Computational and Structural Biotechnology Journal 19 (2021) 4510–4516

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

TSNAD v2.0: A one-stop software solution for tumor-specific neoantigen detection

Zhan Zhou [a,d,1,*], Jingcheng Wu [a,b,1], Jianan Ren [a], Wenfan Chen [a,f], Wenyi Zhao [a,b], Xun Gu [c], Ying Chi [e], Qiaojun He [a,d], Bo Yang [a,d], Jian Wu [b,d,g,*], Shuqing Chen [a,*]

[a] Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China
[b] Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[c] Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA
[d] Innovation Institute for Artificial Intelligence in Medicine, Zhejiang University, Hangzhou 310018, China
[e] Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Alibaba DAMO Academy, Hangzhou 311121, China
[f] Polytechnic Institute, Zhejiang University, Hangzhou 310015, China
[g] Real Doctor AI Research Centre, School of Medicine, Zhejiang University, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 1 April 2021
Received in revised form 10 August 2021
Accepted 10 August 2021
Available online 12 August 2021

Keywords:
Neoantigens
Somatic mutations
Major histocompatibility complex
Tumor immunotherapy
One-stop software

ABSTRACT

TSNAD is a one-stop software solution for predicting neoantigens from the whole genome/exome sequencing data of tumor-normal pairs. Here we present TSNAD v2.0 which provides several new features such as the function of RNA-Seq analysis including gene expression and gene fusion analysis, the support of different versions of the reference genome. Most importantly, we replace the NetMHCpan with DeepHLApan we developed previously, which considers both the binding between peptide and major histocompatibility complex (MHC) and the immunogenicity of the presented peptide-MHC complex (pMHC). TSNAD v2.0 achieves good performamce on a standard dataset. For better usage, we provide the Docker version and the web service of TSNAD v2.0. The source code of TSNAD v2.0 is freely available at https://github.com/jiujiezz/tsnad. And the web service of TSNAD v2.0 is available at http://biopharm.zju.edu.cn/tsnad/.

1. Introduction

Tumor neoantigens which are derived from mutated proteins and presented on the surface of cancer cells are tumor-specific antigens absent from normal cells [1]. To avoid confusion, we treat the complex of mutated peptides and the major histocompatibility complex (peptide-MHC pairs, pMHCs) as tumor neoantigens in this study. Tumor neoantigens have been well acknowledged as the ideal targets for cancer immunotherapies, such as cancer vaccines and T-cell immunotherapies [2–5]. To provide an easy solution for neoantigen prediction, we have previously developed an integrated software for tumor-specific neoantigen detection (TSNAD) [6], which can provide one-stop neoantigen prediction from original whole-exome sequencing (WES) or whole-genome sequencing (WGS) data of normal/tumor samples. TSNAD v1.0 has been adopted by many other groups [7–12] and remains continuously updating.

We present here TSNAD v2.0 that implements new features and improvements including (i) update all the embedded tools into the latest version, (ii) add the function of RNA-Seq data analysis including gene expression and gene fusion analyses, (iii) support two versions of reference genome (GRCh37 and GRCh38) when calling mutations, (iv) add the neoantigen prediction derived from INDELs and gene fusions, (v) replace NetMHCpan with our developed tool DeepHLApan and provide a web service of TSNAD, (vi) provide the installation method of Docker which comprises all the needed tools and reference files.

2. Methods

As illustrated in Fig. 1, the TSNAD v2.0 consists of five modules.

* Corresponding authors at: College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China (Z. Zhou, S. Chen). College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (J. Wu).
E-mail addresses: zhanzhou@zju.edu.cn (Z. Zhou), wujian2000@zju.edu.cn (J. Wu), chenshuqing@zju.edu.cn (S. Chen).
[1] Authors who contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.08.016

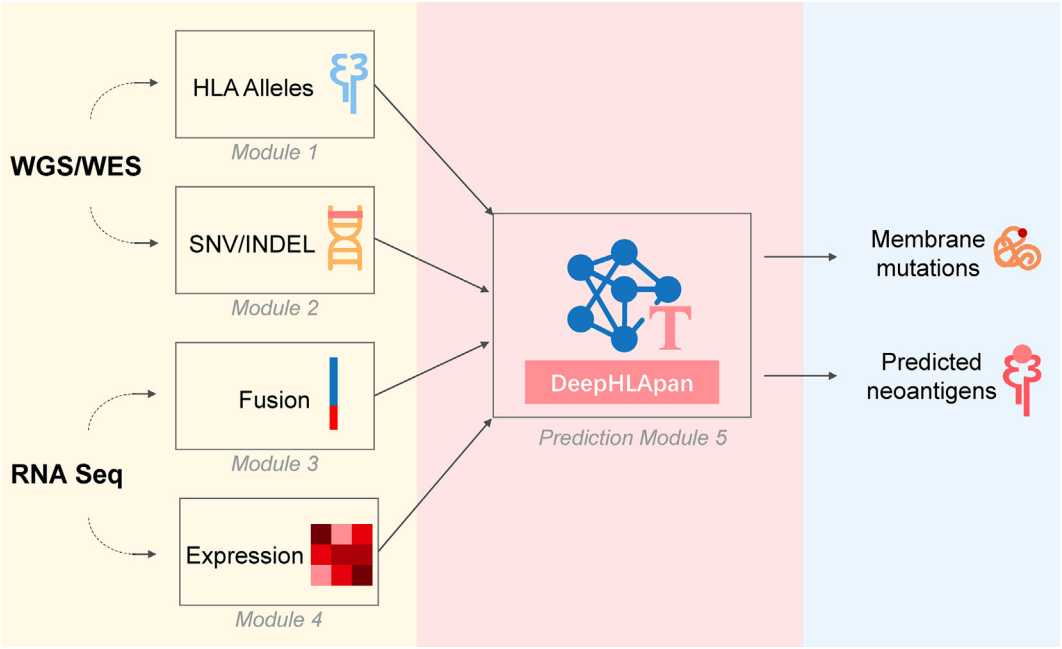

Fig. 1. Workflow of TSNAD v2.0. Including SNV/INDEL detection, HLA allele detection, gene fusion detection, gene expression detection, and neoantigen prediction. WGS, whole-genome sequence. WES, whole-exon sequence. SNV, single nucleotide variant. HLA, human leukemia antigen.

2.1. Module 1: SNV/INDEL detection

Firstly, TSNAD v2.0 uses Trimmomatic (v0.39) [13] to trim and crop raw WES/WGS reads and remove artifacts that will be harmful to the subsequent data processing. Then, BWA (v0.7.17) [14] is used for mapping short sequences to a reference genome (GRCh37 or GRCh38). SAMtools (v1.13) [15] is used to transform sequencing data format from SAM (sequence alignment/map) to BAM (binary alignment/map), which could save an enormous amount of storage space. GATK (v4.2.0.0) [16] is used to remove repeat sequences and recalibrate the base quality score to create the final BAM file. The Mutect2 module of GATK is used to call SNVs/INDELs.

2.2. Module 2: HLA allele detection

The input data used for human leukocyte antigen (HLA) allele detection is the cleaned fastq files created by Trimmomatic in the first module. OptiType (v1.3.5) [17] is used for HLA allele identification.

2.3. Module 3: Fusion detection

The detection of fusions is based on the RNA-Seq data. STAR (v2.7) aligner [18] is used for mapping reads to the reference genome, Arriba (v1.1.0) [19] is then used for fusion detection. Compared with other fusion detectors such as STAR-fusion [20], Arriba is both quicker and more accurate.

2.4. Module 4: Gene expression detection

The gene expression detection is achieved by HISAT2 (v2.2.1) [28], SAMtools (v1.13), and Stringtie (v2.1.6) [29]. HISAT2 is used to map next-generation sequencing reads to human genomes, SAMtools plays the same role as in module 1 and Stringtie is used to assemble the RNA-Seq alignments into potential transcripts and provide the expression result.

2.5. Module 5: Neoantigen prediction

Neoantigen prediction module is the most important part of TSNAD v2.0. With the identified SNVs/INDELs, VEP (v104) [21] is used to annotate them to obtain the mutation information at the amino acid level. Take the advantage of in-home scripts, all mutant peptides cover the mutations with length ranges from 8 to 11 are then extracted. Combining the mutant peptides with detected HLA alleles, DeepHLApan (v1.1) [22] is used for the final neoantigen prediction. In general, all the peptide-MHC pairs with binding score >0.5 and immunogenic score > 0.5 are potential neoantigens. Further, we suggest that the peptide-MHC pairs with immunogenic score >0.5 and binding score rank top 20 across the different lengths of peptides are high-confidence neoantigens within one sample (i.e. 80 high-confidence neoantigens each sample). The prediction process of fusion-derived neoantigens is similar, with the difference that the neo-peptides generates by gene fusions are provided by Arriba. The gene expression detected in module 4 is a filter to remove the neoantigens generated by the mutations in unexpressed genes (TPM < 1), which could improve the accuracy of the final prediction results.

The extracellular mutations of membrane proteins are not the typical neoantigens as defined, however, they could also be potential tumor-specific targets so we retained the function of extracellular mutations identification that TSNAD v1.0 provides. The SNVs/INDELs located in the extracellular domains of membrane proteins would be stored in a separate file.

3. Results

3.1. Features update from TSNAD v1.0

The first version of TSNAD takes advantage of the best practices of the GATK [16] and NetMHCpan [23] to predict neoantigens from the whole genome/exome sequencing data of tumor-normal pairs, comprising other tools such as Trimmomatic [13], BWA [14], SAMtools [15], Picard [24], ANNOVAR [25]. SOAP-HLA [26] and TMHMM [27]. As time goes by, some of the tools are updated or

Table 1
The number of SNVs, INDELs, and Fusions of five TESLA samples, respectively.

Reference genome	ID	#SNVs	#INDELs	#Fusions
GRCh38	1_TESLA	122	16	11
	2_TESLA	440	24	18
	3_TESLA	826	19	10
	12_TESLA	62	14	12
	16_TESLA	59	14	6
GRCh37	1_TESLA	128	15	5
	2_TESLA	450	24	19
	3_TESLA	862	20	10
	12_TESLA	65	15	8
	16_TESLA	58	13	10

Table 2
The 23 validated peptide-MHC pairs from 400 predicted neoantigens under the reference genome GRCh38. The detailed information of the 'Mutation' column includes the gene name, mutation position in a protein, and the position of mutated amino acid in the peptide.

ID	Mutation	HLA	Peptide	Binding score	Immunogenic score	Rank	Validation
1_TESLA	ACE_S167F_1	HLA-A02:01	FLDPDLTNI	0.9972	0.7211	16	TRUE
2_TESLA	PBRM1_K1072E_5	HLA-B57:01	KSFKEIKLW	0.9999	0.9160	0	TRUE
2_TESLA	ATP13A3_S899F_1	HLA-A02:01	FLSELEASV	0.9996	0.7721	1	FALSE
2_TESLA	SEC61A1_R231W_9	HLA-B57:01	RTDKVRALW	0.9995	0.9772	2	FALSE
2_TESLA	CFAP20_P74S_8	HLA-B57:01	KTLGIKLSF	0.9994	0.9450	3	FALSE
2_TESLA	ME1_E227K_4	HLA-A02:01	FLDKFMEAV	0.9992	0.7303	4	FALSE
2_TESLA	G6PD_A149V_9	HLA-A02:01	ALPPTVYEV	0.9976	0.7410	14	TRUE
2_TESLA	OPA1_V585G_5	HLA-B57:01	LSLAGSDCFW	0.9996	0.9944	7	FALSE
3_TESLA	SLC4A2_P363L_4	HLA-B08:01	WGKLHVASL	0.9979	0.7932	8	FALSE
3_TESLA	PLXNA3_E1312K_3	HLA-A03:01	GIKAHPVLK	0.9979	0.6438	9	FALSE
3_TESLA	KDM6B_P776L_3	HLA-A03:01	ALLPPPPLAK	0.9993	0.5449	10	FALSE
16_TESLA	NAA25_R14L_8	HLA-C05:01	VQDPNDRLL	0.9999	0.9719	3	FALSE
16_TESLA	HMGB3_P96R_2	HLA-B27:05	RRPSGFFLF	0.9995	0.7617	8	FALSE
16_TESLA	POFUT2_V241L_5	HLA-B27:05	RRSMLFARH	0.9994	0.9010	12	TRUE
16_TESLA	PSD4_R868H_5	HLA-C05:01	TADWHLYLF	0.9993	0.9647	15	FALSE
16_TESLA	EBF4_G327R_2	HLA-B27:05	KRCPGRFVY	0.9992	0.8832	17	FALSE
16_TESLA	COL5A2_P1266T_4	HLA-C05:01	KTDTGVHATL	1.0000	0.9491	0	TRUE
16_TESLA	HMGB3_P96R_3	HLA-B27:05	KRRPSGFFLF	1.0000	0.7072	1	FALSE
16_TESLA	SNX7_R234I_8	HLA-B27:05	SRMGQTVIAV	0.9999	0.5118	5	FALSE
16_TESLA	PLXDC1_V290F_10	HLA-B27:05	HRIELDPSKF	0.9999	0.9010	6	FALSE
16_TESLA	POFUT2_V241L_5	HLA-B27:05	RRSMLFARHL	0.9999	0.9052	7	FALSE
16_TESLA	PSD4_R868H_7	HLA-B27:05	LRTADWHLYL	0.9999	0.9148	8	FALSE
16_TESLA	ATR_D1243Y_3	HLA-B27:05	NRYAVQDFLH	0.9998	0.6015	13	FALSE

deprecated so we improve their version or replace them with other tools, respectively. The detailed adjustment is as follows: (i) Trimmomatic, BWA, SAMtools, and GATK are updated to the latest version. (ii) Picard is embedded in the latest version of GATK as a module. (iii) ANNOVAR is replaced by the latest version of VEP [21], which is more suitable in our pipeline. (iv) NetMHCpan is replaced by DeepHLApan [22] which would be discussed later.

TSNAD v1.0 only supports the reference genome GRCh37, we add the selection of GRCh38 when identifying neoantigens by replacing SOAP-HLA with OptiType which is able to detect HLA alleles under both reference genomes [17]. Normally, the expression of genes in tumors should be considered when predicting the potential neoantigens of patients since no neoantigens would be generated from unexpressed genes. We provide the new function of RNA-Seq analysis by combining HISAT2 [28] and Stringtie [29] into the pipeline to detect gene expression at the transcription level. Besides, we also provide the new function to predict neoantigens derived from mutations more than single nucleotide variants (SNVs), such as INDELs and gene fusions. The INDEL calling is also called by the Mutect2 module of GATK. The gene fusion analysis is achieved by combining STAR [18] and Arriba [19]. To note, the gene expression and gene fusion analysis would only be performed with RNA-Seq data.

The core part of TSNAD is the prediction that whether one pair of peptide-MHC is potential neoantigen and NetMHCpan is used to achieve this purpose in TSNAD v1.0. However, the mechanism of how neoantigens inducing T-cell response is complex and more than peptide-MHC binding which is the function that NetMHCpan provides. So we replace NetMHCpan with our recently developed tool DeepHLApan [22] for neoantigen prediction, which considers both peptide-MHC binding and the potential immunogenicity of the presented peptide-MHC pairs. It significantly reduced the false positives when predicting neoantigens and was more suitable to be embedded in the updated version of TSNAD.

3.2. TSNAD v2.0 achieves high performance on TESLA data

Standard datasets are of critical importance to provide a benchmark to evaluate software performance. Recently, the Tumor Neoantigen Selection Alliance (TESLA), a global community provide a standard dataset for the comparison of neoantigen prediction tools [30]. In their study, 608 predicted neoantigens, which are derived from six patients and predicted by tools from 28 unique teams are tested for immunogenicity and 37 (6.1%) of them were found to be immunogenic.

Due to the data availability, we obtained the WES, tumor RNA-Seq, and clinical-grade HLA typing of five patients (three melanoma patients with ID 1_TESLA, 2_TESLA, and 3_TESLA and two non-small cell lung carcinoma patients with ID 12_TESLA and 16_TESLA) from Synapse with identifier syn21048999. Under different selection of the reference genome, the number of mutations of each patient identified by TSNAD v2.0 is similar (Table1). Results

Table 3
The 30 validated peptide-MHC pairs from 400 predicted neoantigens under the reference genome GRCh37. The detailed information of the 'Mutation' column includes the gene name, mutation position in a protein, and the position of mutated amino acid in the peptide.

ID	Mutation	HLA	Peptide	Binding score	Immunogenic score	Rank	Validation
1_TESLA	ACE_S167F_1	HLA-A02:01	FLDPDLTNI	0.9972	0.7211	19	TRUE
2_TESLA	ATP13A3_S899F_1	HLA-A02:01	FLSELEASV	0.9996	0.7721	1	FALSE
2_TESLA	SEC61A1_R231W_9	HLA-B57:01	RTDKVRALW	0.9995	0.9772	2	FALSE
2_TESLA	C16orf80_P74S_8	HLA-B57:01	KTLGIKLSF	0.9994	0.9450	3	FALSE
2_TESLA	ME1_E227K_4	HLA-A02:01	FLDKFMEAV	0.9992	0.7303	4	FALSE
2_TESLA	ABCC1_T645I_1	HLA-B57:01	IVRNATFTW	0.9991	0.9437	5	FALSE
2_TESLA	OPA1_V585G_5	HLA-B57:01	LSLAGSDCFW	0.9996	0.9944	7	FALSE
2_TESLA	ABCC1_T645I_2	HLA-B57:01	IIVRNATFTW	0.9995	0.9404	10	FALSE
2_TESLA	PBRM1_K1047E_5	HLA-B57:01	KSFKEIKLW	0.9999	0.9160	0	TRUE
2_TESLA	G6PD_A179V_9	HLA-A02:01	ALPPTVYEV	0.9976	0.7410	15	TRUE
3_TESLA	SLC4A2_P363L_4	HLA-B08:01	WGKLHVASL	0.9979	0.7932	8	FALSE
3_TESLA	PLXNA3_E1312K_3	HLA-A03:01	GIKAHPVLK	0.9979	0.6438	9	FALSE
3_TESLA	KDM6B_P776L_3	HLA-A03:01	ALLPPPPLAK	0.9993	0.5449	10	FALSE
12_TESLA	CYP27A1_G367C_8	HLA-A02:01	ALHEEVVCV	0.9973	0.7153	11	FALSE
12_TESLA	MFSD7_N179D_4	HLA-A02:01	LVADVLSPV	0.9956	0.5472	15	FALSE
12_TESLA	HELZ2_A2160V_9	HLA-A02:01	KLNPSQNVV	0.9947	0.6994	19	FALSE
12_TESLA	A1BG_H362R_6	HLA-A02:01	ALFELRNISV	0.9982	0.5865	14	FALSE
16_TESLA	NAA25_R14L_8	HLA-C05:01	VQDPNDRLL	0.9999	0.9719	3	FALSE
16_TESLA	HMGB3_P96R_2	HLA-B27:05	RRPSGFFLF	0.9995	0.7617	8	FALSE
16_TESLA	PLCL1_P827H_5	HLA-B27:05	YRHVHLRSF	0.9994	0.9491	10	FALSE
16_TESLA	PSD4_R868H_5	HLA-C05:01	TADWHLYLF	0.9993	0.9647	14	FALSE
16_TESLA	EBF4_G327R_2	HLA-B27:05	KRCPGRFVY	0.9992	0.8832	16	FALSE
16_TESLA	HMGB3_P96R_3	HLA-B27:05	KRRPSGFFLF	1.0000	0.7072	1	FALSE
16_TESLA	SNX7_R234I_8	HLA-B27:05	SRMGQTVIAV	0.9999	0.5118	5	FALSE
16_TESLA	PLXDC1_V290F_10	HLA-B27:05	HRIELDPSKF	0.9999	0.9010	6	FALSE
16_TESLA	POFUT2_V241L_5	HLA-B27:05	RRSMLFARHL	0.9999	0.9052	7	FALSE
16_TESLA	PSD4_R868H_7	HLA-B27:05	LRTADWHLYL	0.9999	0.9148	8	FALSE
16_TESLA	ATR_D1243Y_3	HLA-B27:05	NRYAVQDFLH	0.9998	0.6015	13	FALSE
16_TESLA	POFUT2_V241L_5	HLA-B27:05	RRSMLFARH	0.9994	0.9010	11	TRUE
16_TESLA	COL5A2_P1266T_4	HLA-C05:01	KTDTGVHATL	1.0000	0.9491	0	TRUE

Table 4
The performance of TSNAD v2.0 under different selection criterias.

Reference genome	Selection criteria	#Tested	#TRUE	Accuracy
GRCh38	Top10	18	2	11.1%
	Top20	23	5	21.7%
	Top30	27	5	18.5%
	Top40	33	6	18.2%
	Top50	40	6	15.0%
GRCh37	Top10	17	2	11.8%
	Top20	30	5	16.7%
	Top30	35	5	14.3%
	Top40	43	5	11.6%
	Top50	49	6	12.2%

show that melanoma patients have more SNVs than non-small cell lung carcinoma patients while having a similar number of INDELs and gene fusions.

80 high-confidence neoantigens for each sample and 400 in total for five samples are predicted by TSNAD v2.0. However, among the 400 predicted neoantigens, only 23/30 peptide-MHC pairs have been tested by TESLA, and 5/5 of them (21.7%/16.7%) are validated as immunogenic under the reference genome GRCh38/GRCh37, respectively (Tables 2, 3, S1 and S2). Besides, we also evaluate the prediction performance of TSNAD v2.0 when selecting peptide-MHC pairs ranked top 10, 30, 40, and 50 as high-confidence neoantigens, respectively. The poorest performance of TSNAD v2.0 by selecting the top 10 peptide-MHC pairs under the reference genome GRCh38 (11.1%) is also greater than 6.1%, the overall positive rate of TESLA dataset (Table 4). We also evaluated the predicted binding scores for the 608 tested peptide-MHC pairs by applying TSNAD v2.0. The predicted binding scores of immunogenic peptide-MHC pairs are significantly greater than non-immunogenic peptide-MHC pairs no matter in all TESLA dataset (608 peptide-MHC pairs, Wilcoxon test, $p = 1.2 \times 10^{-7}$) or the peptide-MHC pairs with predicted immunogenic scores greater than 0.5 (241 peptide-MHC pairs, Wilcoxon test, $p = 0.026$) (Fig. 2). All the results support the predicted reliability of TSNAD v2.0.

3.3. Improved usability of TSNAD v2.0

For better usage of TSNAD v2.0, we provide the Docker version of TSNAD v2.0 for local installation and usage. It's easy to install TSNAD v2.0 by one command: docker pull biopharm/tsnad:latest

And given a directory *samples* including WES/WGS files (the files would be better to rename as *normal_R1.fastq.gz*, *normal_R2.fastq.gz*, *tumor_R1.fastq.gz*, and *tumor_R2.fastq.gz*), the following commands could achieve the purpose of neoantigen prediction:

docker run -it -v [dir of WES/WGS]/:/home/tsnad/samples -v [dir of RNA-Seq]:/home/tsnad/RNA-seq -v [output dir]:/home/tsnad/results biopharm/tsnad:latest /bin/bash

cd /home/tsnad

bash uncompress.sh

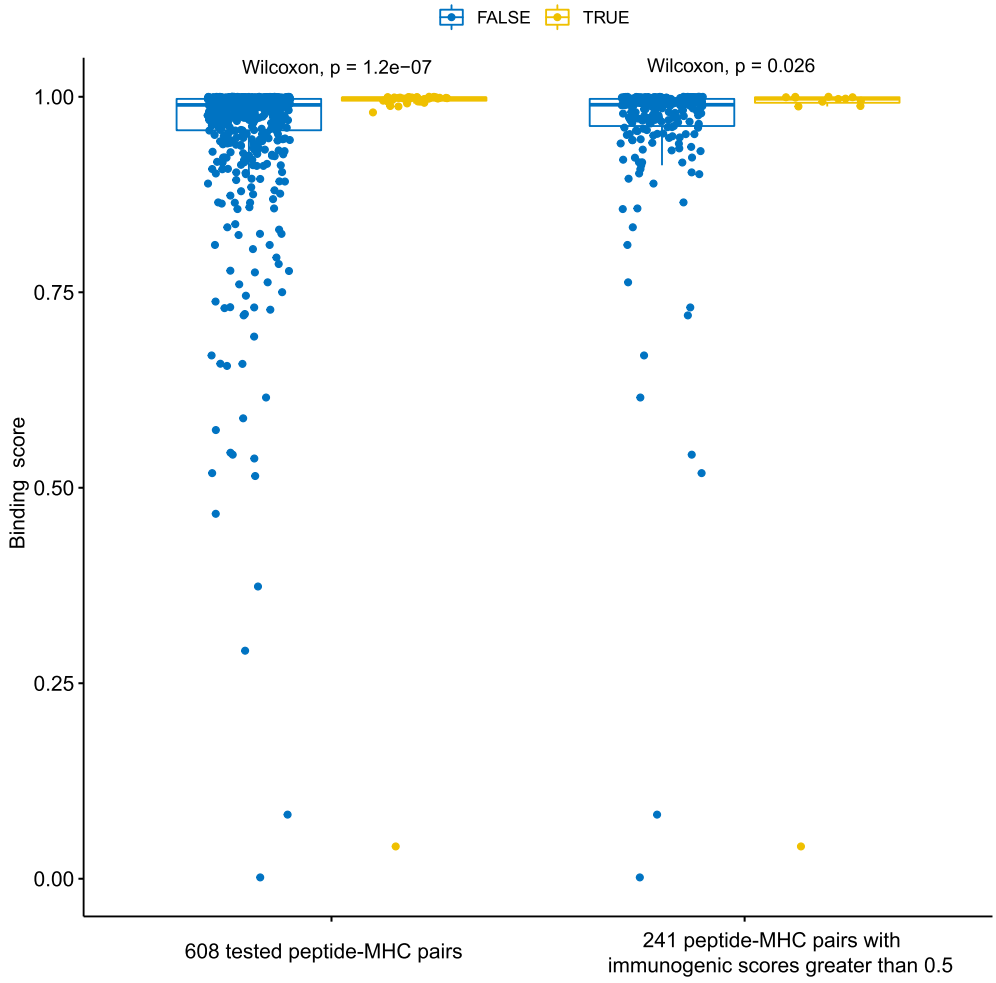

Fig. 2. The comparison of predicted binding score between immunogenic peptide-MHC pairs and non-immunogenic peptide-MHC pairs. There are 37 immunogenic peptide-MHC pairs and 571 non-immunogenic peptide-MHC pairs in all TESLA dataset, among which 12 immunogenic peptide-MHC pairs and 229 non-immunogenic peptide-MHC pairs with predicted immunogenic score greater than 0.5.

python TSNAD.py -I samples/ -R RNA-seq/ -V [grch37/grch38] -O results/

All the results would be stored in the directory *results*, and the predicted neoantigen would be stored in the directory *deephlapan_results*.

To generate useable neoantigen predictions, the minimum depth should be 15 × for WGS and 50 × for WES, the recommended depth should be 30 × for WGS and 100 × for WES. For sample TESLA_1 with WES tumor/normal data and RNA-Seq data, it takes about 50 h to finish neoantigen prediction in the Ubuntu system with 64G memory and 512G hard disk space.

Besides, we also provide a web service of TSNAD v2.0 which supports neoantigens prediction given mutations and HLA alleles (Fig. 3). The mutations should be formatted in the VCF file which is the file format of mutation results from the widely used best practice of GATK. We will support more file formats if required in the future. Compared with stand-alone pipelines, the web service of TSNAD v2.0 only including DeepHLApan and some in-house scripts for processing data (i.e. the function that module 5 provides), it's more suitable for the situation that users already have mutation file and HLA alleles and do not need to analyze the potential fusion-derived neoantigens or gene expression. We would try to provide the full function of TSNAD in the future update of the web service by taking advantage of cloud storage and cloud computing.

4. Conclusions

TSNAD v2.0 could provide one-stop neoantigen prediction from original WES/WGS and RNA-Seq data of normal/tumor samples. In this version, we update or replace most of the embedded tools with the latest versions or more suitable tools. Besides, it supports the choice of different versions of reference genomes when calling mutations, and also provides the analysis of INDELs, gene fusions, and gene expression. The most important change is the integration of DeepHLApan, which is the deep learning based prediction method for both peptide-MHC binding and the potential immunogenicity of the presented peptide-MHC complex. And the web service and Docker version of TSNAD v2.0 would provide an easy solution for tumor-specific neoantigen prediction.

Though DeepHLApan considers both binding and immunogenicity of peptide-MHC pairs, more effort should be put into T-cell receptor (TCR)-pMHC interaction to understand the mechanism that pMHC inducing T-cell response for predicting high-confidence neoantigens. Therefore, the future update of TSNAD should be a more complex and precise neoantigen prediction pipeline containing TCR sequencing and TCR-pMHC binding prediction model.

Code availability

The source code of TSNAD v2.0 is freely available at https://github.com/jiujiezz/tsnad. The Docker version of TSNAD v2.0 is

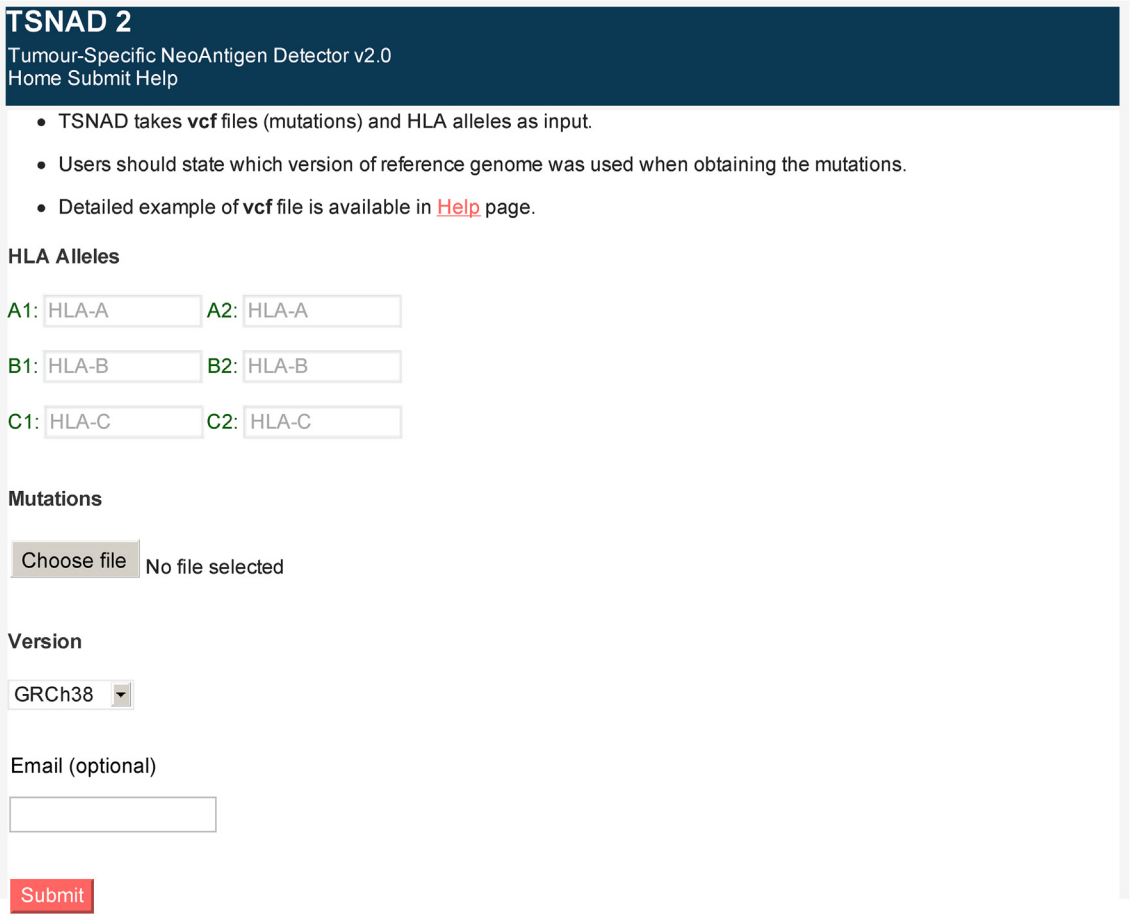

Fig. 3. The web service of TSNAD v2.0. Users could provide the mutations (VCF format) and HLA alleles to predict potential neoantigens. The version of the reference genome is selectable and email is optional.

available at https://hub.docker.com/r/biopharm/tsnad. The web service of TSNAD v2.0 is available at http://biopharm.zju.edu.cn/tsnad.

CRediT authorship contribution statement

Zhan Zhou: Conceptualization, Funding acquisition, Methodology, Software, Supervision, Writing – original draft, Writing - review & editing. **Jingcheng Wu:** Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing - review & editing. **Jianan Ren:** Investigation, Software. **Wenfan Chen:** Formal analysis. **Wenyi Zhao:** Formal analysis. **Xun Gu:** Conceptualization, Methodology. **Ying Chi:** Conceptualization, Methodology. **Qiaojun He:** Resources, Validation. **Bo Yang:** Resources, Validation. **Jian Wu:** Conceptualization, Supervision, Writing - review & editing. **Shuqing Chen:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Grant No. U20A20409, 31971371), the Key R&D Program of Zhejiang Province (Grant No. 2020C03010), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LY19H300003), the Fundamental Research Funds for the Central Universities of China, and Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare. We also thank the Research and Service Center, College of Pharmaceutical Sciences, Zhejiang University, and Alibaba Cloud for technical assistance, and Ms. Wenlin Zhang for language editing which has greatly improved the manuscript. We gratefully acknowledge the clinical contributors and data producers from the TESLA for referencing the TESLA dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.016.

References

[1] Lee CH, Yelensky R, Jooss K, Chan TA. Update on tumor neoantigens and their utility: why it is good to be different. Trends Immunol 2018;39:536–48. https://doi.org/10.1016/j.it.2018.04.005.
[2] Lu YC, Robbins PF. Cancer immunotherapy targeting neoantigens. Semin Immunol 2016;28:22–7. https://doi.org/10.1016/j.smim.2015.11.002.
[3] Tran E, Robbins PF, Lu Y-C, Prickett TD, Gartner JJ, Jia L, et al. T-cell transfer therapy targeting mutant KRAS in cancer. N Engl J Med 2016;375:2255–62. https://doi.org/10.1056/NEJMoa1609279.
[4] Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Löwer M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature 2017;547:222–6. https://doi.org/10.1038/nature23003.

[5] Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 2017;547:217–21. https://doi.org/10.1038/nature22991.
[6] Zhou Z, Lyu X, Wu J, Yang X, Wu S, Zhou J, et al. TSNAD: An integrated software for cancer somatic mutation and tumour-specific neoantigen detection. R Soc Open Sci 2017;4:. https://doi.org/10.1098/rsos.170050170050.
[7] Chen K, Ye H, Lu X, Jie, Sun B, Liu Q. Towards In silico prediction of the immune-checkpoint blockade response. Trends Pharmacol Sci 2017;38:1041–51. https://doi.org/10.1016/j.tips.2017.10.002.
[8] González S, Volkova N, Beer P, Gerstung M. Immuno-oncology from the perspective of somatic evolution. Semin Cancer Biol 2018;52:75–85. https://doi.org/10.1016/j.semcancer.2017.12.001.
[9] Wu J, Zhao W, Zhou B, Su Z, Gu X, Zhou Z, et al. TSNAdb: A database for tumor-specific neoantigens from immunogenomics data analysis. Genom. Proteom. Bioinforma 2018;16:276–82. https://doi.org/10.1016/j.gpb.2018.06.003.
[10] Zhou C, Zhu C, Liu Q. Toward in silico identification of tumor neoantigens in immunotherapy. Trends Mol Med 2019;25:980–92. https://doi.org/10.1016/j.molmed.2019.08.001.
[11] Smith CC, Selitsky SR, Chai S, Armistead PM, Vincent BG, Serody JS. Alternative tumour-specific antigens. Nat Rev Cancer 2019;19:465–78. https://doi.org/10.1038/s41568-019-0162-4.
[12] Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. Best practices for bioinformatic characterization of neoantigens for clinical utility. Genome Med 2019;11:56. https://doi.org/10.1186/s13073-019-0666-2.
[13] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
[14] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95. https://doi.org/10.1093/bioinformatics/btp698.
[15] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27:2987–93. https://doi.org/10.1093/bioinformatics/btr509.
[16] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297–303. https://doi.org/10.1101/gr.107524.110.
[17] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 2014;30:3310–6. https://doi.org/10.1093/bioinformatics/btu548.
[18] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21. https://doi.org/10.1093/bioinformatics/bts635.
[19] Uhrig S, Ellermann J, Walther T, Burkhardt P, Fröhlich M, Hutter B, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Res 2021:gr.257246.119. 10.1101/gr.257246.119.
[20] Haas BJ, Dobin A, Li B, Stransky N, Pochet N, Regev A. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Genome Biol 2019;20:1–16. https://doi.org/10.1186/s13059-019-1842-9.
[21] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4.
[22] Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. Front Immunol 2019;10:2559. https://doi.org/10.3389/fimmu.2019.02559.
[23] Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class i binding prediction beyond humans. Immunogenetics 2009;61:1–13. https://doi.org/10.1007/s00251-008-0341-z.
[24] Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–501. https://doi.org/10.1038/ng.806.
[25] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:1–7. https://doi.org/10.1093/nar/gkq603.
[26] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009;25:1966–7. https://doi.org/10.1093/bioinformatics/btp336.
[27] Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Int. Conf. Intell. Syst. Mol. Biol. ; ISMB., vol. 6, 1998, p. 175–82. 9783223.
[28] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60. https://doi.org/10.1038/nmeth.3317.
[29] Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33:290–5. https://doi.org/10.1038/nbt.3122.
[30] Wells DK, van Buuren MM, Dang KK, Hubbard-Lucey VM, Sheehan KCF, Campbell KM, et al. Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. Cell 2020;183:818–834.e13. https://doi.org/10.1016/j.cell.2020.09.015.