cells

MDPI

Article

PGNneo: A Proteogenomics-Based Neoantigen Prediction Pipeline in Noncoding Regions

Xiaoxiu Tan [1,2], Linfeng Xu [2], Xingxing Jian [2], Jian Ouyang [2], Bo Hu [3], Xinrong Yang [3], Tao Wang [1,*] and Lu Xie [2,*]

[1] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Shanghai-MOST Key Laboratory of Health and Disease Genomics & Institute of Genome and Bioinformatics, Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai 200237, China
[3] Liver Cancer Institute, Fudan University, Shanghai 200032, China
* Correspondence: neowangtao@sjtu.edu.cn (T.W.); xielu@sibpt.com (L.X.)

Abstract: The development of a neoantigen-based personalized vaccine has promise in the hunt for cancer immunotherapy. The challenge in neoantigen vaccine design is the need to rapidly and accurately identify, in patients, those neoantigens with vaccine potential. Evidence shows that neoantigens can be derived from noncoding sequences, but there are few specific tools for identifying neoantigens in noncoding regions. In this work, we describe a proteogenomics-based pipeline, namely PGNneo, for use in discovering neoantigens derived from the noncoding region of the human genome with reliability. In PGNneo, four modules are included: (1) noncoding somatic variant calling and HLA typing; (2) peptide extraction and customized database construction; (3) variant peptide identification; (4) neoantigen prediction and selection. We have demonstrated the effectiveness of PGNneo and applied and validated our methodology in two real-world hepatocellular carcinoma (HCC) cohorts. TP53, WWP1, ATM, KMT2C, and NFE2L2, which are frequently mutating genes associated with HCC, were identified in two cohorts and corresponded to 107 neoantigens from noncoding regions. In addition, we applied PGNneo to a colorectal cancer (CRC) cohort, demonstrating that the tool can be extended and verified in other tumor types. In summary, PGNneo can specifically detect neoantigens generated by noncoding regions in tumors, providing additional immune targets for cancer types with a low tumor mutational burden (TMB) in coding regions. PGNneo, together with our previous tool, can identify coding and noncoding region-derived neoantigens and, thus, will contribute to a complete understanding of the tumor immune target landscape. PGNneo source code and documentation are available at Github. To facilitate the installation and use of PGNneo, we provide a Docker container and a GUI.

Keywords: neoantigen; noncoding regions; proteogenomics; prediction pipeline; tumor immunotherapy

Citation: Tan, X.; Xu, L.; Jian, X.; Ouyang, J.; Hu, B.; Yang, X.; Wang, T.; Xie, L. PGNneo: A Proteogenomics-Based Neoantigen Prediction Pipeline in Noncoding Regions. *Cells* **2023**, *12*, 782. https://doi.org/10.3390/cells12050782

Academic Editor: Yu Xue

Received: 27 January 2023
Revised: 26 February 2023
Accepted: 27 February 2023
Published: 1 March 2023

1. Introduction

Neoantigens are considered to be promising therapeutic targets owing to their tumor specificity, and to their neither being affected by pre-existing immune tolerance nor generating autoimmunity [1]. Thus, neoantigens can be used as potential targets for therapeutic vaccines. A neoantigen vaccine is designed to trigger de novo T cell response and broaden the endogenous repertoire of tumor-specific T cells [2]. A phase-I trial of a neoantigen-based peptide vaccine showed that four patients with stage III melanoma induced CD4+ T cell and CD8+ T cell responses after vaccination and remained disease-free for a median follow-up of 25 months after vaccination. This demonstrates the potent tumor-specific immunogenicity and antitumor activity of neoantigen vaccines [3]. However, one major challenge in neoantigen vaccine design is the rapid and accurate identification of neoantigens that can induce T cell responses in patients.

The advent of next-generation sequencing has provided opportunities to efficiently identify tumor-specific antigens in individual patients, leading to the exploration of clinical target therapies. In fact, several pipelines have been developed to predict neoantigens, such as pVACtools [4], NeoPredPipe [5], Neopepsee [6], etc. Although the development of these tools has opened the way for identifying potentially immunogenic neoantigens [7], limitations to these tools exist. First, most traditional prediction pipelines were developed based on genomic and transcriptomic data. Many false-positive neoantigens inevitably occur, due to the large number of mutations in individual patients and the limited performance of MHC ligand binding prediction [8]. In addition, studies have shown that the mRNA measurements of many genes correlate poorly with protein expression [9,10]. With advances in mass spectrometry (MS)-based proteomics, the combination of proteomics and genomics, i.e., proteogenomics, has been a major force in driving personalized neoantigen vaccine identification [11–13]. It allows the presence verification of those peptides that are most likely to generate an immune response based on neoantigen prediction pipelines; thus, such peptides may be moved into subsequent functional selection processes. Proteogenomics has greatly reduced the number of false positives for predicted neoantigens and has eased the burden of experimental validation. Our group previously developed proteogenomics neoantigen prediction pipelines, ProGeo-neo [14] and ProGeo-neo2.0 [15], and WEN B et al. developed NeoFlow [16].

Another limitation of the currently existing neoantigen prediction pipelines is that they almost all focus on genomic coding regions. While variants in protein-coding regions have received the most attention, numerous studies have noted the importance of noncoding variants in cancers [17]. Exomes only account for 2% of the human genome, whereas up to 75% of the genome has been shown to be transcribed and potentially translated [18]. Therefore, many allegedly noncoding regions are actually protein-coding. For example, long noncoding RNAs (LncRNA) are a type of noncoding RNA with a length of more than 200 nt, lacking a protein-coding function due to the lack of a complete open read frame (ORF) [19]. Intriguingly, several recent studies have noted LncRNAs as a source of new peptides [20,21]. Of particular relevance to tumor neoantigen discovery, 99% of cancer mutations are located in noncoding regions [17]. Therefore, focusing on the exome as the only source of tumor neoantigens is very restrictive. Notably, peptides derived from the noncoding regions have been shown to bind to MHC molecules, some of which were identified as the targets of T cells [22–24]. Subsequently, landmark studies demonstrated that the noncoding regions are the main source of targetable tumor-specific antigens [25,26]. However, an efficient and easy-to-use tool to predict and investigate the personalized neoantigens from noncoding regions is still lacking.

Herein, we present PGNneo, an integrated computational pipeline to predict noncoding neoantigens from RNA-seq and MS data. We demonstrated the effectiveness of PGNneo and validated our methodology in two real-world hepatocellular carcinoma (HCC) cohorts. In addition, we applied PGNneo to a colorectal cancer (CRC) cohort, demonstrating that the tool can be extended and verified in other tumor types. PGNneo is an efficient tool to identify noncoding neoantigens and can be easily installed and deployed at https://github.com/tanxiaoxiu/PGNneo. To be more user-friendly, we also provide a Docker version at (https://hub.docker.com/r/xiaoxiutan/pgnneo) and a GUI.

2. Methods

2.1. Data Collection

The paired-end sequencing data of lncRNA from 5 HCC patient-derived xenograft (PDX) samples, including tumor and normal tissues, were obtained from Hu et al. [27]. The proteomics datasets of the HCC cell line were downloaded from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) (accessed on 21 September 2020) with the identifier, PXD000529. This cohort is hereinafter referred to as HCC_HF. Another HCC cohort (hereinafter referred to as HCC_HT) from a previous collaboration with Jiang et al. [28] included RNA-seq files and MS raw data from 10 patients (early stage

of HCC, subtype 3) were downloaded from the Gene Expression Omnibus (GEO) (accession number GSE124535) (accessed on 15 November 2021) and iProX database (http://www.iprox.org, accession number IPX0000937000) (accessed on 15 November 2021), respectively. Detailed sample information is provided in Table S1 in the Supplementary Materials. In addition, we collected a CRC cohort [29], including RNA-seq data and MS raw data from three CRC cell lines and RNA-seq data from one normal fetal small intestine cell line. This can be downloaded from the GEO (accession number GSE195985) (accessed on 2 October 2022) and the ProteomeXchange Consortium (Identifier PXD028309) (accessed on 2 October 2022), respectively. Detailed sample information of this cohort is presented in Table S12 in the Supplementary Materials.

The human reference genome (hg38) and Proteome (version 101) were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz) (accessed on 1 September 2020) and the Ensembl database (http://www.ensembl.org/) (accessed on 1 September 2020), respectively. Contaminated protein sequences were available in a FASTA format from the common repository of adventitious proteins (cRAP) (http://www.thegpm.org/crap/) (accessed on 1 September 2020).

2.2. RNA-Seq Data Processing

RNA-seq raw data were cleaned by Trimmomatic (v0.39) (Max Planck Institute of Molecular Plant Physiology, Golm, Germany) [30], with the standard adapters trimmed and low-quality reads filtered. All clean reads were aligned to the human reference genome using the Burrows–Wheeler alignment tool (BWA, v0.7.17) (Wellcome Trust Sanger Institute, Cambridge, UK) [31] with the default parameters. The resulting .sam file was converted to .bam, sorted, and indexed using samtools [32]. The Picard [33] tool, MarkDuplicates (Broad Institute, Cambridge, MA, USA), was used to identify duplicates. To correct as many systematic errors in the sequencing process as possible, we performed base quality score recalibration. The Picard AddorReplacereAdgroups tool (Broad Institute, Cambridge, MA, USA) was used to modify the headers of BAM files for subsequent processing. Somatic single nucleotide variants (SNVs), and insertions and deletions (Indels), were detected by GATK Mutect2 (v4.1.9) (The Broad Institute of Harvard and MIT, Cambridge, MA, USA) [34] from the BAM files of paired tumor and normal samples. The GATK FilterMutectCalls (The Broad Institute of Harvard and MIT, Cambridge, MA, USA) tool was used to filter somatic mutations using default parameters, with true positive mutations marked with "PASS" and we selected "PASS" mutations.

Since affinity predictions for the peptide-MHC interface are MHC-specific, it is critical to know the patient HLA types. HLA alleles in each sample were inferred from trimmed RNA-seq data using OptiType (University of Tübingen, Tübingen, Germany) with the default settings, which has been demonstrated to achieve HLA typing with ~97% accuracy [35,36].

2.3. Mutation Annotation and Peptide Extraction

The annotation of mutations and the extraction of the peptide are shown in Figure 1. Somatic mutation data that were obtained based on RNA-seq data were annotated using Annovar [37] to identify noncoding-region somatic mutations, including SNVs and indels. The noncoding regions that we used specifically included: "downstream", "intergenic", "intronic", "ncRNA_exonic", "ncRNA_intronic", "ncRNA_splicing", "splicing", "upstream", "UTR3", and "UTR5". After mutation filtering, nucleotide sequences with a set interval length were obtained according to the mutation sites and reference genome using Bedtools (University of Virginia School of Medicine, Charlottesville, VA, USA) [38]. Specifically, 100 nucleotide sequences were taken from along each side of the mutation site. To generate the nucleotide sequences containing the mutation, we replaced the reference bases with mutant bases, i.e., we replaced the "REF" column with the "ALT" column in the mutation table.

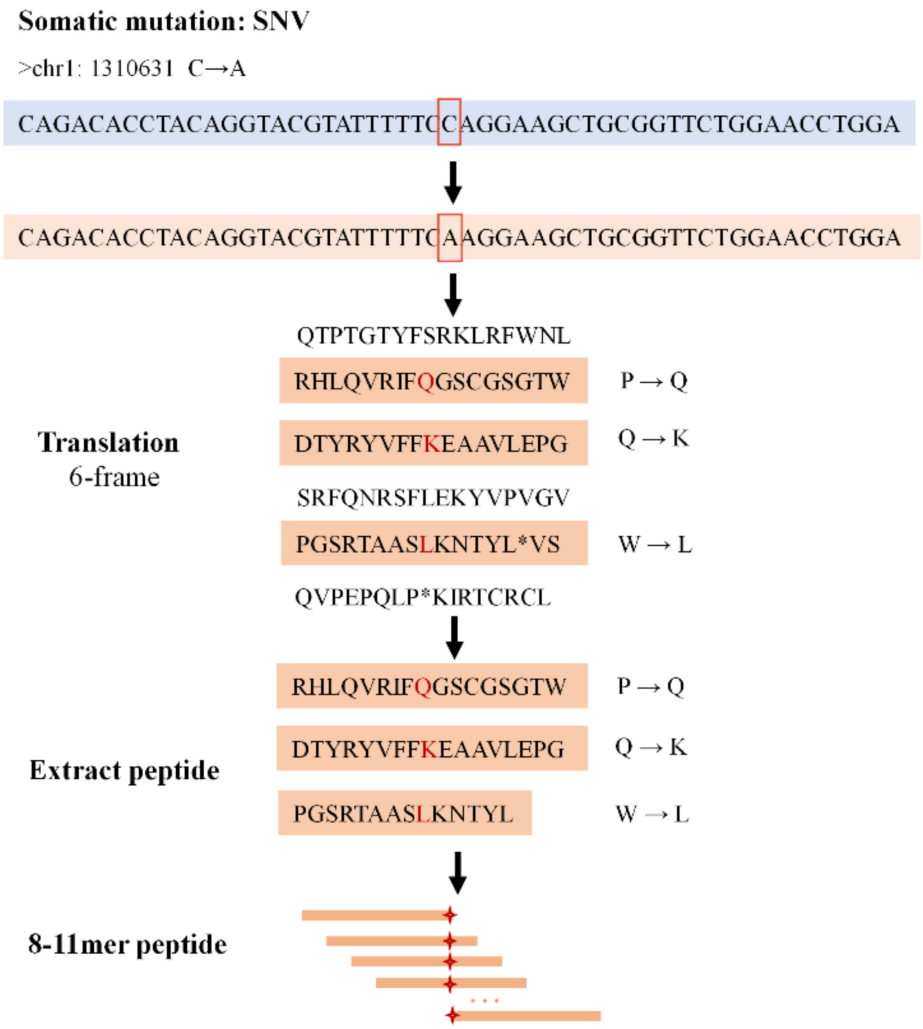

Figure 1. Generation of the variant peptides.

For the translation of nucleotide sequences in the genome, six-frame translation is the classical method. Six-frame translation has the advantage of being independent of any a priori annotation of the nucleotide sequence [39,40]. Thus, according to the 64 codons, the mutant nucleotide sequences were translated into novel proteins via six-frame translation. The termination codons were replaced with "*" and the protein sequences were cleaved into short peptides according to "*". The short peptides that did not contain the mutations were then filtered out. Finally, we obtained tumor protein sequences containing mutations from noncoding regions.

2.4. Database Construction and Peptide Identification

Identifying a mutant peptide expressed at the protein level is a crucial step. In this study, MaxQuant (Max-Planck Institute for Biochemistry, Martinsried, Germany) [41], a proteomics identification quantitative tool, was used to identify the peptides. To search the proteomics data, we first constructed a customized database for each individual tumor sample, including human reference proteins, common contaminant protein sequences in the laboratory (cRAP), and cancer-specific proteomes.

Then, to filter for true peptides, all MS/MS spectra were searched using MaxQuant in the customized peptide database. A separate target-decoy search strategy was used. Decoy peptides were generated from the peptides of corresponding target databases using a reversed tryptic approach. The parameters of MaxQuant were set as follows: (1) the variable modifications included protein N-terminal acetylation with methionine oxidation;

(2) strict trypsin specificity was required, allowing up to two missed cleavages; (3) the carbamidomethylation of cysteine was set as a fixed modification. In addition, false discovery rate (FDR) thresholds for protein peptides were specified at 1%. The minimum required peptide length was set to 7 amino acids. Finally, we extracted the cancer-specific mutant peptides identified by MS data and provided evidence in terms of protein expression level.

2.5. Neoantigen Prediction and Selection

PGNneo uses NetMHCpan 4.1 (Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark) [42] to calculate the binding affinity of peptides to patient-specific HLA alleles. To match the length of the peptides bound by MHC-I molecules, the peptides that were filtered by mass spectrometry were cleaved into short peptides containing mutated 8–11 mers. The percentile rank score was proven to exhibit higher sensitivity and less bias in HLA-peptide identification than the half-maximum inhibitory concentration (IC50) [43]. Therefore, the percentile rank (%rank) value was used as the metric of HLA-peptide binding prediction, and peptides with a %rank < 2 were considered to be candidate neoantigens. Similar sequences often originate from a common ancestral sequence and they are likely to have similar spatial structure and biological function; in fact, tumor-infiltrating T cells were found to exhibit a cross-reactivity that recognizes both tumor neoantigens and homologous non-tumor microbial antigens [44]. Therefore, to filter candidate neoantigens, sequence similarity analysis was performed using the basic local alignment search tool (BLAST) (National Center for Biotechnology Information, Bethesda, MD, USA) [45]. In total, 746 experimentally immunogenic neoantigens, collected from an in-house database, dbPepNeo2.0 [46], were used to build the target sequence database, while candidate neoantigens were treated as retrieval sequences. Then, BLASTp was used to identify homologous sequences so that the degree of homology between candidate neoantigens and immunogenic neoantigens could be established. We adjusted several default options to increase the sensitivity of BLAST searches that were performed with short input sequences. The peptides were reported to have sequence identity to immunogenic neoantigens if the percentage of identical matches was above 60%.

2.6. PGNneo Pipeline Implementation

PGNneo is open-source software written in Python, shell, and R. The software is divided into four toolkits, based on four modules. The user needs to configure the path and parameters of the software before applying the toolkits. After completing the configuration, the pipeline can be run by executing the command line. A more detailed tutorial on the use of PGNneo is available in the User's Manual. The PGNneo source code and documentation are available at https://github.com/tanxiaoxiu/PGNneo. To facilitate the installation and use of PGNneo, we provide a Docker container (Docker: https://hub.docker.com/r/xiaoxiutan/pgnneo) and a GUI.

3. Results

3.1. The Workflow of the PGNneo Pipeline

Here, as shown in Figure 2, a versatile and comprehensive workflow, PGNneo, is presented to identify neoantigens in the noncoding region. In PGNneo, several input datasets are required, including RNA-seq profiles and MS datasets. First, the RNA-seq profiles from the paired tumor and normal samples are used to screen for somatic mutations in the noncoding regions, and the amino acid sequences containing mutant sites in the noncoding region can thus be identified. Then, those expressed sequences can be filtered using MS datasets. Eventually, the resulting peptides are used for neoantigen prediction and selection by using MHC affinity and the database dbPepNeo2.0 [46].

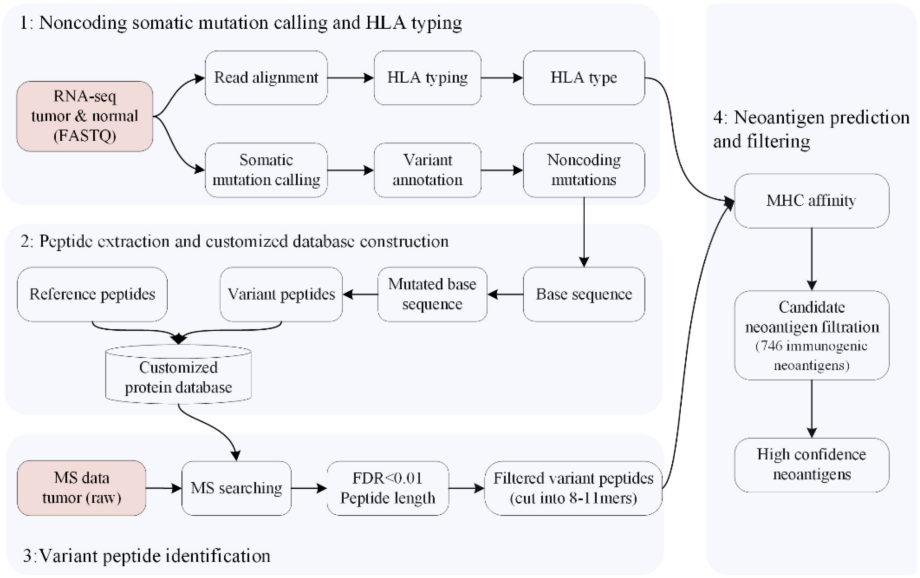

Figure 2. Overview of the PGNneo pipeline for proteogenomics-based noncoding neoantigen prediction.

The general computational framework of PGNneo consists of the following modules. (1) Noncoding somatic variant calling and HLA typing. Identifying somatic mutations in tumor cells is a key step in the neoantigen presentation pathway. For this purpose, paired tumor and normal samples were used for somatic variant calling. Prior to annotation, we removed any low-quality somatic mutations. Eventually, the noncoding mutations were extracted. The prediction of HLA typing was performed, based on the RNA-seq data from tumor samples. (2) Peptide extraction and customized database construction. Working according to the mutation information, the nucleotide sequences were obtained and were then translated into proteins by six-frame translation. The process of extraction of the peptide is shown in Figure 1. Eventually, tumor mutant peptides were obtained. Subsequently, these mutant protein sequences and reference proteins were combined to construct a customized protein database. (3) Variant peptide identification. The resulting peptides were filtered using MS datasets. The proteomic data provided evidence not only for the presence of peptides at protein levels but also for the binding of peptides to MHC molecules. (4) Neoantigen prediction and selection. Candidate neoantigens were predicted, according to peptides and HLA types, using NetMHCpan 4.1. The candidate neoantigens were filtered using the database dbPepNeo 2.0, which is an in-house dataset using 746 experimental immunogenic peptides as a reference. The resulting datasets at different filtering stages could then be obtained and downloaded according to the user's preference. Table 1 summarizes the software used in PGNneo.

Table 1. Summary of the tools available in PGNneo.

Module	Software	Function
(1) Noncoding somatic variant calling and HLA typing	Trimmomatic-0.39 [30]	Trims adapters and filters low-quality reads
	BWA-0.7.17 [31]	Sequence alignment
	SAMtools(V1.7) [32]	Converts .sam files to .bam, sort, and index files
	GATK4.2.0.0 [34]	Call somatic mutation
	Picard-2.23.9 [33]	Modifies the headers of .bam files
	OptiType-1.3.5 [35]	Predicts HLA typing

Table 1. *Cont.*

Module	Software	Function
(2) Peptide extraction and customized database construction	Annovar [37]	Mutation annotation
	Bedtools(v2.29.2) [38]	Sources the nucleotide sequence
(3) Variant peptide identification	MaxQuant [47]	Peptide identification
(4) Neoantigen prediction and selection	NetMHCpan-4.1 [42]	Calculates the binding affinity of peptides to patient-specific HLA alleles
	Blast-2.11.0+ [45]	Sequence similarity analysis

3.2. Evaluation of PGNneo Pipeline Results

To evaluate the performance of PGNneo, we performed sequence similarity analysis using two independent datasets. The positive dataset comprises 746 experimentally validated immunogenic neoantigens (Positive) collected from the dbPepNeo2.0 dataset, and the background control dataset contains 6400 mutant peptides (Random) of length 8–11 residue [48]. We performed sequence similarity analysis on the candidate neoantigens (unfiltered) that were obtained from the two cohorts of HCC_HF and HCC_HT with the positive dataset and random dataset, respectively. Two sequences exhibiting more than 60% of identical matches and an E-value of less than 0.5 are considered to be similar. Figure 3 shows that the results obtained from the positive and random datasets were significantly different in both the HCC_HF and HCC_HT cohorts, with p-values of 0.0278 (<0.05) and 0.01704 (<0.05), respectively (Wilcoxon test). The results indicate that the candidate neoantigens predicted by our method are more likely to have immunogenic potential. Therefore, filtering using the positive datasets was incorporated into the pipeline. This can be considered as an in silico verification step in the pipeline for the prediction of neoantigens.

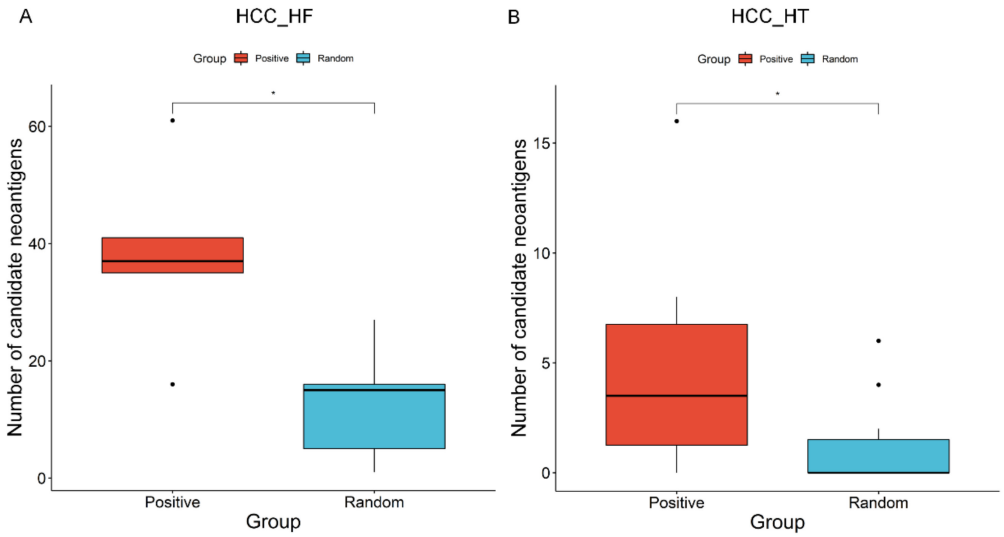

Figure 3. Evaluation of candidate neoantigens predicted by PGNneo, based on independent datasets: boxplot of the positive and random peptides in the HCC_HF cohort (**A**) and the HCC_HT cohort (**B**), with a * p-value < 0.05, ascertained by a Wilcoxon test.

3.3. Neoantigen Prediction, Selection, and Cross-Comparison from HCC Cohorts

PGNneo was applied to two independent HCC cohorts. We statistically analyzed the number of key steps in the pipeline for each sample **(**Figure S1 in the Supplementary Materials). In the HCC_HT cohort, one sample deviated significantly from other samples, possibly due to data quality problems, so this sample was deleted; finally, for this cohort, 9 samples were retained. The average number of key steps in the pipeline on the HCC_HF and HCC_HT cohorts are shown in Figure 4A. After filtering and annotation, an average of 3260 and 1178 noncoding mutations were obtained (Tables S2 and S3 in the Supplementary Materials). At the protein level, an average of 2339 and 610 peptides were identified by MS data analysis. HLA genotypes were predicted from the RNA-seq fastq file using the Optitype, and MHC-I binding predictions for the filtered peptides were predicted with netMHCpan4.1. As a result, an average of 3518 and 932 candidate neoantigens were obtained in the two cohorts, respectively (Tables S4 and S5 in the Supplementary Materials). After screening for HCC_HF, an average of 77 noncoding high-confidence neoantigens were eventually identified in each sample, ranging from 37 to 147 (Table S6 in the Supplementary Materials). However, in the HCC_HT cohort, an average of 13 noncoding high-confidence neoantigens were identified per sample, ranging from 4 to 23 (Table S7 in the Supplementary Materials). Upon comparing the results of the two cohorts, we found 403 overlapping non-coding mutations and 189 overlapping candidate neoantigens in the two cohorts (Figure 4B,C); however, no overlap was found in the high-confidence neoantigens. The results show that neoantigens are unique and the number of neoantigens varies greatly between different datasets.

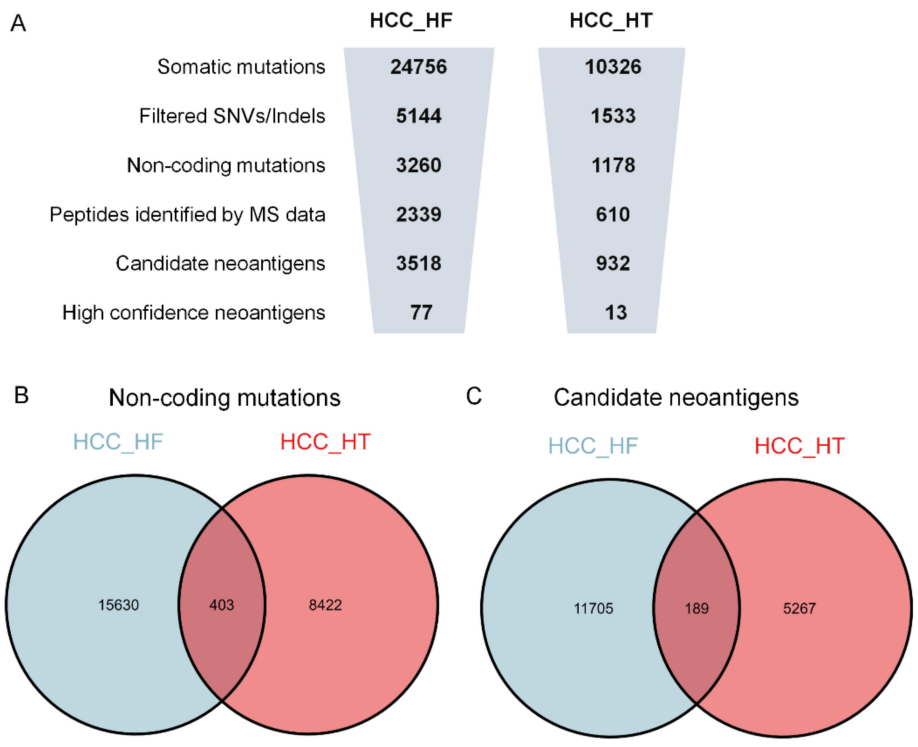

Figure 4. Results overview for the neoantigen discovery process. (**A**) The average number of key steps in the PGNneo on the HCC_HF cohort and HCC_HT cohort, respectively. (**B**) Overlapping non-coding mutations in the HCC_HF cohort and the HCC_HT cohort. (**C**) Overlapping candidate neoantigens in the HCC_HF cohort and the HCC_HT cohort.

In addition, 26 and 28 unique HLA alleles were predicted for the two cohorts, respectively (Table S8 in the Supplementary Materials). We further calculated the frequency of HLA alleles in the sample population using the getHlaFrequencies function in the mi-

dasHLA package of R [49]. Then, the mean value of neoantigens bound by each HLA allele was calculated, based on the count of HLA alleles, thus normalizing the number of neoantigens. The frequency of HLA alleles and the number of corresponding neoantigens are given in Table S9 in the Supplementary Materials. Based on the ranking of the normalized number of neoantigens, the 10 most frequent binding HLA alleles that matched with candidate neoantigens in two cohorts are shown in Figures 5A and 5B, respectively. The results showed that HLA alleles exhibited a preference for the binding of neoantigens, while HLA-A33:03 and HLA-A23:01 accounted for the largest binding proportion in the HCC_HF and HCC_HT cohorts. Moreover, the same candidate neoantigen can bind to different HLA alleles; this neoantigen is more likely to be present and may be applicable to a wider range of individuals.

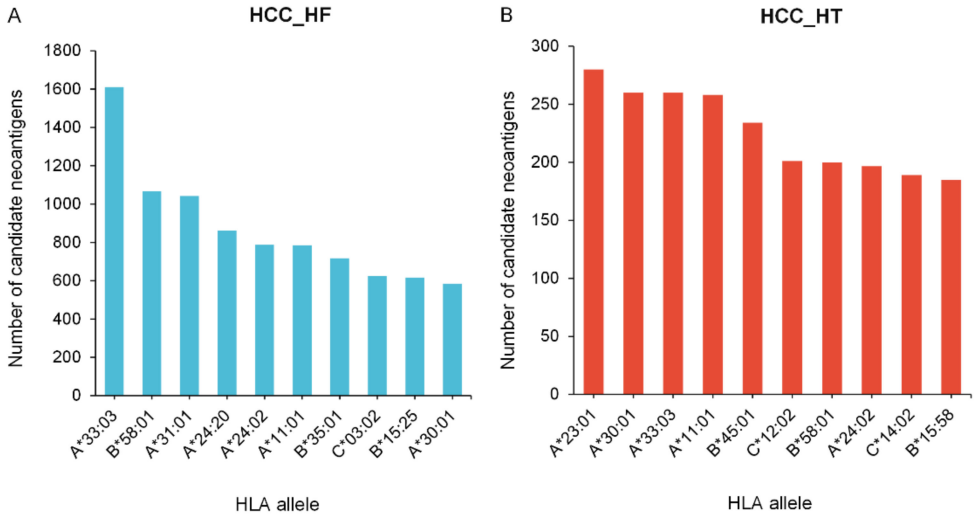

Figure 5. The number of predicted neoantigens binding with each HLA allele. (**A**) The top 10 HLA alleles matched with candidate neoantigens in the HCC_HF cohort. (**B**) The top 10 HLA alleles matched with candidate neoantigens in the HCC_HT cohort.

3.4. The Sharing of Noncoding Neoantigens and Genes in Different Samples

We further analyzed the overlapping neoantigens and their corresponding genes in the two cohorts. In the HCC_HF cohort (lncRNA-seq data from 5 HCC PDX samples and MS data from the HCC cell line), 10 neoantigens were found to be in common in at least 2 patients (Table S10 in the Supplementary Materials). These overlapping neoantigens are called "shared neoantigens" and have the potential to be designed as shared neoantigen vaccines. Conversely, neoantigens across the 5 patients were mapped to 118 unique genes, and 6 of these genes were observed in at least 2 patients (Table S10 in the Supplementary Materials). These overlapping genes are "hot-spot mutations" where the corresponding neoantigens may be in common among multiple patients. Unfortunately, no overlapping neoantigens or genes were found in the HCC_HT cohort (with paired RNA-seq data and MS data from 10 human HCC samples). This may be because the HCC_HF cohort dataset comprises unpaired lncRNA-seq data and MS data; the MS data is cell line data with lower heterogeneity, so that shared neoantigens can be found, while the HCC_HT cohort is of paired RNA-seq data and MS data from patients with a higher degree of individualization. To some extent, this explains the individualization of neoantigens in real patients. Therefore, this also reinforces the necessity for more research on hot-spot mutations for building up data resources for shared neoantigens.

3.5. Function Verification Analysis of Frequently Mutated Genes and Neoantigens in HCC

In order to correlate the predicted neoantigens with the clinical information garnered from patients, we explored the association between the predicted neoantigens and the

pathogenesis of HCC. Rao et al. [50] summarized the frequently mutated genes and their functions that are associated with HCC. We compared candidate neoantigen-associated genes in the HCC_HF cohort and HCC_HT cohort with the frequently mutated genes associated with HCC. The TP53, WWP1, ATM, KMT2C, and NFE2L2 mutant genes associated with HCC were identified in two cohorts and corresponded to 98 neoantigens (Table S11 in the Supplementary Materials). Table 2 only shows information about the 26 candidate neoantigens that bind most strongly to HLA. Among them, TP53 is one of the most studied tumor suppressors, with multiple functions, and is associated with DNA damage checkpoints and repair defects. WWP1 is associated with the activation of oncogenic pathways in HCC; the overexpression of WWP1 promotes tumorigenesis in HCC patients and predicts poor prognosis. The loss of ATM reduces hepatocyte apoptosis and fibrosis, suggesting that the activation of ATM in response to oxidative stress plays a role in hepatic fibrosis development. KMT2C and KMT2D are functionally similar and may be involved in chromatin localization and genomic instability. NFE2L2 deficiency may render cells susceptible to oxidative stress-mediated DNA damage. Genes that are highly mutated in HCC may be attractive potential therapeutic targets.

Table 2. Strongly bound neoantigens that are generated by frequently mutated genes in HCC.

Gene	Neoantigen	HLA	%Rank	Bind Level
WWP1	VSHDGATAL	HLA-C*03:04	0.009	SB
NFE2L2	KTDAQAISL	HLA-C*04:03	0.025	SB
TP53	TMAGQLLHV	HLA-A*02:06	0.052	SB
NFE2L2	SSRPAWPTR	HLA-A*33:03	0.08	SB
KMT2D	QQKNPSLFL	HLA-B*13:02	0.126	SB
NFE2L2	GQHSETPSL	HLA-B*15:01	0.154	SB
NFE2L2	WPGHQFFKY	HLA-B*35:01	0.155	SB
KMT2C	IVSSRFCTR	HLA-A*31:01	0.167	SB
KMT2D	QQKNPSLFLI	HLA-B*13:02	0.185	SB
NFE2L2	GIWPGHQFF	HLA-B*15:25	0.206	SB
NFE2L2	LFFETRSRF	HLA-A*24:02	0.226	SB
ATM	AEAGEPLEP	HLA-B*40:06	0.232	SB
WWP1	YRCVPPHPANF	HLA-C*06:02	0.258	SB
KMT2C	KLGDNHFFM	HLA-A*02:01	0.293	SB
NFE2L2	ATRTGRLWWR	HLA-A*31:01	0.323	SB
NFE2L2	HPKSKQISCTW	HLA-B*58:01	0.362	SB
NFE2L2	IWPGHQFF	HLA-A*24:02	0.376	SB
KMT2D	QKNPSLFLI	HLA-B*13:02	0.379	SB
NFE2L2	RMPVIQAAW	HLA-A*24:20	0.385	SB
NFE2L2	RMPVIQAAW	HLA-B*58:01	0.396	SB
WWP1	FSCLSLSGGW	HLA-B*58:01	0.432	SB
ATM	RACQRQAVGIK	HLA-A*30:01	0.433	SB
NFE2L2	KTDAQAISL	HLA-C*03:04	0.442	SB
NFE2L2	GQHSETPSLLK	HLA-A*11:01	0.46	SB
TP53	ATMAGQLLHV	HLA-A*02:06	0.474	SB
WWP1	VSHDGATAL	HLA-C*04:03	0.48	SB

3.6. Extended Application of PGNneo to Other Tumor Types

In addition, we applied PGNneo to another tumor type with moderate TMB, colorectal cancer (CRC) [29]. Firstly, 4206, 3664, and 5823 non-coding mutations were obtained on COLO205, SW620, and HCT116 cell line data, respectively, and further predictions yielded 217, 330, and 291 candidate neoantigens. In addition, to evaluate the potential immunogenicity of candidate neoantigens, we obtained high-confidence neoantigens based on the filtering of an experimentally validated immunogenic neoantigen database constructed by our group. Detailed results on sample information, noncoding region mutations, HLA typing, candidate neoantigens, and high-confidence neoantigens are provided in Table S12 in the Supplementary Materials. The results demonstrate that our pipeline can be applied

to multiple tumor types. The biological mechanisms of noncoding neoantigens may be cross-verified as the application of PGNneo expands.

3.7. Comparing PGNneo with Other Tools

To complete the identification and comparison of neoantigens from both coding and noncoding regions, we applied four other common neoantigen prediction tools, including ProGeo-neo [14], pVACtools [4], Neopredpipe [5], and TSAFinder [51], to compare their performance with our own tool, PGNneo. pVACtools, Neopredpipe, and TSAFinder require fastq data for RNA-seq and/or VCF data for mutations as input, and ProGeo-neo requires the additional input of MS data. For two HCC cohorts, we randomly selected three patient samples, HCC42, HCC67, and HCC1076, for comparison across five neoantigen prediction tools. Since most neoantigens are composed of 9 amino acids, we only compared the prediction of 9-mer neoantigens [52]. The number of candidate neoantigens obtained by the five tools is shown in Figure 6. It is worth noting that PGNneo introduces MS data and shows candidate neoantigens after MS filtering so that the results of PGNneo are more stringent. Although ProGeo-neo also has a module for MS data filtering, unfortunately, no neoantigens were obtained after MS data identification. This is consistent with studies on neoantigens in the coding region of HCC, which suggested that the tumor mutational burden (TMB) of HCC is relatively low and neoantigens in coding regions are scarce [53]. In contrast, the other tools do not have an MS filtering step, and we only show their candidate neoantigens as predicted by peptide-MHC binding affinity (Table S13 in the Supplementary Materials).

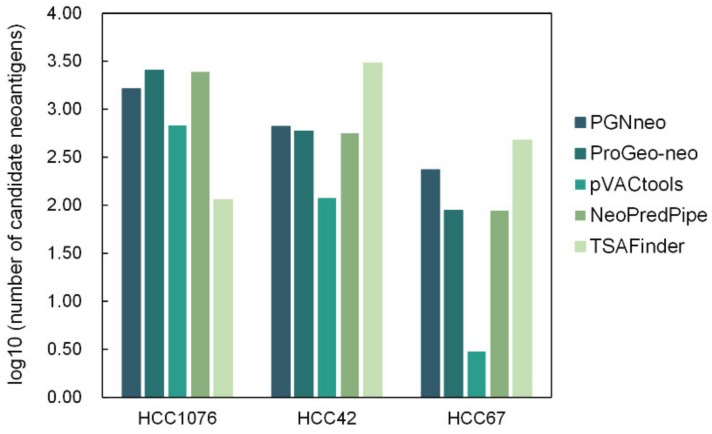

Figure 6. The number of candidate neoantigens predicted by PGNneo, ProGeo-neo, pVACtools, NeoPredPipe, and TSAFinder.

In addition, PGNneo sets up a module for secondary filtering by using 746 experimentally validated neoantigens, resulting in high-confidence neoantigens. Twenty, three, and one high-confidence 9-mer neoantigens were obtained in samples HCC1076, HCC42, and HCC67, respectively (Table S13 in the Supplementary Materials). Furthermore, we explored the association between these high-confidence neoantigens and the pathogenesis of HCC. The high-confidence neoantigen genes TNFSF14, GAD1, STARD1, and DHRS4-AS1 have been reported in the literature to be closely associated with HCC [54–56]. Specifically, TNFSF14 and GAD1 are highly expressed in HCC [54]; STARD1 promotes the progression of non-alcoholic steatohepatitis to HCC via bile acids [55]; DHRS4-AS1 ameliorates HCC by suppressing proliferation and promoting apoptosis via the miR-522-3p/SOCS5 axis [56].

Moreover, we analyzed the overlap of the candidate neoantigens predicted by different tools. For three samples, HCC1076, HCC42, and HCC67, the number of neoantigens identified by at least three tools was 411, 80, and 3, respectively (Table S14 in the Supplementary Materials). We recommend using multiple tools to predict neoantigens, which may yield more reliable results. Our investigation demonstrates that for cancer types

with a low TMB, the source of neoantigens may be enriched when the noncoding region is taken into consideration. Therefore, PGNneo aims to expand the scope of neoantigen prediction and provide a richer neoantigen reference for some cancer types with a low TMB in the coding region.

4. Discussion

Although some algorithms and tools have been developed to tackle the problem of neoantigen prediction, most are based on coding regions. However, in the human genome, 98% of the sequence involves noncoding regions, and most DNA sequence variants occur in the noncoding regions [17]. The general properties of sequence variants are also applicable to noncoding variants, such as SNVs and Indels. What is more, noncoding variants can also generate neoantigens. However, there are few prediction tools that operate on noncoding regions; the majority of tools focus on exonic variant calling, which is based on genomic data rather than transcriptomic data. For this reason, we developed a proteogenomics-based pipeline, PGNneo, to identify those neoantigens derived from noncoding regions, based on transcriptomic data from the human genome. Furthermore, we successfully validated the effectiveness of PGNneo through its application in two HCC cohorts. A total of 386 and 113 high-confidence neoantigens were identified in the two HCC cohorts, respectively. In addition, we applied PGNneo to a CRC cohort, demonstrating that the tool can be extended to multiple tumor types.

Compared with the traditional neoantigen prediction pipeline, PGNneo has several advantages. First, it focuses on neoantigens in noncoding regions, which provides a new source of neoantigens for low-TMB tumor types. Studies have shown that for cancer types with a low TMB, such as liver cancer, the source of neoantigens should be extended to noncoding regions for better applicability to immunotherapy [53]. Second, it combines transcriptomics and proteomics data, furthering the proteogenomics neoantigen prediction pipelines for coding regions in our research group [14,15]. Most of the previously developed neoantigen prediction tools are based on genomic data only, and the predicted false-positive rate of neoantigens is relatively high. Combined with proteomic data, the accuracy of neoantigen prediction can be improved. Moreover, proteogenomics holds the promise of providing deeper mechanistic insights to enable the better matching of patients to targeted therapies than when analyzing each type of omics data separately. In addition, our pipeline uses RNA-seq data from paired tumor and normal tissues to call somatic mutations. Mutations in the RNA-seq data provide a better reference for a proteomics dataset than WES, mainly because of the ability of RNA-seq to identify novel somatic variants, while for oncogenes that are highly expressed in cancer, RNA-seq provides higher sequencing coverage than WES and, therefore, has higher statistical confidence in detecting variants [57]. Thus, in our pipeline, the customized searchable peptide database was derived from tumor RNA-seq data.

Compared to the coding-region-based proteogenomic prediction pipeline established by our group, in terms of data requirements, ProGeo-neo requires data at the genomic, transcriptomic, and proteomic levels of the patients, while PGNneo only requires transcriptomic and proteomic data from patients, and not WES/WGS data. This is because the neoantigen prediction step in ProGneo-neo is performed based on the mutations detected by WES/WGS, whereas in PGNneo, the step is based on RNA-seq data. Therefore, the application scenario of the PGNneo can be further expanded.

There are still some limitations to our study, so we will expand on the following aspects. To further enrich the types of neoantigens, we may later update the tool to predict those neoantigens derived from gene fusion and RNA splicing, which will provide more potential neoantigen targets for developing therapeutic cancer vaccines. In addition, considering the complementary roles of coding-region-based and noncoding-region-based pipelines in identifying neoantigens, we will integrate neoantigen prediction tools such as ProGeo-neo that have already been developed within the group to further facilitate their use by researchers. Moreover, as noncoding neoantigens have been shown to have strong tumor

specificity, more relevant studies have since emerged in the field. Recently, Cai et al. [58] developed IEAtlas, an atlas of HLA-presented immune epitopes derived from noncoding regions, which provides a valuable data resource for studying the immunogenicity of noncoding epitopes. Therefore, we will combine the data from IEAtlas to further improve the predictive power of PGNneo. Even though the subsequent experimental validation of candidate neoantigens is essential for real-world clinical application, our computational methods can narrow down the range of neoantigens to a certain extent and thereby pave the way to improved preclinical vaccine design. Therefore, PGNneo may prove to be a useful tool for cancer researchers and clinicians.

5. Conclusions

In this study, we developed a proteogenomics-based pipeline to predict neoantigens in noncoding regions, namely, PGNneo. PGNneo is a pipeline that integrates state-of-the-art computational tools. It mainly includes four modules: (1) noncoding somatic variant calling and HLA typing; (2) peptide extraction and customized database construction; (3) variant peptide identification; (4) neoantigen prediction and selection. PGNneo can be easily applied to RNA-seq data and MS data drawn from patients of different cancer types. In summary, PGNneo can specifically detect the neoantigens generated by noncoding regions in tumors, providing guidance for cancer types with a low TMB in coding regions. PGNneo, together with our previous tool, can improve the identification of coding and noncoding region-derived neoantigens and will contribute to a more complete understanding of the tumor immune landscape. This capability holds promise for broadening the repertoire of candidates for therapeutic cancer vaccination and T cell-based therapy and may ultimately extend the neoantigen clinical benefits of immunotherapy.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cells12050782/s1, Figure S1: Results of key steps in neoantigen prediction based on PGNneo for each sample; Table S1: Detailed sample information of the HCC_HF cohort and HCC_HT cohort; Table S2: Noncoding mutations in the HCC_HF cohort; Table S3: Noncoding mutations in the HCC_HT cohort; Table S4: Candidate neoantigens in the HCC_HF cohort; Table S5: Candidate neoantigens in the HCC_HT cohort; Table S6: High-confidence neoantigens in the HCC_HF cohort; Table S7: High-confidence neoantigens in the HCC_HT cohort; Table S8: HLA alleles of each sample in the HCC_HF cohort and HCC_HT cohort; Table S9: The frequency of HLA alleles and the number of corresponding neoantigens in the HCC_HF cohort and HCC_HT cohort; Table S10: Overlap of noncoding neoantigens and noncoding genes between samples in the HCC_HF cohort; Table S11: Frequently mutated genes associated with HCC and the corresponding neoantigens; Table S12: Detailed results on sample information, noncoding region mutations, HLA typing, candidate neoantigens, and high-confidence neoantigens in the CRC cohort; Table S13: Candidate neoantigens, as predicted by five tools; Table S14: Overlap of candidate neoantigens, as predicted by five tools; User's Manual: Detailed tutorials for using the PGNneo tool.

Author Contributions: L.X. (Lu Xie) and X.T. conceived and designed this study and drafted the manuscript. L.X. (Lu Xie) and T.W. supervised this study. X.T. carried out data collection and analysis, built the pipeline, and wrote the draft manuscript. X.T., L.X. (Linfeng Xu) and J.O. wrote the software package. L.X. (Lu Xie) and X.J. revised the manuscript. X.Y. and B.H. contributed to data acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (31870829), Shanghai Municipal Health Commission Collaborative Innovation Cluster Project (2019CXJQ02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The noncoding mutation dataset for the lncRNA-seq of HCC samples is available in the Supplementary Materials. The proteomics datasets of the HCC cell line were obtained from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) with the identifier, PXD000529. Another HCC cohort associated with Jiang et al. [28], including RNA-seq files and MS raw data from 10 patients, was downloaded from the Gene Expression Omnibus

(GEO) (accession number GSE124535) and iProX database (http://www.iprox.org, accession number IPX0000937000), respectively. The CRC cohort [29], including RNA-seq data and MS raw data from three CRC cell lines and RNA-seq data from one normal fetal small intestine cell line, were downloaded from the GEO (accession number GSE195985) and ProteomeXchange Consortium (identifier PXD028309), respectively.

Acknowledgments: The authors would like to thank Michael Liebman for his critical reading and native English editing.

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

References

1. Coulie, P.G.; Van den Eynde, B.J.; van der Bruggen, P.; Boon, T. Tumour antigens recognized by T lymphocytes: At the core of cancer immunotherapy. *Nat. Rev. Cancer* **2014**, *14*, 135–146. [CrossRef] [PubMed]
2. Hu, Z.; Ott, P.A.; Wu, C.J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* **2018**, *18*, 168–182. [CrossRef] [PubMed]
3. Ott, P.A.; Hu, Z.; Keskin, D.B.; Shukla, S.A.; Sun, J.; Bozym, D.J.; Zhang, W.; Luoma, A.; Giobbie-Hurder, A.; Peter, L.; et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **2017**, *547*, 217–221. [CrossRef] [PubMed]
4. Hundal, J.; Kiwala, S.; McMichael, J.; Miller, C.A.; Xia, H.; Wollam, A.T.; Liu, C.J.; Zhao, S.; Feng, Y.Y.; Graubert, A.P.; et al. pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol. Res.* **2020**, *8*, 409–420. [CrossRef] [PubMed]
5. Schenck, R.O.; Lakatos, E.; Gatenbee, C.; Graham, T.A.; Anderson, A.R.A. NeoPredPipe: High-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinform.* **2019**, *20*, 264. [CrossRef]
6. Kim, S.; Kim, H.S.; Kim, E.; Lee, M.G.; Shin, E.C.; Paik, S.; Kim, S. Neopepsee: Accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann. Oncol.* **2018**, *29*, 1030–1036. [CrossRef]
7. Blass, E.; Ott, P.A. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nat. Rev. Clin. Oncol.* **2021**, *18*, 215–229. [CrossRef]
8. Müller, M.; Gfeller, D.; Coukos, G.; Bassani-Sternberg, M. 'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization. *Front. Immunol.* **2017**, *8*, 1367. [CrossRef]
9. Lei, J.T.; Zhang, B. Proteogenomics drives therapeutic hypothesis generation for precision oncology. *Br. J. Cancer* **2021**, *125*, 1–3. [CrossRef]
10. Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M.C.; Zimmerman, L.J.; Shaddox, K.F.; Kim, S.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513*, 382–387. [CrossRef]
11. Creech, A.L.; Ting, Y.S.; Goulding, S.P.; Sauld, J.F.K.; Barthelme, D.; Rooney, M.S.; Addona, T.A.; Abelin, J.G. The Role of Mass Spectrometry and Proteogenomics in the Advancement of HLA Epitope Prediction. *Proteomics* **2018**, *18*, e1700259. [CrossRef] [PubMed]
12. Bassani-Sternberg, M.; Braunlein, E.; Klar, R.; Engleitner, T.; Sinitcyn, P.; Audehm, S.; Straub, M.; Weber, J.; Slotta-Huspenina, J.; Specht, K.; et al. Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **2016**, *7*, 13404. [CrossRef] [PubMed]
13. Zhang, X.; Qi, Y.; Zhang, Q.; Liu, W. Application of mass spectrometry-based MHC immunopeptidome profiling in neoantigen identification for tumor immunotherapy. *Biomed. Pharmacother.* **2019**, *120*, 109542. [CrossRef] [PubMed]
14. Li, Y.; Wang, G.; Tan, X.; Ouyang, J.; Zhang, M.; Song, X.; Liu, Q.; Leng, Q.; Chen, L.; Xie, L. ProGeo-neo: A customized proteogenomic workflow for neoantigen prediction and selection. *BMC Med. Genom.* **2020**, *13*, 52. [CrossRef]
15. Liu, C.; Zhang, Y.; Jian, X.; Tan, X.; Lu, M.; Ouyang, J.; Liu, Z.; Li, Y.; Xu, L.; Chen, L.; et al. ProGeo-Neo v2.0: A One-Stop Software for Neoantigen Prediction and Filtering Based on the Proteogenomics Strategy. *Genes* **2022**, *13*, 783. [CrossRef]
16. Wen, B.; Li, K.; Zhang, Y.; Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **2020**, *11*, 1759. [CrossRef]
17. Khurana, E.; Fu, Y.; Chakravarty, D.; Demichelis, F.; Rubin, M.A.; Gerstein, M. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **2016**, *17*, 93–108. [CrossRef]
18. Djebali, S.; Davis, C.A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription in human cells. *Nature* **2012**, *489*, 101–108. [CrossRef]
19. Liao, K.; Xu, J.; Yang, W.; You, X.; Zhong, Q.; Wang, X. The research progress of LncRNA involved in the regulation of inflammatory diseases. *Mol. Immunol.* **2018**, *101*, 182–188. [CrossRef]
20. Ruiz-Orera, J.; Messeguer, X.; Subirana, J.A.; Alba, M.M. Long non-coding RNAs as a source of new peptides. *eLife* **2014**, *3*, e03523. [CrossRef]
21. Lu, S.; Zhang, J.; Lian, X.; Sun, L.; Meng, K.; Chen, Y.; Sun, Z.; Yin, X.; Li, Y.; Zhao, J.; et al. A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res.* **2019**, *47*, 8111–8125. [CrossRef] [PubMed]
22. Laumont, C.M.; Daouda, T.; Laverdure, J.P.; Bonneil, E.; Caron-Lizotte, O.; Hardy, M.P.; Granados, D.P.; Durette, C.; Lemieux, S.; Thibault, P.; et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat. Commun.* **2016**, *7*, 10238. [CrossRef] [PubMed]

23. Laumont, C.M.; Perreault, C. Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell. Mol. Life Sci.* **2018**, *75*, 607–621. [CrossRef] [PubMed]
24. Ehx, G.; Larouche, J.D.; Durette, C.; Laverdure, J.P.; Hesnard, L.; Vincent, K.; Hardy, M.P.; Thériault, C.; Rulleau, C.; Lanoix, J.; et al. Atypical acute myeloid leukemia-specific transcripts generate shared and immunogenic MHC class-I-associated epitopes. *Immunity* **2021**, *54*, 737–752.e710. [CrossRef] [PubMed]
25. Laumont, C.M.; Vincent, K.; Hesnard, L.; Audemard, E.; Bonneil, E.; Laverdure, J.P.; Gendron, P.; Courcelles, M.; Hardy, M.P.; Cote, C.; et al. Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **2018**, *10*, 470. [CrossRef]
26. Xiang, R.; Ma, L.; Yang, M.; Zheng, Z.; Chen, X.; Jia, F.; Xie, F.; Zhou, Y.; Li, F.; Wu, K.; et al. Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens. *Commun. Biol.* **2021**, *4*, 496. [CrossRef]
27. Hu, B.; Li, H.; Guo, W.; Sun, Y.F.; Zhang, X.; Tang, W.G.; Yang, L.X.; Xu, Y.; Tang, X.Y.; Ding, G.H.; et al. Establishment of a hepatocellular carcinoma patient-derived xenograft platform and its application in biomarker identification. *Int. J. Cancer* **2020**, *146*, 1606–1617. [CrossRef] [PubMed]
28. Jiang, Y.; Sun, A.; Zhao, Y.; Ying, W.; Sun, H.; Yang, X.; Xing, B.; Sun, W.; Ren, L.; Hu, B.; et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **2019**, *567*, 257–261. [CrossRef]
29. Cleyle, J.; Hardy, M.P.; Minati, R.; Courcelles, M.; Durette, C.; Lanoix, J.; Laverdure, J.P.; Vincent, K.; Perreault, C.; Thibault, P. Immunopeptidomic analyses of colorectal cancers with and without microsatellite instability. *Mol. Cell. Proteom. MCP* **2022**, *21*, 100228. [CrossRef]
30. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
31. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
32. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
33. "Picard Toolkit" Broad Institute. GitHub Repository. 2019. Available online: https://broadinstitute.github.io/picard/ (accessed on 1 September 2020).
34. 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]
35. Szolek, A.; Schubert, B.; Mohr, C.; Sturm, M.; Feldhahn, M.; Kohlbacher, O. OptiType: Precision HLA typing from next-generation sequencing data. *Bioinformatics* **2014**, *30*, 3310–3316. [CrossRef] [PubMed]
36. Yi, J.; Chen, L.; Xiao, Y.; Zhao, Z.; Su, X. Investigations of sequencing data and sample type on HLA class Ia typing with different computational tools. *Brief. Bioinform.* **2021**, *22*, bbaa143. [CrossRef] [PubMed]
37. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef] [PubMed]
38. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
39. Nesvizhskii, A.I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114–1125. [CrossRef]
40. Zickmann, F.; Renard, B.Y. MSProGene: Integrative proteogenomics beyond six-frames and single nucleotide polymorphisms. *Bioinformatics* **2015**, *31*, i106–i115. [CrossRef]
41. 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*, 1367–1372. [CrossRef]
42. 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] [PubMed]
43. 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*, 33. [CrossRef] [PubMed]
44. Balachandran, V.P.; Luksza, M.; Zhao, J.N.; Makarov, V.; Moral, J.A.; Remark, R.; Herbst, B.; Askan, G.; Bhanot, U.; Senbabaoglu, Y.; et al. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. *Nature* **2017**, *551*, 512–516. [CrossRef] [PubMed]
45. McGinnis, S.; Madden, T.L. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **2004**, *32*, W20–W25. [CrossRef] [PubMed]
46. Lu, M.; Xu, L.; Jian, X.; Tan, X.; Zhao, J.; Liu, Z.; Zhang, Y.; Liu, C.; Chen, L.; Lin, Y.; et al. dbPepNeo2.0: A Database for Human Tumor Neoantigen Peptides From Mass Spectrometry and TCR Recognition. *Front. Immunol.* **2022**, *13*, 855976. [CrossRef] [PubMed]
47. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319. [CrossRef]
48. Kosaloglu-Yalcin, Z.; Lanka, M.; Frentzen, A.; Logandha Ramamoorthy Premlal, A.; Sidney, J.; Vaughan, K.; Greenbaum, J.; Robbins, P.; Gartner, J.; Sette, A.; et al. Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncoimmunology* **2018**, *7*, e1492508. [CrossRef]

49. Migdal, M.; Ruan, D.F.; Forrest, W.F.; Horowitz, A.; Hammer, C. MiDAS-Meaningful Immunogenetic Data at Scale. *PLoS Comput. Biol.* **2021**, *17*, e1009131. [CrossRef]
50. Rao, C.V.; Asch, A.S.; Yamada, H.Y. Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. *Carcinogenesis* **2017**, *38*, 2–11. [CrossRef]
51. Sharpnack, M.F.; Johnson, T.S.; Chalkley, R.; Han, Z.; Carbone, D.; Huang, K.; He, K. TSAFinder: Exhaustive tumor-specific antigen detection with RNAseq. *Bioinformatics* **2022**, *38*, 2422–2427. [CrossRef]
52. Li, L.; Goedegebuure, S.P.; Gillanders, W.E. Preclinical and clinical development of neoantigen vaccines. *Ann. Oncol.* **2017**, *28*, xii11–xii17. [CrossRef] [PubMed]
53. Lu, L.; Jiang, J.; Zhan, M.; Zhang, H.; Wang, Q.T.; Sun, S.N.; Guo, X.K.; Yin, H.; Wei, Y.; Liu, J.O.; et al. Targeting Neoantigens in Hepatocellular Carcinoma for Immunotherapy: A Futile Strategy? *Hepatology* **2021**, *73*, 414–421. [CrossRef] [PubMed]
54. Li, X.; Ramadori, P.; Pfister, D.; Seehawer, M.; Zender, L.; Heikenwalder, M. The immunological and metabolic landscape in primary and metastatic liver cancer. *Nat. Rev. Cancer* **2021**, *21*, 541–557. [CrossRef] [PubMed]
55. Conde de la Rosa, L.; Garcia-Ruiz, C.; Vallejo, C.; Baulies, A.; Nuñez, S.; Monte, M.J.; Marin, J.J.G.; Baila-Rueda, L.; Cenarro, A.; Civeira, F.; et al. STARD1 promotes NASH-driven HCC by sustaining the generation of bile acids through the alternative mitochondrial pathway. *J. Hepatol.* **2021**, *74*, 1429–1441. [CrossRef]
56. Zhou, Y.; Li, K.; Zou, X.; Hua, Z.; Wang, H.; Bian, W.; Wang, H.; Chen, F.; Dai, T. LncRNA DHRS4-AS1 ameliorates hepatocellular carcinoma by suppressing proliferation and promoting apoptosis via miR-522-3p/SOCS5 axis. *Bioengineered* **2021**, *12*, 10862–10877. [CrossRef]
57. Coudray, A.; Battenhouse, A.M.; Bucher, P.; Iyer, V.R. Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data. *PeerJ* **2018**, *6*, e5362. [CrossRef]
58. Cai, Y.; Lv, D.; Li, D.; Yin, J.; Ma, Y.; Luo, Y.; Fu, L.; Ding, N.; Li, Y.; Pan, Z.; et al. IEAtlas: An atlas of HLA-presented immune epitopes derived from non-coding regions. *Nucleic Acids Res.* **2022**, *51*, D409–D417. [CrossRef] [PubMed]