User manual for TRcaller (V1.5.4)

Feb 28, 2023

Contents

1. Software introduction

TRcaller is a software program to precisely and quickly detect tandem repeats (TR) in DNA sequences. With the targeted TR regions defined, TRcaller is able to call the TR alleles from either short or long read sequences, either whole genome or targeted sequences, in just a few seconds with >99% accuracy.

There is no minimum read depth required for the input BAM file. TRcaller only detects the TR alleles that span the whole TR region and does not connect the multiple broken reads to infer potential long TR alleles. We recommend using long-read sequencing technology to generate sequences to detect TR alleles longer than 250bp.

For loci with uncertain ploidy, such as human X chromosome STR, it is recommended to put the maximum number of ploidy in the loci config file. Therefore, TRcaller may detect multiple alleles than expected. For example, TRcaller will output two alleles for an X-STR locus, as the ploidy of X-STR is set as 2 in the predefined loci config file and the gender of the sample is not a parameter for TRcaller. It would be the user's discretion to decide whether the minor allele is a true allele or noise.

2. Data preparation

A. Prepare the TR locus file in BED format

A BED file, which describes the details of the targeted TR loci, is required to run TRcaller. This BED file is a plain text file with fields separated by <Tab> keys. It is basically a tab-separated file, which can be prepared in a spreadsheet software, such as Microsoft Excel. The BED file should include 10 columns, and the first row (or the headline) describes the name of each column. The meaning of each column is explained as follows.

The headline should be:

Chrom	**ChromStart**	**ChromEnd**	**Name**	**Repeat_unit_length**	**Motif**	**Ref_hap_length**	**Ref_allele**	**Stutter_ratio_threshold**	**Ploidy**	**Inner_offset**

Each of the following lines presents a targeted TR locus. For example,

chr2	176093058	176093103	SPD1:HOXD13	3	[GCN]n	39	13	0.25	2	0
chr2	190880872	190880920	GDPAG:GLS	3	[GCA]n	48	16	0.25	2	0
chr3	45540738	45540802	D3S1358	4	[TCTA]n [TCTG]n [TCTA]n	64	16	0.25	2	0
chr3	63912685	63912715	SCA7:ATXN7	3	[CAG]n	30	10	0.25	2	0
chr3	129172576	129172656	DM2:CNBP	4	[CAGG]n	80	20	0.25	2	
chr5	150076323	150076375	CSF1PO	4	[ATCT]n	52	13	0.25	2	0
chr19	29926234	29926298	D19S433	4	[CCTT]n ccta [CCTT]n cttt [CCTT]n	64	14	0.25	2	8
chrY	6993189	6993257	DYS570	4	[TTTC]n	68	17	0.25	1	0
chrY	7547584	7547624	DYS522	4	[ATAG]n	40	10	0.25	1	0
chrX	9402261	9402301	DXS8378	4	[ATAG]n	40	10	0.25	2	0
chrX	134481508	134481560	HPRTB	4	[ATCT]n	52	13	0.25	2	0

The meaning of each column data or field:

Chrom: the name of a chromosomal or reference sequence. It should be exactly as that in the BAM file. For example, chr1, chrX, chrM for human genome reference.

ChromStart: the start coordinate position of a tandem repeat in BED format (0 based position). It is calculated as the start position (1-based coordinate) in chromosomal sequence minus 1. For example, the first nucleotide of D1S1656 in the reference genome (HG38) locates at position 230769616 on chr1, and then the ChromStart of D1S1656 in the BED file should be 230769615.

ChromEnd: the end coordinate position of a tandem repeat (1 based position). It is calculated as the end position (1-based coordinate) in chromosomal sequence. For example, the last nucleotide of D1S1656 in the reference genome (HG38) locates at position 230769683 on chr1, and then the ChromEnd of D1S1656 in the BED file should be 230769683.

Name: the name of the locus. For example, D1S1656. The name itself is not used in the algorithm of allele detection.

Repeat_unit_length: the length in base pair (bp) of the motif unit (i.e., repeat unit). For example, if the motif is [TCTA]n, then the Motif_length is 4, as the length of TCTA is 4.

Motif: the basic motif in nucleotides of a TR in the reference genome. For example, the motif of D1S1656 is CCTA [TCTA]n TCA [TCTA]n, in which the repeated motif is in square brackets and n represents a repeated motif. The repeated motif sequence should be put in square brackets []. If the motif consists of a complex of multiple basic units, each unit should be put in square brackets and separated by a single space (e.g., [TCTA]n [GCAT]n). The motif should not contain any <tab> space.

Ref_hap_length: the length in base pair (bp) of the whole TR in the reference genome. For example, the length of D1S1656 in human HG38 is 68.

Ref_allele: the length-based allele size of repeated times in the reference sequence. For example, the Ref_allele of D1S1656 is 17. Use 0, if it is unknown.

Stutter_ratio_threshold: the maximum read depth ratio to exclude a minor peak as a stutter. This value should be between 0 and 1. It is recommended to set this threshold as 0.25, if you don't know what threshold to set. This threshold will only be used for single source samples and will not be used for mixture samples (i.e., there may be multiple donors contributing to this sample). The read depth ratio can be calculated as the read depth of a minor allele (i.e., the allele with lower read depth) divided by the read depth of the major allele (i.e., the allele with higher read depth). If a stutter ratio is higher than the threshold, the minor allele will be considered as a stutter and will not be output as a true allele.

Ploidy: the number of ploidy of this locus. For example, the ploidy of autosomal loci in the human genome is 2, and the ploidy of Y chromosome loci in the human genome is 1.

Inner_offset: the number of nucleotides which should be excluded in counting length-based allele size of repeated times. For example, there are 8 nucleotides in D19S433 that are not counted toward the length of the TR locus. Thus, the Inner_offset for D19S433 is 8. Inner_offset is not very common. If you are not sure if there is any Inner_offset for a particular locus, please set it as 0.

Multiple predefined BED files can be downloaded on this website, such as CODIS STRs, common human X-STRs and Y-STRs, and common disease-associated TRs. The coordinates of these predefined BED files are for human HG38.

B. Sort and index BAM

If the input BAM file is not sorted or indexed, please use one of the following methods to sort and index the BAM file, because TRcaller only accepts the sorted and indexed BAM file as input.

Method 1:

Use samtools commands, “samtools sort” and “samtools index”. The details of samtools can be found at https://github.com/samtools/samtools.

Command:

```
samtools sort -o input.sort.bam -O bam input.bam
samtools index -b input.sort.bam
```

Method 2:

Download BamAlignSortIndex.jar and run the following command on your local computer. This tool has been tested on Windows, Linux, and MacOS. Please make sure you have the latest version of JAVA installed, https://www.oracle.com/java/technologies/downloads/.

Command:

```
java -jar BamAlignSortIndex.jar input.bam
```

The command will generate two output files, input.sort.bam and input.sort.bam.bai, which are ready to upload to TRcaller website.

C. Reduce the size of the BAM input file

Although TRcaller can process sorted and indexed BAM files with any size, the server may not be able to handle large BAM files due to limited resources. If the BAM file size is larger than the current limit, please reduce the BAM file size before uploading the BAM file using one of the following two methods.

Method 1.

Download BamSubset.jar and run the following command on your local computer. This tool has been tested on Windows, Linux, and MacOS. Please make sure you have the latest version of JAVA installed, https://www.oracle.com/java/technologies/downloads/.

```
java -jar BamSubset.jar    Bed_file    Input.bam    reduced_input.bam
```

For example,

```
java -jar BamSubset.jar    STR.bed    in.bam    in.reduced.bam
```

Method 2.

Use the **samtools** command **view**. The details may be found at http://www.htslib.org/doc/samtools-view.html.

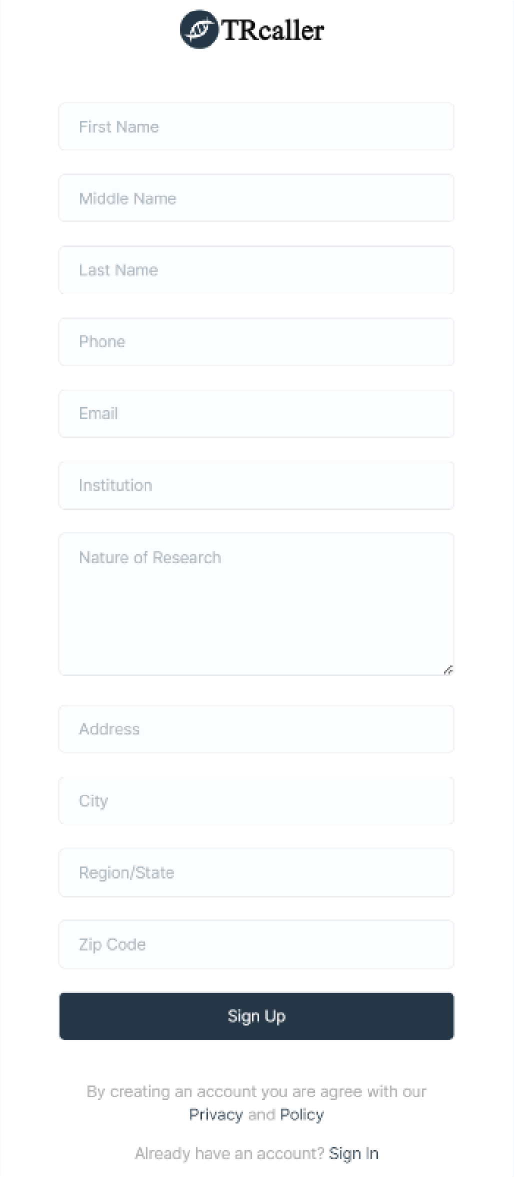
For example,

```
samtools view in.bam --region-file STR.bed --output in.reduced.bam
```

3. Run TRcaller

A. Sign up and sign in

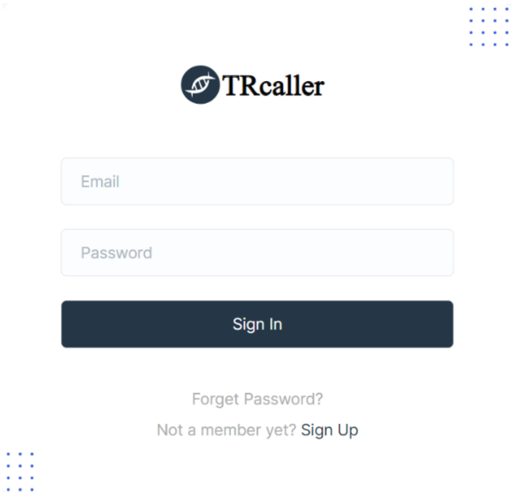
Sign up: https://www.trcaller.com/SignUp.aspx



By creating an account you are agree with our Privacy and Policy

Already have an account? Sign In

Sign in: https://www.trcaller.com/SignIn.aspx



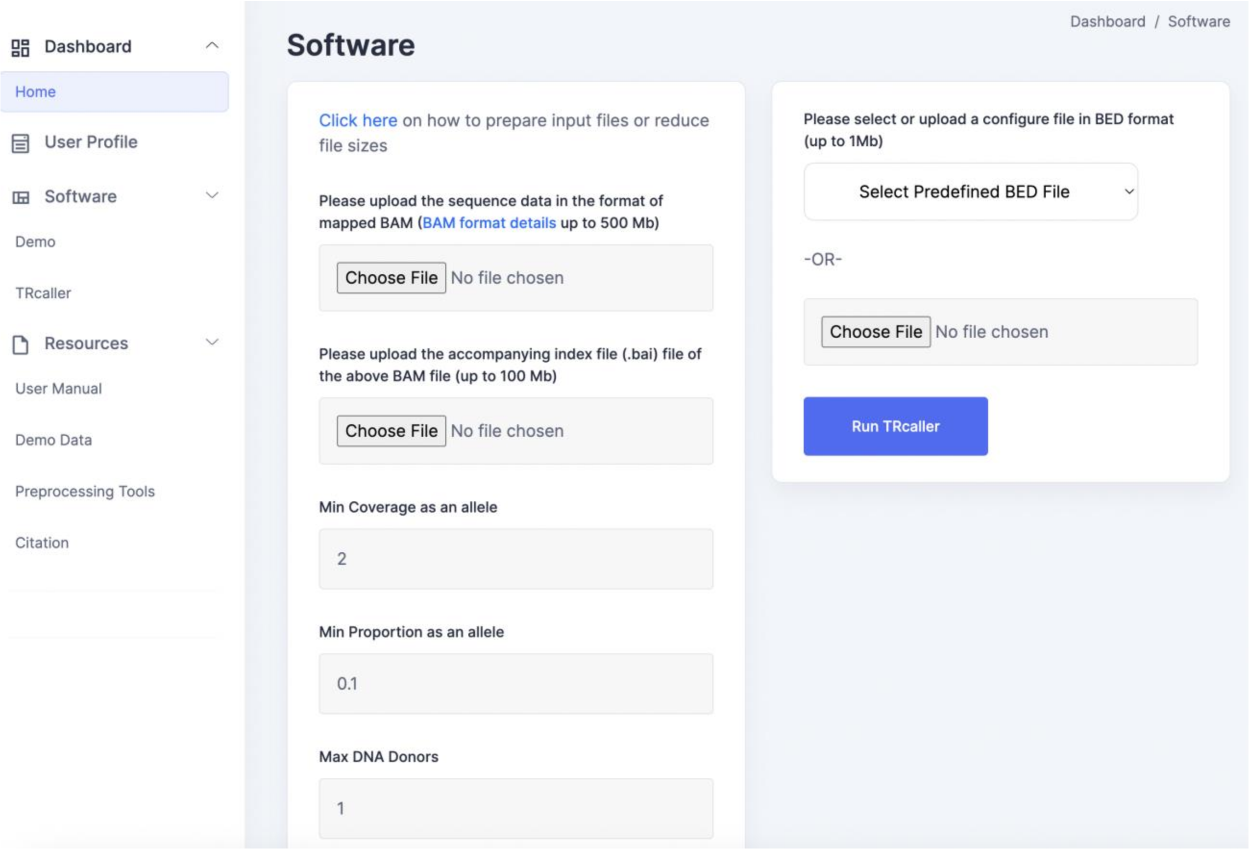
B. Upload files, Parameters, and Run TRcaller

Users need to upload three files to run TRcaller:

1. **BAM** file, which includes the DNA sequences, such as A.bam
2. **BAI** file, which is the index file of the BAM file. The file name has to include the BAM file name and add ".bai", such as A.bam.bai
3. **BED** file, which is the config file to describe the details of the targeted TR loci.

There are three parameters the users need to decide:

1. **Min coverage as an allele**: the minimum number of coverage to be considered as an allele. This threshold should be an integer, ranging from 1 to 1,000,000, and the default is 2.
2. **Min proportion as an allele**: the minimum proportion of an allele among all reads of a targeted TR locus to report as an allele. This threshold should be between 0 and 1, with a default of 0.1. For example, if this threshold is set at 0.1 and 100 reads are detected at a locus; then, a sequence "A" with 9 reads will be considered as noise (i.e., 9/100 = 0.09 < 0.1), and a sequence "B" with 10 reads will be considered as a true allele (i.e., 10/100 = 0.1 ≥ 0.1).
3. **Max DNA donors**: the number of donors in this sample. This number is an integer, ranging from 1 to 10,000, with a default of 1. If the user believes that this sample is from a single person/donor, this number should be 1. If this sample may be a mixture of up to 3 donors, this number should be 3. The maximum number of alleles of a locus is the product of this threshold and the **Ploidy** in the BED file. For example, if Ploidy is 2 and Max DNA donors is 3, TRcaller may report up to 6 alleles. For a special scenario with tri-allele locus, users may set **Ploidy** to 3 in the BED file for that locus.



Click here on how to prepare input files or reduce file sizes

Please upload the sequence data in the format of mapped BAM (BAM format details up to 500 Mb)

Choose File No file chosen

Please upload the accompanying index file (.bai) file of the above BAM file (up to 100 Mb)

Choose File No file chosen

Min Coverage as an allele: 2

Min Proportion as an allele: 0.1

Max DNA Donors: 1

Please select or upload a configure file in BED format (up to 1Mb)

Select Predefined BED File

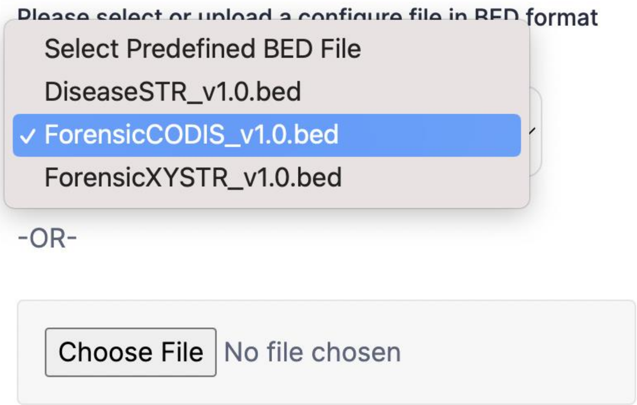
-OR-

Choose File No file chosen

There are three predefined BED files that the user may select. All these BED files are for human HG38 coordinates.

1. DiseaseSTR, including 60 disease-associated TR loci. The configuration details were collected from Chintalaphani et al. (2021) and STRipy.org.
2. ForensicCODIS, including 20 CODIS core loci defined by FBI. The configuration details were collected from Parson et al. (2016)
3. ForensicXYSTR, including 32 Y-STR loci and 7 X-STR loci. The configuration details were collected from Parson et al. (2016)

Users may also upload their own BED files, following the same format described in section 2.A.



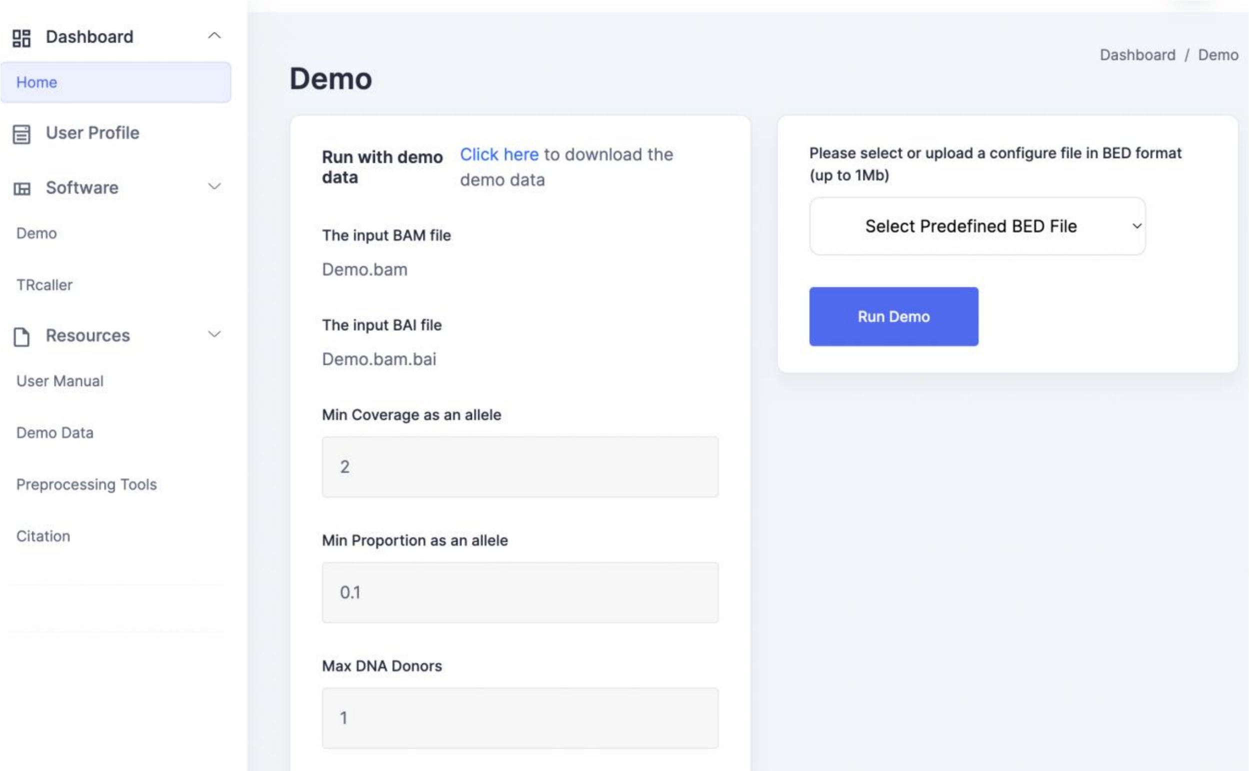
Reference:

[1] Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. Acta neuropathologica communications. 2021 May 25;9(1):98.

[2] Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, De Knijff P. Massively parallel sequencing of forensic STRs: considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. Forensic Science International: Genetics. 2016 May 1;22:54-63.

C. Demo

Users may use demo data to have a test run of TRcaller. The demo data was extracted from the HG002 sample in the Genome In A Bottle (GIAB) project.



Demo

Run with demo data — Click here to download the demo data

The input BAM file: Demo.bam

The input BAI file: Demo.bam.bai

Min Coverage as an allele: 2

Min Proportion as an allele: 0.1

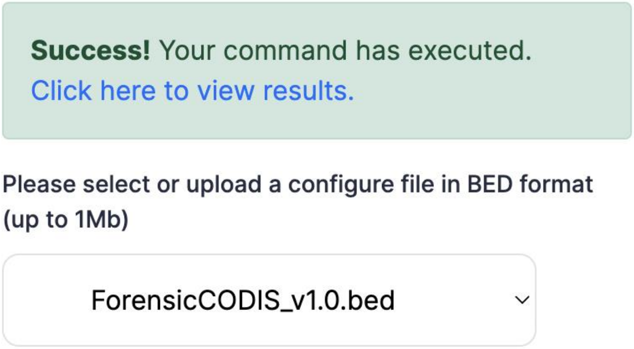
Max DNA Donors: 1

Please select or upload a configure file in BED format (up to 1Mb)

Select Predefined BED File

D. Results

Once TRcaller finishes the analysis, the website should show a link to the analysis results.

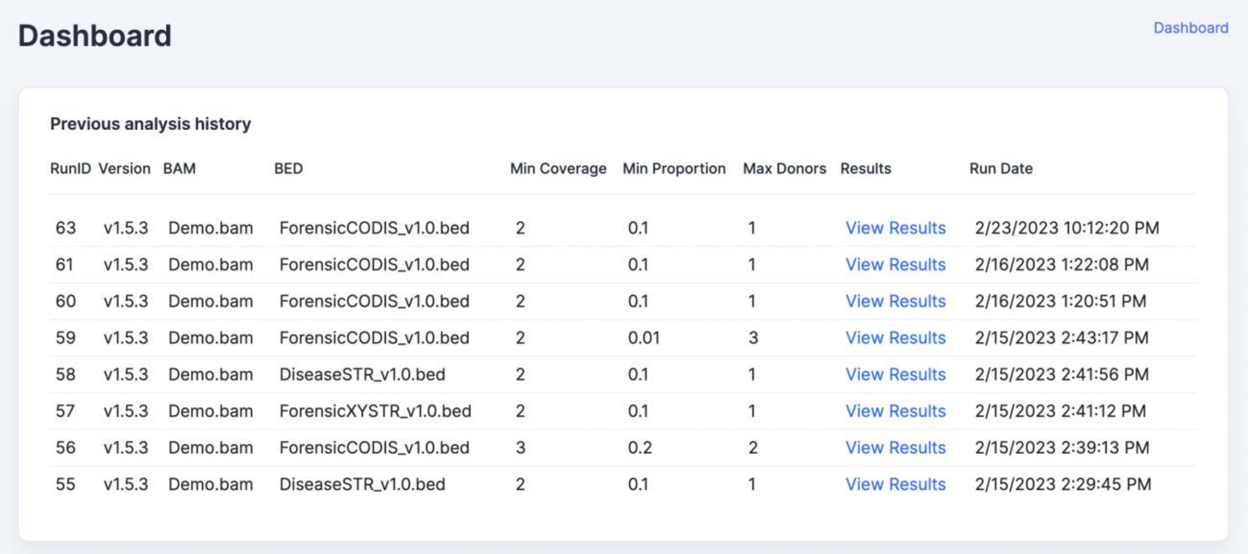


Success! Your command has executed.
Click here to view results.

Please select or upload a configure file in BED format (up to 1Mb)

ForensicCODIS_v1.0.bed

The link will lead the users to the dashboard, in which all the past analyses are listed and sorted by running time.

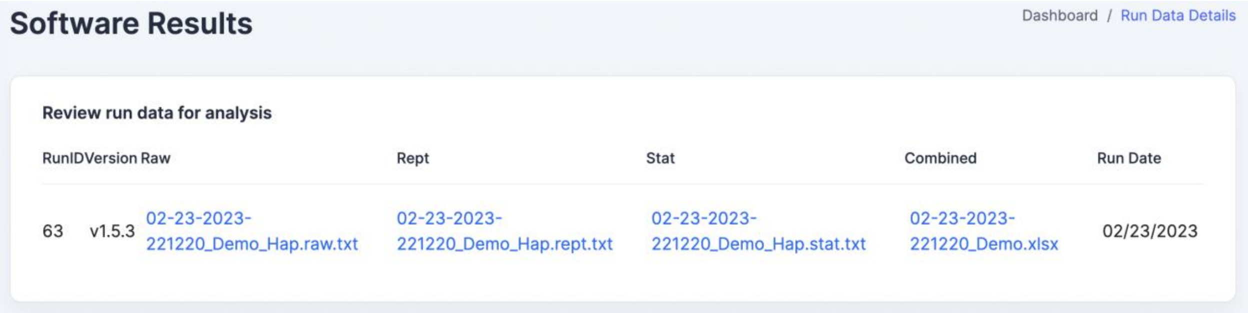


Dashboard

Previous analysis history

RunID	Version	BAM	BED	Min Coverage	Min Proportion	Max Donors	Results	Run Date
63	v1.5.3	Demo.bam	ForensicCODIS_v1.0.bed	2	0.1	1	View Results	2/23/2023 10:12:20 PM
61	v1.5.3	Demo.bam	ForensicCODIS_v1.0.bed	2	0.1	1	View Results	2/16/2023 1:22:08 PM
60	v1.5.3	Demo.bam	ForensicCODIS_v1.0.bed	2	0.1	1	View Results	2/16/2023 1:20:51 PM
59	v1.5.3	Demo.bam	ForensicCODIS_v1.0.bed	2	0.01	3	View Results	2/15/2023 2:43:17 PM
58	v1.5.3	Demo.bam	DiseaseSTR_v1.0.bed	2	0.1	1	View Results	2/15/2023 2:41:56 PM
57	v1.5.3	Demo.bam	ForensicXYSTR_v1.0.bed	2	0.1	1	View Results	2/15/2023 2:41:12 PM
56	v1.5.3	Demo.bam	ForensicCODIS_v1.0.bed	3	0.2	2	View Results	2/15/2023 2:39:13 PM
55	v1.5.3	Demo.bam	DiseaseSTR_v1.0.bed	2	0.1	1	View Results	2/15/2023 2:29:45 PM

Users may click “View Results” to see the analysis result files for each run.



Software Results

Review run data for analysis

RunID	Version	Raw	Rept	Stat	Combined	Run Date
63	v1.5.3	02-23-2023-221220_Demo_Hap.raw.txt	02-23-2023-221220_Demo_Hap.rept.txt	02-23-2023-221220_Demo_Hap.stat.txt	02-23-2023-221220_Demo.xlsx	02/23/2023

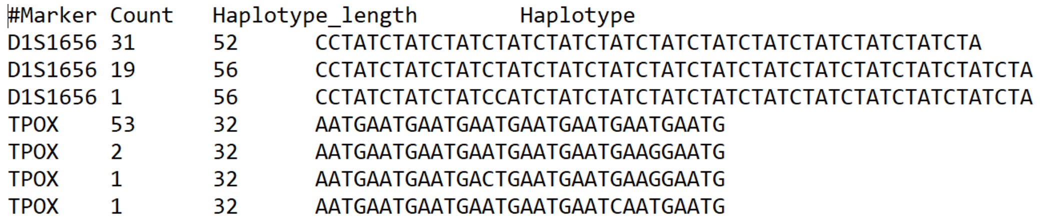
4. Details of the result files

A. Raw file

This file ends with “.raw.txt”, which is a tab-separated values file. In this file, the first row is the header row, which describes the columns.

There are four columns in this file:

1. Marker: marker or locus name
2. Count: the number of supporting reads of the allele
3. Haplotype_length: the length in base pair (bp) of the allele sequence (i.e., haplotype)
4. Haplotype: the allele sequence



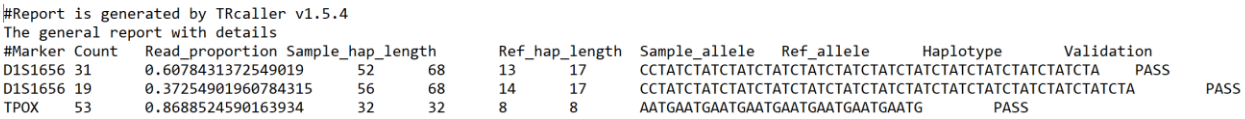
```
#Marker Count   Haplotype_length        Haplotype
D1S1656 31      52      CCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D1S1656 19      56      CCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
D1S1656 1       56      CCTATCTATCTATCCATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA
TPOX    53      32      AATGAATGAATGAATGAATGAATGAATGAATG
TPOX    2       32      AATGAATGAATGAATGAATGAATGAAGGAATG
TPOX    1       32      AATGAATGAATGACTGAATGAATGAAGGAATG
TPOX    1       32      AATGAATGAATGAATGAATGAATCAATGAATG
```

B. Report file

This file ends with “.rept.txt”, which is a tab-separated values file. In this file, the first row is the header row, which describes the columns. Each of the following rows represents one allele that passes the thresholds defined in the parameters.

There are nine columns in this file:

1. Marker: marker or locus name
2. Count: the number of supporting reads of the allele
3. Read_proportion:
4. Sample_hap_length: the length in base pair (bp) of the allele sequence (i.e., haplotype) in this sample
5. Ref_hap_length: the length in base pair (bp) of the allele sequence (i.e., haplotype) in this reference genome (e.g., the HG38)
6. Sample_allele: the length-based allele of the sample
7. Ref_allele: the length-based allele of the reference genome
8. Haplotype: the allele sequence
9. Validation: “PASS” means the allele passes the motif validation; otherwise, it should be “.”. The motif validation requires the haplotype to contain a subsequence with at least two tandem motif repeats. If there are multiple possible motifs at a locus, the validation only requires the haplotype to contain a subsequence generated from one of the motifs.



```
#Report is generated by TRcaller v1.5.4
The general report with details
#Marker Count   Read_proportion Sample_hap_length       Ref_hap_length  Sample_allele   Ref_allele      Haplotype       Validation
D1S1656 31      0.6078431372549019      52      68      13      17      CCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA    PASS
D1S1656 19      0.37254901960784315     56      68      14      17      CCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTA        PASS
TPOX    53      0.8688524590163934      32      32      8       8       AATGAATGAATGAATGAATGAATGAATGAATG        PASS
```

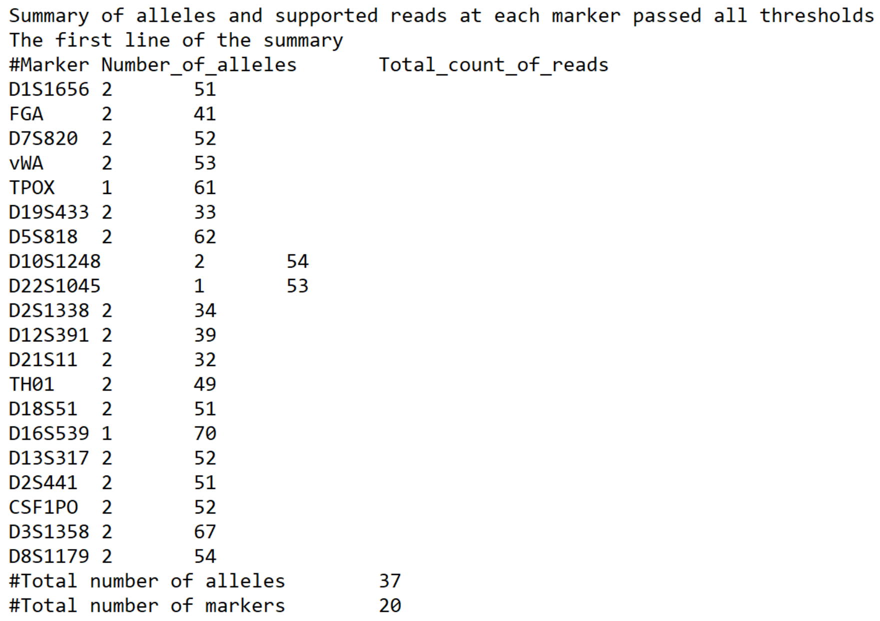
C. Statistics file

This file ends with “.stat.txt”, which is a tab-separated values file. In this file, the first row is the header row, which describes the columns.

There are nine columns in this file:

1. Marker: marker or locus name
2. Number_of_alleles: the number of alleles at this marker
3. Total_count_of_reads: the total number of reads mapped to this marker, including reads both passing the thresholds (i.e., the reads of the true alleles) and not passing the thresholds (i.e., noises). This number represents the read depth or coverage of this marker.

The last two rows show the total number of alleles and the total number of markers in this sample.



```
Summary of alleles and supported reads at each marker passed all thresholds
The first line of the summary
#Marker Number_of_alleles       Total_count_of_reads
D1S1656 2       51
FGA     2       41
D7S820  2       52
vWA     2       53
TPOX    1       61
D19S433 2       33
D5S818  2       62
D10S1248        2       54
D22S1045        1       53
D2S1338 2       34
D12S391 2       39
D21S11  2       32
TH01    2       49
D18S51  2       51
D16S539 1       70
D13S317 2       52
D2S441  2       51
CSF1PO  2       52
D3S1358 2       67
D8S1179 2       54
#Total number of alleles        37
#Total number of markers        20
```

D. Summary in an Excel file

This summary Excel file includes 4 tabs. The first three tabs contain the same contents as the above three text files. The fourth tab contains the running setting, such as the TRcaller version, run time, configure file, and the parameters used in the analysis.

5. Citation

Please cite this work:

Xuewen Wang, Meng Huang, Bruce Budowle, Jianye Ge. 2023. Precise and ultrafast tandem repeat variant detection in massively parallel sequencing reads. bioRxiv 2023.02.15.528687 https://www.biorxiv.org/content/10.1101/2023.02.15.528687v1