

GUAVA Manual

Mayur Divate and Edwin Cheung

Version 1

Released on June 26, 2018

Table of Content

Description	Page
1. About GUAVA	3
2. How to download GUAVA	4
3. How to get help and report bugs	5
4. How to open Terminal	6
5. Installation of GUAVA	7
5.1 Installing dependencies for GUAVA	7
6. How to download a genome fasta file	8
7. How to start GUAVA	9
8. The graphical user interface of GUAVA	10
8.1 How to create a Bowtie/Bowtie2 index for a genome fasta file	10
8.1.1 How to use the Genome Index Builder program	10
8.2 ATAC-seq data analysis program GUI	12
8.2.1 How to use the ATAC-seq data analysis program	12
8.3 Output interface for GUAVA ATAC-seq data analysis	16
8.4 ATAC-seq differential analysis program GUI	24
8.4.1 How to use the ATAC-seq differential analysis program	24
8.5 Output interface of GUAVA ATAC-seq differential analysis	27

1. About GUAVA

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

GUAVA is a standalone GUI tool for the processing, analysis, and visualization of ATAC-seq data from raw sequencing reads to ATAC-seq signals. GUAVA can compare ATAC-seq signals from two conditions to identify genomic loci with differentially enriched ATAC-seq signals. Furthermore, GUAVA provides results on gene ontology and pathways analysis. Since using 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. The 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 was developed in Edwin Cheung's laboratory at the University of Macau.

2. How to download GUAVA

GUAVA is hosted on GitHub and can be downloaded to your computer by performing the following steps.

Step 1: Go to the link: <https://github.com/MayurDivate/GUAVA/releases>.

Step 2: Click on ‘Source code (zip)’.

Step 3: This will save the GUAVA zip package in your computer’s ‘Downloads’ folder.

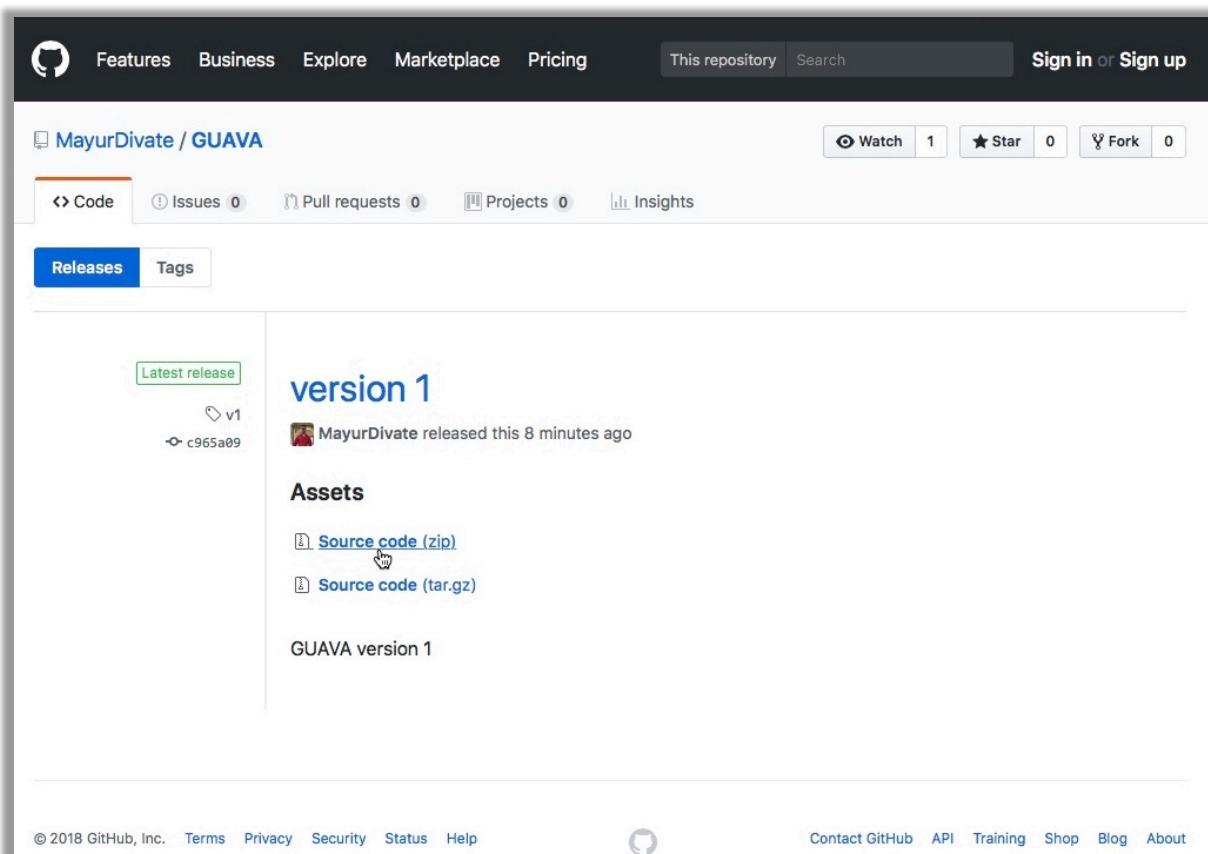


Figure1. GUAVA - GitHub project page. Screenshot showing the download page for GUAVA at GitHub. Users can download the GUAVA package by clicking on ‘Source code (zip)’.

If you would like to download the source code for GUAVA, use the following link:
<https://github.com/MayurDivate/GUAVASourceCode>.

☞ Please follow this link for video demo: <https://goo.gl/kgEZ7u>

3. How to get help and report bugs

Sometimes users face difficulties in installing or using bioinformatic tools. Therefore, it is very important to have an active forum where users can report issues and share information with authors and other users. Thus, GUAVA has a forum at GitHub for this and users can report any issues and share information by going to this link:

[https://github.com/MayurDivate/GUAVA/issues.](https://github.com/MayurDivate/GUAVA/issues)

4. How to open Terminal

After downloading GUAVA, users will need to install it using the Terminal. Here, we describe the procedure to open the Terminal for non-bioinformatics users.

MAC

Step 1: Open Finder

Step 2: Click on ‘Go’ in the menu bar and then select the ‘**Utilities**’ folder

Step 3: **Double-click** on the Terminal icon

Linux

Step 1: Press the ‘windows’ key on the keyboard

Step 2: Type “Terminal” in the search box

Step 3: Click on the Terminal icon

OR press “**Ctrl + Alt + T**” at the same time which will open Terminal

5. Installation of GUAVA

GUAVA can be installed in the home folder. Before proceeding, first copy/move the downloaded GUAVA package to the home folder and unzip the package. If the downloaded package is in the folder ‘Downloads’, then type the following commands in the Terminal to unzip the package.

```
mv ~/Downloads/GUAVA-1.zip ~/  
cd ~/  
unzip GUAVA-1.zip
```

NOTE: If you get an error saying that ‘No such file or directory’, then make sure the downloaded GUAVA package is present in the Downloads folder. Sometimes on the Mac, zip files are unzipped automatically. In this case, remove the ‘.zip’ suffix in the first command and skip the third command.

If the downloaded GUAVA package is in a different folder than ‘Downloads’, you will have to use the complete path of that folder instead of ~/Downloads/GUAVA-1.zip. To copy the path, simply copy the downloaded package and paste it in the Terminal

5.1 Installing dependencies for GUAVA

GUAVA depends on other tools in order to process ATAC-seq data (e.g. Bowtie for alignment). If any of the dependencies are not found on the system, GUAVA will not work properly. Therefore, to help users to install the dependencies, we have written a program called configure.sh, which automatically downloads and installs the dependencies. Please follow the steps given below to run configure.sh.

Run configure.sh

Step 1: Open Terminal

Step 2: Go to the GUAVA-1 folder

```
cd ~/GUAVA-1
```

Step 3: Install dependencies

```
sh ./configure.sh
```

Step 4: Close Terminal after the installation is complete

Note: The installation will take around 15-20 mins to finish. Also, you will need to press ‘enter’ several times to continue. Additionally, answer all questions with ‘yes’. If you see the message about ‘Xcode’, choose ‘install’ to continue. Some users may need to repeat steps 1-3 in a new Terminal if they cannot load ‘conda’ the first time.

This is the end of the installation. GUAVA is now ready to process ATAC-seq data.

6. How to download a genome fasta file

In order to use GUAVA, users will need to have on their computer a genome fasta file which is a file containing the nucleotide sequences from all of the chromosomes of a particular organism. The genome fasta file will be converted to a genome index (see Section 8.1) which is required for mapping reads with a sequence aligner tool. For users who do not know where to find the genome fasta file, they can use the UCSC link below and choose the desired organism to download the fasta file.

- | | |
|---------|--|
| Step 1. | Go to the link: http://hgdownload.soe.ucsc.edu/downloads.html . |
| Step 2. | Click on the desired organism (e.g. human). |
| Step 3. | Click on the 'Full data set' under the appropriate genome build (e.g. hg19). |
| Step 4. | Scroll down and then click on the chromFa.tar.gz to download the genome sequence. This will save the chromFa.tar.gz in the Downloads folder. |
| Step 5. | Open the Terminal. |
| Step 6. | Type the following commands to extract the chromosome files and merge them into a single file. |

```
cd ~/  
mkdir -p Genomes  
cd Genomes/  
mv ~/Downloads/chromFa.tar.gz ./  
tar -zxvf chromFa.tar.gz  
cat *fa > GenomeBuild.fasta
```

NOTE: For the last command, you will need to replace GenomeBuild.fasta with an appropriate name for the output file such as hg19.fasta or mm10.fasta. If chromFa.tar.gz is already converted into chromFa.tar file, then remove its '.gz' suffix in commands.

That's it, your genome fasta file is ready.

☞ Recommendation for first time users:

Genome fasta files are usually very large for organisms such as human and mouse and will take a very long time to generate a genome index (e.g. it takes 4-5 hrs for the human genome). Thus, we highly recommend first time users to perform a quick check of the GUAVA installation and its functions by using the 'Sample Genome' fasta file that we created which contains only data for chromosome 6 and mitochondria and takes only 5 mins to generate a genome index. To download it, please follow this link: http://ec2-52-201-246-161.compute-1.amazonaws.com/guava/sampleddata/Hs_demo.fasta.gz. After downloading is complete, you will have to unzip the file, Hs_demo.fasta.gz. For Mac users, this can be done by simply double-clicking it, while Linux users will have to right click on it and then select the 'Extract Here' option. This will create a new file called, Hs_demo.fasta, which can be used as the 'Sample Genome' fasta file in Step 5 of Section 8.1.1.

7. How to start GUAVA

Users can start GUAVA by using the following commands in the Terminal. This will open the GUAVA home window where users can choose the program they want to use.

```
cd ~/GUAVA-1  
java -jar GUAVA.jar
```

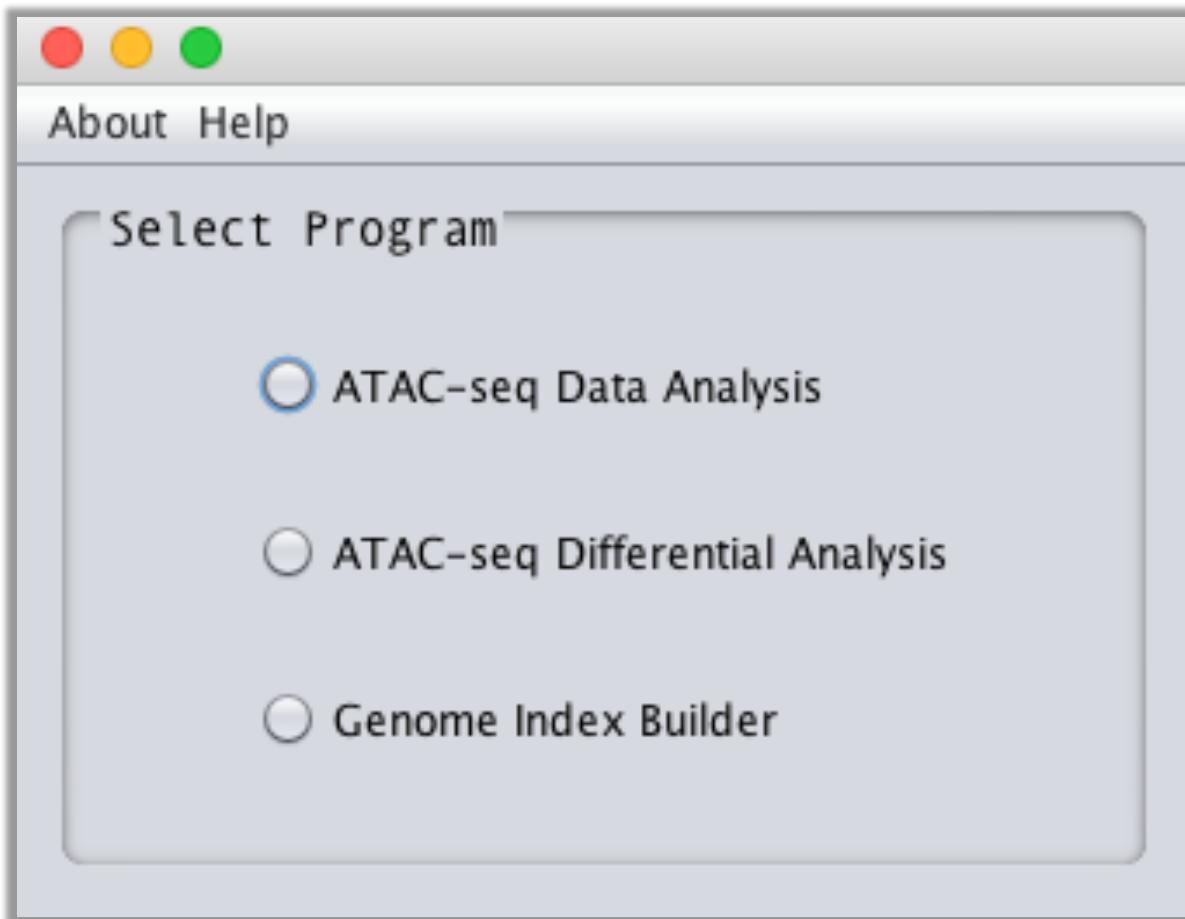


Figure 2. GUAVA-home window. Screenshot showing the GUAVA user interface which allows users to select the desired GUAVA program they wish to run.

8. The graphical user interface of GUAVA

As shown in Figure 2, the GUAVA tool consists of the following three programs:

- 1) **Genome Index Builder**: to create the Bowtie or Bowtie2 index of genome.
- 2) **ATAC-seq Data Analysis**: to process raw ATAC-seq sequencing reads.
- 3) **ATAC-seq Differential Analysis**: to compare ATAC-seq signals.

When the GUAVA GUI is evoked, it will open the GUAVA home window (Figure 2). From here, users can choose one of the above programs to proceed further. Based on the selection of the program, the desired input window will open (Figures 3, 4 and 14).

8.1 How to create Bowtie/Bowtie2 index for a genome fasta file

Read aligners such as Bowtie/Bowtie2 use a special set of files called the index that is generated from the genome fasta file. Index files are used to speed up the read mapping process so that the aligner can map millions of reads within a few hours. Therefore, users need to have a Bowtie/Bowtie2 index before read mapping (please note that the index format is different for each aligner). Thus, if this is your first time using GUAVA and/or you don't have a Bowtie/Bowtie2 index, then you need to first create an index as it is required to run the 'ATAC-seq Data Analysis' program. This is done using the '**Genome Index Builder**' program.

8.1.1 How to use Genome Index Builder program

Below is the step-by-step guide to use Genome Index Builder program.

Step 1: Open Terminal
Step 2: Start GUAVA (see the Section 7)
Step 3: Choose the ' Genome Index Builder ' program from the home window (Figure 2).
Step 4: Select the appropriate aligner (Bowtie or Bowtie2) that you would like use for ATAC-seq data analysis (Figure 3).
Step 5: Click on the ' Genome Fasta ' button to load the genome fasta file.
Step 6: Click on the ' output folder ' button to select the folder to store the index files. For example, you can create a folder called 'Bowtie1_index' or 'Bowtie2_index'.
Step 7: Click on the ' Start ' button to start program.
NOTE: The GUAVA GUI will remain blank until it has finished. It will take around 3-4 hrs to build the index for the hg19 genome. Once the program has finished, it will display a message. After that, click the 'OK' button in the message box to close GUAVA.

That's it, your Bowtie/Bowtie2 genome index is ready.

 **Recommendation for first time users:**

For first time GUAVA users, we highly recommend creating an index from the 'Sample Genome' fasta file (see Section 6), which will take only about 5 mins to generate. To do this, in Step 5 choose the Hs_demo.fasta file that was downloaded in Section 6 and create a Bowtie2 index.

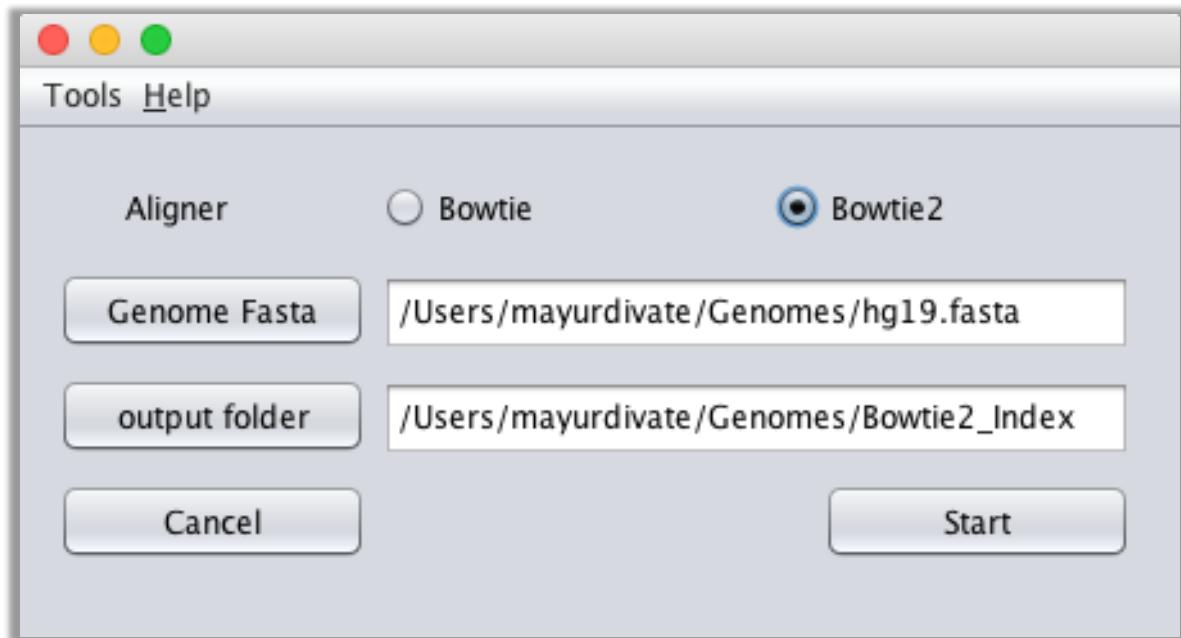


Figure 3. GUAVA Genome Index Builder program. Screenshot showing the interface for the 'Genome Index Builder' program. Here, users select the aligner, genome fasta file, and the output folder, and then click on the 'Start' button to run the program.

8.2 ATAC-seq data analysis program GUI

The ATAC-seq data analysis program accepts raw ATAC-seq reads as an input. Before aligning reads to genome, the program trims adapter sequences from reads using cutadapt if the trimming option has been selected. In addition, it filters unsuitable reads such as duplicate reads. Users can also exclude certain chromosomes from the analysis using the ‘Show Chromosomes’ button (Figure 5), which becomes active only after loading the genome index (Section 8.1). The ATAC-seq data analysis program uses MACS2 to identify ATAC-seq peaks and performs functional annotation of the ATAC-seq peaks.

☞ Recommendation for first time users:

In general, users can use their own fastq files if they generated their own ATAC-seq data or they can download publicly available data from databases such as GEO. For first time GUAVA users, we highly recommend testing the program using the ‘Sample dataset’ that we created from the GSE84515 dataset (contains data for chromosome 6 and mitochondria only). This dataset contains 4 samples from 2 conditions (*i.e.* DMSO and BYL719) with 2 biological replicates for each condition (in total there are 8 files). The ‘Sample dataset’ is located at the GUAVA website (<http://ec2-52-201-246-161.compute-1.amazonaws.com/guava/>), and can be downloaded by clicking on the ‘Sample Data’ option. This will download the ‘SampleData.tar.gz’ file. To extract the fastq files, Mac users can simply double-click on the SampleData.tar.gz file, while Linux users will have to right click and select the ‘Extract Here’ option. In addition, users can compare the results from their ATAC-seq data analysis of the ‘Sample dataset’ by downloading the ‘Sample output’ file from the GUAVA website. The output files provided at the GUAVA website were obtained by processing the ‘Sample dataset’ with all the parameters set at default values except we chose the ‘trim adapter?’ option, selected the hg19 genome assembly and chrM.

8.2.1 How to use ATAC-seq data analysis program

To use ATAC-seq data analysis, users will need provide fastq reads and the Bowtie/Bowtie2 index for genome. Below is the step-by-step guide to using the ATAC-seq data analysis program (**Steps 6-9, 11, and 12 are optional**).

- | | |
|---------|--|
| Step 1. | Open Terminal |
| Step 2. | Start GUAVA (see the Section 7) |
| Step 3. | Choose the “ ATAC-seq Data Analysis ” program from the home window (Figure 2). |
| Step 4. | Click on “ R1 fastq ” and “ R2 fastq ” button to upload Read 1 and Read 2 fastq files, respectively.
(For example, from the ‘Sample dataset’ R1 fastq file=DMSO_Rep1_R1.fastq and R2 fastq file=DMSO_Rep1_R2.fastq) |
| Step 5. | Select ‘Trim adapter?’ option if reads contain adapter. |

- Step 6. Add appropriate number in the ‘**Maximum Ns**’ box.
If any read contains more than the specified number of Ns after the adapter trimming, that read pair will be discarded (default is 2).
- Step 7. Add appropriate number in the ‘**Minimum Read Length**’ box.
If any read is shorter than the specified length after adapter trimming, that read pair will be discarded (default is 30 bp).
- Step 8. Add appropriate number in the ‘**Error Rate**’ box.
This is the allowed number of mismatches as a fraction of length. For example, if the error rate is 0.1 then 1 mismatch is allowed for a 10 bp match of adapter sequence (default is 0.1).
- Step 9. Deselect the ‘**Nextera XT Adapter**’ option if Nextera XT adapter was not used for library preparation and specify the custom adapter in the ‘**Adapter Sequence**’ box.
- Step 10. If users want to use Bowtie for read mapping they should select “**Bowtie index**” from the dropdown menu or select “**Bowtie2 index**” to use Bowtie2. Then, using the ‘**browse**’ button upload the anyone of the file from index (Bowtie or Bowtie2 index). Bowtie index files have suffix ebwt whereas Bowtie2 index file have bt2. Please see Section 8.1 on how to create Bowtie or Bowtie2 index.
For first time users using the ‘Sample genome’, you should select the ‘Hs_demo.fasta.1.bt2’ file from the output folder of the genome index builder tool.
- Step 11. Change the number in the ‘**Maximum insert size**’ box if required.
This is the maximum distance allowed between the read pairs for a paired-end alignment (default is 2,000 bp).
- Step 12. Change the value of **Maximum genomic hits** (Bowtie) or **Minimum Mapping Quality**.
The maximum genomic hit (Bowtie) and Minimum Mapping Quality (Bowtie2) to discard reads pairs which have multiple alignments. Higher mapping quality gives more unique mapping for reads. Similarly, lower number of genomic hits give more unique mapping. The default maximum genomic hits = 1 and the mapping quality = 30.
- Step 13. Select appropriate genome from the **Genome assembly** drop-down menu.
This will be used for peak calling, annotation and functional analysis.
- Step 14. Click on the ‘**Show Chromosomes**’ button to exclude reads mapping to specific chromosomes such as mitochondrial chromosome. After clicking this button, it will open a new window (Figure 5). Then, select the desired chromosome(s) and click on the ‘**add >>**’ button to exclude them from the analysis.
- Step 15. Add the appropriate number in the ‘**RAM (GB)**’ box.
RAM in GB to be used by GUAVA (default is 1).
- Step 16. Add the appropriate number in the ‘**CPU units**’ box.
Number of CPU units to be used by GUAVA (default is 1).

- Step 17. Select the appropriate value (**p/q value**) from the drop-down menu and **specify the cut off value in the box next to it**. This will be used by MACS2 to filter peaks (default is q value).
- Step 18. Click on the '**Output folder**' to select the folder to GUAVA ATAC-seq data analysis results.
- Step 19. Finally, click on the '**Start Analysis**' button to start the 'ATAC-seq data analysis' program.
This will start the program if the provided options are valid. After the analysis has finished, the results are displayed using the output interface which is described in more detail in Section 8.3.
- 'Reset All'** The button to set all parameters to the default value.

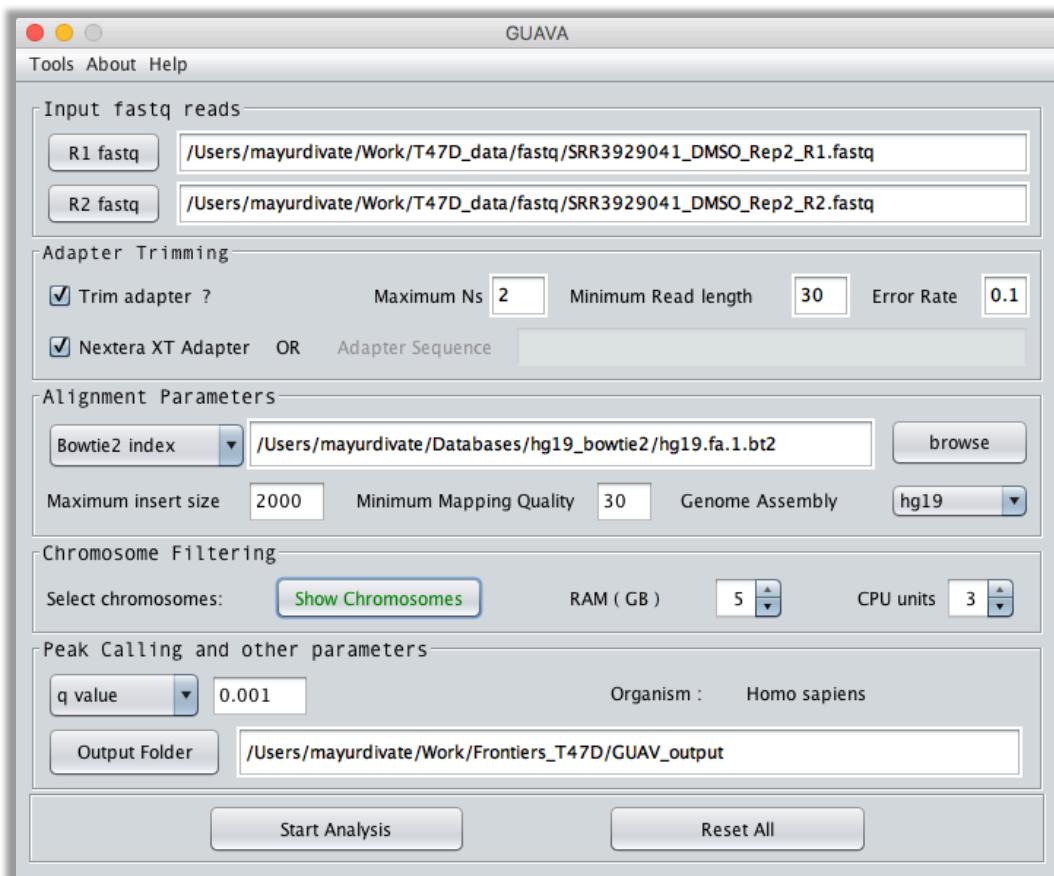


Figure 4. Input window of GUAVA ATAC-seq data analysis program. Screenshot shows the input window of the ATAC-seq data analysis program for uploading input files such as fastq and genome index, and set parameters such as insert size, p/q value, etc.

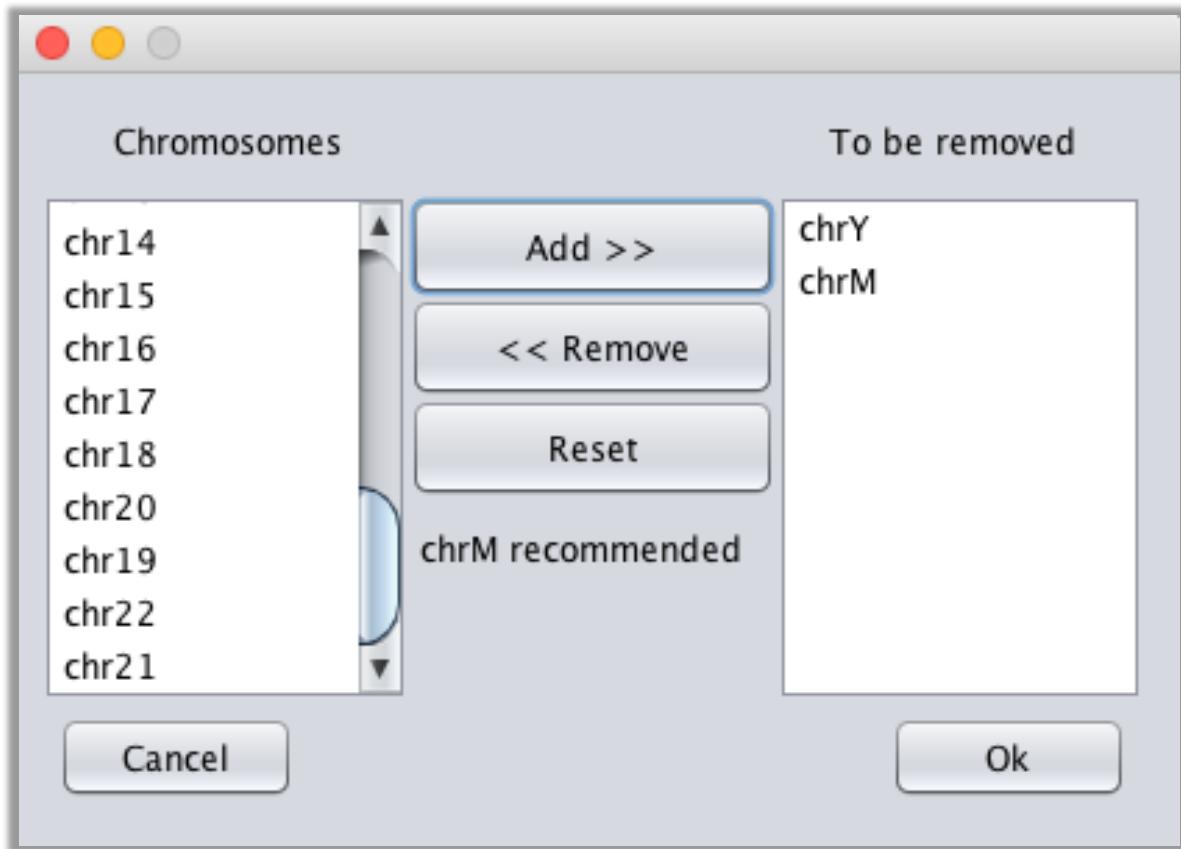


Figure 5. Interface to select chromosomes for alignment filtering. Here, users can add the desired chromosomes to the ‘To be removed’ list to discard reads aligning to those chromosomes they do not want in the analysis.

8.3 Output interface for ATAC-seq data analysis program

Once GUAVA has finished the analysis, it will show the results as a tabular output interface (Figure 6-13). GUAVA also facilitates the visualization of ATAC-seq signals on the IGV browser.

8.3.1 Alignment Statistics

This tab provides the reads mapping statistics (e.g. the total number of reads mapped and not mapped to the genome) along with the summary of the input files and parameters.

Parameter	Value
R1 fastq	SRR3929041_DMSO_Rep2_R1.fastq
R2 fastq	SRR3929041_DMSO_Rep2_R2.fastq
Genome index	hg19.fa
Maximum Insert Size	2000
Minimum Mapping Quality	30
Total Reads	94208170
Total Aligned Reads	75771166 (80.43%)
Total Reads Failed to Align	7719282 (8.19%)
Total Reads with Low Mapping Quality	10717722 (11.38%)
Total chrY Reads	103281 (0.11%)
Total chrM Reads	2907706 (3.09%)

Figure 6. Input summary and alignment statistics.

8.3.2 Alignment Filtering

This tab contains two tables: 1) the alignment filtering statistics (duplicates, useful reads, etc), and 2) a summary of the MACS2 peak calling results.

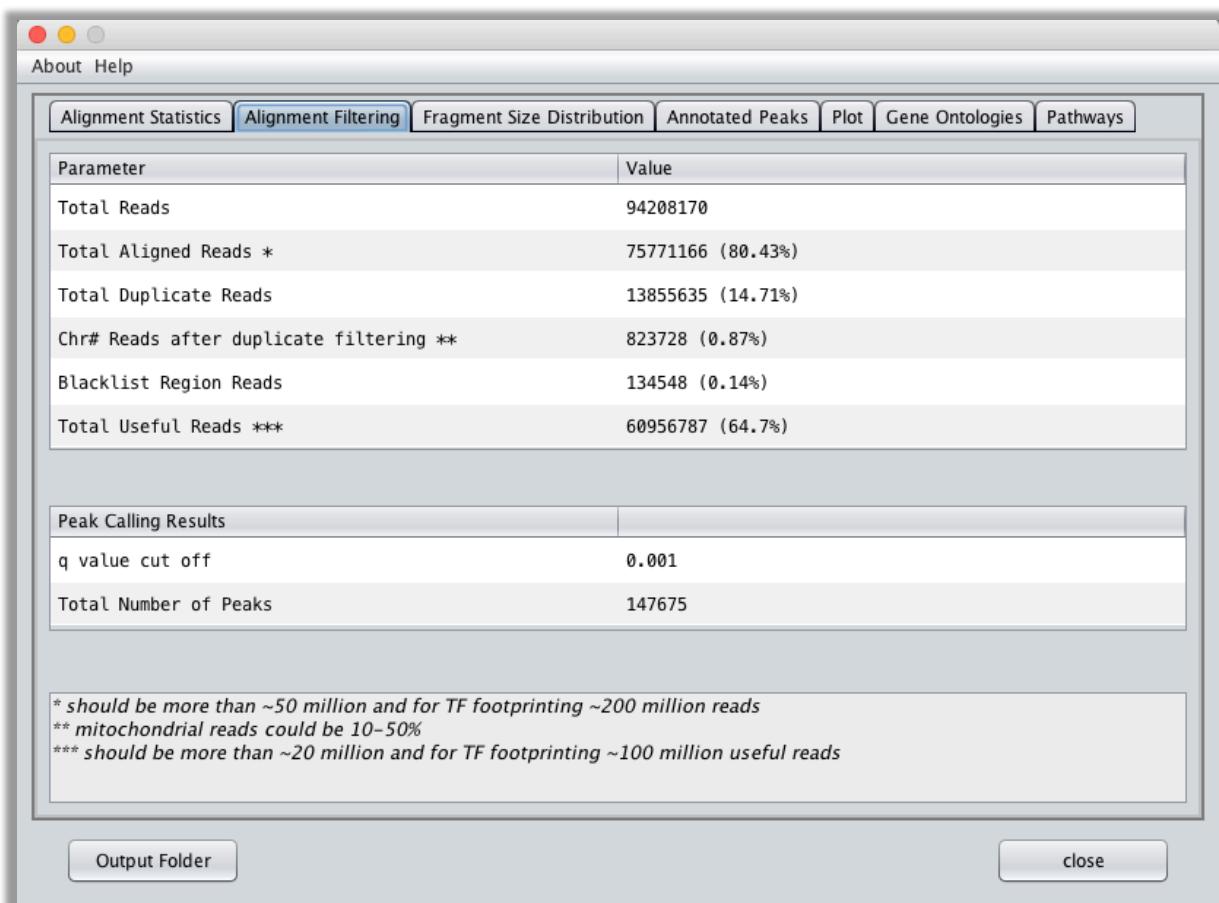
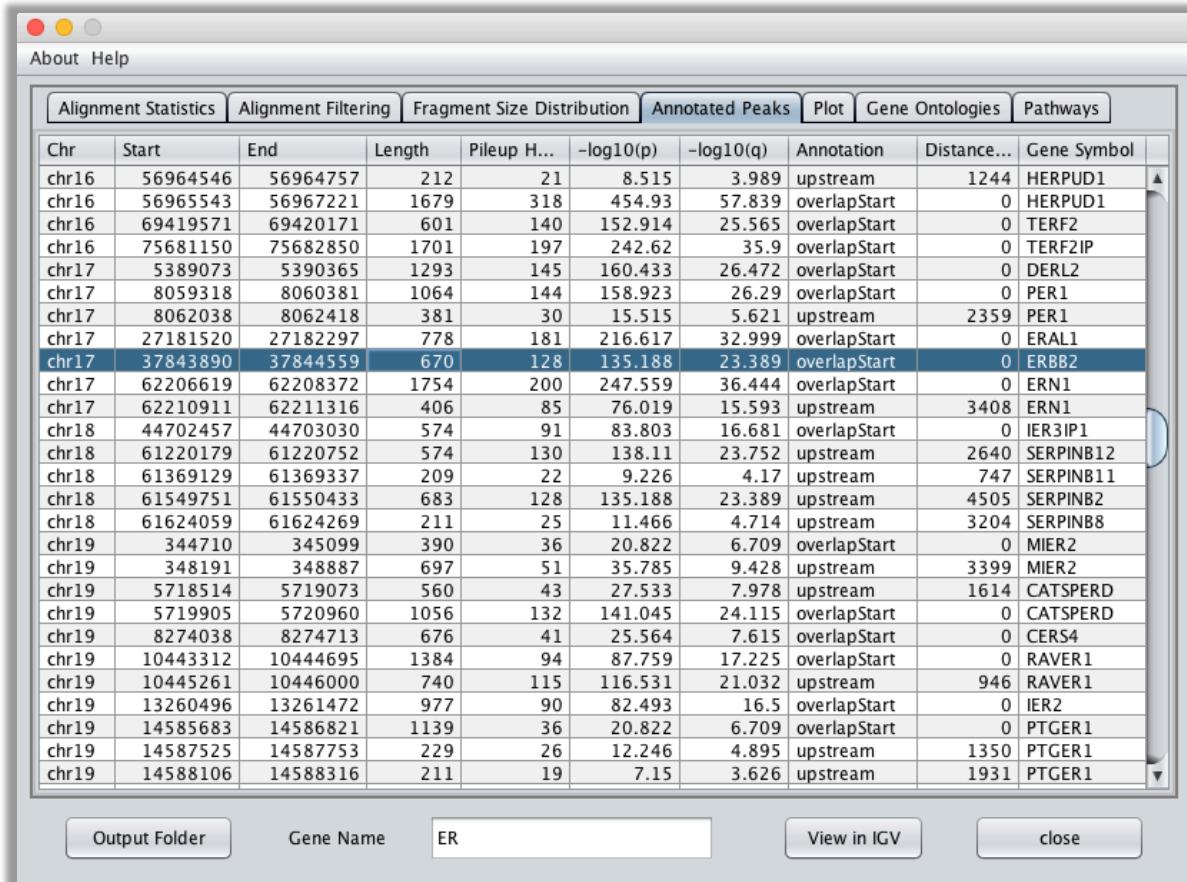


Figure 7. Read filtering and peak calling summary.

8.3.3 Annotated Peaks

This window shows the ATAC-seq annotated peaks. It contains information on peak location, nearest gene, and distance to TSS. This window also provides easy access to the IGV for visualizing peaks and automatically generated normalized ATAC-seq signals by GUAVA. Users can search their peak of interest by typing the symbol of the nearest gene in the search box at the bottom.



The screenshot shows a software interface for managing ATAC-seq peak annotations. At the top, there's a menu bar with 'About' and 'Help' options. Below the menu is a tab bar with several tabs: 'Alignment Statistics', 'Alignment Filtering', 'Fragment Size Distribution', 'Annotated Peaks' (which is currently selected), 'Plot', 'Gene Ontologies', and 'Pathways'. The main area is a table with the following columns:

Chr	Start	End	Length	Pileup H...	$-\log_{10}(p)$	$-\log_{10}(q)$	Annotation	Distance...	Gene Symbol
chr16	56964546	56964757	212	21	8.515	3.989	upstream	1244	HERPUD1
chr16	56965543	56967221	1679	318	454.93	57.839	overlapStart	0	HERPUD1
chr16	69419571	69420171	601	140	152.914	25.565	overlapStart	0	TERF2
chr16	75681150	75682850	1701	197	242.62	35.9	overlapStart	0	TERF2IP
chr17	5389073	5390365	1293	145	160.433	26.472	overlapStart	0	DERL2
chr17	8059318	8060381	1064	144	158.923	26.29	overlapStart	0	PER1
chr17	8062038	8062418	381	30	15.515	5.621	upstream	2359	PER1
chr17	27181520	27182297	778	181	216.617	32.999	overlapStart	0	ERAL1
chr17	37843890	37844559	670	128	135.188	23.389	overlapStart	0	ERBB2
chr17	62206619	62208372	1754	200	247.559	36.444	overlapStart	0	ERN1
chr17	62210911	62211316	406	85	76.019	15.593	upstream	3408	ERN1
chr18	44702457	44703030	574	91	83.803	16.681	overlapStart	0	IER3IP1
chr18	61220179	61220752	574	130	138.11	23.752	upstream	2640	SERPINB12
chr18	61369129	61369337	209	22	9.226	4.17	upstream	747	SERPINB11
chr18	61549751	61550433	683	128	135.188	23.389	upstream	4505	SERPINB2
chr18	61624059	61624269	211	25	11.466	4.714	upstream	3204	SERPINB8
chr19	344710	345099	390	36	20.822	6.709	overlapStart	0	MIER2
chr19	348191	348887	697	51	35.785	9.428	upstream	3399	MIER2
chr19	5718514	5719073	560	43	27.533	7.978	upstream	1614	CATSPERD
chr19	5719905	5720960	1056	132	141.045	24.115	overlapStart	0	CATSPERD
chr19	8274038	8274713	676	41	25.564	7.615	overlapStart	0	CERS4
chr19	104443312	10444695	1384	94	87.759	17.225	overlapStart	0	RAVER1
chr19	10445261	10446000	740	115	116.531	21.032	upstream	946	RAVER1
chr19	13260496	13261472	977	90	82.493	16.5	overlapStart	0	IER2
chr19	14585683	14586821	1139	36	20.822	6.709	overlapStart	0	PTGER1
chr19	14587525	14587753	229	26	12.246	4.895	upstream	1350	PTGER1
chr19	14588106	14588316	211	19	7.15	3.626	upstream	1931	PTGER1

At the bottom of the window, there are four buttons: 'Output Folder', 'Gene Name' (containing 'ER'), 'View in IGV', and 'close'.

Figure 8. A table containing peak annotation information.

8.3.4 Visualization of ATAC-seq peaks using the Integrated Genome Viewer (IGV)

Users can visualize their peak of interest from the table (Figure 8) by selecting it and then clicking the ‘View in IGV’ button. This will automatically load the normalized ATAC-seq signals and peaks to the IGV browser as shown below.

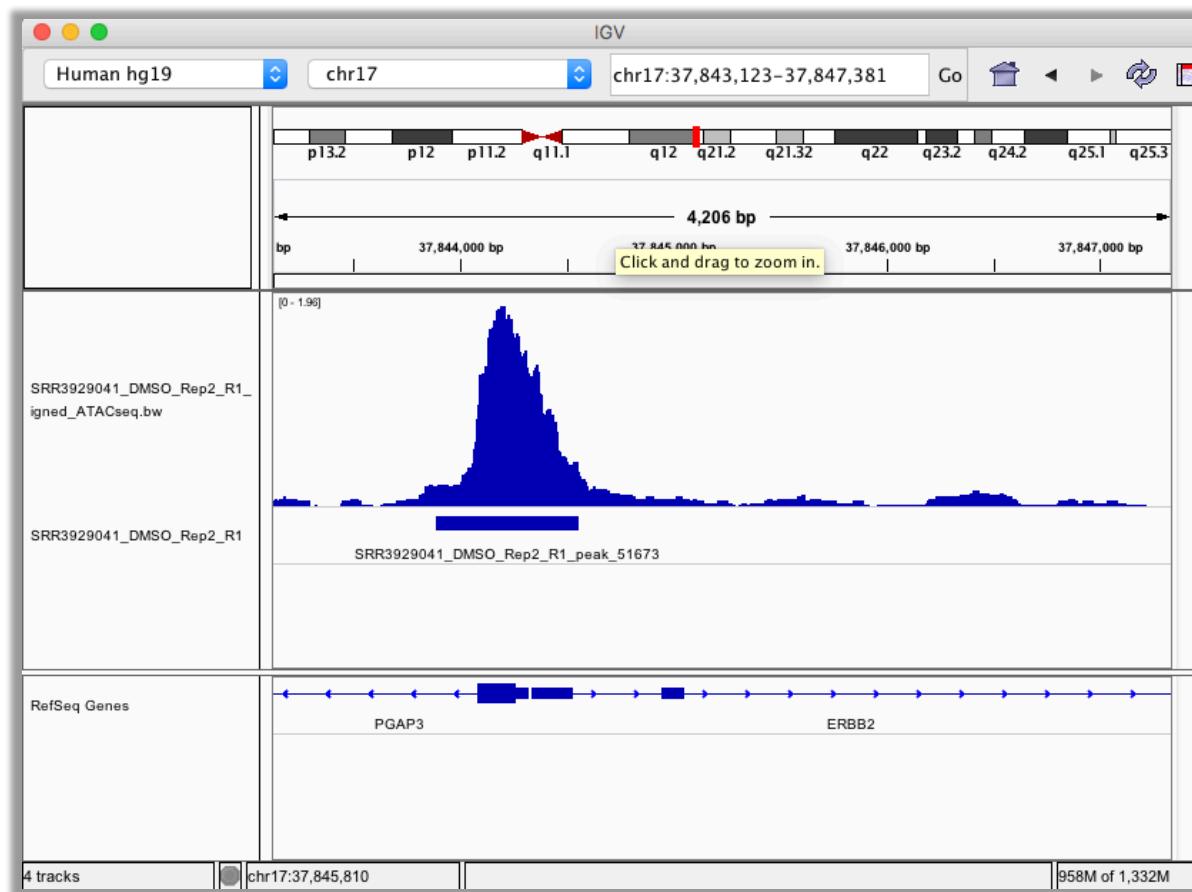


Figure 9. Visualization of ATAC-seq peaks with IGV.

8.3.5 Fragment Size Distribution

This plot shows the observed fragment size distribution for an ATAC-seq sample.

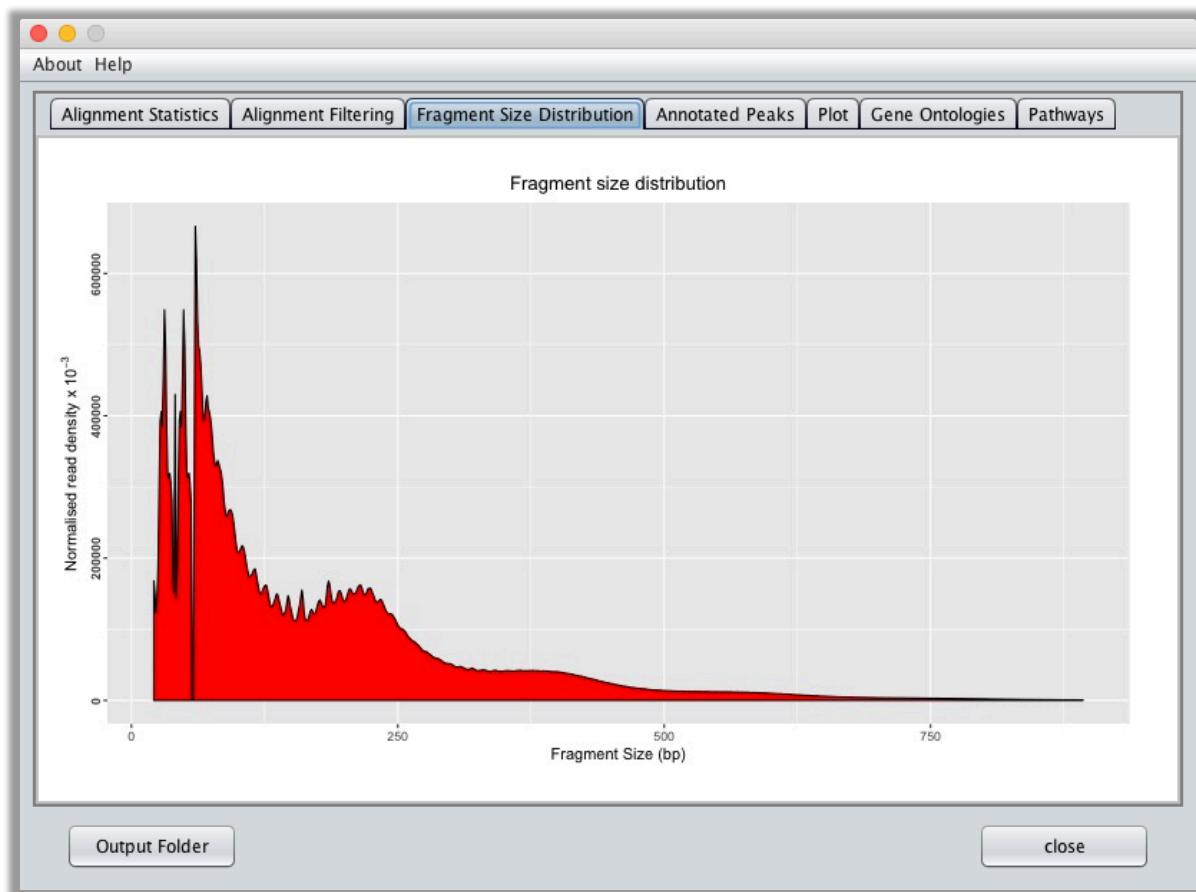


Figure 10. Graph showing the fragment size distribution.

8.3.6 Plot

This tab contains a bar chart which illustrates the distribution of the annotated peaks in various genomic locations such as the promoter, intron, exon, UTR, etc.

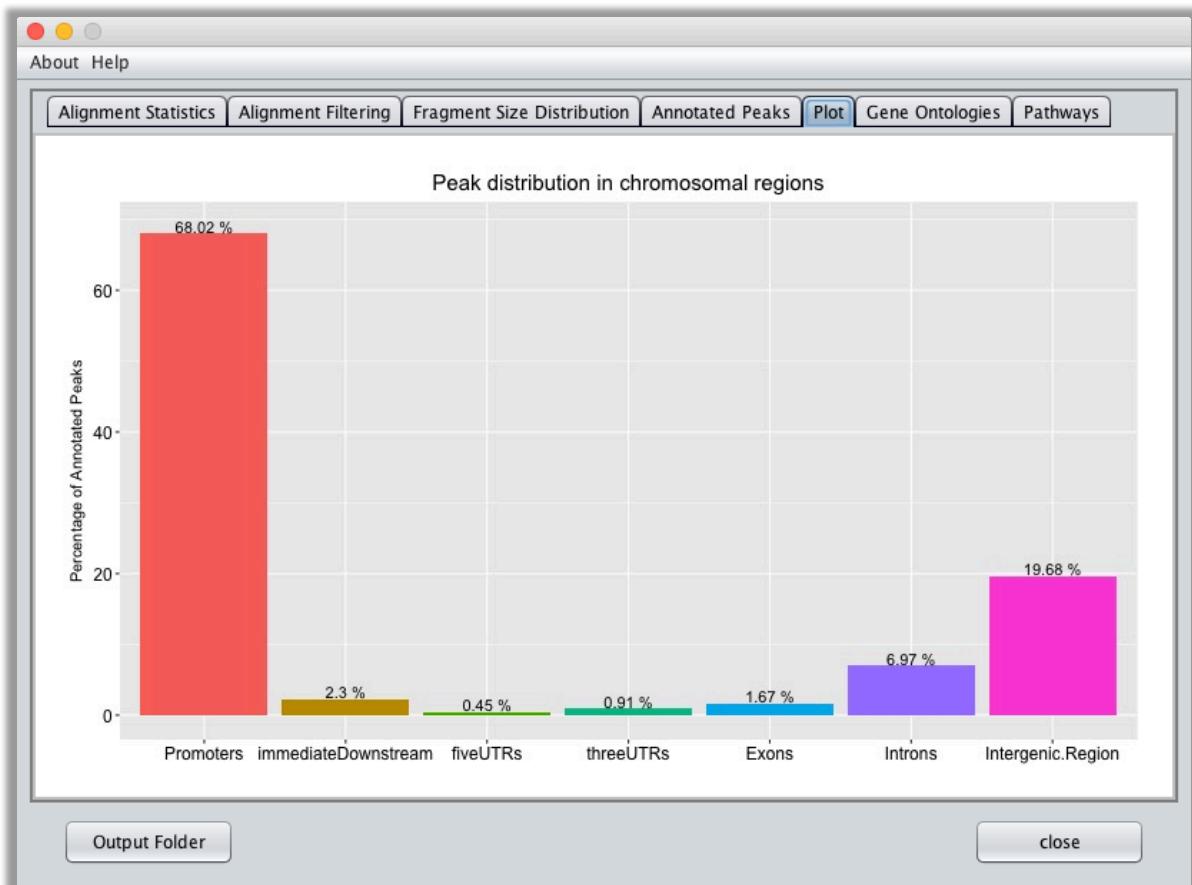


Figure 11: Bar chart showing the distribution of annotated ATAC-seq peaks in various genomic regions.

8.3.7 Gene Ontologies

This tab shows the list of the over-represented gene ontologies associated with ATAC-seq peaks.

The screenshot shows a software window with a title bar 'About Help'. Below the title bar is a horizontal navigation bar with tabs: Alignment Statistics, Alignment Filtering, Fragment Size Distribution, Annotated Peaks, Plot, Gene Ontologies (which is highlighted in blue), and Pathways. The main area of the window contains a table with the following data:

GO ID	GO Term	Type	P value	adj. P value	Gene Symbols
GO:0007389	pattern specification process	BP	3.37E-04	9.70E-03	ERBB4; SHH; DCANP1; BASP1; IFT57...
GO:0009205	purine ribonucleoside triphosphate ...	BP	1.25E-05	6.00E-04	SURF1; IGF1; ATP5B; ATP5C1; PFKF...
GO:0034645	cellular macromolecule biosynthetic ...	BP	1.65E-14	5.37E-12	CRLF3; ZNF143; MDK; CELF4; ATF2;...
GO:0034644	cellular response to UV	BP	2.46E-04	7.51E-03	MC1R; PCNA; AQP1; DDB2; ERCC4; ...
GO:0010557	positive regulation of macromolecule...	BP	2.20E-08	2.13E-06	NPM1; TGFB3; EVX1; ERCC3; ETV1;...
GO:0001841	neural tube formation	BP	4.35E-06	2.35E-04	TGFB1; TSC1; SKI; SEMA4C; IFT122;...
GO:0010556	regulation of macromolecule biosynt...	BP	1.01E-07	8.23E-06	ZNFS74; ISL2; TNFSF4; EFCAB6; ZNF...
GO:0001843	neural tube closure	BP	3.74E-06	2.08E-04	VANGL2; PTCH1; KDM2B; PRKACB; ...
GO:0033554	cellular response to stress	BP	6.06E-22	6.05E-19	FBXW11; INO80E; PCK1; GPX3; GSR...
GO:0017076	purine nucleotide binding	MF	1.08E-11	3.30E-09	PAPSS2; SPHK1; ACVR1; UBE2Q2; K...
GO:0030163	protein catabolic process	BP	1.01E-10	1.73E-08	CUL7; HDAC6; NGLY1; USP15; XPO1...
GO:0033674	positive regulation of kinase activity	BP	1.69E-04	5.43E-03	MNAT1; TRAF7; FRS2; CCNT2; NEK...
GO:0034641	cellular nitrogen compound metaboli...	BP	3.78E-18	2.36E-15	CHKA; PSMC4; ZNF473; PSMA3; PSM...
GO:0071822	protein complex subunit organization	BP	4.13E-11	7.28E-09	PPP6R1; LMAN1; DNM3; TRAPP3; ...
GO:0043009	chordate embryonic development	BP	1.01E-07	8.23E-06	KDM6A; LAT51; DVL2; SKIL; CCNB1...
GO:0071826	ribonucleoprotein complex subunit o...	BP	9.15E-09	9.66E-07	EIF3A; TSR1; ZNHIT6; RPL23A; NIP7...
GO:1905037	autophagosome organization	BP	1.18E-04	4.07E-03	IFT88; RAB5A; RAB7A; MAP1LC3B; P...
GO:0043241	protein complex disassembly	BP	1.18E-06	7.25E-05	ERAL1; MRPL17; MRPL46; MRPS14; ...
GO:0044451	nucleoplasm part	CC	4.79E-21	5.18E-19	POLR1D; ERCC6; POLR3K; ERCC1; A...
GO:0042273	ribosomal large subunit biogenesis	BP	5.51E-06	2.85E-04	ZNF622; RPL3; RPL3L; SURF6; BRIX1...
GO:0042274	ribosomal small subunit biogenesis	BP	1.97E-05	8.71E-04	PDCD11; NGDN; UTP4; KRI1; DCAF...
GO:0045786	negative regulation of cell cycle	BP	3.61E-11	6.69E-09	CNOT8; KANK2; FBXO7; CCNB1; HE...
GO:0044452	nucleolar part	CC	2.81E-08	1.08E-06	TIMM13; DDX46; ANKRD1; WDR36; ...
GO:0044455	mitochondrial membrane part	CC	1.17E-06	3.52E-05	SLC25A44; COX7A2; NDUFV3; COX...
GO:0046872	metal ion binding	MF	9.22E-05	5.04E-03	ALAD; ACE; DTX1; NR2F6; THAP3; L...
GO:0044454	nuclear chromosome part	CC	2.61E-04	4.71E-03	ASH2L; EP400; SS18; TAF5; UHRF2; ...
GO:0044212	transcription regulatory region DNA ...	MF	8.19E-07	7.60E-05	HES7; GMEB2; ARID5B; HINFP; TAF7...

At the bottom left is a button labeled 'Output Folder'. At the bottom right is a button labeled 'close'.

Figure 12: Over-represented gene ontology terms associated with ATAC-seq peaks.

8.3.8 Pathways

This tab shows the list of the over-represented KEGG pathways associated with ATAC-seq peaks.

KEGG ID					Pathway Name	P value	adj. P value	Gene Symbols
hsa04012	ErbB signaling pathway	1.62E-03	1.55E-04	ERBB4; PAK4; RPS6KB2; CRK; MAP2K4; ...				
hsa03040	Spliceosome	3.78E-06	8.24E-08	PRPF8; THOC1; SMNDC1; PHF5A; PRPF4...				
hsa03013	RNA transport	7.84E-04	5.13E-05	EIF3F; EIF1B; EIF4E2; EIF1; THOC1; POP1...				
hsa04310	Wnt signaling pathway	1.57E-03	1.44E-04	WNT9A; FZD9; NFATC3; CUL1; CHP1; PP...				
hsa05219	Bladder cancer	2.28E-03	2.89E-04	FGR3; MAPK3; EGF; MAP2K1; RAF1; RP...				
hsa03018	RNA degradation	3.25E-03	4.26E-04	EXOSC10; CNOT6; EXOSC5; CNOT6L; D...				
hsa04110	Cell cycle	1.63E-07	1.42E-09	MDM2; CCND1; CDK7; HDAC2; ANAPC1...				
hsa04010	MAPK signaling pathway	7.95E-03	1.35E-03	PLA2G12A; CACNA1S; MAPK7; PRKCG; P...				
hsa04146	Peroxisome	2.18E-03	2.57E-04	IDH1; ACOT8; ABCD2; GNPAT; PAOX; A...				
hsa03030	DNA replication	1.85E-03	2.01E-04	RNASEH1; RPA3; RNASEH2A; POLD2; RF...				
hsa04810	Regulation of actin cytoskeleton	2.16E-04	7.10E-06	MYH10; RAC1; IQGAP2; FGF19; WASF2; ...				
hsa05220	Chronic myeloid leukemia	3.97E-04	1.73E-05	BRAF; RAF1; NRAS; SHC1; CDKN1B; CTB...				
hsa04520	Adherens junction	1.25E-04	3.28E-06	IQGAP1; MAP3K7; FGFR1; RAC2; LMO7; ...				
hsa05212	Pancreatic cancer	5.86E-04	3.58E-05	EGFR; BCL2L1; TGFA; RALB; VEGFC; TGF...				
hsa00510	N-Glycan biosynthesis	2.19E-03	2.68E-04	ST6GAL2; DPAGT1; DAD1; ALG11; DOL...				
hsa03410	Base excision repair	3.37E-03	4.70E-04	MPG; SMUG1; POLD2; NEIL2; PCNA; OG...				
hsa04141	Protein processing in endoplasmic reticul...	1.08E-06	1.42E-08	DNAJA1; CANX; ATXN3; FBXO6; UGGT2; ...				
hsa04510	Focal adhesion	5.00E-04	2.62E-05	ITGA8; VAV2; COL2A1; AKT1; CCND2; A...				
hsa00520	Amino sugar and nucleotide sugar meta...	8.49E-03	1.52E-03	GALK2; CYB5R1; GCK; CHIT1; CYB5R2; ...				
hsa04114	Oocyte meiosis	3.61E-03	5.52E-04	CDK1; YWHAZ; PPP3CC; MAD2L2; SMC3; ...				
hsa05211	Renal cell carcinoma	3.55E-03	5.27E-04	ARNT; MAP2K2; MET; ELOC; HGF; PAK1; ...				
hsa04910	Insulin signaling pathway	3.37E-03	4.56E-04	HKDC1; PIK3CA; BRAF; PPP1CB; ELK1; F...				
hsa00310	Lysine degradation	5.75E-03	9.30E-04	DLST; SETDB2; ACAT2; SETD1A; SETMA...				
hsa04360	Axon guidance	9.09E-04	6.58E-05	PLXNB2; SEMA4A; SEMA3B; EFNA4; RAS...				
hsa00020	Citrate cycle (TCA cycle)	6.59E-03	1.09E-03	DLD; SUCLG2; DLAT; FH; PDHB; MDH2; I...				
hsa05016	Huntington's disease	3.97E-04	1.63E-05	POLR2F; NDUFS4; NDUFS3; NDUFS1; TFA...				
hsa00280	Valine, leucine and isoleucine degradation	1.69E-03	1.69E-04	AUH; HADHA; ACADS; PCCB; ALDH9A1; ...				
hsa00521	Choline metabolism in cancer	2.00E-04	5.70E-05	MARCK1; ATCF; CAV1; ADDC1; ADDC1A; ...				

Figure 13: Over represented KEGG pathways associated with ATAC-seq peaks.

NOTE: The above results are stored in the output folder. To access the output folder, click on the '**Output Folder**' button located at the bottom-left corner.

8.4 ATAC-seq differential analysis program

The ATAC-seq differential analysis program compares ATAC-seq signals from two different conditions. It provides users results for the differentially enriched signals, as well as the peak annotation and functional analysis for the differentially enriched peaks. There are two input windows for this program. The first window is to upload the ATAC-seq signals for the two different conditions and replicates (Figure 14). The second window allows users to specify the differential analysis related parameters such as fold change and p value. (Figure 15).

8.4.1 How to use ATAC-seq differential analysis program

Below is the step-by-step guide to use ATAC-seq Differential Analysis program (Figures 14 and 15).

☞ Recommendation for first time users:

For first time GUAVA users, we recommend performing ATAC-seq differential analysis using the output files that were generated with the ‘Sample dataset’ in Section 8.2.

Again, users can compare their results of the differential analysis using the ‘Sample dataset’ by downloading the results we generated using the same dataset. Please follow this link <http://ec2-52-201-246-161.compute-1.amazonaws.com/guava/>. Click on the ‘2) ATAC-seq Differential Analysis Program’. This will download the file named as Sample_Output_Gdiff.tar.gz. To unzip it, Mac users can simply double-click on the downloaded file, while Linux users will have to right click on it and then select ‘Extract Here’ option. This will create new folder called as ‘Sample_Output_Gdiff’ which will contain the differential analysis results. In this analysis, we used all of the parameters with default values and the hg19 genome.

Step 1.	Open Terminal
Step 2.	Start GUAVA (see Section 7)
Step 3.	Choose the “ ATAC-seq Differential Analysis ” program from the home window (Figure 2).
Step 4.	Type the desirable Project Name in the box at top (Figure 14).
Step 5.	Next, click on the ‘ Add File ’ button to upload input files one by one. You have to upload both bam and peak file for each sample. GUAVA accepts peak file in both ‘narrowPeak’ and ‘BED’ format. If you processed ATAC-seq samples using GUAVA then select *.ATAC-seq.bam and narrowPeak files from the GUAVA output folder. As an example for first time users who processed the ‘Sample dataset’, they should upload ‘DMSO_Rep1_R1_aligned_ATACseq.bam’ and ‘DMSO_Rep1_R1_peaks.narrowPeak’ files from the sample output folder (DMSO_Rep1_R1_OUTPUT) for the DMSO replicate 1 sample.
Step 6.	After that, click on the drop-down menu in the Replicate column to select the Replicate number for the sample (e.g. I, II, III or IV).

Step 7.	Similarly, assign the appropriate condition (Control or Treatment) for each sample using the drop-down menu in the condition column.
Step 8.	Repeat steps 5 to 7 until you upload the all the samples. To unload the file, first select the file from the list and then click on the ' Remove ' button.
Step 9.	Click on the ' Next ' button to set differential analysis parameters such as log2 fold change.
Step 10.	Select the correct genome version (e.g. hg19) from the ' Genome ' drop-down menu.
Step 11.	Set the appropriate value in the ' CPUs ' box. Number of CPU units to be used by GUAVA (default is 1).
Step 12.	Select the cut off value from the ' log2 (Fold Change) ' drop-down as a log2 fold change cut off for the differential analysis (default is 2).
Step 13.	Enter the p value cut off for the differentially enriched ATAC-seq peaks in ' P value ' box (default 0.05).
Step 14.	Add a preferred value in ' Upstream of TSS ' box. If a peak is present within a specified distance (in base pair) upstream from the TSS of a gene, then that gene will be associated with the peak for functional analysis. (Default is 5000 bp).
Step 15.	Add a preferred value in ' Downstream of TSS ' box. If a peak is present within a specified distance (in base pair) downstream from the TSS of a gene, then that gene will be associated with the peak for functional analysis. (Default is 5000 bp).
Step 16.	Click on the ' output folder ' button to select folder to store differential analysis results.
Step 17.	Finally, click on the start button. This will initiate differential analysis provided that all parameters are valid. Usually, it takes less than 20 mins to complete. After it has finished the analysis, results are shown in the output interface (Section 8.5).

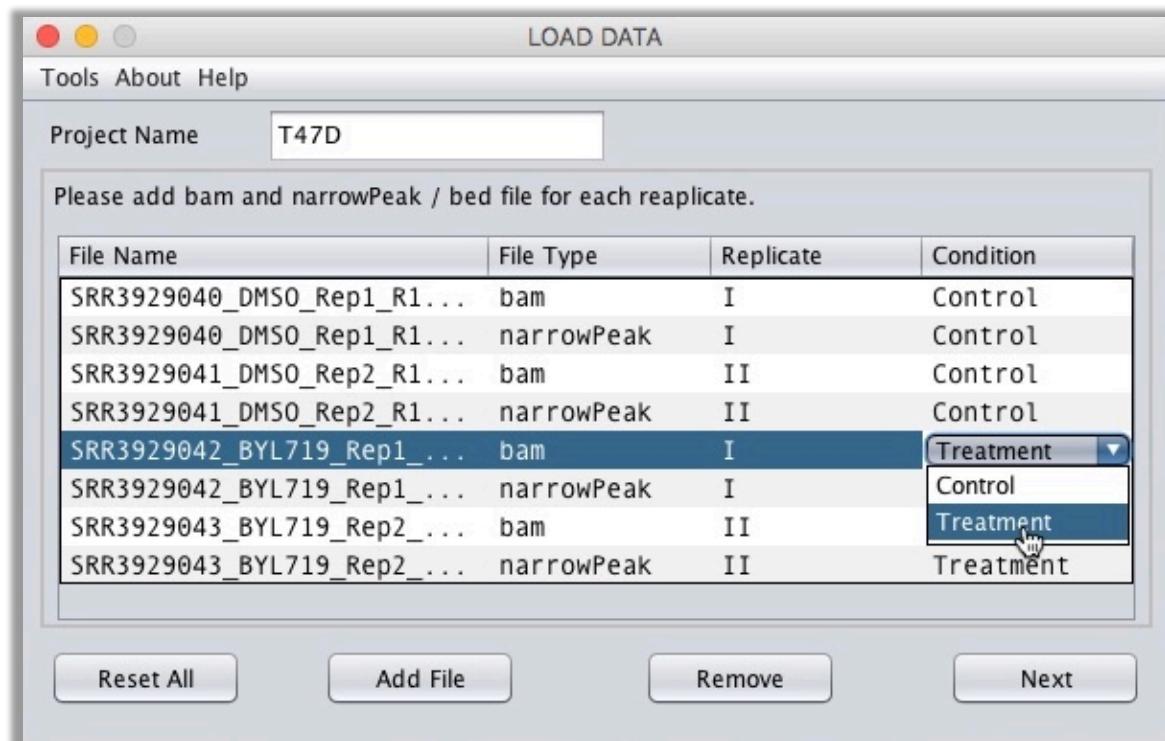


Figure 14. GUAVA ATAC-seq differential analysis input interface 1. Screenshot shows the first input window of the ATAC-seq differential analysis program for uploading input files such as bam and narrowPeak, and assign replicate numbers and conditions.

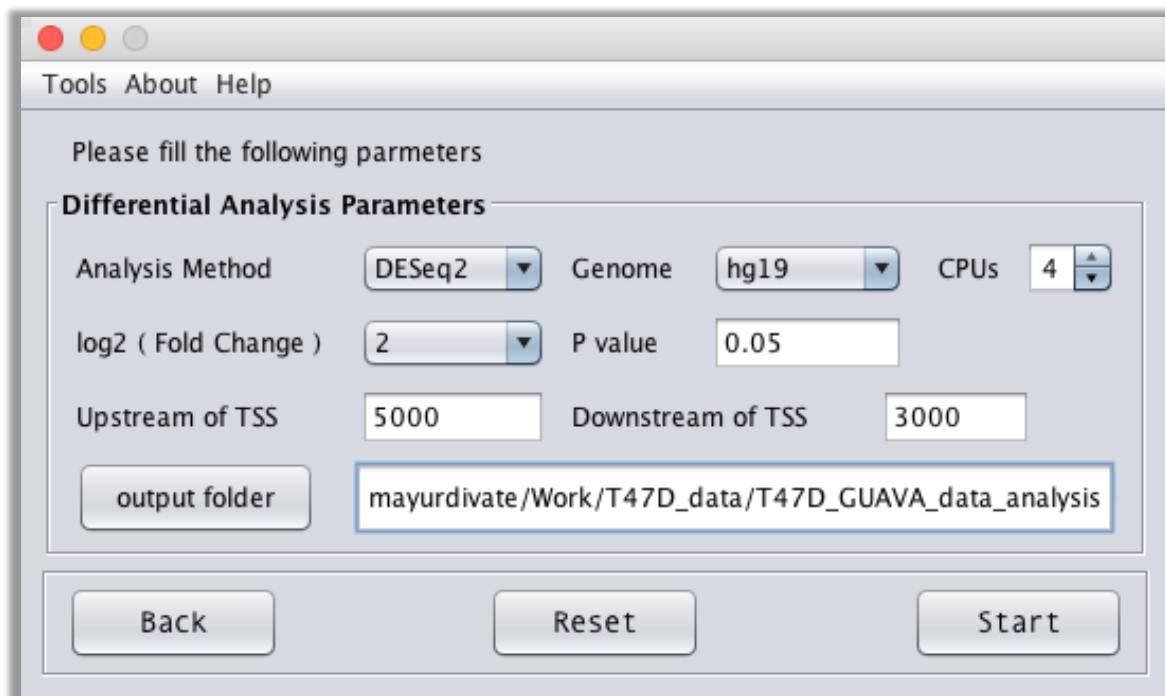


Figure 15. GUAVA ATAC-seq differential analysis input interface 2. Screenshot shows the second input window of the ATAC-seq differential analysis program for setting parameters such as log2 fold change, P value, genome, distance from the TSS, etc.

8.5 Output interface of ATAC-seq differential analysis program

Once differential analysis has finished, GUAVA will show the results as a tabular output interface (Figures 16-23). GUAVA also facilitates the visualization of ATAC-seq signals on the IGV browser.

8.5.1 Summary

This tab provides a summary of the input parameters and files that were used to run differential analysis.

Parameter	value
Project Name	T47D
Genome build	hg19
Analysis Method	DESeq2
P value	0.001
Fold change	2.0
Upstream	5000
Downstream	3000
Control_REPO_I	SRR3929040_DMSO_Rep1_R1_aligned_ATACseq.bam
Control_REPO_I	SRR3929040_DMSO_Rep1_R1_peaks.narrowPeak
Control_REPO_II	SRR3929041_DMSO_Rep2_R1_aligned_ATACseq.bam
Control_REPO_II	SRR3929041_DMSO_Rep2_R1_peaks.narrowPeak
Treatment_REPO_I	SRR3929042_BYL719_Rep1_R1_aligned_ATACseq.bam
Treatment_REPO_I	SRR3929042_BYL719_Rep1_R1_peaks.narrowPeak
Treatment_REPO_II	SRR3929043_BYL719_Rep2_R1_aligned_ATACseq.bam
Treatment_REPO_II	SRR3929043_BYL719_Rep2_R1_peaks.narrowPeak

Figure 16. Input summary. Screenshot showing a summary of the parameters and input files used in the differential analysis.

8.5.2 Volcano Plot

This graph shows the summary of differential analysis. The red and green colors indicate peaks with reduced and increased chromatin accessibility, respectively.

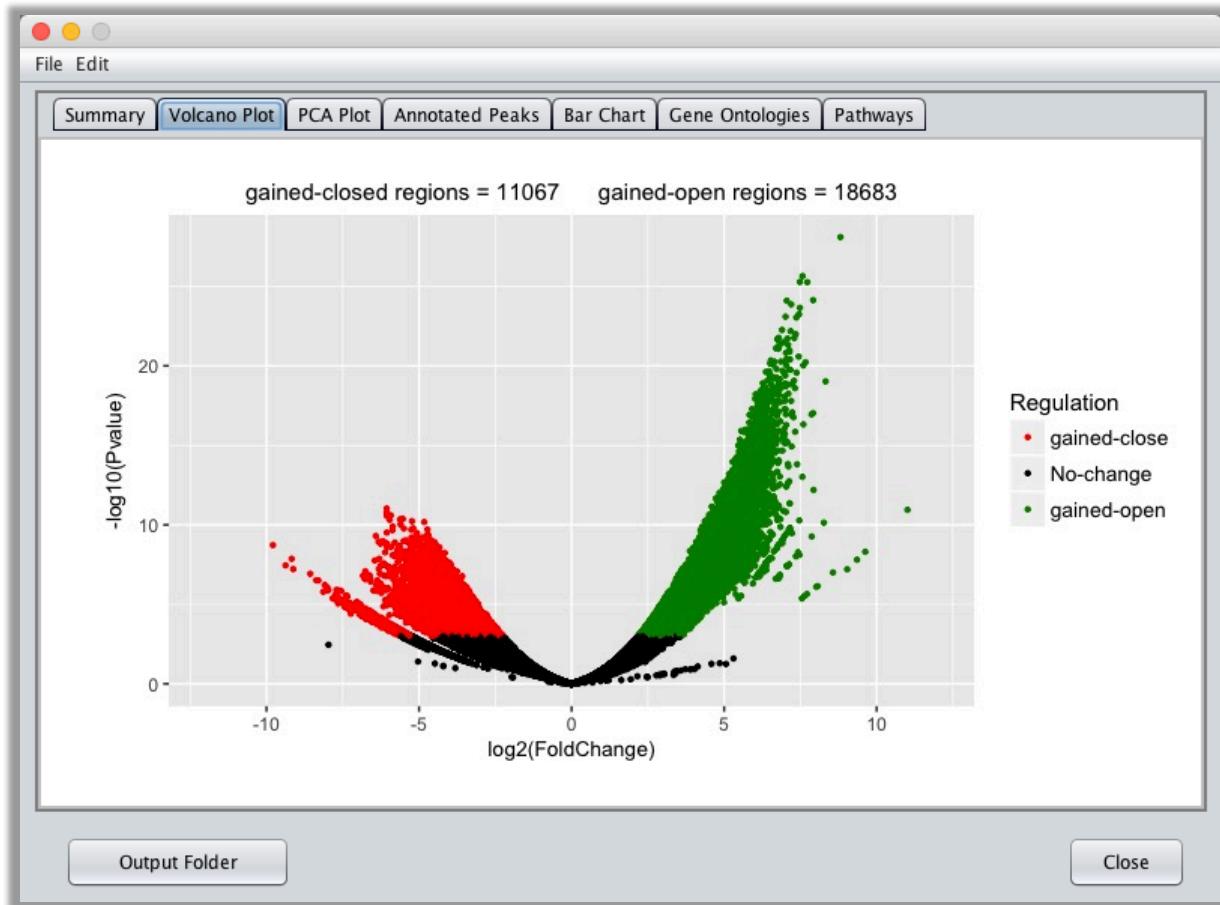


Figure 17. A volcano plot showing the differential ATAC-seq signals.

8.5.3 PCA Plot

This graph shows the principal component analysis (PCA) of the samples used in the differential analysis.

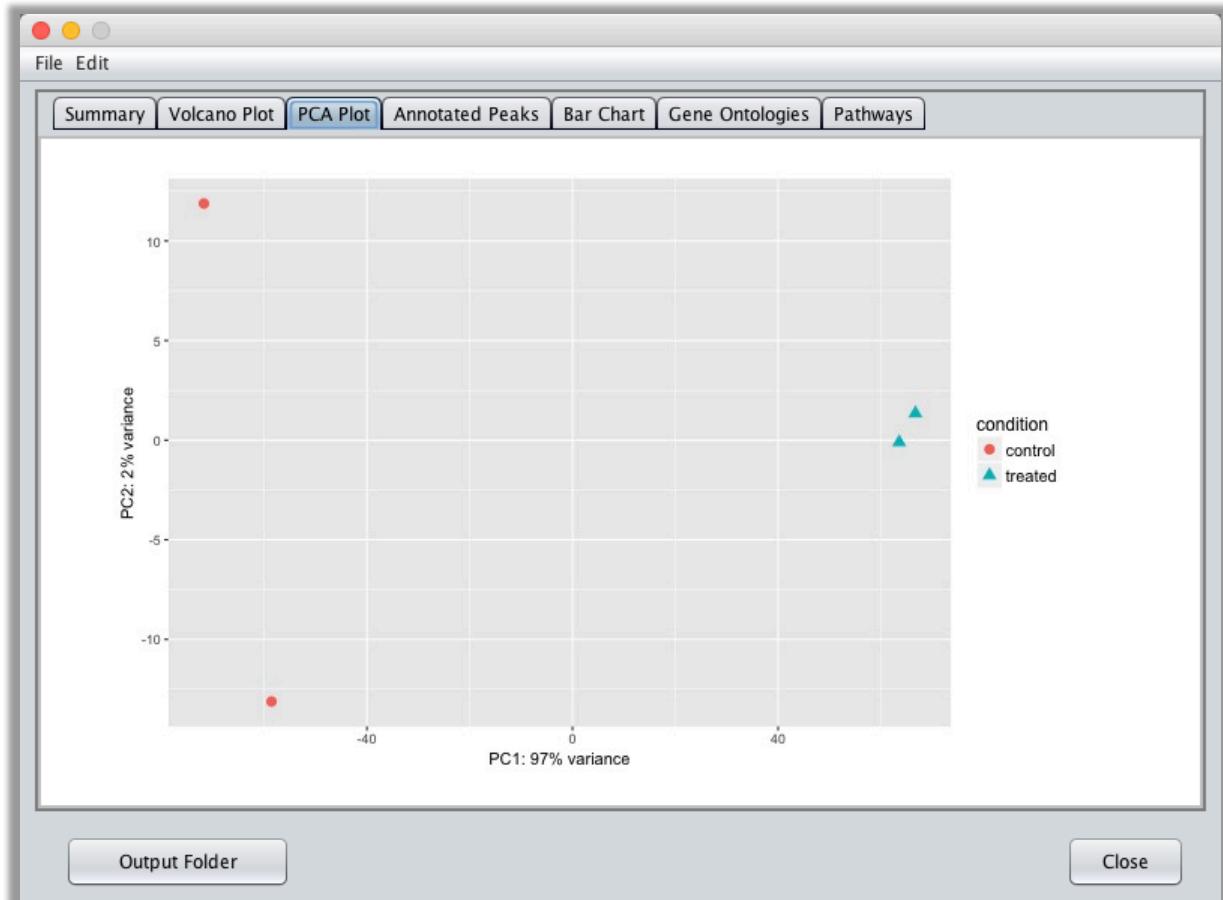
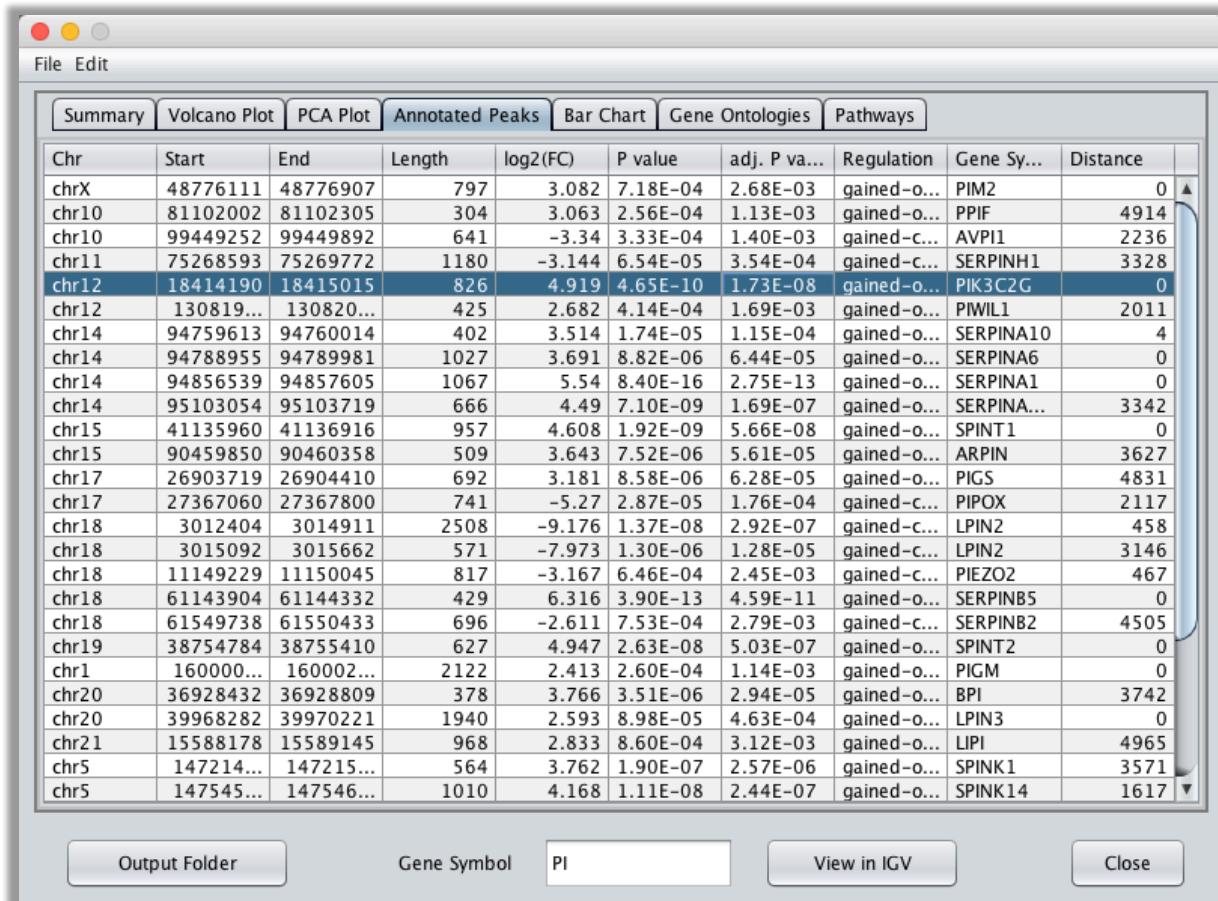


Figure 18. A PCA plot illustrating the variance between the control and treated samples.

8.5.4 Annotated Peaks

This tab provides a table with differentially enriched peaks and easy access to visualize ATAC-seq signals from control and treatment samples.



The screenshot shows a software interface with a title bar "File Edit". Below it is a tab bar with "Summary", "Volcano Plot", "PCA Plot", "Annotated Peaks" (which is selected and highlighted in blue), "Bar Chart", "Gene Ontologies", and "Pathways". The main area is a table with the following columns:

Chr	Start	End	Length	log2(FC)	P value	adj. P va...	Regulation	Gene Sy...	Distance
chrX	48776111	48776907	797	3.082	7.18E-04	2.68E-03	gained-o...	PIM2	0
chr10	81102002	81102305	304	3.063	2.56E-04	1.13E-03	gained-o...	PPIF	4914
chr10	99449252	99449892	641	-3.34	3.33E-04	1.40E-03	gained-c...	AVP11	2236
chr11	75268593	75269772	1180	-3.144	6.54E-05	3.54E-04	gained-c...	SERPINH1	3328
chr12	18414190	18415015	826	4.919	4.65E-10	1.73E-08	gained-o...	PIK3C2G	0
chr12	130819...	130820...	425	2.682	4.14E-04	1.69E-03	gained-o...	PIWI1	2011
chr14	94759613	94760014	402	3.514	1.74E-05	1.15E-04	gained-o...	SERPINA10	4
chr14	94788955	94789981	1027	3.691	8.82E-06	6.44E-05	gained-o...	SERPINA6	0
chr14	94856539	94857605	1067	5.54	8.40E-16	2.75E-13	gained-o...	SERPINA1	0
chr14	95103054	95103719	666	4.49	7.10E-09	1.69E-07	gained-o...	SERPINA...	3342
chr15	41135960	41136916	957	4.608	1.92E-09	5.66E-08	gained-o...	SPINT1	0
chr15	90459850	90460358	509	3.643	7.52E-06	5.61E-05	gained-o...	ARPIN	3627
chr17	26903719	26904410	692	3.181	8.58E-06	6.28E-05	gained-o...	PIGS	4831
chr17	27367060	27367800	741	-5.27	2.87E-05	1.76E-04	gained-c...	PIPOX	2117
chr18	3012404	3014911	2508	-9.176	1.37E-08	2.92E-07	gained-c...	LPIN2	458
chr18	3015092	3015662	571	-7.973	1.30E-06	1.28E-05	gained-c...	LPIN2	3146
chr18	11149229	11150045	817	-3.167	6.46E-04	2.45E-03	gained-c...	PIEZQ2	467
chr18	61143904	61144332	429	6.316	3.90E-13	4.59E-11	gained-o...	SERPINB5	0
chr18	61549738	61550433	696	-2.611	7.53E-04	2.79E-03	gained-c...	SERPINB2	4505
chr19	38754784	38755410	627	4.947	2.63E-08	5.03E-07	gained-o...	SPINT2	0
chr1	160000...	160002...	2122	2.413	2.60E-04	1.14E-03	gained-o...	PIGM	0
chr20	36928432	36928809	378	3.766	3.51E-06	2.94E-05	gained-o...	BPI	3742
chr20	39968282	39970221	1940	2.593	8.98E-05	4.63E-04	gained-o...	LPIN3	0
chr21	15588178	15589145	968	2.833	8.60E-04	3.12E-03	gained-o...	LIP1	4965
chr5	147214...	147215...	564	3.762	1.90E-07	2.57E-06	gained-o...	SPINK1	3571
chr5	147545...	147546...	1010	4.168	1.11E-08	2.44E-07	gained-o...	SPINK14	1617

At the bottom of the window, there are four buttons: "Output Folder", "Gene Symbol" (with a search input field containing "PI"), "View in IGV", and "Close".

Figure 19. Annotation of the differentially enriched ATAC-seq signals. Differentially enriched peaks associated with a particular gene can be searched by typing the gene symbol in the search box provided at the bottom.

8.5.5 Visualization of ATAC-seq signals from the control and treatment samples using IGV

To visualize ATAC-seq signals from the control and treatment samples, select a differentially enriched peak from the ‘Annotated Peaks’ tab and then click on the ‘View in IGV’ button. This will automatically load normalized signals from all the input samples on the IGV browser.

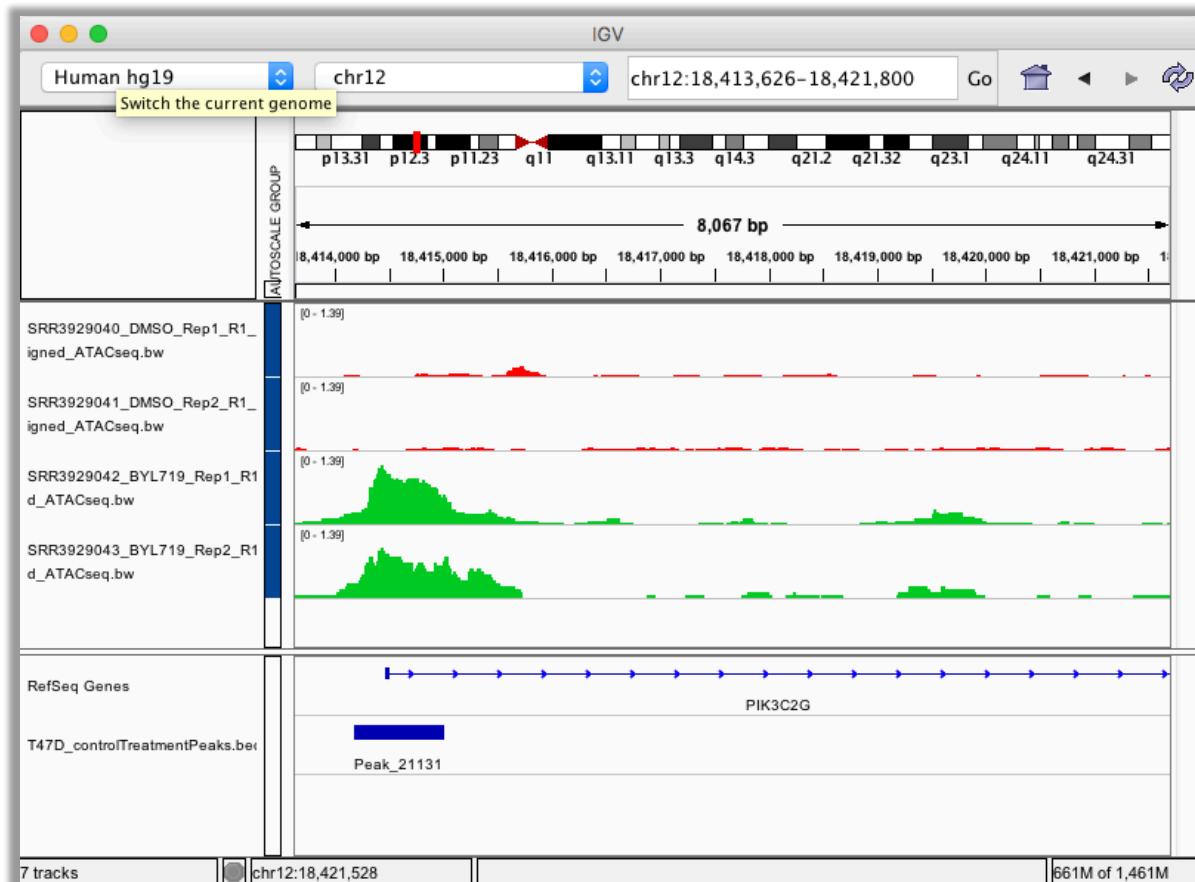


Figure 20. ATAC-seq signal visualization using IGV.

8.5.6 Bar Chart

This tab contains a bar chart that illustrates the distribution of the annotated peaks in various genomic locations such as the promoter, intron, exon, UTR, etc.

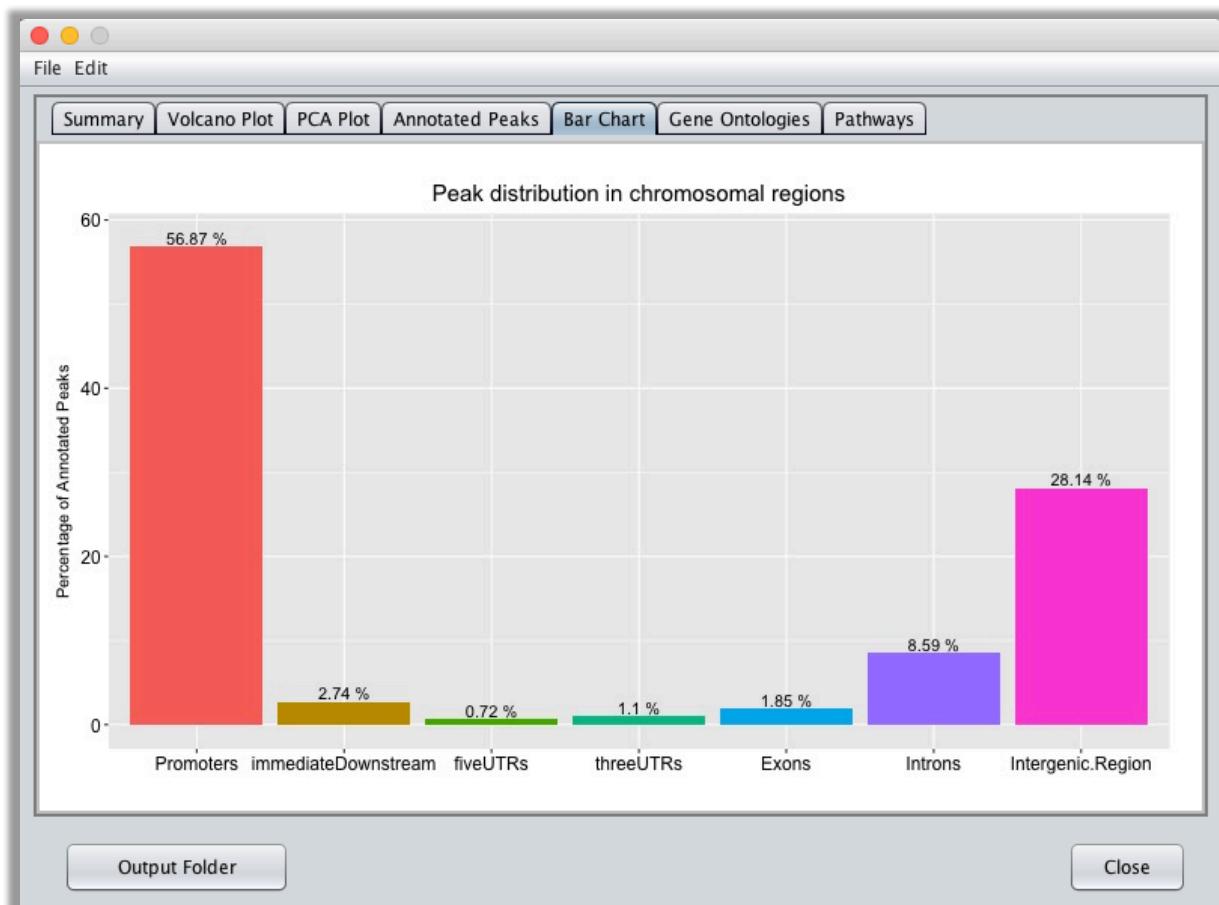


Figure 21. Bar chart showing the distribution of the differentially enriched ATAC-seq peaks in various genomic regions such as promoters, introns, exons, etc.

8.5.7 Gene Ontologies

This tab shows the list of the over-represented gene ontologies associated with differentially enriched peaks.

The table displays the following data:

GO ID	GO Term	Type	Pvalue	adj. P value	Gene Symbol
GO:0014910	regulation of smooth muscle cell ...	BP	4.40E-02	6.41E-01	SORL1; NFE2L2; ADIPOQ; DOCK7...
GO:0019372	lipoxygenase pathway	BP	2.04E-02	6.20E-01	GPX4; PON1; HPGD; GPX1; PON3
GO:0002934	desmosome organization	BP	1.50E-02	6.02E-01	DSP; DSG2; PKP2; JUP
GO:0032101	regulation of response to externa...	BP	4.45E-02	6.41E-01	SERpine1; NR1H4; CD55; CLU; A...
GO:0030166	proteoglycan biosynthetic process	BP	5.25E-03	4.77E-01	NDST2; HS3ST3B1; DSEL; NDST1...
GO:0070861	regulation of protein exit from en...	BP	2.76E-02	6.20E-01	SORL1; YOD1; DERL3; SVIP; SEC1...
GO:0046879	hormone secretion	BP	2.06E-04	1.77E-01	UCP2; TACR1; CPT1A; HIF1A; AD...
GO:0007263	nitric oxide mediated signal trans...	BP	2.76E-02	6.20E-01	NDNF; NEUROD1; ATP2B4; NOS1...
GO:0007266	Rho protein signal transduction	BP	1.09E-04	1.35E-01	ITGA3; VAV1; PLEKHG4; RHOD; ...
GO:0035609	C-terminal protein deglutamylat...	BP	9.73E-03	5.36E-01	FOLH1; AGTPBP1; AGBL4
GO:0043122	regulation of I-kappaB kinase/NF...	BP	1.52E-03	3.10E-01	ZDHHC17; ESR1; SLC20A1; TLE1...
GO:0045785	positive regulation of cell adhesion	BP	1.72E-02	6.20E-01	CCDC88B; ANK3; IL1B; SYK; TGF...
GO:0007267	cell-cell signaling	BP	3.19E-02	6.34E-01	TNKS2; HSPA8; P2RX6; RFX6; A...
GO:0007264	small GTPase mediated signal tr...	BP	4.50E-02	6.41E-01	IGF1; LPAR4; ARL11; ARHGAP40;...
GO:0043124	negative regulation of I-kappaB k...	BP	1.56E-02	6.11E-01	RIPK1; CARD19; TMSB4X; ZC3H1...
GO:0007265	Ras protein signal transduction	BP	2.74E-02	6.20E-01	ARHGEF26; PLEKHG4; ADRA2A; ...
GO:0043123	positive regulation of I-kappaB ki...	BP	3.16E-03	4.22E-01	ADIPOQ; TRADD; CASP1; IL18; LP...
GO:0035608	protein deglutamylation	BP	4.28E-02	6.41E-01	FOLH1; AGBL4; AGTPBP1
GO:0010560	positive regulation of glycoprotei...	BP	2.21E-03	3.62E-01	PLCB1; PAWR; SOAT1; IGF1; ALG...
GO:0019367	fatty acid elongation, saturated f...	BP	2.90E-02	6.20E-01	ELOVL7; ELOVL4; ELOVL1
GO:0019368	fatty acid elongation, unsaturated...	BP	2.90E-02	6.20E-01	ELOVL7; ELOVL1; ELOVL4
GO:0046887	positive regulation of hormone se...	BP	4.09E-02	6.41E-01	HIF1A; FOXL2; TRH; PFKFB2; TAC...
GO:0046888	negative regulation of hormone s...	BP	3.59E-02	6.41E-01	IL1B; EDN1; RAB11FIP1; ADRA2A; ...
GO:0048709	oligodendrocyte differentiation	BP	2.99E-02	6.20E-01	FA2H; HDAC11; OLIG1; NKX6-1; ...
GO:0007270	neuron-neuron synaptic transmis...	BP	4.28E-02	6.41E-01	VDAC1; PTEN; DRD2
GO:0007156	homophilic cell adhesion via plas...	BP	7.97E-08	3.59E-04	PCDHGA10; PCDH15; PCDHAS; ...

Figure 22. Over represented gene ontology terms associated with differentially enriched peaks.

8.5.8 Pathway

This tab shows the list of the over-represented KEGG pathways associated with differentially enriched peaks.

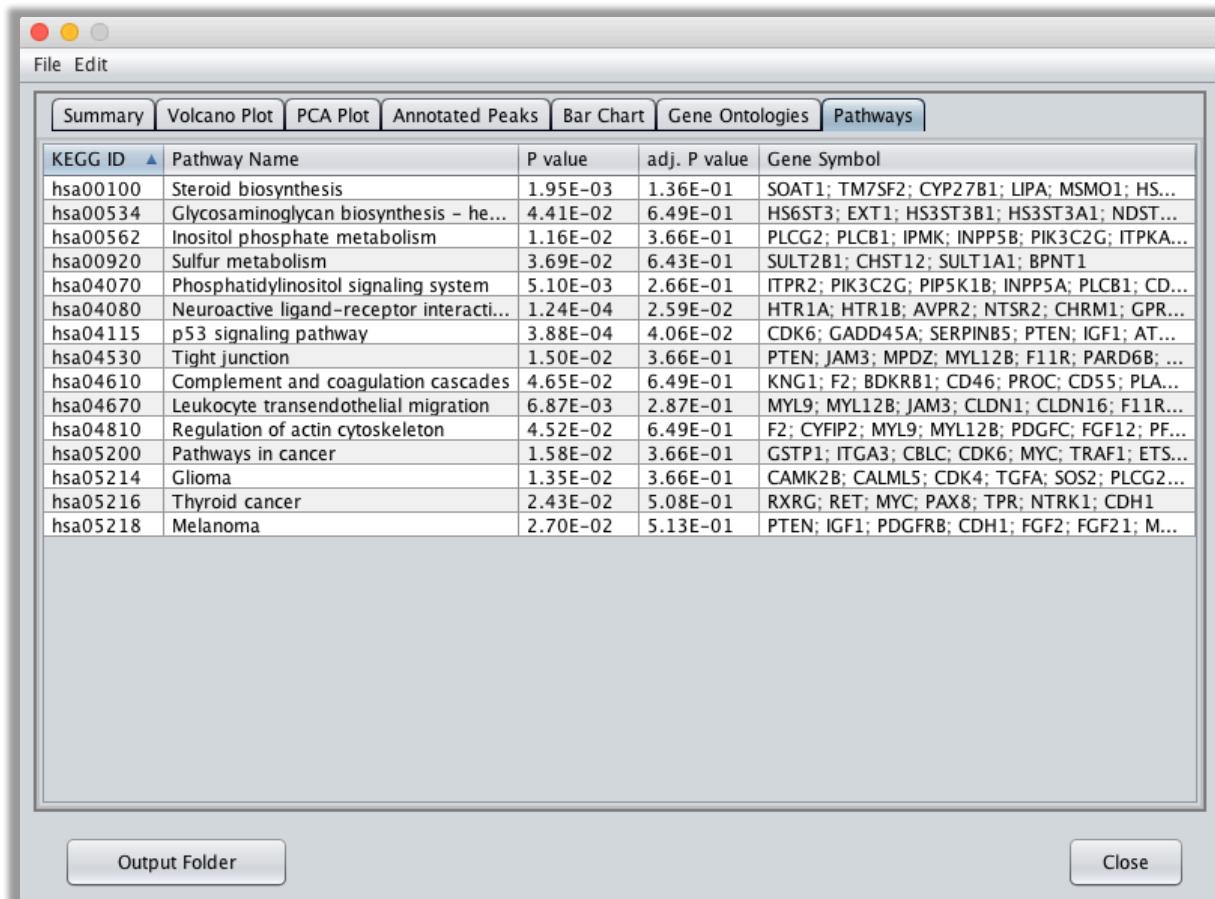


Figure 23. Over represented KEGG pathways associated with differentially enriched peaks.

NOTE: The above results are stored in the output folder. To access output folder, click on the 'Output Folder' button located at the bottom-left corner.