Tumor-Specific NeoAntigen Detector (TSNAD) v2.0 User's Manual

Zhan Zhou, Xingzheng Lyu and Jingcheng Wu

Zhejiang University, CHINA

March, 2016

USER'S MANUAL

TABLE OF CONTENTS

1 GETTING STARTED

Tumor-Specific NeoAntigen Detector (TSNAD) is integrated software used for automatically detecting cancer somatic mutations and predicting potential tumor-specific neoantigens. This section explains how to configure operation system and install required third-party software.

1.1 Copyright

Copyright ©2015-2016 by Zhejiang University. Permission is granted to copy this document provided that no fee is charged for it and that this copyright notice is not removed. TSNAD is distributed free of charge by Zhan Zhou (College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China. Tel/Fax: +86-571-88208410, Email: zhanzhou@zju.edu.cn)

1.2 Running environment

TSNAD requires a Linux operation system (e.g. Ubuntu 15.10) with Python, Perl and Java installed. Recommended software versions are Python 2.7.10, Perl 5.22.1 and Java 7.

1.3 Graphical user interface of TSNAD

TSNAD has a friendly graphical user interface (GUI) and easy to use. It contains several menu bars and buttons. Figure 1 is the main GUI of TSNAD. Processing monitoring area will display the intermediate results and tell user the pipeline progress. User can change the font size and style of the processing monitoring area through font size slider and font style combobox on the top right.

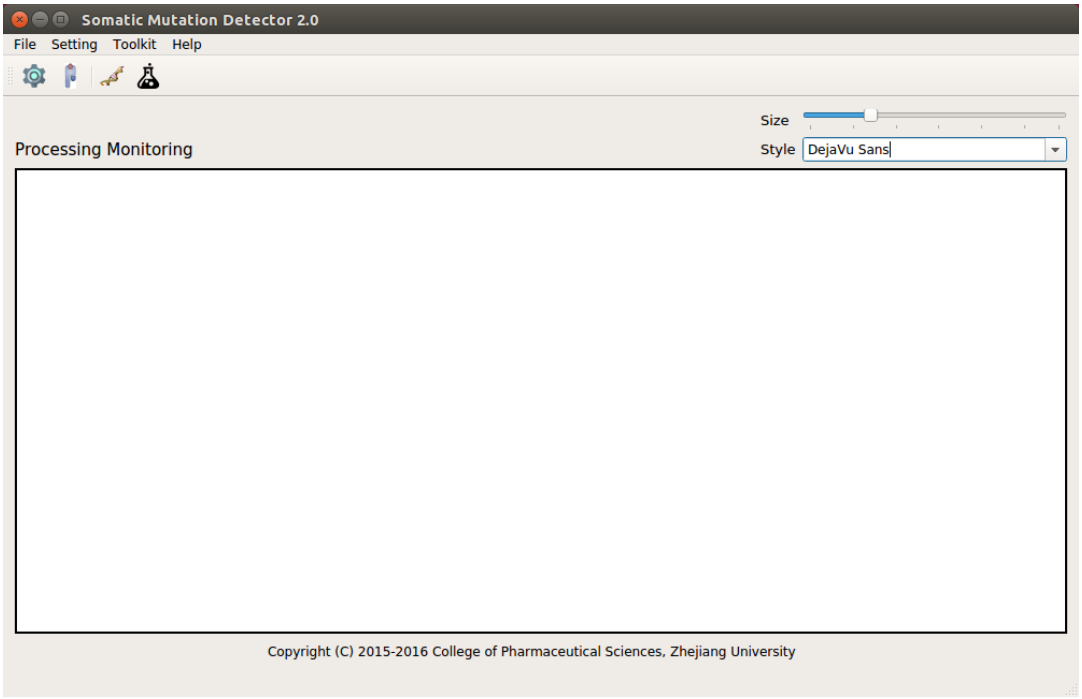


Figure 1. The main GUI of TSNAD

1.4 Required third-party software

TSNAD relies on a series of software for cancer somatic mutation sequencing and antigen prediction. User need preinstall and configure the software correctly. Table 1 summarizes the needed software, major functions and download links.

Table 1. A list of required software for TSNAD pipelines

Software and version	Main function	Download address
Trimmomatic (v0.35)	Filtering raw illumine data, trim crop and remove adaptors.	http://www.usadellab.org/cms/?page=trimmomatic
BWA (v0.7.12)	Mapping a low-divergent, short sequences to a large reference genome, like human genome.	http://bio-bwa.sourceforge.net/
Samtools (v1.3)	File format transformation, alignments manipulation such as sort, remove duplications and index.	http://samtools.sourceforge.net/
Picard (v1.140)	A set of Java command line used to handle with sequencing data, e.g. sort, merge, and mark duplicates.	http://broadinstitute.github.io/picard/

GATK (v3.5)	Identify single nucleotide variants and realign indels in DNA and RNA sequence data; variant callers	http://www.broadinstitute.org/gatk/
Annovar (v20151214)	Annotate genetic variants including start position, end position, reference nucleotide and observed nucleotides, and etc.	http://annovar.openbioinformatics.org/en/latest
SOAP-HLA (v2.2)	Detect human leukocyte antigen (HLA) type for genes	http://soap.genomics.org.cn/SOAP-HLA.html
netMHCpan (v2.8)	Forecast which peptides bind to major histocompatibility complex (MHC) molecules	http://www.cbs.dtu.dk/services/NetMHCpan/

1.5 External reference datasets

In the meanwhile, some third-party software such as GATK and Annovar, need extra databases to run normally. Thus, users have to download these files shown as follows.

(1) GATK

Ftp Address: ftp.broadinstitute.org (user name: gsapubftp-anonymous; password: none)
Path: /bundle/2.8/b37

Necessary files:

human_g1k_v37.fasta

(Notes: BWA software demands this reference sequence that has established index, processing code is

`bwa index -a human_g1k_v37.fasta`)

1000G_phase1.indels.b37.vcf

dbsnp_138.b37.vcf

Mills_and_1000G_gold_standard.indels.b37.vcf

(2) Annovar

During annotating genetic variants, it needs lots of databases including：

refGene, ensGene, cytoBand, genomicSuperDups, esp6500siv2_all, 1000g2015aug_all, avsnp144, dbnsfp30a, cosmic70, nci60, etc. of version hg19, putting them into one folder for the sake of convenience.

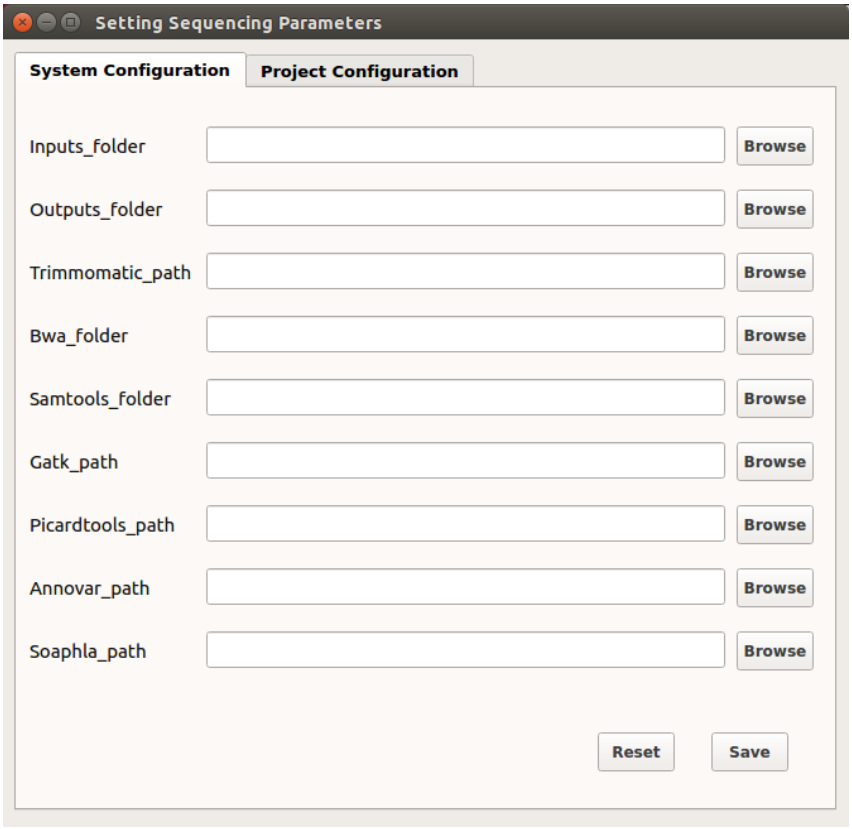
2 SETTING PARAMETERS

2.1 Somatic mutation sequencing

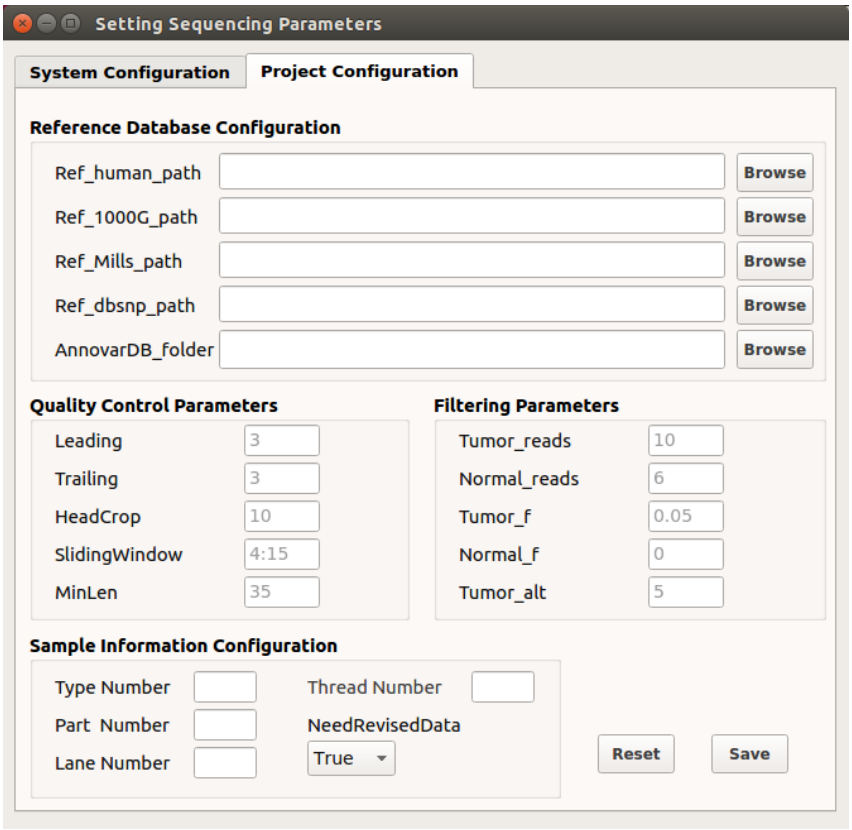
Click on setting menu bar and choose sequencing parameters (or directly click on the toolbar ). Figure 2 is the main GUI of somatic sequencing parameter configuration. On "System Configuration" tab, users point out input and output files folder, third-party software folder or execution file path. On "Project Configuration" tab, users configure the parameters with respect to a specific project. Some recommend parameters are provided in gray color. Also, user can modify these parameters. Table 2 shows meanings of some parameters.

Table 2. Summary of sequencing parameters and corresponding explanations

Parameter	**Meaning**
Type Number	*Number of the type of putting files* (e.g. tumor sample and normal sample or cancer sample and normal sample)
Part Number	*Number of part* (sequence result of forward and reverse direction are two)
Lane Number	*Number of lane* (the sequence result line number of each type of sample)
Thread Number	*Number of threads* (used in a multi-thread mode for Trimmomatic, BWA, Samtools, and GATK)
NeedRevisedData	*Whether need report for Base Quality Score Recalibration in GATK*
Leading	*Cut bases off the start of a read, if below a threshold quality*
Trailing	*Cut bases off the end of a read, if below a threshold quality*
Head crop	*Cut the specified number of bases from the start of the read*
Sliding window	*Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.*
Min len	*Minimum length for each read*
Tumor/Normal_reads	*Minimum sequencing depth of each site for tumor/normal cells*
Tumor/Normal_f	*Minimum mutation frequencies for tumor/normal cells*
Tumor_alt	*Minimum mutated reads for tumor cells*



(a)



(b)

Figure 2. The GUI of setting sequencing parameter. (a) System configuration, (b) project configuration.

TSNAD has its own naming convention for input files. The file name is composed of three strings and connected by a underline. The first string is file type (blood/normal or tumor). The second string is lane number, while the last string denotes the part number. Example names are below

```
blood_L1_R1.fastq
normal_L2_R1.fastq
tumor_L1_R1.fastq
tumor_L2_R3.fastq
```

2.2 Antigen predicting

Choose predicting parameters in the setting menu (or directly click on the toolbar ). Sequencing parameter dialog is shown as Figure 3. In "Path Configurations" groupbox, user need select input file, output files folder and netMHCpan software folder. Input file is the annotated mutations generated by Annovar. In "netMHCpan Parameters" groupbox, a series of parameters can be set by user to qualify the final results.

Table 3. Summary of predicting parameters and corresponding explanations

Parameter	Meaning
HLA_A	*Types of HLA-A alleles, which is the output of* SOAP-HLA
HLA_B	*Types of HLA-B alleles, which is the output of* SOAP-HLA
HLA_C	*Types of HLA-C alleles, which is the output of* SOAP-HLA
strong binding	*Affinity Threshold for Strong binding peptides* (nm)
weak binding	*Affinity Threshold for Weak binding peptides* (nm)
peptide length	*Length of peptides that predict binding to HLA molecules* (mer)

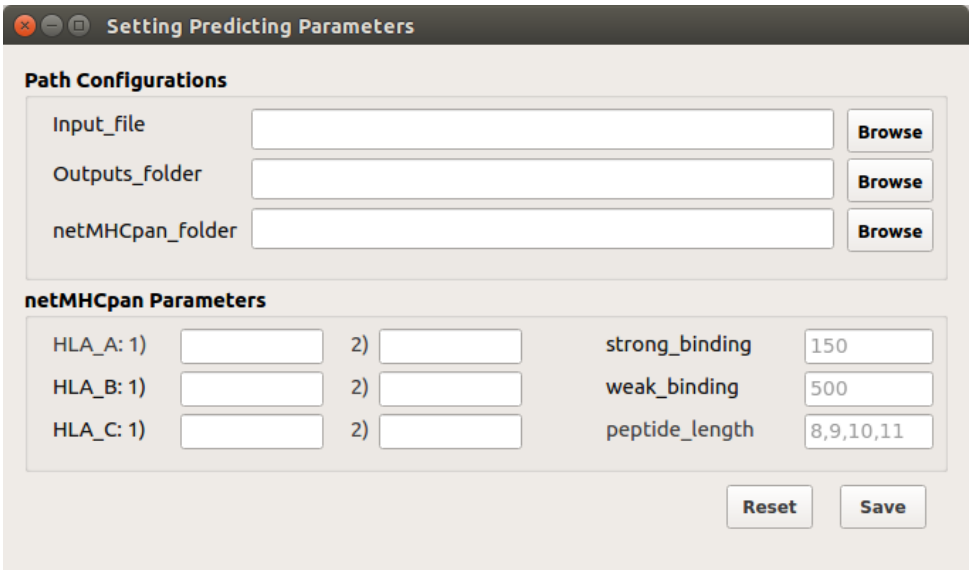


Figure 3. The GUI of setting predicting parameter.

2.3 Settings example

After all the parameters are settled, two configuration files (somatic_mutation_sequencing_parameters.config and antigen_predicting_parameters.config) will be generated automatically in users' TSNAD executable directory. Example settings will look like below.

(1) somatic_mutation_sequencing_parameters.config

```
inputs_folder /home/pub/data/Sequence/Raw_data/
outputs_folder /home/pub/data/Sequence/outputs/
trimmomatic_folder /home/pub/Software/Trimmomatic-0.35/trimmomatic-
0.35.jar
bwa_folder /home/pub/Software/bwa.kit-0.7.12/
samtools_folder /home/pub/Software/samtools-1.3/
gatk_folder /home/pub/Software/GenomeAnalysisTK-3.5/GenomeAnalysisTK.jar
picardtools_folder /home/pub/Software/picard-tools-1.140/picard.jar
annovar_folder /home/pub/Software/annovar20151214/table_annovar.pl
soaphla_folder /home/pub/Software/SOAP-HLA/MHC_autopipeline_b37.pl
ref_human_folder /home/pub/Software/GenomeAnalysisTK-
3.5/resources/b37/human_g1k_v37.fasta
ref_1000G_folder /home/pub/Software/GenomeAnalysisTK-
3.5/resources/b37/1000G_phase1.indels.b37.vcf
ref_Mills_folder /home/pub/Software/GenomeAnalysisTK-
3.5/resources/b37/Mills_and_1000G_gold_standard.indels.b37.vcf
ref_dbsnp_folder /home/pub/Software/GenomeAnalysisTK-
3.5/resources/b37/dbsnp_138.b37.vcf
annovarDB_folder /home/pub/Software/annovar20151214/humandb/
leading 3
trailing 3
headcrop 10
slidingwindow 4:15
minlen 35
normal_reads 6
tumor_reads 10
tumor_f 0.05
normal_f 0
tumor_alt 5
typeNum 2
partNum 2
laneNum 1
threadNum 6
needRevisedData True
```

(2) antigen_predicting_parameters.config

```
Input_file
/home/pub/Test/output/annovar_results/mutect_somatic_anno.hg19_missense.
txt
Outputs_folder /home/pub/Test/output/antigen/
netMHCpan_folder /home/pub/Software/netMHCpan/netMHCpan-2.8/
A1 02:01
A2 33:03
B1 46:01
B2 35:14
C1 01:02
C2 03:02
strong_binding 150
weak_binding 500
peptide_length 8,9,10,11
```

3 RUNNING PROCEDURES

3.1 Starting the mutation sequencing

User can click on run mutation sequencing button in toolkit menu and confirm to run the sequencing pipeline. Figure 4 shows the interface of ensuring to run mutation sequencing.

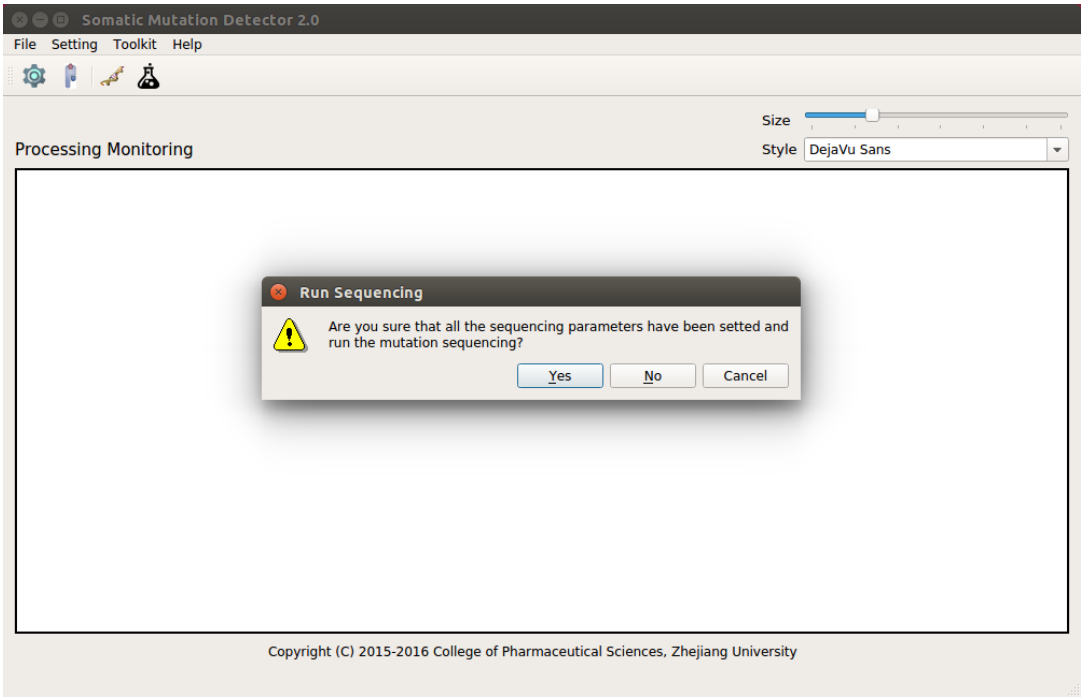


Figure 4. The GUI of confirming the mutation sequencing

3.1.1 Sequencing flowchart

TSNAD sequencing pipeline takes tumor/normal Illumina data (in FASTQ format) as input and process the raw data with some third-party software. The final results are annotated gene mutations which is useful for genetic diagnosis or further analysis. The flow diagram is shown as figure 5.

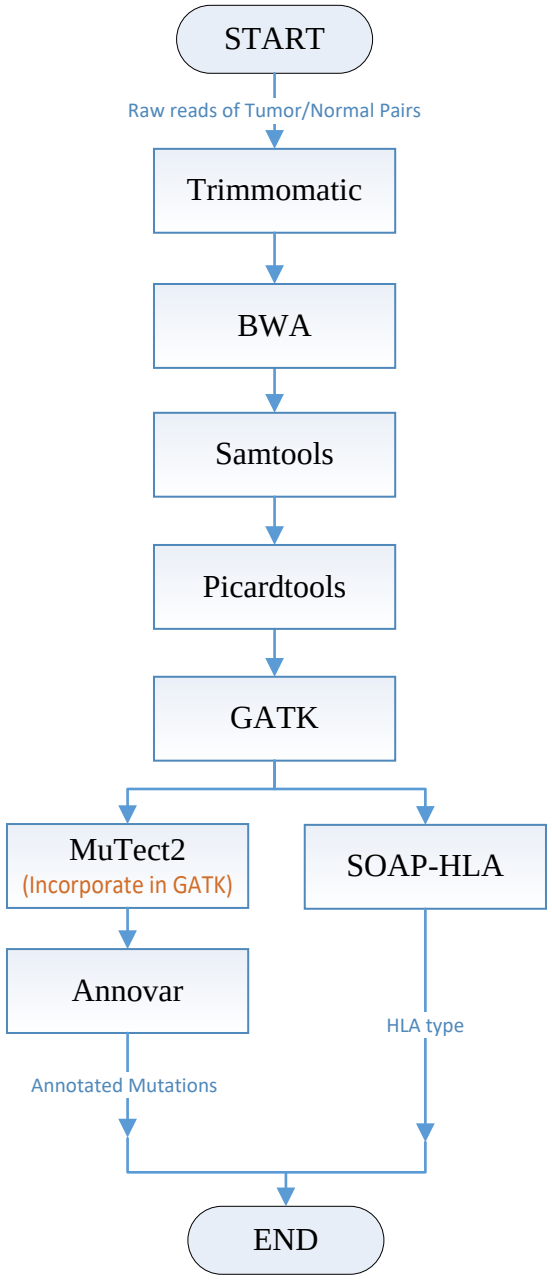
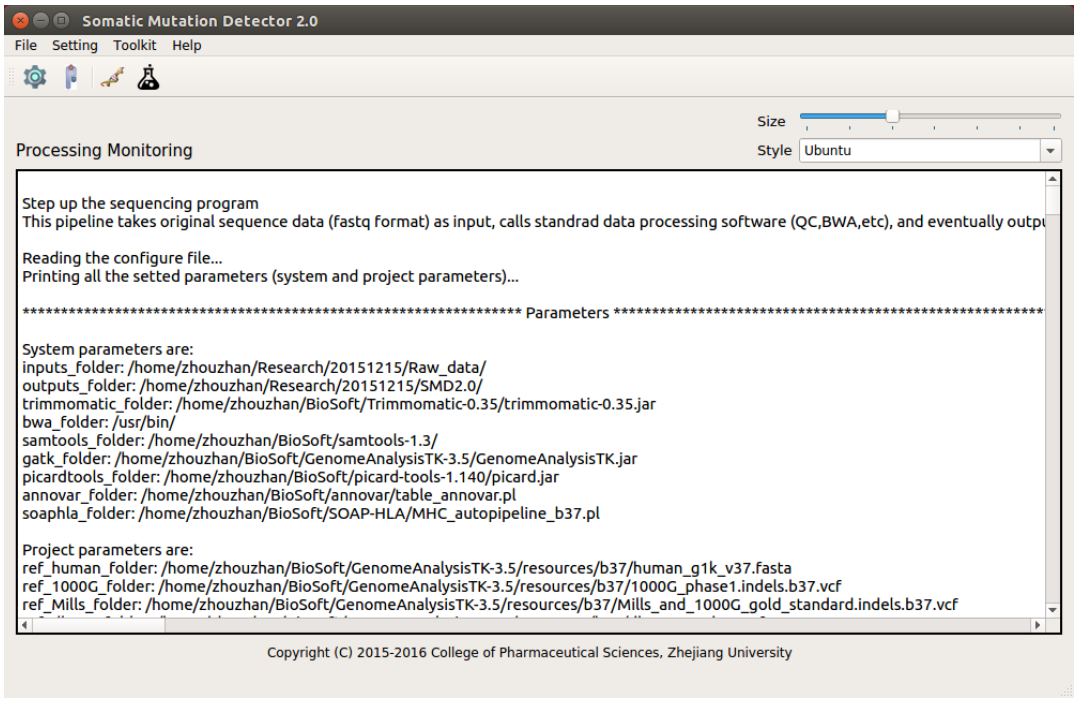


Figure 5. Mutation sequencing pipeline

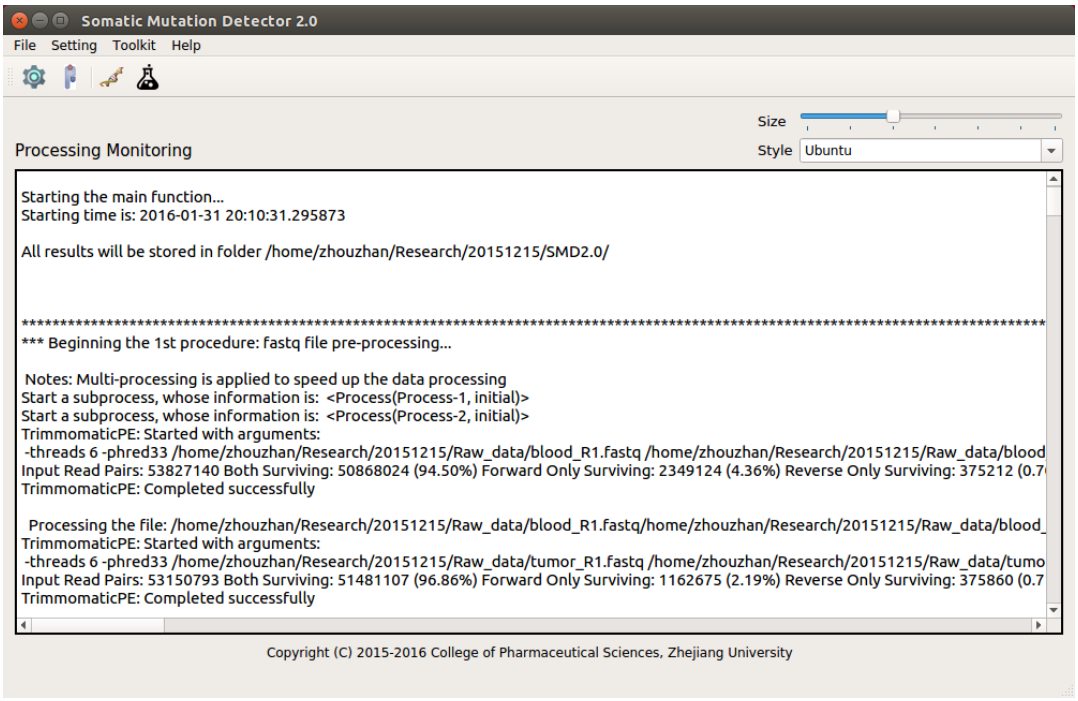
3.1.2 Processing monitor display

Through processing monitoring window, user can clearly observe the current progress of sequencing pipeline. First of all, sequencing parameters set manually by user will display

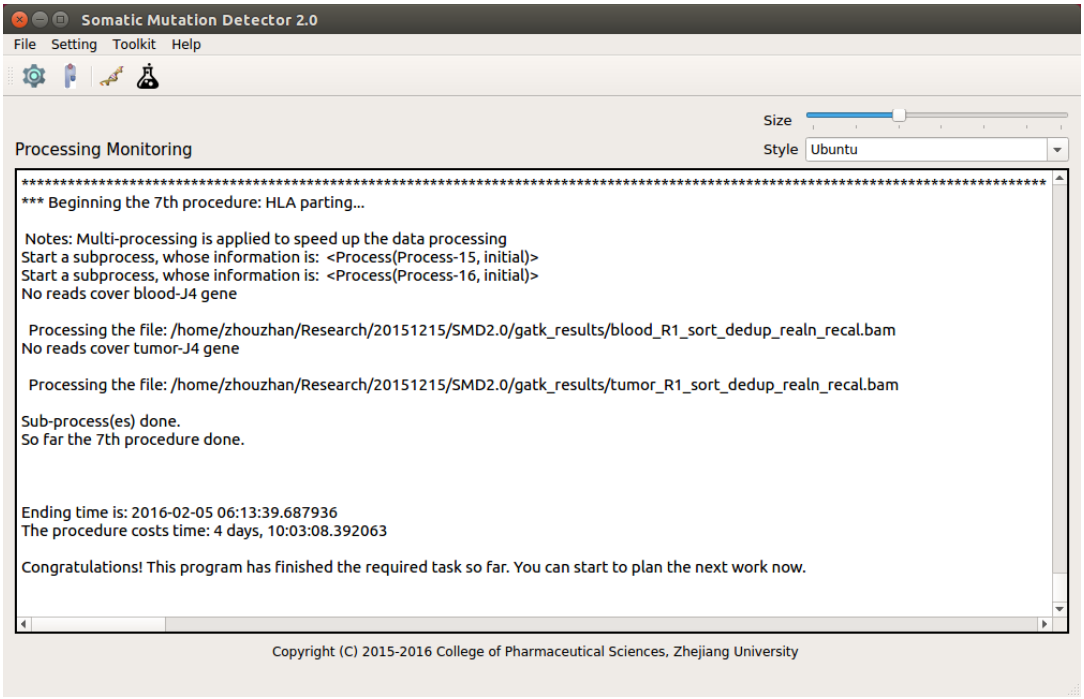
in the processing monitoring window (Figure 6(a)). Then, sequencing pipeline begins to process the raw reads one procedure by one procedure (Figure 6(b)). After the seventh precedure has been done, the whole pipneline ends (Figure 6(c)).



(a)



(b)

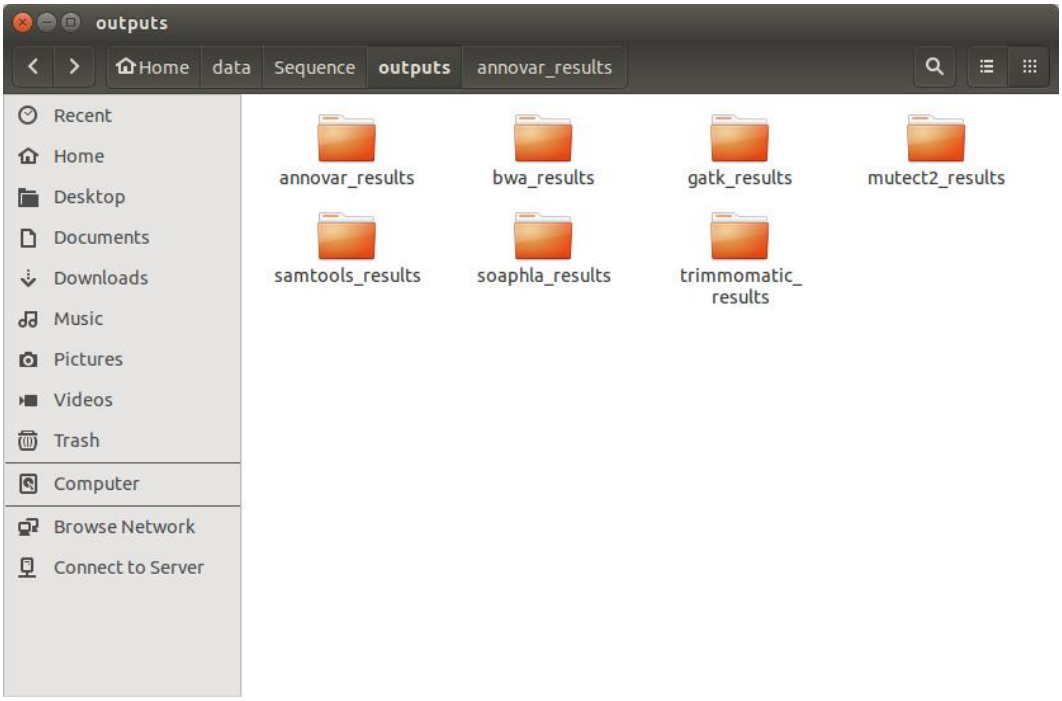


(c)

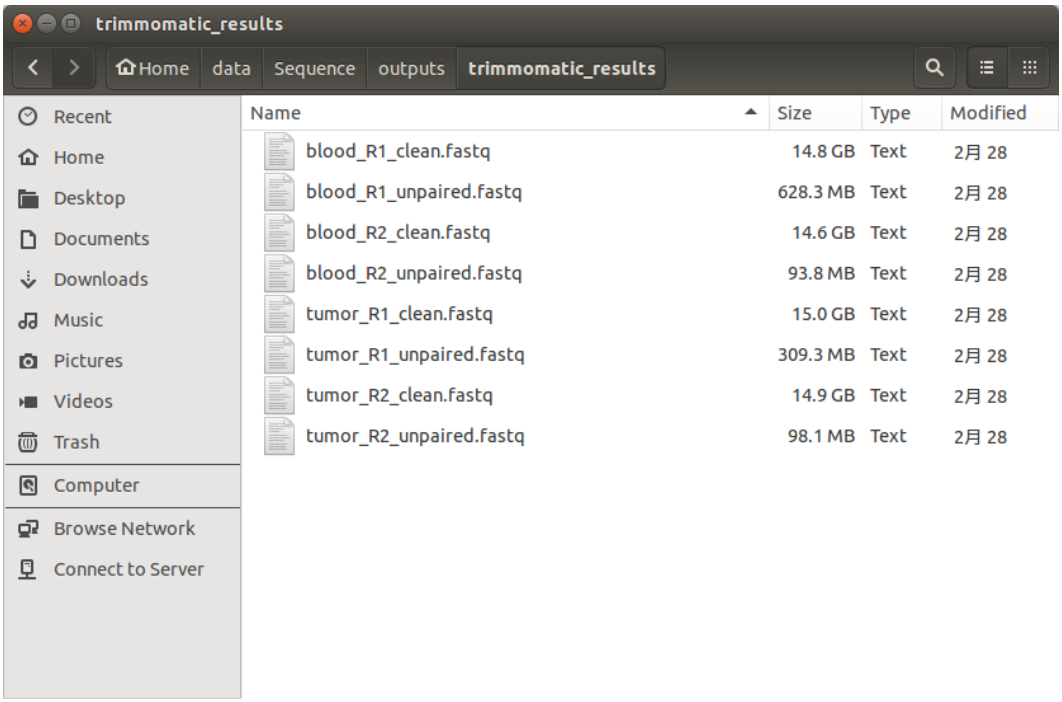
Figure 6. Sequencing pipeline monitoring

3.1.3 Final outputs for mutation sequencing

Results for every stage will be stored in corresponding folder (Figure 7(a)). They are: *trimmomatic_results* (results of quality control and pretreatment), *bwa_results* (results of sequence mapping), *samtools_results* (results of SAM/BAM files handled), *gatk_results* (optimized by GATK), *mutect2_results* (results of detected somatic mutations), *annovar_results* (annotated mutations) and *hla_results* (lists of HLA allele types).



(a)



(b)

Figure 7. Final results for mutation sequencing. (a) Results of every stage saves in a folder; (b) example output of trimmomatic procedure

3.2 Starting the antigen predicting

In toolkit menu, user can click on run antigen predicting button and then confirm to run this processing pipeline. The main GUI of ensuring to run antigen predicting can be showns as

below.

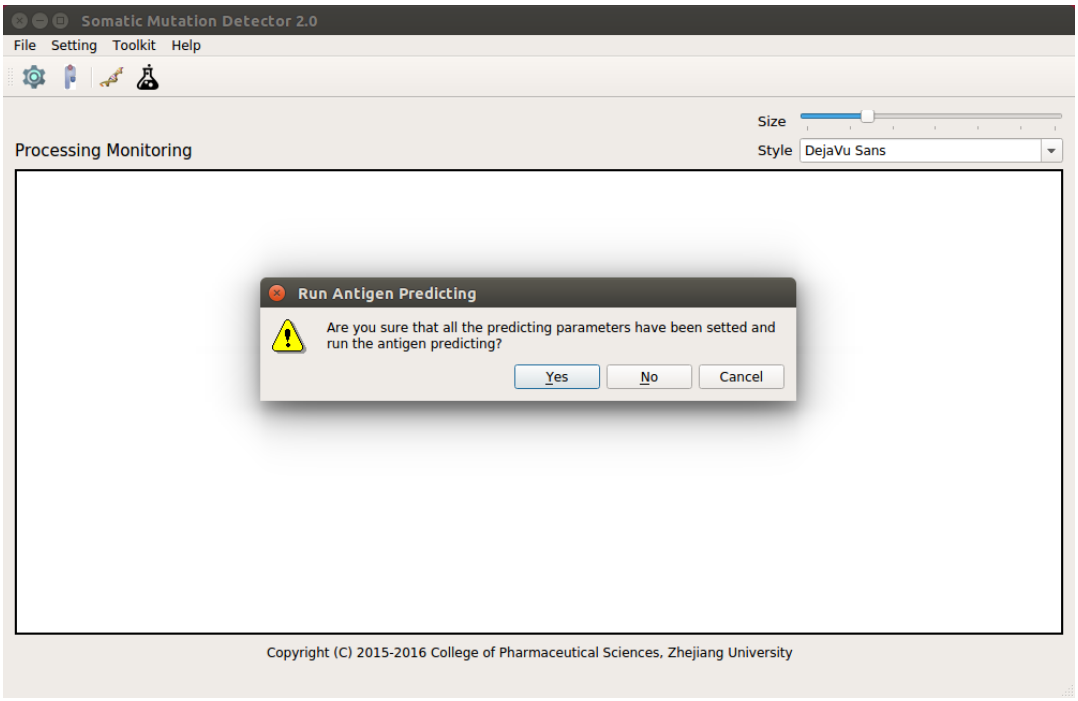


Figure 8. The GUI of confirming the antigen predicting

3.2.1 Predicting flowchart

Antigen predicting pipeline put annotated mutations to external software netMHCpan and our originally developed program AnnovarFilter.pl. Figure 9 is the main flow diagram.

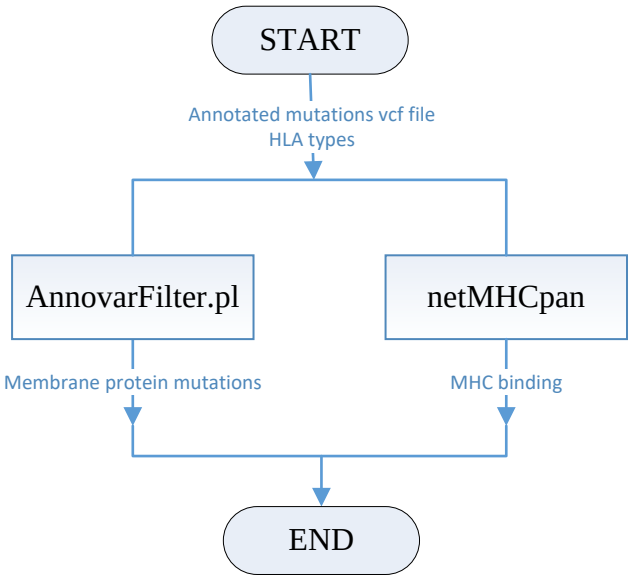


Figure 9. Antigen predicting pipeline

3.2.2 Processing monitor display

In a similar way with sequencing, user can monitor the entire antigen predicting pipneline through processing monitoring window.

3.2.3 Final outputs for antigen predicting

Results for antigen predicting can be shown as Figure 10.

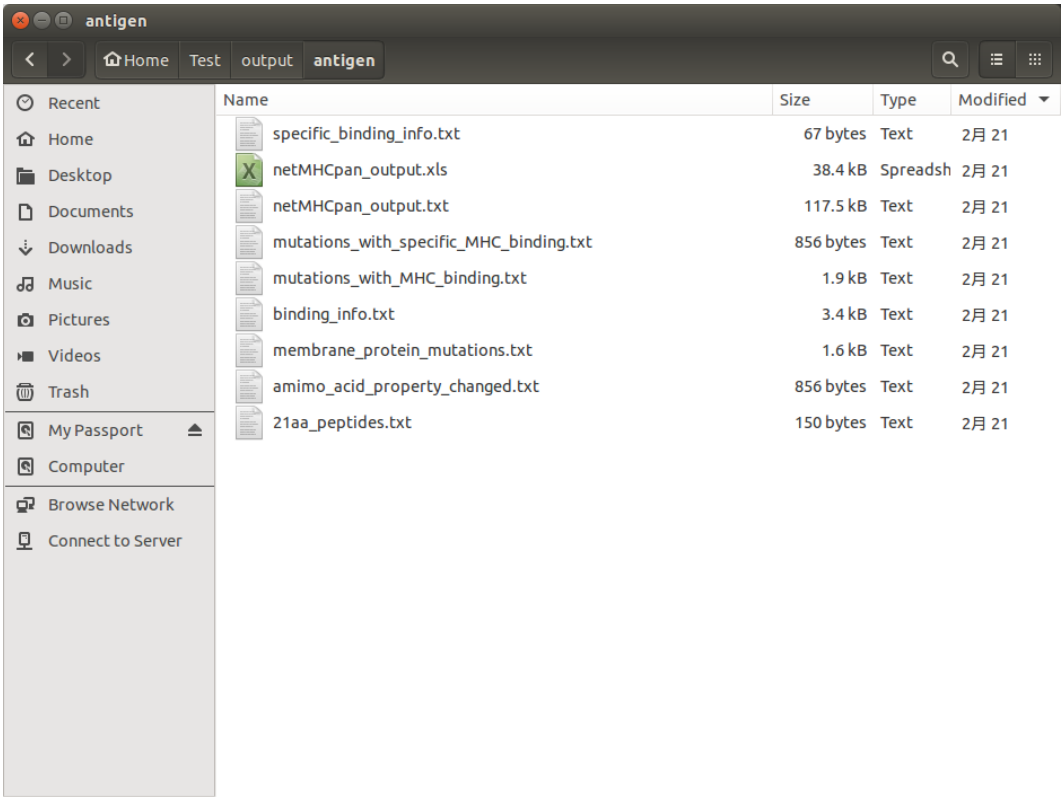


Figure 10. Final results for antigen predicting