GUAVA Manual

Mayur Divate and Edwin Cheung

Index

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

In nutshell, GUAVA is a standalone GUI tool for processing, analyzing and visualizing ATAC-seq data. A user can start GUAVA analysis with raw reads to identify ATAC-seq signals. Then ATAC-seq signals from two or more samples can be compared using GUAVA to identify genomic loci with differentially enriched ATAC-seq signals. Furthermore, GUAVA also provides gene ontology and pathways enrichment analysis. Since to use GUAVA requires only several clicks and no learning curve, it will help novice bioinformatics researchers and biologist with minimal computer skills to analyze ATAC-seq data. Therefore, we believe that GUAVA is a powerful and time saving tool for ATAC-seq data analysis. GUAVA setup contains a script to configure and install dependencies which facilitates the GUAVA installation. GUAVA works on Linux and Mac OS.

This document contains all the information that is required to install and use GUAVA.

GUAVA is developed in the Edwin's laboratory at University of Macau.

Quick Start

Download Software

The GUAVA tool is provided as a Java jar file. You can download the latest release as a zipped GUAVA package from project page on GitHub. The file name will be "GUAVA-master.zip". And the source code is available at project source code page on GitHub.

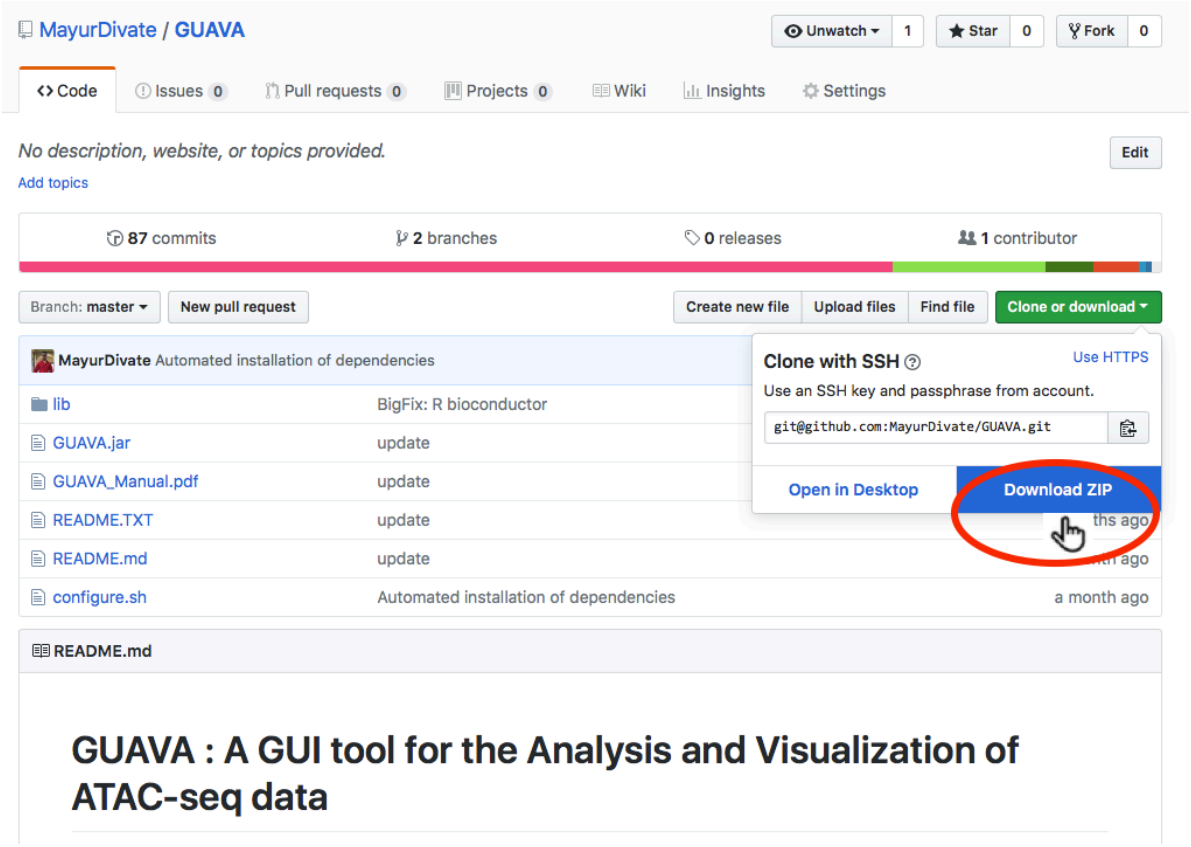

Figure1: GUAVA - GitHub repository

Go to GUAVA project page on GitHub. Click on the 'clone or download' to view option for downloading GUAVA package ZIP file.

Install

Open the downloaded GUAVA package and place the folder containing the jar file in a home directory / folder on your hard drive. It can be placed in any desired folder. But later in this tutorial it is assumed that it is in home directory. It can be achieved by the use following command.

```
cp  /path/to/GUAVA-master.zip ~/
```

Once package is copied to home directory, use command below to unzip and rename it.

```
cd  ~/
unzip GUAVA-master.zip
```

```
mv GUAVA-master GUAVA
```

Install dependencies

GUAVA depends on other tools in order to process ATAC-seq data (*e.g.* bowtie for alignment). If any of the dependency is not found on system, GUAVA will fail to start. To help users, we have written a program (configure.sh) which automatically downloads and installs dependencies. After launching the terminal, users can simply type or copy following command to complete the installation of dependencies.

```
cd ~/GUAVA
sh ./configure.sh
```

Use GUAVA

Once dependencies are configured, the user can use following command to open GUAVA graphical user interface.

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

Graphical user interface of GUAVA

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.

GUAVA tool has two main programs

1) ATAC-seq data analysis: to process raw ATAC-seq sequencing reads

2) ATAC-seq differential analysis: to compare ATAC-seq signals.

When GUAVA GUI is evoked it open GUAVA home window (Figure 2A). Here you can choose between above two programs. Then based on the selection of program, the desire input window will be opened (Figure 2 B and C).

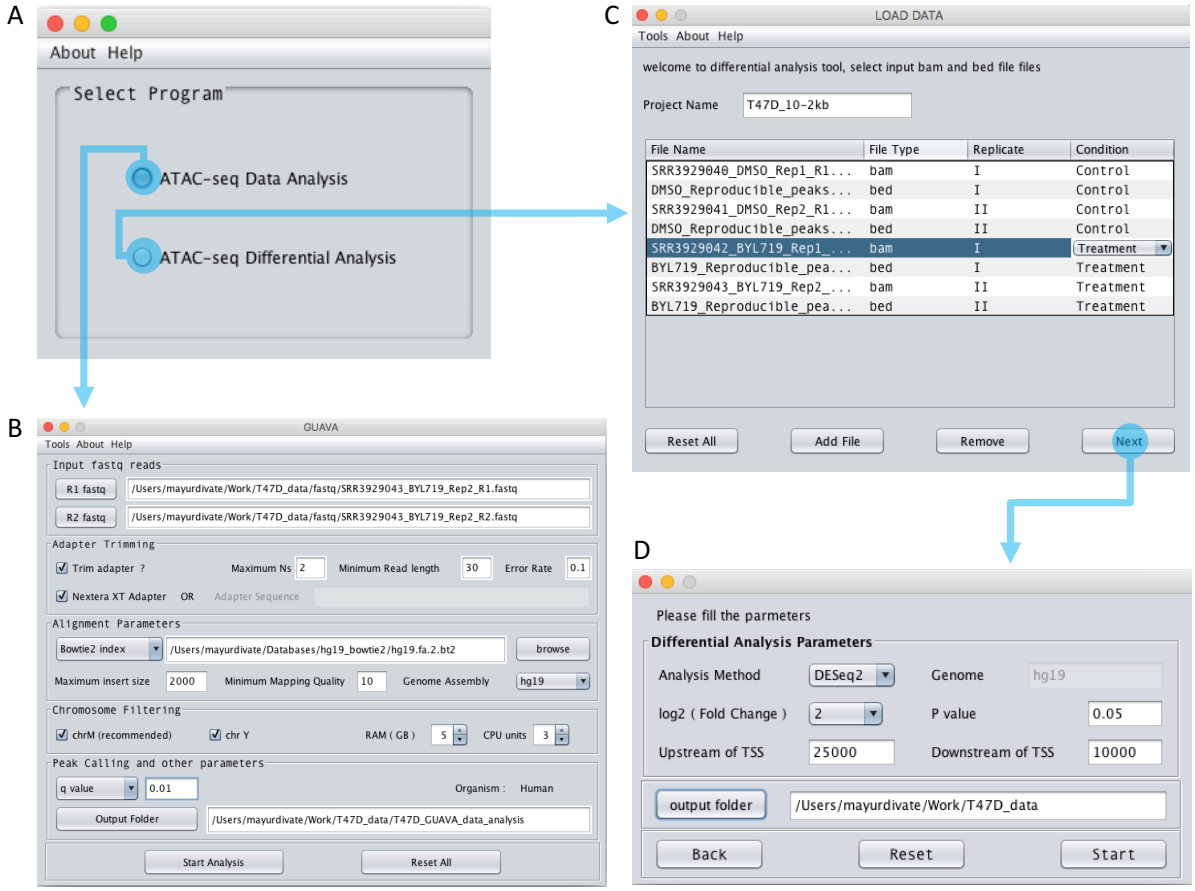

Figure 2. Design of GUAVA Graphical user interface

(A) GUAVA home windows: allows user to choose between available GUAVA program. Once the user has chosen desired program, it opens the input interface for that program. Using input interface user can upload input files such as fastq, bam etc. and set parameters (B) Input window interface of ATAC-seq data analysis program and (C and D) ATAC-seq differential analysis.

ATAC-seq data analysis program

This program accepts raw ATAC-seq reads as an input. Before aligning reads to genome, it trims adapter sequence from reads using cutadapt only if trimming option is selected. After that it filters unsuitable reads for ATAC-seq analysis such as duplicate reads. Next, it uses MACS2 to identify ATAC-seq peaks. Finally, it performs functional annotation on the ATAC-seq peaks.

Parameters

R1 fastq: button to select and upload R1 fastq file ATAC-seq reads

R2 fastq: button to select and upload R2 fastq file ATAC-seq reads

Trim adapter: check this option if reads contains adapter

Maximum Ns: if one of the read in pair contains more than specified number Ns after adapter trimming, that read pair will be discarded (default 2)

Minimum read length: if one of the read in pair is shorter than specified length after adapter trimming, that read pair will be discarded (default 30)

Error Rate: allowed number of mismatches as a fraction of adapter sequence length. For example, if error rate is 0.1 then 1 mismatch is allowed for 10bp match of adapter sequence (default 0.1)

Nextera XT adapter: you can select this option if adapter used for ATAC-seq is Nextera XT adapter (default true)

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

Bowtie V1 or Bowtie V2 index: If you want to use bowtie for read mapping select "Bowtie index" from drop down menu else select "Bowtie2 index" to use bowtie2. Then using browse button upload appropriate genome index (bowtie or bowtie2 index). Please see section 'how to create genome index' to know more about genome index. (default bowtie)

Maximum insert size: Maximum insert size in base pair allowed for paired end alignment (default 2000)

Maximum genomic hits or Mapping quality: Maximum genomic hit (bowtie) and Minimum Mapping quality (bowtie2) to discard reads pairs which has multiple alignments (default Maximum genomic hits =1 and Mapping quality >= 10)

Genome assembly: select the correct genome build from drop down menu e.g. hg19 and same build will be used for peak annotation and functional analysis.

ChrM: if selected, reads aligning to mitochondrial chromosome will be discarded (default true)

ChrY: if selected, reads aligning to chromosome Y will be discarded. (default false)

RAM: RAM in GB to be used by GUAVA (default 1)

CPU units: number of CPU units to be used by GUAVA (default 1)

p or q value: select appropriate value from drop down menu and specify the cut off value in box next to it. This will be used by MACS2 to filter peaks (default q value)

Output folder: select folder to save GUAVA ATAC-seq data analysis results

Reset All: button to set all parameters to default value. select folder to save GUAVA ATAC-seq data analysis results

Start Analysis: click this button to start 'ATAC-seq data analysis' program. If all provided options are valid then GUAVA will start analysis.

Output interface for GUAVA ATAC-seq data analysis

Once GUAVA finishes analysis it shows results on tabular output interface (figure 3). Also facilitates the visualization of ATAC-seq signal on IGV browser.

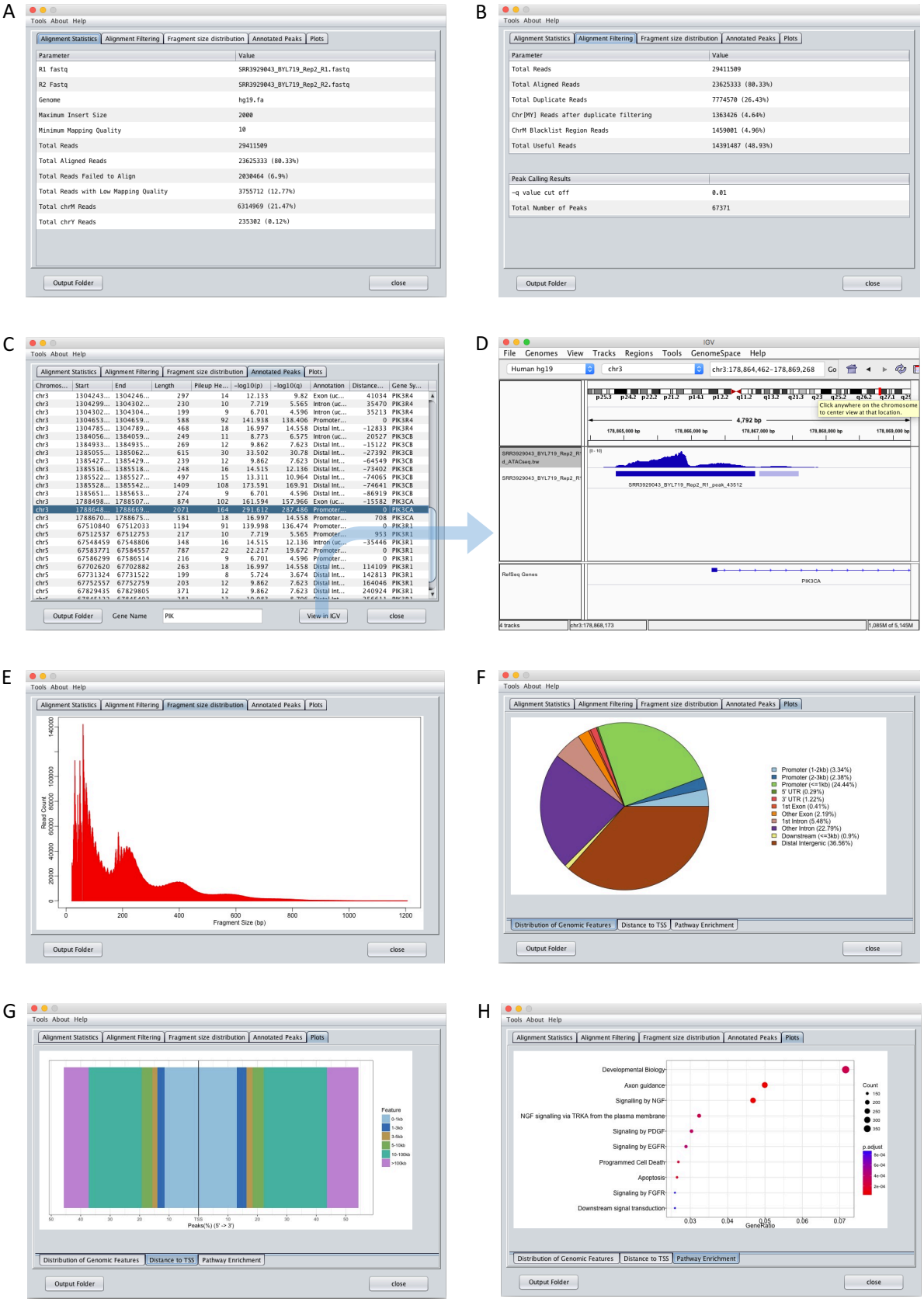

Figure 3: Output interface for GUAVA ATAC-seq data analysis. A) Input summary and alignment statistics. B) Read filtering and peak calling summary. C) Peak annotation table with sorting and filtering functionality. Easy access to IGV for visualizing peaks and automatically generated normalized ATAC-seq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

The output interface of 'ATAC-seq data analysis' program has following five tabs.

1) 'Alignment statistics' tab: This tab provides reads mapping statistics (e.g. total number of reads mapped to genome along) with summary of input files and parameters (Figure 3A).
2) 'Alignment Filtering' tab: It has two tables (Figure 3B). One is to provide figures for various types reads (e.g. useful reads, which are nothing but reads that have passed all the filtering criteria and eligible for the downstream analysis). On the other hand, second table shows summary of MACS2 peak calling.
3) 'Annotated Peaks' tab: This tab provides complete list of ATAC-seq peaks along with annotations such as distance from nearest gene, gene symbol of nearest gene and overlapping genomic feature e.g exon, intron etc. (Figure 3C). The search box is provided at bottom can be used to search peaks (Figure 3C). To view only the list of peaks annotated with a particular gene, type the symbol of that gene in the search box. To visualize peak in the IGV browser, select a peak and then click on the 'view in IGV' button at the next search box. This will open a new IGV browser instance and ATAC-seq signals will be loaded automatically on the browser (Figure 3D).
4) 'Fragment size distribution' tab: This tab displays the fragment size distribution plot for a given ATAC-seq sample (Figure 3E).
5) 'Plots' tab: It has three sub tabs one for the pie chart showing distribution of peaks in the several genomic features (Figure 3H), another for plot that shows proportion of the peaks upstream and downstream of the TSS of the nearest gene (Figure 3F), and the last tab provides top enriched pathways (Figure 3G). Furthermore, these results are stored in the output folder, click the 'output folder' button at the bottom-right to open the output folder.

ATAC-seq differential analysis program

This program compares ATAC-seq signals from two conditions and returns the differentially enriched signals. Additionally, it provides the peak annotation and functional analysis for differentially enriched peaks. There are two input windows for this program. First window is to upload the ATAC-seq signals from different and conditions and replicates (Figure 1C). Use 'add file' and 'remove' buttons to add and delete input files respectively. Once you have uploaded bed file containing ATAC-seq peaks and bam files, specify the condition and replicate number for each file and click 'Next'. Second window allows you to specify differential analysis related parameters e.g. fold change (Figure 1B). Once you have added all the required files and parameters click 'Start' button to run differential analysis.

Parameters

Analysis method: currently we have only implemented DESeq2.

log2 (Fold Change): log2 fold change cut off to define differentially enriched peaks. Default 2.

P value: P value cut off to select most significant differentially enriched peaks. Default 0.05.

Upstream of TSS: if the peak is present within a specified distance (in base pair) from the TSS of a gene, to the upstream. Then that gene will be associated with the peak for functional analysis. Default 25000.

Downstream of TSS: if the peak is present within a specified distance (in base pair) from the TSS of a gene, to the downstream. Then that gene will be associated with the peak for functional analysis. Default 10000.

Output folder: select folder to save GUAVA differential analysis results.

Output interface for GUAVA ATAC-seq differential analysis

The output interface of 'ATAC-seq differential analysis' program is also tabular like 'ATAC-seq data analysis' program.

1) 'Summary' tab: This tab provides summary of input parameters e.g. fold change cut-off, list of input files used for differential analysis (Figure 4A) etc.
2) 'Differential Table' tab: This provides the list of differentially enriched ATAC-seq signals with annotation such as nearest gene to peak and the distance between them (Figure 4B) etc. Same as output interface of 'ATAC-seq data analysis' program, there is a search box and 'view in IGV' button at bottom of window. Which can be used to sort peaks by gene symbol and view peaks in IGV from input samples, respectively (Figure 4D).
3) 'Plot' tab: This provides volcano plot of differentially enriched peaks (Figure 4C).
4) 'Go Analysis' and 5) 'Pathway Analysis' tabs: These tabs provide results of functional analysis i.e. enriched gene ontologies (Figure 4E) and pathways (Figure 4F) respectively.

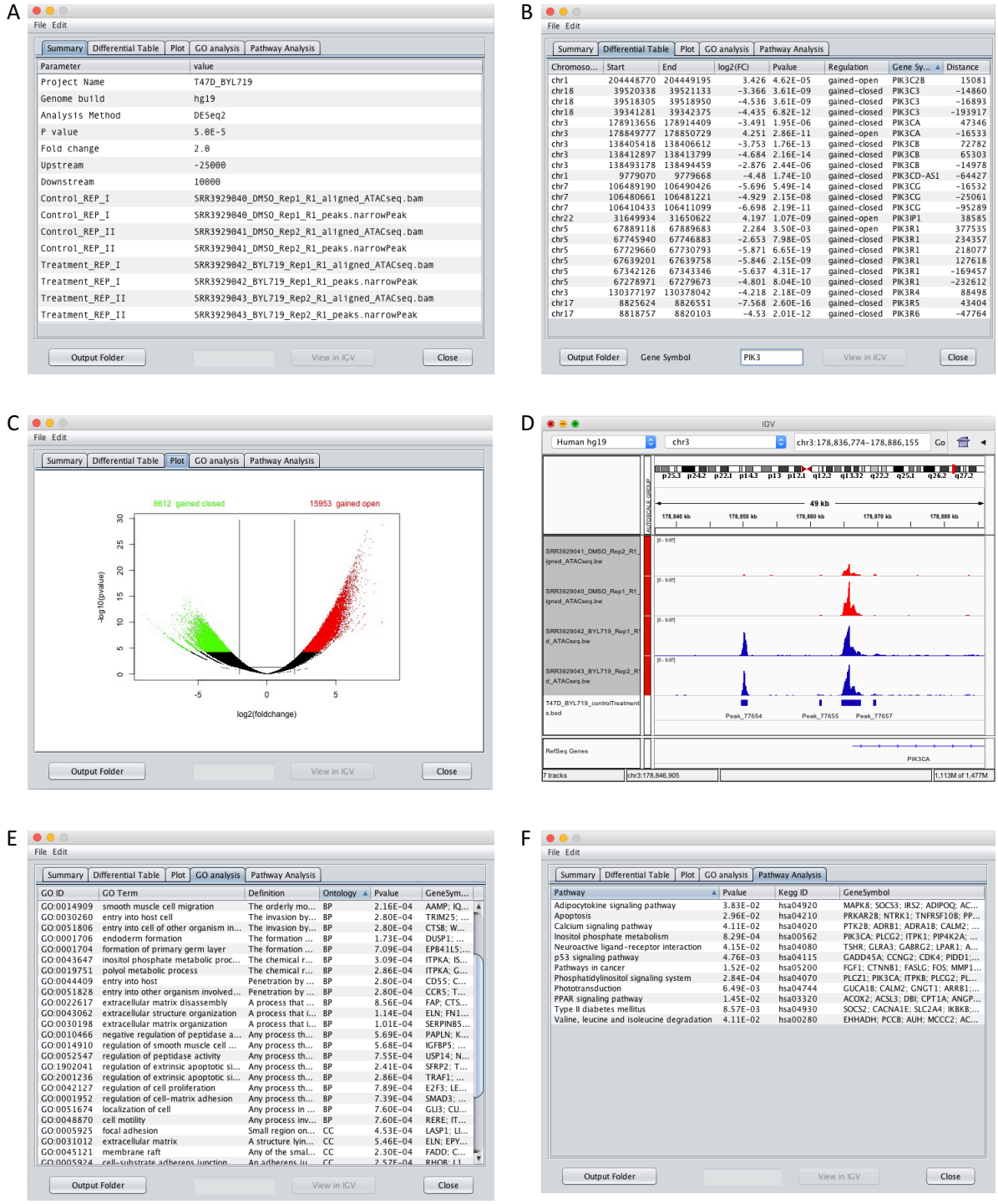

Figure 4: Output interface for GUAVA ATAC-seq differential analysis. A) Input summary. B) Differentially enriched peaks with sorting and filtering functionality. Easy access to IGV to visualize differentially enriched peaks and normalized ATAC-seq signals from each sample. C) Volcano plot indicating 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.

Getting help and reporting issues

If user seeks help any issue that is not covered in this manual, he can first search on GitHub for help (https://github.com/MayurDivate/GUAVASourceCode/issues). If he

does not find it on GitHub he can start new issue. Similarly, user can find any bug, he can also report that on the same page.

Download genome fasta file

Fasta is a text file format for representing nucleotide or protein sequences. Genome fasta file is a fasta file which contains the nucleotide sequences from all of the chromosomes of a particular organism. Genome fasta file is a required for read mapping using any aligner tool. Those users who don't know where they can find genome fasta file please follow the links given below and subsequent instructions,

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

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

Then, click on the 'full data set'. This will open a new page, scroll down and click on the chromFa.tar.gz to download genome sequence.

Use following command to extract chromosome files and merge them into one file.

```
tar -zxvf -d /path/to/chromFa.tar.gz
cat chromFa/*fa > GenomeBuild.fasta
```

How to create a bowtie index of genome fasta file

It is true that the genome fasta file is required for the alignment. But the aligners use special set of files called as genome index, generated using from genome fasta. Index files are used to speed up the read mapping process so that the aligner can map millions of reads within few hours of time. Therefore, you need create genome index file before read mapping. Remember that the index format is different for each aligner. Please refer to 'Download genome fasta file' section to find more information about downloading genome fasta file. If you already have a genome fasta file, follow the commands below to create a bowtie genome index.

To create bowtie index:

```
bowtie-build /path/to/GenomeBuild.fa GenomeBuild.fa
```

To create bowtie2 index:

```
bowtie2-build /path/to/GenomeBuild.fa GenomeBuild.fa
```

For example, suppose the genome fasta file is in 'genomes' directory and which is sub directory of 'database' directory under the home directory then command to create bowtie index will be as follows

```
bowtie-build ~/database/genomes/hg19.fa hg19.fa
```

Note: This is a time-consuming step.