International Journal of
Molecular Sciences

MDPI

Article

Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction

Kaixuan Diao [1,2,3,†], Jing Chen [1,2,3,†], Tao Wu [1], Xuan Wang [1], Guangshuai Wang [1], Xiaoqin Sun [1], Xiangyu Zhao [1], Chenxu Wu [1], Jinyu Wang [1], Huizi Yao [1], Casimiro Gerarduzzi [4] and Xue-Song Liu [1,*]

[1] School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China
[2] Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] Département de Médecine, Faculté de Médecine, Université de Montréal, Montréal, QC H4T 1G2, Canada
* Correspondence: liuxs@shanghaitech.edu.cn
† These authors contributed equally to this work.

Abstract: Neoantigens derived from somatic DNA alterations are ideal cancer-specific targets. In recent years, the combination therapy of PD-1/PD-L1 blockers and neoantigen vaccines has shown clinical efficacy in original PD-1/PD-L1 blocker non-responders. However, not all somatic DNA mutations result in immunogenicity among cancer cells and efficient tools to predict the immunogenicity of neoepitopes are still urgently needed. Here, we present the Seq2Neo pipeline, which provides a one-stop solution for neoepitope feature prediction using raw sequencing data. Neoantigens derived from different types of genome DNA alterations, including point mutations, insertion deletions and gene fusions, are all supported. Importantly, a convolutional neural network (CNN)-based model was trained to predict the immunogenicity of neoepitopes and this model showed an improved performance compared to the currently available tools in immunogenicity prediction using independent datasets. We anticipate that the Seq2Neo pipeline could become a useful tool in the prediction of neoantigen immunogenicity and cancer immunotherapy. Seq2Neo is open-source software under an academic free license (AFL) v3.0 and is freely available at Github.

Keywords: immunogenicity; immunotherapy; bioinformatics pipeline; deep learning

1. Introduction

In recent years, PD-1/PD-L1 blocker immunotherapy has transformed the treatment of cancer. PD-1 is a protein found on T cells that helps to keep the immune systems in check. The combination of PD-1 and PD-L1 helps to stop T cells killing other cells, including cancer cells, which can result in immune evasion [1–4]. Previous studies have reported that only a small proportion of patients present lasting clinical responses while most patients only present transient responses or no response at all [5,6]. The combination of PD-1 blockers and other forms of immunotherapy, such as neoantigen vaccines, has demonstrated favorable development prospects [7,8].

Neoantigens derived from somatic DNA alterations are ideal cancer-specific targets. Neoantigen vaccines have demonstrated therapeutic effects in terms of enhancing immunotherapy efficacy [9]. It has also been reported that the combination of PD-1 antibodies and neoantigen vaccines is safe and effective in the treatment of cancer patients [8]. In addition, TCR-T-targeting neoantigens have shown dramatic effects in clinical practice [10]. However, the success of these neoantigen-related therapies relies on efficient neoantigen prediction tools.

A plethora of peptide–HLA binding prediction algorithms have been developed to predict which peptides would bind to specific cognate HLA alleles [11–14]. However, HLA–peptide binding affinity alone is not sufficient for predicting the immunogenicity of peptides. In addition, the currently available neoantigen prediction tools only provide

check for updates

Citation: Diao, K.; Chen, J.; Wu, T.; Wang, X.; Wang, G.; Sun, X.; Zhao, X.; Wu, C.; Wang, J.; Yao, H.; et al. Seq2Neo: A Comprehensive Pipeline for Cancer Neoantigen Immunogenicity Prediction. *Int. J. Mol. Sci.* **2022**, *23*, 11624. https://doi.org/10.3390/ijms231911624

Academic Editor: Naoe Taira Nihira

Received: 18 August 2022
Accepted: 21 September 2022
Published: 1 October 2022

limited neoantigen features or focus on specific genome alterations, such as point mutations. Methods for the accurate prediction of the immunogenicity of neoantigens based on raw sequence data are still urgently needed. Here, we present an open-source pipeline tool, Seq2Neo, which could provide a one-stop service for raw data preprocessing, HLA typing, mutation labeling and neoantigen prediction as it can support neoantigens derived from point mutations, insertion and deletions (INDELs) and gene fusions and predict various neoantigen features for each candidate peptide, including HLA binding affinity, the transport efficiency of transporters associated with antigen processing (TAP) and gene expression. Importantly, a convolutional neural network (CNN)-based immunogenicity prediction model was also constructed and this model showed an improved performance compared to other known methods.

2. Results and Discussion

2.1. Neoantigen Feature Prediction

In our study, Seq2Neo used a command line-based interface, which allowed users to perform workflows automatically. Seq2Neo used publicly available tools for mutation labeling, HLA typing and HLA affinity binding prediction. Then, a CNN-based model was constructed with features that were generated using Seq2Neo to predict the immunogenicity of peptides and directly stimulate CD8+ T cell response. Finally, Seq2Neo outputted various peptide features, including immunogenicity score, peptide–HLA binding affinity, TAP transport efficiency and gene expression (Figures 1 and S1). The Seq2Neo model (Figure 1) began by importing raw sequencing data in FASTQ, SAM or BAM format and then utilized the user input to select the corresponding workflow to run. Point mutation and INDEL detection was performed using Mutect2 [15] and gene fusion detection was performed using STAR-Fusion [16]. Subsequently, somatic variant data were generated in VCF format. MHC genotyping was performed using HLA-HD [17]. Before the neoantigen prediction, sample somatic variants were annotated using ANNOVAR [18] or Agfusion [19] to obtain potential mutant peptides.

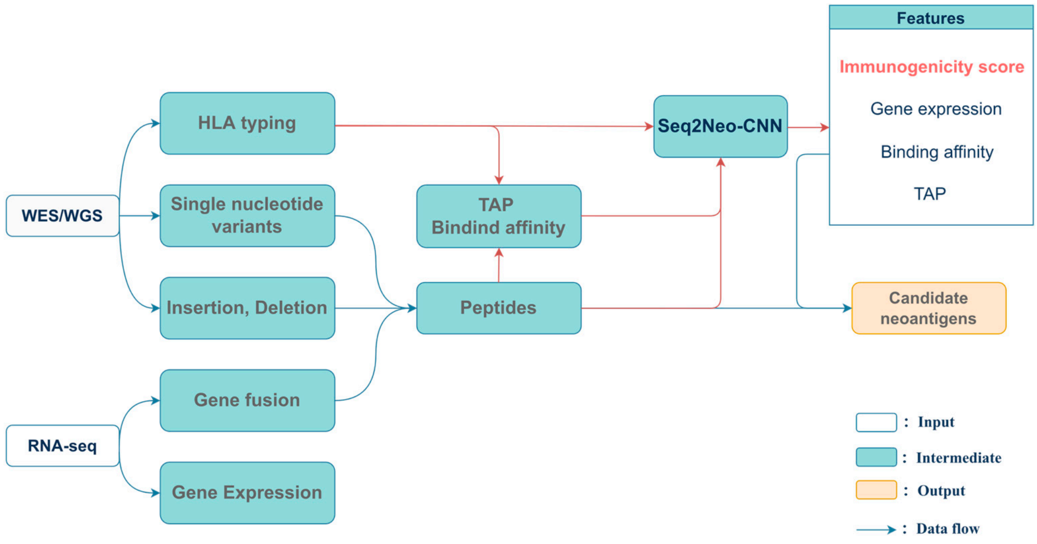
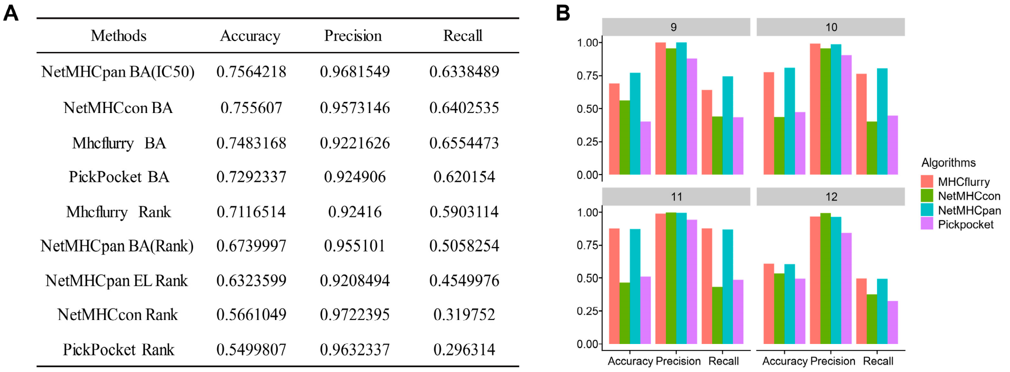


Figure 1. An overview of Seq2Neo. The input of Seq2Neo includes raw WGS, WES, RNA-seq or peptide information. Seq2Neo predicts various peptide features, including CNN-based immunogenicity score, peptide–HLA binding affinity, TAP transport efficiency and gene expression. Then, Seq2Neo uses those features to rank candidate neoantigens.

2.2. Selection of the Best HLA-I Binding Affinity Prediction Algorithms

We used 23319 peptides (14677 positives, with IC50 = 500 nm as the threshold) from the Immune Epitope Database (IEDB) to evaluate the performance of selected peptide–HLA I binding affinity prediction algorithms, including NetMHCpan [20], MHCflurry [21], PickPocket [22] and NetMHCcon [23]. The NetMHCpan BA model obtained the highest accuracy score (0.75) and the highest precision score (0.96) (Figure 2A,B), so this algorithm was selected for peptide–HLA binding prediction in Seq2Neo. In addition to peptide–HLA binding affinity, Seq2Neo used TPMCalculator [24] to detect gene expression and NetCTLpan [25] to obtain TAP transport efficiency.



Methods	Accuracy	Precision	Recall
NetMHCpan BA(IC50)	0.7564218	0.9681549	0.6338489
NetMHCcon BA	0.755607	0.9573146	0.6402535
Mhcflurry BA	0.7483168	0.9221626	0.6554473
PickPocket BA	0.7292337	0.924906	0.620154
Mhcflurry Rank	0.7116514	0.92416	0.5903114
NetMHCpan BA(Rank)	0.6739997	0.955101	0.5058254
NetMHCpan EL Rank	0.6323599	0.9208494	0.4549976
NetMHCcon Rank	0.5661049	0.9722395	0.319752
PickPocket Rank	0.5499807	0.9632337	0.296314

Figure 2. A benchmark analysis of different HLA-I binding affinity prediction algorithms: (**A**) a comparison of the performance of different algorithms in terms of prediction accuracy, precision and recall (the dataset was downloaded from IEDB and the thresholds of an IC50 value less than 500 nM and a rank percentile less than 1% were used to determine positive peptides); (**B**) a comparison of the different algorithms on different lengths of peptides.

2.3. Data Used for Seq2Neo-CNN Model Training

The fundamental feature of neoepitopes is their ability to stimulate cytolytic T cell responses, but this immunogenicity information cannot be predicted using most of the current neoantigen prediction tools. For immunogenicity prediction, we searched the IEDB database for experimental evidence that supported the immunogenicity of peptides and acquired 75496 experimentally evaluated immunogenicity assays (Figure 3). After applying our filter criteria (Section 3), 8975 data points (5342 negative peptides) were retained in the final dataset. We chose an independent dataset for model validation, which included 599 experimentally tested tumor-specific neoantigens from the Tumor Neoantigen Selection Alliance (TESLA) after deduplication and length restriction to 8–11 [26].

2.4. Features Associated with Peptide Immunogenicity

In order to find beneficial features for immunogenicity prediction, we compared the features of immunogenic and non-immunogenic peptides. The features of HLA-binding affinity, TAP transport efficiency and proteasomal C terminal cleavage were considered. The differences in HLA-binding affinity and TAP transport efficiency between the immunogenic and non-immunogenic peptides were significant but those in proteasomal C terminal cleavage were not (Figure 4A–C). HLA-binding affinity and TAP transport efficiency were not correlated (R = 0.02 and P = 0.055; Figure 4D); therefore, we incorporated these two features into our Seq2Neo-CNN model.

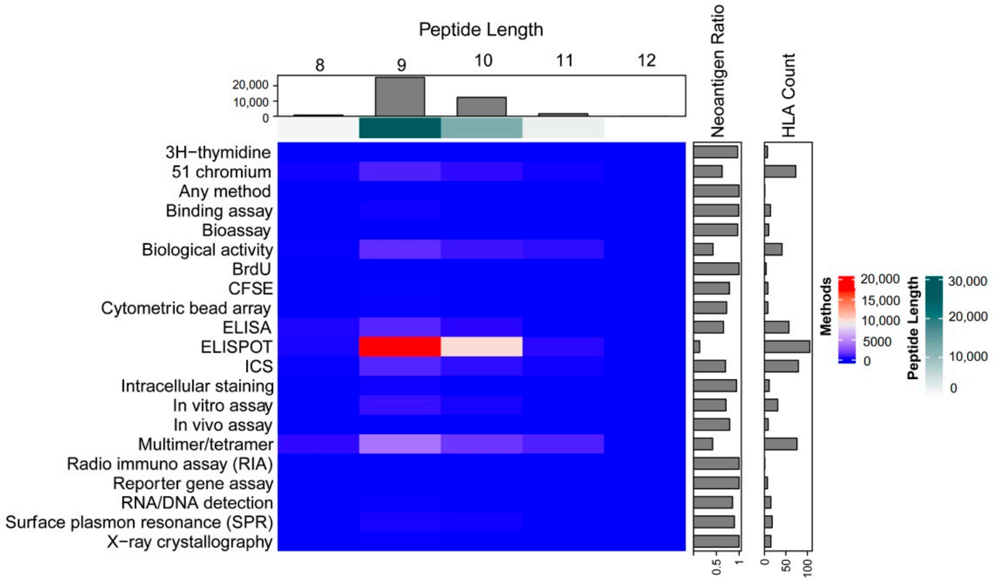


Figure 3. An overview of the data that were used for the Seq2Neo model training, including basic information about the IEDB dataset. We restricted the dataset to peptides with metadata that matched the following keywords: (1) linear epitopes, (2) specific T cell assays, (3) intact MHC I class, (4) originated from humans, (5) any diseases and (6) intact test information for negative peptides. CFSE, carboxyfluorescein succinimidyl amino ester; ELISA, enzyme-linked immunosorbent assay; ELISPOT, enzyme-linked immunosorbent spot; ICS, intracellular cytokine staining.

2.5. *Seq2Neo-CNN Model for Immunogenicity Prediction*

We built a CNN-based model, named Seq2Neo-CNN, to predict peptide immunogenicity (Figure 5A). The performance of the trained CNN model was compared to that of other machine learning models (ExtraTree, random forest, logistic regression, SVM and XGBoost), which were trained with the prediction accuracy, recall and precision data that were collected in this study (Figure 5B) and also data from the independent TESLA dataset (Figure S2). The Seq2Neo-CNN model showed the highest performance compared to the other machine learning models. The performance of Seq2Neo-CNN was also compared to other available neoepitope immunogenicity prediction tools using the independent TESLA dataset, including the DeepHLApan [27], IEDB [28] and DeepImmuno-CNN [29] models. Seq2Neo-CNN also showed the highest performance compared to these selected known methods (Figure 5C). The details of the Seq2Neo-CNN model construction and training are described in Section 3.

2.6. *Seq2Neo Validation*

In recent years, several tools for predicting neoantigens have been reported. Some representative tools are shown in Table 1 [12,14,30–33]. Two of the pipelines (TSNAD2 and Neopepsee) contain immunogenicity prediction functions. However, the other tools call their immunogenicity prediction modules DeepHLApan and IEDB, which proved to be less accurate than Seq2Neo-CNN (Figure 5C). Compared to the other pipelines, ease of use was also an advantage of Seq2Neo, since the Seq2Neo model provides a one-stop solution for neoantigen prediction using raw sequencing data. To demonstrate the performance of Seq2Neo, we applied Seq2Neo to samples from five cancer patients with experimentally validated neoantigenic mutations [34–37]. Those cancer samples contained WES, RNA-seq data and 16 experimentally validated neoantigenic DNA sites (Figure S3A,B). After applying the selection criteria (TAP > 0, IC50 $\leq$ 500, TPM > 0 and immunogenicity > 0.5), Seq2Neo identified 10 out of the 16 validated neoantigenic sites. The ranking of the candidate neoantigens is shown in Figure S3C. We selected three pipelines with detailed documentation, namely pVACseq, TSNAD 2.0 and NeoPredPipe to compare to Seq2Neo. Then, we compared the prediction results of Seq2Neo to those of the other three pipelines.

The ranking of the most validated neoantigenic sites in Seq2Neo was lower than that in the other three pipelines, which meant that Seq2Neo demonstrated an improved performance in terms of identifying the real immunogenic neoantigens (Figure S3C).

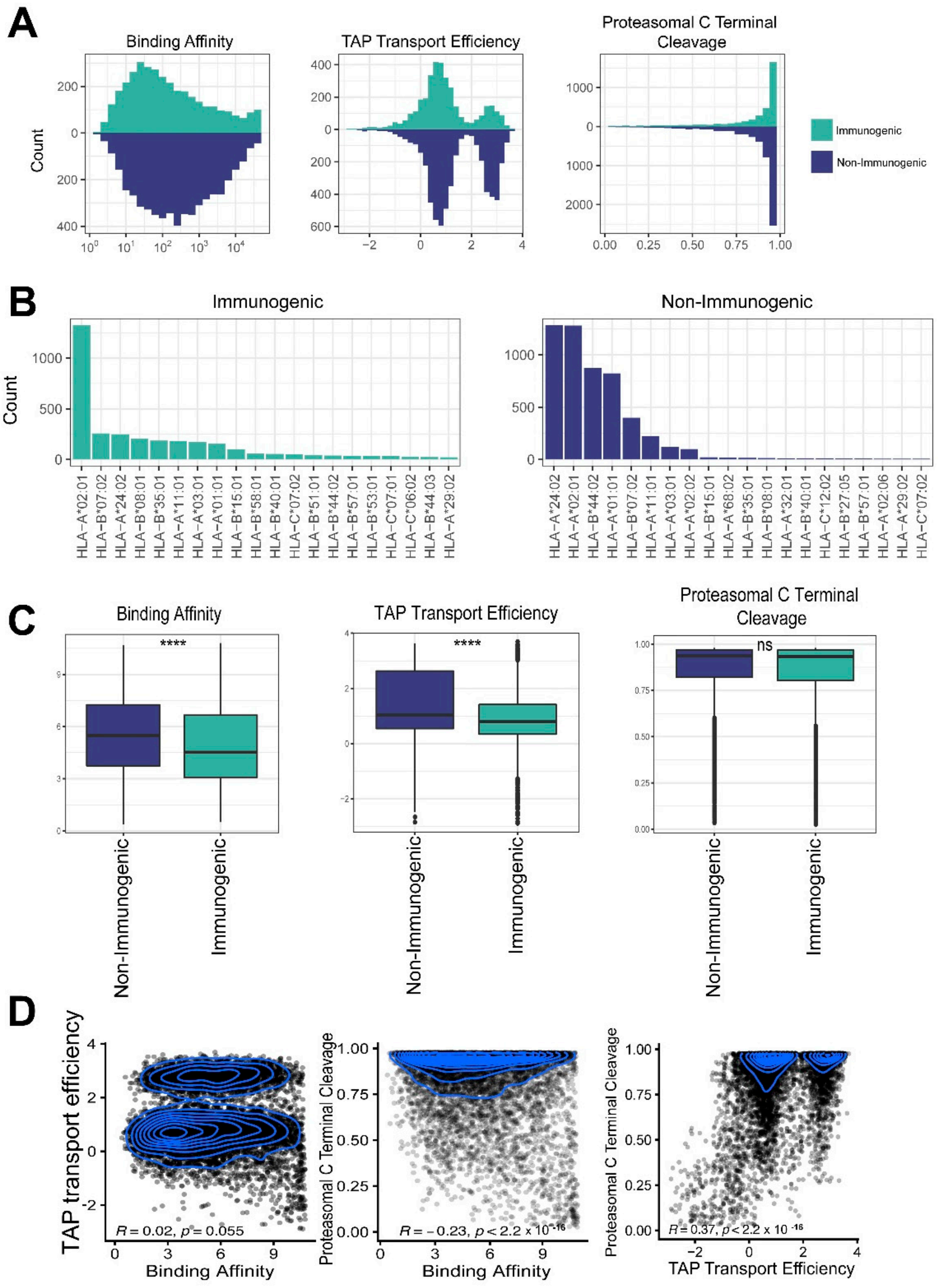


Figure 4. An exploration of the feature differences between immunogenic and non-immunogenic peptides: (**A**) a distribution comparison of HLA I binding affinity, TAP transport efficiency and proteasomal C terminal cleavage between immunogenic and non-immunogenic peptides; (**B**) a distribution comparison of the binding of the 20 most frequent HLA-I alleles to the immunogenic (left) and non-immunogenic mutated peptides (right); (**C**) a comparison of binding affinity, TAP transport efficiency and proteasomal C terminal cleavage between immunogenic and non-immunogenic mutated peptides (****, $p < 10^{-4}$; ns, not significant); (**D**) pairwise correlations between the three neoepitope features (peptide–HLA binding affinity, TAP transport efficiency and proteasomal C terminal cleavage), showing the Pearson correlation coefficients R and p values.

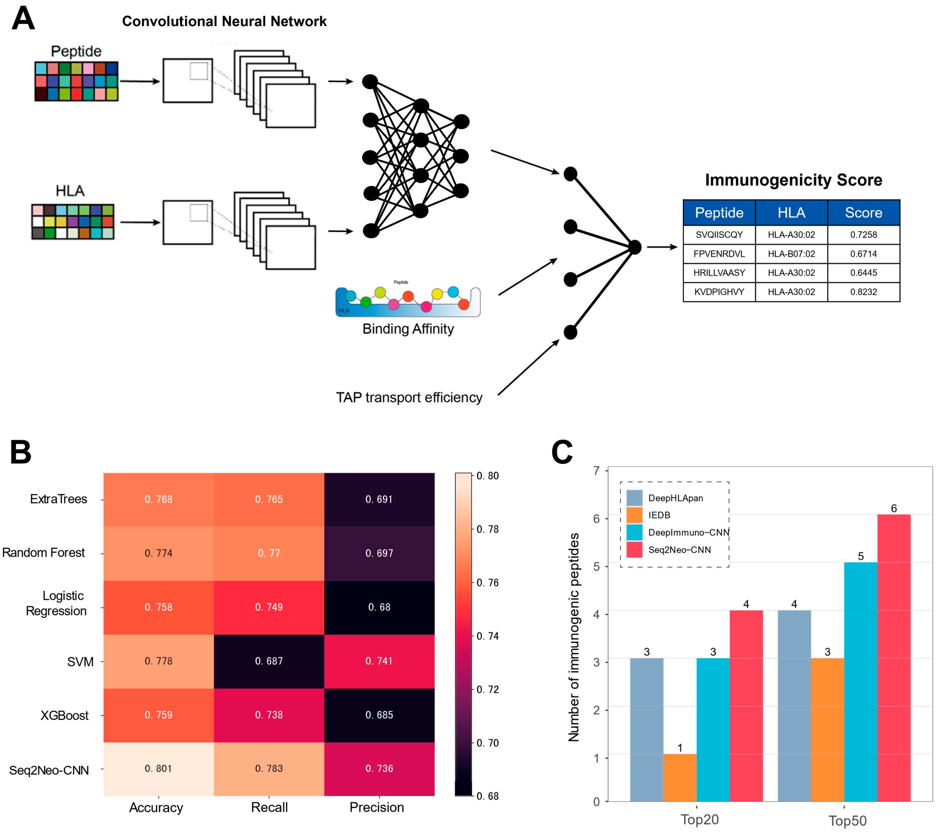


Figure 5. Our convolutional neural network-based model (Seq2Neo-CNN) for peptide immunogenicity prediction: (**A**) a schematic diagram of the Seq2Neo-CNN architecture (in this model, each peptide–MHC pair is subjected to two consecutive convolutional layers, followed by three fully connected dense layers and is then input into two fully connected dense layers, together with TAP transport efficiency and binding affinity information, to output a predicted immunogenicity value); (**B**) a comparison between the Seq2Neo-CNN model and other machine learning algorithms (to select the best predictive model, we constructed five traditional machine learning classifiers (ExtraTree, random forest, logistic regression, SVM and XGBoost) and the accuracy, recall and precision of each method are shown); (**C**) a comparison of the performance of the different models when predicting immunogenic peptides, based on the number of true positive peptides that overlapped with the top 20 or 50 predictions of each algorithm (the Seq2Neo-CNN model outperformed the existing immunogenicity prediction methods using the independent TESLA dataset).

Table 1. Representative tools for predicting neoantigens that have been published in recent years. The neoantigen type, input data, neoantigen class, HLA typing, immunogenicity score, TAP score and programming language that were used are presented.

Method	Neoantigen Types	Input Data	Neoantigen Class	HLA Typing	Immunogenicity Score	TAP Score	Language	Publish Year
Seq2Neo	SNVs, indels, gene fusions	WES/WGS, RNA-seq	Class I	Yes	Yes	Yes	Python	This study
pVACseq	SNVs, indels, gene fusions	VCF	Class I and II	No	No	No	Python	2019
TSNAD 2	SNVs, indels, gene fusions	WES/WGS, RNA-seq	Class I	Yes	Yes	No	Python	2021
NeoPredPipe	SNVs, indels	VCF, HLA types	Class I and II	No	No	No	Python	2019
Neopepsee	SNVs	VCF, RNA-seq, HLA types	Class I	Yes	Yes	No	Java	2018
nextNEOpi	SNVs, indels, gene fusions	WES/WGS, RNA-seq	Class I and II	Yes	No	No	Nextflow	2021
ProTECT	SNVs	WES/WGS, RNA-seq	Class I and II	Yes	No	No	Python	2020

2.7. Seq2Neo Implementation

The Seq2Neo pipeline was developed in Python 3.7.12 following a clean, modular and robust design, in accordance with best practice coding standards. The instructions for installing and running Seq2Neo are presented in a public GitHub repository (https://github.com/XSLiuLab/Seq2Neo accessed on 28 June 2022). This model was designed to run as a command line-based program with a user-friendly interface, thereby allowing non-expert users to become familiarized with its functions quickly. To facilitate the installation of Seq2Neo, Docker containers and Conda packages are provided (Docker: https://hub.docker.com/r/liuxslab/seq2neo accessed on 28 June 2022; Conda: https://anaconda.org/liuxslab/seq2neo accessed on 28 June 2022).

3. Materials and Methods

3.1. Data Preprocessing

The Seq2Neo model began by importing data in FASTQ, SAM and BAM format and then utilized the user input to select the corresponding workflow to run. The FASTQ files were processed for quality control and any adapter sequences at the end of the reads were removed using Fastp [38]. The raw sequence data were aligned to the reference genome (hg38) using the Burrows–Wheeler alignment tool [39]. When the input format was SAM or BAM, GATK best practice was performed first during the data preprocessing [40]. The SAM files were sorted and read group tags were added using Samtools [40]. After being sorting into coordinate order, the BAM files were processed using PICARD MarkDuplicates and the local realignment and quality score recalibration were conducted using the Genome Analysis Toolkit [41].

3.2. Somatic Mutation Detection

Generated or user-inputted co-cleaned BAM files were used for point mutation and insertion and deletion (INDEL) detection using Mutect2 [15] and gene fusions were detected using STAR-Fusion [16]. Then, somatic variant data were generated in VCF format. Additionally, parallel computation was enabled, which significantly reduced the computation time.

3.3. HLA Genotyping

Human leukocyte antigen (HLA) genes play a critical role in antigen presentation and immune signaling. Here, HLA-HD [17] was adapted for HLA genotyping using DNA-seq data and outputted personal HLA types for each patient, including class I and II HLAs.

3.4. Gene Expression Detection

The expression and presentation of tumor antigen-presenting cells on the surface are the prerequisites for neoantigens to be recognized by T cells. Seq2Neo supported the annotation of the expression of neoantigen candidates using TPMCalculator [24].

3.5. Neoepitope Features

In addition to peptide–HLA binding affinity, other features, including TAP transport efficiency, gene expression and immunogenicity score, could also be predicted using Seq2Neo. These neoepitope features could facilitate the filtering of candidate peptides for vaccine or immunotherapy target selection.

3.6. Immunogenicity Prediction (Seq2Neo-CNN Model)

As the core of the Seq2Neo pipeline, Seq2Neo-CNN could predict the immunogenicity of selected peptides. Below, we provide a detailed description of the generation of the Seq2Neo-CNN model.

3.6.1. Dataset Selection

We collected data from the IEDB database for the initial model training and validation (3 August 2021 version) using the following IEDB searching conditions: epitope (linear sequence), assay (positive/negative), T cell assay, MHC restriction (MHC class I), host (Human) and disease (any). In all, we found 75,496 relevant experiments. Although there are different ways to detect the immunogenicity of peptides, some experiments did not detect direct contact with T cells that induced immune responses, so we only selected data that were validated by ELISPOT, 51 Chromium, ICS, Multimer/Tetramer and ELISA. Then, we deleted any instances that did not have four-digit MHC alleles or were repeated. We also limited the length of peptides to 8–11 mer and removed negative peptides that had missing experimental information or less than four test subjects. Finally, we obtained 8975 peptides that met the requirements for the final dataset, among which 3633 were positive reactive instances and the remaining 5342 were negative. We selected an independent dataset for further evaluation, which included 599 experimentally tested tumor-specific neoantigens from the Tumor Neoantigen Selection Alliance (TESLA) after selecting only 8–11 mer peptides and removing duplicates [26].

3.6.2. Allele Representation

In order to input the MHC class I alleles into the neural network in numerical matrix form, we used pseudo-sequences to represent them. The pseudo-sequences were constructed by Nielsen et al. [42] and consisted of amino acid residues that were in contact with the peptides. The selected positions were 79, 24, 45, 59, 62, 63, 66, 67, 69, 70, 73, 74, 76, 77, 80, 81, 84, 95, 97, 99, 114, 116, 118, 143, 147, 150, 152, 156, 158, 159, 163, 167 and 171. We used the following strategy to encode the MHC pseudo-sequences.

3.6.3. Encoding Strategy

We used a one-hot encoding scheme to represent each HLA allele and peptide sequence in numerical matrix form, which were used as the inputs for the following algorithms. The one-hot encoding scheme was realized by assigning a unique integer to each letter in the 21-digit amino acid alphabet that contained padding characters as the index of that letter in the amino acid alphabet. Taking the letter “A” as an example, we obtained the alphabet “ACDEFGHIKLMNPQRSTVWYX” (the unknown amino acid was set to ”X”) and the corresponding index of alanine “A” was 0. Then, the values of the other amino acids were set to 0, but the value of “A” was set to 1. Finally, we obtained the one-hot vector of [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]. For each peptide, the unique one-hot vectors of each amino acid in the amino acid sequence were vertically combined to form a numerical matrix to complete vectorization.

3.6.4. Feature Normalization

Binding affinity and TAP transport efficiency were predicted for all peptide–HLAs using the method that was described previously and then normalized using the maximum and minimum values simultaneously. The basic mathematical form was represented as:

$$y = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

3.6.5. Prediction Model

We used a CNN (convolutional neural network) to predict the immunogenicity of mutant peptides. The proportions of the training set, testing set and validation set were 70%, 20% and 10%, respectively. The peptides and MHCs were processed by two consecutive convolutional layers, followed by three dense layers to execute the affine transformation and then flattened vectors with dimensions of 256 were obtained. NetMHCpan and NetCTLpan were used to calculate the binding affinity (IC50) and TAP transport efficiency of the peptides and those features were used to train the natural network. To incorporate the IC50 and TAP transport efficiency features into our CNN, two dense layers were included. Finally, Seq2Neo outputted the immunogenicity prediction. We used the ReLu function as the activation function. Some hyperparameters were set during the optimization process before the training started: batch size was set to 64, training loss with patience was set to 15, validation loss with patience was set to 20, epochs were set to 200 and the Adam learning rate was set to 0.001. Two early stopping strategies were adopted to ensure that the acquired model was the best possible version. In addition, we adopted batch normalization and dropout strategies to accelerate the model convergence speed and enhance its generalization ability. Since the number of negative reactive instances was significantly higher than that of positive reactive instances, the weight was set according to the proportions of negative and positive instances to eliminate this imbalance. The weight operation was mathematically represented as:

$$w = \frac{1}{S} \times \frac{T}{2}$$

where w is the negative or positive class weight, S is the number of corresponding reactive instances and T is the total number of training instances.

3.7. Other Machine Learning-Based Immunogenicity Prediction Models

In order to select the best model to predict immunogenicity, we compared the Seq2Neo-CNN model to five other machine learning algorithms (logistic regression, SVM, XGBoost, random forest and ExtraTree) after optimizing the parameters for each method. We used accuracy as the evaluation criterion to tune the best parameters for each model. The best parameters for logistic regression were acquired through 10-fold cross-validation (penalty = l2 and C = 2.21). Similar to logistic regression, kernel = rbf, gamma = 0.1 and C = 10 were the best parameters for SVM., whereas max_depth = 10, min_child_weight = 1.0, gamma = 1.625, subsample = 1.0 and colsample_bytree = 1.0 were the best parameters for XGBoost, n_estimators = 200 and min_samples_leaf = 2 were the best parameters for random forest and n_estimators = 1000 and min_samples_leaf = 2 were the best parameters for ExtraTree. Then, the optimized models were compared to the Seq2Neo-CNN model using the testing set and the TESLA dataset.

3.8. Seq2Neo Implementation in Cancer Patient Samples

To test the performance of the Seq2Neo pipeline, WES (normal/tumor exome) and RNA-seq (tumor transcriptome) data from five patients with different solid tumors were downloaded from the NCBI SRA database (bioproject IDs: PRJNA298310, PRJNA298330 and PRJNA298376) [34–36]. Each sample had 2–4 experimentally verified neoantigens derived from point mutations that could induce T cell responses. Here, we used Seq2Neo to predict these neoantigens to verify the performance of Seq2Neo. Then, we compared the rank percentage of Seq2Neo to that of pVACseq using default parameters.

4. Conclusions

As a supplement to PD-1 immunotherapy, neoantigens are ideal cancer-specific targets for precision vaccine design or TCR-T therapy and act as key factors in cancer immunoediting [43]. However, current neoantigen prediction is cumbersome and lacks a comprehensive one-step tool. Furthermore, most neoantigen prediction tools only focus on the binding between peptides and HLA I and accurate tools for directly predicting the immunogenicity of neoepitopes are still lacking. Seq2Neo is a user-friendly and robust tool that could provide a one-stop solution for neoantigen prediction using raw sequencing data. Importantly, various features of neoantigens can be predicted using Seq2Neo, including the immunogenicity capability of neoepitopes.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/ijms231911624/s1.

Author Contributions: Conceptualization, X.-S.L., K.D. and X.W.; methodology, K.D. and X.W.; software, K.D.; validation, K.D., J.C. and T.W.; formal analysis, K.D., J.C. and T.W.; investigation, K.D., J.C., T.W., X.W., G.W., X.S., X.Z., C.W., J.W., H.Y. and C.G.; resources, X.-S.L.; data curation, K.D., J.C. and T.W.; writing—original draft preparation, K.D., J.C. and X.-S.L.; writing—review and editing, K.D., J.C., T.W. and X.-S.L.; visualization, K.D.; supervision, X.-S.L.; project administration, X.-S.L.; funding acquisition, X.-S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by the Shanghai Science and Technology Commission (21ZR1442400), the National Natural Science Foundation of China (31771373) and startup funding from ShanghaiTech University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The HLA-I binding affinity and immunogenicity data were downloaded from the IEDB data portal (https://www.iedb.org/, accessed on 3 August 2021). The WES data and mRNA expression data from the five cancer samples were downloaded from the SRA database of NCBI (https://www.ncbi.nlm.nih.gov/sra, accessed on 20 May 2022).

Acknowledgments: We would like to thank the ShanghaiTech University High-Performance Computing Public Service Platform for their computing services. We would also like to thank the multi-omics facility and molecular cellular facility of ShanghaiTech University for their technical help.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Waldmann, T.A. Immunotherapy: Past, Present and Future. *Nat. Med.* **2003**, *9*, 269–277. [CrossRef]
2. Sangro, B.; Sarobe, P.; Hervás-Stubbs, S.; Melero, I. Advances in Immunotherapy for Hepatocellular Carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **2021**, *18*, 525–543. [CrossRef]
3. Ren, Y.; Song, J.; Li, X.; Luo, N. Rationale and Clinical Research Progress on PD-1/PD-L1-Based Immunotherapy for Metastatic Triple-Negative Breast Cancer. *Int. J. Mol. Sci.* **2022**, *23*, 8878. [CrossRef]
4. Topalian, S.L.; Hodi, F.S.; Brahmer, J.R.; Gettinger, S.N.; Smith, D.C.; McDermott, D.F.; Powderly, J.D.; Carvajal, R.D.; Sosman, J.A.; Atkins, M.B.; et al. Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *N. Engl. J. Med.* **2012**, *366*, 2443–2454. [CrossRef]
5. Wang, S.; He, Z.; Wang, X.; Li, H.; Liu, X.-S. Antigen Presentation and Tumor Immunogenicity in Cancer Immunotherapy Response Prediction. *Elife* **2019**, *8*, e49020. [CrossRef]
6. Boutros, C.; Tarhini, A.; Routier, E.; Lambotte, O.; Ladurie, F.L.; Carbonnel, F.; Izzeddine, H.; Marabelle, A.; Champiat, S.; Berdelou, A.; et al. Safety Profiles of Anti-CTLA-4 and Anti-PD-1 Antibodies Alone and in Combination. *Nat. Rev. Clin. Oncol.* **2016**, *13*, 473–486. [CrossRef]
7. Morand, S.; Devanaboyina, M.; Staats, H.; Stanbery, L.; Nemunaitis, J. Ovarian Cancer Immunotherapy and Personalized Medicine. *Int. J. Mol. Sci.* **2021**, *22*, 6532. [CrossRef]
8. Ott, P.A.; Hu-Lieskovan, S.; Chmielowski, B.; Govindan, R.; Naing, A.; Bhardwaj, N.; Margolin, K.; Awad, M.M.; Hellmann, M.D.; Lin, J.J.; et al. A Phase Ib Trial of Personalized Neoantigen Therapy plus Anti-PD-1 in Patients with Advanced Melanoma, Non-Small Cell Lung Cancer, or Bladder Cancer. *Cell* **2020**, *183*, 347–362. [CrossRef]
9. Hu, Z.; Leet, D.E.; Allesøe, R.L.; Oliveira, G.; Li, S.; Luoma, A.M.; Liu, J.; Forman, J.; Huang, T.; Iorgulescu, J.B.; et al. Personal Neoantigen Vaccines Induce Persistent Memory T Cell Responses and Epitope Spreading in Patients with Melanoma. *Nat. Med.* **2021**, *27*, 515–525. [CrossRef]
10. Leidner, R.; Sanjuan Silva, N.; Huang, H.; Sprott, D.; Zheng, C.; Shih, Y.-P.; Leung, A.; Payne, R.; Sutcliffe, K.; Cramer, J.; et al. Neoantigen T-Cell Receptor Gene Therapy in Pancreatic Cancer. *N. Engl. J. Med.* **2022**, *386*, 2112–2119. [CrossRef]
11. Hasegawa, T.; Hayashi, S.; Shimizu, E.; Mizuno, S.; Niida, A.; Yamaguchi, R.; Miyano, S.; Nakagawa, H.; Imoto, S. Neoantimon: A Multifunctional R Package for Identification of Tumor-Specific Neoantigens. *Bioinformatics* **2020**, *36*, 4813–4816. [CrossRef]
12. Hundal, J.; Carreno, B.M.; Petti, A.A.; Linette, G.P.; Griffith, O.L.; Mardis, E.R.; Griffith, M. PVAC-Seq: A Genome-Guided in Silico Approach to Identifying Tumor Neoantigens. *Genome Med.* **2016**, *8*, 11. [CrossRef]
13. Lang, F.; Riesgo-Ferreiro, P.; Löwer, M.; Sahin, U.; Schrörs, B. NeoFox: Annotating Neoantigen Candidates with Neoantigen Features. *Bioinformatics* **2021**, *37*, 4246–4247. [CrossRef]
14. Rieder, D.; Fotakis, G.; Ausserhofer, M.; René, G.; Paster, W.; Trajanoski, Z.; Finotello, F. NextNEOpi: A Comprehensive Pipeline for Computational Neoantigen Prediction. *Bioinformatics* **2022**, *38*, 1131–1132. [CrossRef]

15. Benjamin, D.; Sato, T.; Cibulskis, K.; Getz, G.; Stewart, C.; Lichtenstein, L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* **2019**. bioRix:861054.
16. Haas, B.J.; Dobin, A.; Stransky, N.; Li, B.; Yang, X.; Tickle, T.; Bankapur, A.; Ganote, C.; Doak, T.G.; Pochet, N.; et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* **2017**. bioRxiv:120295.
17. Kawaguchi, S.; Higasa, K.; Shimizu, M.; Yamada, R.; Matsuda, F. HLA-HD: An Accurate HLA Typing Algorithm for next-Generation Sequencing Data. *Hum. Mutat.* **2017**, *38*, 788–797. [CrossRef]
18. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
19. Murphy, C.; Elemento, O. AGFusion: Annotate and Visualize Gene Fusions. *bioRxiv* **2016**. bioRxiv:080903.
20. Reynisson, B.; Alvarez, B.; Paul, S.; Peters, B.; Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data. *Nucleic Acids Res.* **2020**, *48*, W449–W454. [CrossRef]
21. O'Donnell, T.J.; Rubinsteyn, A.; Bonsack, M.; Riemer, A.B.; Laserson, U.; Hammerbacher, J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst.* **2018**, *7*, 129–132.e4. [CrossRef]
22. Zhang, H.; Lund, O.; Nielsen, M. The PickPocket Method for Predicting Binding Specificities for Receptors Based on Receptor Pocket Similarities: Application to MHC-Peptide Binding. *Bioinformatics* **2009**, *25*, 1293–1299. [CrossRef]
23. Karosiene, E.; Lundegaard, C.; Lund, O.; Nielsen, M. NetMHCcons: A Consensus Method for the Major Histocompatibility Complex Class I Predictions. *Immunogenetics* **2012**, *64*, 177–186. [CrossRef]
24. Vera Alvarez, R.; Pongor, L.S.; Mariño-Ramírez, L.; Landsman, D. TPMCalculator: One-Step Software to Quantify MRNA Abundance of Genomic Features. *Bioinformatics* **2019**, *35*, 1960–1962. [CrossRef]
25. Stranzl, T.; Larsen, M.V.; Lundegaard, C.; Nielsen, M. NetCTLpan: Pan-Specific MHC Class I Pathway Epitope Predictions. *Immunogenetics* **2010**, *62*, 357–368. [CrossRef]
26. Wells, D.K.; van Buuren, M.M.; Dang, K.K.; Hubbard-Lucey, V.M.; Sheehan, K.C.; Campbell, K.M.; Lamb, A.; Ward, J.P.; Sidney, J.; Blazquez, A.B.; et al. Key Parameters of Tumor Epitope Immunogenicity Revealed through a Consortium Approach Improve Neoantigen Prediction. *Cell* **2020**, *183*, 818–834. [CrossRef]
27. Wu, J.; Wang, W.; Zhang, J.; Zhou, B.; Zhao, W.; Su, Z.; Gu, X.; Wu, J.; Zhou, Z.; Chen, S. DeepHLApan: A Deep Learning Approach for Neoantigen Prediction Considering Both HLA-Peptide Binding and Immunogenicity. *Front. Immunol.* **2019**, *10*, 2559. [CrossRef]
28. Calis, J.J.; Maybeno, M.; Greenbaum, J.A.; Weiskopf, D.; De Silva, A.D.; Sette, A.; Keşmir, C.; Peters, B. Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **2013**, *9*, e1003266. [CrossRef]
29. Li, G.; Iyer, B.; Prasath, V.S.; Ni, Y.; Salomonis, N. DeepImmuno: Deep Learning-Empowered Prediction and Generation of Immunogenic Peptides for T-Cell Immunity. *Brief. Bioinform.* **2021**, *22*, bbab160. [CrossRef]
30. Zhou, Z.; Wu, J.; Ren, J.; Chen, W.; Zhao, W.; Gu, X.; Chi, Y.; He, Q.; Yang, B.; Wu, J.; et al. TSNAD v2.0: A One-Stop Software Solution for Tumor-Specific Neoantigen Detection. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4510–4516. [CrossRef]
31. Schenck, R.O.; Lakatos, E.; Gatenbee, C.; Graham, T.A.; Anderson, A.R. NeoPredPipe: High-Throughput Neoantigen Prediction and Recognition Potential Pipeline. *BMC Bioinform.* **2019**, *20*, 264. [CrossRef] [PubMed]
32. Kim, S.; Kim, H.S.; Kim, E.; Lee, M.; Shin, E.-C.; Paik, S. Neopepsee: Accurate Genome-Level Prediction of Neoantigens by Harnessing Sequence and Amino Acid Immunogenicity Information. *Ann. Oncol.* **2018**, *29*, 1030–1036. [CrossRef]
33. Rao, A.A.; Madejska, A.A.; Pfeil, J.; Paten, B.; Salama, S.R.; Haussler, D. ProTECT—Prediction of T-Cell Epitopes for Cancer Therapy. *Front. Immunol.* **2020**, *11*, 483296. [CrossRef] [PubMed]
34. Gros, A.; Parkhurst, M.R.; Tran, E.; Pasetto, A.; Robbins, P.F.; Ilyas, S.; Prickett, T.D.; Gartner, J.J.; Crystal, J.S.; Roberts, I.M.; et al. Prospective Identification of Neoantigen-Specific Lymphocytes in the Peripheral Blood of Melanoma Patients. *Nat. Med.* **2016**, *22*, 433–438. [CrossRef] [PubMed]
35. Gros, A.; Tran, E.; Parkhurst, M.R.; Ilyas, S.; Pasetto, A.; Groh, E.M.; Robbins, P.F.; Yossef, R.; Garcia-Garijo, A.; Fajardo, C.A.; et al. Recognition of Human Gastrointestinal Cancer Neoantigens by Circulating PD-1+ Lymphocytes. *J. Clin. Investig.* **2019**, *129*, 4992–5004. [CrossRef] [PubMed]
36. Tran, E.; Ahmadzadeh, M.; Lu, Y.-C.; Gros, A.; Turcotte, S.; Robbins, P.F.; Gartner, J.J.; Zheng, Z.; Li, Y.F.; Ray, S.; et al. Immunogenicity of Somatic Mutations in Human Gastrointestinal Cancers. *Science* **2015**, *350*, 1387–1390. [CrossRef]
37. Leoni, G.; D'Alise, A.M.; Tucci, F.G.; Micarelli, E.; Garzia, I.; De Lucia, M.; Langone, F.; Nocchi, L.; Cotugno, G.; Bartolomeo, R.; et al. VENUS, a Novel Selection Approach to Improve the Accuracy of Neoantigens' Prediction. *Vaccines* **2021**, *9*, 880. [CrossRef]
38. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
39. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
40. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33.

41. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]
42. Nielsen, M.; Lundegaard, C.; Blicher, T.; Lamberth, K.; Harndahl, M.; Justesen, S.; Røder, G.; Peters, B.; Sette, A.; Lund, O.; et al. NetMHCpan, a Method for Quantitative Predictions of Peptide Binding to Any HLA-A and-B Locus Protein of Known Sequence. *PLoS ONE* **2007**, *2*, e796. [CrossRef] [PubMed]
43. Wu, T.; Wang, G.; Wang, X.; Wang, S.; Zhao, X.; Wu, C.; Ning, W.; Tao, Z.; Chen, F.; Liu, X.-S. Quantification of Neoantigen-Mediated Immunoediting in Cancer Evolution. *Cancer Res.* **2022**, *82*, 2226–2238. [CrossRef] [PubMed]