

# **Somatic Mutation Detector (SMD) v2.0**

## **User's Manual**

Zhan Zhou, Xingzheng Lyu and Jingcheng Wu  
Zhejiang University, CHINA  
March, 2016

# SMD USER'S MANUAL

## TABLE OF CONTENTS

<b>1 GETTING STARTED .....</b>	<b>1</b>
1.1 Copyright .....	1
1.2 Running environment.....	1
1.3 Graphical user interface of SMD .....	1
1.4 Required third-party software .....	2
1.5 External reference datasets.....	3
<b>2 SETTING PARAMETERS .....</b>	<b>4</b>
2.1 Somatic mutation sequencing .....	4
2.2 Antigen predicting.....	6
2.3 Settings example .....	7
<b>3 RUNNING PROCEDURES .....</b>	<b>8</b>
3.1 Starting the mutation sequencing .....	8
3.1.1 Sequencing flowchart.....	9
3.1.2 Processing monitor display .....	9
3.1.3 Final outputs for mutation sequencing.....	11
3.2 Starting the antigen predicting .....	12
3.2.1 Predicting flowchart .....	13
3.2.2 Processing monitor display .....	14
3.2.3 Final outputs for antigen predicting .....	14

# 1 GETTING STARTED

Somatic Mutation Detector (SMD) is integrated software used for automatically detecting cancer somatic mutations and predicting potential tumor-specific antigens. 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. SMD 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

SMD 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 SMD

SMD 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 SMD. 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.

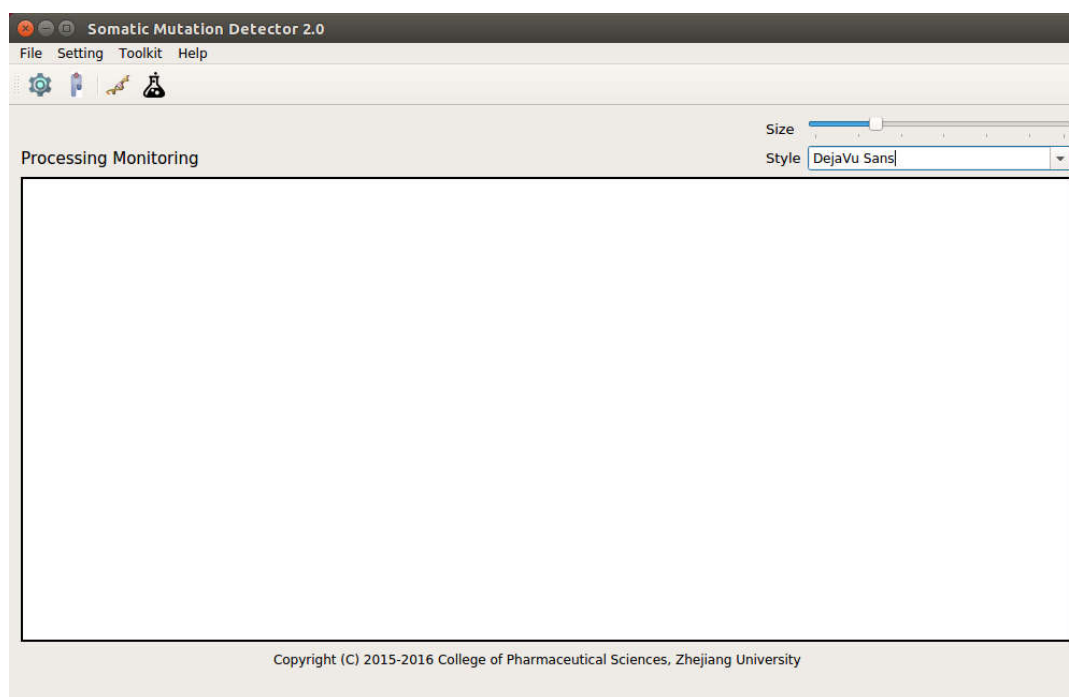


Figure 1. The main GUI of SMD

## 1.4 Required third-party software

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

Table 1. A list of required software for SMD pipelines

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

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

## 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, dbSNP30a, 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. Parameters and corresponding explanations

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

**Setting Sequencing Parameters**

**System Configuration** | **Project Configuration**

Inputs\_folder

Outputs\_folder

Trimmomatic\_path

Bwa\_folder

Samtools\_folder

Gatk\_path

Picardtools\_path

Annovar\_path

SoapHla\_path

(a)

**Setting Sequencing Parameters**

**System Configuration** | **Project Configuration**

**Reference Database Configuration**

Ref\_human\_path

Ref\_1000G\_path

Ref\_Mills\_path

Ref\_dbsnp\_path

AnnovarDB\_folder

**Quality Control Parameters**

Leading

Trailing

HeadCrop

SlidingWindow

MinLen

**Filtering Parameters**

Tumor\_reads

Normal\_reads

Tumor\_f

Normal\_f

Tumor\_alt

**Sample Information Configuration**

Type Number  Thread Number

Part Number  NeedRevisedData

Lane Number  True

(b)

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

SMD 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. Parameters and corresponding explanations

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

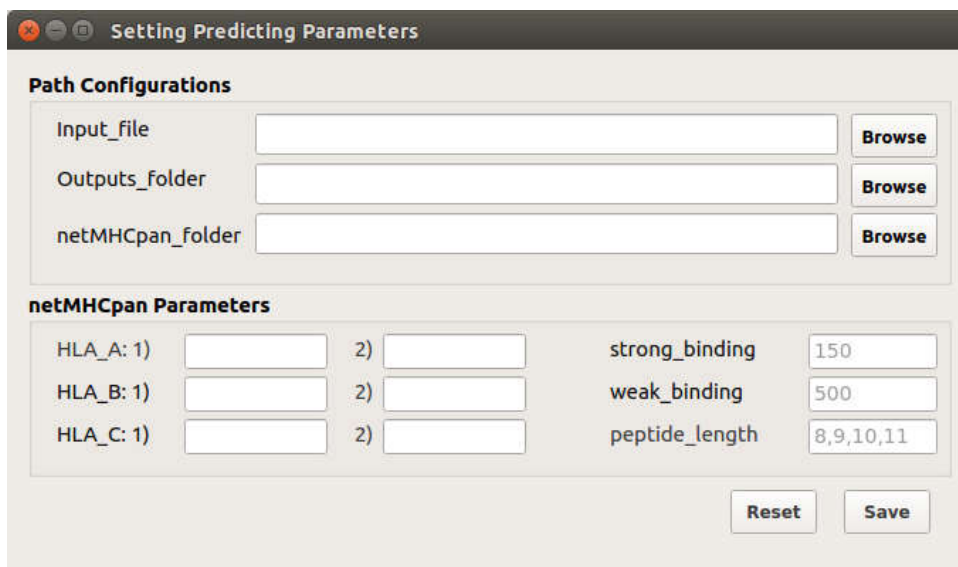


Figure 3. The GUI of setting predicting parameter.



## 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' SMD execution folder. 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.

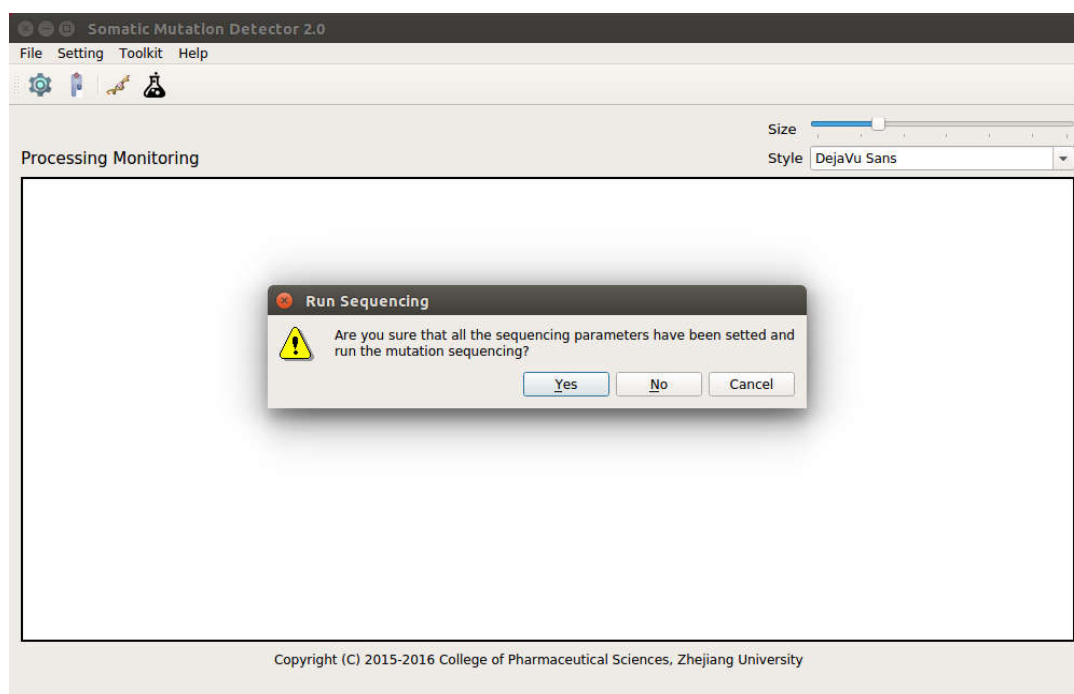


Figure 4. The GUI of confirming the mutation sequencing

### 3.1.1 Sequencing flowchart

SMD 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.

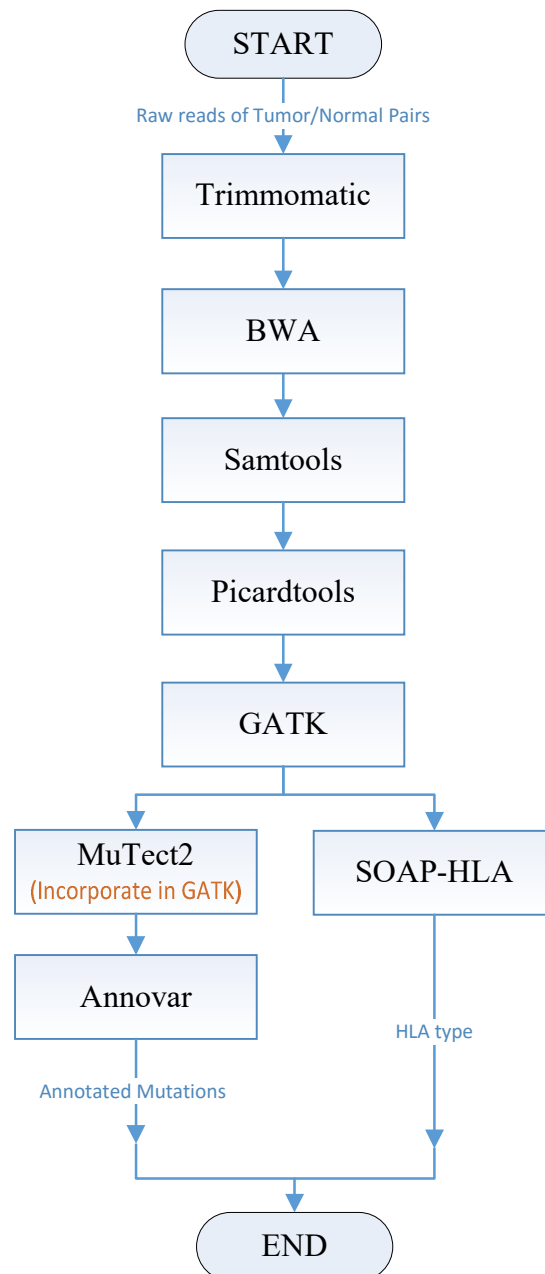
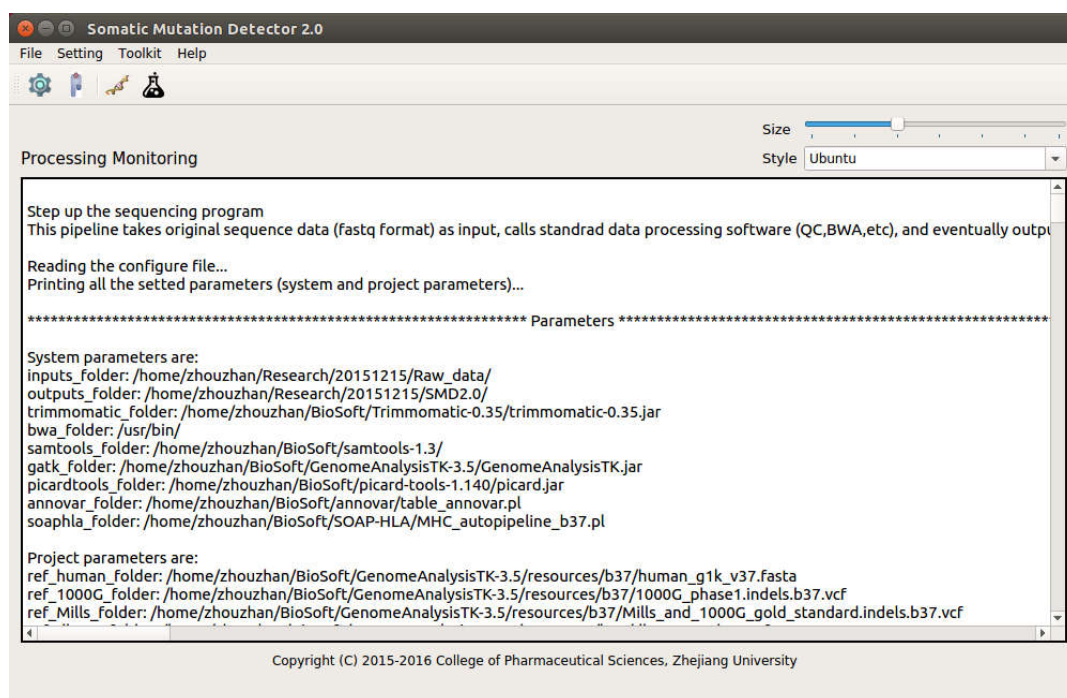


Figure 5. Mutation sequencing pipeline

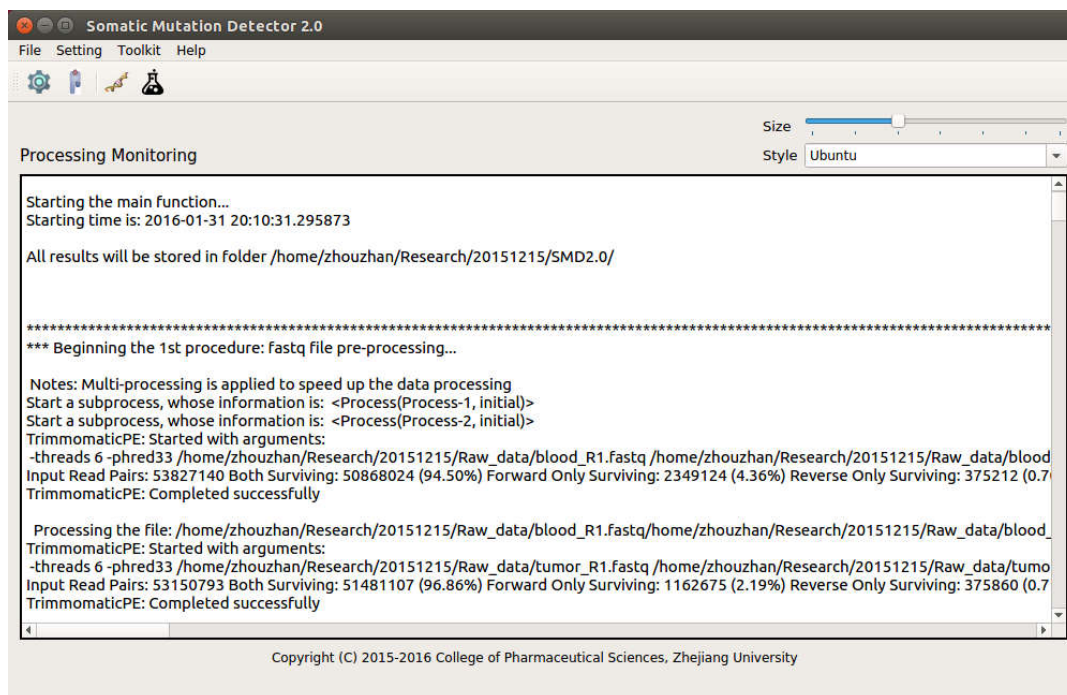
### 3.1.2 Processing monitor display

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

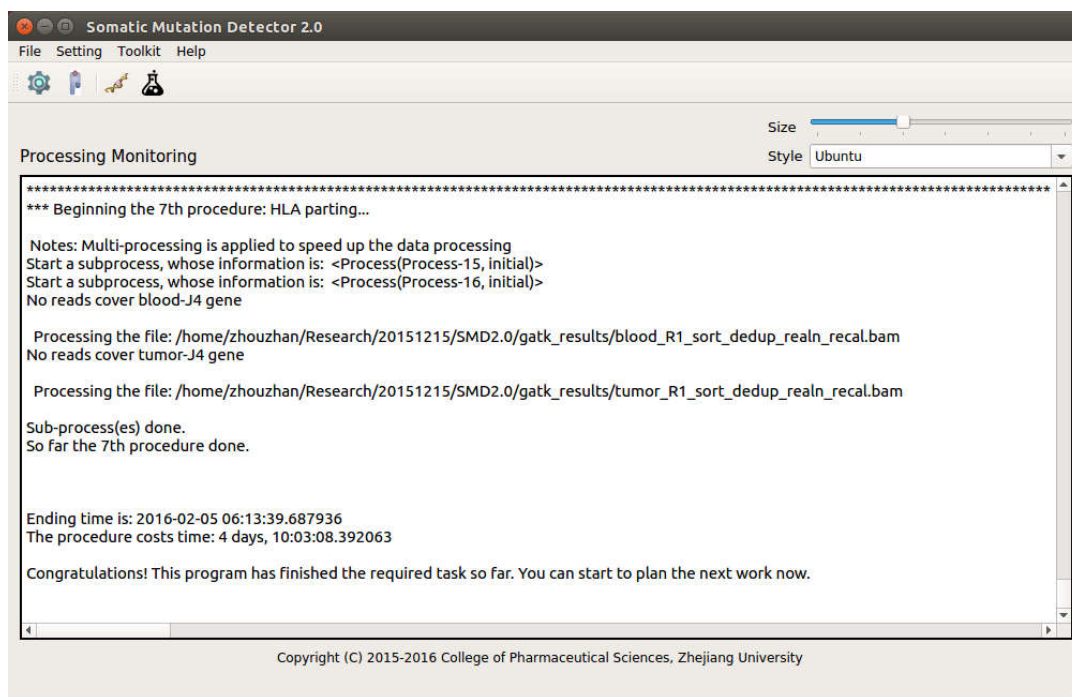
in the processing monitoring window (Figure 6(a)). Then, sequencing pipeline begins to process the raw reads one procedure by one procedure (Figure 6(b)). After the seventh procedure has been done, the whole pipeline ends (Figure 6(c)).



(a)



(b)

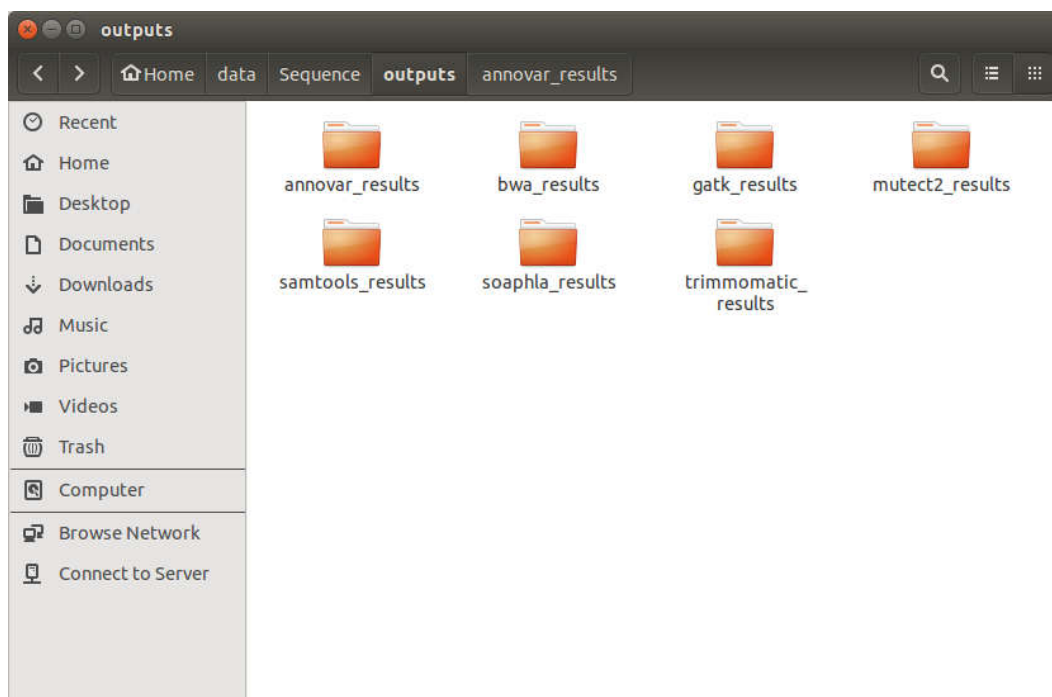


(c)

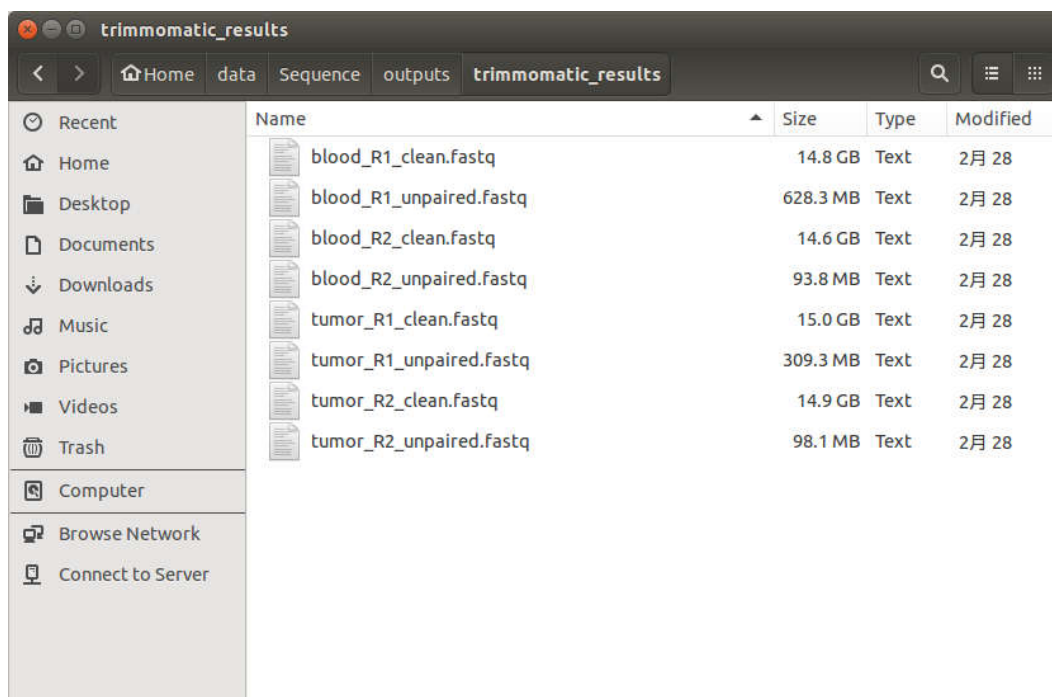
Figure 6. Sequencing pipeline monitoring

### 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

shows as below.

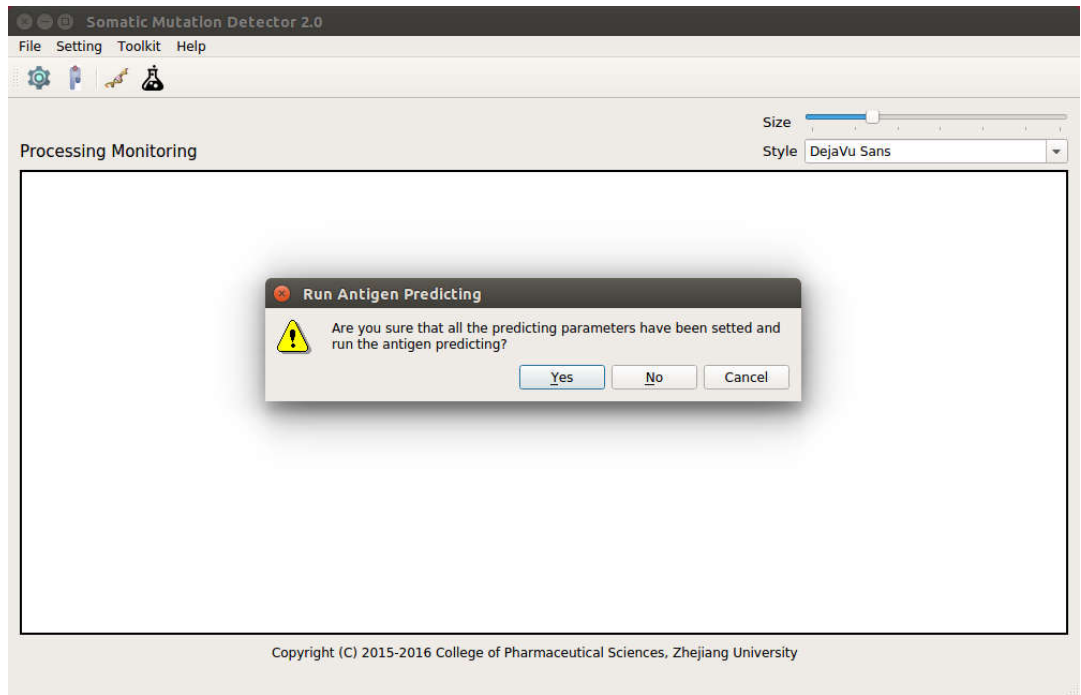


Figure 8. The GUI of confirming the antigen predicting

### 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.

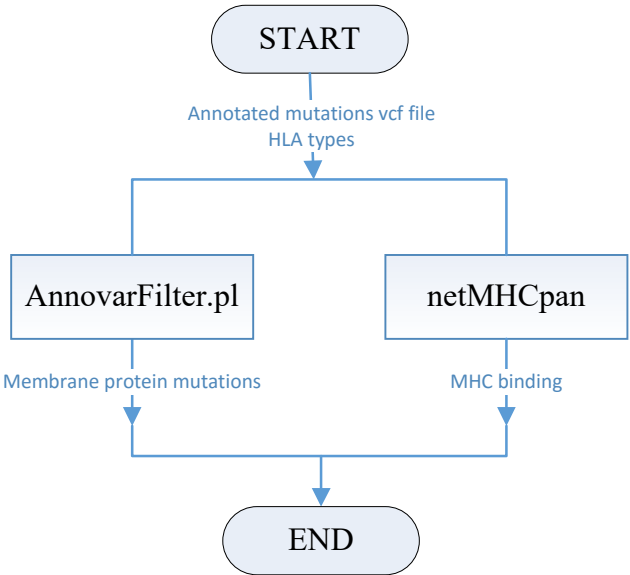


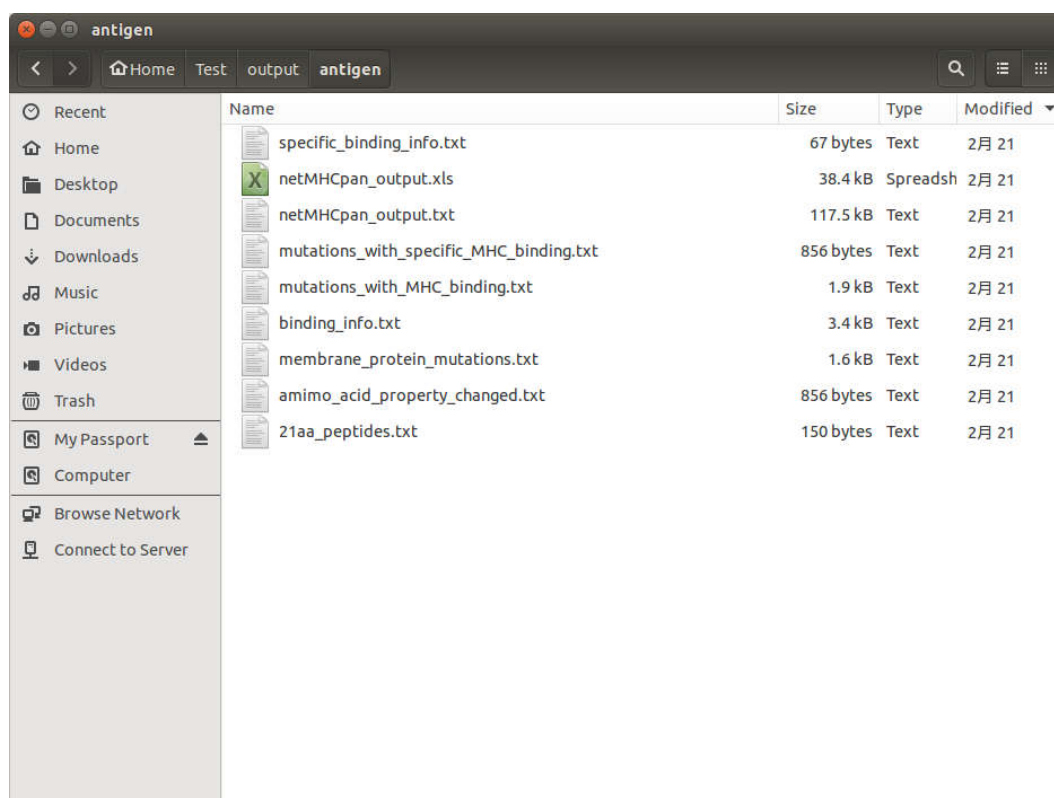
Figure 9. Antigen predicting pipeline

### 3.2.2 Processing monitor display

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

### 3.2.3 Final outputs for antigen predicting

Results for antigen predicting can be shown as Figure 10.



The screenshot shows a file manager window titled 'antigen'. The window has a sidebar on the left with navigation options: Recent, Home, Desktop, Documents, Downloads, Music, Pictures, Videos, Trash, My Passport, Computer, Browse Network, and Connect to Server. The main area displays a list of files with columns for Name, Size, Type, and Modified. The files listed are:

Name	Size	Type	Modified
specific_binding_info.txt	67 bytes	Text	2月 21
netMHCpan_output.xls	38.4 kB	Spreadsh	2月 21
netMHCpan_output.txt	117.5 kB	Text	2月 21
mutations_with_specific_MHC_binding.txt	856 bytes	Text	2月 21
mutations_with_MHC_binding.txt	1.9 kB	Text	2月 21
binding_info.txt	3.4 kB	Text	2月 21
membrane_protein_mutations.txt	1.6 kB	Text	2月 21
amimo_acid_property_changed.txt	856 bytes	Text	2月 21
21aa_peptides.txt	150 bytes	Text	2月 21

Figure 10. Final results for antigen predicting