NARWHAL

NARWHAL Tutorial

Table of Contents

General Instruction

NARWHAL is a comprehensive online tool that aims to identify potential neoantigens for personalized cancer treatment strategies. The name NARWHAL stands for **Neoantigens Recognition Website and HLA Genotyping Tool**, reflecting its purpose. This web-based toolkit is designed to analyze various sample types, such as **DNA-seq**, **RNA-seq**, and **LC-MS/MS data**, with optimized steps to identify both **mutated tumor-specific antigens (mTSAs)**, **aberrantly expressed TSAs (aeTSAs), and tumor-associated antigens (TAAs)**. Additionally, we have developed a web-based application for predicting HLA genotypes to enable more efficient identification of neoantigens for cancer immunotherapy.

The availability of this tool for both neoantigen recognition and HLA genotype analysis is essential for the development of cancer immunotherapies. The header section of NARWHAL features a set of seven interactive buttons, each directing users to distinct pages within the website. The homepage provides users with a comprehensive overview of the employed pipeline, enabling them to commence neoantigen identification analysis through the **"Start Neoantigen"** button or the **"Neoantigen Identification"** option located at the top (Fig. 1).

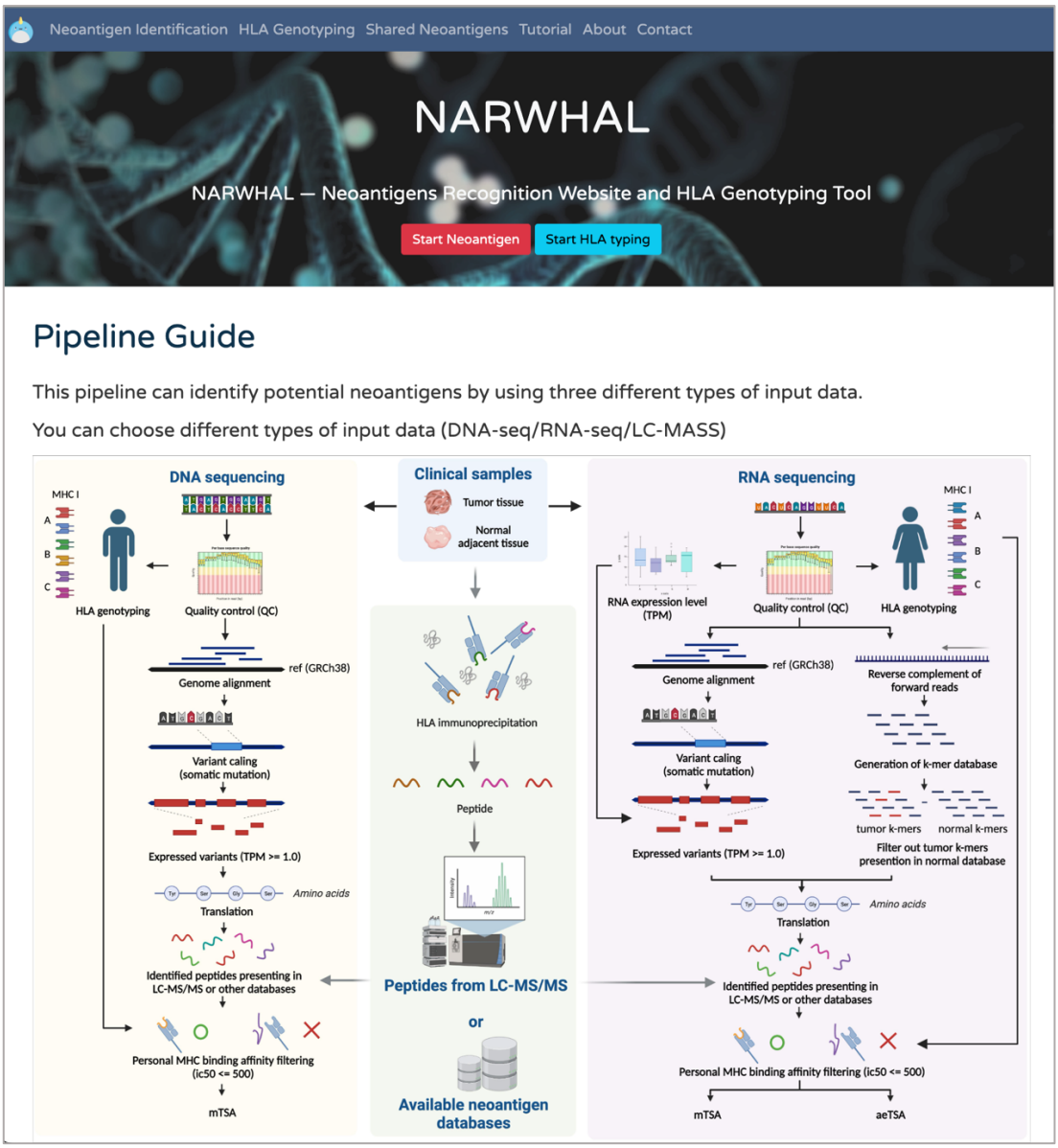


Fig. 1: The homepage of NARWHAL. There is a header section at the top allows users to select different functions.

There are step-by-step instructions on each pipeline's home page. Please follow them. For all the pipelines, including (1) neoantigen identification, (2) HLA genotyping, and (3) shared neoantigen discovery, you have to input your basic information first (i.e., the email address) (Fig. 2), so that NARWHAL can send a unique link to check the status or the final report of the submitted task. The unique link for your task is generated for security reasons. With the correct email address and link, you can come back to view

Email entry

Fig. 2: The section for email entry, enabling users to securely provide and verify their email addresses. This crucial step ensures accurate notification via email.

Neoantigen Identification

Data type selection

There are three different pipelines of neoantigen identification, including using DNA-seq, RNA-seq or both of them. Users select their input data types at first (Fig. 3).

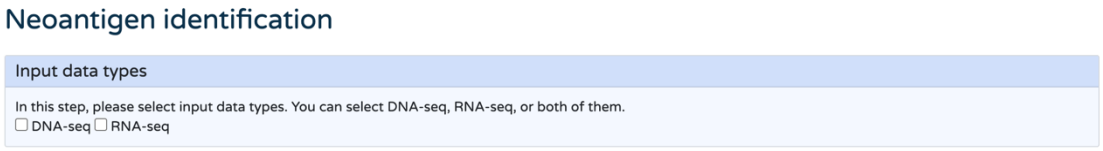


Fig. 3: The section is designed for selecting their input data types.

Data upload

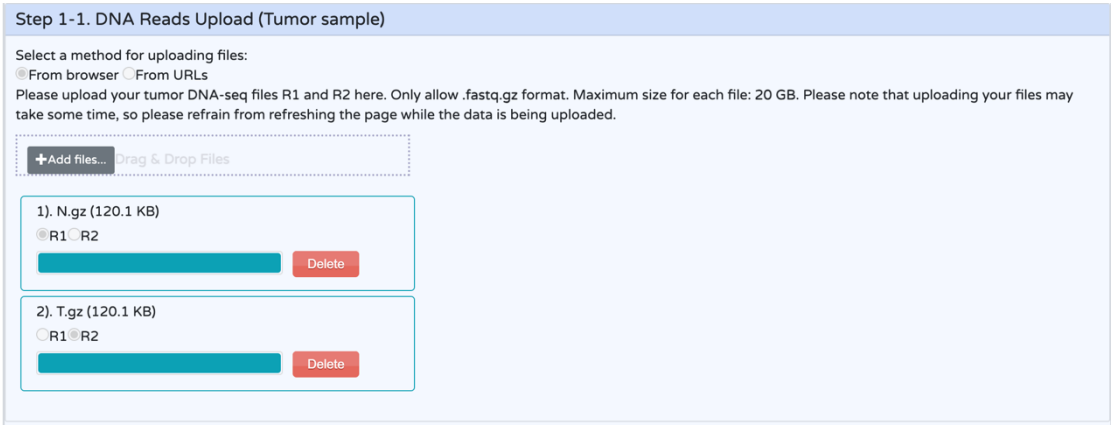
The DNA or/and RNA R1/R2 files of tumor and adjacent normal samples should be uploaded separately in ***.fastq.gz*** format, and the maximum size for each file is 20 GB (Fig. 4). Noted that total number of files is different in various data types (Table 1).

Table 1. Necessary uploaded read files in different combination.

Method	Tumor DNA		Normal DNA		Tumor RNA		Normal RNA		LC-MS/MS
	R1	R2	R1	R2	R1	R2	R1	R2	
Only DNA	o	o	o	o					optional
Only RNA					o	o	o	o	optional
DNA+RNA	o	o	o	o	o	o	o	o	optional

(a)

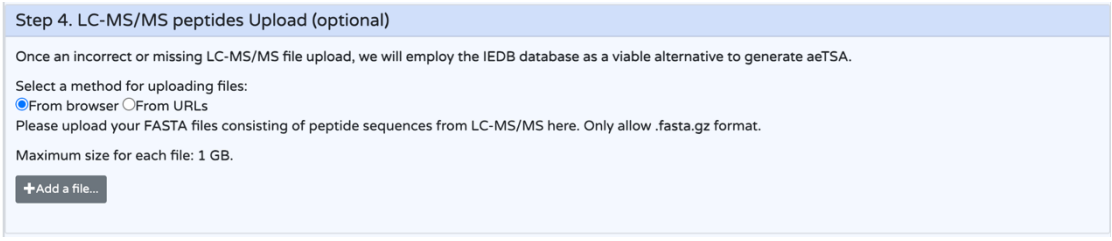
(b)



(c)

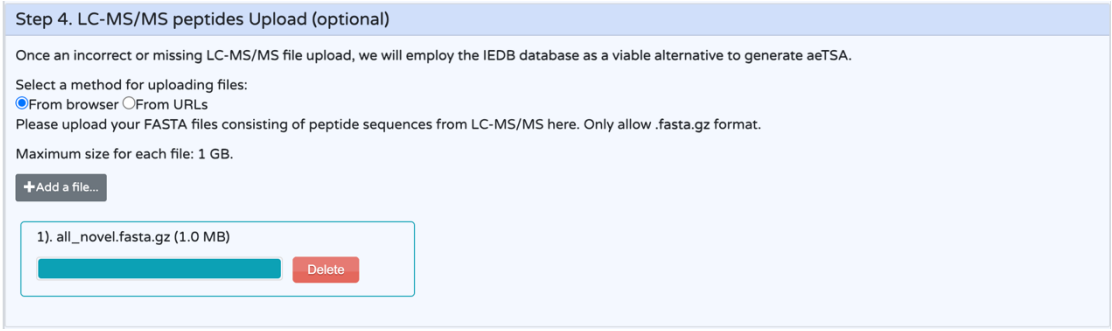
Fig. 4: The section designed for uploading sequencing data (using tumor DNA-seq as an example) offers users the flexibility to employ two distinct methods: (a) direct data upload through their browser interface and (b) file upload using specific Google Drive file IDs. The figure (c) demonstrates the successful upload of data, indicating that users have successfully uploaded their files. The files can be deleted via the **"Delete"** buttons. Note that uploading files may take some time (~a few minutes per file), so please do not refresh the page; otherwise, the files will need to be uploaded again.

Besides, LC-MS/MS files (optional) should be uploaded in ***.fasta.gz*** format, and the maximum size is 1 GB (Fig. 5). Users can choose to upload through their browser or with URLs (e.g. google drive links).



(a)

(b)

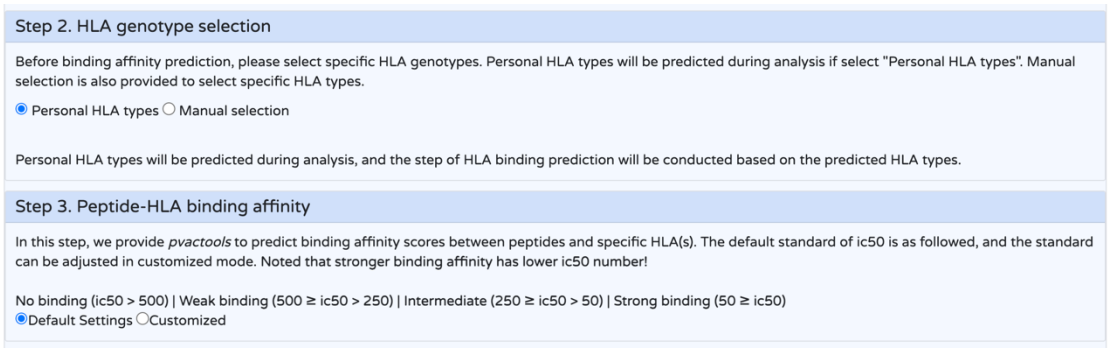


(c)

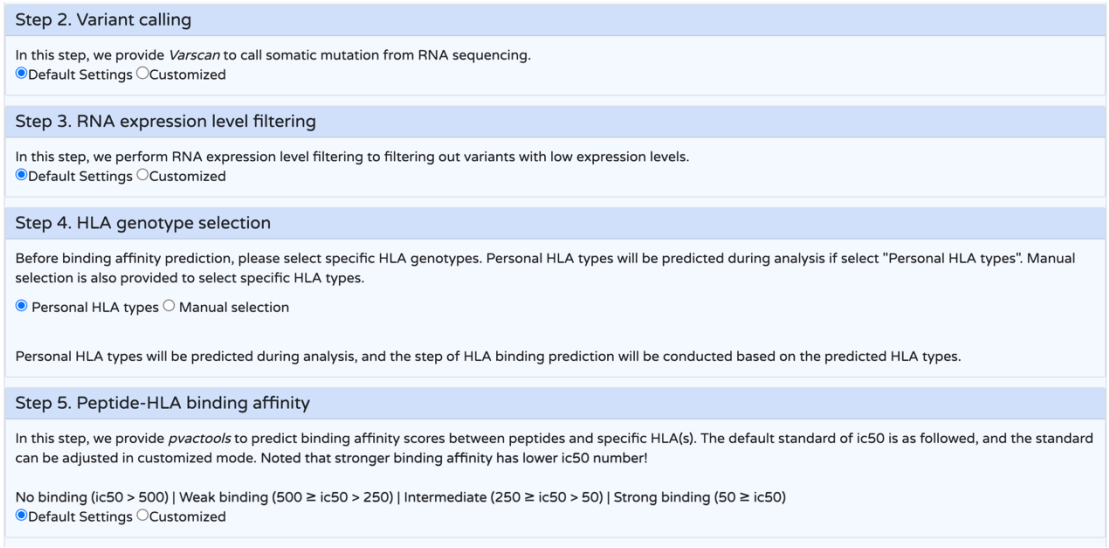
Fig. 5: The section designed for uploading LC-MS/MS data, offering users the flexibility to employ two distinct methods: (a) direct data upload through their browser interface and (b) file upload using specific Google Drive file IDs. (c) The figure demonstrates the successful upload of data, indicating that the users have successfully uploaded their files.

Parameter settings

After the file upload step, the following analysis settings are enabled. Default settings are provided for all the tools used in the analysis process, and the system will analyze patients' HLA types automatically (Fig. 6). Users can also customize the parameters based on their needs. Although we provide default settings, you can use the **"Customized"** option to adjust the parameters (Fig. 7). The parameter settings will differ depending on the selected pipeline, but the adjustable parameters are displayed after you choose the **"Customized"** option. The parameters are listed in Table 2.



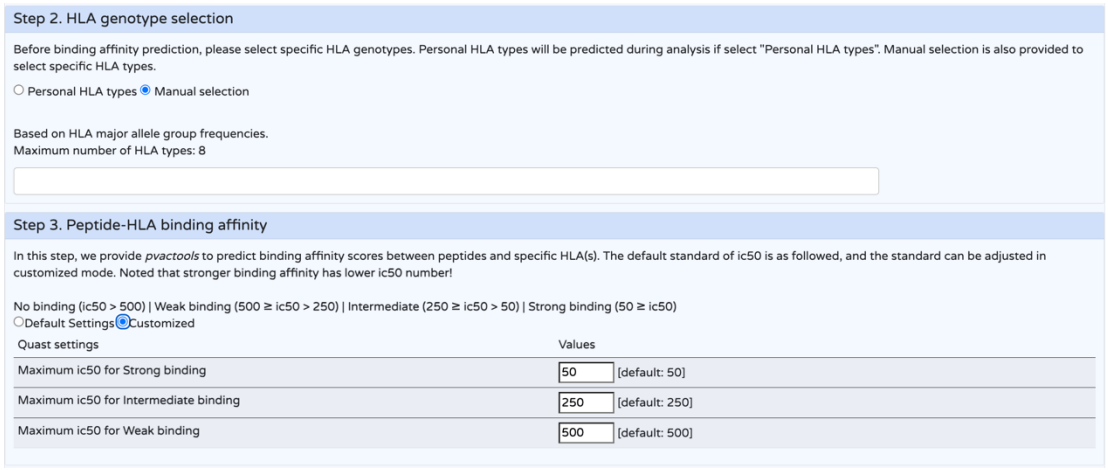
(a)



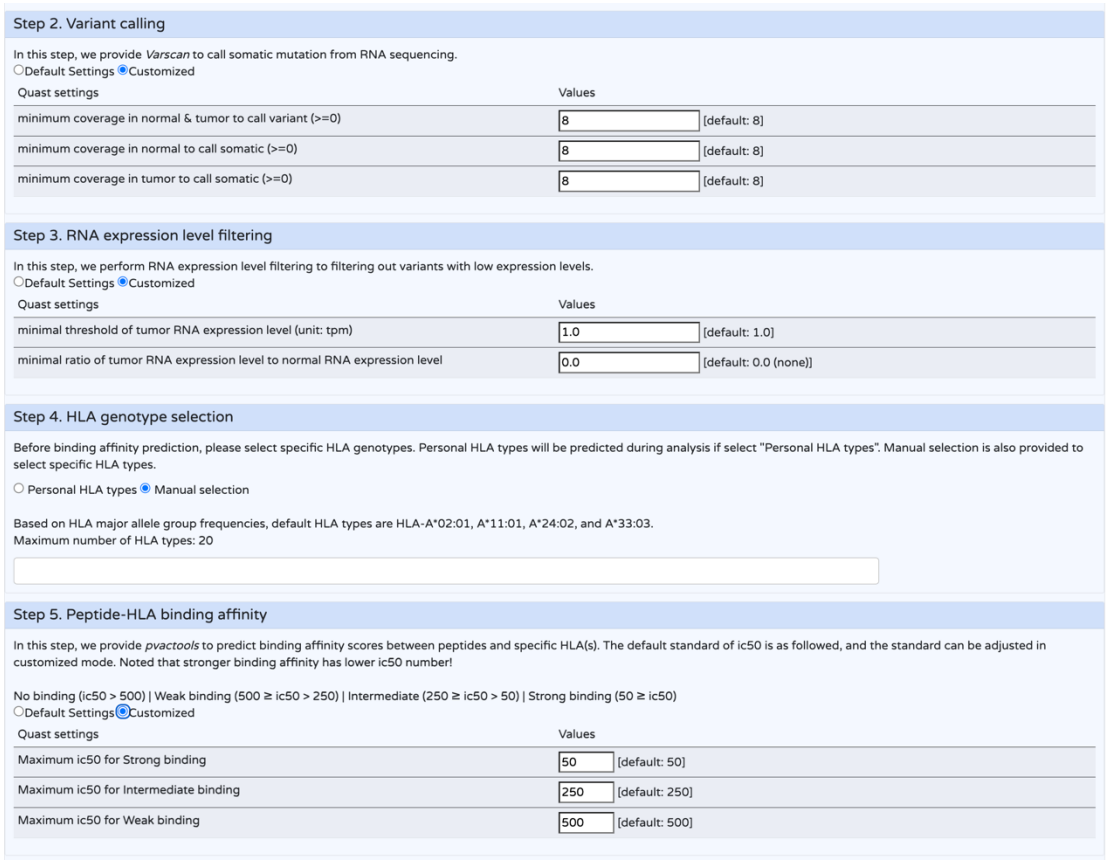
(b)

Fig. 6: (a) Default settings of neoantigen identification with DNA-seq. (b) Default

settings of neoantigen identification with RNA-seq.



(a)



(b)

Fig. 7: (a) Customized parameters for neoantigen identification with DNA-seq. (b) Customized parameters for neoantigen identification with RNA-seq.

Parameter	Pipeline	Applied tool	Default Value		Description
HLA genotype	All	pVACtools	HLA-A*02:01, A*11:01, A*24:02, and A*33:03		Before binding affinity prediction, you should select specific HLA genotypes.
Peptide-HLA binding affinity		In-house script	Maximum ic50 for Strong binding	50	We provide pVACtools to predict binding affinity scores between peptides and specific HLA(s). Noted that stronger binding affinity has lower ic50 number. Please use integers.
			Maximum ic50 for Intermediate binding	250	
			Maximum ic50 for Weak binding	500	
Variant calling minimum coverage	Only RNA	VarScan	Minimum coverage in total reads to call somatic mutations	8	We provide Varscan to call somatic mutation from RNA sequencing. In RNA data, you might specify a lower minimum coverage (default = 8) since the depth of RNA reads is lower than DNA. Please use integers.
			Minimum coverage in normal reads to call somatic mutations	8	
			Minimum coverage in tumor reads to call somatic mutations	8	
RNA expression level filtering		In-house script	Minimal threshold of tumor RNA expression level (unit: tpm)	1.0	We perform RNA expression level filtering to filtering out variants with low expression levels. Please use floats.
			Minimal ratio of tumor to normal RNA expression level	Infinite	

Status tracking

Within the analysis list view, users have access to a comprehensive overview of executed, queued, and currently running analyses, as depicted in Fig. 8. Initially, all steps are designated as **"waiting"** until they are initiated. During the active phase, the status remains **"running"** until completion, at which point it transitions to **"successful"** or **"failed"**. Upon the culmination of all steps, the result page becomes accessible.

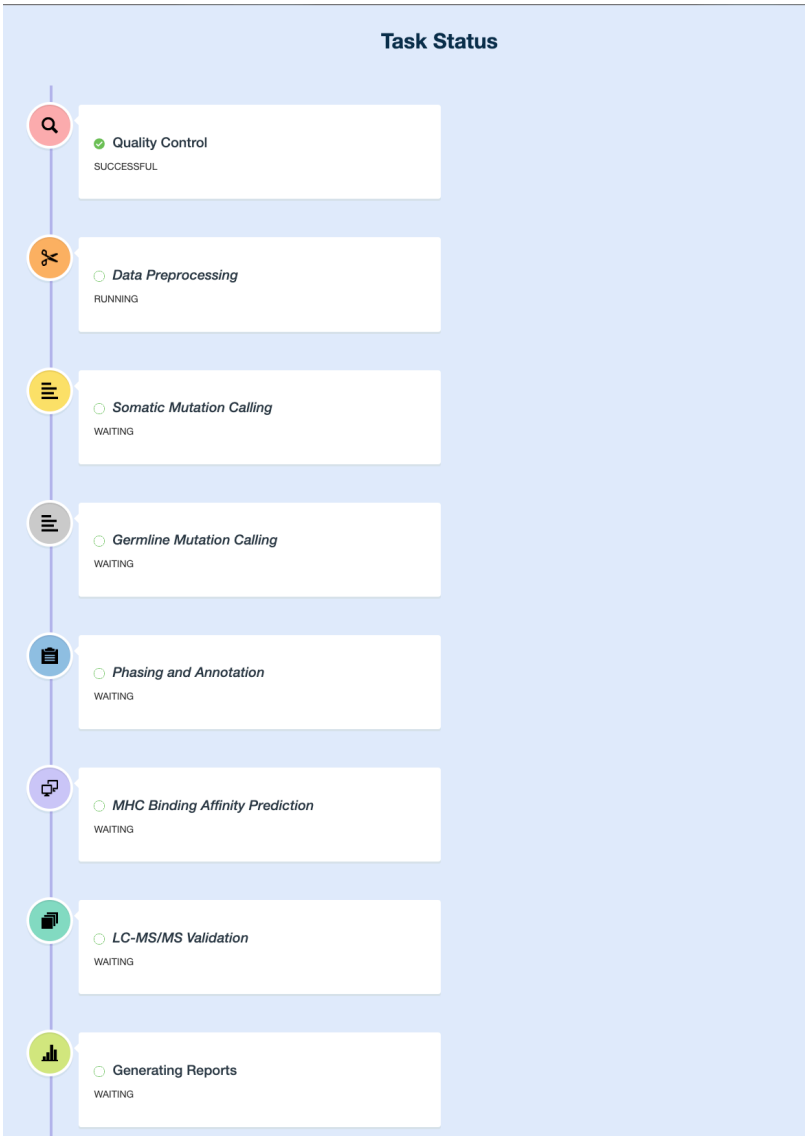
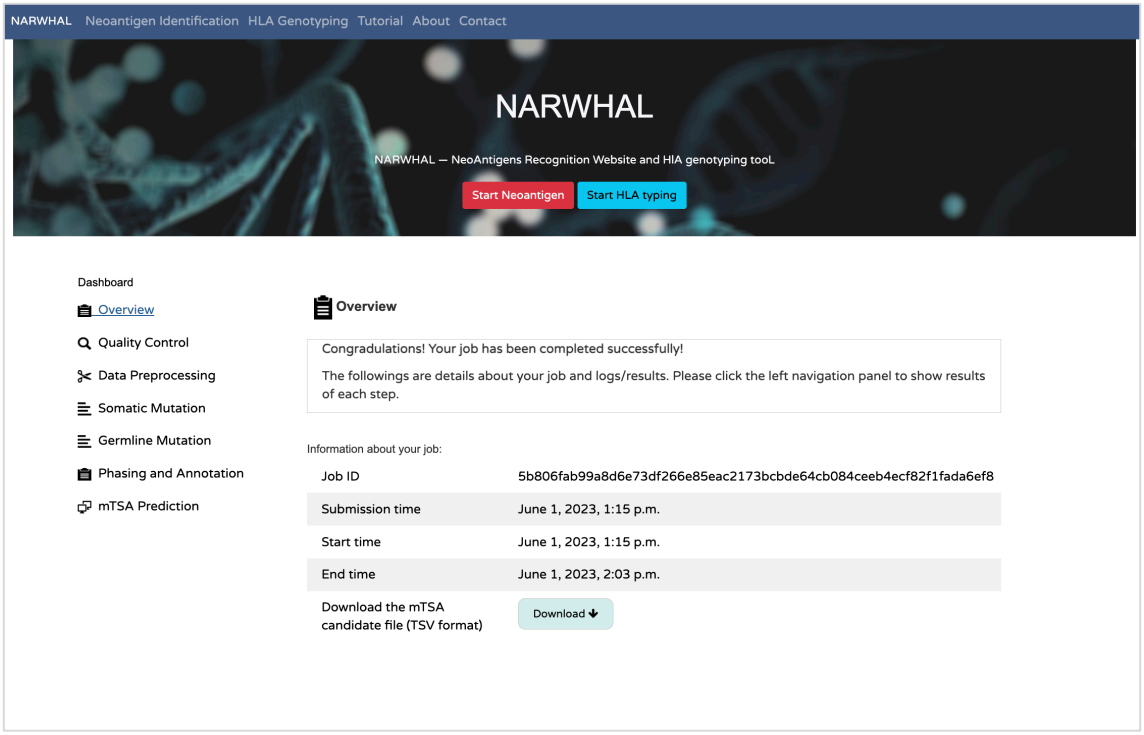


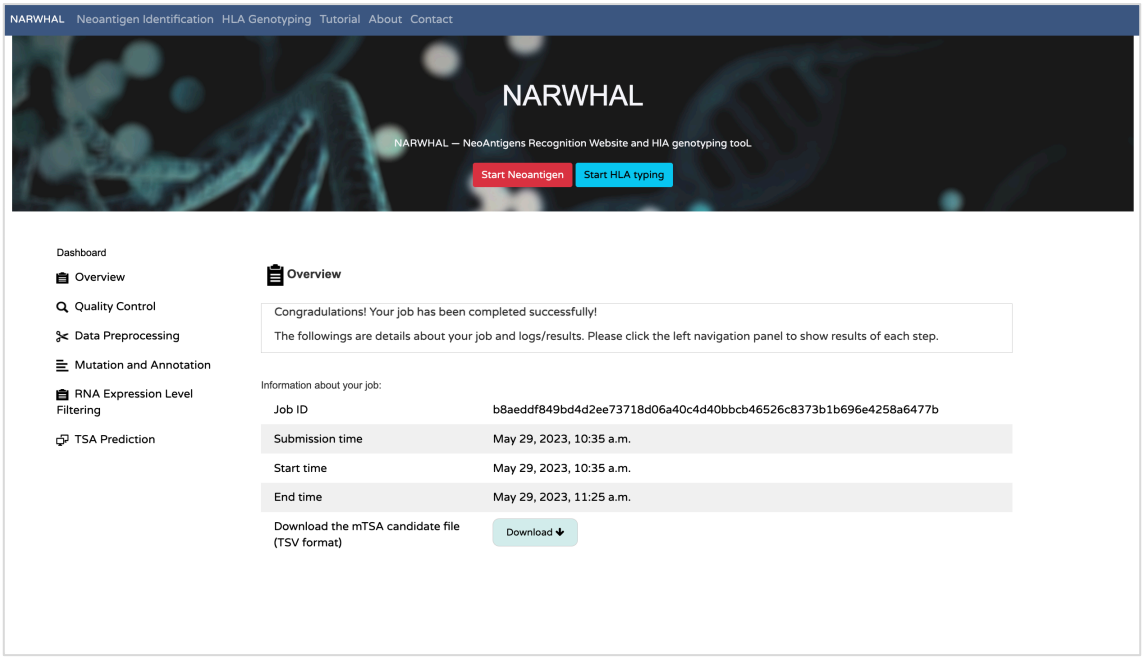
Fig. 8: Task status of neoantigen identification (using tumor DNA-seq as an example).

Report

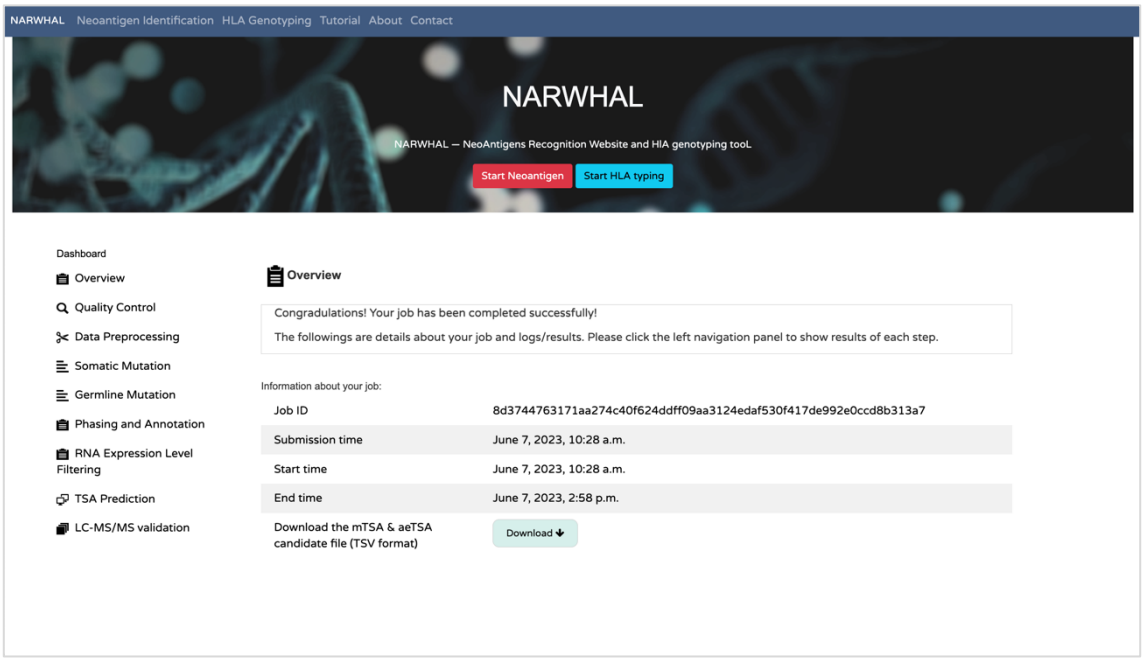
Once the task is successfully submitted to the server, users can conveniently monitor the status and access the results of their task through the website using the link provided via email. The online results page, depicted in Fig. 9, features a user-friendly dashboard on the left-hand panel. This dashboard enables easy navigation through the reports generated at each step of the pipeline. The report pages present the output of each tool in the form of tables or plots, facilitating a visual understanding of the data. Additionally, users are provided with download links for the output files, allowing them to conduct further analyses as needed. If there are errors during the analysis, failed results will be shown.



(a)



(b)



(c)

Fig. 9: The result pages of neoantigen identification using different sequencing inputs: (a) DNA-seq, (b) RNA-seq, (c) and a combination of DNA and RNA-seq.

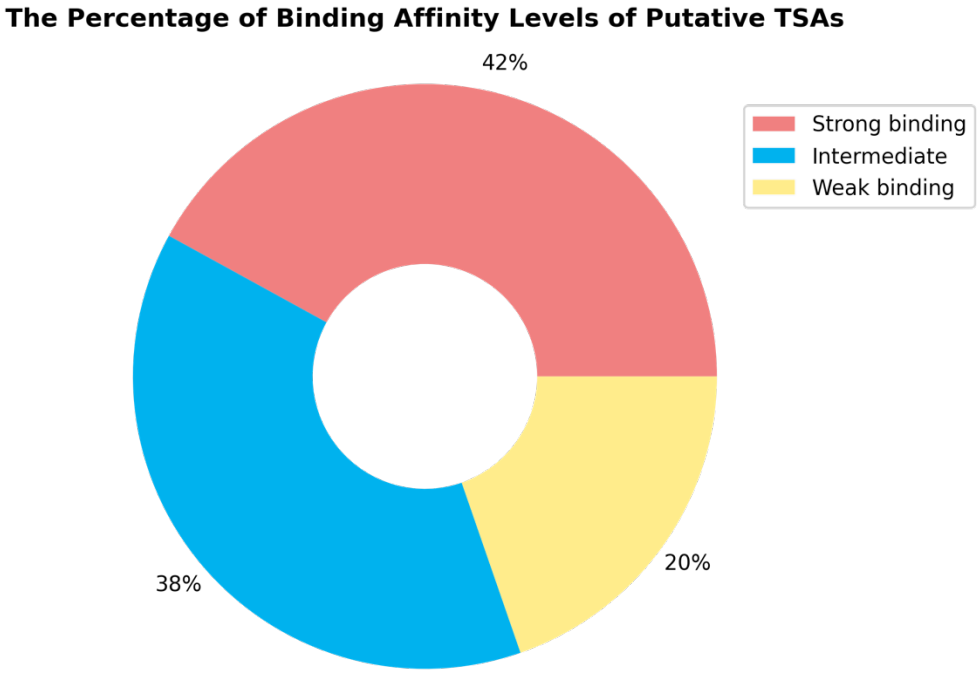
The summary reports offer comprehensive information in the form of a summary table and a pie chart. The summary table in Fig. 10 (a) provides detailed data for each TSA, including the peptide sequence, TSA types, original gene location, binding affinities with HLA, and other relevant details. If only mTSAs are present (only with DNA-seq), the pie chart illustrates the percentage of different binding affinity levels (Fig. 10 (b)). If both mTSAs, aeTSAs, and TAAs are identified, the pie chart showcases the percentage of different neoantigen types (Fig. 10 (c)).

TSA Identification

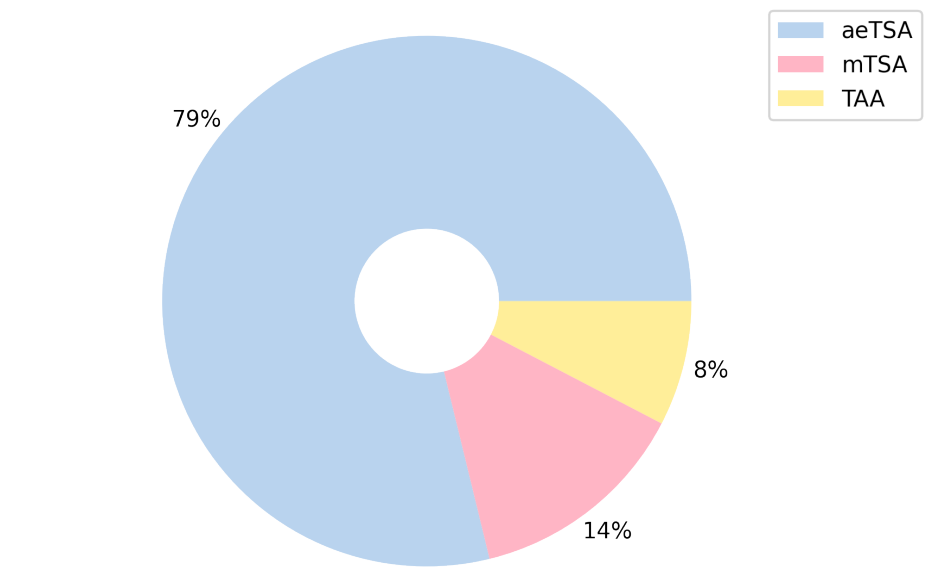
Search this table

TRANSCRIPT_INFO	A*02:11	A*11:01	A*24:02	A*33:03	BEST PEPTIDE	NUM PASSING PEPTIDES	IC50 MT	IC50 WT	%ILE MT	TRANSCRIPT	ENSEMBL GENE ID	BINDING AFFINITY LEVEL	TSA TYPE
chr19-2933544-2933545-T-C	1			1	KAFRYLASL	2	214.705	24.885	1.6	ENST00000314531	ENSG00000175691	Intermediate	mTSA
chr19-2933544-2933545-T-C	1			1	KAFRYLASL	2	214.705	24.885	1.6	ENST00000314531	ENSG00000175691	Intermediate	mTSA
chr19-2933554-2933555-C-A				1	ECKHCGKAFR	1	402.26	929.53	1.2	ENST00000314531	ENSG00000175691	Weak binding	mTSA
chr19-2933554-2933555-C-A				1	ECKHCGKAFR	1	402.26	929.53	1.2	ENST00000314531	ENSG00000175691	Weak binding	mTSA
10T_seq_458		2			SSASPTSPK	2	7.82		0.02			Strong binding	aeTSA
1T_seq_662		2			SSASPTSPK	2	7.82		0.02			Strong binding	aeTSA
7T_seq_582		2			SSASPTSPK	2	7.82		0.02			Strong binding	aeTSA
8T_seq_637		2			SSASPTSPK	2	7.82		0.02			Strong binding	aeTSA
8T_seq_1694		1			SVFEEPLSK	1	8.01		0.02			Strong binding	aeTSA
2T_seq_1509		3			SSALLQLLK	3	8.445		0.02			Strong binding	aeTSA
20T_seq_4364		1		1	ATLGNFAAK	1	8.835		0.025			Strong binding	aeTSA
8T_seq_3862				1	TVKLTLQNR	1	34.39		0.11			Strong binding	aeTSA

(a)



(b)



(c)

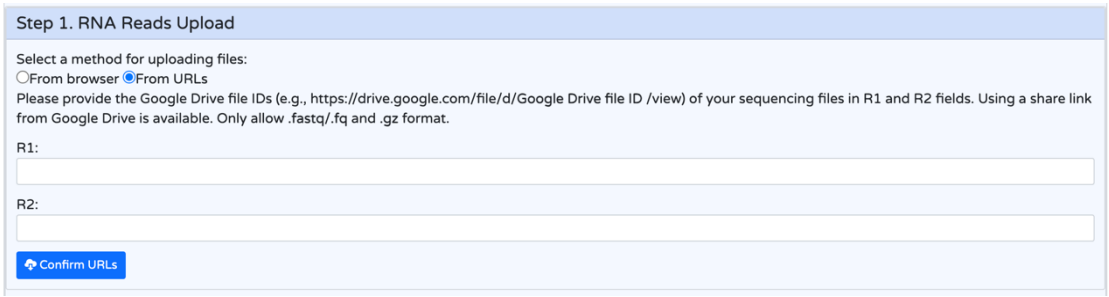
Fig. 10: (a) A comprehensive summary table of neoantigen identification with DNA-seq. (b) A pie chart illustrating the distribution of different binding affinity levels and HLA types. (c) A pie chart demonstrating the distribution of different binding affinity levels and TSA types.

HLA Genotyping

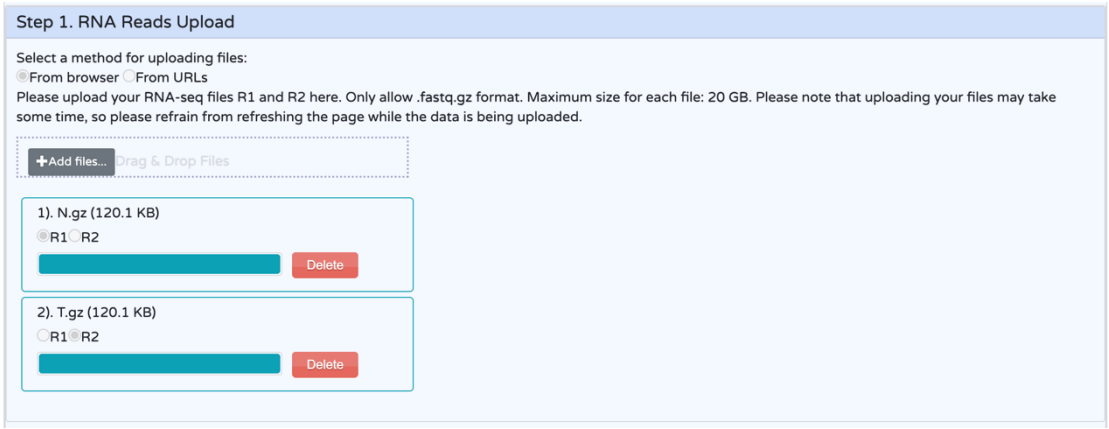
Data upload

The RNA R1/R2 files of tumor and adjacent normal samples should be uploaded separately in ***.fastq.gz*** format, and the maximum size for each file is 20 GB (Fig. 11).

(a)



(b)



(c)

Fig. 11: The section designed for uploading RNA-seq data, offering users the flexibility to employ two distinct methods: (a) direct data upload through their browser interface

and (b) file upload using specific Google Drive file IDs. (c) The figure demonstrates the successful upload of data, indicating that the users have successfully uploaded their files.

Status tracking

HLA genotyping is a valuable feature that allows users to determine a patient's specific HLA types. Once RNA-seq data is uploaded, the status page provides information about the progress of the job, as shown in Fig. 12.

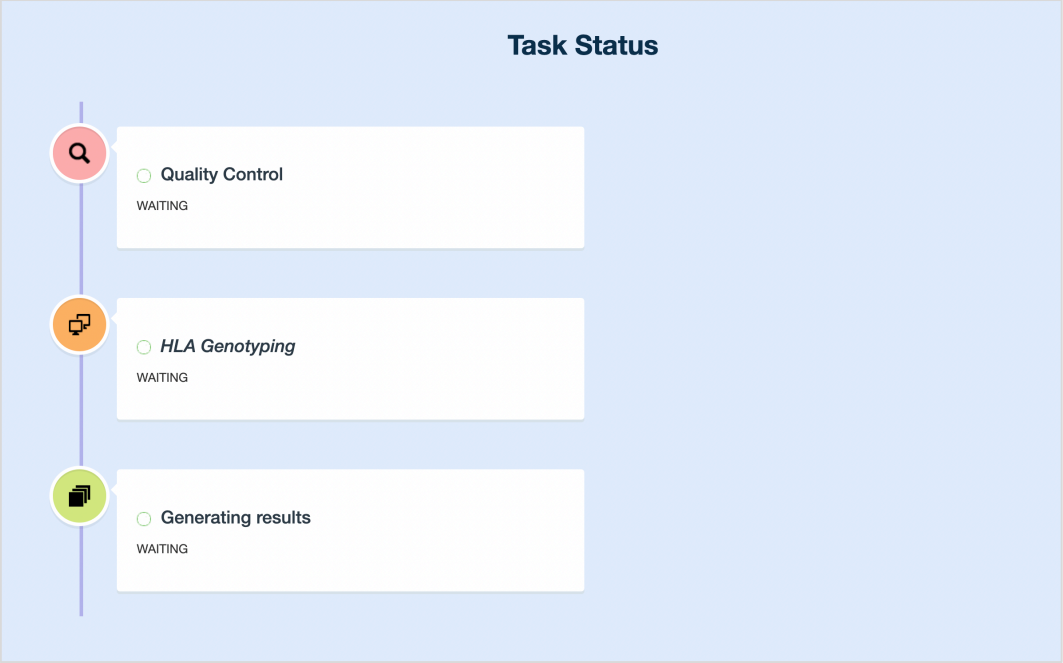


Fig. 12: An ongoing HLA genotyping process is provided as an example. The HLA genotyping and report generation stages are displayed as **"waiting"** since they are queued behind the currently active stage. During the active phase, the status remains **"running"** until completion, at which point it transitions to **"successful"** or **“failed”**. The result page shows accessible after the tasks have been finished.

Report

The final report of the patient's HLA types is generated on the subsequent page, illustrated in Fig. 13. The report includes a table displaying the patient's most common HLA types for both alleles, providing essential information for further analysis and personalized medicine approaches.

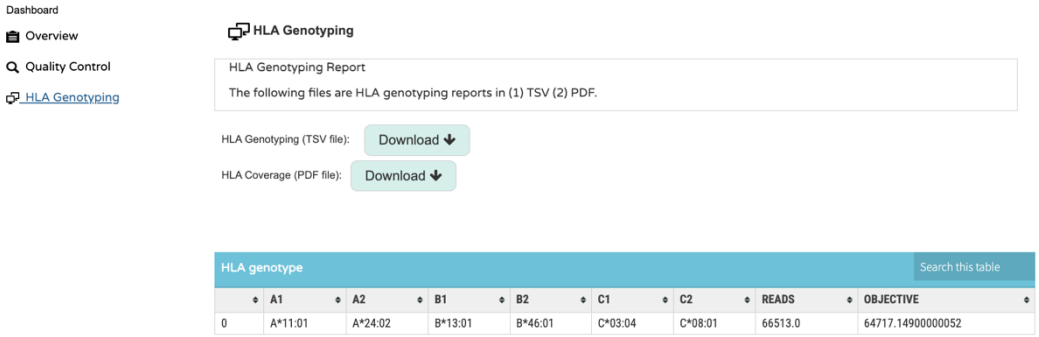
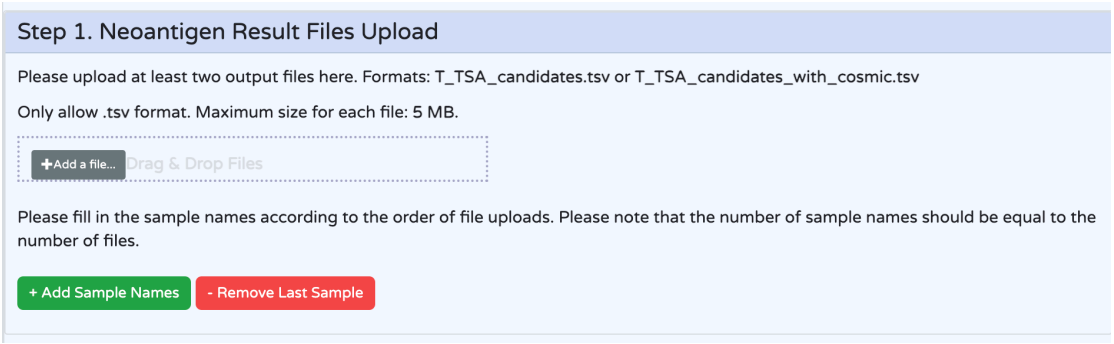


Fig. 13: Summary reports presenting the results of HLA genotyping using RNA-seq.

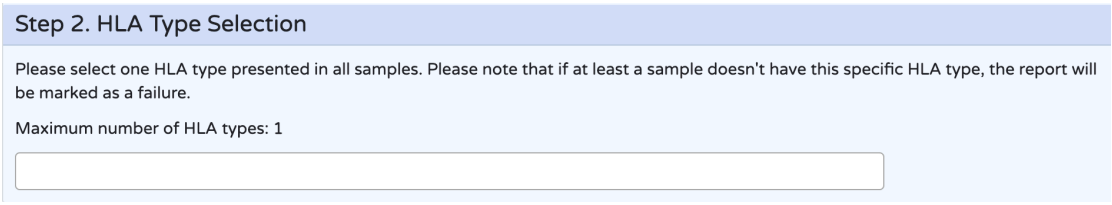
Shared Neoantigen Discovery

Data upload

Users can identify neoantigens presented in multiple samples through the **"Shared Neoantigen Discovery"** function. Please upload at least two output files; the file formats should be either mTSA.tsv or mTSA_and_aeTSA.tsv (see Fig. 14 (a)). Next, specify the HLA types you wish to use for predictions. Please note that if any sample lacks the specified HLA type, the system will exclude that sample from the analysis (see Fig. 14 (b)).



(a)

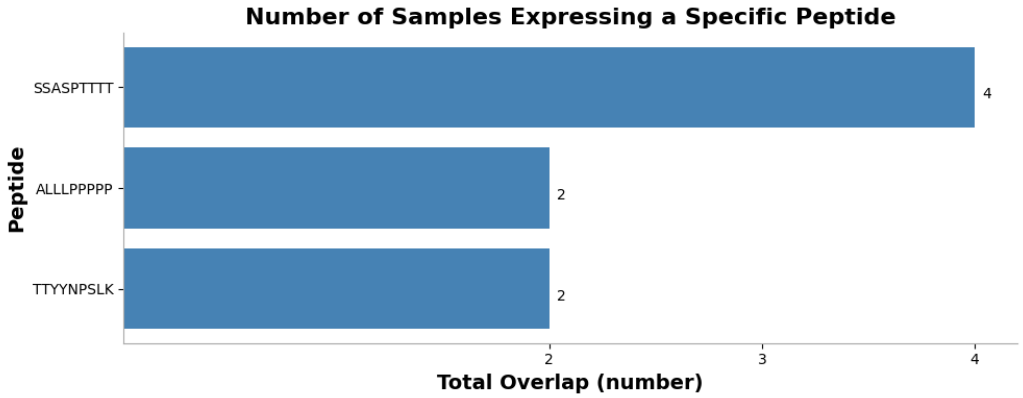


(b)

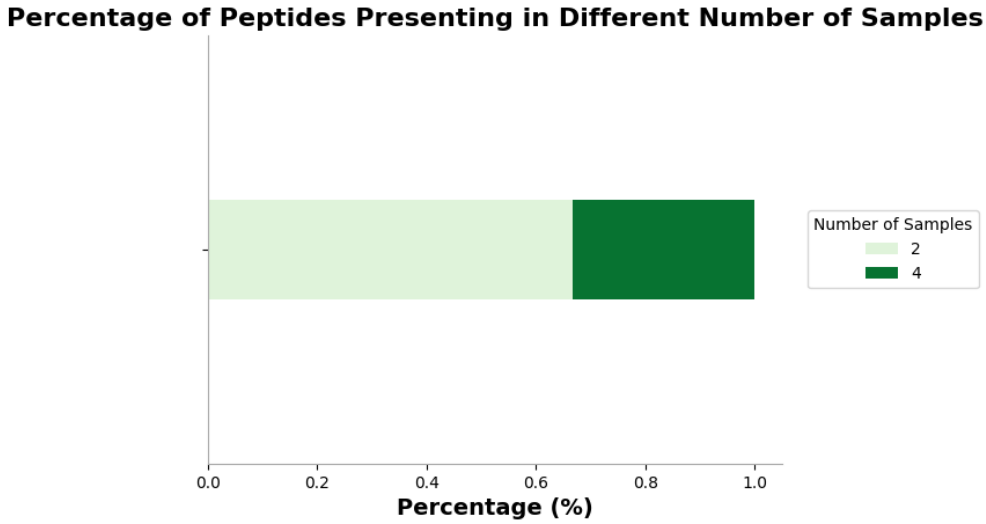
Report

The final report on neoantigens presented in multiple samples (two or more) will be generated on the subsequent page. The report includes a pie chart illustrating the percentage of peptides in different samples (see Fig. 15(a)), a pie chart showing the

number of samples expressing a specific peptide, and a table displaying the neoantigens in details (see Fig. 15(c)). This report provides essential information for further analysis and widely-used medical approaches.



(a)



(b)

Shared Neoantigens — Search this table

PEPTIDE	GROUP ID	SAMPLE1	SAMPLE2	SAMPLE3	SAMPLE4	BEST IC50 MT	BEST %ILE MT	BEST BINDING AFFINITY LEVEL	TOTAL OVERLAP
SSASPTTTT	Cluster 0	v	v	v	v	7.82	0.02	Strong binding	4
TTYYNPSLK	Cluster 6			v	v	8.345	0.02	Strong binding	2
ALLLPPPPP	Cluster 7			v	v	39.615	0.205	Strong binding	2

(c)

System Information

The System Info can be found on the **"About"** section of the website. This page shows information about the NARWHAL version and collected all third-party tools and databases with their versions and executing functions that were conducted in the current pipeline (Fig. 16). This information would help the reproducibility of analysis results from NARWHAL. The names of tools and databases on the tables contain external links to their official websites.

List of Third-party Tools Incorporated into NARWHAL

Tool Name	Version	Executing Function
arcasHLA	0.4.0	HLA genotyping
bedtools	2.30.0	File converter
Biopython	1.68	Tool development
BLAST	2.9.0	Local Alignment (merging with LC-MS/MS peptides)
BWA	0.7.17	Aligment (DNA)
covtobed	1.2.0	File converter
Dragmap	1.2.1	Aligment (DNA)
Ensembl-map	1.2.0	Gene annotation
Ensembl-vep	104.3	Gene annotation
Fastqc	0.11.9	Quality control,
GATK4	4.2.6	Variant calling (DNA)
Jellyfish	2.2.10	k-mer profiling (aeTSA)
kallisto	0.46.0	Quantification of transcripts
MultiQC	1.13	Quality control,
Pandas	1.3.5	Tool development
Picard	2.27.4	MarkDuplicates
pip	22.1.2	Tool development
pVACtools	3.1.0	MHC binding affinity prediction
Pyensembl	2.2.4	k-mer profiling (aeTSA)
Python	3.7	Tool development
Samtools	1.9	VCF filtering
STAR	2.7.10a	Aligment (RNA)
Trimmomatic	0.39	Adaptor trimming
unzip	6	Tool development
VarScan	2.4.4	Variant calling (RNA)

(a)

List of Databases Incorporated into NARWHAL

Database Name	Released Version/Date
COSMIC database	v98, 23rd May 2023
Dragen reference	v8
GRCh38 databases	13rd Dec. 2017
GRCh38.d1.vd1 Reference Sequence	md5: 3ffbcfe2d05d43206f57f81ebb251dc9
IEDB databases	3.1
Illuminaclip	V0.32
VEP cache	104
VEP plugins	109

(b)

Fig. 16: Screenshot of the list of (a) third-party tools and (b) databases used in NARWHAL.

To provide more information to the users, the **"Contact"** section offers users a dedicated email address through which they can reach out to the NARWHAL team for assistance, problem resolution, or to provide valuable feedback and suggestions (Fig. 17).

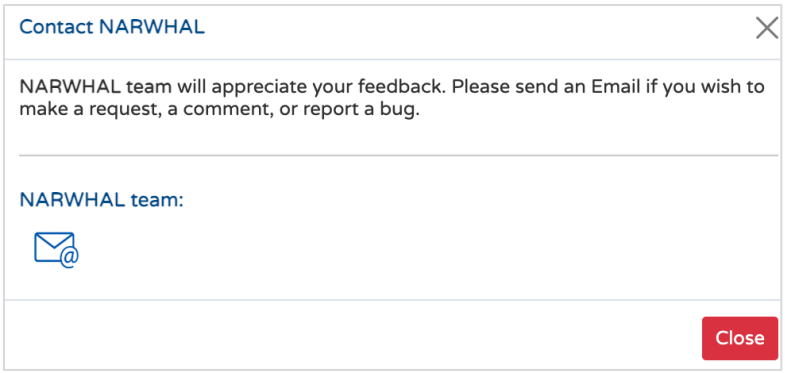


Fig. 17: Contact Information.