(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110799196 B
(45)授权公告日 2024.02.13

(21)申请号 201880019637.1
(22)申请日 2018.03.31
(65)同一申请的已公布的文献号
申请公布号 CN 110799196 A
(43)申请公布日 2020.02.14
(30)优先权数据
62/479,320 2017.03.31 US
(85)PCT国际申请进入国家阶段日
2019.09.20
(86)PCT国际申请的申请数据
PCT/US2018/025597 2018.03.31
(87)PCT国际申请的公布数据
WO2018/183980 EN 2018.10.04
(73)专利权人 行动基因(智财)有限公司
地址 中国香港新界区香港科学园科技大道东5号15W大楼8楼803
(72)发明人 杨沛佳 郑人豪 陈映嘉 陈淑贞 陈华键
(74)专利代理机构 上海德昭知识产权代理有限公司 31204
专利代理师 郁旦蓉
(51)Int.Cl.
C07K 14/00(2006.01)
(56)对比文件
US 2016069895 A1,2016.03.10
US 2016101170 A1,2016.04.14
WO 2016174085 A1,2016.11.03
审查员 王慧研

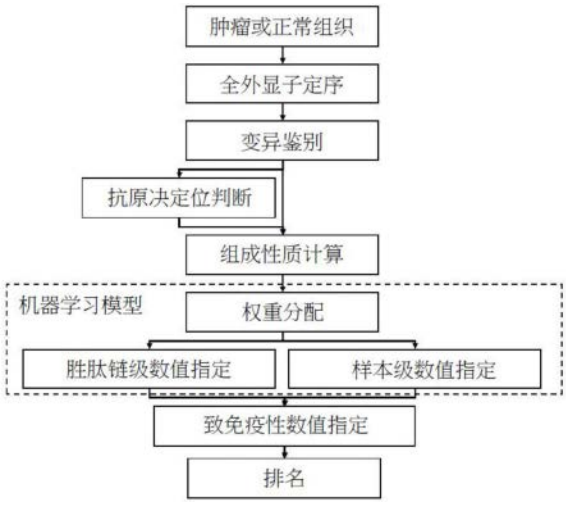
权利要求书3页 说明书28页
序列表20页 附图8页

(54)发明名称

致免疫性的癌症特异抗原决定位的排名系统

(57)摘要

本发明与能判断、预测及排名致免疫性T细胞抗原决定位的系统与方法有关,尤其是判断由疾病相关突变所产生的抗原决定位,且其中的抗原决定位被预测为能引起T细胞免疫反应。具体来说,本发明同时考虑了胜肽链级信息(包含胜肽链在主要组织兼容性复合体类型一及类型二的呈现和辅助性与胞杀性T细胞的免疫反应上的影响)以及样本级信息(包含突变群落性和主要组织兼容性复合体等位基因表现量)。在一些实施例中,上述系统与方法被使用在癌症个人化医疗上。



CN 110799196 B

1.一种筛选肿瘤组织中至少一个致免疫性且具突变信息的胜肽链的方法,其步骤包括:

(a)取得复数个具突变信息的序列;

(b)从疾病相关的突变中判断出至少一个抗原决定位;

(c)将与所述至少一个抗原决定位的免疫性相关的复数个特征整合,其中,所述复数个特征包含:

第一特征,其用于计算胜肽链级数值;以及

第二特征,其用于计算样本级数值,

其中,所述第二特征包含所述疾病相关突变的群落性;

(d)决定所述复数个特征的重要性;

(e)以所述复数个特征的重要性决定所述至少一个抗原决定位的致免疫性数值;

(f)根据所述致免疫性数值的高低,将所述至少一个抗原决定位排名;以及

(g)依所述(f)步骤中的排名结果来选择至少一个所述致免疫性且具突变信息的胜肽链,其中,所述致免疫性且具突变信息的胜肽链包含至少一个抗原决定位且其可能可引起T细胞免疫反应,

其中,所述步骤(c)～(e)为利用机器学习模型达成,

在胜肽链级数值计算中,利用所述第一特征构建四套机器学习模型,利用机器学习回归器和数据分析方法整合这四个模型和其组合,最终模型包含加权后的因素、特征筛选和叠代调整后的最佳化机器学习模型,最后模型利用已知具备致免疫性的抗原决定位进行验证,

所述复数个特征与下述多种有关:

至少一个所述抗原决定位在主要组织兼容性复合体类型一及类型二上的呈现、至少一个所述抗原决定位引起胞杀性T细胞免疫反应的能力有关、至少一个所述抗原决定位引起辅助性T细胞免疫反应的能力、蛋白质量、基因表现量、突变的变异频率、拷贝数、肿瘤纯度、异质性丧失的状态、等位基因量;

并且,所述第二特征包含由下列公式计算得的亚群落纯度S:

$$S = \frac{AF_{observed}}{M}\left(pC_t + C_n(1-p)\right),$$

$AF_{observed}$为观察等位基因频率值,p为肿瘤纯度,Ct为肿瘤变异状况(t)的拷贝数,Cn为正常状况(n)的拷贝数,M为突变等位基因数量。

2.如权利要求1所述的方法,其特征在于:其中,被选择所述抗原决定位的数量为≤100。

3.如权利要求1所述的方法,其特征在于:其中,被选择所述抗原决定位的数量为≤50。

4.如权利要求1所述的方法,其特征在于:其中,被选择所述抗原决定位的数量为≤30。

5.如权利要求1所述的方法,其特征在于:其中,被选择所述抗原决定位的数量为≤10。

6.如权利要求1所述的方法,其特征在于:其中,被选择至少一个所述抗原决定位与所述主要组织兼容性复合体类型一的结合亲和力的半数抑制浓度值为＜1500nM。

7.如权利要求1所述的方法,其特征在于:其中,所述复数个特征包含所述抗原决定位和所述主要组织兼容性复合体类型一的结合稳定性。

8.如权利要求1所述的方法,其特征在于:其中,所述复数个特征包含所述抗原决定位与其自体胜肽链的相似性。

9.如权利要求1所述的方法,其特征在于:其中,所述复数个特征包含至少一个所述抗原决定位与已知抗原的同源性。

10.如权利要求1所述的方法,其特征在于:其中,所述突变的变异频率为至少10%。

11.如权利要求1所述的方法,其特征在于:其中,所述突变的变异频率为至少30%。

12.如权利要求1所述的方法,其特征在于:其中,所述致免疫性数值是由整合所述复数个特征计算而来,所述复数个特征包含至少一个所述抗原决定位在主要组织兼容性复合体类型一及类型二上的呈现能力、所述抗原决定位引起辅助性及胞杀性T细胞免疫反应的能力以及所述疾病相关突变的群落性。

13.一种筛选肿瘤组织中至少一个致免疫性且具突变信息的胜肽链的系统,其中,所述系统使用筛选步骤包括:

(a)取得复数个具突变信息的序列;

(b)从疾病相关的突变中决定出至少一个抗原决定位;

(c)将与所述至少一个抗原定位的免疫性相关的复数个特征整合,其中所述复数个特征包含:

第一特征,其用于计算胜肽链级数值;以及

第二特征,其用于计算样本级数值,

其中,所述第二特征包含所述疾病相关突变的群落性;

(d)决定所述复数个特征的重要性;

(e)以所述复数个特征的重要性决定所述至少一个抗原决定位的致免疫性数值;

(f)根据所述致免疫性数值的高低,将所述至少一个抗原决定位排名;以及

(g)依所述(f)步骤中的排名结果来选择至少一个所述致免疫性且具突变信息的胜肽链,其中,所述至免疫性且具突变信息之胜肽链包含至少一个抗原决定位且其可能可引起T细胞免疫反应,

其中,所述步骤(c)～(e)为利用机器学习模型达成,

在胜肽链级数值计算中,利用所述第一特征建构四套机器学习模型,利用机器学习回归器和数据分析方法整合这四个模型和其组合,最终模型包含加权后的因素、特征筛选和叠代调整后的最佳化机器学习模型,最后本模型利用已知具备致免疫性的抗原决定位进行验证,

所述复数个特征与下述多种有关:

至少一个所述抗原决定位在主要组织兼容性复合体类型一及类型二上的呈现、至少一个所述抗原决定位引起胞杀性T细胞免疫反应的能力有关、至少一个所述抗原决定位引起辅助性T细胞免疫反应的能力、蛋白质量、基因表现量、突变的变异频率、拷贝数、肿瘤纯度、异质性丧失的状态、等位基因量;

并且,所述第二特征包含由下列公式计算得的亚群落纯度S:

$$S = \frac{AF_{observed}}{M}\left(pC_t + C_n(1-p)\right),$$

$AF_{observed}$为观察等位基因频率值,p为肿瘤纯度,Ct为肿瘤变异状况(t)的拷贝数,Cn为

正常状况(n)的拷贝数,M为突变等位基因数量。

致免疫性的癌症特异抗原决定位的排名系统

技术领域

[0001] 本发明关于致免疫性的抗原决定位,尤其是一种判断、预测致免疫性癌症特异抗原决定位并将其排名的系统与方法。

背景技术

[0002] 以肿瘤特异抗原(tumor-specific antigen)引起免疫反应对抗肿瘤细胞为抵抗癌症带来契机。这些抗原被认定为肿瘤基因体学及免疫疗法的临床效用之间的连结。大致上,带有致癌突变的基因会产生带有突变的胜肽链。这些胜肽链接下来会和主要组织兼容性复合体(MHC)类型一和二结合,并且呈现到肿瘤细胞表面而作为抗原。免疫系统中特别是胞杀性T细胞(cytotoxic T cell)及辅助性T细胞(helper T cell)会辨识这些抗原为异己物而引发免疫反应。许多这类的抗原为肿瘤所专属的,而不曾被免疫系统辨识过。因此,它们是作为免疫疗法合适的标的,因为能将治疗用于肿瘤细胞而不伤害正常细胞。

[0003] 用肿瘤特异抗原来引起T细胞反应的方法展现出不同的结果。用这些抗原时,会面对两项障碍:第一项:免疫细胞必须辨识这些抗原为异己物,并在不攻击正常细胞的状况下引发免疫反应;第二项:即便是T细胞辨识到抗原为异己物,我们身体中的细胞(包括带有突变的肿瘤细胞)都有安全检查机制(称为免疫检查点(immune checkpoint))会防止T细胞进行长期高强度的攻击,而免疫疗法的成功需仰赖这些免疫检查点失去功能才有效。上述第二项障碍在近年来成为亮点。随着药物、临床试验、及目标癌种类的成长,免疫检查点抑制剂(immune checkpoint inhibitor)例如:抑制PD1、PDL1、及CTLA4等抗体被研发出来。但是,其中仍有很大的进步空间,因为免疫检查点抑制剂疗法的反应率只有大约20%或更低。因此在进行疗法前,事先筛选会对疗法敏感的病人较理想。带有高质或高量的肿瘤特异抗原,被认为和疗法的反应率及存活率有高度相关。为了加速对实施免疫检查点抑制剂的病人进行筛选,对于能准确地辨识这些抗原有了高度需求。

[0004] 相较之下,开发直接选择T细胞能辨识的抗原的进展仍相对处于停滞状态。T细胞在辨识外部抗原后会引发攻击,但是,目前仍没有有效的方法可以让T细胞辨识肿瘤特异抗原。治疗方法可以大致分为癌症疫苗和细胞输入疗法。治疗型癌症疫苗的目标,是从初始T细胞组成库(naïveT cell repertoire)中扩增T细胞,并重新活化原有T细胞以延缓肿瘤成长并且让肿瘤缩小。此疫苗由肿瘤特异抗原组成,且这些抗原是由其可引起免疫反应的能力而筛选出来的。但是,欠佳的致免疫性抗原筛选使得疫苗的进展被阻碍而效果不佳。细胞输入疗法直接聚焦于训练免疫细胞来攻击肿瘤细胞。免疫细胞(通常是T细胞或是树突细胞(dendritic cell))是从病人中收集并且进一步在实验室中培养。然后,筛选能够透过辨认肿瘤特异抗原而成功去除肿瘤细胞的T细胞,且再输入回病人体内。但是,这个方法因为抗原选择方法效率欠佳而造成低成功率。由前述两种方法中可得知,筛选出最佳的致免疫性肿瘤特异抗原是免疫疗法要达到临床效用所必须的。

[0005] 一种能可靠地判断致免疫性肿瘤特异抗原的方法具有广泛的应用并且对于多种免疫疗法策略皆有关键效用。目前判断肿瘤特异抗原的方法通常包含辨识突变以及预测抗

原决定位(即为抗原当中决定抗原能够引起免疫反应的位置(epitope))和主要组织兼容性复合体的结合亲和力。用来预测抗原决定位的工具有数种,但是彼此间的预测结果并不相符合,且实验只能够验证大约55％被预测出的抗原决定位(Rajasagi M et al., Blood.2014 Jul 17;124(3):453-62.)。典型的方法是基于胜肽链序列,而没有同时考虑主要组织兼容性复合体的两种类型及其相对应的免疫细胞。并且,每位病患或是样本都有其特殊性质会影响预测,而这些样本特异特性并未被目前癌症抗原(neoantigen)排名方法考虑到。这些特性可以概括以等位基因的量(allele dosage)来描述。当带有突变的等位基因及主要组织兼容性复合体的等位基因量较高时,免疫系统将有更高的机会可以辨识到肿瘤特异抗原,进而影响抗原决定位的预测。本发明所揭示判断、预测致免疫性T细胞抗原决定位并将其排序的系统与方法,其中运用到包含胜肽链级信息以及样本级信息。胜肽链级信息同时包含主要组织兼容性复合体类型一及二的呈现、CD4活化、及CD8活化,而样本级信息包含等位基因量,亦即带有突变的等位基因的群落性(clonality)、即主要组织兼容性复合体的等位基因数目。而且,此系统与方法整合了完整的因素清单,每样都是基于细胞生化反应过程、肿瘤特异性质、抗原呈现过程以及免疫活化过程。本发明所揭示用各个因素的权重来达到抗原决定位的最佳筛选方式。本发明还揭示一个抗原决定位的排名方法,可用于研发个人化治疗方法,如癌症疫苗、细胞输入疗法(adoptive cell transfer)、或是免疫检查点抑制剂等。

发明内容

[0006] 本发明揭示一种系统和方法用来从病人的肿瘤组织中判断抗原决定位,并且预测和排名抗原决定位是否可诱发针对疾病的免疫反应。本系统和方法同时考量抗原决定位的胜肽链级(peptide-level)信息和肿瘤组织的样本级(sample-level)信息,胜肽链级特性为胜肽链序列与主要组织兼容性复合体(major histocompatibility complex,MHC)类型一及类型二、辅助型T细胞(helper T cell)的活化、胞杀性T细胞(cytotoxic T cell)的活化相关的特性;样本级信息为肿瘤特异信息,包含突变等位基因之群落性(clonality of mutated allele)和主要组织兼容性复合体的数量。本系统和方法整合上述因素,并针对各种因素计算权重,代表可诱发免疫反应的程度。本系统和方法会给予每个抗原决定位一个致免疫性数值(immunogenicity score),并以该数值排列抗原决定位的优先级,以提供后续个人化医疗的参考依据。

[0007] 本系统需要次世代定序分析所得到的突变位点和拷贝数变异信息、其他定序相关信息包含:原始测序片段、主要组织兼容性复合体型别进行运算。在一些实施例中,主要组织兼容性复合体型可以与包含突变位点的个体为同一个或是不同个体。本系统会输出一组与突变相关的抗原决定位,包含:(a)具有突变的胜肽链序列;(b)胜肽链级数值(peptide-level score),其代表胜肽链可被呈现和活化免疫反应的能力;(c)样本级数值(sample-level score),其代表异质肿瘤中的突变群落性;(d)抗原决定位的排名,其代表使用于免疫疗法时,其预测的疗效的优先级。

[0008] 本案的系统和方法涵盖下述的部分或所有步骤:(1)辨识次世代定序分析所判断的突变,其包含:突变位点分析、突变位点标注、拷贝数分析、异质性丧失(loss of heterozygosity)分析、肿瘤纯度(tumor purity)分析;(2)分析具有突变位点的基因特性;

(3)从公开数据库中组织特异和疾病特异资料判断基因表现量;(4)从公开数据库中组织特异和疾病特异资料判断蛋白质量;(5)取得含有突变位点的胜肽链。与主要组织兼容性复合体类型一相关的胜肽链的长度为8～15个胺基酸,8～11个胺基酸长为首选。与主要组织兼容性复合体类型二相关的胜肽链的长度为9～23个胺基酸;(6)预测胜肽链与主要组织兼容性复合体类型一以及类型二的结合;(7)预测胜肽链能够活化CD8+T细胞和CD4+T细胞免疫反应的能力;(8)预测胜肽链是否会经由抗原呈现程序被呈现于细胞表面;(9)比较有突变的胜肽链和没突变的胜肽链之间的差异;(10)比较胜肽链和已知抗原的差异;(11)判断主要组织兼容性复合体类型一的等位基因量,并加入组织兼容性复合体类型一的分析;(12)结合和整合步骤1～11、计算胜肽链级因素的权重并预测胜肽链级的致免疫性;(13)计算群落突变位点的突变频率(clonal mutation frequency)并用于样本级数值;(14)整合胜肽链级数值和样本级数值为致免疫性数值;(15)判断基因拷贝缺失,当基因失去所有拷贝时致免疫性数值将设为零;(16)将致免疫性数值做排名。

[0009] 判断抗原决定位致免疫性之因素包括下述一种、多种、或任何组合但不限于(i)突变的变异频率;(ii)拷贝数变异;(iii)异质性丧失;(iv)肿瘤纯度;(v)突变等位基因的群落性;(vi)与已知抗原序列的同源性(抗原同源性);(vii)在主要组织兼容性复合体类型一交互作用中与野生型的相似性(自体相似性);(viii)在主要组织兼容性复合体类型二交互作用中与野生型的相似性(自体相似性);(ix)基因表现量;(x)蛋白质量;(xi)蛋白酶体切割位倾向性(proteasome cleavage);(xii)TAP运输效率(TAP transport);(xiii)主要组织兼容性复合体类型一结合亲合力;(xiv)主要组织兼容性复合体类型二结合亲合力;(xv)主要组织兼容性复合体类型一结合稳定性;(xvi)主要组织兼容性复合体类型一的等位基因量(allele dosage);(xvii)胜肽链序列和致免疫性T细胞抗原决定位序列的一致序列矩阵(consensus sequence matrix)的相似性。

[0010] 在一些实施例中,有些个体中含有较高的等位基因量,例如拥有纯和子对(homozygous pair)的主要组织兼容性复合体。较高的的等位基因量可能导致抗原呈现上的增加效果。本模型将等位基因量的增加效果加入主要组织兼容性复合体类型一的计算中。

[0011] 在胜肽链级数值计算中,我们利用上述的因素vi至xv建构四套机器学习模型。模型一预测主要组织兼容性复合体类型一的呈现,包含基因表现、蛋白质量、蛋白酶体切割位倾向性、TAP运输效率、主要组织兼容性复合体类型一结合亲合力、主要组织兼容性复合体类型一结合稳定性、主要组织兼容性复合体的等位基因量。模型二预测主要组织兼容性复合体类型二的呈现,包含主要组织兼容性复合体类型二结合亲合力。模型三预测辅助型T细胞的活化,包含自体相似性和抗原同源性。模型四预测胞杀性T细胞(cytotoxic T cell)的活化,其包含:自体相似性、抗原同源性和主要组织兼容性复合体类型一的致免疫性。模型三和模型四的训练资料来自体外T细胞免疫反应实验的结果。我们利用机器学习回归器和数据分析方法整合这四个模型和其组合。最终模型包含加权后的因素、特征筛选和叠代调整后的最佳化机器学习模型。最后本模型利用已知俱备致免疫性之抗原决定位进行验证。

[0012] 在样本级数值计算中,我们利用上述的因素(i)至(v)计算突变是否为群落突变。肿瘤可能包含数个群落,而每一个群落都拥有独特的基因组成。如一个突变出现于大多数群落中,该突变便被定义为群落突变,代表群落突变发生于癌症演化早期的「主干」(trunk)

时间点中。群落突变衍生的肿瘤特异抗原存在于大多数肿瘤细胞中,因此极有可能遭受免疫攻击。相反的,从亚群落衍生的肿瘤特异抗原属于「分支」(leaf)突变,只存在于少数肿瘤细胞中,就算遭受攻击,其他群落并不会有影响。判断群落突变需要利用最大似然的期望变异频率(maximum likelihood of expected frequency)计算突变等位基因的数量,接着估算亚群落的纯度(subclonal purity)。求得亚群落的纯度和肿瘤纯度后便可计算样本级数值。

[0013] 致免疫性数值包含胜肽链级数值和样本级数值。本系统将每一个抗原决定位依照致免疫性数值高低排名。本系统最终输出抗原决定位、致免疫性数值和排名。

附图说明

[0014] 图1为系统流程图。呈现系统进行运算时的整体流程和主要步骤。

[0015] 图2为输入和输出流程图。本系统需输入次世代定序分析所产出的变异和样本信息以及主要组织兼容性复合体信息。本系统会输出胜肽链序列、胜肽链级数值、样本级数值、致免疫性排名。

[0016] 图3为依据T细胞特征比较免疫反应,即有免疫反应及无反应的胜肽链级数值分布;A)为实施例5的特征包含CD4+细胞相关的自体相似性和抗原同源性;B)为实施例6的特征包含CD8+免疫性。上述图中p值是由独立双样本中位数差异检定所计算出来。

[0017] 图4为依据抗原表现特征和T细胞特征比较免疫反应,即有免疫反应及无反应的胜肽链级数值分布;A)为实施例7的特征包含CD4+细胞相关的自体相似性、抗原同源性和实施例4的特征;B)为实施例8的特征包含CD8+相关的CD8+免疫性和实施例1的特征。以上图中p值是由独立双样本中位数差异检定所计算出来。

[0018] 图5为依据抗原预测数值和T细胞特征比较免疫反应。有免疫反应及无反应的胜肽链级数值分布;A)实施例9的特征包含CD4+细胞相关的自体相似性、抗原同源性和实施例4的模型所预测的数值;B)实施例10的特征包含CD8+相关的CD8+免疫性和实施例1的模型所预测的数值。图中p值是由独立双样本中位数差异检定所计算出来。

[0019] 图6为依据抗原表现信息和两种T细胞特征比较免疫反应。有免疫反应及无反应的胜肽链级数值分布;A)为实施例11包含实施例5和实施例6;B)为实施例8包含实施例9和实施例10。

[0020] 图7为前50名胜肽链级数值所包含的有反应的胜肽链。长条图表示实验证实有CD8+反应的胜肽链数量,胜肽链同时在前50名胜肽链级数值中。每一张图代表一位病人。虚线代表每位病人中有免疫反应的胜肽链的总数。

具体实施方式

[0021] 在一些实施例,本发明揭示了在精准医疗的重大计划中,一个能够判断疾病特异抗原决定位(epitope),以及预测抗原决定位的致免疫性并将抗原决定位排名以更进一步用于病人个人化治疗的整合性系统及方法。该系统及方法整合了以定序为基础的突变位点分析(variant calling)、以定序为基础的拷贝数(copy number)判别、序列比对、相似性矩阵、机器学习、最佳化方法以及数学建模以用于致免疫性抗原决定位的准确且切实的判断(如图1所揭示)。该系统及方法考量了每个组成细胞生化反应过程(cellular process)、肿

瘤特异性质、抗原呈现(antigen presentation)过程以及免疫活化过程的组成性质(component),在每个过程中的组成性质则依照其在细胞中的实际功能来计算为系统中的考量因素(factor),接着每个因素根据其对抗原决定位的致免疫性的贡献程度来给予其权重,而一个给予权重的因素可帮助探讨抗原决定位具致免疫性的成因并促进临床及研究上的进展。上述系统考虑了代表胜肽链级信息(peptide-level information)及样本级信息(sample-level information)的因素并利用该因素将抗原决定位的致免疫性进行评分,在本案中也会根据预测出的抗原决定位的致免疫性数值来将判断出的抗原决定位进行排名。

[0022] 在本案中所使用的术语应被视为以描述实施例及申请专利范围为目的,任何术语的时态变化及字根改变不应被视为限制本案的效果,而任何术语的其他惯用同义词使用也不应被视为限制本案的效果。

[0023] 当有其他可能不同的替代做法时,本发明并不受限于本文中所描述的特定方法或规程或程序,而在本发明中所描述的具体实施例仅为范例,不应被解释为用来限制本案的范畴。

[0024] 在本发明中所使用的单数形式冠词:一个、所述等,具体来说也包含其字词所指内容的复数形式,除非内文中另有明定。

[0025] 「组成性质(component)」这个字词意指为突变的特异性质或基因的特异性质或细胞生化反应过程中的特定步骤或样本的特异性质。

[0026] 「因素(factor)」这个字词意指为组成性质在计算上的代表,其中因素可能以数学公式计算出来或以计算工具预测出来或做为一个类别被分类出来。

[0027] 「胜肽链(peptide)」这个字词意指为各种不同长度的胺基酸序列,其可能具有或不具有致免疫性,也可能是或不是和肿瘤相关。「抗原(antigen)」这个字词意指为具免疫性而能被免疫系统所辨识的胜肽链。「抗原决定位(epitope)」这个字词意指为一个能被呈现在细胞表面上的短片段抗原,抗原决定位可能透过「蛋白酶体(proteasome)」切割长片段抗原所产生。

[0028] 「癌症疫苗(cancer vaccine)」这个字词意指为以治疗癌症为目的并藉由增强人体免疫系统来对抗癌症的治疗性疫苗,其不应与在发生疾病之前以预防为目的而普遍施用的预防性疫苗有所混淆。

[0029] 「主要组织兼容性复合体(major histocompatibility complex,MHC)」这个字词意指为其任何变异型态和名称,其中包括但不限于其类型、其替代名称如「人类白血球抗原(human leukocyte antigen,HLA)」、其种类如A、B、C、DRB1、DPA1、DPB1、DQA1及DQB1等等。

[0030] 「突变(mutation)」这个字词除非另有明定,否则其意指为非同义体细胞突变(nonsynonymous somatic mutation),包含误义突变(missense mutation)、移码突变(frameshift mutation)及剪接位突变(splice site mutation)。「变异(variant)」这个字词包含突变但更进一步包含结构上的变异,包括拷贝数变异(copy number variation)、染色体的重组(rearrangement)、融合(fusion)、易位(translocation)及倒置(inversion)。体细胞变异(somatic variant)被定义为没有出现在生殖细胞中并出现在生命后期,特别是在癌症发展过程中的变异,变异可能导致肿瘤形成或为伴随癌症发生的变异。

[0031] 「定序深度(total depth)」这个字词意指为在特定基因位置定序出的测序片段(read)总量。

[0032] 在一些实施例中,本系统和方法可接收次世代定序资料(如图2所揭示)。次世代定序资料可以是(VCF)档案、(SAM)档案、(BAM)档案、FASTQ档案或任何其他未处理或处理过的档案。VCF档案包含所有突变在基因体上的信息,所述信息包含但不限于突变等位基因、参考等位基因(reference allele)、染色体、染色体上位置、突变的变异频率、和定序深度。在一些实施例中,使用者必须提供大片段变异的信息,包含:拷贝数变异、肿瘤纯度和异质性丧失。在一些实施例中,本系统可接收SAM档或是BAM档,上述的信息都可从SAM档或是BAM档求得。在一些实施例中,本系统可接收FASTQ档,上述的信息都可在与参考基因体(reference genome)进行序列比对后求得。

[0033] 本系统可接收主要组织兼容性复合体的类型(如图2所揭示)。在一些实施例中,主要组织兼容性复合体包含类型一的各种亚型但不限于A型、B型和C型,且需要四位数的分辨率。在一些实施例中,主要组织兼容性复合体包含类型二的各种亚型但不限于DRB1型、DPA1型、DPB1型、DQA1型和DQB1型,且需要四位数的分辨率。在一些实施例中,主要组织兼容性复合体类型可从次世代定序资料求得。

[0034] 本案描述判断抗原决定位和预测致免疫性的系统和方法,系统和方法包含下述一种、多种或任何组合的因素但不限于(i)突变位点分析所判定的突变的变异频率;(ii)拷贝数变异;(iii)突变的异质性丧失;(iv)肿瘤纯度;(v)突变等位基因的群落性;(vi)以序列比对判断与已知抗原序列的同源性(抗原同源性);(vii)计算突变胜肽链与主要组织兼容性复合体类型一的结合亲合力和野生型胜肽链与主要组织兼容性复合体类型一的结合亲合力,接着计算两种结合亲合力的比率决定突变胜肽链和野生型胜肽链的相似性;(viii)计算突变胜肽链与主要组织兼容性复合体类型二的结合亲合力和野生型胜肽链与主要组织兼容性复合体类型二的结合亲合力,接着计算两种结合亲合力的比率决定突变胜肽链和野生型胜肽链的相似性;(ix)由公开数据库中取得组织特异和疾病特异实验资料所判定的基因表现量;(x)由公开数据库中取得组织特异和疾病特异实验资料所判定的蛋白质量;(xi)以蛋白质降解资料所判定的蛋白酶体切割位倾向性;(xii)以TAP运输速率资料所判定的TAP运输效率;(xiii)以体外实验结果所判定的主要组织兼容性复合体类型一的结合亲合力;(xiv)以体外实验结果所判定的主要组织兼容性复合体类型二的结合亲合力;(xv)主要组织兼容性复合体类型一的结合稳定性;(xvi)主要组织兼容性复合体类型一的等位基因量;(xvii)以体外(in vitro)和离体(ex vivo)T细胞扩增实验结果所判定的胜肽链序列免疫性。

[0035] 抗原决定位会经由抗原呈现程序表现在细胞表面上。癌症特异抗原决定位的呈现程序会先从基因突变所衍生的突变胜肽练开始,经由蛋白酶体切割成小片段胜肽链,接着透过TAP进入内质网。在内质网中,胜肽链会与主要组织兼容性复合体结合,再一起被呈现于细胞表面上以提供免疫细胞辨识。上述抗原呈现程序中的每一个步骤都会影响抗原决定位的致免疫性。

[0036] 肿瘤中的突变并不见得会出现于所有肿瘤细胞中。如果一个突变所衍生的致免疫抗原决定位出现于大部分的肿瘤细胞中,免疫细胞更有机会辨识并攻击大部分肿瘤细胞,进而消灭肿瘤。因此,含有突变的细胞比例(以0~100%的变异频率作为代表)就成为一个判断抗原决定位致免疫性的重要依据。一个较高的变异频率代表一个突变存在于大部分肿瘤中,进而影响免疫攻击的效用。其他变异信息,包含拷贝数变异、异质性丧失、肿瘤纯度、

突变等位基因的分群性等都同样反映肿瘤细胞是否会产出代有突变的抗原决定位,导致肿瘤细胞遭受免疫攻击。

[0037] 产生出抗原决定位的一项先决条件是必须有基因表现。侦测肿瘤样本内的基因表现量可通过次世代定序(例如RNA定序)、微阵列(microarray)、实时聚合酶连锁反应(quantitative real-time PCR)或北方墨点法(Northern Blot)等实验测得。组织和癌症特异基因表现资料可从公开数据库取得。利用公开数据库中的资料可免除低表现基因所造成的噪声,得到真正有表现的基因的信息。虽然转录有一套复杂的调控机制,但目前已知基因体任何位置都可被转录,低表现量的基因仍然可被实验侦测到而造成过多噪声。因此,多数人在相同的疾病组织中都有表现的基因可以代表所述基因普遍会表现于疾病细胞中。基因表现后再经过转译才能产生抗原决定位。在一些资料集中,基因表现量可以是定性的表现方式,例如:低、中、高。在这些资料集中,定性资料可以被转成数值,例如:0、1、2、3。在其他资料集中,基因表现量可以是任何单位的数值,例如为一个比例或是自行判断的一个单位。在一些实施例中,本系统的机器学习模型可接收数值或转换而成的数值。一个没有被表现的基因会以低、0、或无表现表示。在其他实施例中,没有被表现的基因会被筛除。相反的,被上述实验资料所判定为表现量高的基因会有助于决定抗原决定位的量。一个高表现量的抗原决定位有较高的机会接触到主要组织兼容性复合体,也更容易被呈现于细胞表面。

[0038] 蛋白质量信息可利用质谱分析、免疫荧光法、免疫组织化学法或是西方墨点法(Western Blot)侦测。蛋白质量可从公开数据库中取得。带有突变的抗原决定位的蛋白质量有助于判断抗原决定位和主要组织兼容性复合体的结合。一个抗原决定位虽然可能俱备非常高的致免疫性,但其含量可能非常少而因此无法引起免疫反应。在一些资料集中,蛋白质量可以是定性的表现方式例如:低、中、高。在这些资料集中,定性资料可以被转换成数值,例如:0、1、2、3。在其他资料集中,蛋白质量可以是各种单位的数值,例如为一个比例或是自行判断的一个单位。在一些实施例中,本系统的机器学习模型接收数值或转换而成的数值。一个没有被侦测到的蛋白质会以低、0、或无表现代表。在其他实施例中,没有被侦测到的蛋白质会被筛除。相反的,被上述实验资料所判定为量高的蛋白质会有助于判断抗原决定位的量。一个高蛋白质量的抗原决定位有较高的机会接触到主要组织兼容性复合体,也更容易被呈现于细胞表面。

[0039] 本系统和方法会判断突变的胜肽链和没突变的野生型胜肽链之间的相似性。如果一个突变的胜肽链和野生型胜肽链非常相似,免疫细胞有可能会认为突变的胜肽链是自体的并且容忍它的存在。判断突变和野生型胜肽链的相似性可计算两者和主要组织兼容性复合体结合亲合力,并计算两个结合亲和力的比例。主要组织兼容性复合体类型一及类型二都会进行计算。

[0040] 本系统和方法会判断突变胜肽链和已知抗原的同源性。已知抗原来自细菌、病毒、或其他病原体,并且在大部分情况会引起T细胞的免疫反应。如果一个突变胜肽链和已知抗原非常相似,便更有可能引起免疫反应。我们利用序列比对分析决定突变胜肽链和已知抗原两者序列的同一性(identity)和有同一性的序列长度来判断抗原同源性。同源性是指突变胜肽链中包含有同一性抗原序列的比例。

[0041] 抗原决定位会在内质体接触到主要组织兼容性复合体。进入内质体前,突变胜肽链必须先被蛋白酶体切割成适合大小的抗原决定位。蛋白酶体切割位预测是一个0到1的数

值。在最佳情况下,抗原决定位内不包含可能被蛋白酶体切割的位置,有较小可能性在被呈现前就被分解。接着抗原决定位需要透过TAP蛋白质运输进内质体。TAP运输效率可用IC50数值表示,其中数值越低的IC50代表运输越有效率。可被有效运输的抗原决定位便有较高可能性接触主要组织兼容性复合体。

[0042] 抗原决定位必须和主要组织兼容性复合体结合才可以被呈现于细胞表面。主要组织兼容性复合体类型一可与8~15个胺基酸长度的抗原决定位结合,但8~11个胺基酸长度为首选。主要组织兼容性复合体类型二可与9~23个胺基酸长度的抗原决定位结合,但15和16个胺基酸长度为首选。抗原决定位和主要组织兼容性复合体的接合位(anchorposition)会随着不同主要组织兼容性复合体的类型而改变。抗原决定位上特定胺基酸与接合位的结合能力的对抗原呈现非常重要,对结合亲合力预测非常重要。IC50数值小于1500nM或是1000nM代表与主要组织兼容性复合体类型一及类型二有较好的结合亲合力,小于500nM更好,其代表抗原决定位极有可能与主要组织兼容性复合体结合和呈现于细胞表面上。

[0043] 除了结合亲合力之外,结合稳定性也是抗原呈现程序上重要的因子。抗原决定位或许可和主要组织兼容性复合体形成非常强的结合力,但若是其与主要组织兼容性复合体的结合时间长度不足则可能无法使其被呈现。亦即如果抗原决定位在被呈现之前就与主要组织兼容性复合体分离,此抗原决定位也就无法被呈现于细胞表面上。主要组织兼容性复合体的结合稳定性(其半衰期数值为0到1)代表抗原决定位与主要组织兼容性复合体结合的时间。上述结合时间越长则抗原决定位越有机会被呈现于细胞表面上。

[0044] 主要组织兼容性复合体类型一的免疫性代表抗原决定位的组成可引起免疫反应的能力。特定抗原决定位序列可能会对活化胞杀性T细胞的T细胞受体(T-cell receptor)有生化反应。触发T细胞受体是免疫攻击的第一步。主要组织兼容性复合体类型一免疫性是一个从-1到1的数值。较高的主要组织兼容性复合体类型一免疫性代表所述的抗原决定位较有可能引起T细胞扩增。

[0045] 有些个体拥有同样类型的主要组织兼容性复合体,代表来自父母的主要组织兼容性复合体是同一种类型,称作纯和子对(homozygous pair)。纯和子主要组织兼容性复合体等位基因(homozygous MHC allele)可能因为等位基因量(allele dosage)的效应造成加成效果。纯和子主要组织兼容性复合体等位基因拥有较高量可与抗原决定位结合的等位基因,因此提高抗原决定位呈现于细胞表面上的可能性。此外,细胞表面上有较高量的主要组织兼容性复合体也会提高T细胞辨识抗原决定位的可能性。因此,此加乘效果也被纳入计算中。

[0046] 依照上述任一有关于个体接受免疫疗法而得到的免疫反应资料的方法,没有免疫反应可能是因为负责抗原呈现机制的因子有缺陷。这些缺陷会令抗原呈现程序失去功能,因此就算抗原决定位具备致免疫性,他们还是无法呈现于细胞表面上。这些状况是决定免疫疗法疗效计算中的混乱因子(confounding factors),因此具有抗原呈现机制缺陷的个体将不纳入计算中。

[0047] 依照上述任何一种方法,每个因素的权重是从本判断系统所决定。本判断系统包含特征选择(feature selection)、机器学习(machine learning),验证(validation)、叠代模型调整和最佳化(iterative model tuning for optimization)。特征挑选后所包含的特征如下:

[0048] 胜肽链级特征

[0049] 主要组织兼容性复合体类型一呈现:基因表现量、蛋白质量、蛋白酶体切割位倾向性、TAP运输、主要组织兼容性复合体类型一结合亲合力、主要组织兼容性复合体类型一稳定性。

[0050] 主要组织兼容性复合体类型二呈现:主要组织兼容性复合体类型二结合亲合力。

[0051] 辅助型T细胞活动:自体相似性、抗原同源性。

[0052] 胞杀性T细胞活动:自体相似性、抗原同源性、主要组织兼容性复合体类型一免疫性

[0053] 样本级特征

[0054] 突变等位基因的群落性、主要组织兼容性复合体类型一的等位基因量。

[0055] 利用上述对应的特征,可计算出四种模型的胜肽链级数值,其中模型包含:主要组织兼容性复合体类型一、主要组织兼容性复合体类型二、辅助型T细胞活化、胞杀性T细胞活化。此外,主要组织兼容性复合体类型一包含样本级特征主要组织兼容性复合体类型一的等位基因量。我们利用机器学习回归器的结果整合这四种模型和它们的组合。我们还运用叠代模型调整和最佳化方法计算胜肽链级数值,并利用已知致免疫性抗原决定位验证上述模型。我们接者以数学或分析方式(例如:乘积)整合任二或多个模型得到最终模型。

[0056] 我们从突变等位基因的群落性之计算得到样本级数值。在癌症早期发展阶段所形成的突变是群落突变或是「主干」突变,其代表它们在癌症突变演化中发生于主干的时段而不在分支的时段。群落突变会出现于大部分癌细胞中。判断群落突变需要求得期望变异频率和观察变异频率的统计显著性,再计算突变等位基因的期望数量,接着用突变等位基因数量计算亚群落纯度。样本级数值便是亚群落纯度和肿瘤纯度的比例。

[0057] 将胜肽链数值和样本级数值整合便可得到致免疫性数值。本模型利用叠代计算进行参数调整,每一代将会重新建构和训练模型。拥有最好效能的模型就是最终模型。每一个抗原决定位将会计算一个致免疫性数值,而致免疫性数值的大小代表本系统对每一个癌症特异抗原决定位的排名。

[0058] 上述模型是利用机器学习方法所建构,其包含:主要组织兼容性复合体类型一、主要组织兼容性复合体类型二、辅助型T细胞活化、胞杀性T细胞活化,以及整合这些模型和样本级数值而求得的最终致免疫性数值。各种机器学型模型都可用于训练本模型上,例如:回归类模型(regression-based models)、树类模型(tree-based models)、贝氏类模型(Bayesian models)、支援矢量机(support vector machines)、提升类模型(boosting models)和神经网络类模型(neural network models)。

[0059] 本案的系统和方法可有助于癌症免疫学。本系统提供一套可以帮助病患的疗法的做法。本系统所判断的致免疫性抗原决定位可用于各人化医疗和各种免疫疗法,例如:免疫检察点抑制剂(immune-checkpoint inhibitor)、癌症疫苗(cancer vaccine)、或是细胞输入疗法(adoptive cell transfer)。癌症疫苗和细胞输入疗法中,排名好的抗原决定位可以提供制作疫苗或是训练免疫细胞时较高潜力的胜肽链选择。在免疫检察点抑制剂疗法中,致免疫性抗原决定位的数量可用于预测用药后的药效。本系统适用于针对个人的精准医疗或是用于广大群众的疗法上。

[0060] 实验示例

[0061] 实施例1:以胜肽链和主要组织兼容性复合体类型一(MHC class I)的结合亲和力(binding affinity)及结合稳定性(binding stability)来预测胜肽链被主要组织兼容性复合体类型一呈现的情形。

[0062] 一个胜肽链若要成为一个抗原则该胜肽链需要能被主要组织兼容性复合体呈现在细胞表面上且进而被免疫细胞所辨识。上述过程包含:胜肽链在抗原呈现细胞中由主要组织兼容性复合体类型二(MHC class II)呈现给CD4+T细胞、胜肽链在抗原呈现细胞中由主要组织兼容性复合体类型一呈现给CD8+T细胞以及胜肽链在肿瘤细胞中由主要组织兼容性复合体类型一呈现给CD8+T细胞。在本实施例中,我们以所选特征建构出一个模型来预测胜肽链被主要组织兼容性复合体类型一呈现的情形。

[0063] 我们以胜肽链结合亲和力及胜肽链结合稳定性这两个能影响胜肽链和主要组织兼容性复合体类型一结合的特性来建构出一个用以预测胜肽链被主要组织兼容性复合体类型一呈现情形的模型。我们使用NetMHC4.0软件(Andreatta M and Nielsen M,Bioinformatics(2016)Feb 15;32(4):511-7;Nielsen M,et al.,Protein Sci.,(2003)12:1007-17)来计算胜肽链和主要组织兼容性复合体类型一的结合亲和力(IC50),若主要组织兼容性复合体类型一不适用NetMHC4.0软件则改用NetMHCpan3.0软件(Nielsen M and Andreatta M,Genome Medicine(2016):8:33;Hoof I,et al.,Immunogenetics 61.1(2009):1-13)。我们使用NetMHCstabpan1.0软件(Rasmussen M,et al.,J Immunol.2016 Aug 15;197(4):1517-24)来计算胜肽链和主要组织兼容性复合体类型一的结合稳定性。我们从Bassani-Sternberg等人的研究(Bassani-Sternberg et al.,Molecular&Cellular Proteomics,2015 and Bassani-Sternberg et al.,Nature Communications,2016)中搜集训练资料来训练机器学习模型。而资料中的胜肽链只要是由不只一个基因所产生出来的胜肽链或是没有被标注为未修饰的胜肽链或是其胜肽链长度不介于9～11的胜肽链均会被从训练资料中移除。资料中若胜肽链及其对应的人类白血球抗原(HLA)类型被确认可形成胜肽链与人类白血球抗原类型的复合物且被呈现出来,则该资料则将被作为阳性(Positive)资料;若同一个胜肽链对应其他人类白血球抗原类型无法被确认可形成复合物且被呈现出来,则该资料则会被作为阴性(Negative)资料。每一个胜肽链与其对应的人类白血球抗原类型的结合亲和力及结合稳定性之计算方式则依上述进行计算。我们以胜肽链对主要组织兼容性复合体类型一的结合亲和力及结合稳定性作为特征来建构一个逻辑回归模型,并且利用Scikit-learn(Fabian Pedregosa et al.,JMLR(2011)Oct 12:2825-2830)中的LogisticRegression套件来预测主要组织兼容性复合体类型一的胜肽链呈现情形。在执行十次交叉验证(ten-fold cross-validation)后,训练资料中测试资料集的预测准确度(accuracy)及接收者操作特征曲线下面积(AUC ofROC)如表1中所揭示。在此模型中,我们发现特征中结合亲和力对于预测结果的影响力比结合稳定性还要强。

[0064] 实施例2:以胜肽链和主要组织兼容性复合体类型一的结合亲和力及胜肽链其所属基因的表现程度来预测胜肽链被主要组织兼容性复合体类型一呈现的情形。

[0065] 除了胜肽链和主要组织兼容性复合体类型一的结合能力之外,胜肽链被表现的程度对于胜肽链是否可被呈现也是很重要的。在本实施例中,我们以胜肽链其所属基因的表现情形以及胜肽链和主要组织兼容性复合体类型一的结合亲和力建构出一个模型来预测胜肽链被主要组织兼容性复合体类型一呈现的情形。

[0066] 我们以实施例1中所揭示的方式来计算胜肽链和主要组织兼容性复合体类型一的结合亲和力,并通过计算胜肽链其所属基因的RNA表现量以作为所述胜肽链其所属基因的表现程度。而各个胜肽链其所属基因的表现程度则从Illumina Body Map数据库(Petryszak R et al.,Nucleic Acids Res.2016 Jan 4;44(D1):D746-52)中获得。如实施例1中所揭示,我们从Bassani-Sternberg等人的研究中搜集训练资料来训练机器学习模型,并且以同样的方式进行筛选。若胜肽链及其对应的人类白血球抗原类型被确认可行成所述复合物且可被呈现出来,则该资料将被作为阳性资料;若同一个胜肽链对应其他人类白血球抗原类型无法被确认可行成所述复合物且被呈现出,则该资料将被作为阴性资料。每一个胜肽链和其对应的人类白血球抗原类型的结合亲和力以及胜肽链其所属基因的表现程度的产生方式则如前述。进一步地,我们以胜肽链对主要组织兼容性复合体类型一的结合亲和力以及胜肽链其所属基因表现程度作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测主要组织兼容性复合体类型一的胜肽链呈现情形。在执行十次交叉验证后,训练资料中的测试资料集的预测准确度及接收者操作特征曲线下面积如表1中所揭示。在此模型中,我们发现结合亲和力对于预测结果的影响力比基因表现程度还要强。

[0067] 实施例3:以胜肽链和主要组织兼容性复合体类型一的结合亲和力以及胜肽链其所属蛋白质表现量来预测胜肽链被主要组织兼容性复合体类型一呈现的情形。

[0068] 除了胜肽链和主要组织兼容性复合体类型一的结合能力之外,胜肽链的表现量也会影响到胜肽链被主要组织兼容性复合体所呈现的量。在本实施例中,我们以胜肽链和主要组织兼容性复合体类型一的结合亲和力以及胜肽链的表现量作为所选特征来建构出一个用以预测所述胜肽链被主要组织兼容性复合体类型一呈现情形的模型。

[0069] 我们以两种会影响到胜肽链和主要组织兼容性复合体类型一的结合能力及结合机率的特性(分别为胜肽链结合亲和力及胜肽链表现量)来建构出模型以预测胜肽链被主要组织兼容性复合体类型一呈现的情形。我们以实施例1中所揭露的方式来计算胜肽链和主要组织兼容性复合体类型一的结合亲和力。而在本实施例中,胜肽链其所属蛋白质表现量被用以作为代表胜肽链的表现量,且其更进一步被定义为一个基因所产生含有所述胜肽链的蛋白质表现量中最大的蛋白质表现量。我们从PaxDb蛋白质表现量数据库中的H.sapiens-Whole organism(Integrated)数据库(Wang,M.et al.,Proteomics 2015,10.1002/pmic.201400441)取得胜肽链其所属蛋白质的表现量。如实施例1中所揭示,我们从Bassani-Sternberg等人的研究中搜集训练资料来训练所述机器学习模型,并且以同样的方式进行筛选。若资料中的胜肽链及其对应的人类白血球抗原类型被确认可形成所述复合物且可被呈现出,则该资料将被作为阳性资料;若同一个胜肽链对应其他人类白血球抗原类型无法被确认可形成复合物且无法被呈现出所述胜肽链,则该资料则将被作为阴性资料。每一个胜肽链和其对应的人类白血球抗原类型的结合亲和力以及胜肽链其所属蛋白质的表现量的计算方式则如前所述。我们以胜肽链对主要组织兼容性复合体类型一的结合亲和力以及胜肽链其所属蛋白质表现量作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测主要组织兼容性复合体类型一的胜肽链呈现情形。在执行十次交叉验证后,训练资料中的测试资料集的预测准确度及接收者操作特征曲线下面积如表1中所揭示。在模型中,我们发现结合亲和力对于预测结果的影响力比蛋白质

表现量还要强。

[0070] 实施例4:以胜肽链和主要组织兼容性复合体类型二的结合亲和力来预测胜肽链被主要组织兼容性复合体类型二呈现的情形。

[0071] 一个胜肽链若要成为一个抗原则所述胜肽链需能被主要组织兼容性复合体呈现在细胞表面上进而被免疫细胞所辨识。上述过程包含:胜肽链在抗原呈现细胞中由主要组织兼容性复合体类型二呈现给CD4+T细胞、胜肽链在抗原呈现细胞中由主要组织兼容性复合体类型一呈现给CD8+T细胞以及胜肽链在肿瘤细胞中由主要组织兼容性复合体类型一呈现给CD8+T细胞。在本实施例中,我们建构出一个模型来预测胜肽链被主要组织兼容性复合体类型二呈现的情形。

[0072] 我们以胜肽链和主要组织兼容性复合体类型二的结合亲和力来建构出一个用以预测胜肽链被主要组织兼容性复合体类型二呈现情形的模型。我们使用NetMHCII2.2软件(Nielsen M,et al.,BMC Bioinformatics.2007 Jul 4;8:238)来计算胜肽链和主要组织兼容性复合体类型二的结合亲和力,若主要组织兼容性复合体类型二不适用NetMHCII2.2软件则改用NetMHCIIpan3.1软件(Andreatta M,et al.,Immunogenetics.2015 Nov;67(11-12):641-50)。我们从Chong等人的研究(Chong et al.,Molecular&Cellular Proteomics,2017)中搜集训练资料来训练机器学习模型,而资料中长度小于9的胜肽链则被从训练资料中移除。资料中若胜肽链及其对应的人类白血球抗原类型被确认可形成胜肽链与人类白血球抗原类型的复合物且被呈现出来,则资料则将被作为阳性资料;若同一个胜肽链对应其他人类白血球抗原类型无法被确认可形成所述复合物且被呈现出来,则资料则会被作为阴性资料。我们以胜肽链对主要组织兼容性复合体类型二的结合亲和力为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测主要组织兼容性复合体类型二的胜肽链呈现情形。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表1中所揭示。

[0073] 表1:实施例1～4的模型表现结果。

[0074]

	实施例 1	实施例 2	实施例 3	实施例 4

[0075]

AUC of ROC	0.88	0.88	0.88	0.78
Accuracy	0.82	0.81	0.81	0.71

[0076] 实施例5:以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性(self-similarity for MHC Class II)以及与已知抗原的同源性(homology)来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0077] 一个胜肽链若要成为一个具致免疫性的抗原,除了被主要组织兼容性复合体类型一及类型二呈现之外,胜肽链引起CD4+T细胞及CD8+T细胞免疫反应的能力也是很重要的。在本实施例中,我们以所选特征建构出一个模型来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0078] 我们以两个能影响胜肽链被CD4+T细胞辨识的特性来建构出一个用以预测胜肽链引起CD4+T细胞免疫反应能力的模型。两个特性为胜肽链和人类蛋白质序列在与主要组织

兼容性复合体类型二结合上的相似性(简称为胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性)以及胜肽链与已知抗原的同源性。我们以三个步骤来计算胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性,首先,我们从ENSEMBL GRch37数据库中获取所有人类蛋白质序列并且裁剪出在9～23胺基酸长度中所有可能的胺基酸序列。由于可能有胜肽链并非由人类蛋白质序列突变而来,因此我们模仿了突变胜肽链和野生型胜肽链间的关系,将胜肽链和裁剪出来的人类蛋白质序列做比对并从裁剪出来的序列中挑选出具有相同长度且只有一个胺基酸差异的人类蛋白质序列作为自体胜肽链(self-peptide)。其次,我们以实施例4中所揭露的方式分别计算所述胜肽链以及其对应的自体胜肽链和主要组织兼容性复合体类型二的结合亲和力。最后,我们以前述两个计算出的结合亲和力中较小的数值除以较大的数值来定义胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性。若有胜肽链只能从所述裁剪出来的序列中比对出具有两个或两个以上胺基酸差异的序列,则将胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性定为0。我们利用BLAST(https://blast.ncbi.nlm.nih.gov/Blast.cgi)的方法将胜肽链和已知抗原序列做比对来计算胜肽链与已知抗原的同源性。我们从IEDB数据库(www.iedb.org)的抗原资料集中选择被标注为病毒或细菌的抗原序列作为已知抗原序列。若一个胜肽链其序列具有较高比例能比对上一个已知抗原序列,则胜肽链被视为和已知抗原序列有同源性。若有胜肽链无法比对上任何已知抗原,则将胜肽链与已知抗原的同源性定为0。我们从IEDB数据库档名为“tcell_full_v3.csv”的档案中,搜集主要组织兼容性复合体被标注为类型二且有标注CD4+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD4+T细胞免疫反应的能力。而资料中若其胜肽链长度不介于9～30个胺基酸或是其细胞类型并非一个正常T细胞或是其试验类别组并不是被标注为侦测免疫讯号释放或T细胞活化或T细胞与抗原呈现细胞的结合,则前述资料均会被从训练资料中移除。我们以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性以及与已知抗原的同源性作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测胜肽链引起CD4+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。

[0079] 在完成模型训练后,我们从OttP等人的研究(Ott P et al.,Nature,2017)中搜集经实验测试过T细胞免疫反应并有标注CD4+T细胞免疫反应结果的资料作为测试资料来测试所述模型。我们将测试资料中的每个胜肽链如前进行特征计算并且以所计算出的特征和所训练出的参数来计算所述模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图(boxplot)以及两笔资料的预测数值分布利用独立双样本中位数差异检定(Wilcoxon rank-sumtest)所计算出来的p值如图3A所揭示。

[0080] 实施例6:以胜肽链的免疫性来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0081] 一个胜肽链若要成为一个具致免疫性的抗原,除了被主要组织兼容性复合体类型一及类型二呈现之外,所述胜肽链引起CD4+T细胞及CD8+T细胞免疫反应的能力也是很重要的。在本实施例中,我们以所选特征建构出一个模型来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0082] 我们以胜肽链的免疫性此能影响胜肽链被CD8+T细胞辨识的特性来建构出一个用以预测胜肽链引起CD8+T细胞免疫反应能力之模型。胜肽链的免疫性为利用IEDB

immunogenicity predictor软件(Calis JJ,PLoS Comput Biol.(2013)Oct 9(10):e1003266)所计算而来。我们从IEDB数据库档名为“tcell_full_v3.csv”的档案中,搜集主要组织兼容性复合体被标注为类型一且有标注CD8+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD8+T细胞免疫反应的能力。而资料中若其胜肽链长度不介于8～11个胺基酸或是其细胞类型并非一个正常T细胞或是其试验类别组并不是被标注为侦测免疫讯号释放或T细胞活化或T细胞与抗原呈现细胞的结合,则前述资料均会被从训练资料中移除。我们以IEDB immunogenicity predictor软件预测出来的胜肽链的免疫性作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测胜肽链引起CD8+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。

[0083] 在完成模型训练后,我们从OttP等人的研究(Ott P et al.,Nature,2017)中搜集经实验测试过T细胞免疫反应并有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试模型。我们将测试资料中的每个胜肽链如前所述进行特征计算并且以所计算出的特征和所训练出的参数来计算所述模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图3B所揭示。

[0084] 实施例7:以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、与已知抗原的同源性以及实施例4中的特征来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0085] 一个抗原决定位要能引起CD4+T细胞的免疫反应需要能先被抗原呈现细胞呈现出来给CD4+T细胞辨识。在本实施例中,我们考虑了胜肽链引起免疫反应的能力以及胜肽链被主要组织兼容性复合体类型二呈现在抗原呈现细胞上的能力并且以所选特征建构出一个模型来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0086] 我们以影响胜肽链被CD4+T细胞辨识的特性以及影响胜肽链被主要组织兼容性复合体类型二呈现的特性来建构出一个用以预测胜肽链引起CD4+T细胞免疫反应能力的模型。该特性为胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、胜肽链与已知抗原的同源性以及胜肽链和主要组织兼容性复合体类型二的结合亲和力。我们以实施例5中所揭示的方式来计算胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性以及胜肽链与已知抗原的同源性,并且以实施例4中所揭示的方式来计算胜肽链和主要组织兼容性复合体类型二的结合亲和力。如实施例5中所揭示,我们从IEDB数据库中,搜集有标注CD4+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD4+T细胞免疫反应的能力。其资料筛选方式如实施例5中所揭示。我们以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、与已知抗原的同源性以及和主要组织兼容性复合体类型二的结合亲和力作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测胜肽链引起CD4+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。在此模型中,我们发现胜肽链和主要组织兼容性复合体类型二的结合亲和力对于预测结果的影响力比胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性以及与已知抗原的同源性还要强。

[0087] 在完成模型训练后,如实施例5中所揭示,我们从Ott P等人的研究中搜集有标注

CD4+T细胞免疫反应结果的资料作为测试资料来测试所述模型。我们将测试资料中的每个胜肽链如前所述进行特征计算并且以所计算出的特征和所训练出的参数来计算所述模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图4A所揭示。

[0088] 实施例8:以胜肽链的免疫性以及实施例1中的特征来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0089] 一个抗原决定位要能引起CD8+T细胞的免疫反应需要能先被主要组织兼容性复合体类型一呈现出来给CD8+T细胞辨识。在本实施例中,我们考虑了胜肽链引起免疫反应的能力以及胜肽链被主要组织兼容性复合体类型一呈现的能力并且以所选特征建构出一个模型来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0090] 我们以影响胜肽链被CD8+T细胞辨识的特性以及影响胜肽链被主要组织兼容性复合体类型一呈现的特性来建构出一个用以预测胜肽链引起CD8+T细胞免疫反应能力之模型。该特性为胜肽链的免疫性、胜肽链和主要组织兼容性复合体类型一的结合亲和力以及结合稳定性。我们以实施例6中所揭示的方式来计算胜肽链的免疫性,并且以实施例1中所揭示的方式来计算胜肽链和主要组织兼容性复合体类型一的结合亲和力以及结合稳定性。如实施例6中所揭示,我们从IEDB数据库中,搜集有标注CD8+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD8+T细胞免疫反应的能力。其资料筛选方式如实施例6中所揭示。我们以胜肽链之免疫性、胜肽链和主要组织兼容性复合体类型一的结合亲和力以及结合稳定性作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测胜肽链引起CD8+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。在此模型中,我们发现胜肽链和主要组织兼容性复合体类型一的结合稳定性对于预测结果的影响力最强,其次为胜肽链和主要组织兼容性复合体类型一的结合亲和力,最后则为胜肽链的免疫性。

[0091] 在完成模型训练后,如实施例6中所揭示,我们从Ott P等人的研究中搜集有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试模型。我们将测试资料中的每个胜肽链如前所述进行特征计算并且以所计算出的特征和所训练出的参数来计算模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图4B所揭示。

[0092] 实施例9:以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、与已知抗原的同源性以及实施例4中训练的模型所计算出的预测数值来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0093] 一个抗原决定位要能引起CD4+T细胞的免疫反应需要能先被抗原呈现细胞呈现出来给CD4+T细胞辨识。在本实施例中,我们考虑了胜肽链引起免疫反应的能力以及以实施例4中所建构的模型考虑胜肽链被主要组织兼容性复合体类型二呈现在抗原呈现细胞上的能力并且以所选特征建构出一个模型来预测胜肽链引起CD4+T细胞免疫反应的能力。

[0094] 我们以影响胜肽链被CD4+T细胞辨识的特性以及影响胜肽链被主要组织兼容性复合体类型二呈现的特性来建构出一个用以预测胜肽链引起CD4+T细胞免疫反应能力的模型。该特性为胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、胜肽链与已

知抗原的同源性以及实施例4中揭示的主要组织兼容性复合体类型二抗原呈现模型所计算出的预测数值。我们以实施例5中所揭示的方式来计算胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性以及胜肽链与已知抗原的同源性。为了要计算主要组织兼容性复合体类型二抗原呈现模型的预测数值,我们以实施例4中所揭示的方式来计算胜肽链和主要组织兼容性复合体类型二的结合亲和力,并且我们以前述计算出的特征以及实施例4中所训练出的参数来计算主要组织兼容性复合体类型二抗原呈现模型的预测数值。如实施例5中所揭示,我们从IEDB数据库中,搜集有标注CD4+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD4+T细胞免疫反应的能力。其资料筛选方式如实施例5中所揭示。我们以胜肽链在与主要组织兼容性复合体类型二结合上的自体相似性、与已知抗原的同源性以及主要组织兼容性复合体类型二抗原呈现模型的预测数值作为特征来建构逻辑回归模型,并且利用Scikit-learn中的Logistic Regression套件来预测胜肽链引起CD4+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。

[0095] 在完成模型训练后,如实施例5中所揭示,我们从Ott P等人的研究中搜集有标注CD4+T细胞免疫反应结果的资料作为测试资料来测试所述模型。我们将测试资料中的每个胜肽链如前所述进行特征计算并且以所计算出的特征和所训练出的参数来计算所述模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图以及所述两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图5A所揭示。

[0096] 实施例10:以胜肽链的免疫性以及实施例1中训练的模型所计算出的预测数值来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0097] 一个抗原决定位要能引起CD8+T细胞的免疫反应需要能先被主要组织兼容性复合体类型一呈现出来给CD8+T细胞辨识。在本实施例中,我们考虑了胜肽链引起免疫反应的能力以及如实施例1中的模型所揭示的胜肽链被主要组织兼容性复合体类型一呈现的能力并且以所选特征建构出一个模型来预测胜肽链引起CD8+T细胞免疫反应的能力。

[0098] 我们以影响胜肽链被CD8+T细胞辨识的特性以及影响胜肽链被主要组织兼容性复合体类型一呈现的特性来建构出一个用以预测胜肽链引起CD8+T细胞免疫反应能力的模型。该特性为胜肽链的免疫性以及实施例1中揭示的主要组织兼容性复合体类型一抗原呈现模型所计算出的预测数值。我们以实施例6中所揭示的方式来计算胜肽链的免疫性。为了要计算主要组织兼容性复合体类型一抗原呈现模型的预测数值,我们以实施例1中所揭示的方式来计算胜肽链和主要组织兼容性复合体类型一的结合亲和力以及结合稳定性,并且我们以前述计算出的特征以及实施例1中所训练出的参数来计算主要组织兼容性复合体类型一抗原呈现模型的预测数值。如实施例6中所揭示,我们从IEDB数据库中,搜集有标注CD8+T细胞免疫反应结果的资料作为训练资料来训练机器学习模型以预测胜肽链引起CD8+T细胞免疫反应的能力。其资料筛选方式如实施例6中所揭示。我们以胜肽链之免疫性以及主要组织兼容性复合体类型一抗原呈现模型的预测数值作为特征来建构逻辑回归模型,并且利用Scikit-learn中的LogisticRegression套件来预测胜肽链引起CD8+T细胞免疫反应的能力。在执行十次交叉验证后,训练资料中测试资料集的预测准确度及接收者操作特征曲线下面积如表2中所揭示。在此模型中,我们发现主要组织兼容性复合体类型一抗原呈现模型的预测数值对于预测结果的影响力比所述特征胜肽链之免疫性还要强。

[0099] 在完成模型训练后,如实施例6中所揭示,我们从Ott P等人的研究中搜集有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试所述模型。我们将测试资料中的每个胜肽链如前所述进行特征计算并且以所计算出的特征和所训练出的参数来计算模型的预测数值。其阳性反应和阴性反应资料各自的预测数值盒形图以及所述两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图5B所揭示。

[0100] 表2:实施例5～10的模型表现结果。

[0101]

	实施例5	实施例6	实施例7	实施例8	实施例9	实施例10
AUC of ROC	0.59	0.56	0.58	0.64	0.59	0.61
准确度	0.59	0.55	0.6	0.62	0.63	0.59

[0102] 实施例11:以实施例1及实施例4～6中所揭示的抗原呈现能力以及引起免疫反应能力来预测一个胜肽链是否为一个免疫原(immunogen)。

[0103] 一个胜肽链若要能成为一个免疫原而能引起免疫反应需要具备能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞和CD8+T细胞免疫反应的能力。因此在本实施例中,我们整合实施例1及实施例4～6中所揭示的内容并计算胜肽链级数值(peptide-level score)来预测胜肽链是否为免疫原。

[0104] 我们通过整合实施例1和实施例4中抗原呈现能力的信息以及实施例5和实施例6中引起CD4+T细胞和CD8+T细胞免疫反应能力的信息建构了一个整合模型组来计算胜肽链级数值。胜肽链级数值代表了一个胜肽链能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞和CD8+T细胞免疫反应的能力。我们通过将实施例1及实施例4～6中每个所揭示模型的预测数值相乘来计算胜肽链级数值。需要注意的是,在执行主要组织兼容性复合体类型二呈现预测以及CD4+T细胞免疫反应预测时的胜肽链长度会长于执行主要组织兼容性复合体类型一呈现预测以及CD8+T细胞免疫反应预测时的胜肽链长度。为了解决这个问题,我们对每个可能的较长胜肽链且所述较长胜肽链须包含CD8+T细胞免疫反应预测中的较短胜肽链,去计算其在主要组织兼容性复合体类型二呈现预测以及在CD4+T细胞免疫反应预测中的特征,并从较长胜肽链及其特征中选出引起CD4+T细胞免疫反应能力最强的胜肽链及其特征来整合CD4+T细胞和CD8+T细胞间免疫反应预测的信息。我们从PatrickA.Ott等人的研究中搜集有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试胜肽链级数值。我们计算测试资料中每个胜肽链的胜肽链级数值,而其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图6A所揭示。

[0105] 实施例12:以实施例7～8中所揭示的抗原呈现能力以及引起免疫反应能力来预测一个胜肽链是否为一个免疫原。

[0106] 同理实施例11,要预测一个胜肽链为一个免疫原,胜肽链需要具备能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞和CD8+T细胞免疫反应的能力。在本实施例中,我们整合实施例7～8中所揭示的内容并计算胜肽链级数值来预测胜肽链是否为免疫原。

[0107] 我们通过整合实施例7～8中抗原呈现能力的信息以及引起免疫反应能力的信息建构了一个整合模型组来计算胜肽链级数值。胜肽链级数值代表了一个胜肽链能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞

和CD8+T细胞免疫反应的能力。我们通过将实施例7～8中每个所揭示模型的预测数值相乘来计算胜肽链级数值。需要注意的是,在执行主要组织兼容性复合体类型二呈现预测以及CD4+T细胞免疫反应预测时的胜肽链长度会长于执行主要组织兼容性复合体类型一呈现预测以及CD8+T细胞免疫反应预测时的胜肽链长度。为了解决这个问题,我们对每个可能的较长胜肽链且所述较长胜肽链须包含CD8+T细胞免疫反应预测中的较短胜肽链,去计算其在主要组织兼容性复合体类型二呈现预测以及在CD4+T细胞免疫反应预测中的特征,并从较长胜肽链及其特征中选出引起CD4+T细胞免疫反应能力最强的胜肽链及其特征来整合CD4+T细胞和CD8+T细胞间免疫反应预测的信息。我们从PatrickA.Ott等人的研究中搜集有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试所述胜肽链级数值。我们计算测试资料中每个胜肽链的胜肽链级数值,而其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图6B所揭示。

[0108] 实施例13:以实施例9～10中所揭示的抗原呈现能力以及引起免疫反应能力来预测一个胜肽链是否为一个免疫原。

[0109] 同理实施例11,要预测一个胜肽链为一个免疫原,胜肽链需要具备能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞和CD8+T细胞免疫反应的能力。在本实施例中,我们整合实施例9～10中所揭示的内容并计算胜肽链级数值来预测胜肽链是否为免疫原。

[0110] 我们通过整合实施例9～10中抗原呈现能力的信息以及引起免疫反应能力的信息建构了一个整合模型组来计算胜肽链级数值。胜肽链级数值代表了一个胜肽链能被主要组织兼容性复合体类型一和主要组织兼容性复合体类型二呈现的能力以及能引起CD4+T细胞和CD8+T细胞免疫反应的能力。我们通过将实施例9～10中每个所揭示模型的预测数值相乘来计算胜肽链级数值。需要注意的是,在执行主要组织兼容性复合体类型二呈现预测以及CD4+T细胞免疫反应预测时的胜肽链长度会长于执行主要组织兼容性复合体类型一呈现预测以及CD8+T细胞免疫反应预测时的胜肽链长度。为了解决这个问题,我们对每个可能的较长胜肽链且该较长胜肽链须包含CD8+T细胞免疫反应预测中的较短胜肽链,去计算其在主要组织兼容性复合体类型二呈现预测以及在CD4+T细胞免疫反应预测中的特征,并从较长胜肽链及其特征中选出引起CD4+T细胞免疫反应能力最强的胜肽链及其特征来整合CD4+T细胞和CD8+T细胞间免疫反应预测的信息。我们从PatrickA.Ott等人的研究中搜集有标注CD8+T细胞免疫反应结果的资料作为测试资料来测试所述胜肽链级数值。我们计算测试资料中每个胜肽链的胜肽链级数值,而其阳性反应和阴性反应资料各自的预测数值盒形图以及两笔资料的预测数值分布利用独立双样本中位数差异检定所计算出来的p值如图6C所揭示。

[0111] 实施例14:以胜肽链级数值从病人检体中将胜肽链作为免疫原进行判断、评分、排名与评估。

[0112] 为了将一个病人的胜肽链作为免疫原进行判断、评分、排名与评估,我们从OttP等人的研究中获取胜肽链资料并以我们的评分方法来比较所述胜肽链资料。OttP等人根据6个黑色素瘤病人的突变从中设计出免疫长胜肽链(immunizing long peptides,IMP),并以实验测试所述免疫长胜肽链在免疫细胞上的致免疫性。免疫长胜肽链其胺基酸长度介于15

～30之间,用以测试CD8+T细胞免疫反应的胜肽链则较短,其胺基酸长度为9～10,而能引起免疫反应的胜肽链则在测试后被辨识出来。因此我们从免疫长胜肽链中获取胺基酸长度为9～10、带有突变的胺基酸位点、较短且重叠的胜肽链,并以实施例11～13中所揭示的方法来计算每个病人的较短胜肽链的胜肽链级数值。因为50为一个在疫苗胜肽链数量选择上常见的数字,所以我们挑选出了胜肽链级数值前50高的胜肽链。以实施例12中揭示的方法所计算出1号病人的资料如表3中所揭示,而每个病人以实施例11～13所揭示的方法来计算出的前50名胜肽链中,其能引起CD8+T细胞免疫反应的胜肽链数量如图7所揭示。

[0113] 表3:以实施例12所揭示的方法计算出1号病人中胜肽链级数值前50高的资料。

[0114]

排名	病人编号	IMP 编号	序列	胜肽链级数值	CD8 免疫反应
1	1	1-IMP04	KLKFVTLVF	0.345	1
2	1	1-IMP07	RFLEYLPLRF	0.336	1
3	1	1-IMP02	VQKVASKIPF	0.328	0
4	1	1-IMP06	TLFHTFYEL	0.321	0
5	1	1-IMP06	TLFHTFYELL	0.317	0
6	1	1-IMP11	KFGDLTNNF	0.303	0
7	1	1-IMP26	PREEFLRLC	0.302	0
8	1	1-IMP08	KLFESKAEL	0.299	0
9	1	1-IMP17	LCPREEFLR	0.298	0
10	1	1-IMP02	PFPDRITEES	0.297	0

[0115]

11	1	1-IMP04	VLAKKLKFV	0.295	0
12	1	1-IMP03	KKKWFLFQD	0.292	0
13	1	1-IMP02	PFPDRITEE	0.291	0
14	1	1-IMP07	HTELERFLE	0.288	0
15	1	1-IMP08	KLFESKAELA	0.287	0
16	1	1-IMP07	TELERFLEY	0.286	0
17	1	1-IMP07	LLHTELERF	0.286	0
18	1	1-IMP02	FPDRITEES	0.285	0
19	1	1-IMP19	VSVGDFSQEF	0.283	0
20	1	1-IMP02	IPFPDRITEE	0.280	0
21	1	1-IMP17	CPREEFLRLC	0.280	0
22	1	1-IMP12	ALFASRPRF	0.279	0
23	1	1-IMP03	FLFQDSKKI	0.277	0
24	1	1-IMP08	DKLFESKAE	0.277	0
25	1	1-IMP03	SKKKWFLFQD	0.276	0
26	1	1-IMP02	IPFPDRITE	0.276	0
27	1	1-IMP12	GGALFASRP	0.276	0
28	1	1-IMP07	LHTELERFLE	0.275	0
29	1	1-IMP26	LSPREEFLR	0.275	0
30	1	1-IMP26	SPREEFLRLC	0.274	0
31	1	1-IMP19	VGDFSQEFS	0.272	0
32	1	1-IMP03	KKKWFLFQDS	0.271	0
33	1	1-IMP10	DSGIPENSFN	0.271	0
34	1	1-IMP19	VGDFSQEFSP	0.271	0
35	1	1-IMP12	RGGALFASRP	0.270	0
36	1	1-IMP19	SVGDFSQEF	0.269	0
37	1	1-IMP10	LADSGIPEN	0.267	0
38	1	1-IMP04	KKLKFVTLV	0.266	0
39	1	1-IMP02	SKIPFPDRIT	0.265	0
40	1	1-IMP10	GIPENSFNV	0.265	0

[0116]

41	1	1-IMP02	KIPFPDRIT	0.264	0
42	1	1-IMP11	GKFGDLTNN	0.264	0
43	1	1-IMP08	DSDKLFESK	0.263	0
44	1	1-IMP08	EDSDKLFES	0.263	0
45	1	1-IMP07	LLHTELERFL	0.262	0
46	1	1-IMP12	RRGGALFASR	0.262	0
47	1	1-IMP10	IPENSFNVS	0.262	0
48	1	1-IMP02	ASKIPFPDR	0.261	0
49	1	1-IMP06	LFHTFYELLI	0.261	0
50	1	1-IMP07	ILLHTELERF	0.259	0

[0117] 实施例15:从突变等位基因(mutated allele)的群落性(clonality)来判断样本级数值(sample-level score)。

[0118] 我们计算样本特异资料中突变等位基因的群落性。我们从每个病人身上取得福尔马林固定石蜡包埋(formalin-fixed paraffin-embedded,FFPE)组织样本及配对的周边血液单核细胞(peripheral blood mononuclear cell)样本,使用QIAamp®DNA FFPE Tissue试剂组(QIAGENR,Hilden,Germany)从样本中抽取基因体DNA,将抽取出的DNA针对18,136组扩增子(amplicon)位置进行多重PCR放大,且其外显子组(exome)利用Ion Proton™(Thermo Fisher Scientific,Waltham,MA)系统以及Ion PI芯片(Thermo Fisher Scientific,Waltham,MA)并接着以厂商推荐的规程进行定序。在定序之后,我们将原始测序片段透过厂商提供的Torrent Variant Caller v.4.4软件(TVC)进行处理并产生.bam和.vcf档案,TVC软件也计算出了每个变异的变异频率,而变异则利用Variant Effect Predictor v.74软件进行标注,我们接着利用dbSNP 138数据库、1000Genome数据库以及正常的配对血液样本过滤掉单核苷酸多型性(single nucleotide polymorphism,SNP)及生殖细胞突变(germline mutation),剩余的变异则以人工方式进行检查。我们利用ONCOCNV软件及ADTEx软件分别从.bam档案中判断出其拷贝数(copy number)及肿瘤纯度(tumor purity),而异质性丧失(loss of heterozygosity,LOH)的判断则定为福尔马林固定石蜡包埋样本及正常的配对血液样本中同一单核苷酸多型性位点其等位基因频率(allele frequency)差异大于8%。

[0119] 突变等位基因其群落性的判断方式为先基于突变等位基因的期望等位基因频率值(expected allele frequency)和观察等位基因频率值(observed allele frequency)之间差异的统计显著性来给定突变等位基因数量的期望值,并接着估算其亚群落纯度(subclonal purity)。期望等位基因频率值的计算方式(McGranahan et al.,Science (2016)Mar 25;351(6280):1463-9)如下所揭示:

[0120] $$AF_{expected} = \frac{pM}{pC_t + C_n(1-p)}$$

[0121] 其中AFexpected表示期望等位基因频率值,t表示肿瘤变异之状况条件,p表示肿

瘤纯度,C表示拷贝数,n表示正常状况条件,M表示突变等位基因数量。M值的给定为基于将期望等位基因频率值和观察等位基因频率值之间的差异以卡方检定(χ2test)进行统计显著性检定,并给定检定结果中和观察等位基因频率值最接近的期望等位基因频率值的情况下所对应的M值,其中期望等位基因频率值如表4中所揭示(由Sun等人的研究(Sun et al., CancerRes(2014)74(19S):1893)中修改而来)。

[0122] 表4:在不同拷贝数、异质性丧失及肿瘤纯度下突变等位基因数量及期望等位基因频率值之数值。

[0123]

拷贝数*	LOH状态	M	肿瘤纯度**										
			0	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1

[0124]

1	LOH	1	0	0.05	0.11	0.18	0.25	0.33	0.43	0.54	0.67	0.82	1
2	LOH	2	0	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1
2	het	1	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
3	LOH	3	0	0.14	0.27	0.39	0.50	0.60	0.69	0.78	0.86	0.93	1
3	het	1	0	0.05	0.09	0.13	0.17	0.20	0.23	0.26	0.29	0.31	0.33
3	het	2	0	0.10	0.18	0.26	0.33	0.40	0.46	0.52	0.57	0.62	0.67
4	LOH	4	0	0.18	0.33	0.46	0.57	0.67	0.75	0.82	0.89	0.95	1
4	Het	3	0	0.14	0.25	0.35	0.43	0.50	0.56	0.62	0.67	0.71	0.75
4	Het	1	0	0.05	0.08	0.12	0.14	0.17	0.19	0.21	0.22	0.24	0.25
4	Het	2	0	0.09	0.17	0.23	0.29	0.33	0.38	0.41	0.44	0.47	0.50
...													

[0125] *虽然表4只揭示了1～4的拷贝数,但任意拷贝数皆可做计算

[0126] **虽然表4只以间距0.1揭示了肿瘤纯度,但任意肿瘤纯度皆可做计算

[0127] 使用给定的M值,亚群落纯度s之计算方式如下所揭示:

[0128] $$s = \frac{AF_{observed}}{M}\left(pC_t + C_n(1-p)\right)$$

[0129] 我们以亚群落纯度除以肿瘤纯度来计算样本级数值,样本级数值即代表突变等位基因在肿瘤组织中实际的量,样本级数值之计算方式如下所揭示:

[0130] sample-level score＝s/p

[0131] 需要注意的是当观察等位基因频率值大于期望等位基因频率值时,我们假设此变异为一个群落突变(clonal mutation)而不进行卡方统计检定,此外在此情况下,亚群落纯度会大于肿瘤纯度,因而直接给定样本级数值为1。一个样本的样本级数值资料如表5中所揭示。

[0132] 表5:一个样本的样本级数值资料

[0133]

变异	变异频率	肿瘤纯度	LOH	拷贝数	群落种类	p 值	期望频率值	亚群落纯度	样本级数值
ANK1 chr8_41552157_T>C	0.723	66.8	LOH	2	clonal	obs>exp	0.668	0.723	1
CHD5 chr1_6194848_G>A	0.684	66.8	LOH	2	clonal	obs>exp	0.668	0.684	1
SLC3A2 chr11_62652688_G>T	0.408	66.8	HET	2	clonal	obs>exp	0.334	0.816	1
MRC1 chr10_18122699_G>A	0.407	66.8	HET	2	clonal	obs>exp	0.334	0.814	1
TGS1 chr8_56698861_A>T	0.392	66.8	HET	2	clonal	obs>exp	0.334	0.784	1
SPACA1 chr6_88769199_C>A	0.391	66.8	HET	2	clonal	obs>exp	0.334	0.782	1
ARHGAP21 chr10_24896718_G>T	0.38	66.8	HET	2	clonal	obs>exp	0.334	0.760	1
NSRP1 chr17_28511815_G>T	0.375	66.8	HET	2	clonal	obs>exp	0.334	0.750	1
SCN3A chr2_165984466_C>T	0.373	66.8	HET	2	clonal	obs>exp	0.334	0.746	1
TET3 chr2_74328527_G>A	0.373	66.8	HET	2	clonal	obs>exp	0.334	0.746	1
ZDHHC1 chr16_67440152_C>T	0.366	66.8	HET	2	clonal	obs>exp	0.334	0.732	1
G0S2 chr1_209849154_T>G	0.363	66.8	HET	2	clonal	obs>exp	0.334	0.726	1

[0134]

FANCM										
chr14_45645307_A>T	0.354	66.8	HET	2	clonal	obs>exp	0.334	0.708	1
NFASC										
chr1_204957844_G>A	0.326	66.8	HET	2	clonal	0.637	0.334	0.652	0.977
SON										
chr21_34921835_A>G	0.636	66.8	LOH	2	clonal	0.585	0.668	0.636	0.953
SSTR4										
chr20_23017227_A>C	0.636	66.8	LOH	2	clonal	0.209	0.668	0.636	0.953
TMEM104										
chr17_72832365_C>T	0.317	66.8	HET	2	clonal	0.555	0.334	0.634	0.950
ZC3H3										
chr8_144557605_C>A	0.315	66.8	HET	2	clonal	0.432	0.334	0.630	0.944
KDELC2										
chr11_108361810_T>C	0.315	66.8	HET	2	clonal	0.765	0.334	0.630	0.944
OSBPL5										
chr11_3125480_G>A	0.621	66.8	LOH	2	clonal	0.197	0.668	0.621	0.930
TP53										
chr17_7578445_A>T	0.584	66.8	LOH	2	subclonal	0.028	0.668	0.584	0.875
SP5										
chr2_171572922_C>G	0.291	66.8	HET	2	clonal	0.262	0.334	0.582	0.872
COL6A5										
chr3_130150786_C>A	0.57	66.8	LOH	2	subclonal	0.006	0.668	0.570	0.854
LRRC8E										
chr19_7964551_G>A	0.564	66.8	LOH	2	subclonal	0.000	0.668	0.564	0.845
NUP160										
chr11_47809795_T>C	0.274	66.8	HET	2	clonal	0.213	0.334	0.548	0.821
ANO2										
chr12_5685076_A>C	0.41	66.8	HET	3	subclonal	0.010	0.5	0.547	0.819
MUC16										
chr19_9086656_C>T	0.539	66.8	LOH	2	subclonal	0	0.668	0.539	0.807
ARFGEF1										
chr8_68208826_G>T	0.265	66.8	HET	2	clonal	0.207	0.334	0.530	0.794
APC2										
chr19_1456374_GA>TT	0.524	66.8	LOH	2	subclonal	0.000	0.668	0.524	0.785
CSE1L										
chr20_47675019_A>T	0.611	66.8	LOH	4	subclonal	0	0.801	0.509	0.763
KCNA5										
chr12_5154521_G>A	0.382	66.8	HET	3	subclonal	0.003	0.5	0.510	0.763
TRIM34										
chr11_5664563_G>T	0.454	66.8	LOH	2	subclonal	0	0.668	0.454	0.680
CCDC141										
chr2_179730534_C>T	0.399	66.8	LOH	2	subclonal	0	0.668	0.399	0.598
KIAA1429										
chr8_95500959_C>A	0.395	66.8	LOH	2	subclonal	0	0.668	0.395	0.592
RNF17										
chr13_25338358_C>T	0.39	66.8	LOH	2	subclonal	0	0.668	0.390	0.584
GBF1										
chr10_104126933_A>G	0.379	66.8	LOH	2	subclonal	0	0.668	0.379	0.568
PRDX6	0.367	66.8	LOH	2	subclonal	0	0.668	0.367	0.550	

	chr1_173454623_G>A GPR115									
	chr6_47681616_C>T DCHS1	0.359	66.8	LOH	2	subclonal	0	0.668	0.359	0.538
	chr11_6644247_C>A VPS13A	0.354	66.8	LOH	2	subclonal	0	0.668	0.354	0.530
	chr9_79865083_A>T COL12A1	0.35	66.8	LOH	2	subclonal	0	0.668	0.350	0.524
	chr6_75851835_G>T EPN3	0.345	66.8	LOH	2	subclonal	0	0.668	0.345	0.517
	chr17_48616588_A>G SF3B1	0.341	66.8	LOH	2	subclonal	0	0.668	0.341	0.511
	chr2_198272747_T>C FAM65B	0.33	66.8	LOH	2	subclonal	0	0.668	0.330	0.494
	chr6_24861205_C>A NF1	0.326	66.8	LOH	2	subclonal	0	0.668	0.326	0.488
	chr17_29509534_G>A C4BPA	0.323	66.8	LOH	2	subclonal	0	0.668	0.323	0.484
	chr1_207317200_G>T DST	0.317	66.8	LOH	2	subclonal	0	0.668	0.317	0.475
	chr6_56494089_T>A NALCN	0.303	66.8	LOH	2	subclonal	0	0.668	0.303	0.454
	chr13_101944438_A>C NEK2	0.284	66.8	LOH	2	subclonal	0	0.668	0.284	0.425
[0135]	chr1_211836839_G>A HES1	0.261	66.8	LOH	2	subclonal	0	0.668	0.261	0.391
	chr3_193854752_C>G DOPEY1	0.293	66.8	LOH	3	subclonal	0	0.751	0.261	0.390
	chr6_83839042_C>T TATDN3	0.245	66.8	LOH	2	subclonal	0	0.668	0.245	0.367
	chr1_212981098_C>T CCDC34	0.111	66.8	HET	2	subclonal	0.001	0.334	0.222	0.333
	chr11_27360421_G>A SNX19	0.148	66.8	LOH	2	subclonal	0	0.668	0.148	0.222
	chr11_130785825_C>G CPS1	0.145	66.8	LOH	2	subclonal	0	0.668	0.145	0.217
	chr2_211441138_G>A PLOD1	0.12	66.8	LOH	2	subclonal	0	0.668	0.120	0.180
	chr1_12025537_G>A RBBP6	0.118	66.8	LOH	2	subclonal	0	0.668	0.118	0.177
	chr16_24579145_C>T DUT	0.117	66.8	LOH	2	subclonal	0	0.668	0.117	0.175
	chr15_48634228_G>A LAMC1	0.114	66.8	LOH	2	subclonal	0	0.668	0.114	0.171
	chr1_183103819_G>A PELP1	0.109	66.8	LOH	2	subclonal	0	0.668	0.109	0.163
	chr17_4575375_C>T HAUS2	0.109	66.8	LOH	2	subclonal	0	0.668	0.109	0.163
	chr15_42858894_A>C	0.103	66.8	LOH	2	subclonal	0	0.668	0.103	0.154

[0136]

RFC1 chr4_39322052_C>T	0.103	66.8	LOH	2	subclonal	0	0.668	0.103	0.154
TSSK2 chr22_19119814_C>T	0.102	66.8	LOH	2	subclonal	0	0.668	0.102	0.153
KBTBD12 chr3_127642629_G>A	0.1	66.8	LOH	2	subclonal	0	0.668	0.100	0.150
DCTN2 chr12_57939846_C>T	0.109	66.8	LOH	3	subclonal	0	0.751	0.097	0.145

[0137] 实施例16:以胜肽链级数值和样本级数值从癌症样本中将胜肽链作为免疫原进行判断、评分与排名。

[0138] 为了将一个癌症样本的胜肽链作为免疫原进行判断、评分与排名,我们从癌症样本中获取胜肽链资料并将我们的评分方法应用在胜肽链资料上,样本的外显子组定序流程如实施例15所揭示。在确认了样本中的体细胞突变后,我们获取了胺基酸长度为8～23且带有突变胺基酸位点的胜肽链,接着将胜肽链以实施例12中所揭示的方法计算胜肽链级数值以及以实施例15中所揭示的方法计算样本级数值。为了整合胜肽链以及样本相关的信息来将胜肽链作为免疫原进行排名,我们将每个胜肽链的胜肽链级数值和样本级数值相乘来计算其致免疫性数值(immunogenic score)。在一个胃癌的样本中,所选致免疫性数值前50高的胜肽链资料如表6中所揭示。

[0139] 表6:在一个胃癌的样本中致免疫性数值前50高的胜肽链资料。

[0140]

排名	变异 ID	序列	胜肽链级数值	样本级数值	致免疫性数值
1	chr16_67440152_C>T	LYLFFAVIGFEILVPLLPHHW	0.393	1.000	0.393
2	chr16_67440152_C>T	LYLFFAVIGFEILVPLL	0.390	1.000	0.390
3	chr16_67440152_C>T	HPLQIVAWLLYLFFAVIGFEILV	0.381	1.000	0.381
4	chr17_28511815_G>T	RVNFRREKVIETPENDFKHHR	0.379	1.000	0.379
5	chr16_67440152_C>T	LYLFFAVIGFEI	0.379	1.000	0.379
6	chr10_18122699_G>A	STLTWHQAR	0.369	1.000	0.369
7	chr16_67440152_C>T	LYLFFAVIGFEIL	0.364	1.000	0.364
8	chr16_67440152_C>T	LYLFFAVIGFEILVPLLPHHWV	0.363	1.000	0.363
9	chr2_165984466_C>T	KMQECFQKAFFR	0.359	1.000	0.359
10	chr16_67440152_C>T	LYLFFAVIGFEILV	0.358	1.000	0.358
11	chr17_28511815_G>T	RVNFRREKVIETPENDFK	0.352	1.000	0.352
12	chr16_67440152_C>T	AWLLYLFFAVIGFEILVPLL	0.350	1.000	0.350
13	chr16_67440152_C>T	FFAVIGFEILVPLLPHHW	0.350	1.000	0.350
14	chr17_28511815_G>T	RVNFRREK	0.349	1.000	0.349
15	chr17_72832365_C>T	MYTLNFACCDVVGLAAVRFFL	0.366	0.950	0.348
16	chr10_18122699_G>A	TLTWHQAR	0.345	1.000	0.345
17	chr8_41552157_T>C	KLVPLVQAAFPENAVTKR	0.344	1.000	0.344

[0141]

18	chr10_18122699_G>A	KSTLTWHQAR	0.343	1.000	0.343
19	chr16_67440152_C>T	AWLLYLFFAVIGFEI	0.343	1.000	0.343
20	chr8_41552157_T>C	KSKLVPLVQAAFPENAVTKR	0.340	1.000	0.340
21	chr1_209849154_T>G	ALFGVVLGRMETVCSPFTAAR	0.338	1.000	0.338
22	chr1_209849154_T>G	VVLGRMETVCSPFTAAR	0.337	1.000	0.337
23	chr1_209849154_T>G	KLYVLGSVLALFGVVLGR	0.337	1.000	0.337
24	chr10_18122699_G>A	STLTWHQARK	0.336	1.000	0.336
25	chr1_209849154_T>G	VVLGRMETVCSPFTAARRLR	0.336	1.000	0.336
26	chr10_24896718_G>T	KTSAPLIRR	0.336	1.000	0.336
27	chr17_72832365_C>T	MYTLNFACCDVVGLAAVRFF	0.351	0.950	0.333
28	chr8_41552157_T>C	VPLVQAAFPENAV	0.333	1.000	0.333
29	chr2_165984466_C>T	KMQECFQKAFFRK	0.331	1.000	0.331
30	chr17_72832365_C>T	MYTLNFACCDVVGLAAVRF	0.348	0.950	0.331
31	chr1_209849154_T>G	ALFGVVLGR	0.330	1.000	0.330
32	chr2_74328527_G>A	SLTEKPWALR	0.329	1.000	0.329
33	chr17_28511815_G>T	KSSADDEIEETRVNFR	0.329	1.000	0.329
34	chr8_56698861_A>T	KYLVEIVQESW	0.329	1.000	0.329
35	chr1_209849154_T>G	KMVKLYVLGSVLALFGVVLGR	0.328	1.000	0.328
36	chr1_209849154_T>G	ALFGVVLGRMETVCSPFTAARR	0.328	1.000	0.328
37	chr17_28511815_G>T	DEIEETRVNFRREKVIETPEND	0.327	1.000	0.327
38	chr17_72832365_C>T	MYTLNFACCDVVGLAAVRFFLGL	0.344	0.950	0.327
39	chr2_74328527_G>A	STSALAGPSLTEKPWALR	0.326	1.000	0.326
40	chr6_88769199_C>A	KYMWKLLRQDQQSIILVNDSEI	0.326	1.000	0.326
41	chr17_28511815_G>T	DDEIEETRVNFRREKVIETPEND	0.326	1.000	0.326
42	chr1_209849154_T>G	RMETVCSPFTAAR	0.326	1.000	0.326
43	chr2_165984466_C>T	KMQECFQKAFFRKPK	0.324	1.000	0.324
44	chr10_24896718_G>T	KTSAPLIR	0.324	1.000	0.324
45	chr1_209849154_T>G	MVKLYVLGSVLALFGVVLGR	0.323	1.000	0.323

[0142]

46	chr1_209849154_T>G	VLALFGVVLGRMETVCSPFTAAR	0.323	1.000	0.323
47	chr17_28511815_G>T	VNFRREKVIETPEND	0.322	1.000	0.322
48	chr17_28511815_G>T	ADDEIEETRVNFRREKVIETPEN	0.320	1.000	0.320
49	chr2_74328527_G>A	ALAGPSLTEKPWALR	0.320	1.000	0.320
50	chr1_209849154_T>G	SVLALFGVVLGR	0.320	1.000	0.320

[0001] 序列表

[0002] <110> 行动基因(智财)有限公司(ACT Genomics (IP) Co., Ltd.)

[0003] <120> 致免疫性的癌症特异抗原决定位的排名系统

[0004] <130> ACTG-1PCT

[0005] <140> PCT/US2018/025597

[0006] <141> 2018-03-31

[0007] <150> 62/479,320

[0008] <151> 2017-03-31

[0009] <160> 100

[0010] <170> SIPOSequenceListing 1.0

[0011] <210> 1

[0012] <211> 9

[0013] <212> PRT

[0014] <213> Artificial sequence

[0015] <400> 1

[0016] Lys Leu Lys Phe Val Thr Leu Val Phe

[0017] 1 5

[0018] <210> 2

[0019] <211> 10

[0020] <212> PRT

[0021] <213> Artificial sequence

[0022] <400> 2

[0023] Arg Phe Leu Glu Tyr Leu Pro Leu Arg Phe

[0024] 1 5 10

[0025] <210> 3

[0026] <211> 10

[0027] <212> PRT

[0028] <213> Artificial sequence

[0029] <400> 3

[0030] Val Gln Lys Val Ala Ser Lys Ile Pro Phe

[0031] 1 5 10

[0032] <210> 4

[0033] <211> 9

[0034] <212> PRT

[0035] <213> Artificial sequence

[0036] <400> 4

[0037] Thr Leu Phe His Thr Phe Tyr Glu Leu

[0038] 1 5

[0039] <210> 5

[0040] <211> 10

[0041] <212> PRT

[0042] <213> Artificial sequence

[0043] <400> 5

[0044] Thr Leu Phe His Thr Phe Tyr Glu Leu Leu

[0045] 1 5 10

[0046] <210> 6

[0047] <211> 9

[0048] <212> PRT

[0049] <213> Artificial sequence

[0050] <400> 6

[0051] Lys Phe Gly Asp Leu Thr Asn Asn Phe

[0052] 1 5

[0053] <210> 7

[0054] <211> 9

[0055] <212> PRT

[0056] <213> Artificial sequence

[0057] <400> 7

[0058] Pro Arg Glu Glu Phe Leu Arg Leu Cys

[0059] 1 5

[0060] <210> 8

[0061] <211> 9

[0062] <212> PRT

[0063] <213> Artificial sequence

[0064] <400> 8

[0065] Lys Leu Phe Glu Ser Lys Ala Glu Leu

[0066] 1 5

[0067] <210> 9

[0068] <211> 9

[0069] <212> PRT

[0070] <213> Artificial sequence

[0071] <400> 9

[0072] Leu Cys Pro Arg Glu Glu Phe Leu Arg

[0073] 1 5

[0074] <210> 10

[0075] <211> 10

[0076] <212> PRT

[0077] <213> Artificial sequence

[0078] <400> 10

[0079] Pro Phe Pro Asp Arg Ile Thr Glu Glu Ser

[0080] 1 5 10

[0081] <210> 11

[0082] <211> 9

[0083] <212> PRT

[0084] <213> Artificial sequence

[0085] <400> 11

[0086] Val Leu Ala Lys Lys Leu Lys Phe Val

[0087] 1 5

[0088] <210> 12

[0089] <211> 9

[0090] <212> PRT

[0091] <213> Artificial sequence

[0092] <400> 12

[0093] Lys Lys Lys Trp Phe Leu Phe Gln Asp

[0094] 1 5

[0095] <210> 13

[0096] <211> 9

[0097] <212> PRT

[0098] <213> Artificial sequence

[0099] <400> 13

[0100] Pro Phe Pro Asp Arg Ile Thr Glu Glu

[0101] 1 5

[0102] <210> 14

[0103] <211> 9

[0104] <212> PRT

[0105] <213> Artificial sequence

[0106] <400> 14

[0107] His Thr Glu Leu Glu Arg Phe Leu Glu

[0108] 1 5

[0109] <210> 15

[0110] <211> 10

[0111] <212> PRT

[0112] <213> Artificial sequence

[0113] <400> 15

[0114] Lys Leu Phe Glu Ser Lys Ala Glu Leu Ala

[0115] 1 5 10

[0116] <210> 16

[0117] <211> 9

[0118] <212> PRT

[0119] <213> Artificial sequence

[0120] <400> 16

[0121] Thr Glu Leu Glu Arg Phe Leu Glu Tyr

[0122] 1 5

[0123] <210> 17

[0124] <211> 9

[0125] <212> PRT

[0126] <213> Artificial sequence

[0127] <400> 17

[0128] Leu Leu His Thr Glu Leu Glu Arg Phe

[0129] 1 5

[0130] <210> 18

[0131] <211> 9

[0132] <212> PRT

[0133] <213> Artificial sequence

[0134] <400> 18

[0135] Phe Pro Asp Arg Ile Thr Glu Glu Ser

[0136] 1 5

[0137] <210> 19

[0138] <211> 10

[0139] <212> PRT

[0140] <213> Artificial sequence

[0141] <400> 19

[0142] Val Ser Val Gly Asp Phe Ser Gln Glu Phe

[0143] 1 5 10

[0144] <210> 20

[0145] <211> 10

[0146] <212> PRT

[0147] <213> Artificial sequence

[0148] <400> 20

[0149] Ile Pro Phe Pro Asp Arg Ile Thr Glu Glu

[0150] 1 5 10

[0151] <210> 21

[0152] <211> 10

[0153] <212> PRT

[0154] <213> Artificial sequence

[0155] <400> 21

[0156] Cys Pro Arg Glu Glu Phe Leu Arg Leu Cys

[0157] 1 5 10

[0158] <210> 22

[0159] <211> 9

[0160] <212> PRT

[0161] <213> Artificial sequence

[0162] <400> 22

[0163] Ala Leu Phe Ala Ser Arg Pro Arg Phe

[0164] 1 5

[0165] <210> 23

[0166] <211> 9

[0167] <212> PRT

[0168] <213> Artificial sequence

[0169] <400> 23

[0170] Phe Leu Phe Gln Asp Ser Lys Lys Ile

[0171] 1 5

[0172] <210> 24

[0173] <211> 9

[0174] <212> PRT

[0175] <213> Artificial sequence

[0176] <400> 24

[0177] Asp Lys Leu Phe Glu Ser Lys Ala Glu

[0178] 1 5

[0179] <210> 25

[0180] <211> 10

[0181] <212> PRT

[0182] <213> Artificial sequence

[0183] <400> 25

[0184] Ser Lys Lys Lys Trp Phe Leu Phe Gln Asp

[0185] 1 5 10

[0186] <210> 26

[0187] <211> 9

[0188] <212> PRT

[0189] <213> Artificial sequence

[0190] <400> 26

[0191] Ile Pro Phe Pro Asp Arg Ile Thr Glu

[0192] 1 5

[0193] <210> 27

[0194] <211> 9

[0195] <212> PRT

[0196] <213> Artificial sequence

[0197] <400> 27

[0198] Gly Gly Ala Leu Phe Ala Ser Arg Pro

[0199] 1 5

[0200] <210> 28

[0201] <211> 10

[0202] <212> PRT

[0203] <213> Artificial sequence

[0204] <400> 28

[0205] Leu His Thr Glu Leu Glu Arg Phe Leu Glu

[0206] 1 5 10

[0207] <210> 29

[0208] <211> 9

[0209] <212> PRT

[0210] <213> Artificial sequence

[0211] <400> 29

[0212] Leu Ser Pro Arg Glu Glu Phe Leu Arg

[0213] 1 5

[0214] <210> 30

[0215] <211> 10

[0216] <212> PRT

[0217] <213> Artificial sequence

[0218] <400> 30

[0219] Ser Pro Arg Glu Glu Phe Leu Arg Leu Cys

[0220] 1 5 10

[0221] <210> 31

[0222] <211> 9

[0223] <212> PRT

[0224] <213> Artificial sequence

[0225] <400> 31

[0226] Val Gly Asp Phe Ser Gln Glu Phe Ser

[0227] 1 5

[0228] <210> 32

[0229] <211> 10

[0230] <212> PRT

[0231] <213> Artificial sequence

[0232] <400> 32

[0233] Lys Lys Lys Trp Phe Leu Phe Gln Asp Ser

[0234] 1 5 10

[0235] <210> 33

[0236] <211> 10

[0237] <212> PRT

[0238] <213> Artificial sequence

[0239] <400> 33

[0240] Asp Ser Gly Ile Pro Glu Asn Ser Phe Asn

[0241] 1 5 10

[0242] <210> 34

[0243] <211> 10

[0244] <212> PRT

[0245] <213> Artificial sequence

[0246] <400> 34

[0247] Val Gly Asp Phe Ser Gln Glu Phe Ser Pro

[0248] 1 5 10

[0249] <210> 35

[0250] <211> 10

[0251] <212> PRT

[0252] <213> Artificial sequence

[0253] <400> 35

[0254] Arg Gly Gly Ala Leu Phe Ala Ser Arg Pro

[0255] 1 5 10

[0256] <210> 36

[0257] <211> 9

[0258] <212> PRT

[0259] <213> Artificial sequence

[0260] <400> 36

[0261] Ser Val Gly Asp Phe Ser Gln Glu Phe

[0262] 1 5

[0263] <210> 37

[0264] <211> 9

[0265] <212> PRT

[0266] <213> Artificial sequence

[0267] <400> 37

[0268] Leu Ala Asp Ser Gly Ile Pro Glu Asn

[0269] 1 5

[0270] <210> 38

[0271] <211> 9

[0272] <212> PRT

[0273] <213> Artificial sequence

[0274] <400> 38

[0275] Lys Lys Leu Lys Phe Val Thr Leu Val

[0276] 1 5

[0277] <210> 39

[0278] <211> 10

[0279] <212> PRT

[0280] <213> Artificial sequence

[0281] <400> 39

[0282] Ser Lys Ile Pro Phe Pro Asp Arg Ile Thr

[0283] 1 5 10

[0284] <210> 40

[0285] <211> 9

[0286] <212> PRT

[0287] <213> Artificial sequence

[0288] <400> 40

[0289] Gly Ile Pro Glu Asn Ser Phe Asn Val

[0290] 1 5

[0291] <210> 41

[0292] <211> 9

[0293] <212> PRT

[0294] <213> Artificial sequence

[0295] <400> 41

[0296] Lys Ile Pro Phe Pro Asp Arg Ile Thr

[0297] 1 5

[0298] <210> 42

[0299] <211> 9

[0300] <212> PRT

[0301] <213> Artificial sequence

[0302] <400> 42

[0303] Gly Lys Phe Gly Asp Leu Thr Asn Asn

[0304] 1 5

[0305] <210> 43

[0306] <211> 9

[0307] <212> PRT

[0308] <213> Artificial sequence

[0309] <400> 43

[0310] Asp Ser Asp Lys Leu Phe Glu Ser Lys

[0311] 1 5

[0312] <210> 44

[0313] <211> 9

[0314] <212> PRT

[0315] <213> Artificial sequence

[0316] <400> 44

[0317] Glu Asp Ser Asp Lys Leu Phe Glu Ser

[0318] 1 5

[0319] <210> 45

[0320] <211> 10

[0321] <212> PRT

[0322] <213> Artificial sequence

[0323] <400> 45

[0324] Leu Leu His Thr Glu Leu Glu Arg Phe Leu

[0325] 1 5 10

[0326] <210> 46

[0327] <211> 10

[0328] <212> PRT

[0329] <213> Artificial sequence

[0330] <400> 46

[0331] Arg Arg Gly Gly Ala Leu Phe Ala Ser Arg

[0332] 1 5 10

[0333] <210> 47

[0334] <211> 9

[0335] <212> PRT

[0336] <213> Artificial sequence

[0337] <400> 47

[0338] Ile Pro Glu Asn Ser Phe Asn Val Ser

[0339] 1 5

[0340] <210> 48

[0341] <211> 9

[0342] <212> PRT

[0343] <213> Artificial sequence

[0344] <400> 48

[0345] Ala Ser Lys Ile Pro Phe Pro Asp Arg

[0346] 1 5

[0347] <210> 49

[0348] <211> 10

[0349] <212> PRT

[0350] <213> Artificial sequence

[0351] <400> 49

[0352] Leu Phe His Thr Phe Tyr Glu Leu Leu Ile

[0353] 1 5 10

[0354] <210> 50

[0355] <211> 10

[0356] <212> PRT

[0357] <213> Artificial sequence

[0358] <400> 50

[0359] Ile Leu Leu His Thr Glu Leu Glu Arg Phe

[0360] 1 5 10

[0361] <210> 51

[0362] <211> 21

[0363] <212> PRT

[0364] <213> Artificial sequence

[0365] <400> 51

[0366] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu Val Pro Leu

[0367] 1 5 10 15

[0368] Leu Pro His His Trp

[0369] 20

[0370] <210> 52

[0371] <211> 17

[0372] <212> PRT

[0373] <213> Artificial sequence

[0374] <400> 52

[0375] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu Val Pro Leu

[0376] 1 5 10 15

[0377] Leu

[0378] <210> 53

[0379] <211> 23

[0380] <212> PRT

[0381] <213> Artificial sequence

[0382] <400> 53

[0383] His Pro Leu Gln Ile Val Ala Trp Leu Leu Tyr Leu Phe Phe Ala Val

[0384] 1 5 10 15

[0385] Ile Gly Phe Glu Ile Leu Val

[0386] 20

[0387] <210> 54

[0388] <211> 21

[0389] <212> PRT

[0390] <213> Artificial sequence

[0391] <400> 54

[0392] Arg Val Asn Phe Arg Arg Glu Lys Val Ile Glu Thr Pro Glu Asn Asp

[0393] 1 5 10 15

[0394] Phe Lys His His Arg

[0395] 20

[0396] <210> 55

[0397] <211> 12

[0398] <212> PRT

[0399] <213> Artificial sequence

[0400] <400> 55

[0401] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile

[0402] 1 5 10

[0403] <210> 56

[0404] <211> 9

[0405] <212> PRT

[0406] <213> Artificial sequence

[0407] <400> 56

[0408] Ser Thr Leu Thr Trp His Gln Ala Arg

[0409] 1 5

[0410] <210> 57

[0411] <211> 13

[0412] <212> PRT

[0413] <213> Artificial sequence

[0414] <400> 57

[0415] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu

[0416] 1 5 10

[0417] <210> 58

[0418] <211> 22

[0419] <212> PRT

[0420] <213> Artificial sequence

[0421] <400> 58

[0422] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu Val Pro Leu

[0423] 1 5 10 15

[0424] Leu Pro His His Trp Val

[0425] 20

[0426] <210> 59

[0427] <211> 12

[0428] <212> PRT

[0429] <213> Artificial sequence

[0430] <400> 59

[0431] Lys Met Gln Glu Cys Phe Gln Lys Ala Phe Phe Arg

[0432] 1 5 10

[0433] <210> 60

[0434] <211> 14

[0435] <212> PRT

[0436] <213> Artificial sequence

[0437] <400> 60

[0438] Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu Val

[0439] 1 5 10

[0440] <210> 61

[0441] <211> 18

[0442] <212> PRT

[0443] <213> Artificial sequence

[0444] <400> 61

[0445] Arg Val Asn Phe Arg Arg Glu Lys Val Ile Glu Thr Pro Glu Asn Asp

[0446] 1 5 10 15

[0447] Phe Lys

[0448] <210> 62

[0449] <211> 20

[0450] <212> PRT

[0451] <213> Artificial sequence

[0452] <400> 62

[0453] Ala Trp Leu Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile Leu

[0454] 1 5 10 15

[0455] Val Pro Leu Leu

[0456] 20

[0457] <210> 63

[0458] <211> 18

[0459] <212> PRT

[0460] <213> Artificial sequence

[0461] <400> 63

[0462] Phe Phe Ala Val Ile Gly Phe Glu Ile Leu Val Pro Leu Leu Pro His

[0463] 1 5 10 15

[0464] His Trp

[0465] <210> 64

[0466] <211> 8

[0467] <212> PRT

[0468] <213> Artificial sequence

[0469] <400> 64

[0470] Arg Val Asn Phe Arg Arg Glu Lys

[0471] 1 5

[0472] <210> 65

[0473] <211> 21

[0474] <212> PRT

[0475] <213> Artificial sequence

[0476] <400> 65

[0477] Met Tyr Thr Leu Asn Phe Ala Cys Cys Asp Val Val Gly Leu Ala Ala

[0478] 1 5 10 15

[0479] Val Arg Phe Phe Leu

[0480] 20

[0481] <210> 66

[0482] <211> 8

[0483] <212> PRT

[0484] <213> Artificial sequence

[0485] <400> 66

[0486] Thr Leu Thr Trp His Gln Ala Arg

[0487] 1 5

[0488] <210> 67

[0489] <211> 18

[0490] <212> PRT

[0491] <213> Artificial sequence

[0492] <400> 67

[0493] Lys Leu Val Pro Leu Val Gln Ala Ala Phe Pro Glu Asn Ala Val Thr

[0494] 1 5 10 15

[0495] Lys Arg

[0496] <210> 68

[0497] <211> 10

[0498] <212> PRT

[0499] <213> Artificial sequence

[0500] <400> 68

[0501] Lys Ser Thr Leu Thr Trp His Gln Ala Arg

[0502] 1 5 10

[0503] <210> 69

[0504] <211> 15

[0505] <212> PRT

[0506] <213> Artificial sequence

[0507] <400> 69

[0508] Ala Trp Leu Leu Tyr Leu Phe Phe Ala Val Ile Gly Phe Glu Ile

[0509] 1 5 10 15

[0510] <210> 70

[0511] <211> 20

[0512] <212> PRT

[0513] <213> Artificial sequence

[0514] <400> 70

[0515] Lys Ser Lys Leu Val Pro Leu Val Gln Ala Ala Phe Pro Glu Asn Ala

[0516] 1 5 10 15

[0517] Val Thr Lys Arg

[0518] 20

[0519] <210> 71

[0520] <211> 21

[0521] <212> PRT

[0522] <213> Artificial sequence

[0523] <400> 71

[0524] Ala Leu Phe Gly Val Val Leu Gly Arg Met Glu Thr Val Cys Ser Pro

[0525] 1 5 10 15

[0526] Phe Thr Ala Ala Arg

[0527] 20

[0528] <210> 72

[0529] <211> 17

[0530] <212> PRT

[0531] <213> Artificial sequence

[0532] <400> 72

[0533] Val Val Leu Gly Arg Met Glu Thr Val Cys Ser Pro Phe Thr Ala Ala

[0534] 1 5 10 15

[0535] Arg

[0536] <210> 73

[0537] <211> 18

[0538] <212> PRT

[0539] <213> Artificial sequence

[0540] <400> 73

[0541] Lys Leu Tyr Val Leu Gly Ser Val Leu Ala Leu Phe Gly Val Val Leu

[0542] 1 5 10 15

[0543] Gly Arg

[0544] <210> 74

[0545] <211> 10

[0546] <212> PRT

[0547] <213> Artificial sequence

[0548] <400> 74

[0549] Ser Thr Leu Thr Trp His Gln Ala Arg Lys

[0550] 1 5 10

[0551] <210> 75

[0552] <211> 20

[0553] <212> PRT

[0554] <213> Artificial sequence

[0555] <400> 75

[0556] Val Val Leu Gly Arg Met Glu Thr Val Cys Ser Pro Phe Thr Ala Ala

[0557] 1 5 10 15

[0558] Arg Arg Leu Arg

[0559] 20

[0560] <210> 76

[0561] <211> 9

[0562] <212> PRT

[0563] <213> Artificial sequence

[0564] <400> 76

[0565] Lys Thr Ser Ala Pro Leu Ile Arg Arg

[0566] 1 5

[0567] <210> 77

[0568] <211> 20

[0569] <212> PRT

[0570] <213> Artificial sequence

[0571] <400> 77

[0572] Met Tyr Thr Leu Asn Phe Ala Cys Cys Asp Val Val Gly Leu Ala Ala

[0573] 1 5 10 15

[0574] Val Arg Phe Phe

[0575] 20

[0576] <210> 78

[0577] <211> 13

[0578] <212> PRT

[0579] <213> Artificial sequence

[0580] <400> 78

[0581] Val Pro Leu Val Gln Ala Ala Phe Pro Glu Asn Ala Val

[0582] 1 5 10

[0583] <210> 79

[0584] <211> 13

[0585] <212> PRT

[0586] <213> Artificial sequence

[0587] <400> 79

[0588] Lys Met Gln Glu Cys Phe Gln Lys Ala Phe Phe Arg Lys

[0589] 1 5 10

[0590] <210> 80

[0591] <211> 19

[0592] <212> PRT

[0593] <213> Artificial sequence

[0594] <400> 80

[0595] Met Tyr Thr Leu Asn Phe Ala Cys Cys Asp Val Val Gly Leu Ala Ala

[0596] 1 5 10 15

[0597] Val Arg Phe

[0598] <210> 81

[0599] <211> 9

[0600] <212> PRT

[0601] <213> Artificial sequence

[0602] <400> 81

[0603] Ala Leu Phe Gly Val Val Leu Gly Arg

[0604] 1 5

[0605] <210> 82

[0606] <211> 10

[0607] <212> PRT

[0608] <213> Artificial sequence

[0609] <400> 82

[0610] Ser Leu Thr Glu Lys Pro Trp Ala Leu Arg

[0611] 1 5 10

[0612] <210> 83

[0613] <211> 16

[0614] <212> PRT

[0615] <213> Artificial sequence

[0616] <400> 83

[0617] Lys Ser Ser Ala Asp Asp Glu Ile Glu Glu Thr Arg Val Asn Phe Arg

[0618] 1 5 10 15

[0619] <210> 84

[0620] <211> 11

[0621] <212> PRT

[0622] <213> Artificial sequence

[0623] <400> 84

[0624] Lys Tyr Leu Val Glu Ile Val Gln Glu Ser Trp

[0625] 1 5 10

[0626] <210> 85

[0627] <211> 21

[0628] <212> PRT

[0629] <213> Artificial sequence

[0630] <400> 85

[0631] Lys Met Val Lys Leu Tyr Val Leu Gly Ser Val Leu Ala Leu Phe Gly

[0632] 1 5 10 15

[0633] Val Val Leu Gly Arg

[0634] 20

[0635] <210> 86

[0636] <211> 22

[0637] <212> PRT

[0638] <213> Artificial sequence

[0639] <400> 86

[0640] Ala Leu Phe Gly Val Val Leu Gly Arg Met Glu Thr Val Cys Ser Pro

[0641] 1 5 10 15

[0642] Phe Thr Ala Ala Arg Arg

[0643] 20

[0644] <210> 87

[0645] <211> 22

[0646] <212> PRT

[0647] <213> Artificial sequence

[0648] <400> 87

[0649] Asp Glu Ile Glu Glu Thr Arg Val Asn Phe Arg Arg Glu Lys Val Ile

[0650] 1 5 10 15

[0651] Glu Thr Pro Glu Asn Asp

[0652] 20

[0653] <210> 88

[0654] <211> 23

[0655] <212> PRT

[0656] <213> Artificial sequence

[0657] <400> 88

[0658] Met Tyr Thr Leu Asn Phe Ala Cys Cys Asp Val Val Gly Leu Ala Ala

[0659] 1 5 10 15

[0660] Val Arg Phe Phe Leu Gly Leu

[0661] 20

[0662] <210> 89

[0663] <211> 18

[0664] <212> PRT

[0665] <213> Artificial sequence

[0666] <400> 89

[0667] Ser Thr Ser Ala Leu Ala Gly Pro Ser Leu Thr Glu Lys Pro Trp Ala

[0668] 1 5 10 15

[0669] Leu Arg

[0670] <210> 90

[0671] <211> 22

[0672] <212> PRT

[0673] <213> Artificial sequence

[0674] <400> 90

[0675] Lys Tyr Met Trp Lys Leu Leu Arg Gln Asp Gln Gln Ser Ile Ile Leu

[0676] 1 5 10 15

[0677] Val Asn Asp Ser Glu Ile

[0678] 20

[0679] <210> 91

[0680] <211> 23

[0681] <212> PRT

[0682] <213> Artificial sequence

[0683] <400> 91

[0684] Asp Asp Glu Ile Glu Glu Thr Arg Val Asn Phe Arg Arg Glu Lys Val

[0685] 1 5 10 15

[0686] Ile Glu Thr Pro Glu Asn Asp

[0687] 20

[0688] <210> 92

[0689] <211> 13

[0690] <212> PRT

[0691] <213> Artificial sequence

[0692] <400> 92

[0693] Arg Met Glu Thr Val Cys Ser Pro Phe Thr Ala Ala Arg

[0694] 1 5 10

[0695] <210> 93

[0696] <211> 15

[0697] <212> PRT

[0698] <213> Artificial sequence

[0699] <400> 93

[0700] Lys Met Gln Glu Cys Phe Gln Lys Ala Phe Phe Arg Lys Pro Lys

[0701] 1 5 10 15

[0702] <210> 94
[0703] <211> 8
[0704] <212> PRT
[0705] <213> Artificial sequence
[0706] <400> 94
[0707] Lys Thr Ser Ala Pro Leu Ile Arg
[0708] 1 5
[0709] <210> 95
[0710] <211> 20
[0711] <212> PRT
[0712] <213> Artificial sequence
[0713] <400> 95
[0714] Met Val Lys Leu Tyr Val Leu Gly Ser Val Leu Ala Leu Phe Gly Val
[0715] 1 5 10 15
[0716] Val Leu Gly Arg
[0717] 20
[0718] <210> 96
[0719] <211> 23
[0720] <212> PRT
[0721] <213> Artificial sequence
[0722] <400> 96
[0723] Val Leu Ala Leu Phe Gly Val Val Leu Gly Arg Met Glu Thr Val Cys
[0724] 1 5 10 15
[0725] Ser Pro Phe Thr Ala Ala Arg
[0726] 20
[0727] <210> 97
[0728] <211> 15
[0729] <212> PRT
[0730] <213> Artificial sequence
[0731] <400> 97
[0732] Val Asn Phe Arg Arg Glu Lys Val Ile Glu Thr Pro Glu Asn Asp
[0733] 1 5 10 15
[0734] <210> 98
[0735] <211> 23
[0736] <212> PRT
[0737] <213> Artificial sequence
[0738] <400> 98
[0739] Ala Asp Asp Glu Ile Glu Glu Thr Arg Val Asn Phe Arg Arg Glu Lys
[0740] 1 5 10 15

[0741] Val Ile Glu Thr Pro Glu Asn

[0742] 20

[0743] <210> 99

[0744] <211> 15

[0745] <212> PRT

[0746] <213> Artificial sequence

[0747] <400> 99

[0748] Ala Leu Ala Gly Pro Ser Leu Thr Glu Lys Pro Trp Ala Leu Arg

[0749] 1 5 10 15

[0750] <210> 100

[0751] <211> 12

[0752] <212> PRT

[0753] <213> Artificial sequence

[0754] <400> 100

[0755] Ser Val Leu Ala Leu Phe Gly Val Val Leu Gly Arg

[0756] 1 5 10

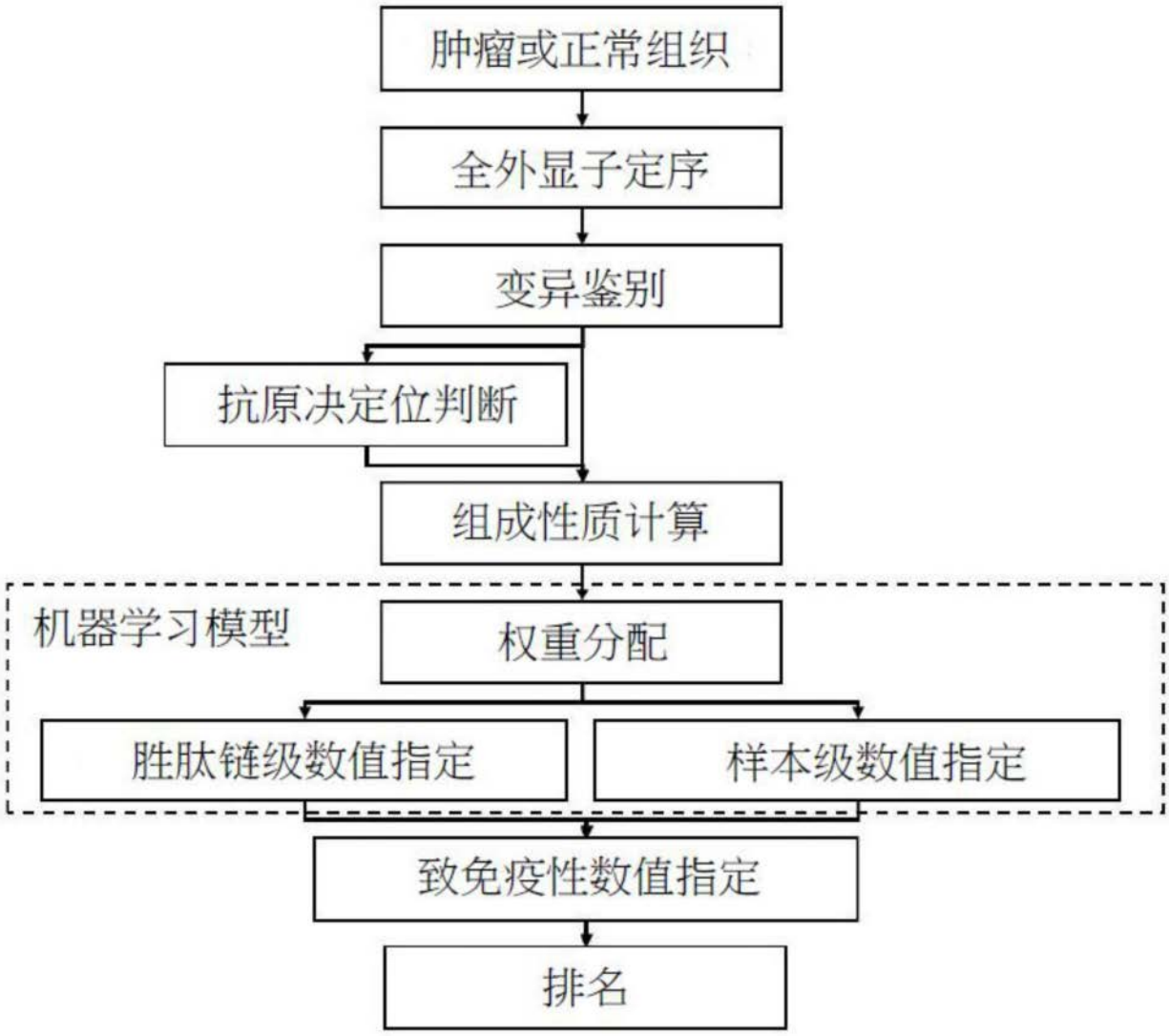

肿瘤或正常组织
全外显子定序
变异鉴别
抗原决定位判断
组成性质计算
机器学习模型
权重分配
胜肽链级数值指定
样本级数值指定
致免疫性数值指定
排名

图1

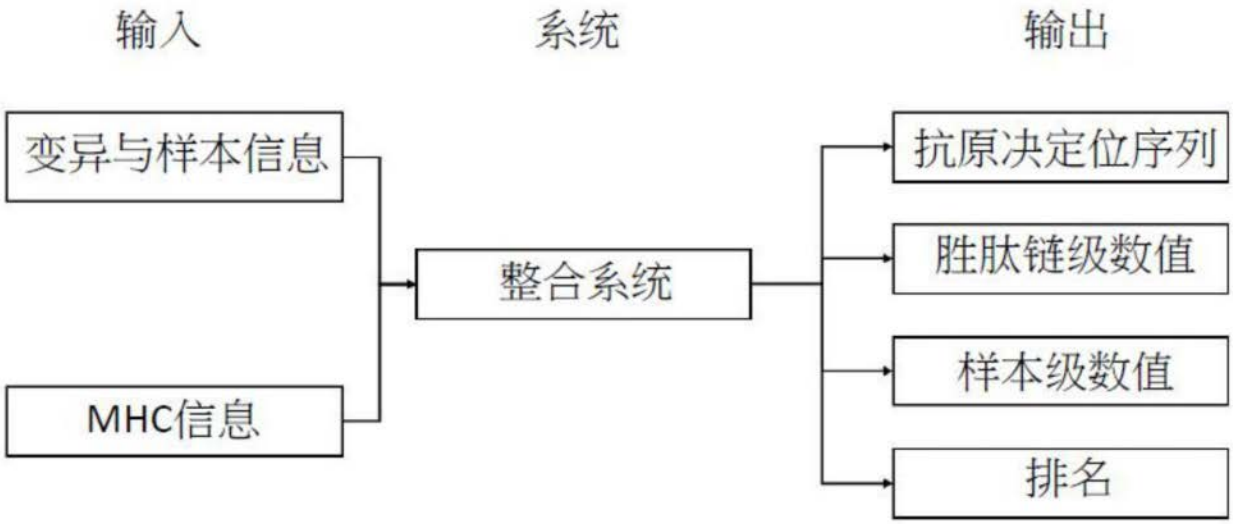

输入
系统
输出
变异与样本信息
MHC信息
整合系统
抗原决定位序列
胜肽链级数值
样本级数值
排名

图2

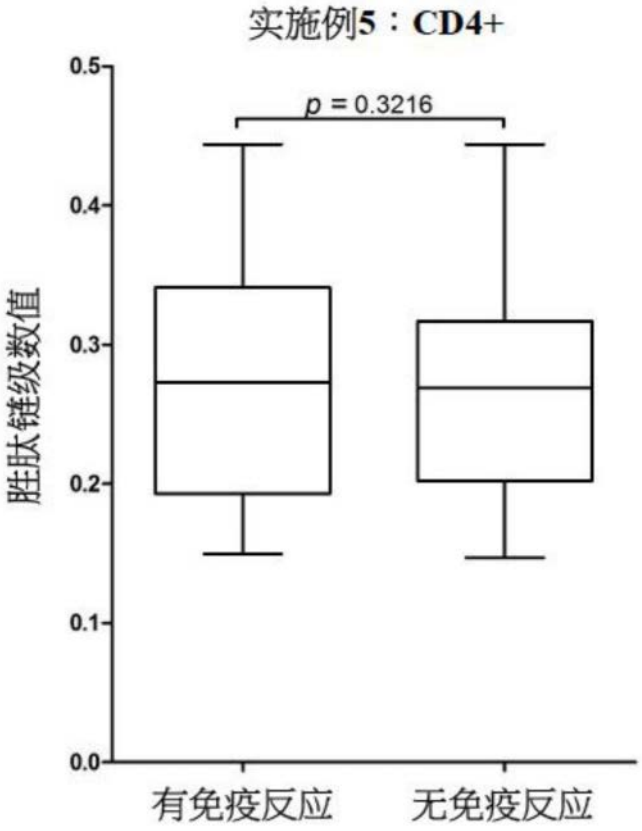

实施例5：CD4+
p = 0.3216
胜肽链级数值
0.5
0.4
0.3
0.2
0.1
0.0
有免疫反应
无免疫反应

图3A

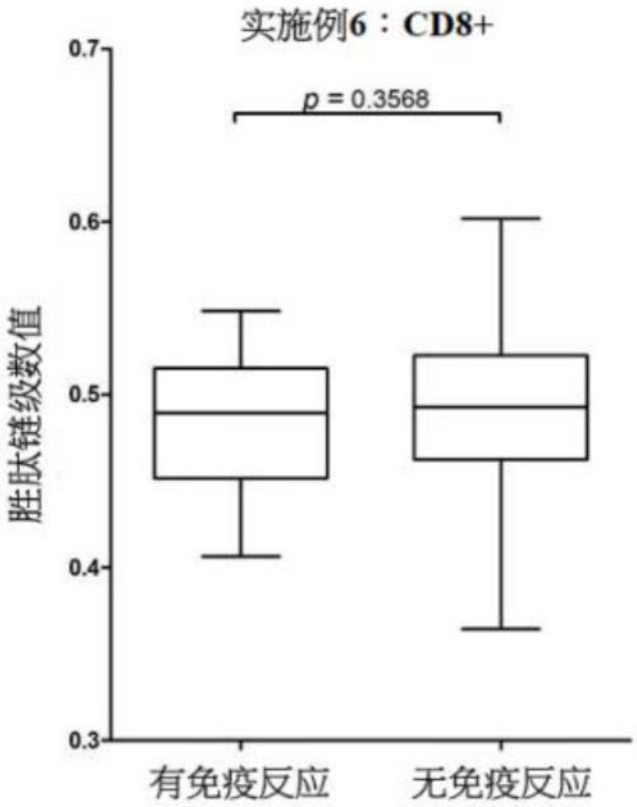


图3B

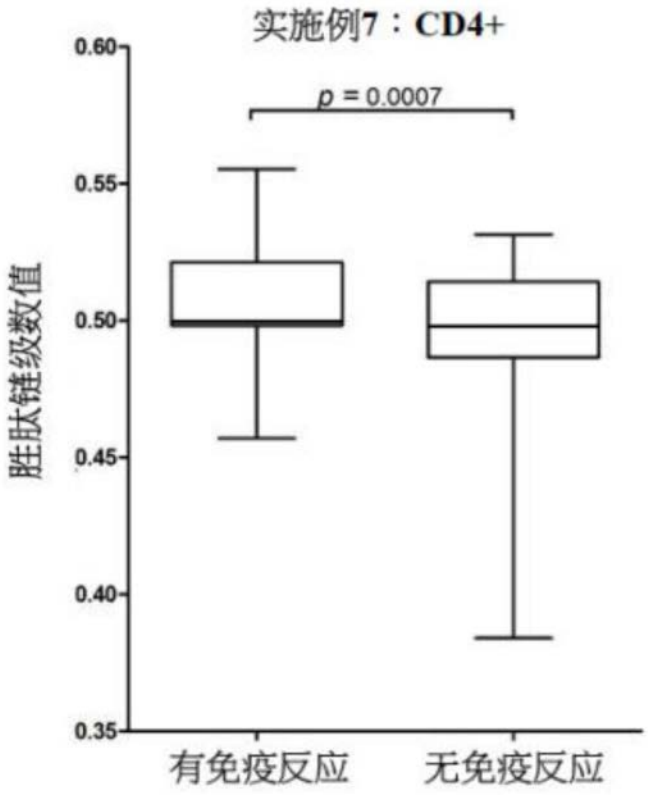

实施例7：CD4+
p = 0.0007
0.60
0.55
0.50
0.45
0.40
0.35
胜肽链级数值
有免疫反应
无免疫反应

图4A

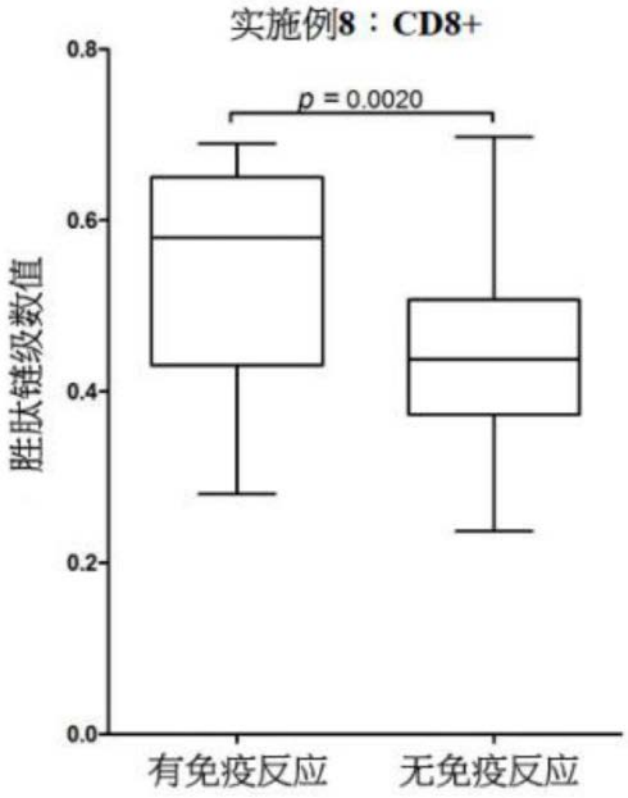

实施例8：CD8+
p = 0.0020
0.8
0.6
0.4
0.2
0.0
胜肽链级数值
有免疫反应
无免疫反应

图4B

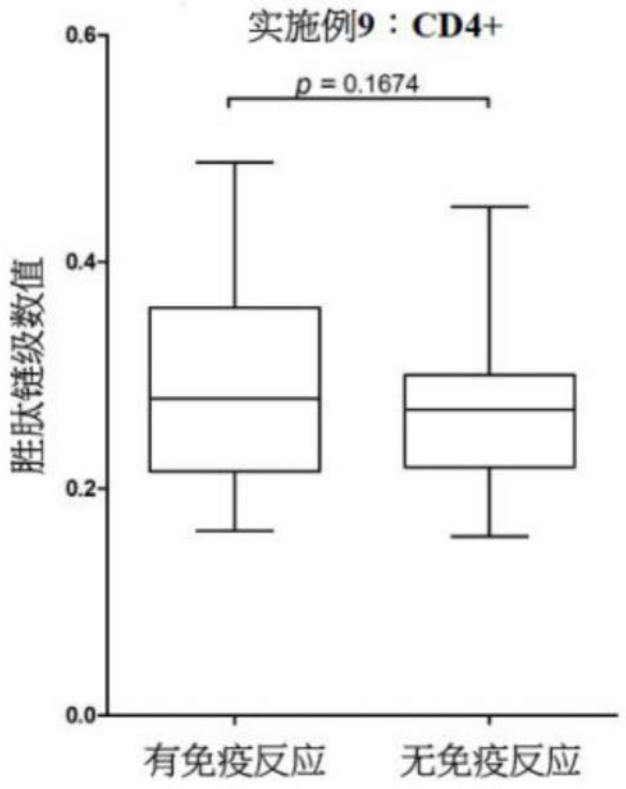

实施例9：CD4+
p = 0.1674
胜肽链级数值
0.6
0.4
0.2
0.0
有免疫反应
无免疫反应

图5A

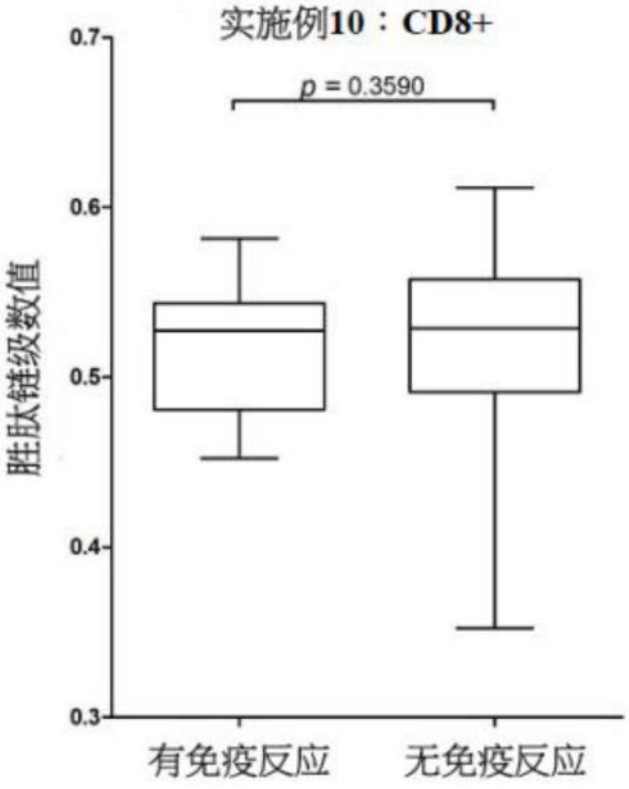

实施例10：CD8+
p = 0.3590
胜肽链级数值
0.7
0.6
0.5
0.4
0.3
有免疫反应
无免疫反应

图5B

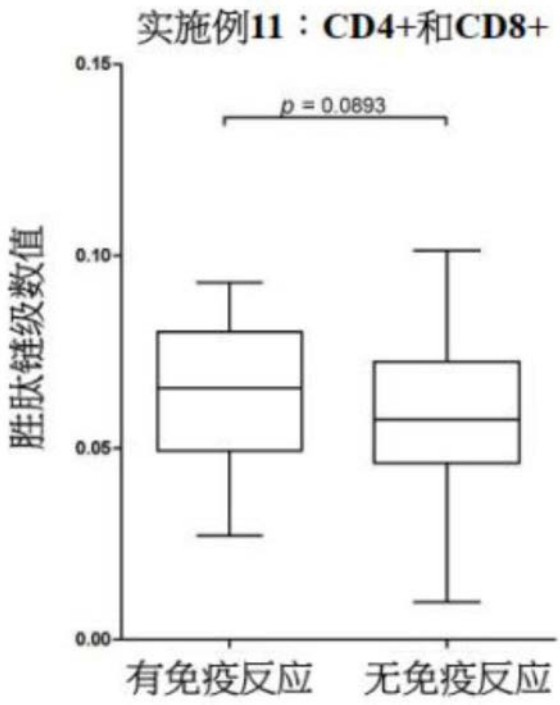

实施例11：CD4+和CD8+
p = 0.0893
胜肽链级数值
0.15
0.10
0.05
0.00
有免疫反应
无免疫反应

图6A

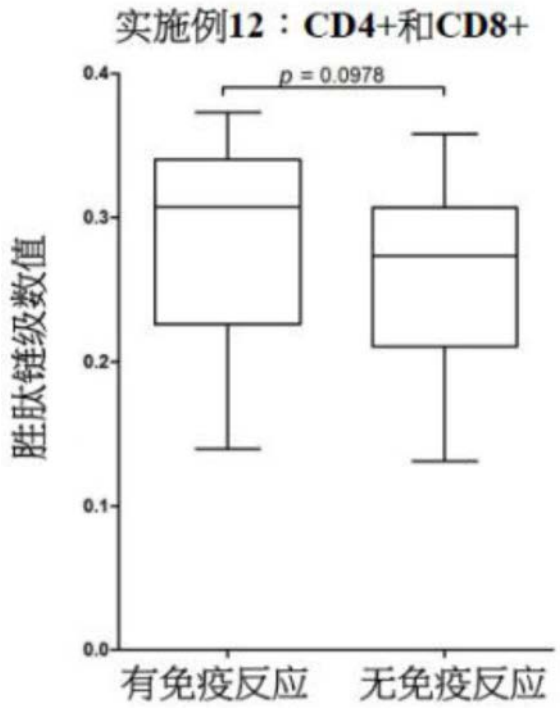

实施例12：CD4+和CD8+
p = 0.0978
胜肽链级数值
0.4
0.3
0.2
0.1
0.0
有免疫反应
无免疫反应

图6B

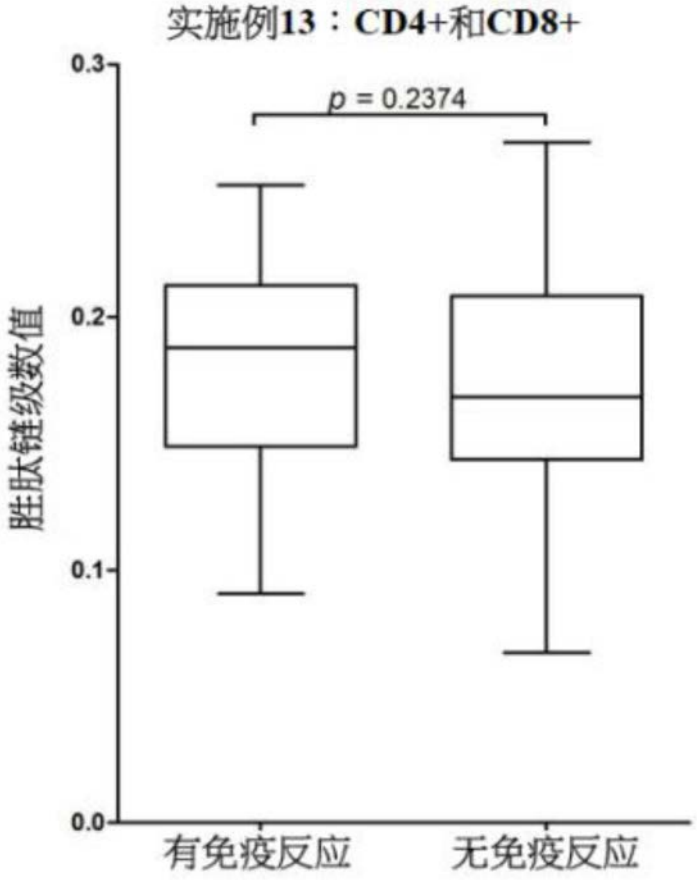


图6C

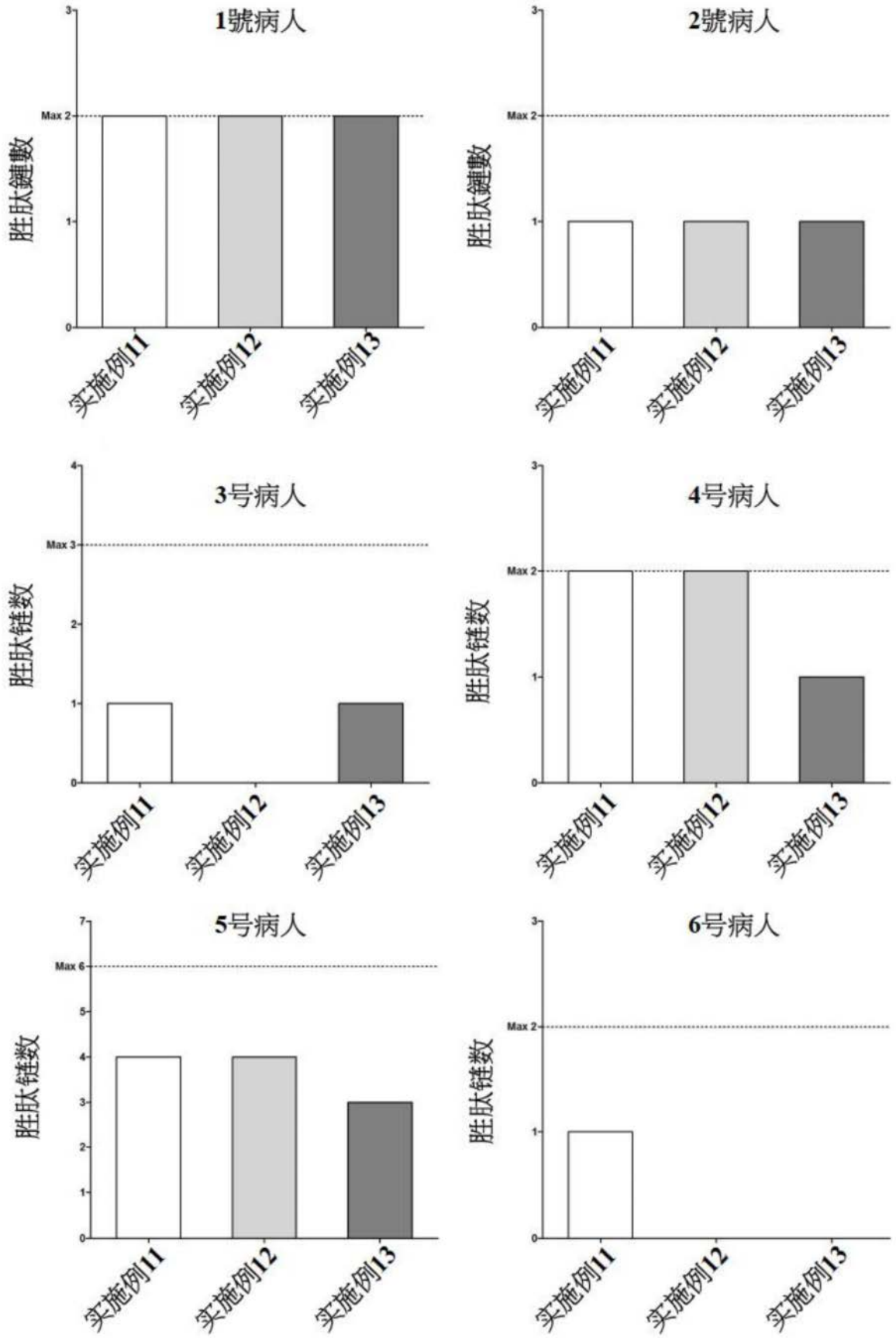

1號病人
胜肽鏈數
Max 2
实施例11
实施例12
实施例13
2號病人
胜肽鏈數
Max 2
实施例11
实施例12
实施例13
3号病人
胜肽链数
Max 3
实施例11
实施例12
实施例13
4号病人
胜肽链数
Max 2
实施例11
实施例12
实施例13
5号病人
胜肽链数
Max 6
实施例11
实施例12
实施例13
6号病人
胜肽链数
Max 2
实施例11
实施例12
实施例13

图7