Analysis of ChIP-seq data (answers to questions)

Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1: mapping statistics

- 2.
    - Click on the button ⚙ an 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 📓 to History
    - Select history: ChIP-seq data analysis
    - Click on Import
    - 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 ✓ Execute

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 ✔ Execute
  - 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

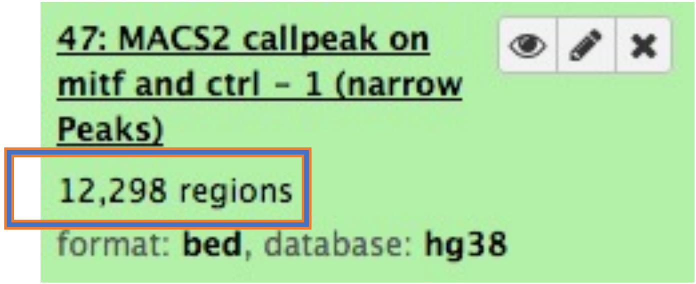
- 1.
  - Idh1 -> No peak
  - NPAS2 -> peak
  - AP1S2 -> Peak,
  - PABPC1l -> No peak
  - Park7 -> No peak
  - Pmel -> Peak
  - Cdk2 -> Peak
  - Actb -> No peak

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

Exercise 4: peak calling

- 2.
  - There is 12,298 peaks

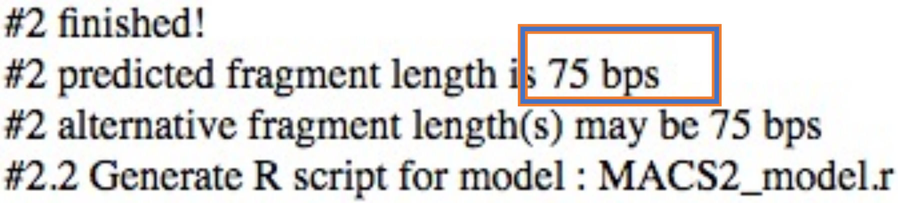


47: MACS2 callpeak on mitf and ctrl – 1 (narrow Peaks)

12,298 regions

format: **bed**, database: **hg38**

- 3. Look at the HTML dataset

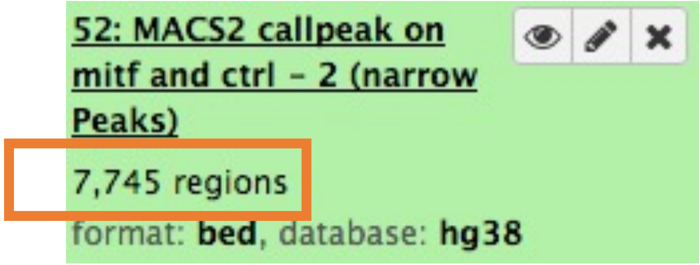


```
#2 finished!
#2 predicted fragment length is 75 bps
#2 alternative fragment length(s) may be 75 bps
#2.2 Generate R script for model : MACS2_model.r
```

  - 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 ✓ Execute
- 2.
  - 7,745 peaks are now found



52: MACS2 callpeak on mitf and ctrl - 2 (narrow Peaks)

7,745 regions

format: **bed**, database: **hg38**

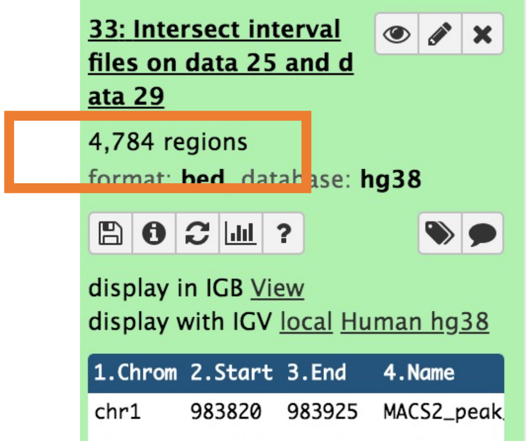
  - NOTE: the graphs (showing the d values estimate) are no longer generated

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 ✓ Execute

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

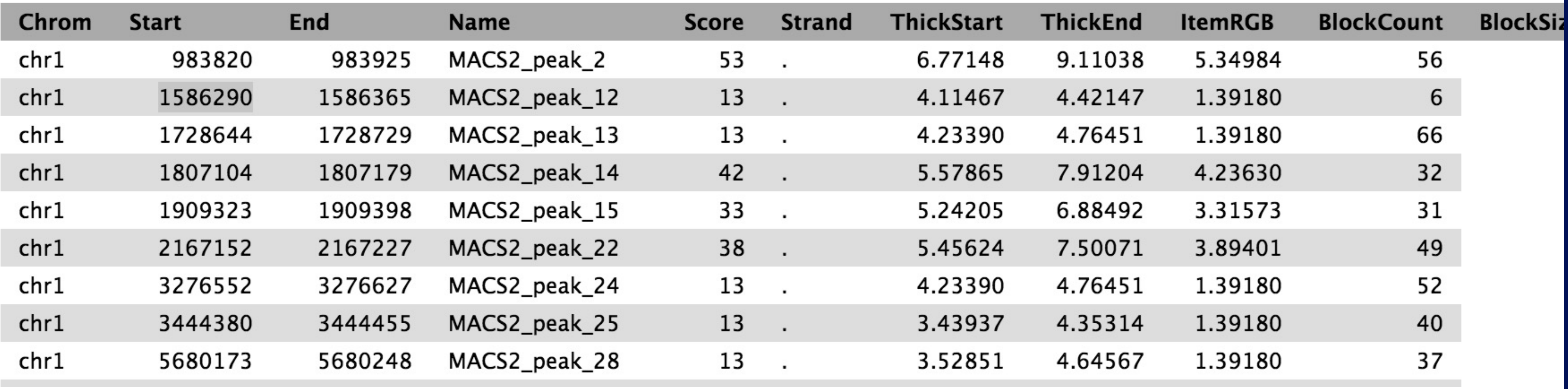
display with IGV local Human hg38

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

Exercise 6: compare the two runs of MACS

2.

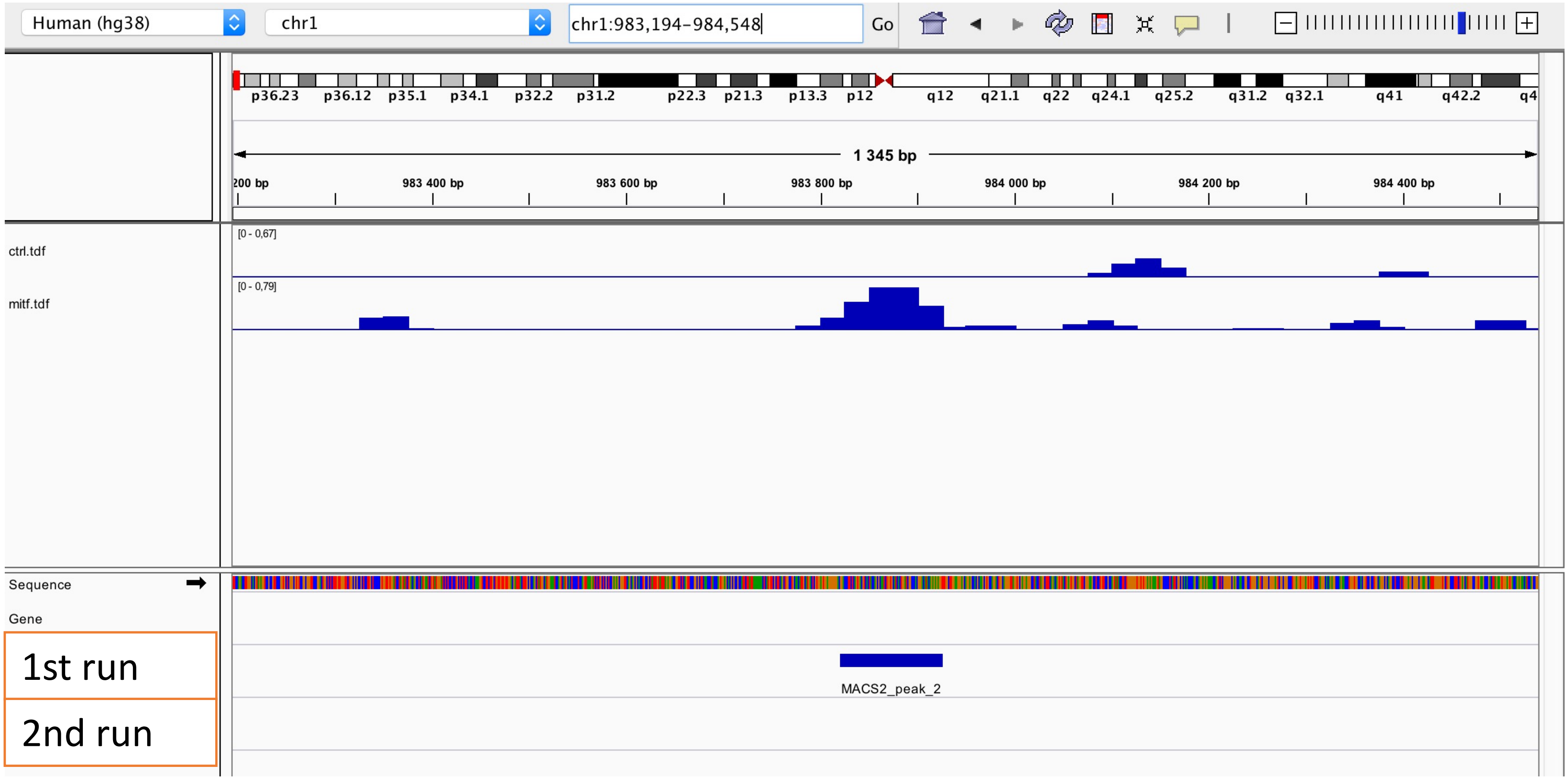
1. In Galaxy, click on 💾 for the two datasets named « MACS2 callpeak on data 1 and data 2 (narrow Peaks) » and save the files onto your computer
2. Go to IGV and load the two files along with the two tdf files already loaded (mitf.tdf and ctrl.tdf)



Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount	BlockSiz
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

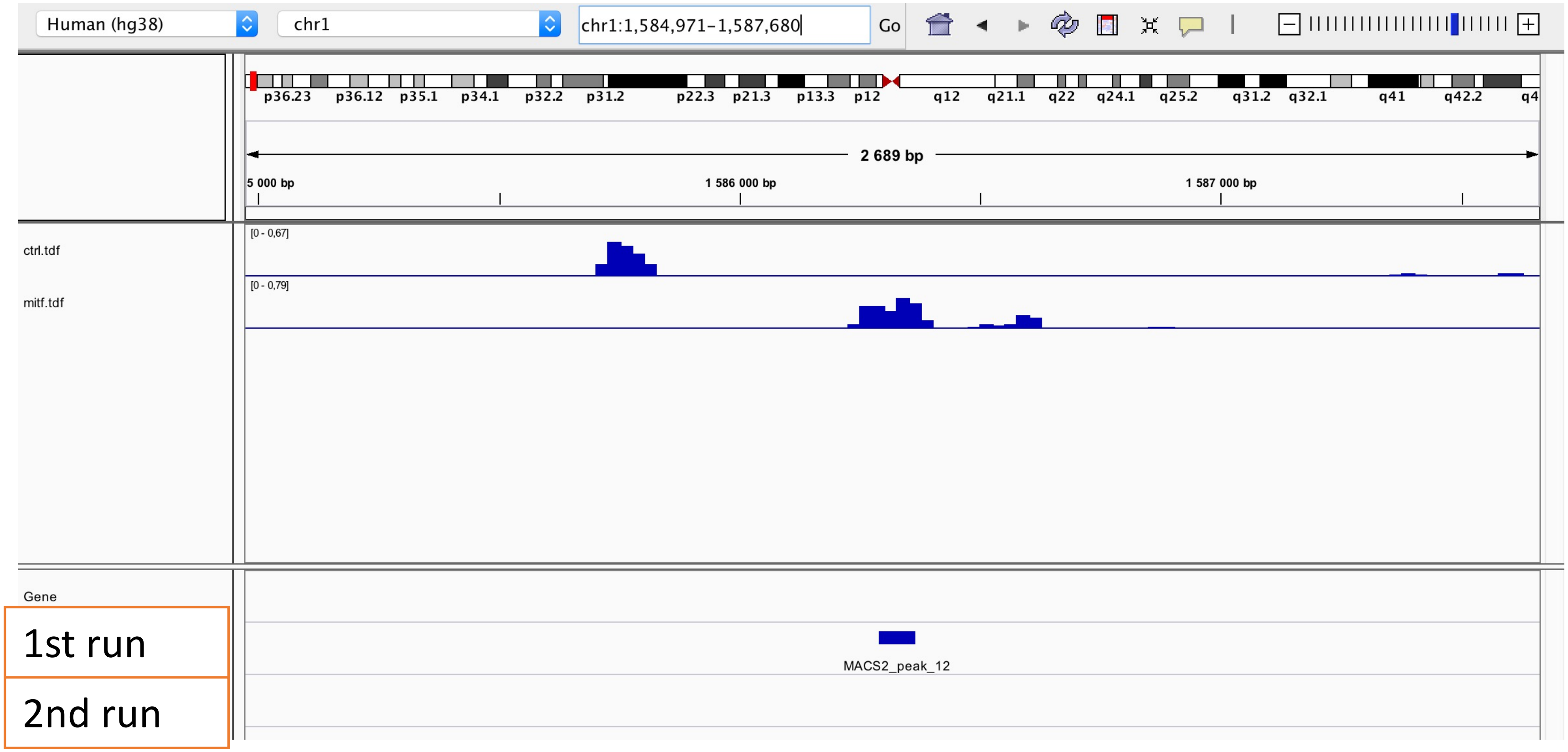


1st run

2nd run

Exercise 6: compare the two runs of MACS

chr1:1586290-1586365



Human (hg38)

chr1

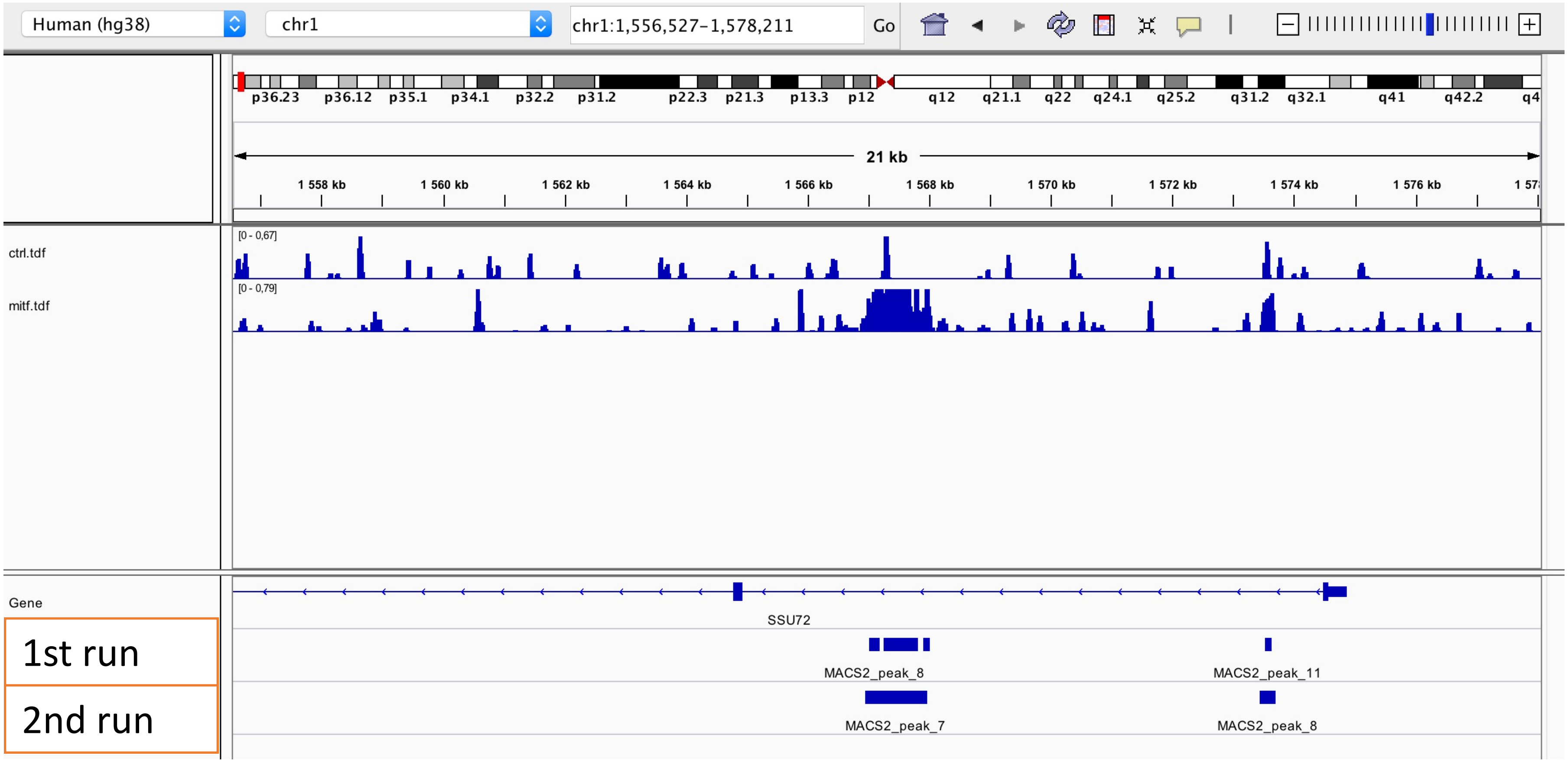
chr1:1,584,971-1,587,680

1st run

2nd run

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 (2nd run of Macs2)
    - Genome version: hg38
  - Click on ✔ Execute
- 2.
  - The Homer annotatePeaks tool generates two datasets: a log file and a tabular file cont✎ning annotated peaks.
  - Click on the of the dataset which contain annotated peaks.
  - Click on the Datatype tab
  - Select **tabul** Save he 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

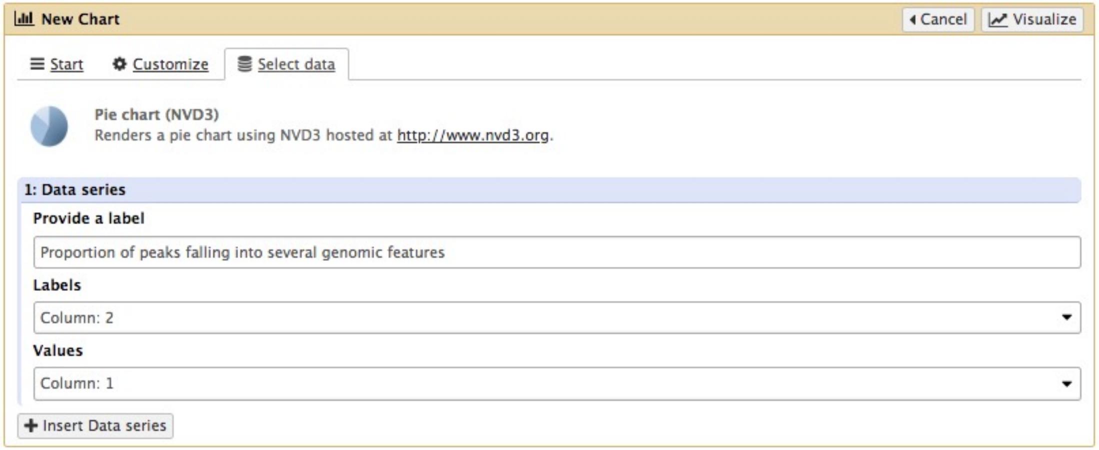
Exercise 7: peak annotation

- 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



Pie chart (NVD3)
Renders a pie chart using NVD3 hosted at http://www.nvd3.org.

1: Data series

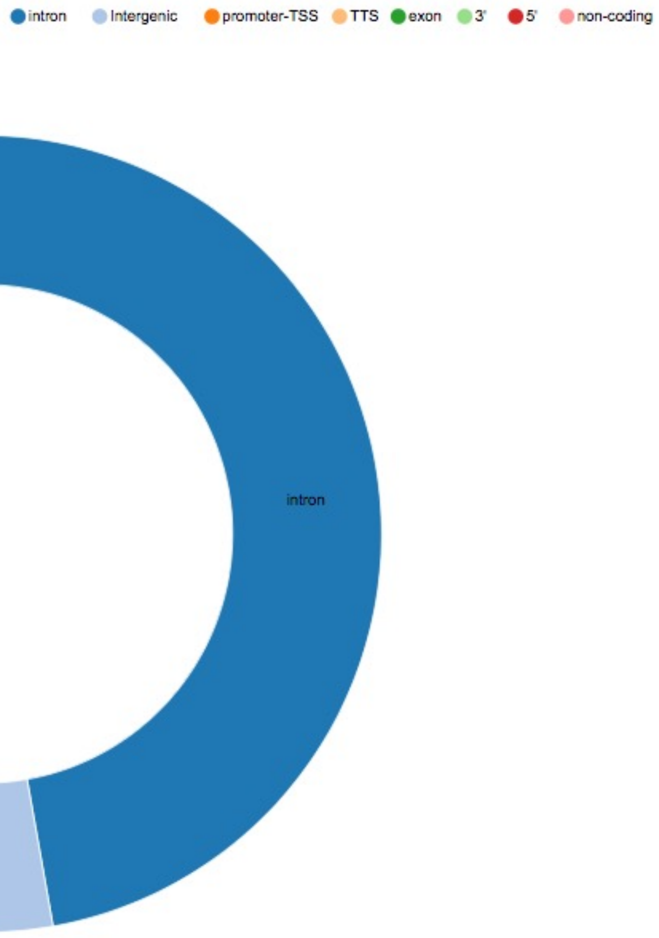
Provide a label: Proportion of peaks falling into several genomic features

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

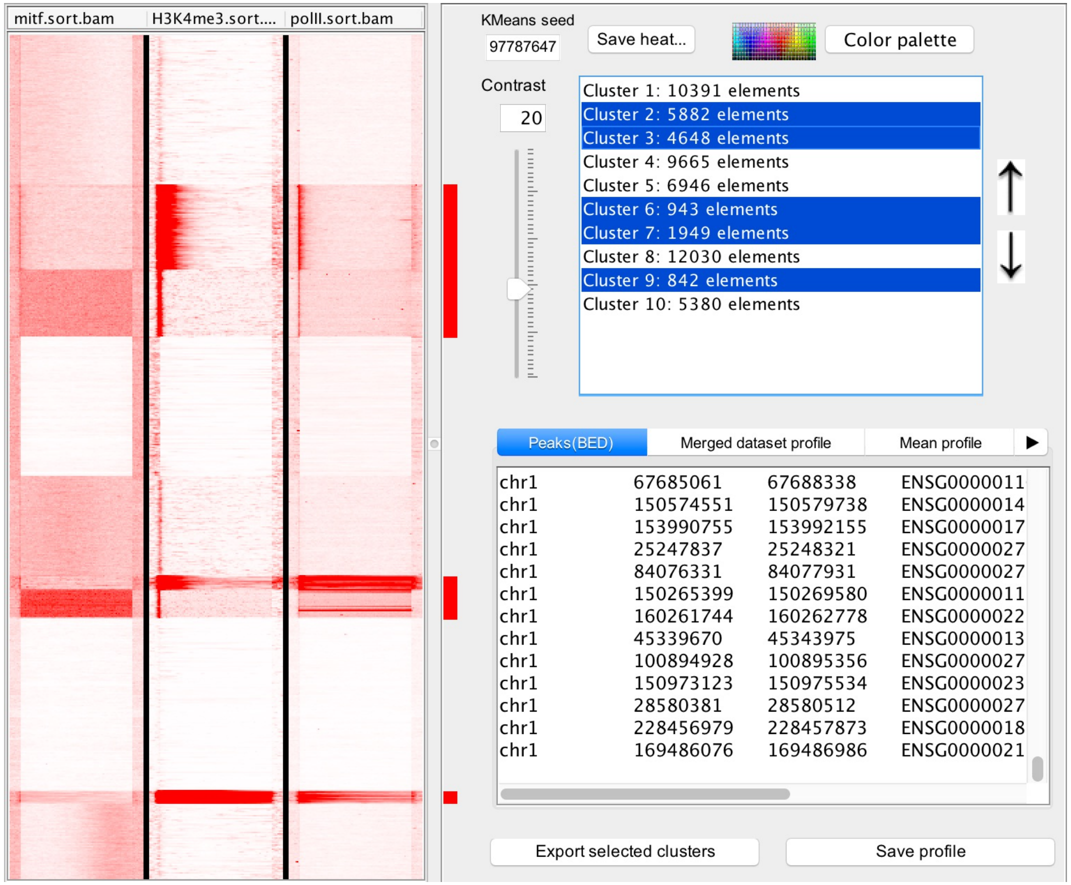
- 2.a
  - Import the file which contains chromosome lengths
  - Click on Shared Data (top menu) and select “Data Libraries”
  - Click on “Chromosome length”
  - Select the dataset named hg38.len (tick boxes beside dataset names)
  - Click on the button “To history”
  - Select history: ChIP-seq data analysis
  - 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 ✓ Execute
- 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



KMeans seed 97787647

Contrast 20

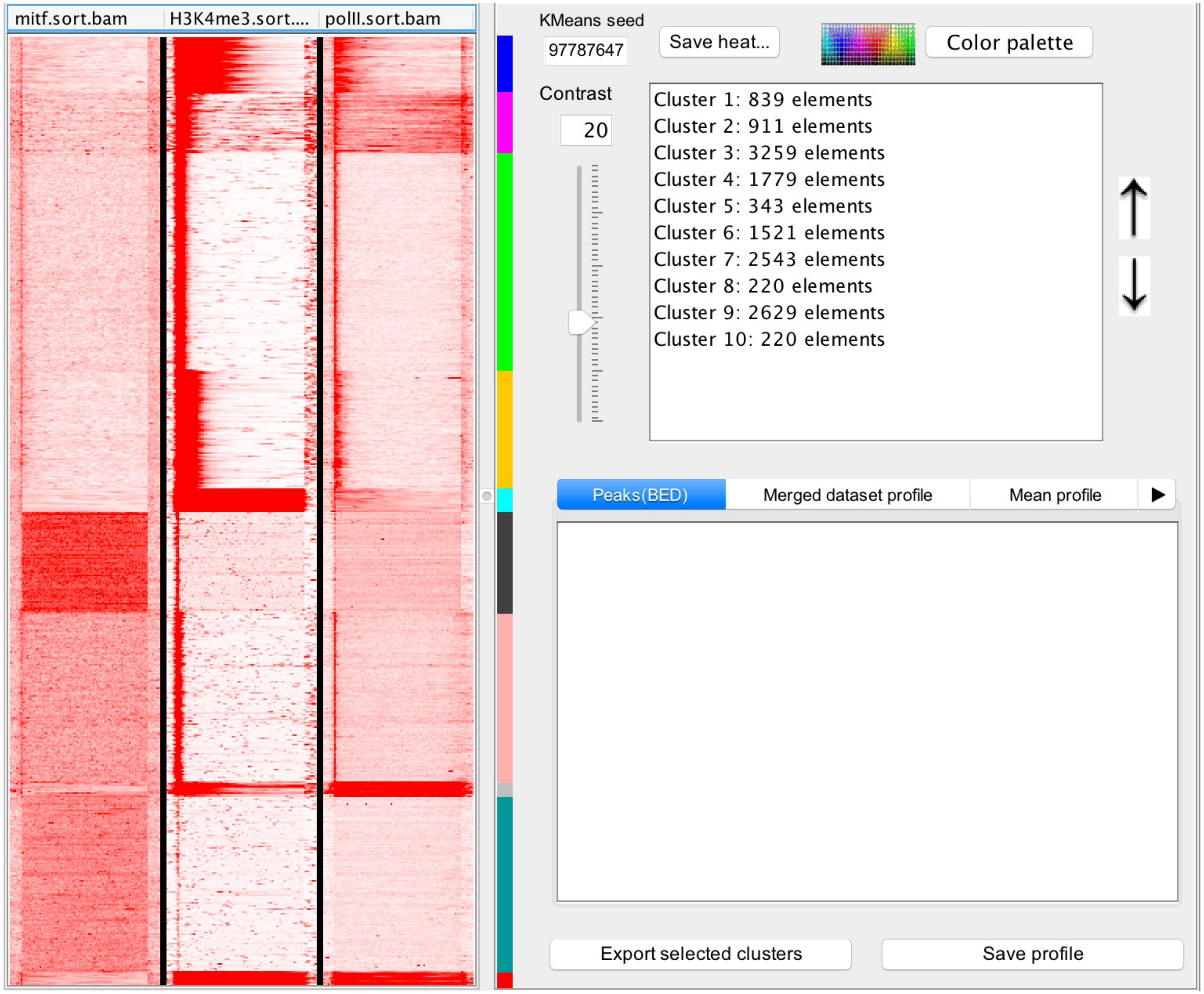
Cluster 1: 10391 elements
Cluster 2: 5882 elements
Cluster 3: 4648 elements
Cluster 4: 9665 elements
Cluster 5: 6946 elements
Cluster 6: 943 elements
Cluster 7: 1949 elements
Cluster 8: 12030 elements
Cluster 9: 842 elements
Cluster 10: 5380 elements

```
chr1        67685061     67688338     ENSG0000011
chr1        150574551    150579738    ENSG0000014
chr1        153990755    153992155    ENSG0000017
chr1        25247837     25248321     ENSG0000027
chr1        84076331     84077931     ENSG0000027
chr1        150265399    150269580    ENSG0000011
chr1        160261744    160262778    ENSG0000022
chr1        45339670     45343975     ENSG0000013
chr1        100894928    100895356    ENSG0000027
chr1        150973123    150975534    ENSG0000023
chr1        28580381     28580512     ENSG0000027
chr1        228456979    228457873    ENSG0000018
chr1        169486076    169486986    ENSG0000021
```

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



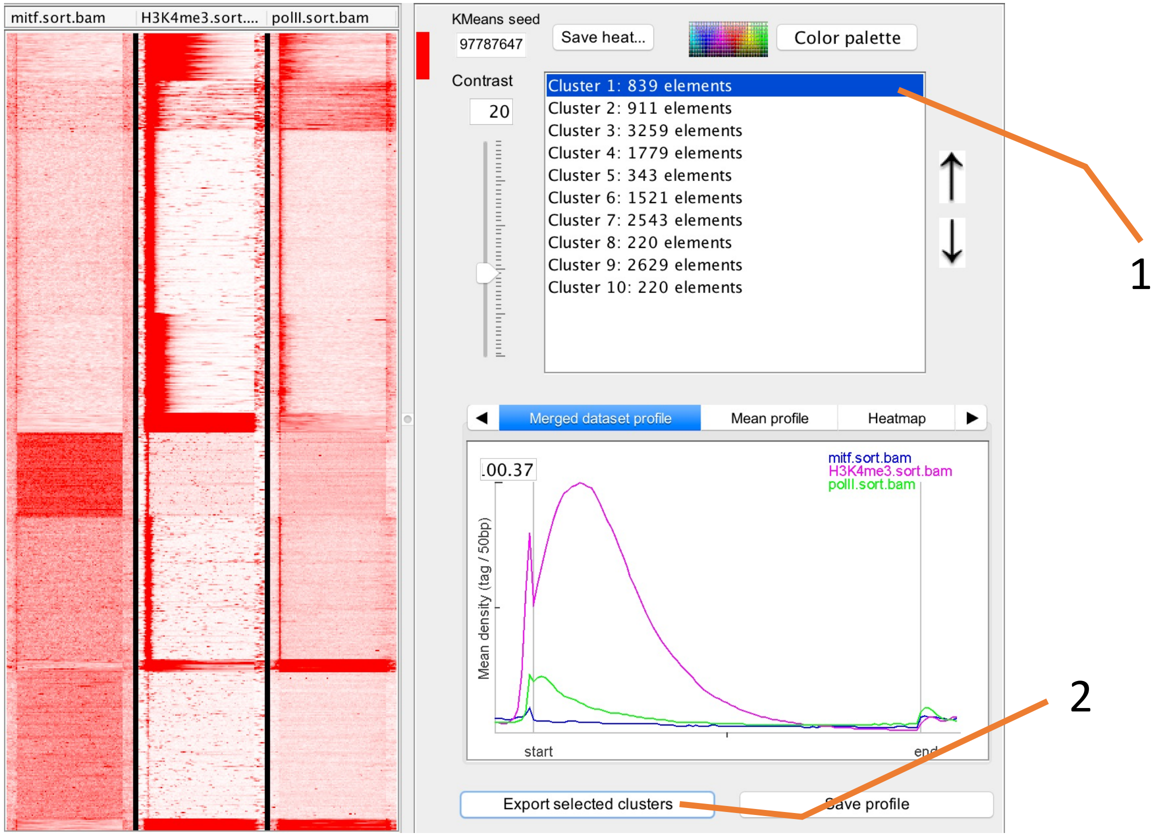
KMeans seed 97787647

Contrast 20

Cluster 1: 839 elements
Cluster 2: 911 elements
Cluster 3: 3259 elements
Cluster 4: 1779 elements
Cluster 5: 343 elements
Cluster 6: 1521 elements
Cluster 7: 2543 elements
Cluster 8: 220 elements
Cluster 9: 2629 elements
Cluster 10: 220 elements

Exercise 9: Clustering

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



KMeans seed 97787647

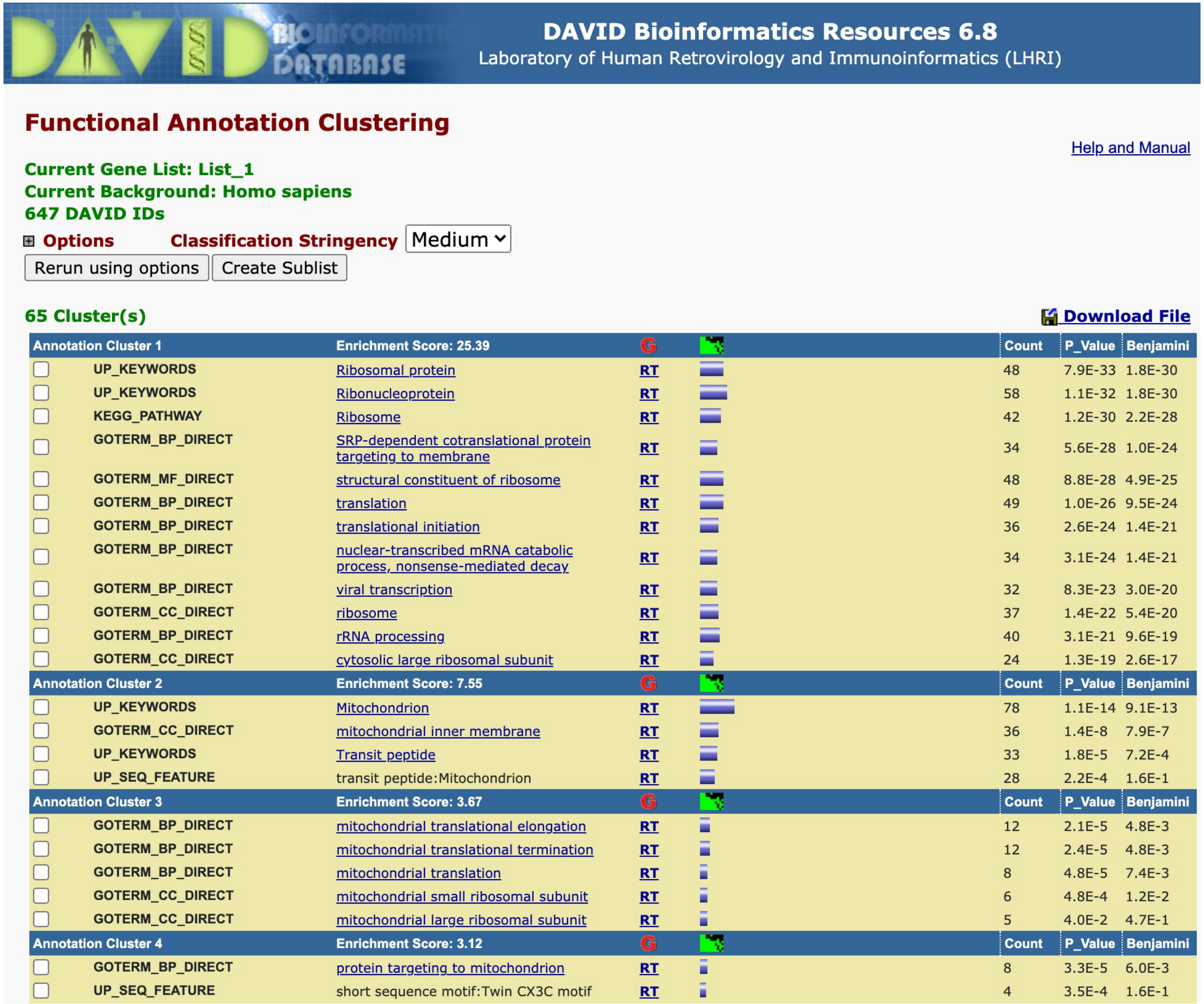
Contrast 20

Cluster 1: 839 elements
Cluster 2: 911 elements
Cluster 3: 3259 elements
Cluster 4: 1779 elements
Cluster 5: 343 elements
Cluster 6: 1521 elements
Cluster 7: 2543 elements
Cluster 8: 220 elements
Cluster 9: 2629 elements
Cluster 10: 220 elements

Exercise 9: Clustering

- Go to DAVID website https://david.ncifcrf.gov/
- Click on Shortcut to DAVID Tools (top menu)/Function Annotation
- 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 (647)
    - Click on Select Species
- Click on Functional Annotation Tool
- Keep all default
- Click on Functional Annotation Clustering

Exercise 9: Clustering



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
647 DAVID IDs

⊞ **Options** **Classification Stringency** Medium

Rerun using options | Create Sublist

65 Cluster(s)

Download File

					Count	P_Value	Benjamini
Annotation Cluster 1		**Enrichment Score: 25.39**	G		**Count**	**P_Value**	**Benjamini**
□	UP_KEYWORDS	Ribosomal protein	RT		48	7.9E-33	1.8E-30
□	UP_KEYWORDS	Ribonucleoprotein	RT		58	1.1E-32	1.8E-30
□	KEGG_PATHWAY	Ribosome	RT		42	1.2E-30	2.2E-28
□	GOTERM_BP_DIRECT	SRP-dependent cotranslational protein targeting to membrane	RT		34	5.6E-28	1.0E-24
□	GOTERM_MF_DIRECT	structural constituent of ribosome	RT		48	8.8E-28	4.9E-25
□	GOTERM_BP_DIRECT	translation	RT		49	1.0E-26	9.5E-24
□	GOTERM_BP_DIRECT	translational initiation	RT		36	2.6E-24	1.4E-21
□	GOTERM_BP_DIRECT	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	RT		34	3.1E-24	1.4E-21
□	GOTERM_BP_DIRECT	viral transcription	RT		32	8.3E-23	3.0E-20
□	GOTERM_CC_DIRECT	ribosome	RT		37	1.4E-22	5.4E-20
□	GOTERM_BP_DIRECT	rRNA processing	RT		40	3.1E-21	9.6E-19
□	GOTERM_CC_DIRECT	cytosolic large ribosomal subunit	RT		24	1.3E-19	2.6E-17
Annotation Cluster 2		**Enrichment Score: 7.55**	G		**Count**	**P_Value**	**Benjamini**
□	UP_KEYWORDS	Mitochondrion	RT		78	1.1E-14	9.1E-13
□	GOTERM_CC_DIRECT	mitochondrial inner membrane	RT		36	1.4E-8	7.9E-7
□	UP_KEYWORDS	Transit peptide	RT		33	1.8E-5	7.2E-4
□	UP_SEQ_FEATURE	transit peptide:Mitochondrion	RT		28	2.2E-4	1.6E-1
Annotation Cluster 3		**Enrichment Score: 3.67**	G		**Count**	**P_Value**	**Benjamini**
□	GOTERM_BP_DIRECT	mitochondrial translational elongation	RT		12	2.1E-5	4.8E-3
□	GOTERM_BP_DIRECT	mitochondrial translational termination	RT		12	2.4E-5	4.8E-3
□	GOTERM_BP_DIRECT	mitochondrial translation	RT		8	4.8E-5	7.4E-3
□	GOTERM_CC_DIRECT	mitochondrial small ribosomal subunit	RT		6	4.8E-4	1.2E-2
□	GOTERM_CC_DIRECT	mitochondrial large ribosomal subunit	RT		5	4.0E-2	4.7E-1
Annotation Cluster 4		**Enrichment Score: 3.12**	G		**Count**	**P_Value**	**Benjamini**
□	GOTERM_BP_DIRECT	protein targeting to mitochondrion	RT		8	3.3E-5	6.0E-3
□	UP_SEQ_FEATURE	short sequence motif:Twin CX3C motif	RT		4	3.5E-4	1.6E-1