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 is developed to help researcher with minimal or no Linux background for analyzing ATAC-seq data. This document contains all the information which is required to install dependencies and use GUAVA graphical and command line interface. This document also explains the GUAVA graphical user interface using published ATAC-seq data. It also provides procedure to create bowtie index from genome fasta for bioinformatics novice.

1. How to install dependencies?

GUAVA depends on others tools in order to process ATAC-seq data e.g. bowtie for alignment. Those dependencies need to be installed on machine to use GUAVA. If any of the dependency is not found GUAVA will fail to start. Text followed by "$" is a command. Once terminal (command line provided by OS) is launched, user can simply copy and paste on terminal those command to complete installation.

1.1 Java 1.8 or latest

As GUAVA is developed in Java, it is required.

To innstall java on Mac OS,

- To Download java go 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 displays 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 Linux OS

copy paste following command to the terminal

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

Or follow the link:
https://java.com/en/download/help/linux_x64_install.xml

1.2 Bowtie version 1.1.2

- 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. location and to paste it on terminal use,
  Mac => command + v
  Linux => ctrl + shift + v

- Launch the terminal and Use following commands to terminal to install bowtie

```
$ cp <bowtie file path> ~/
$ cd ~/
$ unzip bowtie-1.1.2*.zip
$ cd bowtie-1.1.2/
```

Mac OS

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

Linux OS

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

1.3 Python version 2.7

Python is required for macs2 installation.

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.

Linux OS, Use following command to install python

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

1.4 MACS2 version 2.1.1.20160309

To install MACS2 on Mac or Linux OS use command below

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

1.5 SAMtools Version: 1.3.1

- Download samtools-1.3.1.tar.bz2 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
  To paste path on terminal use
  Mac => command + v
  Linux => ctrl + shift + v
- Open the terminal
- Use copy 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
```

- Mac

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

- Linux

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

1.6 R Version: >= 3.3.0

Mac user can follow the link below for video tutorial 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 installation

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

2. How to download and launch GUAVA?

To download GUAVA use from GitHub

https://github.com/MayurDivate/GUAVA

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

Unzip downloaded GUAVA package

```
$ cd <GUAVA folder>
```

To launch GUAVA graphical user interface use following command

```
$ cd <GUAVA folder>
$ java –jar GUAVA.jar
```

3. GUAVA graphical user interface

We used SRR891275[1] dataset to demonstrate GUAVA graphical user interface.

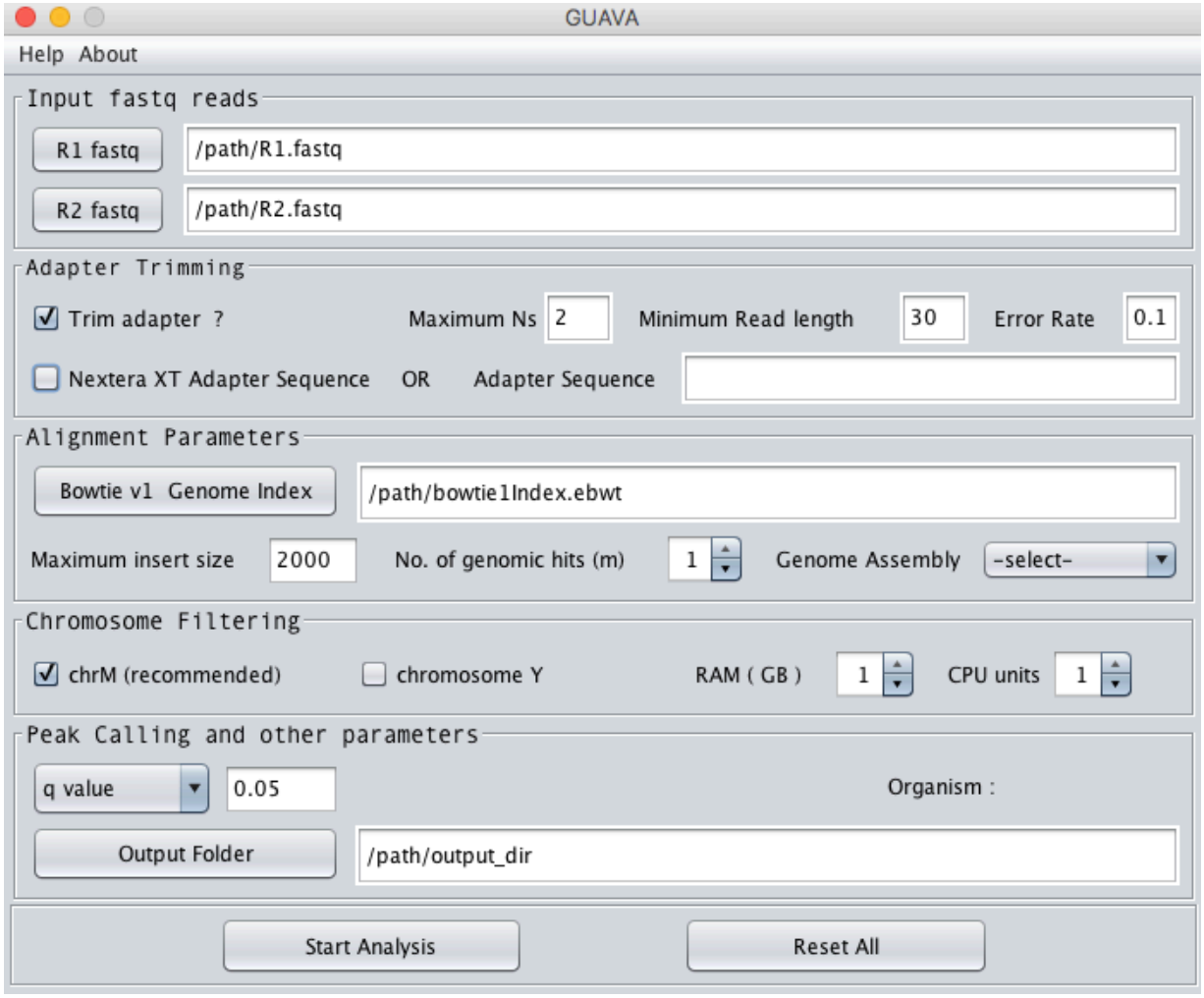

Figure 1: Graphical user interface of GUAVA

User can input the FASTQ files and bowtie index of genome fasta file using the buttons for R1 fastq, R2 fastq and Bowtie v1 Genome Index, respectively. If user select “Trim Data?” check box, then adapter will be trimmed from sequencing reads using cutadapt(Martin, 2011) before aligning them to genome. User can either select Nextera XT adapter or can provide custom adapter sequence. Reads with more “Maximum Ns” and less than “Minimum Read length” after adapter trimming will be filtered. Maximum distance for mapping of read mates can be set using “Maximum insert size” input field. To exclude reads pairs which are exceeding desired number of alignments (m) limits can be set

using “No. of genomic hits(m)” input field. User can specify the assembly or version name using genome assembly drop down menu. If chromosome checkbox is selected, reads mapping to that chromosome will be discarded. Dropdown menu is provided to select p or q value and input field next it can be used to set cutoff to filter MACS2(Feng, et al., 2012) peaks. RAM and threads can be set using box 5 and 6 respectively. Start analysis button will start ATAC-seq data processing, if provided inputs values are valid else in valid values will be highlighted with red color.

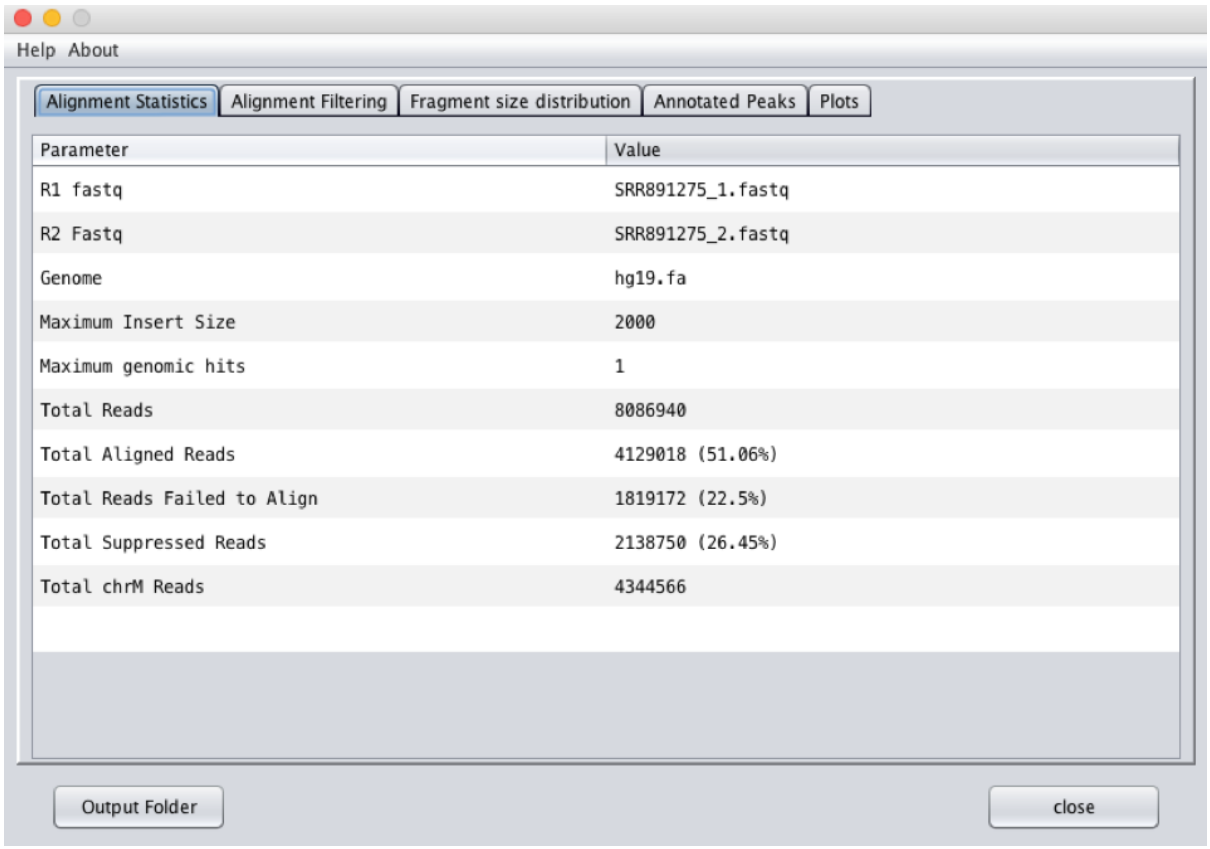

Figure 2: Result interface tab for alignment

This includes the input summary: names of FASTQ files and genome assembly used for analysis, maximum insert size and genomic hits. Total number of reads, aligned reads un aligned reads , fail to align due to m limit. Reads mapping to chr M and If chrY selected.

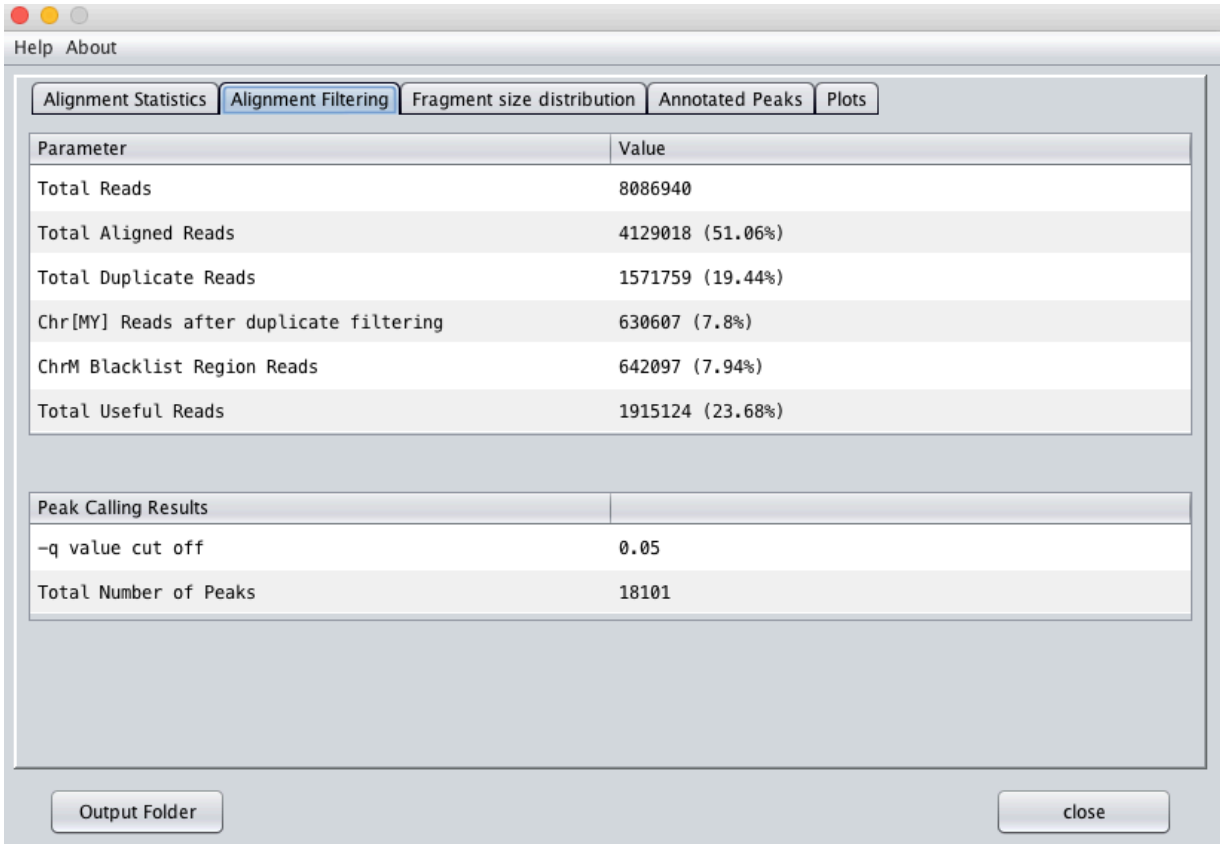

Figure 3: Alignment filtering and peak calling summary

The first table provides the number and percentage of reads relative to the total input reads i) aligned reads ii) duplicate reads iii) reads aligned to chromosome M and Y (based on selection) iv) reads mapped in blacklisted regions v) Useful reads represents the actual reads that qualifies for further processing. Second table gives you total number of peaks called and p/q value cutoff as per input

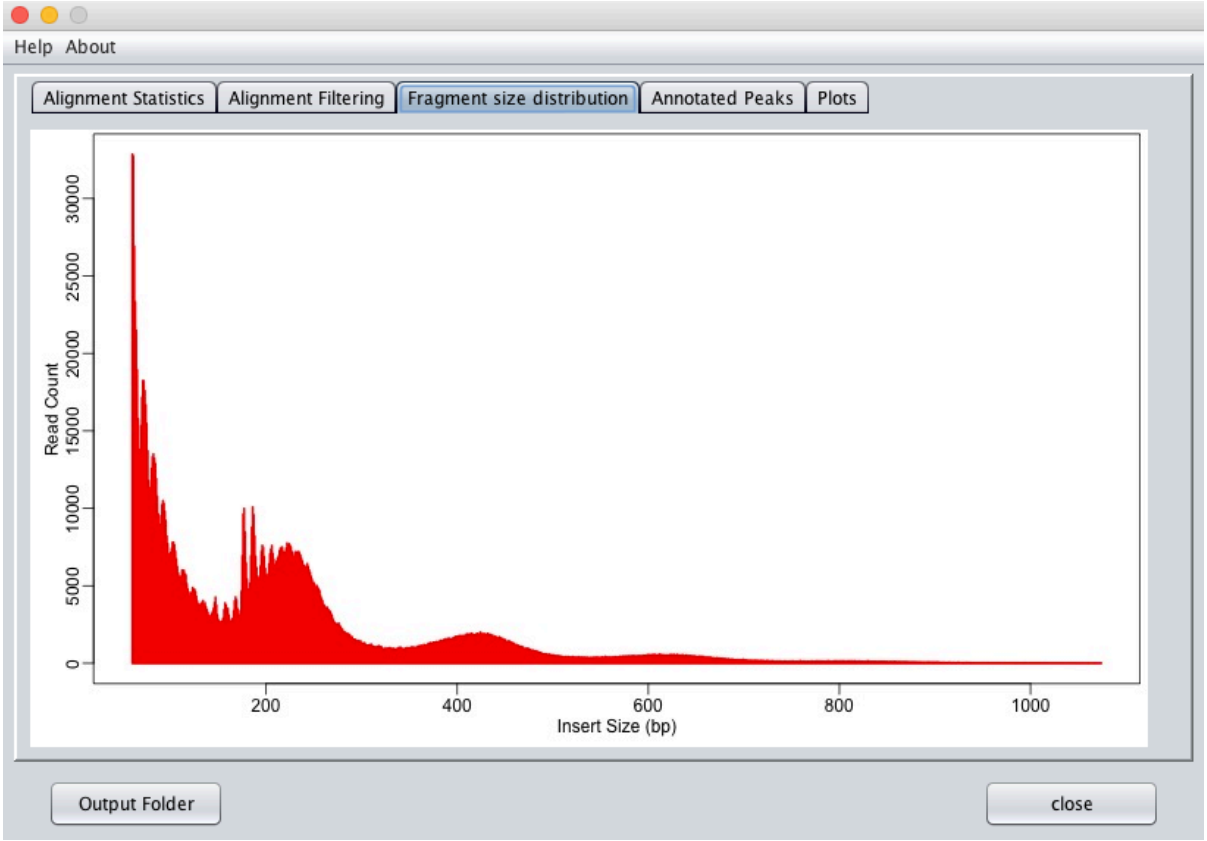

Figure 4: Fragment size distribution graph

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

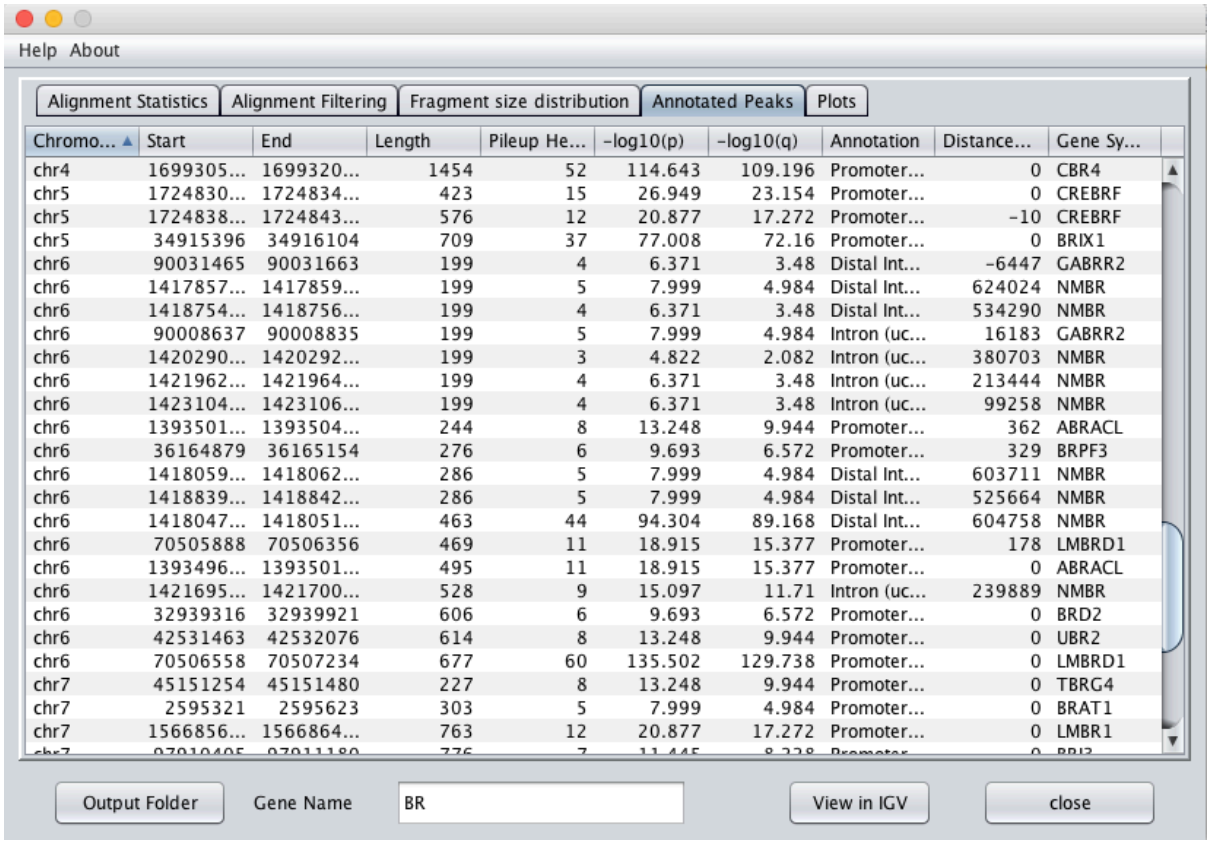

Figure 5: Peak annotation table

Annotated peaks can be browsed using peak annotation table. Gene name search boxed helps to filter peaks using gene symbol column. Table can can be sorted by clicking on header header of column. View in IGV button load the data tracks in IGV(Robinson, et al., 2011) and sets IGV location same as peak coordinates.

Figure 6: Peak visualization using IGV

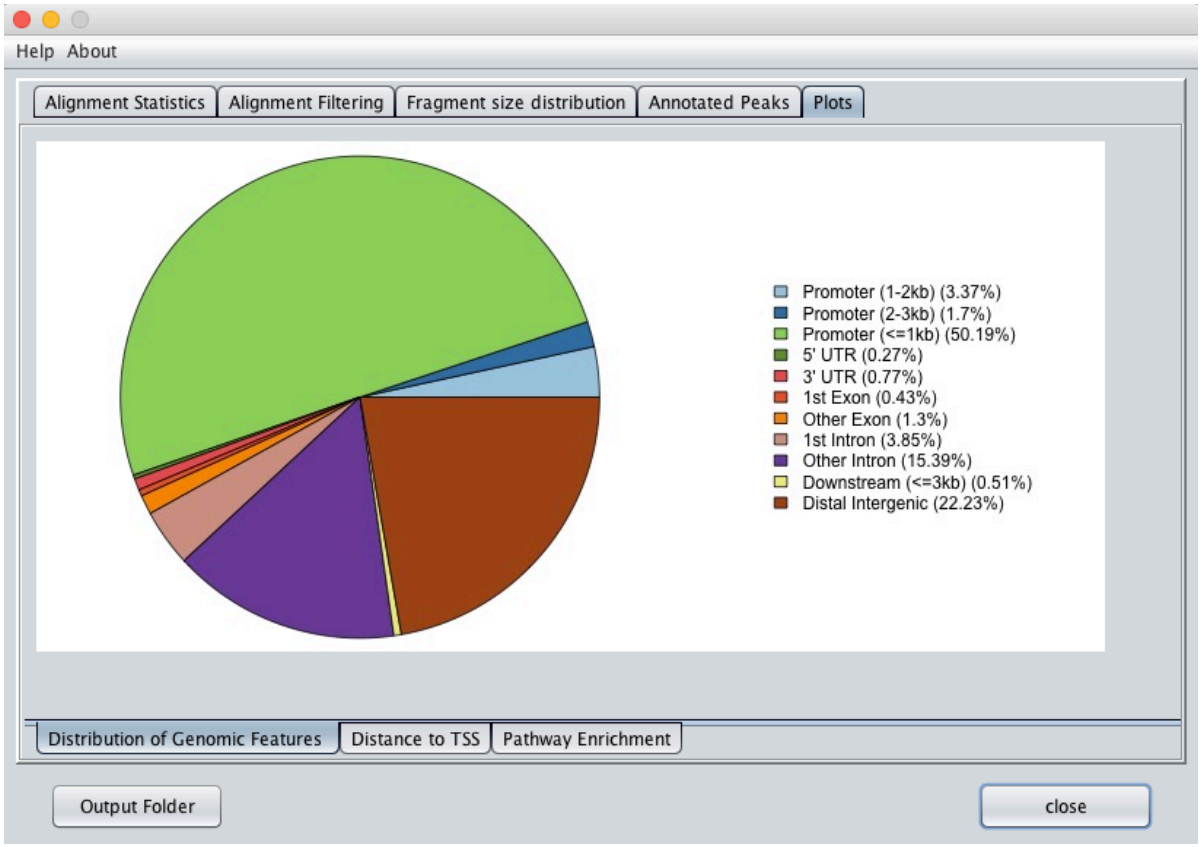

Figure 7: Peak annotation pie chart

Pie chart showing genomic feature wise peak distribution obtained using ChIPseeker bioconductor package. In this example, more than 50% of the peaks are in the promoter region. The promoter region is divided into three sub-categories: 0 to 1 kb, 1 to 2 kb and 2 to 3 kb from the TSS.

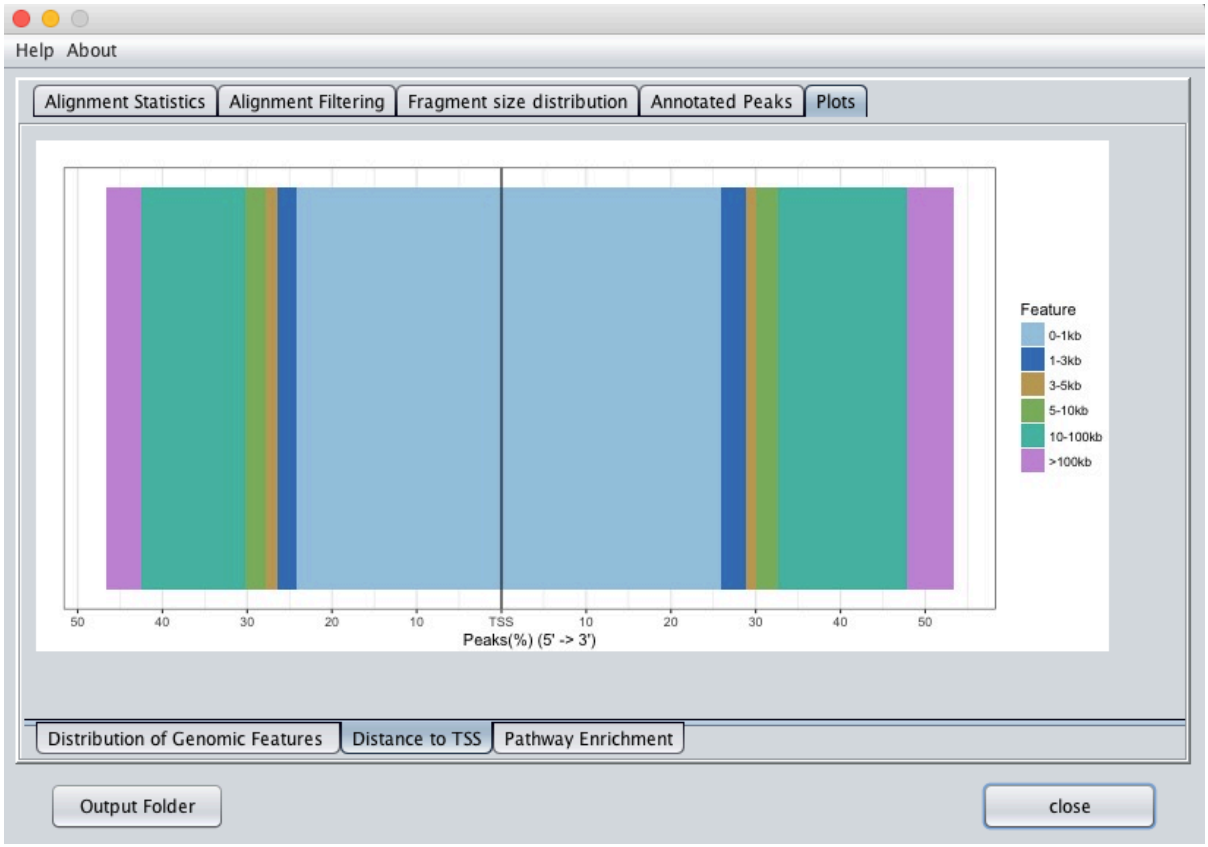

Figure 8: Peak distribution based on relative distance to the TSS

Graph showing the percentage of peaks in a given range from the TSS. Different colors indicate the different range for distance and y-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

This graph shows most significantly enriched pathway in given sample using ReactomePA(Yu and He, 2016) bioconductor package. Here, Cell cycle pathway enriched with $1x10^{-6}$ adjusted p-value.

4. GUAVA command-line interface

Like graphical user interface also has command line interface. Command line user interface make GUAVA flexible and therefore it can be easily integrated into existing pipelines. Also provides flexibility for running GUAVA through resource manger or job scheduler systems such as SLURM.

To print help message

```
$ java –jar GUAVA.jar –h
```

4.1 Usage and option summary

Usage: $ java -jar ATAC_GUI.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 GUAVA command line interface. Compulsory options are shown in blue color.

5. How to download genome fasta file?

To download genome fasta files please follow the following links.

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

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

Select 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 file use the following command:

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

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

e.g.

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

6. How to create bowtie index of genome fasta file?

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

```
$ cd <path to genome fasta file>
$ bowtie-build <genome.fasta> <genome>
```

Note: This is a time consuming step

References

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

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

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.