MEGASAT: infer microsatellite genotypes from short-read sequence data

User's Manual

December, 2015

Luyao Zhan[1], Ian G. Paterson[2], Bonnie A. Fraser[3], Beth Watson[2], Ian R. Bradbury[4], Praveen N. Ravindran[1], David Reznick[5], Robert G. Beiko[1], Paul Bentzen[2].

[1] Faculty of Computer Science, Dalhousie University. 6050 University Avenue, Halifax, Nova Scotia B3H 4R2, Canada
[2] Marine Gene Probe Laboratory, Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax, Nova Scotia B3H 4R2, Canada
[3] Department of Molecular Biology Max Planck Institute for Developmental Biology, Spemannstrasse 37-39, 72076 Tübingen Germany
[4] Salmonids Section, Science Branch, Department of Fisheries and Oceans Canada, 80 East White Hills Road, St. John's, Newfoundland A1C 5X1, Canada
[5] Department of Biology, University of California, Riverside, California 92521 USA

USER'S MANUAL

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 Overview

MEGASAT is a program that enables genotyping of microsatellite loci using next-generation sequencing data. MEGASAT reads FASTQ or FASTA files and automatically scores microsatellite genotypes using sequence depth with decision rules to account for amplification artifacts. Moreover, it can generate histograms of length frequency distributions for manual verification of genotypes.

MEGASAT was written in Perl & R and offers a simple graphical user interface (GUI) to users who use Windows or Mac OS X systems, which enables easy operation of MEGASAT for users who are unfamiliar with Perl.

The current version of MEGASAT is 1.0. The Perl scripts in MEGASAT should work with any relatively recent version of Perl and have been tested with versions 5.18.2 and 5.16.3.

The GUI in MEGASAT should work with any relatively recent version of Java and have been tested with versions 1.6.0_65 and 1.8.0_40.

The R script in MEGASAT should work with any relatively recent version of R and have been tested with version 3.2.0.

2.0 Installing MEGASAT

MEGASAT is free available at https://github.com/beiko-lab/MEGASAT. You can download the whole compressed zip file or just the folder that contains the scripts for your operating system. If you want to test MEGASAT, you can also download "test data.zip" that contains a test data set (30 fastq files) and "primers.txt" (a text file that contains information of PCR primers, flanking regions and repeated motifs for 43 guppy loci).

2.1 Windows

Decompress the file "MEGASAT_1.0 for Windows.zip" and there are two Perl executable files, one executable Jar files and one R script in the decompressed folder. "MEGASAT_GUI.jar" is the graphical user interface for running those two Perl executable files ("MEGASAT_Genotype.exe" and "MEGASAT_Update.exe") and the Rscript "Mplot.R". "Mplot.R" can generate bar plots to display sequence length frequency distribution for each individual and each locus.

In order to run "MEGASAT_GUI.jar", Java needs to be installed in your computer. Here is the link of downloading Java: https://java.com/en/download/.

There are two options to invoke the Perl scripts. One option is to use the GUI (you do not need to install Perl) and another option is to call the Perl scripts from command line (you need to have Perl installed in your computer). If Perl is already installed in your computer, you can just click the Start button and go to your Perl interpreter to run the Perl scripts. If you don't have Perl installed but still want to run the Perl scripts from command line, here are the two main distributions for Windows: ActivePerl (http://www.activestate.com/activePerl) and Strawberry Perl (http://strawberryPerl.com/). We have used the latter in development and testing of MEGASAT, and recommend its use. The script "MEGASAT_Genotype.pl" and "MEGASAT_Update.pl" uses no complicated library functions.

In order to run "Mplot.R", R interpreter needs to be installed in your computer. Since "Mplot.R" has command line arguments and RStudio cannot access command line arguments, one way to run "Mplot.R" is to call it from command line, the other way is to use "MEGASAT_GUI.jar" to invoke "Mplot.R". Here is the link of downloading R: http://cran.r-project.org/bin/windows/base/.

2.2 Macintosh

If you are using a Macintosh system, Perl should already be installed; type "Perl –v" at the command line to ensure this is the case. So Perl scripts can be easily invoked from the terminal

on Mac system. But if you don't want to run scripts in terminal, a simple GUI ("MEGASAT_GUI.jar") is also offered to invoke those two Perl scripts and R script.

In order to run "MEGASAT_GUI.jar", Java needs to be installed in your computer. Here is the link of downloading Java: https://java.com/en/download/.

To run "Mplot.R", R interpreter also needs to be installed in your computer. Here is the link of downloading R: http://cran.r-project.org/bin/macosx/.

2.3 Linux

Perl should already be installed on Linux system. So it's easy to go to terminal to invoke Perl scripts.

On Linux system, R interpreter also needs to be installed to run "Mplot.R". Here is the link of downloading R: http://cran.r-project.org.

3.0 Preparing input files

3.1 MEGASAT_Genotype.pl

"MEGASAT_Genotype.pl" requires an input file with information about PCR primers, and a set of .fastq or .fasta files representing reads from each sampled individual.

3.1.1 Primer file

The primer file must be in a tab-separated format, with the following headers:

- **Column1: locus name**
- **Column2: 5' microsatellite primer**
- **Column3: reverse-complement of 3' microsatellite primer**
- **Column4: 5' flank**
- **Column5: 3' flank**
- **Column6: the repeat_unit_sequence**
- **Column7: the ratios group (You don't need to write this column if you want to use all the default ratios of MEGASAT).**

Table 1 An example of the input primer file, which shows the file format to run in MEGASAT. The column 7 is optional.

Locus Name	5' microsatellite primer	reverse-complement of 3' microsatellite primer	5' flank	3' flank	repeat_unit sequence	Ratios group (optional)
Locus 1	AACCTGC	GGCCTAC	GGCC	CATGCT	AC	
Locus 2	CCTGACC	TTAACGT	ATAGCG	TGACC	TG	
Locus n	TCGACTT	ACCTGCT	TAGCCA	CCT	ACG	

In this primer text file, a header line is required to specify the column name. If one locus doesn't have 3' flank and 5' flank, a character "X" needs to be written in the 3' flank column and 5' flank column in that text file. Column 7 has six ratios that are separated by comma. The details of how we implement these six ratios to distinguish true alleles from artifacts are illustrated in supplementary materials (S2 Scoring Rules).

If you want to use all the default ratios which is (0.15,0.4,0.7,0.6,0.8,0.2) to predict genotypes, you don't need to write this column in the primer file. But if you want to change part of these six ratios, you can write your own ratios in the corresponding positions in column7. For other ratios you don't want to change, a space can be used in the corresponding position. For example, if the user just want to change the first ratio to 0.3, the column7 format will be (0.3, , , , ,). In the column7, you don't need to write brackets.

3.1.2 Input sequence file

Input sequence read files could be in standard FASTQ format or FASTA format.

3.2 MEGASAT_Update.pl

"MEGASAT_Update.pl" is a short script that provides the function to replace the wrong genotypes by your preferred genotypes in seconds. It requires an input scores file and an original genotypes file (called "Genotype.txt") that is generated by "MEGASAT_Genotype.pl".

3.2.1 Scores file

The input scored file contains some original genotypes automatically but wrongly predicted by the script "MEGASAT_Genotype.pl" and new genotypes you want to update to. This scores text file must be in comma-separated format, with the following headers:

- **Column1: locus name**
- **Column2 & Column3: original genotype**
- **Column4 & Column5: new genotype**

3.2.2 Original genotypes file

Another input file is the original genotypes file (called “Genotype.txt”) that is automatically generated by “MEGASAT_Genotype.pl”. This “Genotype.txt” file contains all the inferred genotypes for different individuals and loci.

3.3 Mplot.R

“Mplot.R” is a R script that generate plot files (histograms of sequence length-frequency distributions) for each locus and each individual. It requires an input folder which is the output folder (always named as “Ouput_” followed by your FASTQ or FASTA data folder name) generated by “MEGASAT_Genotype.pl”.

4.0 Running MEGASAT

4.1 Running "MEGASAT_Genotype.pl"

If you don't want to use command line to invoke scripts, a simple GUI is provided for Windows, Mac users. Double click the "MEGASAT_GUI" will display a pop-up page. Figure 1 below shows the simple GUI for calling the script "MEGASAT_Genotype.pl".

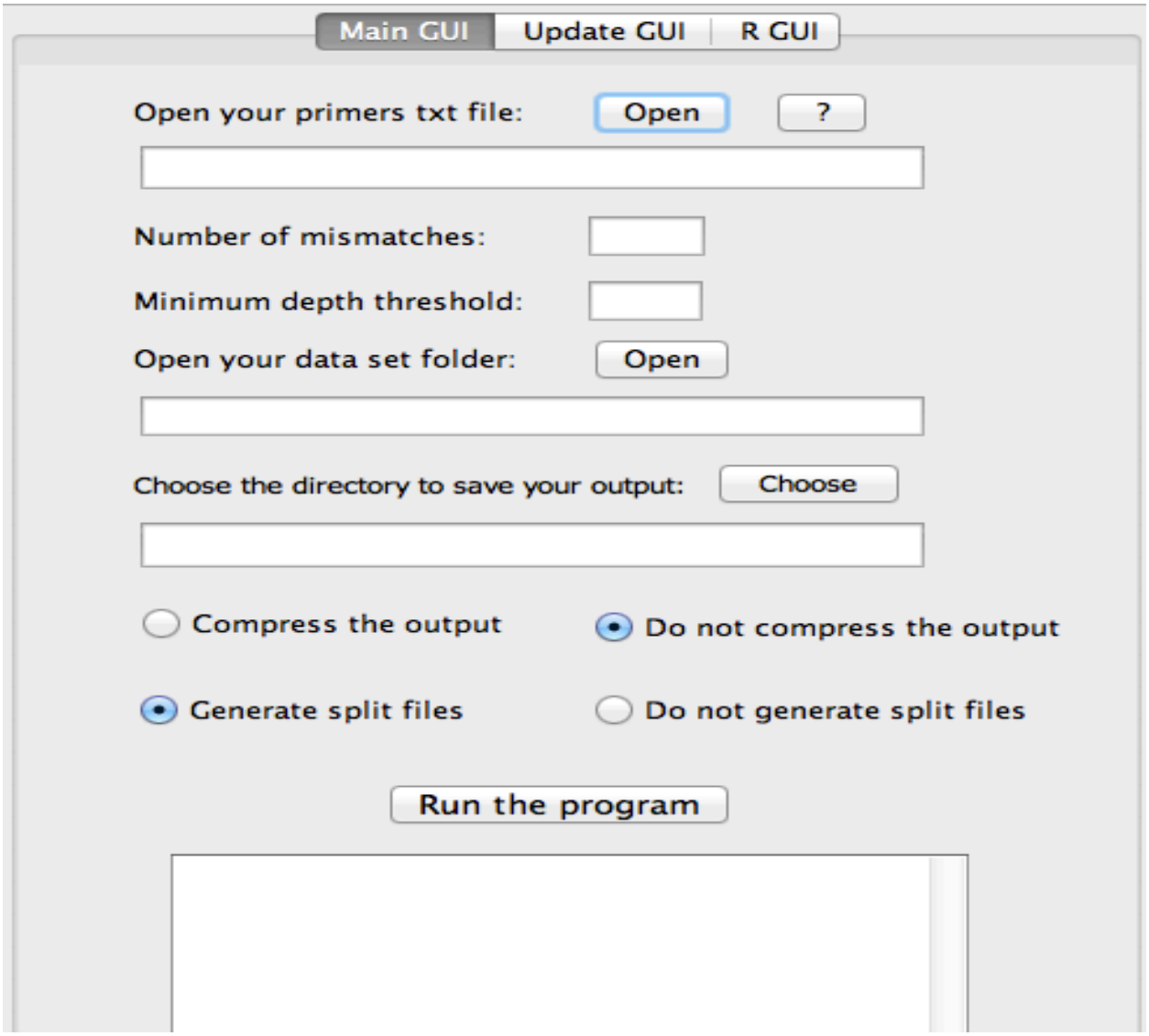

Figure 1 The GUI to call the script "MEGASAT_Genotype.pl"

In the tab page "main GUI", you can click the first “Open” button to open your input primers file. The “?” button is to display the required format of primer text file. The text field under the “Open” button will display the directory of your primers file. The second small text field is for typing the number of mismatches that gives the error tolerance to forward primers and reverse primers. For five prime flank and three prime flank, the number of mismatches is set based on their lengths. The next small text field is for typing the minimum depth threshold (we set it to 50 in our experiment). The second “Open” button is to open the data set folder that contains the input sequence read files (can either be FASTQ or FASTA files). The “Choose” button is to choose the directory to save your output folder. Two radio buttons in this page offer two options- compress the output folder or not compress the output folder. Another two radio buttions provide the option for the generation of split files (see 5.0 Output files). After all these parameters are filled, click the “Run the program” to run the Perl scripts “MEGASAT_Genotype.pl”.

If you want to run the scripts from command line, for Windows users, make sure you already have Perl installed in your system. We assume that “MEGASAT_Genotype.pl” and your primers txt file “primers.txt” are saved in the directory “C:\Users\Andy\Downloads”. And the data set folder “dataset” is also saved in the directory “C:\Users\Andy\Downloads”. In order to run the Perl script, first step is go back to the command prompt (terminal for Mac users). If the users prefer to generate all the split files, the following step is to type “perl C:\Users\Andy\Downloads\ MEGASAT_Genotype.pl C:\Users\Andy\Downloads\primers.txt 2 50 C:\Users\Andy\Downloads\dataset C:\Users\Andy\Desktop”. If you prefer to not generate split files, “1” needs to be assigned as the last command-line argument.

All the command-line arguments correspond to the parameters set in the GUI. The last command-line argument is optional (no values need to be assigned for generating all the split files; “1” needs to be assigned as the last command-line argument in order to not generate split files in the output). After this script is completed, an output folder called “Output_dataset” will be in the saving directory you type in the command line.

4.2 Running “MEGASAT_Update.pl”

Double click the “MEGASAT_GUI” will display a pop-up page. In the tab page "update GUI" (as shown in Figure 2), the first “Open” button is to open your scores txt file that contains the original genotypes and new genotypes you want to update to. The second “Open” button is to open the original genotyping text file (called “Genotype.txt”) generated by "MEGASAT_Genotype.pl". The “Choose” button is to choose the directory to save the output text file. After all these parameters are filled, click the “Run the program” to run “MEGASAT_Update.pl”.

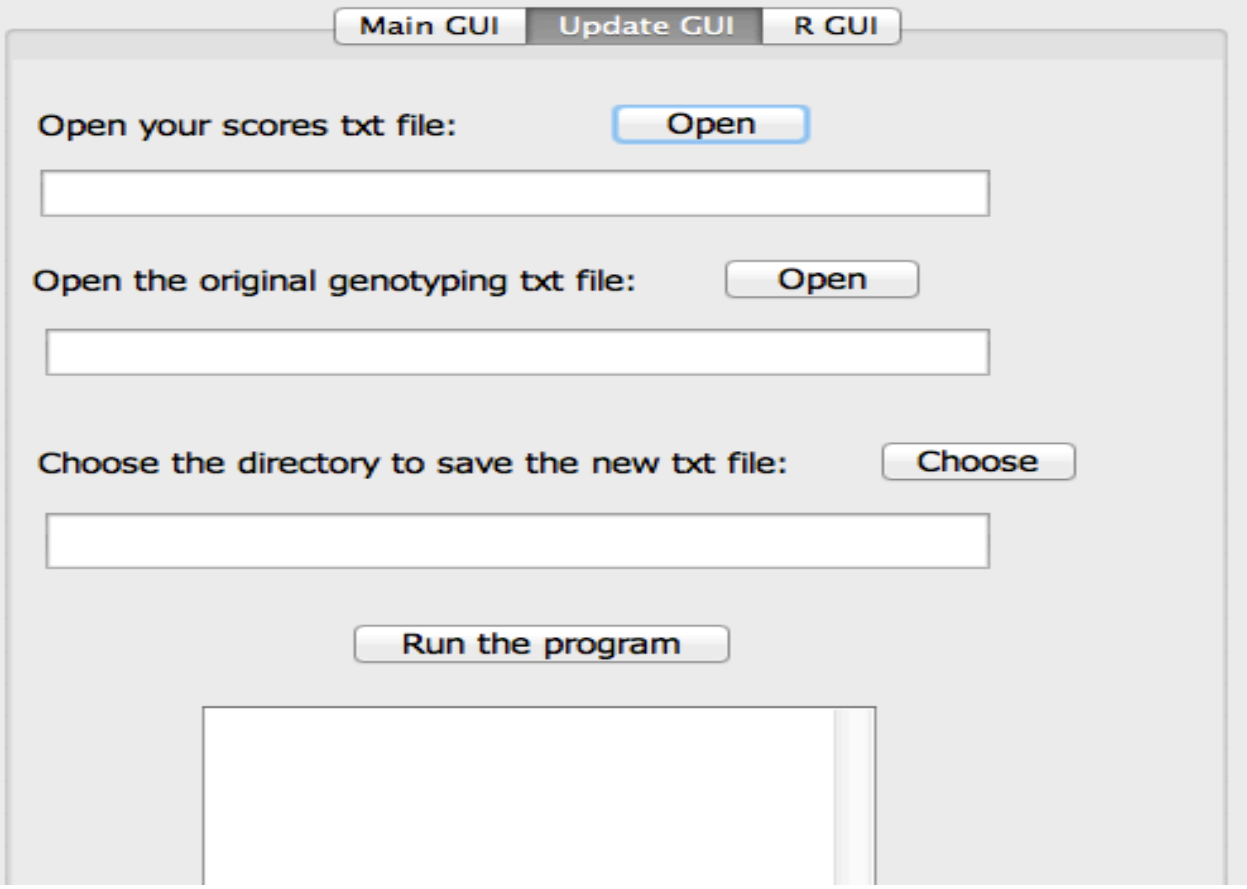

Figure 2 The GUI to invoke the script "MEGASAT_Update.pl"

If you want to run the scripts from command line, for Windows users, make sure you already have Perl installed in your system. We assume that "MEGASAT_Update.pl" and your scores text file "Scores.txt" are saved in the directory "C:\Users\Andy\Downloads". And the original genotyping text file is also saved in the directory "C:\Users\Andy\Downloads". In order to run the Perl script, first step is go back to the command prompt (terminal for Mac users) and type "perl C:\Users\Andy\Downloads\ MEGASAT_Update.pl C:\Users\Andy\Downloads\Scores.txt C:\Users\Andy\Downloads\Genotype.txt C:\Users\Andy\Desktop".

The first command-line argument is the directory of scores txt file. The second command-line argument is the directory of original genotyping text file "Genotype.txt" generated by "MEGASAT_Genotype.pl". The last command-line argument specifies the directory where you want to save your new txt file. After this script is completed, a new tab-separated txt file called "NewGenotype.txt" will be in the saving directory you typed in the command line.

4.3 Running “Mplot.R”

You can type "R" in command prompt (or terminal for MAC users) to check if R is correctly installed on your computer. If R is correctly installed, it will show the version information of R. For Windows users, don't forget to add the path of "R" to environment variables (here is the link of instructions that tells you how to set the path and environment variables in Windows http://www.computerhope.com/issues/ch000549.htm).

Double click "MEGASAT_GUI" will display a pop-up page. In the tab page "R GUI" (as shown in Figure 3), the "Open" button is to open the output folder automatically generated by "MEGASAT_Genotype.pl". The "Choose" button is to choose the directory to save the histogram output folder. After all these parameters are filled, click the “Run the program” to run “Mplot.R”.

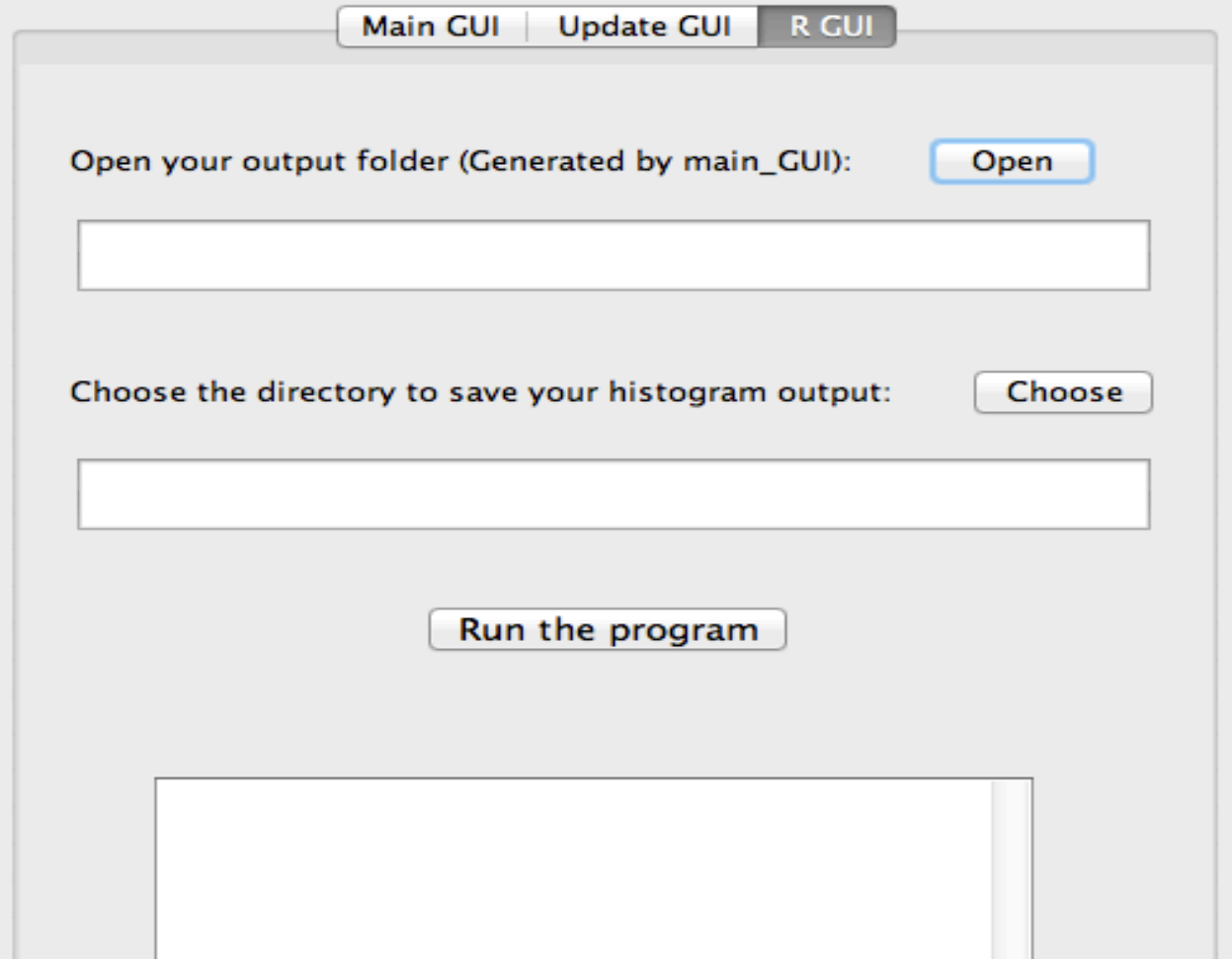

Figure 3 GUI to call the R script “Mplot.R”

If you want to run the R scripts from command line, then type the following command to invoke the R script: rscript /Users/Alex/Documents/Mplot.R /Users/Alex/Desktop/Output /Users/Alex/Documents. We assume that the "Mplot.R" is saved in "/Users/Alex/Documents". The first argument is the directory of input folder which is the output folder generated by "MEGASAT_Genotype.pl". The second argument is the directory to save your output plot folder. A folder called "Plots_Output" will be generated in the directory "/Users/Alex/Documents" when the program is completed.

5.0 Output files

5.1 MEGASAT_Genotype.pl

The output folder generated by "MEGASAT_Genotype.pl" has three types of files. In this folder, "Genotype.txt" is a tab-separated text file that gives all the genotypes for all the individuals and loci. In this Genotype.txt file, "X X" means that this locus doesn't occur in this individual. "0 0" means that the depth of alleles is too small to score. "Unscored Unscored" means that there are three possible real alleles, which makes the genotype difficult to be determined. The "Genotype.txt" file is organized as in the following example (Figure 4). The first row specifies the loci names and the first column represents the individuals' names. The genotype for each individual and each locus is split in two columns for easy review.

Sample_idx1_idx2	Locus1	Locus1-b	Locus2	Locus2-b	Locus3	Locus3-b
sample1	27	51	Unscored	Unscored	115	250
sample2	27	27	49	57	106	115
sample3	30	48	46	49	X	X
sample4	51	51	57	57	99	115
sample5	X	X	45	57	99	99

Figure 4 An example of the output file "Genotype.txt"

"Number_Discarded.txt" is a tab-separated txt file that counts the number of discarded sequences for all the individuals and loci. Those discarded sequences are sequences that have 5' microsatellite primers but have no flank, repeat_unit_sequence and reverse-complement of 3' microsatellite primers. In "Number_Discarded.txt" file, "X" means that there are no discarded sequences for this individual at this locus. You can use Microsoft Excel to open these tab-separated txt files, which makes these files easier to read.

Those text files (stored in subfolder "length_distribution") whose names start with "Genotype" and follow by the individual names show the sequence length frequency distribution for each microsatellite locus. In each of those text files, the first row illustrates the different sequence length for all the loci in one individual. Each row under the first row shows the count of length variants for each locus. The last column is the genotypes information for all loci in one individual. The text files whose names start with "Ratios_Threshold" and follow by the data set name lists the minimum depth threshold set for each run and depth ratios group used for each microsatellite locus.

Sequences which contain the 5' microsatellite primers (defined in the primer.txt file above) are retained in *.split files. These files are necessary for reviewing the performance of MEGASAT. The three *.split file types are "Sorted", "Trimmed" and "Discarded" that are stored in three subfolders. Each file name begins with the file type followed by the sample name and locus name.

1. Sorted.split: The input sequences are sorted into locus specific files by identifying and removing the 5' locus-specific oligonucleotide. These sequences retain all bases downstream of the 5' oligo. The number of sorted.split files = (# of samples) x (# of loci).

2. Trimmed.split: All sequences that pass MEGASAT's trimming criteria (see Zahn *et al*.2016 Fig 2, copied below) are written to Trimmed.split files. These sequences have had both oligos removed and are the sequences used to determine genotypes. The number of Trimmed.split files = (# of samples) x (# of loci)

3. Discarded.split: These are all the sequences rejected by MEGASAT, for any of the reasons outlined in Fig 5.

Note. All sequences in the Sorted.split files are either 'good' sequences used for scoring (i.e. are copied to Trimmed.split) or 'bad' (copied to Discard.split). For example: Sorted_Sample1-BF47.split is a file containing of all sequences for Sample1 at locus BF47. Trimmed_Sample1-BF47.split is all the good sequences from Sorted_Sample1-BF47.split, with the 3' oligos trimmed off, while Discarded_Sample1-BF47.split contains the discarded sequences.

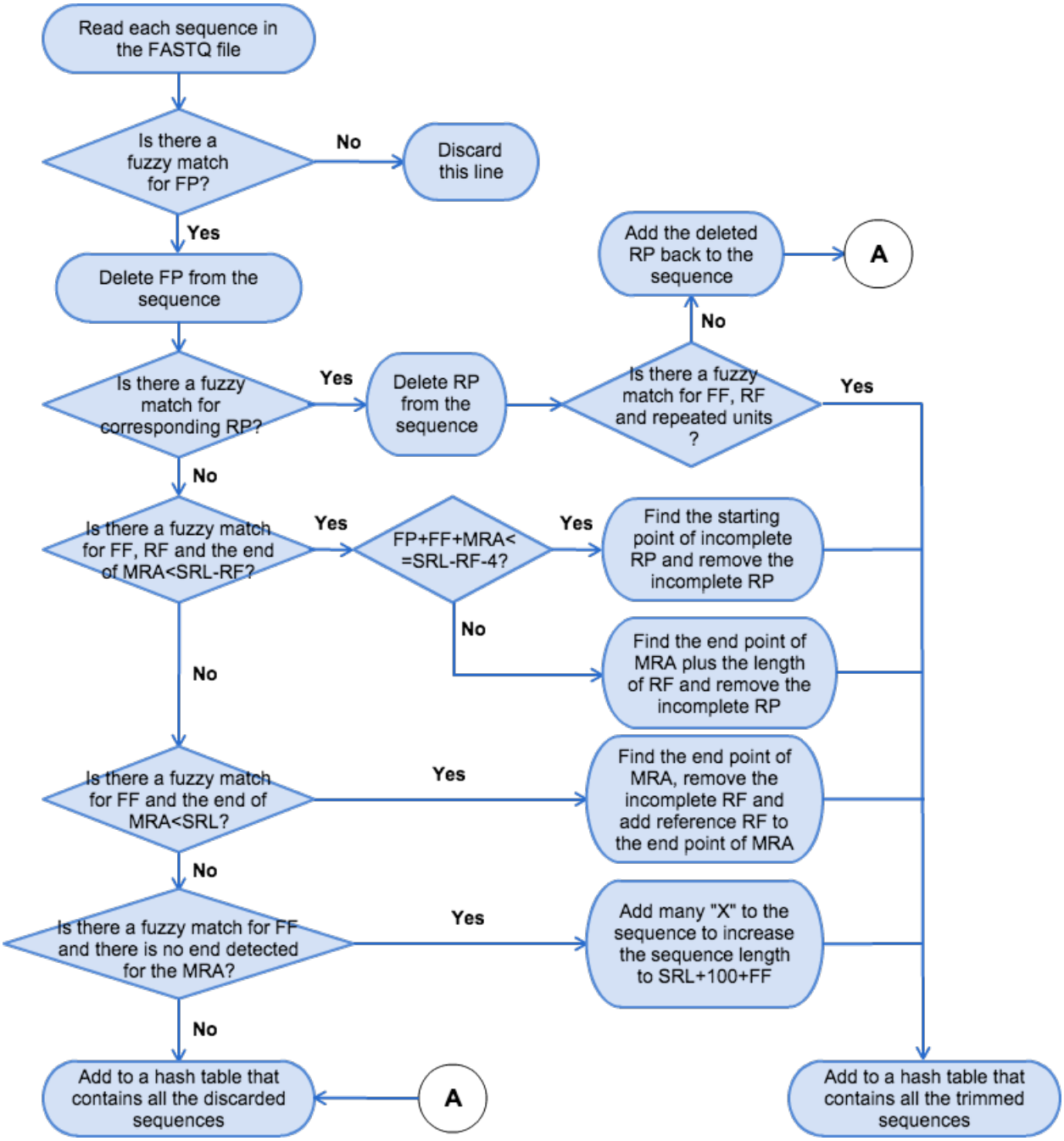

Figure 5 Flowchart of the algorithm that MEGASAT uses to trim off microsatellite primers (copied from Zhan *et al.* 2016).

5.2 MEGASAT_Update.pl

For another Perl script "MEGASAT_Update.pl" that helps to update the "Genotype.txt" very fast, its output is a tab-separated text file called "NewGenotype.txt". This "NewGenotype.txt" has all the updated genotypes.

5.3 Mplot.R

The output folder of "Mplot.R" contains pdfs for each locus. Each pdf shows the sequence-length frequency distributions of each individual for each locus. The plots are a graphical representation of the allele calls MEGASAT has made, and are an important tool for quickly reviewing the veracity of the genotypes. The plots are colour coded for easy review. Histograms bars are either blue, pink or grey based on the depth per locus. Grey indicates a sample below the minimum depth threshold and are scored with a "0 0" genotype in the GENOTYPE.TXT file. Pink histograms serve as a visual clue that the depth is just marginally above the minimum threshold (<threshold +10) and blue indicates a high depth (>threshold +10). Allele calls are plotted in deep blue, allowing the reviewer to scan quickly over the plot files to see if MEGASAT has called the alleles correctly. Example of pdf histograms is illustrated in the following Figure 6 (copied from Zhan *et al.* 2016).

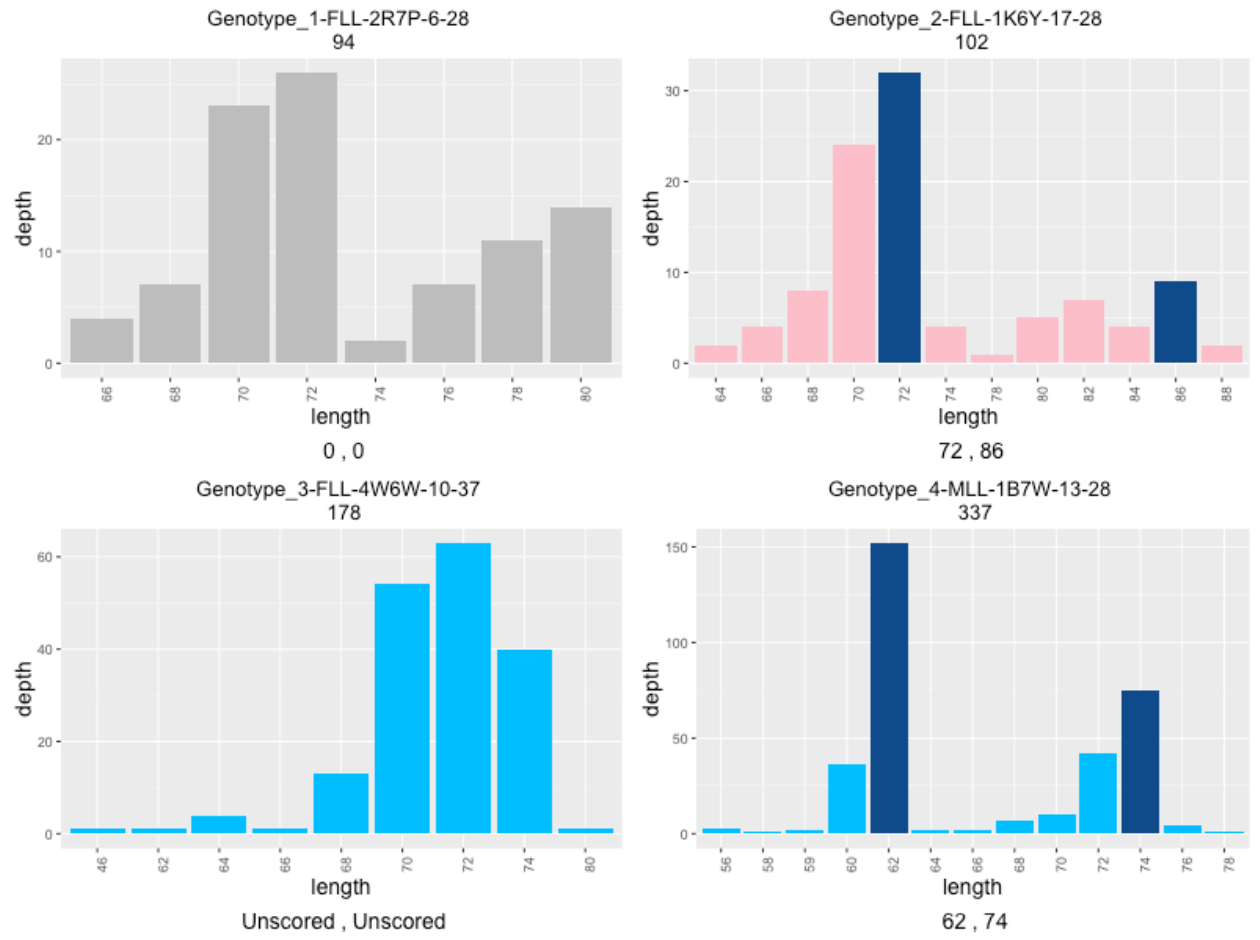

Figure 6 Examples of MEGASAT portable document format (pdf) histograms that show the frequencies of sequence length variants per microsatellite locus per individual. Sample IDs title each plot, followed by the total depth. Genotypes are listed under the x-axis. Colour codes include a) grey for samples below the minimum depth threshold, no alleles called. b) pink warning that the depth is close to the minimum, deep blue indicates allele calls (72/86). c) blue for acceptably high depths, no alleles called. d) Deep blue indicates allele calls (62/74).