(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113474840 B
(45)授权公告日 2024.12.17

(21)申请号 201980092864.1
(22)申请日 2019.12.20
(65)同一申请的已公布的文献号
申请公布号 CN 113474840 A
(43)申请公布日 2021.10.01
(30)优先权数据
62/783,914 2018.12.21 US
62/826,827 2019.03.29 US
62/855,379 2019.05.31 US
62/891,101 2019.08.23 US
(85)PCT国际申请进入国家阶段日
2021.08.23
(86)PCT国际申请的申请数据
PCT/US2019/068084 2019.12.20
(87)PCT国际申请的公布数据
WO2020/132586 EN 2020.06.25

(73)专利权人 百欧恩泰美国公司
地址 美国马萨诸塞州
(72)发明人 迈克尔·史蒂文·鲁尼
詹尼弗·格雷斯·阿贝林
杜米尼克·巴特尔梅
罗伯特·卡门
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 韦昌金

(51)Int.Cl.
G16B 20/00(2006.01)
(56)对比文件
US 2017212984 A1,2017.07.27
WO 2017184590 A1,2017.10.26
审查员 刘秀

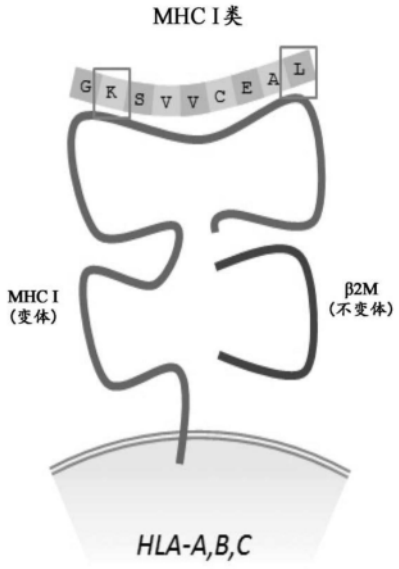
权利要求书7页 说明书150页 附图111页

(54)发明名称

用于预测HLA II类特异性表位及表征CD4+ T细胞的方法和系统

(57)摘要

本公开提供了制备个性化癌症疫苗的方法。本公开还提供了一种训练机器学习HLA-肽呈递预测模型的方法。



CN 113474840 B

1.一种方法,其包括:

(a)使用机器学习HLA肽呈递预测模型处理多个候选肽序列的氨基酸信息,以生成多个呈递预测,其中所述多个候选肽序列中的每个候选肽序列由受试者的基因组或外显子组或受试者中的病原体或病毒编码,其中所述多个呈递预测包括针对所述多个候选肽序列中的每一个的HLA呈递预测,其中每个HLA呈递预测指示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个候选肽序列中的给定候选肽序列的可能性,

其中使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白呈递;以及

(b)至少基于所述多个呈递预测,将所述多个候选肽序列中的肽序列鉴定为由所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种来呈递;

其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的呈递阳性预测值。

2.根据权利要求1所述的方法,其中当处理多个测试肽序列的氨基酸信息以生成多个测试呈递预测,每个测试呈递预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个测试肽序列中的给定测试肽序列的可能性时,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值,其中所述多个测试肽序列包含至少500个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少499个包含在由生物体基因组编码的蛋白质内的诱饵肽序列,其中所述生物体和受试者是同一物种,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少499个诱饵肽序列之比为1:499,并且根据机器学习HLA肽呈递预测模型,排名靠前的百分比的所述多个测试肽序列被预测为将由在细胞中表达的HLA蛋白所呈递。

3.根据权利要求1所述的方法,其中所述至少一个命中肽序列包含至少5、10、20、50或100个命中肽序列。

4.根据权利要求1所述的方法,其中所述至少499个诱饵肽序列包含至少2500、5000、10000、25000、50000或100000个诱饵肽序列。

5.根据权利要求1所述的方法,其中所述至少500个测试肽序列包含至少2500、5000、10000、25000、50000或100000个测试肽序列。

6.根据权利要求1或2所述的方法,其中由HLA蛋白呈递的肽具有15-40个氨基酸的长度。

7.一种方法,其包括:

(a)使用机器学习HLA肽呈递预测模型处理由受试者的基因组或外显子组编码的多个候选肽序列的氨基酸信息,以生成多个呈递预测,其中所述多个呈递预测包括对于所述多个候选肽序列中的每一个的HLA呈递预测,每个呈递预测指示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个候选肽序列中的给定候选肽序列结合的可能性,

其中使用训练数据训练所述机器学习HLA肽呈递预测模型,所述训练数据包含经鉴定与HLA II类蛋白或HLA II类蛋白类似物结合的肽序列的序列信息;以及

(b)至少基于所述多个呈递预测,鉴定所述多个候选肽序列中的肽序列,该肽序列与所

述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种相结合的概率大于阈值呈递预测概率值;

其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.1的呈递阳性预测值。

8.根据权利要求7所述的方法,其中使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白所呈递。

9.根据权利要求1或2所述的方法,其中所述方法包括基于所述多个呈递预测,对被鉴定为由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递的至少两种肽进行排名。

10.根据权利要求9所述的方法,其中所述方法包括从所述至少两种排名的肽中选择一种或多种肽。

11.根据权利要求1或2所述的方法,其中所述方法包括从所述多个候选肽序列中选择一种或多种肽,所述肽被鉴定为由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递。

12.根据权利要求1或2所述的方法,其中所述方法包括从基于所述多个呈递预测排名的两种或更多种肽中选择一种或多种肽。

13.根据权利要求2所述的方法,其中当处理多个测试肽序列的氨基酸信息以生成多个测试呈递预测,每个测试呈递预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个测试肽序列中的给定测试肽序列结合的可能性时,所述机器学习HLA肽呈递预测模型具有至少0.1的阳性预测值,其中所述多个测试肽序列包含至少20个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少19个包含在蛋白质内的诱饵肽序列,所述蛋白质包含至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白如细胞中表达的单个HLA蛋白所呈递的肽序列,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少19个诱饵肽序列之比为1:19,并且根据机器学习HLA肽呈递预测模型,排名靠前的百分比的所述多个测试肽序列被预测为将与在细胞中表达的HLA蛋白结合。

14.根据权利要求13所述的方法,其中所述至少一个命中肽序列与所述至少19个诱饵肽序列之间不存在氨基酸序列重叠。

15.根据权利要求1或2所述的方法,其中所述机器学习HLA肽呈递预测模型的阳性预测值为至少0.08、0.09、0.1、0.11、0.12、0.13、0.14、0.15、0.16、0.17、0.18、0.19、0.2、0.21、0.22、0.23、0.24、0.25、0.26、0.27、0.28、0.29、0.3、0.31、0.32、0.33、0.34、0.35、0.36、0.37、0.38、0.39、0.4、0.41、0.42、0.43、0.44、0.45、0.46、0.47、0.48、0.49、0.5、0.51、0.52、0.53、0.54、0.55、0.56、0.57、0.58、0.59、0.6、0.61、0.62、0.63、0.64、0.65、0.66、0.67、0.68、0.69、0.7、0.71、0.72、0.73、0.74、0.75、0.76、0.77、0.78、0.79、0.8、0.81、0.82、0.83、0.84、0.85、0.86、0.87、0.88、0.89、0.9、0.91、0.92、0.93、0.94、0.95、0.96、0.97、0.98或0.99。

16.根据权利要求2或13所述的方法,其中所述排名靠前的百分比是前0.2%、0.3%、0.4%、0.5%、0.6%、0.7%、0.8%、0.9%、1.00%或2.00%。

17.根据权利要求13所述的方法,其中所述至少一个命中肽序列包含至少5、10、20、50或100个命中肽序列。

18.根据权利要求9所述的方法,其中所述至少19个诱饵肽序列包含至少500、1000、2000、5000或10000个诱饵肽序列。

19.根据权利要求9所述的方法,其中所述至少20个测试肽序列包含至少500、1000、2000、5000或10000个测试肽序列。

20.根据权利要求2或13所述的方法,其中所述排名靠前的百分比是前5%、6%、7%、8%、9%、10%、15%或20%。

21.根据权利要求1或2所述的方法,其中对于由表11的相应HLA等位基因编码的蛋白质,阳性预测值大于表11的第2列的阳性预测值。

22.根据权利要求1或7所述的方法,其中对于由HLA II类等位基因编码的蛋白质,阳性预测值大于表16的第2列的阳性预测值。

23.根据权利要求1或2所述的方法,其中所述受试者是单个受试者。

24.根据权利要求1或2所述的方法,其中所述细胞是表达由受试者细胞的II类HLA等位基因编码的单一蛋白质的细胞。

25.根据权利要求1或2所述的方法,其中所述细胞是单等位基因HLA细胞,或表达具有亲和标签的HLA等位基因的细胞。

26.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列与癌症相关。

27.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的至少一个肽序列被受试者的癌细胞过表达。

28.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列被受试者的癌细胞过表达。

29.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的至少一个肽序列是癌细胞特异性肽。

30.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列是癌细胞特异性肽。

31.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列被受试者的癌细胞表达。

32.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的至少一个肽序列不由受试者的非癌细胞编码。

33.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列不由受试者的非癌细胞编码。

34.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的至少一个肽序列不被受试者的非癌细胞表达。

35.根据权利要求1或2所述的方法,其中所述多个候选肽序列中的每个肽序列不被受试者的非癌细胞表达。

36.根据权利要求1或2所述的方法,其中所述方法包括获得所述受试者的所述多个候选肽序列。

37.根据权利要求1或2所述的方法,其中所述方法包括获得所述受试者的多个多核苷酸序列。

38.根据权利要求1或2所述的方法,其中所述方法包括获得所述受试者的多个多核苷酸序列,所述多核苷酸序列编码由所述受试者的基因组或外显子组编码的多个候选肽序列。

39.根据权利要求1或2所述的方法,其中所述方法包括通过计算机处理器获得所述受试者的多个多核苷酸序列,所述多核苷酸序列编码由所述受试者的基因组或外显子组编码的多个候选肽序列。

40.根据权利要求1或2所述的方法,其中所述方法包括通过基因组或外显子组测序获得所述受试者的多个多核苷酸序列。

41.根据权利要求1或2所述的方法,其中所述方法包括通过全基因组测序或全外显子组测序获得所述受试者的多个多核苷酸序列。

42.根据权利要求1或2所述的方法,其中所述方法进一步配制包含所选择的所述多个候选肽序列的子集中的一种或多种肽序列的组合物。

43.根据权利要求1或2所述的方法,其中HLA II类蛋白包括HLA-DR、HLA-DQ或HLA-DP蛋白。

44.根据权利要求1或2所述的方法,其中HLA II类蛋白选自:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01和HLA-DRB5*01:01;选自下组的HLA-DR蛋白:DPA*01:03/DPB*04:01、DRB1*01:01、DRB1*01:02、DRB1*03:01、DRB1*04:01、DRB1*04:02、DRB1*04:04、DRB1*04:05、DRB1*07:01、DRB1*08:01、DRB1*08:02、DRB1*08:03、DRB1*09:01、DRB1*11:01、DRB1*11:02、DRB1*11:04、DRB1*12:01、DRB1*13:01、DRB1*13:02、DRB1*13:03、DRB1*14:01、DRB1*15:01、DRB1*15:02、DRB1*15:03、DRB1*16:02、DRB3*01:01、DRB3*02:01、DRB3*02:02、DRB3*03:01、DRB4*01:01、DRB4*01:03和DRB5*01:01;选自下组的HLA-DP蛋白:DPB1*01:01、DPB1*02:01、DPB1*02:02、DPB1*03:01、DPB1*04:01、DPB1*04:02、DPB1*05:01、DPB1*06:01、DPB1*11:01、DPB1*13:01、DPB1*17:01;或选自下组的HLA-DQ蛋白复合物:A1*01:01+B1*05:01、A1*01:02+B1*06:02、A1*01:02+B1*06:04、A1*01:03+B1*06:03、A1*02:01+B1*02:02、A1*02:01+B1*03:03、A1*03:01+B1*03:02、A1*03:03+B1*03:01、A1*05:01+B1*02:01和A1*05:05+B1*03:01。

45.根据权利要求42所述的方法,其中所述组合物进一步包括佐剂。

46.根据权利要求42所述的方法,其中所述组合物进一步包括免疫检查点抑制剂。

47.一种鉴定用于受试者的免疫治疗的HLA II类特异性肽的方法,其包括:

(a)通过计算机处理器获得包含表位的候选肽和多个肽序列,每个肽序列均包含所述表位;

(b)使用机器学习HLA-肽呈递预测模型,通过所述计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个向免疫细胞的呈递预测,每个呈递预测指示由HLA II类等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息,并且其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值;

(c)从所述受试者细胞的HLA II类等位基因编码的一种或多种蛋白质中选择蛋白质,所述蛋白质通过机器学习HLA-肽呈递预测模型被预测为与所述候选肽结合,其中所述蛋白质将所述候选肽呈递给免疫细胞的概率大于阈值呈递预测概率值;

(d)使所述候选肽与所选蛋白质接触,使得所述候选肽竞争与所选蛋白质相关联的占位肽;以及

(e)根据所述候选肽是否取代所述占位肽,将所述候选肽鉴定为对所选蛋白质具有特异性的用于免疫治疗的肽。

48.一种测定MHC II类结合肽的免疫原性的方法,其包括:

(a)选择通过机器学习HLA-肽呈递预测模型被预测为与MHCII类结合肽结合的、由HLA II类等位基因编码的蛋白质,其中所述机器学习HLA-肽呈递预测模型被配置为生成针对给定肽序列的呈递预测,该呈递预测指示由所述HLA II类等位基因编码的一种或多种蛋白质能够呈递给定肽序列的可能性,并且其中所述蛋白质呈递所述MHC II类结合肽的概率大于阈值呈递预测概率值,并且其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值;

(b)使所述MHC II类结合肽与所选蛋白质接触,使得所述MHC II类结合肽竞争与所选蛋白质相关联的占位肽,并取代所述占位肽,从而形成包含所述HLA II类等位基因和所述MHC II类结合肽的复合物;

(c)使所述复合物与CD4+T细胞接触,以及

(d)测定所述CD4+T细胞的一种或多种活化参数,所述参数选自:细胞因子的诱导、趋化因子的诱导和细胞表面标志物的表达。

49.一种方法,该方法包括:

(a)鉴定与癌症相关并包含癌症突变的肽序列,其中鉴定所述肽序列包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较;

(b)选择由HLA II类等位基因编码的蛋白质,该蛋白质由受试者的细胞正常表达,并且通过机器学习HLA-肽呈递预测模型被预测为与所述肽序列结合;其中所述机器学习HLA-肽呈递预测模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值,并且其中所述蛋白质呈递所鉴定的肽序列的概率大于阈值呈递预测概率值;

(c)使所鉴定的肽序列与由HLA II类等位基因编码的所选蛋白质接触,以验证所鉴定的肽序列是否竞争与由HLA II类等位基因编码的所选蛋白质相关联的占位肽,从而以小于500nM的IC50值取代该占位肽;以及

(d)任选地,纯化所鉴定的肽。

50.一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,其包括:

(a)通过计算机处理器获得所述多肽序列的多个肽序列;

(b)使用机器学习HLA-肽呈递预测模型,通过计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含与在细胞中表达的HLA蛋白相关的序列信息;以及

(c)基于所述呈递预测,确定或预测所述多肽序列的所述多个肽序列中的每一个对所述受试者不是免疫原性的,并且其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值。

51.一种通过缀合四个单独的HLA蛋白α1和β1异二聚体来制备HLA II类四聚体的方法,该方法包括:

(a)在真核细胞中表达包含编码HLA蛋白的α链和β链的核酸序列、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签的载体,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所述异二聚体与占位肽相关联,

(b)从细胞培养基中纯化所分泌的异二聚体,

(c)使用肽交换测定验证肽结合活性,

(d)加入链霉亲和素,从而将异二聚体缀合成四聚体,

(e)纯化所述四聚体,并具有大于1mg/L的产量。

52.一种包含HLA-DR或HLA-DP或HLA-DQ异二聚体的HLA II类四聚体或多聚体,每个异二聚体包含α链和β链,其中所述异二聚体被纯化并以大于1mg/L的浓度存在。

53.根据权利要求52所述的HLA II类四聚体或多聚体,其中所述HLA II类四具体或多聚体是选自下组的HLA II类蛋白:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01和HLA-DRB5*01:01。

54.一种载体,其包含编码权利要求51-53中任一项的HLA蛋白的α链和β链的核酸序列、

分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,任选地,其中所述分泌的异二聚体与占位肽相关联。

55.一种细胞,其包含权利要求54的载体。

56.一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,该方法包括:

(a)使用计算机处理器将所述多肽序列的肽序列的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的II类MHC等位基因编码的一种或多种蛋白质呈递给定肽序列的表位序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;

(b)基于该组呈递预测,确定或预测所述多肽序列的肽序列中的至少一个对所述受试者是免疫原性的,并且其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的呈递阳性预测值。

57.组合物在制备用于治疗有需要的受试者中的自身免疫性疾病或病况的药物的用途,所述组合物包括表达与复合物结合的T细胞受体的调节性T细胞,该复合物包含:(i)通过机器学习HLA肽呈递预测模型被鉴定或预测为由受试者细胞的II类MHC呈递的所表达蛋白质的表位,其中根据呈递阳性预测值确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的呈递阳性预测值,和(ii)II类MHC,其中所述复合物被所述受试者的CD8或CD4 T细胞所靶向。

用于预测HLA II类特异性表位及表征CD4+ T细胞的方法和系统

[0001] 交叉引用

[0002] 本申请要求2018年12月21日提交的第62/783,914号美国临时申请、2019年3月29日提交的第62/826,827号美国临时申请、2019年5月31日提交的第62/855,379号美国临时申请和2019年8月23日提交的第62/891,101号美国临时申请的权益,所述临时申请中的每一个均通过引用整体并入本文。

背景技术

[0003] 主要组织相容性复合物(MHC)是编码人类白细胞抗原(HLA)基因的基因复合物。HLA基因被表达为在人类细胞表面上展示给循环T细胞的蛋白质异二聚体。HLA基因是高度多态性的,允许其微调适应性免疫系统。适应性免疫应答部分地取决于T细胞鉴别并消除展示与人类白细胞抗原(HLA)异二聚体结合的疾病相关肽抗原的细胞的能力。

[0004] 在人类中,内源性和外源性蛋白质可以被蛋白酶体以及被胞质和内体/溶酶体蛋白酶和肽酶加工成肽,并由MHC基因编码的两类细胞表面蛋白质呈递。这些细胞表面蛋白质被称为人类白细胞抗原(HLA I类和II类),与它们结合并引发免疫应答的一组肽被称为HLA表位。HLA表位是使免疫系统能够探测到危险信号如病原体感染和自身转化的关键成分。CD4+ T细胞识别展示在抗原呈递细胞(APC)如树突细胞和巨噬细胞上的II类MHC(HLA-DR、HLA-DQ和HLA-DP)表位。HLA II类配体的内源性加工和呈递是一个复杂的过程,涉及多种分子伴侣和酶的子集,它们尚未全部充分表征。HLA II类肽呈递激活辅助性T细胞,随后促进B细胞分化和抗体产生以及CTL应答。活化的辅助性T细胞还分泌激活并诱导其他T细胞分化的细胞因子和趋化因子。

[0005] 了解每种HLA II类异二聚体的肽结合偏好是成功预测哪些癌症或肿瘤特异性抗原可能引发癌症或肿瘤特异性T细胞应答的关键。对鉴定和分离特定HLA II类关联肽(例如,新抗原肽)的方法存在需求。这样的方法和分离的分子可用于例如开发治疗剂,包括但不限于基于免疫的治疗剂。

发明内容

[0006] 本文描述的方法和组合物可用于广泛的应用中。例如,本文描述的方法和组合物可以用来鉴定免疫原性抗原肽,并且可以用来开发药物,如个性化药物,以及抗原特异性T细胞的分离和表征。

[0007] CD4+ T细胞应答可具有抗肿瘤活性。可以在不使用II类预测的情况下显示高CD4+ T细胞响应率(例如,NeoVax研究中60%的SLP表位(NT-001中的49%,参见Ott等人,Nature,2017年7月13日;547(7662):217-221),以及Biontech研究中48%的mRNA表位,参见Sahin等人,Nature,2017年7月13日;547(7662):222-226)。可能不清楚这些表位是否通常(由肿瘤或吞噬性DC)天然呈递。可能希望通过改进对真正呈递的HLA II类结合表位的鉴定,将高CD4+ T响应率转化为治疗效果。

[0008] 基因表达、酶切和途径/定位偏好的作用可能尚未得到强有力的量化。可能尚不清楚是自噬(肿瘤细胞的HLA II类呈递)还是吞噬作用(APC对肿瘤表位的HLA II类呈递)是更相关的途径,尽管大多数现有MS数据可能被推定来源于自噬。NetMHCIIpan可能是当前的预测标准,但可能认为其并不准确。在三个HLA II类基因座(DR、DP和DQ)中,可能仅HLA-DR的某些常见等位基因存在数据。

[0009] 可能有不同的数据生成方法来学习HLA II类呈递的规则,包括现场标准和提议的方法。现场标准可包括亲和力测量,这可能是NetMHCIIpan预测器的基础,提供低通量并且需要放射性试剂,并且它省略了处理的作用。提议的方法可包括质谱法,其中来自细胞系/组织/肿瘤的数据可以帮助确定自噬的处理规则,而单等位基因MS可以允许确定等位基因特异性结合规则(假设多等位基因MS数据对于有效学习而言过于复杂(Bassani-Sternberg.MCP.2018))。

[0010] 可能有不同的方法来验证新的HLA II类预测器:保留的MS数据的验证,这可能是默认设置;疫苗研究的回顾(例如NT-001),其中免疫监测数据可以评估APC上的疫苗肽负载,而不是肿瘤呈递,并且数据可以在许多不同的等位基因中稀疏地扩展;生化亲和力测量,其可被配置为获得不一致预测的肽的测量值(仅用于2-3个等位基因);T细胞诱导,其可被配置为测试Neon偏好的和NetMHCIIpan偏好的表位诱导离体T细胞应答的速率。

[0011] 对于通过T细胞诱导进行的验证,默认方法可包括评估来自TCGA的预测不一致的neoORF,其中诱导材料可包括健康供体APC和T细胞,并且诱导和读出可以通过SLP(约15聚体肽)。随机肽可产生高反应率,而SLP可能无法充分解决处理问题。可能的解决方案可包括通过mRNA进行的诱导。

[0012] 本文公开的方法可包括生成LC-MS/MS单等位基因数据,用于训练用于表位预测的等位基因特异性机器学习方法。此类方法可包括使用一组质量度量来提高LC-MS/MS数据质量,以严格去除增加预测模型性能的假阳性;从HLA-配体组LC-MS/MS数据集鉴定等位基因特异性HLA II类结合核心;利用机器学习算法改进HLA II类配体和表位预测;并且/或者鉴定影响HLA II类-配体呈递并改善HLA II类表位预测的生物变量,如基因表达、可切割性、基因偏好、细胞定位和二级结构。

[0013] 本文提供了一种方法,其包括:(a)使用机器学习HLA肽呈递预测模型处理多个候选肽序列的氨基酸信息,以生成多个呈递预测,其中所述多个候选肽序列中的每个候选肽序列由受试者的基因组或外显子组编码,其中所述多个呈递预测包括针对所述多个候选肽序列中的每一个的HLA呈递预测,其中每个HLA呈递预测指示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个候选肽序列中的给定候选肽序列的可能性,其中使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白呈递;以及(b)至少基于所述多个呈递预测,将所述多个肽序列中的肽序列鉴定为由所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种来呈递;其中根据呈递阳性预测值(PPV)确定方法,所述机器学习HLA肽呈递预测模型具有至少0.07的PPV。

[0014] 本文提供了一种方法,其包括:(a)使用机器学习HLA肽结合预测模型处理由受试者的基因组或外显子组编码的多个肽序列的氨基酸信息,以生成多个结合预测,其中所述多个结合预测包括对于所述多个候选肽序列中的每一个的HLA结合预测,每个结合预测指

示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个候选肽序列中的给定候选肽序列结合的可能性,其中使用训练数据训练所述机器学习HLA肽结合预测模型,所述训练数据包含经鉴定与HLA II类蛋白或HLA II类蛋白类似物结合的肽序列的序列信息;以及(b)至少基于所述多个结合预测,鉴定所述多个肽序列中的肽序列,该肽序列与所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种相结合的概率大于阈值结合预测概率值;其中根据结合阳性预测值(PPV)确定方法,所述机器学习HLA肽结合预测模型具有至少0.1的PPV。

[0015] 在一些实施方案中,使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白呈递。

[0016] 在一些实施方案中,所述方法包括基于所述呈递预测,对被鉴定为由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递的至少两种肽进行排名。

[0017] 在一些实施方案中,所述方法包括从所述两种或更多种排名的肽中选择一种或多种肽。

[0018] 在一些实施方案中,所述方法包括从所述多种肽中选择一种或多种肽,所述肽被鉴定为由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递。

[0019] 在一些实施方案中,所述方法包括从基于所述呈递预测排名的两种或更多种肽中选择一种或多种肽。

[0020] 在一些实施方案中,当处理多个测试肽序列的氨基酸信息以生成多个测试呈递预测,每个测试呈递预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个测试肽序列中的给定测试肽序列的可能性时,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值(PPV),其中所述多个测试肽序列包含至少500个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少499个包含在由生物体基因组编码的蛋白质内的诱饵肽序列,其中所述生物体和受试者是同一物种,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少499个诱饵肽序列之比为1:499,并且根据机器学习HLA肽呈递预测模型,排名靠前的百分比的所述多个测试肽序列被预测为将由在细胞中表达的HLA蛋白所呈递。

[0021] 在一些实施方案中,当处理多个测试肽序列的氨基酸信息以生成多个测试结合预测,每个测试结合预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个测试肽序列中的给定测试肽序列结合的可能性时,所述机器学习HLA肽呈递预测模型具有至少0.1的阳性预测值(PPV),其中所述多个测试肽序列包含至少20个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少19个包含在蛋白质内的诱饵肽序列,所述蛋白质包含至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白如细胞(例如,单等位基因细胞)中表达的单个HLA蛋白所呈递的肽序列,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少19个诱饵肽序列之比为1:19,并且根据机器学习HLA肽呈递预测模型,排名靠前的百分比的所述多个测试肽序列被预测为将与在细胞中表达的HLA蛋白结合。

[0022] 在一些实施方案中,所述至少一个命中肽序列与所述诱饵肽序列之间不存在氨基酸序列重叠。

[0023] 在一些实施方案中,所述机器学习HLA肽呈递预测模型的阳性预测值(PPV)为至少0.08、0.09、0.1、0.11、0.12、0.13、0.14、0.15、0.16、0.17、0.18、0.19、0.2、0.21、0.22、0.23、0.24、0.25、0.26、0.27、0.28、0.29、0.3、0.31、0.32、0.33、0.34、0.35、0.36、0.37、0.38、0.39、0.4、0.41、0.42、0.43、0.44、0.45、0.46、0.47、0.48、0.49、0.5、0.51、0.52、0.53、0.54、0.55、0.56、0.57、0.58、0.59、0.6、0.61、0.62、0.63、0.64、0.65、0.66、0.67、0.68、0.69、0.7、0.71、0.72、0.73、0.74、0.75、0.76、0.77、0.78、0.79、0.8、0.81、0.82、0.83、0.84、0.85、0.86、0.87、0.88、0.89、0.9、0.91、0.92、0.93、0.94、0.95、0.96、0.97、0.98或0.99。

[0024] 在一些实施方案中,所述至少一个命中肽序列包含至少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或100个命中肽序列。

[0025] 在一些实施方案中,所述至少499个诱饵肽序列包含至少500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个诱饵肽序列。本领域技术人员能够认识到,改变命中:诱饵比改变了PPV。

[0026] 在一些实施方案中,所述至少500个测试肽序列包含至少600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、

35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个测试肽序列。

[0027] 在一些实施方案中,所述排名靠前的百分比是前0.20%、0.30%、0.40%、0.50%、0.60%、0.70%、0.80%、0.90%、1.00%、1.10%、1.20%、1.30%、1.40%、1.50%、1.60%、1.70%、1.80%、1.90%、2.00%、2.10%、2.20%、2.30%、2.40%、2.50%、2.60%、2.70%、2.80%、2.90%、3.00%、3.10%、3.20%、3.30%、3.40%、3.50%、3.60%、3.70%、3.80%、3.90%、4.00%、4.10%、4.20%、4.30%、4.40%、4.50%、4.60%、4.70%、4.80%、4.90%、5.00%、5.10%、5.20%、5.30%、5.40%、5.50%、5.60%、5.70%、5.80%、5.90%、6.00%、6.10%、6.20%、6.30%、6.40%、6.50%、6.60%、6.70%、6.80%、6.90%、7.00%、7.10%、7.20%、7.30%、7.40%、7.50%、7.60%、7.70%、7.80%、7.90%、8.00%、8.10%、8.20%、8.30%、8.40%、8.50%、8.60%、8.70%、8.80%、8.90%、9.00%、9.10%、9.20%、9.30%、9.40%、9.50%、9.60%、9.70%、9.80%、9.90%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%。

[0028] 在一些实施方案中,所述至少一个命中肽序列包含至少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或100个命中肽序列。

[0029] 在一些实施方案中,所述至少19个诱饵肽序列包含至少30、40、50、60、70、80、90、100、110、120、130、140、150、160、170、180、190、200、210、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个诱饵肽序列。

[0030] 在一些实施方案中,所述至少20个测试肽序列包含包含至少30、40、50、60、70、80、90、100、110、120、130、140、150、160、170、180、190、200、210、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个测试肽序列。

[0031] 在一些实施方案中,所述排名靠前的百分比是前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%或40%。

[0032] 在一些实施方案中,对于由表11的相应HLA等位基因编码的蛋白质,PPV大于表11的第2列的相应PPV。在一些实施方案中,对于由表11的相应HLA等位基因编码的蛋白质,PPV至少等于表11的第3列的相应PPV。

[0033] 在一些实施方案中,对于由HLA II类等位基因编码的蛋白质,PPV等于或大于表12的第2列的相应PPV。

[0034] 在一些实施方案中,对于由HLA II类等位基因编码的蛋白质,PPV大于表16的第2列的相应PPV。

[0035] 在一些实施方案中,所述受试者是单个受试者。

[0036] 在一些实施方案中,所述受试者是哺乳动物。

[0037] 在一些实施方案中,所述受试者是人。

[0038] 在一些实施方案中,所述训练细胞是表达由受试者细胞的II类HLA等位基因编码的单一蛋白质的细胞。

[0039] 在一些实施方案中,所述训练细胞是单等位基因HLA细胞,或表达具有亲和标签的HLA等位基因的细胞。

[0040] 在一些实施方案中,所述受试者的细胞包括癌细胞。

[0041] 在一些实施方案中,所述方法用于鉴定肽序列。

[0042] 在一些实施方案中,所述方法用于选择肽序列。

[0043] 在一些实施方案中,所述方法用于准备癌症治疗。

[0044] 在一些实施方案中,所述方法用于准备受试者特定的癌症治疗。

[0045] 在一些实施方案中,所述方法用于准备癌细胞特定的癌症治疗。

[0046] 在一些实施方案中,所述多个肽序列中的每个肽序列与癌症相关。

[0047] 在一些实施方案中,所述多个肽序列中的至少一个肽序列被受试者的癌细胞过表达。

[0048] 在一些实施方案中,所述多个肽序列中的每个肽序列被受试者的癌细胞过表达。

[0049] 在一些实施方案中,所述多个肽序列中的至少一个肽序列是癌细胞特异性肽。

[0050] 在一些实施方案中,所述多个肽序列中的每个肽序列是癌细胞特异性肽。

[0051] 在一些实施方案中,所述多个肽序列中的每个肽序列被受试者的癌细胞表达。

[0052] 在一些实施方案中,所述多个肽序列中的至少一个肽序列不由受试者的非癌细胞编码。

[0053] 在一些实施方案中,所述多个肽序列中的每个肽序列不由受试者的非癌细胞编码。

[0054] 在一些实施方案中,所述多个肽序列中的至少一个肽序列不被受试者的非癌细胞表达。

[0055] 在一些实施方案中,所述多个肽序列中的每个肽序列不被受试者的非癌细胞表达。

[0056] 在一些实施方案中,所述方法包括获得受试者的所述多个肽序列。

[0057] 在一些实施方案中,所述方法包括获得受试者的多个多核苷酸序列。

[0058] 在一些实施方案中,所述方法包括获得受试者的多个多核苷酸序列,所述多核苷酸序列编码由受试者的基因组或外显子组或由受试者中的病原体或病毒编码的多个肽序列。

[0059] 在一些实施方案中,所述方法包括通过计算机处理器获得受试者的多个多核苷酸序列,所述多核苷酸序列编码由受试者的基因组或外显子组编码的多个肽序列。

[0060] 在一些实施方案中,所述方法包括通过基因组或外显子组测序获得受试者的多个多核苷酸序列。

[0061] 在一些实施方案中,所述方法包括通过全基因组测序或全外显子组测序获得受试者的多个多核苷酸序列。

[0062] 在一些实施方案中,处理包括用计算机处理器处理。

[0063] 在一些实施方案中,处理包括至少基于所述多个肽序列的氨基酸信息生成多个预测变量。

[0064] 在一些实施方案中,使用机器学习HLA-肽呈递预测模型处理所述多个预测变量。

[0065] 在一些实施方案中,由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质是由受试者表达的II类HLA等位基因编码的一种或多种蛋白质。

[0066] 在一些实施方案中,由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质是由受试者的癌细胞表达的II类HLA等位基因编码的一种或多种蛋白质。

[0067] 在一些实施方案中,由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质是由受试者细胞的II类HLA等位基因编码的单一蛋白质。

[0068] 在一些实施方案中,由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质是由受试者细胞的II类HLA等位基因编码的两种、三种、四种、五种或六种或更多种蛋白质。

[0069] 在一些实施方案中,由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质是由受试者细胞的II类HLA等位基因编码的每种蛋白质。

[0070] 在一些实施方案中,所述方法进一步包括向受试者施用包含所选择的肽序列子集中的一种或多种肽序列的组合物。

[0071] 在一些实施方案中,鉴定所述多个肽序列包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较,其中所述多个肽中的每一个包含至少一个突变,所述突变存在于受试者的癌细胞中,但不存在于受试者的正常细胞中。

[0072] 在一些实施方案中,所述机器学习HLA-肽呈递预测模型包括至少基于训练数据鉴定的多个预测变量,其中所述训练数据包括包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表所述氨基酸位置信息与基于所述氨基酸位置信息和所述多个预测变量作为输出生成的呈递可能性之间的关系的函数。

[0073] 在一些实施方案中,鉴定包括至少基于所述多个呈递预测,鉴定所述多个肽序列中的肽序列,该肽序列被所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种呈递的概率大于阈值呈递预测概率值。

[0074] 在一些实施方案中,通过机器学习HLA肽呈递预测模型被预测为将被呈递的所述多个测试肽序列的0.2%中的一个或多个被受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递的概率大于阈值呈递预测概率值。

[0075] 在一些实施方案中,通过机器学习HLA肽呈递预测模型被预测为将被呈递的所述多个测试肽序列的0.2%中的每一个被受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种所呈递的概率大于阈值呈递预测概率值。

[0076] 在一些实施方案中,阳性的数目被限制为等于命中的数目。

[0077] 在一些实施方案中,所述质谱法是单等位基因质谱法。

[0078] 在一些实施方案中,所述肽通过自噬由在细胞中表达的HLA蛋白呈递。

[0079] 在一些实施方案中,所述肽通过吞噬作用由在细胞中表达的HLA蛋白呈递。

[0080] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的表达水平预测值。

[0081] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的稳定性预测值。

[0082] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的降解速率预测值。

[0083] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的蛋白质可切割性预测值。

[0084] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的细胞或组织定位预测值。

[0085] 在一些实施方案中,所述多个预测变量包括包含肽的源蛋白质的细胞内加工模式的预测值,其中源蛋白质的加工模式尤其包括源蛋白质是否经历自噬、吞噬作用和细胞内转运的预测值。

[0086] 在一些实施方案中,通过使用多个质量度量来提高训练数据的质量。

[0087] 在一些实施方案中,所述多个质量度量包括常见污染物肽去除、高评分峰强度、高

评分和高质量准确度。

[0088] 在一些实施方案中,评分峰值强度至少为50%。

[0089] 在一些实施方案中,评分峰值强度至少为60%。

[0090] 在一些实施方案中,评分至少为7。

[0091] 在一些实施方案中,质量准确度至多为5ppm。

[0092] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一免疫沉淀的HLA蛋白呈递的肽。

[0093] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一外源HLA蛋白呈递的肽。

[0094] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一重组HLA蛋白呈递的肽。

[0095] 在一些实施方案中,所述多个预测变量包括肽-HLA亲和力预测变量。

[0096] 在一些实施方案中,由HLA蛋白呈递的肽包括通过搜索无酶特异性无修饰肽数据库而鉴定的肽。

[0097] 在一些实施方案中,由HLA蛋白呈递的肽包括通过使用反转数据库搜索策略搜索肽数据库而鉴定的肽。

[0098] 在一些实施方案中,所述HLA蛋白包括HLA-DR、HLA-DQ或HLA-DP蛋白。

[0099] 在一些实施方案中,所述HLA蛋白包括选自下组的HLA II类蛋白:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01、HLA-DRB5*01:01。

[0100] 在一些实施方案中,所述HLA-DR与DRA*01:01配对。

[0101] 在一些实施方案中,所述HLA蛋白是选自下组的HLA II类蛋白:DPA*01:03/DPB*04:01、DRB1*01:01、DRB1*01:02、DRB1*03:01、DRB1*04:01、DRB1*04:02、DRB1*04:04、DRB1*04:05、DRB1*07:01、DRB1*08:01、DRB1*08:02、DRB1*08:03、DRB1*09:01、DRB1*11:01、DRB1*11:02、DRB1*11:04、DRB1*12:01、DRB1*13:01、DRB1*13:02、DRB1*13:03、DRB1*14:01、DRB1*15:01、DRB1*15:02、DRB1*15:03、DRB1*16:02、DRB3*01:01、DRB3*02:01、DRB3*02:02、DRB3*03:01、DRB4*01:01、DRB4*01:03和DRB5*01:01。

[0102] 在一些实施方案中,所述HLA-DR蛋白包含二聚体形式的DRA*01:01。

[0103] 在一些实施方案中,所述HLA蛋白包括选自下组的HLA-DP蛋白:DPB1*01:01、DPB1*02:01、DPB1*02:02、DPB1*03:01、DPB1*04:01、DPB1*04:02、DPB1*05:01、DPB1*06:01、DPB1*

11:01、DPB1*13:01、DPB1*17:01。

[0104] 在一些实施方案中,所述HLA-DP蛋白配对包含DPA1*01:03。

[0105] 在一些实施方案中,所述HLA蛋白包括选自下组的HLA-DQ蛋白复合物:A1*01:01+B1*05:01、A1*01:02+B1*06:02、A1*01:02+B1*06:04、A1*01:03+B1*06:03、A1*02:01+B1*02:02、A1*02:01+B1*03:03、A1*03:01+B1*03:02、A1*03:03+B1*03:01、A1*05:01+B1*02:01和A1*05:05+B1*03:01。

[0106] 在一些实施方案中,由HLA蛋白呈递的肽包括通过将HLA-肽的MS/MS谱与肽或蛋白质数据库中的一种或多种肽或蛋白质的MS/MS谱进行比较而鉴定的肽。

[0107] 在一些实施方案中,所述突变选自点突变、剪接位点突变、移码突变、通读突变和基因融合突变。

[0108] 在一些实施方案中,所述由HLA蛋白呈递的肽具有15-40个氨基酸的长度。

[0109] 在一些实施方案中,由HLA蛋白呈递的肽包括通过将HLA-肽的MS/MS谱与肽或蛋白质数据库中的一种或多种肽或蛋白质的MS/MS谱进行比较而鉴定由HLA蛋白呈递的肽而鉴定的肽。

[0110] 在一些实施方案中,所述个性化癌症治疗进一步包括佐剂。

[0111] 在一些实施方案中,所述个性化癌症治疗进一步包括免疫检查点抑制剂。

[0112] 在一些实施方案中,所述训练数据包括结构化数据、时间序列数据、非结构化数据、关系数据或其任何组合。

[0113] 在一些实施方案中,所述非结构化数据包括图像数据。

[0114] 在一些实施方案中,所述关系数据包括来自客户系统、企业系统、操作系统、网站、网络可访问应用程序界面(API)或其任何组合的数据。

[0115] 在一些实施方案中,所述训练数据被上传到基于云的数据库。

[0116] 在一些实施方案中,所述训练使用卷积神经网络来进行。

[0117] 在一些实施方案中,所述卷积神经网络包含至少两个卷积层。

[0118] 在一些实施方案中,所述卷积神经网络包括至少一个批归一化步骤。

[0119] 在一些实施方案中,所述卷积神经网络包括至少一个空间丢弃(spatial dropout)步骤。

[0120] 在一些实施方案中,所述卷积神经网络包括至少一个全局最大池化步骤。

[0121] 在一些实施方案中,所述卷积神经网络包含至少一个密集层。

[0122] 在一些实施方案中,鉴定肽序列包括鉴定在受试者的癌细胞中表达的具有突变的肽序列。

[0123] 在一些实施方案中,鉴定肽序列包括鉴定在受试者的正常细胞中不表达的肽序列。

[0124] 在一些实施方案中,鉴定肽序列包括鉴定病毒肽序列。

[0125] 在一些实施方案中,鉴定肽序列包括鉴定过表达的肽序列。

[0126] 本文提供了一种鉴定用于受试者的免疫治疗的HLA II类特异性肽的方法,其包括:通过计算机处理器获得包含表位的候选肽和多个肽序列,每个肽序列均包含所述表位;使用机器学习HLA-肽呈递预测模型,通过计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个向免疫细胞的呈递预测,每个呈递预测指示由

HLAII类等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;从受试者细胞的HLA II类等位基因编码的一种或多种蛋白质中选择蛋白质,所述蛋白质通过机器学习HLA-肽呈递预测模型被预测为与所述候选肽结合,其中所述蛋白质将所述候选肽呈递给免疫细胞的概率大于阈值呈递预测概率值;使所述候选肽与所选蛋白质接触,使得所述候选肽竞争与所选蛋白质相关联的占位肽;以及根据所述候选肽是否取代所述占位肽,将所述候选肽鉴定为对所选蛋白质具有特异性的用于免疫治疗的肽。

[0127] 在一些实施方案中,获得包括鉴定候选肽,其中鉴定候选肽包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较。

[0128] 在一些实施方案中,处理包括至少基于所述多个肽序列的氨基酸信息确定多个预测变量,并使用机器学习HLA-肽呈递预测模型处理所述多个预测变量。

[0129] 在一些实施方案中,所述机器学习HLA-肽呈递预测模型包括至少基于训练数据鉴定的多个预测变量,其中所述训练数据包括包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表所述氨基酸位置信息与基于所述氨基酸位置信息和所述多个预测变量作为输出生成的呈递可能性之间的关系的函数。

[0130] 在一些实施方案中,阳性的数目被限制为等于命中的数目。

[0131] 在一些实施方案中,所述质谱法是单等位基因质谱法。

[0132] 在一些实施方案中,所述多个预测变量包括以下任一个或多个:包含肽的源蛋白质的表达水平预测值、稳定性预测值、降解速率预测值、可切割性预测值、细胞或组织定位预测值,以及包括自噬、吞噬作用和细胞内转运在内的细胞内加工模式预测值。

[0133] 在一些实施方案中,通过使用多个质量度量来提高训练数据的质量。

[0134] 在一些实施方案中,所述多个质量度量包括常见污染物肽去除、高评分峰强度、高评分和高质量准确度。

[0135] 在一些实施方案中,评分峰值强度至少为50%。

[0136] 在一些实施方案中,评分峰值强度至少为60%。

[0137] 在一些实施方案中,所述占位肽是CLIP肽。

[0138] 在一些实施方案中,所述占位肽是CMV肽。

[0139] 在一些实施方案中,所述方法进一步包括测量所述占位肽被目标肽取代的IC50。

[0140] 在一些实施方案中,所述占位肽被目标肽取代的IC50小于500nM。

[0141] 在一些实施方案中,来自由受试者细胞的HLA II类等位基因编码的一种或多种蛋白质的至少一种蛋白质是HLA II类四聚体或多聚体。

[0142] 在一些实施方案中,所述目标肽进一步通过质谱法来鉴定。

[0143] 在一些实施方案中,所述由受试者细胞的HLA II类等位基因编码的至少一种蛋白质是重组蛋白。

[0144] 在一些实施方案中,所述由受试者细胞的HLA II类等位基因编码的至少一种蛋白质在真核细胞中表达。

[0145] 在一些实施方案中,所述肽通过自噬由在细胞中表达的HLA蛋白呈递。

[0146] 在一些实施方案中,所述肽通过吞噬作用由在细胞中表达的HLA蛋白呈递。

[0147] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一免疫沉淀的HLA蛋白呈递的肽。

[0148] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一外源HLA蛋白呈递的肽。

[0149] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一重组HLA蛋白呈递的肽。

[0150] 在一些实施方案中,所述多个预测变量包括肽-HLA亲和力预测变量。

[0151] 在一些实施方案中,由HLA蛋白呈递的肽包括通过搜索无酶特异性无修饰肽数据库而鉴定的肽。

[0152] 在一些实施方案中,由HLA蛋白呈递的肽包括通过使用反转数据库搜索策略搜索肽数据库而鉴定的肽。

[0153] 在一些实施方案中,所述HLA蛋白包括HLA-DR、HLA-DQ或HLA-DP蛋白。

[0154] 在一些实施方案中,所述免疫治疗是癌症免疫治疗。

[0155] 在一些实施方案中,所述表位是癌症特异性表位。

[0156] 在一些实施方案中,所述由HLAII类等位基因编码的至少一种蛋白质至少包含以二聚体形式存在的HLA蛋白的α1亚单位和β1亚单位。

[0157] 在一些实施方案中,所述肽的身份是已知的。

[0158] 在一些实施方案中,所述肽的身份是未知的。

[0159] 在一些实施方案中,所述肽的身份通过质谱法确定。

[0160] 在一些实施方案中,肽交换测定包括检测肽荧光探针或标签。

[0161] 在一些实施方案中,所述占位肽是CLIP肽。在一些实施方案中,所述占位肽具有氨基酸序列PVSKMRMATPLLMQA。

[0162] 在一些实施方案中,所述多核酸构建体包含表达载体,该表达载体进一步包含以下一种或多种:启动子、分泌信号、二聚化因子、核糖体跳跃序列、一种或多种用于纯化和/或检测的标签。

[0163] 在一些实施方案中,所述占位肽序列由载体内的核酸序列编码。

[0164] 在一些实施方案中,将编码可切割结构域的序列置于编码占位肽的序列与编码HLA β1肽的序列之间。

[0165] 本文提供了一种测定MHC II类结合肽的免疫原性的方法,其包括:选择通过机器学习HLA-肽呈递预测模型被预测为与MHC II类结合肽结合的、由HLA II类等位基因编码的蛋白质,其中所述机器学习HLA-肽呈递预测模型被配置为生成针对给定肽序列的呈递预测,该呈递预测指示由所述HLA II类等位基因编码的一种或多种蛋白质能够呈递给定肽序列的可能性,并且其中所述蛋白质呈递MHC II类结合肽的概率大于阈值呈递预测概率值;使所述肽与所选蛋白质接触,使得所述肽竞争与所选蛋白质相关联的占位肽,并取代所述占位肽,从而形成包含所述HLA II类蛋白和所述MHC II类结合肽的复合物;使所述复合物与CD4+ T细胞接触,以及测定所述CD4+ T细胞的一种或多种活化参数,所述参数选自:细胞因子的诱导、趋化因子的诱导和细胞表面标志物的表达。

[0166] 在一些实施方案中,所述HLA II类等位基因是四聚体或多聚体。

[0167] 在一些实施方案中,所述细胞因子是IL-2。

[0168] 本文提供了一种在受试者中诱导CD4+ T细胞活化以用于癌症免疫治疗的方法,该方法包括:鉴定与癌症相关并包含癌症突变的肽序列,其中鉴定所述肽序列包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较;选择由HLA II类等位基因编码的蛋白质,该蛋白质由受试者的细胞正常表达,并且通过机器学习HLA-肽呈递预测模型被预测为与所述肽结合;其中所述预测模型在至少0.1%、0.1%-50%或至多50%的判定率(recall rate)下具有至少0.1的阳性预测值,并且其中所述蛋白质呈递所鉴定的肽序列的概率大于阈值呈递预测概率值;使所鉴定的肽与由HLA II类等位基因编码的所选蛋白质接触,以验证所鉴定的肽是否竞争与由HLA II类等位基因编码的所选蛋白质相关联的占位肽,从而以小于500nM的IC50值取代该占位肽;任选地,纯化所鉴定的肽;以及向所述受试者施用有效量的包含所鉴定的肽的序列的多肽或编码该多肽的多核苷酸。

[0169] 本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,其包括:通过计算机处理器获得所述多肽序列的多个肽序列;使用机器学习HLA-肽呈递预测模型,通过计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含与在细胞中表达的HLA蛋白相关的序列信息;基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的每一个对所述受试者不是免疫原性的;以及向所述受试者施用包含所述药物的组合物。

[0170] 本文提供了一种通过缀合四个单独的HLA蛋白α1和β1异二聚体来制备HLA II类四聚体或多聚体的方法,该方法包括:在真核细胞中表达包含编码HLA蛋白的α链和β链的核酸序列、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签的载体,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所述异二聚体与占位肽相关联,从细胞培养基中纯化所分泌的异二聚体,使用肽交换测定验证肽结合活性,加入链霉亲和素,从而将异二聚体缀合成四聚体,纯化所述四聚体,并具有大于1mg/L的产量。也可以同样产生多聚体,例如五聚体、六聚体或八聚体,它们在本文中同样被考虑。

[0171] 在一些实施方案中,所述载体包含CMV启动子。

[0172] 在一些实施方案中,所述载体包含编码通过可切割位点与β1链连接的占位肽的序列。

[0173] 在一些实施方案中,肽交换测定涉及从β链预先切割占位肽。

[0174] 在一些实施方案中,所述可切割位点是凝血酶切割位点。

[0175] 在一些实施方案中,肽交换测定是FRET测定。

[0176] 在一些实施方案中,所述纯化是通过以下任一种进行的:柱色谱法、离子交换色谱法、大小排阻色谱法、亲和色谱法或LC-MS。

[0177] 本文提供了一种包含HLA-DR或HLA-DP或HLA-DQ异二聚体的HLA II类四聚体或多聚体,每个异二聚体包含α和β链,其中所述异二聚体被纯化并以大于1mg/L的浓度存在。

[0178] 在一些实施方案中,所述HLA II类四聚体选自表8A-8C。

[0179] 在一些实施方案中,所述HLA II类四聚体包含选自下组的异二聚体对:HLA-DR、

HLA-DP和HLA-DQ蛋白。

[0180] 在一些实施方案中,所述HLA蛋白是选自下组的HLA II类蛋白:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01和HLA-DRB5*01:01)。

[0181] 在一些实施方案中,所述异二聚体对在真核细胞中表达。

[0182] 在一些实施方案中,所述异二聚体对由载体编码。

[0183] 本文提供了一种载体,其中所述载体包含编码本文所述HLA蛋白的α链和β链的核酸序列、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所分泌的异二聚体任选地与占位肽相关联。

[0184] 本文提供了一种细胞,其包含本文所述的载体。

[0185] 在一些实施方案中,所述HLA II类异二聚体从真核细胞分泌到细胞培养基中,其进一步通过以下任一种方法纯化:柱色谱法、离子交换色谱法、大小排阻色谱法、亲和色谱法或LC-MS。

[0186] 本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,其包括:通过计算机处理器获得所述多肽序列的多个肽序列;使用机器学习HLA-肽呈递预测模型,通过计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;以及基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的至少一个对所述受试者是免疫原性的。

[0187] 本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,该方法包括:使用计算机处理器将所述多肽序列的肽序列的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质呈递给定肽序列的表位序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;基于该组呈递预测,确定或预测

所述多肽序列的肽序列中的每一个对所述受试者不是免疫原性的;以及向所述受试者施用包含所述药物的组合物。

[0188] 本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,该方法包括:使用计算机处理器将所述多肽序列的肽序列的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质呈递给定肽序列的表位序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;基于该组呈递预测,确定或预测所述多肽序列的肽序列中的至少一个对所述受试者是免疫原性的。

[0189] 本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,其包括:通过计算机处理器获得所述多肽序列的多个肽序列;使用机器学习HLA-肽呈递预测模型,通过计算机处理器处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含与在细胞中表达的HLA蛋白相关的序列信息;基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的每一个对所述受试者不是免疫原性的;以及向所述受试者施用包含所述药物的组合物。

[0190] 在一些实施方案中,所述方法进一步包括决定不向所述受试者施用所述药物。

[0191] 在一些实施方案中,所述药物包含抗体或其结合片段。

[0192] 在一些实施方案中,所述多肽序列的肽序列具有8、9、10、11或12个氨基酸的长度,并且其中由受试者细胞的I类或II类MHC等位基因编码的蛋白质是由受试者细胞的I类MHC等位基因编码的蛋白质。

[0193] 在一些实施方案中,所述多肽序列的肽序列具有15、16、17、18、19、20、21、22、23、24或25个氨基酸的长度,并且其中由受试者细胞的I类或II类MHC等位基因编码的蛋白质是由受试者细胞的II类MHC等位基因编码的蛋白质。

[0194] 本文提供了一种治疗患有自身免疫性疾病或病况的受试者的方法,其包括:(a)鉴定或预测由受试者细胞的I类或II类MHC呈递的所表达蛋白质的表位,其中包含所鉴定或预测的表位和I类或II类MHC的复合物被所述受试者的CD8或CD4 T细胞所靶向;(b)鉴定与所述复合物结合的T细胞受体(TCR);(c)在来自所述受试者的调节性T细胞或同种异体调节性T细胞中表达所述TCR;以及(d)向所述受试者施用表达所述TCR的调节性T细胞。

[0195] 在一些实施方案中,所述自身免疫性疾病或病况是糖尿病。

[0196] 在一些实施方案中,所述细胞是胰岛细胞。

[0197] 本文提供了一种治疗患有自身免疫性疾病或病况的受试者的方法,其包括向所述受试者施用表达与复合物结合的T细胞受体(TCR)的调节性T细胞,该复合物包含:(i)被鉴定或预测为由受试者细胞的I类或II类MHC呈递的所表达蛋白质的表位,和(ii)I类或II类

MHC,其中所述复合物被所述受试者的CD8或CD4 T细胞所靶向。

[0198] 本文提供了一种用于鉴定用于受试者的个性化癌症治疗的肽序列的计算机系统,其包括:配置为存储受试者的多个肽序列的数据库;以及一个或多个可操作地耦合到所述数据库的计算机处理器,其中所述一个或多个计算机处理器单独地共同地被编程为:使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;以及至少基于所述多个呈递预测,选择所述多个肽序列的子集用于所述受试者的个性化癌症治疗。

[0199] 本文提供了一种用于鉴定用于受试者的免疫治疗的HLA II类特异性肽的计算机系统,其包括:配置为存储包含表位的候选肽和多个肽序列的数据库,每个肽序列均包含所述表位;以及一个或多个可操作地耦合到所述数据库的计算机处理器,其中所述一个或多个计算机处理器单独地共同地被编程为:使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个向免疫细胞的呈递预测,每个呈递预测指示由HLA II类等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;从受试者细胞的HLA II类等位基因编码的一种或多种蛋白质中选择蛋白质,所述蛋白质通过机器学习HLA-肽呈递预测模型被预测为与所述候选肽结合,其中所述蛋白质将所述候选肽呈递给免疫细胞的概率大于阈值呈递预测概率值;以及在候选肽与所选蛋白质接触,使得候选肽竞争与所选蛋白质相关联的占位肽时,根据候选肽是否取代占位肽,将所述候选肽鉴定为对所选蛋白质具有特异性的用于免疫治疗的肽。

[0200] 本文提供了一种用于针对在受试者中的免疫原性筛选包含多肽序列的药物的计算机系统,其包括:配置为存储所述多肽序列的多个肽序列的数据库;以及一个或多个可操作地耦合到所述数据库的计算机处理器,其中所述一个或多个计算机处理器单独地共同地被编程为:使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含与在细胞中表达的HLA蛋白相关的序列信息;以及基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的每一个对所述受试者不是免疫原性的,其中向所述受试者施用包含所述药物的组合物。

[0201] 本文提供了一种用于针对在受试者中的免疫原性筛选包含多肽序列的药物的计算机系统,其包括:配置为存储所述多肽序列的多个肽序列的数据库;以及一个或多个可操作地耦合到所述数据库的计算机处理器,其中所述一个或多个计算机处理器单独地共同地被编程为:使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位

序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;以及基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的至少一个对所述受试者是免疫原性的。

[0202] 本文提供了一种包含机器可执行代码的非暂时性计算机可读介质,该代码在被一个或多个计算机处理器执行时实施用于鉴定用于受试者的个性化癌症治疗的肽序列的方法,所述方法包括:获得所述受试者的多个肽序列;使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;以及至少基于所述多个呈递预测,选择所述多个肽序列的子集用于所述受试者的个性化癌症治疗。

[0203] 本文提供了一种包含机器可执行代码的非暂时性计算机可读介质,该代码在被一个或多个计算机处理器执行时实施用于鉴定用于受试者的免疫治疗的HLA II类特异性肽的方法,所述方法包括:获得包含表位的候选肽和多个肽序列,每个肽序列均包含所述表位;使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个向免疫细胞的呈递预测,每个呈递预测指示由HLA II类等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;从受试者细胞的HLA II类等位基因编码的一种或多种蛋白质中选择蛋白质,所述蛋白质通过机器学习HLA-肽呈递预测模型被预测为与所述候选肽结合,其中所述蛋白质将所述候选肽呈递给免疫细胞的概率大于阈值呈递预测概率值;以及在候选肽与所选蛋白质接触,使得候选肽与占位肽竞争时,根据候选肽是否取代占位肽,将所述候选肽鉴定为对所选蛋白质具有特异性的用于免疫治疗的肽。

[0204] 本文提供了一种包含机器可执行代码的非暂时性计算机可读介质,该代码在被一个或多个计算机处理器执行时实施针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,所述方法包括:获得所述多肽序列的多个肽序列;使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含与在细胞中表达的HLA蛋白相关的序列信息;以及基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的每一个对所述受试者不是免疫原性的,其中向所述受试者施用包含所述药物的组合物。

[0205] 本文提供了一种包含机器可执行代码的非暂时性计算机可读介质,该代码在被一个或多个计算机处理器执行时实施针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,所述方法包括:获得所述多肽序列的多个肽序列;使用机器学习HLA-肽呈递预测模型处理所述多个肽序列的氨基酸信息,以生成关于所述多个肽序列中的每一个的呈递预测,每个呈递预测指示由受试者细胞的I类或II类MHC等位基因编码的一种或多种蛋白质能

够呈递所述多个肽序列中的给定肽序列的表位序列的可能性,其中使用训练数据训练所述机器学习HLA-肽呈递预测模型,所述训练数据包含由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;以及基于所述多个呈递预测,确定或预测所述多肽序列的所述多个肽序列中的至少一个对所述受试者是免疫原性的。

[0206] 本文提供了一种方法,其包括:使用机器学习HLA肽呈递预测模型处理多个候选肽序列的氨基酸信息,以生成多个呈递预测,其中所述多个候选肽序列中的每个候选肽序列由受试者的基因组或外显子组编码,其中所述多个呈递预测包括针对所述多个候选肽序列中的每一个的HLA呈递预测,其中每个呈递预测指示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个候选肽序列中的给定候选肽序列的可能性,其中使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白呈递;以及至少基于所述多个呈递预测,鉴定所述多个肽序列中的肽序列,所述肽序列被由受试者细胞的II类HLA等位基因编码的一种或多种蛋白质中的至少一种蛋白质呈递的概率大于阈值呈递预测概率值;其中当处理多个测试肽序列的氨基酸信息以生成多个测试呈递预测,每个测试呈递预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质能够呈递所述多个测试肽序列中的给定测试肽序列的可能性时,所述机器学习HLA肽呈递预测模型具有至少0.07的阳性预测值(PPV),其中所述多个测试肽序列包含至少500个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少499个包含在由生物体基因组编码的蛋白质内的诱饵肽序列,其中所述生物体和受试者是同一物种,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少499个诱饵肽序列之比为1:499,并且根据机器学习HLA肽呈递预测模型,所述多个测试肽序列的0.2%被预测为将由在细胞中表达的HLA蛋白所呈递。

[0207] 本文提供了一种方法,其包括:使用机器学习HLA肽结合预测模型处理由受试者的基因组或外显子组编码的多个肽序列的氨基酸信息,以生成多个结合预测,其中所述多个结合预测包括对于所述多个候选肽序列中的每一个的HLA结合预测,每个结合预测指示由所述受试者的细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个候选肽序列中的给定候选肽序列结合的可能性,其中使用训练数据训练所述机器学习HLA肽结合预测模型,所述训练数据包含经鉴定与HLA II类蛋白或HLA II类蛋白类似物结合的肽序列的序列信息;以及至少基于所述多个结合预测,鉴定所述多个肽序列中的肽序列,该肽序列与所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种相结合的概率大于阈值结合预测概率值;其中当处理多个测试肽序列的氨基酸信息以生成多个测试结合预测,每个测试结合预测指示由所述受试者细胞的II类HLA等位基因编码的一种或多种蛋白质与所述多个测试肽序列中的给定测试肽序列结合的可能性时,所述机器学习HLA肽结合预测模型具有至少0.1的阳性预测值(PPV),其中所述多个测试肽序列包含至少50个测试肽序列,所述序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少19个包含在蛋白质内的诱饵肽序列,所述蛋白质包含通过质谱法被鉴定为将由细胞中表达的HLA蛋白呈递的肽序列,其中生物体和受试者是同一物种,其中所述多个测试肽序列中所述至少一个命中肽序列与所述至少19个诱饵肽序列之比为1:19,并且根据机器学习HLA肽呈递预测模型,所述多个测试肽序列的5%被预测为将与在细胞中

表达的HLA蛋白结合。

[0208] 在一些实施方案中,使用包含训练肽序列的序列信息的训练数据训练所述机器学习HLA肽呈递预测模型,所述训练肽通过质谱法被鉴定为将由在训练细胞中表达的HLA蛋白呈递。

[0209] 在一些实施方案中,通过机器学习HLA肽呈递预测模型被预测为将被呈递的所述多个测试肽序列的0.2%中的一个或多个被所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种呈递的概率大于阈值呈递预测概率值。

[0210] 在一些实施方案中,通过机器学习HLA肽呈递预测模型被预测为将被呈递的所述多个测试肽序列的0.2%中的每一个被所述受试者细胞的II类HLA等位基因所编码的一种或多种蛋白质中的至少一种呈递的概率大于阈值呈递预测概率值。

[0211] 在一些实施方案中,对于由表13的相应HLA等位基因编码的蛋白质,PPV大于表11的第2列的相应PPV。在一些实施方案中,对于由表11的相应HLA等位基因编码的蛋白质,PPV至少等于表11的第3列的相应PPV。

[0212] 在一些实施方案中,对于由HLA II类等位基因编码的蛋白质,PPV大于表12的第2列的相应PPV。

[0213] 在一些实施方案中,对于由表16的相应HLA等位基因编码的蛋白质,PPV至少等于表16的第2列的相应PPV。

[0214] 本文提供了一种准备个性化癌症治疗的方法,该方法包括:鉴定肽序列,其中所述肽序列与癌症相关,其中鉴定包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较;使用计算机处理器将所鉴定的肽序列的氨基酸位置信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所鉴定的肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的HLA II类等位基因编码的一种或多种蛋白质呈递所鉴定的肽序列中的给定肽序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;以及基于该组呈递预测选择所鉴定的肽序列的子集用于准备个性化癌症治疗;其中所述预测模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值。

[0215] 本文提供了一种方法,其包括训练机器学习HLA-肽呈递预测模型,其中训练包括使用计算机处理器将从来自表达HLAII类等位基因的细胞的一个或多个HLA-肽复合物中分离的HLA-肽的氨基酸位置信息序列输入到所述HLA-肽呈递预测模型中;所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量,所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含训练肽的氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数。

[0216] 在一些实施方案中,所述呈递模型在至少0.1%、0.1%-50%或至多50%的判定率

下具有至少0.25的阳性预测值。

[0217] 在一些实施方案中,所述呈递模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.4的阳性预测值。

[0218] 在一些实施方案中,所述呈递模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.6的阳性预测值。

[0219] 在一些实施方案中,所述质谱法是单等位基因质谱法。

[0220] 在一些实施方案中,所述肽通过自噬由在细胞中表达的HLA蛋白呈递。

[0221] 在一些实施方案中,所述肽通过吞噬作用由在细胞中表达的HLA蛋白呈递。

[0222] 在一些实施方案中,通过使用多个质量度量来提高训练数据的质量。

[0223] 在一些实施方案中,所述多个质量度量包括常见污染物肽去除、高评分峰强度、高评分和高质量准确度。

[0224] 在一些实施方案中,评分峰值强度至少为50%。

[0225] 在一些实施方案中,评分峰值强度至少为60%。

[0226] 在一些实施方案中,评分至少为7。

[0227] 在一些实施方案中,质量准确度至多为5ppm。

[0228] 在一些实施方案中,质量准确度至多为2ppm。

[0229] 在一些实施方案中,骨架切割评分至少为5。

[0230] 在一些实施方案中,骨架切割评分至少为8。

[0231] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一免疫沉淀的HLA蛋白呈递的肽。

[0232] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一外源HLA蛋白呈递的肽。

[0233] 在一些实施方案中,由细胞中表达的HLA蛋白呈递的肽是由细胞中表达的单一重组HLA蛋白呈递的肽。

[0234] 在一些实施方案中,所述多个预测变量包括肽-HLA亲和力预测变量。

[0235] 在一些实施方案中,所述多个预测变量包括源蛋白质表达水平预测变量。

[0236] 在一些实施方案中,所述多个预测变量包括肽可切割性预测变量。

[0237] 在一些实施方案中,所述训练肽序列信息包括来自由HLA蛋白呈递的肽的序列,所述肽包括通过搜索无酶特异性无修饰肽数据库而鉴定的肽。在一些实施方案中,由HLA蛋白呈递的肽包括通过搜索从头肽测序工具而鉴定的肽。

[0238] 在一些实施方案中,由HLA蛋白呈递的肽包括通过使用反转数据库搜索策略搜索肽数据库而鉴定的肽。

[0239] 在一些实施方案中,所述HLA蛋白包括HLA-DR和HLA-DP或HLA-DQ蛋白。在一些实施方案中,所述HLA蛋白包括选自HLA-DR和HLA-DP或HLA-DQ蛋白的HLA-DR蛋白。在一些实施方案中,所述HLA蛋白包括选自下组的HLA-DR蛋白:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-

DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01和HLA-DRB5*01:01)。

[0240] 在一些实施方案中,由HLA蛋白呈递的肽包括通过将HLA-肽的MS/MS谱与肽数据库中的一种或多种HLA-肽的MS/MS谱进行比较而鉴定的肽。

[0241] 在一些实施方案中,所述突变选自点突变、剪接位点突变、移码突变、通读突变和基因融合突变。

[0242] 在一些实施方案中,由HLA蛋白呈递的肽具有15-40个氨基酸的长度。

[0243] 在一些实施方案中,由HLA蛋白呈递的肽包括通过以下方式鉴定的肽:(a)从表达单个HLAII类等位基因的细胞系中分离一种或多种HLA复合物;(b)从所述一种或多种分离的HLA复合物中分离一种或多种HLA-肽;(c)获得所述一种或多种分离的HLA-肽的MS/MS谱;以及(d)从肽数据库中获得与所述一种或多种分离的HLA-肽的MS/MS谱相对应的肽序列;其中从步骤(d)获得的一种或多种序列鉴定所述一种或多种分离的HLA-肽的序列。

[0244] 在一些实施方案中,所述个性化癌症治疗进一步包括佐剂。

[0245] 在一些实施方案中,所述个性化癌症治疗进一步包括免疫检查点抑制剂。

[0246] 在一些实施方案中,所述训练数据包括结构化数据、时间序列数据、非结构化数据、关系数据或其任何组合。

[0247] 在一些实施方案中,所述非结构化数据包括图像数据。

[0248] 在一些实施方案中,所述关系数据包括来自客户系统、企业系统、操作系统、网站、网络可访问应用程序界面(API)或其任何组合的数据。

[0249] 在一些实施方案中,所述训练数据被上传到基于云的数据库。

[0250] 在一些实施方案中,所述训练使用卷积神经网络来进行。

[0251] 在一些实施方案中,所述卷积神经网络包含至少两个卷积层。

[0252] 在一些实施方案中,所述卷积神经网络(CNN)包括至少一个批归一化步骤。

[0253] 在一些实施方案中,所述卷积神经网络包括至少一个空间丢弃(spatial dropout)步骤。

[0254] 在一些实施方案中,所述卷积神经网络包括至少一个全局最大池化步骤。

[0255] 在一些实施方案中,所述卷积神经网络包含至少一个密集层。

[0256] 在一些实施方案中,鉴定肽序列包括鉴定在受试者的癌细胞中表达的具有突变的肽序列。

[0257] 在一些实施方案中,鉴定肽序列包括鉴定在受试者的正常细胞中不表达的肽序列。

[0258] 在一些实施方案中,鉴定肽序列包括鉴定过表达的肽序列。

[0259] 在一些实施方案中,鉴定肽序列包括鉴定病毒肽序列。在一方面,本文提供了一种鉴定对受试者具有特异性的用于免疫治疗的HLA II类特异性肽的方法,该方法包括:鉴定

包含表位的候选肽;使用计算机处理器将多个肽序列(各自包含表位)的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列向免疫细胞的一组HLA呈递预测,每个呈递预测代表由受试者细胞的HLA II类等位基因编码的一种或多种蛋白质呈递包含所述表位的给定肽序列的概率;其中所述预测模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值,从受试者细胞的HLA II类等位基因编码的一种或多种蛋白质中选择蛋白质,所述蛋白质被所述预测模型预测为与所述候选肽结合,其中所述蛋白质将所述候选肽呈递给免疫细胞的概率大于阈值呈递预测概率值;使所述候选肽与HLA II类等位基因编码的蛋白质接触,使得所述候选肽竞争与HLA II类等位基因编码的蛋白质相关联的占位肽;以及根据所述候选肽是否取代所述占位肽,将所述候选肽鉴定为对HLA II类等位基因编码的蛋白质具有特异性的用于免疫治疗的肽。

[0260] 在一些实施方案中,所述免疫治疗是癌症免疫治疗。

[0261] 在一些实施方案中,鉴定包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较。在一些实施方案中,所述表位是癌症特异性表位。

[0262] 在一些实施方案中,所述由HLA II类等位基因编码的至少一种蛋白质至少包含以二聚体形式存在的HLA蛋白或其片段的α1亚单位和β1亚单位。在一些实施方案中,所述占位肽是CLIP肽。在一些实施方案中,所述占位肽是CMV肽。在一些实施方案中,所述方法进一步包括测量所述占位肽被目标肽取代的IC50。在一些实施方案中,所述占位肽被目标肽取代的IC50小于500nM。在一些实施方案中,来自由受试者细胞的HLA II类等位基因编码的一种或多种蛋白质的至少一种蛋白质是HLA II类四聚体或多聚体。在一些实施方案中,所述目标肽进一步通过质谱法来鉴定。在一些实施方案中,所述由受试者细胞的HLA II类等位基因编码的至少一种蛋白质是重组蛋白。在一些实施方案中,所述由受试者细胞的HLAII类等位基因编码的至少一种蛋白质在真核细胞中表达。

[0263] 在一方面,本文提供了用于验证候选肽结合HLA II类蛋白的特异性的测定方法,该方法包括:在真核细胞中表达包含编码HLA II类蛋白的核酸序列的多核酸构建体,该蛋白包含α链和β链或其部分,能够结合包含MHC-II结合表位的肽,并且其中所表达的HLA II类蛋白或其部分保持与占位肽相关联;分离在所述真核细胞中表达的HLA II类蛋白或其部分;通过以下方式进行肽交换测定:(a)加入渐增量的候选肽,以确定该候选肽是否取代与所述HLA II类蛋白或其部分相关联的占位肽;以及(b)计算取代反应的IC50,以确定所述候选肽对所述HLA II类蛋白或其部分相对于所述占位肽的亲和力,从而验证所述候选肽结合HLA II类蛋白的特异性。

[0264] 在一些实施方案中,所述肽的身份是已知的。在一些实施方案中,所述肽的身份是未知的。在一些实施方案中,所述肽的身份通过质谱法确定。

[0265] 在一些实施方案中,所述肽交换测定包括检测肽荧光探针或标签。在一些实施方案中,所述占位肽是CLIP肽。

[0266] 在一些实施方案中,所述多核酸构建体包含表达载体,该表达载体进一步包含以下一种或多种:启动子、连接体、一个或多个蛋白酶切割位点、分泌信号、二聚化因子、核糖体跳跃序列、一种或多种用于纯化和/或检测的标签。

[0267] 在一方面,本文提供了一种用于测定MHC II类结合肽的免疫原性的方法,该方法

包括:选择通过机器学习HLA-肽呈递预测模型预测与肽结合的由HLA II类等位基因编码的蛋白质;其中所述预测模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值,并且其中所述蛋白质呈递所鉴定的肽序列的概率大于阈值呈递预测概率值;使所述肽与由HLA II类等位基因编码的所选蛋白质接触,使得所述肽竞争与HLA II类等位基因编码的所选蛋白质相关联的占位肽,并取代所述占位肽,从而形成包含所述HLA II类蛋白和所鉴定的肽的复合物;使所述HLA II类蛋白和所鉴定的肽的复合物与CD4+ T细胞接触,测定所述CD4+ T细胞的一个或多个活化参数,所述参数选自细胞因子的诱导、趋化因子的诱导和细胞表面标志物的表达。

[0268] 在一些实施方案中,所述HLA II类等位基因是四聚体或多聚体。在一些实施方案中,所述细胞因子是IL-2。在一些实施方案中,所述细胞因子是IFN-γ。

[0269] 在一方面,本文提供了一种在受试者中诱导CD4+ T细胞活化以用于癌症免疫治疗的方法,该方法包括:鉴定与癌症相关并包含癌症突变的肽序列,其中鉴定包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较;选择由HLA II类等位基因编码的蛋白质,该蛋白质由受试者的细胞正常表达,并且通过机器学习HLA-肽呈递预测模型被预测为与所述肽结合;其中所述预测模型在至少0.1%、0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值,并且其中所述蛋白质呈递所鉴定的肽序列的概率大于阈值呈递预测概率值;使所鉴定的肽与由HLA II类等位基因编码的所选蛋白质接触,以验证所鉴定的肽是否竞争与由HLA II类等位基因编码的所选蛋白质相关联的占位肽,从而以小于500nM的IC50值取代该占位肽;纯化所鉴定的肽;以及向所述受试者施用有效量的所鉴定的肽。

[0270] 在一方面,本文提供了一种制备HLA II类四聚体或多聚体的方法,该方法包括:在真核细胞中表达包含编码HLA蛋白的α链和β链的核酸序列、连接体、一个或多个蛋白酶切割位点、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签的载体,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所述异二聚体与占位肽相关联,从细胞培养基中纯化所分泌的异二聚体,使用肽交换测定验证肽结合活性,加入链霉亲和素,从而将异二聚体缀合成四聚体,纯化所述四聚体,并具有大于1mg/L的产量。

[0271] 在一些实施方案中,所述载体包含CMV启动子。在一些实施方案中,所述载体包含编码通过可切割位点与β1链连接的占位肽的序列。在一些实施方案中,肽交换测定涉及从β链预先切割占位肽。在一些实施方案中,所述可切割位点是凝血酶切割位点。在一些实施方案中,肽交换测定是FRET测定。在一些实施方案中,所述纯化是通过以下任一种进行的:柱色谱法、批量色谱法、离子交换色谱法、大小排阻色谱法、亲和色谱法或LC-MS。

[0272] 在一方面,本文提供了一种包含HLA II类四聚体的组合物,该四聚体包含HLA-DR或HLA-DP或HLA-DQ异二聚体,每个异二聚体包含α和β链,其被纯化并以大于0.25mg/L的浓度存在。在一些实施方案中,所述HLA II类四聚体包含选自下组的异二聚体对:蛋白质可以选自HLA-DR和HLA-DP或HLA-DQ蛋白。在一些实施方案中,所述HLA蛋白选自:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:

01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01、HLA-DRB5*01:01)。

[0273] 在一些实施方案中,所述异二聚体对在真核细胞中表达。在一些实施方案中,所述异二聚体对由载体编码。在一些实施方案中,所述载体包含:编码HLA蛋白的α链和β链的核酸序列、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所分泌的异二聚体与占位肽相关联。在一些实施方案中,所述载体包含:编码HLA蛋白的α链和β链的核酸序列、分泌信号、生物素化基序和至少一个用于鉴定或纯化的标签,使得每个HLA蛋白α1和β1异二聚体以二聚化状态分泌,其中所分泌的异二聚体与占位肽相关联。

[0274] 在一些实施方案中,HLA II类异二聚体从真核细胞分泌到细胞培养基中,并且通过以下任一种方法纯化:柱色谱法或批量色谱法、离子交换色谱法、大小排阻色谱法、亲和色谱法或LC-MS。

[0275] 在一方面,本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,该方法包括:使用计算机处理器将所述多肽序列的肽序列的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的HLA I类或II类等位基因编码的一种或多种蛋白质呈递给定肽序列的表位序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;(b)基于该组呈递预测,确定或预测所述多肽序列的肽序列中的每一个对所述受试者不是免疫原性的;以及(c)向所述受试者施用包含所述药物的组合物。

[0276] 在一方面,本文提供了一种针对在受试者中的免疫原性筛选包含多肽序列的药物的方法,该方法包括:(a)使用计算机处理器将所述多肽序列的肽序列的氨基酸信息输入到机器学习HLA-肽呈递预测模型中,以生成关于所述肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的HLA I类或II类等位基因编码的一种或多种蛋白质呈递给定肽序列的表位序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量;其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;(b)基于该组呈递预测,确定或预测所述多肽序列的肽序列中的至少一个对所述受试者是免疫原性的。

[0277] 在一个实施方案中,所述方法进一步包括决定不向所述受试者施用所述药物。

[0278] 在一个实施方案中,所述药物包含抗体或其结合片段。

[0279] 在一个实施方案中,所述多肽序列的肽序列包含所述多肽序列的每个连续肽序列,其具有8、9、10、11或12个氨基酸的长度,并且其中由受试者细胞的HLA I类或II类等位基因编码的蛋白质是由受试者细胞的HLA I类等位基因编码的蛋白质。

[0280] 在一个实施方案中,所述多肽序列的肽序列包含所述多肽序列的每个连续肽序列,其具有15、16、17、18、19、20、21、22、23、24或25个氨基酸的长度,并且其中由受试者细胞的HLA I类或II类等位基因编码的蛋白质是由受试者细胞的II类MHC等位基因编码的蛋白质。

[0281] 在一方面,本文提供了一种治疗患有自身免疫性疾病或病况的受试者的方法,其包括:(a)鉴定或预测由受试者细胞的HLA I类或II类呈递的所表达蛋白质的表位,其中包含所鉴定或预测的表位和HLA I类或II类的复合物被所述受试者的CD8或CD4 T细胞所靶向;(b)鉴定与所述复合物结合的T细胞受体(TCR);(c)在来自所述受试者的调节性T细胞或同种异体调节性T细胞中表达所述TCR;以及(d)向所述受试者施用表达所述TCR的调节性T细胞。

[0282] 在一个实施方案中,所述自身免疫性疾病或病况是糖尿病。

[0283] 在一个实施方案中,所述细胞是胰岛细胞。

[0284] 在一方面,本文提供了一种治疗患有自身免疫性疾病或病况的受试者的方法,其包括向所述受试者施用表达与复合物结合的T细胞受体(TCR)的调节性T细胞,该复合物包含:(i)被鉴定或预测为由受试者细胞的HLA I类或II类呈递的所表达蛋白质的表位,和(ii)HLA I类或II类,其中所述复合物被所述受试者的CD8或CD4 T细胞所靶向。

[0285] 基于仅示出并描述了本公开的说明性实施方案的以下详细描述,本公开的其他方面和优点对本领域技术人员而言将变得显而易见。应当认识到,本公开能够具有其他不同的实施方案,并且其若干细节能够在各个明显的方面进行修改,所有这些都不脱离本公开内容。因此,附图和说明书在本质上将被视为说明性的,而非限制性的。

[0286] MAPTAC™可用于高通量肽结合测定,其中在不同时间点和不同条件(例如在37℃下加热)下使用MAPTAC™构建体分离后测量与HLA II类结合的肽,以使用LC-MS/MS获得具有不同稳定性的肽群体的序列。

[0287] 在一方面,本文提供了一种治疗受试者的癌症的方法,该方法包括:鉴定肽序列,其中所述肽序列与癌症相关,其中鉴定包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较;使用计算机处理器将所鉴定的肽序列的氨基酸信息输入机器学习HLA-肽呈递预测模型中,以生成关于所鉴定的肽序列的一组呈递预测,每个呈递预测代表由受试者细胞的HLA II类等位基因编码的一种或多种蛋白质呈递所鉴定的肽序列的给定序列的概率;其中所述机器学习HLA-肽呈递预测模型包括:至少基于训练数据确定的多个预测变量,其中所述训练数据包含:由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数;以及基于该组呈递预测选择所鉴定的肽序列的子集用于准备个性化癌症治疗;以及向所述受试者施用包含一种或多种所述肽的组合物,其中所述预测模型在至少0.1%、

0.1%-50%或至多50%的判定率下具有至少0.1的阳性预测值。

[0288] 在一些实施方案中,所述机器学习HLA-肽呈递预测模型包含由细胞中表达的HLA蛋白呈递并在进行反相离线分级分离后通过质谱法鉴定的肽的序列的序列信息。

[0289] 在一些实施方案中,与NetMHCIIpan相比,所述预测模型表现出1.1x至100x倍的改善。在一些实施方案中,与NetMHCIIpan相比,所述预测模型表现出1.1、2、3、4、5、6、7、7.4、8、9、10、11、12、13、14、15、16、17、18、18、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、50、55、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、8、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98、99、100倍或更多的改善。

[0290] 援引并入

[0291] 本说明书中所提及的所有出版物、专利和专利申请均通过引用并入本文,其程度犹如具体地且单独地指出每个单独的出版物、专利或专利申请均通过引用而并入。如果通过引用而并入的出版物和专利或专利申请与本说明书中包含的公开内容存在矛盾,则本说明书旨在取代和/或优先于任何这样的矛盾材料。

附图说明

[0292] 本发明的新颖特征在所附权利要求书中具体阐述。通过参考以下对利用本发明原理的说明性实施方案加以阐述的详细描述以及附图(在本文中也称为“图”),将会对本发明的特征和优点获得更好的理解,在这些附图中:

[0293] 图1A是表示停靠在MHC I类蛋白上的肽的图示。

[0294] 图1B描绘了表示停靠在MHC II类蛋白上的肽的示例性图示。

[0295] 图2描绘了用于生成单等位基因HLA II类结合肽数据的示例性实验方法。将HLA II类肽引入任何细胞,包括不表达HLA II类的细胞中,以便在细胞中表达特定的HLA II类等位基因。收获经遗传工程化的表达HLA的细胞群体,裂解,标记(例如,生物素化)它们的HLA-肽复合物,并免疫纯化(例如,使用生物素-链霉亲和素相互作用)。可以从其标记的(例如,生物素化的)复合物中洗脱对单HLA具有特异性的HLA关联肽,并进行评价(例如,使用高分辨率LC-MS/MS进行测序)。

[0296] 图3描绘了HLA II类-DRB1*11:01关联肽在以下中的示例性序列标识表示:Neon BAP,Expi293细胞系;Neon BAP,A375细胞系;IEDB,亲和力<50nM;和全HLA II类抗体,纯合LCL。图3显示,MS衍生的基序的示例与已知模式匹配,并在转染的细胞系之间显示出一致性。

[0297] 图4是HLA II类结合预测器性能的示例性描绘。图4是一个条形图,其显示了应用于验证数据集的结合预测器(neonmhc2)和NetMHC IIpan的性能,该验证数据集由观察到的质谱肽和诱饵肽组成,其通过随机改组命中肽以1:19(命中:诱饵)的比率生成。对于NEON结合预测器neonmhc2,为示出的每个MHC II等位基因构建单独的模型。条形的高度显示阳性预测值(PPV),其被定义为验证集中预测的结合物的分数,这些结合物实际上是命中肽。在预测该等位基因时,等位基因按模型的性能排序。

[0298] 图5描绘了评分峰值强度(SPI)阈值对结合预测器验证的示例性影响。图5显示了当在具有不同评分峰值强度(SPI)截止值的成组肽上训练/验证时,HLA II类结合预测器的

性能。对于被训练的每个等位基因特异性模型,显示了模型在3种设置下的性能:使用观察到的大于或等于70SPI的MS命中肽对数据集进行训练和评估,对大于或等于50SPI的肽进行训练并对大于或等于70SPI的肽进行验证,以及对大于或等于50SPI的肽进行训练和验证。

[0299] 图6描绘了示例性条形图,其显示了用大于或等于70的评分峰值强度(SPI)截止值通过LC-MS/MS进行等位基因概况分析,来自观察到的肽的数目的代表性数据。每个条形代表观察到的等位基因的肽的总数。收集了35个HLA-DR等位基因的数据。收集到的35个HLA-DR等位基因的数据具有>95%的HLA-DR群体覆盖率(美国等位基因频率)。

[0300] 图7A显示了对于所示的HLA II类等位基因,当应用于数据的测试分区时,模型的PPV。使用的诱饵肽是阳性(命中)肽序列的乱序序列,命中与诱饵之比为1:19。通过鉴定测试分区中评分排名前5%的肽并确定它们之中与相应HLA II类等位基因编码的蛋白质结合呈阳性的分数来确定PPV。

[0301] 图7B描绘了作为训练集大小(通过人工地对训练集进行降采样获得的曲线)的函数的示例性预测性能。图7B显示,通常,对于收集的35个HLA-DR等位基因,当训练集大小增加时,PPV的值增加。

[0302] 图8描绘了示例性图示,其证明了处理相关变量可以进一步改进预测。区分MS观察到的肽,可以区分选自蛋白质编码外显子组的随机序列。在训练数据分区上,逻辑回归可能适合于使用结合强度(NetMHCIIpan或Neon预测器)和处理特征(RNA-Seq表达和衍生的基因水平偏项)来预测HLA II类呈递。在单独的评估分区中,与MS观察到的MHC II肽重叠的外显子位置(“命中”)可以与MS中未观察到的随机外显子位置一起评分(1:499比率)。前0.2%(1/500)可被称为阳性,阳性预测值可评估该阈值。

[0303] 图9描绘了示例性神经网络架构。输入肽被表示为20聚体,较短的肽用“缺失”字符填充。每个肽都有31维的嵌入,因此向神经网络中的输入是20x31矩阵。在被神经网络处理之前,根据训练集中的特征值平均值和标准偏差对20x31矩阵进行特征归一化。第一个卷积层具有9个氨基酸的内核和50个过滤器(也称为通道),具有修正线性单元(ReLU)激活函数。接着是批量归一化,然后是空间丢弃,丢弃率为20%。接着是另一个卷积层,其具有3个氨基酸的内核和20个过滤器,具有ReLU激活函数,然后再次是批量归一化和空间丢弃,丢弃率为20%。然后应用全局最大池化,在20个过滤器中的每个过滤器中获取最大激活的神经元;然后使用S形激活函数使这20个值进入具有单个神经元的完全连接的(密集)层。该层的输出被视为结合/非结合预测。将L2正则化应用于第一卷积层、第二卷积层和密集层的权重,权重分别为0.05、0.1和0.01。使用的其他模型改变了卷积层的数目和每个层的内核大小。

[0304] 图10描绘了被编程或以其他方式配置为实现本文提供的方法的示例性计算机控制系统。

[0305] 图11A描绘了MAPTAC™实验工作流程的示例性概览。

[0306] 图11B描绘了跨重复合并的示例性每个等位基因的肽计数。

[0307] 图11C描绘了通过MAPTAC™分析的HLA I类和HLA II类等位基因的示例性肽长度分布。

[0308] 图11D描绘了针对MAPTAC™和IEDB(等位基因DRB1*01:01、DRB1*03:01、DRB1*09:01和DRB1*11:01)、人类蛋白质组和来自先前出版物的多等位基因MS数据观察到的示例性每个残基的半胱氨酸频率。

[0309] 图12A描绘了存在于>1％的个体中的HLA-DR、-DP和-DQ等位基因的高加索人频率，以及来自被测量为强结合物(<50nM)的所示来源的肽的计数。

[0310] 图12B描绘了具有相关HLA II类亲和力测量的IEDB肽的示例性长度分布。

[0311] 图12C描绘了分别用两个HLA I类和两个HLA II类等位基因——HLA-A*02:01、HLA-B*45:01、HLA-DRB1*01:01和HLA-DRB1*11:01——转染的(1)Expi293、(2)HeLa和(3)A375细胞系的示例性Western印迹。用抗生物素连接酶表位标签对膜进行印迹，以显示生物素接受肽(BAP)和抗β-微管蛋白作为加载对照。泳道对应于在MAPTAC™方案期间收集的以下级分：泳道1，输入；泳道2，生物素化的输入；泳道3，下拉后的输入。

[0312] 图12D描绘了针对MAPTAC™和IEDB(等位基因DRB1*01:01、DRB1*03:01，DRB1*09:01和DRB1*11:01)、人类蛋白质组和来自先前出版物的多等位基因MS数据观察到的示例性每个残基的氨基酸频率。

[0313] 图12E描绘了存在于>1％的个体中的HLA-DR、-DP和-DQ等位基因的高加索人频率，以及来自被测量为强结合物(<50nM)的所示来源的肽的计数。该图包括相对于图12A的额外数据。该额外数据获自：tools.iedb.org/main/datasets/。

[0314] 图12F描绘了针对MAPTAC™(还原的和烷基化的)、MAPTAC™(无处理)和IEDB(等位基因DRB1*01:01、DRB1*03:01，DRB1*09:01和DRB1*11:01)、人类蛋白质组和来自先前出版物的多等位基因MS数据观察到的示例性每个残基的氨基酸频率。

[0315] 图13描绘了每个MAPTAC™和IEDB的MHC II等位基因的核心结合序列标识的示例性表示。序列标识是图形表示，其中每个氨基酸的高度与其在与等位基因编码的MHC蛋白结合的肽中的出现频率成比例。具有最低熵的位置用颜色表示，其中颜色对应于氨基酸性质。肽来自所示的数据集，并根据基于CNN的预测器(方法)进行对齐。标识代表所有肽，包括那些与整体基序不紧密匹配的肽(例如，没有肽被隔离在“垃圾”簇中)。

[0316] 图14A描绘了在两种不同的细胞系(A375和expi293)中使用不同HLA-配体概况分析技术分析的HLA-A*02:01结合肽(配体)的示例性序列标识，所述概况分析技术包括结合测定、稳定性测定、可溶性HLA(sHLA)质谱法、单等位基因质谱法和MAPTAC™。

[0317] 图14B描绘了展示0、1、2、3和4个启发式定义的锚的MAPTAC™肽的示例性级分。

[0318] 图14C描绘了针对MAPTAC™观察到的肽(每个等位基因20个肽，每个具有SPI>70和大小>＝2的嵌套组)和从蛋白质组采样的长度匹配的诱饵的NetMHCIIpan预测的结合亲和力的示例性分布。

[0319] 图15A描绘了卷积神经网络(CNN)的示例性架构，该CNN被训练以区分单等位基因MHC肽与乱序长度匹配的诱饵。示意图示出了使用氨基酸特征嵌入，具有不同过滤器大小的2个卷积层，以及使用全局最大池化作为最终逻辑输出节点的输入。

[0320] 图15B是示例性结果，其显示了测量的IEDB亲和力与来自neonmhc2或NetMHCIIpan的结合预测的相关性的Kendall Tau统计学。评估的肽仅包括在NetMHCIIpan发布后一年发布到IEDB的肽。

[0321] 图16是作为训练数据集大小的函数的neonmhc2的性能的示例性描述。

[0322] 图17A描绘了掺入泛DR和泛II类MHC MS数据集中的MAPTAC™肽(每个等位基因20个)的示例性簇分配。使用GibbsCluster对数据集进行去卷积。每个方框代表一个MAPTAC™肽。方框的颜色指示它被分配到哪个簇，灰色条形指示该肽实际上来自哪个等位基因。

Gibbs簇方案(右侧)中的簇总数使用交互信息(MI)度量进行选择。MI评分还决定了样品的排序方式;具有高MI方案的样品出现在顶部。

[0323] 图17B描绘了通过GibbsCluster去卷积的多等位基因MS数据的示例性核心结合序列标识。每组肽对应于与MAPTAC™掺加物最佳匹配的簇。

[0324] 图17C描绘了使用MAPTAC™数据或去卷积的多等位基因数据预测保留MAPTAC™肽的模型的代表性性能。对于每个等位基因,对两个数据源中较大的一个(通常是MAPTAC™)进行降采样,以使预测器基于相同数目的训练示例。NetMHCIIpan性能显示为额外的比较。

[0325] 图17D描绘了从来自所示来源的多等位基因MS数据导出的示例性核心结合序列标志。

[0326] 图18A描绘了对于MS观察到的肽和随机蛋白质组诱饵(从Schuster等人,2017重新绘制的数据),肽与源基因表达的分数(每百万个中的转录物(TPM))的示例性图示。

[0327] 图18B描绘了通过结直肠癌、黑素瘤和卵巢癌数据集的联合分析所确定的,每个基因的II类肽的示例性观察数目与预期数目(Löffler等人,2018和Schuster等人,2017)。预期计数是通过将基因长度乘以表达水平得出的。将相关样品的预期计数和观察计数相加。已知存在于血浆中的基因根据其浓度(插图)进行标记。

[0328] 图18C描绘了对于与自噬相关的基因,富集评分的示例性分布(观察到的与预期的观察的比率,如图18B)。

[0329] 图18D描绘了根据每个源基因的定位,富集评分的示例性分布。使用Uniprot(uniprot_sprot.dat)确定源基因定位。

[0330] 图18E描绘了示例性数据,其代表基于它们的细胞定位性质分离的、具有MHC-II结合亲和力的肽的总数的分数的预期与观察频率的比较。

[0331] 图18F描绘了观察中的肽对于两个不同基因表达谱的相对一致性的示例性代表性数据。对于每个样品,基因水平的肽计数被建模为成块(bulk)肿瘤基因表达和专职APC(巨噬细胞)基因表达谱的线性组合。系数的比率决定了每个表达谱与肽组库(repertoire)的相对一致性。误差条对应于通过自举重采样计算的95%置信区间。

[0332] 图19A描绘了五个示例研究中HLA-DRB1表达水平的示例性代表性数据。每个斑点代表在细胞中平均的、个体患者中个体细胞类型的表达。

[0333] 图19B描绘了从TCGA患者的RNA-Seq输入的肿瘤和基质衍生的HLA-DRB1表达的示例性代表性数据。水平条对应个体患者并按肿瘤类型分组。如果如通过基于DNA的突变判定所确定的,患者在HLA II类途径基因(CIITA、CD74或CTSSS)中具有突变,则包括这些患者。对于每名患者,归因于肿瘤的HLA-DRB1表达的分数被估计为min(1,2f),其中f是显示突变的CIITA、CD74或CTSS中RNA-Seq读取的分数。

[0334] 图19C描绘了额外的单细胞RNA-Seq研究的示例性代表性数据,所述研究包括检查点阻断免疫治疗之前和之后的活检。

[0335] 图20描绘了评估对天然供体组织的预测总体性能的示例性代表性实验数据。

[0336] 图21A描绘了示例性的代表性数据,其显示整合的呈递模型预测细胞HLA II类配体组(ligandomes)。它代表了泛DR数据集在命中:诱饵比为1:499时的PPV(也在图30B和图32E中进行了分析)。预测器使用结合预测(NetMHCIIpan或neonmhc2)并任选地使用基因表达、基因偏好(根据图32A),并与先前观察到的HLA-DQ肽重叠。对于每个候选肽,结合评分被

计算为样品基因型中HLA-DR等位基因中的最大值。

[0337] 图21B描绘了示例性代表性数据,其显示了使用SILAC鉴定的肿瘤衍生肽的预测性能,所述肽由树突细胞呈递(从细胞裂解物分析),其中使用与图21A中相同的命中:诱饵比和性能度量,使用和不使用处理特征。

[0338] 图21C描绘了与轻标记的肽(灰色斑点,根据DC表达绘制)相比,在UV处理实验中观察到的重标记的肽(红色斑点,根据K562表达绘制)的示例性表达和基因偏好评分。

[0339] 图21D描绘了代表根据裂解物和UV处理实验,重标记的肽源基因的重叠的示例图示。基因名称按功能类别着色。

[0340] 图22A描绘了代表本文公开的测定方案的示例性流程图,以验证HLA II类驱动的CD4+ T细胞和T细胞应答。

[0341] 图22B描绘了用于肽交换测定的示例性HLA蛋白二聚体构建体设计(上图)和示例性测定工作流程的图示(下图)。

[0342] 图23描绘了用于筛选新结合肽的用于MHC-II表达的示例性载体设计的示例性图示,以及表达的蛋白质产物的表示。

[0343] 图24描绘了转染、纯化和从β链上切割占位肽的示例性流程图。

[0344] 图25A描绘了示例性图示,其显示编码与表达的MHC-II肽的分泌增加相关的CLIP肽的载体。

[0345] 图25B描绘了分别具有编码CLIP0和CLIP1的核酸的较短和较长形式的示例性图示。

[0346] 图25C描绘了具有或不具有较长夹子(clip)的α和β链的考马斯凝胶分析的示例性代表性结果。

[0347] 图26A描绘了TR-FRET测定的示例性图示。

[0348] 图26B描绘了使用特定肽,使用荧光共振能量转移(FRET)测定,HLA II类肽结合测定的示例性代表性偏振数据。

[0349] 图26C描绘了使用特定肽,使用荧光共振能量转移(FRET)测定,HLA II类肽结合测定的示例性代表性偏振数据。

[0350] 图26D描绘了从荧光增加计算的MHC-构建体结合肽的示例性取代百分比。

[0351] 图26E描绘了从荧光增加计算的MHC-构建体结合肽的示例性取代百分比。

[0352] 图26F描绘了使用采用差示扫描荧光法(DSF)的测定获得的示例性肽交换。图示显示了检测采用加热使肽从MHC II类上解离的示例性机制,其也会解离MHC II类异二聚体,导致荧光团的结合和高荧光。还描绘了被表位肽移位的占位肽的示例性示意图。还描绘了随温度绘制的示例性解链曲线。

[0353] 图26G描绘了示例性的可溶性HLA-DM构建体及其用于进行MHC II类肽交换的用途。所描绘的构建体包含CMV启动子、HLA-DMβ链的编码序列,和分泌序列(前导区)下游的HLA-DMα链的编码序列,以及β链编码序列3'端的BAP序列;位于α链编码序列3'端的His标签。这两条链被中间的核糖体跳跃序列隔开。在Expi-CHO细胞中表达该构建体,并纯化分泌到培养基中的蛋白质。

[0354] 图26H显示了使用HLA-sDM进行肽交换的示例性大小排阻色谱数据。

[0355] 图27A描绘了示例性DRB四聚体组库构建的示例性图示。

[0356] 图27B描绘了示例性II类四聚体组库构建的示例性图示。

[0357] 图27C描绘了用于肽交换的DRB1等位基因的DRB四聚体组库覆盖的总结的示例性图示。

[0358] 图27D描绘了人MHC II类等位基因产生的示例性覆盖。

[0359] 图27E显示了来自用流感表位(记忆应答)或HIV表位(原始应答)诱导的样品的四聚体染色的示例性结果。

[0360] 图28A描绘了通过荧光偏振测定评估肽的HLA II类限制性的方法的示例性图示,该测定使筛选方法能够快速鉴定表位肽的等位基因限制性。图28A中描绘的测定原理允许亲和力测量和肽交换的明确测量。

[0361] 图28B描绘了在采用DRB1*01:01的荧光偏振测定中探索的多个测定条件的示例性总结(上图)。还描绘了可溶性MHC II类等位基因和在去污剂胶束中具有跨膜域的全长MHC II类等位基因(下图),两者均用具有可切割连接体的占位肽构建以用于测定。

[0362] 图28C描绘了用于研究前面在图28B下图中显示的全长和可溶性等位基因的测定的示例性图示。简而言之,全长和可溶性等位基因都在细胞中表达。膜结合的全长等位基因形式通过对膜进行透化来收集,而分泌形式从细胞上清液中收集。收获的II类HLA等位基因蛋白通过穿过镍(Ni^{2+})柱进行纯化。

[0363] 图28D描绘了显示纯化方法不影响肽效力的示例性数据。左侧显示的是来自使用L243纯化的全长HLA-DR1和Ni^{2+}纯化的全长HLA-DR1的实验的平均IC50值。

[0364] 图28E描绘了显示可溶形式(sDR1)或全长形式(fDR1)的选择不影响肽效力的示例性数据。左侧显示的是来自使用sDR1形式或fDR1的实验的平均IC50值。FP,荧光偏振。

[0365] 图28F描绘了neonmhc2和NetMHCIIpan预测的肽在结合测定和不一致肽的鉴定中的示例性评估的示例性图形视图。

[0366] 图28G描绘了用于评估neonmhc2预测的肽的示例性荧光偏振结合筛选数据;显示为热图以及针对所用肽的每个浓度指示的探针结合抑制百分比。绿色描绘了与颜色强度成比例的良好结合。黄色描绘了中等结合,红色描绘了较差结合,还由相应的抑制百分比值来指示。

[0367] 图28H描绘了neonmhc2预测的肽在示例性结合测定中的评估的总结。

[0368] 图29描绘了对于每个HLA等位基因,来自平均MAPTAC™实验重复(5000万个细胞)的示例性平均肽计数。

[0369] 图30A-30C描绘了HLA II类MAPTAC™等位基因+/-HLA-DM和多等位基因去卷积保真度的示例性结合核心分析。图30A描绘了根据MAPTAC™,采用和不采用HLA-DM共转染(expi293细胞系)和IEDB(其中每个氨基酸的高度与其频率成比例),一个代表性HLA-DR、-DQ和-DP等位基因的示例性序列标识。频率大于10%的氨基酸根据化学性质用颜色表示;所有其他的都显示为灰色。肽根据GibbsCluster工具(补充方法)进行比对,标识代表所有肽,包括那些与整体基序不紧密匹配的肽(例如,没有肽被隔离在“垃圾”簇中)。图30B描绘了掺入泛DR MS数据集中的MAPTAC™肽(每个等位基因为20个)的簇分配的示例性描述。使用GibbsCluster对数据集进行去卷积。每个彩色框代表一个MAPTAC™肽。方框的颜色指示它被分配到哪个簇,灰色条形指示该肽来自哪个等位基因。图30C描绘了示例性图示,其显示对于图30B中所示的等位基因,在锚定位置显示0、1、2、3或4个预期残基的肽的份额。锚定位置

被定义为具有最低熵的四个位置,“预期”残基被定义为在这些位置具有≥10%频率的那些残基。

[0370] 图31A-31F描绘了neonmhc2结合预测算法的示例性架构和基准化。图31A描绘了卷积神经网络(CNN)的示例性架构,其被训练以区分单等位基因HLA II类肽与乱序的长度匹配诱饵。该示意图表明使用氨基酸特征嵌入层、2个宽度为6的卷积层、跳到末端连接的存在以及作为最终逻辑输出节点输入的平均池化和最大池化操作的组合。图31B描绘了在未用于训练或超参数优化的MAPTAC™数据分区上评估的NetMHCIIpan和neonmhc2的示例性阳性预测值(PPV)。对于每个等位基因,n个MS观察到的肽与从同一组源基因中采样的19n个长度匹配的诱饵一起评分,并且每个预测器的n个排名靠前的肽(例如前5%)被判定为阳性。根据此评估方案,PPV与判定率(recall)相同,因为假阳性和假阴性的数目必然相等。图31C描绘了TGEM数据集上的NetMHCIIpan和neonmhc2的示例性PPV。对于每个等位基因,将n个排名靠前的肽判定为阳性,其中n是评估的集合中已确认的免疫原性表位的数目。图31D描绘了新抗原肽的示例性离体T细胞诱导结果。基于HLA-DRB1*11:01的高neonmhc2评分和弱NetMHCIIpan评分选择肽。图31E描绘了在单等位基因MAPTAC数据上训练的模型与在保留单等位基因数据上评估的去卷积多等位基因数据的比较。值如neonmhc2所示,其中对训练数据集进行降采样以匹配去卷积训练集的大小。图31F显示了NetMHCIIpan-v3.1、去卷积训练的预测器和neonmhc2(进行和不进行降采样)的TGEM数据集上的PPV。对于每个等位基因,将n个排名靠前的肽判定为阳性,其中n是评估的集合中确认的免疫原性表位的数目。

[0371] 图32A-32E描绘了HLA II类肿瘤肽组(peptidomes)中的示例性基因表示和蛋白质加工。图32A描绘了通过结直肠癌、黑素瘤和卵巢癌数据集的联合分析确定的每个基因的HLA II类肽的观察数目与预期数目的示例性结果。预期计数是通过将基因长度乘以表达水平得出的。将相关样品的预期计数和观察计数相加。已知存在于血浆中的基因根据其浓度进行标记。图32B描绘了每个细胞定位的肽的预期频率与观察频率的示例性结果。图32C描绘了由蛋白酶体调节的基因的富集评分(观察到的与预期的观察的比率,部分如图32B)分布的示例性结果。基因集合包括那些具有已知泛素化位点的基因和那些在应用蛋白酶体抑制剂后丰度增加的基因。图32D描绘了呈现关于如何处理HLA II类肽的三个示例性工作模型的图示,根据该图,i)组织蛋白酶和其他酶将蛋白质裂解成随后被HLA结合的肽片段,ii)蛋白质或解折叠的多肽结合HLA并且随后被切割为肽长度,iii)蛋白质在结合前被部分消化,并在结合后进一步修剪。每个模型对应不同的预测方法。图32E描绘了与仅使用结合预测的模型相比,对于包括处理相关变量和neonmhc2结合预测的逻辑回归模型观察到的PPV的绝对增加。对通过HLA-DR抗体分析的11个样品进行评估(图30B中分析的相同样品);每个点对应于一个样品。星号表示根据双尾配对t检验的显著改善(*:$p<0.01$,**:$p<0.001$,***:$p<0.0001$)。相同的分析显示在图40B中,而是使用NetMHCIIpan作为基本预测器。用于诱饵选择和PPV计算的方法与图31B中使用的方法相同。

[0372] 图33A-33G描绘了鉴定和预测由树突细胞呈递的肿瘤抗原的示例性结果。图33A描绘了用于鉴定源自癌细胞(K562)的DC呈递的HLA-II配体的实验工作流程的示例性图示。癌细胞在SILAC培养基中生长至完全掺入,裂解或照射,然后接种单核细胞衍生的树突细胞。呈递的肽用泛DR抗体分离并通过LC-MS/MS进行测序。图33B描绘了代表使用与图21A中相同的命中:诱饵比和性能度量对由树突细胞呈递的肿瘤衍生肽的预测性能的示例性数据。显

示了使用和不使用处理特征的基于NetMHCIIpan和neonmhc2的模型的性能。图33C描绘了与轻标记的肽的源基因(灰色曲线,根据DC表达绘制)相比,在UV处理实验中观察到的重标记的肽的源基因(红色曲线,根据K562表达绘制)的示例性基因表达分布。图33D显示了使用具有和不具有处理特征的基于NetMHCIIpan和neonmhc2的模型,以1:499的命中:诱饵比预测呈递的肿瘤抗原的PPV的示例性图示。从左到右的数据点代表以下样品:供体1HOCl处理的细胞:NetMHCIIpan连续表达;NetMHCIIpan连续表达+基因偏好;NetMHCIIpan连续表达+基因偏好+DQ重叠,完全处理模式;供体1,UV处理的:neonmhc2;neonmhc2+阈值表达;neonmhc2+连续表达;neonmhc2+连续表达+基因偏好;neonmhc2+连续表达+基因偏好+DQ重叠。图33E描绘了各种基因定位和功能类别在分别预测重(K562衍生的)和轻(DC衍生的)肽中的显著性。根据控制neonmhc2结合评分和源基因表达的逻辑回归计算P值。条形颜色指示与回归系数相关的符号。图33F描绘了结果的示例性图示,其显示了在UV和HOCl处理的实验中肿瘤细胞衍生的肽源基因(按功能类别着色)的重叠。图33G描绘了示例性数据,其显示了用于使用在第一个供体中观察到的重标记肽上的逻辑模型拟合来预测在第二个供体中呈递的肿瘤抗原的PPV。单独使用neonmhc2结合;使用结合和表达;或结合、表达和指示肽是否来自线粒体基因的二元变量,来拟合模型。

[0373] 图34A-34B描绘了与图29相关的MAPTAC™数据的示例性表征。图34A描绘了通过用编码亲和标记的HLA-A*02:01-BAP的MAPTAC™构建体转染的Expi293细胞系的FACS进行的示例性HLA细胞表面分析。图34B描绘了通过用编码亲和标记的HLA-DRB1*11:01-BAP的MAPTAC™构建体转染的Expi293细胞系的FACS进行的示例性HLA细胞表面分析(下图)。将转染的Expi293细胞(橙色)的HLA细胞表面表达与染色的未转染的Expi293(蓝色)、未染色的未转染的Expi293(红色)、染色的PBMC(深绿色)和未染色的PBMC(浅绿色)进行比较。所有HLA I类染色剂均使用W6/32(泛HLA I类),而HLA II类染色剂使用REA332(泛HLA II类)。

[0374] 图35描绘了与图30A相关的MAPTAC™和IEDB标识的示例性比较。MS观察到的肽的测量亲和力和NetMHCIIpan预测的亲和力,其没有表现出良好NetMHCIIpan评分,但得到了MS的良好支持(评分峰值强度>70且嵌套组大小≥1)。

[0375] 图36A-36C描绘了与图30A-30C相关的HLA-DR1 MAPTAC™数据保真度的示例性分析。图36A描绘了对于常见等位基因,与从蛋白质组随机取样的50,000长度匹配的诱饵肽(蓝色)相比,HLA-DR1MAPTAC™肽(绿色)(长度12-23)的示例性NetMHCIIpan3.1评分。图36B描绘了示例性MS观察到的肽的示例性测量亲和力和NetMHCIIpan预测的亲和力,这些肽没有表现出良好的NetMHCIIpan评分,但得到了MS的良好支持(评分峰强度>70且嵌套组大小≥1)。图36C描绘了如通过MAPTAC™在不同细胞类型中所确定的,HLA-DRB1等位基因的示例性HLA II类序列标识。

[0376] 图37A-37C描绘了与图30A-30C相关的MAPTAC™基序的额外示例性分析。图37A描绘了用于采用和不采用HLA-DM共转染(expi293细胞系)的实验的MAPTAC™衍生的序列标识。图37B描绘了根据MAPTAC™和IEDB的若干HLA I类等位基因的序列标识。请注意,A*32:01在P2处未显示高频率Q,而C*03:03在P9处未显示高频率Y,这与之前使用多等位基因去卷积的研究不同;B*52:01的标识以前未公开。图37C描绘了MAPAC™观察到的肽与CD74的基因序列的示例性比对。

[0377] 图38Ai-38D描绘了与图31A-31D相关的示例性neonmhc2性能统计学和T细胞流染

色。图38Ai描绘了作为训练数据集大小的函数的neonmhc2的示例性性能。以与图31B相同的方式并使用相同的评估肽评估PPV;然而,训练数据被随机降采样以模拟较小的训练数据集。图38Aii描绘了使用GibbsCluster(默认设置;“允许垃圾簇”)衍生自多等位基因HLA-DR配体组的肽簇的示例性序列标识。图38B描绘了来自用neonmhc2预测的新抗原肽判定的诱导样品的CD4+细胞表达IFN-γ的示例性代表性流式细胞图。通过从在不存在新抗原(无肽)的情况下判定时表达IFN-γ的CD4+的百分比减去用新抗原(+肽)判定时表达IFN-γ的CD4+细胞的百分比来计算Delta值。左边的两个流式图代表诱导CD4+ T细胞T细胞应答的新抗原(PEASLYGALSKGSGG)和不诱导T细胞应答的新抗原(PATYILILKEFCLVG)。图38C描绘了来自用单个neonmch2新抗原肽判定的孔的示例性delta值。如果肽具有阳性应答(高于3%的delta应答,突出显示),则将其视为诱导命中。图38D显示了使用GibbsCluster(默认设置;允许“垃圾”簇)为多等位基因HLA-DR配体组衍生的肽簇的示例性序列标识。

[0378] 图39A-39C描绘了与图32A-32E相关的HLA II类的额外示例性起源细胞分析。图39A描绘了根据肽源基因是否存在于人血浆中,在通过泛DR抗体(来自图30B的RG1248、RG1104、RG1095和HDSC)分析的4个PBMC样品中观察到的HLA II类肽的示例性百分排名neonmhc2评分。对于每个肽,在供体中存在的等位基因中使用最佳(最低)百分排名。显示了随机长度匹配的蛋白质组诱饵的评分以供比较。箱线图标记了第5个、第25个、第50个、第75个和第95个百分位数。图39B描绘了使用与图32A中相同的方法对HLA I类每个基因观察到的肽与预期肽的示例性计数。数据对应于相同的肿瘤类型(结直肠癌、卵巢癌和黑素瘤)。人类血浆中存在的基因以蓝色突出显示,并根据其浓度确定大小。图39C描绘了肽观察对于两个不同基因表达谱的示例性相对一致性。对于每个样品,基因水平的肽计数被建模为成块肿瘤基因表达和专职APC基因表达谱的线性组合。系数的比率决定了每个表达谱与肽组库的相对一致性。误差条对应于通过自举重采样计算的95%置信区间。

[0379] 图40A-40B描绘了与图32A-32E相关的处理基序的额外示例性分析。图40A描绘了如在供体PBMC、单核细胞衍生的树突细胞、结直肠癌、黑素瘤、卵巢癌和expi293细胞系(用于大多数MAPTAC™数据生成)中观察到的,相对于平均蛋白质组频率(适用于上游位置U3-U1和下游位置D1-D3)或相对于平均肽频率(适用于内部位置N1-C1),靠近N末端和C末端肽切割位点的示例性氨基酸频率。图40B描绘了除使用NetMHCIIpan作为基本预测器外与图32E相同的分析。对于通过HLA-DR抗体分析的8个样品(图31B中分析的相同样品),除了NetMHCIIpan预测之外还包括处理相关变量的逻辑回归模型所观察到的PPV绝对增加(与仅使用NetMHCIIpan的模型相比)。星号表示根据双尾配对t检验的显著改善(*:$p<0.01$,**:$p<0.001$,***:$p<0.0001$)。

[0380] 图41描绘了用于指代肽上游、肽内和肽下游位置的示例性命名系统。

[0381] 图42A描绘的图示代表通过nLC-MS/MS分析内源加工的以及HLA-1和HLA II类呈递的肽的示例性工作流程。

[0382] 图42B描绘的图示显示了来自采用或不采用FAIMS的胰蛋白酶肽的nLC-MS/MS分析的示例性实验结果。还描绘了在所示分析规模下通过采用或不采用FAIMS的nLC-MS/MS分析检测HLA-1和HLA II类肽的代表性重叠。

[0383] 图43A描绘了采用或不采用FAIMS的示例性HLA I类酸性和碱性反相分级分离肽检测。

[0384] 图43B描绘了示例性实验结果,其显示了对保留时间绘制的HLA I类结合的独特肽的检测。

[0385] 图44A描绘了采用或不采用FAIMS的示例性HLA II类酸性和碱性反相分级分离肽检测。

[0386] 图44B描绘了示例性实验结果,其显示了对保留时间绘制的HLA II类结合的独特肽的检测。

[0387] 图45描绘了使用所示方法检测到的HLA I类结合肽的交叉大小的示例性图示(左)以及用于HLA I类结合肽的LC-MS/MS检测的示例性标准工作流程和优化工作流程的维恩图(右)。

[0388] 图46描绘了使用所示方法检测到的HLA II类结合肽的交叉大小的示例性图示(左)以及用于HLA II类结合肽的LC-MS/MS检测的示例性标准工作流程和优化工作流程的维恩图(右)。

具体实施方式

[0389] 所有术语均应按照它们将被本领域技术人员所理解的那样来理解。除非另有定义,否则本文使用的所有技术和科学术语均具有与本公开所属领域的普通技术人员通常所理解的含义相同的含义。

[0390] 本文所用的章节标题仅用于组织编排的目的,而不应解释为限制所描述的主题。

[0391] 尽管本公开的各个特征可以在单个实施方案的语境中描述,但是这些特征也可以单独提供或以任何合适的组合提供。相反,尽管为了清楚起见,本文可以在单独的实施方案的语境中描述本公开,但是本公开也可以在单个实施方案中实现。

[0392] 本公开基于以下重要发现:使用新的基于计算机的机器学习HLA-肽呈递预测模型——该模型允许使用HLA II类特异性肽来改进免疫治疗,可以高度置信地预测抗原,特别是癌症抗原,被特定HLA II类α和β链对的呈递。

[0393] 在一方面,本公开提供了预测肽的方法,所述肽可以与特定的HLA II类α和β链异二聚体准确配对或结合,使得所述肽与HLA II类蛋白(包括α和β链异二聚体)的高保真度结合确保将特定肽呈递给T淋巴细胞,从而引发特定的免疫应答并避免任何交叉反应或免疫混杂性。最近的几项研究表明,CD4+ T细胞还可以识别HLA II类呈递的配体并有助于控制肿瘤。理想情况下,癌症疫苗和其他免疫疗法会利用引导CD4+ T细胞应答,但目前的努力已经完全放弃了HLA II类抗原预测,因为当前预测工具的准确性不足。

[0394] 在一方面,本公开提供了预测肽的方法,所述肽可以与特定HLA II类蛋白准确地结合,使得当将肽治疗性地施用于表达特定同源HLA II类蛋白的受试者时,借助于HLA II类蛋白激活CD4+ T细胞并刺激免疫记忆的能力,可以用该肽激活更持续且有力的免疫应答。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约1.1倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约2倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约3倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特

定HLA II类蛋白预测方面表现出至少约4倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约5倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约6倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约7倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约8倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约9倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约10倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约15倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约20倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约30倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约40倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约50倍的改善。在一些实施方案中,相对于当前可用的预测器,本文提供的方法在特定HLA II类蛋白预测方面表现出至少约60倍的改善。

[0395] 在一方面,本文提出了针对特定受试者定制或个性化的免疫治疗方法。每个受试者或患者都表达特定的一组HLA I类和HLA II类蛋白。HLA分型是一种众所周知的技术,其允许确定受试者表达的HLA蛋白的特定组库(repertoire)。一旦已知由特定受试者表达的HLA异二聚体,具有如本文所述的用于以高保真度预测可与特定HLA II类α和β链异二聚体结合的肽的改进的、复杂的和可靠的方法,可以确保能生成专门为该受试者定制的特异性免疫应答。

[0396] 在本申请中,除非另有特别说明,否则单数形式的使用包括复数形式。必须指出,除非上下文另有明确说明,否则如本说明书中所用的,单数形式“一个”、“一种”和“该”包括复数指示物。在本申请中,除非另有说明,否则“或”的使用意指“和/或”。此外,术语“包括”以及其他形式如“包含”、“含有”和“具有”的使用不是限制性的。术语“一个或多个”或“至少一个”,例如一组成员中的一个或多个或至少一个成员,本身是清楚的,通过进一步举例说明,该术语尤其包括提及任何一个所述成员,或任何两个或更多个所述成员,例如,所述成员中的任意3、4、5、6或7个等,直至所有所述成员。

[0397] 本说明书中提及“一些实施方案”、“实施方案”、“一个实施方案”或“其他实施方案”意指与该实施方案相关描述的特征、结构或特性包含在本公开的至少一些实施方案中,但不一定包含在所有实施方案中。

[0398] 如在本说明书和权利要求书中所用的,词语“包含”(和任何形式的包含)、“具有”(和任何形式的具有)、“包括”(和任何形式的包括)或“含有”(和任何形式的含有)是包含性的或开放式的,并不排除其他未列举的要素或方法步骤。可以想到,本说明书中讨论的任何实施方案可以采用本公开的任何方法或组合物来实施,反之亦然。此外,本公开的组合物可以用来实现本公开的方法。

[0399] 当涉及诸如参数、量、持续时间等可测量的值时,如本文所用的术语“约”或“大约”

旨在涵盖指定值的+/-20％或更小、+/-10％或更小、+/-5％或更小或+/-1％或更小的变化，只要这样的变化适合在本公开中进行。应当理解，修饰语“约”或“大约”所指的值本身也被具体公开。

[0400] 术语“免疫应答”包括受T细胞共刺激调节影响的T细胞介导的和/或B细胞介导的免疫应答。示例性的免疫应答包括T细胞应答，例如细胞因子的产生和细胞的细胞毒性。另外，术语免疫应答包括受T细胞活化间接影响的免疫应答，例如抗体产生(体液应答)和细胞因子应答性细胞例如巨噬细胞的活化。

[0401] “受体”应被理解为是指能够结合配体的生物分子或分子分组。受体可以用来在细胞、细胞形成或生物体中传递信息。受体包含至少一个受体单元，并且可以含有两个或更多个受体单元，其中每个受体单元可以由蛋白质分子例如糖蛋白分子组成。受体具有与配体的结构互补的结构，并且可以作为结合配偶体与配体复合。信号信息可以通过受体与细胞表面上的配体结合后的构象变化来传递。根据本公开，受体可指能够与配体例如适当长度的肽或肽片段形成受体/配体复合物的MHC I类和II类蛋白。由HLA I类和II类等位基因编码的I类和II类MHC肽在此通常分别称为HLA I类和HLA II类肽，或HLA I类和HLA II类肽，或HLA I类II类蛋白，或HLA I类和HLA II类蛋白，或HLA I类和II类分子，或其此类常见变体，如本领域普通技术人员在讨论的上下文中所熟知的。

[0402] “配体”是能够与受体形成复合物的分子。根据本公开，配体应被理解为是指例如在其氨基酸序列中具有合适的长度和合适的结合基序的肽或肽片段，从而该肽或肽片段能够与MHC I类或MHC II类蛋白(即，HLA I类和HLA II类蛋白)结合并形成复合物。

[0403] “抗原”是能够刺激免疫应答的分子，并且可以由癌细胞或传染原或自身免疫性疾病产生。被T细胞(无论是辅助性T淋巴细胞(T辅助(TH)细胞)还是细胞毒性T淋巴细胞(CTL))识别的抗原不是作为完整蛋白质被识别，而是作为与细胞表面上的HLA I类或II类蛋白缔合的小肽被识别。在自然发生的免疫应答的过程中，与抗原呈递细胞(APC)上的HLA II类分子缔合而被识别的抗原从细胞外获取，内化，并加工成与HLA II类分子缔合的小肽。APC还可以通过加工外源抗原并将加工后的抗原呈递给HLA I类分子来交叉呈递肽抗原。产生与HLA I类MHC分子缔合而被识别的肽的抗原通常是在细胞内产生的肽，并且这些抗原被加工并与I类MHC分子缔合。现已理解，与给定的HLA I类或II类分子缔合的肽被表征为具有共同的结合基序，并且已经确定了针对大量不同的HLA I类和II类分子的结合基序。也可以合成与给定抗原的氨基酸序列相对应并含有针对给定HLA I类或II类分子的结合基序的合成肽。然后可以将这些肽添加至适当的APC，并且可以使用该APC在体外或体内刺激T辅助细胞或CTL应答。结合基序、合成肽的方法和刺激T辅助细胞或CTL应答的方法都是本领域普通技术人员已知的并且容易获得。

[0404] 在本说明书中，术语“肽”与“突变肽”和“新抗原肽”可互换使用。类似地，在本说明书中，术语“多肽”与“突变多肽”和“新抗原多肽”可互换使用。“新抗原”或“新表位”是指由表达的蛋白质中的肿瘤特异性突变产生的一类肿瘤抗原或肿瘤表位。本公开进一步包括包含肿瘤特异性突变的肽，包含已知肿瘤特异性突变的肽，以及通过本公开的方法鉴定的突变多肽或其片段。这些肽和多肽在本文中被称为“新抗原肽”或“新抗原多肽”。这些多肽或肽可具有多种长度，可以是其中性(不带电荷的)形式，也可以是盐形式，并且不含修饰，如糖基化、侧链氧化、磷酸化或任何翻译后修饰，或含有这些修饰，条件是该修饰不会破坏本

文所述的多肽的生物学活性。在一些实施方案中,本公开的新抗原肽可包括:对于HLA I类,长度为22个或更少的残基,例如,约8个至约22个残基,约8个至约15个残基,或9或10个残基;对于HLA II类,长度为40个或更少的残基,例如,长度为约8个至约40个残基,长度为约8个至约24个残基,约12个至约19个残基,或约14个至约18个残基。在一些实施方案中,新抗原肽或新抗原多肽包含新表位。

[0405] 术语“表位”包括能够与本文所定义的抗体、抗体肽和/或抗体样分子(包括但不限于T细胞受体)特异性结合的任何蛋白质决定簇。表位决定簇通常由分子的化学活性表面基团如氨基酸或糖侧链组成,并且通常具有特定的三维结构特征以及特定的电荷特征。

[0406] “T细胞表位”是这样的肽序列,其可以以呈递肽的MHC分子或MHC复合物的形式被I或II类MHC分子结合,然后以这种形式分别被细胞毒性T淋巴细胞或T辅助细胞识别并结合。

[0407] 如本文所用的术语“抗体”包括IgG(包括IgG1、IgG2、IgG3和IgG4)、IgA(包括IgA1和IgA2)、IgD、IgE、IgM和IgY,并且意在包括完整抗体,包括单链完整抗体,及其抗原结合(Fab)片段。抗原结合抗体片段包括但不限于Fab、Fab'和F(ab')2、Fd(由VH和CH1组成)、单链可变片段(scFv)、单链抗体、二硫键连接的可变片段(dsFv)和包含VL或VH结构域的片段。抗体可以来自任何动物来源。抗原结合抗体片段,包括单链抗体,可以包含单独的或与以下全部或部分组合的可变区:铰链区、CH1、CH2和CH3结构域。还包括可变区和铰链区、CH1、CH2和CH3结构域的任何组合。抗体可以是例如特异性结合HLA关联多肽或HLA-HLA结合肽(HLA-肽)复合物的单克隆抗体、多克隆抗体、嵌合抗体、人源化抗体以及人单克隆和多克隆抗体。本领域技术人员将会认识到,多种免疫亲和技术适合于富集可溶性蛋白质,如可溶性HLA-肽复合物或膜结合的HLA关联多肽,例如,其已通过蛋白水解从膜上切割下来。这包括以下技术,其中(1)将一种或多种能够与可溶性蛋白质特异性结合的抗体固定在固定的或可移动的基底(例如,塑料孔或树脂、乳胶或顺磁珠)上,以及(2)使含有来自生物样品的可溶性蛋白质的溶液通过抗体包被的基底,从而使可溶性蛋白质与抗体结合。从溶液中分离具有抗体和结合的可溶性蛋白质的基底,并且任选地例如通过改变浸浴抗体的溶液的pH和/或离子强度和/或离子组成来使抗体和可溶性蛋白质解离。或者,可以使用免疫沉淀技术,其中将抗体和可溶性蛋白质组合并形成大分子聚集体。该大分子聚集体可通过大小排阻技术或通过离心从溶液中分离。

[0408] 术语“免疫纯化(IP)”(或免疫亲和纯化或免疫沉淀)是本领域公知的方法,并广泛应用于从样品中分离所需抗原。通常,该方法包括使含有所需抗原的样品与亲和基质接触,该亲和基质包含共价附接至固相的针对该抗原的抗体。样品中的抗原通过免疫化学键与亲和基质结合。然后洗涤亲和基质以除去任何未结合的物质。通过改变与亲和基质接触的溶液的化学组成,从亲和基质中取出抗原。免疫纯化可以在含有亲和基质的柱上进行,在这种情况下溶液是洗脱液。或者,免疫纯化可以是分批过程,在这种情况下,亲和基质保持为溶液中的悬浮液。该过程中的重要步骤是从基质中取出抗原。这通常通过增加与亲和基质接触的溶液的离子强度来实现,例如通过添加无机盐。pH的改变也可以有效地解离抗原与亲和基质之间的免疫化学键。

[0409] “药剂”是任何小分子化合物、抗体、核酸分子或多肽或其片段。

[0410] “改变”或“变化”是增加或降低。改变可能少至1%、2%、3%、4%、5%、10%、20%、30%或40%、50%、60%,甚至多达70%、75%、80%、90%或100%。

[0411] “生物样品”是来源于生物体的任何组织、细胞、流体或其他物质。如本文所用的，术语“样品”包括生物样品，例如来源于生物体的任何组织、细胞、流体或其他物质。“特异性结合”是指识别并结合分子(例如，多肽)但基本上不识别并结合样品(例如，生物样品)中的其他分子的化合物(例如，肽)。

[0412] “捕获试剂”是指特异性结合分子(例如，核酸分子或多肽)以选择或分离该分子(例如，核酸分子或多肽)的试剂。

[0413] 如本文所用的，术语“确定”、“评估”、“测定”、“测量”、“检测”及其语法等同语是指定量和定性确定，因此，术语“确定”与“测定”、“测量”等在本文中可互换使用。在意欲定量确定的情况下，使用短语“确定分析物等的量”。在意欲定性和/或定量确定的情况下，使用短语“确定分析物的水平”或“检测”分析物。

[0414] “片段”是与参考蛋白质或核酸基本相同的蛋白质或核酸的一部分。在一些实施方案中，该部分保留本文所述参考蛋白质或核酸的生物学活性的至少50％、75％或80％或90％、95％，乃至99％。

[0415] 术语“分离的”、“纯化的”、“生物学纯的”及其语法等同语是指从在其天然状态下通常与其伴随的组分中不同程度地释放出来的物质。“分离”表示与原始来源或环境的分离程度。“纯度”表示高于分离的分离程度。“纯化的”或“生物学纯的”蛋白质充分不含其他物质，使得任何杂质均不会实质性地影响该蛋白质的生物学性质或引起其他不利后果。即，如果当通过重组DNA技术生产时基本上不含细胞材料、病毒材料或培养基，或者当化学合成时基本上不含化学前体或其他化学品，则本公开的核酸或肽是纯化的。纯度和均质性通常使用分析化学技术来确定，例如聚丙烯酰胺凝胶电泳或高效液相色谱法。术语“纯化的”可以表示核酸或蛋白质在电泳凝胶中基本上产生一个条带。对于可以进行修饰例如磷酸化或糖基化的蛋白质，不同的修饰可以产生不同的分离的蛋白质，可以将其分别纯化。

[0416] “分离的”多肽(例如，来自HLA-肽复合物的肽)或多肽复合物(例如，HLA-肽复合物)是已经与天然伴随的组分分离的本公开的多肽或多肽复合物。通常，当多肽或多肽复合物至少60重量％不含与之天然关联的蛋白质和天然存在的有机分子时，其为分离的。该制品可以是按重量计至少75％、至少90％或至少99％的本公开的多肽或多肽复合物。本公开的分离的多肽或多肽复合物可以例如通过从天然来源中提取、通过表达编码该多肽或多肽复合物的一种或多种组分的重组核酸或通过化学合成该多肽或该多肽复合物的一种或多种组分而获得。纯度可以通过任何适当的方法来测量，例如柱色谱法、聚丙烯酰胺凝胶电泳或通过HPLC分析。在一些情况下，HLA等位基因编码的MHC II类蛋白(即MHC II类肽)在本文件中可互换地称为HLA II类蛋白(或HLA II类肽)。

[0417] 术语“载体”是指能够转运或介导异源核酸表达的核酸分子。质粒是术语“载体”所涵盖的种类中的一种。载体通常是指含有复制起点和在宿主细胞中复制和/或维持所必需的其他实体的核酸序列。能够指导与其可操作地连接的基因和/或核酸序列的表达的载体在本文中被称为“表达载体”。通常，有用的表达载体通常是“质粒”的形式，其是指环状双链DNA分子，其在载体形式下不与染色体结合，并且通常包含用于稳定或瞬时表达的实体或编码的DNA。可以在本文公开的方法中使用的其他表达载体包括但不限于质粒、附加体、细菌人工染色体、酵母人工染色体、噬菌体或病毒载体，并且此类载体可以整合到宿主的基因组中或在细胞中自主复制。载体可以是DNA或RNA载体。也可以使用本领域技术人员已知的发

挥等效功能的其他形式的表达载体,例如,自我复制的染色体外载体或能够整合到宿主基因组中的载体。示例性载体是能够自主复制和/或表达与其连接的核酸的载体。

[0418] 对于融合蛋白使用的术语“间隔体”或“连接体”是指连接包含融合蛋白的蛋白质的肽。通常,间隔体除了连接或保持蛋白质或RNA序列之间的某种最小距离或其他空间关系以外,没有具体的生物学活性。然而,在一些实施方案中,可以选择间隔体的组成氨基酸以影响分子的某些性质,如分子的折叠、净电荷或疏水性。供本公开的实施方案使用的合适的连接体是本领域技术人员公知的,并且包括但不限于直链或支链碳连接体、杂环碳连接体或肽连接体。连接体用来将两个抗原肽分开一定的距离,该距离足以确保在一些实施方案中每个抗原肽正确地折叠。示例性的肽连接体序列采用柔性的延伸构象,并且不表现出发展有序二级结构的倾向。柔性蛋白质区域中的典型氨基酸包括Gly、Asn和Ser。实际上,预期含有Gly、Asn和Ser的氨基酸序列的任何排列都将满足上述针对连接体序列的标准。其他接近中性的氨基酸,如Thr和Ala,也可以在连接体序列中使用。在Maratea等人(1985),Gene 40:39-46、Murphy等人(1986)Proc.Nat'l.Acad.Sci.USA 83:8258-62、美国专利4,935,233和美国专利4,751,180中公开了可用作连接体的其他氨基酸序列。

[0419] 术语“瘤形成”是指导致不适当地高水平的细胞分裂、不适当地低水平的凋亡或两者或由其引起的任何疾病。胶质母细胞瘤是瘤形成或癌症的一个非限制性实例。术语“癌症”或“肿瘤”或“过度增殖性病症”是指具有致癌细胞的典型特征如不受控制的增殖、无限增殖性、转移潜能、快速生长和增殖速率以及某些特有形态特征的细胞的存在。癌细胞通常为肿瘤形式,但此类细胞可以单独存在于动物体内,也可以是非致瘤性癌细胞,如白血病细胞。癌症包括但不限于B细胞癌症,例如,多发性骨髓瘤、瓦尔登斯特伦巨球蛋白血症(Waldenstrom's macroglobulinemia)、重链病(例如α链疾病、γ链疾病及μ链疾病)、良性单克隆丙种球蛋白病和免疫细胞淀粉样变性、黑素瘤、乳腺癌、肺癌、支气管癌、结直肠癌、前列腺癌(例如,转移性、激素难治性前列腺癌)、胰腺癌、胃癌、卵巢癌、膀胱癌、脑或中枢神经系统癌症、周围神经系统癌、食管癌、宫颈癌、子宫癌或子宫内膜癌、口腔或咽部的癌症、肝癌、肾癌、睾丸癌、胆道癌、小肠或阑尾癌、唾液腺癌、甲状腺癌、肾上腺癌、骨肉瘤、软骨肉瘤、血液组织的癌症等。适用于本公开所涵盖的方法的癌症类型的其他非限制性实例包括人类肉瘤和癌,例如,纤维肉瘤、粘液肉瘤、脂肪肉瘤、软骨肉瘤、骨肉瘤、脊索瘤、血管肉瘤、内皮肉瘤、淋巴管肉瘤、淋巴管内皮肉瘤、滑膜瘤、间皮瘤、尤因瘤(Ewing's tumor)、平滑肌肉瘤、横纹肌肉瘤、结肠癌、结直肠癌、胰腺癌、乳腺癌、卵巢癌、鳞状细胞癌、基底细胞癌、腺癌、汗腺癌、皮脂腺癌、乳头状癌、乳头状腺癌、囊腺癌、髓样癌、支气管癌、肾细胞癌、肝细胞瘤、胆管癌、肝癌、绒毛膜癌、精原细胞瘤、胚胎性癌、维尔姆斯瘤、宫颈癌、骨癌、脑瘤、睾丸癌、肺癌、小细胞肺癌、膀胱癌、上皮癌、神经胶质瘤、星形细胞瘤、髓母细胞瘤、颅咽管瘤、室管膜瘤、松果体瘤、血管母细胞瘤、听神经瘤、少突胶质细胞瘤、脑膜瘤、黑素瘤、神经母细胞瘤、视网膜母细胞瘤;白血病,例如,急性淋巴细胞白血病及急性髓细胞性白血病(成髓细胞性、早幼粒细胞性、粒单核细胞性、单核细胞性和红白血病);慢性白血病(慢性髓细胞性(粒细胞性)白血病及慢性淋巴细胞白血病);以及真性红细胞增多症、淋巴瘤(霍奇金病及非霍奇金病)、多发性骨髓瘤、瓦尔登斯特伦巨球蛋白血症和重链病。在一些实施方案中,该癌症是上皮癌,例如但不限于膀胱癌、乳腺癌、宫颈癌、结肠癌、妇科癌症、肾癌、喉癌、肺癌、口腔癌、头颈癌、卵巢癌、胰腺癌、前列腺癌或皮肤癌。在其他实施方案中,该癌症是乳腺癌、前列

腺癌、肺癌或结肠癌。在另外其他的实施方案中,该上皮癌是非小细胞肺癌、非乳头状肾细胞癌、宫颈癌、卵巢癌(例如,浆液性卵巢癌)或乳腺癌。上皮癌可以用各种其他方式表征,包括但不限于浆液性、子宫内膜样、粘液性、透明细胞、布伦纳型(brenner)或未分化的。在一些实施方案中,本公开用于淋巴瘤或其亚型(包括但不限于套细胞淋巴瘤)的治疗、诊断和/或预后。淋巴组织增生性病症也被认为是增殖性疾病。

[0420] 术语“疫苗”应被理解为是指用于产生免疫力以预防和/或治疗疾病(例如,瘤形成/肿瘤/传染原/自身免疫性疾病)的组合物。因此,疫苗是包含抗原的药物,并且旨在用于人类或动物中以通过接种产生特异性防御和保护性物质。“疫苗组合物”可以包含药学上可接受的赋形剂、载体或稀释剂。本公开的方面涉及该技术在制备基于抗原的疫苗中的应用。在这些实施方案中,疫苗是指一种或多种疾病特异性抗原肽(或编码它们的相应核酸)。在一些实施方案中,基于抗原的疫苗含有至少两种、至少三种、至少四种、至少五种、至少六种、至少七种、至少八种、至少九种、至少10种、至少11种、至少12种、至少13种、至少14种、至少15种、至少16种、至少17种、至少18种、至少19种、至少20种、至少21种、至少22种、至少23种、至少24种、至少25种、至少26种、至少27种、至少28种、至少29种、至少30种或更多种抗原肽。在一些实施方案中,基于抗原的疫苗含有2至100、2至75、2至50、2至25、2至20、2至19、2至18、2至17、2至16、2至15、2至14、2至13、2至12、2至10、2至9、2至8、2至7、2至6、2至5、2至4、3至100、3至75、3至50、3至25、3至20、3至19、3至18、3至17、3至16、3至15、3至14、3至13、3至12、3至10、3至9、3至8、3至7、3至6、3至5、4至100、4至75、4至50、4至25、4至20、4至19、4至18、4至17、4至16、4至15、4至14、4至13、4至12、4至10、4至9、4至8、4至7、4至6、5至100、5至75、5至50、5至25、5至20、5至19、5至18、5至17、5至16、5至15、5至14、5至13、5至12、5至10、5至9、5至8或5至7种抗原肽。在一些实施方案中,基于抗原的疫苗含有2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种抗原肽。在一些情况下,该抗原肽是新抗原肽。在一些情况下,该抗原肽包含一种或多种新表位。

[0421] 术语“药学上可接受的”是指由联邦或州政府的监管机构批准或可批准,或在美国药典或其他公认的药典中列出用于动物,包括人类。“药学上可接受的赋形剂、载体或稀释剂”是指可以与药剂一起施用于受试者并且不会破坏其药理活性并且当以足以递送治疗量的药剂的剂量施用时无毒的赋形剂、载体或稀释剂。如本文所述,合并的疾病特异性抗原的“药学上可接受的盐”可以是本领域中通常认为适用于与人类或动物组织接触而没有过度毒性、刺激性、变态反应或其他问题或并发症的酸式盐或碱式盐。这样的盐包括碱性残基如胺的无机和有机酸盐,以及酸性残基如羧酸的碱金属或有机盐。具体的药物盐包括但不限于诸如以下酸的盐:盐酸、磷酸、氢溴酸、苹果酸、乙醇酸、富马酸、硫酸、对氨基磺酸、氨基苯磺酸、甲酸、甲苯磺酸、甲烷磺酸、苯磺酸、乙烷二磺酸、2-羟乙基磺酸、硝酸、苯甲酸、2-乙酰氧基苯甲酸、柠檬酸、酒石酸、乳酸、硬脂酸、水杨酸、谷氨酸、抗坏血酸、帕莫酸、琥珀酸、富马酸、马来酸、丙酸、羟基马来酸、氢碘酸、苯乙酸、链烷酸如乙酸、HOOC-(CH2)n-COOH,其中n为0-4,等等。类似地,药学上可接受的阳离子包括但不限于钠、钾、钙、铝、锂和铵。本领域普通技术人员从本公开内容和本领域中的知识将会认识到,本文提供的合并的疾病特异性抗原的其他药学上可接受的盐,包括Remington’s Pharmaceutical Sciences,第17版,Mack Publishing Company,Easton,PA,p.1418(1985)列出的那些。通常,可通过任何常规化学方法由含有碱性或酸性部分的母体化合物合成药学上可接受的酸式盐或碱式盐。简言之,可

以通过在合适的溶剂中使这些化合物的游离酸或碱形式与化学计量量的适当碱或酸反应来制备这类盐。

[0422] 可用于本公开的方法的核酸分子包括编码本公开的多肽或其片段的任何核酸分子。这样的核酸分子不必与内源核酸序列100％相同,但是通常显示出实质的同一性。与内源序列具有实质同一性的多核苷酸通常能够与双链核酸分子的至少一条链杂交。“杂交”是指核酸分子在各种严格性条件下在互补的多核苷酸序列或其部分之间配对形成双链分子。(参见,例如,Wahl,G.M.和S.L.Berger(1987)Methods Enzymol.152:399;Kimmel,A.R.(1987)Methods Enzymol.152:507)。例如,严格的盐浓度通常可以小于约750mM NaCl和75mM柠檬酸三钠,小于约500mM NaCl和50mM柠檬酸三钠,或小于约250mM NaCl和25mM柠檬酸三钠。在不存在有机溶剂例如甲酰胺的情况下可以获得低严格性杂交,而在至少约35％的甲酰胺或至少约50％的甲酰胺的存在下可以获得高严格性杂交。严格的温度条件通常可以包括至少约30C、至少约37C或至少约42C的温度。改变其他参数,如杂交时间,例如十二烷基硫酸钠(SDS)的去污剂的浓度,以及载体DNA的包含或排除,是本领域技术人员公知的。通过根据需要组合这些各种条件来实现各种严格性水平。在一个示例性实施方案中,杂交可以在30C下在750mM NaCl、75mM柠檬酸三钠和1％SDS中发生。在另一个示例性实施方案中,杂交可以在37C下在500mM NaCl、50mM柠檬酸三钠、1％SDS、35％甲酰胺和100μg/ml变性鲑精DNA(ssDNA)中进行。在另一个示例性实施方案中,杂交可以在42C下在250mM NaCl、25mM柠檬酸三钠、1％SDS、50％甲酰胺和200μg/ml ssDNA中发生。关于这些条件有用的变化对于本领域技术人员将是显而易见的。对于大多数应用,杂交后的洗涤步骤在严格性上也可能有所不同。洗涤严格性条件可以通过盐浓度和温度来定义。如上所述,可以通过降低盐浓度或通过提高温度来提高洗涤严格性。例如,洗涤步骤的严格盐浓度可以小于约30mM NaCl和3mM柠檬酸三钠,或小于约15mM NaCl和1.5mM柠檬酸三钠。洗涤步骤的严格温度条件可以包括至少约25C、至少约42C或至少约68C的温度。在示例性实施方案中,洗涤步骤可以在25C下在30mM NaCl、3mM柠檬酸三钠和0.1％SDS中进行。在其他示例性实施方案中,洗涤步骤可以在42C下在15mM NaCl、1.5mM柠檬酸三钠和0.1％SDS中进行。在另一个示例性实施方案中,洗涤步骤可以在68C下在15mM NaCl、1.5mM柠檬酸三钠和0.1％SDS中进行。这些条件的其他变化对于本领域技术人员将是显而易见的。杂交技术是本领域技术人员公知的,并且描述于例如Benton和Davis(Science 196:180,1977);Grunstein和Hogness(Proc.Natl.Acad.Sci.,USA 72:3961,1975);Ausubel等人(Current Protocols in Molecular Biology,Wiley Interscience,New York,2001);Berger和Kimmel(Guide to Molecular Cloning Techniques,1987,Academic Press,New York);和Sambrook等人,Molecular Cloning:A Laboratory Manual,Cold Spring Harbor Laboratory Press,New York。

[0423] “基本上相同”是指与参考氨基酸序列(例如,本文所述的任何一种氨基酸序列)或核酸序列(例如,本文所述的任何一种核酸序列)显示出至少50％同一性的多肽或核酸分子。这样的序列在氨基酸水平或核酸水平上与用于比较的顺序可以至少60％、80％或85％、90％、95％、96％、97％、98％乃至99％或更高水平地相同。通常使用序列分析软件(例如,Sequence Analysis Software Package of the Genetics Computer Group,University of Wisconsin Biotechnology Center,1710University Avenue,Madison,Wis.53705,

BLAST,BESTFIT,GAP,或PILEUP/PRETTYBOX程序)来测量序列同一性。这样的软件通过将同源性程度指定给各种置换、缺失和/或其他修饰来匹配相同或相似的序列。保守置换通常包括以下组内的置换:甘氨酸、丙氨酸;缬氨酸、异亮氨酸、亮氨酸;天冬氨酸、谷氨酸、天冬酰胺、谷氨酰胺;丝氨酸、苏氨酸;赖氨酸、精氨酸;和苯丙氨酸、酪氨酸。在确定同一性程度的示例性方法中,可以使用BLAST程序,其中e-3与e-m°之间的概率得分表示密切相关的序列。"参考"是比较标准。

[0424] 术语"受试者"或"患者"是指作为治疗、观察或实验对象的动物。仅作为示例,受试者包括但不限于哺乳动物,包括但不限于人类或非人类哺乳动物,如非人类灵长类、鼠、牛、马、犬、绵羊或猫。

[0425] 术语"治疗"、"处理"等意指减少、预防或改善病症和/或与之相关的症状(例如,瘤形成或肿瘤或传染原或自身免疫性疾病)。"治疗"可以指在疾病(例如,癌症或传染原的感染或自身免疫性疾病)发作或怀疑发作之后向受试者施用治疗。"治疗"包括"减轻"的概念,后者是指降低与疾病有关的任何症状或其他不良作用和/或与治疗相关的副作用的发生或复发频率或严重程度。术语"治疗"还涵盖"管理"的概念,后者是指降低患者中的疾病或病症的严重程度,例如,延长该疾病患者的寿命或延长其生存期,或延迟其复发,例如,延长已罹患该疾病的患者的缓解期。应当理解,尽管不排除,但治疗疾病或病况并不需要完全消除该病症、病况或与其相关的症状。

[0426] 如本文所用的,术语"预防"、"防止"及其语法等同语是指在开始施用药剂或化合物时尚未发展出此类症状的受试者中避免或延迟与疾病或病况相关的症状的发作。

[0427] 术语"治疗效果"是指病症(例如,瘤形成、肿瘤或传染原的感染或自身免疫性疾病)的一种或多种症状或其相关病理学的一定程度的减轻。如本文所用的"治疗有效量"是指在向细胞或受试者单次或多次剂量施用后有效延长此类病症患者的生存期、减轻该病症的一种或多种体征或症状、预防或延迟等超出在没有此类治疗的情况下所预期的程度的药剂量。"治疗有效量"旨在限定达到治疗效果所需的量。本领域具有普通技能的医师或兽医可以容易地确定并开出所需药物组合物的"治疗有效量"(例如,ED50)。例如,医师或兽医可以以低于获得所需治疗效果所需水平的水平开始在药物组合物中使用的本发明化合物的剂量,并逐渐增加剂量直至获得所需效果。疾病、病况和病症在本文中可互换使用。

[0428] 本领域普通技术人员将会认识到,术语"肽标签"、"亲和标签"、"表位标签"或"亲和受体标签"在本文中可互换使用。如本文所用的,术语"亲和受体标签"是指允许例如通过亲和纯化容易地检测或纯化所标记的蛋白质的氨基酸序列。亲和受体标签通常(但不是必须)放置在HLA等位基因N-或C-末端处或附近。各种肽标签是本领域公知的。非限制性实例包括聚组氨酸标签(例如,4至15个连续的His残基,如8个连续的His残基);聚组氨酸-甘氨酸标签;HA标签(例如,Field等人,Mol.Cell.Biol.,8:2159,1988);c-myc标签(例如,Evans等人,Mol.Cell.Biol.,5:3610,1985);单纯疱疹病毒糖蛋白D(gD)标签(例如,Paborsky等人,Protein Engineering,3:547,1990);FLAG标签(例如,Hopp等人,BioTechnology,6:1204,1988;美国专利4,703,004号和第4,851,341);KT3表位标签(例如,Martine等人,Science,255:192,1992);微管蛋白表位标签(例如,Skinner,Biol.Chem.,266:15173,1991);T7基因10蛋白肽标签(例如,Lutz-Freyemuth等人,Proc.Natl.Acad.Sci.USA,87:6393,1990);链霉亲和素标记(StrepTagTM或StrepTagIITM;参见,例如,Schmidt等人,

J.Mol.Biol.,255(5):753-766,1996或美国专利5,506,121;也可以从Sigma-Genosys商购获得);或衍生自水疱性口炎病毒糖蛋白的VSV-G表位标签;或衍生自猿猴病毒5(SV5)副粘病毒P和V蛋白上发现的小表位(Pk)的V5标签。在一些实施方案中,所述亲和受体标签是"表位标签",这是一种类型的肽标签,其向HLA蛋白添加可识别的表位(抗体结合位点)以提供相应抗体的结合,从而允许鉴定或亲和纯化所标记的蛋白质。表位标签的非限制性实例是可与IgG结合的蛋白A或蛋白G。在一些实施方案中,IgG Sepharose 6Fast Flow色谱树脂的基质与人IgG共价偶联。该树脂允许高流速,快速且方便地纯化用蛋白A标记的蛋白质。许多其他标签部分是普通技术人员已知的并且可以设想的,并且在此想到。可以使用任何肽标签,只要它能够被表达为亲和受体标记的HLA-肽复合物的元件即可。

[0429] 如本文所用的,术语"亲和分子"是指以化学特异性与亲和受体肽结合的分子或配体。化学特异性是蛋白质结合位点结合特定配体的能力。蛋白质可以结合的配体越少,其特异性越高。特异性描述了给定蛋白质与配体之间的结合的强度。这种关系可以通过解离常数(KD)来描述,该常数表征蛋白质-配体系统的结合状态与未结合状态之间的平衡。

[0430] 术语"亲和受体标记的HLA-肽复合物"是指包含与包含亲和受体肽的单等位基因重组HLA I类或II类肽特异性结合的HLA I类或II类关联肽或其部分的复合物。

[0431] 当用于亲和分子和亲和受体标签或表位与HLA肽的相互作用时,术语"特异结合"或"特异性结合"是指该相互作用取决于蛋白质上的特定结构(例如,抗原决定簇或表位)的存在;换句话说,亲和分子识别并结合特定的亲和受体肽结构,而不是总体上结合蛋白质。

[0432] 如本文所用的,术语"亲和力"是指结合对的两个成员(例如,"亲和受体标签"和"亲和分子"以及HLA结合肽和HLA I类或II类分子)之间的结合强度的量度。KD是解离常数并且具有摩尔浓度的单位。亲和常数是解离常数的倒数。亲和常数有时用作描述该化学实体的通用术语。它是结合能量的直接量度。亲和力可以例如使用市售Biacore SPR单元通过表面等离子体共振(SPR)经实验确定。亲和力也可以被表示为抑制浓度50(IC50),即50%的肽被替代时的浓度。同样,lnIC50是指IC50的自然对数。K_{off}是指解离速率常数,例如,亲和分子从亲和受体标记的HLA-肽复合物上解离的速率常数。

[0433] 在一些实施方案中,亲和受体标记的HLA-肽复合物包含生物素受体肽(BAP),并使用链霉亲和素/NeutrAvidin珠从复合细胞混合物中免疫纯化。生物素-亲和素/链霉亲和素结合是自然界中已知的最强非共价相互作用。该性质作为生物学工具用于广泛的应用,例如与生物素共价连接的蛋白质的免疫纯化。在一个示例性实施方案中,编码HLA等位基因的核酸序列将生物素受体肽(BAP)作为用于免疫纯化的亲和受体标签使用。BAP可以在体内或体外在标签内的单个赖氨酸残基处进行特异性生物素化(例如,美国专利5,723,584;5,874,239;和5,932,433;以及英国专利GB2370039)。BAP通常长15个氨基酸,并且含有一个赖氨酸作为生物素受体残基。在一些实施方案中,将BAP置于单等位基因HLA肽的N-或C-末端处或附近。在一些实施方案中,将BAP置于HLA I类肽的重链结构域和β2微球蛋白结构域之间。在一些实施方案中,将BAP置于HLA II类肽的β链结构域和α链结构域之间。在一些实施方案中,将BAP置于HLA I类重链的α1、α2和α3结构域之间的环区域中,或者分别位于HLA II类的α链和β链的α1与α2和β1与β2结构域之间的环区域中。图2中描述了针对HLA I和II类表达而设计的示例性构建体,其使用用于生物素化和免疫纯化的BAP。

[0434] 如本文所用的,术语"生物素"是指化合物生物素本身及其类似物、衍生物和变体。

因此,术语“生物素”包括生物素(顺式六氢-2-氧代-1H-噻吩并[3,4]咪唑-4-戊酸)及其任何衍生物和类似物,包括生物素样化合物。这类化合物包括例如生物素-e-N-赖氨酸、生物胞素酰肼、2-亚氨基生物素和生物素基-E-氨基己酸-N-羟基琥珀酰亚胺酯的氨基或巯基衍生物、磺基琥珀酰亚胺亚氨基生物素、生物素溴乙酰基酰肼、对重氮苯甲酰基生物胞素、3-(N-马来酰亚胺基丙酰基)生物胞素、脱硫生物素等。术语“生物素”还包括可以与Rhizavidin、亲和素、链霉亲和素、tamavidin部分或其他亲和素样肽中的一种或多种特异性结合的生物素变体。

[0435] 如本文所用的,“PPV确定方法”可指呈递PPV确定方法。例如,“PPV确定方法”可指包括以下步骤的方法:(a)使用HLA肽呈递预测模型,如机器学习HLA肽呈递预测模型,处理多个测试肽序列的氨基酸信息,以生成多个测试呈递预测,每个测试呈递预测指示由细胞的II类HLA等位基因(如受试者细胞的II类HLA等位基因)编码的一种或多种蛋白质能够呈递所述多个测试肽序列的给定测试肽序列的可能性,其中所述多个测试肽序列包含至少500个测试肽序列,所述测试肽序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少499个包含在由生物体(例如与受试者属于同一物种的生物体)基因组编码的蛋白质内的诱饵肽序列,其中所述多个测试肽序列中命中肽序列的数目与诱饵肽序列的数目之比小于1,例如所述至少一个命中肽序列与所述至少499个诱饵肽序列之比为1:499;(b)将排名靠前的百分比的所述多个测试肽序列,如前0.2%的所述多个测试肽序列,鉴定或判定为由细胞的II类HLA等位基因呈递;(c)计算所述HLA肽呈递预测模型的PPV,其中PPV是所述多个测试肽序列中被鉴定或判定为由细胞的II类HLA等位基因所呈递的测试肽序列的分数,这些肽是通过质谱法观察到由细胞的II类HLA等位基因所呈递的肽。在一些实施方案中,诱饵肽具有相同长度,即包含与命中肽相同数目的氨基酸。在一些实施方案中,与命中肽相比,诱饵肽可包含多一个或少一个氨基酸。在一些实施方案中,诱饵肽是作为内源性肽的肽。在一些实施方案中,诱饵肽是合成肽。在一些实施方案中,诱饵肽是已通过质谱法被鉴定为与第一MHC I类或II类蛋白结合的内源性肽,其中第一MHC I类或II类蛋白不同于与命中肽结合的第二MHC I类或II类蛋白。在一些实施方案中,诱饵肽可以是乱序肽,例如,诱饵肽可以包含这样的氨基酸序列,其中在肽的长度内,相对于命中肽的氨基酸位置,其氨基酸位置重排。在一些实施方案中,PPV确定方法可以是呈递PPV确定方法。在一些实施方案中,命中肽序列的数目与诱饵肽序列的数目之比为约1:10、1:20、1:50、1:100、1:250、1:500、1:1000、1:1500、1:2000、1:2500、1:5000、1:7500、1:10000、1:25000、1:50000或1:100000。在一些实施方案中,所述至少一个命中肽序列包含至少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或100个命中肽序列。在一些实施方案中,所述至少499个诱饵肽序列包含至少500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、

7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个诱饵肽序列。在一些实施方案中,所述至少500个测试肽序列包含至少600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个测试肽序列。在一些实施方案中,将排名靠前的百分比的所述多个测试肽序列鉴定或判定为由细胞的II类HLA等位基因呈递包括将前0.20%、0.30%、0.40%、0.50%、0.60%、0.70%、0.80%、0.90%、1.00%、1.10%、1.20%、1.30%、1.40%、1.50%、1.60%、1.70%、1.80%、1.90%、2.00%、2.10%、2.20%、2.30%、2.40%、2.50%、2.60%、2.70%、2.80%、2.90%、3.00%、3.10%、3.20%、3.30%、3.40%、3.50%、3.60%、3.70%、3.80%、3.90%、4.00%、4.10%、4.20%、4.30%、4.40%、4.50%、4.60%、4.70%、4.80%、4.90%、5.00%、5.10%、5.20%、5.30%、5.40%、5.50%、5.60%、5.70%、5.80%、5.90%、6.00%、6.10%、6.20%、6.30%、6.40%、6.50%、6.60%、6.70%、6.80%、6.90%、7.00%、7.10%、7.20%、7.30%、7.40%、7.50%、7.60%、7.70%、7.80%、7.90%、8.00%、8.10%、8.20%、8.30%、8.40%、8.50%、8.60%、8.70%、8.80%、8.90%、9.00%、9.10%、9.20%、9.30%、9.40%、9.50%、9.60%、9.70%、9.80%、9.90%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%鉴定或判定为由细胞的II类HLA等位基因呈递。在一些实施方案中,该细胞是单等位基因细胞。

[0436] 如本文所用的,“PPV确定方法”可指结合PPV确定方法。例如,“PPV确定方法”可指包括以下步骤的方法:(a)使用HLA肽结合预测模型,如机器学习HLA肽结合预测模型,处理多个测试肽序列的氨基酸信息,以生成多个结合呈递预测,每个结合呈递预测指示由细胞

的II类HLA等位基因(如受试者细胞的II类HLA等位基因)编码的一种或多种蛋白质与所述多个测试肽序列的给定测试肽序列结合的可能性,其中所述多个测试肽序列包含至少20个测试肽序列,所述测试肽序列包含(i)至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的命中肽序列,和(ii)至少19个包含在蛋白质内的诱饵肽序列,所述蛋白质包含至少一个通过质谱法被鉴定为将由细胞中表达的HLA蛋白所呈递的肽序列,其中所述多个测试肽序列中命中肽序列的数目与诱饵肽序列的数目之比小于1,例如所述至少一个命中肽序列与所述至少19个诱饵肽序列之比为1:19;(b)将排名靠前的百分比的所述多个测试肽序列,如前5%的所述多个测试肽序列,鉴定或判定为与所述HLA蛋白结合;(c)计算所述HLA肽结合预测模型的PPV,其中PPV是所述多个测试肽序列中被鉴定或判定为与细胞的II类HLA等位基因结合的测试肽序列的分数,这些肽是通过质谱法观察到由细胞的II类HLA等位基因所呈递的肽。在一些实施方案中,命中肽序列的数目与诱饵肽序列的数目之比为约1:2、1:3、1:4、1:5、1:10、1:20、1:25、1:30、1:40、1:50、1:75、1:100、1:200、1:250、1:500或1:1000。在一些实施方案中,所述至少一个命中肽序列包含至少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或100个命中肽序列。在一些实施方案中,所述至少19个诱饵肽序列包含至少30、40、50、60、70、80、90、100、110、120、130、140、150、160、170、180、190、200、210、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个诱饵肽序列。在一些实施方案中,所述至少20个测试肽序列包含至少30、40、50、60、70、80、90、100、110、120、130、140、150、160、170、180、190、200、210、220、230、240、250、260、270、280、290、300、310、320、330、340、350、360、370、380、390、400、410、420、430、440、450、460、470、480、490、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900、2000、2100、2200、2300、2400、2500、2600、2700、2800、2900、3000、3100、3200、3300、3400、3500、3600、3700、3800、3900、4000、4100、4200、4300、4400、4500、4600、

4700、4800、4900、5000、5100、5200、5300、5400、5500、5600、5700、5800、5900、6000、6100、6200、6300、6400、6500、6600、6700、6800、6900、7000、7100、7200、7300、7400、7500、7600、7700、7800、7900、8000、8100、8200、8300、8400、8500、8600、8700、8800、8900、9000、9100、9200、9300、9400、9500、9600、9700、9800、9900、10000、11000、12000、13000、14000、15000、16000、17000、18000、19000、20000、21000、22000、23000、24000、25000、26000、27000、28000、29000、30000、31000、32000、33000、34000、35000、36000、37000、38000、39000、40000、41000、42000、43000、44000、45000、46000、47000、48000、49000、50000、52500、55000、57500、60000、62500、65000、67500、70000、72500、75000、77500、80000、82500、85000、87500、90000、92500、95000、97500、100000、125000、150000、175000、200000、225000、250000、275000、300000、325000、350000、375000、400000、425000、450000、475000、500000、600000、700000、800000、900000或1000000个测试肽序列。在一些实施方案中,将排名靠前的百分比的所述多个测试肽序列鉴定或判定为由细胞的II类HLA等位基因呈递包括将前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%或40%鉴定或判定为由细胞的II类HLA等位基因呈递。在一些实施方案中,该细胞是单等位基因细胞。

[0437] 人类白细胞抗原(HLA)系统

[0438] 免疫系统可以被分类为两个功能子系统:先天性免疫系统和适应性免疫系统。先天性免疫系统是抵抗感染的第一道防线,大多数潜在病原体会在引起例如明显感染之前被该系统快速中和。适应性免疫系统对入侵生物体的分子结构(称为抗原)起反应。与先天性免疫系统不同,适应性免疫系统对病原体是高度特异性的。适应性免疫还可以提供持久的保护;例如,从麻疹中康复的人现在终生免受麻疹的侵害。有两种类型的适应性免疫反应,包括体液免疫反应和细胞介导的免疫反应。在体液免疫反应中,由B细胞分泌到体液中的抗体与病原体衍生的抗原结合,导致病原体通过多种机制被消除,例如补体介导的裂解。在细胞介导的免疫反应中,能够破坏其他细胞的T细胞被激活。例如,如果与疾病相关的蛋白质存在于细胞中,它们会在细胞内被蛋白水解破碎成肽。然后,特定的细胞蛋白质自身附着于以这种方式形成的抗原或肽,并将其转运至细胞表面,在此呈递给体内T细胞中的分子防御机制。细胞毒性T细胞识别这些抗原并杀死带有该抗原的细胞。

[0439] 术语“主要组织相容性复合物(MHC)”、“MHC分子”或“MHC蛋白”是指这样的蛋白质,其能够结合由蛋白质抗原的蛋白水解切割产生并代表潜在的T细胞表位的肽,将其转运至细胞表面,并将所述肽呈递给特定细胞,例如在细胞毒性T淋巴细胞或T辅助细胞中。人类MHC也被称为HLA复合物。因此,术语“人类白细胞抗原(HLA)系统”、“HLA分子”或“HLA蛋白”是指编码人类MHC蛋白的基因复合物。术语MHC在鼠类物种中被称为”H-2”复合物。本领域普通技术人员将会认识到,术语“主要组织相容性复合物(MHC)”、“MHC分子”、“MHC蛋白”和“人类白细胞抗原(HLA)系统”、“HLA分子”、“HLA蛋白”在本文中可互换使用。

[0440] HLA蛋白被分类为两种类型,被称为HLA I类和HLA II类。两种HLA类型的蛋白质的结构非常相似;但是,它们具有非常不同的功能。HLA I类蛋白存在于身体几乎所有细胞的表面,包括大多数肿瘤细胞。HLA I类蛋白负载有抗原,这些抗原通常起源于内源性蛋白质或细胞内存在的病原体,然后被呈递给幼稚或细胞毒性T淋巴细胞(CTL)。HLA II类蛋白存

在于抗原呈递细胞(APC)上,包括但不限于树突细胞、B细胞和巨噬细胞。它们主要向辅助T细胞呈递从外部抗原来源,例如细胞外部加工的肽。被HLA I类蛋白结合的大多数肽都起源于在生物体自身的健康宿主细胞中产生的胞质蛋白质,并且通常不会刺激免疫反应。

[0441] HLA I类分子(图1)由两条非共价连接的多肽链组成:一条HLA编码的α链(重链,44至47kD)和一个非HLA编码的亚单位,被称为β2微球蛋白(或β2m),(12kD)。α链具有三个胞外域——α1、α2和α3以及跨膜区,其中α1和α2区能够结合约7至13个氨基酸(例如,约8至11个氨基酸,或9或10个氨基酸)的肽。HLA 1类分子与具有合适的结合基序的肽结合,并将其呈递给细胞毒性T淋巴细胞。HLA 1类重链可以是HLA-A等位基因的蛋白质产物,也称为HLA-A单体,或HLA-B等位基因的蛋白质产物(同样,HLA-B单体)或HLA-C等位基因的蛋白质产物(HLA-C单体),其中每一个都与β-2-微球蛋白复合。α1依赖于非HLA蛋白β2m;β2m由位于人类15号染色体上的β-2-微球蛋白基因编码。α3结构域连接至跨膜区,将HLA I类分子锚定至细胞膜。所呈递的肽被肽结合沟的底部保持在α1/α2异二聚体(由两个不同的亚单位组成的分子)的中心区域。HLA I-A类、HLA I-B类或HLA I-C类是高度多态性的。HLA 1-A类基因(称为HLA-A基因)、HLA 1-B类基因(称为HLA-B基因)和HLA 1-C类基因(称为HLA-C基因)各包含8个外显子,外显子1编码前导肽,外显子2和3编码α1和α2结构域,外显子5编码跨膜区,外显子6和7编码胞质尾。外显子2和外显子3的多态性决定了每个1类分子的肽结合特异性。HLA I-B类基因(HLA-B)有许多可能的变异、表达模式和呈递的抗原。该组被再分为在HLA基因座内编码的组,例如HLA-E、HLA-F、HLA-G,以及不在其中编码的那些,例如应激配体,如ULBP、Rae1和H60。这些分子中许多分子的抗原/配体仍然未知,但它们可以与CD8+ T细胞、NKT细胞和NK细胞相互作用。

[0442] 在一些实施方案中,本公开利用非经典的HLA I-E类等位基因。HLA-E分子被自然杀伤(NK)细胞和CD8+ T细胞识别。HLA-E在几乎所有组织中表达,包括肺、肝、皮肤和胎盘细胞。在实体瘤(例如,骨肉瘤和黑素瘤)中也检测到HLA-E表达。HLA-E分子与CD8+ T细胞上表达的TCR结合,从而导致T细胞活化。还已知HLA-E结合在NK细胞和CD8+ T细胞上表达的CD94/NKG2受体。CD94可以与NKG2的几种不同的同种型配对,以形成具有抑制(NKG2A、NKG2B)或促进(NKG2C)细胞活化的潜力的受体。HLA-E可以结合由大多数HLA-A、-B、-C和-G分子前导序列的氨基酸残基3-11衍生的肽,但不能结合其自身的前导肽。还已经证明HLA-E呈递由类似于HLA-A、-B和-C等位基因的内源性蛋白质衍生的肽。在生理条件下,CD94/NKG2A与负载有来自HLA I类前导序列的肽的HLA-E的接合通常会诱导抑制信号。巨细胞病毒(CMV)通过表达UL40糖蛋白(模拟HLA-A前导序列)利用逃避NK细胞免疫监视的机制。然而,也报道了CD8+ T细胞可以识别负载有来源于CMV Toledo株的UL40肽的HLA-E,并在防御CMV中起作用。大量研究揭示了HLA-E在感染性疾病和癌症中的几个重要功能。

[0443] 肽抗原在呈递于细胞表面上之前,通过内质网内的竞争性亲和力结合将自身附接到HLA I类分子上。在此,单个肽抗原的亲和力与其氨基酸序列以及在氨基酸序列内限定位置上特异性结合基序的存在直接相关。如果这样的肽的序列是已知的,则可以使用例如肽疫苗来操纵免疫系统对抗病变细胞。

[0444] MHC分子是高度多态性的,即,存在许多MHC变体。每个变体由编码蛋白质的基因的变体编码,并且每个这样的变体基因被称为等位基因。对于人类,MHC被称为人类白细胞抗原(HLA),它涉及三种类型的HLA II类分子:DP、DQ和DR。HLA II类肽(图1)具有两条链:α和

β,它们各自具有两个结构域——α1和α2以及β1和β2——每条链分别具有跨膜结构域:α2和β2,将HLA II类分子锚定至细胞膜。肽结合沟由α1和β1的异二聚体形成。研究最广泛的HLA-DR分子具有DRA和DRB,分别对应于α和β结构域。DRB是多样的,DRA几乎相同。因此,DRB等位基因的结合特异性指示相应HLA-DR的结合特异性。每个MHC蛋白都有其自身的结合特异性,这意味着与MHC分子结合的一组肽可能不同于与另一MHC分子结合的肽。经典分子将肽呈递给CD4+淋巴细胞。具有细胞内功能的非经典分子,附件,不在细胞膜上暴露,而是暴露在溶酶体的内膜中,通常将抗原肽加载到经典HLA II类分子上。

[0445] 在HLA II类系统中,吞噬细胞如巨噬细胞和未成熟的树突细胞通过向吞噬体中的吞噬作用摄取实体——尽管B细胞表现出更普遍的向内体中的内吞作用——吞噬体与溶酶体融合,溶酶体的酸性酶将摄取的蛋白质裂解成许多不同的肽。自噬是HLA II类肽的另一个来源。通过与宿主携带的HLA II类变体(在宿主基因组中编码)的分子相互作用的物理化学动力学,特定的肽表现出免疫显性并负载在HLA II类分子上。它们被运输到细胞表面并在细胞表面外化。研究最多的HLA II类基因的亚类是:HLA-DPA1、HLA-DPB1、HLA-DQA1、HLA-DQB1、HLA-DRA和HLA-DRB1。

[0446] HLA II类分子向CD4+辅助T细胞呈递肽是对外来抗原的免疫应答所必需的(Roche和Furuta,2015)。一旦被激活,CD4+ T细胞就会促进B细胞分化和抗体产生,以及CD8+ T细胞(CTL)应答。CD4+ T细胞还分泌激活并诱导其他免疫细胞分化的细胞因子和趋化因子。HLA II类分子是α和β链的异二聚体,α和β链相互作用以形成比HLA I类肽结合沟更开放的肽结合沟(Unanue等人,2016)。与HLA II类分子结合的肽被认为具有9个氨基酸的结合核心,该结合核心在N端或C端侧具有从结合沟突出的侧翼残基(Jardetzky等人,1996;Stern等人,1994)。这些肽的长度通常为12-16个氨基酸,并且通常在结合部分的P1、P4、P6/7和P9位置处含有3-4个锚残基(Rossjohn等人,2015)。

[0447] HLA等位基因以共显性方式表达,这意味着从父母双方继承的等位基因(变体)同等地表达。例如,每个人携带3个I类基因(HLA-A、HLA-B和HLA-C)中每个基因的2个等位基因,因此可以表达六种不同类型的HLA II类。在HLA II类基因座中,每个人继承一对HLA-DP基因(DPA1和DPB1,编码α和β链)、HLA-DQ(对于α和β链为DQA1和DQB1)、一个HLA-DRα基因(DRA1)和一个或多个HLA-DRβ基因(DRB1和DRB3、-4或-5)。例如,HLA-DRB1具有超过近400个已知的等位基因。这意味着一个杂合个体可以继承六个或八个功能性HLA II类等位基因,每个亲本三个或更多个。因此,HLA基因是高度多态性的;群体内的不同个体中存在许多不同的等位基因。编码HLA蛋白的基因具有许多可能的变异,从而使每个人的免疫系统能够对众多外来入侵物发生反应。一些HLA基因具有数百个已鉴定的形式(等位基因),给予每个形式特定的编号。在一些实施方案中,HLA I类等位基因是HLA-A*02:01、HLA-B*14:02、HLA-A*23:01、HLA-E*01:01(非经典的)。在一些实施方案中,HLA II类等位基因是HLA-DRB*01:01、HLA-DRB*01:02、HLA-DRB*11:01、HLA-DRB*15:01和HLA-DRB*07:01。

[0448] 受试者的受试者特异性HLA等位基因或HLA基因型可以通过本领域已知的任何方法来确定。在示例性实施方案中,通过在通过引用整体并入本文的国际专利申请PCT/US2014/068746(2015年6月11日作为WO2015085147公开)中描述的任何方法确定HLA基因型。简言之,所述方法包括确定多态性基因类型,其可以包括产生从测序数据集提取的读取与包含多态性基因的等位基因变体的基因参考集的比对,在比对中确定每个等位基因变体

的第一后验概率或由后验概率得出的评分,将具有最大第一后验概率或由后验概率得出的评分的等位基因变体鉴定为第一等位基因变体,鉴定与第一等位基因变体和一个或多个其他等位基因变体比对的一个或多个重叠读取,使用权重因子对所述一个或多个其他等位基因变体确定第二后验概率或由后验概率得出的评分,通过选择具有最大第二后验概率或由后验概率得出的评分的等位基因变体来确定第二等位基因变体,第一和第二等位基因变体定义了多态性基因的基因类型,并提供第一和第二等位基因变体的输出。

[0449] 在一些实施方案中,本文描述的MHC II类肽:抗原肽结合和呈递预测方法具有从由单独HLA等位基因编码的大组库MHC II类肽预测结合物的能力。在一些实施方案中,MAPTAC技术用质谱法验证的HLA匹配肽的大型数据库进行训练。在一些实施方案中,质谱法验证的HLA匹配肽的大型数据库包含多于1.2x10^6个这样的HLA匹配肽。在一些实施方案中,质谱法验证的HLA匹配肽的大型数据库覆盖超过150个HLA等位基因,包括MHC I类和II类等位基因亚型。在一些实施方案中,该数据库对于HLA-I和HLA-II(DR亚型)覆盖至少95%的美国人口。

[0450] 如本文所述,在动物和人类中有大量证据表明突变的表位可有效诱导免疫应答,并且自发性肿瘤消退或长期存活的情况与CD8+ T细胞对突变表位的应答相关,并且“免疫编辑”可以追踪小鼠和人类显性突变抗原表达的变化。

[0451] 测序技术揭示,每个肿瘤含有多个患者特异性突变,这些突变改变基因的蛋白质编码内容。此类突变产生改变的蛋白质,范围从单氨基酸改变(由错义突变引起)到由于移码、终止密码子的通读或内含子区域的翻译(新的开放阅读框突变;neoORF)而增加新氨基酸序列的长区域。这些突变蛋白质是宿主对肿瘤的免疫应答的有价值的靶标,因为与天然蛋白质不同,它们不受自身耐受性的免疫抑制作用的影响。因此,与患者的正常细胞相比,突变的蛋白质更可能具有免疫原性,并且对肿瘤细胞也更具特异性。本质上,含有癌症相关突变的短肽(8-24个氨基酸长)是癌症免疫疗法的候选物。

[0452] 在一些实施方案中,驱动预测方法的算法可进一步用于对肽进行突变判定。在一些实施方案中,该预测方法可用于确定驱动突变状态,和/或RNA表达状态,和/或肽内的切割预测。

[0453] 术语“T细胞”包括CD4+ T细胞和CD8+ T细胞。术语T细胞还包括T辅助1型T细胞和T辅助2型T细胞。本文所用的T细胞通常按功能和也有助于T细胞受体与抗原结合的细胞表面抗原(簇分化抗原或CD)分类为两大类:辅助性T(TH)细胞和细胞毒性T淋巴细胞(CTL)。

[0454] 成熟的辅助性T(TH)细胞表达表面蛋白质CD4,并且被称为CD4+ T细胞。在T细胞发育后,成熟的幼稚T细胞离开胸腺并开始在全身扩散,包括淋巴结。幼稚T细胞是从未暴露于它们被编程以对其发生应答的抗原的T细胞。像所有T细胞一样,它们表达T细胞受体-CD3复合物。T细胞受体(TCR)由恒定区和可变区组成。可变区决定了T细胞可以对什么抗原发生应答。CD4+ T细胞具有对MHC II类蛋白具有亲和力的TCR,CD4参与确定胸腺成熟过程中的MHC亲和力。MHC II类蛋白通常仅可见于特化抗原呈递细胞(APC)的表面上。特化抗原呈递细胞(APC)主要是树突细胞、巨噬细胞和B细胞,尽管树突细胞是组成型(始终)表达MHC II类的唯一一组细胞。一些APC也将天然(或未加工的)抗原结合到它们的表面,如滤泡树突细胞,但是未加工的抗原不与T细胞相互作用,也不参与其活化。与HLA I类蛋白结合的肽抗原通常比与HLA II类蛋白结合的肽抗原短。

[0455] 细胞毒性T淋巴细胞(CTL),也被称为细胞毒性T细胞、溶细胞性T细胞、CD8+ T细胞或杀伤性T细胞,是指在所靶向的细胞中诱导凋亡的淋巴细胞。CTL通过TCR与靶细胞表面上加工的抗原(Ag)的相互作用,与靶细胞形成抗原特异性偶联物,从而导致靶细胞凋亡。凋亡小体被巨噬细胞消除。术语"CTL应答"用于指由CTL细胞介导的原发性免疫应答。细胞毒性T淋巴细胞在其表面上具有T细胞受体(TCR)和CD8分子两者。T细胞受体能够识别并结合与HLA I类分子复合的肽。每个细胞毒性T淋巴细胞表达独特的T细胞受体,该T细胞受体能够结合特定的MHC/肽复合物。大多数细胞毒性T细胞表达可以识别特定抗原的T细胞受体(TCR)。为了使TCR与HLA I类分子结合,前者必须伴有被称为CD8的糖蛋白,CD8与HLA I类分子的恒定部分结合。因此,这些T细胞被称为CD8+ T细胞。CD8与MHC分子之间的亲和力使T细胞和靶细胞在抗原特异性激活过程中紧密结合在一起。CD8+ T细胞一旦被激活,就被识别为T细胞,并且通常被分类为在免疫系统中具有预定的细胞毒性作用。然而,CD8+ T细胞也具有产生某些细胞因子的能力。

[0456] "T细胞受体(TCR)"是参与T细胞响应于抗原呈递而活化的细胞表面受体。TCR通常由α和β两条链构成,这两条链装配形成异二聚体,并与CD3转导亚单位缔合以形成存在于细胞表面上的T细胞受体复合物。TCR的每条α和β链由免疫球蛋白样N-末端可变(V)和恒定(C)区、疏水性跨膜结构域和短胞质区组成。至于免疫球蛋白分子,α和β链的可变区是通过V(D)J重组产生的,从而在T细胞群体中产生极大多样性的抗原特异性。然而,与识别完整抗原的免疫球蛋白相比,T细胞被与MHC分子缔合的加工的肽片段激活,从而为T细胞对抗原的识别引入了额外的维度,被称为MHC限制。通过T细胞受体识别供体与受体之间的MHC差异会导致T细胞增殖和GVHD的潜在发展。已经表明,TCR的正常表面表达依赖于复合物所有七种组分的协调合成和装配(Ashwell和Klusner 1990)。TCRα或TCRβ的失活可以导致TCR从T细胞表面消除,从而阻止同种抗原的识别,因而防止GVHD。然而,TCR破坏通常导致CD3信号传导成分的消除,并改变进一步T细胞扩充的方式。

[0457] 术语"HLA肽组"是指与特定HLA类别特异性相互作用的一组肽,并且可以包含数千种不同的序列。HLA肽组包括多样化的肽,它们来源于在细胞中表达的正常和异常蛋白质。因此,可以研究HLA肽组以鉴定癌症特异性肽,以供开发肿瘤免疫疗法,并作为关于癌细胞内蛋白质合成和降解方案的信息来源。在一些实施方案中,HLA肽组是一组可溶性HLA肽(sHLA)。在一些实施方案中,HLA肽组是一组膜结合的HLA(mHLA)。

[0458] "抗原呈递细胞"或"APC"包括专职抗原呈递细胞(例如,B淋巴细胞、巨噬细胞、单核细胞、树突细胞、朗格汉斯细胞),以及其他抗原呈递细胞(例如,角质形成细胞、内皮细胞、星形胶质细胞、成纤维细胞、少突细胞、胸腺上皮细胞、甲状腺上皮细胞、神经胶质细胞(脑)、胰腺β细胞和血管内皮细胞)。"抗原呈递细胞"或"APC"是表达主要组织相容性复合物(MHC)分子并可以在其表面上展示与MHC复合的外来抗原的细胞。

[0459] 单等位基因HLA细胞系

[0460] 可以通过用编码单个HLA等位基因的多核酸(例如载体)转导或转染合适的细胞群体来生成表达单个HLA I类等位基因、一对HLA II类等位基因或单个HLA I类等位基因和一对HLA II类等位基因的单等位基因细胞系(图2)。合适的细胞群体包括,例如,其中外源表达单个HLA I类等位基因的HLA I类缺陷细胞系,其中外源表达一对HLA II类等位基因的HLA II类缺陷细胞系,或其中外源表达单个HLA I类和/或一对II类等位基因的I类和I类II

型缺陷细胞系。作为示例性实施方案,HLA I类缺陷B细胞系是B721.221。然而,对于技术人员而言清楚的是,可以产生HLA I类和/或HLA II类缺陷的其他细胞群体。用于删除/灭活内源HLA I类或HLA II类基因的示例性方法包括CRISPR-Cas9介导的基因组编辑,例如在THP-1细胞中。在一些实施方案中,所述细胞群体是专职抗原呈递细胞,如巨噬细胞、B细胞和树突细胞。所述细胞可以是B细胞或树突细胞。在一些实施方案中,所述细胞是肿瘤细胞或来自肿瘤细胞系的细胞。在一些实施方案中,所述细胞是从患者中分离的。在一些实施方案中,所述细胞含有传染原或其一部分。在一些实施方案中,所述细胞群体包含至少10^7个细胞。在一些实施方案中,所述细胞群体被进一步修饰,例如通过增加或减少至少一个基因的表达和/或活性。在一些实施方案中,该基因编码免疫蛋白酶体的成员。已知免疫蛋白酶体参与HLA I类结合肽的加工,并且包括LMP2(β1i)、MECL-1(β2i)和LMP7(β5i)亚单位。免疫蛋白酶体也可以被干扰素-γ诱导。因此,在一些实施方案中,所述细胞群体可以与一种或多种细胞因子、生长因子或其他蛋白质接触。可以用诸如干扰素-γ、IL-10、IL-6和/或TNF-α等炎性细胞因子刺激所述细胞。所述细胞群体也可以经受各种环境条件,如应激(热应激、缺氧、葡萄糖饥饿、DNA破坏剂等)。在一些实施方案中,使细胞与化疗药物、放射疗法、靶向疗法或免疫疗法中的一种或多种接触。因此,本文公开的方法可以用来研究各种基因或条件对HLA肽加工和呈递的影响。在一些实施方案中,选择所使用的条件以匹配待鉴定其HLA-肽群体的患者的状况。

[0461] 可以使用基于病毒的系统(例如,腺病毒系统、腺相关病毒(AAV)载体、痘病毒或慢病毒)来编码并表达本公开的单个HLA等位基因。先前已经描述了可用于腺相关病毒、腺病毒和慢病毒递送的质粒(参见,例如,美国专利6,955,808和6,943,019,以及美国专利申请20080254008,在此引入作为参考)。在可以在本公开的实践中使用的载体中,利用逆转录病毒基因转移方法可以实现向细胞的宿主基因组中的整合,通常导致所插入的转基因的长期表达。在一个示例性实施方案中,该逆转录病毒是慢病毒。另外,已经在许多不同的细胞类型和靶组织中观察到高转导效率。逆转录病毒的向性可以通过并入外来包膜蛋白、扩充靶细胞的潜在目标群体来改变。逆转录病毒也可以被工程改造以允许所插入的转基因的条件表达,使得仅有某些细胞类型被慢病毒感染。细胞类型特异性启动子可用来靶向在特定细胞类型中的表达。慢病毒载体是逆转录病毒载体(因此慢病毒和逆转录病毒载体均可以在本公开的实践中使用)。此外,慢病毒载体能够转导或感染非分裂细胞并且通常产生高病毒滴度。

[0462] 逆转录病毒基因转移系统的选择可以取决于靶组织。逆转录病毒载体由顺式作用长末端重复序列组成,其包装能力可达6-10kb的外来序列。最小顺式作用LTR足以复制和包装载体,然后用其将所需核酸整合到靶细胞中以提供永久表达。可以在本公开的实践中使用的广泛使用的逆转录病毒载体包括基于鼠白血病病毒(MuLV)、长臂猿白血病病毒(GaLV)、猿猴免疫缺陷病毒(SIV)、人类免疫缺陷病毒(HIV)及其组合的载体(参见,例如,Buchscher等人(1992)J.Virol.66:2731-2739;Johann等人(1992)J.Virol.66:1635-1640;Sommnerfelt等人(1990)Virol.176:58-59;Wilson等人(1998)J.Virol.63:2374-2378;Miller等人(1991)J.Virol.65:2220-2224;PCT/US94/05700)。另外,在本公开的实践中有用的是最小的非灵长类动物慢病毒载体,如基于马传染性贫血病毒(EIAV)的慢病毒载体(参见,例如,Balagaan,(2006)J Gene Med;8:275-285,于2005年11月21日在线发表于

Wiley InterScience DOI:10.1002/jgm.845)。所述载体可以具有驱动靶基因表达的巨细胞病毒(CMV)启动子。因此,本公开涉及可用于实施本公开内容的一种或多种载体:病毒载体,包括逆转录病毒载体和慢病毒载体。

[0463] 任何HLA等位基因均可以在细胞群体中表达。在一个示例性实施方案中,该HLA等位基因是HLA I类等位基因。在一些实施方案中,该HLA I类等位基因是HLA-A等位基因或HLA-B等位基因。在一些实施方案中,该HLA等位基因是HLA II类等位基因。HLA I类和II类等位基因的序列可见于IPD-IMGT/HLA数据库中。示例性的HLA等位基因包括但不限于HLA-A*02:01、HLA-B*14:02、HLA-A*23:01、HLA-E*01:01、HLA-DRB*01:01、HLA-DRB*01:02、HLA-DRB*11:01、HLA-DRB*15:01和HLA-DRB*07:01。

[0464] 在一些实施方案中,选择HLA等位基因以对应于目的基因型。在一些实施方案中,该HLA等位基因是突变的HLA等位基因,其可以是患病患者中非天然存在的等位基因或天然存在的等位基因。本文公开的方法具有为与各种疾病相关的HLA等位基因以及以低频率存在的等位基因鉴定HLA结合肽的进一步的优点。因此,在一些实施方案中,本文提供的方法可以鉴定HLA等位基因,即使其以低于1%的频率存在于群体如高加索人群体中。

[0465] 在一些实施方案中,编码HLA等位基因的核酸序列进一步包含可用来免疫纯化HLA蛋白的亲和受体标签。合适的标签是本领域公知的。在一些实施方案中,亲和受体标签是聚组氨酸标签、聚组氨酸-甘氨酸标签、聚精氨酸标签、聚天冬氨酸标签、聚半胱氨酸标签、聚苯丙氨酸、c-myc标签、单纯疱疹病毒糖蛋白D(gD)标签、FLAG标签、KT3表位标签、微管蛋白表位标签、T7基因10蛋白肽标签、链霉亲和素标签、链霉亲和素结合肽(SPB)标签、Strep-标签、Strep-标签II、白蛋白结合蛋白(ABP)标签、碱性磷酸酶(AP)标签、蓝舌病毒标签(B-tag)、钙调蛋白结合肽(CBP)标签、氯霉素乙酰转移酶(CAT)标签、胆碱结合域(CBD)标签、壳多糖结合域(CBD)标签、纤维素结合域(CBP)标签、二氢叶酸还原酶(DHFR)标签、半乳糖结合蛋白(GBP)标签、麦芽糖结合蛋白(MBP)、谷胱甘肽-S-转移酶(GST)、Glu-Glu(EE)标签、人类流感血凝素(HA)标签、辣根过氧化物酶(HRP)标签、NE-标签、HSV标签、酮类固醇异构酶(KSI)标签、KT3标签、LacZ标签、萤光素酶标签、NusA标签、PDZ结构域标签、AviTag、钙调蛋白标签、E-标签、S-标签、SBP-标签、Softag 1、Softag 3、TC标签、VSV-标签、Xpress标签、Isopeptag、SpyTag、SnoopTag、Profinity eXact标签、蛋白C标签、S1-标签、S-标签、生物素-羧基载体蛋白(BCCP)标签、绿色荧光蛋白(GFP)标签、小泛素样修饰物(SUMO)标签、串联亲和纯化(TAP)标签、HaloTag、Nus-标签、硫氧还蛋白标签、Fc-标签、CYD标签、HPC标签、TrpE标签、泛素标签、衍生自水疱性口炎病毒糖蛋白的VSV-G表位标签,或衍生自猿猴病毒5(SV5)副粘病毒P和V蛋白上发现的小表位(Pk)的V5标签。在一些实施方案中,所述亲和受体标签是“表位标签”,这是一种类型的肽标签,其向HLA蛋白添加可识别的表位(抗体结合位点)以提供相应抗体的结合,从而允许鉴定或亲和纯化所标记的蛋白质。表位标签的非限制性实例是可与IgG结合的蛋白A或蛋白G。在一些实施方案中,亲和受体标签包括生物素受体肽(BAP)或人类流感血凝素(HA)肽序列。许多其他标签部分是普通技术人员已知的并且可以设想的,并且在此想到。可以使用任何肽标签,只要它能够被表达为亲和受体标记的HLA-肽复合物的元件即可。

[0466] 本文提供的方法包括从用HLA构建体的亲和下拉转染或转导的细胞中分离HLA-肽复合物(图3)。在一些实施方案中,可以使用本领域已知的标准免疫沉淀技术与市售抗体分

离复合物。可以先裂解细胞。可以使用HLA I类特异性抗体如W6/32抗体分离HLA I类-肽复合物,而使用HLA II类特异性抗体如M5/114.15.2单克隆抗体分离HLA II类-肽复合物。在一些实施方案中,单个(或一对)HLA等位基因被表达为具有肽标签的融合蛋白,并且使用识别该肽标签的结合分子分离HLA-肽复合物。

[0467] 所述方法进一步包括从所述HLA-肽复合物中分离肽并对该肽进行测序。通过本领域技术人员已知的任何方法,如酸洗脱,从复合物中分离肽。尽管可以使用任何测序方法,但是在一些实施方案中,采用使用质谱法的方法,如液相色谱-质谱法(LC-MS或LC-MS/MS,或者HPLC-MS或HPLC-MS/MS)。这些测序方法是技术人员公知的,并且在Medzihradszky KF和Chalkley RJ.Mass Spectrom Rev.2015年1月-2月;34(1):43-63中进行了综述。

[0468] 在一些实施方案中,所述细胞群体表达一个或多个内源HLA等位基因。在一些实施方案中,所述细胞群体是缺乏一个或多个内源HLA I类等位基因的工程化细胞群体。在一些实施方案中,所述细胞群体是缺乏内源HLA I类等位基因的工程化细胞群体。在一些实施方案中,所述细胞群体是缺乏一个或多个内源HLA II类等位基因的工程化细胞群体。在一些实施方案中,所述细胞群体是缺乏内源HLA II类等位基因的工程化细胞群体或缺乏内源HLA I类等位基因和内源HLA II类等位基因的工程化细胞群体。在一些实施方案中,所述细胞群体包含已经富集或分选的细胞,例如通过荧光激活细胞分选(FACS)。在一些实施方案中,使用荧光激活的细胞分选(FACS)来分选细胞群体。在一些实施方案中,预先针对HLA I类或II类或者HLA I类和II类两者的细胞表面表达对细胞群体进行FACS分选。例如,可以使用FACS针对HLA I类等位基因、HLA II类等位基因或其组合的细胞表面表达分选细胞群体。

[0469] 制备个性化癌症疫苗的方法

[0470] 一旦鉴定出针对癌症的特异性突变,使得该突变存在于癌细胞的DNA中,但不存在于同一人类受试者的正常细胞中,并且该突变导致该DNA所编码蛋白质中的一个或多个氨基酸发生变化,则该突变可以是宿主免疫应答的靶标。天然免疫应答可以针对突变的蛋白质,从而导致表达该蛋白质的癌细胞遭到破坏。由于癌组织中存在天然的耐受性应答和免疫功能低下的环境,因此免疫疗法是试图增强这种免疫应答以超过身体的耐受性和免疫抑制作用的临床路径。因此,包含上述突变的蛋白质或肽是免疫疗法的合适的候选物。

[0471] 突变的蛋白质被充当抗原呈递细胞(APC)的专职吞噬细胞摄取,切碎,并作为抗原展示在细胞表面上,以供在包含主要组织相容性复合物(MHC)蛋白的抗原呈递复合物中的T细胞活化。人类MHC蛋白被称为人类白细胞抗原,HLA。MHC蛋白可以是MHC I类或II类蛋白,一些功能差异归因于I类或II类MHC蛋白(HLA I类和HLA II类蛋白)对肽的呈递,一个显著的区别在于HLA I类-肽复合物将抗原呈递给细胞毒性CD8+ T细胞,而HLA II类肽复合物还能够激活CD4+ T细胞,从而导致延长的免疫应答。CD8+ T细胞在逐细胞清除病病细胞(例如受感染细胞或肿瘤细胞)的任务中必不可少。CD4+ T细胞在激活后具有更持久的影响,其中最重要的是产生免疫记忆。根据免疫威胁的类型不同地招募CD4子集,并且可以共同招募具有重叠或不同功能的多个子集。这有助于平衡与病原体威胁有关的免疫应答。在这些方面,HLA II类肽介导的抗原呈递实现持续和定制的免疫应答。另一方面,HLA II类与肽的结合可能是混杂的,因此非特异性肽结合和向免疫系统的呈递会导致异常免疫应答,如自身免疫。

[0472] 在一方面,本公开提供了预测肽的方法,该肽可以与特定HLA II类α和β链异二聚

体准确地配对或结合,使得该肽与HLA II类蛋白(包括α和β链异二聚体)的高保真度结合确保将特定肽呈递给T淋巴细胞,从而引发特定的免疫应答并避免任何交叉反应或免疫混杂性。

[0473] 在一方面,本公开提供了预测肽的方法,该肽可以与特定HLA II类蛋白准确地结合,使得当将肽治疗性地施用于表达特定同源HLA II类蛋白的受试者时,借助于HLA II类蛋白激活CD4+ T细胞并刺激免疫记忆的能力,可以用该肽激活更持续且有力的免疫应答。在一些实施方案中,预测以高特异性与HLA II类蛋白结合的给定肽是包含突变的肽,其中该突变在受试者的癌症或肿瘤细胞中普遍存在;而预测与突变肽结合的相同HLA II类蛋白要么(a)不结合,要么(b)与相应的非突变野生型肽结合的亲和力显著低于与受试者的突变肽结合的亲和力。HLA与突变肽的优先结合在免疫治疗剂的开发中是有利的,因为表达野生型肽的细胞将免受对HLA呈递的肽具有反应性的T细胞的免疫攻击。在一些实施方案中,与HLA II类蛋白特异性结合的预测肽是具有翻译后修饰的肽。示例性的翻译后修饰包括但不限于:磷酸化、泛素化、去磷酸化、糖基化、甲基化或乙酰化。在一些实施方案中,预测的肽在用于免疫治疗之前进行翻译后修饰。

[0474] 在一些实施方案中,本文公开的免疫治疗方法和策略还可适用于抑制不需要的免疫激活,例如在自身免疫反应中。具体来说,被鉴定为特定HLA亚型的潜在结合物的肽可以被定制为与特定HLA分子结合并诱导耐受而不是引起免疫原性反应。

[0475] 在一方面,本文提出了针对特定受试者定制或个性化的免疫治疗方法。每个受试者或患者都表达特定的一组HLA I类和HLA II类蛋白。HLA分型是一种众所周知的技术,其允许确定受试者表达的HLA蛋白的特定组库(repertoire)。一旦已知由特定受试者表达的HLA异二聚体,具有如本文所述的用于以高保真度预测可与特定HLA II类α和β链异二聚体结合的肽的改进的、复杂的和可靠的方法,可以确保能生成专门为该受试者定制的特异性免疫应答。

[0476] 编码HLA异二聚体的基因是高度多态性的,在整个人类群体中鉴定出超过4,000个HLA II类等位基因变体。对于HLA II类基因座中的每一个,个体可以从母本和父本HLA单倍型继承不同的等位基因,并且每个HLA II类异二聚体由α链和β链组成。由于大量的α链和β链配对组合,特别是对于HLA-DP和HLA-DQ等位基因,可能的HLA异二聚体的群体非常复杂。HLA II类异二聚体在内质网(ER)中翻译,并与源自蛋白质CD74的不变链(Ii)组装成稳定的复合物。Ii通过允许正确的蛋白质折叠来稳定II类复合物,并使HLA II类异二聚体能够输出到内体/溶酶体区室中。在这些HLA II类加载区室内,Ii被组织蛋白酶蛋白水解切割成被称为CLIP的占位肽。然后CLIP在低pH环境中通过分子伴侣HLA-DM(一种非经典的HLA II类异二聚体)交换为更高亲和力的肽。加载有高亲和力肽的HLA II类复合物然后到达反式高尔基体,最后到达细胞表面以供CD4+ T细胞展示。

[0477] 估计每个HLA异二聚体以等位基因特异性结合偏好结合数千个肽。实际上,估计每个HLA等位基因结合并向T细胞呈递约1,000-10,000种独特的肽。考虑到HLA结合的这种多样性,对肽是否可能与特定HLA等位基因结合的准确预测非常具有挑战性。对HLA II类分子的等位基因特异性肽结合特性知之甚少,这是因为α和β链配对的异质性,数据的复杂性限制了确信地指定核心结合表位的能力,以及缺乏高分辨率生化分析所需的免疫沉淀等级的等位基因特异性抗体。此外,当在细胞表面呈递多个HLA等位基因时,分析源自给定HLA等位

基因的肽表位会引起不确定性。

[0478] 候选新抗原的预测主要针对HLA I类表位进行(考虑到I类预测算法的实验数据与II类相比的可获得性),但在临床前和临床个性化新抗原疫苗接种研究中经常观察到CD4+T细胞应答。这些观察表明,HLA II类表位加工和呈递也可能在癌症治疗中发挥关键作用。尽管存在HLA II类预测算法,但它们是不准确的,因为HLA II类异二聚体上的开放式肽结合沟允许更长的肽(通常为15-40个氨基酸)结合,这增加了表位呈递的异质性和复杂性。因此,需要进一步的工作以更好地了解HLA II类肽结合核心的特性以及II类表位加工和呈递所涉及的细胞过程。蛋白质组学领域目前受到HLA II类异二聚体形成的复杂性和用于HLA II类-肽复合物分离的免疫沉淀级抗体的可用性的限制。为了克服这些挑战,开发了单等位基因HLA概况分析工作流程,该工作流程依赖于LC-MS/MS来表征等位基因特异性HLA II类-配体组和II类表位预测方法。以下定义补充了本领域中的定义,并且针对本申请,而不应归于任何相关或不相关的情况,例如任何共同拥有的专利或申请。本文描述了示例性材料和方法,但是与本文描述的方法和材料类似或等同的任何方法和材料可以在测试本公开的实践中使用。因此,本文使用的术语仅仅是为了描述特定实施方案的目的,而并非旨在限制。

[0479] 本文公开了制备个性化癌症疫苗的方法。制备个性化癌症疫苗的方法可以包括鉴定在受试者的癌细胞中表达的具有突变的肽序列;使用计算机处理器将所鉴定的肽序列的氨基酸位置信息输入到机器学习HLA-肽呈递预测模型中,以生成针对所鉴定的肽序列的一组呈递预测,每个呈递预测代表由受试者癌细胞的II类MHC等位基因编码的一种或多种蛋白质将呈递所鉴定的肽序列的给定序列的概率;并选择基于这组呈递预测鉴定的肽序列的子集,以用于制备个性化癌症疫苗。

[0480] 在一些实施方案中,从本文描述的方法获得的一个或多个结果可提供指示以下一项或多项的一个或多个定量值:诊断准确性的可能性、受试者中存在病况的可能性、受试者发生病况的可能性、特定治疗成功的可能性或其任何组合。在一些实施方案中,如本文所述的方法可以预测发生病况的风险或可能性。在一些实施方案中,如本文所述的方法可以是发生病况的早期诊断指示。在一些实施方案中,如本文所述的方法可以确认病况的诊断或存在。在一些实施方案中,如本文所述的方法可以监测病况的进展。在一些实施方案中,如本文所述的方法可以监测治疗对受试者的病况的功效。

[0481] MHC-II肽的鉴定方法

[0482] 在一方面,本文提供了一种鉴定由MHC-II蛋白呈递以用于免疫活化的一种或多种肽的方法。在一些实施方案中,所述一种或多种肽包含表位。在一些实施方案中,该方法涉及特定表位由MHC-II蛋白呈递的可能性的计算预测。在一些实施方案中,该方法涉及表位对于MHC-II呈递的特异性的计算预测。在一些实施方案中,计算预测方法涉及肽-MHC相互作用的评估。在一些实施方案中,计算预测方法涉及预测肽对抗原呈递的等位基因特异性。

[0483] 在一些实施方案中,计算预测方法涉及生物信息学信息的整合,例如核苷酸序列、生物分子的结构基序、蛋白质-蛋白质相互作用特征和功能效力如免疫原性。在一些实施方案中,计算预测方法涉及机器学习。基于机器学习方法,如简单模式基序、支持向量机(SVM)、隐马尔可夫模型(HMM)、神经网络(NN)模型、定量构效关系(QSAR)分析、基于结构的方法和生物物理学方法,已经针对MHC I类和II类开发了许多用于预测肽-MHC相互作用的免疫信息学方法。这些方法可以分为两类,即等位基因内(等位基因特异性)和跨等位基因

(泛特异性)方法。等位基因内方法在有限的一组实验肽结合数据上针对特定MHC分子进行训练,并应用于预测与该分子结合的肽。由于MHC分子的极端多态性,数千个等位基因变体的存在,加之缺乏足够的实验结合数据,无法对每个等位基因建立预测模型。因此,已经使用在许多等位基因上或跨物种扩展的肽结合数据开发了跨等位基因和通用方法,如NetMHCIIpan(Karosiene E等人,NetMHCIIpan-3.0,a common pan-specific MHC class II prediction method including all three human MHC class II isotypes,HLA-DR,HLA-DP and HLADQ.Immunogenetics(2013)65(10):711-24)和TEPITOPEpan(Zhang L等人,TEPITOPEpan:extending TEPITOPE for peptide binding prediction covering over 700HLA-DR molecules.PLoS One(2012)7(2):e30483)。用于MHC-I的类似方法也可以获得,如NetMHCpan和KISS。

[0484] 在一些实施方案中,肽序列可能不在受试者的正常细胞中表达。在一些实施方案中,受试者的每个和每一个细胞可能不是癌细胞。所述癌细胞可以由不同的癌症产生,包括但不限于甲状腺癌、肾上腺皮质癌、肛门癌、再生障碍性贫血、胆管癌、膀胱癌、骨癌、骨转移、中枢神经系统(CNS)癌、周围神经系统(PNS)癌、乳腺癌、Castleman病、宫颈癌、儿童非霍奇金淋巴瘤、淋巴瘤、结肠直肠癌、子宫内膜癌、食管癌、尤因肿瘤家族(例如尤因肉瘤)、眼癌、胆囊癌、胃肠道类癌瘤、胃肠道间质肿瘤、妊娠滋养细胞疾病、毛细胞白血病、霍奇金病、卡波西肉瘤、肾癌、喉癌和下咽癌、急性淋巴细胞性白血病、急性髓样白血病、儿童白血病、慢性淋巴细胞性白血病、慢性髓样白血病、肝癌、肺癌、肺类癌瘤、非何杰金淋巴瘤、男性乳腺癌、恶性间皮瘤、多发性骨髓瘤、脊髓增生异常综合征、骨髓增生性疾病、鼻腔和鼻旁癌、鼻咽癌、神经母细胞瘤、口腔和口咽癌、骨肉瘤、卵巢癌、胰腺癌、阴茎癌、垂体瘤、前列腺癌、视网膜母细胞瘤、横纹肌肉瘤、唾液腺癌、肉瘤(成人软组织癌)、黑素瘤皮肤癌、非黑素瘤皮肤癌、胃癌、睾丸癌、胸腺癌、子宫癌(例如子宫肉瘤)、阴道癌、外阴癌或瓦尔登斯特伦巨球蛋白血症。

[0485] 所述鉴定可包括将来自受试者癌细胞的DNA、RNA或蛋白质序列与来自受试者正常细胞的DNA、RNA或蛋白质序列进行比较。来自受试者癌细胞的DNA、RNA或蛋白质序列可以不同于来自受试者正常细胞的DNA、RNA或蛋白质序列。所述鉴定可以以高灵敏度鉴定核酸变体。

[0486] 所述机器学习HLA-肽呈递预测模型可以包含至少基于训练数据鉴定的多个预测变量。所述训练数据可以包含由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数。

[0487] 在一些实施方案中,所述训练数据可以进一步包括结构化数据、时间序列数据、非结构化数据和关系数据。非结构化数据可以包括音频数据、图像数据、视频、机械数据、电气数据、化学数据及其任何组合,用于准确地模拟或训练机器人或模拟。时间序列数据可以包括来自智能仪表、智能电器、智能设备、监控系统、遥测设备或传感器中的一个或多个的数据。关系数据包括来自客户系统、企业系统、操作系统、网站、网络可访问应用程序界面(API)或其任何组合的数据。这可以由用户通过将文件或其他数据格式输入到软件或系统中的任何方法来完成。

[0488] 在一些实施方案中,训练数据可以存储在数据库中。数据库可以以计算机可读格式存储。计算机处理器可以被配置为访问存储在计算机可读存储器中的数据。在一些实施方案中,计算机系统可用于分析数据以获得结果。该结果可以远程或内部存储在存储介质上,并传送给诸如药物专家之类的人员。在一些实施方案中,计算机系统可以与用于传输结果的组件可操作地耦合。用于传输的组件可以包括有线和无线组件。有线通信组件的示例可以包括通用串行总线(USB)连接、同轴电缆连接、以太网电缆如Cat5或Cat6电缆、光纤电缆或电话线。无线通信组件的示例可以包括Wi-Fi接收器、用于访问诸如3G或4G LTE数据信号等移动数据标准的组件或蓝牙接收器。在一些实施方案中,存储介质中的所有这些数据被收集并存档以构建数据仓库。

[0489] 在一些实施方案中,数据库包括外部数据库。外部数据库可以是医学数据库,例如但不限于药物不良反应数据库、AHFS补充文件、变应原选择列表文件、平均WAC定价文件、品牌概率文件、加拿大药物文件v2、综合价格历史、受控物质文件、药物过敏交叉参考文件、药物申请文件、药物给药和施用数据库、药物图像数据库v2.0/药物印记数据库v2.0、药物无效日期文件、药物适应症数据库、药物实验室冲突数据库、药物治疗监测系统(DTMS)v2.2/DTMS消费者专著、重复治疗数据库、联邦政府定价文件、医疗保健通用程序编码系统代码(HCPCS)数据库、ICD-10映射文件、免疫交叉参考文件、综合A到Z药物事实模块、综合患者教育、主参数数据库、半跨度电子药物文件(MED-File)v2、Medicaid Rebate文件、医疗照护计划文件、医疗条件选择列表文件、医疗条件主数据库、药物订单管理数据库(MOMD)、监测参数数据库、患者安全计划文件、支付限额-部分B(PAF-B)v2.0、预防措施数据库、RxNorm交叉参考文件、标准药物标识符数据库、替换组文件、补充名称文件、统一分类交叉参考文件或警告标签数据库。

[0490] 在一些实施方案中,训练数据也可以通过其他数据源获得。数据源可包括传感器或智能设备,如电器、智能仪表、可穿戴设备、监测系统、数据存储、客户系统、计费系统、金融系统、人群源数据、天气数据、社交网络或任何其他传感器、企业系统或数据存储。智能仪表或传感器的示例可以包括位于客户地点的仪表或传感器,或者位于客户与生成或源位置之间的仪表或传感器。通过整合来自广泛来源的数据,该系统可能能够执行复杂和详细的分析。在一些实施方案中,数据源可以包括但不限于用于其他医疗平台的传感器或数据库。

[0491] HLA分型常规地通过使用抗体的血清学方法或通过基于PCR的方法如序列特异性寡核苷酸探针杂交(SSOP)或基于序列的分型(SBT)来进行。第一个受到潜在的高度交叉反应性和有限的分辨能力的阻碍,而第二个具有与PCR效率相关的困难,这是因为由于多态性位置导致的定位引物的可能性非常有限。

[0492] 在一些实施方案中,通过测序方法或使用质谱法的方法,如液相色谱-质谱法(LC-MS或LC-MS/MS,或者HPLC-MS或HPLC-MS/MS)来鉴定序列信息。这些测序方法可以是技术人员公知的,并且在Medzihradszky KF和Chalkley RJ.Mass Spectrom Rev.2015年1月-2月;34(1):43-63中进行了综述。在一些实施方案中,该质谱法是单等位基因质谱法。在一些实施方案中,该质谱法可以是MS分析、MS/MS分析、LC-MS/MS分析或其组合。在一些实施方案中,可以使用MS分析来确定完整肽的质量。例如,所述确定可以包括确定完整肽的质量(例如,MS分析)。在一些实施方案中,可以使用MS/MS分析来确定肽片段的质量。例如,所述确定可以包括确定肽片段的质量,后者可以用来确定肽或其部分的氨基酸序列(例如,MS/MS分

析)。在一些实施方案中,可以使用肽片段的质量来确定该肽内的氨基酸序列。在一些实施方案中,可以使用LC-MS/MS分析来分离复杂的肽混合物。例如,所述确定可以包括例如通过液相色谱法分离复杂的肽混合物,以及确定完整肽的质量、肽片段的质量或其组合(例如,LC-MS/MS分析)。该数据可用于例如肽测序。

[0493] 在一些实施方案中,训练肽序列信息包括训练肽的氨基酸位置信息。在一些实施方案中,训练肽序列信息包括由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的至多约90%、80%、70%、60%、50%、40%、30%、20%、10%或更少的序列信息。在一些实施方案中,训练肽序列信息可包括由在细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的至少约10%、20%、30%、40%、50%、60%、70%、80%、90%或更多的序列信息。

[0494] 任何信息和数据都可以与作为该信息和数据来源的受试者配对。受试者或医疗专业人员可以通过受试者身份从存储或服务器中检索信息和数据。受试者身份可以包括患者的照片、姓名、地址、社会安全号、生日、电话号码、邮政编码或其任意组合。受试者身份可以加密并编码为可视图形代码。可视图形代码可以是一次性条形码,其可以唯一地与受试者身份相关联。条形码可以是UPC条形码、EAN条形码、Code 39条形码、Code 128条形码、ITF条形码、CodaBar条形码、GS1 DataBar条形码、MSI Plessey条形码、QR条形码、Datamatrix码、PDF417码或Aztec条形码。可视图形代码可以被配置为显示在显示屏上。条形码可以包括可以被机器光学捕获并读取的QR。条形码可以定义诸如条形码的版本、格式、位置、对齐或定时等元素以实现条形码的读取和解码。条形码可以以任何类型的合适格式对各种类型的信息进行编码,例如二进制或字母数字信息。QR码就可以具有各种符号大小,只要QR码可以通过成像设备从合理的距离扫描。QR码可以是任何图像文件格式(例如EPS或SVG矢量图形、PNG、TIF、GIF或JPEG光栅图形格式)。

[0495] 在一些实施方案中,代表作为输入接收的氨基酸位置信息与基于该氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数包括线性或非线性函数。该函数可以是例如修正线性单元(ReLU)激活函数、Leaky ReLu激活函数或其他函数,如饱和双曲正切、恒等式、二进制阶跃函数、逻辑函数、arcTan、softsign、参数修正线性单元、指数线性单元、softPlus、bent identity、softExponential、正弦曲线(Sinusoid)、Sinc、高斯(Gaussian)或S形函数,或其任意组合。

[0496] 在一些实施方案中,线性函数是通过线性回归获得的。在一些实施方案中,线性回归是通过拟合因变量与自变量之间的最佳线性关系来预测目标变量的方法。最佳拟合可能意味着每个点处的形状与实际观察值之间的所有距离之和最小。线性回归可以包括简单线性回归或多元线性回归。简单线性回归可以使用单个自变量来预测因变量。多元线性回归可以使用多于一个自变量通过拟合最佳线性关系来预测因变量。非线性函数可以通过非线性回归获得。非线性回归可以是回归分析的一种形式,其中观察数据由作为模型参数的非线性组合并且取决于一个或多个自变量的函数建模。非线性回归可以包括阶跃函数、分段函数、样条和广义加性模型。

[0497] 在一些实施方案中,呈递可能性由一维值(例如,概率)来呈现。在一些实施方案中,概率被配置为测量事件可能发生的可能性。在一些实施方案中,概率范围为约0至1、0.1至0.9、0.2至0.8、0.3至0.7或0.4至0.6。事件发生的概率越高,该事件就越可能发生。在一些实施方案中,事件包括任何类型的情况,作为非限制性实例,包括HLA-肽是否会呈递具有

特定氨基酸位置信息的某些肽,以及基于氨基酸位置信息,人是否会生病。在一些实施方案中,可能性可以由多维值来呈现。多维值可以通过多维空间、热图或电子表格来呈现。

[0498] 在一个实施方案中,选择基于一组呈递预测鉴定的肽序列的子集被配置为制备个性化癌症疫苗。在一些实施方案中,该子集包含至多约90%、80%、70%、60%、50%、40%、30%、20%、10%或更少的基于该组呈递预测鉴定的肽序列。在其他情况下,该子集可以包含至少约10%、20%、30%、40%、50%、60%、70%、80%、90%或更多的基于该组呈递预测鉴定的肽序列。癌症疫苗可以是治疗现有癌症或预防癌症发展的疫苗。疫苗可以由取自患者的样品制备,并且可能是对于该患者特定的。

[0499] 在一些实施方案中,痘病毒在疾病疫苗或免疫原性组合物中使用。这些包括正痘病毒、禽痘(avipox)、牛痘、MVA、NYVAC、金丝雀痘、ALVAC、禽痘(fowlpox)、TROVAC等。该载体的优点可包括简单构建、适应大量外来DNA的能力和高表达水平。关于可用于实施本公开内容的痘病毒如脊椎动物痘病毒亚科(Chordopoxvirinae)痘病毒(脊椎动物的痘病毒),例如,正痘病毒和禽痘病毒,例如牛痘病毒(例如,Wyeth株、WR株(例如,ATCC® VR-1354)、哥本哈根(Copenhagen)株、NYVAC、NYVAC.1、NYVAC.2、MVA、MVA-BN)、金丝雀痘病毒(例如,Wheatley C93株、ALVAC),禽痘病毒(例如,FP9株、Webster株、TROVAC)、鸽痘(dovepox)、鸽痘(pigeonpox)、鹌鹑痘(quailpox)和浣熊痘,尤其是其合成或非天然存在的重组体、其用途以及制备和使用这类重组体的方法的信息可以在科学和专利文献中找到。

[0500] 在一些实施方案中,痘苗病毒在疾病疫苗或免疫原性组合物中使用以表达抗原。重组牛痘病毒可以能够在感染的宿主细胞的细胞质内复制,因此目的多肽可以诱导免疫应答。

[0501] 在一些实施方案中,ALVAC用作疾病疫苗或免疫原性组合物中的载体。ALVAC可以是这样的金丝雀痘病毒,其可以经修饰以表达外来转基因,并且已用作针对原核和真核抗原的接种方法。

[0502] 在一些实施方案中,修饰的安卡拉牛痘(MVA)病毒用作抗原疫苗或免疫原性组合物的病毒载体。MVA可以是正痘病毒科的成员,并且已经通过牛痘病毒(CVA)的安卡拉株在鸡胚成纤维细胞中的约570次连续传代产生。由于这些传代,所得的MVA病毒含有的基因组信息可以比CVA少31千碱基,并且是高度宿主细胞限制性的。MVA的特征可以在于其极度衰减,即减弱的毒力或感染能力,但仍具有优异的免疫原性。当在多种动物模型中进行测试时,MVA可被证明是无毒的,甚至在免疫抑制的个体中也是无毒的。此外,MVA-BN®-HER2可以是用于治疗HER-2阳性乳腺癌的候选免疫治疗,并且目前正在进行临床试验。

[0503] 在一些实施方案中,阳性预测值(PPV)用作预测模型的一部分。PPV,也称为精度测量值,是通过例如测试或模型被诊断出患有疾病或病况的个体实际患有该疾病或病况的概率。它可以通过将真阳性结果的数目除以返回阳性结果(包括假阳性的结果)的总数来计算。PPV=真阳性/(真阳性+假阳性)。例如,如果在一组100名患者中,模型在50名患者中确定了阳性结果,其中25名是真阳性,则PPV将为25/50=0.5。PPV越接近1,表示诊断方法如测试或模型越精确。PPV可用于确定预测模型的准确性。PPV可用于调整预测模型以适应该模型可能生成的假阳性结果。

[0504] 判定率可以用作预测模型的一部分。判定率可以被认为是真阳性结果占样品集中阳性总数的百分比。判定率=真阳性/(真阳性+假阴性)。例如,如果在一组100名患者中,模

型在50名患者中确定了阳性结果,其中25名是真阳性,而该组患者中总共有75名阳性,则判定率为{25/(25+25)}x100＝50％。判定率可用于确定预测模型的准确性。判定率可用于调整预测模型以适应该模型可能生成的假阳性结果或假阴性结果。

[0505] 在一些实施方案中,所述预测模型在0.1％-10％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在0.1％-10％的判定率下可以具有至多0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1或更低的阳性预测值。所述预测模型在低于0.1％的判定率下可以具有至少0.05、0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在低于0.1％的判定率下可以具有至多0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1或更低的阳性预测值。所述预测模型在高于10％的判定率下可以具有至少0.05、0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在高于10％的判定率下可以具有至多0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1或更低的阳性预测值。

[0506] 在一些实施方案中,所述预测模型在0.1％至10％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在0.1％至0.5％、0.1％至1％、0.1％至2％、0.1％至3％、0.1％至4％、0.1％至5％、0.1％至6％、0.1％至7％、0.1％至8％、0.1％至9％、0.1％至10％、0.5％至1％、0.5％至2％、0.5％至3％、0.5％至4％、0.5％至5％、0.5％至6％、0.5％至7％、0.5％至8％、0.5％至9％、0.5％至10％、1％至2％、1％至3％、1％至4％、1％至5％、1％至6％、1％至7％、1％至8％、1％至9％、1％至10％、2％至3％、2％至4％、2％至5％、2％至6％、2％至7％、2％至8％、2％至9％、2％至10％、3％至4％、3％至5％、3％至6％、3％至7％、3％至8％、3％至9％、3％至10％、4％至5％、4％至6％、4％至7％、4％至8％、4％至9％、4％至10％、5％至6％、5％至7％、5％至8％、5％至9％、5％至10％、6％至7％、6％至8％、6％至9％、6％至10％、7％至8％、7％至9％、7％至10％、8％至9％、8％至10％或9％至10％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在0.1％、0.5％、1％、2％、3％、4％、5％、6％、7％、8％、9％或10％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在至少0.1％、0.5％、1％、2％、3％、4％、5％、6％、7％、8％或9％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在至多0.5％、1％、2％、3％、4％、5％、6％、7％、8％、9％或10％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。

[0507] 在一些实施方案中,所述预测模型在10％至20％的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在10％至11％、10％至12％、10％至13％、10％至14％、10％至15％、10％至16％、10％至17％、10％至18％、10％至19％、10％至20％、11％至12％、11％至13％、11％至14％、11％至15％、11％至16％、11％至17％、11％至18％、11％至19％、11％至20％、12％至13％、12％至14％、12％至15％、12％至16％、12％至17％、12％至18％、12％至19％、12％至20％、13％至14％、13％至15％、13％至16％、13％至17％、13％至18％、13％至19％、13％至20％、14％至15％、14％至16％、14％至17％、14％至18％、14％至19％、14％至20％、15％至16％、

15%至17%、15%至18%、15%至19%、15%至20%、16%至17%、16%至18%、16%至19%、16%至20%、17%至18%、17%至19%、17%至20%、18%至19%、18%至20%或19%至20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在至少10%、11%、12%、13%、14%、15%、16%、17%、18%或19%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。在一些实施方案中,所述预测模型在至多11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。

[0508] 在一些实施方案中,所述预测模型在至少0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至少10%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至少5%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至少20%的判定率下可以具有至少0.9的阳性预测值。

[0509] 在一些实施方案中,所述预测模型在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在约10%的

判定率下可以具有至少0.2的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在约10%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在约5%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在约20%的判定率下可以具有至少0.9的阳性预测值。

[0510] 在一些实施方案中,所述预测模型在小于0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下具有至少0.05、0.1、0.2、0.25、0.3、0.4、0.5、0.6、0.7、0.8、0.9或更高的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至多10%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.2的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至多5%的判定率下可以具有至少0.9的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.1的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.2的阳性预测值。例如,预测模

型在至多20%的判定率下可以具有至少0.3的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.4的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.5的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.6的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.7的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.8的阳性预测值。例如,预测模型在至多20%的判定率下可以具有至少0.9的阳性预测值。

[0511] 在一些实施方案中,在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型具有0.05%至0.6%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有0.05%至0.1%、0.05%至0.15%、0.05%至0.2%、0.05%至0.25%、0.05%至0.3%、0.05%至0.35%、0.05%至0.4%、0.05%至0.45%、0.05%至0.5%、0.05%至0.55%、0.05%至0.6%、0.1%至0.15%、0.1%至0.2%、0.1%至0.25%、0.1%至0.3%、0.1%至0.35%、0.1%至0.4%、0.1%至0.45%、0.1%至0.5%、0.1%至0.55%、0.1%至0.6%、0.15%至0.2%、0.15%至0.25%、0.15%至0.3%、0.15%至0.35%、0.15%至0.4%、0.15%至0.45%、0.15%至0.5%、0.15%至0.55%、0.15%至0.6%、0.2%至0.25%、0.2%至0.3%、0.2%至0.35%、0.2%至0.4%、0.2%至0.45%、0.2%至0.5%、0.2%至0.55%、0.2%至0.6%、0.25%至0.3%、0.25%至0.35%、0.25%至0.4%、0.25%至0.45%、0.25%至0.5%、0.25%至0.55%、0.25%至0.6%、0.3%至0.35%、0.3%至0.4%、0.3%至0.45%、0.3%至0.5%、0.3%至0.55%、0.3%至0.6%、0.35%至0.4%、0.35%至0.45%、0.35%至0.5%、0.35%至0.55%、0.35%至0.6%、0.4%至0.45%、0.4%至0.5%、0.4%至0.55%、0.4%至0.6%、0.45%至0.5%、0.45%至0.55%、0.45%至0.6%、0.5%至0.55%、0.5%至0.6%或0.55%至0.6%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有0.05%、0.1%、0.15%、0.2%、0.25%、0.3%、0.35%、0.4%、0.45%、0.5%、0.55%或0.6%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有至少0.05%、0.1%、0.15%、0.2%、0.25%、0.3%、0.35%、0.4%、0.45%、0.5%或0.55%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有至多0.1%、0.15%、0.2%、0.25%、0.3%、0.35%、0.4%、0.45%、0.5%、0.55%或0.6%的阳性预测值。

[0512] 在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有0.45%至0.98%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有0.45%至0.5%、0.45%至0.55%、0.45%至0.6%、0.45%至0.65%、0.45%至0.7%、0.45%至0.75%、0.45%至0.8%、0.45%至0.85%、0.45%至0.9%、0.45%至0.96%、0.45%至0.98%、0.5%至0.55%、0.5%至0.6%、0.5%至0.65%、0.5%至0.7%、0.5%至

0.75%、0.5%至0.8%、0.5%至0.85%、0.5%至0.9%、0.5%至0.96%、0.5%至0.98%、0.55%至0.6%、0.55%至0.65%、0.55%至0.7%、0.55%至0.75%、0.55%至0.8%、0.55%至0.85%、0.55%至0.9%、0.55%至0.96%、0.55%至0.98%、0.6%至0.65%、0.6%至0.7%、0.6%至0.75%、0.6%至0.8%、0.6%至0.85%、0.6%至0.9%、0.6%至0.96%、0.6%至0.98%、0.65%至0.7%、0.65%至0.75%、0.65%至0.8%、0.65%至0.85%、0.65%至0.9%、0.65%至0.96%、0.65%至0.98%、0.7%至0.75%、0.7%至0.8%、0.7%至0.85%、0.7%至0.9%、0.7%至0.96%、0.7%至0.98%、0.75%至0.8%、0.75%至0.85%、0.75%至0.9%、0.75%至0.96%、0.75%至0.98%、0.8%至0.85%、0.8%至0.9%、0.8%至0.96%、0.8%至0.98%、0.85%至0.9%、0.85%至0.96%、0.85%至0.98%、0.9%至0.96%、0.9%至0.98%或0.96%至0.98%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有0.45%、0.5%、0.55%、0.6%、0.65%、0.7%、0.75%、0.8%、0.85%、0.9%、0.96%或0.98%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有至少0.45%、0.5%、0.55%、0.6%、0.65%、0.7%、0.75%、0.8%、0.85%、0.9%或0.96%的阳性预测值。在约0.1%、0.5%、1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%或20%的判定率下,预测模型可以具有至多0.5%、0.55%、0.6%、0.65%、0.7%、0.75%、0.8%、0.85%、0.9%、0.96%或0.98%的阳性预测值。

[0513] 训练机器学习HLA-肽呈递预测模型的方法

[0514] 在一方面,训练机器学习HLA-肽呈递预测模型的方法可以包括使用计算机处理器的将从表达HLA II类等位基因的细胞的一个或多个HLA-肽复合物中分离的HLA-肽的氨基酸位置信息序列输入到HLA-肽呈递预测模型中;训练机器学习HLA-肽呈递预测模型可以包括调整神经网络节点上的加权值,以最好地匹配所提供的训练数据。

[0515] 所述训练数据可以包含由细胞中表达的HLA蛋白呈递并通过质谱法鉴定的肽的序列的序列信息;包含训练肽的氨基酸位置信息的训练肽序列信息,其中所述训练肽序列信息与细胞中表达的HLA蛋白相关;以及代表作为输入接收的氨基酸位置信息与基于所述氨基酸位置信息和预测变量作为输出生成的呈递可能性之间的关系的函数。训练数据、训练肽序列信息、函数和呈递可能性在本文别处公开。

[0516] 经训练的算法可包括一个或多个神经网络。神经网络可以是一种类型的基于一系列层中几个连接的神经元(或节点)的图形的计算系统。神经网络可以包括输入层,数据被呈递给该输入层;一个或多个内部和/或“隐藏”层;和输出层,从中呈现结果。神经网络可以通过调整一系列连接权重来学习输入数据集与目标数据集之间的关系。神经元可以通过具有权重的连接与其他层中的神经元连接,权重是控制连接强度的参数。每层中神经元的数目可能与待解决的问题的复杂性有关。层中需要的神经元的最小数目可由问题复杂性决定,最大数目可能受神经网络泛化能力的限制。输入神经元可以接收呈现的数据,然后通过连接权重将该数据传输到第一隐藏层中的节点,连接权重在训练期间进行修改。结果节点可以加和所有输入对的乘积及其相关权重。加权和可以偏移出偏差,以调整结果节点的值。节点或神经元的输出可以使用阈值或激活函数进行门控。激活函数可以是线性或非线性函

数。激活函数可以是例如修正线性单元(ReLU)激活函数、Leaky ReLu激活函数或其他函数,如饱和双曲正切、恒等式、二进制阶跃函数、逻辑函数、arcTan、softsign、参数修正线性单元、指数线性单元、softPlus、bent identity、softExponential、正弦曲线(Sinusoid)、Sinc、高斯(Gaussian)或S形函数,或其任意组合。

[0517] 神经网络中的隐藏层可以处理数据并通过第二组加权连接将其结果传输到下一层。每个后续层都可以将来自先前层的结果“汇集”为更复杂的关系。神经网络可以使用训练数据的已知的样本组(从一个或多个传感器收集的数据)进行训练,方法是使它们在训练期间(和之后)修改自己,以便从给定的一组输入中提供所需的输出,如输出值。经过训练的算法可以包括卷积神经网络、循环神经网络、扩张卷积神经网络、完全连接的神经网络、深度生成模型和Boltzmann机。

[0518] 可以在训练阶段使用一组或多组训练数据来“教导”或“学习”神经网络的加权因子、偏好值和阈值或其他计算参数。例如,可以使用来自训练数据集的输入数据和梯度下降或反向传播方法来训练参数,使得来自神经网络的输出值与训练数据集中包括的示例一致。

[0519] 在神经网络的输入层中使用的节点数目可以是至少约10、50、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10,000、20,000、30,000、40,000、50,000、60,000、70,000、80,000、90,000、100,000或更多。在其他情况下,输入层中使用的节点数目可以是至多约100,000、90,000、80,000、70,000、60,000、50,000、40,000、30,000、20,000、10,000、9000、8000、7000、6000、5000、4000、3000、2000、1000、900、800、700、600、500、400、300、200、100、50或10更少。在一些情况下,神经网络中使用的层的总数(包括输入层和输出层)可以是至少约3、4、5、10、15、20或更多。在其他情况下,层的总数可以是至多约20、15、10、5、4、3或更少。

[0520] 在一些情况下,神经网络中使用的可学习或可训练参数(例如,加权因子、偏好或阈值)的总数可以是至少约10、50、100、200、300、400、500、600、700、800、900、1000、2000、3000、4000、5000、6000、7000、8000、9000、10,000、20,000、30,000、40,000、50,000、60,000、70,000、80,000、90,000、100,000或更多。在其他情况下,可学习参数的数目可以是至多约100,000、90,000、80,000、70,000、60,000、50,000、40,000、30,000、20,000、10,000、9000、8000、7000、6000、5000、4000、3000、2000、1000、900、800、700、600、500、400、300、200、100、50或10或更少。

[0521] 神经网络可以包括卷积神经网络。卷积神经网络可以包括一个或多个卷积层、扩张层或完全连接层。卷积层的数目可以是1-10,扩张层的数目可以是0-10。卷积层(包括输入层和输出层)的总数可以是至少约1、2、3、4、5、10、15、20或更多,并且扩张层的总数可以是至少约1、2、3、4、5、10、15、20或更多。卷积层的总数可以是至多约20、15、10、5、4、3或更少,并且扩张层的总数可以是至多约20、15、10、5、4、3或更少。在一些实施方案中,卷积层的数目为1-10,并且完全连接层的数目为0-10。卷积层(包括输入和输出层)的总数可以是至少约1、2、3、4、5、10、15、20或更多,并且完全连接层的总数可以是至少约1、2、3、4、5、10、15、20或更多。卷积层的总数可以是至多约20、15、10、5、4、3或更少,并且完全连接层的总数可以是至多约20、15、10、5、4、3或更少。

[0522] 卷积神经网络(CNN)可以是深度和前馈人工神经网络。CNN可以适用于分析视觉表

像。CNN可以包含输入层、输出层和多个隐藏层。CNN的隐藏层可以包括卷积层、池化层、完全连接层和归一化层。层可以按3个维度组织:宽度、高度和深度。

[0523] 卷积层可以对输入应用卷积操作并将卷积操作的结果传递到下一层。对于处理图像,卷积操作可以减少自由参数的数目,从而使网络以更少的参数变得更深。在卷积层中,神经元可以仅从前一层的受限分区接收输入。卷积层的参数可以包括一组可学习的过滤器(或内核)。可学习的过滤器可以具有小的感受野并延伸到输入体积的整个深度。在前向传递期间,每个过滤器可以在输入体积的宽度和高度上进行卷积,计算过滤器条目和输入之间的点积,并生成该过滤器的二维激活图。因此,网络可以学习过滤器,当在输入的某个空间位置检测到某种特定类型的特征时,这些过滤器就会激活。

[0524] 池化层可以包括全局池化层。全局池化层可以将一层神经元簇的输出组合成下一层中的单个神经元。例如,最大池化层可以使用来自前一层神经元簇中的每一个的最大值;并且平均池化层可以使用来自前一层神经元簇中的每一个的平均值。完全连接层可以将一层中的每个神经元连接到另一层中的每个神经元。在完全连接层中,每个神经元都可以接收来自前一层的每个元素的输入。归一化层可以是批量归一化层。批量归一化层可以提高神经网络的性能和稳定性。批量归一化层可以为神经网络中的任何层提供作为零均值/单位方差的输入。使用批量归一化层的优点可包括更快的训练网络、更高的学习率、更容易初始化权重、更多可行的激活函数以及更简单的创建深度网络的过程。

[0525] 神经网络可以包括循环神经网络。循环神经网络可以被配置为接收顺序数据作为输入,如连续数据输入,并且循环神经网络软件模块可以在每个时间步更新内部状态。循环神经网络可以使用内部状态(记忆)来处理输入序列。循环神经网络可能适用于诸如手写识别或语音识别、下一个单词预测、音乐创作、图像字幕、时间序列异常检测、机器翻译、场景标记和股票市场预测等任务。循环神经网络可以包括完全循环神经网络、独立循环神经网络、Elman网络、Jordan网络、Echo状态、神经历史压缩器、长短期记忆、门控循环单元、多时标模型、神经图灵机、可微分神经计算机、神经网络下推自动机或其任意组合。

[0526] 经过训练的算法可以包括监督或无监督学习方法,例如SVM、随机森林、聚类算法(或软件模块)、梯度提升、逻辑回归和/或决策树。监督学习算法可以是依赖于使用一组标记的、成对的训练数据示例来推断输入数据与输出数据之间关系的算法。无监督学习算法可以是用于从训练数据集推断输出数据的算法。无监督学习算法可包括聚类分析,其可用于探索性数据分析以发现过程数据中的隐藏模式或分组。无监督学习方法的一个示例可以包括主成分分析。主成分分析可以包括减少一个或多个变量的维数。给定变量的维数可以是至少1、5、10、50、100、200、300、400、500、600、700、800、900、1000、1100、1200 1300、1400、1500、1600、1700、1800或更多。给定变量的维数可以是至多1800、1600、1500、1400、1300、1200、1100、1000、900、800、700、600、500、400、300、200、100、50、10或更少。

[0527] 训练算法可以通过统计技术获得。在一些实施方案中,统计技术可包括线性回归、分类、重采样方法、子集选择、收缩、降维、非线性模型、基于树的方法、支持向量机、无监督学习或其任何组合。

[0528] 线性回归可以是通过拟合因变量与自变量之间的最佳线性关系来预测目标变量的方法。最佳拟合可能意味着每个点处的形状与实际观察值之间的所有距离之和最小。线性回归可以包括简单线性回归和多元线性回归。简单的线性回归可以使用单个自变量来预

测因变量。多元线性回归可以使用多个自变量通过拟合最佳线性关系来预测因变量。

[0529] 分类可以是一种数据挖掘技术,它为数据集合分配类别,以实现准确的预测和分析。分类技术可以包括逻辑回归和判别分析。当因变量是二分的(二进制)时,可以使用逻辑回归。逻辑回归可用于发现并描述一个二元因变量与一个或多个名义、有序、区间或比率水平的自变量之间的关系。重采样可以是包括从原始数据样本中抽取重复样本的方法。重采样可能不涉及使用通用分布表来计算近似概率值。重采样可以根据实际数据生成唯一的采样分布。在一些实施方案中,重新采样可以使用实验方法而不是分析方法来生成独特的采样分布。重采样技术可以包括自举(bootstrapping)和交叉验证。自举可以通过从原始数据中采样和替换来执行,并将“未选择”的数据点作为测试例。可以通过将训练数据分成多个部分来执行交叉验证。

[0530] 子集选择可以鉴定与响应相关的预测器的子集。子集选择可以包括最佳子集选择、前向逐步选择、后向逐步选择、混合方法或其任何组合。在一些实施方案中,收缩拟合涉及所有预测变量的模型,但是相对于最小二乘估计,估计的系数向零收缩。这种收缩可以减少方差。收缩可以包括脊回归和套索(lasso)。降维可以将估计n+1个系数的问题简化为m+1个系数的更简单问题,其中m<n。它可以通过计算变量的n个不同的线性组合或投影来获得。然后将这n个投影用作预测变量,以通过最小二乘法拟合线性回归模型。降维可以包括主成分回归和偏最小二乘法。主成分回归可用于从一大组变量中导出一组低维特征。在主成分回归中使用的主成分可以在随后的正交方向上使用数据的线性组合来捕获数据中的最大方差。偏最小二乘可以是主成分回归的监督替代,因为偏最小二乘可以利用响应变量来鉴定新特征。

[0531] 非线性回归可以是回归分析的一种形式,其中观察数据由函数建模,该函数是模型参数的非线性组合并且取决于一个或多个自变量。非线性回归可以包括阶跃函数、分段函数、样条、广义加性模型或其任何组合。

[0532] 基于树的方法可用于回归和分类问题。回归和分类问题可能涉及将预测变量空间分层或分割为多个简单区域。基于树的方法可以包括装袋(bagging)、推进(boosting)、随机森林或其任何组合。装袋可以通过使用重复组合从原始数据集生成用于训练的额外数据来减少预测的方差,以产生与原始数据具有相同camality/大小的多步。推进可以使用几种不同的模型计算输出,然后使用加权平均方法对结果进行平均。随机森林算法可以抽取训练集的随机自举样本。支持向量机可以是分类技术。支持向量机可以包括找到以最大边际最好地分离两类点的超平面。支持向量机可以约束优化问题,使得边际被最大化,受到它完美分类数据的约束。

[0533] 无监督方法可以是从包含没有标记的响应的输入数据的数据集得出推论的方法。无监督方法可以包括聚类、主成分分析、k-均值聚类、层次聚类或其任何组合。

[0534] 所述质谱法可以是单等位基因质谱法。在一些实施方案中,该质谱法可以是MS分析、MS/MS分析、LC-MS/MS分析或其组合。在一些实施方案中,可以使用MS分析来确定完整肽的质量。例如,所述确定可以包括确定完整肽的质量(例如,MS分析)。在一些实施方案中,可以使用MS/MS分析来确定肽片段的质量。例如,所述确定可以包括确定肽片段的质量,后者可以用来确定肽或其部分的氨基酸序列(例如,MS/MS分析)。在一些实施方案中,可以使用肽片段的质量来确定该肽内的氨基酸序列。在一些实施方案中,可以使用LC-MS/MS分析来

分离复杂的肽混合物。例如,所述确定可以包括例如通过液相色谱法分离复杂的肽混合物,以及确定完整肽的质量、肽片段的质量或其组合(例如,LC-MS/MS分析)。该数据可用于例如肽测序。

[0535] 所述肽可以通过自噬由在细胞中表达的HLA蛋白呈递。自噬可允许细胞组分的有序降解和再循环。自噬可包括巨自噬、微自噬和分子伴侣介导的自噬。所述肽可以通过吞噬作用由在细胞中表达的HLA蛋白呈递。吞噬作用可以是用来去除病原体和细胞碎片的主要机制。例如,当巨噬细胞摄入病原微生物时,病原体陷入吞噬体中,然后吞噬体与溶酶体融合形成吞噬溶酶体。在HLA II类中,吞噬细胞如巨噬细胞和未成熟的树突细胞可以通过向吞噬体中的吞噬作用摄取实体——尽管B细胞表现出更普遍的向内体中的内吞作用——吞噬体与溶酶体融合,溶酶体的酸性酶将摄取的蛋白质裂解成许多不同的肽。

[0536] 可以通过使用多个质量度量来提高训练数据的质量。所述多个质量度量可包括常见污染物肽去除、高评分峰强度、高评分和高质量准确度。可以在进行评分之前使用评分的峰强度。MS/MS搜索首先使用简单过滤器针对候选序列筛选MS/MS谱。该过滤器可以是最小评分峰强度。一旦检查了足够数目的谱峰并且发现其不满足由该过滤器建立的阈值,则通过使候选序列被快速且概括地拒绝,使用评分峰强度可以提高搜索速度。评分峰强度可以是至少50%。评分峰强度可以是至少70%。评分峰值强度可以是至少10%、20%、30%、40%、50%、60%、70%、80%、90%或更大。在一些情况下,评分峰强度可以是至多90%、80%、70%、60%、50%、40%、30%、20%、10%或更小。评分可以至少为7。评分可以至少为1、2、3、4、5、6、7、8、9、10、15、20或更高。在一些情况下,评分可以是至多约20、15、10、9、8、7、6、5、4、3、2、1或更小。质量准确度可以是至多5ppm。质量准确度可以是至多10ppm、9ppm、8ppm、7ppm、6ppm、5ppm、4ppm、3ppm、2ppm、1ppm或更低。质量准确度可以是至少1ppm、2ppm、3ppm、4ppm、5ppm、6ppm、7ppm、8ppm、9ppm、10ppm或更高。

[0537] 在一些实施方案中,质量准确度至多为2ppm。在一些实施方案中,骨架切割评分至少为5。在一些实施方案中,骨架切割评分至少为8。

[0538] 由细胞中表达的HLA蛋白呈递的肽可以是由细胞中表达的单一免疫沉淀的HLA蛋白呈递的肽。免疫沉淀(IP)可以是使用与特定蛋白质特异性结合的抗体从溶液中沉淀蛋白质抗原的技术。该过程可以用来从含有数千种不同蛋白质的样品中分离并浓缩特定蛋白质。免疫沉淀可能需要在过程中的某个时间点将抗体偶联至固体基底上。

[0539] 由细胞中表达的HLA蛋白呈递的肽可以是由细胞中表达的单一外源HLA蛋白呈递的肽。可以通过将一种或多种外源肽引入细胞群体来产生单一外源HLA蛋白。在一些实施方案中,所述引入包括使所述细胞群体与所述一种或多种外源肽接触或在所述细胞群体中表达所述一种或多种外源肽。在一些实施方案中,所述引入包括使所述细胞群体与编码所述一种或多种外源肽的一种或多种核酸接触。在一些实施方案中,编码所述一种或多种肽的所述一种或多种核酸是DNA。在一些实施方案中,编码所述一种或多种肽的所述一种或多种核酸是RNA,任选地其中该RNA是mRNA。在一些实施方案中,所述富集不包括使用四聚体(或多聚体)试剂。

[0540] 由细胞中表达的HLA蛋白呈递的肽可以是由细胞中表达的单一重组HLA蛋白呈递的肽。该重组HLA蛋白可由重组HLA I类或HLA II类等位基因编码。所述HLA I类可以选自HLA-A、HLA-B、HLA-C。所述HLA I类可以是非经典的I-b类组。所述HLA I类可以选自HLA-E、

HLA-F和HLA-G。所述HLA I类可以是选自HLA-E、HLA-F和HLA-G的非经典I-b类组。在一些实施方案中,所述HLA II类包含HLA II类α链、HLA II类β链或其组合。

[0541] 所述多个预测变量可包括肽-HLA亲和力预测变量。所述多个预测变量可包括源蛋白质表达水平预测变量。源蛋白质表达水平可以是细胞内肽的源蛋白质的表达水平。在一些实施方案中,所述表达水平可以通过测量来源蛋白质的量或编码所述来源蛋白质的RNA的量来确定。所述多个预测变量可包括肽序列、氨基酸物理性质、肽物理性质、细胞内肽的来源蛋白质的表达水平、蛋白质稳定性、蛋白质翻译速率、泛素化位点、蛋白质降解速率、来自核糖体概况分析的翻译效率、蛋白质可切割性、蛋白质定位、促进TAP转运的宿主蛋白质的基序、经历自噬的宿主蛋白质、有利于核糖体停顿的基序(例如,聚脯氨酸或聚赖氨酸段)、有利于NMD的蛋白质特征(例如长3'UTR、最后一个外显子:外显子连接上游>50nt的终止密码子和肽可切割性)。

[0542] 所述多个预测变量可包括肽可切割性预测变量。肽可切割性可以与可切割连接体或切割序列相关。在一些实施方案中,该可切割连接体是核糖体跳跃位点或内部核糖体进入位点(IRES)元件。在一些实施方案中,当在细胞中表达时,该核糖体跳跃位点或IRES被切割。在一些实施方案中,该核糖体跳跃位点选自F2A、T2A、P2A和E2A。在一些实施方案中,该IRES元件选自常见的细胞或病毒IRES序列。切割序列,如F2A,或内部核糖体进入位点(IRES),可以放置在α链与β2-微球蛋白之间(HLA I类),或者α链与β链之间(HLA II类)。在一些实施方案中,单个HLA I类等位基因是HLA-A*02:01、HLA-A*23:01和HLA-B*14:02或HLA-E*01:01,而HLA II类等位基因是HLA-DRB*01:01、HLA-DRB*01:02和HLA-DRB*11:01、HLA-DRB*15:01或HLA-DRB*07:01。在一些实施方案中,切割序列是T2A、P2A、E2A或F2A序列。例如,切割序列可以是E G R G S L T C G D V E N P G P(T2A)、A T N F S L K Q A G D V E N P G P(P2A)、Q C T N Y A L K L A G D V E S N P G P(E2A)或V K Q T L N F D L K L A G D V E S N P G P(F2A)。

[0543] 在一些实施方案中,切割序列可以是凝血酶切割位点CLIP。

[0544] 由HLA蛋白呈递的肽可以包括通过搜索无酶特异性无修饰肽数据库而鉴定的肽。所述肽数据库可以是无酶特异性的肽数据库,如无修饰的数据库或有修饰(例如,磷酸化或半胱氨酸化)的数据库。在一些实施方案中,所述肽数据库是多肽数据库。在一些实施方案中,该多肽数据库可以是蛋白质数据库。在一些实施方案中,所述方法进一步包括使用反向数据库搜索策略搜索肽数据库。在一些实施方案中,所述方法进一步包括使用反向数据库搜索策略搜索蛋白质数据库。在一些实施方案中,进行从头搜索,例如,以发现不包含在正常肽或蛋白质数据库中的新肽。所述肽数据库可以通过以下方法生成:提供第一和第二细胞群体,每个细胞群体包含一个或多个包含亲和受体标记的HLA的细胞,其中该亲和受体标记的HLA的序列包含与亲和受体肽可操作地连接的由不同HLA等位基因编码的不同重组多肽;富集亲和受体标记的HLA-肽复合物;表征与来自该富集的亲和受体标记的HLA-肽复合物结合的肽或其部分;以及生成HLA等位基因特异性肽数据库。

[0545] 由HLA蛋白呈递的肽可包括通过将HLA-肽的MS/MS谱与肽数据库中的一种或多种HLA-肽的MS/MS谱进行比较而鉴定的肽。

[0546] 肽或编码肽的核酸上可存在突变。该突变可以选自点突变、剪接位点突变、移码突变、通读突变和基因融合突变。该点突变可以是其中从DNA或RNA序列中改变、插入或删除单

个核苷酸碱基的基因突变。该剪接位点突变可以是在前体信使RNA加工成成熟信使RNA过程中发生剪接的特定位点处插入、删除或改变许多核苷酸的基因突变。该移码突变可以是由不能被三整除的DNA序列中的许多核苷酸的插入缺失(插入或缺失)引起的基因突变。该突变还可包括插入、缺失、置换突变、基因重复、染色体易位和染色体倒位。

[0547] 在一些实施方案中,所述HLA II类蛋白包括HLA-DR蛋白。

[0548] 在一些实施方案中,所述HLA II类蛋白包括HLA-DP蛋白。

[0549] 在一些实施方案中,所述HLA II类蛋白包括HLA-DQ蛋白。

[0550] 在一些实施方案中,所述HLA II类蛋白可以选自HLA-DR和HLA-DP或HLA-DQ蛋白。在一些实施方案中,所述HLA蛋白是选自下组的HLA II类蛋白:HLA-DPB1*01:01/HLA-DPA1*01:03、HLA-DPB1*02:01/HLA-DPA1*01:03、HLA-DPB1*03:01/HLA-DPA1*01:03、HLA-DPB1*04:01/HLA-DPA1*01:03、HLA-DPB1*04:02/HLA-DPA1*01:03、HLA-DPB1*06:01/HLA-DPA1*01:03、HLA-DQB1*02:01/HLA-DQA1*05:01、HLA-DQB1*02:02/HLA-DQA1*02:01、HLA-DQB1*06:02/HLA-DQA1*01:02、HLA-DQB1*06:04/HLA-DQA1*01:02、HLA-DRB1*01:01、HLA-DRB1*01:02、HLA-DRB1*03:01、HLA-DRB1*03:02、HLA-DRB1*04:01、HLA-DRB1*04:02、HLA-DRB1*04:03、HLA-DRB1*04:04、HLA-DRB1*04:05、HLA-DRB1*04:07、HLA-DRB1*07:01、HLA-DRB1*08:01、HLA-DRB1*08:02、HLA-DRB1*08:03、HLA-DRB1*08:04、HLA-DRB1*09:01、HLA-DRB1*10:01、HLA-DRB1*11:01、HLA-DRB1*11:02、HLA-DRB1*11:04、HLA-DRB1*12:01、HLA-DRB1*12:02、HLA-DRB1*13:01、HLA-DRB1*13:02、HLA-DRB1*13:03、HLA-DRB1*14:01、HLA-DRB1*15:01、HLA-DRB1*15:02、HLA-DRB1*15:03、HLA-DRB1*16:01、HLA-DRB3*01:01、HLA-DRB3*02:02、HLA-DRB3*03:01、HLA-DRB4*01:01、HLA-DRB5*01:01)。由HLA蛋白呈递的肽可具有15-40个氨基酸的长度。由HLA蛋白呈递的肽可具有至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27个或更多个氨基酸的长度。在一些实施方案中,由HLA蛋白呈递的肽可具有至多30、29、28、27、26、25、24、23、22、21、20、19、18、17、16、15、14、13、12、11个或更少的氨基酸的长度。

[0551] 由HLA蛋白呈递的肽可包括通过以下方式鉴定的肽:(a)从表达单个HLA II类等位基因的细胞系中分离一种或多种HLA复合物;(b)从所述一种或多种分离的HLA复合物中分离一种或多种HLA-肽;(c)获得所述一种或多种分离的HLA-肽的MS/MS谱;以及(d)从肽数据库中获得与所述一种或多种分离的HLA-肽的MS/MS谱相对应的肽序列;其中从步骤(a、b、c)和(d)获得的一种或多种序列鉴定所述一种或多种分离的HLA-肽的序列。

[0552] 所述分离可包括从用亲和标记的HLA构建体转染或转导的细胞中分离HLA-肽复合物。在一些实施方案中,可以使用本领域已知的标准免疫沉淀技术与市售抗体分离复合物。可以先裂解细胞。可以使用HLA II类特异性抗体,如M5/114.15.2单克隆抗体,分离HLA II类-肽复合物。在一些实施方案中,单个(或一对)HLA等位基因被表达为具有肽标签的融合蛋白,并且使用识别该肽标签的结合分子分离HLA-肽复合物。

[0553] 所述分离可包括从HLA-肽复合物中分离肽并对所述肽进行测序。通过本领域技术人员已知的任何方法,如酸洗脱,从复合物中分离肽。尽管可以使用任何测序方法,但是在一些实施方案中,采用使用质谱法的方法,如液相色谱-质谱法(LC-MS或LC-MS/MS,或者HPLC-MS或HPLC-MS/MS)。这些测序方法可以是技术人员公知的,并且在Medzihradszky KF和Chalkley RJ.Mass Spectrom Rev.2015年1月-2月;34(1):43-63中进行了综述。

[0554] 适合于分离或纯化的另外的候选组分和分子可包括结合分子,如生物素(生物素-亲和素特异性结合对)、抗体、受体、配体、凝集素或包含固体支持物的分子,包括例如塑料或聚苯乙烯珠子、板或珠子、磁珠、测试条和膜。诸如阳离子交换色谱法的纯化方法可以用于通过电荷差异分离偶联物,其有效地将偶联物分离成各种分子量。通过阳离子交换色谱法获得的级分的含量可以使用常规方法通过分子量来鉴定,例如,质谱法、SDS-PAGE或用于通过分子量分离分子实体的其他已知方法。

[0555] 在一些实施方案中,所述方法进一步包括在表征之前从亲和受体标记的HLA-肽复合物中分离肽。在一些实施方案中,使用抗HLA抗体分离HLA-肽复合物。在一些情况下,使用抗HLA抗体分离具有或不具有亲和标签的HLA-肽复合物。在一些情况下,从细胞培养物的培养基中分离具有或不具有亲和标签的可溶性HLA(sHLA)。在一些情况下,使用抗HLA抗体分离具有或不具有亲和标签的可溶性HLA(sHLA)。例如,可以使用含有抗HLA抗体的珠子或柱子分离HLA,如具有或不具有亲和标签的可溶性HLA(sHLA)。在一些实施方案中,使用抗HLA抗体分离肽。在一些情况下,使用抗HLA抗体分离具有或不具有亲和标签的可溶性HLA(sHLA)。在一些情况下,使用含有抗HLA抗体的柱子分离具有或不具有亲和标签的可溶性HLA(sHLA)。在一些实施方案中,所述方法进一步包括从与亲和受体标记的HLA-肽复合物结合的肽的末端去除一个或多个氨基酸。

[0556] 所述个性化癌症疫苗可以进一步包含佐剂。例如,聚-ICLC——TLR3以及MDA5和RIG3的RNA解旋酶结构域的激动剂,已显示出疫苗佐剂的几种理想的性质。这些性质可包括诱导体内免疫细胞的局部和全身活化、产生刺激性趋化因子和细胞因子以及通过DC刺激抗原呈递。此外,聚-ICLC可以在人体中诱导持久的CD4+和CD8+应答。重要的是,在接种聚-ICLC的受试者和接受过高效、具有复制能力的黄热病疫苗的志愿者中可观察到转录和信号转导途径上调的惊人相似性。此外,在最近的1期研究中,>90%的(除Montanide外还)使用聚-ICLC联合NYESO-1肽疫苗免疫的卵巢癌患者显示出CD4+和CD8+ T细胞的诱导以及对肽的抗体应答。

[0557] 所述个性化癌症疫苗可以进一步包含免疫检查点抑制剂。免疫检查点抑制剂可以包括一种类型的药物,其阻断由某些类型的免疫系统细胞如T细胞和某些癌细胞产生的某些蛋白质。这些蛋白质有助于控制免疫应答,并可以防止T细胞杀死癌细胞。当这些蛋白质被阻断时,免疫系统上的“制动”被释放,T细胞能够更好地杀死癌细胞。在T细胞或癌细胞上发现的检查点蛋白质的示例包括PD-1/PD-L1和CTLA-4/B7-1/B7-2。一些免疫检查点抑制剂用于治疗癌症。

[0558] 训练数据还可以包括结构化数据、时间序列数据、非结构化数据和关系数据。非结构化数据可以包括音频数据、图像数据、视频、机械数据、电气数据、化学数据及其任何组合,用于准确地模拟或训练机器人或模拟。时间序列数据可以包括来自智能仪表、智能电器、智能设备、监控系统、遥测设备或传感器中的一个或多个的数据。关系数据包括来自客户系统、企业系统、操作系统、网站、网络可访问应用程序界面(API)或其任何组合的数据。这可以由用户通过将文件或其他数据格式输入到软件或系统中的任何方法来完成。

[0559] 训练数据可以上传到基于云的数据库。可以从运行基于机器学习的传感器信号处理算法的本地和/或远程计算机系统访问基于云的数据库。基于云的数据库和相关软件可用于归档电子数据、共享电子数据和分析电子数据。本地生成的数据或数据集可以上传到

基于云的数据库,从中可以访问并用于训练同一地点或不同地点的其他基于机器学习的检测系统。本地生成的传感器设备和系统测试结果可以上传到基于云的数据库,并用于实时更新训练数据集,以持续改进传感器设备和检测系统的测试性能。

[0560] 可以使用卷积神经网络来执行训练。卷积神经网络(CNN)在本文别处描述。卷积神经网络可以包含至少两个卷积层。卷积层的数目可以是1-10,扩张层的数目可以是0-10。卷积层(包括输入层和输出层)的总数可以是至少约1、2、3、4、5、10、15、20或更多,并且扩张层的总数可以是至少约1、2、3、4、5、10、15、20或更多。卷积层的总数可以是至多约20、15、10、5、4、3或更少,并且扩张层的总数可以是至多约20、15、10、5、4、3或更少。在一些实施方案中,卷积层的数目为1-10,并且完全连接层的数目为0-10。卷积层(包括输入和输出层)的总数可以是至少约1、2、3、4、5、10、15、20或更多,并且完全连接层的总数可以是至少约1、2、3、4、5、10、15、20或更多。卷积层的总数可以是至多约20、15、10、5、4、3或更少,并且完全连接层的总数可以是至多约20、15、10、5、4、3或更少。

[0561] 卷积神经网络可以包括至少一个批量归一化步骤。批量归一化层可以提高神经网络的性能和稳定性。批量归一化层可以为神经网络中的任何层提供作为零均值/单位方差的输入。批量归一化层的总数可以是至少约3、4、5、10、15、20或更多。批量归一化层的总数可以是至多约20、15、10、5、4、3或更少。

[0562] 卷积神经网络可以包括至少一个空间丢弃步骤。空间丢弃步骤的总数可以是至少约3、4、5、10、15、20或更多,并且空间丢弃步骤的总数可以是至多约20、15、10、5、4、3或更少。

[0563] 卷积神经网络可以包括至少一个全局最大池化步骤。全局池化层可以将一层神经元簇的输出组合成下一层中的单个神经元。例如,最大池化层可以使用来自前一层神经元簇中的每一个的最大值;而平均池化层可以使用来自前一层神经元簇中的每一个的平均值。卷积神经网络可以包括至少约1、2、3、4、5、10、15、20个或更多的全局最大池化步骤。卷积神经网络可以包括至多约20、15、10、5、4、3个或更少的全局最大池化步骤。

[0564] 卷积神经网络可包含至少一个密集层。卷积神经网络可包含至少约1、2、3、4、5、10、15、20个或更多个密集层。卷积神经网络可包含至多约20、15、10、5、4、3个或更少的密集层。

[0565] 治疗方法

[0566] 已经描述了使用肿瘤特异性肽的个性化免疫疗法。由恶性细胞内的遗传改变(例如,倒位、易位、缺失、错义突变、剪接位点突变等)引起的肿瘤新抗原代表最具肿瘤特异性的一类抗原。由于鉴定新抗原、选择优化的抗原和产生用于疫苗或免疫原性组合物的新抗原的技术困难,新抗原很少用于癌症疫苗或免疫原性组合物。有效地选择使用哪些特定肽作为免疫原需要具有预测哪些肿瘤特异性肽将有效地结合患者中存在的HLA等位基因并有效地呈递给患者的免疫系统以诱导抗肿瘤免疫的能力。开发治愈性且肿瘤特异性免疫疗法的关键障碍之一是鉴定和选择用以避免自身免疫的高度特异性且受限制的肿瘤抗原。这对于由MHC II类抗原呈递的用于免疫治疗的候选肿瘤特异性肽尤为重要,因为在MHC II类肽与免疫系统的结合和呈递中存在一定程度的混杂性。同时,不仅细胞毒性细胞的活化,而且CD4+ve记忆T细胞的活化也需要MHC II类呈递的肽。因此,为了在肿瘤保护中具有更高的有效性,需要MHC II类介导的免疫原性应答来提供强大的长期免疫原性。这些问题可以通过

以下方式解决:拥有可靠的肽-MHC预测算法,以及拥有用于测定和验证肽-MHC相互作用和免疫原性的可靠系统。因此,在一些实施方案中,可以通过以下方法产生高效且免疫原性的癌症疫苗:鉴定肿瘤中以DNA水平存在但不存于来自高比例癌症受试者的匹配种系样品中的瘤形成/肿瘤中的突变;用一种或多种肽-MHC结合预测算法分析所鉴定的突变,以鉴定哪种MHC(对于人类,为人类白细胞抗原或HLA)与高比例的患者HLA等位基因结合;以及合成多种选自所有新抗原肽和预测的结合肽的组的新抗原肽,以用于适用于治疗高比例的癌症受试者的癌症疫苗或免疫原性组合物。

[0567] 例如,将肽测序信息翻译成治疗性疫苗可以包括预测可以结合高比例个体的HLA肽的突变肽。有效地选择使用哪些特定突变作为免疫原需要具有预测哪些突变肽将有效地结合高比例的患者的HLA等位基因的能力。最近,采用经验证的结合和非结合肽的基于神经网络的学习方法提高了主要HLA-A和HLA-B等位基因的预测算法的准确性。然而,即使使用先进的基于神经网络的算法来编码HLA-肽结合规则,若干因素仍然限制了预测HLA等位基因上呈递的肽的能力。

[0568] 例如,将肽测序信息翻译成治疗性疫苗可以包括将药物配制为长肽的多表位疫苗。靶向实际上尽可能多的突变表位利用了免疫系统的巨大能力,通过下调免疫靶向基因产物来防止免疫逃逸的机会,并补偿表位预测方法的已知不准确性。合成肽提供了有效制备多种免疫原以及快速地将突变表位的鉴定转化为有效的疫苗的有用方法。肽可轻松化学合成且易于使用不含污染细菌或动物物质的试剂进行纯化。小尺寸允许清楚地聚焦蛋白质的突变区域并且还减少了与其他组分(未突变的蛋白质或病毒载体抗原)的不相关的抗原竞争。

[0569] 例如,将肽测序信息翻译成治疗性疫苗可以包括与强疫苗佐剂的组合。有效的疫苗可能需要强佐剂来引发免疫应答。例如,聚-ICLC——TLR3以及MDA5和RIG3的RNA解旋酶结构域的激动剂,已显示出疫苗佐剂的几种理想的性质。这些性质包括诱导体内免疫细胞的局部和全身活化、产生刺激性趋化因子和细胞因子以及通过DC刺激抗原呈递。此外,聚-ICLC可以在人体中诱导持久的CD4+和CD8+应答。重要的是,在接种聚-ICLC的受试者和接受过高效、具有复制能力的黄热病疫苗的志愿者中观察到转录和信号转导途径上调的惊人相似性。此外,在最近的1期研究中,>90%的(除Montanide外还)使用聚-ICLC联合NYESO-1肽疫苗免疫的卵巢癌患者显示出CD4+和CD8+ T细胞的诱导以及对肽的抗体应答。同时,聚-ICLC迄今已在超过25项临床试验中进行了广泛测试,并表现出相对良好的毒性概况。

[0570] 在一些实施方案中,可以从来自患有疾病或病况的受试者的细胞中鉴定免疫原性肽。在一些实施方案中,免疫原性肽可以对患有疾病或病况的受试者是特异性的。在一些实施方案中,免疫原性肽可以结合与患有疾病或病况的受试者的HLA单元型匹配的HLA。

[0571] 在一些实施方案中,肽文库可以在细胞中表达。在一些实施方案中,该细胞包含待鉴定或表征的肽。在一些实施方案中,待鉴定或表征的肽是内源肽。在一些实施方案中,该肽是外源肽。例如,可以从编码肽文库的多个序列表达待鉴定或表征的肽。

[0572] 在本说明书公开之前,大多数对HLA肽组进行的LC-MS/MS研究使用表达多个HLA肽的细胞,这需要使用现有的生物信息学预测器将肽指定给多达6个HLA I类等位基因中的1个或“解卷积”(Bassani-Sternberg和Gfeller,2016年)。因此,不能可信地将与已知基序不紧密匹配的肽报告为给定HLA等位基因的结合物。

[0573] 本文提供了预测可以结合个体的HLA肽的肽如突变肽的方法。在一些实施方案中，本申请提供了从给定的一组包含抗原的肽中鉴定最合适的肽以供制备用于受试者的免疫原性组合物的方法，所述方法包括从给定的一组肽中选择能够结合受试者的HLA蛋白的多个肽，其中所述结合HLA蛋白的能力通过用已经用肽序列数据库训练的机器分析肽的序列来确定，该肽序列数据库对应于所述受试者的每个HLA等位基因的特定HLA结合肽。本文提供了从给定的一组包含抗原的肽中鉴定最合适的肽以供制备用于受试者的免疫原性组合物的方法，所述方法包括从给定的一组肽中选择被确定为能够结合受试者的HLA蛋白的多个肽，结合HLA蛋白的能力通过用已经用肽序列数据库训练的机器分析肽的序列来确定，该肽序列数据库是通过进行本文以上描述的方法而获得的。因此，在一些实施方案中，本公开提供了鉴定用于制备受试者特异性免疫原性组合物的多个受试者特异性肽的方法，其中该受试者具有肿瘤，并且所述受试者特异性肽对该受试者和该受试者的肿瘤是特异性的，所述方法包括：对受试者的肿瘤样品和受试者的非肿瘤样品进行测序；基于核酸测序确定：在受试者的癌细胞基因组中存在但在受试者的正常组织中不存在的非沉默突变，以及受试者的HLA基因型；以及从鉴定的非沉默突变中选择多个受试者特异性肽，每个肽具有不同的肿瘤表位，该肿瘤表位对受试者的肿瘤是特异的，并且每个肽被鉴定为能够结合受试者的HLA蛋白，如通过在本文所述的预测HLA结合的方法中分析源自非沉默突变的肽的序列所确定的。

[0574] 在一些实施方案中，本文公开了一种表征对个体特异的HLA-肽复合物的方法。

[0575] 在一些实施方案中，表征个体特异性HLA-肽复合物的方法用来在有需要的个体如患有病况或疾病的个体中开发免疫治疗剂。

[0576] 本文提供了一种在哺乳动物中提供抗肿瘤免疫力的方法，其包括向该哺乳动物施用包含编码根据所述方法鉴定的肽的序列的多核酸。本文提供了一种在哺乳动物中提供抗肿瘤免疫力的方法，其包括向该哺乳动物施用有效量的具有根据本文所述方法鉴定的肽序列的肽。本文提供了一种在哺乳动物中提供抗肿瘤免疫力的方法，其包括向该哺乳动物施用包含肽的细胞，该肽包含根据本文所述方法鉴定的肽的序列。本文提供了一种在哺乳动物中提供抗肿瘤免疫力的方法，其包括向该哺乳动物施用包含多核酸的细胞，该多核酸包含编码包含根据本文所述方法鉴定的肽序列的肽的序列。在一些实施方案中，该细胞将该肽以HLA-肽复合物的形式呈递。

[0577] 本文提供了一种治疗受试者的疾病或病症的方法，该方法包括向该受试者施用多核酸，该多核酸包含编码根据本文所述的方法鉴定的肽的序列。本文提供了一种治疗受试者的疾病或病症的方法，该方法包括向该受试者施用有效量的肽，该肽包含根据本文所述的方法鉴定的肽的序列。本文提供了一种治疗受试者的疾病或病症的方法，该方法包括向该受试者施用包含肽的细胞，该肽包含根据本文所述的方法鉴定的肽的序列。本文提供了一种治疗受试者的疾病或病症的方法，该方法包括向该受试者施用包含多核酸的细胞，该多核酸包含编码包含根据本文所述方法鉴定的肽序列的肽的序列。在一些实施方案中，所述疾病或病症为癌症。在一些实施方案中，该方法进一步包括向受试者施用免疫检查点抑制剂。

[0578] 在一些实施方案中，本文公开了通过表征HLA-肽复合物为有需要的个体开发免疫治疗剂的方法，该方法包括：a)提供来源于有需要的个体的细胞群体，其中该细胞群体的一

个或多个细胞包含多核酸,该多核酸包含编码亲和受体标记的HLA I类或HLA II类等位基因的序列,其中编码亲和受体标记的HLA等位基因的序列包含:i)编码重组HLA I类或HLA II类等位基因的序列,其可操作地连接至ii)编码亲和受体肽的序列;b)在所述细胞群体的一个或多个细胞中的至少一个细胞中表达亲和受体标记的HLA,从而在所述至少一个细胞中形成亲和受体标记的HLA-肽复合物;c)富集所述亲和受体标记的HLA-肽复合物;表征对有需要的个体特异的HLA-肽复合物;以及d)基于对有需要的个体特异的HLA-肽复合物开发免疫治疗剂;其中该个体患有疾病或病况。

[0579] 在一些实施方案中,所述免疫治疗剂是核酸或肽治疗剂。

[0580] 在一些实施方案中,所述方法包括将一种或多种肽引入细胞群体。在一些实施方案中,所述方法包括使所述细胞群体与所述一种或多种肽接触或在所述细胞群体中表达所述一种或多种肽。在一些实施方案中,所述方法包括使所述细胞群体与编码所述一种或多种肽的一种或多种核酸接触。

[0581] 在一些实施方案中,所述方法包括基于结合患者特异性HLA鉴定的肽开发免疫治疗剂。在一些实施方案中,所述细胞群体来源于有需要的个体。

[0582] 在一些实施方案中,所述方法包括在细胞群体中表达肽的文库。在一些实施方案中,所述方法包括表达亲和受体标记的HLA-肽复合物的文库。在一些实施方案中,该文库包含与疾病或病况相关的肽的文库。在一些实施方案中,该疾病或病况是癌症或传染原的感染或自身免疫性疾病。在一些实施方案中,所述方法包括将传染原或其部分引入细胞群体中的一个或多个细胞中。在一些实施方案中,所述方法包括从对有需要的个体特异的HLA-肽复合物中表征一种或多种肽,任选地,其中所述肽来自传染原或自身免疫性疾病的一种或多种靶蛋白。在一些实施方案中,所述方法包括从传染原或自身免疫性疾病的一种或多种靶蛋白表征肽的一个或多个区域。在一些实施方案中,所述方法包括从衍生自传染原或自身免疫性疾病的HLA-肽复合物中鉴定肽。

[0583] 在一些实施方案中,所述传染原是病原体。在一些实施方案中,该病原体是病毒、细菌或寄生虫。

[0584] 在一些实施方案中,所述病毒选自:BK病毒(BKV)、登革病毒(DENV-1、DENV-2、DENV-3、DENV-4、DENV-5)、巨细胞病毒(CMV)、乙型肝炎病毒(HBV)、丙型肝炎病毒(HCV)、EB病毒(EBV)、腺病毒、人类免疫缺陷病毒(HIV)、人类T细胞淋巴营养性病毒(HTLV-1)、流感病毒、RSV、HPV、狂犬病、腮腺炎风疹病毒、脊髓灰质炎病毒、黄热病、甲型肝炎、乙型肝炎、轮状病毒、水痘病毒、人乳头瘤病毒(HPV)、天花、带状疱疹及其组合。

[0585] 在一些实施方案中,所述细菌选自:克雷伯氏菌属的种、惠普尔吸收障碍菌、麻风分枝杆菌、Mycobacterium lepromatosis和结核分枝杆菌。在一些实施方案中,所述细菌选自:伤寒、肺炎球菌、脑膜炎球菌、嗜血杆菌B、炭疽、破伤风类毒素、脑膜炎球菌B组、卡介苗(bcg)、霍乱及其组合。

[0586] 在一些实施方案中,所述寄生虫是蠕虫或原生动物。在一些实施方案中,所述寄生虫选自:利什曼原虫属的种(例如,硕大利什曼原虫(L.major)、婴儿利什曼原虫(L.infantum)、巴西利什曼原虫(L.braziliensis)、杜氏利什曼原虫(L.donovani)、恰氏利什曼原虫(L.chagasi)、墨西哥利什曼原虫(L.mexicana))、疟原虫属的种(例如,恶性疟原虫(P.falciparum)、间日疟原虫(P.vivax)、卵形疟原虫(P.ovale)、三日疟原虫

(P.malariae))、克氏锥虫、似蚓蛔线虫、鞭形鞭虫、美洲板口线虫和血吸虫属的种(曼森血吸虫(S.mansoni)、埃及血吸虫(S.haematobium)、日本血吸虫(S.japonicum))。

[0587] 在一些实施方案中,所述免疫治疗剂是工程化受体。在一些实施方案中,该工程化受体是嵌合抗原受体(CAR)、T细胞受体(TCR)或B细胞受体(BCR)、过继性T细胞治疗(ACT)或其衍生物。在其他方面,该工程化受体是嵌合抗原受体(CAR)。在一些方面,该CAR是第一代CAR。在其他方面,该CAR是第二代CAR。在另外其他方面,该CAR是第三代CAR。

[0588] 在一些方面,所述CAR包含细胞外部分、跨膜部分和细胞内部分。在一些方面,该细胞内部分包含至少一个T细胞共刺激域。在一些方面,该T细胞共刺激域选自CD27、CD28、TNFRS9(4-1BB)、TNFRSF4(OX40)、TNFRSF8(CD30)、CD40LG(CD40L)、ICOS、ITGB2(LFA-1)、CD2、CD7、KLRC2(NKG2C)、TNFRS18(GITR)、TNFRSF14(HVEM)或其任意组合。

[0589] 在一些方面,所述工程化受体结合靶标。在一些方面,该结合对于从表征对患有疾病或病况的个体特异的HLA-肽复合物的方法鉴定的肽是特异性的。

[0590] 在一些方面,所述免疫治疗剂是如本文所详述的细胞。在一些方面,该免疫治疗剂是包含特异性结合从表征对患有疾病或病况的个体特异的HLA-肽复合物的方法鉴定的肽的受体的细胞。在一些方面,该免疫治疗剂是与本发明所述的肽/核酸联合使用的细胞。在一些实施方案中,该细胞是患者细胞。在一些实施方案中,该细胞是T细胞。在一些实施方案中,该细胞是肿瘤浸润淋巴细胞。

[0591] 在一些方面,基于受试者的T细胞受体组库(repertoire)治疗患有病况或疾病的受试者。在一些实施方案中,基于受试者的T细胞受体组库选择抗原疫苗。在一些实施方案中,用表达对使用本文所述方法鉴定的抗原或肽具有特异性的TCR的T细胞治疗受试者。在一些实施方案中,用使用本文所述方法鉴定的对TCR(例如受试者特异性TCR)具有特异性的抗原或肽治疗受试者。在一些实施方案中,用使用本文所述方法鉴定的对表达TCR(例如,受试者特异性TCR)的T细胞具有特异性的抗原或肽治疗受试者。在一些实施方案中,用使用本文所述方法鉴定的对受试者特异性TCR具有特异性的抗原或肽治疗受试者。

[0592] 在一些实施方案中,基于在受试者中鉴定的TCR选择免疫原性抗原组合物或疫苗。在一个实施方案中,使用T细胞组库的鉴定和在功能分析中的测试来确定待施用于患有病况或疾病的受试者的免疫原性组合物或疫苗。在一些实施方案中,该免疫原性组合物是抗原疫苗。在一些实施方案中,该抗原疫苗包含受试者特异性抗原肽。在一些实施方案中,基于与抗原结合的受试者特异性TCR的定量来选择待包含在抗原疫苗中的抗原肽。在一些实施方案中,基于肽与TCR的结合亲和力选择抗原肽。在一些实施方案中,该选择基于量和结合亲和力的组合。例如,在功能分析中与抗原强力结合但在TCR组库中没有高度代表性的TCR可能是抗原疫苗的良好候选者,因为表达TCR的T细胞将有利地扩增。

[0593] 在一些实施方案中,基于与TCR的结合来选择供施用于受试者的抗原。在一些实施方案中,可以扩充T细胞,如来自患有疾病或病况的受试者的T细胞。经扩充的表达对使用本文所述方法鉴定的免疫原性抗原肽具有特异性的TCR的T细胞可以施用回受试者。在一些实施方案中,用多核苷酸转导或转染合适的细胞,例如,PBMC,以用于表达对使用本文所述方法鉴定的免疫原性抗原肽具有特异性的TCR,并将其施用于受试者。可以扩充表达对使用本文所述方法鉴定的免疫原性抗原肽具有特异性的TCR的T细胞并将其施用回受试者。在一些实施方案中,可以扩充在与自体病变组织一起孵育时产生细胞溶解活性且表达对使用本文

所述方法鉴定的免疫原性抗原肽具有特异性的TCR的T细胞,并将其施用于受试者。在一些实施方案中,可以扩充在功能分析中使用的导致与使用本文所述方法鉴定的免疫原性抗原肽结合的T细胞,并将其施用于受试者。在一些实施方案中,可以在T细胞中表达已经确定与使用本文所述方法鉴定的受试者特异性免疫原性抗原肽结合的TCR并将其施用于受试者。

[0594] 本文所述的方法可涉及对选定抗原如肿瘤或病原体相关抗原具有特异性的免疫系统细胞如T细胞的过继转移。可以采用各种策略,通过改变T细胞受体(TCR)的特异性,例如通过向新的TCRα和β链引入对使用本文所述方法鉴定的免疫原性抗原肽的特异性,来遗传修饰T细胞(参见,例如,美国专利8,697,854;PCT专利公开:WO2003020763、WO2004033685、WO2004044004、WO2005114215、WO2006000830、WO2008038002、WO2008039818、WO2004074322、WO2005113595、WO2006125962、WO2013166321、WO2013039889、WO2014018863、WO2014083173;美国专利8,088,379)。

[0595] 可以使用嵌合抗原受体(CAR)生成对所选靶标具有特异性的免疫应答细胞,如T细胞,该靶标例如是使用本文所述的方法鉴定的免疫原性抗原肽,该细胞具有多种受体嵌合构建体(参见,例如,美国专利5,843,728;5,851,828;5,912,170;6,004,811;6,284,240;6,392,013;6,410,014;6,753,162;8,211,422;以及PCT公开WO9215322)。可替代的CAR构建体可以被表征为属于连续的世代。第一代CAR通常由对抗原具有特异性的抗体的单链可变片段组成,例如,包含与特异性抗体的VH连接的VL,其通过柔性连接体,例如通过CD8a铰链结构域和CD8a跨膜结构域,连接至CD3ζ或FcRy或scFv-FcRy的跨膜和胞内信号传导域(参见,例如,美国专利7,741,465;美国专利5,912,172;美国专利5,906,936)。第二代CAR将一个或多个共刺激分子如CD28、OX40(CD134)或4-1BB(CD137)的胞内结构域并入内结构域内,例如scFv-CD28/OX40/4-1BB-CD3(参见,例如,美国专利8,911,993;8,916,381;8,975,071;9,101,584;9,102,760;9,102,761)。第三代CAR包括共刺激性内域的组合,如CD3C-链、CD97、GDI la-CD18、CD2、ICOS、CD27、CD154、CDS、OX40、4-1BB或CD28信号域,例如scFv-CD28-4-1BB-CD3C或scFv-CD28-OX40-CD3Q(参见,例如,美国专利8,906,682;美国专利8,399,645;美国专利5,686,281;PCT公开WO2014134165;PCT公开WO2012079000)。在一些实施方案中,可以通过在抗原特异性T细胞中表达CAR来协调共刺激,该细胞被选择为在例如与专职抗原呈递细胞上的抗原相互作用后被激活并扩充。可以在免疫应答性细胞上提供其他工程化受体,例如,以改善T细胞攻击的靶向和/或使副作用最小化。

[0596] 可使用替代技术转化目标免疫应答性细胞,如原生质体融合、脂质转染、转染或电穿孔。可以使用各种各样的载体,如逆转录病毒载体、慢病毒载体、腺病毒载体、腺相关病毒载体、质粒或转座子,如睡美人(Sleeping Beauty)转座子(参见美国专利6,489,458;7,148,203;7,160,682;7,985,739;和5,243,337),可用来引入CAR,例如使用通过CD3ζ和CD28或CD137的第二代抗原特异性CAR信号传导。例如,病毒载体可以包括基于HIV、SV40、EBV、HSV或BPV的载体。

[0597] 例如,作为转化目标的细胞可以包括T细胞、自然杀伤(NK)细胞、细胞毒性T淋巴细胞(CTL)、调节性T细胞、人胚胎干细胞、肿瘤浸润淋巴细胞(TIL)或可以由其分化出淋巴样细胞的多能干细胞。表达期望的CAR的T细胞可以例如通过与γ-辐射的活化和繁殖细胞(APC)共培养来选择,所述细胞共表达癌症抗原和共刺激分子。例如可以通过在可溶性因子如IL-2和IL-21的存在下在APC上共培养来扩充工程化CAR T细胞。例如,可以进行这种扩充

以提供记忆性CAR T细胞(例如,其可以通过非酶数字阵列和/或多组流式细胞术来测定)。以这种方式,可以提供对带有抗原的肿瘤具有特异性细胞毒活性的CAR T细胞(任选地与所需趋化因子如干扰素-γ的产生相结合)。这种CAR T细胞可以例如在动物模型中使用,例如威胁肿瘤异种移植物。

[0598] 诸如前述的方法可以适合于提供治疗患有疾病如瘤形成或病原性感染的受试者和/或增加其生存期的方法,例如通过施用有效量的包含结合选定抗原的抗原识别受体的免疫应答性细胞,其中所述结合激活免疫应答性细胞,从而治疗或预防该疾病(如瘤形成、病原体感染、自身免疫性疾病或同种异体移植反应)。在CAR T细胞疗法中的给药可以例如包括以10^6至10^9个细胞/kg的剂量施用,进行或不进行淋巴消耗过程,例如使用环磷酰胺。

[0599] 为了防止可能的不良反应,工程化免疫应答性细胞可以配备转基因安全开关,其形式为使细胞易于暴露于特定信号的转基因。例如,单纯疱疹病毒胸苷激酶(TK)基因可以以这种方式使用,例如通过在干细胞移植后引入作为供体淋巴细胞输注使用的同种异体T淋巴细胞中。在此类细胞中,施用诸如更昔洛韦或阿昔洛韦等核苷前药会导致细胞死亡。备选的安全开关构建体包括可诱导的胱天蛋白酶9,例如通过施用将两个无功能icasp9分子聚集在一起形成活性酶的小分子二聚体而触发。已经描述了用于实施细胞增殖控制的多种替代方法(参见,例如,美国专利公开20130071414;PCT专利公开WO2011146862;PCT专利公开WO2014011987;PCT专利公开WO2013040371)。在过继疗法的进一步改进中,可以使用基因组编辑使免疫应答性细胞适应替代实施方式,例如提供编辑的CAR T细胞。

[0600] 细胞治疗方法还可以涉及T细胞的离体活化和扩充。在一些实施方案中,T细胞可以在施用于有需要的受试者之前被激活。这些类型的治疗的实例包括使用肿瘤浸润淋巴细胞(TIL)(参见美国专利5,126,132)、细胞毒性T细胞(参见美国专利6,255,073和美国专利5,846,827)、扩充的肿瘤引流淋巴结细胞(参见美国专利6,251,385)和各种其他淋巴细胞制品(参见美国专利6,194,207;美国专利5,443,983;美国专利6,040,177;和美国专利5,766,920)。

[0601] 离体激活的T细胞群体可以处于最大程度地协调对癌症、感染性疾病或其他疾病状态例如自身免疫性疾病状态的免疫应答的状态。为了激活,可以将至少两个信号传递至T细胞。通常,第一个信号通常通过T细胞表面的T细胞受体(TCR)传递。TCR第一信号通常在TCR与抗原呈递细胞(APC)表面上与MHC复合物一起表达的肽抗原相互作用时触发。第二个信号通常通过T细胞表面的共刺激受体传递。共刺激受体通常由APC表面上表达的相应配体或细胞因子触发。

[0602] 预期可以获得对使用本文所述方法鉴定的免疫原性抗原肽具有特异性的T细胞,并将其用于治疗或预防疾病的方法中。就此而言,本公开提供了一种治疗或预防受试者的疾病或病况的方法,其包括以有效治疗或预防受试者的疾病的量向该受试者施用包含对使用本文所述方法鉴定的免疫原性抗原肽具有特异性的细胞的细胞群体。在一些实施方案中,治疗或预防受试者的疾病的方法包括以有效治疗或预防哺乳动物的癌症的量向受试者施用富含疾病反应性T细胞的细胞群体。所述细胞可以是与所述受试者同种异体或自体的细胞。

[0603] 本公开进一步提供了一种通过向受试者施用抗原肽或疫苗在受试者中诱导疾病特异性免疫应答、针对疾病进行疫苗接种、治疗和/或减轻疾病症状的方法。

[0604] 本公开的肽或组合物可以以足以诱导CTL应答的量施用。抗原肽或疫苗组合物可以单独施用或与其他治疗剂联合施用。示例性治疗剂包括但不限于化学治疗剂或生物治疗剂、放射或免疫疗法。可以施用对特定疾病合适的任何治疗性治疗。化学治疗剂和生物治疗剂的实例包括但不限于阿地白介素、六甲蜜胺、氨磷汀、天冬酰胺酶、博来霉素、卡培他滨、卡铂、卡莫司汀、克拉屈滨、西沙必利、顺铂、环磷酰胺、阿糖胞苷、达卡巴嗪(DTIC)、更生霉素、多西他赛、多柔比星、屈大麻酚、红细胞生成素α、依托泊苷、非格司亭、氟达拉滨、氟尿嘧啶、吉西他滨、格拉司琼、羟基脲、伊达比星、异环磷酰胺、干扰素α、伊立替康、兰索拉唑、左旋咪唑、醛氢叶酸、甲地孕酮、美司钠、甲氨蝶呤、甲氧氯普胺、丝裂霉素、米托坦、米托蒽醌、奥美拉唑、昂丹司琼、紫杉醇(Taxol®)、毛果芸香碱、丙氯拉嗪、利妥昔单抗、他莫昔芬、泰素、盐酸拓扑替康、曲妥珠单抗、长春碱、长春新碱和酒石酸长春瑞滨。另外,可以进一步向受试者施用抗免疫抑制剂或免疫刺激剂。例如,可以进一步向受试者施用抗CTLA抗体或抗PD-1或抗PD-L1。

[0605] 将要包含在疫苗组合物中的每种肽的量和给药方案可以由本领域技术人员确定。例如,可以制备肽或其变化形式用于静脉内(i.v.)注射、皮下(s.c.)注射、皮内(i.d.)注射、腹膜内(i.p.)注射、肌肉内(i.m.)注射。肽注射的示例性方法包括s.c.、i.d.、i.p.、i.m.和i.v.。DNA注射的示例性方法包括i.d.、i.m.、s.c.、i.p.和i.v.。疫苗组合物的其他给药方法是本领域技术人员已知的。

[0606] 可以调整药物组合物,使得该组合物中存在的肽的选择、数目和/或量是疾病和/或患者特异性的。例如,可以由亲本蛋白质在给定组织中的表达模式指导肽的确切选择,以避免副作用。该选择可以取决于疾病的具体类型、疾病的状态、早先的治疗方案、患者的免疫状态以及患者的HLA单元型。此外,根据特定患者的个人需求,本公开的疫苗可以含有个性化的组分。实例包括根据特定患者中相关抗原的表达、由于个人变态反应或其他治疗导致的不良副作用以及在第一轮治疗或治疗方案后对二次治疗的调整来改变肽的量。

[0607] 计算机控制系统

[0608] 本公开提供了计算机控制系统,其被编程用于实现本公开的方法。图10显示了计算机系统(1001),其被编程或以其他方式配置用于训练机器学习HLA-肽呈递预测模型。计算机系统(1001)可以调节本公开的各个方面,例如输入氨基酸位置信息,将估算的信息传输到数据集,以及用数据集生成经训练的算法。计算机系统(1001)可以是用户电子设备或远程计算机系统。该电子设备可以是移动电子设备。

[0609] 计算机系统(1001)包括中央处理单元(CPU,本文中也称为“处理器”和“计算机处理器”)(1005),中央处理单元(1005)可以是单核或多核处理器,通过顺序处理或并行处理。计算机系统(1001)还包括存储器单元或设备(1010)(例如,随机存取存储器、只读存储器、闪速存储器)、存储单元(1015)(例如,硬盘)、用于与一个或多个其他系统通信的通信接口(1020)(例如,网络适配器)和外围设备(1025),其为外部的或内部的或两者,如打印机、监视器、USB驱动器和/或CD-ROM驱动器。存储器(1010)、存储单元(1015)、接口(1020)和外围设备(1025)通过诸如主板的通信总线(实线)与CPU(1005)通信。存储单元(1015)可以是用于存储数据的数据存储单元(或数据存储库)。计算机系统(1001)可以借助于通信接口(1020)可操作地耦合至计算机网络(“网络”)(1030)。网络(1030)可以是因特网、互联网和/或外联网,或者与因特网通信的内联网和/或外联网。网络(1030)在一些情况下是电信和/

或数据网络。网络(1030)可以包括一个或多个计算机服务器,其能够实现支持分布式计算的对等网络。在一些情况下,借助于计算机系统(1001),网络(1030)可以实现客户端-服务器结构,这可以使得耦合至计算机系统(1001)的设备能够起到客户端或服务器的作用。

[0610] CPU(1005)可以执行一系列可以在程序或软件中体现的机器可读指令。所述指令可以存储在存储器(1010)中。所述指令可被导向CPU(1005),其随后可对CPU(1005)进行编程或以其他方式进行配置,以实现本公开的方法。由CPU(1005)执行的操作的实例可以包括获取、解码、执行和写回。

[0611] CPU(1005)可以是电路如集成电路的一部分。系统(1001)中的一个或多个其他组件可被包括在该电路中。在一些情况下,该电路是专用集成电路(ASIC)。

[0612] 存储单元(1015)可以存储文件,如驱动程序、文库和保存的程序。存储单元(1015)可以存储用户数据,例如,用户偏好和用户程序。在一些情况下,计算机系统(1001)可以包括位于计算机系统(1001)外部(诸如位于通过内联网或因特网与计算机系统(1001)通信的远程服务器上)的一个或多个附加数据存储单元。

[0613] 计算机系统(1001)可以通过网络(1030)与一个或多个远程计算机系统通信。例如,计算机系统(1001)可与远程计算机系统或用户进行通信。远程计算机系统的实例包括个人计算机(例如,便携式PC)、平板或平板PC(例如,Apple®iPad、Samsung®Galaxy Tab)、电话、智能电话(例如,Apple®iPhone、支持Android的设备、Blackberry®)或个人数字助理。用户可以通过网络(1030)访问计算机系统(1001)。

[0614] 如本文所述的方法可通过存储在计算机系统(1001)的电子存储位置上,例如存储器(1010)或数据存储单元(1015)中的机器(例如,计算机处理器)可执行代码来实现。该机器可执行代码或机器可读代码可以以软件的形式提供。在使用期间,该代码可以由处理器(1005)执行。在一些情况下,该代码可从存储单元(1015)中检索并存储在存储器(1010)中,以备处理器(1005)访问。在一些情况下,可以不包括存储单元(1015),而将机器可执行指令存储在存储器(1010)中。

[0615] 可将该代码预编译并配置用于与具有适于执行该代码的处理器的机器一起使用,或者可以在运行过程中对其进行编译。该代码可以以编程语言的形式提供,该编程语言可以被选择为使得该代码能够以预编译或实时编译的方式执行。

[0616] 本文提供的系统和方法的各方面,如计算机系统(1001),可以在编程中体现。该技术的各个方面可以被认为是“产品”或“制品”,其一般为在某种类型的机器可读介质中携带或体现的机器(或处理器)可执行代码和/或关联数据的形式。机器可执行代码可以存储在存储单元上,如硬盘上,或存储器(例如,只读存储器、随机存取存储器、闪速存储器)中。“存储”型介质可以包括计算机、处理器等的任何或全部有形存储器,或其相关模块,如各种半导体存储器、磁带驱动器、磁盘驱动器等,其可以随时为软件编程提供非暂时性存储。软件的全部或部分可以不时地通过因特网或各种其他电信网络进行通信。例如,这样的通信可以使软件能够从一台计算机或处理器加载到另一台计算机或处理器,例如,从管理服务器或主机加载到应用服务器的计算机平台。因此,可以承载软件元件的另一类型的介质包括光波、电波和电磁波,诸如跨越本地设备之间的物理接口、通过有线和光学陆线网络以及经由各种空中链路所使用的。携带这类波的物理元件,如有线或无线链路、光学链路等,也可

以被认为是承载软件的介质。除非局限于非暂时性有形“存储”介质,否则如本文所用的诸如计算机或机器“可读介质”等术语是指参与向处理器提供指令以供执行的任何介质。

[0617] 因此,机器可读介质,如计算机可执行代码,可以采取许多形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括,例如,光盘或磁盘,如任何计算机中的任何存储设备等,例如可用来实现附图中所示的数据库等。易失性存储介质包括动态存储器,如这样的计算机平台的主存储器。有形传输介质包括同轴电缆、铜线和光纤,包括构成计算机系统内的总线的导线。载波传输介质可以采取电信号或电磁信号或者声波或光波的形式,诸如在射频(RF)和红外线(IR)数据通信期间生成的那些信号或波。因此,计算机可读介质的常见形式包括,例如:软盘、柔性盘、硬盘、磁带、其他任何磁性介质、CD-ROM、DVD或DVD-ROM、其他任何光学介质、穿孔卡片纸带、其他任何具有孔洞图案的物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、其他任何存储器芯片或匣盒、传输数据或指令的载波、传输这类载波的线缆或链路,或者计算机可以从中读取编程代码和/或数据的其他任何介质。这些计算机可读介质形式中的许多形式可以参与将一个或多个指令的一个或多个序列运载到处理器以供执行。

[0618] 计算机系统(1001)可包括电子显示器(1035)或与电子显示器(1035)通信,电子显示器(1035)包括用户界面(UI)(1040),以提供例如由受试者癌细胞的II类MHC等位基因编码的一种或多种蛋白质将呈递所鉴定的肽序列的给定序列的概率。UI的实例包括但不限于图形用户界面(GUI)和基于网络的用户界面。

[0619] 本公开的方法和系统可以通过一个或多个算法来实现。算法可以通过软件在由中央处理单元(1005)执行时实现。例如,该算法可以输入氨基酸位置信息,将估算的信息传输到数据集,并使用数据集生成经训练的算法。

[0620] 实施例

[0621] 以下提供的实施例仅用于说明目的,并非限制本文提供的权利要求的范围。

[0622] 实施例1.HLA II类结合预测器性能

[0623] 在该实施例中,通过随机改组命中肽以1:19(命中:诱饵)的比率生成的、包含观察到的质谱肽和诱饵肽的验证数据集用于分析结合预测器neonmhc2(NEON)和NetMHCIIpan的性能(图4)。对于NEON结合预测器,为显示的每个MHC II等位基因构建了单独的模型。条形的高度显示阳性预测值(PPV)。在预测该等位基因时,等位基因按模型的性能排序。与NetMHCIIpan相比,NEON结合预测器在所有等位基因中显示出更高的PPV。

[0624] 在该实施例中,还测试了SPI阈值对结合预测器验证的影响(图5)。显示了当在具有不同评分峰值强度(SPI)截止值的成组肽上进行训练/验证时,HLA II类结合预测器的性能。使用了不同的SPI截止条件:使用观察到的大于或等于70SPI的MS命中肽对数据集进行训练和评估,对大于或等于50SPI的肽进行训练并对大于或等于70SPI的肽进行验证,以及对大于或等于50SPI的肽进行训练和验证。

[0625] 该实施例中,收集了35个HLA-DR等位基因的数据,这些等位基因具有>95%的HLA-DR群体覆盖率(美国等位基因频率),以显示使用大于或等于70的评分峰值强度(SPI)截止值通过LC-MS/MS进行的等位基因概况分析观察到的肽的数目(图6)。

[0626] 在一个示例性设置中,将模型PPV分析应用于测试迄今为止为neonmhc2程序生成的每个II类等位基因的分区数据。测试分区数据由作为MS观察到的II类结合物的阳性示例

(例如命中样品肽)和作为阳性示例的乱序版本的阴性示例(例如诱饵样品肽)组成。命中:诱饵比率保持为1:19,例如,对于每个阳性样品,包括19个阴性样品(即5%的阳性样品),并执行测试分区以进行验证。PPV评分是通过选择测试分区中评分最高的5%的肽并询问其中哪些部分为阳性来生成的。结果示于图7A中。

[0627] 观察到对于收集的HLA-DR等位基因,当训练集大小增加时,PPV的值也增加(图7B)。

[0628] 该实施例中,处理相关变量进一步改进了预测(图8)。在训练数据分区上,逻辑回归可能适合于使用结合强度(NetMHCIIpan或Neon预测器)和处理特征(RNA-Seq表达和衍生的基因水平偏项)来预测HLA II类呈递。在单独的评估分区中,与MS观察到的MHC II肽重叠的外显子位置("命中")与MS中未观察到的随机外显子位置一起评分(1:499比率)。通常,具有处理相关变量的Neon比NetMHCIIpan、Neon的预测器和具有处理相关变量的NetMHCIIpan显示出更高的PPV。

[0629] 实施例2.神经网络架构

[0630] 在该实施例中,使用神经网络来获得训练算法(图9)。输入肽被表示为20聚体,较短的肽用"缺失"字符填充。每个肽都有31维的嵌入,因此向神经网络中的输入是20x31矩阵。在被神经网络处理之前,根据训练集中的特征值平均值和标准偏差对20x31矩阵进行特征归一化。第一个卷积层具有9个氨基酸和50个过滤器的内核(也称为通道),具有ReLU激活函数。接着是批量归一化,然后是空间丢弃,丢弃率为20%。接着是另一个卷积层,其具有3的内核和20个过滤器,以及ReLU激活函数,然后再次是批量归一化和空间丢弃,丢弃率为20%。然后应用全局最大池化,在20个过滤器中的每个过滤器中获取最大激活的神经元,然后使用S形激活函数使这20个值进入具有单个神经元的完全连接的(密集)层。该输出被视为结合/非结合预测。将L2正则化应用于第一卷积层、第二卷积层和密集层的权重,权重分别为0.05、0.1和0.01。

[0631] 实施例3.用于单等位基因MHC II类配体概况分析的可扩展方案

[0632] 目前关于MHC II类结合基序的知识可以基于两种体外结合试验,一种使用细胞MHC计算EC50,另一种使用纯化的MHC计算IC50。领先的HLA II类预测算法NetMHCIIpan专门在这些数据上进行训练。

[0633] 有限数目的人类HLA II类等位基因目前得到已确认结合肽(亲和力<100nM)的超过200个实例的支持(图12E),它们几乎都是15聚体。这些实验仅涵盖最常见的高加索人HLA-DR等位基因,对非高加索人群体特定的等位基因(例如,HLA-DRB1*15:02)覆盖范围有限,并且几乎没有覆盖常见的HLA-DP和HLA-DQ等位基因。目前的HLA II类预测性能,即使在常见的高加索人等位基因上,也明显落后于MHC I类的准确性;ROC曲线仅略好于随机曲线。

[0634] 考虑到这些限制,本文开发了一种新型生物技术,称为通过标记等位基因捕获的单等位基因捕获(MAPTAC™),它能够有效分离HLA II类结合肽,该肽结合由单个等位基因编码的MHC蛋白,用于基于MS的鉴定(图11A和11B);这种方法也适用于HLA I类。当应用于HLA II类时,所选等位基因的α和β链在基因构建体上编码,将生物素接受肽(BAP)序列置于β链的C末端。然后裂解这些细胞并与BirA酶一起孵育,以对捕获等位基因的β链的C末端进行生物素化。NeutrAvidin下拉纯化MHC结合的肽的群体,这些肽进一步通过大小排阻分离,并使用一流的LC-MS/MS方案进行测序。

[0635] 在一些实施方案中,使用高场不对称波形离子迁移光谱法(FAIMS)来评估LC-MS/MS分析。在一些实施方案中,肽在通过nLC-MS/MS分析之前经历酸性反相(aRP)和碱性反相(bRP)离线分级分离。

[0636] 两天的转染足以实现构建体的稳健表达(图12B),在三种不同的细胞系(expi293、A375和B721)中对四个不同的等位基因进行适当的细胞表面定位(图12C)。

[0637] 由于HLA-DRA在功能上是不变的,即使捕获β链与内源性α链配对,这种方法也能实现单等位基因分辨率。这意味着该方法可用于对HLA-DR等位基因进行概况分析,而不管给定细胞系中预先存在的HLA基因型和表达水平如何。

[0638] 对于HLA-DP和HLA-DQ,α和β链都是可变的,并且都有助于肽结合,因此只有当天然α链不表达或天然等位基因是纯合的并与捕获等位基因匹配时,才能预期单等位基因分辨率。或者,可以使用仅β链捕获来建立与天然α链相对应的肽的背景。

[0639] 进行概况分析的等位基因包括五个HLA-DR等位基因(DRB1*03:01、DRB1*09:01、DRB1*11:01、DRB3*01:01和DRB3*02:02)以及一个HLA-DP等位基因(DPB1*01:01/DPA1*01:03)、一个HLA-DQ等位基因(DQB1*06:02/DQA1*01:02)和两个I类等位基因(表1)。在所有情况下,2-3次重复足以观察到至少1500个独特的肽(图11B)。在进行概况分析的等位基因中,只有一小部分命中对应于已知污染物或完美的胰蛋白酶解物;另一方面,模拟转染返回的肽相对较少,这些肽大多可被鉴定为已知污染物或完美的胰蛋白酶解物(图11B)。

[0640] 表1显示了示例性实验中使用的样品的总结。

[0641] 表1

[0642]

供体	DR1	DR3/4/5
D001000763	DRB1*03:01, DRB1*11:01	DRB3*01:01, DRB3*02:02
HD84	DRB1*01:01	不存在
RG1248	DRB1*03:01	DRB3*01:01

[0643]

RG1095	DRB1*03:01, DRB1*11:01	DRB3*01:01, DRB3*02:02
RG1104	DRB1*01:01, DRB1*11:01	DRB3*02:02
2010113472	DRB1*01:01, DRB1*11:01	DRB3*02:02
2010113438	DRB1*03:01, DRB1*11:01	DRB3*01:01, DRB3*02:02

[0644] 由于MHC II结合肽的末端不需要拟合在MHC结合沟内,如果多个不同的肽种类共享相同的核心结合序列,则它们可以同样好地结合。当肽与重叠序列合并为“嵌套组”时,每个HLA II类等位基因观察到500-700个独特的嵌套组;这些通常来自500-600个独特的基因。HLA I类和HLA II类结合肽的长度分布与先前使用基于抗体的下拉的MS研究中观察到的那些相匹配(图11C)。

[0645] 在推定的MHC结合肽中,大多数氨基酸的表示水平与其来源蛋白质组频率一致。例外情况包括半胱氨酸、甲硫氨酸和色氨酸,它们已耗尽,这与之前基于MS的MHC II肽研究一致。在来自IEDB的等位基因匹配的高亲和力肽(<50nM)中未观察到半胱氨酸、甲硫氨酸和色氨酸的消耗;然而,就蛋白质组而言,IEDB肽确实显示出富含亮氨酸和甲硫氨酸,以及脯氨酸、天冬氨酸和谷氨酸的消耗。

[0646] 实施例4.MAPTAC™方案揭示了已知的和新颖的MHC II结合基序

[0647] 由于II类肽的MHC结合子序列不在相对于N末端或C末端的固定位置,因此准确的II类基序发现必须动态地考虑每个结合肽的不同结合寄存器(binding register)的可能性。Gibb's Cluster工具通过期望最大化(EM)算法解决了这一挑战。探索了使用卷积神经网络(CNN)的新基序发现方法的使用。CNN在计算机视觉领域已经取得了成功,该领域同样寻求实现平移不变的模式识别。对CNN进行训练以区分MHC结合肽与它们自己的乱序形式,然后根据在倒数第二个网络层中实现最大节点激活的子序列对齐阳性示例。当应用于单等位基因MS数据时,该方法产生与Gibbs聚类一致的基序,并在相对位置1、4、6和9处显示锚(图13)。这些基序与针对来自IEDB的高亲和力结合物观察到的CNN衍生基序高度一致(亲和力<50nM;图13)。对于DRB1*11:01,进一步验证了该基序跨细胞系是稳定的,并且与先前使用泛DR抗体分析的DRB1*11:01纯合细胞系一致。类似地,源自MHC I类等位基因的基序与来自基于亲和力的方法和先前基于MS的研究的基序一致(图14A)。

[0648] 尽管所有MHC II类等位基因都显示出可区分的基序,但锚位置处的熵明显高于MHC I类等位基因所观察到的熵。因此,定义了每个MHC II类等位基因的每个锚位置处的优选氨基酸,并观察到只有10-20%的肽在所有四个锚位置表现出理想的残基,多达60%的肽表现出两个或更少的预期锚(图14B和图30C)。使用NetMHCIIpan对13-17聚体的结合潜力进行评分,虽然MS观察到的肽在所有情况下都针对预测的结合潜力进行了富集,但与长度匹配的随机肽的评分存在显著重叠(图14C和图36A)。

[0649] 实施例5.在单等位基因MHC II MS数据上训练的算法预测免疫原性

[0650] 接下来,考虑来自单等位基因MS平台的数据是否可以产生改进的MHC II类结合预测器。在CNN方法的基础上,创建了具有过滤器大小、跳跃连接和总感受野的多层网络(图31A)。为了训练和评估这个称为neonmhc2的深度学习模型,将蛋白质组分为三个分区,分别代表75%、12.5%和12.5%的基因。第一个分区用于通过随机梯度下降训练CNN,第二个分区用于架构和超参数优化。第三个分区仅用于在分析结束时评估一次性能。为了确保评估的完整性,小心地将所有基因放在同一分区中的旁系同源基因分组中。

[0651] 由于MS表现出一定程度的残基偏好,特别是针对半胱氨酸(图12D),这个问题通过使用通过随机排列阳性示例的序列而生成的阴性训练示例(称为诱饵)来缓解。由于这种方法存在学习天然蛋白质序列性质的风险,这可能会人为地提高预测性能,因此模型评估采用了独特的诱饵生成策略,其中诱饵从未观察到的肽源基因的子序列中随机采样。以1:19

命中:诱饵比计算阳性预测值(PPV)表明,neonmhc2相对于NetMHCIIpan在预测评估分区中的MS肽方面具有改进的PPV(图4和图31B)。人为地对neonmhc2训练数据集的大小进行降采样的实验表明,其性能受数据限制,并且会随着更深的覆盖数据而提高(图16)。

[0652] 探索了neonmhc2预测结合亲和力的能力,这是训练NetMHCIIpan的数据类型。为了剥夺NetMHCpan对相同肽测量进行训练和评估的好处,使用存储到IEDB的稍旧版本的NetMHCIIpan评分肽进行评估。使用Kendall Tau统计来评估预测准确性,在所有情况下,NetMHCIIpan评分与基于MS的预测器相似或略好(图15B)。有趣的是,性能取决于所进行的亲和力测定的类型。neonmhc2在Sette和同事预测亲和力测量值时适度落后于NetMHCIIpan,而在Buus和同事预测测量值时更明显落后于NetMHCIIpan。综合考虑这些结果,这些平台之间似乎存在内在差异,但尚不清楚哪种方法更正确。

[0653] 为了达到改善的清晰度,评估了预测自然CD4 T细胞应答的能力。来自IEDB的数据通常不适合此目的,因为响应的等位基因限制几乎总是未定义或估算的。因此,组装了四聚体引导的表位定位(TGEM)数据的大型数据集。这些研究都使用了全面的重叠肽筛选而不是预测优先级,从而消除了有利于NetMHCIIpan的观察偏差。同时,等位基因限制是明确的。对于有足够数据进行评估的所有等位基因,neonmhc2的表现明显优于NetMHCpan,后者的表现仅略好于随机。因此,MAPTAC™平台可能是用于鉴定免疫原性MHC II类表位的训练模型的最佳平台。

[0654] 实施例6.在多等位基因MS数据上训练的算法较差

[0655] 鉴于公共领域中有许多基于标准泛DR和泛II抗体纯化的多等位基因II类数据库,测试是否可以仅使用多等位基因数据训练合适的预测器。几个小组在从多等位基因I类数据中去卷积MHC I类等位基因基序方面取得了成功,尽管这些努力尚未转化为公开可用的预测器。由于需要同时解析每个肽的结合寄存器和簇成员,II类基序的去卷积也变得更加复杂。虽然Gibbs Cluster工具已被用于探索II类去卷积的可能性,但这种方法的保真度尚未得到广泛验证。

[0656] 为了评估II类去卷积的准确性,选择了具有已知基因型的可公开获得的泛DR数据集。对于每个数据集,为供体基因型中的每个等位基因(1-2个DR1等位基因加上0-2个DR3/4/5等位基因,取决于单元型和接合性)掺加了我们的单等位基因数据的20个肽。在每个数据集上运行Gibbs聚类工具,并根据其已知的来源等位基因观察掺加肽是否适当地共聚类。在该分析的早期版本中,要么固定等位基因数的簇数,要么允许Gibbs簇自动确定最佳簇数;然而,这两种方法似乎都不能准确地将肽去卷积。为了给算法提供帮助,通过计算掺加肽的真实源等位基因与其分配的簇之间的调整后的交互信息来选择最佳簇计数。尽管如此,在除了几种情况之外的所有情况下,肽在不同的簇中分布,而不论它们的源等位基因如何(图17A)。这些结果表明,当前的去卷积方案对于MHC II类可能不是可靠准确的。

[0657] 该分析的一个警告是一些肽可能能够结合一个以上的等位基因。与此一致,下一个问题是源自多等位基因数据的结合基序是否仍然可以合理地匹配从单等位基因数据观察到的那些。为了评估这一点,选择与每个单个等位基因的捕获肽具有最佳对应性的簇,并建立基于这些群体的基序(参见例如图17B)。许多基序清楚地展示了一些已知的锚,而其他位置与单等位基因基序不一致或源数据集之间不一致。此外,也有明显的出现假锚的情况。最后,我们评估了去卷积的数据是否可用于训练可以预测我们单等位基因数据集评估分区

中的肽的CNN。在去卷积的多等位基因数据上训练的模型在所有情况下都达不到MAPTAC™训练的模型(图17C)。

[0658] 实施例7.源蛋白质特征影响呈递可能性

[0659] 对于MHC I类,蛋白酶体在确定所呈递的表位的组库中起着重要作用;因此,对蛋白质到肽的加工如何塑造II类组库进行了表征。

[0660] 首先,重点关注在几个基于组织的肽概况分析数据集中观察到的MHC II类肽的N端和C端的确切位置。相对于诱饵肽比较基于位置的氨基酸频率,观察到显著的富集和消耗。这种模式与最近的观察一致。有趣的是,整体模式与组织蛋白酶S的已知切割偏好([RPI][FMLW][KQTR][ALS])不匹配,这是最佳表征的II类加工酶。

[0661] 为了确定该基序的预测潜力,构建了用于N端和C端的基于NN的预测器,并拟合使用两个切割变量以及预测的结合潜力(每个MS训练的CNN)的逻辑回归,以区分来自从相同源基因采样的长度匹配的诱饵肽的真实MS肽。

[0662] 与单独考虑结合潜力的模型相比,该预测器在肽预测方面提供了适度的改进;然而,由于MHC II类结合表位(可互换地称为II类表位)的免疫原性可能不取决于肽切割的确切位置,因此问题是,如果切割的确切位点未知,该模型是否仍会增加价值。因此,第二次运行该预测方案,保留命中和诱饵的确切切割位置,而是对估算的结合核心附近(+/-15个氨基酸)的蛋白质位置的复合可切割性评分进行评分。有趣的是,与仅结合的预测器相比,性能没有提高。这些结果与之前的工作一致,这表明增加II类切割预测可以提高对MS观察到的配体的预测,但不能提高对T细胞识别的预测,这对确切的肽末端而言可能是不可知的。

[0663] 提出了一种模型,其中MHC II肽的显著部分在MHC结合后从它们的N端和C端“回缩”。在该模型下,出现倒数第二个脯氨酸特征,这是因为脯氨酸阻止了外肽酶的进程。在这种情况下,源自直接分析MHC配体末端的基序可能会错误引导,因为它反映了下游编辑,而不是肽片段生成的初始步骤。因此,在II类肽附近确定了其他序列特征,这些特征可能能够解释它们的产生。首先,搜索了经典的组织蛋白酶S特征,但在MS观察到的II类肽与从肽源基因采样的长度匹配的诱饵肽附近的组织蛋白酶S位点没有富集。由于这种处理特征可能反映了复杂的酶集合,因此基于观察到的肽和诱饵周围的上游和下游蛋白质环境(+-25个氨基酸)训练了从头(de novo)CNN。

[0664] 考虑了第三种模型,其中肽的可用性由蛋白质的折叠或半解折叠状态而不是其一级序列决定。基于同源性的ACCPRO用于预测二级结构和溶剂可及性区域,并使用一组预测器来鉴定本质上无序的结构域。

[0665] 如果加工偏好的区域本质上难以预测,则可能简单地构建由先前公布的多等位基因II类MS数据的大集合中至少一个肽覆盖的所有蛋白质区域的目录,并使用重叠作为预测特征。诚然,重叠特征受到结合信息的污染,因为先前公布的数据中表示的等位基因可能具有相同或相似的结合基序。尽管如此,即使这个特征也只是适度地改善了对呈递肽的预测,这表明MHC II类肽可能不受强加工热点的影响。

[0666] 下一个问题是哪些基因对II类结合肽组库的贡献最大。在预测MHC I类配体时,已经知道基因水平的特征如表达水平可以提供很大的提升。利用先前公布的分析人类组织的II类结合组库的MS数据集,观察到MS观察到的肽比随机诱饵肽(从蛋白质组采样)更高度表达,相差一个数目级(图18A)。尽管如此,注意到大约5%的II类肽映射到根据代表性RNA-

Seq数据表面上不表达的基因。基于这种模式,通过提出基线预期,即每个基因的观察数应与其长度和表达水平的乘积成比例,试图量化每个基因在II类肽组库中被过度表示或表示不足的程度(图18B)。在过度表示的基因中,人类组织血清中表达的蛋白质明显富集,其产生许多II类结合肽,但表面上不在天然组织中表达。这与MHC II类在呈递从细胞外环境取样的抗原中的已知作用是一致的。

[0667] 由于自噬是另一种成熟的II类加工途径,因此确定每个基因的观察肽与预期肽之比(不包括观察肽少于5个和预期肽少于5个的任何基因),并确定是否存在相对于已知自噬基因或通过敲除小鼠中的Atg5稳定的基因的物理配偶体的富集(图18C)。两个基因集似乎都没有在II类数据中富集;事实上,自噬基因的物理配偶体似乎适度地表示不足。

[0668] 纵观所有细胞定位(图18D和图18E),很少有区室被明确地过度表示或表示不足。两个最丰富的区室是细胞膜和溶酶体,每个产生的II类肽的数目大约是预期数目的两倍。目前尚不清楚膜蛋白的富集是否与膜再循环到自噬体或膜蛋白直接进入自噬途径的高尔基体路径有关。溶酶体蛋白的富集与先前观察到的自噬基因消耗之间的明显矛盾表明,这些趋势对所考虑的自噬相关基因的特定子集高度敏感。图18F显示了关于两种不同基因表达谱、成块肿瘤细胞和专职抗原呈递细胞的肽观察结果的相对一致性。

[0669] 实施例8.准确的MHC II预测需要了解内吞途径

[0670] 除了了解II类基因的来源途径外,了解哪些细胞类型负责大多数II类表达可能也很重要。在癌症的情况下,非专职APC,包括成纤维细胞和肿瘤本身,被认为在发炎的肿瘤微环境(TME)中呈递II类。为了获得进一步的了解,在最近发表的三个单细胞RNA-Seq数据集中分析了HLA-DRB1的表达,这些数据集对肺癌、头颈癌和黑素瘤进行了概况分析。在细胞之间取平均到患者细胞类型水平,显然经典APC(巨噬细胞、树突细胞和B细胞)比肿瘤和其他基质细胞类型呈递高得多的水平的II类,并且这种趋势在多个患者和肿瘤类型中是一致的。

[0671] 为了探究免疫疗法是否破坏了这一趋势,分析了来自检查点阻断反应性肿瘤类型的额外单细胞RNA-Seq,并在治疗前后评估了HLA-DRB1表达。包括一名确认为反应者的黑素瘤群组在治疗前和治疗后活检中均显示均一地很低的肿瘤细胞的HLA-DRB1表达(图19C)。对抗PD-1疗法显示出55%临床反应率的基底细胞癌群组同样表现出低的肿瘤细胞来源的HLA-DRB1表达,无论时间点如何(图19C)。

[0672] 这些结果表明,大多数肿瘤内HLA II类呈递主要由专职APC驱动,“热”TME条件并不能保证与一般模式的差异。

[0673] 由于肿瘤细胞在肿瘤微环境中的数目可能超过APC,因此它们较低的MHC II类表达水平可能是免疫学相关的。为了评估总体II类表达中有多少来自肿瘤细胞与基质,鉴定了在II类特异性基因(重点是CIITA、CD74和CTSS)中具有突变的TCGA患者,并且确定显示体细胞(肿瘤特异性)变异的RNA-Seq读取值的分数。该信息用于估算来源于肿瘤与基质的HLA-DRB1表达的比例(图19B)。基于在代表17种不同肿瘤类型的153名患者中鉴定的突变,观察到一种主要模式,其中大多数II类表达似乎来自非肿瘤细胞。仅关注具有最高水平T细胞浸润的患者(前10%,如使用先前公布的18基因特征(Ayers等人,2017)所确定的),低肿瘤HLA-DR表达似乎仍然是常态,16名患者中只有3名表达>1000TPM(肿瘤进展和转移)。

[0674] 为了探究免疫疗法是否破坏了这一趋势,分析了来自检查点阻断反应性肿瘤类型

的额外单细胞RNA-Seq,并在治疗前后评估了HLA-DRB1表达。包括一名确认为反应者的黑素瘤群组在治疗前和治疗后活检中均显示均一地很低的肿瘤细胞的HLA-DRB1表达(图19C)。对抗PD-1疗法显示55%临床反应率的基底细胞癌群组同样表现出低的肿瘤细胞来源的HLA-DRB1表达,无论时间点如何(图19C)。

[0675] 这些结果表明,大多数肿瘤内HLA II类呈递主要由专职APC驱动,“热”TME条件并不能保证与一般模式的差异。

[0676] 实施例9.新的预测概念能够更准确地鉴定免疫原性新抗原

[0677] 为了探索neonmhc2的效用和相关的处理规则,考虑了在几种预测场景中的性能。首先,在来自7名用泛DR抗体进行概况分析的健康供体的PMBC上评估了预测MS鉴定的肽的能力。这种分析可以控制MAPTAC™系统或我们的生产细胞系固有的任何系统偏差。使用1:499的命中与诱饵之比以及从蛋白质编码外显子组中随机对诱饵进行采样,评估了基于neonmhc2和NetMHCIIpan的模型以及包含附加处理特征的模型的阳性预测值(表达,根据图18B的基因水平偏差,以及与之前的MHC II肽的重叠)。这些模型证实了结合和处理预测的实质性改进(图20)。

[0678] 图21A示出了NetMHCIIpan和neonmhc2与所指示的其他处理参数或特征的比较。通过HLA-DR抗体分析八个MS样品的预测性能(在实施例6中分析的相同样品,图17A)。预测器最低限度地使用HLA结合预测(NetMHCIIpan或neonmhc2)并任选地使用额外的处理相关变量:基因表达、基因偏好(例如,根据图18B、图18C、图19B),以及与先前观察到的HLA-DQ肽的重叠。在该实施例中,诱饵从蛋白质组中随机采样(包括从未产生MS观察到的肽的基因),以达到1:499的命中与诱饵之比,这几乎使可用的诱饵序列饱和。以类似于图4的方式计算阳性预测值,例如,前0.2%的肽被判定为阳性,PPV是真正MS观察到的肽的阳性的分数。对于每个样品中的每个候选肽,结合评分被计算为样品基因型中存在的HLA-DR等位基因的最大值。尽管通过两种方法发现的肽的趋势存在相当大的相关性,但本文描述的模型显示出更可靠的结果。图21B(也参见图33B)表示使用与图21A中相同的命中:诱饵比和性能度量对由树突细胞(裂解物)呈递的肿瘤衍生肽的预测性能。显示了NetMHCIIpan和此处描述的模型的性能,其中使用和不使用处理功能。图21C显示了每个重标记的肽的表达水平和基因偏好评分。图21D是表示根据裂解物和UV处理实验,重标记的肽源基因的重叠的图示。

[0679] 实施例10.在细胞系中表达II类HLA肽并分离MHC-II结合的肽

[0680] 构建体设计、细胞培养和HLA肽免疫沉淀

[0681] 在该示例性研究中,通过将单个亲和标记的HLA构建体转染到细胞系(A375、HEK293T、Expi293、HeLa)中来生成单等位基因细胞系,并免疫沉淀亲和标记的HLA肽复合物。在图12A和12E中,MHC II类等位基因频率是从allelefrequencies.net/获得的等位基因频率,除非另有说明。美国人口的等位基因频率是通过假设62.3%的欧洲裔、13.3%的非洲裔、6.8%的亚洲裔和17.6%的西班牙裔的混合群体来估算的。

[0682] 关于图12A和图12E,mhc_ligand_full.csv数据集于2018年9月21日从IEDB数据下载。要求有效的亲和力测量具有等于“细胞MHC/竞争性/荧光”、“细胞MHC/竞争性/放射性”、“细胞MHC/直接/荧光”、“纯化的MHC/竞争性/荧光”、“纯化的MHC/竞争性/放射性”或“纯化的MHC/直接/荧光”的“方法/技术”,和等于“解离常数KD”、“解离常数KD(~EC50)”、“解离常数KD(~IC50)”,“半数最大有效浓度(EC50)”或“半数最大抑制浓度(IC50)”的“测定组”。如

果字符串“Buus”出现在“作者”字段中,则测量归于Søren Buus小组(丹麦哥本哈根大学)。否则,如果作者字段包含字符串“Sette”或“Sidney”,则测量归于Alessandro Sette小组(La Jolla Institute for Immunology,U.S.A)。所有其他测量值都标记为“其他”。为了列举强结合物,仅对测得的亲和力强于50nM的肽进行计数(图12A)。图12E包括来自tools.iedb.org/main/datasets/的额外数据,并列举了亲和力<100nM的强结合物。

[0683] DNA构建体设计

[0684] HLA I类和HLA II类等位基因的基因序列通过IPD-IMGT/HLA网页(ebi.ac.uk/ipd/imgt/hla)确定,并用于设计重组表达构建体。对于HLA I类,α链与C端GSGGSGGSAGG连接体融合,然后是生物素接受肽(BAP)标签序列GLNDIFEAQKIEWHE、终止密码子和可变DNA条形码,并通过NcoI和XbaI限制位点克隆到pSF Lenti载体(Oxford Genetics,Oxford,UK)中。HLA II类构建体类似地通过NcoI和XbaI限制位点克隆到pSF Lenti中,并包含在C端与来自I类构建体的连接体-BAP序列(SGGSGGSAGGGLNDIFEAQKIEWHE)融合的β链序列,然后是另一个短GSG连接体、F2A核糖体跳跃序列(VKQTLNFDLLKLAGDVESNPGP)、α链序列、HA标签(GSYPYDVPDYA)、终止密码子和可变DNA条形码。通过Sanger测序验证所有DNA序列的身份。

[0685] 细胞培养和瞬时转染

[0686] Expi293细胞(Thermo Scientific)在含有8%CO_2的Expi293培养基(Thermo Scientific)中在37℃下以125rpm振荡生长。Expi293细胞保持在0.5x10^6/mL至6x10^6/mL之间的细胞密度,每两周定期传代。30mL Expi293细胞悬液用于以约3x10^6/mL的细胞密度和>90%的活力的瞬时转染。简而言之,将30ug DNA(每毫升细胞悬浮液1μg/mL DNA)稀释到一个管中的1.5mL Opti-MEM培养基(Thermo Scientific)中,同时将80μL ExpiFectamine™ 293转染试剂(Thermo Scientific)稀释到含有1.5mL Opti-MEM的第二管中。这两个管在室温下孵育5分钟,合并,轻轻混合,在室温下孵育30分钟。将DNA和ExpiFectamine混合物加入Expi293细胞,并在37℃、8%CO_2、80%相对湿度下孵育。48小时后,以每管50x10^6个细胞在四次技术重复中收获转染的细胞,离心,用1x Gibco DPBS(Thermo Scientific)洗涤一次,并在液氮中快速冷冻以用于质谱分析。从每个转染批次中收集1x10^6个细胞的等分试样,并通过抗BAP(Rockland Immunochemicals Inc.,Limerick,PA)或抗HA(Bio-Rad,Hercules,CA)Western印迹进行分析,以验证亲和标记的HLA蛋白表达。

[0687] A375细胞(ATCC)在含有10%FBS的DMEM中生长,并以不超过80%的汇合度保持在培养物中,定期传代。对于质谱实验,A375细胞在500cm^2板中以18.5x10^6个细胞/mL的接种密度在100mL中培养,根据70%汇合的细胞数计算。24小时后,按照针对总培养体积调整的TransIT系统方案,用TransIT-X2(Mirus Bio,Madison,WI)转染细胞。48小时后,吸出细胞培养基,并用1X Gibco DPBS(Thermo Scientific)洗涤细胞。为了收获,将A375细胞在37℃下与30mL非酶促细胞解离溶液(Sigma-Aldrich)孵育10分钟,离心,用1X DPBS洗涤,并以每个样品50x10^6个细胞等分。293T和HeLa细胞购自ATCC,并分别在37℃、5%CO_2下在DMEM、10%FBS、2mM L-谷氨酰胺或DMEM+10%FBS中培养。使用TransIT LT1试剂(Mirus Bio,Madison,WI)按照制造商的说明用HLA构建体转染两种细胞系,并在转染后48小时如针对A375细胞所述进行处理。从所有样品中,从每次转染中收集1x10^6个细胞的等分试样,并通过抗BAP(Rockland Immunochemicals Inc.,Limerick,PA)或抗HA(Bio-Rad,Hercules,CA)Western印迹法进行分析,以验证亲和标记的HLA蛋白表达。

[0688] BirA蛋白表达和纯化

[0689] 使用编码与C端六组氨酸标签融合的大肠杆菌BirA的pET19载体。化学感受态大肠杆菌BL21(DE3)细胞(New England Biolabs)用BirA表达质粒转化,在37℃下在LB肉汤加100μg/ml氨苄青霉素中生长至OD_{600}为0.6-0.8,并冷却至30℃,之后通过添加0.4mM异丙基-β-D-硫代吡喃半乳糖苷诱导表达。大肠杆菌细胞在30℃下继续生长4小时。通过在4℃下以8000xg离心30分钟收获大肠杆菌细胞,并在-80℃下储存直至使用。将表达重组BirA的冷冻细胞沉淀物重新悬浮在含有5mM咪唑的IMAC缓冲液(50mM NaH_2PO_4 pH 8.0,300mM NaCl)中,与1mg/ml溶菌酶一起在冰上孵育20分钟,然后通过超声处理裂解。通过在4℃下以16,000xg离心30分钟去除细胞碎片和不溶性物质。随后使用ÄKTA纯色谱系统(GE Healthcare)将澄清的上清液加载到HisTrap HP 5mL柱上,用IMAC缓冲液加25mM和50mM咪唑洗涤,然后用500mM咪唑洗脱。将含有BirA的级分合并,并用25mM NaCl对20mM Tris-HCl pH 8.0进行透析,并加载到HiTrap Q HP 5mL柱(GE Healthcare)上,并通过应用25至600mM NaCl的线性梯度来洗脱。合并含有高纯度BirA的级分,在储存缓冲液(20mM Tris-HCl pH 8.0 100mM NaCl,5%甘油)中进行缓冲液交换,并浓缩至约5-10mg/mL,等分,并在液氮中快速冷冻以在-80℃下储存。使用计算的消光系数ε=47,440$M^{-1}cm^{-1}$,通过紫外光谱法在OD_{280nm}处计算BirA蛋白浓度。

[0690] Western印迹分析方案

[0691] 将样品加入XT样品缓冲液和XT还原剂(Bio-Rad,Hercules,CA),在95℃下加热5分钟,然后将对应于约100,000个细胞的体积加入10%Criterion XT Bis-Tris凝胶(Bio-Rad)中,并使用PowerPac Basic Power Supply(Bio-Rad,Hercules,CA)和XT MES电泳缓冲液(Bio-Rad,Hercules,CA)在200V下电泳35分钟。用水简单冲洗凝胶,然后在iBlot2 Gel Transfer Device(Thermo Scientific)上使用设置P3将蛋白质转移到Invitrogen iBlot Transfer Stacks(Thermo Fisher Scientific)内的PVDF膜上。Precision Plus Protein All Blue Standard(Bio-Rad,Hercules,CA)用于监测分子量。接下来,将膜用TBS Tween 20(TBST)缓冲液(25mM Tris,0.15mM NaCl,0.05%(v/v)Tween 20,pH 7.5)洗涤3x5分钟,在室温下在TBST-M(含有5%(w/v)脱脂速溶奶粉的TBST)中封闭1小时,然后在TBST-B(含有5%(w/v)牛血清白蛋白(Sigma Aldrich)的TBST]和1:5,000稀释的兔抗β微管蛋白抗体(目录号ab6046,Abcam)和兔抗生物素连接酶表位标签抗体(目录号100-401-B21,Rockland Immunochemicals)中孵育过夜。接下来,将膜用TBST洗涤3x5分钟,在含有1:10,000稀释的山羊抗兔IgG(H+L-辣根过氧化物酶偶联抗体(目录号170-6515,Bio-Rad,Hercules,CA)的TBST-M中室温孵育1小时,然后在室温下用TBST洗涤3x5分钟。最后,将膜用Pierce ECL Western Blotting Substrate(Thermo Fisher Scientific,Rockford,IL)冲洗,使用ChemiDoc XRS+Imager(Bio-Rad)显色,并使用Image Lab软件(Bio-Rad)进行可视化。

[0692] 亲和标记的HLA-肽复合物的分离

[0693] 从表达BAP标记的HLA等位基因的细胞和仅表达无BAP标签的内源性HLA-肽复合物的阴性对照细胞系进行亲和标记的HLA-肽复合物的分离。在用于HLA-肽亲和纯化之前,将NeutrAvidin珠状琼脂糖树脂用1mL冷PBS洗涤3次。将含有50x10^6个表达BAP标记的HLA肽的细胞的冷冻沉淀物在冰上解冻20分钟,然后通过在1.2mL冷裂解缓冲液[20mM Tris-Cl pH 8,100mM NaCl,6mM $MgCl_2$,1.5%(v/v)Triton X-100,60mM辛基葡糖苷、0.2mM 2-碘乙酰

胺、1mM EDTA pH 8,1mM PMSF,1X完全不含EDTA的蛋白酶抑制剂混合物(Roche,Basel,Switzerland)]中手工移液轻轻裂解。将裂解物与>250单位的Benzonase核酸酶(Sigma-Aldrich)在4℃下翻滚(end/over/end)孵育15分钟,以降解DNA/RNA,并在4℃下以15,000xg离心20分钟以去除细胞碎片和不溶性物质。将澄清的上清液转移到新管中,并通过在室温下在1.5mL管中与0.56μM生物素、1mM ATP和3μM BirA翻滚孵育10分钟来生物素化BAP标记的HLA肽。将上清液在4℃下以相当于200μL的Pierce高容量NeutrAvidin珠状琼脂糖树脂(Thermo Scientific)浆液的体积在4℃下翻滚孵育30分钟,以亲和富集生物素化的HLA-肽复合物。最后,将HLA结合的树脂用1mL冷洗涤缓冲液(20mM Tris-Cl pH 8,100mM NaCl,60mM辛基葡糖苷、0.2mM 2-碘乙酰胺、1mM EDTA pH8)洗涤四次,然后用1mL的冷10mM Tris-Cl pH8洗涤四次。在洗涤之间,手工轻轻混合HLA结合的树脂,然后通过在4℃下以1,500xg离心一分钟来沉淀。将洗涤过的HLA结合树脂储存在-80℃或立即进行HLA肽洗脱和脱盐。

[0694] 基于抗体的HLA-肽复合物分离

[0695] HLA II类DR-肽复合物从健康供体外周血单核细胞(PBMC)中分离出来。将相当于75μL GammaBind Plus Sepharose树脂的体积用1mL冷PBS洗涤3次,与10μg抗体在4℃下孵育过夜,然后用1mL冷PBS洗涤3次,然后用于HLA肽免疫沉淀。将含有$50x10^6$个细胞的冷冻PBMC沉淀物在冰上解冻20分钟,然后通过在1.2mL冷裂解缓冲液[20mM Tris-Cl pH 8,100mM NaCl,6mM MgCl2,1.5%(v/v)Triton X-100,60mM辛基葡糖苷,0.2mM 2-碘乙酰胺,1mM EDTA pH8,1mM PMSF,1X完全不含EDTA的蛋白酶抑制剂混合物(Roche,Basel,Switzerland)]中轻轻裂解。将裂解物在4℃下与>250单位的Benzonase核酸酶(Sigma-Aldrich)一起孵育15分钟以降解DNA/RNA,并在4℃下以15,000xg离心20分钟以去除细胞碎片和不溶性物质。然后将上清液与结合至GammaBind Plus Sepharose树脂(GE Life Sciences)的抗HLA DR抗体(TAL 1B5,产品号sc-53319;Santa Cruz Biotechnology,Dallas,TX)在4℃下翻滚孵育3小时,以免疫沉淀HLA DR-肽复合物。最后,将HLA结合的树脂用1mL冷洗涤缓冲液(20mM Tris-Cl pH 8,100mM NaCl,60mM辛基葡糖苷、0.2mM 2-碘乙酰胺、1mM EDTA pH8)洗涤四次,然后用1mL冷的10mM Tris-ClpH8洗涤四次。在洗涤之间,轻轻混合HLA结合的树脂,然后通过在4℃下以1,500xg离心1分钟沉淀。将洗涤过的HLA结合的树脂储存在-80℃或立即进行HLA-肽洗脱和脱盐。

[0696] HLA-肽洗脱和脱盐

[0697] 将HLA肽从亲和标记的内源性HLA复合物中洗脱,同时使用Sep-Pak(Waters,Milford,MA)固相萃取系统脱盐。简而言之,将Sep-Pak Vac 1cc(50mg)37-55μm粒径tC18小柱连接到24位提取歧管(Restek),用200μL MeOH活化两次,然后用100μL 50%(v/v)ACN/1%(v/v)FA,然后用500μL 1%(v/v)FA洗涤四次。为了将HLA肽与亲和标记的HLA肽解离并促进肽与tC18固相的结合,将400μL的3%(v/v)ACN/5%(v/v)FA添加到含有HLA-结合的珠状琼脂糖树脂的管中。通过移液混合浆液,然后转移到Sep-Pak柱中。用1%(v/v)FA(2x200μL)冲洗试管和移液器吸头,并将冲洗液转移到柱中。将100fmol的Pierce Peptide Retention Time Calibration(PRTC)混合物(Thermo Scientific)添加到柱中作为上样对照。将珠状琼脂糖树脂与200μL 10%(v/v)AcOH孵育两次,每次五分钟,以进一步将HLA肽与亲和标记的HLA肽解离,然后用500μL 1%(v/v))FA洗涤四次。HLA-肽从tC18洗脱到新的1.5mL微管(Sarstedt)中,方法是用250μL的15%(v/v)ACN/1%(v/v)FA,然后用2x250μL的

30％(v/v)乙腈/1％(v/v)FA进行分步分级分离。用于活化、上样、洗涤和洗脱的溶液通过重力流动,但使用真空(≤-2.5PSI)从柱中去除剩余的洗脱液。将含有HLA-肽的洗脱液冷冻,通过真空离心干燥,并在进行第二次脱盐工作流程之前储存在-80℃。

[0698] 如前所述,使用两个16号Empore C18固相萃取盘(3M,St.Paul,MN)的冲头包装的内部构建的StageTips对HLA-肽样品进行二次脱盐。StageTips用100μL MeOH活化两次,然后用50μL 50％(v/v)ACN/0.1％(v/v)FA活化,然后用100μL 1％(v/v)FA洗涤3次。通过添加200μL 3％(v/v)ACN/5％(v/v)溶解干燥的HLA肽,然后加载到StageTips上。用1％(v/v)FA(2x100μL)冲洗试管和移液器吸头,并将冲洗液转移到StageTips,然后用100μL 1％(v/v)FA冲洗StageTips五次。使用20μL 15％(v/v)ACN/0.1％(v/v)FA,然后两个20μL段30％(v/v)ACN/0.1％(v/v)FA的阶梯梯度洗脱肽。上样、洗涤和洗脱在台式离心机上进行,最大速度为1,500-3,000xg。将洗脱液冷冻,通过真空离心干燥,在-80℃下保存。

[0699] 通过串联质谱法进行的HLA-肽测序

[0700] 所有nanoLC-ESI-MS/MS分析都采用如下所述的相同LC分离条件。使用配备PicoFrit(New Objective,Inc.,Woburn,MA)75μm内径毛细管的Proxeon Easy NanoLC 1200(Thermo Scientific,San Jose,CA)层析分离样品,在采用He的1000psi压力下包装10-μm发射器,至30-40cm,采用1.9μm粒径/200 Å孔径的C18 Reprosil珠子,并在分离过程中在60℃下加热。柱子用10倍床体积的缓冲液A(0.1％(v/v)FA和3％(v/v)ACN)平衡,将样品加至4μL 3％(v/v)ACN/5％(v/v)FA,将肽用82分钟内7-30％缓冲液B(0.1％(v/v)FA和80％(v/v)ACN),6分钟内30-90％缓冲液B的线性梯度洗脱,然后在90％的缓冲液B中保持15分钟,以洗涤柱子。一部分样品用84分钟内6-40％缓冲液B、9分钟内40-60％缓冲液B的线性梯度洗涤,然后在90％缓冲液B中保持5分钟,在50％缓冲液B中保持9分钟,以洗涤柱子。用于样品洗脱的线性梯度以200nL/min的速率运行,产生约13秒的中值峰宽。

[0701] 在依赖于数据的采集过程中,将洗脱的肽引入配备有Nanospray Flex离子源(Thermo Scientific)的Orbitrap Fusion Lumos质谱仪(Thermo Scientific),电压为2.2-2.5kV。完全扫描MS以60,000的分辨率采集,从300到1,700m/z(AGC目标4e5,最大IT为50ms)。每次完全扫描之后是2秒的循环时间,或前10次,分辨率为15,000的数据依赖性MS2扫描,使用隔离宽度为1.0m/z,碰撞能量为34(HLA I类数据)和38(HLA II类数据),ACG目标为5e4,最大填充时间为250ms最大离子时间。使用1.0m/z的分离宽度是因为HLA II类肽往往更长(中值为16个氨基酸,部分肽>40个氨基酸),因此单同位素峰并不总是同位素簇中的最高峰,并且在没有指定偏移量的情况下,质谱仪采集软件将最高的同位素峰放置在隔离窗口的中心。因此,1.0m/z隔离窗口将允许共同隔离单同位素峰,即使它不是同位素簇中的最高峰,因为II类肽的电荷态通常为+2或更高。启用动态排除,重复计数为1,排除持续时间为5秒,以启用每个选定的前体约3个PSM。由于HLA-肽鉴定依赖于PSM质量,因此同位素被排除在每个前体的单个电荷态的依赖扫描被禁用时,因此不同电荷态的多个PSM进一步增加了我们对肽鉴定的信心。HLA II类数据收集的电荷态筛选与使用肽模式的单同位素前体选择(MIPS)一起启用,以防止对电荷态为1(仅适用于具有碱性锚残基的等位基因)、>7或未分配的前体离子触发MS/MS。对于HLA I类数据收集,选择了电荷态为1(质量范围800-1700m/z)和2-4的前体离子,而排除了>4和未分配的电荷态。

[0702] 使用以下方案评估使用高场不对称波形离子迁移光谱法(FAIMS)的肽检测。来自

A375细胞的内源性加工和呈递的HLA I类和HLA II类肽经历酸性反相(aRP)和碱性反相(bRP)离线分级分离,然后使用配备或不配备FAIMS接口的轨道阱融合lumos tribid质谱仪通过nLC-MS/MS进行分析。图42A演示了工作流程。图42B显示了用少量胰蛋白酶样品Jurkat细胞、HeLa细胞对FAIMS进行的基准化。使用FAIMS(分别是图43A和图43B;以及图44A和图44B)分析了HLA I类结合肽和HLA II类结合肽。在每种情况下,肽检测都略有改进,尤其是使用bRP分级分离的肽。图45和图46显示了交叉大小。

[0703] LC-MS/MS数据的解释

[0704] 该部分涉及例如图29。使用Spectrum Mill软件包v6.0预发布版(Agilent Technologies,Santa Clara,CA)解释质谱。如果MS/MS谱没有600-2000(I类)/600-4000(II类)范围内的前体MH+,前体电荷>5(I类)/>7(II类),或至少检测到<5个峰,则从搜索中排除该MS/MS谱。禁止将在同一色谱峰中获得的具有相同前体m/z的相似光谱合并。针对包含所有UCSC Genome Browser基因和基因组hg19注释的数据库及其蛋白质编码转录物(63,691个条目;10,917,867个独特的9聚体肽)以及264种常见污染物的数据库搜索MS/MS谱。在数据库搜索之前,所有MS/MS都必须通过序列标签长度>2的光谱质量过滤器,例如,最少3个质量被氨基酸的链内质量分开。设置了为5的最小骨架切割评分(BCS),并使用了ESI QExactive HLAv2评分方案。使用无酶特异性、将半胱氨酸修饰固定为半胱氨酸化以及以下可变修饰来搜索来自未还原和烷基化的天然HLA肽样品的所有光谱:氧化甲硫氨酸(m)、焦谷氨酸(N-term q)、氨基甲酰甲基化(c)。使用无酶特异性、将半胱氨酸修饰固定为氨基甲酰甲基化以及以下可变修饰来搜索还原和烷基化的HLA肽样品:氧化甲硫氨酸(m)、焦谷氨酸(N-term q)、半胱氨酸化(c)。±10ppm的前体质量容差、±10ppm的产物质量容差和30%的最小匹配峰强度用于天然、还原和烷基化HLA-肽数据集。使用Spectrum Mill自动验证模块将单个光谱的肽谱匹配(PSM)自动指定为可信分配,以在PSM排名上应用基于靶标-诱饵的FDR估计,来设置评分阈值标准。自动阈值策略使用最小序列长度为7、自动可变范围前体质量过滤以及在所有LC-MS/MS运行中针对HLA等位基因优化的评分和delta排名1-排名2评分阈值,对于每个前体电荷态产生<1.0%的PSM FDR估计。

[0705] 通过去除参考数据库中分配给264种常见污染物蛋白质的所有肽以及去除在阴性对照MAPTAC™亲和下拉中鉴定的肽,对通过PSM FDR估计值<1.0%的鉴定肽进一步过滤污染物。此外,映射到参考数据库的计算机胰蛋白酶消化物的所有肽鉴定都被删除,因为不能排除这些肽是uPLC柱上样品残留的胰蛋白酶污染物。

[0706] 使用MAPTAC™方案对HLA-DR、-DQ、DP异二聚体的单等位基因分配

[0707] LC-MS/MS鉴定肽的单等位基因HLA分配遵循两种方法。由于HLA-DRA1中的等位基因变异有限且不被认为会影响肽结合,因此DR实验(概况分析DRB1、3、4和5)的所有数据都被视为单等位基因的,这意味着肽最有可能与包含与捕获α链配对的捕获β链的HLA II类异二聚体结合。然而,仍有可能某些肽可能与包含与不同内源性表达的α链配对的敲入的β链的HLA II异二聚体结合。

[0708] 相反,对于HLA-DP和HLA-DQ基因座,α链表现出重要的等位基因变体,因此敲入和内源性α链等位基因的存在创造了多种异二聚体的潜力。例如,编码不同HLA-DP和HLA-DQ异二聚体的敲入α和β链可以各自与内源表达的α和β链配对,为每个HLA-DP和DQ MAPTAC™构建体构成多达四个独特的异二聚体。因此,纯化的MAPTAC™肽群体之间的结合特异性不是单等

位基因的。为了缓解这种内源性配对问题,使用了一种缺乏α链的构建体(无α敲入),使我们能够鉴定可能与包含内源性α链和MAPTAC™β链的HLA异二聚体结合的肽群。这些肽通过计算从相应的α+β链MAPTAC™实验中减去,以近似估计对单等位基因MAPTAC™α+β组合具有特异性的肽群。

[0709] 将每个肽分配给UCSC hg19基因注释中的一个或多个蛋白质编码转录物。由于许多肽鉴定与其他鉴定重叠,并因此构成大部分冗余信息,因此将肽分组为"嵌套组",每个意味着对应于约1个独特的结合事件,如图11C所示。例如,肽GKAPILIATDVASRGLDV、GKAPILIATDVASRGLD和KAPILIATDVASRGLDV都包含保守序列KAPILIATDVASRGLD,并且可能都在同一寄存器中结合MHC。为了嵌套给定数据集的肽,构建了一个图,其中每个节点对应于一个独特的肽,并且在任何一对共享至少一个9聚体并可映射到至少一个共同转录物的肽之间创建一条边。R程序包igraph中的簇命令用于鉴定连接节点的簇,每个簇被定义为一个嵌套组。此过程可确保将满足边缘标准(≥1个常见9聚体和≥1个常见转录物)的任何两个肽放置在同一嵌套组中。

[0710] 先前公布的MS数据的分析

[0711] 以下部分至少涉及图12A-12F、图35、图36A-36B、图38D、图39A-39C、图40A-40B。已公布的提供.raw文件的LC-MS/MS数据集使用Spectrum Mill软件包v6.0预发布版(Agilent Technologies,Santa Clara,CA)重新处理。在Thermo Orbitrap仪器上收集的在轨道阱(高分辨率)中利用HCD碎裂以及MS和MS/MS数据收集的数据集(例如Velos、QExactive、Fusion、Lumos)使用上述"LC-MS/MS数据的解释"部分中描述的参数进行分析。对于利用CID碎裂的MS和MS/MS高分辨率数据集,将与上述相同的参数与ESI Orbitrap评分方案一起使用。对于在轨道阱中收集MS数据和在离子阱中收集MS/MS数据的数据集,也使用了以下相同参数,但存在以下偏差。对于HCD数据,使用ESI QExactive HLAv2评分方案,而ESI Orbitrap评分方案用于CID数据。使用±10ppm的前体质量容差和±0.5Da的产物质量容差。对于高分辨率和低分辨率MS/MS数据集,使用Spectrum Mill自动验证模块将单个光谱的肽谱匹配(PSM)自动指定为可信分配,以在PSM排名中应用基于靶标-诱饵的FDR估计来设置评分阈值标准。自动阈值策略使用最小序列长度为7、自动可变范围前体质量过滤以及在所有LC-MS/MS运行中针对HLA等位基因优化的评分和Delta排名1-排名2评分阈值,对于每个前体电荷态产生<1.0%的PSM FDR估计。

[0712] 人类蛋白质组中的氨基酸频率是基于UCSC hg19注释中所有蛋白质编码基因的序列计算的(为由多个转录物同种型表示的基因随机选择一个转录物),如图11D所示。IEDB频率通过鉴定一组独特的肽来确定,这些肽具有至少一个≤50nM的亲和力观察结果(不包括一些在其C端具有六价多组氨酸的肽)。MAPTAC™频率首先在跨五个DRB1等位基因(DRB1*01:01、DRB1*03:01、DRB1*09:01和DRB1*11:01)的标准前相方案的背景中考虑,每个嵌套组仅使用一个肽(最长的)。此外,MAPTAC™频率针对多个等位基因的基本反相方案单独计算。不考虑潜在的起源等位基因,同样使用每个嵌套组的最长肽,分析来自外部数据集的MS数据。

[0713] 构建I类(HLA I类结合肽)序列标识

[0714] 对于每个I类等位基因(如图14A所示),通过分析相应肽的前五个位置(映射到标识位置1-5)和最后四个位置(映射到标识位置6-9)的氨基酸频率来创建长度为9的序列标识。以这种方式,无论长度如何,肽都有助于序列标识。与II类标识一样,字母高度与每个位

置的每个氨基酸的频率成正比,低熵位置使用颜色编码。

[0715] MS观察到的肽的预测亲和力

[0716] 该部分至少与图14C相关。对于每个II类等位基因,使用NetMHCIIpan鉴定所有长度为14到17的独特肽并对其结合潜力进行评分。为了比较,从人类蛋白质组中取样随机长度匹配的肽。密度分布(如图14C所示)基于对数转换值来确定。一些等位基因被排除在分析之外,因为NetMHCIIpan不支持它们的预测。

[0717] MS观察到的肽的测得的亲和力

[0718] 如果肽具有较差的预测NetMHCIIpan结合亲和力(对于DRB1*01:01为>100nM或对于DRB1*09:01和DRB1*11:01为>500nM),或者如果它们表现出在先前公布的生化MHC-肽亲和力测定中测试的≤2的启发式定义的锚,则选择肽用于亲和力测量。

[0719] 训练、调整和测试蛋白质组分区的建立

[0720] 该部分至少与图15A(也参见图31A)相关。创建了一个图,其中每个节点代表一个蛋白质编码转录物,并且在所有共享至少5个独特9聚体的氨基酸序列内容物的转录物对之间存在边(UCSC hg19基因注释)。R程序包igraph(cran.r-project.org/web/packages/igraph/citation.html)中的簇命令用于鉴定连接节点的簇,每个簇被定义为一个“转录组”。以这种方式,如果两个转录物共享一条边(>5个共享的9聚体),则保证它们被放置在同一个转录物组中。对转录物组随机采样,将75%放在训练分区中,12.5%放在调整分区中,12.5%放在测试分区中。在所有分析中,MS观察到的肽(和未观察到的诱饵肽)根据其源转录物的分区放置在分区(训练、调整或测试)中。在源转录物映射到多个不同分区的极少数情况下,分配优先顺序是训练(最优先)、调整,然后是测试(最不优先)。在所有基于分区的分析中使用相同的蛋白质组分区。使用基于图的蛋白质组分区方法来最大限度地减少训练和评估过程中出现类似肽序列的可能性,这可能会人为地提高预测性能。

[0721] 基于CNN的II类结合预测器的架构和训练

[0722] 至少关于图15A,虽然氨基酸可以由“独热(one-hot)”编码表示,但其他人选择使用PMBEC矩阵和BLOSUM矩阵来编码氨基酸,其中相似的氨基酸具有相似的特征概况。出于我们肽特征化的目的,在解析的蛋白质结构中生成了基于氨基酸邻近的独特矩阵。这种方法的概念是氨基酸的典型邻居应该反映其化学性质。对于约100,000个DSSP蛋白质结构(cdn.rcsb.org/etl/kabschSander/ss.txt.gz)中的每个氨基酸,确定在3D空间中最接近但在一级序列中至少相距10个氨基酸的残基。使用这些数据,确定丙氨酸的最近邻是丙氨酸的次数,丙氨酸的最近邻是半胱氨酸的次数等,以创建一个20x20的邻近计数矩阵。该矩阵的每个元素除以其相应的列和行总和的乘积,然后对整个矩阵进行对数变换。最后,从每个元素中减去整个矩阵的平均值。三个额外的物理特征——疏水性、电荷和大小——作为额外的列添加,使得每个氨基酸由23个输入特征表示。

[0723] 对MAPTAC™观察到的肽的基准化预测性能,与图21A-B相关。

[0724] 为了评估给定等位基因的预测性能,有必要定义一组可能已经观察到(因为它们存在于蛋白质组中)但未在MS数据中观察到的肽。这些阴性示例被称为“自然诱饵”(与上述“乱序诱饵”相反)。作为指导原则,决定:

[0725] 1.自然诱饵的长度分布应与MS观察到的命中的长度分布相匹配。

[0726] 2.自然诱饵不应包含与其他自然诱饵重复的序列。

[0727] 3.自然诱饵不应与命中重叠。

[0728] 4.自然诱饵应该来自至少产生一次命中的基因。

[0729] 以下伪代码表示创建满足这些原则的评估的过程：

[0730]
```
初始化两个空的命中列表，H最小和H穷举
对于每个嵌套组S的MS观察到的肽：
如果S中没有肽能够映射到训练或调节分区中的转录物：
将S中最短的肽添加到H最小
将S中的所有肽添加到H穷举
初始化一个空的诱饵肽列表，D
对于测试分区中的每个蛋白质编码转录物（先最长，最后最短）：
如果H穷举中没有肽映射到转录物：
跳到下一个转录物
用一组重叠的肽P覆盖转录物的蛋白质序列，其中肽长度从H最小的长度分布中随机采样。重叠是8个氨基酸。（P中的最后一个肽通常会悬挂在蛋白质的末端上。）
在P中的最后一个肽仍然悬挂时：
从P中最长肽的长度减去1个氨基酸
对于P中的每个肽：
如果它不与H穷举中的肽共享9聚体，并且它确实包含在D中的任何肽中均未观察到的9聚体：
将该肽添加到D
否则：
拒绝该肽
H最小和D构成评价数据集
```

[0731] 为了评估该组的性能,所有n个命中肽都由预测器(neonmhc2或NetMHCIIpan)进行评估,并与一组19n个诱饵(从完整的诱饵组中随机采样而不替换)一起评分。组合的组中前5%的肽被标记为阳性判定,阳性预测值(PPV)被计算为作为命中的阳性判定的比例。请注意,由于阳性数被限制为等于命中数,因此在此评估场景中判定率等于PPV。在等位基因之间应用一致的1:19比率有助于稳定性能值,否则性能值会受到每个等位基因观察到的命中数的高度影响。这被认为是合适的,因为假设与等位基因的内在特性相比,命中数更多地与实验条件和重复计数有关。如果未实施降采样,则1:19的比率与要使用的比率相差不远。

[0732] IEDB亲和力测量的基准化预测性能

[0733] 关于图15B,由于NetMHCIIpan是在IEDB亲和力数据上训练的,因此使用稍微过时的版本和IEDB测量来评估其样品外性能。预测和测量亲和力的对应关系由Kendall's tau

系数确定。相同的统计数据用于评估neonmhc2在同一组测量中的性能。评估由等位基因和分布组(Sette或Buus)分别进行。

[0734] 天然CD4+ T细胞应答的基准化预测性能

[0735] 由于IEDB中记录的绝大多数CD4+ T细胞反应具有未知或计算推算的II类等位基因限制,因此记录的子集集中在由II类四聚体通过实验证实的那些。几乎所有此类记录都由William Kwok Laboratory(Benaroya Research Institute,Seattle,WA)保存,该实验室使用免疫应答性个体的血液对不同病原体和变应原进行四聚体引导的表位定位(TGEM)。由于某些研究发布了阴性肽而其他研究没有,因此审查了来源出版物以重建完整的阳性和阴性肽反应性集合。所有20聚体肽均由neonmhc2和NetMHCIIpan评分。为了以跨等位基因的可比较方式计算跨等位基因的阳性预测值(PPV),对每个等位基因的阴性样品进行随机降采样,直到阳性与阴性的比例为1:19。PPV被计算为评分最高的5%肽中经实验证实的阳性分数。性能还通过受试者工作曲线进行评估。

[0736] 评估MHC II肽去卷积的性能

[0737] 为了评估GibbsCluster(v2.0)工具按起源等位基因对多等位基因MHC II类肽数据进行聚类的能力,首先对已知DR基因型的受试者进行的一组多样的已公布DR特异性实验中的肽(表2)进行策划。在某些情况下,原始出版物提供了HLA-DRB1分型,但省略了HLA-DRB3/4/5的分型。为了解决这些情况,假设IMGT提供了DR1:DR3/4/5链接,如果这不足以解决四位数打字,则在“USASanFranciscoCaucasian”(allelefrequencies.net,群体ID3098)中观察到的链接:使用表2。

[0738] 表2

[0739]

DRB1	DRB3/4/5	单元型频率(%)
DRB1*03:01	DRB3*01:01	23.2
DRB1*04:01	DRB4*01:01	9
DRB1*04:02	DRB4*01:01	2.2
DRB1*04:04	DRB4*01:01	3.2
DRB1*07:01	DRB4*01:01	13.2
DRB1*09:01	DRB4*01:01	2.8
DRB1*11:01	DRB3*02:02	7.2
DRB1*11:04	DRB3*02:02	2.8
DRB1*13:01	DRB3*02:02	2.6
DRB1*13:02	DRB3*03:01	2.6
DRB1*14:01	DRB3*02:02	2.2
DRB1*15:01	DRB5*01:01	28.6
注意已表明 DRB4*01:01 与 DRB4*01:03 相同 (ebi.ac.uk/cgi-bin/ipd/imgt/hla/get_allele.cgi?DRB4*0101)		

[0740] 对于每个(推算的)基因型中存在的每个DRB1/3/4/5等位基因,掺加了来自我们单等位基因MAPTAC™数据的20种肽。然后将这些增强的数据集提交给GibbsCluster-v2.0。

[0741] 表征观察到的MHC II肽的切割位点

[0742] 在此通过合并使用免疫纯化来分析人体组织的多项研究中的肽鉴定结果,公开了自然加工和呈递的肽MHC II肽的大型数据集(表2)。由于许多肽共享相同的N端(例如GKAPILIATDVASRGLDV和GKAPILIATDVASRGLD)或相同的C端(例如GKAPILIATDVASRGLD和KAPILIATDVASRGLD),因此策划了两组非冗余剪切位点,一组用于N端,一组用于C端。然后,从产生至少一种MHC II肽的基因组中随机取样等量的独特的未观察到的N端和C端切割位点。这四个数据集被称为N端命中、C端命中、N端诱饵和C端诱饵。此外,命名系统用于指代肽上游、肽内和肽下游的位置,如图41所示。

[0743] 确定每个氨基酸在位置U10到N3的N端命中频率,并将这些频率与N端诱饵观察到的频率进行比较。为了确定命中和诱饵在给定位置的给定氨基酸的比率上是否显示出显著差异,创建2x2表(例如,U1是赖氨酸的命中计数,U1是赖氨酸的诱饵计数;U1不是赖氨酸,以及U1不是赖氨酸的诱饵计数)并通过卡方检验进行评分。类似的方法用于分析C端命中和诱饵的C3到D10位置的氨基酸频率。

[0744] 第二个分析考虑了紧接在切割事件之前和之后的残基之间的统计联系。首先,对U1:N1对(A:A,A:C,A:D,…,Y:V,Y:W,Y:Y)的计数比较N端命中与N-末端诱饵,并且每对富集/消耗的显著性通过2x2列联表(例如,P:K的命中数,P:K的诱饵数;没有P:K的命中数,没有P:K的诱饵计数)的卡方检验来确定。类似的方法用于分析C端命中和诱饵的C1:D1对频率。

[0745] 对各种II类切割预测器的性能进行基准化

[0746] 对来自健康供体的外周血进行DR结合肽的分析。这些样品用于对切割相关变量/预测器的能力进行基准化,以增强对所呈递的II类表位的鉴定。

[0747] 为了构建使用结合潜力和切割潜力预测肽呈递的综合预测器,首先使用图4中描述的相同方法构建数据集,但使用“调整”分区而不是“测试”分区。简而言之,这意味着使用1:20的命中与诱饵之比,其中诱饵与命中的长度匹配,并从产生至少一个命中的一组基因中随机采样。结合潜力使用neonmhc2来计算,因为这些样品是多等位基因的,所以每个候选肽的结合分数被认为是每个供体基因型所指示的1-4个DR等位基因的平均值。这个拟合过程确定了在前向预测中将放置在结合和切割变量上的相对权重。

[0748] 为了确定前向预测的性能,使用刚刚描述的相同方案从“测试”分区获得评估命中和诱饵(1:19比率)。PPV以与图4相同的方式计算。评估了几种不同的切割预测器,如表3所示。

[0749] 表3

具有已知切割位点的 PSM
具有未知切割位点的 PSM 为了对候选肽进行评分，首先估算了 9 聚体核心。 接下来，计算三个区域的最大 PSM 评分：核心上游的 10 个氨基酸（无论真正的 N 末端切割位点的位置如何）、核心序列和核心后的 10 个氨基酸。使用 *neonmhc2* 结合评分以及三个区域特定的 PSM 评分在调整分区（见上文）上训练逻辑回归。
新型神经网络，未知的切割位点
组织蛋白酶基序，未知的切割位点
结构特征 在 SCRATCH 套件（Cheng J, Randall AZ, Sweredoski MJ & Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 33: W72–6 Available at: dx.doi.org/10.1093/nar/gki396）内，工具 ACCpro20 用于预测相对溶剂可及性（RSA），而工具 SSpro8 用于预测二级结构特征（H：α-helix；G：310-螺旋：pi-螺旋 E：延长的链 B：β 桥 T：转角 S：弯曲 C：其他）。序列无序的每个残基评分（d2p2.pro/download）同样在整个蛋白质组上确定，根据将位置标记为无序的预测引擎的数目以 0-5 的等级评分（使用的服务器：anchor、espritz-d、espritz-n、espritz-x、iupred-l 和 iupred-s）。最后，使用 Uniprot（uniprot_sprot.dat）在残基水平上确定拓扑结构域（跨膜、信号、胞外、腔和二硫键）。对于每个候选肽，在以预测的 9 聚体核心为中心的 21 聚体窗口上计算平均 RSA 和平均无序评分。在相同的 21 残基窗口上，计算了二级结构特征和拓扑特征的组成百分比。使用以下输入特征在“调整”分区上训练四个逻辑回归模型： 1. *Neonmhc2* 结合评分和溶剂可及性（RSA） 2. *Neonmhc2* 结合评分和无序评分

[0750]

3. *Neonmhc2* 结合评分和代表二级结构的八个 SSpro8 衍生特征 4. *Neonmhc2* 结合评分和代表拓扑学的五个 Uniprot 特征
[0751] *与先前观察到的肽重叠* 对来自公共领域的许多多等位基因 II 类实验汇集了基于 MS 的肽鉴定。创建了一个新特征，表示新的候选肽是否与这些先前观察到的肽之一重叠。具体而言，如果该特征与一大组先前观察到的 MHC II 配体中的任何肽共享至少一个 9 聚体，则该特征被设置为 1；否则，该特征被设置为 0。使用 *neonmhc2* 结合评分和重叠特征在“调整”分区上训练逻辑回归。

[0752] MHC II类呈递与表达之间的关系

[0753] 肽在先前公布的MS实验中汇总,这些实验描述了人类卵巢组织的HLA-DR配体组。对于具有可用RNA-Seq数据的每个样品,从SRA下载原始fastqs并使用bowtie2与UCSC hg19转录组进行比对。使用RSEM计算的每百万个中的转录物(TPM)进行转录水平基因定量。通过对基因水平求和、删除非编码基因和重新归一化来进一步处理表达估计值,使得总TPM总和为1000000(蛋白质编码基因的重新归一化解释了ncRNA丰度的文库间差异)。

[0754] 对于每个组织样品中的每个基因,考虑其在样品中的表达水平以及它是否在样品中产生至少一种肽。在所有MS实验中,根据表达水平和肽生成状态对这些观察结果进行分箱(bin)(参见图18A)。

[0755] 过度表示和表示不足的基因的鉴定

[0756] 为了鉴定在MHC II配体组中过度表示和表示不足的基因,从五项先前的研究汇编数据,这些研究对卵巢组织、结直肠组织和皮肤黑素瘤、肺癌和头颈癌进行了概况分析。对于每个基因,我们的基线假设是它应该产生与其长度乘以其表达水平成比例的肽。为了确定每个基因的长度,列举了所有转录物同种型的独特9聚体。通过对转录物同种型求和获得基因水平的表达。在嵌套组水平上确定映射到每个基因的肽的观察的数目(例如肽GKAPILIATDVASRGLDV、GKAPILIATDVASRGLD和KAPILIATDVASRGLDV被计为单个观察)。

[0757] 来自卵巢研究的许多样品都有相应的RNA-Seq数据,但有些没有。在这些情况下,使用可用的RNA-Seq数据估计整个样品的平均表达。对于结直肠和黑素瘤研究,任何样品都没有相应的RNA-Seq,因此使用来自GTEx和TCGA的数据计算了替代样品的平均值。在所有情况下,根据与上述卵巢研究的RNA-Seq相同的方案,获得原始fastqs并对其进行比对和量化。

[0758] 创建了两个矩阵,分别表示预期和观察到的计数,分别称为E和O,其中行对应于基因,列对应于样品。首先通过将每个基因的长度乘以其在每个样品中的表达来填充矩阵E;然后重新调整E的列,以使E的列总和与O的列总和匹配。最后,通过比较E的行总和与O的行总和,在基因水平上进行分析。根据其在人血浆中的存在和浓度突出显示基因。

[0759] 自噬相关基因的分析

[0760] 定义了两个自噬相关基因集。第一组包括通过实验被鉴定为已知自噬相关基因的物理相互作用配偶体的蛋白质。对于在自噬交互网络数据库(从besra.hms.harvard.edu/

ipmsmsdbs/cgi-bin/downloads.cgi访问)中记录的IP-MS实验中用作诱饵的每个经典的自噬相关基因(genenames.org/cgi-bin/genefamilies/set/1022),根据“WD”置信度得分(besra.hms.harvard.edu/ipmsmsdbs/cgi-bin/tutorial.cgi)确定前100个蛋白质鉴定。汇集22个实验,获得了一组1004个与至少一个经典自噬相关基因可信地相关的独特基因(图18C)。

[0761] 使用利用SILAC(sciencedirect.com/science/article/pii/S1097276514006121)测量ATG5敲除前后幼小鼠肾上皮(iBMK)细胞中泛蛋白质组蛋白丰度的研究,确定了第二组自噬相关基因。t-统计学>5的基因被归类为通过ATG5敲除稳定的(饥饿前条件;补充数据文件mmc2.xls中的变量“Intercept_t”)。为了将每个小鼠Uniprot ID映射到hg19 UCSC ID,确定与小鼠Uniprot序列共享最多9聚体的人UCSC蛋白质序列(图18C)。

[0762] 基于图18B,计算每个基因观察到的肽与预期肽的比率R,将1的伪计数添加到分子和分母,例如R＝(O+1)/(E+1)。Log(R)被用来代表每个基因的相对富集(Log(R)>0)或耗竭(Log(R)<0),认为没有足够的信息来量化这些基因的相对富集/耗竭。在具有有效Log(R)计算的基因中,针对IP-MS数据集中的基因、SILAC数据集中的基因和两个数据集都没有的基因绘制了Log(R)分布(图18C)。

[0763] 源基因定位的分析,与图18D相关

[0764] 使用与上述相同的log(R)评分,根据每个源基因的定位绘制分布图(图18D)。使用Uniprot(uniprot_sprot.dat)确定源基因定位。

[0765] 单细胞RNA-Seq数据中II类表达数据的分析,与图19A相关

[0766] 单细胞RNA-Seq数据来自三个先前公布的人类肿瘤样品数据集。

[0767] 第一项研究包括来自皮肤黑素瘤的数据。文件“GSE72056_melanoma_single_cell_revised_v2.txt”从Gene Expression Omnibus(ncbi.nlm.nih.gov/geo/;登录:GSE72056)下载。将肿瘤状态标志为“2”的细胞视为肿瘤细胞,将肿瘤状态标志为“1”且免疫细胞类型标志等于“1”至“6”的细胞分别视为T细胞、B细胞、巨噬细胞、内皮、成纤维细胞和NK。所有其他单元都被丢弃。数据本身以log2(TPM/10+1)为单位呈现,因此在数学上转换为TPM标度。一旦在TPM标度上,将每个细胞的数据相对于蛋白质编码UCSC基因符号集重新归一化至总和为1,000,000(未出现在表达矩阵中的蛋白质编码基因被隐含地视为具有零表达)。最后,将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值,以产生患者细胞类型水平上的表达估计。

[0768] 第二项研究包括来自头颈部肿瘤的数据。文件“GSE103322_HNSCC_all_data.txt”从Gene Expression Omnibus(ncbi.nlm.nih.gov/geo/;登录:GSE103322)下载。该表中的数据以log2(TPM/10+1)为单位;因此,这些值在数学上转换为TPM单位。与黑素瘤研究一样,将每个细胞的数据相对于蛋白质编码UCSC基因符号集重新归一化至总和为1,000,000,并且将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值。对应于淋巴结活检的数据被排除在外。第三项研究包括来自未经治疗的非小细胞肺的数据。文件“RawDataLung.table.rds”和“metadata.xlsx”从ArrayExpress(ebi.ac.uk/arrayexpress/;登录:E-MTAB-6149和E-MTAB-6653)下载。如前所述,数据(已经以TPM为单位)相对于先前描述的蛋白质编码基因集重新缩放至总和为1,000,000。最后,将对应于相

同细胞类型和相同来源活检的单细胞观察结果取平均值,以产生患者细胞类型水平上的表达估计。在结直肠癌和卵巢癌中进行了类似的研究。结果显示在图19A中。

[0769] 为简单起见,将细胞类型合并为比表4中原始报告的更粗略的类别。

[0770] 表4

[0771]

粗略分配	组成细胞类型
肺泡	“肺泡”，不包括“立方形肺泡 2 型(AT2)细胞”
FO B 细胞	“滤泡 B 细胞”
浆细胞	“血浆 B 细胞”
CLEC9A+ DCs	“交叉呈递树突细胞”
单核树突细胞	“单核细胞衍生的树突细胞”
pDCs	“类浆树突细胞”
郎格汉斯	“郎格汉斯细胞”
巨噬细胞	“巨噬细胞”
粒细胞	“粒细胞”
内皮	“正常内皮细胞”、“肿瘤内皮细胞”和“较低质量的内皮细胞”，不包括“淋巴 EC”
上皮	“上以细胞”和“较低质量的上皮细胞”
成纤维细胞	“表达 COL12A1 的成纤维细胞”、“表达 COL4A2 的成纤维细胞”、“表达 GABARAP 的成纤维细胞”、“较低质量的成纤维细胞”、“正常肺成纤维细胞”、“表达 PLA2G2A 的成纤维细胞”和“表达 TFPI2 的成纤维细胞”
T 细胞	“调节性 T 细胞”、“CD4+ T 细胞”和“CD8+ T 细胞”
NKs	“自然杀伤细胞”

[0772]

肿瘤	“癌细胞”
从分析中排除	“成红血细胞”和“MALT B 细胞”

[0773] 五项研究中HLA-DRB1的表达水平绘制在图19A中。

[0774] 肿瘤衍生的与基质衍生的II类表达的表征

[0775] 为了确定归因于肿瘤与基质的MHC II类结合肽表达的相对量,在TCGA患者的II类途径基因(根据DNA判定)中鉴定了突变,并且对于每个携带II类突变的患者,对相应RNA-Seq基因量化了突变和非突变拷贝的相对表达。此外,还假设:

[0776] 1.突变的读取来自肿瘤

[0777] 2.肿瘤中的基质或野生型等位基因出现非突变的读取

[0778] 3.肿瘤保留了野生型拷贝,其表达量大约等于突变的拷贝

[0779] 基于此,确定对于观察到的突变等位基因分数f,,归因于肿瘤的II类表达分数约为2f且不大于100%。选择三个基因——CIITA、CD74和CTSS——作为核心II类途径基因,并评估TCGA中的突变(不排除同义和UTR突变)(数据从TumorPortal(tumorportal.org/)下载:BRCA、CRC、HNSC、DLBCL、MM、LUAD;TCGA批量下载(tcga-data.nci.nih.gov):CESC、LIHC、PAAD、PRAD、KIRP、TGCT、UCS;Synapse(synapse.org/#!Synapse:syn1729383):GBM、KIRC、LAML、UCEC、LUSC、OV、SKCM;或原始TCGA出版物(cancergenome.nih.gov/publications):BLCA、RICH、STAD和THCA)。这些基因的选择是基于它们在II类表达中的已知作用以及它们与8500个GTEx样品群组中的HLA-DRB1的紧密相关性。与HLA-DRB1具有同等相关性的其他基因(HLA-DRA1、HLA-DPA1、HLA-DQA1、HLA-DQB1和HLA-DPB1)被排除在外,因为它们的多态性使它们易于出现假阳性突变。自然,只有一小部分患者在CIITA、CD74或CTSS中发生突变,并且对于某些肿瘤类型,没有可供分析的患者。

[0780] 对二进制序列比对/定位(BAM)格式的原始全外显子组测序(WES)的序列进行目视评估(IGV工具),以确认突变存在于肿瘤样品中,而不存在于正常样品中。使用pysam从相应的RNA-Seq获得突变体与野生型读取计数。基于从Genomic Data Commons(gdc.cancer.gov)下载的表达数据确定总体HLA-DRB1表达,其相对于蛋白质编码基因集重新归一化为总和为1,000,000。归因于肿瘤的HLA-DRB1表达的分数(图19B)被估计为min(1,2f),其中f是CIITA、CD74或CTSS中显示突变的RNA-Seq读取的分数。

[0781] 评估对天然供体组织的预测整体性能

[0782] 如上文“基于抗体的HLA-肽复合物的分离”部分所述,用DR特异性抗体分析来自7个健康供体的外周血。基于这些结果,定义了两个数据集:一个用于拟合多元逻辑回归,另一个用于评估回归的预测性能。

[0783] 第一个数据集通过使用先前关于图4描述的命中和诱饵选择算法来构建。简而言之,这意味着用一个命中肽(嵌套组中最短的肽)表示每个嵌套组,并在基因上平铺长度匹配的诱饵,这样它们与命中的重叠最少,彼此之间的重叠也最少。然而,有两个重要细节与图4中概述的算法不同。首先,命中和诱饵是从“调整”分区(而不是“测试”分区)中的基因中选择的,其次,诱饵被允许映射到显示零命中的基因。在这个数据集上训练了具有MHC结合评分(来自NetMHCIIpan或neonmhc2)以及其他输入特征(表达、基因偏好等)的逻辑回归模型。

[0784] 第二个数据集(用于评估)以相同的方式构建,不同之处在于它使用从“测试”分区中提取的命中和诱饵。除了结合评分之外,以下变量还用于回归的子集中,如表5所示。

[0785] 表5

[0786]

NetMHCIIpan 亲和力	源自 NetMHCIIpan。对于每个候选肽，在供体基因型中所有 DR 等位基因中取最高评分。
基于MS的结合评分	源自 *neonmhc2*。对于每个候选肽，在供体基因型中所有 DR 等位基因取最高评分。
表达	基因表达估计是通过分析来自 bowtie2、RSEM 和仅对蛋白质编码基因的重新归一化的数据获得的，值是 N 个样本的平均值

[0787]

热点	候选肽是否与先前公布的任何多等位基因数据集共享至少一个 9 聚体的指示变量（0/1）
基因偏好	(1+观察)/(1+预期)，根据分析
交叉呈递	候选肽是否来自观察到被 DC 在 MHC II 类上交叉呈递的基因的指示变量（0/1）。

[0788] 出于性能评估的目的,所有n个命中肽都通过给定的逻辑回归进行评估,并与一组499n个诱饵(从完整的诱饵组中随机取样,没有替换)一起评分。组合的集合中前0.2%的肽被标记为阳性判定,阳性预测值(PPV)被计算为作为命中的阳性判定的分数。请注意,由于阳性数被限制为等于命中数,因此在此评估场景中,判定率恰好等于PPV。在等位基因之间应用一致的1:499比率有助于稳定性能值,否则性能值会受到每个供体观察到的命中数的高度影响。这被认为是合适的,因为假设与供体细胞的内在特性相比,命中次数更多地与实验条件相关。如果不实施降采样,那么1:499的比率与将使用的比率相差不远。

[0789] 虽然本文已经显示并描述了本发明的优选实施方案,但是对于本领域技术人员显而易见的是,这些实施方案仅以示例的方式提供。并非打算用本说明书中提供的具体实例来限制本发明。尽管已经参照上述说明书对本发明进行了描述,但并不意味着对本文实施方案的描述和说明以限制性的意义来解释。在不脱离本发明的情况下,本领域技术人员现将想到许多变化、改变和替换。此外,应当理解,本发明的所有方面均不限于本文所阐述的具体描述、配置或相对比例,其取决于多种条件和变量。应当理解,在实施本发明的过程中可以采用本文所述的本发明实施方案的各种替代方案。因此可以预期,本发明还应涵盖任何这类替代、改变、变化或等同物。旨在用以下权利要求书限定本发明的范围,并由此涵盖在这些权利要求范围内的方法和结构及其等同物。

[0790] 实施例11:HLA II类等位基因结合表位的高通量鉴定和验证

[0791] 在该实施例中,描述了使用时间分辨荧光能量转移(TR-FRET)来鉴定和验证新型MHC-II等位基因结合肽的代表性可靠、高通量方法。该测定有几个部分,(1)用适合表达和分泌MHC-IIα和β链的载体构建体转染细胞,其具有用于FRET测定的荧光标签,(2)纯化分泌的MHC-II构建体蛋白质产物,(3)进行肽交换测定(图22A)。图22B和图23进一步举例说明了设计和程序。本文所述的测定促进了快速有效的检测和验证方案,因为它可能不需要稳定的细胞系,并且包含简单的分离策略。此外,四聚体或多聚体可用于检测抗原特异性CD4细

胞,例如,在体内施用neonmhc2预测表位后,其后产生的免疫应答用于验证CD4+ T细胞应答。

[0792] 用于鉴定高亲和力MHC II类结合肽的CLIP-TR-FRET测定

[0793] 本文提供了用于在单个构建体中表达由CMV启动子驱动的HLA II类α和β链的示例性载体,其蛋白质产物产生正确折叠的α和β链对。在适当折叠的α和β链形式中,α1亚单体和β1亚单位呈二聚体形式,α1亚单体和β1亚单位形成开放接受端,能够接受肽,类似于生理构型。出于该测定的目的,这些载体表达的具有正确折叠的α和β链形式的HLA蛋白产物被称为HLA单体。表达构建体包含与生物素化基序(BAP)和10X-His-标签连接的连接体、一个或多个肽切割位点、分泌信号、二聚化因子,例如c-Fos和Jun。占位肽用于稳定单体并帮助分泌。占位肽可以是CMV肽。占位肽可以是CLIP肽。占位肽可以是通过基于MS的等位基因配体组鉴定的肽。占位肽可以在开放的α1-β1肽接受端与HLA肽共价结合。

[0794] 本文使用的示例性构建体编码具有置于CLIP和b链之间的凝血酶切割部分的CLIP占位肽,如图23(上图)所示。在转染和培养转染的细胞后,使其达到最佳生长,收集包含分泌蛋白(单体)的细胞培养上清液(培养基)并通过镍(Ni^{2+})柱进行纯化(图24)。通过考马斯染色检查表达水平和纯化。28聚体CLIP肽保持与β链缔合,β链通过用凝血酶处理而被切割(图24),此后可通过与测试(例如,候选)HLA-II类结合肽竞争而脱落。基于取代CLIP肽的能力,如其IC50所测量的,成功去除CLIP肽的测试肽是用于结合构建体的MHC-II异二聚体的同源肽。从头测试肽可用于如上所述的竞争性取代反应,然后可通过质谱法(MS)进行鉴定。

[0795] 大量HLA-DR异二聚体构建体由成功分泌的CLIP占位肽制成,并进行了肽交换测定。

[0796] 据观察,源自CD74的肽占位体CLIP对HLA II类单体的分泌具有显著影响。具有CD74序列PVSKMRMATPLLMQA(在图25A-25C中指定为CLIP0)的经编辑的经典CLIP肽通常用作占位体序列。然而,可以看到一些HLA-DR肽,例如DRB1*12:01和DRB1*13:02,使用经典肽的产率较低(表6)。观察到某些HLA DRB等位基因二聚体具有更长的结合序列,覆盖序列中的全部或部分氨基酸:LPKPPKPVSKMRMATPLLMQALPM(CLIP1)(图25A)。实际上,在DRB1*12:01和DRB1*13:02的情况下使用CLIP1序列代替CLIP0序列提高了HLA二聚体的分泌产量(图25B-25C)。

[0797] 表6

[0798]

DRB1*	肽占位体	产量(mg/L)
01:01	PVSKMRMATPLLMQA (CLIP)	20
	LPLKMLNIPSINVH (CMV)	100
	PKYVKQNTLKLAT (HA)	85
12:01	PVSKMRMATPLLMQA (CLIP)	<3
13:02	PVSKMRMATPLLMQA (CLIP)	<3

[0799] 使用STII-TR-FRET通过成功肽交换测定对肽的从头筛选

[0800] 可以使用涉及在细胞系(例如Expi293细胞)中表达上述HLA单体蛋白的分析从头筛选肽,从上清液中收集并纯化,并进行肽交换测定。使用肽交换测定测试由预测算法预测

的HLA II类结合肽。可以使用涉及荧光偏振的方法进行肽交换测定。例如,任何荧光团均可用于标记占位肽,或标记测试肽,或使用两种不同的荧光团标记两者。可以记录下来由于先前用荧光团标记的结合占位肽的丢失或由释放的荧光团(该荧光团在其HLA结合形式中被生化反应猝灭)的荧光发射而引起的荧光变化,以用于取代反应的定量评估。或者,可以记录用标记的荧光肽取代非荧光占位体肽,以定量地确定取代反应。在示例性测定中,FITC标记的占位体CLIP肽用于取代现有的共价结合肽,例如CMV肽。FITC标记的肽与HLA结合时会诱导高偏振。当用测试肽滴定FITC占位肽时,测试肽会取代FITC-CLIP,从而导致荧光降低。

[0801] 还可以使用时间分辨FRET(TR-FRET)技术代替本文所述的荧光偏振来进行肽交换测定。在本文所述的示例性TR-FRET测定中,用具有占位肽的HLA单体构建体转染细胞,所述占位肽包含Streptag II(STII)部分。STII部分由Alexa-647标记的STII抗体检测到。同时,通过铕III(Eu)化合物偶联的抗His抗体检测与本实施例中前面描述的单体构建体的Jun末端相连的His标签,该标签位于HLA肽的α2-β2末端附近(图26A)。Eu络合物充当能量供体,而当占位肽保持与HLA单体结合时,Alexa647在FRET反应中充当受体。当测试肽取代STII-CLIP占位体时,Alexa-647-αSTII肽被释放,而无法再通过荧光检测到。发现TR-FRET测定比荧光偏振更可靠。此外,该测定的背景信号大大降低。(荧光读出数据显示在图26B-26E中)。该测定为HLA-肽对提供了高通量鉴定平台。如以下表7所示,测试肽(或候选肽)P-156至P-191表现出广泛的取代能力和结合亲和力,如每次运行计算的IC50所确定的。较低的IC50表现出较高的取代能力和较高的结合亲和力。

[0802] 表7

[0803]

肽	IC50 (nm) 运行 1	IC50 (nm) 运行 2	Avg IC50 (nM)	标准偏差 (nM)	备注
P-156	36294	38246	37270	976	
P-157	2550	2243	2397	154	
P-158	5786	5815	5800	15	
P-159	58	94	76	18	
P-160	1668	2020	1844	176	
P-161	13541	14401	13971	430	
P-162	3298	3636	3467	169	

[0804]

P-163	>40,000	>40,000	>40,000	N/A	40μM 时<50%取代
P-164	4553	4353	4453	100	
P-165	357	422	389	32	
P-166	3448	3104	3276	172	
P-167	6612	5906	6259	353	
P-189	4597	----	----	---	
P-190	1137	-----	----	---	
P-191	5167	-----	----	---	
P-192	>40,000				直到40μM时无取代
HA	23	23	23	0	
ASP51	>40,000	>40,000	N/A	N/A	直到40μM时无取代

[0805] 使用差示扫描荧光法(DSF)进行的肽交换验证

[0806] 在该方法中,确定了用于筛选可与特定HLA等位基因结合的肽以及肽与HLA二聚体结合的强度的高通量测定(图26F)。在该测定中,使用荧光探针,它与蛋白质的疏水残基结合,因此可以与MHC等位基因结合,只有当等位基因通过加热而彼此解离时。当MHC II类二聚体与同源肽结合时,二聚体以其二聚体形式结合在一起。当对结合形式的MHC二聚体肽加热时,弱结合肽会更快地从MHC II类蛋白二聚体解离,使荧光团与解离的MHCα和β链结合并产生高荧光。荧光被记录为温度的函数。代表性解链曲线示于图26F中。可以比较解链曲线以确定强结合物(在较高温度下检测到荧光)与弱结合物(在较低温度下检测到荧光)。

[0807] 使用可溶性HLA-DM(HLA-sDM)作为MHC11类肽交换的催化剂:HLA-DM是HLA-DR、-DP和-DQ分子的天然分子伴侣和肽交换催化剂。它是膜内在蛋白,以α和β多肽链(DMA和DMB)的异二聚体形式出现。本节中描述的肽交换是使用可溶形式的HLA-DM(例如HLA-sDM蛋白)作为HLA-DR、-DP和-DQ交换的分子伴侣进行的。HLA-sDM蛋白通过在Expi-CHO细胞中瞬时转染而产生,如图26G所示。简而言之,设计了重组HLA-sDM构建体,如图26G上半部分所示。重组HLA-sDM构建体包含CMV组成型启动子,其位于前导序列的上游并与启动子可操作地连接。前导序列有助于产物的分泌(分泌信号)。在前导序列的3'末端引入了HLA-DMβ链胞外域(并且缺少跨膜域)的编码序列。编码生物素化基序(BAP)的序列连接在β链编码序列的3'端。编码HLA-DMα链胞外域(且缺少跨膜域)的序列在其5'端带有分泌序列(前导序列),通过插入的核糖体跳跃序列与BAP序列分开。HLA-DMα链序列在3'末端与10XHIS标签连接。一旦形成,异二聚体HLA-sDM就会分泌到细胞外。当该构建体在Expi-CHO细胞中表达时,HLA-sDM蛋白被分泌到培养基中。

[0808] 用表达HLA-sDM构建体的质粒载体转染Expi-CHO细胞,并培养约14天的时间。在培

养期间蛋白质被分泌到培养基中。在与纯化MHC-II蛋白非常相似的过程中,从培养物中纯化HLA-sDM蛋白。MHC-II肽交换可以用酸和HLA-sDM有效地进行,或者不采用酸,以及采用辛基葡糖苷。进行大小排阻色谱法以评估肽交换,结果如图26H所示。使用HLA-sDM或辛基-葡糖苷作为催化剂进行所有肽交换测定。

[0809] HLA-II类四聚体(或多聚体)组库

[0810] 为了在生化分析中测试表位:HLA结合和解离动力学,产生了大量HLA II类四聚体。由此产生的这些II类四聚体用于分析肽结合和呈递。例如,四聚体用于肽交换测定中。如图27A所示,以大于15mg/ml的浓度产生并储存12个四聚体;六个四聚体和四个<5mg/ml。HLA四聚体用于流式细胞术以鉴定新抗原反应性CD4+ T细胞。当被HLA四聚体呈递时,测试流感病毒表位(HA)和HIV表位的T细胞识别(图27E)。

[0811] 图27B-27D描绘了产生和纯化的HLA II类四聚体的各种子集。如图27B所示,以大于15mg/L的浓度构建并纯化了大量DRB1异二聚体四聚体。图27C和27D总结了针对基于荧光的肽结合测定产生和验证的人MHC II类等位基因构建体的覆盖范围。表8A、表8B和表8C提供了制备的等位基因四聚体的列表,以及相应的纯化产物的分泌产量浓度。

[0812] 表8A

[0813]

HLA 异二聚体	分泌产量
DRB1*01:01	>15 mg/L
DRB1*04:01	>15 mg/L
DRB1*04:02	>15 mg/L
DRB1*04:04	>15 mg/L
DRB1*04:05	>15 mg/L
DRB1*08:01	>15 mg/L
DRB1*09:01	>15 mg/L
DRB1*11:01	>15 mg/L
DRB1*13:03	>15 mg/L
DRB1*14:01	>15 mg/L
DRB1*15:03	>15 mg/L
DRB1*01:02	>15 mg/L
DRB1*11:04	>15 mg/L
DRB1*15:02	>15 mg/L
DRB4*01:01	>15 mg/L
DRB1*07:01	5-15 mg/L
DRB1*13:01	5-15 mg/L
DRB1*13:02	5-15 mg/L
DRB1*15:01	5-15 mg/L
DRB1*15:02	5-15 mg/L
DRB3*01:01	5-15 mg/L
DRB1*08:03	5-15 mg/L
DRB1*11:02	5-15 mg/L
DRB1*16:02	5-15 mg/L
DRB3*02:01	5-15 mg/L

[0814]

DRB3*02:02	**5-15 mg/L**
DRB3*03:01	**5-15 mg/L**
murine I-Ab	**5-15 mg/L**
DRB1*03:01	**<5 mg/L**
DRB1*12:01	**<5 mg/L**
DRB5*01:01	**<5 mg/L**
DPA*01:03/DPB*04:01	**<5 mg/L**

[0815] 表8B

[0816]

HLA异二聚体	分泌产量
DPB1*05:01	>15mg/L
DPB1*13:01	>15mg/L
DPB1*03:01	5-15mg/L
DPB1*04:02	5-15mg/L
DPB1*06:01	5-15mg/L
DPB1*11:01	5-15mg/L
DPB1*01:01	<5mg/L
DPB1*02:01	<5mg/L
DPB1*02:02	<5mg/L
DPB1*04:01	<5mg/L
DPB1*17:01	<5mg/L

[0817] 表8C

[0818]

HLA 异二聚体	**分泌产量**
A1*02:02+B1*06:02	**>15 mg/L**
A1*02:01+B1*02:02	**>15 mg/L**
A1*01:03+B1*06:03	**>15 mg/L**
A1*02:01+B1*03:03	**>15 mg/L**
A1*01:02+B1*06:04	**5-15 mg/L**
A1*05:01+B1*02:01	**<5 mg/L**
A1*05:05+B1*03:01	**<5 mg/L**

[0819]

A1*01:01+B1*05:01	**<5 mg/L**
A1*03:01+B1*03:02	**<5 mg/L**
A1*03:03+B1*03:01	**<5 mg/L**

[0820] MHC-II四聚体产品管线还包括DRB3、4和5等位基因,以及DP和DQ等位基因。

[0821] 使用荧光偏振(FP)进行的肽交换验证

[0822] 荧光偏振显微术用于区分与MHC II类蛋白结合的肽与游离肽的测定。荧光标记的占位肽与MHC II类二聚体结合时,通过荧光偏振(FP)显微镜观察产生高偏振光,与其释放形式相比,当非荧光团标记的竞争表位肽仍与MHC II类结合时二聚体取代占位肽。图28A通过图形表示展示了该原理。简而言之,该测定以下列通用方法进行,并且变化要么在各自的描述中指出,要么容易被本领域技术人员理解。

[0823] 将表9(下文)中所述的试剂组装在反应管(例如1.5mlEppendorf管)中,充分混合并在37℃下孵育2小时。在孵育时间结束时将25ml10XPBS添加到混合物中以中和肽交换反应。

[0824] 表9

[0825]

成分	储备浓度	终溶液
凝血酶消化的 MHC II 类等位基因	可变	5 M
交换肽/目的肽	10mM	50 M
乙酸钠, pH5.2	1M	100 mM
氯化钠	5M	50 mM
可溶性 DM	可变	5 M
MiliQ 水		Up to 100 l

[0826] 例如,通过染色检测交换的肽;或在-80℃下在液氮中快速冷冻以供日后评估。

[0827] 图28B和28C提供了使用FP的HLA DRB1*01:01的测定开发以及所使用的各种条件的概述。在一些实施方案中,确定了pH对测定的影响。简而言之,全长和可溶性等位基因都在细胞中表达。膜结合的全长等位基因形式是通过对膜进行透化来收集的,而分泌形式是从细胞上清液中收集的。收获的HLA II类蛋白通过镍(Ni^{2+})柱进行纯化。在一些实施方案中,在用于收获全长MHC II等位基因的膜透化中评估了去污剂(1%辛基葡糖苷对比1.6%NP40)的效果。在一些实例中,单独评估温度或所用探针、或纯化方法或目标形式的影响(图28B)。

[0828] 评估了使用构象特异性抗体L243或His标签纯化的纯化方法的效果。结果示于图28D。每个数据点都在左侧的表格中进行了描述,并在右侧的图表中表示为一个点。图中的点与任一轴大致成45度角对齐,r值为0.9621,这表明两种纯化方法的IC50值彼此一致。它还表明肽效价的等级顺序在纯化方法之间没有变化。

[0829] 评估了选择可溶形式(sDR1)与全长形式(fDR1)的HLA II类蛋白的影响,并且图28E显示目标形式的选择不影响肽效力。左侧显示的是来自使用sDR1形式或fDR1的实验的平均IC50值。绘制这些数据以获得右侧的图形。每个数据点都由一个点表示,大致沿与任一轴成45度角对齐,r值为0.9365,这表明所使用的两种形式的IC50值相互关联良好。它还表明肽效价的等级顺序在纯化方法之间没有变化。

[0830] 图28F显示了neonmhc2和NetMHCIIpan预测的肽在结合测定和不一致肽的鉴定中

的示例性评估方法的图解视图。荧光偏振测定用于评估实际肽结合测定中的neonmhc2和NetMHCIIpan预测肽。对于该测定,将60nM凝血酶消化的可溶性HLA-DRB1*15:01与FITC标记的超级结合探针肽(PVVHFFK(FITC)NIVTPRTPPY)(每次测定10nM)和测定肽在测定缓冲液(pH5.2)中在37℃下孵育5小时。检查荧光偏振,由此计算探针位移百分比。如图28G所示,超级结合荧光肽的抑制与预测肽的浓度成正比,表明该测定具有良好的特异性。neonmhc2和NetMHCIIpan预测肽的性能之间存在显著差异。使用neonmhc2预测肽,更多的肽被阳性结合,并且具有更高的抑制程度;而与neonmhc2肽相比,NetMHCIIpan预测的肽总体上表现不佳。图28H总结了neonmhc2预测肽在结合测定中的评估。如饼图所示,在双阴性肽(未被任何NetMHCIIpan或neonmhc2预测的肽)中,只有5%是结合物,95%是非结合物。在NetMHC II预测的肽中,40%是通过探针取代测定的荧光偏振检测的结合物,而在neonmhc2预测的肽中,通过探针取代测定发现100%是真正的结合物。

[0831] FITC标记的探针是通过回顾Sette等人先前公布的肽来制备的,以结合特定的等位基因。然后使用预测的II类结合核心分析这些肽序列以鉴定肽的最小9聚体核心和锚残基。然后在选择用于赖氨酸置换和FITC标记的残基位置时考虑该信息。例如,在下表(表10)中,列出了Sette等人(Sidney J,Southwood S,Moore C等人.Measurement of MHC/peptide interactions by gel filtration or monoclonal antibody capture.Curr Protoc Immunol.2013;Chapter 18:Unit-18.3.doi:10.1002/0471142735.im1803s100)描述的序列(以下简称“Sette序列”)。每个肽的预测II类结合核心在特定等位基因的上下文中加下划线。粗体表示作为表位改进的结果鉴定的锚位置。在某些情况下,相同的肽序列可用于不同的等位基因。

[0832] 表10

[0833]

速写ID	等位基因	Sette序列	选择的探针序列
SB-DR7/11	DRB1*07:01	YATFFIKANSKFIGITE	YATFFIKANSKFIGITE
SB-DR7/11	DRB1*11:01	YATFFIKANSKFIGITE	YATFFIKANSKFIGITE
SB-DR9	DRB1*09:01	TLSVTFIGAAPLILSY	TLSVTFIGAAPKILSY
SB-DR4/15	DRB1*15:01	PVVHFFKNIVTPRTPPY	PVVHFFKNIVTPRTPPY
SB-DR4/15	DRB1*04:01	PVVHFFKNIVTPRTPPY	PVVHFFKNIVTPRTPPY
SB-DR3	DRB1*03:01	YARIRRDGCLLRLVD	YARIKRDGCLLRLVD

[0834] 基于如上所述的定位,通过关注结合核心内非蓝色的位置(下划线)选择用于FITC缀合的内部赖氨酸——以红色突出显示这些位置作为FITC缀合的合适位置。对于没有用于FITC缀合的内部赖氨酸的序列,采用手动方法,将等位基因的结合基序与肽序列进行比较,并为DRB1*09:01和DRB1*03:01肽选择内部赖氨酸置换的位置(见上表)。更具体地,将DRB1*09:01的亮氨酸残基和DRB1*03:01的精氨酸残基置换为赖氨酸以允许FITC缀合。这种置换策略基于MAPTAC衍生的基序,其中手动鉴定没有强氨基酸偏好的位置(也在neonmhc2预测的9聚体核心的中间),因为缀合的荧光团在结合时可能更可能发出偏振光(即,荧光团的运动更受限制)。

[0835] 实施例12:HLA II类结合和加工规则,用于鉴定可治疗靶向的癌症抗原

[0836] 越来越多的证据表明CD4+ T细胞可以识别癌症特异性抗原并控制肿瘤生长。然而,仍然难以预测将由人类白细胞抗原II类分子(HLA II类)呈递的抗原——阻碍了在治疗

上最佳靶向它们的努力。障碍包括不准确的肽结合预测和未解决的HLA II类途径的复杂性。在本实施例中,描述了用于发现HLA II类结合基序的改进技术。此外,本文描述了对肿瘤配体组的综合分析,以学习与肿瘤微环境(TME)相关的加工规则。

[0837] 对40个HLA II类等位基因进行了分析,结果表明结合基序对HLA-DM(一种装载肽的分子伴侣)高度敏感。肿瘤内HLA II类呈递显示由专职抗原呈递细胞(APC)主导,而不是癌细胞。整合这些观察结果,如本文所述开发了算法,其准确预测APC配体组,包括来自被吞噬的癌细胞的肽。这些工具和生物学见解可以增强HLA II类指导的癌症治疗。

[0838] 一类有前景的新疗法寻求通过诱导T细胞对癌症抗原和称为新抗原的体细胞突变序列的反应来治疗癌症。目前,这些努力主要集中在引发CD8+ T细胞对HLA I类(HLA I类)呈递配体的反应。然而,最近的几项研究表明,CD4+ T细胞也可以识别HLA II类呈递的配体并有助于控制肿瘤。理想情况下,癌症疫苗和其他免疫疗法会利用引导CD4+ T细胞应答的优势,但目前的努力已经完全放弃了HLA II类抗原预测,因为当前预测工具的准确性不足。

[0839] 阻止准确鉴定HLA II类癌症抗原的一个关键因素是学习肽结合规则所需的全面、高质量数据的可用性。需要三个高度多态性的典型HLA II类基因座HLA-DR、-DP和-DQ的数据,其中每个等位基因变体表现出不同的肽结合偏好。一种广泛使用的定义肽结合基序的方法是一种生化分析,它在没有生理伴侣(例如HLA-DM)的情况下测量单个肽的亲和力。测量的亲和力数据覆盖范围仅限于常见的高加索人HLA-DR等位基因,即使对于这些等位基因,预测准确性也显著落后于HLA I类。原则上,基于质谱(MS)的配体组学应通过提供可扩展性和内源性来改进预测肽加载条件。尽管如此,天然配体组是多等位基因的,隐藏了获得准确训练数据所需的肽到等位基因的映射信息。使用去卷积和单等位基因HLA II类细胞系或具有纯合HLA-DR等位基因的细胞系解决此问题的HLA I类已取得进展,已经使用低通量转基因小鼠模型HLA II类缺陷细胞系生成了单等位基因HLA II类配体组数据集。

[0840] 另一个挑战是关于哪些肿瘤抗原最有可能进入HLA II类呈递途径的模糊性。最近基于MS的研究调查了肿瘤样品的HLA II类配体组,但没有解决专职APC或癌细胞是否呈递治疗相关的HLA II类抗原。此外,目前尚不清楚肿瘤抗原的HLA II类加工主要依赖于吞噬作用还是自噬作用。根据哪种途径在相关细胞类型中占主导地位,在哪些蛋白质优选作为HLA II类肽配体的来源方面可能存在巨大差异。使问题更加复杂的是,没有系统的方法来确定蛋白质内的哪些区域最有可能产生HLA II类配体,尽管普遍的理论认为蛋白质序列特征应该影响HLA II类的加工潜力。

[0841] 为了研究可治疗靶向的HLA II类抗原的加工和呈递规则,采用了i)改进肽结合预测和ii)确定HLA II类配体在TME中如何加工和呈递的双管齐下的方法。为了学习等位基因特异性肽结合规则,开发了被称为MAPTAC™(带标记等位基因构建体的单等位基因纯化)的可扩展单等位基因HLA配体组分析工作流程,该工作流程利用MS对内源性呈递的HLA II类配体进行测序。MAPTAC™允许清楚地解析40个HLA II类等位基因的肽结合基序,并训练可以准确识别免疫原性病毒表位和新抗原的结合预测算法。为了提高HLA II类加工预测,对肿瘤样品进行了分析,建立了专职APC作为肿瘤内HLA II类表达的主要来源,并定义了这些细胞优先处理的基因和基因区域集。然后证明了整合结合和处理特征的算法可以预测天然APC配体组,更重要的是,可以预测来自内吞癌细胞的HLA II类配体的子集。这些在理解治疗相关HLA II类抗原的加工和呈递规则方面取得的进展将使旨在利用CD4+ T细胞反应的

疗法成为可能。

[0842] 实验程序

[0843] MAPTAC™构建体设计和细胞培养

[0844] 对于HLA I类,α链与C端GSG连接体融合,然后是生物素受体肽(BAP)序列、终止密码子和可变DNA条形码,并克隆到pSF Lenti载体(Oxford Genetics)中。HLA II类构建体也被类似地克隆到pSF Lenti中,其包含β链序列和在C端融合的相同连接体-BAP序列,然后是另一个短GSG连接体、F2A核糖体跳跃序列、带有C端HA标签的α链的序列、终止密码子和可变DNA条形码。MAPTAC™构建体被转染或转导到Expi293、HEK293T、A375、HeLa、KG-1、K562和B721.221细胞中。

[0845] HLA-肽分离方案

[0846] 在冰上将含有$50x10^6$个表达BAP标记的HLA的细胞的快速冷冻细胞沉淀物解冻20分钟,并通过在1.2mL冷裂解缓冲液中手动移液轻轻裂解。清除DNA、RNA和细胞碎片后,将上清液转移到新的1.5mL管中,并通过在室温下与0.56μM生物素、1mM ATP和3μM BirA孵育10分钟来生物素化BAP标记的HLA。将生物素化裂解物与200μL NeutrAvidin树脂在4℃下孵育30分钟,以形成亲和富集的生物素化HLA肽复合物。洗涤后,通过在4℃下以1,500xg离心1分钟将HLA结合树脂沉淀并储存在-80℃,或立即使用Sep-Pak固相萃取进行HLA肽洗脱和脱盐。为了分析健康供体材料的内源性HLA II类配体组,使用内部生成的抗HLA-DR抗体L243或市售的TAL 1B5抗体分离HLA-肽复合物。

[0847] 通过串联质谱法进行的HLA-肽测序

[0848] 所有nanoLC-ESI-MS/MS分析均采用相同的LC分离条件、仪器参数和数据分析。简而言之,使用Proxeon Easy NanoLC 1200色谱分离样品,该系统配备有PicoFrit柱,内部装有C18 Reprosil珠子,并在60℃加热。在依赖数据的采集过程中,将洗脱的肽引入配备有Nanospray Flex离子源的Orbitrap Fusion Lumos质谱仪中。使用Spectrum Mill软件包v6.0 pre-Release解释质谱。通过去除参考数据库中分配给264种常见污染物蛋白质的所有肽以及去除已鉴定的阴性对照MAPTAC™亲和力下拉的肽,对通过PSM FDR估计值<1%的已鉴定的肽进行过滤。此外,所有映射到参考数据库的计算机胰蛋白酶消化物的肽都被删除,以解决胰蛋白酶样品残留。原始质谱数据集将在接受后存放在MassIVE(massive.ucsd.edu)中。

[0849] 结合基序和结合预测的机器学习方法

[0850] 对于每个等位基因,训练了一组卷积神经网络以区分MAPTAC™肽和乱序诱饵。每个网络包含两个ReLU激活的卷积层,每个层都有50个6宽的过滤器。每层每个过滤器的最大和平均激活被路由到具有sigmoid激活的最终密集层。正则化通过L2范数、每个卷积层后20%的空间丢弃和提前停止来实现,并根据非冗余肽的保留分区(约12.5%)调整每个等位基因。在性能基准化中,NetMHCIIpan-v3.1预测被计算为每个查询肽内的最高评分的15聚体,这种方法的表现均优于原始NetMHCIIpan-v3.1预测。

[0851] CD4+ T细胞诱导测定

[0852] PBMC与肽脉冲的mDC以1:10的比例共培养,总共3次刺激。然后将诱导的T细胞用之前描述的独特的双色条形码标记,并与肽脉冲和成熟的自体mDC以1:10的比例培养过夜。随后通过流式细胞术评估细胞响应于肽的IFN-γ产生。对肽有阳性反应的诱导样品是比无肽

对照以高3%诱导IFN-γ产生的样品。

[0853] SILAC标记的肿瘤细胞的APC内吞作用

[0854] K562细胞(ATCC,Manassas,VA)在含有重同位素氨基酸L-赖氨酸2HCl$^{13}C_6{}^{15}N_2$(Life Technologies)和L-亮氨酸$^{13}C_6$(Life Technologies)的SILAC(ThermoFisher)的RPMI培养基中生长5次倍增。单核细胞衍生的树突细胞(mDC)与紫外线处理的K562细胞以1:3的比例共培养过夜,或与HOCl处理后产生的裂解物共培养5小时。收获细胞、沉淀并在液氮中快速冷冻以用于蛋白质组学分析。

[0855] 结果

[0856] MAPTAC™:用于单等位基因HLA II类配体概况分析的可扩展平台

[0857] 目前对HLA II类结合基序的了解主要基于使用两种生化结合测定产生的数据。在前一种方法中,将测定肽和放射性标记的竞争肽与细胞来源的HLA提取物共同孵育以确定IC50。在另一种方法中,构象特异性抗体测量与测定肽结合的HLA的比例以确定EC50。来自这些测定的数据被编译在免疫表位数据库(IEDB)中,并用于训练HLA II类预测算法,如NetMHCIIpan。五个最常见的高加索人HLA-DRB1等位基因在IEDB中得到了很好的支持(每个3326-8967个肽),尽管其中只有约29%是强结合物(亲和力<100nM),并且85%的IEDB肽总体上是精确的15聚体(图12B、图12E)。HLA-DP和HLA-DQ等位基因和非高加索人HLA-DR等位基因(例如HLA-DRB1*15:02)的数据要少得多。

[0858] 为了创建具有等位基因宽度的高质量数据集以支持不同的患者群体,开发了MAPTAC™,该技术能够有效分离结合单个等位基因的HLA II类肽,用于基于MS的鉴定(图11A)。选定的HLA II类异二聚体的α和β链由遗传构建体编码,其中生物素受体肽(BAP)序列位于β链的C端。由于HLA-DRA在功能上是不变的,MAPTAC™产生单等位基因HLA-DR结果,而不管外源性β链和内源性α链之间的潜在配对如何。对于表现出一组有限功能性α链变体的HLA-DP和HLA-DQ,选择具有匹配或未表达的α链等位基因的细胞系。重要的是,这种方法也适用于HLA I类,将BAP标签附加到HLA I类重链上。

[0859] 48小时转染实现了MAPTAC™构建体的稳健表达(图12C),其具有正常水平的细胞表面呈递(图34A和34B)。这在7个不同的细胞系(expi293、A375、KG-1、K562、HeLa、HEK293和B72.221)中并对40个HLA II类等位基因中得到证实,提供了所有三个典型HLA II类基因座的数据:HLA-DR、-DP和-DQ。通过质量控制过滤器的每个重复(约5000万个细胞)的独特肽鉴定的平均数目跨等位基因为236至2580个(图29),中值为1319个肽。采用了几种工艺变化来增加数据深度,包括HLA-DM过表达、肽还原和烷基化。只有一小部分MS命中对应于已知污染物、胰蛋白酶肽和模拟转染(空质粒)(图11B和图29)。MAPTAC™ HLA I类和HLA II类肽的长度分布与之前使用基于抗体的下拉的MS研究中观察到的相匹配(图11C)。

[0860] 在MAPTAC™ HLA II类肽中,大多数氨基酸以与源蛋白质组频率一致的水平表示(图12D和图12F)。例外情况包括C、M和W,它们分别消耗了85%、34%和42%,与之前基于MS的HLA II类肽研究一致。HLA II类肽的还原和烷基化几乎使C的频率增加了三倍,尽管它在蛋白质组方面仍然表示不足(图12F)。在来自IEDB的等位基因匹配的高亲和力肽(<100nM)中未观察到C、M和W的消耗。相反,与IEDB非结合物(>5000nM)相比,IEDB结合物表现出D(-39%)和E(-37%)的消耗以及M(+65%)的富集。因此,MAPTAC™表现出与其他技术观察到的偏差一致的定义偏差。

[0861] MAPTAC™解析HLA II类肽结合基序

[0862] MAPTAC™用于解析等位基因特异性HLA II类结合基序。分析了40个HLA II类等位基因,其中15个以前未表征(在IEDB中具有<100nM亲和力的<30个肽),包括在非高加索人群中常见的等位基因(DRB1*12:02、DRB1*15:03和DRB1*04:07)。由于HLA II类肽可能比结合沟中的残基数更长,因此无法立即确定每个肽的哪个部分是HLA相互作用("核的心")与突出部分;然而,解析结合核心对于表征结合基序至关重要。为了鉴定结合核心,使用工具GibbsCluster-2.0将肽与共有结合核心对齐,该工具使用期望最大化算法为每个肽迭代指定一个结合寄存器,并重新学习跨肽的结合基序。除了少数例外,常见HLA-DR等位基因的结合核心基序与基于IEDB的基序非常一致(图35)。MAPTAC™观察到的肽对于常见等位基因并不总是显示出强的NetMHCIIpan评分(图36A);然而,观察到的被NetMHCIIpan不佳预测的结合物显示具有非常强的测量亲和力(图36B),表明这些观察结果不太可能是假阳性。值得注意的是,MAPTAC™基序在多个细胞系中总是稳定的(图36C)。

[0863] 通常,MAPTAC™和IEDB在锚位置(~4个最高度保守的位置)的最高频率氨基酸上一致,但MAPTAC™基序通常表现出较低的熵(表现为序列标识中较高的字母高度)。有趣的是,当细胞用MAPTAC™构建体和HLA-DM共转染时,大多数等位基因的锚位置的熵甚至进一步降低(图30A和图37A)。这在12个HLA-DR等位基因中一致地观察到,显示HLA-DM作为组库"编辑器"的普遍作用,并表明基于缺乏HLA-DM和其他加载伴侣的亲和力测定的模型可以学习在体内不适用的结合规则。HLA-DM对CLIP肽的存在的影响也很明显。在没有HLA-DM共转染的情况下,在10个HLA-DR等位基因和匹配的已知CLIP变体中观察到CD74衍生肽(图37C);同时,在我们的任何HLA-DM共转染实验中均未观察到CLIP肽。

[0864] 对于分析的HLA-DP等位基因(图30A和图37A),HLA-DM的影响不明显,这可能与以前未报道的异常带正电荷的PI锚的存在有关。HLA-DM被认为主要作用于结合肽的N端侧,因此,不寻常的PI锚不是HLA-DM不敏感的标志,因为具有疏水性PI锚的HLA-DP基序也未因HLA-DM的存在而改变(图37A)。另一方面,HLA-DQB1*06:04/A1*01:02受到HLA-DM的深刻影响(图30A)。如果没有HLA-DM共转染,该等位基因的结合基序是不可区分的,这表明无分子伴侣加载到一些HLA-DQ等位基因上会产生很大比例的非生理结合物。

[0865] 鉴于已发布的多等位基因HLA II类数据集的可用性,我们使用计算机去卷积方法研究了我们的等位基因特异性肽是否可以被有效鉴定。几个研究小组在从多等位基因HLA I类数据中去卷积HLA I类等位基因基序方面取得了成功;然而,由于需要同时解析每个肽的结合核心和等位基因分配,HLA II类基序的去卷积变得复杂。为了评估HLA II类去卷积的准确性,从通过泛DR抗体分析的八个样品(PBMC和已发表的细胞系)中分析了HLA-DR配体组。对于每个数据集,在与样品基因型中的每个等位基因(1-2个DR1等位基因加上0-2个DR3/4/5等位基因,取决于单元型和接合性)匹配的单等位基因数据中掺加了20个肽。GibbsCluster工具(也可用于去卷积)用于将肽分组并观察掺加肽是否根据其已知起源等位基因被适当地共聚簇。在所有情况下,肽分布在不同的簇中,仅显示与正确源等位基因的适度关联(图30B),并表明基于去卷积的HLA II类训练数据可能存在重大错误。

[0866] 为了了解去卷积的不良性能,审查了单等位基因MAPTAC™数据以确定可以作为GibbsCluster路标的"明显"锚的频率。因此,定义了每个HLA II类等位基因的每个锚位置(具有最低熵的四个位置)的明显氨基酸(频率>10%)。只有10-20%的肽在所有四个锚位置

表现出理想的残基,多达50%的肽表现出两个或更少的明显锚(图30C)。鉴于表现出大多数预期锚的肽的频率较低,因此在纯计算基础上很难对大部分肽进行分类也就不足为奇了。因此,MAPTAC™解决了使用计算机方法难以解决的不确定性的主要来源。

[0867] HLA I类等位基因的基序也可以使用MAPTAC™定义。这包括之前未定义结合谱的等位基因(例如B*52:01,日本常见)。对于先前表征的等位基因,可以看出源自基于亲和力的方法和先前的单等位基因MS研究的基序之间存在良好的对应关系。尽管如此,值得注意的是,在基于多等位基因MS的研究中存在一些差异,这些研究采用去卷积方法来定义基序(图37B)。

[0868] 在MAPTAC™数据上训练的算法预测免疫原性

[0869] 考虑了MAPTAC™数据是否可以生成准确度更高的HLA II类结合预测器。由于HLA II类肽的HLA结合子序列相对于N或C端不是处于固定位置,因此学习算法必须动态考虑每个肽的不同结合核心可能性。为了解决这一限制,采用了卷积神经网络(CNN),由于它们精通平移不变模式识别,因此在计算机视觉领域取得了成功。对于每个等位基因,训练了一组CNN(图31A),称为整体预测器neonmhc2。”

[0870] 考虑到MS表现出一定程度的氨基酸残基偏好,特别是针对C,通过随机排列观察到的结合物的序列(称为命中)来生成阴性训练示例(称为诱饵)。由于这种方法存在学习天然蛋白质序列特性的风险,因此从HLA II类配体的肽源基因的未观察到的子序列中随机抽取诱饵。为了计算每个等位基因的阳性预测值(PPV),将n个MS观察到的肽与从同一组源基因中采样的19n个长度匹配的诱饵一起评分,每个预测器的n个排名靠前的肽(即前5%)被称为阳性。在这种情况下,PPV与判定相同,因为假阳性的数目和假阴性的数目相等。以1:19命中:诱饵比计算阳性预测值(PPV)表明,neonmhc2相对于NetMHCIIpan在预测MAPTAC™观察到的肽方面提高了PPV(图31B;表11)。

[0871] 表11

[0872]

MHC II 类等位基因	对于 NetMHCIIpan 的 PPV	对于 *neonmhc2* 的 PPV
DRB1*16:01	0.13	0.66
DRB1*15:01	0.17	0.61
DRB4*01:01	0.23	0.62
DPB1*02:01/DPA1*01:03	0.12	0.52
DRB1*11:04	0.22	0.60
DRB1*14:01	0.14	0.59
DRB1*13:03	0.08	0.58
DPB1*06:01/ DPA1*01:03	0.01	0.48
DRB3*03:01	0.21	0.57
DRB1*03:01	0.35	0.55
DRB1*01:01	0.34	0.56
DRB5*01:01	0.35	0.54
DRB1*01:02	0.27	0.54
DRB3*01:01	0.43	0.54
DPB1*01:01/DPA1*01:03	0.08	0.53
DRB1*07:01	0.29	0.54
DRB1*04:04	0.22	0.55
DRB1*11:01	0.17	0.52
DRB1*15:03	0.02	0.50
DPB1*04:01/DPA1*01:03	0.27	0.52
DPB1*04:02/DPA1*01:03	0.25	0.53

[0873]

DRB1*15:02	0.17	0.50
DRB1*10:01	0.15	0.48
DRB1*08:02	0.21	0.50
DRB1*13:01	0.16	0.49
DRB1*04:05	0.13	0.50
DRB1*09:01	0.19	0.47
DQB1*06:02/DQA1*01:02	0.14	0.49
DRB1*11:02	0.07	0.46
DRB1*12:01	0.1	0.46
DRB1*04:01	0.27	0.45
DRB1*04:02	0.1	0.46
DRB1*04:03	0.18	0.45
DRB3*02:02	0.22	0.45
DRB1*08:04	0.1	0.44
DRB1*12:02	0.12	0.44
DPB1*03:01/DPA1*01:03	0.03	0.48
DRB1*04:07	0.21	0.40
DRB1*08:01	0.11	0.37
DQB1*06:04/DQA1*01:02	0.06	0.37
DRB1*03:02	0.31	0.29
DRB1*08:03	0.1	0.24
DRB1*13:02	0.06	0.22

[0874] 饱和度实验,其中训练数据集大小被不同程度地降采样,表明neonmhc2的性能受数据限制并且可能会随着更多数据而改善(图38Ai)。

[0875] 对图30B中HLA II类去卷积的低保真度观察的分析表明,如果没有单等位基因数据,就无法实现可比较的预测性能。为了测试这一点,遵循了最近发布的计算工作流程,该工作流程使用去卷积来训练多等位基因MS数据上的等位基因特异性结合预测器(Barra等人,2018)。检查十一个多等位基因样品(图30B中的相同样品)的GibbsCluster标识,观察到许多簇(13/32)与样品中已知的等位基因没有任何相似性(图38Aii)。使用关于基序应该是什么样子的预先存在的知识,仅选择合量的簇(标记在图38Aii中)并构建具有相同CNN架构

的预测器。然后在真正的单等位基因数据(未用于训练的MAPTAC™数据的保留分区)上与neonmhc2一起评估这些模型。在去卷积多等位基因数据上训练的模型通常超过NetMHCIIpan,但不如MAPTAC™训练的neonmhc2(图31E)。即使对MAPTAC™数据集进行降采样,使得各个训练数据集的大小相同,单等位基因数据的优越性也得以保持。

[0876] 为了确保在对非MS数据进行评估时,明显的预测改进将保持不变,对等位基因特异性CD4+记忆T细胞反应的大型数据集进行了策划,这些反应是通过四聚体引导的表位作图(TGEM)检测到的。值得注意的是,这些四聚体数据依赖于无伴侣肽交换,因此它们可能会受到与传统亲和力测定相同的偏见(Archila和Kwok,2017年)。尽管如此,neonmhc2对所有等位基因都优于NetMHCIIpan,具有足够的评估数据(至少20个阳性示例)(图31C)。NetMHCIIpan的性能(由PPV衡量)是可变的,对于DRB1*15:01下降到低至5%(相比之下,neonmhc2的性能从未低于30%PPV),并且仅在两个等位基因上接近neonmhc2,包括经过充分研究的HLA-DRB1*01:01。另一方面,neonmhc2在所有其他评估的等位基因上表现出令人信服的改善,包括两个最常见的高加索人HLA-DR等位基因(DRB1*07:01和DRB1*15:01)。这些结果表明,neonmhc2对NetMHCIIpan的预测改进可以在非基于MS的基准测试中得到验证,并且可能扩展到大多数等位基因。

[0877] 为了评估neonmhc2的治疗相关性,确定了neonmhc2是否可以识别能够在离体诱导测定中引发CD4+ T细胞反应的新抗原(参见方法)。关注DRB1*11:01,它是IEDB中具有许多亲和力测定确认的结合物的常见等位基因(仅被DRB1*01:01和DRB1*07:01超越;图12E),一组对癌症基因组图谱(TCGA)观察到的新抗原序列进行评分,并选择了neonmhc2优选(预测的前1%)但NetMHCIIpan未选择的子集(预测的后90%)。通过去除可能结合诱导材料中存在的其他HLA-DR等位基因的肽,进一步完善了该组。大多数neonmhc2-选择的肽(8/12)产生CD4+ T细胞应答,如通过响应于肽回忆的IFNγ表达所测量的(图31D、图38B和图38C)。这些结果表明,MAPTAC™训练的预测器可以识别NetMHCIIpan未识别的免疫原性HLA II类新抗原序列。

[0878] 专职APC是肿瘤微环境中占主导地位的HLA II类呈递者

[0879] 开发了一种能够表征和预测HLA II类等位基因特异性肽结合偏好的技术,试图通过进一步深入了解抗原加工来补充结合预测的改进,这对于确定最有可能产生HLA II类癌症抗原的蛋白质序列至关重要。为了在TME的背景下解决这些问题,我们分析了非MAPTAC™数据集,包括单细胞RNA-Seq和已发表的基于MS的研究,这些研究调查了肿瘤中的HLA II类配体组。考虑了微环境中哪些细胞类型最有可能呈递可治疗靶向的癌症抗原。目前,对于癌症抗原是由具有内吞肿瘤蛋白的专职APC还是由肿瘤细胞本身呈递,尚无共识。为此,在描述肺癌、头颈癌、结直肠癌、卵巢癌和黑素瘤的五个已发表的单细胞RNA-Seq数据集中分析了HLA-DRB1表达,并发现典型的APC(巨噬细胞、树突细胞)和B细胞)比TME中的肿瘤细胞和其他基质细胞类型表达更高水平的HLA II。该观察结果在多个患者和肿瘤类型中是一致的(图19A)。由于肿瘤细胞在TME中的数目可能超过APC,因此它们较低水平的HLA II表达可能与免疫学相关。为了评估总体HLA II类表达中有多少来自肿瘤细胞与基质,鉴定了HLA II类特异性基因(重点是CIITA、CD74和CTSS)突变的TCGA患者,并确定了展示体细胞变异的RNA-Seq读数的分数,以估算源自肿瘤与基质的HLA-DRB1表达的比例(图19B,参见方法)。基于在代表17种不同肿瘤类型的153名患者中鉴定的突变,大多数HLA II类表达似乎来自非

肿瘤细胞。事实上,45％的患者显示出零肿瘤衍生的HLA II类表达。专注于T细胞浸润水平最高的患者(前10％,如使用先前公布的18基因特征(Ayers等人,2017)确定的),低肿瘤HLA-DR表达似乎仍然是常态,只有316名患者表达>1000TPM。为了探究免疫疗法是否破坏了这一趋势,我们分析了来自检查点阻断反应性肿瘤类型的额外单细胞RNA-Seq,并评估了治疗前后HLA-DRB1的表达。包括一名确认为反应者的黑素瘤群组在治疗前和治疗后活检中均显示均一地很低的肿瘤细胞的HLA-DRB1表达(图19C)。对抗PD1疗法显示出55％临床反应率的基底细胞癌群组同样表现出低的肿瘤细胞来源的HLA-DRB1表达,无论时间点如何(图19C)。这些结果表明,大多数肿瘤内HLA II类表现主要由专职APC驱动,“热”TME条件并不能保证与这种一般模式的差异。

[0880] 特定基因有权进入HLA II类呈递途径

[0881] 为了确定优先由肿瘤驻留APC呈递的表位的源基因,以及它们是来自自噬还是内吞作用,对三项已发表的HLA II类配体组研究进行了分析,这些研究是使用肿瘤组织进行的。

[0882] 首先,假设每个基因的观察次数应与其长度和表达水平的乘积成正比,量化每个基因在肿瘤HLA II类配体组中的代表程度(图18B)。尽管未在天然组织中表达,但观察到人血浆中表达的蛋白质明显富集,尤其是白蛋白、纤维蛋白原、补体因子、载脂蛋白和转铁蛋白。考虑到这些鉴定代表HLA配体组中的非特异性结合,评估了四个PBMC HLA-DR配体组中血浆衍生肽的neonmhc2结合评分(图39A);这些肽显示出很强的结合分数,表明它们是HLA结合的。血浆来源的蛋白质在肿瘤HLA I类配体组数据中没有显著富集(图39B)。HLA II类配体组中血浆基因的富集是一致的APC通过微胞饮作用从组织血清中“吸食”细胞外蛋白。还观察到参与白细胞细胞粘附的基因的额外富集,例如ITGAM(11x富集的)、LCP1(8x)、ITGAV(6x)和ICAM1(6x),表明APC正在积极回收它们自己的膜。最近报道称MUC16富含卵巢癌HLA I类配体组,但并没有过度表示。

[0883] 细胞定位也被认为可以进一步询问HLA II类抗原呈递途径中的基因偏好。当基因按定位分组时,分泌基因和膜基因的表达频率是基于基因表达的预期的两倍,强调了巨胞饮作用在塑造HLA II类配体组中的重要作用。尽管如此,超过一半的HLA II类肽来自与巨胞饮作用不一致的区室,例如细胞核和细胞质。有理由认为,如果这些基因中有许多是通过自噬呈递的,那么已知被蛋白酶体清除的基因应该存在相应的缺陷。事实上,已知含有泛素位点的蛋白质产生肽的频率低于基于它们的长度和表达所预期的肽(图32C)。对于已知在蛋白酶体抑制后水平增加的蛋白质,也观察到消耗。这些是自噬的预期模式,但不一定是吞噬作用,这表明APC肽配体组部分代表了它们自己的细胞内蛋白质组。

[0884] 为了解决TME中APC呈递的HLA II类抗原的起源,考虑是否有可能通过确定核和细胞溶质肽鉴定与APC特异性或大块肿瘤基因表达谱(图39C)。尽管估计值存在很大的不确定性(通过基于回归的模型和自举重采样评估;补充方法),HLA II类配体组最好通过肿瘤和APC基因表达谱的混合来解释。结合观察到的蛋白酶体清除蛋白的消耗,该结果表明肿瘤内APC呈递出外源性和内源性蛋白质的混合物。

[0885] 一些基因区域被优先加工但缺乏明显的切割基序

[0886] 关于哪些序列优选用于抗原加工有多种理论(图32D)。根据一种模型,酶在结合II类HLA之前切割源蛋白,就像I类HLA的情况一样(Sercarz和Maverakis,2003)。第二个模型

假设肽结合首先发生,然后结合的肽被外肽酶修剪,直到它们受到HLA II类的空间阻碍。在第三个模型中,肽切割事件发生在HLA结合之前和之后。因为对于如何生成HLA II类肽存在竞争模型,所以生成了三种不同的预测框架(图32D)。第一个假设内肽酶占主导地位("先切割");第二个模型假设HLA II类与全长蛋白质结合,随后被外肽酶向内修剪("先结合");第三个模型假设酶消化发生在HLA结合之前和之后("混合")。每个模型都需要不同的算法方法。具体而言,由先切割观点驱动的算法应关注MS观察到的配体边缘的氨基酸基序;然而,以先结合观点为动机的算法会更好地忽略这些基序并专注于决定HLA结合可及性的局部蛋白质结构特性。受混合模型启发的算法应该在观察到的HLA II类肽的上游和下游寻找候选前体切割位点。

[0887] 在所考虑的三种方法中,只有先切割算法比基线模型产生了可测量的改进(图32E和图40B)。然而,这种方法似乎学习了阳性示例肽中存在的外肽酶修剪的特征(例如,倒数第二个脯氨酸特征(Barra等人,2018)),因为如果查询肽的确切切割位点被掩盖,它就无法增加价值(STAR方法)。

[0888] 转向纯粹经验方法,在已发表的HLA-DQ配体组(Bergseng等人,2015)中观察到的蛋白质区域被编目并使用重叠来预测HLA-DR配体。重叠变量产生了预测性能的适度改进(与单独的neonmhc2相比,PPV平均增加3.1%)(图32E)。假设HLA-DQ和HLA-DR等位基因共享相同的HLA-II加工环境但不共享结合基序,该结果表明某些基因区域确实有利于加工,但与切割基序或构象特性无关。

[0889] 研究小组报告了使用MS观察到的肽的观察末端来训练处理算法的积极结果,这种方法假定"先切割"模型。然而,在审查多种不同细胞系和组织类型中与肽末端相邻的氨基酸富集时(图40A),观察到的模式似乎与结合后修剪更一致。这些包括与HLA II类加工酶——组织蛋白酶S——的已知基序缺乏对应关系,以及在倒数第二个肽位置富集不易切割的P,如果P阻止修剪酶的加工,可能会出现这种基序。为了测试"先切割"假设是否正确,神经网络模型在肽末端进行了训练,并以两种不同的方式对其进行评估:i)对每个肽的确切N和C末端的可切割性进行评分或ii)对每个肽的预测结合核心±15AA范围内的最佳位点进行评分(补充方法)。假设如果先切割模型正确,则两种方法都应该增加预测值,但只有第一种方法是这样(图32E和图40B)。因此,神经网络可以根据配体(例如倒数第二个P)中的明显特征从诱饵中辨别出HLA II类,但是当切割位点未知时则无关紧要——这在从初级蛋白质序列中预测免疫原性肽时总是如此。这种微妙的区别有可能在该领域引起混乱。

[0890] 根据"先结合"理论,对MS观察到的和诱饵肽的溶剂可及性以及内在无序结构域进行评分。如果蛋白质结构决定了HLA结合的可用性,则可以在HLA II类配体中富集溶剂可及或无序结构域。然而,这些特征也证明是非预测性的(图32E)。然后考虑混合模型,其中酶在肽结合之前部分消化蛋白质,之后发生额外的修剪。在该模型中,前体切割位点存在于观察到的MS配体末端的上游和下游。因此,基于扩展的蛋白质背景(±30AA)对CNN进行了训练,以检测与前体切割相对应的远端信号。该模型也没有显示预测值(图32E)。最后,由于加工优选区域证明难以根据一级序列进行预测,因此对已发表的HLA-DQ配体组中观察到的蛋白质区域进行了编目,并使用重叠来预测HLA-DR配体。重叠变量产生了预测性能的适度改进(与单独的neonmhc2相比,PPV平均增加3.1%)(图32E)。假设HLA-DQ和HLA-DR等位基因共享相同的HLA II类加工环境但不共享肽结合基序,该结果表明某些基因区域确实有利于加工

但不显示明显的切割基序或特殊的构象特性。

[0891] 整合呈递规则大大增强了HLA-DR配体组预测

[0892] 为了量化结合规则如何与处理相关特征协同作用,创建了一个多变量模型来预测HLA II类呈递细胞系、树突细胞和健康供体外周血单核细胞(PBMC)的HLA-DR配体组。尽管所呈递的肽没有突变,但预测场景模拟了新抗原预测,其中必须根据其产生HLA II类肽的能力来评估随机采样的基因组位点。使用1:499的命中与诱饵之比,并从蛋白质编码外显子组中随机对诱饵采样,评估了基于neonmhc2和NetMHCIIpan的模型的性能,以及包含额外处理特征的模型,包括RNA-Seq衍生表达、基因水平偏差(根据图32A,参见相关的图39B),并且与先前观察到的HLA-DQ肽重叠。为了使模型与突变的肿瘤表位靶标在癌症治疗中的优先级一致,修改了基因水平偏差特征以抵消对血浆基因的偏好,这些基因不是新抗原的相关来源。

[0893] 这些综合算法证实了结合和加工预测的实质性改进(图21A)。具体来说,根据评估的数据集,完整模型比单独使用NetMHCIIpan结合预测的模型显示了7.4倍到61倍的变化改进。表达和基因偏好都为预测准确性提供了实质性的独立贡献。DQ重叠功能的贡献较小,但始终提供积极的改进。重要的是,即使提供了处理相关预测变量的全部好处,基于亲和力的模型的准确度也只有基于MAPTAC™的模型的一半。

[0894] 使用专职APC呈递的肿瘤衍生HLA II类肽的基准化预测准确性

[0895] 评估了我们预测HLAII类配体组的准确性后,注意力转移到测试是否可以预测由专职APC内吞的肿瘤衍生配体。我们观察到TME中大多数HLA II类表达来自专职APC,这表明该加工途径可能是与肿瘤抗原呈递最相关的途径。遗憾的是,传统的基于MS的肿瘤组织配体组不能鉴别哪些肽源自内吞的肿瘤蛋白。因此,设计了一项实验,其中分析了已“饲喂”SILAC标记的肿瘤细胞的树突细胞(DC)的HLA-DR配体组(图33A)。

[0896] 为了标记肿瘤来源的蛋白质,HLA II类缺陷的癌细胞系(K562)在含有同位素标记的L和K的培养基中生长,标记效率超过95%。将裂解的肿瘤细胞(模拟肿瘤碎片的巨胞饮作用)或紫外线处理的全肿瘤细胞(模拟全细胞的吞噬作用)饲喂给DC。使用MS对HLA-DR结合肽进行分析,以鉴定带有重标记或轻标记的氨基酸的肽。实验产生了29种重标记的肽,全细胞实验产生了56种重标记的肽,分别用于裂解物和UV实验(表10(数据S1B))。带有超过1个L或K的肽在除两种情况外的所有情况下都显示完全标记,表明重标记的肽源自肿瘤细胞而不是来自新翻译的DC蛋白,这将显示不一致的标记。未处理的DC和与裂解物孵育10分钟后收获的DC均未产生重标记的肽。

[0897] 使用这里公开的整合预测算法,评估了预测肿瘤衍生肽的能力。与我们之前预测天然HLA II类配体组的结果一致,基于neonmhc2的模型实现了比基于NetMHCIIpan的模型高得多的预测准确性(图33D)。

[0898] 与基因表达不同,基因偏好和DQ重叠特征并没有改善对内吞抗原的预测,这表明从大量组织配体组中学习的模式与此类表位无关。分析重标记的肽的源基因,注意到RNA结合蛋白(RBP)、DNA结合蛋白(DBP)、热休克蛋白(HSP)和线粒体蛋白(图21D),而不是在配体组实验中主要看到的分泌蛋白和膜蛋白(图32A)。目前尚不清楚这是否代表了不同的加工偏好。事实上,源蛋白通常在K562中高度表达(中值表达为430TPM,而未标记的肽的中值表达为130TPM),这表明检测极限可能会驱动观察到的基因偏好。

[0899] 为了清楚起见,建立了逻辑回归模型来测试基因定位和功能类别是否可以在已经考虑基因表达的模型之外改进肽预测。当考虑结合和表达时,RBP、DBP和HSP不再显著,但线粒体蛋白仍然显著(p=2.6e-4:图33E)。值得注意的是,富集模式与在轻标记的肽中观察到的完全不同。

[0900] 为了确定线粒体富集是否可以改善预测,我们从新供体收集数据,目的是通过增加细胞输入、仅关注紫外线处理方案并在过夜时间点之外增加24小时孵育时间点来实现更深入的覆盖。该实验在过夜和24小时时间点分别产生了77和59个重标记的肽,并共同鉴定了78个独特的源基因。使用解释线粒体偏好的逻辑回归模型(在原始SILAC数据上训练),我们能够通过比仅包括结合和表达的模型净增加8-12%来提高PPV(图33G)。这些改进是显著的(分别为16小时和24小时,p=1.1e-9和p=1.5e=-8)。这些偏好无法从大量配体组中学习到,可用于实现更准确的表位预测。

[0901] TME中HLA II类呈递的存在与接受癌症免疫疗法治疗的患者的阳性结果相关。不幸的是,HUAII类配体预测的不准确以及肿瘤抗原如何在TME中呈递的模糊性已经减缓了针对HLA II类抗原的疗法的发展。因此,如本文所述开发了称为MAPTAC™的单等位基因分析技术,并全面分析了肿瘤配体组以定义HLA II类配体处理规则。MAPTAC™能够快速分析40个HLA II类等位基因,包括覆盖95%的美国患者的35个HLA-DRB1等位基因。此外,我们在MAPTAC™数据上训练的结合预测算法neonmhc2在预测记忆CD4+ T细胞反应方面的表现优于NetMHCIIpan,即使对于具有最多可用于NetMHCIIpan训练的预先存在的亲和力测量的等位基因也是如此。观察到,在TGEM验证数据集中,neonmhc2在识别记忆CD4+ T细胞反应方面的性能优于NetMHCIIpan。此外,本文公开的算法还擅长预测针对新抗原的离体诱导的CD4+ T细胞反应,成功识别了NetMHCIIpan不会优先考虑的免疫原性新表位。同时,对单细胞RNA-Seq肿瘤数据的分析表明,最相关的肿瘤抗原可能主要通过浸润APC吞噬肿瘤细胞来表达。因此,研究了哪些基因和基因区域优先出现在TME中,并创建了多变量模型,以准确预测吞噬APC提供的HLA-DR配体组和肿瘤衍生配体。这些模型大大超过了NetMHCIIpan的阳性预测值。

[0902] 与传统的肽结合分析相比,使用MAPTAC™直接分析内源性加工和呈递的HLA II类配体的一个优点是存在加载肽的分子伴侣,例如HLA-DM。已知HLA-DM在编辑APC的HLA II类肽组库中发挥作用,这促使我们研究其差异表达对HLA II类配体的影响。当HLA-DM在HLA-DR MAPTAC™实验中过度表达时,结合基序比没有HLA-DM过度表达的实验更清楚。令人惊讶的是,HLA-DM对HLA-DQB1*06:04/A1*01:02具有深远的影响,这表明学习某些HLA-DQ等位基因的准确肽结合规则可能需要存在这种肽加载伴侣。相反,两个HLA-DP等位基因没有显示出影响(Yin等人,2015),表明HLA-DM敏感性和P1锚偏好之间存在关系,这对于这两个HLA-DP等位基因来说是不寻常的。除了HLA-DM之外,MAPTAC™平台提供了一种快速了解参与HLA II类途径的其他关键分子伴侣和蛋白质(如CD74或HLA-DO)如何影响HLA II类等位基因的肽结合库的方法。

[0903] 关于肿瘤生物学,我们最重要的观察结果是APC负责TME中对所评估肿瘤类型的显性HLA II类表达。这表明治疗相关肿瘤抗原的呈递可能取决于凋亡肿瘤细胞的吞噬作用或分泌的肿瘤蛋白的巨胞饮作用。尽管有直接杀死CD4 T细胞的报道,但提供的数据表明CD4 T细胞通常在TME中发挥支持作用,主要识别浸润白细胞上呈递的肿瘤抗原。因此,CD4 T细

胞的抗肿瘤作用可能主要由调节其他免疫细胞(包括具有直接溶细胞功能的免疫细胞)的运输和活化的趋化因子和细胞因子的分泌介导。虽然这在机制上更复杂,但一个好处是肿瘤对HLA II类抗原是否被呈递的控制较少,这表明通过功能丧失突变的免疫逃逸——这是肿瘤避免HLA I类呈递的常见机制,可能HLA II类不那么频繁。未来的研究仔细定义哪些APC群体负责呈递内吞的肿瘤抗原,以及是否有办法增强这些吞噬细胞向TME的募集,将对该领域有益。此外,了解不同的肿瘤细胞死亡模式(如低氧、化疗和放疗)如何导致这些APC捕获不同水平的肿瘤抗原,这可能会导致与HLA II类靶向疗法的最佳治疗组合。

[0904] 最后,对HLA II类配体组的综合分析导致观察到某些基因似乎比其转录物表达水平预测的更频繁地出现。从肿瘤细胞中学习基因水平偏差有助于改进对APC HLA II类配体组的预测;然而,这些信号中的一些可能与新抗原预测不太相关。例如,检测到的富集似乎与APC中的自噬和膜再循环有关,而不是与外源抗原的摄取有关。有趣的是,当在体外将"肿瘤细胞"喂给树突细胞时,源基因鉴定反而显示了RNA结合蛋白的富集。推测优先呈递RNA结合蛋白是很吸引人的,因为这种机制将促进病原体表位的呈递,并可能解释在系统性红斑狼疮和其他自身免疫疾病中观察到的针对RNA结合蛋白的反应性。在任何情况下,重要的是要注意我们基于SILAC的HLA配体组学工作流程的用途不仅限于肿瘤抗原,因为它还可以应用于研究涉及传染病和自身免疫的抗原。

[0905] 总之,HLA II类处理和呈递的规则明显比HLA I类复杂。由于这个原因,驱动CD4+T细胞应答的抗原通常仍未确定。我们在定义HLA II类结合和处理规则方面取得的进展将能够识别可靶向的癌症抗原和其他疾病相关表位,这些表位可以转化为更有效的治疗方法。

[0906] 实施例13:补充信息

[0907] 具有相关元数据的实验和数据源的总结

[0908] 数据集的详尽列表,包括MAPTAC™数据、非MAPTAC™手稿数据和以前发布的数据。在适当的地方提供了相关的相关特征,例如样品基因型。B)在实验性MAPTAC™重复、PBMC供体、细胞系和SILAC饲喂实验中合并的独特肽鉴定。污染物和完美胰蛋白酶肽被去除。参见例如至少图12E-12F、34A-34B等。

[0909] 用于去卷积分析的掺加肽

[0910] 掺加分析中每个等位基因使用的20个示例肽的示例列表。通过要求最小SPI为70,长度在12到20个氨基酸之间,并且不允许与针对其他MAPTAC™分析的DR等位基因观察到的任何结合物有9聚体重叠来选择肽。此外,对于给定的等位基因,没有两个掺加肽共享9聚体。参见例如至少图35、36A-36C等。

[0911] 所选等位基因的整理TGEM数据集,补充实验程序

[0912] 不同病原体和变应原肽的DRB1*01:01、DRB1*03:01、DRB1*04:01、DRB1*07:01、DRB1*11:01和DRB1*15:01的HLA II类四聚体结果及其相应的NetMHCIIpan和neonmhc2预测。数据来自Kwok及其同事发表的论文。参见图38A-38C等。

[0913] 补充方法

[0914] HLA II类等位基因频率和亲和力数据统计,与图12A和12E相关。

[0915] 等位基因频率获自资源bioinformatics.bethematchclinical.org/hla-resources/haplotype-frequencies/high-resolution-hla-alleles-and-haplotypes-

in-the-us-population。mhc_ligand_full.csv数据集于2018年9月21日从IEDB数据(iedb.org/database_export_v3.php)下载。要求有效的亲和力测量具有等于“细胞MHC/竞争性/荧光”、“细胞MHC/竞争性/放射性”、“细胞MHC/直接/荧光”、“纯化的MHC/竞争性/荧光”、“纯化的MHC/竞争性/放射性”或“纯化的MHC/直接/荧光”的“方法/技术”,和等于“解离常数KD”、“解离常数KD(~EC50)”、“解离常数KD(~IC50)”,“半数最大有效浓度(EC50)”或“半数最大抑制浓度(IC50)”的“测定组”。如果字符串“Buus”出现在“作者”字段中,则测量归于Søren Buus小组(丹麦哥本哈根大学)。否则,如果作者字段包含字符串“Sette”或“Sidney”,则测量归于Alessandro Sette小组(La Jolla Institute for Immunology,U.S.A)。所有其他测量值都标记为“其他”。为了列举强结合物,仅计算亲和力大于100nM的肽。

[0916] MAPTAC™方案概述,与图2相关:DNA构建设计

[0917] HLA I类和HLA II类等位基因的基因序列通过IPD-IMGT/HLA网页(ebi.ac.uk/ipd/imgt/hla)确定,并用于设计重组表达构建体。对于HLA I类,α链与C端GSGGSGGSAGG连接体融合,然后是生物素接受肽(BAP)标签序列GLNDIFEAQKIEWHE、终止密码子和可变DNA条形码,并通过NcoI和XbaI限制位点克隆到pSF Lenti载体(Oxford Genetics,Oxford,UK)中。HLA II类构建体类似地通过NcoI和XbaI限制位点克隆到pSF Lenti中,并包含在C端与来自I类构建体的连接体-BAP序列(SGGSGGSAGGGLNDIFEAQKIEWHE)融合的β链序列,然后是另一个短GSG连接体、F2A核糖体跳跃序列(VKQTLNFDLLKLAGDVESNPGP)、α链序列、HA标签(GSYPYDVPDYA)、终止密码子和可变DNA条形码。对于所有DR等位基因,β-链与DRA*01:01配对。HLA-DM构建体的克隆与HLA II类构建体类似,只是它缺少BAP序列和HA标签。将HLA-DM添加到HLA II类实验的子集中。通过Sanger测序验证所有DNA序列的身份。

[0918] 细胞培养和瞬时转染

[0919] Expi293细胞(Thermo Scientific)在含有8%CO_2的Expi293培养基(Thermo Scientific)中在37℃下以125rpm振荡生长。Expi293细胞保持在0.5x10^6/mL至6x10^6/mL之间的细胞密度,每两周定期传代。30mL Expi293细胞悬液用于以约3x10^6/mL的细胞密度和>90%的活力的瞬时转染。简而言之,将30ug DNA(每毫升细胞悬浮液1μg/mL DNA)稀释到一个管中的1.5mL Opti-MEM培养基(Thermo Scientific)中,同时将80μL ExpiFectamine™ 293转染试剂(Thermo Scientific)稀释到含有1.5mL Opti-MEM的第二管中。这两个管在室温下孵育5分钟,合并,轻轻混合,在室温下孵育30分钟。将DNA和ExpiFectamine混合物加入Expi293细胞,并在37℃、8%CO_2、80%相对湿度下孵育。48小时后,以每管50x10^6个细胞在四次技术重复中收获转染的细胞,离心,用1x Gibco DPBS(Thermo Scientific)洗涤一次,并在液氮中快速冷冻以用于质谱分析。从每个转染批次中收集1x10^6个细胞的等分试样,并通过抗BAP(Rockland Immunochemicals Inc.,Limerick,PA)或抗HA(Bio-Rad,Hercules,CA)Western印迹进行分析,以验证亲和标记的HLA蛋白表达。Expi293的内源性HLA II类基因型被确定为DRB1*15:01、DRB1*01:01、DPB1*04:02、DPA1*01:03、DQB1*06:02、DQA1*01:02(Laboratory Corporation of America,Burlington,NC)。在一些实验中,HLA II类等位基因与HLA-DM共转染,在这种情况下,用于两种质粒的DNA浓度降至0.5μg DNA/mL细胞悬液。

[0920] A375细胞(ATCC)在含有10%FBS的DMEM中生长,并以不超过80%的汇合度保持在培养物中,定期传代。对于质谱实验,A375细胞在500cm^2板中以18.5x10^6个细胞/mL的接种密

度在100mL中培养,根据70％汇合的细胞数计算。24小时后,按照针对总培养体积调整的TransIT系统方案,用TransIT-X2(Mirus Bio,Madison,WI)转染细胞。48小时后,吸出细胞培养基,并用1X Gibco DPBS(Thermo Scientific)洗涤细胞。为了收获,将A375细胞在37℃下与30mL非酶促细胞解离溶液(Sigma-Aldrich)孵育10分钟,离心,用1X DPBS洗涤,并以每个样品50x10^6个细胞等分。293T和HeLa细胞购自ATCC,并分别在37℃、5％CO_2下在DMEM、10％FBS、2mM L-谷氨酰胺或DMEM+10％FBS中培养。使用TransIT LT1试剂(Mirus Bio,Madison,WI)按照制造商的说明用HLA构建体转染两种细胞系,并在转染后48小时如针对A375细胞所述进行处理。从所有样品中,从每次转染中收集1x10^6个细胞的等分试样,并通过抗BAP(Rockland Immunochemicals Inc.,Limerick,PA)或抗HA(Bio-Rad,Hercules,CA)Western印迹法进行分析,以验证亲和标记的HLA蛋白表达。B721.221细胞获自Fred Hutchison Cancer Center(Seattle,WA),并在含有10％热灭活胎牛血清和1％青霉素/链霉素(均为Thermo Fisher Scientific)的RPMI-1640加glutamax(Thermo Fisher Scientific)中培养。细胞每周培养两次,传代25次后弃去。K562细胞和KG-1细胞(ATCC,Manassas,VA)在IMDM(Thermo Fisher Scientific)培养基加10％热灭活FBS、1％青霉素/链霉素、1％丙酮酸钠和1％MEM-NEAA培养基中生长。细胞每周培养两次,传代25次后弃去。

[0921] 用于转导B721.221、KG-1和K562细胞的慢病毒在生长至80％汇合的HEK293T细胞中产生。将6微克编码HLA I类或HLA II类的基因组载体psFLenti(在前面的部分中描述)与5.3ug的慢病毒包装载体psPAX2和1.81ug的包膜载体pMD.2混合。DNA与Opti-MEM(Thermo Fisher Scientific)和转染试剂Fugene HD(Promega,Madison,WI)混合,混合物在室温下孵育15分钟。然后将混合物滴加到HEK293T细胞培养皿中并孵育72小时。然后收获上清液,并使用Lenti-X GoStix(Takara Bio Inc.,Japan)测试慢病毒滴度。对于转导,将细胞接种在12孔平底板(Corning Inc.,Corning,NY)中,并与含有6ug/ml聚凝胺(Sigma-Aldrich)的慢病毒上清液混合。将与慢病毒混合的细胞在32℃下以800xg旋转90分钟。将细胞重新悬浮在温热培养基中,并在37℃培养箱中在5％CO_2中培养72小时。然后使用1ug/ml嘌呤霉素选择细胞2周。选择后,收获至少5000万个细胞,离心,用1x Gibco DPBS(Thermo Scientific)洗涤一次,然后在液氮中快速冷冻用于质谱分析。

[0922] BirA蛋白表达和纯化

[0923] 使用编码与C端六组氨酸标签融合的大肠杆菌BirA的pET19载体。化学感受态大肠杆菌BL21(DE3)细胞(New England Biolabs)用BirA表达质粒(编码与C端六组氨酸融合的大肠杆菌BirA的pET19载体)转化,在37℃下在LB肉汤加100μg/ml氨苄青霉素中生长至OD_{600}为0.6-0.8,并冷却至30℃,之后通过添加0.4mM异丙基-β-D-硫代吡喃半乳糖苷诱导表达。大肠杆菌细胞在30℃下继续生长4小时。通过在4℃下以8000xg离心30分钟收获大肠杆菌细胞,并在-80℃下储存直至使用。将表达重组BirA的冷冻细胞沉淀物重新悬浮在含有5mM咪唑的IMAC缓冲液(50mM NaH_2PO_4 pH 8.0,300mM NaCl)中,与1mg/ml溶菌酶一起在冰上孵育20分钟,然后通过超声处理裂解。通过在4℃下以16,000xg离心30分钟去除细胞碎片和不溶性物质。随后使用ÄKTA纯色谱系统(GE Healthcare)将澄清的上清液加载到HisTrap HP 5mL柱上,用IMAC缓冲液加25mM和50mM咪唑洗涤,然后用500mM咪唑洗脱。将含有BirA的级分合并,并用25mM NaCl对20mM Tris-HCl pH 8.0进行透析,并加载到HiTrap Q HP 5mL柱(GE Healthcare,Chicago,IL)上,并通过应用25至600mM NaCl的线性梯度来洗脱。合并含有高

纯度BirA的级分,在储存缓冲液(20mM Tris-HCl pH 8.0 100mM NaCl,5%甘油)中进行缓冲液交换,并浓缩至约5-10mg/mL,等分,并在液氮中快速冷冻以在-80℃下储存。使用计算的消光系数ε=47,440$M^{-1}cm^{-1}$,通过紫外光谱法在OD_{280nm}处计算BirA蛋白浓度。

[0924] Western印迹方案

[0925] 将样品加入XT样品缓冲液和XT还原剂(Bio-Rad,Hercules,CA),在95℃下加热5分钟,然后将对应于约100,000个细胞的体积加入10%Criterion XT Bis-Tris凝胶(Bio-Rad)中,并使用PowerPac Basic Power Supply(Bio-Rad,Hercules,CA)和XT MES电泳缓冲液(Bio-Rad,Hercules,CA)在200V下电泳35分钟。用水简单冲洗凝胶,然后在iBlot2 Gel Transfer Device(Thermo Scientific)上使用设置P3将蛋白质转移到Invitrogen iBlot Transfer Stacks(Thermo Fisher Scientific)内的PVDF膜上。Precision Plus Protein All Blue Standard(Bio-Rad,Hercules,CA)用于监测分子量。接下来,将膜用TBS Tween 20(TBST)缓冲液(25mM Tris,0.15mM NaCl,0.05%(v/v)Tween 20,pH 7.5)洗涤3x5分钟,在室温下在TBST-M(含有5%(w/v)脱脂速溶奶粉的TBST)中封闭1小时,然后在TBST-B(含有5%(w/v)牛血清白蛋白(Sigma Aldrich)的TBST]和1:5,000稀释的兔抗β微管蛋白抗体(产品目录#ab6046,Abcam,Cambridge,MA)和兔抗生物素连接酶表位标签抗体(产品目录#100-401-B21,Rockland Immunochemicals,Limerick,PA)中孵育过夜。接下来,将膜用TBST洗涤3x5分钟,在含有1:10,000稀释的山羊抗兔IgG(H+L-辣根过氧化物酶偶联抗体(目录号170-6515,Bio-Rad)的TBST-M中室温孵育1小时,然后在室温下用TBST洗涤3x5分钟。最后,将膜用Pierce ECL Western Blotting Substrate(Thermo Fisher Scientific)冲洗,使用ChemiDoc XRS+Imager(Bio-Rad)显色,并使用Image Lab软件(Bio-Rad)进行可视化。

[0926] 亲和标记的HLA-肽复合物的分离

[0927] 从表达BAP标记的HLA等位基因的细胞和仅表达无BAP标签的内源性HLA-肽复合物的阴性对照细胞系进行亲和标记的HLA-肽复合物的分离。在用于HLA-肽亲和纯化之前,将NeutrAvidin珠状琼脂糖树脂用1mL冷PBS洗涤3次。将含有$50x10^6$个表达BAP标记的HLA肽的细胞的冷冻沉淀物在冰上解冻20分钟,然后通过在1.2mL冷裂解缓冲液[20mM Tris-Cl pH 8,100mM NaCl,6mM $MgCl_2$,1.5%(v/v)Triton X-100,60mM辛基葡糖苷、0.2mM 2-碘乙酰胺、1mM EDTA pH 8,1mM PMSF,1X完全不含EDTA的蛋白酶抑制剂混合物(Roche,Basel,Switzerland)]中手工移液轻轻裂解。将裂解物与>250单位的Benzonase核酸酶(Sigma-Aldrich)在4℃下翻滚(end/over/end)孵育15分钟,以降解DNA/RNA,并在4℃下以15,000xg离心20分钟以去除细胞碎片和不溶性物质。将澄清的上清液转移到新管中,并通过在室温下在1.5mL管中与0.56μM生物素、1mM ATP和3μM BirA翻滚孵育10分钟来生物素化BAP标记的HLA肽。将上清液在4℃下以相当于200μL的Pierce高容量NeutrAvidin珠状琼脂糖树脂(Thermo Scientific)浆液的体积在4℃下翻滚孵育30分钟,以亲和富集生物素化的HLA-肽复合物。最后,将HLA结合的树脂用1mL冷洗涤缓冲液(20mM Tris-Cl pH 8,100mM NaCl,60mM辛基葡糖苷、0.2mM 2-碘乙酰胺、1mM EDTA pH8)洗涤四次,然后用1mL的冷10mM Tris-Cl pH8洗涤四次。在洗涤之间,手工轻轻混合HLA结合的树脂,然后通过在4℃下以1,500xg离心一分钟来沉淀。将洗涤过的HLA结合树脂储存在-80℃或立即进行HLA肽洗脱和脱盐。

[0928] 基于抗体的HLA肽复合物分离

[0929] HLA II类DR-肽复合物从健康供体外周血单核细胞(PBMC)中分离出来。将相当于

75μL GammaBind Plus Sepharose树脂的体积用1mL冷PBS洗涤3次,与10μg抗体在4℃下孵育过夜,然后用1mL冷PBS洗涤3次,然后用于HLA肽免疫沉淀。将含有50x10^6个细胞的冷冻PBMC沉淀物在冰上解冻20分钟,然后通过在1.2mL冷裂解缓冲液[20mM Tris-Cl pH 8,100mM NaCl,6mM MgCl2,1.5%(v/v)Triton X-100,60mM辛基葡糖苷,0.2mM 2-碘乙酰胺,1mM EDTA pH8,1mM PMSF,1X完全不含EDTA的蛋白酶抑制剂混合物(Roche)]中轻轻裂解。将裂解物在4℃下与>250单位的Benzonase核酸酶(Sigma-Aldrich)一起孵育15分钟以降解DNA/RNA,并在4℃下以15,000xg离心20分钟以去除细胞碎片和不溶性物质。然后将上清液与结合至GammaBind Plus Sepharose树脂(GE Life Sciences)的抗HLA DR抗体(TAL 1B5,产品号sc-53319;Santa Cruz Biotechnology,Dallas,TX)在4℃下翻滚孵育3小时,以免疫沉淀HLA DR-肽复合物。最后,将HLA结合的树脂用1mL冷洗涤缓冲液(20mM Tris-Cl pH 8,100mM NaCl,60mM辛基葡糖苷、0.2mM 2-碘乙酰胺、1mM EDTA pH8)洗涤四次,然后用1mL冷的10mM Tris-ClpH8洗涤四次。在洗涤之间,轻轻混合HLA结合的树脂,然后通过在4℃下以1,500xg离心1分钟沉淀。将洗涤过的HLA结合的树脂储存在-80℃或立即进行HLA-肽洗脱和脱盐。

[0930] HLA-肽洗脱和脱盐

[0931] 将HLA肽从亲和标记的内源性HLA复合物中洗脱,同时使用Sep-Pak(Waters)固相萃取系统脱盐。简而言之,将Sep-Pak Vac 1cc(50mg)37-55μm粒径tC18小柱连接到24位提取歧管(Restek),用200μL MeOH活化两次,然后用100μL 50%(v/v)ACN/1%(v/v)FA,然后用500μL 1%(v/v)FA洗涤四次。为了将HLA肽与亲和标记的HLA肽解离并促进肽与tC18固相的结合,将400μL的3%(v/v)ACN/5%(v/v)FA添加到含有HLA-结合的珠状琼脂糖树脂的管中。通过移液混合浆液,然后转移到Sep-Pak柱中。用1%(v/v)FA(2x200μL)冲洗试管和移液器吸头,并将冲洗液转移到柱中。将100fmol的Pierce Peptide Retention Time Calibration(PRTC)混合物(Thermo Scientific)添加到柱中作为上样对照。将珠状琼脂糖树脂与200μL 10%(v/v)AcOH孵育两次,每次五分钟,以进一步将HLA肽与亲和标记的HLA肽解离,然后用500μL 1%(v/v))FA洗涤四次。HLA-肽从tC18洗脱到新的1.5mL微管(Sarstedt)中,方法是用250μL的15%(v/v)ACN/1%(v/v)FA,然后用2x250μL的30%(v/v)乙腈/1%(v/v)FA进行分步分级分离。用于活化、上样、洗涤和洗脱的溶液通过重力流动,但使用真空(≤-2.5PSI)从柱中去除剩余的洗脱液。将含有HLA-肽的洗脱液冷冻,通过真空离心干燥,并在进行第二次脱盐工作流程之前储存在-80℃。如前所述,使用两个16号Empore C18固相萃取盘(3M,St.Paul,MN)的冲头包装的内部构建的StageTips对HLA-肽样品进行二次脱盐。StageTips用100μL MeOH活化两次,然后用50μL 50%(v/v)ACN/0.1%(v/v)FA活化,然后用100μL 1%(v/v)FA洗涤3次。通过添加200μL 3%(v/v)ACN/5%(v/v)溶解干燥的HLA肽,然后加载到StageTips上。用1%(v/v)FA(2x100μL)冲洗试管和移液器吸头,并将冲洗液转移到StageTips,然后用100μL 1%(v/v)FA冲洗StageTips五次。使用20μL 15%(v/v)ACN/0.1%(v/v)FA,然后两个20μL段30%(v/v)ACN/0.1%(v/v)FA的阶梯梯度洗脱肽。上样、洗涤和洗脱在台式离心机上进行,最大速度为1,500-3,000xg。将洗脱液冷冻,通过真空离心干燥,在-80℃下保存。

[0932] 通过串联质谱法进行的HLA-肽测序

[0933] 所有nanoLC-ESI-MS/MS分析都采用如下所述的相同LC分离条件。使用配备

PicoFrit(New Objective,Inc.,Woburn,MA)75μm内径毛细管的Proxeon Easy NanoLC 1200(Thermo Scientific,San Jose,CA)层析分离样品,在采用He的1000psi压力下包装10-μm发射器,至30-40cm,采用1.9μm粒径/200 Å孔径的C18 Reprosil珠子(Dr.Maisch GmbH,Ammerbuch,Germany),并在分离过程中在60℃下加热。柱子用10倍床体积的缓冲液A(0.1%(v/v)FA和3%(v/v)ACN)平衡,将样品加至4μL 3%(v/v)ACN/5%(v/v)FA,将肽用82分钟内7-30%缓冲液B(0.1%(v/v)FA和80%(v/v)ACN),6分钟内30-90%缓冲液B的线性梯度洗脱,然后在90%的缓冲液B中保持15分钟,以洗涤柱子。一部分样品用84分钟内6-40%缓冲液B、9分钟内40-60%缓冲液B的线性梯度洗涤,然后在90%缓冲液B中保持5分钟,在50%缓冲液B中保持9分钟,以洗涤柱子。用于样品洗脱的线性梯度以250nL/min的速率运行,产生约13秒的中值峰宽。

[0934] 在依赖于数据的采集过程中,将洗脱的肽引入配备有Nanospray Flex离子源(Thermo Scientific)的Orbitrap Fusion Lumos质谱仪(Thermo Scientific),电压为2.2-2.5kV。完全扫描MS以60,000的分辨率采集,从300到1,700m/z(AGC目标4e5,最大IT为50ms)。每次完全扫描之后是2秒的循环时间,或前10次,分辨率为15,000的数据依赖性MS2扫描,使用隔离宽度为1.0m/z,碰撞能量为34(HLA I类数据)和38(HLA II类数据),ACG目标为5e4,最大填充时间为250ms最大离子时间。使用1.0m/z的分离宽度是因为HLA II类肽往往更长(中值为16个氨基酸,部分肽>40个氨基酸),因此单同位素峰并不总是同位素簇中的最高峰,并且在没有指定偏移量的情况下,质谱仪采集软件将最高的同位素峰放置在隔离窗口的中心。因此,1.0m/z隔离窗口将允许共同隔离单同位素峰,即使它不是同位素簇中的最高峰,因为II类肽的电荷态通常为+2或更高。启用动态排除,重复计数为1,排除持续时间为5秒,以启用每个选定的前体约3个PSM。由于HLA-肽鉴定依赖于PSM质量,因此同位素被排除在每个前体的单个电荷态的依赖扫描被禁用时,因此不同电荷态的多个PSM进一步增加了我们对肽鉴定的信心。HLA II类数据收集的电荷态筛选与使用肽模式的单同位素前体选择(MIPS)一起启用,以防止对电荷态为1(仅适用于具有碱性锚残基的等位基因)、>7或未分配的前体离子触发MS/MS。对于HLA I类数据收集,选择了电荷态为1(质量范围800-1700m/z)和2-4的前体离子,而排除了>4和未分配的电荷态。

[0935] LC-MS/MS数据的解释,与图29相关。

[0936] 使用Spectrum Mill软件包v6.0预发布版(Agilent Technologies,Santa Clara,CA)解释质谱。如果MS/MS谱没有600-2000(I类)/600-4000(II类)范围内的前体MH+,前体电荷>5(I类)/>7(II类),或至少检测到<5个峰,则从搜索中排除该MS/MS谱。禁止将在同一色谱峰中获得的具有相同前体m/z的相似光谱合并。针对包含所有UCSC Genome Browser基因和基因组hg19注释的数据库及其蛋白质编码转录物(63,691个条目;10,917,867个独特的9聚体肽)以及264种常见污染物的数据库搜索MS/MS谱。在数据库搜索之前,所有MS/MS都必须通过序列标签长度>2的光谱质量过滤器,例如,最少3个质量被氨基酸的链内质量分开。设置了为5的最小骨架切割评分(BCS),并使用了ESI QExactive HLAv2评分方案。使用无酶特异性、将半胱氨酸修饰固定为半胱氨酸化以及以下可变修饰来搜索来自未还原和烷基化的天然HLA肽样品的所有光谱:氧化甲硫氨酸(m)、焦谷氨酸(N-term q)、氨基甲酰甲基化(c)。使用无酶特异性、将半胱氨酸修饰固定为氨基甲酰甲基化以及以下可变修饰来搜索还原和烷基化的HLA肽样品:氧化甲硫氨酸(m)、焦谷氨酸(N-term q)、半胱氨酸化(c)。±

10ppm的前体质量容差、±10ppm的产物质量容差和30％的最小匹配峰强度用于天然、还原和烷基化HLA-肽数据集。使用Spectrum Mill自动验证模块将单个光谱的肽谱匹配(PSM)自动指定为可信分配,以在PSM排名上应用基于靶标-诱饵的FDR估计,来设置评分阈值标准。自动阈值策略使用最小序列长度为7、自动可变范围前体质量过滤以及在所有LC-MS/MS运行中针对HLA等位基因优化的评分和delta排名1-排名2评分阈值,对于每个前体电荷态产生<1.0％的PSM FDR估计。

[0937] 通过去除参考数据库中分配给264种常见污染物蛋白质的所有肽以及去除在阴性对照MAPTAC™亲和下拉中鉴定的肽,对通过PSM FDR估计值<1.0％的鉴定肽进一步过滤污染物。此外,映射到参考数据库的计算机胰蛋白酶消化物的所有肽鉴定都被删除,因为不能排除这些肽是uPLC柱上样品残留的胰蛋白酶污染物。

[0938] 为了从SILAC DC饲喂实验中去除潜在的假阳性PSM鉴定,将额外的质量过滤器应用于使用上述方法鉴定的PSM。使用以下阈值过滤所有FDR<1％的肽以获得高质量PSM:i)得分峰值强度>60％,ii)骨架切割得分≥8,和iii)ppm质量耐受性与在同一LC-MS/MS重复中的所有PSM鉴定中观察到的中值ppm相差±1ppm。

[0939] 使用MAPTAC™方案对HLA-DR、-DQ、-DP异二聚体的单等位基因分配

[0940] 由于在MAPTAC™方案中仅标记了HLA II类的β链,因此下拉步骤可分离肽-MHC复合物,无论它们是否包含敲入或内源性α链。在HLA-DR的情况下,α链中的等位基因变异不被认为会影响肽结合;因此,与内源性α链配对的相对程度与数据解释无关——数据实际上是单等位基因的。然而,对于HLA-DP和HLA-DQ基因座,α链表现出重要的等位基因变异,因此敲入和内源性α链等位基因的存在创造了1-3种不同特异性的潜力(取决于细胞系是否具有一个或两个α链等位基因以及是否与敲入等位基因匹配)。原则上,这个问题可以通过运行使用和不使用敲入α链的方案并鉴定采用α的实验特定的肽组来缓解。本文采用使用表达与敲入α等位基因匹配的单个α等位基因的细胞系的方法。

[0941] 先前公布的MS数据的分析,与图12A-12F、图30A-30C、图31A-31D、图21A、图3139A-39B和图40A-B相关。

[0942] 已公布的提供.raw文件的LC-MS/MS数据集使用Spectrum Mill软件包v6.0预发布版(Agilent Technologies,Santa Clara,CA)重新处理。在Thermo Orbitrap仪器上收集的在轨道阱(高分辨率)中利用HCD碎裂以及MS和MS/MS数据收集的数据集(例如Velos、QExactive、Fusion、Lumos)使用上述“LC-MS/MS数据的解释”部分中描述的参数进行分析。对于利用CID碎裂的MS和MS/MS高分辨率数据集,将与上述相同的参数与ESI Orbitrap评分方案一起使用。对于在轨道阱中收集MS数据和在离子阱中收集MS/MS数据的数据集,也使用了以下相同参数,但存在以下偏差。对于HCD数据,使用ESI QExactive HLAv2评分方案,而ESI Orbitrap评分方案用于CID数据。使用±10ppm的前体质量容差和±0.5Da的产物质量容差。对于高分辨率和低分辨率MS/MS数据集,使用Spectrum Mill自动验证模块将单个光谱的肽谱匹配(PSM)自动指定为可信分配,以在PSM排名中应用基于靶标-诱饵的FDR估计来设置评分阈值标准。自动阈值策略使用最小序列长度为7、自动可变范围前体质量过滤以及在所有LC-MS/MS运行中针对HLA等位基因优化的评分和Delta排名1-排名2评分阈值,对于每个前体电荷态产生<1.0％的PSM FDR估计。对一些先前公布的数据的肽鉴定分析显示9聚体(>10％)的比率很高。由于这些可能代表污染的HLA I类配体,因此从所有外部数据集中

删除了短肽(长度<12)。

[0943] 将肽映射到基因和“嵌套组”,与图30A-30C、图31A-31D和图32A-32E相关

[0944] 将每个肽分配给UCSC hg19基因注释(genome.ucsc.edu/cgi-bin/hgTables)中的一个或多个蛋白质编码转录物。由于许多肽鉴定与其他鉴定重叠,并因此构成大部分冗余信息,因此将肽分组为“嵌套组”,每个意味着对应于约1个独特的结合事件。例如,肽GKAPILIATDVASRGLDV、GKAPILIATDVASRGLD和KAPILIATDVASRGLDV都包含保守序列KAPILIATDVASRGLD,并且可能都在同一寄存器中结合MHC。为了嵌套给定数据集的肽,构建了一个图,其中每个节点对应于一个独特的肽,并且在任何一对共享至少一个9聚体并可映射到至少一个共同转录物的肽之间创建一条边。R程序包igraph(Team,2014)(cran.r-project.org/web/packages/igraph/citation.html)中的簇命令用于鉴定连接节点的簇,每个簇被定义为一个嵌套组。此过程可确保将满足边缘标准(≥1个常见9聚体和≥1个常见转录物)的任何两个肽放置在同一嵌套组中。嵌套用于序列标识生成(使用每个嵌套组中最短的肽生成标识;图30A-30C,机器学习(嵌套组中肽之间的重要性权重总和为1;图31A-31D),以及基因偏好分析(每个嵌套组被算作一次观察而不是每个单独的肽;图32A-32E)。

[0945] 氨基酸频率分析,与图12F相关

[0946] 人类蛋白质组中的氨基酸频率是基于UCSC hg19注释中所有蛋白质编码基因的序列计算的(为由多个转录物同种型表示的基因随机选择一个转录物),如图11D所示。IEDB频率通过鉴定一组独特的肽来确定,这些肽具有至少一个≤50nM的亲和力观察结果(不包括一些在其C端具有六价多组氨酸的肽)。MAPTAC™频率首先在跨五个DRB1等位基因(DRB1*01:01、DRB1*03:01、DRB1*09:01和DRB1*11:01)的标准前相方案的背景中考虑,每个嵌套组仅使用一个肽(最长的)。此外,MAPTAC™频率针对多个等位基因的基本反相方案单独计算。不考虑潜在的起源等位基因,同样使用每个嵌套组的最长肽,分析来自外部数据集的MS数据。

[0947] 构建HLA I类序列标识,与图37B相关

[0948] 对于每个HLA I类等位基因,通过分析相应肽的前五个位置(映射到标识位置1-5)和最后四个位置(映射到标识位置6-9)的氨基酸频率来创建长度为9的序列标识。以这种方式,无论长度如何,肽都有助于序列标识。与II类标识一样,字母高度与每个位置的每个氨基酸的频率成正比,对频率≥10%的氨基酸使用颜色编码。

[0949] 评估HLA II类肽去卷积的性能,与图30B相关

[0950] 为了评估GibbsCluster(v2.0)工具按来源等位基因对多等位基因HLA II类肽数据进行聚类的能力,分析了其在8个样品上的性能,包括4个PBMC样品、1个黑素瘤细胞系(A375)和3个先前发表的淋巴母细胞系。对于每个样品基因型中存在的每个DRB1/3/4/5等位基因,从我们的单等位基因MAPTAC™数据中掺加了20种肽。掺加的肽被限制为12-20聚体,SPI≥70,与其他HLA-DR等位基因的MAPTAC™数据中的任何肽或目标等位基因的任何掺加肽不共享9聚体。然后将这些增强的数据集提交给GibbsCluster-v2.0,使用默认的HLA类II设置,除了在位置1强制执行疏水性偏好之外,与以前用于去卷积的相同。对于每个样品,手动指定溶液中的簇数,并设置为等于基因型中存在的HLA-DR等位基因的数目。

[0951] 计算具有优选锚残基的肽的分数,与图30C相关

[0952] 锚位置被定义为具有最低熵的四个位置,在这些位置中,“优选的”氨基酸包括所有频率≥10%的氨基酸。在计算在n个位置具有优选氨基酸的肽的分数时,每个嵌套组(最

短的)仅使用一个肽。

[0953] MS观察到的肽的预测亲和力,与图36A相关

[0954] 对于每个HLA II类等位基因,使用NetMHCIIpan-v3.1鉴定所有长度为14到17的独特肽并对其结合潜力进行评分。为了进行比较,从人类蛋白质组中采取了50,000个随机长度匹配的肽。基于对数转换值确定密度分布。

[0955] MS观察到的肽测得的亲和力,与图36B相关

[0956] 如果肽具有较差的预测NetMHCIIpan结合亲和力(对于DRB1*01:01为>100nM或对于DRB1*09:01和DRB1*11:01为>500nM),或者如果它们表现出≤2个启发式定义的锚,则选择肽用于亲和力测量。交叉验证分区的建立,与图31A相关

[0957] 创建了一个图,其中每个节点代表一个蛋白质编码转录物,并且在所有共享至少5个独特9聚体的氨基酸序列内容物的转录物对之间存在边(UCSC hg19基因注释)。R程序包igraph(Team,2014)(cran.r-project.org/web/packages/igraph/citation.html)中的簇命令用于鉴定连接节点的簇,每个簇被定义为一个“转录组”。以这种方式,如果两个转录物共享一条边(>5个共享的9聚体),则保证它们被放置在同一个转录物组中。对转录物组随机采样,将蛋白质组分成八个大致相同大小的分区。MS观察到的肽(和未观察到的诱饵肽)根据其源转录物的分区放置在分区中,这些分区用于交叉验证和超参数调整。使用基于图的蛋白质组分区方法来最大限度地减少训练和评估过程中出现类似肽序列的可能性,这可能会人为地提高预测性能。

[0958] 基于CNN的HLA II类结合预测器neonmhc2的架构和训练,与图31A相关的

[0959] 通过随机改组命中肽的序列来生成用于训练的阴性样品(诱饵)。选择这种诱饵生成方法,而不是从蛋白质组中选择未观察到的区域,以消除可能导致一般氨基酸偏好的MS偏差。以这种方式,例如,我们的结合预测器不知道半胱氨酸的相对消耗(图12F)。类似地,这可以防止我们的模型学习与肽的全局特性(例如整体疏水性)相关的MS偏差。该方法与图31A中描绘的结果有关。

[0960] 针对两种应用场景训练模型:验证内部MAPTAC™数据(图31B)和验证外部数据(图31C、图21A和图21B)。在为前者训练模型时,采用了一个简单的训练过程,使用六个数据分区(训练分区)学习网络权重优化,使用第七个分区(调优分区)执行超参数优化和提前停止,以及模型设计完成后,对第八个分区(评估分区)进行最终验证。在外部验证的情况下,采用交叉验证,为每个数据分区构建模型集合,其中保留该分区用于超参数调整和提前停止,并使用剩余的七个分区进行网络权重优化。此外,当对非MS数据进行评分时(图31C和图31D),对目标肽的每个12-20聚体子串进行评分并保持最高评分。

[0961] 在训练我们的模型时,每个命中和诱饵在损失函数中根据其源嵌套组的大小进行加权,使得每个嵌套组作为一个整体具有相同的权重。在评估超参数调整模型时,每个嵌套组中最短的肽在相关分区中用作阳性示例,并将这些命中的乱序形式用作诱饵。此外,应用了整体加权因子,使得命中的总和等于训练时诱饵的总权重。对于模型的最终评估,如图31B所示,再次从评估分区(第8部分)中的每个嵌套组中选择最短的肽,但从未观察到的肽源基因子序列(“自然诱饵”,在后续部分中描述)中随机采样诱饵。通过这种方式,模型为了简单地将自然序列与乱序序列区分开来而学习的任何偏差都不会提高我们在评估分区上的表现。

[0962] 模型使用Adam优化器进行训练,初始学习率为0.003,beta_1值为0.9,beta_2值为0.999,没有衰减(默认Keras参数,学习率除外),并使用二元交叉熵损失函数。初始模型权重是使用He初始化设置的。每5轮训练后,测量调整分区上的阳性预测值(PPV,在后续部分中描述)并跟踪最大值。每个epoch之后,如果训练损失没有减少,则将学习率乘以1/3。类似地,每次在调整分区上测量PPV时,如果与运行最大值相比它没有增加,则学习率将乘以1/3。实施了提前停止方案,如果连续三个时期的训练损失未能减少,或者连续3次检查时调PPV未能增加到运行最大值以上,则停止训练。在训练模型时,在训练集中使用了1:39的固定命中:诱饵比,在调整分区中使用了1:19。

[0963] 特征化:虽然氨基酸可以由“独热(one-hot)”编码表示,但其他人选择使用PMBEC矩阵和BLOSUM矩阵(Henikoff和Henikoff,1992)来编码氨基酸,其中相似的氨基酸具有相似的特征概况。出于我们肽特征化的目的,在解析的蛋白质结构中生成了基于氨基酸邻近的独特矩阵。这种方法的概念是氨基酸的典型邻居应该反映其化学性质。对于约100,000个DSSP蛋白质结构(cdn.rcsb.org/etl/kabschSander/ss.txt.gz)中的每个氨基酸,确定在3D空间中最接近但在一级序列中至少相距10个氨基酸的残基。使用这些数据,确定丙氨酸的最近邻是丙氨酸的次数,丙氨酸的最近邻是半胱氨酸的次数等,以创建一个20x20的邻近计数矩阵。该矩阵的每个元素除以其相应的列和行总和的乘积,然后对整个矩阵进行对数变换。最后,从每个元素中减去整个矩阵的平均值。

[0964] 每个氨基酸还编码有11个描述氨基酸性质的二元特征,例如它是否是:酸性(N,Q)、脂肪族(I、L、V)、芳香族(H、F、W、Y)、碱性(H、K、R)、带电荷(D、E、H、K、R)、疏水性(A、C、F、H、I、K、L、M、T、V、W、Y)、羟基(S、T)、极性(C、S、N、Q、T、D、E、H、K、R,Y、W)、小(V、P、A、G、C、S,T、N、D)、非常小(A、G、C、S)或含有硫(M、C)。两个特征用于描述每个氨基酸的位置,一个在整个肽中单调增加,一个表示距肽中心的绝对距离,均以位置单位(而非物理距离)为单位。最后,包括一个单一的二元特征来指示氨基酸是否从该位置“缺失”,这会发生在较短肽的边缘之外。结果是每个氨基酸由20个氨基酸邻近特征、11个氨基酸特性特征、2个位置特征和1个缺失字符特征编码,总共34个特征。所有肽都被编码为20聚体,其中中央20个氨基酸用于较长的肽,缺失的字符值对称添加到短于20个氨基酸的肽的边缘。

[0965] 当示例被输入到神经网络中用于训练和评估时,34个特征中的每一个都通过减去它们的平均值并除以它们的标准差来归一化。平均值和标准偏差仅根据训练集计算,不考虑肽内的位置。

[0966] 对于每个等位基因,训练一组卷积神经网络以预测结合。模型架构的草图如图31A所示,图31A描绘了两个卷积层,每个卷积层的内核大小为6和50个过滤器。在每一层之后,应用全局最大和平均池化,并将结果值输入到具有sigmoid激活的最终输出神经元中。暗示但未显示在每个卷积层之后立即应用ReLU激活、批量归一化(Ioffe和Szegedy,2015)和20%空间丢弃。

[0967] 当为每个等位基因训练一组模型时,架构是固定的,但L2正则化的数目是变化的。第一个卷积层使用0.05的基本L2正则化权重,第二个卷积层使用0.1。为了改变L2正则化的数目,这些值乘以0.1、0.5和1。对于集成中的每次迭代,每个正则化级别都训练一个模型,并根据调优分区的性能保持最佳。

[0968] 对MAPTAC™观察到的肽的基准化预测性能,与图7A相关

[0969] 在对由HLA等位基因编码的给定肽或蛋白质的预测性能值的一些示例性评估中，可以使用包括“乱序诱饵”的方法。乱序诱饵是具有与基于例如质谱数据已知结合给定HLA肽或蛋白质的肽相同的肽长度和氨基酸的肽，但是氨基酸的序列是杂乱的。对于通过质谱法鉴定的每个单个肽，使用了19个这样的乱序肽诱饵(命中:诱饵是1:19)，如图7A所示。对呈递预测模型进行了测试，并通过分析测试分区中得分最高的5％的肽并询问其中哪些部分为阳性来确定PPV。由此产生的PPV示于图7A和下面的表12中。

[0970] 表12

[0971]

MHC II 类等位基因	对于 *neonmhc2* 的 PPV
DPB1-0101_DPA1-0103	0.48
DPB1-0101_DPA1-0201	0.41
DPB1-0101_DPA1-0202	0.56
DPB1-0201_DPA1-0103	0.48
DPB1-0202_DPA1-0103	0.27
DPB1-0301_DPA1-0103	0.36
DPB1-0401_DPA1-0103	0.52
DPB1-0402_DPA1-0103	0.54
DPB1-0501_DPA1-0201	0.59
DPB1-0501_DPA1-0202	0.46
DPB1-0601_DPA1-0103	0.43
DPB1-0901_DPA1-0201	0.53
DPB1-1001_DPA1-0201	0.49
DPB1-1101_DPA1-0201	0.31
DPB1-1301_DPA1-0201	0.58
DPB1-1401_DPA1-0201	0.49
DPB1-1701_DPA1-0201	0.39
DQB1-0201_DQA1-0201	0.49
DQB1-0201_DQA1-0501	0.08
DQB1-0202_DQA1-0201	0.3
DQB1-0301_DQA1-0501	0.27
DQB1-0301_DQA1-0505	0.13
DQB1-0301_DQA1-0601	0.28
DQB1-0302_DQA1-0301	0.1
DQB1-0303_DQA1-0201	0.46
DQB1-0303_DQA1-0301	0.39
DQB1-0401_DQA1-0301	0.47

[0972]

DQB1-0402_DQA1-0401	0.53
DQB1-0501_DQA1-0101	0.62
DQB1-0502_DQA1-0101	0.26
DQB1-0502_DQA1-0102	0.19
DQB1-0601_DQA1-0102	0.17
DQB1-0601_DQA1-0103	0.35
DQB1-0602_DQA1-0102	0.42
DQB1-0603_DQA1-0103	0.56
DQB1-0604_DQA1-0102	0.28
DRB1_0101	0.48
DRB1_0102	0.49
DRB1_0301	0.47
DRB1_0302	0.29
DRB1_0401	0.38
DRB1_0402	0.49
DRB1_0403	0.43
DRB1_0404	0.42
DRB1_0405	0.41
DRB1_0407	0.36
DRB1_0410	0.61
DRB1_0701	0.49
DRB1_0801	0.36
DRB1_0802	0.46
DRB1_0803	0.30
DRB1_0804	0.36
DRB1_0901	0.35
DRB1_1001	0.54
DRB1_1101	0.43
DRB1_1102	0.42

[0973]

DRB1_1104	0.51
DRB1_1201	0.47
DRB1_1202	0.55
DRB1_1301	0.42
DRB1_1302	0.23
DRB1_1303	0.56
DRB1_1401	0.56
DRB1_1501	0.53
DRB1_1502	0.43
DRB1_1503	0.41
DRB1_1601	0.61
DRB3_0101	0.57
DRB3_0201	0.6
DRB3_0202	0.48
DRB3_0301	0.53
DRB4_0103	0.67
DRB5_0101	0.54
DRB5_0102	0.65
DRB5_0202	0.63

[0974] 对MAPTAC™观察到的肽的基准化预测性能,与图31B相关

[0975] 为了评估给定等位基因的预测性能,有必要定义一组可能已经观察到(因为它们存在于蛋白质组中)但未在MS数据中观察到的肽。这些阴性示例被称为“自然诱饵”(与上述“乱序诱饵”相反)。作为指导原则,决定:自然诱饵的长度分布应与MS观察到的命中的长度分布相匹配,自然诱饵不应包含与其他自然诱饵重复的序列,自然诱饵不应与命中重叠,和/或自然诱饵应该来自至少产生一次命中的基因。

[0976] 以下伪代码表示为创建满足这些原则的评估而实施的过程:

[0977] 初始化两个空的命中列表,$H_{最小}$和$H_{穷举}$

[0978] 对于每个嵌套组S的MS观察到的肽:

[0979] 如果S中没有肽能够映射到训练或调节分区中的转录物:

[0980] 将S中最短的肽添加到$H_{最小}$

[0981] 将S中的所有肽添加到$H_{穷举}$

[0982] 初始化一个空的诱饵肽列表,D

[0983] 对于测试分区中的每个蛋白质编码转录物(先最长,最后最短):

[0984] 如果$H_{穷举}$中没有肽映射到转录物:

[0985] 跳到下一个转录物

[0986] 用一组重叠的肽P覆盖转录物的蛋白质序列,其中肽长度从$H_{最小}$的长度分布中随机采样。重叠是8个氨基酸。(P中的最后一个肽通常会悬挂在蛋白质的末端上。)

[0987] 在P中的最后一个肽仍然悬挂时:

[0988] 从P中最长肽的长度减去1个氨基酸

[0989] 对于P中的每个肽:

[0990] 如果它不与$H_{穷举}$中的肽共享9聚体,并且它确实包含在D中的任何肽中均未观察到的9聚体:

[0991] 将该肽添加到D

[0992] 否则:

[0993] 拒绝该肽

[0994] $H_{最小}$和D构成评价数据集

[0995] 为了评估该组的性能,所有n个命中肽都由预测器(neonmhc2或NetMHCIIpan)进行评估,并与一组19n个诱饵(从完整的诱饵组中随机采样而不替换)一起评分。组合的组中前5%的肽被标记为阳性判定,阳性预测值(PPV)被计算为作为命中的阳性判定的比例。请注意,由于阳性数被限制为等于命中数,因此在此评估场景中判定率等于PPV。在等位基因之间应用一致的1:19比率有助于稳定性能值,否则性能值会受到每个等位基因观察到的命中数的高度影响。这被认为是合适的,因为假设与等位基因的内在特性相比,命中数更多地与实验条件和重复计数有关。

[0996] 计算非15聚体的NetMHCIIpan亲和力,与图31A-D、图40A-B和33A-D相关

[0997] 在早期分析中,NetMHCIIpan-v3.1亲和力和非15聚体百分排名预测在基准上表现不佳。然而,以下方法显著提高了性能:如果肽长度超过15个氨基酸,则所有成分都被评分为15聚体,并选择最强的预测作为总体肽评分;如果肽短于15个氨基酸,则在N端填充G以强制肽长度为15,并对得到的延伸肽进行评分。

[0998] 作为训练集大小的函数的性能,与图38A相关

[0999] 为了了解我们模型的性能如何受到数据集大小的限制,我们进行了饱和度分析。这涉及重新训练模型集合,同时改变所用训练数据的比例,以了解这如何影响保留分区的性能。图38A显示了作为训练集中使用的命中肽数目的函数的评估分区(分区8)PPV。每个数据点显示10个模型集合的平均PPV,误差条表示标准偏差。

[1000] 天然CD4+ T细胞应答的基准化预测性能,与图31C相关

[1001] 由于IEDB中记录的绝大多数CD4+ T细胞应答具有未知或计算推算的II类等位基因限制,因此记录的子集集中在由II类四聚体通过实验证实的那些。几乎所有此类记录都由William Kwok Laboratory(Benaroya Research Institute,Seattle,WA)保存,该实验室使用免疫应答性个体的血液对不同病原体和变应原进行四聚体引导的表位定位(TGEM)。由于某些研究发布了阴性肽而其他研究没有,因此审查了来源出版物以重建完整的阳性和阴性肽反应性集合。在某些情况下,来源出版物明确列出了阴性肽。在其他情况下,通过遵

循出版物方法中指定的平铺程序并确认肽边界与已知阳性示例一致来估算阴性值。在图31C所示的该测定中,从流感病毒和鼻病毒的病毒基因绘制病毒表位,并使用包含表位的肽序列来预测每个表位的HLA II类蛋白结合物。在这种情况下,针对相应的HLA-DRB1蛋白预测了CD4+记忆T细胞对肽的反应的PPV。PPV是通过询问阳性结合物中排名靠前的表位的哪一部分是真正命中来确定的。鉴于HLA II类蛋白分子与肽的阳性配对,并且HLA分子存在于被相应病毒感染的受试者中,将在受试者中产生CD4应答。Neonmhc2和公开可用的预测器(NetMHCIIpan)之间的预测效率(PPV,换言之,预测真实命中数)的比较按该示例性研究中测试的每个DRB1蛋白质显示在图31C中。Neonmhc2在测试的六个等位基因中的每一个都优于NetMHCIIpan。

[1002] 所有20聚体肽都由neonmhc2和NetMHCIIpan-v3.1评分。PPV被计算为n个得分最高的肽中经实验证实为阳性的分数,其中总共有n个实验证实的肽(图31C)。

[1003] T细胞诱导方案和免疫原性读出,与图31D相关

[1004] 为了产生单核细胞衍生的树突细胞(mDC),按照制造商的方案(Miltenyi Biotec),通过磁分离使用人类CD14微珠从HLA-DRB1*11:01+健康供体外周血单核细胞(PBMC)中分离CD14+单核细胞。分离的CD14+细胞在补充有800U/ml rh GM-CSF和400U/ml rh IL-4(Cellgenix)的Cellgenix GMP DC培养基中分化5天。在第5天,收获mDC并在37摄氏度下用0.4mM肽脉冲1小时,然后使用10ng/ml TNF-α、10ng/ml IL-1β、10ng/ml IL-6(Cellgenix)和0.5ug/ml PGE1(Cayman Pharma)成熟。四十八小时后,将mDC与自体PBMC以1:10的比例在含有AIMV/RPMI(ThermoFisher)、10％人血清(Sigma-Aldrich)、1％Pen/Strep(ThermoFisher)和补充有5ng/ml的IL7和IL15(Cellgenix)的培养基中共培养。在第12天,收获T细胞并在0.4mM肽脉冲的成熟DC上重新刺激7天,进行两次额外的刺激,总共3次刺激。

[1005] 诱导的T细胞用如前所述的独特的双色条形码标记系统进行标记,并与如上所述的源自CD14+单核细胞的肽脉冲的和成熟的自体mDC以1:10的比例培养过夜。第二天早上,通过流式细胞术评估细胞响应于肽的IFN-γ的产生。细胞在37℃下用Golgi Plug/Golgi Stop(BD Biosciences)处理四小时。然后用针对CD19、CD16、CD14、CD3、CD4、CD8(BD Biosciences,SanJose,CA)的表面标记抗体以及活/死可固定死细胞染色剂(ThermoFisher)对细胞进行染色;见下表13。然后根据制造商的方案使用BD Cytofix/Cytoperm试剂盒(BD Biosciences)对样品进行透化和固定,并用细胞内抗IFN-γ抗体(BD Biosciences)染色。样品在BD Fortessa X-20流式细胞仪上运行并使用FlowJo软件(Treestar)进行分析。对肽有阳性反应的诱导样品是比无肽对照高3％诱导IFN-γ产生的样品。

[1006] 表13

[1007]

标志物	荧光团	供应商	目录号	克隆
活死染色剂	IR 染料	ThermoFisher	L34976	
CD19	BUV395	BD	563551	SJ25C1
CD16	BUV395	BD	563784	3G8
CD14	BUV395	BD	563562	MφP9
CD3	BUV805	BD	565511	SK7
CD4	AF700	BD	557922	RPA-T4
CD8	PerCP-Cy5.5	BD	565310	SK1
IFN-γ	APC	BD	554702	B27

[1008] 单细胞RNA-Seq中HLA II类表达数据的分析,与图19A相关

[1009] 单细胞RNA-Seq数据来自三个先前公布的人类肿瘤样品数据集。第一项研究包括来自皮肤黑素瘤的数据。文件“GSE72056_melanoma_single_cell_revised_v2.txt”从Gene Expression Omnibus(ncbi.nlm.nih.gov/geo/;登录:GSE72056)下载。将肿瘤状态标志为“2”的细胞视为肿瘤细胞,将肿瘤状态标志为“1”且免疫细胞类型标志等于“1”至“6”的细胞分别视为T细胞、B细胞、巨噬细胞、内皮、成纤维细胞和NK。所有其他单元都被丢弃。数据本身以log2(TPM/10+1)为单位呈现,因此在数学上转换为TPM标度。一旦在TPM标度上,将每个细胞的数据相对于蛋白质编码UCSC基因符号集重新归一化至总和为,000,000(未出现在表达矩阵中的蛋白质编码基因被隐含地视为具有零表达)。最后,将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值,以产生患者细胞类型水平上的表达估计。

[1010] 第二项研究包括来自头颈部肿瘤的数据。文件“GSE103322_HNSCC_all_data.txt”从Gene Expression Omnibus(ncbi.nlm.nih.gov/geo/;登录:GSE103322)下载。该表中的数据以log2(TPM/10+1)为单位;因此,这些值在数学上转换为TPM单位。与黑素瘤研究一样,将每个细胞的数据相对于蛋白质编码UCSC基因符号集重新归一化至总和为1,000,000,并且将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值。对应于淋巴结活检的数据被排除在外。

[1011] 第三项研究包括来自未经治疗的非小细胞肺的数据。文件“RawDataLung.table.rds”和“metadata.xlsx”从ArrayExpress(ebi.ac.uk/arrayexpress/;登录:E-MTAB-6149和E-MTAB-6653)下载。如前所述,数据(已经以TPM为单位)相对于先前描述的蛋白质编码基因集重新缩放至总和为1,000,000。最后,将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值,以产生患者细胞类型水平上的表达估计。为简单起见,将细胞类型合并为比表14中原始报告的更粗略的类别。

[1012] 表14

[1013]

粗略分配	组成细胞类型
肺泡	“肺泡”，不包括“立方形肺泡 2 型(AT2)细胞”
FO B 细胞	“滤泡 B 细胞”
浆细胞	“血浆 B 细胞”
CLEC9A+ DCs	“交叉呈递树突细胞”
单核树突细胞	“单核细胞衍生的树突细胞”
pDCs	“类浆树突细胞”
郎格汉斯	“郎格汉斯细胞”
巨噬细胞	“巨噬细胞”
粒细胞	“粒细胞”
内皮	“正常内皮细胞”、“肿瘤内皮细胞”和“较低质量的内皮细胞”，不包括“淋巴 EC”
上皮	“上以细胞”和“较低质量的上皮细胞”
成纤维细胞	“表达 COL12A1 的成纤维细胞”、“表达 COL4A2 的成纤维细胞”、“表达 GABARAP 的成纤维细胞”、“较低质量的成纤维细胞”、“正常肺成纤维细胞”、“表达 PLA2G2A 的成纤维细胞”和“表达 TFPI2 的成纤维细胞”

[1014]

T 细胞	“调节性 T 细胞”、“CD4+ T 细胞”和“CD8+ T 细胞”
NKs	“自然杀伤细胞”
肿瘤	“癌细胞”
从分析中排除	“成红血细胞”和“MALT B 细胞”

[1015] 第四项研究包括来自结直肠肿瘤的数据。文件“GSE81861_CRC_tumor_all_cells_FPKM.csv”从Gene Expression Omnibus(ncbi.nlm.nih.gov/geo/;登录:GSE81861)下载。如前所述,数据(已经以TPM为单位)相对于先前描述的蛋白质编码基因集重新缩放至总和为1,000,000。最后,将对应于相同细胞类型和相同来源活检的单细胞观察结果取平均值,以产生患者细胞类型水平上的表达估计。对于这项研究,假定标记为“上皮”的细胞代表肿瘤细胞和正常上皮的混合物。

[1016] 第五项研究包括来自浆液性卵巢癌肿瘤的数据。2名低级别浆液性卵巢癌患者(LG1、LG2)和4名高级别浆液性卵巢癌患者(HG1、HG2F、HG3、HG4)的6例卵巢上皮癌的单细胞RNA测序数据来自别处。质量过滤、聚类和分析遵循Shih等人,2018概述的步骤。简而言之,Seurat分析工具用于对通过质量过滤的细胞进行聚类(最少200个表达的基因,其中每个基

因必须在至少3个不同的细胞中检测到;总共2258个细胞)。细胞周期和独特的转录物计数的影响被回归。在主成分分析后对细胞进行聚类,并根据原始出版物中基因特征的表达将聚类分配给细胞类型。HLA-DRB1基因的TPM根据每个患者的每种细胞类型的蛋白质编码基因的标准化唯一转录物计数来计算。四项研究中HLA-DRB1的表达水平绘制在图19A中。

[1017] 肿瘤来源与基质来源的HLA II类表达的表征,与图19B相关

[1018] 为了确定归因于肿瘤与基质的MHC II类结合肽表达的相对量,在TCGA患者的II类途径基因(根据DNA判定)中鉴定了突变,并且对于每个携带II类突变的患者,对相应RNA-Seq基因量化了突变和非突变拷贝的相对表达。此外,还假设突变的读取来自肿瘤,肿瘤中的基质或野生型等位基因出现非突变的读取,并且肿瘤保留了野生型拷贝,其表达量大约等于突变的拷贝。

[1019] 基于此,确定对于观察到的突变等位基因分数f,,归因于肿瘤的II类表达分数约为2f且不大于100%。选择三个基因——CIITA、CD74和CTSS——作为核心II类途径基因,并评估TCGA中的突变(不排除同义和UTR突变)(数据从TumorPortal(tumorportal.org/)下载:BRCA、CRC、HNSC、DLBCL、MM、LUAD;TCGA批量下载(tcga-data.nci.nih.gov):CESC、LIHC、PAAD、PRAD、KIRP、TGCT、UCS;Synapse(synapse.org/#!Synapse:syn1729383):GBM、KIRC、LAML、UCEC、LUSC、OV、SKCM;或原始TCGA出版物(cancergenome.nih.gov/publications):BLCA、RICH、STAD和THCA)。这些基因的选择是基于它们在II类表达中的已知作用以及它们与8500个GTEx样品群组中的HLA-DRB1的紧密相关性。与HLA-DRB1具有同等相关性的其他基因(HLA-DRA1、HLA-DPA1、HLA-DQA1、HLA-DQB1和HLA-DPB1)被排除在外,因为它们的多态性使它们易于出现假阳性突变。自然,只有一小部分患者在CIITA、CD74或CTSS中发生突变,并且对于某些肿瘤类型,没有可供分析的患者。

[1020] 对二进制序列比对/定位(BAM)格式的原始全外显子组测序(WES)的序列进行目视评估(IGV),以确认突变存在于肿瘤样品中,而不存在于正常样品中。使用pysam从相应的RNA-Seq获得突变体与野生型读取计数。基于从Genomic Data Commons(gdc.cancer.gov)下载的表达数据确定总体HLA-DRB1表达,其相对于蛋白质编码基因集重新归一化为总和为1,000,000。归因于肿瘤的HLA-DRB1表达的分数(图19B)被估计为min(1,2f),其中f是CIITA、CD74或CTSS中显示突变的RNA-Seq读取的分数。

[1021] 过度表示和表示不足的基因的鉴定,与图32A和图39B相关

[1022] 从先前发表的MS实验中分析了样品,这些实验描述了卵巢癌、结直肠癌和黑素瘤的MHC-II配体组。来自卵巢癌数据集的许多样品都有可用的RNA-Seq;这些样品的数据从SRA(NCBIBioProjectPRJNA398141)下载,并使用STAR aligner与UCSC hg19转录组进行比对。对于没有可用RNA-Seq的卵巢样品,使用可用RNA-Seq估计所有样品的平均表达。对于结直肠和黑素瘤研究,任何样品都没有相应的RNA-Seq,因此使用来自TCGA(癌症基因组图谱网络)的数据计算了替代样品的平均值。使用RSEM版本1.2.31计算的每百万个中的转录物(TPM)进行转录水平基因定量。通过对基因水平求和、删除非编码基因和重新归一化来进一步处理表达估计值,使得总TPM总和为1000000(蛋白质编码基因的重新归一化解释了ncRNA丰度的文库间差异)。

[1023] 为了鉴定在MHC II配体组中过度表示和表示不足的基因,对表达分析中使用的三个相同的数据集进行了分析。对于每个基因,我们的基线假设是它应该产生与其长度乘以

其表达水平成比例的肽。为了确定每个基因的长度,列举了所有转录物同种型的独特9聚体。通过对转录物同种型求和获得基因水平的表达。在嵌套组水平上确定映射到每个基因的肽的观察的数目(例如肽GKAPILIATDVASRGLDV、GKAPILIATDVASRGLD和KAPILIATDVASRGLDV被计为单个观察)。

[1024] 创建了两个矩阵,分别表示预期和观察到的计数,分别称为E和O,其中行对应于基因,列对应于样品。O中的值是通过在嵌套组级别计算每个样品的肽来确定的。首先通过将每个基因的长度乘以其在每个样品中的表达来填充矩阵E;然后重新调整E的列,以使E的列总和与O的列总和匹配。最后,通过比较E的行总和与O的行总和,在基因水平上进行分析(图32A)。基因根据其在人血浆中的存在和浓度被突出显示。使用来自同一组研究的黑素瘤、结直肠癌和卵巢癌数据,使用相同的方法来鉴定HLA I类数据中过度表示和表示不足的基因。对于HLA I类分析,没有应用嵌套,而只计算独特的肽。

[1025] 过度表示的基因中结合评分的评估,与图39A相关

[1026] 据观察,许多过度表示的基因是血浆基因。获得了一份全面的血清基因列表,并将来自血浆基因的HLADR结合肽与来自非血清基因的HLADR结合肽以及从免疫肽组中表示的基因中取样的长度匹配的非结合(例如在MS中未观察到)肽的neonmhc2结合评分进行了比较。对于具有用泛DR抗体分析的HLA II类肽的基因分型、多等位基因数据集(图30B中分析的相同样品),对样品表达的每个DR等位基因的具有neonmhc2的肽进行评分。neonmhc2在所有表达的等位基因上输出的最佳评分被视为每个肽的代表评分。将所有可用数据集中的数据汇总在一起,并使用箱线图将每个类别的肽的分数分布可视化。

[1027] 与蛋白质周转相关的基因的分析,与图32C相关

[1028] 确定了两个基因集,其代表其周转受蛋白酶体调节的蛋白质。第一组基因包含在细胞系KG1、Jurkat或MM1S中具有至少一个观察到的泛素化位点的基因。第二组包含在应用蛋白酶体抑制剂硼替佐米(BTZ)后水平增加的基因,应用0.01的p值过滤器并选择具有最大向上倍数变化的300个基因。

[1029] 比较成块肿瘤与抗原呈递细胞基因表达的解释力,与图39C相关

[1030] 创建了四个基因表达谱。第一个旨在代表APC,并通过对来自上述单细胞RNA-Seq实验的细胞类型特异性谱进行平均来估计。平均值包括“巨噬细胞”(来自头颈研究、肺研究和黑素瘤研究)、“CLEC9A DCs”(来自肺研究)和“单核树突细胞”(来自肺研究)。其他三个表达谱对应于来自卵巢癌、结直肠癌和黑素瘤的成块肿瘤谱(数据图19A)。卵巢特征是Schuster等人公布的样品的平均值,其他特征来自每个肿瘤类型具有最高肿瘤细胞数目的五个TCGA样品,如先前使用“绝对”算法推断的那样。对于每种肿瘤类型,计算每个基因的肽数目(在嵌套组级别),并使用线性回归将每个基因的肽数目建模为基因长度、APC特异性基因表达和肿瘤特异性基因表达的函数。输出变量和所有输入变量通过log(x+1)进行转换。使用模型的参数估计,肿瘤的贡献计算为$\beta_{tumor}/(\beta_{tumor}+\beta_{APC})$,APC的贡献计算为$\beta_{APC}/(\beta_{tumor}+\beta_{APC})$。对于每个样品,使用基因水平的自举重采样(M=100)来计算解释比例的置信区间。

[1031] 表征观察到的HLA II类肽的切割位点,与图40A相关

[1032] 从六个数据集中分析了自然加工和呈递的HLA II类肽:PBMC绘图、DC样MUTZ3细胞系、结直肠癌组织、黑素瘤、卵巢癌和expi293细胞系。由于许多肽共享相同的N端(例如GKAPILIATDVASRGDV和GKAPILIATDVASRGLD)或相同的C端(例如GKAPILIATDVASRGLD和

KAPILIATDVASRGLD),策划了两组非冗余剪切位点,一组用于N端,一组用于C端。如41图所示的命名系统用于表示肽上游、肽内和肽下游的位置。上行和下行频率(...U1和D1...)与蛋白质组氨基酸频率进行比较,并通过卡方检验对显著偏差进行评分。肽位置(N1...C1)与在MS肽中观察到的频率进行比较。

[1033] 对各种HLA II类切割预测器的性能进行基准化,与图40B相关

[1034] 四个PBMC样品和公开的数据集用于基准切割相关变量/预测器的能力以增强对呈递的HLA II类表位的鉴定。

[1035] 为了构建使用结合潜力和切割潜力预测肽呈递的综合预测器,首先使用图31B中描述的相同方法构建数据集。这意味着使用1:20的命中与诱饵之比,其中诱饵与命中的长度匹配,并从产生至少一个命中的一组基因中随机采样。以这种方式构建了不同的数据集,用于三个不同的目的:

[1036] 1.对于基于溶剂可及性和基于无序的切割预测器,使用来自人类肿瘤组织的HLA II类配体组数据拟合逻辑模型。据推测,对于在配体组实验中观察到的肽,它必须已成功加工。(对于神经网络和基于CNN的分裂预测器,训练数据是使用相同的数据集以不同的方式生成的,如下表所述)。

[1037] 2.为了评估给定的切割预测器是否比单独结合提高了性能,使用B721和KG1细胞生成的单等位基因MAPTAC™数据拟合模型,并询问了功能最强的APC样细胞系。使用neonmhc2计算结合潜力,逻辑回归确定了在正向预测中将置于结合和切割变量上的相对权重。

[1038] 3.为了评估前向预测的性能,以与以前相同的方式为PBMC样品和已发布的数据集构建数据集。然而,因为这些样品是多等位基因的,每个候选肽的结合评分被认为是每个供体基因型所指示的1-4个DR等位基因的最大评分。PPV的计算方法如图31B所述。.

[1039] 评估了几种不同的切割预测器

[1040] 先切割模型,切割位置已知(神经网络)

[1041] 为了从MS观察到的切割位点了解切割信号,从U3到N3和C3到D3的所有独特的6聚体氨基酸序列(使用部分中介绍的命名法,“表征观察到的HLA II类肽的切割位点,与图40A相关”)在肿瘤组织衍生的HLA II类配体组中用作阳性示例,用于训练两个不同的神经网络,分别对N端切割和C端切割建模。和以前一样,通过从蛋白质组的氨基酸频率中提取上下文,以及从MS观察到的配体组,合成生成了相同数目的独特的未观察到的N端和C端切割位点(阴性示例)。氨基酸序列用neonmhc2中使用的相同特征的子集编码,特别是基于蛋白质结构和氨基酸特性(例如酸性、脂肪族等)的氨基酸邻近性。对于每个N端和C端观察到的切割点模型,(lr=0.0005,Adam优化器,二元交叉熵作为损失函数),然后用两个隐藏层(一层中有20个神经元)训练一个完全连接的神经网络,接下来是10个神经元)和ReLu激活,然后是最后的sigmoid层。对于正则化,使用0.001的L2范数(仅适用于C终端模型)和4的最大范数约束使用20%的辍学率。

[1042] 为了对候选肽进行评分,将N端模型应用于与肽相关的6聚体序列U3至N3,并将C端模型应用于C3至D3。N端和C端模型也应用于跨候选肽平铺的6聚体序列,以评估肽本身内序列的切割倾向。使用neonmhc2结合评分以及四个神经网络输出(对应于N端和C端模型的N端、C端和最大评分切割位点)对MAPTAC™数据进行逻辑回归训练肽。

[1043] 先切割模型,切割位点未知(+/-15AA)(神经网络)

[1044] 为了确定当肽的精确末端未知时,从观察到的切割位点学习的切割模型是否具有预测性,将上面学习的相同神经网络应用于扩展上下文,即肽末端之外的15个氨基酸。在这种情况下,为了对候选肽进行评分,计算了三个区域的最大评分:肽上游的15个氨基酸(无论真实N端切割位点的位置如何),使用N端模型对其进行评分,肽序列,用N端和C端模型评分,以及肽下游的15个氨基酸,用C端模型评分。使用neonmhc2结合评分以及四个区域特异性(因为肽本身贡献两组值,来自N端和C端模型)评分对MAPTAC™数据进行逻辑回归训练。

[1045] 先结合模型,溶剂可及性

[1046] 在SCRATCH套件中,工具ACCpro20用于预测相对溶剂可及性。给定肽的平均溶剂可及性分数的肽被加工的可能性然后与使用肿瘤组织数据的逻辑回归拟合。最后,使用neonmhc2结合评分和肿瘤组织训练预测器的输出对单等位基因数据进行逻辑回归训练。

[1047] 先结合模型,无序

[1048] 在整个蛋白质组中确定序列无序的每个残基分数,根据将位置标记为无序的预测引擎的数目以0-5的等级评分(使用的服务器:锚、espritz-d、espritz-n、espritz-x、iupred-1和iupred-s)。计算每个候选肽的平均无序分数,并将六个无序预测器输出相加。与溶剂可及性一样,首先使用该总体无序评分与肿瘤组织数据拟合逻辑模型。然后使用neonmhc2结合评分和肿瘤组织训练预测器的输出对单等位基因数据进行逻辑回归训练。

[1049] 混合模型,前体切割扫描(+/-30AA)(CNN)

[1050] 命中的训练数据按照“先切割,已知切割位点”切割预测器的描述生成,不同之处在于,不是使用从U3到N3和C3到D3的独特6聚体序列,而是使用肽侧翼的30个氨基酸(U30至U1和D1至D30)作为模型输入。此外,没有区分30聚体序列是来自N端还是C端侧翼,而是将数据合并以训练单个模型来学习前体切割信号,该信号假设可能出现在观察到的肽的任一侧。在这种情况下,不是使用合成诱饵,而是使用来自相同源基因的未观察到的肽的侧翼序列作为阴性示例。使用基于蛋白质结构和氨基酸特性(例如酸性、脂肪族等)的氨基酸邻近性,如前所述编码序列。CNN的架构由两个卷积层组成,第一层的内核大小为2,有48个过滤器,然后是内核大小为3和40个过滤器的层。这些层具有ReLu激活。卷积层之后是全局最大池化层,之后是具有sigmoid激活的最终密集层。CNN以0.001的学习率进行训练,以Adam优化和二元交叉熵作为损失函数。

[1051] 为了对候选肽进行评分,将CNN应用于肽上游的30个氨基酸和下游的30个氨基酸,产生N端侧翼评分和C端侧翼评分。使用neonmhc2结合分数和两个CNN分数对MAPTAC™数据进行逻辑回归训练。

[1052] DQ重叠

[1053] Bergseng等人,2015在HLA-DQ配体组中汇总了基于MS的肽鉴定。创建了一个新特征,表示新的候选肽是否与这些先前观察到的肽之一重叠。具体而言,如果该特征与先前观察到的HLA-DQ配体组中的任何肽共享至少一个9聚体,则该特征被设置为1;否则该特征被设置为0。使用neonmhc2结合评分和重叠特征对单等位基因数据进行逻辑回归训练。

[1054] 整合的结合和切割模型也全部拟合并使用NetMHCIIpan作为结合预测器进行评估,而不是在图40B中。

[1055] 评估对天然供体组织的预测整体性能,与图21A-21B相关。

[1056] 如上文“基于抗体的HLA-肽复合物分离”部分所述,使用DR特异性抗体对来自七名健康供体的外周血进行分析。训练和评估数据集是使用先前关于图31B描述的命中和诱饵选择算法构建的。简而言之,这意味着用一个命中肽(嵌套组中最短的肽)表示每个嵌套组,并在基因上平铺长度匹配的诱饵,这样它们与命中的重叠最少,彼此之间的重叠也最少。在这种情况下,诱饵选择不限于MS观察到的基因,而是从整个蛋白质组中随机取样而不取代。使用了1:499的诱饵:命中比,反映了蛋白质组中HLA-DR呈递肽频率的粗略估计。具有MHC结合评分(来自NetMHCIIpan或neonmhc2)以及其他输入特征(表达、基因偏好和DQ重叠)的逻辑回归模型在来自KG1和B721细胞系的MAPTAC™数据上进行了训练。

[1057] 表15中的以下变量用于回归的子集中。

[1058] 表15

[1059]

NetMHCIIpan 百分排名	源自 NetMHCIIpan-v3.1。对于每个候选肽,在供体基因型中所有 DR 等位基因中取最高评分。
neonmhc2 百分排名	源自 *neonmhc2*。对于每个候选肽,在供体基因型中所有 DR 等位基因取最高评分。
表达	基因表达估计是通过分析来自(bowtie2、RSEM 和仅对蛋白质编码基因的重新归一化)的数据获得的,值是 N 个样本的平均值。表达被阈值化(0/1,表示表达是否为非零)或被视为连续变量。
基因偏好	(1+观察)/(1+预期),根据图 32A 中的分析
DQ 重叠	候选肽是否与任何 HLA-DQ 数据集共享至少一个 9 聚体的指示变量(0/1)。HLA-DQ 结合肽与 HLA-DR 结合肽具有不同的结合基序,因此不应通过此特征来学习结合倾向。

[1060] 然后评估这些模型在来自天然供体组织(PBMC样品等)的HLA-DR配体组上的性能。诱饵是从蛋白质组中随机采样的(包括从未产生MS观察到的肽的基因),以达到1:499的命中与诱饵之比,这几乎使可用的诱饵序列饱和。1:499的诱饵:命中比用于评估(以及训练)。评估数据集中得分最高的0.2%肽被标记为阳性判定,并且PPV被计算为作为命中的阳性检测的分数(参见例如图21A和表15)。请注意,由于阳性数被限制为等于命中数,因此在此评估场景中,判定率恰好等于PPV。在等位基因之间应用一致的1:499比率有助于稳定性能值,否则性能值会受到每个供体观察到的命中数的高度影响。这被认为是合适的,因为假设与供体细胞的内在特性相比,命中次数更多地与实验条件相关。

[1061] 表16

[1062]

MS 样品	对于 NetMHCIIpan 的 PPV	对于 *neonmhc2* 的 PPV
肺	0.05	>0.3
脾脏	0.025	~0.3

[1063]

Heyder LCLs	0.01	~0.2
A375	0.01	~0.3
mDCs	0.05	>0.44
PBMC 样品#1	0.01	~0.3
PBMC 样品#2	0.01	>0.3
PBMC 样品#3	0.025	>0.3
PBMC 样品#4	0.0025	~0.275
Ritz DOHH2	0.01	0.3
Ritz Maver1	0.015	~0.175

[1064] DC呈递的肿瘤肽的基于SILAC的鉴定,与图33A相关。

[1065] 为了产生单核细胞衍生的树突细胞(mDC),按照制造商的方案(Miltenyi Biotec),使用人CD14微珠通过磁分离从健康供体外周血单核细胞(PBMC)中分离出CD14+单核细胞。分离的细胞在补充有800U/ml rh GM-CSF和400U/ml rh IL-4(CellGenix,德国)的CellGenix GMP DC培养基中分化6天。K562细胞(ATCC,Manassas,VA)使用稳定同位素标记与细胞培养中的氨基酸(SILAC)进行同位素标记。细胞在含有重同位素氨基酸L-赖氨酸2HCl 13C6 15N2(Life Technologies,Carlsbad,CA)和L-亮氨酸13C6(Life Technologies,Carlsbad,CA)与15%热灭活、透析的胎牛血清(ThermoFisher)的SILAC用RPMI 1640培养基(ThermoFisher)的存在下生长5次倍增。SILAC标记的K562细胞如前所述使用60mM次氯酸(HOCl)裂解,或在室温下用紫外线处理3小时以诱导细胞凋亡并静置过夜。7,500万个mDC与紫外线处理的SILAC标记的K562细胞以1:3的比例在37℃下共培养14小时,或以1:3的比例与HOCl裂解的K562在37℃下共培养10分钟或5小时。共培养后,收获细胞、沉淀并在液氮中快速冷冻用于蛋白质组学分析。

[1066] DC呈递的肿瘤肽的预测和表达分析,与图33B和图21C相关。

[1067] 为了计算用于预测重标记(肿瘤衍生的)肽的PPV,使用与图33B中使用的相同的模型和评估方法。K562细胞系的表达基于来自ENCODE的数据(encodeproject.org/experiments/ENCSR545DKY/;文库ENCLB075GEK和ENCLB365AUY;(ENCODE Project Consortium,2012))来确定。树突细胞的表达基于GSE116412(对GSM3231102、GSM3231111、GSM3231121、GSM3231133、GSM3231145进行平均)来确定。

[1068] 实施例14.用胰蛋白酶对FAIMS进行基准化

[1069] 该实施例中,描述了使用高场不对称波形离子迁移光谱法(FAIMS)的标准HLA肽组学工作流程。表征了来自A375细胞的内源性加工和呈递的HLA I类和HLA II类肽。肽在通过

nLC-MS/MS分析之前经历酸性反相(aRP)和碱性反相(bRP)离线分级分离,其中使用未配备(-)和配备有(+)FAIMS Pro接口的Thermo Scientific Orbitrap Fusion Lumos Tribrid质谱仪。工作流程在图42A中描绘的图表中指示。图42B显示的结果表明FAIMS改进了低至10ng的胰蛋白酶样品中的肽检测。尽管较低的MS1强度,FAIMS增加了整个LC梯度的HLA-1和HLA II类肽检测(图43A和图44A,分别为HLA-1和HLA II类肽的数据)。引人注目的是,采用FAIMS评估,在酸性和碱性反相样品中观察到独特肽检测的增加(图43B和图44B,分别为HLA-1和HLA II类肽的数据)。该研究表明,使用FAIMS在整个LC梯度中对HLA I类和HLA II类肽的检测增加,尽管MSI强度较低。结合离线分级分离和FAIMS增加了HLA肽组库的分析深度,如图45和图46所示(分别为HLA-1和HLA II类肽)。

[1070] 实施例15-差示扫描荧光法(DSF)肽交换测定

[1071] 如下进行肽交换测定:将以下试剂(表17)合并并在37℃下混合18小时。

[1072] 表17

[1073]

成分	储备浓度	终浓度
凝血酶消化的 DRB1*15:01 (CLIP0), “DR15”	85.6mM	5 mM
交换肽/目的肽 [A]	10mM	100 mM
乙酸钠	1M	100mM
氯化钠	5M	50mM

[1074]

辛基葡糖苷	10%	1%
MiliQ 水	-	加至 100 ml

[1075] [A]使用具有以下序列的DMSO或肽:PPIDGYPNHPCFEPE(M230)、PQILPYPAPEEAQEN(M231)、PQLRQWWAQGADPLA(M247)、LLRPGQIVAFDSTAQ(M248)或ASLRSWPSTWAPWAS(M371)。

[1076] 然后使用PD minitrap G-25脱盐柱更换缓冲液。Sypro橙色染料(Fisher S6651)在100%DMSO中稀释至1000倍。在脱盐缓冲液中制备50μL 100X Sypro橙色染料工作储备液。将2μL 100X sypro橙色染料和18μL脱盐肽交换的样品转移到384白色PCR微孔板的孔中并混合。然后将板用透明板密封剂覆盖,并在Roche lightcycler 480中对板进行以下程序:(1)加热至25℃,保持10秒;(2)将温度升至99℃,读板20次/1℃;(3)将温度降至25℃并保持10秒。然后计算解链温度。示例性结果显示在以下表18中。

[1077] 表18

[1078]

样品	DT_m(样品T_m-对照T_m)	平均T_m	斜率	初始	低	峰值
DR15+M230	21.4	79.7	0.5	2.8	2.0	4.8
DR15+M231	15.6	73.9	0.3	2.7	1.9	5.8
DR15+M247	20.1	78.4	0.4	3.0	2.3	5.0
DR15+M248	5.6	63.9	0.4	1.8	1.4	5.1
DR15+M371	19.9	78.2	0.4	2.5	1.9	4.9
DR15+DMSO(对照)	NA	58.3	0.8	2.1	1.7	9.2

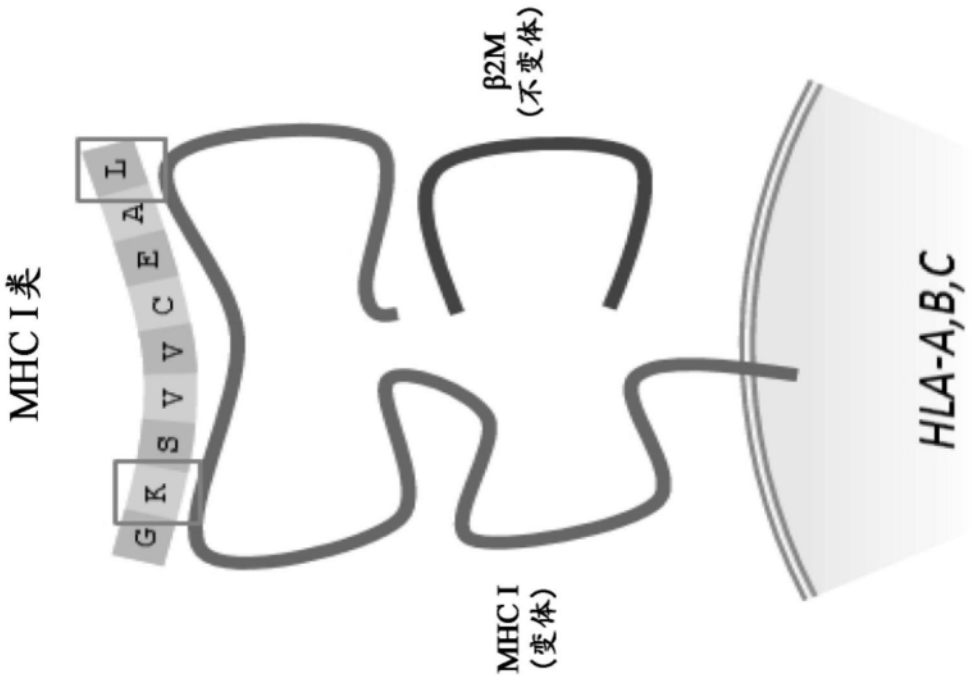

MHC I类
G K S V V C E A L
β2M
(不变体)
MHC I
(变体)
HLA-A,B,C

图1A

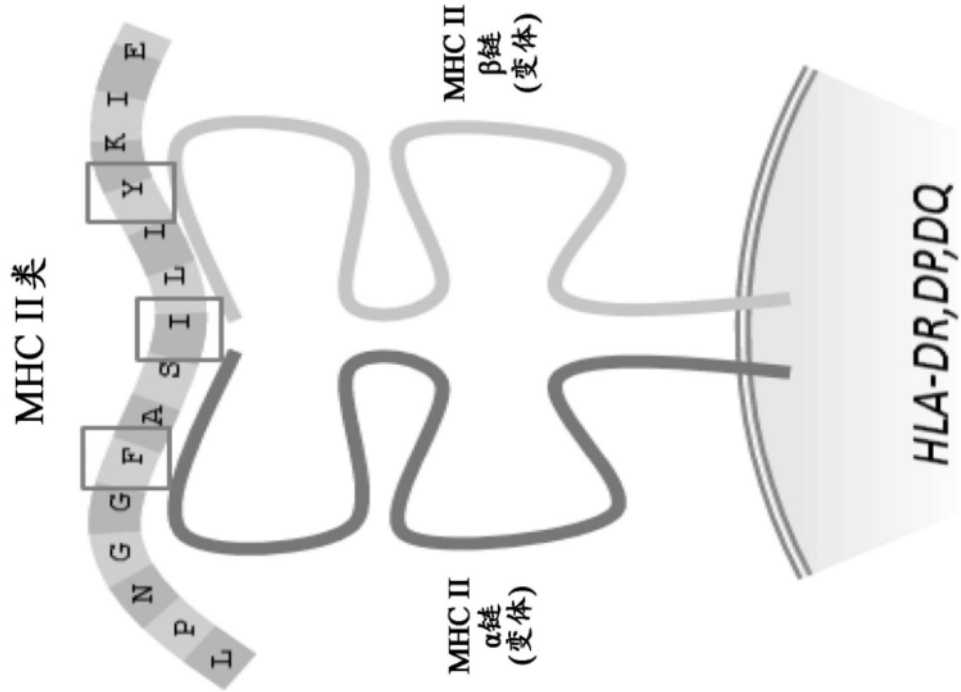

MHC II类
G G F A S I L I Y K I E
N P L
MHC II
β链
(变体)
MHC II
α链
(变体)
HLA-DR,DP,DQ

图1B

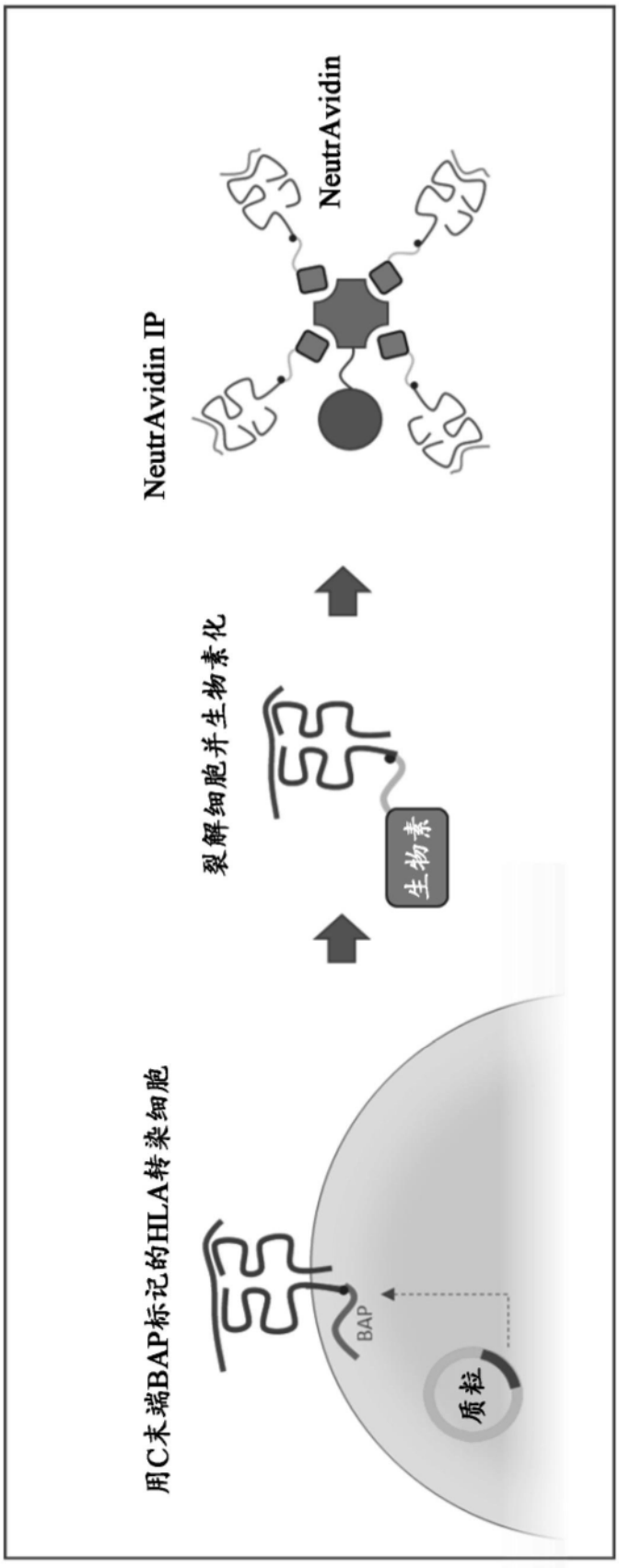

用C末端BAP标记的HLA转染细胞
BAP
质粒
裂解细胞并生物素化
生物素
NeutrAvidin IP
NeutrAvidin

图2

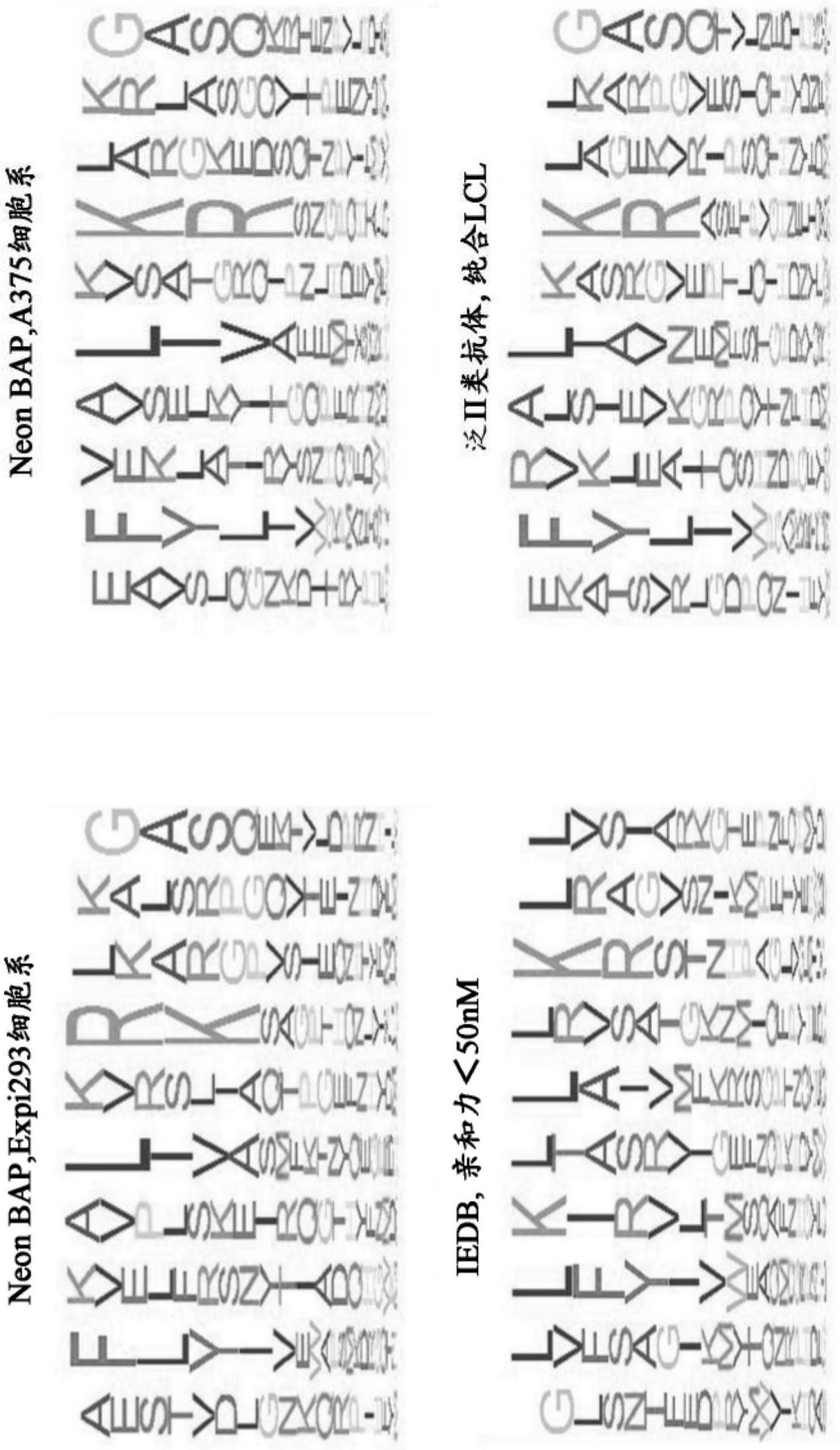

Neon BAP,A375细胞系
泛II类抗体,纯合LCL
Neon BAP,Expi293细胞系
IEDB,亲和力<50nM

图3

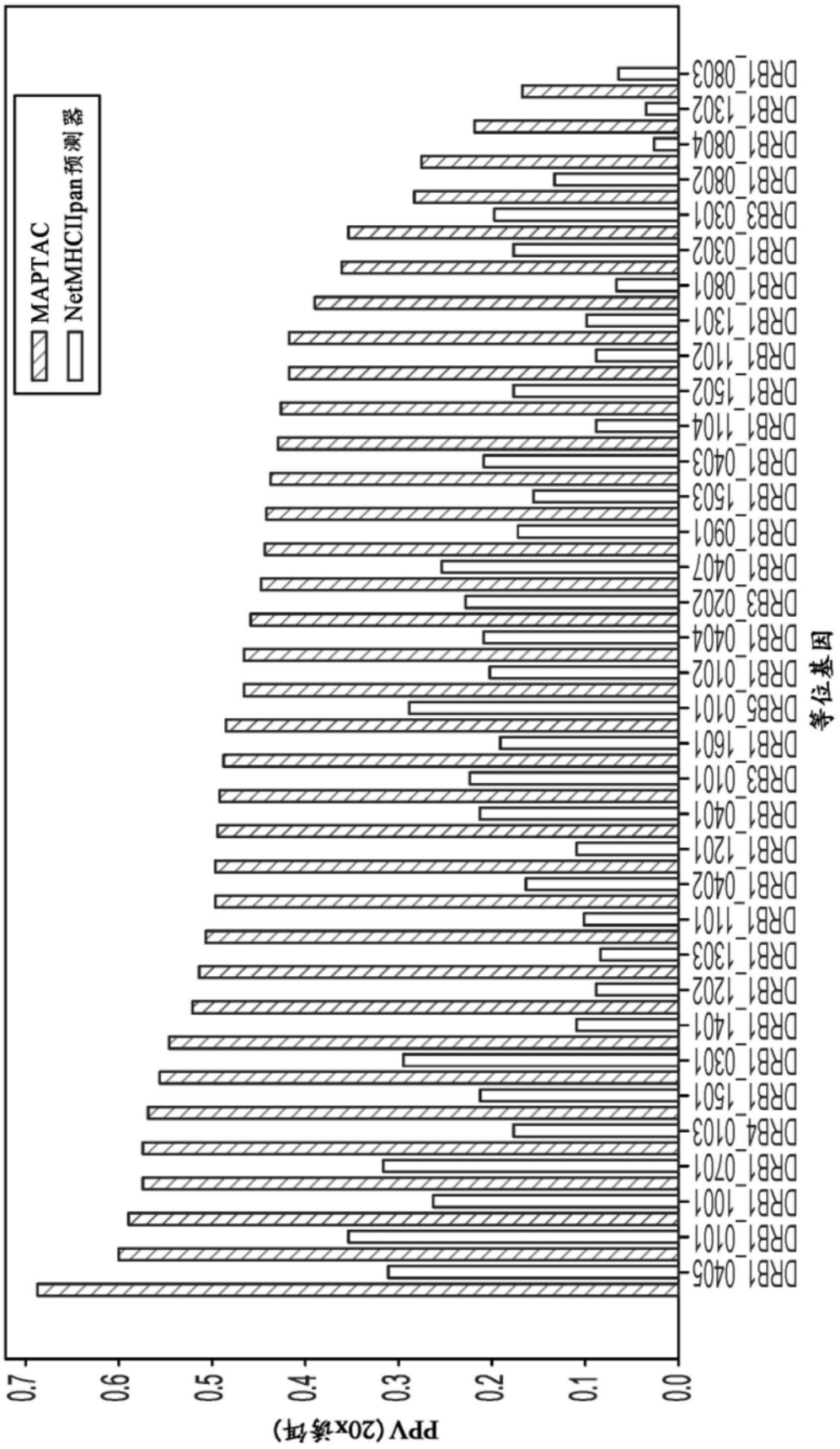

MAPTAC
NetMHCIIpan预测器
PPV(20x诱饵)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
等位基因
DRB1_0405
DRB1_0101
DRB1_1001
DRB1_0701
DRB4_0103
DRB1_1501
DRB1_0301
DRB1_1401
DRB1_1202
DRB1_1303
DRB1_1101
DRB1_0402
DRB1_1201
DRB1_0401
DRB3_0101
DRB1_1601
DRB5_0101
DRB1_0102
DRB1_0404
DRB3_0202
DRB1_0407
DRB1_0901
DRB1_1503
DRB1_0403
DRB1_1104
DRB1_1502
DRB1_1102
DRB1_1301
DRB1_0801
DRB1_0302
DRB3_0301
DRB1_0802
DRB1_0804
DRB1_1302
DRB1_0803

图4

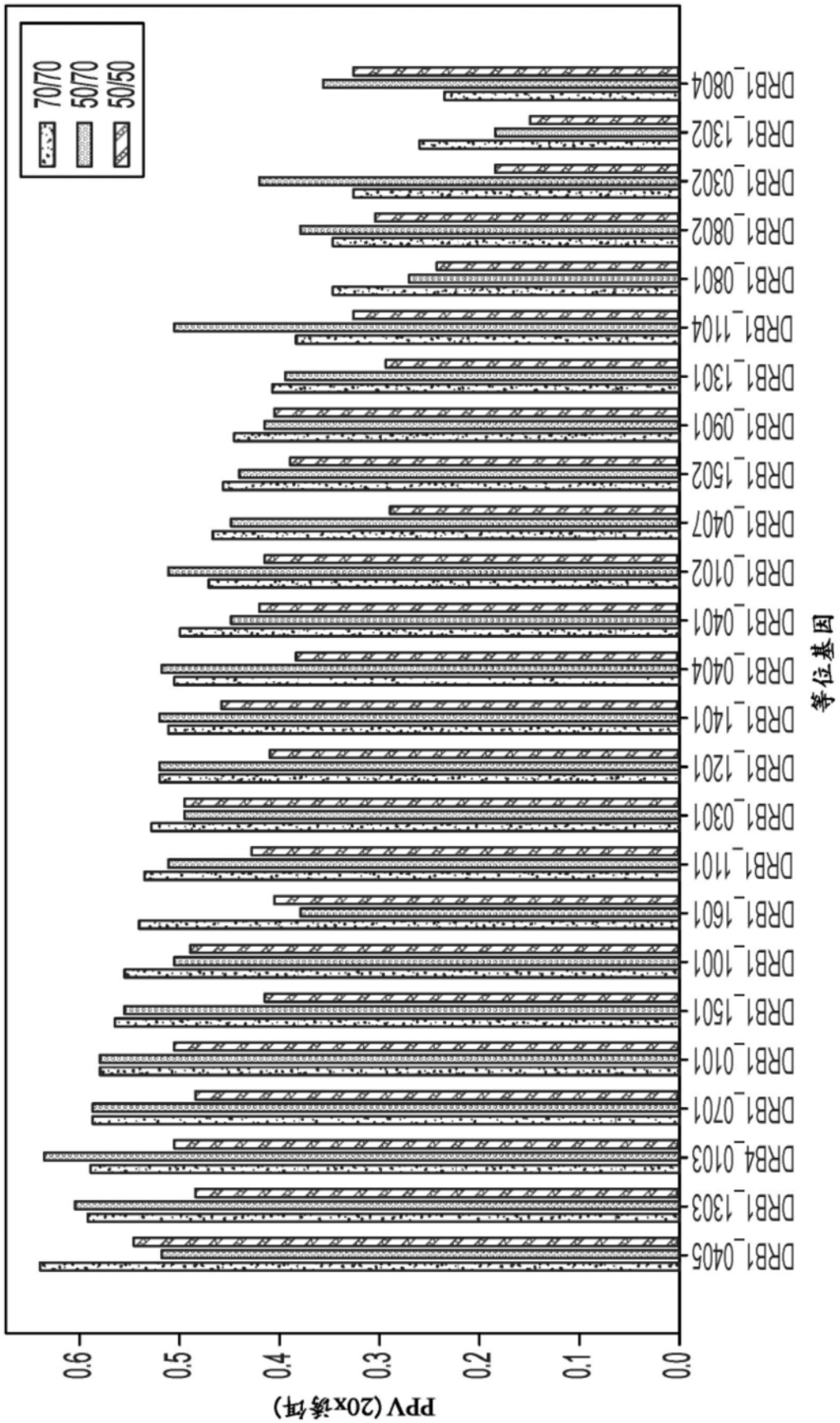

70/70
50/70
50/50
DRB1_0405
DRB1_1303
DRB4_0103
DRB1_0701
DRB1_0101
DRB1_1501
DRB1_1001
DRB1_1601
DRB1_1101
DRB1_0301
DRB1_1201
DRB1_1401
DRB1_0404
DRB1_0401
DRB1_0102
DRB1_0407
DRB1_1502
DRB1_0901
DRB1_1301
DRB1_1104
DRB1_0801
DRB1_0802
DRB1_0302
DRB1_1302
DRB1_0804
等位基因
0.0
0.1
0.2
0.3
0.4
0.5
0.6
PPV (20x诱饵)

图5

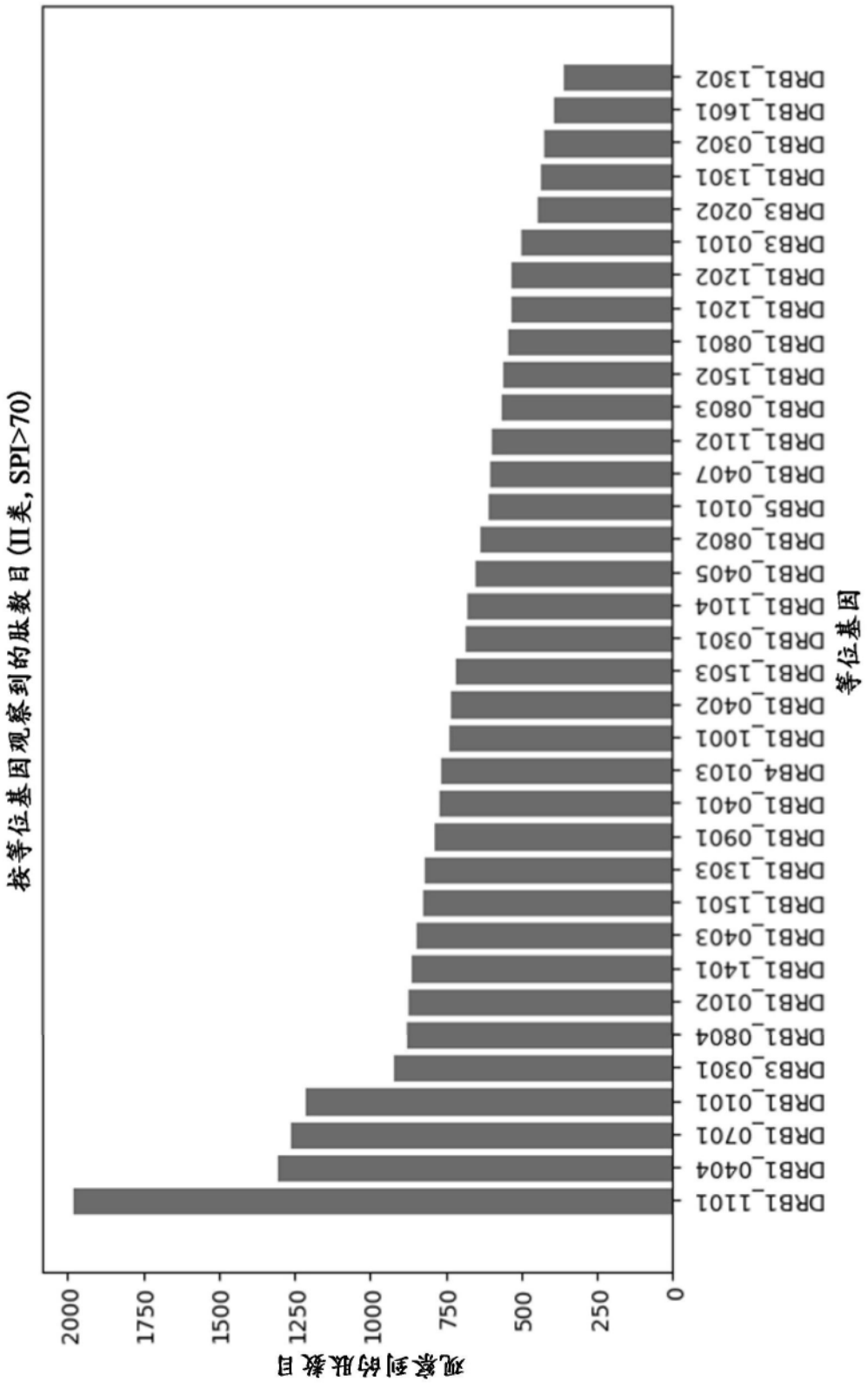

按等位基因观察到的肽数目(II类, SPI>70)
观察到的肽数目
等位基因
2000
1750
1500
1250
1000
750
500
250
0
DRB1_1101
DRB1_0404
DRB1_0701
DRB1_0101
DRB3_0301
DRB1_0804
DRB1_0102
DRB1_1401
DRB1_0403
DRB1_1501
DRB1_1303
DRB1_0901
DRB1_0401
DRB4_0103
DRB1_1001
DRB1_0402
DRB1_1503
DRB1_0301
DRB1_1104
DRB1_0405
DRB1_0802
DRB5_0101
DRB1_0407
DRB1_1102
DRB1_0803
DRB1_1502
DRB1_0801
DRB1_1201
DRB1_1202
DRB3_0101
DRB3_0202
DRB1_1301
DRB1_0302
DRB1_1601
DRB1_1302

图6

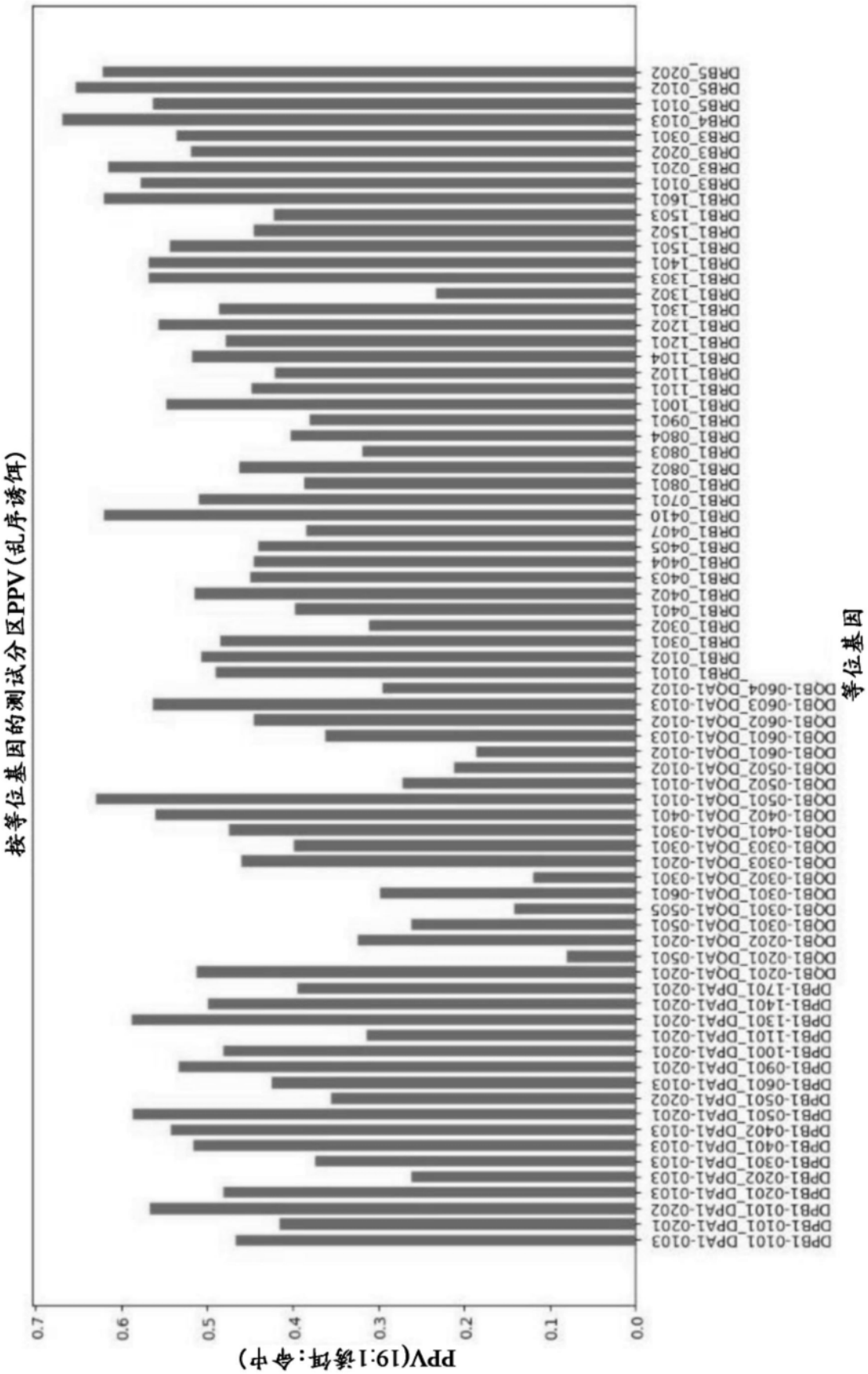

按等位基因的测试分区PPV(乱序诱饵)
PPV(19:1诱饵:命中)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
等位基因
DPB1-0101_DPA1-0103
DPB1-0101_DPA1-0201
DPB1-0101_DPA1-0202
DPB1-0201_DPA1-0103
DPB1-0202_DPA1-0103
DPB1-0301_DPA1-0103
DPB1-0401_DPA1-0103
DPB1-0402_DPA1-0103
DPB1-0501_DPA1-0201
DPB1-0501_DPA1-0202
DPB1-0601_DPA1-0103
DPB1-0901_DPA1-0201
DPB1-1001_DPA1-0201
DPB1-1101_DPA1-0201
DPB1-1301_DPA1-0201
DPB1-1401_DPA1-0201
DPB1-1701_DPA1-0201
DQB1-0201_DQA1-0201
DQB1-0201_DQA1-0501
DQB1-0202_DQA1-0201
DQB1-0301_DQA1-0501
DQB1-0301_DQA1-0505
DQB1-0301_DQA1-0601
DQB1-0302_DQA1-0301
DQB1-0303_DQA1-0201
DQB1-0303_DQA1-0301
DQB1-0401_DQA1-0301
DQB1-0402_DQA1-0401
DQB1-0501_DQA1-0101
DQB1-0502_DQA1-0101
DQB1-0502_DQA1-0102
DQB1-0601_DQA1-0102
DQB1-0601_DQA1-0103
DQB1-0602_DQA1-0102
DQB1-0603_DQA1-0103
DQB1-0604_DQA1-0102
DRB1_0101
DRB1_0102
DRB1_0301
DRB1_0302
DRB1_0401
DRB1_0402
DRB1_0403
DRB1_0404
DRB1_0405
DRB1_0407
DRB1_0410
DRB1_0701
DRB1_0801
DRB1_0802
DRB1_0803
DRB1_0804
DRB1_0901
DRB1_1001
DRB1_1101
DRB1_1102
DRB1_1104
DRB1_1201
DRB1_1202
DRB1_1301
DRB1_1302
DRB1_1303
DRB1_1401
DRB1_1501
DRB1_1502
DRB1_1503
DRB1_1601
DRB3_0101
DRB3_0201
DRB3_0202
DRB3_0301
DRB4_0103
DRB5_0101
DRB5_0102
DRB5_0202

图7A

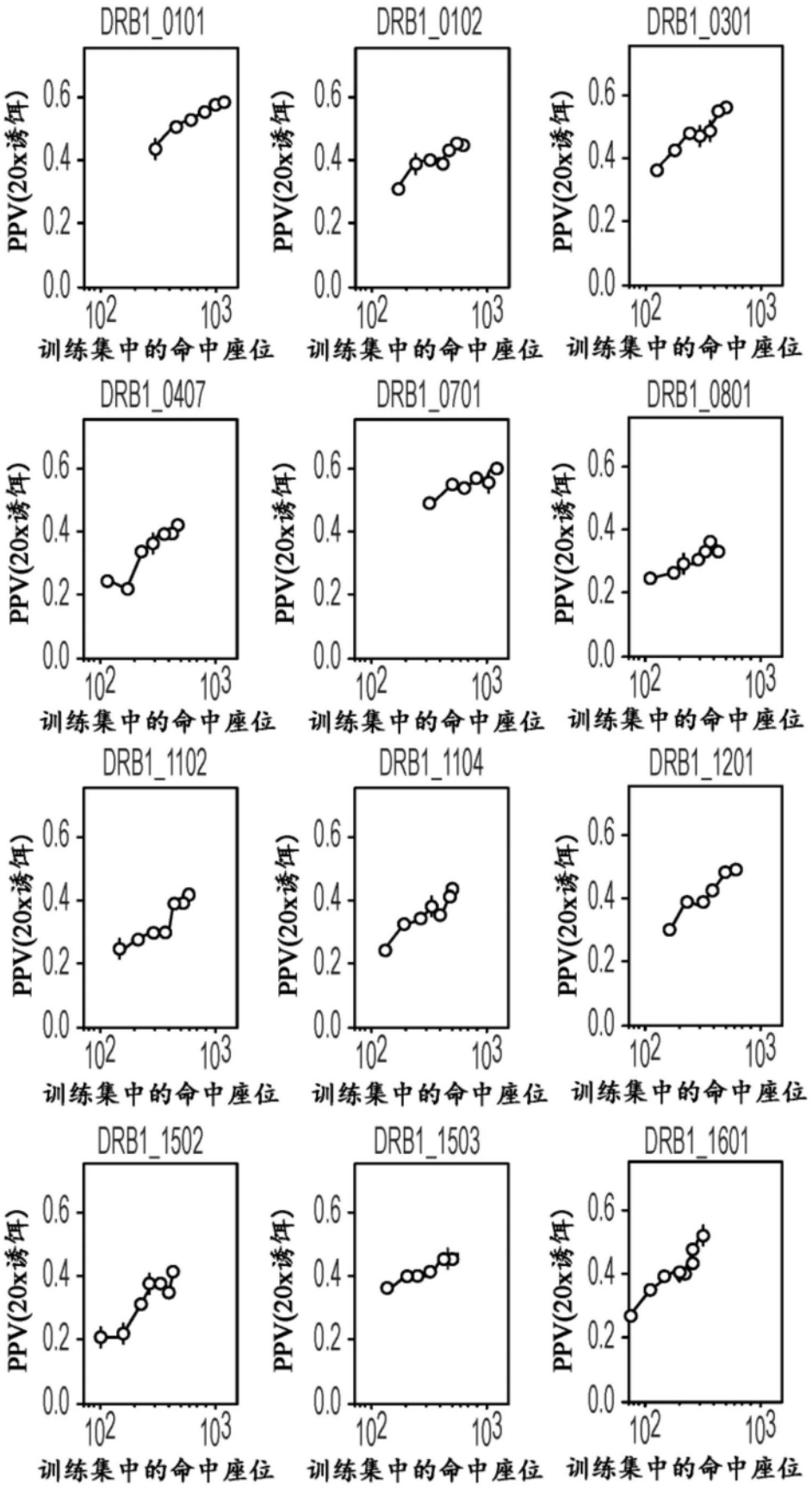

DRB1_0101
DRB1_0102
DRB1_0301
DRB1_0407
DRB1_0701
DRB1_0801
DRB1_1102
DRB1_1104
DRB1_1201
DRB1_1502
DRB1_1503
DRB1_1601
PPV(20x诱饵)
0.0
0.2
0.4
0.6
10^2
10^3
训练集中的命中座位

图7B

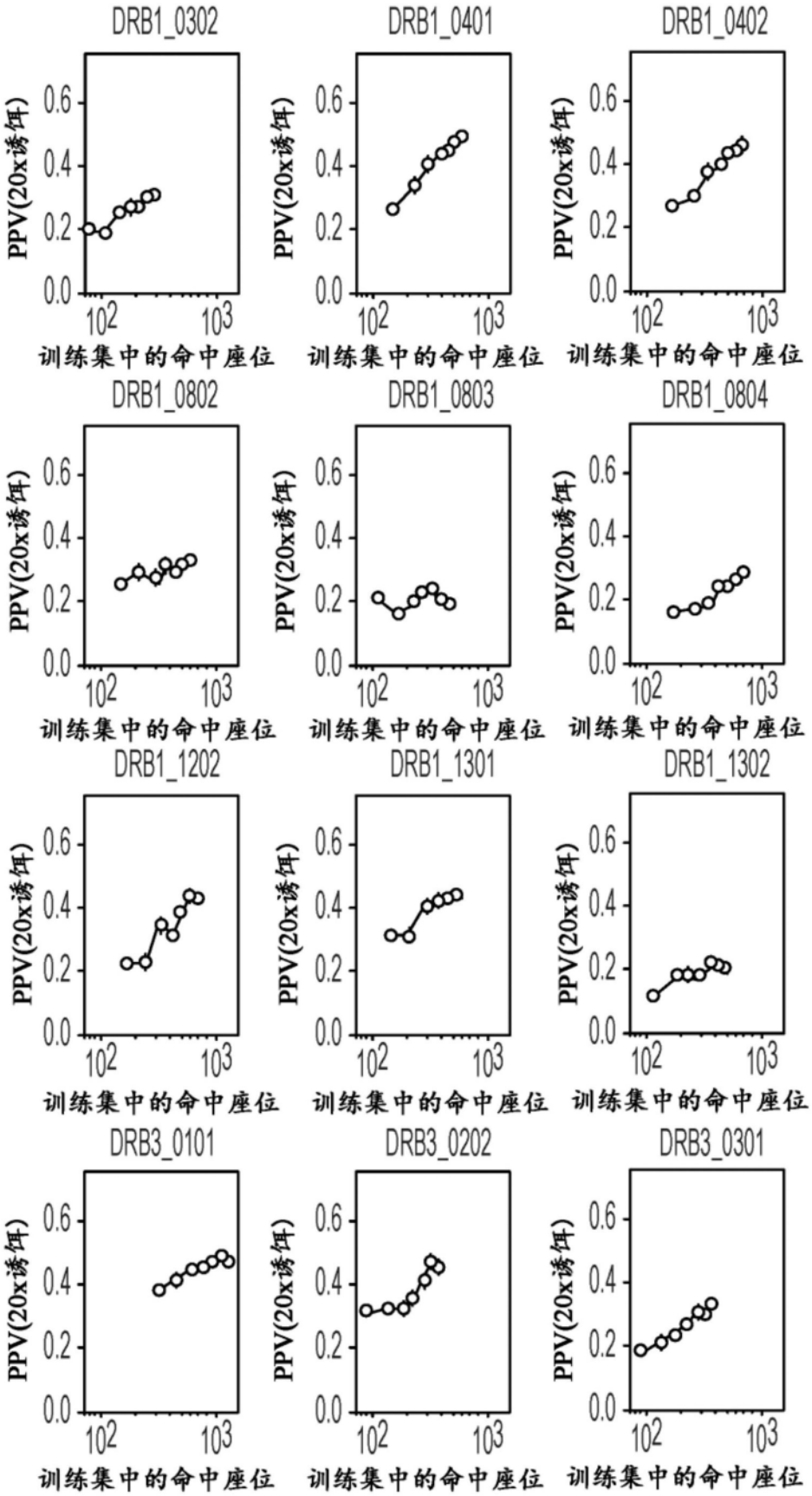

DRB1_0302
DRB1_0401
DRB1_0402
DRB1_0802
DRB1_0803
DRB1_0804
DRB1_1202
DRB1_1301
DRB1_1302
DRB3_0101
DRB3_0202
DRB3_0301
PPV(20x诱饵)
0.0
0.2
0.4
0.6
10^2
10^3
训练集中的命中座位

图7B续

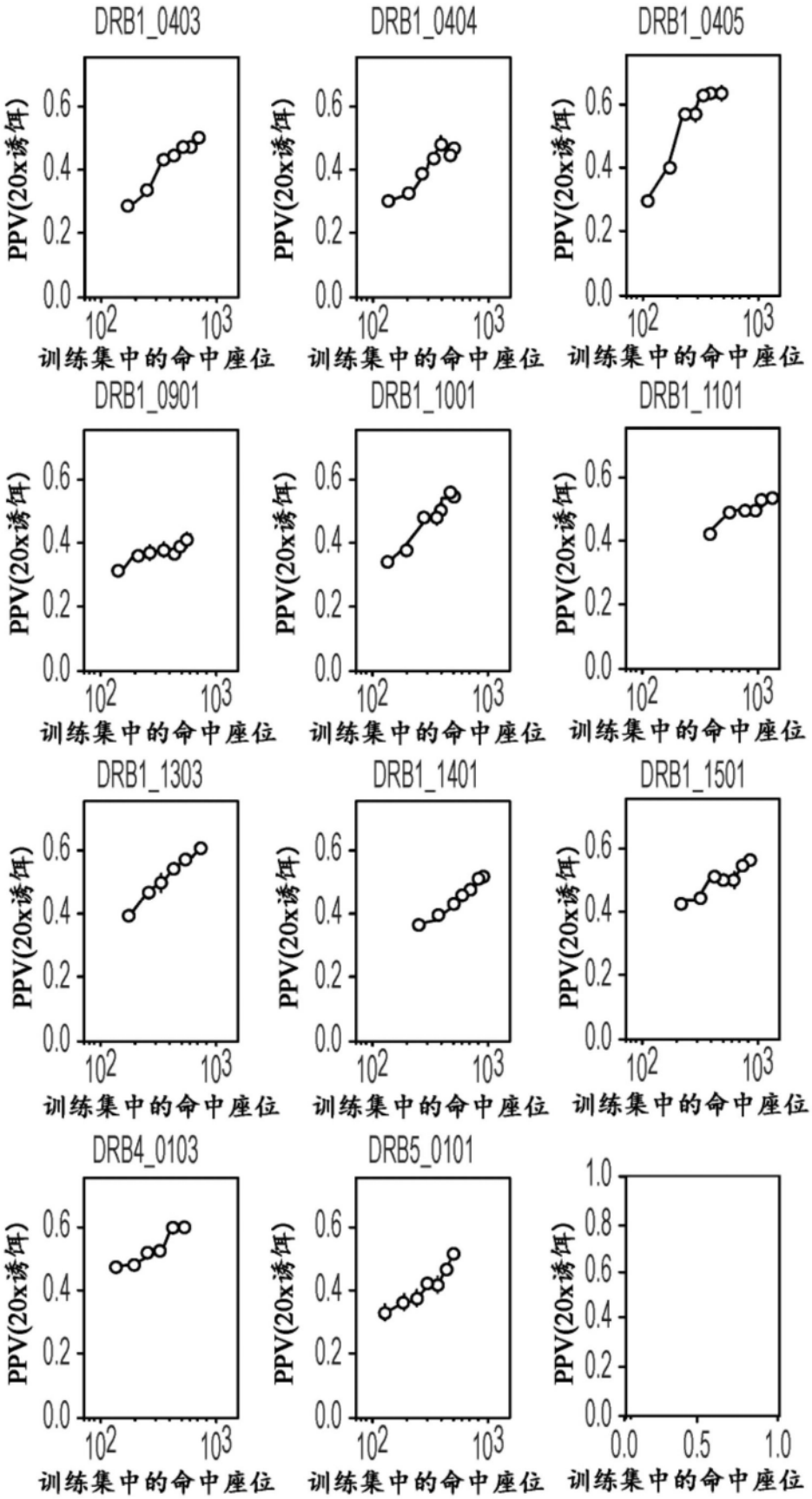

DRB1_0403
DRB1_0404
DRB1_0405
DRB1_0901
DRB1_1001
DRB1_1101
DRB1_1303
DRB1_1401
DRB1_1501
DRB4_0103
DRB5_0101
PPV(20x诱饵)
训练集中的命中座位

图7B续

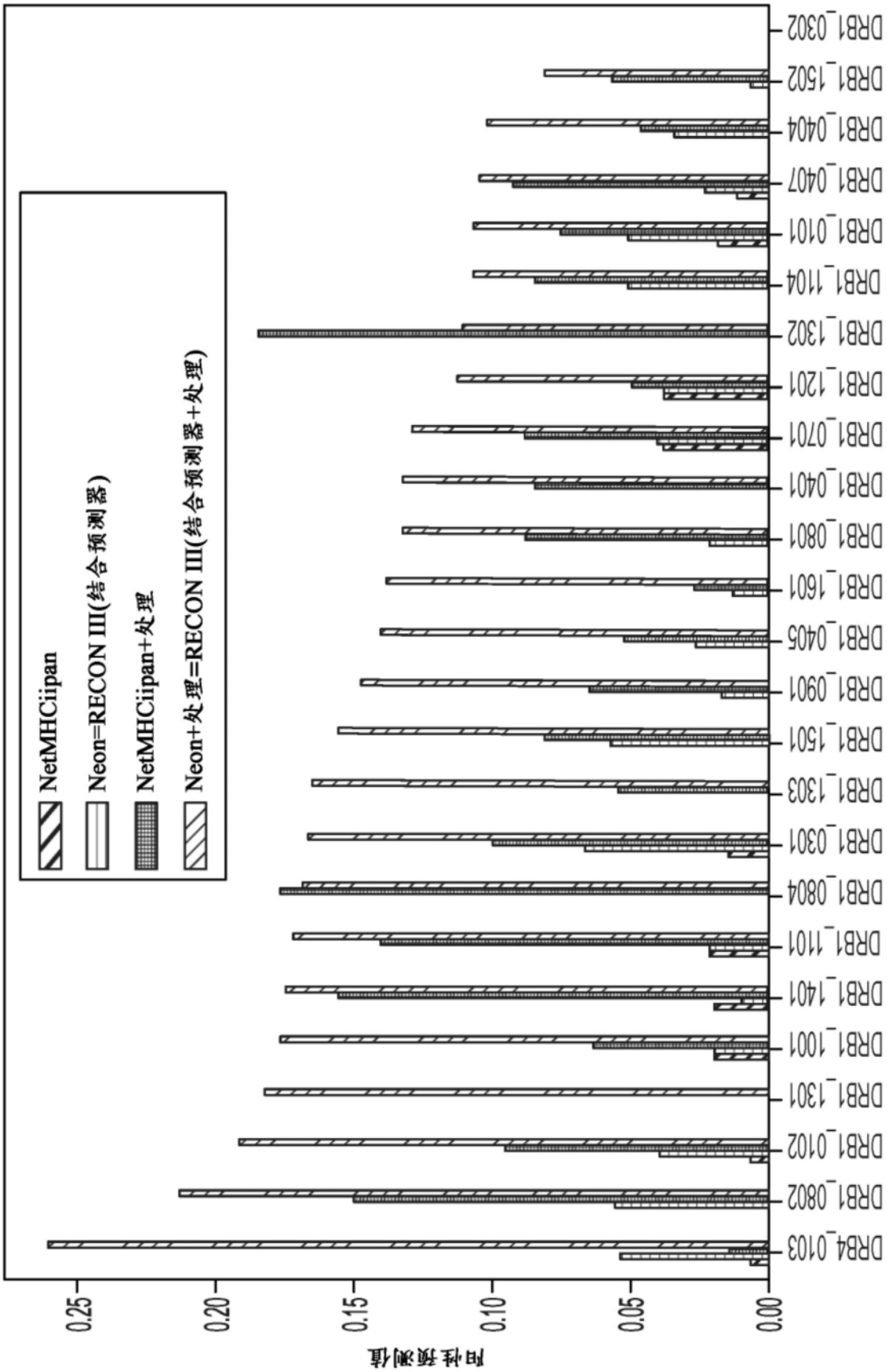

NetMHCiipan
Neon=RECON III(结合预测器)
NetMHCiipan+处理
Neon+处理=RECON III(结合预测器+处理)
阳性预测值
0.25
0.20
0.15
0.10
0.05
0.00
DRB4_0103
DRB1_0802
DRB1_0102
DRB1_1301
DRB1_1001
DRB1_1401
DRB1_1101
DRB1_0804
DRB1_0301
DRB1_1303
DRB1_1501
DRB1_0901
DRB1_0405
DRB1_1601
DRB1_0801
DRB1_0401
DRB1_0701
DRB1_1201
DRB1_1302
DRB1_1104
DRB1_0101
DRB1_0407
DRB1_0404
DRB1_1502
DRB1_0302

图8

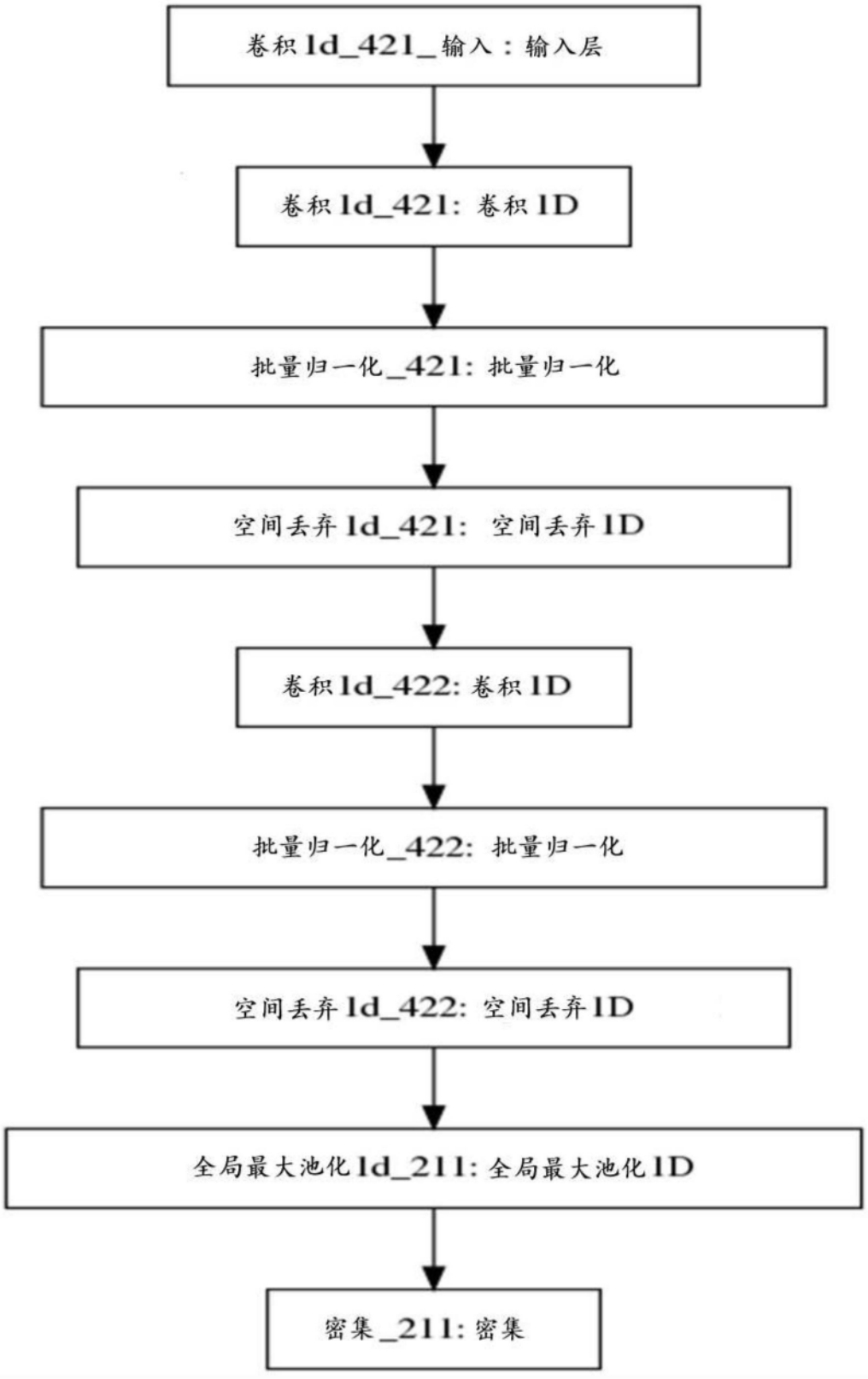

卷积1d_421_输入：输入层
卷积1d_421: 卷积1D
批量归一化_421: 批量归一化
空间丢弃1d_421: 空间丢弃1D
卷积1d_422: 卷积1D
批量归一化_422: 批量归一化
空间丢弃1d_422: 空间丢弃1D
全局最大池化1d_211: 全局最大池化1D
密集_211: 密集

图9

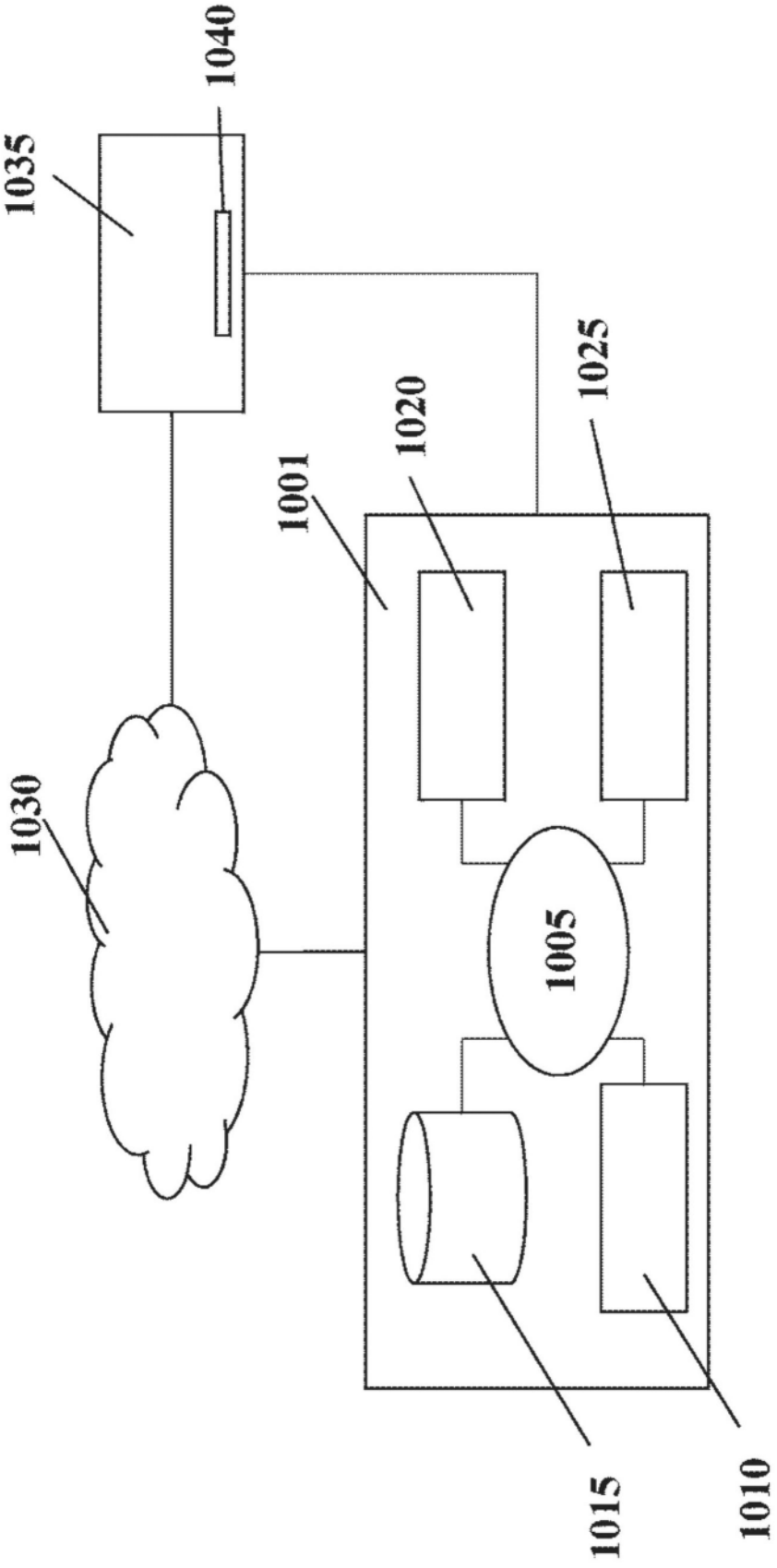


图10

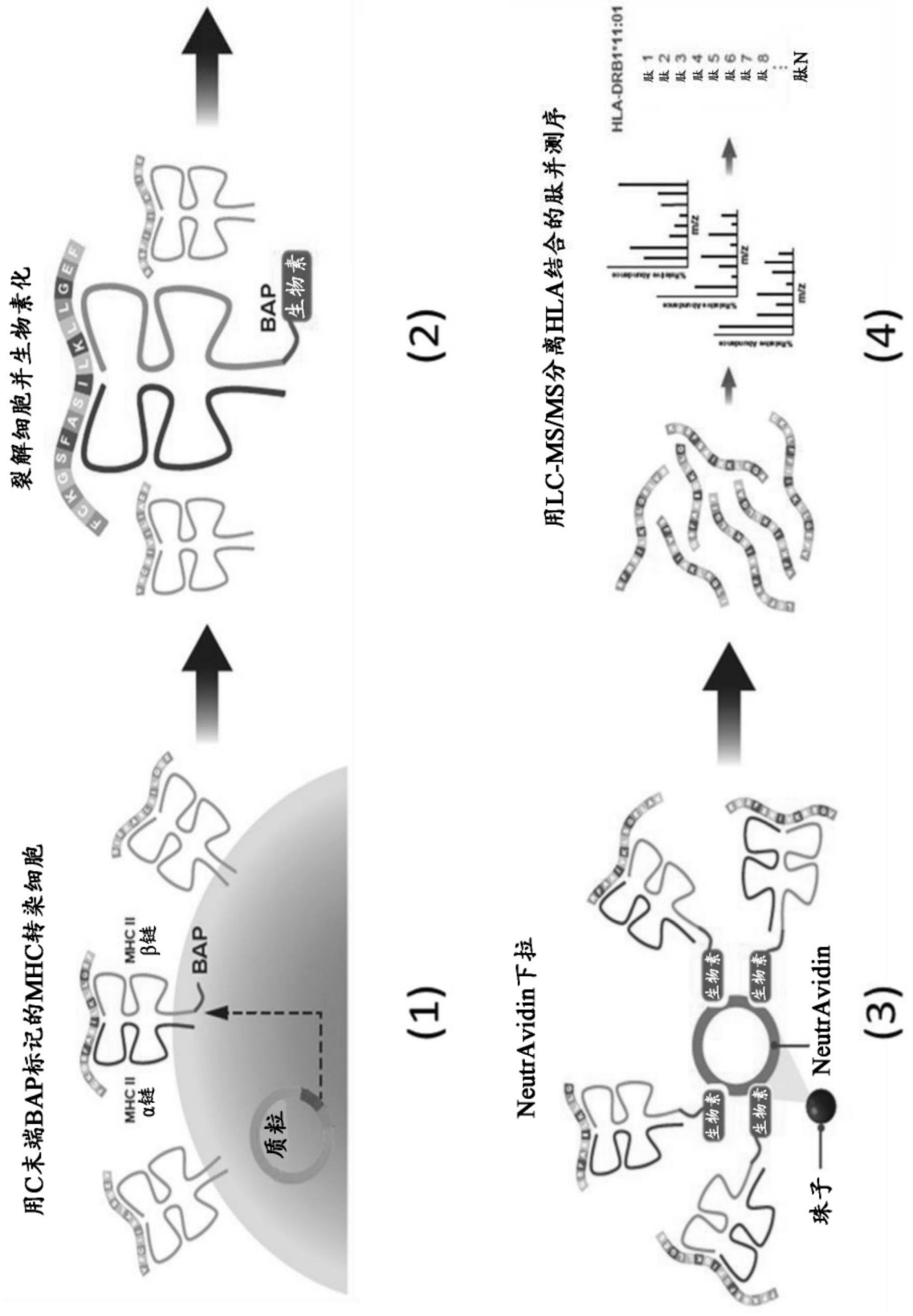

用C末端BAP标记的MHC转染细胞
MHC II α链
MHC II β链
BAP
质粒
(1)
裂解细胞并生物素化
BAP
生物素
(2)
NeutrAvidin下拉
生物素
生物素
生物素
生物素
NeutrAvidin
珠子
(3)
用LC-MS/MS分离HLA结合的肽并测序
HLA-DRB1*11:01
肽 1
肽 2
肽 3
肽 4
肽 5
肽 6
肽 7
肽 8
肽N
m/z
m/z
m/z
(4)

图11A

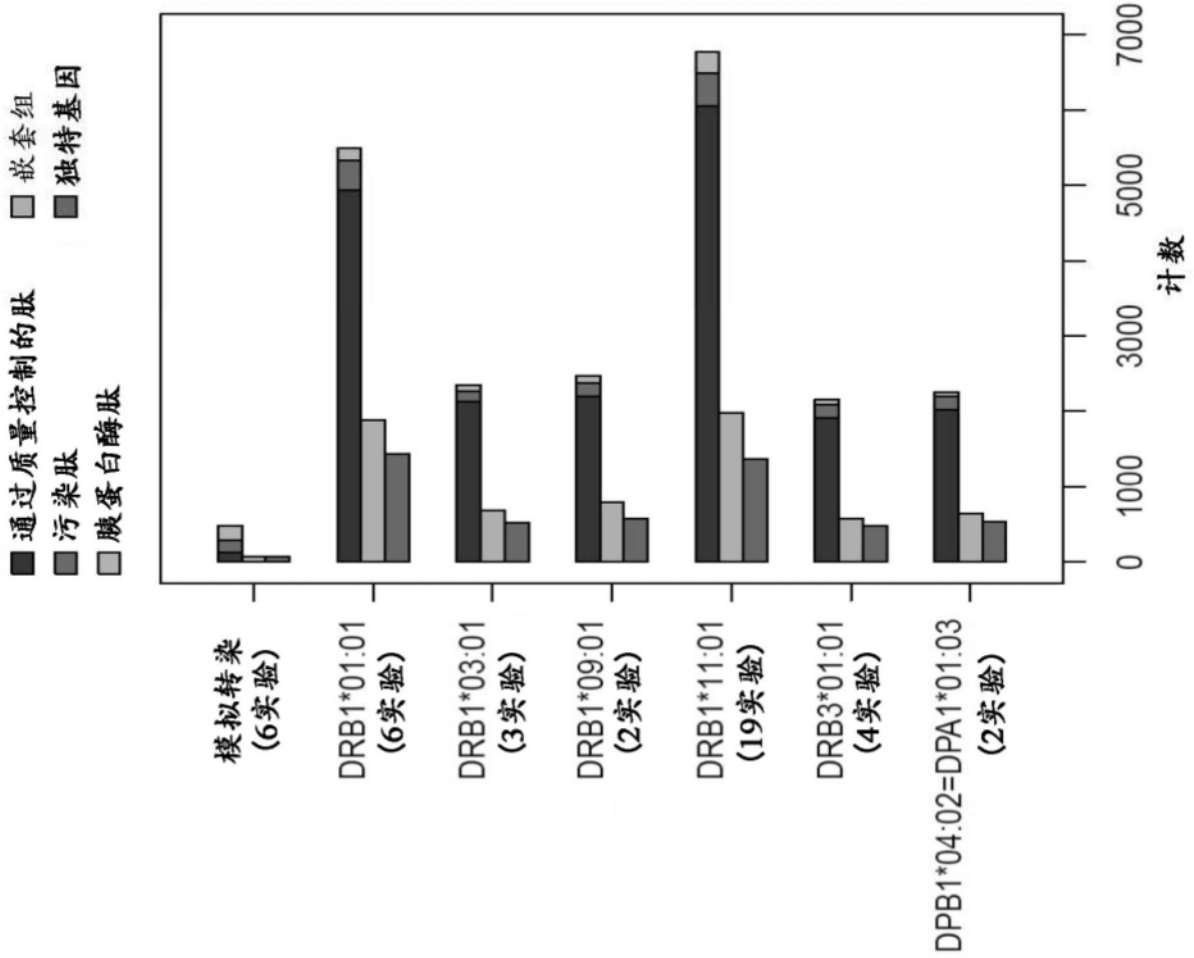

通过质量控制的肽
污染肽
胰蛋白酶肽
嵌套组
独特基因
计数
0
1000
3000
5000
7000
模拟转染
(6实验)
DRB1*01:01
(6实验)
DRB1*03:01
(3实验)
DRB1*09:01
(2实验)
DRB1*11:01
(19实验)
DRB3*01:01
(4实验)
DPB1*04:02=DPA1*01:03
(2实验)

图11B

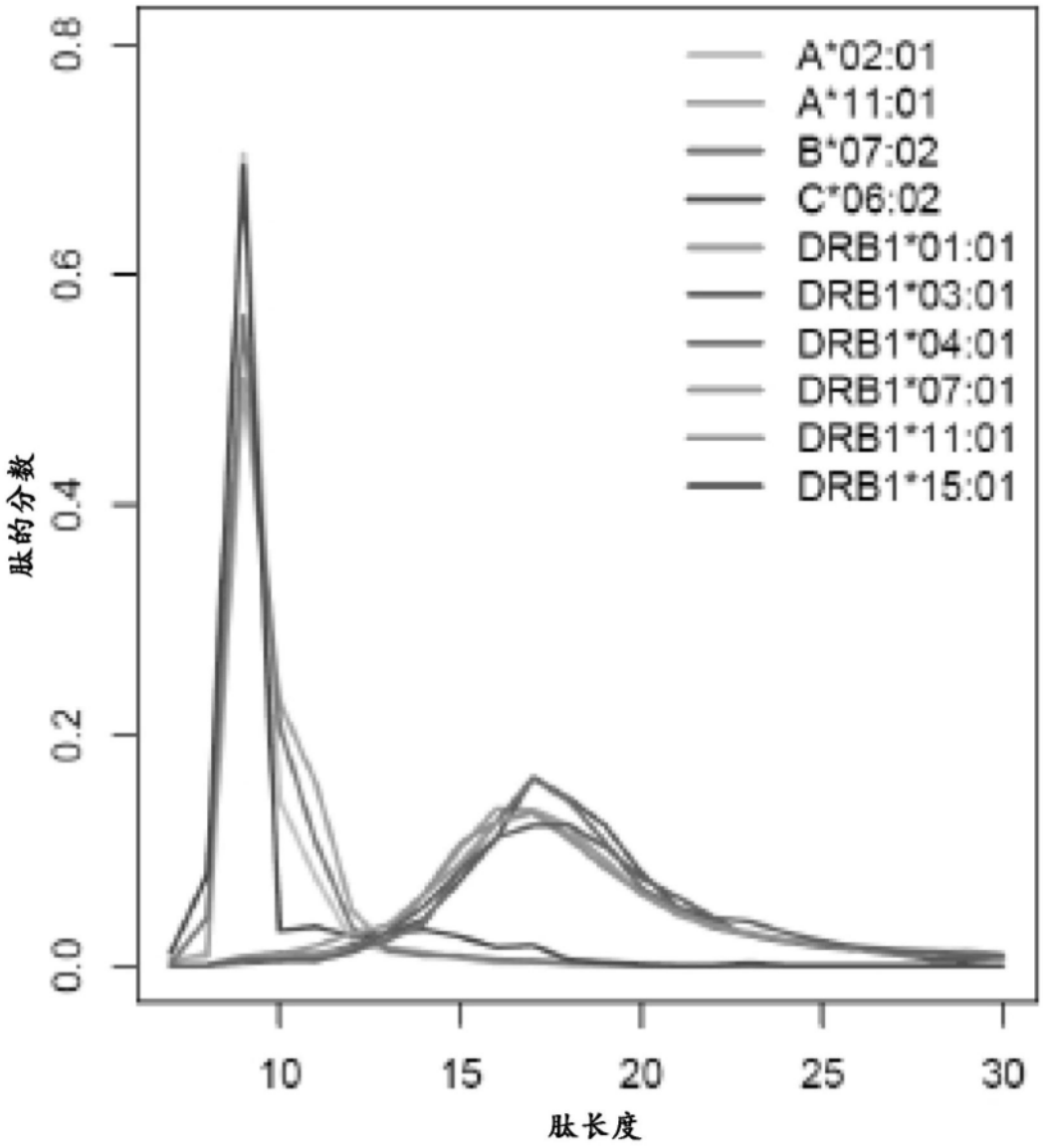

A*02:01
A*11:01
B*07:02
C*06:02
DRB1*01:01
DRB1*03:01
DRB1*04:01
DRB1*07:01
DRB1*11:01
DRB1*15:01
肽的分数
0.0
0.2
0.4
0.6
0.8
10
15
20
25
30
肽长度

图11C

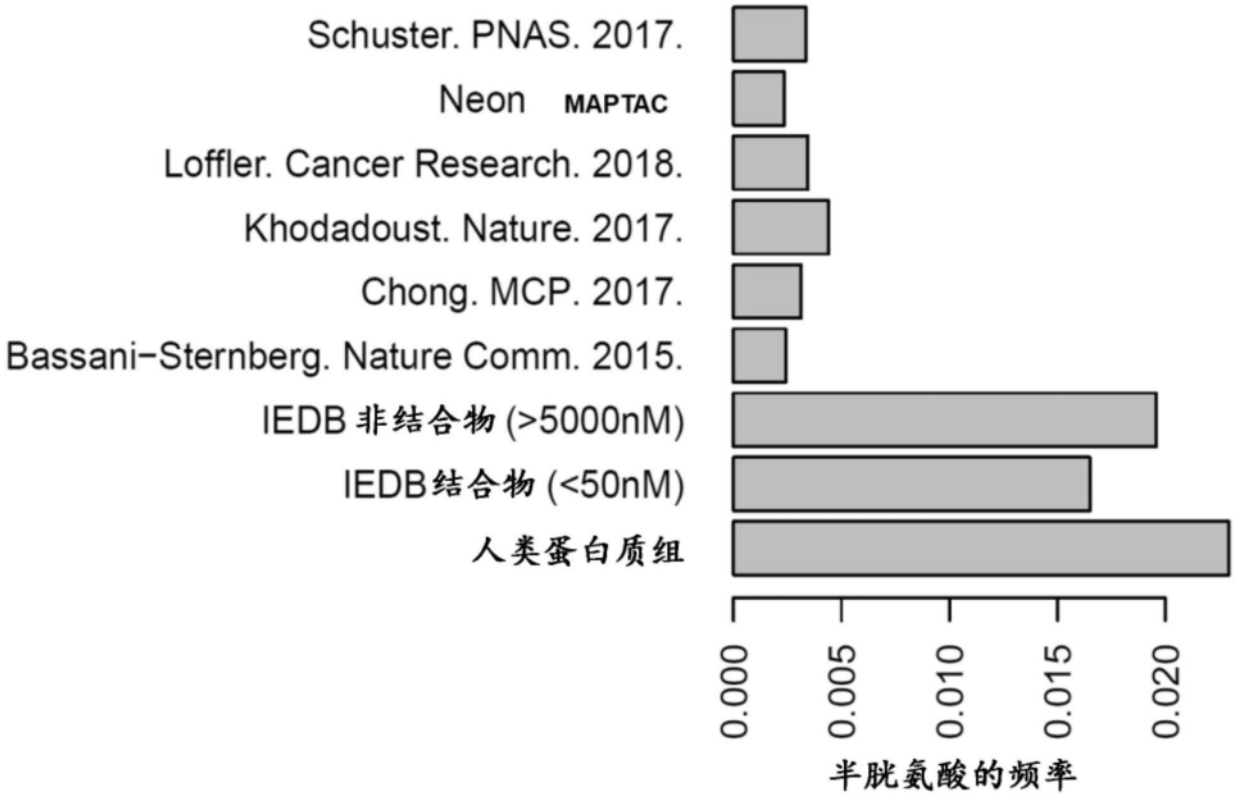

Schuster. PNAS. 2017.
Neon MAPTAC
Loffler. Cancer Research. 2018.
Khodadoust. Nature. 2017.
Chong. MCP. 2017.
Bassani-Sternberg. Nature Comm. 2015.
IEDB 非结合物 (>5000nM)
IEDB结合物 (<50nM)
人类蛋白质组
0.000
0.005
0.010
0.015
0.020
半胱氨酸的频率

图11D

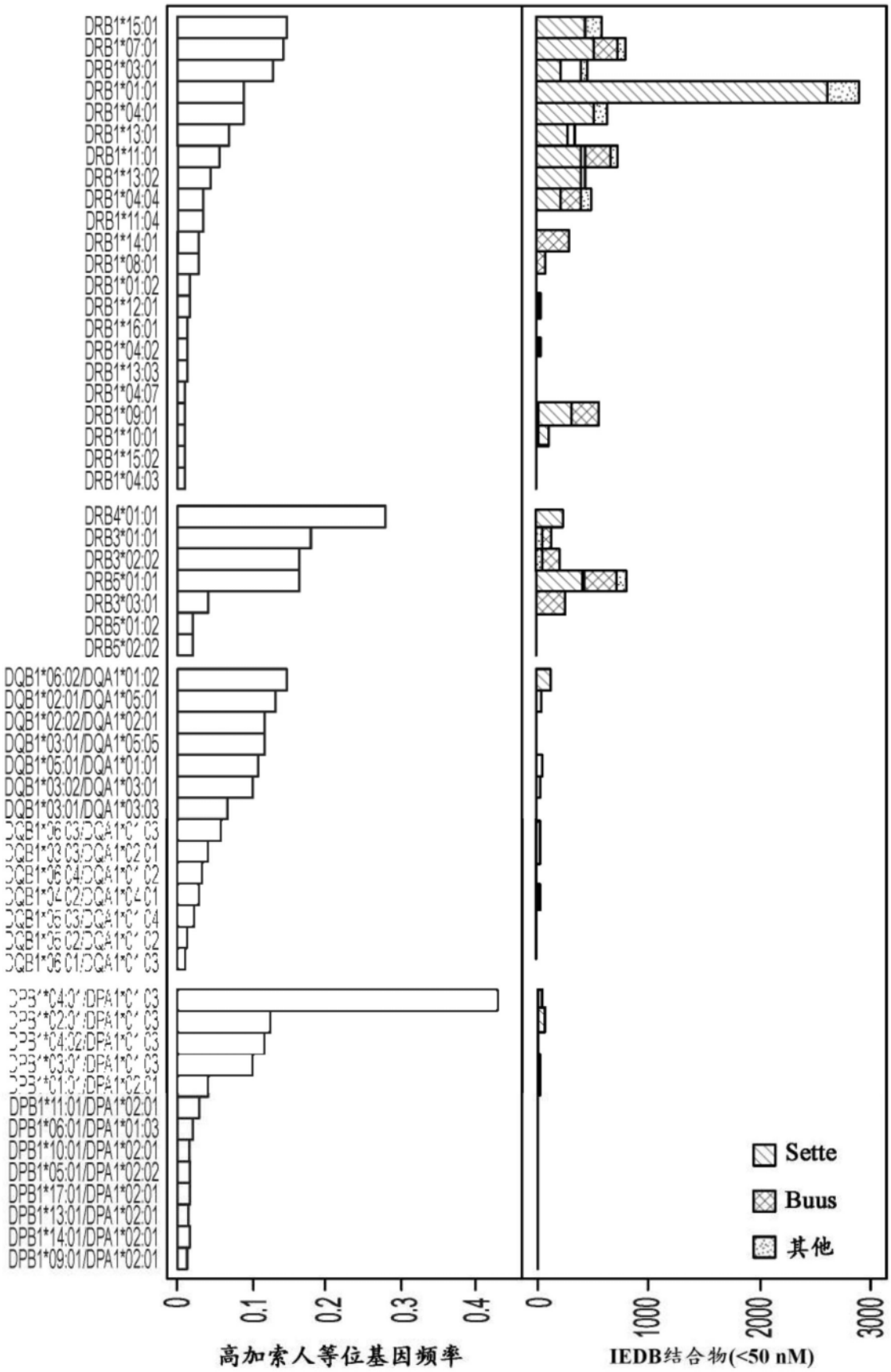

DRB1*15:01
DRB1*07:01
DRB1*03:01
DRB1*01:01
DRB1*04:01
DRB1*13:01
DRB1*11:01
DRB1*13:02
DRB1*04:04
DRB1*11:04
DRB1*14:01
DRB1*08:01
DRB1*01:02
DRB1*12:01
DRB1*16:01
DRB1*04:02
DRB1*13:03
DRB1*04:07
DRB1*09:01
DRB1*10:01
DRB1*15:02
DRB1*04:03
DRB4*01:01
DRB3*01:01
DRB3*02:02
DRB5*01:01
DRB3*03:01
DRB5*01:02
DRB5*02:02
DQB1*06:02/DQA1*01:02
DQB1*02:01/DQA1*05:01
DQB1*02:02/DQA1*02:01
DQB1*03:01/DQA1*05:05
DQB1*05:01/DQA1*01:01
DQB1*03:02/DQA1*03:01
DQB1*03:01/DQA1*03:03
DPB1*11:01/DPA1*02:01
DPB1*06:01/DPA1*01:03
DPB1*10:01/DPA1*02:01
DPB1*05:01/DPA1*02:02
DPB1*17:01/DPA1*02:01
DPB1*13:01/DPA1*02:01
DPB1*14:01/DPA1*02:01
DPB1*09:01/DPA1*02:01
Sette
Buus
其他
0
0.1
0.2
0.3
0.4
高加索人等位基因频率
0
1000
2000
3000
IEDB结合物(<50 nM)

图12A

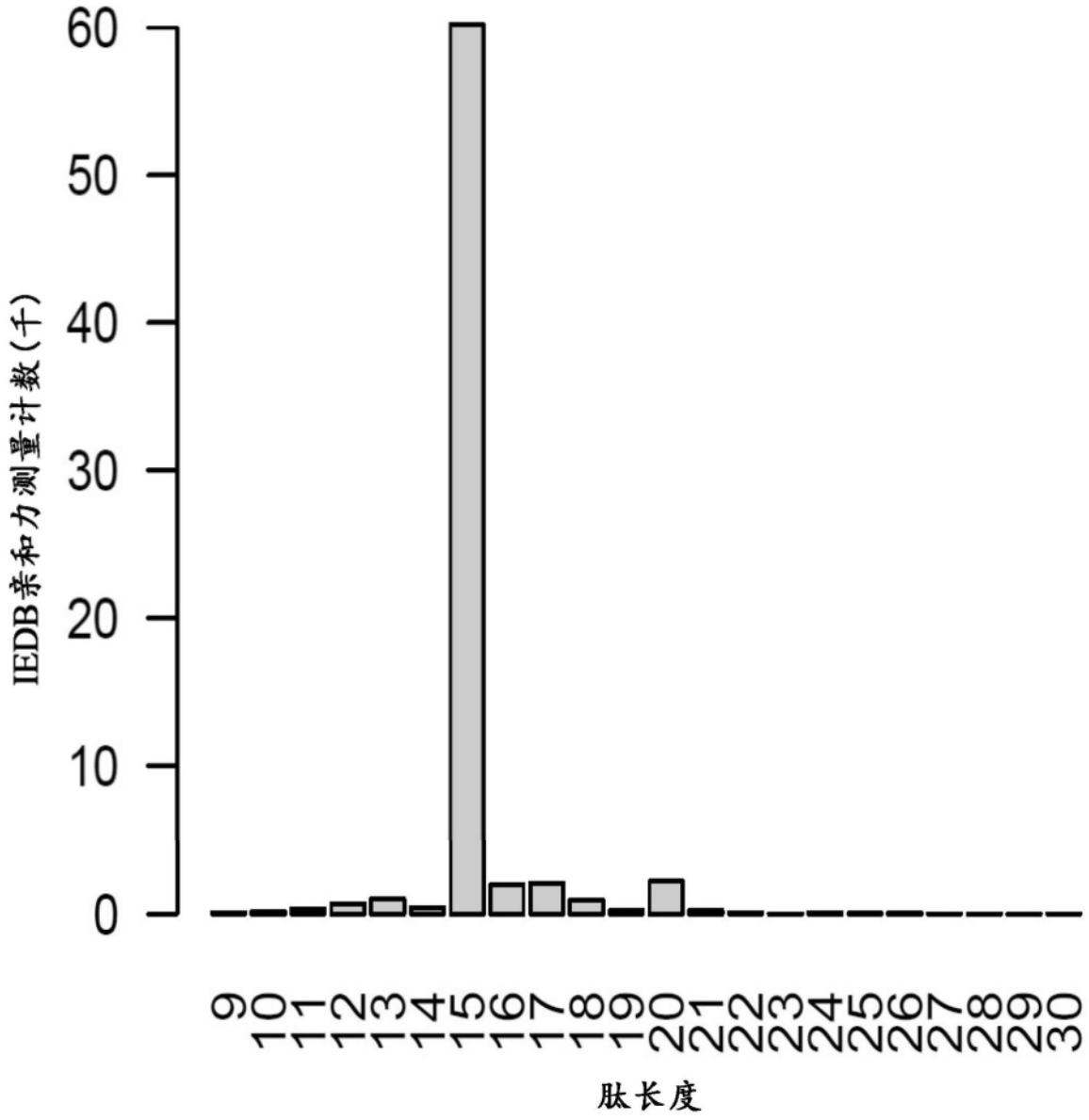

60
50
40
30
20
10
0
IEDB亲和力测量计数(千)
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
肽长度

图12B

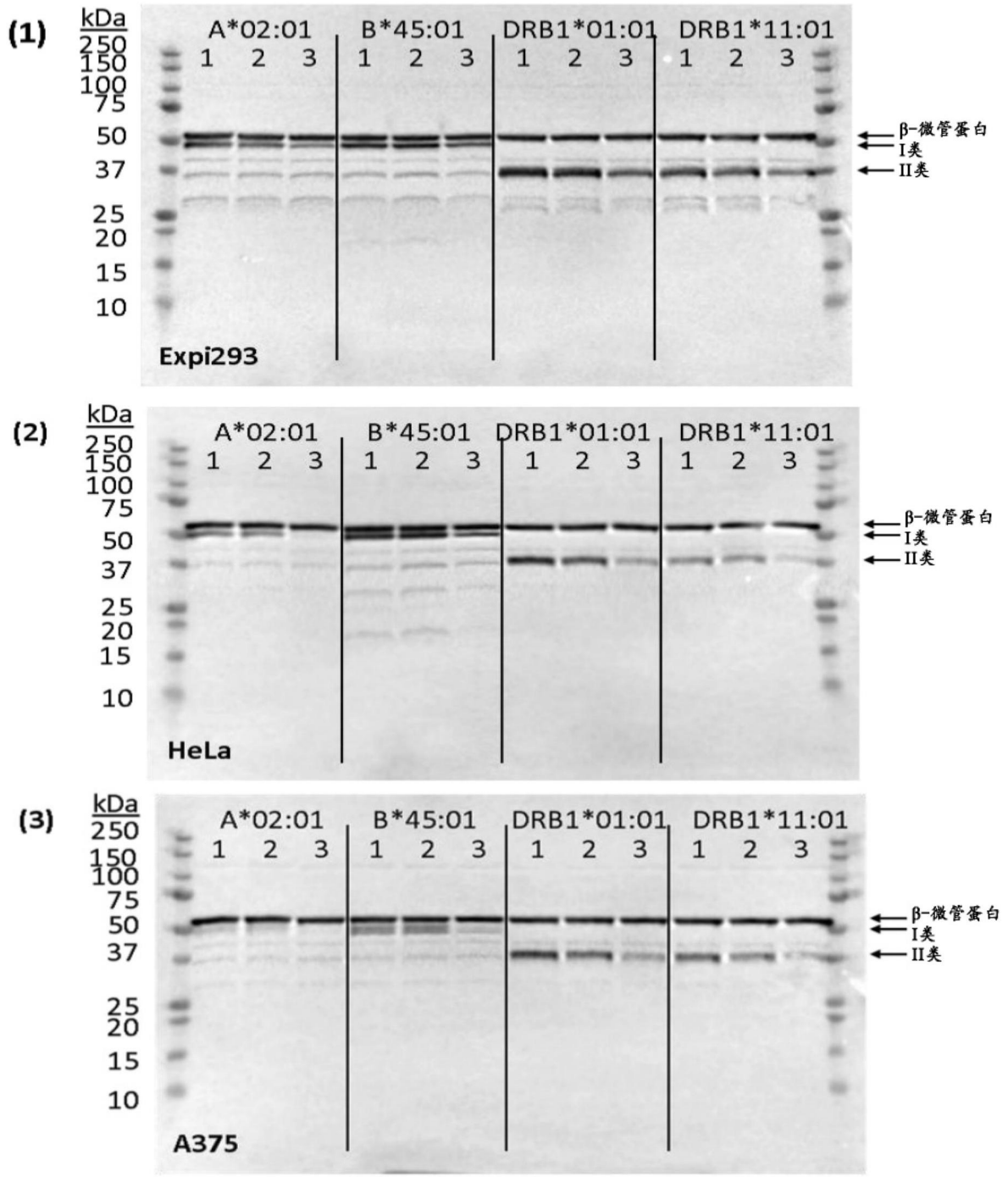

(1)
kDa
250
150
100
75
50
37
25
20
15
10
A*02:01
B*45:01
DRB1*01:01
DRB1*11:01
1 2 3 1 2 3 1 2 3 1 2 3
β-微管蛋白
I类
II类
Expi293
(2)
kDa
250
150
100
75
50
37
25
20
15
10
A*02:01
B*45:01
DRB1*01:01
DRB1*11:01
1 2 3 1 2 3 1 2 3 1 2 3
β-微管蛋白
I类
II类
HeLa
(3)
kDa
250
150
100
75
50
37
25
20
15
10
A*02:01
B*45:01
DRB1*01:01
DRB1*11:01
1 2 3 1 2 3 1 2 3 1 2 3
β-微管蛋白
I类
II类
A375

图12C

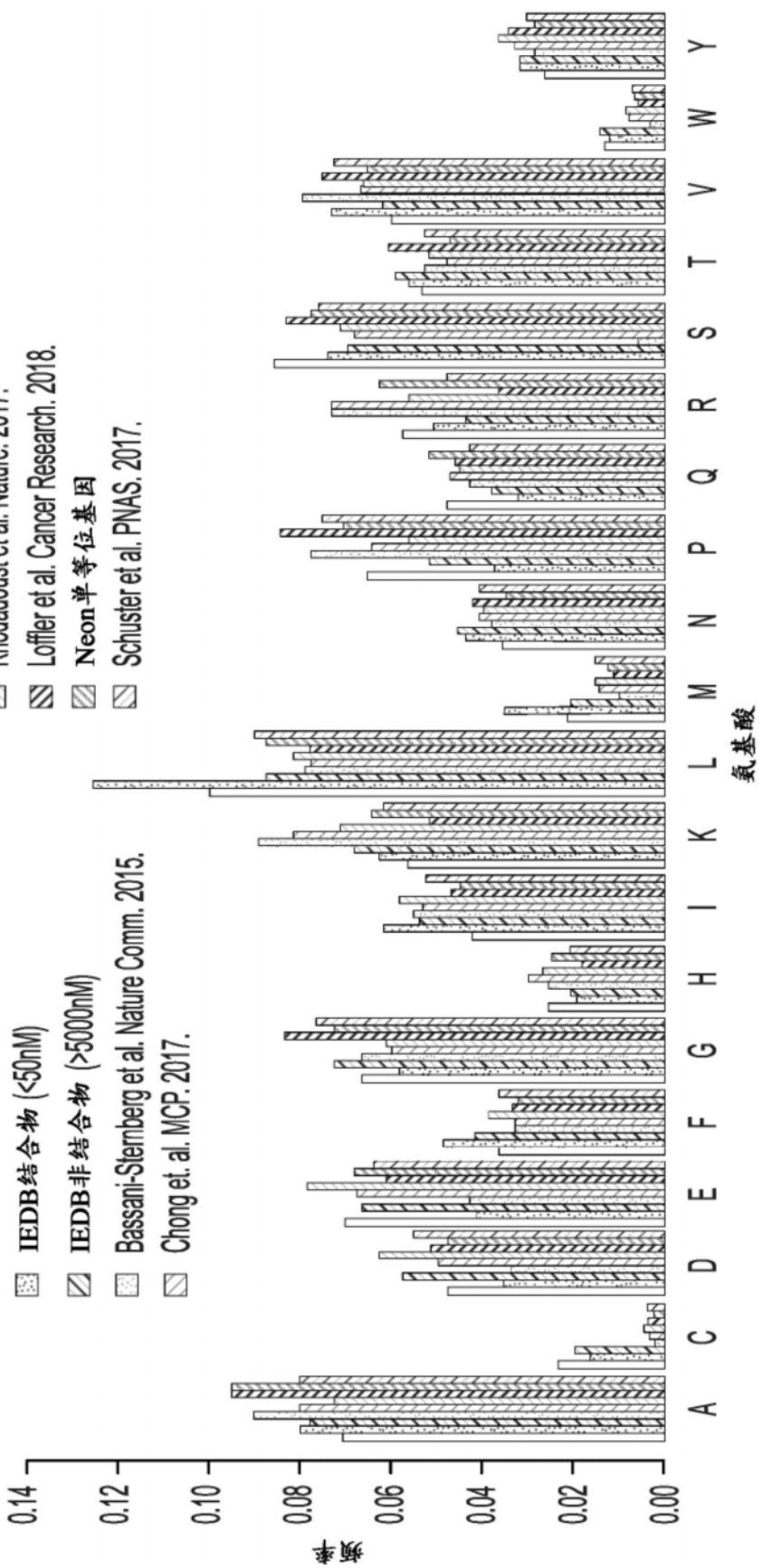

人类蛋白质组
IEDB结合物 (<50nM)
IEDB非结合物 (>5000nM)
Bassani-Sternberg et al. Nature Comm. 2015.
Chong et. al. MCP. 2017.
Khodadoust et al. Nature. 2017.
Loffler et al. Cancer Research. 2018.
Neon单等位基因
Schuster et al. PNAS. 2017.
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
频率
A C D E F G H I K L M N P Q R S T V W Y
氨基酸

图12D

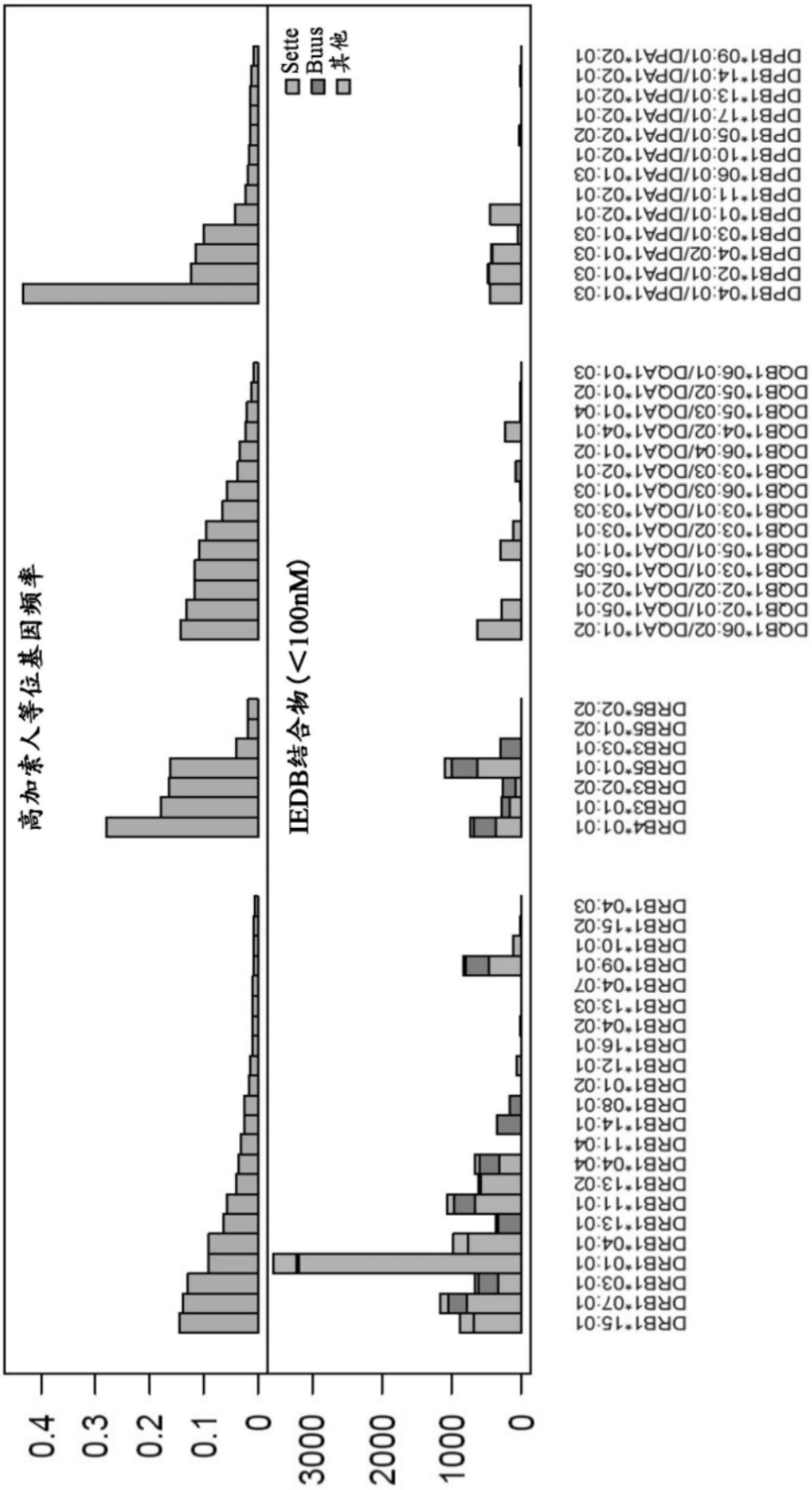

高加索人等位基因频率
IEDB结合物(<100nM)
Sette
Buus
其他
0.4
0.3
0.2
0.1
0
3000
2000
1000
0
DRB1*15:01
DRB1*07:01
DRB1*03:01
DRB1*01:01
DRB1*04:01
DRB1*13:01
DRB1*11:01
DRB1*13:02
DRB1*04:04
DRB1*11:04
DRB1*14:01
DRB1*08:01
DRB1*01:02
DRB1*12:01
DRB1*16:01
DRB1*04:02
DRB1*13:03
DRB1*04:07
DRB1*09:01
DRB1*10:01
DRB1*15:02
DRB1*04:03
DRB4*01:01
DRB3*01:01
DRB3*02:02
DRB5*01:01
DRB3*03:01
DRB5*01:02
DRB5*02:02
DQB1*06:02/DQA1*01:02
DQB1*02:01/DQA1*05:01
DQB1*02:02/DQA1*02:01
DQB1*03:01/DQA1*05:05
DQB1*05:01/DQA1*01:01
DQB1*03:02/DQA1*03:01
DQB1*03:01/DQA1*03:03
DQB1*06:03/DQA1*01:03
DQB1*03:03/DQA1*02:01
DQB1*06:04/DQA1*01:02
DQB1*04:02/DQA1*04:01
DQB1*05:03/DQA1*01:04
DQB1*05:02/DQA1*01:02
DQB1*06:01/DQA1*01:03
DPB1*04:01/DPA1*01:03
DPB1*02:01/DPA1*01:03
DPB1*04:02/DPA1*01:03
DPB1*03:01/DPA1*01:03
DPB1*01:01/DPA1*02:01
DPB1*11:01/DPA1*02:01
DPB1*06:01/DPA1*01:03
DPB1*10:01/DPA1*02:01
DPB1*05:01/DPA1*02:02
DPB1*17:01/DPA1*02:01
DPB1*13:01/DPA1*02:01
DPB1*14:01/DPA1*02:01
DPB1*09:01/DPA1*02:01

图12E

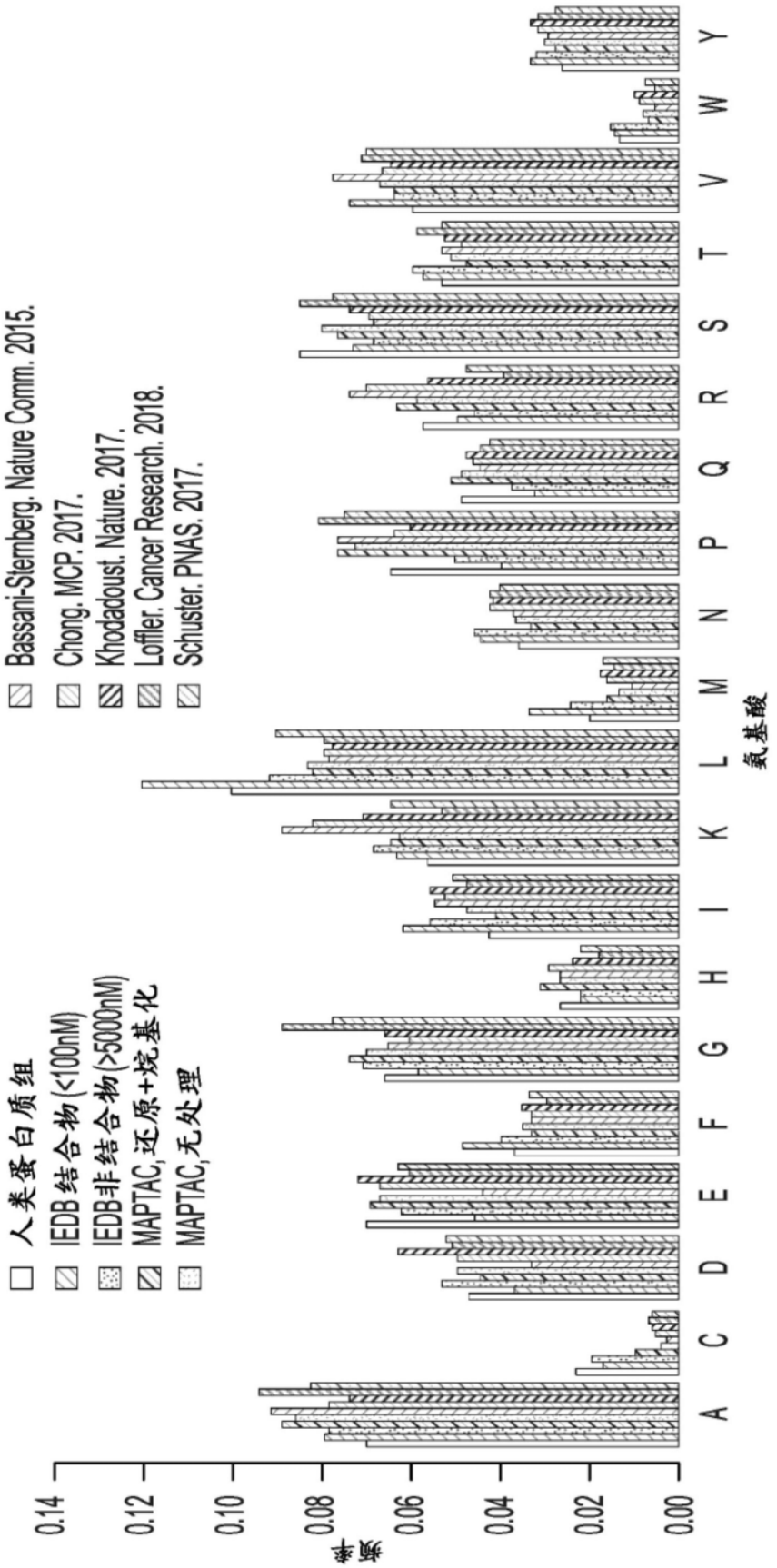

人类蛋白质组
IEDB结合物(<100nM)
IEDB非结合物(>5000nM)
MAPTAC,还原+烷基化
MAPTAC,无处理
Bassani-Sternberg. Nature Comm. 2015.
Chong. MCP. 2017.
Khodadoust. Nature. 2017.
Loffler. Cancer Research. 2018.
Schuster. PNAS. 2017.
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
A C D E F G H I K L M N P Q R S T V W Y
氨基酸

图12F

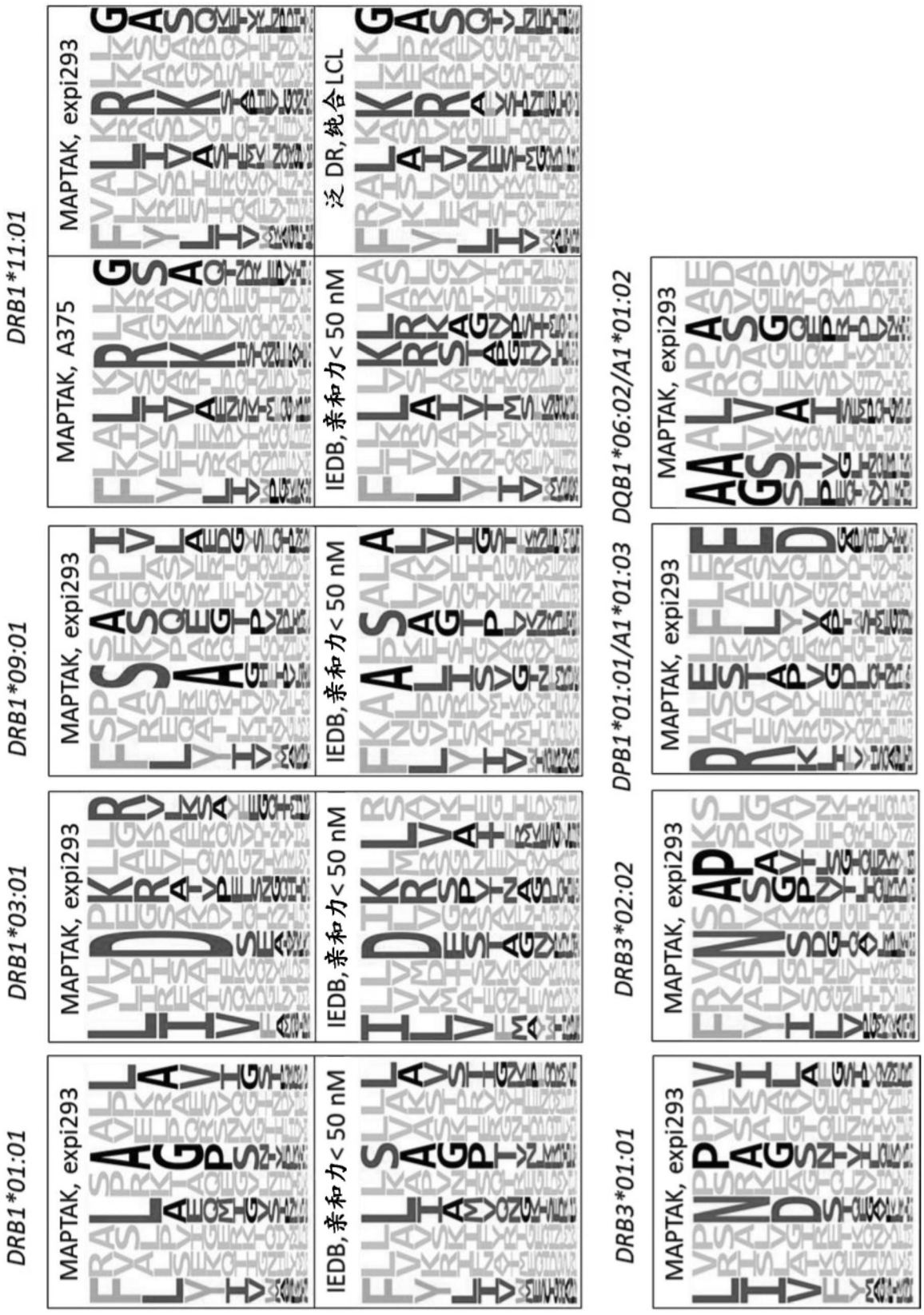

DRB1*01:01
DRB1*03:01
DRB1*09:01
DRB1*11:01
DRB3*01:01
DRB3*02:02
DPB1*01:01/A1*01:03
DQB1*06:02/A1*01:02
MAPTAK, expi293
IEDB,亲和力< 50 nM
MAPTAK, A375
泛 DR,钝合LCL

图13

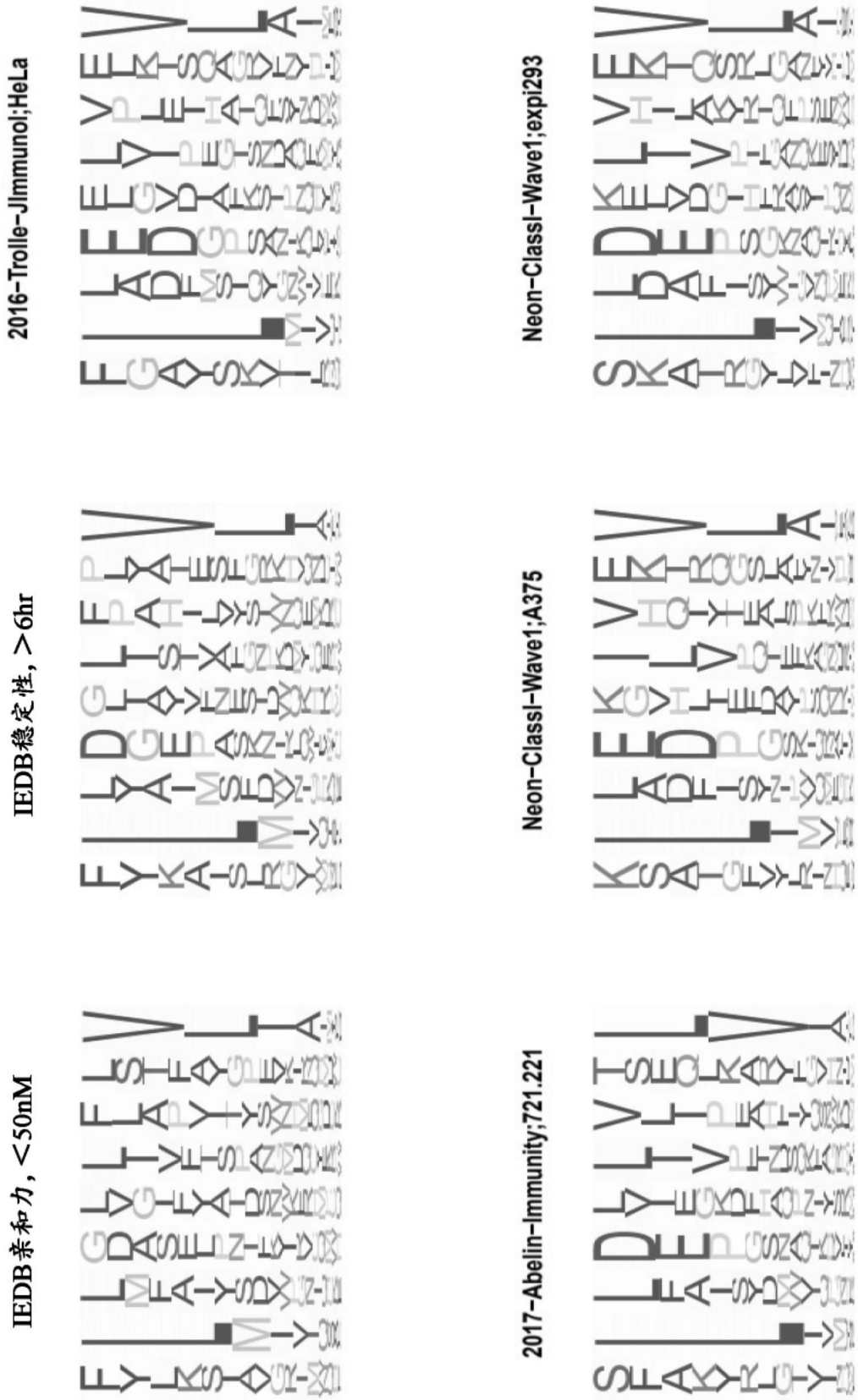

2016-Trolle-JImmunol;HeLa
Neon-ClassI-Wave1;expi293
IEDB稳定性，>6hr
Neon-ClassI-Wave1;A375
IEDB亲和力，<50nM
2017-Abelin-Immunity;721.221

图14A

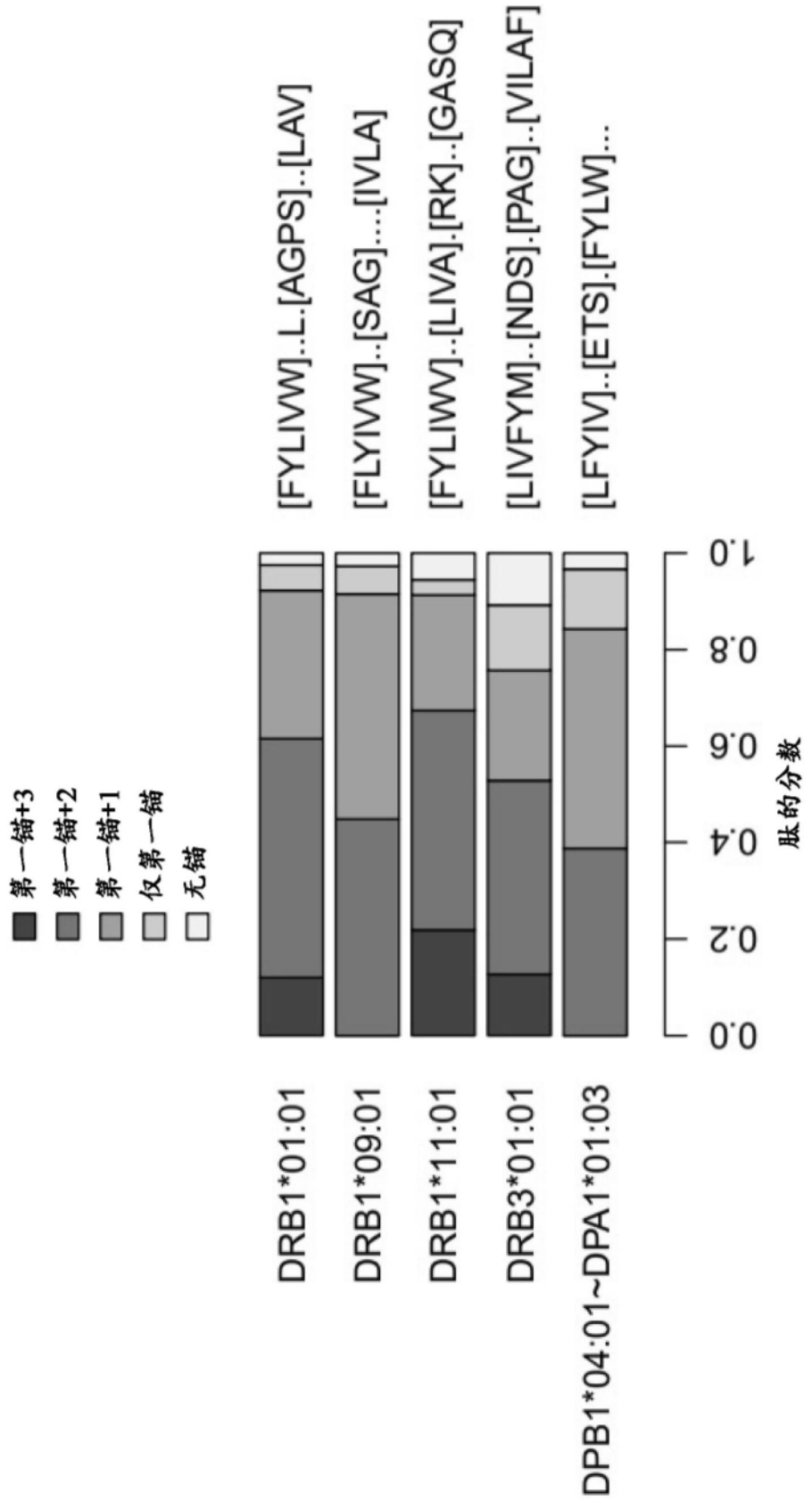

第一锚+3
第一锚+2
第一锚+1
仅第一锚
无锚
[FYLIVW]..L..[AGPS]..[LAV]
[FLYIVW]..[SAG]....[IVLA]
[FYLIWV]..[LIVA].[RK]..[GASQ]
[LIVFYM]..[NDS].[PAG]..[VILAF]
[LFYIV]..[ETS].[FYLW]...
DRB1*01:01
DRB1*09:01
DRB1*11:01
DRB3*01:01
DPB1*04:01~DPA1*01:03
0.0
0.2
0.4
0.6
0.8
1.0
肽的分数

图14B

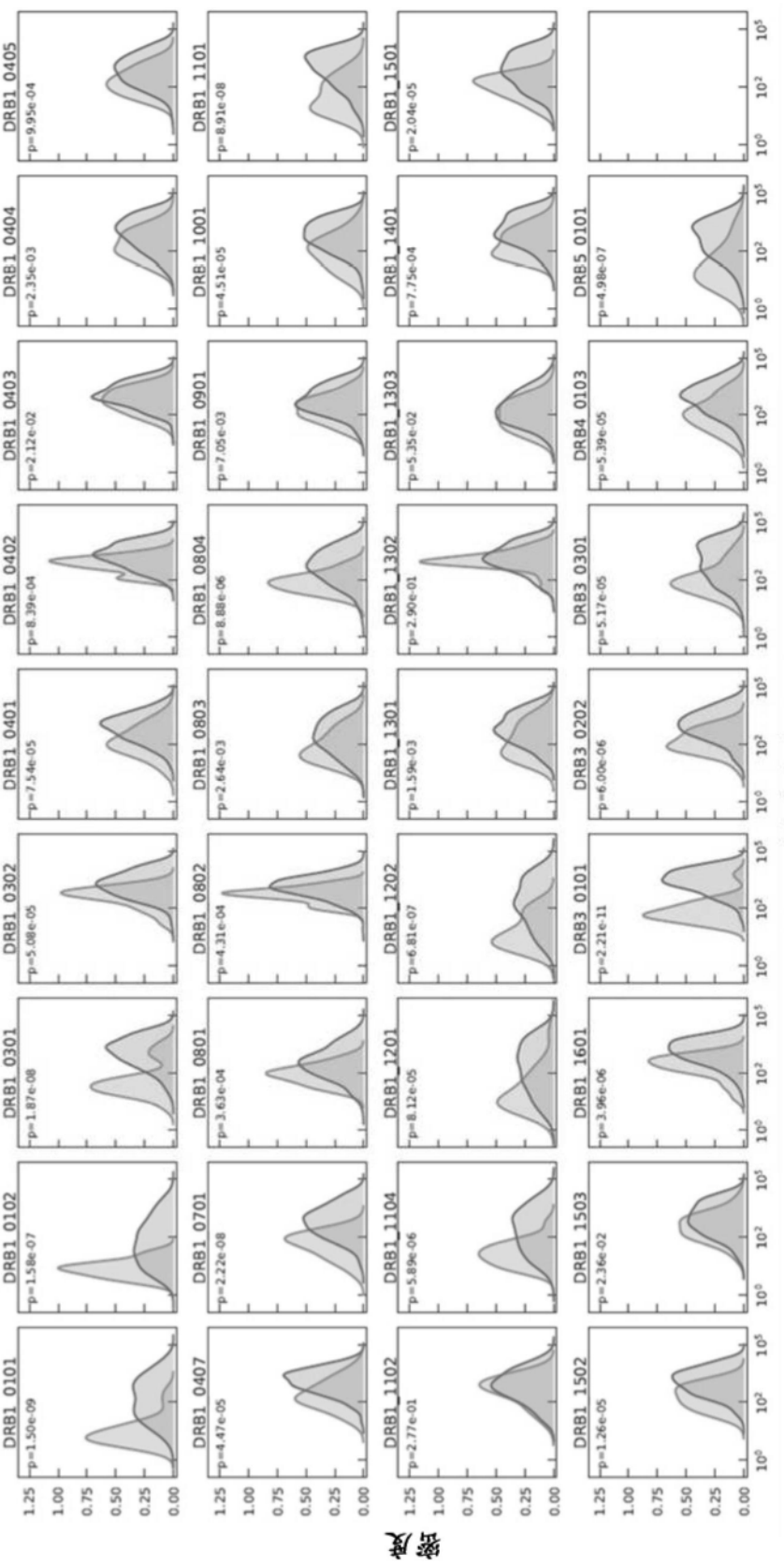

DRB1_0101
DRB1_0102
DRB1_0301
DRB1_0302
DRB1_0401
DRB1_0402
DRB1_0403
DRB1_0404
DRB1_0405
DRB1_0407
DRB1_0701
DRB1_0801
DRB1_0802
DRB1_0803
DRB1_0804
DRB1_0901
DRB1_1001
DRB1_1101
DRB1_1102
DRB1_1104
DRB1_1201
DRB1_1202
DRB1_1301
DRB1_1302
DRB1_1303
DRB1_1401
DRB1_1501
DRB1_1502
DRB1_1503
DRB1_1601
DRB3_0101
DRB3_0202
DRB3_0301
DRB4_0103
DRB5_0101
密度
NetMHCIIpan预测的亲和力（nM）

图14C

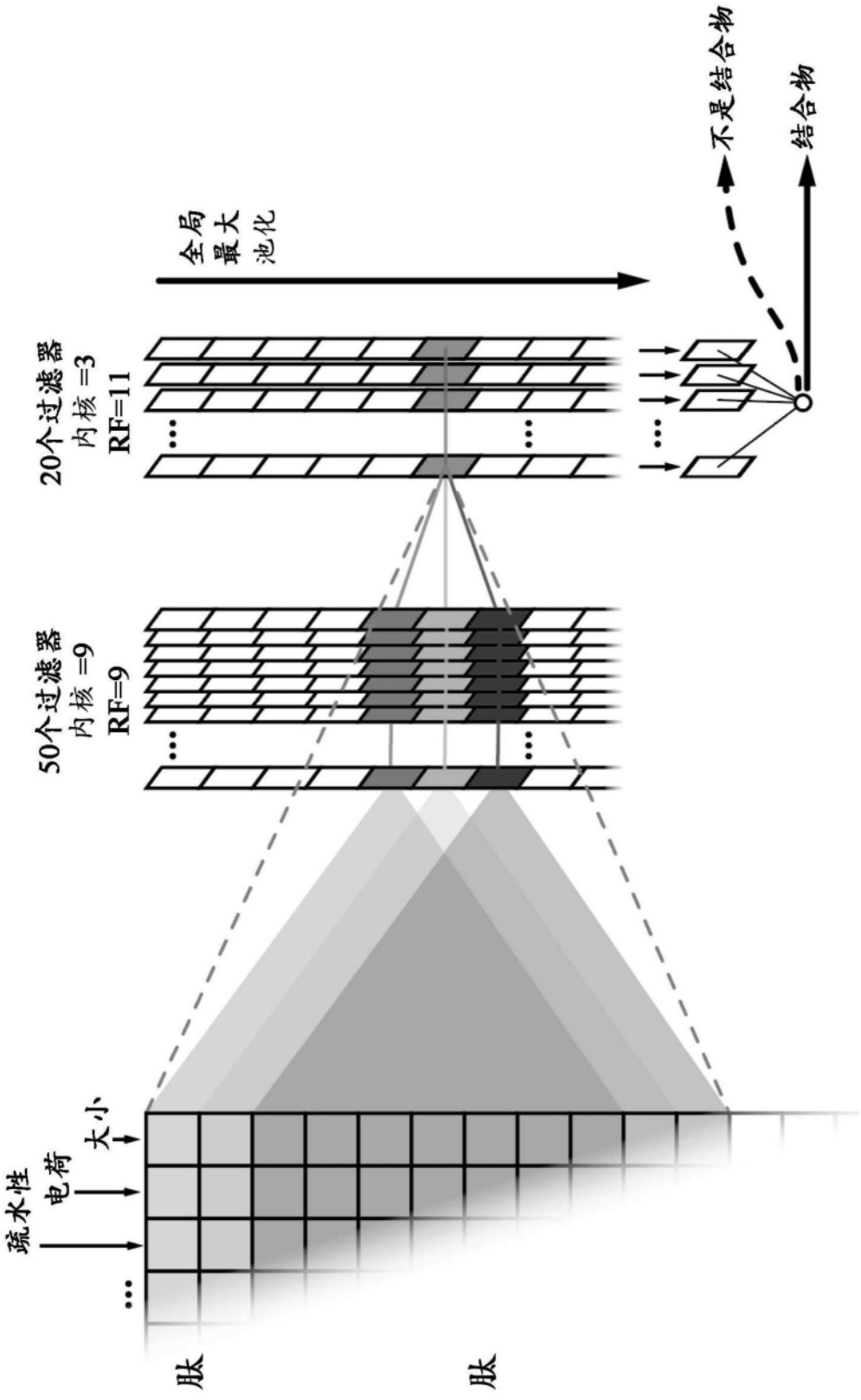


图15A

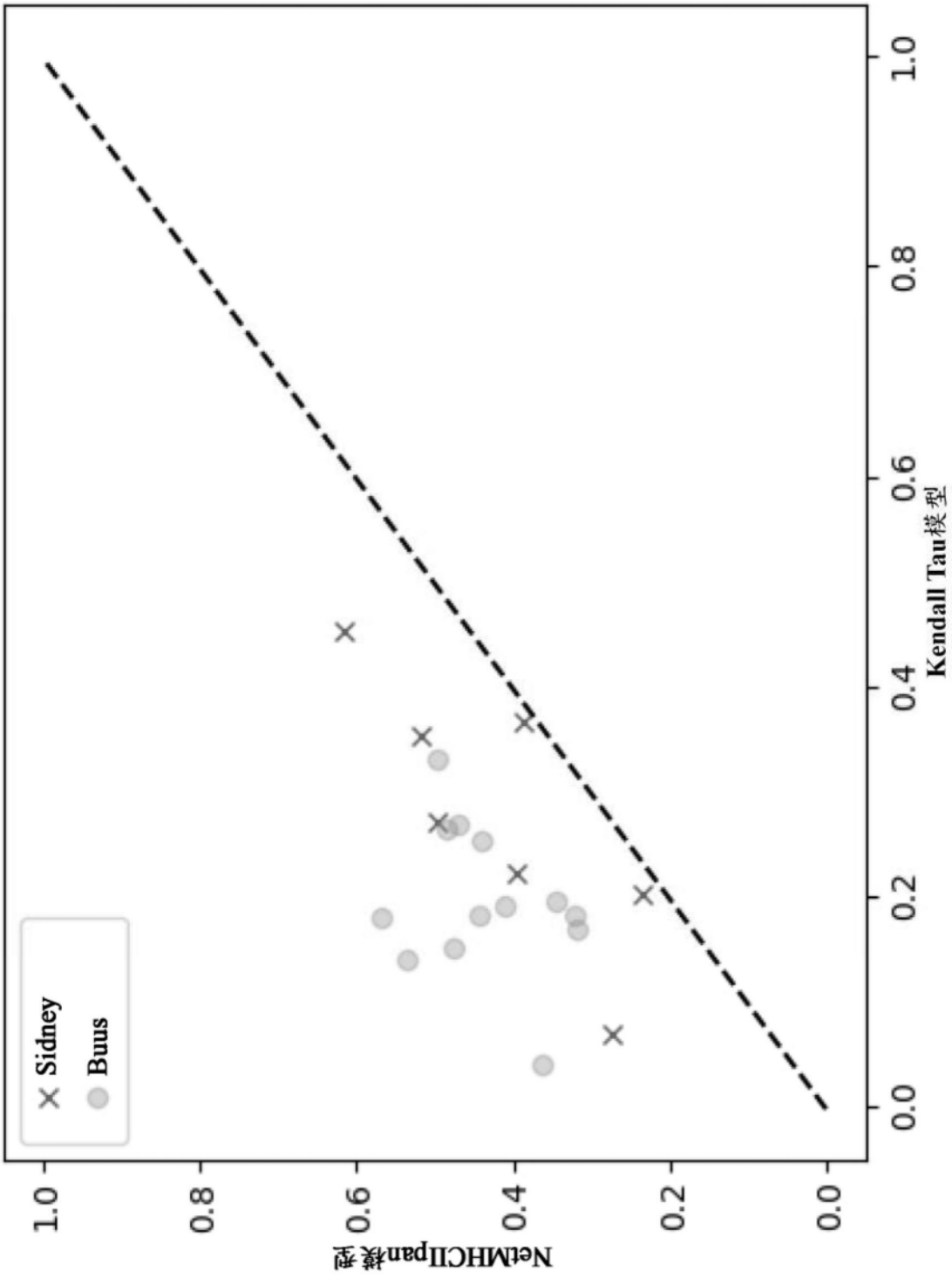


图15B

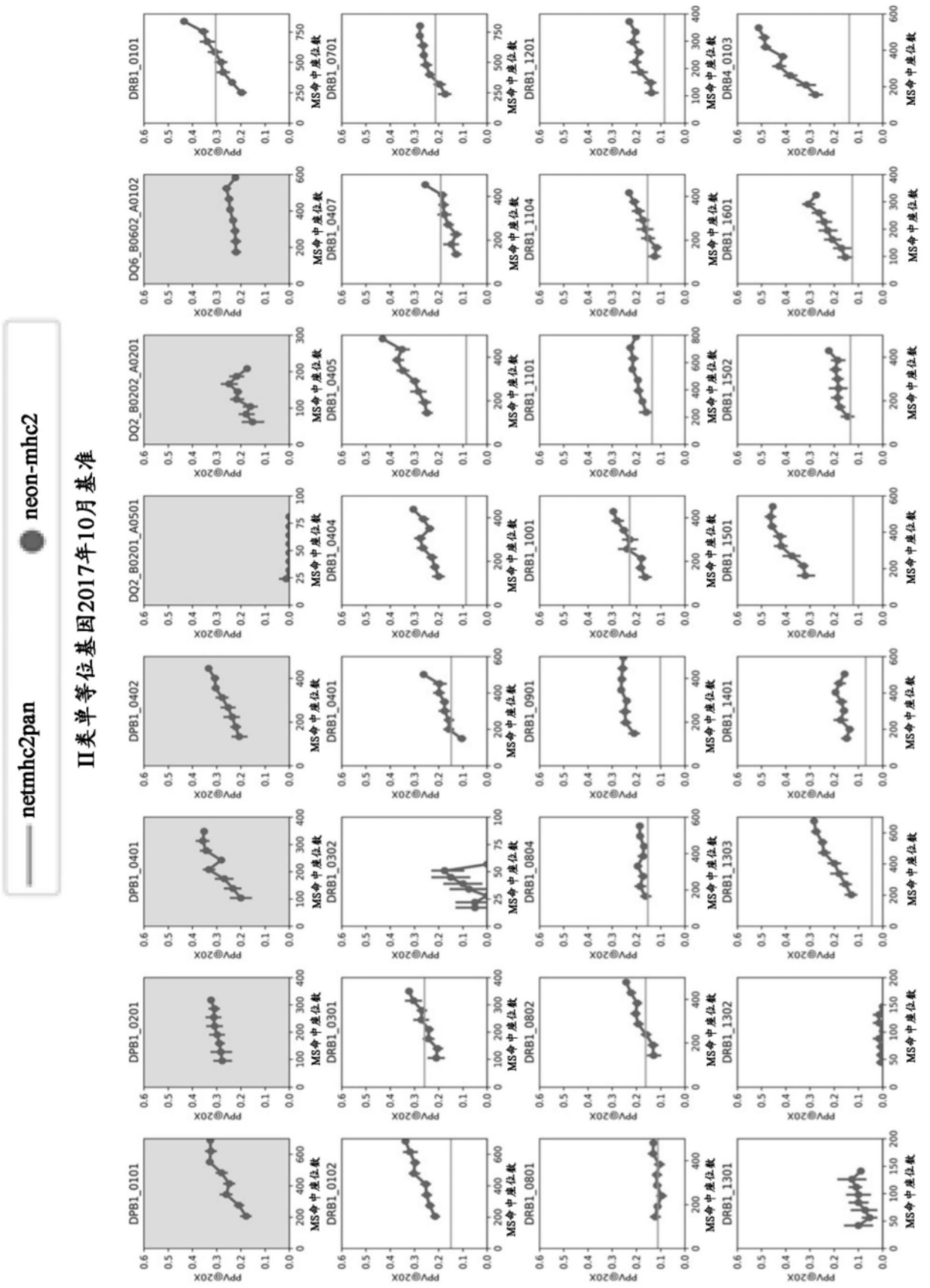

netmhc2pan
neon-mhc2
II类单等位基因2017年10月基准
DPB1_0101
DPB1_0201
DPB1_0401
DPB1_0402
DQ2_B0201_A0501
DQ2_B0202_A0201
DQ6_B0602_A0102
DRB1_0101
DRB1_0102
DRB1_0301
DRB1_0302
DRB1_0401
DRB1_0404
DRB1_0405
DRB1_0407
DRB1_0701
DRB1_0801
DRB1_0802
DRB1_0804
DRB1_0901
DRB1_1001
DRB1_1101
DRB1_1104
DRB1_1201
DRB1_1301
DRB1_1302
DRB1_1303
DRB1_1401
DRB1_1501
DRB1_1502
DRB1_1601
DRB4_0103
PPV@20X
MS命中基因数

图16

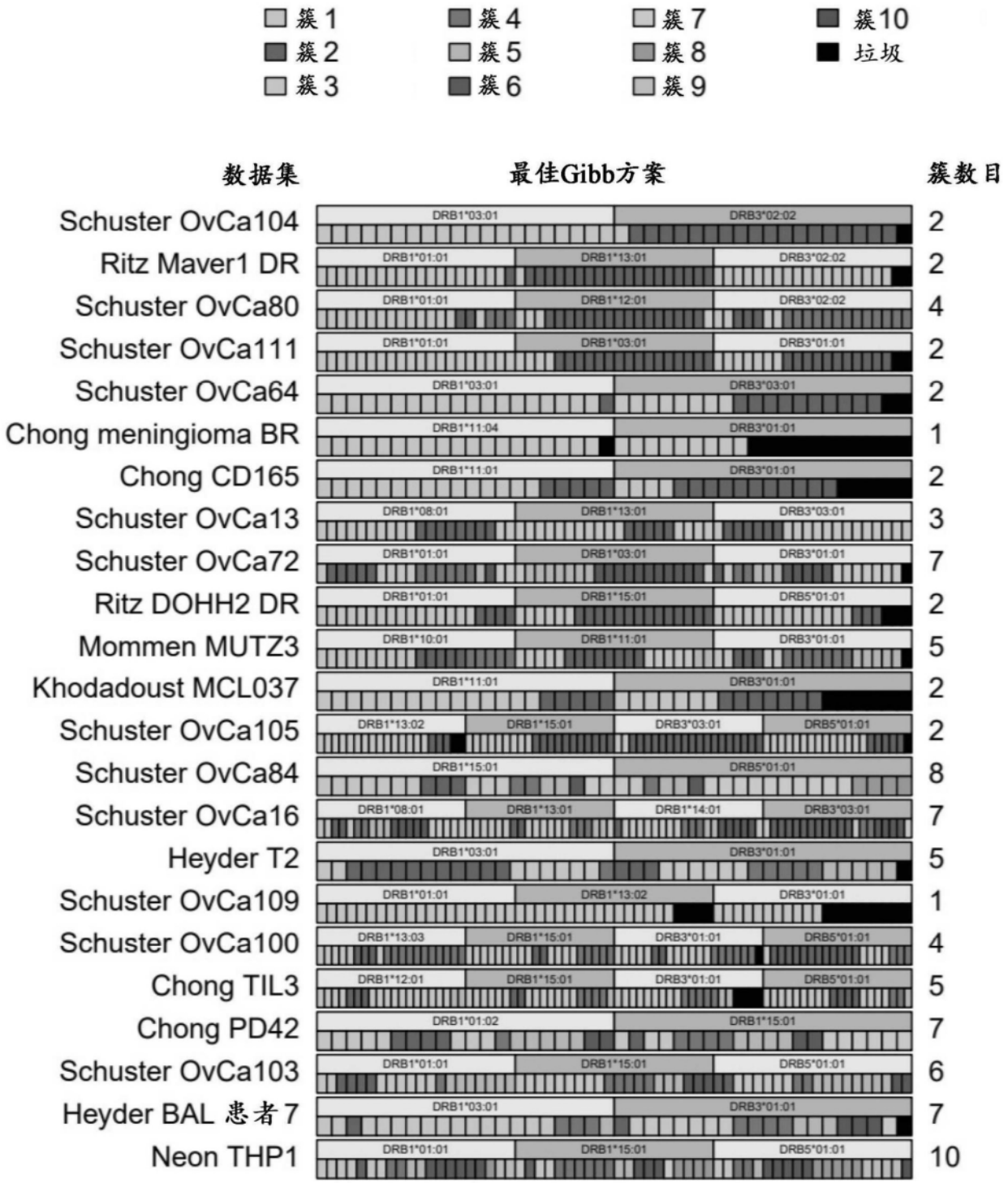

簇1
簇2
簇3
簇4
簇5
簇6
簇7
簇8
簇9
簇10
垃圾
数据集
最佳Gibb方案
簇数目
Schuster OvCa104 DRB1*03:01 DRB3*02:02 2
Ritz Maver1 DR DRB1*01:01 DRB1*13:01 DRB3*02:02 2
Schuster OvCa80 DRB1*01:01 DRB1*12:01 DRB3*02:02 4
Schuster OvCa111 DRB1*01:01 DRB1*03:01 DRB3*01:01 2
Schuster OvCa64 DRB1*03:01 DRB3*03:01 2
Chong meningioma BR DRB1*11:04 DRB3*01:01 1
Chong CD165 DRB1*11:01 DRB3*01:01 2
Schuster OvCa13 DRB1*08:01 DRB1*13:01 DRB3*03:01 3
Schuster OvCa72 DRB1*01:01 DRB1*03:01 DRB3*01:01 7
Ritz DOHH2 DR DRB1*01:01 DRB1*15:01 DRB5*01:01 2
Mommen MUTZ3 DRB1*10:01 DRB1*11:01 DRB3*01:01 5
Khodadoust MCL037 DRB1*11:01 DRB3*01:01 2
Schuster OvCa105 DRB1*13:02 DRB1*15:01 DRB3*03:01 DRB5*01:01 2
Schuster OvCa84 DRB1*15:01 DRB5*01:01 8
Schuster OvCa16 DRB1*08:01 DRB1*13:01 DRB1*14:01 DRB3*03:01 7
Heyder T2 DRB1*03:01 DRB3*01:01 5
Schuster OvCa109 DRB1*01:01 DRB1*13:02 DRB3*01:01 1
Schuster OvCa100 DRB1*13:03 DRB1*15:01 DRB3*01:01 DRB5*01:01 4
Chong TIL3 DRB1*12:01 DRB1*15:01 DRB3*01:01 DRB5*01:01 5
Chong PD42 DRB1*01:02 DRB1*15:01 7
Schuster OvCa103 DRB1*01:01 DRB1*15:01 DRB5*01:01 6
Heyder BAL 患者7 DRB1*03:01 DRB3*01:01 7
Neon THP1 DRB1*01:01 DRB1*15:01 DRB5*01:01 10

图17A

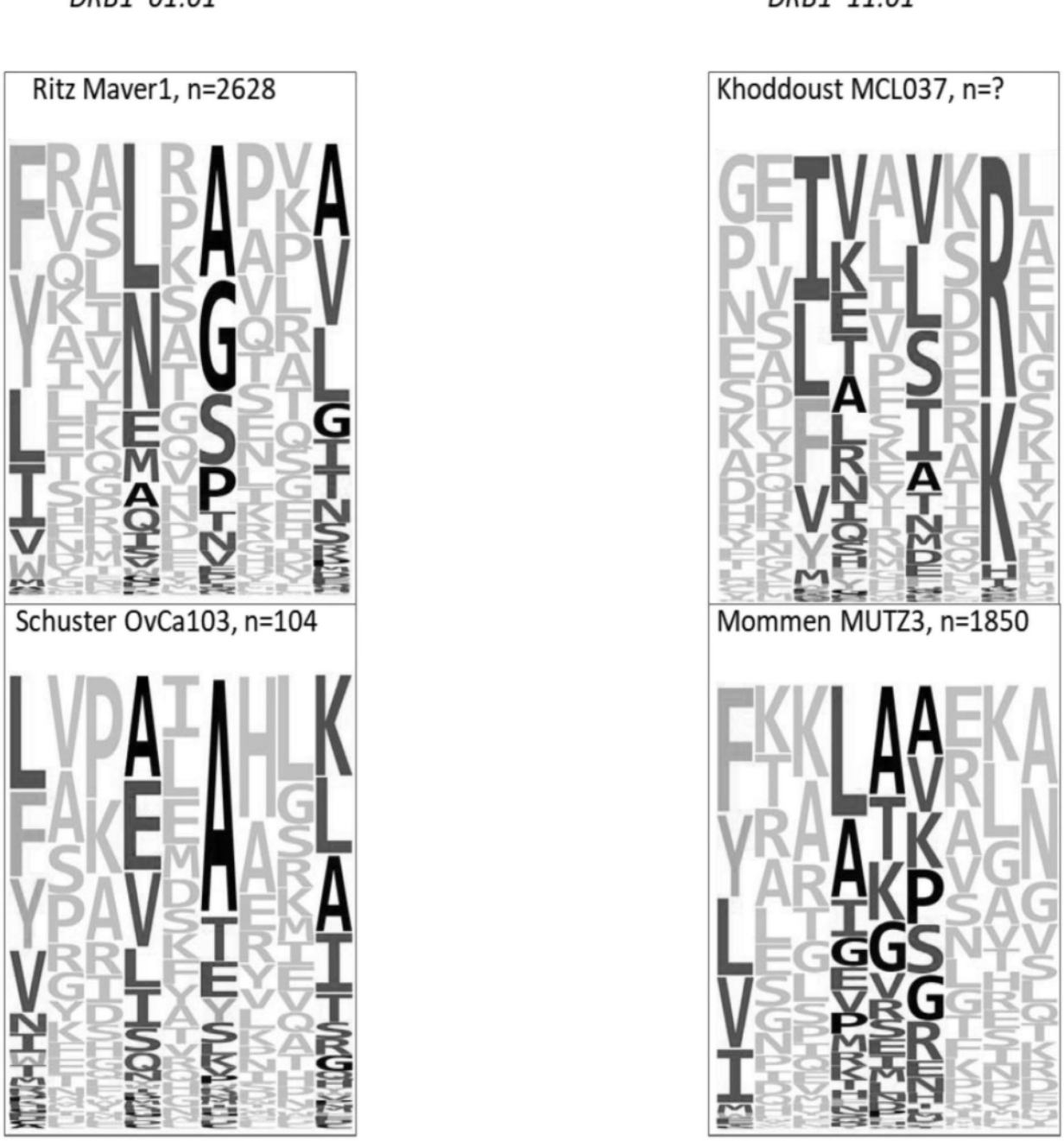

DRB1*01:01
DRB1*11:01
Ritz Maver1, n=2628
Khoddoust MCL037, n=?
Schuster OvCa103, n=104
Mommen MUTZ3, n=1850

图17B

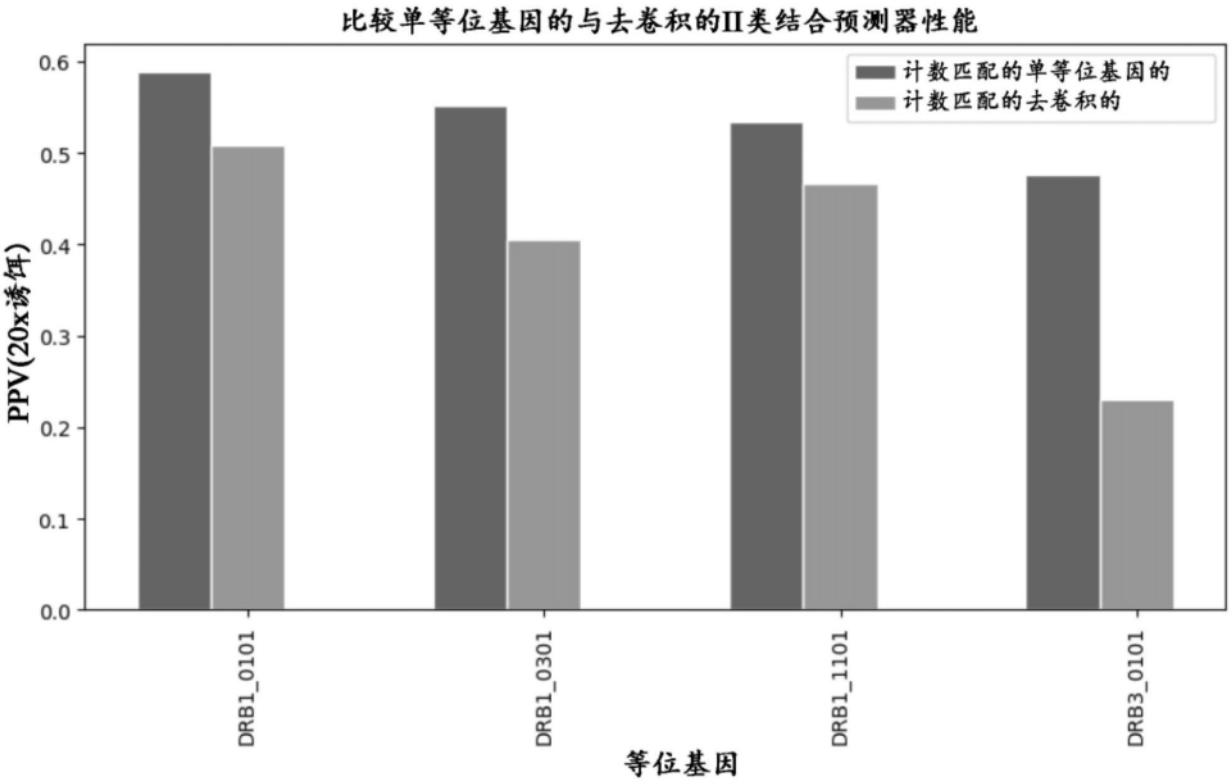

比较单等位基因的与去卷积的II类结合预测器性能
计数匹配的单等位基因的
计数匹配的去卷积的
PPV(20x诱饵)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
DRB1_0101
DRB1_0301
DRB1_1101
DRB3_0101
等位基因

图17C

*DRB1*01:01* *DRB1*01:01* *DRB1*11:01*

（续）

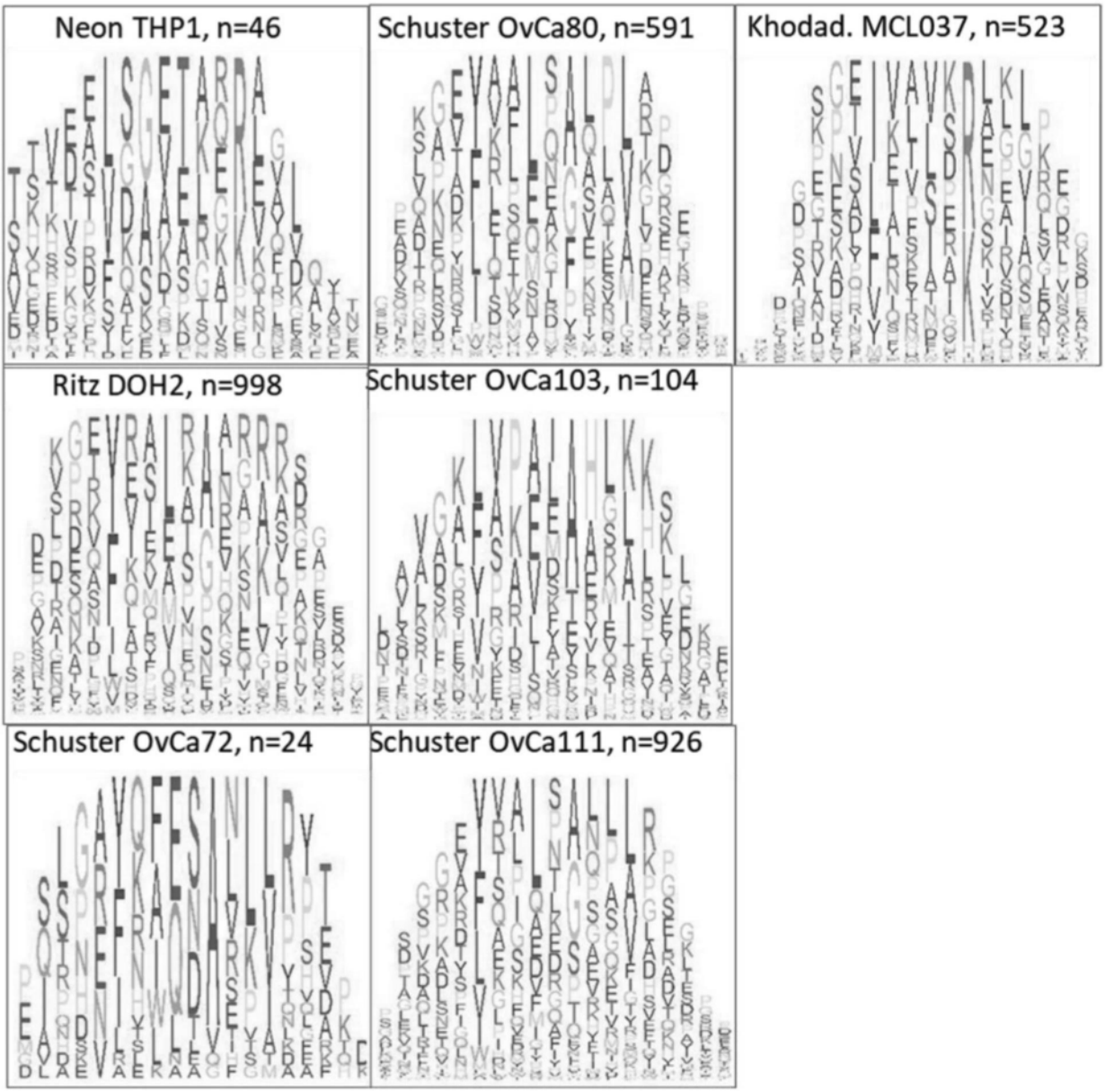


图17D

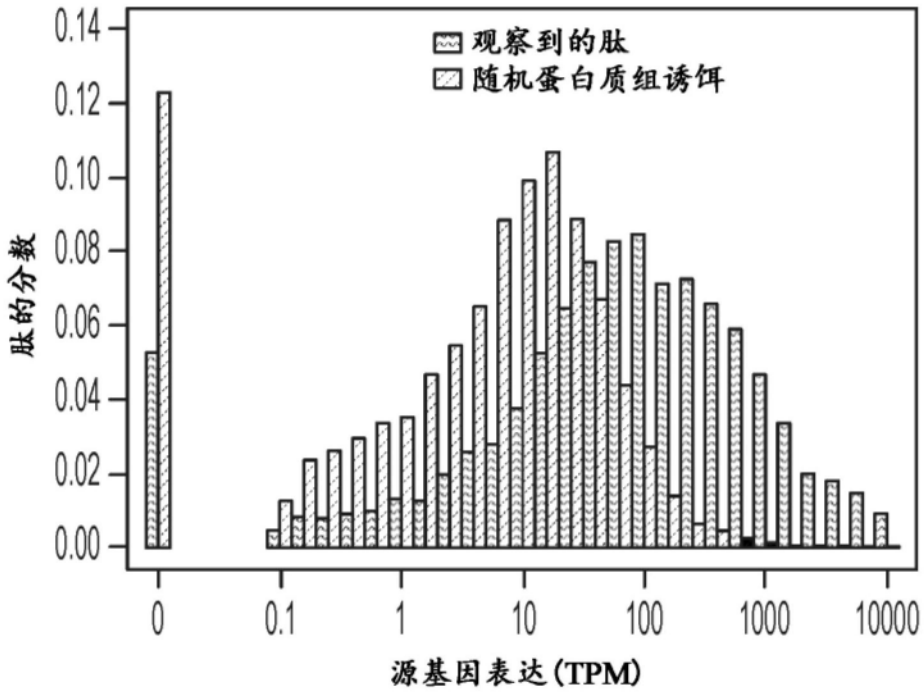

观察到的肽
随机蛋白质组诱饵
肽的分数
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
0.1
1
10
100
1000
10000
源基因表达(TPM)

图18A

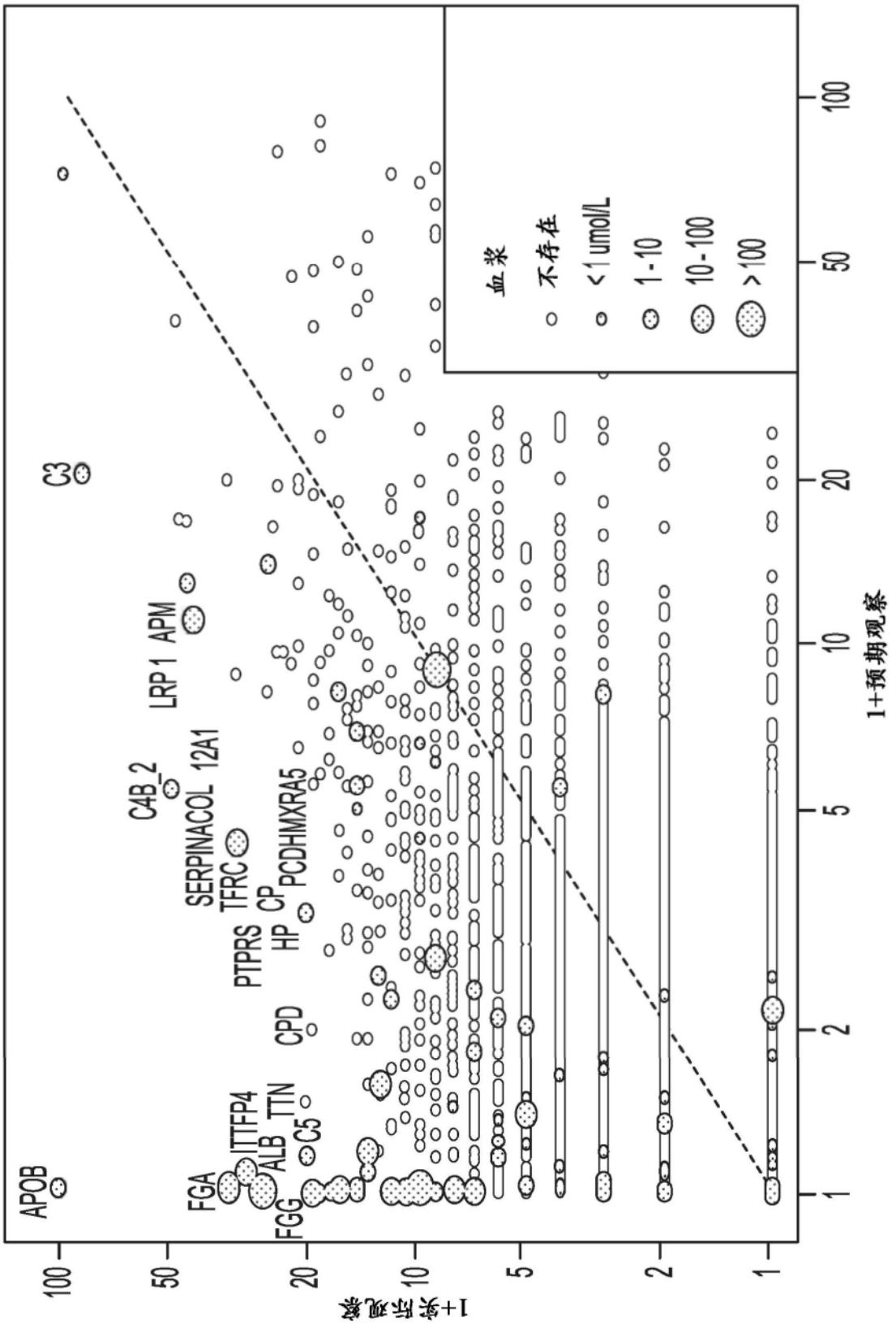

血浆
不存在
<1 umol/L
1-10
10-100
>100
1+预期观察
1+实际观察
1
2
5
10
20
50
100
C3
LRP1
APM
C4B_2
12A1
SERPINACOL
TFRC
PCDHMXRA5
CP
HP
PTPRS
CPD
ITTFP4
ALB
TTN
C5
FGA
FGG
APOB

图18B

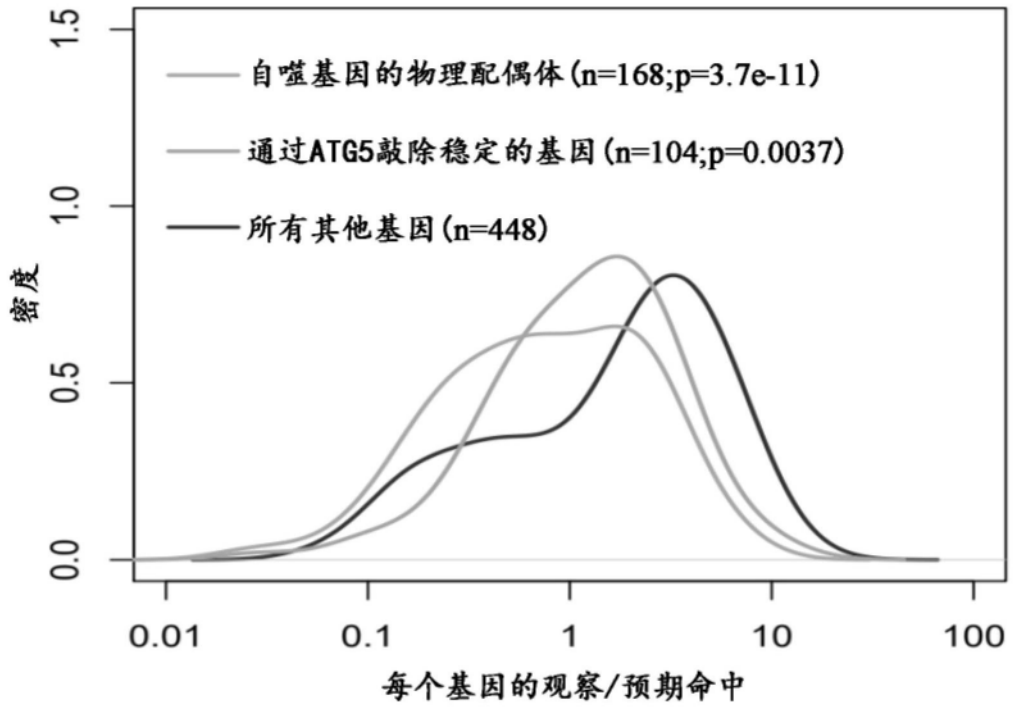

自噬基因的物理配偶体(n=168;p=3.7e-11)
通过ATG5敲除稳定的基因(n=104;p=0.0037)
所有其他基因(n=448)
密度
0.0
0.5
1.0
1.5
0.01
0.1
1
10
100
每个基因的观察/预期命中

图18C

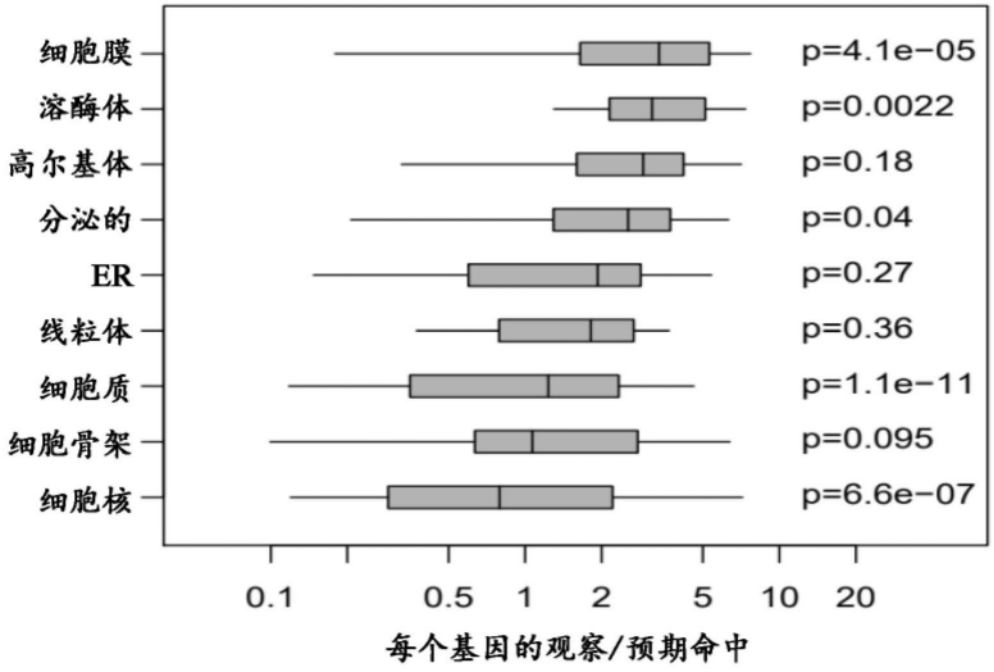

细胞膜
溶酶体
高尔基体
分泌的
ER
线粒体
细胞质
细胞骨架
细胞核
p=4.1e−05
p=0.0022
p=0.18
p=0.04
p=0.27
p=0.36
p=1.1e−11
p=0.095
p=6.6e−07
0.1
0.5
1
2
5
10
20
每个基因的观察/预期命中

图18D

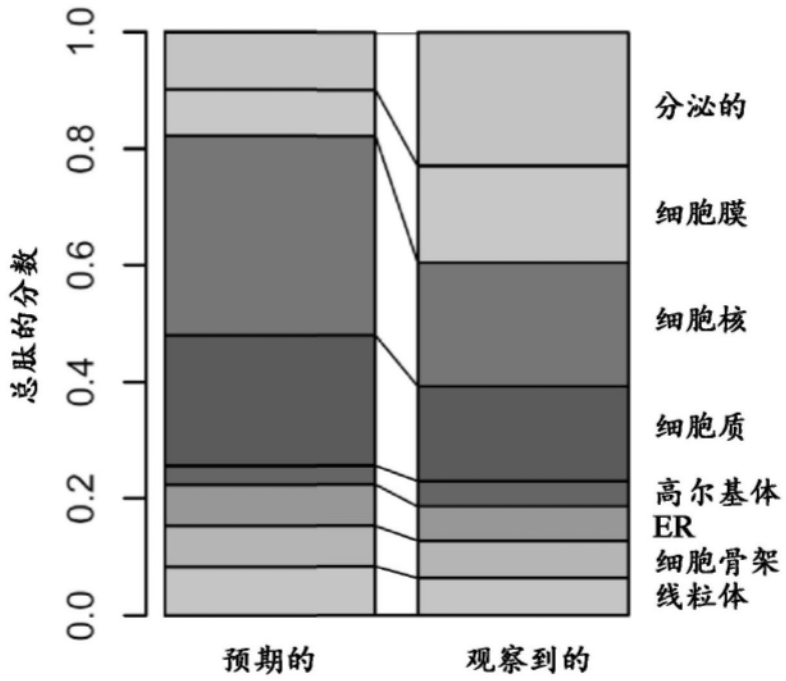


图18E

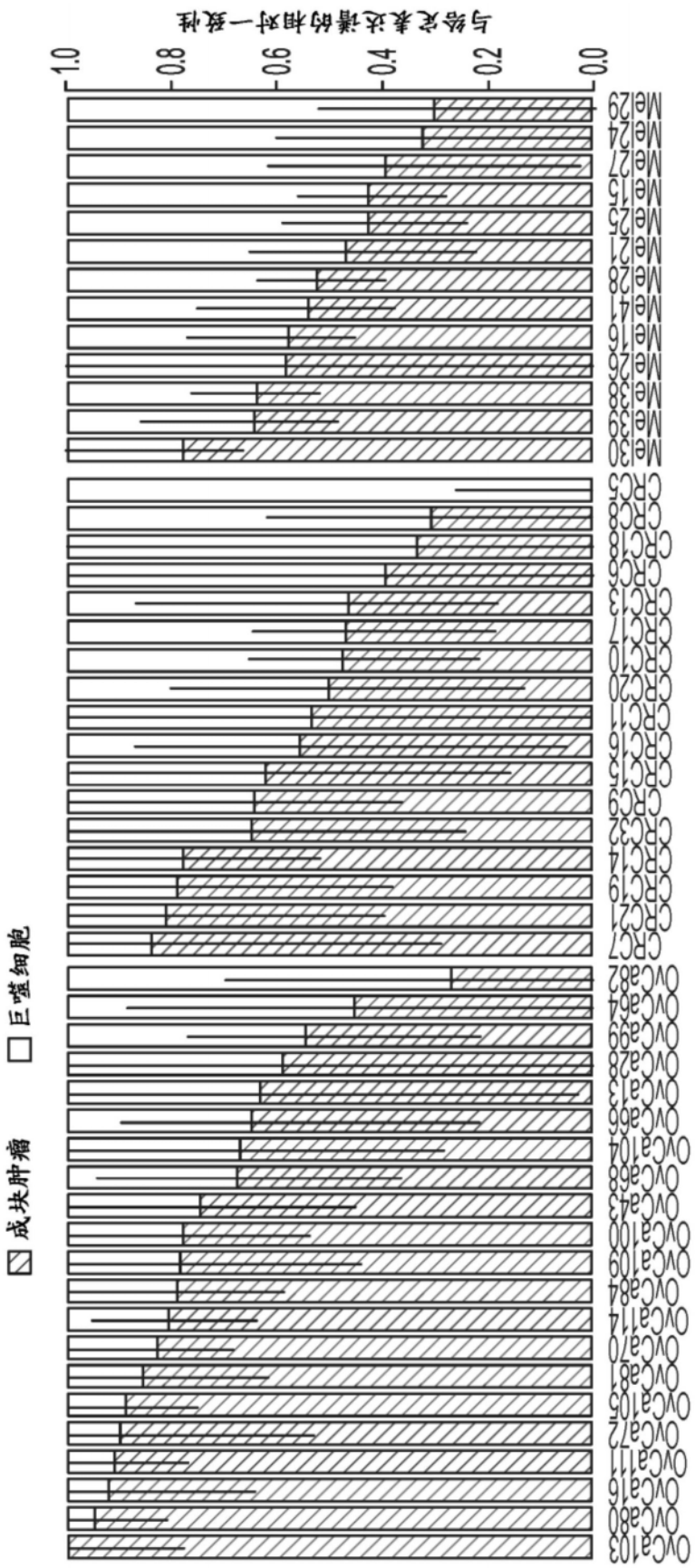

与给定表达谱的相对一致性
1.0
0.8
0.6
0.4
0.2
0.0
成块肿瘤
巨噬细胞
Mel29
Mel24
Mel27
Mel15
Mel25
Mel21
Mel28
Mel41
Mel16
Mel26
Mel33
Mel30
CRC5
CRC8
CRC18
CRC6
CRC13
CRC17
CRC10
CRC20
CRC11
CRC16
CRC15
CRC9
CRC12
CRC14
CRC19
CRC21
CRC7
OVCa82
OVCa64
OVCa99
OVCa28
OVCa13
OVCa69
OVCa104
OVCa68
OVCa43
OVCa100
OVCa109
OVCa84
OVCa114
OVCa70
OVCa81
OVCa105
OVCa72
OVCa111
OVCa116
OVCa80
OVCa103

图18F

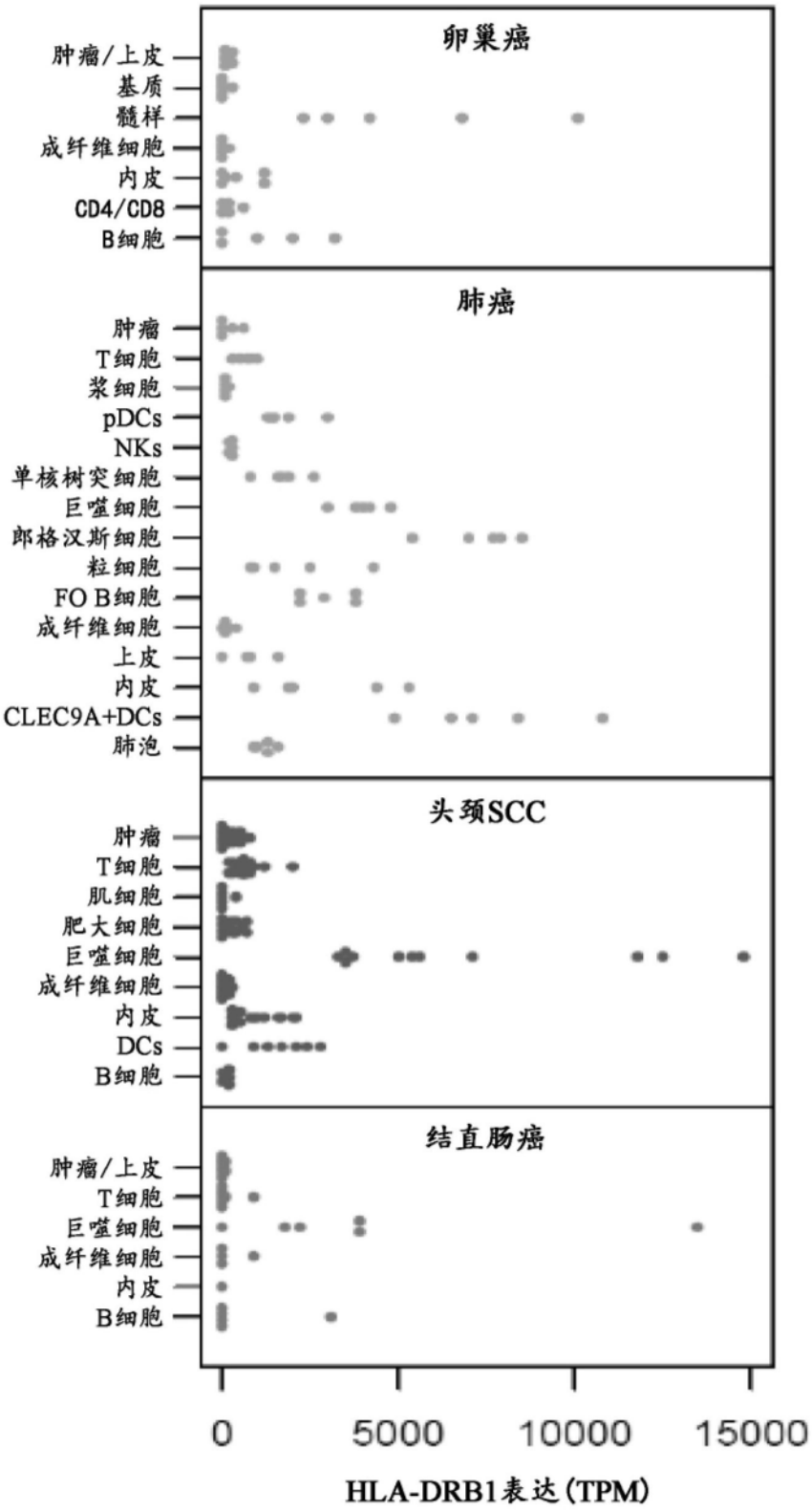

卵巢癌
肿瘤/上皮
基质
髓样
成纤维细胞
内皮
CD4/CD8
B细胞
肺癌
肿瘤
T细胞
浆细胞
pDCs
NKs
单核树突细胞
巨噬细胞
郎格汉斯细胞
粒细胞
FO B细胞
成纤维细胞
上皮
内皮
CLEC9A+DCs
肺泡
头颈SCC
肿瘤
T细胞
肌细胞
肥大细胞
巨噬细胞
成纤维细胞
内皮
DCs
B细胞
结直肠癌
肿瘤/上皮
T细胞
巨噬细胞
成纤维细胞
内皮
B细胞
0
5000
10000
15000
HLA-DRB1表达(TPM)

图19A

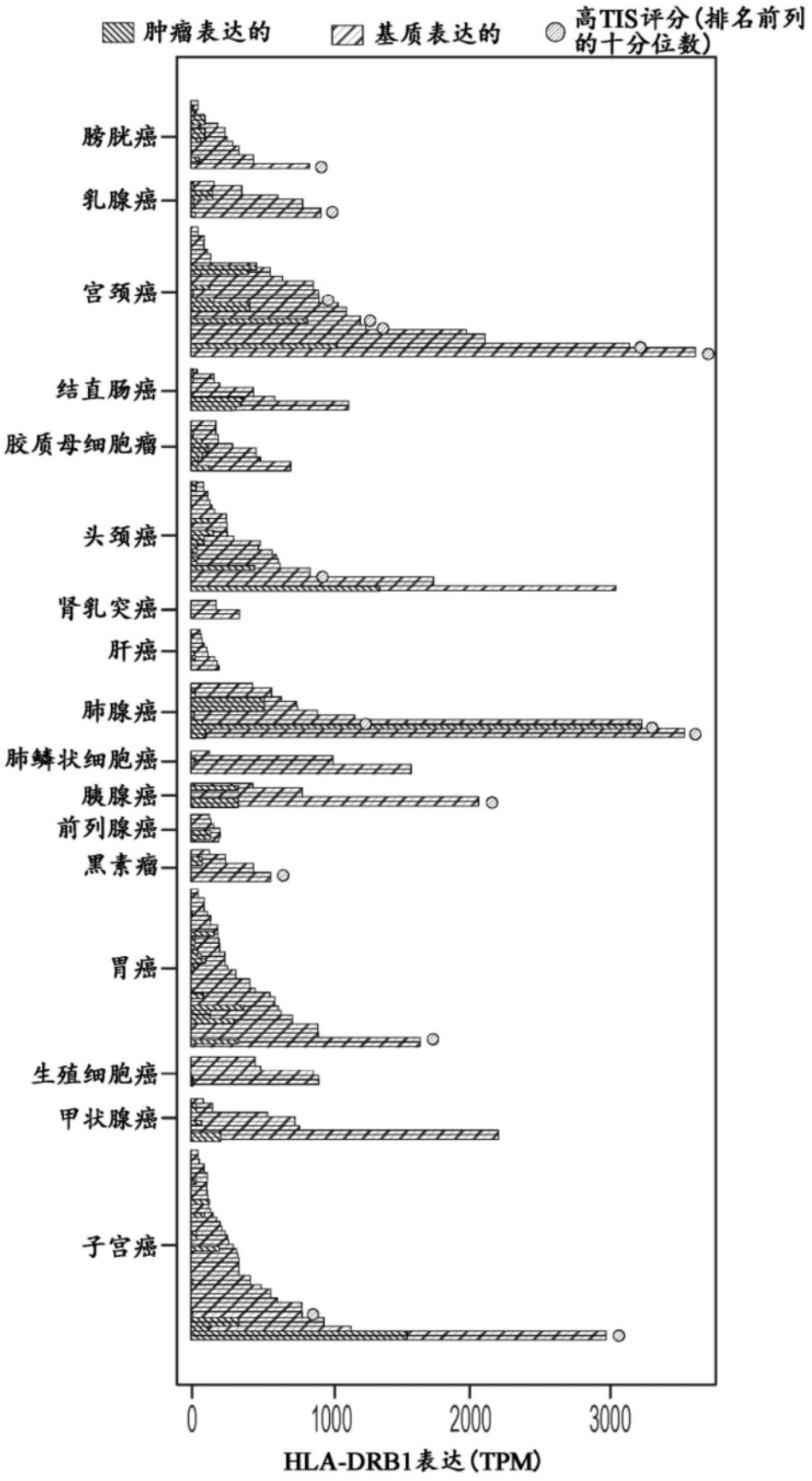

肿瘤表达的
基质表达的
高TIS评分(排名前列的十分位数)
膀胱癌
乳腺癌
宫颈癌
结直肠癌
胶质母细胞瘤
头颈癌
肾乳突癌
肝癌
肺腺癌
肺鳞状细胞癌
胰腺癌
前列腺癌
黑素瘤
胃癌
生殖细胞癌
甲状腺癌
子宫癌
0
1000
2000
3000
HLA-DRB1表达(TPM)

图19B

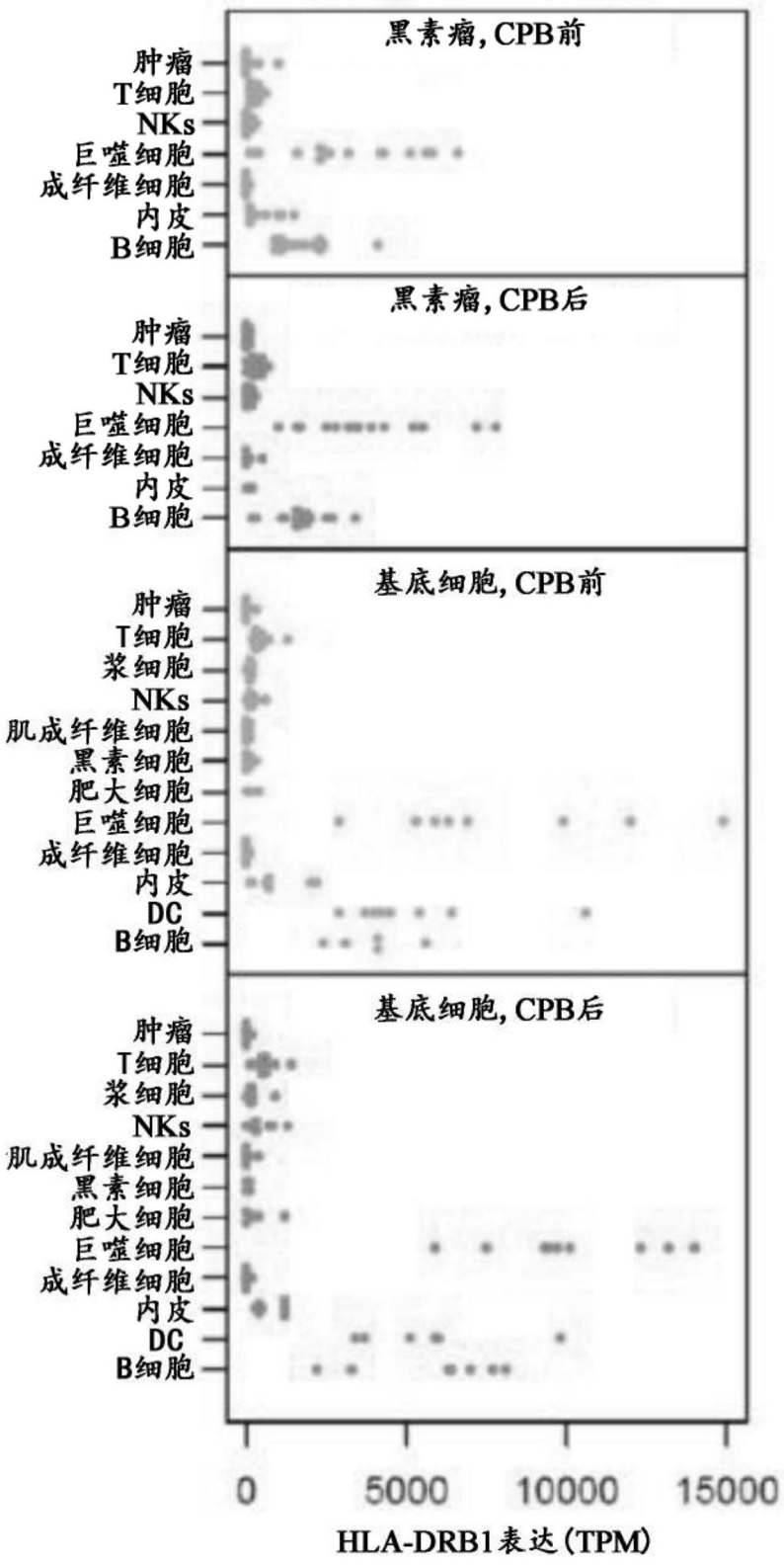

黑素瘤,CPB前
肿瘤
T细胞
NKs
巨噬细胞
成纤维细胞
内皮
B细胞
黑素瘤,CPB后
肿瘤
T细胞
NKs
巨噬细胞
成纤维细胞
内皮
B细胞
基底细胞,CPB前
肿瘤
T细胞
浆细胞
NKs
肌成纤维细胞
黑素细胞
肥大细胞
巨噬细胞
成纤维细胞
内皮
DC
B细胞
基底细胞,CPB后
肿瘤
T细胞
浆细胞
NKs
肌成纤维细胞
黑素细胞
肥大细胞
巨噬细胞
成纤维细胞
内皮
DC
B细胞
0
5000
10000
15000
HLA-DRB1表达(TPM)

图19C

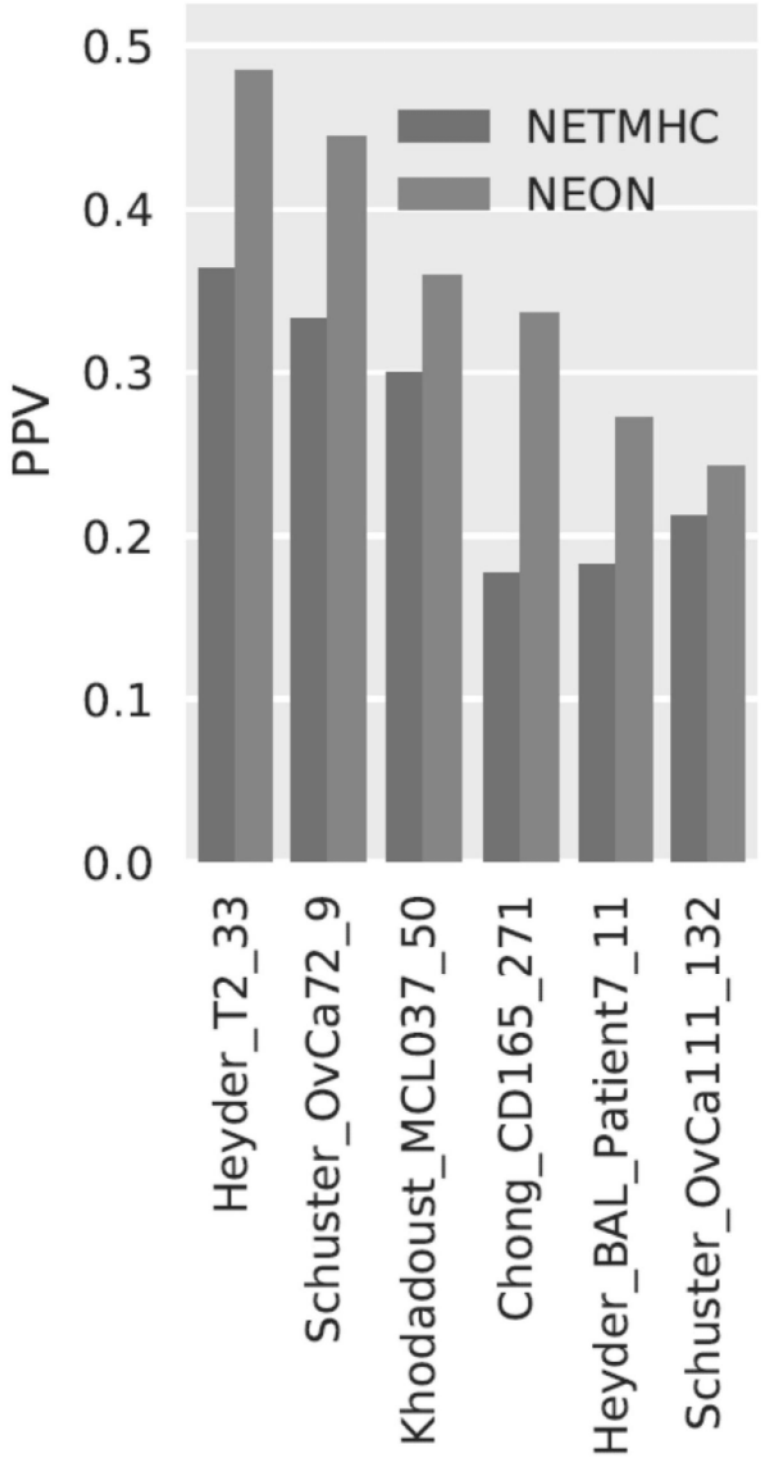

0.5
0.4
0.3
0.2
0.1
0.0
PPV
NETMHC
NEON
Heyder_T2_33
Schuster_OvCa72_9
Khodadoust_MCL037_50
Chong_CD165_271
Heyder_BAL_Patient7_11
Schuster_OvCa111_132

图20

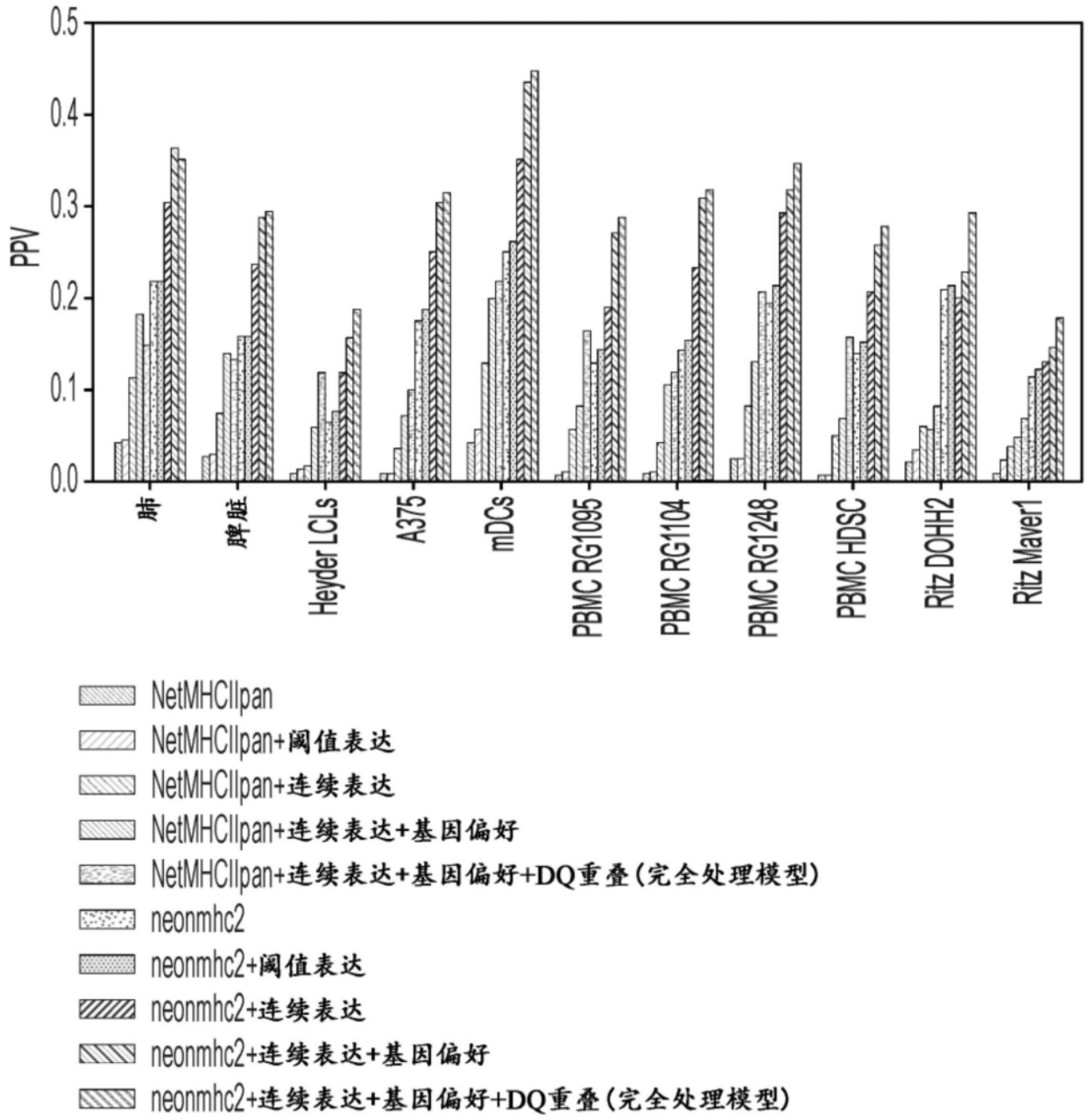

0.5
0.4
0.3
0.2
0.1
0.0
PPV
肺
脾脏
Heyder LCLs
A375
mDCs
PBMC RG1095
PBMC RG1104
PBMC RG1248
PBMC HDSC
Ritz DOHH2
Ritz Maver1
NetMHCIIpan
NetMHCIIpan+阈值表达
NetMHCIIpan+连续表达
NetMHCIIpan+连续表达+基因偏好
NetMHCIIpan+连续表达+基因偏好+DQ重叠(完全处理模型)
neonmhc2
neonmhc2+阈值表达
neonmhc2+连续表达
neonmhc2+连续表达+基因偏好
neonmhc2+连续表达+基因偏好+DQ重叠(完全处理模型)

图21A

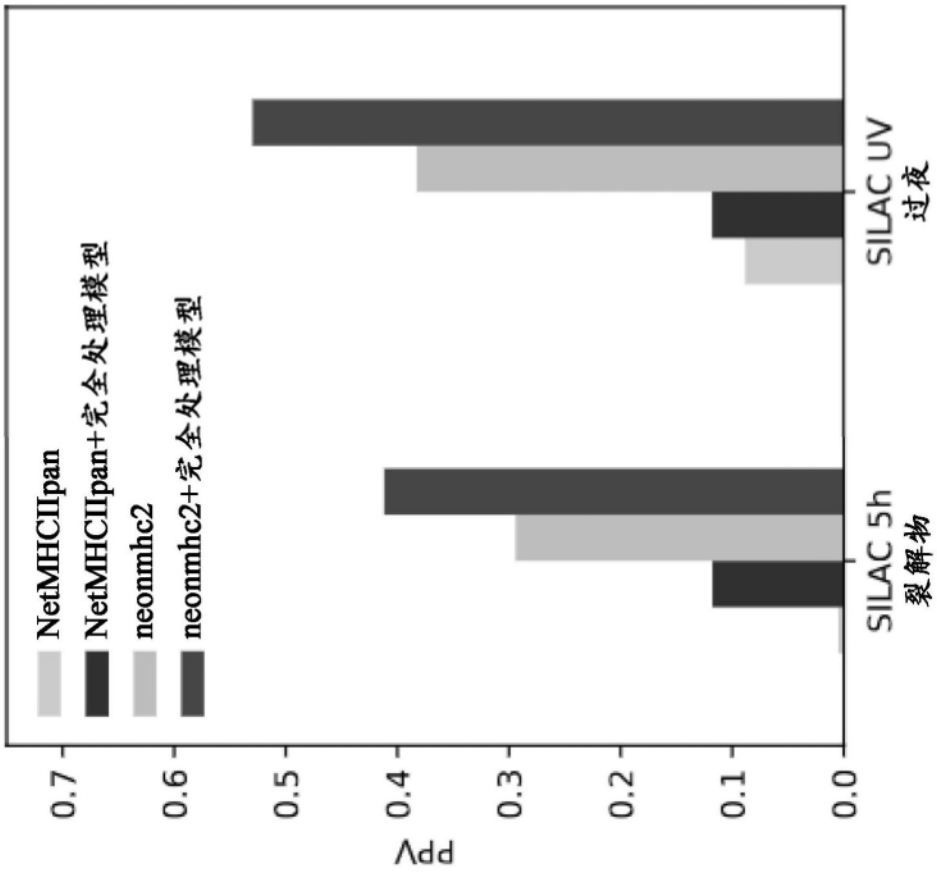

NetMHCIIpan
NetMHCIIpan+完全处理模型
neonmhc2
neonmhc2+完全处理模型
PPV
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
SILAC 5h
裂解物
SILAC UV
过夜

图21B

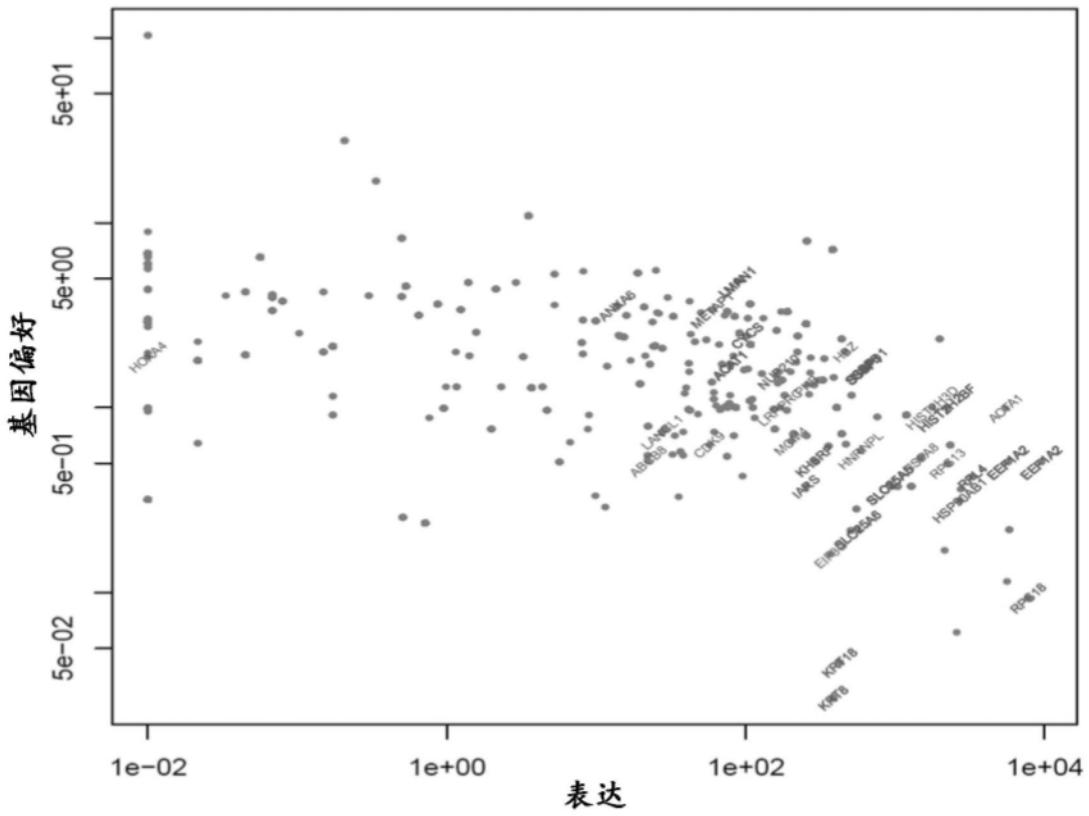


图21C

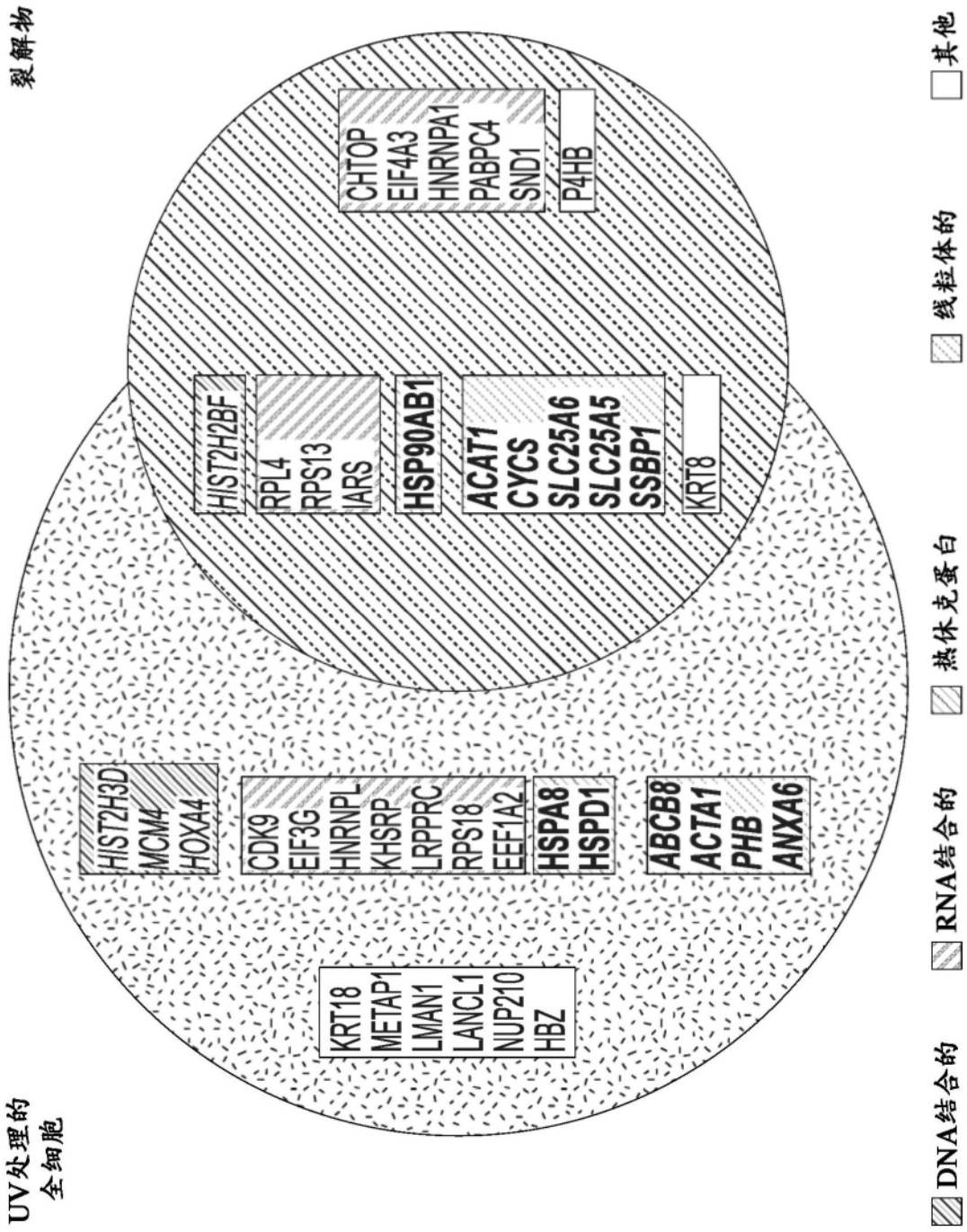

裂解物
UV处理的
全细胞
CHTOP
EIF4A3
HNRNPA1
PABPC4
SND1
P4HB
HIST2H2BF
RPL4
RPS13
IARS
HSP90AB1
ACAT1
CYCS
SLC25A6
SLC25A5
SSBP1
KRT8
HIST2H3D
MCM4
HOXA4
CDK9
EIF3G
HNRNPL
KHSRP
LRPPRC
RPS18
EEF1A2
HSPA8
HSPD1
ABCB8
ACTA1
PHB
ANXA6
KRT18
METAP1
LMAN1
LANCL1
NUP210
HBZ
DNA结合的
RNA结合的
热休克蛋白
线粒体的
其他

图21D

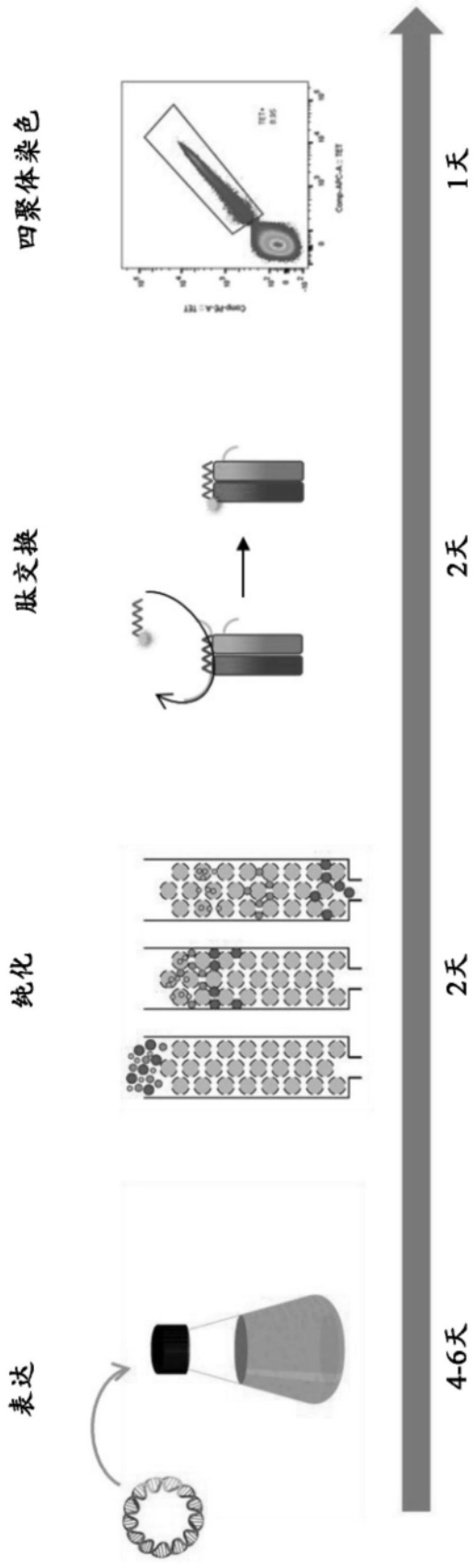


图22A

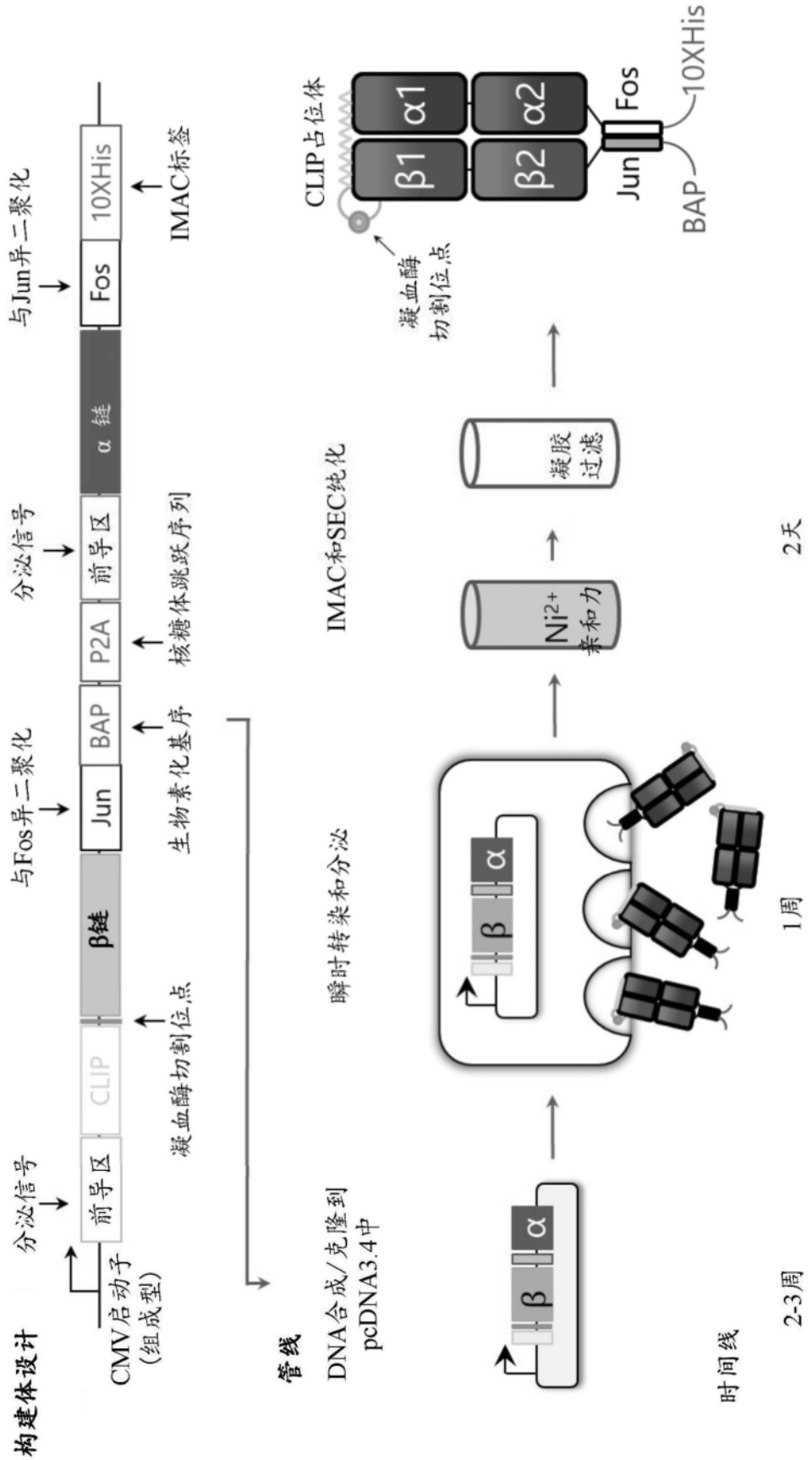

构建体设计
分泌信号
前导区
CMV启动子
(组成型)
CLIP
凝血酶切割位点
βlink
与Fos异二聚化
Jun
BAP
生物素化基序
P2A
分泌信号
前导区
核糖体跳跃序列
α 链
与Jun异二聚化
Fos
10XHis
IMAC标签
管线
DNA合成/克隆到
pcDNA3.4中
瞬时转染和分泌
IMAC和SEC纯化
Ni2+
亲和力
凝胶
过滤
CLIP占位体
凝血酶
切割位点
β1
α1
β2
α2
Fos
Jun
BAP
10XHis
时间线
2-3周
1周
2天

图22B

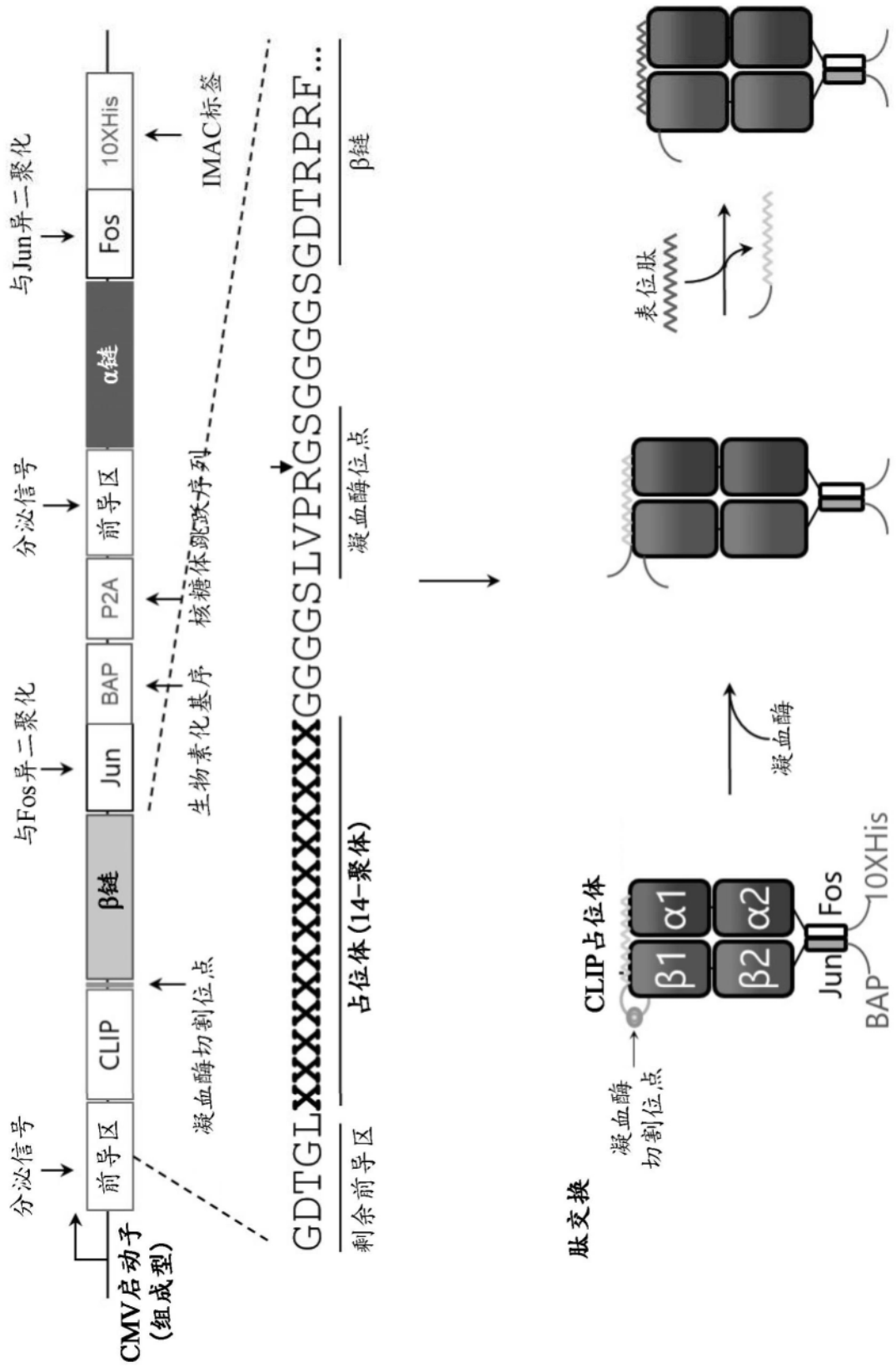

分泌信号
前导区
CLIP
凝血酶切割位点
剩余前导区
CMV启动子
(组成型)
β链
与Fos异二聚化
Jun
BAP
生物素化基序
P2A
核糖体跳跃序列
分泌信号
前导区
α链
与Jun异二聚化
Fos
10XHis
IMAC标签
GDTGLXXXXXXXXXXXXXXXGGGGSLVPRGSGGGGSGDTRPRF...
占位体(14-聚体)
凝血酶位点
β链
肽交换
CLIP占位体
凝血酶
切割位点
β1
α1
β2
α2
Jun
Fos
BAP
10XHis
凝血酶
表位肽

图23

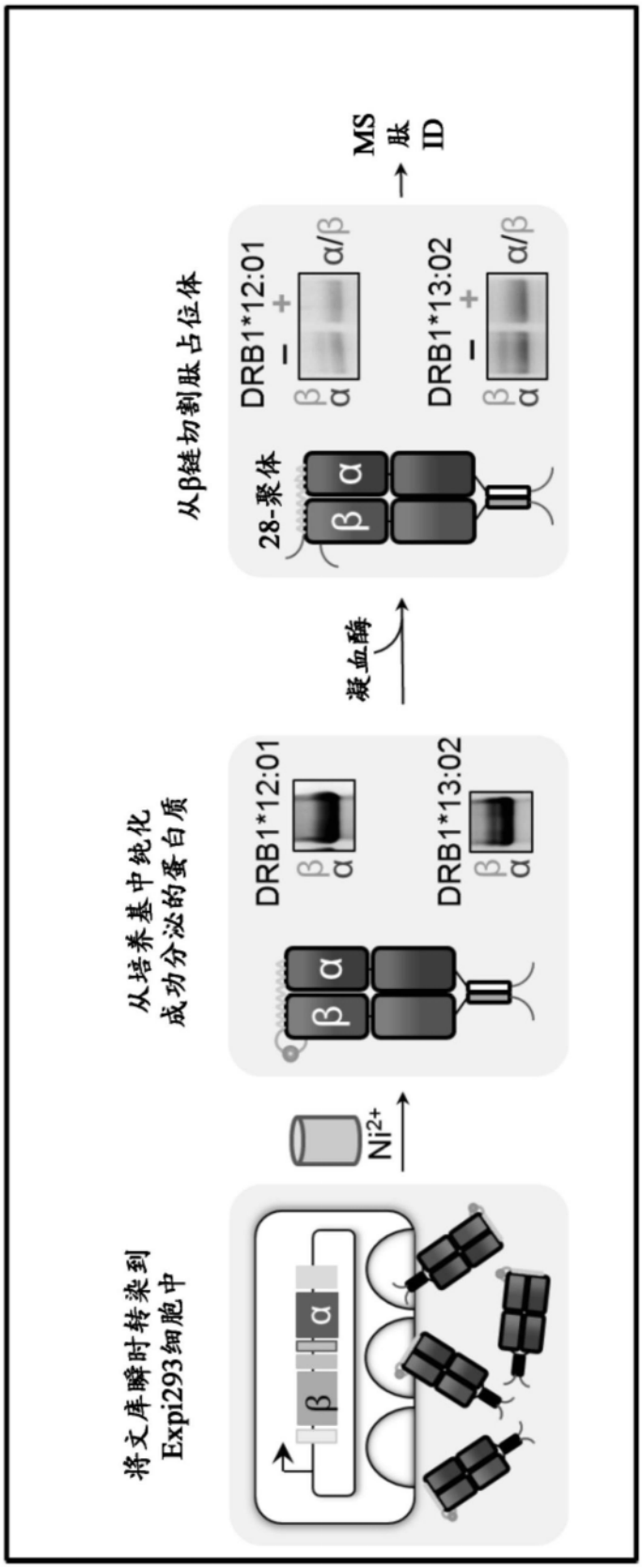

将文库瞬时转染到
Expi293细胞中
Ni2+
从培养基中纯化
成功分泌的蛋白质
DRB1*12:01
DRB1*13:02
凝血酶
从β链切割肽占位体
28-聚体
DRB1*12:01
DRB1*13:02
MS
肽
ID

图24

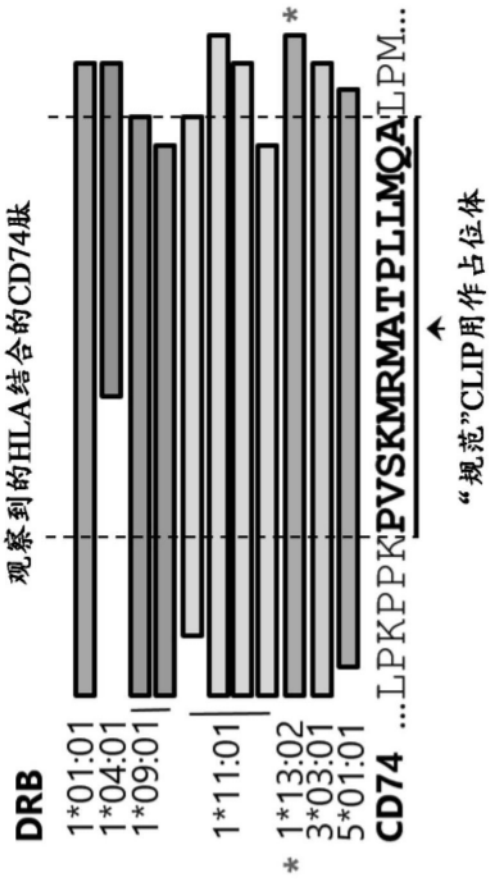

观察到的HLA结合的CD74肽
DRB
1*01:01
1*04:01
1*09:01
1*11:01
1*13:02
3*03:01
5*01:01
CD74 ...LPKPPKPVSKMRMATPLLMQALPM...
“规范”CLIP用作占位体

图25A

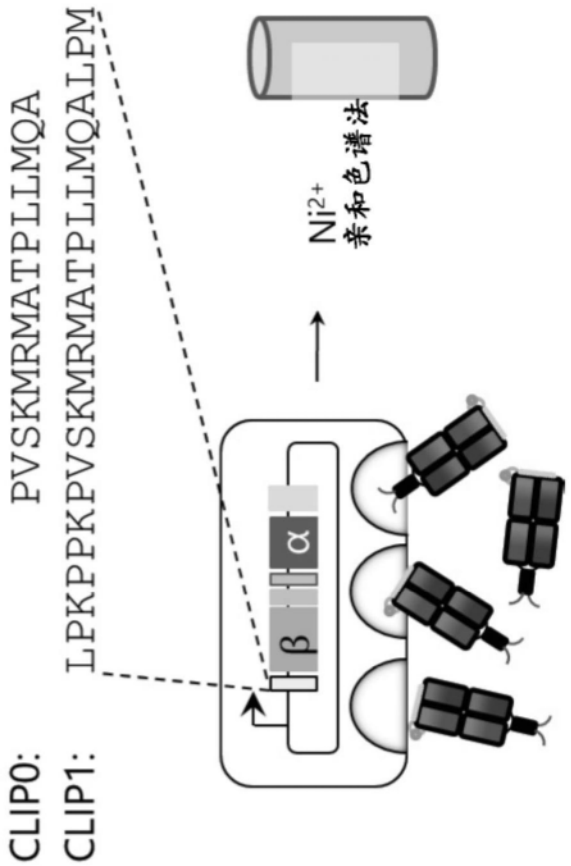

CLIP0: PVSKMRMATPLLMQA
CLIP1: LPKPPKPVSKMRMATPLLMQALPM
Ni2+
亲和色谱法
β
α

图25B

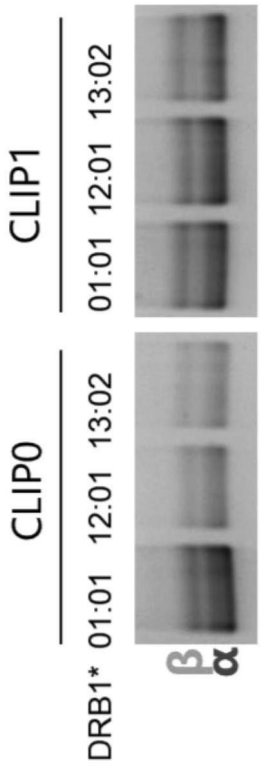


图25C

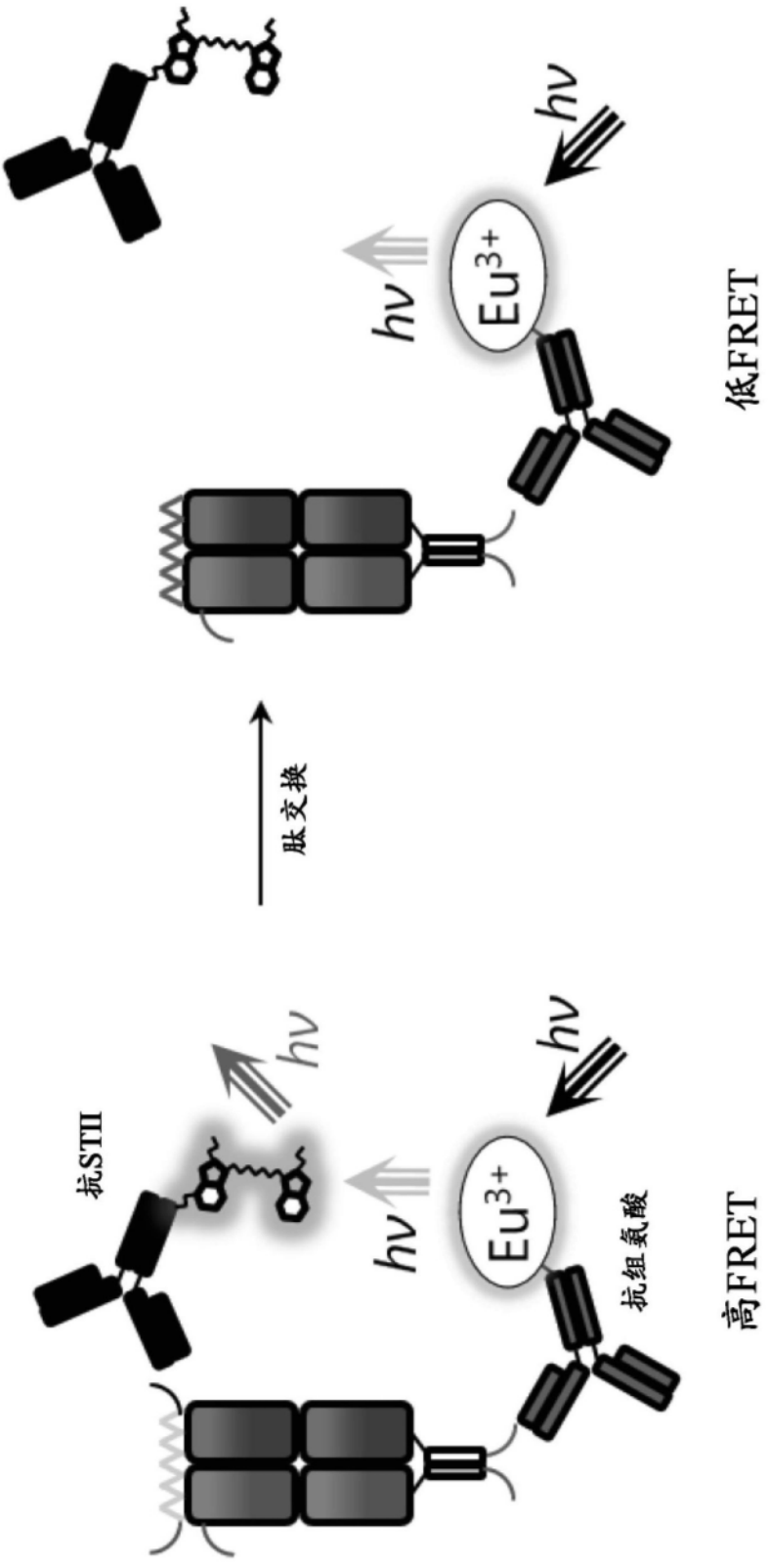


图26A

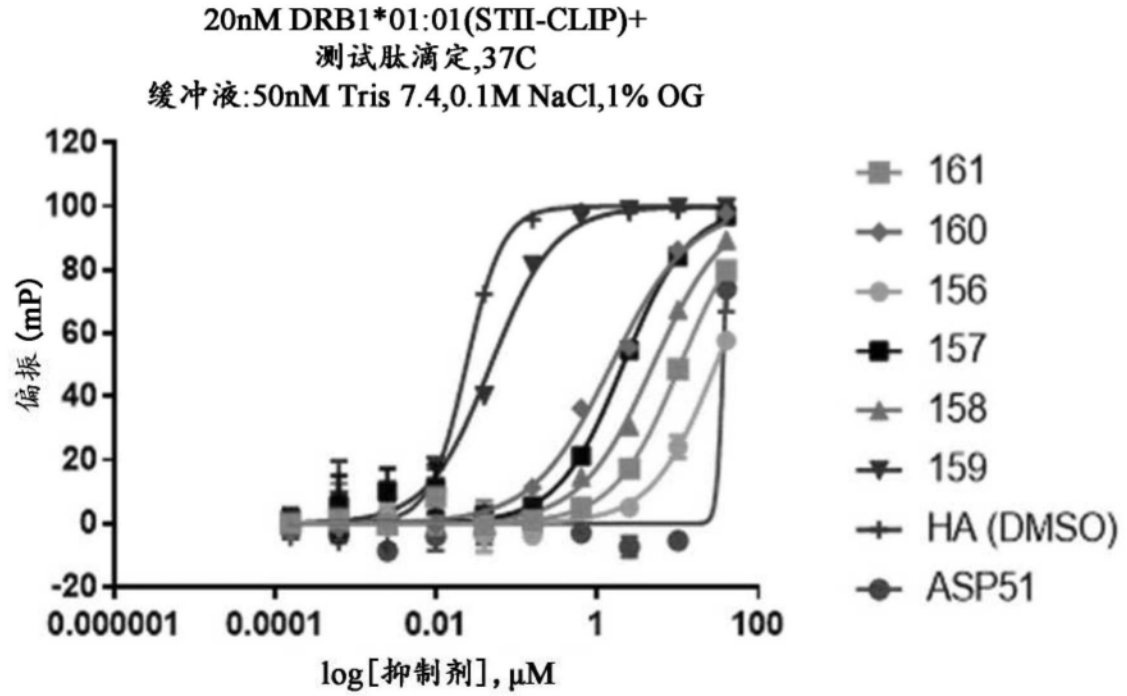

20nM DRB1*01:01(STII-CLIP)+
测试肽滴定,37C
缓冲液:50nM Tris 7.4,0.1M NaCl,1% OG
偏振 (mP)
120
100
80
60
40
20
0
-20
0.000001
0.0001
0.01
1
100
log[抑制剂], μM
161
160
156
157
158
159
HA (DMSO)
ASP51

图26B

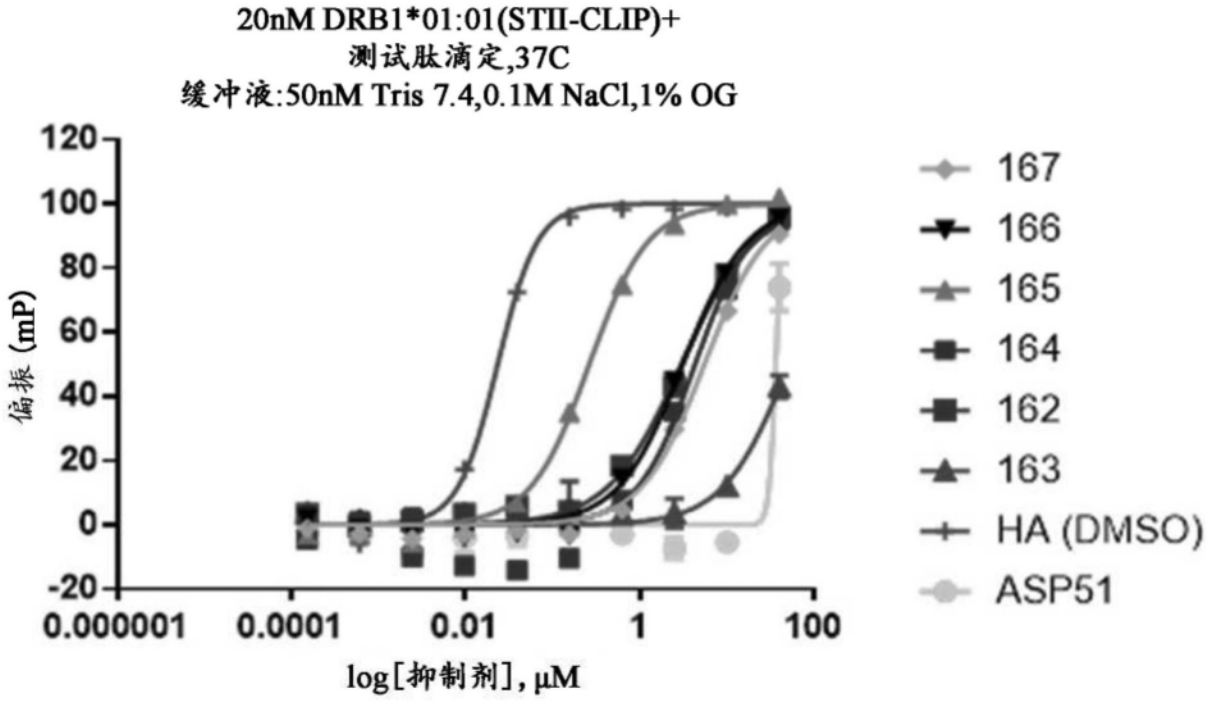

20nM DRB1*01:01(STII-CLIP)+
测试肽滴定,37C
缓冲液:50nM Tris 7.4,0.1M NaCl,1% OG
偏振 (mP)
120
100
80
60
40
20
0
-20
0.000001
0.0001
0.01
1
100
log[抑制剂], μM
167
166
165
164
162
163
HA (DMSO)
ASP51

图26C

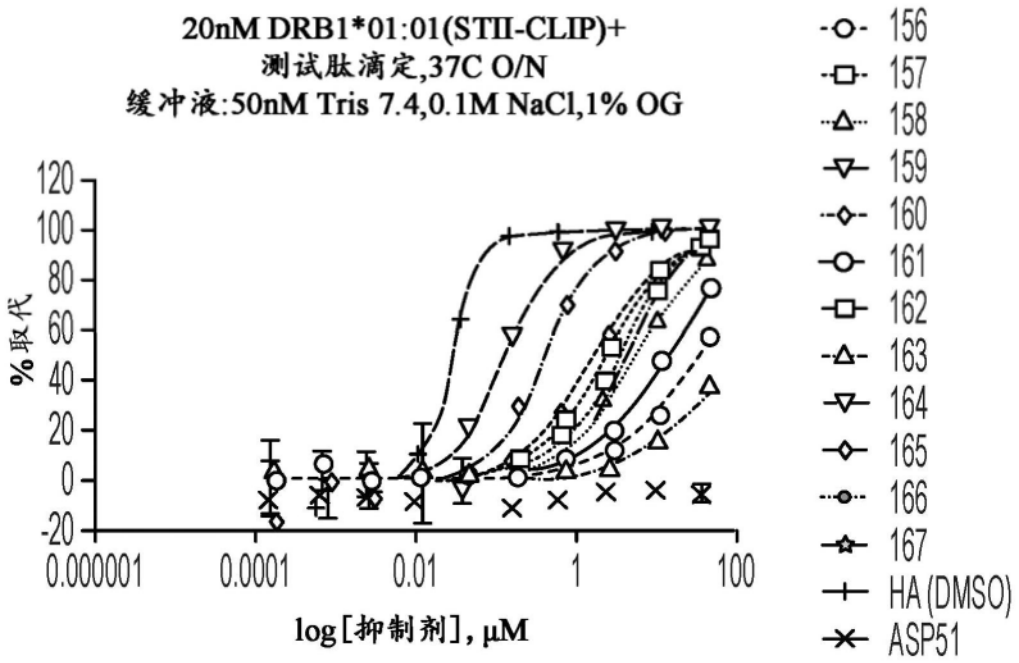

20nM DRB1*01:01(STII-CLIP)+
测试肽滴定,37C O/N
缓冲液:50nM Tris 7.4,0.1M NaCl,1% OG
% 取代
log[抑制剂], μM
156
157
158
159
160
161
162
163
164
165
166
167
HA (DMSO)
ASP51

图26D

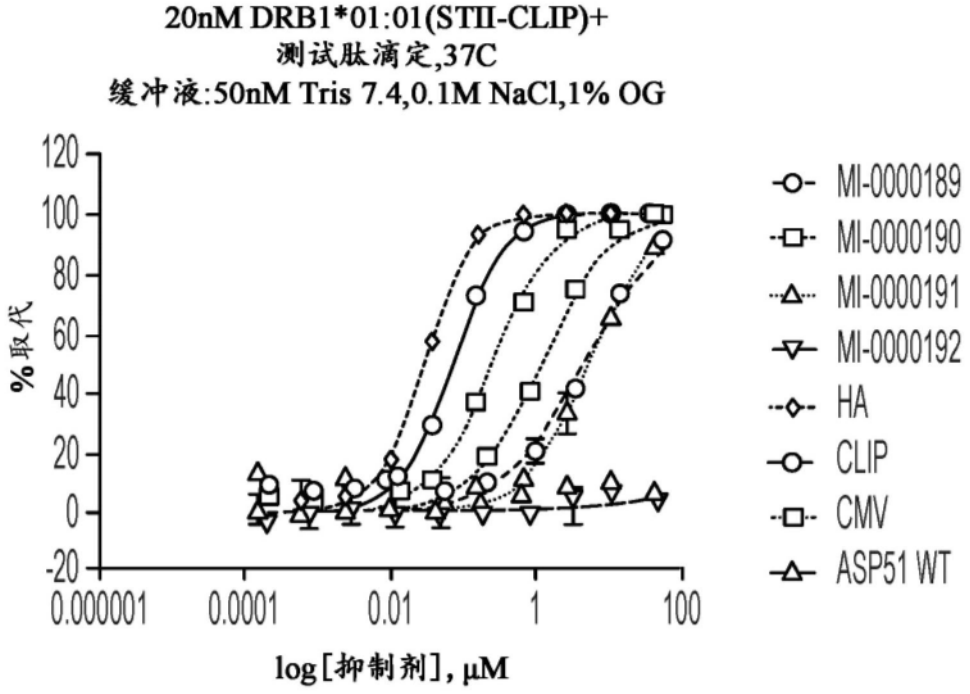

20nM DRB1*01:01(STII-CLIP)+
测试肽滴定,37C
缓冲液:50nM Tris 7.4,0.1M NaCl,1% OG
% 取代
log[抑制剂], μM
MI-0000189
MI-0000190
MI-0000191
MI-0000192
HA
CLIP
CMV
ASP51 WT

图26E

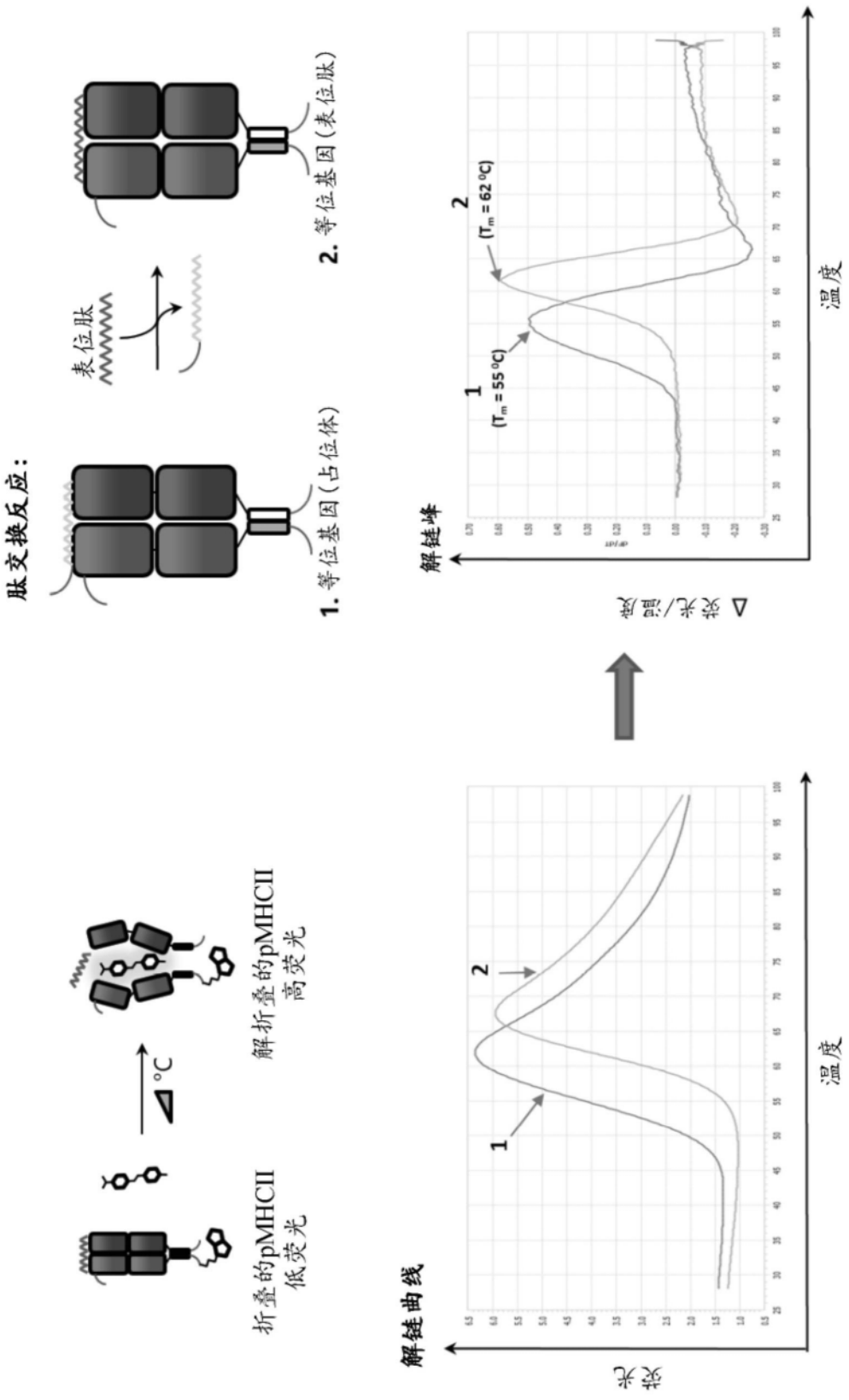

肽交换反应：
表位肽
1. 等位基因(占位体)
2. 等位基因(表位肽)
折叠的pMHCII
低荧光
°C
解折叠的pMHCII
高荧光
解链曲线
荧光
温度
1
2
解链峰
Δ 荧光/温度
(Tm = 55 °C)
(Tm = 62 °C)

图26F

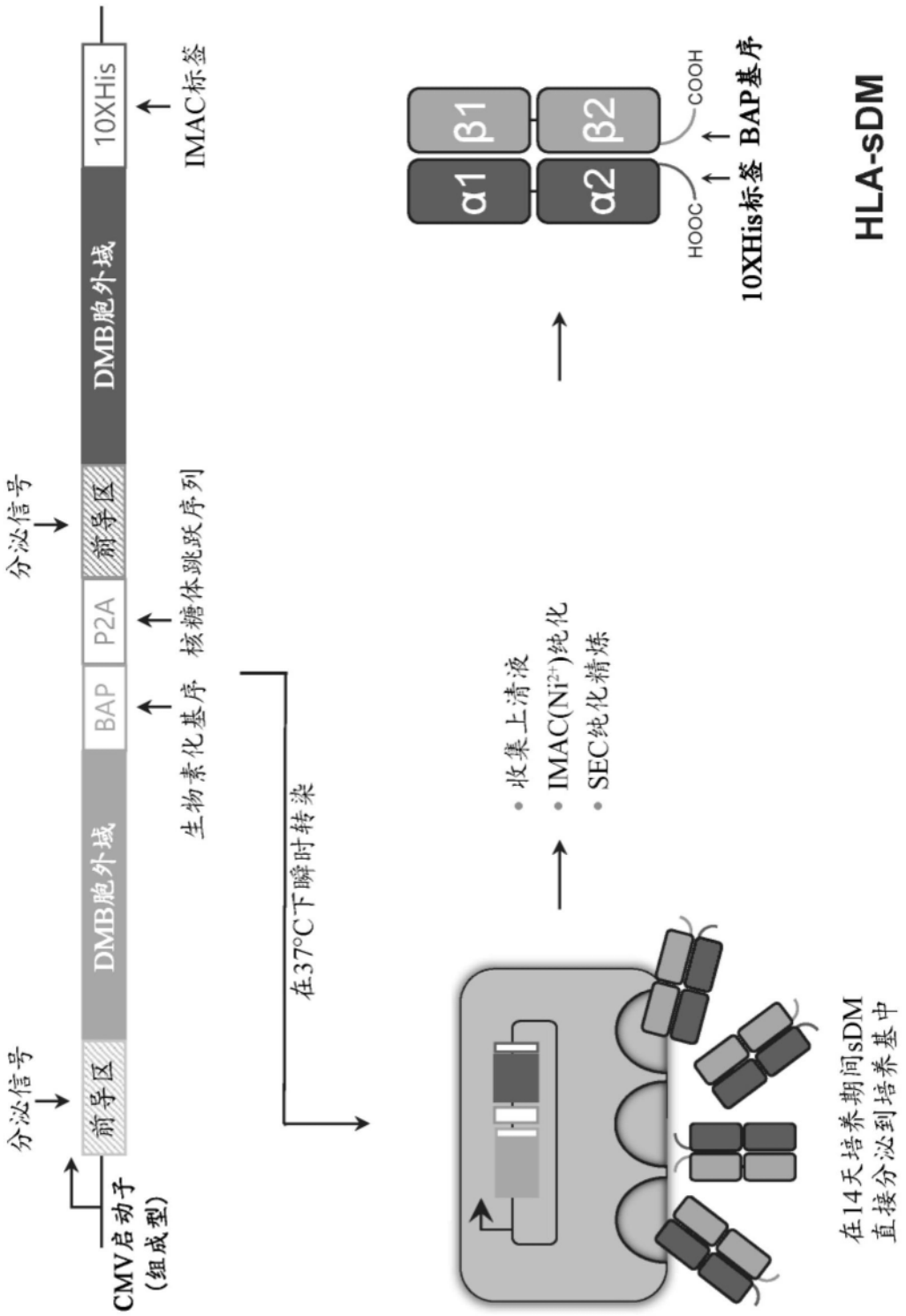

CMV启动子
(组成型)
分泌信号
前导区
DMB胞外域
BAP
生物素化基序
P2A
核糖体跳跃序列
分泌信号
前导区
DMB胞外域
10XHis
IMAC标签
在37°C下瞬时转染
在14天培养期间sDM
直接分泌到培养基中
• 收集上清液
• IMAC(Ni2+)纯化
• SEC纯化精炼
α1
β1
α2
β2
HOOC
COOH
10XHis标签
BAP基序
HLA-sDM

图26G

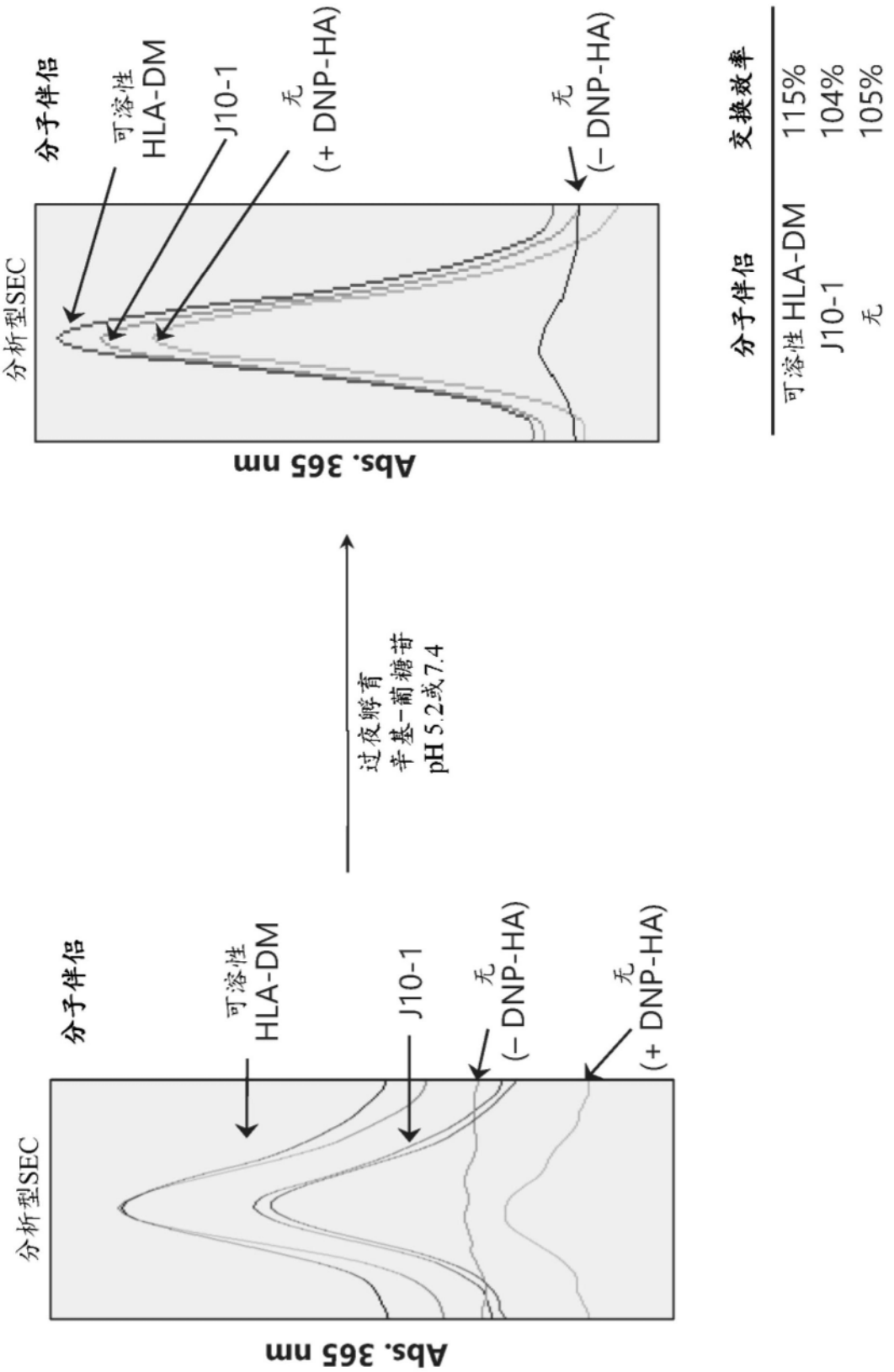

分子伴侣
可溶性 HLA-DM
J10-1
无 (+ DNP-HA)
无 (– DNP-HA)
分析型SEC
Abs. 365 nm
过夜孵育
辛基-葡糖苷
pH 5.2或7.4
分子伴侣 交换效率
可溶性 HLA-DM 115%
J10-1 104%
无 105%

图26H

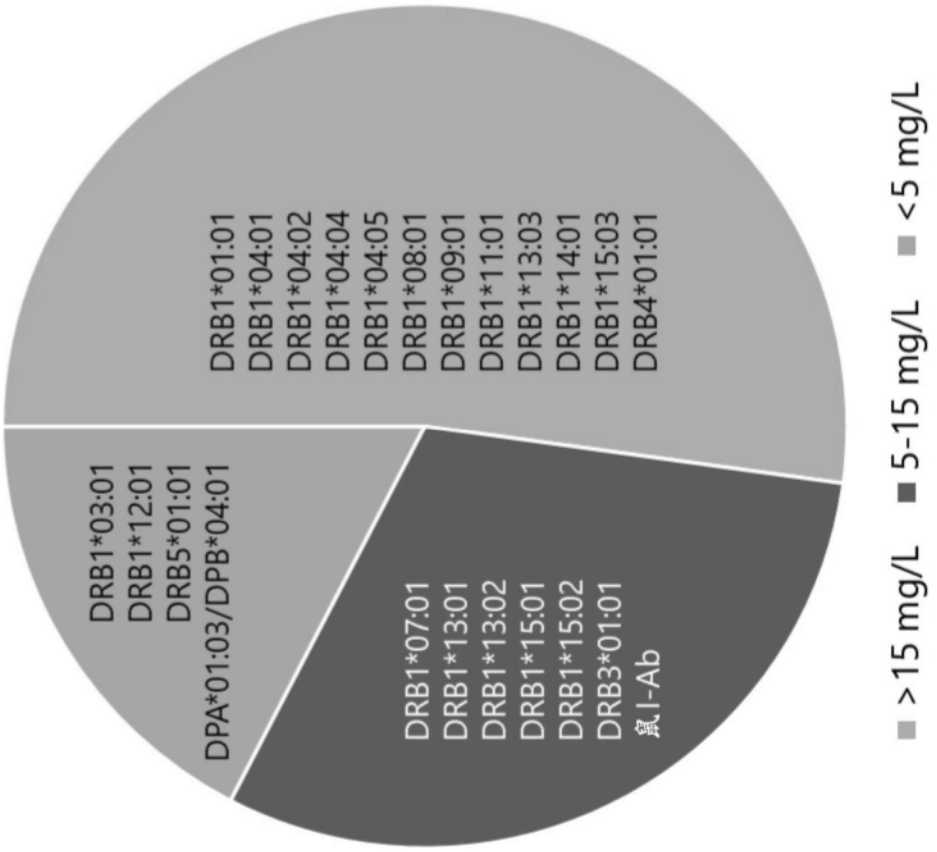

DRB1*01:01
DRB1*04:01
DRB1*04:02
DRB1*04:04
DRB1*04:05
DRB1*08:01
DRB1*09:01
DRB1*11:01
DRB1*13:03
DRB1*14:01
DRB1*15:03
DRB4*01:01
DRB1*07:01
DRB1*13:01
DRB1*13:02
DRB1*15:01
DRB1*15:02
DRB3*01:01
鼠I-Ab
DRB1*03:01
DRB1*12:01
DRB5*01:01
DPA*01:03/DPB*04:01
>15 mg/L
5-15 mg/L
<5 mg/L

图27A

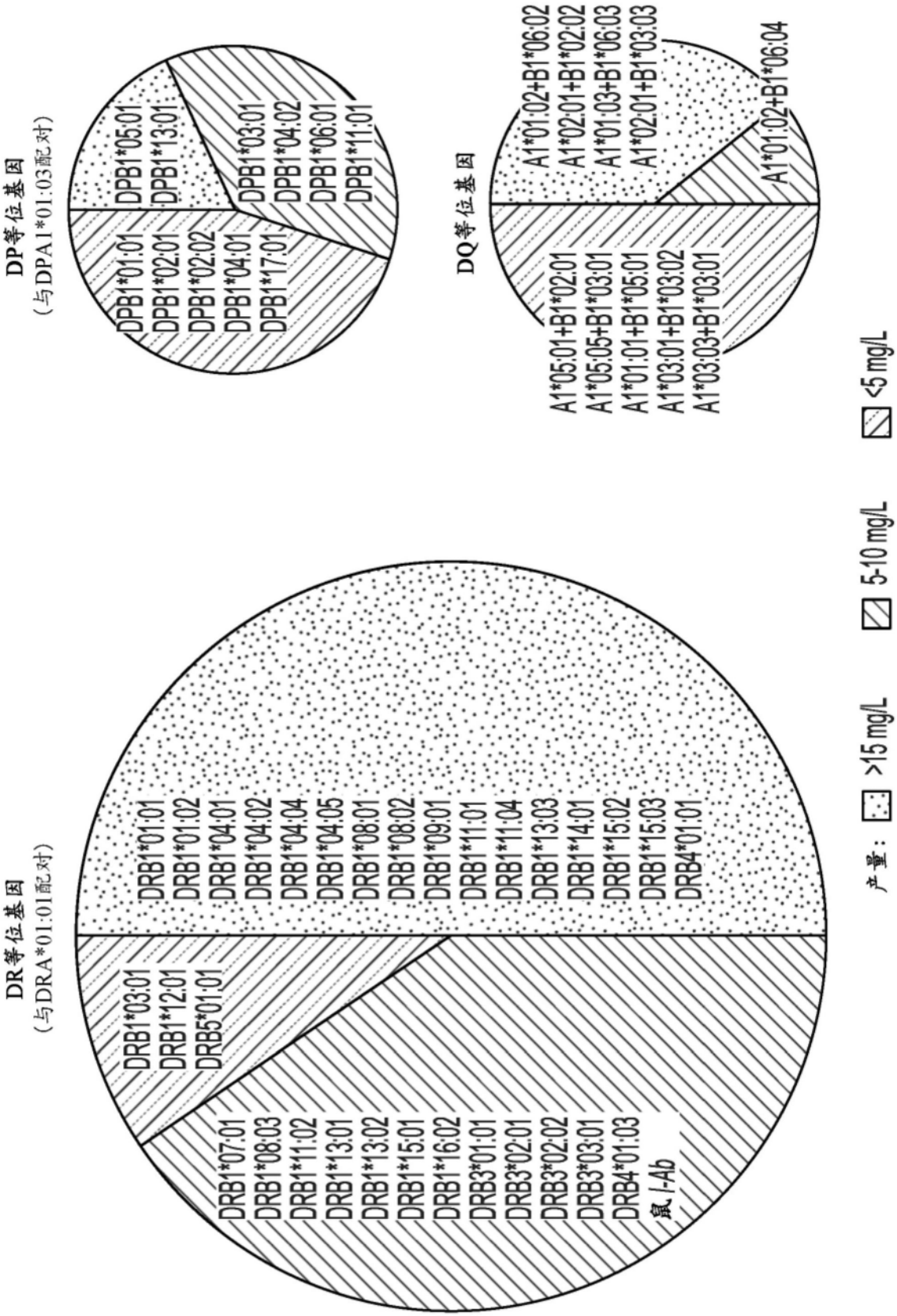

DP等位基因
(与DPA1*01:03配对)
DPB1*05:01
DPB1*13:01
DPB1*03:01
DPB1*04:02
DPB1*06:01
DPB1*11:01
DPB1*01:01
DPB1*02:01
DPB1*02:02
DPB1*04:01
DPB1*17:01
DQ等位基因
A1*01:02+B1*06:02
A1*02:01+B1*02:02
A1*01:03+B1*06:03
A1*02:01+B1*03:03
A1*01:02+B1*06:04
A1*05:01+B1*02:01
A1*05:05+B1*03:01
A1*01:01+B1*05:01
A1*03:01+B1*03:02
A1*03:03+B1*03:01
DR等位基因
(与DRA*01:01配对)
DRB1*01:01
DRB1*01:02
DRB1*04:01
DRB1*04:02
DRB1*04:04
DRB1*04:05
DRB1*08:01
DRB1*08:02
DRB1*09:01
DRB1*11:01
DRB1*11:04
DRB1*13:03
DRB1*14:01
DRB1*15:02
DRB1*15:03
DRB4*01:01
DRB1*03:01
DRB1*12:01
DRB5*01:01
DRB1*07:01
DRB1*08:03
DRB1*11:02
DRB1*13:01
DRB1*13:02
DRB1*15:01
DRB1*16:02
DRB3*01:01
DRB3*02:01
DRB3*02:02
DRB3*03:01
DRB4*01:03
鼠 I-Ab
产量:
>15 mg/L
5-10 mg/L
<5 mg/L

图27B

等位基因	验证肽的交换
DRB1*07:01	3
DRB1*15:01	7
DRB1*01:01	8
DRB1*11:01	12
DRB1*13:02	1

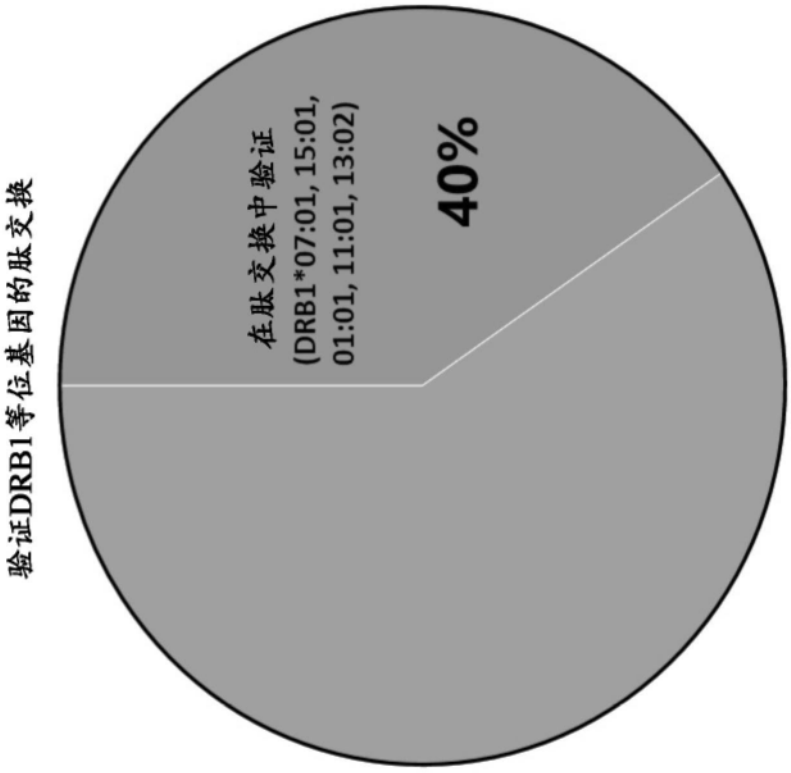


图27C

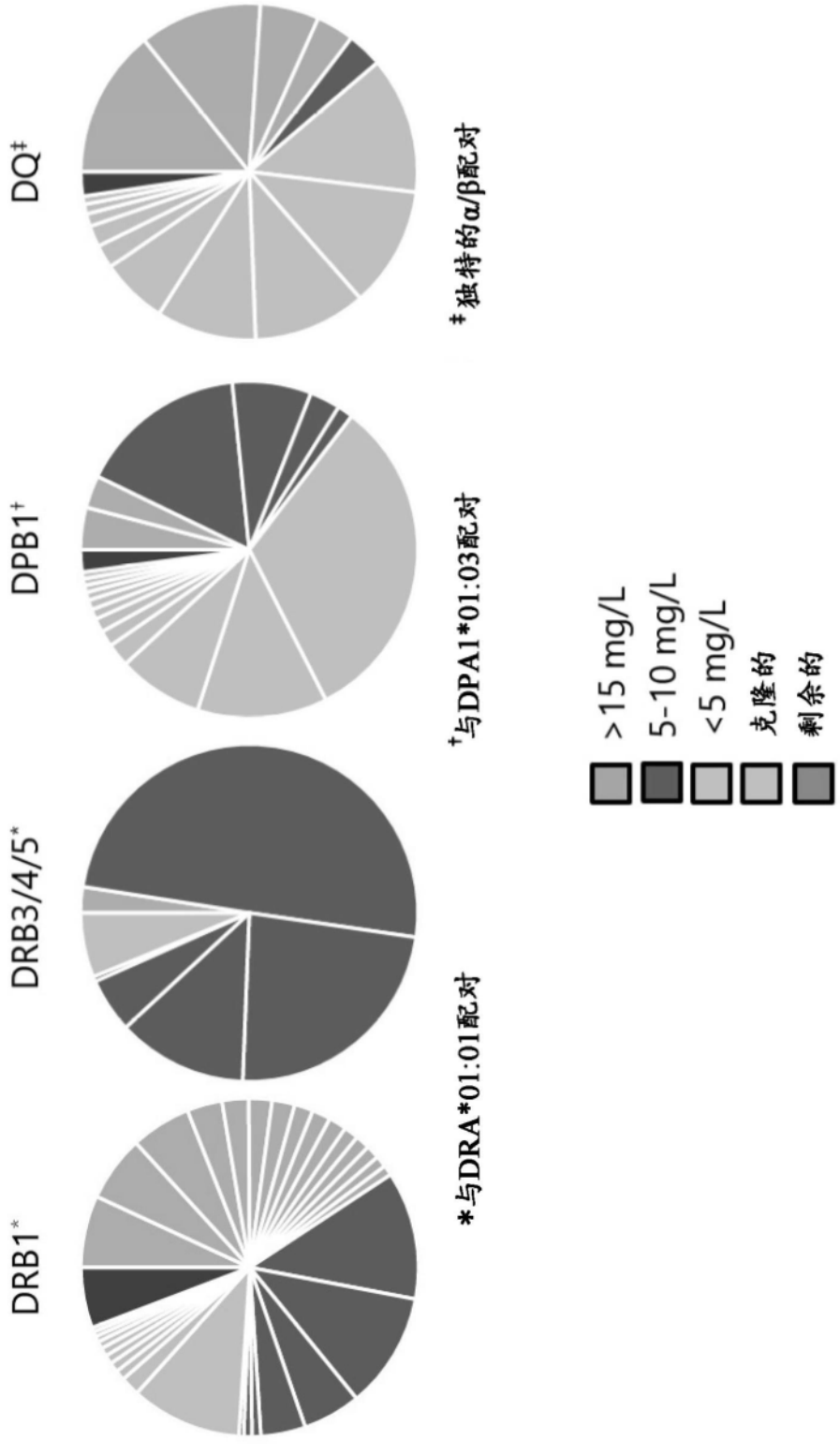

DRB1*
DRB3/4/5*
DPB1†
DQ‡
*与DRA*01:01配对
†与DPA1*01:03配对
‡独特的α/β配对
>15 mg/L
5-10 mg/L
<5 mg/L
克隆的
剩余的

图27D

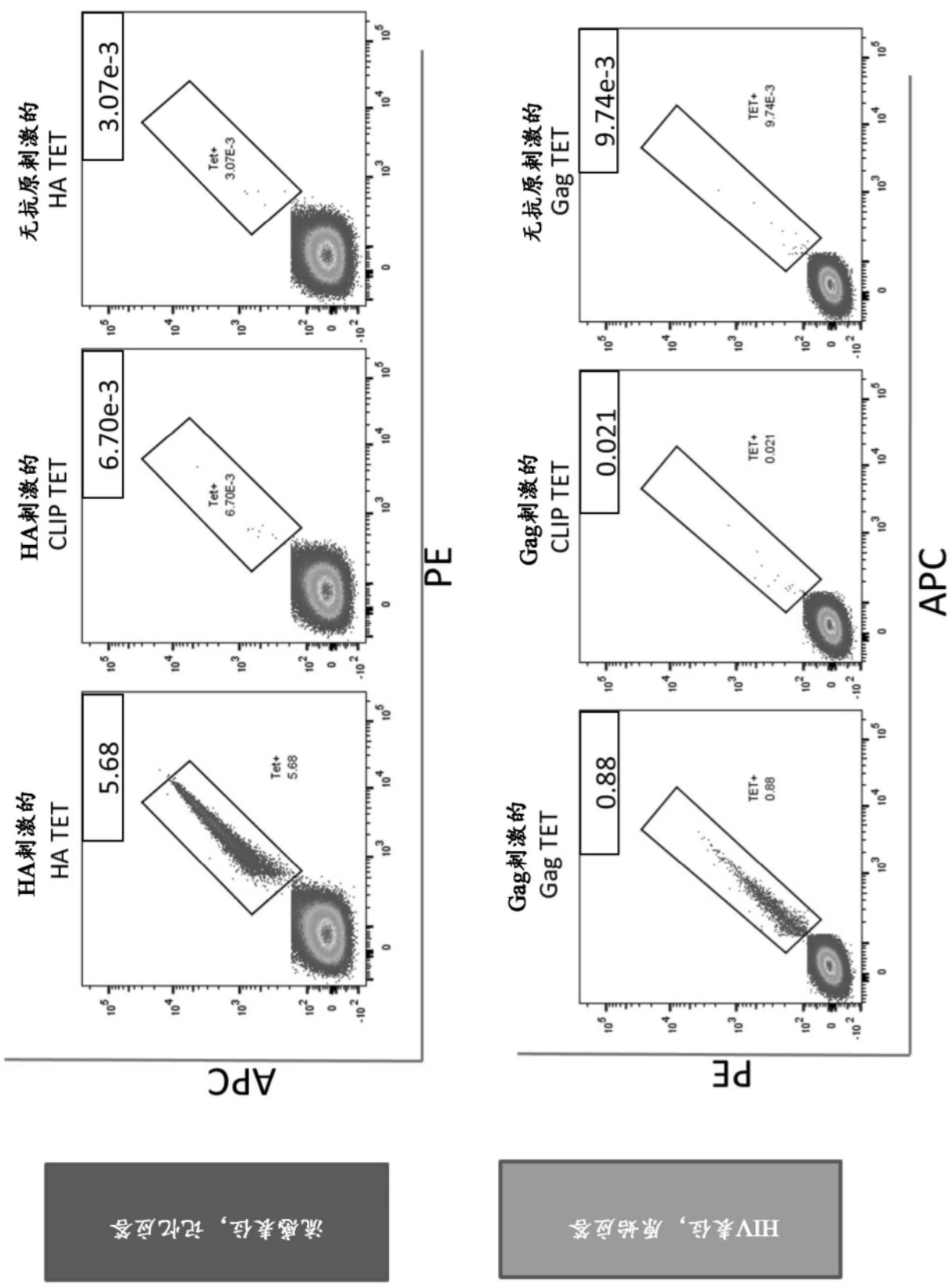

HA刺激的
HA TET
5.68
Tet+
5.68
HA刺激的
CLIP TET
6.70e-3
Tet+
6.70E-3
无抗原刺激的
HA TET
3.07e-3
Tet+
3.07E-3
APC
PE
Gag刺激的
Gag TET
0.88
TET+
0.88
Gag刺激的
CLIP TET
0.021
TET+
0.021
无抗原刺激的
Gag TET
9.74e-3
TET+
9.74E-3
PE
APC
流感表位，记忆应答
HIV表位，原始应答

图27E

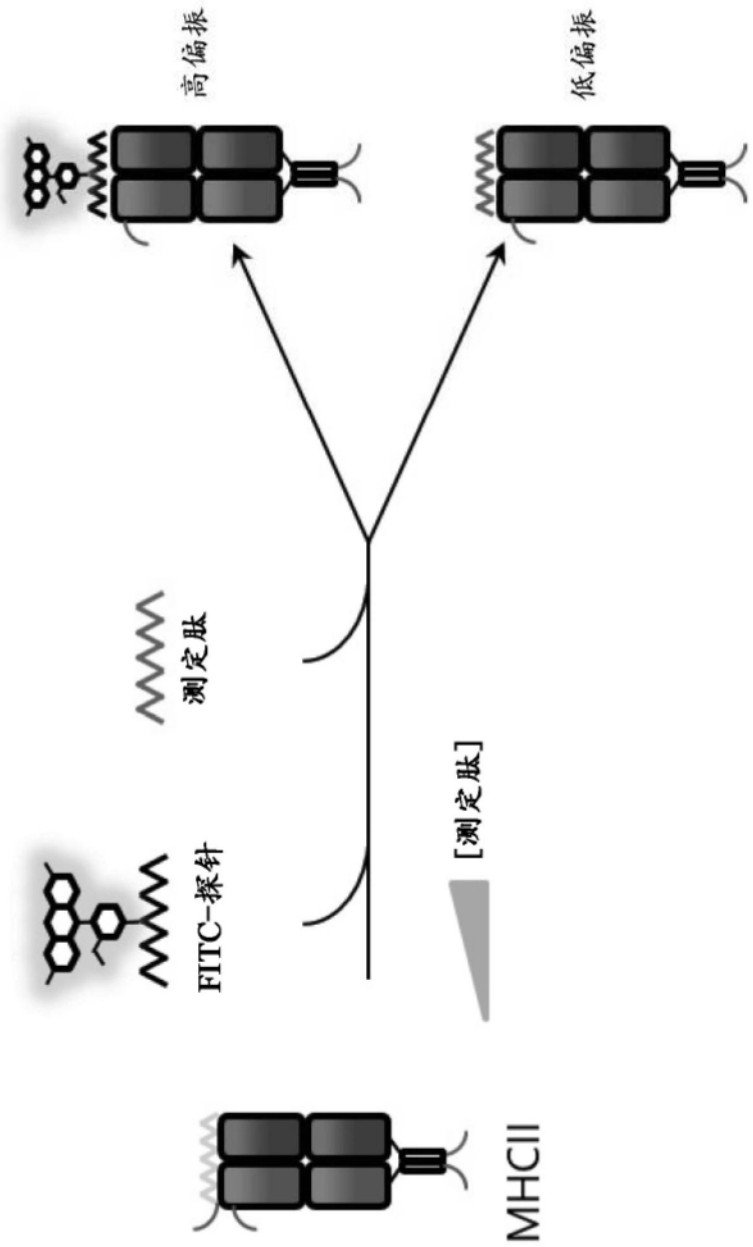


图28A

变量	评价的条件
pH	5.2 vs 7.4
去污剂	1%辛基葡糖苷 vs 1.6% NP40
温度	37°C vs 室温
探针	FITC-CLIP0 vs FITC-HA（超级结合物）
纯化方法	His标签 vs 构象特异性抗体(L243)
目标格式	可溶性 vs 全长等位基因

图28B

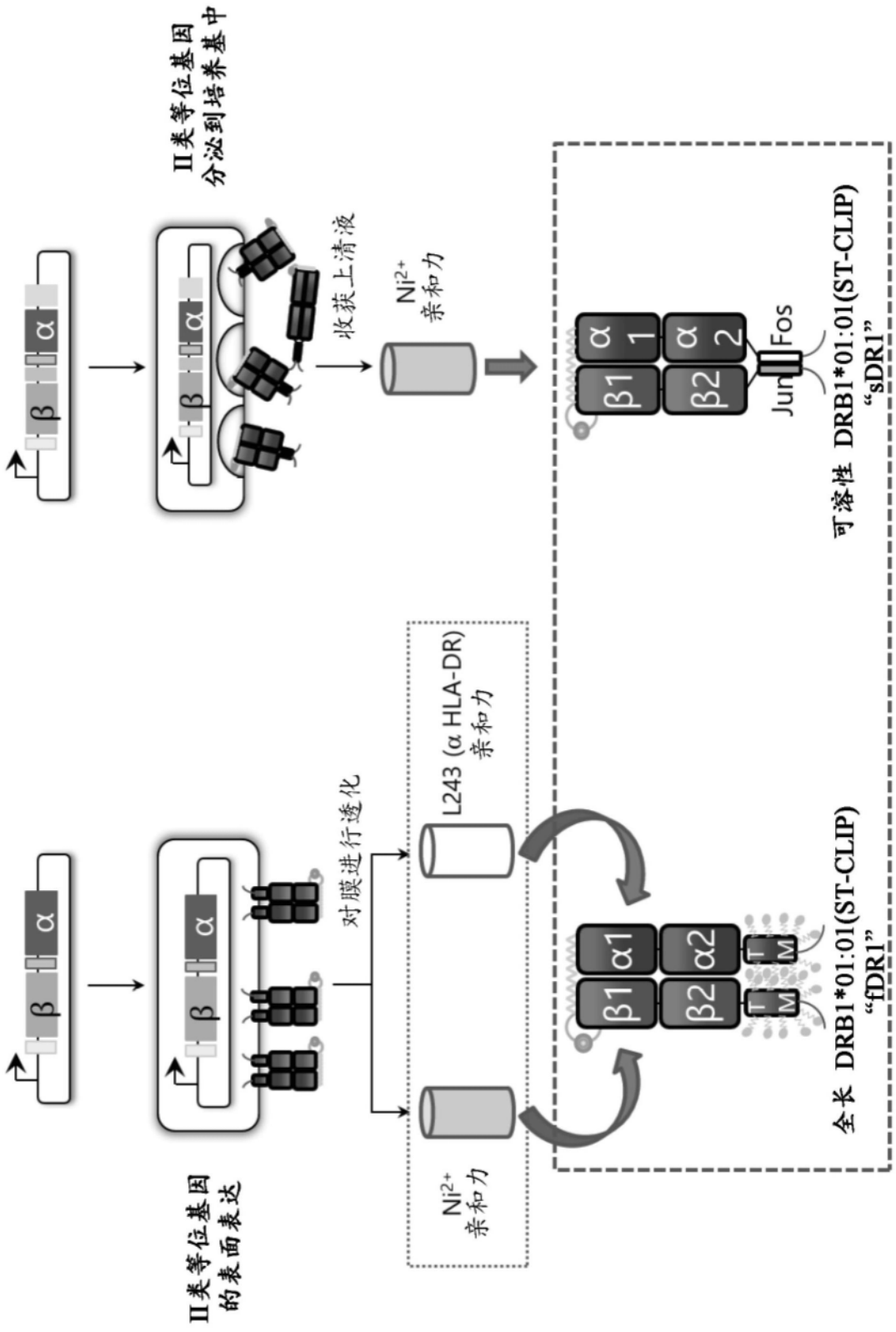
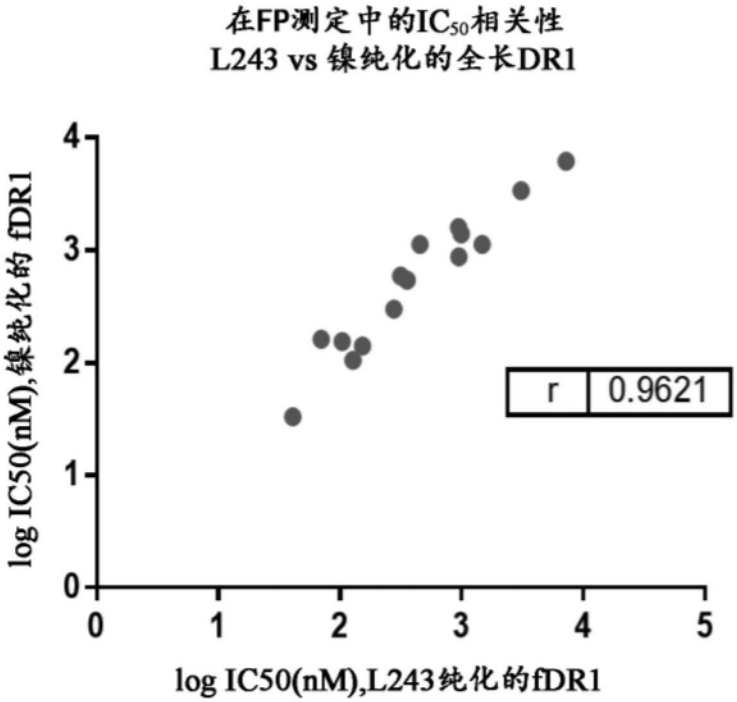
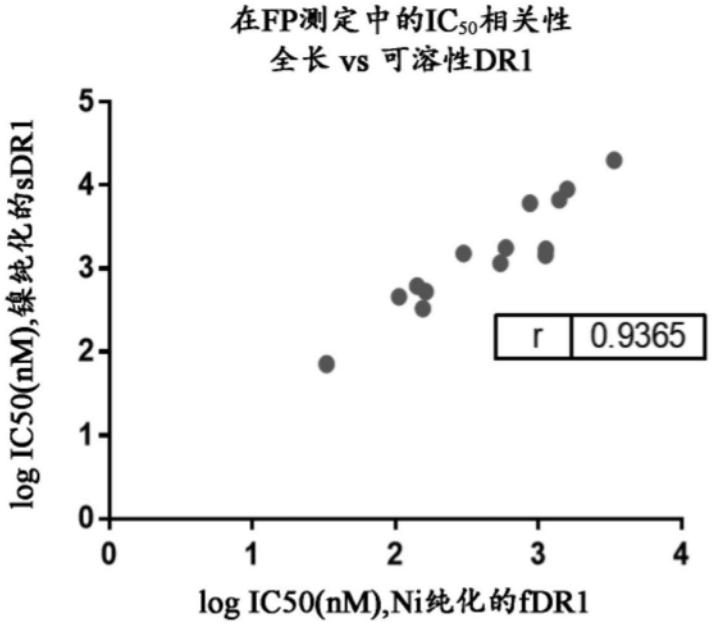

II类等位基因
分泌到培养基中
收获上清液
Ni2+
亲和力
α
β1
1
α
β2
2
Jun
Fos
可溶性 DRB1*01:01(ST-CLIP)
"sDR1"
II类等位基因
的表面表达
对膜进行透化
L243 (α HLA-DR)
亲和力
Ni2+
亲和力
β1
α1
β2
α2
TM
TM
全长 DRB1*01:01(ST-CLIP)
"fDR1"

图28C



	平均IC50, nM (n=2)		IC50变化倍数
肽	L234纯化的fDR1	Ni 纯化的fDR1	L234 fDR1至Ni fDR1
肽166	453	1128	2.5
肽165	70	163	2.3
肽157	315	591	1.9
肽163	941	1585	1.7
肽158	357	542	1.5
肽159	104	156	1.5
肽190	992	1398	1.4
肽162	3073	3384	1.1
肽164	278	300	1.1
肽161	946	875	0.9
肽167	153	142	0.9
肽189	7173	6132	0.9
肽160	128	106	0.8
肽191	1470	1124	0.8
HA(对照)	41	33	0.8

图28D



肽	平均IC50, nM (n=2) 全长DR1 (fDR1)	可溶性DR1 (sDR1)	IC50变化倍数 sDR1至 fDR1
肽161	875	6076	**6.9**
肽162	3384	20076	**5.9**
肽163	1585	8915	**5.6**
肽164	300	1517	**5.1**
肽190	1398	6695	**4.8**
肽160	106	457	**4.3**
肽167	142	614	**4.3**
肽165	163	529	**3.2**
肽157	591	1773	**3**
HA(对照)	33	72	**2.2**
肽158	542	1155	**2.1**
肽159	156	331	**2.1**
肽166	1128	1706	**1.5**
肽191	1124	1444	**1.3**
肽189	6132	>50,000	ND

图28E

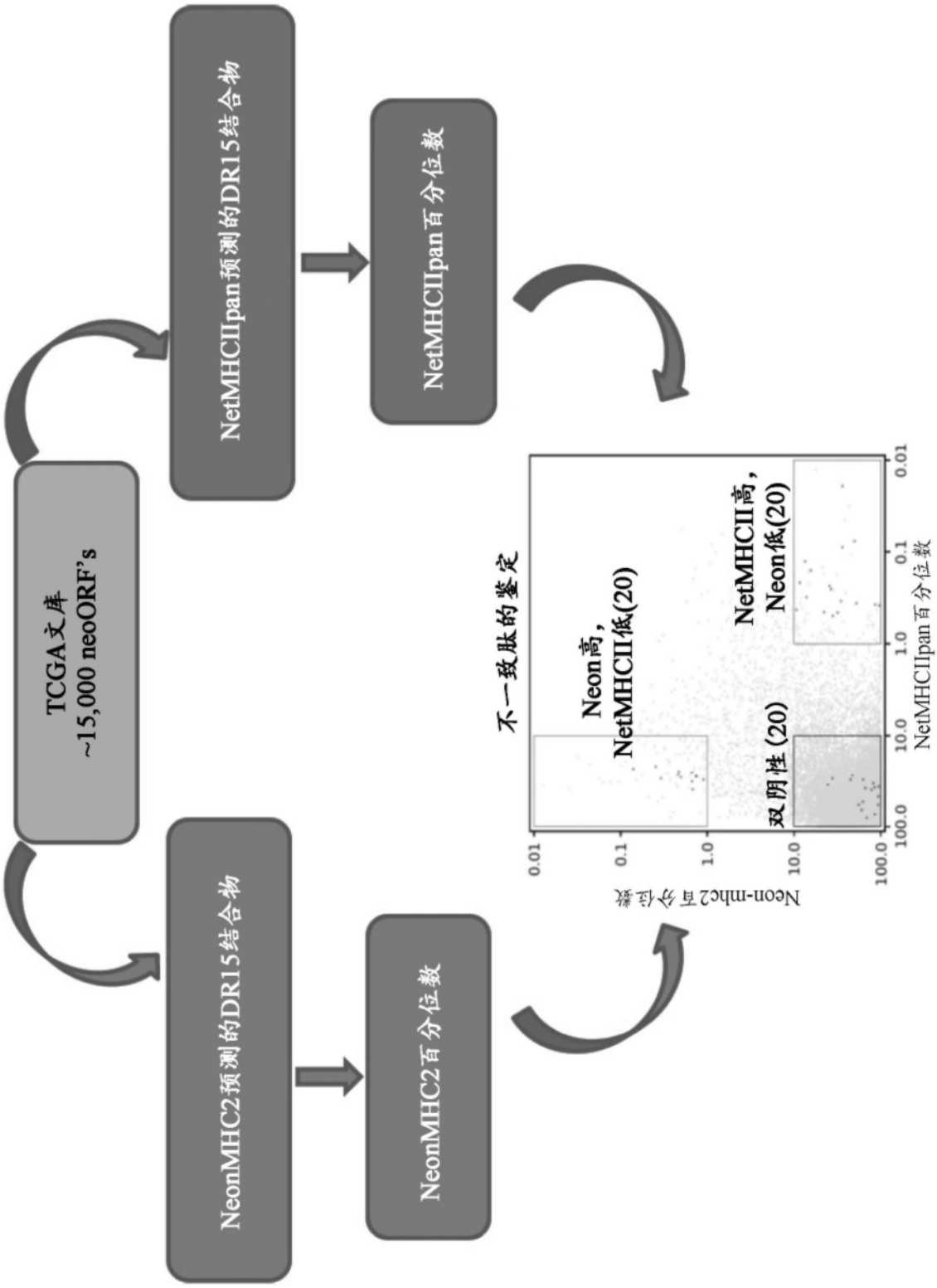
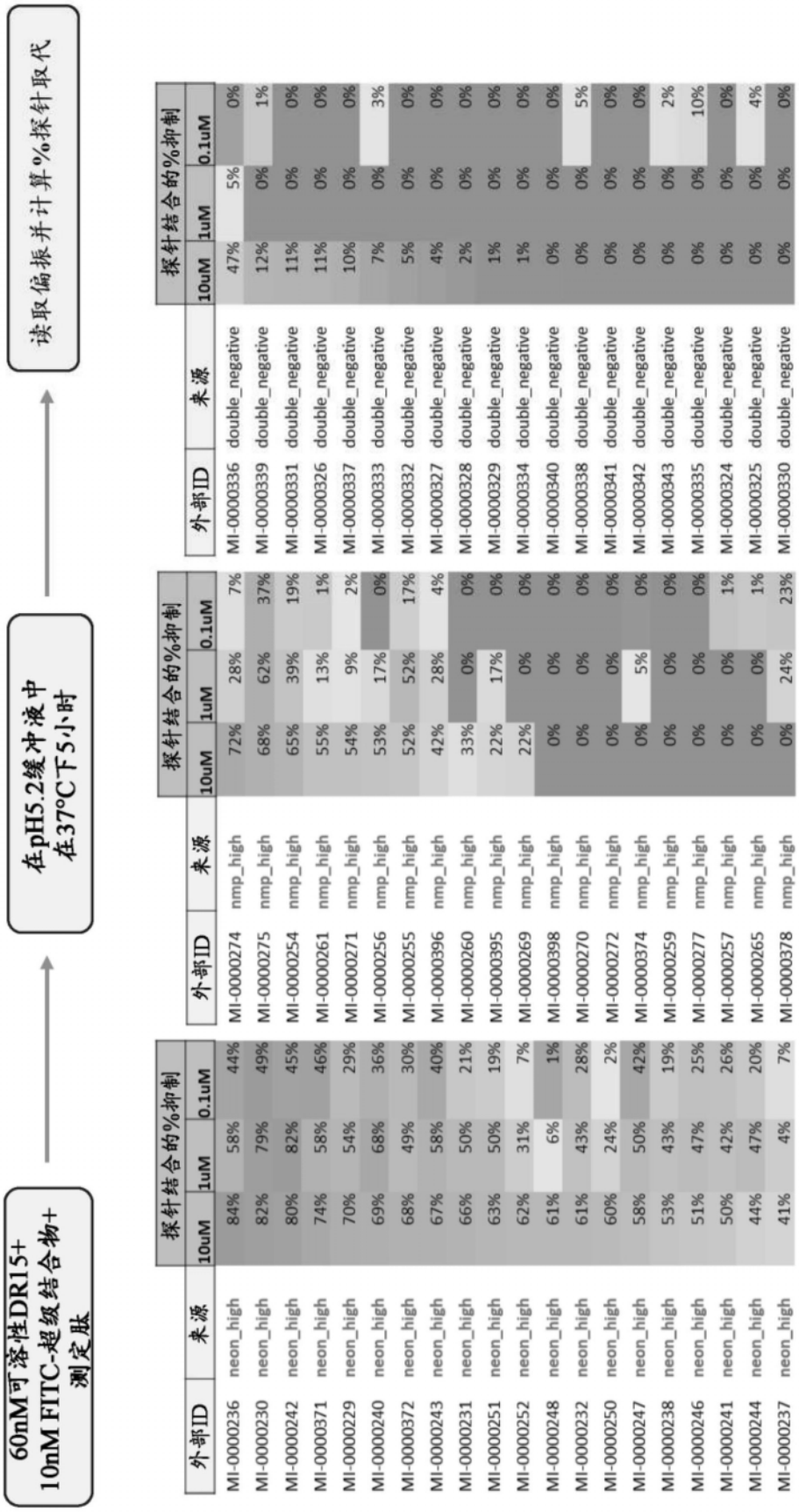

TCGA文库
~15,000 neoORF's
NeonMHC2预测的DR15结合物
NeonMHC2百分位数
NetMHCIIpan预测的DR15结合物
NetMHCIIpan百分位数
不一致肽的鉴定
Neon高，
NetMHCII低(20)
双阴性 (20)
NetMHCII高，
Neon低(20)
Neon-mhc2百分位数
NetMHCIIpan百分位数
0.01
0.1
1.0
10.0
100.0

图28F



外部ID	来源	探针结合的%抑制			外部ID	来源	探针结合的%抑制			外部ID	来源	探针结合的%抑制		
		10uM	1uM	0.1uM			10uM	1uM	0.1uM			10uM	1uM	0.1uM
MI-0000236	neon_high	84%	58%	44%	MI-0000274	nmp_high	72%	28%	7%	MI-0000336	double_negative	47%	5%	0%
MI-0000230	neon_high	82%	79%	49%	MI-0000275	nmp_high	68%	62%	37%	MI-0000339	double_negative	12%	0%	1%
MI-0000242	neon_high	80%	82%	45%	MI-0000254	nmp_high	65%	39%	19%	MI-0000331	double_negative	11%	0%	0%
MI-0000371	neon_high	74%	58%	46%	MI-0000261	nmp_high	55%	13%	1%	MI-0000326	double_negative	11%	0%	0%
MI-0000229	neon_high	70%	54%	29%	MI-0000271	nmp_high	54%	9%	2%	MI-0000337	double_negative	10%	0%	0%
MI-0000240	neon_high	69%	68%	36%	MI-0000256	nmp_high	53%	17%	0%	MI-0000333	double_negative	7%	0%	3%
MI-0000372	neon_high	68%	49%	30%	MI-0000255	nmp_high	52%	52%	17%	MI-0000332	double_negative	5%	0%	0%
MI-0000243	neon_high	67%	58%	40%	MI-0000396	nmp_high	42%	28%	4%	MI-0000327	double_negative	4%	0%	0%
MI-0000231	neon_high	66%	50%	21%	MI-0000260	nmp_high	33%	0%	0%	MI-0000328	double_negative	2%	0%	0%
MI-0000251	neon_high	63%	50%	19%	MI-0000395	nmp_high	22%	17%	0%	MI-0000329	double_negative	1%	0%	0%
MI-0000252	neon_high	62%	31%	7%	MI-0000269	nmp_high	22%	0%	0%	MI-0000334	double_negative	1%	0%	0%
MI-0000248	neon_high	61%	6%	1%	MI-0000398	nmp_high	0%	0%	0%	MI-0000340	double_negative	0%	0%	0%
MI-0000232	neon_high	61%	43%	28%	MI-0000270	nmp_high	0%	0%	0%	MI-0000338	double_negative	0%	0%	5%
MI-0000250	neon_high	60%	24%	2%	MI-0000272	nmp_high	0%	0%	0%	MI-0000341	double_negative	0%	0%	0%
MI-0000247	neon_high	58%	50%	42%	MI-0000374	nmp_high	0%	5%	0%	MI-0000342	double_negative	0%	0%	0%
MI-0000238	neon_high	53%	43%	19%	MI-0000259	nmp_high	0%	0%	0%	MI-0000343	double_negative	0%	0%	2%
MI-0000246	neon_high	51%	47%	25%	MI-0000277	nmp_high	0%	0%	0%	MI-0000335	double_negative	0%	0%	10%
MI-0000241	neon_high	50%	42%	26%	MI-0000257	nmp_high	0%	0%	1%	MI-0000324	double_negative	0%	0%	0%
MI-0000244	neon_high	44%	47%	20%	MI-0000265	nmp_high	0%	0%	1%	MI-0000325	double_negative	0%	0%	4%
MI-0000237	neon_high	41%	4%	7%	MI-0000378	nmp_high	0%	24%	23%	MI-0000330	double_negative	0%	0%	0%

图28G

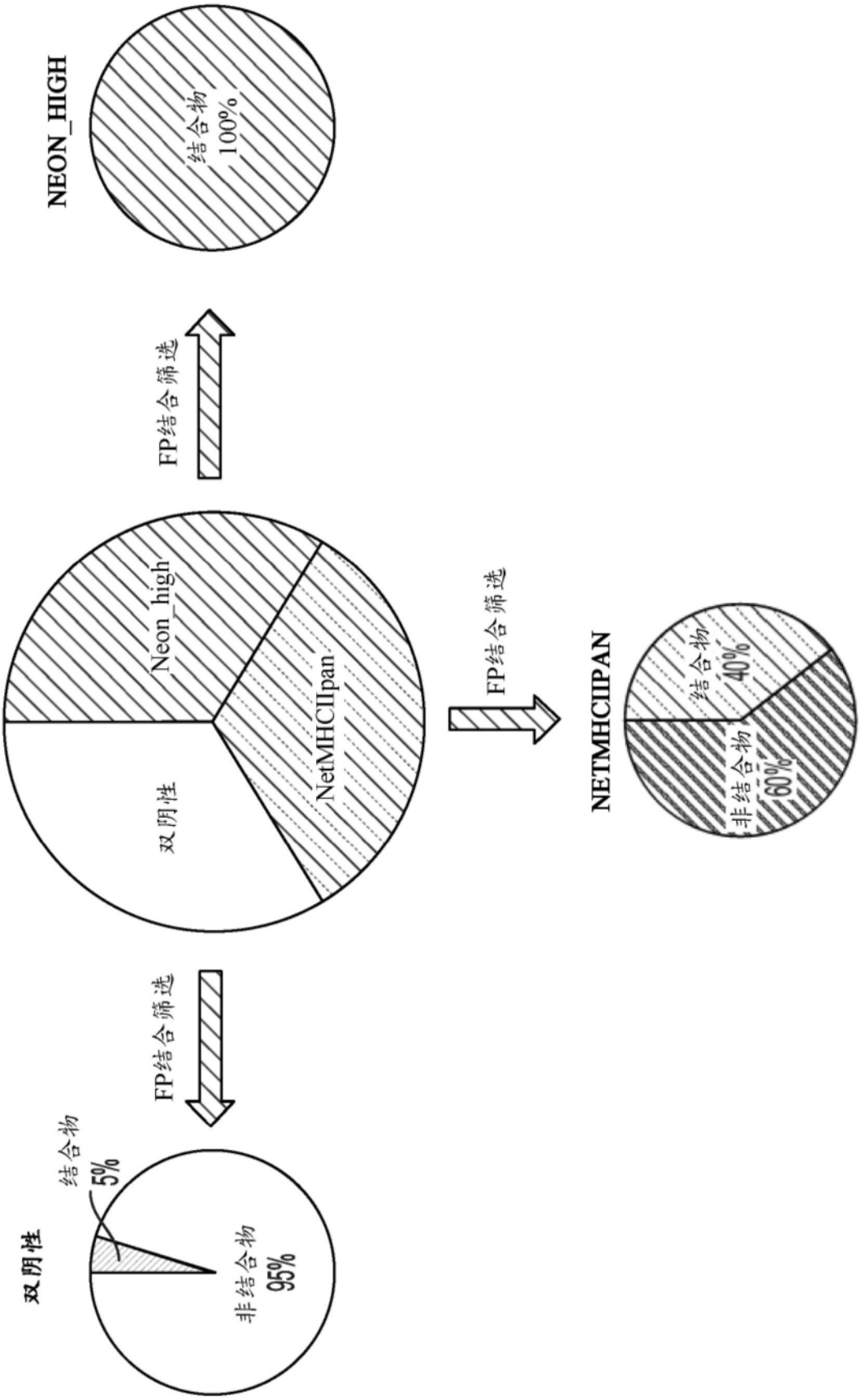

双阴性
结合物
5%
非结合物
95%
FP结合筛选
双阴性
Neon_high
NetMHCIIpan
FP结合筛选
NEON_HIGH
结合物
100%
FP结合筛选
NETMHCIIPAN
结合物
40%
非结合物
60%

图28H

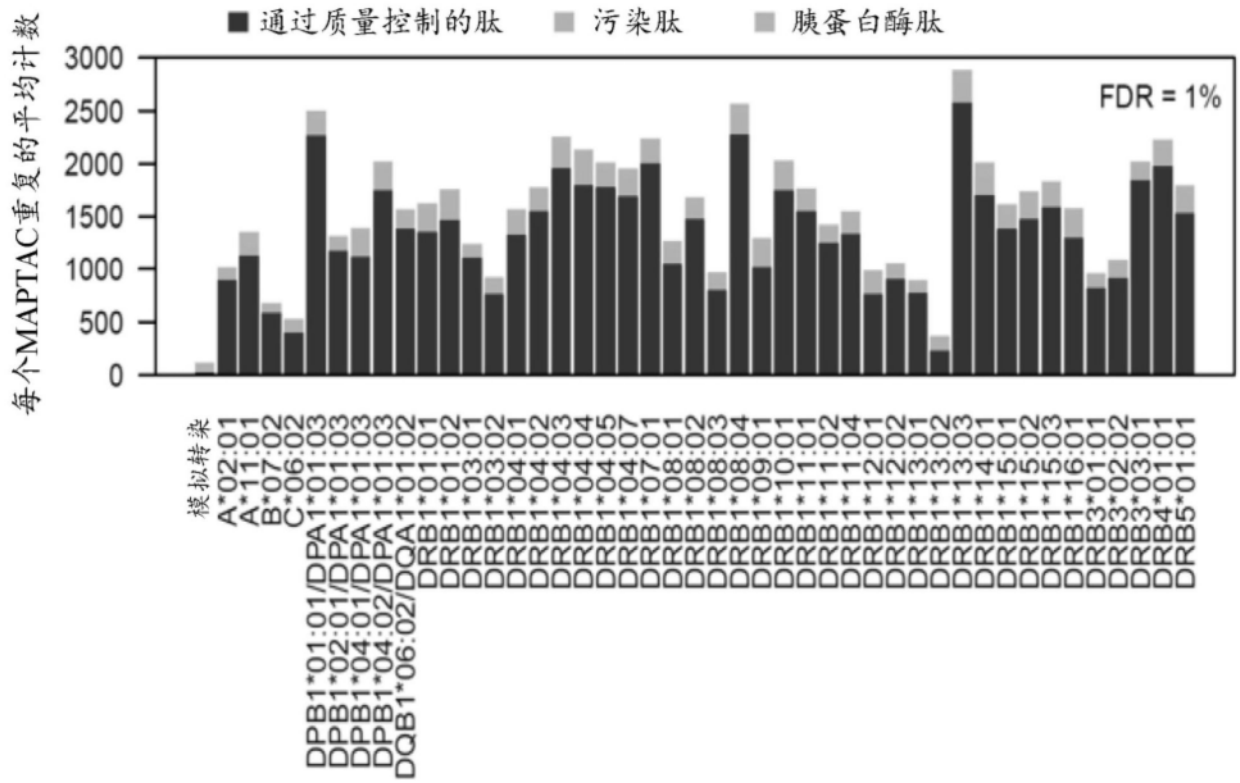

通过质量控制的肽
污染肽
胰蛋白酶肽
每个MAPTAC重复的平均计数
3000
2500
2000
1500
1000
500
0
FDR = 1%
模拟转染
A*02:01
A*11:01
B*07:02
C*06:02
DPB1*01:01/DPA1*01:03
DPB1*02:01/DPA1*01:03
DPB1*04:01/DPA1*01:03
DPB1*04:02/DPA1*01:03
DQB1*06:02/DQA1*01:02
DRB1*01:01
DRB1*01:02
DRB1*03:01
DRB1*03:02
DRB1*04:01
DRB1*04:02
DRB1*04:03
DRB1*04:04
DRB1*04:05
DRB1*04:07
DRB1*07:01
DRB1*08:01
DRB1*08:02
DRB1*08:03
DRB1*08:04
DRB1*09:01
DRB1*10:01
DRB1*11:01
DRB1*11:02
DRB1*11:04
DRB1*12:01
DRB1*12:02
DRB1*13:01
DRB1*13:02
DRB1*13:03
DRB1*14:01
DRB1*15:01
DRB1*15:02
DRB1*15:03
DRB1*16:01
DRB3*01:01
DRB3*02:02
DRB3*03:01
DRB4*01:01
DRB5*01:01

图29

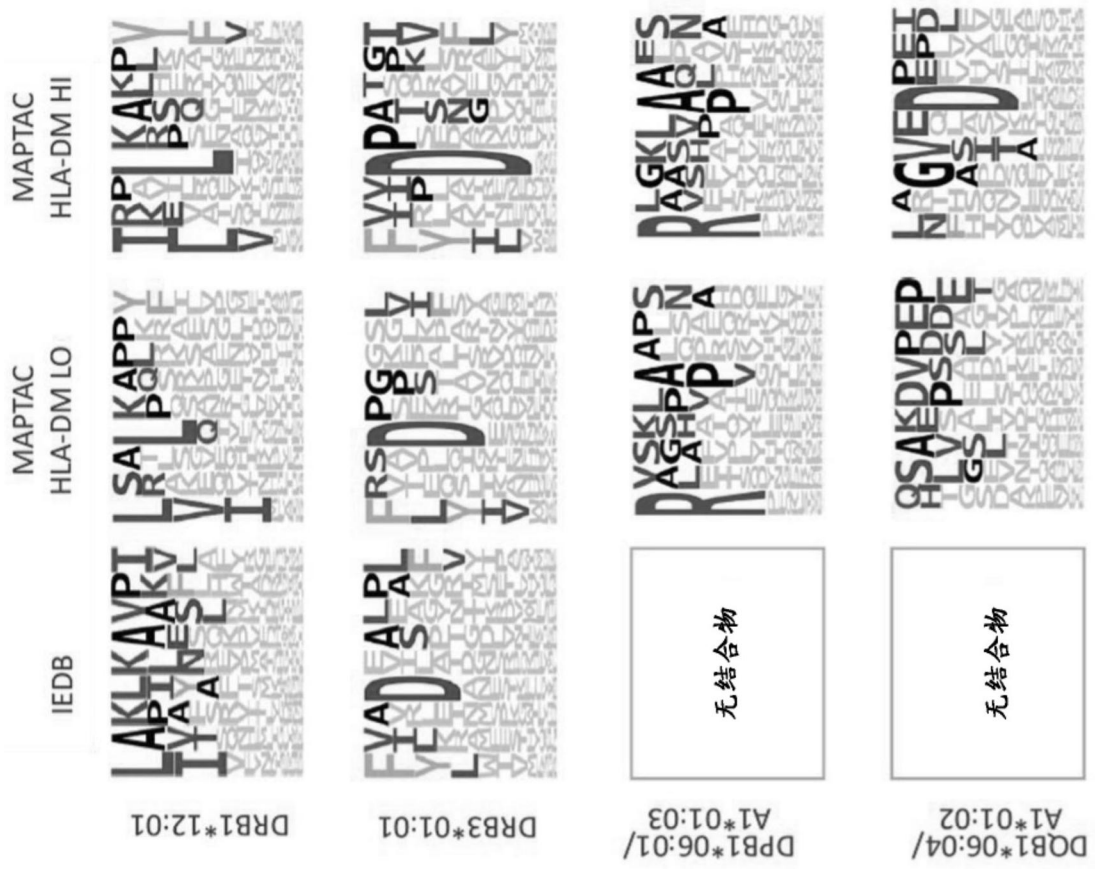

MAPTAC HLA-DM HI
MAPTAC HLA-DM LO
IEDB
DRB1*12:01
DRB3*01:01
DPB1*06:01/ A1*01:03
DQB1*06:04/ A1*01:02
无结合物
无结合物

图30A

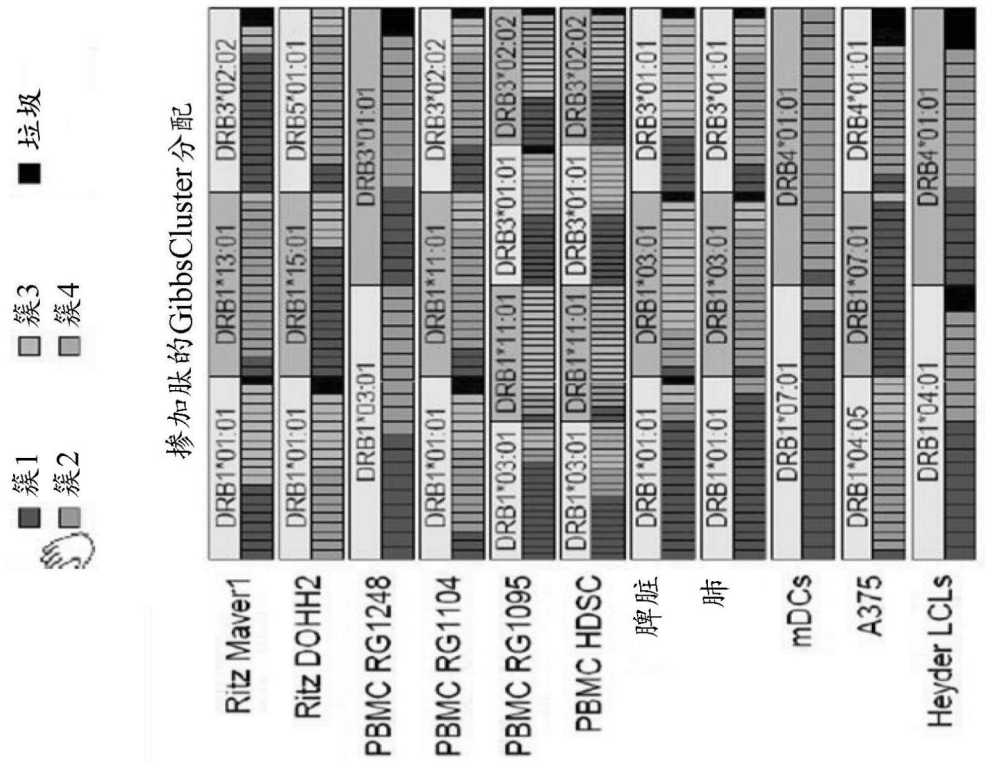

簇1
簇2
簇3
簇4
垃圾
掺加肽的GibbsCluster分配
DRB1*01:01
DRB1*13:01
DRB3*02:02
DRB1*01:01
DRB1*15:01
DRB5*01:01
DRB1*03:01
DRB3*01:01
DRB1*01:01
DRB1*11:01
DRB3*02:02
DRB1*03:01
DRB1*11:01
DRB3*01:01
DRB1*03:01
DRB1*11:01
DRB3*01:01
DRB3*02:02
DRB1*01:01
DRB1*03:01
DRB3*01:01
DRB1*01:01
DRB1*03:01
DRB3*01:01
DRB1*07:01
DRB4*01:01
DRB1*04:05
DRB1*07:01
DRB4*01:01
DRB1*04:01
DRB4*01:01
Ritz Maver1
Ritz DOHH2
PBMC RG1248
PBMC RG1104
PBMC RG1095
PBMC HDSC
脾脏
肺
mDCs
A375
Heyder LCLs

图30B

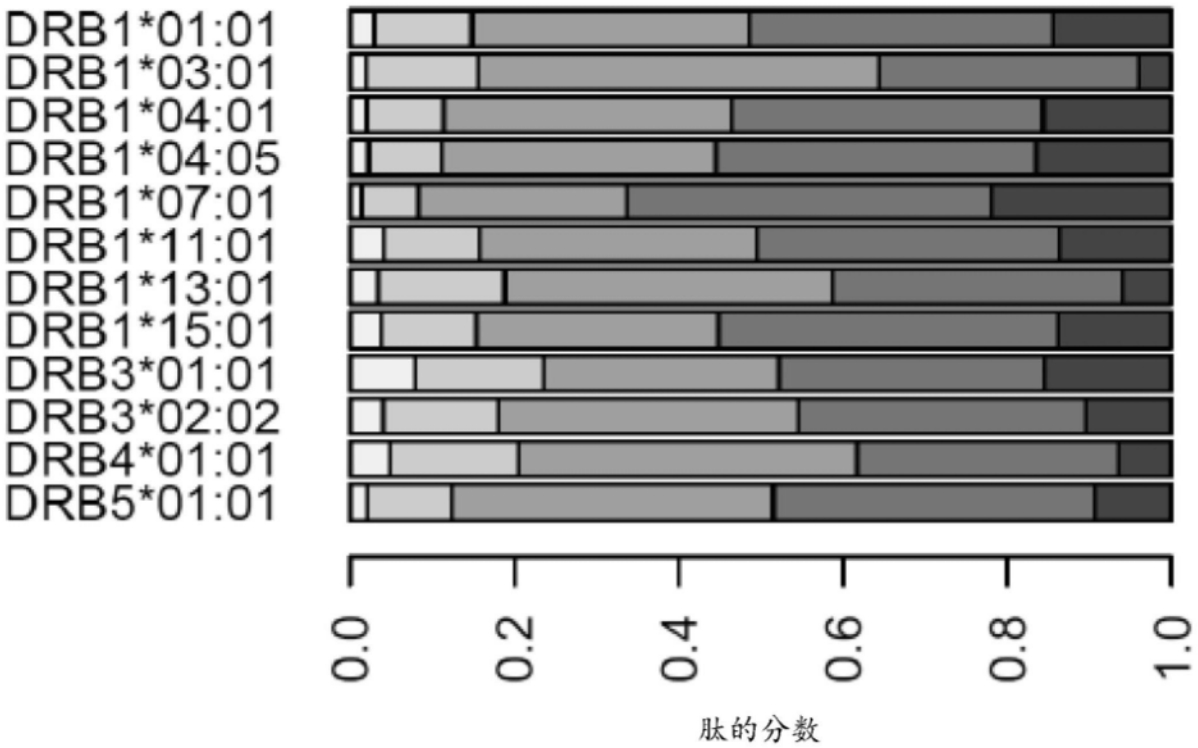

DRB1*01:01
DRB1*03:01
DRB1*04:01
DRB1*04:05
DRB1*07:01
DRB1*11:01
DRB1*13:01
DRB1*15:01
DRB3*01:01
DRB3*02:02
DRB4*01:01
DRB5*01:01
0.0
0.2
0.4
0.6
0.8
1.0
肽的分数

图30C

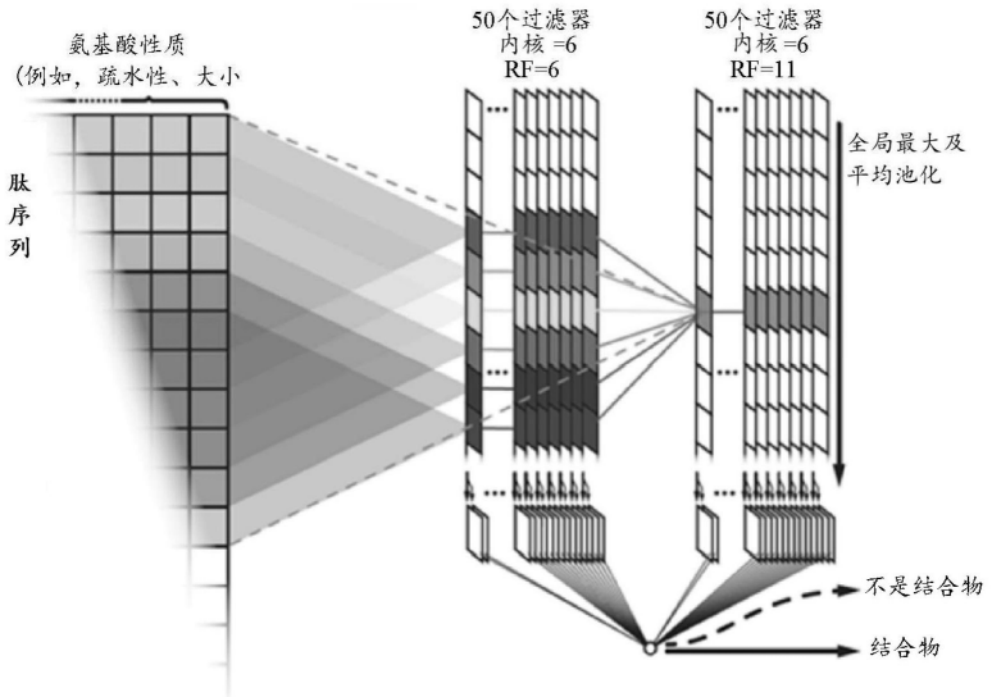

氨基酸性质
(例如，疏水性、大小
肽
序
列
50个过滤器
内核 =6
RF=6
50个过滤器
内核 =6
RF=11
全局最大及
平均池化
不是结合物
结合物

图31A

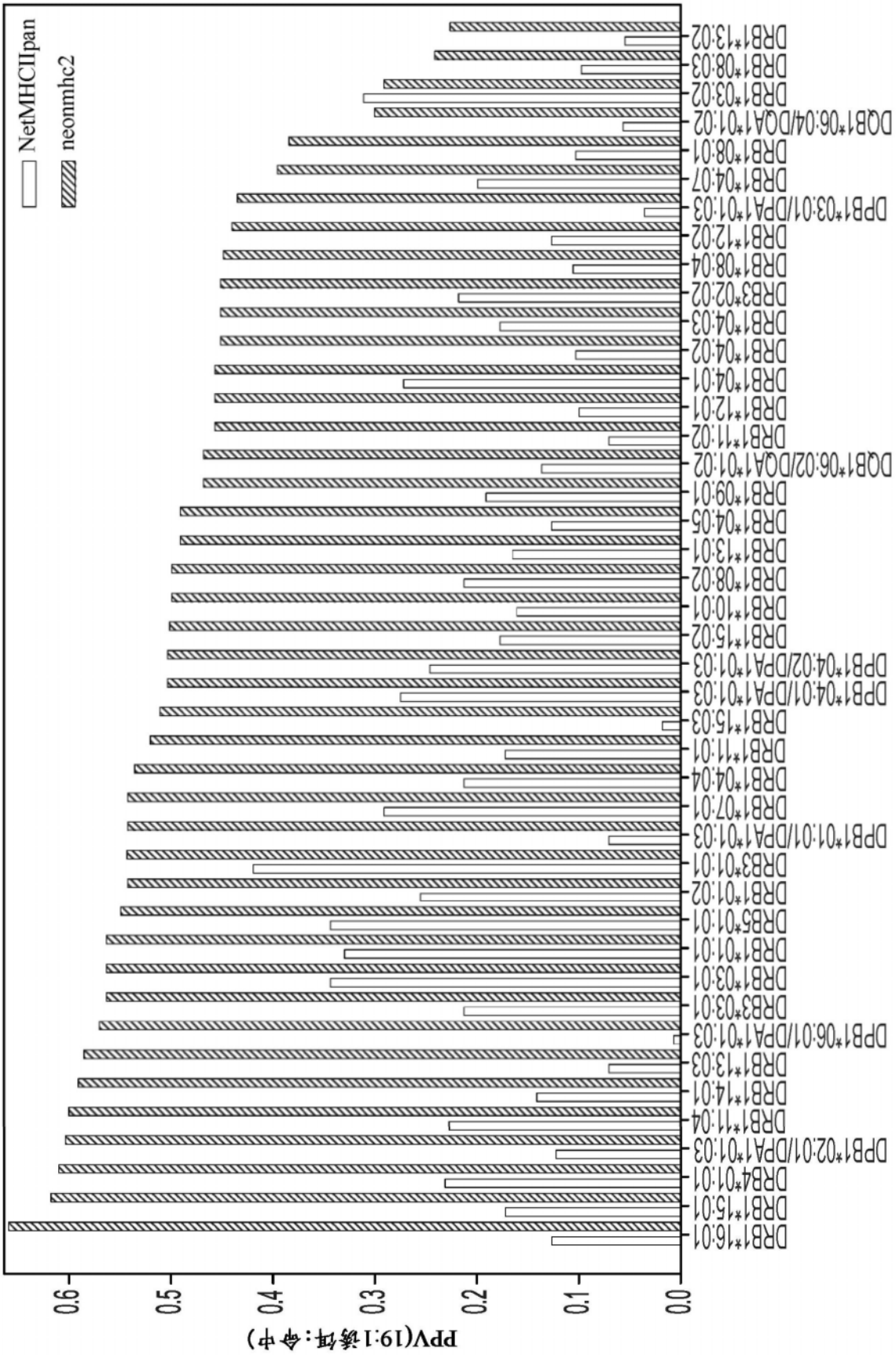

NetMHCIIpan
neonmhc2
PPV(19:1诱饵:命中)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
DRB1*16:01
DRB1*15:01
DRB4*01:01
DPB1*02:01/DPA1*01:03
DRB1*11:04
DRB1*14:01
DRB1*13:03
DPB1*06:01/DPA1*01:03
DRB3*03:01
DRB1*03:01
DRB1*01:01
DRB5*01:01
DRB1*01:02
DRB3*01:01
DPB1*01:01/DPA1*01:03
DRB1*07:01
DRB1*04:04
DRB1*11:01
DRB1*15:03
DPB1*04:01/DPA1*01:03
DPB1*04:02/DPA1*01:03
DRB1*15:02
DRB1*10:01
DRB1*08:02
DRB1*13:01
DRB1*04:05
DRB1*09:01
DQB1*06:02/DQA1*01:02
DRB1*11:02
DRB1*12:01
DRB1*04:01
DRB1*04:02
DRB1*04:03
DRB3*02:02
DRB1*08:04
DRB1*12:02
DPB1*03:01/DPA1*01:03
DRB1*04:07
DRB1*08:01
DQB1*06:04/DQA1*01:02
DRB1*03:02
DRB1*08:03
DRB1*13:02

图31B

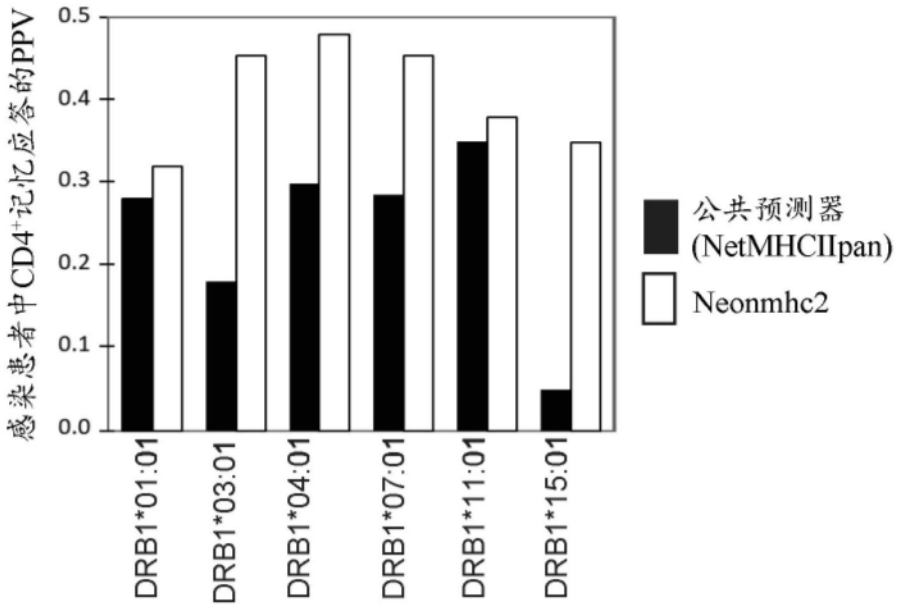

感染患者中CD4+记忆应答的PPV
0.5
0.4
0.3
0.2
0.1
0.0
DRB1*01:01
DRB1*03:01
DRB1*04:01
DRB1*07:01
DRB1*11:01
DRB1*15:01
公共预测器
(NetMHCIIpan)
Neonmhc2

图31C

新抗原肽	NetMHCIIpan 亲和力 (nM)	NetMHCIIpan %排名	neonmhc2 %排名	状态
GEGLFQPAHRYPDAG	744	39.00%	0.43%	命中
RNQWKCLGKPVGAEM	426	29.00%	0.10%	测定的
VPAWTRAWRNSSPKG	395	27.00%	0.51%	命中
DGGSYFSLWKIWTQV	354	26.00%	0.11%	命中
LHPELLPLWRLLPDG	367	26.00%	0.23%	命中
PEASLYGALSKGSGG	353	26.00%	0.24%	命中
HEGGFPPLLRRAAED	325	24.00%	0.63%	命中
LSPVWCLQWKLSGTD	293	23.00%	0.31%	命中
KNTIVYTTKQVQSCQ	259	21.00%	0.57%	测定的
PATYILILKEFCLVG	246	20.00%	0.32%	测定的
NNEQFQWKIRHVGPE	222	19.00%	0.50%	命中
YKGGYELVKKSQTEL	200	17.00%	0.37%	测定的

图31D

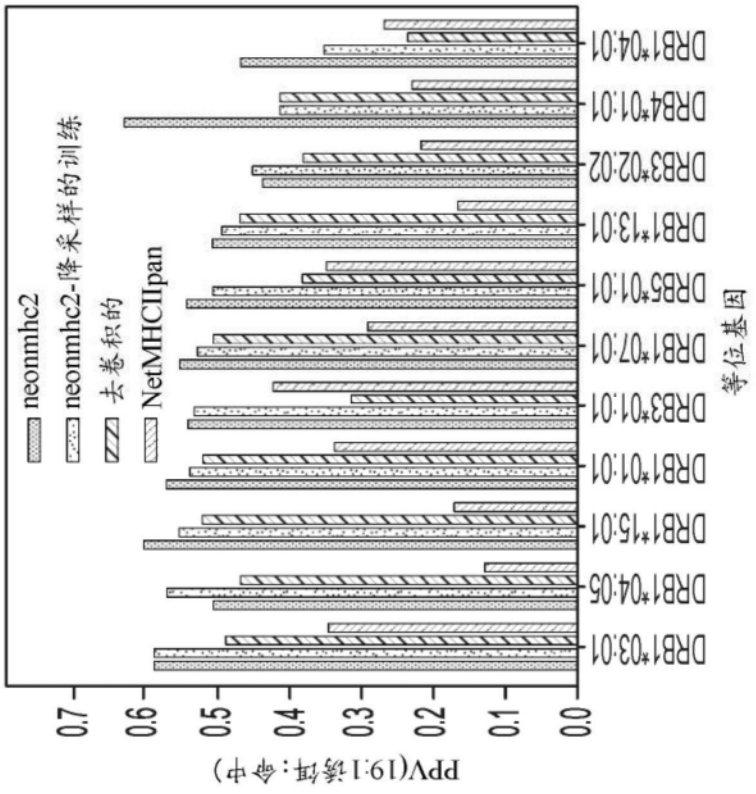

neonmhc2
neonmhc2-降采样的训练
去卷积的
NetMHCIIpan
DRB1*03:01
DRB1*04:05
DRB1*15:01
DRB1*01:01
DRB3*01:01
DRB1*07:01
DRB5*01:01
DRB1*13:01
DRB3*02:02
DRB4*01:01
DRB1*04:01
等位基因
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
PPV(19:1诱饵:命中)

图31E

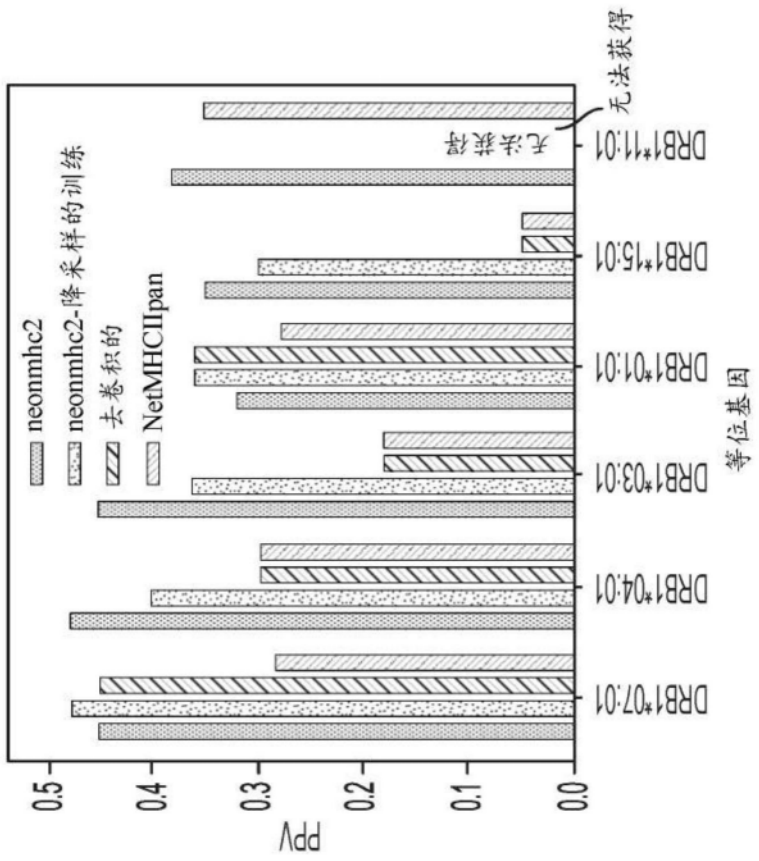

neonmhc2
neonmhc2-降采样的训练
去卷积的
NetMHCIIpan
PPV
0.5
0.4
0.3
0.2
0.1
0.0
DRB1*07:01
DRB1*04:01
DRB1*03:01
DRB1*01:01
DRB1*15:01
DRB1*11:01
无法获得
等位基因

图31F

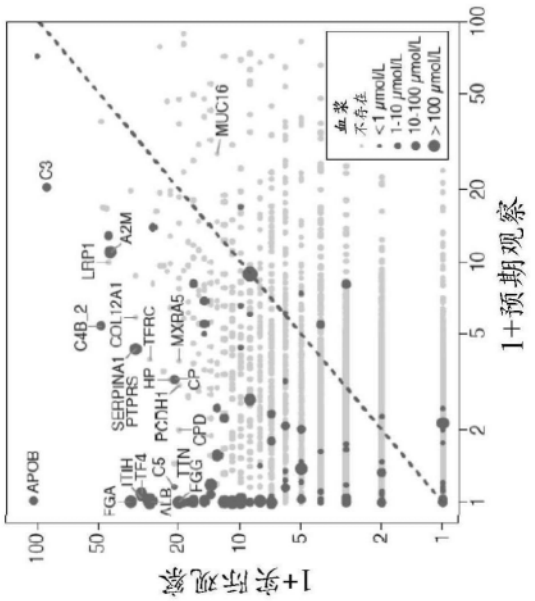

1+实际观察
1+预期观察
血浆
不存在
<1 μmol/L
1-10 μmol/L
10-100 μmol/L
>100 μmol/L
APOB
C3
A2M
LRP1
C4B_2
COL12A1
TFRC
MXRA5
SERPINA1
PTPRS
HP
PCDH1
CP
CPD
MUC16
ITIH4
TF
C6
ALB
TTN
FGG
FGA

图32A

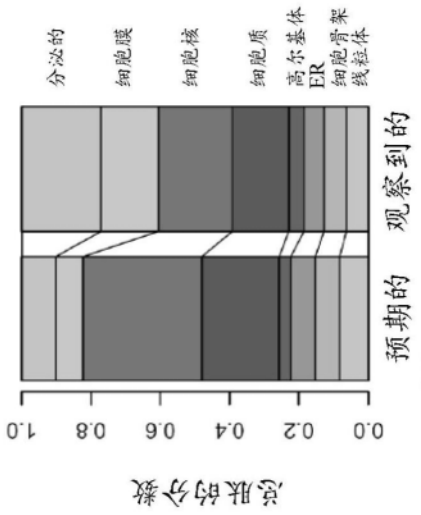

分泌的
细胞膜
细胞核
细胞质
高尔基体
ER
细胞骨架
线粒体
预期的 观察到的
0.0
0.2
0.4
0.6
0.8
1.0
总肽的分数

图32B

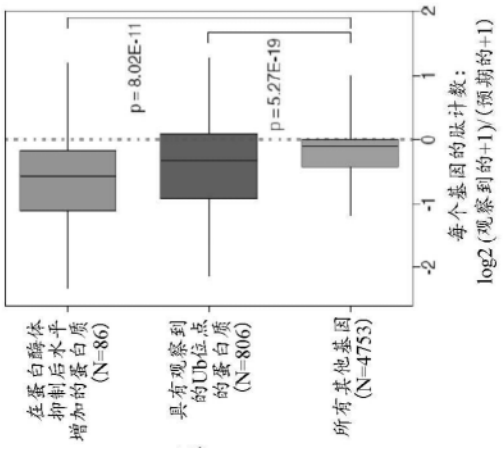

p=8.02E-11
p=5.27E-19
2
1
0
-1
-2
每个基因的肽计数：
log2(观察到的+1)/(预期的+1)
在蛋白酶体抑制后水平增加的蛋白质 (N=86)
具有观察到的Ub位点的蛋白质 (N=806)
所有其他基因 (N=4753)

图32C

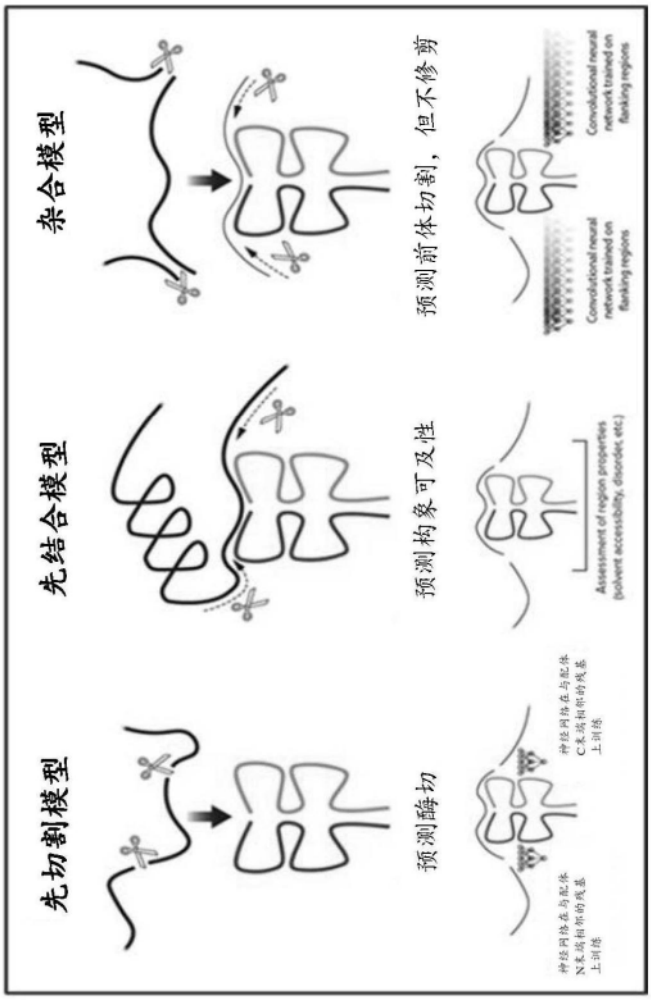

先切割模型
先结合模型
杂合模型
预测酶切
预测构象可及性
预测前体切割，但不修剪
神经网络在与配体N末端相邻的残基上训练
神经网络在与配体C末端相邻的残基上训练
Assessment of region properties (solvent accessibility, disorder, etc.)
Convolutional neural network trained on flanking regions
Convolutional neural network trained on flanking regions

图32D

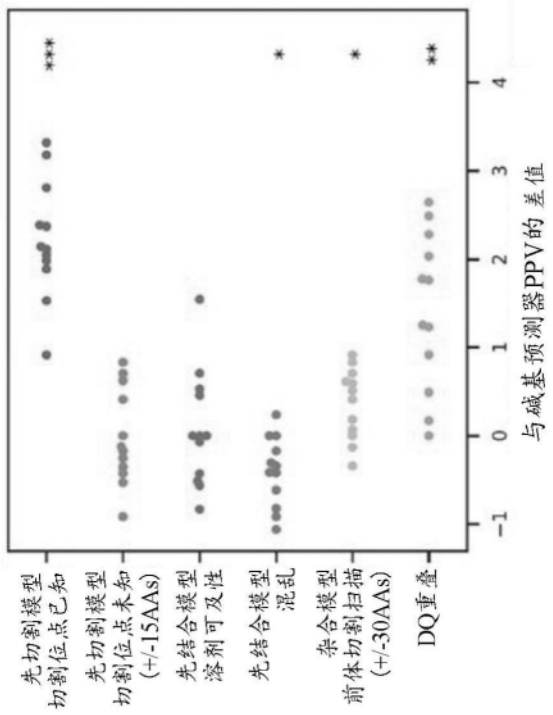

先切割模型
切割位点已知
先切割模型
切割位点未知
(+/-15AAs)
先结合模型
溶剂可及性
先结合模型
混乱
杂合模型
前体切割扫描
(+/-30AAs)
DQ重叠
***
*
*
**
−1
0
1
2
3
4
与碱基预测器PPV的差值

图32E

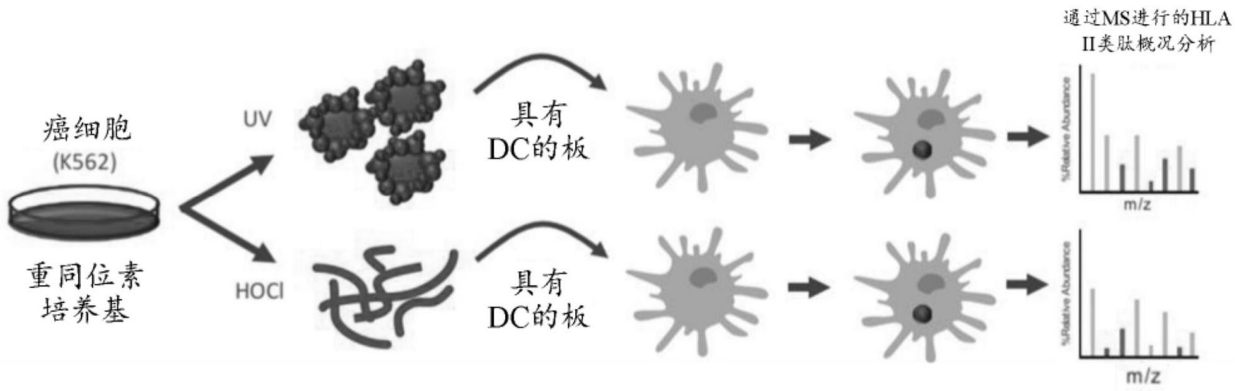

癌细胞
(K562)
重同位素
培养基
UV
HOCl
具有
DC的板
具有
DC的板
通过MS进行的HLA
II类肽概况分析
%Relative Abundance
m/z
%Relative Abundance
m/z

图33A

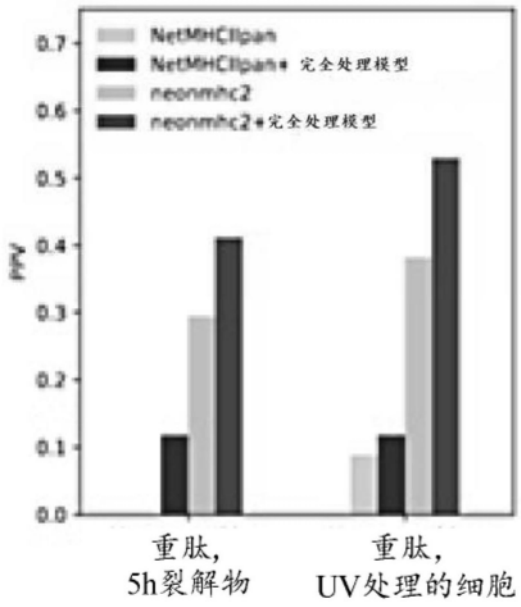

NetMHCIIpan
NetMHCIIpan* 完全处理模型
neonmhc2
neonmhc2*完全处理模型
PPV
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
重肽，
5h裂解物
重肽，
UV处理的细胞

图33B

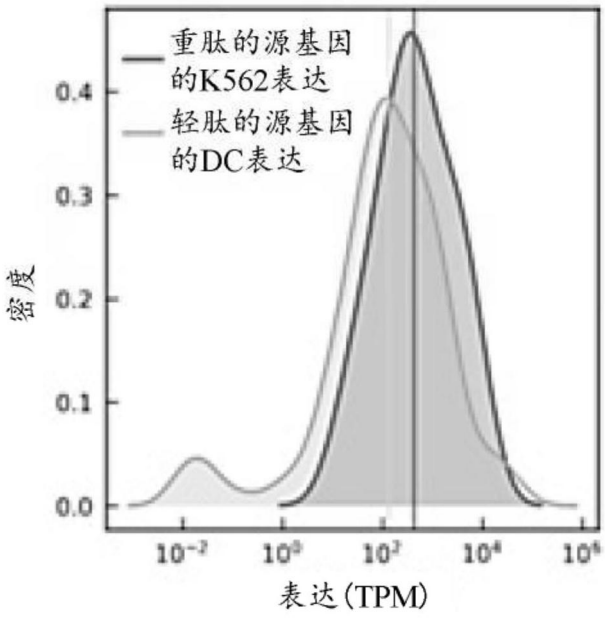

重肽的源基因
的K562表达
轻肽的源基因
的DC表达
密度
0.0
0.1
0.2
0.3
0.4
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}
表达(TPM)

图33C

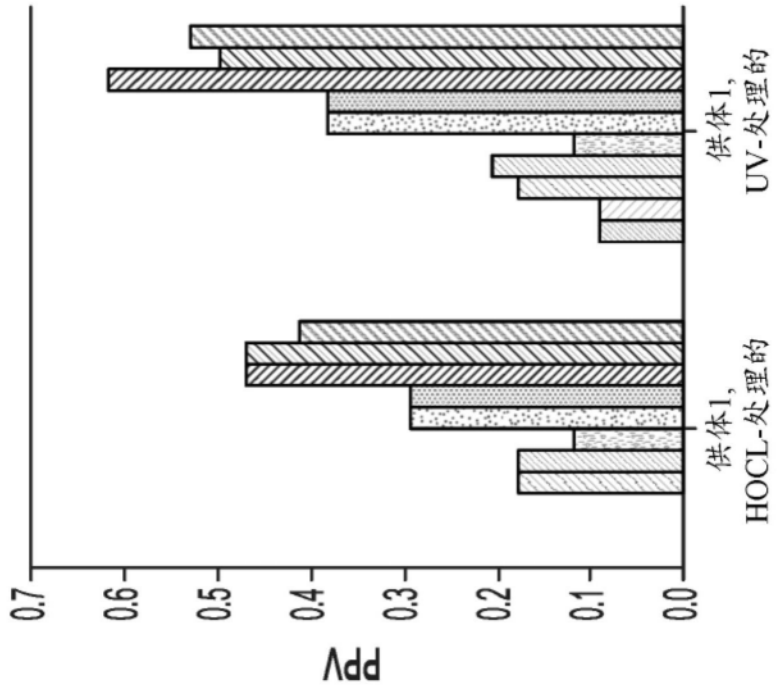

PPV
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
供体1,
HOCL-处理的
供体1,
UV-处理的

图33D

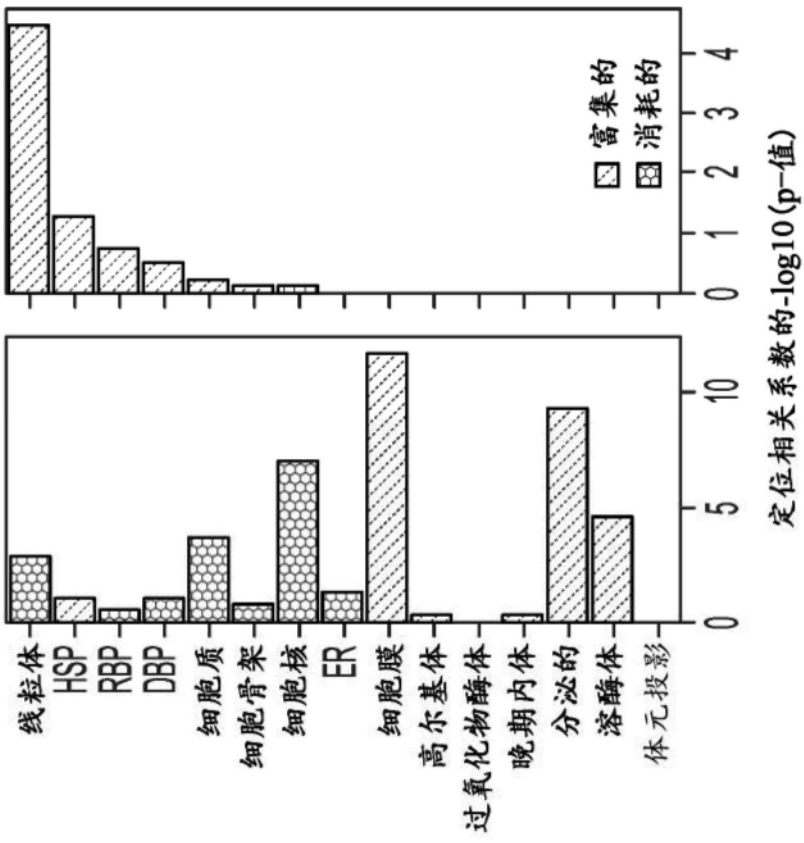

线粒体
HSP
RBP
DBP
细胞质
细胞骨架
细胞核
ER
细胞膜
高尔基体
过氧化物酶体
晚期内体
分泌的
溶酶体
体元投影
0
5
10
0
1
2
3
4
富集的
消耗的
定位相关系数的-log10(p-值)

图33E

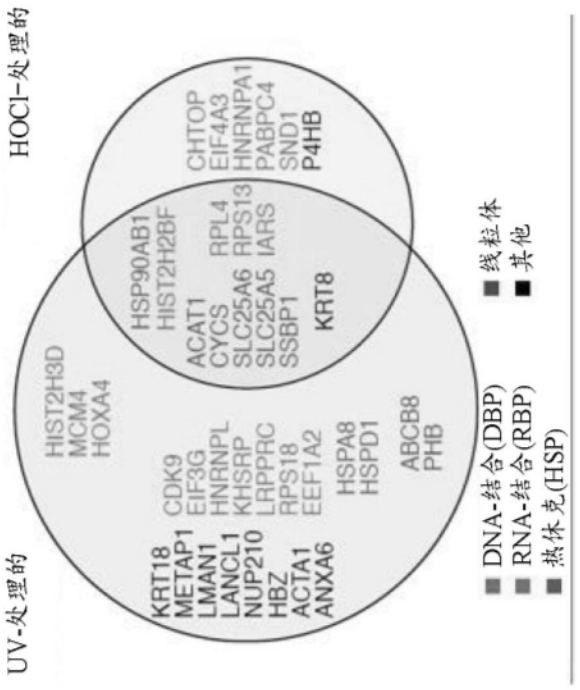

HOCl-处理的
UV-处理的
HIST2H3D
MCM4
HOXA4
CDK9
EIF3G
HNRNPL
KHSRP
LRPPRC
RPS18
EEF1A2
HSPA8
HSPD1
ABCB8
PHB
KRT18
METAP1
LMAN1
LANCL1
NUP210
HBZ
ACTA1
ANXA6
HSP90AB1
HIST2H2BF
ACAT1
CYCS
SLC25A6
SLC25A5
SSBP1
KRT8
RPL4
RPS13
IARS
CHTOP
EIF4A3
HNRNPA1
PABPC4
SND1
P4HB
线粒体
共他
DNA-结合(DBP)
RNA-结合(RBP)
热休克(HSP)

图33F

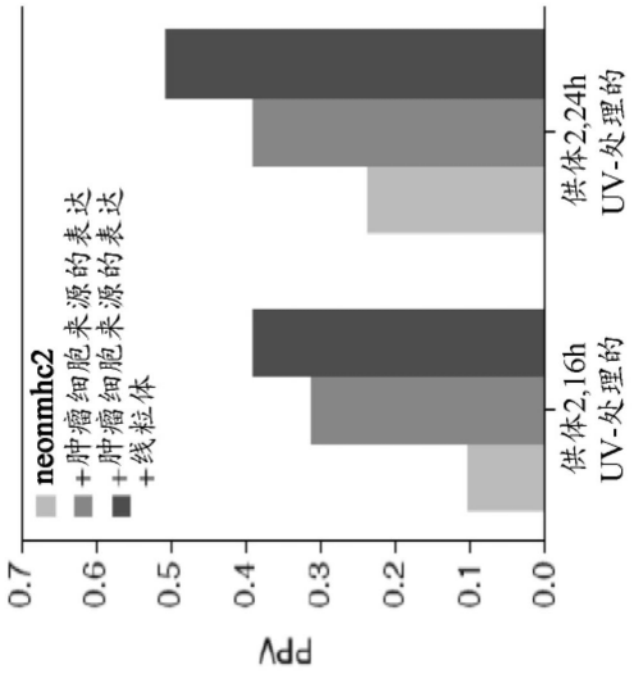

neonmhc2
+肿瘤细胞来源的表达
+肿瘤细胞来源的表达
+线粒体
PPV
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
供体2,16h
UV-处理的
供体2,24h
UV-处理的

图33G

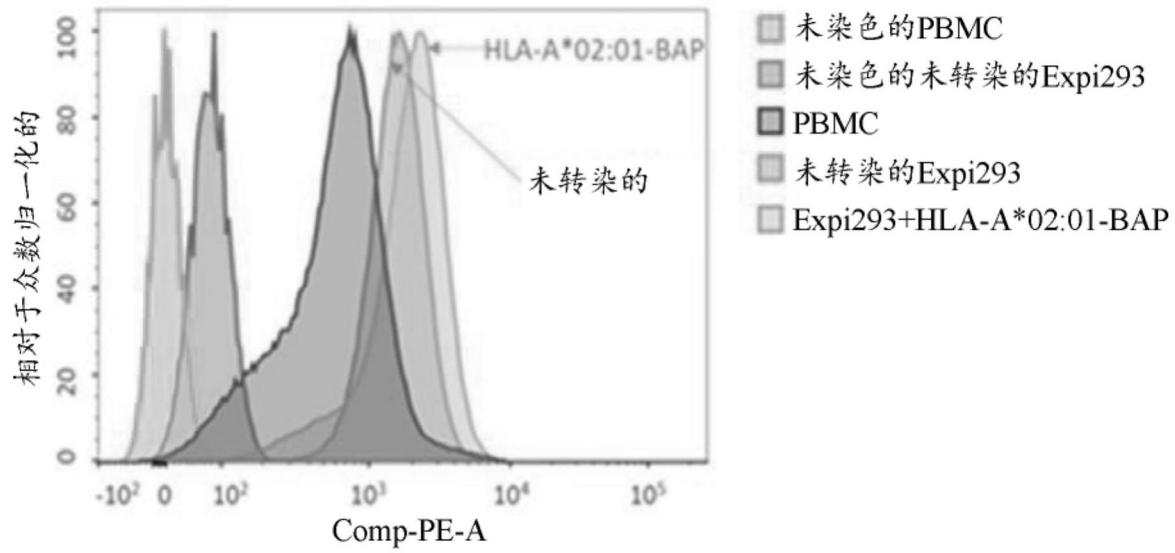

HLA-A*02:01-BAP
未转染的
相对于众数归一化的
Comp-PE-A
未染色的PBMC
未染色的未转染的Expi293
PBMC
未转染的Expi293
Expi293+HLA-A*02:01-BAP

图34A

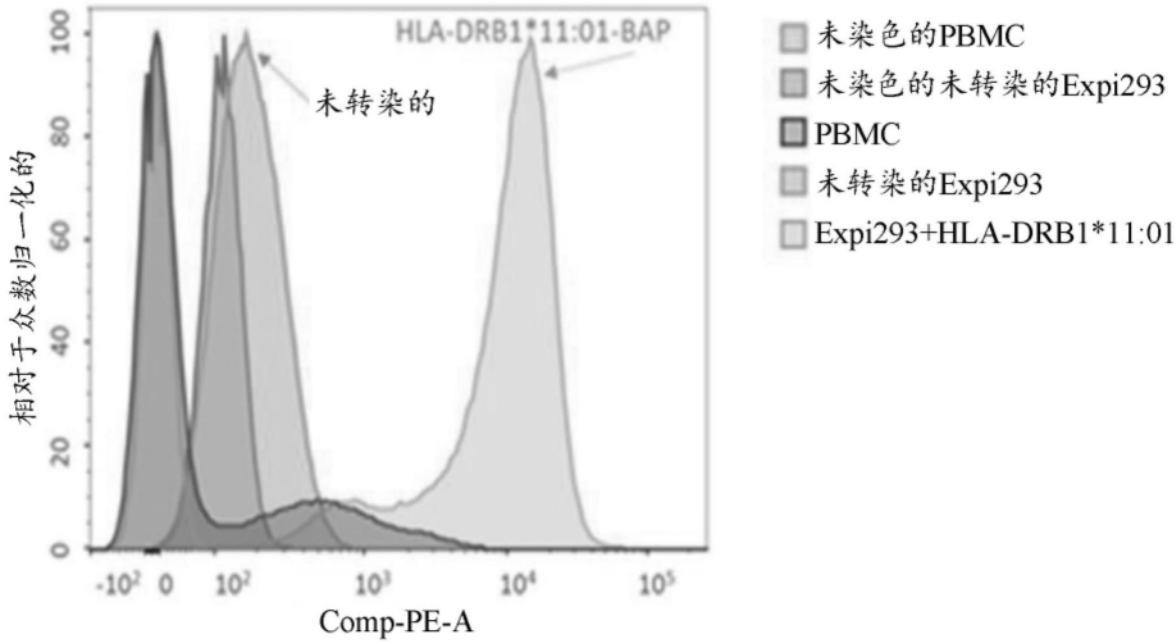

HLA-DRB1*11:01-BAP
未转染的
相对于众数归一化的
Comp-PE-A
未染色的PBMC
未染色的未转染的Expi293
PBMC
未转染的Expi293
Expi293+HLA-DRB1*11:01

图34B


MAPTAC
IEDB
DRB1*01:01
DRB1*04:02
DRB1*07:01
DRB1*09:01
DRB1*12:01
DRB1*14:01
DRB3*01:01
DPB1*01:01 / DPA1*01:03
DPB1*06:01 / DPA1*01:03
n = 2886
n = 998
n = 1888
n = 1499
n = 919
n = 1677
n = 824
n = 1172
n = 1127
n = 3473
n = 1155
n = 828
n = 70
n = 344
n = 279
<30个结合物
DRB1*01:02
DRB1*04:03
DRB1*08:01
DRB1*10:01
DRB1*12:02
DRB1*15:01
DRB3*02:02
DPB1*02:01 / DPA1*01:03
DQB1*06:02 / DQA1*01:02
n = 1016
n = 1138
n = 680
n = 870
n = 599
n = 1154
n = 774
n = 767
n = 936
n = 156
n = 117
n = 862
n = 250
n = 81
n = 77
DRB1*03:01
DRB1*04:04
DRB1*08:02
DRB1*11:01
DRB1*13:01
DRB1*15:02
DRB3*03:01
DPB1*03:01 / DPA1*01:03
DQB1*06:04 / DQA1*01:02
n = 849
n = 873
n = 924
n = 3407
n = 856
n = 692
n = 1163
n = 1586
n = 985
n = 646
n = 649
n = 487
n = 1047
n = 348
n = 277
n = 32
DRB1*03:02
DRB1*04:05
DRB1*08:03
DRB1*11:02
DRB1*13:02
DRB1*15:03
DRB4*01:01
DPB1*04:01 / DPA1*01:03
n = 397
n = 804
n = 613
n = 871
n = 444
n = 888
n = 886
n = 552
n = 495
n = 616
n = 415
n = 79
DRB1*04:01
DRB1*04:07
DRB1*08:04
DRB1*11:04
DRB1*13:03
DRB1*16:01
DRB5*01:01
DPB1*04:02 / DPA1*01:03
n = 864
n = 786
n = 1096
n = 770
n = 1129
n = 486
n = 792
n = 771
n = 953
n = 1070
n = 87

图35

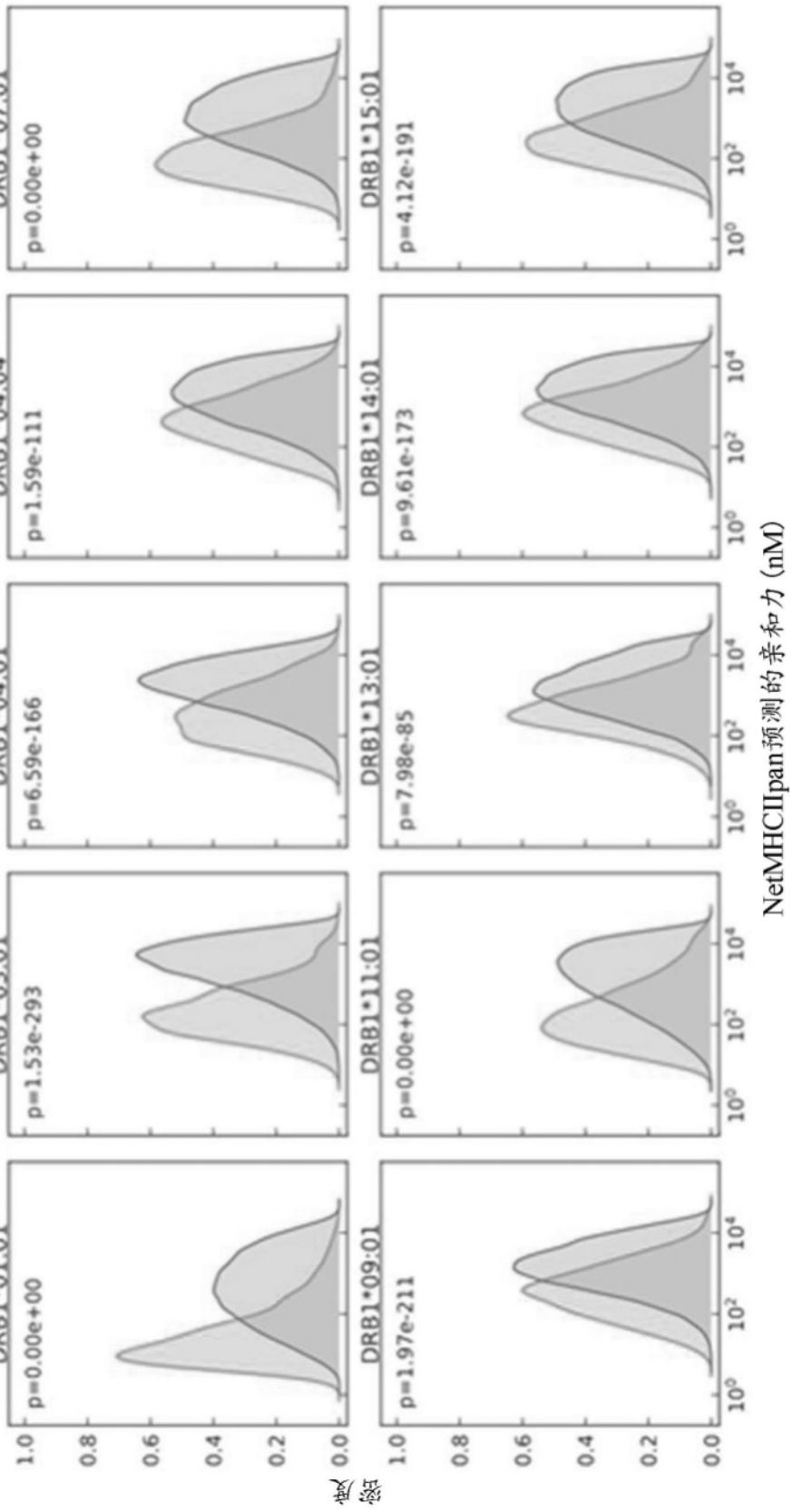

DRB1*01:01
p=0.00e+00
DRB1*03:01
p=1.53e-293
DRB1*04:01
p=6.59e-166
DRB1*04:04
p=1.59e-111
DRB1*07:01
p=0.00e+00
DRB1*09:01
p=1.97e-211
DRB1*11:01
p=0.00e+00
DRB1*13:01
p=7.98e-85
DRB1*14:01
p=9.61e-173
DRB1*15:01
p=4.12e-191
密度
1.0
0.8
0.6
0.4
0.2
0.0
10^0
10^2
10^4
NetMHCIIpan预测的亲和力(nM)

图36A

HLA-DRB1*01:01				
肽和NetMHCIIpan估算的结合核心	NetMHCIIpan亲和力	NetMHCIIpan百分排名	测得的亲和力	
DRYEMEDGKVIER	218	31	54	
GGHMTTLSGEEISYTG	239	48	17	
KTFDQLTPEESKER	193	34	12	
LPRYEALRGEQPPD	179	33	22	
DKKNIILEEGKEILVG	354	60	27	
KEAAYHPEVAPDVR	352	46	17	
DAEFRHDSGYEVHHQ	332	55	74	
GHPDLQGQPAEEIFES	212	46	189	
LGKNFDFQKSDRINSE	131	36	16	
MPSFVPSDGRQAAD	143	29	3	
GLRYKKLHDPKGWIT	189	40	30	
TIEKFEKEAAEMGKG	234	44	39	
GVQRGLVGEIIKR	120	20	8	
TWFNQPARKIR	155	7	8	
EFKKYEMMKEHERR	529	55	114	
SPHAFKTESGEETDL	1100	75	333	

HLA-DRB1*09:01				
肽和NetMHCIIpan估算的结合核心	NetMHCIIpan亲和力	NetMHCIIpan百分排名	测得的亲和力	
NNLCPSGSNIISN	996	32	5	
ERPYWDMSNQDVIN	1132	43	39	
VGGTMVRSGQDPYA	3415	80	40	
KEALEPSGENVIQ	4850	85	23	
RQRRLLGSVQQDLER	850	41	22	
IDRALNEACESVIQ	2919	75	262	
NENNLESAKGLLDDLR	758	42	64	
DSKSLRTALQKEITTR	847	45	23	
TGGDINAAIERLL	937	30	21	
SLDNLKASVSQVEADL	619	35	17	
DERRFKAADLNGDL	1025	45	127	
KPAPALRSARSAPEN	591	38	15	
RTSYVTSVEENTVD	1759	60	76	
RAVEFQEAQAYADDN	725	44	72	
RDLAQYDAAHHEEF	1227	50	237	

HLA-DRB1*11:01				
肽和NetMHCIIpan估算的结合核心	NetMHCIIpan亲和力	NetMHCIIpan百分排名	测得的亲和力	
EGEYQGIPRAESGG	1103	43	159	
SPEEFDEVSRIVGSVE	1704	65	59	
IAGEWQVLHREGAIT	561	34	35	
EPAEFIIDTRDAGYG	1654	60	195	
DLEELEVLERKPAAG	587	35	63	
GMNIVEAMERFGSRNG	853	45	79	
GFNWNWINKQQGKRG	611	36	12	
GPFSFSVIDKPPG	814	32	209	
LNELKPISKGGHSS	977	40	67	
DVNEYAPVFKEKSYK	588	35	223	
LAPTWEELSKKEFPG	8213	95	174	
TRNEVIPMSHPGAVD	1193	44	36	
ENPYFAPNPKIIRQ	421	18	128	
YEDKFRNNLKGKRLD	297	18	254	
KNGAYKVETKKYDFY	308	19	12	
GQQWTYEQRKIVE	685	20	145	

图36B

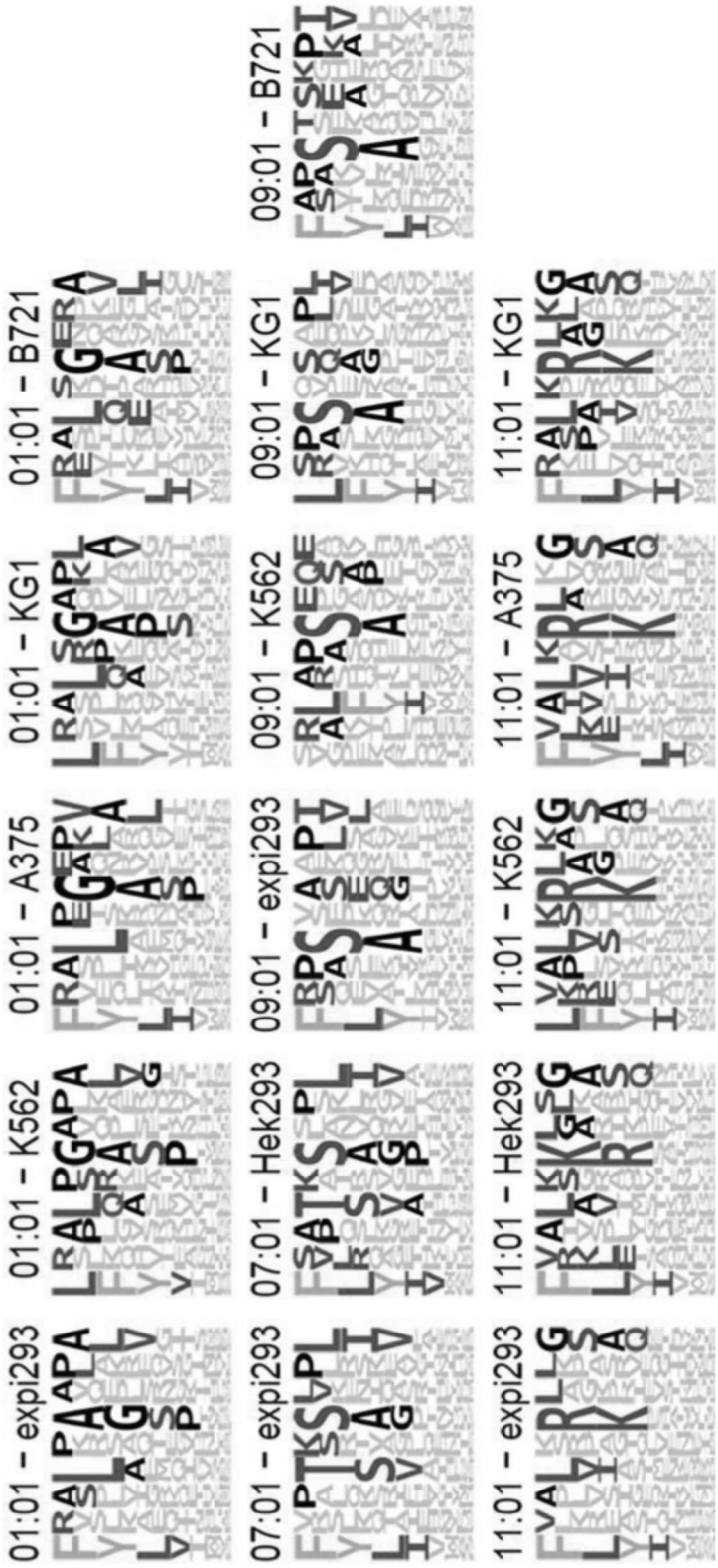

09:01 - B721
01:01 - B721
09:01 - KG1
11:01 - KG1
01:01 - KG1
09:01 - K562
11:01 - A375
01:01 - A375
09:01 - expi293
11:01 - K562
01:01 - K562
07:01 - Hek293
11:01 - Hek293
01:01 - expi293
07:01 - expi293
11:01 - expi293

图36C

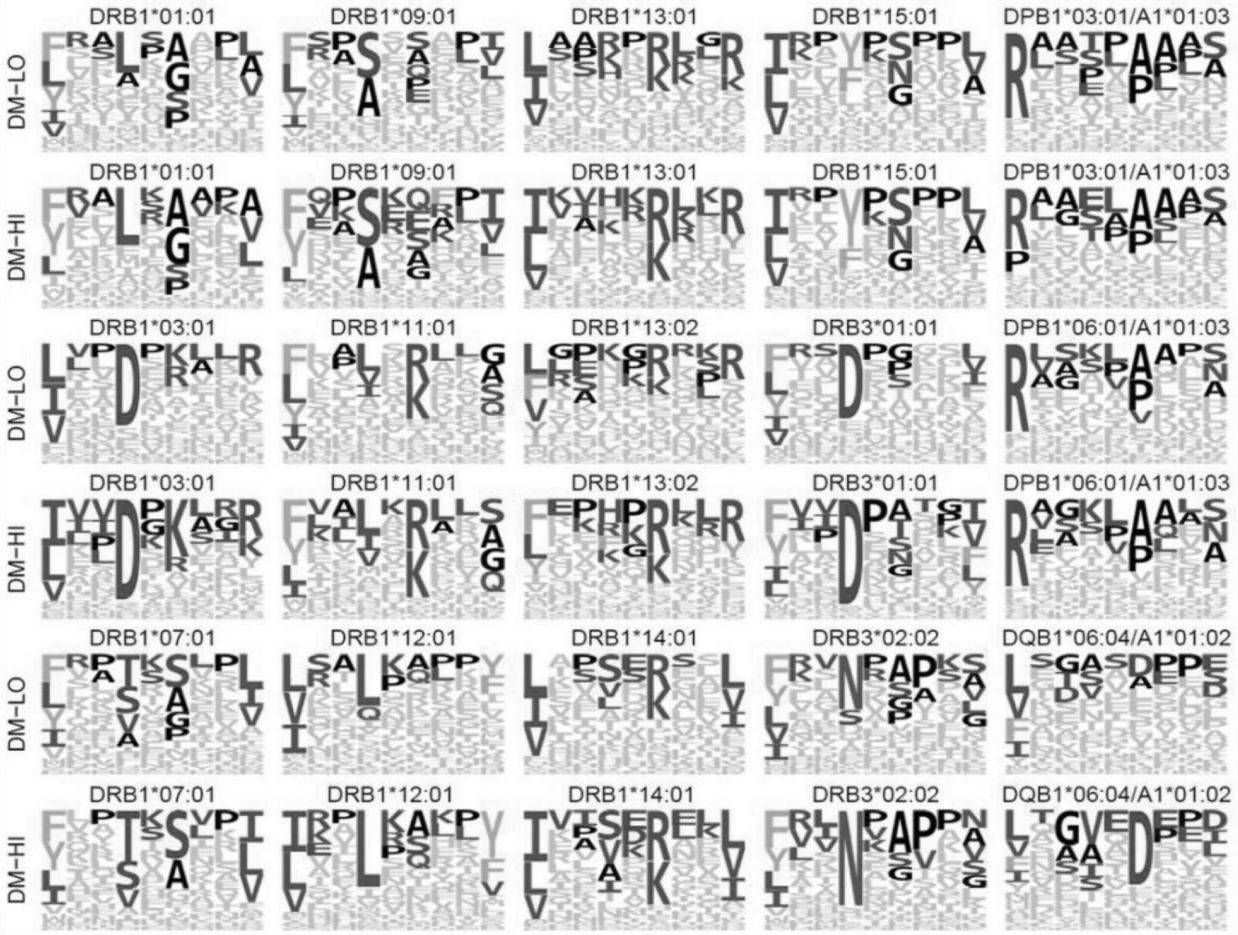

DM-LO
DM-HI
DRB1*01:01
DRB1*09:01
DRB1*13:01
DRB1*15:01
DPB1*03:01/A1*01:03
DRB1*03:01
DRB1*11:01
DRB1*13:02
DRB3*01:01
DPB1*06:01/A1*01:03
DRB1*07:01
DRB1*12:01
DRB1*14:01
DRB3*02:02
DQB1*06:04/A1*01:02

图37A

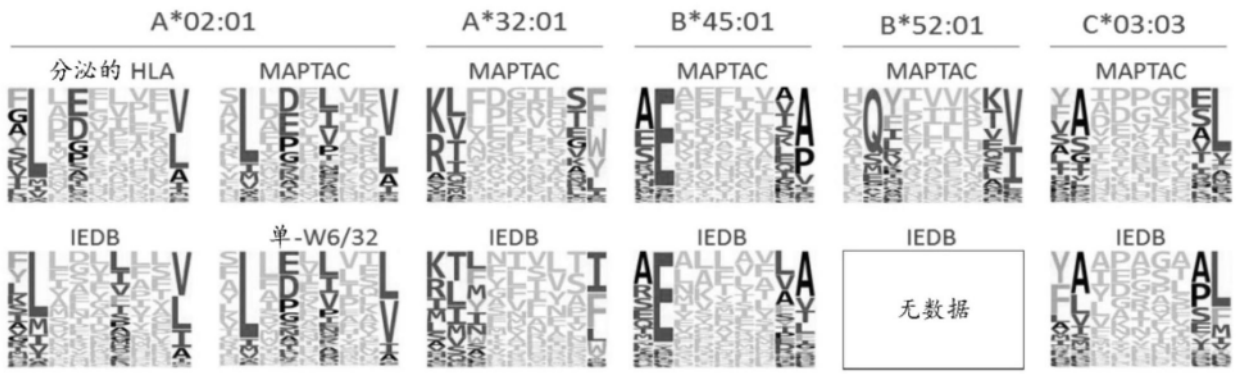

A*02:01
A*32:01
B*45:01
B*52:01
C*03:03
分泌的 HLA
MAPTAC
IEDB
单-W6/32
无数据

图37B


DRA1*01:01=DRB5*01:01
DRA1*01:01=DRB4*01:03
DRA1*01:01=DRB3*03:01
DRA1*01:01=DRB1*15:03
DRA1*01:01=DRB1*13:02
DRA1*01:01=DRB1*11:01
DRA1*01:01=DRB1*09:01
DRA1*01:01=DRB1*04:05
DRA1*01:01=DRB1*01:01
DQA1*02:01=DQB1*02:02

图37C

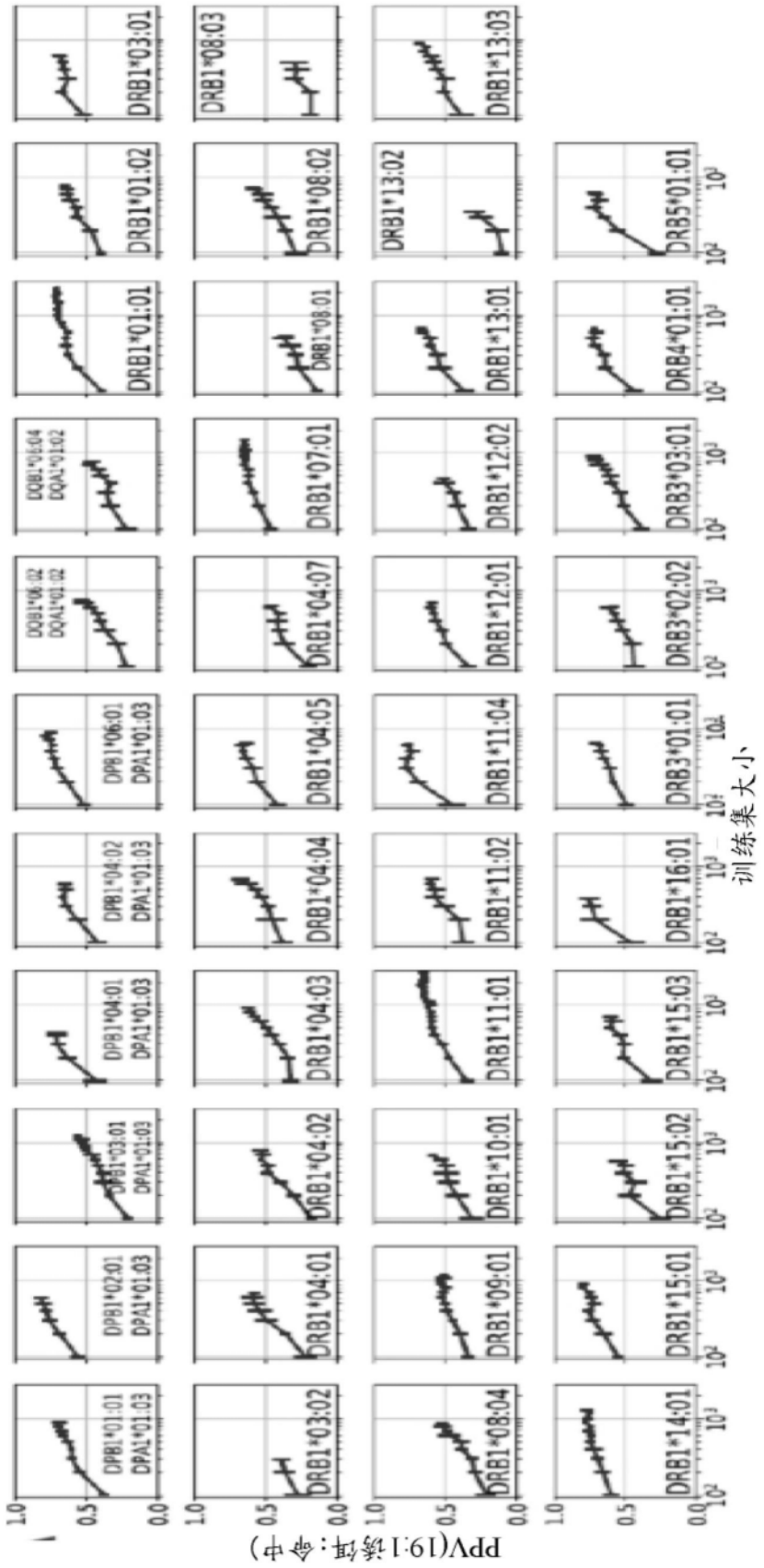

DRB1*03:01
DRB1*01:02
DRB1*08:03
DRB1*08:02
DRB1*13:02
DRB1*13:03
DRB1*01:01
DRB1*08:01
DRB1*13:01
DRB5*01:01
DRB1*07:01
DRB1*12:02
DRB4*01:01
DRB1*04:07
DRB1*12:01
DRB3*03:01
DRB1*04:05
DRB1*11:04
DRB3*02:02
DRB1*04:04
DRB1*11:02
DRB3*01:01
DRB1*04:03
DRB1*11:01
DRB1*16:01
DRB1*04:02
DRB1*10:01
DRB1*15:03
DRB1*04:01
DRB1*09:01
DRB1*15:02
DRB1*03:02
DRB1*08:04
DRB1*15:01
DRB1*14:01
DPB1*01:01
DPA1*01:03
DPB1*02:01
DPA1*01:03
DPB1*03:01
DPA1*01:03
DPB1*04:01
DPA1*01:03
DRB1*04:02
DPA1*01:03
DPB1*06:01
DPA1*01:03
DQB1*06:02
DQA1*01:02
DQB1*06:04
DQA1*01:02
1.0
0.5
0.0
10^2
10^3
PPV(19:1诱饵:命中)
训练集大小

图38Ai

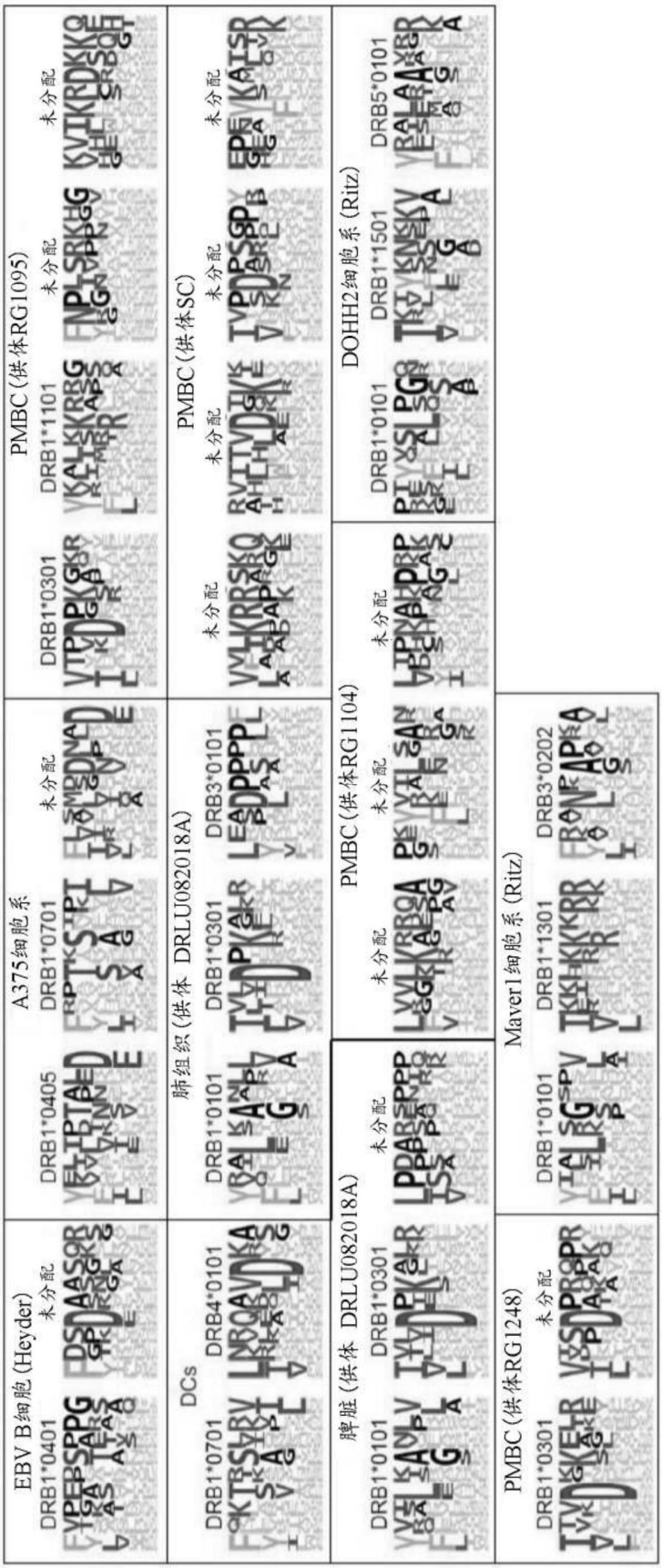

EBV B细胞 (Heyder)
DRB1*0401
未分配
A375细胞系
DRB1*0405
DRB1*0701
未分配
PMBC (供体RG1095)
DRB1*0301
DRB1*1101
未分配
未分配
DCs
DRB1*0701
DRB4*0101
肺组织 (供体 DRLU082018A)
DRB1*0101
DRB1*0301
DRB3*0101
未分配
PMBC (供体SC)
未分配
未分配
脾脏 (供体 DRLU082018A)
DRB1*0101
DRB1*0301
未分配
PMBC (供体RG1104)
未分配
未分配
DOHH2细胞系 (Ritz)
DRB1*0101
DRB1*1501
DRB5*0101
PMBC (供体RG1248)
DRB1*0301
未分配
Maver1细胞系 (Ritz)
DRB1*0101
DRB1*1301
DRB3*0202

图38Aii

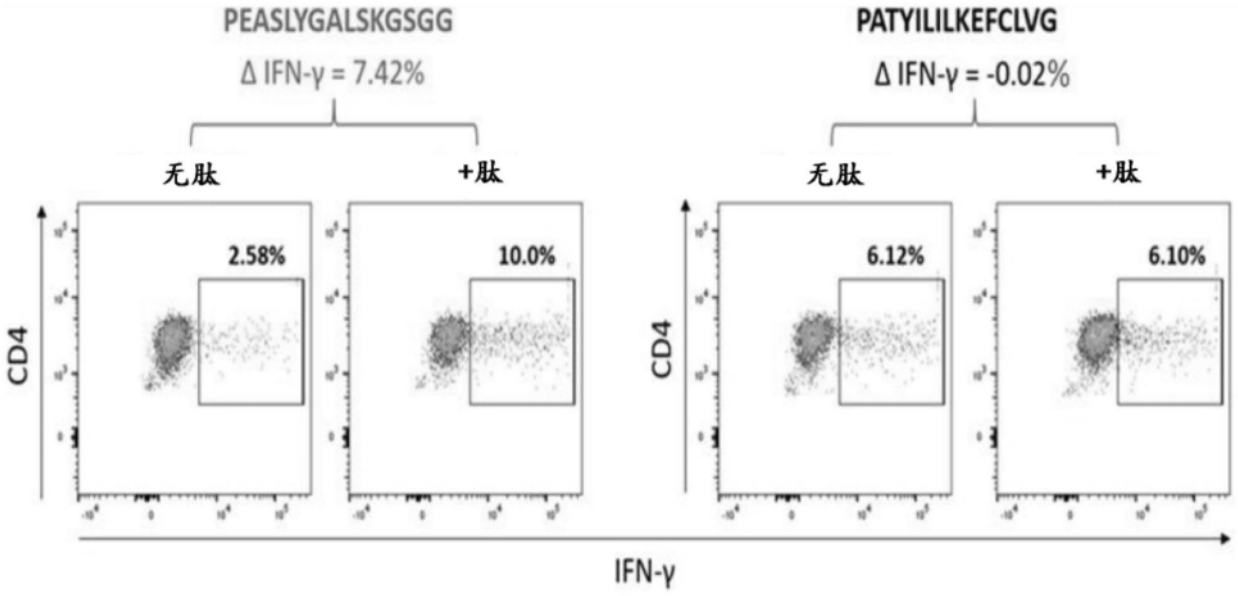


图38B

肽	状态	阳性孔数目	孔1	孔2	孔3	孔4	孔5	孔6	孔7	孔8
GEGLFQPAHRYPDAG	诱导_命中	4	3.28	3.5	3.87	3.33	0.81			
RNQWKCLGKPVGAEM	测定的	0	1.88	-1.44	-1.27	-15.43	-10.18	-0.09		
VPAWTRAWRNSSPKG	诱导_命中	2	3.53	0.4	-0.36	2.93	2.2	6.3		
LHPELLPLWRLLPDG	诱导_命中	1	-1.26	0.46	-0.36	-1.74	0.75	3.92		
DGGSYFSLWKIWTQV	诱导_命中	5	4.83	3.84	4.34	13.8	8.1			
PEASLYGALSKGSGG	诱导_命中	5	1.96	7.42	1.09	5.24	6.03	1.72	12.93	9.74
HEGGFPPLLRRAAED	诱导_命中	5	-0.36	3.58	2.58	-22.75	6.59	3.5	8.44	10.39
LSPVWCLQWKLSGTD	诱导_命中	2	-0.08	2.52	0.71	13.27	11.32			
KNTIVYTTKQVQSCQ	测定的	0	-1.19							
PATYILILKEFCLVG	测定的	0	-0.02	0.69	1.47	-0.17				
NNEQFQWKIRHVGPE	诱导_命中	4	5.44	5.7	0.93	2.97	5.1	4.5		
YKGGYELVKKSQTEL	测定的	0	2.28	-2.01	-2.31					

图38C

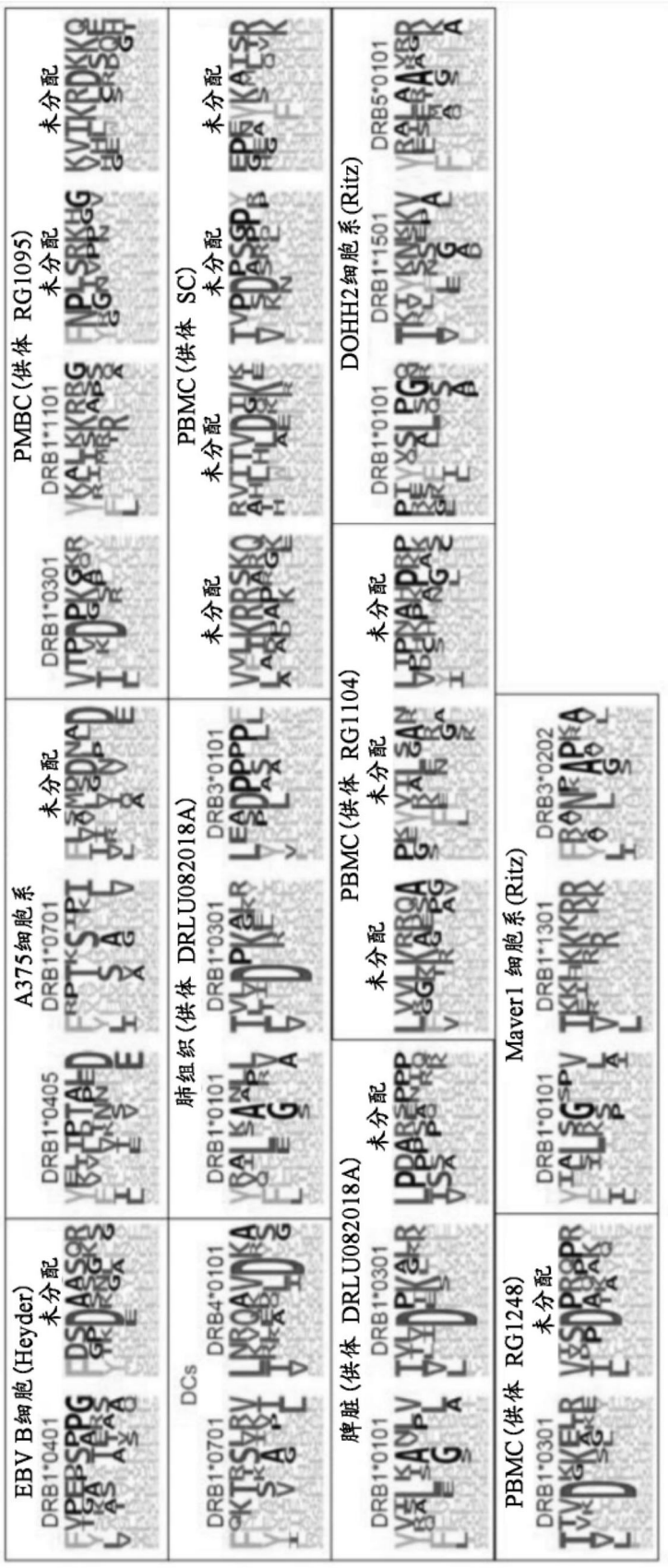

EBV B细胞 (Heyder)
DRB1*0401
未分配
DCs
DRB1*0701
DRB4*0101
脾脏 (供体 DRLU082018A)
DRB1*0101
DRB1*0301
未分配
PBMC (供体 RG1248)
DRB1*0301
未分配
A375细胞系
DRB1*0405
DRB1*0701
未分配
肺组织 (供体 DRLU082018A)
DRB1*0101
DRB1*0301
DRB3*0101
未分配
PBMC (供体 RG1104)
未分配
未分配
Maver1 细胞系 (Ritz)
DRB1*0101
DRB1*1301
DRB3*0202
PMBC (供体 RG1095)
DRB1*0301
未分配
DRB1*1101
未分配
PBMC (供体 SC)
未分配
未分配
DOHH2细胞系 (Ritz)
DRB1*0101
DRB1*1501
DRB5*0101

图38D

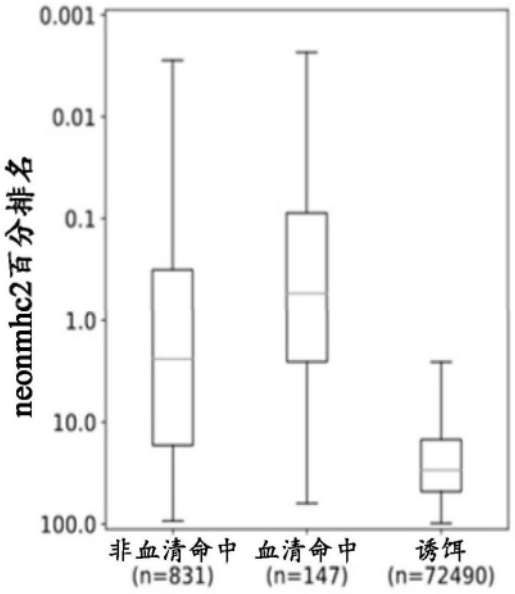

0.001
0.01
0.1
1.0
10.0
100.0
neonmhc2百分排名
非血清命中
(n=831)
血清命中
(n=147)
诱饵
(n=72490)

图39A

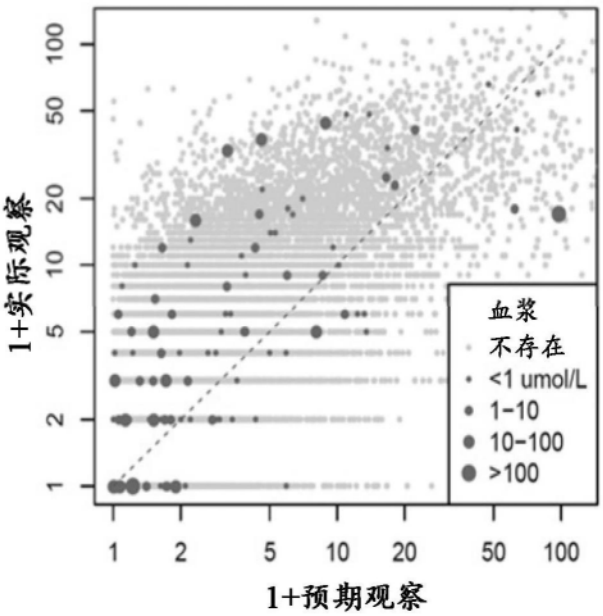

1
2
5
10
20
50
100
1+实际观察
1+预期观察
血浆
不存在
<1 umol/L
1-10
10-100
>100

图39B

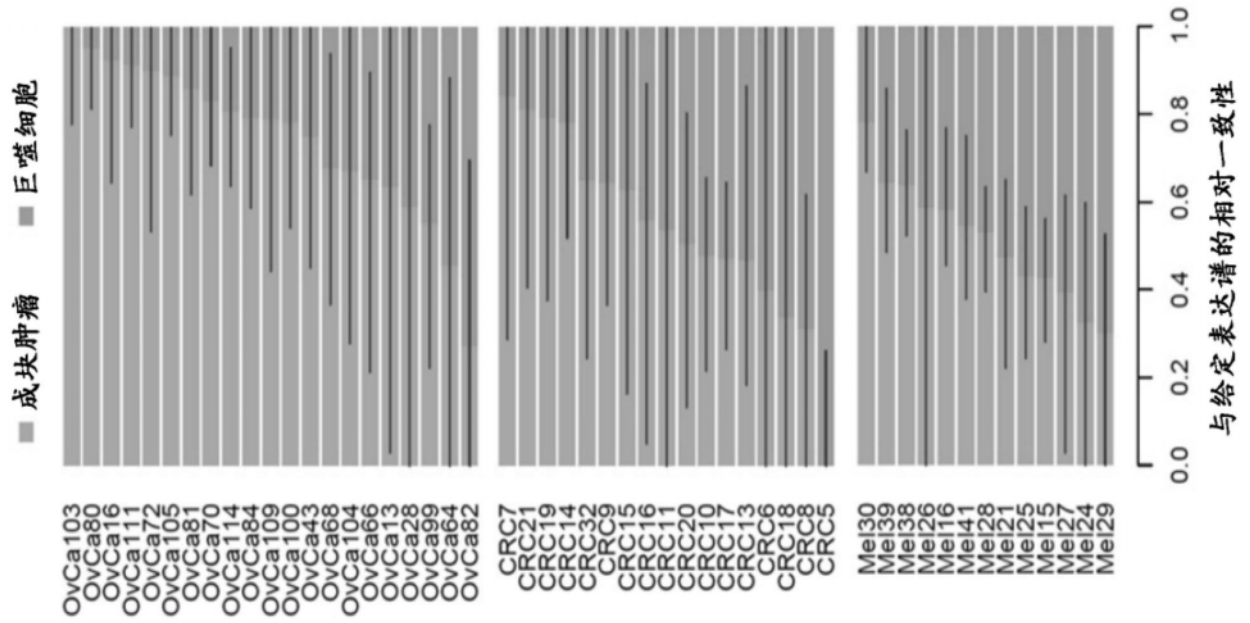

成块肿瘤
巨噬细胞
OvCa103
OvCa80
OvCa16
OvCa111
OvCa72
OvCa105
OvCa81
OvCa70
OvCa114
OvCa84
OvCa109
OvCa100
OvCa43
OvCa68
OvCa104
OvCa66
OvCa13
OvCa28
OvCa99
OvCa64
OvCa82
CRC7
CRC21
CRC19
CRC14
CRC32
CRC9
CRC15
CRC16
CRC11
CRC20
CRC10
CRC17
CRC13
CRC6
CRC18
CRC8
CRC5
Mel30
Mel39
Mel38
Mel26
Mel16
Mel41
Mel28
Mel21
Mel25
Mel15
Mel27
Mel24
Mel29
0.0
0.2
0.4
0.6
0.8
1.0
与给定表达谱的相对一致性

图39C

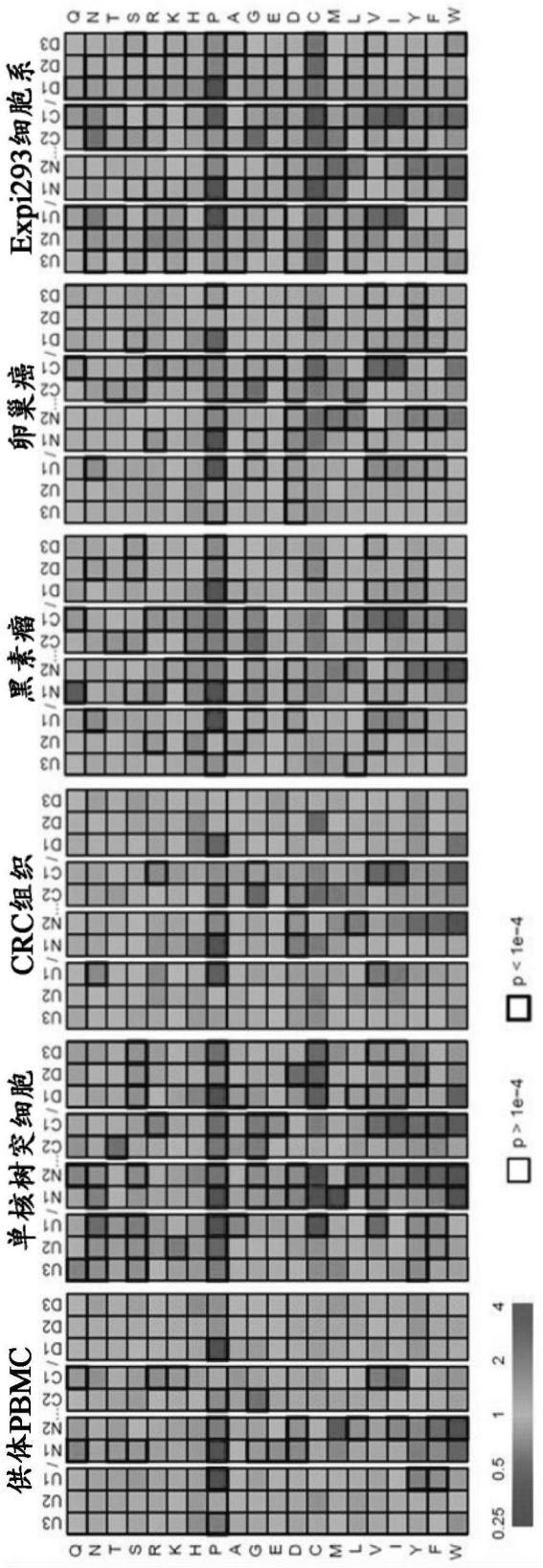

Expi293细胞系
卵巢癌
黑素瘤
CRC组织
单核树突细胞
供体PBMC
Q N T S R K H P A G E D C M L V I Y F W
p < 1e-4
p > 1e-4
0.25 0.5 1 2 4

图40A

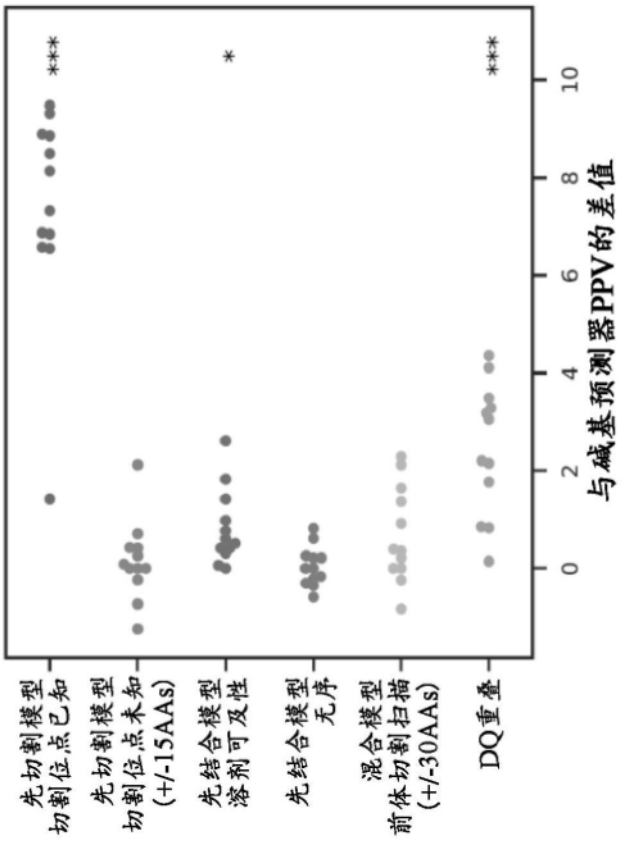

先切割模型
切割位点已知
先切割模型
切割位点未知
(+/-15AAs)
先结合模型
溶剂可及性
先结合模型
无序
混合模型
前体切割扫描
(+/-30AAs)
DQ重叠
***
*
***
0
2
4
6
8
10
与碱基预测器PPV的差值

图40B

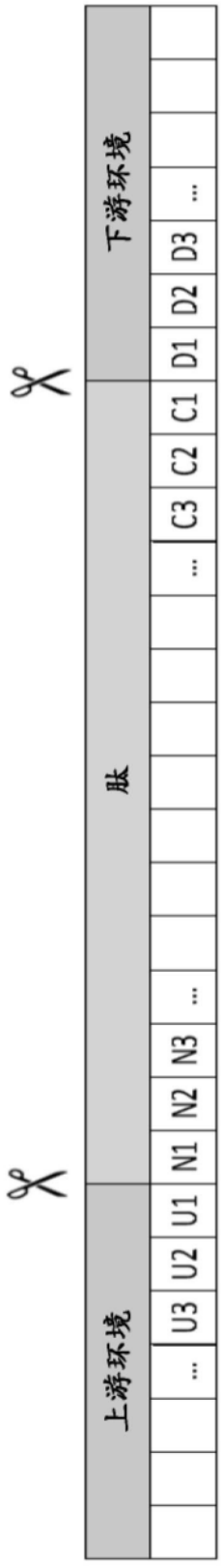


图41

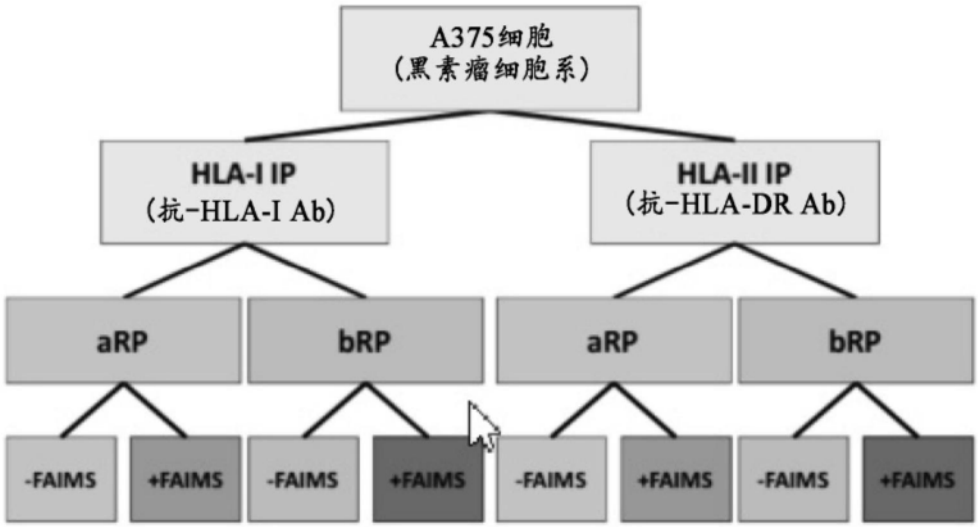

A375细胞
(黑素瘤细胞系)
HLA-I IP
(抗-HLA-I Ab)
HLA-II IP
(抗-HLA-DR Ab)
aRP
bRP
aRP
bRP
-FAIMS
+FAIMS
-FAIMS
+FAIMS
-FAIMS
+FAIMS
-FAIMS
+FAIMS

图42A

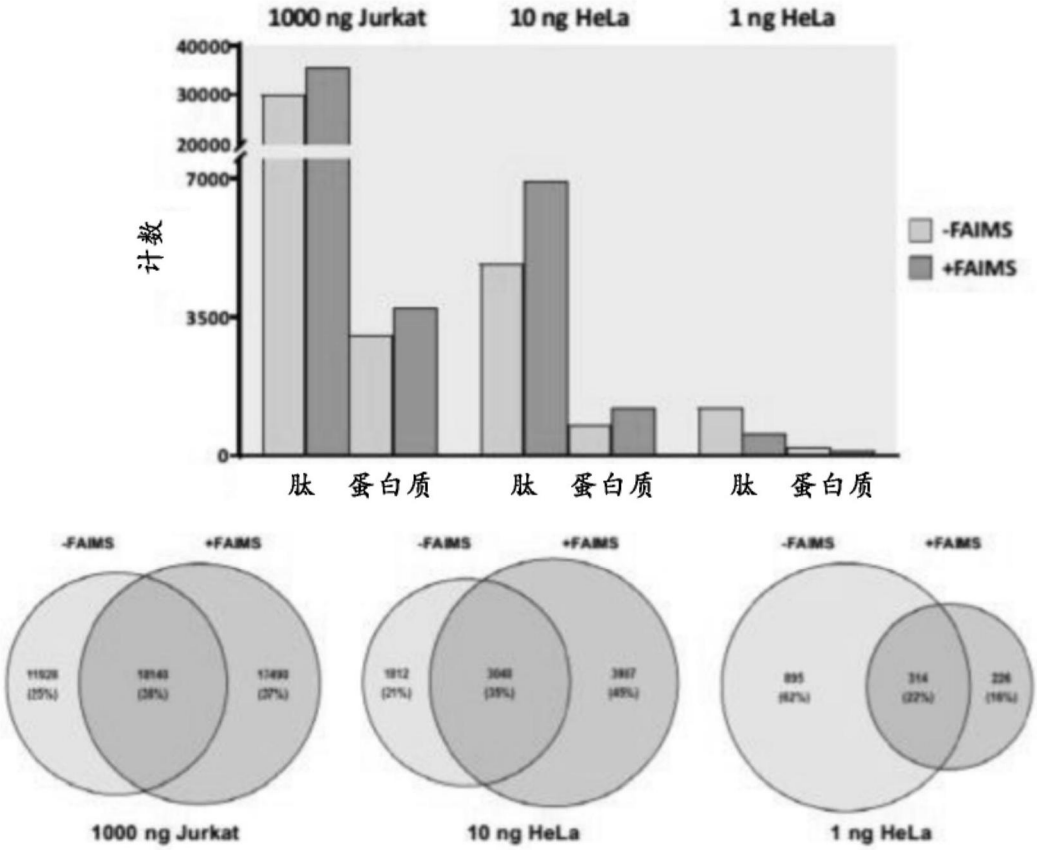

1000 ng Jurkat
10 ng HeLa
1 ng HeLa
40000
30000
20000
7000
3500
0
计数
-FAIMS
+FAIMS
肽
蛋白质
肽
蛋白质
肽
蛋白质
-FAIMS
+FAIMS
-FAIMS
+FAIMS
-FAIMS
+FAIMS
1000 ng Jurkat
10 ng HeLa
1 ng HeLa

图42B

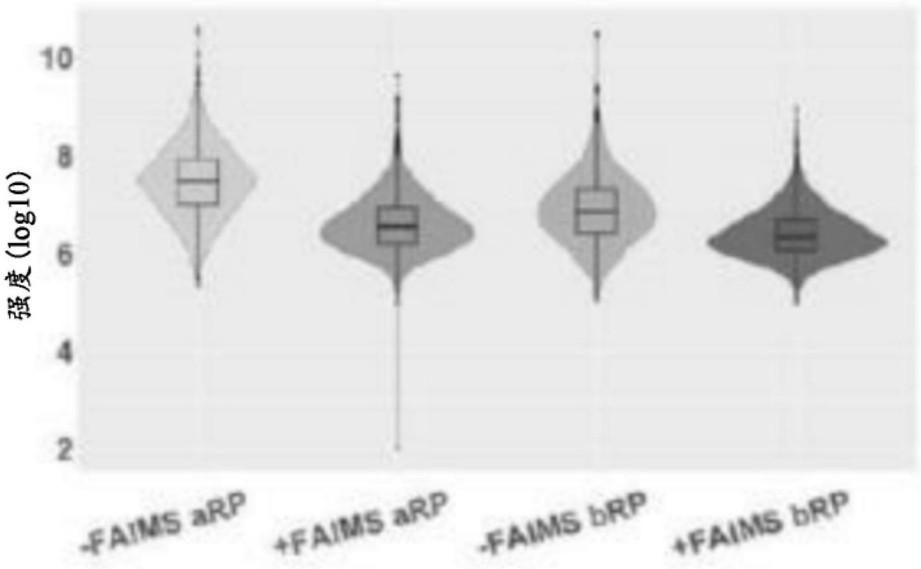

强度(log10)
10
8
6
4
2
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP

图43A

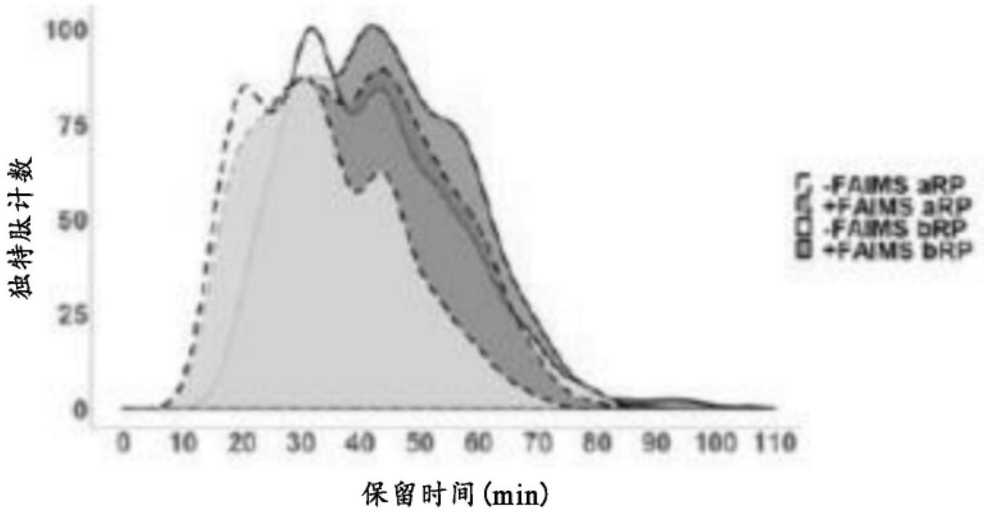

独特肽计数
100
75
50
25
0
0 10 20 30 40 50 60 70 80 90 100 110
保留时间(min)
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP

图43B

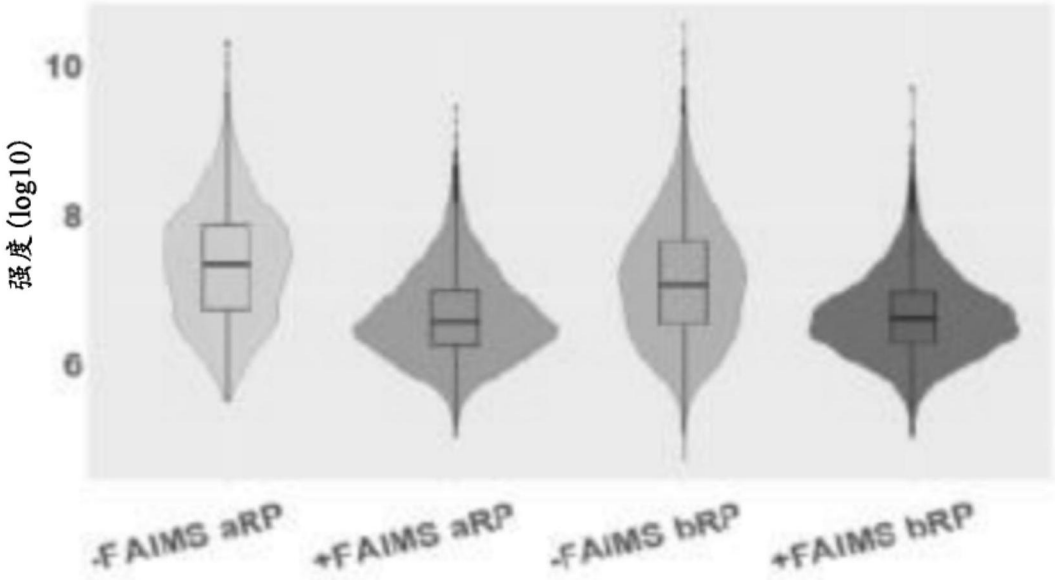

强度(log10)
10
8
6
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP

图44A

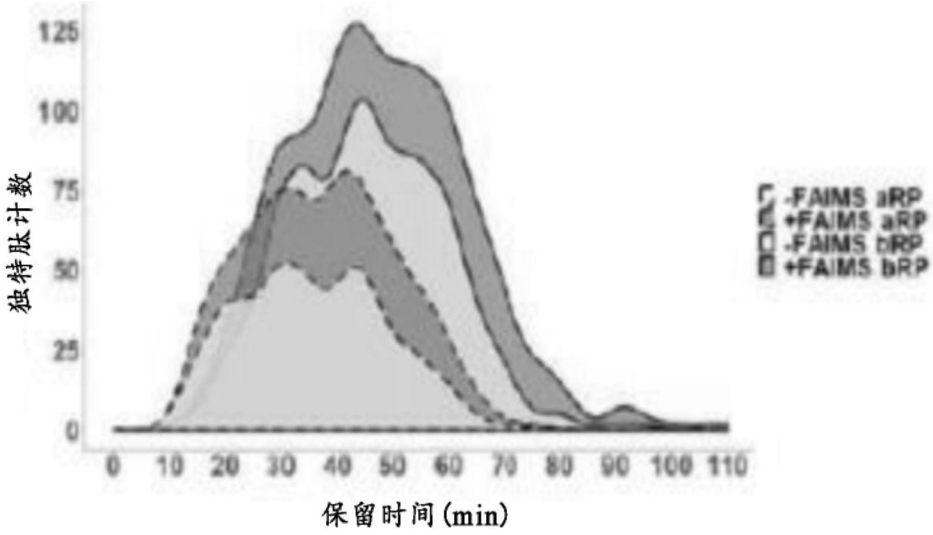

独特肽计数
125
100
75
50
25
0
0 10 20 30 40 50 60 70 80 90 100 110
保留时间(min)
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP

图44B

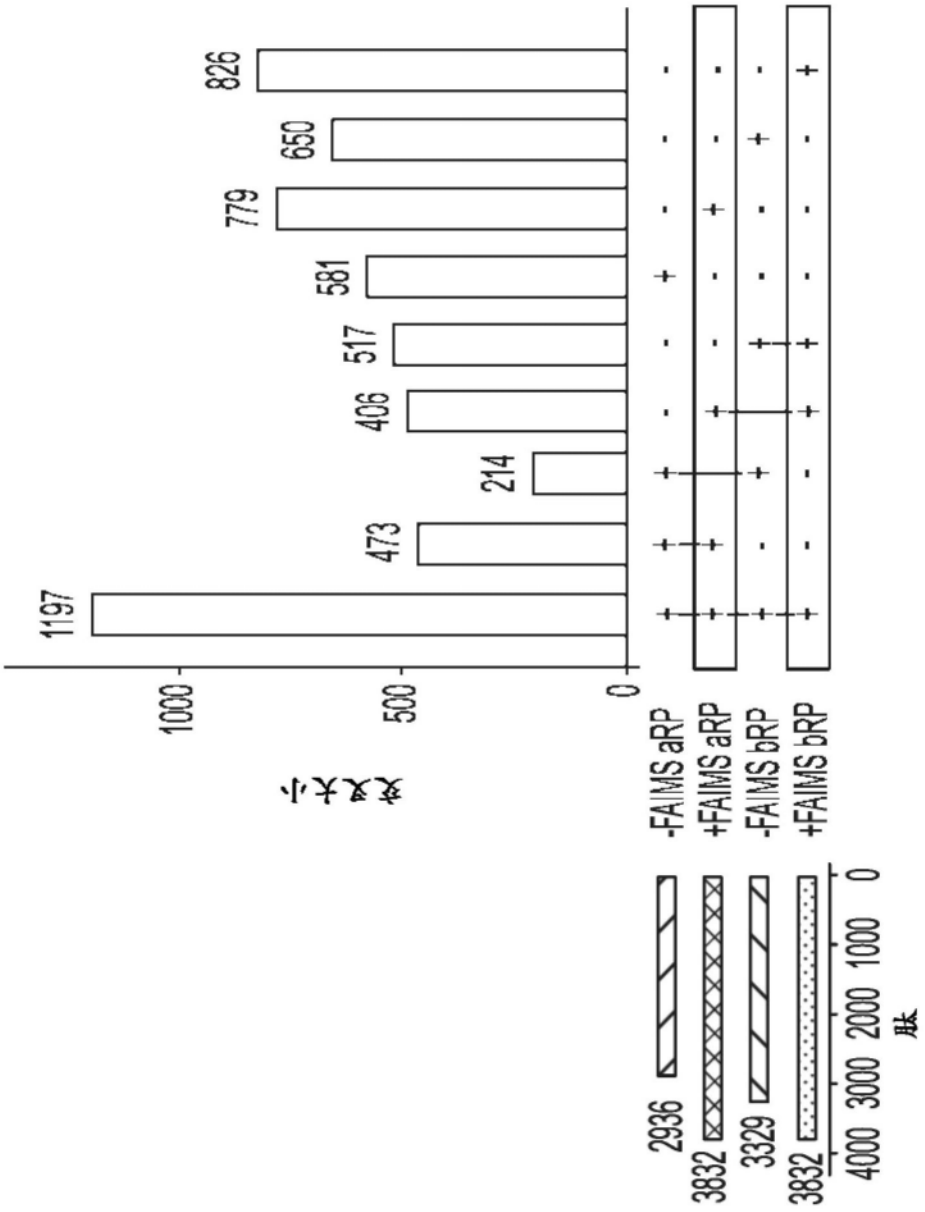

826
650
779
581
517
406
214
473
1197
1000
500
0
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP
2936
3832
3329
3832
4000 3000 2000 1000 0

图45

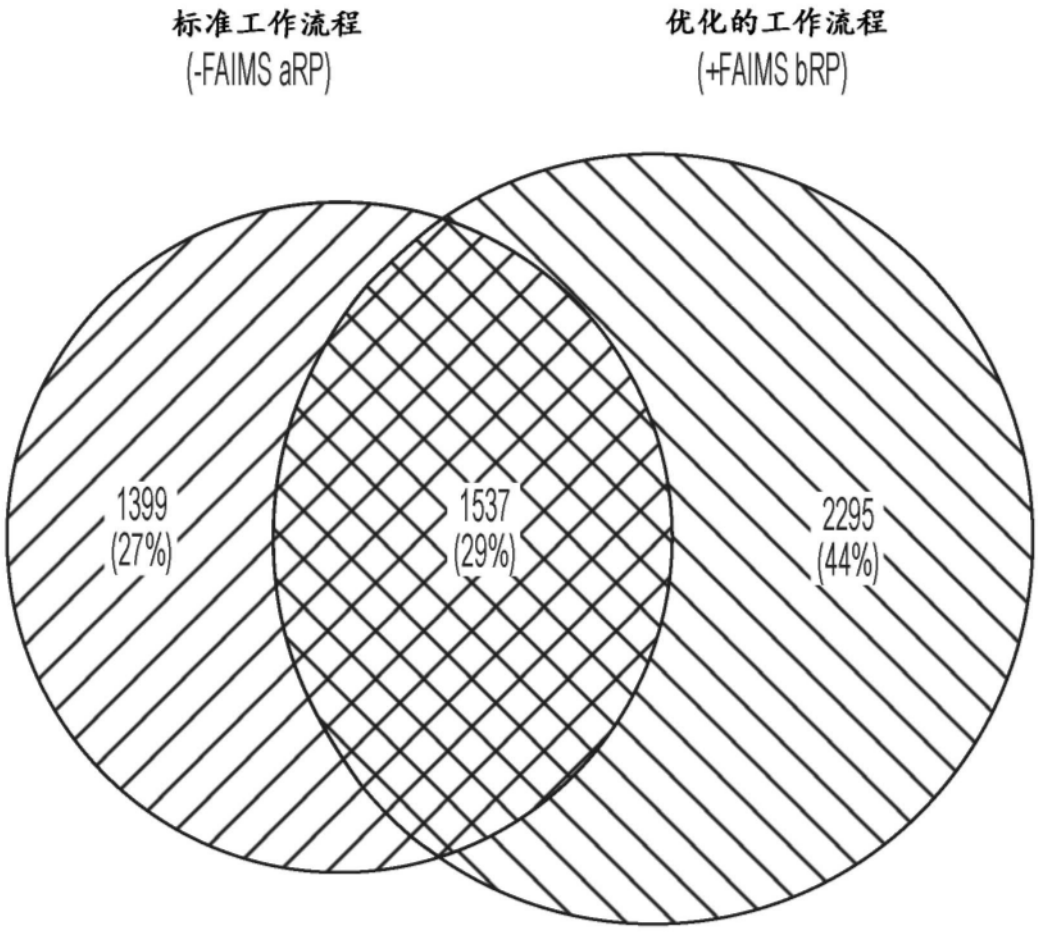


图45(续)

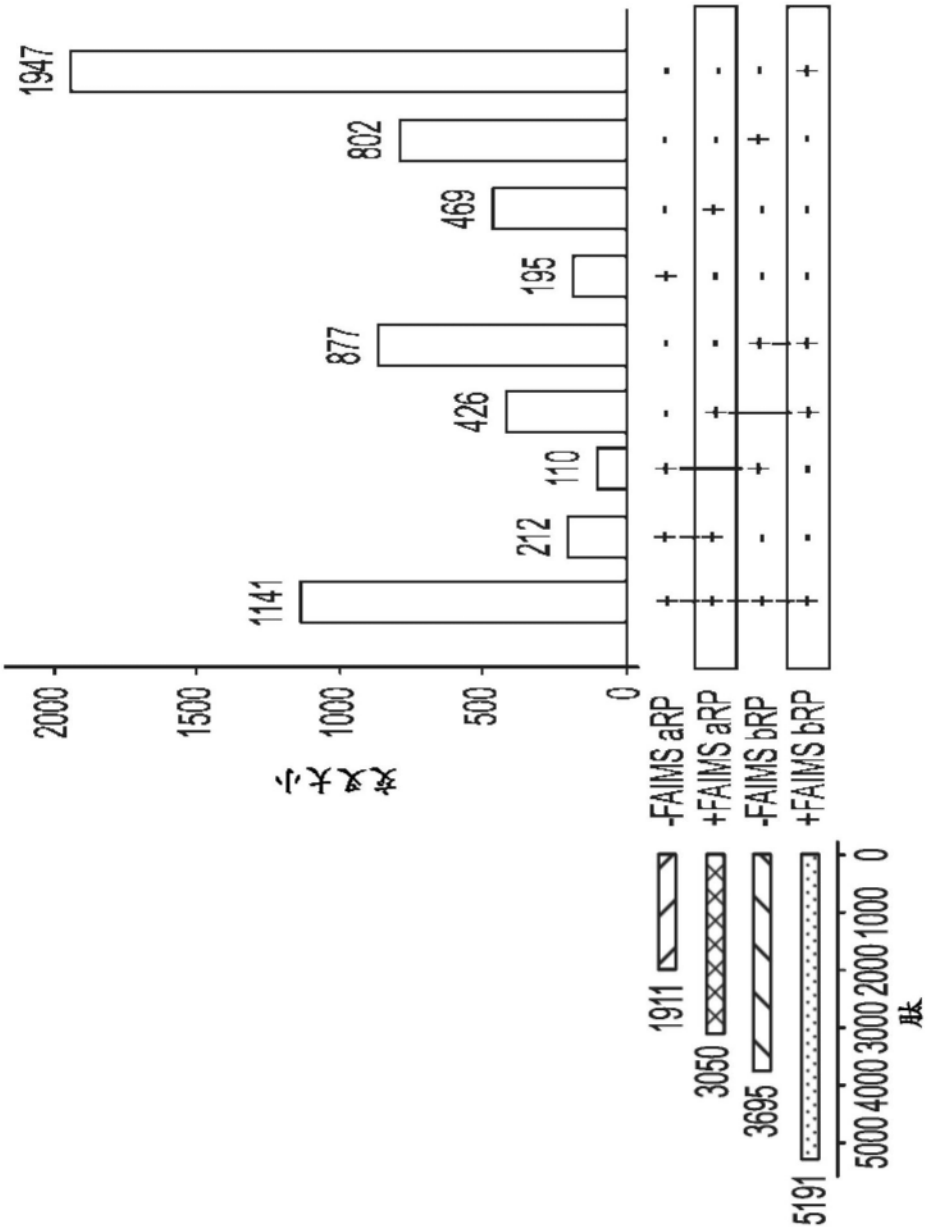

1947
802
469
195
877
426
110
212
1141
2000
1500
1000
500
0
-FAIMS aRP
+FAIMS aRP
-FAIMS bRP
+FAIMS bRP
1911
3050
3695
5191
5000 4000 3000 2000 1000 0

图46

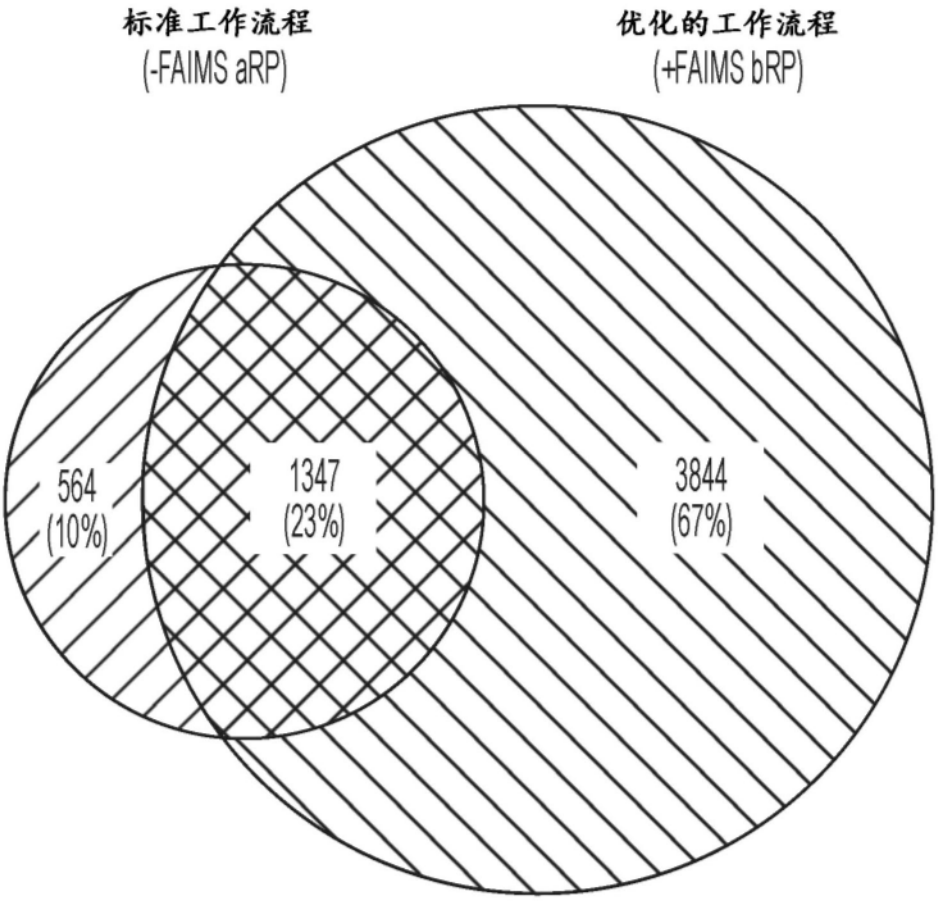


图46(续)