Mrinal Vashisth, M4135, ИТМО, биоинформатика и системная биология, ITMO, 2019

Unicorns are real. They are called Rhinoceros.

EPIGENOMICS

Chipseq data analysis and peak calling.

Taken from Systems Biology 2019, module on Epigenomics by Oleg Shpynov and Roman Chernyatchik.

Personal summary of slides

What is Epigenomics?

Evolution does not work on blank slate. It works with the preexisting information and builds upon it. In context of epigenomics, we have **regulatory DNA**.

Genomics is the sum total of genes and genic product (Transcriptomics - different types of RNA). We do NGS, RNAseq, Variant, Copy Number Variation etc. analyses for these.

In Epigenomics we are concerned with the regulation of the genome. Here is where environment/ lifestyle can actually affect the individual.

Epigenomics also plays a key role in development e.g. different tissues of the body share although the same DNA but different epigenetic patterns that results in specificity of their function. Thus differentiation, de-differentiation, re-differentiation etc. can be seen as Epigentic changes. It is also noteworthy that these mechanisms vary greatly between plants and animals, an early evolutionary split.

Examples: 1.) the Dutch mass starvation during second world war resulted in epigenetic changes which can still be seen 6 generations after the incident. Their metabolism shifted to a conservative mode i.e. preserving energy from diet. Other exampl

2.) X barr inactivation in females. There are two copies of X chromosomes, one of these is inactivated during embryonic development and remains highly condensed.

3.) Mixed coat of Cats, red fur in these furry animals, mostly females. Sometimes male cats also show this condition, they have XXY condition and are sterile. Note: When it is XXXY male cat they are not sterile. An analogy could be bacterial Lac operon. Where if Lactose is present it will switch on the gene responsible for lactose breakdown.

https://en.wikipedia.org/wiki/DNA_methylation (https://en.wikipedia.org/wiki/DNA_methylation)

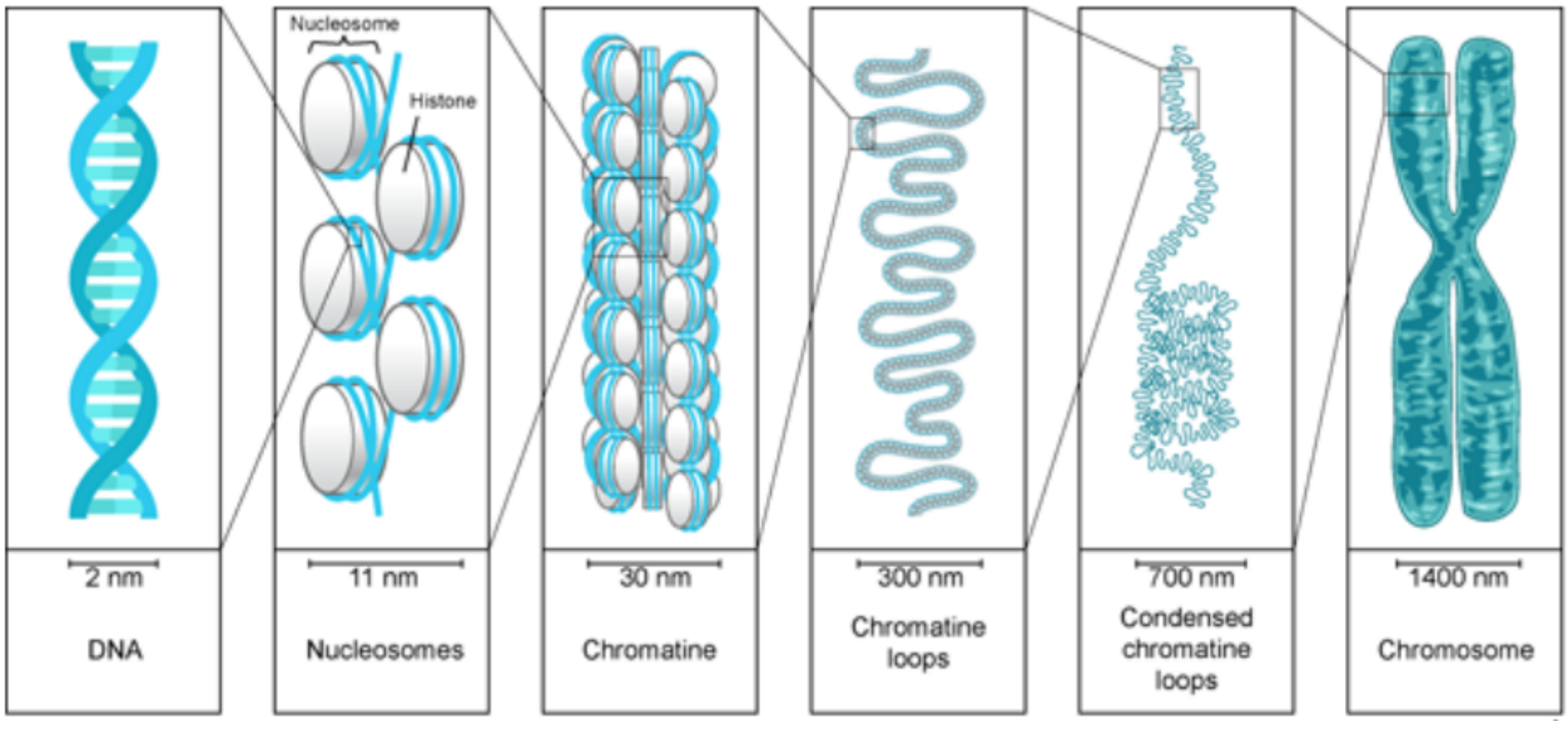
In January Sys Biol module there was a lecture where segnor Oleg described their research about whether Epigenetic markers change in Humans. The answer was 'no', they were surprised by the result. Another lab in some other part of the world also came upon the same conclusion.

This is surprising because in Mice these markers change with age. Suggesting that in Humans this process of Epiregulation is complicated.

Mechanisms of Epigenetic regulation

Packaging of DNA

If we look at the packaging:



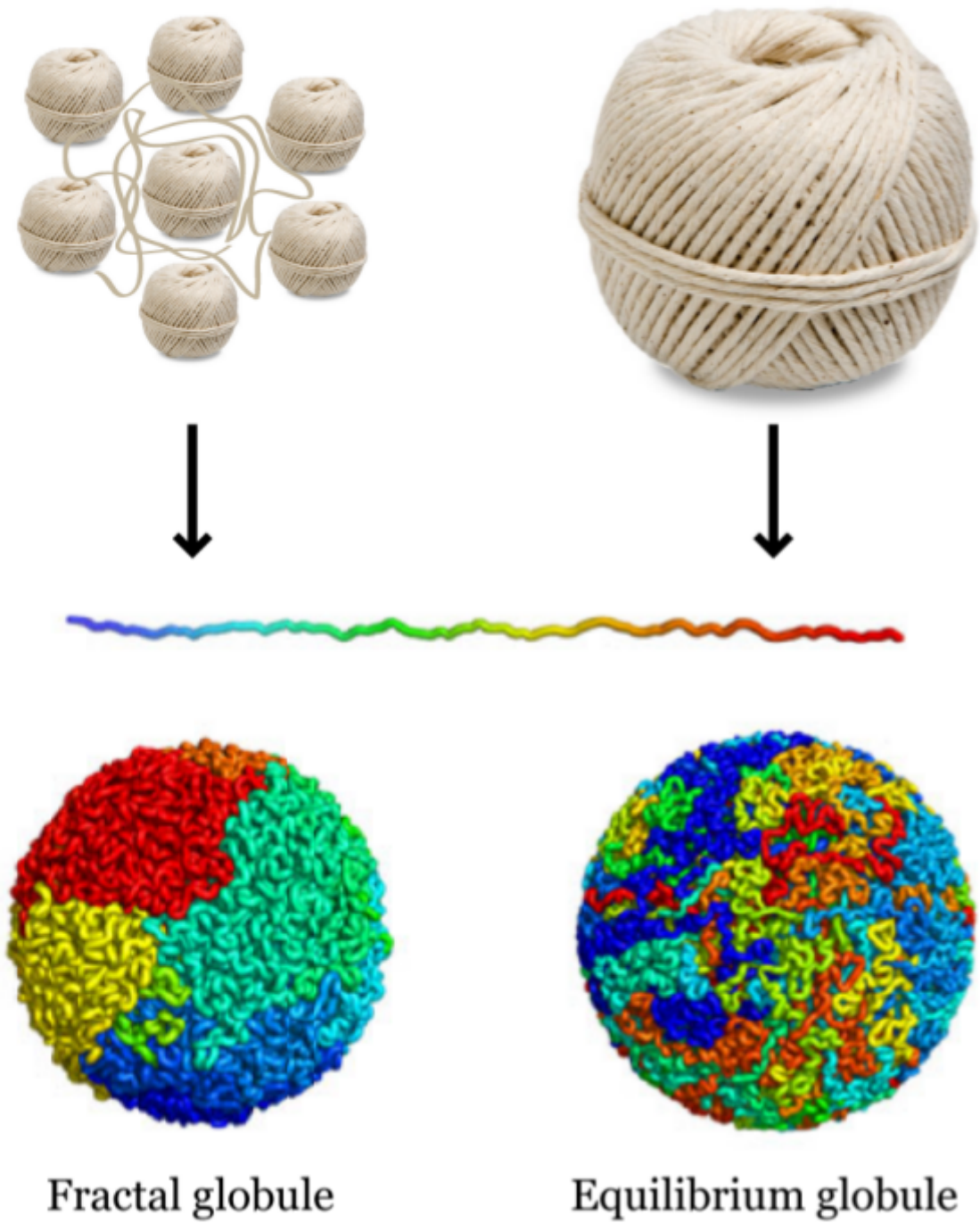
DNA -wrapped around---> **histone octamer** -multiple histones form---> **nucleosomes** --multiple nucleosomes---> **solenoids** (cylindrical supra-assemblies) --forms----> **chromatin** --form---> **chromatin loops** (if they are too tight: hetro; else: euchromatine) ----further condense to form-----> **chromosome**

Epigentic regulation can occur at any of the stages given above.

Examples:

1.) For DNA, at individual nucleotide base level, we can have methylation (methyltransferases enzyme), hydroxy-methylation, sulphynation etc. This results in physical blockage of DNA structure, such that replication motor cannot move over the strand and transcription is not possible. This can be thought of as a similar process to protein activation/ inactivation by mechanism such as phosphorylation, only that DNA methylation is the boss of all.

2.) Histone level includes modifications of histone molecules which activates them to wrap tightly or loosely around the DNA. The resulting subsequent structures will also be tightly bound.



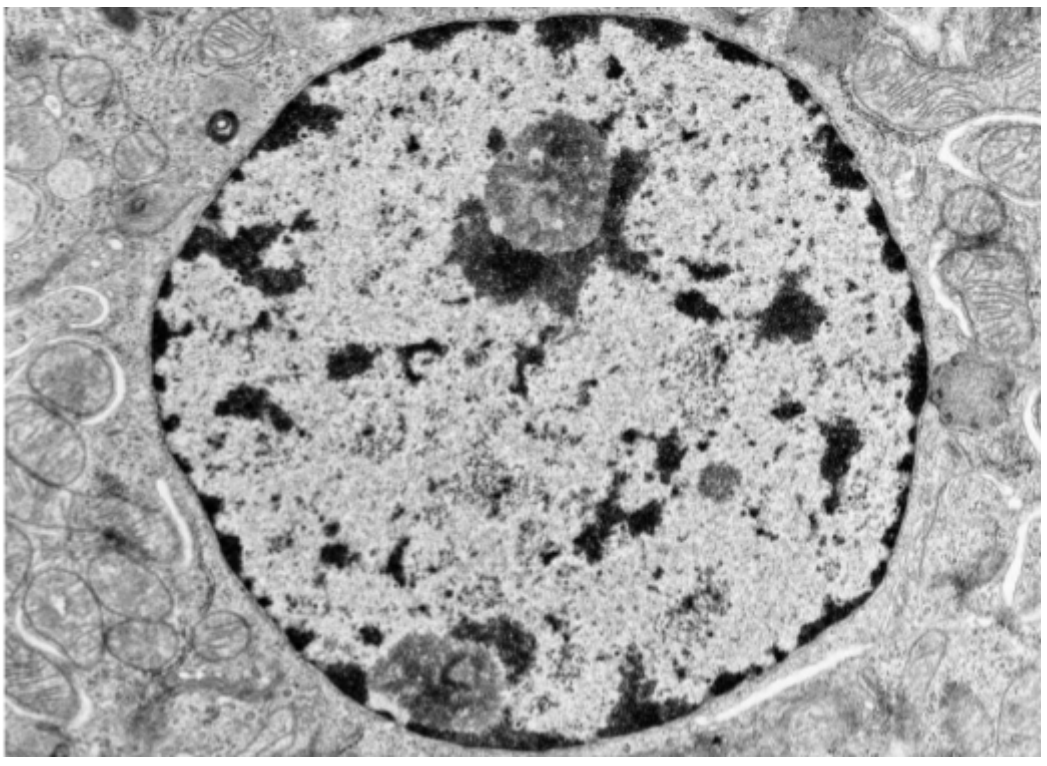
Mirny, L.A., Chromosome Res 19, pp 37–51 (2011)

Nucleosome is a equilibrium globule i.e. it allows for easier access of DNA relatively independent of regions. Many such Nucleosomes come together to form a fractal globule.

There are two types of chromatine: Heterochromatin and Euchromatine. 'Eu' means 'true' i.e. his region of chromatine is open for transcription as opposed to 'hetro' which is tightly bound. So the inactivated X (barr body) can be seen as heterochromosome as the entire chromatin on this chromosome is heterochromatine.

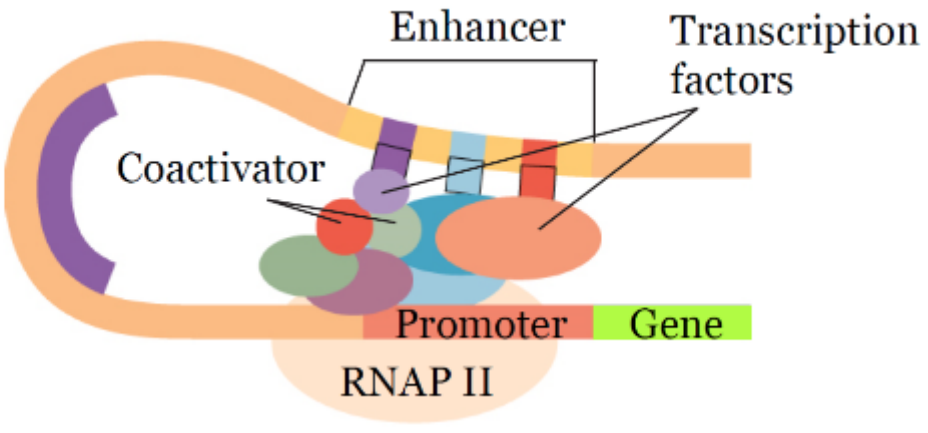
Thus the process of chromatin activation (conversion of hetero to Eu chromatin). Note: this is not the same as saying decondensation of chromosome, decondensation is the process of unpacking of chromatids into the original state after cell division [https://www.sciencedirect.com/science/article/pii/S0955067416300059 (https://www.sciencedirect.com/science/article/pii/S0955067416300059)]. Activation means that they are actually ready for transcription. REMEMBER!! **Constitutive heterochromatin is irreversable** and **Facultative heterochromatin is reverible**. This means that if heterochromatine is constituent of the organism it won't change to euchromaine, thus any gene which has been constitutively heterochromatisised is irreversibly silenced forever. Check out the image below.

- Gene silencing can be reversible and (relatively) irreversible
- irreversible = heterochromatin
- reversible can be removed in the process of cell differentiation



At loop level, distant regions of DNA come together for transcription of certain genes. If this process is interrupted, say for example, by binding of a protein, we will see no transcription. Thus looping is also crucial not just for condensation but transcription of certain genes. Observe the picture below to understand.

Basal transcription: **general transcription factors** bind the promoter and RNA polymerase II
Activator proteins bind DNA spots named **enhancers**
Enhancers are often located far and have to **loop**

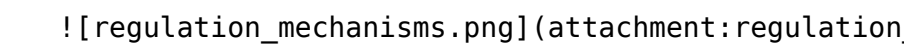


To summarize: There are 2 major definations of epigenetics:

○ Heritable changes that do not involve DNA modification (still heavily debated) -- changes in metabolic patterns, onset of diseases
○ DNA modification and packaging changes that influence expression (well established) -- regulation of genic function

Also depending on the process:

○ differentiating cells -- switching on/off of genes to achieve functional roles
○ pluripotent cells -- completely open genome, for the sake of differentiating in any other kind of cell in the body
○ differentiated cells -- after achieving a particular role, changes in the metabolic patterns. An analogy can be
bacterial Lac, His operons etc.

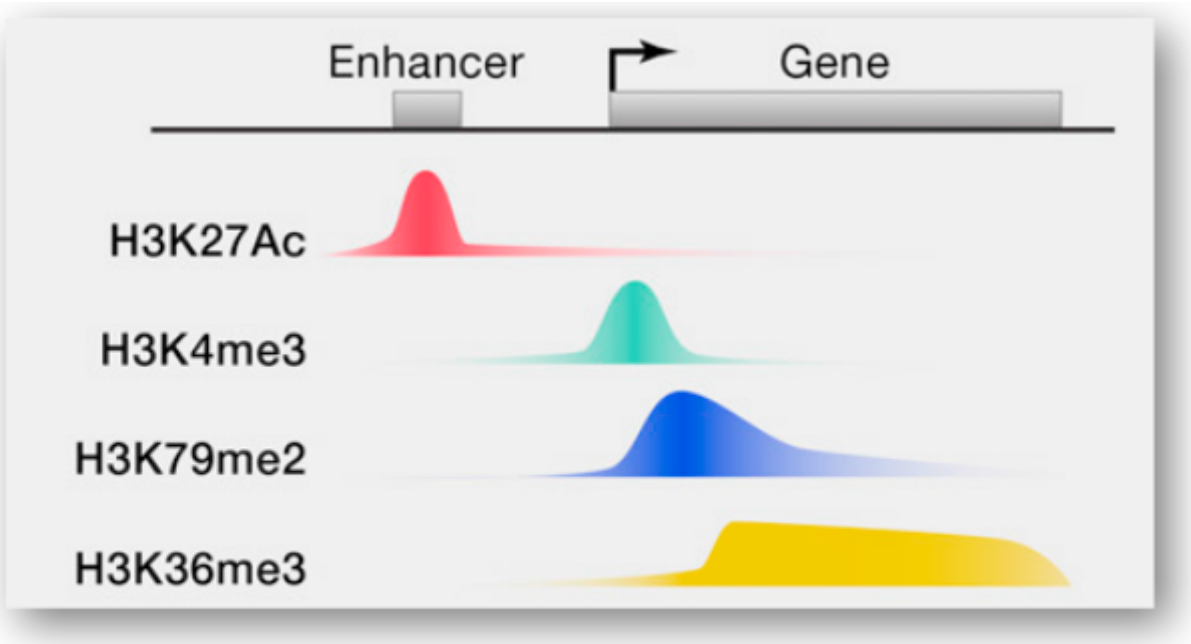


Epigenetics Regulation via:

Histone Modifications
Chromatin accessibility and 3D structure
Non-Coding RNAs
DNA cytosine methylation

Histone modifications

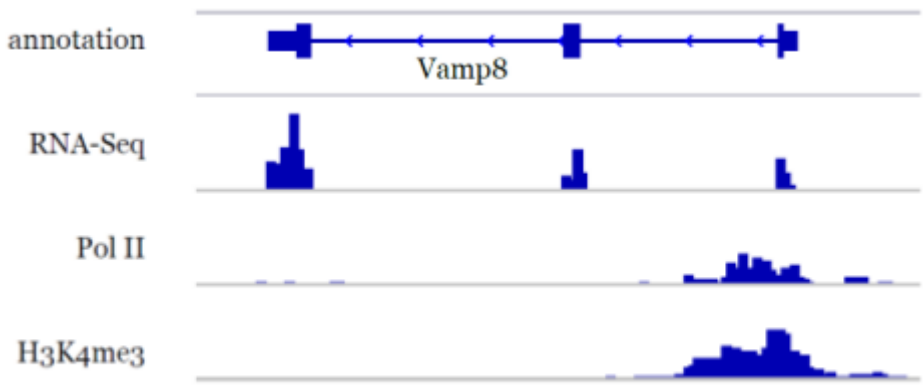
Look at this image, these are the modifications we are also going to analyse for our data:



Lee, T.I.; Young, R.A., Cell, 152(6), pp 1237-51 (2013)

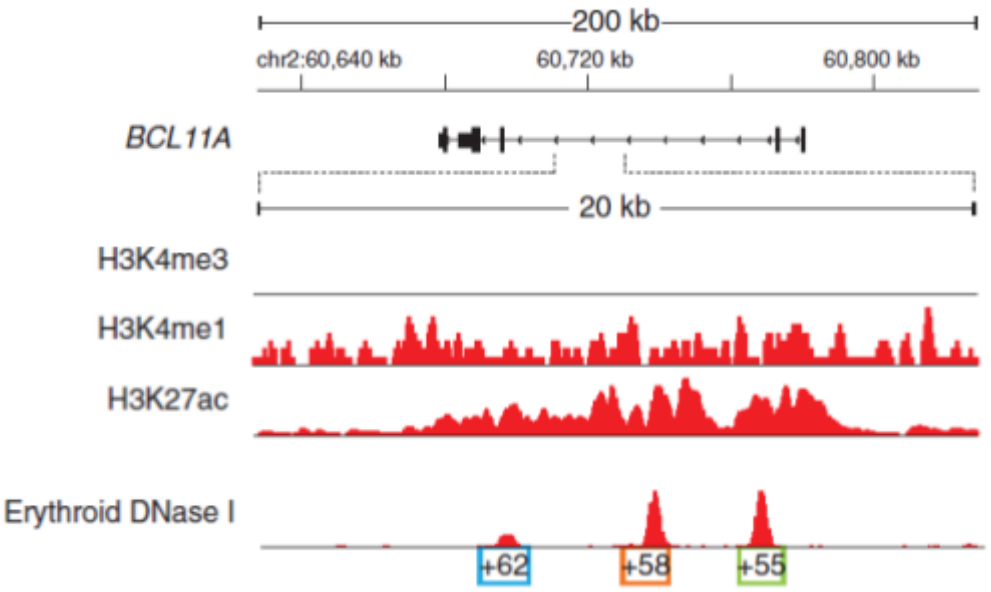
Cut to the chase summary:

1.)



Narrow peaks of H3K4me3 mark promoters. Enzyme that methylates K4 binds only to non-CpG-methylated promoters (mouse heart)!

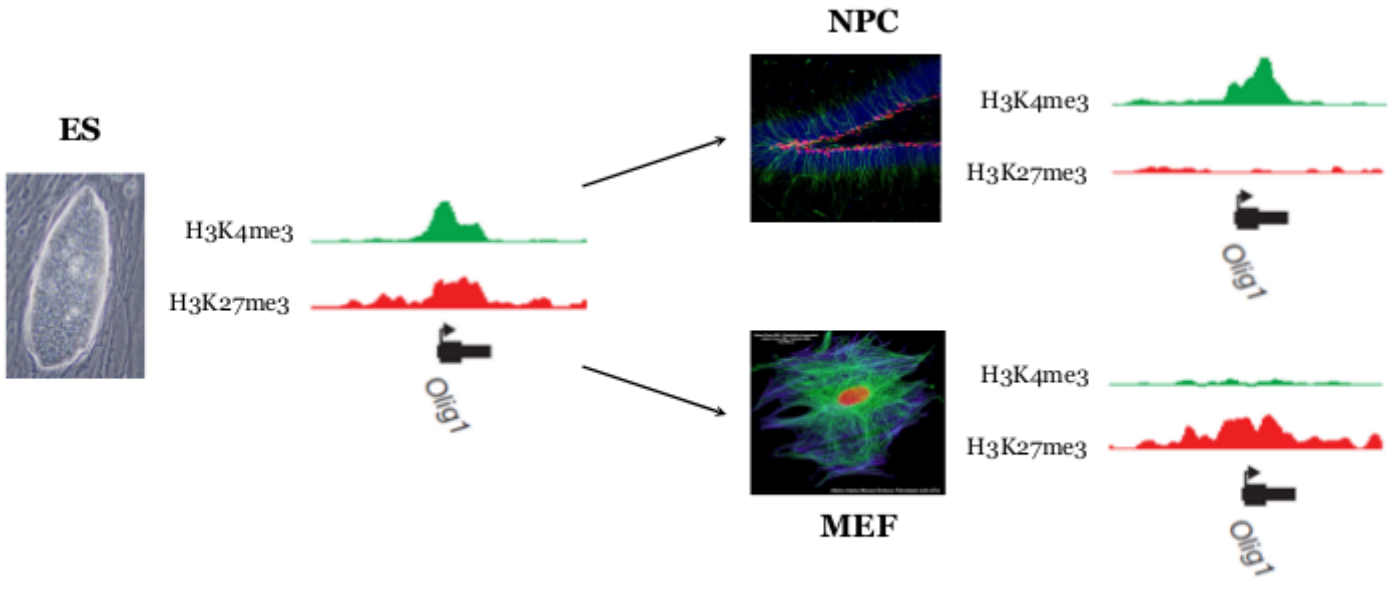
2.)



Enhancers are associated with H3K4me1 and H3K27ac. H3K27ac is thought to distinguish active enhancers from poised. The word poised means the gracefully static, just like the ballerina when she is en-point. Thus a poised gene is a gene that was suppressed but is now ready to be activated.

H3K27me3 marks suppressed genes poised to be activated.

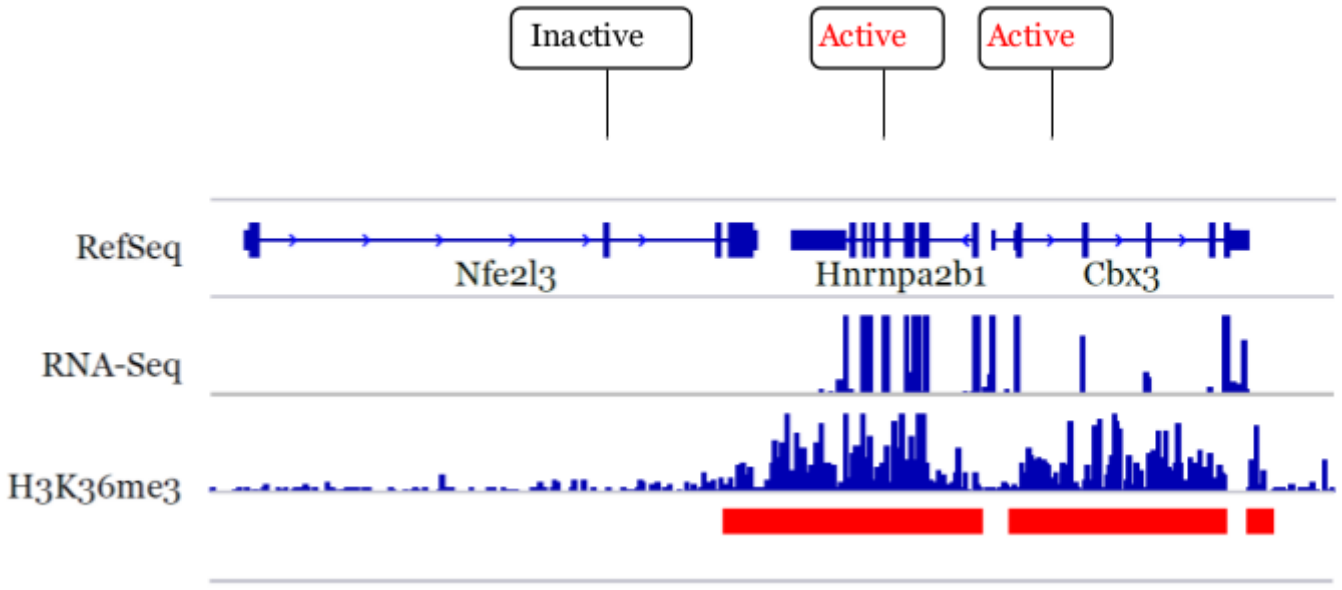
Stem cells can have both H3K4me3 and H3K27me3 – unique!



Mikkelsen, T.S.; Ku, M. et al, Nature 448, pp 553-560 (2007)
Vastenhouw N.L.; Zhang Y. et al, Nature 464, pp 922-6 (2010)

In this image, same promoter region has both tri-methylation marks. Such region is a bivalent chromatin domain and are shown to exist also in embryonic stem cells. These poise gene for activation while also keeping them repressed!

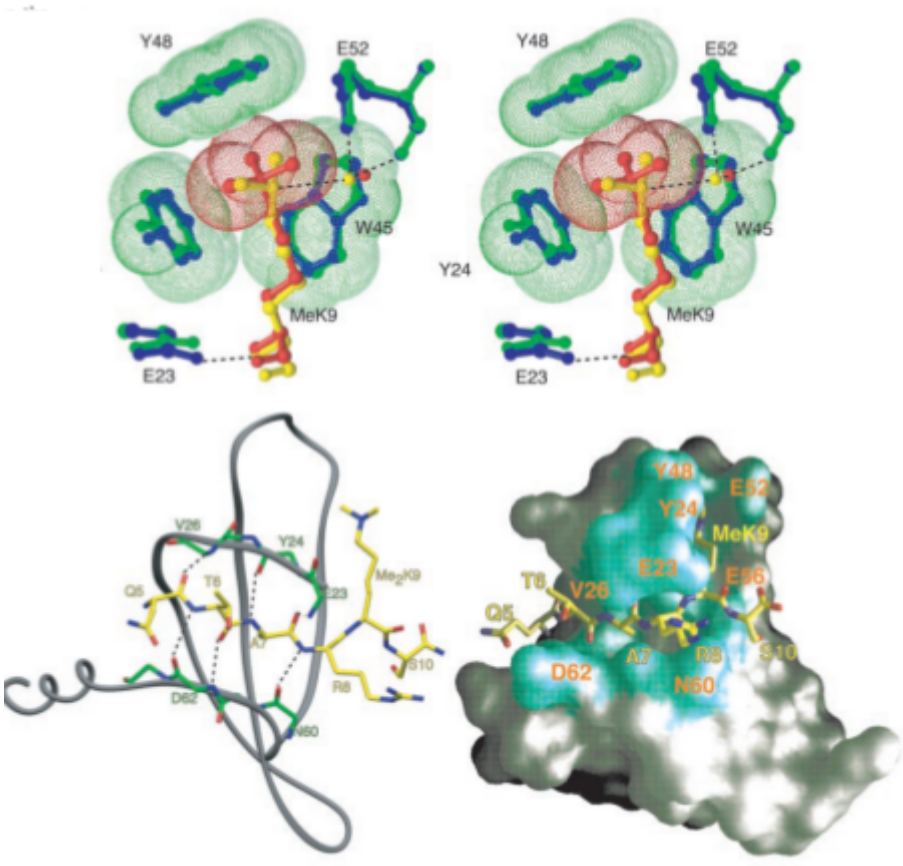
3.)



Transcriptional elongation is marked my H3K36me3 and H3K79me2 (Prob cells, Mouse).

4.)

Marks H3K9me2 and H3K9me3 are strongly associated with heterochromatin.

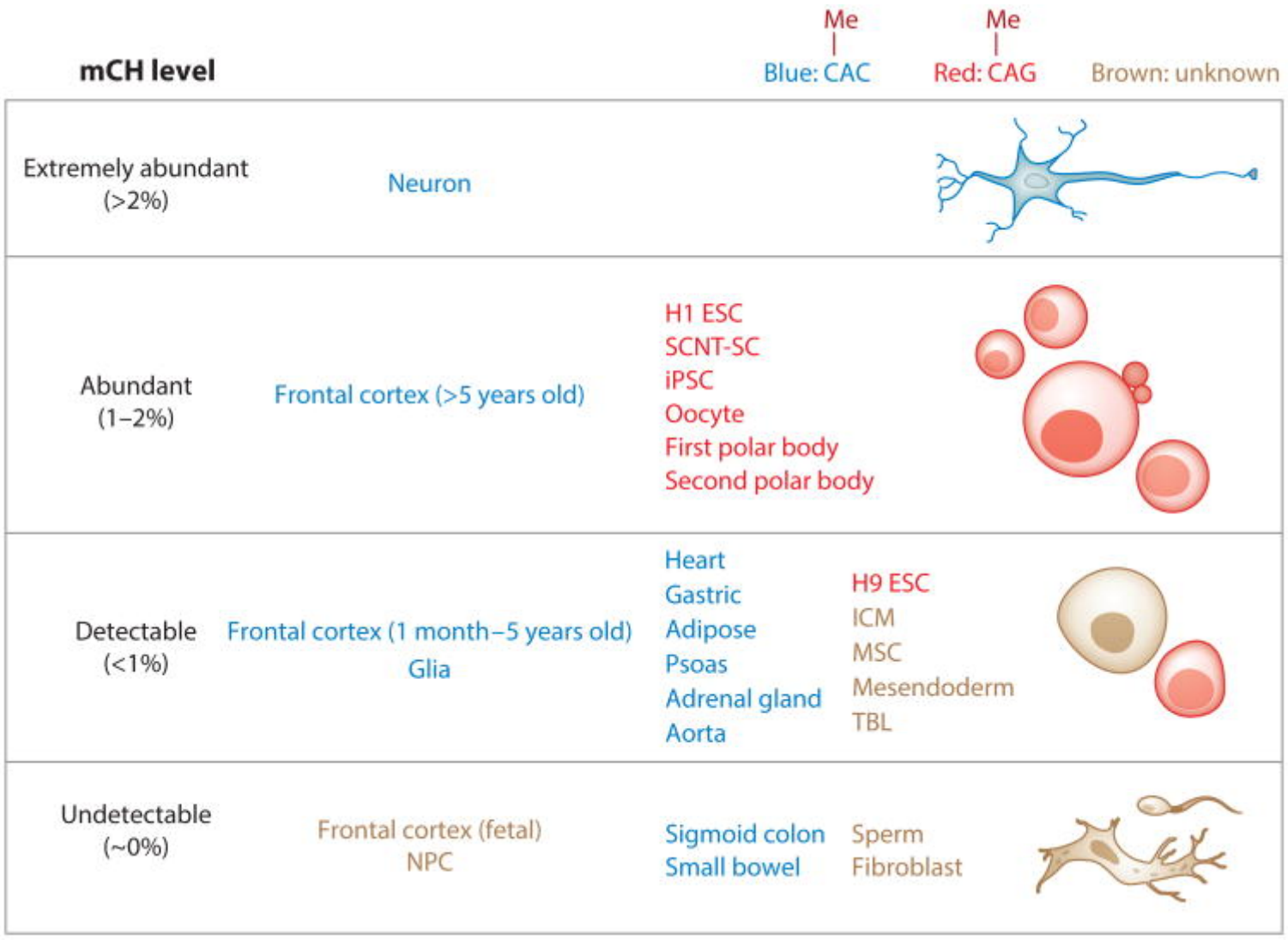


DNA Methylation landscape

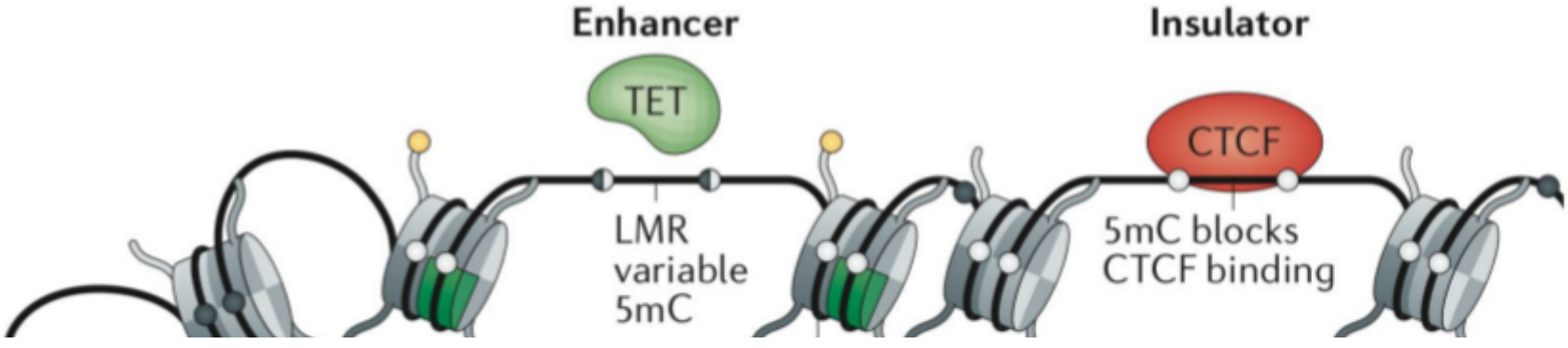
When we think of methylation, we mostly think about CpG islands. The CpG (C preceding G) islands is a region with high frequency of CG's. 60% human genes has CGI in promoter. Methylation/ demethlyation of these islands can turn genic activity on/off. At transcription start site (TSS) we find minimum number of CG islands, even in unexpressed genes.

But of-course CpG islands or mCG (modification at CG) are not the only way methylation works. We have methylation at other bases called *modification at CH* where H is either A, T, or C.

The following image shows that mCH are most common in neurons. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4729449/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4729449/)]

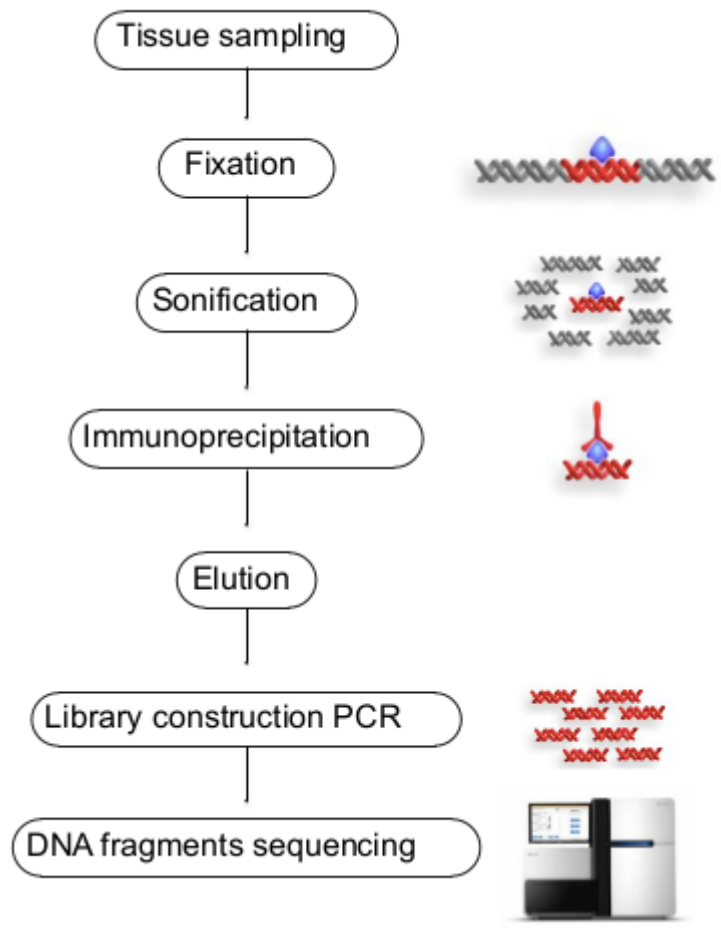


Observe the following image:


Enhancer
Insulator
TET
CTCF
LMR
variable
5mC
5mC blocks
CTCF binding

PIPELINE

ChIP-Seq protocol from Chip-seq to data analyses



Wet lab

1.) Destroy the cells 2.) Add formaldehyde to chemically bind proteins to DNA 3.) Break the DNA into smaller fragments (using ultrasound -- sonification) 4.) Precipitate by binding to antibodies (immunoprecipitation) 5.) Elute (extract) the fragments 6.) Construct the library for PCR 7.) Sequencing of the fragments

Dry lab

7.5) Quality checking 8.) Alignment of reads to a reference genome using bowtie2 9.) Sorting and indexing of bams using samtools 9.5) Visualisation using JBR (JetBrains Research (correct me if I am wrong) -Genome Browser) 10.) Peak calling:

10.1.) Macs2 10.2.) Sicer 10.3.) SPAN

11.) Getting most confident peaks 12.) Annotaing the peaks in JBR 13.) Model training 14.) Model tuning

15.) Interpretation

Data Preprocessing Pipeline

The data is taken from following paper:

LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation [https://www.nature.com/articles/s41467-018-05841-x (https://www.nature.com/articles/s41467-018-05841-x)]

If you search for keyword 'GSE' on this paper you'll get **Gene Expression Omnibus accession: GSE112622**

From there you can get an SRP (Sequence Read Project) ID.

We are intrested in Chip-Seq samples, particularly: ● GSM3074494 ● GSM3074495

We can download GSM3074495 FASTQ Files using *fastq dump*.

But we already have the data, since the data is large only 15th chromosome is taken.

All the respective tracks are also downloaded in the data folder: ENCODE project https://www.encodeproject.org (https://www.encodeproject.org)

Please look at the presentation for complete pipeline. Let's discuss the script line by line with results.

Open terminal and get tree. Let's look at the directory structure.

We need to modify the scripts according to this directory structure. We will only modify the **main.sh** file. **Please extract this zip in you any directory.**

```
sudo apt-get install tree
```

```
tree
```

In order for the script to work, the structure should look like this.

```
$ tree
.
|-- Notes_Mrinal.ipynb
|-- data
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15.fastq
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15.fastq
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15.fastq
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15.fastq
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15.fastq
|   |-- GSM1102807_CD14_input_hg19.chr15.fastq
|   `-- index
|       |-- hg19.1.ebwt
|       |-- hg19.2.ebwt
|       |-- hg19.3.ebwt
|       |-- hg19.4.ebwt
|       |-- hg19.fa
|       |-- hg19.rev.1.ebwt
|       `-- hg19.rev.2.ebwt
|-- main
|   |-- main.sh
|   |-- scripts
|   |   |-- bigwig.sh
|   |   |-- bowtie.sh
|   |   |-- closest_gene.sh
|   |   |-- fastqc.sh
|   |   |-- get_genome.sh
|   |   |-- index_bowtie.sh
|   |   |-- macs2.sh
|   |   |-- pipeline.sh
|   |   |-- remove_duplicates.sh
|   |   |-- sicer.sh
|   |   |-- span.sh
|   |   |-- util.py
|   |   `-- util.sh
|   `-- tools
|       |-- picard-2.10.7.jar
|       `-- span-0.10.0.4787.jar
```

```
##########################################################################################

# DEPENDENCIES

# fastqc for individual file quality check

sudo apt-get install fastqc

# multiqc for combined quality check

pip install multiqc

# bowtie for alignment

sudo apt-get install bowtie

# deeptools for analysis of deep-seq data

conda install -c bioconda deeptools

# latest samtools

conda install -c bioconda samtools

# pybigwig: it is an extension for deeptools for quick accessing bigWig and bigBed files

conda install -c bioconda/label/cf201901 pybigwig

# deeptools_intervels: it is another python extension for deeptools for constructing interval trees with

associated exon/annotation information
conda install -c bioconda deeptoolsintervals

# pysam: Pysam is a python module for reading and manipulating Samfiles

conda install -c bioconda pysam

# Macs2 peak caller

conda install -c bioconda macs2

# Scicer peak caller

conda install -c bioconda scicer

##########################################################################################
```

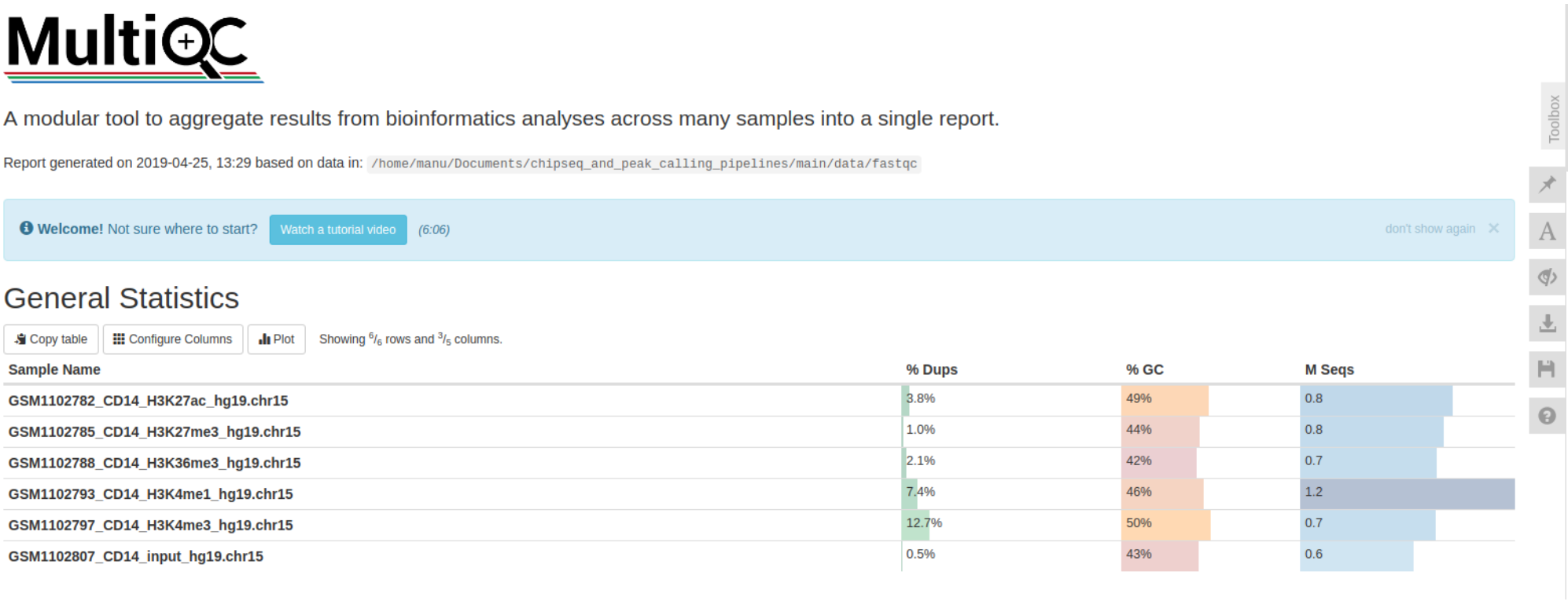
```
##########################################################################################

$PWD =pwd
# bash: {current directory}: Is a directory

mkdir $PWD/main/data
ln -s $PWD/data/* $PWD/main/data/

bash $PWD/main/scripts/fastqc.sh $PWD/main/data

multiqc -f -o $PWD/main/data/fastqc $PWD/main/data/fastqc

    bash $PWD/main/scripts/bowtie.sh hg19 $PWD/data/index/ 0 $PWD/main/data

##########################################################################################
```

The tree should look like this now.

```
.
|-- data
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15.fastq
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15.fastq
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15.fastq
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15.fastq
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15.fastq
|   |-- GSM1102807_CD14_input_hg19.chr15.fastq
|   |-- genes.zip
|   `-- index
|       |-- hg19.1.ebwt
|       |-- hg19.2.ebwt
|       |-- hg19.3.ebwt
|       |-- hg19.4.ebwt
|       |-- hg19.fa
|       |-- hg19.rev.1.ebwt
|       `-- hg19.rev.2.ebwt
`-- main
    |-- data
    |   |-- GSM1102782_CD14_H3K27ac_hg19.chr15.fastq -> /home/manu/Documents/ch
    |   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_fastqc.log
    |   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19.bam
    |   |-- GSM1102785_CD14_H3K27me3_hg19.chr15.fastq -> /home/manu/Documents/c
    |   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_fastqc.log
    |   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19.bam
    |   |-- GSM1102788_CD14_H3K36me3_hg19.chr15.fastq -> /home/manu/Documents/c
    |   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_fastqc.log
    |   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19.bam
    |   |-- GSM1102793_CD14_H3K4me1_hg19.chr15.fastq -> /home/manu/Documents/ch
    |   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_fastqc.log
    |   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19.bam
    |   |-- GSM1102797_CD14_H3K4me3_hg19.chr15.fastq -> /home/manu/Documents/ch
    |   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_fastqc.log
    |   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19.bam
    |   |-- GSM1102807_CD14_input_hg19.chr15.fastq -> /home/manu/Documents/chip
    |   |-- GSM1102807_CD14_input_hg19.chr15_bowtie_hg19.log
    |   |-- GSM1102807_CD14_input_hg19.chr15_fastqc.log
    |   |-- GSM1102807_CD14_input_hg19.chr15_hg19.bam
    |   `-- fastqc
    |       |-- GSM1102782_CD14_H3K27ac_hg19.chr15_fastqc.html
    |       |-- GSM1102782_CD14_H3K27ac_hg19.chr15_fastqc.zip
    |       |-- GSM1102785_CD14_H3K27me3_hg19.chr15_fastqc.html
    |       |-- GSM1102785_CD14_H3K27me3_hg19.chr15_fastqc.zip
    |       |-- GSM1102788_CD14_H3K36me3_hg19.chr15_fastqc.html
    |       |-- GSM1102788_CD14_H3K36me3_hg19.chr15_fastqc.zip
    |       |-- GSM1102793_CD14_H3K4me1_hg19.chr15_fastqc.html
    |       |-- GSM1102793_CD14_H3K4me1_hg19.chr15_fastqc.zip
    |       |-- GSM1102797_CD14_H3K4me3_hg19.chr15_fastqc.html
    |       |-- GSM1102797_CD14_H3K4me3_hg19.chr15_fastqc.zip
    |       |-- GSM1102807_CD14_input_hg19.chr15_fastqc.html
    |       |-- GSM1102807_CD14_input_hg19.chr15_fastqc.zip
    |       |-- multiqc_data
    |       |   |-- multiqc.log
    |       |   |-- multiqc_data.json
    |       |   |-- multiqc_fastqc.txt
    |       |   |-- multiqc_general_stats.txt
    |       |   `-- multiqc_sources.txt
```

```
##########################################################################################

multiqc -f -o $PWD/main/data/bams_qc $PWD/main/data/*_bowtie*.log

##########################################################################################
```

The multiqc report should be generated now.



Now we are going to use convert our data into bigWig format. BigWig is a cool format that allows for fast visualisation of the alignment data. But the trade off is that these files are quite big in size. On UCSC browser what we see is actually huge condensed data as a graph. For the sake of definition:

bigWig files are indexed binary files which contain alignment data to the genome. It is similar to bed file which contains location of the genome where the read aligned to. Because it is indexed, it can be used to quickly find the location you want.

We will use JBR browser to visualise our data.

```
##########################################################################################

bash $PWD/main/scripts/bigwig.sh $PWD/main/data

##########################################################################################
```

Our data is now ready for peak calling.

We will run peak calling on two settings, at 0.05 with NARROW mode (abscence of --broad parameter) and 0.1 which has BROAD mode on by specifying --broad parameter. Narrow mode has a stringent cutoff while broad mode has a relaxed cutoff,

Both have their pros and cons, for example, BROAD mode covers more peaks but NARROW mode will give less peak with good quality.

```
##########################################################################################

bash $PWD/main/scripts/macs2.sh $PWD/main/data hg19 "q0.05" "-q 0.05"

ls $PWD/main/data/*macs*.log

bash $PWD/main/scripts/macs2.sh $PWD/main/data hg19 "broad_0.1" "--broad --broad-cutoff 0.1"

##########################################################################################
```

Let's look at the executed command.

```
$ bash $PWD/main/scripts/macs2.sh $PWD/main/data hg19 "broad_0.1" "--broad --broad-cutoff 0.1"
Batch macs2 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/data hg19 broad_0.1 --broad --broad-cutoff 0.1
GSM1102782_CD14_H3K27ac_hg19.chr15_hg19.bam: control file: GSM1102807_CD14_input_hg19.chr15_hg19.bam
Macs2 TMP_DIR: /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/data/macs2.DeZUYd2h
GSM1102782_CD14_H3K27ac_hg19.chr15_hg19.bam: control file found: GSM1102807_CD14_input_hg19.chr15_hg19.bam
INFO  @ Thu, 25 Apr 2019 23:37:20:
# Command line: callpeak --tempdir /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/data/macs2.DeZUYd2h -t GSM1
bam -f BAM -g hs -n GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1 --broad --broad-cutoff 0.1
# ARGUMENTS LIST:
# name = GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1
# format = BAM
# ChIP-seq file = ['GSM1102782_CD14_H3K27ac_hg19.chr15_hg19.bam']
# control file = ['GSM1102807_CD14_input_hg19.chr15_hg19.bam']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff for narrow/strong regions = 5.00e-02
# qvalue cutoff for broad/weak regions = 1.00e-01
# The maximum gap between significant sites is assigned as the read length/tag size.
# The minimum length of peaks is assigned as the predicted fragment length "d".
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is on
# Paired-End mode is off
```

Now is a good time to organise the folders for further analysis. Because we are going to do some mental gymnastics.

Peak Calling

Peak calling programs help to define sites of Protein:DNA binding by identifying regions where sequence reads are enriched in the genome after mapping. The common assumption is that the ChIP-seq process is relatively unbiased so reads should accumulate at sites of protein binding faster than in background regions of the genome.

Peak calling with MACS2

MACS is (for TF peaks) one of the most popular peak callers, it is also one of the oldest and this probably contributes to its success. It is a good method, good enough for many experimental conditions and requires very little justification if cited as the tool used in a publication. MACS performs removal of redundant reads, read-shifting to account for the offset in forward or reverse strand reads. It uses control samples and local statistics to minimize bias and calculates an empirical FDR.

[https://epigenie.com/guide-peak-calling-for-chip-seq/ (https://epigenie.com/guide-peak-calling-for-chip-seq/)]

We specified the CHIP-seq file and control file. We will have only 5 outputs because one of the bams serve as a control. In our case it's CD4 lymphocytes. Defining the read length parameter is optional because MACS2 can detect read length automatically.

It will then remove the duplicate reads. IT odes so by calculating the maximum number of duplicate reads in a single position taking into account the sequencing depth and will remove any read which is in excess of the calculated sequencing depth. This is done to prevent any bias in the further caluclations.

We have an effective genome size is about 10 % because we only have the 15th chromosome. It's nothing to concern in our situation.

Our bandwidth is 300. In other words, a window will slide where the algorithm will try to find enriched regions which have M-fold sensitivity. The size of window is 2 times the bandwidth. The expected background is the number of reads times their length divided by the mappable genome size. Note that the mappable genome size is always less than the real genome size because of repetitive sequence.

The regions' fold enrichment must be higher than 10 and less than 30, but you can change these values if not enough regions are found. A smaller value for the lower cutoff provides more regions for model building, but it can also include spurious data into the model and thereby adversely affect the peak finding results. MACS2 uses 1000 enriched regions to model the distance d between the forward and reverse strand peaks.

In the actual peak detection phase, MACS2 extends the reads in the 3' direction to the fragment length obtained from modeling. If the model building failed or if it was switched off, the reads are extended to the value of the extension size parameter.

If a control sample is available (as in our case), MACS2 scales the samples linearly to the same read number. It then selects candidate peaks by scanning the genome again, now using a window size which is twice the fragment length.

MACS2 calculates a p-value for each peak using a dynamic Poisson distribution to capture local biases in read background levels. **If there is a control sample, it calculates the local background.** In case we are calling peaks with no control, we need to change the mode.

Finally, q-values are calculated using the Benjamini-Hochberg correction. We see that as model fold [5, 50] for tuning this parameter.

```
################################################################################

# Organise our data by folders

mkdir $PWD/main/macs2
mv $PWD/main/data/*q0.05* $PWD/main/macs2/

mkdir $PWD/main/macs2_broad
mv $PWD/main/data/*broad* $PWD/main/macs2_broad/

mkdir $PWD/main/bw
mv $PWD/main/data/*.bw $PWD/main/bw

############
# Use this command (and modify) to get most confident peaks
############

cat $PWD/main/macs2_broad/GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak | sort -k9,9r

################################################################################
```

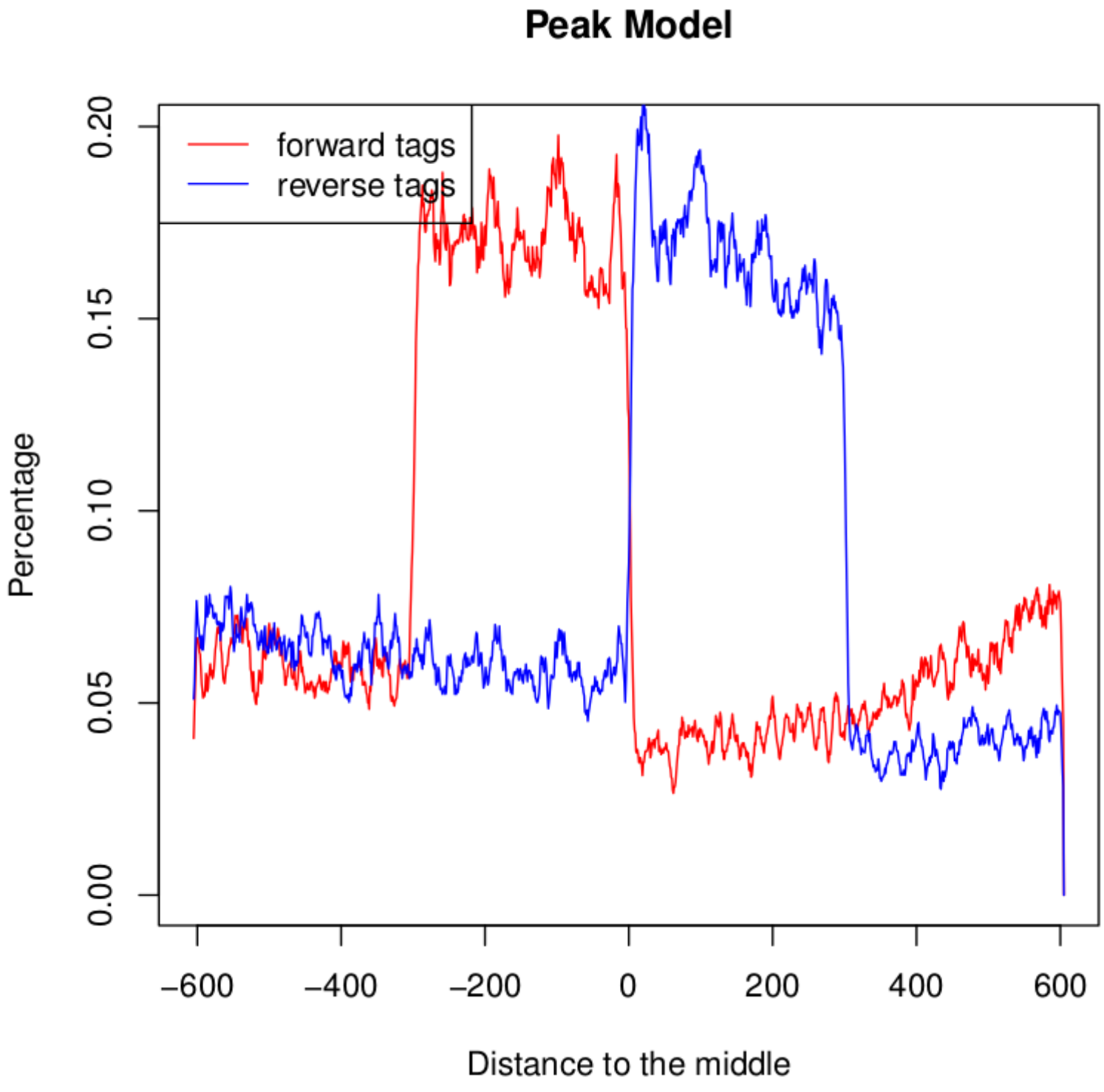
Whatever the selected mode, we can always get the most confident peaks by sorting according the quality and extracting the best peaks.

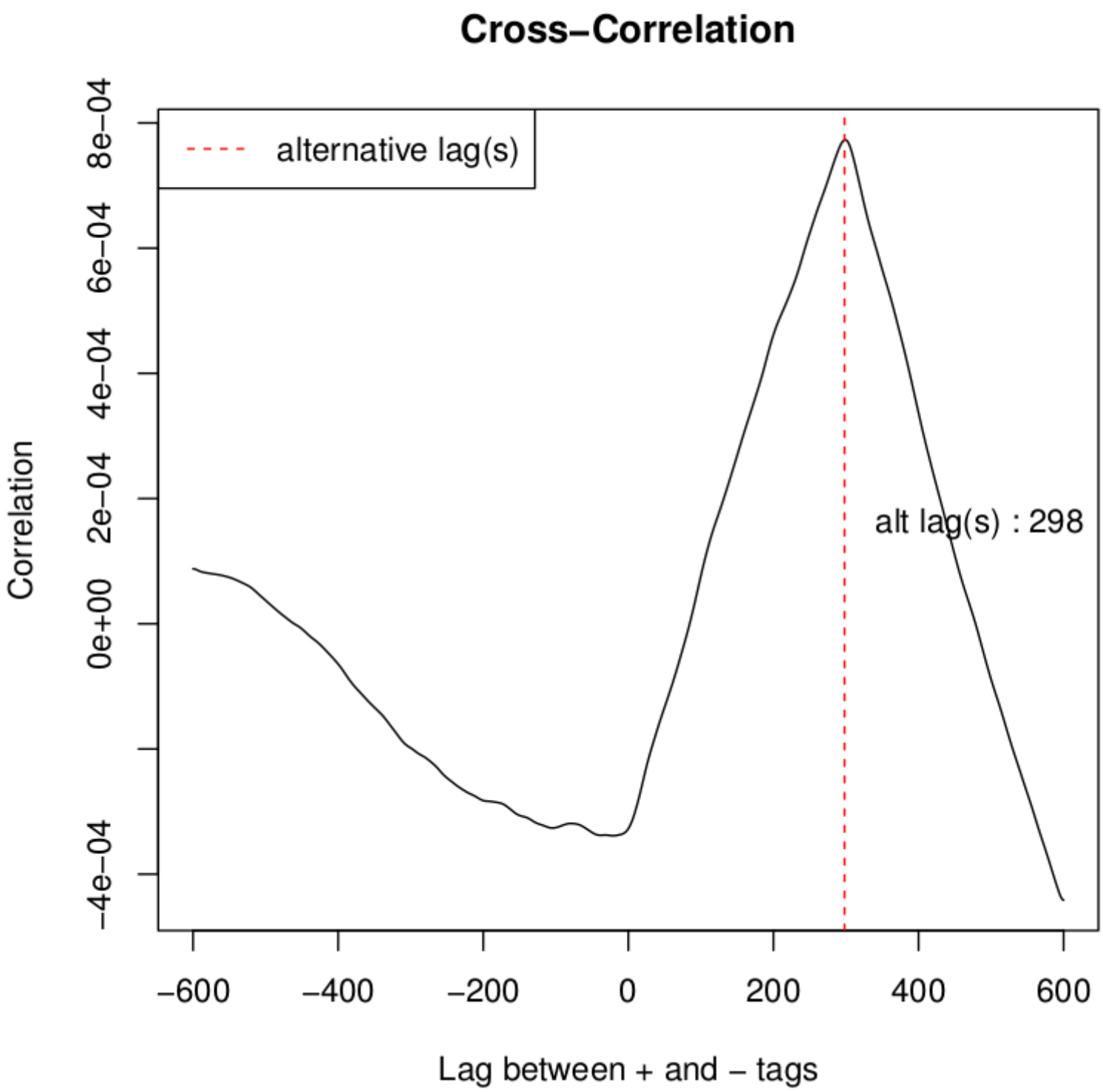
```
##########################################################################################

# Organise our data by folders
mkdir ~/macs2
mv $PWD/main/data/*q0.05* ~/macs2/

mkdir ~/macs2_broad
mv $PWD/main/data/*broad* ~/macs2_broad/

mkdir ~/bw
mb $PWD/main/data/*.bw ~/bw

##########################################################################################
```

Before we move further let's discuss the peak results a bit. Here is the tree structure of the main directory now.

```
$ tree --dirsfirst --filelimit 7
.
|-- bw
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19.bw
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19.bw
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19.bw
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19.bw
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19.bw
|   `-- GSM1102807_CD14_input_hg19.chr15_hg19.bw
|-- data [38 entries exceeds filelimit, not opening dir]
|-- macs2 [30 entries exceeds filelimit, not opening dir]
|-- macs2_broad [30 entries exceeds filelimit, not opening dir]
|-- scripts [13 entries exceeds filelimit, not opening dir]
|-- tools
|   |-- picard-2.10.7.jar
|   `-- span-0.10.0.4787.jar
`-- main.sh
```

```
|-- macs2
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_macs2.log
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_model.pdf
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_model.r
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_peaks.narrowPeak
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_peaks.xls
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_q0.05_summits.bed
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_macs2.log
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_model.pdf
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_model.r
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_peaks.narrowPeak
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_peaks.xls
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_q0.05_summits.bed
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_macs2.log
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_model.pdf
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_model.r
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_peaks.narrowPeak
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_peaks.xls
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_q0.05_summits.bed
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_macs2.log
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_model.pdf
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_model.r
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_peaks.narrowPeak
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_peaks.xls
|   |-- GSM1102793_CD14_H3K4me1_hg19.chr15_hg19_q0.05_summits.bed
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_macs2.log
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_model.pdf
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_model.r
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_peaks.narrowPeak
|   |-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_peaks.xls
|   `-- GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_summits.bed
|-- macs2_broad
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_macs2.log
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_model.pdf
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_model.r
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_peaks.broadPeak
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_peaks.gappedPeak
|   |-- GSM1102782_CD14_H3K27ac_hg19.chr15_hg19_broad_0.1_peaks.xls
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_macs2.log
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_model.pdf
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_model.r
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_peaks.gappedPeak
|   |-- GSM1102785_CD14_H3K27me3_hg19.chr15_hg19_broad_0.1_peaks.xls
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_macs2.log
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_model.pdf
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_model.r
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_peaks.gappedPeak
|   |-- GSM1102788_CD14_H3K36me3_hg19.chr15_hg19_broad_0.1_peaks.xls
```


Peak Model
forward tags
reverse tags
Percentage
0.00
0.05
0.10
0.15
0.20
−600
−400
−200
0
200
400
600
Distance to the middle



The red peak is from Watson (forward) strand, while blue is from Crick (reverse) strand.

The two key features of MACS are: empirical modeling of 'd' (distance) and tag shifting by d/2 to putative protein-DNA interaction site; and the use of a dynamic λ (as in the lambda parameter of a distribution) local to capture local biases in the genome.

You can imagine this as fitting different curves (smoothening procedure for peak) by altering the width and lambda of the curve.

The effectiveness of the dynamic λ{local} is assessed by comparing MACS to a procedure that uses a uniform λ{background} from the genome background.

In other words background helps in maximising signal to noise ratio.

The spatial resolution, defined as the average distance from the peak summit to the nearest FKHR motif, are greatly improved by using tag shifting and the dynamic λ{local}.

In other words, if you do a tag shifting and tune the lambda, you can see two peaks joined together or different based.

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592715/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2592715/)] section: model evaluation.

BUT GUESS WHAT!

All of this can be done by manually annotating peak (10 is enough!) and then do a Machine Learning model to do this fine tuning for us.

Here is why JBR browser is so good, according to segnor Oleg, all you need to do is load the tuned model into JBR and then manually tweak the parameters (e.g. FDR) to see an optimal model.

Sick!!

By the way, FDR is false discovery rate or Number of control peaks / Number of ChIP peaks. **λlocal is critical for ChIP-Seq studies when matching control samples are not available.** In such cases FDR can remain pretty low (say 4%) but for λBackGround it can go upto 50%.

"Macs2 is a Gold Standard when we have control. SPAN is a better choice when we have not a well annotated genome"-- according to Segnor Oleg. (correct me please)

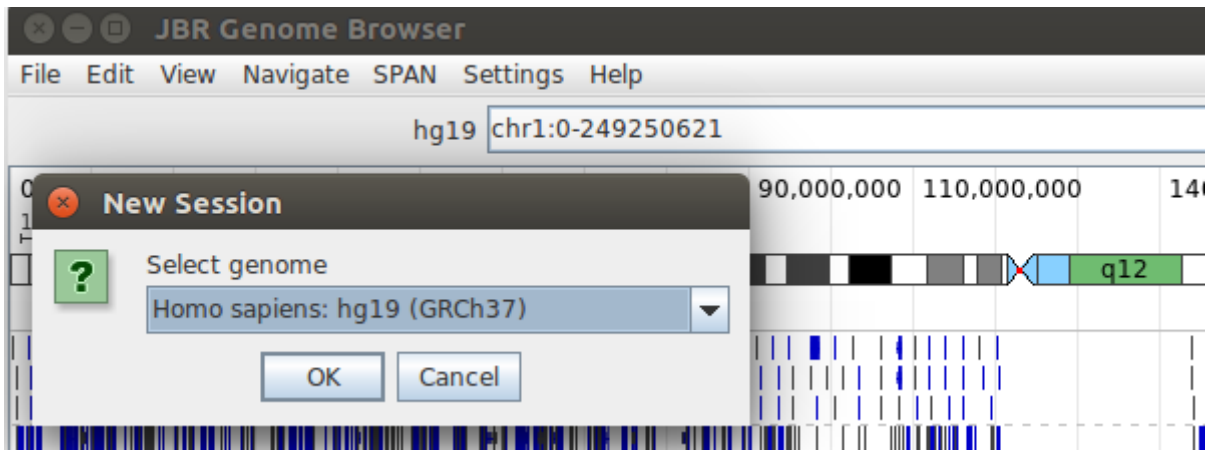
Here we are going to do SPAN and Sicer peak calling as well. The pros and cons will be listed accordingly. But before moving on let's observe our results in JBR genome browser. You can download it from here:

https://research.jetbrains.org/groups/biolabs/tools/jbr-genome-browser (https://research.jetbrains.org/groups/biolabs/tools/jbr-genome-browser)

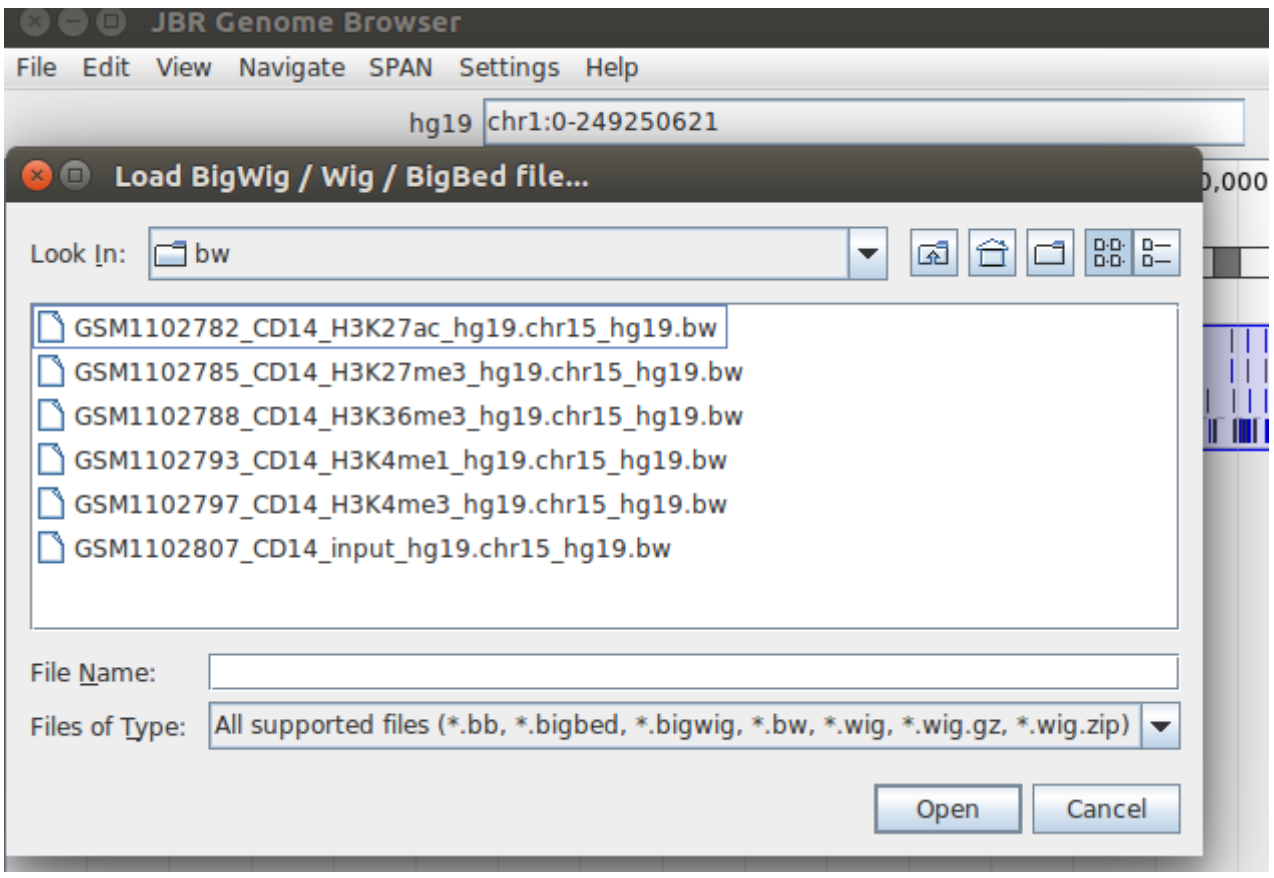
The instructions for installation is posted on the site.

Visualisation in JBR for MACS2 result

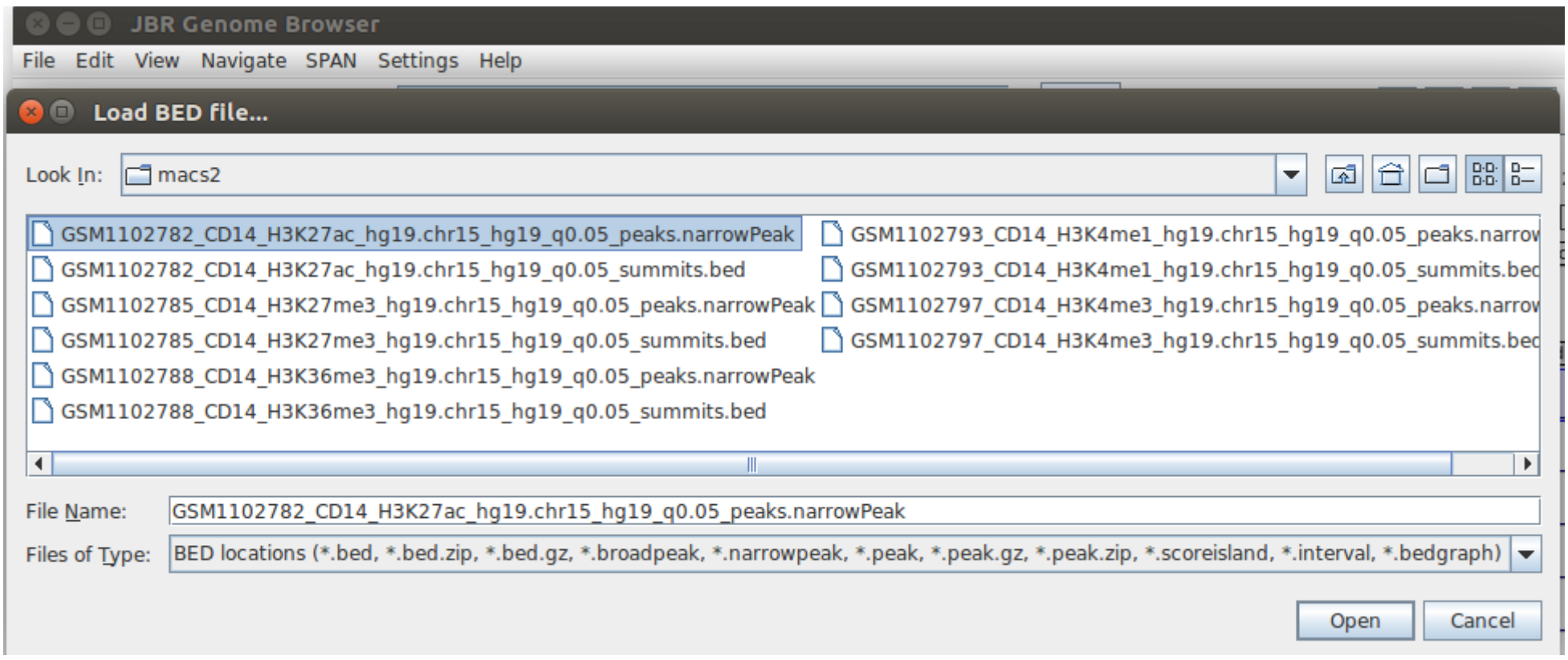
Open JBR and select hg19 reference genome.



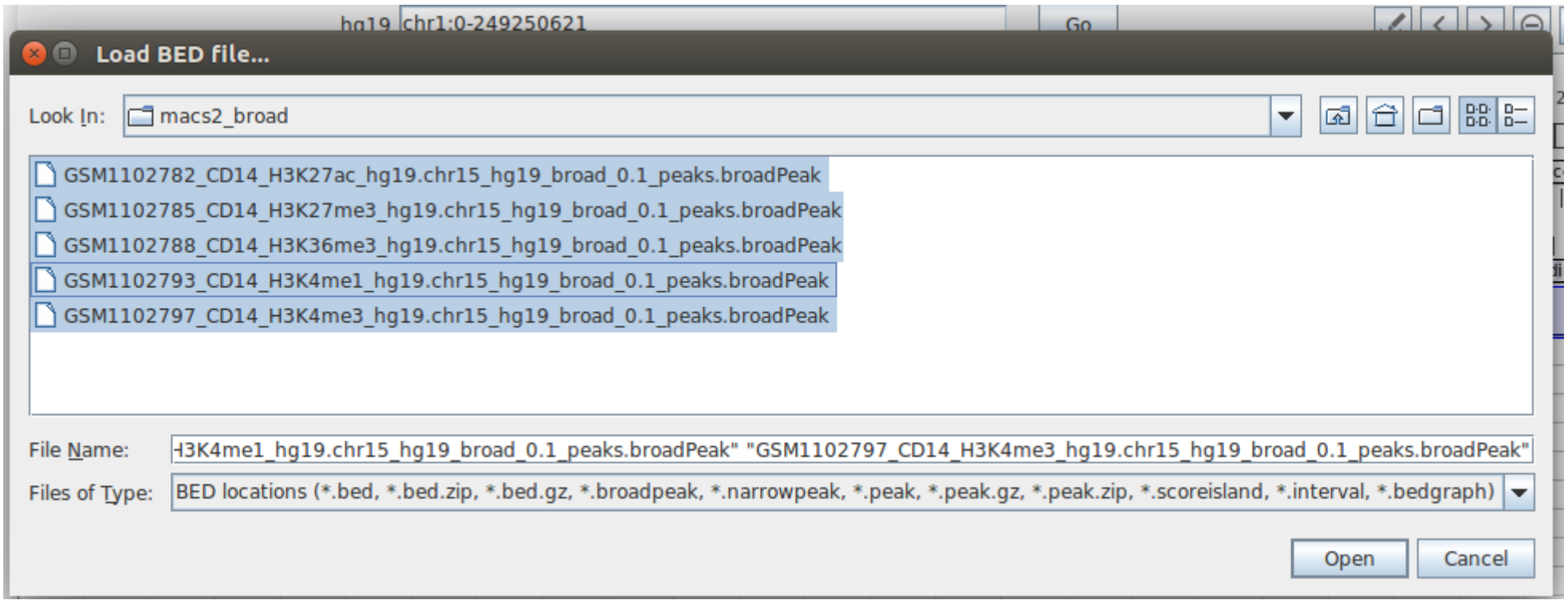
Select all the bigWig files for Samples and Control



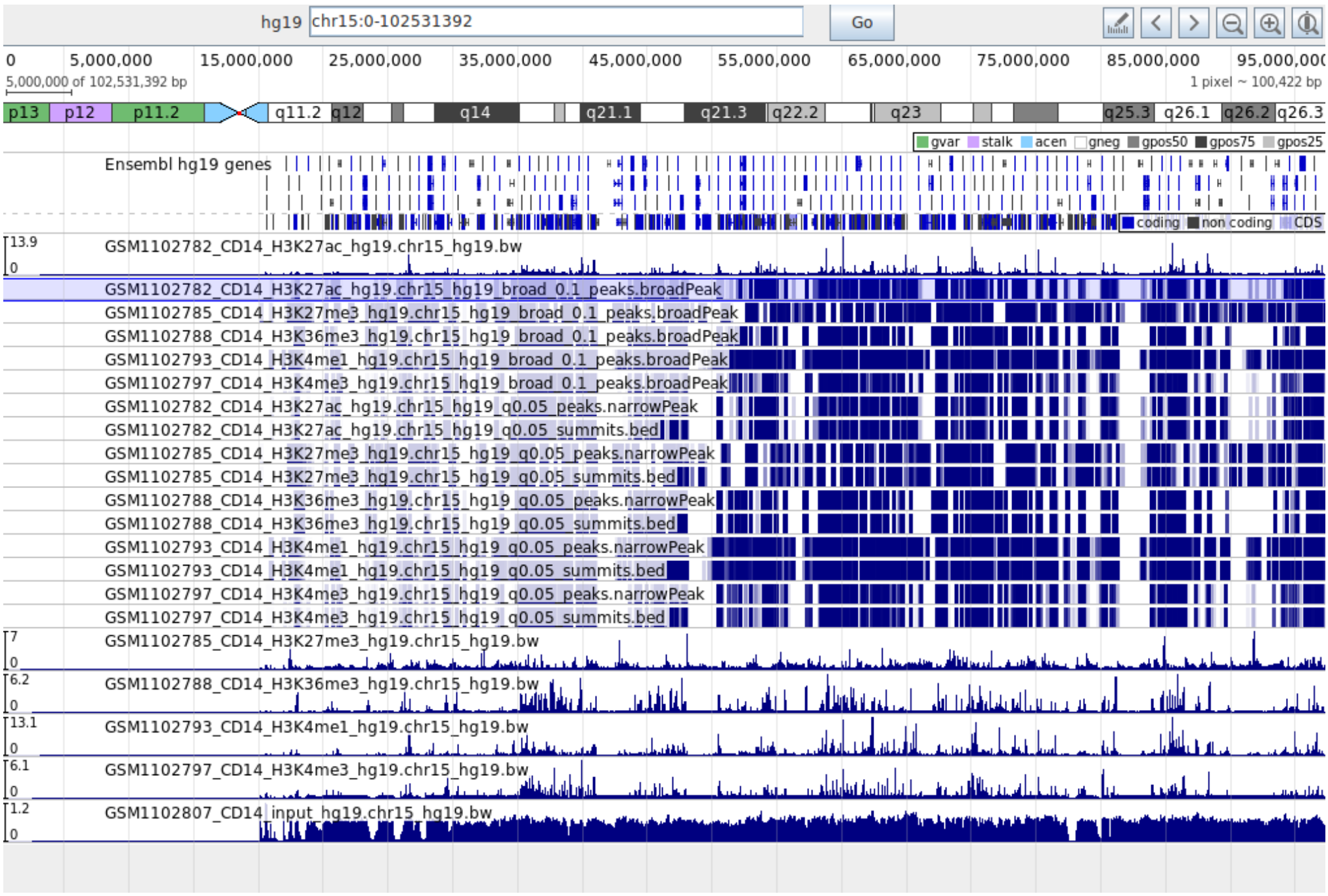
Select all of the files from macs folder.



Select all of the files from macs broad folder.

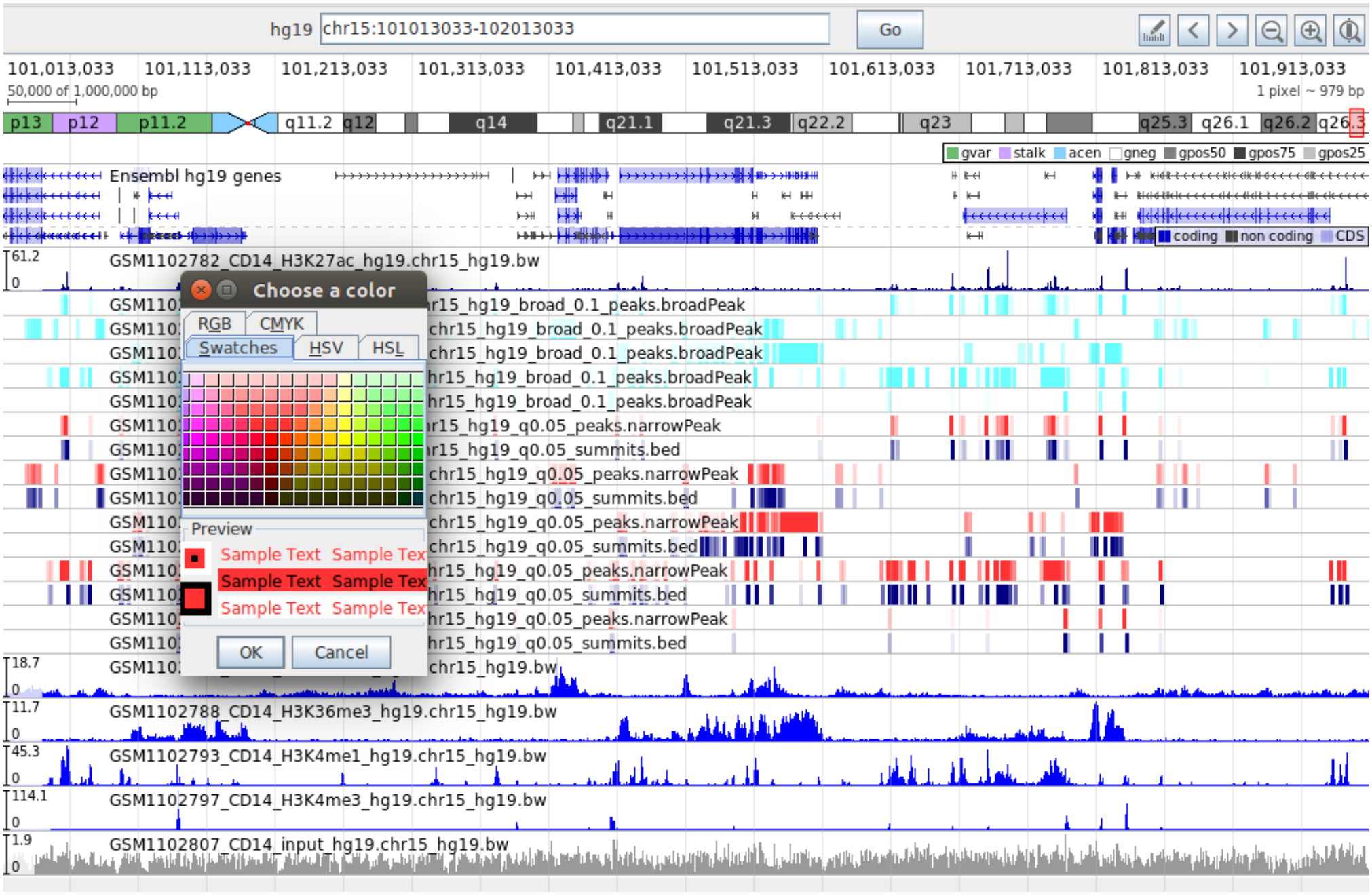


Go to chromosome 15 (why?)



VOILA!

Let's make the results pretty by selecting colours for the tracks.



Peak calling with SCICER

Not all ChIP-seq users are interested in the “peaky” data as seen with transcription factors. However nearly all peak callers were developed for exactly this kind of data. SCICER was developed for more diffuse chromatin modifications that can span kilobases or megabases of the genome. Their method scans the genome in windows and identifies clusters of spatial signals that are unlikely to appear by chance. These clusters or “islands” are used rather than fixed length windows, gaps in the islands are allowed to overcome technical issues (under-saturated experiments, repeat regions, etc). And this gap size can be adjusted for different types of chromatin modification. The program makes use of control data or a random background model.

In order to go further we need the hg19 version 30 annotation file (50.6 MB in size, this is for the **interpretation step**).

```
################################################################################

mkdir $PWD/data/genes

$PWD/data/genes/ wget 'ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_30/GRCh37_mapping/gencod
e.v30lift37.annotation.gtf.gz'

gunzip $PWD/data/genes/gencode.v30lift37.annotation.gtf.gz

cd $PWD

################################################################################
```

```
##########################################################################################

bash $PWD/main/scripts/sicer.sh $PWD/main/data/ hg19 $PWD/data/index/hg19.chrom.sizes 0.05

mkdir $PWD/main/sicer
ls $PWD/main/data/* | grep island | xargs -I {} mv {} $PWD/main/sicer
ls $PWD/main/data/* | grep sicer | xargs -I {} mv {} $PWD/main/sicer

# Command to find out the bed file with biggest number of lines
ls $PWD/main/sicer/*island.bed | xargs wc -l | grep -v total | sort -k1,1nr

# Sort BED file by chromosome and by start position
cat $PWD/main/sicer/GSM1102793_CD14_H3K4me1_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed | sort -k1,1 -k2,2n > s
orted.bed

##########################################################################################
```

lsPWD/main/sicer/*island.bed | xargs wc -l | grep -v total | sort -k1,1nr

2742 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/sicer/**GSM1102793_CD14_H3K4me1**_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed

2527 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/sicer/**GSM1102785_CD14_H3K27me3**_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed

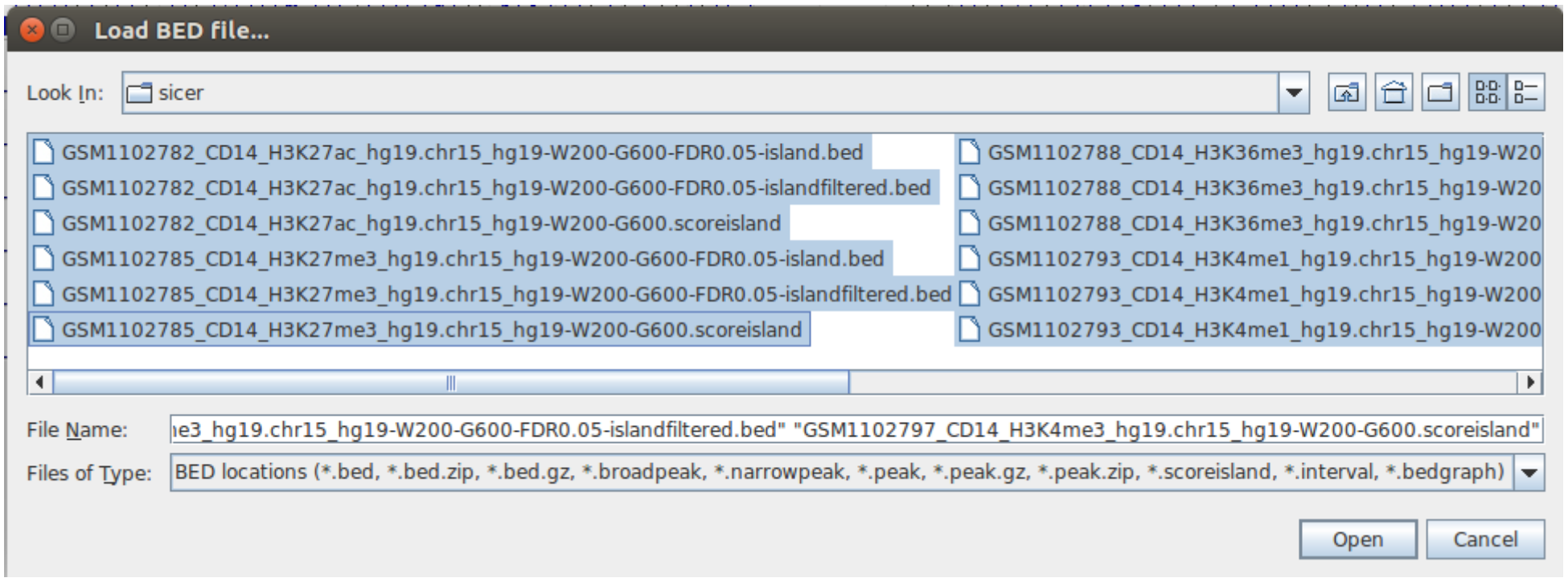
1776 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/sicer/**GSM1102782_CD14_H3K27ac**_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed

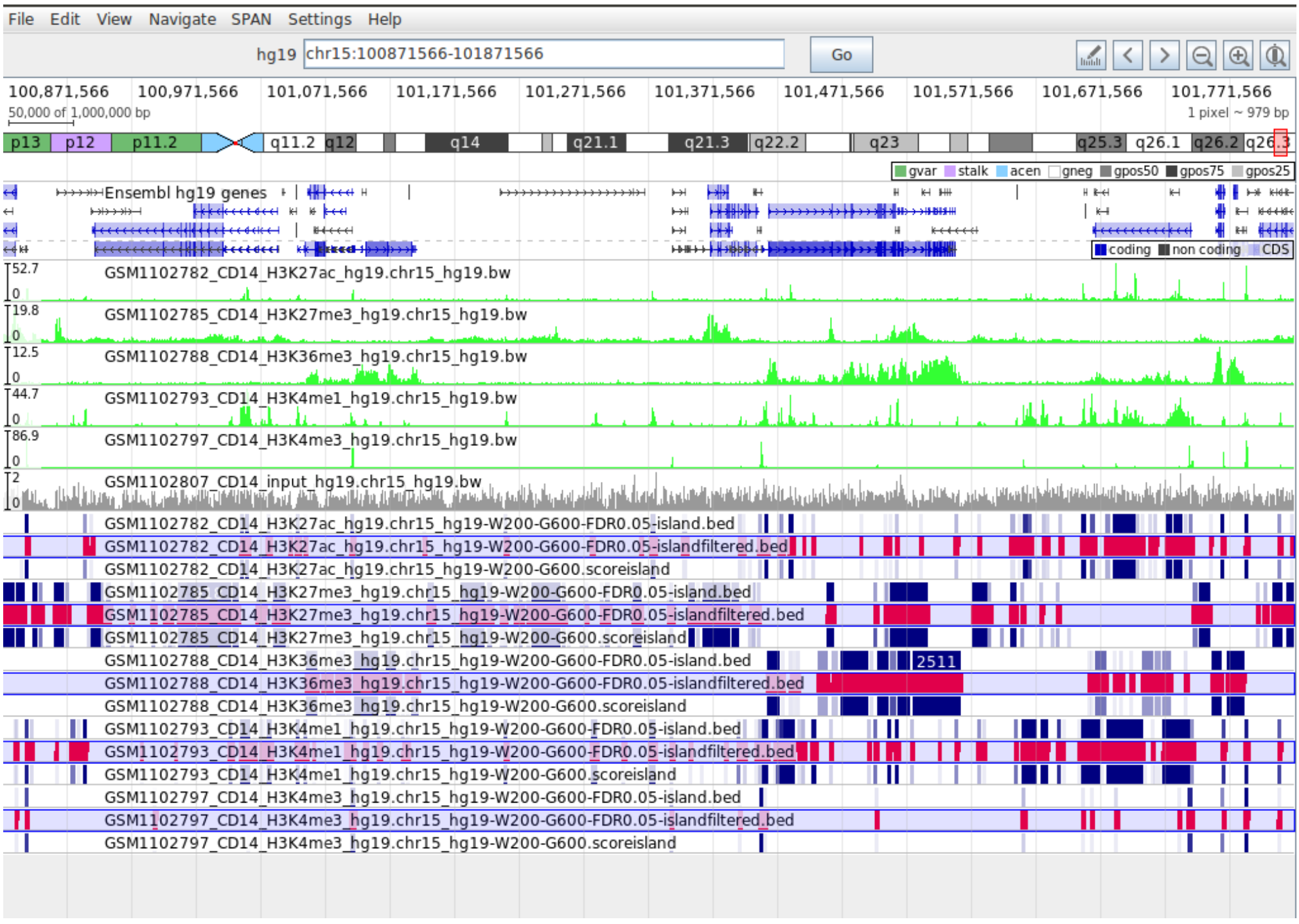
1401 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/sicer/**GSM1102788_CD14_H3K36me3**_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed

747 /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/sicer/**GSM1102797_CD14_H3K4me3**_hg19.chr15_hg19-W200-G600-FDR0.05-island.bed

(base)

Let's visualise **sorted.bed** file in JBR browser with controls and samples.





Peak calling in SPAN

```
+---------------------------------+
|SPAN Semi-supervised Peak Analyzer|
+-----------------------------|/----+
           ,        ,
      __.-'|'-.__.-'|'-.__
    ='=====|========|====='=
    ~_^~-^~~_~^-^~-~~^_~^~^~^
```

SPAN is a tool for analyzing ChIP-seq / ATAC-seq data supporting ultra-low and single-cell input.

https://research.jetbrains.org/groups/biolabs/tools/span-peak-analyzer (https://research.jetbrains.org/groups/biolabs/tools/span-peak-analyzer)

We have the files for SPAN and Picard in /main/tools.

If you read the model fitting section, you would have read the following:

Model fitting

SPAN workflow consists of several steps:

1.) Convert raw reads to tags using user-supplied FRAGMENT parameter or maximum cross-correlation estimate. 2.) Compute coverage for all genome tiled into bins of BIN base pairs. 3.) Fit 3-state hidden Markov model that classifies bins as ZERO states with no coverage, LOW states of non-specific binding, and HIGH states of the specific binding. 4.) Compute posterior HIGH state probability of each bin. 5.) Trained model is saved into .span binary format. 6.) Peaks are computed using trained model and FDR and GAP parameters. 7.) If LABELS are provided, optimal parameters are computed to conform with them.

Model fitting mode produces trained model file in binary format as output, which can be:

1.) visualized directly in JBR Genome Browser used in integrated peak calling pipeline 2.) used in integrated peak calling pipeline

If you launch just: $**bash$PWD/main/scripts/span.sh**

Batch SPAN Need at least 4 parameters! [] [] []

So our command has three additional parameters, which we don't really need to specify. We will do this tuning directly in JBR.

```
################################################################################

# Command to launch SPAN peak caller

bash $PWD/main/scripts/span.sh $PWD/main/tools/span-0.10.0.4787.jar $PWD/main/data/ hg19 $PWD/data/index/hg19.ch
rom.sizes

# Command to get all SPAN models

tree $PWD/main/data/ | grep "\.span"

# Move all the SPAN models to a single folder

mkdir $PWD/main/span_models
find $PWD/main/data/ -name "*.span" | xargs -I {} mv {} $PWD/main/span_models/

################################################################################
```

The basic idea is that once we have peaks, we can open JBR browser and manually annotate upto 10 peaks and train the model to identify peak. The end result is getting a model which we can 'autotune' in JBR for best settings.

If someone is facing troubles with the first command of SPAN (as I did too), the results are present in **span models** directory in $PWD/data.

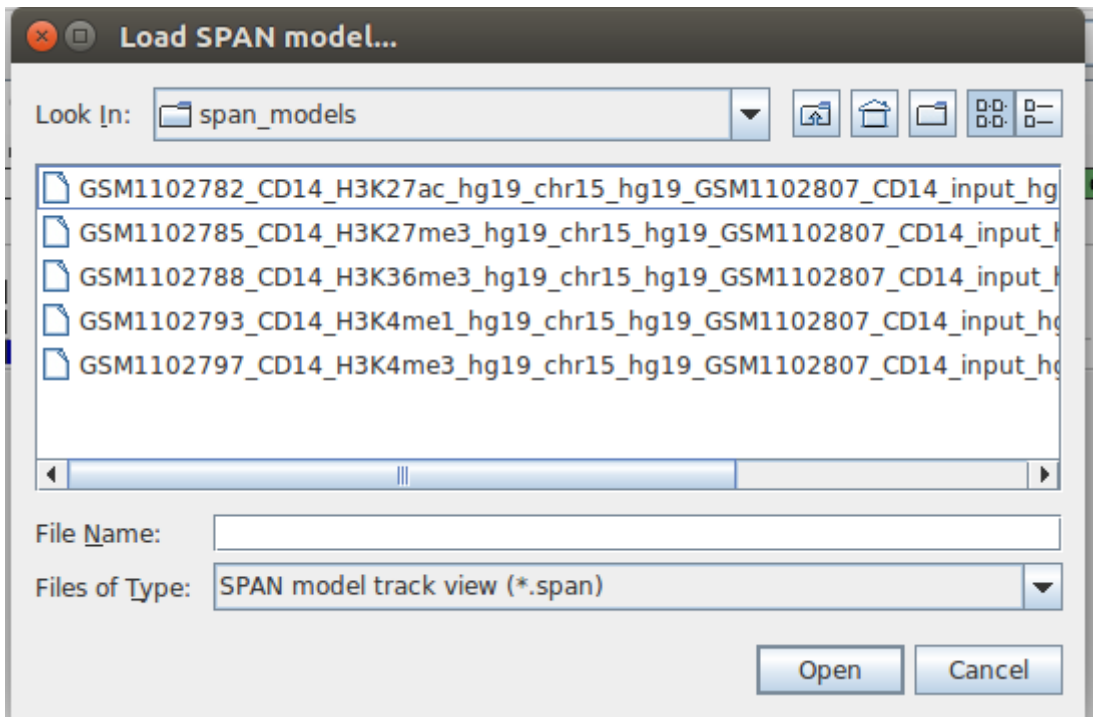
```
mkdir $PWD/main/span_models
cp $PWD/data/span_models/* $PWD/main/span_models
```

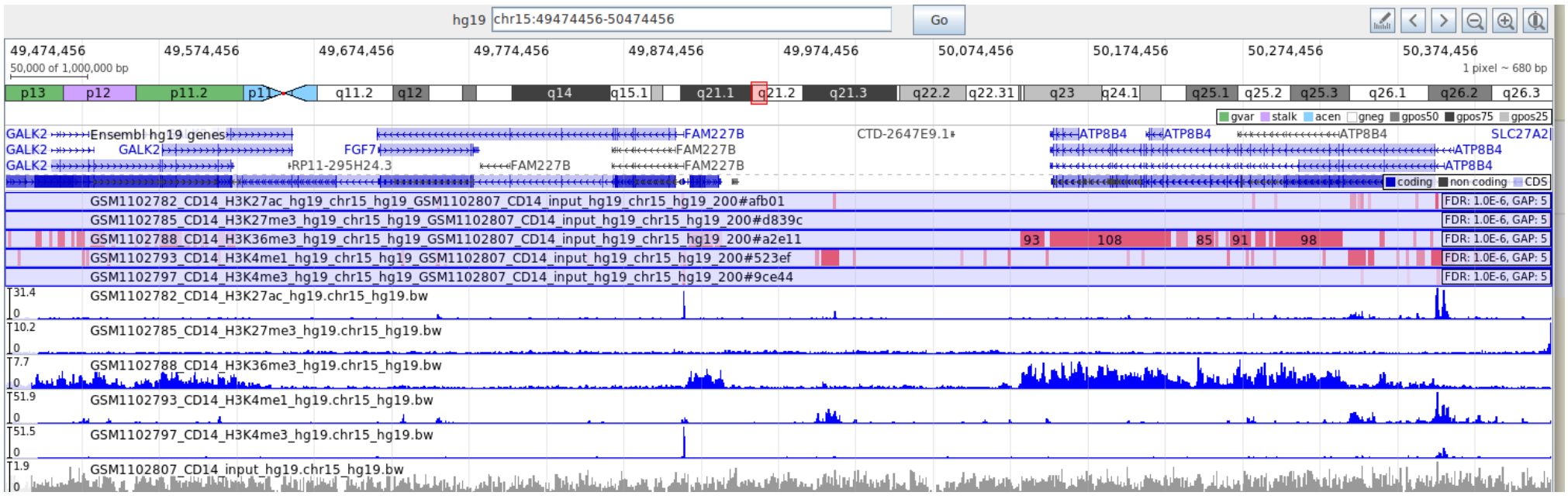
```
$ls $PWD/main/span_models
```

GSM1102782_CD14_H3K27ac_hg19_chr15_hg19_GSM1102807_CD14_input_hg19_chr15_hg19_200#afb01.span
GSM1102785_CD14_H3K27me3_hg19_chr15_hg19_GSM1102807_CD14_input_hg19_chr15_hg19_200#d839c.span
GSM1102788_CD14_H3K36me3_hg19_chr15_hg19_GSM1102807_CD14_input_hg19_chr15_hg19_200#a2e11.span
GSM1102793_CD14_H3K4me1_hg19_chr15_hg19_GSM1102807_CD14_input_hg19_chr15_hg19_200#523ef.span
GSM1102797_CD14_H3K4me3_hg19_chr15_hg19_GSM1102807_CD14_input_hg19_chr15_hg19_200#9ce44.span

(base)

Let's load these models in JBR.

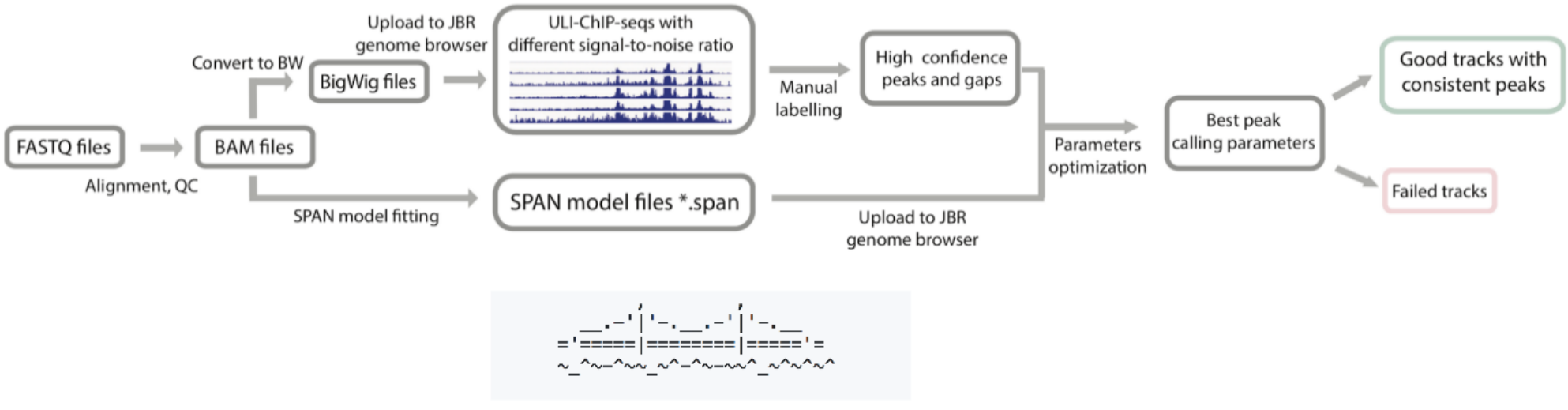




Semi supervised peak calling in SPAN

To exploit the complete power of SPAN, we can load model and manually annotate it. Here is the complete pipleline available: [https://artyomovlab.wustl.edu/aging/howto.html (https://artyomovlab.wustl.edu/aging/howto.html)]

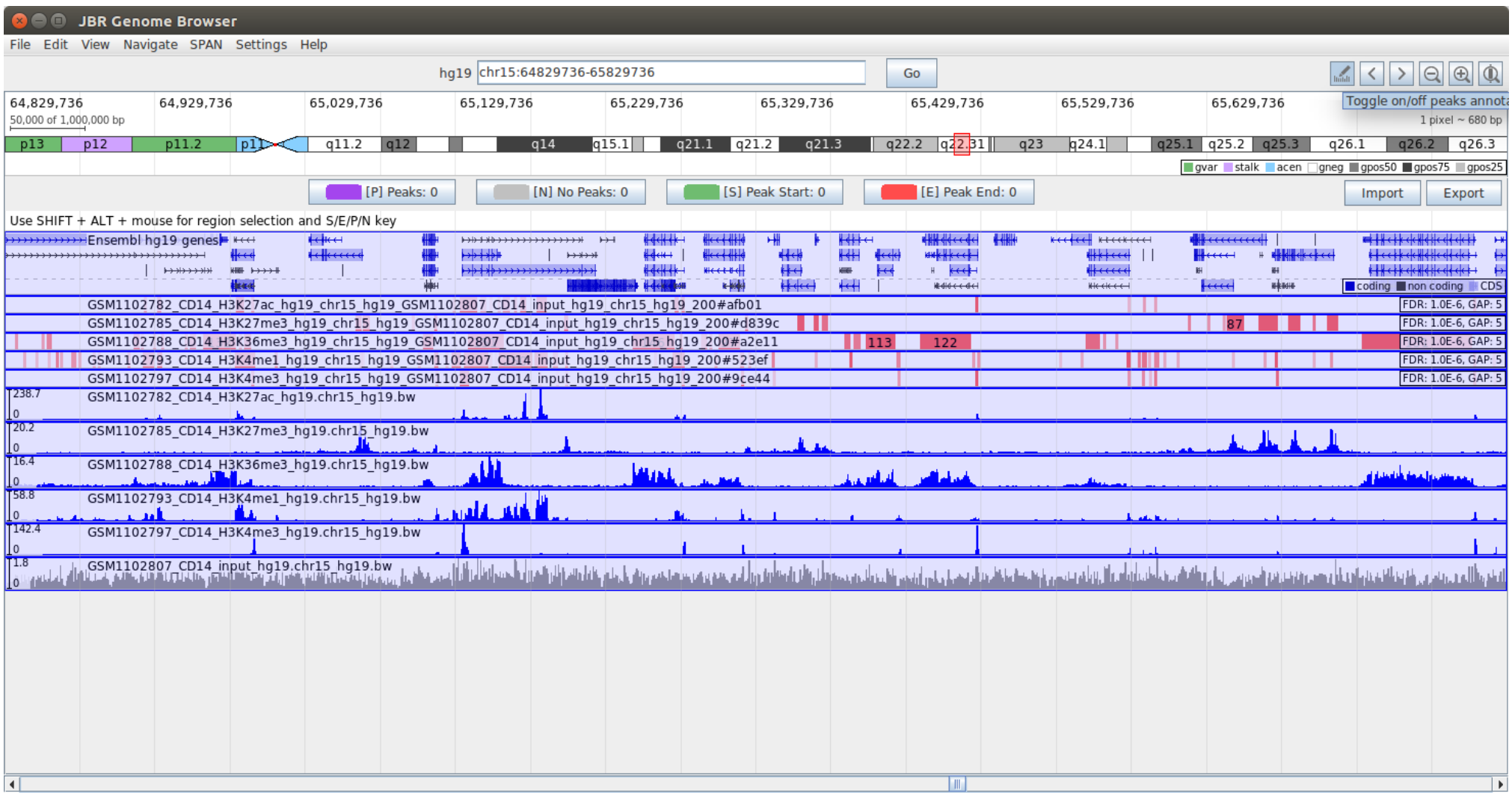
But I am going to give a basic idea of the steps anyways:



To do this we start to annotate some peaks (say 10) within JBR browser, by observing the bigWig peaks.

PLEASE NOTE THAT WE ARE DOING ANNOTATION OF PEAKS BASED ON BIGWIG FILES AND NOT SPAN MODELS.

The annotations on bigWig will then help to improve the models.




```
To add an annotation, first select region:

Move cursor into any track
Press and hold SHIFT key + click and hold left mouse button + move mouse
Release SHIFT key and mouse button
Then set the annotation type:

Click on one of four label buttons or press s/e/n/p key on keyboard
To clear highlighting press ESC
```

Enrichment analysis

This step of the pipline inculdes annotating the results from peaks we get. In other words, whatever peaks we are observing with positions we try to superimpose them over the gene tracks.

We know already that:

▶ H3K27ac – distinguishes active enhancers from poised ▶ H3K27me3 – repression of transcription; has only one methyltransferase EZH2, which is a part of PRC2 ▶ H3K4me1 – enhancer mark ▶ H3K4me3 – promoters, active transcription ▶ H3K36me3 – gene body, active transcription

Let's begin!

```
################################################################################

# Convert GTF file to BED file, FILTER out the transcripts and SORTING the result and FILTER out all the non-chr
omosome positions and TAB separated. ONLY 15 chromosome is taken into account (this step takes a lot of time whe
n all chromosomes are selected)

cat $PWD/data/genes/gencode.v30lift37.annotation.gtf | grep -v "#" | grep "^chr15" | awk -v OFS='\t'  '($3=="gen
e") {print $1,$4-1,$5,$10}' | sort -k1,1 -k2,2n > $PWD/main/gencode.v30lift37.annotation.bed

################################################################################
```

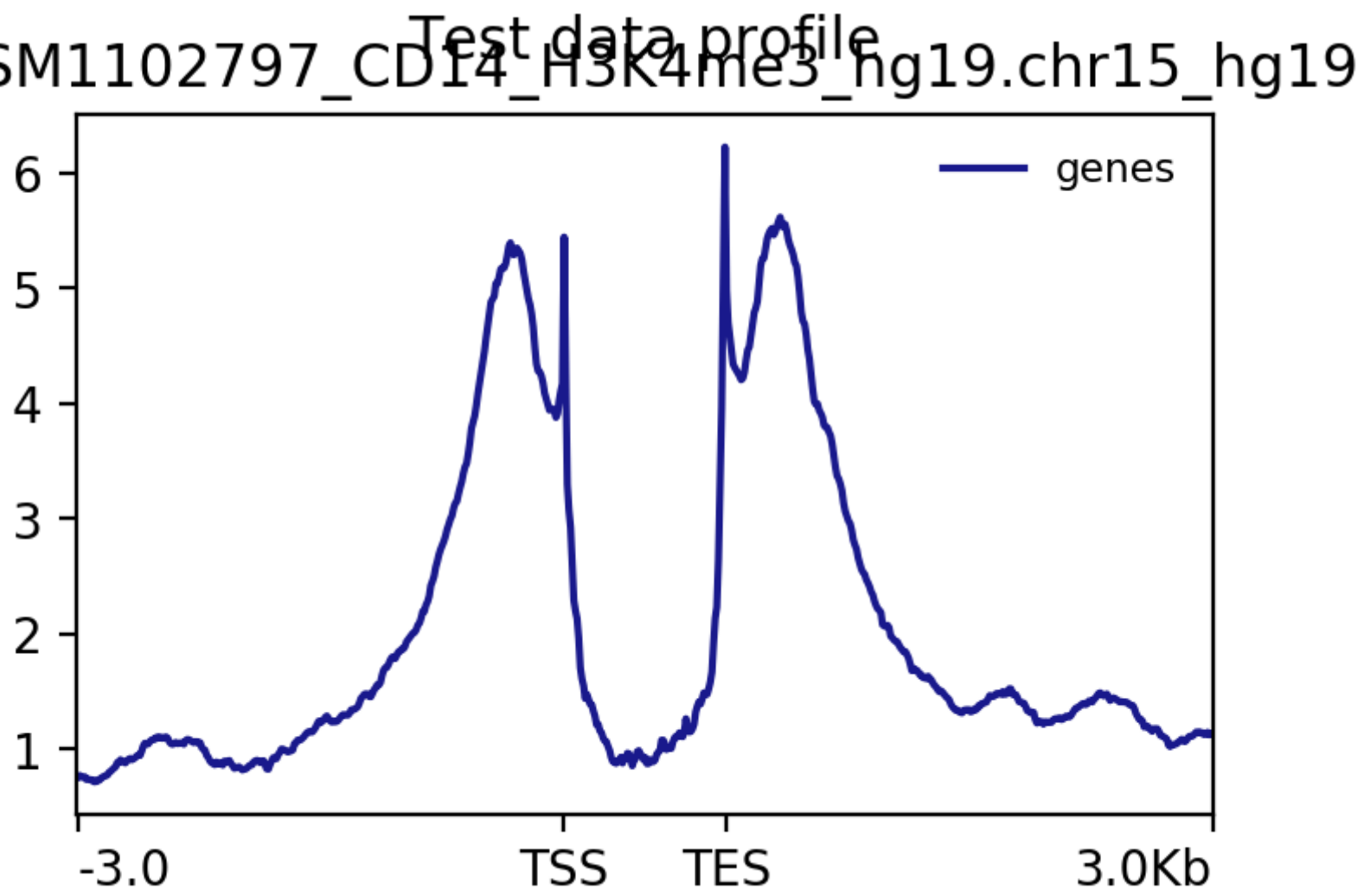
```
$ head /home/manu/Documents/chipseq_and_peak_calling_pipelines/main/gencode.
chr15   20083807        20093067        "ENSG00000215567.5_5";
chr15   20088866        20088969        "ENSG00000201241.1";
chr15   20104586        20104812        "ENSG00000258463.1_5";
chr15   20160552        20160855        "ENSG00000274347.1_5";
chr15   20169918        20170354        "ENSG00000188403.7_2";
chr15   20172666        20173605        "ENSG00000258706.1_4";
chr15   20178034        20178471        "ENSG00000259337.4_2";
chr15   20192908        20193370        "ENSG00000259490.2_2";
chr15   20209092        20209115        "ENSG00000270961.1_4";
chr15   20210049        20210068        "ENSG00000271317.1_2";
(base)
```

Finally we have our .bed file annotated, now we can turn our attention to find the genes and superimpose the results from peak caller.

```
################################################################################

# Find the closest genes for the peak file
bedtools closest -a $PWD/main/macs2_broad/GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak -b
 $PWD/main/gencode.v30lift37.annotation.bed -D ref | head -n 1

################################################################################
```

```
$ bedtools closest -a $PWD/main/macs2_broad/GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak -b $PWD/main/gencode.v30lift37.annotation.bed -D ref | head -n 1
chr15   20711430        20711984        GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_broad_0.1_peak_1        44      .       4.93696 6.54585 4.45528 chr15   20584747        20711433        "ENSG00000180229.12_
3";     0
(base)
```

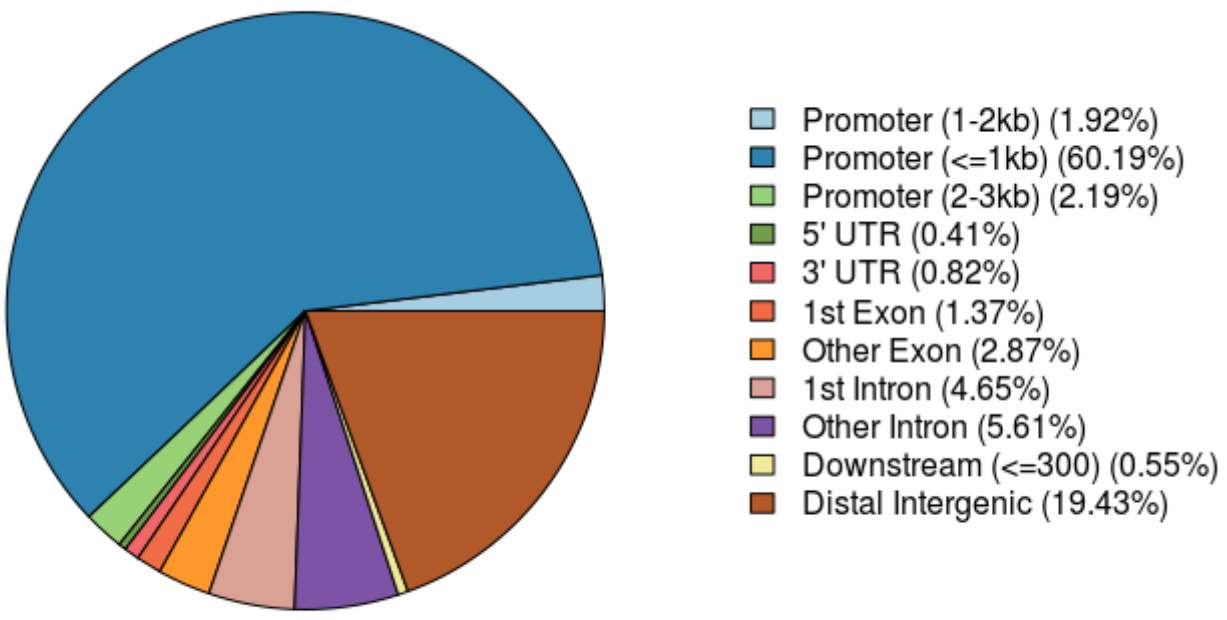
```
##########################################################################################

# Plot profile signal, which consists of 2 steps

# Step 1 - matrix computation

computeMatrix scale-regions -S $PWD/main/bw/GSM1102797_CD14_H3K4me3_hg19.chr15_hg19.bw -R $PWD/main/gencode.v30l
ift37.annotation.bed -a 3000 -b 3000 -out matrix

# Step2 - plot profile

plotProfile -m matrix -out ExampleProfile.png --plotTitle "Test data profile"

##########################################################################################
```

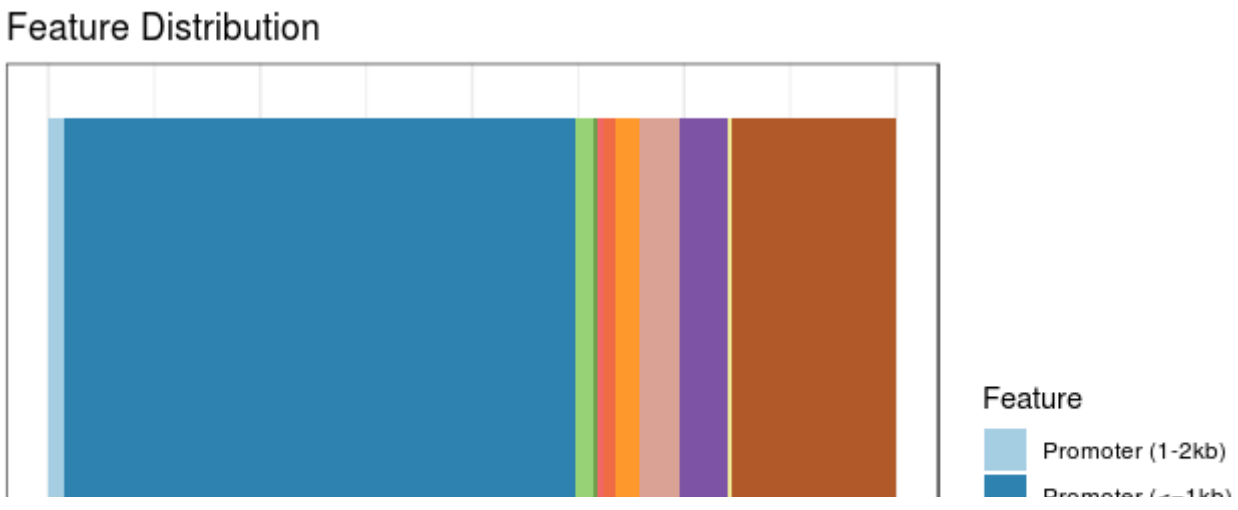


Now we are going to do some R analysis:

Please remember that half of the internet will be installed while trying to install ChIPseeker library (according to segnor Oleg). It took me about 30 minutes to get everything up and running. Just use BiocManager to install the packages.

```
#_____________________________________________________________________________________________

install.packages('BiocManager')
# .libPaths("/tmp/RtmpYs45an/downloaded_packages")
library(BiocManager)

BiocManager::install('ChIPseeker')
BiocManager::install('TxDb.Hsapiens.UCSC.hg19.knownGene')
BiocManager::install('org.Hs.eg.db')

library(ChIPseeker)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)

peaks <- readPeakFile("/home/manu/Documents/chipseq_and_peak_calling_pipelines/main/macs2_broad/GSM1102797_CD14_
H3K4me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak")
peaks

txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
peakAnno <- annotatePeak("/home/manu/Documents/chipseq_and_peak_calling_pipelines/main/macs2_broad/GSM1102797_CD
14_H3K4me3_hg19.chr15_hg19_broad_0.1_peaks.broadPeak", tssRegion=c(-3000, 3000), TxDb=txdb, annoDb="org.Hs.eg.d
b")

plotAnnoPie(peakAnno)
plotAnnoBar(peakAnno)

#_____________________________________________________________________________________________
```





Let's return to bash to do some more analysis. We want to prepare our files for gene-enrichment analysis in GREAT.

GREAT is a 'great' tool developed at harvard which does some cool stuff (Boris can explain), to give you all the genes and pathways observed in your data.

Another great enrichment tools is Mount Sinai MC's Enricher. In-fact Enricher encorporates more pathways than GREAT but GREAT has a 'great' feature up it's sleeve i.e. it can take a **background file**. This acts as kind of a filter and gives us more specific results.

The best thing to do is to combine the results from both of the enrichment analysis tools.

For GSEA (Gene Set Enrichment Analysis) in GREAT we need to reformat our BED file, sadly. Because GREAT is quite strict about the formats, we will do some clipping to our bed file. Additionally, we are going to select top and bottom 100 (total 200) most differentially expressed genes.

```
################################################################################

# Switch back to BASH
# Preparing my peaks file for uploading to GREAT, picking first 3 columns
cat $PWD/main/macs2/GSM1102797_CD14_H3K4me3_hg19.chr15_hg19_q0.05_peaks.narrowPeak | awk -v OFS='\t' '{print
($1,$2,$3)}' > $PWD/main/h3k4me3.bed

cd $PWD/main/

# Pay attention to appending results to a file
head -n 100 h3k4me3.bed > h3k4me3_200.bed
tail -n 100 h3k4me3.bed >> h3k4me3_200.bed

################################################################################
```

GSEA in GREAT

http://great.stanford.edu/public/cgi-bin/greatWeb.php (http://great.stanford.edu/public/cgi-bin/greatWeb.php)

Let's upload our **h3k4me3.bed** file and observe the results. For background we selected the whole genome.

Job Description

Job ID:	20190426-public-3.0.0-qcYsRc
Display name:	h3k4me3.bed
Test set:	h3k4me3.bed (751 genomic regions) Show in UCSC genome browser. *How do I look at my regions in the genome?*
Background:	Whole genome background
Assembly:	Human: GRCh37 (UCSC hg19, Feb 2009) *What gene set does GREAT use?*
Associated genomic regions:	Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 1000.0 kb max extension). Curated regulatory domains are included. 2 of all 751 genomic regions (0.3%) are not associated with any genes. View all genomic region-gene associations. *Which genes are my regions associated with?* Revise the region-gene association rule. *How are my regions associated with genes?*

Overview News Use GREAT Demo Video How to Cite Help Forum — Bejerano Lab, Stanford University

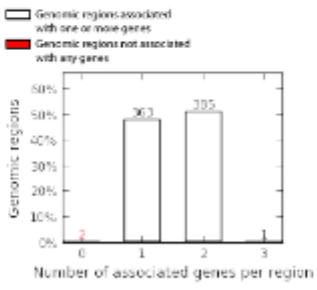
GREAT version 3.0.0 current (02/15/2015 to now)

Job Description

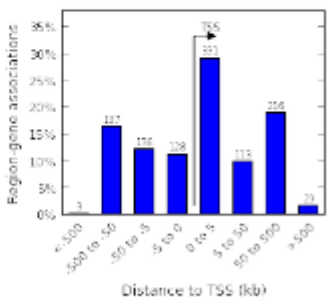
Region-Gene Association Graphs

What do these graphs illustrate?

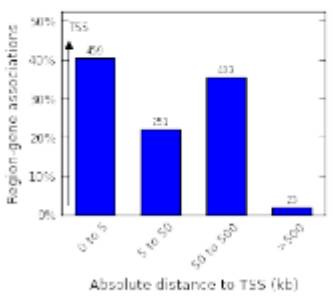
Number of associated genes per region — Download as PDF.



Binned by orientation and distance to TSS — Download as PDF.



Binned by absolute distance to TSS — Download as PDF.



Global Controls — Global Export — *Which data is exported by each option?*

GO Molecular Function (no terms) — Global controls

GO Biological Process (7 terms) — Global controls

Table controls: Export — Shown top rows in this table: 20 Set — Term annotation count: Min: 1 Max: Inf Set — Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
melanosome localization	2	8.6927e-14	4.5376e-10	14.3084	16	2.13%	1	4.8421e-3	13.4836	7	21	1.57%
pigment granule localization	3	1.6278e-13	5.6648e-10	13.7204	16	2.13%	2	3.4756e-3	12.8707	7	22	1.57%
melanosome transport	4	1.9499e-13	5.0893e-10	15.3316	15	2.00%	3	7.5430e-3	14.2767	6	17	1.35%
establishment of melanosome localization	5	3.8004e-13	7.9351e-10	14.6220	15	2.00%	4	8.3092e-3	13.4836	6	18	1.35%
pigment granule transport	6	3.8234e-13	6.6527e-10	14.6158	15	2.00%	4	8.3092e-3	13.4836	6	18	1.35%
establishment of pigment granule localization	8	7.2138e-13	9.4141e-10	13.9695	15	2.00%	6	7.9277e-3	12.7739	6	19	1.35%
cellular pigmentation	40	2.9002e-8	7.5695e-6	5.8978	16	2.13%	7	4.3500e-2	7.6528	7	37	1.57%

The test set of 751 genomic regions picked 446 (2%) of all 18,041 genes.
GO Biological Process has 10,440 terms covering 15,441 (86%) of all 18,041 genes, and 950,065 term - gene associations.
10,440 ontology terms (100%) were tested using an annotation count range of [1, Inf].

GO Cellular Component (no terms) — Global controls

Mouse Phenotype (no terms) — Global controls

Human Phenotype (no terms) — Global controls

Disease Ontology (1 term) — Global controls

Table controls: Export — Shown top rows in this table: 20 Set — Term annotation count: Min: 1 Max: Inf Set — Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
Prader-Willi syndrome	18	4.2294e-7	5.2515e-5	7.4951	11	1.46%	1	1.8891e-4	16.6562	7	17	1.57%

The test set of 751 genomic regions picked 446 (2%) of all 18,041 genes.
Disease Ontology has 2,235 terms covering 7,886 (44%) of all 18,041 genes, and 232,324 term - gene associations.
2,235 ontology terms (100%) were tested using an annotation count range of [1, Inf].

MSigDB Cancer Neighborhood (no terms) — Global controls

Placenta Disorders (no terms) — Global controls

PANTHER Pathway (1 term) — Global controls

Table controls: Export — Shown top rows in this table: 20 Set — Term annotation count: Min: 1 Max: Inf Set — Visualize this table: [select one]

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
Alzheimer disease-amyloid secretase pathway	4	4.9511e-6	1.8814e-4	3.4458	19	2.53%	1	2.3435e-2	5.1366	8	63	1.79%

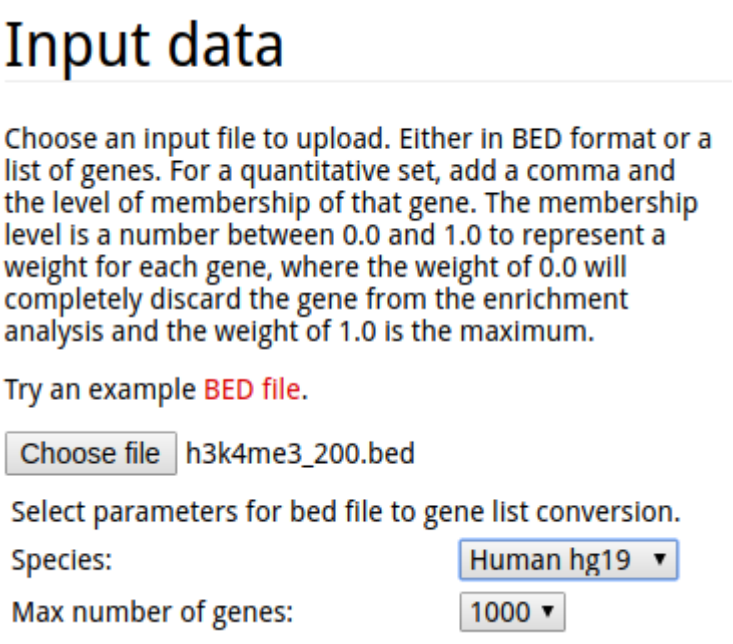
great.stanford.edu/public/cgi-bin/greatWeb.php 1/3

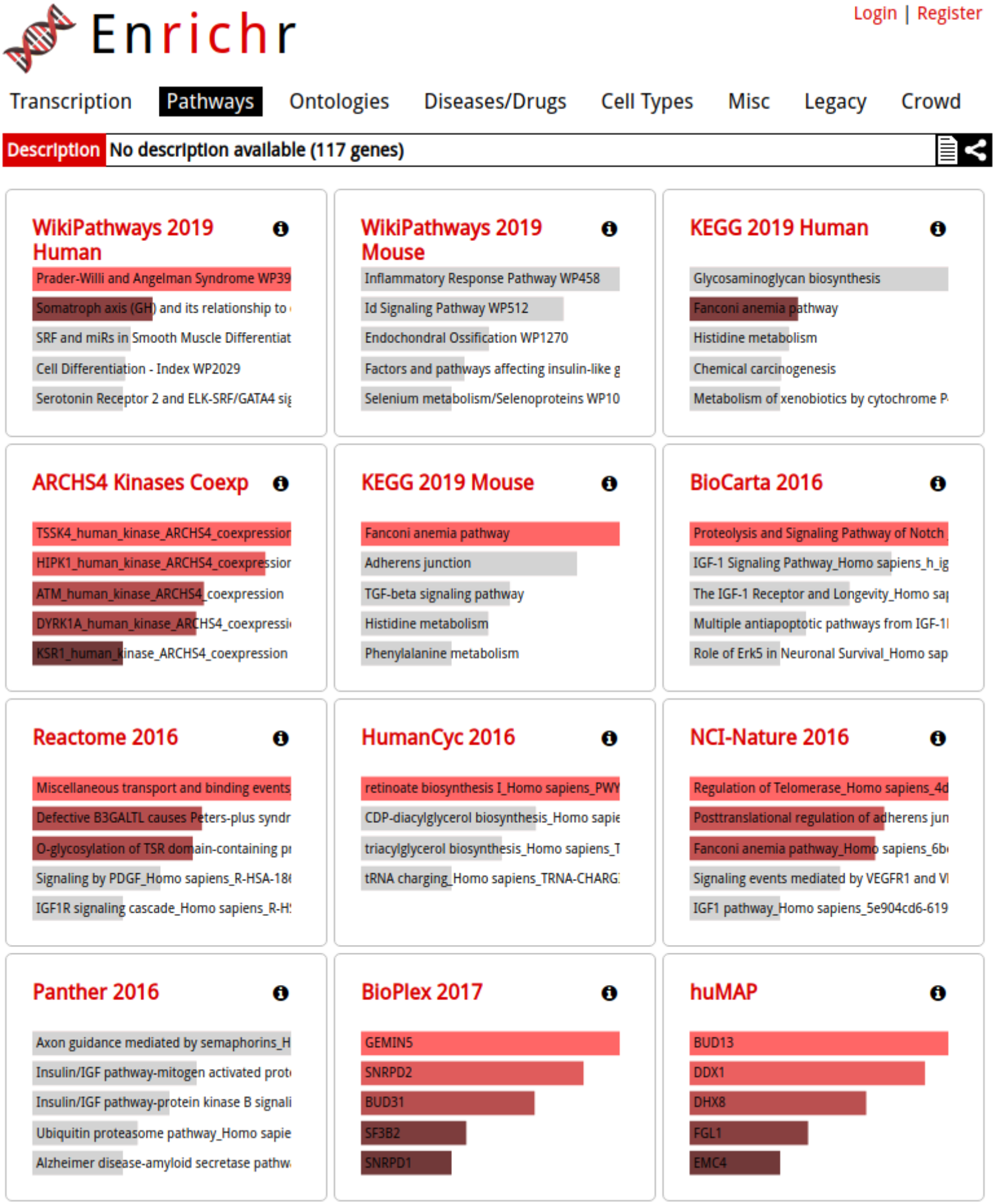
As an exercise, you can upload the top_200 genes bed file, the result should seem surprising. When we were using all genes from Chr15 we were getting lesser pathways enriched, but when we used only 200 genes we get more pathways enriched (why?).

GSEA in Enrichr

http://amp.pharm.mssm.edu/Enrichr/ (http://amp.pharm.mssm.edu/Enrichr/)

Let's upload our **h3k4me3.bed** file and observe the results.





Chip Atlas

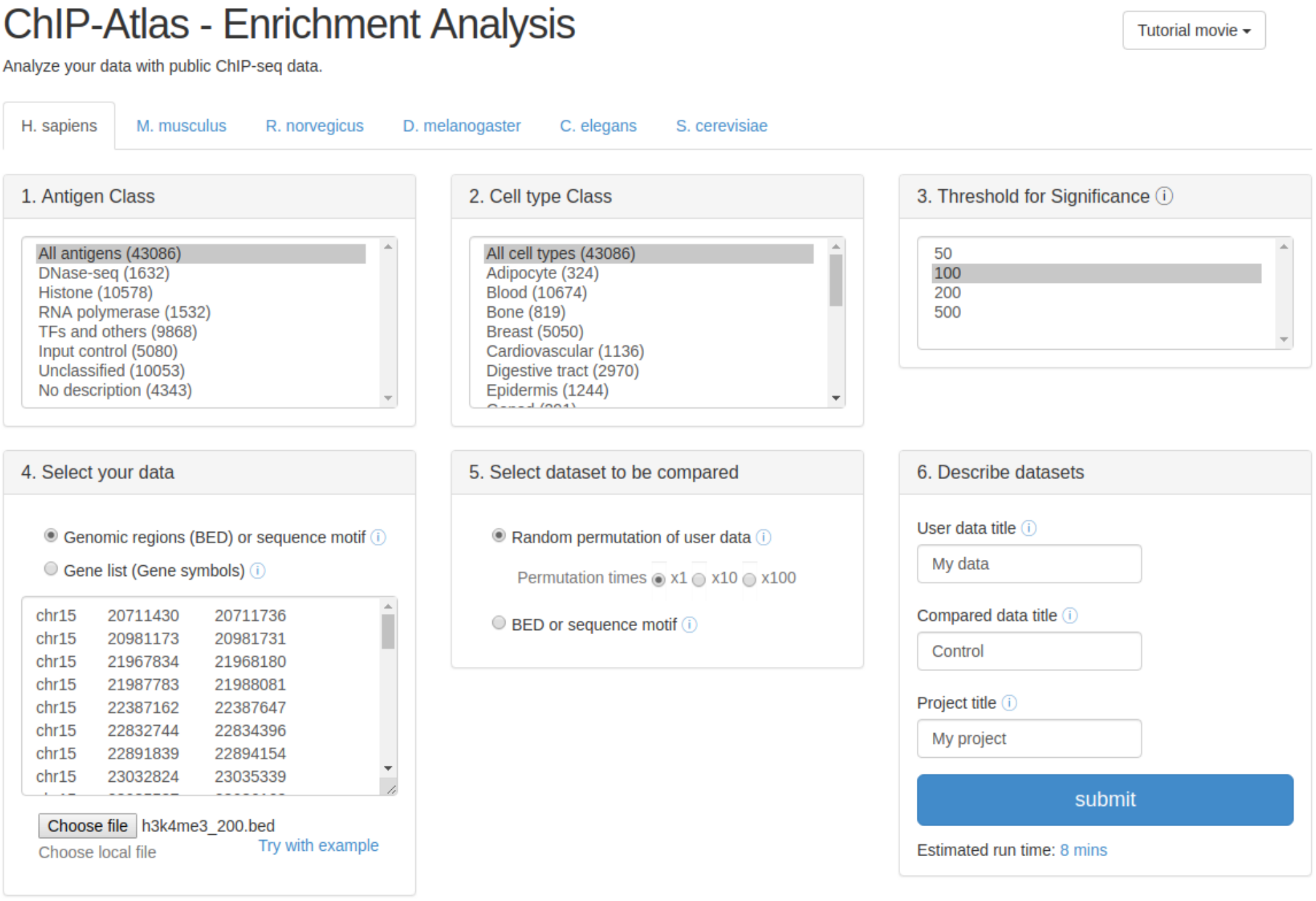
As a final step we are going to use Chip Atlas. It contains all the GEOsets reanalysed for MACS2. We are doing this to get our results in a bigger picrture. We want to find out if there is data out there to confirm our finding.

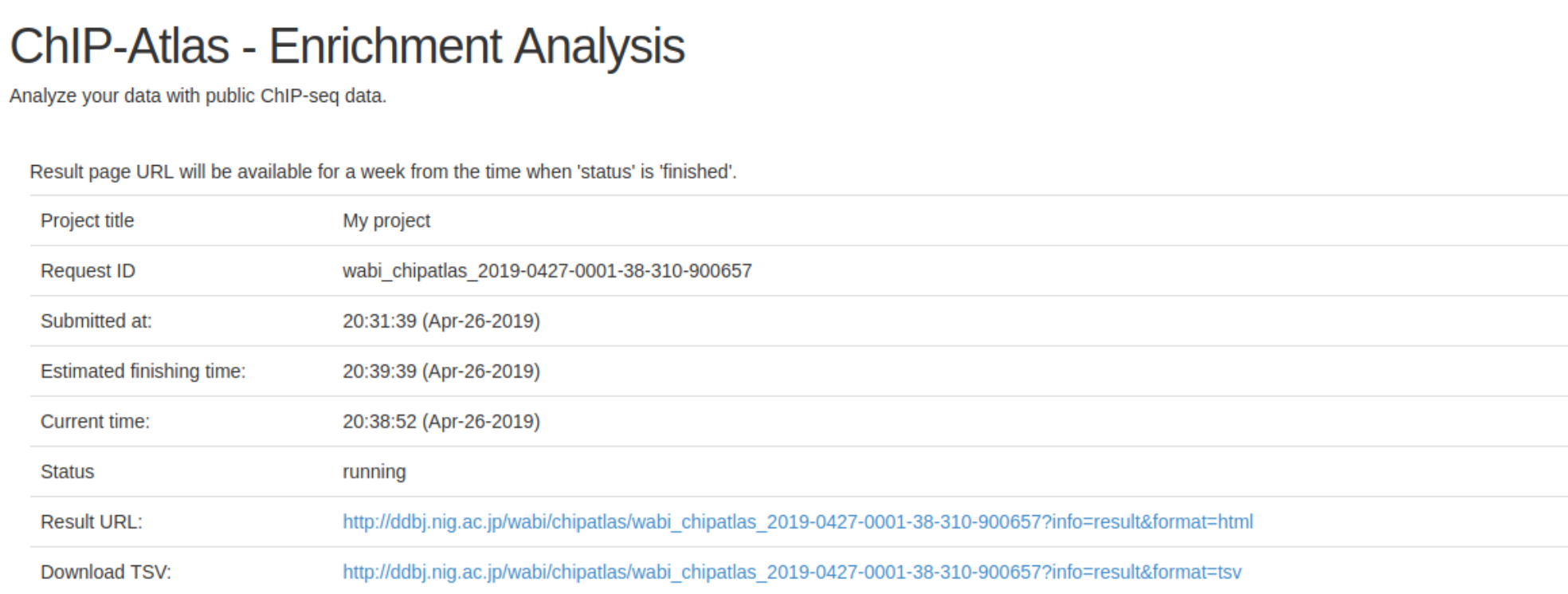
The intuition is that, different studies have the data but their question is different. Having our data compared to these studies greatly helps in hypothesis generation. We can find intresting literature, if the questions are similar to ours, we can also take the considerations from the experiments that have already been performed before us.

Our results include:

▶ Enrichment analysis against multiple ChIP-Seq experiments ▶ Matched peaks and genes

https://chip-atlas.org/enrichment_analysis (https://chip-atlas.org/enrichment_analysis)





ChIP-Atlas / Enrichment Analysis

Search for proteins significantly bound to your data.

Show 100 entries Search:

Telegram

My project

ID	Antigen class	Antigen	Cell class	Cell	Num of peaks	Overlaps / My Experiment	Overlaps / Control	Log P-val	Log Q-val	Fold Enrichment	FE > 1?
SRX2770848	Histone	H3K4me3	Blood	Macrophages	46619	192/200	5/200	-95.5	-91.2	38.40	TRUE
SRX651510	Histone	H3K4me3	Blood	Monocytes	37608	188/200	4/200	-92.0	-87.9	47.00	TRUE
SRX186712	Histone	H3K4me2	Blood	Monocytes-CD14+	42900	188/200	5/200	-90.4	-86.6	37.60	TRUE
SRX749795	Histone	H3K4me3	Blood	Monocytes-CD14+	33331	185/200	4/200	-88.5	-84.7	46.25	TRUE
SRX2004318	Histone	H3K4me1	Blood	Monocytes	28800	184/200	4/200	-87.4	-83.7	46.00	TRUE
SRX2004207	Histone	H3K4me1	Blood	Monocytes	29370	182/200	3/200	-86.8	-83.3	60.67	TRUE
SRX2004218	Histone	H3K4me1	Blood	Monocytes	29791	183/200	4/200	-86.3	-82.9	45.75	TRUE
SRX2004265	Histone	H3K4me1	Blood	Monocytes	28464	183/200	4/200	-86.3	-82.9	45.75	TRUE
SRX2004305	Histone	H3K4me1	Blood	Monocytes	29211	183/200	4/200	-86.3	-82.9	45.75	TRUE
SRX2004238	Histone	H3K4me1	Blood	Monocytes	29230	182/200	4/200	-85.2	-81.9	45.50	TRUE
SRX186732	Histone	H3K4me3	Blood	Monocytes-CD14+	31435	178/200	2/200	-84.5	-81.2	89.00	TRUE
SRX2004222	Histone	H3K4me1	Blood	Monocytes	29543	181/200	4/200	-84.2	-81.0	45.25	TRUE
SRX2004229	Histone	H3K4me1	Blood	Monocytes	29820	181/200	4/200	-84.2	-81.0	45.25	TRUE
SRX2004315	Histone	H3K4me1	Blood	Monocytes	29721	181/200	4/200	-84.2	-81.0	45.25	TRUE
SRX2770849	Histone	H3K4me3	Blood	Macrophages	37964	181/200	4/200	-84.2	-81.0	45.25	TRUE
SRX2004192	Histone	H3K4me1	Blood	Monocytes	28168	179/200	3/200	-83.7	-80.7	59.67	TRUE
SRX2004321	Histone	H3K4me1	Blood	Monocytes	31122	179/200	3/200	-83.7	-80.7	59.67	TRUE
SRX2004201	Histone	H3K4me1	Blood	Monocytes	27646	178/200	4/200	-81.1	-78.1	44.50	TRUE
SRX2004214	Histone	H3K4me1	Blood	Monocytes	27657	174/200	2/200	-80.7	-77.6	87.00	TRUE

Showing 1 to 100 of 15,176 entries Previous 1 2 3 4 5 … 152 Next

The most intresting thing about our result is that ChIP-Atlas was able to pinpoint the Cell Type! **Monocytes-CD14+**

That's all folks!

Please contribute or elaborate to some sections.