Proteogenomics Study for Tumor Neoantigen Prediction and Identification

谢鹭

上海生物信息技术研究中心/上海市生物医药技术研究院
(Shanghai Institute for Biomedical and Pharmaceutical Technologies，SIBPT)

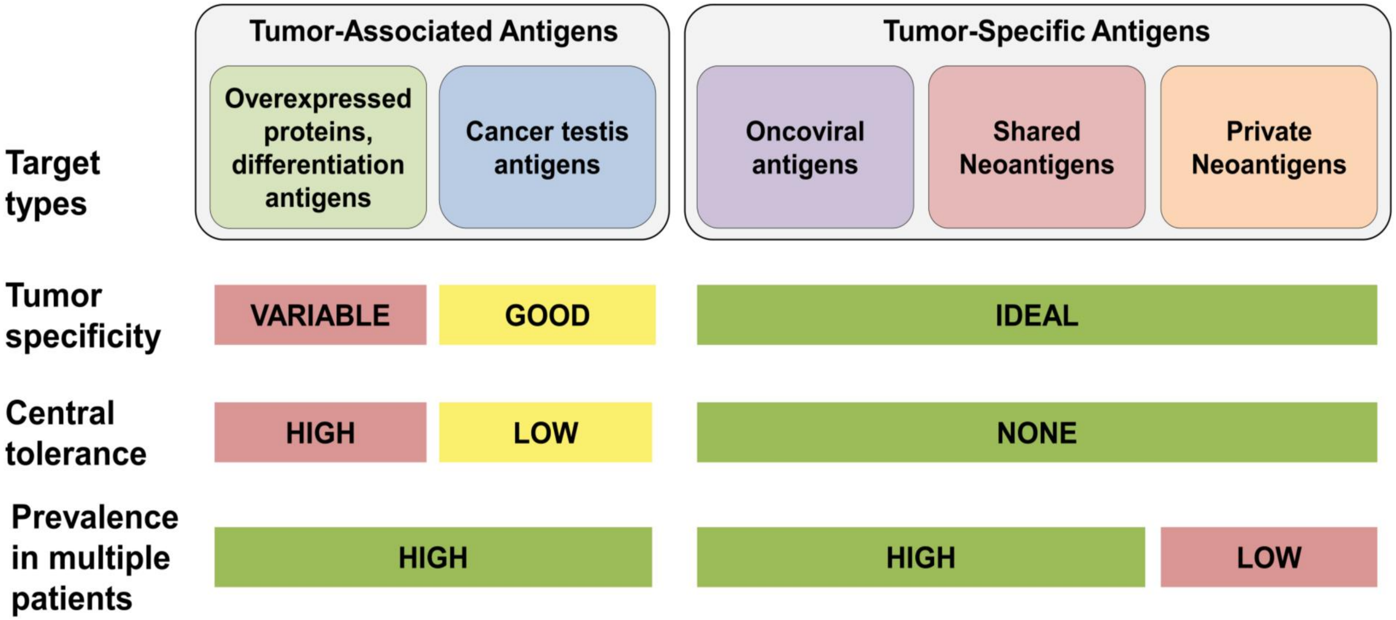
生物信息应用医学研究组（IAM）

2021.8.26

The 6th China Workshop on Computational Proteomics(CNCP 2021)，
2021.8.25-8.26
腾讯会议 763 2900 0243

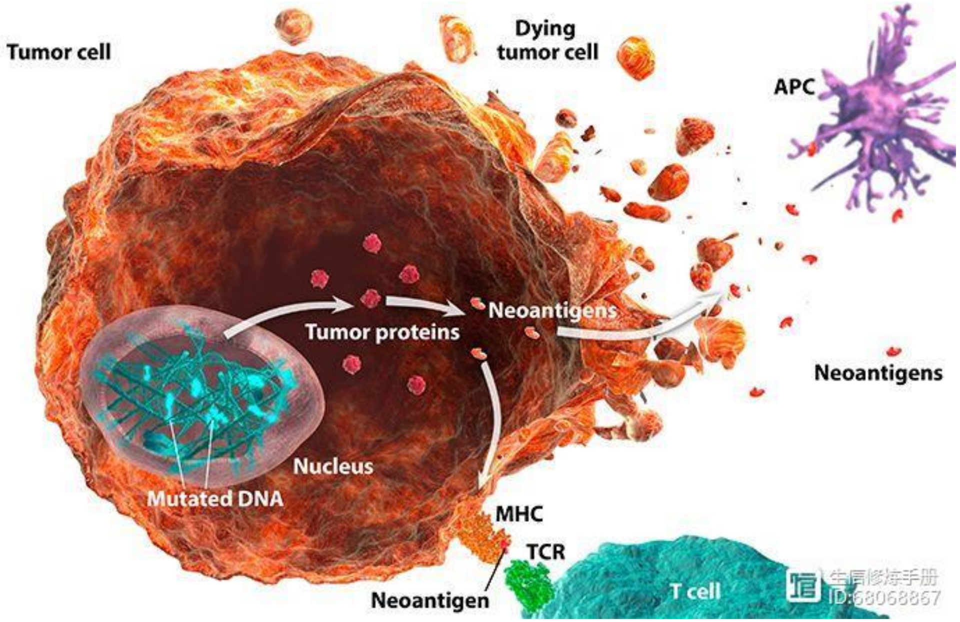
肿瘤抗原

抗原的选择是癌症疫苗设计中最重要的一个组成部分



	Tumor-Associated Antigens		Tumor-Specific Antigens		
Target types	Overexpressed proteins, differentiation antigens	Cancer testis antigens	Oncoviral antigens	Shared Neoantigens	Private Neoantigens
Tumor specificity	VARIABLE	GOOD	IDEAL		
Central tolerance	HIGH	LOW	NONE		
Prevalence in multiple patients	HIGH		HIGH		LOW

Hollingsworth RE, et al. Turning the corner on therapeutic cancer vaccines. 2019,NPJ vaccines



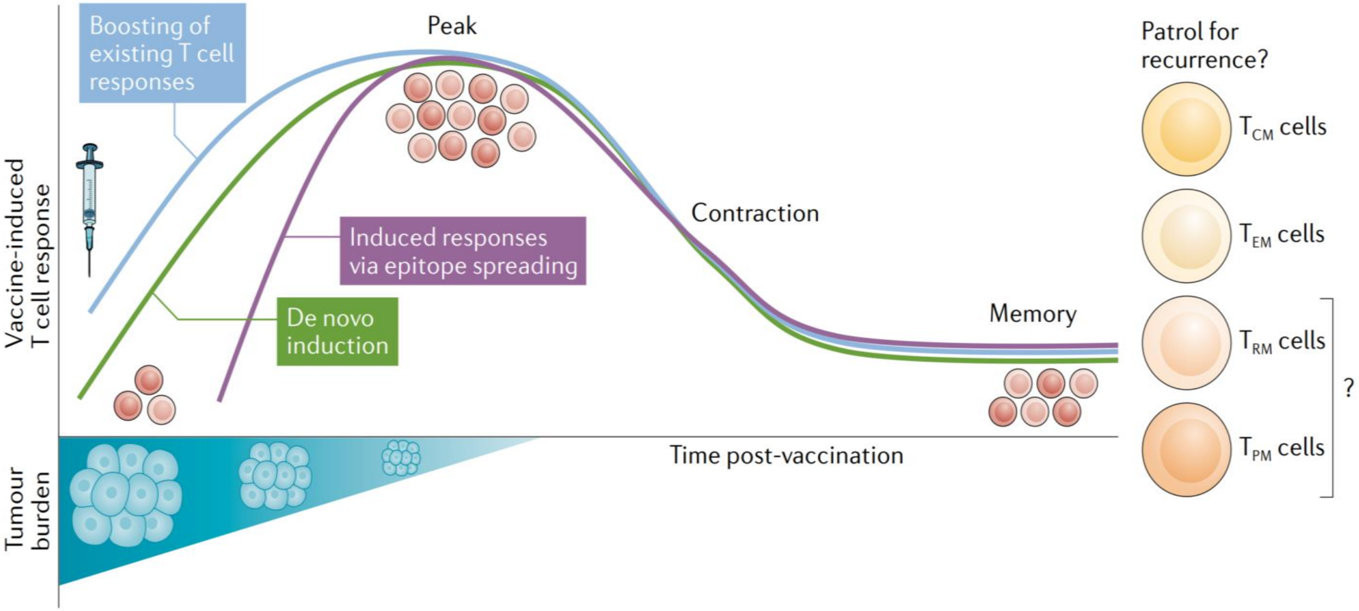
肿瘤特异性抗原 （tumor- specificantigen，TSAs） 又称**新抗原**（Neoantigen)是由肿瘤细胞突变基因编码的新生抗原，主要由基因点突变、删除/插入突变、基因融合等产生的与正常细胞表达的蛋白不一样的新的异常蛋白。

基于TSAs（新抗原）而不是传统使用的 TAAs 的疫苗有几个优点

新抗原仅由肿瘤细胞表达，可以引发真正的肿瘤特异性 T 细胞反应

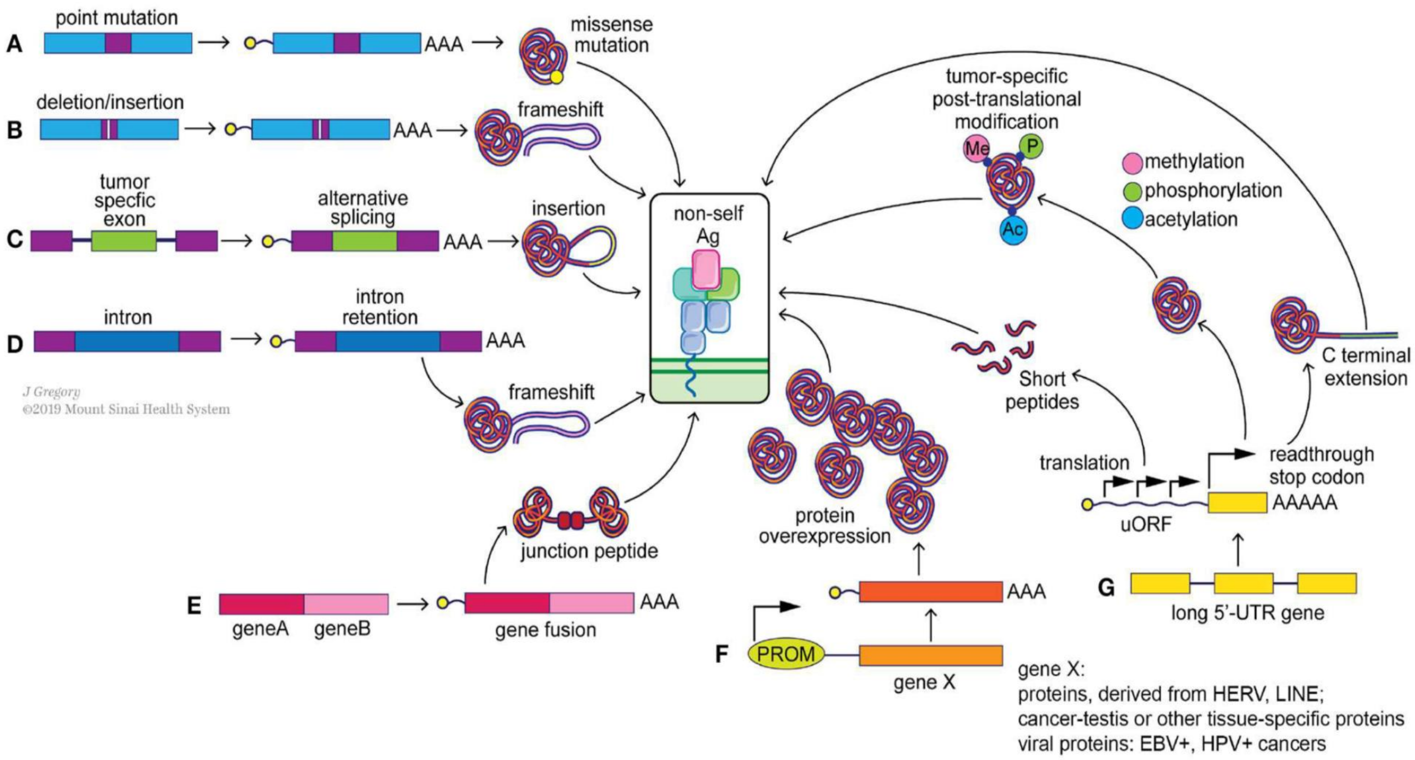
绕过自身表位的 T 细胞中心耐受性，从而有效诱导对肿瘤的免疫反应

个性化新抗原的疫苗可能诱导持久的肿瘤特异性记忆 T 细胞群



Blass E and Ott PA, Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. 2021，Nature Reviews|Clinical Oncology

“非自身”肿瘤新抗原的潜在来源



新抗原来自多种遗传和表观遗传畸变，其充分表征来源是体细胞错义和插入缺失突变，或基因组重排，例如可变剪切、基因融合。由于与人类编码基因组中的序列缺乏相似性，移码新抗原可能被证明比错义突变产生抗原更具免疫原性，目前正在积极研究中 。源自基因融合的新抗原最近通过了免疫原性测试，并且在突变负荷低时可能具有特殊意义。

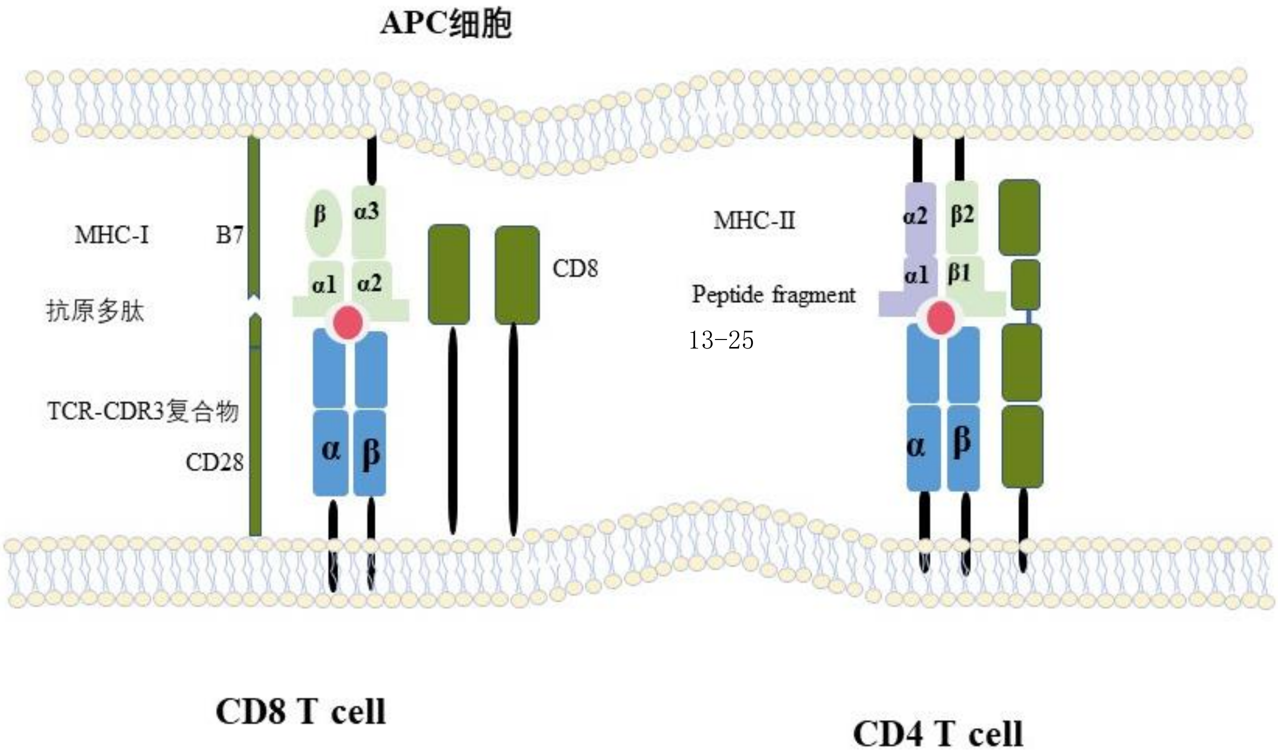
Roudko V et al, Computational Prediction and Validation of Tumor-Associated Neoantigens. 2020, Frontiers in Immunology

具有免疫原性的新抗原至少符合两个标准：

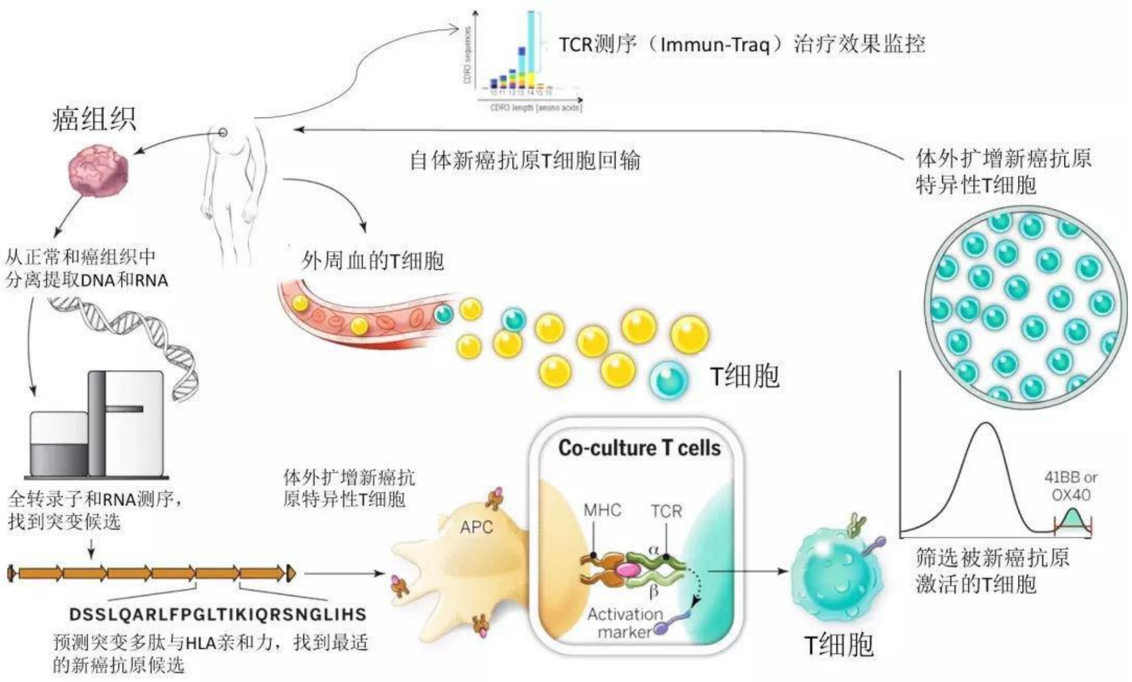
1）**被MHC提呈**到肿瘤细胞表面；

2）可形成的p-MHC复合物**被T细胞受体识别**

这两个标准都离不开组织相容性受体（MHC，在人中又叫HLA）的结合。分为I型和II型。HLA-I型分子结合长约8-11氨基酸的内源性抗原并把它们提呈给细胞毒性CD8$^+$ T细胞；HLA-II型分子则结合长约13-25氨基酸的外源性抗原并把它们提呈给辅助性CD4$^+$ T细胞。



个性化癌症疫苗治疗



新一代测序技术极大推动了新抗原疫苗的可行性，但从测序识别肿瘤的体细胞突变到T细胞受体（T cell receptor ,TCR）识别新抗原产生免疫反应，这一过程中存在大量的候选假阳性新抗原多肽，这对于个性化癌症疫苗的制备无疑是难以跨越的障碍。一套**有效合理的筛选新抗原多肽的方法**在新抗原疫苗制备过程中就显得尤为重要

PART ONE

一

研究内容

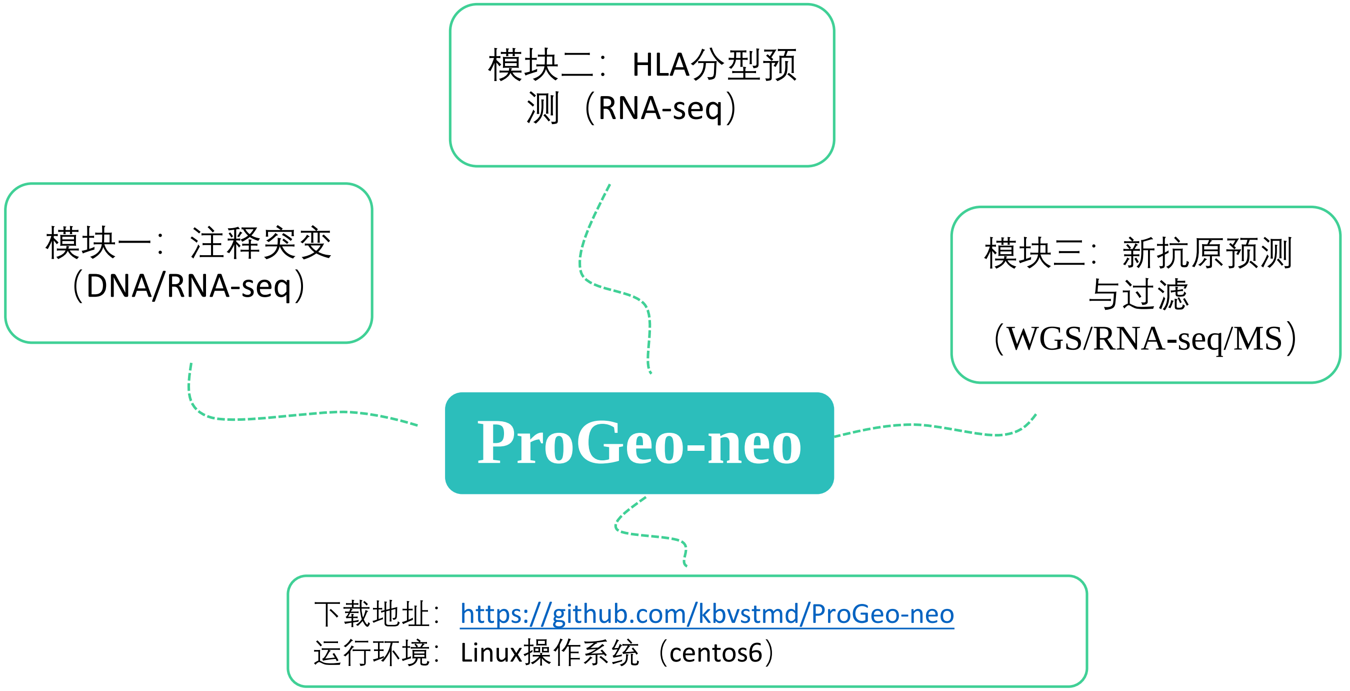
ProGeo-neo

基于蛋白质基因组学的新抗原鉴定和筛选流程

目前国内外所开发的大多数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细胞受体仅对其中非常小部分（约1%）预测的新抗原有反应。

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，也可以直接鉴定或验证人类肿瘤组织来源的新抗原。

基于蛋白质基因组学方法预测肿瘤新抗原



模块二：HLA分型预测（RNA-seq）

模块一：注释突变（DNA/RNA-seq）

模块三：新抗原预测与过滤（WGS/RNA-seq/MS）

ProGeo-neo

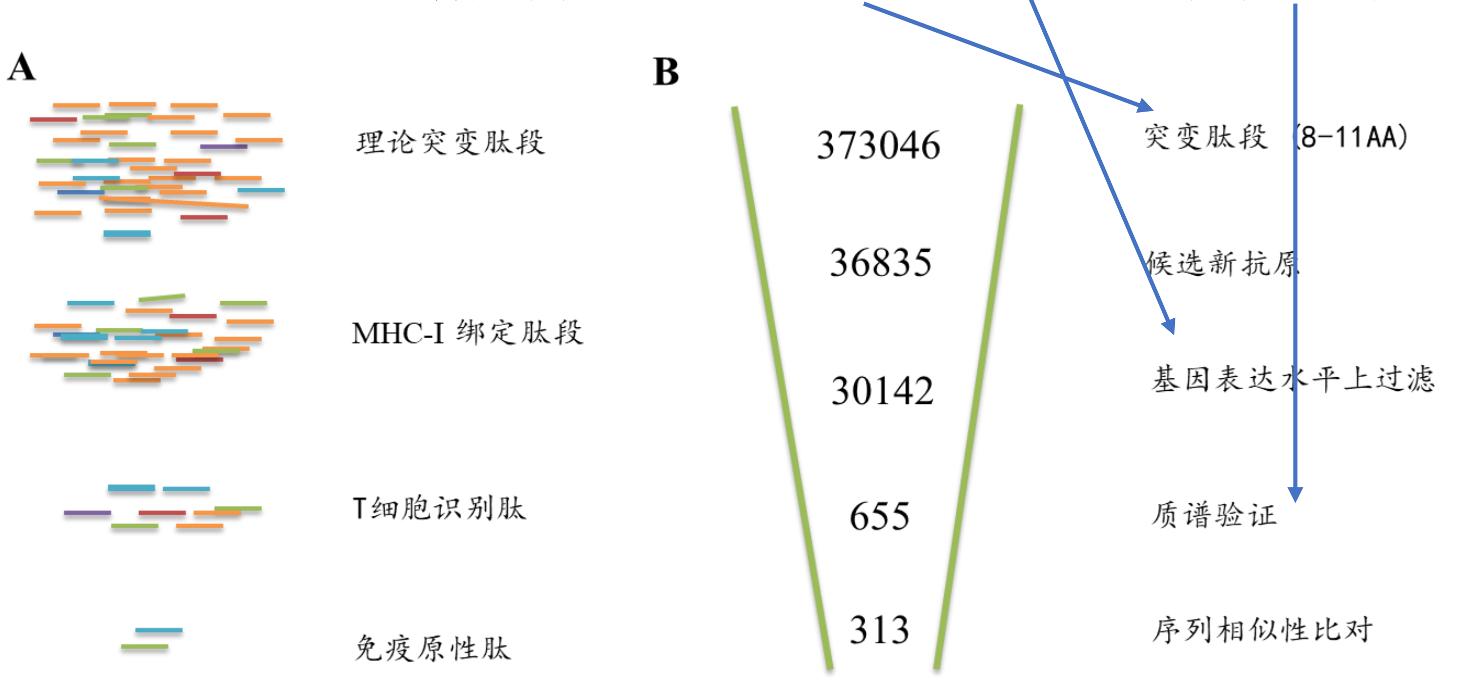
下载地址：https://github.com/kbvstmd/ProGeo-neo
运行环境：Linux操作系统（centos6）

李雨雨,王广志,陈兰明,谢鹭.基于蛋白质基因组学方法的新抗原鉴定流程[J].生物化学与生物物理进展,2019

Yuyu Li et al. ProGeo-neo: a customized proteogenomic workflow for neoantigen prediction and selection. BMC Med Genomics, 2020

基于蛋白质基因组学方法预测肿瘤新抗原

基于Jurkat白血病细胞系数据预测的新抗原在基因组、转录组以及蛋白质组水平上的鉴定



A. 基因突变产生的肽段只有一小部分被提呈到肿瘤细胞表面并激活CD8+T细胞 B. 本研究中利用蛋白质基因组学方法预测新抗原的数量结果概括

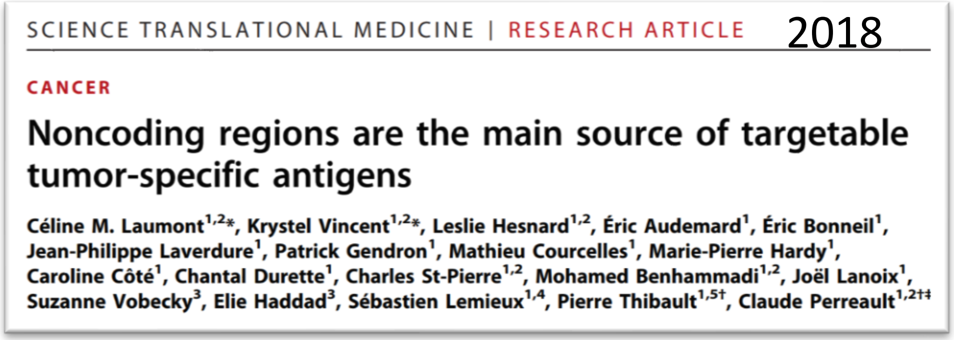
新抗原与非编码区的研究

外显子组(所有蛋白质编码序列)仅占人类基因组的2%，而多达75%的基因组可以转录和潜在翻译。(*Sarah Djebali, et al. Nature, 2012.*)

99%的肿瘤突变位于非编码区，因此将外显子组作为新抗原的唯一来源是非常局限性的。(*Ekta Khurana, et al*. Nat Rev Genet, 2016.)

随后有研究进一步证明，非编码产生的肽段能与MHC分子结合，而且可以作为T细胞的靶点。 (*Céline M Laumont, et al*. Nat Commun, 2016; *Céline M Laumont , et al*. Cell Mol Life Sci, 2018; *Maria J L Kracht, et al*. Nat Med. 2017.)

新抗原与非编码区的研究



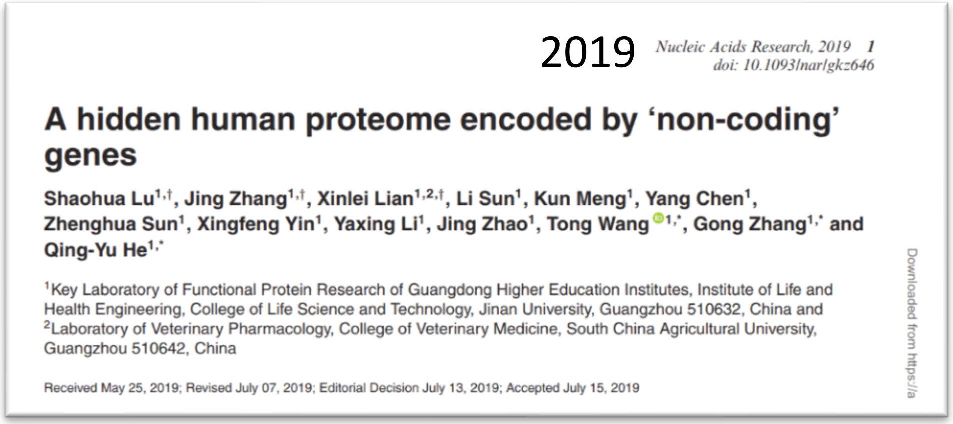
SCIENCE TRANSLATIONAL MEDICINE | RESEARCH ARTICLE 2018

CANCER

Noncoding regions are the main source of targetable tumor-specific antigens

Céline M. Laumont[1,2]*, Krystel Vincent[1,2]*, Leslie Hesnard[1,2], Éric Audemard[1], Éric Bonneil[1], Jean-Philippe Laverdure[1], Patrick Gendron[1], Mathieu Courcelles[1], Marie-Pierre Hardy[1], Caroline Côté[1], Chantal Durette[1], Charles St-Pierre[1,2], Mohamed Benhammadi[1,2], Joël Lanoix[1], Suzanne Vobecky[3], Elie Haddad[3], Sébastien Lemieux[1,4], Pierre Thibault[1,5]†, Claude Perreault[1,2]†‡

90%的TSA来自非编码区域，标准的外显子方法可能会遗漏



2019 *Nucleic Acids Research, 2019 1*
doi: 10.1093/nar/gkz646

A hidden human proteome encoded by 'non-coding' genes

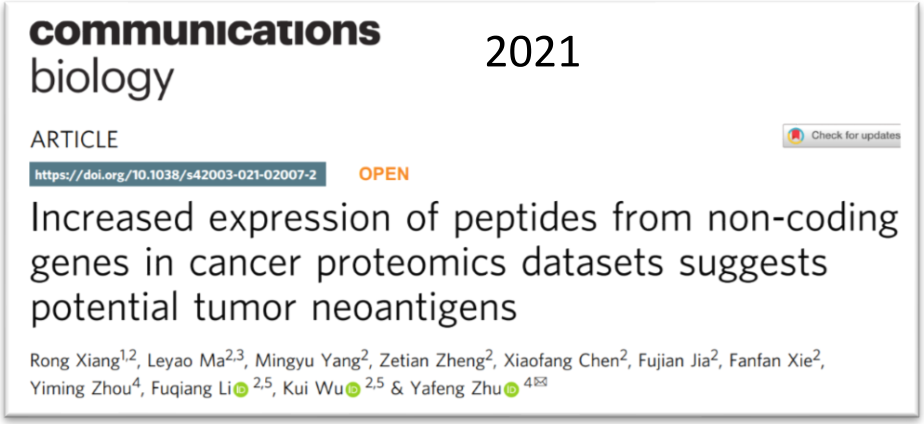
Shaohua Lu[1,†], Jing Zhang[1,†], Xinlei Lian[1,2,†], Li Sun[1], Kun Meng[1], Yang Chen[1], Zhenghua Sun[1], Xingfeng Yin[1], Yaxing Li[1], Jing Zhao[1], Tong Wang[1,*], Gong Zhang[1,*] and Qing-Yu He[1,*]

[1]Key Laboratory of Functional Protein Research of Guangdong Higher Education Institutes, Institute of Life and Health Engineering, College of Life Science and Technology, Jinan University, Guangzhou 510632, China and [2]Laboratory of Veterinary Pharmacology, College of Veterinary Medicine, South China Agricultural University, Guangzhou 510642, China

Received May 25, 2019; Revised July 07, 2019; Editorial Decision July 13, 2019; Accepted July 15, 2019

308个lncRNA编码的新蛋白质

从肿瘤的非编码区域，预测了大量的新抗原，这可能为新抗原提供了除标准肿瘤特异性突变以外的新来源



communications biology 2021

ARTICLE Check for updates

https://doi.org/10.1038/s42003-021-02007-2 OPEN

Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens

Rong Xiang[1,2], Leyao Ma[2,3], Mingyu Yang[2], Zetian Zheng[2], Xiaofang Chen[2], Fujian Jia[2], Fanfan Xie[2], Yiming Zhou[4], Fuqiang Li[2,5], Kui Wu[2,5] & Yafeng Zhu[4]✉

PART TWO

研究内容

Ineo-EPP

抗原及新抗原肽免疫原性预测算法

肽免疫原性预测算法

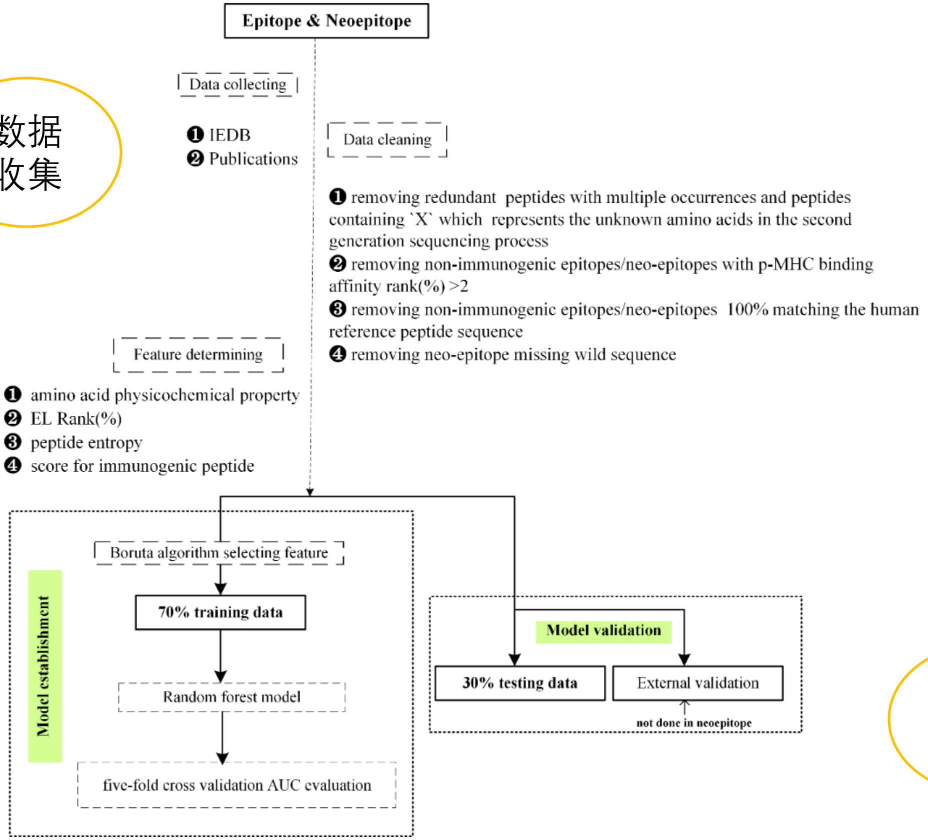
收集实验验证的真实表位数据探寻**免疫原性**与**非免疫原性**表位差异，促进抗原及新抗原表位免疫原性识别

数据收集

数据清洗

特征筛选

构建模型


Epitope & Neoepitope

Data collecting

❶ IEDB
❷ Publications

Data cleaning

❶ removing redundant peptides with multiple occurrences and peptides containing `X` which represents the unknown amino acids in the second generation sequencing process
❷ removing non-immunogenic epitopes/neo-epitopes with p-MHC binding affinity rank(%) >2
❸ removing non-immunogenic epitopes/neo-epitopes 100% matching the human reference peptide sequence
❹ removing neo-epitope missing wild sequence

Feature determining

❶ amino acid physicochemical property
❷ EL Rank(%)
❸ peptide entropy
❹ score for immunogenic peptide

Model establishment

Boruta algorithm selecting feature

70% training data

Random forest model

five-fold cross validation AUC evaluation

Model validation

30% testing data

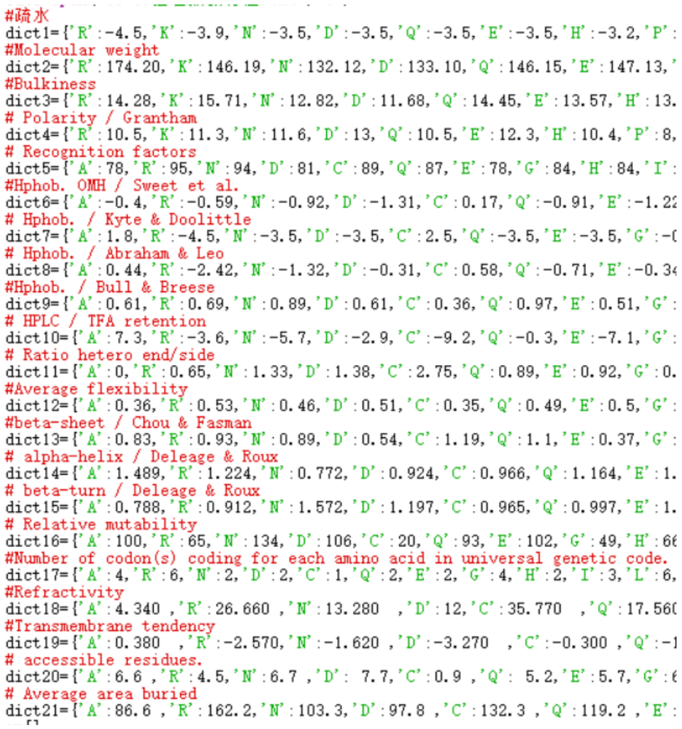
External validation

not done in neoepitope

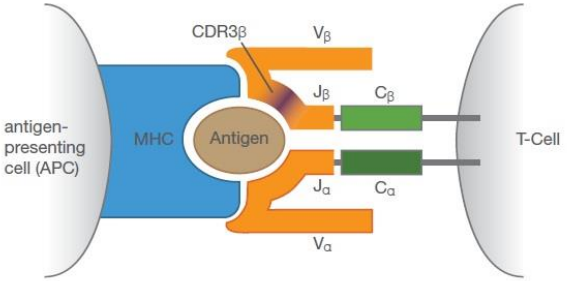
肽免疫原性预测算法

表位序列氨基酸理化性质特征构建

氨基酸理化性质主来源： SIB ExPASy Bioinformatics Resource Portal

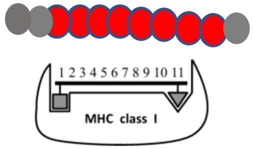


```
#疏水
dict1={'R':-4.5,'K':-3.9,'N':-3.5,'D':-3.5,'Q':-3.5,'E':-3.5,'H':-3.2,'P':
#Molecular weight
dict2={'R':174.20,'K':146.19,'N':132.12,'D':133.10,'Q':146.15,'E':147.13,'
#Bulkiness
dict3={'R':14.28,'K':15.71,'N':12.82,'D':11.68,'Q':14.45,'E':13.57,'H':13.
# Polarity / Grantham
dict4={'R':10.5,'K':11.3,'N':11.6,'D':13,'Q':10.5,'E':12.3,'H':10.4,'P':8,
# Recognition factors
dict5={'A':78,'R':95,'N':94,'D':81,'C':89,'Q':87,'E':78,'G':84,'H':84,'I':
#Hphob. OMH / Sweet et al.
dict6={'A':-0.4,'R':-0.59,'N':-0.92,'D':-1.31,'C':0.17,'Q':-0.91,'E':-1.2
# Hphob. / Kyte & Doolittle
dict7={'A':1.8,'R':-4.5,'N':-3.5,'D':-3.5,'C':2.5,'Q':-3.5,'E':-3.5,'G':-
# Hphob. / Abraham & Leo
dict8={'A':0.44,'R':-2.42,'N':-1.32,'D':-0.31,'C':0.58,'Q':-0.71,'E':-0.3
#Hphob. / Bull & Breese
dict9={'A':0.61,'R':0.69,'N':0.89,'D':0.61,'C':0.36,'Q':0.97,'E':0.51,'G':
# HPLC / TFA retention
dict10={'A':7.3,'R':-3.6,'N':-5.7,'D':-2.9,'C':-9.2,'Q':-0.3,'E':-7.1,'G':
# Ratio hetero end/side
dict11={'A':0,'R':0.65,'N':1.33,'D':1.38,'C':2.75,'Q':0.89,'E':0.92,'G':0.
#Average flexibility
dict12={'A':0.36,'R':0.53,'N':0.46,'D':0.51,'C':0.35,'Q':0.49,'E':0.5,'G':
#beta-sheet / Chou & Fasman
dict13={'A':0.83,'R':0.93,'N':0.89,'D':0.54,'C':1.19,'Q':1.1,'E':0.37,'G':
# alpha-helix / Deleage & Roux
dict14={'A':1.489,'R':1.224,'N':0.772,'D':0.924,'C':0.966,'Q':1.164,'E':1.
# beta-turn / Deleage & Roux
dict15={'A':0.788,'R':0.912,'N':1.572,'D':1.197,'C':0.965,'Q':0.997,'E':1.
# Relative mutability
dict16={'A':100,'R':65,'N':134,'D':106,'C':20,'Q':93,'E':102,'G':49,'H':6
#Number of codon(s) coding for each amino acid in universal genetic code.
dict17={'A':4,'R':6,'N':2,'D':2,'C':1,'Q':2,'E':2,'G':4,'H':2,'I':3,'L':6,
#Refractivity
dict18={'A':4.340 ,'R':26.660 ,'N':13.280  ,'D':12,'C':35.770  ,'Q':17.56
#Transmembrane tendency
dict19={'A':0.380  ,'R':-2.570,'N':-1.620 ,'D':-3.270  ,'C':-0.300 ,'Q':-1
# accessible residues.
dict20={'A':6.6 ,'R':4.5,'N':6.7 ,'D': 7.7,'C':0.9 ,'Q': 5.2,'E':5.7,'G':6
# Average area buried
dict21={'A':86.6 ,'R':162.2,'N':103.3,'D':97.8 ,'C':132.3 ,'Q':119.2 ,'E':
```



Characteristics calculation of peptides based on amino acid sequences. The formula for calculating peptide characteristics is shown in (1). P_N, P_2, P_C (N-terminal, position 2, C-terminal as anchored sites by default) are considered to be embedded in HLA molecules and no contact with TCRs, therefore not evaluated.

$$P_c = \{ \sum_{x\in Pos(P)}^{x\notin(N,2,C)} P_{A_c} \} \Big/ (len(P)-3) \qquad (1)$$

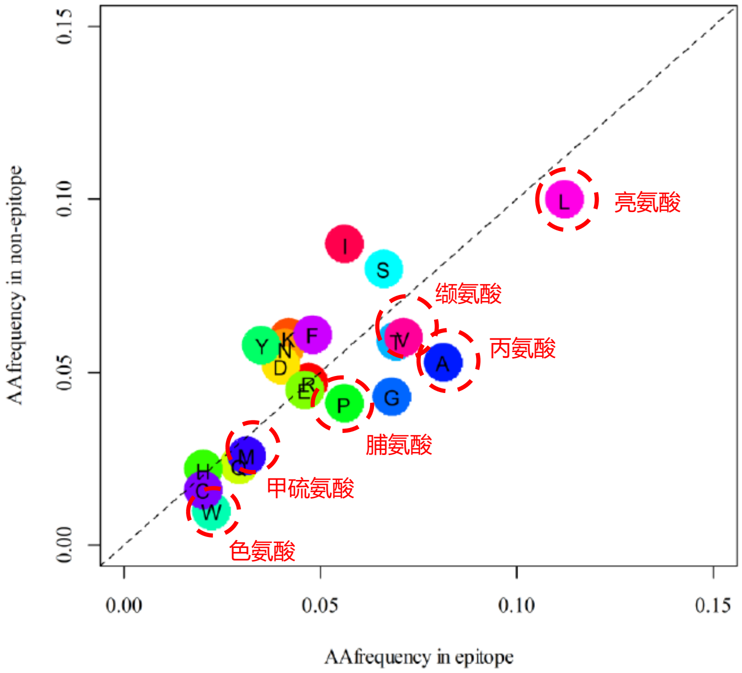


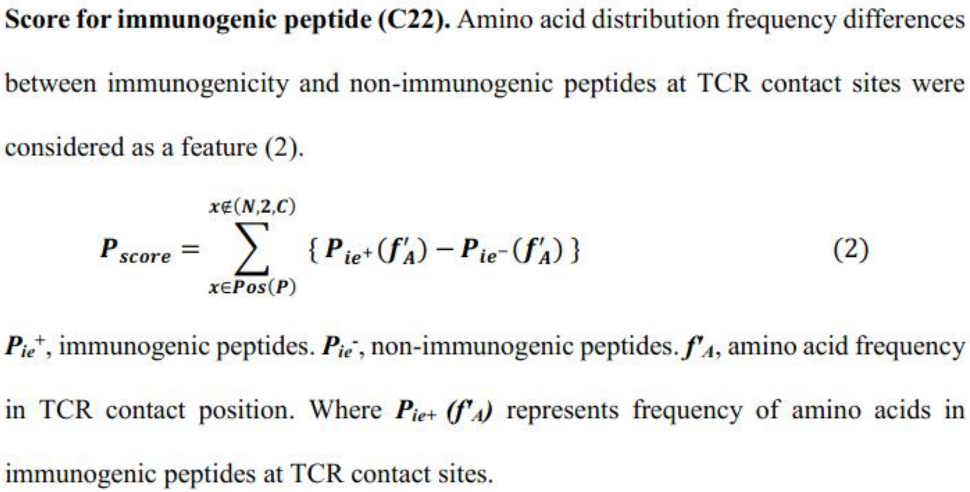
P, peptide. ***c***, characteristic. Where P_c represents characteristics of peptides. ***A***, amino acid. ***N***, N-terminal in a peptide. ***C***, C-terminal in a peptide. Pos, amino acid position in peptide. Where P_{Ac} represents characteristics of amino acids in peptides.

序列转化为氨基酸性质值的叠加

肽免疫原性预测算法

免疫原性多肽氨基酸偏好频率打分





Score for immunogenic peptide (C22). Amino acid distribution frequency differences between immunogenicity and non-immunogenic peptides at TCR contact sites were considered as a feature (2).

$$P_{score} = \sum_{x\in Pos(P)}^{x\notin(N,2,C)} \{ P_{ie^+}(f'_A) - P_{ie^-}(f'_A) \} \qquad (2)$$

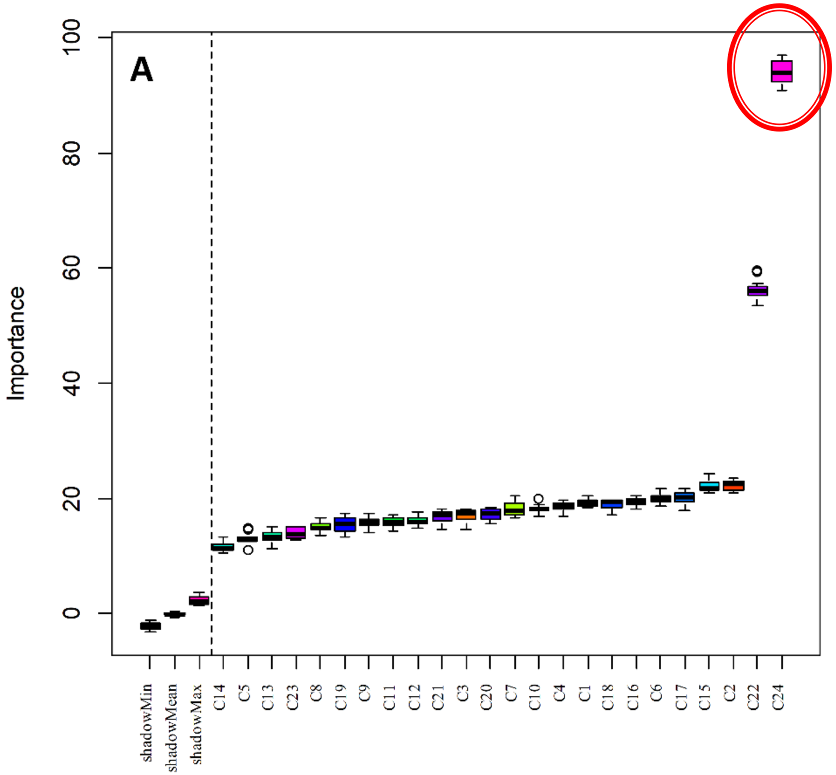
P_{ie}^{+}, immunogenic peptides. P_{ie}^{-}, non-immunogenic peptides. f'_A, amino acid frequency in TCR contact position. Where P_{ie+} (f'_A) represents frequency of amino acids in immunogenic peptides at TCR contact sites.

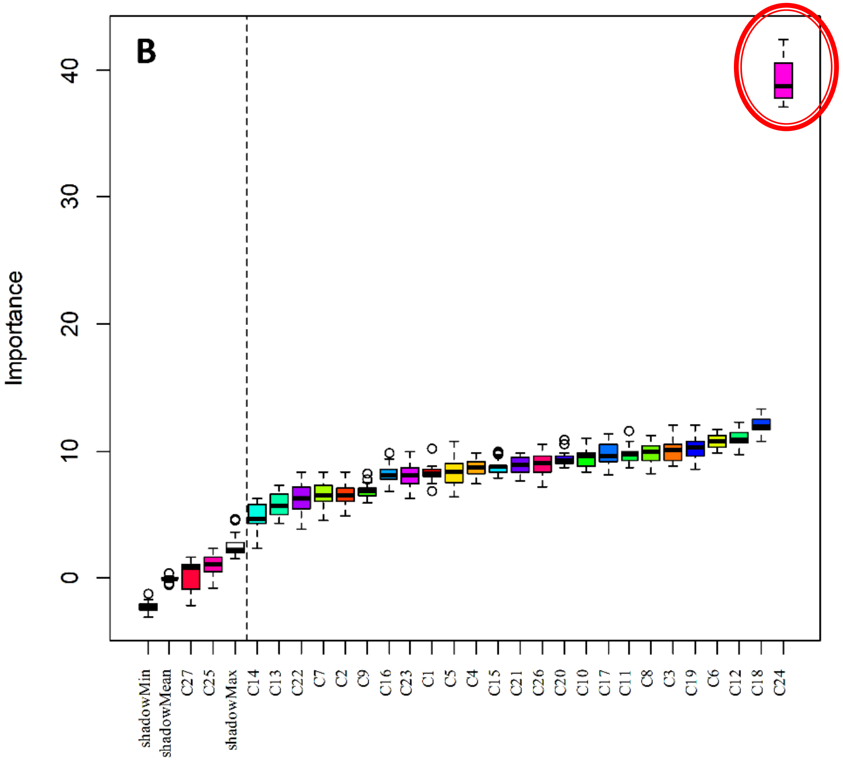
肽免疫原性预测算法

Boruta算法筛选特征

通过对信息系统中重要和不重要的属性进行无偏和稳定的选择，从而找到所有的相关变量。它迭代地删除被统计测试证明不如随机探测相关的特性。

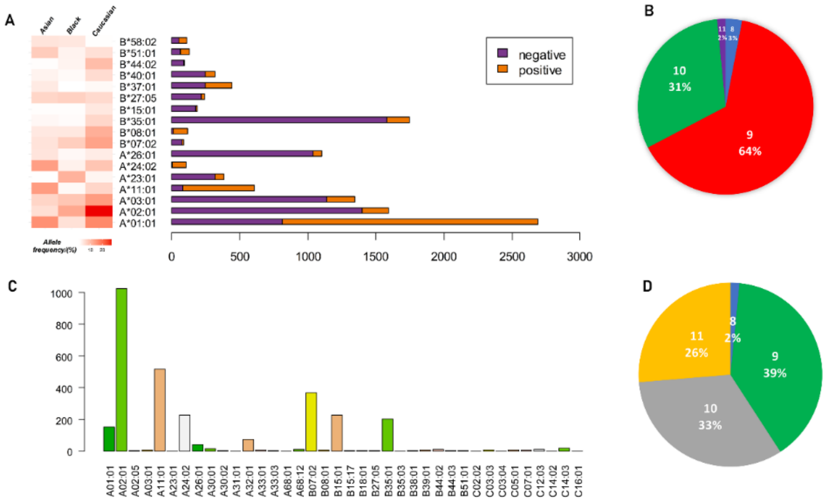
%rank值



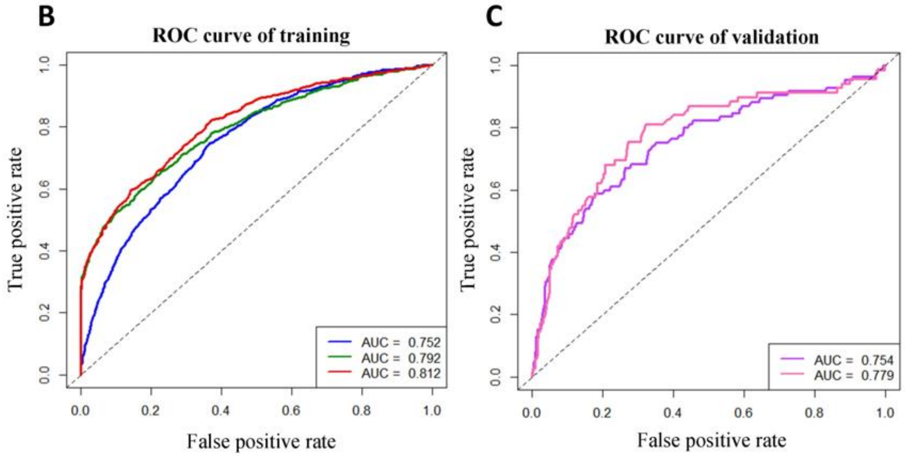


%rank值

肽免疫原性预测算法



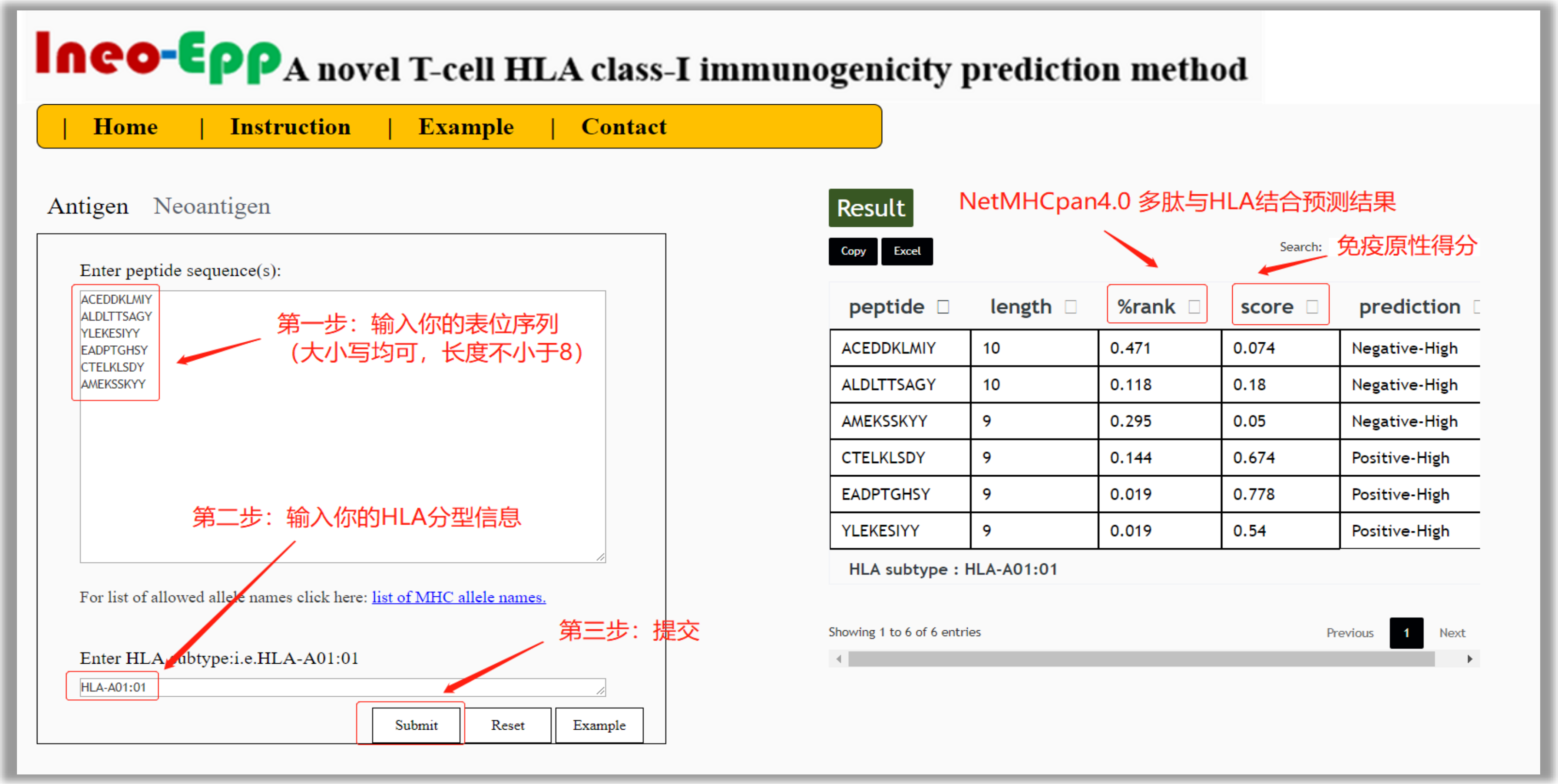
人类相关的免疫研究：肽段数量大于10000条



平均最优AUC=0.81，外部验证最优AUC=0.77
（之前的研究平均AUC=0.65，外部验证AUC=0.62）

Ineo-Epp

http://www.biostatistics.online/INeo-Epp/neoantigen.php



Ineo-Epp A novel T-cell HLA class-I immunogenicity prediction method

Enter peptide sequence(s):

ACEDDKLMIY
ALDLTTSAGY
YLEKESIYY
EADPTGHSY
CTELKLSDY
AMEKSSKYY

第一步：输入你的表位序列（大小写均可，长度不小于8）

第二步：输入你的HLA分型信息

For list of allowed allele names click here: list of MHC allele names.

Enter HLA subtype:i.e.HLA-A01:01

HLA-A01:01

第三步：提交

NetMHCpan4.0 多肽与HLA结合预测结果

免疫原性得分

peptide □	length □	%rank □	score □	prediction □
ACEDDKLMIY	10	0.471	0.074	Negative-High
ALDLTTSAGY	10	0.118	0.18	Negative-High
AMEKSSKYY	9	0.295	0.05	Negative-High
CTELKLSDY	9	0.144	0.674	Positive-High
EADPTGHSY	9	0.019	0.778	Positive-High
YLEKESIYY	9	0.019	0.54	Positive-High

HLA subtype：HLA-A01:01

Showing 1 to 6 of 6 entries

王广志,李雨雨,谢鹭.个性化肿瘤新抗原疫苗中抗原肽预测研究进展[J]. 生物化学与生物物理进展, 2019

Guangzhi wang et al. INeo-Epp: T-cell HLA class I immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features. Biomed Research International, 2020

基于高通量测序数据及免疫原性预测算法进行肿瘤新抗原排序

- 合作同期发表的预测流程：**pTuneos**

- 用于从高通量测序数据中预测新抗原并**对其真实免疫原性进行评估和排序**。

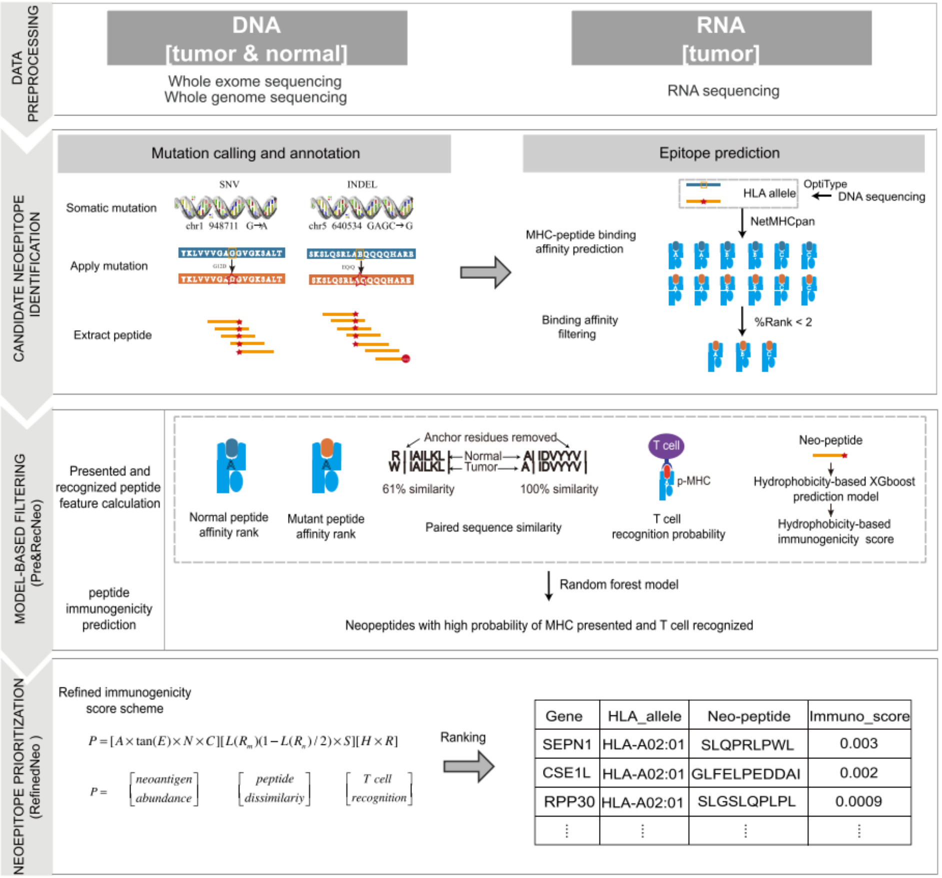
- **该计算框架包含四个步骤：**

  1）数据预处理

  2）候选新抗原鉴定

  3）**基于体外实验新抗原免疫活性数据训练特定机器学习模型**并处理数据不平衡性，该模型可以初步筛选可被MHC复合物递呈和T细胞表面受体识别的新抗原

  4）新抗原排序优化



DATA PREPROCESSING
DNA [tumor & normal]
Whole exome sequencing
Whole genome sequencing
RNA [tumor]
RNA sequencing
CANDIDATE NEOEPITOPE IDENTIFICATION
Mutation calling and annotation
SNV
INDEL
Somatic mutation
chr1 948711 G→A
chr5 640534 GAGC→G
Apply mutation
YKLVVVGAGGVGKSALT
G12D
YKLVVVGADGVGKSALT
SKSLQSRLAEQQQQHARE
EQQ
SKSLQSRLACQQQQHARE
Extract peptide
Epitope prediction
HLA allele
OptiType
DNA sequencing
NetMHCpan
MHC-peptide binding affinity prediction
Binding affinity filtering
%Rank < 2
MODEL-BASED FILTERING (Pre&RecNeo)
Presented and recognized peptide feature calculation
Normal peptide affinity rank
Mutant peptide affinity rank
Anchor residues removed
R IAILKL
W IAILKL
Normal
Tumor
A IDVYYV
A IDVYYV
61% similarity
100% similarity
Paired sequence similarity
T cell
p-MHC
T cell recognition probability
Neo-peptide
Hydrophobicity-based XGboost prediction model
Hydrophobicity-based immunogenicity score
Random forest model
peptide immunogenicity prediction
Neopeptides with high probability of MHC presented and T cell recognized
NEOEPITOPE PRIORITIZATION (RefinedNeo)
Refined immunogenicity score scheme
Ranking

$$P = [A \times \tan(E) \times N \times C][L(R_m)(1 - L(R_n)/2) \times S][H \times R]$$

$$P = \begin{bmatrix} neoantigen \\ abundance \end{bmatrix} \begin{bmatrix} peptide \\ dissimilariy \end{bmatrix} \begin{bmatrix} T\ cell \\ recognition \end{bmatrix}$$

Gene	HLA_allele	Neo-peptide	Immuno_score
SEPN1	HLA-A02:01	SLQPRLPWL	0.003
CSE1L	HLA-A02:01	GLFELPEDDAI	0.002
RPP30	HLA-A02:01	SLGSLQPLPL	0.0009
⋮	⋮	⋮	⋮

Zhou C et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. Genome Med. 2019

PART THREE

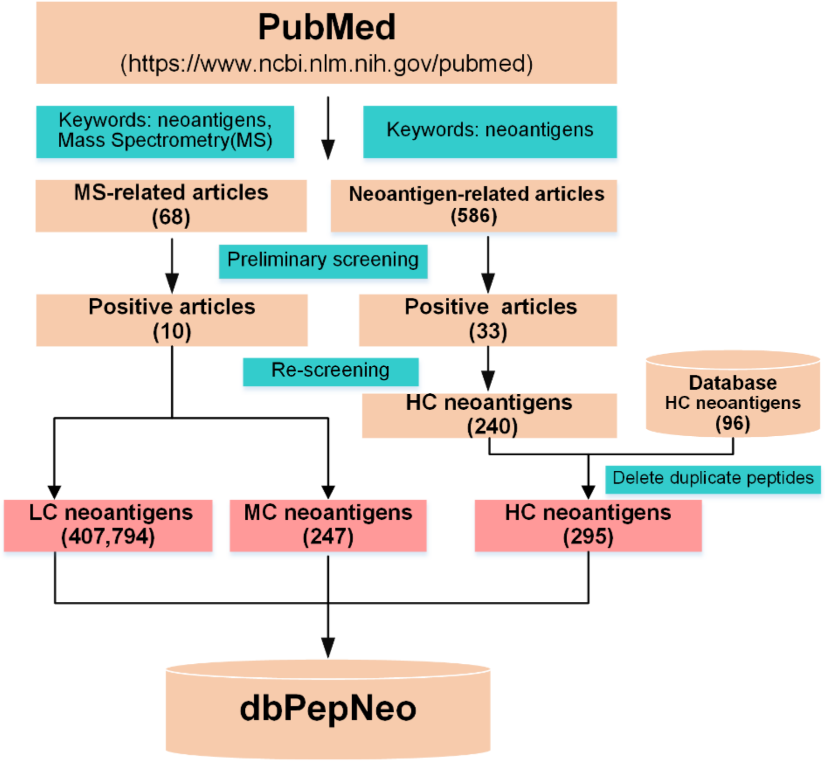
研究内容

dbPepNeo

新抗原肽数据库及工具平台

肿瘤新抗原肽段数据库的构建

数据收集与处理



PubMed
(https://www.ncbi.nlm.nih.gov/pubmed)

Keywords: neoantigens, Mass Spectrometry(MS)

Keywords: neoantigens

MS-related articles (68)

Neoantigen-related articles (586)

Preliminary screening

Positive articles (10)

Positive articles (33)

Re-screening

HC neoantigens (240)

Database HC neoantigens (96)

Delete duplicate peptides

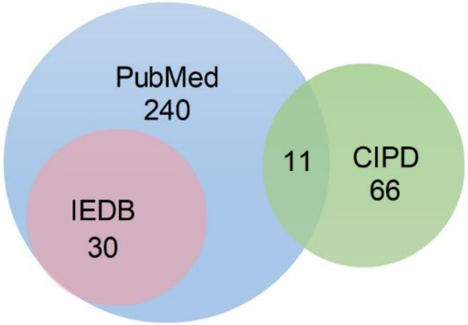
LC neoantigens (407,794)

MC neoantigens (247)

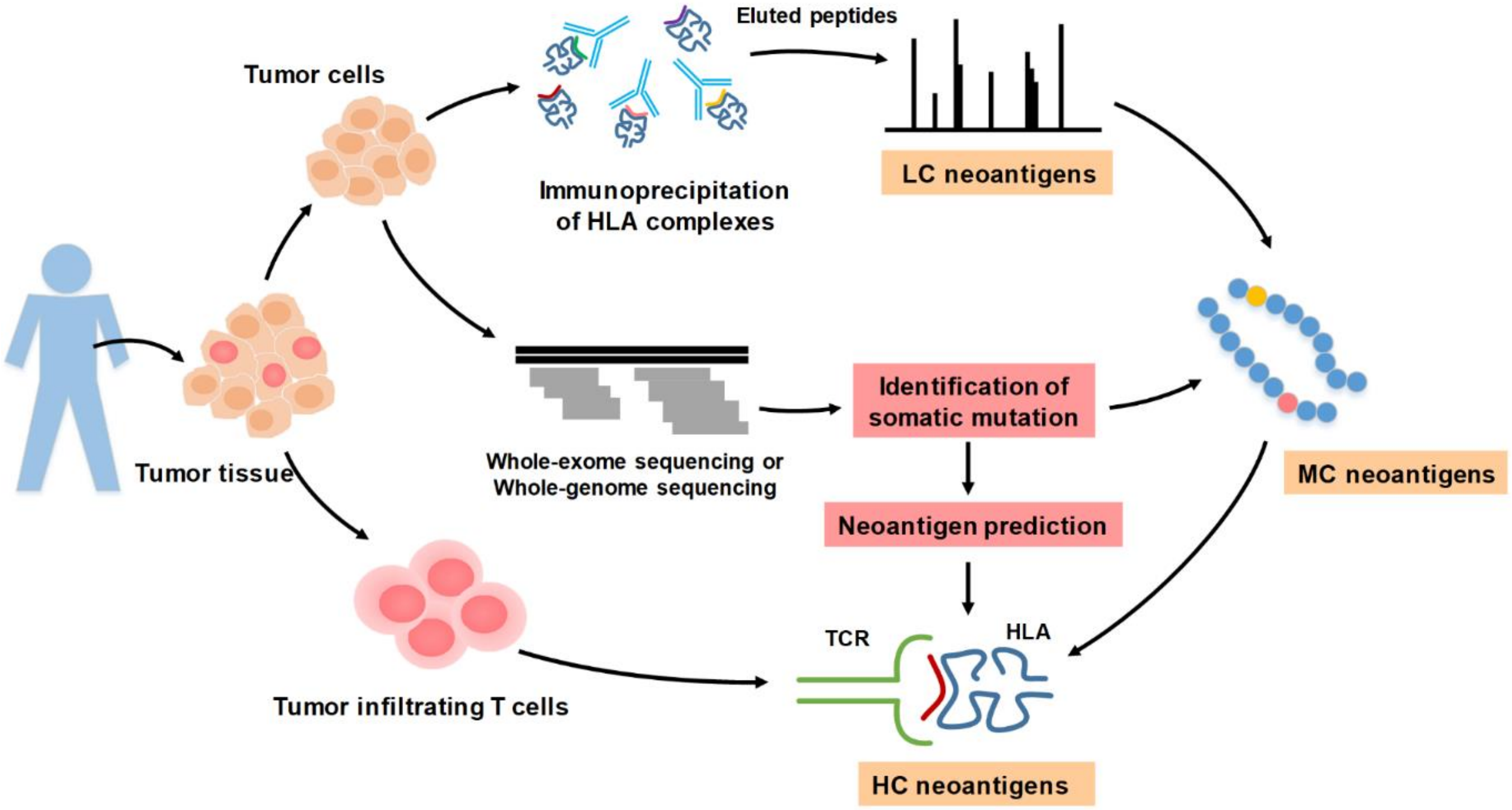
HC neoantigens (295)

dbPepNeo

HC neoantigens



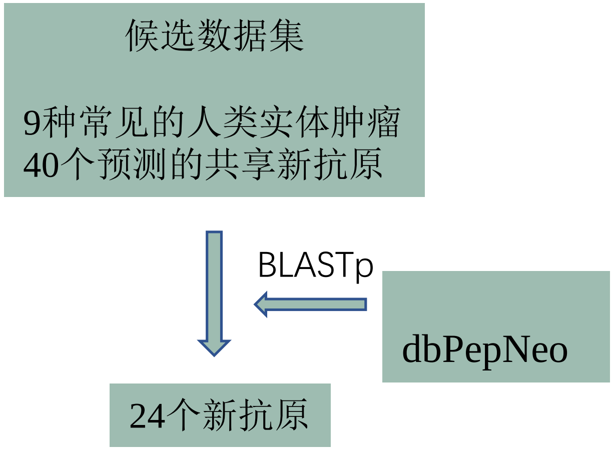
dbPepNeo 中新抗原数据的纳入标准基于以下几个关键的新抗原呈现步骤


Tumor cells
Eluted peptides
Immunoprecipitation of HLA complexes
LC neoantigens
Tumor tissue
Whole-exome sequencing or Whole-genome sequencing
Identification of somatic mutation
Neoantigen prediction
MC neoantigens
TCR
HLA
Tumor infiltrating T cells
HC neoantigens

通过 MS 鉴定并与 HLA-I 分子结合的原始肿瘤肽定义为 LC (low confidence)新抗原；含有体细胞突变并经MS和WES/WGS确认的肽段定义为中等置信度（MC）新抗原；特异性 TCR 识别实验验证的免疫原性肽被认为是高可信度 (HC) 新抗原。

肿瘤新抗原肽段数据库的构建

dbPepNeo在肿瘤新抗原研究中的应用：候选新抗原的过滤筛选



候选数据集

9种常见的人类实体肿瘤
40个预测的共享新抗原

BLASTp

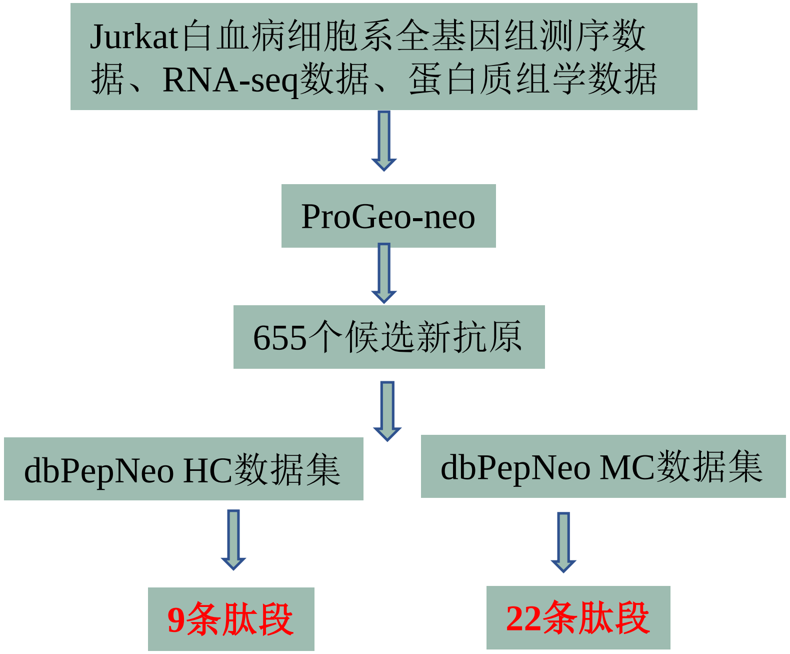
dbPepNeo

24个新抗原

1. 结果显示24个新抗原与HC新抗原具有序列相似性，**其中包括6条完全匹配的免疫验证的新抗原**，这也从侧面说明了HC新抗原作为高可信度筛选数据集的可靠性。
2. 24个新抗原和HC新抗原的序列一致性百分比在78%至100%之间。通常，TCR识别新抗原的可能性与序列一致性百分比成正比。

肿瘤新抗原肽段数据库的构建

dbPepNeo在肿瘤新抗原研究中的应用：流程预测的新抗原过滤筛选



Jurkat白血病细胞系全基因组测序数据、RNA-seq数据、蛋白质组学数据

ProGeo-neo

655个候选新抗原

dbPepNeo HC数据集

dbPepNeo MC数据集

9条肽段

22条肽段

结果表明**dbPepNeo的过滤**降低了后续免疫实验的负担，可以提高新抗原验证的效率，为患者争取到更多的时间。

Xiaoxiu Tan et al. dbPepNeo: a manually curated database for human tumor neoantigen peptides, Database, 2020

总结

Updated software, algorithm and database will be available in near future at: https://www.researchgate.net/profile/Lu-Xie-9/research

cancers

MDPI

Review

Main Strategies for the Identification of Neoantigens

Alexander V. Gopanenko, Ekaterina N. Kosobokova and Vyacheslav S. Kosorukov *

N.N. Blokhin National Medical Research Center of Oncology, Ministry of Health of the Russian Federation, 115478 Moscow, Russia; alexandr.gopanenko@yandex.ru (A.V.G.); ekkos@mail.ru (E.N.K.)
* Correspondence: kosorukov@ronc.ru; Tel.: +7-499-324-2274

Received: 4 September 2020; Accepted: 5 October 2020; Published: 7 October 2020

Simple Summary: This review provides an overview of currently available approaches applied for neoantigens discovery—tumor-specific peptides that appeared due to the mutation process and distinguish tumors from normal tissues. Focusing on genomics-based approaches and computational pipelines, we cover all steps required for selecting appropriate candidate peptides starting from NGS-derived data. Moreover, additional approaches such as mass-spectrometry-based and structure-based methods are discussed highlighting their advantages and disadvantages. This review also provides a description of available complex bioinformatics pipelines ensuring automated data processing resulting in a list of neoantigens. We propose the possible ideal pipeline that could be implemented in the neoantigens identification process. We discuss the integration of results generated by different approaches to improve the accuracy of neoantigens selection.

Table 1. *Cont.*

Pipeline	Source, Required Input Data and Otput:	Workflow and Features:	Refs.
ProGeo-neo 2020	Source: https://github.com/kbvstmd/ProGeo-neo Description: Neoantigen prediction workflow that integrates genomic and mass spectrometry data. It consists of three modules: construction of customized protein sequence database, HLA allele prediction, neoantigen prediction and filtration. Input: RNA-seq data (FASTQ format), Genomic variants (VCF format), LC-MS/MS data (Raw format). Output: List of candidate peptides	1. HLA typing (OptiType) 2. Identification tumor-specific antigens for NGS data (WES/RNA-seq) (BWA, GATK tools) 3. MHC binding prediction (NetMHCpan 4.0) 4. Verifying MHC-peptides using mass spectrometry data (MaxQuant) 5. Checking potential immunogenicity of T-cell-recognition	[150]
Neoepiscope 2020	Source: https://github.com/pdxgx/neoepiscope Description: Neoepitope identification pipeline that incorporates germline context and considers variant phasing for SNV and indels. Requires DNA-sequencing data. Input: Set of somatic and germline mutations (VCF format), BAM files. Output: TSV file with the information of mutations and neoepitopes	1. VCF files preprocessing (merging somatic and germline variants) 2. Haplotype phasing (HapCUT2) 3. Neoepitope prediction (MHCflurry, MHCnuggets, etc.)	[80]
neoANT-HILL 2020	Source: https://github.com/neoanthill/neoANT-HILL Description: User-friendly python-based toolkit that combines several pipelines that ensure fully-automated identification of potential neoantigens with a graphical interface. It allows starting from raw NGS data as well as ready-to-use variant calls. Input: Somatic variants (VCF format) and/or RNA-seq data (raw or aligned) Output: User-defined generic directory that contains variant calling data, FASTA with WT and MT sequences, predicted HLA types, gene expression estimates, tumor-infiltrating immune cells quantifications.	1. Expression estimation (kallisto) 2. Variant discovery (GATK tools) 3. HLA typing (OptiType) 4. Tumor-infiltrating immune-cell estimation (quanTIseq) 5. Variant annotation (snpEff) 6. MHC binding affinity prediction (IEDB tools, MHCflurry)	[151]
INeo-Epp 2020	Source: http://www.biostatistics.online/INeo-Epp/antigen.php Description: User-friendly web-tool implementing T-cell HLA class I immunogenicity prediction method based on sequence-related amino acid features utilizing the random forest algorithm. Input: Candidate peptide sequences (8-12 aa recommended), HLA allotype Output: Table containing peptides sequences annotated with score, %rank and prediction.	1. Providing peptide sequences and HLA types. 2. Annotation of peptides with score metrics. 3. Selecting immunogenic peptides with a score > 0.5 as recommended.	[152]

* The descriptions of the pipelines presented in the table are based on information provided in associated articles and obtained from the web-based source descriptions that are available on source websites. It is limited by highlighting the main features that distinguished the pipelines from each other. The date of the pipeline appearance is based on the publishing date of the supported article if other information is not provided. The source link is cited as "not available" if the website was not available at the time of writing. The output and input descriptions are presented as described in supporting articles or web-based sources (if available). In cases where a clear description was lacking, these fields were cited as "Not described". "Workflow and features" field contains information on the main steps that are available within the workflow. The main tools utilized as a part of the described workflows are also provided if they are described in supporting articles or in web-based sources.

Table 2. *Cont.*

Database, Year of Appearance	Source and Description	Refs.
dbPepNeo 2020	Source: http://www.biostatistics.online/dbPepNeo/ Description: dbPepNeo is a manually curated database of experimentally confirmed human tumor antigens that bind specifically to HLA class I, which contains information extracted from peer-reviewed articles and the publicly available data sources. The database relies on mass spectrometry (MS) validation and specific T-cell immunoassays. The peptides were classified according to validation methods: 1. Low confidence (407794): validated by MS only; 2. Medium confidence (247): contain a somatic mutation and are validated by MS and WES/WGS; 3. High confidence (295): immunogenicity was validated directly by utilizing specific T-cell response experiments. dbPepNeo also includes the following tools: ProGeo-neo (see Table 1) and INeo-Epp, a machine learning algorithm for neoepitope immunogenicity prediction using neoantigen peptide features.	[184]

* The information presented here is based on the introductions to these databases provided in the respective articles as well as on details specified on source websites. The database creation date is based on the publication date of the supporting article unless specified otherwise.