GUAVA Manual

Mayur Divate and Edwin Cheung

Contact: mdivate@umac.mo

Index

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 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.)

1.1 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

1.2 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
```

1.3 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
```

1.4 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

1.5 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
```

1.6 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-eJTNBDdKgJgJzaQcY6OXmsXAHU

Linux user can use following link for installing R:

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

2. 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
```

3. 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 following ATAC-seq dataset, SRR891275(Buenrostro, et al., 2013).

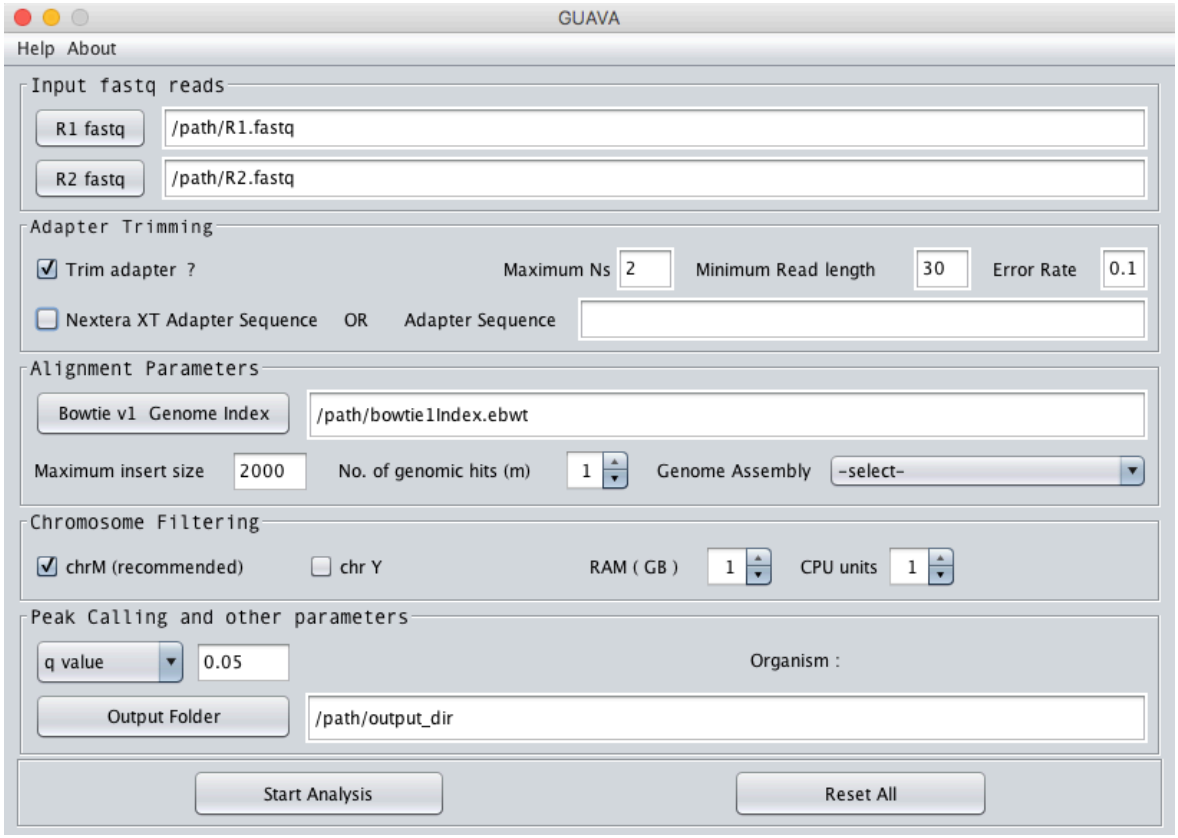

Figure 1. Graphical user interface of GUAVA

Users can input FASTQ files and the bowtie index of genome fasta file by using the buttons for R1 fastq, R2 fastq and Bowtie v1 Genome Index, respectively. If user select the “Trim adapter?” check box, then adapters will be trimmed from sequencing reads using cutadapt (Martin, 2011) before aligning them to genome. Users can select either Nextera XT adapter or provide custom adapter sequences. Reads with more than “Maximum Ns” and less than “Minimum Read length” after adapter trimming will be filtered. The maximum distance for mapping of read mates can be set using “Maximum insert size” input field. To exclude read pairs which are exceeding the desired number of alignments (m), limits can be set using “No. of genomic hits(m)” input field. Users can specify the assembly or version name by using genome assembly drop down menu. If the checkboxes of chr M and/or Y are selected, reads mapping to those chromosomes will be discarded. A dropdown menu is provided for selecting the p- or q-value and the input field next to it can be used for setting the cutoff for filtering MACS2 (Feng, et al., 2012) peaks. The RAM memory in gigabytes and threads can be set using “RAM (GB)” and “CPU units”, respectively. Clicking on the Start Analysis button will initiate the processing of ATAC-seq data, however, no processing will be performed if any invalid values (highlighted with red) are entered.

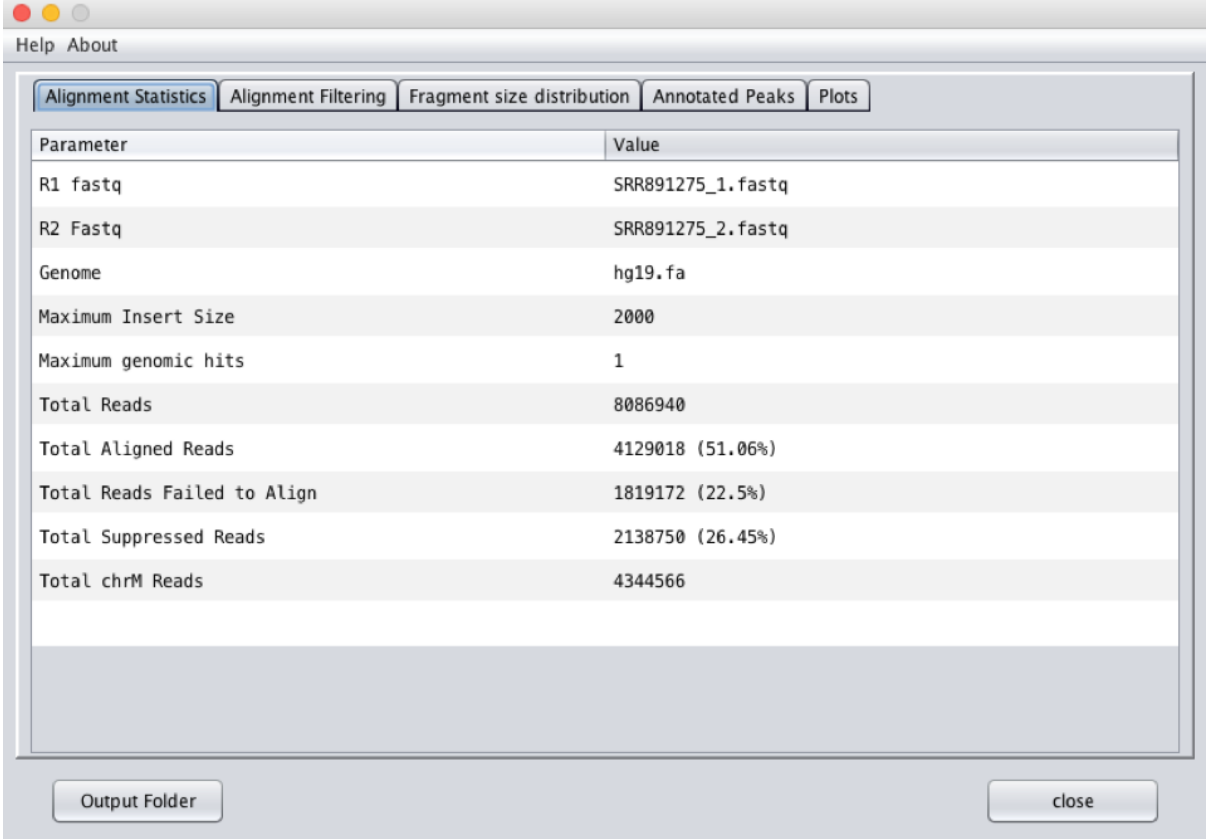

Figure 2. Summary of input information and alignment statistics

This tab shows the input summary: the names of the FASTQ files and the genome assembly used for the analysis, the maximum insert size and genomic hits. The following alignment statistics are found here: the total number of reads sequenced, the number and percentage of aligned, un-aligned reads, and reads failed to align due to the m limit. The reads mapping to chr M and chr Y are also shown if selected.

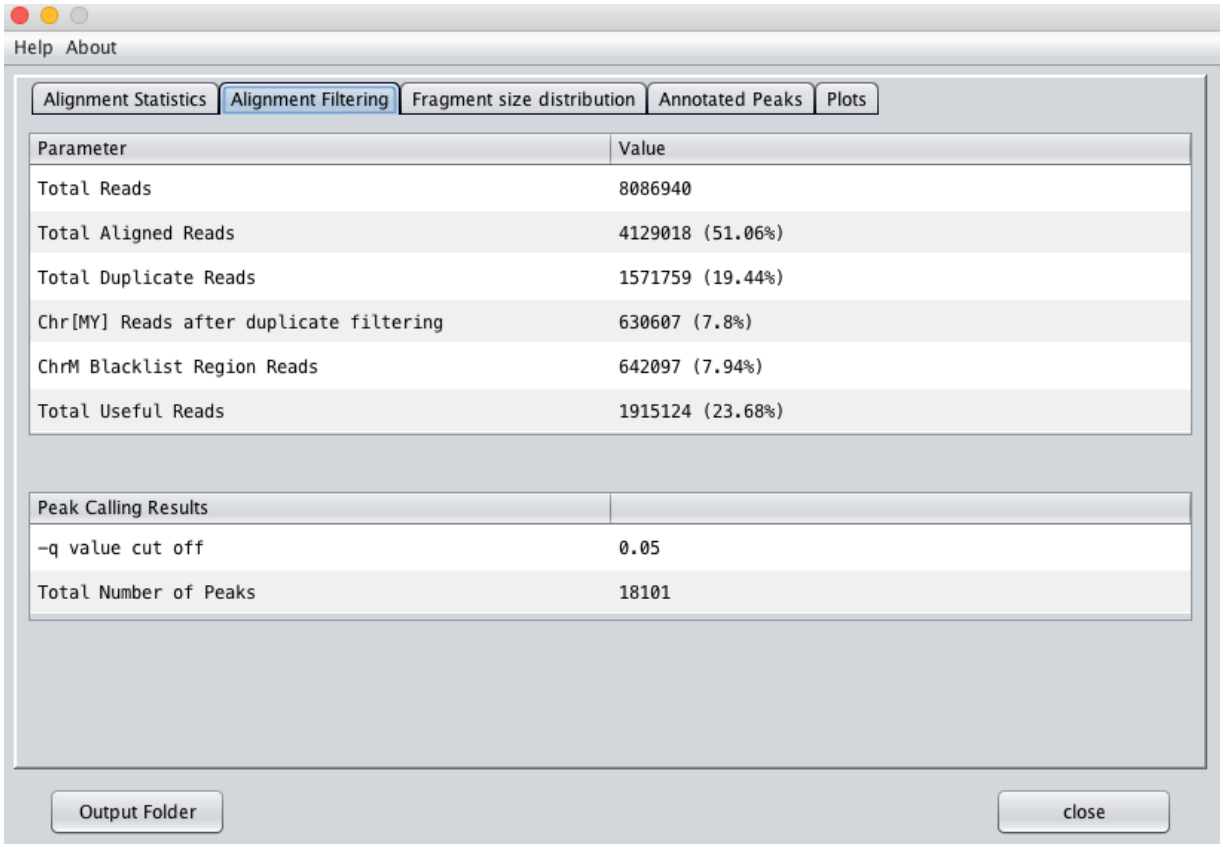

Figure 3. Summary of alignment filtering and peak calling information

The first table provides the number and percentage of reads relative to the total input reads for i) aligned reads, ii) duplicate reads, iii) reads aligned to chromosome M and Y (based on user selection), iv) reads mapped in the blacklisted regions, and v) useful reads that qualifies for further processing. The second table shows the total number of peaks called and the p- or q-value cutoff as per input.

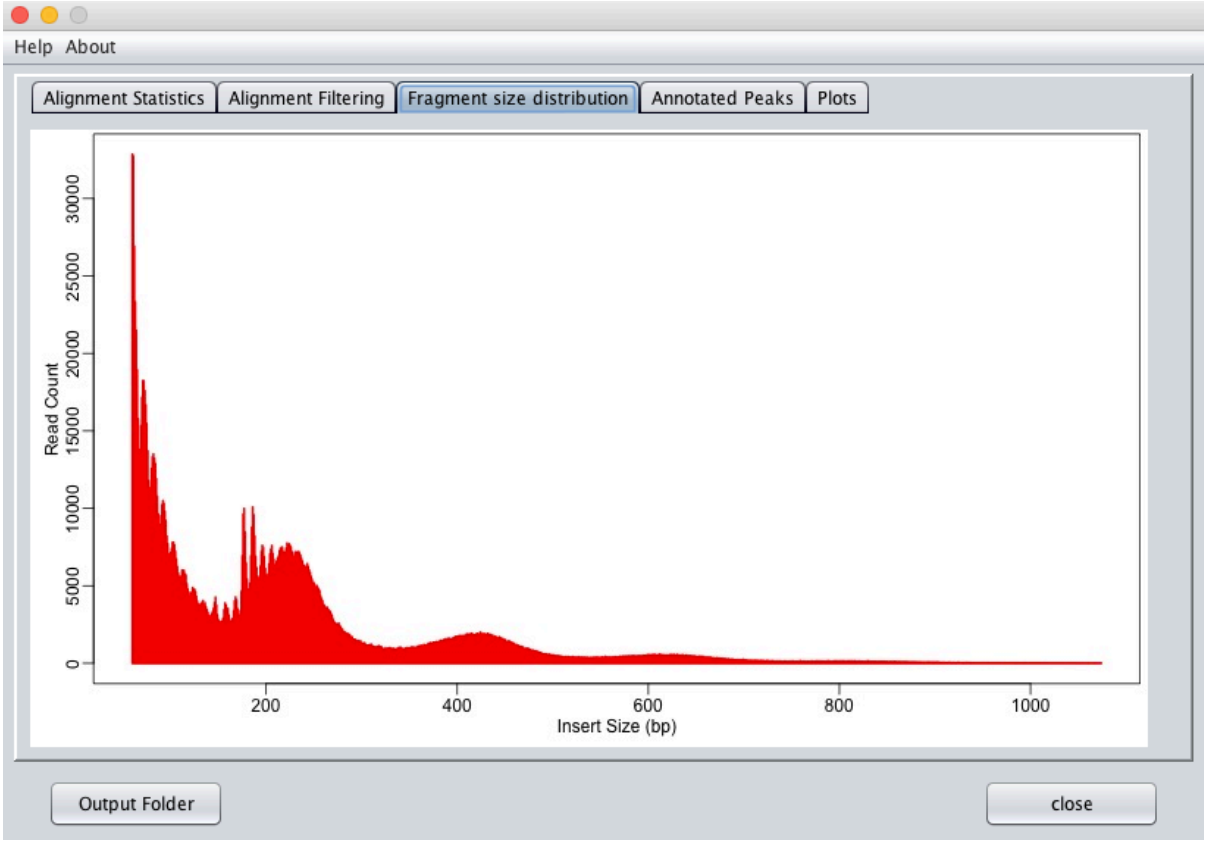

Figure 4. Fragment size distribution graph

This graph shows the fragment size distribution calculated using the Picard (Broad Institute) tool. The x-axis shows the size of fragment in bp and while the y-axis shows the read count.

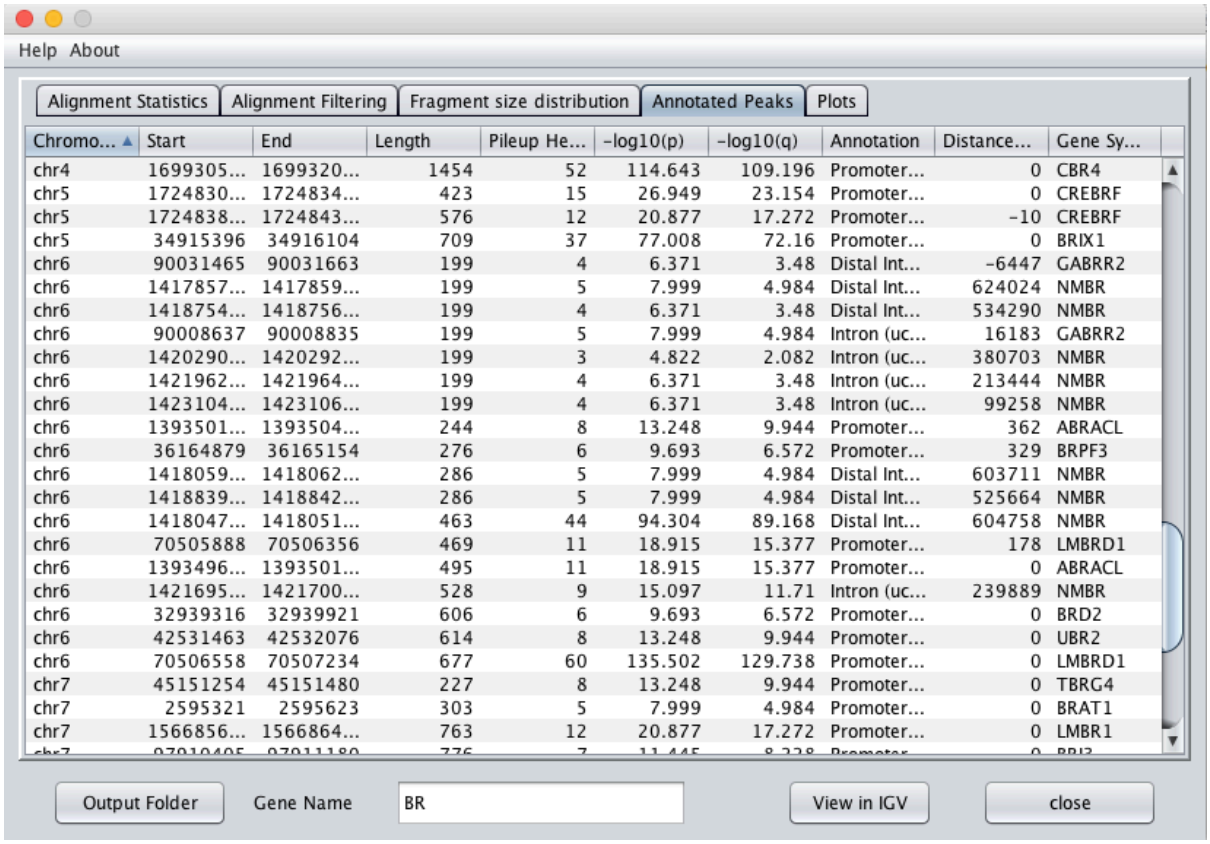

Figure 5. Table of Annotated peaks

Information regarding annotated peaks can be browsed using the Annotated Peaks table. The Gene Name search box can be used to help filter peaks according to the gene symbol. The table can be sorted by clicking on the header for each column. Clicking the “View in IGV” button will load the data tracks in the IGV (Robinson, et al., 2011) genome browser and set the IGV location same as the peak coordinates.

Figure 6. Peak visualization using IGV

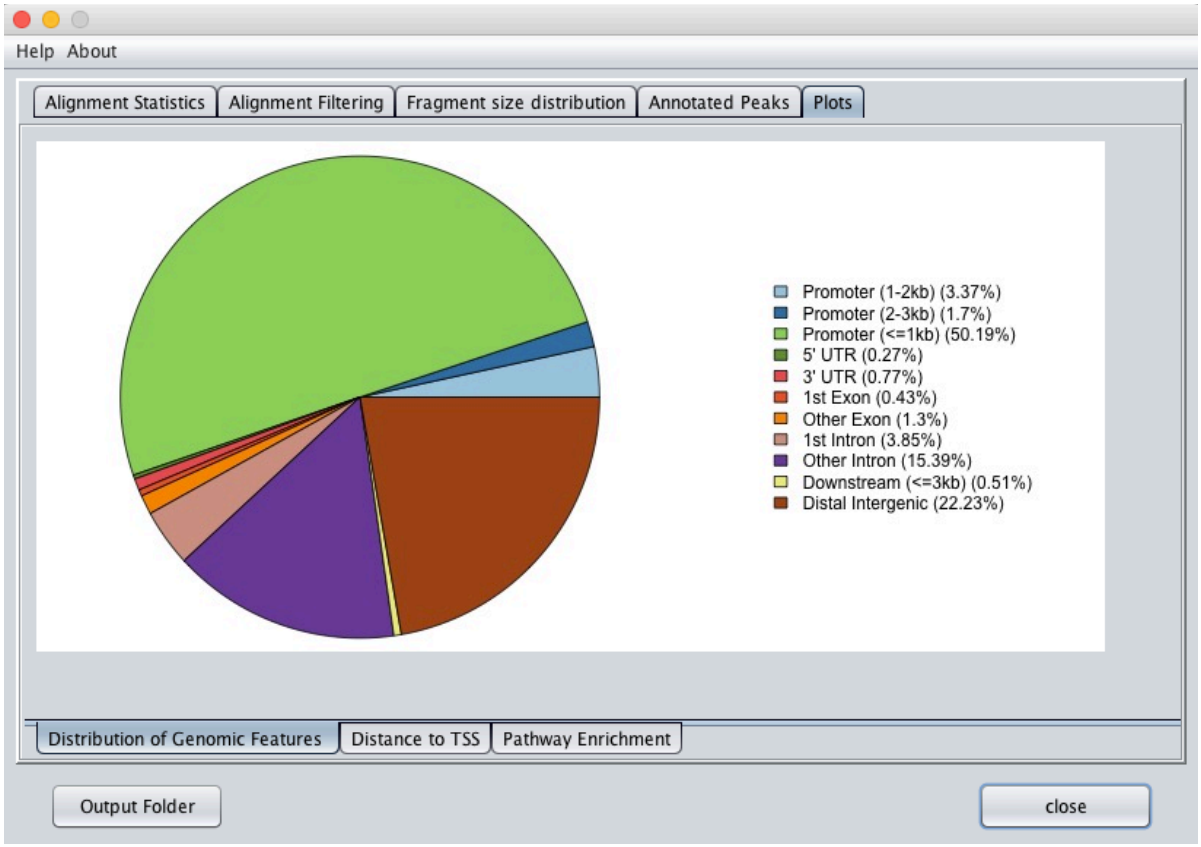

Figure 7. Distribution of ATAC-seq peaks in the genome

Pie chart showing the genomic distribution of ATAC-seq peaks obtained using the ChIPseeker bioconductor package. In this example, more than 50% of the peaks are located in the promoter region.

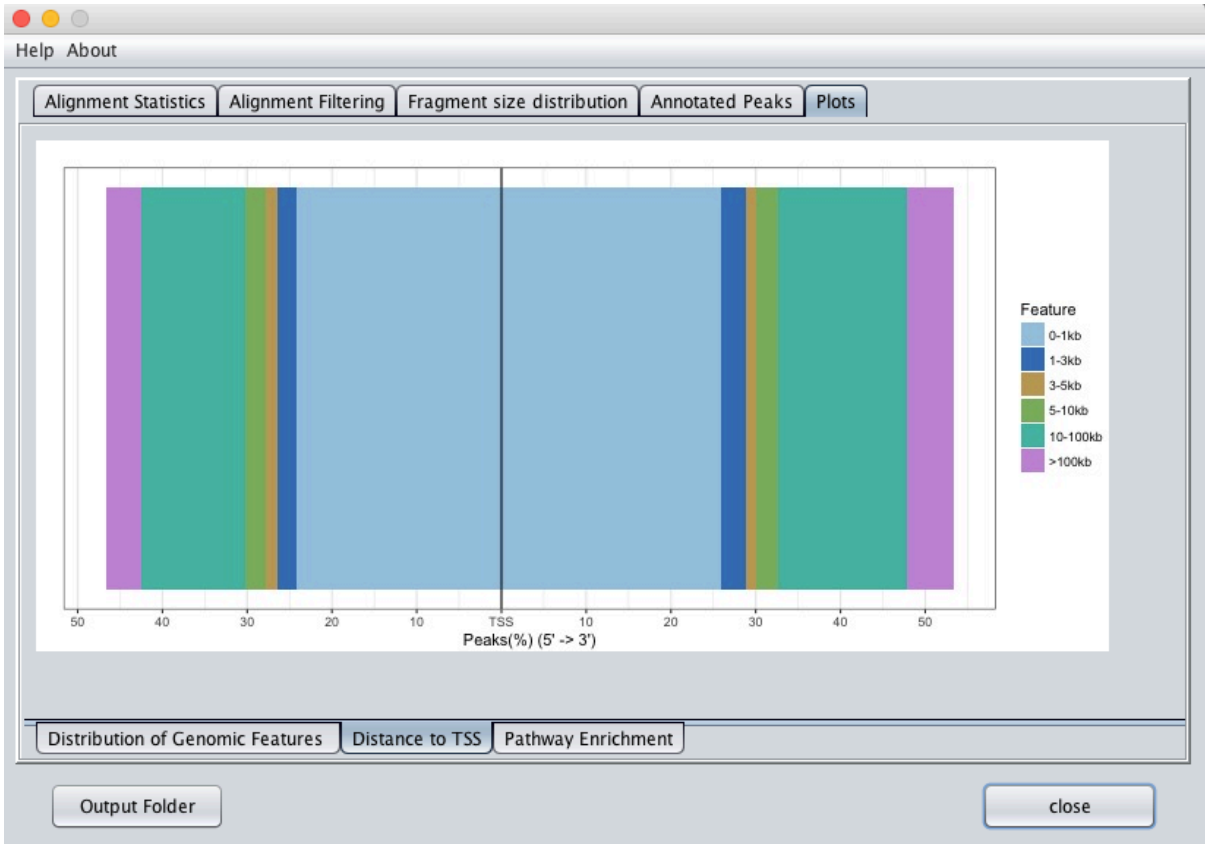

Figure 8. Distribution of ATAC-seq peaks relative to the TSS

Graph showing the percentage of peaks in a given distance range from the TSS. Different colors indicate the different range for distance and x-axis shows the percentage of total peaks. From above graph, around 50% peaks are within 0-1 kb (sky blue color) from the TSS.

Figure 9. Pathway enrichment analysis of ATAC-seq peaks

This graph shows the most significantly enriched pathways from given sample using the ReactomePA (Yu and He, 2016) bioconductor package. For example, the cell-cycle pathway is significantly enriched with an adjusted p-value of $1x10^{-6}$. ATAC-seq peaks and gene association (nearest gene) as shown in "Table of Annotated Peaks" (Figure 5) were used for the pathway enrichment analysis, which was obtained using the ChIPseeker bioconductor package.

4. 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
```

4.1 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 mesage

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

5. 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
```

6. 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

References

Broad Institute., Picard. http://broadinstitute.github.io/picard/.

Buenrostro, J.D., et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;10(12):1213-1218

Feng, J., et al. Identifying ChIP-seq enrichment using MACS. Nat Protoc 2012;7(9):1728-1740.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal, North America 2011.

Robinson, J.T., et al. Integrative genomics viewer. Nat Biotechnol 2011;29(1):24-26.

Yu, G. and He, Q.Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol Biosyst 2016;12(2):477-479.