User manual for USAT

Version 2022-April-12

Contents

1. Software introduction

Universal STR Allele Toolkit (USAT) a standalone bioinformatic software package for Tander repeat (TR) haplotype analysis. USAT has a user-friendly graphic interface and runs in all major computing operating systems at a fast speed with parallel computing enabled. USAT is able to facilitate the interpretation, visualization, and comparisons of TR haplotypes.

Availability: https://github.com/XuewenWangUGA or https://github.com/Ge-Lab

Free for all academic and educational purposes. A license is needed to obtained from us for any industrial and any other purposes.

Contact: HSC Center for Human Identification CBH, 6th Floor 3500 Camp Bowie Boulevard Fort Worth, TX 76107 Local: 817-735-0606 Fax: 817-735-0553

Major features and functions:

1) allele size conversion;
2) sequence comparison of haplotypes;
3) figure plotting for allele distribution;
4) interactive visualization

This manual is for version (1.1) and may work for all later versions. Technical support: xwwang@ymail.com

Important: read through the manual before you use this software.

2. Installation

USAT is released with all other dependency package. Just a JAVA runtime environment, which exists in most computers, is needed. Please update the Java runtime environment into the newest version (JDK 17 or higher) from the link for Java Run Time SE.

https://www.oracle.com/java/technologies/downloads/#jdk17-windows

Download the software and subdirectories from Github. e.g. for software, using the following command in terminal

```
git clone https://github.com/XuewenWangUGA/USAT
```

or

Download the software release zip file from Github into your computer, unzip it into a directory called "USAT".

Go to the directory. Then the installation is completed.

3. Input files

A. Input haplotype file

USAT takes a sequence file with haplotype sequence for each STR and an optional BED file for specific information at each locus. The format is a tabular text file with data like marker1 haplotype sequence SampleID, one haplotype per line. If there are multiple haplotypes, the same marker ID could be used for each locus. Lines with # can be used annotation or comments.

e.g., #CODIS core STR loci for HG002

#Marker_Name Sample_haplotype SampleID

MK1 CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAT S1

MK1 CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAT S0

A test dataset is provided with the software release for testing.

B. Prepare the TR locus file in BED format

The TR locus file in BED format should be provided before running USAT. This BED file is a plain text file, with fields separated with a <Tab> key. Or a user can prepare it in a spread sheet software like the Microsoft Excel and then save it as a table delimited text file. In the BED file, 12 columns of data are presented in the BED file and one column head line is presented.

Then the head line is followed by one or more locus information lines. An example locus file in BED format is provided for Window and Linux system in the subfolder "settings" of software package.

USAT will require the column of Name, Basic_motif_period and Inner_offsetB. Other columns are could be any number or text if match the example below.

The locus information is given in BED format in plain text file (https://genome.ucsc.edu/FAQ/FAQformat.html#format1). Fields are separated by tab. It starts with a head line and then one marker locus per line. Multiple markers can be used. e.g.,

Chrom	ChromStart	ChromEnd	**Name**	Left_offset	Right_offset	**Basic_motif_period**	Ref_hap_length	Motif	Ref_allele	**Inner_offset**	Min_stutter_threshold
chr1	230769615	230769683	D1S1656	3	3	4	68	CCTA	[TCTA]n TCA [TCTA]n	17	0	0.1

Meaning of each column data or field:

Chrom: the name of chromosomal sequence or reference sequence. E.g., chr1, chrX, chrM for human genome reference h38.

ChromStart: This is related to the start coordinate position of a tandem repeat. It is calculated as the start position (1-based coordinate) in chromosomal sequence minus 1, e.g., 230769616 on chr1 of h38 should be 230769615.

ChromEnd: This is the end coordinate position of a tandem repeat.

Name: the name of the locus or any alternative name of the locus, e.g., the STR on chr1 between coordinate 230769616 and 230769683 is named as D1S1656.

Left_offset: a positive integer value of the distance in bp from the start position of a TR, e.g. 3.

Right_offset: a positive integer value of the distance in bp from the end position of a TR, e.g. 3.

Basic_motif_period: the count of bp in the basic motif unit, e.g. 4 for motif [TCTA]n.

Ref_hap_length: the length in bp of the whole TR in the reference sequence, e.g. 68 for D1S1656

Motif: the basic motif in nucleotides of a TR in the reference sequence, e.g. CCTA [TCTA]n TCA [TCTA]n where the repeated motif in a square parenthesis, n for a repeated motif.

Ref_allele: the length-based allele size of repeated times in the reference sequence, e.g., 17. If unknown, use 0 or any positive integer instead.

Inner_offset: the number of nucleotides which should be excluded in counting length-based allele size of repeated times, e.g. 0 for D1S1656, 8 for forensic D19S433 in human CODIS core STR loci. Use 0 if unknown.

Min_stutter_threshold: the minimum proportion or frequency cutoff to filter out the noisy allele for this locus. If unknown, this value should be provided as -1 or any negative number. Then the tool will use the general minimum frequency e.g., 0.01 for this locus.

4. Run USAT

Here demonstrate how to run analysis with USAT in Windows. For other operating system like MacOS or Linux, just follow the same operation as that for Windows system.

Step 1. start to run USAT

One of the following two methods can be used to start the run of USAT.

method 1:

double click the USAT.jar file in the installed directory USAT to run

or

mehotd 2 in command terminal, type the command and then press the <enter> key

```
java -jar USAT.jar
```

After start USAT, a graphic window will appear (Figure 1).

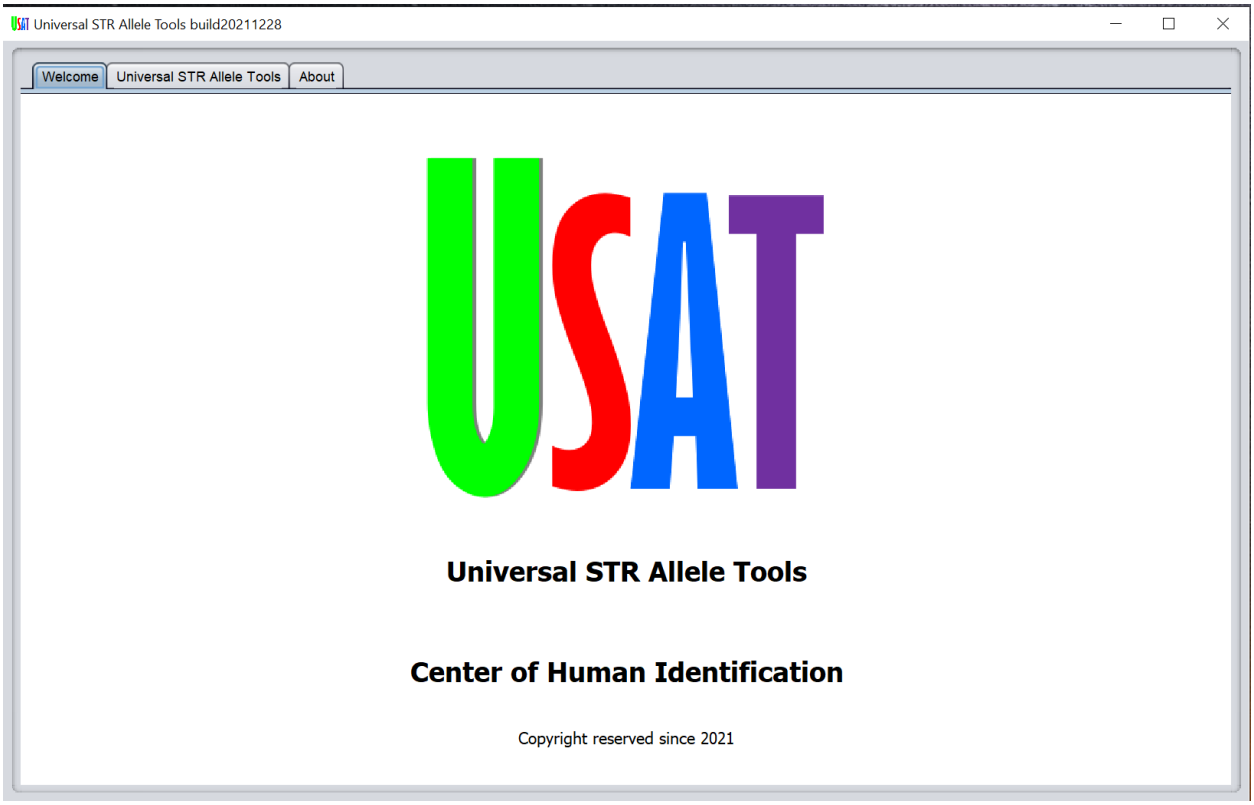

Figure 1 Start window of USAT

Step 2. Load input data and run an initial analysis

Load the haplotype sequence data and BED file as prepared previously. Here we used the demonstrate data.

Click the <**Universal STR Allele Tools**> tab next to <**Welcome**> tab, then an interface for selected input data shows up (Figure 2).

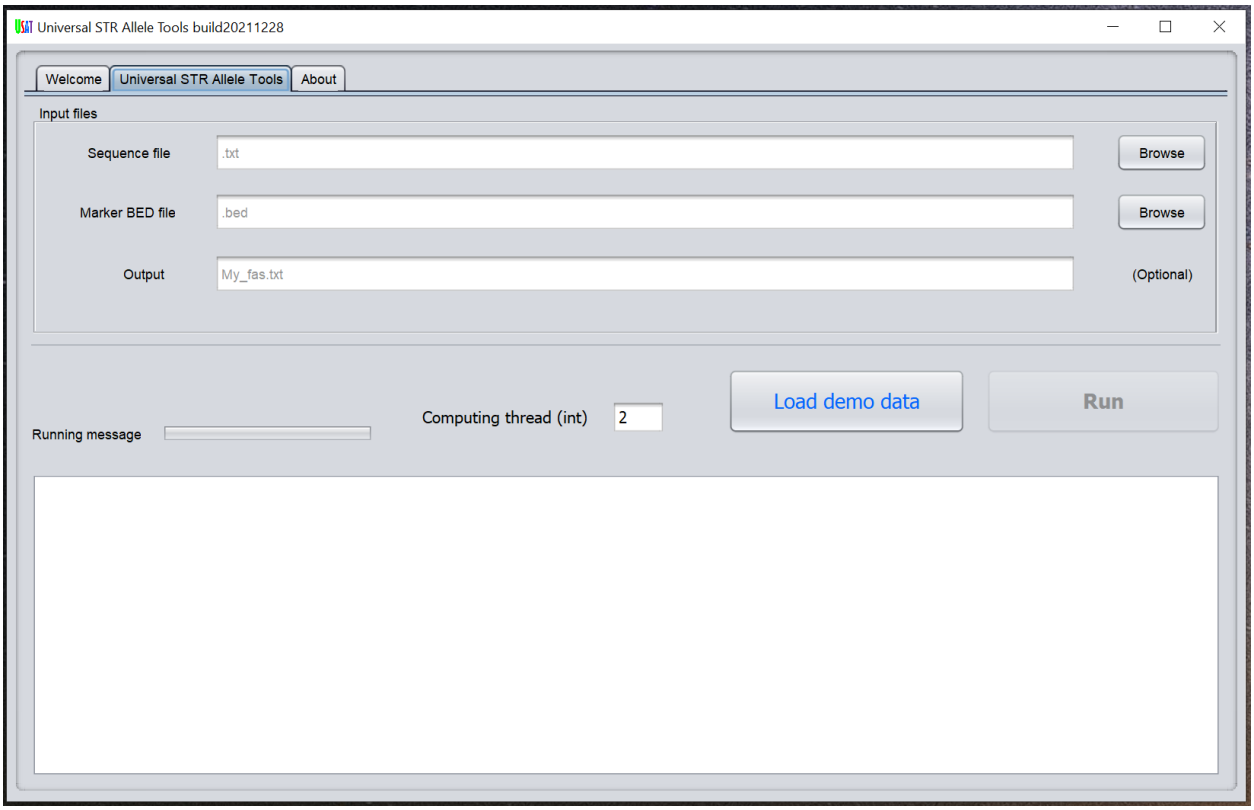

Figure 2 Input interface of USAT

Then click <**Browse**> button to select the input file in popped window. Here we selected the **MkHaplotype.txt** in **testData** folder released with USAT (Figure 3). Click **Open** to confirm the file selection. After that, the file and path to the file is selected and displayed in the box next to <**Browse**> (Figure 4).

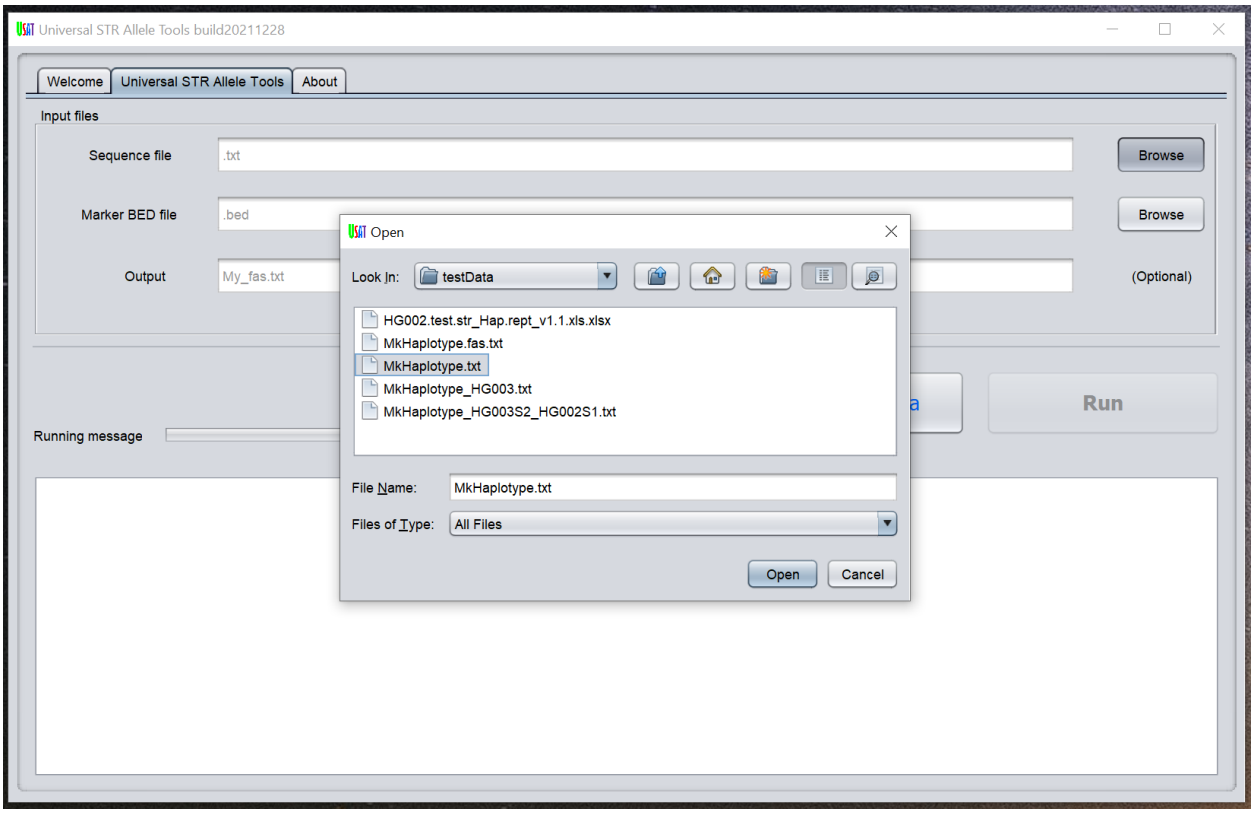

Figure 3 Selecting of Input sequence file

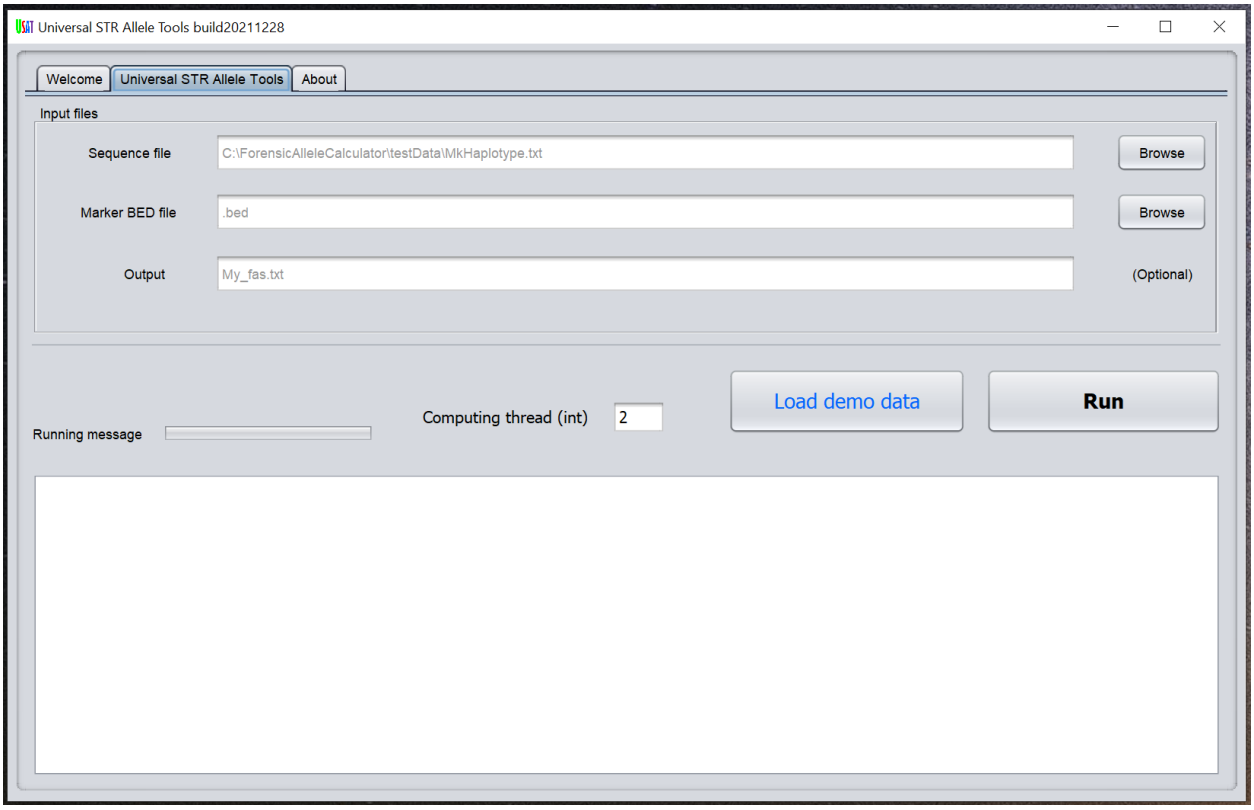

Figure 4 Selected sequence file

Use the similar way to select the BED file next to <**Browse**> for the second input file.

For beginner, the demo data can be automatic selected after click the button <Load demo data> (Figure 5).

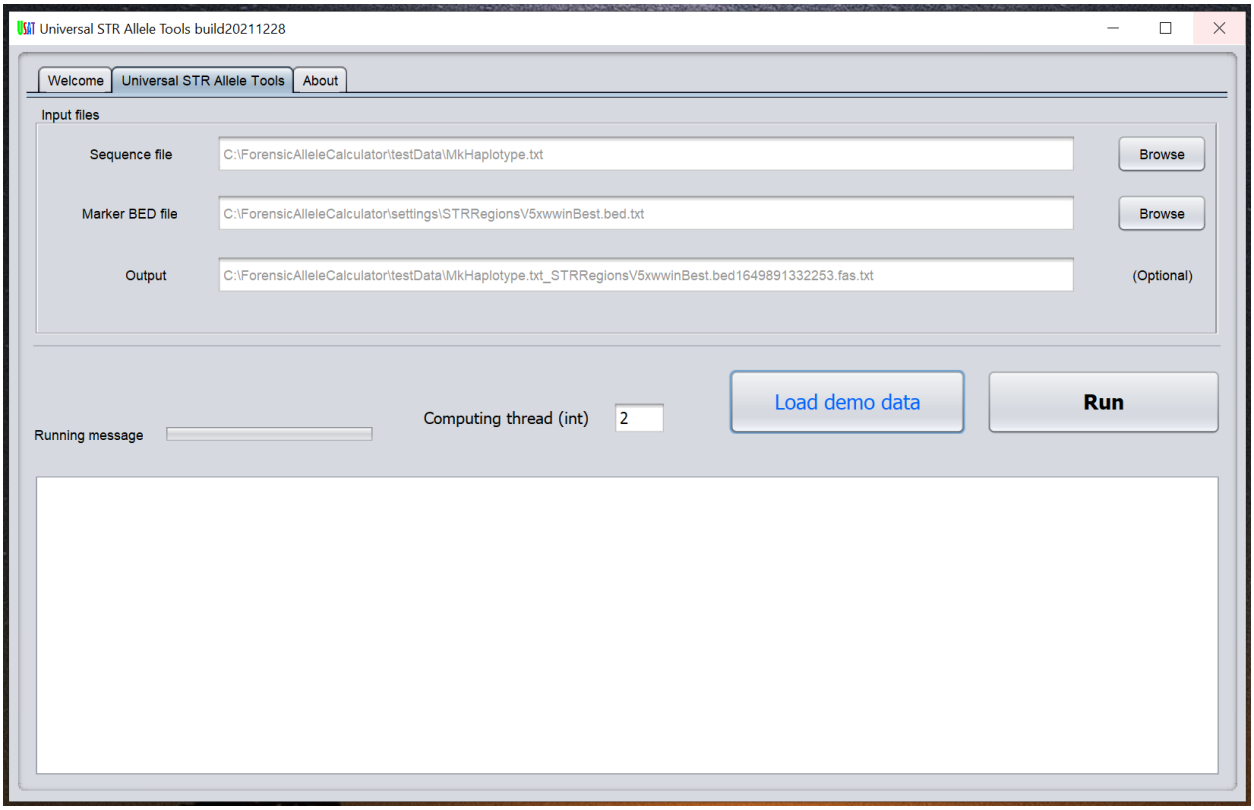

Figure 5 input files loaded from demo data

The third box below the BED file box is automatically generated for holding the temporary data in a file, and doesn't need to be changed. If you want to change it, you can specify the path and file name as needed.

Then you can click <**Run**> button to start TR haplotype analyses , and a view and plot window will pop up showing the results. The previous window will display the status, log and summary information in an interactive table for the run (Figure 6). The number of computing thread is set to 2 by default which is enough for most dataset. For a large data set, greater number of the computing thread will result in a faster speed. Thus, user can increase the number as needed.

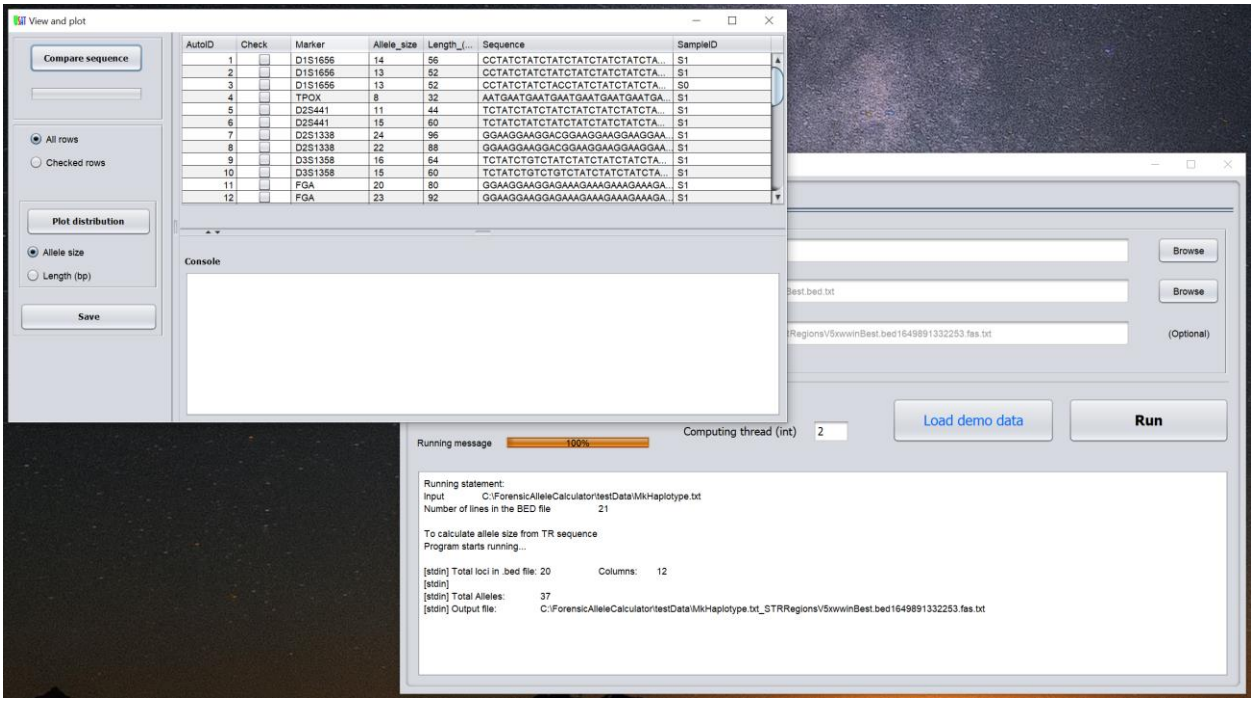

Figure 6 Interface after click Run

Step 3. Compare the sequences of haplotypes

To compare sequences of haplotypes, if want to compare all sequences in the interactive table, check <**All rows**>. If want to compared specific sequences, just click the check box(es) in Check column of the interested rows and check <**Checked rows**> circle. At least two rows should be selected, and multiple number of rows can be selected. Here we checked three sequences and then click button <**Compare sequence**>. Then comparing between sequences will be running. When it is finished, the result is displayed in the following box, including the running log and sequence alignment. The alignment shows the difference and similarity of haplotypes (Figure 7).

There are several possible and correct alignment of TR haplotypes. The user can edit to adjust the alignment into their preferred in the box if needed.

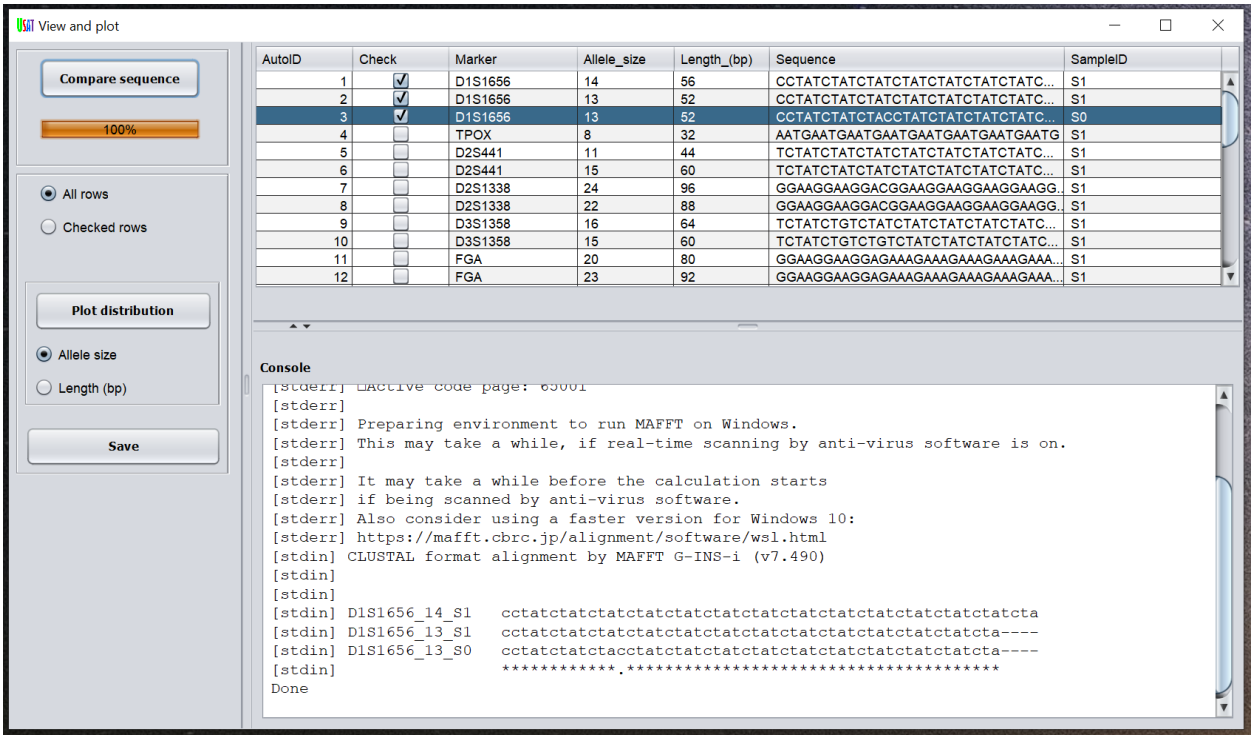

Figure 7 Haplotype comparison

Step 4. Plot and save distribution figure, comparison allele size

For all plot in USAT, to zoom in the plot, press the left key of mouse, draw the rectangle to right side. To zoom out the plot, press the left key of mouse, draw the rectangle to left side. To change the ratio and size of plot area, drag the plot window to change. To save or export the plot, right click the mouse to save in preferred format, e.g. jpg, svg.

To compare and see the atlas of length or size of haplotypes for all sequences in the interactive table, check option <**All rows**> and option <**Allele size**>, then click button "**Plot distribution**". A new window will display the allele-size atlas of all haplotypes in the table (Figure 8).

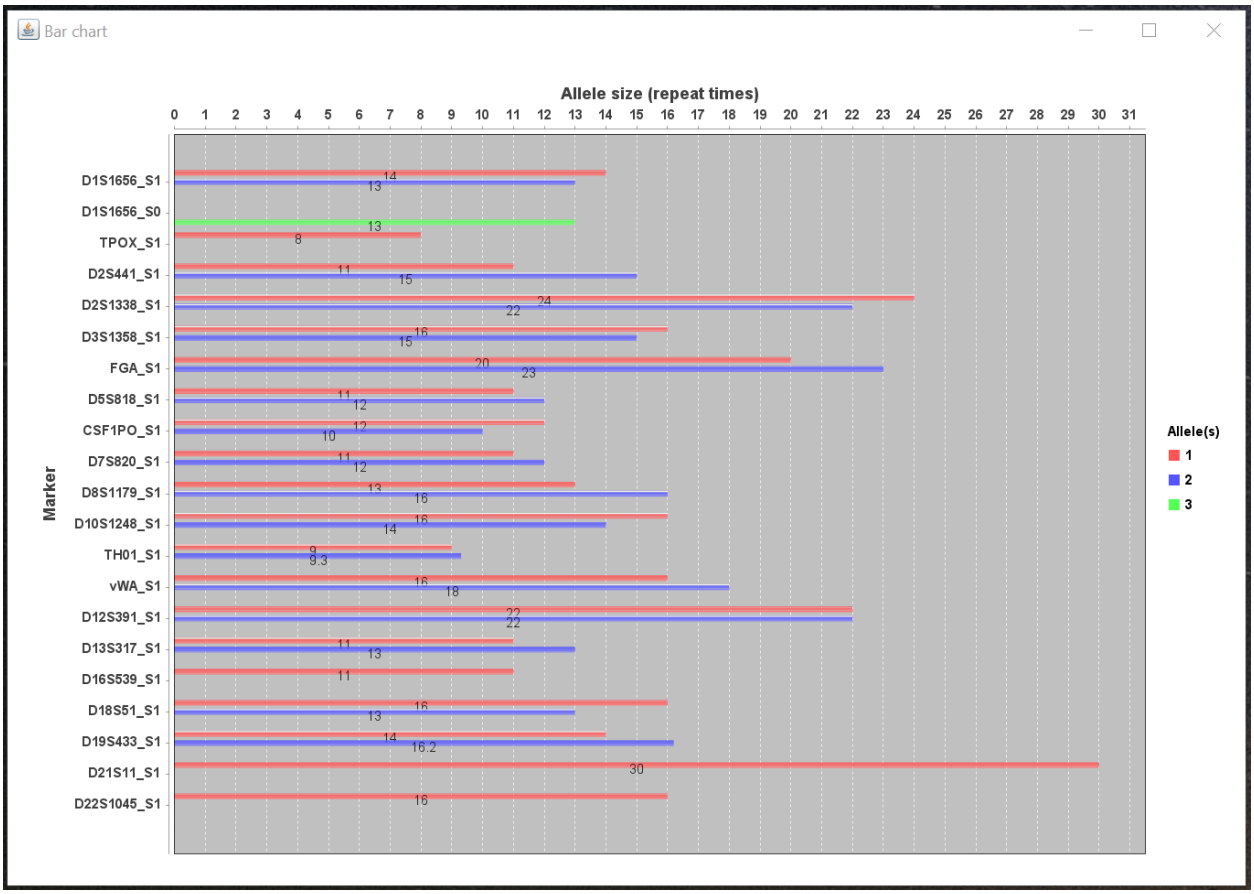

Figure 8 Comparison of atlas of allele size

check option <**Checked rows**>, option <**Allele size**>, and check the interested rows; then click button "**Plot distribution**". A new window will display the allele-size atlas of haplotypes in checked rows (Figure 9).

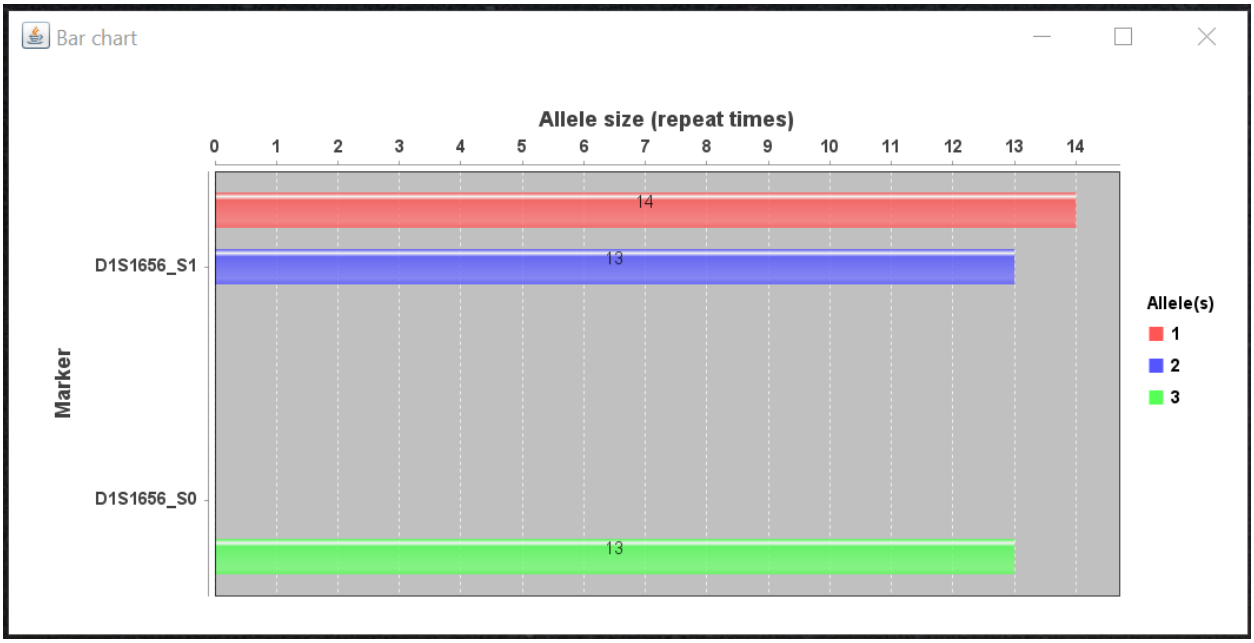

Figure 9 Plot and comparison of allele size of checked haplotypes

Similarly, if option <**Length (bp)**> is checked, the plot will be in base pair, instead of allele size (Figure 10).

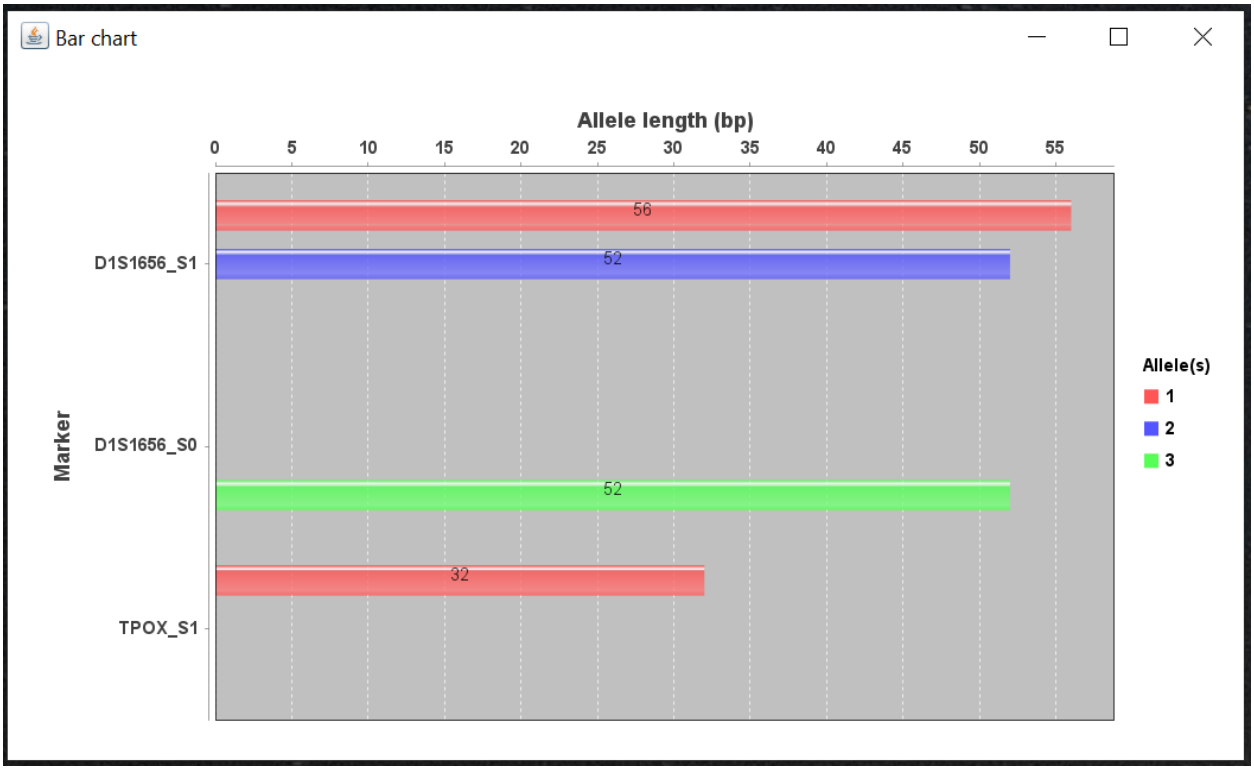

Figure 10 Plot and comparison of allele length of checked haplotypes

Step 5. Change the display of the interactive table

The displaying table is an interactive table. Following the annotation to interact with the table.

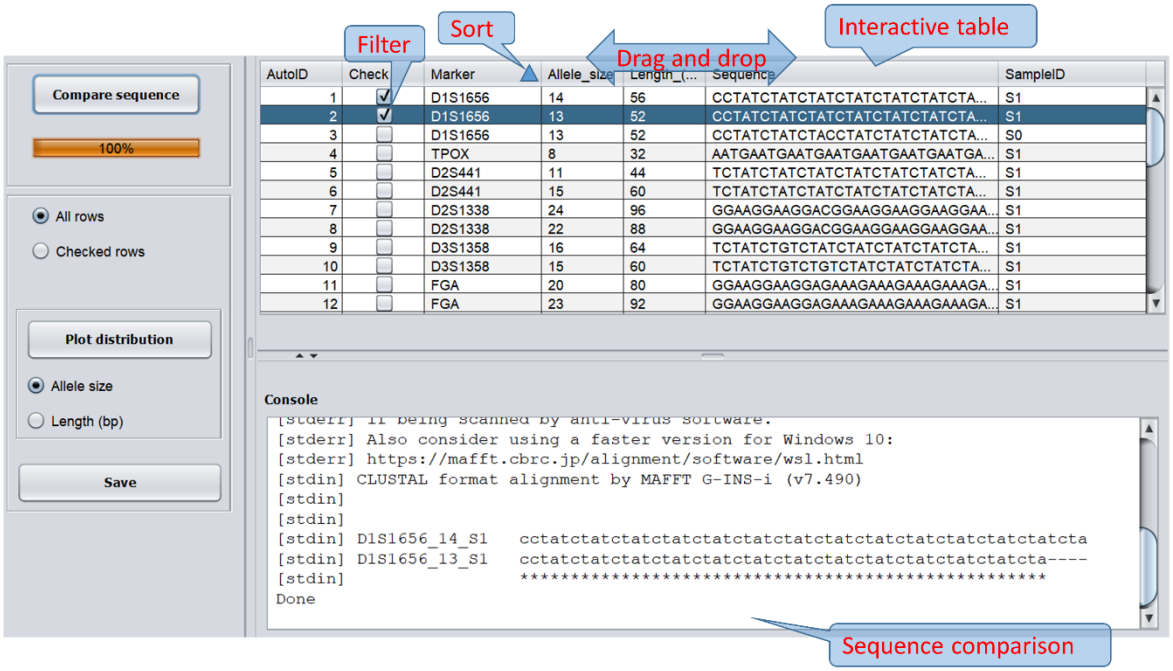

Figure 11 Interactive table

Step 6. Options to export the data in interactive table

Select the option <**All rows**> and <**Checked rows**> to save the full interactive table and selected rows into a user defined file in plain text with tabular delimit (Figure 12).

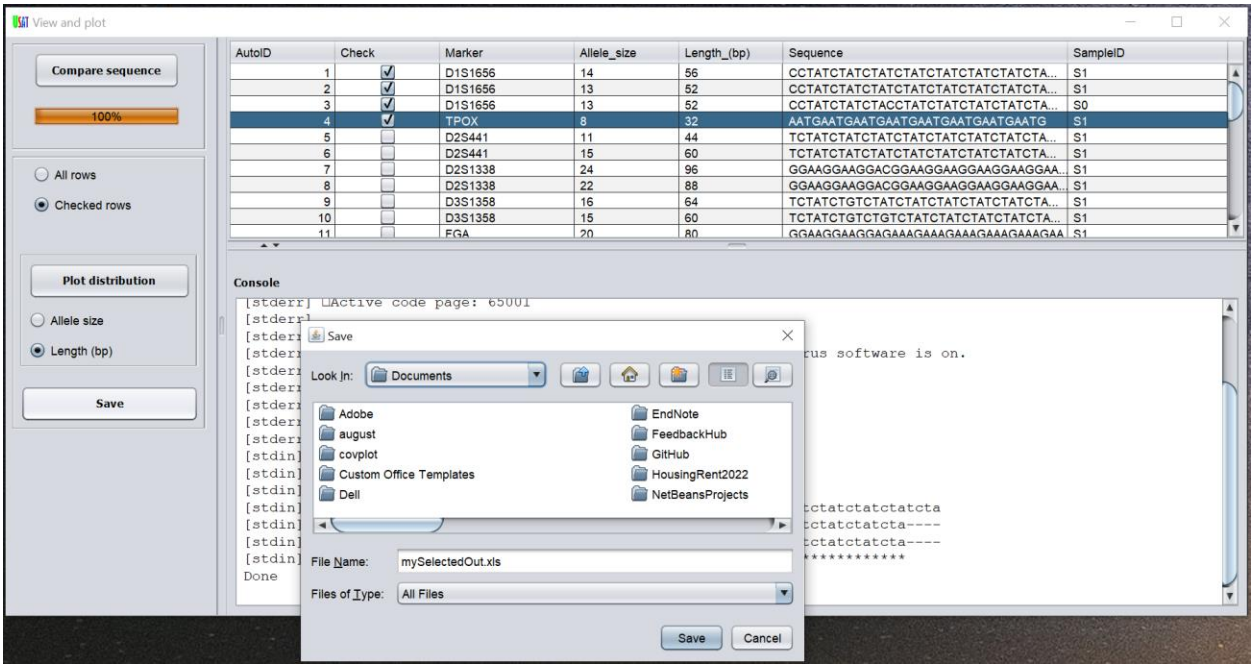

Figure 12 Save the customized content interactive table