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 conditions can be compared using GUAVA to identify genomic loci with differentially enriched ATAC-seq signals. Furthermore, GUAVA also provides gene ontology and pathways 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.

Download

Step 1: Go to the link: https://github.com/MayurDivate/GUAVA/releases
Step 2: Click on the `Source Code (zip)`
Step 3: This will save GUAVA zip package in the downloads folder

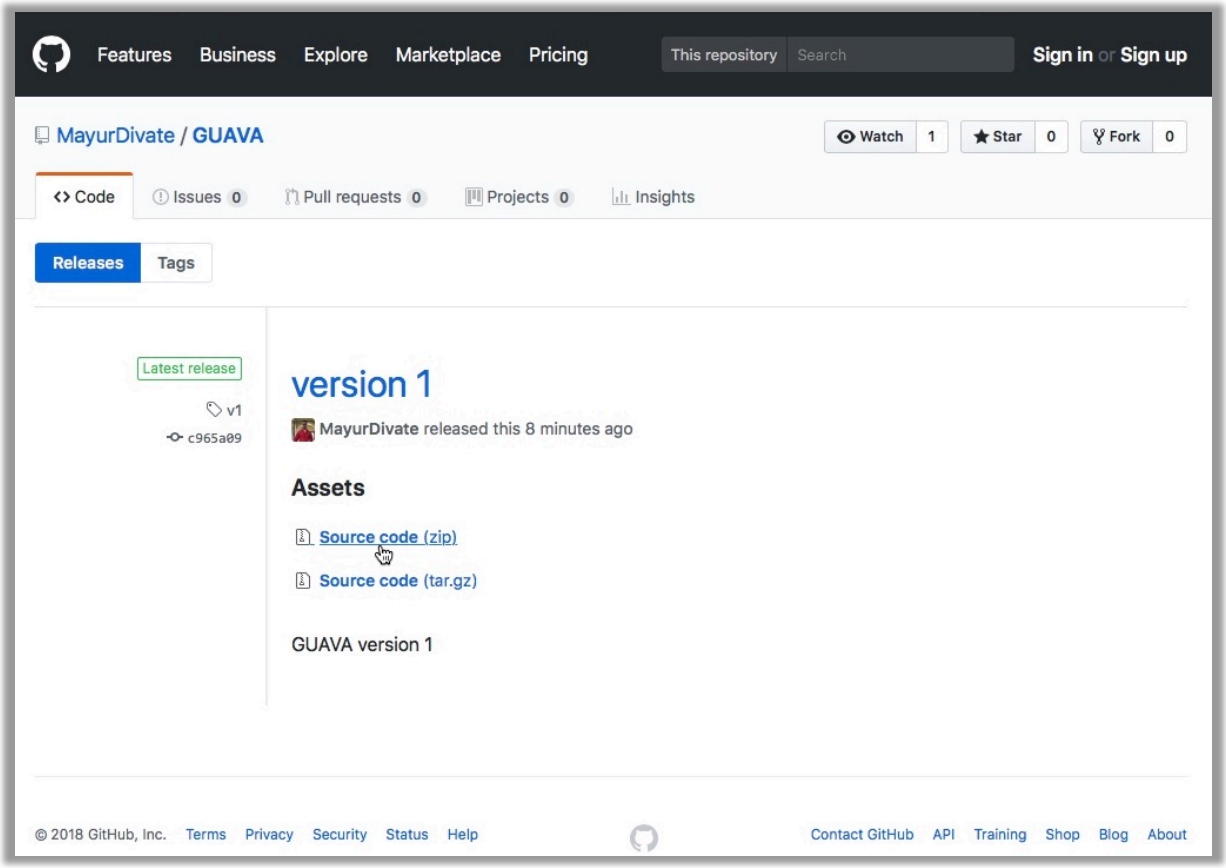


Figure1: GUAVA - GitHub project page

Go to GUAVA project page on GitHub and click on the ‘releases’ option. After that, click on the ‘Source code (zip)’ under the latest release.

If want download source code, follow the link below.
https://github.com/MayurDivate/GUAVASourceCode.

Getting help and reporting issues

If you face difficulties in installing or using GUAVA please report the issue here
https://github.com/MayurDivate/GUAVASourceCode/issues

How to open Terminal?

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

MAC

1. Open the Finder
2. Click on the `Go` in the menu bar then select the `Utilities` folder
3. Double-click on the Terminal icon

Linux

1. Press windows key on keyboard
2. Type “terminal” in the search box
3. Click on the Terminal icon

OR press “**Ctrl** + **Alt** + **T**” this will open then terminal

Install

Let’s us assume that the user wants to install GUAVA in the home folder. Therefore, before proceeding, first copy / move downloaded GUAVA package to the home folder. If the downloaded package is in the folder `Downloads`, then type the following commands to unzip package on the terminal.

```
mv ~/Downloads/GUAVA-1.zip ~/
cd ~/
unzip GUAVA-1.zip
mv GUAVA-v1 GUAVA
```

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

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 not work properly. Therefore, to help users to install dependencies, we have written a program (configure.sh) which automatically downloads and installs dependencies. However, the user need to install R and MACS2 manually due to the technical reasons.

Step 1: Install R

MAC

1. Download R follow this link => https://cran.r-project.org/bin/macosx/
2. Click on the R-X.X.X.pkg file link (e.g. R-3.4.3.pkg)
3. Double click on the downloaded file and follow the instructions

Linux

1. Open the terminal
2. Type command ` sudo apt-get install r-base ` and press enter

To know more about R, follow the link https://cran.r-project.org/bin/linux/. Choose appropriate Linux OS type

Step 2: Install other dependencies

To run configure.sh use the following commands on the terminal.

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

Note: This may take a while to finish. Also, you will need to press 'enter' several times to continue. Additionally, answer all question with 'yes'

Step 3: Install MACS2

Now, to install last dependency MACS2 use following commands on the terminal

```
cd ~/GUAVA
python get-pip.py
pip install MACS2
```

Error: Sometime MACS2 fails to install **Numpy**. In such a situation run `pip install numpy` first then try to install MACS2.

NOTE: If **permission denied,** type 'sudo' at the beginning of the commands. Then, in order to proceed you have enter your password

That is the end of installation part. Now, GUAVA is ready to process ATAC-seq data.

How to start GUAVA?

After successful installation of GUAVA, user can start GUAVA using following command on the terminal. This will open home window of GUAVA where you can choose program to proceed further.

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

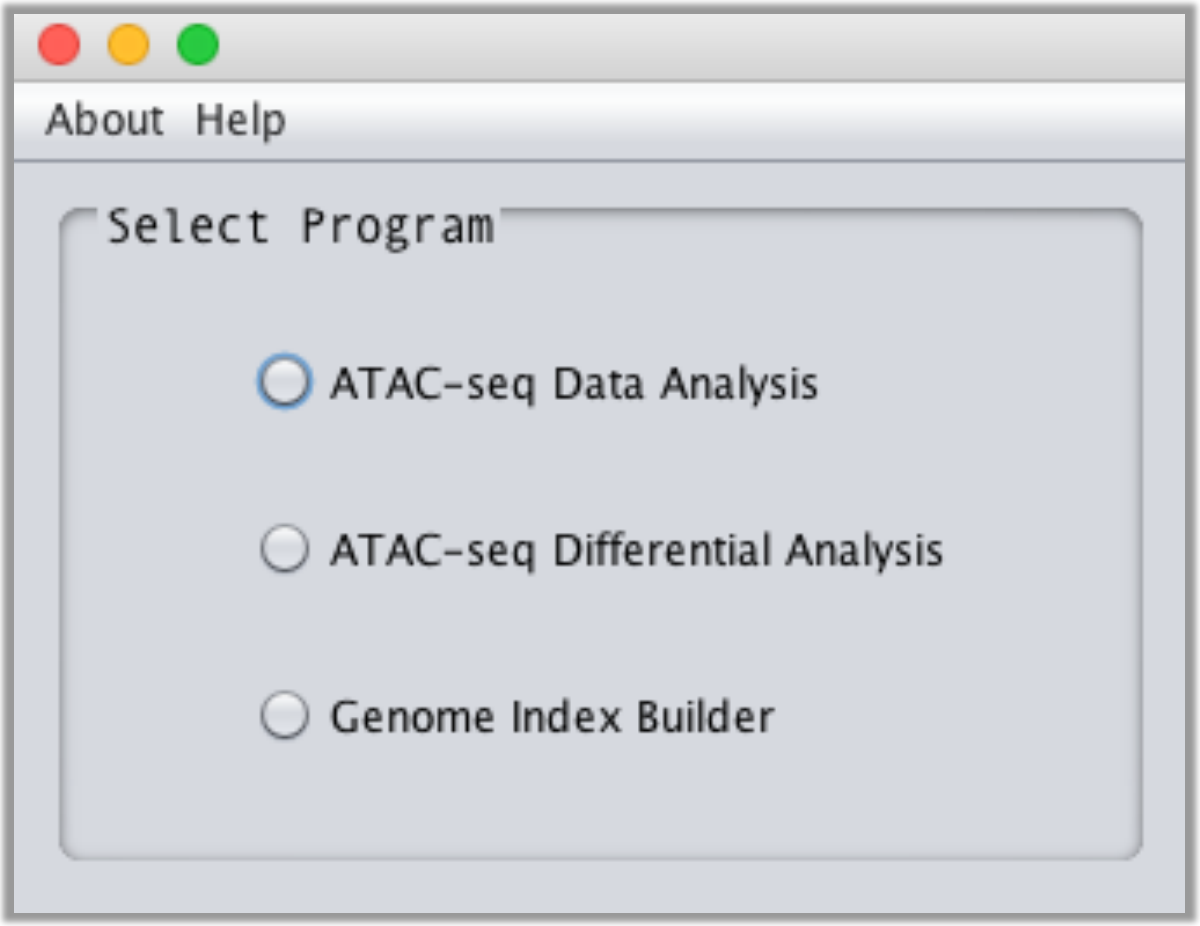


Figure 2: GUAVA – home window

Choose ATAC-seq Data Analysis to process ATAC-seq data, ATAC-seq Differential Analysis to compare ATAC-seq datasets or Genome Index Builder to generate genome index.

Graphical user interface of GUAVA

GUAVA tool has three programs

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

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

3) Genome index builder

When GUAVA GUI is evoked it open GUAVA home window (Figure 2). Then user have to choose one of the above programs to proceed further. Finally, based on the selection of program, the desire input window will be opened (Figure 3, 13 and 22).

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. Additionally, the user can exclude certain chromosomes from the analysis using 'show chromosomes' button (Figure 4). Next, it uses MACS2 to identify ATAC-seq peaks. Finally, it performs functional annotation on the ATAC-seq peaks.

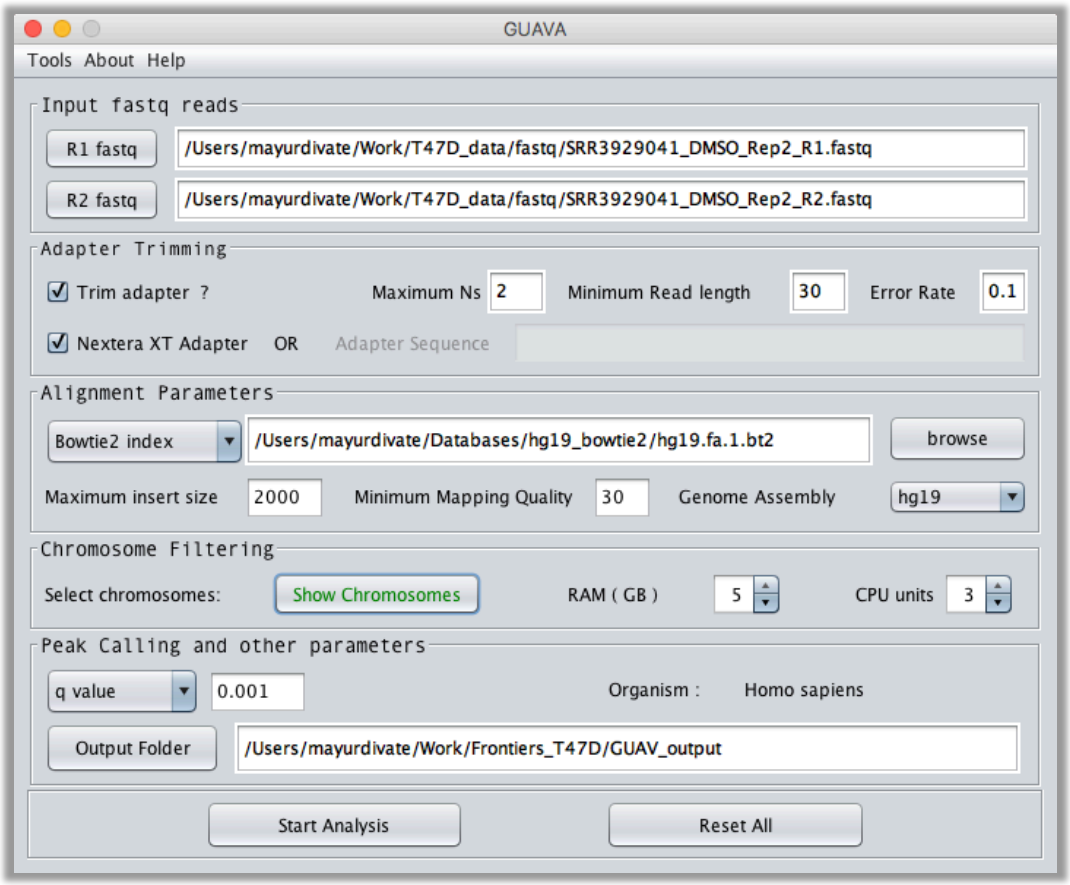


Figure 3. Input form of GUAVA ATAC-seq data analysis program
Input window of ATAC-seq data analysis program to upload input files such as fastq, genome index and set parameters such as insert size, p/q value etc.

R1 fastq and R2 fastq: Buttons to upload read1 and read2 fastq files of ATAC-seq data
Trim adapter: check this option if reads contains adapter
Maximum Ns: If any read contains more than specified number Ns after the adapter trimming, that read pair will be discarded (default 2)
Minimum read length: If any read 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 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: User can select this option if adapter used for ATAC-seq is a Nextera XT adapter (default adapter)
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 the appropriate genome index file (bowtie or bowtie2 index). Please see section 'how to create genome index' to know more about genome index and genome index builder tool.
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 = 30
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.
Show chromosomes: Button to exclude reads mapping to specific chromosomes such mitochondrial chromosome. After clicking this button, it will open a new window (Figure 4) there user can select desired chromosomes.
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.
Start Analysis: click this button to start 'ATAC-seq data analysis' program. If all provided options are valid then GUAVA will start analysis.

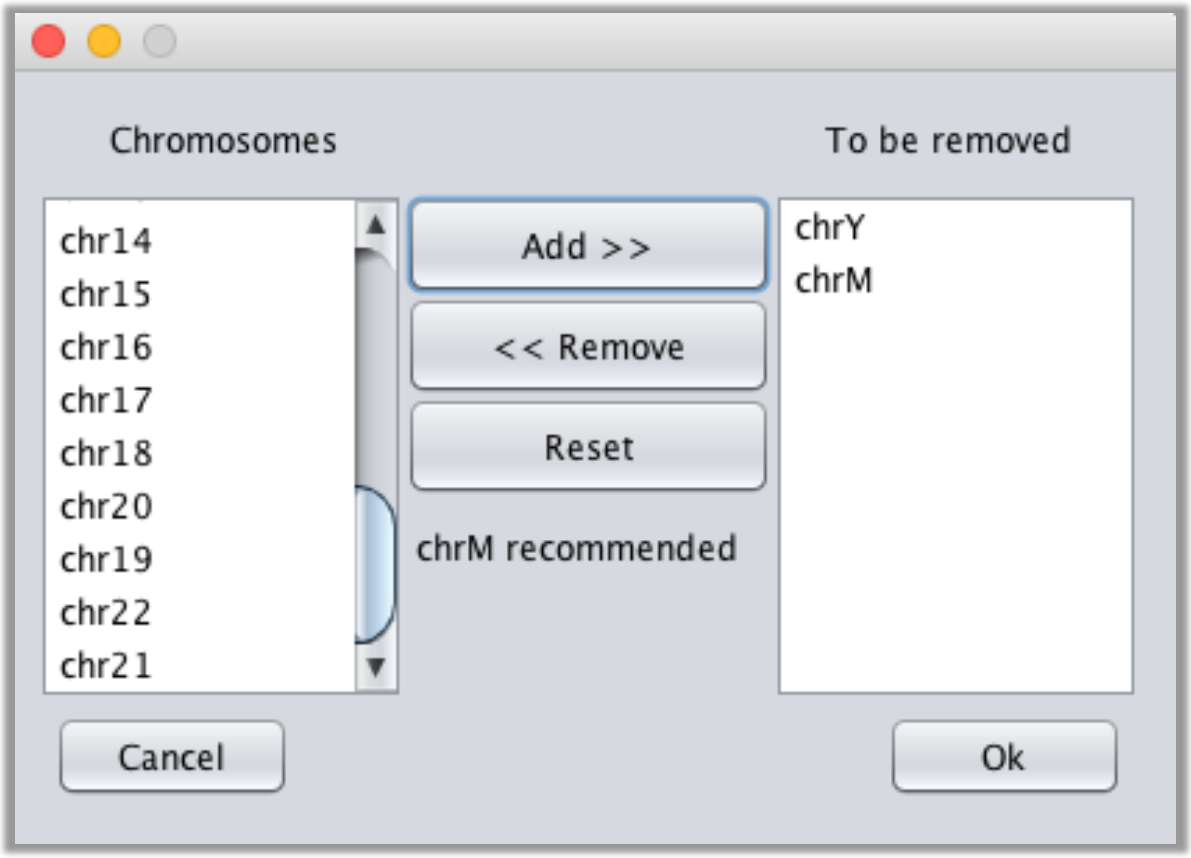


Figure 4. interface to select chromosomes for alignment filtering.

Here user can add desired chromosomes to the ‘to be removed’ list to discard reads aligning to those chromosomes.

Output interface for ATAC-seq data analysis program

Once GUAVA finishes analysis it shows results on tabular output interface (Figure 5-12). Also facilitates the visualization of ATAC-seq signal on IGV browser.

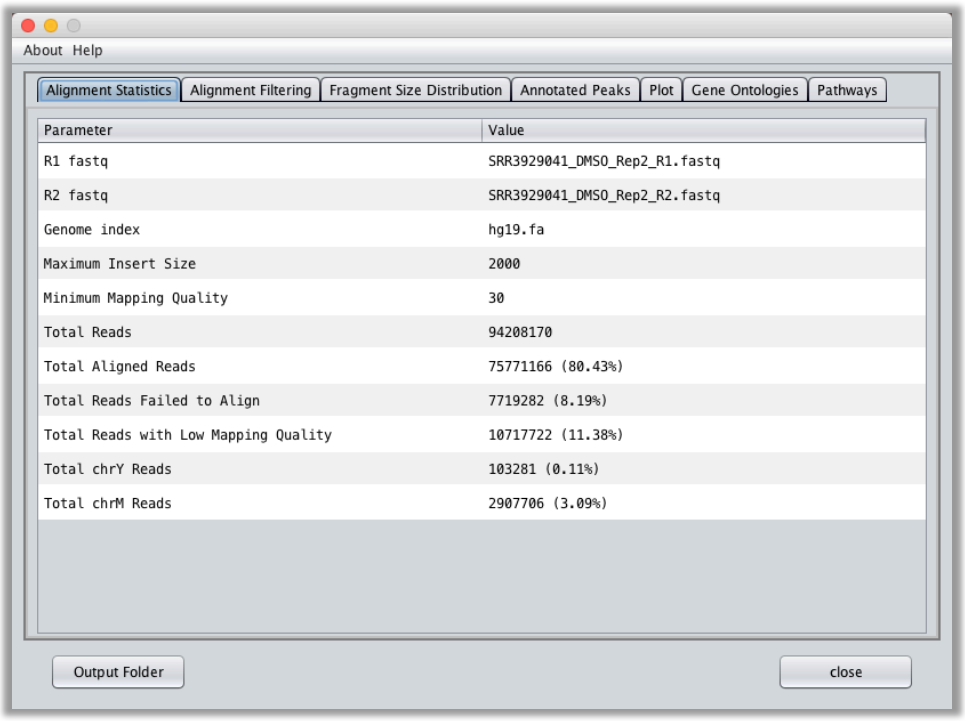


Figure 5: Input summary and alignment statistics This tab provides reads mapping statistics (e.g. total number of reads mapped to genome along) with summary of input files and parameters

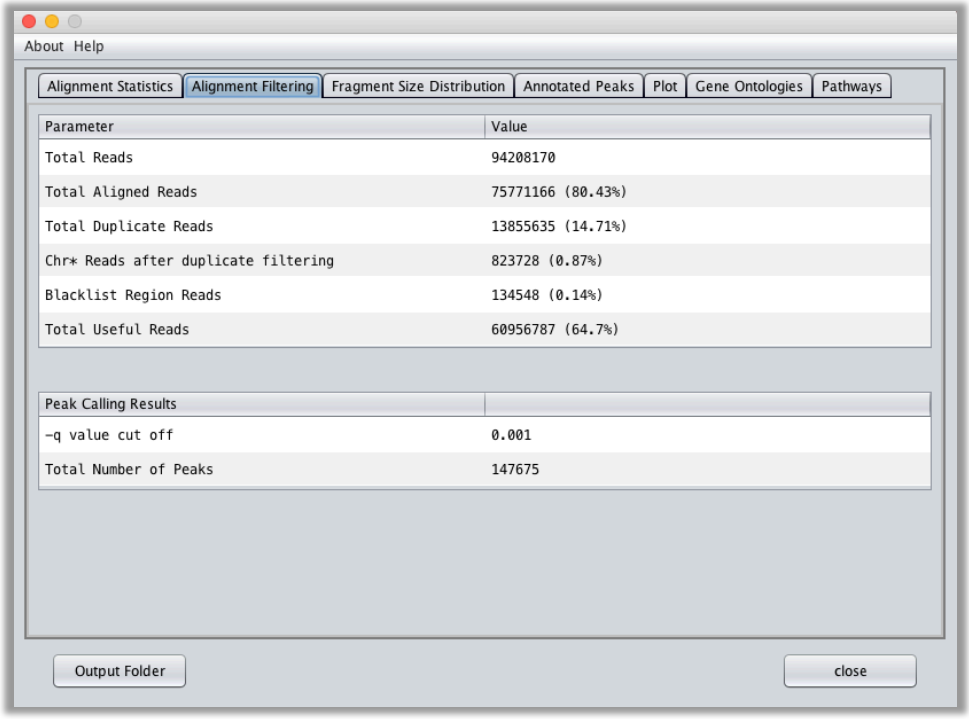
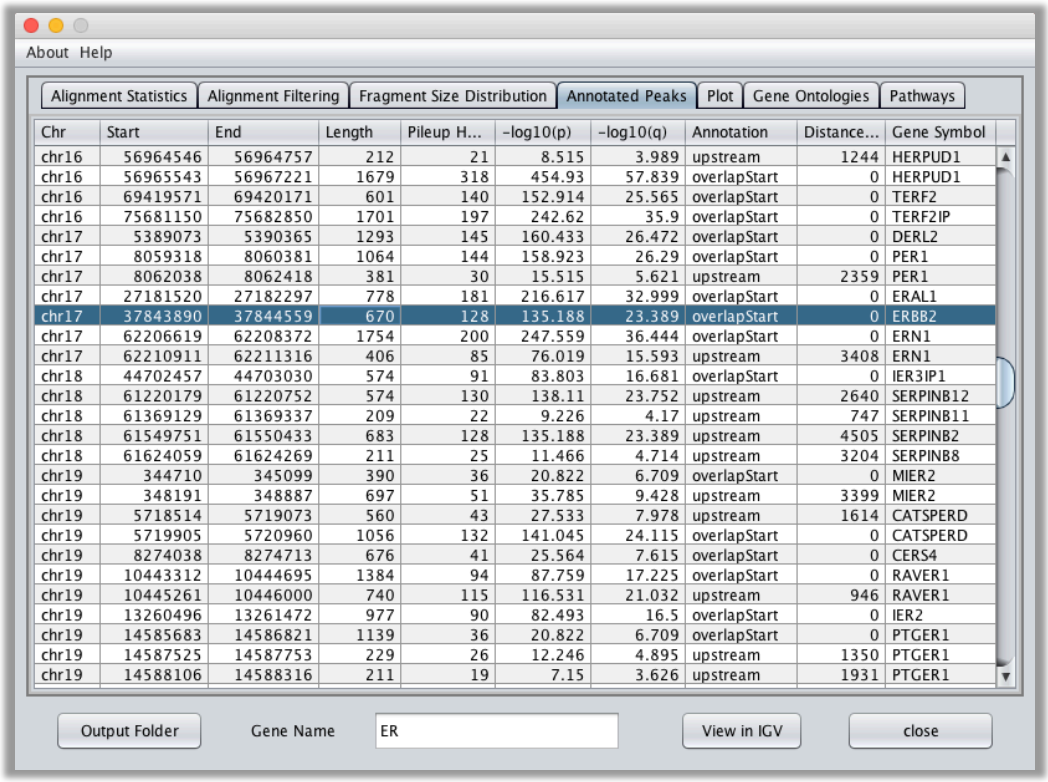


Figure 6: Read filtering and peak calling summary. This tab has two tables 1) to provide alignment filtering statistics e.g. duplicate, useful reads etc. 2) to provide summary of MACS2 peak calling



Chr	Start	End	Length	Pileup H...	-log10(p)	-log10(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	10443312	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

Figure 7: Annotation table. This window exhibits annotated peaks. It contains peak location, nearest gene and distance to TSS. This window also provides, easy access to IGV for visualizing peaks and automatically generated normalized ATAC-seq signal by GUAVA. One can search peak by symbol of the nearest gene using search box provided at the bottom.

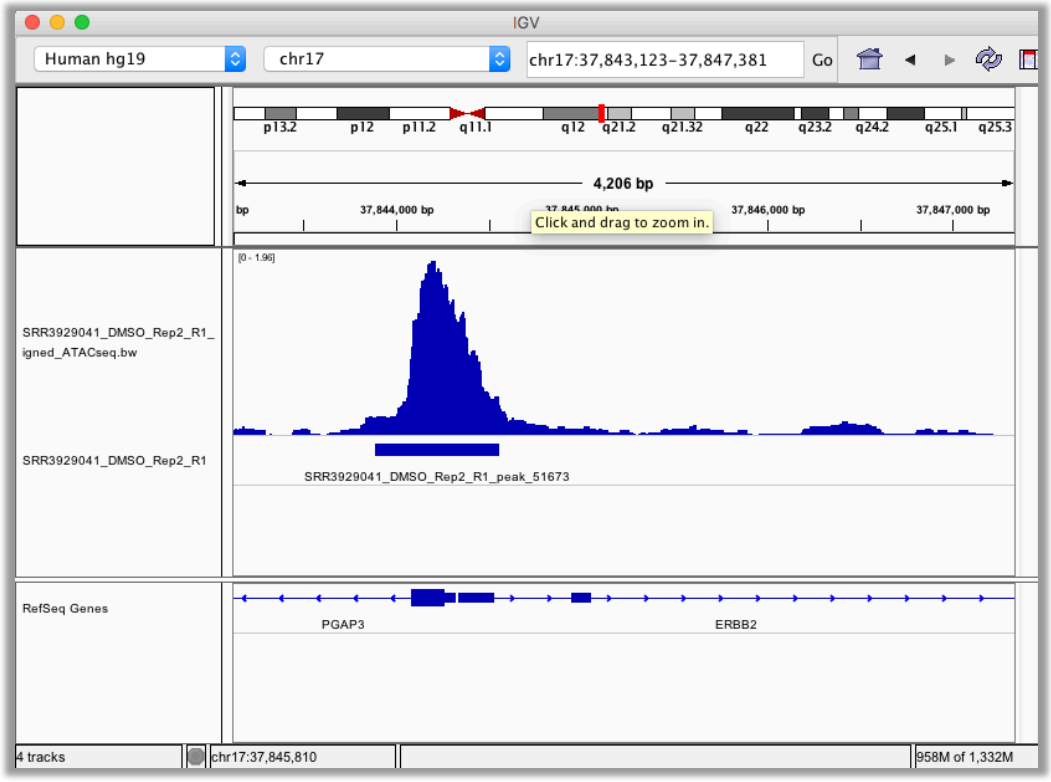


Figure 8: Visualization of ATAC-seq peaks with IGV.

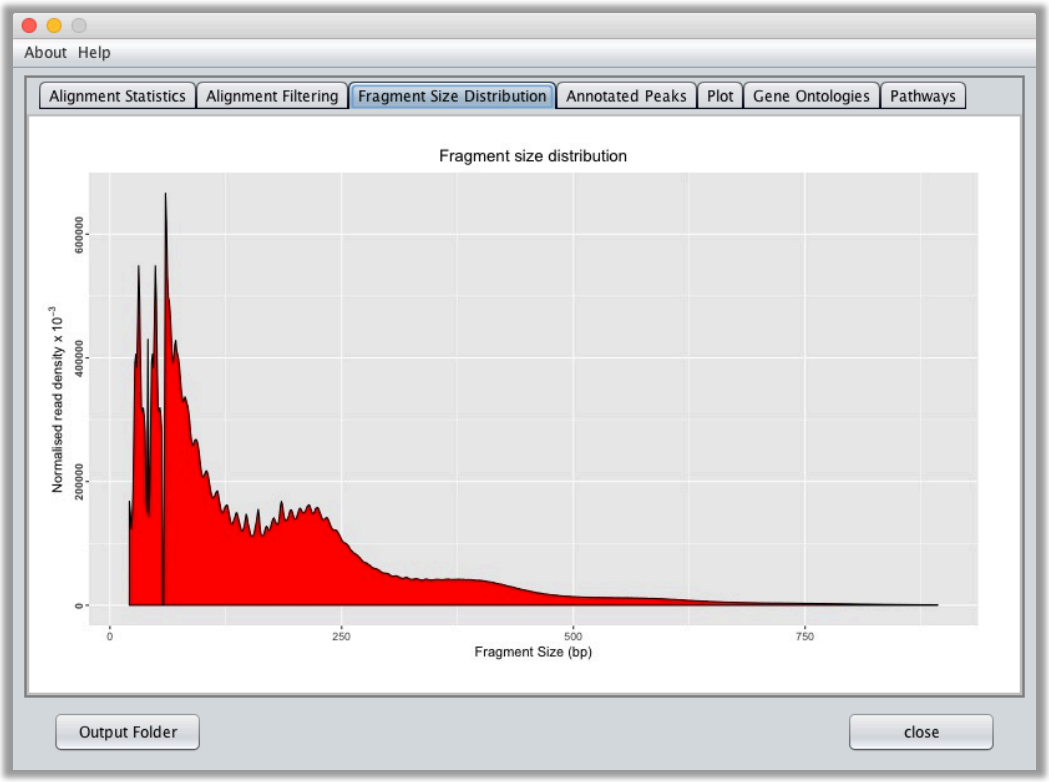


Figure 9: Graph showing the fragment size distribution.

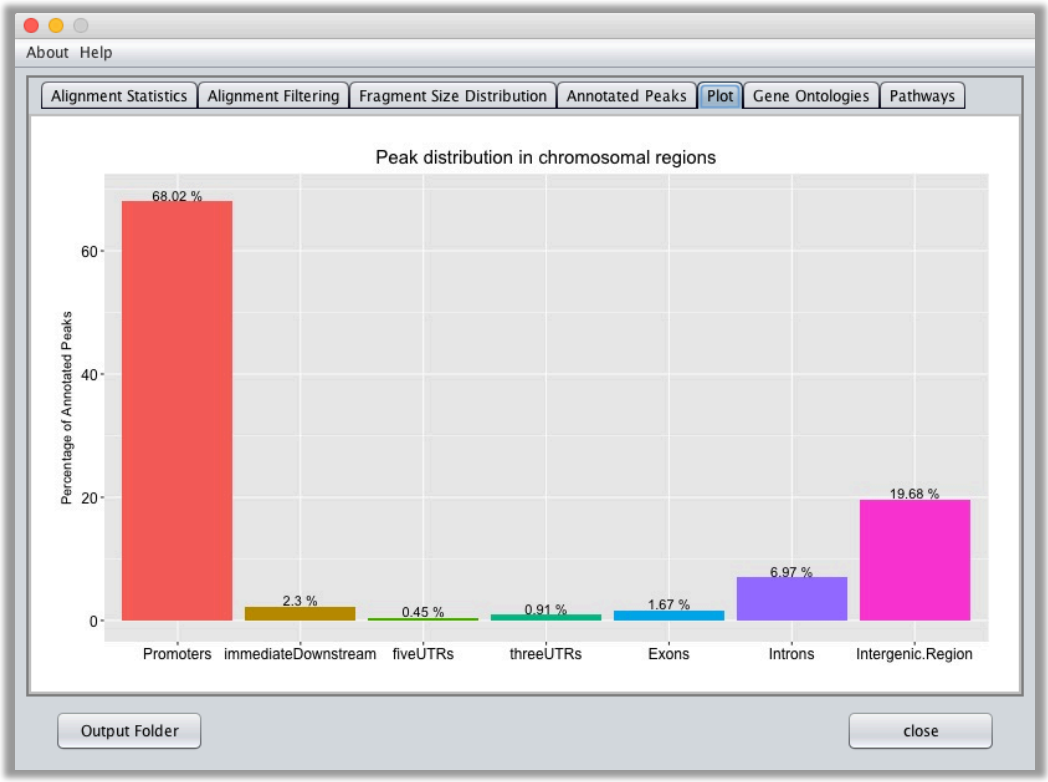
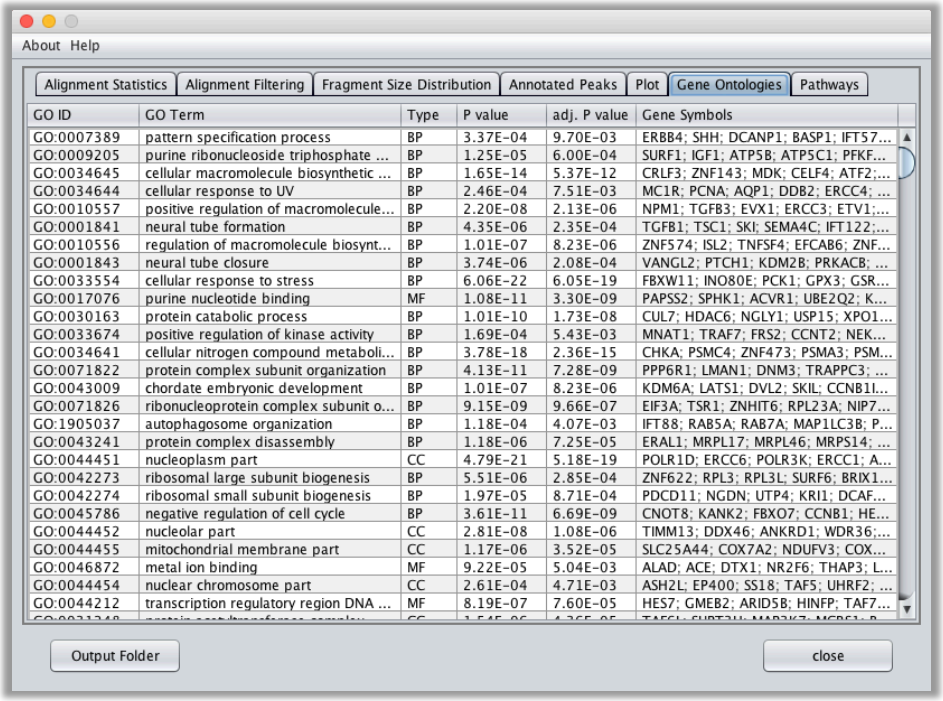
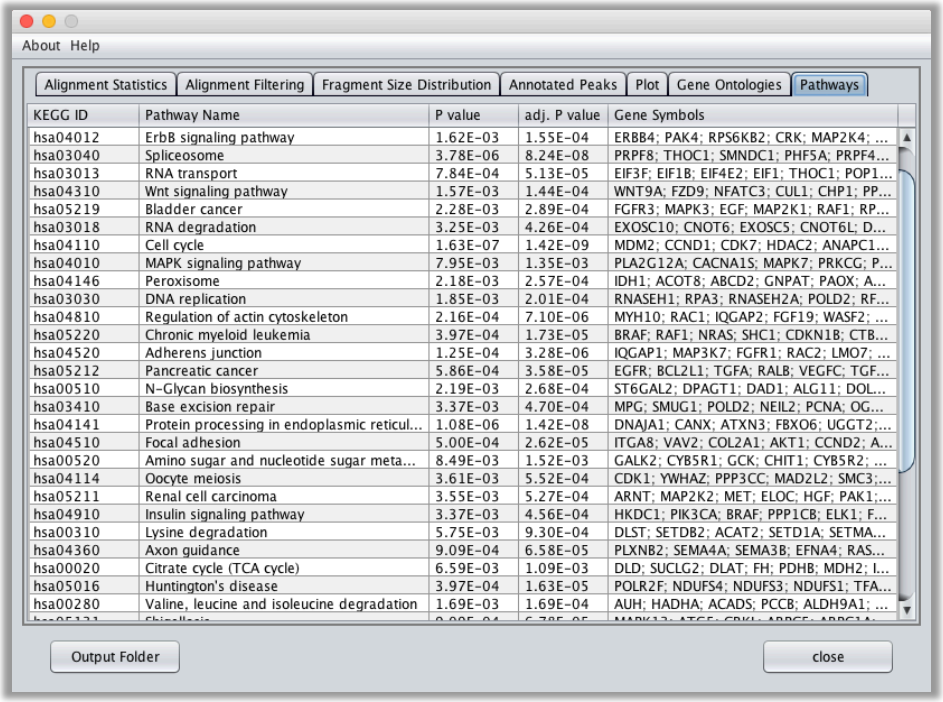


Figure 10: Bar chart showing the percentage of peaks in various genomic locations such as promoter, intron, exon, UTR, etc.



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	ZNF574; 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; TRAPPC3; ...
GO:0043009	chordate embryonic development	BP	1.01E-07	8.23E-06	KDM6A; LATS1; DVL2; SKIL; CCNB1I...
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...

Figure 11: Over represented gene ontology terms.



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	FGFR3; 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; ...

Figure 12: Over represented pathways.

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 13 and 14). Second window allows you to specify differential analysis related parameters e.g. fold change (Figure 14).

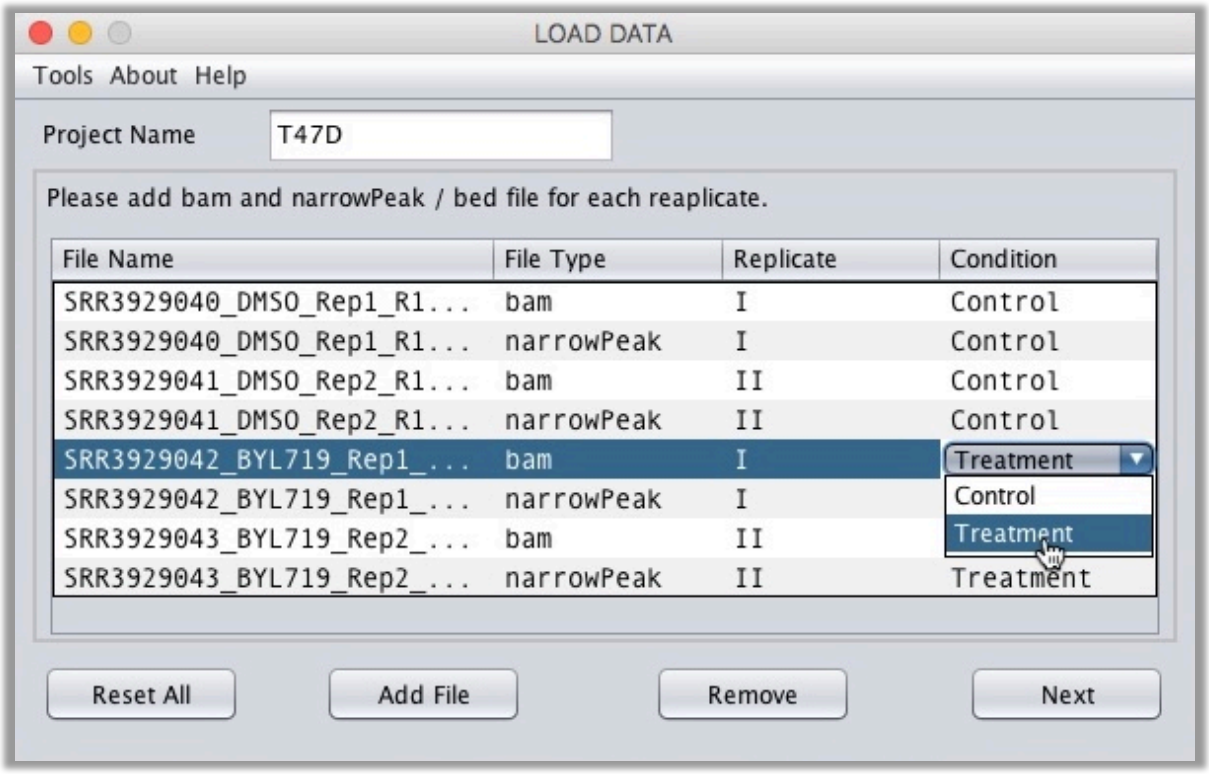


Figure 13: GUAVA ATAC-seq differential analysis input interface 1

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. Click ‘Next’ to specify differential analysis parameters.

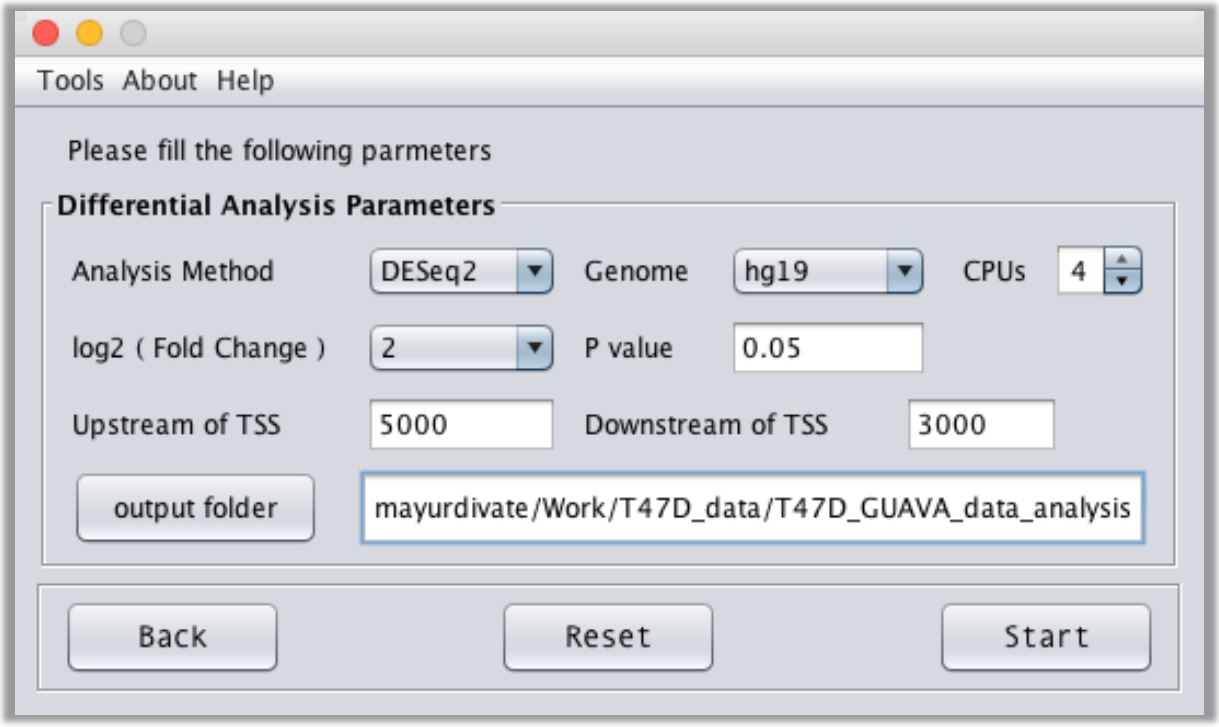


Figure 14: GUAVA ATAC-seq differential analysis input interface 2
Once you have filled all the required parameters click 'Start' button to run differential analysis.

Analysis method: select method for differential analysis 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 5000.
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 3000.
Output folder: Select folder to save GUAVA differential analysis results.

Output interface of ATAC-seq differential analysis program

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 15: Input summary.

Figure 16: 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.

Figure 17: 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.

Figure 18: PCA plot

Figure 19: Peak visualization in IGV.

Figure 20: over-represented gene ontologies and

Figure 21: over-represented pathways.

How to download genome fasta file?

Fasta is a flat 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. For users who don't know where they can use the UCSC link given below and choose desired organism to download fasta file,

UCSC link: http://hgdownload.soe.ucsc.edu/downloads.html

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 genome index from 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 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 genome 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, then use Genome index builder program to create genome index. This program just takes genome fasta file and output folder as an input.

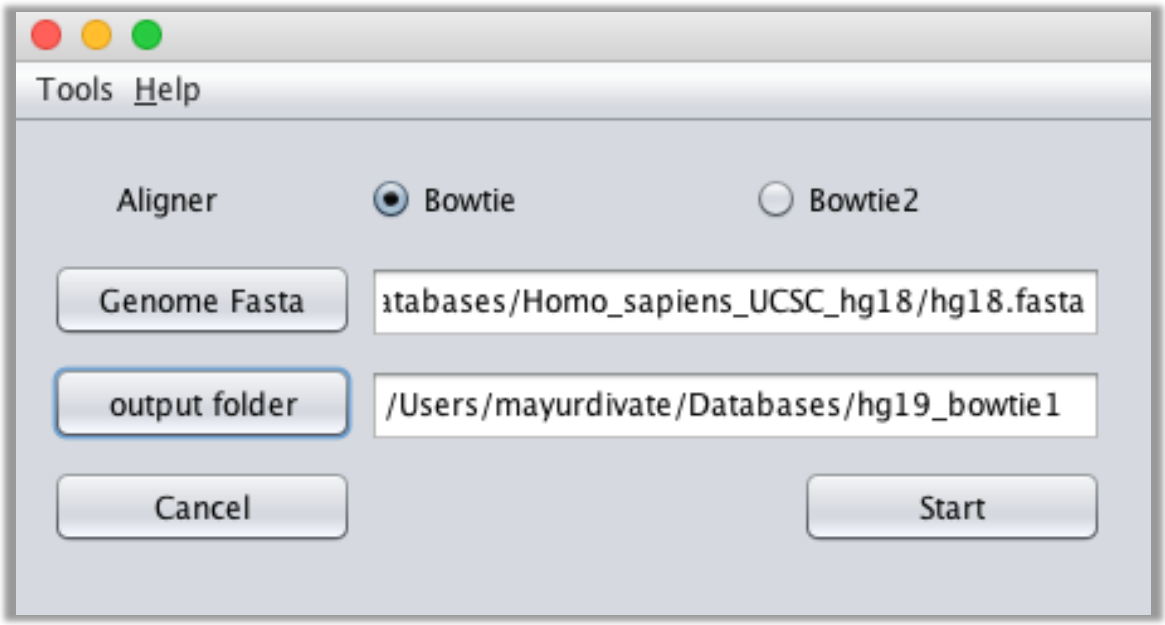


Figure 22: GUAVA Genome index builder program.