GUAVA Manual

Mayur Divate and Edwin Cheung

Contact: mdivate@umac.mo

Index

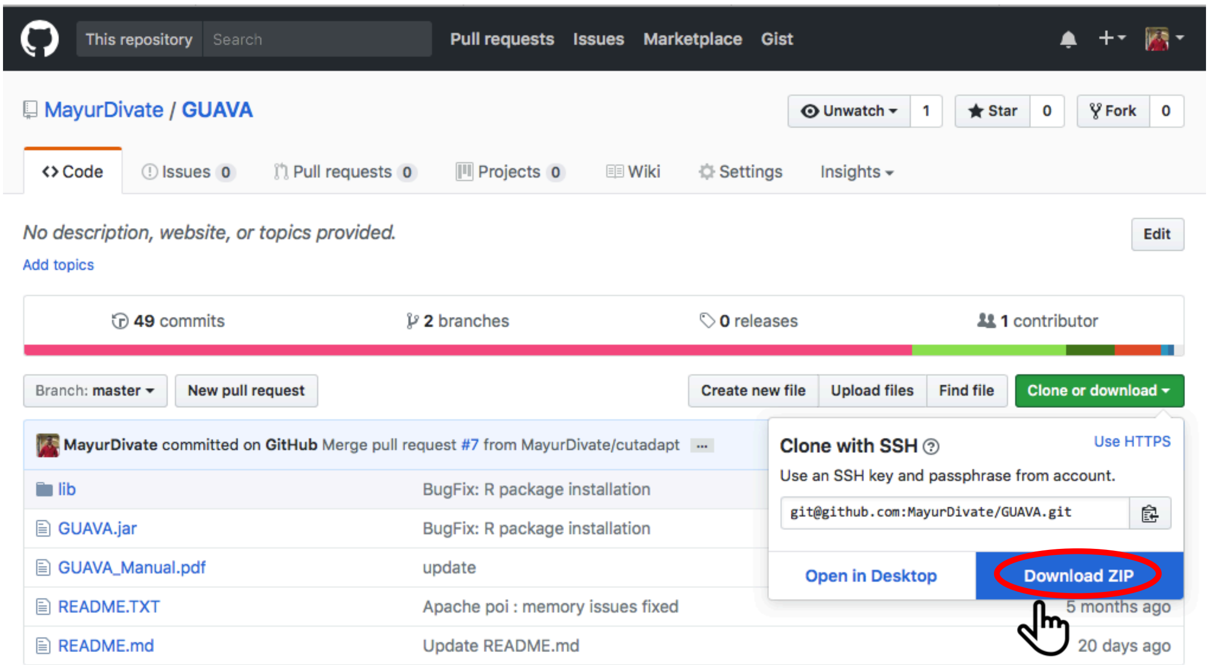
GUAVA: a GUI tool for the Analysis and Visualization of ATAC-seq data

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 download and launch GUAVA

To download GUAVA, go to GitHub via this link:

https://github.com/MayurDivate/GUAVA



No description, website, or topics provided.

MayurDivate committed on GitHub Merge pull request #7 from MayurDivate/cutadapt

lib	BugFix: R package installation	
GUAVA.jar	BugFix: R package installation	
GUAVA_Manual.pdf	update	
README.TXT	Apache poi : memory issues fixed	5 months ago
README.md	Update README.md	20 days ago

Clone with SSH

Use an SSH key and passphrase from account.

```
git@github.com:MayurDivate/GUAVA.git
```

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:

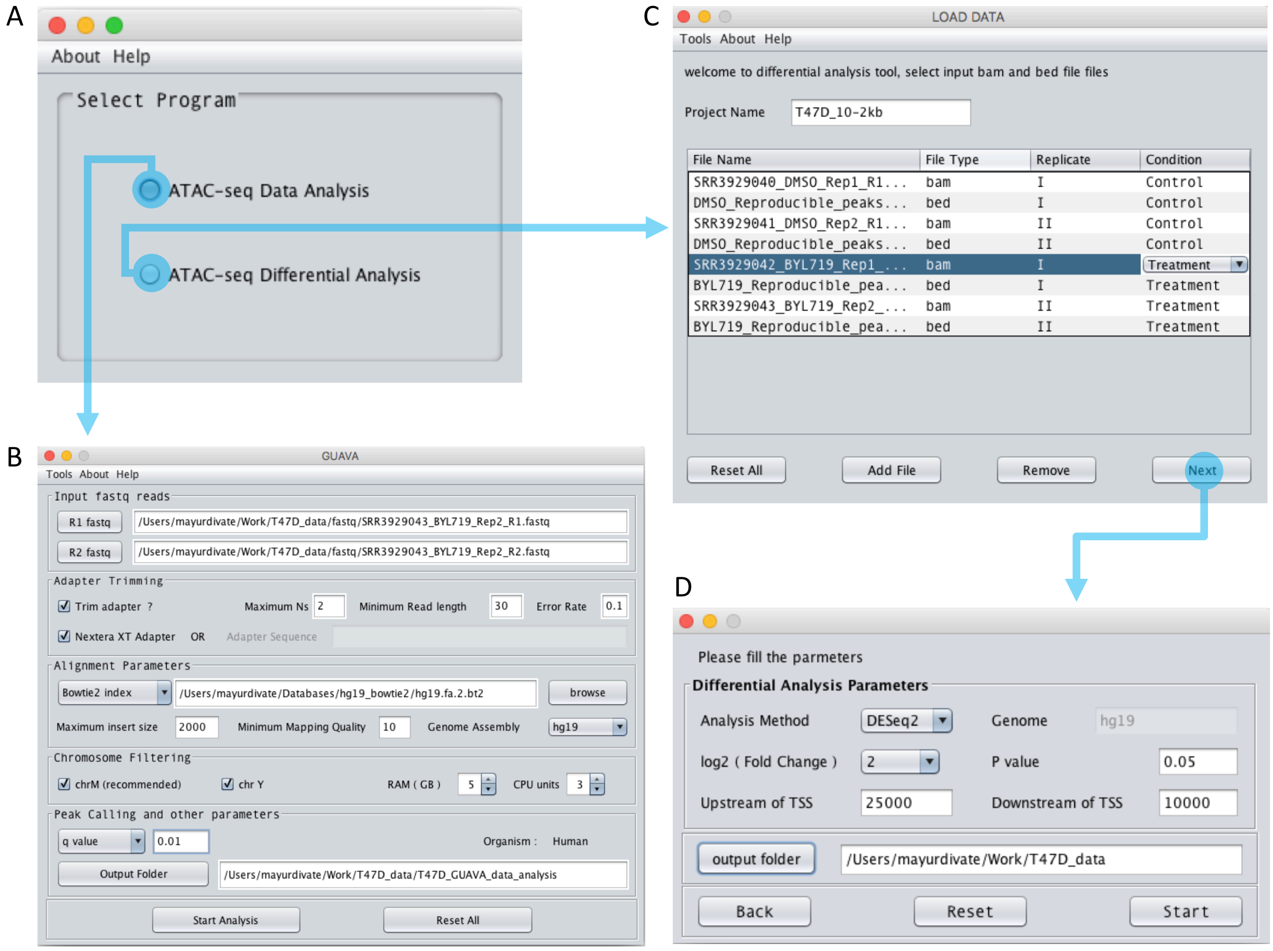
```
$ cp </path/to/ GUAVA-master.zip> ~
$ cd ~
$ unzip GUAVA-master.zip
```

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.



A

Select Program

○ ATAC-seq Data Analysis

○ ATAC-seq Differential Analysis

B — GUAVA

Input fastq reads

R1 fastq: /Users/mayurdivate/Work/T47D_data/fastq/SRR3929043_BYL719_Rep2_R1.fastq

R2 fastq: /Users/mayurdivate/Work/T47D_data/fastq/SRR3929043_BYL719_Rep2_R2.fastq

Adapter Trimming

☑ Trim adapter ? Maximum Ns 2 Minimum Read length 30 Error Rate 0.1

☑ Nextera XT Adapter OR Adapter Sequence

Alignment Parameters

Bowtie2 index /Users/mayurdivate/Databases/hg19_bowtie2/hg19.fa.2.bt2

Maximum insert size 2000 Minimum Mapping Quality 10 Genome Assembly hg19

Chromosome Filtering

☑ chrM (recommended) ☑ chr Y RAM (GB) 5 CPU units 3

Peak Calling and other parameters

q value 0.01 Organism : Human

Output Folder /Users/mayurdivate/Work/T47D_data/T47D_GUAVA_data_analysis

C — LOAD DATA

welcome to differential analysis tool, select input bam and bed file files

Project Name T47D_10-2kb

File Name	File Type	Replicate	Condition
SRR3929040_DMSO_Rep1_R1...	bam	I	Control
DMSO_Reproducible_peaks...	bed	I	Control
SRR3929041_DMSO_Rep2_R1...	bam	II	Control
DMSO_Reproducible_peaks...	bed	II	Control
SRR3929042_BYL719_Rep1_...	bam	I	Treatment
BYL719_Reproducible_pea...	bed	I	Treatment
SRR3929043_BYL719_Rep2_...	bam	II	Treatment
BYL719_Reproducible_pea...	bed	II	Treatment

D

Please fill the parmeters

Differential Analysis Parameters

Analysis Method DESeq2 Genome hg19

log2 (Fold Change) 2 P value 0.05

Upstream of TSS 25000 Downstream of TSS 10000

output folder /Users/mayurdivate/Work/T47D_data

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 botiwe 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:

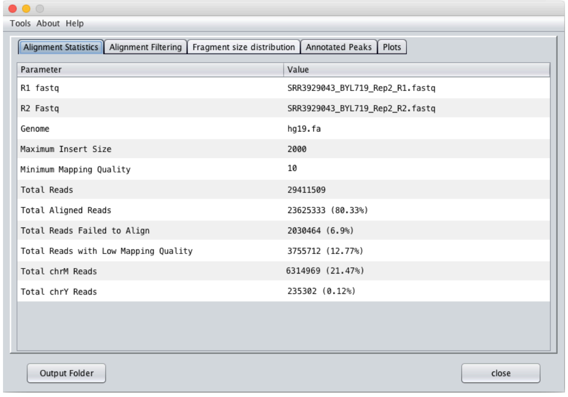
Log2(Fold Change): log2 fold change cut off

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.

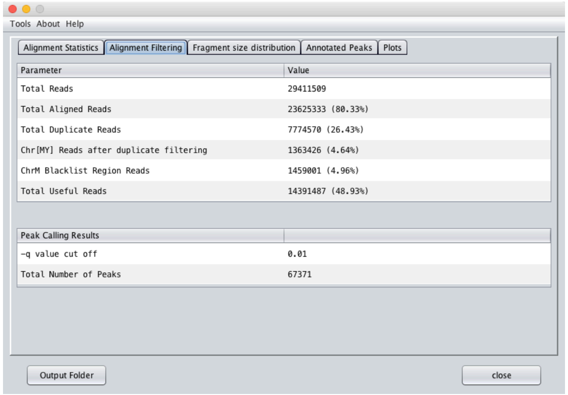
3. Output interface for GUAVA ATAC-seq data analysis

A



Parameter	Value
R1 fastq	SRR3929043_BYL719_Rep2_R1.fastq
R2 Fastq	SRR3929043_BYL719_Rep2_R2.fastq
Genome	hg19.fa
Maximum Insert Size	2000
Minimum Mapping Quality	10
Total Reads	29411509
Total Aligned Reads	23625333 (80.33%)
Total Reads Failed to Align	2030464 (6.9%)
Total Reads with Low Mapping Quality	3755712 (12.77%)
Total chrM Reads	6314969 (21.47%)
Total chrY Reads	235302 (0.12%)

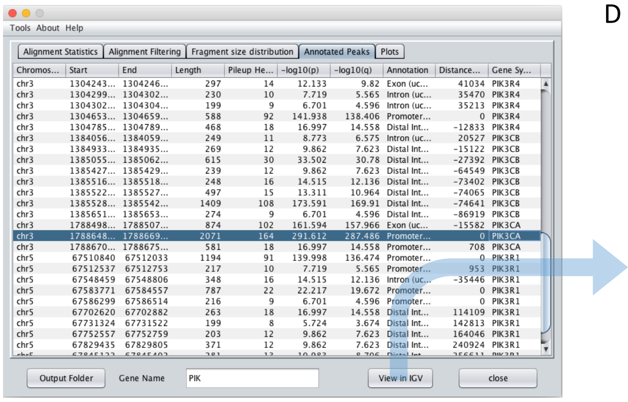
B



Parameter	Value
Total Reads	29411509
Total Aligned Reads	23625333 (80.33%)
Total Duplicate Reads	7774570 (26.43%)
Chr[MY] Reads after duplicate filtering	1363426 (4.64%)
ChrM Blacklist Region Reads	1459001 (4.96%)
Total Useful Reads	14391487 (48.93%)

Peak Calling Results	
-q value cut off	0.01
Total Number of Peaks	67371

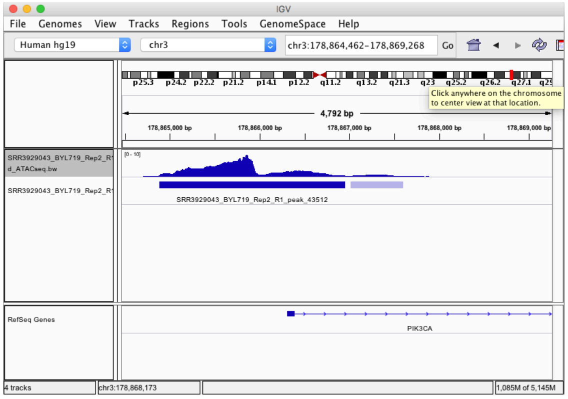
C



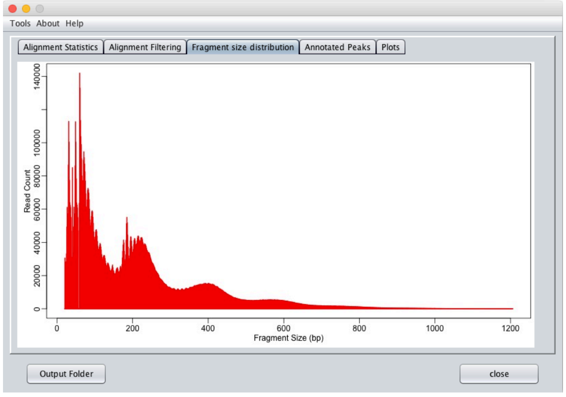
Gene Name: PIK

Chromos...	Start	End	Length	Pileup He...	-log10(p)	-log10(q)	Annotation	Distance...	Gene Sy...
chr3	1304243...	1304246...	297	14	12.133	9.82	Exon (uc...	41034	PIK3R4
chr3	1304299...	1304302...	230	10	7.719	5.565	Intron (uc...	35470	PIK3R4
chr3	1304302...	1304304...	199	9	6.701	4.596	Intron (uc...	35213	PIK3R4
chr3	1304653...	1304659...	588	92	141.938	138.406	Promoter...	0	PIK3R4
chr3	1304785...	1304789...	468	18	16.997	14.558	Distal Int...	-12833	PIK3R4
chr3	1384056...	1384059...	249	11	8.773	6.575	Intron (uc...	20527	PIK3CB
chr3	1384933...	1384935...	269	12	9.862	7.623	Distal Int...	-15122	PIK3CB
chr3	1385055...	1385062...	615	30	33.502	30.78	Distal Int...	-27392	PIK3CB
chr3	1385427...	1385429...	239	12	9.862	7.623	Distal Int...	-64549	PIK3CB
chr3	1385516...	1385518...	248	16	14.515	12.136	Distal Int...	-73402	PIK3CB
chr3	1385522...	1385527...	497	15	13.311	10.964	Distal Int...	-74065	PIK3CB
chr3	1385528...	1385542...	1409	108	173.591	169.91	Distal Int...	-74641	PIK3CB
chr3	1385651...	1385653...	274	9	6.701	4.596	Distal Int...	-86919	PIK3CB
chr3	1788498...	1788507...	874	102	161.594	157.966	Exon (uc...	-15582	PIK3CA
chr3	1788648...	1788669...	2071	164	291.612	287.486	Promoter...	0	PIK3CA
chr3	1788670...	1788675...	581	18	16.997	14.558	Promoter...	708	PIK3CA
chr5	67510840	67512033	1194	91	139.998	136.474	Promoter...	0	PIK3R1
chr5	67512537	67512753	217	10	7.719	5.565	Promoter...	953	PIK3R1
chr5	67548459	67548806	348	16	14.515	12.136	Intron (uc...	-35446	PIK3R1
chr5	67583771	67584557	787	22	22.217	19.672	Promoter...	0	PIK3R1
chr5	67586299	67586514	216	9	6.701	4.596	Promoter...	0	PIK3R1
chr5	67702620	67702882	263	18	16.997	14.558	Distal Int...	114109	PIK3R1
chr5	67731324	67731522	199	8	5.724	3.674	Distal Int...	142813	PIK3R1
chr5	67752557	67752759	203	12	9.862	7.623	Distal Int...	164046	PIK3R1
chr5	67829435	67829805	371	12	9.862	7.623	Distal Int...	240924	PIK3R1
chr5	67845123	67845402	281	13	10.983	8.706	Distal Int...	256611	PIK3R1

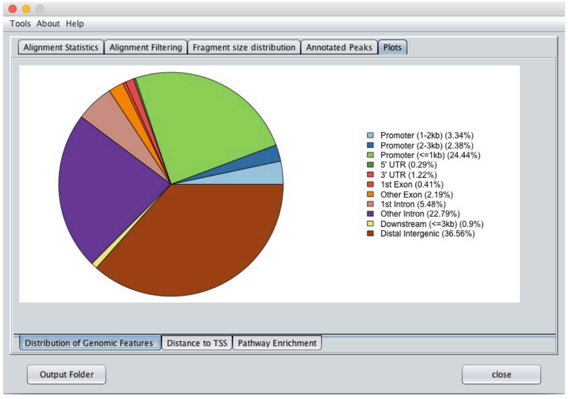
D


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr3
chr3:178,864,462-178,869,268
Go
p25.3 p24.2 p22.2 p21.2 p14.1 p12.2 q11.2 q13.2 q21.3 q23 q25.2 q26.2 q27.1 q29
Click anywhere on the chromosome to center view at that location.
4,792 bp
178,865,000 bp
178,866,000 bp
178,867,000 bp
178,868,000 bp
178,869,000 bp
SRR3929043_BYL719_Rep2_R1
d_ATACseq.bw
[0 - 10]
SRR3929043_BYL719_Rep2_R1
SRR3929043_BYL719_Rep2_R1_peak_43512
RefSeq Genes
PIK3CA
4 tracks
chr3:178,868,173
1,085M of 5,145M

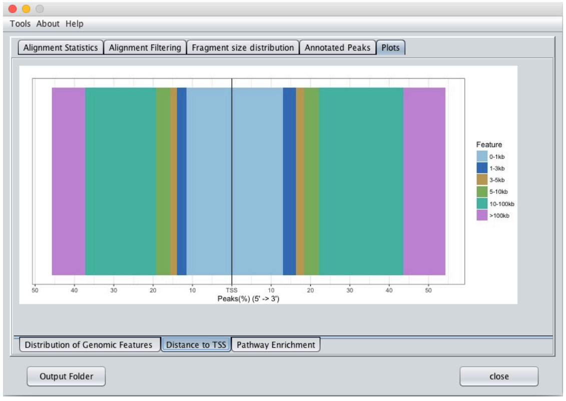
E



F



G



H

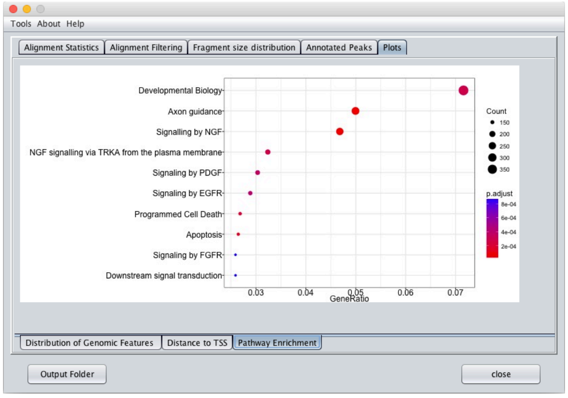
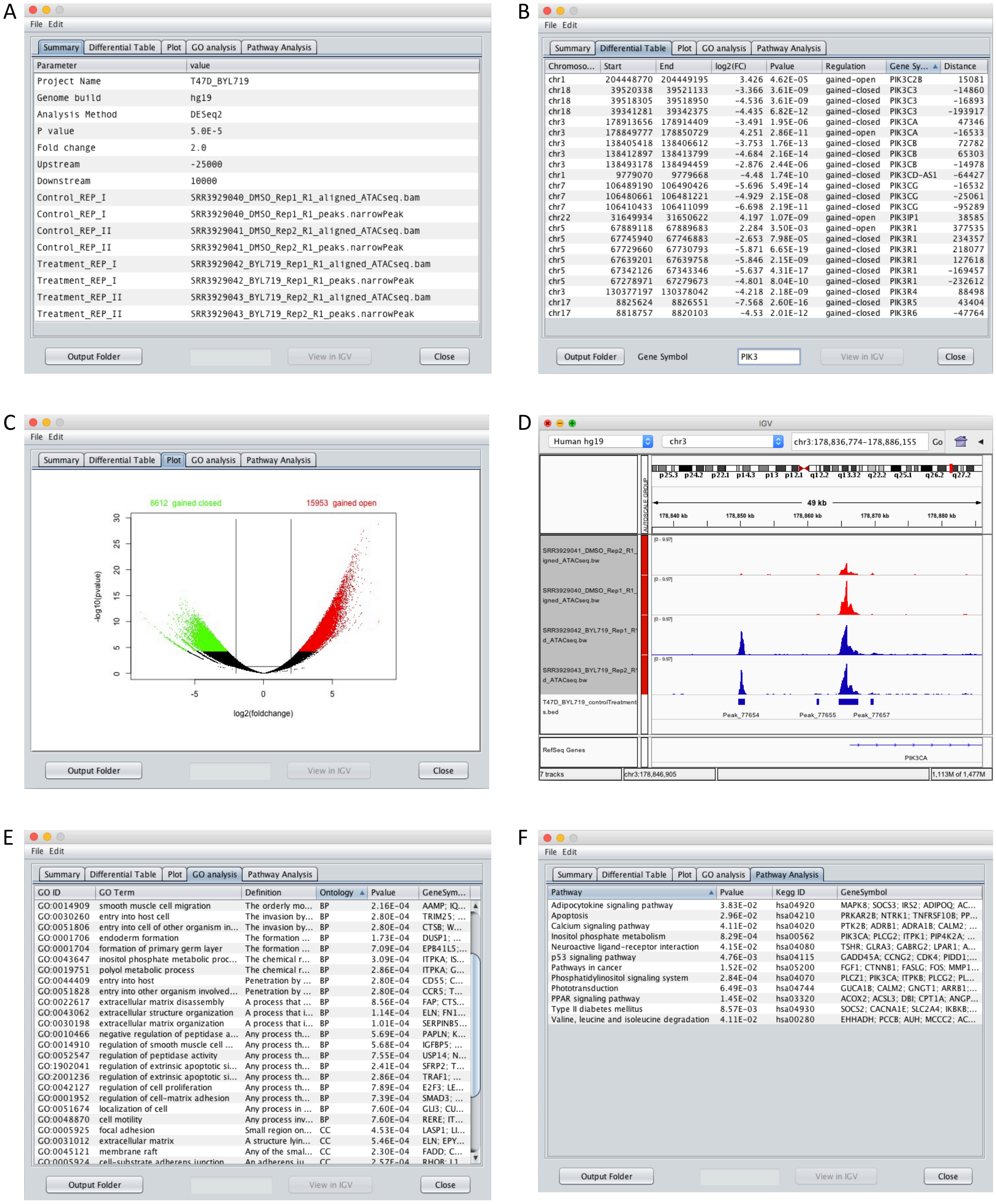


Figure 3: Output interface for GUAVA ATAC-seq data analysis. A) Input summary and alignment statistics. B) Read filtering and peak calling summary. C) Peak annotation table with sorting and filtering functionality. Easy access to IGV for visualizing peaks and automatically generated normalized ATAC-seq 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



Parameter	value
Project Name	T47D_BYL719
Genome build	hg19
Analysis Method	DESeq2
P value	5.0E-5
Fold change	2.0
Upstream	-25000
Downstream	10000
Control_REP_I	SRR3929040_DMSO_Rep1_R1_aligned_ATACseq.bam
Control_REP_I	SRR3929040_DMSO_Rep1_R1_peaks.narrowPeak
Control_REP_II	SRR3929041_DMSO_Rep2_R1_aligned_ATACseq.bam
Control_REP_II	SRR3929041_DMSO_Rep2_R1_peaks.narrowPeak
Treatment_REP_I	SRR3929042_BYL719_Rep1_R1_aligned_ATACseq.bam
Treatment_REP_I	SRR3929042_BYL719_Rep1_R1_peaks.narrowPeak
Treatment_REP_II	SRR3929043_BYL719_Rep2_R1_aligned_ATACseq.bam
Treatment_REP_II	SRR3929043_BYL719_Rep2_R1_peaks.narrowPeak

Chromoso...	Start	End	log2(FC)	Pvalue	Regulation	Gene Sy... ▲	Distance
chr1	204448770	204449195	3.426	4.62E-05	gained-open	PIK3C2B	15081
chr18	39520338	39521133	-3.366	3.61E-09	gained-closed	PIK3C3	-14860
chr18	39518305	39518950	-4.536	3.61E-09	gained-closed	PIK3C3	-16893
chr18	39341281	39342375	-4.435	6.82E-12	gained-closed	PIK3C3	-193917
chr3	178913656	178914409	-3.491	1.95E-06	gained-closed	PIK3CA	47346
chr3	178849777	178850729	4.251	2.86E-11	gained-open	PIK3CA	-16533
chr3	138405418	138406612	-3.753	1.76E-13	gained-closed	PIK3CB	72782
chr3	138412897	138413799	-4.684	2.16E-14	gained-closed	PIK3CB	65303
chr3	138493178	138494459	-2.876	2.44E-06	gained-closed	PIK3CB	-14978
chr1	9779070	9779668	-4.48	1.74E-10	gained-closed	PIK3CD-AS1	-64427
chr7	106489190	106490426	-5.696	5.49E-14	gained-closed	PIK3CG	-16532
chr7	106480661	106481221	-4.929	2.15E-08	gained-closed	PIK3CG	-25061
chr7	106410433	106411099	-6.698	2.19E-11	gained-closed	PIK3CG	-95289
chr22	31649934	31650622	4.197	1.07E-09	gained-open	PIK3IP1	38585
chr5	67889118	67889683	2.284	3.50E-03	gained-open	PIK3R1	377535
chr5	67745940	67746883	-2.653	7.98E-05	gained-closed	PIK3R1	234357
chr5	67729660	67730793	-5.871	6.65E-19	gained-closed	PIK3R1	218077
chr5	67639201	67639758	-5.846	2.15E-09	gained-closed	PIK3R1	127618
chr5	67342126	67343346	-5.637	4.31E-17	gained-closed	PIK3R1	-169457
chr5	67278971	67279673	-4.801	8.04E-10	gained-closed	PIK3R1	-232612
chr3	130377197	130378042	-4.218	2.18E-09	gained-closed	PIK3R4	88498
chr17	8825624	8826551	-7.568	2.60E-16	gained-closed	PIK3R5	43404
chr17	8818757	8820103	-4.53	2.01E-12	gained-closed	PIK3R6	-47764

GO ID	GO Term	Definition	Ontology ▲	Pvalue	GeneSym...
GO:0014909	smooth muscle cell migration	The orderly mo...	BP	2.16E-04	AAMP; IQ...
GO:0030260	entry into host cell	The invasion by...	BP	2.80E-04	TRIM25; ...
GO:0051806	entry into cell of other organism in...	The invasion by...	BP	2.80E-04	CTSB; W...
GO:0001706	endoderm formation	The formation ...	BP	1.73E-04	DUSP1; ...
GO:0001704	formation of primary germ layer	The formation ...	BP	7.09E-04	EPB41L5;...
GO:0043647	inositol phosphate metabolic proc...	The chemical r...	BP	3.09E-04	ITPKA; IS...
GO:0019751	polyol metabolic process	The chemical r...	BP	2.86E-04	ITPKA; G...
GO:0044409	entry into host	Penetration by ...	BP	2.80E-04	CD55; C...
GO:0051828	entry into other organism involved...	Penetration by ...	BP	2.80E-04	CCR5; T...
GO:0022617	extracellular matrix disassembly	A process that ...	BP	8.56E-04	FAP; CTS...
GO:0043062	extracellular structure organization	A process that i...	BP	1.14E-04	ELN; FN1...
GO:0030198	extracellular matrix organization	A process that i...	BP	1.01E-04	SERPINB5...
GO:0010466	negative regulation of peptidase a...	Any process th...	BP	5.69E-04	PAPLN; K...
GO:0014910	regulation of smooth muscle cell ...	Any process th...	BP	5.68E-04	IGFBP5; ...
GO:0052547	regulation of peptidase activity	Any process th...	BP	7.55E-04	USP14; N...
GO:1902041	regulation of extrinsic apoptotic si...	Any process th...	BP	2.41E-04	SFRP2; T...
GO:2001236	regulation of extrinsic apoptotic si...	Any process th...	BP	2.86E-04	TRAF1; ...
GO:0042127	regulation of cell proliferation	Any process th...	BP	7.89E-04	E2F3; LE...
GO:0001952	regulation of cell-matrix adhesion	Any process th...	BP	7.39E-04	SMAD3; ...
GO:0051674	localization of cell	Any process in ...	BP	7.60E-04	GLI3; CU...
GO:0048870	cell motility	Any process inv...	BP	7.60E-04	RERE; IT...
GO:0005925	focal adhesion	Small region on...	CC	4.53E-04	LASP1; LI...
GO:0031012	extracellular matrix	A structure lyin...	CC	5.46E-04	ELN; EPY...
GO:0045121	membrane raft	Any of the smal...	CC	2.30E-04	FADD; C...
GO:0005924	cell-substrate adherens junction	An adherens ju...	CC	2.57E-04	RHOB; L1...

Pathway ▲	Pvalue	Kegg ID	GeneSymbol
Adipocytokine signaling pathway	3.83E-02	hsa04920	MAPK8; SOCS3; IRS2; ADIPOQ; AC...
Apoptosis	2.96E-02	hsa04210	PRKAR2B; NTRK1; TNFRSF10B; PP...
Calcium signaling pathway	4.11E-02	hsa04020	PTK2B; ADRB1; ADRA1B; CALM2; ...
Inositol phosphate metabolism	8.29E-04	hsa00562	PIK3CA; PLCG2; ITPK1; PIP4K2A; ...
Neuroactive ligand-receptor interaction	4.15E-02	hsa04080	TSHR; GLRA3; GABRG2; LPAR1; A...
p53 signaling pathway	4.76E-03	hsa04115	GADD45A; CCNG2; CDK4; PIDD1;...
Pathways in cancer	1.52E-02	hsa05200	FGF1; CTNNB1; FASLG; FOS; MMP1...
Phosphatidylinositol signaling system	2.84E-04	hsa04070	PLCZ1; PIK3CA; ITPKB; PLCG2; PL...
Phototransduction	6.49E-03	hsa04744	GUCA1B; CALM2; GNGT1; ARRB1;...
PPAR signaling pathway	1.45E-02	hsa03320	ACOX2; ACSL3; DBI; CPT1A; ANGP...
Type II diabetes mellitus	8.57E-03	hsa04930	SOCS2; CACNA1E; SLC2A4; IKBKB;...
Valine, leucine and isoleucine degradation	4.11E-02	hsa00280	EHHADH; PCCB; AUH; MCCC2; AC...

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-eJTNBDdKgJgJzaQcY6OXmsXAHU

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