一多肽。不同的多肽意味着,该肽的长度、氨基酸序列或两者不同。这些肽来源于已知或被发现含有肿瘤特异性突变的任何多肽。可以作为新抗原肽的来源的适合多肽可以见于例如COSMIC数据库。COSMIC策划了有关人癌症中的体细胞突变的全面信息。肽含有肿瘤特异性突变。在一些方面,肿瘤特异性突变是特定癌症类型的驱动突变。

[0182] 具有所希望的活性或特性的新抗原肽和多肽可以被修饰成用于提供某些所希望的属性,例如改良的药理学特征,同时增加或至少保持未修饰肽的大体上所有生物活性以结合所希望的MHC分子并活化适当T细胞。举例来说,新抗原肽和多肽可以经历各种变化,如保守性或非保守性取代,其中此类变化可能在其使用中提供某些优势,如改良的MHC结合、稳定性及呈递。保守性取代意思指氨基酸残基被在生物上和/或化学上类似的另一氨基酸残基置换,例如一个疏水性残基被另一个置换,或一个极性残基被另一个置换。取代包括如Gly、Ala;Val、Ile、Leu、Met;Asp、Glu;Asn、Gln;Ser、Thr;Lys、Arg;及Phe、Tyr等的组合。单氨基酸取代的影响还可以使用D-氨基酸探测。此类修饰可以使用众所周知的肽合成程序进行,如例如Merrifield, *Science* 232:341-347(1986), Barany&Merrifield, *The Peptides*, Gross&Meienhofer编辑(N.Y., Academic Press), 第1-284页(1979);及Stewart和Young, *Solid Phase Peptide Synthesis*, (Rockford, Ill., Pierce), 第2版(1984)中所述。

[0183] 用各种氨基酸模拟物或非天然氨基酸修饰肽和多肽特别适用于增加该肽和多肽的体内稳定性。稳定性可以通过多种方式测定。举例来说,使用肽酶和各种生物介质如人血浆和血清测试稳定性。参见例如,Verhoef等人, *Eur. J. Drug Metab Pharmacokin.* 11:291-

302(1986)。肽的半衰期可以使用25%人血清(v/v)测定,按常规方式测定。方案大致如下。在使用前,通过离心使汇集的人血清(AB型,未热灭活)脱脂。接着,用RPMI组织培养基将该血清稀释至25%并用于测试肽稳定性。按预定时间间隔,取出少量反应溶液并添加至6%三氯乙酸水溶液或乙醇中。冷却混浊的反应样品(4℃),保持15分钟,然后离心以使沉淀的血清蛋白聚结。接着,通过反相HPLC,使用稳定性特异性色谱条件测定肽的存在。

[0184] 这些肽和多肽可以经过修饰以提供除改良的血清半衰期外的所希望的属性。举例来说,可以通过将这些肽连接至含有至少一个能够诱导T辅助细胞应答的表位的序列来增强其诱导CTL活性的能力。免疫原性肽/T辅助偶联物可以借助于间隔子分子连接。间隔子通常包含在生理条件下大体上不带电荷的相对较小的中性分子,如氨基酸或氨基酸模拟物。这些间隔子通常选自例如Ala、Gly或由非极性氨基酸或中性极性氨基酸组成的其它中性间隔子。应理解,任选存在的间隔子无需包含相同残基且因此可以是异低聚物或同低聚物。当存在时,间隔子通常是至少一个或二个残基,更通常是三个至六个残基。或者,可以在无间隔子情况下将肽连接至T辅助肽。

[0185] 新抗原肽可以直接地或通过间隔子在肽的氨基或羧基末端连接至T辅助肽。新抗原肽或T辅助肽的氨基末端可以被酰基化。示例性T辅助肽包括破伤风类毒素830-843、流感307-319、疟疾环孢子382-398和378-389。

[0186] 蛋白质或肽可以通过本领域技术人员已知的任何技术制备,包括通过标准分子生物学技术表达蛋白质、多肽或肽、从天然来源分离蛋白质或肽,或化学合成蛋白质或肽。先前已公开对应于各种基因的核苷酸和蛋白质、多肽和肽序列,并且可以见于本领域普通技术人员已知的计算机化数据库。一种此类数据库是位于美国国家卫生研究院(National Institutes of Health)网站的国家生物技术信息中心(National Center for Biotechnology Information)的Genbank和GenPept数据库。已知基因的编码区可以使用本文所公开或本领域普通技术人员已知的技术扩增和/或表达。或者,本领域技术人员已知蛋白质、多肽和肽的各种市售制剂。

[0187] 在另一方面,新抗原包括了编码新抗原肽或其部分的核酸(例如多核苷酸)。该多核苷酸可以是例如单链和/或双链DNA、cDNA、PNA、CAN、RNA(例如mRNA),或多核苷酸的天然或稳定化形式,如例如具有硫代磷酸酯主链的多核苷酸,或其组合,并且该多核苷酸可以含有或不含有内含子。又另一方面提供了一种能够表达多肽或其部分的表达载体。用于不同细胞类型的表达载体是本领域众所周知的并且可以在无过度实验情况下进行选择。一般来说,将DNA以适当取向和正确的表达阅读框插入表达载体,如质粒中。必要时,可以将DNA连接至能被所希望的宿主识别的适当转录和翻译调控性控制核苷酸序列,不过此类控制一般在表达载体中可用。接着,通过标准技术将载体插入宿主中。相关指导可见于例如 Sambrook等人(1989)Molecular Cloning,A Laboratory Manual,Cold Spring Harbor Laboratory,Cold Spring Harbor,N.Y.

[0188] IV. 疫苗组合物

[0189] 本文还公开了一种能够引起特异性免疫应答,例如肿瘤特异性免疫应答的免疫原性组合物,例如疫苗组合物。疫苗组合物通常包含多个例如使用本文所描述的方法选择的新抗原。疫苗组合物又可以称为疫苗。

[0190] 疫苗可以含有个数在1个与30个之间的肽,即2、3、4、5、6、7、8、9、10、11、12、13、14、

15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个不同的肽；6、7、8、9、10 11、12、13或14个不同肽；或12、13或14个不同的肽。肽可以包括翻译后修饰。疫苗可以含有个数在1个与100个之间或更多个核苷酸序列，即2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100或更多个不同的核苷酸序列；6、7、8、9、10 11、12、13或14个不同的核苷酸序列；或12、13或14个不同的核苷酸序列。疫苗可以含有个数在1个与30个之间的新抗原序列，即2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100或更多个不同的新抗原序列；6、7、8、9、10 11、12、13或14个不同的新抗原序列；或12、13或14个不同的新抗原序列。

[0191] 在一个实施方案中，不同肽和/或多肽或编码其的核苷酸序列的选择使得这些肽和/或多肽能够与不同MHC分子，如不同的I类MHC分子和/或不同的II类MHC分子缔合。在一些方面，一种疫苗组合物包含能够与最常出现的I类MHC分子和/或II类MHC分子缔合的肽和/或多肽的编码序列。因此，疫苗组合物可以包含能够与至少2个优选的、至少3个优选的或至少4个优选的I类MHC分子和/或II类MHC分子缔合的不同片段。

[0192] 该疫苗组合物能够引起特异性细胞毒性T细胞应答和/或特异性辅助T细胞应答。

[0193] 疫苗组合物还可以包含佐剂和/或载体。有用的佐剂和载体的实例提供于下文中。组合物可以与载体缔合，如例如蛋白质或抗原呈递细胞，如能够将肽呈递至T细胞的树突状细胞(DC)。

[0194] 佐剂是混合至疫苗组合物中增加或以其它方式改变针对新抗原的免疫应答的任何物质。载体可以是能够与新抗原缔合的支架结构，例如多肽或多糖。任选地，佐剂是共价或非共价缀合的。

[0195] 佐剂增加针对抗原的免疫应答的能力通常通过免疫介导的反应的显著或实质上增加，或疾病症状的减少来表现。举例来说，体液免疫的增加通常表现为针对抗原所产生的抗体的效价的显著增加，并且T细胞活性增加通常表现为细胞增殖，或细胞毒性，或细胞因子分泌的增加。佐剂也可以通过例如将主要体液或Th反应变成主要细胞或Th反应来改变免疫应答。

[0196] 适合的佐剂包括但不限于1018ISS、矾、铝盐、Amplivax、AS15、BCG、CP-870,893、CpG7909、CyaA、dSLIM、GM-CSF、IC30、IC31、咪喹莫特(Imiquimod)、ImuFactIMP321、IS Patch、ISS、ISCOMATRIX、JuvImmune、LipoVac、MF59、单磷酸脂质A、Montanide IMS1312、Montanide ISA 206、Montanide ISA 50V、Montanide ISA-51、OK-432、OM-174、OM-197-MP-EC、ONTAK、PepTel载体系统、PLG微粒、雷西莫特(resiquimod)、SRL172、病毒颗粒和其它类病毒颗粒、YF-17D、VEGF捕捉剂、R848、 β -葡聚糖、Pam3Cys、Aquila的来源于皂素的QS21刺激子(Aquila Biotech, Worcester, Mass., USA)、分枝杆菌提取物和合成细菌细胞壁模拟物，以及其它专用佐剂，如Ribi的Detox.Quil或Superfos。佐剂，如不完全弗氏佐剂或GM-CSF是

有用的。先前已描述若干专用于树突状细胞的免疫佐剂(例如MF59)和其制备方法(Dupuis M等人,Cell Immunol.1998;186(1):18-27;Allison A C;Dev Biol Stand.1998;92:3-11)。也可以使用细胞因子。若干细胞因子与以下直接相关:影响树突状细胞向淋巴组织(例如TNF- α)的迁移;加速树突状细胞成熟成为T淋巴细胞的有效抗原呈递细胞(例如GM-CSF、IL-1及IL-4)(美国专利第5,849,589号,特定地以引用的方式整体并入本文中)及充当免疫佐剂(例如IL-12)(Gabrilovich D I等人,JImmunother Emphasis Tumor Immunol.1996(6):414-418)。

[0197] 也已经报导过CpG免疫刺激性寡核苷酸能增强佐剂在疫苗环境中的作用。也可以使用其它TLR结合分子,如RNA结合性TLR 7、TLR8和/或TLR 9。

[0198] 有用佐剂的其它实例包括但不限于化学修饰的CpG(例如CpR、Idera)、聚(I:C)(例如聚i:CI2U)、非CpG细菌DNA或RNA以及免疫活性小分子和抗体,如环磷酰胺、舒尼替尼(sunitinib)、贝伐单抗(bevacizumab)、西乐葆(celebrex)、NCX-4016、西地那非(sildenafil)、他达那非(tadalafil)、伐地那非(vardenafil)、索拉非尼(sorafenib)、XL-999、CP-547632、帕佐盘尼(pazopanib)、ZD2171、AZD2171、伊匹单抗(ipilimumab)、曲美单抗(tremelimumab)及SC58175,这些可以起到治疗作用和/或充当佐剂。佐剂和添加剂的量和浓度可以由熟练技术人员容易地确定,无需过度实验。其它佐剂包括集落刺激因子,如粒细胞巨噬细胞集落刺激因子(GM-CSF,沙格司亭(sargramostim))。

[0199] 疫苗组合物可以包含超过一种不同的佐剂。此外,治疗组合物可以包含任何佐剂物质,包括上述任一种或其组合。另外,预期疫苗和佐剂可以一起施用或按任何适当的次序分开施用。

[0200] 载体(或赋形剂)可以独立于佐剂而存在。载体的功能可以是例如增加特定突变体的分子量以增加活性或免疫原性;赋予稳定性、增加生物活性或增加血清半衰期。此外,载体可以帮助将肽呈递至T细胞。载体可以是本领域技术人员已知的任何适合的载体,例如蛋白质或抗原呈递细胞。载体蛋白可以是但不限于匙孔血蓝蛋白、血清蛋白如转铁蛋白、牛血清白蛋白、人血清白蛋白、甲状腺球蛋白或卵白蛋白、免疫球蛋白或激素,如胰岛素或棕榈酸。对于人的免疫,载体一般是对人生理学上可接受的载体并且是安全的。不过,破伤风类毒素及/或白喉类毒素是适合的载体。或者,载体可以是葡聚糖,例如琼脂糖。

[0201] 细胞毒性T细胞(CTL)识别呈结合至MHC分子的肽形式的抗原,而非整个外来抗原本身。MHC分子本身位于抗原呈递细胞的细胞表面上。因此,如果存在肽抗原、MHC分子和APC的三聚体复合物,则可能活化CTL。相应地,如果该肽不仅用于活化CTL,而且如果另外添加具有相应MHC分子的APC,则其可以增强免疫应答。因此,在一些实施方案中,疫苗组合物另外含有至少一种抗原呈递细胞。

[0202] 新抗原也可以被包括在基于病毒载体的疫苗平台中,如牛痘、禽痘、自复制型 α 病毒、马拉巴病毒(marabavirus)、腺病毒(参见例如Tatsis等人,Adenoviruses,Molecular Therapy (2004) 10,616—629)或慢病毒,包括但不限于第二代、第三代和/或混合第二/第三代慢病毒和设计成靶向特定细胞类型或受体的任何一代重组慢病毒(参见例如,Hu等人,Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases,Immunol Rev. (2011) 239(1):45-61;Sakuma等人,Lentiviral vectors:basic to translational,Biochem J. (2012) 443(3):603-18;Cooper等人,Rescue of splicing-

mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter, *Nucl. Acids Res.* (2015) 43(1):682-690; Zufferey 等人, Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery, *J. Virol.* (1998) 72(12):9873-9880)。取决于以上提到的基于病毒载体的疫苗平台的包装能力,此方法可以递送编码一种或多种新抗原肽的一个或多个核苷酸序列。这些序列可以侧接非突变序列,可以由连接子分开,或者可以在前面具有一个或多个靶向亚细胞区室的序列(参见例如, Gros 等人, Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients, *Nat Med.* (2016) 22(4):433-8; Stronen 等人, Targeting of cancer neoantigens with donor-derived T cell receptor repertoires, *Science.* (2016) 352(6291):1337-41; Lu 等人, Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions, *Clin Cancer Res.* (2014) 20(13):3401-10)。在引入宿主中后,受感染的细胞表达新抗原,并由此引起针对肽的宿主免疫(例如 CTL)反应。可用于免疫方案的牛痘载体和方法描述于例如美国专利第 4,722,848 号中。另一载体是卡介苗(Bacille Calmette Guerin, BCG)。BCG 载体描述于 Stover 等人(Nature 351:456-460 (1991))中。根据本文的描述,本领域技术人员将显而易见可用于新抗原的治疗性施用或免疫的多种其它疫苗载体,例如,伤寒沙门氏菌(*Salmonella typhi*)载体。

[0203] IV.A. 有关疫苗设计和制造的其它考虑因素

[0204] IV.A.1. 确定涵盖所有肿瘤亚克隆的肽集合

[0205] 躯干肽(Truncal peptide),意思指由所有或大部分肿瘤亚克隆呈递的肽,将优先被包括在疫苗中。⁵³任选地,如果不存在预测会以较高概率呈递并具有免疫原性的躯干肽,或者如果预测能够以较高概率呈递并具有免疫原性的躯干肽的数量足够小以致可以在疫苗中包括其它非躯干肽,则可以通过估计肿瘤亚克隆的数量和属性并选择肽以使该疫苗所涵盖的肿瘤亚克隆的数量最大来对其它肽进行优先排序。⁵⁴

[0206] IV.A.2. 新抗原优先排序

[0207] 与疫苗技术可以支持的量相比,在应用所有以上新抗原过滤器后,仍有许多候选新抗原可包括在疫苗中。另外,可以保留有关新抗原分析的各个方面的不确定性,并且在候选疫苗新抗原的不同特性之间可能存在折中。因此,可以考虑用整合式多维模型代替在选择过程的每个步骤中的预定过滤器,该多维模型将候选新抗原放入具有至少以下轴的空间中并使用整合方法优化选择。

[0208] 1. 自体免疫或耐受的风险(生殖系的风险)(通常优选较低的自体免疫风险)。

[0209] 2. 测序伪影的概率(通常优选较低的伪影概率)。

[0210] 3. 免疫原性的概率(通常优选较高的免疫原性概率)。

[0211] 4. 呈递概率(通常优选较高的呈递概率)。

[0212] 5. 基因表达(通常优选较高表达)。

[0213] 6. HLA 基因的覆盖率(参与呈递新抗原集合的 HLA 分子数量增多可以降低肿瘤通过 HLA 分子下调或突变而逃避免疫攻击的概率)。

[0214] 7. HLA 类别的覆盖率(同时覆盖 HLA-I 和 HLA-II 可能会增加治疗反应的概率并降低肿瘤逃逸的概率)。

[0215] V. 治疗和制造方法

[0216] 还提供了一种通过向受试者施用一种或多种新抗原,如使用本文所公开的方法鉴别的多个新抗原来诱导受试者的肿瘤特异性免疫应答、针对肿瘤接种疫苗、治疗和或缓解受试者的癌症症状的方法。

[0217] 在一些方面,受试者被诊断患有癌症或有发生癌症的风险。受试者可以是需要肿瘤特异性免疫应答的人、狗、猫、马或任何动物。肿瘤可以是任何实体肿瘤,如乳房肿瘤、卵巢肿瘤、前列腺肿瘤、肺肿瘤、肾肿瘤、胃肿瘤、结肠肿瘤、睾丸肿瘤、头颈部肿瘤、胰腺肿瘤、脑肿瘤、黑素瘤及其它组织器官肿瘤;以及血液肿瘤,如淋巴瘤和白血病,包括急性骨髓性白血病、慢性骨髓性白血病、慢性淋巴细胞性白血病、T细胞淋巴细胞性白血病及B细胞淋巴瘤。

[0218] 新抗原的施用量应足以诱导CTL反应。

[0219] 新抗原可以单独施用或与其它治疗剂组合施用。治疗剂是例如化学治疗剂、放射或免疫疗法。针对特定癌症的任何适合的治疗性治疗都可以施用。

[0220] 此外,还可以向受试者施用抗免疫抑制/免疫刺激剂,如检查点抑制剂。举例来说,还可以向受试者施用抗CTLA抗体或抗PD-1或抗PD-L1。抗体阻断CTLA-4或PD-L1可以增强针对患者体内癌细胞的免疫应答。确切地说,经显示,当遵循疫苗接种方案时,有效阻断CTLA-4。

[0221] 可以确定包括在疫苗组合物中的每种新抗原的最佳量和最佳剂量方案。举例来说,可以制备供静脉内(i.v.)注射、皮下(s.c.)注射、皮内(i.d.)注射、腹腔内(i.p.)注射、肌肉内(i.m.)注射的新抗原或其变体。注射方法包括皮下(s.c.)、皮内(i.d.)、腹腔(i.p.)、肌肉(i.m.)及静脉内。DNA或RNA注射方法包括皮内、肌肉、皮下、腹腔及静脉内。本领域技术人员已知施用疫苗组合物的其它方法。

[0222] 疫苗可以被设计成使得组合物中存在的新抗原的选择、数量和/或量具有组织、癌症和/或患者特异性。举例来说,肽的确切选择可以由给定组织中亲本蛋白质的表达模式来指导。该选择可以取决于癌症的具体类型、疾病状态、先前的治疗方案、患者的免疫状态及当然要患者的HLA单倍型。此外,根据特定患者的个人需要,疫苗还可以含有个性化组分。实例包括根据特定患者体内新抗原的表达来改变新抗原的选择或遵循第一轮治疗方案调整后续治疗。

[0223] 对于打算用作癌症疫苗的组合物,在正常组织中大量表达的具有类似正常自身肽的新抗原应当避免或以少量存在于本文所描述的组合物中。另一方面,如果已知患者的肿瘤大量表达某一新抗原,则用于治疗此癌症的相应药物组合物可以大量存在和/或可以包括超过一种对于此特定新抗原或此新抗原的路径具有特异性的新抗原。

[0224] 可以将包含新抗原的组合物施用给患上癌症的个体。在治疗应用中,组合物是以足以引起针对肿瘤抗原的有效CTL反应及治愈或至少部分停滞症状和/或并发症的量施用给患者。适于实现此目的的量定义为“治疗有效剂量”。有效用于此用途的量将取决于例如组成、施用方式、所治疗的疾病的分期和严重程度、患者的体重和一般健康状态,以及处方医师的判断。应了解,组合物一般可以用于严重疾病状态,也就是说,危及生命或可能危及生命的状况,特别是当癌症已经转移的时候。在此类情况下,考虑到要使外来物质最少以及新抗原的相对无毒性质,治疗医师有可能并且会感觉需要施用大体上过量的这些组合物。

[0225] 对于治疗用途,施用可以在检测到或手术移除肿瘤时开始。这之后是增加剂量,直到至少症状大体上减轻并且之后持续一段时间。

[0226] 用于治疗性治疗的药物组合物(例如疫苗组合物)意图用于肠胃外、表面、鼻、口或局部施用。药物组合物可以通过肠胃外施用,例如静脉内、皮下、皮内或肌肉内施用。这些组合物可以施用到手术切除的部位处以诱导针对肿瘤的局部免疫应答。本文公开了供肠胃外施用的组合物,这些组合物包含新抗原溶液并且疫苗组合物被溶解或悬浮于可接受的载体,例如水性载体中。可以使用多种水性载体,例如水、缓冲水、0.9%生理盐水、0.3%甘氨酸、透明质酸等。这些组合物可以通过众所周知的常规灭菌技术灭菌,或者可以经历无菌过滤。由此得到的水溶液可以被包装起来按原样使用,或者被冻干;冻干的制剂在施用之前与无菌溶液组合。必要时,这些组合物可以含有药学上可接受的辅助物质以接近生理条件,如pH调节剂和缓冲剂、张力调节剂、润湿剂等,例如乙酸钠、乳酸钠、氯化钠、氯化钾、氯化钙、脱水山梨糖醇单月桂酸酯、三乙醇胺油酸酯等。

[0227] 新抗原还可以通过脂质体施用,使脂质体靶向特定细胞组织,如淋巴组织。脂质体还可用于增加半衰期。脂质体包括乳液、泡沫状物、胶束、不溶性单层、液晶、磷脂分散体、薄层状层等。在这些制剂中,待递送的新抗原是单独或与结合至例如淋巴细胞间普遍存在的受体的分子如结合至CD45抗原的单克隆抗体,或其它治疗或免疫原性组合物缀合作为脂质体的一部分并入。因此,填充有所希望的新抗原的脂质体可以被引导至淋巴细胞部位,接着脂质体递送所选治疗性/免疫原性组合物。脂质体可以由标准囊泡形成脂质形成,这些脂质一般包括中性和带负电的磷脂以及固醇如胆固醇。脂质的选择一般通过考虑例如脂质体尺寸、酸不稳定性和脂质体在血流中的稳定性来指导。如例如S z o k a等人, Ann.Rev.Biophys.Bioeng.9;467(1980);美国专利第4,235,871号、第4,501,728号、第4,501,728号、第4,837,028号及第5,019,369号中所述,有多种可用于制备脂质体的方法。

[0228] 为靶向免疫细胞,打算并入脂质体中的配体可以包括例如对所希望的免疫系统细胞的细胞表面决定子具有特异性的抗体或其片段。脂质体悬浮液可以经静脉内、局部、表面等途径施用,其剂量尤其根据施用方式、所递送的肽及所治疗疾病的分期等而变化。

[0229] 出于治疗或免疫接种目的,还可以向患者施用编码肽的核酸和任选地一种或多种本文所描述的肽。常常使用多种方法将核酸递送给患者。举例来说,核酸可以直接被递送,如“裸DNA”。这一方法描述于例如Wolff等人, Science247:1465-1468(1990)以及美国专利第5,580,859号和第5,589,466号。核酸还可以使用例如美国专利第5,204,253号中所描述的冲击递送法(ballistic delivery)施用。可以施用仅包含DNA的粒子。或者,可以使DNA附着至粒子,如金粒子。用于递送核酸序列的方法可以包括病毒载体、mRNA载体及DNA载体,利用或不利用电穿孔。

[0230] 核酸也可以与阳离子性化合物,如阳离子性脂质形成复合物来递送。脂质介导的基因递送方法描述于例如9618372WOAWO 96/18372;9324640WOAWO 93/24640;Mannino和Gould-Fogerite, BioTechniques 6(7):682-691(1988);美国专利第5,279,833号;Rose美国专利第5,279,833号;9106309WOAWO 91/06309;及Felgner等人, Proc.Natl.Acad.Sci.USA 84:7413-7414(1987)。

[0231] 新抗原也可以被包括在基于病毒载体的疫苗平台中,如牛痘、禽痘、自复制型 α 病毒、马拉巴病毒、腺病毒(参见例如Tatsis等人, Adenoviruses, Molecular Therapy (2004))

10,616—629) 或慢病毒,包括但不限于第二代、第三代和/或混合第二/第三代慢病毒和设计成靶向特定细胞类型或受体的任何一代重组慢病毒(参见例如,Hu等人,Immunization Delivered by Lentiviral Vectors for Cancer and Infectious Diseases,Immunol Rev. (2011) 239(1):45-61;Sakuma等人,Lentiviral vectors:basic to translational, Biochem J. (2012) 443(3):603-18;Cooper等人,Rescue of splicing-mediated intron loss maximizes expression in lentiviral vectors containing the human ubiquitin C promoter,Nucl.Acids Res. (2015) 43(1):682-690;Zufferey等人,Self-Inactivating Lentivirus Vector for Safe and Efficient In Vivo Gene Delivery, J.Virol. (1998) 72(12):9873-9880)。取决于以上提到的基于病毒载体的疫苗平台的包装能力,此方法可以递送编码一种或多种新抗原肽的一个或多个核苷酸序列。这些序列可以侧接非突变序列,可以由连接子分开,或者可以在前面具有一个或多个靶向亚细胞区室的序列(参见例如,Gros等人,Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients,Nat Med. (2016) 22(4):433-8;Stronen等人,Targeting of cancer neoantigens with donor-derived T cell receptor repertoires,Science. (2016) 352(6291):1337-41;Lu等人,Efficient identification of mutated cancer antigens recognized by T cells associated with durable tumor regressions,Clin Cancer Res. (2014) 20(13):3401-10)。在引入宿主中后,受感染的细胞表达新抗原,并由此引起针对肽的宿主免疫(例如CTL)反应。可用于免疫方案的牛痘载体和方法描述于例如美国专利第4,722,848号中。另一载体是卡介苗(BCG)。BCG载体描述于Stover等人(Nature 351:456-460(1991))中。根据本文的描述,本领域技术人员将显而易见可用于新抗原的治疗性施用或免疫的多种其它疫苗载体。

[0232] 施用核酸的方式使用了编码一个或多个表位的微型基因构建体。为了产生用于在人细胞中表达的编码所选CTL表位的DNA序列(微型基因),对这些表位的氨基酸序列进行逆翻译。使用人密码子用法表指导各氨基酸的密码子选择。将这些表位编码DNA序列直接邻接,产生连续多肽序列。为了优化表达和/或免疫原性,可以将另外的元件并入微型基因设计中。可以被逆翻译并且包括在微型基因序列中的氨基酸序列的实例包括:辅助T淋巴细胞、表位、前导(信号)序列及内质网滞留信号。此外,通过邻近CTL表位包括合成(例如聚丙烯氨酸)或天然存在的侧接序列可以改善CTL表位的MHC呈递。通过组装编码微型基因正链和负链的寡核苷酸,将微型基因序列转化成DNA。使用众所周知的技术,在适当条件下合成、磷酸化、纯化重叠寡核苷酸(30-100个碱基长)并使其退火。使用T4 DNA连接酶接合寡核苷酸的末端。接着,可以将这一编码CTL表位多肽的合成微型基因克隆至所希望的表达载体中。

[0233] 可以使用多种配制物制备注射用纯化质粒DNA。这些方法中最简单的方法是在无菌磷酸盐缓冲生理盐水(PBS)中使冻干的DNA复水。多种方法已有描述,并且新技术也可以使用。如上文所述,核酸宜用阳离子性脂质配制。此外,还可以使统称为保护性、相互作用性、非缩合性(PINC)的糖酯、促融脂质体、肽和化合物与纯化的质粒DNA形成复合物以影响各种变量,如稳定性、肌肉内分散或向特定器官或细胞类型的运输。

[0234] 还公开了一种制造肿瘤疫苗的方法,该方法包括执行本文所公开的方法的各个步骤;及产生包含多个新抗原或该多个新抗原的子集的肿瘤疫苗。

[0235] 本文所公开的新抗原可以使用本领域中已知的方法制造。举例来说,本文所公开

的产生新抗原或载体(例如包括至少一个编码一种或多种新抗原的序列的载体)的方法可以包括在适于表达该新抗原或载体的条件下培养宿主细胞,其中该宿主细胞包含至少一个编码该新抗原或载体的多核苷酸;及纯化该新抗原或载体。标准纯化方法包括色谱技术、电泳技术、免疫技术、沉淀、透析、过滤、浓缩及等电聚焦技术。

[0236] 宿主细胞可以包括中国仓鼠卵巢(CHO)细胞、NS0细胞、酵母或HEK293细胞。宿主细胞可以用一个或多个多核苷酸转化,该一个或多个多核苷酸包含至少一个编码本文所公开的新抗原或载体的核酸序列,任选地其中分离的多核苷酸另外包含可操作地连接到该至少一个编码新抗原或载体的核酸序列的启动子序列。在某些实施方案中,该分离的多核苷酸可以是cDNA。

[0237] VI. 新抗原鉴别

[0238] VI.A. 新抗原候选物的鉴别。

[0239] 有关以NGS分析肿瘤和正常外显子组和转录组的研究方法已有描述且被应用于新抗原鉴别领域中。^{6,14,15}以下实施例考虑了在临床环境中对于新抗原鉴别具有较高灵敏度和特异性的某些优化措施。这些优化措施可以分为两个领域,即与实验室方法有关的优化和与NGS数据分析有关的优化。

[0240] VI.A.1. 实验室方法优化

[0241] 此处提出的方法改进通过将所开发的有关可靠地评估靶癌症组中的癌症驱动基因的概念¹⁶扩展至新抗原鉴别所需的全外显子组和全转录组环境,解决了从肿瘤含量较低并且体积较小的临床试样中高准确性发现新抗原的难题。确切地说,这些改进包括:

[0242] 1. 靶向整个肿瘤外显子组的深度(>500×)独特平均覆盖率,以检测由于肿瘤含量低或处于亚克隆状态而以低突变等位基因频率存在的突变。

[0243] 2. 靶向整个肿瘤外显子组的均匀覆盖率,其中在<100×下覆盖<5%的碱基,由此通过例如以下方式使遗漏新抗原的可能性最低:

[0244] a. 采用基于DNA的捕捉探针和个别探针QC¹⁷

[0245] b. 包括针对覆盖较少的区域的额外诱饵

[0246] 3. 靶向整个正常外显子组的均匀覆盖率,其中在<20×下覆盖<5%的碱基,由此对于体细胞/生殖系状态可能有最少的新抗原未被分类(并因此不能用作TSNA)

[0247] 4. 为了使需要测序的总量减到最少,序列捕捉探针应被设计成仅针对基因编码区,因为非编码RNA不会产生新抗原。其它优化包括:

[0248] a. 针对HLA基因的补充探针,这些基因富含GC并且通过标准外显子组测序很难捕捉¹⁸

[0249] b. 排除由于如表达水平不足、蛋白酶体消化欠佳或不常见的序列特征等因素而被预测产生极少或不产生候选新抗原的基因。

[0250] 5. 肿瘤RNA将通常同样在高深度(>100M个读段)下测序,以便能够进行变体检测、基因和剪接变体(“同功型”)表达水平的定量,及融合物检测。来自FFPE样品的RNA将使用基于探针的富集方法¹⁹,使用与捕捉DNA中的外显子组相同或类似的探针进行提取。

[0251] VI.A.2. NGS数据分析优化

[0252] 分析方法的改进解决了常用研究突变调用方法灵敏度和特异性欠佳的问题,并且特别考虑到了在临床环境中与新抗原鉴别相关的定制。这些包括:

[0253] 1. 使用HG38参考人基因组或后续版本进行比对,因为相对于先前的基因组版本,该基因组含有多个MHC区域组装体,较佳地反映了群体多态性。

[0254] 2. 通过合并由不同程序得到的结果⁵,克服单个变体调用程序的局限性²⁰

[0255] a. 利用一套工具,检测肿瘤DNA、肿瘤RNA及正常DNA中的单核苷酸变体和插入缺失,该套工具包括:基于肿瘤与正常DNA的比较大的程序,如Strelka²¹和Mutect²²;和并入了肿瘤DNA、肿瘤RNA及正常DNA的程序,如UNCeQr,特别适用于低纯度样品²³。

[0256] b. 插入缺失将利用执行局部再组装的程序测定,如Strelka和ABRA²⁴。

[0257] c. 结构重排将使用专用工具测定,如Pindel²⁵或Breakseq²⁶。

[0258] 3. 为了检测并防止样品调换,将在选定的多态性位点数量下,比较来自同一患者的样品中的变体调用。

[0259] 4. 针对伪调用的广泛过滤将例如通过以下方式进行:

[0260] a. 移除在正常DNA中发现的变体,在低覆盖率下可能使用不严格的检测参数,并且在插入缺失情况下使用容许的接近标准。

[0261] b. 移除由低定位质量或低碱基质量引起的变体²⁷。

[0262] c. 移除来源于反复出现的测序伪影的变体,即使在相应的正常情况下未观察到²⁷。实例包括主要在一链上检测到的变体。

[0263] d. 移除不相关的对照物集合中检测到的变体²⁷

[0264] 5. 使用seq2HLA²⁸、ATHLATES²⁹或Optitype之一,从正常外显子组中准确地调用HLA,并且还将外显子组与RNA测序数据组合²⁸。其它可能的优化包括采用专用于HLA分型的分析,如长读段DNA测序³⁰,或调适用于接合RNA片段的方法以保持连续性³¹。

[0265] 6. 针对由肿瘤特异性剪接变体产生的neo-ORF的稳健检测将通过使用CLASS³²、Bayesemblem³³、StringTie³⁴或类似程序以其参考引导的模式,根据RNA-seq数据组装转录物来进行(即,使用已知的转录物结构而非尝试在每个实验中重新构建整个转录物)。尽管Cufflinks³⁵通常被用于此目的,但它常常会不合情理地产生大量剪接变体,其中有许多比全长基因要短得多,并且无法回收简单的阳性对照。编码序列及无义介导的衰变可能性将通过如SpliceR³⁶和MAMBA³⁷等工具,利用重新引入的突变序列测定。基因表达将利用如Cufflinks³⁵或Express (Roberts和Pachter, 2013) 等工具测定。野生型和突变体特异性表达计数和/或相对水平将利用开发用于这些目的的工具,如ASE³⁸或HTSeq³⁹测定。可能的过滤步骤包括:

[0266] a. 移除被认为表达不足的候选neo-ORF。

[0267] b. 移除被预测会触发无义介导的衰变(NMD)的候选neo-ORF。

[0268] 7. 仅在RNA中观察到的无法直接验证为肿瘤特异性抗原的候选新抗原(例如neoORF)将根据额外参数,例如通过考虑以下因素而归类为可能是肿瘤特异性的:

[0269] a. 存在仅支持肿瘤DNA的顺式作用移码或剪接位点突变

[0270] b. 在剪接因子中存在仅证实肿瘤DNA的反式作用突变。举例来说,在利用R625突变型SF3B1进行的三个独立公布的实验中,尽管一个实验检查到葡萄膜黑素瘤患者⁴⁰,第二个实验检查到葡萄膜黑素瘤细胞系⁴¹,而第三个实验检查到乳癌患者⁴²,但展现最大剪接差异的基因是一致的。

[0271] c. 对于新剪接同功型,在RNASeq数据中存在确证的“新”剪接-接合读段。

[0272] d.对于新重排,有确证在肿瘤DNA中存在而在正常DNA中不存在的近似外显子读段

[0273] e.基因表达概略中缺乏,如GTEx⁴³(即,使得不太可能为生殖系起源)

[0274] 8.通过直接比较组装的DNA肿瘤与正常读段(或来自这些读段的k-mer)来补充基于参考基因组比对的分析以避免基于比对和注释的错误和伪影。(例如对于在生殖系变体或重复序列插入缺失附近出现的体细胞变体)

[0275] 在具有聚腺苷酸化RNA的样品中,RNA-seq数据中病毒和微生物RNA的存在将使用RNA CoMPASS⁴⁴或类似方法评估,以鉴别可以预测患者响应的其它因素。

[0276] VI.B.HLA肽的分离和检测

[0277] HLA-肽分子的分离在溶胞和溶解组织样品之后,使用经典免疫沉淀(IP)方法进行⁵⁵⁻⁵⁸。使用澄清的溶解产物进行HLA特异性IP。

[0278] 免疫沉淀是使用偶合至珠粒的抗体进行,其中该抗体对HLA分子具有特异性。对于全I类HLA免疫沉淀,使用全I类CR抗体,对于II类HLA-DR,使用HLA-DR抗体。在过夜培育期间,将抗体共价连接至NHS-琼脂糖珠粒。在共价连接后,洗涤珠粒并等分试样用于IP。^{59,60}免疫沉淀也可以使用未共价结合至磁珠的抗体进行。通常,使用包被有蛋白A和/或蛋白G的琼脂糖或磁珠将抗体固定在色谱柱上来完成此操作。下面列出了一些可用于选择性富集MHC/肽复合物的抗体。

[0279]	抗体名称	特异性
	W6/32	I类HLA-A,B,C
	L243	II类-HLA-DR
	Tu36	II类-HLA-DR
	LN3	II类-HLA-DR
	Tu39	II类-HLA-DR,DP,DQ

[0280] 将澄清的组织溶解产物添加至抗体珠粒中进行免疫沉淀。免疫沉淀后,从溶解产物移除珠粒,并储存溶解产物用于另外的实验,包括另外的IP。洗涤IP珠粒以移除非特异性结合并使用标准技术,从珠粒洗脱下HLA/肽复合物。使用分子量旋转柱或C18分级分离,从肽移除蛋白质组分。通过SpeedVac蒸发使所得肽变干并且在一些情形中在-20C下储存以待MS分析。

[0281] 干燥的肽在适于反相色谱法的HPLC缓冲液中复水并装载至C-18微毛细管HPLC柱上以在Fusion Lumos质谱仪(Thermo)中进行梯度洗脱。在Orbitrap检测器中在高分辨率下收集肽质/荷比(m/z)的MS1谱,随后在所选离子经历HCD片段化后,在离子阱检测器中收集MS2低分辨率扫描谱。另外,可以使用CID或ETD片段化方法,或三种技术的任何组合获得MS2谱,以达到该肽的较高氨基酸覆盖率。还可以在Orbitrap检测器中用高分辨率质量精确率测量MS2谱。

[0282] 使用Comet^{61,62},针对蛋白质数据库搜索由各分析得到的MS2谱并使用Percolator⁶³⁻⁶⁵对肽鉴别进行评分。可以使用PEAKSstudio(Bioinformatics Solutions Inc.)进行另外的测序,并且可以使用其它搜索引擎或其它测序方法,包括光谱匹配和从头测序⁷⁵。

[0283] VI.B.1.支持全面HLA肽测序的MS检测限研究。

[0284] 使用肽YVYVADVAAK,利用装载至LC柱上的不同量的肽确定检测限。测试肽的量是

1pmol、100fmol、10fmol、1fmol及100amol。(表1)结果显示于图1F中。这些结果表明,最低检测限(LoD)是埃摩尔(attomol)范围(10^{-18}),动态范围跨五个数量级,并且信噪比看来足以在低飞摩尔(femtomol)范围(10^{-15})内进行测试。

	肽 m/z	装载于柱上	在 1e9 个细胞中的拷贝数/细胞
[0285]	566.830	1 pmol	600
	562.823	100 fmol	60
	559.816	10 fmol	6
	556.810	1 fmol	0.6
[0286]	553.802	100 amol	0.06

[0287] VII. 呈递模型

[0288] VII.A. 系统综述

[0289] 图2A是根据一个实施方案,用于鉴别患者体内肽呈递的可能性的环境100的概述。环境100提供背景以便引入呈递鉴别系统160,该系统本身包括呈递信息存储器165。

[0290] 呈递鉴别系统160是一个或多个在如以下关于图38所论述的计算系统中体现的计算机模型,其接收与MHC等位基因集合有关的肽序列并测定这些肽序列将被该相关MHC等位基因集合中的一个或多个MHC等位基因呈递的可能性。呈递鉴别系统160可以应用于I类和II类MHC等位基因两者。这在多种情形中都适用。呈递鉴别系统160的一个具体使用情形是,它能够接收与来自患者110的肿瘤细胞的MHC等位基因集合有关的候选新抗原的核苷酸序列,并测定这些候选新抗原将被该肿瘤的相关MHC等位基因中的一个或多个呈递和/或在患者110的免疫系统中诱导免疫原性反应的可能性。可以选出被系统160测定具有高可能性的候选新抗原用于包括在疫苗118中,此类抗肿瘤免疫应答可以由提供肿瘤细胞的患者110的免疫系统引发。另外,可以产生具有TCR的T细胞以用于T细胞疗法,所述T细胞对具有高呈递可能性的候选新抗原具有响应,从而还引起来自患者110的免疫系统的抗肿瘤免疫应答。

[0291] 呈递鉴别系统160通过一个或多个呈递模型测定呈递可能性。确切地说,呈递模型生成给定肽序列是否将由相关MHC等位基因集合呈递的可能性,并且这是基于存储在存储器165中的呈递信息生成的。举例来说,呈递模型可以生成肽序列“YVYVADVAAK”是否将被等位基因HLA-A*02:01、HLA-A*03:01、HLA-B*07:02、HLA-B*08:03、HLA-C*01:04的集合呈递在样品的细胞表面上的可能性。作为另一个实例,呈递模型还可以产生肽序列“YVYVADVAAK”是否将被具有HLA等位基因序列“AYANGPW”、“UIIKNFDL”、“WRTSAOGH”的HLA等位基因呈递的可能性。呈递信息165含有关于肽是否结合至不同类型的MHC等位基因以使得这些肽被MHC等位基因呈递的信息,该信息在模型中是根据肽序列中氨基酸的位置确定。呈递模型可以基于呈递信息165预测未被识别的肽序列的呈递是否会与相关MHC等位基因集合相关联。如前所述,呈递模型可以应用于I类和II类MHC等位基因两者。

[0292] 在本说明书中使用了术语“HLA覆盖率”。如在整个说明书中所使用的,“HLA覆盖率”可以应用于个体和/或个体群体。当应用于个体时,“HLA覆盖率”是指在个体的基因组内发现的存在呈递模型的HLA等位基因的比例。例如,对于具有HLA类型A*02:01,A*02:01,B*

07:02,B*07:02,C*07:02,C*07:02的纯合子个体,如果等位基因A*02:01和B*07:02存在呈递模型,但C*07:02不存在呈递模型,则该个体的HLA覆盖率为4/6。

[0293] 当应用于个体群体时,“HLA覆盖率”是指对于存在呈递模型的个体HLA覆盖率的每个可能水平而言的群体中个体的比例。在人类个体的情况下,每个人类基因组包含六个HLA等位基因。因此,个体HLA覆盖率的可能水平包括0/6,1/6,2/6,……,6/6。因此,例如,在个体的群体中,如果群体中一半的个体具有2/6的个体HLA覆盖率,并且群体中一半的个体具有6/6的个体HLA覆盖率,那么对于个体HLA覆盖率0/6而言,群体的HLA覆盖率为0%,对于个体HLA覆盖率1/6而言,群体的HLA覆盖率为0%,对于个体HLA覆盖率2/6而言,群体的HLA覆盖率为50%,对于个体HLA覆盖率3/6而言,群体的HLA覆盖率为0%,对于个体HLA覆盖率4/6而言,群体的HLA覆盖率为0%,对于个体HLA覆盖率5/6而言,群体的HLA覆盖率为0%,对于个体HLA覆盖率6/6而言,群体的HLA覆盖率为50%。

[0294] 如以下关于VIII部分更详细描述,训练呈递模型的目的是使群体的每个个体具有最高的可能的HLA覆盖率,并且因此对于群体的HLA覆盖率,使得就较高的个体HLA覆盖率而言,群体中个体的比例尽可能高。

[0295] VII.B. 呈递信息

[0296] 图2A示出了根据一个实施方案的获得呈递信息的方法。呈递信息165包括两个通用信息类别:等位基因相互作用信息和等位基因非相互作用信息。等位基因相互作用信息包括影响与MHC等位基因的类型相关的肽序列的呈递的信息。等位基因非相互作用信息包括影响与MHC等位基因的类型无关的肽序列的呈递的信息。

[0297] VII.B.1. 等位基因相互作用信息

[0298] 等位基因相互作用信息主要包括经过鉴别的肽序列,已知这些肽序列已经被来自人、小鼠等的一个或多个经过鉴别的MHC分子呈递。值得注意的是,这可能包括或可能不包括从肿瘤样品获得的数据。可以从表达单个MHC等位基因的细胞鉴别出所呈递的肽序列。在这一情形中,所呈递的肽序列一般是从单个等位基因细胞系收集,这些细胞系被工程改造成表达预定MHC等位基因并且随后暴露于合成蛋白质。在MHC等位基因上呈递的肽是通过如酸洗脱等技术分离并通过质谱法鉴别。图2B示出了这一情形的一个实施例,其中分离出在预定MHC等位基因HLA-DRB1*12:01上呈递的示例肽YEMFNDKSQRAPDDKMF并通过质谱法鉴别。由于在此情况下,肽是通过被工程改造成表达单一预定MHC蛋白质的细胞鉴别,故呈递的肽与其所结合的MHC蛋白质之间的直接关联是确定已知的。

[0299] 也可以从表达多个MHC等位基因的细胞收集所呈递的肽序列。通常,在人体中,一种细胞表达6种不同类型的MHC-I和至多12种不同类型的MHC-II分子。如此呈递的肽序列可以从被工程改造成表达多个预定MHC等位基因的多等位基因细胞系鉴别到。还可以从组织样品,如正常组织样品或肿瘤组织样品鉴别如此呈递的肽序列。特别就这一情形来说,MHC分子可以从正常或肿瘤组织免疫沉淀。在多个MHC等位基因上呈递的肽可类似地通过如酸洗脱等技术分离并通过质谱法鉴别。图2C示出了此种情形的一个实施例,其中将六个示例肽YEMFNDKSF、HROEIFSHDFJ、FJIEJFOESS、NEIOREIREI、JFKSIFEMMSJDSSUIFLKSJFIEIFJ及KNFLENFIESOFI呈递于所鉴别的I类MHC等位基因HLA-A*01:01、HLA-A*02:01、HLA-B*07:02、HLA-B*08:01及II类MHC等位基因HLA-DRB1*10:01、HLA-DRB1:11:01并且分离,并通过质谱法鉴别。相对于单等位基因细胞系,呈递的肽与其所结合的MHC蛋白质之间的直接关联可能

是未知的,因为结合肽是在鉴别之前与MHC分子分离。

[0300] 等位基因相互作用信息还可以包括质谱离子流,其取决于肽-MHC分子复合物的浓度和肽电离效率。电离效率以序列依赖性方式随肽而变化。一般来说,电离效率随肽而在约两个数量级内变化,而肽-MHC复合物的浓度在比其更大的范围内变化。

[0301] 等位基因相互作用信息还可以包括给定MHC等位基因与给定肽之间结合亲和力的测量或预测。(72,73,74)一个或多个亲和力模型可以生成此类预测。举例来说,再看回图1D中所示的实施例,呈递信息165可以包括肽YEMFNDKSF与等位基因I类HLA-A*01:01之间的1000nM的结合亲和力预测值。IC₅₀>1000nm的肽很少被MHC呈递,且较低的IC₅₀值使呈递概率增加。呈递信息165可以包括肽KNFLENFIESOFI和II类等位基因HLA-DRB1:11:01之间的结合亲和力预测。

[0302] 等位基因相互作用信息也可以包括该MHC复合物稳定性的测量或预测。一个或多个稳定性模型可以生成此类预测。较稳定的肽-MHC复合物(即,半衰期较长的复合物)比较可能在肿瘤细胞上及在遭遇疫苗抗原的抗原呈递细胞上以高拷贝数呈递。举例来说,再看回图2C中所示的实施例,呈递信息165可以包括I类分子HLA-A*01:01的半衰期是1小时的稳定性预测值。呈递信息165可以包括II类分子HLA-DRB1:11:01的半衰期的稳定性预测值。

[0303] 等位基因相互作用信息也可以包括测量或预测的肽-MHC复合物的形成反应速率。以较高速率形成的复合物比较可能以高浓度呈递于细胞表面上。

[0304] 等位基因相互作用信息还可以包括肽的序列和长度。I类MHC分子通常偏好呈递长度介于8与15个肽之间的肽。所呈递的肽中有60-80%的长度是9个。II类MHC分子通常更优先呈递介于6到30个肽之间的肽。

[0305] 等位基因相互作用信息还可以包括新抗原编码肽上激酶序列基元的存在,以及新抗原编码肽上特定翻译后修饰的不存在或存在。激酶基元的存在会影响翻译后修饰的概率,该翻译后修饰可能增强或干扰MHC结合。

[0306] 等位基因相互作用信息还可以包括翻译后修饰过程中所涉及的蛋白质,例如激酶的表达水平或活性水平(如由RNAseq、质谱法或其它方法所测量或预测)。

[0307] 等位基因相互作用信息还可以包括来自表达特定MHC等位基因的其它个体的细胞中具有相似序列的肽的呈递概率,这可通过质谱蛋白组学或其它手段评估。

[0308] 等位基因相互作用信息还可以包括所讨论的个体中特定MHC等位基因的表达水平(例如,如通过RNA-seq或质谱法测量)。相较于最强地结合至以低水平表达的MHC等位基因的肽,最强地结合至以高水平表达的MHC等位基因的肽比较可能被呈递。

[0309] 等位基因相互作用信息还可以包括不依赖于总体新抗原编码肽序列而在表达特定MHC等位基因的其它个体中由特定MHC等位基因呈递的概率。

[0310] 等位基因相互作用信息还可以包括不依赖于总体肽序列而在其它个体中由同一家族分子(例如HLA-A、HLA-B、HLA-C、HLA-DQ、HLA-DR、HLA-DP)中的MHC等位基因呈递的概率。举例来说,HLA-C分子的表达水平通常低于HLA-A或HLA-B分子,且由此可推断,由HLA-C呈递肽的概率低于由HLA-A或HLA-B呈递的概率。再举一个例子,HLA-DP的表达水平通常低于HLA-DR或HLA-DQ,且由此可推断,由HLA-DP呈递肽的概率低于由HLA-DR或HLA-DQ呈递的概率。

[0311] 等位基因相互作用信息还可以包括特定MHC等位基因的蛋白质序列。

[0312] 以下部分中所列的任何MHC等位基因非相互作用信息也可以按MHC等位基因相互作用信息的方式进行建模。

[0313] VII.B.2.等位基因非相互作用信息

[0314] 等位基因非相互作用信息可以包括在源蛋白质序列内侧接新抗原编码肽的C末端序列。对于MHC-I,C末端侧接序列可能影响肽的蛋白酶体加工。不过,C末端侧接序列是在肽转运至内质网并遇到细胞表面上的MHC等位基因之前,在蛋白酶体作用下自该肽裂解得到。因此,MHC分子接收不到有关C末端侧接序列的信息,且由此,C末端侧接序列的影响不会随MHC等位基因类型而变化。举例来说,再参看图2C中所示的实施例,呈递信息165可以包括从肽的源蛋白鉴别到的呈递肽FJIEJFOESS的C末端侧接序列FOEIFNDKSLDKFJI。

[0315] 等位基因非相互作用信息也可以包括mRNA定量测量。举例来说,可以获得与提供质谱训练数据相同的样品的mRNA定量数据。如稍后描述的,RNA表达水平被鉴别为肽呈递的强预测因子。在一个实施方案中,mRNA定量测量值是由软件工具RSEM鉴别得到。有关RSEM软件工具的详细实施方式可见于Bo Li及Colin N.Dewey.RSEM:accurate transcript quantification from RNA-Seq data with or without a reference genome.BMC Bioinformatics,12:323,2011年8月。在一个实施方案中,mRNA定量是以每一百万条定位读段数中每千碱基转录物的片段数(FPKM)为单位度量。

[0316] 等位基因非相互作用信息还可以包括在源蛋白质序列内侧接所述肽的N末端序列。

[0317] 等位基因非相互作用信息还可以包括肽序列的源基因。可以将源基因定义为肽序列的Ensembl蛋白家族。在另一些例子中,源基因可以被定义为肽序列的源DNA或源RNA。可以例如将源基因表示为编码蛋白质的一串核苷酸,或者基于已知编码特定蛋白质的已知DNA或RNA序列的命名集合将更直接地表示。在另一个例子中,等位基因非相互作用信息还可以包括从数据库如Ensembl或RefSeq中提取的肽序列的源转录本或同工型或潜在的源转录本或同工型的集合。

[0318] 等位基因非相互作用信息还可以包括肽序列来源的细胞的组织类型、细胞类型或肿瘤类型。

[0319] 等位基因非相互作用信息还可以包括在该肽中蛋白酶裂解基元的存在,任选地根据肿瘤细胞中相应蛋白酶的表达(如通过RNA-seq或质谱法测量)加权。含有蛋白酶裂解基元的肽不太可能被呈递,因为这些肽比较容易被蛋白酶降解,并因此在细胞内不太稳定。

[0320] 等位基因非相互作用信息还可以包括如在适当细胞类型中测量的源蛋白的转换率。转换率较快(即,半衰期较短)会增加呈递概率;不过,如果在不相似的细胞类型中测量,则此特征的预测能力较低。

[0321] 等位基因非相互作用信息还可以包括如通过RNA-seq或蛋白质组质谱法所测量,或如根据在DNA或RNA序列数据中检测到的生殖系或体细胞剪接突变的注释所预测的源蛋白的长度,任选地考虑在肿瘤细胞中表达水平最高的特定剪接变体(“同工型”)。

[0322] 等位基因非相互作用信息还可以包括肿瘤细胞中蛋白酶体、免疫蛋白酶体、胸腺蛋白酶体或其它蛋白酶的表达水平(可以通过RNA-seq、蛋白质组质谱法或免疫组织化学分析测量)。不同的蛋白酶体具有不同的裂解位点偏好。与表达水平成比例的各类型蛋白酶体的裂解偏好将被给予较大权重。

[0323] 等位基因非相互作用信息还可以包括肽的源基因的表达水平(例如通过RNA-seq或质谱法测量)。可能的优化措施包括调整表达水平测量值以说明肿瘤样品内基质细胞和肿瘤浸润淋巴细胞的存在。来自表达水平较高的基因的肽比较可能被呈递。来自表达水平不可检测的基因的肽可以不予考虑。

[0324] 等位基因非相互作用信息还可以包括如由无义介导的衰变模型,例如来自Rivas等人,Science 2015的模型所预测的新抗原编码肽的源mRNA将经历无义介导的衰变的概率。

[0325] 等位基因非相互作用信息还可以包括在各种细胞周期阶段期间肽的源基因的典型肿瘤特异性表达水平。以总体较低水平表达(如通过RNA-seq或质谱蛋白质组学所测量)但已知在特定细胞周期阶段期间高水平表达的基因所产生的呈递肽可能多于以极低水平稳定表达的基因。

[0326] 等位基因非相互作用信息还可以包括例如UniProt或PDB <http://www.rcsb.org/pdb/home/home.do>中提供的源蛋白特征的综合目录。这些特征尤其可以包括:蛋白质的二级和三级结构、亚细胞定位、基因本体(Gene ontology, GO)项。确切地说,这一信息可以含有在蛋白质水平上起作用的注释,例如5' UTR长度;及在特定残基水平上起作用的注释,例如在残基300与310之间的螺旋基元。这些特征还可以包括转角基元、折叠基元和无序残基。

[0327] 等位基因非相互作用信息还可以包括描述含有该肽的源蛋白的结构域的特性的特征,例如:二级或三级结构(例如 α 螺旋对比 β 折叠);选择性剪接。

[0328] 等位基因非相互作用信息还可以包括新抗原的肽序列与新抗原源基因的多个k聚体单元中的一个或多个k聚体单元之间的关联(如受试者的核苷酸测序数据中所存在的)。在呈递模型的训练期间,新抗原的肽序列和新抗原核苷酸测序数据的k聚体单元之间的这些关联被输入模型中,并且部分地被模型用来学习代表与训练肽序列相关联的k聚体单元的呈递热点的存在或不存在的模型参数。然后,在训练后的模型使用期间,将测试肽序列与测试肽序列的源基因的一个或多个k聚体单元之间的关联输入模型中,并且模型在训练过程中所学习的参数使得呈递模型能够对测试肽序列的呈递可能性做出更准确的预测。

[0329] 一般来说,表示k聚体单元的呈递热点存在或不存在的模型参数表示在控制了所有其他变量(如肽序列、RNA表达,通常在HLA结合肽等中发现的氨基酸)后,k聚体单元将产生所呈递的肽的残余倾向。表示k聚体单元的呈递热点的存在或不存在的参数可以是二进制系数(例如,0或1)或沿着标度的模拟系数(例如,在0和1之间,包括0和1)。在这两种情况的任一情况下,较大的系数(例如,接近1或1)表示k聚体单元将产生控制其他因素的所呈递肽的可能性较大,而较低的系数(例如,接近0或0)表示k聚体单元将产生所呈递肽的可能性较低。例如,具有低热点系数的k聚体单元可能是来自具有高RNA表达的基因的k聚体单元,其中氨基酸通常存在于HLA结合肽中,在这里源基因产生许多其他所呈递肽,但所呈递肽很少出现在k聚体单元中。由于其他肽存在来源可能已经由其他参数(例如,通常在HLA结合肽中发现的在k聚体单元或更大单元基础上的RNA表达)所解释,因此这些热点参数提供了新的、单独的信息,这些信息不会“双重计数”由其他参数捕获的信息。

[0330] 等位基因非相互作用信息还可以包括其它个体中来自相关肽的源蛋白的肽的呈递概率(在调整这些个体中源蛋白的表达水平及这些个体的不同HLA类型的影响之后)。

[0331] 等位基因非相互作用信息还可以包括由于技术偏差而无法通过质谱法检测到或过量表示该肽的概率。

[0332] 通过基因表达测定如RNASeq、微阵列、靶向组如Nanostring所测量的各种基因模块/路径的表达,或通过如RT-PCR等测定(无需含有该肽的源蛋白)所测量的基因模块的单基因/多基因代表提供了有关肿瘤细胞、基质或肿瘤浸润淋巴细胞(TIL)的状态的信息。

[0333] 等位基因非相互作用信息还可以包括肿瘤细胞中肽的源基因的拷贝数。举例来说,在肿瘤细胞中经历纯合子缺失的基因的肽可以指定为呈递概率是零。

[0334] 等位基因非相互作用信息还可以包括肽结合至TAP的概率或肽与TAP的结合亲和力测量值或预测值。比较可能结合至TAP的肽,或以较高亲和力结合TAP的肽比较可能被MHC-I呈递。

[0335] 等位基因非相互作用信息还可以包括肿瘤细胞中TAP的表达水平(可以通过RNA-seq、蛋白质组质谱法、免疫组织化学分析测量)。对于MHC-I,较高的TAP表达水平会增加所有肽的呈递概率。

[0336] 等位基因非相互作用信息还可以包括肿瘤突变的存在或不存在,包括但不限于:

[0337] i. 已知癌症驱动基因,如EGFR、KRAS、ALK、RET、ROS1、TP53、CDKN2A、CDKN2B、NTRK1、NTRK2、NTRK3中的驱动突变

[0338] ii. 编码抗原呈递机器中所涉及的蛋白质的基因(例如B2M、HLA-A、HLA-B、HLA-C、TAP-1、TAP-2、TAPBP、CALR、CNX、ERP57、HLA-DM、HLA-DMA、HLA-DMB、HLA-DO、HLA-DOA、HLA-DOBHLA-DP、HLA-DPA1、HLA-DPB1、HLA-DQ、HLA-DQA1、HLA-DQA2、HLA-DQB1、HLA-DQB2、HLA-DR、HLA-DRA、HLA-DRB1、HLA-DRB3、HLA-DRB4、HLA-DRB5或编码蛋白酶体或免疫蛋白酶体的组分的任何基因)中的突变。呈递依赖于肿瘤中经历功能丧失性突变的抗原呈递机器组分的肽具有降低的呈递概率。

[0339] 存在或不存在功能性生殖系多态现象,包括但不限于:

[0340] i. 编码抗原呈递机器中所涉及的蛋白质的基因(例如B2M、HLA-A、HLA-B、HLA-C、TAP-1、TAP-2、TAPBP、CALR、CNX、ERP57、HLA-DM、HLA-DMA、HLA-DMB、HLA-DO、HLA-DOA、HLA-DOBHLA-DP、HLA-DPA1、HLA-DPB1、HLA-DQ、HLA-DQA1、HLA-DQA2、HLA-DQB1、HLA-DQB2、HLA-DR、HLA-DRA、HLA-DRB1、HLA-DRB3、HLA-DRB4、HLA-DRB5或编码蛋白酶体或免疫蛋白酶体的组分的任何基因)中的功能性生殖系多态现象

[0341] 等位基因非相互作用信息还可以包括肿瘤类型(例如NSCLC、黑素瘤)。

[0342] 等位基因非相互作用信息还可以包括HLA等位基因的已知功能,如由例如HLA等位基因的后缀所反映。举例来说,等位基因名称HLA-A*24:09N中的N后缀指示未表达并因此不可能呈递表位的无效等位基因;完整HLA等位基因后缀命名法描述于<https://www.ebi.ac.uk/ipd/imgt/hla/nomenclature/suffixes.html>。

[0343] 等位基因非相互作用信息还可以包括临床肿瘤亚型(例如鳞状肺癌对比非鳞状肺癌)。

[0344] 等位基因非相互作用信息也可以包括吸烟史。

[0345] 等位基因非相互作用信息还可以包括晒伤史、太阳曝晒史或暴露于其它诱变剂的历史。

[0346] 等位基因非相互作用信息还可以包括肽的源基因在相关肿瘤类型或临床亚型中

的典型表达,任选地利用驱动基因突变分层。通常在相关肿瘤类型中高水平表达的基因比较可能被呈递。

[0347] 等位基因非相互作用信息还可以包括所有肿瘤中,或同一类型肿瘤中,或来自具有至少一个共有MHC等位基因的个体的肿瘤中,或具有至少一个共有MHC等位基因的个体体内的同一类型肿瘤中的突变频率。

[0348] 就突变的肿瘤特异性肽而言,用于预测呈递概率的特征清单也可以包括突变注释(例如错义、通读、移码突变、融合等)或预测该突变是否会引起无义介导的衰变(NMD)。举例来说,来自因纯合子早期终止突变而在肿瘤细胞中不翻译的蛋白质区段的肽可以指定为呈递概率是零。NMD使mRNA翻译减少,由此降低呈递概率。

[0349] VII.C. 呈递鉴别系统

[0350] 图3是一个高级框图,示出了根据一个实施方案的呈递鉴别系统160的计算机逻辑组件。在本示例实施方案中,呈递鉴别系统160包括数据管理模块312、编码模块314、训练模块316及预测模块320。呈递鉴别系统160还包括训练数据存储器170和呈递模型存储器175。该模型管理系统160的一些实施方案具有与此处所描述不同的模块。类似地,这些模块的功能分布可能不同于此处描述的模块。

[0351] VII.C.1. 数据管理模块

[0352] 数据管理模块312根据呈递信息165生成数组训练数据170。每组训练数据含有多个数据实例,其中每个数据实例 i 含有一组自变量 z^i ,这些自变量包括至少一个呈递或未呈递的肽序列 p^i 、一个或多个与该肽序列 p^i 相关联的相关MHC等位基因 a^i ,和/或一个或多个与该肽序列 p^i 相关联的MHC等位基因序列 d^i ;和一个因变量 y^i ,该因变量表示呈递鉴别系统160有意预测自变量的新值的信息。

[0353] 在本说明书其余部分通篇提到的一个特定的实施方式中,因变量 y^i 是一种二元标记,其指示肽 p^i 是否被该一个或多个相关MHC等位基因 a^i 并且/或者被一个或多个与一个或多个MHC等位基因序列 d^i 相关联的MHC等位基因呈递。不过,应理解,在其它实施方式中,取决于自变量 z^i ,因变量 y^i 可以表示呈递鉴别系统160有意进行预测的任何其它类别的信息。举例来说,在另一实施方案中,因变量 y^i 还可以是指示所鉴别的数据实例的质谱离子电流的数值。

[0354] 数据实例 i 的肽序列 p^i 是具有 k_i 个氨基酸的序列,其中 k_i 可以在随数据实例 i 而在一定范围内变化。举例来说,该范围对于I类MHC可以是8-15,或对于II类MHC是6-30。在系统160的一个具体实施方案中,一个训练数据集的所有肽序列 p^i 可以具有相同长度,例如9。肽序列中氨基酸的数量可以取决于MHC等位基因的类型(例如人体中的MHC等位基因等)而变化。数据实例 i 的MHC等位基因 a^i 指示存在的与相应肽序列 p^i 相关的MHC等位基因。类似地,在一些实施方案中,数据实例 i 的MHC等位基因序列 d^i 指示哪些MHC等位基因序列与对应的肽序列 p^i 相关联地存在。

[0355] 数据管理模块312还可以包括另外的等位基因相互作用变量,如与训练数据170中所包含的肽序列 p^i 和相关MHC等位基因 a^i 有关的结合亲和力 b^i 和稳定性预测值 s^i 。举例来说,训练数据170可以含有肽 p^i 与以 a^i 指示的各相关MHC分子之间的结合亲和力预测值 b^i 。作为另一个实施例,训练数据170可以含有以 a^i 指示的各MHC等位基因的稳定性预测值 s^i 。

[0356] 数据管理模块312还可以包括等位基因非相互作用变量 w^i ,如与肽序列 p^i 有关的C

[illegible]

[0363] 尽管以上编码序列数据的方法是参照具有氨基酸序列的序列描述,但该方法可以类似地扩展至其它类型的序列数据,如DNA或RNA序列数据等。

[0364] 编码模块314还将数据实例i的一个或多个MHC等位基因 a^i 编码成具有m个元素的行向量,其中每个元素 $h=1,2,\dots,m$ 对应于唯一鉴别的MHC等位基因。对应于所鉴别的数据实例i的MHC等位基因的元素的值是1。另外,其余元素的值是0。举个例子, $m=4$ 种唯一鉴别的MHC等位基因类型{HLA-A*01:01,HLA-C*01:08,HLA-B*07:02,HLA-DRB1*10:01}当中对应于多等位基因细胞系的数据实例i的等位基因HLA-B*07:02和HLA-DRB1*10:01可以由具有4个元素的行向量表示: $a^i=[0\ 0\ 1\ 1]$,其中 $a_3^i=1$ 和 $a_4^i=1$ 。尽管本文中用4种鉴别的MHC等位基因类型描述该实施例,但MHC等位基因类型的数量实际上可以是数百或数千。如先前所述,每个数据实例i通常含有最多6种不同的与肽序列 p_i 相关的MHC等位基因类型。

[0365] 编码模块314还将每个数据实例i的标记 y_i 编码为具有来自集合{0,1}的值的二元变量,其中值1指示肽 x^i 由相关的MHC等位基因 a^i 中的一个呈递,而值0指示肽 x^i 不被任何相关的MHC等位基因 a^i 呈递。当因变量 y_i 表示质谱离子电流时,编码模块314可以另外使用各种函数,如对 $[0, \infty)$ 之间的离子电流具有 $(-\infty, \infty)$ 范围的对数函数等缩放这些值。

[0366] 编码模块314可以将有关肽 p_i 和相关MHC等位基因 h 的一对等位基因相互作用变量 x_h^i 表示为行向量,其中等位基因相互作用变量的数字表示相继地串接。举例来说,编码模块314可以将 x_h^i 表示为等于 $[p^i]$ 、 $[p^i b_h^i]$ 、 $[p^i s_h^i]$ 或 $[p^i b_h^i s_h^i]$ 的行向量,其中 b_h^i 是肽 p_i 和相关MHC等位基因 h 的结合亲和力预测值,并且类似地 s_h^i 是关于稳定性。或者,等位基因相互作用变量的一个或多个组合可以个别地存储(例如以个别向量或矩阵形式)。

[0367] 在一个实例中,编码模块314通过将结合亲和力的测量值或预测值并入等位基因相互作用变量 x_{bi} 中表示结合亲和力信息。

[0368] 在一个实例中,编码模块314通过将结合稳定性的测量值或预测值并入等位基因相互作用变量 x_{ij} 中表示结合稳定性信息。

[369] 在一个实例中,编码模块314通过将结合缔合速率的测量值或预测值并入等位基因相互作用变量 x_b^i 中表示结合缔合速率信息。

[0370] 在一个实例中,对于由I类MHC分子呈递的肽,编码模块314将肽长度表示为向量 $T_k=[1(L_k=8) \ 1(L_k=9) \ 1(L_k=10) \ 1(L_k=11) \ 1(L_k=12) \ 1(L_k=13) \ 1(L_k=14) \ 1(L_k=15)]$,

其中 $\mathbf{1}$ 是指示函数,并且 L_k 表示肽 p_k 的长度。向量 T_k 可以被包括在等位基因相互作用变量 x_h^i

中。在另一个实例中,对于由II类MHC分子呈递的肽,编码模块314将肽长度表示为向量 $T_k = [1(L_k=6) \ 1(L_k=7) \ 1(L_k=8) \ 1(L_k=9) \ 1(L_k=10) \ 1(L_k=11) \ 1(L_k=12) \ 1(L_k=13) \ 1(L_k=14) \ 1(L_k=15) \ 1(L_k=16) \ 1(L_k=17) \ 1(L_k=18) \ 1(L_k=19) \ 1(L_k=20) \ 1(L_k=21) \ 1(L_k=22) \ 1(L_k=23) \ 1(L_k=24) \ 1(L_k=25) \ 1(L_k=26) \ 1(L_k=27) \ 1(L_k=28) \ 1(L_k=29) \ 1(L_k=30)]$, 其中 1 是指示函数,并且 L_k 表示肽 p_k 的长度。向量 T_k 可以被包括在等位基因相互作用变量 x_h^i 中。

[0371] 在一个实例中,编码模块314通过将基于RNA-seq的MHC等位基因表达水平并入等位基因相互作用变量 x_h^i 中表示MHC等位基因的RNA表达信息。

[0372] 类似地,编码模块314可以将等位基因非相互作用变量 w^i 表示为行向量,其中等位基因非相互作用变量的数字表示相继地串接。举例来说, w^i 可以是等于 $[c^i]$ 或 $[c^i \ m^i \ w^i]$ 的行向量,其中 w^i 是除肽 p^i 的C末端侧接序列和与该肽相关的mRNA定量测量值 m^i 外,还表示任何其它等位基因非相互作用变量的行向量。或者,等位基因非相互作用变量的一个或多个组合可以个别地存储(例如以个别向量或矩阵形式)。

[0373] 在一实例中,编码模块314通过将转换率或半衰期并入等位基因非相互作用变量 w^i 中表示肽序列的源蛋白的转换率。

[0374] 在一个实例中,编码模块314通过将蛋白质长度并入等位基因非相互作用变量 w^i 中表示源蛋白或同功型的长度。

[0375] 在一个实例中,编码模块314通过将包括 $\beta 1_i$ 、 $\beta 2_i$ 、 $\beta 5_i$ 亚单元在内的免疫蛋白酶体特异性蛋白酶体亚单元的平均表达水平并入等位基因非相互作用变量 w^i 中表示免疫蛋白酶体的活化情况。

[0376] 在一个实例中,编码模块314通过将源蛋白的丰度并入等位基因非相互作用变量 w^i 中表示肽的源蛋白或者肽的基因或转录物的RNA-seq丰度(通过如RSEM等技术以FPKM、TPM为单位定量)。

[0377] 在一个实例中,编码模块314通过将利用Rivas等人,Science,2015中的模型估计的肽的源转录物会经历无义介导的衰变(NMD)的概率并入等位基因非相互作用变量 w^i 中表示此概率。

[0378] 在一个实例中,编码模块314例如通过使用例如路径中每个基因的RSEM,以TPM为单位定量该路径中基因的表达水平,接着计算该路径中所有基因的概括统计量,例如平均值,以此表示经RNA-seq评估的基因模块或路径的活化状态。该平均值可以并入等位基因非相互作用变量 w^i 中。

[0379] 在一个实例中,编码模块314通过将拷贝数并入等位基因非相互作用变量 w^i 中表示源基因的拷贝数。

[0380] 在一个实例中,编码模块314通过将测量的或预测的TAP结合亲和力例如以纳摩尔浓度为单位)包括在等位基因非相互作用变量 w^i 中表示TAP结合亲和力。

[0381] 在一个实例中,编码模块314通过将利用RNA-seq测量(并利用例如RSEM,以TPM为单位定量)的TAP表达水平包括在等位基因非相互作用变量 w^i 中表示TAP表达水平。

[0382] 在一个实例中,编码模块314在等位基因非相互作用变量 w^i 中将肿瘤突变表示为指示变量的向量(即,如果肽 p^k 来自具有KRAS G12D突变的样品,则 $d^k=1$,否则是0)。

[0383] 在一个实例中,编码模块314将抗原呈递基因中的生殖系多态性表示为指示变量的向量(即,如果肽 p^k 来自在TAP中具有物种生殖系多态性的样品,则 $d^k=1$)。这些指示变量都可以被包括在等位基因非相互作用变量 w^i 中。

[0384] 在一个实例中,编码模块314根据肿瘤类型(例如NSCLC、黑素瘤、结肠直肠癌等)的字母表将肿瘤类型表示为长度一独热编码的向量。这些独热编码的变量都可以被包括在等位基因非相互作用变量 w^i 中。

[0385] 在一个实例中,编码模块314通过用不同后缀处理有4个数字的HLA等位基因来表示MHC等位基因后缀。举例来说,出于该模型的目的,HLA-A*24:09N被认为是与HLA-A*24:09不同的等位基因。或者,由于以N后缀结尾的HLA等位基因不表达,故可以将以N为后缀的MHC等位基因对所有肽的呈递概率设置成零。

[0386] 在一个实例中,编码模块314根据肿瘤亚型(例如肺腺癌、肺鳞状细胞癌等)的字母表将肿瘤亚型表示为长度一独热编码的向量。这些独热编码的变量都可以被包括在等位基因非相互作用变量 w^i 中。

[0387] 在一个实例中,编码模块314将吸烟史表示为二元指示变量(如果患者有吸烟史,则 $d^k=1$,否则是0),该变量可以包括在等位基因非相互作用变量 w^i 中。或者,可以根据吸烟严重程度的字母表,将吸烟史编码为长度一独热编码的变量。举例来说,吸烟状态可以在1-5级量表上评级,其中1指示非吸烟者,并且5指示当前多量吸烟者。由于吸烟史主要与肺部肿瘤相关,故当训练有关多种肿瘤类型的模型时,此变量也可以在患者有吸烟史时定义为等于1并且肿瘤类型是肺部肿瘤,否则是零。

[0388] 在一个实例中,编码模块314将晒伤史表示为二元指示变量(如果患者有重度晒伤史,则 $d^k=1$,否则是0),该变量可以包括在等位基因非相互作用变量 w^i 中。由于重度晒伤主要与黑素瘤相关,故当训练有关多种肿瘤类型的模型时,此变量也可以在患者有重度晒伤史时定义为等于1并且肿瘤类型是黑素瘤,否则是零。

[0389] 在一个实例中,编码模块314通过使用参考数据库如TCGA将有关人基因组中各基因或转录物的特定基因或转录物的表达水平分布表示为表达水平分布的概括统计量(例如平均值、中值)。确切地说,对于肿瘤类型是黑素瘤的样品中的肽 p^k ,不仅可以由肽 p^k 的源基因或转录物的基因或转录物表达水平测量值包括在等位基因非相互作用变量 w^i 中,而且还包括通过TCGA测量的黑素瘤中肽 p^k 的源基因或转录物的平均和/或中值基因或转录物表达水平。

[0390] 在一个实例中,编码模块314根据突变类型(例如错义突变、移码突变、NMD诱导的突变等)的字母表将突变类型表示为长度一独热编码的变量。这些独热编码的变量都可以被包括在等位基因非相互作用变量 w^i 中。

[0391] 在一个实例中,编码模块314在等位基因非相互作用变量 w^i 中将蛋白质的蛋白质水平特征表示为源蛋白的注释值(例如5' UTR长度)。在另一个实例中,编码模块314通过在等位基因非相互作用变量 w^i 中包括指示变量来表示 p^i 的残基水平的源蛋白注释,即,如果肽 p^i 与螺旋基元重叠则等于1,否则是0,或者如果肽 p^i 完全包含在螺旋基元内则等于1。在另一个实例中,表示肽 p^i 中包含在螺旋基元注释内的残基的比例的特征可以包括在等位基因非相互作用变量 w^i 中。

[0392] 在一个实例中,编码模块314将人蛋白质组中蛋白质或同功型的类型表示为指示

向量 o^k ,该向量的长度等于人蛋白质组中蛋白质或同功型的数量,并且如果肽 p^k 来自蛋白质 i ,则相应元素 o_i^k 是1,否则是0。

[0393] 在一个实例中,编码模块314将肽 p^i 的源基因 $G=\text{gene}(p^i)$ 表示为具有 L 个可能类别的分类变量,其中 L 表示索引的源基因1、2、... L 的数目的上限。

[0394] 在一个实例中,编码模块314将肽 p^i 的组织类型、细胞类型、肿瘤类型或肿瘤组织学类型 $T=\text{tissue}(p^i)$ 表示为具有 M 个可能类别的分类变量,其中 M 表示索引类型1、2、... M 的数目的上限。组织的类型可以包括,例如,肺组织、心脏组织、肠组织、神经组织等。细胞的类型可以包括树突状细胞、巨噬细胞、CD4T细胞等。肿瘤类型可以包括肺腺癌、肺鳞状细胞癌、黑色素瘤、非霍奇金淋巴瘤等。

[0395] 编码模块314还可以将有关肽 p_i 和相关MHC等位基因 h 的变量 z^i 的总体集合表示为行向量,其中等位基因相互作用变量 x^i 和等位基因非相互作用变量 w^i 的数字表示相继地串接。举例来说,编码模块314可以将 z_h^i 表示为等于 $[x_h^i w^i]$ 或 $[w_i x_h^i]$ 的行向量。

[0396] VIII. 训练模块

[0397] 训练模块316构建一个或多个呈递模型,这些模型生成肽序列是否会被与这些肽序列相关的MHC等位基因呈递的可能性。具体来说,给定肽序列 p^k 及与肽序列 p^k 相关联的一组MHC等位基因 a^k 和/或MHC等位基因序列 d^k ,每个呈递模型生成估计值 u_k ,指示肽序列 p^k 会被与一个或多个相关MHC等位基因 a^k 呈递的可能性。

[0398] VIII.A. 综述

[0399] 训练模块316基于由存储在165中的呈递信息产生的存储于存储器170中的训练数据集来构建一个或多个呈递模型。一般来说,不管呈递模型的具体类型如何,所有呈递模型都捕捉训练数据170中自变量与因变量之间的相关性以使损失函数减到最小。确切地说,损失函数 $\ell(y_{i \in S}, u_{i \in S}; \theta)$ 表示训练数据170中一个或多个数据实例 S 的因变量 $y_{i \in S}$ 与由呈递模型生成的数据实例 S 的估计可能性 $u_{i \in S}$ 值之间的偏差。在本说明书其余部分通篇所提到的一个特定实施方式中,损失函数 $(y_{i \in S}, u_{i \in S}; \theta)$ 是由以下等式(1a)提供的负对数可能性函数:

$$[0400] \quad \ell(y_{i \in S}, u_{i \in S}; \theta) = \sum_{i \in S} (y_i \log u_i + (1 - y_i) \log(1 - u_i)). \quad (1a)$$

[0401] 不过,实际上,可以使用另一损失函数。举例来说,当对质谱离子电流进行预测时,损失函数是由以下等式1b提供的均方损失:

$$[0402] \quad \ell(y_{i \in S}, u_{i \in S}; \theta) = \sum_{i \in S} (\|y_i - u_i\|_2^2). \quad (1b)$$

[0403] 呈递模型可以是一种参数模型,其中一个或多个参数 θ 在数学上指明自变量与因变量之间的相关性。通常,使损失函数 $(y_{i \in S}, u_{i \in S}; \theta)$ 最小的参数型呈递模型的各种参数是通过基于梯度的数值优化算法,如批量梯度算法、随机梯度算法等来确定。或者,呈递模型可以是非参数模型,其中模型结构是由训练数据170决定并且并不严格基于固定参数集合。

[0404] VIII.B. 独立等位基因模型

[0405] 训练模块316可以在独立等位基因(per-allele)基础上构建呈递模型以预测肽的呈递可能性。在此情况下,训练模块316可以基于由表达单个MHC等位基因的细胞产生的训练数据170中的数据实例 S 训练呈递模型。

[0406] 在一种实施方式中,训练模块316通过下式使特定等位基因 h 对于肽 p^k 的估计呈递可能性 u_k 建模:

$$[0407] \quad u_k^h = \Pr(p^k \text{ 被呈递 ; MHC 等位基因 } h) = f(g_h(x_h^k; \theta_h)), \quad (2)$$

[0408] 其中 x_h^k 表示编码的有关肽 p^k 和相应MHC等位基因 h 的等位基因相互作用变量, $f(\cdot)$ 是任何函数,并且为便于说明,在本文通篇称为变换函数。此外, $g_h(\cdot)$ 是任何函数,为便于说明,在本文通篇称为依赖性函数(dependency function),并且基于所测定的MHC等位基因 h 的参数集合 θ_h 产生等位基因相互作用变量 x_h^k 的依赖性分数。有关各MHC等位基因 h 的参数集合 θ_h 的值可以通过使关于 θ_h 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因 h 的细胞所产生的训练数据170的子集 S 中的每个实例。

[0409] 依赖性函数 $g_h(x_h^k; \theta_h)$ 的输出值表示至少基于等位基因相互作用特征 x_h^k ,并且确切地说,基于肽 p^k 的肽序列中氨基酸的位置的针对MHC等位基因 h 的依赖性分数,其指示MHC等位基因 h 将呈递相应新抗原。举例来说,如果MHC等位基因 h 可能呈递肽 p^k ,则MHC等位基因 h 的依赖性分数可能具有较高值,而如果不可能呈递,则可能具有较低值。变换函数 $f(\cdot)$ 将输入,并且更确切地说,在此情形中将由 $g_h(x_h^k; \theta_h)$ 生成的依赖性分数变换成适当值以指示肽 p^k 将由MHC等位基因呈递的可能性。

[0410] 在本说明书其余部分通篇提到的一个特定实施方式中, $f(\cdot)$ 是对于适当域范围具有在 $[0, 1]$ 内的范围的函数。在一个实施例中, $f(\cdot)$ 是由下式提供的expit函数:

$$[0411] \quad f(z) = \frac{\exp(z)}{1 + \exp(z)}. \quad (3)$$

[0412] 作为另一个实施例,当域 z 的值等于或大于0时, $f(\cdot)$ 也可以是由下式提供的双曲正切函数:

$$[0413] \quad f(z) = \tanh(z) \quad (4)$$

[0414] 或者,当质谱离子电流的预测值超出范围 $[0, 1]$ 时, $f(\cdot)$ 可以是任何函数,如恒等函数、指数函数、对数函数等。

[0415] 因此,可以通过将有关MHC等位基因 h 的依赖性函数 $g_h(\cdot)$ 应用于肽序列 p^k 的编码形式以产生相应依赖性分数来产生肽序列 p^k 将由MHC等位基因 h 呈递的独立等位基因可能性。依赖性分数可以由变换函数 $f(\cdot)$ 变换以产生肽序列 p^k 将由MHC等位基因 h 呈递的独立等位基因可能性。

[0416] VIII.B.1有关等位基因相互作用变量的依赖性函数

[0417] 在整个说明书中提到的一个特定实施方案中,依赖性函数 $g_h(\cdot)$ 是由下式提供的仿射函数:

$$[0418] \quad g_h(x_h^i; \theta_h) = x_h^i \cdot \theta_h. \quad (5)$$

[0419] 该函数将 x_h^k 中的每个等位基因相互作用变量与所测定的相关MHC等位基因 h 的参数集合 θ_h 中的相应参数线性地组合。

[0420] 在本说明书通篇提到的另一个特定实施方式中,依赖性函数 $g_h(\cdot)$ 是由下式提供的网络函数:

$$[0421] \quad g_h(x_h^i; \theta_h) = NN_h(x_h^i; \theta_h). \quad (6)$$

[0422] 该网络函数由具有布置在一个层或多个层中的一系列节点的网络模型 $NN_h(\cdot)$ 表示。一个节点可以通过连接物而连接至其它节点,该连接物各自在参数集合 θ_h 中具有相关参数。在一个特定节点处的值可以表示为连接至该特定节点的节点值的总和,该值由与该特定节点相关联的激活函数所映射的相关参数加权。由于呈递模型可以并入具有不同氨基酸序列长度的非线性和工艺数据,与仿射函数相比,网络模型是有利的。具体而言,通过非线性建模,网络模型可以捕捉在肽序列不同位置处的氨基酸之间的相互作用,并且可以捕捉这些相互作用如何影响肽呈递。

[0423] 一般来说,网络模型 $NN_h(\cdot)$ 可以被构造成前馈网络,如人工神经网络(ANN)、卷积神经网络(CNN)、深度神经网络(DNN),和/或循环网络,如长短期记忆网络(LSTM)、双向循环网络、深度双向循环网络等。

[0424] 在本说明书其余部分通篇提到的一个实例中, $h=1,2,\dots,m$ 中的每个MHC等位基因与独立网络模型相关联,并且 $NN_h(\cdot)$ 表示来自与MHC等位基因 h 相关联的网络模型的输出。

[0425] 图5示出了与任意MHC等位基因 $h=3$ 相关联的示例网络模型 $NN_3(\cdot)$ 。如图5中所示,关于MHC等位基因 $h=3$ 的网络模型 $NN_3(\cdot)$ 包括在层 $l=1$ 处的三个输入节点、在层 $l=2$ 处的四个节点、在层 $l=3$ 处的两个节点及在层 $l=4$ 处的一个输出节点。网络模型 $NN_3(\cdot)$ 与一组十个参数 $\theta_3(1), \theta_3(2), \dots, \theta_3(10)$ 相关。网络模型 $NN_3(\cdot)$ 接收关于MHC等位基因 $h=3$ 的三个等位基因相互作用变量 $x_3^k(1), x_3^k(2)$ 及 $x_3^k(3)$ 的输入值(包括编码的多肽序列数据和所用任何其它训练数据的个别数据实例)并输出值 $NN_3(x_3^k)$ 。网络函数还可以包括一个或多个网络模型,每个网络模型采用不同的等位基因相互作用变量作为输入。

[0426] 在另一个实施例中,鉴别的MHC等位基因 $h=1,2,\dots,m$ 与单个网络模型 $NN_h(\cdot)$ 相关联,并且 $NN_h(\cdot)$ 表示与MHC等位基因 h 相关的单个网络模型的一个或多个输出。在此类实例中,参数集合 θ_h 可以对应于该单个网络模型的参数集合,并因此,参数集合 θ_h 可以是所有MHC等位基因共有的。

[0427] 图6示出了MHC等位基因 $h=1,2,\dots,m$ 共享的示例网络模型 $NN_h(\cdot)$ 。如图6中所示,网络模型 $NN_h(\cdot)$ 包括 m 个输出节点,各自对应于MHC等位基因。网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并输出 m 值,包括对应于MHC等位基因 $h=3$ 的值 $NN_3(x_3^k)$ 。

[0428] 在又一个实施例中,依赖性函数 $g_h(\cdot)$ 可以表示为:

$$[0429] \quad g_h(x_h^k; \theta_h) = g'_h(x_h^k; \theta'_h) + \theta_h^0$$

[0430] 其中 $g'_h(x_h^k; \theta'_h)$ 是具有参数集合 θ'_h 的仿射函数、网络函数等,其中有关MHC等位基因的等位基因相互作用变量的参数集合的偏差参数 θ_h^0 表示MHC等位基因 h 的基线呈递概率。

[0431] 在另一种实施方式中,偏差参数 θ_h^0 可以是根据MHC等位基因 h 的基因家族共享的。也就是说,MHC等位基因 h 的偏差参数 θ_h^0 可以等于 $\theta_{\text{基因}(h)}^0$,其中基因(h)是MHC等位基因 h 的基因家族。举例来说,I类MHC等位基因HLA-A*02:01、HLA-A*02:02及HLA-A*02:03可以指定给“HLA-A”基因家族,并且这些MHC等位基因各自的偏差参数 θ_h^0 可以是共享的。作为另一个实例,II类MHC等位基因HLA-DRB1:10:01、HLA-DRB1:11:01及HLA-DRB3:01:01可以指定给“HLA-DRB”基因家族,并且这些MHC等位基因各自的偏差参数 θ_h^0 可以是共享的。

[0432] 再回到等式(2),作为一个实例,在使用仿射依赖性函数 $g_h(\cdot)$ 鉴别的 $m=4$ 种不同

的MHC等位基因当中,肽 p^k 将被MHC等位基因 $h=3$ 呈递的可能性可以由下式得出:

$$[0433] \quad u_k^3 = f(x_3^k \cdot \theta_3),$$

[0434] 其中 x_3^k 是鉴别的MHC等位基因 $h=3$ 的等位基因相互作用变量,并且 θ_3 是通过损失函数最小化测定的MHC等位基因 $h=3$ 的参数集合。

[0435] 作为另一个实施例,在使用独立网络变换函数 $g_h(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=3$ 呈递的可能性可以由下式得到:

$$[0436] \quad u_k^3 = f(NN_3(x_3^k; \theta_3)),$$

[0437] 其中 x_3^k 是鉴别的MHC等位基因 $h=3$ 的等位基因相互作用变量,并且 θ_3 是测定的与MHC等位基因 $h=3$ 相关联的网络模型 $NN_3(\cdot)$ 的参数集合。

[0438] 图7示出了使用示例网络模型 $NN_3(\cdot)$ 生成与MHC等位基因 $h=3$ 相关联的肽 p^k 的呈递可能性。如图7中所示,网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$ 。该输出由函数 $f(\cdot)$ 映射以产生估计的呈递可能性 u_k 。

[0439] VIII.B.2.具有等位基因非相互作用变量的独立等位基因

[0440] 在一种实施方式中,训练模块316并入等位基因非相互作用变量并通过下式使肽 p^k 的估计呈递可能性 u_k 建模:

$$[0441] \quad u_k^h = \Pr(p^k \text{ 被呈递}) = f(g_w(w^k; \theta_w) + g_h(x_h^k; \theta_h)), \quad (7)$$

[0442] 其中 w^k 表示肽 p^k 的编码的等位基因非相互作用变量, $g_w(\cdot)$ 是基于测定的等位基因非相互作用变量的参数集合 θ_w 的等位基因非相互作用变量 w^k 的函数。确切地说,有关各MHC等位基因 h 的参数集合 θ_h 和有关等位基因非相互作用变量的参数集合 θ_w 的值可以通过使关于 θ_h 和 θ_w 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因的细胞所产生的训练数据170的子集 S 中的每个实例。

[0443] 依赖性函数 $g_w(w^k; \theta_w)$ 的输出表示基于等位基因非相互作用变量的影响的等位基因非相互作用变量的依赖性分数,其指示肽 p^k 是否会由一个或多个MHC等位基因呈递。举例来说,如果肽 p^k 与已知会积极地影响肽 p^k 的呈递的C末端侧接序列相关,则等位基因非相互作用变量的依赖性分数可能具有较高值,并且如果肽 p^k 与已知会不利地影响肽 p^k 的呈递的C末端侧接序列相关,则可能具有较低值。

[0444] 根据等式(7),可以通过将有关MHC等位基因 h 的函数 $g_h(\cdot)$ 应用于肽序列 p^k 的编码形式以产生等位基因相互作用变量的相应依赖性分数来产生肽序列 p^k 将由MHC等位基因 h 呈递的独立等位基因可能性。有关等位基因非相互作用变量的函数 $g_w(\cdot)$ 也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的依赖性分数。将两个分数合并,并通过变换函数 $f(\cdot)$ 变换该合并的分数以产生肽序列 p^k 将由MHC等位基因 h 呈递的独立等位基因可能性。

[0445] 或者,训练模块316可以通过将等位基因非相互作用变量 w^k 添加至等式(2)中的等位基因非相互作用变量 x_h^k 中,来将等位基因非相互作用变量 w^k 包括在预测值中。因此,呈递可能性可以由下式得到:

$$[0446] \quad u_k^h = \Pr(p^k \text{ 被呈递; 等位基因 } h) = f(g_h([x_h^k w^k]; \theta_h)). \quad (8)$$

[0447] VIII.B.3有关等位基因非相互作用变量的依赖性函数

[0448] 与有关等位基因相互作用变量的依赖性函数 $g_h(\cdot)$ 类似,有关等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可以是仿射函数或网络函数,其中独立网络模型与等位基因非相互作用变量 w^k 相关联。

[0449] 确切地说,依赖性函数 $g_w(\cdot)$ 是由下式提供的仿射函数:

$$[0450] \quad g_w(w^k; \theta_w) = w^k \cdot \theta_w。$$

[0451] 该函数将等位基因非相互作用变量 w^k 与参数集合 θ_w 中的相应参数线性地组合。

[0452] 依赖性函数 $g_w(\cdot)$ 还可以是由下式提供的网络函数:

$$[0453] \quad g_h(w^k; \theta_w) = NN_w(w^k; \theta_w)。$$

[0454] 该函数是由具有参数集合 θ_w 中的相关参数的网络模型 $NN_w(\cdot)$ 表示。网络函数可能还包括一个或多个网络模型,每个网络模型采用不同的等位基因非相互作用变量作为输入。

[0455] 在另一个实施例中,有关等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可以由下式提供:

$$[0456] \quad g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + h(m^k; \theta_w^m), \quad (9)$$

[0457] 其中 $g'_w(w^k; \theta'_w)$ 是仿射函数,具有等位基因非相互作用参数集合 θ'_w 的网络函数等, m^k 是肽 p^k 的mRNA定量测量值, $h(\cdot)$ 是变换该定量测量值的函数,并且 θ_w^m 是有关等位基因非相互作用变量的参数集合中的一个参数,该参数与mRNA定量测量值组合以生成有关mRNA定量测量值的依赖性分数。在本说明书其余部分通篇所提到的一个特定实施方案中, $h(\cdot)$ 是对数函数,不过实际上, $h(\cdot)$ 可以是多种不同函数中的任一种。

[0458] 在又一个实例中,有关等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可以由下式提供:

$$[0459] \quad g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + \theta_w^o \cdot o^k, \quad (10)$$

[0460] 其中 $g'_w(w^k; \theta'_w)$ 是仿射函数、具有等位基因非相互作用参数集合 θ'_w 的网络函数等, o^k 是VII.C.2部分中描述的表示人蛋白质组中有关肽 p^k 的蛋白质和同功型的指示向量,并且 θ_w^o 是有关等位基因非相互作用变量的参数集合中的参数集合,其与指示向量组合。在一种变化形式中,当 o^k 的维度和参数集合 θ_w^o 明显较高时,可以在测定参数值时将参数正则项,如 $\lambda \cdot \|\theta_w^o\|$ 添加至损失函数中,其中 $\|\cdot\|$ 表示L1范数、L2范数、组合等。超参数 λ 的最佳值可以通过适当方法测定。

[0461] 在又一个实例中,有关等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可以由下式提供:

$$[0462] \quad g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + \sum_{l=1}^L \mathbf{1}(\text{基因}(p^k) = l) \cdot \theta_w^l, \quad (11)$$

[0463] 其中 $g'_w(w^k; \theta'_w)$ 是仿射函数、具有等位基因非相互作用参数集合 θ'_w 的网络函数等, $\mathbf{1}(\text{基因}(p^k)=l)$ 是指示函数,如上文对于等位基因非相互作用变量所述,如果肽 p^k 来自源基因1,则其等于1,并且 θ_w^1 是指示源基因1的“抗原性”的参数。在一种变化形式中,当L显著较高并且因此参数 $\theta_w^{1=1,2,\dots,L}$ 数量也显著较高时,可以在测定参数值时将参数正则项,如 $\lambda \cdot \|\theta_w^o\|$ 添加至损失函数中,其中 $\|\cdot\|$ 表示L1范数、L2范数、组合等。可以通过适当的方

法来确定超参数 λ 的最优值。

[0464] 在又一个实例中,有关等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可以由下式提供:

$$[0465] \quad g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + \sum_{m=1}^M \sum_{l=1}^L \mathbb{1}(\text{基因}(p^k) = l, \text{组织}(p^k) = m) \cdot \theta_w^{lm}, \quad (12a)$$

[0466] 其中 $g'_w(w^k; \theta'_w)$ 是仿射函数、具有等位基因非相互作用参数集合 θ'_w 的网络函数等, $\mathbb{1}(\text{基因}(p^k)=l, \text{组织}(p^k)=m)$ 是指示函数,如上文对于等位基因非相互作用变量所述,如果肽 p^k 来自源基因1并且如果肽 p^k 来自组织类型 m ,则其等于1,并且 θ_w^{lm} 是指示源基因1和组织类型 m 的组的抗原性的参数。具体地,组织类型 m 的基因1的抗原性可以表示在控制RNA表达和肽序列背景之后,组织 m 的细胞呈递来自基因1的肽的残余倾向。

[0467] 在一种变化形式中,当 L 或 M 显著较高并且因此参数 $\theta_w^{lm=1,2,\dots,LM}$ 数量也显著较高时,可以在测定参数值时将参数正则项,如 $\lambda \cdot \|\theta_w^{lm}\|$ 添加至损失函数中,其中 $\|\cdot\|$ 表示L1范数、L2范数、组合等。可以通过适当的方法来确定超参数 λ 的最优值。在另一种变化形式中,可以在测定参数值时将参数正则项添加至损失函数中,使得相同源基因的参数不会在组织类型之间有显著差异。例如,惩罚项如:

$$[0468] \quad \lambda \cdot \sum_{l=1}^L \sqrt{\sum_{m=1}^M (\theta_w^{lm} - \overline{\theta_w^l})^2}$$

[0469] 可以惩罚损失函数中不同组织类型之间抗原性的标准偏差,其中 $\overline{\theta_w^l}$ 是源基因1的组织类型之间的平均抗原性。

[0470] 在又一个实例中,等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 可由下式给出:

$$[0471] \quad g_w(w^k; \theta_w) = g'_w(w^k; \theta'_w) + \sum_{l=1}^L \mathbb{1}(\text{基因}(p^k) = l) \cdot \theta_w^l + \sum_{m=1}^M \mathbb{1}(\text{loc}(p^k) = m) \cdot \theta_w^m \quad (12b)$$

[0472] 其中 $g'_w(w^k; \theta'_w)$ 是仿射函数、具有等位基因非相互作用参数 θ'_w 的集合的网络函数等; $\mathbb{1}(\text{基因}(p^k)=l)$ 是这样的指示函数,如果肽 p^k 来自如上参考等位基因非相互作用变量所描述的源基因1,则该指示函数等于1;并且 θ_w^l 是指示源基因1的“抗原性”的参数;并且 $\mathbb{1}(\text{loc}(p^k)=m)$ 是这样的指示函数,如果肽 p^k 来自蛋白质组位置 m ,则该指示函数等于1;并且 θ_w^m 是指示蛋白质组位置 m 是呈递“热点”的程度的参数。在一个实施方案中,蛋白质组位置可以包含来自相同蛋白质的 n 个相邻肽的单元,其中 n 是经由适当方法例如网格搜索交叉验证确定的模型的超参数。

[0473] 实际上,等式(9)、(10)、(11)、(12a)和(12b)中的任一个的附加项可以组合以产生

等位基因非相互作用变量的依赖性函数 $g_w(\cdot)$ 。例如,可以将等式(9)中表示mRNA定量测量的项 $h(\cdot)$ 和等式(11)、(12a)和(12b)中表示源基因抗原性的项与任何其它仿射或网络函数一起相加,以生成等位基因非相互作用变量的依赖性函数。

[0474] 再回到等式(7),作为一个实例,在使用仿射变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将被MHC等位基因 $h=3$ 呈递的可能性可以由下式得出:

$$[0475] \quad u_k^3 = f(w^k \cdot \theta_w + x_3^k \cdot \theta_3),$$

[0476] 其中 w^k 是所鉴别的肽 p^k 的等位基因非相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0477] 作为另一个实施例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=3$ 呈递的可能性可以由下式得到:

$$[0478] \quad u_k^3 = f(NN_w(w^k; \theta_w) + NN_3(x_3^k; \theta_3))$$

[0479] 其中 w^k 是所鉴别的肽 p^k 的等位基因相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0480] 图8示出了使用示例网络模型 $NN_3(\cdot)$ 和 $NN_w(\cdot)$ 生成与MHC等位基因 $h=3$ 相关联的肽 p^k 的呈递可能性。如图8中所示,网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$ 。网络模型 $NN_w(\cdot)$ 接收有关肽 p^k 的等位基因非相互作用变量 w^k 并生成输出 $NN_w(w^k)$ 。将输出合并,并由函数 $f(\cdot)$ 映射以产生估计的呈递可能性 u_k 。

[0481] VIII.C.多等位基因模型

[0482] 训练模块316还可以在存在两个或更多个MHC等位基因的多等位基因环境中构建呈递模型以预测肽的呈递可能性。在此情况下,训练模块316可以基于由表达单个MHC等位基因的细胞、表达多个MHC等位基因的细胞或其组合产生的训练数据170中的数据实例S训练呈递模型。

[0483] VIII.C.1.实施例1:独立等位基因模型的最大值

[0484] 在一种实施方式中,训练模块316使与一组多个MHC等位基因 H 相关联的肽 p^k 的估计呈递可能性 u_k 随基于表达单等位基因的细胞所测定的集合 H 中每个MHC等位基因 h 的呈递可能性 $u_k^{h \in H}$ 的变化建模,如上文结合等式(2)-(10)所描述。确切地说,呈递可能性 u_k 可以是 $u_k^{h \in H}$ 的任何函数。在一种实施方式中,如等式(11)、(12a)和(12b)中所示,该函数是最大值函数,并且呈递可能性 u_k 可以测定为集合 H 中每个MHC等位基因 h 的呈递可能性最大值。

$$[0485] \quad u_k = \Pr(p^k \text{ 被呈递; 等位基因 } H) = \max(u_k^{h \in H}).$$

[0486] VIII.C.2.实施例2.1:和的函数(Funciton-of-Sums)模型

[0487] 在一种实施方式中,训练模块316通过下式使肽 p^k 的估计呈递可能性 u_k 建模:

$$[0488] \quad u_k = \Pr(p^k \text{ 被呈递 }) = f\left(\sum_{h=1}^m a_h^k \cdot g_h(x_h^k; \theta_h)\right), \quad (13)$$

[0489] 其中元素 a_h^k 对于与肽序列 p^k 相关的多个MHC等位基因 H 是1,并且 x_h^k 表示编码的有关肽 p^k 和相应MHC等位基因的等位基因相互作用变量。有关各MHC等位基因 h 的参数集合 θ_h 的

值可以通过使关于 θ_h 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集 S 中的每个实例。依赖性函数 g_h 可以呈以上VIII.B.1部分中介绍的依赖性函数 g_h 中的任一种的形式。

[0490] 根据等式(13),可以通过将依赖性函数 $g_h(\cdot)$ 应用于有关MHC等位基因 H 中的每一个的肽序列 p^k 的编码形式以产生等位基因相互作用变量的相应分数来产生肽序列 p^k 将由一个或多个MHC等位基因 h 呈递的呈递可能性。将每个MHC等位基因 h 的分数合并,并通过变换函数 $f(\cdot)$ 变换以产生肽序列 p^k 将由MHC等位基因集合 H 呈递的呈递可能性。

[0491] 等式(13)的呈递模型与等式(2)的独立等位基因模型的不同之处在于,每个肽 p^k 的相关等位基因的数量可以大于1。换句话说,对于与肽序列 p^k 相关的多个MHC等位基因 H , a_h^k 中超过一个元素值可以是1。

[0492] 举个例子,在使用仿射变换函数 $g_h(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0493] \quad u_k = f(x_2^k \cdot \theta_2 + x_3^k \cdot \theta_3),$$

[0494] 其中 x_2^k 、 x_3^k 是鉴别的MHC等位基因 $h=2$ 、 $h=3$ 的等位基因相互作用变量,并且 θ_2 、 θ_3 是测定的MHC等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0495] 作为另一个实例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0496] \quad u_k = f(NN_2(x_2^k; \theta_2) + NN_3(x_3^k; \theta_3)),$$

[0497] 其中 $NN_2(\cdot)$ 、 $NN_3(\cdot)$ 是鉴别的MHC等位基因 $h=2$ 、 $h=3$ 的网络模型,并且 θ_2 、 θ_3 是测定的MHC等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0498] 图9示出了使用示例网络模型 $NN_2(\cdot)$ 和 $NN_3(\cdot)$ 生成与MHC等位基因 $h=2$ 、 $h=3$ 相关联的肽 p^k 的呈递可能性。如图9中所示,网络模型 $NN_2(\cdot)$ 接收有关MHC等位基因 $h=2$ 的等位基因相互作用变量 x_2^k 并生成输出 $NN_2(x_2^k)$,并且网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$ 。将输出合并,并由函数 $f(\cdot)$ 映射以产生估计的呈递可能性 u_k 。

[0499] VIII.C.3. 实施例2.2: 利用等位基因非相互作用变量的和的函数模型

[0500] 在一种实施方式中,训练模块316并入等位基因非相互作用变量并通过下式使肽 p^k 的估计呈递可能性 u_k 建模:

$$[0501] \quad u_k = \Pr(p^k \text{ 被呈递 }) = f\left(g_w(w^k; \theta_w) + \sum_{h=1}^m a_h^k \cdot g_h(x_h^k; \theta_h)\right), \quad (14)$$

[0502] 其中 w^k 表示编码的有关肽 p^k 的等位基因非相互作用变量。确切地说,有关各MHC等位基因 h 的参数集合 θ_h 和有关等位基因非相互作用变量的参数集合 θ_w 的值可以通过使关于 θ_h 和 θ_w 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集 S 中的每个实例。依赖性函数 g_w 可以呈以上VIII.B.3部分中介绍的依赖性函数 g_w 中的任一种的形式。

[0503] 因此,根据等式(14),可以通过将函数 $g_h(\cdot)$ 应用于有关MHC等位基因 H 中的每一个的肽序列 p^k 的编码形式以产生有关每个MHC等位基因 h 的等位基因相互作用变量的相应

依赖性分数来产生肽序列 p^k 将由一个或多个MHC等位基因 H 呈递的呈递可能性。有关等位基因非相互作用变量的函数 $g_w(\cdot)$ 也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的依赖性分数。将分数合并,并通过变换函数 $f(\cdot)$ 变换该合并的分数以产生肽序列 p^k 将由MHC等位基因 H 呈递的呈递可能性。

[0504] 在等式(14)的呈递模型中,每个肽 p^k 的相关等位基因的数量可以大于1。换句话说,对于与肽序列 p^k 相关的多个MHC等位基因 H , a_h^k 中超过一个元素的值可以是1。

[0505] 举个例子,在使用仿射变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0506] \quad u_k = f(w^k \cdot \theta_w + x_2^k \cdot \theta_2 + x_3^k \cdot \theta_3),$$

[0507] 其中 w^k 是所鉴别的肽 p^k 的等位基因非相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0508] 作为另一个实施例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0509] \quad u_k = f(NN_w(w^k; \theta_w) + NN_2(x_2^k; \theta_2) + NN_3(x_3^k; \theta_3))$$

[0510] 其中 w^k 是所鉴别的肽 p^k 的等位基因相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0511] 图10示出了使用示例网络模型 $NN_2(\cdot)$ 、 $NN_3(\cdot)$ 及 $NN_w(\cdot)$ 生成与MHC等位基因 $h=2$ 、 $h=3$ 相关联的肽 p^k 的呈递可能性。如图10中所示,网络模型 $NN_2(\cdot)$ 接收有关MHC等位基因 $h=2$ 的等位基因相互作用变量 x_2^k 并生成输出 $NN_2(x_2^k)$ 。网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$ 。网络模型 $NN_w(\cdot)$ 接收有关肽 p^k 的等位基因非相互作用变量 w^k 并生成输出 $NN_w(w^k)$ 。将输出合并,并由函数 $f(\cdot)$ 映射以产生估计的呈递可能性 u_k 。

[0512] 或者,训练模块316可以通过将等位基因非相互作用变量 w^k 添加至等式(15)的等位基因非相互作用变量 x_h^k 中,来将等位基因非相互作用变量 w^k 包括在预测值中。因此,呈递可能性可以由下式得到:

$$[0513] \quad u_k = \Pr(p^k \text{ 被呈递 }) = f\left(\sum_{h=1}^m a_h^k \cdot g_h([x_h^k w^k]; \theta_h)\right). \quad (15)$$

[0514] VIII.C.4. 实施例3.1:使用隐式独立等位基因可能性的模型

[0515] 在另一种实施方式中,训练模块316通过下式使肽 p^k 的估计呈递可能性 u_k 建模:

$$[0516] \quad u_k = \Pr(p^k \text{ 被呈递 }) = r\left(s\left(v = \left[a_1^k \cdot u_k^1(\theta) \dots a_m^k \cdot u_k^m(\theta)\right]\right)\right), \quad (16)$$

[0517] 其中元素 a_h^k 对于与肽序列 p^k 相关联的多个MHC等位基因 $h \in H$ 是1, u_k^h 是MHC等位基因 h 的隐式独立等位基因呈递可能性,向量 v 是其中元素 v_h 对应于 $a_h^k \cdot u_k^h$ 的向量, $s(\cdot)$ 是映射元素 v 的函数,并且 $r(\cdot)$ 是限幅函数(clipping function),其将输入值削减至给定范围内。如以下更详细地描述, $s(\cdot)$ 可以是求和函数或二阶函数,但应理解在其它实施方案中, $s(\cdot)$ 可以是任何函数,如最大值函数。有关隐式独立等位基因可能性的参数集合 θ 的值可以通过使关于 θ 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因的细胞和/

或表达多个MHC等位基因的细胞所产生的训练数据170的子集S中的每个实例。

[0518] 使等式(16)的呈递模型中的呈递可能性随各自对应于肽 p^k 将由个别MHC等位基因 h 呈递的可能性的隐式独立等位基因呈递可能性 $u'_k{}^h$ 的变化建模。隐式独立等位基因可能性与VIII.B部分的独立等位基因呈递可能性的不同之处在于,有关隐式独立等位基因可能性的参数可以从多等位基因环境习得,其中除单等位基因环境外,呈递肽与相应MHC等位基因之间的直接关联也是未知的。因此,在多等位基因环境中,呈递模型不仅可以估计肽 p^k 是否会由作为整体的一组MHC等位基因 H 呈递,而且还可以提供指示最可能呈递肽 p^k 的MHC等位基因 h 的个别可能性 $u'_k{}^h \in \mathbb{H}$ 。其优势在于,呈递模型可以在无有关表达单MHC等位基因的细胞的训练数据存在下产生隐式可能性。

[0519] 在本说明书其余部分通篇提到的一个特定实施方式中, $r(\cdot)$ 是具有范围 $[0,1]$ 的函数。举例来说, $r(\cdot)$ 可以是限幅函数:

$$[0520] \quad r(z) = \min(\max(z, 0), 1),$$

[0521] 其中选择 z 与1之间的最小值作为呈递可能性 u_k 。在另一种实施方式中,当域 z 的值等于或大于0时, $r(\cdot)$ 是由下式提供的双曲正切函数:

$$[0522] \quad r(z) = \tanh(z).$$

[0523] VIII.C.5. 实施例3.2:函数的和(Sum-of-Functions)模型

[0524] 在一个特定实施方式中, $s(\cdot)$ 是求和函数,并且呈递可能性是通过对隐式独立等位基因呈递可能性求和得到:

$$[0525] \quad u_k = \Pr(p^k \text{ 被呈递 }) = r\left(\sum_{h=1}^m a_h^k \cdot u'_k{}^h(\theta)\right). \quad (17)$$

[0526] 在一种实施方式中,MHC等位基因 h 的隐式独立等位基因呈递可能性是由下式得到:

$$[0527] \quad u'_k{}^h = f(g_h(x_h^k; \theta_h)), \quad (18)$$

[0528] 由此通过下式估计出呈递可能性:

$$[0529] \quad u_k = \Pr(p^k \text{ 被呈递 }) = r\left(\sum_{h=1}^m a_h^k \cdot f(g_h(x_h^k; \theta_h))\right). \quad (19)$$

[0530] 根据等式(19),可以通过将函数 $g_h(\cdot)$ 应用于有关MHC等位基因 H 中的每一个的肽序列 p^k 的编码形式以产生等位基因相互作用变量的相应依赖性分数来产生肽序列 p^k 将由一个或多个MHC等位基因 H 呈递的呈递可能性。每个依赖性分数都先通过函数 $f(\cdot)$ 变换以产生隐式独立等位基因呈递可能性 $u'_k{}^h$ 。将独立等位基因可能性 $u'_k{}^h$ 合并,并且可以将限幅函数应用于合并的可能性以将值削减至范围 $[0,1]$ 中以产生肽序列 p^k 将由MHC等位基因集合 H 呈递的呈递可能性。依赖性函数 g_h 可以呈以上VIII.B.1部分中介绍的依赖性函数 g_h 中的任一种的形式。

[0531] 举个例子,在使用仿射变换函数 $g_h(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0532] \quad u_k = r\left(f(x_2^k \cdot \theta_2) + f(x_3^k \cdot \theta_3)\right),$$

[0533] 其中 x_2^k 、 x_3^k 是鉴别的MHC等位基因 $h=2$ 、 $h=3$ 的等位基因相互作用变量,并且 θ_2 、 θ_3 是测定的MHC等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0534] 作为另一个实施例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0535] \quad u_k = r \left(f(NN_2(x_2^k; \theta_2)) + f(NN_3(x_3^k; \theta_3)) \right),$$

[0536] 其中 $NN_2(\cdot)$ 、 $NN_3(\cdot)$ 是鉴别的MHC等位基因 $h=2$ 、 $h=3$ 的网络模型,并且 θ_2 、 θ_3 是测定的MHC等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0537] 图11示出了使用示例网络模型 $NN_2(\cdot)$ 和 $NN_3(\cdot)$ 生成与MHC等位基因 $h=2$ 、 $h=3$ 相关联的肽 p^k 的呈递可能性。如图11中所示,网络模型 $NN_2(\cdot)$ 接收有关MHC等位基因 $h=2$ 的等位基因相互作用变量 x_2^k 并生成输出 $NN_2(x_2^k)$,并且网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$ 。每个输出由函数 $f(\cdot)$ 映射以产生估计的呈递可能性 u_k 。

[0538] 在另一种实施方式中,当预测质谱离子电流的对数时, $r(\cdot)$ 是对数函数并且 $f(\cdot)$ 是指数函数。

[0539] VIII.C.6. 实施例3.3: 利用等位基因非相互作用变量的函数的和模型

[0540] 在一种实施方式中,MHC等位基因 h 的隐式独立等位基因呈递可能性是由下式得到:

$$[0541] \quad u_k^h = f \left(g_h(x_h^k; \theta_h) + g_w(w^k; \theta_w) \right), \quad (20)$$

[0542] 由此通过下式产生呈递可能性:

$$[0543] \quad u_k = \Pr(p^k \text{ 被呈递 }) = r \left(\sum_{h=1}^m a_h^k \cdot f \left(g_w(w^k; \theta_w) + g_h(x_h^k; \theta_h) \right) \right), \quad (21)$$

[0544] 以并入等位基因非相互作用变量对肽呈递的影响。

[0545] 根据等式(21),可以通过将函数 $g_h(\cdot)$ 应用于有关MHC等位基因 H 中的每一个的肽序列 p^k 的编码形式以产生有关每个MHC等位基因 h 的等位基因相互作用变量的相应依赖性分数来产生肽序列 p^k 将由一个或多个MHC等位基因 H 呈递的呈递可能性。有关等位基因非相互作用变量的函数 $g_w(\cdot)$ 也应用于等位基因非相互作用变量的编码形式以产生等位基因非相互作用变量的依赖性分数。将等位基因非相互作用变量的分数与等位基因相互作用变量的各个依赖性分数合并。每个合并的分数都通过函数 $f(\cdot)$ 变换以产生隐式独立等位基因呈递可能性。将隐式可能性合并,并且可以将限幅函数应用于合并的输出以将值削减至范围 $[0,1]$ 中以产生肽序列 p^k 将由MHC等位基因集合 H 呈递的呈递可能性。依赖性函数 g_w 可以呈以上VIII.B.3部分中介绍的依赖性函数 g_w 中的任一种的形式。

[0546] 举个例子,在使用仿射变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0547] \quad u_k = r \left(f(w^k \cdot \theta_w + x_2^k \cdot \theta_2) + f(w^k \cdot \theta_w + x_3^k \cdot \theta_3) \right),$$

[0548] 其中 w^k 是所鉴别的肽 p^k 的等位基因非相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0549] 作为另一个实施例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的MHC等位基因当中,肽 p^k 将由MHC等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0550] \quad u_k = r \left(f(NN_w(w^k; \theta_w) + NN_2(x_2^k; \theta_2)) + f(NN_w(w^k; \theta_w) + NN_3(x_3^k; \theta_3)) \right)$$

[0551] 其中 w^k 是所鉴别的肽 p^k 的等位基因相互作用变量,并且 θ_w 是测定的等位基因非相互作用变量的参数的集合。

[0552] 图12示出了使用示例网络模型 $NN_2(\cdot)$ 、 $NN_3(\cdot)$ 及 $NN_w(\cdot)$ 生成与MHC等位基因 $h=2$ 、 $h=3$ 相关联的肽 p^k 的呈递可能性。如图12中所示,网络模型 $NN_2(\cdot)$ 接收有关MHC等位基因 $h=2$ 的等位基因相互作用变量 x_2^k 并生成输出 $NN_2(x_2^k)$ 。网络模型 $NN_w(\cdot)$ 接收有关肽 p^k 的等位基因非相互作用变量 w^k 并生成输出 $NN_w(w^k)$ 。将输出合并,并且通过函数 $f(\cdot)$ 映射。网络模型 $NN_3(\cdot)$ 接收有关MHC等位基因 $h=3$ 的等位基因相互作用变量 x_3^k 并生成输出 $NN_3(x_3^k)$,再次将该输出与同一网络模型 $NN_w(\cdot)$ 的输出 $NN_w(w^k)$ 合并,并且通过函数 $f(\cdot)$ 映射。将两个输出合并以产生估计的呈递可能性 u_k 。

[0553] 在另一种实施方式中,MHC等位基因 h 的隐式独立等位基因呈递可能性由下式得到:

$$[0554] \quad u_k'^h = f(g_h([x_h^k w^k]; \theta_h)). \quad (22)$$

[0555] 由此通过下式得到呈递可能性:

$$[0556] \quad u_k = \Pr(p^k \text{ 被呈递 }) = r \left(\sum_{h=1}^m a_h^k \cdot f(g_h([x_h^k w^k]; \theta_h)) \right).$$

[0557] VIII.C.7. 实施例4:二阶模型

[0558] 在一种实施方式中, $s(\cdot)$ 是二阶函数,并且肽 p^k 的估计呈递可能性 u_k 由下式得到:

$$[0559] \quad u_k = \Pr(p^k \text{ 被呈递 }) = \sum_{h=1}^m a_h^k \cdot u_k'^h(\theta) - \sum_{h=1}^m \sum_{j < h} a_h^k \cdot a_j^k \cdot u_k'^h(\theta) \cdot u_k'^j(\theta) \quad (23)$$

[0560] 其中元素 $u_k'^h$ 是MHC等位基因 h 的隐式独立等位基因可能性。有关隐式独立等位基因可能性的参数集合 θ 的值可以通过使关于 θ 的损失函数减到最小来测定,其中 i 是由表达单个MHC等位基因的细胞和/或表达多个MHC等位基因的细胞所产生的训练数据170的子集 S 中的每个实例。隐式独立等位基因呈递可能性可以呈以上描述的等式(18)、(20)及(22)中所示的任何形式。

[0561] 在一方面,等式(23)的模型可以暗示存在肽 p^k 将同时由两个MHC等位基因呈递的可能,其中两个HLA等位基因的呈递在统计学上是独立的。

[0562] 根据等式(23),肽序列 p^k 将由一个或多个MHC等位基因 H 呈递的呈递可能性可以通过组合隐式独立等位基因呈递可能性并自总和中减去每对MHC等位基因将同时呈递肽 p^k 的可能性以产生肽序列 p^k 将由MHC等位基因 H 呈递的呈递可能性来产生。

[0563] 举个例子,在使用仿射变换函数 $g_h(\cdot)$ 鉴别的 $m=4$ 种不同的HLA等位基因当中,肽 p^k 将由HLA等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0564] \quad u_k = f(x_2^k \cdot \theta_2) + f(x_3^k \cdot \theta_3) - f(x_2^k \cdot \theta_2) \cdot f(x_3^k \cdot \theta_3),$$

[0565] 其中 x_2^k 、 x_3^k 是鉴别的HLA等位基因 $h=2$ 、 $h=3$ 的等位基因相互作用变量,并且 θ_2 、 θ_3

是测定的HLA等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0566] 作为另一个实施例,在使用网络变换函数 $g_h(\cdot)$ 、 $g_w(\cdot)$ 鉴别的 $m=4$ 种不同的HLA等位基因当中,肽 p^k 将由HLA等位基因 $h=2$ 、 $h=3$ 呈递的可能性可以由下式得到:

$$[0567] \quad u_k = f(NN_2(x_2^k; \theta_2)) + f(NN_3(x_3^k; \theta_3)) - f(NN_2(x_2^k; \theta_2)) \cdot f(NN_3(x_3^k; \theta_3)),$$

[0568] 其中 $NN_2(\cdot)$ 、 $NN_3(\cdot)$ 是鉴别的HLA等位基因 $h=2$ 、 $h=3$ 的网络模型,并且 θ_2 、 θ_3 是测定的HLA等位基因 $h=2$ 、 $h=3$ 的参数的集合。

[0569] VIII.D. 泛等位基因模型

[0570] 与独立等位基因模型相比,泛等位基因模型是能够基于泛等位基因预测肽的呈递可能性的呈递模型。具体而言,与独立等位基因模型(其能够预测肽将被一个或多个先前已用于训练该独立等位基因模型的已知MHC等位基因呈递的概率)不同,泛等位基因模型是这样的呈递模型,其能够预测肽将被任何MHC等位基因(包括该模型先前在训练期间未遇到过的未知MHC等位基因)呈递的概率。

[0571] 简要地,泛等位基因模型由训练模块316训练。类似于独立等位基因模型的训练,训练模块316可以基于由表达单个MHC等位基因的细胞、表达多个MHC等位基因的细胞或其组合产生的训练数据170中的数据实例S来训练泛等位基因呈递模型。然而,训练模块316不是使用特定的MHC等位基因或特定的一组MHC等位基因 a_h^k 来训练泛等位基因呈递模型,而是使用训练数据170中所有可用的MHC等位基因肽序列 d_h 来训练泛等位基因呈递模型。具体而言,训练模块316基于训练数据170中可用的MHC等位基因的氨基酸位置来训练泛等位基因呈递模型。

[0572] 在泛等位基因模型已被训练之后,当将肽序列和已知或未知的MHC等位基因肽序列输入该模型中以确定已知或未知的MHC等位基因将呈递该肽的概率时,该模型能够通过使用在用类似MHC等位基因肽序列训练期间学习到的信息来准确地预测这种概率。例如,使用不包含任何A*02:07等位基因的出现的训练数据170训练的泛等位基因模型仍可以通过利用在用类似等位基因(例如,A*02基因家族中的等位基因)训练期间学习到的信息来准确地预测A*02:07等位基因对肽的呈递。这样,单个呈递泛等位基因模型可以预测肽在任何MHC等位基因上的呈递可能性。

[0573] VIII.D.2. 泛等位基因模型的优势

[0574] 泛等位基因呈递模型的主要优势是,泛等位基因呈递模型比独立等位基因呈递模型具有更大的通用性。如上所提到的,独立等位基因模型能够预测肽将被用于训练该独立等位基因模型的一个或多个已鉴别的MHC等位基因呈递的概率。换句话说,独立等位基因模型与一个或多个已知MHC等位基因的有限集合相关联。

[0575] 因此,给定一个含有一个或多个MHC等位基因的特定集合的样品,为了确定肽被该MHC等位基因的特定集合呈递的概率,选择使用利用该MHC等位基因的特定集合训练的独立等位基因模型。换句话说,当依靠独立等位基因模型预测肽将被MHC等位基因呈递的概率时,只能为训练数据170中已出现的MHC等位基因作出预测。因为存在大量的MHC等位基因(特别是对于同一基因家族内的微小变异),所以将需要非常大的数量的训练样品来训练要装备的独立等位基因呈递模型,以便为所有MHC等位基因作出肽呈递预测。

[0576] 相比之下,泛等位基因模型不限于对在其上进行训练的一个或多个MHC等位基因的特定集合进行预测。取而代之的是,在使用过程中,泛等位基因模型能够通过使用在用类

似MHC等位基因肽序列训练期间学习到的信息来准确地预测先前见过的和/或先前未见过的MHC等位基因将呈递给定肽的概率。结果是,泛等位基因模型不与一个或多个MHC等位基因的特定集合相关联,并且能够预测肽将被任何MHC等位基因呈递的概率。泛等位基因模型的这种多功能性意味着可以使用单个模型来预测任何肽将被任何MHC等位基因呈递的可能性。因此,使用泛等位基因模型减少了最大程度地增大如上文VII.A部分中所定义的个体HLA覆盖率和群体HLA覆盖率所需的训练数据的量。

[0577] VIII.D.3.泛等位基因模型的使用

[0578] 以下在VIII.D.4.-VIII.D.7.部分中的讨论涉及使用泛等位基因模型预测肽将被一个或多个MHC等位基因呈递的概率。为简单起见,该讨论是在训练模块316已经训练了该泛等位基因模型的假设下进行的。下面将在VIII.D.8.部分中详细讨论泛等位基因模型的训练。

[0579] 此外,以下在VIII.D.4.-VIII.D.6.部分中的讨论涉及到使用泛等位基因模型预测肽将被给定样品中单个MHC等位基因和/或被给定样品中多个MHC等位基因呈递的可能性。然而,如下面在VIII.D.7.部分中详细描述,在使用泛等位基因模型预测肽将被样品中单个MHC等位基因呈递的可能性与使用泛等位基因模型预测肽将被样品中多个MHC等位基因呈递的可能性之间存在细微差异。

[0580] 简要地,当使用泛等位基因模型预测肽将被单个MHC等位基因呈递的可能性时,如下所述地,一组输入被提供给泛等位基因模型,并且泛等位基因模型产生单个输出。

[0581] 另一方面,当使用泛等位基因模型预测肽将被多个MHC等位基因呈递的可能性时,针对多个MHC等位基因中的每个MHC等位基因迭代使用泛等位基因模型。具体而言,当使用泛等位基因模型预测肽将被多个MHC等位基因呈递的可能性时,与多个MHC等位基因中的第一MHC等位基因相关联的第一组输入被提供给泛等位基因模型,并且泛等位基因模型产生针对第一MHC等位基因的第一输出。然后,与多个MHC等位基因中的第二MHC等位基因相关联的第二组输入被提供给泛等位基因模型,并且泛等位基因模型产生针对第二MHC等位基因的第二输出。针对多个MHC等位基因中的每个MHC等位基因迭代地执行这个过程。最后,泛等位基因模型产生的针对多个MHC等位基因中的每个MHC等位基因的输出被组合,以产生多个MHC等位基因呈递给定肽的单一概率,如VIII.D.7.部分所述的。

[0582] VIII.D.4.泛等位基因模型概述

[0583] 在一个实施方式中,泛等位基因模型被用于估计等位基因 h 对肽 p^k 的呈递可能性 u_k 。在一些实施方案中,泛等位基因模型由下面的等式表示:

$$[0584] \quad u_k^h = \Pr(p^k \text{ 被呈递 }; \text{MHC 等位基因 } h) = f(g_h([p^k \ d_h]; \theta_h)), \quad (24)$$

[0585] 其中 p^k 表示肽序列, d_h 表示MHC等位基因 h 的肽序列, $f(\cdot)$ 是任何变换函数,而 $g_h(\cdot)$ 是任何依赖性函数。泛等位基因模型基于为所有MHC等位基因确定的一组共享参数 θ_h 产生肽序列 p^k 和MHC等位基因肽序列 d_h 的依赖性分数。该组共享参数 θ_h 的值是在泛等位基因模型训练期间学习到的,并将在下面的VIII.D.8.部分中详细讨论。

[0586] 依赖性函数 $g_h([p^k \ d_h]; \theta_h)$ 的输出表示MHC等位基因 h 的依赖性分数,其指示MHC等位基因 h 是否将至少基于肽序列 p^k 的氨基酸的位置和MHC等位基因肽序列 d_h 的氨基酸的位置来呈递肽 p^k 。例如,如果MHC等位基因 h 在给定输入MHC等位基因肽序列 d_h 的情况下有可能呈

递肽 p^k , 则MHC等位基因h的依赖性分数可能具有高的值, 而如果不太可能呈递则可能具有低的值。变换函数 $f(\cdot)$ 变换输入, 更具体地说, 将在这种情况下由 $g_H([p^k d_h]; \theta_H)$ 产生的依赖性分数变换为合适的值, 以指示肽 p^k 将被MHC等位基因h呈递的可能性。

[0587] 在整个说明书的其余部分中提到的一个特定实施方式中, $f(\cdot)$ 是对于适当域范围具有在 $[0, 1]$ 内的范围的函数。在一个实例中, $f(\cdot)$ 是expit函数。作为另一个实例, 当域 z 的值等于或大于0时, $f(\cdot)$ 也可以是双曲正切函数。可替代地, 当预测到质谱离子电流具有超出范围 $[0, 1]$ 的值时, $f(\cdot)$ 可以是任何函数, 例如恒等函数、指数函数、对数函数等。

[0588] 因此, 可以通过将依赖性函数 $g_H(\cdot)$ 应用于肽序列 p^k 的编码形式以及应用于MCH等位基因肽序列 d_h 的编码形式以产生相应依赖性分数来产生肽序列 p^k 将被MHC等位基因h呈递的可能性。依赖性分数可以由变换函数 $f(\cdot)$ 变换以产生肽序列 p^k 将被MHC等位基因h呈递的可能性。

[0589] VIII.D.5. 等位基因相互作用变量的依赖性函数

[0590] 在整个说明书中提到的一个特定实施方式中, 依赖性函数 $g_H(\cdot)$ 是由下式得到的仿射函数:

$$[0591] \quad g_H([p^k d_h]; \theta_H) = \sum_{i=1}^{n_{pep}} \sum_{j=1}^{n_{MHC}} \sum_{k=1}^{20} \sum_{l=1}^{20} 1[p_i^k = k] 1[d_{hj} = l] \theta_{H,ijkl} + \alpha, \quad (25)$$

[0592] 其中 α 是截距; p_i^k 表示在肽 p^k 的位置 i 处的残基; d_{hj} 表示在MHC等位基因h的位置 j 处的残基; $1[\cdot]$ 表示指示变量, 如果方括号内的条件是真, 则该指示变量的值为1, 否则为0; 如果肽 p^k 的位置 i 处的氨基酸为氨基酸 k , 则 $p_i^k = k$ 为真, 否则为假; 如果MHC等位基因h的位置 j 处的氨基酸是氨基酸 l , 则 $d_{hj} = l$ 为真, 否则为假; n_{pep} 表示所建模的肽的长度; n_{MHC} 表示模型中所考虑的MHC残基的数目; 并且 $\theta_{H,ijkl}$ 是描述了在肽的位置 i 处具有残基 k 和在MHC等位基因的位置 j 处具有残基 l 对呈递可能性的贡献的系数。这是一个热编码的肽序列和一个热编码的MHC等位基因序列中的线性模型, 其具有针对所有肽残基和MHC等位基因残基的肽-残基-MHC-残基相互作用。

[0593] 在整个说明书中提到的一个特定实施方式中, 依赖性函数 $g_H(\cdot)$ 是由下式得到的网络函数:

$$[0594] \quad g_H([p^k d_h]; \theta_H) = NN_H([p^k d_h]; \theta_H) \quad (26)$$

[0595] 该网络函数由具有布置在一个或多个层中的一系列节点的网络模型 $NN_H(\cdot)$ 表示。一个节点可以通过连接而连接至其它节点, 这些连接各自在参数集合 θ_H 中具有相关参数。在一个特定节点处的值可以表示为连接至该特定节点的节点值的总和, 该值由与该特定节点相关联的激活函数映射的相关参数加权。由于呈递模型可以并入具有不同氨基酸序列长度的非线性和工艺数据, 与仿射函数相比, 网络模型是有利的。具体而言, 通过非线性建模, 网络模型可以捕捉在肽序列中不同位置处的氨基酸之间的相互作用以及在MHC等位基因肽序列中的不同位置处的氨基酸之间的相互作用, 并且可以捕捉这些相互作用如何影响肽呈递。

[0596] 一般来说, 网络模型 $NN_H(\cdot)$ 可以被构造成前馈网络, 如人工神经网络(ANN)、卷积神经网络(CNN)、深度神经网络(DNN), 和/或循环网络, 如长短期记忆网络(LSTM)、双向循环网络、深度双向循环网络等。

[0597] 在一个实例中,单个网络模型 $NN_h(\cdot)$ 可以是在被给予MHC等位基因h的编码肽序列 p^k 和编码蛋白序列 d_h 时能输出依赖性分数的网络模型。在这样的实例中,参数集合 θ_h 可以对应于单个网络模型的参数集合,因此,参数集合 θ_h 可以被所有MHC等位基因共享。因此,在这样的实例中,在向单个网络模型提供任何输入 $[p^k d_h]$ 时, $NN_h(\cdot)$ 可以表示单个网络模型 $NN_h(\cdot)$ 的输出。如上所讨论的,此种网络模型是有利的,因为仅通过鉴别MHC等位基因的蛋白质序列就可以预测训练数据中未知的MHC等位基因的肽呈递概率。

[0598] 图13示出了MHC等位基因共享的示例网络模型 $NN_h(\cdot)$ 。如图13所示,网络模型 $NN_h(\cdot)$ 接收MHC等位基因h的肽序列 p^k 和蛋白质序列 d_h 作为输入,并输出对应于MHC等位基因h的依赖性分数 $NN_h([p^k d_h])$ 。

[0599] 图14示出了示例网络模型 $NN_h(\cdot)$ 。如图14中所示,网络模型 $NN_h(\cdot)$ 包括在层 $l=1$ 处的四个输入节点、在层 $l=2$ 处的五个节点、在层 $l=3$ 处的两个节点以及在层 $l=4$ 处的一个输出节点。在可替代实施方案中,网络模型 $NN_h(\cdot)$ 可含有任意数量的层,并且每一层可以含有任意数量的节点。网络模型 $NN_h(\cdot)$ 与13个非零参数 $\theta_h(1), \theta_h(2), \dots, \theta_h(13)$ 的集合相关联。这些参数用于变换通过网络模型从一个节点传播到另一个节点的值。

[0600] 如图14所示,在网络模型 $NN_h(\cdot)$ 的层 $l=1$ 处的四个输入节点接收包括编码的多肽序列数据和编码的MHC等位基因肽序列数据的输入值。编码的多肽序列数据含有肽的氨基酸序列,而编码的MHC等位基因肽序列数据含有可能(或可能不)呈递该肽的MHC等位基因的氨基酸序列。在某些实施方案中,一旦经由在层 $l=1$ 处的输入节点输入网络模型 $NN_h(\cdot)$ 中,编码的多肽序列就被串接至网络模型 $NN_h(\cdot)$ 的一层内的编码的MHC等位基因肽序列的前面。这些输入值然后根据参数值通过网络模型 $NN_h(\cdot)$ 传播。在一些实施方案中,网络模型 $NN_h(\cdot)$ 的层包括两个完全连接的密集网络层。在进一步的实施方案中,这两个完全连接的密集网络层中的第一层包括64-128个节点,具有已校正的线性单元激活函数。在甚至进一步的实施方案中,这两个完全连接的密集网络层中的第二层包括具有线性输出的单个节点。在这样的实施方案中,该单个节点可以是网络模型 $NN_h(\cdot)$ 的输出节点。最后,网络模型 $NN_h(\cdot)$ 输出值 $NN_h([p^k d_h])$ 。该输出代表MHC等位基因h的依赖性分数,指示MHC等位基因h是否将呈递肽序列 p^k 。该网络函数还可以包括一个或多个网络模型,每个网络模型采用不同的等位基因相互作用变量(例如,肽序列)作为输入。

[0601] 在又一个实例中,依赖性函数 $g_h(\cdot)$ 可以表示为:

$$[0602] \quad g_h([p^k d_h]; \theta_h) = g'_h([p^k d_h]; \theta'_h) + \theta_h^0$$

[0603] 其中 $g'_h([p^k d_h]; \theta'_h)$ 是具有参数集合 θ'_h 的仿射函数、网络函数等,其中等位基因相互作用变量的共享参数集合 θ_h 中的偏差参数 θ_h^0 表示任何MHC等位基因的基线呈递概率。

[0604] 在另一种实施方式中,偏差参数 θ_h^0 可以根据MHC等位基因h的基因家族共享的。也就是说,MHC等位基因h的偏差参数 θ_h^0 可以等于 $\theta_{\text{基因}(h)}^0$,其中基因(h)是MHC等位基因h的基因家族。举例来说,I类MHC等位基因HLA-A*02:01、HLA-A*02:02及HLA-A*02:03可以指定给“HLA-A”基因家族,并且这些MHC等位基因各自的偏差参数 θ_h^0 可以是共享的。作为另一个实例,II类MHC等位基因HLA-DRB1:10:01、HLA-DRB1:11:01及HLA-DRB3:01:01可以指定给“HLA-DRB”基因家族,并且这些MHC等位基因中的每一种的偏差参数 θ_h^0 可以是共享的。如上所讨论的,基因家族可能是与MHC等位基因h相关联的等位基因相互作用变量之一。

[0605] 再回到等式(23),作为一个实例,在使用仿射依赖性函数 $g_H(\cdot)$ 时,肽 p^k 将被MHC等位基因 h 呈递的可能性可以由下式得出:

$$[0606] \quad u_k^h = f \left(\sum_{l=1}^{n_{pep}} \sum_{j=1}^{n_{MHC}} \sum_{k=1}^{20} \sum_{l=1}^{20} 1[p_l^k = k] 1[d_{hj} = l] \theta_{H,ijkl} + \alpha \right),$$

[0607] 其中 α 是截距; p_i^k 表示在肽 p^k 的位置 i 处的残基; d_{hj} 表示在MHC等位基因 h 的位置 j 处的残基; $1[\cdot]$ 表示指示变量,如果方括号内的条件是真,则该指示变量的值为1,否则为0;如果肽 p^k 的位置 i 处的氨基酸为氨基酸 k ,则 $p_i^k = k$ 为真,否则为假;如果MHC等位基因 h 的位置 j 处的氨基酸是氨基酸 l ,则 $d_{hj} = l$ 为真,否则为假; n_{pep} 表示所建模的肽的长度; n_{MHC} 表示模型中所考虑的MHC残基的数目;并且 $\theta_{H,ijkl}$ 是描述了在肽的位置 i 处具有残基 k 和在MHC等位基因的位置 j 处具有残基 l 对呈递可能性的贡献的系数。这是一个热编码的肽序列和一个热编码的MHC等位基因序列中的线性模型,其具有针对所有肽残基和MHC等位基因残基的肽-残基-MHC-残基相互作用。

[0608] 作为另一个实例,使用网络变换函数 $g_H(\cdot)$ 时,肽 p^k 将被MHC等位基因 h 呈递的可能性可以由下式得出:

$$[0609] \quad u_k^h = f(NN_H([p^k \ d_h]; \theta_H)),$$

[0610] 其中 p^k 表示肽序列, d_h 表示MHC等位基因 h 的肽序列, θ_H 是为与所有MHC等位基因相关联的网络模型 $NN_H(\cdot)$ 确定的参数的集合。

[0611] 图15示出了使用示例共享网络模型 $NN_H(\cdot)$ 产生与MHC等位基因 h 相关联的肽 p^k 的呈递可能性。如图15所示,共享网络模型 $NN_H(\cdot)$ 接收肽序列 p^k 和MHC等位基因肽序列 d_h ,并产生输出 $NN_H([p^k \ d_h])$ 。输出由函数 $f(\cdot)$ 映射,以生成估计的呈递可能性 u_k 。

[0612] VIII.D.6.等位基因非相互作用变量

[0613] 如上所述,等位基因非相互作用变量包括影响与MHC等位基因的类型无关的肽的呈递的信息。例如,等位基因非相互作用变量可包括该肽的N端上和C端上的蛋白质序列、所呈递肽的蛋白质家族、该肽的源基因的RNA表达水平,以及任何另外的等位基因非相互作用变量。

[0614] 在一种实施方式中,训练模块316以与针对独立等位基因模型和复等位基因模型(multiple allele model)所描述的相似的方式,将等位基因非相互作用变量并入到泛等位基因呈递模型中。例如,在一些实施方案中,等位基因非相互作用变量可作为输入被键入到与用于等位基因相互作用变量的依赖性函数分开的依赖性函数中。在这样的实施方案中,两个单独的依赖性函数的输出可以被求和,并且所得的总和可以被输入变换函数中以生成呈递预测。上文VIII.B.2.、VIII.B.3.、VIII.C.3.和VIII.C.6.部分中讨论了将等位基因非相互作用变量并入到泛等位基因模型以及其他模型中的此类实施方案。

[0615] VIII.D.7.复等位基因样品

[0616] 如上所述,测试样品可以含有多个MHC等位基因而不是单个MHC等位基因。实际上,大多数取自自然界的样品都包含多于一个MHC等位基因。例如,每个人类基因组均含有六个I类MHC基因座。因此,含有人类基因组的样品可以含有多达六个不同的I类MHC等位基因。因

此,含有多个MHC等位基因而不是单个MHC等位基因的样品是现实生活中测试案例的典型样品。

[0617] 在测试样品含有多个MHC等位基因的实施方案中,上文VIII.D.4.-VIII.D.6.部分中描述的泛等位基因模型可以用来确定来自测试样品的给定肽被多个MHC等位基因呈递的概率。然而,如上所简要描述的,当使用泛等位基因模型预测肽将被多个MHC等位基因呈递的可能性时,针对多个MHC等位基因中的每个MHC等位基因迭代地使用上述泛等位基因模型。换句话说,对于多个MHC等位基因中的每个MHC等位基因,该MHC等位基因肽序列和该肽序列被独立地输入所有MHC等位基因共享的依赖性函数中。基于这些输入,依赖性函数生成与MHC等位基因相对应的输出。对多个MHC等位基因中的每个MHC等位基因迭代地执行此过程。因此,多个MHC等位基因中的每个MHC等位基因独立地与依赖性函数的输出相关联。然后,将对与多个MHC等位基因中的每个MHC等位基因相关联的输出进行组合。

[0618] 与多个MHC等位基因中的每个MHC等位基因相关联的依赖性函数的输出可以如关于VIII.C.-VIII.C.7.部分所述的那样进行组合。如关于VIII.C.-VIII.C.7.部分所述的,依赖性函数的多个输出组合的方式可以改变。例如,在一些实施方案中,可以对依赖性函数迭代的输出进行求和,并且可以将所得的总和输入变换函数中以生成呈递预测。捕获此类实施方案的等式可写为:

$$[0619] \quad u_k^h = \Pr(p^k \text{ 被呈递}; \text{MHC等位基因 } h) = f\left(\sum_{h=1}^T g_H([p^k d_h]; \theta_H)\right), \quad (27)$$

[0620] 其中T是含有多个等位基因的样品中的独特MHC等位基因的总数。在可替代实施方案中,依赖性函数迭代的每个单独的输出可以被输入变换函数中,并且来自变换函数的结果输出可以被求和以产生呈递预测。捕获此可替代实施方案的等式可写为:

$$[0621] \quad u_k^h = \Pr(p^k \text{ 被呈递}; \text{MHC等位基因 } h) = \left(\sum_{h=1}^T f(g_H([p^k d_h]; \theta_H))\right), \quad (28)$$

[0622] 这样的实施方案以及其中依赖性函数的多个输出被组合以预测肽将在复等位基因环境中被呈递的可能性的其他实施方案在上文VIII.C.-VIII.C7.部分中有进一步的讨论。

[0623] VIII.D.8.泛等位基因模型的训练

[0624] 训练泛等位基因模型涉及对与依赖性函数相关联的共享参数集合 θ_H 中的每个参数的值进行优化。具体而言,优化参数 θ_H ,以使得依赖性函数能够输出准确地指示一个或多个给定MHC等位基因是否将呈递给定肽序列的依赖性分数。

[0625] 为了优化参数 θ_H 的值,使用了训练数据170。如上所提到的,用于训练模型的训练数据170可以包括含有表达单个MHC等位基因的细胞的训练样品、含有表达多个MHC等位基因的细胞的训练样品,或含有表达单个MHC等位基因和多个MHC等位基因二者的组合的细胞的训练样品。因此,来自训练数据170的每个数据实例i被输入泛等位基因模型中,并且更具体而言,被输入泛等位基因模型的依赖性函数中。例如,在某些实施方案中,可以将MHC等位基因肽序列和肽序列输入泛等位基因模型中。然后,泛等位基因模型对这些输入进行处理,就好像该模型是按上文关于VIII.D.3-VIII.D.7.部分所述的那样被常规地使用。然而,与VIII.D.3-VIII.D.7.部分中描述的泛等位基因模型的运行过程不同,在泛等位基因模型训

练期间,还将该肽呈递的已知结果输入该模型中。换句话说,标记 y^i 也被输入该模型中。在输入泛等位基因模型中的训练样品含有表达多个MHC等位基因的细胞的实施方案中,针对样品中多个MHC等位基因中的每个等位基因,将 y^i 设置为1。

[0626] 在使用数据实例 i 的泛等位基因模型的每次迭代之后,该模型确定所预测的MHC等位基因呈递肽的概率与已知标记 y^i 之间的差值。然后,为了最大程度地减小这种差值,泛等位基因模型会修改参数 θ_H 。换句话说,泛等位基因模型通过最大程度地减小关于 θ_H 的损失函数来确定参数 θ_H 的值。当泛等位基因模型达到一定水平的预测准确率时,训练就完成了,并且该模型就已准备就绪,可以按VIII.D.3-VIII.D.7.部分中所述的那样使用。

[0627] VIII.D.9.泛等位基因模型实例

[0628] 以下实例比较了示例独立等位基因呈递模型和示例泛等位基因呈递模型的预测精确率(即阳性预测值)。在这个实例中,使用相同的训练数据集来训练等位基因呈递模型和泛等位基因呈递模型。训练后,使用六个测试样品对独立等位基因呈递模型和泛等位基因呈递模型进行测试。注意,训练数据集含有在每个测试样品中测试的每个MHC等位基因的充足训练数据。下表2示出了在使用独立等位基因模型和泛等位基因模型时在40%召回率下的预测精确率(或阳性预测值)。由于在六个样品中测试的每个MHC等位基因都具有充足训练数据,因此独立等位基因模型以平均精确率高出0.04而略优于泛等位基因模型。

	样品 1	样品 2	样品 3	样品 4	样品 5	样品 6	均值
[0629] 泛等位基因	0.310	0.446	0.341	0.433	0.547	0.304	0.397
独立等位基因	0.363	0.517	0.349	0.458	0.565	0.368	0.437

[0630] 表2

[0631] 然而,可以在针对图16-22讨论的可替代实验中观察到泛等位基因模型预测未包括在用于训练模型的训练数据集中的MHC等位基因的呈递可能性的能力。

[0632] 图16-22描绘了被设计用于测试泛等位基因模型预测未训练的MHC等位基因将呈递给定肽的概率的实验的结果。特别地,图16-18描绘了被设计用于测试包含神经网络模型的泛等位基因模型预测未训练的MHC等位基因将呈递给定肽的概率的实验的结果。另一方面,图19-22描绘了被设计用于测试包含非神经网络模型的泛等位基因模型预测未训练的MHC等位基因将呈递给定肽的概率的实验的结果。

[0633] 首先来看与图16-18相关联的实验,为了证明包含神经网络模型的泛等位基因模型预测未训练的MHC等位基因将呈递给定肽的概率的能力,将由未用正在测试的MHC等位基因训练的包含神经网络模型的泛等位基因模型产生的预测与已用正在测试的MHC等位基因训练的相同的泛等位基因模型产生的预测进行比较。换句话说,该泛等位基因模型之间的唯一差异是用来对其进行训练的训练数据集。尚未在包含测试的HLA等位基因的样品上进行训练的泛等位基因模型的预测精确率相对于已经在包含测试的HLA等位基因的样品上进行训练的泛等位基因模型的预测精确率越高,则泛等位基因模型预测未用于训练该泛等位基因模型的MHC等位基因的呈递可能性的能力越大。

[0634] 如上所注意到的,在与图16-18相关联的实验内使用的泛等位基因模型在用不同

的训练数据集训练之前是相同的。同样如上所注意到的,在与图16-18相关联的实验内使用的每个泛等位基因模型均包含神经网络模型作为其依赖性函数。泛等位基因模型中使用的神经网络模型含有单个隐藏层。在神经网络模型的隐藏层之间的激活函数是已校正的线性单元 (ReLU) 函数, $f(x) = \max(0, x)$ 。神经网络模型的最后一层包含线性激活层 $f(x) = x$ 。神经网络模型的每个子网络的隐藏单元的数目取决于对神经网络模型的输入。具体而言,对于被配置为接收mRNA丰度的神经网络模型,该神经网络模型的mRNA丰度子网络中的隐藏单元数目为16。对于被配置为接收编码的侧接序列的神经网络模型,该神经网络模型的侧接序列子网络中的隐藏单元数目为32。对于被配置为接收编码的多肽序列的神经网络模型,神经网络模型的多肽序列子网络中的隐藏单元数目为256。对于被配置为接收编码的多肽序列和编码的MHC等位基因肽序列的神经网络模型(如在泛等位基因模型的情况下),神经网络模型的多肽和MHC等位基因肽序列子网络中的隐藏单元数目为128。

[0635] 与图16-18相关联的每个实验均包括唯一的测试样品,每个唯一的测试样品均包含不同的HLA等位基因。为了证明这些实验产生的结果不限于特定的基因座,从三个基因座A、B和C的每一个中选择出等位基因。因此,第一测试样品含有HLA-A等位基因,第二样品含有HLA-B等位基因,第三样品含有HLA-C等位基因。具体而言,第一测试样品含有HLA等位基因A*02:03,第二测试样品含有HLA等位基因B*54:01,第三测试样品含有HLA等位基因C*08:02。这些HLA等位基因中每一种的蛋白质序列均从Anthony Nolan Research Institute维护的HLA蛋白质序列数据库 (<https://www.ebi.ac.uk/ipd/imgt/hla/>) 中获得。

[0636] 针对三个样品中的每一个,将该特定HLA等位基因的蛋白质序列和所讨论的肽的蛋白质序列输入尚未使用该HLA等位基因训练的第一泛等位基因模型中,并输入已使用该HLA等位基因训练的第二相同泛等位基因模型中。该泛等位基因模型输出所预测的HLA等位基因将呈递该肽的概率。将这些预测的概率与该肽呈递的已知结果(即标记 y^i)进行比较,以产生图16-18所示的精确率/召回率曲线。具体而言,图16对应于由泛等位基因模型输出的针对第一测试样品的数据输出。图17对应于由泛等位基因模型输出的针对第二测试样品的数据输出,而图18对应于由泛等位基因模型输出的针对第三测试样品的数据输出。在每个图中,蓝色线表示已在包含测试的HLA等位基因的样品上训练的泛等位基因模型的精确率/召回率曲线,橙色线表示尚未在包含测试的HLA等位基因的任何样品上训练的泛等位基因模型的精确率/召回率曲线。另外,每个图均指示出训练和未训练的泛等位基因模型的平均预测精确率(即阳性预测值)。例如,如在图18中可以看到,已经在包含测试的HLA等位基因的样品上训练的泛等位基因模型的平均预测精确率是0.256,而尚未在包含测试的HLA等位基因的样品上训练的泛等位基因模型的平均预测精确率是0.231。

[0637] 如图16-18所示,即使由橙色线表示的泛等位基因模型从未见过正在测试的HLA等位基因,这些泛等位基因模型仍能够获得与在培训期间已见过正在测试的HLA等位基因的由蓝色线表示的泛等位基因模型相当的性能。因此,这些结果证明了包含神经网络模型的泛等位基因模型能够准确地预测未用于训练该泛等位基因模型的HLA等位基因的呈递可能性。

[0638] 接下来看与图19-22相关联的实验,为了证明包含非神经网络模型的泛等位基因模型预测未被训练的MHC等位基因将呈递给定肽的概率的能力,在每个实验内比较了四种模型的性能。该四种模型包括:如上针对图16-18所述的包含神经网络模型的泛等位基因呈

递模型、由1,000棵树组成的现用随机森林模型、适合多元高斯函数 (multivariate Gaussians) 的现用二次判别分析 (QDA) 模型,以及适合每个等位基因的独特前馈的完全连接的神经网络的当前最先进的MHC 1类结合亲和力模型MHCFlurry。随机森林模型和二次判别模型都基于包含非神经网络模型的泛等位基因模型架构。

[0639] 与图19-22相关联的每个实验均包括测试样品,并且每个测试样品均含有HLA等位基因。为了证明这些实验产生的结果不限于特定的基因座,从三个基因座A、B和C的每一个中选择出等位基因。因此,第一测试样品和第二测试样品含有HLA-A等位基因,第三样品含有HLA-B等位基因,第四样品含有HLA-C等位基因。具体而言,第一测试样品和第二测试样品含有HLA等位基因A*02:01,第三测试样品含有HLA等位基因B*44:02,第四测试样品含有HLA等位基因C*08:02。这些HLA等位基因中的每一种的蛋白质序列均从AnthonyNolanResearch Institute维护的HLA蛋白质序列数据库 (<https://www.ebi.ac.uk/ipd/imgt/hla/>) 中获得。

[0640] 在针对四个测试样品中的每个样品训练用于预测呈递可能性的四个模型期间,泛等位基因呈递模型、随机森林模型和二次判别模型各自在由来自31个不同等位基因的9-聚体组成并包含HLA-A、HLA-B和HLA-C的单等位基因数据上进行训练。另一方面,MHCFlurry模型由其作者使用包括来自HLA-A、HLA-B和HLA-C的等位基因的IEDB和BD2013结合亲和力数据集的子集进行训练。每个等位基因均利用8个神经网络的集合体单独建模,等位基因名称直接传递给模型,以选择要使用哪个等位基因子模型来生成呈递预测。[76]。

[0641] 用于针对四个测试样品中的每个样品训练该四个模型的特定等位基因取决于给定测试样品中含有的HLA等位基因。具体而言,对于含有HLA等位基因A*02:01的第一测试样品,用于训练该四个模型以预测HLA等位基因A*02:01的呈递可能性的训练数据包括HLA等位基因A*02:01。对于含有HLA等位基因A*02:01的第二测试样品,用于训练该四个模型以预测HLA等位基因A*02:01的呈递可能性的训练数据不包括HLA等位基因A*02:01。对于含有HLA等位基因B*44:02的第三测试样品,用于训练该四个模型以预测HLA等位基因B*44:02的呈递可能性的训练数据不包括HLA等位基因B*44:02。对于含有HLA等位基因C*08:02的第四测试样品,用于训练该四个模型以预测HLA等位基因C*08:02的呈递可能性的训练数据不包括HLA等位基因C*08:02。

[0642] 在对该四个样品中的每个样品进行测试期间,在留出的单等位基因数据集上测试每个模型,所述单等位基因数据集包括给定样品中的HLA等位基因,并且由约250,000个肽(对呈递肽和非呈递肽二者进行计数)组成。具体而言,在对该四个样品中的每个样品进行测试期间,泛等位基因呈递模型、随机森林模型和二次判别模型各自接收相同的输入。特别地,对于该四个样品中的每个样品,泛等位基因呈递模型、随机森林模型和二次判别模型各自接收了样品内HLA等位基因的34聚体的独热编码HLA等位基因蛋白质序列以及所讨论的肽的9聚体的独热编码(即,二值化)的蛋白质序列。另一方面,对于该四个样品中的每个样品,MHCFlurry模型都接收样品内HLA等位基因的名称,以及所讨论肽的9聚体的独热编码(即二值化)蛋白质序列。如上所述,该模型之间的输入差异是以下事实的结果:MHCFlurry模型被配置为使用等位基因的名称来选择使用哪个等位基因子模型来生成呈递预测。

[0643] 在这些输入被输入该四个模型中之后,该四个模型中的每个模型然后均输出HLA等位基因将呈递肽的预测概率。将这些预测概率与肽呈递的已知结果(即标记 y^i)进行比

较,以产生图19-22所示的精确率/召回率曲线。具体而言,图19对应于由该四个模型中的每个模型输出的针对第一测试样品的数据,图20对应于由该四个模型中的每个模型输出的针对第二测试样品的数据,图21对应于由该四个模型中的每个模型输出的针对第三测试样品的数据,而图22对应于由该四个模型中的每个模型输出的针对第四测试样品的数据。在每个图中,蓝色线显示了泛等位基因模型的精确率/召回率曲线,橙色线显示了MHCFlurry模型的精确率/召回率曲线,绿色线显示了随机森林模型的精确率/召回率曲线,而红色线显示了二次判别模型的精确率/召回率曲线。另外,每个图指示出每个模型的平均预测精确率(即阳性预测值)。例如,如在图19中可以看到,泛等位基因模型的平均预测精确率为0.32。

[0644] 如图19-22中所示,随机森林模型和二次判别模型都使用了包含非神经网络模型的泛等位基因模型架构,它们的性能大约是MHCFlurry模型的两倍。此外,包含神经网络模型的泛等位基因呈递模型的性能大约是使用包含非神经网络模型的泛等位基因模型架构的随机森林模型和二次判别模型的两倍。换句话说,相对于其他模型,包含神经网络模型的泛等位基因呈递模型获得了最高的精确率。然而,使用包含非神经网络模型的泛等位基因模型架构的随机森林模型和二次判别模型的性能仍然优于定制的独立等位基因结合亲和力模型MHCFlurry。因此,这些结果表明,泛等位基因模型架构可以很好地推广到其他非神经网络机器学习模型(这些模型与基于决策树的随机森林和贝叶斯方法(Bayesianmethod)(如二次判别分析)一样多种多样)中,同时仍提供高水平的预测精确率。

[0645] 另外,如图20-22中进一步所示,尽管泛等位基因呈递模型、随机森林模型和二次判别模型从未见过正在测试的HLA等位基因,但包括使用了包含非神经网络模型的泛等位基因模型架构的随机森林模型和二次判别模型在内的这些模型能够获得与对应于图19的在训练期间已看到过正在测试的HLA等位基因的模型相当的性能。因此,这些结果证明了包含非神经网络的泛等位基因模型架构能够准确地预测未用于训练该模型的HLA等位基因的呈递可能性。

[0646] IX. 实施例5:预测模块

[0647] 预测模块320接收序列数据并使用呈递模型在序列数据中选择候选新抗原。确切地说,序列数据可以是来自患者的肿瘤组织细胞中提取的DNA序列、RNA序列和/或蛋白质序列。预测模块320将序列数据处理成对于MHC-I具有8-15个氨基酸或对于MHC-II具有6-30个氨基酸的多个肽序列 p^k 。举例来说,预测模块320可以将给定序列“IEFROEIFJEF”处理成具有9个氨基酸的三个肽序列“IEFROEIFJ”、“EFROEIFJE”及“FROEIFJEF”。在一个实施方案中,预测模块320可以通过将从患者的正常组织细胞提取的序列数据与从患者的肿瘤组织细胞提取的序列数据相比较以鉴别含有一个或多个突变的部分,由此鉴别出作为突变肽序列的候选新抗原。

[0648] 预测模块320将一个或多个呈递模型应用于处理的肽序列以估计这些肽序列的呈递可能性。确切地说,预测模块320可以通过将呈递模型应用于候选新抗原来选择一个或多个可能被呈递于肿瘤HLA分子上的候选新抗原肽序列。在一种实施方式中,预测模块320选出估计呈递可能性超过预定阈值的候选新抗原序列。在另一种实施方式中,呈递模块选出 $\underline{v}$ 个具有最高估计呈递可能性的候选新抗原序列(其中 $\underline{v}$ 一般是在疫苗中递送的表位的最大数量)。包括选择用于给定患者的候选新抗原的疫苗可以注射到患者体内以诱导

免疫应答。

[0649] X. 实施例6: 患者选择模块

[0650] 患者选择模块324基于患者是否满足纳入标准来选择用于疫苗治疗和/或T细胞疗法的患者子集。在一个实施方案中,基于由呈递模块产生的患者新抗原候选物的呈递可能性来确定纳入标准。通过调整纳入标准,患者选择模块324可以基于患者新抗原候选物的呈递可能性来调整将要接受疫苗和/或T细胞疗法的患者数量。具体地,严格的纳入标准导致较少数量的患者将被用疫苗和/或T细胞疗法治疗,但是可以导致较高比例的疫苗和/或T细胞疗法所治疗患者接受有效治疗(例如接受一种或多种肿瘤特异性新抗原(TSNA)和/或一种或多种新抗原响应性T细胞)。另一方面,宽松的纳入标准导致较大数量的患者将被用疫苗和/或T细胞疗法治疗,但是可导致较低比例的疫苗和/或T细胞疗法所治疗患者接受有效治疗。患者选择模块324基于将要接受治疗的患者的目标比例与接受有效治疗的患者比例之间的期望平衡来修改纳入标准。

[0651] 在一些实施方案中,用于选择接受疫苗治疗的患者的纳入标准与用于选择接受T细胞疗法的患者的纳入标准相同。然而,在替代实施方案中,用于选择接受疫苗治疗的患者的纳入标准可能不同于用于选择接受T细胞疗法的患者的纳入标准。以下X.A和X.B部分分别讨论了用于选择接受疫苗治疗的患者的纳入标准和用于选择接受T细胞疗法的患者的纳入标准。

[0652] X.A. 对于疫苗治疗的患者的选择

[0653] 在一个实施方案中,患者与 v 种新抗原候选物的相应治疗子集相关联,所述 v 种新抗原候选物可以潜在地包含在具有疫苗容量 v 的为患者定制疫苗中。在一个实施方案中,患者的治疗子集是通过呈递模型确定的具有最高呈递可能性的新抗原候选物。例如,如果疫苗可以包含 $v=20$ 种表位,则疫苗可以包含每个患者的通过呈递模型确定的具有最高呈递可能性的治疗子集。但是,应当理解,在另一些实施方案中,可以基于其它方法来确定患者的治疗子集。例如,患者的治疗子集可以从患者的新抗原候选物集合中随机选择,或者可以部分地基于对肽序列的结合亲和力和稳定性建模的当前现有技术模型,或者包括来自呈递模型的呈递可能性和关于这些肽序列的亲和力或稳定性信息的一些因素组合来确定。

[0654] 在一个实施方案中,如果患者的肿瘤突变负荷等于或高于最小突变负荷,则患者选择模块324确定患者满足纳入标准。患者的肿瘤突变负荷(TMB)指示肿瘤外显子组中非同义突变的总数。在一种实施方式中,如果患者的TMB的绝对数量等于或高于预定阈值,则患者选择模块324可以选择患者进行疫苗治疗。在另一种实施方式中,如果患者的TMB在为患者集合确定的TMB中的阈值百分位数之内,则患者选择模块324可以选择患者用于进行疫苗治疗。

[0655] 在另一个实施方案中,如果基于患者治疗子集的患者效用分数等于或高于最小效用分数,则患者选择模块324确定患者满足纳入标准。在一种实施方式中,效用分数是对治疗子集中被呈递的新抗原的估计数目的度量。

[0656] 可以通过将新抗原呈递建模为一种或多种概率分布的随机变量来预测被呈递的新抗原的估计数目。在一种实施方式中,患者 i 的效用分数是治疗子集中被呈递的新抗原候选物的预期数目,或其一些函数。作为示例,每种新抗原的呈递可以建模为伯努利随机变量,其中呈递(成功)概率由新抗原候选物的呈递可能性给出。具体来说,对于 v 种新抗原候

选物 p^{i1} 、 p^{i2} 、 $\dots$ 、 p^{iv} 的治疗子集 S_i ,每种新抗原候选物具有最高呈递可能性 u_{i1} 、 u_{i2} 、 $\dots$ 、 u_{iv} ,则新抗原候选物 p^{ij} 的呈递由随机变量 A_{ij} 给出,其中:

$$[0657] \quad P(A_{ij}=1)=u_{ij}, P(A_{ij}=0)=1-u_{ij}。 \quad (29)$$

[0658] 被呈递的新抗原的预期数目由每种新抗原候选物的呈递可能性之和给出。换句话说,患者i的效用分数可以表示为:

$$[0659] \quad \text{util}_i(S_i) = \mathbb{E} \left[\sum_{j=1}^v A_{ij} \right] = \sum_{j=1}^v u_{ij}。 \quad (30)$$

[0660] 患者选择模块324选择效用分数等于或高于最小效用的患者子集进行疫苗治疗。

[0661] 在另一种实施方式中,患者i的效用分数是至少阈值数目k的新抗原将被呈递的概率。在一个示例中,将新抗原候选物的治疗子集 S_i 中被呈递的新抗原的数目建模为泊松二项式随机变量,其中呈递(成功)概率由每种表位的呈递可能性给出。具体来说,患者i的被呈递的新抗原的数目可以通过随机变量 N_i 给出,其中:

$$[0662] \quad N_i = \sum_{j=1}^v A_{ij} \sim \text{PBD}(u_{i1}, u_{i2}, \dots, u_{iv})。 \quad (31)$$

[0663] 其中 $\text{PBD}(\cdot)$ 表示泊松二项式分布。至少阈值数目k的新抗原将被呈递的概率由被呈递的新抗原的数目 N_i 等于或大于k的概率的总和给出。换句话说,患者i的效用分数可以表示为:

$$[0664] \quad \text{util}_i(S_i) = \mathbb{P}[N_i \geq k] = \sum_{m=k}^v \mathbb{P}[N_i = m]。 \quad (32)$$

[0665] 患者选择模块324选择效用分数等于或高于最小效用的患者子集进行疫苗治疗。

[0666] 在另一种实施方式中,患者i的效用分数是新抗原候选物的治疗子集 S_i 中具有低于对一种或多种患者HLA等位基因的固定阈值(例如500nM)的结合亲和力或预测的结合亲和力的新抗原的数目。在一个实例中,固定阈值在1000nM至10nM的范围。任选地,效用分数可以仅计数通过RNA-seq检测到的那些新抗原。

[0667] 在另一种实施方式中,患者i的效用分数是新抗原候选物的治疗子集 S_i 中对一种或多种患者HLA等位基因的结合亲和力等于或低于随机肽对该HLA等位基因的结合亲和力的阈值百分位数的新抗原的数量。在一个实例中,阈值百分位数是从第10个百分位数到第0.1个百分位数的范围。任选地,效用分数可以仅计数通过RNA-seq检测到的那些新抗原。

[0668] 应当理解,关于等式(25)和(27)示出的生成效用分数的实例仅是示例性的,并且患者选择模块324可以使用其它统计学或概率分布来生成效用分数。

[0669] X.B. 对于T细胞疗法的患者选择

[0670] 在另一个实施方案中,代替接受疫苗治疗或除了接受疫苗治疗之外,患者可以接受T细胞疗法。像疫苗治疗一样,在患者接受T细胞疗法的实施方案中,患者可以如上所述与v种新抗原候选物的相应治疗子集相关联。v种新抗原候选物的该治疗子集可用于体外鉴别来自患者的对v种新抗原候选物中的一种或多种有响应的T细胞。然后可以扩增鉴别的T细胞并且输注回患者体内以进行定制的T细胞疗法。

[0671] 可以选择患者以在两个不同的时间点接受T细胞疗法。第一个点是在已经使用模型为患者预测了v种新抗原候选物的治疗子集之后,但在体外筛选对v种新抗原候选物的预测的治疗子集具有特异性的T细胞之前。第二个点是在体外筛选对v种新抗原候选物的预测的治疗子集具有特异性的T细胞之后。

[0672] 首先,可以在已经为患者预测了v种新抗原候选物的治疗子集之后,但在体外鉴别来自患者的对v种新抗原候选物的预测的子集具有特异性的T细胞之前来选择患者以接受T细胞疗法。具体地,由于来自患者的新抗原特异性T细胞的体外筛选可能是昂贵的,所以可能期望的是仅在患者可能具有新抗原特异性T细胞的情况下才选择患者来筛选新抗原特异性T细胞。为了在体外T细胞筛选步骤之前选择患者,可以使用与用于选择对于疫苗治疗的患者相同的标准。具体地,在一些实施方案中,如果如上所述患者的肿瘤突变负荷等于或高于最小突变负荷,则患者选择模块324可选择患者以接受T细胞疗法。在另一个实施方案中,如果如上所述基于患者的v种新抗原候选物治疗子集的患者效用分数等于或高于最小效用分数,则患者选择模块324可选择患者以接受T细胞疗法。

[0673] 第二,除了或代替在体外鉴别来自患者的对v种新抗原候选物的预测的子集具有特异性的T细胞之前来选择患者以接受T细胞疗法,还可在体外鉴别对v种新抗原候选物的预测的治疗子集具有特异性的T细胞之后选择患者以接受T细胞疗法。具体地,如果在患者T细胞的新抗原识别的体外筛选过程中对患者鉴别了至少阈值量的新抗原特异性TCR,则可以选择患者以接受T细胞疗法。例如,仅在对患者鉴别了至少两种新抗原特异性TCR或者仅在针对两种不同新抗原鉴别了新抗原特异性TCR时,才可选择患者以接受T细胞疗法。

[0674] 在另一个实施方案中,仅在患者的v种新抗原候选物的治疗子集中的至少阈值量的新抗原被患者TCR识别时,才可选择患者以接受T细胞疗法。例如,仅在患者的v种新抗原候选物的治疗子集中的至少一种新抗原被患者TCR识别时,才可选择患者以接受T细胞疗法。在另一些实施方案中,仅在患者的至少阈值量的TCR被鉴别为对特定HLA限制类别的新抗原肽具有新抗原特异性时,才可选择患者以接受T细胞疗法。例如,仅在患者的至少一种TCR被鉴别为对I类HLA限制的新抗原肽具有新抗原特异性时,才可选择患者以接受T细胞疗法。

[0675] 在甚至另一些实施方案中,仅在特定HLA限制类别的至少阈值量的新抗原肽被患者TCR识别时,才可选择患者以接受T细胞疗法。例如,仅在至少一种I类HLA限制的新抗原肽被患者TCR识别时,才可选择患者接受T细胞疗法。又例如,仅在至少两种II类HLA限制的新抗原肽被患者TCR识别时,才可选择患者以接受T细胞疗法。在体外鉴别对患者的v种新抗原候选物的预测治疗子集具有特异性的T细胞之后,上述标准的任何组合也可以用于选择患者以接受T细胞疗法。

[0676] XI. 实施例7:显示示例患者选择表现的实验结果

[0677] 通过对模拟患者集合进行患者选择来测试在X部分中描述的患者选择方法的有效性,每个模拟患者均与模拟新抗原候选物的测试集相关联,其中已知模拟新抗原的子集存在于质谱数据中。具体地,测试集中的每个模拟新抗原候选物都与指示新抗原是否存在于来自Bassani-Sternberg数据集(数据集“D1”)(数据可在www.ebi.ac.uk/pride/archive/projects/PXD0000394上找到)的多等位基因JY细胞系HLA-A*02:01和LA-B*07:02质谱数据集的标记相关联。如以下结合图23A更详细地描述的,基于非小细胞肺癌(NSCLC)患者中突

变负荷的已知频率分布,从人蛋白质组中取样了模拟患者的许多新抗原候选物。

[0678] 使用训练集训练相同HLA等位基因的独立等位基因呈递模型,所述训练集是来自IEDB(数据集“D2”)(数据可以在http://www.iedb.org/doc/mhc_ligand_full.zip中找到)的单等位基因HLA-A*02:01和HLA-B*07:02质谱数据的子集。具体地,每个等位基因的呈递模型是等式(8)中所示的独立等位基因模型,其并入了N末端和C末端侧接序列作为等位基因非相互作用变量,以及网络依赖性函数 $g_h(\cdot)$ 和 $g_w(\cdot)$,及expit函数 $f(\cdot)$ 。等位基因HLA-A*02:01的呈递模型产生了给定肽将在等位基因HLA-A*02:01上呈递的呈递概率,给出了肽序列作为等位基因相互作用变量,并且N末端和C末端侧接序列作为等位基因非相互作用变量。等位基因HLA-B*07:02的呈递模型产生了给定肽将在等位基因HLA-B*07:02上呈递的呈递概率,给出了肽序列作为等位基因相互作用变量,并且N末端和C末端侧接序列作为等位基因非相互作用变量。

[0679] 如在以下实施例中并且参照图23A-23E所阐述的,将用于肽结合预测的多种模型例如经训练的呈递模型和当前现有技术模型应用于每个模拟患者的新抗原候选物的测试集,以基于预测来鉴别患者的不同治疗子集。选择满足纳入标准的患者进行疫苗治疗,并将其与包含患者治疗子集中的表位的定制疫苗相关联。治疗子集的大小根据不同疫苗容量而变化。在用于训练呈递模型的训练集和模拟新抗原候选物的测试集之间没有引入重叠。

[0680] 在以下实施例中,分析了具有在疫苗中所包含的表位中的至少一定数量的被呈递新抗原的选定患者的比例。该统计数据指示了模拟疫苗递送将在患者中引发免疫应答的潜在新抗原的有效性。具体地,如果新抗原存在于质谱数据集D2中,则呈递在测试集中的模拟新抗原。高比例的具有被呈递新抗原的患者指示通过诱导免疫应答通过新抗原疫苗成功治疗的潜力。

[0681] XI.A. 实施例7A: NSCLC癌症患者的突变负荷的频率分布

[0682] 图23A示出了NSCLC患者中突变负荷的样品频率分布。包括NSCLC在内的不同肿瘤类型的突变负荷和突变可以在例如癌症基因组图谱(TCGA)(<https://cancergenome.nih.gov>)上找到。x轴代表每个患者中非同义突变的数目,并且y轴代表具有给定数目的非同义突变的样品患者的比例。图23A的样品频率分布显示了一系列3-1786种突变,其中30%的患者具有少于100个突变。尽管未在图23A中示出,研究表明吸烟者的突变负荷比不吸烟者高,而且突变负荷可能是患者中新抗原负荷的强指标。

[0683] 如在以上XI部分开始时所介绍的,许多模拟患者中的每一个都与新抗原候选物的测试集相关联。对于每个患者,通过从图23A中所示的频率分布对突变负 m_i 采样来产生每个患者的测试集。对于每个突变,随机选择来自人类蛋白质组的21聚体肽序列以代表模拟突变序列。通过鉴别跨越21聚体中的突变的每种(8、9、10、11)聚体肽序列,为每个患者i产生新抗原候选物序列的测试集。每种新抗原候选物与都指示新抗原候选物序列是否存在于质谱D1数据集中的标记相关联。例如,存在于数据集D1中的新抗原候选物序列可以与标记“1”相关联,而不存在于数据集D1中的序列可以与标记“0”相关联。如以下更详细描述,图23B至23E示出了基于测试集中患者的被呈递新抗原进行的患者选择的实验结果。

[0684] XI.B. 实施例7B: 基于肿瘤突变负荷纳入标准具有新抗原呈递的选定患者的比例

[0685] 图23B示出了对于基于患者是否满足最小突变负荷的纳入标准选择的患者,模拟疫苗中被呈递的新抗原的数目。确定在相应的测试中具有至少一定数量的被呈递的新抗原

的选定患者的比例。

[0686] 在图23B中,x轴表示基于最小突变负荷(如通过标记“最小突变数”指示的)从疫苗治疗中排除的患者的比例。例如,200“最小突变数”处的数据点表示患者选择模块324仅选择具有至少200个突变的突变负荷的模拟患者的子集。又例如,300“最小突变数”处的数据点表示患者选择模块324选择了具有至少300个突变的较低比例的模拟患者。y轴表示在没有任何疫苗容量 v 的测试集中与至少一定数量的被呈递的新抗原相关联的选定患者的比例。具体来说,上方的图显示了呈递至少一种新抗原的选定患者的比例,中间的图显示了呈递至少两种新抗原的选定患者的比例,并且底部的图显示了呈递至少三种新抗原的选定患者的比例。

[0687] 如图23B所示,随着更高的突变负荷,具有被呈递的新抗原的选定患者的比例显著增加。这表明突变负荷作为纳入标准可以有效地选择新抗原疫苗更有可能诱导成功免疫应答的患者。

[0688] XI.C. 实施例7c:通过呈递模型对比于现有技术模型鉴别的疫苗的新抗原呈递的比较

[0689] 图23C在与包含基于呈递模型鉴别的治疗子集的疫苗相关的选定患者和与包含通过当前现有技术模型鉴别的治疗子集的疫苗相关的选定患者之间,比较了模拟疫苗中被呈递的新抗原的数目。左图假定有限的疫苗容量 $v=10$,而右图假定有限的疫苗容量 $v=20$ 。基于效用分数选择患者,所述效用分数指示被呈递新抗原的预期数目。

[0690] 在图23C中,实线表示与包含基于等位基因HLA-A*02:01和HLA-B*07:02的呈递模型鉴别治疗子集的疫苗相关的患者。通过将每个呈递模型应用于测试集中的序列,并且鉴别具有最高呈递可能性的 v 个新抗原候选物,来鉴别每个患者的治疗子集。虚线表示与包含基于单等位基因HLA-A*02:01的当前现有技术模型NETMHCpan鉴别的治疗子集的疫苗相关的患者。NETMHCpan的实现细节在<http://www.cbs.dtu.dk/services/NetMHCpan>中详细提供。通过将NETMHCpan模型应用于测试集中的序列,并且鉴别具有最高估计结合亲和力的 v 个新抗原候选物,来鉴别每个患者的治疗子集。两个图的x轴表示基于预期效用分数从疫苗治疗中排除的患者的比例,所述预期效用分数指示基于呈递模型鉴别的治疗子集中被呈递的新抗原的预期数目。预期效用分数的确定如参考X部分中的等式(25)所述。y轴表示呈递疫苗中所包含的至少一定数量的新抗原(1、2或3种新抗原)的选定患者的比例。

[0691] 如图23C中所示,同与包含基于现有技术模型的治疗子集的疫苗相关的患者相比,与包含基于呈递模型的治疗子集的疫苗相关的患者以显著更高的比例接受包含被呈递新抗原的疫苗。例如,如右图所示,与基于呈递模型的疫苗相关的选定患者中有80%接受疫苗中至少一种被呈递的新抗原,相比之下,与基于当前现有技术模型的疫苗相关的选定患者中仅有40%。结果表明,如本文所述的呈递模型对于选择可能引发用于治疗肿瘤的免疫应答的疫苗的新抗原候选物是有效的。

[0692] XI.D. 实施例7D:HLA覆盖率对通过呈递模型鉴别的疫苗的新抗原展呈递的影响

[0693] 图23D在与包含基于HLA-A*02:01的单独等位基因呈递模型鉴别的治疗子集的疫苗相关的选定患者和与包含基于HLA-A*02:01和HLA-B*07:02的双独立等位基因呈递模型鉴别的治疗子集的疫苗相关的选定患者之间,比较了模拟疫苗中被呈递的新抗原的数目。疫苗容量设定为 $v=20$ 个表位。对于每个实验,根据基于不同治疗子集确定的期望效用

分数选择患者。

[0694] 在图23D中,实线表示与包含基于HLA等位基因HLA-A*02:01和HLA-B*07:02的双呈递模型的治疗子集的疫苗相关的患者。通过将每个呈递模型应用于测试集中的序列,并且鉴别具有最高呈递可能性的 v 个新抗原候选物,来鉴别每个患者的治疗子集。虚线表示与包含基于HLA等位基因HLA-A*02:01的单呈递模型的治疗子集的疫苗相关的患者。通过将仅单个HLA等位基因的呈递模型应用于测试集中的序列,并且鉴别具有最高呈递可能性的 v 个新抗原候选物,来鉴别每个患者的治疗子集。对于实线图,x轴表示基于通过双呈递模型鉴别的治疗子集的预期效用分数从疫苗治疗中排除的患者的比例。对于虚线图,x轴表示基于通过单呈递模型鉴别的治疗子集的预期效用分数从疫苗治疗中排除的患者的比例。y轴表示呈递至少一定数量的新抗原(1、2或3种新抗原)的选定患者的比例。

[0695] 如图23D中所示,同与包含通过单呈递模型鉴别的治疗子集的疫苗相关的患者相比,与包含通过双HLA等位基因的呈递模型鉴别治疗子集的疫苗相关的患者以显著更高的比例呈递新抗原。结果表明了建立具有高HLA等位基因覆盖率的呈递模型的重要性。

[0696] XI.E. 实施例7E:通过肿瘤突变负荷对比于通过被呈递抗原的预期数目选择的患者的新抗原呈递的比较

[0697] 图23E在基于肿瘤突变负荷选择的患者和通过预期效用分数选择的患者之间,比较了模拟疫苗中被呈递的新抗原的数量。基于通过具有 $v=20$ 个表位的呈递模型鉴别的治疗子集确定预期效用分数。

[0698] 在图23E中,实线表示与包含通过呈递模型鉴别的治疗子集的疫苗相关的基于预期效用分数选择的患者。通过将呈递模型应用于测试集中的序列,并且鉴别具有最高呈递可能性的 $v=20$ 个新抗原候选物,来鉴别每个患者的治疗子集。基于根据X部分中的等式(25)的鉴别的治疗子集的呈递可能性来确定预期效用分数。虚线表示与还包含通过呈递模型鉴别的治疗子集的疫苗相关的基于突变负荷选择的患者。实线图的x轴表示基于预期效用分数从疫苗治疗中排除的患者的比例,虚线图的x轴表示基于突变负荷从疫苗治疗中排除的患者的比例。y轴表示接受包含至少一定数量的被呈递的抗原(1、2或3种新抗原)的疫苗的选定患者的比例。

[0699] 如图23E中所示,与基于突变负荷选择的患者相比,基于预期效用分数选择的患者以更高的比率接受包含被呈递新抗原的疫苗。然而,基于突变负荷选择的患者比未选择的患者以更高的比率接受包含被呈递新抗原的疫苗。因此,突变负荷是成功的新抗原疫苗治疗的有效患者选择标准,尽管预期效用分数更有效。

[0700] XII. 实施例8:根据留出质谱数据的质谱训练模型的评估

[0701] 由于肿瘤细胞的HLA肽呈递是抗肿瘤免疫的关键要求^{91,96,97},产生了具有配对的I类HLA肽序列的人类肿瘤和正常组织样品、HLA类型和转录组RNA-seq(方法)的大的($N=74$ 位患者)综合数据集,目的是使用这些和可公开获得的数据^{92,98,99}训练新型深度学习模型¹⁰⁰,以预测人癌症中的抗原呈递。从数种感兴趣的肿瘤类型中选择样品以进行免疫疗法开发并根据组织的可用性进行选择。质谱在肽水平 $FDR<0.1$ (范围344-11,301)鉴别了每个样品平均3,704个肽。肽遵循特征性I类HLA长度分布:长度8-15aa,模态长度为9(占肽的56%)。与以前的报告一致,通过MHCflurry⁹⁰预测大多数肽(中值79%)以标准500nM亲和力和阈值结合至少一个患者HLA等位基因,但样品之间存在很大差异(例如,一个样品中33%的

肽具有预测亲和力 $>500\text{nM}$ 。阈值为 50nM 的常用¹⁰¹“强结合剂”捕获中值为仅42%的呈递肽。转录组测序平均每个样品产生131M独特读段,并且至少一个样品中68%的基因以每百万至少1个转录本(TPM)的水平表达,突显了大型多样的样品组对于观察最大数量的基因表达的价值。HLA的肽呈递与mRNA表达密切相关。观察到肽呈递速率的显著且可重现的基因间差异,超出了仅通过RNA表达或序列差异可以解释的。观察到的HLA类型符合主要来自欧洲血统患者组的试样预期。

[0702] 对于每位患者,阳性标记的数据点是通过质谱检测到的肽,并且阴性标记的数据点是来自参考蛋白质组(SwissProt)的在该样品中未通过质谱检测到的肽。将数据分为训练集、验证集和测试集(方法)。训练集由来自101个样品(本研究中新描述的69个样品和先前发表的32个样品)的142,844个HLA呈递肽($\text{FDR}<\text{约}0.02$)组成。验证集(用于早期停止)由来自相同101个样品的18,004个呈递的肽组成。使用两个质谱数据集进行测试:(1)肿瘤样品测试集,其由来自5个另外的肿瘤样品(2个肺样品、2个结肠样品、1个卵巢样品)的571个呈递肽组成,这些肽不包含在训练数据中;和(2)一个单等位基因细胞系测试集,其由来自与训练数据中包含的单等位基因肽的位置相邻但有区别的2,128个来自基因组定位窗口(单元)的呈递肽组成(参见方法以获取有关训练/测试分类的更多详细信息)。

[0703] 使用这些和可公开获得的HLA肽数据^{92,98,99},训练神经网络(NN)模型以预测HLA抗原呈递。具体而言,在实施例9中,上文VIII.D部分中所讨论的泛等位基因模型使用以上数据来进行训练,以预测HLA抗原呈递。另一方面,在实施例11中,下面详细描述的特异性模型使用以上数据来进行训练,以预测HLA抗原呈递。在实施例10中,上文VIII.D部分中讨论的泛等位基因模型和下面详细描述的特异性模型都使用以上数据来进行训练,以预测HLA抗原呈递。

[0704] 特别地,在实施例10和11中,为了从肿瘤质谱数据(其中每种肽可能已经由六个HLA等位基因中的任何一个呈递)中学习等位基因特异性模型,开发了能够共同学习等位基因肽图谱和等位基因特异性呈递基序的新型网络架构(见下文XVII.B部分)。训练数据鉴别了53个HLA等位基因的预测模型。与先前的工作^{92,104}相反,这些模型捕获HLA呈递对多种长度的肽的每个序列位置的依赖性。该模型还正确地学习了对基因RNA表达和基因特异性呈递倾向的关键依赖性,其中mRNA丰度和学习的独立等位基因呈递倾向独立地组合,以产生最低表达最小呈递倾向和最高表达最大呈递倾向基因之间的呈递率的高达约60倍的差异。进一步观察到,即使在控制预测的结合亲和力之后(对于10个测试的等位基因中的8个, $p<0.05$),模型预测的HLA/肽复合物的测量的稳定性为IEDB⁸⁸(对于10个等位基因, $p<1\text{e}-10$)。这些特征共同构成了改进的免疫原性I类HLA肽预测的基础。

[0705] XIII. 实施例9: 包括呈递热点建模在内的实验结果

[0706] 为了具体评价在HLA呈递建模中使用呈递热点参数的益处,将并入呈递热点参数的泛等位基因神经网络呈递模型的性能与未并入呈递热点参数的泛等位基因神经网络呈递模型的性能进行了比较。这两个泛等位基因模型的基本神经网络架构相同,并且与上面在VII-VIII部分中描述的泛等位基因呈递模型相同。简言之,该泛等位基因模型包括肽和侧接氨基酸序列参数、RNA测序转录数据(TPM)、蛋白质家族数据、每样品鉴别以及HLA-A、B、C类型。每个泛等位基因模型使用5个网络的集合体。包含呈递热点参数的泛等位基因模型使用上面在VIII.B.3.部分中描述的等式12b,独立基因蛋白质组的单元大小为10,肽长度

为8-12。

[0707] 通过使用上面在XII部分中描述的质谱数据集进行实验,比较了两种泛等位基因模型。具体来说,为了公平评价竞争模型,从模型训练和验证中留出了5个样品。其余样品随机分为用于模型训练的90%和用于验证训练的10%。

[0708] 图24比较了当在五个留出测试样品上测试泛等位基因模型时,使用呈递热点参数的泛等位基因呈递模型和不使用呈递热点参数的泛等位基因呈递模型在40%召回率下的阳性预测值(PPV)。如图24所示,并入呈递热点参数的泛等位基因模型的性能始终优于未并入呈递热点参数的泛等位基因模型。

[0709] XIV. 实施例10:回顾性新抗原T细胞数据的模型评价

[0710] 我们然后评价了泛等位基因模型的HLA肽呈递的准确预测是否可以转化为鉴别人类肿瘤CD8T细胞表位(即免疫疗法靶标)的能力。为该评估定义合适的测试数据集具有挑战性,因为它需要既可被T细胞识别又可被HLA呈递在肿瘤细胞表面上的肽。另外,正式的性能评估不仅需要阳性标记的(即,T细胞识别的)肽,而且还需要足够数量的阴性标记的(即,经测试但未被识别的)肽。质谱数据集解决肿瘤呈递,但不能解决T细胞识别;相反地,引发或疫苗接种后的T细胞测定解决了T细胞前体的存在以及T细胞识别,但是不能解决肿瘤呈递(例如,源基因在肿瘤中的表达水平太低以致于不能支持该肽的呈递的强结合肽可能会在疫苗施用后引起强烈的CD8T细胞应答,但不是治疗上有用的靶标,因为它不能被肿瘤呈递)。

[0711] 为了获得合适的数据集,我们收集了4项符合要求标准的近期研究中已公开的CD8 T细胞表位:研究A¹⁴⁰在自体DC中使用串联小基因(TM)方法检查了9位患有胃肠道肿瘤的患者TIL,并且报道了通过IFN- γ ELISPOT测试的12/1,053个体细胞SNV突变的T细胞识别。研究B⁸⁴也使用TM,并报道了来自4名黑色素瘤患者的CD8+PD-1+循环淋巴细胞对6/574SNV的T细胞识别。研究C¹⁴¹使用冲击肽刺激评估了来自3名黑色素瘤患者的TIL,并且发现对5/381测试的SNV突变的响应。研究D¹⁰⁸使用TM测定和利用最小表位肽的冲击的组合评估了来自一名乳腺癌患者的TIL,并报道了对2/62SNV的识别。被组合的数据集由来自17位患者的2,023个测定的SNV(包括26个具有预先存在的T细胞应答的TSNA)组成。重要的是,由于该数据集主要包含肿瘤浸润淋巴细胞对新抗原的识别,因此成功的预测暗示着不仅能够鉴别如文献^{81, 82, 141}中的能够引发T细胞的新抗原,而且(更严格地说)能够鉴别肿瘤向T细胞呈递的新抗原。

[0712] 我们使用标准的HLA结合亲和力预测,以>2TPM阈值在如通过RNA-seq、VIII.B部分中描述的等位基因特异性神经网络模型和VIII.D部分中描述的泛等位基因神经网络模型测定的基因表达上,按照呈递概率的顺序对突变进行了排名。因为抗原特异性免疫疗法的能力在作为靶的特异性的数量上受到限制(例如,当前的个性化疫苗编码约10-20个突变^{6, 81, 82}),因此我们通过对每个患者的前5位、前10位或前20位突变中的预先存在的T细胞应答的数目进行计数来比较预测方法。这些结果描绘在图25A中。具体而言,图25A比较了对于包含12个不同测试样品(每个测试样品均取自具有至少一种预先存在的T细胞应答的患者)的测试集而言,在排名前5位、前10位和前20位的体细胞突变中由T细胞(例如,预先存在的T细胞应答)识别的体细胞突变所占的比例,所述排名前5位、前10位和前20位的体细胞突变是使用标准HLA结合亲和力预测以>2TPM阈值对如通过RNA-seq、等位基因特异性神经网络

络模型和泛等位基因神经网络模型测定的基因表达鉴别出的。

[0713] 如所预期的,结合亲和力预测在优先的突变当中仅包括少数预先存在的T细胞应答,例如在前20位当中有9/26 (35%)的预先存在的T细胞应答。相比之下,大多数(19/26, 73%)预先存在的T细胞应答排在由等位基因特异性模型和泛等位基因NN模型鉴别的前20位中(图25A)。这些结果证实了泛等位基因模型能够以与等位基因特异性模型相当的准确率(统计学上无显著性)鉴别人类肿瘤CD8T细胞表位。

[0714] 然后,我们在最小新表位水平下评价了突变(即,识别出与突变重叠的8-11聚体),这对于鉴别用于细胞疗法的T细胞/TCR可能是有用的。换句话说,使用标准HLA结合亲和力预测,以>2TPM阈值在如通过RNA-seq、VIII.B部分中描述的等位基因特异性神经网络模型和VIII.D部分中描述的泛等位基因神经网络模型鉴别的基因表达上,按呈递概率的顺序对最小新表位进行排序。如上所提到的,由于抗原特异性免疫疗法在作为靶的特异性的数目上受到技术限制,因此通过对具有至少一种预先存在的T细胞应答的每位患者的排名前5位、前10位或前20位最小新表位中预先存在的T细胞应答的数目进行计数来比较预测方法。阳性标记的表位是通过基于肽的方法(而不是基于TMG的测定法,或者外加基于TMG的测定法)被确认为具有免疫原性最小表位的那些,而阴性实例是在基于肽的测定法中未被识别的所有表位,以及包含在未被识别的微小基因中的所有跨越突变的表位(mutation-spanning epitope)。结果描绘在图25B中。

[0715] 具体而言,图25B比较了对于包含12个不同测试样品(每个测试样品均取自具有至少一种预先存在的T细胞应答的患者)的测试集而言,在排名前5位、前10位和前20位的最小新表位中由T细胞(例如,预先存在的T细胞应答)识别的最小新表位所占的比例,所述排名前5位、前10位和前20位的最小新表位是使用标准HLA结合亲和力预测以>2TPM阈值对如通过RNA-seq、等位基因特异性神经网络模型和泛等位基因神经网络模型测定的基因表达鉴别出的。

[0716] 如图25B所示,当在最小表位水平下评价突变时,泛等位基因模型继续表现得与等位基因特异性模型相当。

[0717] XIV.A. 数据

[0718] 我们从Gros等人⁸⁴、Tran等人¹⁴⁰、Stronen等人¹⁴¹和Zacharakis等人的补充信息中获得了突变调用、HLA类型和T细胞识别数据。患者特异性RNA-seq数据是无法获得的。推测肿瘤RNA的表达在具有相同肿瘤类型的不同患者之间是相关的,因此替换了TCGA中的肿瘤类型匹配的患者的RNA-seq数据,将其用于神经网络预测以及结合亲和力预测之前的RNA表达过滤。肿瘤类型匹配的RNA-seq数据的添加改善了预测性能。

[0719] 对于突变水平分析(图25A),Gros等人、Tran等人和Zacharakis等人的阳性标记的数据点是在TMG测定或最小表位肽冲击测定中被患者T细胞识别的突变。阴性标记的数据点是在TMG测定中测试的所有其它突变。对于Stronen等人,阳性标记的突变是被至少一个识别肽跨越的突变,并且阴性数据点是经过测试但是在四聚体测定中未被识别的所有突变。对于Gros、Tran和Zacharakis数据,通过对所有跨突变肽的呈递概率求和或采用最小结合亲和力对突变进行排名,因为突变的25聚体TMG测定测试了所有跨突变肽的T细胞识别。对于Stronen数据,通过对在四聚体测定中测试的所有跨突变肽的呈递概率求和或采用最小结合亲和力来对突变进行排名。

[0720] 对于表位水平分析,阳性标记的数据点是在肽冲击或四聚体测定中被患者T细胞识别的所有最小表位,并且阴性数据点是在肽冲击或四聚体测定中未被T细胞识别的所有最小表位以及来自测试的TMG的未被患者T细胞识别的所有跨突变肽。在Gros等人,Tran等人和Zacharakis等人的情况下,从分析中除去了未通过肽冲击测定进行测试的在TMG分析中识别的跨突变的最小表位肽,因为这些肽的T细胞识别状态尚未通过实验确定。

[0721] XV. 实施例11:在癌症患者中鉴别新抗原反应性T细胞

[0722] 该实施例证明改进的预测可以从常规患者样品中鉴别新抗原。为此,分析了接受抗PD(L)1疗法的9位转移性NSCLC患者的存档FFPE肿瘤活检和5-30ml外周血(补充表1:图26A-C中研究的N=9位患者的患者人口统计学和治疗信息。关键字段包括肿瘤分期和亚型、接受的抗PD1疗法以及NGS结果的概述)。肿瘤全外显子组测序、肿瘤转录组测序和匹配的正常外显子组测序产生每位患者平均198个体细胞突变(SNV和短插入缺失),其中平均表达118个(方法,补充表1)。应用完全MS模型对每位患者确定20个新表位的优先级,以针对预先存在的抗肿瘤T细胞应答进行测试。为了将分析重点放在可能的CD8应答上,将优先肽合成为8-11聚体最小表位(方法),然后在短的体外刺激(IVS)培养中用合成的肽培养外周血单核细胞(PBMC)以扩增新抗原反应性T细胞(补充表2)。两周后,使用IFN- γ ELISpot针对优先的新表位评估了抗原特异性T细胞的存在。在有足够PBMC可用的7位患者中,还进行了单独的实验以对所识别的特异性抗原进行完全或部分去卷积。结果在图26A-C和27A-30中示出。

[0723] 图26A描绘了对于9位患者的患者特异性新抗原肽库的T细胞应答的检测。对于每位患者,根据模型排名和任何序列同源性,将预测的新抗原组合成10个肽的2个库(同源肽分到不同的库中)。然后,对于每位患者,在IFN- γ ELISpot中用2个患者特异性新抗原肽库刺激该患者的体外扩增PBMC。图26A中的数据表示为减去了背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的斑点形成单位(SFU)。背景测量(DMSO阴性对照)显示在图30中。对于患者1-038-001、1-050-001、1-001-002、CU04、1-024-001、1-024-002和CU05,显示了针对同源肽库#1和#2的单孔(患者1-038-001、CU02、CU03和1-050-001)或具有均值和标准差的重复(所有其它患者)的应答。对于患者CU02和CU03,仅允许针对特定肽库#1测试细胞数。值比背景增加>2倍的样品被认为是阳性的,并标有星号(响应性供体包括患者1-038-001、CU04、1-024-001、1-024-002和CU02)。非响应性供体包括患者1-050-001、1-001-002、CU05和CU03。图26C描绘了具有来自患者CU04的体外扩增的PBMC的ELISpot孔的照片,其在IFN- γ ELISpot中用DMSO阴性对照、PHA阳性对照、CU04特异性新抗原肽库#1、CU04特异性肽1、CU04特异性肽6和CU04特异性肽8进行刺激。

[0724] 图27A-B描绘了来自利用HLA匹配的健康供体中的患者新抗原的对照实验的结果。这些实验的结果证实了体外培养条件仅扩增预先存在的体内引发的记忆T细胞,而不能够在体外从头引发。

[0725] 图28描绘了对于图26A中描绘的每个供体和每个体外扩增,对PHA阳性对照的T细胞应答的检测。对于图26A中的每个供体和每个体外扩增,用PHA刺激体外扩增的患者PBMC以使T细胞激活最大化。图28中的数据表示为减去了背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的斑点形成单位(SFU)。对于患者1-038-001、1-050-001、1-001-002、CU04、1-024-001、1-024-002、CU05和CU03,显示了单孔或生物学重复的应答。未对患者CU02进行PHA测试。在分析中包括来自患者CU02的细胞,因为针对肽库#1的阳性应答(图26A)表明有活力和

功能性T细胞。如图26A所示,对肽库有响应的供体包括患者1-038-001、CU04、1-024-001和1-024-002。同样如图26A所示,对肽库无响应的供体包括患者1-050-001、1-001-002、CU05和CU03。

[0726] 图29A描绘了患者CU04对库#2中的每个个体患者特异性新抗原肽的T细胞应答的检测。图29A还描绘了患者CU04对PHA阳性对照的T细胞应答的检测。(该阳性对照数据也显示在图28中。)对于患者CU04,用来自患者CU04的库#2的患者特异性个体新抗原肽在IFN- γ ELISpot中刺激患者体外扩增的PBMC。还用PHA作为阳性对照在IFN- γ ELISpot中刺激患者的体外扩增的PBMC。数据表示为减去背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的斑点形成单位(SFU)。

[0727] 图29B描绘了对于患者CU04的三次访问中的每一次以及对于患者1-024-002的两次访问中的每一次,对个体患者特异性新抗原肽的T细胞应答的检测,每次访问发生在不同的时间点。对于这两位患者,使用患者特异性个体新抗原肽在IFN- γ ELISpot中刺激患者的体外扩增PBMC。对于每位患者,每次访问的数据表示为减去背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的累积(增加的)斑点形成单位(SFU)。患者CU04的数据显示为3次访问的减去背景的累积SFU。对于患者CU04,显示了初次访问(T0)以及初次访问(T0)之后2个月(T0+2个月)和14个月(T0+14个月)的后续访问的减去背景的SFU。患者1-024-002的数据显示为2次访问的减去背景的累积SFU。对于患者1-024-002,显示了初次访问(T0)以及初次访问(T0)之后1个月(T0+1个月)的后续访问的减去背景的SFU。值比背景增加>2倍的样品被认为是阳性的,并标有星号。

[0728] 图29C描绘了对于患者CU04的两次访问中的每一次以及对于患者1-024-002的两次访问中的每一次,对个体患者特异性新抗原肽和患者特异性新抗原肽库的T细胞应答的检测,每次访问发生在不同的时间点。对于两位患者,使用患者特异性个体新抗原肽以及患者特异性新抗原肽库在IFN- γ ELISpot中刺激患者体外扩增的PBMC。具体而言,对于患者CU04,用CU04特异性个体新抗原肽6和8以及CU04特异性新抗原肽库在IFN- γ ELISpot中刺激患者CU04的体外扩增PBMC,并且对于患者1-024-002,用1-024-002特异性个体新抗原肽16以及1-024-002特异性新抗原肽库在IFN- γ ELISpot中刺激患者1-024-002的体外扩增PBMC。图29C的数据表示为对于具有平均值和范围的每个技术重复的减去背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的斑点形成单位(SFU)。患者CU04的数据显示为2次访问的减去背景的SFU。对于患者CU04显示了初次访问(T0;技术一式三份)以及初次访问(T0)之后2个月(T0+2个月;技术一式三份)的后续访问的减去背景的SFU。患者1-024-002的数据显示为2次访问的减去背景的SFU。对于患者1-024-002显示了初次访问(T0;技术一式三份)以及初次访问(T0)之后1个月(T0+1个月;技术重复,除了用患者1-024-002特异性新抗原肽库刺激的样品外)的后续访问的减去背景的SFU。

[0729] 图30描绘了对于图26A的患者,对两个患者特异性新抗原肽库和DMSO阴性对照的T细胞应答的检测。对于每位患者,在IFN- γ ELISpot中用两个患者特异性新抗原肽库刺激患者的体外扩增PBMC。对于每个供体和每个体外扩增,还用DMSO作为阴性对照在IFN- γ ELISpot中刺激体外扩增的患者PBMC。图30中的数据表示为对患者特异性新抗原肽库和相应DMSO对照的包括背景(相应的DMSO阴性对照)的每 10^5 个铺板细胞的斑点形成单位(SFU)。对于患者1-038-001、1-050-001、1-001-002、CU04、1-024-001、1-024-002和CU05,显示了针

对同源肽库#1和#2的单孔(1-038-001、CU02、CU03和1-050-001)或生物学重复的具有标准偏差的平均值(所有其它样品)的应答。对于患者CU02和CU03,仅允许针对特定肽库#1测试细胞数。值比背景增加>2倍的样品被认为是阳性的,并标有星号(响应性供体包括患者1-038-001、CU04、1-024-001、1-024-002和CU02)。非响应性供体包括患者1-050-001、1-001-002、CU05和CU03。

[0730] 如以上关于图27A-B所简要讨论的,为了证实了体外培养条件仅扩增预先存在的体内引发的记忆T细胞,而不能够在体外从头引发,在HLA匹配的健康供体中用新抗原进行了一系列对照实验。这些实验的结果示出在图27A-B和补充表4中。这些实验的结果证实了使用IVS培养技术在健康供体中不存在从头引发和不存在可检测到的新抗原特异性T细胞应答。

[0731] 相比之下,使用IFN- γ ELISpot在用患者特异性肽库(图26A和29-30)测试的大多数患者(5/9,56%)中鉴别出预先存在的新抗原反应性T细胞。在其细胞数允许完全或部分测试个体新抗原同源肽的7位患者中,有4位患者对至少一种所测的新抗原肽有响应,并且所有这些患者均具有相应的库响应(图26B)。其余3位用个体新抗原测试的患者(患者1-001-002、1-050-001和CU05)对单个肽没有可检测到的响应(数据未显示),证实了这些患者缺乏对新抗原库的响应(图26A)。在这4位响应患者中,对于2位具有响应的患者(患者1-024-001和1-038-001)可获得来自单次访问的样品,而对于另外2位具有响应的患者(CU04和1-024-002)可获得来自多次访问的样品。对于具有来自多次访问的样品的2位患者,来自3次访问(患者CU04)或2次访问(患者1-024-002)的累积(增加的)斑点形成单位(SFU)示出在图26B中,并且通过图29B中的访问分解。对于患者1-024-002和CU04,来自相同访问的另外的PBMC样品也是可获得的,并且重复IVS培养和ELISpot证实了对患者特异性新抗原的应答(图29C)。

[0732] 总的来说,在如通过图26A中的10个肽的库的应答所示鉴别了其至少一种T细胞识别表位的患者中,识别的新表位的数目平均为每位患者至少2个(在5位患者中鉴别了最少10个表位,对不能去卷积为1个识别肽的识别库计数)。除了通过ELISpot测试IFN- γ 应答外,还通过ELISA测试培养上清液中的颗粒酶B,并且通过MSD细胞因子多重测定测试TNF- α 、IL-2和IL-5。来自5位具有阳性ELISpot的患者中的4位的细胞分泌3种或更多种分析物,包括颗粒酶B(补充表3),表明新抗原特异性T细胞具有多功能性。重要的是,由于组合预测和IVS方法不依赖于可用的MHC多聚体的有限集合,因此对限制性HLA等位基因进行了广泛的应答测试。此外,与鉴别已识别的突变并且需要单独的去卷积步骤以识别最小表位的串联小基因筛选相反,该方法直接识别最小表位。总的来说,新抗原鉴别的产量与先前最好的方法⁹⁶相当,后者利用单采血液分离样品测试针对所有突变的TIL,同时利用常规5-30mL全血仅筛选20种合成肽。

[0733] XV.A. 肽

[0734] 定制的重组冻干肽购自JPT Peptide Technologies (Berlin, Germany)或Genscript (Piscataway, NJ, USA),并以10-50mM的浓度在无菌DMSO (VWR International, Pittsburgh, PA, USA)中重构,等分并储存在-80℃。

[0735] XV.B. 人外周血单个核细胞(PBMC)

[0736] 来自健康供体的冷冻保存的HLA型PBMC(已确认的HIV、HCV和HBV血清阴性)购自

Precision for Medicine(Gladstone,NJ,USA)或Cellular Technology,Ltd.(Cleveland, OH,USA),并储存在液氮中直到使用。新鲜血液样品购自Research Blood Components (Boston,MA,USA),leukopak购自AllCells (Boston,MA,USA),并且在冷冻保存之前通过Ficoll-Paque密度梯度法分离PBM(C GE HealthcareBio,Marlborough,MA,USA)。根据当地临床标准操作程序(SOP)和IRB批准的方案,在当地临床处理中心对患者PBM进行了处理。批准IRB是Quorum Review IRB、Comitato Etico Interaziendale A.O.U.,San Luigi Gonzaga di Orbassano和Comité Ético de la Investigación del Grupo Hospitalario Quirón en Barcelona。

[0737] 简单地说,通过密度梯度离心分离PBM,洗涤,计数并以 5×10^6 个细胞/ml的密度冷冻保存在CryoStorCS10(STEMCELL Technologies,Vancouver,BC,V6A 1B6,Canada)中。将冷冻保存的细胞在cryoport中运输,并且在到达后转移到LN₂中储存。补充表1中列出了患者的人口统计学资料。将冷冻保存的细胞解冻,并在含有Benzonase(EMD Millipore, Billerica,MA,USA)的OpTmizerT细胞扩增基础培养基(Gibco,Gaithersburg,MD,USA)中洗涤两次,并在无Benzonase的情况下洗涤一次。使用GuavaViaCount试剂和Guava easyCyte HT细胞计数器(EMD Millipore)上的模块评估细胞计数和活力。随后将细胞以适合进行测定的浓度和培养基进行重悬(参见下一部分)。

[0738] XV.C. 体外刺激(IVS)培养

[0739] 以与Ott等人⁸¹相似的方法,在同源肽和IL-2存在下扩增来自健康供体或患者样品的预先存在的T细胞。简单地说,将解冻的PBM放置过夜,并且在24孔组织培养板中在含10IU/ml rhIL-2(R&D Systems Inc.,Minneapolis,MN)的ImmunoCult™-XF T细胞扩增培养基(STEMCELL Technologies)中在肽库(每种肽10μM,每个库10种肽)的存在下刺激14天。以 2×10^6 个细胞/孔接种细胞,并且每2-3天通过更换2/3的培养基来补料。一个患者样品显示出偏离方案,并且应视为潜在的假阴性:患者CU03在解冻后未产生足够数量的细胞,并且以每个肽库 2×10^5 个细胞接种细胞(比每个方案少10倍)。

[0740] XV.D. IFN γ 酶联免疫斑点(ELISpot)测定

[0741] 通过ELISpot测定¹⁴²进行产生IFN γ 的T细胞的检测。简单地说,收获PBM(离体或体外扩增),在无血清RPMI(VWR International)中洗涤,并在用抗人IFN γ 捕获抗体(Mabtech,Cincinnati,OH,USA)包被的ELISpot Multiscreen板(EMD Millipore)中在OpTmizer T细胞扩增基础培养基(离体)或ImmunoCult™-XF T细胞扩增培养基(扩增培养物)中在对照或同源肽的存在下培养。在5%CO₂、37℃的潮湿培养箱中孵育18小时后,将细胞从板中移出,并使用抗人IFN γ 检测抗体(Mabtech)、Vectastain Avidin过氧化物酶复合物(VectorLabs,Burlingame,CA,USA)和AEC底物(BD Biosciences,San Jose,CA,USA)检测膜结合的IFN γ。使ELISpot板干燥,避光保存,然后发送至Zellnet Consulting,Inc., FortLee,NJ,USA)以进行标准化评估¹⁴³。数据表示为每铺板细胞数的斑点形成单位(SFU)。

[0742] XV.E. 颗粒酶B ELISA和MSD多重测定

[0743] 使用3重测定MSD U-PLEX Biomarker测定(目录号K15067L-2)进行ELISpot上清液中分泌的IL-2、IL-5和TNF-α的检测。根据制造商的说明进行测定。对于每种细胞因子,使用已知标准品的系列稀释液计算分析物浓度(pg/ml)。对于图形数据表示,低于标准曲线最小范围的值表示为零。根据制造商的说明,使用Granzyme BDuoSet®ELISA(R&D Systems,

Minneapolis, MN) 进行ELISpot上清液中颗粒酶B的检测。简单地说,将ELISpot上清液在样品稀释液中以1:4稀释,并与颗粒酶B标准品的系列稀释液一起运行以计算浓度(pg/ml)。对于图形数据表示,低于标准曲线最小范围的值表示为零。

[0744] XV.F. IVS测定的阴性对照实验-在健康供体中测试的来自肿瘤细胞系的新抗原

[0745] 图27A示出了用于在健康供体中测试的来自肿瘤细胞系的新抗原的IVS测定的阴性对照实验。在IVS培养中,用含有阳性对照肽(先前暴露于感染性疾病)、源自肿瘤细胞系的HLA匹配的新抗原(未暴露)和源自所述供体为血清阴性的病原体的肽的肽库刺激健康供体PBMC。在用DMSO(阴性对照,黑色圆圈)、PHA和常见感染性疾病多肽(阳性对照,红色圆圈)、新抗原(未暴露,浅蓝色圆圈)或HIV和HCV肽(已确认供体是血清阴性的,藏青色,A和B)刺激之后,随后通过IFN γ ELISpot (10^5 个细胞/孔)分析扩增的细胞。数据显示为每 10^5 个接种细胞的斑点形成单位(SFU)。显示了具有均值和SEM的生物学重复。没有观察到对新抗原或对源自供体未暴露的病原体(血清阴性)的肽的应答。

[0746] XV.G. IVS测定的阴性对照实验-在健康供体中测试的来自患者的新抗原

[0747] 图27A示出了用于在健康供体中测试反应性的来自患者的新抗原的IVS测定的阴性对照实验。评估健康供体中对HLA匹配的新抗原肽库的T细胞应答。左图:在离体IFN- γ ELISpot中用对照(DMSO、CEF和PHA)或HLA匹配的患者来源的新抗原肽刺激健康供体PBMC。数据表示为一式三份孔的每 2×10^5 个铺板细胞的斑点形成单位(SFU)。右图:在IFN- γ ELISpot中用对照(DMSO、CEF和PHA)或HLA匹配的患者来源的新抗原肽库刺激在新抗原库或CEF库的存在下扩增的IVS培养后的健康供体PBMC。数据表示为一式三份孔的每 1×10^5 个铺板细胞的SFU。在健康的供体中未见对新抗原的应答。

[0748] XV.H. 补充表2: NSCLC患者中测试T细胞识别的肽

[0749] 在图26A-C中研究了对N=9位患者测试的新抗原肽的细节(来自NSCLC患者的新抗原反应性T细胞的鉴别)。关键字段包括来源突变、肽序列、库和所观察到的个体肽响应。“最可能的限制(most_probable_restriction)”列表示模型预测的哪个等位基因最有可能呈递每种肽。还包括通过结合亲和力预测(方法)计算的每位患者的所有突变肽中这些肽的排名。

[0750] 有四种肽通过完全MS模型排名很高并且被CD8T细胞识别,其通过结合亲和力预测具有低的预测结合亲和力或排名较低。

[0751] 对于这些肽中的三种,这是由该模型和MHCflurry 1.2.0之间的HLA覆盖率差异引起的。肽YEHEDVKEA被HLA-B*49:01预测为呈递的,其未被MHCflurry 1.2.0覆盖。类似地,肽SSAAAPFPL和FVSTSDIKSM被HLA-C*03:04预测为呈递的,其也未被MHCflurry 1.2.0覆盖。在线NetMHCpan4.0 (BA) 预测器(一种原则上覆盖所有等位基因的泛特异性结合亲和力预测器)将SSAAAPFPL排名为HLA-C*03:04的强结合剂(23.2nM,对于患者1-024-002排名第二),预测了FVSTSDIKSM与HLA-C*03:04的弱结合(943.4nM,对于患者1-024-002排名第39)以及YEHEDVKEA与HLA-B*49:01的弱结合(3387.8nM),但是与HLA-B*41:01的更强结合(208.9nM,对于患者1-038-001排名第11),其也存在于该患者中,但未被模型覆盖。因此,在这三种肽中,FVSTSDIKSM被结合亲和力预测错过了,SSAAAPFPL已被捕获并且YEHEDVKEA的HLA限制尚不确定。

[0752] 对其发展了肽特异性T细胞应答的其余五种肽来自这样的患者,其中通过模型确

定的最可能的呈递等位基因也被MHCflurry1.2.0所覆盖。在这五种肽中,4/5的预测结合亲和力和强于标准500nM阈值,并排名前20,尽管其排名比模型中的排名稍低(肽DENITTIQF、QDVSVQVER、EVADAATLTM、DTVEYPYTSF被模型分别排名为0、4、5、7,对比于被MHCflurry排名为2、14、7和9)。肽GTKKDVDVLK被CD8 T细胞识别并且通过模型排名第1,但通过MHCflurry排名第70,且预测的结合亲和力为2169nM。

[0753] 总体上,通过完全MS模型排名高的个体识别肽中的6/8使用结合亲和力预测也排名高,并且预测的结合亲和力<500nM,而如果使用结合亲和力预测而不是完全MS模型,则个体识别肽中的2/8将被错过。

[0754] XV.I. 补充表3:对来自NSCLC新抗原肽的ELISpot上清液的MSD细胞因子多重和ELISA测定

[0755] 显示了从阳性ELISpot (IFN γ) 孔的上清液中检测到的分析物颗粒酶B (ELISA)、TNF α 、IL-2和IL-5 (MSD)。值显示为来自技术重复的平均pg/ml。阳性值以斜体显示。颗粒酶BELISA:值比DMSO背景高 ≥ 1.5 倍被认为是阳性。U-PlexMSD测定:值比DMSO背景高 ≥ 1.5 倍被认为是阳性。

[0756] XV.J. 补充表4:IVS对照实验中的新抗原和感染性疾病抗原表位

[0757] 在图27A-B中示出了IVS对照实验中测试的肿瘤细胞系新抗原和病毒肽的细节。关键字段包括源细胞系或病毒、肽序列以及预测的呈递HLA等位基因。

[0758] XV.K. 数据

[0759] 用于训练和测试预测模型的MS肽数据集(图25A-B)可从MassIVE档案馆(massive.ucsd.edu)获得,登录号MSV000082648。文件中包含通过ELISpot测试的新抗原肽(图26A-C和27A-B)(补充表2和4)。

[0760] XVI. 实施例8-11的方法

[0761] XVI.A. 质谱

[0762] XVI.A.1. 试样

[0763] 用于质谱分析的存档冷冻组织试样获自商业来源,包括BioServe (Beltsville, MD)、ProteoGenex (Culver City, CA)、iSpecimen (Lexington, MA) 和Indivumed (Hamburg, Germany)。根据Comité de Protection des Personnes, Ile-de-France VII批准的研究方案下,还从Hopital Marie Lannelongue (Le Plessis-Robinson, France) 的患者中前瞻性地收集了一部分试样。

[0764] XVI.A.2. HLA免疫沉淀

[0765] 在组织样品裂解和溶解之后,使用建立的免疫沉淀(IP)方法进行HLA-肽分子的分离^{87,124-126}。将新鲜冷冻组织粉碎(CryoPrep; Covaris, Woburn, MA),加入裂解缓冲液(1% CHAPS, 20mM Tris-HCl, 150mM NaCl, 蛋白酶和磷酸酶抑制剂, pH=8)以溶解组织,并将所得溶液在4℃下离心2小时以沉淀碎屑。将澄清的裂解物用于HLA特异性IP。如先前所描述的使用抗体W6/32进行免疫沉淀¹²⁷。将裂解物添加至抗体珠粒,并在4℃旋转过夜以进行免疫沉淀。免疫沉淀后,将珠粒从裂解物中除去。洗涤IP珠粒以除去非特异性结合,并用2N乙酸从珠粒上洗脱HLA/肽复合物。使用分子量旋转柱从肽中去除蛋白质组分。将所得肽通过SpeedVac蒸发至干,并在MS分析之前保存在-20℃。

[0766] XVI.A.3. 肽测序

[0767] 将干燥的肽在HPLC缓冲液A中复原,并加载到C-18微毛细管HPLC柱上,以梯度洗脱到质谱仪中。使用180分钟内0-40%B(溶剂A-0.1%甲酸,溶剂B-80%乙腈的0.1%甲酸)的梯度将肽洗脱到Fusion Lumos质谱仪(Thermo)中。在所选离子的HCD裂解后,在Orbitrap检测器中以120,000的分辨率收集肽质量/电荷(m/z)的MS1图谱,然后在Orbitrap或离子阱检测器中收集了20个MS2低分辨率扫描。MS2离子的选择使用依赖数据的采集模式进行,并且在离子的MS2选择之后30秒进行动态排除。MS1扫描的自动增益控制(AGC)设置为 4×10^5 ,并且MS2扫描的设置设置为 1×10^4 。对于测序HLA肽,可以选择+1、+2和+3电荷状态用于MS2片段化。

[0768] 使用Comet^{128,129}针对蛋白质数据库搜索每个分析的MS2图谱,并使用Percolator¹³⁰⁻¹³²对肽鉴定进行评分。

[0769] XVI.B.机器学习

[0770] XVI.B.1.数据编码

[0771] 对于每个样品,训练数据点是来自参考蛋白质组的所有8-11聚体(含)肽,这些肽正确映射到样品中表达的一个基因。通过将每个训练样品的训练数据集连接形成整体训练数据集。选择长度8-11,因为该长度范围捕获了所有I类HLA呈递肽的约95%;但是,可以使用相同的方法来为模型增加长度12-15,但要以适度增加计算需求为代价。使用独热编码方案将肽和侧接序列向量化。通过使用填充字符扩展氨基酸字母并将所有肽填充到最大长度11来将多种长度(8-11)的肽表示为固定长度的向量。训练肽的源蛋白的RNA丰度表示为从RSEM¹³³获得的同工型水平每百万转录物(TPM)估计值的对数。对于每个肽,将每个肽的TPM计算为对于包含肽的每个同工型的每同工型TPM估计值的总和。训练数据中排除了来自以0TPM表达的基因的肽,并且在测试时,为未表达基因的肽的分配0的呈递概率,最后,为每个肽分配Ensembl蛋白质家族ID,并且每个唯一Ensembl蛋白家族ID对应于每基因呈递倾向截距(参见下一部分)

[0772] XVII.B.2.模型架构的说明

[0773] 完整呈递模型具有以下功能形式:

[0774] (等式1) $Pr(\text{呈递的肽}i) = \sum_{k=1}^m a_k^i$ 。
 $Pr(\text{由等位基因}\alpha\text{呈递的肽}i),$

[0775] 其中k索引数据集中的HLA等位基因,范围从1到m,并且 a_k^i 是指示变量,如果肽i来源的样品中存在等位基因k,则其值为1,否则为0。注意,对于给定的肽i,所有但最多6个 a_k^i (6对应于肽i来源的样品中的HLA类型)将为零。概率之和固定为 $1 - \epsilon$,例如 $\epsilon = 10^{-6}$ 。

[0776] 独立等位基因呈递概率建模如下:

[0777] $Pr(\text{由等位基因}\alpha\text{呈递的肽}i) = \text{sigmoid}\{NN_{\alpha}(\text{肽}i) +$
 $NN_{\text{侧接}}(\text{侧接}i) + NN_{RNA}(\log(TPM_i)) + \alpha_{\text{样品}(i)} + \beta_{\text{蛋白}(i)}\},$

[0778] 其中变量具有以下含义:sigmoid是sigmoid(又称expit)函数,肽_i是肽i的独热编码的中间填充氨基酸序列,NN_α是具有线性最后一层激活的神经网络,其模拟了肽序列对呈递概率的贡献,侧接_i是其源蛋白中肽i的独热编码的侧接序列,NN_{侧接}是具有线性最后一层激活的神经网络,其模拟了侧接序列对呈递概率的贡献,TPM_i是肽i的源mRNA的以TPM为单位的表达,样品(i)是肽i来源的样品(即患者), $\alpha_{\text{样品}(i)}$ 是每样品截距,蛋白(i)是肽i的源蛋

白,并且 $\beta_{\text{蛋白}(i)}$ 是每蛋白质截距(也就是每基因的呈递倾向)。

[0779] 对于结果部分中描述的模型,组件神经网络具有以下架构:

[0780] • 每个 NN_{α} 是单隐藏层多层感知器(MLP)的一个输出节点,具有输入维度231(11个残基 $\times$ 每个残基21种可能的字符(包括填充字符)),宽度256,隐藏层中的已校正的线性单元(ReLU)激活,以及训练数据集中每HLA等位基因 α 的一个输出节点。

[0781] • $NN_{\text{侧接}}$ 是单隐藏层MLP,具有输入维度210(N末端侧接序列的5个残基+C末端侧接序列的5个残基 $\times$ 每个残基21种可能的字符(包括填充字符)),宽度32,隐藏层中的已校正的线性单元(ReLU)激活和输出层中的线性激活。

[0782] • NN_{RNA} 是单隐藏层MLP,具有输入维度1,宽度16,隐藏层中的已校正的线性单元(ReLU)激活和输出层中的线性激活。

[0783] 应注意,模型的一些组件(例如 NN_{α})取决于特定的HLA等位基因,但是许多组件($NN_{\text{侧接}}$ 、 NN_{RNA} 、 $\alpha_{\text{样品}(i)}$ 、 $\beta_{\text{蛋白}(i)}$)不是。前者称为“等位基因相互作用”,后者称为“等位基因非相互作用”。根据生物学现有技术知识选择模拟为等位基因相互作用或非相互作用的特征:HLA等位基因能识别肽,因此应将肽序列建模为等位基因相互作用,但是没有关于源蛋白、RNA表达或侧接序列的信息被传递至HLA分子(因为肽在其在内质网中遇到HLA时已从同源蛋白分离),因此这些特征应建模为等位基因非相互作用。该模型在Keras v2.0.4¹³⁴和Theano v0.9.0¹³⁵中实现。

[0784] 肽MS模型使用与完全MS模型相同的去卷积程序(等式1),但是使用仅考虑肽序列和HLA等位基因的简化的等位基因模型生成了独立等位基因呈递概率:

[0785] $\text{Pr}(\text{由等位基因}\alpha\text{呈递的肽}i) = \text{sigmoid}\{NN_{\alpha}(\text{肽}_i)\}$ 。

[0786] 肽MS模型使用与结合亲和力预测相同的特征,但是模型的权重是在不同的数据类型上进行训练的(即质谱数据相比于HLA肽结合亲和力数据)。因此,肽MS模型和完全MS模型的预测性能的比较揭示了非肽特征(即RNA丰度、侧接序列、基因ID)对总体预测性能的贡献,并且肽MS模型和结合亲和力模型的预测性能的比较揭示了改进肽序列建模对整体预测性能的重要性。

[0787] XVI.B.3. 训练/验证/测试分组

[0788] 我们使用以下程序确保没有肽出现在多于一个训练/验证/测试集中:首先从参考蛋白质组中去除出现在多于一种蛋白质中的所有肽,然后将蛋白质组划分为10个相邻的肽单元。每个单元都被唯一分配到训练、验证或测试集。这样,没有肽出现在多于一个训练、验证或测试集中。验证集仅用于提前停止。

[0789] XVI.B.4. 模型训练

[0790] 对于模型训练,将所有肽建模为独立的,其中每肽损失是负的伯努利对数似然损失函数(又称对数损失)。形式上,肽 i 对总损失的贡献为

[0791] 损失(i) = $-\log(\text{伯努利}(y_i | \text{Pr}(\text{呈递的肽}i)))$,

[0792] 其中 y_i 是肽 i 的标记;即,如果肽 i 被呈递,则 $y_i = 1$,否则为0,并且伯努利($y|p$)表示考虑i.i.d.二进制观测向量 y 的参数 $p \in [0, 1]$ 的伯努利似然性。通过使损失函数最小化来训练模型。

[0793] 为了减少训练时间,可通过随机去除90%的负标记训练数据来调整类平衡,从而产生每约2000个未呈递肽中一个呈递肽的总体训练集类平衡。将模型权重使用Glorot统一

程序61初始化,并在Nvidia Maxwell TITAN X GPU上使用具有标准参数的ADAM62随机优化器进行训练。由总数据的10%组成的验证集用于早期停止。在每个四分之一周期对验证集进行模型评价,并在验证损失(即验证集上的负伯努利对数似然)未能降低时在第一个四分之一周期后停止模型训练。

[0794] 完全呈递模型是10个模型副本的集合体,每个副本在相同训练数据的混洗副本上独立训练,其中集合体中每个模型的模型权重都有不同的随机初始化。在测试时,通过取模型副本输出的概率平均值来生成预测。

[0795] XVI.B.5. 基序徽标

[0796] 使用weblogolib Python API v3.5.0¹³⁸产生基序徽标。为了产生结合亲和力徽标,2017年7月从免疫表位数据库(IEDB⁸⁸)下载了mhc_ligand_full.csv文件,并保留了符合以下标准的肽:以纳摩尔(nM)为单位的测量,2000年后的参考日期,对象类型等于“线性肽”并且肽中的所有残基均来自规范的20个字母的氨基酸字母表。使用具有低于500nM的常规结合阈值的测量的结合亲和力的经过滤肽的子集产生徽标。对于在IEDB中具有太少结合剂的等位基因对,不产生徽标。为了产生代表学习的呈递模型的徽标,针对每个等位基因和每个肽长度预测了2,000,000个随机肽的模型预测。对于每个等位基因和每个长度,通过学习的呈递模型使用排名前1%(即前20,000)的肽产生徽标。重要的是,来自IEDB的这种结合亲和力数据未用于模型训练或测试,而仅用于比较学习的基序。

[0797] XVI.B.6. 结合亲和力预测

[0798] 我们使用来自MHCflurry v1.2.0¹³⁹的仅结合亲和力预测器预测了肽-MHC结合亲和力,MHCflurry v1.2.0¹³⁹是一种开源、GPU兼容的I类HLA结合亲和力预测器,其性能与NetMHC系列模型相当。为组合多个HLA等位基因中的单个肽的结合亲和力预测,选择最小结合亲和力。为了组合多个肽的结合亲和力(即,为了对被多个突变肽所跨越的突变进行排名,如图25A-B所示),选择了肽中的最小结合亲和力。对于T细胞数据集上的RNA表达阈值,使用了从TCGA到TPM>1阈值的肿瘤类型匹配RNA-seq数据。在原始出版物中,所有原始T细胞数据集均以TPM>0进行过滤,因此未使用要以TPM>0进行过滤的TCGA RNA-seq数据。

[0799] XVI.B.7. 呈递预测

[0800] 为了组合多个HLA等位基因的单个肽的呈递概率,如等式1中所示鉴别了概率的总和。为了组合多个肽的呈递概率(即,为了对被多个突变肽所跨越的突变进行排名,如图25A-B所示),鉴别了呈递概率的总和。概率上,如果肽的呈递被认为是i.i.d.伯努利随机变量,则概率的总和对应于所呈递的突变肽的预期数目:

$$[0801] \quad E[\text{跨越突变}i\text{的呈递的新抗原的数目}] = \sum_{j=1}^{n_i} \text{Pr}[\text{被呈递的表位}j],$$

[0802] 其中Pr[被呈递的表位j]是通过将经训练的呈递模型应用于表位j获得的, n_i 表示跨越突变i的突变表位的数目。例如,对于远离其源基因末端的SNV i,有8个跨8聚体、9个跨9聚体、10个跨10聚体和11个跨11聚体,总共 $n_i = 38$ 个跨越突变的表位。

[0803] XVI.C. 下一代测序

[0804] XVI.C.1. 试样

[0805] 对于对冷冻切除的肿瘤的转录组分析,从用于MS分析的相同组织试样(肿瘤或邻

近的正常组织)中获得RNA。对于进行抗PD1疗法的患者中的新抗原外显子组和转录组分析,从存档FFPE肿瘤活检中获得DNA和RNA。使用邻近的正常、相配的血浆或PBMC获得用于正常外显子组和HLA分型的正常DNA。

[0806] XVI.C.2.核酸提取与文库构建

[0807] 使用Qiagen DNeasy柱(Hilden,Germany)按照制造商推荐的程序分离来自血液的正常/生殖细胞DNA。使用Qiagen Allprep DNA/RNA分离试剂盒按照制造商推荐的程序分离来自组织试样的DNA和RNA。分别通过Picogreen和Ribogreen荧光(Molecular Probes)对DNA和RNA进行定量,并且将产量>50ng的试样进行文库构建。按照制造商推荐的方案,通过声学剪切(Covaris,Woburn,MA)和随后的DNA Ultra II(NEB,Beverly,MA)文库制备试剂盒产生DNA测序文库。通过热裂解和利用RNA Ultra II(NEB)的文库构建来产生肿瘤RNA测序文库。通过Picogreen(MolecularProbes)定量得到的文库。

[0808] XVI.C.3.全外显子组捕获

[0809] 使用xGEN Whole Exome Panel(Integrated DNA Technologies)对DNA和RNA测序文库进行外显子富集。将1至1.5μg正常DNA或肿瘤DNA或RNA来源的文库用作输入,并使其杂交超过12小时,然后进行链霉亲和素纯化。将捕获的文库通过PCR进行最少扩增,并通过NEBNext文库定量试剂盒(NEB)进行定量。将捕获的文库以等摩尔浓度合并,并使用c-bot(Illumina)进行聚类,并在HiSeq4000(Illumina)上以75个碱基配对的末端进行测序,以达到靶标独特的平均覆盖率>500x肿瘤外显子组,>100x正常外显子组和>100M读段肿瘤转录组。

[0810] XVI.C.4.分析

[0811] 使用BWA-MEM¹⁴⁴(v.0.7.13-r1126)将外显子组读段(FFPE肿瘤和匹配的正常)与参考人基因组(hg38)进行比对。使用STAR(v.2.5.1b)将RNA-seq读段(FFPE和冷冻的肿瘤组织样品)与基因组和GENCODE转录本(v.25)进行比对。使用RSEM¹³³(v.1.2.31)和相同的参考转录本对RNA表达进行定量。Picard(v.2.7.1)用于标记重复的比对并计算比对量度。对于用GATK¹⁴⁵(v.3.5-0)对碱基质量评分进行重新校准后的FFPE肿瘤样品,利用FreeBayes¹⁴⁶(1.0.2)使用配对肿瘤-正常外显子组确定取代和短插入缺失变体。过滤器包括等位基因频率>4%;中值碱基质量>25,支持读段的最小映射质量30和正常中的替代读段计数≤2且获得足够的覆盖率。还必须在两条链上都检测到变体。排除发生在重复区域的体细胞变体。使用RefSeq转录本用snpEff¹⁴⁷(v.4.2)进行翻译和注释。在肿瘤RNA比对中鉴别的非同义、非终止变体进入新抗原预测。Optitype¹⁴⁸1.3.1用于产生HLA类型。

[0812] XVI.C.5.图27A-B:用于IVS控制实验的肿瘤细胞系和匹配的正常细胞系

[0813] 肿瘤细胞系H128、H122、H2009、H2126、CoLo829及其正常供体匹配的对照细胞系BL128、BL2122、BL2009、BL2126和CoLo829BL均购自ATCC(Manassas,VA),根据销售者的说明生长至 10^{83} - 10^{84} 个细胞,然后速冻用于核酸提取和测序。NGS程序基本上如上所述,只是MuTect¹⁴⁹(3.1-0)仅用于取代突变检测。IVS对照测定中使用的肽列在补充表4中。

[0814] XVI.D.II类模型概念验证

[0815] 为了证明泛等位基因神经网络(NN)模型预测被II类MHC分子呈递的能力,使用人B细胞淋巴瘤样品(n=39)进行了实验。39个样品中的每一个均包含HLA-DR分子,更具体地,HLA-DRB1分子、HLA-DRB3分子、HLA-DRB4分子和/或HLA-DRB5分子。将其中四个样品留作测

试集,其余35个样品用于训练和验证。训练集由长度为9-20个氨基酸(AA)(包括端值)的20,136种呈递的肽组成,具有13个氨基酸和14个氨基酸长度的模式。验证集和测试集分别由2,279种和301种呈递的肽组成。

[0816] II类MHC泛等位基因NN模型架构与I类MHC泛等位基因NN模型架构相同,但有3个例外:(1)II类模型接受每个样品的最多4个独特HLA-DRB等位基因(而不是HLA-A、HLA-B、HLA-C的6个等位基因), (2)II类模型在更长的肽序列(9-20聚体而不是8-11聚体)上进行训练,以及(3)独立等位基因模型适合每个等位基因的不同的子网络模型,而泛等位基因模型通过使用所有等位基因的共享密集网络而在等位基因之间共享知识。将泛等位基因模型的性能与等位基因特异性NN模型进行了比较。这两种模型都在相同的肽上进行训练。这两种NN模型之间的模型输入的唯一差异是,泛等位基因模型使用34个长度的AA序列来描述HLA类型,而等位基因特异性模型则使用标准HLA命名法(例如,HLA-DRB1*01:01)。

[0817] 图31A-D显示了泛等位基因模型和等位基因特异性模型的每个测试样品的精确率-召回率曲线。具体而言,图31A描绘了泛等位基因模型和等位基因特异性模型的针对每个测试样品0的精确率-召回率曲线。图31B描绘了泛等位基因模型和等位基因特异性模型的针对每个测试样品1的精确率-召回率曲线。图31C描绘了泛等位基因模型和等位基因特异性模型的针对每个测试样品2的精确率-召回率曲线。图31D描绘了泛等位基因模型和等位基因特异性模型的针对每个测试样品4的精确率-召回率曲线。如图31A-D所示,这两种NN模型都获得相当的(统计上无显著性的)阳性预测分数,并且接收器工作特性曲线下的面积(ROCAUC)也同样相当(另见下表3和4)。这证明了在II类MHC肽呈递预测任务中,泛等位基因模型的能力能够比得上等位基因特异性模型的性能。

		在 40%召回率下的阳性预测值				
		测试样品 0	测试样品 1	测试样品 2	测试样品 3	均值
[0818]	等位基因特异性	6.4%	6.5%	1.4%	6.8%	5.3%
	泛等位基因	3.8%	10.9%	3.9%	3.2%	5.4%

[0819] 表3

		ROC 曲线下的面积				
		测试样品 0	测试样品 1	测试样品 2	测试样品 3	均值
[0820]	等位基因 特异性	0.99	0.98	0.96	0.97	0.98
	泛等位基 因	0.99	0.98	0.98	0.98	0.98

[0821] 表4

[0822] XVII. 实施例12:对来自NSCLC患者的外周血的新抗原特异性记忆T细胞的TCR进行测序

[0823] 图32描绘了对来自NSCLC患者的外周血的新抗原特异性记忆T细胞的TCR进行测序的方法。在ELISpot孵育后,收集来自NSCLC患者CU04的外周血单核细胞(PBMC)(以上关于图26A-30所述)。具体地,如上所述,用CU04特异性个体新抗原肽(图29C)、CU04特异性新抗原肽库(图29C)和DMSO阴性对照(图30)在IFN- γ ELISpot中刺激来自患者CU04的2次访问的体外扩增的PBMC。在孵育之后并且添加检测抗体之前,将PBMC转移至新的培养板中,并在完成ELISpot分析期间保持在孵育箱中。根据ELISpot结果鉴别阳性(响应性)孔。如图32所示,鉴别出的阳性孔包括用CU04特异性个体新抗原肽8刺激的孔和用CU04特异性新抗原肽库刺激的孔。合并来自这些阳性孔和阴性对照(DMSO)孔的细胞,并用磁性标记的抗体对CD137进行染色,以用于使用Miltenyi磁性分离柱进行富集。

[0824] 使用10xGenomics单细胞分辨率配对免疫TCR分析方法对如上所述分离和扩增的富含CD137和耗尽CD137的T细胞级分进行测序。具体来说,将活T细胞分配到单细胞乳剂中,以用于随后的单细胞cDNA产生和全长TCR分析(5' UTR至恒定区——确保 α 和 β 配对)。一种方法使用在转录物的5'端的分子条形码化模板转换寡核苷酸,第二种方法使用在3'端的分子条形码化恒定区寡核苷酸,并且第三种方法是将RNA聚合酶启动子与TCR的5'端或3'端偶联。所有这些方法能够在单细胞水平上进行 α 和 β TCR对的鉴别和去卷积。所得条形码化的cDNA转录本经历优化的酶和文库构建工作流程,以减少偏差并确保细胞库内克隆型的准确表示。在Illumina的MiSeq或HiSeq4000仪器(配对末端150个循环)上对文库进行测序,目标测序深度为每个细胞约五千至五万个读段。所得TCR核酸序列描述于补充表5中。补充表5中描述的TCRa和TCR β 链的存在通过基于正交锚定PCR的TCR测序方法(Archer)证实。与基于10x Genomics的TCR测序相比,此特定方法的优势在于使用有限的细胞数作为输入,并且酶操作较少。

[0825] 使用10x软件和定制生物信息学管线分析测序输出,以鉴别T细胞受体(TCR) α 和 β 链对,如补充表5所示。补充表5进一步列出了最常见的TCR克隆型的 α 和 β 可变(V)、连接(J)、恒定(C)和 β 多样性(D)区域以及CDR3氨基酸序列。克隆型定义为独特的CDR3氨基酸序列的 α 、 β 链对。针对以大于2个细胞的频率出现的单 α 和单 β 链对,对克隆型进行过滤,以产生患者CU04中每个靶标肽的克隆型的最终列表(补充表5)。

[0826] 总而言之,使用关于图32的上述方法,鉴别了来自患者CU04的外周血的记忆CD8+T细胞,所述细胞对于如上文XV部分中关于实施例11所讨论而鉴别的CU04的肿瘤新抗原是新抗原特异性的。对这些鉴别的新抗原特异性T细胞的TCR进行测序。此外,还鉴别了对于通过上述呈递模型鉴别的患者CU04的肿瘤新抗原是新抗原特异性的经测序TCR。

[0827] XVIII. 实施例13:新抗原特异性记忆T细胞在T细胞疗法中的用途

[0828] 在鉴别出对患者肿瘤呈递的新抗原具有新抗原特异性的T细胞和/或TCR之后,这些鉴别出的新抗原特异性T细胞和/或TCR可用于患者的T细胞疗法。具体地,这些鉴别出的新抗原特异性T细胞和/或TCR可用于产生治疗量的用于在T细胞疗法期间输注到患者体内的新抗原特异性T细胞。在本文XVIII.A.和XVIII.B.部分中讨论了两种用于产生治疗量的用于患者中的T细胞疗法的新抗原特异性T细胞的方法。第一种方法包括从患者样品中扩增鉴别出的新抗原特异性T细胞(XVIII.A.部分)。第二种方法包括对已鉴别的新抗原特异性T

细胞的TCR进行测序,并将经测序TCR克隆到新的T细胞中(XVIII.B.部分)。本文未明确提及的用于产生用于T细胞疗法的新抗原特异性T细胞的替代方法也可以用于产生治疗量的用于T细胞疗法的新抗原特异性T细胞。一旦通过一种或多种这些方法获得了新抗原特异性T细胞,就可以将这些新抗原特异性T细胞输注到患者体内以用于T细胞疗法。

[0829] XVIII.A.从患者样品中鉴别和扩增新抗原特异性记忆T细胞以用于T细胞疗法

[0830] 产生治疗量的用于患者中的T细胞疗法的新抗原特异性T细胞的第一种方法包括扩增从患者样品中鉴别出的新抗原特异性T细胞。

[0831] 具体地,为了将新抗原特异性T细胞扩增至用于患者的T细胞疗法中的治疗量,使用上述呈递模型鉴别最有可能由患者癌细胞呈递的新抗原肽集。另外,从患者获得包含T细胞的患者样品。患者样品可能包含患者的外周血、肿瘤浸润淋巴细胞(TIL)或淋巴结细胞。

[0832] 在其中患者样品包含患者外周血的实施方案中,可以使用以下方法将新抗原特异性T细胞扩增至治疗量。在一个实施方案中,可以进行引发。在另一个实施方案中,可以使用一种或多种上述方法鉴别已激活T细胞。在另一个实施方案中,可以进行引发和已激活T细胞的鉴别二者。引发和鉴别已激活T细胞二者的优点是使所代表的特异性的数目最大化。引发和鉴别已激活T细胞二者的缺点是这种方法是困难的和费时的。在另一个实施方案中,可以分离不一定被激活的新抗原特异性细胞。在这样的实施方案中,也可以进行这些新抗原特异性细胞的抗原特异性或非特异性扩增。在收集这些引发的T细胞之后,可以对引发的T细胞进行快速扩增方案。例如,在一些实施方案中,可以对引发的T细胞进行Rosenberg快速扩增方案

[0833] (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2978753/>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2305721/>)^{153,154}。

[0834] 在其中患者样品包含患者的TIL的实施方案中,可以使用以下方法将新抗原特异性T细胞扩增至治疗量。在一个实施方案中,新抗原特异性TIL可以在体外进行四聚体/多聚体分选,然后可以对分选的TIL进行如上所述的快速扩增方案。在另一个实施方案中,可以进行TIL的新抗原非特异性扩增,然后可以对新抗原特异性TIL进行四聚体分选,然后可以对分选的TIL进行如上所述的快速扩增方案。在另一个实施方案中,可以在使TIL经历快速扩增方案之前进行抗原特异性培养。

[0835] (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4607110/>, <https://onlinelibrary.wiley.com/doi/pdf/10.1002/eji.201545849>)^{155,156}。

[0836] 在一些实施例中,可以修改Rosenberg快速扩增方案。例如,可以将抗PD1和/或抗41BB添加到TIL培养物中以模拟更快速的扩增。

[0837] (<https://jitc.biomedcentral.com/articles/10.1186/s10425-016-0164-7>)¹⁵⁷。

[0838] XVIII.B.鉴别新抗原特异性T细胞,对鉴别的新抗原特异性T细胞的TCR进行测序并且将经测序的TCR克隆到新的T细胞中

[0839] 用于产生治疗量的用于患者中的T细胞疗法的新抗原特异性T细胞的第二种方法包括从患者样品中鉴别新抗原特异性T细胞,对鉴别的新抗原特异性T细胞的TCR进行测序,并且将经测序的TCR克隆到新的T细胞中。

[0840] 首先,从患者样品中鉴别新抗原特异性T细胞,并对鉴别的新抗原特异性T细胞的TCR进行测序。可以从其分离出T细胞的患者样品可以包含血液、淋巴结或肿瘤中的一种或

多种。更具体地,可以从其分离T细胞的患者样品可以包含外周血单核细胞(PBMC)、肿瘤浸润细胞(TIL)、离体肿瘤细胞(DTC)、体外引发的T细胞和/或分离自淋巴结的细胞中的一种或多种。这些细胞可以是新鲜的和/或冷冻的。PBMC和体外引发的T细胞可获自癌症患者和/或健康受试者。

[0841] 在获得患者样品之后,可以扩增和/或引发样品。可以实施各种方法来扩增和引发患者样品。在一个实施方案中,可以在肽或串联小基因的存在下模拟新鲜和/或冷冻的PBMC。在另一个实施方案中,可以在肽或串联小基因的存在下用抗原呈递细胞(APC)模拟和引发新鲜和/或冷冻的分离的T细胞。APC的实例包括B细胞、单核细胞、树突状细胞、巨噬细胞或人工抗原呈递细胞(例如呈递相关HLA和共刺激分子的细胞或珠粒,在<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929753>中有综述)。在另一个实施方案中,可以在细胞因子(例如IL-2、IL-7和/或IL-15)存在下刺激PBMC、TIL和/或分离的T细胞。在另一个实施方案中,可以在最大刺激物、细胞因子和/或饲养细胞的存在下刺激TIL和/或分离的T细胞。在这样的实施方案中,可以通过激活标志物和/或多聚体(例如,四聚体)分离T细胞。在另一个实施方案中,可以用刺激性和/或共刺激性标志物(例如,CD3抗体、CD28抗体和/或珠粒(例如,DynaBeads)刺激TIL和/或分离的T细胞。在另一个实施方案中,可以在富培养基中以高剂量IL-2在饲养细胞上使用快速扩增方案扩增DTC。

[0842] 然后,鉴别和分离新抗原特异性T细胞。在一些实施方案中,从患者离体样品分离T细胞,而无需事先扩增。在一个实施方案中,以上关于XVII.部分描述的方法可用于从患者样品中鉴别新抗原特异性T细胞。在另一个实施方案中,通过阳性选择富集特定细胞群或通过阴性选择耗尽特定细胞群来进行分离。在一些实施方案中,通过将细胞与一种或多种抗体或其它结合剂孵育来实现阳性或阴性选择,所述抗体或其它结合剂与分别在阳性或阴性选择的细胞上表达(标志物+)或以相对高的水平表达(标志物高)的一种或多种表面标志物特异性结合。

[0843] 在一些实施方案中,通过非T细胞(例如B细胞、单核细胞或其它白细胞)上表达的标志物(例如CD14)的阴性选择从PBMC样品分离T细胞。在一些方面,CD4⁺或CD8⁺选择步骤用于分离CD4⁺辅助细胞和CD8⁺细胞毒性T细胞。可以通过对在一种或多种天然、记忆和/或效应T细胞亚群上表达或以相对较高的程度表达的标志物的阳性或阴性选择将这样的CD4⁺和CD8⁺种群进一步分选为亚群。

[0844] 在一些实施方案中,例如通过基于与各个亚群相关的表面抗原的阳性或阴性选择进一步富集或耗尽CD8⁺细胞的天然、中枢记忆、效应记忆和/或中枢记忆干细胞。在一些实施方案中,进行中枢记忆T(TCM)细胞的富集以提高效力,例如改善施用后的长期存活、扩增和/或植入,其在一些方面在这样的亚群中特别强效。参见Terakura等人(2012) Blood. 1: 72-82; Wang等人(2012) J Immunother. 35(9): 689-701。在一些实施方案中,组合富含TCM的CD8⁺T细胞和CD4⁺T细胞进一步增强了功效。

[0845] 在一些实施方案中,记忆T细胞存在于CD8⁺外周血淋巴细胞的CD62L⁺和CD62L⁻亚群二者中。可以富集或耗尽PBMC的CD62L⁻CD8⁺和/或CD62L⁺CD8⁺级分,例如使用抗CD8和抗CD62L抗体。

[0846] 在一些实施方案中,中枢记忆T(TCM)细胞的富集是基于CD45RO、CD62L、CCR7、CD28、CD3和/或CD127的阳性或高表面表达;在一些方面,其基于对表达或高表达CD45RA和/

或颗粒酶B的细胞的阴性选择。在一些方面,通过表达CD4、CD14、CD45RA的细胞的耗尽以及表达CD62L的细胞的阳性选择或富集来进行富集TCM细胞的CD8+群的分离。在一方面,从基于CD4表达选择的细胞的阴性级分开始,对其进行基于CD14和CD45RA表达的阴性选择以及基于CD62L的阳性选择来进行中枢记忆T (TCM) 细胞的富集。在一些方面,这样的选择同时进行,而在另一些方面,以任一顺序依次进行。在一些方面,用于制备CD8+细胞群或亚群的相同的基于CD4表达的选择步骤也用于产生CD4+细胞群或亚群,使得任选地在一个或多个阳性或阴性选择步骤之后,保留来自基于CD4-的分离的阳性和阴性级分二者并且用于方法的后续步骤。

[0847] 在特定的实例中,对PBMC样品或其它白细胞样品进行CD4+细胞的选择,其中保留阴性级分和阳性级分二者。然后,对阴性级分进行基于CD14和CD45RA或ROR1的表达的阴性选择,以及基于中枢记忆T细胞的标志物特征(例如CD62L或CCR7)的阳性选择,其中阳性和阴性选择以任一顺序进行。

[0848] 通过鉴别具有细胞表面抗原的细胞群,将CD4+T辅助细胞分选为天然、中枢记忆和效应细胞。CD4+淋巴细胞可以通过标准方法获得。在一些实施方案中,天然CD4+T淋巴细胞是CD45RO⁻、CD45RA⁺、CD62L⁺、CD4+T细胞。在一些实施方案中,中枢记忆CD4+细胞是CD62L⁺和CD45RO⁺。在一些实施方案中,效应CD4+细胞是CD62L⁻和CD45RO⁻。

[0849] 在一个实施方案中,为了通过阴性选择富集CD4+细胞,单克隆抗体混合物通常包括针对CD14、CD20、CD11b、CD16、HLA-DR和CD8的抗体。在一些实施方案中,抗体或结合配偶体结合至固体支持物或基质,例如磁珠或顺磁珠,以允许分离细胞以用于阳性和/或阴性选择。例如,在一些实施方案中,使用免疫-磁性(或亲和-磁性)分离技术分离或分隔细胞和细胞群(综述于Methods in Molecular Medicine,第58卷:Metastasis Research Protocols,第2卷:Cell Behavior In Vitro and In Vivo,第17-25页,编辑:S.A.Brooks and U.Schumacher Humana Press Inc.,Totowa,N.J.)。

[0850] 在一些方面,将待分离的样品或细胞组合物与小的可磁化或磁响应的材料,例如磁响应颗粒或微粒,例如顺磁珠(例如,Dynabeads或MACS珠粒)一起孵育。磁响应材料(例如颗粒)通常直接或间接附着至结合配偶体(例如抗体),该结合配偶体特异性结合存在于期望分离(例如期望阴性或阳性选择)的细胞、多个细胞或细胞群上的分子(例如表面标志物)。

[0851] 在一些实施方案中,磁性颗粒或珠粒包含结合至特异性结合成员(例如抗体或其它结合配偶体)的磁响应材料。存在许多用于磁分离方法中的众所周知的磁响应材料。合适的磁性颗粒包括在Molday的美国专利第4,452,773号以及欧洲专利说明书EP 452342B中描述的那些,所述专利通过引用并入本文。胶体大小的颗粒,例如在Owen的美国专利第4,795,698号和Liberti等人的美国专利第5,200,084号中描述的那些是其它示例。

[0852] 孵育通常在一定的条件下进行,所述条件使得附着于磁性颗粒或磁珠的抗体或结合配偶体或者与这样的抗体或结合配偶体特异性结合的分子(例如二抗或其它试剂)特异性地结合细胞表面分子(如果存在于样品中的细胞上的话)。

[0853] 在一些方面,将样品放置在磁场中,并且具有附着于其上的磁响应或可磁化颗粒的那些细胞将被磁体吸引并与未标记的细胞分离。对于阳性选择,保留被磁体吸引的细胞。对于阴性选择,保留未被吸引的细胞(未标记的细胞)。在一些方面,在同一选择步骤期间进

行阳性选择和阴性选择的组合,其中保留阳性和阴性级分并进一步处理或经受进一步的分离步骤。

[0854] 在某些实施方案中,将磁响应颗粒包被在一抗或其它结合配偶体、二抗、凝集素、酶或链霉亲和素中。在某些实施方案中,磁性颗粒通过对一种或多种标志物具有特异性的一抗涂层附着到细胞上。在某些实施方案中,用一抗或结合配偶体标记细胞而不是珠粒,然后添加细胞类型特异性的二抗或其它结合配偶体(例如链霉亲和素)包被的磁性颗粒。在某些实施方案中,链霉亲和素包被的磁性颗粒与生物素化的一抗或二抗结合使用。

[0855] 在一些实施方案中,使磁响应颗粒附着于待随后孵育、培养和/或工程改造的细胞;在一些方面,使颗粒附着于用于向患者施用的细胞。在一些实施方案中,从细胞中去除可磁化或磁响应颗粒。用于从细胞中去除可磁化颗粒的方法是已知的,并且包括例如使用竞争性未标记的抗体、与可切割接头缀合的可磁化颗粒或抗体等。在一些实施方案中,可磁化颗粒是可生物降解的。

[0856] 在一些实施方案中,基于亲和力的选择是通过磁激活细胞分选(MACS)(Miltenyi Biotech, Auburn, Calif.)进行的。磁激活细胞分选(MACS)系统能够高纯度选择附着有磁化颗粒的细胞。在某些实施方案中,MACS以其中在施加外部磁场之后非靶物质和靶物质顺序洗脱的模式操作。即,附着于磁化颗粒上的细胞被保持在原位,而未附着的物质被洗脱。然后,在完成该第一洗脱步骤后,以某种方式将捕获在磁场中并被阻止洗脱的物质释放出来,以便可以将其洗脱并回收。在某些实施方案中,非靶T细胞被标记并从异质细胞群中耗尽。

[0857] 在某些实施方案中,使用执行所述方法的分离、细胞制备、分隔、加工、孵育、培养和/或配制步骤中的一个或多个的系统、装置或设备来执行所述分离或分隔。在一些方面,系统用于在封闭或无菌环境中执行这些步骤中的每一个,例如以使错误、用户操作和/或污染最小化。在一个实例中,系统是如国际专利申请公开号W02009/072003或US 20110003380A1中所述的系统。

[0858] 在一些实施方案中,系统或设备在集成或自包含系统中和/或以自动化或可编程方式执行分离、加工、工程改造和配制步骤中的一个或多个,例如全部。在一些方面,系统或设备包括与该系统或设备通信的计算机和/或计算机程序,其允许用户对加工、分离、工程改造和配制步骤进行编程、控制、评估结果和/或调整多个方面。

[0859] 在一些方面,使用CliniMACS系统(Miltenyi Biotec)进行分离和/或其它步骤,例如,用于在封闭和无菌系统中在临床规模水平上自动分离细胞。组件可以包括集成的微型计算机、磁分离单元、蠕动泵和各种夹管阀。在一些方面,集成计算机控制仪器的所有组件,并指示系统以标准化顺序执行重复程序。在一些方面,磁分离单元包括可移动的永磁体和用于选择柱的保持器。蠕动泵控制整个管组的流速,并与夹管阀一起确保缓冲液通过系统的受控流动和细胞的连续悬浮。

[0860] 在一些方面,CliniMACS系统使用提供在无菌、无热原溶液中的抗体偶联的可磁化颗粒。在一些实施方案中,在用磁性颗粒标记细胞后,洗涤细胞以除去过量的颗粒。然后将细胞制备袋连接到管组,该管组又连接到包含缓冲液的袋和细胞收集袋。管组由预组装的无菌管道组成,包括前置柱和分离柱,并且仅供一次性使用。启动分离程序后,系统会自动将细胞样品上样到分离柱上。标记的细胞保留在柱内,而未标记的细胞通过一系列洗涤步骤除去。在一些实施方案中,用于本文所述方法的细胞群体是未标记的并且不保留在柱中。

在一些实施方案中,用于本文所述方法的细胞群体是标记的并保留在柱中。在一些实施方案中,在除去磁场后从柱中洗脱用于本文所述方法的细胞群体,并收集在细胞收集袋中。

[0861] 在某些实施方案中,使用CliniMACS Prodigy系统(Miltenyi Biotec)进行分离和/或其它步骤。在一些方面,CliniMACS Prodigy系统配备了细胞处理单元,该单元允许通过离心自动洗涤和分级分离细胞。CliniMACS Prodigy系统还可以包括机载摄像头和图像识别软件,该软件通过辨别源细胞产品的宏观层来确定最佳的细胞分级分离终点。例如,可以将外周血自动分离为红细胞、白细胞和血浆层。CliniMACS Prodigy系统还可以包括集成的细胞培养室,该室完成细胞培养方案,例如细胞分化和扩增、抗原加载和长期细胞培养。输入口可以允许无菌去除和补充培养基,并且可以使用集成显微镜监控细胞。参见,例如,Klebanoff等人.(2012)J Immunother.35(9):651-660, Terakura等人.(2012)Blood.1:72-82,和Wang等人.(2012)J Immunother.35(9):689-701。

[0862] 在一些实施方案中,通过流式细胞术收集和富集(或耗尽)本文所述的细胞群,其中针对多种细胞表面标志物染色的细胞被携带在流体流中。在一些实施方案中,通过制备规模(FACS)分选收集和富集(或耗尽)本文所述的细胞群。在某些实施方案中,通过使用与基于FACS的检测系统组合的微机电系统(MEMS)芯片来收集和富集(或耗尽)本文所述的细胞群(参见,例如,WO 2010/033140,Cho等人(2010)Lab Chip 10,1567-1573;和Godin等人.(2008)J Biophoton.1(5):355-376。在两种情况下,可以用多种标志物标记细胞,从而允许以高纯度分离定义明确的T细胞亚群。

[0863] 在一些实施方案中,用一种或多种可检测的标志物标记抗体或结合配偶体,以促进用于阳性和/或阴性选择的分离。例如,分离可以基于与荧光标记的抗体的结合。在一些实例中,基于对一种或多种细胞表面标志物特异性的抗体或其它结合配偶体的结合的细胞的分离在流体流中进行,例如通过荧光激活细胞分选(FACS),包括制备规模(FACS)和/或微机电系统(MEMS)芯片,例如与流式细胞术检测系统组合使用。这样的方法允许同时基于多种标志物进行阳性和阴性选择。

[0864] 在一些实施方式中,制备方法包括在分离、孵育和/或工程改造之前或之后冷冻(例如冷冻保存)细胞的步骤。在一些实施方案中,冷冻和随后的解冻步骤去除了细胞群中的粒细胞,并且在一定程度上去除了单核细胞。在一些实施方案中,例如在洗涤步骤以除去血浆和血小板之后,将细胞悬浮在冷冻溶液中。在一些方面,可以使用多种已知的冷冻溶液和参数中的任何一种。一个实例涉及使用含有20%DMSO和8%人血清白蛋白(HSA)或其它合适的细胞冷冻介质的PBS。然后可以用培养基将其1:1稀释,以使DMSO和HSA的终浓度分别为10%和4%。其它实例包括Cryostor®、CTL-Cryo™ABC冷冻介质等。然后将细胞以每分钟1度的速率冷冻至-80摄氏度,并存储在液氮储罐的蒸汽相中。

[0865] 在一些实施方案中,所提供的方法包括培养、孵育、培养和/或基因工程步骤。例如,在一些实施方案中,提供了用于孵育和/或工程改造耗尽的细胞群体和培养起始组合物的方法。

[0866] 因此,在一些实施方案中,将细胞群体在培养起始组合物中孵育。孵育和/或工程改造可以在培养容器中进行,例如单元、腔室、孔、柱、管、管组、阀、小瓶、培养皿、袋或用于培养或培养细胞的其它容器。

[0867] 在一些实施方案中,在基因工程之前或与基因工程相结合孵育和/或培养细胞。孵

育步骤可包括培养、培养、刺激、激活和/或繁殖。在一些实施方案中,将组合物或细胞在刺激条件或刺激剂的存在下孵育。这样的条件包括被设计以诱导群体中细胞的增殖、扩增、活化和/或存活、模拟抗原暴露和/或引发细胞以进行基因工程(例如用于引入重组抗原受体)的那些条件。

[0868] 条件可以包括以下一种或多种:特定培养基、温度、氧含量、二氧化碳含量、时间、剂(例如营养素、氨基酸、抗生素、离子)和/或刺激因子(例如细胞因子、趋化因子、抗原、结合配偶体、融合蛋白、重组可溶性受体)和旨在激活细胞的任何其它剂。

[0869] 在一些实施方案中,刺激条件或剂包括能够激活TCR复合物的细胞内信号传导结构域的一种或多种剂,例如配体。在一些方面,该剂打开或启动T细胞中的TCR/CD3细胞内信号传导级联。这样的剂可以包括抗体,例如对TCR组分和/或共刺激受体具有特异性的抗体,例如抗CD3、抗CD28,其例如与固体支持物如珠粒和/或一种或多种细胞因子结合。任选地,扩增方法可以进一步包括将抗CD3和/或抗CD28抗体添加到培养基中的步骤(例如,以至少约0.5ng/ml的浓度)。在一些实施方案中,刺激剂包括IL-2和/或IL-15,例如,IL-2浓度为至少约10单位/mL。

[0870] 在一些方面,根据例如以下中描述的那些技术进行孵育:Riddell等人的美国专利第6,040,177号,Klebanoff等人(2012)J Immunother.35(9):651-660,Terakura等人(2012)Blood.1:72-82,和/或Wang等人(2012)J Immunother.35(9):689-701。

[0871] 在一些实施方案中,通过向培养起始组合物添加饲养细胞,例如非分裂外周血单核细胞(PBMC)(例如,使得对于待扩增的起始群体中的每个T淋巴细胞,所得细胞群体包含至少约5、10、20或40或更多个PBMC饲养细胞);以及孵育培养物(例如足以扩增T细胞的时间)来扩增T细胞。在一些方面,非分裂饲养细胞可以包含 γ 辐照的PBMC饲养细胞。在一些实施方案中,用约3000至3600拉德范围内的 γ 射线辐照PBMC以防止细胞分裂。在一些实施方案中,将PBMC饲养细胞用丝裂霉素C灭活。在一些方面,在添加T细胞群体之前将饲养细胞添加至培养基。

[0872] 在一些实施方案中,刺激条件包括适合于人T淋巴细胞生长的温度,例如,至少约25摄氏度,通常至少约30摄氏度,并且通常在或约37摄氏度。任选地,孵育还可以包括添加非分裂的EBV转化的类淋巴母细胞(LCL)作为饲养细胞。LCL可以用约6000至10,000拉德范围内的 γ 射线辐照。在一些方面,以任何合适的量提供LCL饲养细胞,例如LCL饲养细胞与初始T淋巴细胞的比率为至少约10:1。

[0873] 在一些实施方案中,通过用抗原刺激天然或抗原特异性T淋巴细胞来获得抗原特异性T细胞,例如抗原特异性CD4+和/或CD8+T细胞。例如,可以通过从受感染的受试者中分离T细胞并用相同的抗原体外刺激细胞来产生巨细胞病毒抗原的抗原特异性T细胞系或克隆。

[0874] 在一些实施方案中,在用功能测定(例如,ELISpot)刺激后,鉴别和/或分离新抗原特异性T细胞。在一些实施方案中,通过细胞内细胞因子染色对多功能细胞进行分选来分离新抗原特异性T细胞。在一些实施方案中,使用活化标志物(例如,CD137、CD38、CD38/HLA-DR双阳性和/或CD69)鉴别和/或分离新抗原特异性T细胞。在一些实施方案中,使用I类或II类多聚体和/或活化标志物鉴别和/或分离新抗原特异性CD8+、自然杀伤T细胞、记忆T细胞和/或CD4+T细胞。在一些实施方案中,使用记忆标志物(例如,CD45RA、CD45RO、CCR7、CD27和/或

CD62L) 鉴别和/或分离新抗原特异性CD8+和/或CD4+T细胞。在一些实施方案中,鉴别和/或分离增殖细胞。在一些实施方案中,鉴别和/或分离活化的T细胞。

[0875] 从患者样品中鉴别出新抗原特异性T细胞之后,对鉴别出的新抗原特异性T细胞中的新抗原特异性TCR进行测序。为了对新抗原特异性TCR进行测序,必须首先鉴别TCR。鉴别T细胞的新抗原特异性TCR的一种方法可以包括使T细胞与包含至少一种新抗原的HLA-多聚体(例如,四聚体)接触;以及通过HLA-多聚体和TCR之间的结合鉴别TCR。鉴别新抗原特异性TCR的另一种方法可以包括获得包含TCR的一种或多种T细胞;用在至少一种抗原呈递细胞(APC)上呈递的至少一种新抗原活化所述一种或多种T细胞;以及通过选择通过与至少一种新抗原相互作用而活化的一种或多种细胞来鉴别TCR。

[0876] 在鉴别出新抗原特异性TCR之后,可以对TCR进行测序。在一个实施方案中,以上关于XVII.部分描述的方法可用于对TCR进行测序。在另一个实施方案中,可以对TCR的TCRa和TCRb进行批量测序,然后基于频率进行配对。在另一个实施方案中,可以使用Howie等人, Science Translational Medicine 2015(doi:10.1126/scitranslmed.aac5624)的方法对TCR进行测序和配对。在另一个实施方案中,可以使用Han等人., Nat Biotech 2014(PMID 24952902, doi 10.1038/nbt.2938)的方法对TCR进行测序和配对。在另一个实施方案中,可以使用以下中描述的方法获得配对的TCR序列:

[0877] <https://www.biorxiv.org/conent/early/2017/05/05/134841>和<https://patents.google.com/patent/US20160244825A1/>。^{158,159}

[0878] 在另一个实施方案中,可以通过有限稀释产生T细胞的克隆群,然后可以对T细胞的克隆群的TCRa和TCRb进行测序。在又一个实施方案中,可以将T细胞分选到具有孔的板上,使得每个孔有一个T细胞,然后可以对每个孔中每个T细胞的TCRa和TCRb进行测序和配对。

[0879] 接下来,在从患者样品中鉴别出新抗原特异性T细胞并且对所鉴别的新抗原特异性T细胞的TCR进行测序之后,将经测序的TCR克隆到新的T细胞中。这些经克隆的T细胞含有新抗原特异性受体,例如含有细胞外结构域,包括TCR。还提供了这样的细胞的群体以及包含这样的细胞的组合物。在一些实施方案中,使组合物或群体富集这样的细胞,例如其中表达TCR的细胞占某类型的组合物或细胞(例如T细胞或CD8+或CD4+细胞)中总细胞的至少1、5、10、20、30、40、50、60、70、80、90、91、92、93、94、95、96、97、98、99或超过99百分比。在一些实施方案中,组合物包含至少一种包含本文公开的TCR的细胞。组合物包括用于施用,例如用于过继细胞疗法的药物组合物和制剂。还提供了用于将细胞和组合物施用给受试者(例如患者)的治疗方法。

[0880] 因此,还提供了表达TCR的基因工程细胞。细胞通常是真核细胞,例如哺乳动物细胞,并且通常是人细胞。在一些实施方案中,细胞源自血液、骨髓、淋巴或淋巴器官,是免疫系统的细胞,例如先天或适应性免疫的细胞,例如髓样或淋巴样细胞,包括淋巴细胞,通常为T细胞和/或NK细胞。其它示例性细胞包括干细胞,例如多能和多潜能干细胞,包括诱导性多能干细胞(iPSC)。细胞通常是原代细胞,例如直接从受试者中分离和/或从受试者中分离并冷冻的细胞。在一些实施方案中,细胞包括T细胞或其它细胞类型的一个或多个子集,例如整个T细胞群体、CD4+细胞、CD8+细胞及其亚群,例如由功能、活化状态、成熟度、潜能分化、扩增、再循环、定位和/或持久能力、抗原特异性、抗原受体的类型、在特定器官或区室中

的存在、标志物或细胞因子的分泌谱和/或分化程度定义的那些。关于待治疗的受试者,细胞可以是同种异体的和/或自体的。这些方法包括现成的方法。在一些方面,例如对于现成技术,细胞是多能的和/或多潜能的,例如干细胞,例如诱导多能干细胞(iPSC)。在一些实施方案中,该方法包括如本文所述从受试者分离细胞、制备、加工、培养和/或工程改造它们,以及在冷冻保存之前或之后将其重新引入同一患者。

[0881] T细胞和/或CD4+和/或CD8+T细胞的亚型和亚群是天然T(TN)细胞、效应T细胞(TEFF)、记忆T细胞及其亚型,例如干细胞记忆T(TSCM)、中枢记忆T(TCM)、效应记忆T(TEM)或终末分化的效应记忆T细胞、肿瘤浸润淋巴细胞(TIL)、未成熟T细胞、成熟T细胞、辅助T细胞、细胞毒性T细胞、粘膜相关性不变T(MALT)细胞、天然和适应性调节性T(Treg)细胞、辅助T细胞(例如TH1细胞、TH2细胞、TH3细胞)、TH17细胞、TH9细胞、TH22细胞、滤泡辅助T细胞、 α/β T细胞和 δ/γ T细胞。

[0882] 在一些实施方案中,细胞是自然杀伤(NK)细胞。在一些实施方案中,细胞是单核细胞或粒细胞,例如髓样细胞、巨噬细胞、嗜中性粒细胞、树突状细胞、肥大细胞、嗜酸性粒细胞和/或嗜碱性粒细胞。

[0883] 可以对细胞进行遗传修饰以减少表达或敲除内源TCR。这样的修饰描述在以下中: Mol Ther Nucleic Acids.2012 Dec;1(12):e63;Blood.2011 Aug 11;118(6):1495-503; Blood.2012 Jun 14;119(24):5697-5705;Torikai,Hiroki等人"HLA and TCR Knockout by Zinc Finger Nucleases:Toward"off-the-Shelf"Allogeneic T-Cell Therapyfor CD19+ Malignancies.."Blood 116.21(2010):3766;Blood.2018 Jan 18;131(3):311-322.doi:10.1182/blood-2017-05-787598;和W02016069283,所述文献通过引用整体并入。

[0884] 可以对细胞进行遗传修饰以促进细胞因子的分泌。这样的修饰描述在以下中:Hsu C,Hughes MS,Zheng Z,Bray RB,Rosenberg SA,Morgan RA.Primary human T lymphocytes engineered with a codon-optimized IL-15 gene resist cytokine withdrawal-induced apoptosis and persist long-term in the absence of exogenous cytokine.J Immunol.2005;175:7226-34;Quintarelli C,Vera JF,Savoldo B,Giordano Attianese GM,Pule M,Foster AE,Co-expression of cytokine and suicide genes to enhance the activity and safety of tumor-specific cytotoxic T lymphocytes.Blood.2007;110:2793-802;和Hsu C,Jones SA,Cohen CJ,Zheng Z,Kerstann K,Zhou J,Cytokine-independent growth and clonal expansion of a primary human CD8+ T-cell clone following retroviral transduction with the IL-15 gene.Blood.2007;109:5168-77。

[0885] 已显示T细胞上趋化因子受体和肿瘤分泌的趋化因子的错配是造成T细胞向肿瘤微环境的次佳运输的原因。为了提高治疗效果,可以对细胞进行遗传修饰,以提高对肿瘤微环境中趋化因子的识别。这样的修饰描述在以下中: Moon,EKCarpenito,CSun,JWang, LCKapoor,VPredina,J Expression of a functional CCR2 receptor enhances tumor localization and tumor eradication by retargeted human T-cells expressing a mesothelin-specific chimeric antibody receptor.Clin Cancer Res.2011;17:4719-4730;和Craddock,JALu,ABear,APule,MBrenner,MKRooney,CM et al.Enhanced tumor trafficking of GD2 chimeric antigen receptor T-cells by expression of the

chemokine receptor CCR2b.J Immunother.2010;33:780-788。

[0886] 可以对细胞进行遗传修饰以增强共刺激/增强受体(例如CD28和41BB)的表达。

[0887] T细胞疗法的不良反应可包括细胞因子释放综合征和延长的B细胞耗竭。在受体细胞中引入自杀/安全开关可以改善基于细胞的疗法的安全性谱。因此,可以对细胞进行遗传修饰以包含自杀/安全开关。自杀/安全开关可以是这样的基因,其在表达该基因的细胞上赋予对剂例如药物的敏感性,并且当细胞与该剂接触或暴露于该剂时导致该细胞死亡。示例性的自杀/安全开关描述于Protein Cell.2017 Aug;8(8):573-589中。自杀/安全开关可以是HSV-TK。自杀/安全开关可以是胞嘧啶脱氨酶、嘌呤核苷磷酸化酶或硝基还原酶。自杀/安全开关可以是美国专利申请公开No.US20170166877A1中描述的RapaCIDE™。自杀/安全开关系统可以是Haematologica.2009 Sep;94(9):1316-1320中描述的CD20/利妥昔单抗。这些参考文献通过引用整体并入。

[0888] TCR可以作为分裂受体(split receptor)引入受体细胞,分裂受体仅在异二聚化小分子的存在下组装。这样的系统描述在Science.2015年10月16日;350(6258):aab4077和美国专利第9,587,020号中,所述文献通过引用并入。

[0889] 在一些实施方案中,细胞包含一种或多种核酸,例如编码本文公开的TCR的多核苷酸,其中所述多核苷酸通过基因工程引入,并因此表达本文公开的重组或基因工程TCR。在一些实施方案中,核酸是异源的,即,通常不存在于从所述细胞获得的细胞或样品中,例如是从另一种生物或细胞获得的,例如其通常不在所工程改造的细胞和/或这样的细胞所来源的生物中发现。在一些实施方案中,核酸不是天然存在的,例如自然界中不存在的核酸,包括包含编码来自多种不同细胞类型的多个结构域的核酸的嵌合组合的核酸。

[0890] 核酸可包括密码子优化的核苷酸序列。不受特定理论或机制的束缚,据信核苷酸序列的密码子优化增加了mRNA转录物的翻译效率。核苷酸序列的密码子优化可以包括将天然密码子替换为另一种密码子,所述另一种密码子编码相同氨基酸,但是可以通过在细胞内更容易获得的tRNA进行翻译,从而提高翻译效率。核苷酸序列的优化还可以减少将会干扰翻译的二级mRNA结构,从而提高翻译效率。

[0891] 可使用构建体或载体将TCR引入受体细胞。本文描述了示例性构建体。编码TCR的 α 和 β 链的多核苷酸可以在单个构建体中或在分开的构建体中。编码 α 和 β 链的多核苷酸可以可操作地连接至启动子,例如异源启动子。异源启动子可以是强启动子,例如EF1 α 、CMV、PGK1、Ubc、 β 肌动蛋白、CAG启动子等。异源启动子可以是弱启动子。异源启动子可以是诱导型启动子。示例性诱导型启动子包括但不限于TRE、NFAT、GAL4、LAC等。其它示例性诱导型表达系统描述于美国专利第5,514,578号、第6,245,531号、第7,091,038号和欧洲专利第0517805号中,所述专利通过引用整体本文。

[0892] 用于将TCR引入受体细胞的构建体还可包含编码信号肽的多核苷酸(信号肽元件)。信号肽可以促进引入的TCR的表面运输。示例性信号肽包括但不限于CD8信号肽、免疫球蛋白信号肽,其中具体实例包括GM-CSF和IgG κ 。这样的信号肽在以下中进行了描述:Trends Biochem Sci.2006年10月;31(10):563-71.Epub2006年8月21日;和An,等人“Construction of a New Anti-CD19 Chimeric Antigen Receptor and the Anti-Leukemia Function Study of the Transduced T-cells.”Oncotarget 7.9(2016):10638-10649.PMC.Web.2018年8月16日,所述文献通过引用并入本文。

[0893] 在一些情况下,例如,在从单个构建体或开放阅读框表达 α 和 β 链的情况,或在该构建体中包含标志基因的情况,该构建体可包含核糖体跳读序列。核糖体跳读序列可以是2A肽,例如P2A或T2A肽。示例性的P2A和T2A肽在Scientific Reports第7卷,文章编号:2193 (2017)中进行了描述,所述文献通过引用整体并入。在一些情况下,在2A元件上游引入了FURIN/PACE切割位点。FURIN/PACE切割位点描述于例如<http://www.nuolan.net/substrates.html>中。切割肽也可以是因子Xa的切割位点。在从单个构建体或开放阅读框表达 α 和 β 链的情况下,该构建体可包含内部核糖体进入位点(IRES)。

[0894] 构建体可以进一步包含一种或多种标志基因。示例性标志基因包括但不限于GFP、荧光素酶、HA、lacZ。如本领域技术人员已知的,标志物可以是可选择的标志物,例如抗生素抗性标志物、重金属抗性标志物或抗生物杀灭剂标志物。标志物可以是用于营养缺陷宿主的互补标志物。示例性的互补标志物和营养缺陷的宿主在Gene. 2001年1月24日;263(1-2): 159-69中进行了描述。这样的标志物可以通过IRES、移码序列、2A肽接头、与TCR融合表达,或者由单独的启动子分开表达。

[0895] 用于将TCR引入受体细胞的示例性载体或系统包括但不限于腺相关病毒、腺病毒、腺病毒+修饰的牛痘病毒、安卡拉病毒(MVA)、腺病毒+逆转录病毒、腺病毒+仙台病毒、腺病毒+牛痘病毒、甲病毒(VEE)复制子疫苗、反义寡核苷酸、长双歧杆菌(*Bifidobacterium longum*)、CRISPR-Cas9、大肠杆菌(*E. coli*)、黄病毒、基因枪、疱疹病毒、单纯疱疹病毒、乳酸乳球菌、电穿孔、慢病毒、脂质体转染、单核细胞性李斯特菌(*Listeria monocytogenes*)、麻疹病毒、修饰的牛痘安卡拉病毒(MVA)、mRNA电穿孔、裸/质粒DNA、裸/质粒DNA+腺病毒、裸/质粒DNA+修饰的牛痘安卡拉病毒(MVA)、裸/质粒DNA+RNA转移、裸/质粒DNA+牛痘病毒、裸/质粒DNA+水泡性口炎病毒、新城疫病毒、非病毒、PiggyBac™(PB)转座子、基于纳米颗粒的系统、脊髓灰质炎病毒、痘病毒、痘病毒+牛痘病毒、逆转录病毒、RNA转移、RNA转移+裸/质粒DNA、RNA病毒、酿酒酵母(*Saccharomyces cerevisiae*)、鼠伤寒沙门氏菌(*Salmonella typhimurium*)、塞姆利基森林病毒(Semliki forest virus)、仙台病毒、痢疾志贺氏菌(*Shigella dysenteriae*)、猿猴病毒、siRNA、睡美人转座子、变形链球菌(*Streptococcus mutans*)、牛痘病毒、委内瑞拉马脑炎病毒复制子、水泡性口炎病毒和霍乱弧菌(*Vibrio cholera*)。

[0896] 在优选的实施方案中,将TCR通过腺相关病毒(AAV)、腺病毒、CRISPR-CAS9、疱疹病毒、慢病毒、脂转染、mRNA电穿孔、PiggyBac™(PB)转座子、逆转录病毒、RNA转移或睡美人转座子引入受体细胞。

[0897] 在一些实施方案中,用于将TCR引入受体细胞的载体是病毒载体。病毒载体的实例包括腺病毒载体、腺相关病毒(AAV)载体、慢病毒载体、疱疹病毒载体、逆转录病毒载体等。这样的载体在本文中描述。

[0898] 用于将TCR引入受体细胞的TCR构建体的示例性实施方案在图33中示出。在一些实施方案中,TCR构建体在5'-3'方向包含以下多核苷酸序列:启动子序列、信号肽序列、TCR β 可变(TCR β v)序列、TCR β 恒定(TCR β c)序列、切割肽(例如,P2A)、信号肽序列、TCR α 可变(TCR α v)序列和TCR α 恒定(TCR α c)序列。在一些实施方案中,构建体的TCR β c和TCR α c序列包含一个或多个鼠区域,例如,如本文所述的完整鼠恒定序列或人鼠氨基酸交换。在一些实施方案中,构建体进一步在TCR α c序列的3'包含切割肽序列(例如,T2A)然后是报告基因。在一个实

施方案中,该构建体在5'-3'方向包含以下多核苷酸序列:启动子序列、信号肽序列、TCR β 可变(TCR β v)序列、包含一个或多个鼠区域的TCR β 恒定(TCR β c)序列、切割肽(例如P2A)、信号肽序列、TCR α 可变(TCR α v)序列和包含一个或多个鼠区域的TCR α 恒定(TCR α c)序列、切割肽(例如(T2A)和报告基因。

[0899] 图34描绘了用于将TCR克隆到表达系统中以进行疗法开发的示例性P526构建体骨架核苷酸序列。

[0900] 图35描绘了用于将患者新抗原特异性TCR,克隆型1克隆到表达系统中以进行疗法开发的示例性构建体序列。

[0901] 图36描绘了用于将患者新抗原特异性TCR,克隆型3克隆到表达系统中以进行疗法开发的示例性构建体序列。

[0902] 还提供了编码TCR的分离的核酸,包含所述核酸的载体以及包含所述载体和核酸的宿主细胞,以及用于产生TCR的重组技术。

[0903] 核酸可以是重组的。可通过将天然或合成核酸片段连接至可在活细胞中复制的核酸分子或其复制产物来在活细胞外部构建重组核酸。出于本文的目的,复制可以是体外复制或体内复制。

[0904] 为了重组产生TCR,可以分离编码它的核酸并将其插入可复制的载体中以进一步克隆(即,DNA的扩增)或表达。在一些方面,核酸可以通过同源重组产生,例如,如美国专利第5,204,244号中所述,其通过引用整体并入本文。

[0905] 许多不同的载体是本领域已知的。载体组分通常包括以下一种或多种:信号序列、复制起点、一种或多种标志基因、增强子元件、启动子和转录终止序列,例如美国专利第5,534,615号中所述,其通过引用并入本文。

[0906] 适用于表达TCR、抗体或其抗原结合片段的示例性载体或构建体包括,例如,pUC系列(Fermentas Life Sciences)、pBluescript系列(Stratagene,LaJolla,CA)、pET系列(Novagen,Madison,WI)、pGEX系列(Pharmacia Biotech,Uppsala,Sweden)和pEX系列(Clontech,Palo Alto,CA)。噬菌体载体,例如AGT10、AGT11、AZapII(Stratagene)、AEMBL4和ANM1149也适用于表达本文公开的TCR。

[0907] XIX. 治疗概述流程图

[0908] 图37是根据一个实施方案的用于向患者提供定制的新抗原特异性治疗的方法的流程图。在其它实施方案中,该方法可以包括与图37所示的步骤不同的步骤和/或另外的步骤。另外,该方法的步骤可以以与多个实施方案中结合图37描述的顺序不同的顺序执行。

[0909] 如上所述,使用质谱数据训练呈递模型3701。获得患者样品3702。在一些实施方案中,患者样品包含肿瘤活检和/或患者的外周血。对在步骤3702中获得的患者样品进行测序,以鉴别输入呈递模型中的数据,以预测来自患者样品的肿瘤抗原肽将被呈递的可能性。使用训练的呈递模型来预测在步骤3702中获得的来自患者样品的肿瘤抗原肽的呈递可能性3703。基于预测的呈递可能性为患者鉴别治疗新抗原3704。接下来,获得另一个患者样品3705。该患者样品可以包含患者的外周血、肿瘤浸润淋巴细胞(TIL)、淋巴、淋巴结细胞和/或任何其它T细胞来源。将步骤3705中获得的患者样品在体内筛选3706新抗原特异性T细胞。

[0910] 在治疗过程中的这一点上,患者可以接受T细胞疗法和/或疫苗治疗。为了接受疫

苗治疗,鉴别患者的T细胞对其特异的新抗原3714。然后,产生包含已鉴别的新抗原的疫苗3715。最后,向患者施用疫苗3716。

[0911] 为了接受T细胞疗法,对新抗原特异性T细胞进行扩增和/或对新的新抗原特异性T细胞进行基因工程改造。为了扩增新抗原特异性T细胞以用于T细胞疗法,简单地将细胞扩增3707并输注3708到患者。

[0912] 为了对新的新抗原特异性T细胞进行基因工程改造以用于T细胞疗法,对在体内鉴别出的新抗原特异性T细胞的TCR进行测序3709。接下来,将这些TCR序列克隆到表达载体中3710。然后将表达载体3710转染到新的T细胞中3711。扩增转染的T细胞3712。最后,将扩增的T细胞输注到患者体内3713。

[0913] 患者可以同时接受T细胞疗法和疫苗疗法。在一个实施方案中,患者首先接受疫苗疗法,然后接受T细胞疗法。这种方法的一个优点是疫苗疗法可以增加肿瘤特异性T细胞的数量和由可检测水平的T细胞识别的新抗原的数量。

[0914] 在另一个实施方案中,患者可以接受T细胞疗法后进行疫苗疗法,其中疫苗中包含的表位集包含由T细胞疗法靶向的一个或多个表位。该方法的一个优点是疫苗的施用可以促进治疗性T细胞的扩增和持久性。

[0915] XX. 示例计算机

[0916] 图38示出了用于实施图1和3中所示实体的示例计算机3800。计算机3800包括耦合至芯片组3804的至少一个处理器3802。芯片组3804包括内存控制器集线器3820和输入/输出(I/O)控制器集线器3822。内存3806和图形适配器3812耦合至内存控制器集线器3820,并且显示器3818耦合至图形适配器3812。存储装置3808、输入装置3814和网络适配器3816耦合至I/O控制器集线器3822。计算机3800的其它实施方案具有不同的架构。

[0917] 存储装置3808是非暂时性计算机可读存储介质,如硬盘驱动器、致密光盘只读存储器(CD-ROM)、DVD或固态内存装置。内存3806保存处理器3802所使用的指令和数据。输入接口3814是触摸屏界面、鼠标、轨迹球或其它类型的指向装置、键盘或其某一组合,并且用于将数据输入计算机3800中。在一些实施方案中,计算机3800可以被配置成通过用户的示意动作从输入接口3814接收输入(例如,命令)。图形适配器3812将图像和其它信息显示于显示器3818上。网络适配器3816将计算机3800耦合至一个或多个计算机网络。

[0918] 计算机3800被调适成执行计算机程序模块以提供本文所述的功能。如本文所使用,术语“模块”是指用于提供指定功能的计算机程序逻辑。因此,模块可以在硬件、固件和/或软件中实施。在一个实施方案中,程序模块被存储于存储装置3808上,装载至内存3806中并由处理器3802执行。

[0919] 图1的实体所使用的计算机3800的类型可以根据实施方案和实体所需的处理能力而变化。举例来说,呈递鉴别系统160可以在单一计算机3800或在通过网络,如在服务器群中彼此通信的多台计算机3800中运行。计算机3800可以缺少以上描述的组件中的一些,如图形适配器3812和显示器3818。

[0920]

补充表 1
NSCLS 患者中的人口统计

患者 ID	年龄范围 (岁)	性别	种族	初始(肺癌)诊断年 份	肿瘤阶段(登记 时)	原发肿瘤位置	组织学类型
1-001-002	81-90	男	白人	2010	IIIB	肺	非鳞状
1-024-001	81-90	男	白人	2016	IV	肺	肺肉瘤样癌
1-024-002	51-60	女	白人	2016	IV	肺	腺癌
1-038-001	61-70	男	白人	2016	IV	肺	腺癌

[0921]

补充表 1

NSCLS 患者中的人口统计

系统性 NSCLC 定向疗法	当前抗 PD(L)-1 疗法	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C	表达的突 变
卡铂	纳武单抗	A*01:01	A*01:01	B*08:01	B*51:01	C*01:02	C*07:01	122
		A*32:01	A*03:01	B*27:05	B*27:05	C*02:02	C*02:02	83
多西他赛、贝伐珠单抗、雷 莫芦单抗、培美曲塞二钠	纳武单抗							
		A*68:01	A*68:01	B*40:02	B*40:27	C*03:04	C*03:04	38
培美曲塞、顺铂	纳武单抗	A*69:01	A*01:02	B*41:01	B*49:01	C*17:01	C*07:01	158

[0922]

补充表 1
NSCLS 患者中的人口统计

非同义突变	正常 DNA 中值外 子覆盖率	肿瘤 DNA 中值外 子覆盖率	RNA PF 独 特读段(M)	已知驱动	可能驱动	中值 VAF
232	145	552	173	KRAS_G12D, TP53_R213*	STK11_G52fs	0.22
143	165	508	131.9	KRAS_G12C, TP53_R280T	PML_E43*, NF2_R341*	0.093
69	190	454	114.4	KRAS_G12S, TP53_Q331*	STK11_E199*	0.182
265	158	983	311.8	KRAS_G12V	KDM5C_E303 *	0.19

[0923]

补充表 1

NSCLS 患者中的人口统计

患者 ID	年龄范围 (岁)	性别	种族	初始(肺癌)诊断年 份	肿瘤阶段(登记 时)	原发肿瘤位置	组织学类型
1-050-001	71-80	女	白人	2015	IIIB	肺	腺癌
CU05	71-80	女	白人	2013	IV	肺	肺鳞状细胞癌
CU04	61-70	女	西班牙裔或 拉丁裔	2013	I	肺	腺癌
CU03	61-70	男	非裔美国人	2016	I	肺	肺鳞状细胞癌
CU02	61-70	男	白人	2016	I	肺	肺鳞状细胞癌

[0924]

补充表 1
NSCLS 患者中的人口统计

系统性 NSCLC 定向疗法	当前抗 PD(L)-1 疗法	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C	表达的突 变
依托泊苷、顺铂	纳武单抗	A*29:02	A*26:01	B*44:03	B*07:05	C*16:01	C*15:05	53
卡铂+培美曲塞	纳武单抗	A*24:02	A*68:02	B*14:02	B*15:17	C*07:01	C*08:02	65
	德瓦鲁单抗 (durvalumab)+替西木单抗 (tremelimumab)							
	n/a	A*24:26	A*26:01	B*18:01	B*38:01	C*12:03	C*12:03	336
	n/a	A*23:01	A*01:01	B*08:01	B*15:03	C*01:02	C*12:03	105
卡铂+吉西他滨	n/a	A*02:01	A*03:01	B*07:02	B*57:01	C*07:02	C*06:02	102

[0925]

补充表 1

NSCLS 患者中的人口统计

非同义突变	正常 DNA 中值外显子覆盖率	肿瘤 DNA 中值外显子覆盖率	RNA PF 独特读段(M)	已知驱动	可能驱动	中值 VAF
92	117	556	119			0.059
109	191	448	83.6			0.095
					NFKBIE_G41fs, CDH1_Q346*, NF1_D2163fs, MED12_R730*	
511	213	552	240.4	TP53_R158G		0.224
187	114	830	182.1			0.242
174	105	738	185.3	TP53_R175H	ATR_Q195*	0.32

[0926]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应 (任何时间 点)	个体肽 响应注 释	库 ID	库 响应(任何时 间点)	突变
1-001-002	HSPFTATSL	N		1-001-002_库_1	N	chr15_28215653_C_A
1-001-002	DPEEVLVTV	N		1-001-002_库_1	N	chr17_59680958_C_T
1-001-002	ELDPDIQLEY	N		1-001-002_库_1	N	chr13_30210371_C_A
1-001-002	TPLTKDVTI	N		1-001-002_库_1	N	chr5_78100974_A_T
1-001-002	DGVGKSAL	N		1-001-002_库_1	N	chr12_25245350_C_T
1-001-002	YTTVRALTI	N		1-001-002_库_1	N	chr17_28339664_G_T
1-001-002	TPSAAVKLI	N		1-001-002_库_1	N	chr15_81319417_T_C
1-001-002	WPVLLLV	N		1-001-002_库_1	N	chr3_179025167_AAC_A
1-001-002	ELNARRCSF	N		1-001-002_库_1	N	chr18_79943341_G_A

[0927]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效 果	TPM	由完全 MS 模 型覆盖的最可 能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的 限制
snp	HERC2	A2060S	41.9	HLA-C*01:02	0		955169.68205	假
snp	CLTC	S989L	272.1	HLA-B*51:01	1		613455.25069	真
snp	KATNAL 1	D407Y	12.81	HLA-A*01:01	2		124.2177849	真
snp	AP3B1	S817T	44.4	HLA-B*08:01	3		248.9740194	真
snp	KRAS	G12D	40.75	HLA-B*08:01	4		894714.29522	真
snp	TNFAIP1	R48L	45.62	HLA-B*08:01	5		26973.417701	真
snp	STARD5	M108V	1.95	HLA-B*51:01	6		392030.48603	真
del fs	ZMAT3	V240fs	14.99	HLA-B*51:01	7		16600.564752	真
snp	PQLC1	R109C	33.89	HLA-B*08:01	8		562.0439997	真

[0928]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应 (任何时间点)	个体肽响应 应注释	库 ID	库 响应(任何时 间点)	突变
1-001-002	QMKNPILEL	N		1-001-002_库_1	N	chr9_127663287_G_T
1-001-002	LTEKVSLLK	N		1-001-002_库_2	N	chr9_92719180_C_T
1-001-002	SPFTATSL	N		1-001-002_库_2	N	chr15_28215653_C_A
1-001-002	NVDMRTISF	N		1-001-002_库_2	N	chr9_121353262_T_A
1-001-002	TSIVVSQTL	N		1-001-002_库_2	N	chr4_39205691_C_T
1-001-002	HIKIEPVAI	N		1-001-002_库_2	N	chr13_73062087_C_T
1-001-002	DSPDGSNGL	N		1-001-002_库_2	N	chr20_44197575_C_T
1-001-002	YTAVHYAASY	N		1-001-002_库_2	N	chr12_56248788_C_A
1-001-002	VGADGVGKSA L	N		1-001-002_库_2	N	chr12_25245350_C_T

[0929]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	STXBP1	R171L	38.76	HLA-B*08:01	9	20	674.64733	真
snp	BICD2	E489K	42.66	HLA-A*01:01	10	10	428.744925	真
snp	HERC2	A2060S	41.9	HLA-B*08:01	11	4	59.1155419	真
snp	STOM	K93N	360.6	HLA-B*08:01	12	30	1490.72261	真
snp	WDR19	A282V	18.12	HLA-B*08:01	13	176	9862.33009	真
snp	KLF5	T163I	25.77	HLA-B*08:01	14	27	1122.27455	真
snp	OSER1	S119N	20.7	HLA-C*01:02	15	471	21598.414	假
	ANKRD5							
snp		A559S	18.32	HLA-A*01:01	16	0	11.5906737	真
snp	KRAS	G12D	40.75	HLA-C*01:02	17	370	17985.3612	假

[0930]

补充表 2
NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应 (任何时间点)	个体肽响应 应注释	库 ID	库 响应(任何时间点)	突变
1-001-002	MMPPLPGI	N		1-001-002_库_2	N	chr17_32369404_A_T
1-001-002	FPYPGMTNQ	N		1-001-002_库_2	N	chr5_109186272_G_T
1-024-001	VTNHAPLSW	N		1-024-001_库_1	Y	chr3_125552370_C_A
1-024-001	GTKKDDVDVL	Y		1-024-001_库_1	Y	chr20_56513366_G_A
1-024-001	GLNVPVQSN	N		1-024-001_库_1	Y	chr4_88390868_G_T
1-024-001	VVVGACGV	N		1-024-001_库_1	Y	chr12_25245351_C_A
1-024-001	AQFAGKDQT	N		1-024-001_库_1	Y	chr9_89045819_C_A
1-024-001	KVVLPSDVT	N		1-024-001_库_1	Y	chr3_48591778_G_T
1-024-001	MLMKNISTK	N		1-024-001_库_1	Y	chr12_6959976_G_A

[0931]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类 型	基因	蛋白质 效果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆 盖的最可能的限制
snp	ZNF207	Q409L	186	HLA-B*51:01	18	136	7609.76602	真
snp	FER	C759F	67.36	HLA-B*51:01	19	38	1999.07208	真
snp	OSBPL1	G489W	24.12	HLA-A*32:01	0	7	77.009026	真
snp	RTFDC1	E177K	61.32	HLA-A*03:01	1	70	2168.51668	真
snp	HERC6	R218L	8.7	HLA-A*03:01	2	4	59.675168	真
snp	KRAS	G12C	40.05	HLA-A*03:01	3	11	133.648023	真
snp	SHC3	E376D	8.88	HLA-A*32:01	4	91	3715.42819	真
snp	COL7A1	R468S	25.42	HLA-A*32:01	6	85	3234.15772	真
snp	PTPN6	E471K	105.4	HLA-A*03:01	7	0	12.2301919	真

[0932]

补充表 2
NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间点)	个体肽响应注释	库 ID	库响应(任何时间点)	突变
1-024-001	DLAGGTFDV	N		1-024-001_库_1	Y	chr11_123059991_C_G
1-024-001	LIFDLAGGTF	N		1-024-001_库_1	Y	chr11_123059991_C_G
1-024-001	NVLIFDLA	N		1-024-001_库_1	Y	chr11_123059991_C_G
1-024-001	VVGACGVGK	N		1-024-001_库_2	N	chr12_25245351_C_A
1-024-001	VIMLNGTKK	N		1-024-001_库_2	N	chr20_56513366_G_A
1-024-001	LAGGTFDV	N		1-024-001_库_2	N	chr11_123059991_C_G
1-024-001	LRNSGGEVF	N		1-024-001_库_2	N	chr14_80906012_TC_T
1-024-001	VVLPSDVTSY	N		1-024-001_库_2	N	chr3_48591778_G_T
1-024-001	IFDLAGGTF	N		1-024-001_库_2	N	chr11_123059991_C_G

[0933]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	HSPA8	G201A	736.6	HLA-B*27:05	9	353	18290.7955	真
snp	HSPA8	G201A	736.6	HLA-C*02:02	11	57	1716.74204	假
snp	HSPA8	G201A	736.6	HLA-A*32:01	17	621	27984.1357	真
snp	KRAS	G12C	40.05	HLA-A*03:01	5	19	197.846108	真
snp	RTFDC1	E177K	61.32	HLA-A*03:01	8	10	122.750322	真
snp	HSPA8	G201A	736.6	HLA-C*02:02	10	632	28384.8834	假
del fs	CEP128	R102fs	11.31	HLA-B*27:05	12	46	1020.95087	真
snp	COL7A1	R468S	25.42	HLA-A*32:01	13	62	1925.29397	真
snp	HSPA8	G201A	736.6	HLA-C*02:02	14	427	21255.2074	假

[0934]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应 (任何时间点)	个体肽响应 (任何时间点)	库 ID	库 响应(任何 时间点)	突变
1-024-001	GLLDEAKRLLY	N		1-024-001_库_2	N	chr19_57575861_G_T
1-024-001	SVLLPENYITK	N		1-024-001_库_2	N	chr11_122789248_G_T
1-024-001	DLAGGTFDVS	N		1-024-001_库_2	N	chr11_123059991_C_G
1-024-001	IFDLAGGTFDV	N		1-024-001_库_2	N	chr11_123059991_C_G
1-024-002	AEWRNGSTSSL	N		1-024-002_库_1	Y	chr3_122703943_C_G
1-024-002	YVSEKDVISAK	N		1-024-002_库_1	Y	chr2_43889858_G_A
1-024-002	EGSLGISHTR	N		1-024-002_库_1	Y	chr18_62157782_C_A
1-024-002	IPASVSAPK	N		1-024-002_库_1	Y	chr13_109784018_C_A
1-024-002	QDVSVQVER	Y		1-024-002_库_1	Y	chr9_64411223_T_G

[0935]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质 效果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆 盖的最可能的限制
snp	ZNF416	Q49K	11.89	HLA-A*03:01	15	24	354.82068	真
snp	UBASH3	G307V	12.11	HLA-A*03:01	16	23	228.127132	真
snp	HSPA8	G201A	736.6	HLA-A*32:01	18	487	23357.3292	真
snp	HSPA8	G201A	736.6	HLA-C*02:02	19	563	25887.4267	假
snp	PARP14	P1095A	129.5	HLA-A*68:01	0	8	126.397714	真
snp	LRPPRC	T1335I	79.08	HLA-A*68:01	1	9	136.482978	真
snp	PIGN	W83L	20.74	HLA-A*68:01	2	6	88.2623459	真
snp	IRS2	S679I	63.55	HLA-A*68:01	3	16	224.278982	真
snp	ANKRD2 0A4	M646R	8.92	HLA-A*68:01	4	14	193.974327	真

[0936]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间点)	个体肽响 应注释	库 ID	库 响应(任何 时间点)	突变
1-024-002	LVVVGASGVG K	N		1-024-002_库_1	Y	chr12_25245351_C_T
1-024-002	RATIVPEL	N		1-024-002_库_1	Y	chr7_131463253_A_T
1-024-002	SSAAAPFPL	Y		1-024-002_库_1	Y	chr6_13711102_T_A
1-024-002	GVSKIIGGNPK	N		1-024-002_库_1	Y	chr4_10116175_C_T
1-024-002	EQNFVSTSDIK	未单独测试		1-024-002_库_1	Y	chr3_25791346_A_C
1-024-002	RTQDVSVQVE R	N		1-024-002_库_2	Y	chr9_64411223_T_G
1-024-002	EAGNNSRVPR	N		1-024-002_库_2	Y	chr2_74046630_G_T
1-024-002	RYVLHVVA	N		1-024-002_库_2	Y	chr3_122703943_C_G
1-024-002	VSKIIGGNPK	N		1-024-002_库_2	Y	chr4_10116175_C_T

[0937]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效 果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模 型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的 限制
snp	KRAS	G12S	72.77	HLA-A*68:01	6		411238.56407	真
snp	MKLN1	D521V	84.08	HLA-C*03:04	7		26616010.7063	假
snp	RANBP9	H135L	43.5	HLA-C*03:04	8		1034565.97417	假
snp	WDR1	D26N	134.5	HLA-A*68:01	9		1256797.60699	真
snp	OXSM	K109T	12.82	HLA-A*68:01	17		1569099.70986	真
snp	ANKRD2 0A4	M646R	8.92	HLA-A*68:01	5		531847.42359	真
snp	TET3	G238V	56.35	HLA-A*68:01	10		13161.242762	真
snp	PARP14	P1095A	129.5	HLA-A*68:01	11		176 10453.627	真
snp	WDR1	D26N	134.5	HLA-A*68:01	12		38954.724495	真

[0938]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间点)	个体肽响 应注释	库 ID	库 响应(任何 时间点)	突变
1-024-002	QPSGVPTSL	N		1-024-002_库_2	Y	chr12_14478436_GG_ TT
1-024-002	DVSVQVER	N		1-024-002_库_2	Y	chr9_64411223_T_G
1-024-002	FVSTSDIKSM	Y		1-024-002_库_2	Y	chr3_25791346_A_C
1-024-002	FPVVNSHSL	N		1-024-002_库_2	Y	chr1_116062776_G_C
1-024-002	APFPLGDSAL	N		1-024-002_库_2	Y	chr6_13711102_T_A
1-024-002	ATIVPELNEI	N		1-024-002_库_2	Y	chr7_131463253_A_T
1-038-001	QEFAPLGTV	N	见库结果	1-038-001_库_1	Y	chr2_219501883_G_T
1-038-001	MNQVLHAY	未单独测试	见库结果	1-038-001_库_1	Y	chr14_100354547_C_ G
1-038-001	HEDVKEAI	未单独测试	见库结果	1-038-001_库_1	Y	chr8_96231911_C_G

[0939]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
inmp	ATF7IP	G1021L	123.2	HLA-A*68:01	13	139	7795.97025	真
snp	ANKRD20A4	M646R	8.92	HLA-A*68:01	14	7	123.489687	真
	OXSM	K109T	12.82	HLA-C*03:04	15	128	7025.56581	假
snp	SLC22A15	A396P	8.57	HLA-C*03:04	16	155	9082.40652	假
	RANBP9	H135L	43.5	HLA-A*68:01	18	196	11590.601	真
snp	MKLN1	D521V	84.08	HLA-A*68:01	19	365	19785.1419	真
snp	GMPPA	G92V	21.6	HLA-B*49:01	0	31	3481.07375	假
snp	WARS	D148H	757.2	HLA-C*07:01	12	422	27180.1513	假
snp	UQCRB	D41H	174.8	HLA-B*49:01	16	300	24830.2411	假

[0940]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间点)	个体肽响 应注释	库 ID	库 响应(任何 时间点)	突变
1-038-001	GPYPFVQAV	未单独测试	见库结果	1-038-001_库_1	Y	chr1_111242326_C_T
1-038-001	YEHEDVKEAI	未单独测试	见库结果	1-038-001_库_1	Y	chr8_96231911_C_G
1-038-001	EESVMLLTV	未单独测试	见库结果	1-038-001_库_1	Y	chr1_15583354_CC_A G
1-038-001	IEEDSAEKI	未单独测试	见库结果	1-038-001_库_1	Y	chr6_84215849_C_A
1-038-001	TEEDVKIKF	未单独测试	见库结果	1-038-001_库_1	Y	chr7_93105459_C_A
1-038-001	NEQSKLLKV	未单独测试	见库结果	1-038-001_库_1	Y	chrX_70375298_C_G
1-038-001	VDNIIQSI	未单独测试	见库结果	1-038-001_库_1	Y	chr20_2654879_G_T
1-038-001	YEHEDVKEA	Y		1-038-001_库_2	Y	chr8_96231911_C_G
1-038-001	YVSEVPVSV	未单独测试		1-038-001_库_2	Y	chr17_2330604_G_A

[0941]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质 效果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模 型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能 的限制
snp	CHI3L2	L379F	122.3	HLA-B*49:01	1	19	1176.97782	假
snp	UQCRB	D41H	174.8	HLA-B*49:01	2	212	22559.0306	假
mnp	AGMAT	G105L	1.03	HLA-B*49:01	3	109	17185.8013	假
snp	CEP162	E82D	15.62	HLA-B*49:01	4	171	20568.515	假
snp	SAMD9	M213I	68.23	HLA-B*49:01	5	226	22894.2742	假
snp	KIF4A	L625V	19.51	HLA-B*49:01	6	141	19054.8385	假
snp	NOP56	M167I	89.39	HLA-B*49:01	7	119	17928.6022	假
snp	UQCRB	D41H	174.8	HLA-B*49:01	9	250	23419.567	假
snp	TSR1	H561Y	48.21	HLA-C*17:01	10	0	6.07874308	假

[0942]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间 点)	个体肽响应 注释	库 ID	库 响应(任何时 间点)	突变
1-038-001	SELTVHQRI	未单独测试		1-038-001_库_2	Y	chr19_37564705_G_C
1-038-001	VGVGKSAL	未单独测试		1-038-001_库_2	Y	chr12_25245350_C_A
1-038-001	DMNQVLHAY	未单独测试		1-038-001_库_2	Y	chr14_100354547_C_G
1-038-001	NEKGKALIY	未单独测试		1-038-001_库_2	Y	chr17_51294040_G_T
1-038-001	TEYKLVVVGAV	未单独测试		1-038-001_库_2	Y	chr12_25245350_C_A
1-038-001	QEFAPLGTVG	未单独测试		1-038-001_库_2	Y	chr2_219501883_G_T
1-038-001	QEVNRTLNV	未单独测试		1-038-001_库_2	Y	chr17_4085728_C_A
1-038-001	VEMLGLISC	未单独测试		1-038-001_库_2	Y	chr4_168427109_C_A
1-050-001	LFHDMNVSY	N		1-050-001_库_1	N	chr1_193097666_T_C

[0943]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	ZNF571	L575V	19.07	HLA-B*49:01	11	159	19886.0407	假
snp	KRAS	G12V	91.89	HLA-C*17:01	13	388	26432.7668	假
snp	WARS	D148H	757.2	HLA-C*07:01	14	64	10286.4383	假
snp	UTP18	M547I	63.21	HLA-C*07:01	15	339	25564.2874	假
snp	KRAS	G12V	91.89	HLA-B*49:01	17	233	23113.572	假
snp	GMPPA	G92V	21.6	HLA-B*49:01	18	338	25558.5468	假
snp	ZZEF1	G863V	63	HLA-B*49:01	19	124	18359.7482	假
snp	DDX60L	A631S	44.71	HLA-B*49:01	8	267	23949.2398	假
snp	GLRX2	N94S	17.92	HLA-A*29:02	0	1	44.54051	真

[0944]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间,点)	个体肽响应注释	库 ID	库响应(任何时间,点)	突变
1-050-001	ISTFRQCAL	未单独测试		1-050-001_库_1	N	chr17_80346815_G_T
1-050-001	YNTDDIEFY	未单独测试		1-050-001_库_1	N	chr15_26580447_G_T
1-050-001	EETPPFSNY	N		1-050-001_库_1	N	chr21_31266125_T_A
1-050-001	QASGNHHVW	未单独测试		1-050-001_库_1	N	chr22_30893501_T_C
1-050-001	EEVTPILAI	未单独测试		1-050-001_库_1	N	chr18_5419733_G_A
1-050-001	IEHNIRNAKY	未单独测试		1-050-001_库_1	N	chr3_52617347_T_G
1-050-001	AERLDVKAI	未单独测试		1-050-001_库_1	N	chr14_103339252_G_T
1-050-001	LFQGGKDLQQ Y	未单独测试		1-050-001_库_1	N	chr17_80346815_G_T
1-050-001	DTSPVAVAL	未单独测试		1-050-001_库_1	N	chr5_73074790_T_C

[0945]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效 果	TPM	由完全 MS 模型覆 盖的最可能的限 制	MS 模 型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能 的限制
snp	RNF213	R2827L	330.6	HLA-C*16:01	10	322	22721.4424	假
snp	GABRB3	T185N	2.2	HLA-A*29:02	16	20	447.152559	真
snp	TIAM1	Y283F	13.99	HLA-B*44:03	1	26	537.02592	真
snp	OSBP2	Y677H	7.86	HLA-B*44:03	19	109	7506.81856	真
snp	EPB41L3	S495L	51.69	HLA-B*44:03	2	17	390.306194	真
snp	PBRM1	D578A	65.68	HLA-B*44:03	3	10	186.953378	真
snp	EIF5	M275I	89.97	HLA-B*44:03	5	34	1075.19965	真
snp	RNF213	R2827L	330.6	HLA-A*29:02	6	54	2855.46701	真
snp	FCHO2	L543S	43.6	HLA-A*26:01	8	91	5750.39585	真

[0946]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间)	个体肽响应注释	库 ID	库响应(任何时间)	突变
1-050-001	AEETPPFSNY	N		1-050-001_库_2	N	chr21_31266125_T_A
1-050-001	AAKAALEDF	未单独测试		1-050-001_库_2	N	chr3_47661451_C_G
1-050-001	EVTPIAIR	未单独测试		1-050-001_库_2	N	chr18_5419733_G_A
1-050-001	DVKAIGPLV	未单独测试		1-050-001_库_2	N	chr14_103339252_G_T
1-050-001	NETPVAVLT	未单独测试		1-050-001_库_2	N	chr7_79453094_C_A
1-050-001	LFVVFQTVY	未单独测试		1-050-001_库_2	N	chr1_159535913_A_T
1-050-001	AEAERLDVKAI	未单独测试		1-050-001_库_2	N	chr14_103339252_G_T
1-050-001	ASGNHHVW	未单独测试		1-050-001_库_2	N	chr22_30893501_T_C
1-050-001	KLFHDMNVS	未单独测试		1-050-001_库_2	N	chr1_193097666_T_C

[0947]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效 果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模 型排名	MHCFlurry y 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能 的限制
snp	TIAM1	Y283F	13.99	HLA-B*44:03	9	16	364.187996	真
snp	SMARCC1	E721D	39.53	HLA-C*16:01	11	307	22125.437	假
snp	EPB41L3	S495L	51.69	HLA-A*26:01	12	125	9269.11767	真
snp	EIF5	M275I	89.97	HLA-A*26:01	13	90	5692.75283	真
snp	MAGI2	G76V	2.29	HLA-B*44:03	14	13	253.431553	真
snp	OR10J5	L32Q	0.9	HLA-A*29:02	15	9	139.510048	真
snp	EIF5	M275I	89.97	HLA-B*44:03	17	38	1465.22509	真
snp	OSBP2	Y677H	7.86	HLA-C*16:01	18	173	13216.9384	假
snp	GLRX2	N94S	17.92	HLA-A*29:02	4	21	453.621334	真

[0948]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间点)	个体肽响应注释	库 ID	库响应(任何时间点)	突变
I-050-001	ETPPFSNYNTL	未单独测试		1-050-001_库_2	N	chr21_31266125_T_A
CU04	DENITTIQF	Y		CU04_库_1	Y	chr4_22413213_C_A
CU04	MELKVESF	N		CU04_库_1	Y	chr1_37874128_G_C
CU04	EHIPESAGF	N		CU04_库_1	Y	chr3_9943508_G_C
CU04	YHGDPMPCL	N		CU04_库_1	Y	chr12_7066530_C_T
CU04	DEERIPVL	N		CU04_库_1	Y	chr7_5752914_T_C
CU04	EVADAAATLTM	Y		CU04_库_1	Y	chr1_52268541_A_C
CU04	IEVEVNEI	N		CU04_库_1	Y	chr7_135598004_C_G
CU04	DTVEYPYTSF	Y		CU04_库_1	Y	chr14_34713369_C_A

[0949]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	TIAM1	Y283F	13.99	HLA-A*26:01	7	172	13162.6216	真
snp	ADGRA3	C734F	20.67	HLA-B*18:01	0	2	8.27203164	真
snp	INPP5B	Q606E	36.85	HLA-B*18:01	1	5	13.0510076	真
snp	CRELD1	Q347H	29.9	HLA-B*38:01	2	103	4218.0095	真
snp	C1S	P295L	157.5	HLA-B*38:01	3	12	76.7416543	真
snp	RNF216	M45V	49.2	HLA-B*18:01	4	29	387.328968	真
snp	ZFYVE9	K845T	70.08	HLA-A*26:01	5	7	38.7340629	真
snp	NUP205	L691V	42.37	HLA-B*18:01	6	21	209.301169	真
snp	CFL2	D66Y	16.65	HLA-A*26:01	7	9	42.7267485	真

[0950]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间)	个体肽响应注释	库 ID	库响应(任何时间)	突变
CU04	VEIEQLTY	N		CU04_库_1	Y	chr11_62827178_C_G
CU04	LELKAVHAY	N		CU04_库_1	Y	chr7_138762364_G_T
CU04	EEADFLLAY	N		CU04_库_2	N	chr6_10556704_C_T
CU04	ENITTIQFY	N		CU04_库_2	N	chr4_22413213_C_A
CU04	FHATNPLNL	N		CU04_库_2	N	chr14_75117203_C_G
CU04	VFKDLSVTL	N		CU04_库_2	N	chrX_40597563_G_A
CU04	QAVAAVQKL	N		CU04_库_2	N	chr17_42104792_T_A
CU04	IQDQIQNCI	N		CU04_库_2	N	chr2_67404159_G_C
CU04	VAKGFISRM	N		CU04_库_2	N	chr2_85395579_C_T

[0951]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	STX5	E134Q	83.43	HLA-B*18:01	8	3	11.6727539	真
snp	ATP6V0A4	P163H	47.21	HLA-B*18:01	9	0	3.63590379	真
snp	GCNT2	P94L	25.19	HLA-B*18:01	10	1	6.48490966	真
snp	ADGRA3	C734F	20.67	HLA-A*26:01	11	16	135.44155	真
snp	NEK9	D252H	20.29	HLA-B*38:01	12	8	39.1165673	真
snp	ATP6AP2	E145K	88.26	HLA-B*38:01	13	45	1080.8332	真
snp	DHX58	M513L	35.87	HLA-C*12:03	14	136	6872.44	真
snp	ETAA1	E493Q	38.47	HLA-B*38:01	15	59	1665.0162	真
snp	CAPG	E314K	151.7	HLA-C*12:03	16	107	5236.61406	真

[0952]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间点)	个体肽响 应注释	库 ID	库 响应(任何 时间点)	突变
CU04	QTKPASLLY	N		CU04_库_2	N	chr2_32487684_AG_A
CU04	DHFETIIKY	N		CU04_库_2	N	chr1_220024376_C_G
CU04	VEYPYTSF	N		CU04_库_2	N	chr14_34713369_C_A
CU05	SVSDISEYRV	N		CU05_库_1	N	chr12_15670870_G_C
CU05	YTFEIQGVNG V	N		CU05_库_1	N	chr1_22865138_C_G
CU05	IYTSSGQLQLF	N		CU05_库_1	N	chr10_73293336_T_C
CU05	FATPSLHTSV	N		CU05_库_1	N	chr17_80345147_A_T
CU05	AVSKPGLDYEL	N		CU05_库_1	N	chr14_77026556_T_A
CU05	KYINKTIRV	N		CU05_库_1	N	chr19_2328426_C_T

[0953]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
del fs	BIRC6	G2619fs	111.7	HLA-A*26:01	17	47	1143.73481	真
snp	EPRS	M277I	76.64	HLA-B*18:01	18	6	29.8996386	真
snp	CFL2	D66Y	16.65	HLA-B*18:01	19	4	12.3783994	真
snp	EPS8	Q64E	52.56	HLA-A*68:02	0	1	6.0399624	真
snp	EPHB2	A410G	74.99	HLA-A*68:02	1	22	132.877429	真
snp	CFAP70	E636G	30.45	HLA-A*24:02	2	17	46.3526841	真
snp	RNF213	D2271V	735.3	HLA-A*68:02	4	16	43.8761927	真
snp	IRF2BP1	M413L	58.51	HLA-A*68:02	5	274	13566.6012	真
snp	LSM7	D20N	76.01	HLA-A*24:02	8	32	318.671051	真

[0954]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何 时间点)	个体肽响应 应注释	库 ID	库 响应(任何时间点)	突变
CU05	ETTEEMKYVL	N		CU05_库_1	N	chr6 80040624 G_A
CU05	VVSHPHLVYW	N		CU05_库_1	N	chr4 106232956 C_G
CU05	DIFQVVKAI	N		CU05_库_1	N	chr1 198754369 C_A
CU05	FAFDVSKPGL	N		CU05_库_1	N	chr14 77026556 T_A
CU05	SVSDISEYR	N		CU05_库_2	N	chr12 15670870 G_C
CU05	YTFEIQGV	N		CU05_库_2	N	chr1 22865138 C_G
CU05	ATPSLHTSV	N		CU05_库_2	N	chr17 80345147 A_T
CU05	DFATPSLHTSV	N		CU05_库_2	N	chr17 80345147 A_T
CU05	KYINKTIRVKF	N		CU05_库_2	N	chr19 2328426 C_T

[0955]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	TTK	G804E	17.14	HLA-A*68:02	9	37	398.324158	真
snp	TBCK	D478H	71.17	HLA-A*68:02	11	235	10875.8686	真
snp	PTPRC	L1204I	104.6	HLA-A*68:02	13	36	394.198029	真
snp	IRF2BPL	M413L	58.51	HLA-A*68:02	18	65	1067.11951	真
snp	EPS8	Q64E	52.56	HLA-A*68:02	3	94	2050.45825	真
snp	EPHB2	A410G	74.99	HLA-A*68:02	6	11	26.6362167	真
snp	RNF213	D2271V	735.3	HLA-A*68:02	7	25	177.027506	真
snp	RNF213	D2271V	735.3	HLA-A*68:02	10	185	7619.02631	真
snp	LSM7	D20N	76.01	HLA-A*24:02	12	42	538.209517	真

[0956]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任何时间点)	个体肽响应注释	库 ID	库响应(任何时间点)	突变
CU05	SVKPHLCSL	N		CU05_库_2	N	chr17_35363437_C_T
CU05	DISEYRVEHL	N		CU05_库_2	N	chr12_15670870_G_C
CU05	WVVSHPHLV	N		CU05_库_2	N	chr4_106232956_C_G
CU05	KVFKLGNKV	N		CU05_库_2	N	chrX_24810777_G_A
CU05	VSKPGLDYEL	N		CU05_库_2	N	chr14_77026556_T_A
CU02	SPSKTSLTL	未单独测试	见库结果	CU02_库_1	Y	chr12_132750694_G_T
CU02	ASADGTVKWLW	未单独测试	见库结果	CU02_库_1	Y	chr16_1977246_A_G
CU02	LVGPAQLSHW	未单独测试	见库结果	CU02_库_1	Y	chr8_143930249_G_A
CU02	QTAAAVGVLK	未单独测试	见库结果	CU02_库_1	Y	chr7_77773271_A_G

[0957]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类型	基因	蛋白质效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	SLFN11	R124H	91.5	HLA-A*68:02	14	88	1897.58723	真
snp	EPS8	Q64E	52.56	HLA-A*68:02	15	59	885.161001	真
snp	TBCK	D478H	71.17	HLA-A*68:02	16	15	40.725305	真
snp	POLA1	E1017K	19.31	HLA-A*68:02	17	61	954.869111	真
snp	IRF2BP1	M413L	58.51	HLA-A*68:02	19	258	12457.5646	真
snp	ANKLE2	P266T	43.78	HLA-B*07:02	0	7	20.5140939	真
snp	TBL3	I545V	26.23	HLA-B*57:01	1	20	77.5504026	真
snp	PLEC	P863L	528.5	HLA-B*57:01	4	42	287.473059	真
snp	RSBN1L	T584A	25.89	HLA-A*03:01	5	19	76.1012011	真

[0958]

补充表 2
NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间)	个体肽响 应注释	库 ID	库 响应(任何时间 点)	突变
CU02	FPSPSKTSLTL	未单独测试	见库结果	CU02_库_1	Y	chr12_132750694_G_T
CU02	SSTSNRSTW	未单独测试	见库结果	CU02_库_1	Y	chr10_96604023_G_A
CU02	LVYGPLGAGK	未单独测试	见库结果	CU02_库_1	Y	chr13_33821175_C_T
CU02	HSYSELCTW	未单独测试	见库结果	CU02_库_1	Y	chr8_119802006_C_G
CU02	VTLDVILER	未单独测试	见库结果	CU02_库_1	Y	chr9_108979413_T_G
CU02	HSKPEDTDW	未单独测试	见库结果	CU02_库_1	Y	chr12_133057238_A_G
CU03	IAASRSVVM	未单独测试		CU03_库_1	N	chr1_230868472_G_A
CU03	AAIAASRSV	未单独测试		CU03_库_1	N	chr1_230868472_G_A
CU03	AASRSVVM	未单独测试		CU03_库_1	N	chr1_230868472_G_A

[0959]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类 型	基因	蛋白质 效果	TPM	由完全 MS 模型覆 盖的最可能的限 制	完全 MS 模 型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的 限制
snp	ANKLE2	P266T	43.78	HLA-B*07:02	6	26	131.765585	真
snp	PIK3AP1	R733W	9.84	HLA-B*57:01	7	30	162.029882	真
snp	RFC3	S44L	9.76	HLA-A*03:01	8	2	8.21211585	真
snp	TAF2	D194H	29.74	HLA-B*57:01	9	3	10.120376	真
snp	CTNNAL 1	E323D	32.44	HLA-B*57:01	10	136	2107.24068	真
snp	ZNF84	T175A	29.84	HLA-B*57:01	11	23	90.7546185	真
snp	C1orf19 8	A14V	36.47	HLA-C*12:03	0	19	146.699014	真
snp	C1orf19 8	A14V	36.47	HLA-C*12:03	2	42	492.404622	真
snp	C1orf19 8	A14V	36.47	HLA-C*12:03	6	116	3437.73836	真

[0960]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

患者	肽	个体肽响应(任 何时间点)	个体肽响 应注释	库 ID	库 响应(任何 时间点)	突变
CU03	EMDMHLSDY	未单独测试		CU03_库_1	N	chr5_37180032_T_A
CU03	VENQKHSL	未单独测试		CU03_库_1	N	chr12_30728769_C_T
CU03	QYMDSSLVKI	未单独测试		CU03_库_1	N	chr10_60788061_G_T
CU03	SASLHPATV	未单独测试		CU03_库_1	N	chr2_25929006_C_T
CU03	VPDQKSKQL	未单独测试		CU03_库_1	N	chr6_63685063_T_G
CU03	IVFIATSEF	未单独测试		CU03_库_1	N	chr11_65976483_A_T
CU03	YPAPQPPVL	未单独测试		CU03_库_1	N	chr20_44066022_C_A

[0961]

补充表 2

NSCLS 患者中的用于 T 细胞识别的测试肽

突变类 型	基因	蛋白质 效果	TPM	由完全 MS 模型覆盖的最可能的限制	完全 MS 模型排名	MHCFlurry 排名	MHCFlurry (nM)	由 MHCFlurry 覆盖的最可能的限制
snp	C5orf42	I1908L	14.78	HLA-A*01:01	8	7	35.7275148	真
snp	CAPRI N2	S554N	6.69	HLA-B*08:01	10	124	3970.47602	真
snp	CDK1	S107I	26.84	HLA-A*23:01	7	8	50.3301427	真
snp	KIF3C	R785H	17.29	HLA-C*12:03	9	30	260.370195	真
snp	PHF3	N447K	47.53	HLA-B*08:01	13	130	4071.14261	真
snp	SART1	N554I	70.53	HLA-B*15:03	5	3	17.4168253	真
snp	TOX2	S382Y	11.56	HLA-B*08:01	11	101	2455.95947	真

[0962]

补充表 3

分析物(平均值)	刺激	供体 ID				
		1-038-001	CU04	1-024-001	1-024-002	CU02
颗粒酶 B (pg/ml)*	DMSO	1786.73	1383.53	2639.03	854.78	1449.74
	肽库 1	1672.60	4269.64	2449.23	1281.54	1132.49
	DMSO	1874.02	3747.71	2382.01	626.20	n/a
	肽库 2	3118.30	3191.90	2006.73	872.89	n/a
TNFα (pg/ml)#	DMSO	37.58	34.64	21.76	38.07	1.22
	肽库 1	53.02	217.57	42.05	57.13	7.44
	DMSO	16.58	80.81	24.98	24.77	n/a
	肽库 2	61.54	75.70	33.70	48.84	n/a
IL-2 (pg/ml)#	DMSO	1.78	3.86	4.24	0.23	6.67
	肽库 1	15.53	9.88	7.75	0.00	0.00
	DMSO	26.66	27.25	5.72	10.20	n/a
	肽库 2	0.00	19.15	11.48	0.00	n/a
IL-5 (pg/ml)#	DMSO	26.47	5.20	20.92	11.96	18.91
	肽库 1	10.48	14.65	26.72	9.42	17.64
	DMSO	27.31	19.65	11.01	29.93	n/a
	肽库 2	26.47	25.43	20.11	40.11	n/a

阳性值以斜体显示。*颗粒酶 B ELISA: 相比于 DMSO 背景, 值≥1.5 倍被认为是阳性。#U-Plex MSD 测定: 相比于 DMSO 背景, 值≥1.5 倍被认为是阳性。

[0963]

补充表 4
IVS 对照实验中的 TSNA 和感染性疾病表位

肽名称	序列	来源(细胞系, 基 因)	预测的 HLA 限制	预测的结合亲和力	突变 位置	突变核苷酸
新抗原_A1	APKKKSIKL	H2009 PPFIA3	B*07:02	125	chr19- 12549140014	C-到-T
新抗原_A2	LLLEVVVHL	H128 FANCA	A*02:01	6	chr16- 689808348	C-到-T
新抗原_A3	FTDEKVKAY	H2122 PDE10A	A*01:01	41	chr6- 41165543564	G-到-T
新抗原_A6	RTAKQNPLTK	H2122GPR183	A*03:01	138	chr13- 99295446	G-到-A
新抗原_A7	FLAPTGVVPV	H128 NTM	A*02:01	8	chr11- 131911555	T-到-C
新抗原_A10	RLADAEKLFQL	H128 PLEKHG4	A*02:01	201	chr16- 67284435	G-到-A
新抗原_A11	RTAKQNPLTKK	H2122GPR183	A*03:01	131	chr13- 99295446	G-到-A
新抗原_B2	IMYLTGMV NK	H2009 GSPT1	A*03:01	33	chr16- 11891120	G-到-A
新抗原_B3	TLQELSHAL	H128 PRPF19	A*02:01	106	chr11- 60902829	G-到-T
新抗原_B6	VSQPVAPSY	Colo829 KIAA0319L	A*01:01	948	chr1- 35479047	C-到-T

[0964]

补充表 4

IVS 对照实验中的 TSNA 和感染性疾病表位

肽名称	序列	来源(细胞系，基因)	预测的 HLA 限制	预测的结合亲和力	突变位置	突变核苷酸
新抗原_B7	RLFTPISAGY	H2126CYP26B1	A*03:01		chr2-15772133060	G-到-C
新抗原_B8	ITEEPILMTY	H2122 RP1L1	A*01:01		chr8-30810611205	C-到-A
新抗原_B10	KVTGHRWLK	H2009 BSG	A*03:01		51chr19-579577	G-到-A
新抗原_B12	KLSEQILKK	H2009 TLR5	A*03:01		chr1-39223110532	C-到-G
新抗原_C3	GTKPNPHVY	H2126 OAS3	A*03:01		chr12-7336112961105	G-到-T
新抗原_C4	QQQQVVTNK	H2126 LRP1	A*03:01		chr12-236157162861	G-到-T
新抗原_C5	KVLGKGSAK	H2126 PLK2	A*03:01		chr5-4058459089	G-到-A
新抗原_C6	SVQAPVPPK	H2009 ENGASE	A*03:01		chr17-27979084548	C-到-G
EBV RAKF	RAKFKQLL	EBV BZLF-1	B*08:01		457Nan	Nan
Flu CTCL	CTELKLSDY	流感病毒 NP	A*01:01		39Nan	Nan
Flu ELRS	ELRSRYWAI	流感病毒 A	B*08:01		12Nan	Nan
CMV NLVP	NLVPMVATV	CMV pp65	A*02:01		45Nan	Nan
Flu GILG	GILGFVFTL	流感病毒 MP	A*02:01		20Nan	Nan

[0965]

补充表 4

IVS 对照实验中的 TSNA 和感染性疾病表位

肽名称	序列	来源(细胞系, 基 因)	预测的 HLA 限制	预测的结合亲 和力	突变 位置	突变核苷酸
HCV KLVA	KLVALGINA V	HCV NS3	A*02:01		49Nan	Nan
HIVILKE	ILKEPVHGV	HIV pol	A*02:01		144Nan	Nan
RSV NPKA	NPKASLLSL	RSVNP	B*07:02		60Nan	Nan
*新抗原序列中 的突变肽具有 下划线.						
**肿瘤细胞系: Colo829, H128, H2009, H2122, H2126				NaN	Nan	Nan
				NaN	Nan	Nan

[0966]

补充表 5

克隆型	频率	比例	TRAV	TRAJ	TRAC	TRBV	TRBD	TRBJ	TRBC
克隆型 1		3860.49171975	TRAV8-4	TRAJ5	TRAC	TRBV2	TRBD2	TRBJ2-5	TRBC2
克隆型 3		530.06751592	TRAV6	TRAJ31	TRAC	TRBV6-1	TRBD2	TRBJ1-1	TRBC1
克隆型 9		7	0.0089172	TRAJ33	TRAC	TRBV20-1	TRBD1	TRBJ1-5	TRBC1
克隆型 10		50.00636943	TRAV17	TRAJ57	TRAC	TRBV7-6	TRBD1	TRBJ2-3	TRBC2
克隆型 14		40.00509554	1	TRAJ33	TRAC	TRBV28	TRBD2	TRBJ2-7	TRBC2

[0967]

补充表 5

α CDR3	β CDR3	全长 α VJ
CAVTVTGRRALTF	CASNPPDAARGQ ETQYF	MLLLVPVLEVFTLGGTGAQSVTQLGSHVSVSEGALVLLRCNYSSSVPPYLFWYV QYPNQGLQLLLKYTTGATLVKGINGFEAEFKKSETSFHLTKPSAHMSDAAEYFCAV TVTGRRALTFGSGTRLQVQ
CALNARLMF	CASSYREYNTEAF F	MAFWLRRRLGLHFRPHLGRRMESFLGGVLLILWLQVDWVKSQKIEQNSEALNIQE GKTATLTCNYTNYSPAYLQWYRQDPGRGPVFLLLIRENEKEKREKRLKVTFDITTLK QSLFHITASQPADSATYLCALNARLMFGDGTQLVVK
CAVVLDSDNYQLIW	CSATRGHLSNQP QHF	MKRILGALLGLLSAQVCCVRGHIQVEQSPDDLILQEGANSTLRCNFSDSVNNLQWF HQNPWGQLINLFYIPSGTKQNGRLSATTVATERYSLLYISSQTDDSGVYFCAVVLD SNYQLIWGAGTKLIK
CATASRQGGSEKLV F	CASSRGGGTD DTQYF	METLLGVSLVILWLQARVNSQQGEEDPPQALSIQEGENATMNCYKTSINNQLW YRQNSGRGLVHLILIRSNEREKHSGRRLRVTLDTSKKSSLLITASRAADTASYFCATA SRQGGSEKLVFGKGTCLTVN
CAASSNYQLIW	CASSLGLAYEQYF	MTSIRAVFIPLWLQDLVNGENVEQHPSTLSVQEGDSA VIKCTYSDSASNYFPWYK QELGKGPQLIIDIRSNVGEKKDQRIAVTLNKTAKHFS LHITETQPEDSAVYFCAASS NYQLIWGAGTKLIK

[0968]

补充表 5

全长 β V(D) _J
MDTWLVCWAIFSLKAGLTEPEVTQTPSHQVTQMGEVILRCVPISNHLYFYWYR QILGQKVEFLVSFYNNSEIKSEIFDDQFSVERPDGSNFTLKIRSTKLEDSAMYFCAS NPPDAARGQETQYFGPGTRLVL
MSIGLLCCVAFSLLWASPVNAGVTQTPKFQVLKTGQSMTLQCAQDMNHNMY WYRQDPGMGLRLIYYSASEGTTDKGEVPNGYNVSRNLKREFSLRLESAAPSQTSVY FCASSYREYNTEAFFGQGTRLTVV
MLLLLLGPGSGLGAVVSQHPSRVICKSGTSVKIECRSLDFQATTMFWRQFPKQ SLMLMATSNESGSKATYEQQVEKDKFLINHASLTLSTLTVTSAHPEDSSFYICSATRG HLSNQPHFGDGRSL
MGTSLLCWVVLGFLGTDHTGAGVSQSPRYKVTKRGQDVALRCDPISGHVSLYWY RQALGQGPEFLTYFNYEAOQDKSGLPNDRFSAERPEGSISTLTIQRTQORDSAMYR CASSRGGGTDQYFGPGTRLTVL
MGIRLLCRVAFCLAVGLVDVKVTQSSRYLVKRTGEKVFLQVQDMMDHENMFWY RQDPGLGLRLIYFSYDVKMKEKGDIPEGYSVSREKKERFSLILESASTNQTSMYLCAS SLGLAYEQYFGPGTRLTVT

[0969] 参考文献

[0970] 1.Desrichard,A.,Snyder,A.&Chan,T.A.Cancer Neoantigens and Applications for Immunotherapy.Clin.Cancer Res.Off.J.Am.Assoc.Cancer Res.(2015).doi:10.1158/1078-0432.CCR-14-3175

[0971] 2.Schumacher,T.N.&Schreiber,R.D.Neoantigens in cancer immunotherapy.Science 348,69-74(2015).

[0972] 3.Gubin,M.M.,Artyomov,M.N.,Mardis,E.R.&Schreiber,R.D.Tumor neoantigens:building a framework for personalized cancer immunotherapy.J.Clin.Invest.125,3413-3421(2015).

- [0973] 4.Rizvi,N.A.et al.Cancer immunology.Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer.Science 348,124-128(2015) .
- [0974] 5.Snyder,A.et al.Genetic basis for clinical response to CTLA-4 blockade in melanoma.N.Engl.J.Med.371,2189-2199(2014) .
- [0975] 6.Carreno,B.M.et al.Cancer immunotherapy.A dendritic cell vaccine increases the breadth and divesity of melanoma neoantigen-specific T-cells.Science 348,803-808(2015) .
- [0976] 7.Tran,E.et al.Cancer immunotherapy based on mutation-specific CD4+T-cells in a patient with epithelial cancer.Science 344,641-645(2014) .
- [0977] 8.Hacohen,N.&Wu,C.J.-Y.United States Patent Application:0110293637-COMPOSITIONS AND METHODS OF IDENTIFYING TUMOR SPECIFIC NEOANTIGENS. (A1) .at<<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PT01&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1=20110293637.PGNR>>
>
- [0978] 9.Lundegaard,C.,Hoof,I.,Lund,O.&Nielsen,M.State of the art and challenges in sequence based T-cell epitope prediction.Immunome Res.6Suppl 2, S3(2010) .
- [0979] 10.Yadav,M.et al.Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.Natnre 515,572-576(2014) .
- [0980] 11.Bassani-Sternberg,M.,Pletscher-Frankild,S.,Jensen, L.J.&Mann, M.Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.Mol .Cell.Proteomics MCP 14,658-673(2015) .
- [0981] 12.Van Allen, E.M.et al.Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.Science 350,207-211(2015) .
- [0982] 13.Yoshida,K.&Ogawa,S.Splicing factor mutations and cancer.Wiley Interdiscip.Rev.RNA 5,445-459(2014) .
- [0983] 14.Cancer Genome Atlas Research Network.Comprehensive molecular profiling of lung adenocarcinoma.Nature 511,543-550(2014) .
- [0984] 15.Rajasagi,M.et al.Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia.Blood 124,453-462(2014) .
- [0985] 16.Downing,S.R.et al.United States Patent Application:0120208706-OPTIMIZATION OFMULTIGENE ANALYSIS OF TUMOR SAMPLES. (A1) .at<<http://appft1.uspto.gov/netacgi/nph-Parser?Sect1=PT01&Sect2=HITOFF&d=PG01&p=1&u=/netahtml/PTO/srchnum.html&r=1&f=G&l=50&s1=20120208706.PGNR>>
>
- [0986] 17.Target Capture for NextGen Sequencing-IDT.at<<http://www.idtdna.com/pages/products/nextgen/target-capture>>
- [0987] 18.Shukla,S.A.et al.Comprehensive analysis of cancer-associated

somatic mutations in class IHLA genes. *Nat. Biotechnol.* 33,1152-1158(2015) .

[0988] 19.Cieslik,M.et al.The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* 25, 1372-1381(2015) .

[0989] 20.Bodini,M.et al.The hidden genomic landscape of acute myeloid leukemia:subclonal structure revealed by undetected mutations. *Blood* 125,600-605(2015) .

[0990] 21.Salunders,C.T.et al.Strelka:accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinforma.Oxf.Engl.* 28,1811-1817(2012) .

[0991] 22.Cibulskis,K.et al.Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31,213-219(2013) .

[0992] 23.Wilkerson,M.D.et al.Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* 42,e107(2014) .

[0993] 24.Mose,L.E.,Wilkerson,M.D.,Hayes,D.N.,perou,C.M.&Parker,J.S.ABRA: improved coding indel detection via assembly-based realignment. *Bioinforma.Oxf.Engl.* 30,2813-2815(2014) .

[0994] 25.Ye,K.,Schulz,M.H.,Long,Q.,Apweiler,R.&Ning,Z.Pindel:a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinforma.Oxf.Engl.* 25,2865-2871 (2009) .

[0995] 26.Lam,H.Y.K.et al.Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* 28,47-55 (2010) .

[0996] 27.Frampton,G.M.et al.Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* 31,1023-1031(2013) .

[0997] 28.Boegel,S.et al.HLA typing from RNA-Seq sequence reads. *Genome Med.* 4,102(2012) .

[0998] 29.Liu,C.et al.ATHLATES:accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* 41,e142(2013) .

[0999] 30.Mayor,N.P.et al.HLA Typing for the Next Generation. *PloS One* 10, e0127153(2015) .

[1000] 31.Roy,C.K.,Olson,S.,Graveley,B.R.,Zamore,P.D.&Moore,M.J.Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation. *eLife* 4, (2015) .

[1001] 32.Song,L.&Florea,L.CLASS:constrained transcript assembly of RNA-seq reads. *BMC Bioinformatics* 14 Suppl 5,S14(2013) .

[1002] 33.Maretty,L,Sibbesen,J.A.&Krogh,A.Bayesian transcriptome

assembly. *Genome Biol.* 15,501(2014) .

[1003] 34.Pertea,M.et al.StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat.Biotechnol.* 33,290-295(2015) .

[1004] 35.Roberts,A.,Pimentel,H.,Trapnell,C.&Pachter,L.Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinforma.Oxf.Engl.* (2011).doi:10.1093/bioinformatics/btr355

[1005] 36.Vitting-Seerup,K.,Porse,B.T.,Sandelin,A.&Waage,J.spliceR:an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* 15,81(2014) .

[1006] 37.Rivas,M.A.et al.Human genomics.Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348,666-669 (2015) .

[1007] 38.Skelly,D.A.,Johansson, M.,Madeoy,J.,Wakefield,J.&Akey,J.M.A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* 21,1728-1737(2011) .

[1008] 39.Anders,S.,Pyl,P.T.&Huber,W.HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinforma.Oxf.Engl.* 31,166-169(2015) .

[1009] 40.Furney,S.J.et al.SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov.* (2013).doi:10.1158/2159-8290.CD-13-0330

[1010] 41.Zhou,Q.et al.A chemical genetics approach for the functional assessment of novel cancer genes. *Cancer Res.* (2015).doi:10.1158/0008-5472.CAN-14-2930

[1011] 42.Maguire,S.L.et al.SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J.Pathol.* 235,571-580(2015) .

[1012] 43.Carithers,L.J.et al A Novel Approach to High-Quality postmortem Tissue Procurement:The GTEx Project. *Biopreservation Biobanking* 13,311-319 (2015) .

[1013] 44.Xu,G.et al.RNA CoMPASS:a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PloS One* 9,e89445(2014) .

[1014] 45.Andrealta,M.&Nielsen,M.Gapped sequence alignment using artificial neural networks:application to the MHC class I system. *Bioinforma,Oxf.Engl.* (2015).doi:10.1093/bioinformatics/btv639

[1015] 46.Jorgensen,K.W.,Rasmussen,M.,Buus,S.&Nielsen,M.NetMHCstab- predicting stability of peptide-MHC-I complexes;impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* 141,18-26(2014) .

[1016] 47.Larsen, M.V.et al.An integrative approach to CTL epitope prediction:a combined algorithm integrating MHC class I binding,TAP transport efficiency,and proteasomal cleavage predictions. *Eur.J.Immunol.* 35,2295-2303

(2005) .

[1017] 48.cytotoxic T-cell epitopes:insights obtained from improved predictions of proteasomal cleavage.Immunogenetics 57,33-41(2005) .

[1018] 49.Boisvert,F.-M.et al.A Quantitative Spatial Proteomics Analysis of Proteome Turnover in Human Cells.Mol.Cell.Proteomics 11,M111.011429-M111.011429(2012) .

[1019] 50.Duan,F.et al.Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.J.Exp.Med.211,2231-2248(2014) .

[1020] 51.Janeway's Immunobiology:9780815345312;Medicine&Health Science Book@Amazon.com.at<<http://www.amazon.com/Janeways-Immunobiology-Kenneth-Murphy/dp/0815345313>>

[1021] 52.Calis,J.J.A.et al.Properties of MHC Class I presented peptides That Enhance Immunogenicity.PLoS Comput.Biol.9,e1003266(2013) .

[1022] 53.Zhang,J.et al.Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing.Science 346,256-259 (2014)

[1023] 54.Walter,M.J.et al.Clonal architecture of secondary acute myeloid leukemia.N.Engl.J.Med.366,1090-1098(2012) .

[1024] 55.Hunt DF,Henderson RA,Shabanowitz J,Sakaguchi K,Michel H,Sevilir N, Cox AL,Appella E,Engelhard VH.Chalaclerization of peptides bound to the class I MHC molecule HLA-A2.1by mass spectrometry.Science 1992.255:1261-1263.

[1025] 56.Zarling AL,Polefrone JM.Evans AM,Mikesh LM,Shabanowitz J,Lewis ST, Engelhard VH,Hunt DF.Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy.Proc Natl Acad Sci U S A.2006 Oct 3;103 (40):14889-94.

[1026] 57.Bassani-Sternberg M,Pletscher-Frankild S,Jensen LJ,Mann M.Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.Mol Cell Proteomics.2015 Mar;14(3):658-73.doi:10.1074/mcp.M114.042812.

[1027] 58.Abelin JG,Trantham PD,Penny SA,Patterson AM,Ward ST,Hildebrand WH, Cobbold M,Bai DL,Shabanowitz J,Hunt DF.Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry.Nat Protoc.2015 Sep;10(9):1308-18.doi:10.1038/nprot.2015.086.Epub 2015 Aug 6

[1028] 59.Barnstable CJ,Bodmer WE,Brown G,Galfre G,Milstein C,Williams AF, Ziegler A.Production of monoclonal antibodies to group A erythrocytes,HLA and other human cell surface antigens-new tools for genetic analysis.Cell.1978 May;14(1):9-20.

[1029] 60.Goldman JM,Hibbin J,Kearney L,Orchard K,Th'ng KH.HLA-DR monoclonal

antibodies inhibit the proliferation of normal and chronic granulocytic leukaemia myeloid progenitor cells.Br J Haematol.1982 Nov;52(3):411-20.

[1030] 61.Eng JK,Jahan TA,Hoopmann MR.Comet:an open-source MS/MS sequence database search tool.Proteomics.2013 Jan;13(1):22-4.doi:10.1002/pmic.201200439.Epub 2012 Dec 4.

[1031] 62.Eng JK,Hoopmann MR,Jahan TA,Egertson ID.Noble WS,MacCoss MJ.A deeper look into Comet-implementation and features.J Am Soc Mass Spectrom.2015 Nov;26(11):1865-74.doi:10.1007/s13361-015-1179-x.Epub 2015 Jun 27.

[1032] 63.LukasKall,Jesse Canterbury,Jason Weston,William Stafford Noble and Michael J.MacCoss.Semi-supervised learning for peptide identification from shotgun proteomics datasets.Nature Methods 4:923-925,November 2007

[1033] 64.LukasKall,John D.Storey,Michael J.MacCoss and William Stafford Noble.Assigning confidence measures to peptides identified by tandem mass spectrometry.Journal of Proteome Research,7(1):29-34,January 2008

[1034] 65.LukasKall,John D.Storey and William Stafford Noble.Nonparametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry.Bioinformatics,24(16):i42-i48,August 2008

[1035] 66.Bo Li and C.olin N.Dewey.RSEM:accurate transcript quantification from RNA-Seq data with or without a referenfe genome.BMC Bioinformatics,12:323,August 2011

[1036] 67.Hillary Pearson,Tariq Daouda,Diana Paola Granados,Chantal Durette, Eric Bonneil,Mathieu Courcelles,Anja Rodenbrock,Jean-Philippe Laverdure, CarolineCôté,Sylvie Mader,Sebastien Lemieux,Pierre Thibault,and Claude Perreault.MHC class I-associated peptides derive from selective regions of the human genome.The Journal of Clinical Investigation,2016,

[1037] 68.Juliane Liepe,Fabio Marino,John Sidney,Anita Jeko,Daniel E.Bunting,Alessandro Sette,Peter M.Kloetzel,Michael P.H.Stumpf,Albert J.R.Heck,Michele Mishto.A large fraction of HLA class I ligands are proteasome-generated spliced peptides.Science,21,October 2016.

[1038] 69.Mommen GP.,Marino,F.,Meiring HD.,Poelen,MC.,van Gaans-van den Brink,JA.,Mohammed S.,Heck AJ.,and van Els CA.Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR)Ligandome Proceeds Via High Specificity.Mol Cell Proteomics 15(4):1412-1423,April 2016.

[1039] 70.Sebastian Kreiter,Mathias Vormehr,Niels van de Roemer,Mustafa Diken,MartinLöwer,Jan Diekmann,Sebastian Boegel,BarbaraSchrörs,Fulvia Vascotto, John C.Castle,Arbel D.Tadmor,Stephen P.Schoenberger,Christoph Huber,ÖzlemTüreci,and Ugur Sahin Mutant MHC class II epitopes drive therapeutic immune

responses to cancer. *Nature* 520,692-696, April 2015.

[1040] 71. Tran E., Turcotte S., Gros A., Robbins P.F., Lu Y.C., Dudley M.E., Wunderlich J.R., Somerville R.P., Hogan K., Hinrichs C.S., Parkhurst M.R., Yang J.C., Rosenberg S.A. Cancer immunotherapy based on mutation-specific CD4+T-cells in a patient with epithelial cancer. *Science* 344(6184)641-645, May 2014.

[1041] 72. Andreatta M., Karosiene E., Rasmussen M., Stryhn A., Buus S., Nielsen M. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* 67(11-12)641-650, November 2015.

[1042] 73. Nielsen, M., Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 10:296, September 2009.

[1043] 74. Nielsen, M., Lundegaard, C., Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* 8:238, July 2007.

[1044] 75. Zhang, J., et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & Cellular Proteomics*. 11(4):1-8. 1/2/2012.

[1045] 76. Snyder, A. et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371, 2189-2199 (2014).

[1046] 77. Rizvi, N.A. et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124-128 (2015).

[1047] 78. Gubin, M.M., Artyomov, M.N., Mardis, E.R. & Schreiber, R.D. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J. Clin. Invest.* 125, 3413-3421 (2015).

[1048] 79. Schumacher, T.N. & Schreiber, R.D. Neoantigens in cancer immunotherapy. *Science* 348, 69-74 (2015).

[1049] 80. Carreno, B.M. et al. Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T-cells. *Science* 348, 803-808 (2015).

[1050] 81. Ott, P.A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217-221 (2017).

[1051] 82. Sahin, U. et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222-226 (2017).

[1052] 83. Tran, E. et al. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *N. Engl. J. Med.* 375, 2255-2262 (2016).

[1053] 84. Gros, A. et al. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat. Med.* 22, 433-438

- (2016) .
- [1054] 85.The problem with neoantigen prediction.Nat.Biotechnol.35,97-97 (2017) .
- [1055] 86.Vitiello,A.&Zanetti,M.Neoantigen prediction and the need for validation.Nat.Biotechnol.35,815-817(2017) .
- [1056] 87.Bassani-Sternberg,M.,Pletscher-Frankild,S.,Jensen,L.J.&Mann,M.Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation.Mol.Cell.Proteomics MCP 14,658-673(2015) .
- [1057] 88.Vita,R.et al.The immune epitope database (IEDB) 3.0.Nucleic Acids Res.43,D405-412(2015) .
- [1058] 89.Andreatta,M.&Nielsen,M.Gapped sequence alignment using artificial neural networks:application to the MHC class I system.Bioinforma.Oxf.Engl.32, 511-517(2016) .
- [1059] 90.O'Donnell,T.J.et al.MHCflurry:Open-Source Class I MHC Binding Affinity Prediction.Cell Syst.(2018) .doi:10.1016/j.cels.2018.05.014
- [1060] 91.Bassani-Sternberg,M.et al.Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry.Nat.Commun.7,13404(2016) .
- [1061] 92.Abelin,J.G.et al.Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction.Immunity 46,315-326(2017) .
- [1062] 93.Yadav,M.et al.Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing.Nature 515,572-576(2014) .
- [1063] 94.Stranzl,T.,Larsen,M.V.,Lundegaard,C.&Nielsen,M.NetCTLpan:pan-specific MHC class I pathway epitope predictions.Immunogenetics 62,357-368 (2010) .
- [1064] 95.Bentzen,A.K.et al.Large-scale detection of antigen-specific T-cells using peptide-MHC-I multimers labeled with DNA barcodes.Nat.Biotechnol.34,1037-1045(2016) .
- [1065] 96.Tran,E.et al.Immunogenicity of somatic mutations in human gastrointestinal cancers.Science 350,1387-1390(2015) .
- [1066] 97.Stronen,E.et al.Targeting of cancer neoantigens with donor-derived T-cell receptor repertoires.Science 352,1337-1341(2016) .
- [1067] 98.Trolle,T.et al.The Length Distribution of Class I-Restricted T-cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference.J.Immunol.Baltim.Md 1950 196,1480-1487(2016) .
- [1068] 99.Di Marco,M.et al.Unveiling the Peptide Motifs of HLA-C and HLA-G from Naturally Presented Peptides and Generation of Binding Prediction

Matrices.J.Immunol.Baltim.Md 1950199,2639-2651(2017) .

[1069] 100.Goodfellow,L.,Bengio,Y.&Courville,A.Deep Learning. (MIT Press, 2016) .

[1070] 101.Sette,A.et al.The relationship between class I binding affinity and immunogenicity of potential cytotoxic T-cell epitopes.J.Immunol.Baltim.Md 1950 153,5586-5592(1994) .

[1071] 102.Fortier,M.-H.et al.The MHC class I peptide repertoire is molded by the transcriptome.J.Exp.Med.205,595-610(2008) .

[1072] 103.Pearson,H.et al.MHC class I-associated peptides derive from selective regions of the human genome.J.Clin.Invest.126,4690-4701(2016) .

[1073] 104.Bassani-Sternberg,M.et al.Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allosteric regulating HLA specificity.PLoS Comput.Biol.13,e1005725(2017) .

[1074] 105.Andreatta,M.,Lund,O.&Nielsen,M.Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach.Bioinforma.Oxf.Engl.29,8-14(2013) .

[1075] 106.Andreatta,M.,Alvarez,B.&Nielsen,M.GibbsCluster:unsupervised clustering and alignment of peptide sequences.Nucleic Acids Res. (2017) .doi: 10.1093/nar/gkx248

[1076] 107.Gros,A.et al.Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients.Nat.Med.22,433-438 (2016) .

[1077] 108.Zacharakis,N.et al.Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer.Nat.Med.24,724-730(2018) .

[1078] 109.Chudley,L.et al.Harmonisation of short-term in vitro culture for the expansion of antigen-specific CD8+T-cells with detection by ELISPOT and HLA-multimer staining.Cancer Immunol.Immunother.63,1199-1211(2014) .

[1079] 110.Van Allen,E.M.et al.Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.Science 350,207-211(2015) .

[1080] 111.Anagnostou,V.et al.Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non-Small Cell Lung Cancer.Cancer Discov.7,264-276(2017) .

[1081] 112.Carreno,B.M.et al.Cancer immunotherapy.A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T-cells.Science 348,803-808(2015) .

[1082] 113.Stevanovic,S.et al.Landscape of immunogenic tumor antigens in successful immunotherapy of virally induced epithelial cancer.Science 356, 200-205(2017) .

- [1083] 114.Pasetto,A.et al.Tumor-and Neoantigen-Reactive T-cell Receptors Can Be Identified Based on Their Frequency in Fresh Tumor.Cancer Immunol.Res.4,734-743(2016) .
- [1084] 115.Gillette,M.A.&Carr,S.A.Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry.Nat.Methods 10,28-34 (2013) .
- [1085] 116.Boegel,S.,Löwer,M.,Bukur,T.,Sahin,U.&Castle,J.C.A catalog of HLA type,HLA expression,and neo-epitope candidates in human cancer cell lines.Oncoimmunology 3,e954893(2014) .
- [1086] 117.Johnson,D.B.et al.Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy.Nat.Commun.7,10582(2016) .
- [1087] 118.Robbins,P.F.et al.A Pilot Trial Using Lymphocytes Genetically Engineered with an NY-ESO-1-Reactive T-cell Receptor:Long-term Follow-up and Correlates with Response.Clin.Cancer Res.21,1019-1027(2015) .
- [1088] 119.Snyder,A.et al.Genetic basis for clinical response to CTLA-4 blockade in melanoma.N.Engl.J.Med.371,2189-2199(2014) .
- [1089] 120.Calis,J.J.A.et al.Properties of MHC class I presented peptides that enhance immunogenicity.PLoS Comput.Biol.9,e1003266(2013) .
- [1090] 121.Duan,F.et al.Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity.J.Exp.Med.211,2231-2248(2014) .
- [1091] 122.Glanville,J.et al.Identifying specificity groups in the T-cell receptor repertoire.Nature 547,94-98(2017) .
- [1092] 123.Dash,P.et al.Quantifiable predictive features define epitope-specific T-cell receptor repertoires.Nature 547,89-93(2017) .
- [1093] 124.Hunt,D.F.et al.Pillars article:Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry.Science 1992.255: 1261-1263.J.Immunol.Baltim.Md 1950179,2669-2671(2007) .
- [1094] 125.Zarling,A.L.et al.Identification of class I MHC-associated phosphopeptides as targets for cancer immunotherapy.Proc.Natl.Acad.Sci.U.S.A .103,14889-14894(2006) .
- [1095] 126.Abelin,J.G.et al.Complementary IMAC enrichment methods for HLA-associated phosphopeptide identification by mass spectrometry.Nat.Protoc.10, 1308-1318(2015) .
- [1096] 127.Barnstable,C.J.et al.Production of monoclonal antibodies to group A erythrocytes,HLA and other human cell surface antigens-new tools for genetic analysis.Cell 14,9-20(1978) .
- [1097] 128.Eng,J.K.,Jahan,T.A.&Hoopmann,M.R.Comet:an open-source MS/MS

sequence database search tool. *Proteomics* 13,22-24(2013) .

[1098] 129.Eng,J.K.et al.A deeper look into Comet--implementation and features. *J.Am.Soc.Mass Spectrom.*26,1865-1874(2015) .

[1099] 130.Kall,L.,Storey,J.D.,MacCoss,M.J.&Noble,W.S.Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J.Proteome Res.*7,29-34(2008) .

[1100] 131.Kall,L.,Storey,J.D.&Noble,W.S.Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. *Bioinforma.Oxf.Engl.*24,i42-48(2008) .

[1101] 132.Kall,L.,Canterbury,J.D.,Weston,J.,Noble,W.S.&MacCoss,M.J.Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat.Methods* 4,923-925(2007) .

[1102] 133.Li,B.&Dewey,C.N.RSEM:accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12,323(2011) .

[1103] 134.Chollet,F.&others.Keras. (2015) .

[1104] 135.Bastien,F.et al.Understanding the difficulty of training deep feedforward neural networks. *Proc.Thirteen.Int.Conf.Artif.Intell.Stat.*249-256 (2010) .

[1105] 136.Glorot,X.&Bengio,Y.Understanding the difficulty of training deep feedforward neural networks.in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249-256(2010) .

[1106] 137.Kingma,D.&Ba,J.Adam:A method for stochastic optimization. *ArXiv Prepr.ArXiv14126980*(2014) .

[1107] 138.Schneider,T.D.&Stephens,R.M.Sequence logos:a new way to display consensus sequences. *Nucleic Acids Res.*18,6097-6100(1990) .

[1108] 139.Rubinsteyn,A.,O'Donnell,T.,Damaraju,N.&Hammerbacher,J.Predicting Peptide-MHC Binding Affinities With Imputed Training Data. *biorxiv*(2016).doi: <https://doi.org/10.1101/054775>

[1109] 140.Tran,E.et al.Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350,1387-1390(2015) .

[1110] 141.Stronen,E.et al.Targeting of cancer neoantigens with donor-derived T-cell receptor repertoires. *Science* 352,1337-1341(2016) .

[1111] 142.Janetzki,S.,Cox,J.H.,Oden,N.&Ferrari,G.Standardization and validation issues of the ELISPOT assay. *Methods Mol.Biol.Clifton NJ* 302,51-86 (2005) .

[1112] 143.Janetzki,S.et al.Guidelines for the automated evaluation of Elispot assays. *Nat.Protoc.*10,1098-1115(2015) .

[1113] 144.Li,H.&Durbin,R.Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma.Oxf.Engl.*25,1754-1760(2009) .

- [1114] 145.DePristo,M.A.et al.A framework for variation discovery and genotyping using next-generation DNA sequencing data.Nat.Genet.43,491-498 (2011) .
- [1115] 146.Garrison,E.&Marth,G.Haplotype-based variant detection from short-read sequencing.arXiv(2012) .
- [1116] 147.Cingolani,P.et al.A program for annotating and predicting the effects of single nucleotide poly morphisms, SnpEff:SNPs in the genome of *Drosophila melanogaster* strain w1118;iso-2;iso-3.Fly(Austin)6,80-92(2012) .
- [1117] 148.Szolek,A.et al.OptiType:precision HLA typing from next-generation sequencing data.Bioinforma.Oxf.Engl.30,3310-3316(2014) .
- [1118] 149.Cibulskis,K.et al.Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.Nat.Biotechnol.31,213-219(2013) .
- [1119] 150.Scholz,E.M.et al.Human Leukocyte Antigen(HLA) -DRB1*15:01 and HLA-DRB5*01:01 Present Complementary Peptide Repertoires.Front.Immunol.8,984 (2017) .
- [1120] 151.Ooi,J.D.et al.Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T-cells.Nature 545,243-247(2017) .
- [1121] 152.Karosiene,E.et al.NetMHCpan-3.0,a common pan-specific MHC class II prediction method including all three human MHC class II isotypes,HLA-DR, HLA-DP and HLA-DQ.Immunogenetics 65,711-724(2013) .
- [1122] 153.Dudley ME,Gross CA,Langhan MM,et al.CD8+enriched“young”tumor infiltrating lymphocytes can mediate regression of metastatic melanoma.Clinical cancer research:an official journal of the Americam Association for Cancer Research.2010;16(24) :6122-6131.doi:10.1158/1078-0432.CCR-10-1297.
- [1123] 154.Dudley ME,Wunderlich JR,Shelton TE,Even J,Rosenberg SA.Generation of Tumor-Infiltrating Lymphocyte Cultures for Use in Adoptive Transfer Therapy for Melanoma Patients.Journal of immunotherapy(Hagerstown,Md:1997) .2003;26(4) :332-342.
- [1124] 155.Cohen CJ,Gartner JJ,Horovitz-Fried M,et al.Isolation of neoantigen-specific T cells from tumor and peripberal lymphocytes.The Journal of Clinical Investigation.2015;125(10) :3981-3991.doi:10.1172/JC182416.
- [1125] 156.Kelderman,S.,Heemskerk,B.,Fanchi,L.,Philips,D.,Toebes,M., Kvistborg,P.,Buuren,M.M.,Rooij,N.,Michels,S,Germeroth,L.,Haanen,J.B.and Schumacher,N.M. (2016),Antigen-specific TIL therapy for melanoma:A flexible platform for personalized cancer immunotherapy.Eur.J.Immunol.,46:1351-1360.doi:10.1002/eji.201545849.
- [1126] 157.Hall M,Liu H.Malafa M,et al.Expansion of tumor-infiltrating lymphocytes(TIL)from human pancreatic tumors.Journal for Immunotherapy of

Cancer.2016;4:61.doi:10.1186/s40425-016-0164-7.

[1127] 158.Briggs A,Goldfless S,Timberlake S,et al.Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion.bioRxiv.2017.doi.org/10.1101/134841.

[1128] 159.US Patent Application No.20160244825A1.

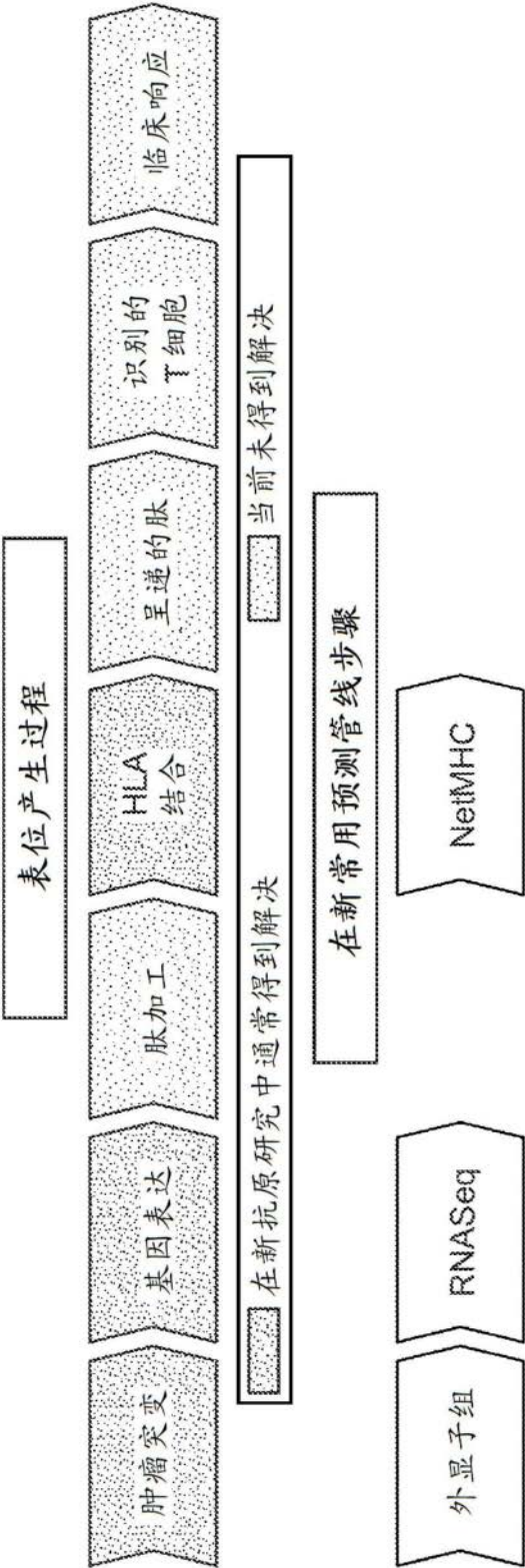


图1A

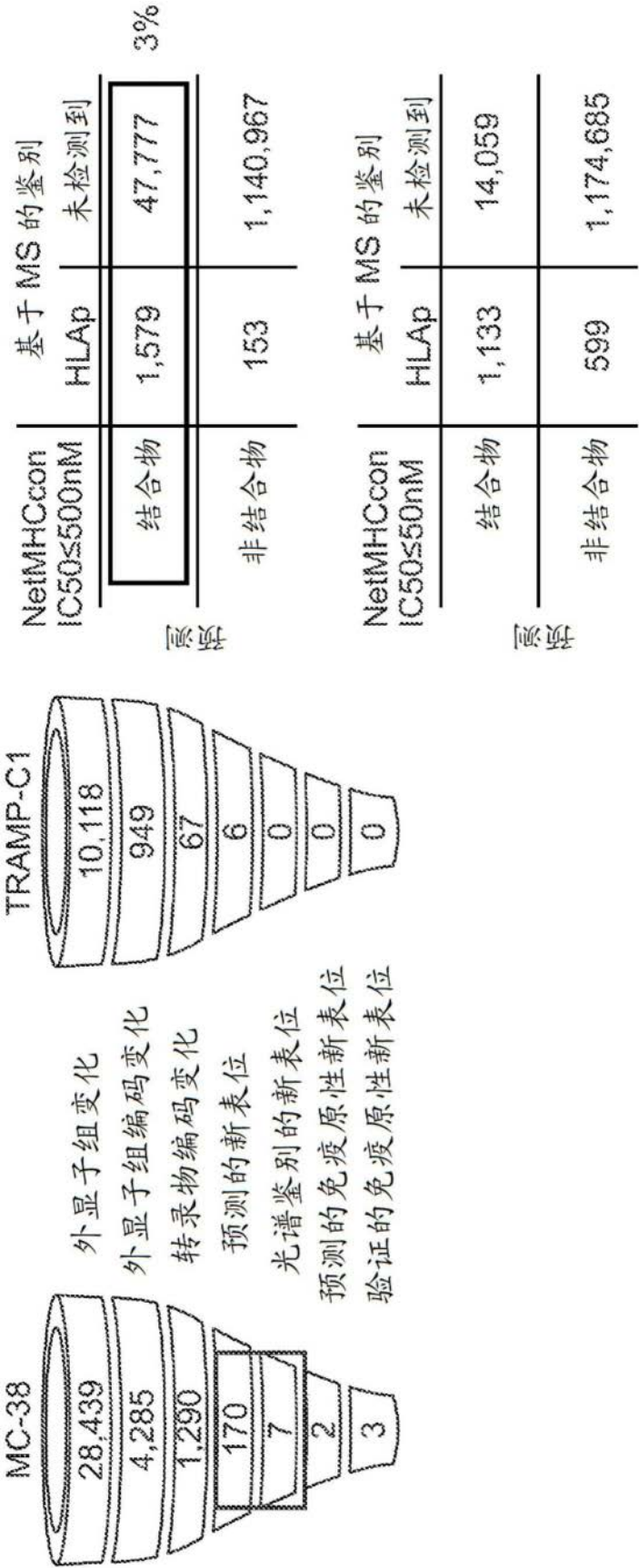
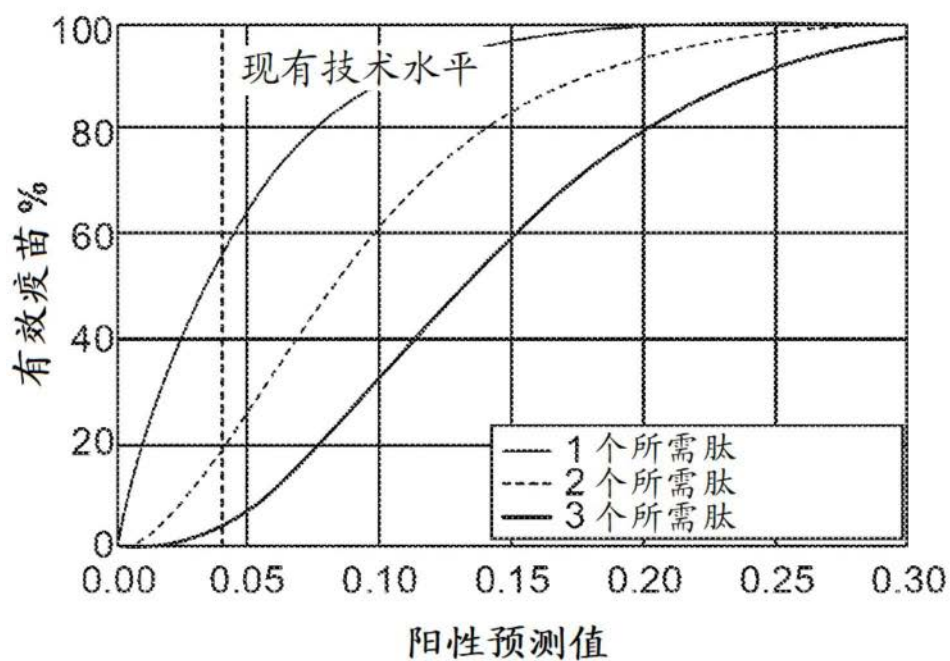


图1B

20 个肽/疫苗



10 个肽/疫苗

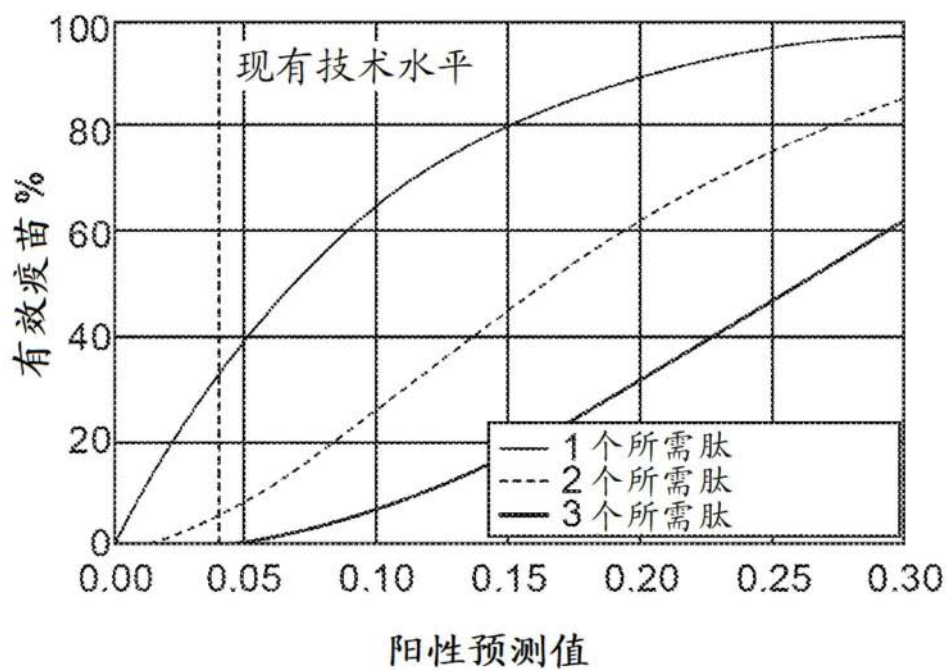


图1C

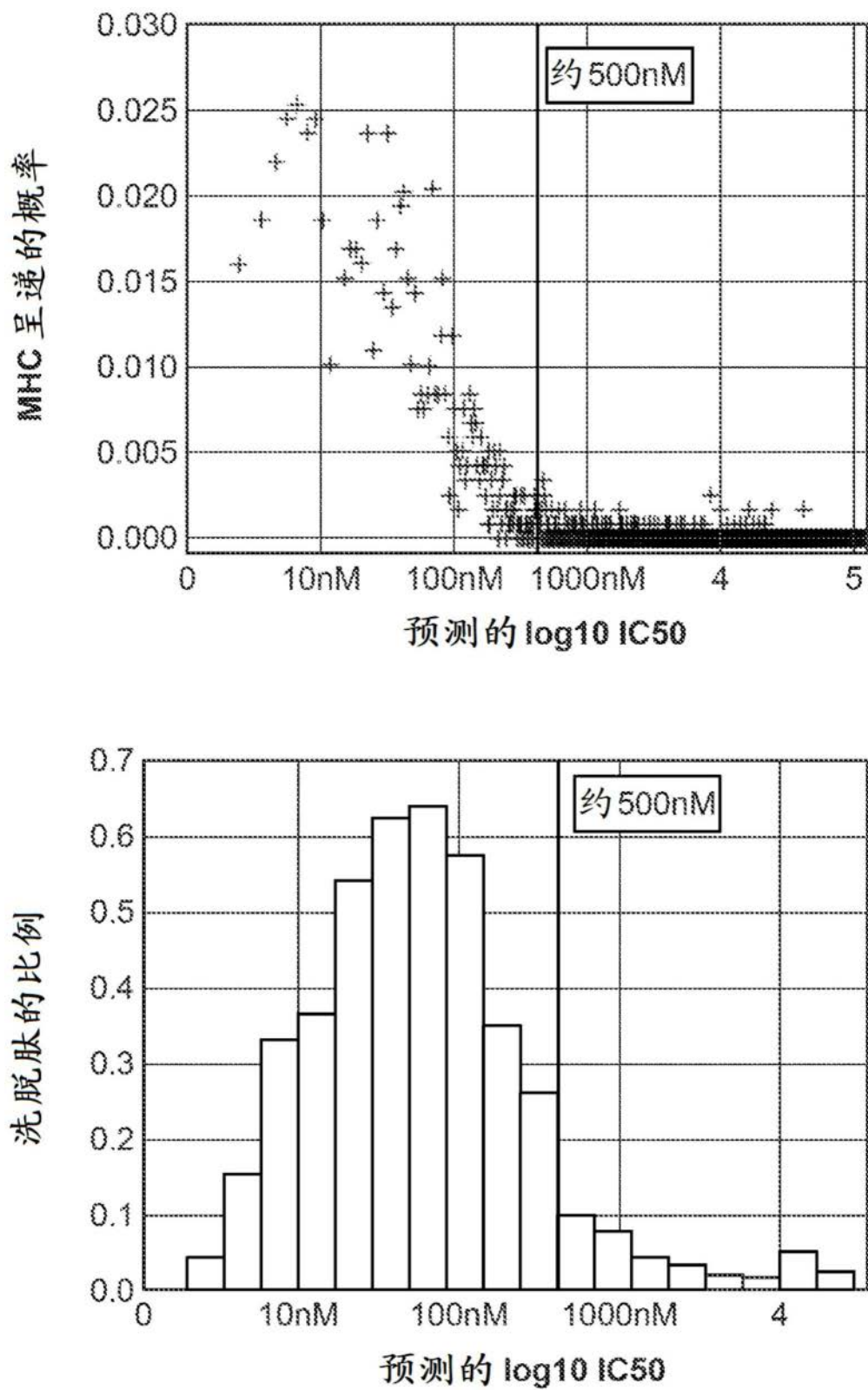


图1D

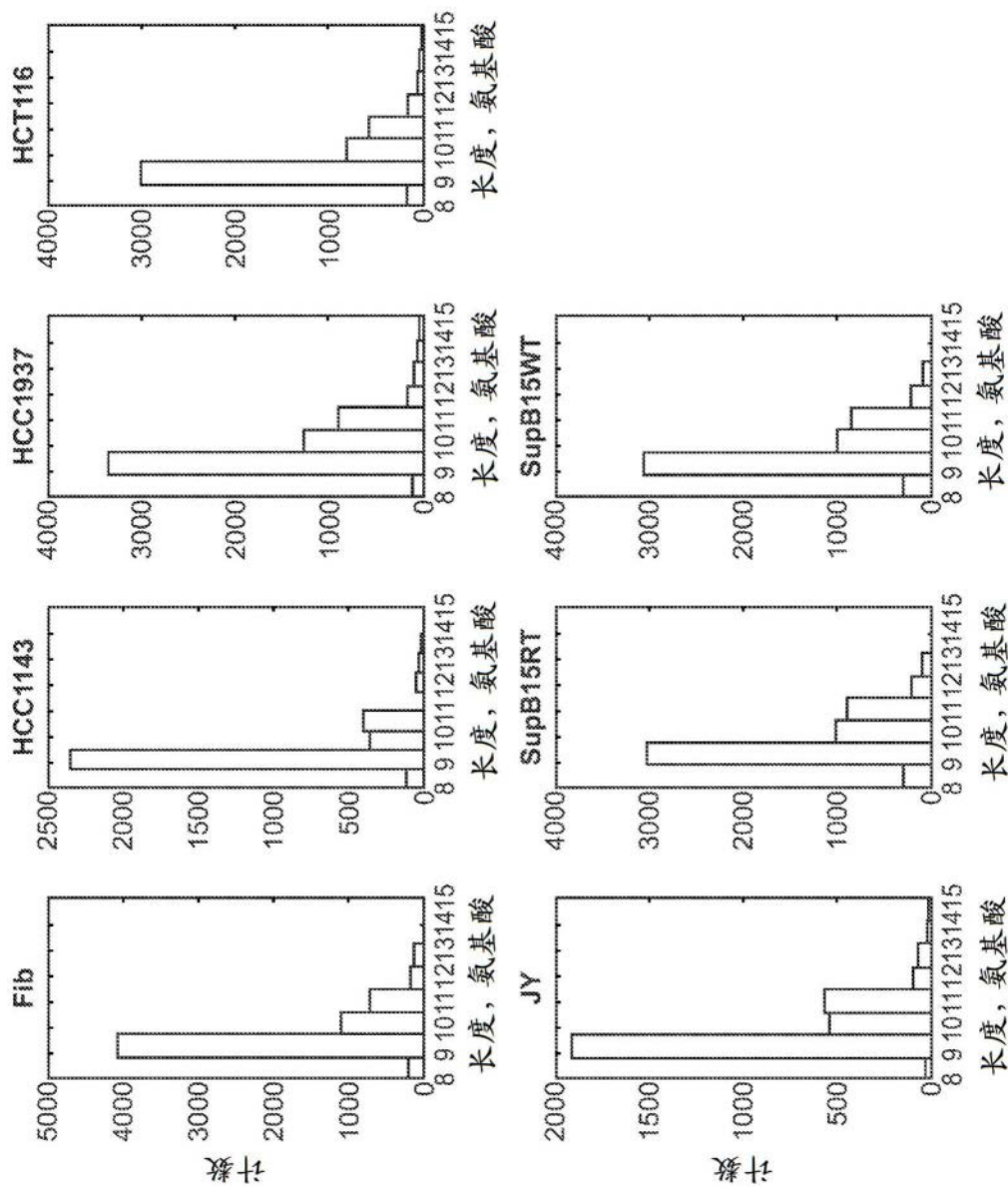


图1E

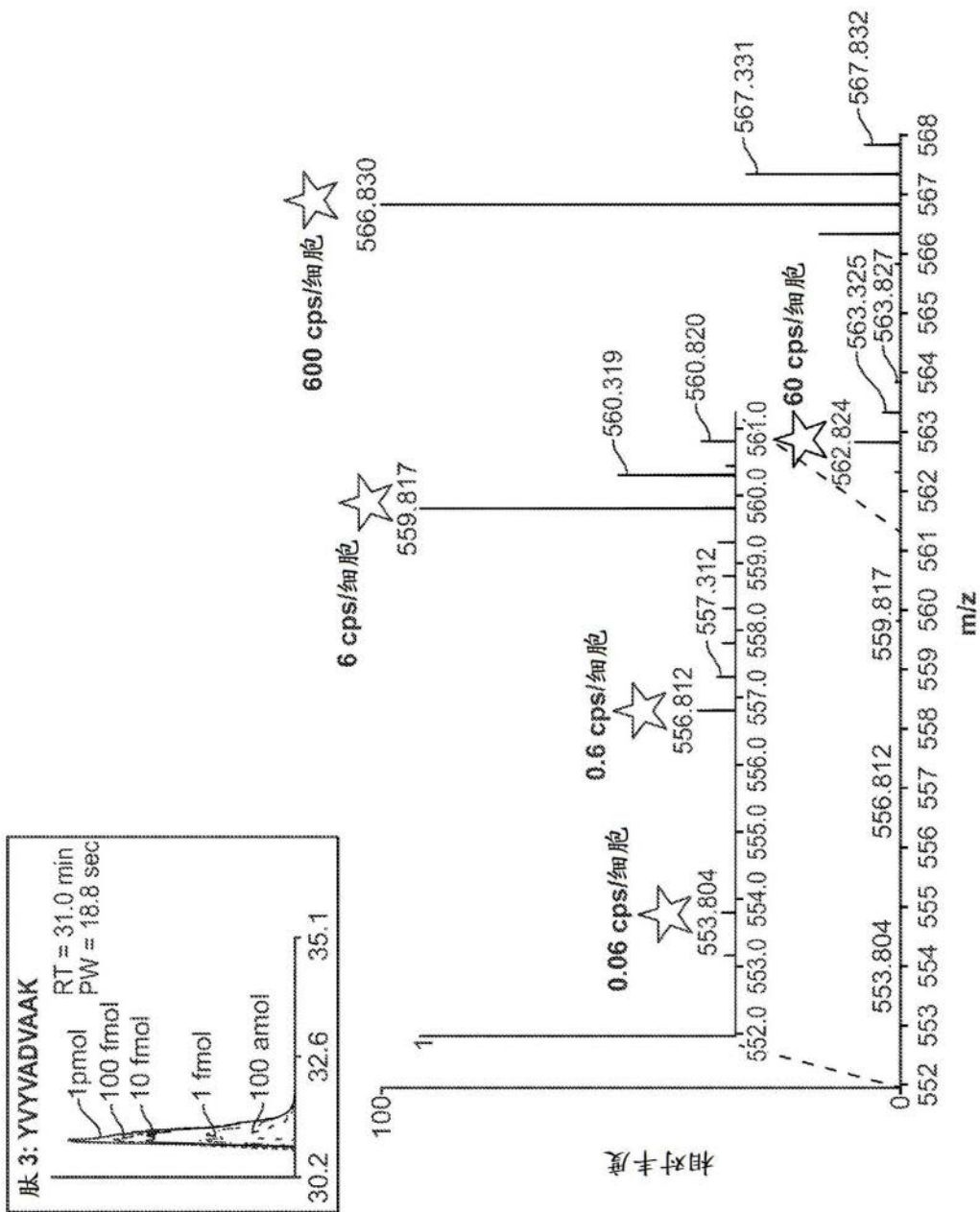


图1F

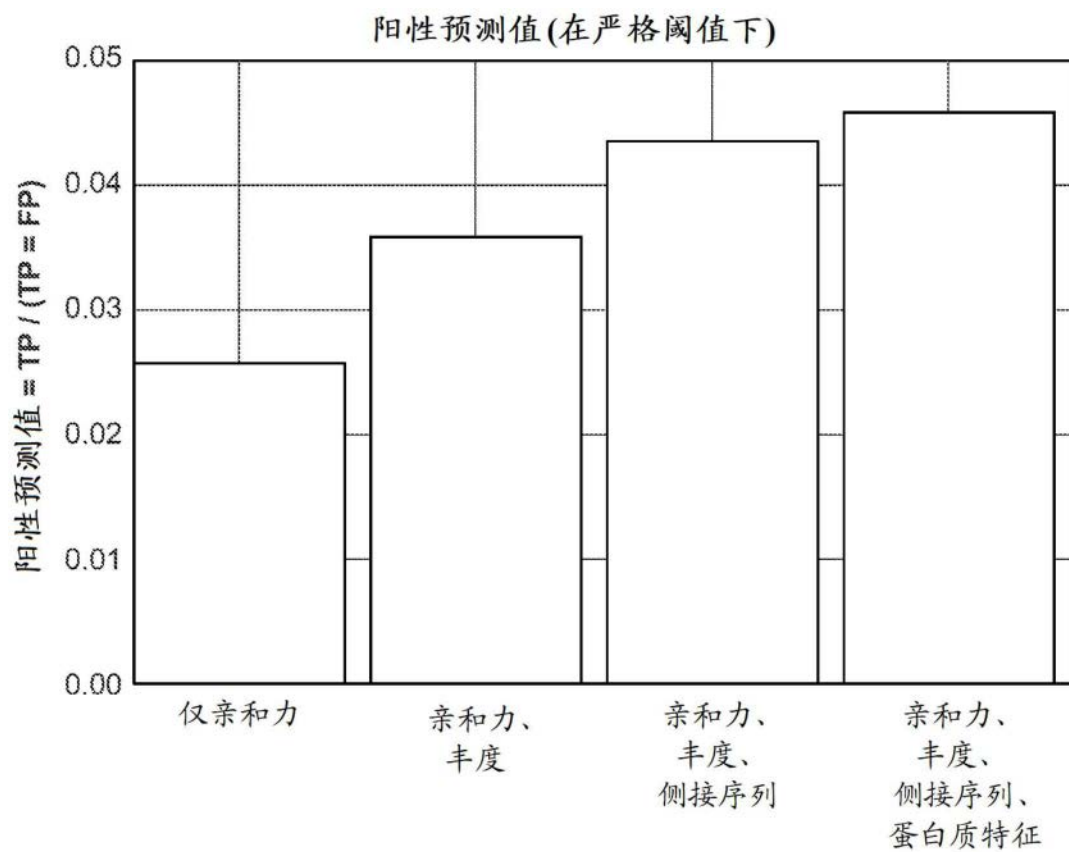


图1G

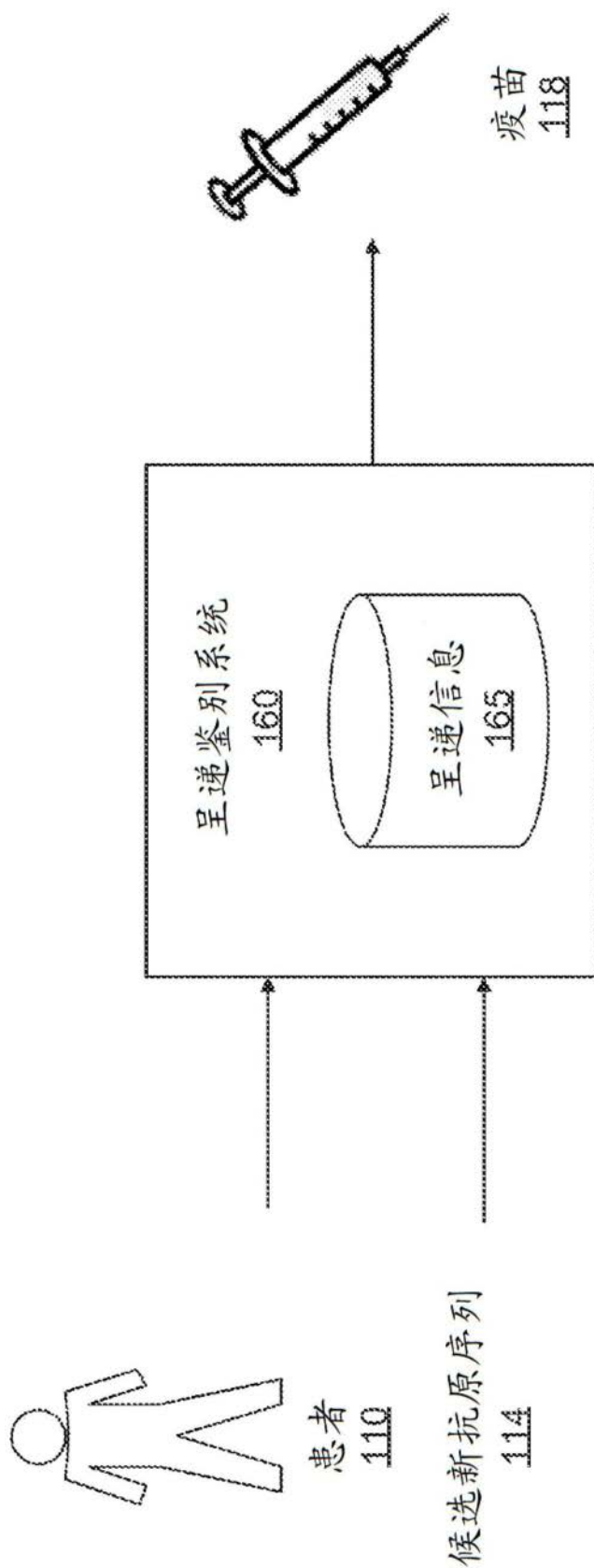


图2A

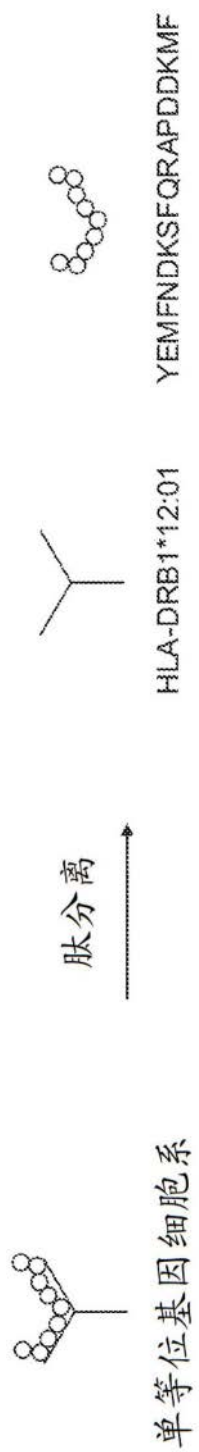


图2B

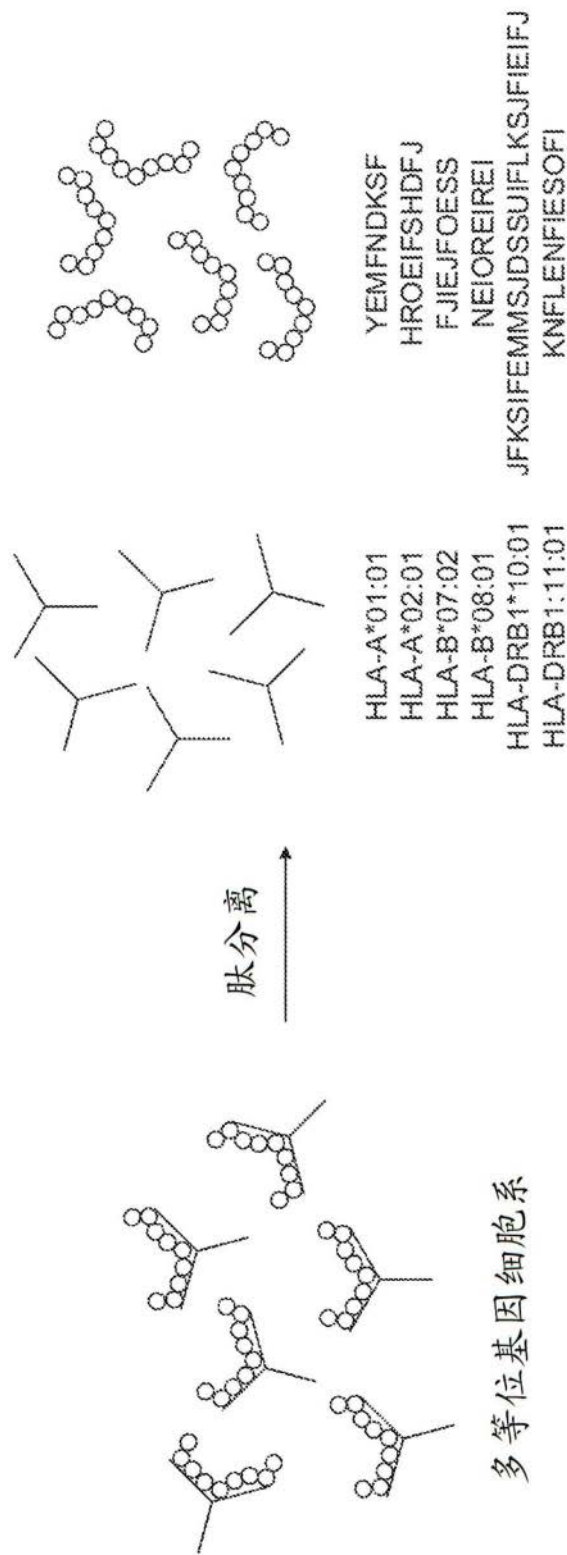


图2C

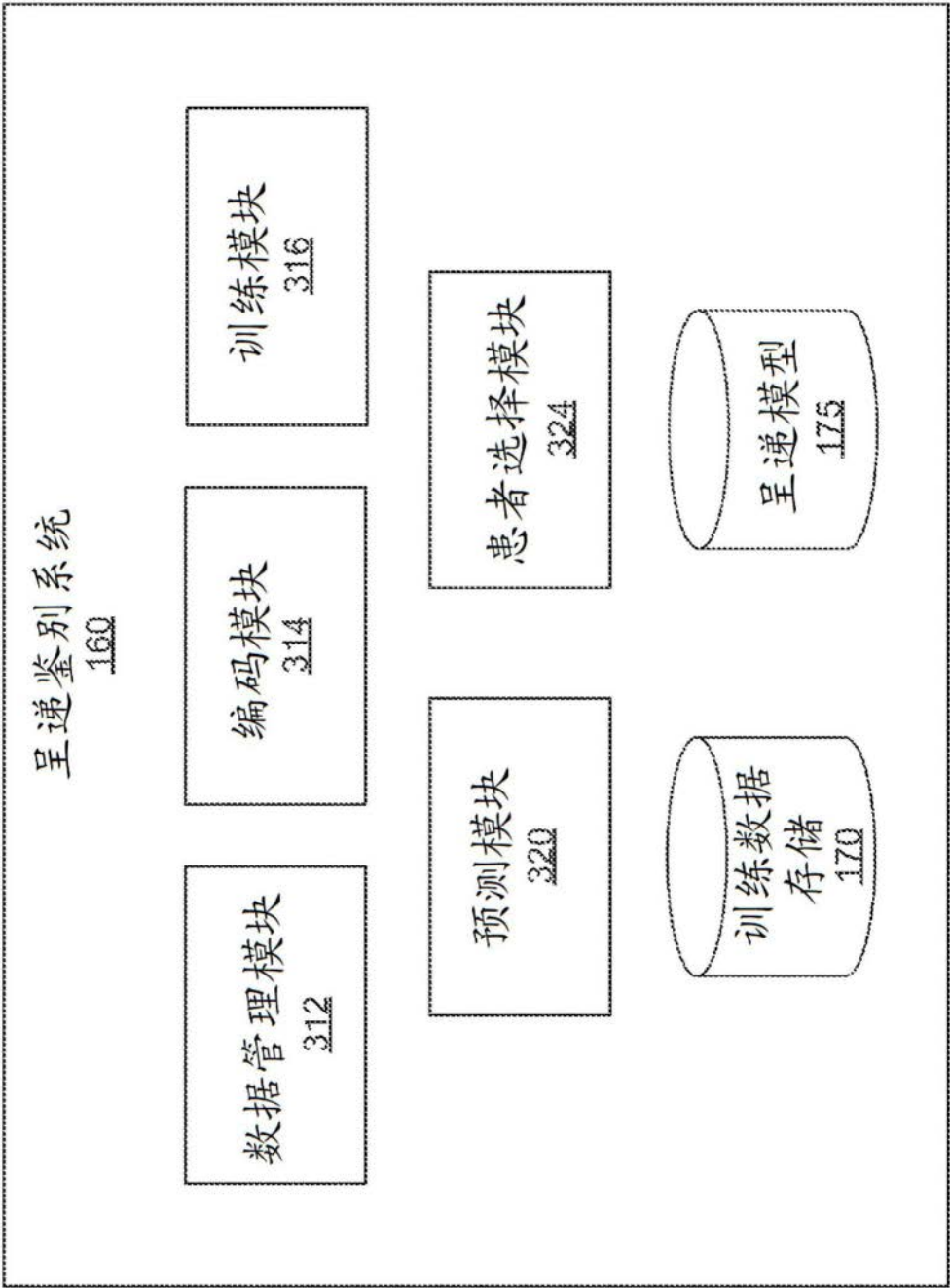


图3

训练数据
170A

依赖于等位基因 (x')				不依赖于等位基因 (w')		
肽序列 (p')	亲和力 (b'-nM)	稳定性 (s'-h)	等位基因 (a')	C 侧接序列 (c')	mRNA Q. (m'-TPM)	标记 (y')
QCEIOWAREFLKEIGJ	1000	1	HLA-DRB3:01:01	FJELFISBOSJFIE	10 ²	未呈递
FIEUHFWI	1500	15	HLA-C*01:03	FEGRKUOOI	10 ⁻³	呈递
FEWRHRJTRUJR	650	20	HLA-C*01:03	PJFIOEJOIJGEIO	10 ¹	呈递
QIEJOEIE	500	1	HLA-B*07:02	PJFIOEJOIJGEIO	1	呈递
	600	14	HLA-C*01:03			
	1200	7	HLA-A*01:01			

图4

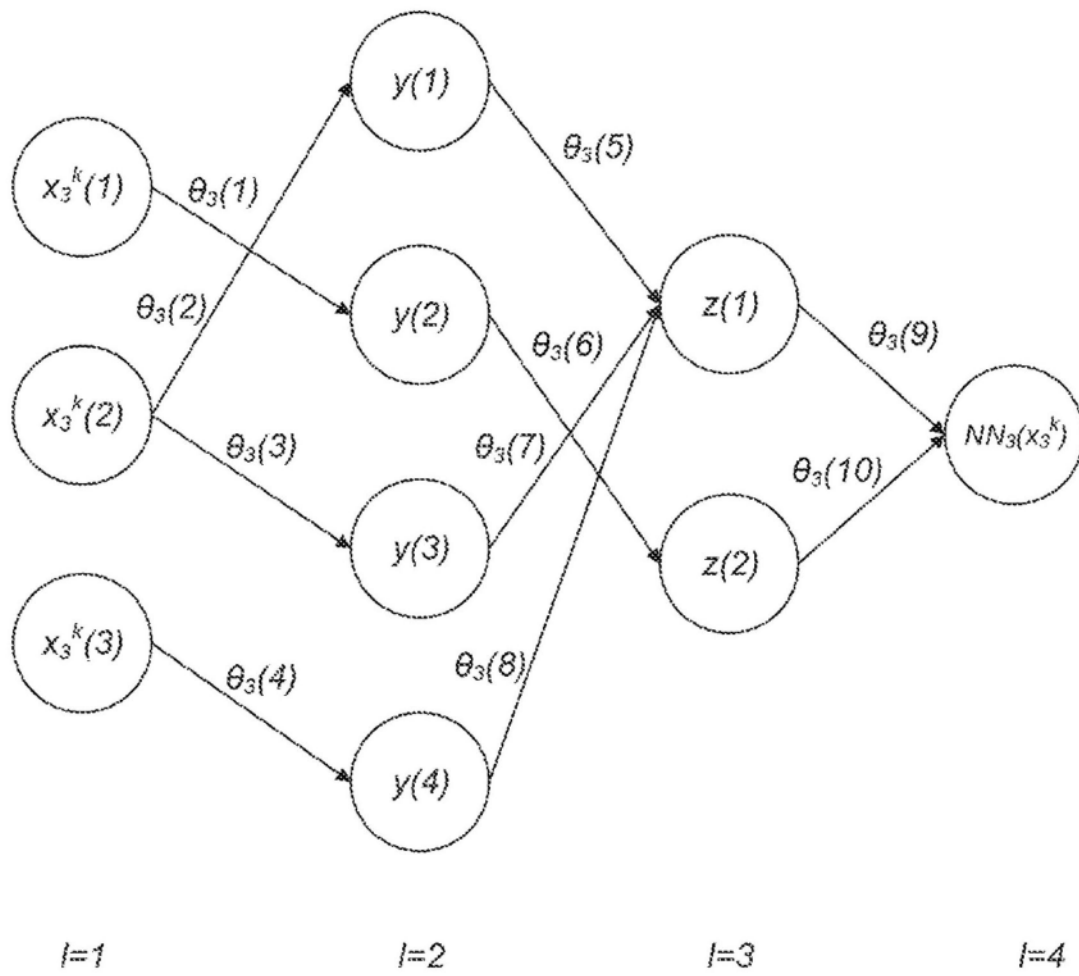


图5

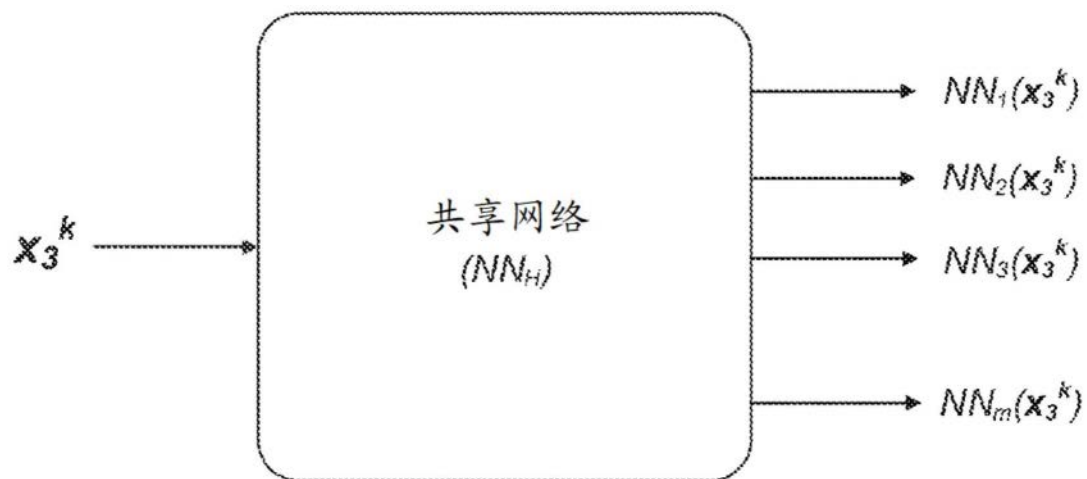


图6

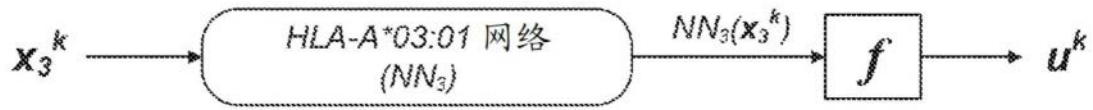


图7

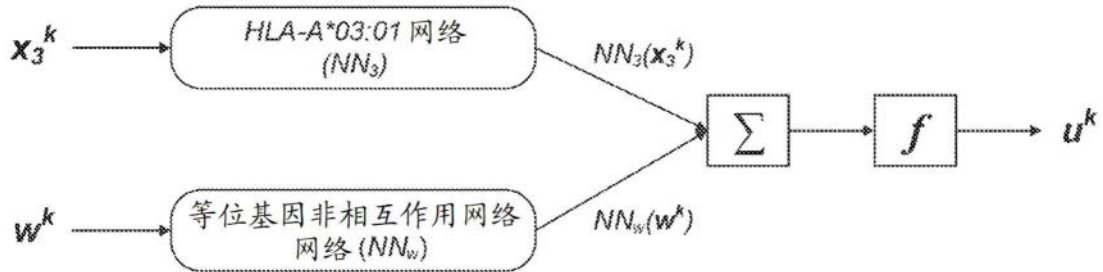


图8

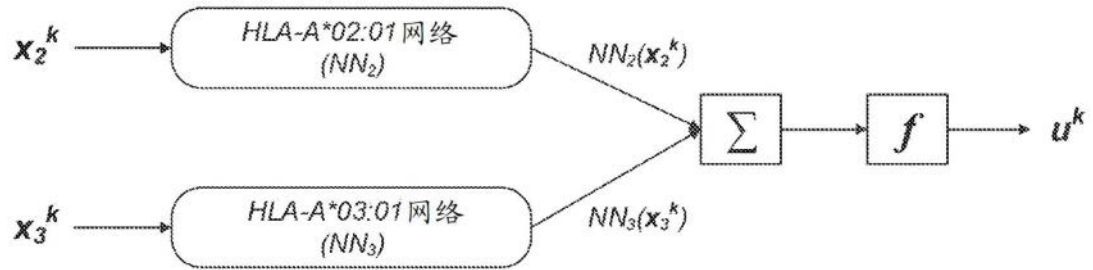


图9

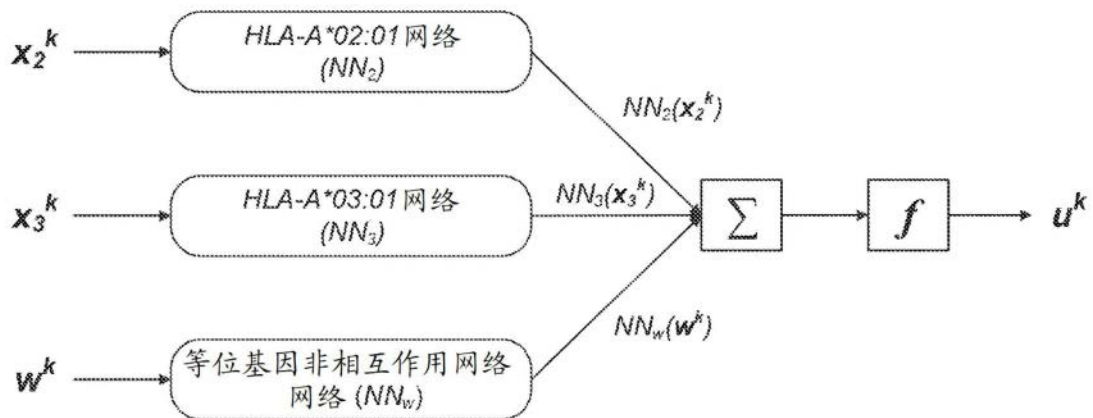


图10

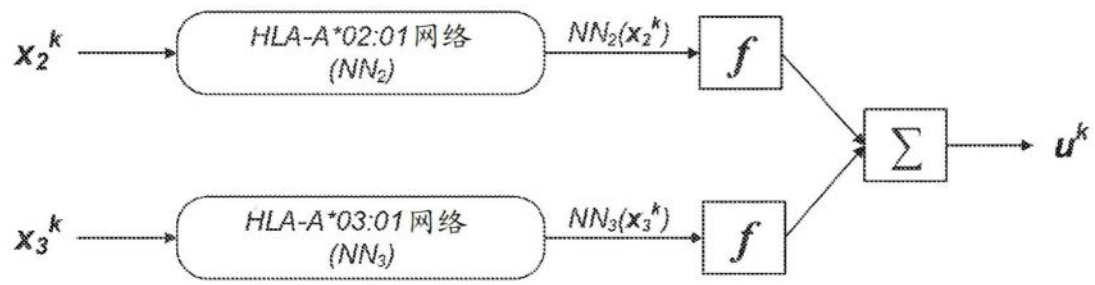


图11

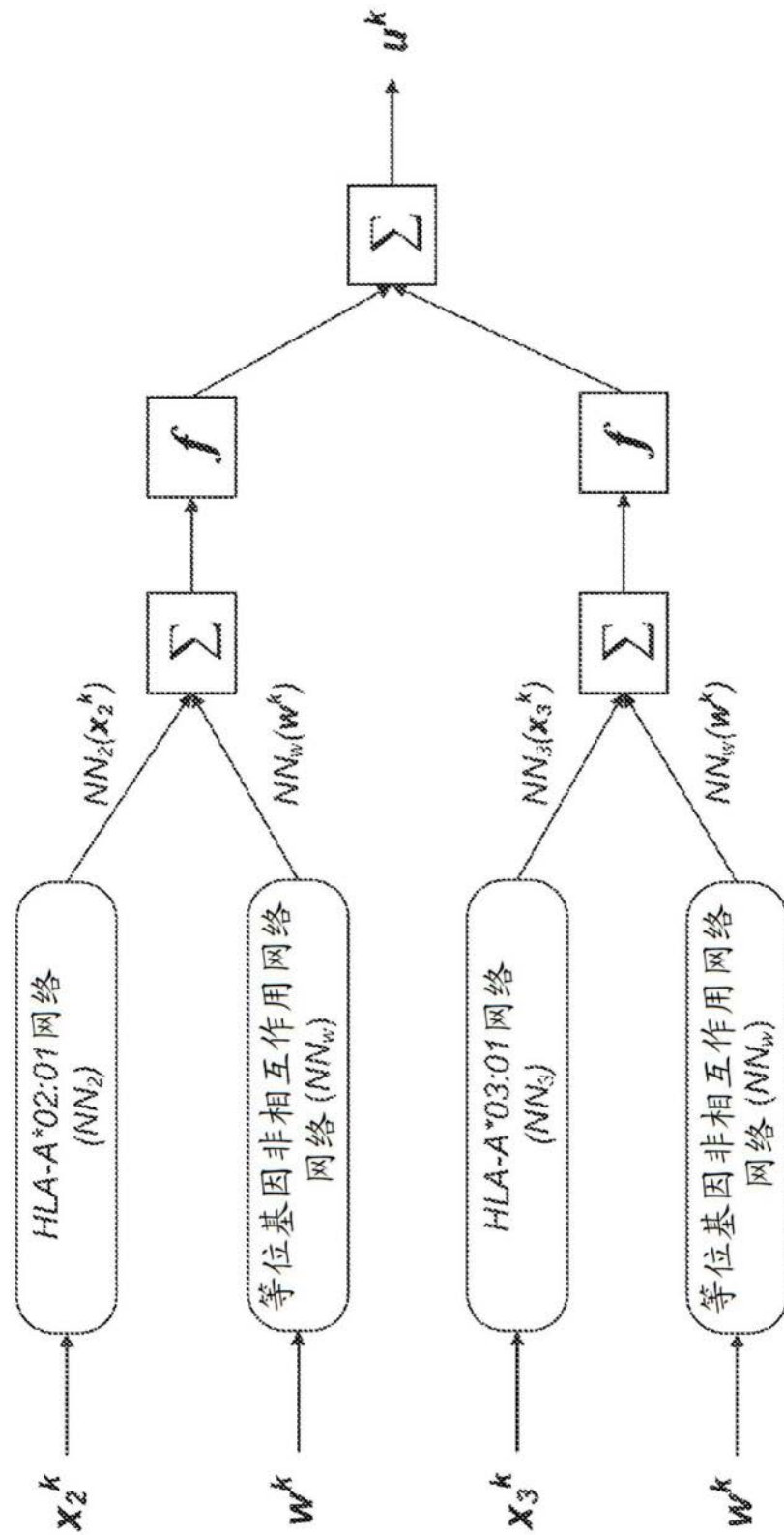


图12

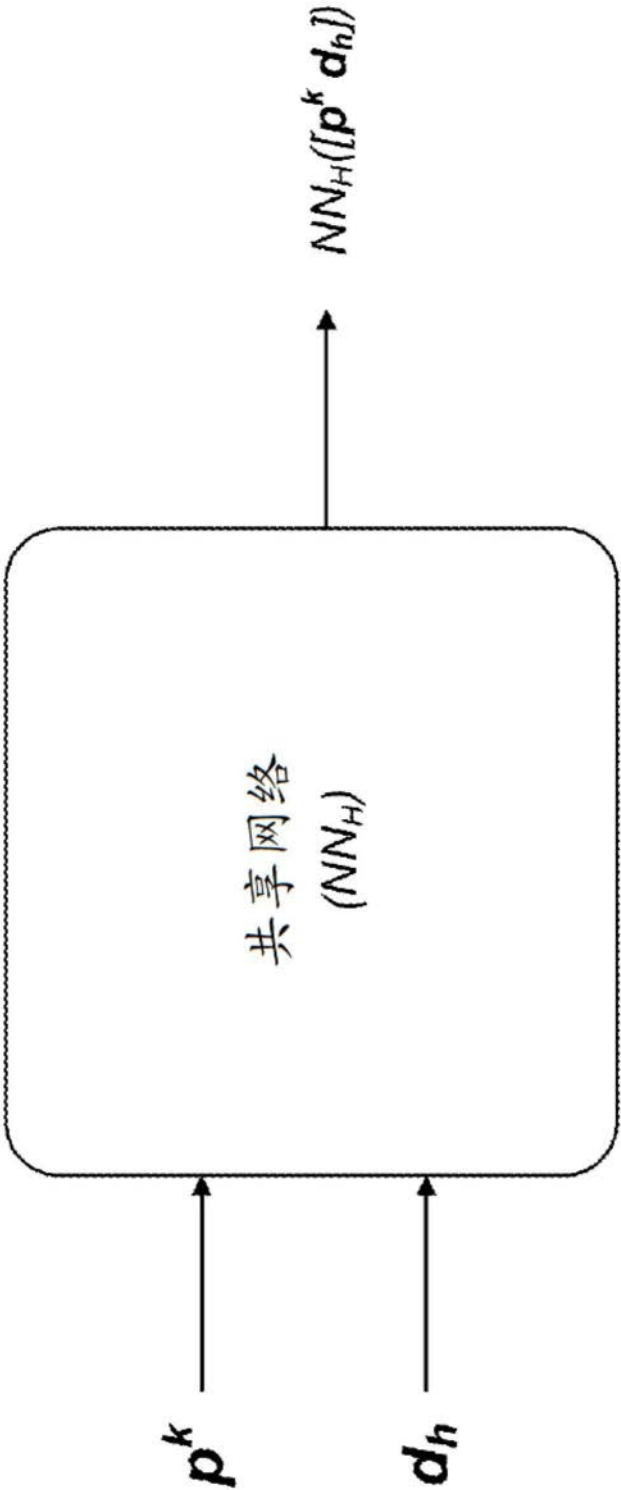


图13

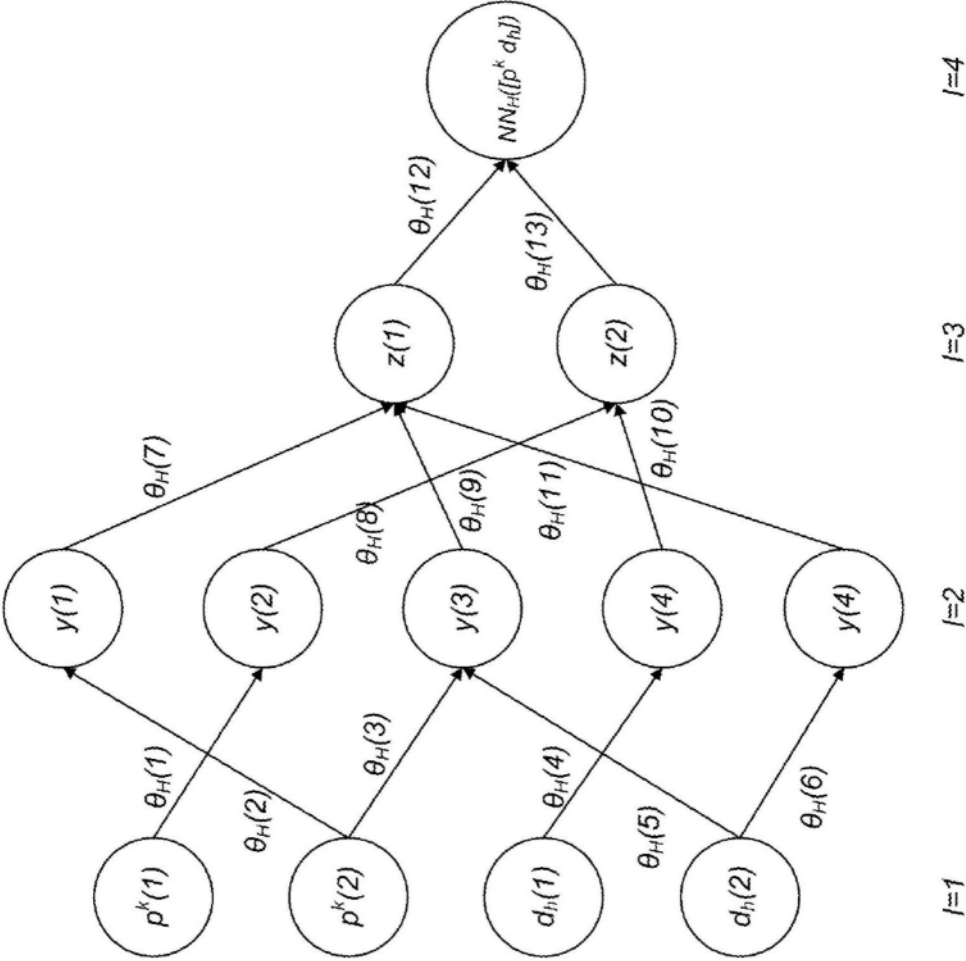


图14

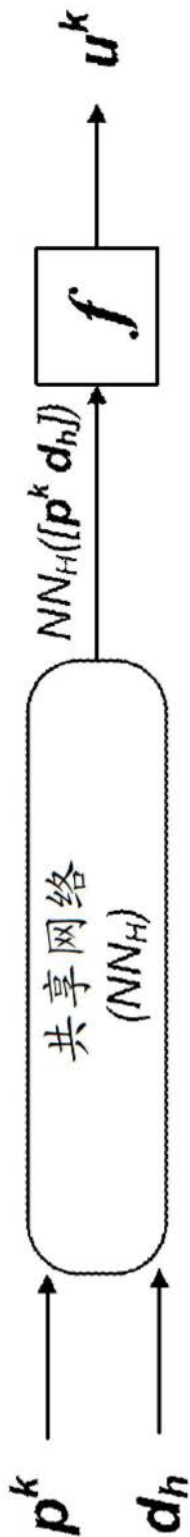


图15

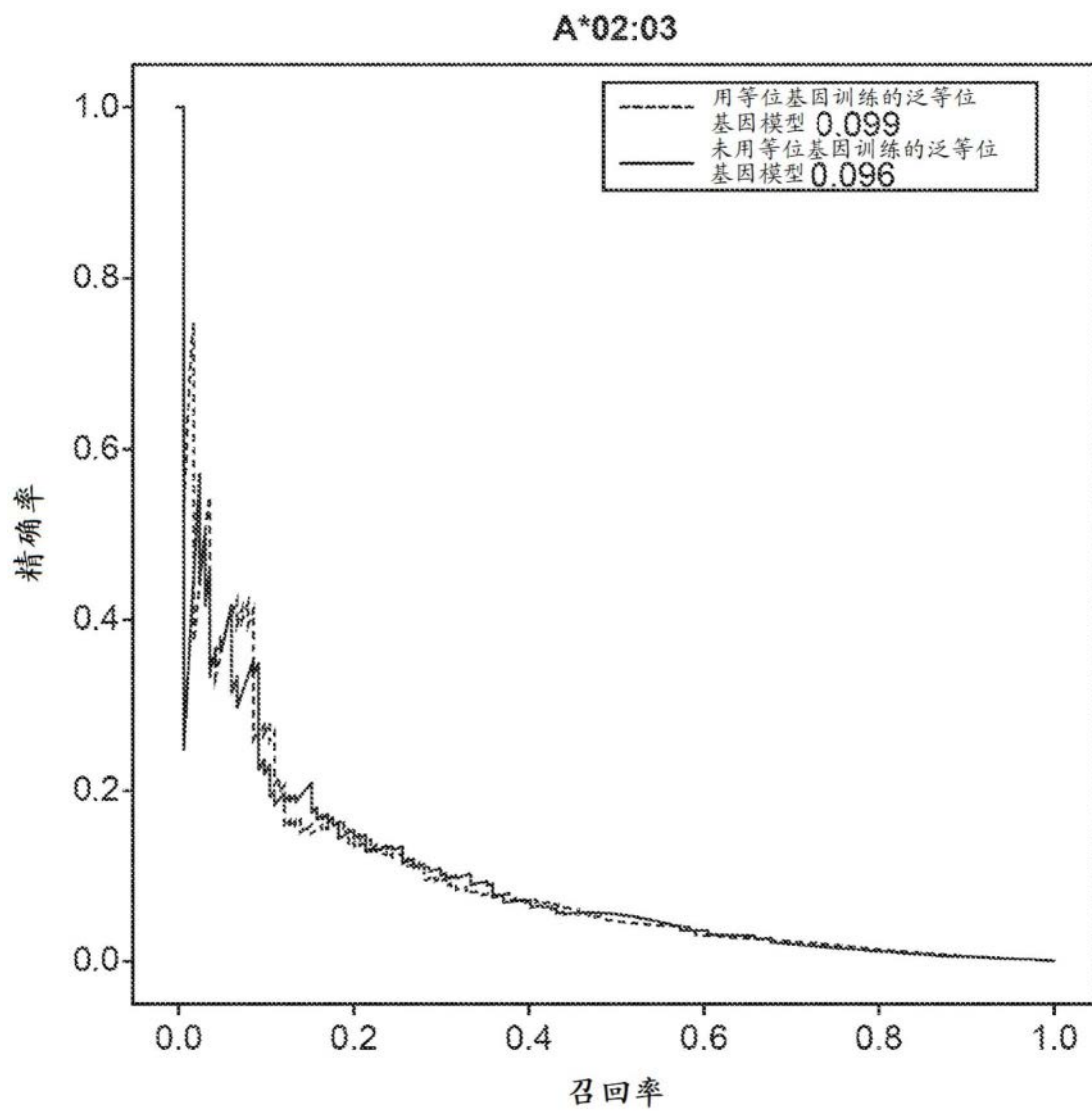


图16

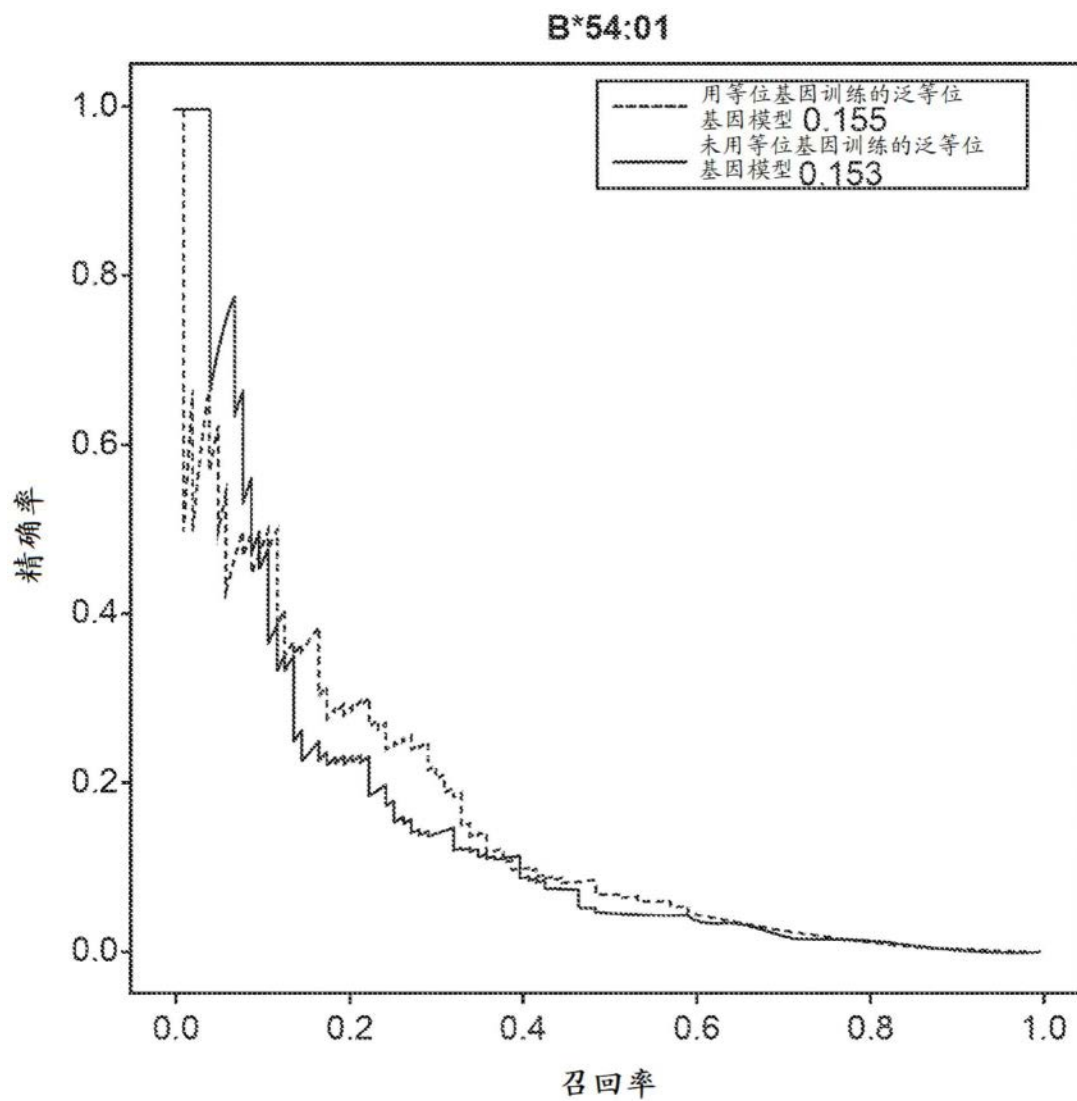


图17

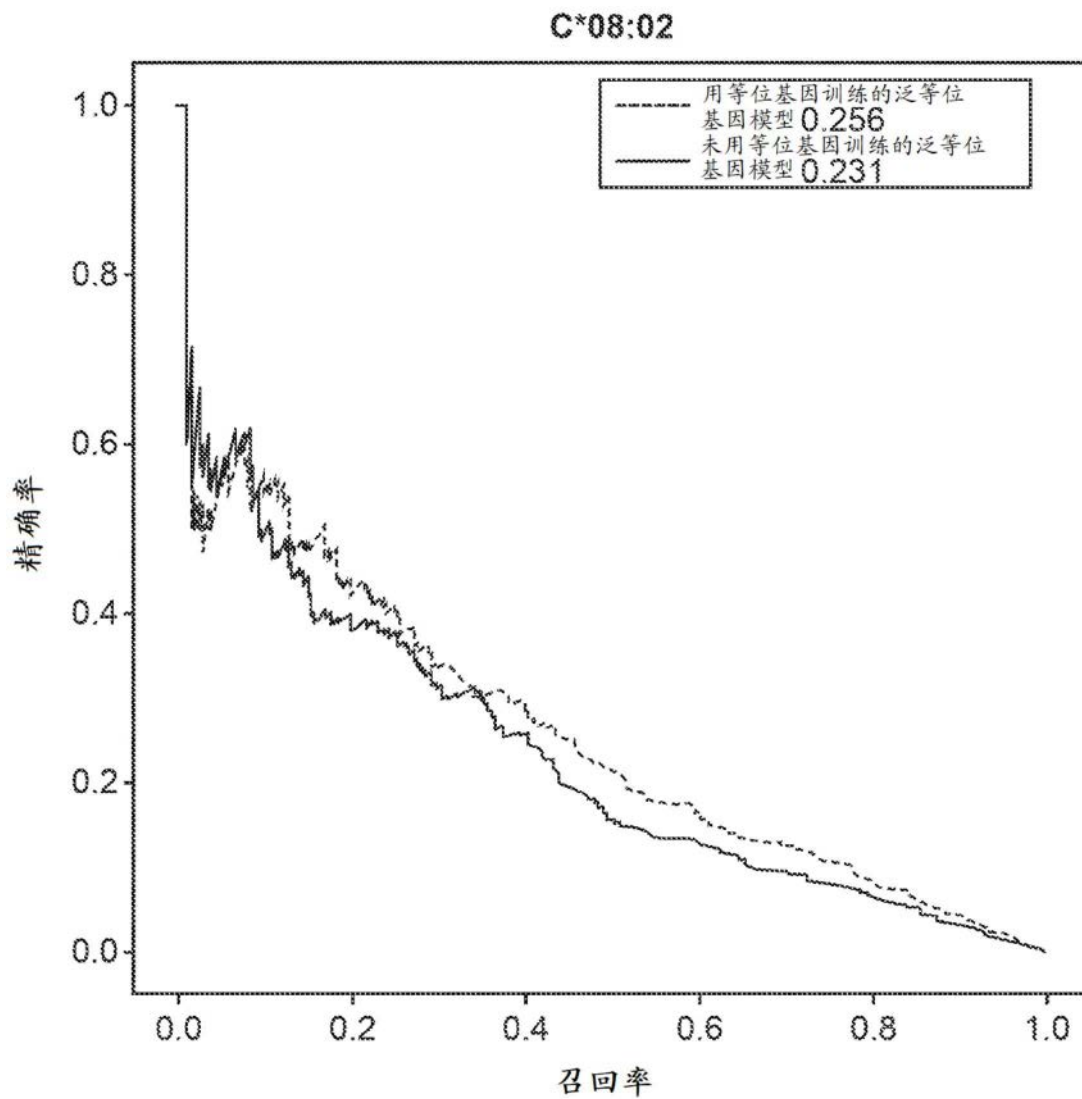


图18

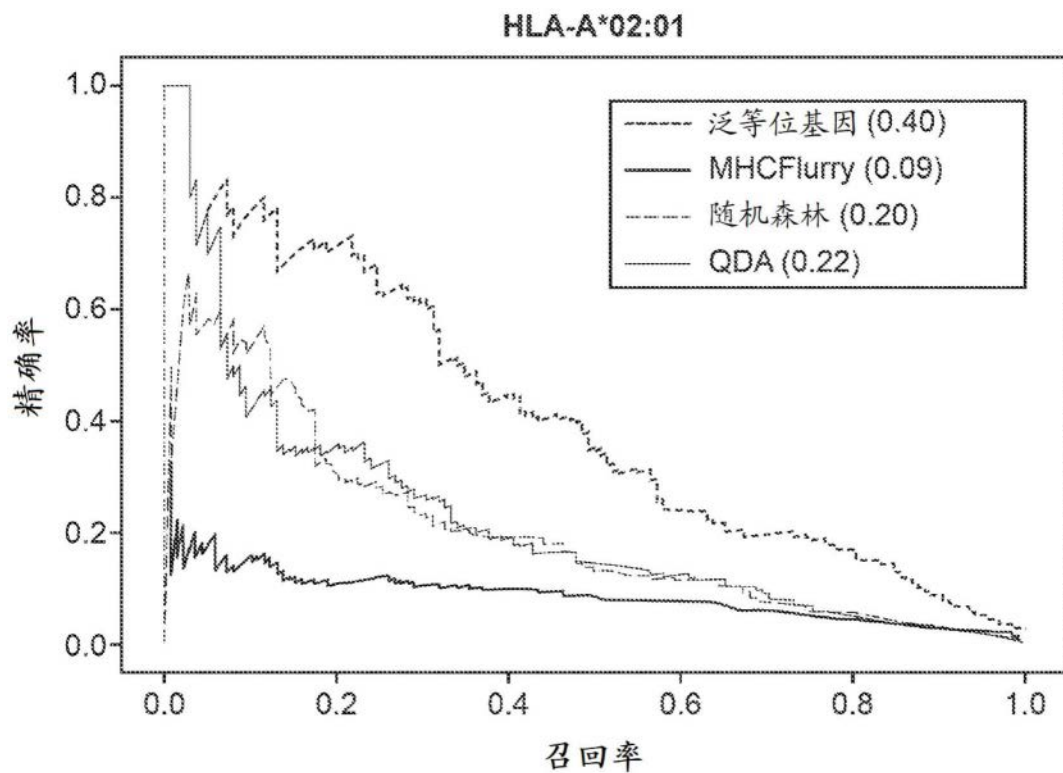


图19

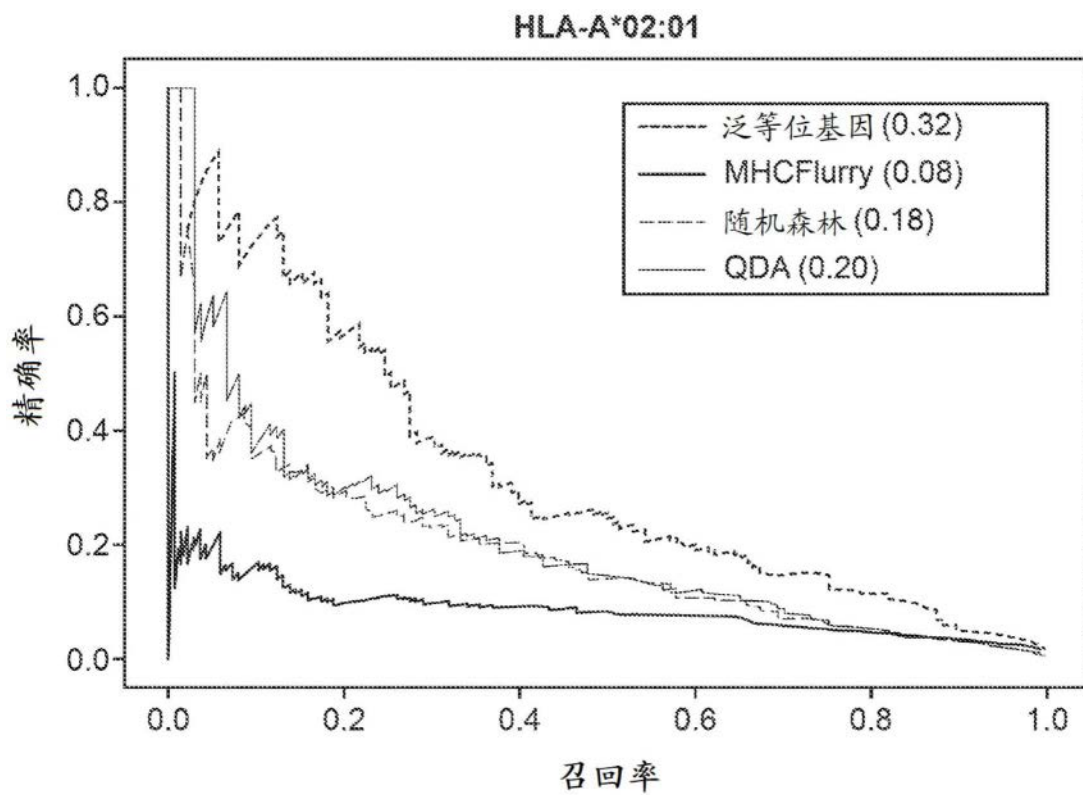


图20

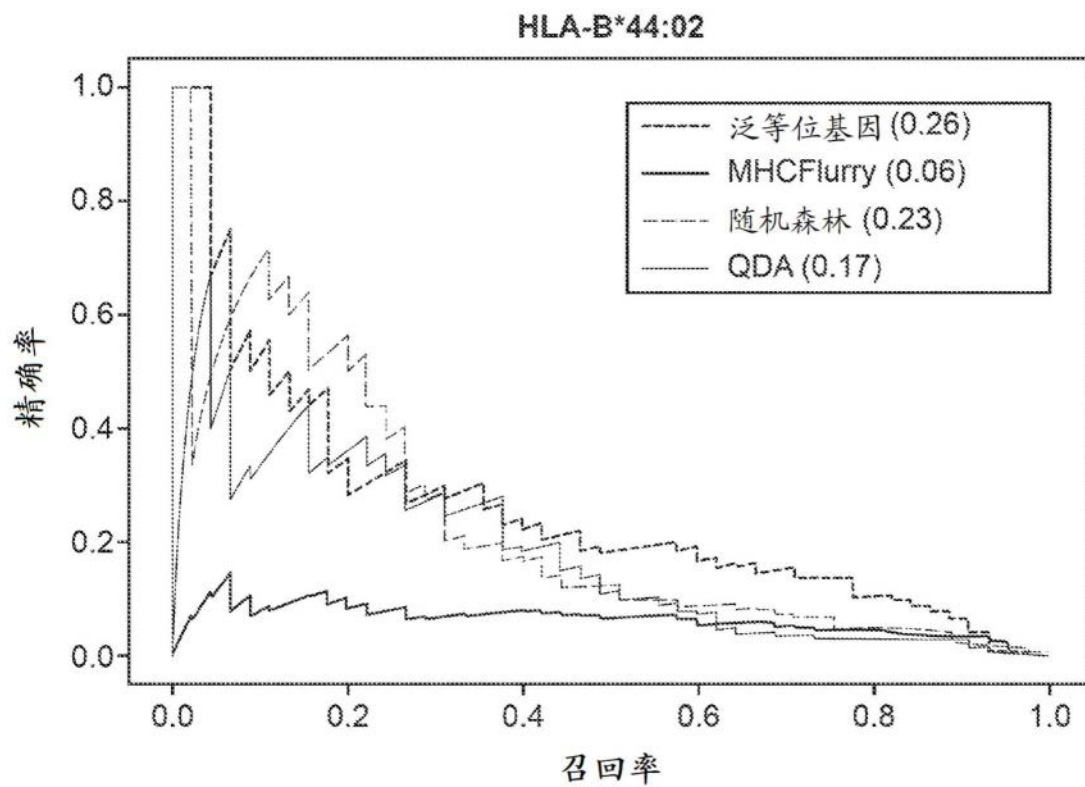


图21

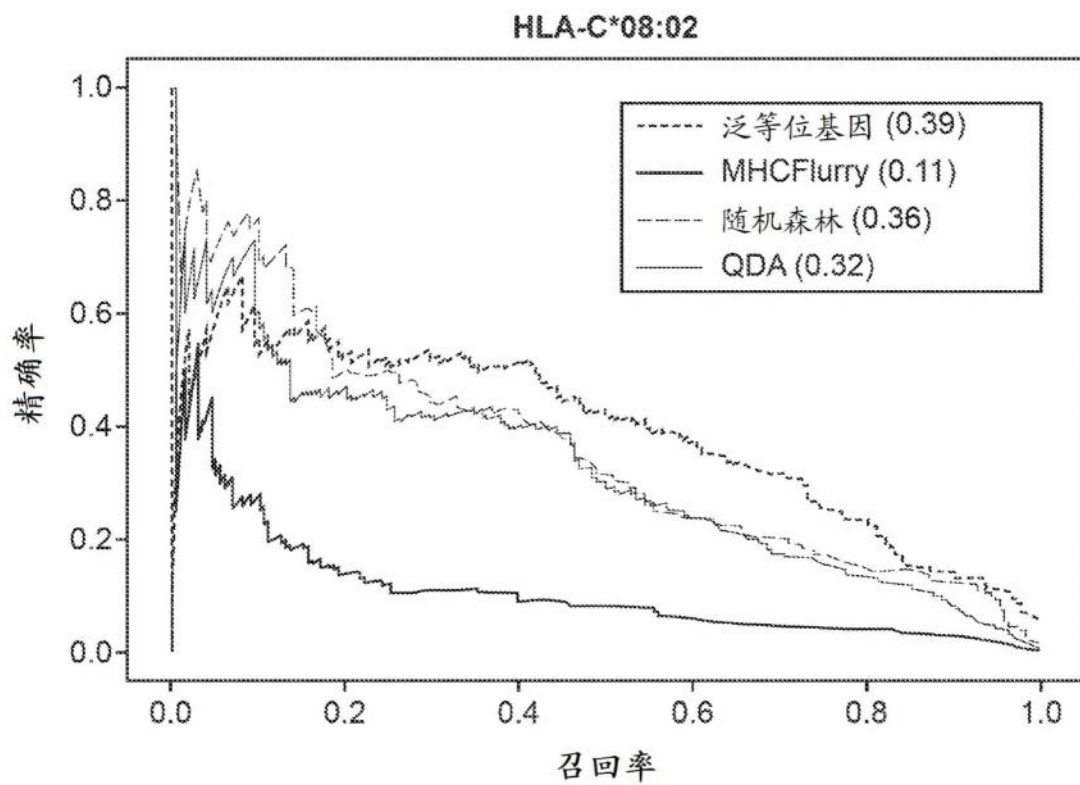


图22

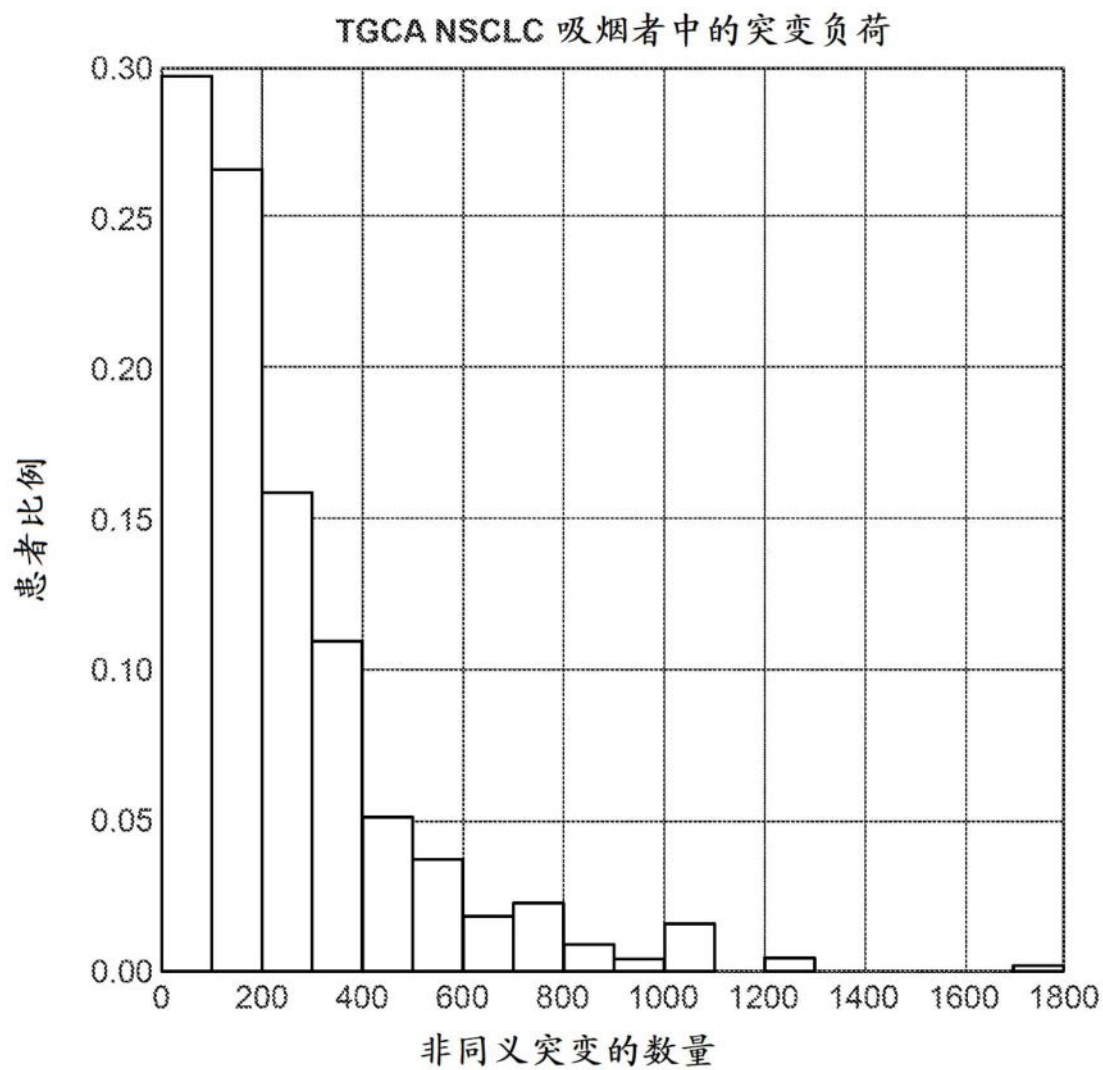


图23A

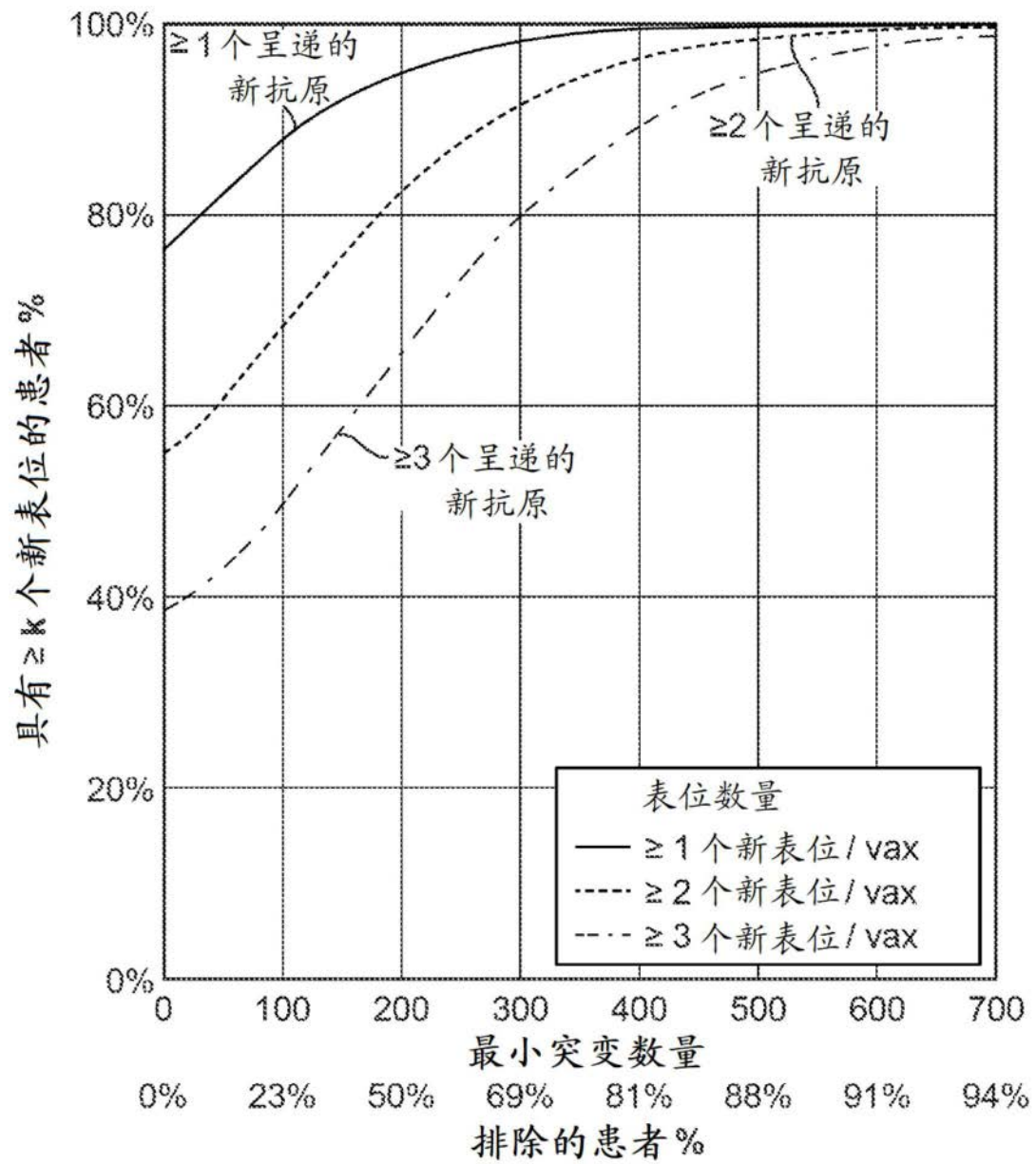


图23B

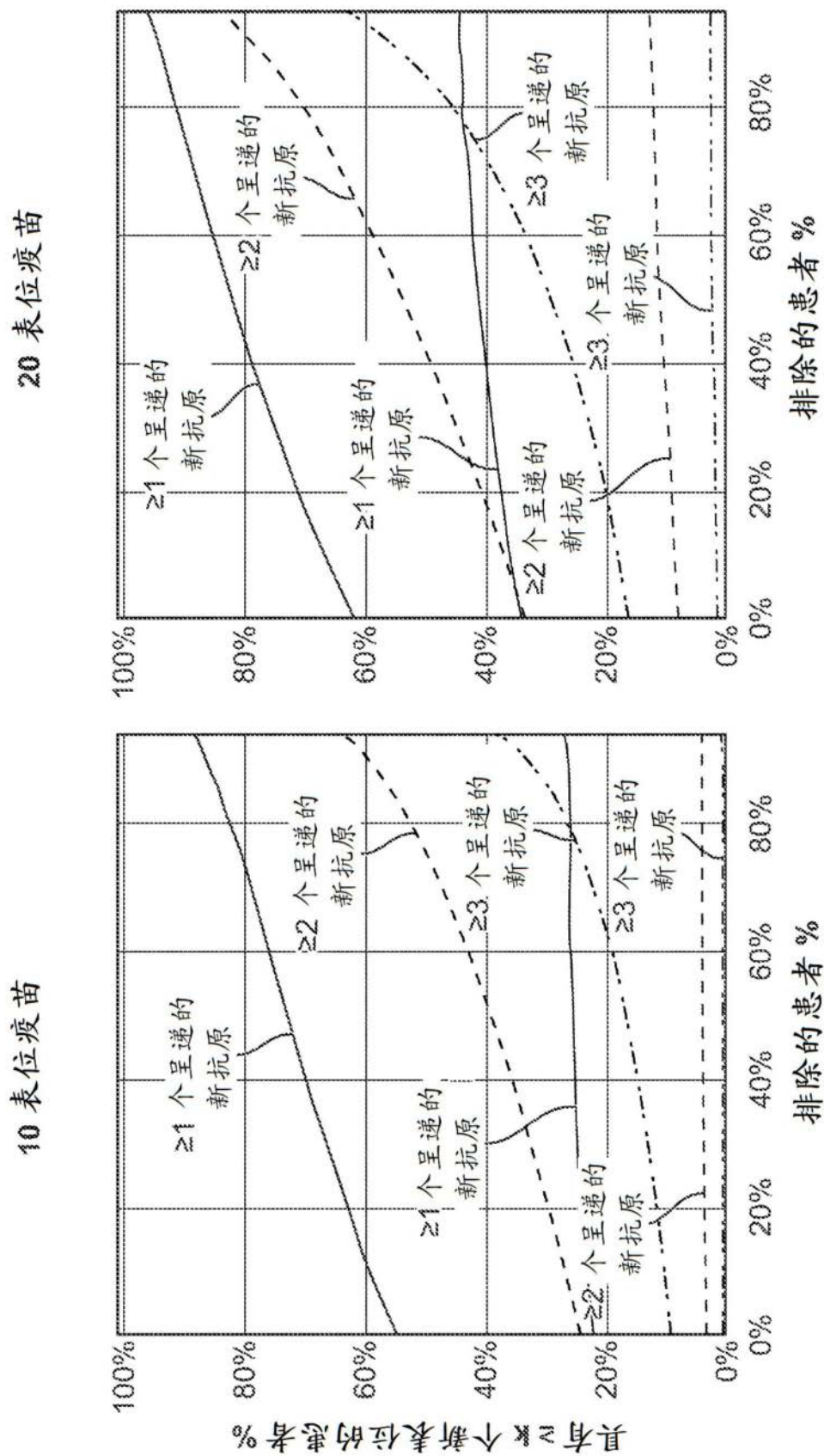


图23C

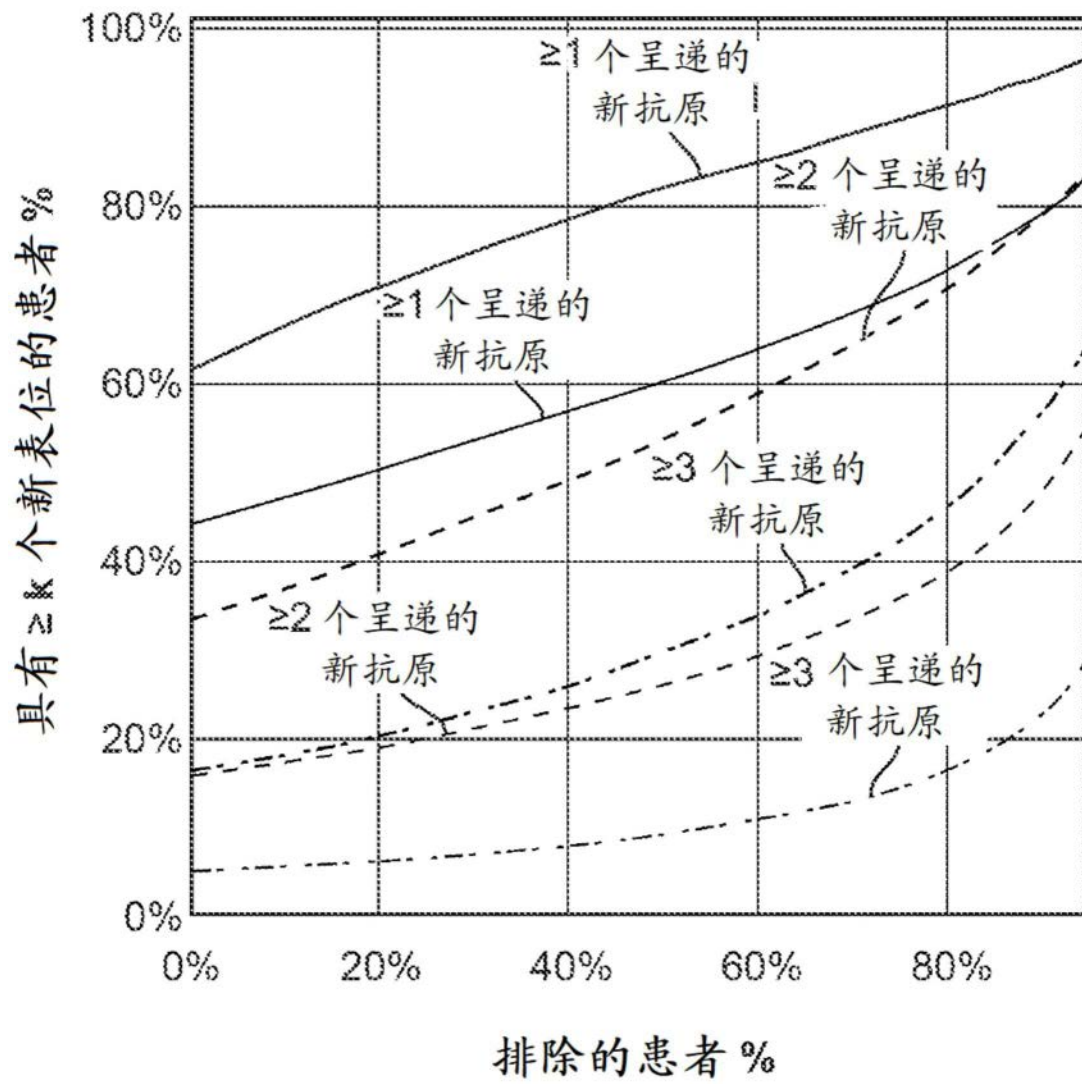


图23D

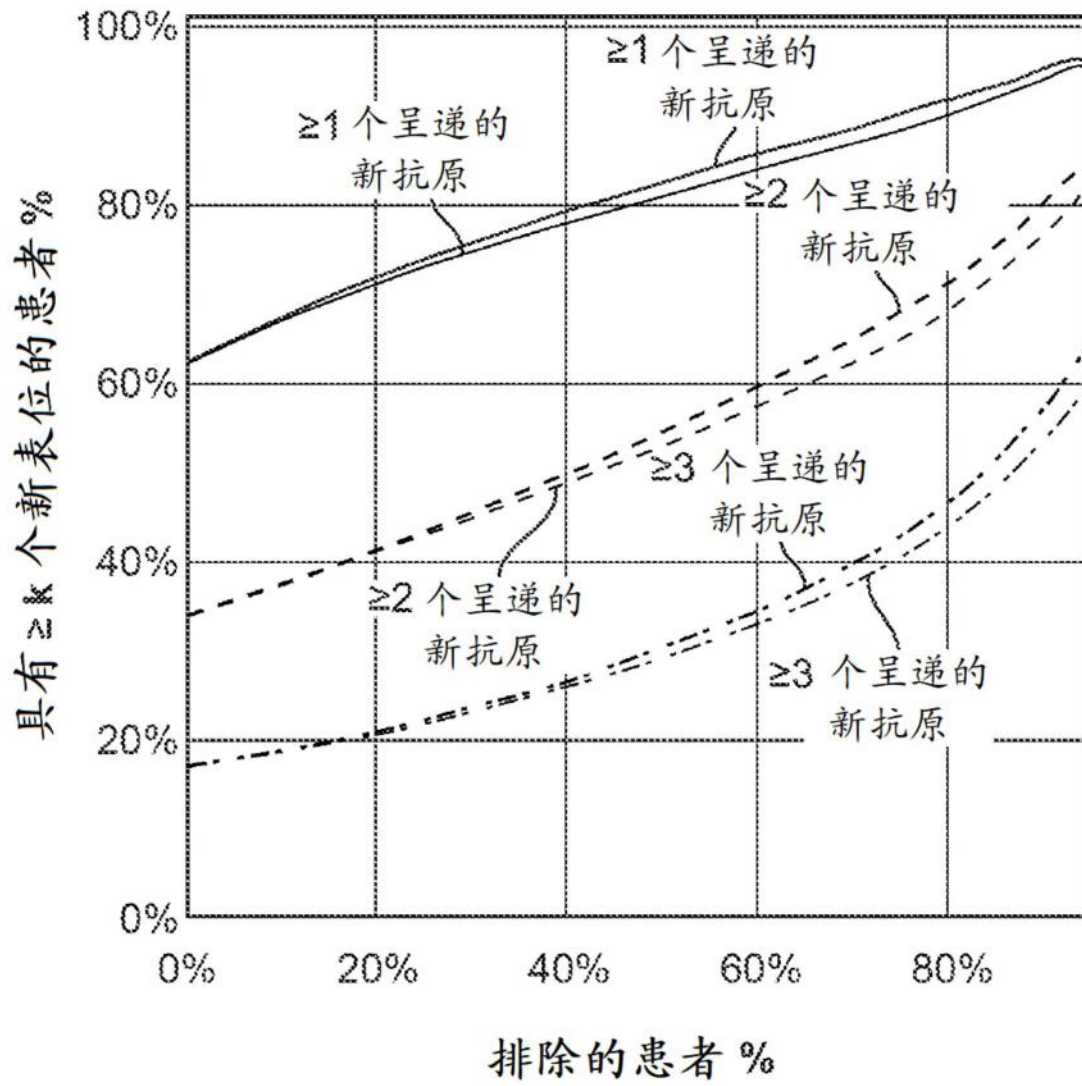


图23E

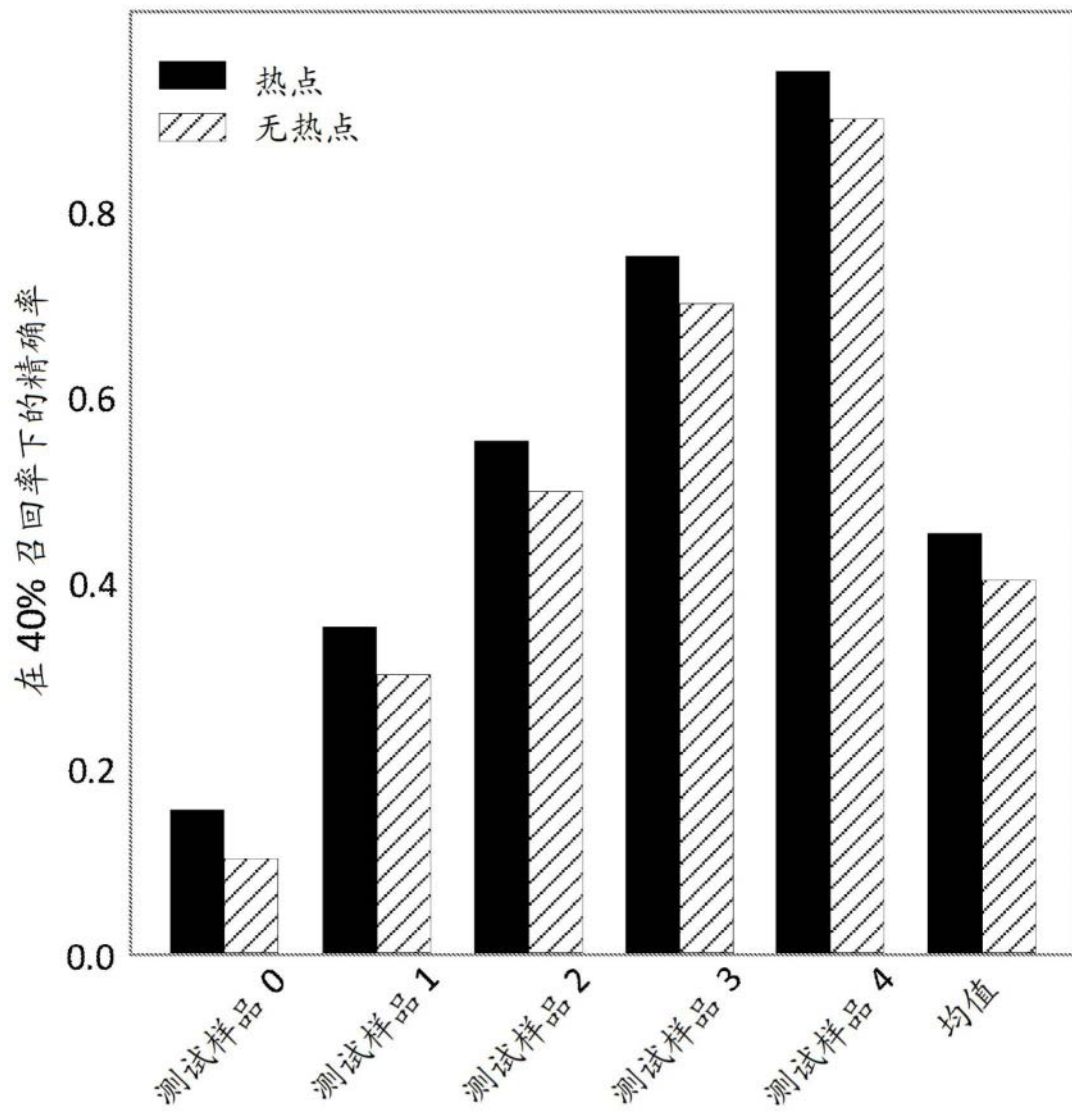


图24

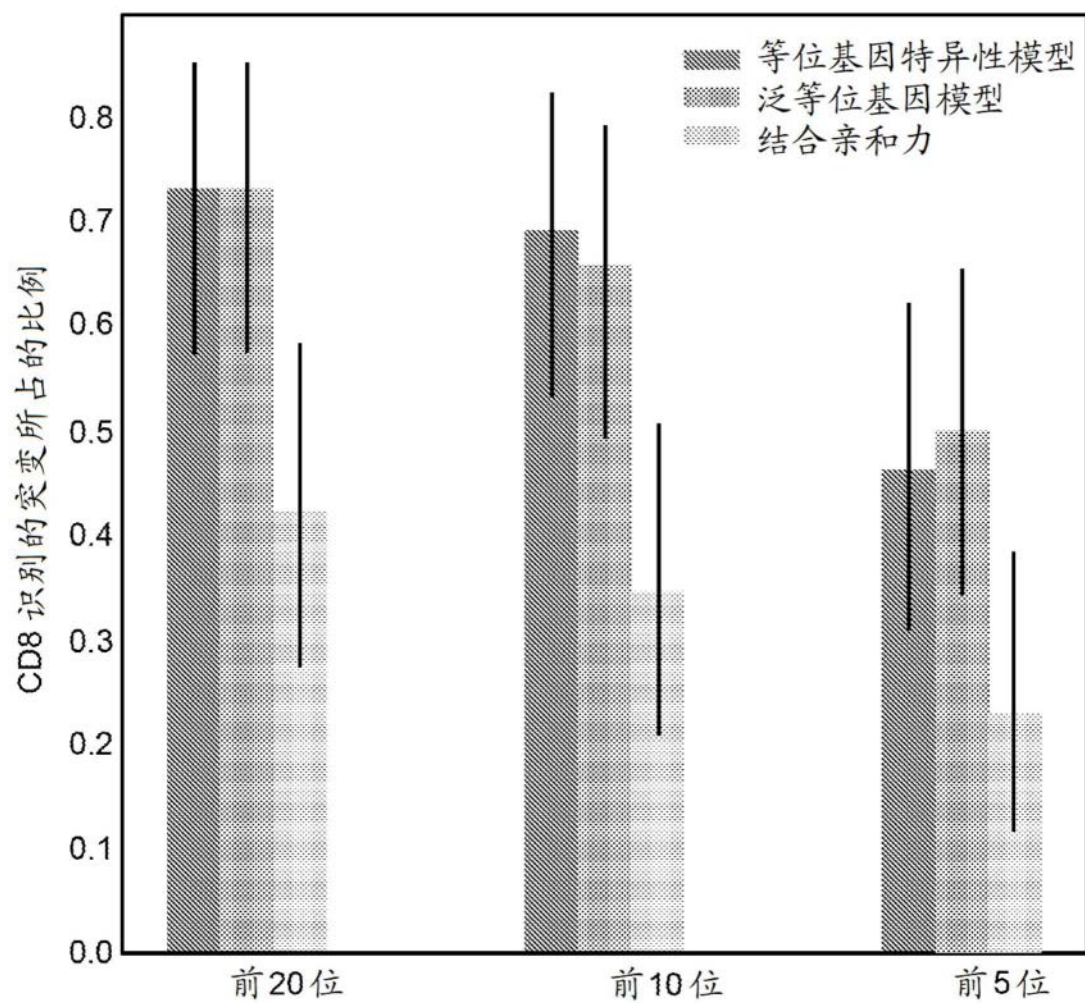


图25A

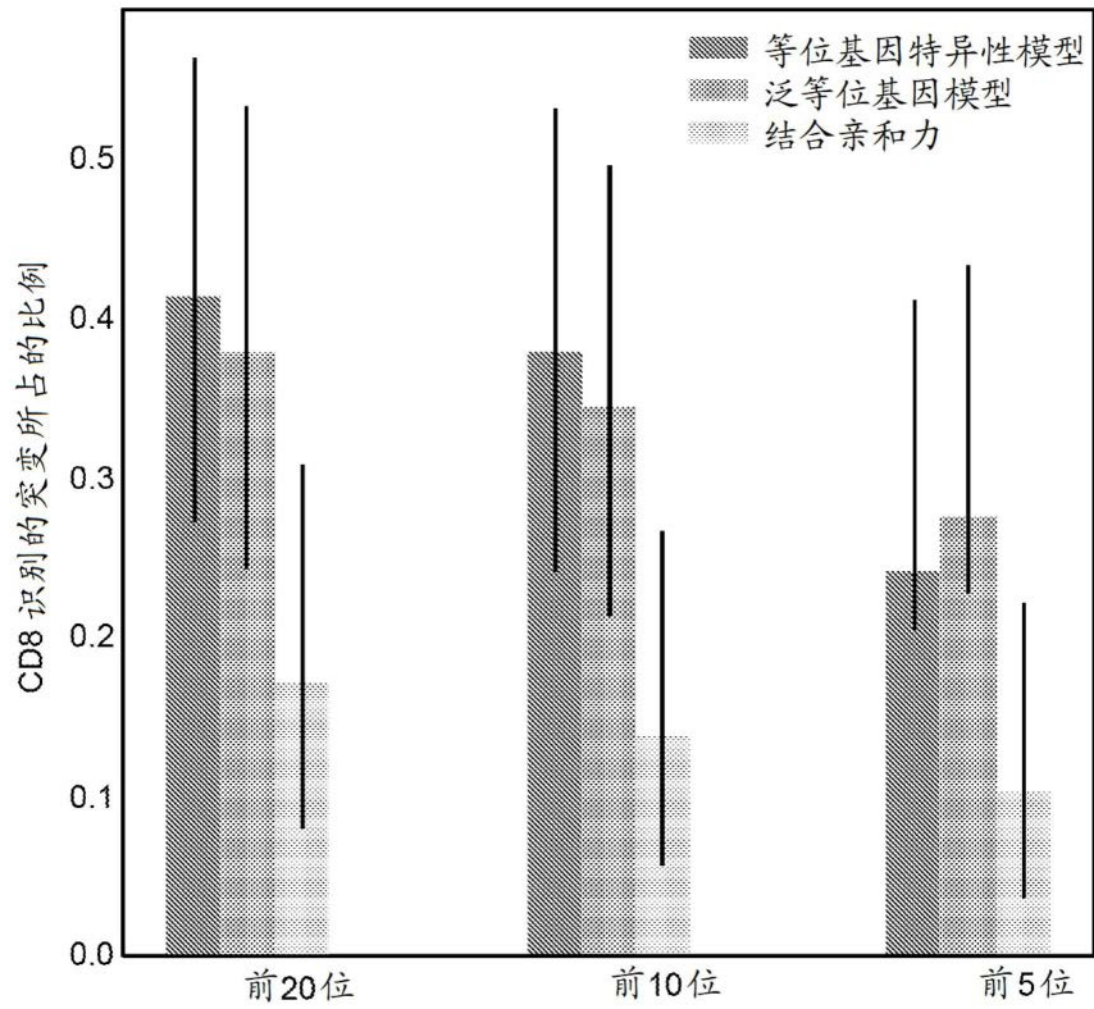


图25B

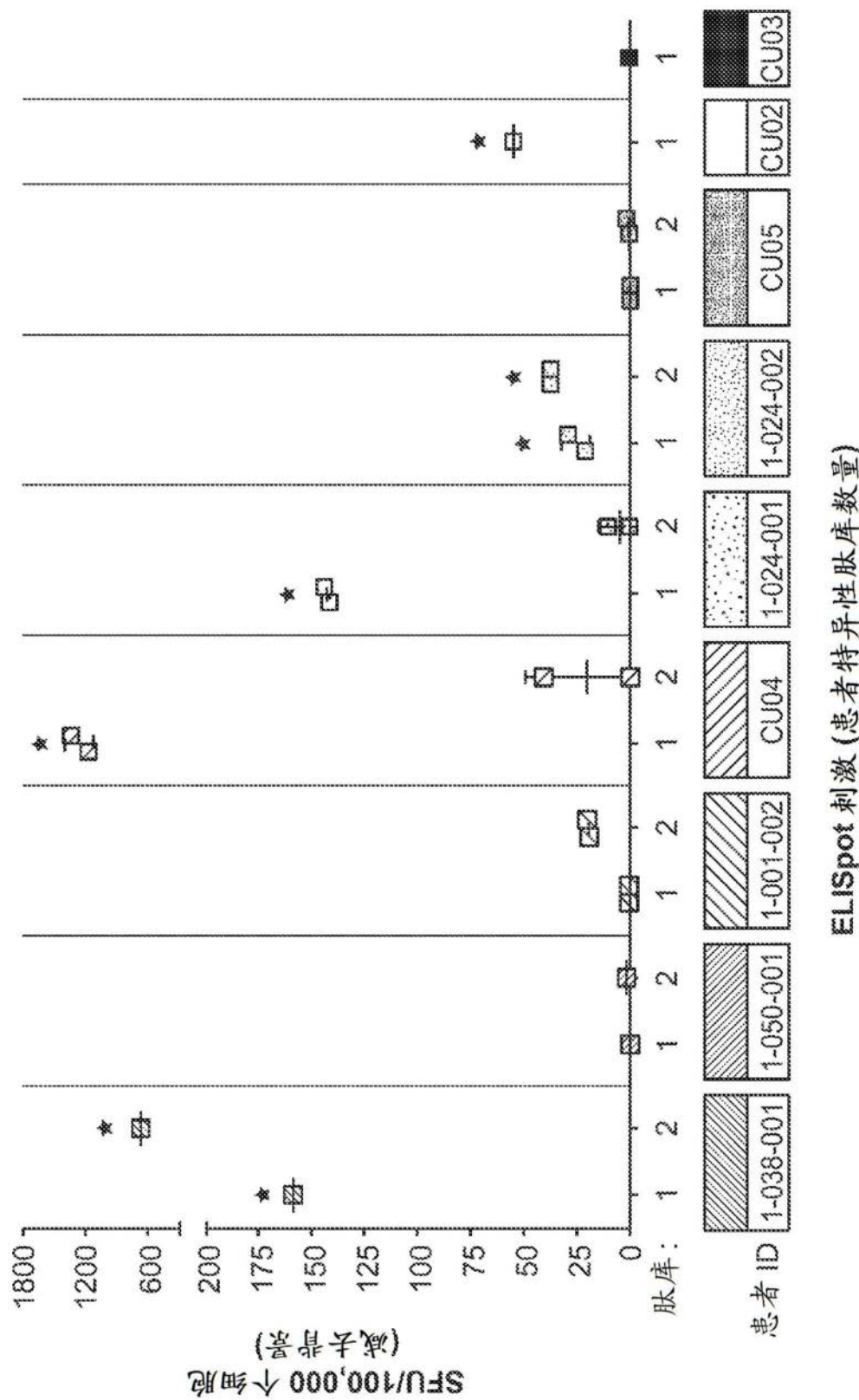


图26A

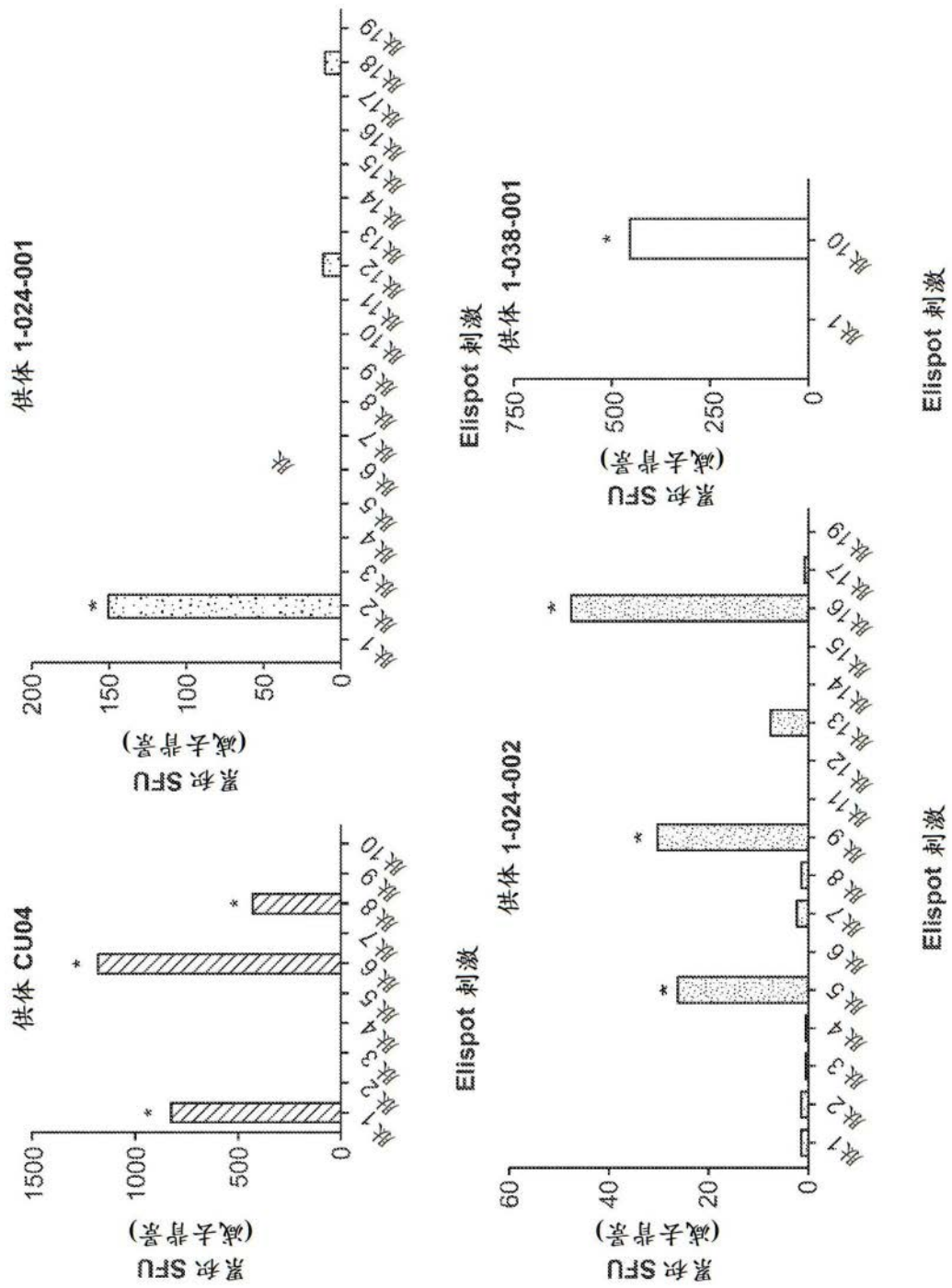


图26B

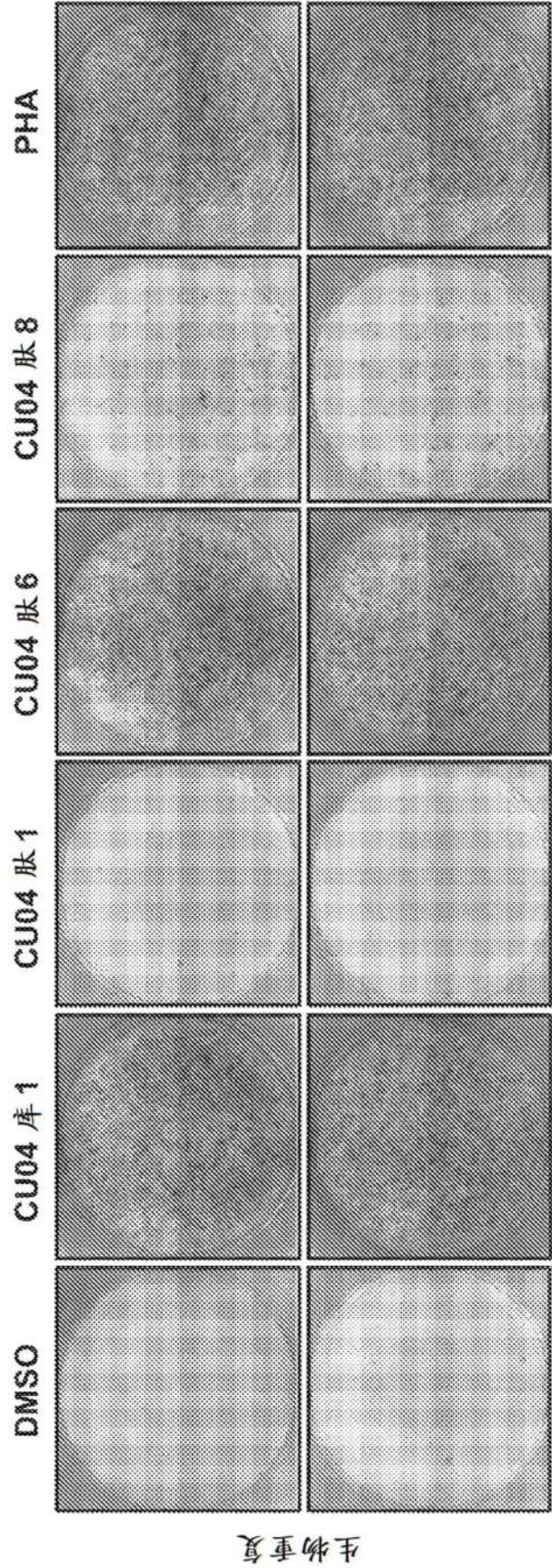


图26C

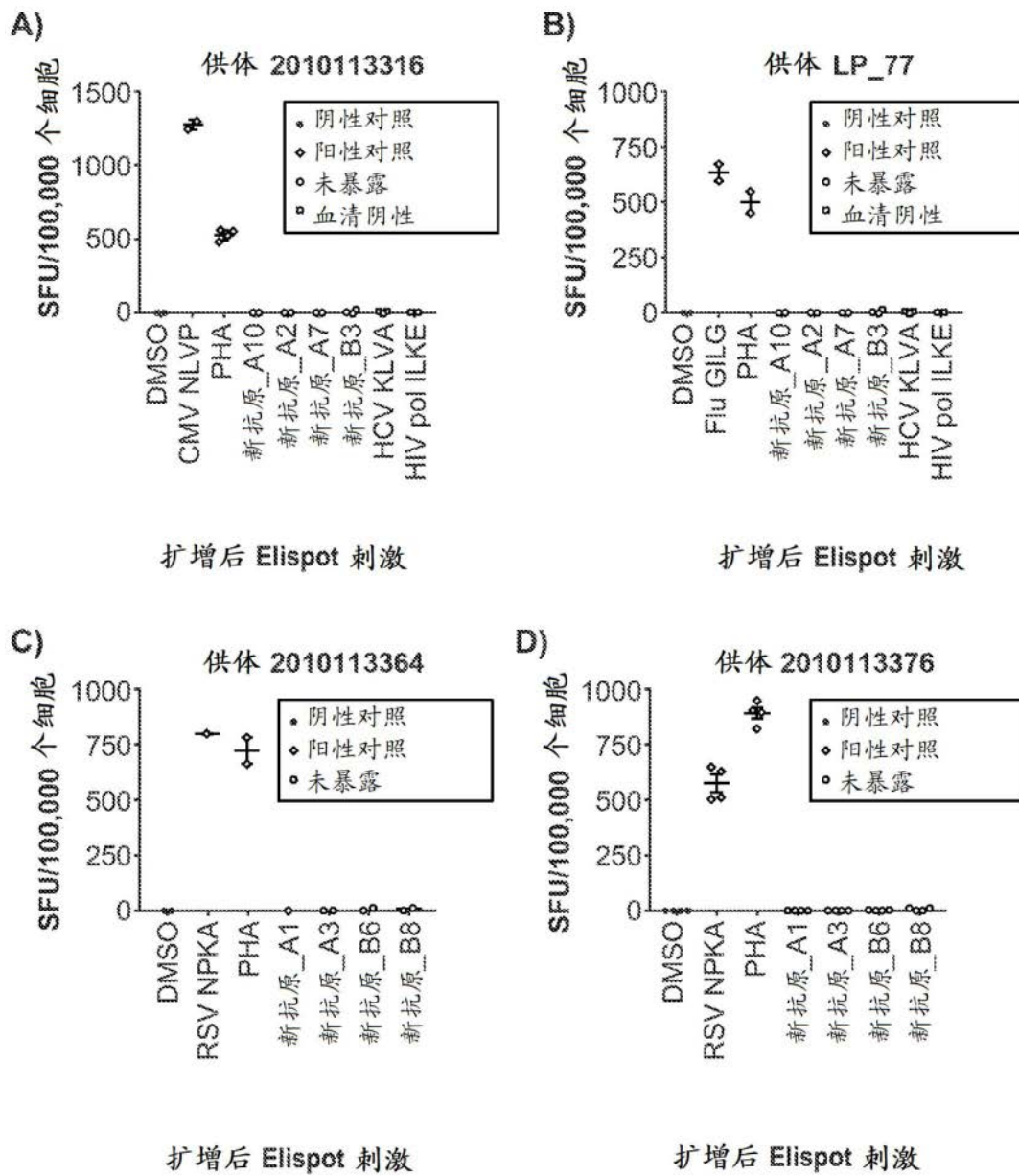


图27A

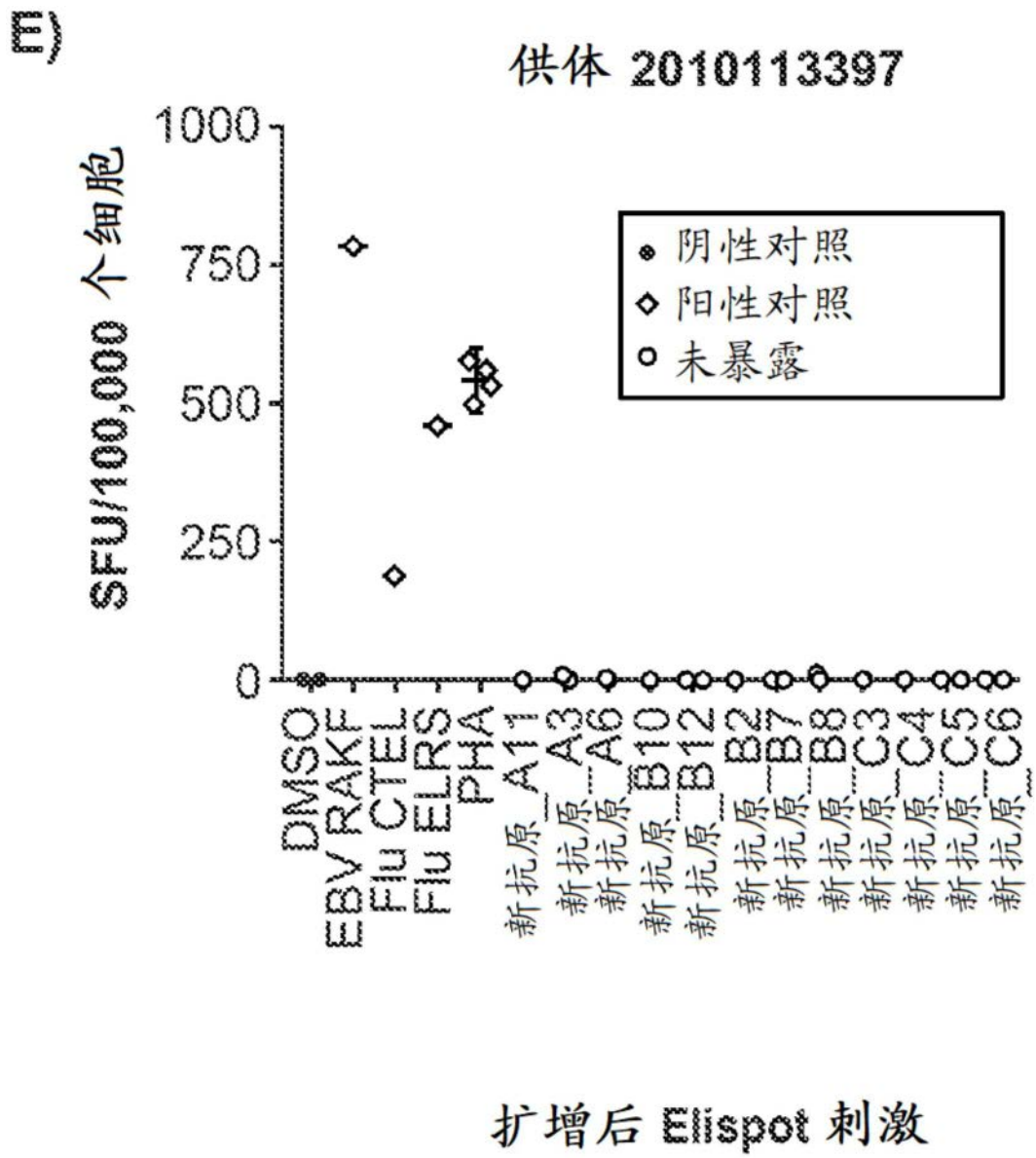


图27A(续)

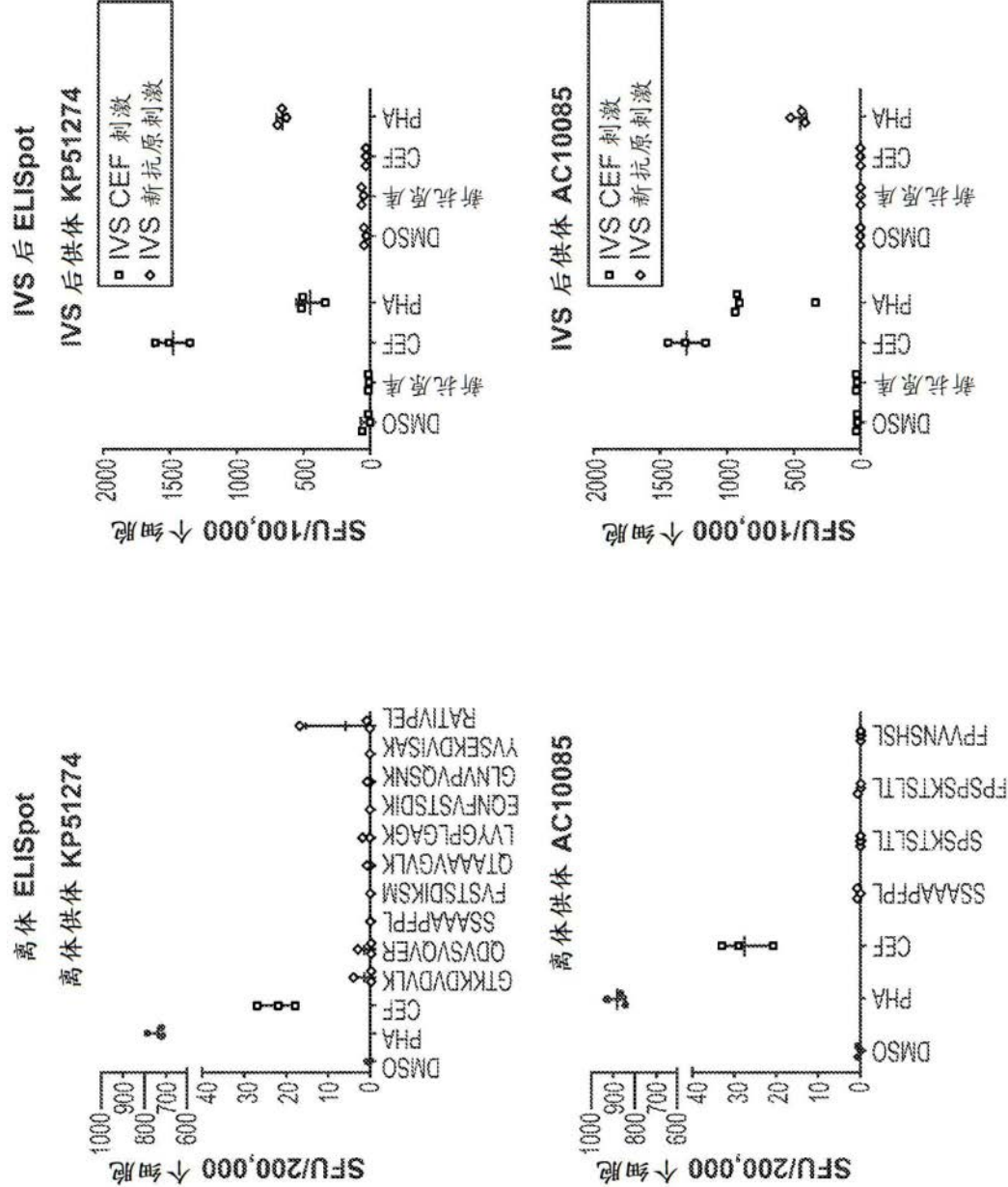


图27B

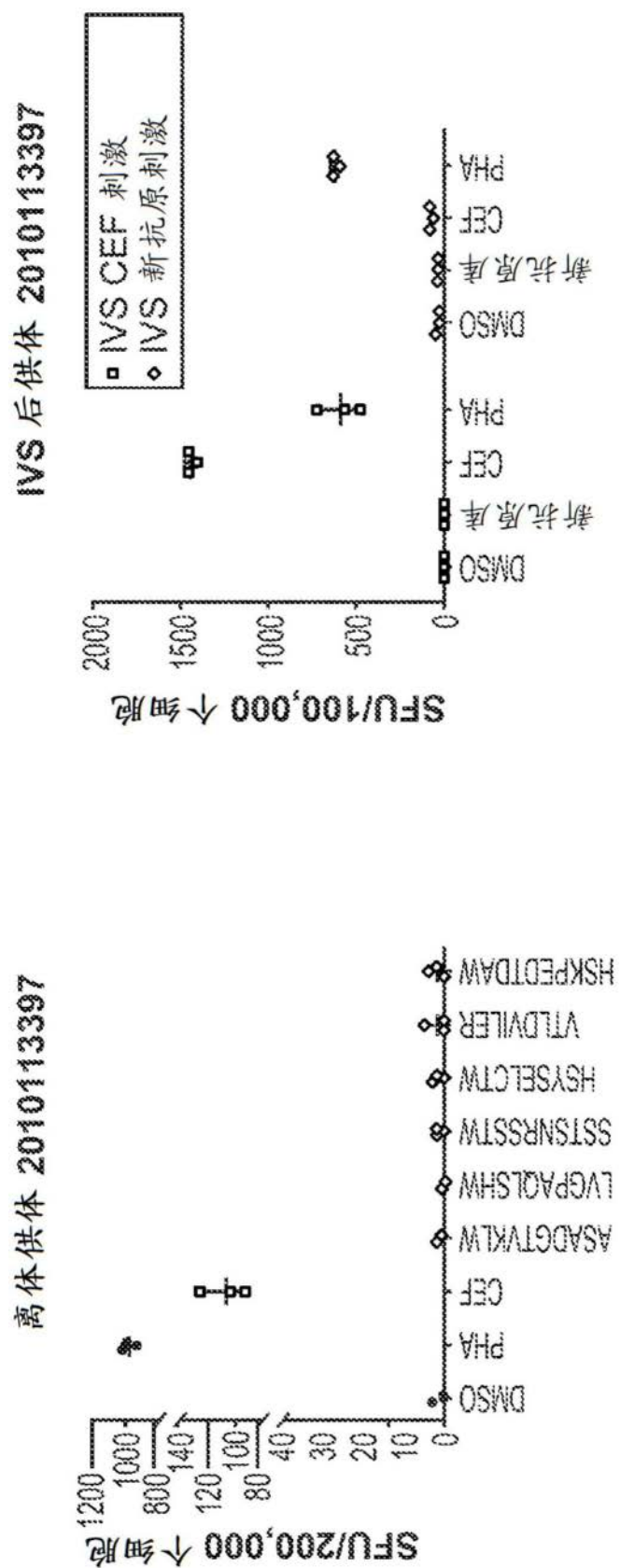


图27B (续)

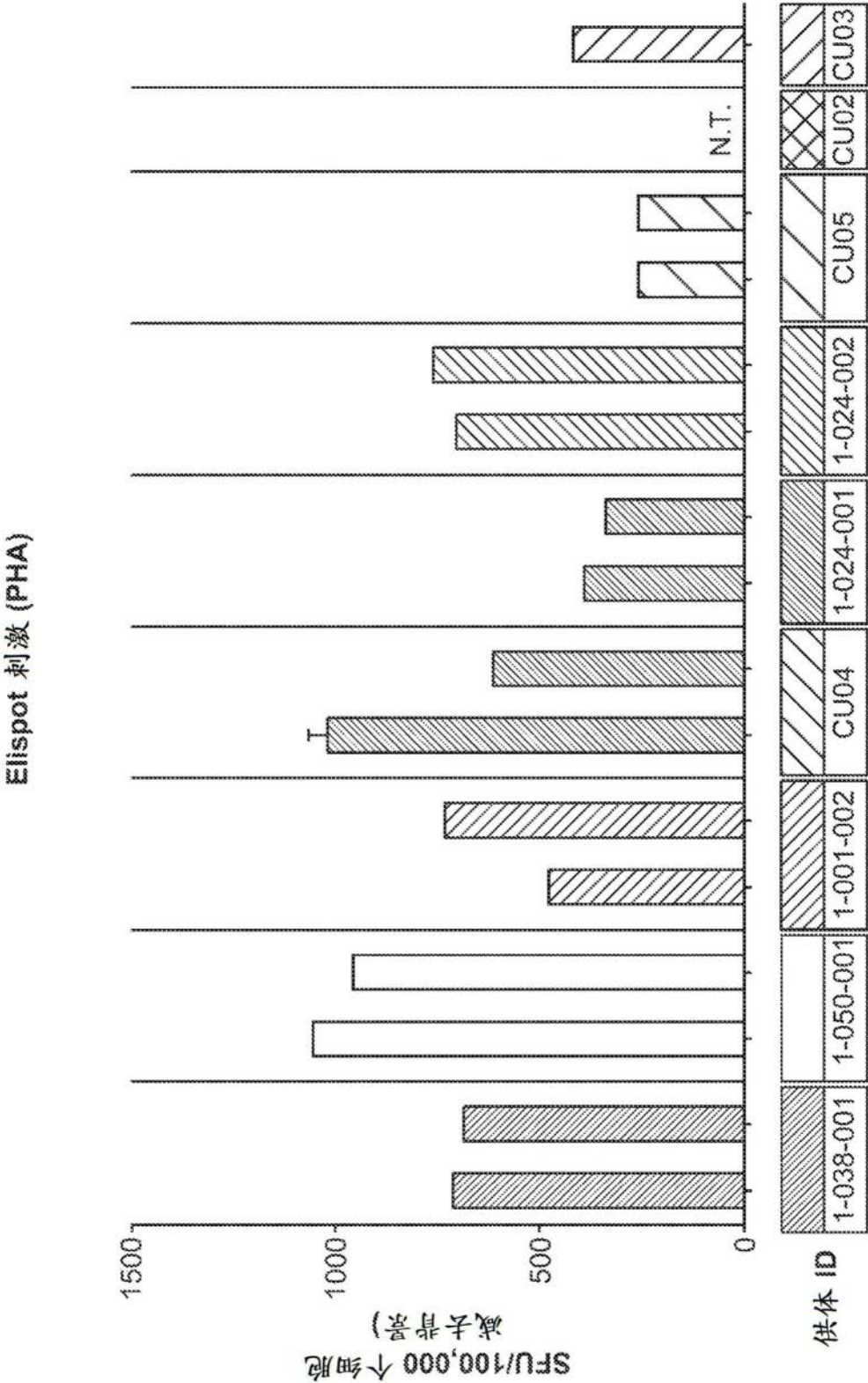


图28

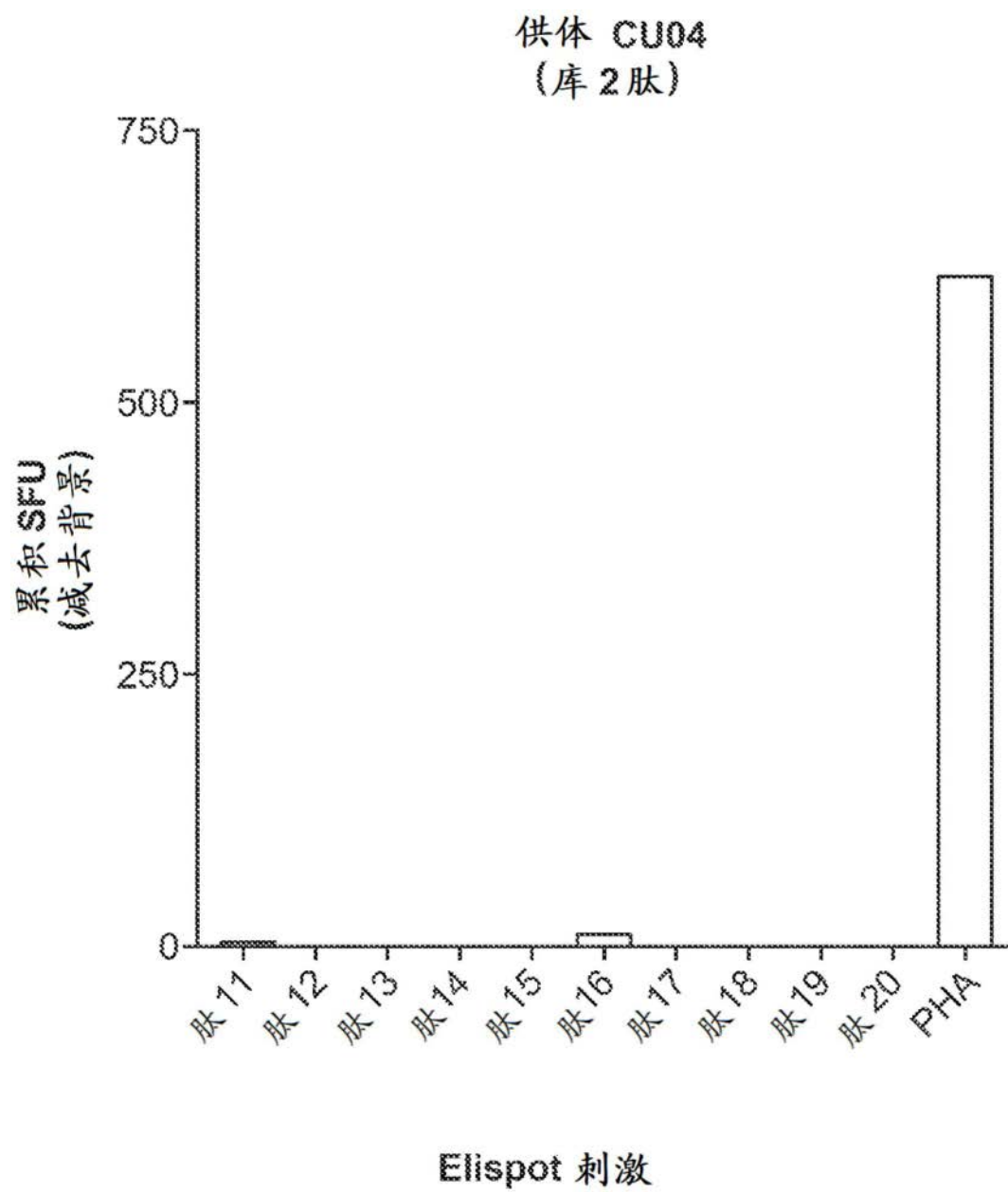


图29A

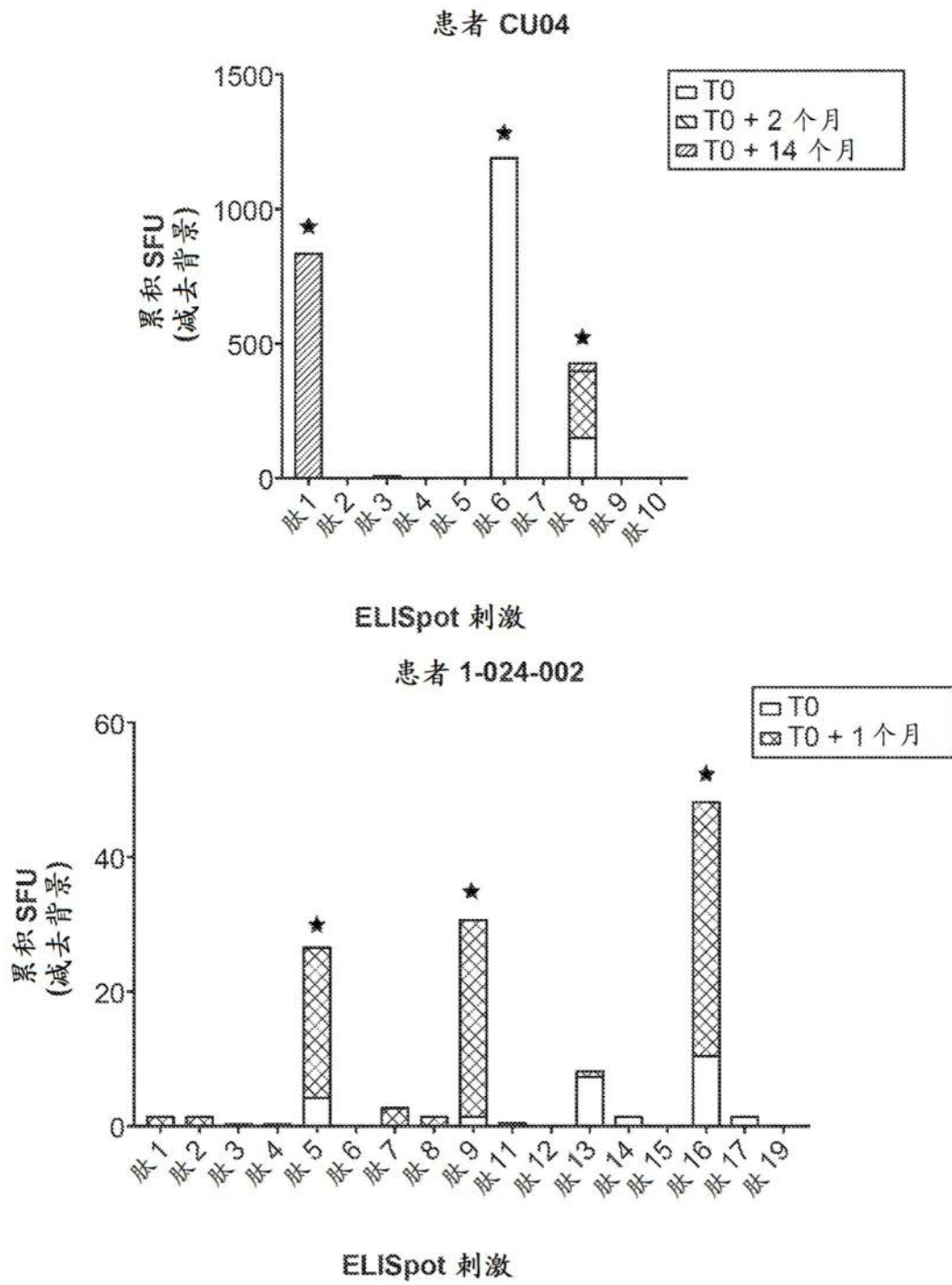


图29B

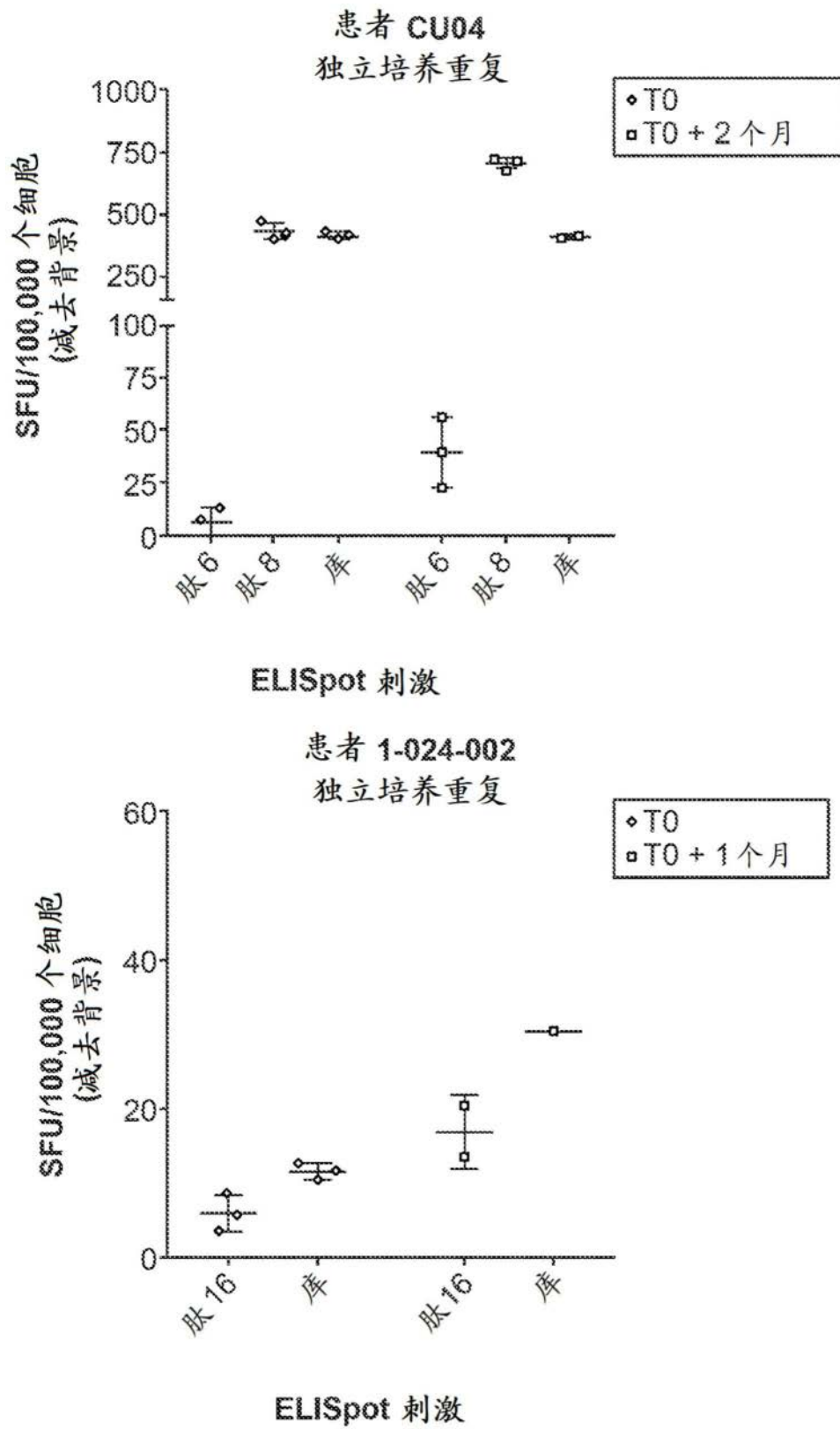


图29C

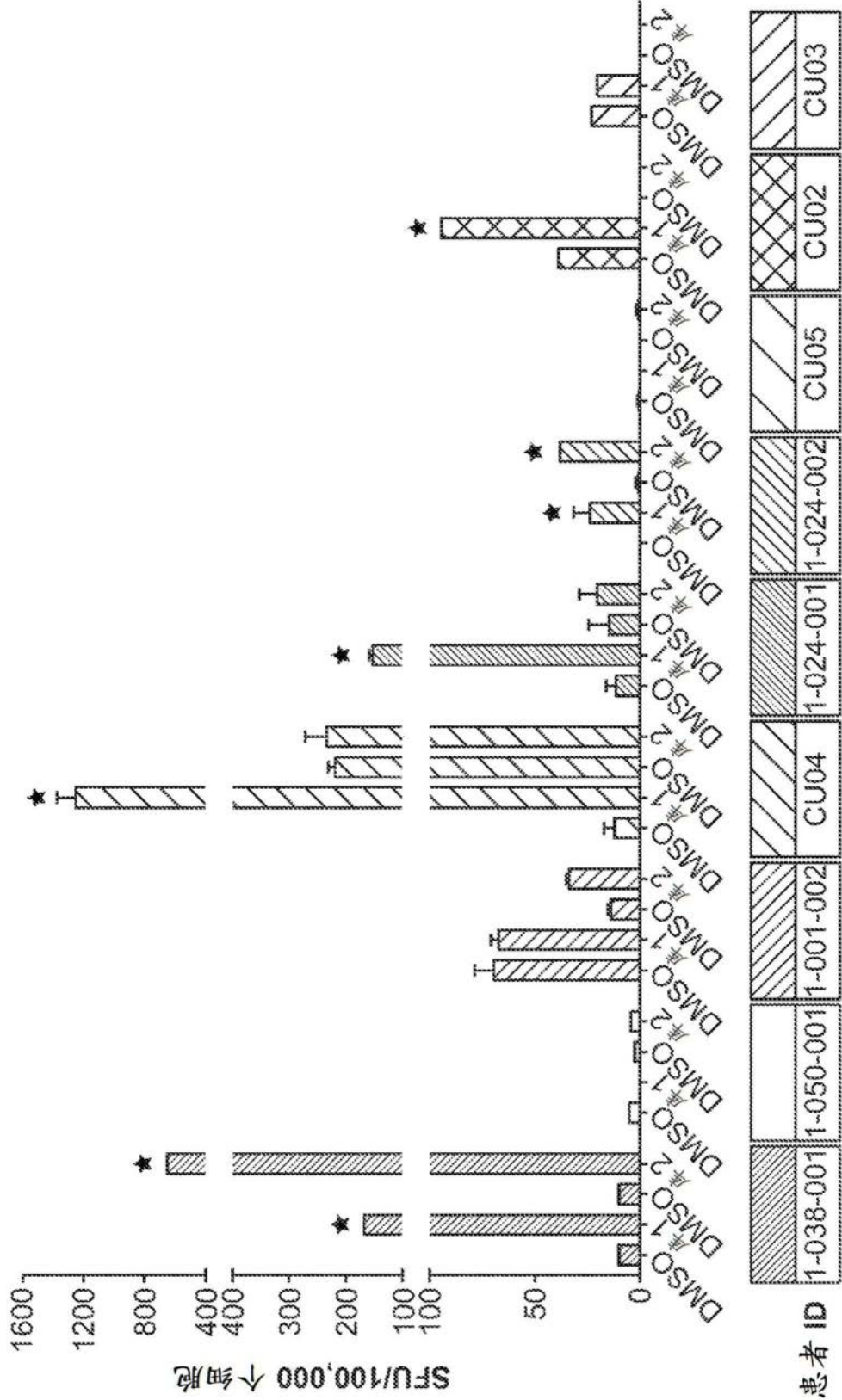


图30

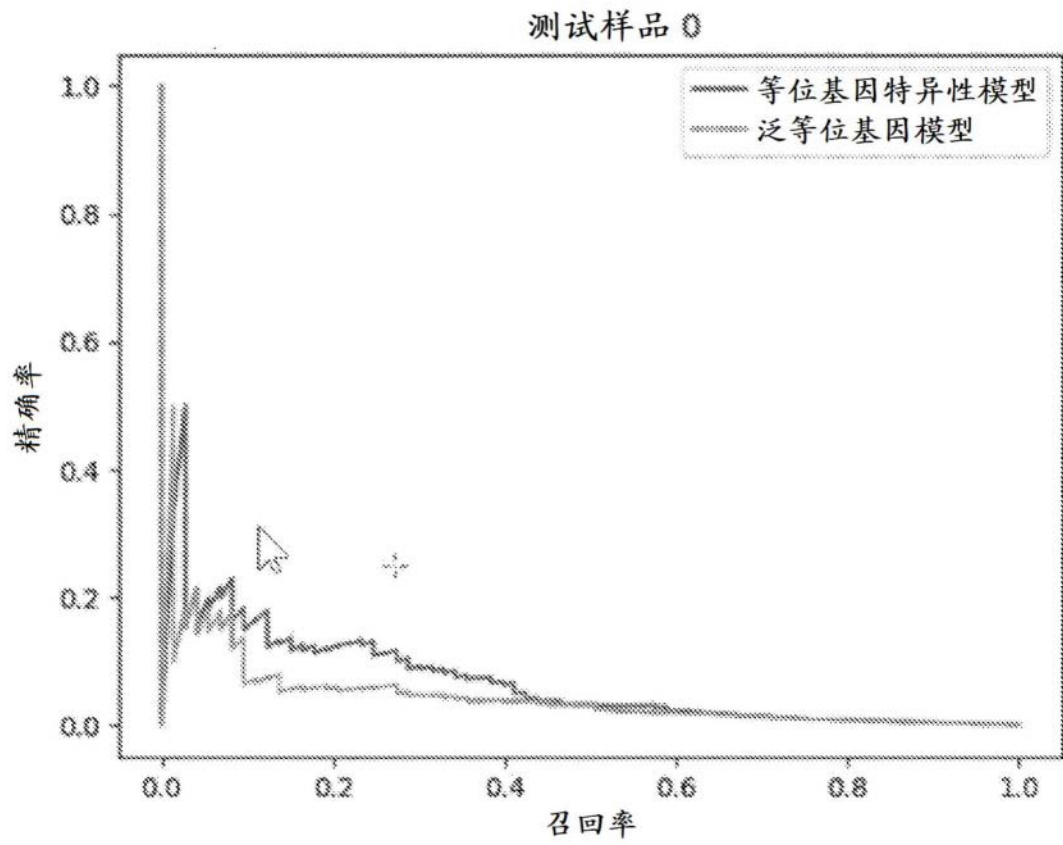


图31A

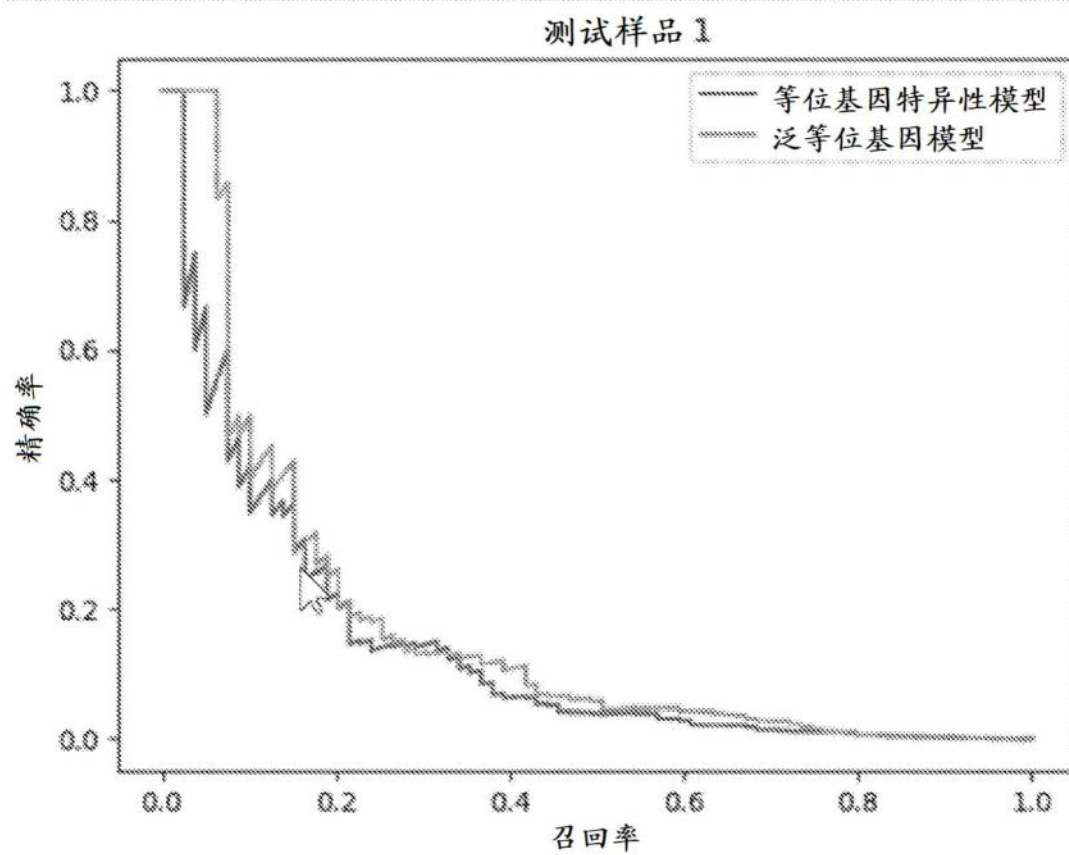


图31B

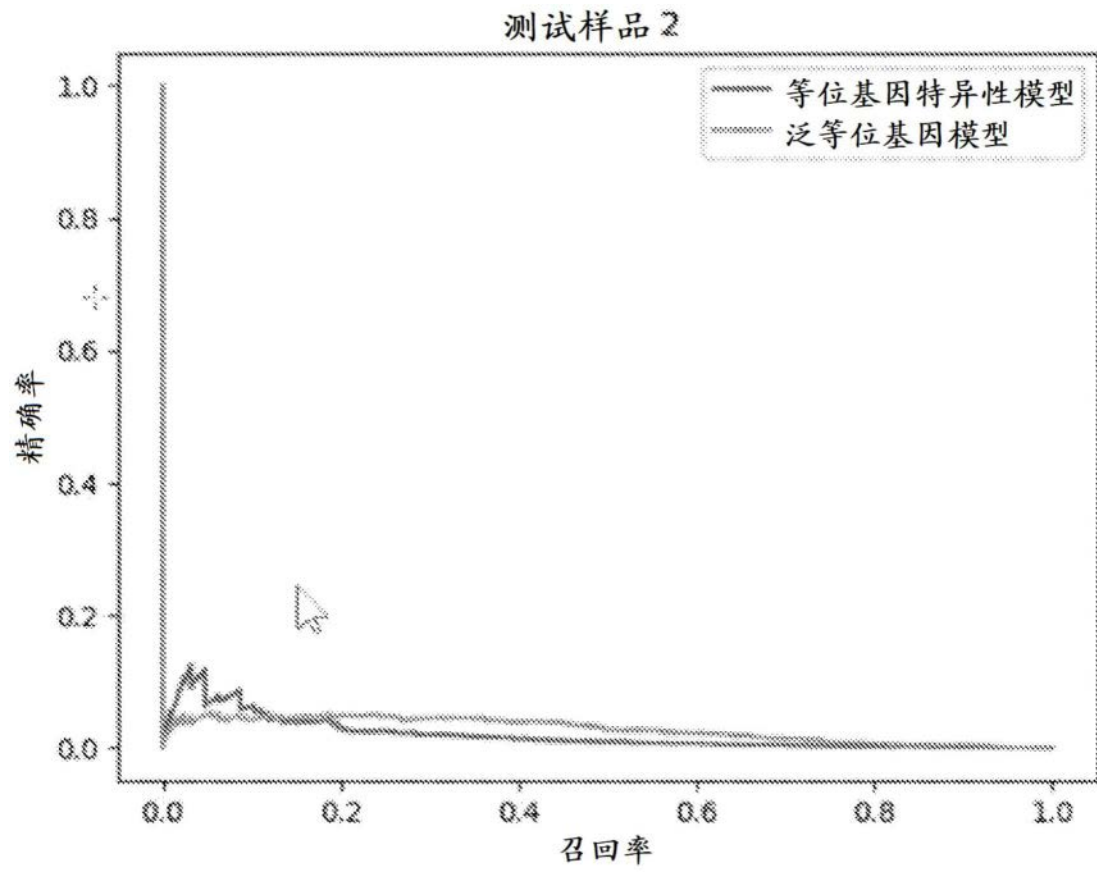


图31C

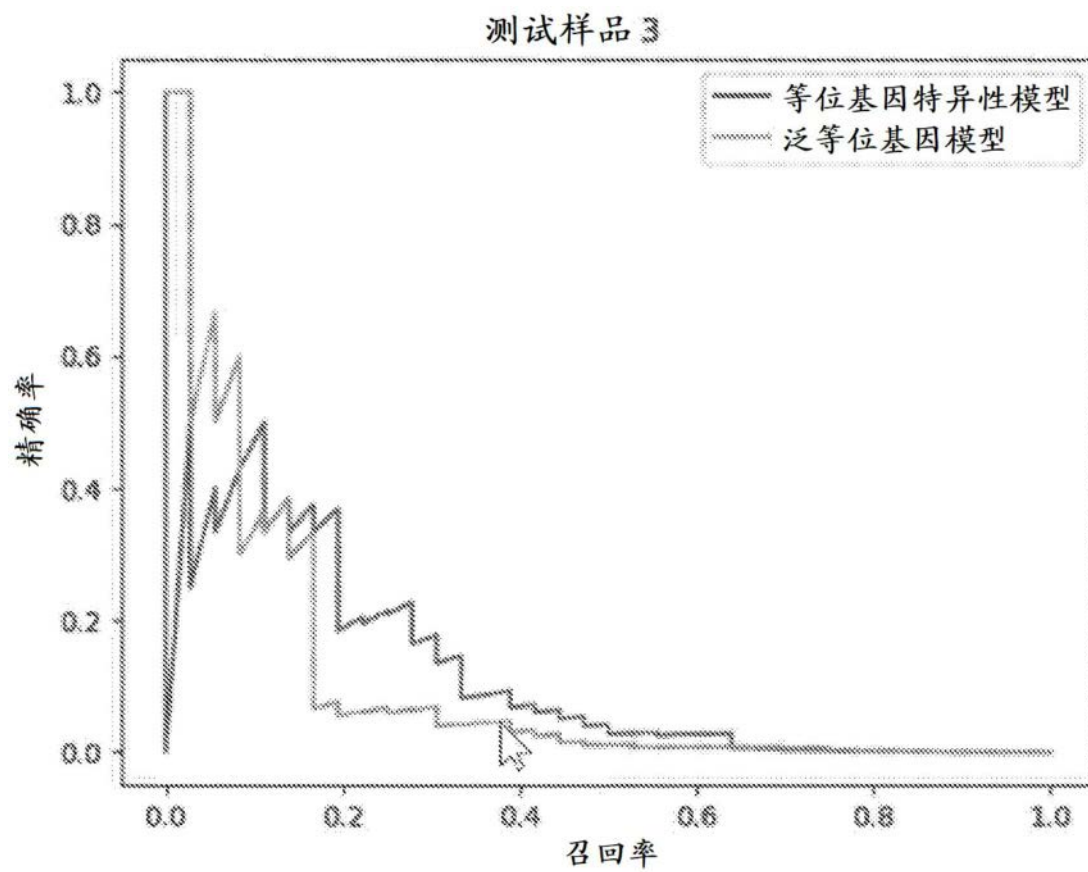


图31D

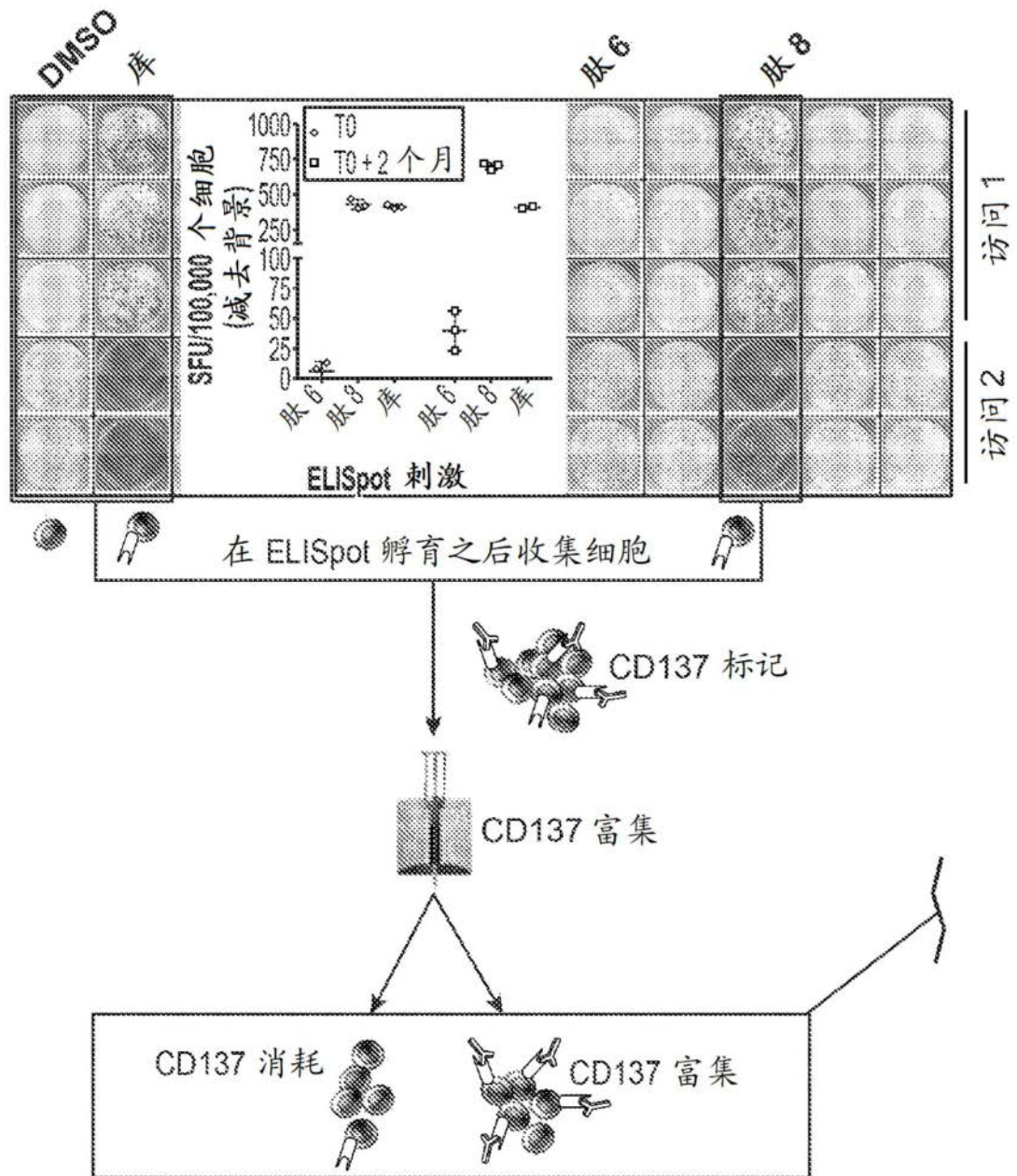


图32

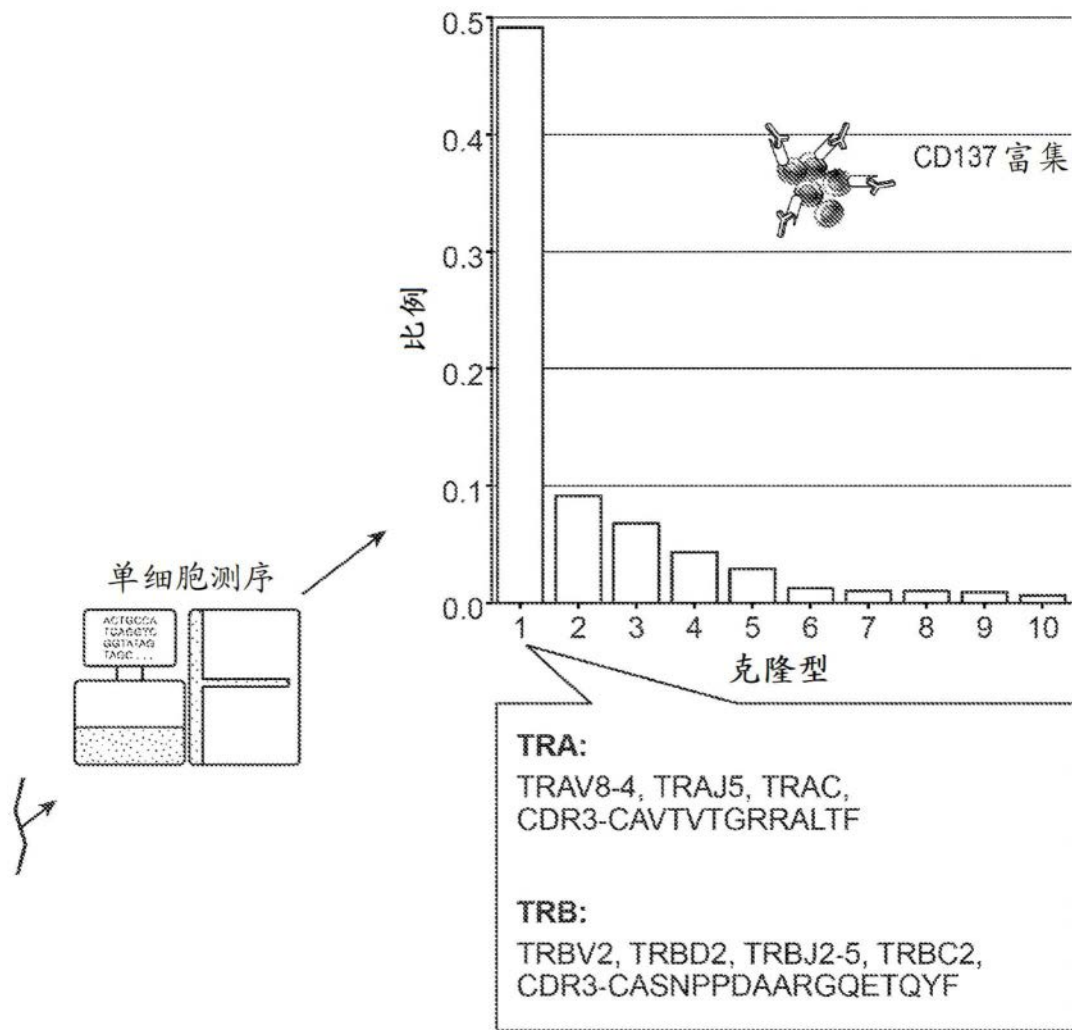


图32(续)

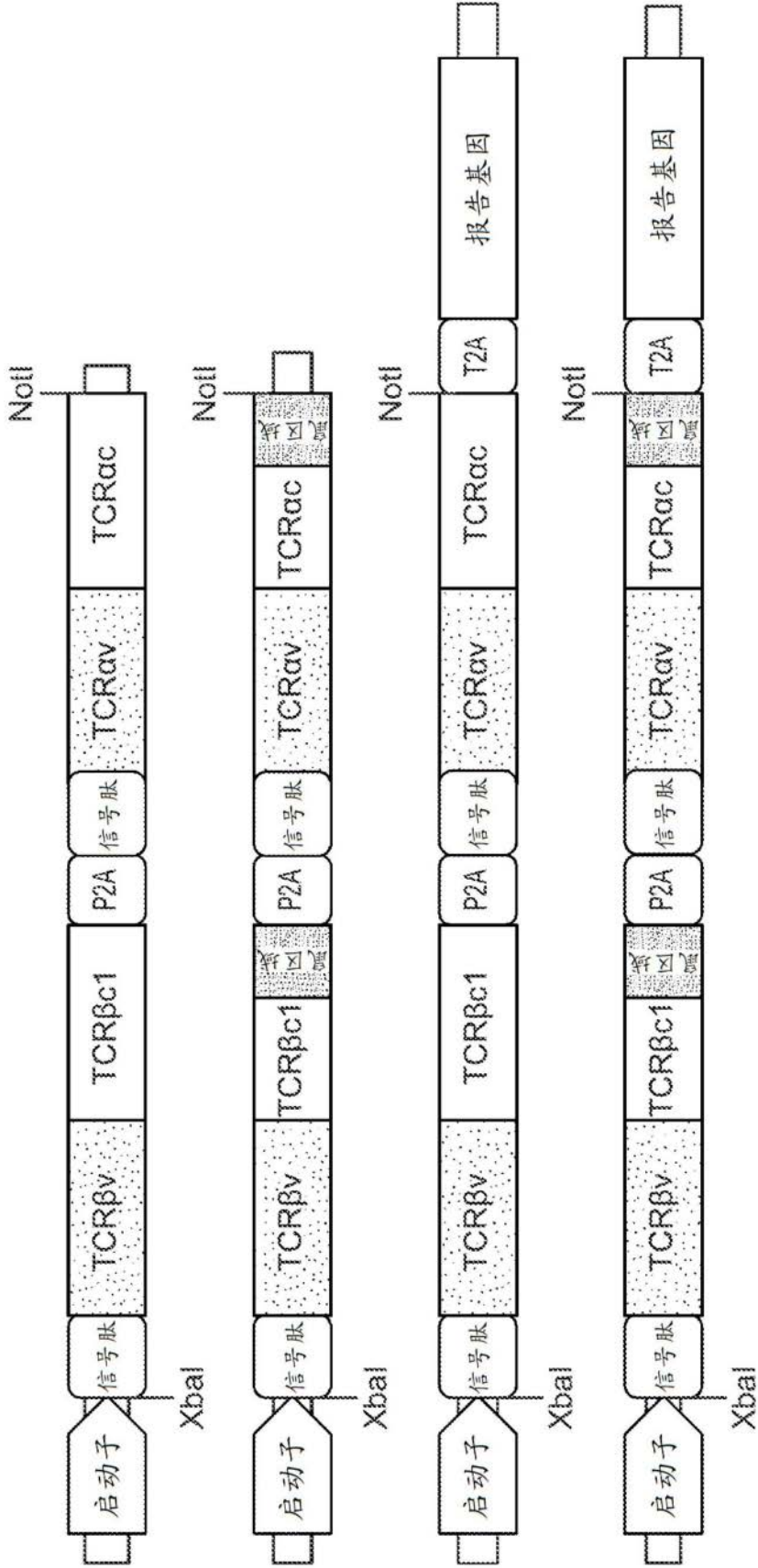


图33

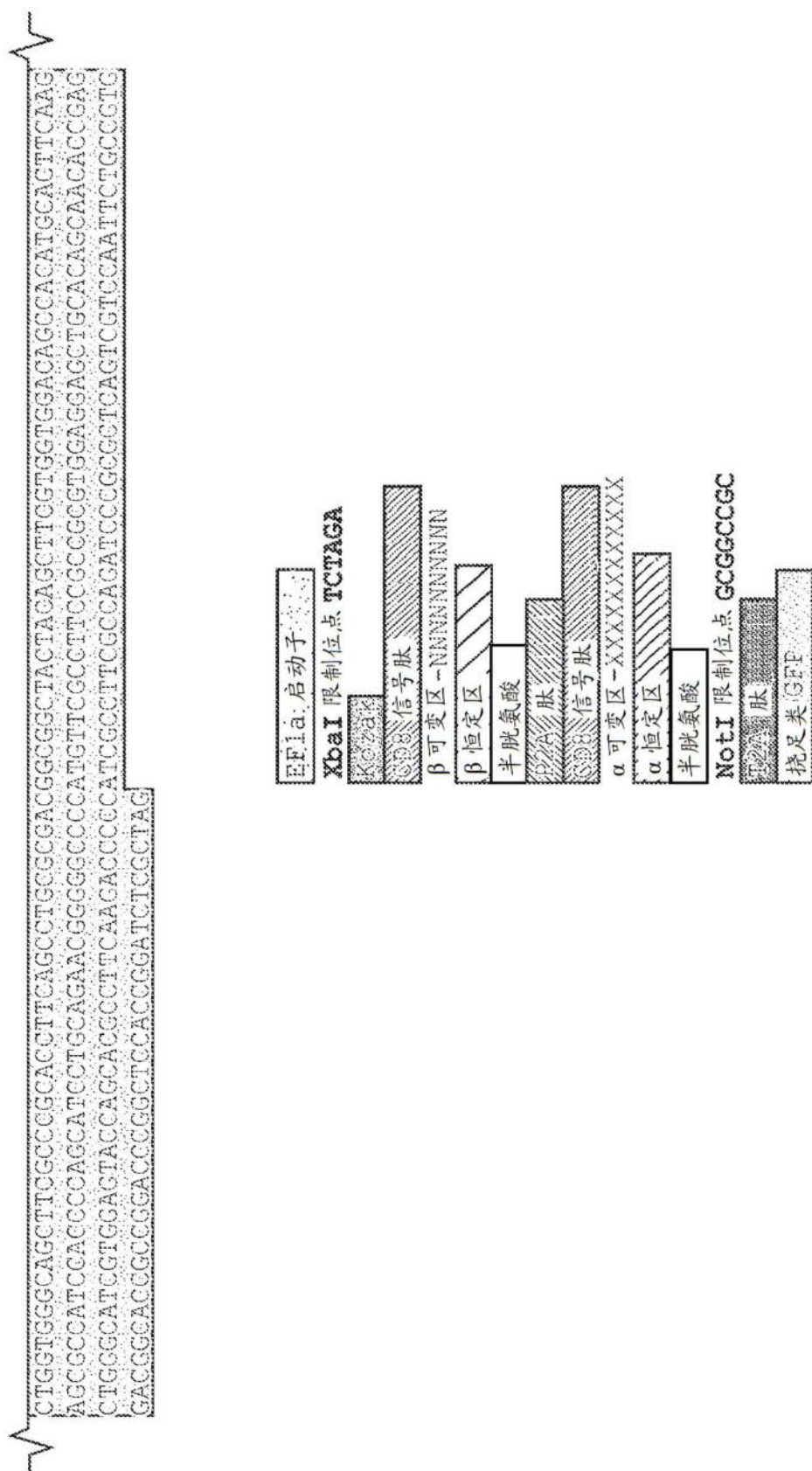


图34(续)

[illegible]

图35

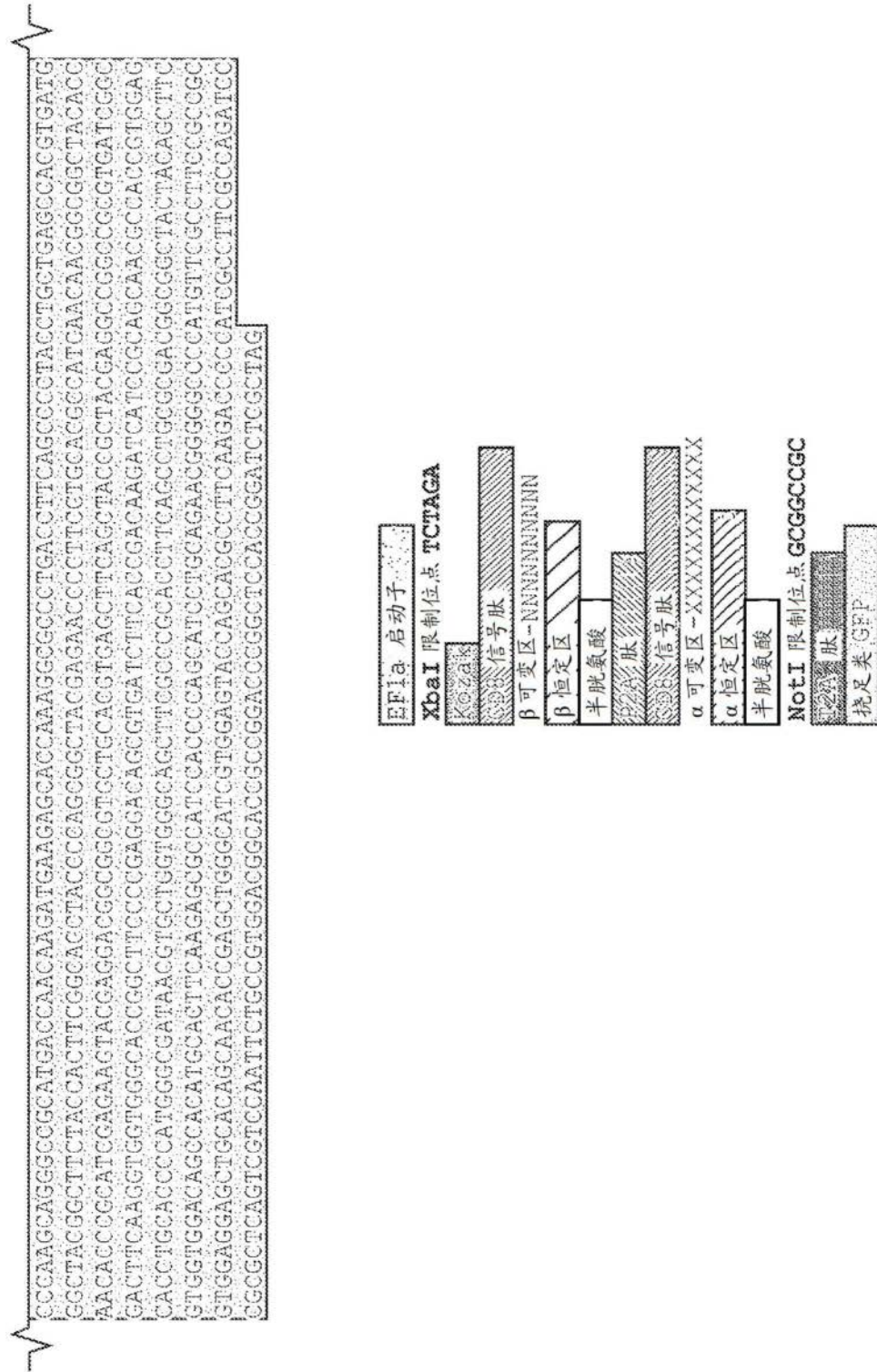


图35(续)

[illegible]

图36

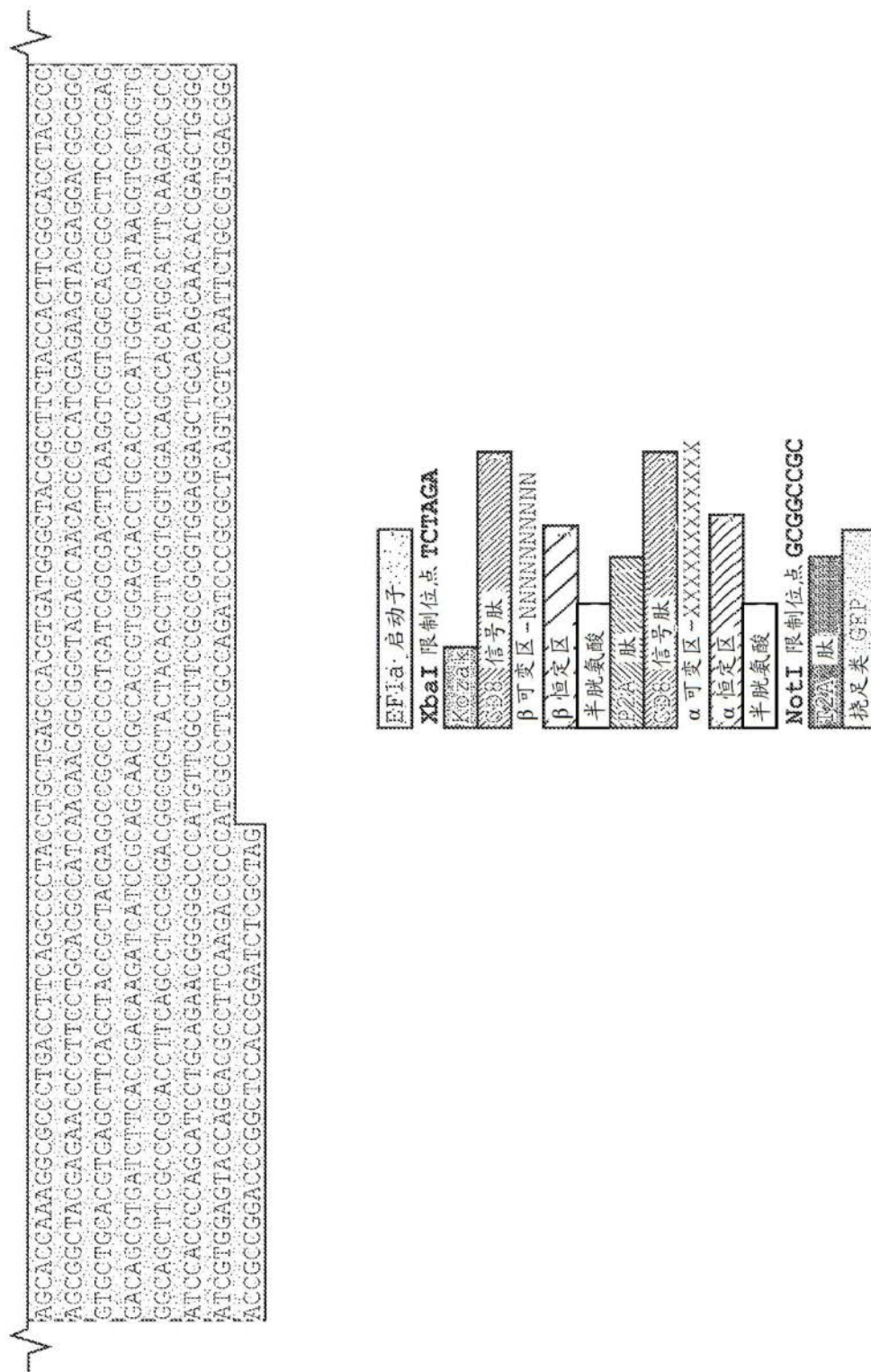


图36(续)

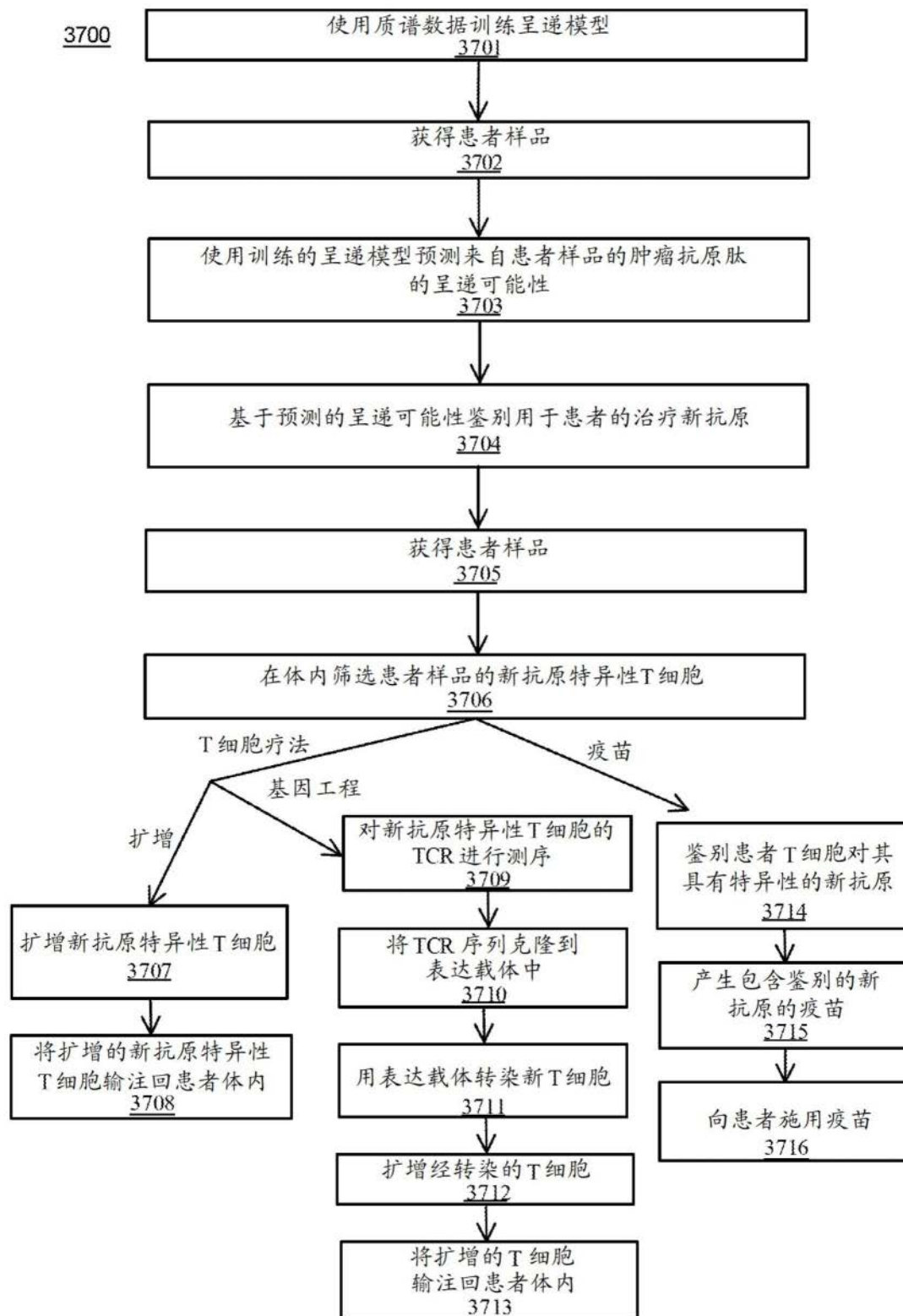


图37

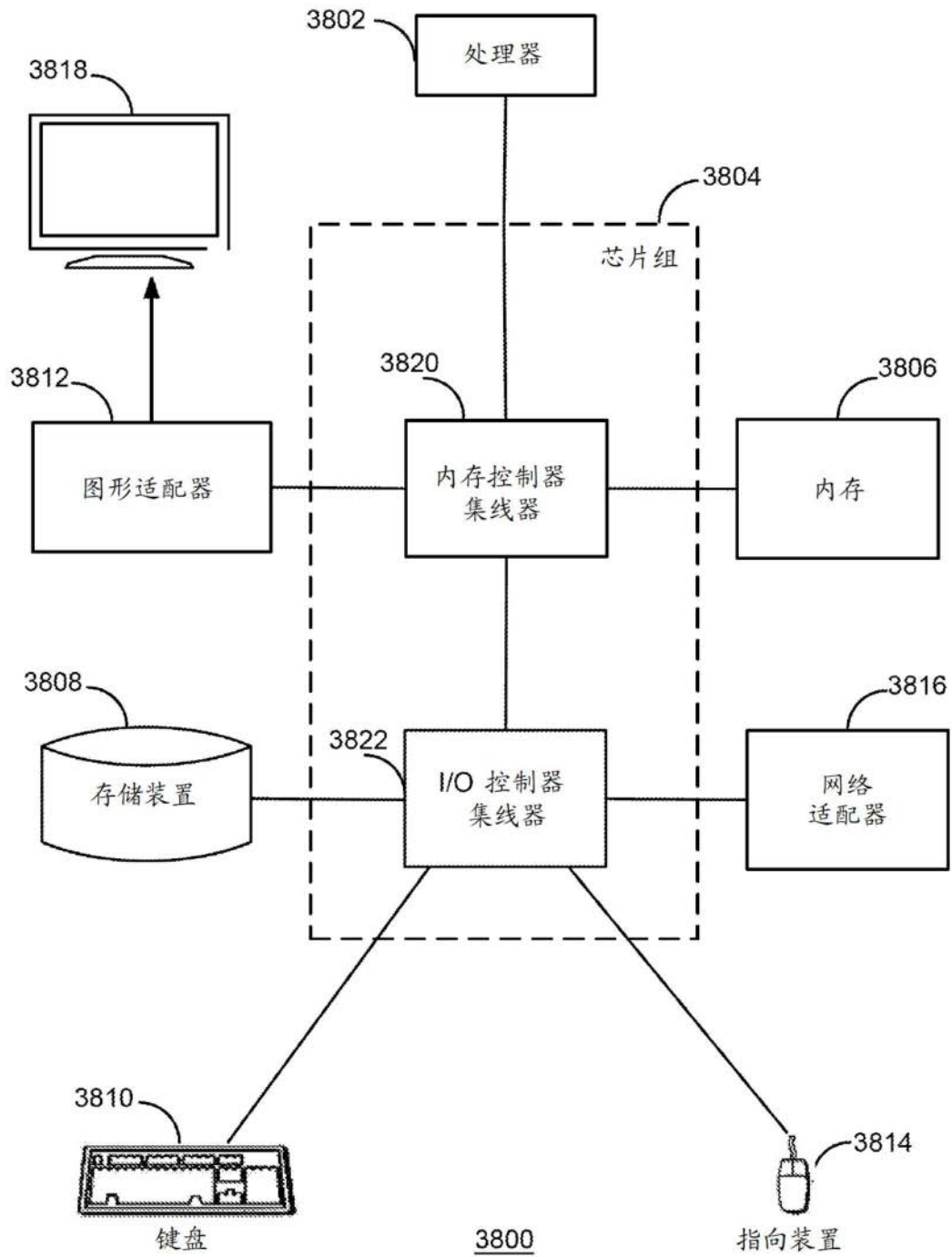


图38