DOI:10.12113/202504006

Neo-Pred:全变异来源的肿瘤新生抗原检测流程

杜　航[1,2],唐景玲[1],周　玲[2],杨　远[1]
(1.贵州医科大学附属医院 临床医学研究中心 贵阳 550004;
2.贵州生诺生物科技有限公司 贵阳 550004)

摘　要:基于体细胞突变产生的新生抗原因其肿瘤特异性高、免疫原性强且不表达于正常组织的特点,成为激活抗肿瘤T细胞应答的理想靶点。目前能全面检测来自不同变异来源的新生抗原的生物信息学工具仍然很匮乏。基于Snakemake流程管理工具,我们开发了Neo-Pred肿瘤新生抗原检测流程,它可以读取高通量测序数据,检测单核苷酸变异(Single nucleotide variant, SNV)、插入缺失(Insertion-deletion, InDel)、基因融合、可变剪接多种变异衍生的新生抗原。我们在肿瘤新生抗原筛选联盟提供的基准数据集上进行了测试,其新生抗原检出的性能为精确率-召回率曲线下面积(Area under the precision-recall curve, AUPRC) 0.71,领先于肿瘤新生抗原筛选联盟其他参与团队(全部参与机构均值为0.221,其中表现最好的团队均值为0.540),筛选性能提升31.5%~221.3%,展示出领先的新生抗原检测能力。通过Singularity容器化和模块化设计,Neo-Pred实现了良好的稳定性、可移植性与动态扩展性。

关键词:新生抗原;单核苷酸突变;基因融合;可变剪接;流程
中图分类号:R73　　**文献标志码**:A　　**文章编号**:1672-5565(2026)01-095-06

Neo-Pred: A comprehensive workflow for detecting tumor neoantigens from all types of mutation sources

DU Hang[1,2], TANG Jingling[1], ZHOU Ling[2], YANG Yuan[1]
(1. *Clinical Medical Research Center, The Affiliated Hospital of Guizhou Medical University, Guiyang* 550004, *China*; 2. *Guizhou Sinorda Biotechnology Co., Ltd, Guiyang* 550004, *China*)

Abstract: Neoantigens derived from somatic mutations have emerged as ideal targets for activating anti-tumor T-cell responses due to their high tumor specificity, strong immunogenicity, and absence of expression in normal tissues. Current bioinformatics tools remain limited in comprehensively detecting neoantigens originating from diverse genomic variations. To address this challenge, we developed Neo-Pred, a tumor neoantigen detection pipeline based on the Snakemake workflow management system. This pipeline processes high-throughput sequencing data to identify neoantigens derived from multiple variant types, including single nucleotide variants (SNVs), insertions-deletions (InDels), gene fusions, and alternative splicing. When evaluated on the benchmark dataset from the Tumor Neoantigen Screening Consortium, Neo-Pred demonstrated superior performance with an Area Under the Precision-Recall Curve (AUPRC) of 0.71 (mean AUPRC: 0.221 for all teams; 0.540 for the top-performing team). This represents a performance improvement of 31.5% to 221.3%, highlighting its leading-edge detection capabilities. The implementation of Singularity containerization and modular architecture ensures remarkable stability, portability, and dynamic scalability. These technical advancements establish Neo-Pred as a cutting-edge solution for neoantigen detection, providing critical support for precision cancer immunotherapy research.

Keywords: Neoantigen; Single nucleotide variant; Gene fusion; Alternative splicing; Workflow

收稿日期:2025-04-12;**修回日期**:2025-05-19;**网络首发日期**:2025-07-11.
网络首发地址:https://link.cnki.net/urlid/23.1513.Q.20250710.1829.002
基金项目:国家自然科学基金(No.82260584);贵州省科技厅项目(No.黔科合支撑[2022]一般193、黔科合基础-ZK[2023]一般359、黔科合支撑[2023]一般373);贵州医科大学附属医院2024年国家自然科学基金培育计划(地区基金)(No.gyfynsfc[2024]-21).
*通信作者:杜航,男,主管技师,研究方向:生物信息学、肿瘤免疫.E-mail:dh274205085@gmc.edu.cn.
引用格式:杜航,唐景玲,周玲,等. Neo-Pred:全变异来源的肿瘤新生抗原检测流程[J].生物信息学,2026,24(1):95-100.
DU Hang,TANG Jingling,ZHOU Ling,et al. Neo-Pred: A comprehensive workflow for detecting tumor neoantigens from all types of mutation sources[J].Chinese Journal of Bioinformatics,2026,24(1):95-100.

肿瘤新生抗原是体细胞突变衍生出的肿瘤特异性的抗原。以肿瘤新生抗原为靶点的免疫治疗方案受到广泛的关注,目前有多项临床试验取得了令人瞩目的进展[1-3]。基于高通量测序和生物信息学手段间接检测新生抗原是常见的选择靶点的方法,然而现有的多种新生抗原的检测工具均存在多种限制,例如 ImmuneMirror、ProGeo-Neo 等[4-5]工具只能提供 SNV 或者 InDel 类型的突变,未能涵盖全面的变异类型。同时,这些工具也缺少检测新生抗原全面的必要的功能,例如 SIGANE[6]需要另行准备包括了表达量信息、突变频率信息以及多种工具预测的亲和力的结果才能进行候选肿瘤抗原的筛选,它仅执行肿瘤抗原筛选的众多步骤中的一个环节。而 pVACtools、NeoPredPipe 等[7-8]工具需要提供变异列表和 HLA 分型列表,pVACtools 还需要新生抗原的表达谱等可选信息,经过多个分析步骤才能完成肿瘤新生抗原的检测工作。目前也有一些多变异类型来源的肿瘤新生抗原发现的贯穿生物信息学流程被开发出来,比如 ScanNeo2[9],但是其整个流程封装于同一个 docker 容器内部,不利于规模化计算集群分发子任务实现规模化生产。基于此,我们开发了 Neo-Pred,一个基于 Snakemake 和容器化技术的肿瘤新生抗原检测生物信息流程,可以实现规模化的肿瘤新生抗原检测。

1 方法

Neo-Pred 是使用 Sankemake 编写的涵盖了 SNV/InDel、基因融合、可变剪接的全变异来源的肿瘤新生抗原检测生物信息学流程(图 1),其调用的软件和工具均已经封装为 Singularity 镜像。

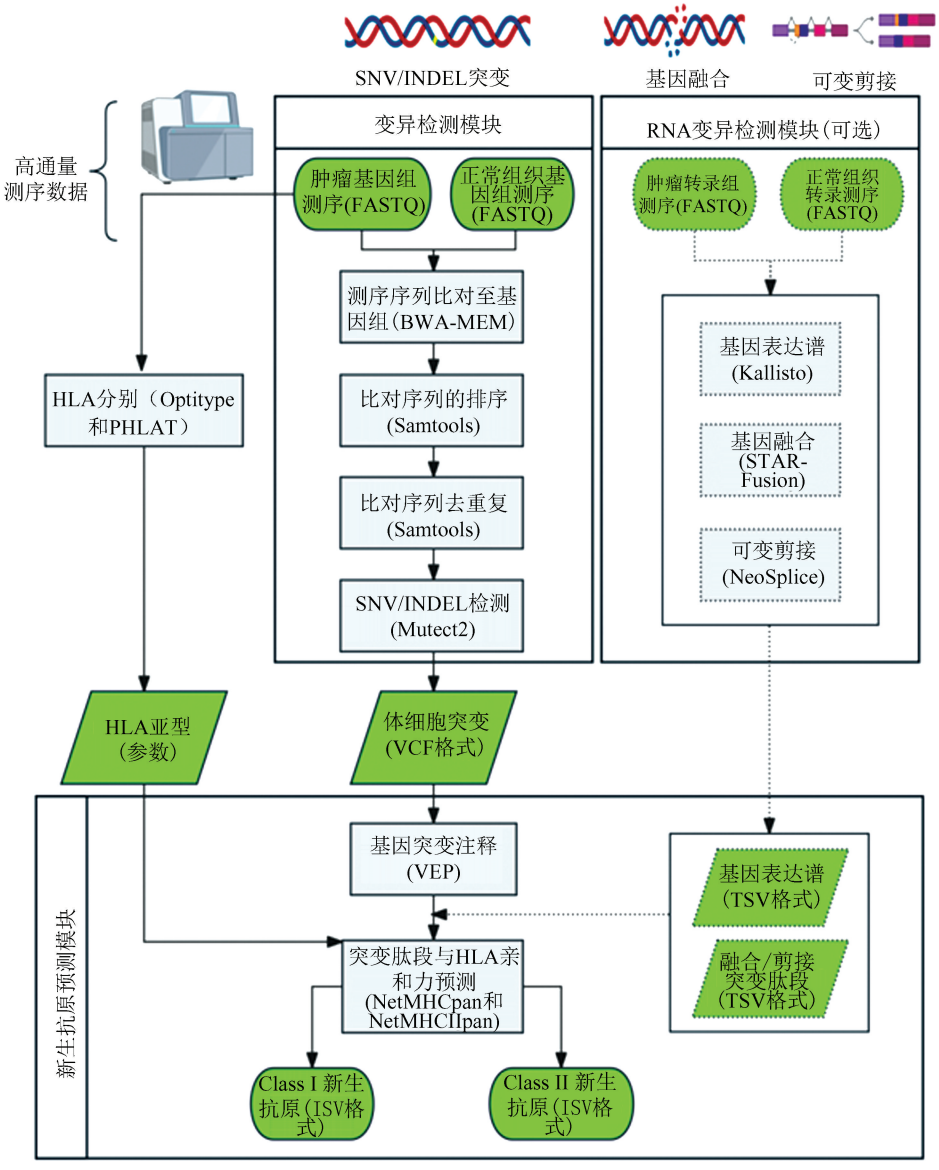

图 1 Neo-Pred 流程设计总览

Fig.1 Overview of Neo-Pred workflow

流程主要包括了三个部分:SNV/InDel 变异检测及其对应的 MHC Class I 和 MHC Class II 的肿瘤新生抗原检测、基因融合变异检测及其对应的 MHC Class I 和 MHC Class II 的肿瘤新生抗原检测以及肿瘤特异性可变剪接检测及其对应的 MHC Class I 和 MHC Class II 的肿瘤新生抗原检测。

流程接收肿瘤及正常组织的高通量测序数据作为输入,同时还需要人类基因组序列及其对应版本的 VEP 注释结果作为参考数据库进行分析。Neo-Pred可接受不同的测序方案:仅提供肿瘤及正常组织的外显子测序文件,或者可以在此基础上再增加肿瘤的 RNA-seq 数据;也可以同时提供肿瘤及正常组织的外显子与 RNA-seq 测序文件。Neo-Pred 在不同的测序方案下会自动判断可执行的模块(表 1)。

表 1　Neo-Pred 支持的测序组合

Table 1　Sequencing combinations supported by Neo-Pred

测序检测方案	肿瘤及正常组织的外显子测序	肿瘤及正常组织的外显子测序与肿瘤的 RNA-seq 测序	肿瘤及正常组织的外显子与 RNA-seq 测序
SNV/InDel 变异检测	是	是	是
基因融合检测	否	否	是
可变剪接检测	否	否	是
变异注释	是	是	是
HLA 分型	是	是	是
新生抗原表达谱分析	否	是	是
新生抗原检测	是(无表达信息)	是	是

1.1　SNV/InDel 变异检测及新生抗原检测

SNV/InDel 的变异检测基于 GATK best practice 进行,获得变异检测的结果后,使用 VEP 软件进行注释。同时,Neo-Pred 还调用 Optitype[10] 和 PHLAT[11] 分别完成 MHC Class I 和 MHC Class II 的二级亚型分型的工作。

在提供 RNA-seq 文件的基础上,还调用 kallisto[12] 程序对肿瘤组织表达量进行分析,可根据表达量水平设定阈值对新生抗原进行筛选。

最后新生抗原的预测部分是调用 pVACtools 来进行,生成含有肿瘤新生抗原关键信息的 tsv 格式文本文件供用户参考。

1.2　基因融合检测及新生抗原检测

基因融合变异检测使用 STAR-Fusion[13] 软件完成,在提供了肿瘤和正常组织 RNA-seq 数据的情况下,分别对肿瘤和正常组织进行变异检测,选择肿瘤特异性的融合来进行新生抗原检测。使用 NetMHCpan-4.0[14] 和 NetMHCIIpan-3.2[15] 来计算融合衍生的突变肽段序列,筛选与 MHC 结合阳性的肽段作为候选肽段。

1.3　可变剪接检测及新生抗原检测

可变剪接来源的新生抗原检测基于 NeoSplice[16] 流程进行。Neo-Pred 在接收到肿瘤和正常组织 RNA-seq 数据时,会自动执行此模块。NeoSplice 基于融合检测步骤的 bam 格式文件分析。使用 NetMHCpan-4.0 和 NetMHCIIpan-3.2 来计算可变剪接衍生的突变肽段序列,筛选与 MHC 结合阳性的肽段作为候选肽段。

2　结果

2.1　变异检测基准数据与检测性能

2005 年美国食品药品监督管理局(FDA)领导成立了微阵列质量控制(MicroArray quality control, MAQC)联盟,旨在解决不断发展的高通量基因组技术的应用相关的再现性问题。其第四个项目,也称为测序质量控制第二阶段(SEQC II)项目提供肿瘤变异检测的基准物质与肿瘤变异检测的最佳实践[17]。这里我们也使用这一项目的基准数据评估 Neo-Pred 变异检测模块的性能(http://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/seqc/Somatic_Mutation_WG)。同时,我们也注意到有使用多 SNV 变异检测工具联合检测的策略[1],在这里,我们也进行了对比 Neo-Pred 单用 Mutect2 的策略以评估使用多个软件联合检测的必要性。使用相同的经过比对、排序、去重复后的 bam 数据,使用 Mutect2、Varscan、Strelka 三个不同软件联合检测,选取两个以上软件支持变异作为最终检测到的变异位点。各样本上的两种变异检测策略检出情况如下表 2 所示。

从结果中可以看出,使用三个常用的肿瘤变异检测软件联合检测的策略 F1 分数对比 Neo-Pred 的单一软件策略,提升相当有限,甚至在 LL_T_1 样本、

NV_T_1 样本和 NV_T_3 样本中出现性能下降的情况,但此步骤需要资源和计算时间却会增加,基于此 Neo-Pred 选择使用单一软件检测的策略搭建变异检测模块。同时,Neo-Pred 也提供使用筛选好的变异信息文件(VCF 格式,与 mutect2 保持一致)作为输入的选择,以灵活应用不同的需求。

表 2 Neo-Pred 变异检测模块在基准数据集上的性能评估

Table 2 Evaluation of the Neo-Pred variant detection module on benchmark datasets

样本	仅 mutect2 检测			mutect2、Varscan 和 Strelka 三个软件联合检测		
	精准度	召回率	F1 分数	精准度	召回率	F1 分数
EA_T_1	0.859 8	0.933 3	0.895 0	0.856 4	0.951 4	0.901 4
FD_T_1	0.753 0	0.867 7	0.806 3	0.747 3	0.876 8	0.806 9
FD_T_2	0.739 3	0.868 0	0.798 5	0.757 5	0.879 0	0.813 7
FD_T_3	0.739 4	0.865 8	0.797 6	0.759 2	0.869 5	0.810 6
LL_T_1	0.805 4	0.850 0	0.827 1	0.767 4	0.847 0	0.805 3
NV_T_1	0.911 8	0.972 5	0.941 2	0.873 5	0.992 9	0.929 4
NV_T_2	0.900 6	0.963 7	0.931 0	0.890 2	0.992 0	0.938 4
NV_T_3	0.911 9	0.975 1	0.942 4	0.876 0	0.992 0	0.930 4

2.2 肿瘤新生抗原基准数据

肿瘤新生抗原筛选联盟(The tumor neoantigen selection alliance,TESLA)提供了一组肿瘤基因组测序的基准数据用于评估新生抗原的检测能力[18](https://www.synapse.org/#!Synapse:syn21048999)。选择其中五例提供抗原免疫原性验证信息的样本数据用于评估 Neo-Pred 的表现。

2.3 Neo-Pred 在基准数据上的性能

尽管 Neo-Pred 的设计提供了灵活地阈值接口,以便于实际使用者调整筛选参数,方便地对接实际肿瘤抗原的治疗策略(例如是否考虑 HLA Class Ⅱ的新生抗原、每个患者筛选肿瘤新生抗原的数量等),但我们仍测试了 Neo-Pred 在默认参数下的性能表现,使用 TESLA 提供的原始测序数据(FASTQ 格式文件)及 TESLA 提供的患者 HLA Class Ⅰ亚型的信息进行测试。由于 TESLA 基准数据中未提供基因融合、可变剪接以及 HLA Class Ⅱ的信息,因此这里仅评估了 Neo-Pred 在 HLA Class Ⅰ的肿瘤新生抗原筛选能力。

筛选过程中对肿瘤新生抗原的表达量和与 HLA Class Ⅰ分子的亲和力设定阈值。表达量要求每百万转录本(Transcripts per million,TPM)≥1,HLA Class Ⅰ分子的亲和力的要求为 IC50 ≤500 nM。

剩余其他参数均使用默认设置。从对 FASTQ 格式文件的处理开始运行完整的 Neo-Pred 流程,抗原筛选完成后,使用 TESLA 官方提供的代码计算 Neo-Pred 的筛选性能(https://github.com/ParkerICI/tesla)。使用 TESLA 提供的其他参与机构的筛选肿瘤新生抗原的 AUPRC 性能表现来横向对比(https://www.cell.com/cms/10.1016/j.cell.2020.09.015/attachment/352a5039-810b-4a9d-923d-d2b5493ee686/mmc5.xlsx),计算 Neo-Pred 在 P1、P2、P16 三例样本上的 AUPRC 性能表现分别为 0.712、0.961、0.856,领先于肿瘤新生抗原筛选联盟其他参与团队(全部参与机构均值为 0.221,其中表现最好的团队均值为 0.540),筛选性能提升 31.5%~221.3%,展示出领先的新生抗原检测能力。与 TESLA 各参与机构的比较参见图 2。

可以看到 Neo-Pred 在基准数据上取得良好的性能,而不同的机构预测表现相差较大。一方面是由于 TESLA 参与团队在构建此基准数据时,有自身工艺的考量,筛选的方案及最终汇报的抗原数均表现出较大的差异[18]。而另一方面,Neo-Pred 则是调用性能表现良好的生物信息工具完成各个关键环节的分析。

同时我们也注意到 P3 和 P12 未能检出目标的阳性肽段。我们从突变肽段预测、突变肽段与 MHC 结合及 RNA 表达三个关键维度分析 Neo-Pred 未能检出的新生抗原情况,汇总如表 3 所示。

从未检出原因分析的统计来看,在默认参数下,MHC 结合预测方法和阈值以及突变表达的分析阈值,是影响检出能力的重要因素,优化 MHC 结合的预测方法和阈值,调整 RNA 表达阳性的阈值及筛选标准是提升流程检测性能的重要方向。

由于缺少基准数据,我们未开展 Neo-Pred 的 HLA Class Ⅱ的新生抗原筛选性能评估,但在实践中,有多项展示出巨大潜力的临床试验均在靶点设计策略中使用了 HLA Class Ⅱ来源的新生抗原[1-3]。并且有证据表明,最终激活的来自 HLA Class Ⅱ的免疫反应占比要远高于 HLA Class Ⅰ[1]。我们也从

这个角度评估了如果考虑 HLA Class Ⅱ的新生抗原,能否改善 P3 和 P12 的预测效果。如果纳入 HLA Class Ⅱ的新生抗原,在 P3 和 P12 中将分别有 3 个和 1 个能将 HLA Class Ⅰ的表位纳入其中。而如果将表达量的阈值放宽到 TPM>0,则能找回全部因为 RNA 表达量阈值筛选掉的抗原肽。这说明在新生抗原筛选中,也许有必要考虑患者实际的基因组数据状态并采用动态的阈值筛选机制。

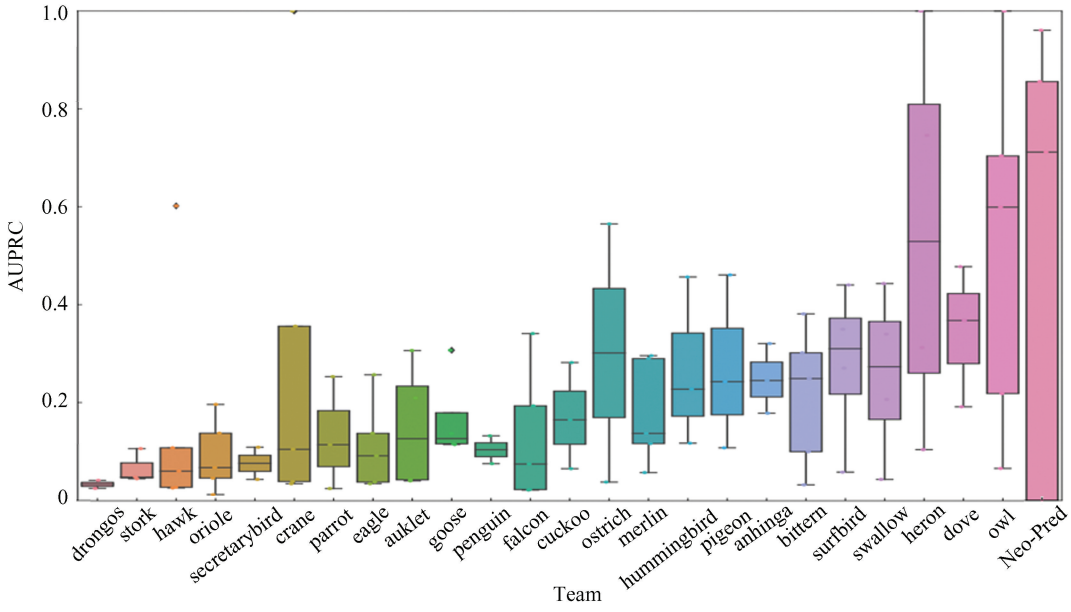

图 2　Neo-Pred 在 TESLA 数据上的测试性能与其它机构的比较

Fig.2　Testing performance of Neo-Pred on the TESLA dataset compared with other institutions

表 3　Neo-Pred 未检出新生抗原原因统计

Table 3　Statistics on the reasons for undetected neoantigens by Neo-Pred

样本	经验证的阳性抗原数量	突变肽段未检出数量(占比/%)	MHC 结合未检出数量(占比/%)	RNA 表达未检出数量(占比/%)
P1	9	1(11.11)	3(33.33)	1(11.11)
P2	4	1(25.00)	0(0)	1(25.00)
P3	13	2(15.38)	5(38.46)	6(46.15)
P12	4	1(25.00)	1(25.00)	2(50.00)
P16	4	0(0)	2(50.00)	0(0)

3　结　论

以肿瘤新生抗原为靶点的个性化肿瘤免疫治疗方法在临床上已经表现出巨大的潜力,这也催生了可用于大规模临床生产的肿瘤新生抗原流程的需求。基于这种需求,我们开发了 Neo-Pred 以满足相关研究及临床试验的开展,并在评估 Neo-Pred 的过程中对新生抗原预测的策略进行了思考与总结:

1)目前也有部分研究使用多种 SNV/InDel 变异检测工具同时检测肿瘤的体细胞突变[19-20],在本研究中,Neo-Pred 的变异检测模块仅用 Mutect2 软件取得了理想的变异检测性能。基于计算速度与资源的考量,未采取多款软件同时检测的策略。研究者可根据自己的试验设计和工艺需求,考虑使用多软件联合检测 SNV/InDel 变异的必要性。

2)在评估中我们发现,同时采用 HLA Class Ⅰ和 HLA Class Ⅱ新生抗原预测的策略,不仅有望激发 $CD4^+$的免疫反应,提升在临床应用的价值,同时这种策略还能改善 HLA Class Ⅰ新生抗原的预测效果,这为应用 Neo-Pred 开发肿瘤疫苗等基于新生抗原的治疗产品提供了良好的参考。

3)Neo-Pred 当前仅针对两类新生抗原(MHC Class Ⅰ和 Class Ⅱ)实现预测功能。用户可根据具体的治疗目标(例如所需新生抗原的种类、数量及组合方式),通过灵活调整筛选参数(如 MHC 结合亲和力阈值、表达丰度阈值等),定制适配个体需求的抗原筛选方案。

参考文献(References)

[1]BRAUN D A, MORANZONI G, CHEA V, et al. A neoantigen vaccine generates antitumour immunity in renal cell carcinoma[J]. Nature, 2025, 639(8054):474-482. DOI: 10.1038/s41586-024-08507-5.

[2]LOPEZ J, POWLES T, BRAITEH F, et al. Autogene cevumeran with or without atezolizumab in advanced solid tumors: A phase 1 trial[J]. Nature Medicine, 2025,31:152-164. DOI: 10.1038/s41591-024-03334-7.

[3]ROJAS L A, SETHNA Z, SOARES K C, et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer[J]. Nature,2023, 618(7963):144-150. DOI: 10.1038/s41586-023-06063-y.

[4]CHUWDHURY G S, GUO Yunshan, CHIANG C L, et al. ImmuneMirror: A machine learning-based integrative pipeline and web server for neoantigen prediction[J]. Briefings in Bioinformatics, 2024,25(2): bbae024. DOI: 10.1093/bib/bbae024.

[5]LI Yuyu, WANG Guangzhi, TAN Xiaoxiu, et al. ProGeoneo: A customized proteogenomic workflow for neoantigen prediction and selection[J]. BMC Medical Genomics, 2020, 13(Suppl 5):52. DOI: 10.1186/s12920-020-0683-4.

[6]YE Yilin, SHEN Yiming, WANG Jian, et al. SIGANEO: Similarity network with GAN enhancement for immunogenic neoepitope prediction[J]. Computational and Structural Biotechnology Journal, 2023, 21:5538-5543. DOI: 10.1016/j.csbj.2023.10.050.

[7]HUNDAL J, KIWALA S, MCMICHAEL J, et al. pVACtools: A computational toolkit to identify and visualize cancer neoantigens[J]. Cancer Immunology Research, 2020, 8(3):409-420. DOI: 10.1158/2326-6066.

[8]SCHENCK R O, LAKATOS E, GATENBEE C, et al. NeoPredPipe: High-throughput neoantigen prediction and recognition potential pipeline[J]. BMC Bioinformatics,2019,20:264. DOI: 10.1186/s12859-019-2876-4.

[9]SCHÄFER R A, GUO Qingxiang, YANG Rrendong. ScanNeo2: A comprehensive workflow for neoantigen detection and immunogenicity prediction from diverse genomic and transcriptomic alterations[J]. Bioinformatics, 2023, 39(11): btad659. DOI: 10.1093/bioinformatics/btad659.

[10]SZOLEK A, SCHUBERT B, MOHR C, et al. OptiType: Precision HLA typing from next-generation sequencing data[J]. Bioinformatics,2014, 30(23):3310-3316. DOI: 10.1093/bioinformatics/btu548.

[11]BAI Yu, WANG D, FURY W. PHLAT: Inference of high-resolution HLA types from RNA and whole exome sequencing[J]. HLA Typing: Methods and Protocols, 2018, 1802:193-201. DOI: 10.1007/978-1-4939-8546-3_13.

[12]BRAY N L, PIMENTEL H, MELSTED P, et al. Near-optimal probabilistic RNA-seq quantification[J]. Nature Biotechnology, 2016, 34(5):525-527. DOI: 10.1038/nbt.3519.

[13]HAAS B J, DOBIN A, LI B, et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods[J].Genome Biology,2019,20:213.DOI: 10.1186/s13059-019-1842-9.

[14]JURTZ V, PAUL S, ANDREATTA M, et al. NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data[J]. The Journal of Immunology, 2017, 199(9):3360-3368. DOI: 10.4049/jimmunol.1700893.

[15]JENSEN K K, ANDREATTA M, MARCATILI P, et al. Improved methods for predicting peptide binding affinity to MHC class II molecules[J]. Immunology, 2018, 154(3):394-406. DOI: 10.1111/imm.12889.

[16]CHAI Shengjie, SMITH C C, KOCHAR T K, et al. NeoSplice: A bioinformatics method for prediction of splice variant neoantigens[J]. Bioinformatics Advances, 2022, 2(1): vbac032. DOI: 10.1093/bioadv/vbac032.

[17]XIAO Wenming, REN Luyao, CHEN Zhong, et al. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing[J]. Nature Biotechnology, 2021,39(9):1141-1150. DOI: 10.1038/s41587-021-00994-5.

[18]WELLS D K, VAN BUUREN M M, DANG K K, et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction[J]. Cell, 2020, 183(3):818-834.e13. DOI: 10.1016/j.cell.2020.09.015.

[19]RIEDER D, FOTAKIS G, AUSSERHOFER M, et al. nextNEOpi: A comprehensive pipeline for computational neoantigen prediction[J]. Bioinformatics, 2022, 38(4):1131-1132. DOI: 10.1093/bioinformatics/btab759.

[20]VENSKO S P, OLSEN K, BORTONE D, et al. LENS: Landscape of effective neoantigens software[J]. Bioinformatics, 2023, 39(6): btad322. DOI: 10.1093/bioinformatics/btad322.