Protocol: peak-calling for ChIP-seq data / segmentation analysis for histone modification data

Table of Contents

Introduction

Objectives : given a dataset of aligned reads on a reference genome for ChIP-seq dataset, define peak regions, i.e. regions with a higher number of reads than expected by chance. For transcription factor ChIP-seq, peak regions are considered as putative binding locations. For histone modification datasets, peak regions reveal local enrichment in this particular histone modification (e.g. H3K4me1).

Relevant questions:

- comparing behavior of different algorithms;
- exploring the effect of some options;
- scaling behavior of algorithms (what happens if we double the number of reads ?);
- evaluation of peak quality using various criteria (inspecting raw data; motif enrichment; functional enrichment,...).

Dataset used: based on the publication “GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility.”, Theodorou et al., Genome Research (2012) (PMID: 23172872).
We will use ChIP-seq data for estrogen-receptor alpha (Eralpha) in two conditions:

1. wild-type condition after estrogen stimulation,
2. Eralpha binding after treatment with siRNA anti-GATA3.

Three replicates are available for each condition, and an input dataset for the considered cell-line (MCF-7) (check the dataset tableon the website for a description of the datasets and links to the files).

Starting point : files of aligned reads obtained from a read aligner (e.g. Bowtie, BWA, etc...) in BAM format; ideally, we should have separate files for the ChIP-seq (called here "*treatment*") and for the "*control"* or "*input"*(in general, sequencing of naked DNA or mock IP).

We have aligned the reads on the human genome hg19 using Bowtie2 with standard options, using a trimming of 10 bases on the 5' side (based on FastQC report).
Only uniquely aligned reads have been kept !
Note: *for reasons of computational time, for this practical we have split the datasets into different chromosomes; usually, one would analyze the full dataset at once. The full-genome datasets are provided for further analysis after the training school.*

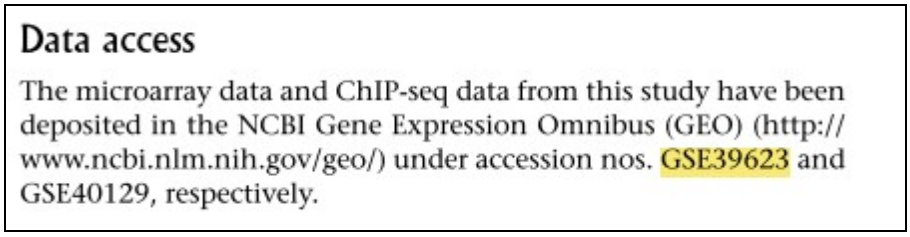
Protocol

Program

- Step 0 : Find datasets on Gene Expression Omnibus
- Step 1 : retrieving data from Galaxy
- Step 2 : data inspection
- Step 3 : peak calling using MACS
- Step 4 : splitting peaks with PeakSplitter
- Step 5 : comparing BED files
- Additional exercises
-
-

Step 0 : retrieve dataset from Gene Expression Omnibus

Goal: how do we obtain the datasets referenced in a publication ?



Data access

The microarray data and ChIP-seq data from this study have been deposited in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession nos. GSE39623 and GSE40129, respectively.

- Go to Gene Expression Omnibus. http://www.ncbi.nlm.nih.gov/geo/
- Give the reference of the dataset (GSE40129).
- *What sequencing platform was used ? How many samples are available ?*
- Click on the GSM link for one of the samples;
- then click on the SRX link to go to the small read archive (SRA), then click on the SRR link (bottom right) to access the record of the run.
- On the new page, click on the “Reads” tab to view the read sequence (you can display the quality clicking on *Customize*).
- From there, you might also download the dataset as a .sra file, but we will not do it in the context of this practical (beware, this would take time and occupy disk space, since SRA files typically weight several hundred Mb !).

This format can be converted to a FASTQ file using fastq-dump, which can then be aligned on a reference genome using a read aligner.

We will skip the read alignment for reasons of computational time, and analyse the already aligned BAM files.

Step 1 : retrieve data on Galaxy

- Log into your Galaxy account.
- Create a new History for this practical and name it "*ChIP-seq peak calling*".
- On the top of the page, click on *Shared data → Data libraries*, then click on *tp-mercredi-chipseq-nicolas.*

- Click on the blue arrow next to BAM to expand its content.

- The datasets have been split into distinct chromosomes (1 to 9) to speed up analysis for this practical; select your favorite chromosome, click on the blue arrow next to it, and tick all 9 datasets. **Note**: human chromosomes are numbered by decreasing size.
  For example:
  - Chr1 250Mb
  - Chr9 141Mb
  - Chr22 50Mb
  - ...

Data Library "tp-mercredi-chipseq-nicolas"

tp-mercredi-chipseq-nicolas

Name	Message	Data type	Date uploaded	File size
BAM	Automatically created by upload tool			
chr1	Automatically created by upload tool			
MCF-7_input_r3_SRR540220.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	47.7 MB
siGATA3_H3K4me1_SRR540211.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	75.9 MB
siGATA_ER_E2_r1_SRR540189.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	29.8 MB
siGATA_ER_E2_r2_SRR540191.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	24.2 MB
siGATA_ER_E2_r3_SRR540193.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	57.8 MB
siNT_ER_E2_r1_SRR540188.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	51.8 MB
siNT_ER_E2_r2_SRR540190.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	24.6 MB
siNT_ER_E2_r3_SRR540192.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	56.0 MB
siNT_H3K4me1_SRR540210.hg19.UNIQALIGN.chr1.bam		bam	2013-01-11	70.4 MB
chr2	Automatically created by upload tool			
chr3	Automatically created by upload tool			

- At the bottom of the page, select "import into current history" then click GO (You should have created a new history before this step !).Finally, click on *Analyze data* at the top of the page; you will go back to your Galaxy account in the current history, and the datasets should appear in the right column in green.
-
-

Step 2 : data inspection

Before starting the peak calling analysis, it is interesting to determine the **number of reads in each BAM file**, and the level of duplicates

- In the “search tools” box, search for the *flagstats* tool
- Run this tool on all the BAM files you imported, and note the number of reads.

One interesting aspect is the level of duplicates in the dataset;

- Run *FastQC* on the BAM file; check for biases and levels of duplicate reads

Step 3 : peakcalling using MACS

We will perform peak calling for the ChIP-seq datasets on ESR1 (hence siNT_ER_E2 and siGATA3_ER_E2 datasets), using the input dataset as a control.

- In the search tool on the left side, type MACS; click on the MACS tool in the left column.

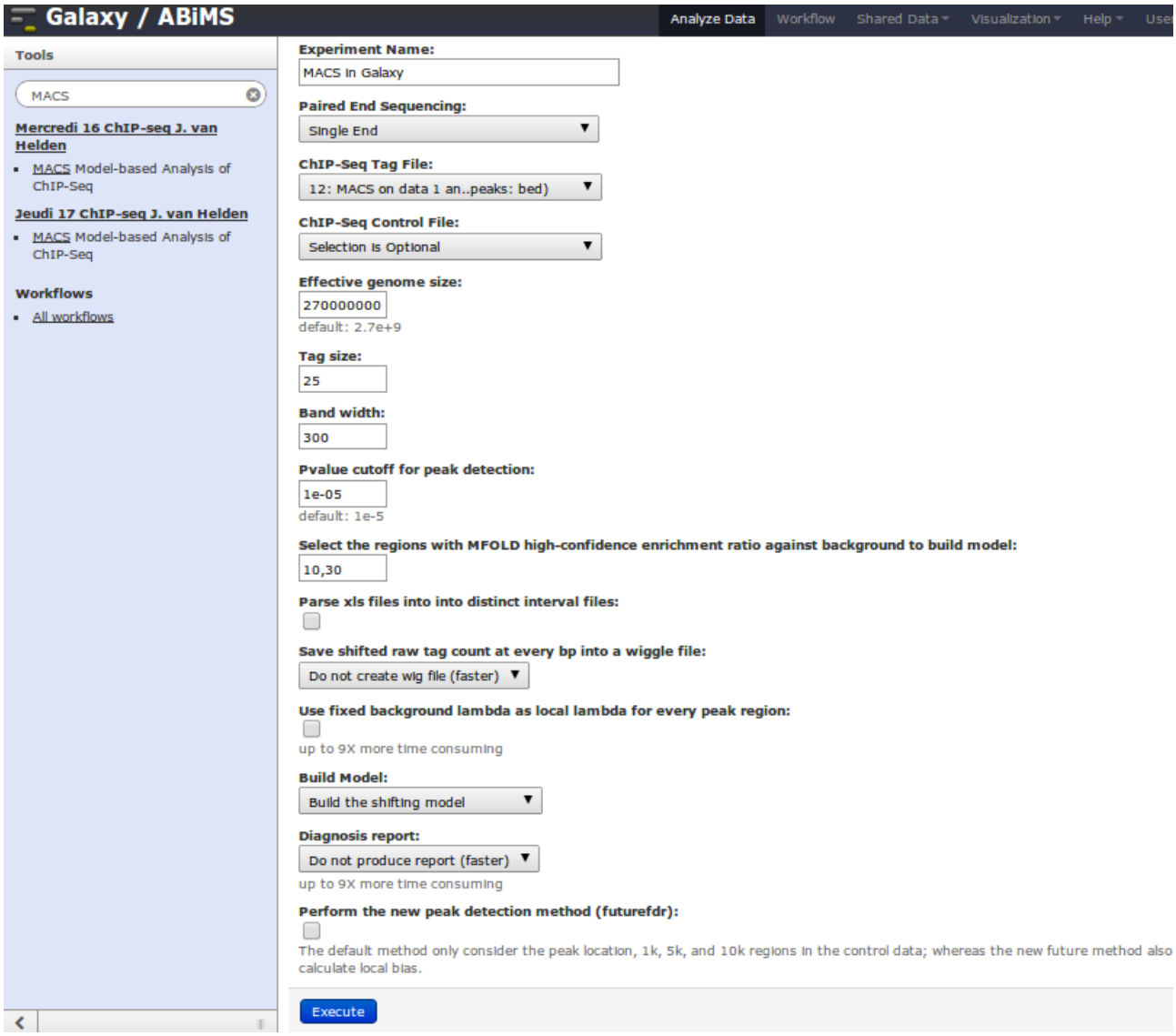


Illustration 1: Default view of the MACS implementation

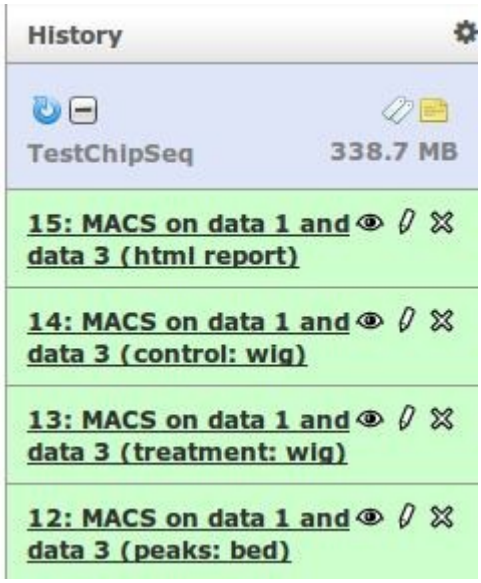
- Options to be used
  - *Experiment name* : give a name for the MACS run.
  - *Paired end sequencing*: MACS can handle single or paired-end data; here we select single end.
  - *ChIP-seq tag file* : select a BAM file containing the treatment.
  - *ChIP-seq control file* : select the BAM file for the input (MCF-7).
  - *Effective genome size*: since you are working on individual chromosomes, you should

indicate the size of the hg19 chromosome (see table below).
    - *Tag size* : these are Illumina datasets of read size 36; however, trimming of 10 positions was performed prior to aligning; hence, this should be 26.
    - *Save shifted raw tag count at every bp into a wiggle file* : select save; this will output WIG files for the treatment and the input.
    - All other options should be default.

Each MACS run using these options should generate 4 result files:

1. html report describing the model built by MACS;
2. 2 wig files for the treatment and control dataset;
3. a bed file containing the peaks.

Click on the eye symbol in the dataset containing the html report; an additional series of files shows up :



- click on the pdf file to see the model determined by MACS
- have a look at the XXX_summits.bed file

Repeat the MACS analysis for the following datasets :

1. siNT_ER_E2_r1
2. siNT_ER_E2_r3
3. siGATA3_ER_E2_r1
4. siGATA3_ER_E2_r3

Things you should consider:

- How many peaks have been called by MACS for each dataset ? Relate this to the number of reads in the BAM files (end of Step 1), in particular for the replicates; relate this to the size of the chromosome (we will compare numbers between participants).

Step 4 : Splitting peaks

Given the algorithm used, peaks called by MACS can be rough on the edge … Peaks might span long regions which, by eye, one would have subdivided into smaller peaks (see figure). We will now use a tool (PeakSplitter) which subdivides large peaks into smaller, more precise one.

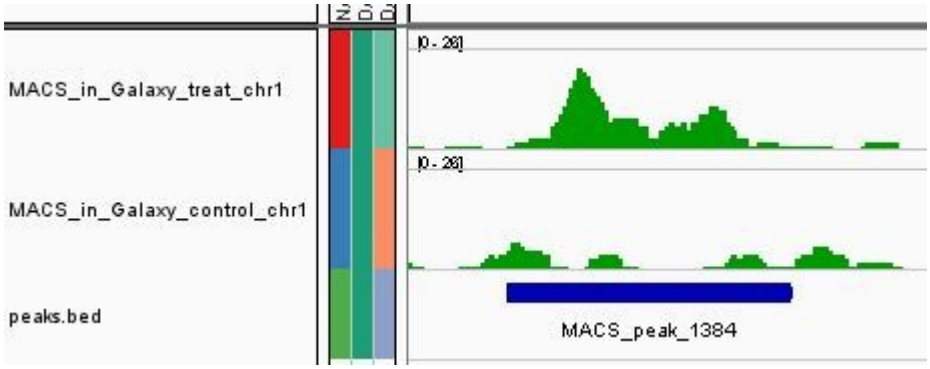


Illustration 2: Example of broad peak called by MACS viewed in IGV

Chromosome	Size
Full genome	2.7e9
chr1	249250621
chr2	243199373
chr3	198022430
chr4	191154276
chr5	180915260
chr6	171115067
chr7	159138663
chr8	146364022
chr9	141213431

Table 1: Chromosome sizes for hg19

First, we are going to compute some statistics on the peaks called by MACS, for example the distribution of peak sizes.

- In Galaxy, go to the *Text manipulation* section in the left column, and select *Compute an expression on every row.*
- As expression, input c3-c2 (i.e. end coordinate minus start coordinate = peak size).
- Select one of the bed files obtained from MACS for one of the datasets.

- Now select the tool *Summary statistics* on numerical column, select the dataset just obtained (which contains an additional column), and compute the statsitics on this additional column.
- Display the distribution of sizes using the *histogram* tool.

Troubleshooting:

if some tools are not available on the Galaxy instance of SBR, you can use the public Galaxy at PSU http://galaxyproject.org/

- Download the BED files you want to analyze (click on the small blue floppy icon) on your computer, then connect to the public Galaxy instance.
- Choose *Get data* in the left column, then *upload file from your computer*,and upload the BED file. Then run the tools.

We are now going to run PeakSplitter in order to obtain a refined set of peaks:

- search for the tool PeakSplitter;
- select the bed file containing the peaks for one of the datasets;
- select the wig file for this dataset (treatment file, not the control file !);
- run the tool.

Repeat this operation for the various datasets.

Things to consider :

- What is the number of peaks obtained with PeakSplitter, as compared to the initial number ?
- Compute the statistics on the new peak file (as previously) and compare the mean, median peak sizes.

Step 5 : Comparing bed files

Goal : compute the intersection of 2 bed files to compare 2 sets of peaks; this can be applied to the comparison of 2 datasets, of various parameter settings, of different peak callers, etc...

As an illustration, we will compare the set of peaks obtained (I) between 2 different replicates, and (ii) between the wild-type and siGATA3 conditions.

- Select the tool *Intersect intervals of two datasets*, and apply it to the output of PeakSplitter for 2 replicates in the siNT and siGATA3 cases.
- Select *return overlapping intervals* (see the figure at the bottom of the tool to understand what this means).

Create 2 additional output files containing the intersection of the 2 replicates for siGATA3 and siNT. Check how many peaks these datasets contain. You can represent this as a Venn Diagramm (on a sheet of paper ...)

Redo the same analysis, this time to compare the consensus datasets for both conditions (siNT and siGATA3; see illustration 3)

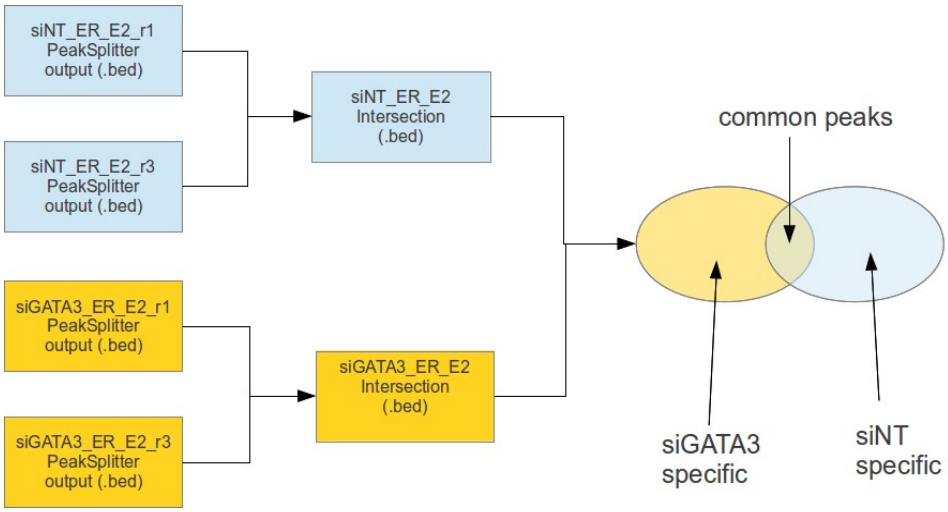


Illustration 3: Workflow diagram for comparing datasets

In the same way, use the tool *Subtract the intervals of two datasets* to determine the set of consensus peaks that are specific to one and the other condition.

In conclusion, you should have obtained 3 sets of peaks:

1. common peaks to siNT and siGATA3;
2. peaks specific to siNT;
3. peaks specific to siGATA3.

Step 5: Visualizing data in the genome browser

Goal: visualizing the different datasets in a genome browser (IGV): BAM files, wig files, bed files for MACS and MACS+PeakSplitter

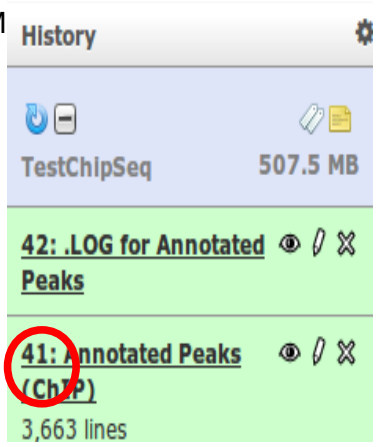


Illustration 4: Downloading a dataset

- choose one dataset
- for this dataset, download the BAM file, the wig file for treatment and control, and the bed files obtained from MACS and PeakSplitter; this is done by clicking on the little floppy-disk symbol close to the dataset in the right column (for the BAM file, download the read file and the index file !)
- launch the IGV genome browser; on the top bar, select the chromosome of interest (see illustration 5); in the File menu, select *Load from file*, and upload the different datasets you obtained from Galaxy (ignore the warnings when uploading the WIG files)

Things to consider:

- Navigate through the genome, and try to identify cases where PeakSplitter has subdivided a single peak into several smaller ones.
- Look at the profile of reads (separate the reads according to strand using a right click).

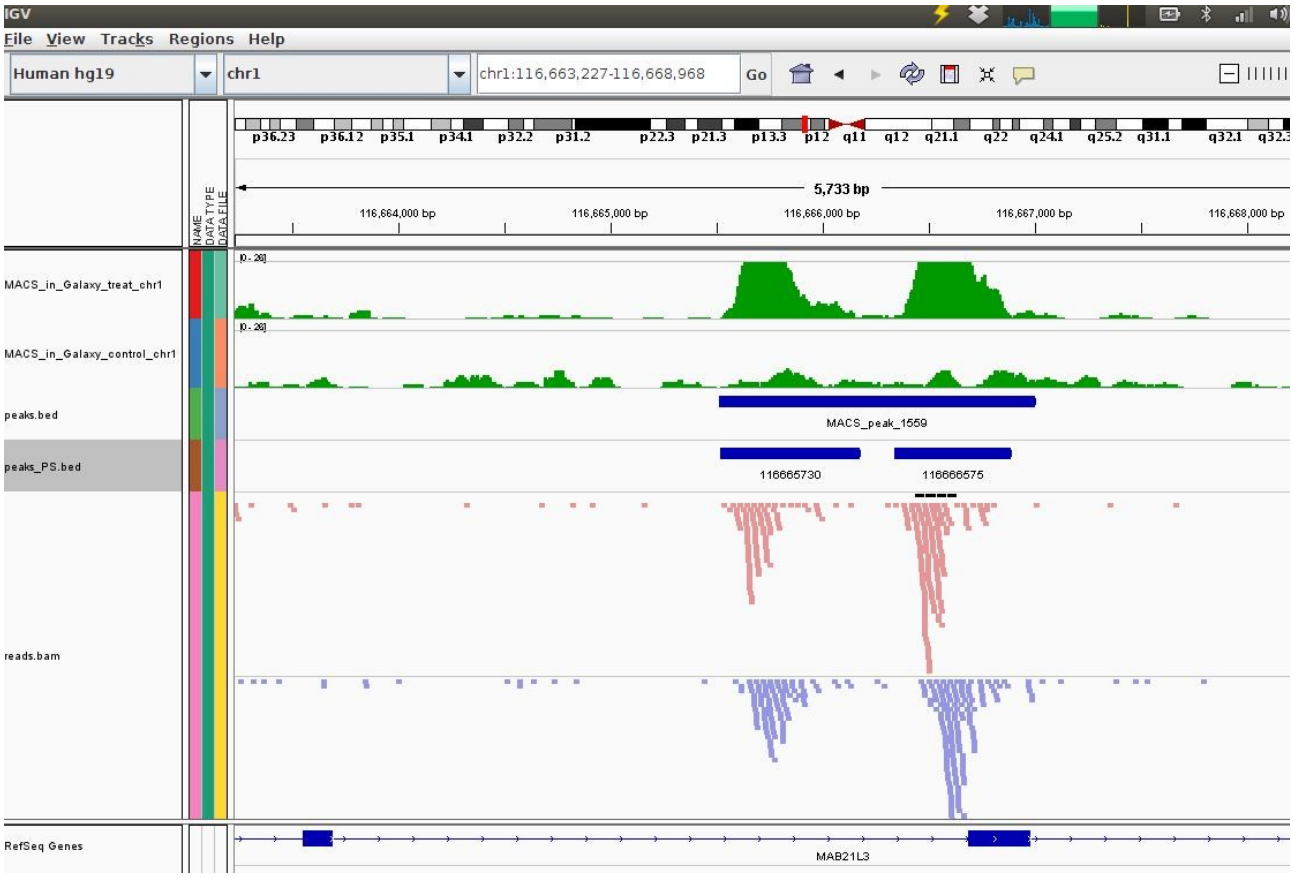


Illustration 5: IGV browser window with BAM, WIG and BED tracks

Exercices

Exercice 1 : how to deal with replicates ?

Several strategies can be applied when replicates are available : calling peaks independently on each replicate, and taking the intersection of the peaks, or merging the replicates into one single BAM file, and calling peaks on the merged dataset.

- Choose one dataset for ESR1
- Do the intersection of peaks called using the different replicates independently (take 2 replicates)
- Merge the BAM files for the replicates using the tool Merge BAM files, then call peaks on the merged file
- Compare the results.

Exercice 2: using MACS without input

Run MACS on one of the datasets without input dataset (use only treatment).
Compare the set of peaks obtained to the set obtained previously using input data (Number of peaks, sizes, etc...)

Exercice 3 : using a different peak caller

Goal : compare the peaks identified using MACS with a different peak caller; here, we will use CCAT as an example (Xu et al., 2010). CCAT will also be used to identify enriched regions for histone modification marks.

- CCAT does not accept BAM input files, but only BED files; search for the tool *Convert from BAM to BED*, and convert the treatment and input file for one of the datasets into BED format.
- when the conversion is done, click on the little pen icon the the newly created datasets to edit the attributes of these datasets; in database/build, select “hg19” as the reference genome.
- select the CCAT tool, and run the tool
- compare

Exercice 4 : H3K4me1 enriched regions

Use CCAT to determine enriched regions for H3K4me1 in both conditions (siNT and siGATA3); how many ESR1 peaks in both conditions are located into regions enriched in H3K4me1 ?

Exercice 5 : full genome analysis

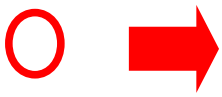
Redo the analysis, but this time on the full dataset

- MACS + PeakSplitter for ESR1 in siNT and siGATA3 condition
- ESR1 peaks specific of siNT; of siGATA3; common peaks
-

-
-
  -
  -
  -
  -
  -
  -
  -
  -

1.
2.
3.

Illustration 6: Default view of the MACS implementation



-
-

1.
2.
3.
4.

-

Illustration 7: Example of broad peak called by MACS viewed in IGV

Table 2

-
-
-
-
-

Troubleshooting:

if some tools are not available on the Galaxy instance of SBR, you can use the public Galaxy at PSU http://galaxyproject.org/

-
-
-
-

-
-

-
-

Illustration 8: Workflow diagram for comparing datasets

1.
2.
3.

Illustration 9: Downloading a dataset

-
-

-
-
-

Illustration 10: IGV browser window with BAM, WIG and BED tracks

-
-
-
-

-
-
- regions enriched in H3K4me1 in both conditions