

GUAVA Manual

Mayur Divate and Edwin Cheung

Contact: mdivate@umac.mo

Index

1. How to download and launch GUAVA	3
2. GUAVA graphical user interface	4
2.1 ATAC-seq data analysis parameters	4
2.2 ATAC-seq differential analysis parameters	5
3. Output interface for GUAVA ATAC-seq data analysis	6
4. Output interface for GUAVA ATAC-seq differential analysis	8
5. GUAVA command-line interface	9
4.1 Usage and option summary	9
6. How to install dependencies	10
6.1 Java 1.8 or latest	10
6.2 Bowtie version 1.1.2	11
6.3 Python version 2.7	11
6.4 MACS2 version 2.1.1.20160309	11
6.5 SAMtools Version: 1.3.1	11
6.6 R Version: >= 3.3.0	12
7. How to download genome fasta file	12
8. How to create bowtie index of genome fasta file	12

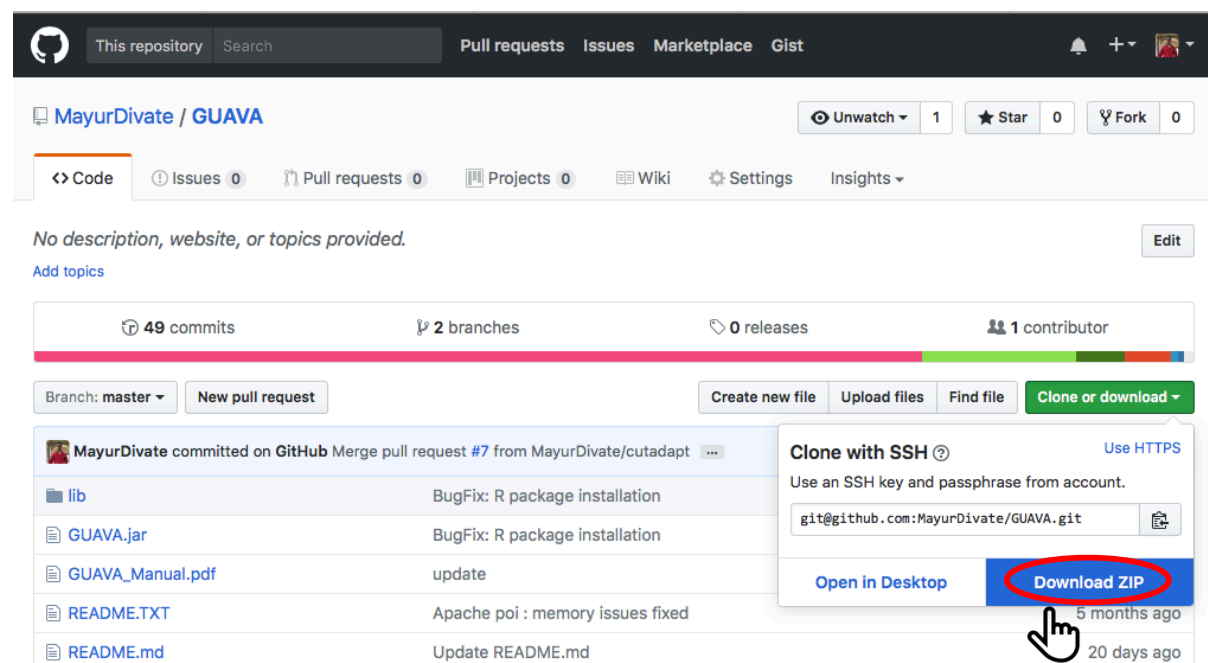
GUAVA: a GUI tool for the Analysis and Visualization of ATAC-seq data

GUAVA is a standalone GUI application for analyzing ATAC-seq data. GUAVA works on Linux and Mac OS. GUAVA was developed to help researchers with minimal or no Linux background to analyze ATAC-seq data. This document contains all the information that is required to install dependencies and use the GUAVA graphical and command line interfaces. This document also explains the GUAVA graphical user interface using a published ATAC-seq data. Finally, we have also provided the procedure on how to create bowtie index from genome fasta for novice bioinformaticians.

1. How to download and launch GUAVA

To download GUAVA, go to GitHub via this link:

<https://github.com/MayurDivate/GUAVA>



There is no need to install GUAVA. It can be easily launched as described below.

Move the GUAVA package to the home folder. Unzip the package by launching Terminal and use following commands:

```
$ cp </path/to/ GUAVA-master.zip> ~  
$ cd ~  
$ unzip GUAVA-master.zip
```

Launch GUAVA using the following commands:

```
$ cd ~/GUAVA-master  
$ java -jar GUAVA.jar
```

2. GUAVA graphical user interface

We demonstrate how to use the GUAVA graphical user interface and show typical results that are obtained from the program by using the GSE84515 ATAC-seq dataset.

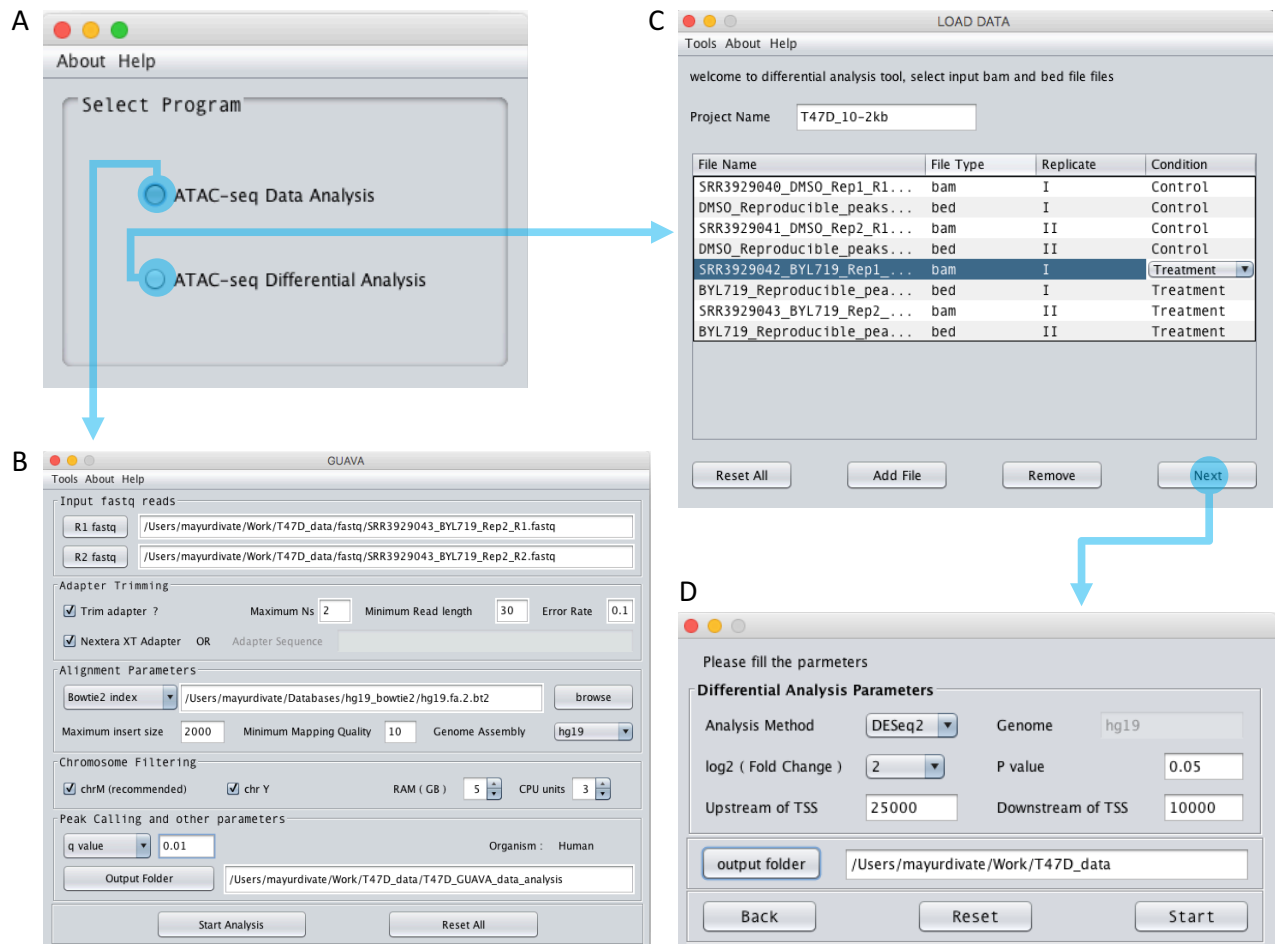


Figure 1. Design of GUAVA Graphical user interface

GUAVA opens the input window for the selected programs (ATAC-seq data analysis or ATAC-seq differential analysis) from the main window. For ATAC-seq data analysis, users load sequencing data using R1 fastq and R2 fastq buttons. Users then choose between bowtie and bowtie 2 from the dropdown menu in the 'Alignment Parameters' section. To start differential analysis users set the project name, load the bam and bed files. Lastly users set the differential analysis parameters.

2.1 ATAC-seq data analysis parameters:

Maximum Ns: Maximum number of Ns to discard pair if one reads of the two reads has too many Ns.

Minimum read length: Reads with length less than 'minimum read length' will be discarded.

Error Rate: Percentage to calculate number of mismatches allowed to match adapter sequences. For example, if error rate is 0.1 then 1 mismatch is allowed for every 10bp match of adapter sequence.

Adapter sequence: Option to specify custom adapter sequence when Nextera XT adapter is not used for library preparation.

Bowtie V1 or Bowtie V2 index: Select Bowtie V1 index to use bowtie for alignment or select bowtie v2 index to bowtie2 for alignment using drop down menu. Then upload bowtie/bowtie2 genome index using browse button.

Maximum insert size: Maximum insert size to call valid paired end alignment.

Maximum genomic hits or Mapping quality: Maximum genomic hit (bowtie) and Minimum Mapping quality (bowtie2) to discard reads pairs which has multiple alignments.

Genome assembly: select the correct genome build (hg19, mm10, mm9) and same build will be used for peak annotation and functional analysis.

ChrM: If selected, reads aligning to mitochondrial chromosome will be discarded.

ChrY: If selected, reads aligning to chromosome Y will be discarded.

RAM: RAM in GB to be used by GUAVA.

CPU units: Number of CPU units used by GUAVA.

p/q value: Select p/q value from drop down menu and specify the cut off value in box next to it. Peaks with less significant p/q value than the specified value will be filtered.

2.2 ATAC-seq differential analysis parameters:

Log2(Fold Change): log2 fold change cut off

P value: P value cut off to select most significant differentially enriched peaks

Upstream and downstream of TSS: Upstream and downstream distance from TSS to peak to associate peak with gene for functional annotation.

3. Output interface for GUAVA ATAC-seq data analysis

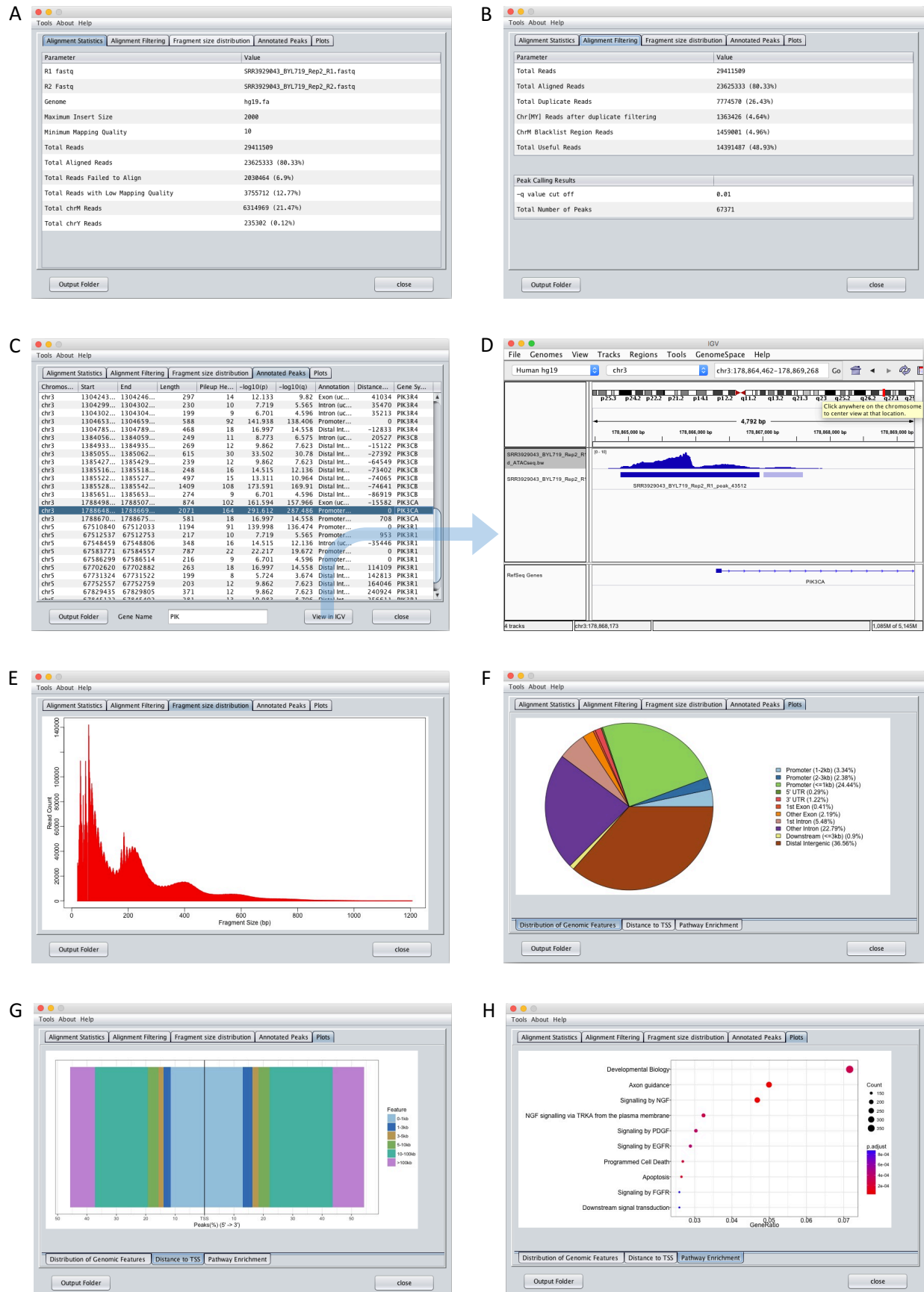


Figure 3: Output interface for GUAVA ATAC-seq data analysis. A) Input summary and alignment statistics. B) Read filtering and peak calling summary. C) Peak annotation table with sorting and filtering functionality. Easy access to IGV for visualizing peaks and automatically generated normalized ATAC-seq signal by GUAVA. D) Visualization of ATAC-seq peaks with IGV. E) Graph showing the fragment size distribution. F) Pie chart showing the percentage of peaks in various genomic locations such as promoter, intron, exon, UTR, etc. G) Plot showing the percentage of the peaks upstream and downstream of the TSS of the nearest genes. Different colors indicate different ranges of distances from the TSS. H) Enriched pathways obtained using ReactomePA bioconductor package.

4. Output interface for GUAVA ATAC-seq differential analysis

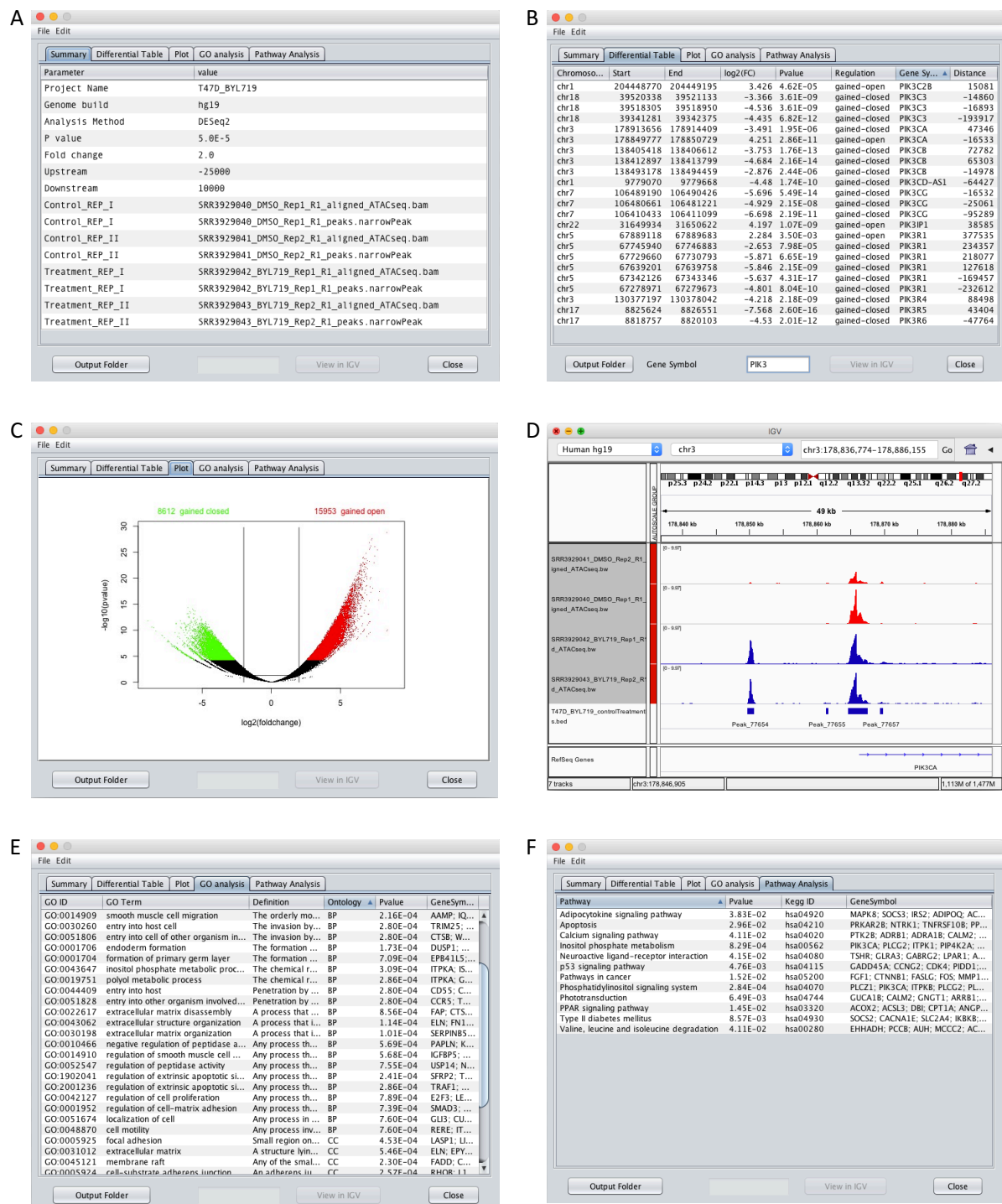


Figure 4: Output interface for GUAVA ATAC-seq differential analysis. A) Input summary. B) Differentially enriched peaks with sorting and filtering functionality. Easy access to IGV to visualize differentially enriched peaks and normalized ATAC-seq signals from each sample. C) Volcano plot indicating the differentially enriched peaks. Red: peaks with increased chromatin accessibility, green: peaks with reduced chromatin accessibility and black: peaks with no significant change in

chromatin accessibility. D) Peak visualization in IGV. E) Enriched gene ontologies and F) enriched pathways.

5. GUAVA command-line interface

In addition to a graphical user interface, GUAVA can also be used via a command-line interface. The command-line user interface makes GUAVA flexible by allowing it to be easily integrated into existing pipelines. Also, it provides flexibility for running GUAVA through a resource manager or a job scheduler system such as SLURM.

Type the following command to print the help message for GUAVA:

```
$ java -jar GUAVA.jar -h
```

5.1 Usage and option summary Usage: \$ java -jar GUAVA.jar [options]*

Options	Description
R1	Path to the FASTQ file containing upstream mates
R2	Path to the FASTQ file containing upstream mates
g	Path to bowtie index of genome fasta file
a	Genome assembly version [hg18,hg18,hg38,mm9,mm10]
value	p q value for MACS2 peak filtering default: q
c cutoff	Cutoff for p/q value e.g. 0.05, 5E-2 default: 0.05
X	Maximum distance from each other at which read mates can map to the genome default: 2000
m	Report alignment for pair, if maximum number of reportable alignments for pair is less or equal to m default: 1
O outdir	Path to the output directory default: current directory
ram	RAM memory to use in GBs default: 1
cpu	Number of threads to use default: 1
chrM	Remove(T) or keep(F) reads mapping to mitochondrial chromosome default: T
chrY	Remove(T) or keep(F) reads mapping to chromosome Y default: F
H help	Print help message

Table 1. Usage and options for the GUAVA command-line interface. Compulsory options are shown in blue color.

6. How to install dependencies

GUAVA depends on others tools/dependencies in order to process ATAC-seq data (e.g. bowtie for alignment). These dependencies need to be installed on the machine that will run GUAVA. If any of the dependencies are not found, GUAVA will fail to start. Dependencies are installed from the Terminal (the command line program provided by the OS). After launching Terminal, users can simply input commands by typing or copy and pasting into the program to complete the installation. (Note: text that is followed by "\$" is a command.)

6.1 Java 1.8 or latest

As GUAVA is developed in Java, this needs to be installed.

To install Java on a Mac OS:

- Download Java by going to <https://java.com/en/download/>
- Double-click the pkg file to launch it
- Double-click on the package icon to launch Install Wizard
- The Install Wizard will display the Welcome to Java installation screen. Click Next
- Click the Next button to continue the installation
- Click Close to finish the installation process

For more details, please follow this link:

https://www.java.com/en/download/help/mac_install.xml

To install Java on a Linux OS, simply copy and paste following command to the Terminal:

```
$ sudo apt-get install oracle-java8-installer
```

Alternatively, follow this link:

https://java.com/en/download/help/linux_x64_install.xml

6.2 Bowtie version 1.1.2

To install bowtie:

- Download bowtie from here:
<https://sourceforge.net/projects/bowtie-bio/files/bowtie/1.1.2/>
Linux OS : bowtie-1.1.2-linux-x86_64.zip
Mac OS : bowtie-1.1.2-macos-x86_64.zip
- Copy downloaded bowtie file or the file path (*i.e.* the file location) and paste it in the Terminal,
Mac => command + v
Linux => ctrl + shift + v
- Launch Terminal and use the following commands in Terminal to install bowtie:

```

$ cp <bowtie file path> ~/
$ cd ~/
$ unzip bowtie-1.1.2*.zip
$ cd bowtie-1.1.2/
For Mac OS use:
$ echo "export PATH=\$PATH:`pwd` | cat - >>
  ~/.bash_profile
$ source ~/.bash_profile
For Linux OS use:
$ echo "export PATH=\$PATH:`pwd` | cat - >> ~/.bashrc
$ source ~/.bashrc

```

6.3 Python version 2.7

Python is required for MACS2 installation.

To install Python on a Mac OS:

- Download the Mac OS X 64-bit/32-bit installer (not the PPC installer) from the Python website, <https://www.python.org/downloads/release/python-2711/>.
- Double-click the python-2.7.11-macosx10.6.pkg file in the Downloads folder.
- If you have Gatekeeper enabled, the installation will be blocked. Open System Preferences > Security & Privacy and click Open Anyway.
- Click Continue, Agree and Install buttons in the Install Python window.

To install Python on a Linux OS, use following command:

```
$ sudo apt-get install python
```

6.4 MACS2 version 2.1.1.20160309

To install MACS2 on either a Mac or Linux OS use the command below:

```
$ pip install --user MACS2
```

Or follow this link to install MACS2

<https://pypi.python.org/pypi/MACS2>

6.5 SAMtools Version: 1.3.1

To install SAMtools:

- Download samtools-1.3.1.tar.bz2 via this link: <https://sourceforge.net/projects/samtools/files/samtools/1.3.1/>
- Copy downloaded SAMtools file or copy the samtools-1.3.1.tar.bz2 file path
Paste path on Terminal:
Mac => command + v
Linux => ctrl + shift + v
- Open Terminal
- Copy the following commands to Terminal and hit enter

```
$ cp <samtools file path> ~/
$ cd ~/
$ tar jxvf samtools-1.3.1.tar.bz2
$ cd samtools-1.3.1
$ make
```

- For Mac:


```
$ echo "export PATH=\$PATH:"`pwd` | cat - >>
  ~/.bash_profile
$ ~/.bash_profile
```
- For Linux:


```
$ echo "export PATH=\$PATH:"`pwd` | cat - >> ~/.bashrc
$ source ~/.bashrc
```

6.6 R Version: >= 3.3.0

Mac users can click on the link below and follow the video tutorial on how to install R. If R installation is found on system, GUAVA will install bioconductor packages automatically.

- https://youtu.be/cX532N_XLIs?list=PLqzoL9-eJTNBDdKgJgJzaQcY60XmsXAHU

Linux user can use following link for installing R:

- <https://cran.r-project.org>

7. How to download genome fasta file

To download genome fasta files, follow the following links.

Human: <http://hgdownload.soe.ucsc.edu/downloads.html#human>

Mouse: <http://hgdownload.soe.ucsc.edu/downloads.html#mouse>

Select the genome assembly that you want to download, then click on “Full data set” and download *.fa.gz or chromFa.tar.gz (one chromosome per file) file.

To decompress the file, use the following command:

```
$ gzip -d <file name>
```

If it is one chromosome per file then make one single fasta file containing all chromosomes using the cat command:

```
$ cat file1.fa file2.fa > genomeName.fasta
```

8. How to create a bowtie index of genome fasta file

If you already have a genome fasta file, follow the commands below to create a bowtie genome index. Bowtie uses this index to speed up the alignment process.

```
$ cd <path to genome fasta file>
```

```
$ bowtie-build <genome.fasta> <genome>
```

Note: This is a time consuming step