

GUAVA Manual

Mayur Divate and Edwin Cheung

Contact: mdivate@umac.mo

Index

1. How to install dependencies	3
1.1 Java 1.8 or latest	3
1.2 Bowtie version 1.1.2	4
1.3 Python version 2.7	4
1.4 MACS2 version 2.1.1.20160309	4
1.5 SAMtools Version: 1.3.1	5
1.6 R Version: >= 3.3.0	5
2. How to download and launch GUAVA	5
3. GUAVA graphical user interface	6
4. GUAVA command-line interface	15
4.1 Usage and option summary	16
5. How to download genome fasta file	16
6. How to create bowtie index of genome fasta file	17
7. References	17

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 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. location and to paste it in the Terminal,
Mac => command + v
Linux => ctrl + shift + v
- Launch the 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 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-eJTNBDdKgJgJzaQcY60XmsXAHU

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 GUAVA package to home folder and unzip it packages. To do this, Launch the 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

To demonstrate how to use the GUAVA graphical user interface and show the typical results that are obtained from the program, we have used an ATAC-seq dataset from xxx SRR891275[1].

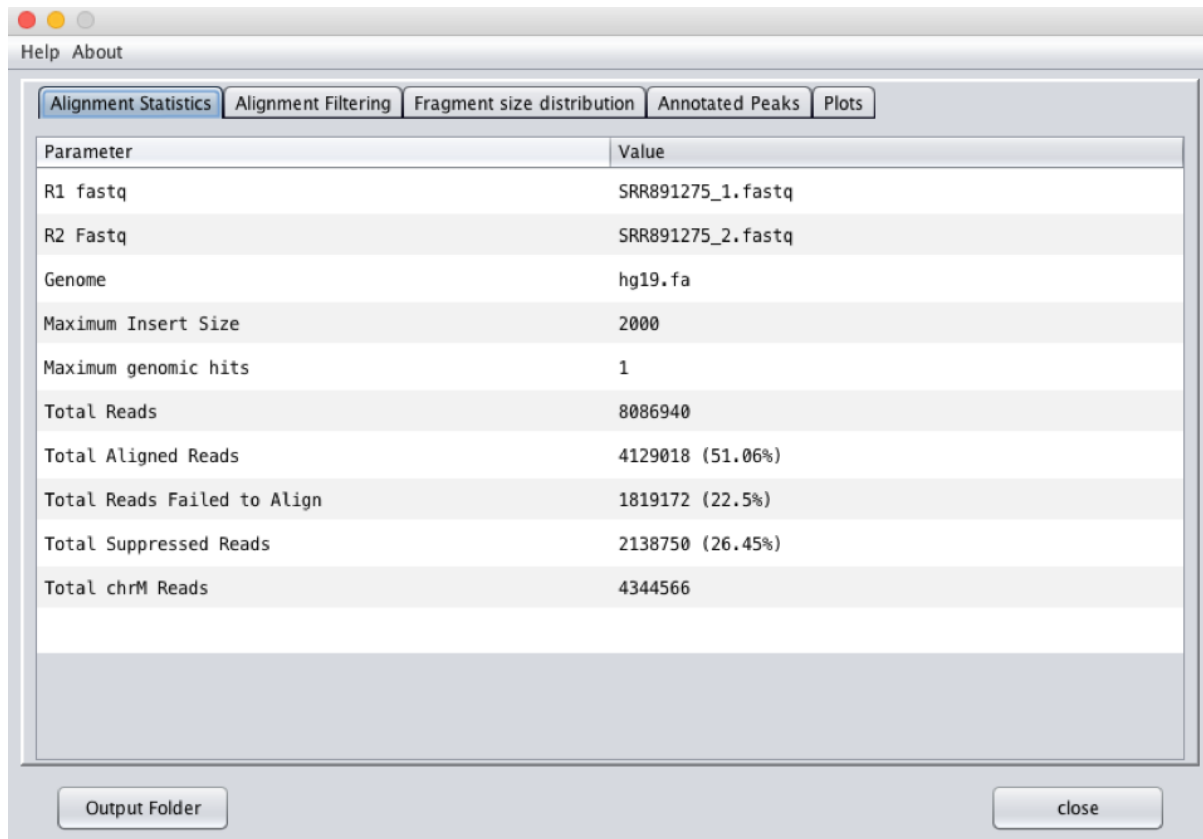
The screenshot shows the GUAVA graphical user interface (GUI) window. The window has a title bar with the text "GUAVA" and a menu bar with "Help" and "About" options. The main content area is divided into several sections:

- Input fastq reads:** Contains two text input fields. The first is labeled "R1 fastq" and contains the path "/path/R1.fastq". The second is labeled "R2 fastq" and contains the path "/path/R2.fastq".
- Adapter Trimming:** Contains a checked checkbox labeled "Trim adapter ?". To its right are three input fields: "Maximum Ns" with the value "2", "Minimum Read length" with the value "30", and "Error Rate" with the value "0.1". Below these is an unchecked checkbox labeled "Nextera XT Adapter Sequence" followed by the text "OR Adapter Sequence" and an empty text input field.
- Alignment Parameters:** Contains a button labeled "Bowtie v1 Genome Index" next to a text input field containing "/path/bowtie1Index.ebwt". Below this are three input fields: "Maximum insert size" with the value "2000", "No. of genomic hits (m)" with the value "1", and "Genome Assembly" with a dropdown menu showing "--select--".
- Chromosome Filtering:** Contains two checkboxes: "chrM (recommended)" which is checked, and "chromosome Y" which is unchecked. To the right are two input fields: "RAM (GB)" with the value "1" and "CPU units" with the value "1".
- Peak Calling and other parameters:** Contains a dropdown menu labeled "q value" with the value "0.05". To its right is a text input field labeled "Organism :". Below these is a text input field labeled "Output Folder" containing the path "/path/output_dir".

At the bottom of the window are two buttons: "Start Analysis" and "Reset All".

Figure 1. Graphical user interface of GUAVA

Users can input the FASTQ files and bowtie index of genome fasta file by using the buttons for R1 fastq, R2 fastq and Bowtie v1 Genome Index, respectively. If user select "Trim Data?" checkbox, 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, User can change the default. 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 chromosome checkbox is selected, reads mapping to that chromosome will be discarded. Dropdown menu is provided for selecting p or q value and input field next it can be used for setting 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. Start analysis button will initiate the processing of ATAC-seq data, however, no processing will be performed if any invalid values (highlighted with red) were entered.



Parameter	Value
R1 fastq	SRR891275_1.fastq
R2 Fastq	SRR891275_2.fastq
Genome	hg19.fa
Maximum Insert Size	2000
Maximum genomic hits	1
Total Reads	8086940
Total Aligned Reads	4129018 (51.06%)
Total Reads Failed to Align	1819172 (22.5%)
Total Suppressed Reads	2138750 (26.45%)
Total chrM Reads	4344566

Output Folder

close

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 m limit. The reads mapping to chr M and chrY are also shown if selected.

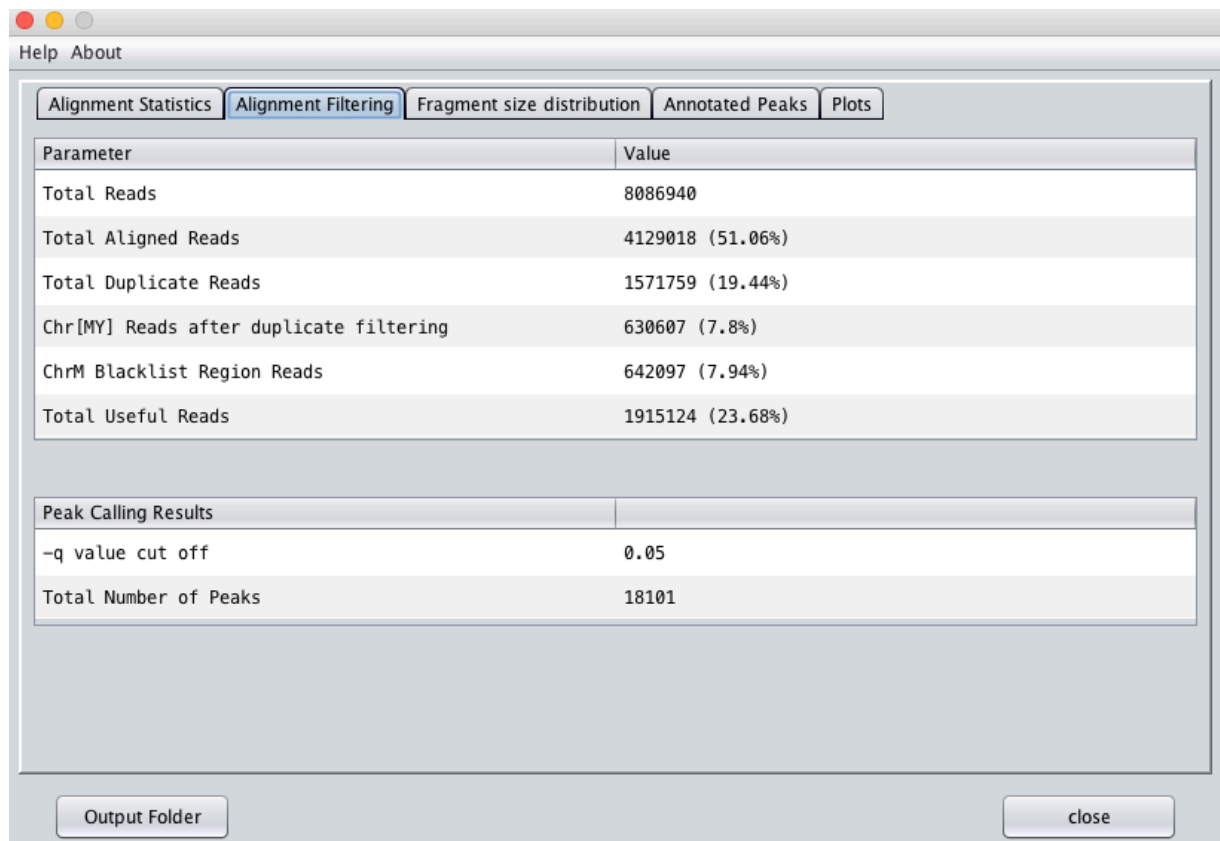


Figure 3. Summary of alignment filtering and peak calling information

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 p- or q-value cutoff as per input.

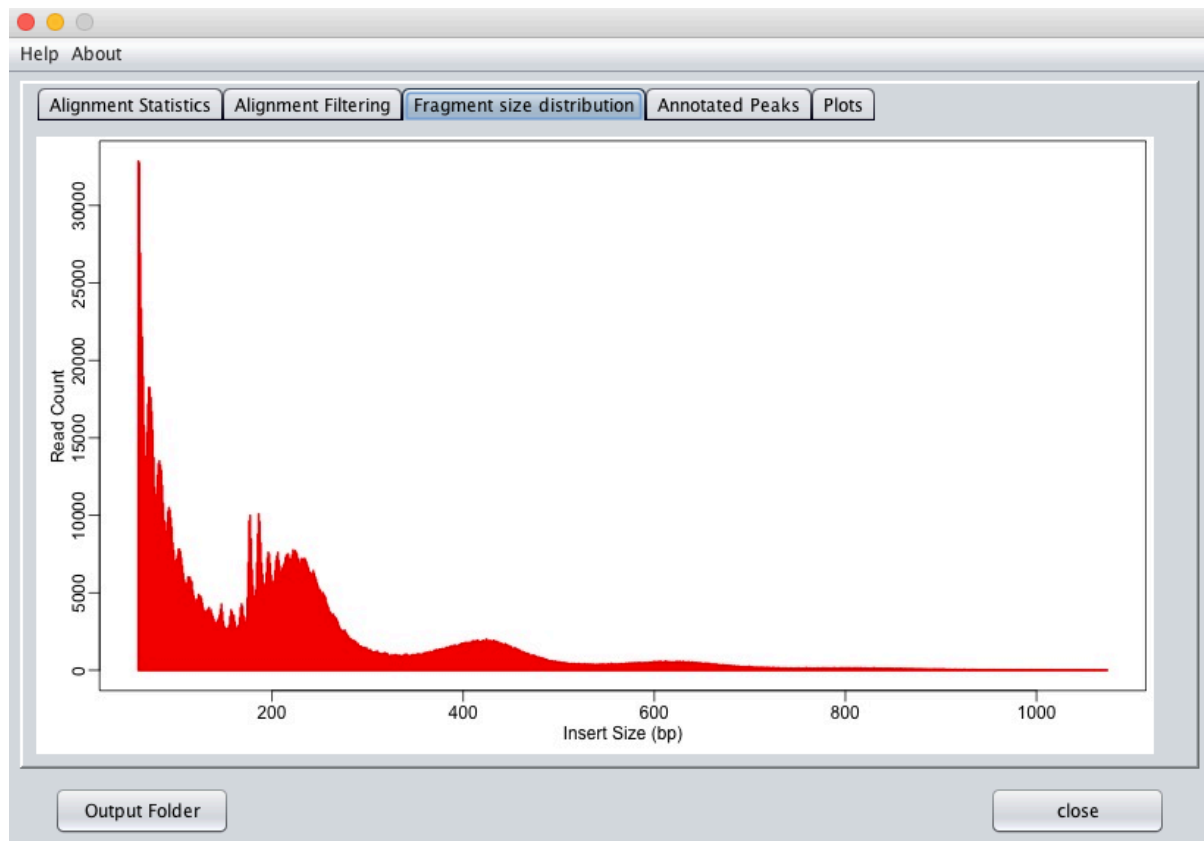


Figure 4. Fragment size distribution graph

This graph gives 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.

Help About

Alignment Statistics Alignment Filtering Fragment size distribution Annotated Peaks Plots

Chromo...	Start	End	Length	Pileup He...	-log10(p)	-log10(q)	Annotation	Distance...	Gene Sy...
chr4	1699305...	1699320...	1454	52	114.643	109.196	Promoter...	0	CBR4
chr5	1724830...	1724834...	423	15	26.949	23.154	Promoter...	0	CREBRF
chr5	1724838...	1724843...	576	12	20.877	17.272	Promoter...	-10	CREBRF
chr5	34915396	34916104	709	37	77.008	72.16	Promoter...	0	BRX1
chr6	90031465	90031663	199	4	6.371	3.48	Distal Int...	-6447	GABRR2
chr6	1417857...	1417859...	199	5	7.999	4.984	Distal Int...	624024	NMBR
chr6	1418754...	1418756...	199	4	6.371	3.48	Distal Int...	534290	NMBR
chr6	90008637	90008835	199	5	7.999	4.984	Intron (uc...	16183	GABRR2
chr6	1420290...	1420292...	199	3	4.822	2.082	Intron (uc...	380703	NMBR
chr6	1421962...	1421964...	199	4	6.371	3.48	Intron (uc...	213444	NMBR
chr6	1423104...	1423106...	199	4	6.371	3.48	Intron (uc...	99258	NMBR
chr6	1393501...	1393504...	244	8	13.248	9.944	Promoter...	362	ABRACL
chr6	36164879	36165154	276	6	9.693	6.572	Promoter...	329	BRPF3
chr6	1418059...	1418062...	286	5	7.999	4.984	Distal Int...	603711	NMBR
chr6	1418839...	1418842...	286	5	7.999	4.984	Distal Int...	525664	NMBR
chr6	1418047...	1418051...	463	44	94.304	89.168	Distal Int...	604758	NMBR
chr6	70505888	70506356	469	11	18.915	15.377	Promoter...	178	LMBRD1
chr6	1393496...	1393501...	495	11	18.915	15.377	Promoter...	0	ABRACL
chr6	1421695...	1421700...	528	9	15.097	11.71	Intron (uc...	239889	NMBR
chr6	32939316	32939921	606	6	9.693	6.572	Promoter...	0	BRD2
chr6	42531463	42532076	614	8	13.248	9.944	Promoter...	0	UBR2
chr6	70506558	70507234	677	60	135.502	129.738	Promoter...	0	LMBRD1
chr7	45151254	45151480	227	8	13.248	9.944	Promoter...	0	TBRG4
chr7	2595321	2595623	303	5	7.999	4.984	Promoter...	0	BRAT1
chr7	1566856...	1566864...	763	12	20.877	17.272	Promoter...	0	LMBR1
chr7	87810405	87811180	776	7	11.445	8.228	Promoter...	0	BRD2

Output Folder Gene Name BR View in IGV close

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 IGV (Robinson, et al., 2011) genome browser and set the IGV location same as the peak coordinates.



Figure 6. Peak visualization using IGV

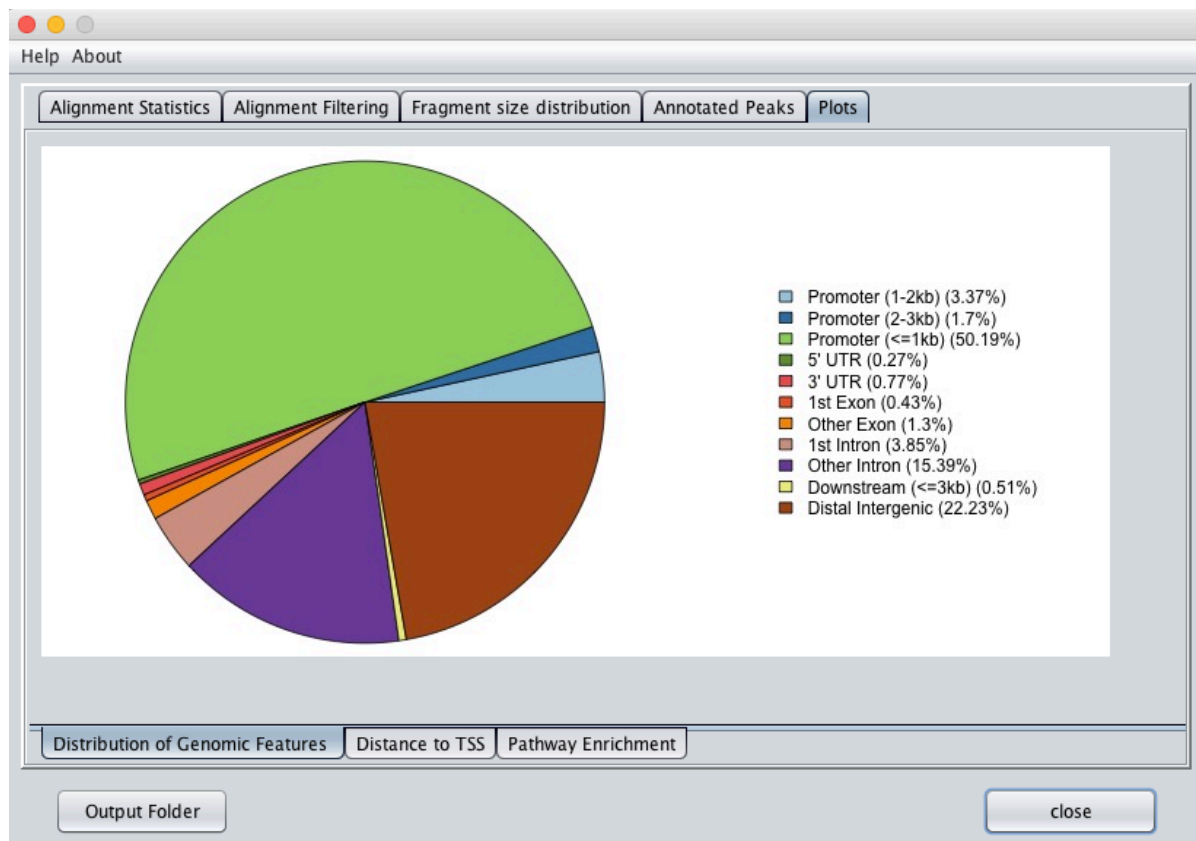


Figure 7. Distribution of ATAC-seq peaks in the genome

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.

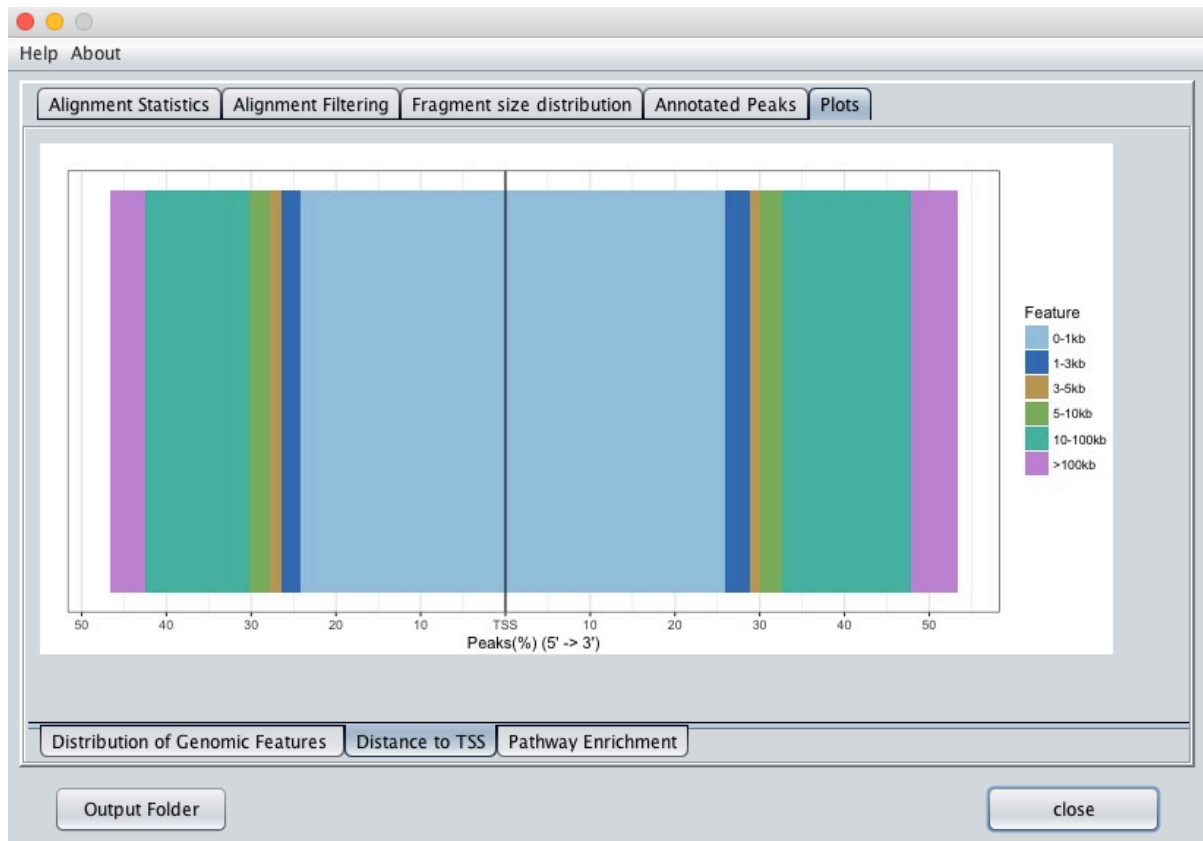


Figure 8. Distribution of ATAC-seq peaks relative to the TSS

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 pathway in given sample using ReactomePA (Yu and He, 2016) bioconductor package. Here, the cell-cycle pathway is significantly enriched with an adjusted p-value of 1×10^{-6} . ATAC-seq peak and gene association (nearest gene) as shown in “Table of Annotated Peaks”(figure 5), used for pathway enrichment analysis, which was obtained using 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 message

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

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

Broad Institute., 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.