Correlation of RNA-seq and ChIP-seq data

Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1

We want to know how many up-regulated genes contain a peak for MITF. Compare **Gene names** of the chIPseq data (annotation step) and the RNAseq data (up-regulated genes).

- Tool: use Venny (https://bioinfogp.cnb.csic.es/tools/venny/)
- Datasets:
  - Use the file siMitfvssiLuc.up.annot.txt (upregulated genes detected by SARtools annotated with BioMart).
    - Download it from your history « RNA-seq data analysis ».
    - Open it with Excel and copy paste the content of column « Gene name » in Venny. You can name this part of the diagramme « RNA-seq »
  - Use the file mitf_peaks.annot.tsv (all chIPseq peaks detected in the second run of MACS2 and annotated with BEDtools closest)
    - Download it from your history « ChIP-seq data analysis »
    - Open it with Excel and copy paste the content of column « O » (the one with Gene names) in Venny. You can name this part of the diagramme « ChIP-seq »

Exercise 2

Use seqMINER to visualize at the same time chIP-seq data along with RNA-seq data.

Tool: seqMINER

Datasets:

- Reference coordinated: MITF_peak_summits.bed (from your history "ChIP-seq data analysis")
- Density
- seqMINER accept a tab separated file formatted like:
  - Gene ID <tab> Expression Values
  - Aligned reads (available in chipseq/mapping):
    - Mitf.sort.bam
    - H3K4me3.sort.bam
    - polII.sort.bam
  - RNA-seq data:
    - seqMINER accept a tab separated file formatted like: Gene ID <tab> Expression Values
      - Expression values used in our example are normalized read counts divided by gene length in Kb.
    - **Let's generate this file!**

Prepare exercise 2

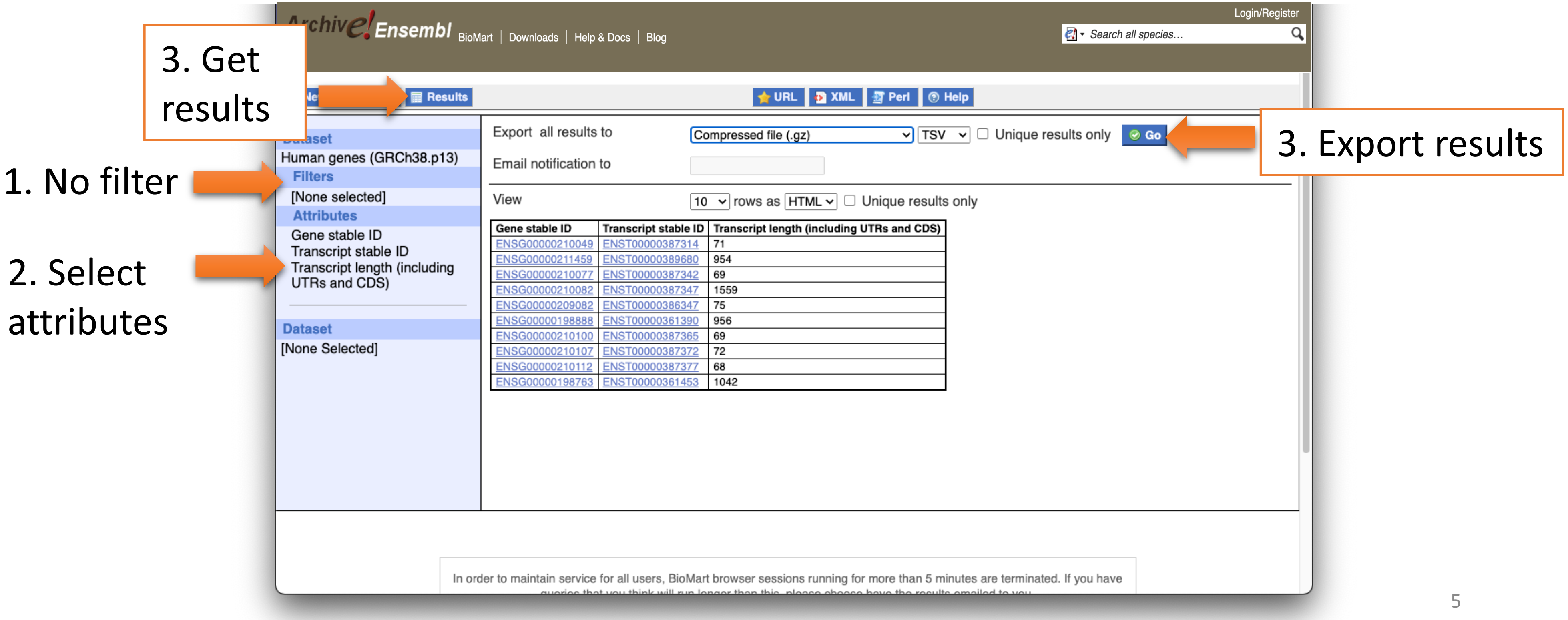
In seqMINER, we are going to visualize and compare expression values of different genes within the same sample. Read count per gene can not be directly compared together as they are correlated to the gene length. In the file siMitfvssiLuc.complete.txt, you can find normalized values. We are going to scale them as if genes were all of the same size.

Steps:

1. Extract transcript lengths [ensembl/BioMart]
2. As we are working on read counts per gene and not per transcript, compute a median of transcript lengths per gene [Galaxy]
3. Add the median of transcript lengths to the dataset siMitfvssiLuc.complete.txt [Galaxy]
4. Compute normalized and divided by median of transcripts length in kb values per gene on normalized data for siLuc (rounded mean of normalized counts) [Excel]

Prepare exercise 2

1. In Ensembl/BioMart
   1. For all genes, extract the following information:
      - Gene stable IDs
      - Transcript stable IDs
      - Transcript length (including UTRs and CDS)
   2. Rename the file: hg38_ens105_transcriptLength.txt.gz (compressed .gz)



Dataset
Human genes (GRCh38.p13)
Filters
[None selected]
Attributes
Gene stable ID
Transcript stable ID
Transcript length (including UTRs and CDS)

Dataset
[None Selected]

Export all results to Compressed file (.gz) TSV □ Unique results only Go

Email notification to

View 10 rows as HTML □ Unique results only

Gene stable ID	Transcript stable ID	Transcript length (including UTRs and CDS)
ENSG00000210049	ENST00000387314	71
ENSG00000211459	ENST00000389680	954
ENSG00000210077	ENST00000387342	69
ENSG00000210082	ENST00000387347	1559
ENSG00000209082	ENST00000386347	75
ENSG00000198888	ENST00000361390	956
ENSG00000210100	ENST00000387365	69
ENSG00000210107	ENST00000387372	72
ENSG00000210112	ENST00000387377	68
ENSG00000198763	ENST00000361453	1042

In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.

1. No filter

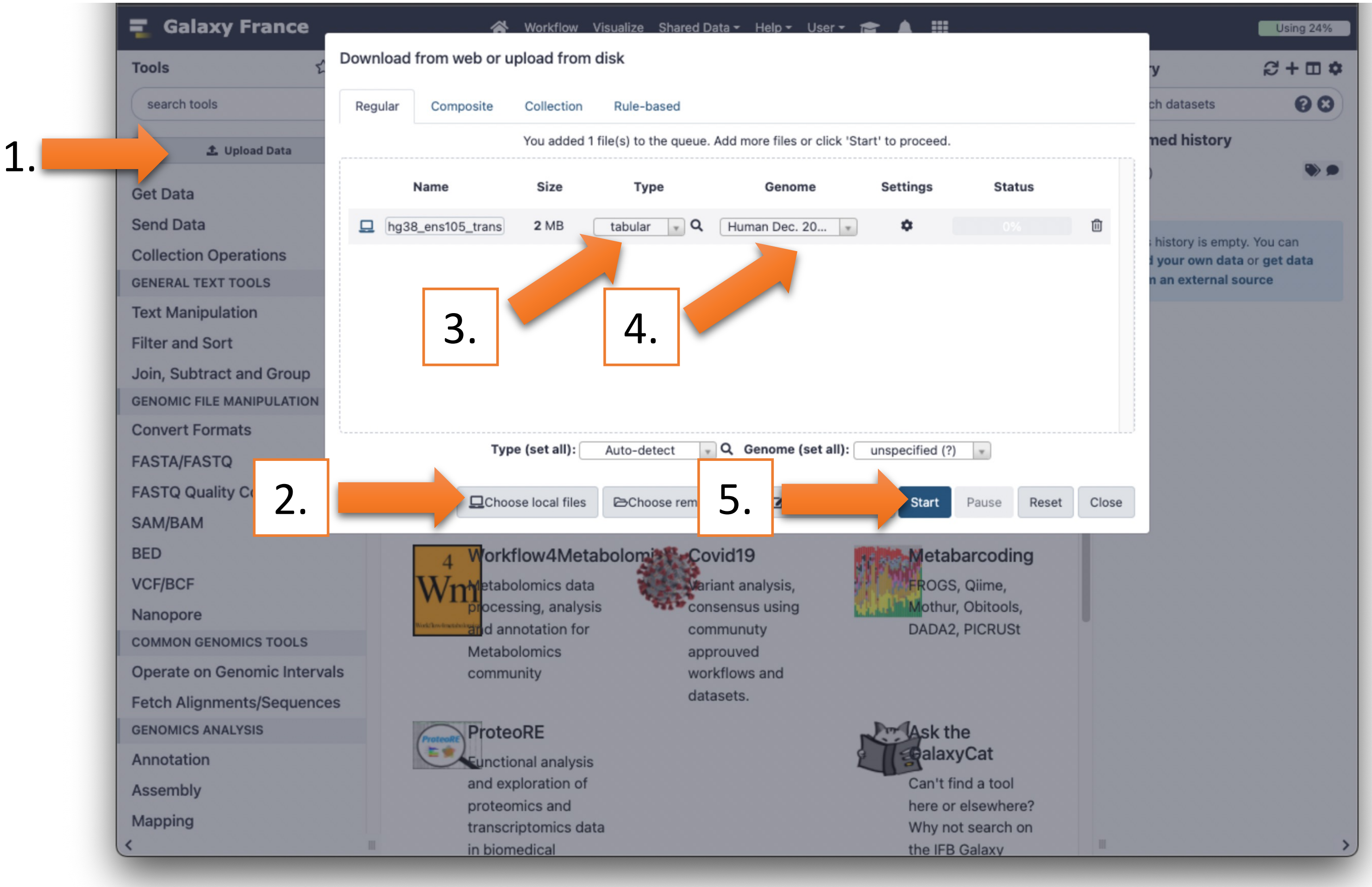
2. Select attributes

3. Get results

3. Export results

Prepare exercise 2

3. Create a new history « Prepare RNA-seq data for seqMINER ». Import the file hg38_ens105_transcriptLength.txt.gz to Galaxy (type: tabular (3), Genome: hg38 (4))



Download from web or upload from disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
hg38_ens105_trans	2 MB	tabular	Human Dec. 20...	⚙	0%

Type (set all): Auto-detect Genome (set all): unspecified (?)

history is empty. You can your own data or get data n an external source

Workflow4Metabolomics Metabolomics data processing, analysis and annotation for Metabolomics community

Covid19 Variant analysis, consensus using community approved workflows and datasets.

Metabarcoding FROGS, Qiime, Mothur, Obitools, DADA2, PICRUSt

ProteoRE Functional analysis and exploration of proteomics and transcriptomics data in biomedical

Ask the GalaxyCat Can't find a tool here or elsewhere? Why not search on the IFB Galaxy

Prepare exercise 2

2.1. Use the tool « **Datamash** (operations on tabular data) » to group gene by Ensembl Gene Ids and compute the median on transcript length:

- **Input tabular dataset:** hg38_ens105_transcriptLength.txt.gz
- **Group by fields:** [column with Gene stable ID] *(1)*
- **Input file has a header line:** Yes
- **Print header line:** Yes
- **Operation to perform on each group:**
  - **Type:** Median
  - **On column:** [column with Transcript length (including UTRs and CDS)] *(3)*

2.2. Use the tool « **Compute** » to round median values:

- **Add expression:** round(c2) *(change accordingly if needed)*
- **Input has a header line with column names?** Yes
  - **The new column name:** rounded median(Transcript length (including UTRs and CDS))

2.3. Use the tool « **Advanced Cut** » to extract only column of interest:

- **File to cut:** [Compute on data *] *(result of step 2.2)*
- **Cut by:** fields
  - **List of Fields:** Column: 1 Column: 3 *(change accordingly if needed)*

Rename resulting dataset: Median_of_transcript_length.tsv

Prepare exercise 2

3.

1. Import the file siMitfvssiLuc.complete.txt (history “RNA-seq data analysis”) to Galaxy (type: tabular, Genome: hg38).
2. Use the tool “**Join two Datasets**” to join the two datasets siMitfvssiLuc.complete.txt and Median_of_transcript_length.tsv
   - **Join:** siMitfvssiLuc.complete.txt
   - **Using column:** 1
   - **With:** Median_of_transcript_length.tsv
   - **and column**: 1
   - **Keep the header lines**: Yes
3. Rename the file siMitfvssiLuc.complete.wTranscriptLength.tsv
4. Download siMitfvssiLuc.complete.wTranscriptLength.tsv and open it in excel.

Prepare exercise 2

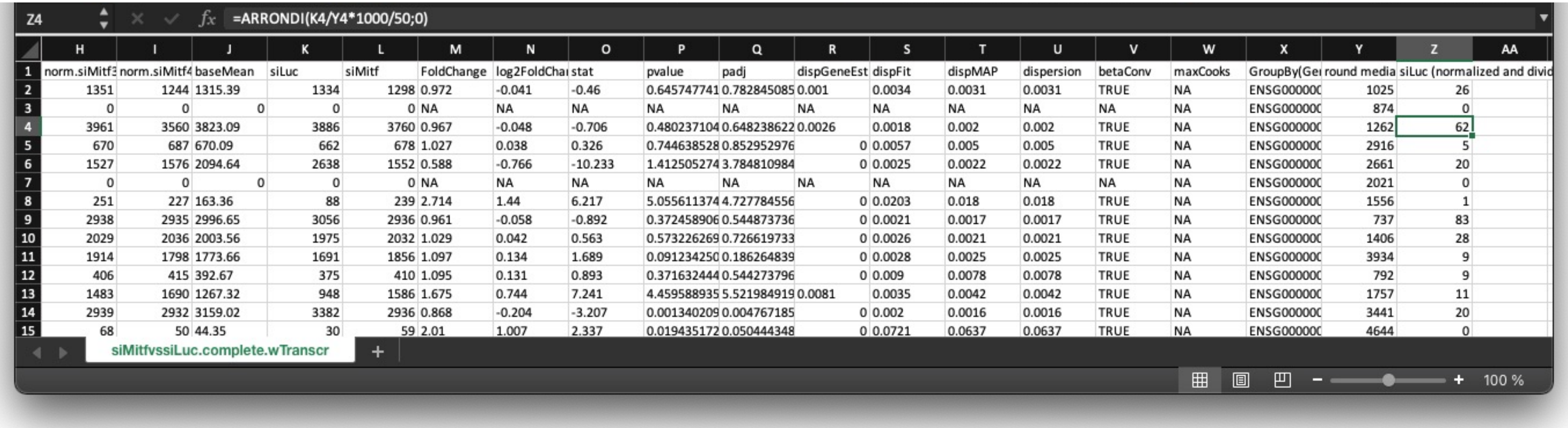
4. In Excel (or an equivalent), add 1 column:

   1. siLuc (normalized and divided by median of transcripts length in kb) filled with the following formula (French ; English)

      =ARRONDI(K2/Y2*1000/50;0) (K is the column : siLuc – Y is the column with round median)

      =ROUND(K2/Y2*1000/50;0)

Hint: here we divide values by 50 to get them in the range of the chIP-seq data that we are going to visualize along with them.



Z4 =ARRONDI(K4/Y4*1000/50;0)

	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	norm.siMitf3	norm.siMitf4	baseMean	siLuc	siMitf	FoldChange	log2FoldChai	stat	pvalue	padj	dispGeneEst	dispFit	dispMAP	dispersion	betaConv	maxCooks	GroupBy(Gei	round media	siLuc (normalized and divid	
2	1351	1244	1315.39	1334	1298	0.972	-0.041	-0.46	0.645747741	0.782845085	0.001	0.0034	0.0031	0.0031	TRUE	NA	ENSG00000C	1025	26	
3	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	ENSG00000C	874	0	
4	3961	3560	3823.09	3886	3760	0.967	-0.048	-0.706	0.480237104	0.648238622	0.0026	0.0018	0.002	0.002	TRUE	NA	ENSG00000C	1262	62	
5	670	687	670.09	662	678	1.027	0.038	0.326	0.744638528	0.852952976	0	0.0057	0.005	0.005	TRUE	NA	ENSG00000C	2916	5	
6	1527	1576	2094.64	2638	1552	0.588	-0.766	-10.233	1.412505274	3.784810984	0	0.0025	0.0022	0.0022	TRUE	NA	ENSG00000C	2661	20	
7	0	0	0	0	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	ENSG00000C	2021	0	
8	251	227	163.36	88	239	2.714	1.44	6.217	5.055611374	4.727784556	0	0.0203	0.018	0.018	TRUE	NA	ENSG00000C	1556	1	
9	2938	2935	2996.65	3056	2936	0.961	-0.058	-0.892	0.372458906	0.544873736	0	0.0021	0.0017	0.0017	TRUE	NA	ENSG00000C	737	83	
10	2029	2036	2003.56	1975	2032	1.029	0.042	0.563	0.573226269	0.726619733	0	0.0026	0.0021	0.0021	TRUE	NA	ENSG00000C	1406	28	
11	1914	1798	1773.66	1691	1856	1.097	0.134	1.689	0.091234250	0.186264839	0	0.0028	0.0025	0.0025	TRUE	NA	ENSG00000C	3934	9	
12	406	415	392.67	375	410	1.095	0.131	0.893	0.371632444	0.544273796	0	0.009	0.0078	0.0078	TRUE	NA	ENSG00000C	792	9	
13	1483	1690	1267.32	948	1586	1.675	0.744	7.241	4.459588935	5.521984919	0.0081	0.0035	0.0042	0.0042	TRUE	NA	ENSG00000C	1757	11	
14	2939	2932	3159.02	3382	2936	0.868	-0.204	-3.207	0.001340209	0.004767185	0	0.002	0.0016	0.0016	TRUE	NA	ENSG00000C	3441	20	
15	68	50	44.35	30	59	2.01	1.007	2.337	0.019435172	0.050444348	0	0.0721	0.0637	0.0637	TRUE	NA	ENSG00000C	4644	0	

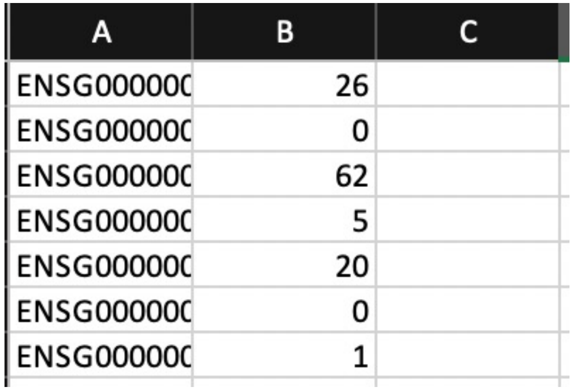
Prepare exercise 2

In Excel (or an equivalent), create a new file with the following columns **without headers**:

- Id (contains ENSEMBL IDs)
- siLuc (normalized and divided by median of transcripts length in kb)

Hint: copy and paste normalized data with a special paste - **by value** – so that it doesn't copy the formula

Save the file as a Text (separator: tabs) (.txt) file named RNAseq_data_ready_for_seqMINER.txt

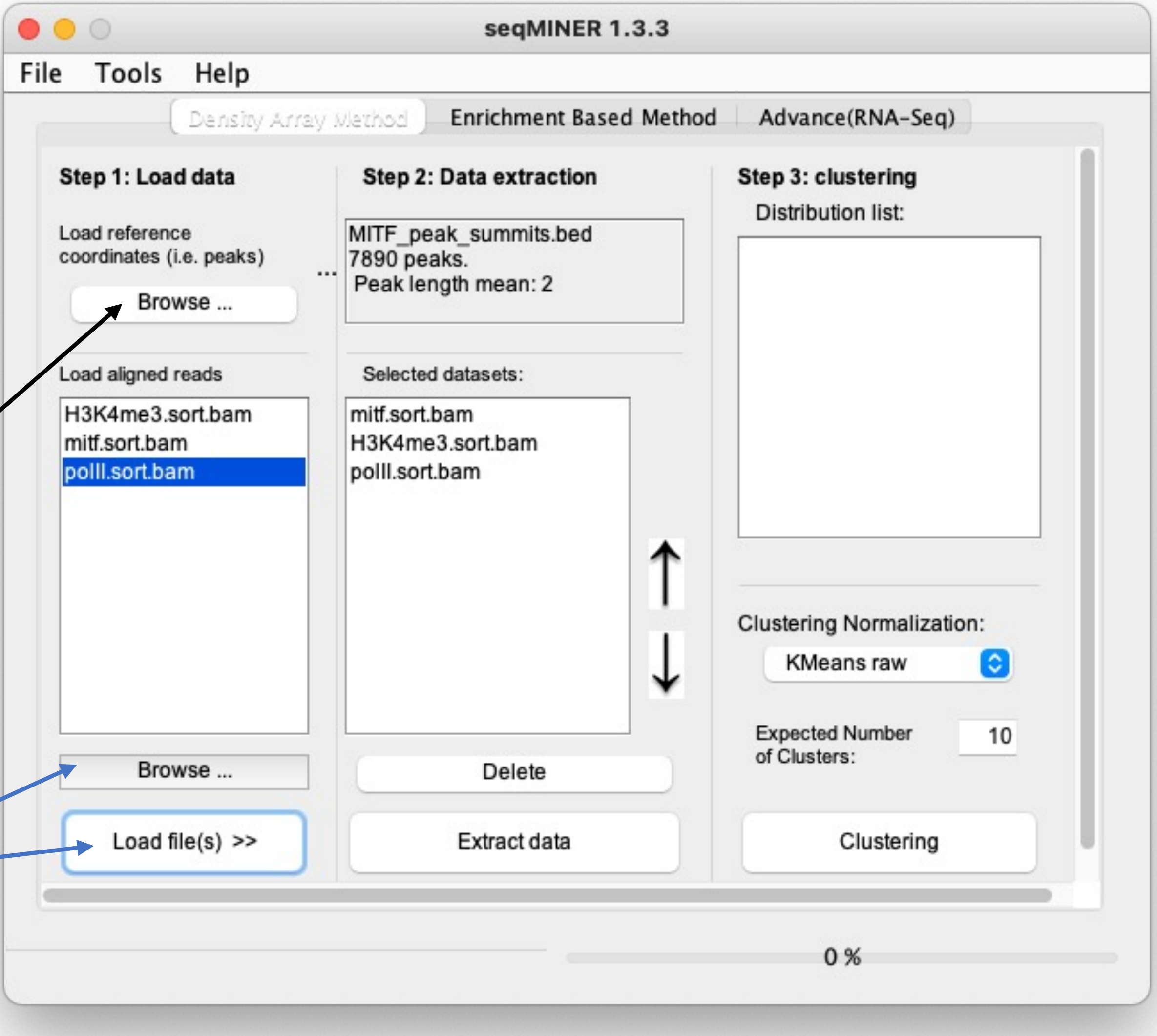


A	B	C
ENSG00000C	26	
ENSG00000C	0	
ENSG00000C	62	
ENSG00000C	5	
ENSG00000C	20	
ENSG00000C	0	
ENSG00000C	1	

Exercise 2

Use seqMINER to visualize at the same time chIP-seq data along with RNA-seq data

- Load MITF_peak_summits.bed (from your history "ChIP-seq data analysis") as reference coordinates.
- Load the 3 bam files (directory chipseq/mapping):
    - mitf.sort.bam
    - H3K4me3.sort.bam
    - polII.sort.bam



Load aligned reads:
H3K4me3.sort.bam
mitf.sort.bam
polII.sort.bam

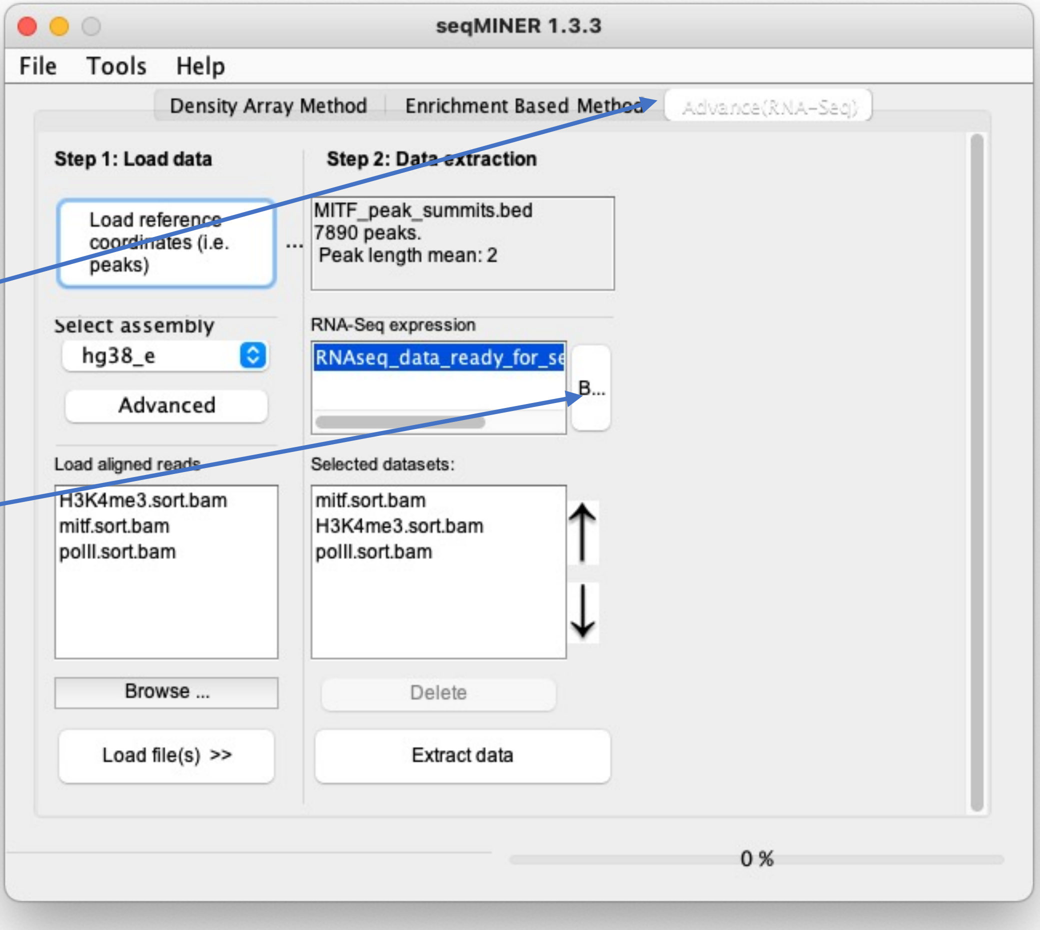
MITF_peak_summits.bed
7890 peaks.
Peak length mean: 2

Selected datasets:
mitf.sort.bam
H3K4me3.sort.bam
polII.sort.bam

Exercise 2

In the Advance (RNAseq) tab:

1. Upload the file RNAseq_data_ready_for_seqMINER.txt:
   - 1st column contains Ensembl Gene IDs
   - 2nd column contains normalized read counts of siLuc divided by gene length in Kb.
2. In advanced:
   - Load a reference file: click on Browse and select the file hg38_ens105.bed.
   - Choose a reference from database: hg38_ens105.bed



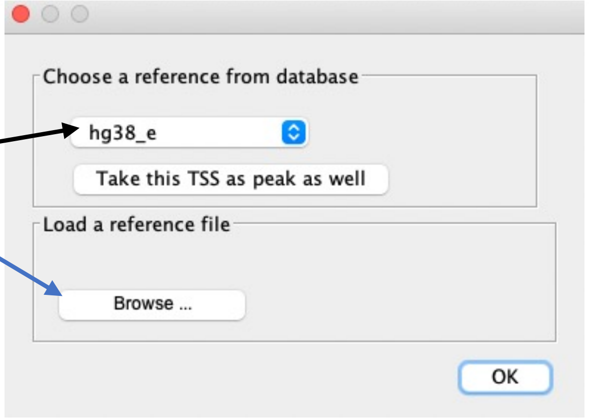
MITF_peak_summits.bed
7890 peaks.
Peak length mean: 2

Select assembly: hg38_e

RNA-Seq expression: RNAseq_data_ready_for_se

Load aligned reads:
H3K4me3.sort.bam
mitf.sort.bam
polII.sort.bam

Selected datasets:
mitf.sort.bam
H3K4me3.sort.bam
polII.sort.bam

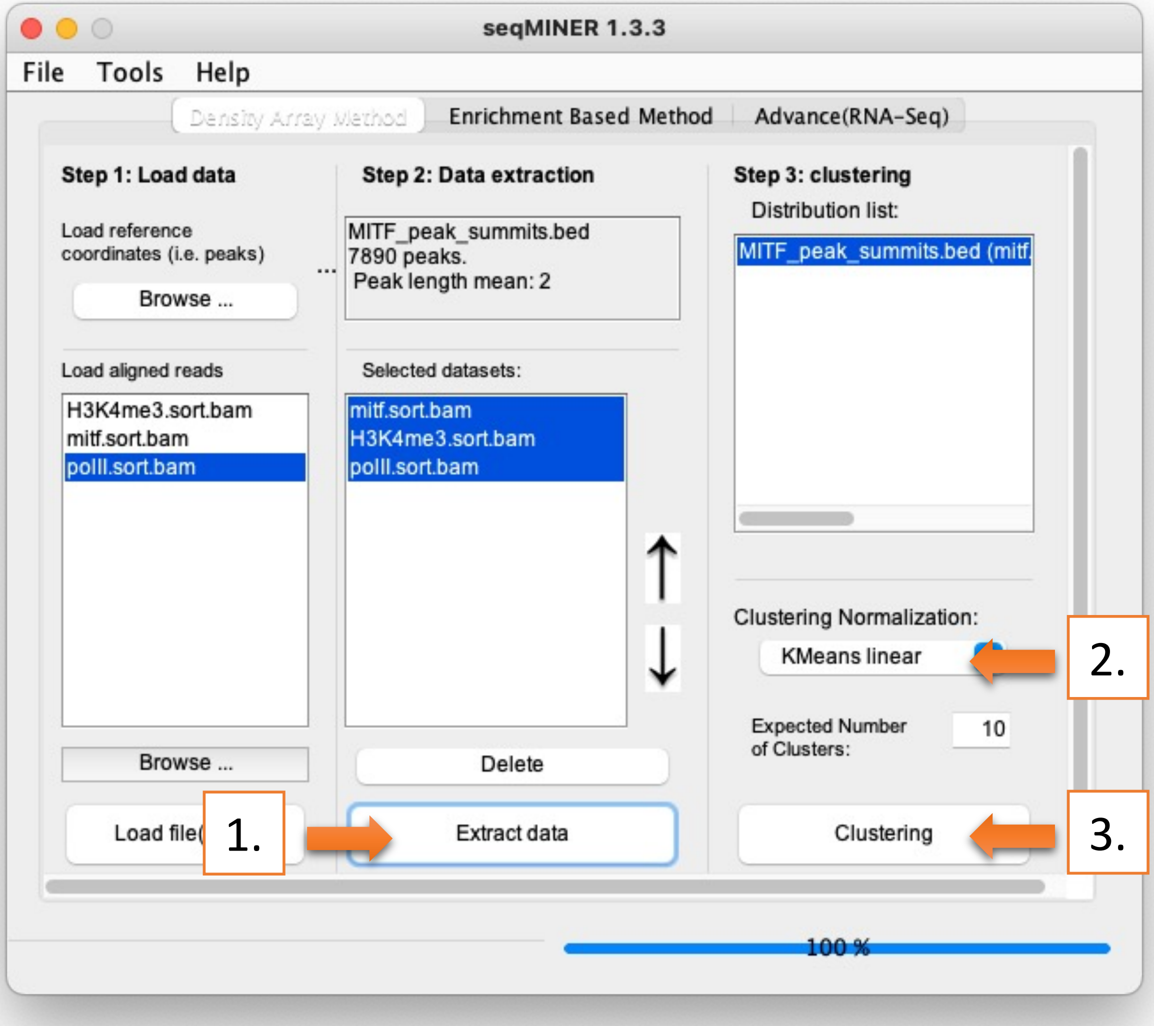


Choose a reference from database: hg38_e

Load a reference file

Exercise 2

Be careful, make sure that in Options > Gene profile, Gene profile analysis is not selected before clicking on « **Extract data** ».



MITF_peak_summits.bed
7890 peaks.
Peak length mean: 2

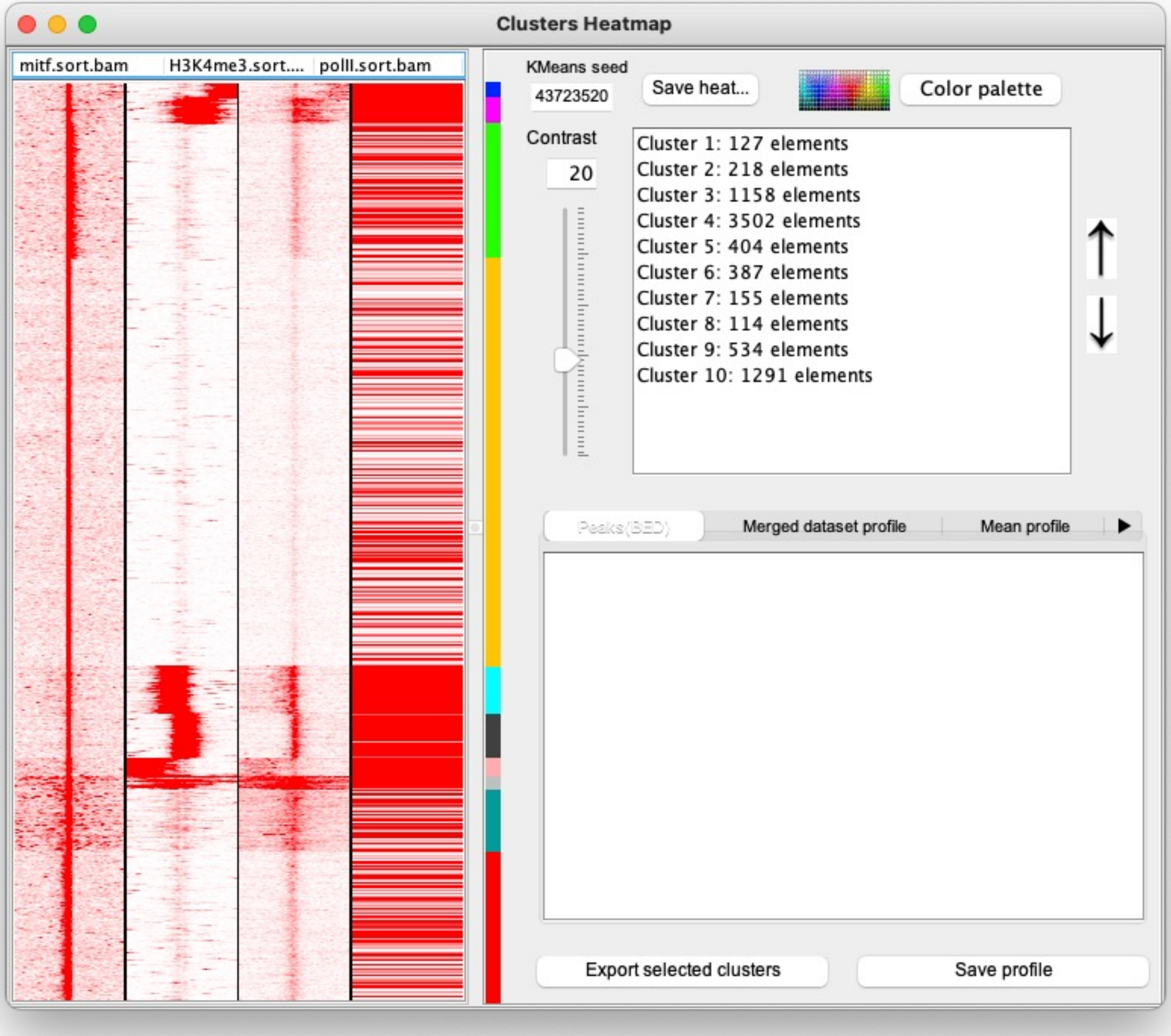
Load aligned reads: H3K4me3.sort.bam, mitf.sort.bam, polII.sort.bam

Selected datasets: mitf.sort.bam, H3K4me3.sort.bam, polII.sort.bam

Distribution list: MITF_peak_summits.bed (mitf

Expected Number of Clusters: 10

Exercise 2



KMeans seed 43723520

Contrast 20

Cluster 1: 127 elements
Cluster 2: 218 elements
Cluster 3: 1158 elements
Cluster 4: 3502 elements
Cluster 5: 404 elements
Cluster 6: 387 elements
Cluster 7: 155 elements
Cluster 8: 114 elements
Cluster 9: 534 elements
Cluster 10: 1291 elements