PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2019,46(7):711~718
www.pibb.ac.cn

基于蛋白质基因组学方法的新抗原鉴定流程*

李雨雨[1,2] 王广志[1,2] 陈兰明[1)**] 谢 鹭[2)**]
([1)] 上海海洋大学食品科学与技术学院，上海 201306；[2)] 上海科学院，上海生物信息技术研究中心，上海 201203）

摘要 肿瘤新抗原是免疫治疗的重要靶点，但基因组数据产生的候选新抗原数量庞大，预测假阳性肽段过多，实验验证费时费力，影响肿瘤新抗原的临床应用. 本研究以乳腺癌为例，使用比转录组水平筛选更严格、比细胞学实验更省时的蛋白质基因组学方法来预测和筛选新抗原. 研究发现，C2（IFN-γ dominant）免疫表型的新抗原数量最多. C2免疫表型显示出最高的M1/M2巨噬细胞极化，较强的CD8信号和最大的T细胞受体多样性，这可能导致产生更多具有良好免疫原性的新抗原. 另外，我们还观察到乳腺癌肿瘤突变负荷与新抗原数目之间呈正相关. 通过同批样本的质谱数据进一步筛选发现，可将两万级别的预测新抗原肽候选降至几十条表达肽段，进而可分析其对应的特异或广谱突变基因. 最后，我们进一步分析了新抗原的免疫原性，即被T细胞受体识别的可能性. 本文利用蛋白质组学数据对基因组数据计算预测得到的候选新抗原进一步筛选，提高了新抗原预测准确性，大大缩小后续实验验证范围. 该流程可为肿瘤新抗原预测与筛选研究提供参考.

关键词 蛋白质基因组学，新抗原，乳腺癌
中图分类号 Q81，R73 **DOI**：10.16476/j.pibb.2018.0304

在肿瘤细胞中，异常蛋白质序列在蛋白酶体的作用下被切割成短肽，然后通过人类白细胞抗原（human lymphocyte antigen，HLA）分子将这些异常肽呈现在细胞表面，最终被T细胞特异性识别为外源性抗原[1-2]，又被称为“新抗原”. 人类的HLA分子也称为主要组织相容性复合体（major histocompatibility complex，MHC），可以广泛地分为MHC-Ⅰ类和MHC-Ⅱ类亚型. 由于正常细胞不会产生和表达肿瘤新抗原，所以新抗原能更有效地激发机体免疫反应. 研究证实，一些肿瘤特异性抗原是检查点阻断疗法和个性化疫苗疗法的靶点[3-4]. Anagnostou等[5]发现在非小细胞肺癌中，新抗原是对检查点阻断的初始反应的相关靶点. Khodadoust等[6]利用抗原递呈谱有效地鉴定了临床相关的肿瘤抗原，揭示了淋巴瘤免疫治疗的有效靶点. 2017年7月，《自然》（*Nature*）杂志同期发表两项独立临床Ⅰ期试验结果，通过对肿瘤细胞及外周血进行基因组测序，寻找基因突变而特异表达的新抗原，然后构建个性化肿瘤疫苗，回输到患者体内激活免疫细胞，杀死带有上述新抗原的肿瘤细胞[7-8]，这是首次在临床试验中取得成功的新抗原疫苗研究. 随后，Keskin等[9]发现新抗原疫苗在Ⅰb期胶质母细胞瘤试验中发生肿瘤内T细胞应答，并有潜力有利地改变胶质母细胞瘤的免疫环境.

与其他肿瘤免疫疗法相比，新抗原疫苗被认为是一种极其有效且安全的免疫治疗，但是其制备复杂耗时，采用免疫学实验筛选与鉴定HLA分子结合肽是个耗时耗力的过程，面对抗原表位的大规模筛选时，单纯采用实验方法鉴定几乎不可能完成. 生物信息学工具的发展则提高了肿瘤新生抗原的筛选能力，基因组大数据和计算机算法加速了肿瘤表位预测，使得抗原表位的大规模筛选成为可能. 目前，国内外已经开发出多种用于表位预测的生物信息学软件，如PSSMHCpan[10]、IEDB[11]、SYFPEITHI[12]、RANKPEP[13]、NetMHCpan[14]等. PSSMHCpan是中国华大基因集团依托全外显子及转录组测序自主开发的肿瘤新抗原检测工具.

* 国家自然科学基金（31870829）资助项目.
** 通讯联系人.
谢鹭. Tel：021-20283705，E-mail：luxiex2017@outlook.com
陈兰明. Tel：021-61900504，E-mail：lmchen@shou.edu.cn
收稿日期：2018-11-24，接受日期：2019-05-22

IEDB作为全球免疫表位信息的门户，提供了一些配套的在线工具用于表位预测与分析. SYFPEITHI数据库提供了一个基于基序打分的表位预测界面，能够预测人类及小鼠的多种MHC分子配体. RANKPEP虽然能够进行MHC-肽结合预测，但是精确度非常低，预测效果较差. NetMHCpan是目前广泛认可的MHC-肽结合预测软件，它是针对常见HLA亚型开发的等位基因特异性预测算法，相对于其他软件效果最佳.

目前国内外所开发的预测工具仅利用基因组和转录组数据预测新抗原，产生的新抗原数量庞大，假阳性高，而T细胞受体仅对非常小部分（约1%）预测的新抗原有反应，这一点对新抗原进一步的筛选是必不可少的. 结合质谱（MS）技术进行新抗原预测与筛选能够提高预测的准确率，质谱技术不仅能鉴定到从肿瘤相关抗原和翻译后修饰产生的肽段，也可以直接鉴定出人类肿瘤组织来源的新抗原. Bassani-Sternberg等[15]利用高精度质谱技术检测到25个黑色素瘤病人相关的免疫肽段组，深度达到95 500个肽段. 不仅发现了大量的包括癌-睾丸抗原以及磷酸肽在内的肿瘤相关抗原，而且直接鉴定到了11个体细胞突变肽段，其中仅有4条突变肽段被证实具有免疫原性. Kalaora等[16]研究发现，在一名黑色素瘤患者的1 019条体细胞突变肽段中，有2条能够在蛋白质水平上得到鉴定，其中1条可刺激机体发生特异性免疫反应. Yadav等[17]将基因组学分析与质谱技术相结合，在小鼠癌细胞系模型中鉴定到2个能够刺激CD8+T细胞并产生免疫应答的新抗原表位. 以上研究表明，目前基于蛋白质基因组学方法预测新抗原仍处于初级阶段，仅仅在少数样本中得到实现，以黑色素瘤居多，而且鉴定到的具有免疫原性的抗原表位数量极少. 国内虽然越来越多的学者开始关注新抗原，但是较之国外还有很大的不足，目前国内未见乳腺癌新抗原疫苗预测与筛选方法的详细报道以及新抗原疫苗临床试验相关的报道. 本研究在利用基因组学数据预测乳腺癌新抗原的基础上，结合质谱技术将突变肽段鉴定引入肿瘤新抗原发现工作流程，提供了一个比转录组水平筛选更严格、比细胞学实验更省时的筛选方法，保证只有那些被MHC-Ⅰ提呈和足够表达的肽段，即最有可能产生免疫应答的肽段进入后续研究. 该工作流程将为肿瘤新抗原预测与筛选研究提供有效的参考.

1　材料与方法

图1为本研究中基于蛋白质基因组学方法的新抗原预测与筛选流程. 对TCGA[18]（https：//cancer.me.nih.gov/）数据库中乳腺癌样本的外显子突变数据进行处理，筛选其中的错义突变；从UniProt[19]（http：//www.uniprot.org/uniprot/）数据库中下载人类标准蛋白质序列，截取包含21个氨基酸长度的突变肽段；用NetMHCpan-3.0预测其中能与常见HLA-Ⅰ型表位结合的抗原肽，保留与HLA-Ⅰ型分子具有亲和力的突变型肽段，即为预测的新抗原肽，用作后续质谱验证；从CPTAC[20]（https：//cptac-data-portal. georgetown. edu/cptacPublic/）公共资源中下载TCGA乳腺癌样本对应的原始质谱数据，对预测的肿瘤突变肽进行鉴定；最后，通过与交叉反应微生物肽进行序列一致性比较，筛选出高可信度的更可能被T细胞识别的新抗原.

1.1　新抗原预测

从TCGA数据库中下载105例乳腺癌样本的体细胞突变信息，筛选出4 780个错义突变位点. 从Uniprot数据库中下载人类标准蛋白质序列（选择UniProtKB/Swiss-Prot中高质量的、手工注释的、非冗余的数据集）. 然后根据错义突变信息将下载的原始蛋白质序列替换成突变后的蛋白质序列. 最后，将错义突变位点附近共21个氨基酸处理成突变肽段.

除了突变肽段之外，乳腺癌样本的HLA基因型也应作为NetMHCpan的输入. 根据千人基因组计划[21]，我们筛选了分布频率>5%的16个HLA分型，包括5个HLA-A等位基因（HLA-A*01:01、HLA-A*02: 01、HLA-A*03: 01、HLA-A*11: 01和HLA-A*24: 02），4个HLA-B等位基因（HLA-B*07: 02、HLA-B*35: 01、HLA-B*40: 01、HLA-B*51:01）和7个HLA-C等位基因（HLA-C*01:02、HLA-C*03: 03、HLA-C*03: 04、HLA-C*04: 01、HLA-C*06: 02、HLA-C*07: 01和HLA-C*07: 02）. 利用NetMHCpan预测其中能与这16个HLA-Ⅰ型分子结合的抗原肽，保留能够与HLA-Ⅰ型分子结合的突变型肽段，即为预测的候选新抗原肽，用作后续质谱验证.

1.2　新抗原过滤

1.2.1　新抗原在蛋白质水平上的鉴定

首先从CPTAC数据库下载使用iTRAQ

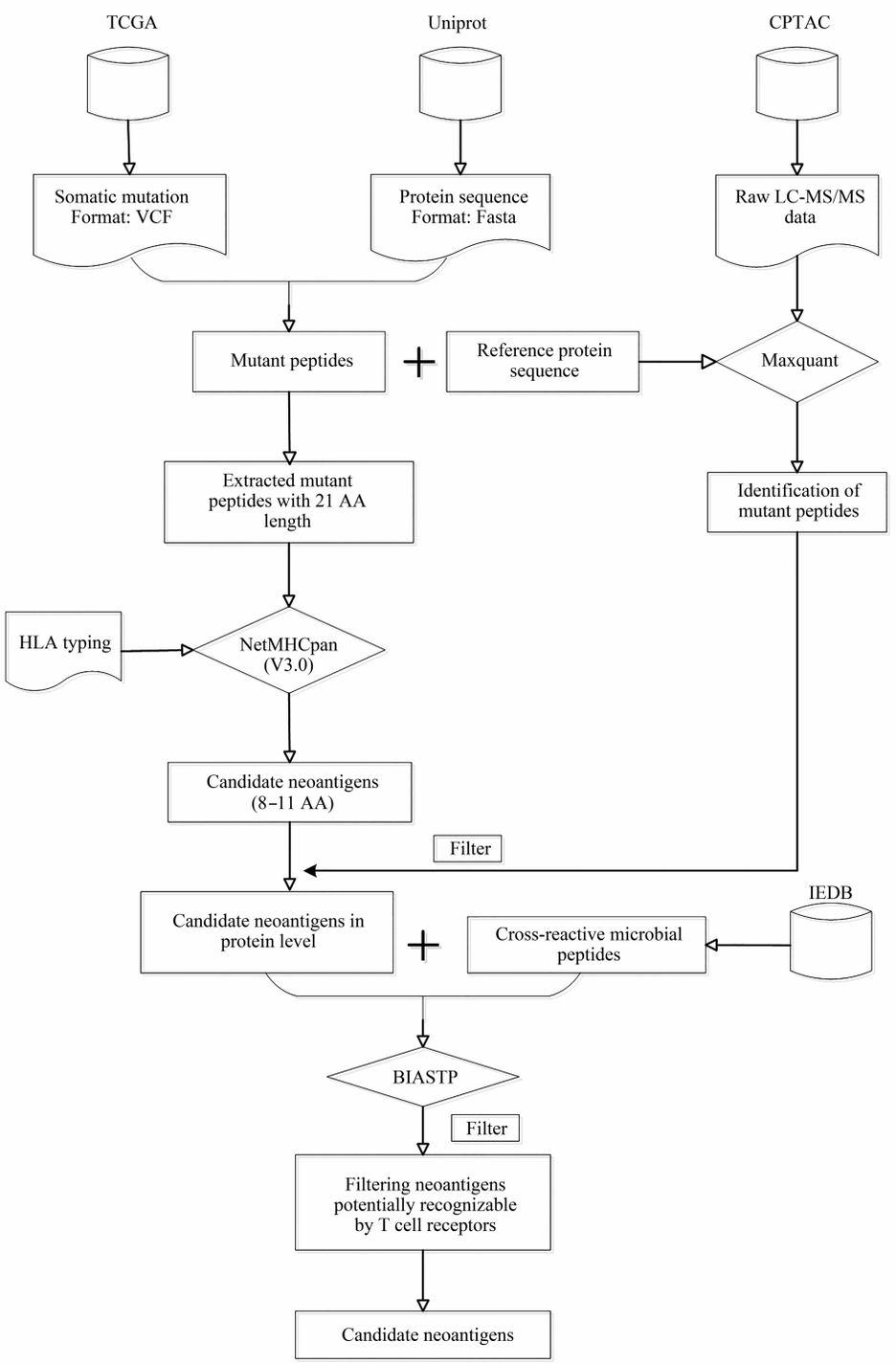

Fig. 1　Neoantigen discovery with proteogenomic method in breast cancer

（isobaric tags for relative and absolute quanti）蛋白质定量方法获得的105个TCGA乳腺癌样本原始质谱数据，然后将突变蛋白质序列以及相应的人类标准蛋白质序列进行建库．最后，我们使用MaxQuant对质谱数据进行搜库以鉴定突变肽段，使用鉴定出的突变肽段对新抗原进行蛋白质水平上的验证．MaxQuant软件参数设置如下：多肽鉴定当中的可变修饰为蛋白质N末端乙酰化，甲硫氨酸氧化、半胱氨酸的氨基甲酰甲基化作为固定修饰，最大漏切位点设置为2个，最小肽段长度设定为7个氨基酸，FDR设置为1%.

1.2.2　T细胞受体识别新抗原的可能性

新抗原鉴定基于两个主要特征：结合MHC-Ⅰ类分子的能力以及被T细胞识别的可能性. 如果新抗原与由传染病衍生的表位具有同源性，这些新抗原就很有可能被人类T细胞受体识别. 肿瘤浸润性T细胞可交叉反应识别肿瘤新抗原和同源非肿瘤微生物抗原. 新抗原和交叉反应性微生物肽的序列一致性越高，T细胞识别新抗原的可能性就越大. 因此，Blastp方法用于研究新抗原和交叉反应性微生物肽的序列一致性，所使用的阳性表位（即交叉反应性微生物肽）数据集来自免疫表位数据库（IEDB，http：//www.iedb.org/）. IEDB数据库收录的表位数量最多、质量好，包含的表位相关的各种背景信息最为丰富，甚至连实验细节也包括在内.

1.3　广谱蛋白质基因组学新抗原预测流程的构建

a. 根据某种肿瘤基因组体细胞突变产生的突变肽段，去掉重复突变，融合得到该种肿瘤广谱的蛋白质突变肽段数据库（本研究使用TCGA乳腺癌105个样本的体细胞突变数据）.

b. 将肿瘤突变肽段数据库、对应的标准蛋白质数据库进行合并，构建广谱肿瘤蛋白质理论数据库（如果有文献中搜集得到的HLA-Ⅰ理论肽段数据，可添加至理论肽库中）.

c. 将广谱肿瘤蛋白质理论数据库进行逆序处理，形成阴性集合，构建肿瘤蛋白质理论肽段数据库反库（如果质谱搜库软件自动构建反库，可忽略此步骤）.

d. 对该肿瘤的群体质谱数据，利用质谱搜库软件进行理论搜库（本研究使用Maxquant软件），采用严格的质量控制标准（FDR设置为0.01），降低假阳性率.

e. 使用Python脚本对搜库结果进行处理，根据基因组突变信息得到可能存在的突变肽段. 将该部分突变肽段与基因组得到的候选新抗原进行匹配筛选能够在蛋白质水平表达的候选新抗原.

2　结果与分析

2.1　新抗原肽段的预测结果

由NetMHCpan-3.0软件所鉴定的新抗原根据%rank值进行分类，%rank值是对亲和力打分（所预测肽段亲和力与一组随机天然肽亲和力的比值）的一个矫正，%rank值越小说明偏离的越小，肽段与MHC-Ⅰ类分子的结合力越强. 本研究中以0.5和2作为%rank的阈值，若%rank<0.5，认为该短肽是MHC-Ⅰ类分子的强结合力抗原肽；%rank <2，则认为该短肽是MHC-Ⅰ类分子的弱结合力抗原肽；%rank>2，预测肽无结合力，不给予考虑. 预测肽与MHC-Ⅰ类分子的结合力越强，其成为新抗原表位的可能性越大. 根据%rank值进行筛选后，我们得到7 487个高结合肽（%rank<0.5），20 915个低结合肽（0.5<%rank<2）（附件1）. 在这部分候选的新抗原中会出现一条肽段在高低结合肽段中同时存在的情况，这是因为一条肽段跟不同分型结合时的亲和力不同，在后续的分析中，我们将其视为一条肽段，经过去除重复肽段后，共得到23 817条候选新抗原. 这些新抗原肽段根据乳腺癌亚型（Basal、Her2、LumA、LumB）和免疫表型（C1: wound healing、 C2: IFN- γ dominant、C3: inflammatory、 C4: lymphocyte depleted、C5: immunologically quiet、C6:TGF-β dominant）进一步分类，我们发现新抗原数目与乳腺癌亚型之间无显著性关联（表1）. 然而，C2（IFN- γ dominant）免疫表型具有最多的抗原数目. 据报道，C2免疫表型显示出最高的M1/M2巨噬细胞极化，较强的CD8信号和最大的TCR多样性[22]，这可能导致产生更多而且具有良好免疫原性的新抗原. 图2显示了每个乳腺癌亚型的每个样本中新抗原数目的分布并显示了很大的差异. 我们观察到每个样本都有独特的新抗原集，这与肿瘤异质性相关，也提示基于新抗原的乳腺癌患者的免疫治疗可能是个性化的.

据我们所知，肿瘤突变负荷（tumor mutation burden，TMB）已成为一种潜在的生物标志物，有助于预测免疫治疗的有效性[23]. 因此，我们使用Pearson相关算法计算了105个乳腺癌样本的肿瘤突变负荷与新抗原数量之间的相关性，从图3中可以发现新抗原数量与肿瘤特异性非沉默体细胞突变负荷之间存在正线性关系（R^2=0.7164，P=2.2e−16）. 突变负荷比较高的肿瘤（例如黑色素瘤），即TMB>10/Mb时，经常会产生新抗原，因此对免疫检查点抑制剂敏感，治疗效果更好. 本研究所使用的105例乳腺癌样本TMB值在0.3/Mb~8/Mb之间. 当1/Mb<TMB<10/Mb时，对应的乳腺癌样本能够产生新抗原，对PD-1/PD-L1抑制剂可能敏感，需要联合治疗来适当增加新抗原的产生. 还观察到有一部分样本中的肿瘤TMB<1/Mb，几乎不太可能产生新抗原，因此这部分样本对PD-1/PD-L1抑制剂不敏感. 在临床研究中，可以通过检

Table 1　The number of neoantigen in each breast cancer subtype （columns） and each immune subtype （rows）

	Basal	Her2	LumA	LumB	Total
C1	1 647（9）	1 951（8）	2 193（10）	1 788（7）	7 543（34）
C2	4 743（15）	3 295（9）	778（7）	4 142（17）	12 957（48）
C3	0	1 106（1）	513（6）	0	1 619（7）
C4	0（1）	0	289（4）	832（7）	1 121（12）
C5	0	0	0	0	0
C6	0	0	263（2）	277（2）	540（4）
Total	6 390（25）	6 352（18）	4 036（29）	7 039（33）	23 817（105）

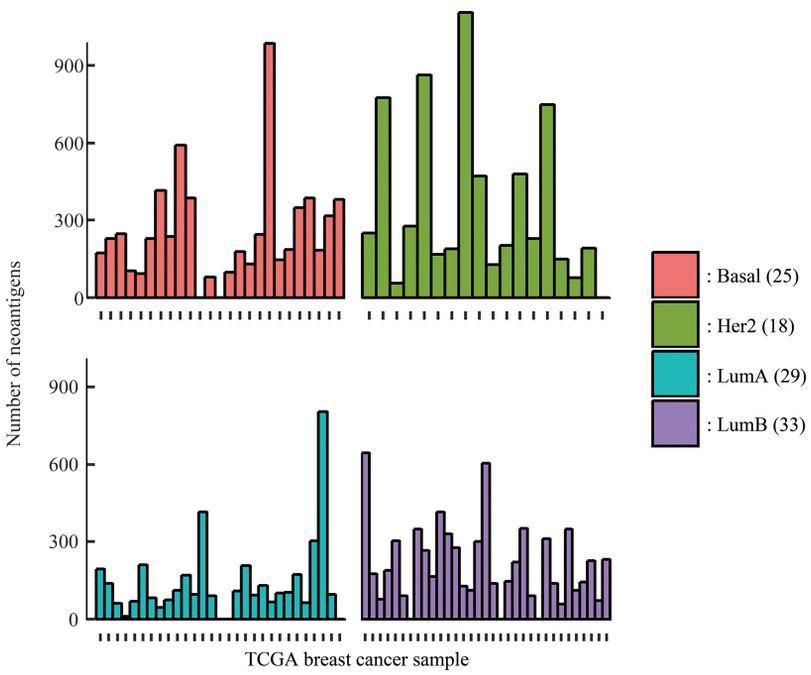

Fig. 2　The number of neoantigens in each sample of each breast cancer subtype

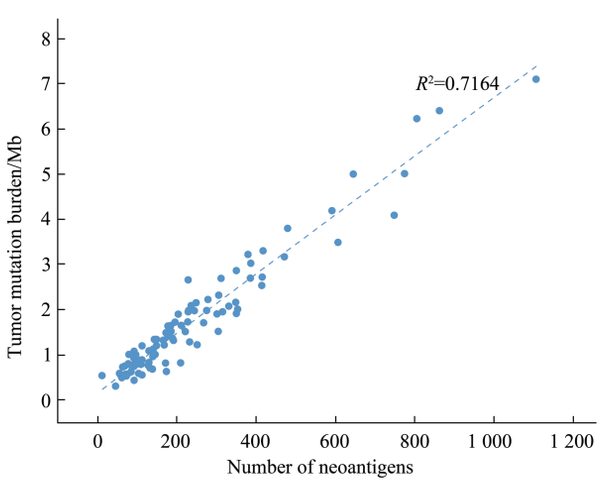

Fig. 3　Correlation between neoantigen number and tumor mutation burden

测 TMB 来预测肿瘤突变所产生的新抗原数量，采用更合理的治疗方法. TMB 越高，肿瘤细胞所产生的能够被 T 细胞识别的新抗原越多，免疫细胞对肿瘤的杀伤力越大.

2.2　新抗原在蛋白质水平上的鉴定

通过 MaxQuant [24] 软件对质谱文件进行搜库分析后得到 77 条突变肽段（8~37AA）. 基于 NetMHCpan 预测的新抗原肽（8~11AA），我们使用 Python 脚本进行新抗原过滤，筛选基因突变信息一致以及肽段相同的候选新抗原，发现只有 75 个候选新抗原能够在肽段水平得到鉴定（网络版附件）. 据报道，这些候选新抗原的许多基因与乳腺癌的发生发展以及转移预后等过程相关 [25-27]，包括 ERBB2IP （Erbin）、 HIST1H3C、 YWHAB、TP53BP1、IL16、ADRBK1 等. 例如，TP53BP1 的

表达与乳腺癌的预后有关，其表达水平的增加抑制了乳腺癌的侵袭和转移[28]. 研究表明[29]，乳腺癌组织中Erbin基因的敲除能够显著促进细胞迁移. Tran等[30]已证实，从一位晚期胆管癌患者肿瘤组织分离出来的一部分CD4阳性T细胞能够识别ERBB2IP基因突变所产生的异常蛋白质. 他们将T淋巴细胞扩增、激活，重新回输给了患者后，发现病人的肿瘤缩小至完全消失，成功地治疗了这位晚期转移性胆管癌患者.

2.3 新抗原与交叉反应性微生物肽的序列一致性

在75个候选新抗原中，我们通过blastp方法得到11个新抗原与交叉反应性微生物肽具有序列一致性，结果显示1条抗原肽可能与2条不同的微生物肽具有序列一致性（网络版附件）. 这11个突变肽段属于高可信度新抗原，可能更易被T细胞受体所识别. 新抗原和交叉反应肽序列一致性得分在20～60之间. 通常，序列一致性越高，新抗原被T细胞受体识别的可能性越大.

3 讨 论

基因组数据产生的候选新抗原肽段的数量庞大、假阳性肽段过多、实验验证费时费力是影响肿瘤新抗原临床应用的一大挑战，质谱检测可以很好地改善这一问题. 本研究综合了基因组学以及蛋白质组学鉴定流程，用蛋白质组学质谱数据直接验证肿瘤新抗原是否有肽段表达的证据. 首先，NetMHCpan软件用来预测肽段/MHC-Ⅰ类分子的亲和力，得到大量的候选新抗原. 然后通过同批样本的质谱数据进一步筛选预测的突变肽，这些突变肽能够在蛋白质水平上得到鉴定，进而可分析其对应的特异或广谱突变基因. 此外，为了进一步检验这些候选新抗原能否被T细胞受体识别，通过突变肽与交叉反应性微生物肽序列一致性研究，发现11个高可信度、可能具有良好免疫原性的新抗原. 本研究的肿瘤新抗原预测工作流程可大大缩小后续实验验证范围，提供了一个比转录组筛选更严格、比细胞学实验更省时的筛选工具，保证只有那些被MHC-Ⅰ提呈和足够表达的肽段，即最有可能产生免疫应答的肽段进入后续研究. 随着大规模平行测序和蛋白质组学越来越适用，这种分析会为肿瘤免疫治疗提供参考.

本研究观察到乳腺癌突变负荷与新抗原数目之间呈正相关关系，可通过检测TMB预测新抗原产生的数量，判断是否需要联合治疗来增加新抗原的产生，使得癌症患者对免疫检查点抑制剂更敏感. TMB越高，肿瘤组织产生的新抗原数量越多，肿瘤的免疫杀伤活性越大. Charoentong等[31]分析了20种实体肿瘤的新抗原，在911 548个独特的新抗原中，仅有24条肽段在5%的病人中能够共享，这24条新抗原并未在乳腺癌中出现. 不同肿瘤或同种肿瘤之间存在的共享新抗原大多来源于驱动突变基因，如BRAF、RAS以及PIK3CA. 驱动突变形成新抗原是最理想的状态，如果大部分肿瘤细胞都具这种突变，一旦产生新抗原，那么针对新抗原的T细胞能消灭大部分肿瘤. 遗憾的是，驱动突变很少产生新抗原，Schumacher等[32]在约20 000个黑色素瘤中发现的20种新抗原中，只有8%的新抗原来自驱动突变，而92%的新抗原来自非驱动突变. 本研究中经质谱鉴定得到的候选新抗原的突变并没有来源于乳腺癌的驱动基因，而是来自乘客基因. 在105个乳腺癌样本中，每个肿瘤突变产生的新抗原是独一无二的，每个样本的新抗原具有特异性，样本之间不存在共享相同的新抗原，与上述文献中所报道的新抗原很少在患者之间共享的结果一致，体现了肿瘤新抗原预测需在个性化水平上进行.

肿瘤新抗原预测的影响因素有很多，未知因素超过40%，新抗原与MHC分子的亲和力仅占28%，仅仅使用预测算法进行理论预测，即使达到理论预测极限也只有28%的准确率. 结合质谱技术将突变肽段鉴定引入肿瘤新抗原，发现工作流程可以提高预测准确率，大大缩小后续实验的验证范围，但候选新抗原需要进一步的实验验证，才能确定其潜在引起免疫治疗应答反应的能力. 质谱法的灵敏度目前仍然有限，因此可能导致假阴性. 以目前的国内外研究现状来看，结合蛋白质基因组学方法预测新抗原仍然属于新的研究领域，即使分析高精度质谱得到HLA-抗体富集的数据，预测能被提呈的肿瘤新抗原肽段中也只有很少的数量能被质谱鉴定到. 但是，目前无论是新抗原预测算法的开发还是新抗原预测流程的建立，研究者们都试图将质谱分析纳入新抗原预测流程，因为蛋白质基因组学策略已经是最大程度地利用了最新质谱技术及多组学整合生物信息分析技术进行新抗原预测的最佳途径.

附件 20180304S1.xlsx见本文网络版(http://www.ibp.ac.cn或http://www.cnki.net).

参 考 文 献

[1] Gfeller D, Bassani-Sternberg M, Schmidt J, *et al*. Current tools for predicting cancer-specific T cell immunity. Oncoimmunology, 2016, **5**(7): e1177691

[2] Van Rooij N, Van Buuren M M, Philips D, *et al*. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. J Clin Oncol, 2013, **31**(32): e439-442

[3] Gubin M M, Zhang X, Schuster H, *et al*. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature, 2014, **515**(7528): 577-581

[4] Matsushita H, Vesely M D, Koboldt D C, *et al*. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. Nature, 2012, **482**(7385): 400-404

[5] Anagnostou V, Smith K N, Forde P M, *et al*. Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. Cancer Discov, 2017, **7**(3): 264-276

[6] Khodadoust M S, Olsson N, Wagar L E, *et al*. Antigen presentation profiling reveals recognition of lymphoma immunoglobulin neoantigens. Nature, 2017, **543**(7647): 723-727

[7] Ott P A, Hu Z, Keskin D B, *et al*. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature, 2017, **547**(7662): 217-221

[8] Sahin U, Derhovanessian E, Miller M, *et al*. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature, 2017, **547**(7662): 222-226

[9] Keskin D B, Anandappa A J, Sun J, *et al*. Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. Nature, 2019, **565**(7738): 234-239

[10] Liu G, Li D, Li Z, *et al*. PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. Gigascience, 2017, **6**(5): 1-11

[11] Vita R, Overton J A, Greenbaum J A, *et al*. The immune epitope database (IEDB) 3.0. Nucleic Acids Res, 2015, **43**(Database issue): D405-412

[12] Rammensee H, Bachmann J, Emmerich N P, *et al*. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics, 1999, **50**(3-4): 213-219

[13] Reche P A, Reinherz E L. Prediction of peptide-MHC binding using profiles. Methods Mol Biol, 2007, **409**:185-200

[14] Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med, 2016, **8**(1): 33

[15] Bassani-Sternberg M, Braunlein E, Klar R, *et al*. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. Nat Commun, 2016, **7**:13404

[16] Kalaora S, Barnea E, Merhavi-Shoham E, *et al*. Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. Oncotarget, 2016, **7**(5): 5110-5117

[17] Yadav M, Jhunjhunwala S, Phung Q T, *et al*. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature, 2014, **515**(7528): 572-576

[18] Tomczak K, Czerwinska P, Wiznerowicz M. The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn), 2015, **19**(1A): A68-77

[19] Uniprot C. UniProt: a hub for protein information. Nucleic Acids Res, 2015, **43**(Database issue): D204-212.

[20] Mertins P, Mani D R, Ruggles K V, *et al*. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature, 2016, **534**(7605): 55-62

[21] Gourraud P A, Khankhanian P, Cereb N, *et al*. HLA diversity in the 1000 genomes dataset. Plos One, 2014, **9**(7): e97282

[22] Thorsson V, Gibbs D L, Brown S D, *et al*. The immune landscape of cancer. Immunity, 2018, **48**(4): 812-830

[23] Chalmers Z R, Connelly C F, Fabrizio D, *et al*. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med, 2017, **9**(1): 34

[24] Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p. p. b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol, 2008, **26**(12): 1367-1372

[25] Connolly R M, Fackler M J, Zhang Z, *et al*. Tumor and serum DNA methylation in women receiving preoperative chemotherapy with or without vorinostat in TBCRC008. Breast Cancer Res Treat, 2018, **167**(1): 107-116

[26] Milke L, Schulz K, Weigert A, *et al*. Depletion of tristetraprolin in breast cancer cells increases interleukin-16 expression and promotes tumor infiltration with monocytes/macrophages. Carcinogenesis, 2013, **34**(4): 850-857

[27] Zhang C, Chen X, Li Y, *et al*. si-RNA-mediated silencing of ADRBK1 gene attenuates breast cancer cell proliferation. Cancer Biother Radiopharm, 2014, **29**(8): 303-309

[28] De Gregoriis G, Ramos J A, Fernandes P V, *et al*. DNA repair genes PAXIP1 and TP53BP1 expression is associated with breast cancer prognosis. Cancer Biol Ther, 2017, **18**(6): 439-449

[29] Liu D, Shi M, Duan C, *et al*. Downregulation of Erbin in Her2-overexpressing breast cancer cells promotes cell migration and induces trastuzumab resistance. Mol Immunol, 2013, **56**(1-2): 104-112

[30] Tran E, Turcotte S, Gros A, *et al*. Cancer immunotherapy based on mutation-specific CD4+ T cells in a patient with epithelial cancer. Science, 2014, **344**(6184): 641-645

[31] Charoentong P, Finotello F, Angelova M, *et al*. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep, 2017, **18**(1): 248-262

[32] Schumacher T N, Schreiber R D. Neoantigens in cancer immunotherapy. Science, 2015, **348**(6230): 69-74

The Workflow of Neoantigen Identification Based on Proteogenomic Methodology*

LI Yu-Yu[1,2], WANG Guang-Zhi[1,2], CHEN Lan-Ming[1)**], XIE Lu[2)**]

([1)]*College of Food Science and Technology, Shanghai Ocean University, Shanghai* 201306, China;

[2)]*Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai* 201203, China)

Abstract Tumor neoantigens are important targets for immunotherapy. Based on high-throughput tumor genomic analysis, each missense mutation can potentially give rise to multiple neopeptides, numerous false positive neoantigens will be produced, experimental verification would require immense time and experience. Specific identification of immunogenic candidate neoantigens is consequently a major challenge. Here we introduce a workflow to predict and filter neoantigens of breast cancer with proteogenomic methodology, which can be beneficial to high quality identification of neoantigens. We found that C2 (IFN - γdominant) immunophenotype possessed the most number of neoantigens. C2 immunophenotype shows the highest M1/M2 macrophage polarization, a strong CD8 signal and the greatest T cell receptor(TCR) diversity, which may lead to more neoantigen number than in other immunophenotypes of breast cancer. In addition, we also observed that there is a positive linear relationship between neoantigen number and tumor mutation burden. By further screening for predicted tumor mutant peptides using mass spectral data of breast cancer, we found that more than 20 000 predicted neoantigens were reduced to dozens of mutant peptides in protein expression level, the corresponding mutant genes could be further analyzed. Finally, in order to define which neoantigens were more likely to be immunogenic, TCR recognition probability was calculated using blastp method. In this study, proteomics data was used to further screen the predicted neoantigens, which improved the prediction accuracy of neoantigens, and could greatly reduce the validation scope of potential subsequent experiments. This workflow provides a new insight for tumor neoantigen prediction and screening.

Key words proteogenomic, neoantigen, breast cancer

DOI：10.16476/j.pibb.2018.0304

---

* This work was supported by a grant from The National Natural Science Foundation of China（31870829）.

** Corresponding author.

CHEN Lan-Ming. Tel：86-21-61900504，E-mail：lmchen@shou.edu.cn

XIE Lu. Tel：86-21-20283705，E-mail：luxiex2017@outlook.com

Received：November 24，2018 Accepted：May 22，2019