Analysis of ChIP-seq peaks

Stéphanie Le Gras
(slegras@igbmc.fr)

Guidelines

Raw data — Fastq file

Mapping — BAM file

Visualization

BED/WIG files

Peak detection

Clustering
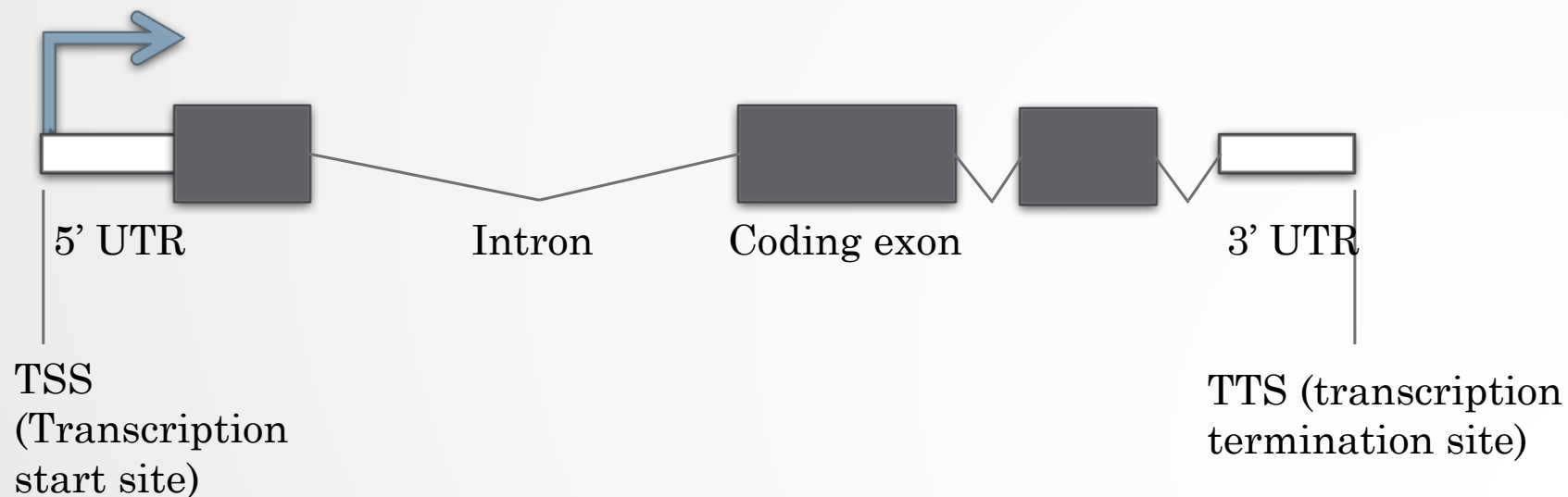
Annotation

Motif discovery
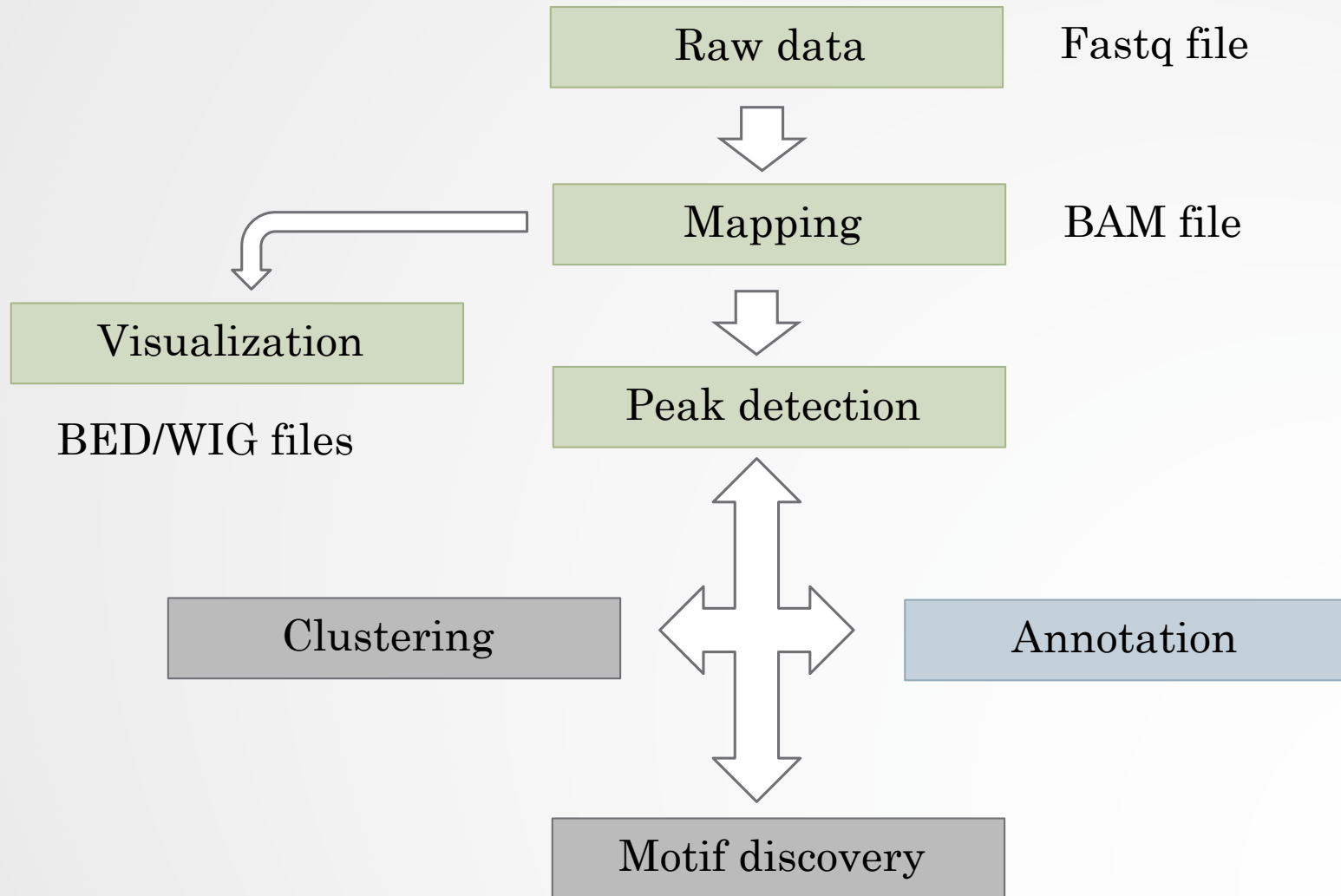
Peak annotation

- Goal: assigning a peak to one or many genome features

- Always be careful on the database used to annotate the peaks (either RefSeq or Ensembl)

- Many tools exist (GPAT, CEAS, CisGenome, Homer…)
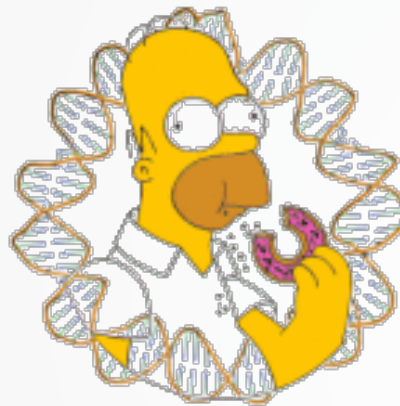
Peak annotation (Homer)

- Works in two parts:
  - Determines the distance to the nearest TSS and assigns the peak to that gene
  - Determines the genomic annotation of the region occupied by the center of the peak/region

- Default behaviour is to use RefSeq annotations

5' UTR Intron Coding exon 3' UTR

TSS
(Transcription
start site)

TTS (transcription
termination site)

Peak annotation (Homer)

- Rank:

1. TSS (by default defined from -1kb to +100bp)

2. TTS (by default defined from -100 bp to +1kb)

3. CDS Exons

4. 5' UTR Exons

5. 3' UTR Exons

6. **CpG Islands

7. **Repeats

8. Introns

9. Intergenic

Exercise 1: peak annotation

Now that we have called peaks, we would like associated the peaks with nearby genes.

- 1. Use the **homer_annotatePeaks** tool to perform the peak annotation.
  - Homer peaks OR BED format: MITF peaks narrow peaks dataset (2nd run of Macs2)
  - Genome version: hg38

- 2. The Homer annotatePeaks tool generates two datasets: a log file and a tabular file which contains annotated peaks. Change datatype of the dataset with the annotated peaks from csv to **tabular**. NOTE: the tool falsely set the output format as csv (comma separated values file) while it's a tsv (tab separated values file). Tsv format is called tabular in Galaxy.
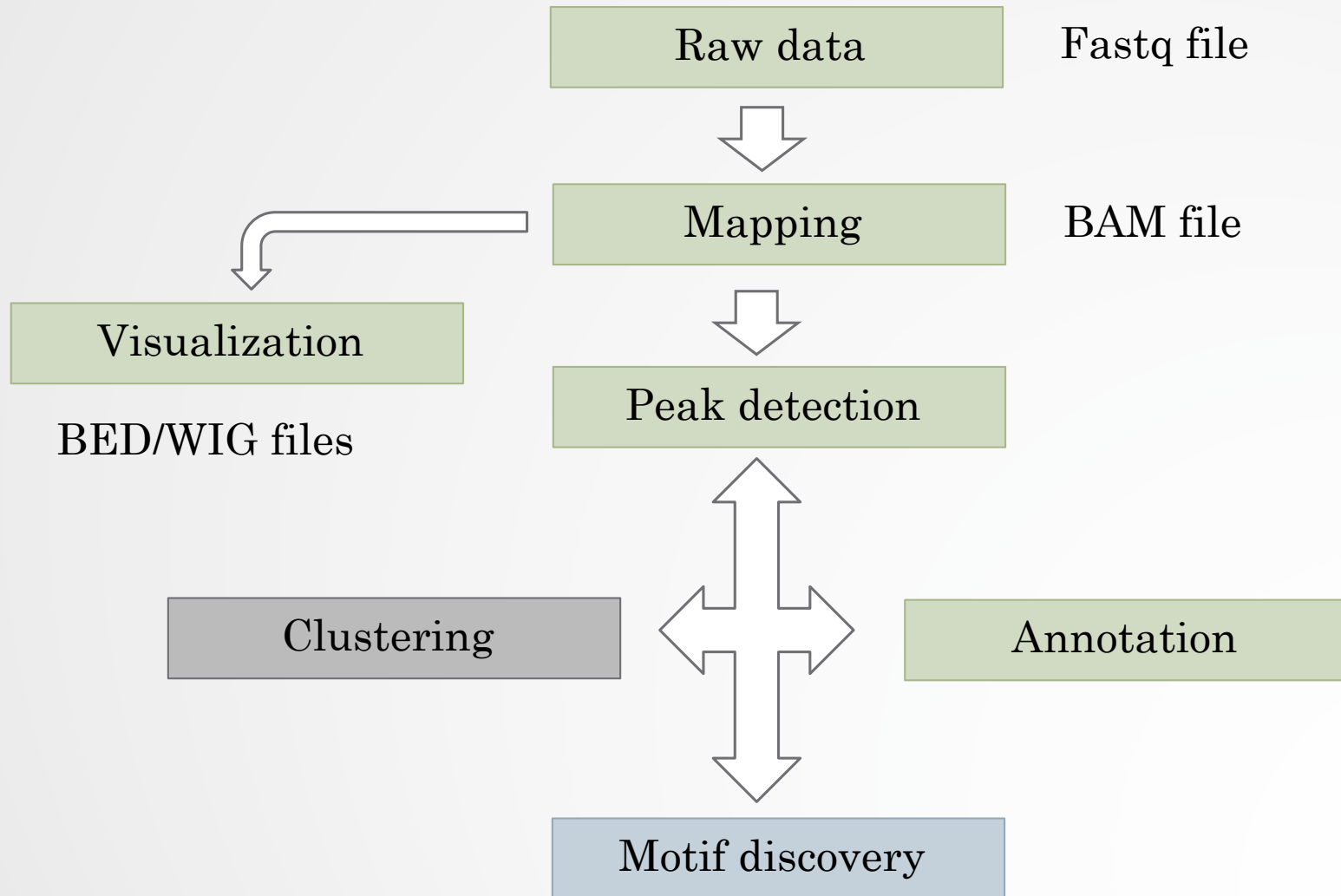
Common plots generated after the annotation steps are:
  - An histogram of the distances Peak <-> TSS
  - A pie chart presenting the proportion of genomic features

- 3. Generate an histogram of the distance Peak <-> TSS using the tool **Histogram**

6

Exercise 1: peak annotation

- 4. Draw a pie chart presenting the proportion of genomic features associated to the MITF peaks. To achieve this, we are going to count the number of time the genomic features (intron, exon…) are found in the Annotation column of the dataset (tabular) generated in 1.
  - 4.a. Use the tool **Cut** to extract the column "Annotation" from the dataset which contains the annotated peaks.
  - 4.b. In the column Annotation, genomic features (exon, intron…) are associated to gene names. We would like to have a table which contains a column with only the genomic features. Split the data contained in the Annotation column using whitespaces with the tool **Convert**
  - 4.c. the column containing genomic features starts with the header « Annotation ». Remove the first line with the tool **Remove beginning**.
  - 4.d. Use the tool **Count** to count the number of each of the genomic features. Sort in descending order.
  - 4.e. Expand the box of the dataset generated in 4.d and click on ▥ to generate a pie chart on the data. You can name the pie chart "Proportion of peaks falling into several genomic features."

Guidelines

Raw data — Fastq file

Mapping — BAM file

Visualization

BED/WIG files

Peak detection

Clustering

Annotation

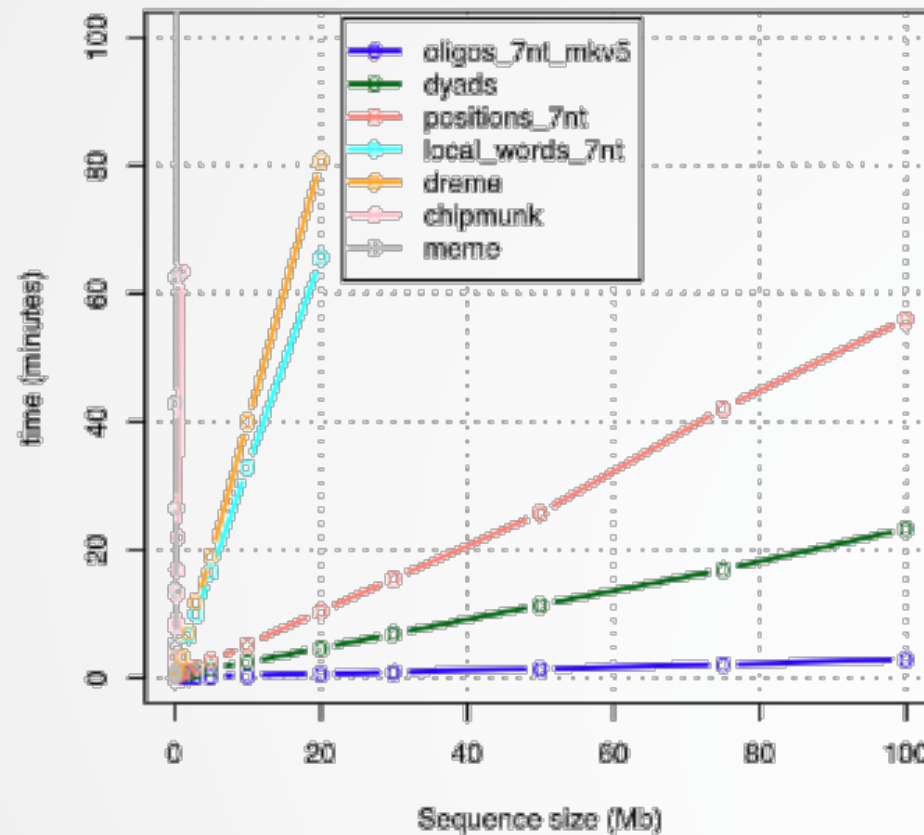Motif discovery

Motif discovery

- Sequence to which the protein of interest may be bound

- Search for enriched nucleotide sequences (i.e motifs) within peak sequences.
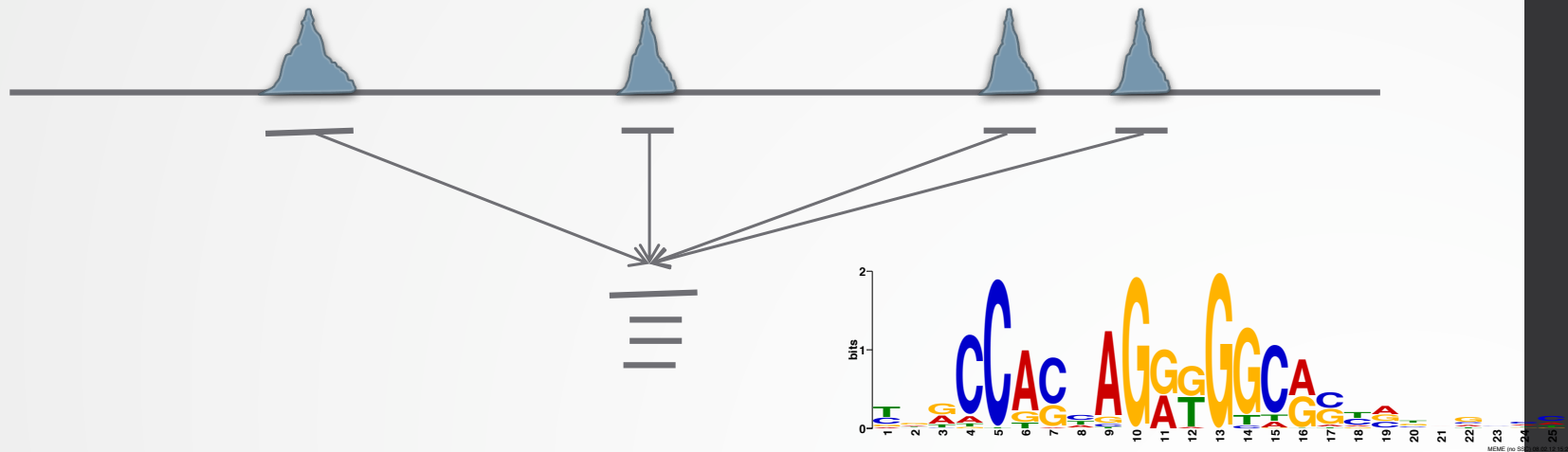
- De novo motif discovery
- Motif searching based on motif databases (JASPAR, Transfac)

De novo motif searching

- Lot of tools exist (Homer, RSAT, MEME-suite…)

- Be careful on the complexity of the algorithms

Morgane Thomas Chollier et al, 2011, NAR

De novo motif discovery

- MEME-suite:
  - MEME (Bailey et al. 1994)
    - Long motifs
    - Complexes of TFs
    - Complexity of the algorithm!
  - DREME (Bailey et al. 2011)
    - Faster than MEME
    - Can have more input sequences (but shorter ~100b)
    - Find regular expression (not PSSM)
    - Short motifs (3 to 8 nucleotides by default)
  - MEME-chIP (Machanick et al. 2011)
    - Pipeline based on the use of several tools from the MEME-suite including DREME, MEME, TOMTOM (Gupta et al, 2007)
    - Only 100b sequences are analyzed
    - A maximum of 600 sequences (randomly selected from the input) are input to the MEME algorithm

MEME-chIP

- MEME and DREME: discover novel DNA-binding motifs

- CentriMo: determine which motifs are most centrally enriched

- Tomtom: analyze them for similarity to known binding motifs

- SpaMo: perform a motif spacing analysis

- MEME-chIP automatically group significant motifs by similarity

Exercise 2: *de novo* motif discovery

We would like to know if there are over-represented nucleotide sequences (i.e motifs) in MITF peaks. Use MEME-chIP (http://meme-suite.org/tools/meme-chip) to perform *de novo* motif discovery in nucleotide sequences located +/- 100b around MITF peak summits

- 1. Extract the top 800 peak summits (ranked by -log10pvalue)
    - 1.a. Sort the peak summits by decreased -log10pvalue using the tool **Sort**
    - 1.b. Extract the top 800 peak summits using the tool **Select first**

- 2. In Galaxy, compute the coordinates of the peak summits +/- 100b using the dataset which contains MITF peak summits (2nd run of Macs2)
    - 2.a. Use the chromosome length file hg38.len from the data library "Chromosome length"
    - 2.b. Use the tool called **SlopBed**

- 3. Extract fasta sequences from the coordinates of the peak summits using the tool **Extract Genomic DNA**

- 4. Download the file, go to MEME-chIP and run MEME-chIP with default parameters on the data

PWM

- **position weight matrix (PWM)**, also known as a **position-specific weight matrix (PSWM)** or **position-specific scoring matrix (PSSM)**

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

http://weblogo.berkeley.edu/logo.cgi

14

Known motif searching

- Charles E. Grant, Timothy L. Bailey, and William Stafford Noble, "FIMO: Scanning for occurrences of a given motif", *Bioinformatics* 27(7):1017–1018, 2011

- Scan nucleotide sequences of interest for PWMs.

- JASPAR, Transfac databases

- Some PWMs are provided by MEME.

Known motif searching

http://meme-suite.org/meme-software/Databases/motifs/motif_databases.12.7.tgz

```
MOTIF MA0001.1 AGL3

log-odds matrix: alength= 4 w= 10 E= 0
-10000      195     -460     -360
  -301      163   -10000      -35
   170     -260     -301     -114
    72     -301     -260      104
   144     -460     -460       26
    99     -360   -10000       95
   142     -228     -228      -14
  -114     -360     -301      174
   142     -301       21     -460
-10000     -301      186     -201

letter-probability matrix: alength= 4 w= 10 nsites= 97 E= 0
 0.000000        0.969072        0.010309        0.020619
 0.030928        0.773196        0.000000        0.195876
 0.814433        0.041237        0.030928        0.113402
 0.412371        0.030928        0.041237        0.515464
 0.680412        0.010309        0.010309        0.298969
 0.494845        0.020619        0.000000        0.484536
 0.670103        0.051546        0.051546        0.226804
 0.113402        0.020619        0.030928        0.835052
 0.670103        0.030928        0.288660        0.010309
 0.000000        0.030928        0.907216        0.061856
```

Guidelines

Raw data — Fastq file

Mapping — BAM file

Visualization

BED/WIG files

Peak detection

Clustering

Annotation

Motif discovery
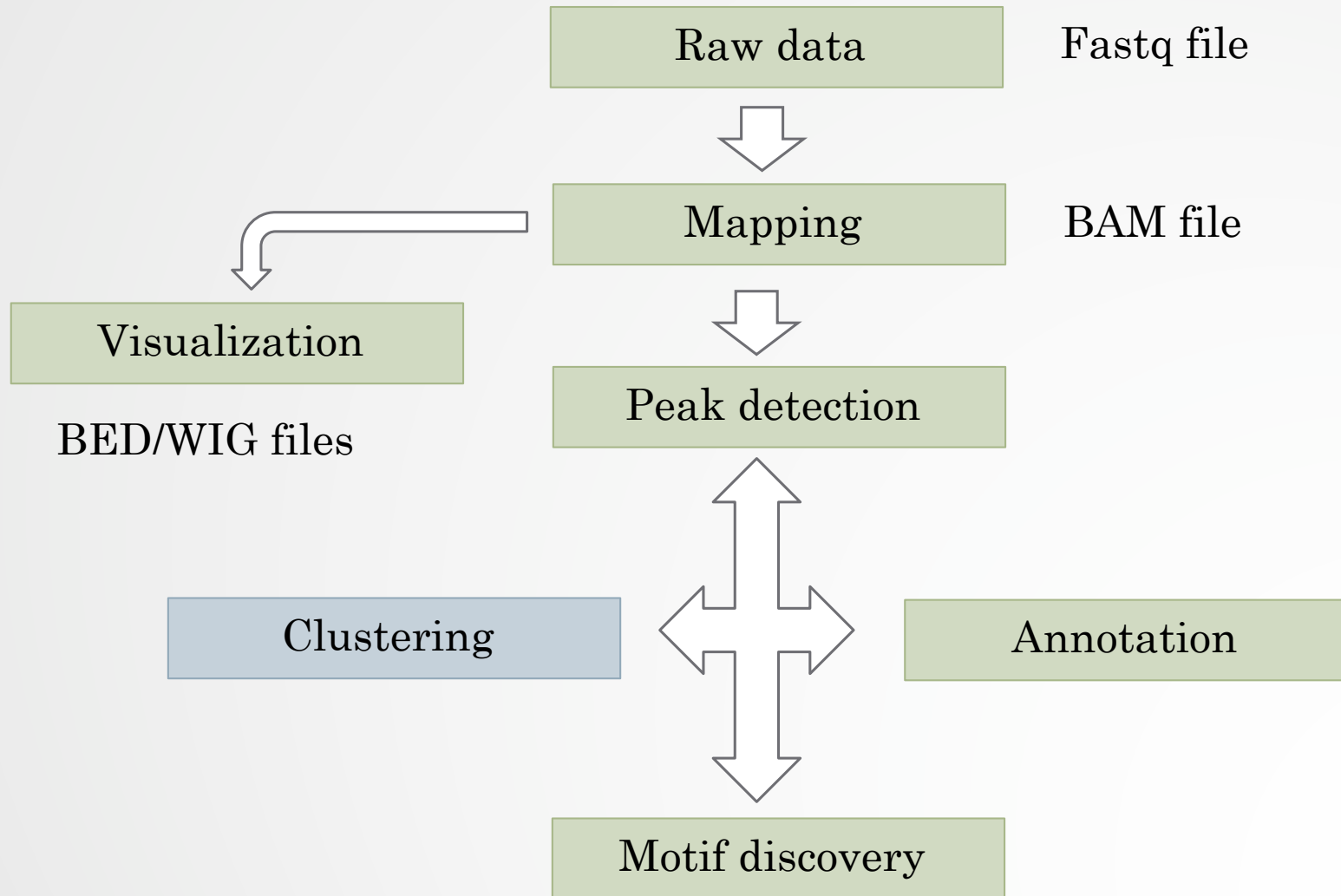
Meta-profiles

- Global visualization of the data

- Need:
  - Regions of interest
    - Regions around a reference point e.g TSS +/- 1Kb,…
    - Scaled regions e.g peaks, gene bodies,…
  - Signal data (mapped reads)
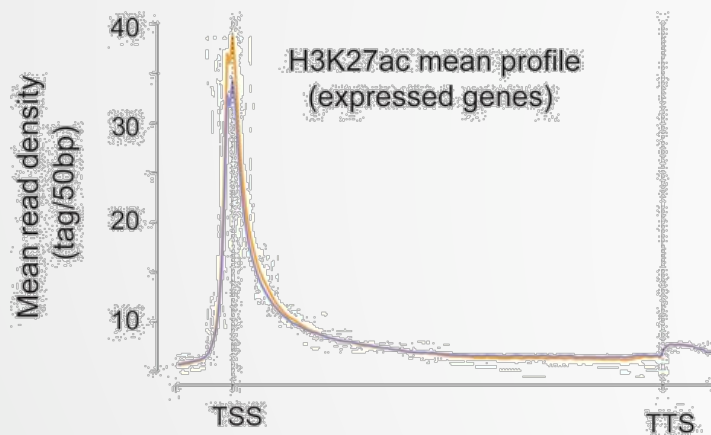
Mean profile   Heatmap
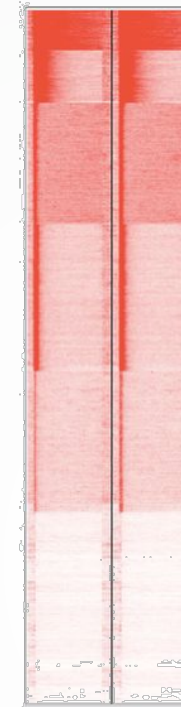
Computing meta-profiles
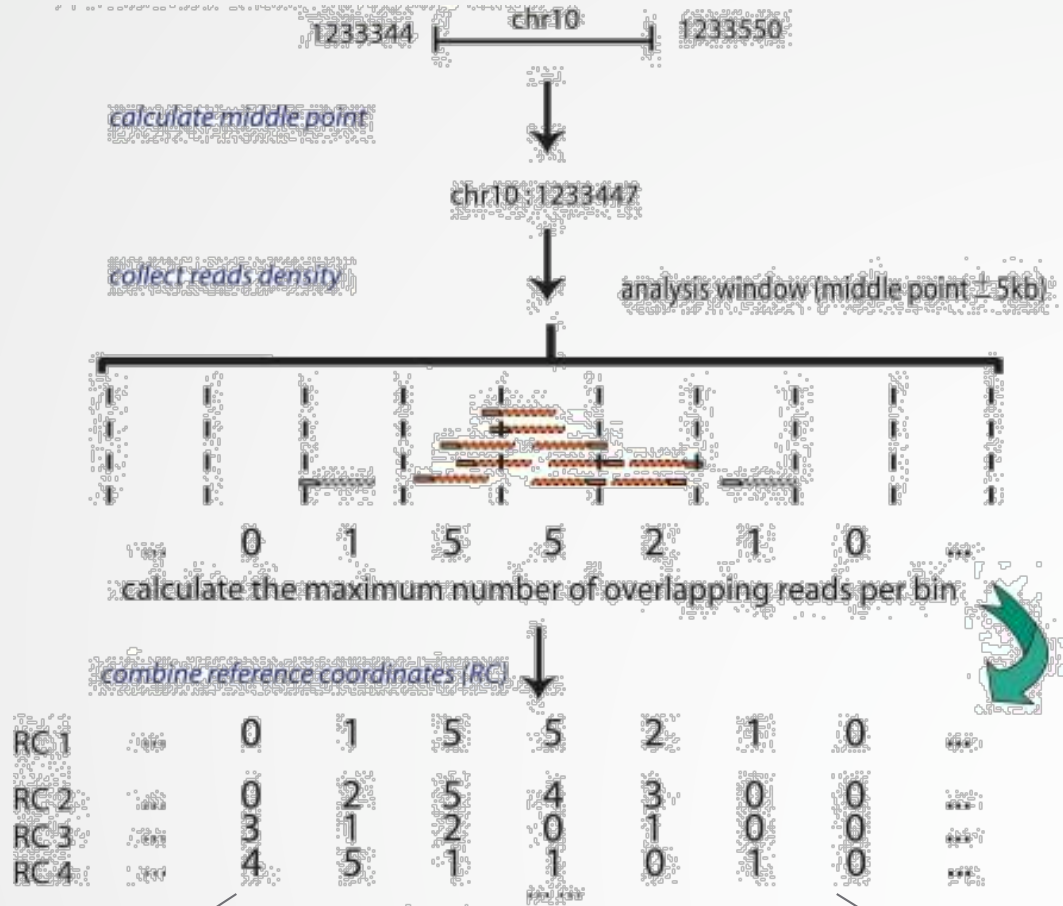
Ye et al, 2011

- (Clustering)
- Heatmap

- Mean of each column
  -> Mean profile

Clustering (heatmap)

- Group together genomic regions with similar enrichments

- In a single sample or multiple samples

- E.g:

TF

H3K4me3

RNA pol II
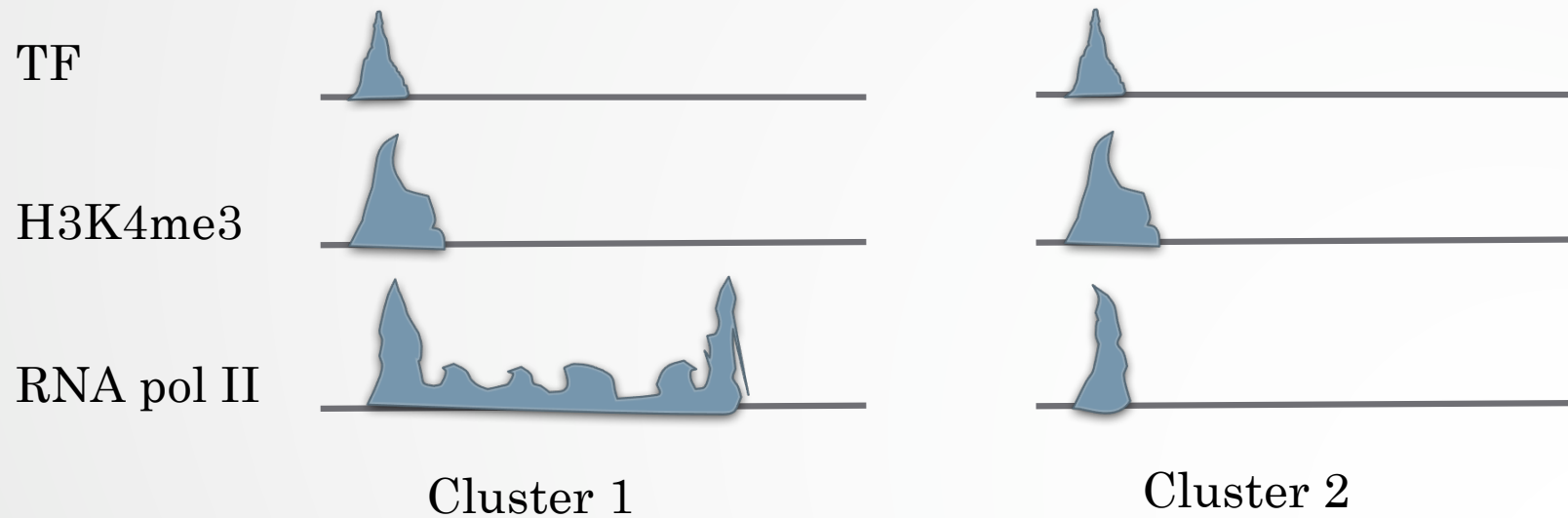
Cluster 1

Cluster 2

Clustering (heatmap)
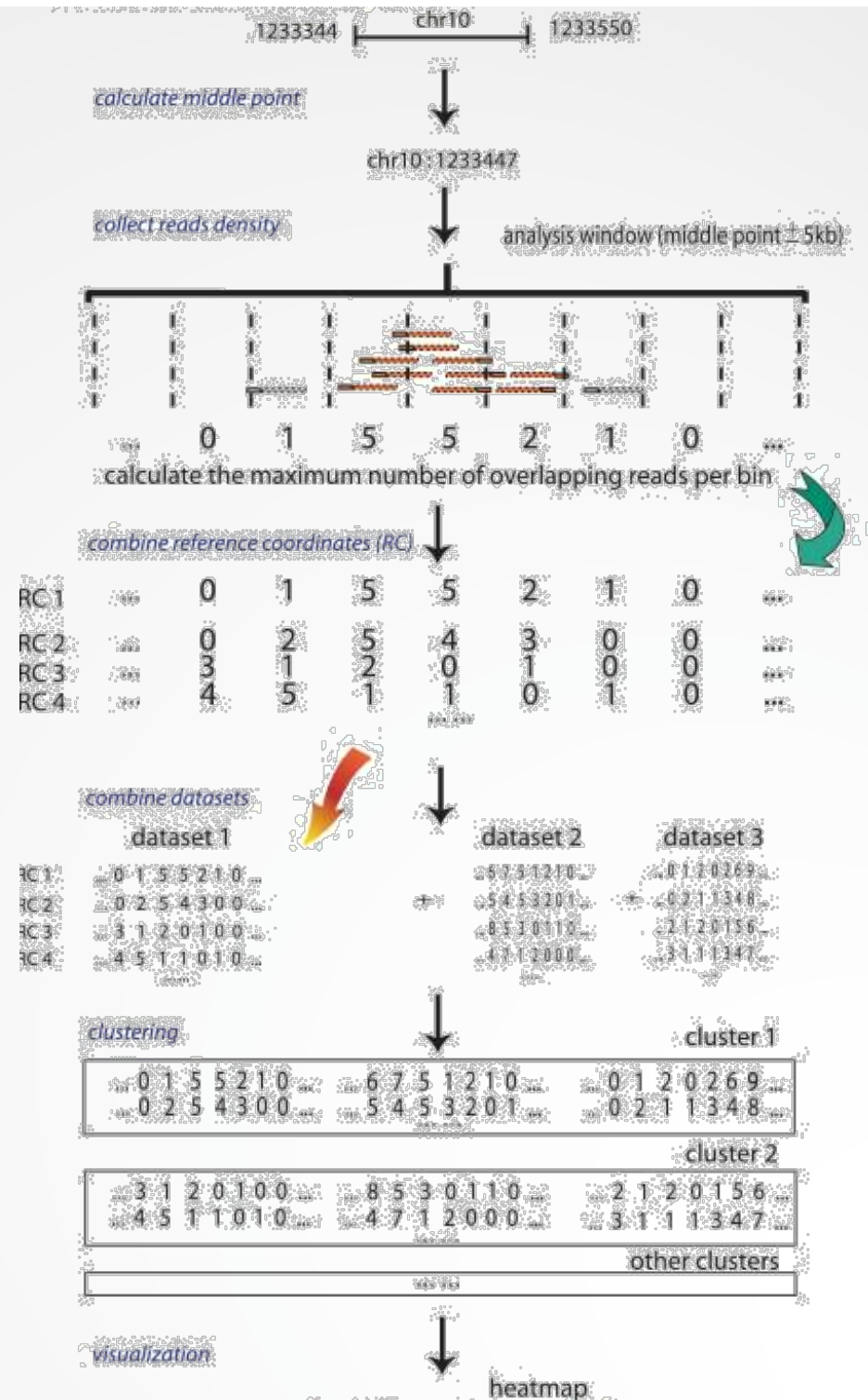
SeqMINER [Ye et al, 2011]
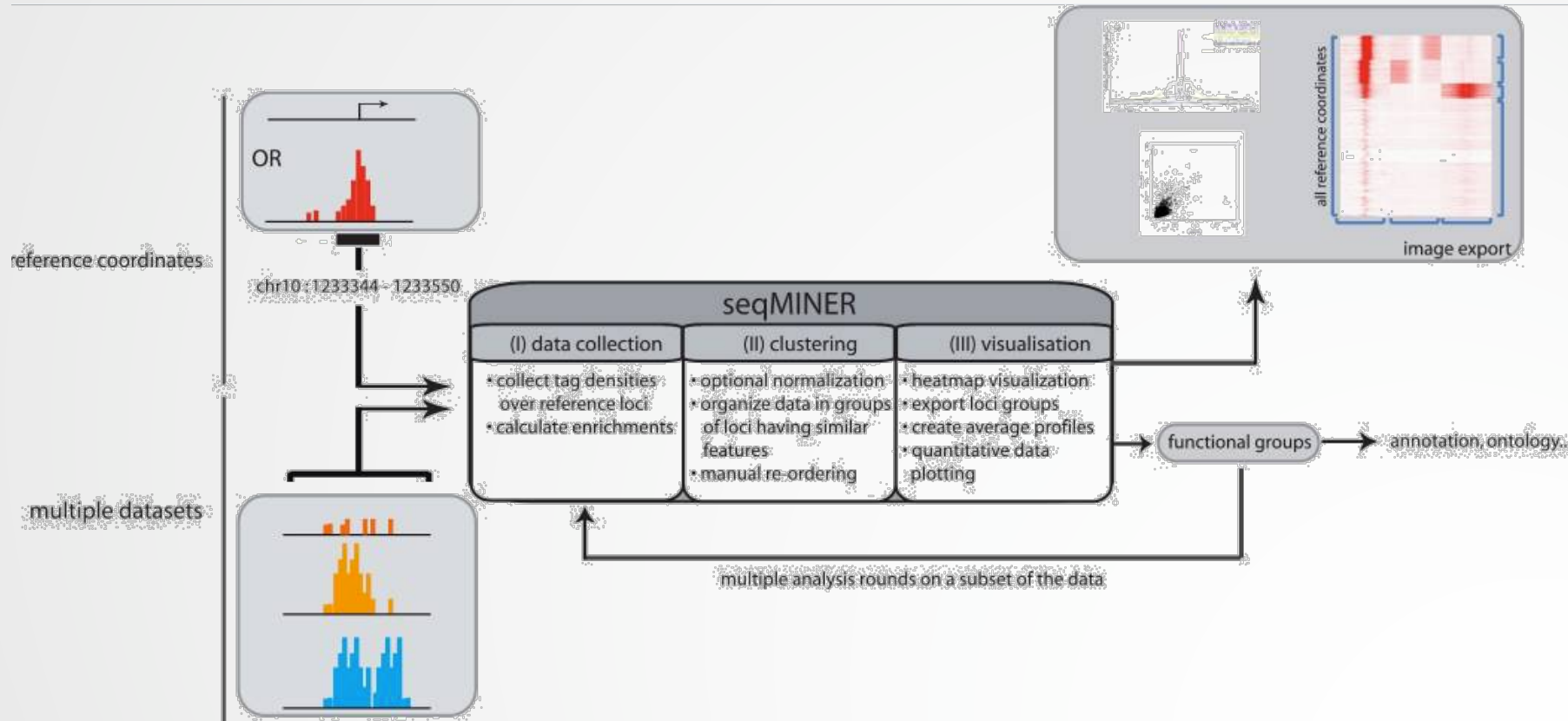
SeqMINER [Ye et al, 2011]
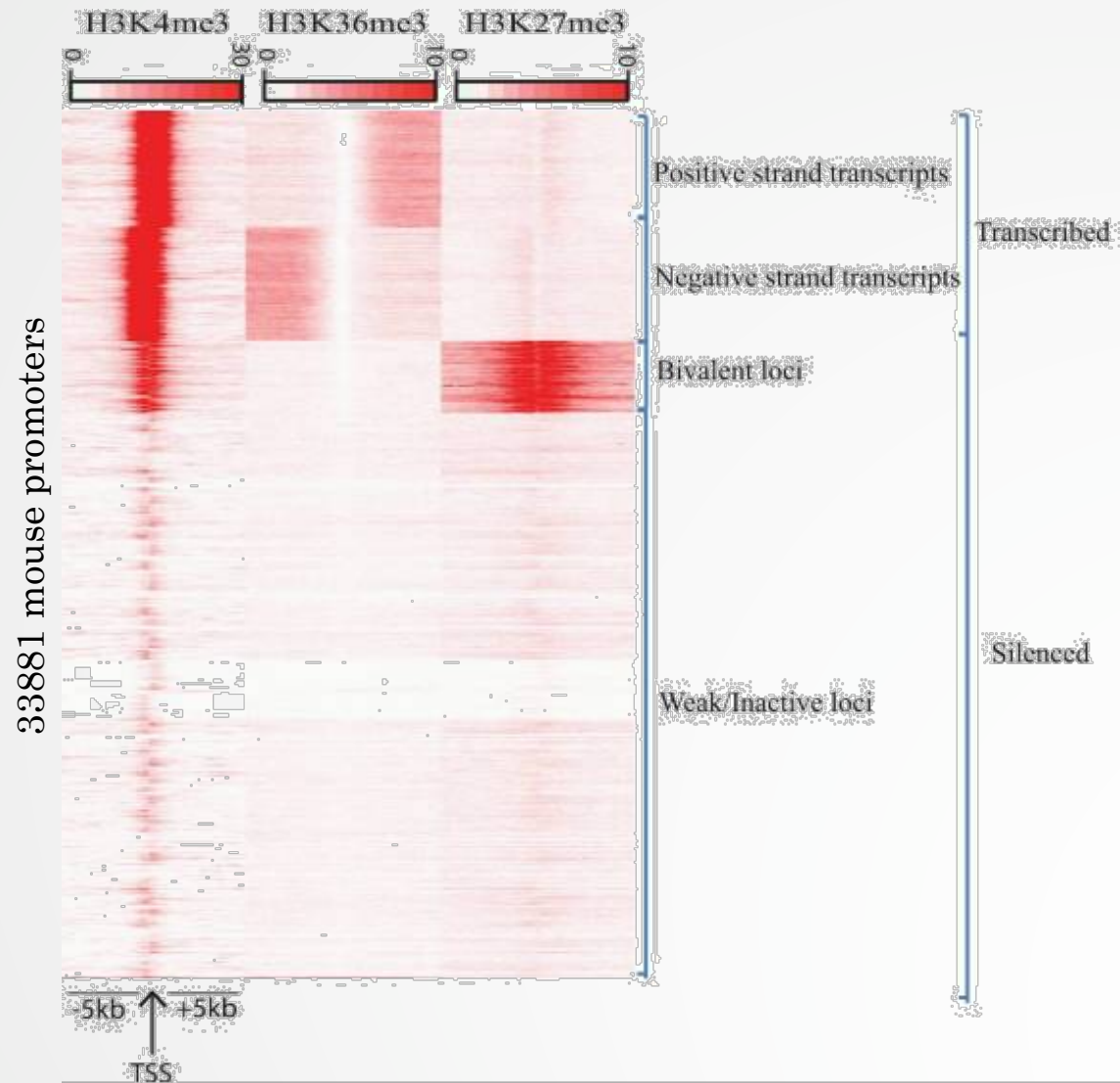
The darker the red the higher the read enrichment

Example
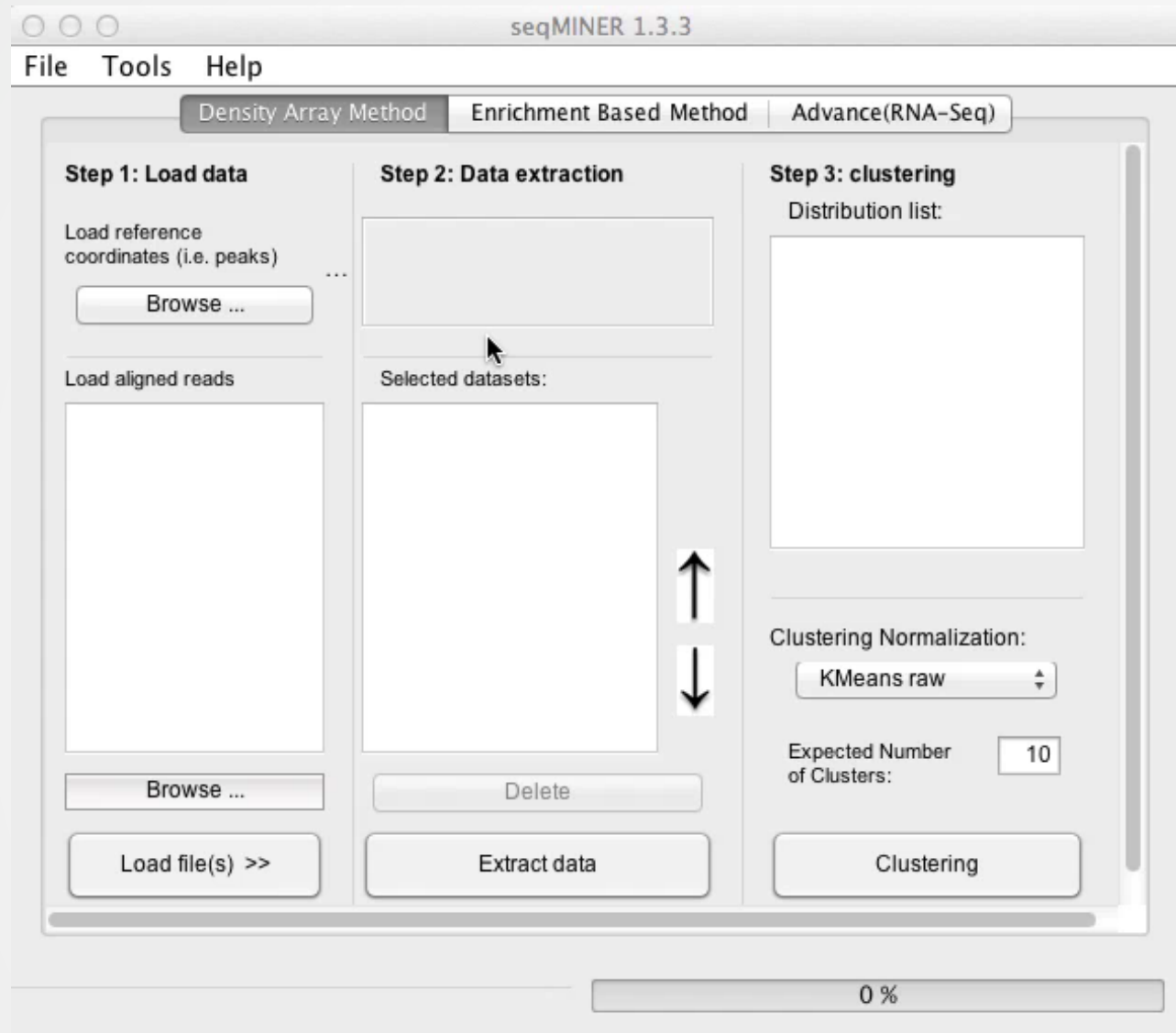
Exercise 3: Clustering

We have 2 additional datasets to those of MITF and the control : H3K4me3 and polII. Use seqMINER to have a look at the correlation between MITF, H3K4me3 and polII.

The tool is in the directory chipseq/seqMINER_1.3.3g. Go to this directory and run the tool by double-clicking on run_in_windows.bat.
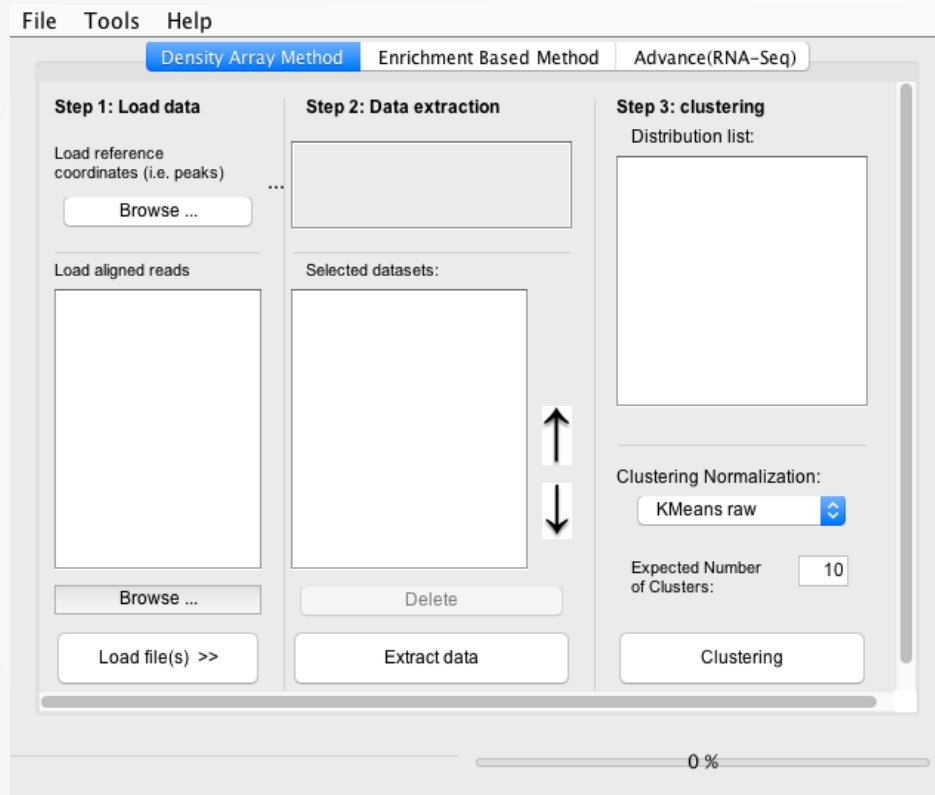
Exercise 3: Clustering

- We are going to have a look at MITF, H3K4me3, polII data at the TSS positions.

- To load the TSS positions of the human genome (hg38 assembly)
  - go to the tab Advance (RNA-Seq)
  - In the drop down list Select Assembly, select hg38_ensembl85. NOTE, selecting the assembly here is used to annotate the reference coordinates when visualizing the clusterings
  - Click on Advanced
  - Click on Take this TSS as peak as well
  - Click on Density Array Method. You now have :

Step 1: Load data

Load reference
coordinates (i.e. peaks)

Browse ...

...

Step 2: Data extraction

hg38_ensembl85.seqminer
57992 peaks.
Peak length mean: 0

Exercise 3: Clustering

- Load the datasets
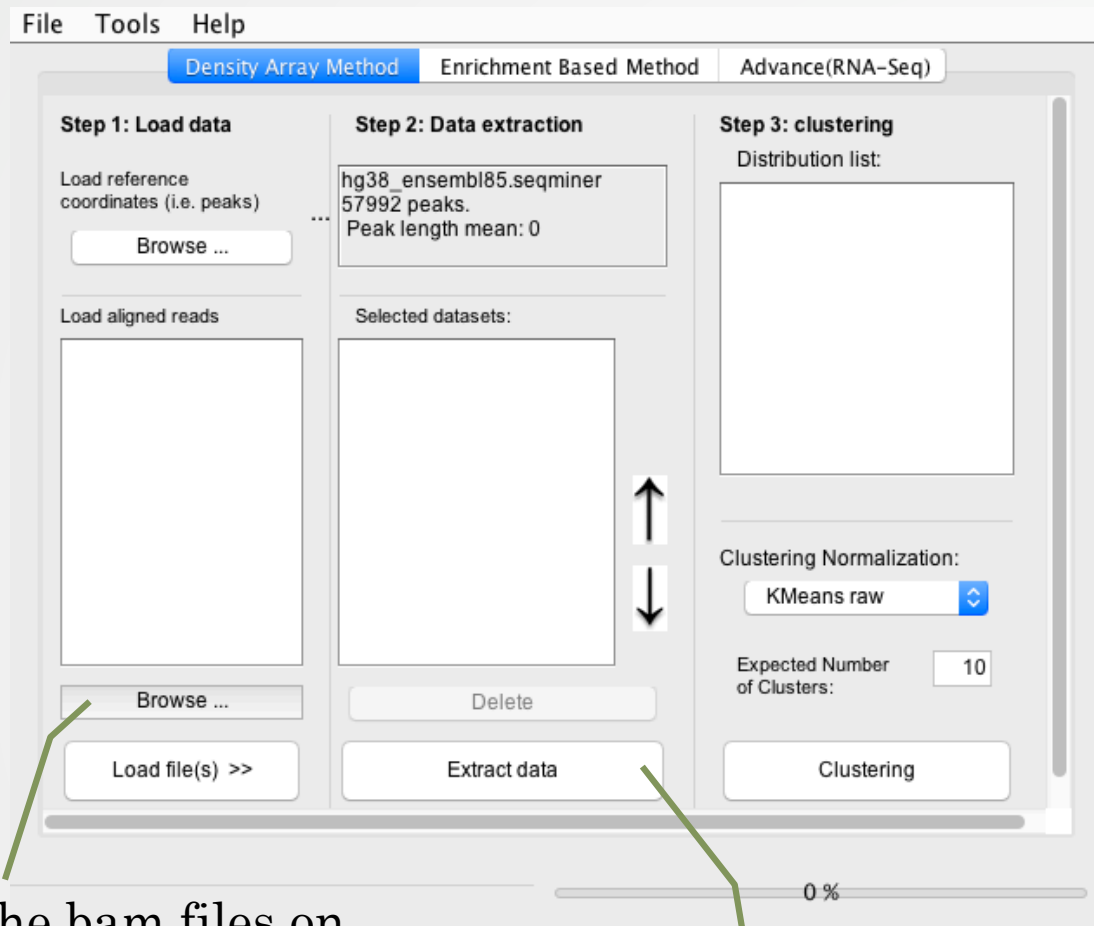
1. Load the bam files on MITF, polII, H3K4me3. Click on Browse, then on Load files. One by one.

2. Once step 1, is done, click on Extract data.

Exercise 3: Clustering

- In Clustering Normalization: select KMeans linear
- Click on Clustering

NOTE: we will all have different results, as the clustering method is Kmean. To have all the same results, we can use a Kmeans seed before running the clustering. To set the seed, go to Tools > options, select Run Kmeans with a given value and enter a value. For instance, the clustering below can be obtained with a Kmeans seed value of 9759332.
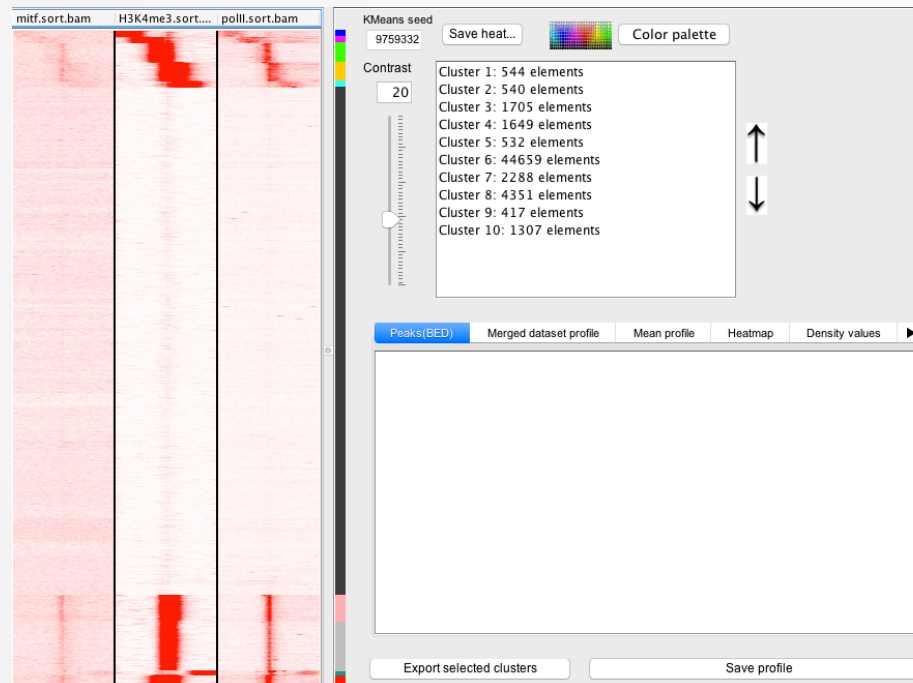
Exercise 3: Clustering

Heatmap

Cluster definition

Kmeans seed value

Clusters, click on one or multiple cluster names to display information in the panel below.
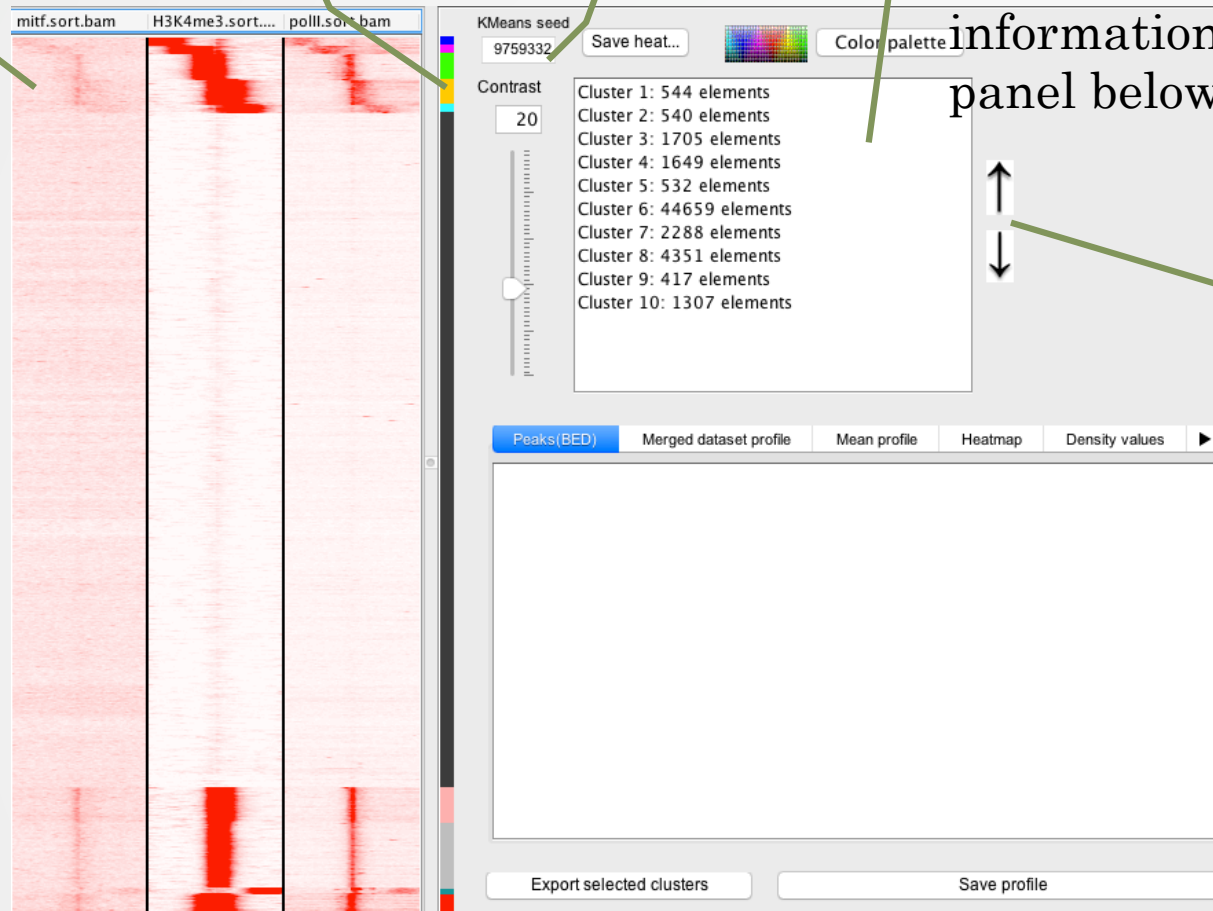
Change position of selected cluster in the heatmap and in the list

Exercise 3: Clustering

- Peaks (BED) : display the reference coordinates of the selected cluster(s)

- Merge dataset profile: display dataset mean profiles in one graph

- Mean profile: display mean profiles side by side

- Heatmap: Display mean profiles as heatmaps side by side. Useful to assess how dispersed the density values are

- Density values: Density values used to plot the heatmaps and the mean profiles

- Annotation: annotation of references coordinates (if annotation is filled in the advance(RNAseq) tab)

- Distance: Histogram of the distances TSS <-> reference coordinates

Exercise 3: Clustering

We are going to do a sub-clustering on reference coordinates (TSS) that have signal.

- Select all the clusters that have signal (1) and export the clusters (reference coordinates) into a file (2).
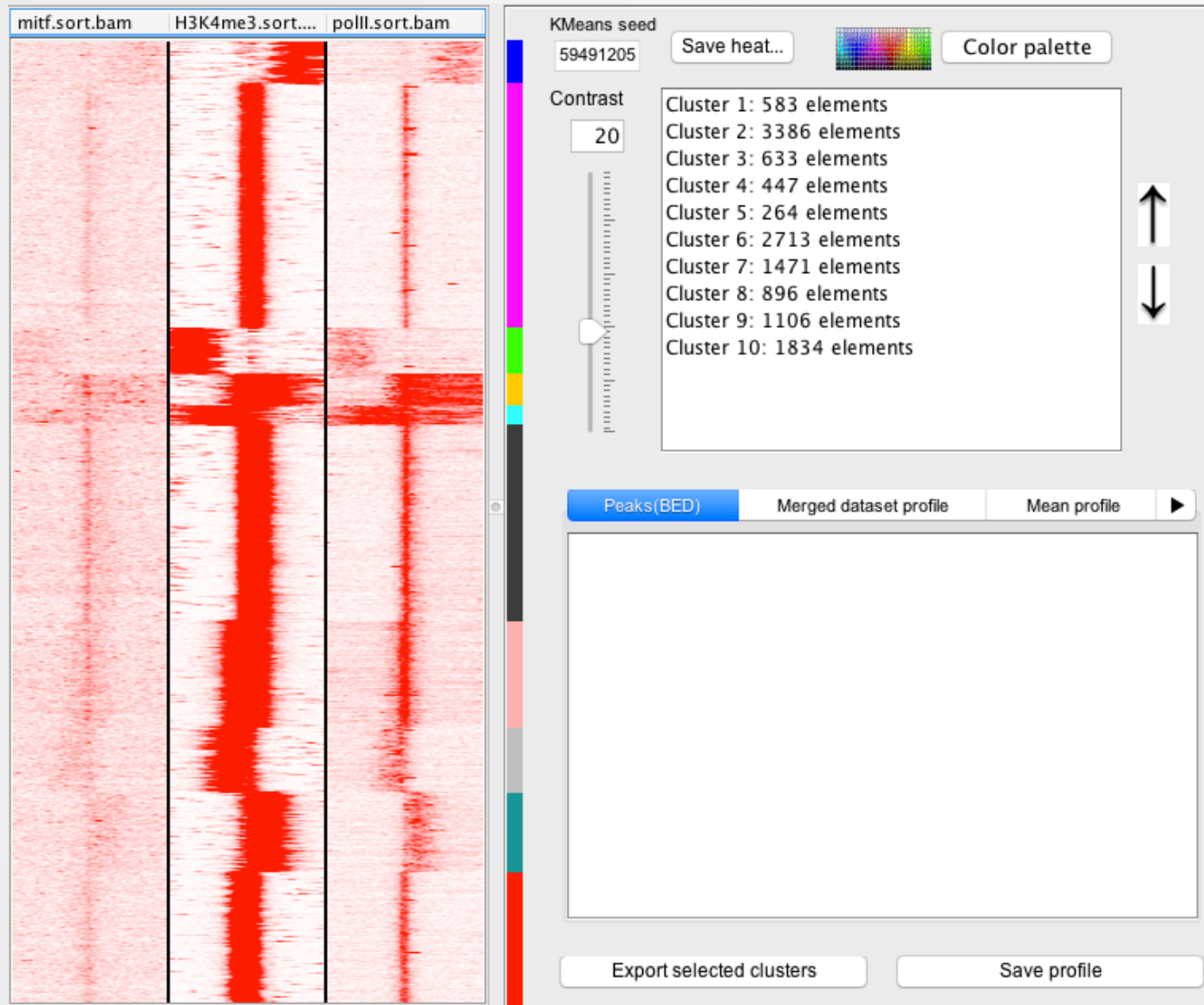
Exercise 3: Clustering

- Load the file previously generated (with cluster coordinated) as reference coordinates (1).

- Extract data (2)

- Run the clustering analysis (3)

Exercise 3: Clustering

Exercise 3: Clustering

- Before running any other analysis remove all the distributions from the distribution list (done to save memory)
  - Select a distribution, Click right on the name of a distribution and click on Delete.

- Run SeqMINER on all Ensembl (v85) genes.
  - Reference coordinates : the file is in chipseq > seqMINER_1.3.3g > lib > hg38_ensembl85.seqminer. NOTE: to be able to select the file, while browsing the file, click on file format, all type of file. SeqMINER limits by default reference coordinates file formats to (SAM, BAM, BED files). Load the file even if you're warned that the file is too big.
  - Go to Tools > Options, click on the Gene profile tab, select Gene profile analysis. Set parameters:
    - Inside bin number: 100
    - Outside bin number (left): 10
    - (right): 10
  - In the general tab, select Run Kmeans with a given value : 97787647
  - Click on OK. NOTE: this option makes SeqMINER to run the analysis on entire reference regions instead of on the middle of the regions +/- 5kb. All regions are normalized to a region of the same length.
  - Click on Extract data
  - Click on Clustering

Exercise 3: Clustering
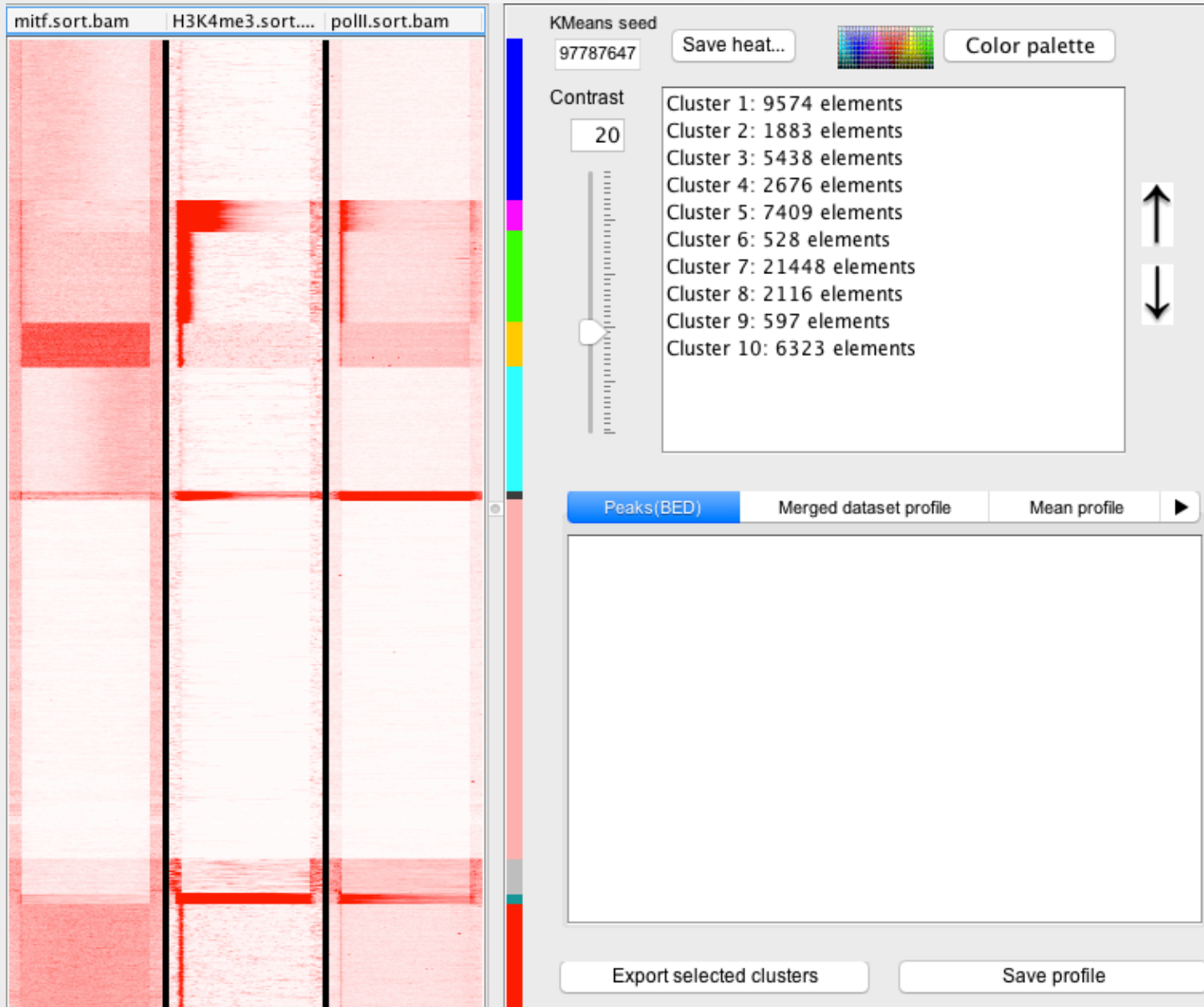
Exercise 3: Clustering

- 1. Select all clusters which contains MITF, polII and H3K4me3 (clusters 2, 3, 4, 6, 9)
    - Do a sub-clustering (keep same Kmeans seed)

- 2. Additional question:
    - 2.a. Export cluster 5.
    - 2.b. Open the file with Excel, open a web browser to DAVID (https://david-d.ncifcrf.gov/), run a functional annotation analysis (functional annotation clustering) with the Ensembl Gene IDs from the file in excel.

Guidelines

Raw data — Fastq file

Mapping — BAM file

Visualization

BED/WIG files

Peak detection

Clustering

Annotation

Motif discovery