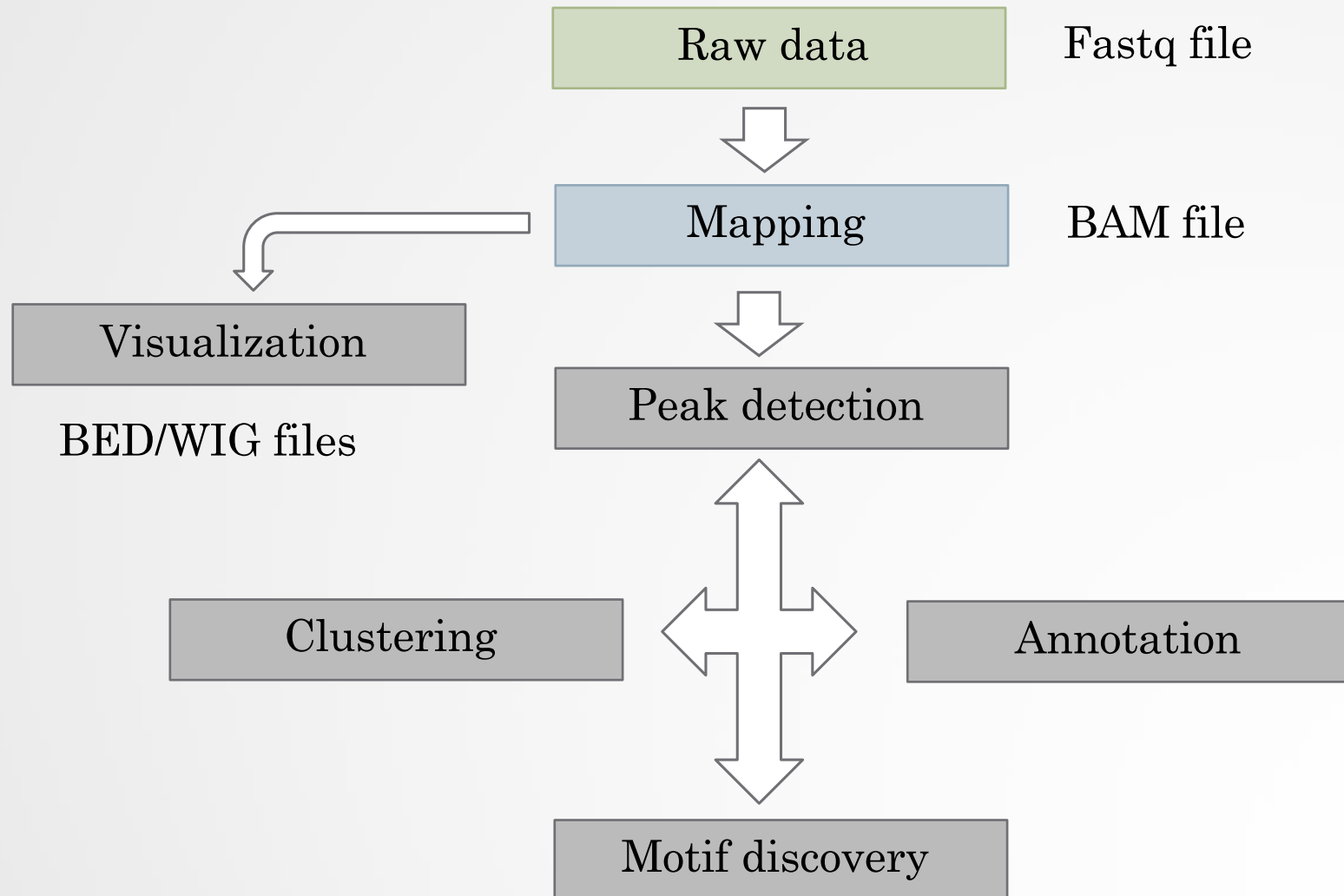


# Analysis of ChIP-seq data

Stéphanie Le Gras  
([slegras@igbmc.fr](mailto:slegras@igbmc.fr))

# Guidelines



# Mapping and visualization of ChIP-seq data

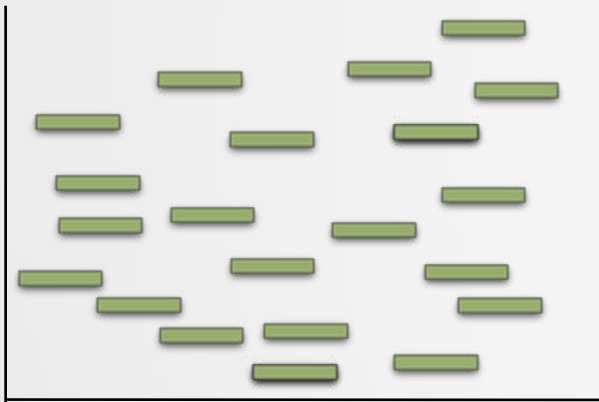
# Mapping

- Find out the position of the reads within the genome

Ref. Genome



Reads

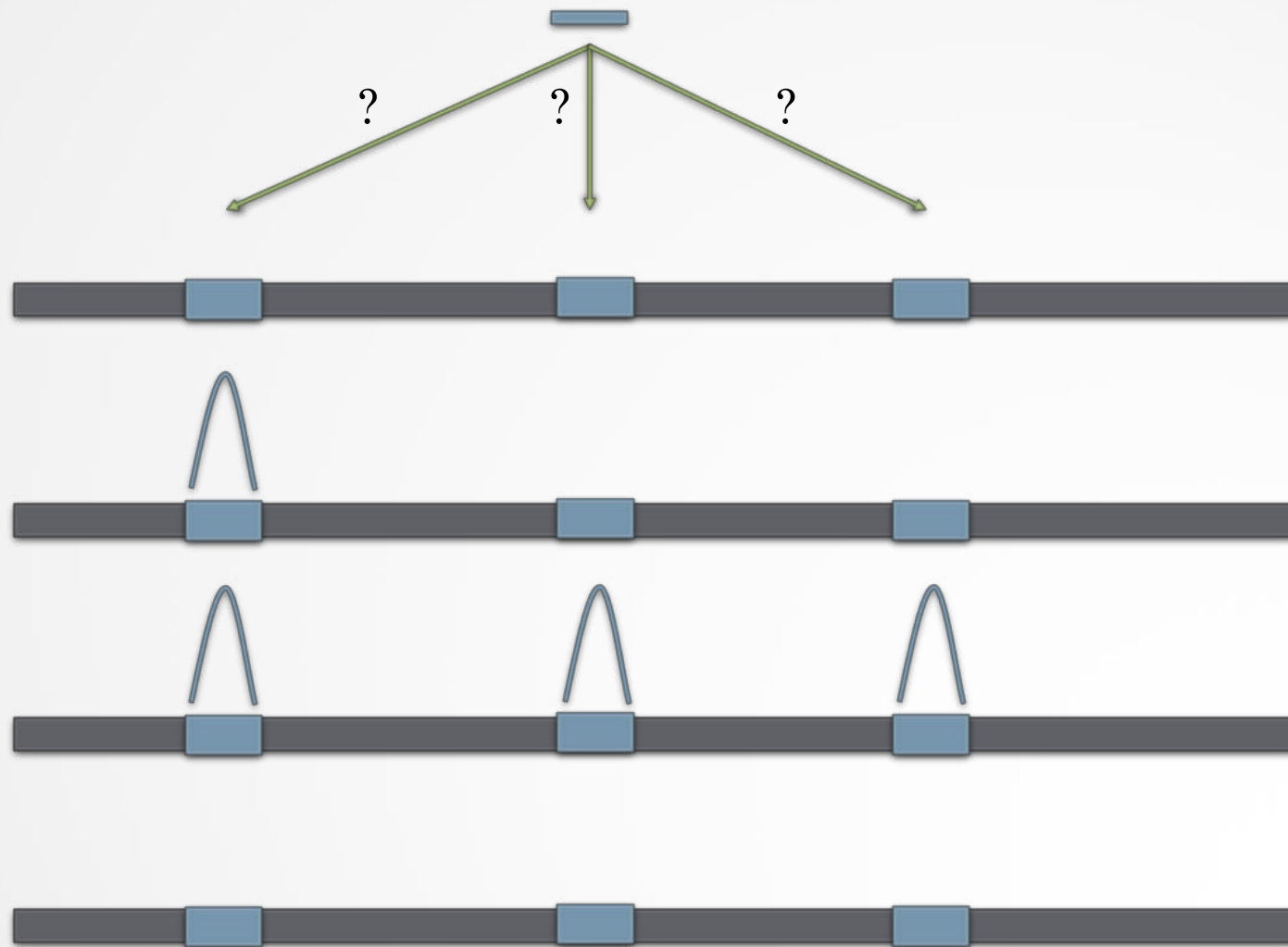


- One position in the genome
- Many possible positions  
(Repeat regions, duplicate regions, pseudogenes...)

# Mapping tool used: Bowtie

- Designed to align reads if:
  - many of the reads have at least one good, valid alignment,
  - many of the reads are relatively high-quality
  - the number of alignments reported per read is small (close to 1)
- Langmead B. et al, Genome Biology 2009
- Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11: Unit 11 17

# Duplicated genomic regions



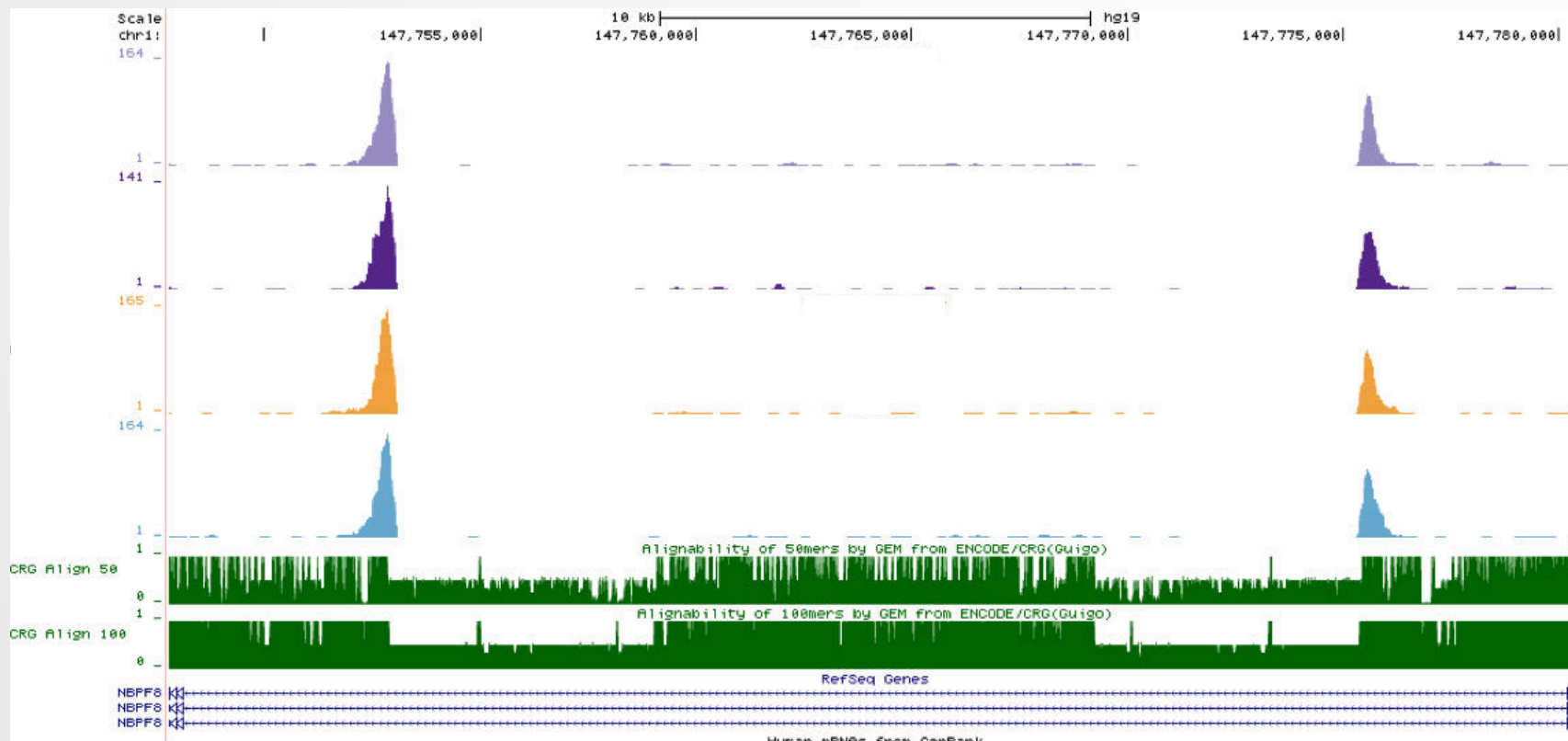
Keep 1 position  
randomly

Keep all possible  
position

Keep none

# Mappability

- Mappability (a): how many times a read of a given length can align at a given position in the genome
  - $a=1$  (read align once)
  - $a=1/n$  (read align n times)
  - Regions are empty or poorly covered if the mappability is low



# Exercise 1: mapping statistics

We used Bowtie 1 with the following parameters “-m 1 --strata --best” to align the reads. How many reads are aligned for each of the samples?

- 1. go to GalaxEast (<http://use.galaxeast.fr/>)
- 2. create a new history named “ChIP-seq data analysis”
- 3. import 2 BAM files (mitf.bam and ctrl.bam) from the data library NGS data analysis training > ChIPseq > mapping
- 4. use the tool **Flagstat** from the “NGS: Sam Tools” section to compute the number of aligned reads in the samples. The tools gives alignment statistics on a BAM file.

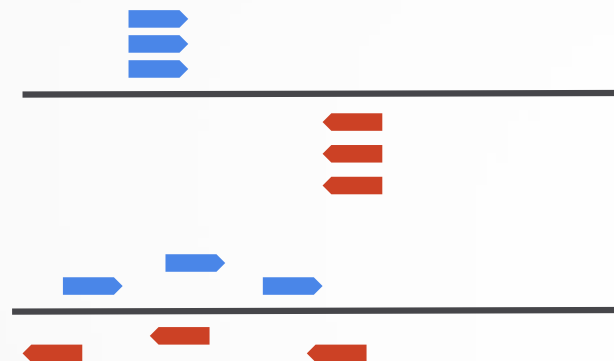


# PCR duplicates

- PCR duplicates
  - Related to poor library complexity
  - The same set of fragments are amplified
    - Indicates that Immuno-precipitation failed
  - Tools to check for
    - FastQC report (duplicate diagram)
    - PCR bottleneck metric (ENCODE)

# QC : PBC (PCR bottleneck coefficient)

- An approximate measure of library complexity
- $PBC = N1/Nd$ 
  - $N1$  = number of positions with EXACTLY 1 read mapped
  - $Nd$  = number of positions with 1 OR MORE reads mapped
- Value :
  - 0-0.5: severe bottlenecking (PCR bias, or a biological finding, such as a very rare genomic feature)
  - 0.5-0.8: moderate bottlenecking
  - 0.8-0.9: mild bottlenecking
  - 0.9-1.0: no bottlenecking (Control or IP with a good library complexity)



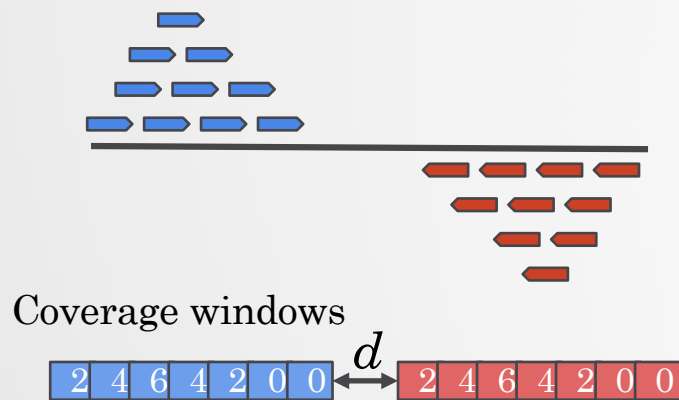
# Exercise 2: duplicate reads estimate

We want to assess the number of duplicate reads

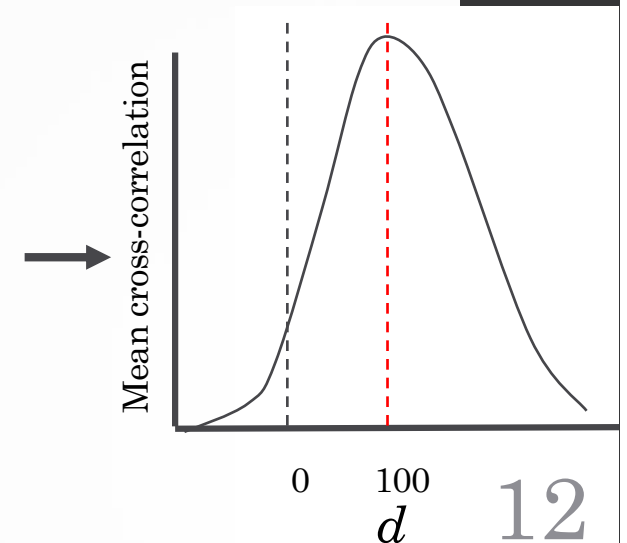
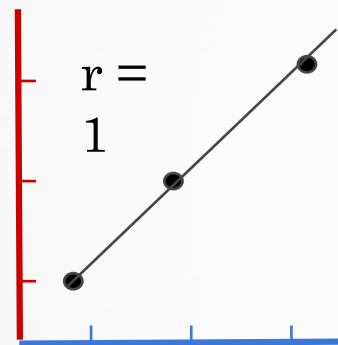
1. Use the tool **MarkDuplicates** to assess the complexity of the libraries (i.e the number of unique sequences).  
Use default parameters except for:
  - Select validation stringency: Silent (The picard tools validation strategy of BAM file is very stringent. So we turn off validation stringency)
  - The tool generates two datasets:
    - A log/metric file that contains statistics on the tool processing (number of input reads, number of duplicate reads)
    - A BAM file in which duplicated reads are flagged
  - Look at the log/metric file (in excel)

# QC: Strand cross-correlation

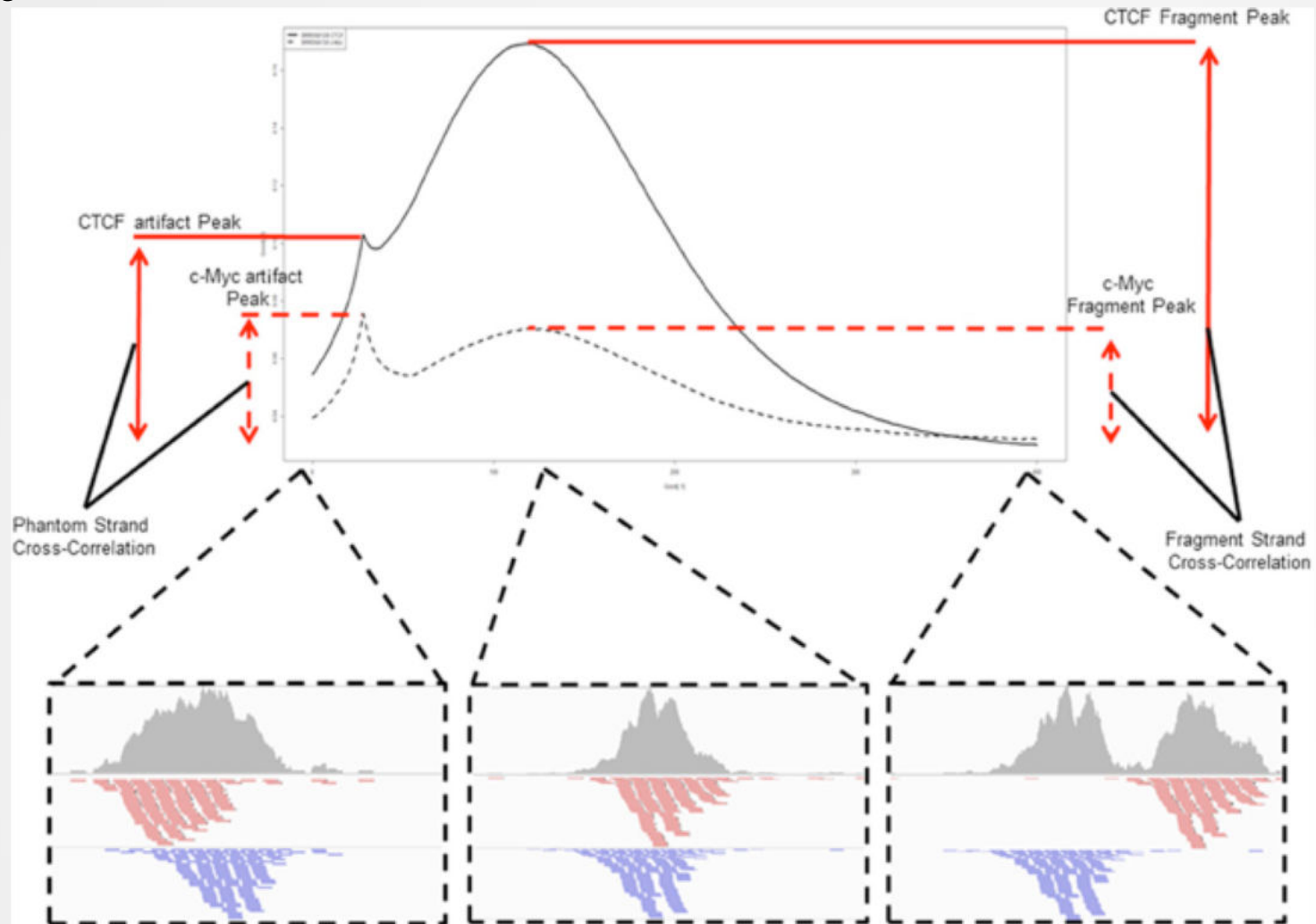
- Compute strand cross correlation for each window  $w$  across the genome.
- Use various distance  $d$  and compute the mean cross-correlation observed



Strand cross-correlation for each window and various  $d$  values

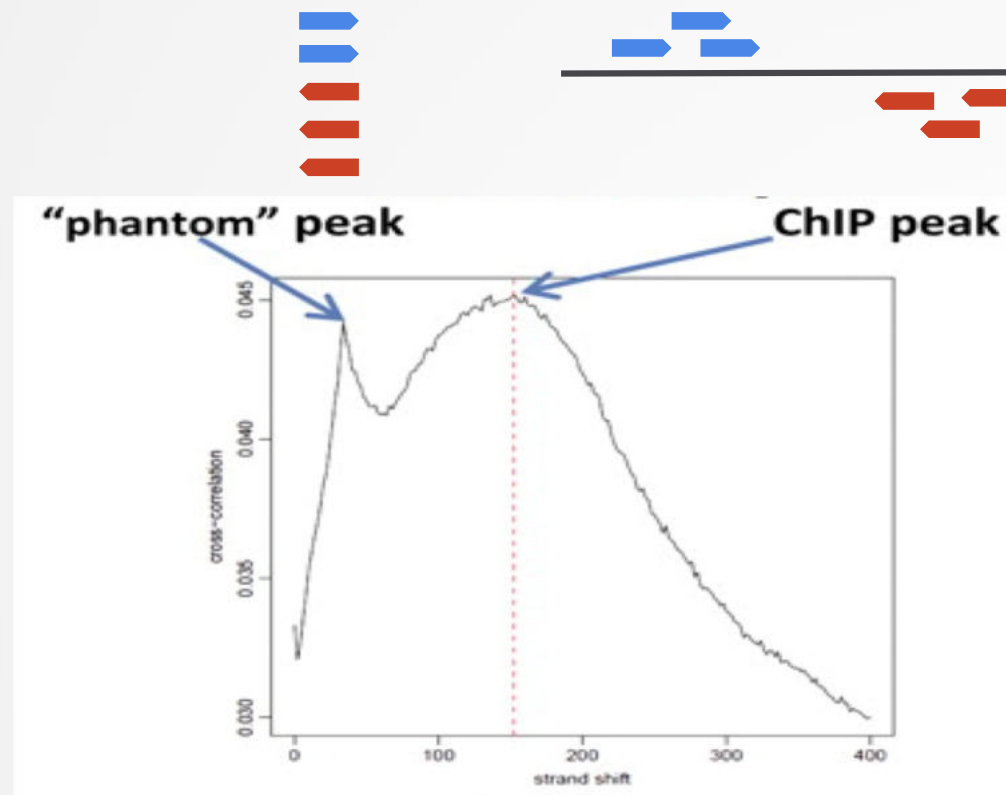


# QC: Strand cross-correlation

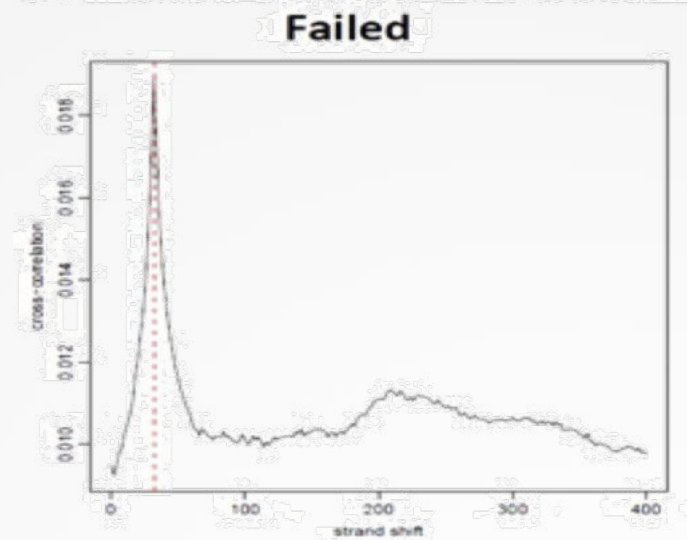
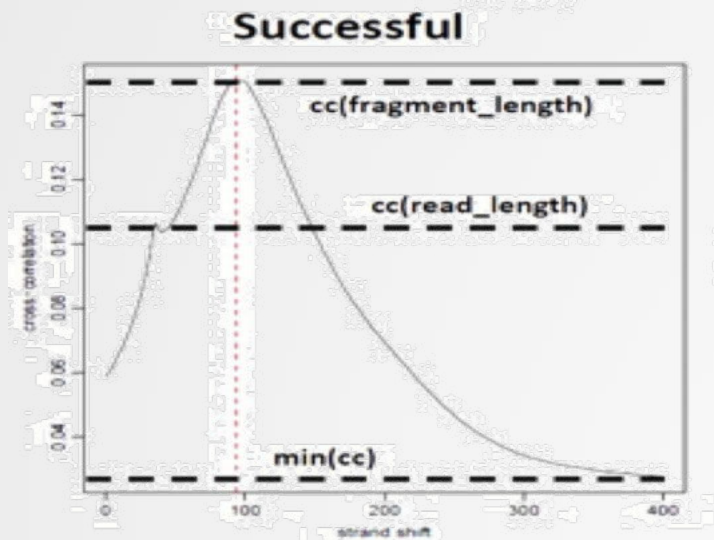


Carroll et al, *Frontiers in genetics*, 2014

# QC: Strand cross-correlation



# QC: Strand cross-correlation



NSC: normalized strand coefficient

$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

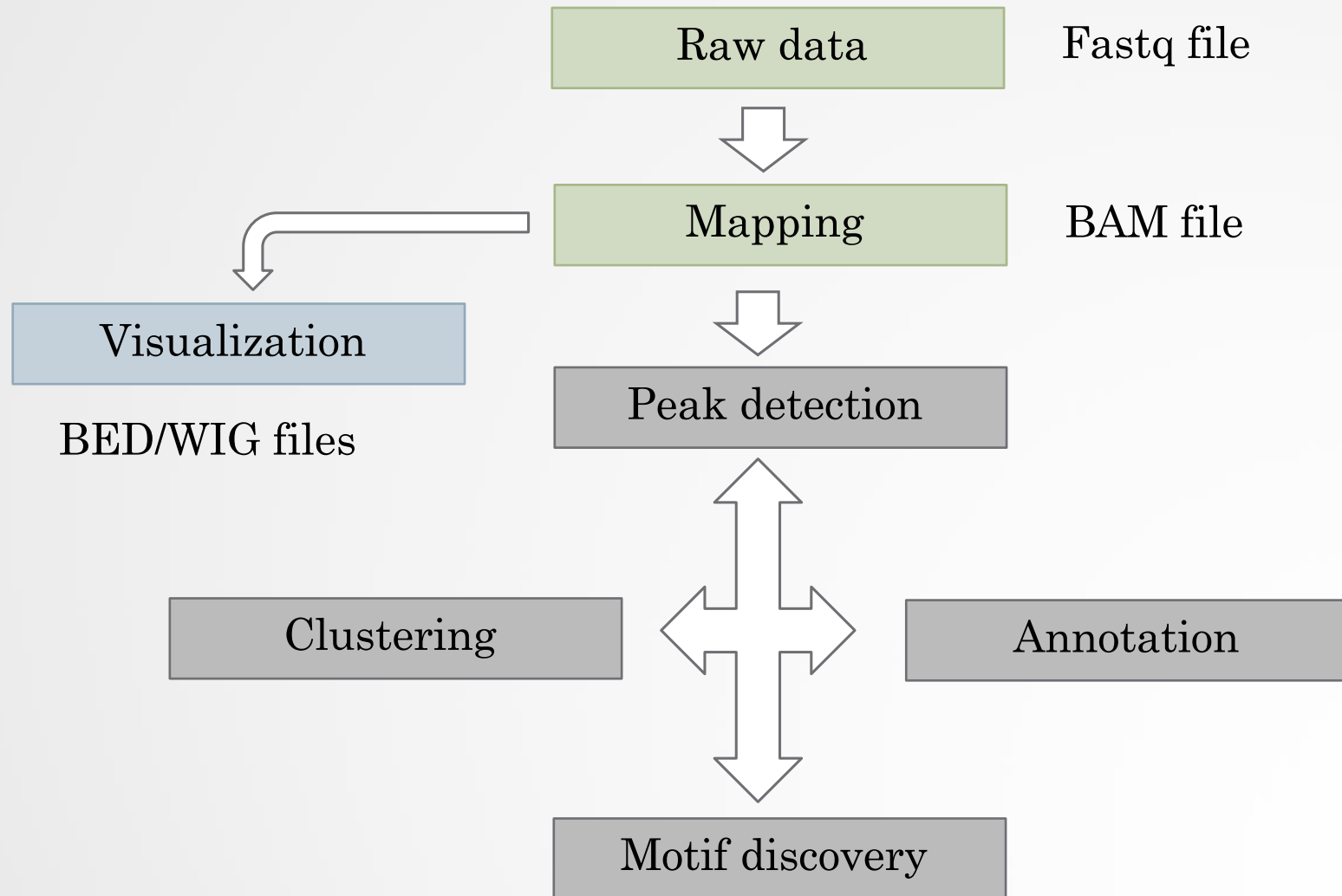
NSC  $\geq$  1.05 is recommended

Relative strand correlation (RSC)

$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

RSC  $\geq$  0.8 is recommended

# Guidelines



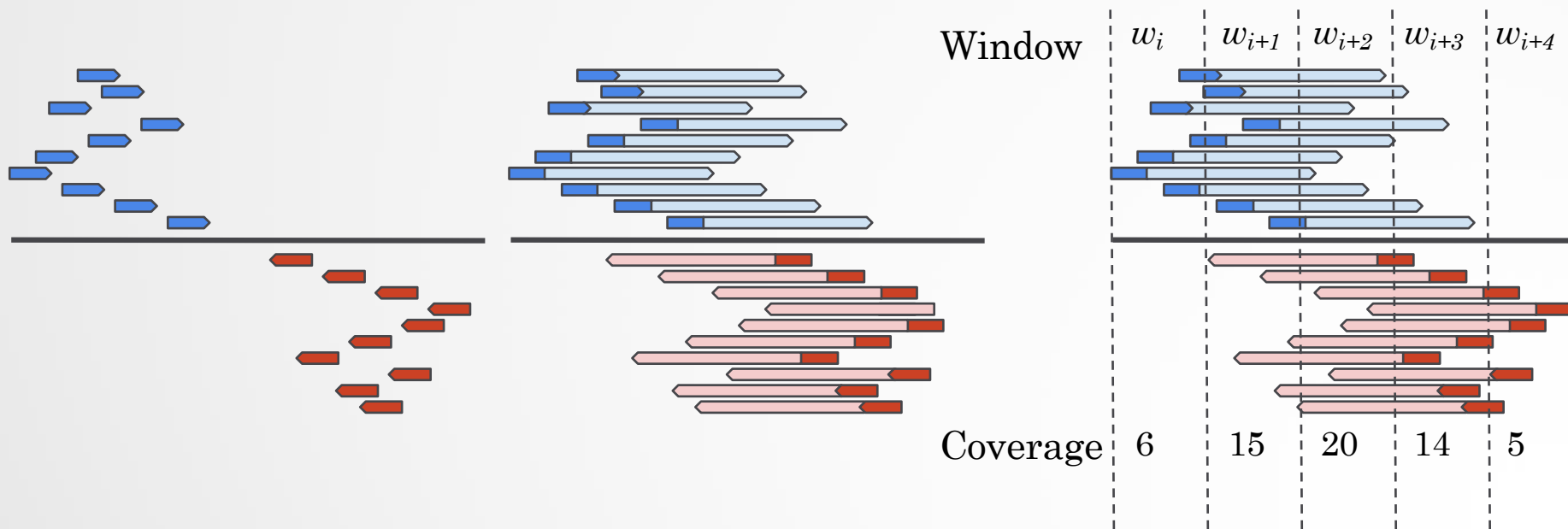


# Bam files are fat

- **BAM files are fat** as they do contain exhaustive information about read alignments.
  - Memory issues (can only visualize fraction of the BAM).
- Need a more **lightweight file format containing only genomic coverage information**:
  - **Wig (not compressed, not indexed)**
  - **TDF (compressed, indexed)**
  - **BigWig (compressed, indexed)**

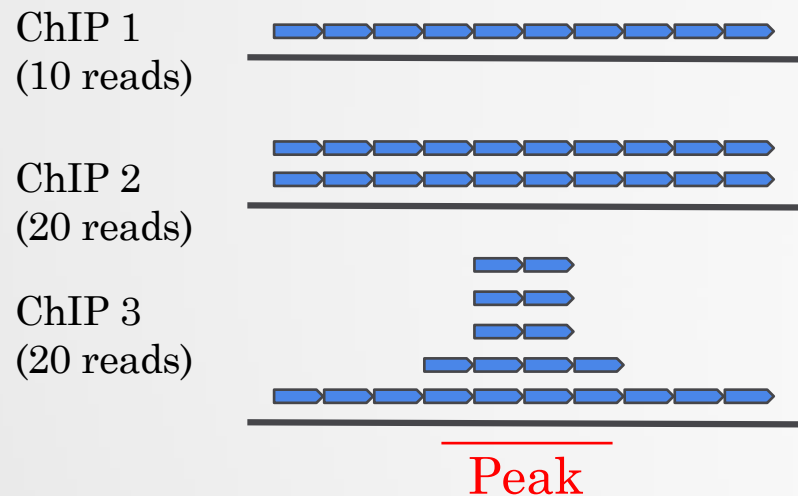
# Coverage file and read extension

- BAM files **do not contain fragment location** but read location
- We need to extend reads to compute fragments coordinates before coverage analysis
- Not required for PE



# Library size normalization

- **Signal need to be normalized**
  - E.g. Normalize coverage to 1x
    - Popular but not optimal



Already normalized to 1x  
coverage

Should be decreased by 2 fold to  
get 1x coverage

Decreasing by 2 fold would  
underestimate peak signal. Problem

...

# Exercise 3: Visualization of the data

## 1. Upload the two tdf files in IGV

You can find them in the directory `chipseq > visualization`

Tip1: They have been generated using IGVtools from the bam files

Tip2: Check that Normalize coverage data (.tdf files only) is selected in `View > Preferences... > Tracks`

Tip3: Select the two datasets, click right on them and select Group Autoscale

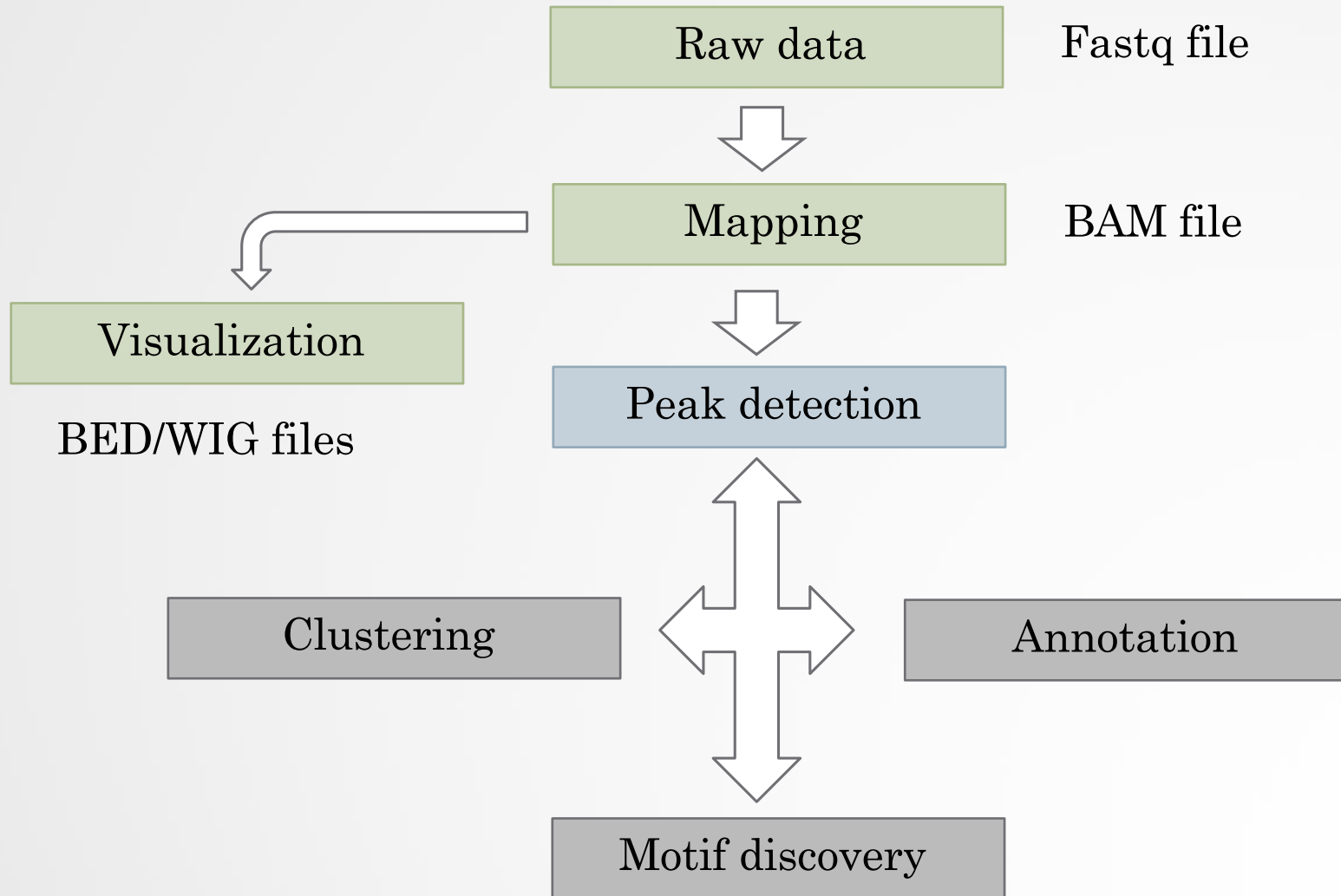
## 2. Go check the following genes:

- Idh1, Eef2, AP1S2, PABPC11, Park7, Pmel, Cdk2, Actb

Do you see peaks at these locations?

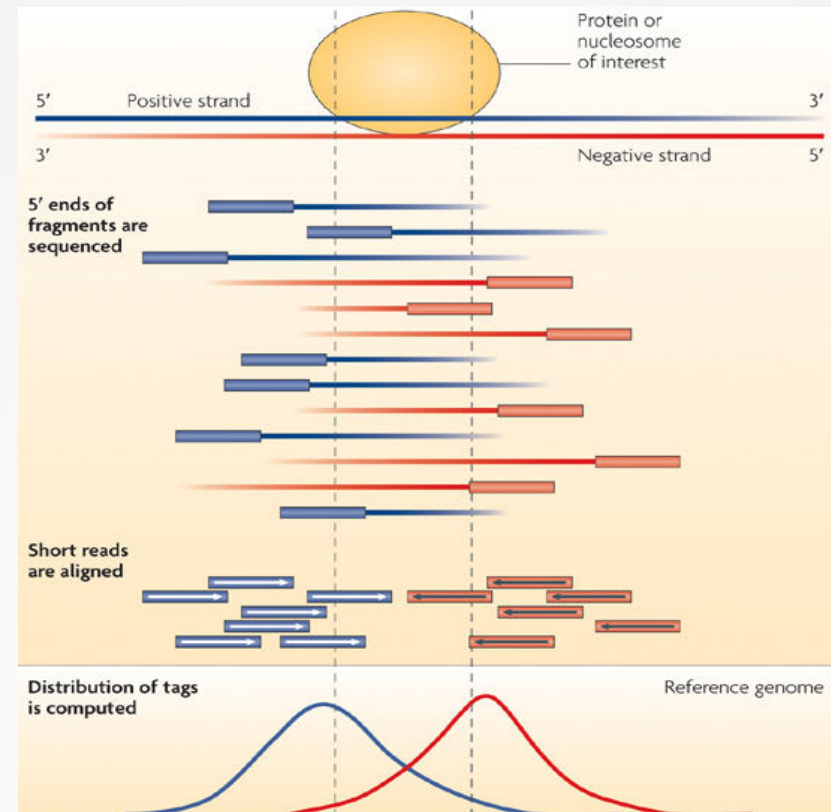
# Peak Calling

# Guidelines



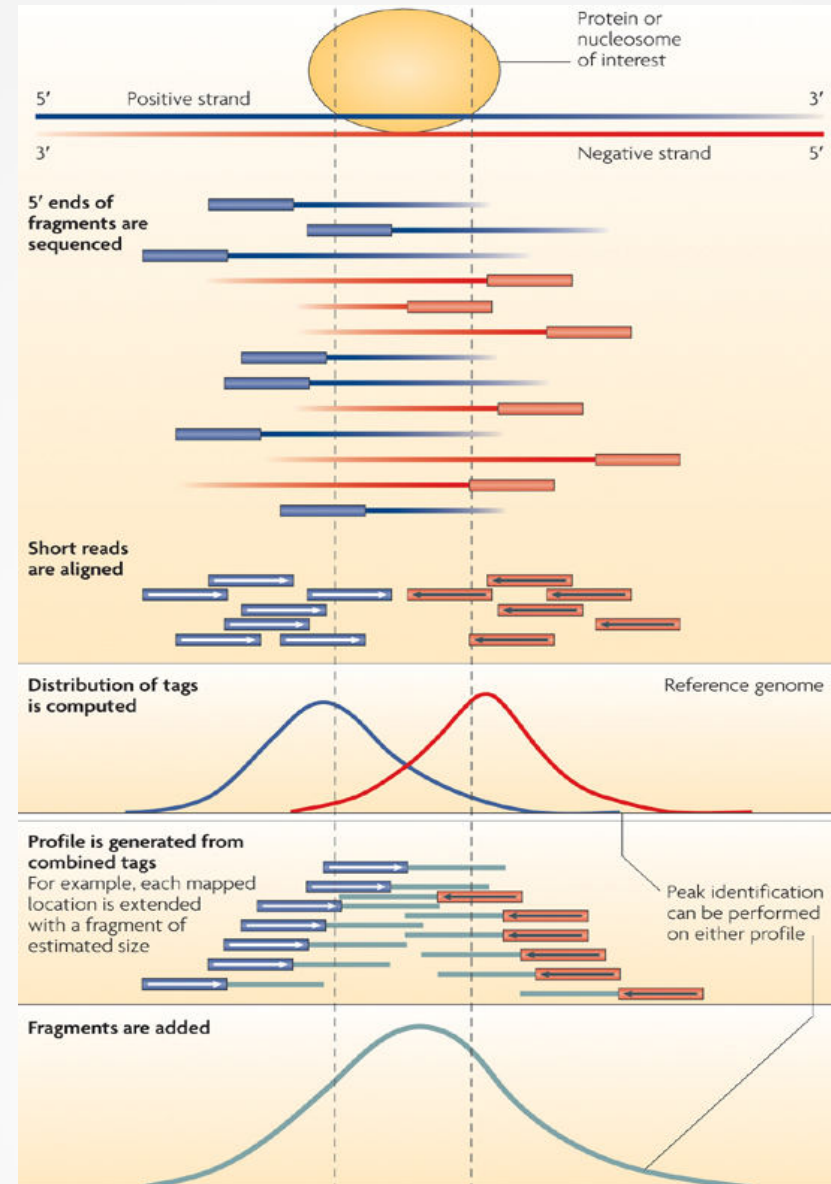
# From reads to peaks

- Chip-seq peaks are a mixture of two signals:
  - + strand reads (Watson)
  - - strand reads (Crick)
- The sequence tag density accumulates on forward and reverse strands centered around the binding site



# From reads to peaks

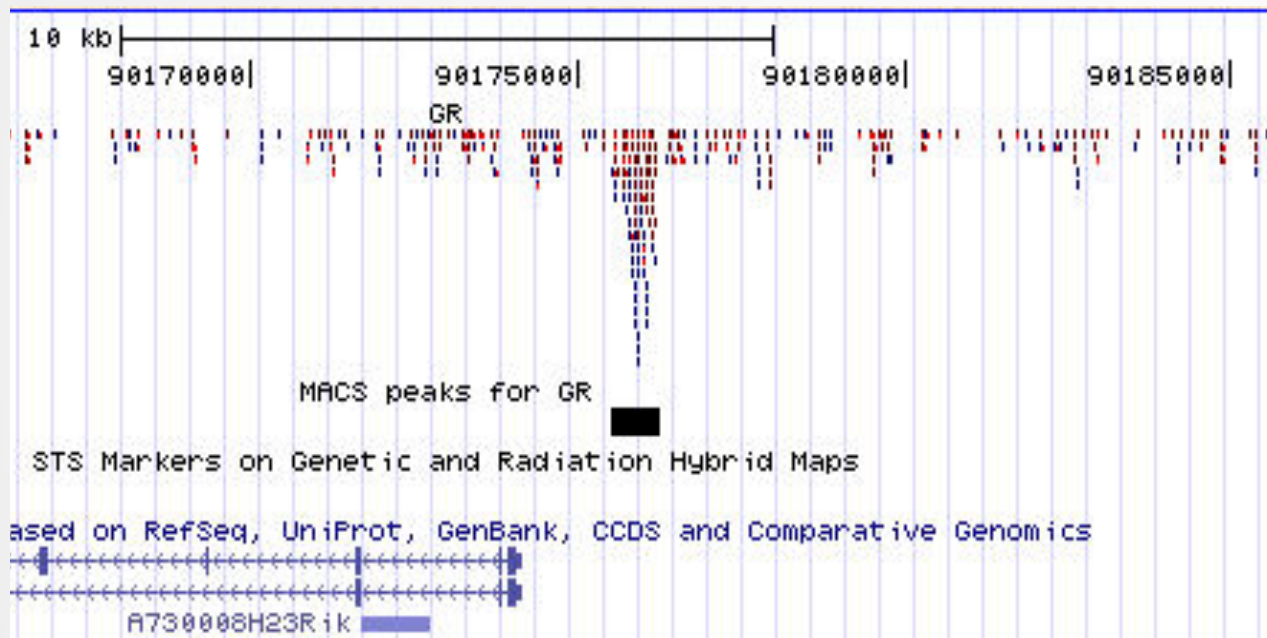
- Get the signal at the right position
  - Read shift
  - Extension
- Estimate the fragment size
- Do paired-end





# Peak detection

- Discover interaction sites from aligned reads
- Idea: loci with a lot of reads/fragments = signal site



# A variety of peak callers

- 60 programs listed on OMICTOOLS
- Most support a control



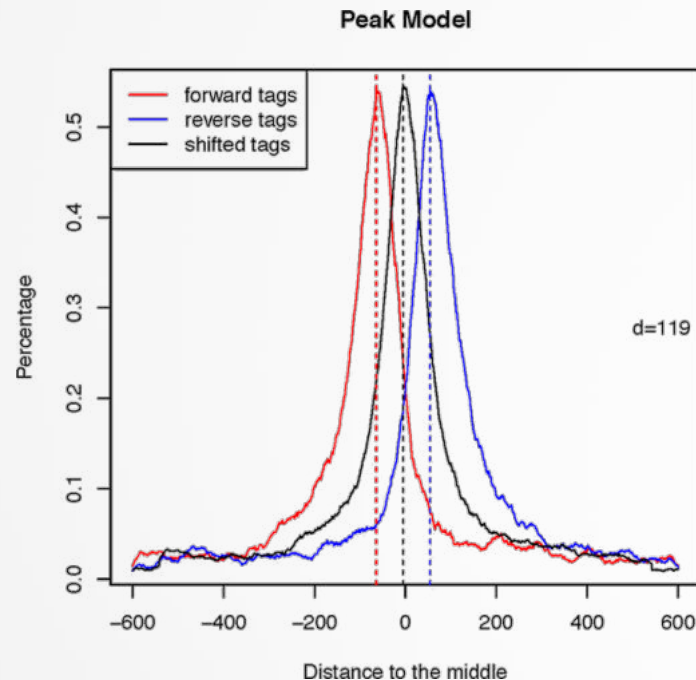
The screenshot displays the OMICTOOLS website interface. At the top, there is a search bar with the text "Find the best of bioinformatics". Below the search bar, a navigation menu shows "HIGH-THROUGHPUT SEQUENCING > CHIP-SEQ ANALYSIS > PEAK CALLING". The main heading is "PEAK CALLING SOFTWARE TOOLS | CHIP SEQUENCING DATA ANALYSIS". A brief description follows: "Identification of genomic regions of interest in CHIP-seq data, commonly referred to as peak-calling, aims to find the locations of transcription factor binding sites, modified histones or nucleosomes. Source text:(Cairns et al., 2011) BayesPeak-an... Read more". A "FILTERS" button is visible. The list of tools includes:

- MACS** / Model-based Analysis for CHIP-Seq: A software to analyze data generated by short read sequencers. MACS empirically models the shift size of CHIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. It... (1 star, 1 discussion, 4 favorites)
- HOMER** / Hypergeometric Optimization of Motif EnRichment: A suite of tools for Motif Discovery and next-gen sequencing analysis. HOMER contains many useful tools for analyzing CHIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of... (5 stars, 0 discussions, 4 favorites)
- SICER**: A clustering approach for identification of enriched domains from histone modification CHIP-Seq data. (1 star, 0 discussions, 1 favorite)
- SPP**: An R package for analysis of CHIP-seq and other functional sequencing data. SPP has been designed to detect protein binding positions with high accuracy. SPP can also examine the saturation level of... (0 stars, 0 discussions)
- Scripture**: A method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome ab initio. The statistical methods to estimate read coverage... (0 stars, 0 discussions)

# MACS [Zhang et al, 2008]

## 1. Modeling the shift size of ChIP-Seq tags

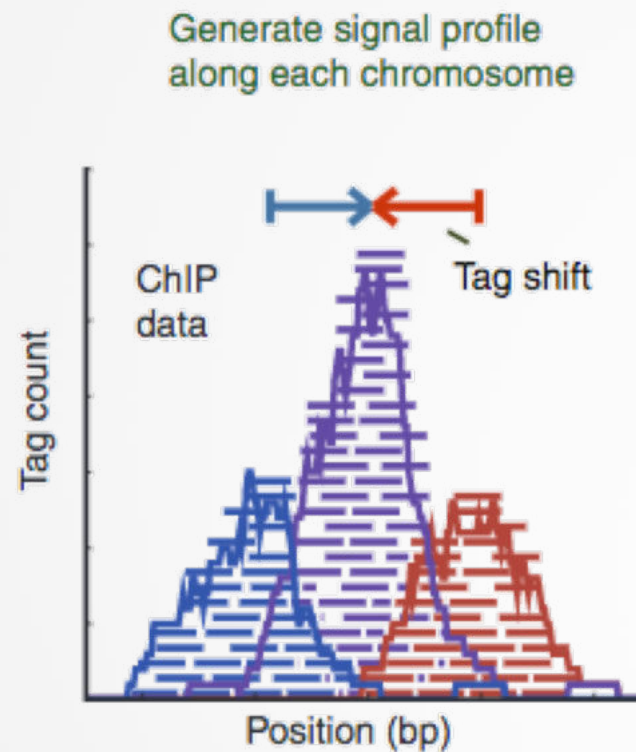
- slides  $2bandwidth$  windows across the genome to find regions with tags more than  $mfold$  enriched relative to a random tag genome distribution
- randomly samples 1,000 of these highly enriched peaks
- separates their Watson and Crick tags, and aligns them by the midpoint between their Watson and Crick tag centers
- define  $d$  as the distance in bp between the summit of the two distributions



# MACS [Zhang et al, 2008]

## • 2. Peak detection

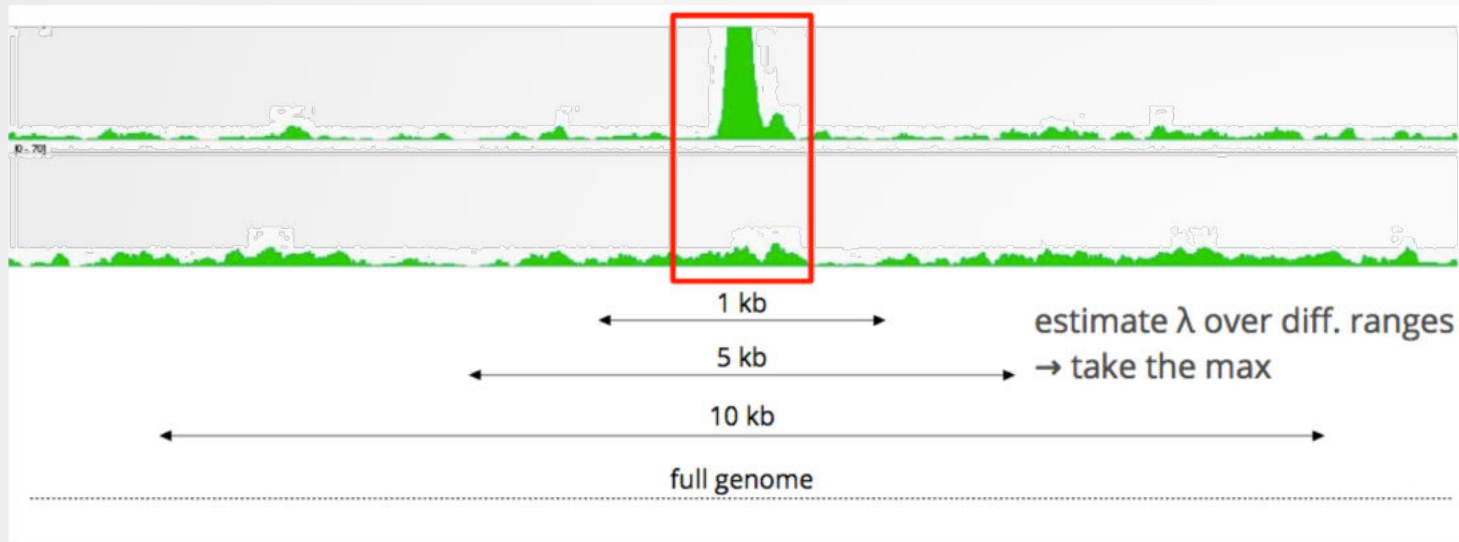
- Normalization: linearly scales the total control read count to be the same as the total ChIP read count
- Duplicate read removal
- Tags are shifted by  $d/2$



Pepke et al, 2009

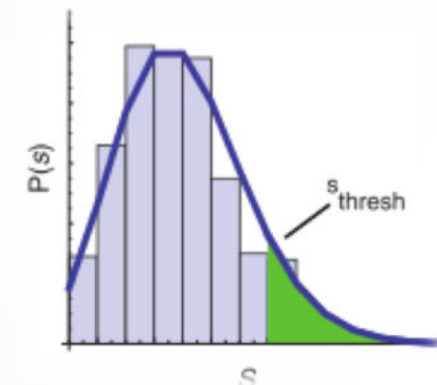
# MACS [Zhang et al, 2008]

- Slides 2d windows across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution  $p$ -value based on  $\lambda_{BG}$ , default  $10^{-5}$ )
- Estimate parameter  $\lambda_{local}$  of Poisson distribution



Source:  
C. Herrmann


- Keep peaks significant under  $\lambda_{BG}$  and  $\lambda_{local}$  and with  $p$ -value  $<$  threshold



# MACS [Zhang et al, 2008]

## 3. Multiple testing correction (FDR)

- Swap treatment and input and call negative peaks
- Take all the peaks (neg + pos) and sort them by increasing p-values

$$\text{FDR}(p) = \frac{\# \text{ Negative peaks with p-value} < p}{\# \text{ Selected peaks}}$$


FDR = 2/27 = 0.074

# Exercise 4: peak calling

We now want to call MITF peaks.

- 1. Use **Macs2 callpeak** to perform the peak calling on the data. Use default parameters except for
  - ChIP-Seq Treatment File: mitf.bam
  - ChIP-Seq Control File: ctrl.bam
  - Effective genome size: Human
  - Outputs:
    - Peaks as tabular file,
    - summits,
    - Summary page (html),
    - Plot in PDF

# Exercise 4: peak calling

- 2. Macs2 callpeak generates 5 datasets:
  - List of the peaks (tabular format)

List of arguments  
used to run Macs2

	A	B	C	D	E	F	G	H	I	J
1	# This file is generated by MACS version 2.1.0.20151222									
2	# Command line: callpeak --name MACS2 -t /galaxy13/files/052/dataset_52866.dat -c /galaxy22/files/052/dataset_52865.dat --fr									
3	# ARGUMENTS LIST:									
4	# name = MACS2									
5	# format = BAM									
6	# ChIP-seq file = ['/galaxy13/files/052/dataset_52866.dat']									
7	# control file = ['/galaxy22/files/052/dataset_52865.dat']									
8	# effective genome size = 2.45e+09									
9	# band width = 300									
10	# model fold = [5, 50]									
11	# qvalue cutoff = 5.00e-02									
12	# Larger dataset will be scaled towards smaller dataset.									
13	# Range for calculating regional lambda is: 1000 bps and 10000 bps									
14	# Broad region calling is off									
15	# tag size is determined as 54 bps									
16	# total tags in treatment: 23124393									
17	# tags after filtering in treatment: 6223075									
18	# maximum duplicate tags at the same position in treatment = 1									
19	# Redundant rate in treatment: 0.73									
20	# total tags in control: 19949607									
21	# tags after filtering in control: 4798380									
22	# maximum duplicate tags at the same position in control = 1									
23	# Redundant rate in control: 0.76									
24	# d = 75									
25	# alternative fragment length(s) may be 75 bps									
26	chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
27	chr1	980686	980816	131	980745	8.48	10.38277	7.29361	6.46786	MACS2_peak_1
28	chr1	983821	983925	105	983877	6.94	9.11038	6.77148	5.34984	MACS2_peak_2
29	chr1	1031344	1031475	132	1031406	6.17	6.82634	5.21345	3.25879	MACS2_peak_3
30	chr1	1079424	1079564	141	1079490	12.34	18.30659	10.88735	13.88358	MACS2_peak_4
31	chr1	1304817	1304958	142	1304891	13.11	20.10101	11.51679	15.56374	MACS2_peak_5

Peaks



# Exercise 4: peak calling

- 2. Macs2 callpeak generates 5 datasets:
  - List of the peaks (tabular format)

26	chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
27	chr1	980686	980816	131	980745	8.48	10.38277	7.29361	6.46786	MACS2_peak_1
28	chr1	983821	983925	105	983877	6.94	9.11038	6.77148	5.34984	MACS2_peak_2
29	chr1	1031344	1031475	132	1031406	6.17	6.82634	5.21345	3.25879	MACS2_peak_3
30	chr1	1079424	1079564	141	1079490	12.34	18.30659	10.88735	13.88358	MACS2_peak_4
31	chr1	1304817	1304958	142	1304891	13.11	20.10101	11.51679	15.56374	MACS2_peak_5

- chr: chromosome name
- start: start position of peak
- end: end position of peak
- length: length of peak region
- abs\_summit: absolute peak summit position
- pileup: pileup height at peak summit
- -log10(pvalue): -log10(pvalue) for the peak summit (e.g. pvalue =1e-10, then this value should be 10)
- fold\_enrichment: fold enrichment for this peak summit against random Poisson distribution with local lambda
- -log10(qvalue): -log10(qvalue) at peak summit
- name: peak name

# Exercise 4: peak calling

- List of the peaks (Narrowpeak format)

1	2	3	4	5	6	7	8	9	10
chr1	980685	980816	MACS2_peak_1	64	.	7.29361	10.38277	6.46786	59
chr1	983820	983925	MACS2_peak_2	53	.	6.77148	9.11038	5.34984	56
chr1	1031343	1031475	MACS2_peak_3	32	.	5.21345	6.82634	3.25879	62
chr1	1079423	1079564	MACS2_peak_4	138	.	10.88735	18.30659	13.88358	66
chr1	1304816	1304958	MACS2_peak_5	155	.	11.51679	20.10101	15.56374	74
chr1	1441082	1441181	MACS2_peak_6	124	.	10.25923	16.71260	12.40068	71

1. chr

2. Start of peak

3. End of peak

4. Peak name

5. Integer score for display

7. fold-change

8.  $-\log_{10}p$ value

9.  $-\log_{10}q$ value

10. Relative summit position to peak start

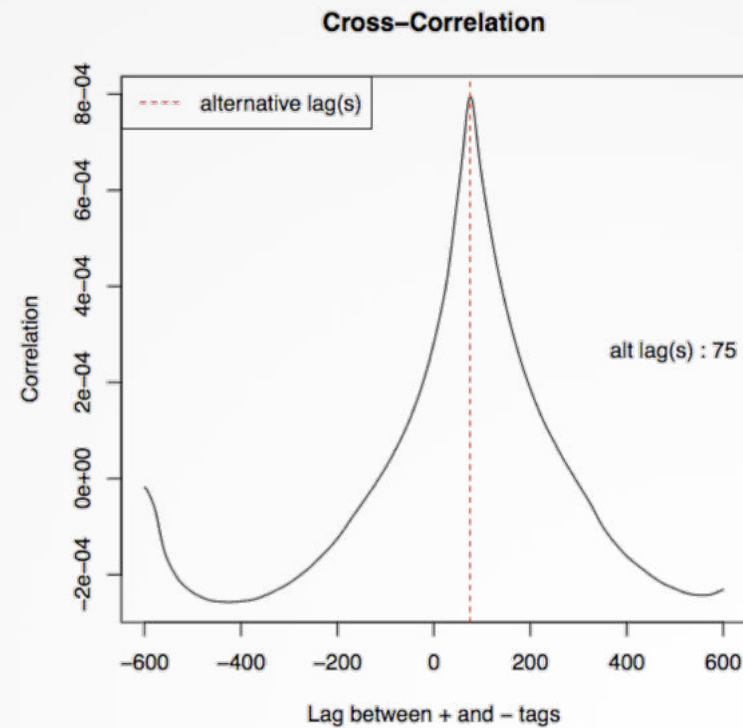
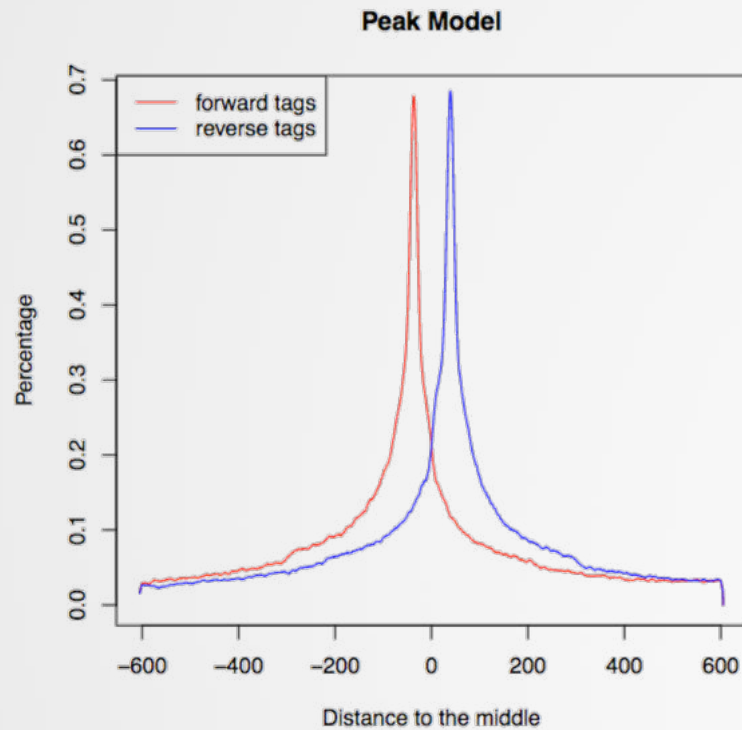
# Exercise 4: peak calling

- List of the peak summits (BED): contains the peak summit location for each peak.

1. chr	2. Start of peak	3. End of peak	4. Peak name	5. $-\log_{10}p$ value
1	2	3	4	5
chr1	980744	980745	MACS2_peak_1	6.46786
chr1	983876	983877	MACS2_peak_2	5.34984
chr1	1031405	1031406	MACS2_peak_3	3.25879
chr1	1079489	1079490	MACS2_peak_4	13.88358
chr1	1304890	1304891	MACS2_peak_5	15.56374
chr1	1441153	1441154	MACS2_peak_6	12.40068

# Exercise 4: peak calling

- PDF images about the model based on your data



- Log of MACS - output during Macs2 run (HTML)

# Exercise 5: peak calling

We now want to call MITF peaks.

- 1. Use **Macs2 callpeak** to perform the peak calling on the data. Use default parameters except for
  - CHIP-Seq Treatment File: mitf.bam
  - CHIP-Seq Control File: ctrl.bam
  - Effective genome size: Human
  - Outputs: Peaks as tabular file, summits, Summary page (html), Plot in PDF
- 2. Look at the resulting datasets. How many peaks are found?
- 3. What is the fragment size estimated by Macs2? What do you think of the value?
- 4. Rerun **Macs2** using the same parameters as before but changing the shift size:
  - Build Model: Do not build the shifting model (--nomodel)
  - The arbitrary extension size in bp: 200
- 5. How many peaks are now found?

# Comparing the two MACS runs

- Check out PLEKHF2



# Comparing the two MACS runs

- Check out ASAP1



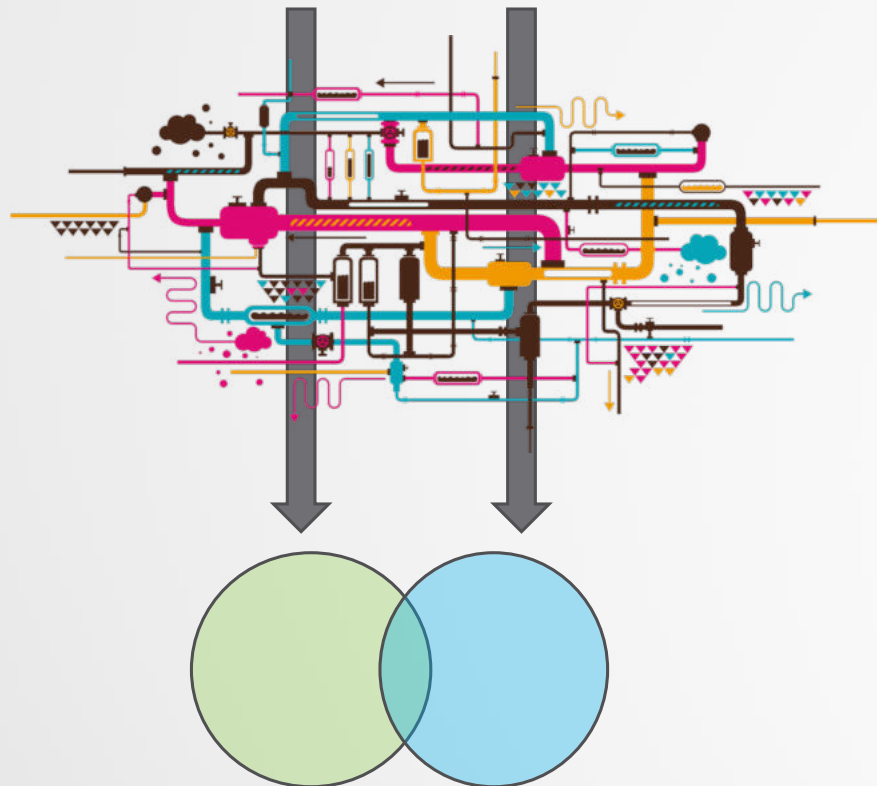
# How to deal with replicates

Analyze samples separately  
and takes union or  
intersection of resulting peaks

Merge samples prior to the  
peak calling

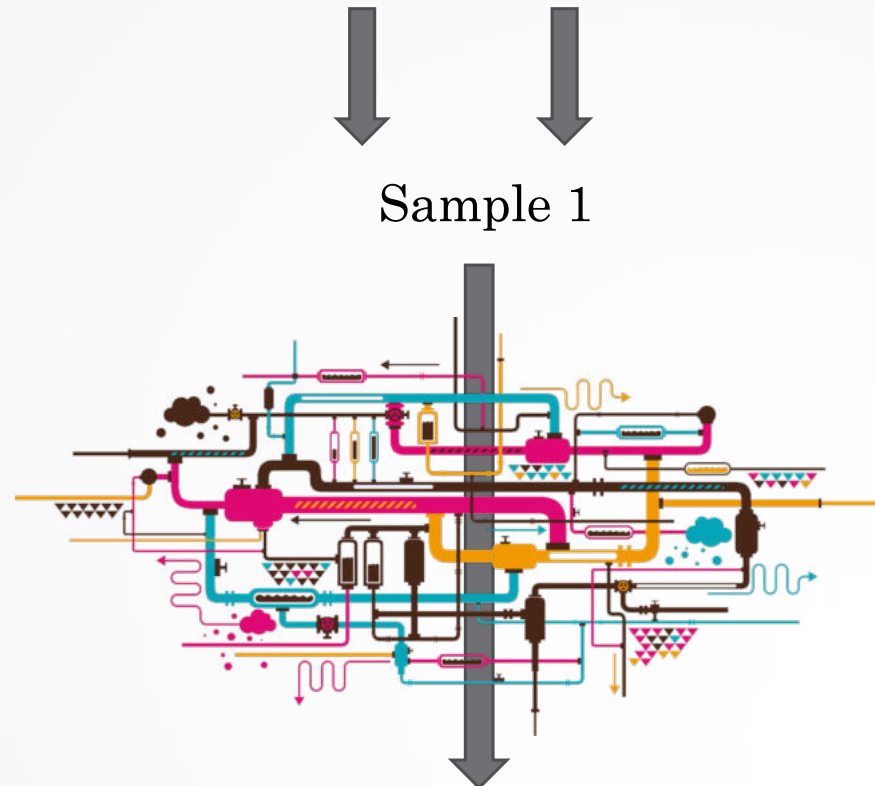
Sample 1.a

Sample 1.b



Sample 1.a

Sample 1.b



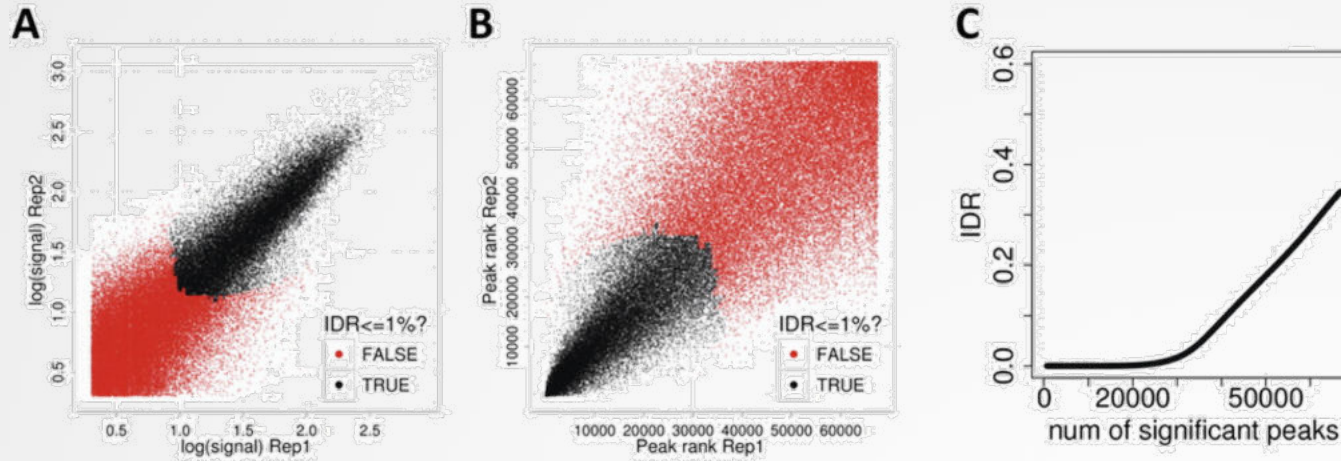


# IDR

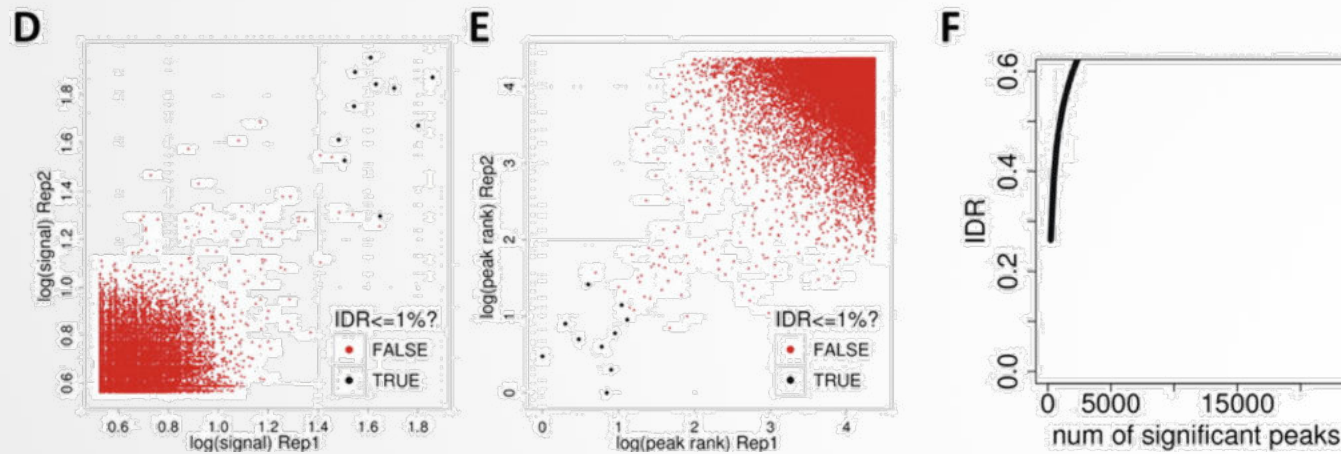
- Measures consistency between replicates
- Uses reproducibility in score rankings between peaks in each replicate to determine an optimal cutoff for significance.
- Idea:
  - The most significant peaks are expected to have high consistency between replicates
  - The peaks with low significance are expected to have low consistency

# IDR

## RAD21 Replicates (high reproducibility)



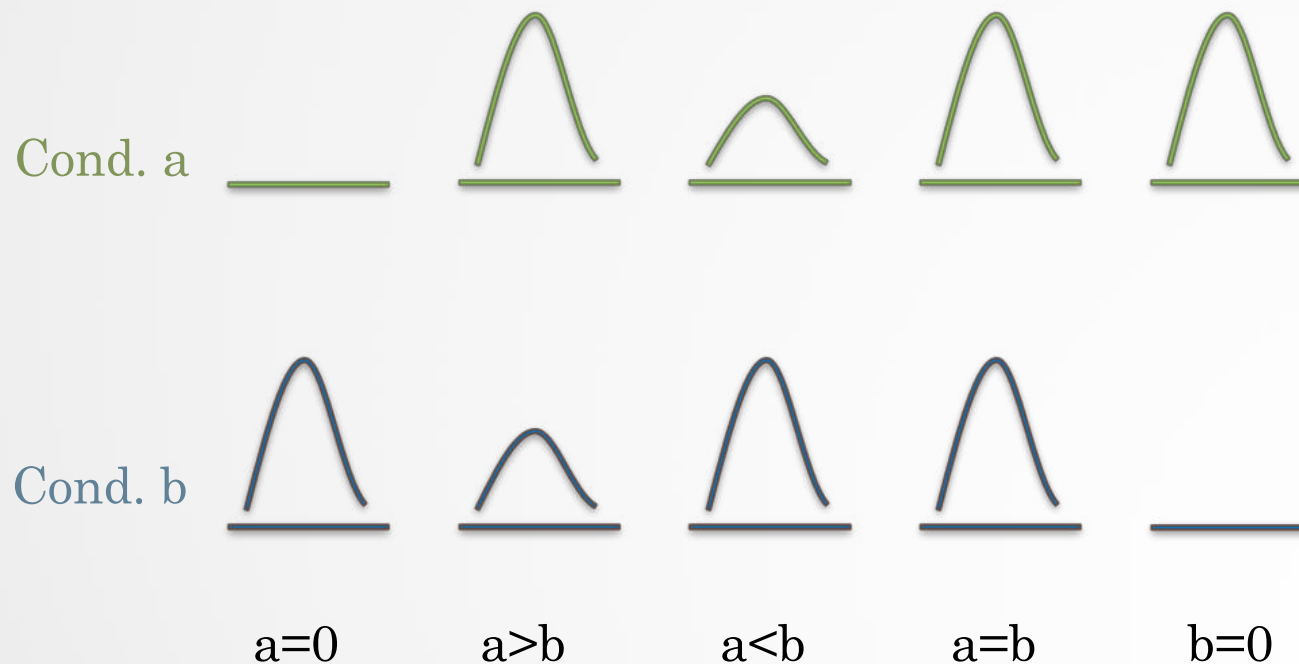
## SPT20 Replicates (low reproducibility)



(!) IDR doesn't work on broad source data!

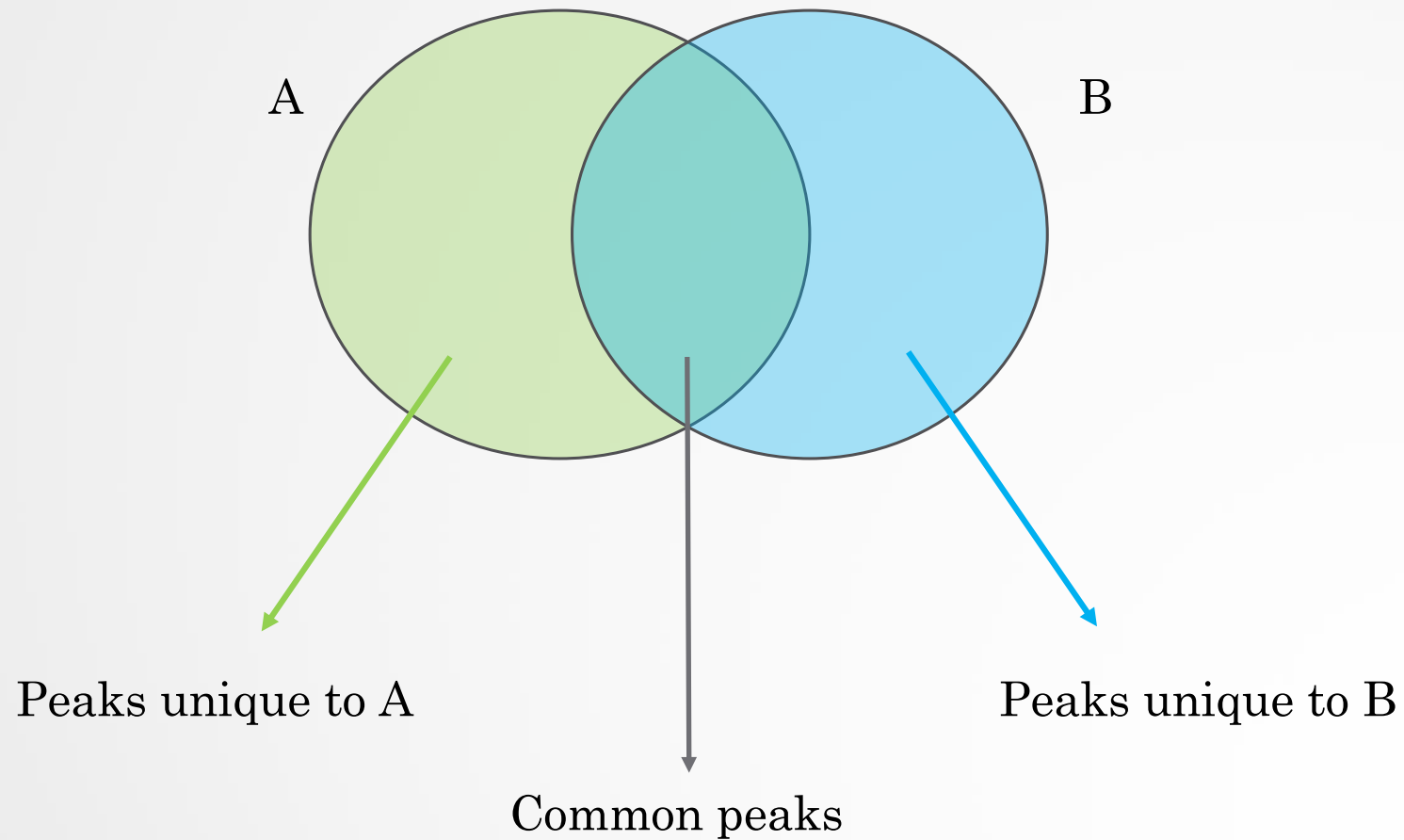
# Differential binding analysis

- Find differential binding events by comparing different conditions
  - qualitative analysis: binding vs no binding
  - quantitative analysis: weak binding vs strong binding



# Differential binding analysis

Qualitative approach



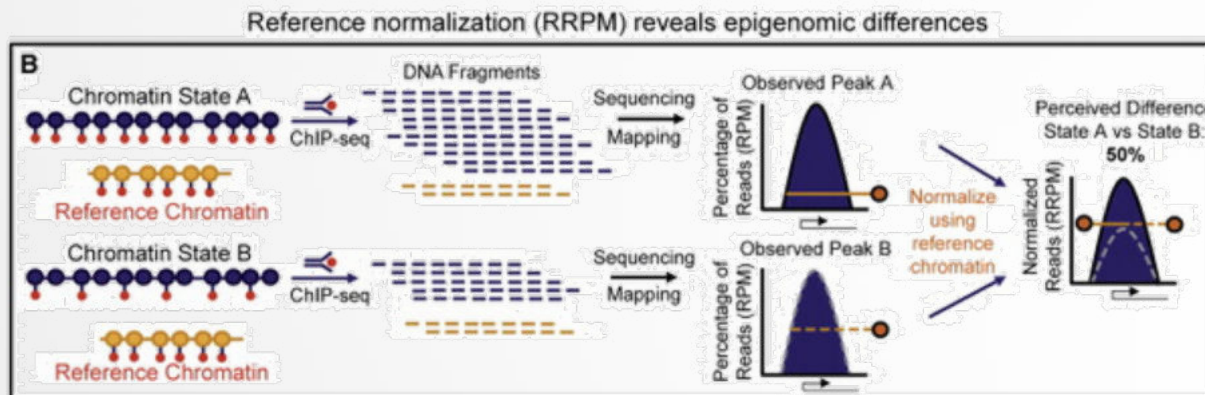
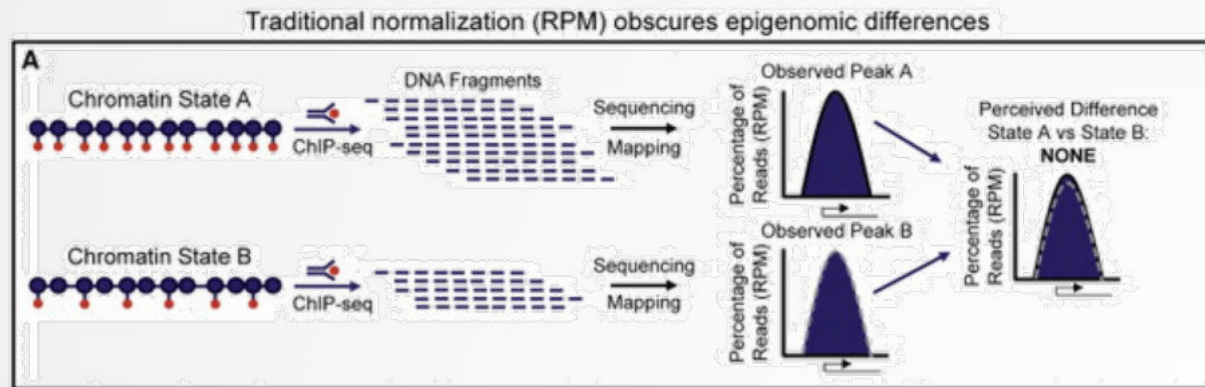
# Differential binding analysis

## Quantitative approach

- Do the peak calling on all data
- Take union of all peaks
- Do quantitative analysis of differential binding events based on read counts
- Statistical models
  - No replicates: assume simple Poisson model
  - With replicates: perform differential test using DE tools from RNA-seq (EdgeR, DESeq,...) based on read counts

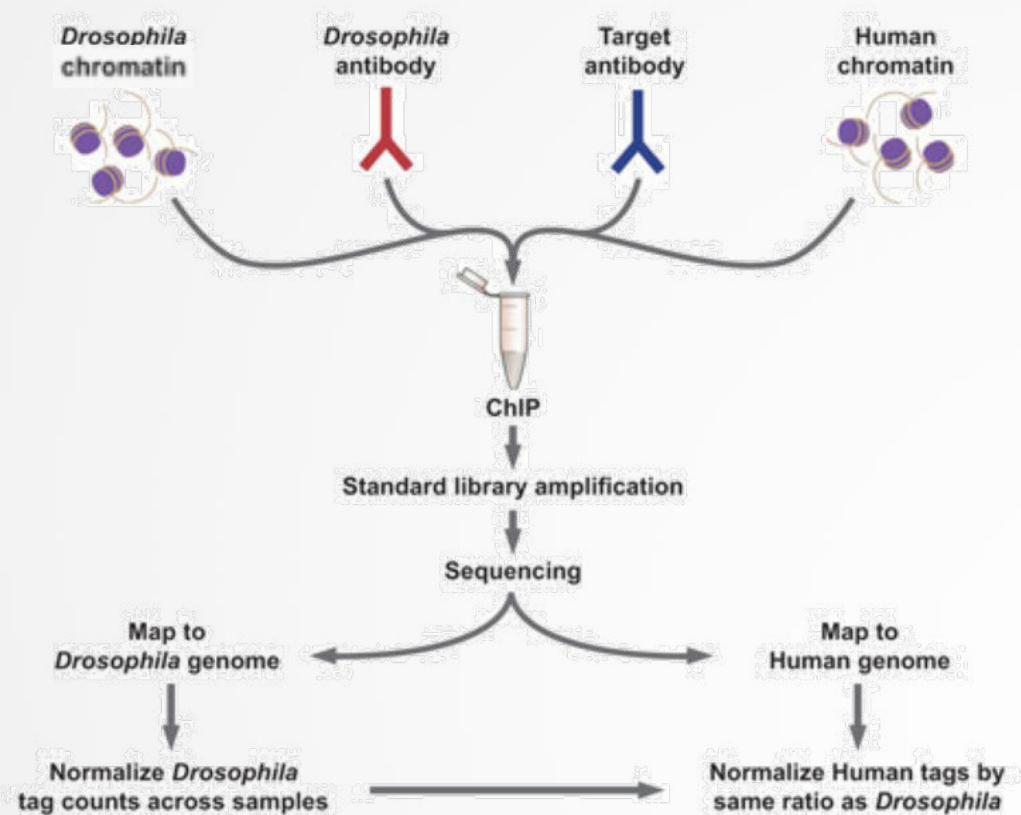
# Spike-in

- Current normalization methods fail to detect global changes as they make the assumption that globally nothing change but a small portion of the genome
- Insert external chromatin used as reference chromatin

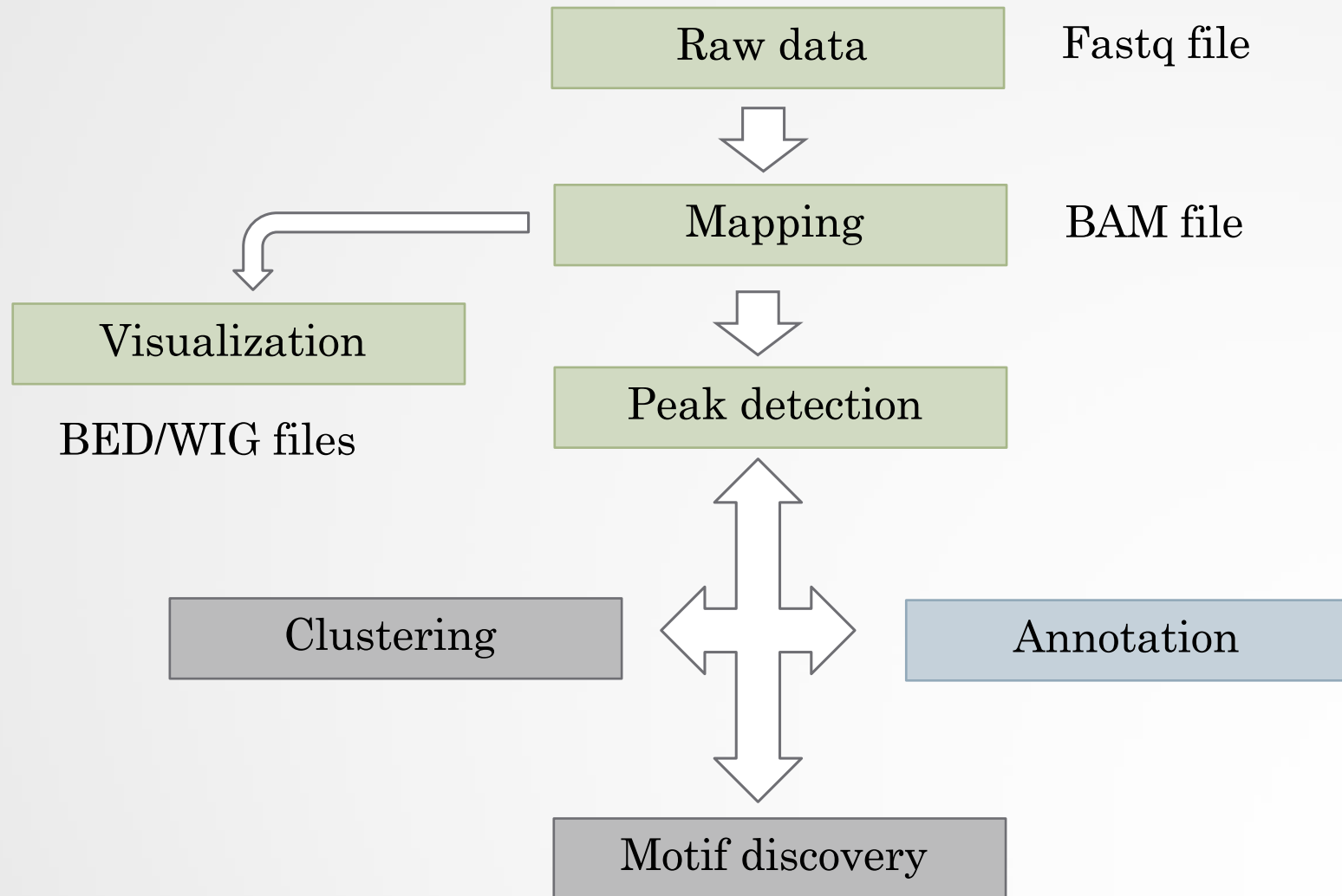


# Spike-in

- Spike-in normalization can be applied to ChIP-Seq data to reduce the effects of technical variation and sample processing bias



# Guidelines





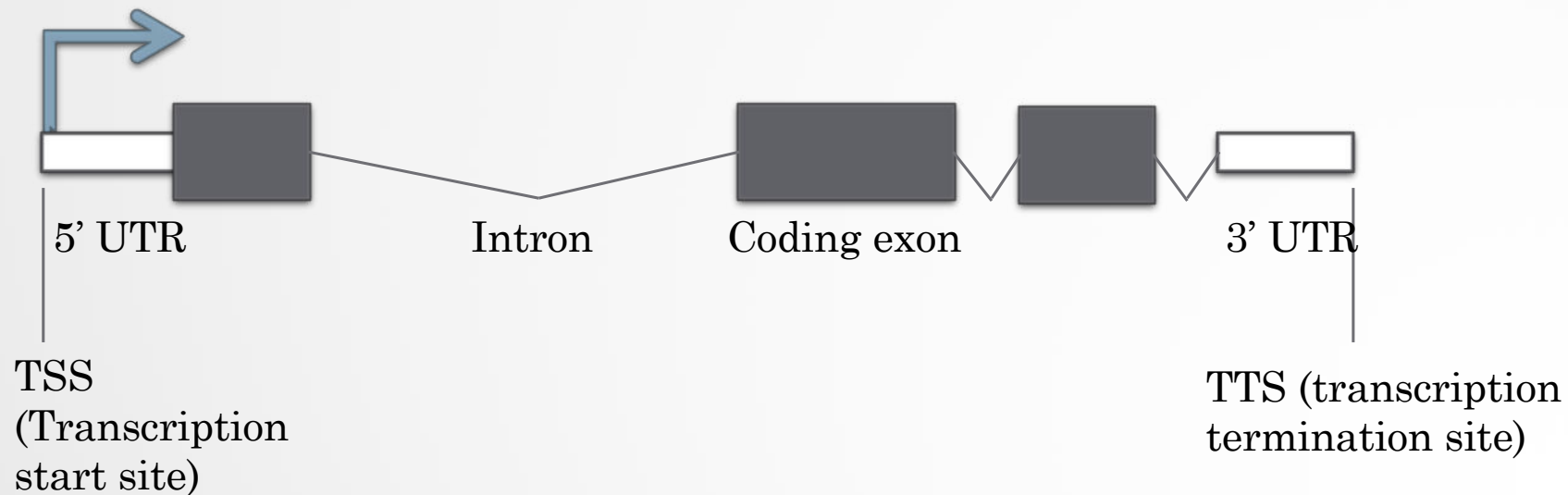
# Peak annotation

- 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



# 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 6: 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 (2<sup>nd</sup> 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.

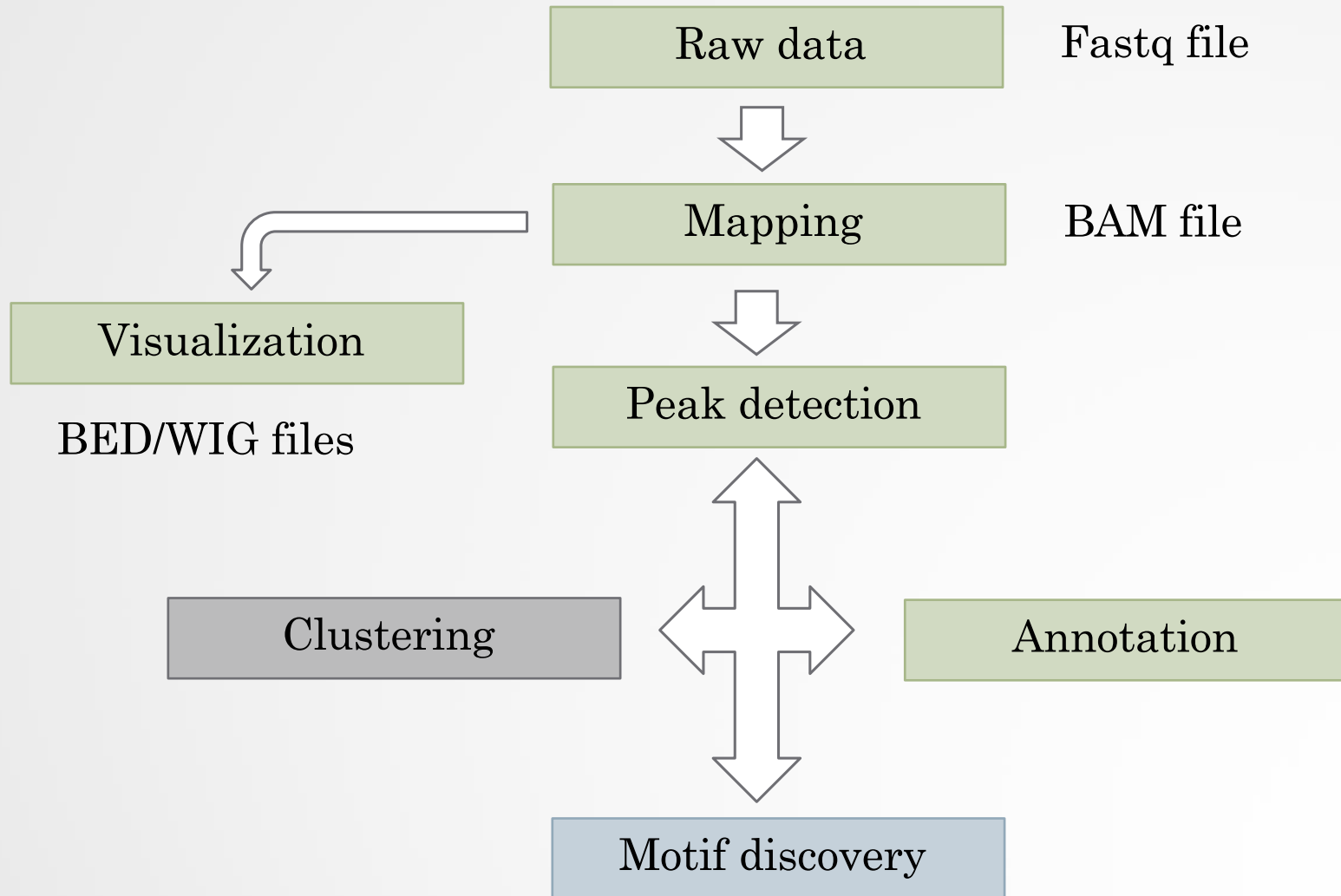
Common plots generated after the annotation steps are:

- 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**
  - Name the plot: Frequency of peaks relative to TSS
  - Name the X axis: Distance to TSS

# Exercise 6: peak annotation

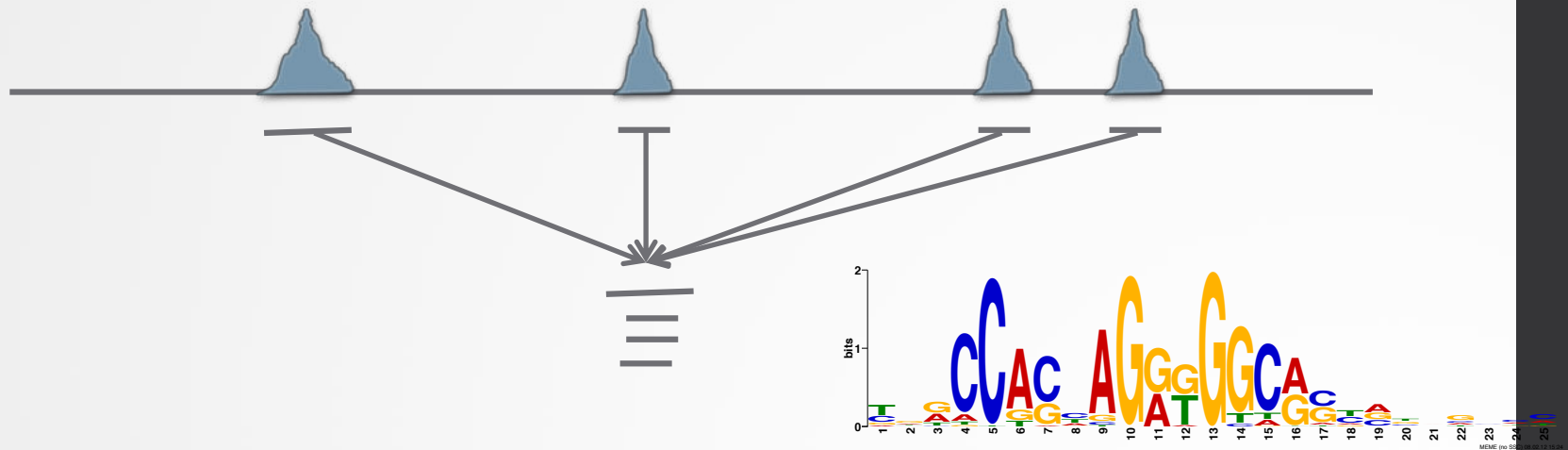
- 4. Draw a pie chart presenting the proportion of genomic features associated to the MITF peaks. To achieve this, we are going to count the number of time the genomic features (intron, exon...) are found in the Annotation column of the dataset (tabular) generated in 1.
  - 4.a. Use the tool **Cut** to extract the column “Annotation” from the dataset which contains the annotated peaks.
  - 4.b. the column containing genomic features starts with the header « Annotation ». Remove the first line with the tool **Remove beginning**.
  - 4.c. Use the tool **Count occurrences of each record** to count the number of each of the genomic features.
    - Sort in descending order.
    - Delimited by: Whitespace
  - 4.d. Expand the box of the dataset generated in 4.d and click on  **Charts** and select **Pie Chart (NVD3)** to generate a pie chart on the data. You can name the pie chart “Proportion of peaks falling into several genomic features.” You can click on “Select data” and “Customize” to select the right columns to plot and to edit features of the plot.

# Guidelines



# Motif discovery

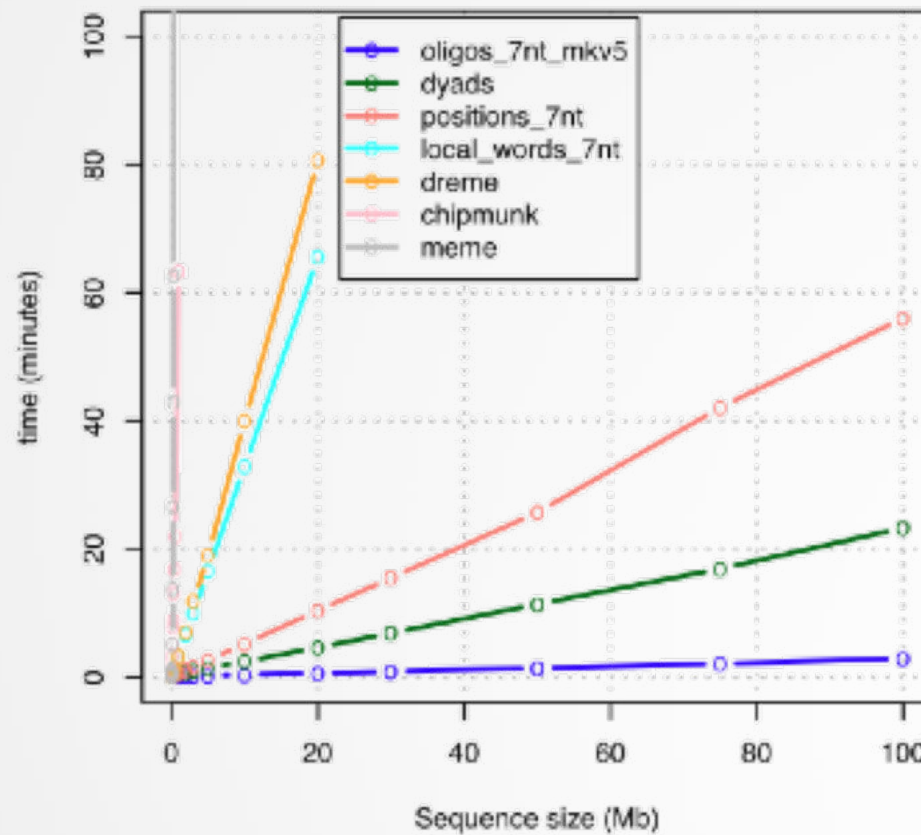
- Sequence to which the protein of interest may be bound
- Search for enriched nucleotide sequences (i.e motifs) within peak sequences.



- 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





# 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. The input sequences should be centered on a 100 character region expected to contain motifs.

# 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 7: *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 +/- 50b around MITF peak summits

- 1. Extract the top 800 peak summits (ranked by -log<sub>10</sub>pvalue)
  - 1.a. Sort the peak summits by decreased -log<sub>10</sub>pvalue 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 +/- 50nt using the dataset which contains MITF peak summits (2<sup>nd</sup> 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 (<http://meme-suite.org/tools/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)**

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

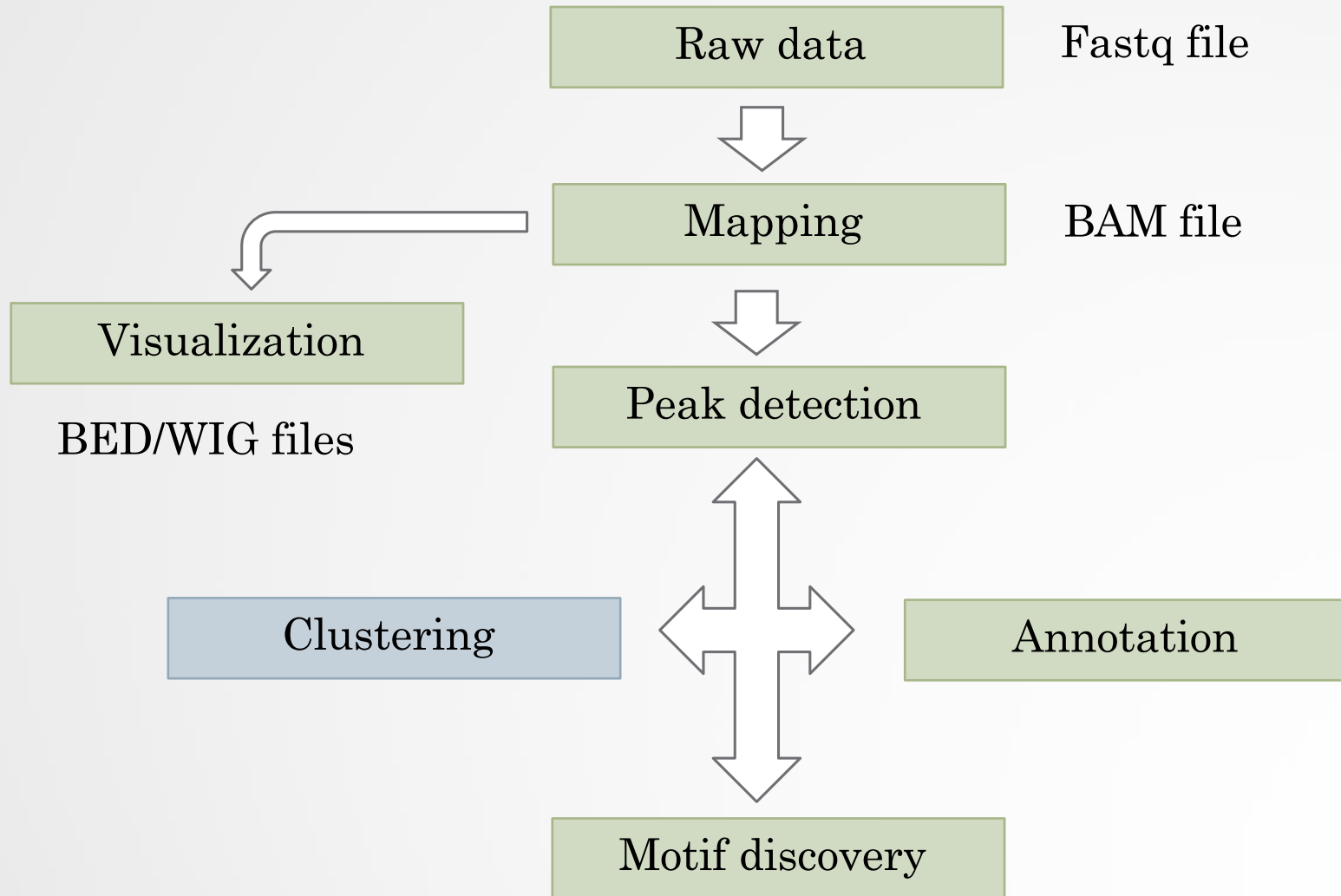


<http://weblogo.berkeley.edu/logo.cgi>

# 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.

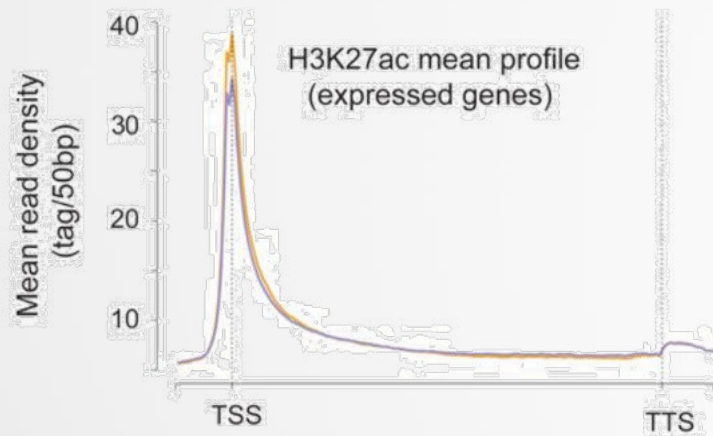
# Guidelines



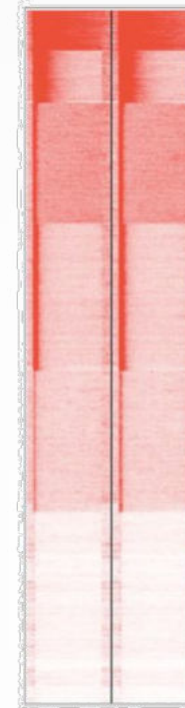
# Meta-profiles

- 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)

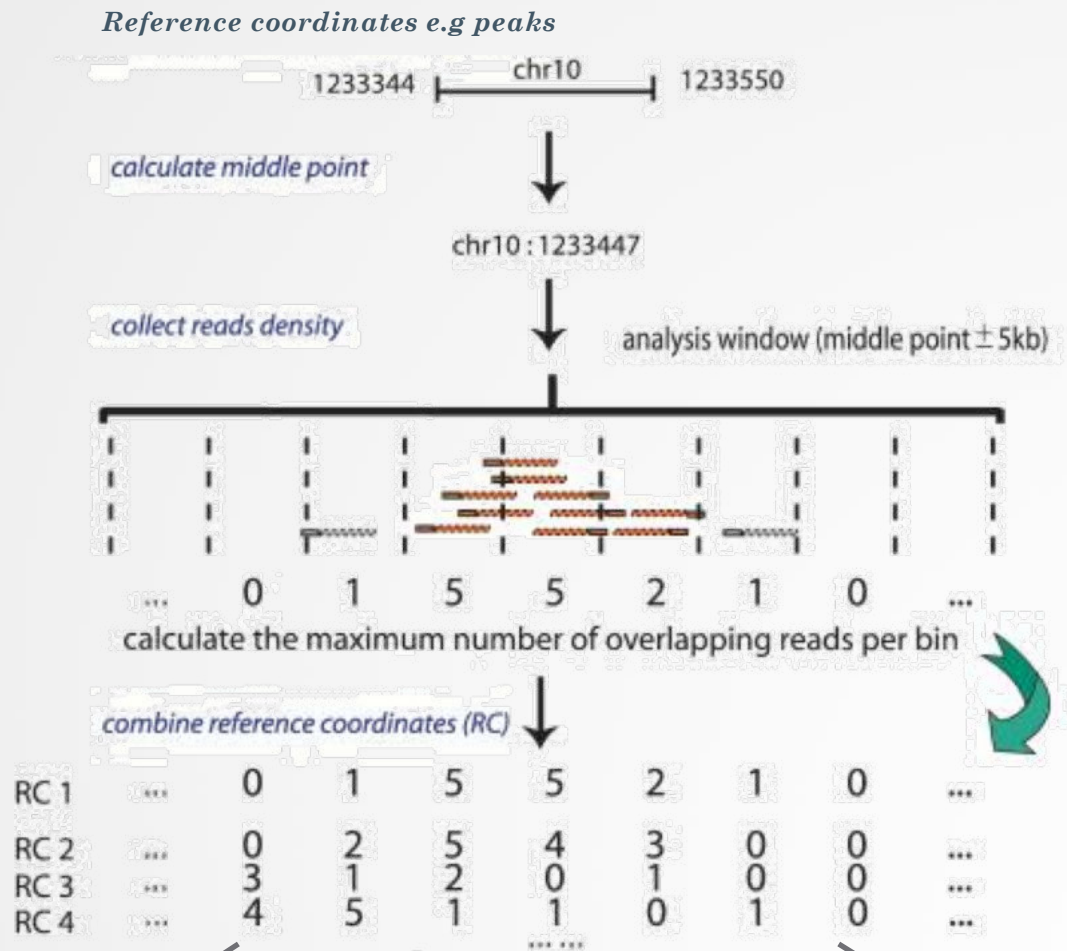
Mean profile



Heatmap



# Computing meta-profiles



Ye et al, 2011

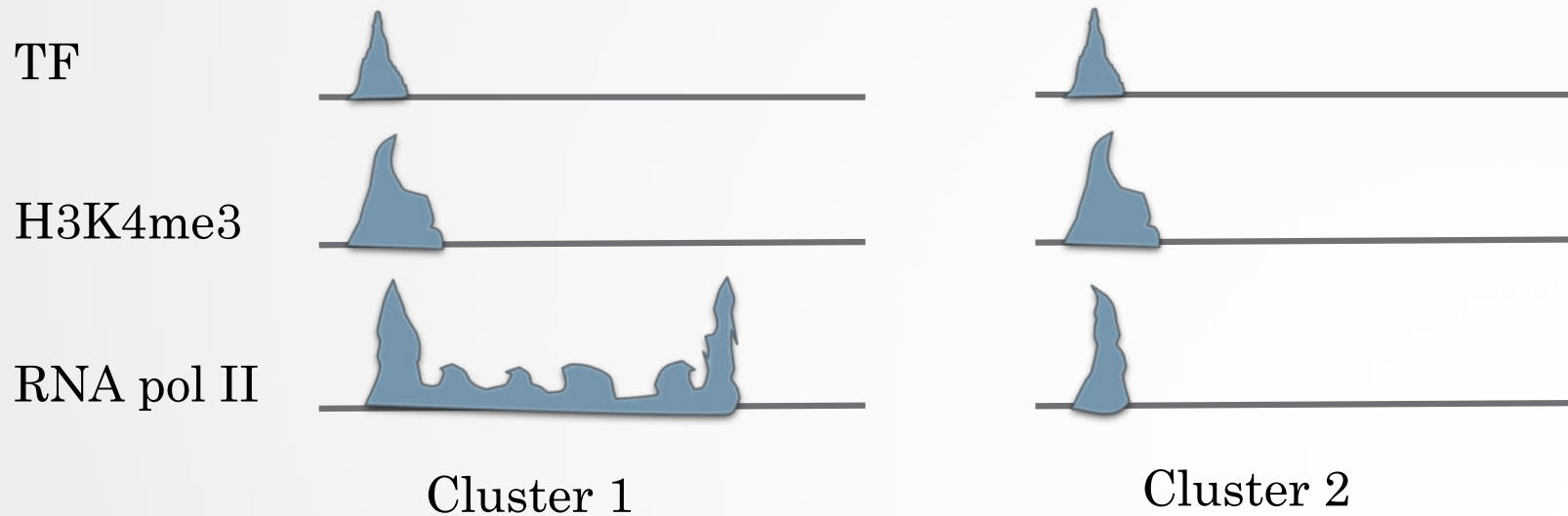
- (Clustering)
- Heatmap

- Mean of each column  
-> Mean profile

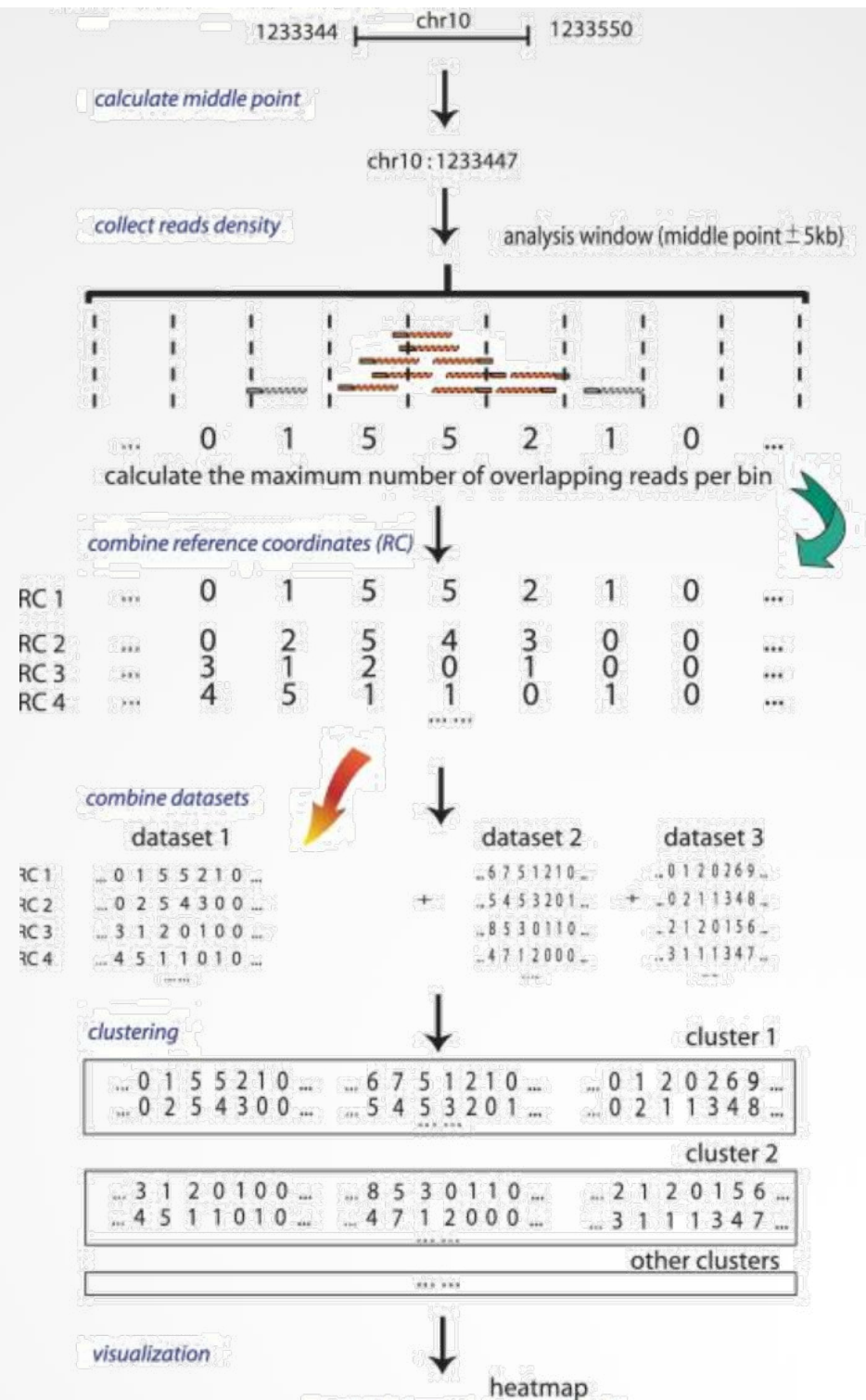


# Clustering (heatmap)

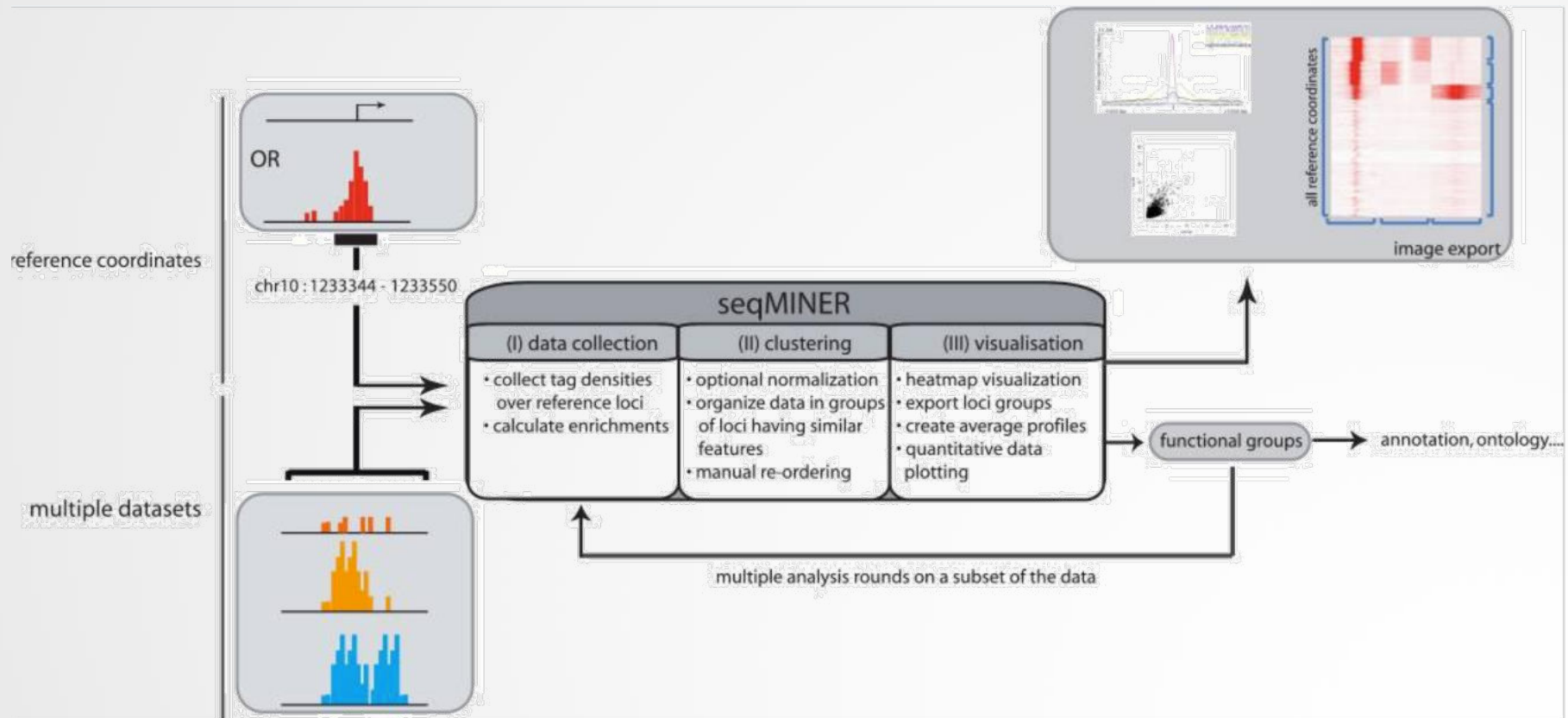
- Group together genomic regions with similar enrichments
- In a single sample or multiple samples
- E.g:



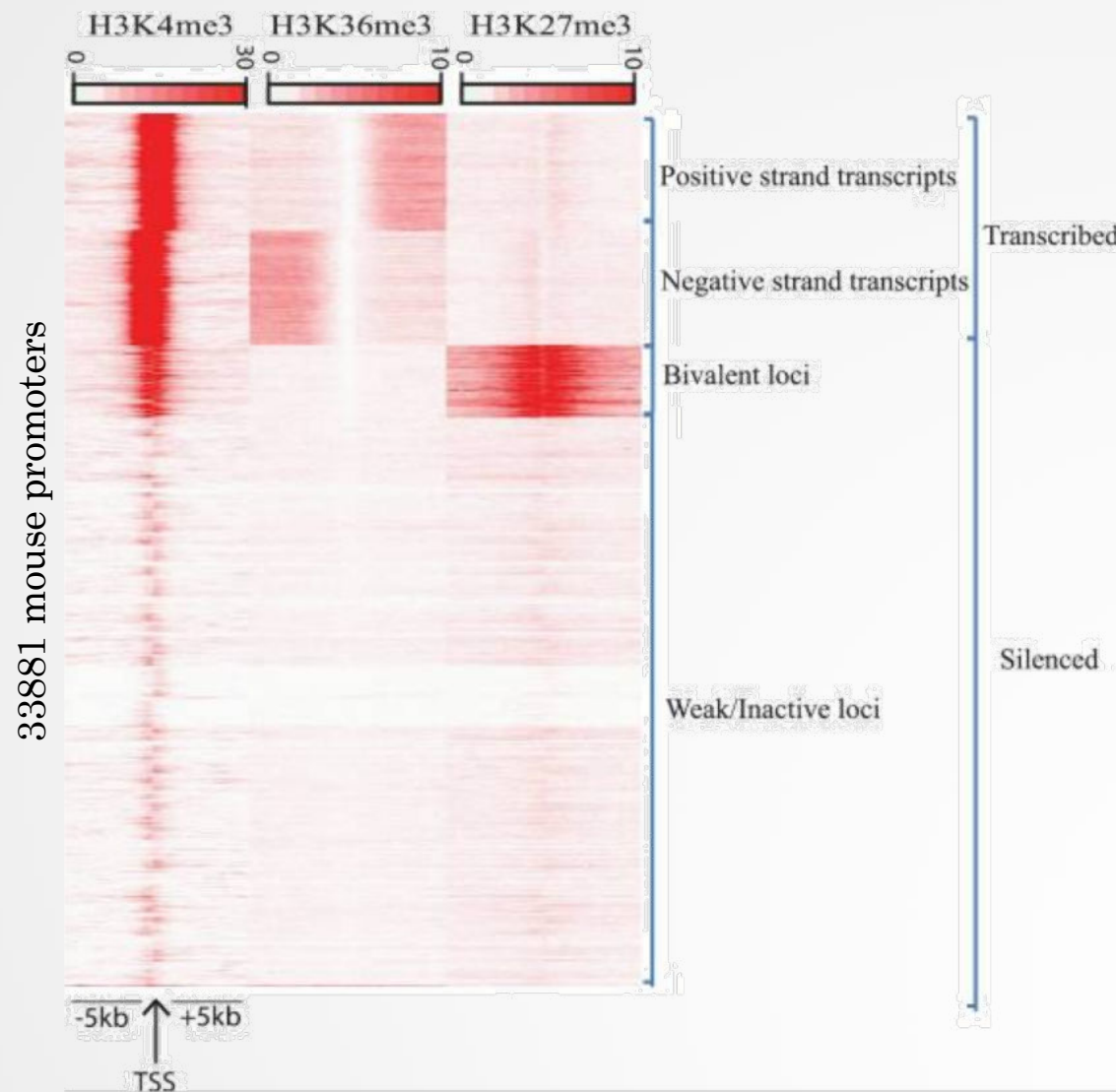
# Clustering (heatmap)



# SeqMINER [Ye et al, 2011]

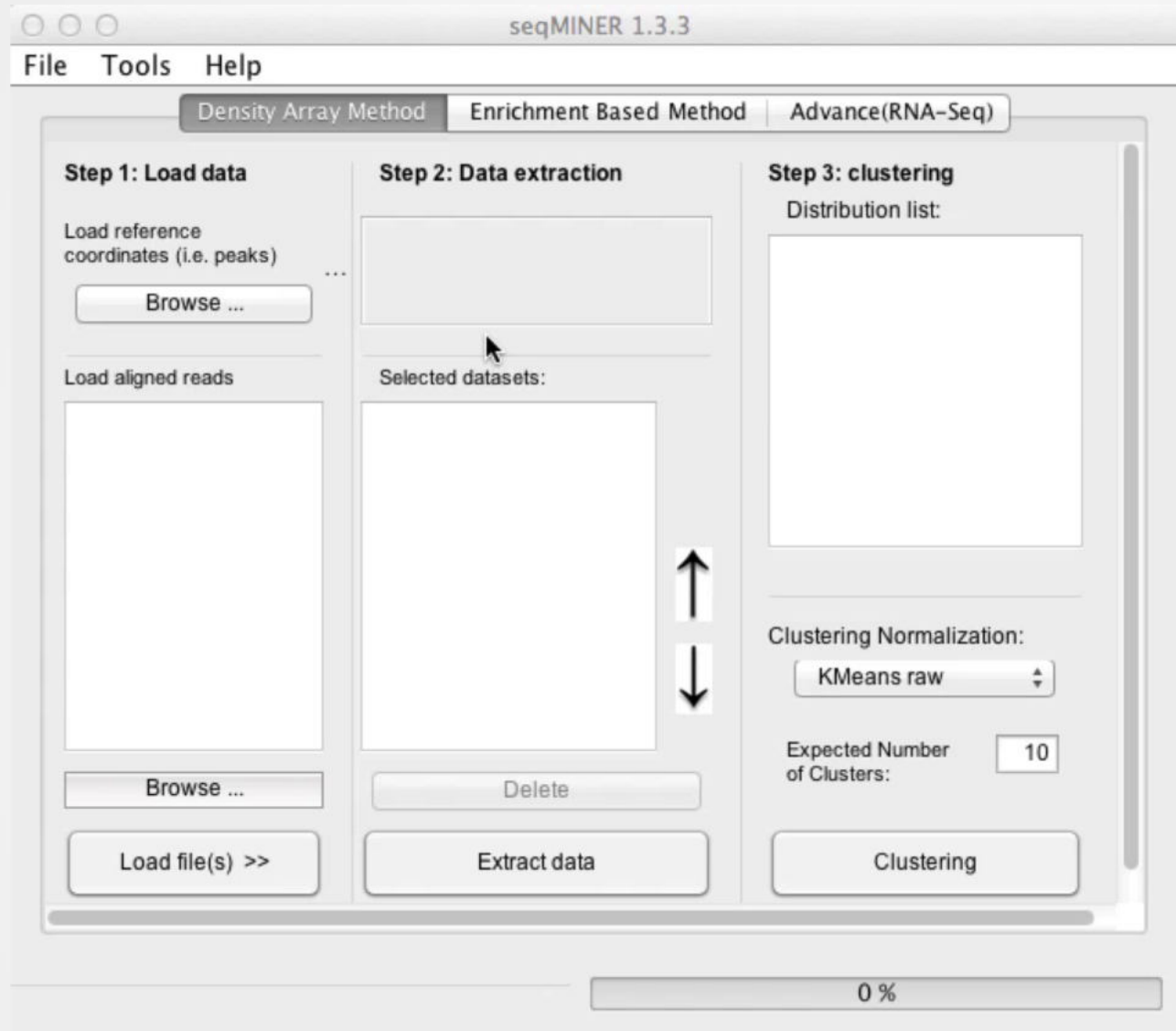


# SeqMINER [Ye et al, 2011]



The darker the red the higher the read enrichment

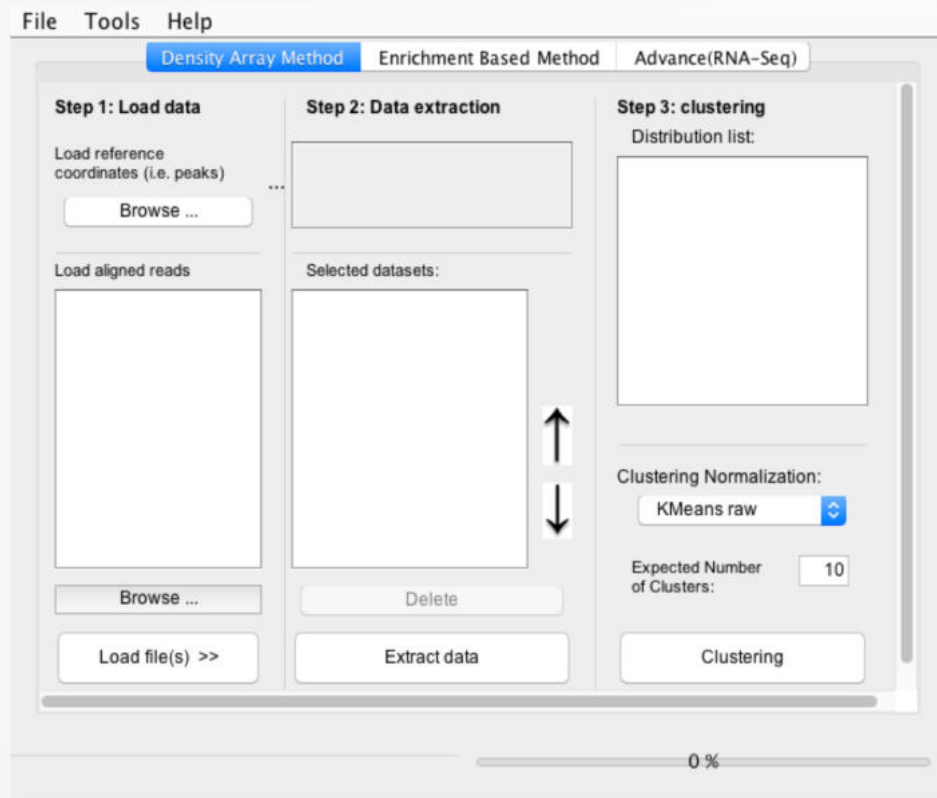
# Example



# Exercise 8: 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.



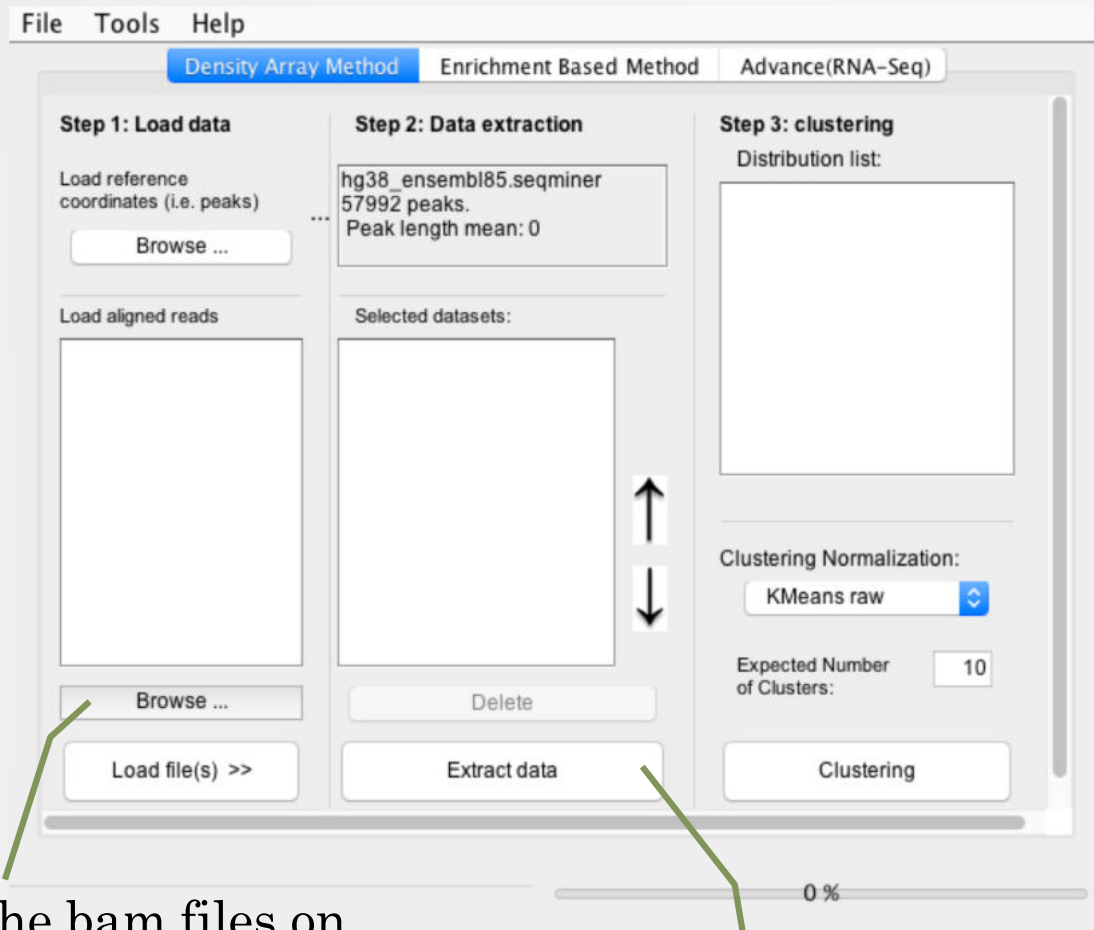
# Exercise 8: Clustering

- 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\_ensembl95. 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 :

The screenshot shows a two-step process. Step 1, 'Load data', includes a label 'Load reference coordinates (i.e. peaks)' and a 'Browse ...' button. Step 2, 'Data extraction', shows a text box with the following information: 'hg38\_ensembl95.seqminer', '58676 peaks.', and 'Peak length mean: 0'. An ellipsis '...' is positioned between the two steps.

# Exercise 8: Clustering

- Load the datasets



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

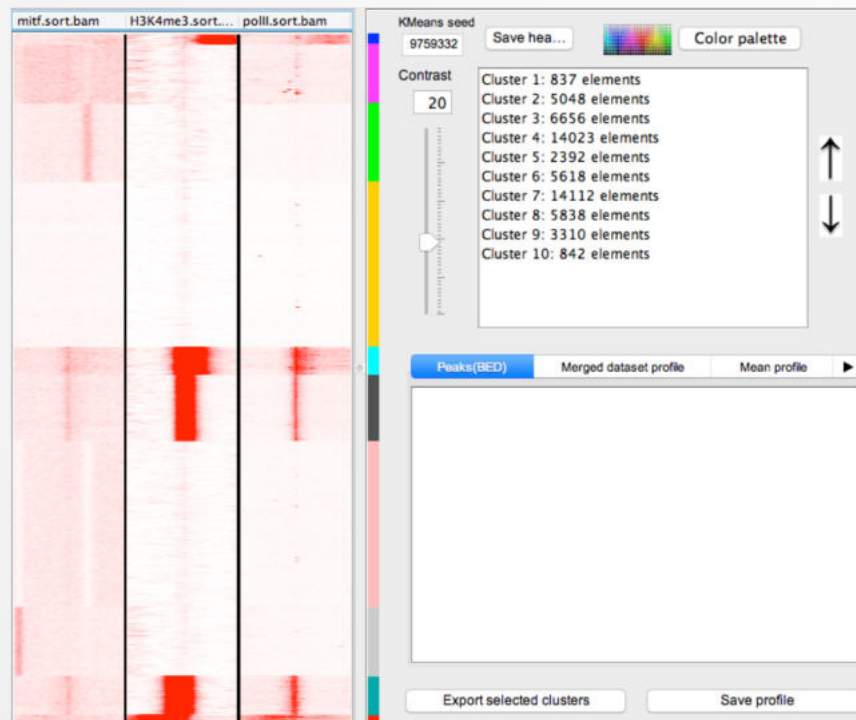
2. Once step 1, is done, click on Extract data.



# Exercise 8: 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.



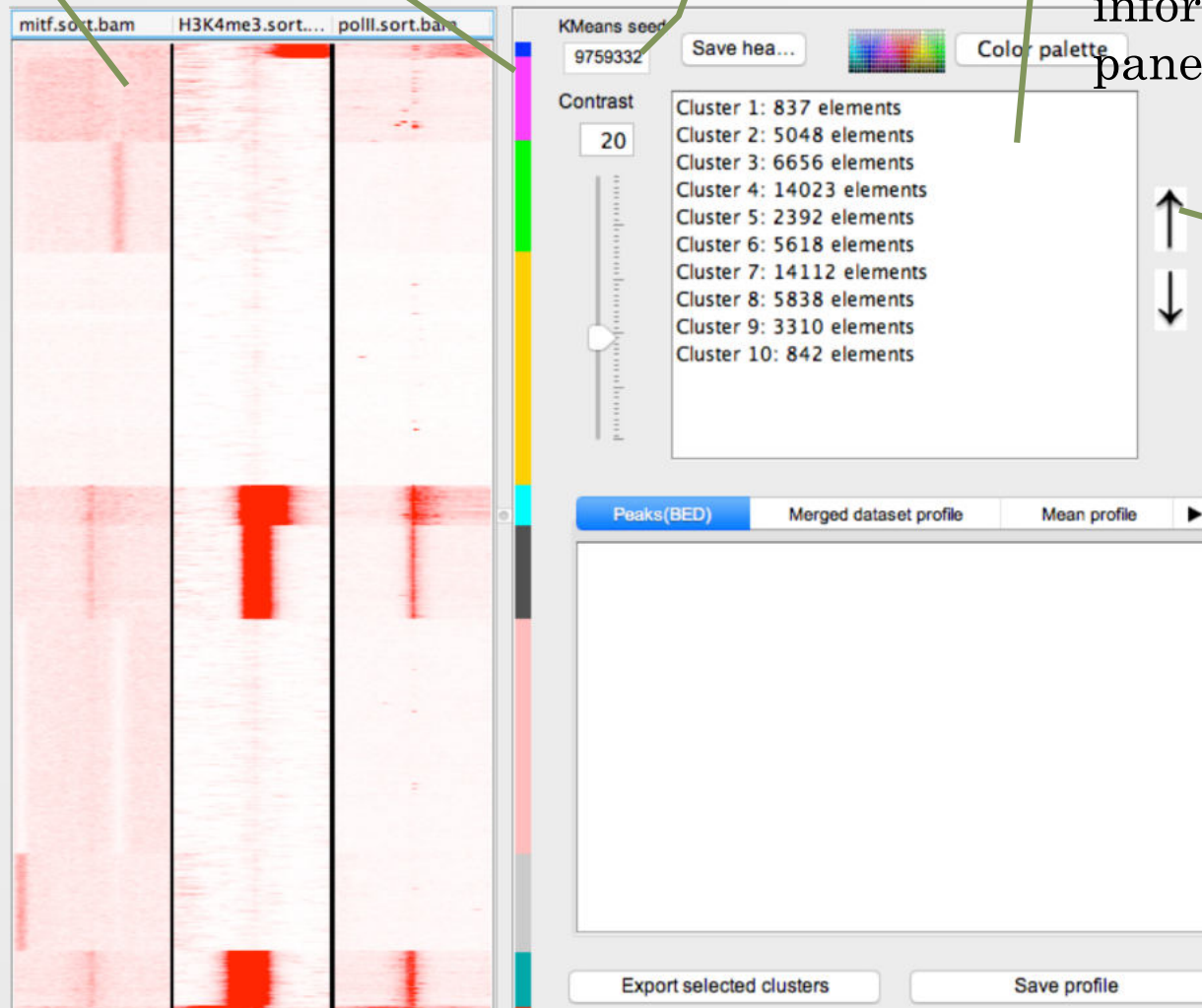
# Exercise 8: Clustering

Heatmap

Cluster definition

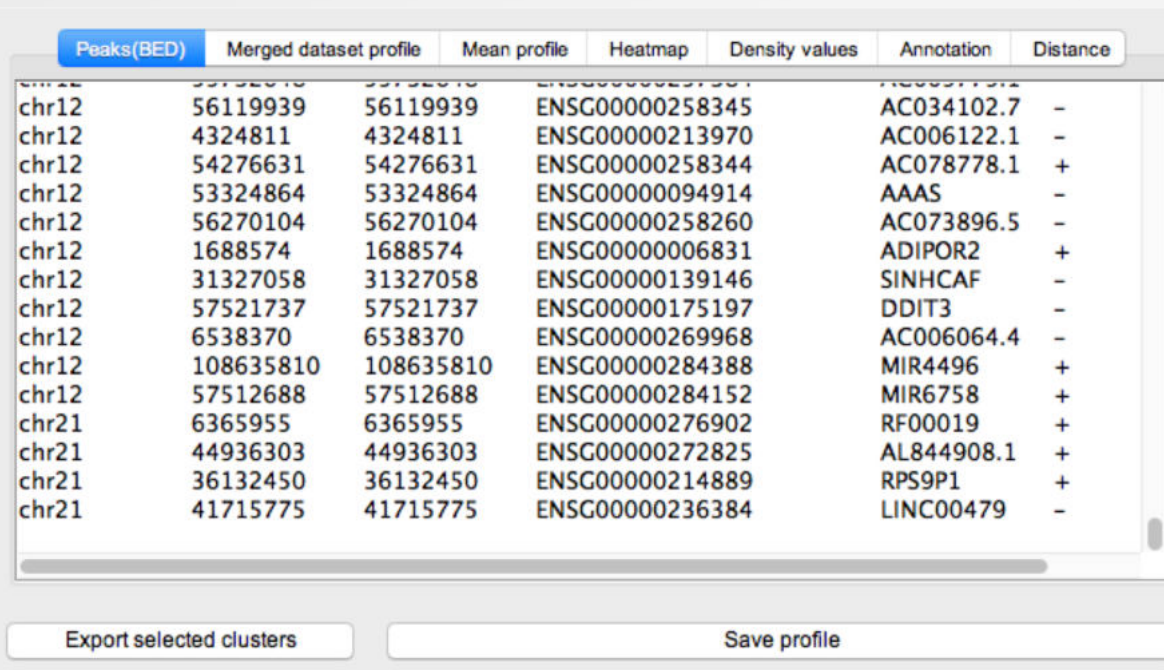
Kmeans seed value

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 8: Clustering



	Peaks(BED)	Merged dataset profile	Mean profile	Heatmap	Density values	Annotation	Distance
chr12	56119939	56119939	ENSG00000258345		AC034102.7	-	
chr12	4324811	4324811	ENSG00000213970		AC006122.1	-	
chr12	54276631	54276631	ENSG00000258344		AC078778.1	+	
chr12	53324864	53324864	ENSG00000094914		AAAS	-	
chr12	56270104	56270104	ENSG00000258260		AC073896.5	-	
chr12	1688574	1688574	ENSG00000006831		ADIPOR2	+	
chr12	31327058	31327058	ENSG00000139146		SINHCAF	-	
chr12	57521737	57521737	ENSG00000175197		DDIT3	-	
chr12	6538370	6538370	ENSG00000269968		AC006064.4	-	
chr12	108635810	108635810	ENSG00000284388		MIR4496	+	
chr12	57512688	57512688	ENSG00000284152		MIR6758	+	
chr21	6365955	6365955	ENSG00000276902		RF00019	+	
chr21	44936303	44936303	ENSG00000272825		AL844908.1	+	
chr21	36132450	36132450	ENSG00000214889		RPS9P1	+	
chr21	41715775	41715775	ENSG00000236384		LINC00479	-	

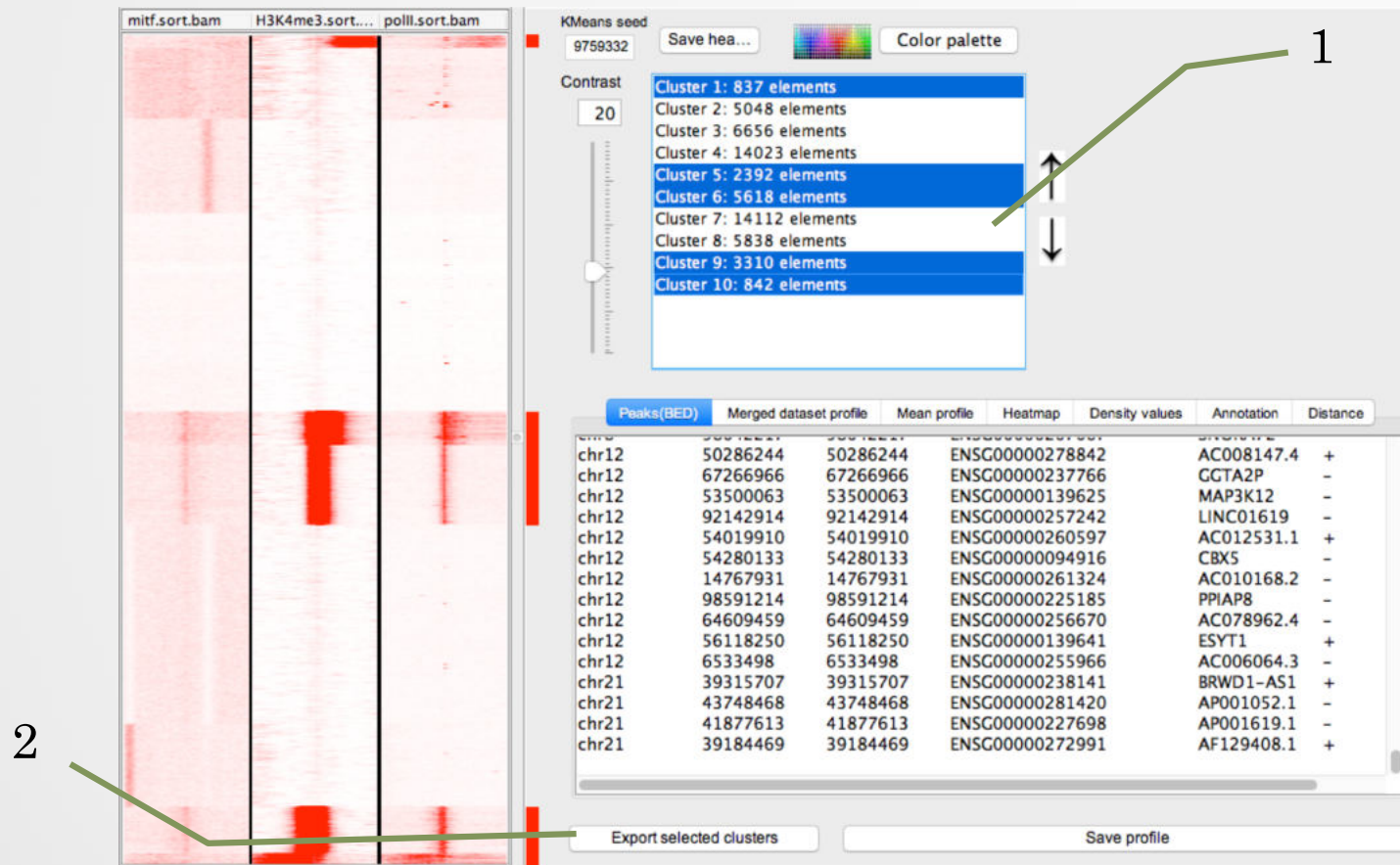
Export selected clusters      Save profile

- 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 8: 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 8: Clustering

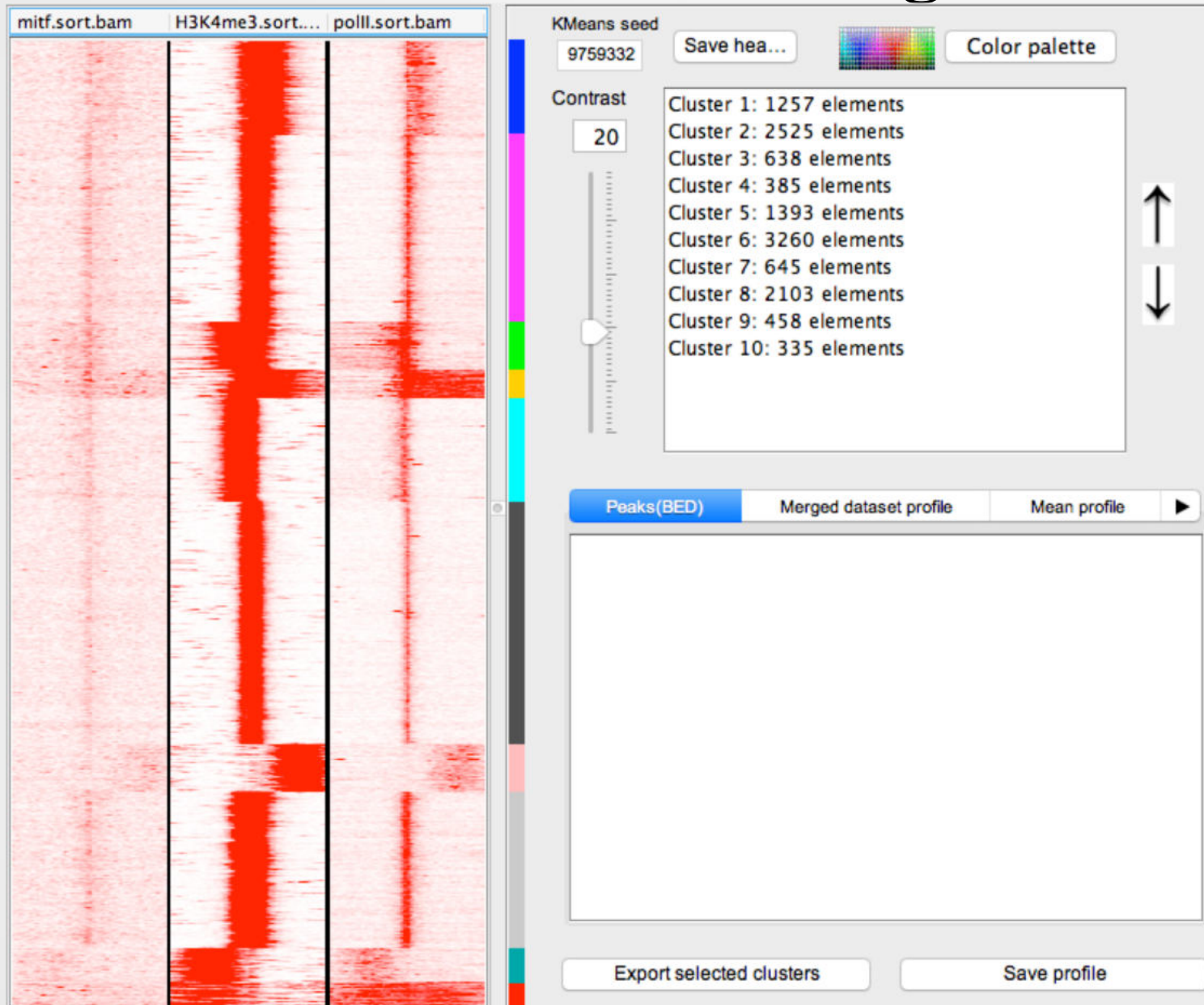
- Load the file previously generated (with cluster coordinates) as reference coordinates (1).
- Extract data (2)
- Run the clustering analysis (3)

The screenshot shows a software interface with three main steps:

- Step 1: Load data**: Includes a "Load reference coordinates (i.e. peaks)" section with a "Browse ..." button (labeled 1) and a "Load aligned reads" section with a list of files (ctrl.sort.bam, H3K4me3.sort.bam, mitf.sort.bam, poll.sort.bam) and a "Load file(s) >>" button (labeled 2).
- Step 2: Data extraction**: Shows "sub-clustering-tss.bed" with "12999 peaks" and "Peak length mean: 1". It has a "Selected datasets" list containing "mitf.sort.bam", "H3K4me3.sort.bam", and "poll.sort.bam" (labeled 2). Below the list are "Delete" and "Extract data" buttons.
- Step 3: clustering**: Shows a "Distribution list" with "hg38\_ensembl95.seqminer (m)" selected. It includes "Clustering Normalization" set to "KMeans linear" and "Expected Number of Clusters" set to "10". A "Clustering" button (labeled 3) is at the bottom.

The interface also features a menu bar (File, Tools, Help), a progress bar at the bottom (100%), and a vertical scrollbar on the right side.

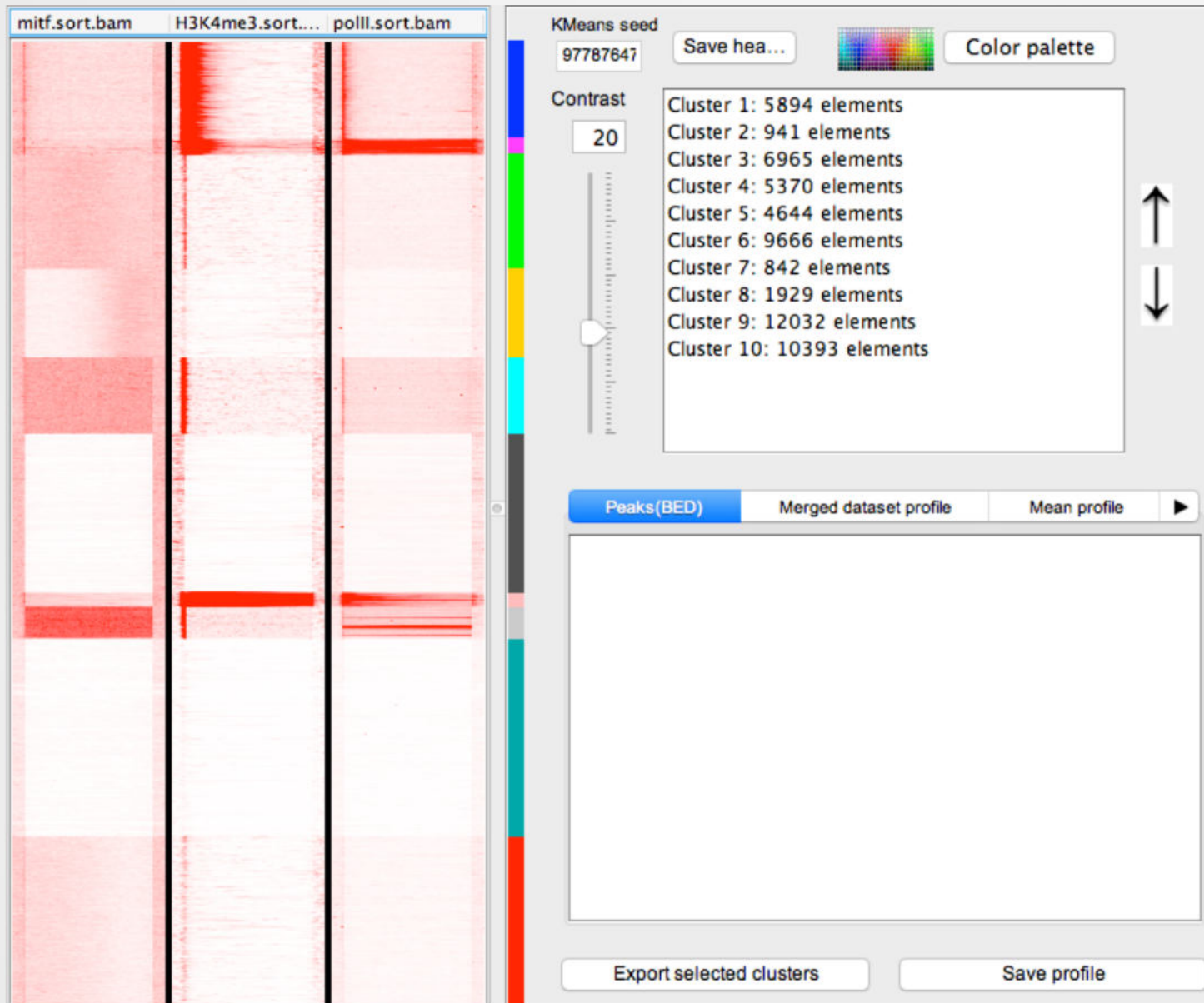
# Exercise 8: Clustering



# Exercise 8: 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 select Delete.
- Run SeqMINER on all Ensembl (v95) genes.
  - Reference coordinates : the file is in chipseq > seqMINER\_1.3.3g > lib > hg38\_ensembl95.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 8: Clustering





# Exercise 8: Clustering

- 1. Select all clusters which contains MITF, polII and H3K4me3 (clusters 1, 5, 7, 8)
  - Do a sub-clustering (keep same Kmeans seed)
- 2. Additional question:
  - 2.a. Export cluster 6. Save the file as cluster6.xls.
  - 2.b. Open the file with Excel, open a web browser to DAVID (<https://david.ncifcrf.gov/>), run a functional annotation analysis (functional annotation clustering) with the Ensembl Gene IDs from the file in excel.

# Guidelines

