
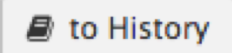
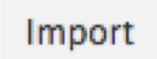



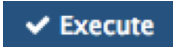
# Analysis of ChIP-seq data (answers to questions)

Stéphanie Le Gras  
(slegras@igbmc.fr)

# Exercise 1: mapping statistics


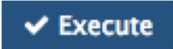
- 2.
  - Click on the button  and select “create new”
  - Click on the history name “Unnamed history”, erase “Unnamed history”, enter “ChIP-seq data analysis” and press enter
- 3.
  - Click on Shared Data (top menu) and select “Data Libraries”
  - Click on “NGS data analysis training ” > “ChIPseq” > “mapping”
  - Select mitf.bam and ctrl.bam datasets (tick boxes beside dataset names)
  - Click on the button 
  - Select history: ChIP-seq data analysis
  - Click on 
  - Go back to the main page by clicking on “Analyzed data” (top menu)

# Exercise 1: mapping statistics

- 4
  - Search for “flagstat” in the search field (tool panel)
  - Click on the name of the tool
  - Click on  to select multiple datasets
  - Select all 2 datasets
  - Click on 

Sample name	No. of raw reads	No. of aligned reads
MITF	31,334,257	23,124,393
Ctrl	29,433,042	19,949,607

# Exercise 2: duplicate reads estimate

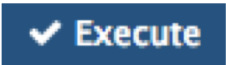
- 1.
  - Search for “markdup” in the search field (tool panel)
  - Click on the name of the tool
  - Click on  to select multiple datasets
  - Select the 2 bam files
  - Select validation stringency: Silent
  - Click on 
  - Open the datasets “MarkDuplicates on data \* : MarkDuplicate metrics”

Sample name	No. of raw reads	No. of aligned reads	No. of duplicate reads
MITF	31,334,257	23,124,393	16,901,318
Ctrl	29,433,042	19,949,607	15,151,227

# Exercise 3: Visualization of the data

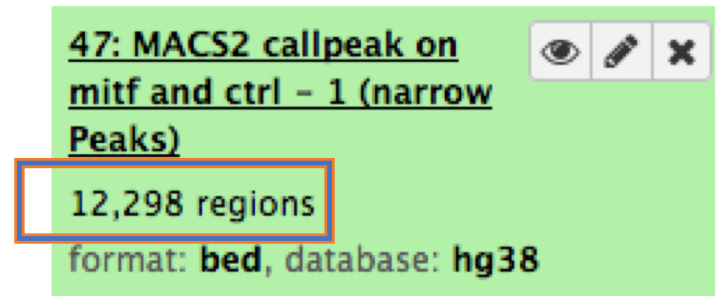
- 1.
  - Idh1 -> No peak
  - NPAS2 -> peak
  - AP1S2 -> Peak,
  - PABPC1I -> No peak
  - Park7 -> No peak
  - Pmel -> Peak
  - Cdk2 -> Peak
  - Actb -> No peak

# Exercise 4: peak calling

- 1.
  - Search for “macs2 callpeak” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - CHIP-Seq Treatment File: mitf.bam
    - CHIP-Seq Control File: ctrl.bam
    - Effective genome size: Human
    - Outputs: select Peaks as tabular file, summits, Summary page (html), Plot in PDF
    - Click on  Execute

# Exercise 4: peak calling

- 2.
  - There is 12,298 peaks




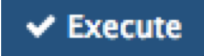
```
47: MACS2 callpeak on mitf and ctrl - 1 (narrow Peaks)
12,298 regions
format: bed, database: hg38
```

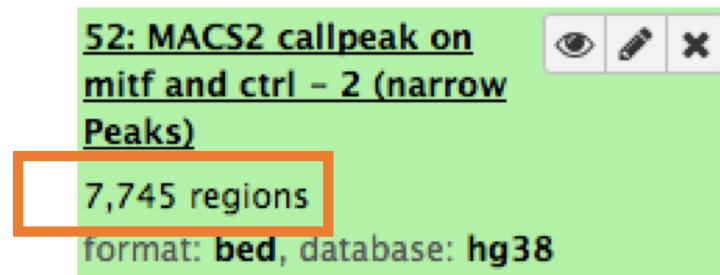
- 3. Look at the HTML dataset

```
#2 finished!
#2 predicted fragment length is 75 bps
#2 alternative fragment length(s) may be 75 bps
#2.2 Generate R script for model : MACS2_model.r
```

- The d value estimated by MACS seems a bit small. Let's try to re-run MACS with the expected fragment size : 200

# Exercise 5: peak calling

- 1.
  - Click on the name of one of the datasets generated by Macs2.
  - Click on  to display Macs2 form with the same parameters as for the previous run of Macs2
  - In Build Model, select Do not build the shifting model (--nomodel)
  - Enter 100 in the text box “The arbitrary extension size in bp”
  - Click on 
- 2.
  - 7,745 peaks are now found

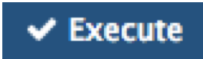


- NOTE: the graphs (showing the d values estimate) are no longer generated

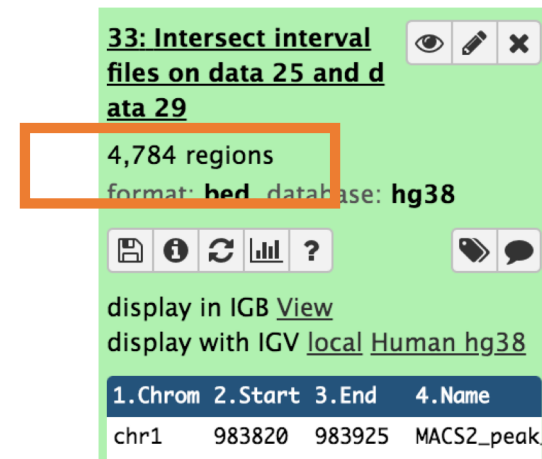


# Exercise 6: compare the two runs of MACS

1.

- Search for “**Intersect**” in the search field (tool panel)
- Click on the name of the tool **Intersect interval files** of the section **NGS: BEDtools**
- Set parameters:
  - **BED/VCF/GFF/BAM file:** MACS2 callpeak on data 1 and data 2 (narrow Peaks) (GalaxEast – 1st run of MACS)
  - **One or more BAM/BED/GFF/VCF file(s):** MACS2 callpeak on data 1 and data 2 (narrow Peaks) (GalaxEast – 2nd run of MACS)
  - **Report only those alignments that **\*\*do not\*\*** overlap the BED file:** Yes
- Click on 

**4,784 regions are found**



33: **Intersect interval files on data 25 and data 29**


4,784 regions  
format: **bed** database: **hg38**

display in IGB [View](#)  
display with IGV [local](#) [Human](#) [hg38](#)

1. Chrom	2. Start	3. End	4. Name
chr1	983820	983925	MACS2_peak

# Exercise 6: compare the two runs of MACS

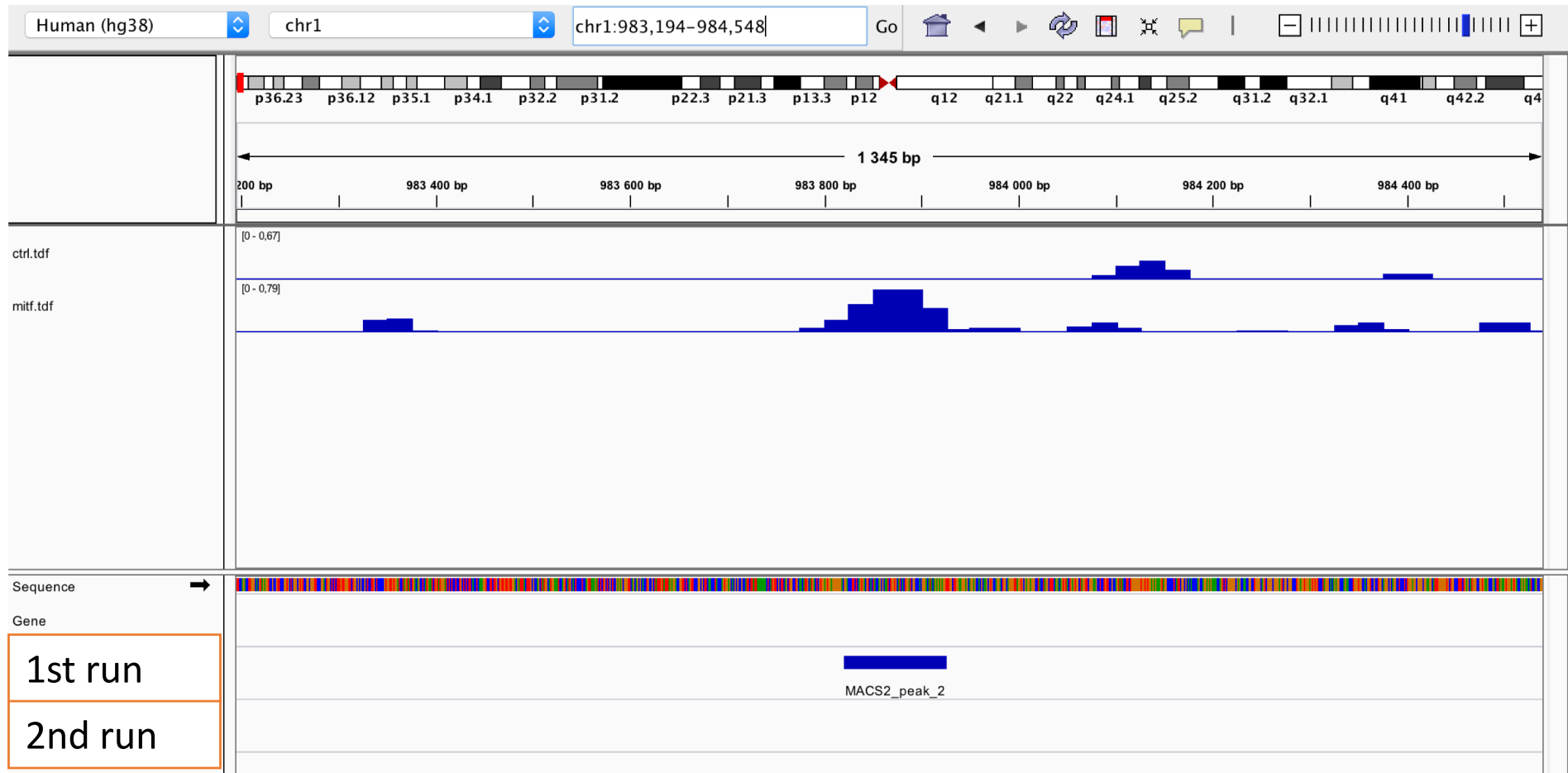
2.

1. In Galaxy, click on  for the two datasets named « MACS2 callpeak on data 1 and data 2 (narrow Peaks) » and save the files onto your computer
2. Go to IGV and load the two files along with the two tdf files already loaded (mitf.tdf and ctrl.tdf)

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSize
chr1	983820	983925	MACS2_peak_2	53	.	6.77148	9.11038	5.34984	56	
chr1	1586290	1586365	MACS2_peak_12	13	.	4.11467	4.42147	1.39180	6	
chr1	1728644	1728729	MACS2_peak_13	13	.	4.23390	4.76451	1.39180	66	
chr1	1807104	1807179	MACS2_peak_14	42	.	5.57865	7.91204	4.23630	32	
chr1	1909323	1909398	MACS2_peak_15	33	.	5.24205	6.88492	3.31573	31	
chr1	2167152	2167227	MACS2_peak_22	38	.	5.45624	7.50071	3.89401	49	
chr1	3276552	3276627	MACS2_peak_24	13	.	4.23390	4.76451	1.39180	52	
chr1	3444380	3444455	MACS2_peak_25	13	.	3.43937	4.35314	1.39180	40	
chr1	5680173	5680248	MACS2_peak_28	13	.	3.52851	4.64567	1.39180	37	

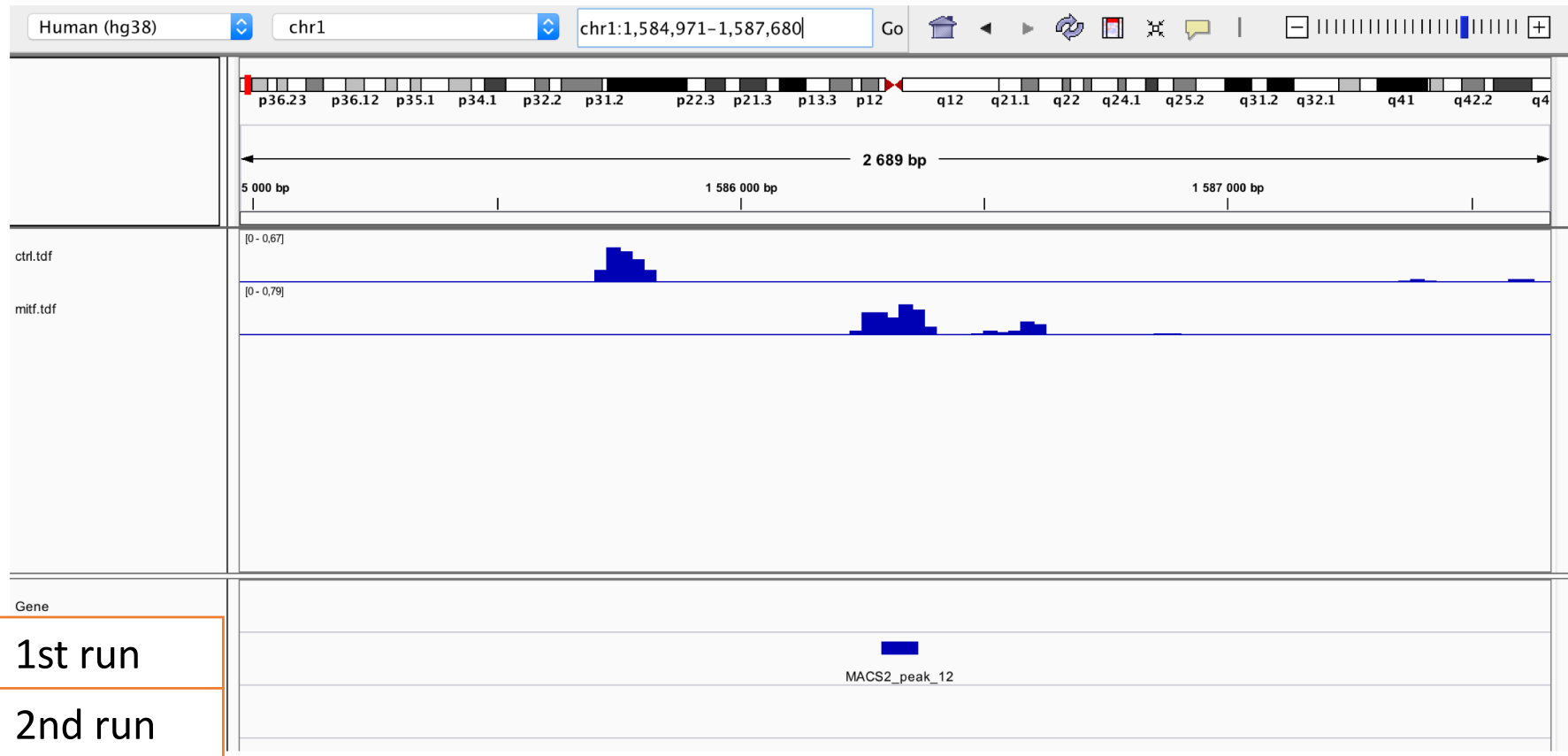
# Exercise 6: compare the two MACS runs

chr1:983820-983925



# Exercise 6: compare the two runs of MACS

chr1:1586290-1586365






# Exercise 6: compare the two runs of MACS

SSU72 (chr1:1556527-1578211)

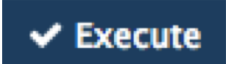
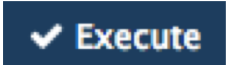


We are going to keep the second run of MACS

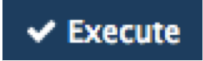

# Exercise 7: peak annotation

- 1.
  - Search for “homer annot” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Homer peaks OR BED format: MITF peaks - narrow peaks dataset (2<sup>nd</sup> run of Macs2)
    - Genome version: hg38
  - Click on 
- 2.
  - The Homer annotatePeaks tool generates two datasets: a log file and a tabular file containing annotated peaks.
  - Click on the  of the dataset which contain annotated peaks.
  - Click on the Datatype tab
  - Select **tabu**  the drop down list “New Type:”
  - Click on

# Exercise 7: peak annotation


- 3.
  - Search for “histogra” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Dataset: tabular file which contains annotated peaks
    - Numerical column for x axis: column: 10
    - Plot title: Frequency of peaks relative to TSS
    - Label for x axis: Distance to TSS
  - Click on 
- 4.a.
  - Search for “Cut” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Cut columns: c8
    - Delimited by: Tab
    - From: tabular file which contains annotated peaks
  - Click on 

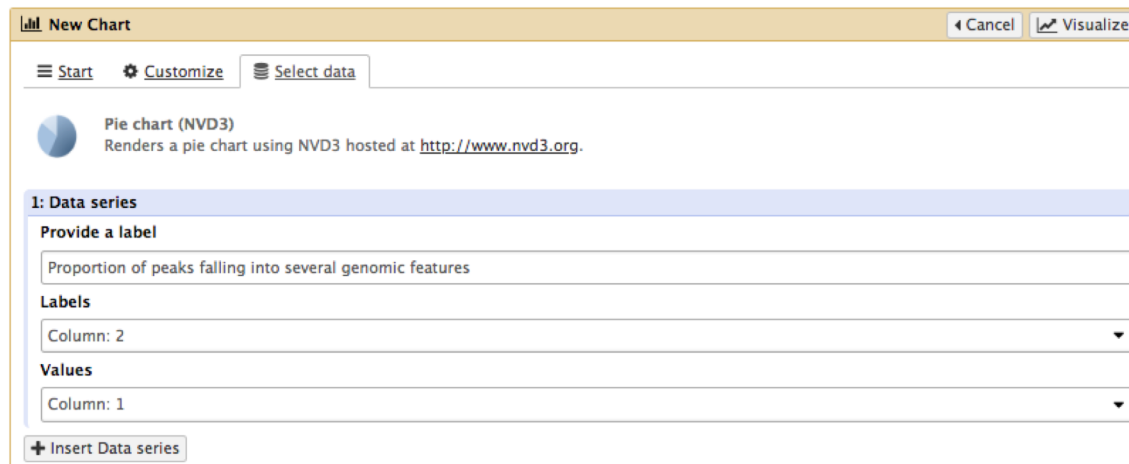
# Exercise 7: peak annotation

- 4.b.
  - Search for “Remove” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Remove first: 1
    - From: resulting dataset after 4.b
  - Click on 
- 4.c.
  - Search for “Count” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - from dataset: resulting dataset after 4.c
    - Count occurrences of values in column(s): column: 1
    - Delimited by: Whitespaces
    - How should the results be sorted?: With the most common values first
  - Click on 



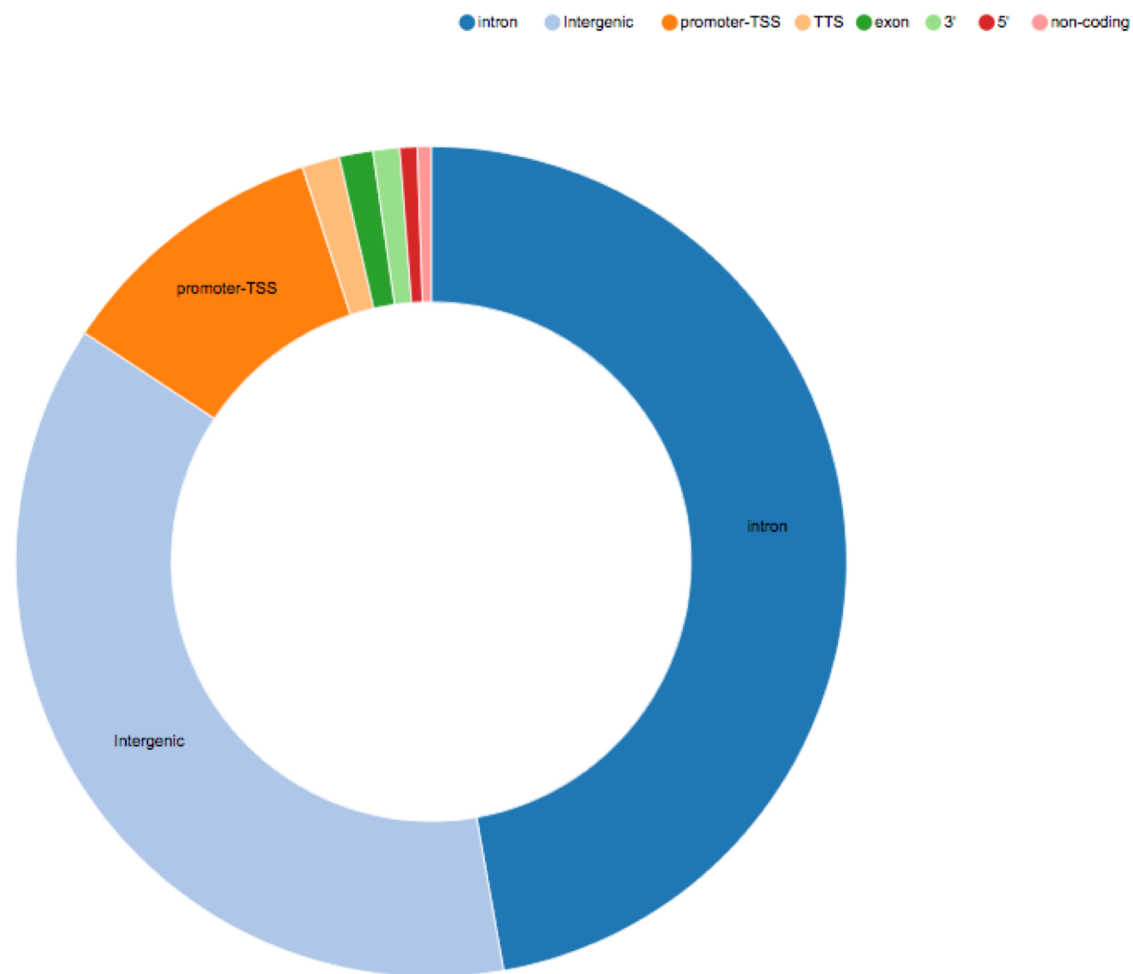
# Exercise 7: peak annotation

- 4.d.
  - Expand the box of the dataset generated in 4.d, click on  and select Charts
  - Double click on Pie charts
  - Click on editor (top right)
  - Go to the Select data tab:
    - Provide a label: Proportion of peaks falling into several genomic features.
    - Labels: Column: 2
    - Values: Column: 1



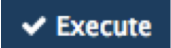
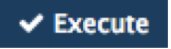
- Click on Visualize

# Exercise 7: peak annotation

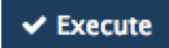


0: Proportion of peaks falling into several genomic features.

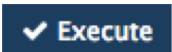

# Exercise 8: *de novo* motif discovery

- 1.a
  - Search for “Sort” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Sort Dataset: dataset with peak summits
    - on column: Column: 5
    - with flavor: Numerical sort
    - everything in: Descending order
  - Click on 
- 1.b
  - Search for “select first” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Select first: 800
    - From: dataset generated in 1.a
  - Click on 

# Exercise 8: *de novo* motif discovery

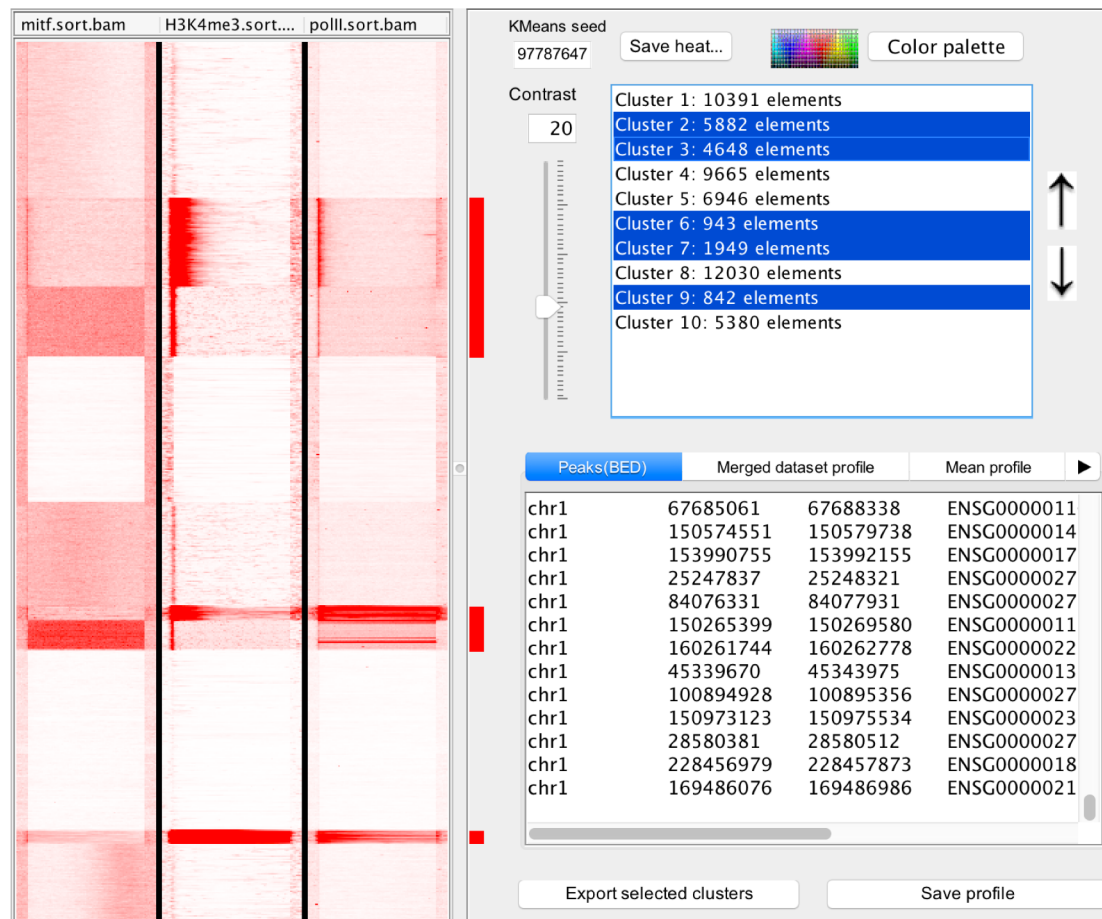
- 2.a
  - Import the file which contains chromosome lengths
  - Click on Shared Data (top menu) and select “Data Libraries”
  - Click on “Chromosome length”
  - Select the dataset named hg38.len (tick boxes beside dataset names)
  - Click on the button “To history”
  - Select history: ChIP-seq data analysis
  - Click on “Import”
  - Go back to the main page by clicking on “Analyzed data” (top menu)
- Run slopBed
  - BED/VCF/GFF file: MACS14\_in\_Galaxy\_summits.bed
  - Genome file: hg38.len
  - Choose what you want to do: Increase the BED/VCF/GFF entry by the same number of base pairs in each direction. (default)
  - Number of base pairs: 50
  - Click on 

## Exercise 8: *de novo* motif discovery

- 3.
  - Search for “extract” in the search field (tool panel)
  - Click on the name of the tool
  - Set parameters:
    - Fetch sequences for intervals in: the dataset generated in 2.c
    - Interpret features when possible: No
    - Click on 
- 4.
  - Expand the box of the dataset generated in 3 and click on  to download the file
- 5.
  - Go to MEME-chIP website and run the tool with the fasta file you’ve just downloaded and with default parameters.

# Exercise 9: Clustering

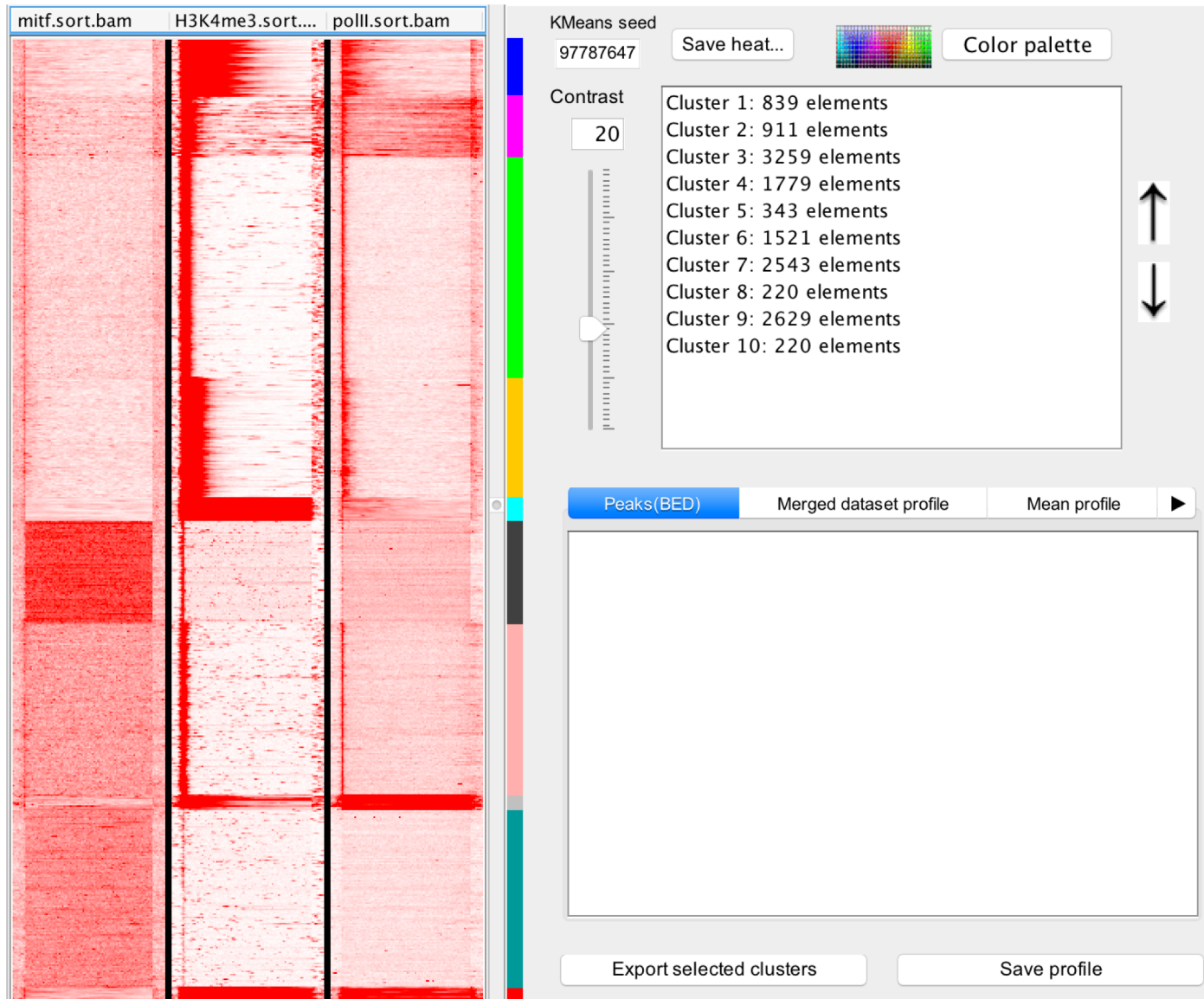
- 1.
  - Select clusters 2, 3, 6, 7, 9 and click on Export Selected clusters



# Exercise 9: Clustering

- 1.
  - Import the file previously exported as reference coordinates. You can use the one provided in `chipseq/seqminer/sub-clustering-gene.bed`. Click on browse, go to the directory which contains the file and click on open.
  - Click on Extract data
  - Click on Clustering

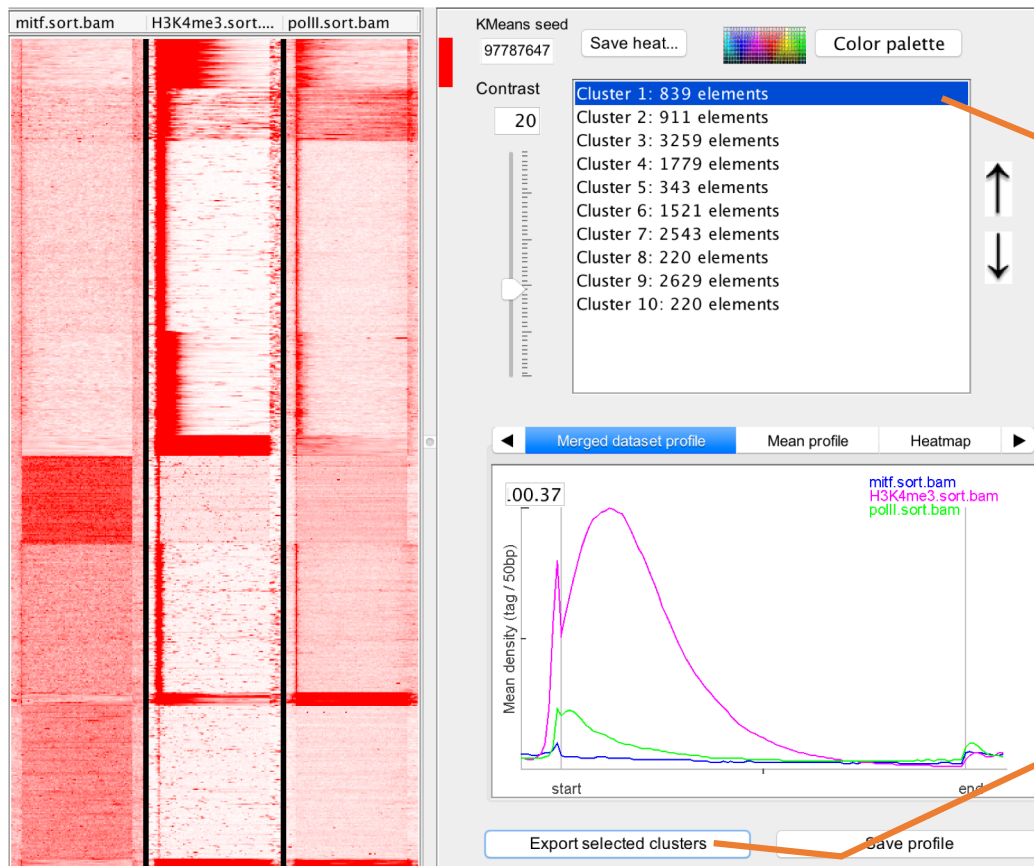
# Exercise 9: Clustering





# Exercise 9: Clustering

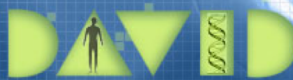
- 2.
  - Click on Cluster 1 (1)
  - Click on Export selected clusters (2)



# Exercise 9: Clustering

- Go to DAVID website <https://david.ncifcrf.gov/>
- Click on Function Annotation (left menu)
- Fill in the form:
  - Copy and paste Ensembl Gene IDs from the Cluster1.xls file in the Paste a list text field
  - Select Identifier (drop down list): ENSEMBL\_GENE\_ID
  - List Type: Gene List
  - Submit List
- Select: Continue to Submit IDs That DAVID Could Map
- Select to limit annotations by one or more species (left panel)
  - Select Homo sapiens (526)
  - Click on Select Species
- Click on Functional Annotation Clustering
- Keep all default
- Click on Functional Annotation Clustering

# Exercise 9: Clustering


**DAVID Bioinformatics Resources 6.7**  
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

\*\*\* You are currently using DAVID 6.7. \*\*\*  
\*\*\* If you are looking for [DAVID 6.8](#), please visit our [production site](#). \*\*\*

## Functional Annotation Clustering

[Help and Manual](#)

**Current Gene List:** List\_1  
**Current Background:** Homo sapiens  
**526 DAVID IDs**

**Options**    **Classification Stringency** Medium ▾

**132 Cluster(s)**  [Download File](#)

Annotation Cluster 1		Enrichment Score: 19.33			Count	P_Value	Benjamini
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">ribosomal protein</a>	RT		44	3.6E-28	1.3E-25
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">ribonucleoprotein</a>	RT		51	2.4E-27	4.2E-25
<input type="checkbox"/>	GOTERM_MF_FAT	<a href="#">structural constituent of ribosome</a>	RT		41	3.2E-27	1.6E-24
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">ribosome</a>	RT		45	3.4E-27	1.1E-24
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">ribonucleoprotein complex</a>	RT		66	3.6E-27	5.9E-25
<input type="checkbox"/>	KEGG_PATHWAY	<a href="#">Ribosome</a>	RT		32	3.2E-26	3.6E-24
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">ribosome</a>	RT		28	2.9E-24	3.3E-22
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">translational elongation</a>	RT		32	7.3E-24	1.2E-20
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">ribosomal subunit</a>	RT		32	9.9E-22	1.1E-19
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">translation</a>	RT		50	1.1E-21	9.2E-19
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">cytosolic ribosome</a>	RT		26	1.5E-20	1.2E-18
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">cytosolic part</a>	RT		31	2.3E-18	1.5E-16
<input type="checkbox"/>	SP_PIR_KEYWORDS	<a href="#">protein biosynthesis</a>	RT		34	3.9E-18	2.7E-16
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">cytosol</a>	RT		82	1.5E-13	7.3E-12
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">small ribosomal subunit</a>	RT		18	2.8E-13	1.2E-11
<input type="checkbox"/>	GOTERM_MF_FAT	<a href="#">structural molecule activity</a>	RT		50	1.1E-11	1.9E-9
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">cytosolic large ribosomal subunit</a>	RT		13	1.2E-10	3.1E-9