GUAVA Demo

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.

Installation

To install and configure GUAVA please refer to the GUAVA manual,

Link: https://github.com/MayurDivate/GUAVA/blob/master/GUAVA_Manual.pdf

How to download sample data?

We demonstrate how to use the GUAVA graphical user interface and show typical results that are obtained from the program by using sample dataset.

To download the **"sample data and sample genome"** please follow the following link,

http://ec2-52-201-246-161.compute-1.amazonaws.com/guava/

Decompress the SampleData.tar.gz (sample data) and Hs_demo.fasta.gz (sample genome) files by using following commands.

```
cd /path/to/sampledata/
tar -zxvf SampleData.tar.gz
gzip -d Hs_demo.fasta.gz
```

Create bowtie2 genome index using Hs_demo.fasta file. Find more about how to create genome index in the manual.

Basic Linux Commands

(source: https://diyhacking.com/linux-commands-for-beginners/)

cd: "cd" is the command used to go to a directory. For example, if you are in the home folder, and you want to go to the Downloads folder, then you can type in "cd Downloads". Remember, this command is case sensitive and you have to type in the name of the folder exactly as it is. But there is a problem with these commands. Imagine you have a folder named "Raspberry Pi". In this case, when you type in "cd Raspberry Pi", the shell will take the second argument of the command as a different one, so you will get an error saying that the directory does not exist. Here, you can use a backward slash. That is, you can use "cd Raspberry\ Pi" in this case. Spaces are denoted like this: If you just type "cd" and press Enter, it takes you to the home directory. To go back from a folder to the folder before that, you can type "cd .." . The two dots represent back.

```
cd /path/to/destination
```

cp: The cp command is used to copy files through the command line. It takes two arguments, the first one is location of the file to be copied, the second is where to copy. For example, to copy 'demo.txt' from Downloads to Documents command will be: cp ~/Downloads/demo.txt ~/Documents/

```
cp /path/to/the/file/to/be/copied /path/to/destination/folder
```

How to start GUAVA?

Step 1) Start GUAVA

To start GUAVA, first change directory to the GUAVA package directory. For example, if GUAVA package is in the home directory then use following command.

```
cd ~/GUAVA
```

Then use following command to start GUAVA graphical user interface

```
java -jar GUAVA.jar
```

This will open the home window of GUAVA as shown below in Fig 1. This allows user to choose between GUAVA programs.

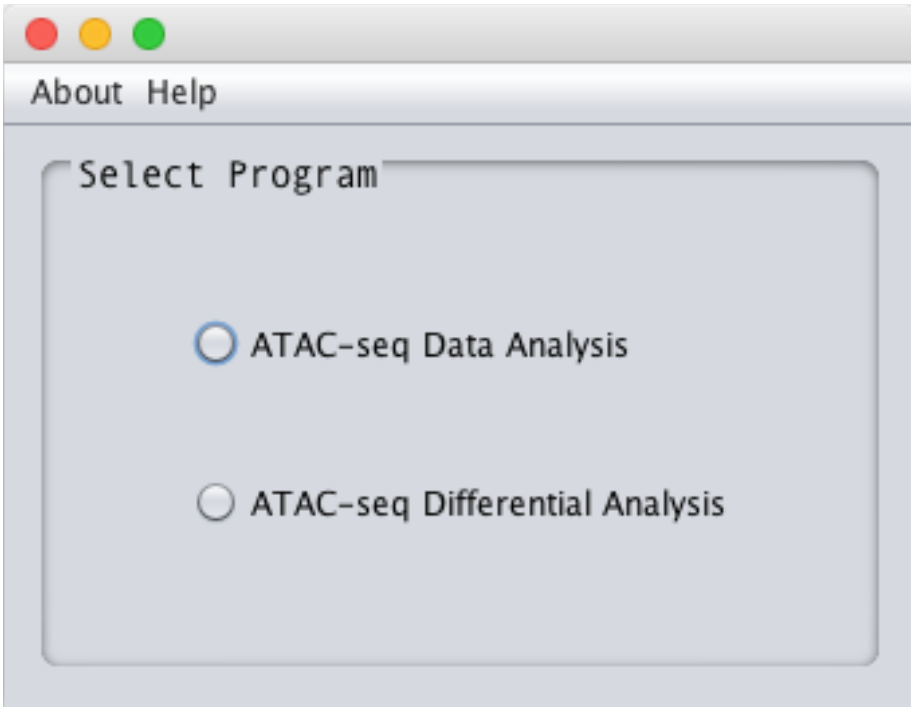


Figure 1: GUAVA home window.
Select desired GUAVA program, to open its input interface.

Step 2) Select GUAVA program

In this tutorial, first we are going to use `ATAC-seq Data Analysis` program to identify ATAC-seq signals from each sample. Then `ATAC-seq Differential Analysis program` will be used to compare these signals from different conditions. Therefore, please select the "ATAC-seq Data Analysis" option. This will open the input interface for ATAC-seq data analysis program (Fig 2).

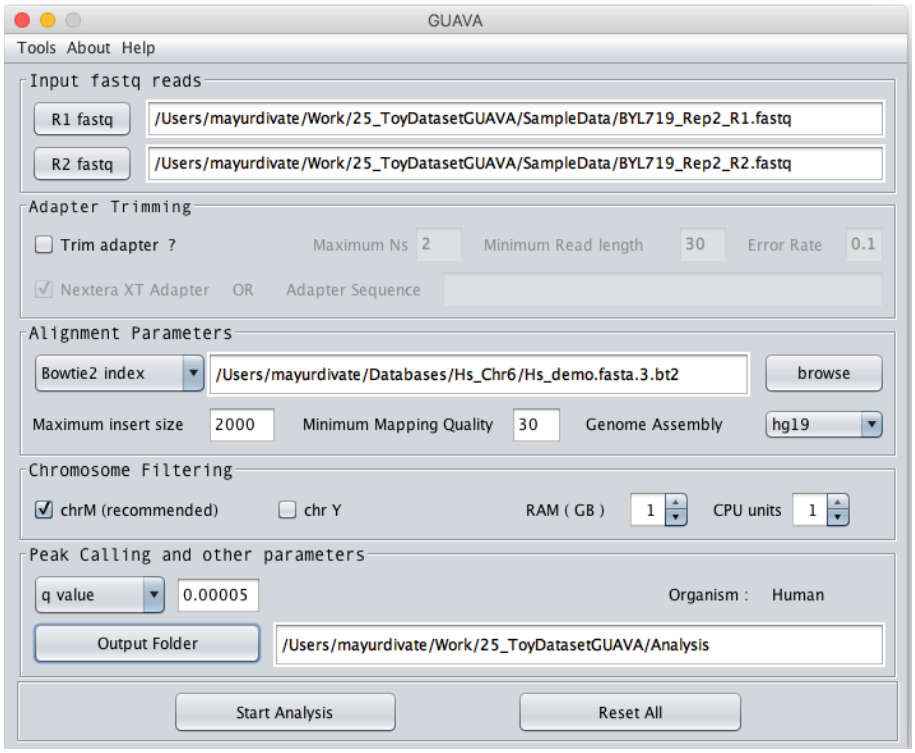


Figure 2. Input interface of GUAVA ATAC-seq data analysis
Above picture shows the input interface for ‘ATAC-seq data analysis’ program. Using input interface user can upload input files such as fastq, bam etc., set parameters and start analysis.

Running ATAC-seq Data analysis program

Step 3) Upload input FASTQ files and set input parameters for the ATAC-seq data analysis program

Follow steps 1 to 3 three to start “GUAVA ATAC-seq Data Analysis” program. Using its input interface upload input files and set parameters as described below.

- **Upload the ATAC-seq sequencing reads.**
  Click on the `**R1 fastq**` button, this will open a file browser (Fig 3). Go to the location of sample data and select fastq file and click on the `open` button. This will upload the it to the GUAVA. Similarly upload the R2 FASTQ file using `**R2 fastq**` button. Here in the example we have used BYL719_Rep2 sample (Fig 2 and 3).
- **Upload the genome index.**
  Before proceeding further make sure you have created bowtie2 genome index using Hs_demo.fasta file (refer manual). Select `Bowtie2 index` from the drop-down menu present in the `alignment parameter` section (Similarly if you intend to use bowtie for alignment, select “bowtie v1 index”). This will allow you to upload bowtie2 genome index. After that, click on the `**browse**` button, this will open a file browser. Go to the location of genome index. and select any one of the index file and click on the `open` button. This will upload the index to the GUAVA.
- Set **Minimum Mapping Quality** to **30.**
- Set **Genome Assembly** to **hg19.**
- Set **RAM** and **CPU units** as per computational power of system.
- Set **q value** to **0.00005.**
  User can choose between q or p value to filter peaks by clicking on the drop-down menu.
- Set **Output Folder**
  User can choose any desired folder to save output files of analysis.
- **Start Analysis**
  Click on the “Start Analysis” button to start processing of data.

Keep rest of the thing default and process remaining three samples in similar manner as described above. Each sample may take a while to finish. Which depends on various factors such as size of input data, RAM and CPU units used.

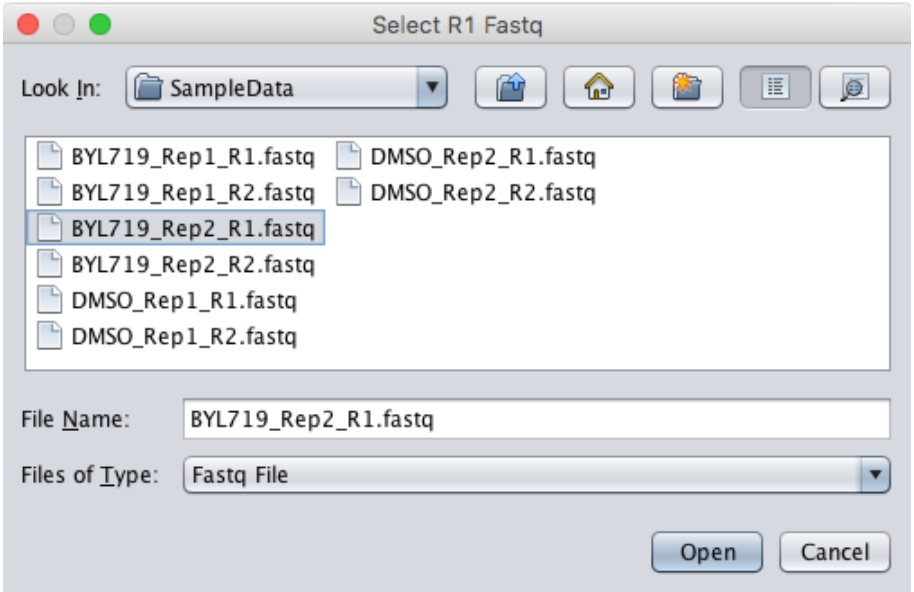


Figure 3. File browser window
This interface helps user to select and upload a desired file to the GUAVA program.

Output Interface of ATAC-seq Data Analysis Program

Step 4) Browse and visualize ATAC-seq data analysis results

There are total five tabs. User can navigate through these tabs simply by clicking on it.

Tab 1) Alignment Statistics

This tab provides selected input parameters and read mapping results e.g. alignment rate. This helps to understand quality of data. The higher the alignment rate, higher the data quality. In given example below (Fig 4), Total 1,360,988 (99.9%) were aligned to Hs_demo.fasta genome out of 1362349 input reads from BYL719 replicate 1 sample. 246 read failed to align whereas 1115 reads did not map to genome because of poor quality of alignment i.e. mapping quality less than 30. More than three thousand reads mapped to the mitochondrial chromosome (chrM).

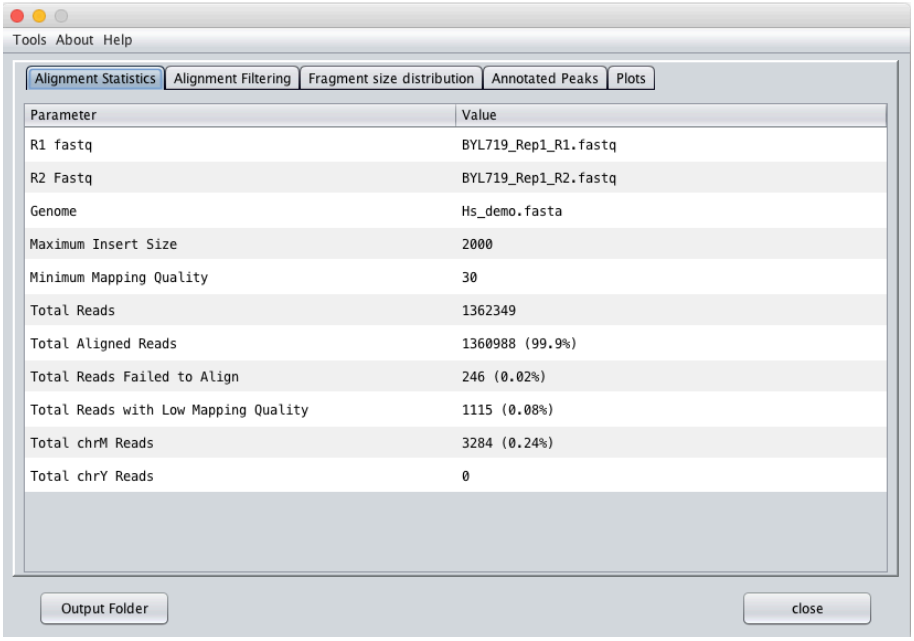


Figure 4) Alignment statistics tab of ATAC-seq data analysis output interface

Tab 2) Alignment Filtering

Tables on this tab contains results of alignment post-processing and peak calling. There were total 336342 duplicates reads. After discarding reads mapping to the chrM, this figure dropped to the 99 reads only. 0.57% of reads mapped to blacklist region of hg19 genome. There are around 74.64% useful reads after filtering duplicate reads, reads aligning to the chrM and blacklist region. These useful reads were used to call peaks by using MACS2 peak caller. Total 32652 peaks were identified with 0.00005 q value cut-off.

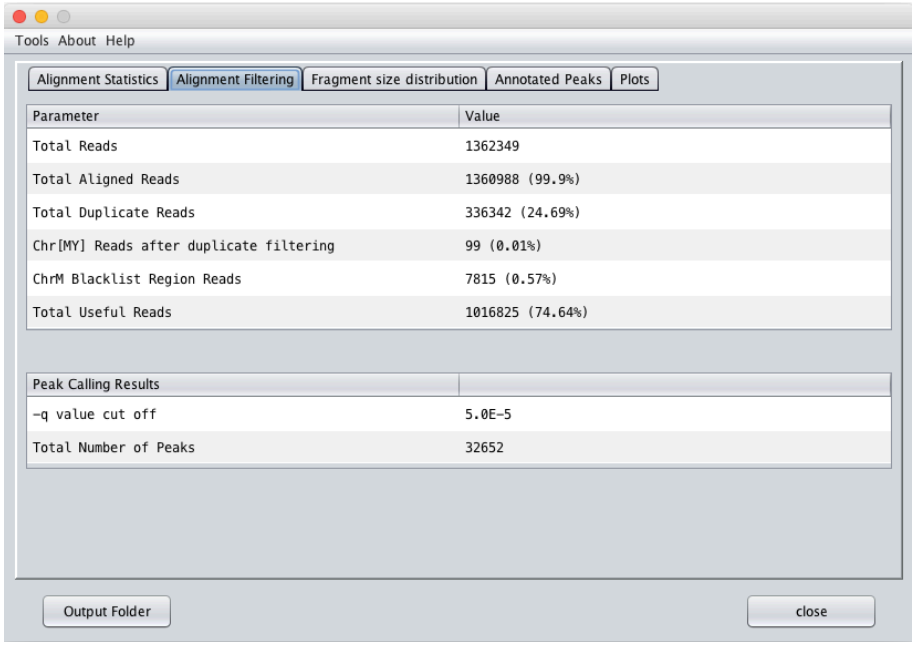


Figure 5) Alignment Filtering tab of ATAC-seq data analysis output interface

Tab 3) Fragment size distribution

Fragment size distribution graph is helps to assess the quality of ATAC-seq library. Peaks at ~200 bp intervals representing mono-, di-, and tri-nucleosomes, an indication that the ATAC-seq library is of good quality. Fragment size distribution graph for BYL719 replicate 1(Fig 6) is an example of good quality ATAC-seq library.

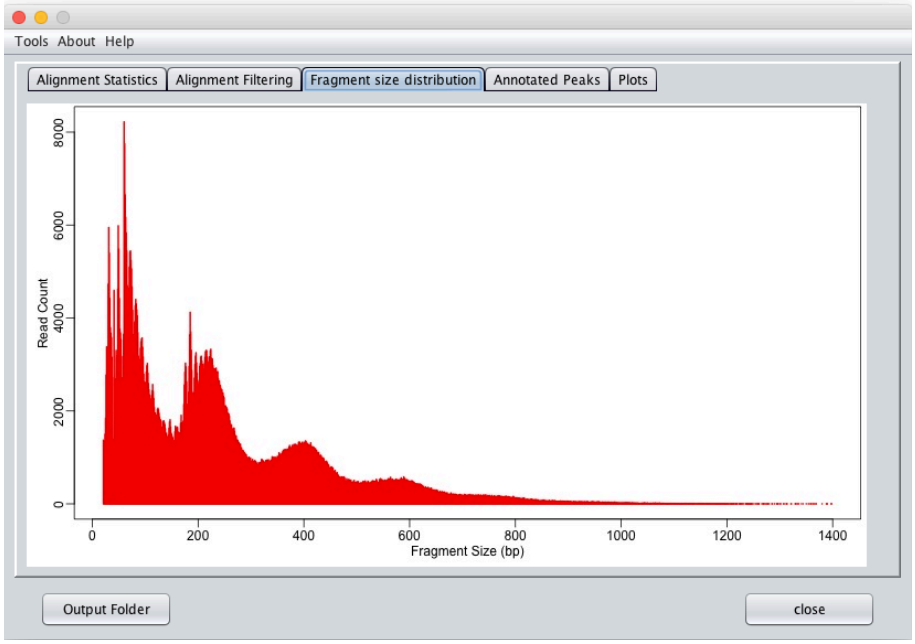


Figure 6) Fragment size distribution graph

Tab 4) Annotated Peaks

This tab provides the list of peaks with the name of nearest downstream gene, its distances from the TSS of the gene. Also, the region in which peak is falling such as promotor, UTR etc. You can search for the peak by typing gene symbol in the search box, which is provided at bottom of tab (Fig 7). A desired peak can be in visualized in the IGV browser (Fig 8). To view a peak in the IGV browser select the peak and click on the "View in IGV" button.

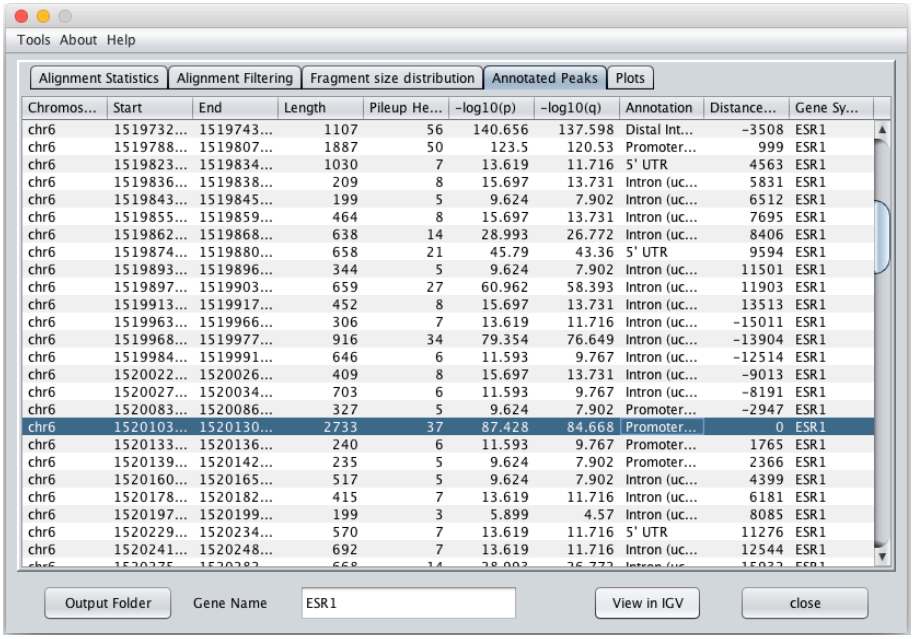


Figure 7) Annotated peaks of BYL719 replicate 1.

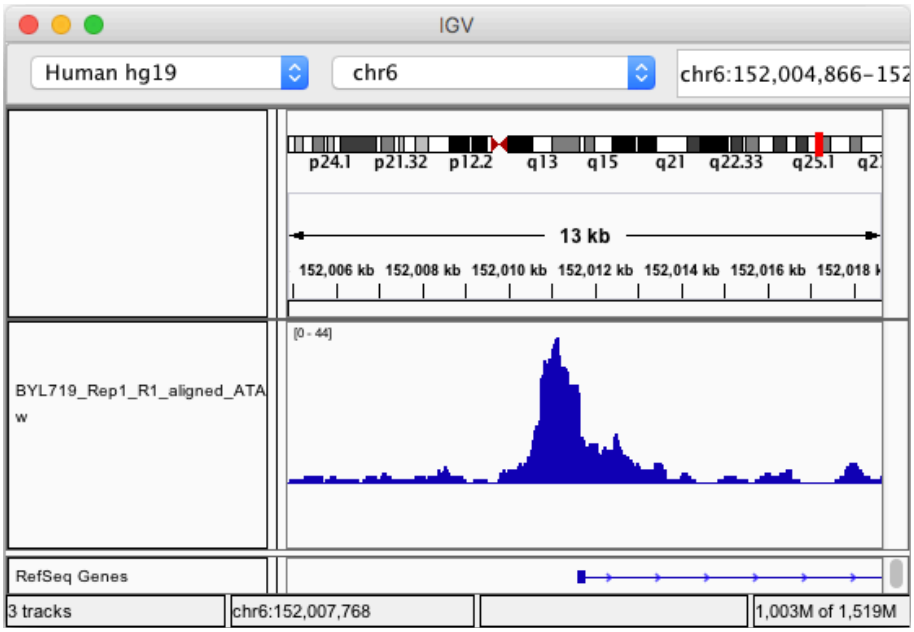


Figure 8) Visualizing ATAC-seq peak in the IGV browser

Tab 4) Plots

This tab has three sub-tabs, each of them is dedicated specific graph.

a) Distribution of Genomic features
   This graph shows the distribution of the peaks around the TSS. Different color indicates the different distance ranges from the TSS.

b) Distance to TSS
   This graph shows the distribution of the peaks in the various regions of the genome such as promotor, UTR, intron.

c) Pathway Enrichment
   This graph shows the distribution of the peaks in the various regions of the genome such as promotor, UTR, intron.

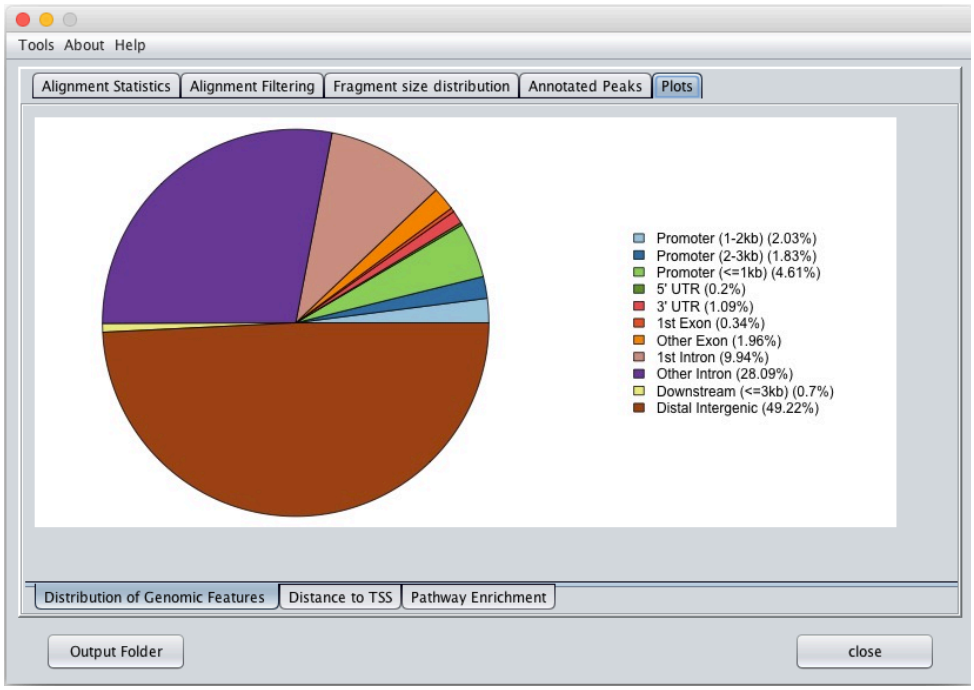


Figure 9) Distribution of Genome features.

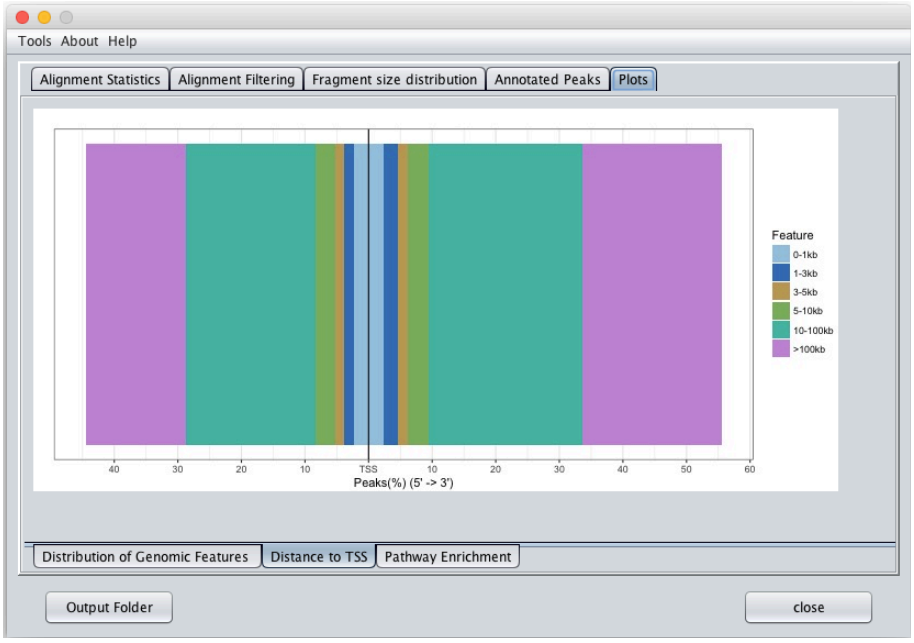


Figure 10) Distance to TSS

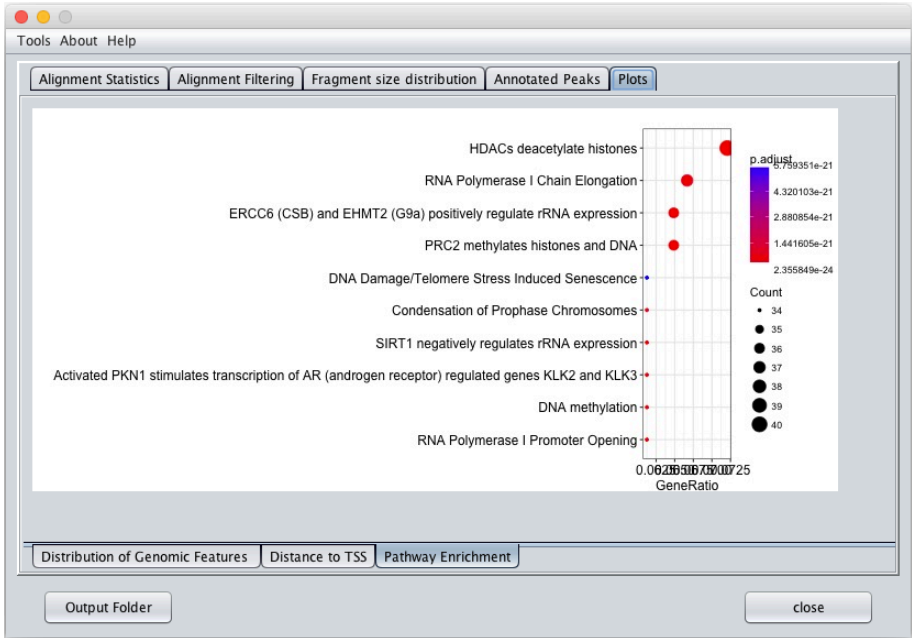


Figure 11) Top enriched pathways for BYL719 replicate 1

Step 5) Output folder

"Output Folder" button is provided at bottom input interface. To access intermediate files (bam, bed etc.) along with all the tables and plots show on the output interface. Click on the "output folder", this will open the output folder automatically.

Running ATAC-seq Differential analysis program

Step 1) Start ATAC-seq Differential analysis program

Once you have finished analysis all the samples using above procedure. Start GUAVA as described in step 1 but this time select "ATAC-seq differential analysis". This will open input interface of ATAC-seq differential analysis program (Fig 12). Here we will compare ATAC-seq signals of BYL719 treated T47D cells with DMSO treated T47D cells (control). Type "SampleData_BYL719" or any desire name in the project name text box.

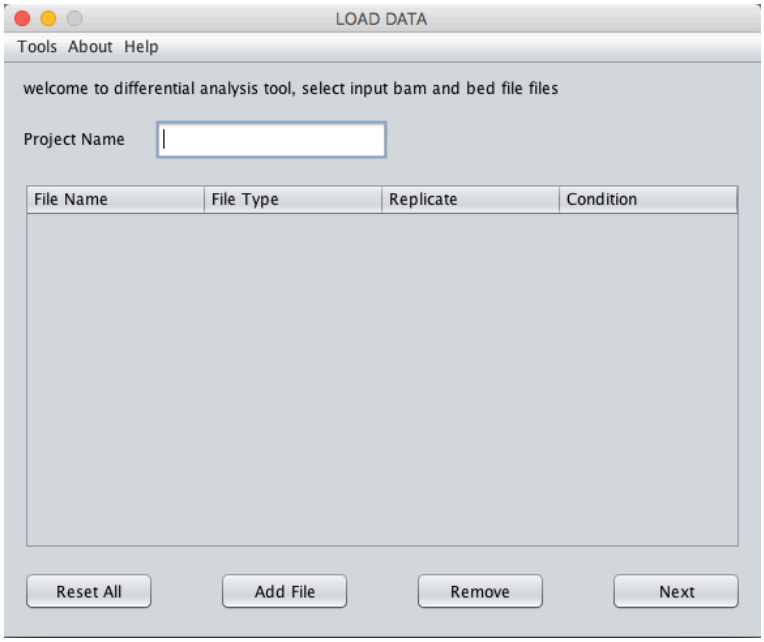


Figure 12) Input interface of ATAC-seq differential analysis program.

Step 2) Add input files

By using "Add file" button, upload the *.ATACseq.bam file from the output folder of ATAC-seq data analysis program (Fig 13). Next upload the *.narrowPeak file which can be found under subfolder *PEAK_CALLING. Upload these two files from each sample.

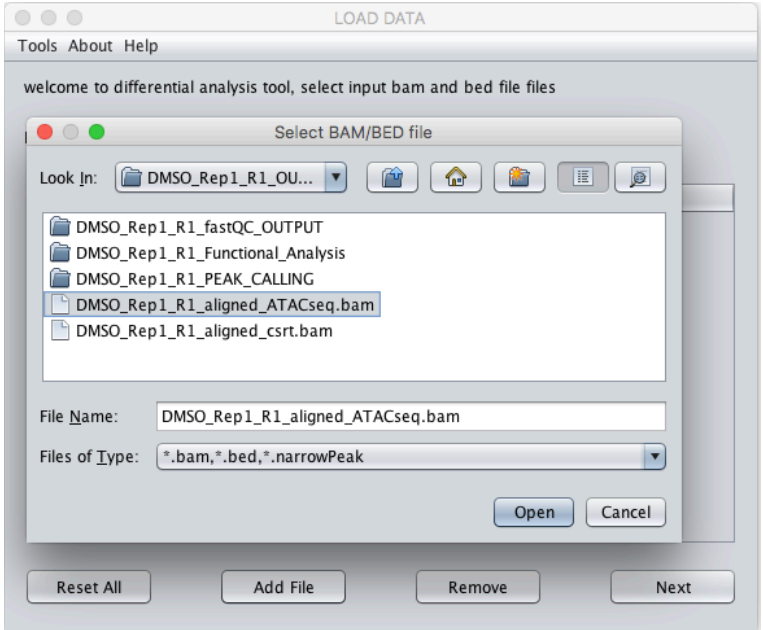


Figure 13) Adding input files ATAC-seq differential analysis program.

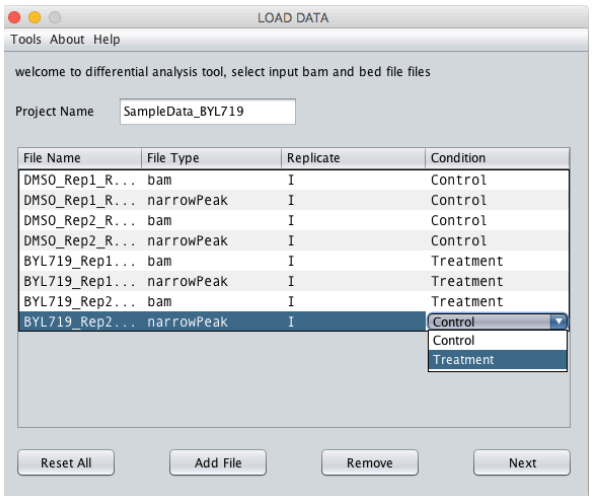
Step 3) Define replicates, control and treatment files

Each uploaded file appears in the table provided on the input interface. Program automatically detects file type. There are total four columns. The user should define replicate number and condition by using column number three and four, respectively (Fig 14).

a) Set condition "Treatment" for all BYL719 samples files.
b) Set condition "Control" for all BYL719 samples files.
c) Set replicate "I" for the all files containing "Rep1" in their name.
d) Set replicate "II" for the all files containing "Rep1" in their name.
e) Click on the "Next" button.

Remember that to proceed further user need to upload both bam and peak file from each replicate. There should be at least two replicates from each condition.

A



B

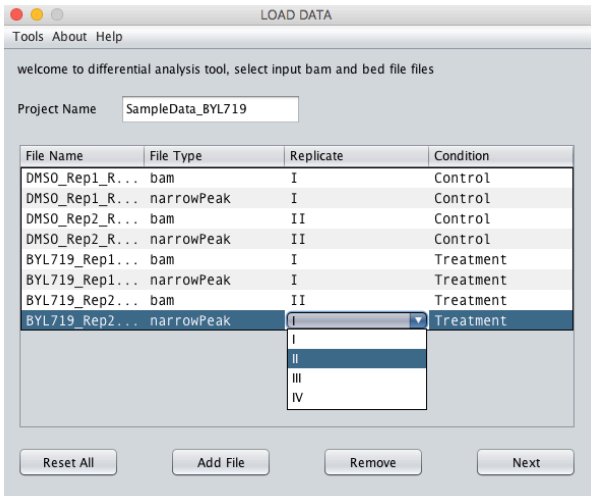


Figure 14) Define replicate number and condition

Step 4) Set Differential analysis parameters

Once you click on the "Next button" in the previous step it will open a input interface window 2 (Fig 15)

a) Set log2(fold change) to 2
b) Set "P value" to 0.005
c) Set "Upstream of TSS" to 10000
d) Set "Downstream of TSS" to 10000
e) Set "Output folder" to any desired location
f) Finally, Click on the "Start" button

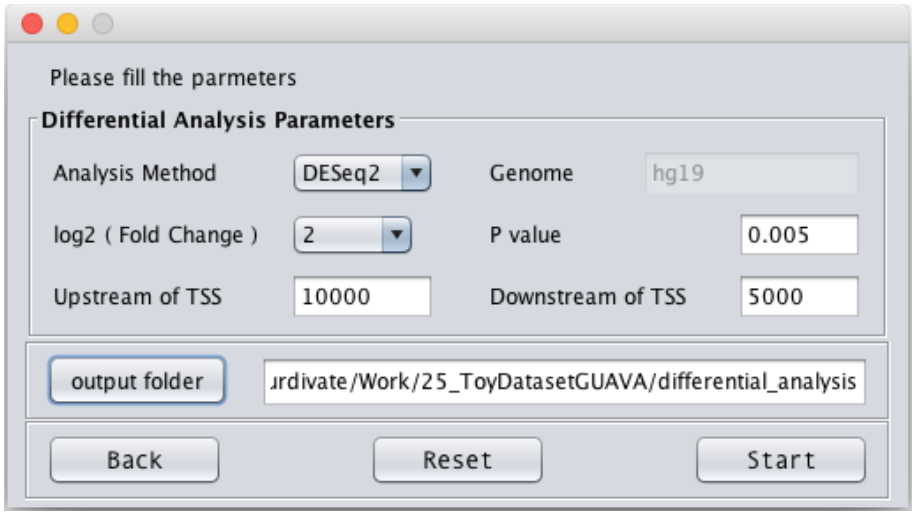


Figure 15) ATAC-seq differential analysis input interface-2

This will take some time to complete and on completion, results will be shown on tabular output interface.

Output interface of ATAC-seq Differential analysis program

Step 5) Browsing differential analysis results

The output interface of ATAC-seq differential analysis program has five tabs (Fig 16). User can browse each by simply clicking on its title.

Tab 1) Summary tab

This provides input summary including p value cut-off, fold change cut-off, input file names etc. (Fig 16).

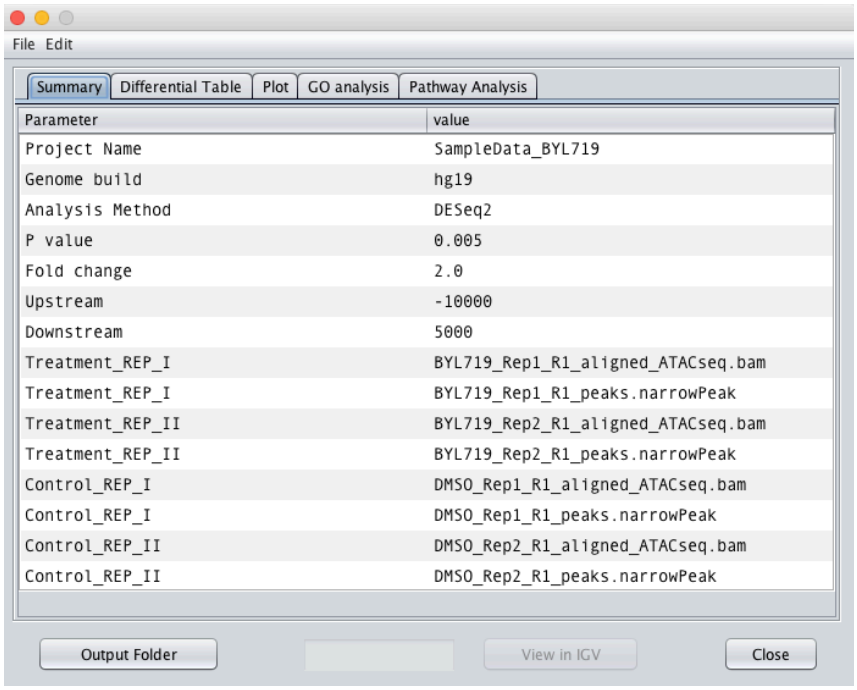


Figure 16) Summary tab gives summary of input such input files used and corresponding condition and replicate number, fold change and p value cut off

Tab 2) Differential Table tab

This provides list of differentially enriched peaks along with annotations such log2(fold change), p value, Regulation (gained open or closed), nearest gene symbol etc. (Fig 17). User can search peak using gene symbol by typing the gene symbol in search box given at bottom (Fig 17).

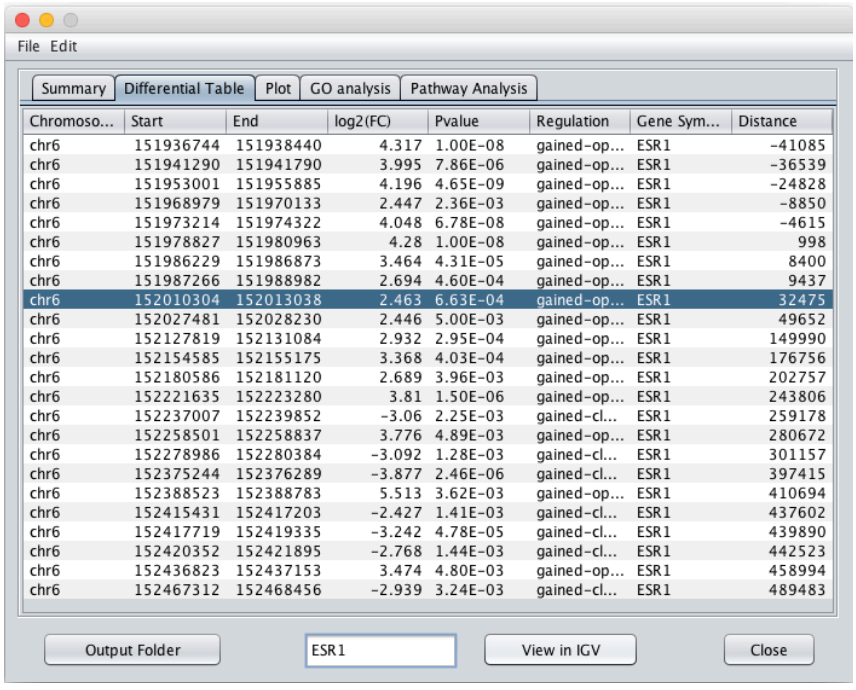


Figure 17) Differential Table tab lists the all differentially enriched peaks.

Select the peak of interest from the table and click on the "View in IGV button".

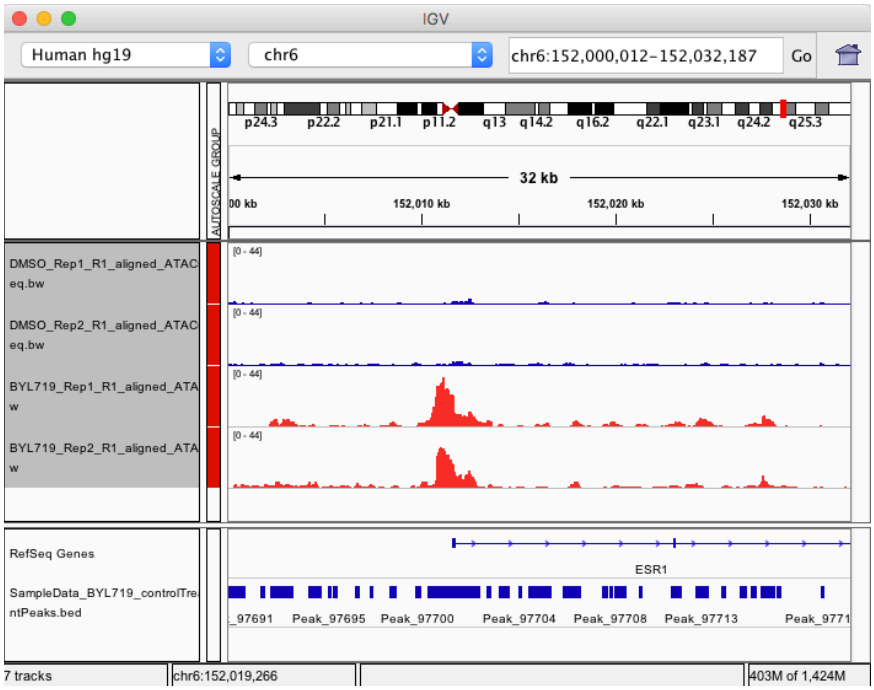


Figure 18) Visualization of differentially enriched peak in IGV.

Tab 3) Plot

This provides graphical representation of the differentially enriched ATAC-seq peaks in the form volcano plot (Fig 19).

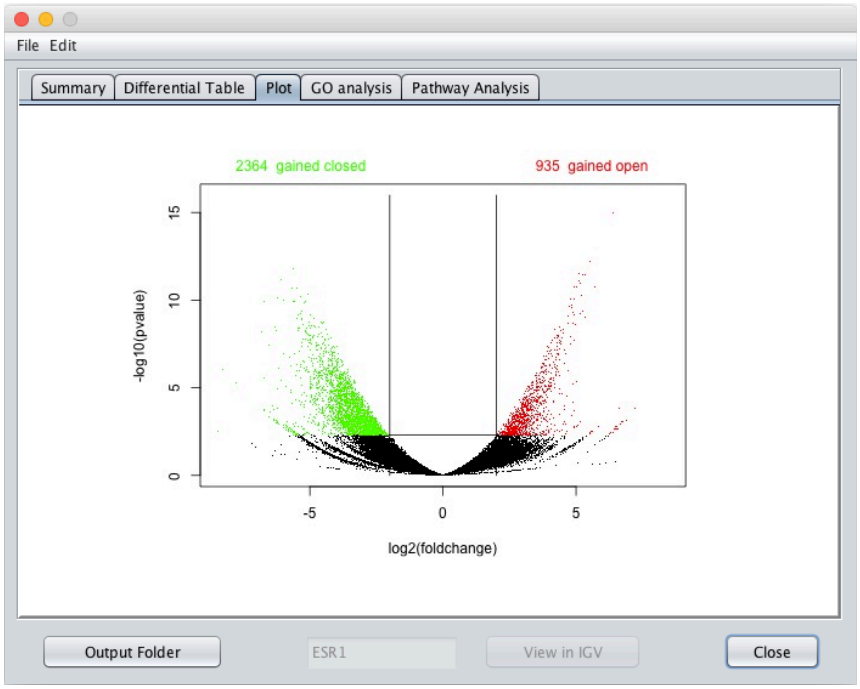


Figure 19) Volcano plot showing differentially enriched ATAC-seq peaks between BYL719 and DMSO treated cells.

Tab 4) GO analysis

This provides list of enriched gene ontologies upon comparison of ATAC-seq signals from two conditions (Fig 20).

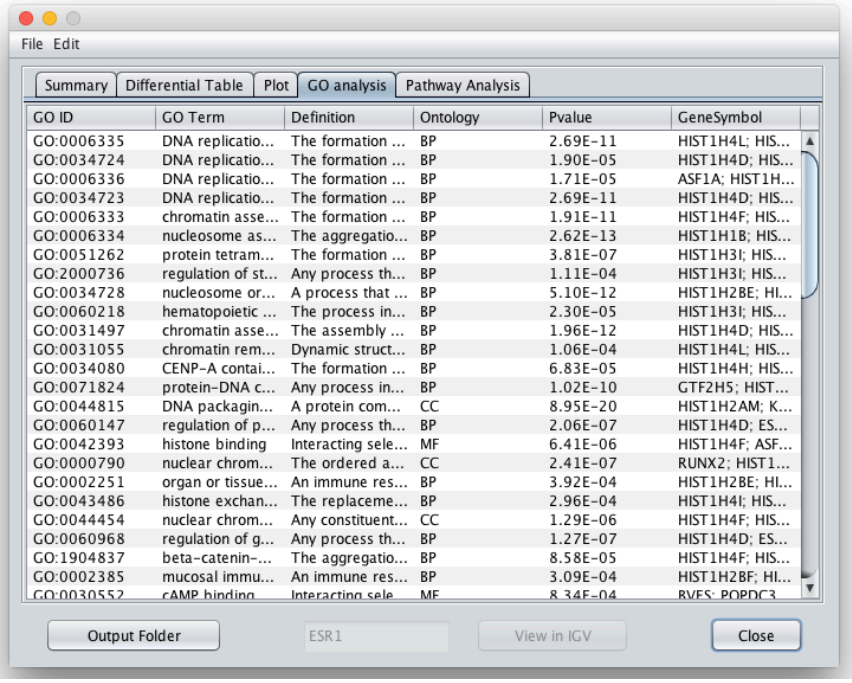


Figure 20) List of enriched gene ontology terms.

Tab 5) Pathway analysis

This provides list of enriched pathways upon comparison of ATAC-seq signals from two conditions (Fig 21).

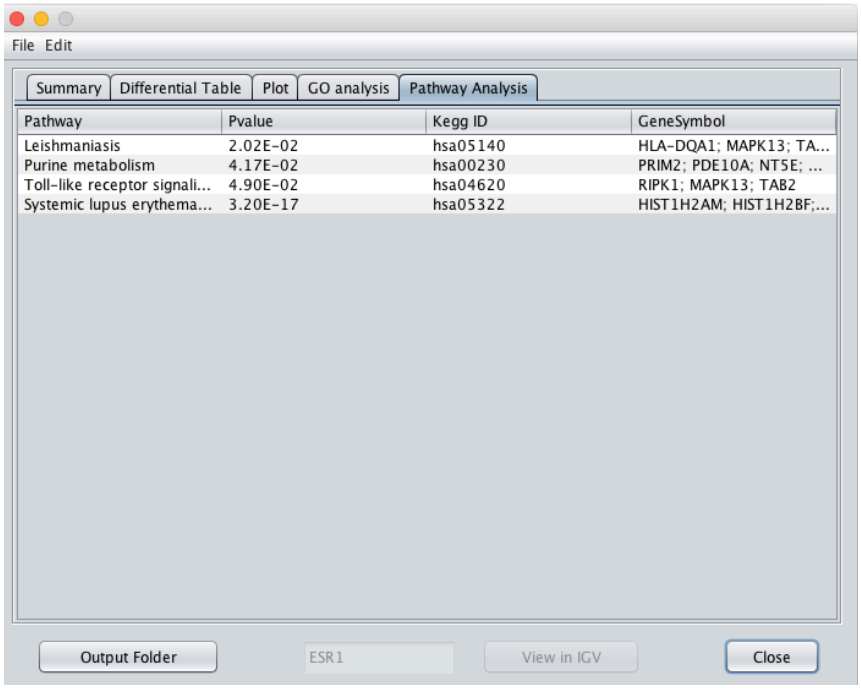


Figure 21) List of the enriched pathways.

Sample Output files

User can download sample output for the sample data from here: http://ec2-52-201-246-161.compute-1.amazonaws.com/guava/