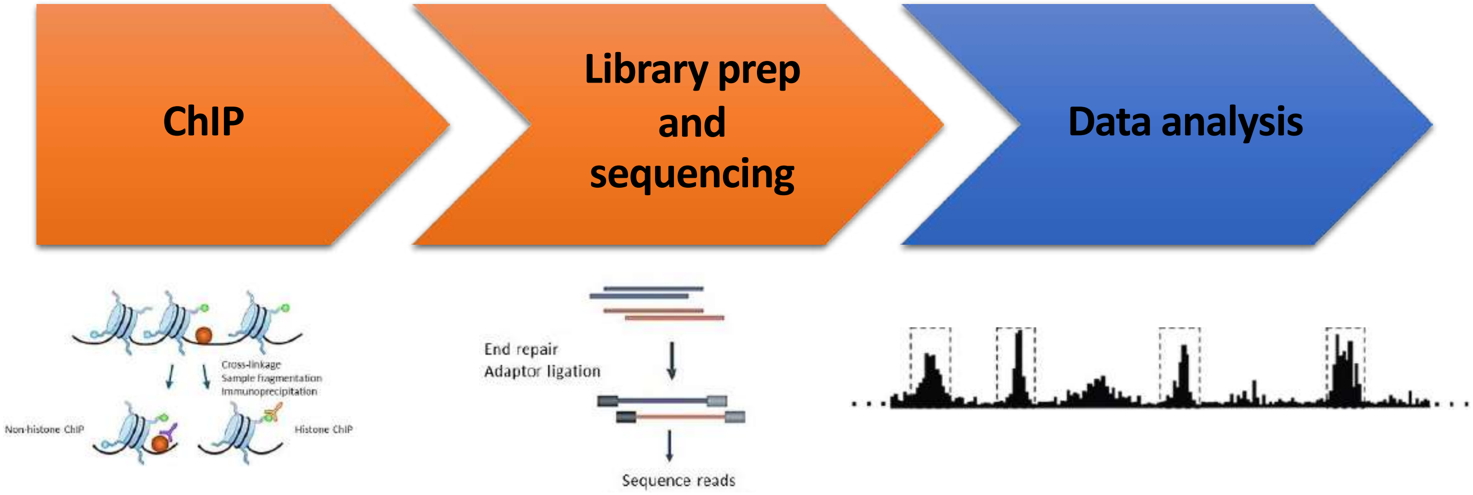
Analysis of ChIP-seq data

Stéphanie Le Gras
(slegras@igbmc.fr)

Analysis of ChIP-seq data

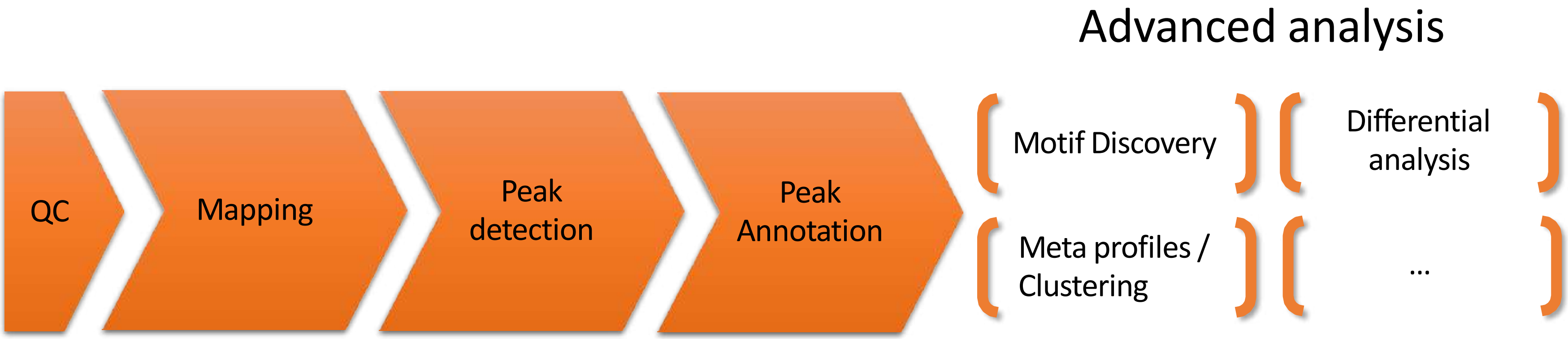


ChIP

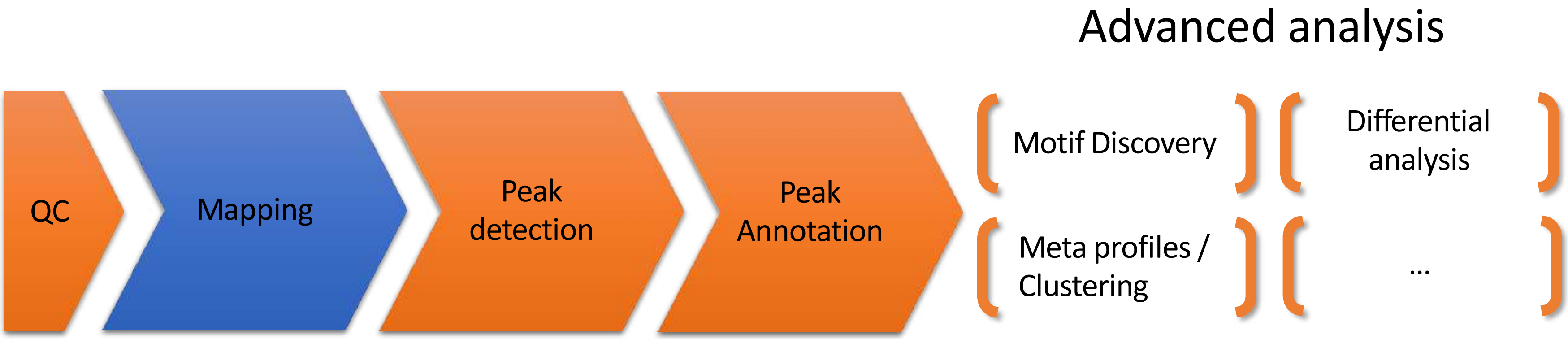
Library prep and sequencing

Data analysis

Analysis of ChIP-seq data

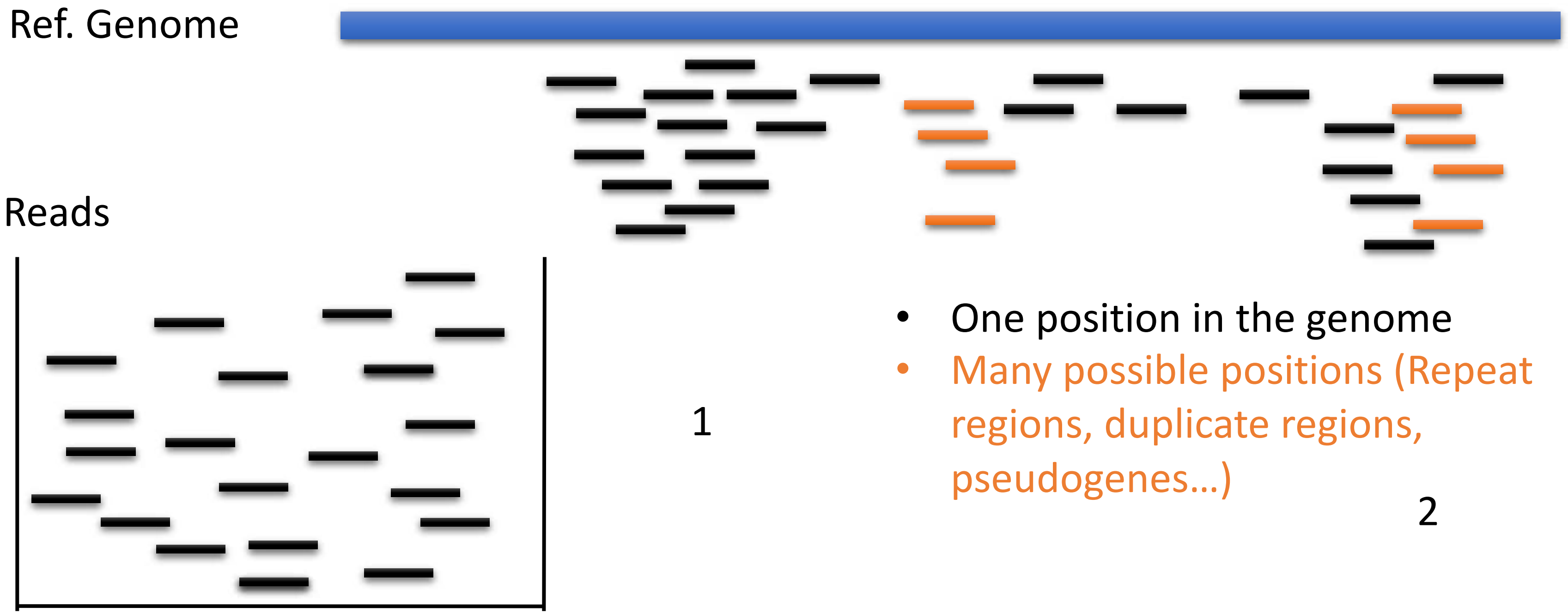


Analysis of ChIP-seq data



Mapping

- Find out the position of the reads within the genome

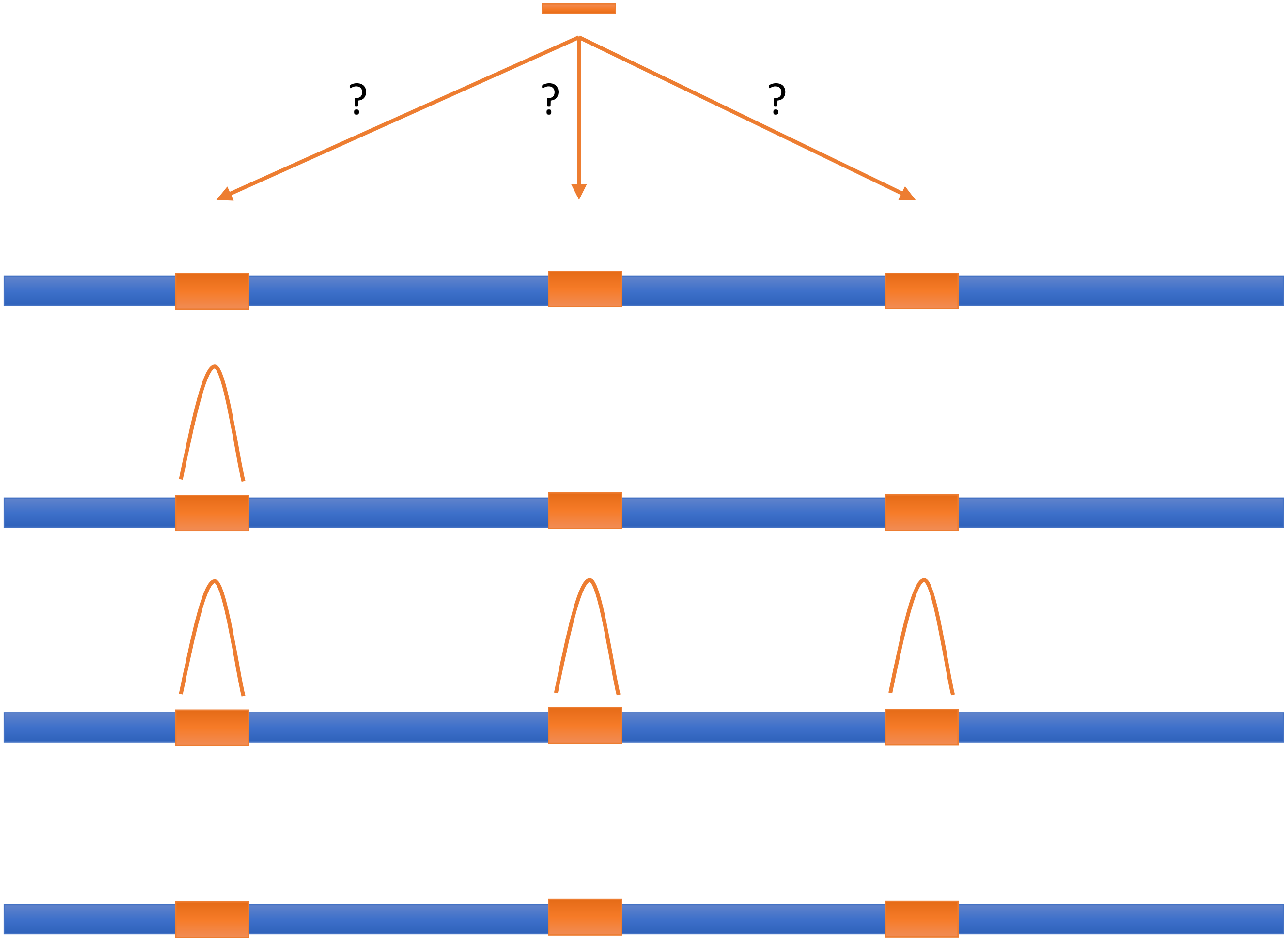


- One position in the genome
- Many possible positions (Repeat regions, duplicate regions, pseudogenes…)

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



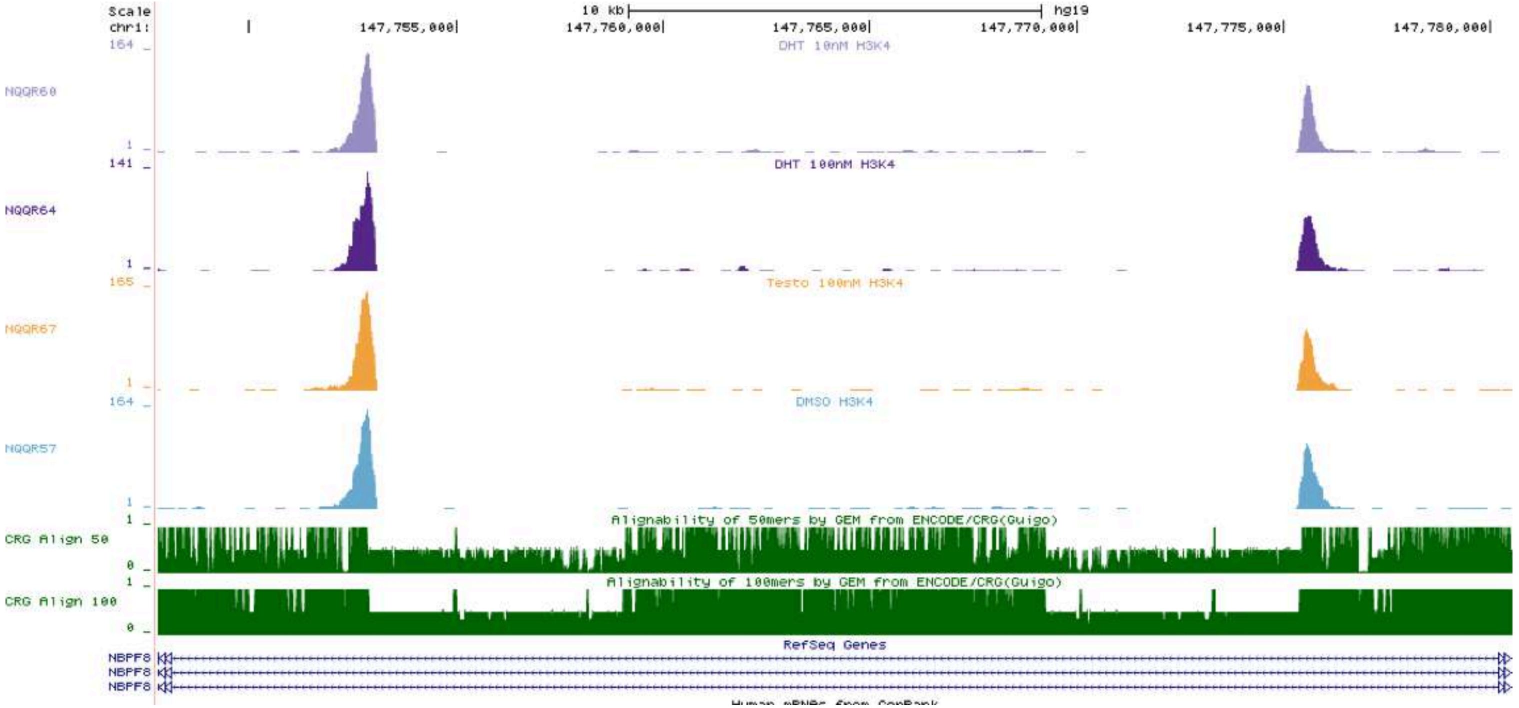
Keep 1 position randomly

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

Data were aligned using Bowtie v1 with parameters allowing to get the best possible **unique** alignment. How many reads are aligned for each of the samples?

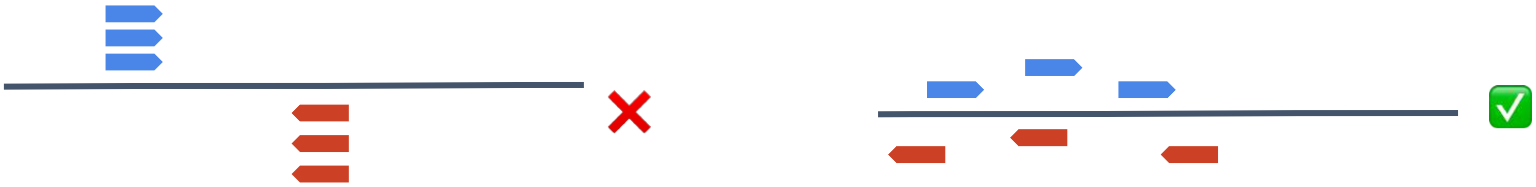
- 1. go to Galaxy France (https://usegalaxy.fr/)
- 2. create a new history named “ChIP-seq data analysis”
- 3. import 2 BAM files (22:mitf.bam and 23:ctrl.bam) from the imported history “NGS data analysis training Strasbourg”
- 4. use the tool ***Samtools flagstat*** *tabulate descriptive stats for BAM dataset* to compute the number of aligned reads in the samples.
  - The tool gives alignment statistics on a BAM file.

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= Genomic position with 1 read aligned
  - Nd = Genomic position with ≧ 1 read aligned
- 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)



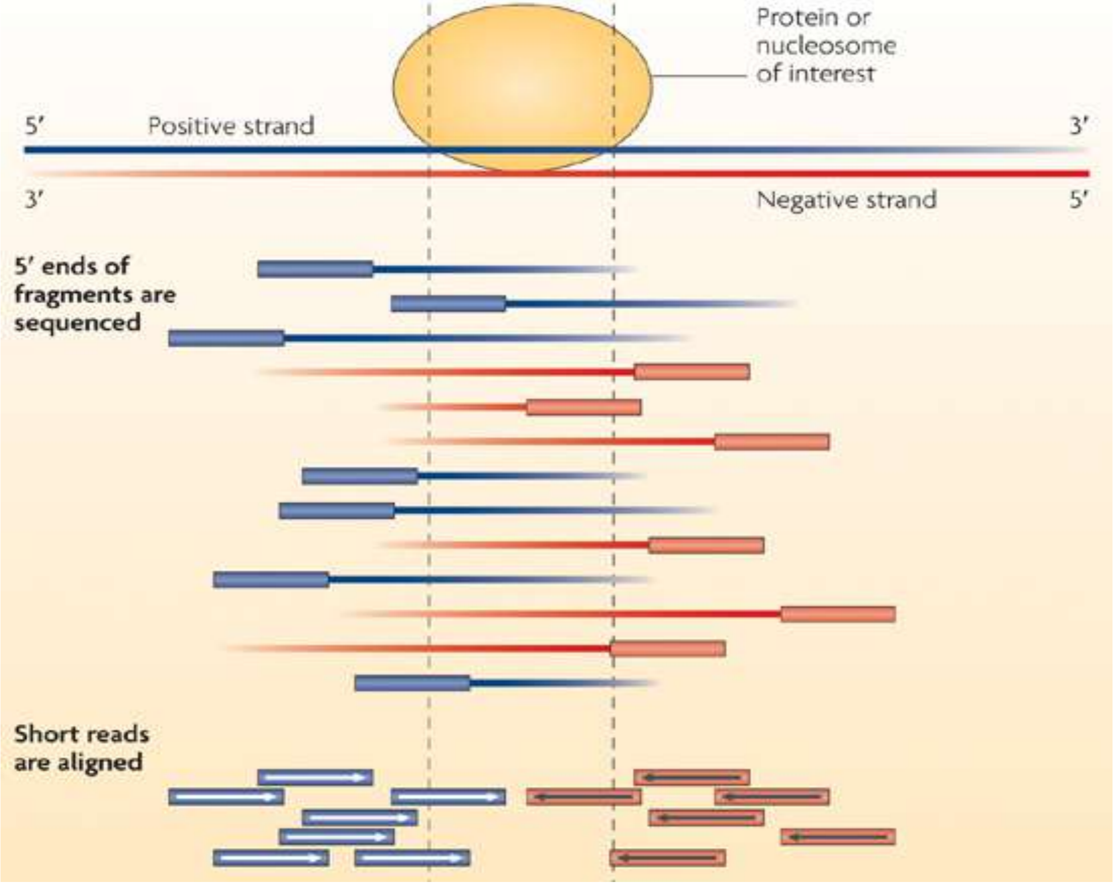
https://genome.ucsc.edu/ENCODE/qualityMetrics.html

Exercise 2: duplicate reads estimate

We want to assess the number of duplicate reads

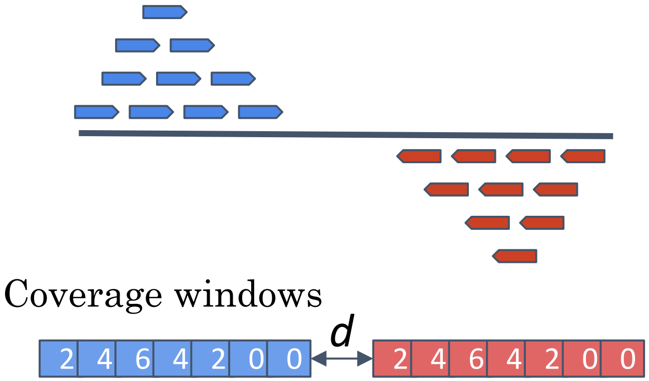
1. Use the tool **MarkDuplicates** to assess the complexity of the libraries (i.e the number of unique sequences). Use default parameters except for:
   - Select validation stringency: Silent (The picard tools validation strategy of BAM file is very stringent. So we turn off validation stringency)
   - 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

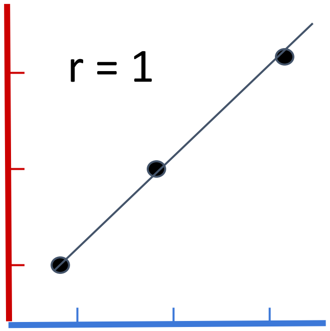

Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned

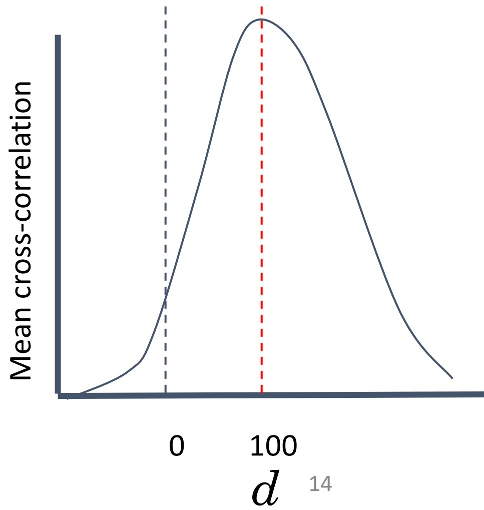
QC: Strand cross-correlation

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

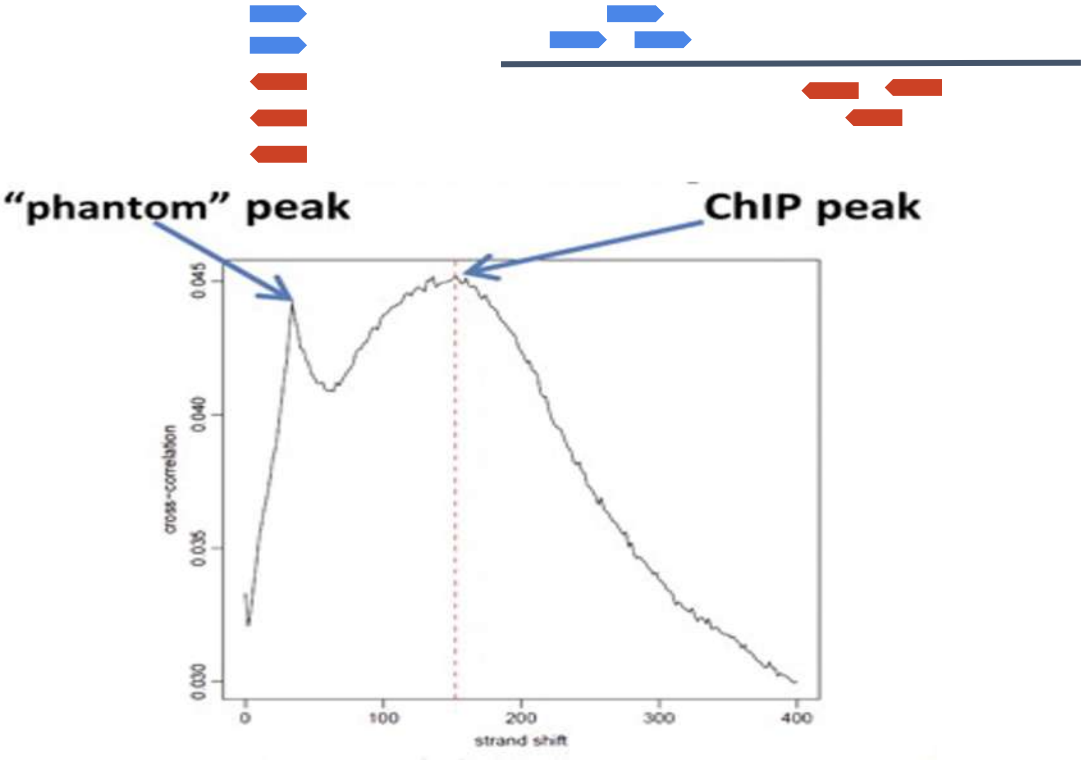


Strand cross-correlation
for each window and
various *d* values





QC: Strand cross-correlation

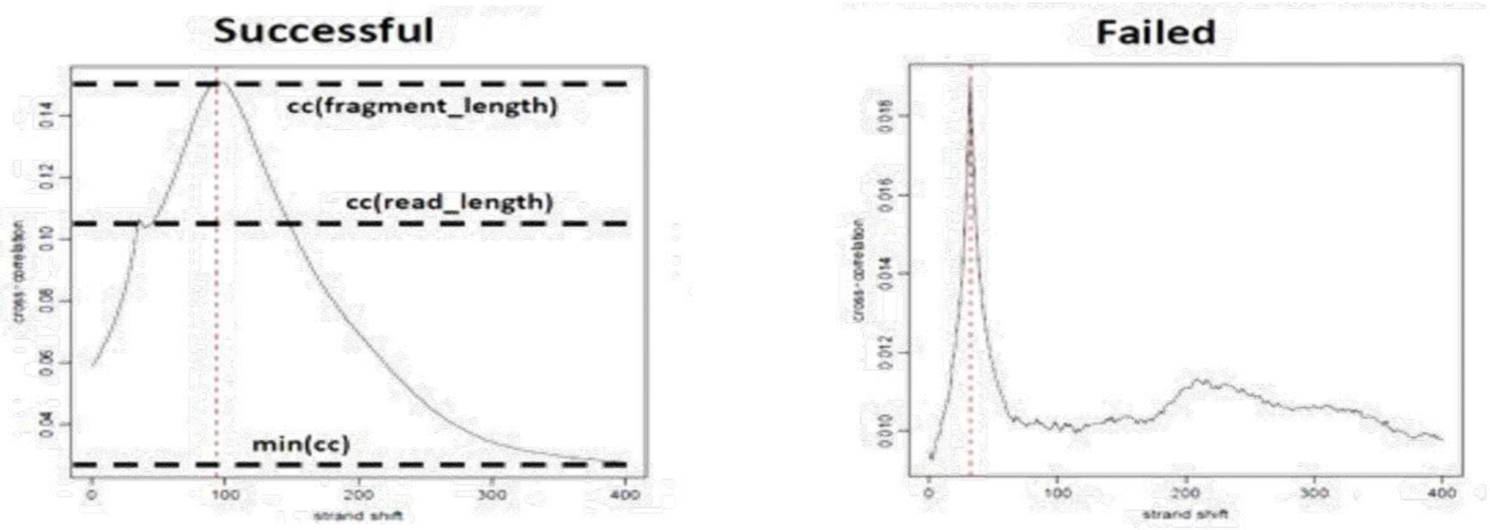


“phantom” peak

ChIP peak

Landt et al, 2012

QC: Strand cross-correlation



Landt et al, 2012

NSC: normalized strand coefficient

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

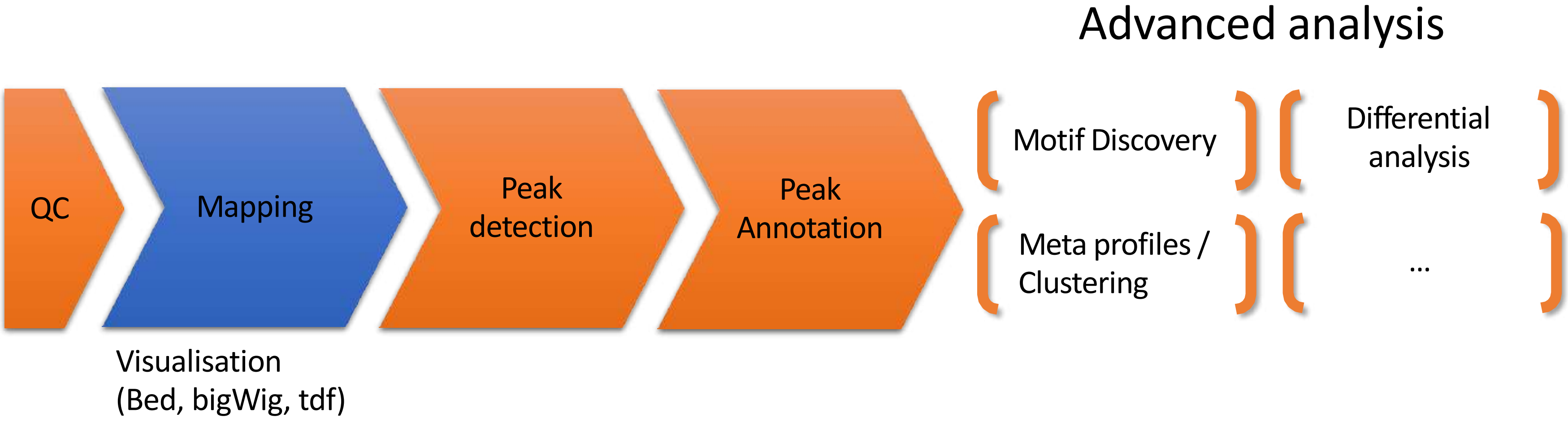
NSC ≥ 1.05 is recommended

Relative strand correlation (RSC)

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

RSC ≥ 0.8 is recommended

Analysis of ChIP-seq data



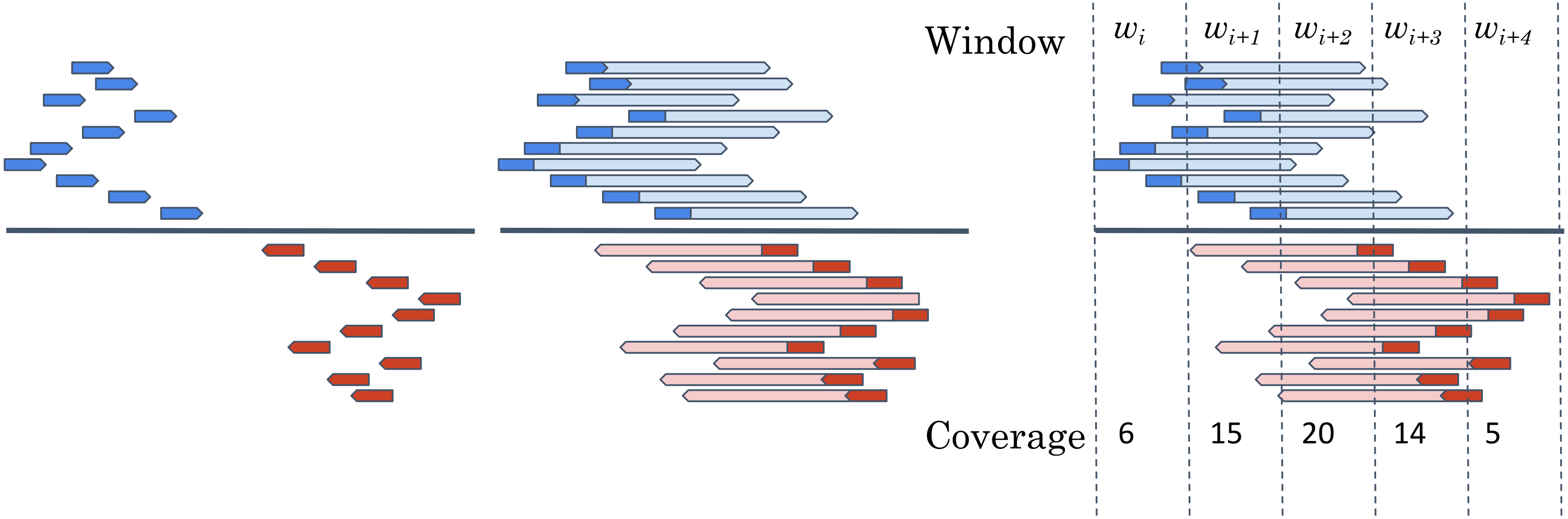
Bam files are fat

- **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)**

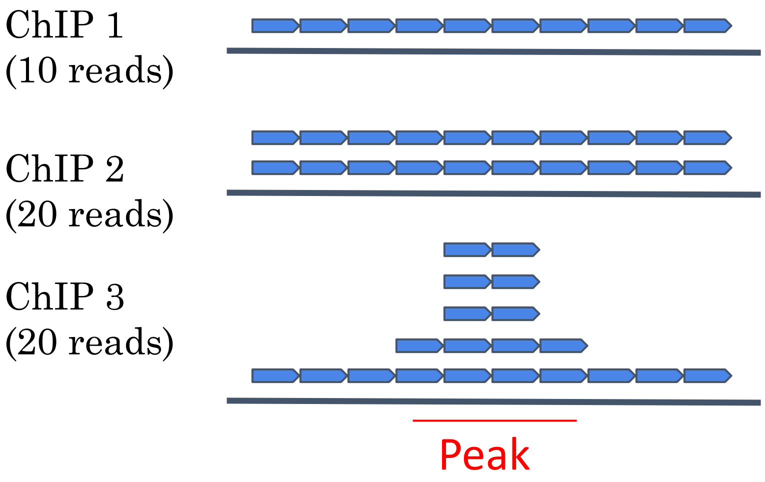
Coverage file and read extension

- 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 needs 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 using 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

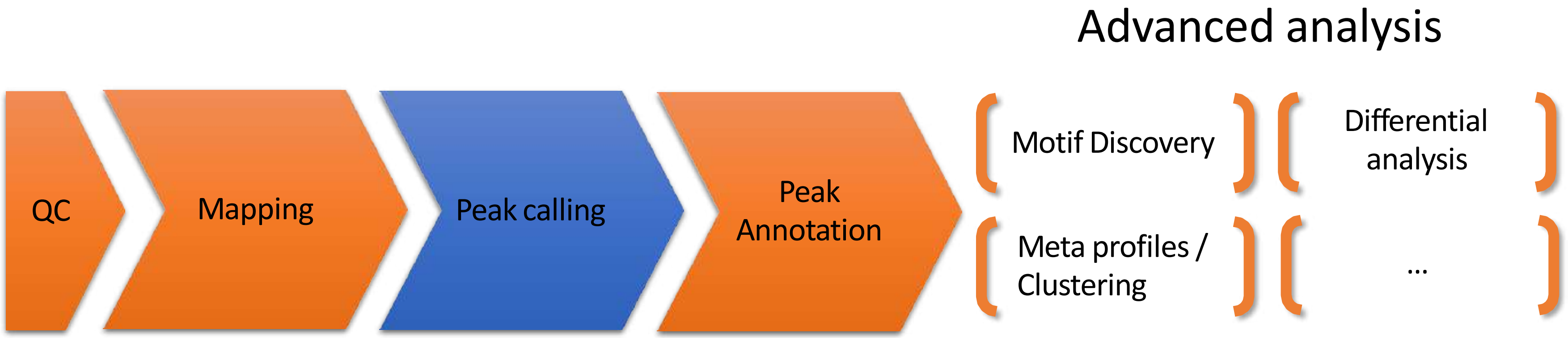
2. Check the following genes:

   - Idh1, NPAS2, AP1S2, PABPC1l, Park7, Pmel, Cdk2, Actb

   Do you see peaks at these locations?

Keep IGV opened with this two datasets

Analysis of ChIP-seq data



Advanced analysis

QC

Mapping

Peak calling

Peak Annotation

Motif Discovery

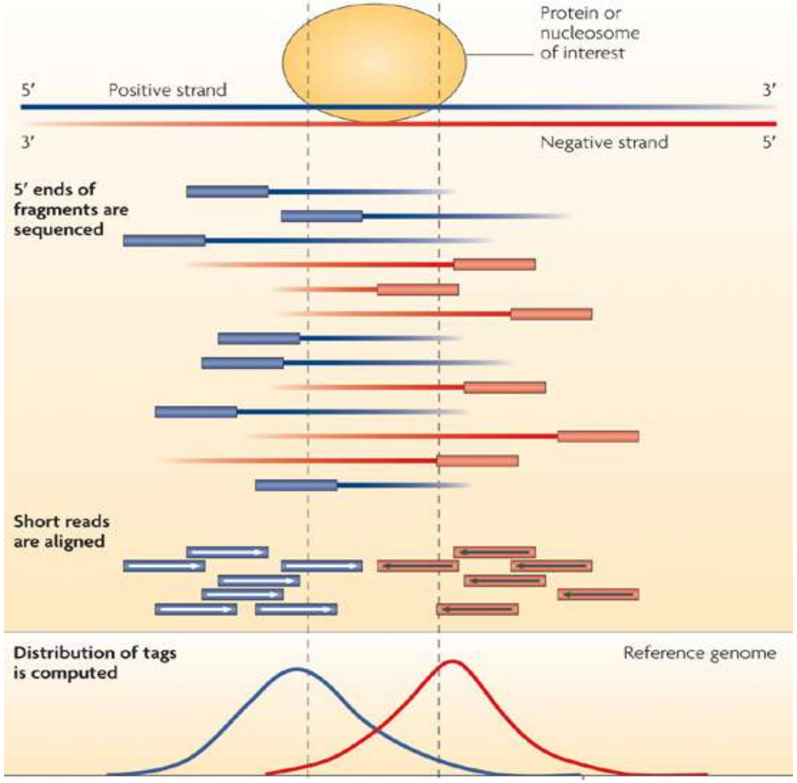
Differential analysis

Meta profiles / Clustering

…

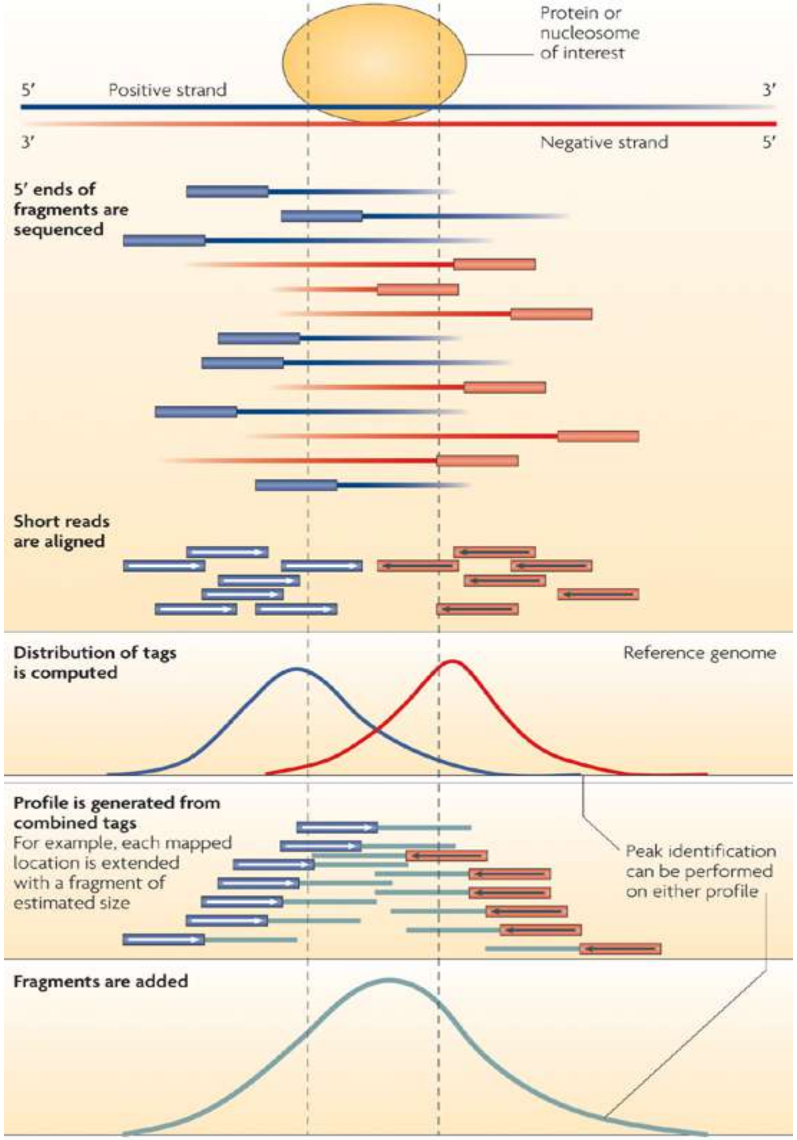
From reads to peaks

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


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome

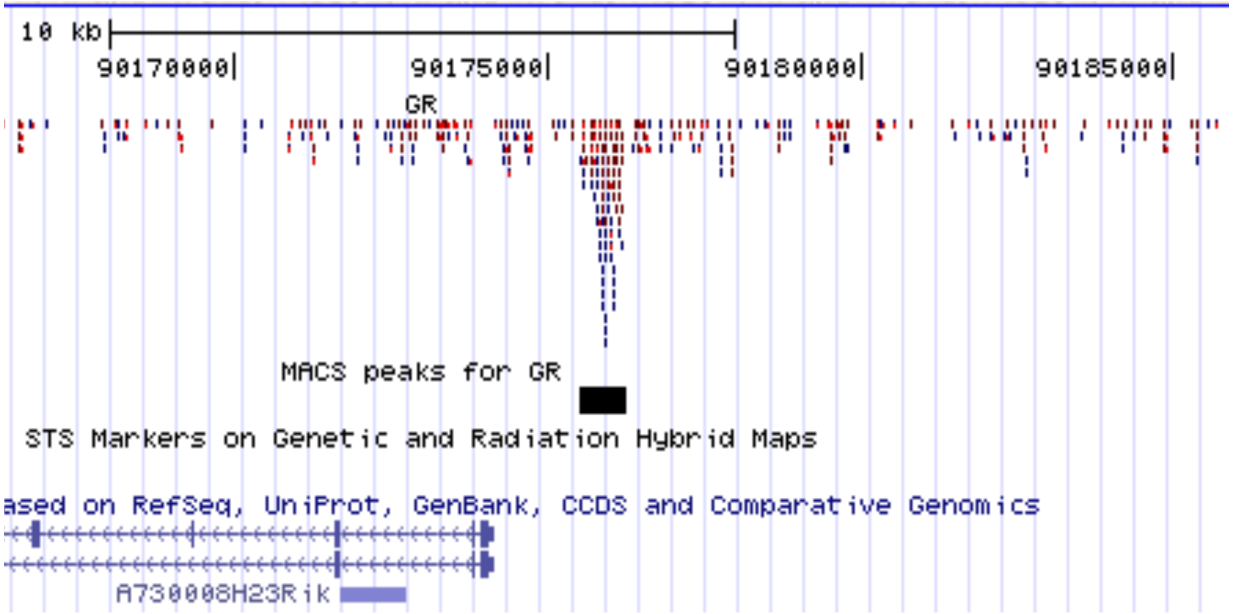
From reads to peaks

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

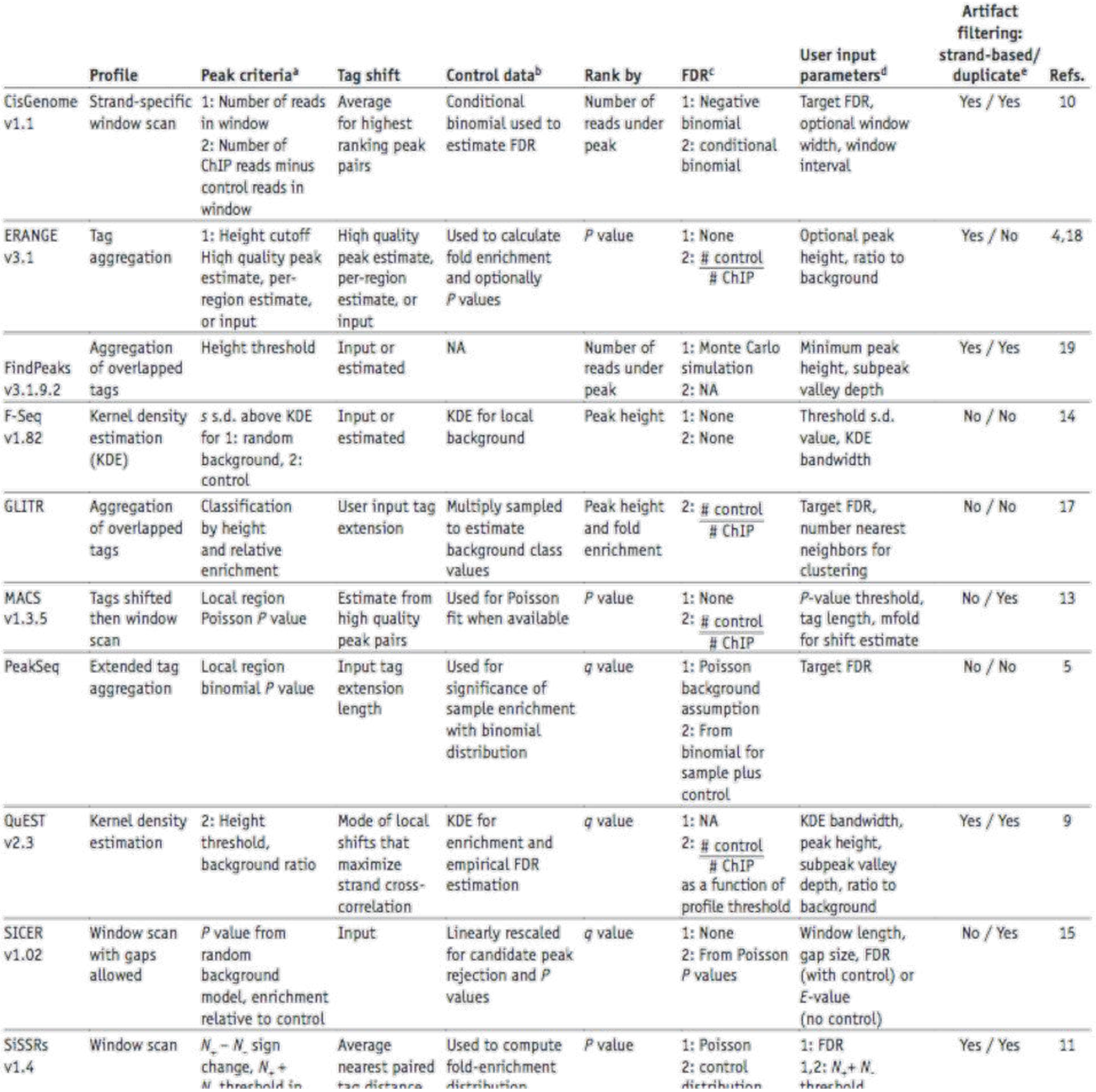

Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Peak detection

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


10 kb
90170000|
90175000|
90180000|
90185000|
GR
MACS peaks for GR
STS Markers on Genetic and Radiation Hybrid Maps
ased on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
A730008H23Rik

Peak finders



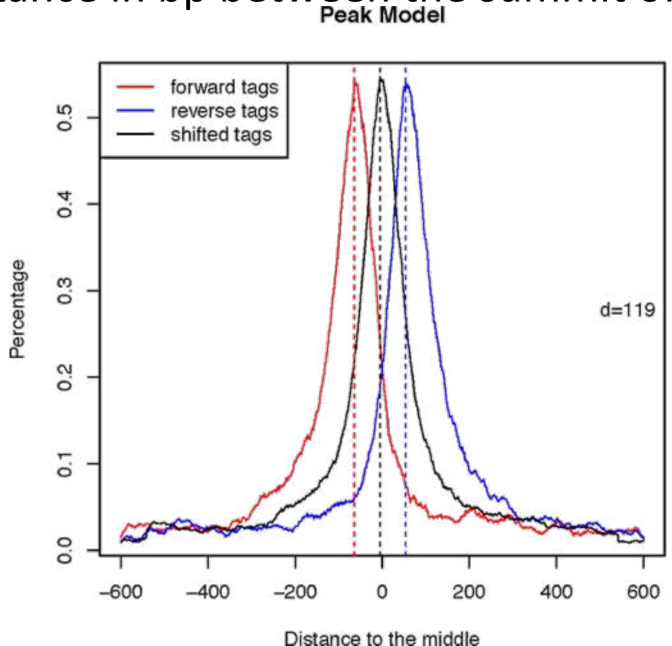
	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\text{ control}}{\#\text{ ChIP}}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\text{ control}}{\#\text{ ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\text{ control}}{\#\text{ ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\text{ control}}{\#\text{ ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ +$ N_- threshold in	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11

Pepke et al, 2009

MACS [Zhang et al, 2008]

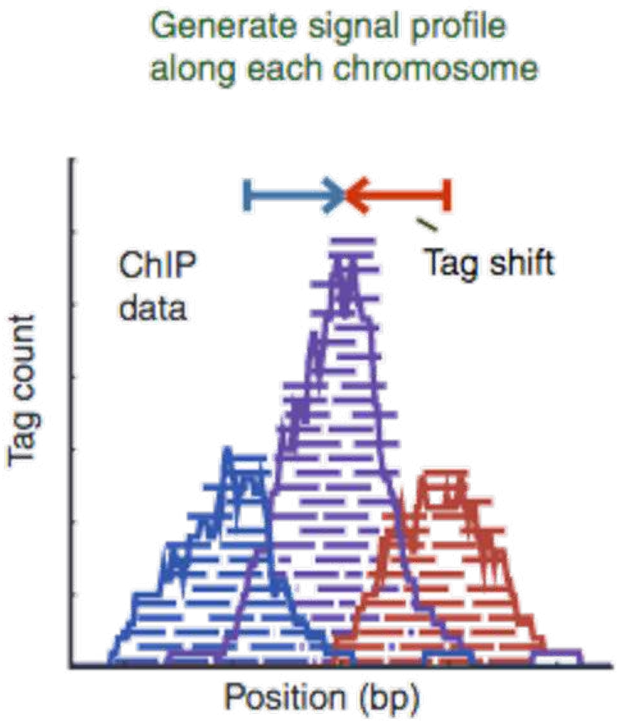
1. Modeling the shift size of ChIP-Seq tags

- 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]

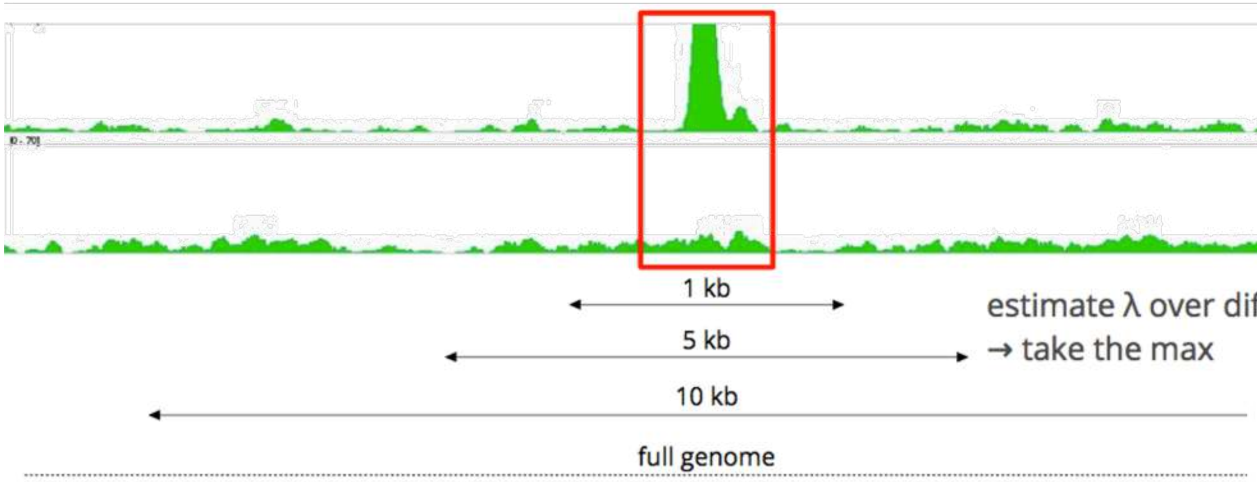
- **2. Peak detection**
  - Normalization: linearly scales the total control read count to be the same as the total ChIP read count
  - Duplicate read removal
  - Tags are shifted by *d*/2



Pepke et al, 2009

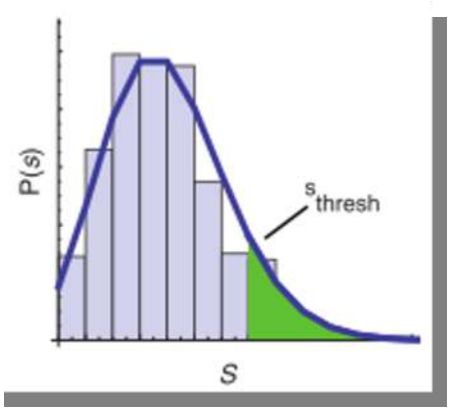
MACS [Zhang et al, 2008]

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



estimate λ over diff. ranges → take the max

Source:
C. Herrmann



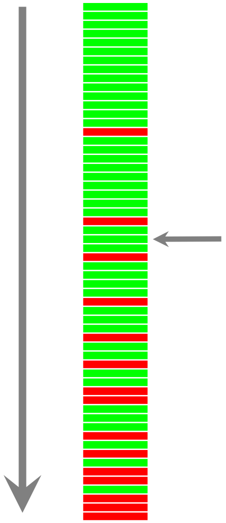
- Keep peaks significant under λ_{BG} and λ_{local} and with p-value

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

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



FDR = 2/27 = 0.074

Exercise 4: peak calling

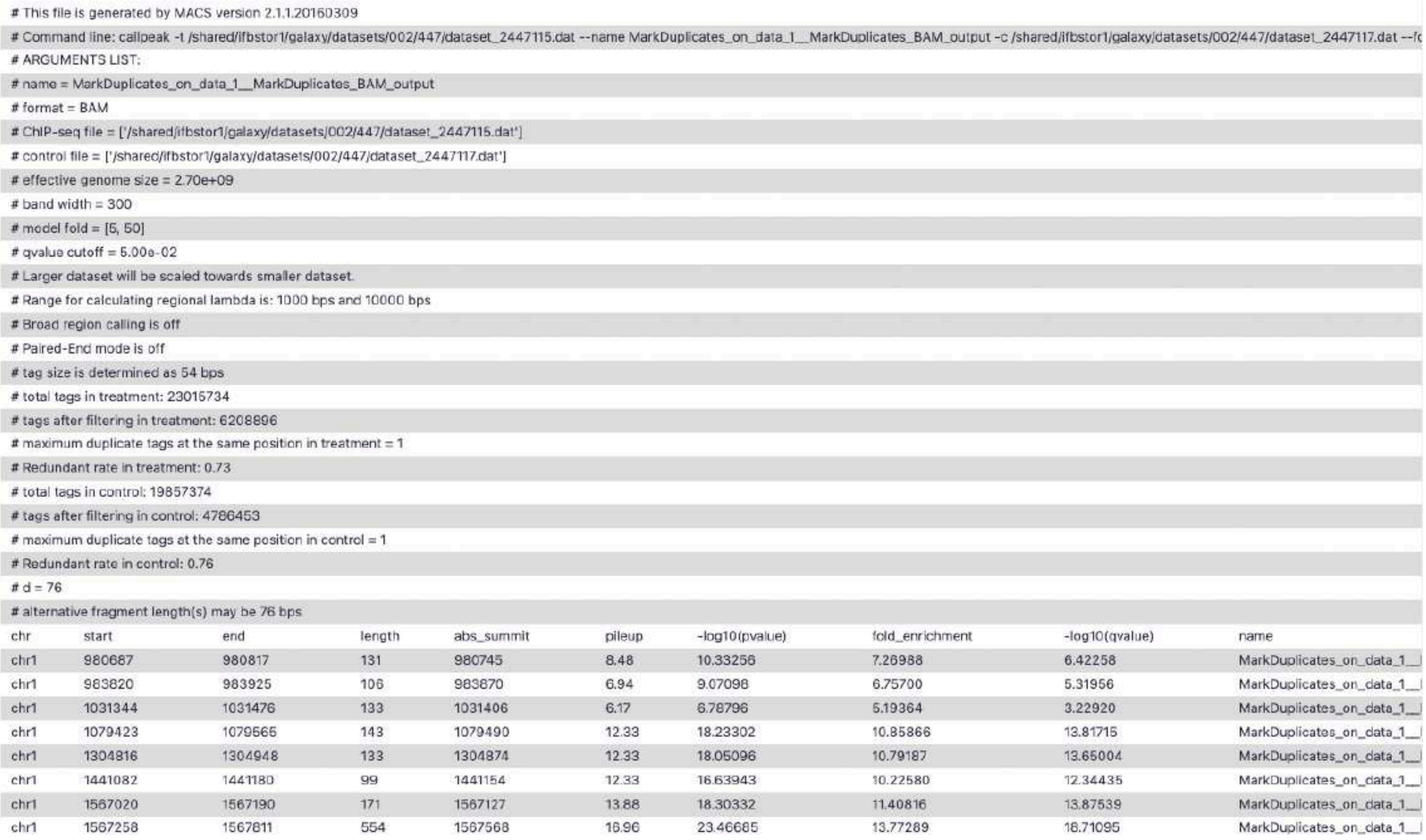
We now want to call MITF peaks.

- 1. Use **Macs2 callpeak** to perform the peak calling on the data. Use default parameters except for
  - **Are you pooling Treatment Files?** No
    - ChIP-Seq Treatment File: [mitf bam file marked by MarkDuplicates] *(1)*
  - **Do you have a Control File?** Yes
    - Are you pooling Control files? No
    - ChIP-Seq Control File: [control bam file marked by MarkDuplicates] *(2)*
  - Effective genome size: H.Sapiens (2.7e9)
  - Outputs:
    - Peaks as tabular file (compatible with MultiQC)
    - Peak summits
    - Summary page (html)
    - Plot in PDF (only available if a model is created and if BAMPE is not used)

Exercise 4: peak calling

Macs2 callpeak generates 5 datasets:

- List of the peaks (tabular format)



List of arguments used to run Macs2

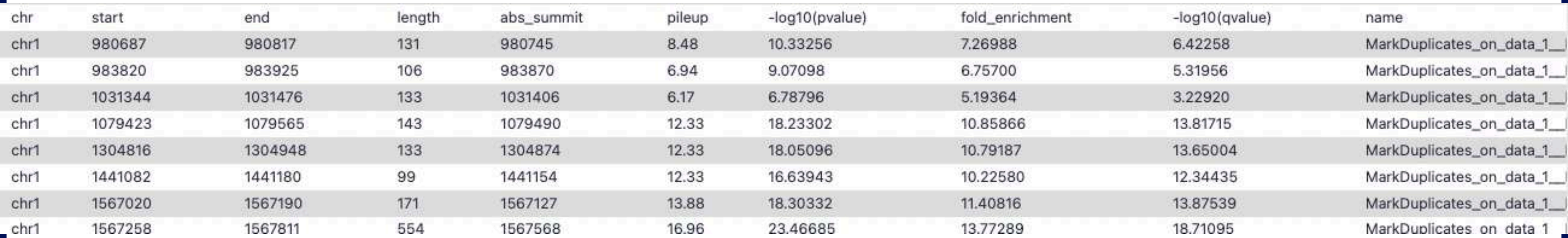
```
# This file is generated by MACS version 2.1.1.20160309
# Command line: callpeak -t /shared/ifbstor1/galaxy/datasets/002/447/dataset_2447115.dat --name MarkDuplicates_on_data_1__MarkDuplicates_BAM_output -c /shared/ifbstor1/galaxy/datasets/002/447/dataset_2447117.dat --fo
# ARGUMENTS LIST:
# name = MarkDuplicates_on_data_1__MarkDuplicates_BAM_output
# format = BAM
# ChIP-seq file = ['/shared/ifbstor1/galaxy/datasets/002/447/dataset_2447115.dat']
# control file = ['/shared/ifbstor1/galaxy/datasets/002/447/dataset_2447117.dat']
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off
# Paired-End mode is off
# tag size is determined as 54 bps
# total tags in treatment: 23015734
# tags after filtering in treatment: 6208896
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.73
# total tags in control: 19857374
# tags after filtering in control: 4786453
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.76
# d = 76
# alternative fragment length(s) may be 76 bps
```

Peaks

chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	980687	980817	131	980745	8.48	10.33256	7.26988	6.42258	MarkDuplicates_on_data_1__
chr1	983820	983925	106	983870	6.94	9.07098	6.75700	5.31956	MarkDuplicates_on_data_1__
chr1	1031344	1031476	133	1031406	6.17	6.78796	5.19364	3.22920	MarkDuplicates_on_data_1__
chr1	1079423	1079565	143	1079490	12.33	18.23302	10.85866	13.81715	MarkDuplicates_on_data_1__
chr1	1304816	1304948	133	1304874	12.33	18.05096	10.79187	13.65004	MarkDuplicates_on_data_1__
chr1	1441082	1441180	99	1441154	12.33	16.63943	10.22580	12.34435	MarkDuplicates_on_data_1__
chr1	1567020	1567190	171	1567127	13.88	18.30332	11.40816	13.87539	MarkDuplicates_on_data_1__
chr1	1567258	1567811	554	1567568	16.96	23.46685	13.77289	16.71095	MarkDuplicates_on_data_1__

Exercise 4: peak calling

- List of the peaks (tabular format)

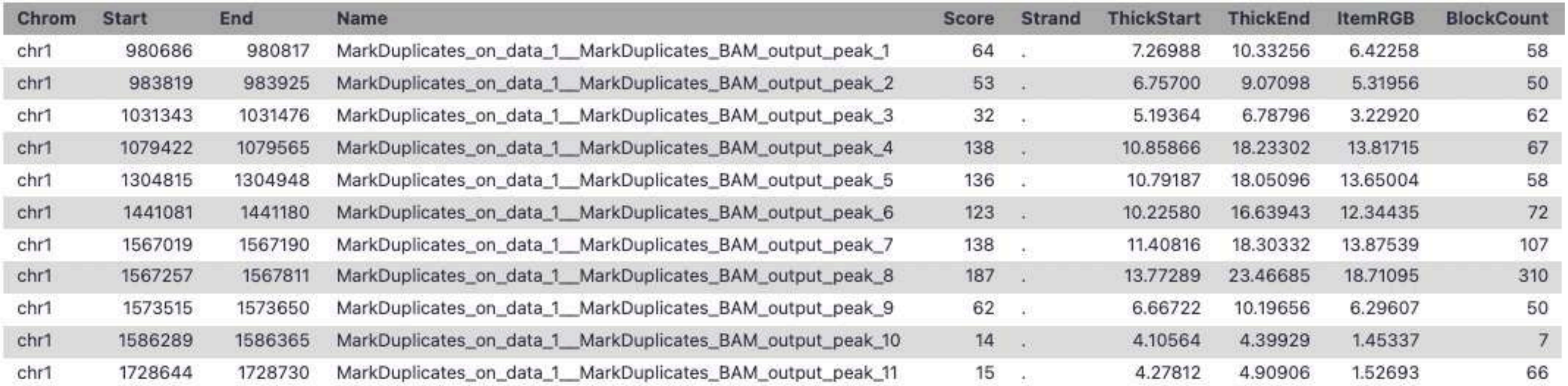


chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	980687	980817	131	980745	8.48	10.33256	7.26988	6.42258	MarkDuplicates_on_data_1__
chr1	983820	983925	106	983870	6.94	9.07098	6.75700	5.31956	MarkDuplicates_on_data_1__
chr1	1031344	1031476	133	1031406	6.17	6.78796	5.19364	3.22920	MarkDuplicates_on_data_1__
chr1	1079423	1079565	143	1079490	12.33	18.23302	10.85866	13.81715	MarkDuplicates_on_data_1__
chr1	1304816	1304948	133	1304874	12.33	18.05096	10.79187	13.65004	MarkDuplicates_on_data_1__
chr1	1441082	1441180	99	1441154	12.33	16.63943	10.22580	12.34435	MarkDuplicates_on_data_1__
chr1	1567020	1567190	171	1567127	13.88	18.30332	11.40816	13.87539	MarkDuplicates_on_data_1__
chr1	1567258	1567811	554	1567568	16.96	23.46685	13.77289	18.71095	MarkDuplicates_on_data_1

- 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)



Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount
chr1	980686	980817	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_1	64	.	7.26988	10.33256	6.42258	58
chr1	983819	983925	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_2	53	.	6.75700	9.07098	5.31956	50
chr1	1031343	1031476	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_3	32	.	5.19364	6.78796	3.22920	62
chr1	1079422	1079565	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_4	138	.	10.85866	18.23302	13.81715	67
chr1	1304815	1304948	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_5	136	.	10.79187	18.05096	13.65004	58
chr1	1441081	1441180	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_6	123	.	10.22580	16.63943	12.34435	72
chr1	1567019	1567190	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_7	138	.	11.40816	18.30332	13.87539	107
chr1	1567257	1567811	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_8	187	.	13.77289	23.46685	18.71095	310
chr1	1573515	1573650	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_9	62	.	6.66722	10.19656	6.29607	50
chr1	1586289	1586365	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_10	14	.	4.10564	4.39929	1.45337	7
chr1	1728644	1728730	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_11	15	.	4.27812	4.90906	1.52693	66

1. chr

2. Start of peak

3. End of peak

4. Peak name

5. Integer score for display

7. fold-change

8. -log10pvalue

9. -log10qvalue

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.

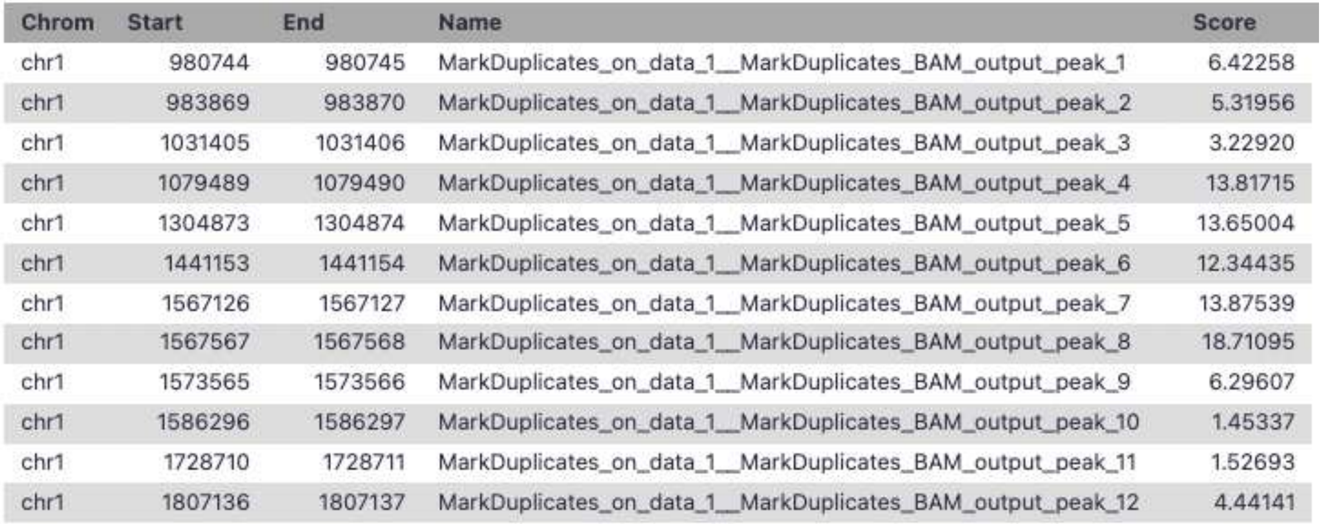
1. chr

2. Start of peak

3. End of peak

4. Peak name

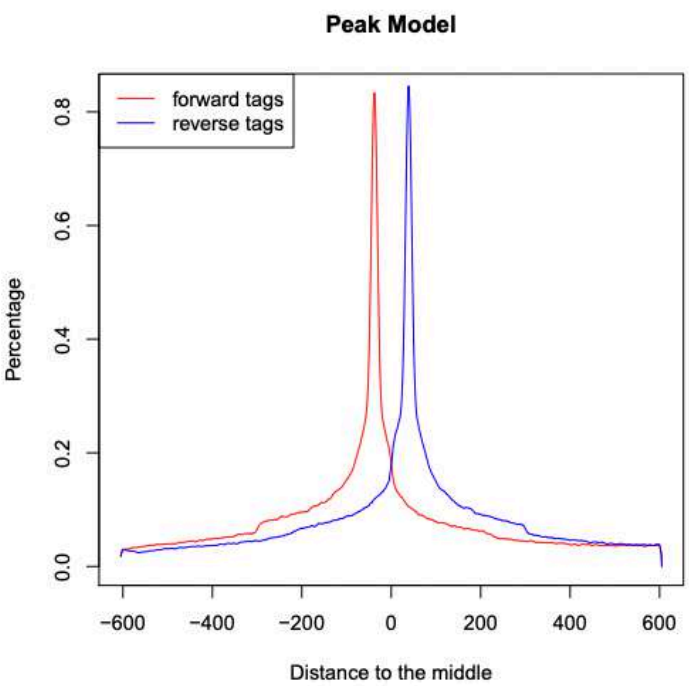
5. -log10pvalue

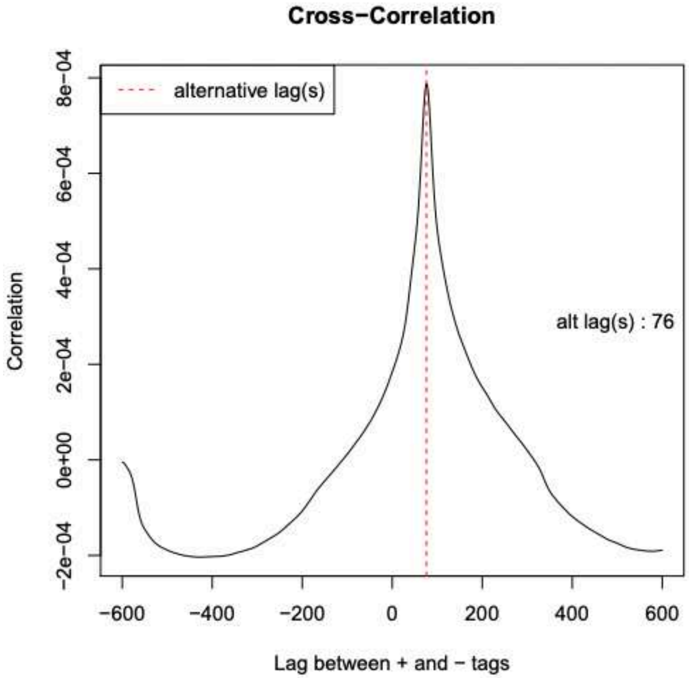


Chrom	Start	End	Name	Score
chr1	980744	980745	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_1	6.42258
chr1	983869	983870	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_2	5.31956
chr1	1031405	1031406	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_3	3.22920
chr1	1079489	1079490	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_4	13.81715
chr1	1304873	1304874	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_5	13.65004
chr1	1441153	1441154	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_6	12.34435
chr1	1567126	1567127	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_7	13.87539
chr1	1567567	1567568	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_8	18.71095
chr1	1573565	1573566	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_9	6.29607
chr1	1586296	1586297	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_10	1.45337
chr1	1728710	1728711	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_11	1.52693
chr1	1807136	1807137	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_12	4.44141

Exercise 4: peak calling

- PDF images about the model based on your data





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

Exercise 4: peak calling

- 2. Look at MACS2 results. How many peaks are found?
- 3. What is the fragment size estimated by Macs2? What do you think of the value?

Exercise 5: peak calling

- 1. Rerun **Macs2** using the same parameters as before but change the shift size:
  - Build Model: Do not build the shifting model (--nomodel)
  - The arbitrary extension: 200
- 2. How many peaks are now found?

Exercise 6: compare the two runs of MACS

To assess which peak calling is best, we are going to:

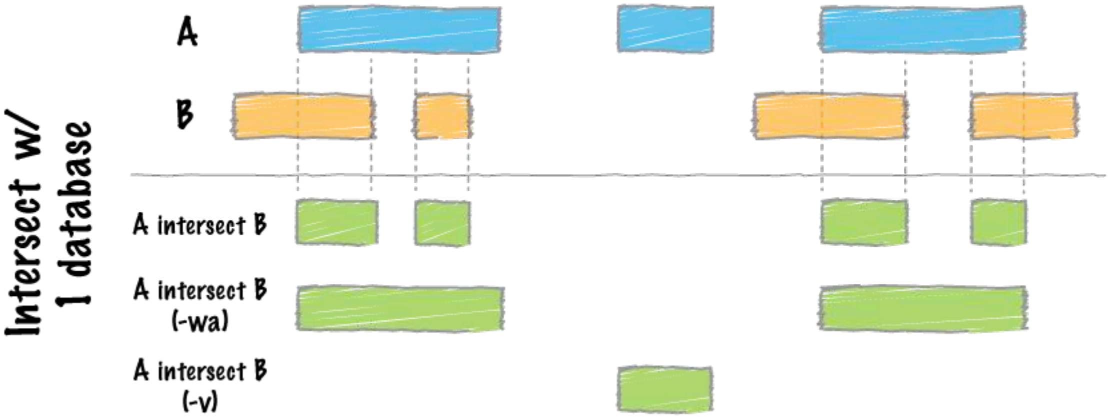
1. Extract regions that are unique to the first peak sets [Galaxy]
2. Look at peaks called in the two peak sets in a genome browser and check whether the peaks are fine [IGV]
3. Keep the best peak set

Exercise 6: compare the two runs of MACS

Bedtools is a collection of tools for genome arithmetic

- *intersect*, *merge*, *count*, *get closest, shuffle (…)* genomic intervals of one or multiple files
- Supported formats: BAM, BED, GFF/GTF, VCF
- https://bedtools.readthedocs.io/en/latest/

- Bedtools intersect



https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

Exercise 6: compare the two runs of MACS

1. Extract regions that are unique to the first peak sets

Use bedtools intersect (**bedtools Intersect intervals**) in Galaxy to extract peaks found in the first peak set and not in the second.

Parameters

- **File A to intersect with B: [**MACS2 callpeak on data * and data * (narrow Peaks)] *(1st run of MACS)*
- **Combined or separate output files**
    - One output per file 'input B' file
    - **File B to intersect with A: [**MACS2 callpeak on data * and data * (narrow Peaks)] *(2nd run of MACS)*
- **Calculation based on strandedness?** Overlaps on either strand
- **What should be written to the output file?** Write the original entry in A for each overlap (-wa)
- **Report only those alignments that **do not** overlap the BED file:** Yes

How many regions are found only in the first run of MACS?

Exercise 6: compare the two runs of MACS

2. Look at peaks called in the two peak sets in a genome browser a check whether the peaks are fine

   1. Download the narrowpeak files of the two runs of MACS
   2. Load in IGV :
      1. mitf.tdf (folder chipseq/visualization)
      2. ctrl.tdf (folder chipseq/visualization)
      3. [MACS2 callpeak on data * and data * (**narrow Peaks**)] (1st run of MACS)
      4. [MACS2 callpeak on data * and data * (**narrow Peaks**)] (2nd run of MACS)
   3. Look at the dataset resulting from Bedtools intersect and check at genomic locations found in the file
      1. Look at the peaks in the gene SSU72 (chr1:1556527-1578211)
      2. Look at the peak in the gene HIVEP3 (chr1:41882599-41882681)
      3. Look at the peak in the region chr1:1586290-1586365

Would you keep peaks found in the 1st run of MACS or the 2nd run of MACS?

For select MACS2 results, rename the datasets:

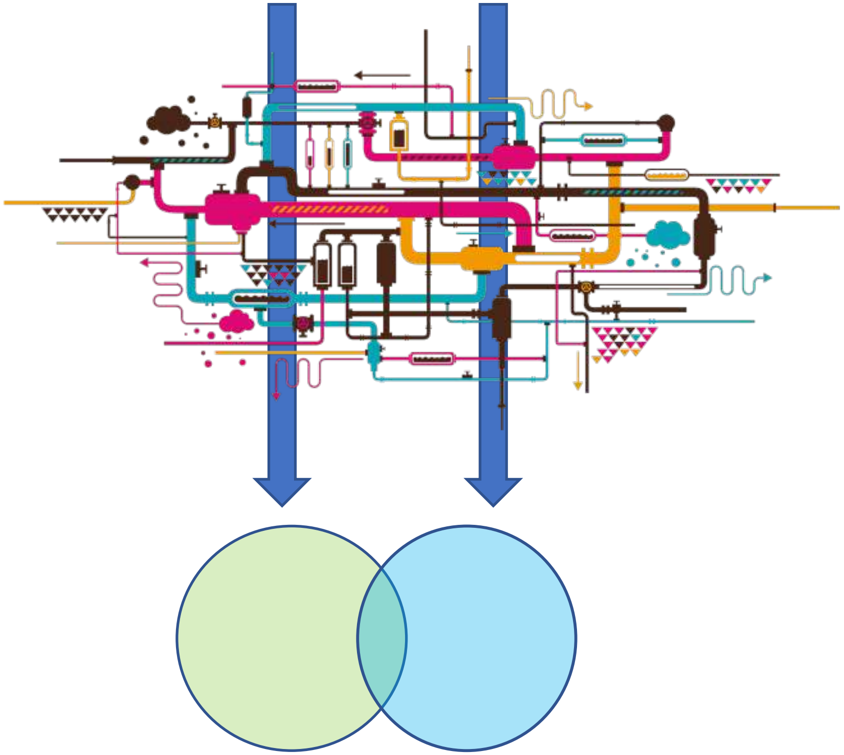
- [MACS2 callpeak on data * and data * (summits in BED)] -> MITF_peak_summits.bed
- [MACS2 callpeak on data * and data * (narrow Peaks)] -> MITF_peaks.narrowPeak

How to deal with replicates

Analyze samples separately and takes union or intersection of resulting peaks

Sample 1.a

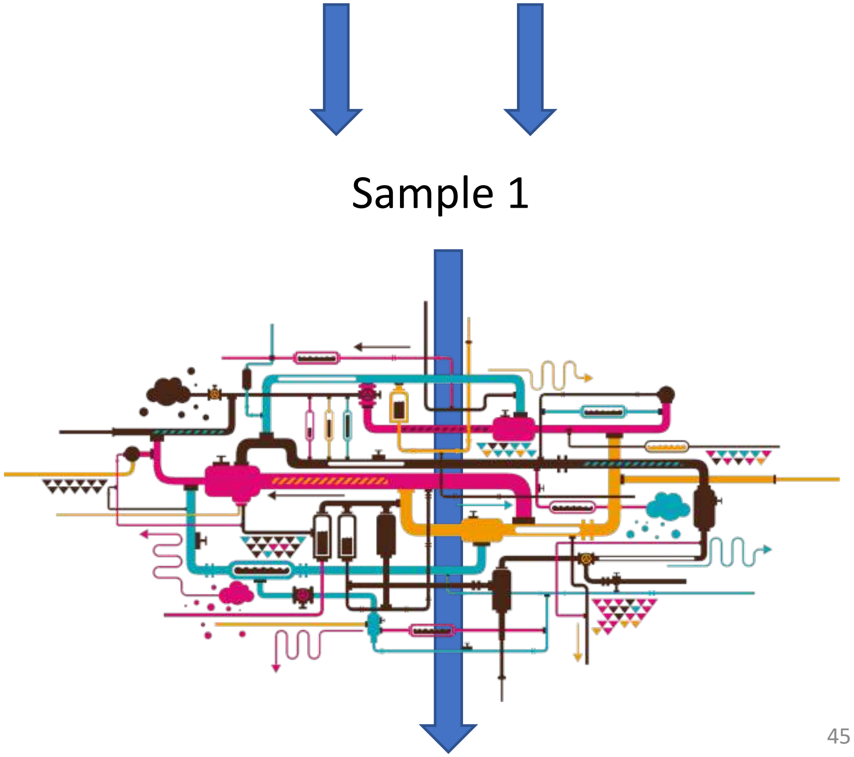
Sample 1.b



Merge samples prior to the peak calling (e.g recommended by MACS)

Sample 1.a

Sample 1.b



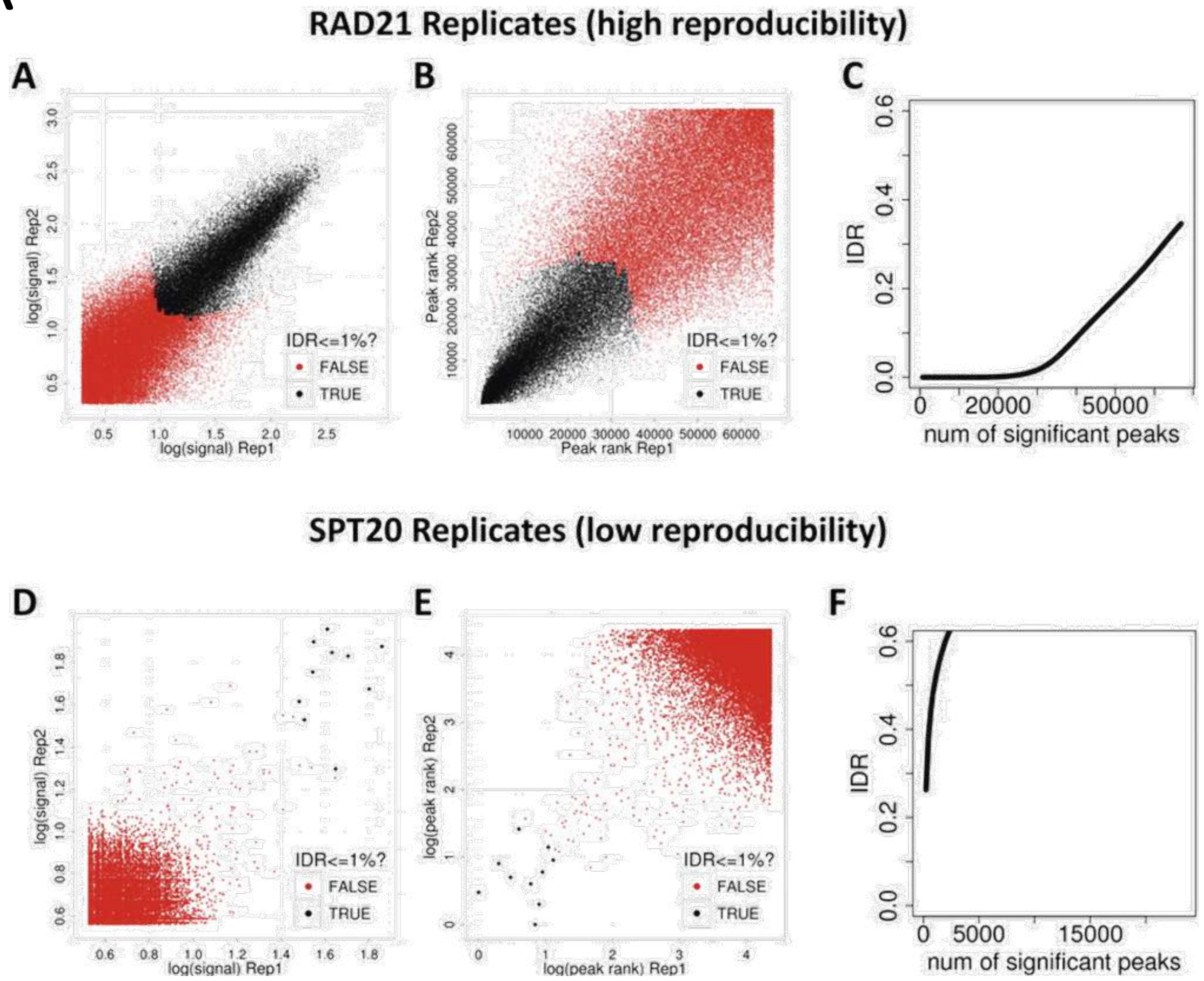
Sample 1

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

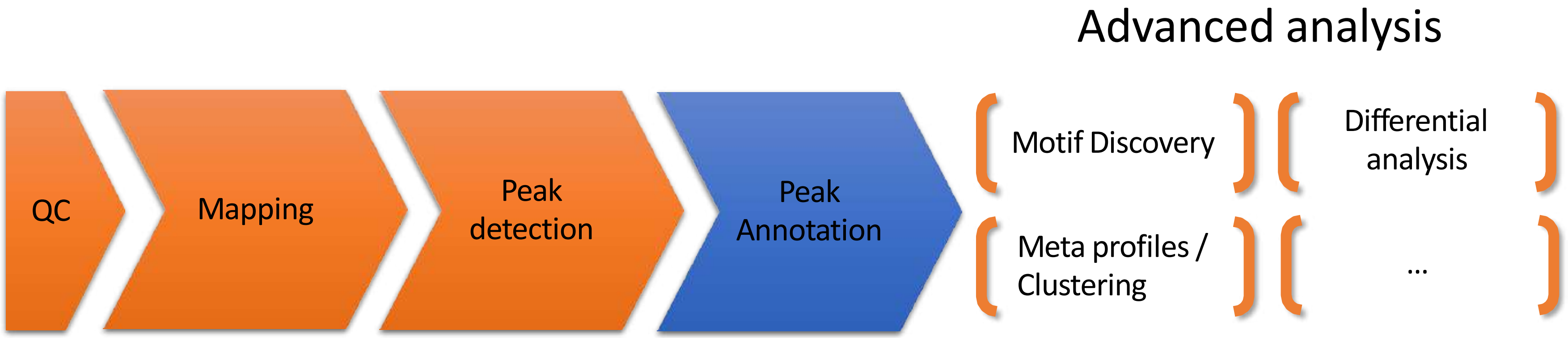
https://sites.google.com/site/anshulkundaje/projects/idr
https://github.com/ENCODE-DCC/chip-seq-pipeline2

IDR



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

Analysis of ChIP-seq data



Advanced analysis

QC

Mapping

Peak detection

Peak Annotation

Motif Discovery

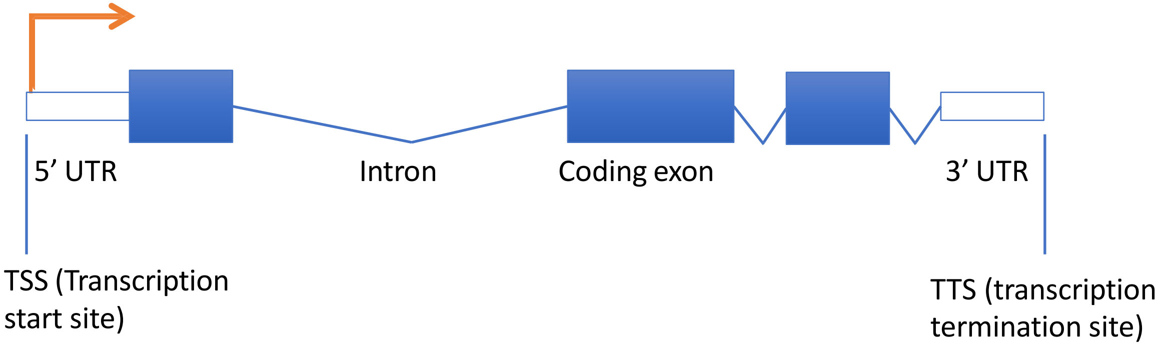
Differential analysis

Meta profiles / Clustering

…

Peak annotation

- Goal: assigning a peak to one or many genome features (genes/transcripts) to understand which genes are possibly regulated by the binding of the protein of interest
- The name of the gene is important as well as the genic region where the peak is located
- Example of Homer tools:
  - Determines the distance to the nearest Transcription Start Site (TSS) and assigns the peak to that gene
  - Determines the genomic annotation of the region **occupied by the center** of the peak/region. Possible genomic annotation:



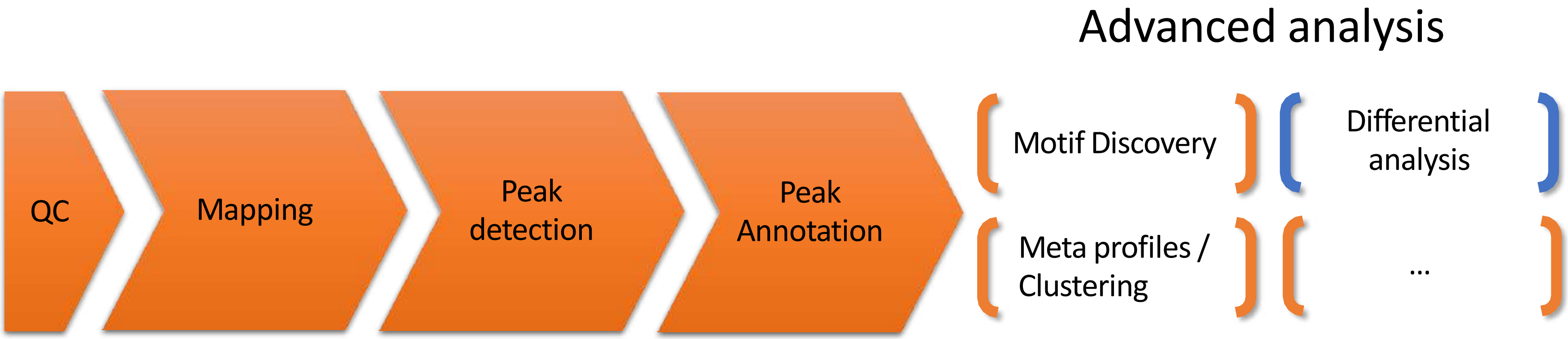
5’ UTR
Intron
Coding exon
3’ UTR
TSS (Transcription start site)
TTS (transcription termination site)

Exercise 7: peak annotation

Most of the peak annotation tools assign peaks to the closest gene. Use the tool **bedtools ClosestBed** to find the closest gene for each detected peak.

- Import to your current history the dataset 25:hg38_ens105_ucsc.bed from the imported history « NGS data analysis training Strasbourg ».
- Then, Here are the parameters to use:
  - **BED/bedGraph/GFF/VCF/EncodePeak file:** MITF_peaks.narrowPeak *(second run of MACS2)*
  - **Overlap with: will you select a BED/bedGraph/GFF/VCF/EncodePeak file from your history or use a built-in GFF file?**
    - Use a BED/bedGraph/GFF/VCF/EncodePeak file from the history
    - **Select a BED/bedGraph/GFF/VCF/EncodePeak file:** hg38_ens105_ucsc.bed
  - **How ties for closest feature should be handled:** first – Report the first tie that occurred in the B file
  - **In addition to the closest feature in B, report its distance to A as an extra column:** Yes
  - **Add additional columns to report distance to upstream feature. Distance defintion:**
    - Report distance with respect to A. When A is on the – strand, « upstream » means B has a higher (start,stop). (-a)
    - **Choose first from features in B that are upstream of feature in A:** Yes
- Rename the file: mitf_peaks.annot.tsv.

Analysis of ChIP-seq data



QC

Mapping

Peak detection

Peak Annotation

Advanced analysis

Motif Discovery

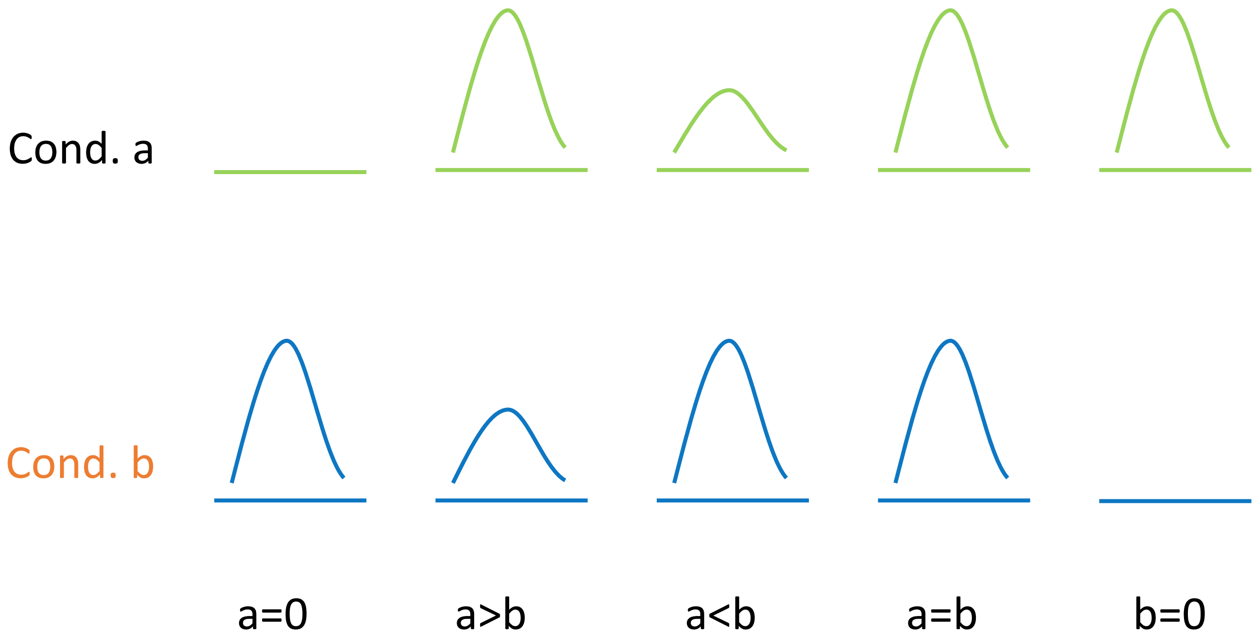
Differential analysis

Meta profiles / Clustering

…

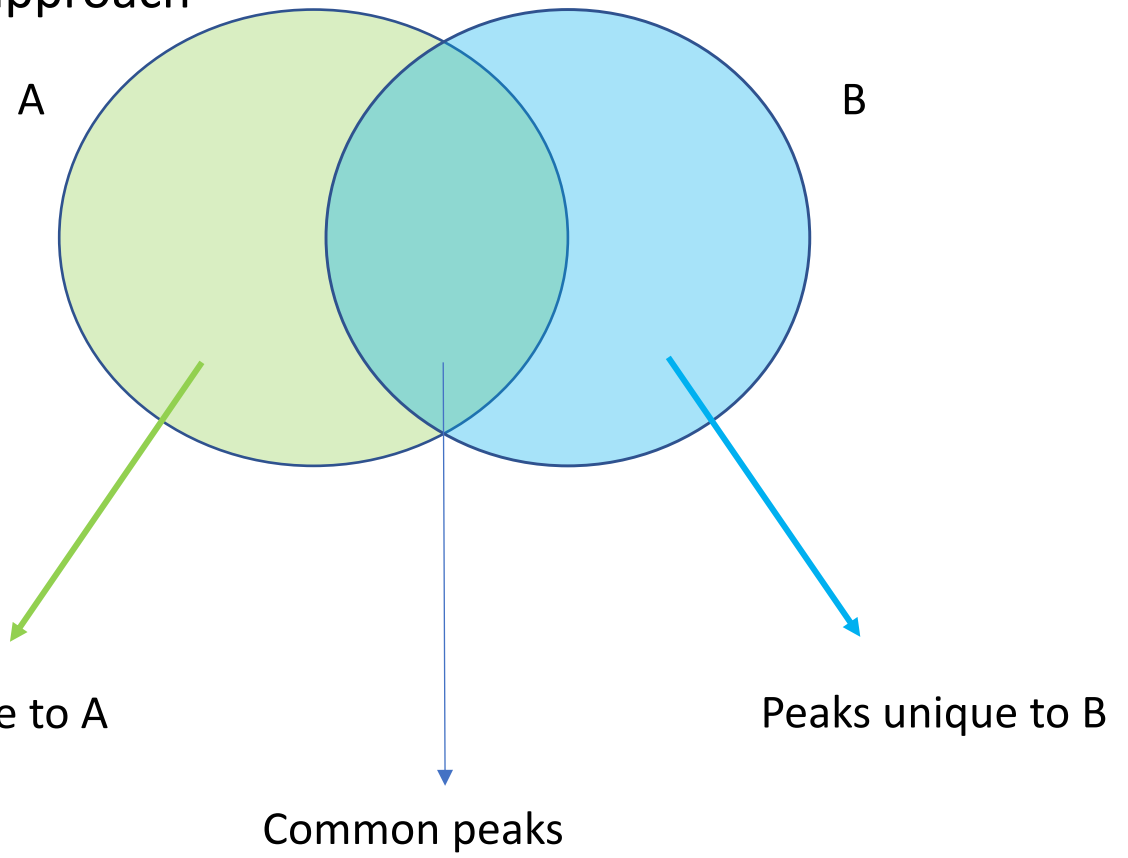
Differential binding analysis

- 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



Peaks unique to A

Peaks unique to B

Common peaks

Differential binding analysis

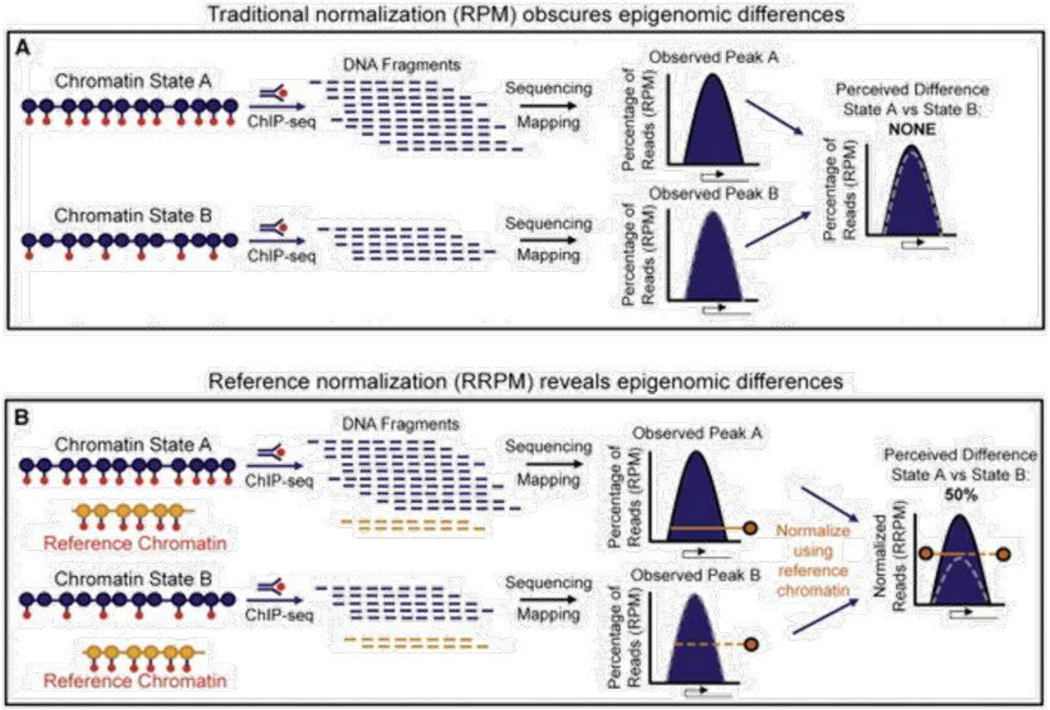
Quantitative approach

- 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

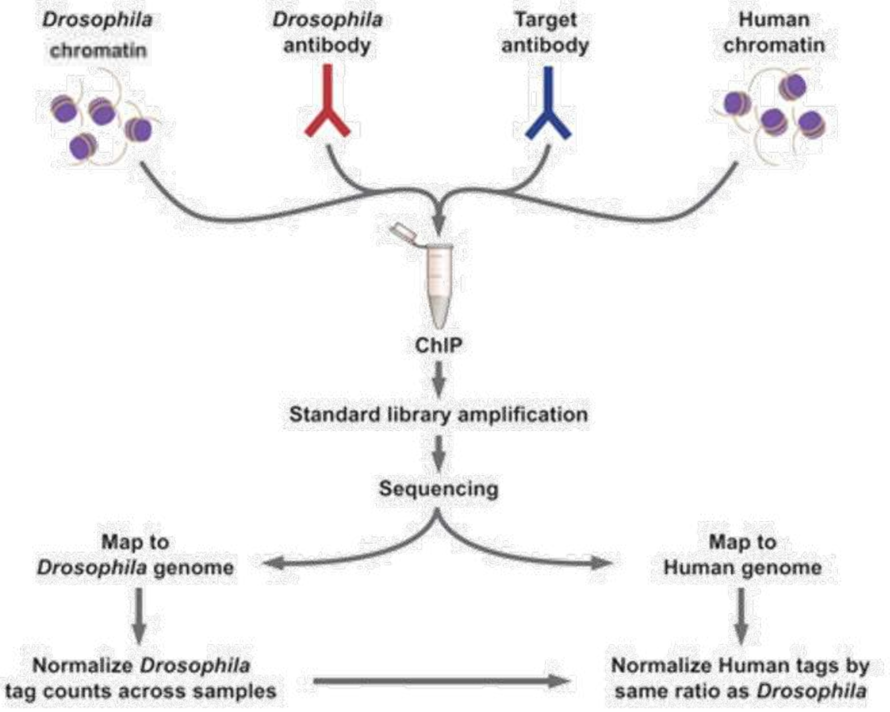
- Current normalization methods fail to detect global changes as they make the assumption that globally nothing change but a small portion of observations
- Insert external chromatin used as reference chromatin


Traditional normalization (RPM) obscures epigenomic differences
A
Chromatin State A
ChIP-seq
DNA Fragments
Sequencing
Mapping
Observed Peak A
Percentage of Reads (RPM)
Perceived Difference State A vs State B:
NONE
Chromatin State B
Observed Peak B
Reference normalization (RRPM) reveals epigenomic differences
B
Reference Chromatin
Normalize using reference chromatin
Normalized Reads (RRPM)
Perceived Difference State A vs State B:
50%

Orlando et al, 2014

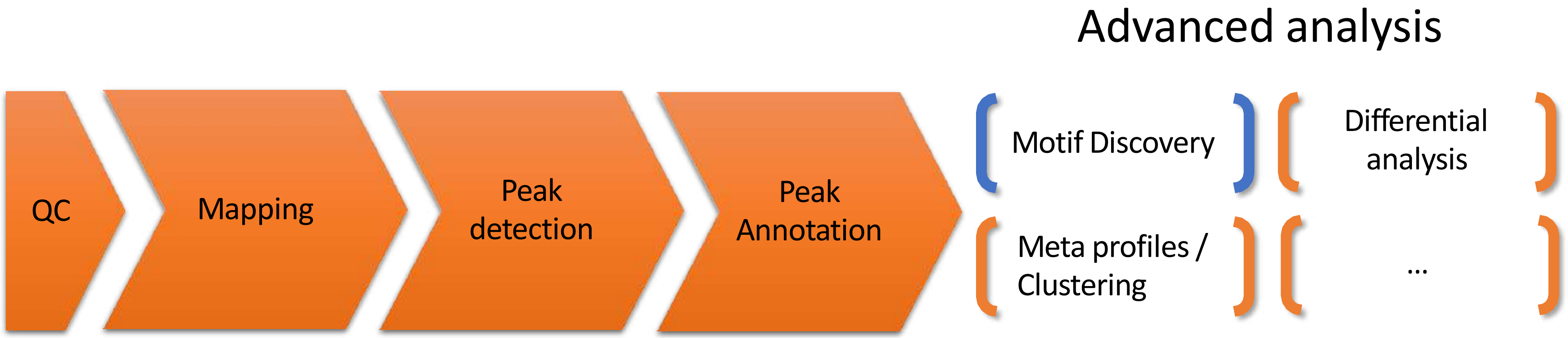
Spike-in

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


Drosophila chromatin
Drosophila antibody
Target antibody
Human chromatin
ChIP
Standard library amplification
Sequencing
Map to *Drosophila* genome
Map to Human genome
Normalize *Drosophila* tag counts across samples
Normalize Human tags by same ratio as *Drosophila*

http://www.activemotif.com/catalog/1091/chip-normalization

ChIP-seq data analysis



Advanced analysis

QC

Mapping

Peak detection

Peak Annotation

Motif Discovery

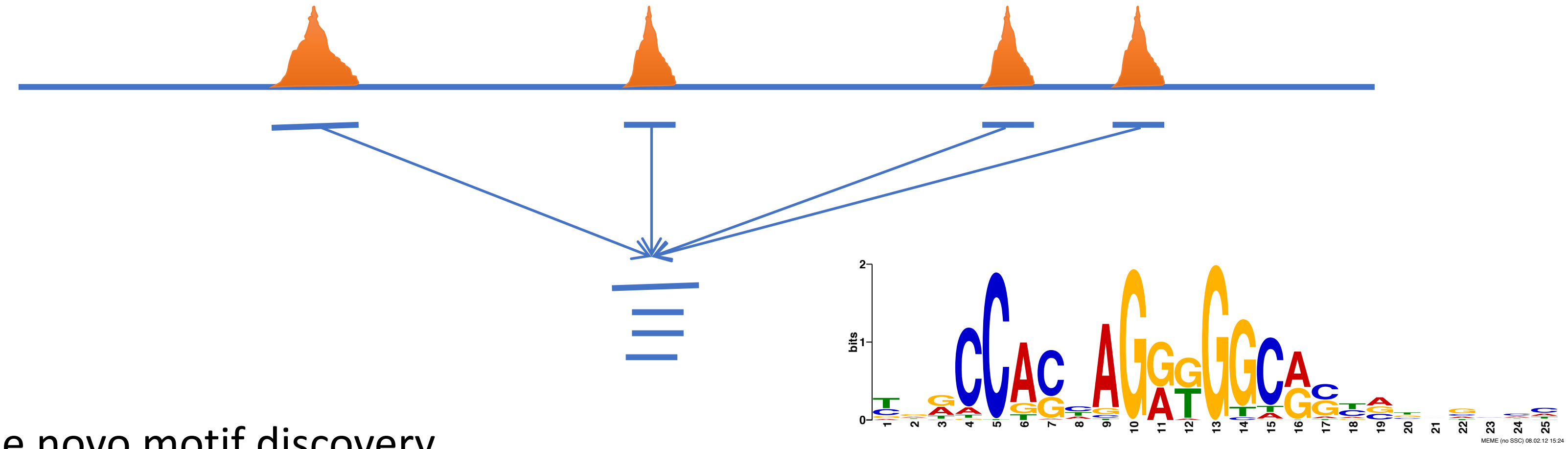
Differential analysis

Meta profiles / Clustering

…

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 discovery

- Lot of tools exist (Homer, RSAT, MEME-suite…)
- 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 8: *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 -log10pvalue) [Galaxy]
  - 1.a. Sort the peak **summits (**MITF_peak_summits.bed) by decreasing -log10pvalue using the tool **Sort**
  - 1.b. Extract the top 800 peak summits using the tool **Select first** on sorted peak summits

Tip: we limit the analysis to the first top 800 peaks to speed up the analysis and to increase the probability to have true positive peaks and thus to have peaks with motifs

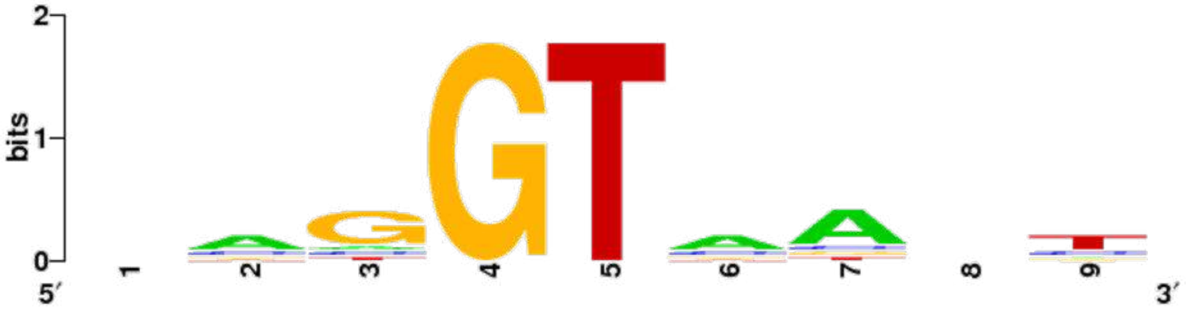
Exercise 8: *de novo* motif discovery

- 2. In Galaxy, compute the coordinates of the peak summits +/- 50nt using the dataset which contains the top 800 MITF peak **summits** (2nd run of Macs2) using the tool **SlopBed.**
  - Hint: use a genome locally installed (hg38)
  - Hint 2: you want to extend genomic coordinates in each direction
- 3. Extract fasta sequences from the coordinates of the peak summits using the tool **bedtools GetFastaBed.** Rename the dataset peakSummits_+/-50nt_top800.fasta.
- 4. Download the file peakSummits_+/-50nt_top800.fasta, 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)**

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

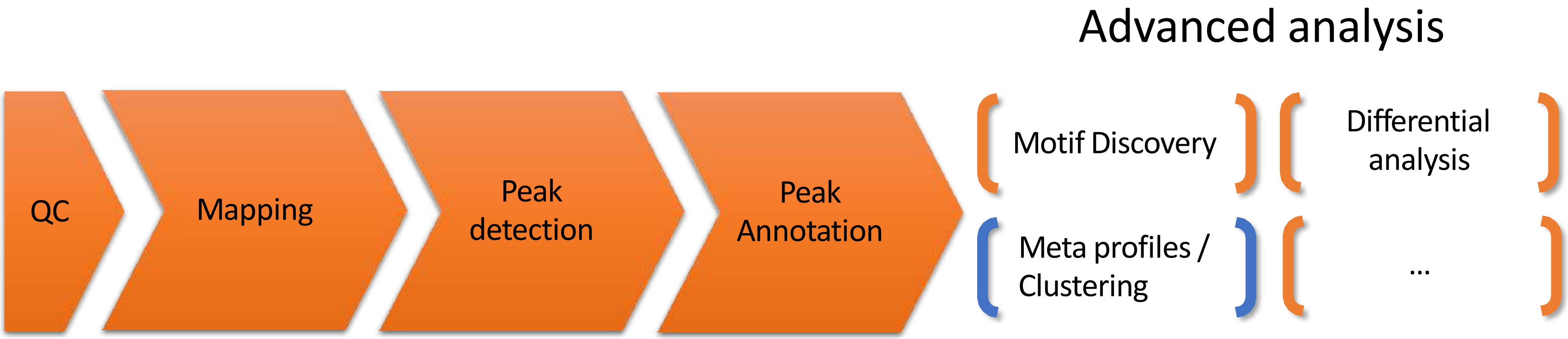


http://weblogo.berkeley.edu/logo.cgi

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.

Analysis of ChIP-seq data

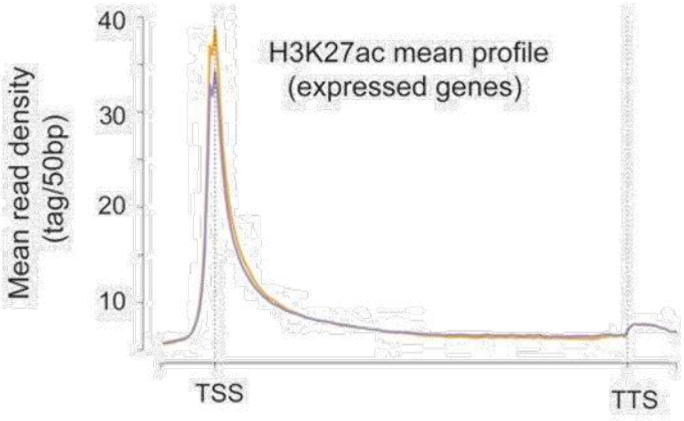


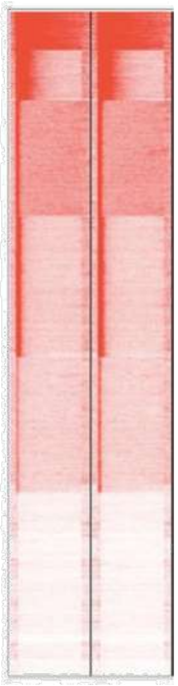
Meta-profiles

- 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)

Heatmap

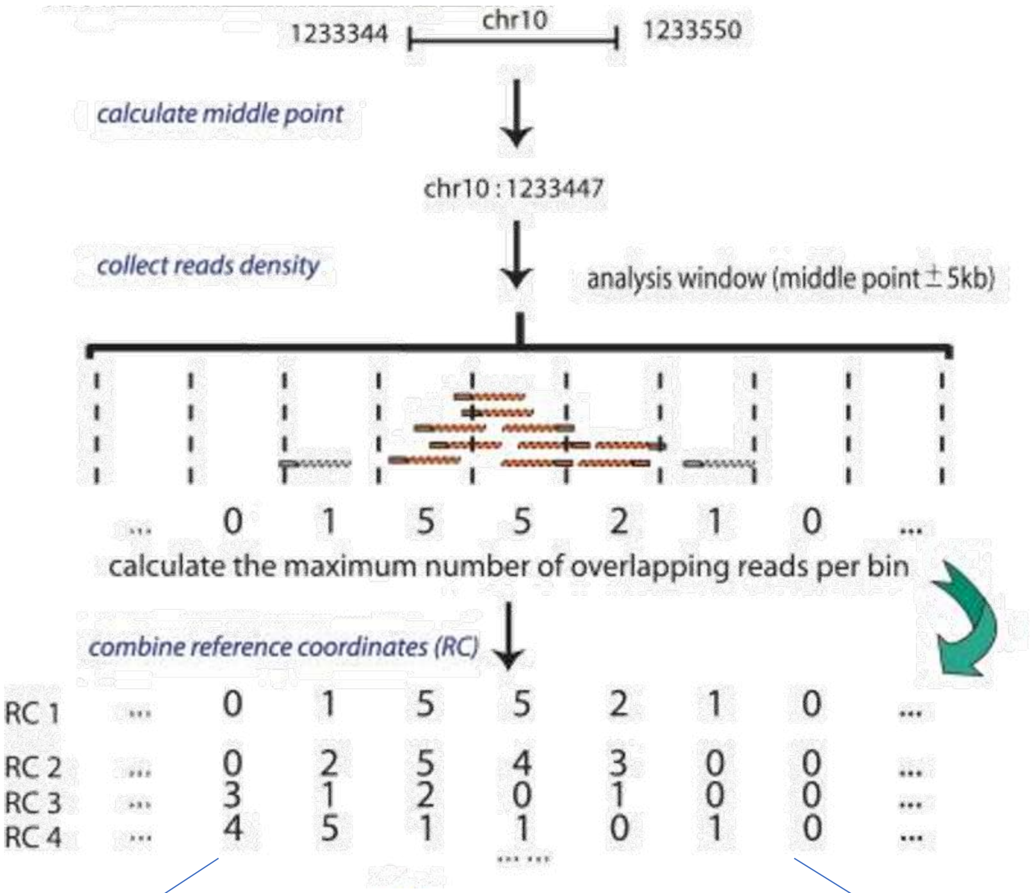
Mean profile





Computing meta-profiles

Reference coordinates e.g peaks


1233344
chr10
1233550
calculate middle point
chr10 : 1233447
collect reads density
analysis window (middle point ± 5kb)
... 0 1 5 5 2 1 0 ...
calculate the maximum number of overlapping reads per bin
combine reference coordinates (RC)

RC 1	...	0	1	5	5	2	1	0	...
RC 2	...	0	2	5	4	3	0	0	...
RC 3	...	3	1	2	0	1	0	0	...
RC 4	...	4	5	1	1	0	1	0	...
					... ...				

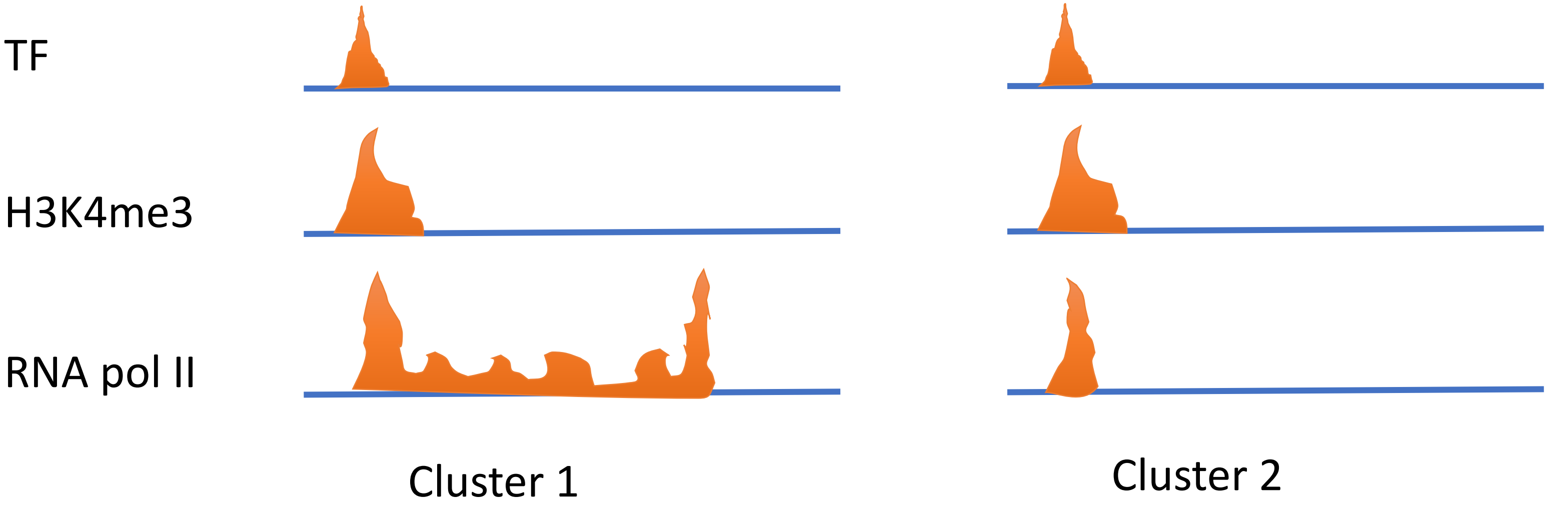
Ye et al, 2011

- (Clustering)
- Heatmap

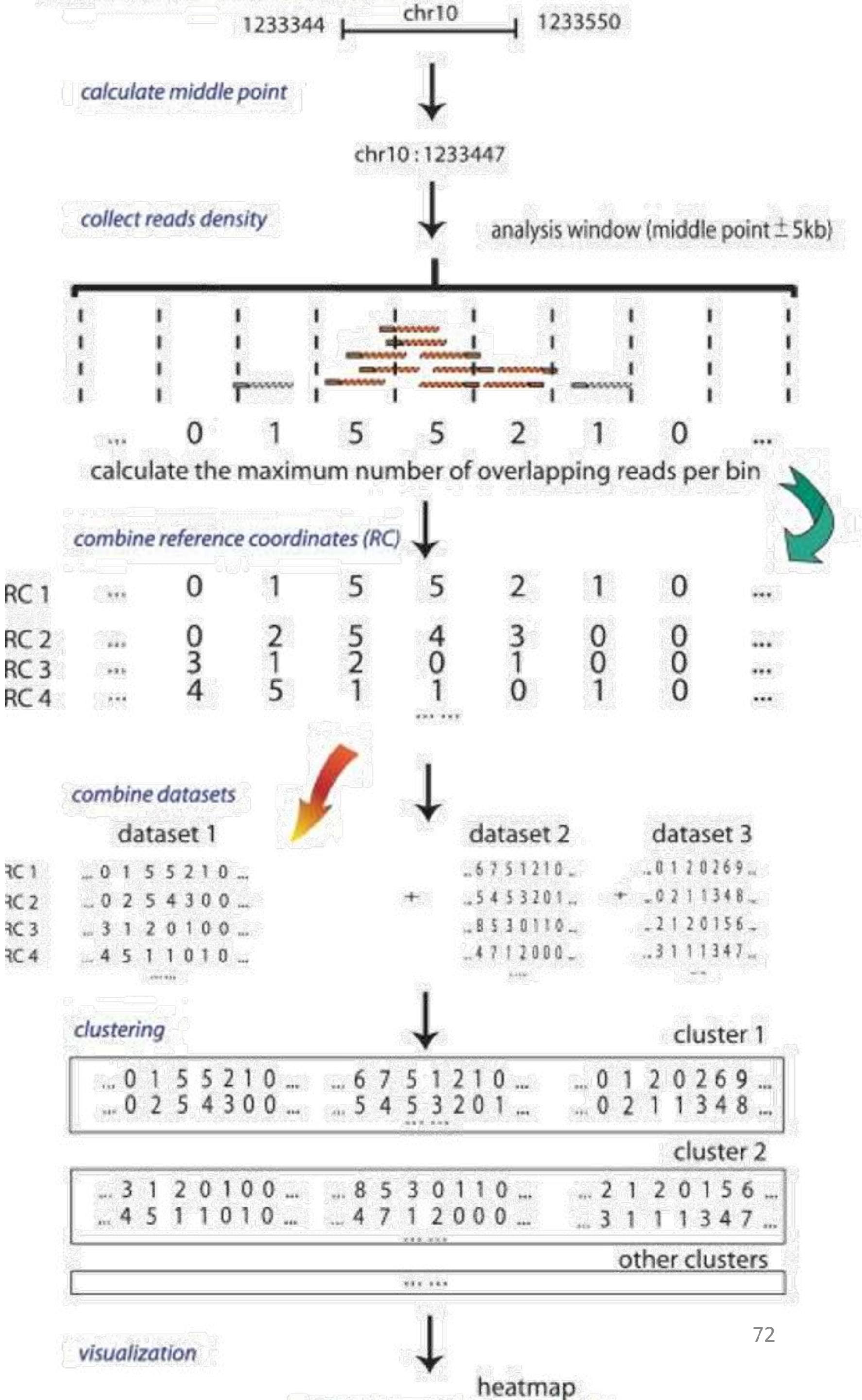
- Mean of each column -> Mean profile

Heatmap (clustering)

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



Heatmap (clustering)



1233344 chr10 1233550

calculate middle point

chr10 : 1233447

collect reads density

analysis window (middle point ± 5kb)

... 0 1 5 5 2 1 0 ...

calculate the maximum number of overlapping reads per bin

combine reference coordinates (RC)

RC 1 ... 0 1 5 5 2 1 0 ...

RC 2 ... 0 2 5 4 3 0 0 ...

RC 3 ... 3 1 2 0 1 0 0 ...

RC 4 ... 4 5 1 1 0 1 0 ...

combine datasets

	dataset 1		dataset 2		dataset 3
RC 1	... 0 1 5 5 2 1 0 ...		... 6 7 5 1 2 1 0 ...		... 0 1 2 0 2 6 9 ...
RC 2	... 0 2 5 4 3 0 0 ...	+	... 5 4 5 3 2 0 1 ...	+	... 0 2 1 1 3 4 8 ...
RC 3	... 3 1 2 0 1 0 0 ...		... 8 5 3 0 1 1 0 ...		... 2 1 2 0 1 5 6 ...
RC 4	... 4 5 1 1 0 1 0 ...		... 4 7 1 2 0 0 0 ...		... 3 1 1 1 3 4 7 ...

clustering

cluster 1

... 0 1 5 5 2 1 0 ... ... 6 7 5 1 2 1 0 ... ... 0 1 2 0 2 6 9 ...

... 0 2 5 4 3 0 0 ... ... 5 4 5 3 2 0 1 ... ... 0 2 1 1 3 4 8 ...

cluster 2

... 3 1 2 0 1 0 0 ... ... 8 5 3 0 1 1 0 ... ... 2 1 2 0 1 5 6 ...

... 4 5 1 1 0 1 0 ... ... 4 7 1 2 0 0 0 ... ... 3 1 1 1 3 4 7 ...

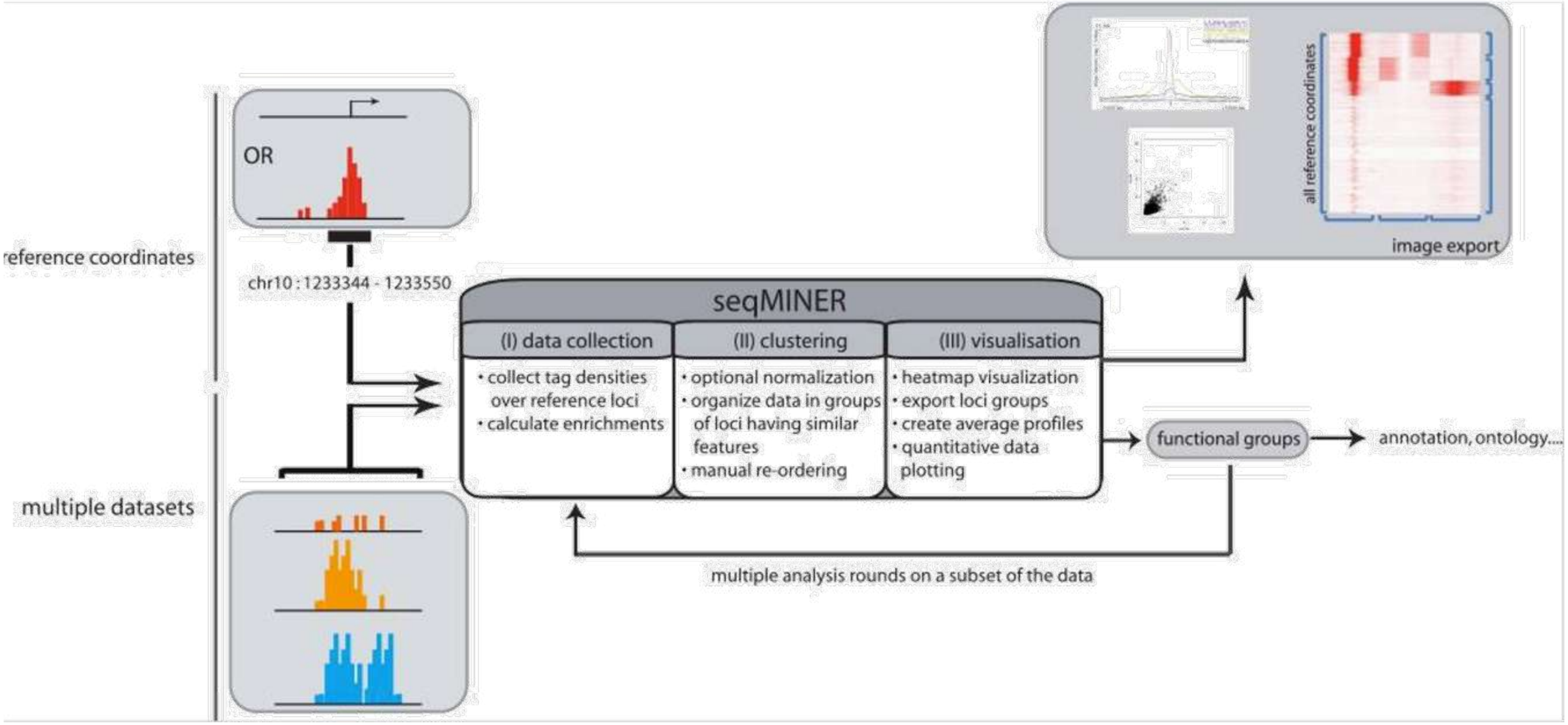
other clusters

... ...

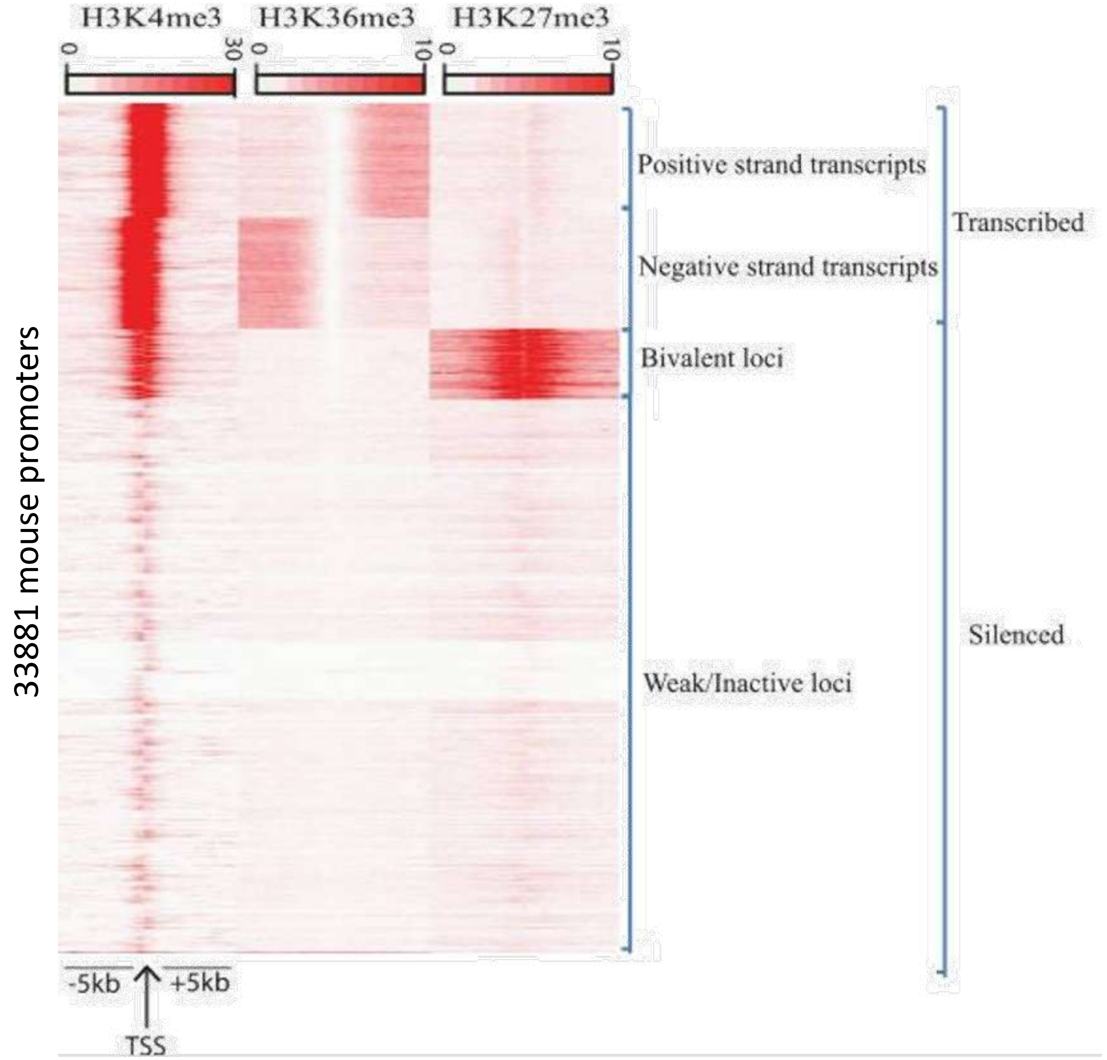
visualization

heatmap

SeqMINER [Ye et al, 2011]

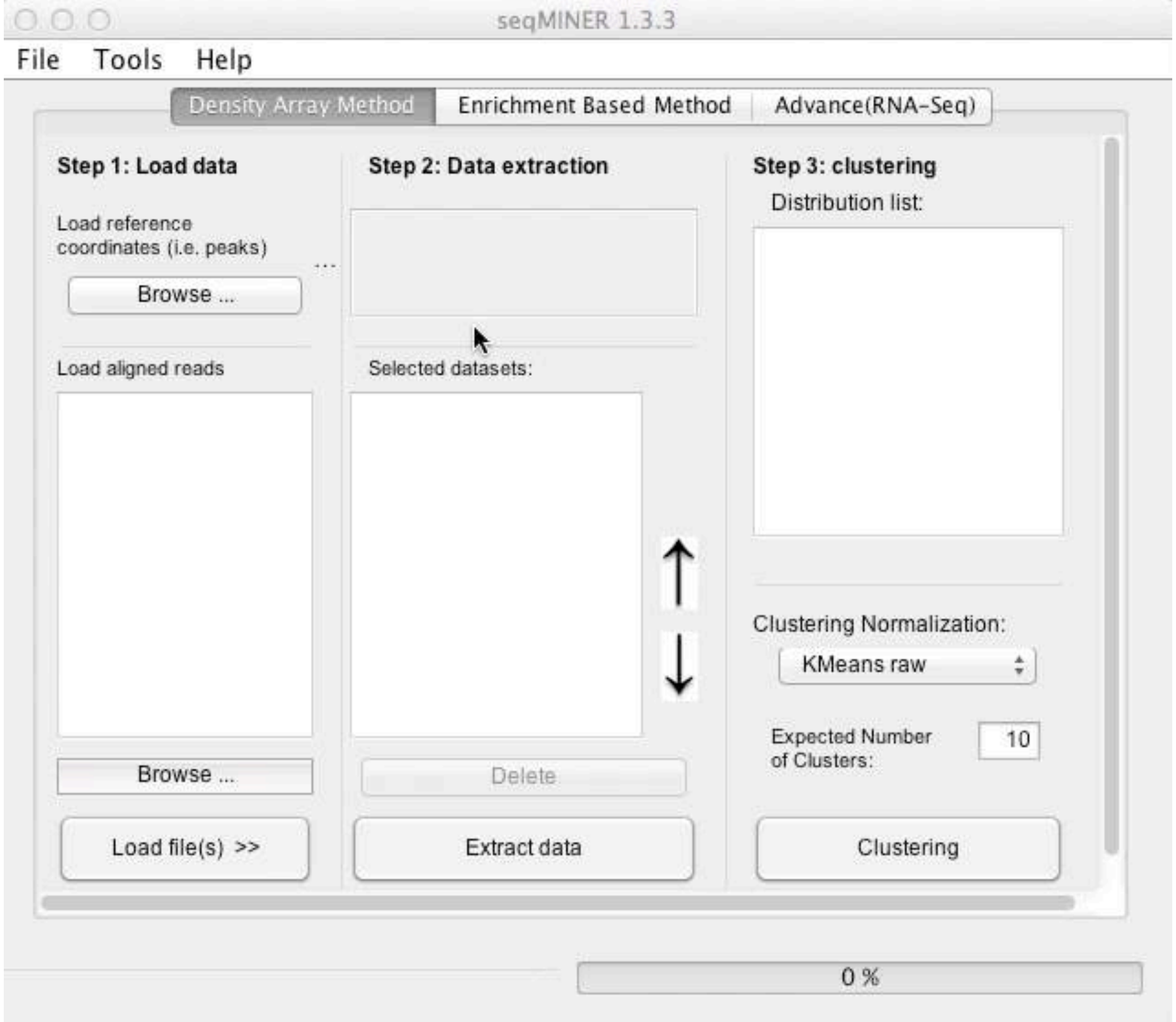


SeqMINER [Ye et al, 2011]



The darker the red the higher the read enrichment

Example



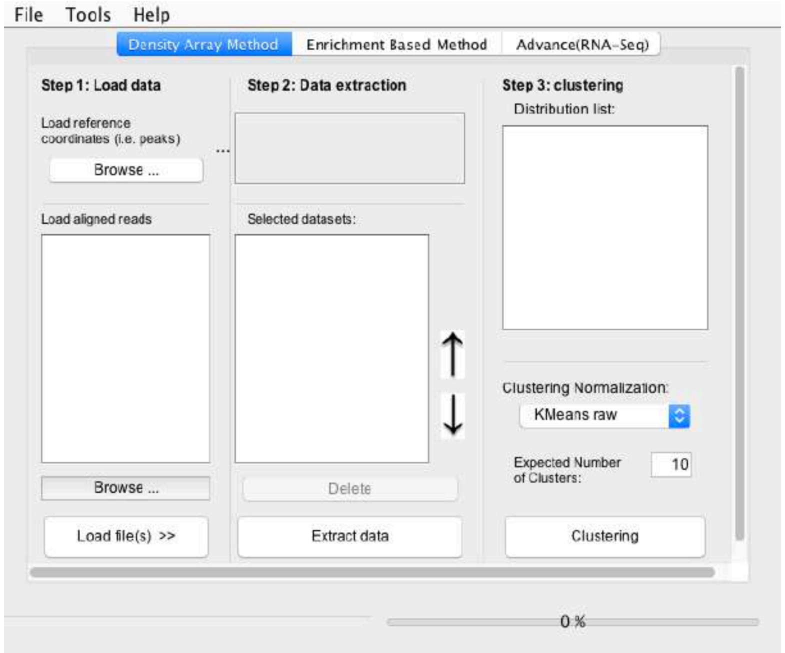
Clustering Normalization: KMeans raw

Expected Number of Clusters: 10

Exercise 9: Clustering

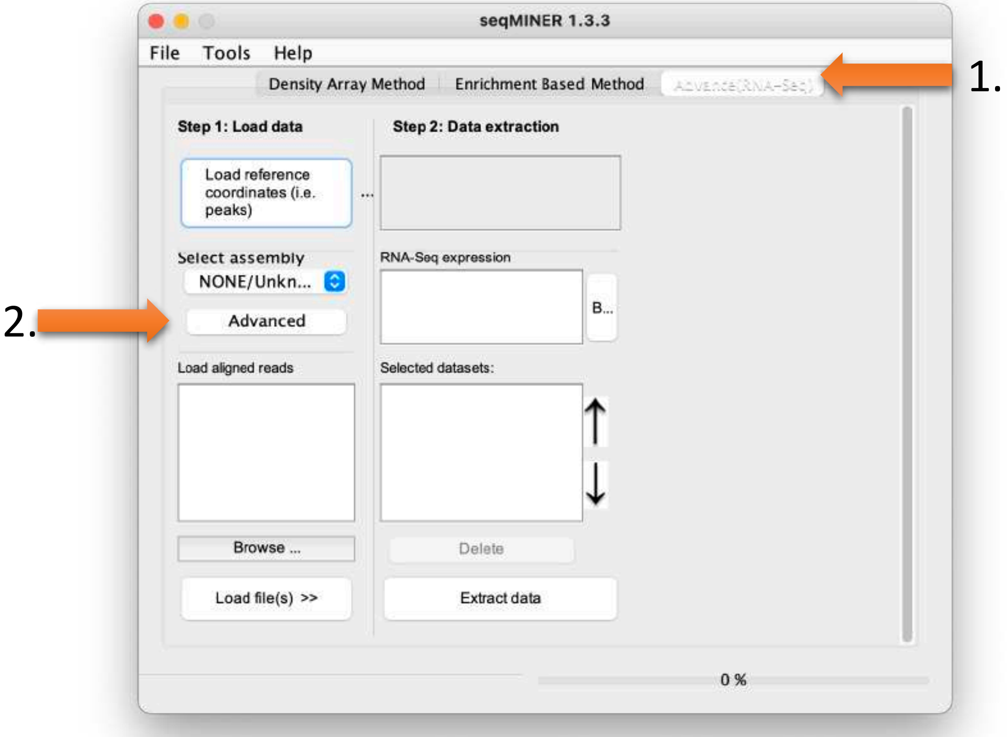
We have 2 additional datasets to those of MITF and the control : H3K4me3 and polII. Use seqMINER to have a look at the correlation between MITF, H3K4me3 and polII.

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 for Windows users or run_in_mac.command for Mac users.

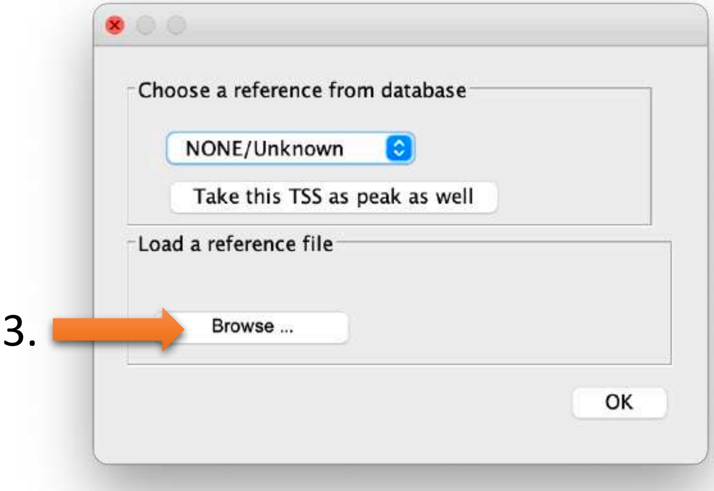


Exercise 9: Clustering

- 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) (1)
  - Click on Advanced (2)



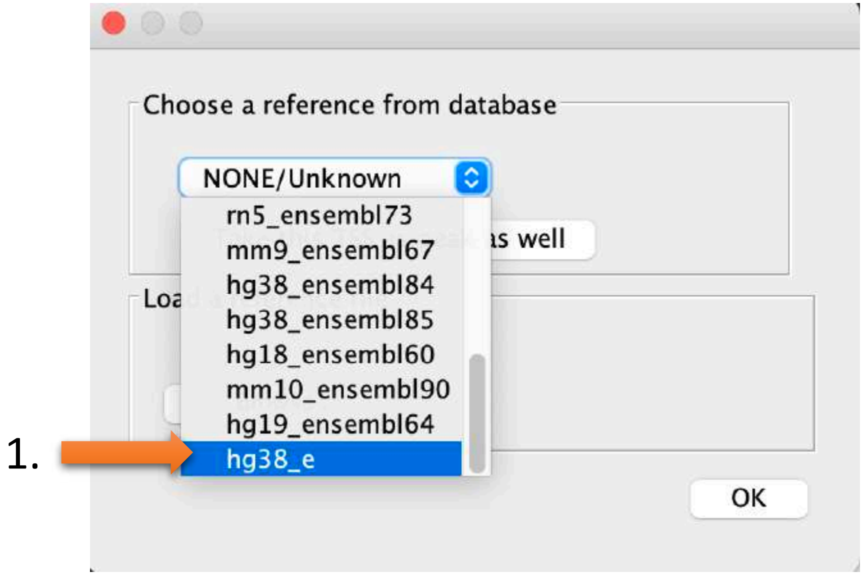
Click on Browse (3), and select the file extracted from Ensembl/Biomart hg38_ens105.bed



Choose a reference from database
NONE/Unknown
Take this TSS as peak as well
Load a reference file
3.
Browse ...
OK

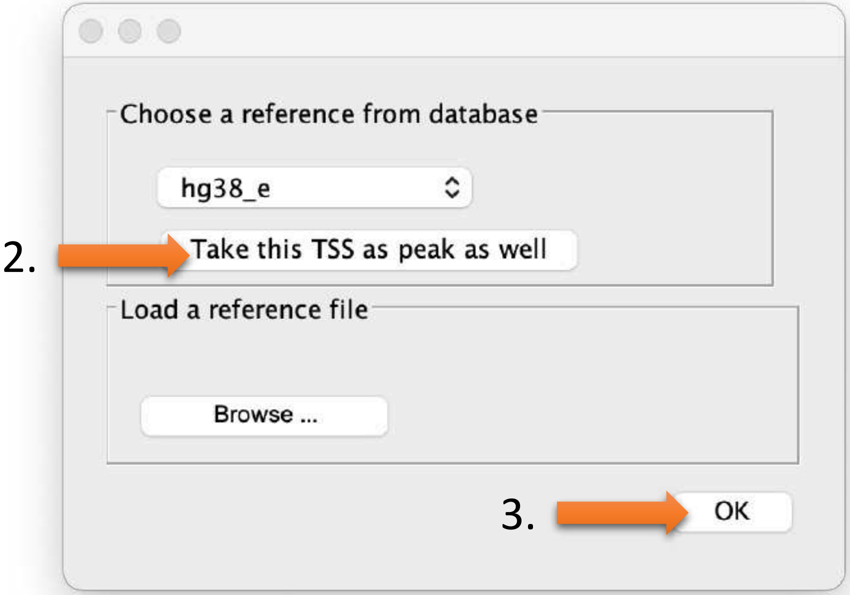
Exercise 9: Clustering

- Now in Choose a reference from database (1), select the last entry of the drop down list (note that name may be truncated).
- Then, click on Take this TSS as peak as well (2) and OK.



1.

Choose a reference from database: NONE/Unknown



2.

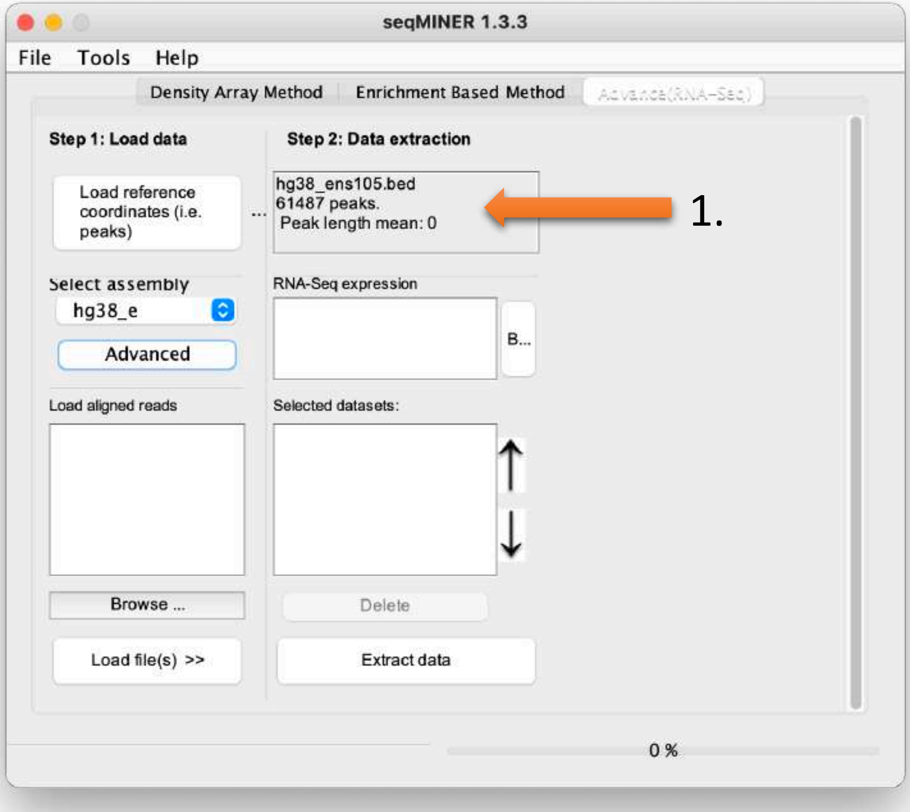
3.

Choose a reference from database: hg38_e

Load a reference file

Exercise 9: Clustering

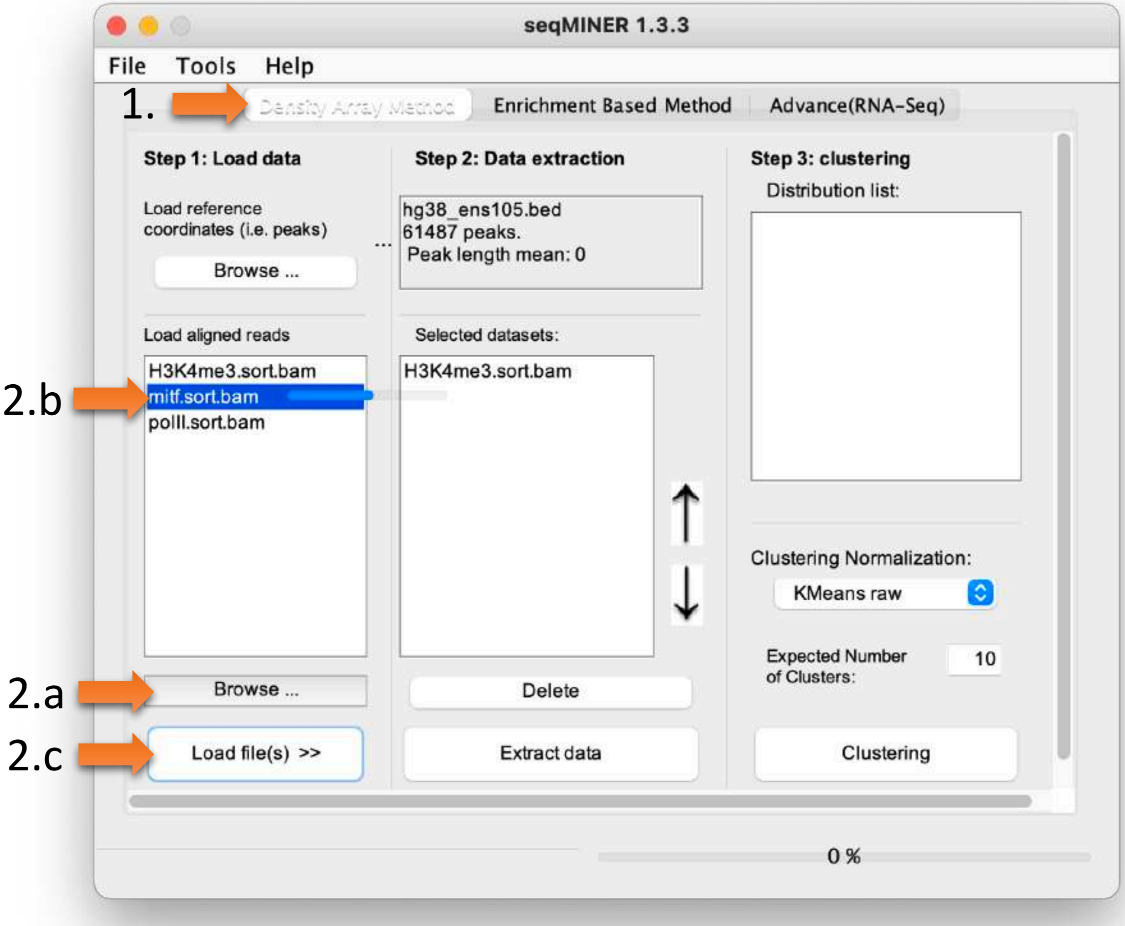
- seqMINER has now extracted the coordinates of TSS from the BED file from genes coordinates (1).



hg38_ens105.bed
61487 peaks.
Peak length mean: 0

Exercise 9: Clustering

- Click on Density Array Method (1).
- Load the datasets (2)
  - Click on Browse… (2.a) and select the files in the browser. Select the bam files of MITF, polII, H3K4me3 (in the directory chipseq/mapping).
  - Select one file in the list (2.b) and click on Load files >> (2.c). **Do it for all files, one at a time.**



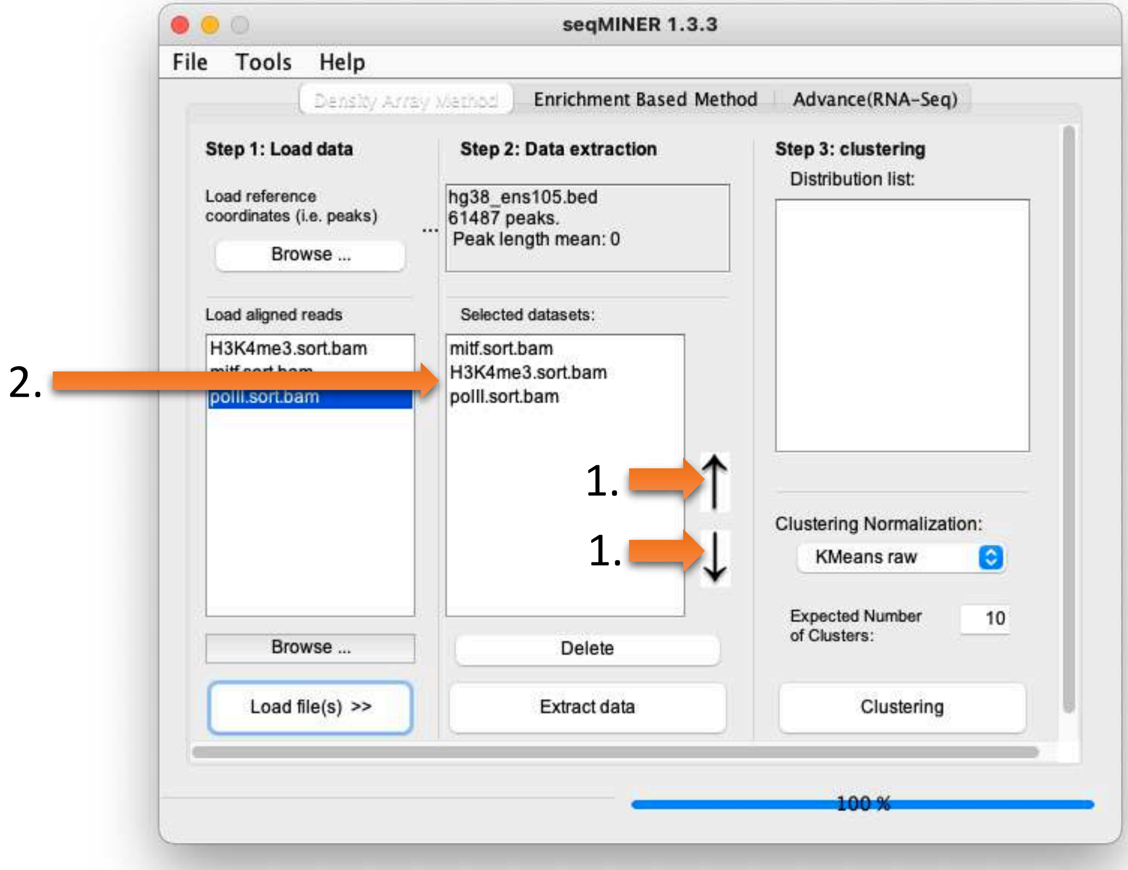
hg38_ens105.bed
61487 peaks.
Peak length mean: 0

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

Selected datasets:
H3K4me3.sort.bam

Exercise 9: Clustering

- Note you can change the track order using the arrows (1.). Set this specific order:
  - MITF,
  - H3K4me3,
  - PolII
  - (See 2.)



hg38_ens105.bed
61487 peaks.
Peak length mean: 0

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

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

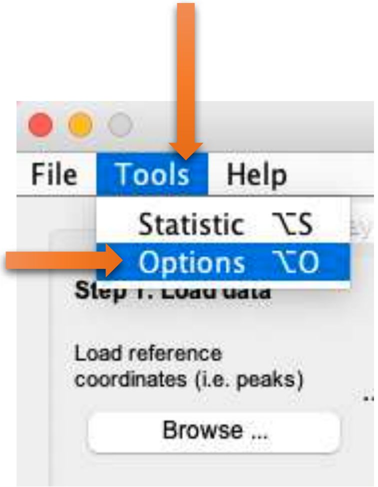
Clustering Normalization: KMeans raw

Expected Number of Clusters: 10

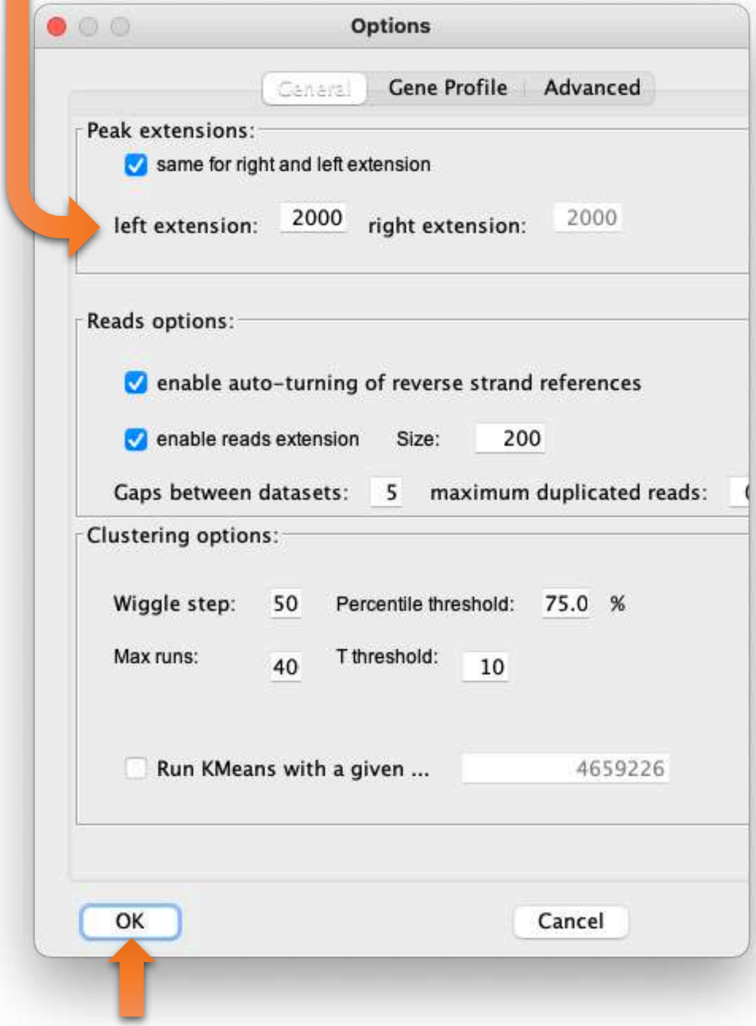
Exercise 9: Clustering

- We are going to restrict the analysis to the +/2Kb region around TSS. Let's edit the parameters accordingly:

1. Click on Tools > Options



2. Enter 2000 instead of 5000 for left extension. Note that the same value is used for right and left extension



Peak extensions:

☑ same for right and left extension

left extension: 2000 right extension: 2000

Reads options:

☑ enable auto-turning of reverse strand references

☑ enable reads extension Size: 200

Gaps between datasets: 5 maximum duplicated reads: (

Clustering options:

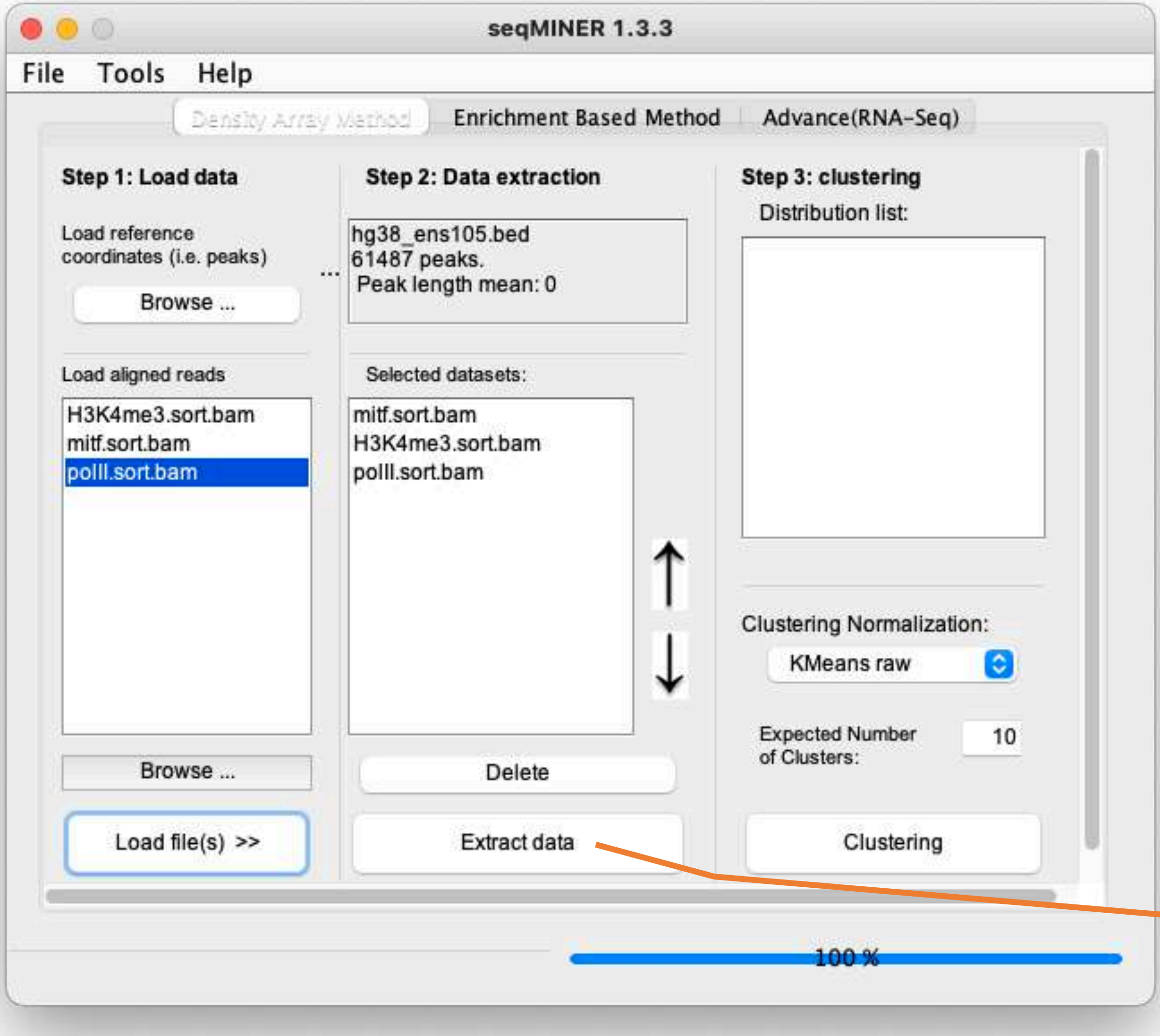
Wiggle step: 50 Percentile threshold: 75.0 %

Max runs: 40 T threshold: 10

□ Run KMeans with a given ... 4659226

3.

Exercise 9: Clustering



Step 1: Load data

Load reference coordinates (i.e. peaks)

Load aligned reads

H3K4me3.sort.bam
mitf.sort.bam
polll.sort.bam

Step 2: Data extraction

hg38_ens105.bed
61487 peaks.
Peak length mean: 0

Selected datasets:

mitf.sort.bam
H3K4me3.sort.bam
polll.sort.bam

Step 3: clustering

Distribution list:

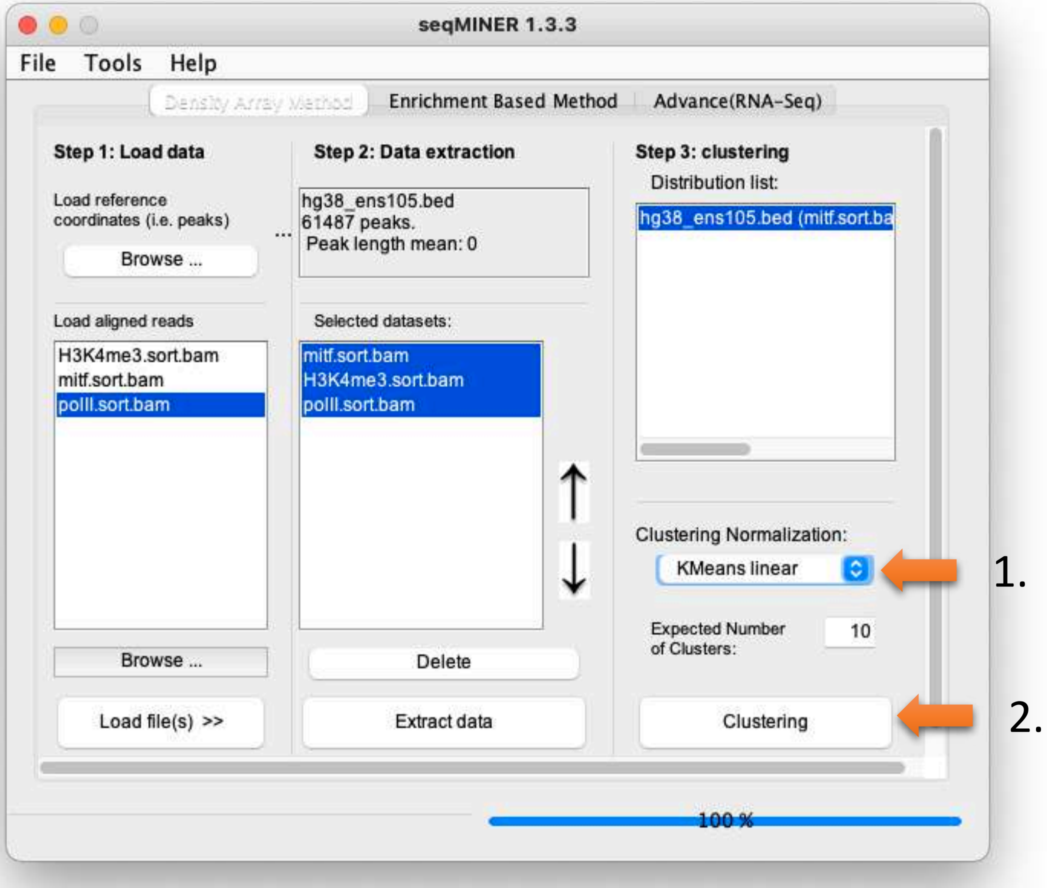
Clustering Normalization: KMeans raw

Expected Number of Clusters: 10

Click on Extract data.

Exercise 9: Clustering

- In Clustering Normalization: select KMeans linear (1.)
- Click on Clustering (2.)



seqMINER 1.3.3

Step 1: Load data

Load reference coordinates (i.e. peaks)

Load aligned reads

H3K4me3.sort.bam
mitf.sort.bam
polII.sort.bam

Step 2: Data extraction

hg38_ens105.bed
61487 peaks.
Peak length mean: 0

Selected datasets:

mitf.sort.bam
H3K4me3.sort.bam
polII.sort.bam

Step 3: clustering

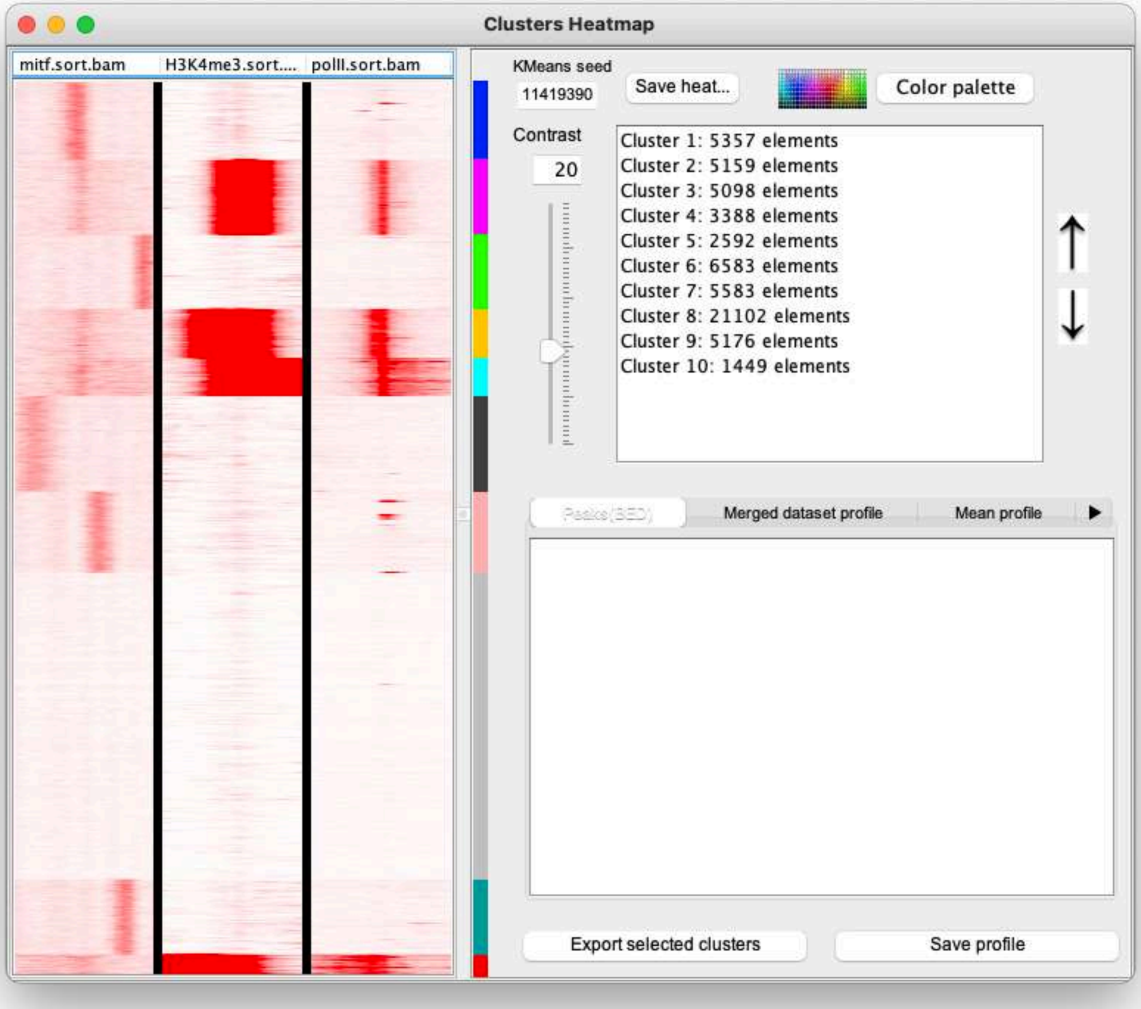
Distribution list:

hg38_ens105.bed (mitf.sort.ba

Clustering Normalization: KMeans linear

Expected Number of Clusters: 10

Exercise 9: Clustering



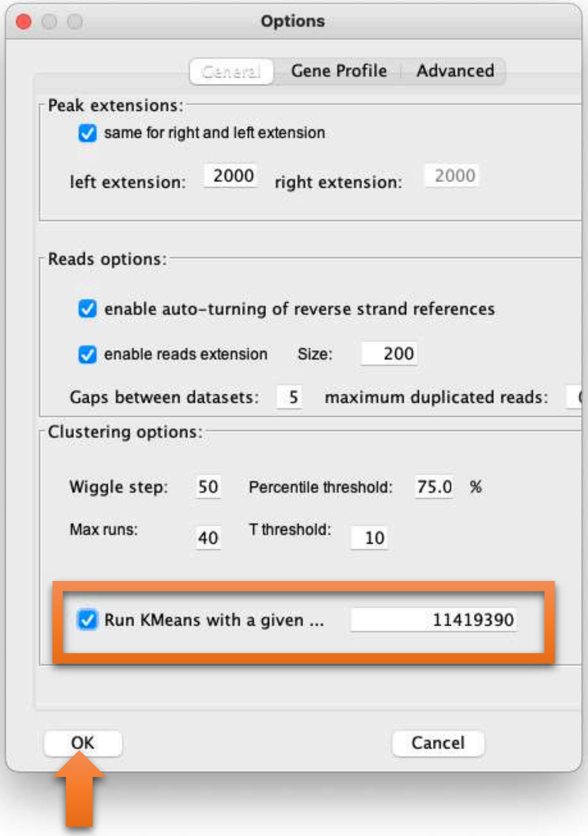
KMeans seed 11419390

Contrast 20

Cluster 1: 5357 elements
Cluster 2: 5159 elements
Cluster 3: 5098 elements
Cluster 4: 3388 elements
Cluster 5: 2592 elements
Cluster 6: 6583 elements
Cluster 7: 5583 elements
Cluster 8: 21102 elements
Cluster 9: 5176 elements
Cluster 10: 1449 elements

Exercise 9: 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. Then, click on Clustering in the main window and you'll get the same results. For instance, the clustering below can be obtained with a Kmeans seed value of 11419390.



Peak extensions:

☑ same for right and left extension

left extension: 2000 right extension: 2000

Reads options:

☑ enable auto-turning of reverse strand references

☑ enable reads extension Size: 200

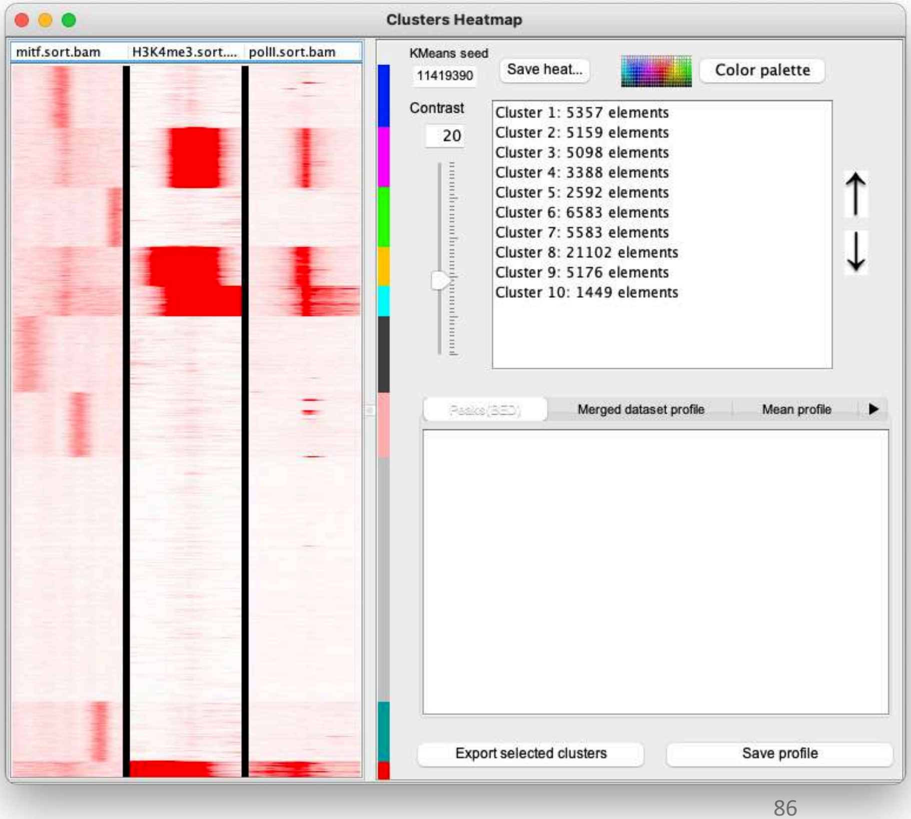
Gaps between datasets: 5 maximum duplicated reads:

Clustering options:

Wiggle step: 50 Percentile threshold: 75.0 %

Max runs: 40 T threshold: 10

☑ Run KMeans with a given ... 11419390



KMeans seed 11419390

Contrast 20

Cluster 1: 5357 elements
Cluster 2: 5159 elements
Cluster 3: 5098 elements
Cluster 4: 3388 elements
Cluster 5: 2592 elements
Cluster 6: 6583 elements
Cluster 7: 5583 elements
Cluster 8: 21102 elements
Cluster 9: 5176 elements
Cluster 10: 1449 elements

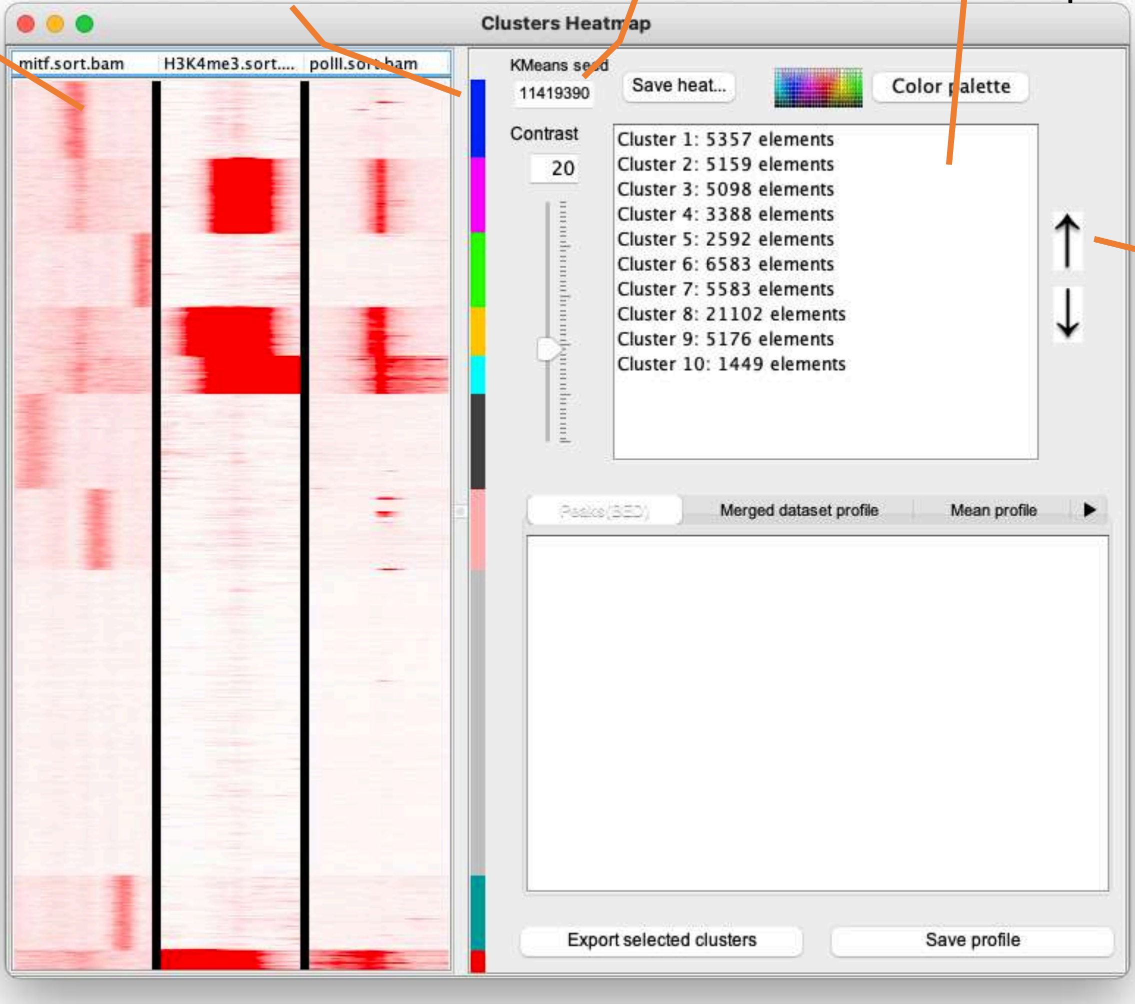
Exercise 9: Clustering

Heatmap

Cluster definition

Kmeans seed value

Clusters, click on one or multiple cluster names to display information in the panel below.



KMeans seed 11419390

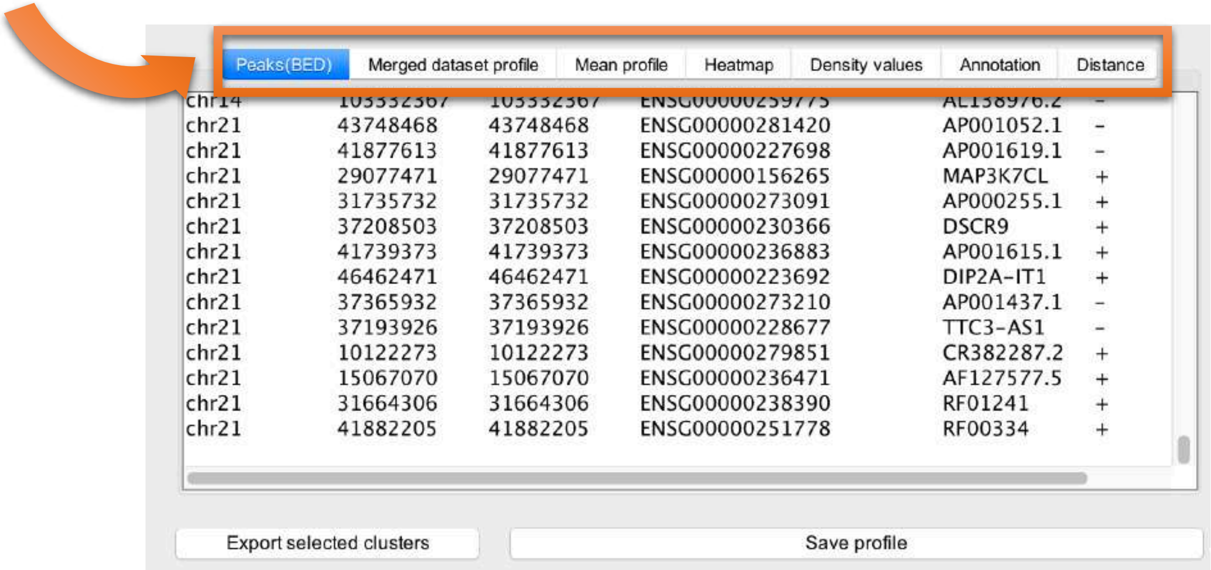
Contrast 20

```
Cluster 1: 5357 elements
Cluster 2: 5159 elements
Cluster 3: 5098 elements
Cluster 4: 3388 elements
Cluster 5: 2592 elements
Cluster 6: 6583 elements
Cluster 7: 5583 elements
Cluster 8: 21102 elements
Cluster 9: 5176 elements
Cluster 10: 1449 elements
```

Change position of selected cluster in the heatmap and in the list

Exercise 9: 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

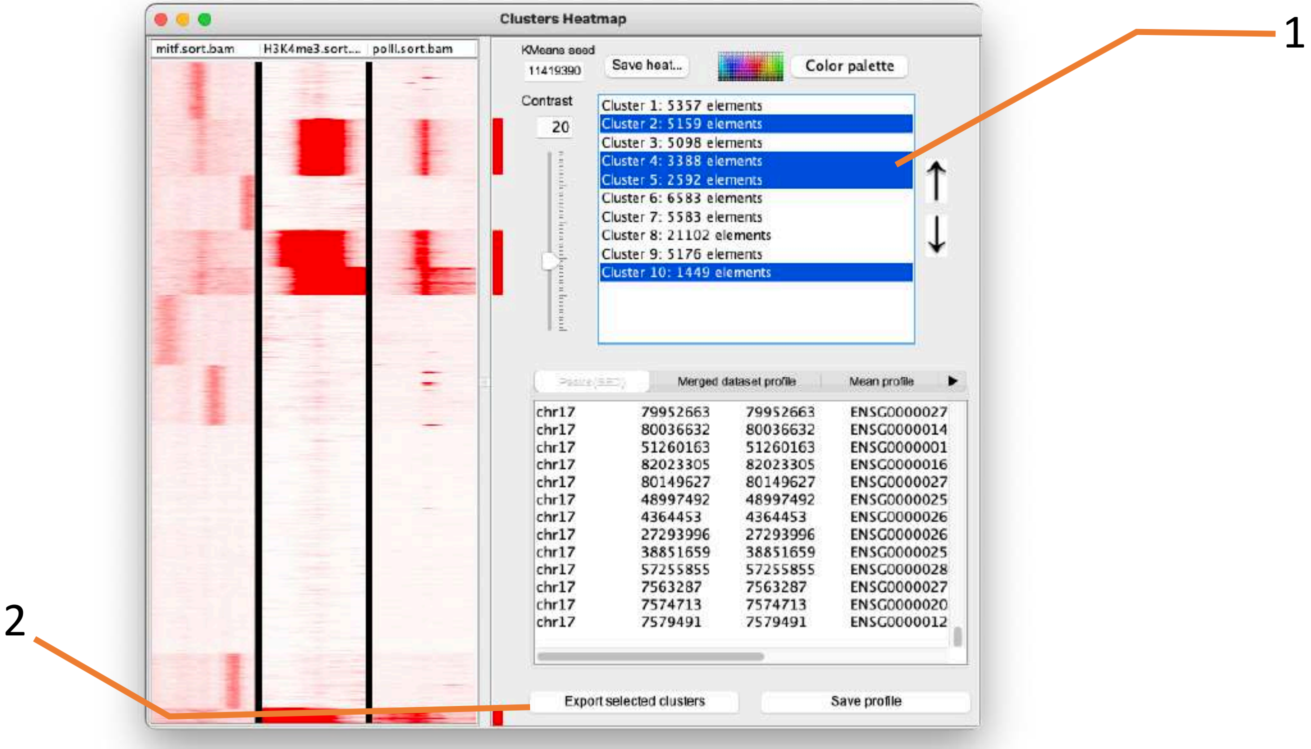


chr14	103332367	103332367	ENSG00000259775	AL138976.2	-
chr21	43748468	43748468	ENSG00000281420	AP001052.1	-
chr21	41877613	41877613	ENSG00000227698	AP001619.1	-
chr21	29077471	29077471	ENSG00000156265	MAP3K7CL	+
chr21	31735732	31735732	ENSG00000273091	AP000255.1	+
chr21	37208503	37208503	ENSG00000230366	DSCR9	+
chr21	41739373	41739373	ENSG00000236883	AP001615.1	+
chr21	46462471	46462471	ENSG00000223692	DIP2A-IT1	+
chr21	37365932	37365932	ENSG00000273210	AP001437.1	-
chr21	37193926	37193926	ENSG00000228677	TTC3-AS1	-
chr21	10122273	10122273	ENSG00000279851	CR382287.2	+
chr21	15067070	15067070	ENSG00000236471	AF127577.5	+
chr21	31664306	31664306	ENSG00000238390	RF01241	+
chr21	41882205	41882205	ENSG00000251778	RF00334	+

Exercise 9: Clustering

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

- Select all clusters that have signal at TSS (1) and export the clusters (2) (reference coordinates) into a file called sub-clustering-tss.bed.



KMeans seed: 11419390

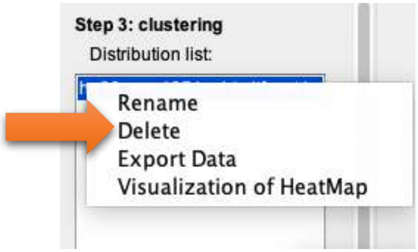
Contrast: 20

Cluster 1: 5357 elements
Cluster 2: 5159 elements
Cluster 3: 5098 elements
Cluster 4: 3388 elements
Cluster 5: 2592 elements
Cluster 6: 6583 elements
Cluster 7: 5583 elements
Cluster 8: 21102 elements
Cluster 9: 5176 elements
Cluster 10: 1449 elements

```
chr17    79952663    79952663    ENSG0000027
chr17    80036632    80036632    ENSG0000014
chr17    51260163    51260163    ENSG0000001
chr17    82023305    82023305    ENSG0000016
chr17    80149627    80149627    ENSG0000027
chr17    48997492    48997492    ENSG0000025
chr17    4364453     4364453     ENSG0000026
chr17    27293996    27293996    ENSG0000026
chr17    38851659    38851659    ENSG0000025
chr17    57255855    57255855    ENSG0000028
chr17    7563287     7563287     ENSG0000027
chr17    7574713     7574713     ENSG0000020
chr17    7579491     7579491     ENSG0000012
```

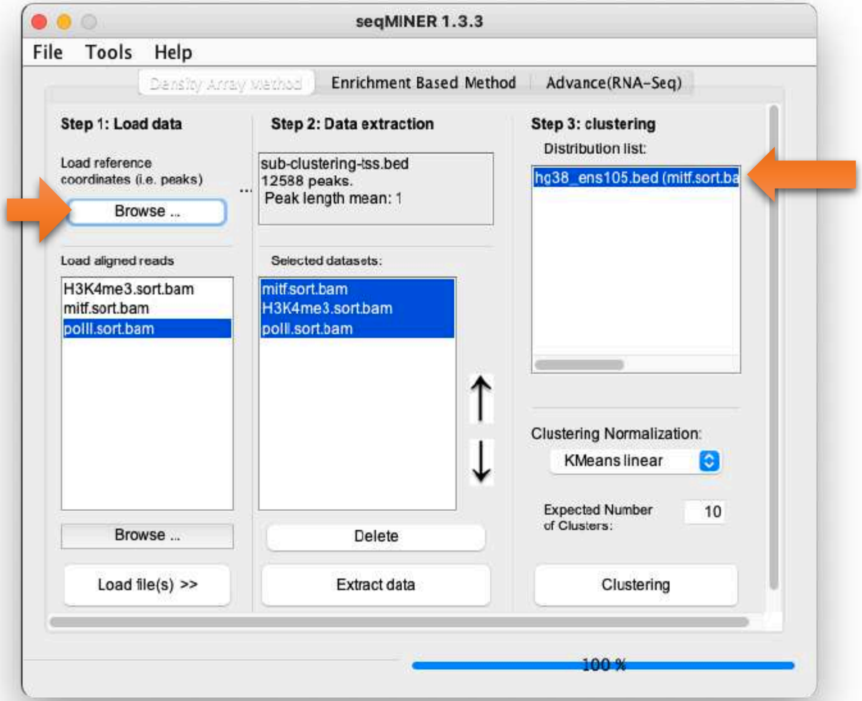
Exercise 9: Clustering

- Load the file sub-clustering-tss.bed as reference coordinates (1). Or use the one I generated (see chipseq/seqminer/sub-clustering-tss.bed)
- Remove previous distribution (to save memory) (2)
  - Select the distribution (2.a)
  - Click right on the name of a distribution
  - Select Delete (2.b)
- Extract data (3)
- Run the clustering analysis (4)



2.b

1



sub-clustering-tss.bed
12588 peaks.
Peak length mean: 1

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

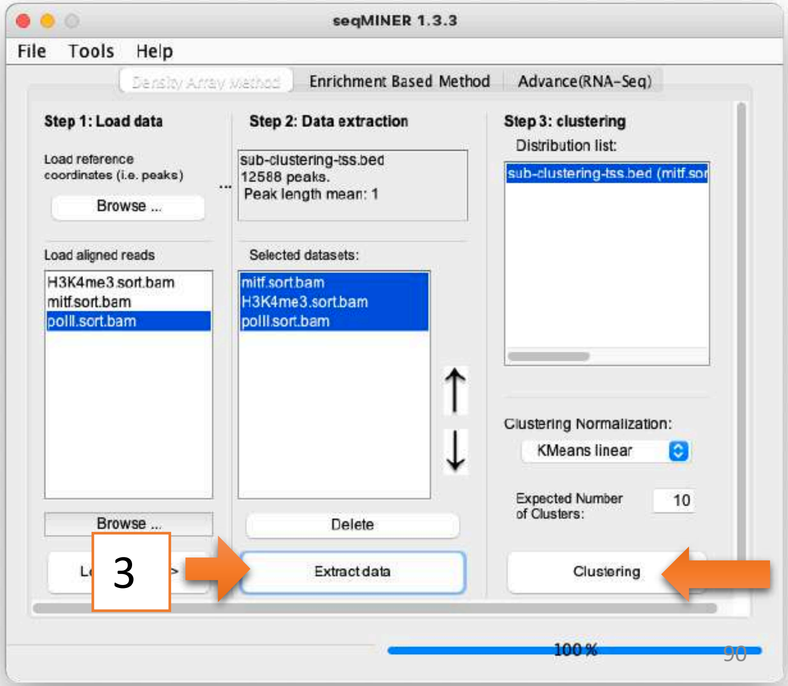
Selected datasets: mitf.sort.bam, H3K4me3.sort.bam, polII.sort.bam

Distribution list: hg38_ens105.bed (mitf.sort.ba

Clustering Normalization: KMeans linear

Expected Number of Clusters: 10

2.a



sub-clustering-tss.bed
12588 peaks.
Peak length mean: 1

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

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

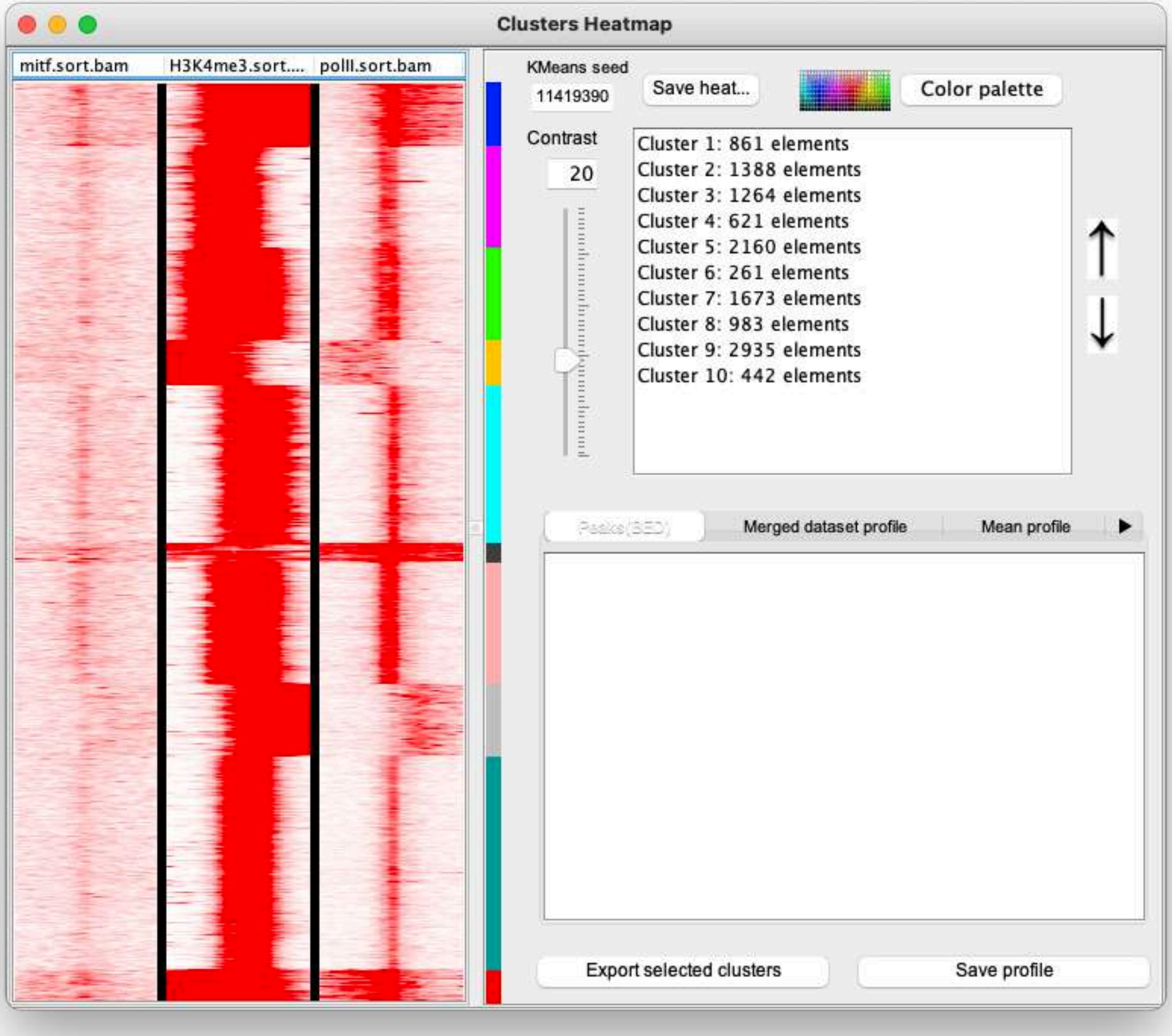
Distribution list: sub-clustering-tss.bed (mitf.sor

Clustering Normalization: KMeans linear

Expected Number of Clusters: 10

4

Exercise 9: Clustering



Clusters Heatmap

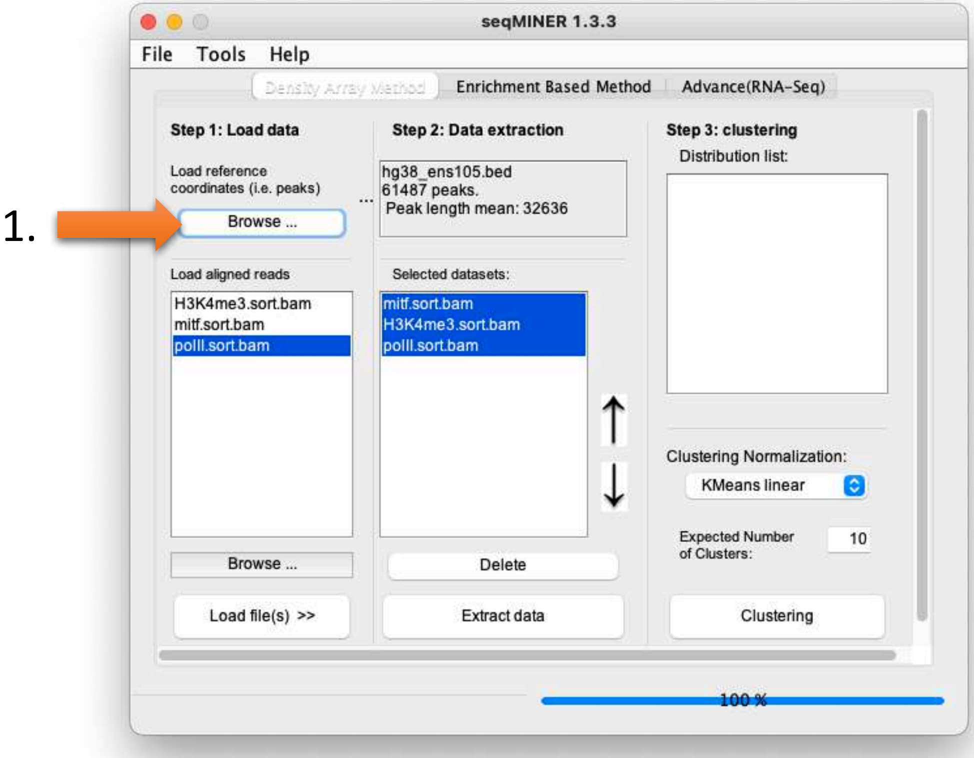
KMeans seed 11419390

Contrast 20

```
Cluster 1: 861 elements
Cluster 2: 1388 elements
Cluster 3: 1264 elements
Cluster 4: 621 elements
Cluster 5: 2160 elements
Cluster 6: 261 elements
Cluster 7: 1673 elements
Cluster 8: 983 elements
Cluster 9: 2935 elements
Cluster 10: 442 elements
```

Exercise 9: Clustering

- Before running any other analysis remove all the distributions from the distribution list (done to save memory)
- Run SeqMINER on all Ensembl (v105) genes from TSS to TTS.
  - Reference coordinates : the file is the one you generated using Ensembl/BioMART (hg38_ens105.bed). Click on Browse to load it. (1)



hg38_ens105.bed
61487 peaks.
Peak length mean: 32636

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

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

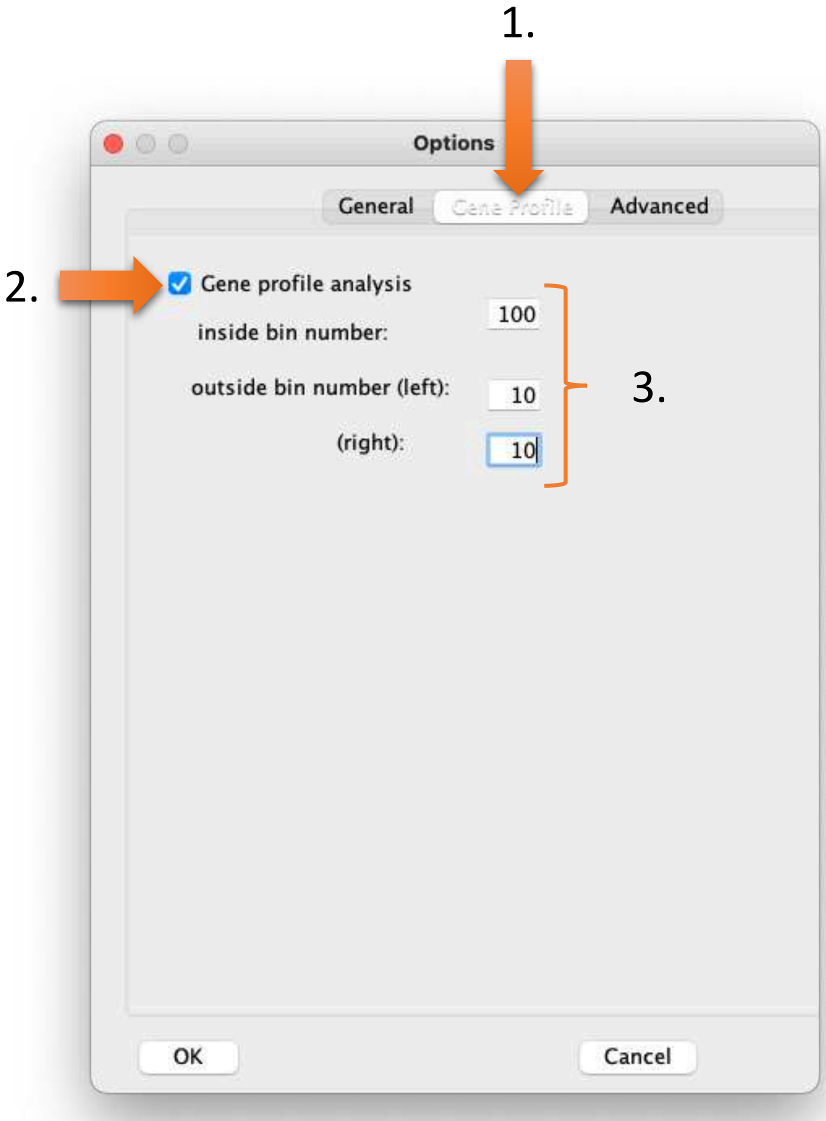
Clustering Normalization: KMeans linear

Expected Number of Clusters: 10

Exercise 9: Clustering

Now we are going to tell seqMINER to work with scaled regions so that they are all considered to be of the same size.

- Go to Tools > Options
- Click on the Gene profile tab (1), select Gene profile analysis. Set parameters (3):
  - Inside bin number: 100
  - Outside bin number (left): 10
  - (right): 10



☑ Gene profile analysis

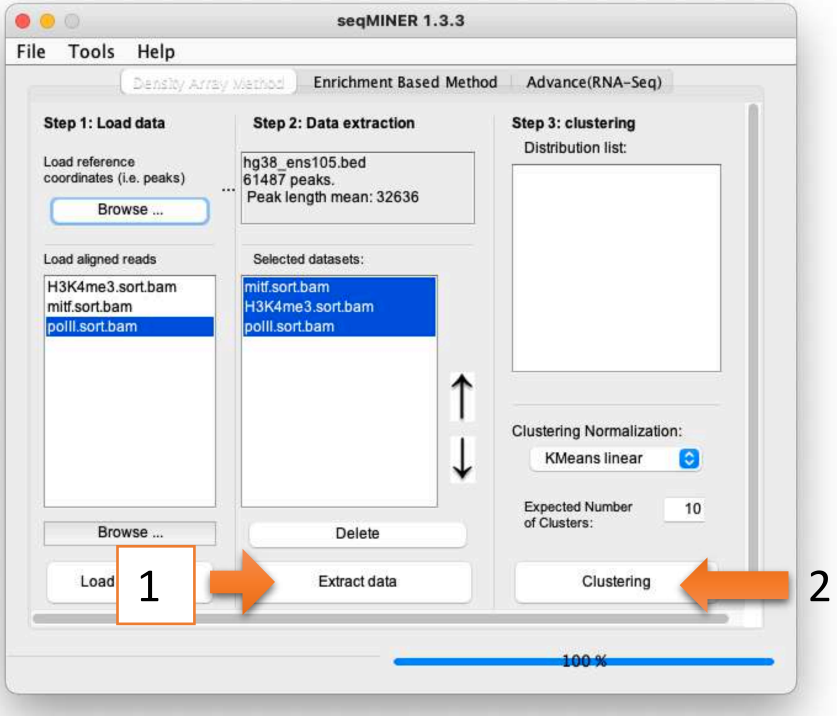
inside bin number: 100

outside bin number (left): 10

(right): 10

Exercise 9: Clustering

- In the tab Options > General, make sure that “Run Kmeans with a given value” is set to 11419390
- Click on OK.
- Click on Extract data (1)
- Click on Clustering (2)



seqMINER 1.3.3

Step 1: Load data

Load reference coordinates (i.e. peaks)

Load aligned reads

H3K4me3.sort.bam
mitf.sort.bam
polII.sort.bam

Step 2: Data extraction

hg38_ens105.bed
61487 peaks.
Peak length mean: 32636

Selected datasets:

mitf.sort.bam
H3K4me3.sort.bam
polII.sort.bam

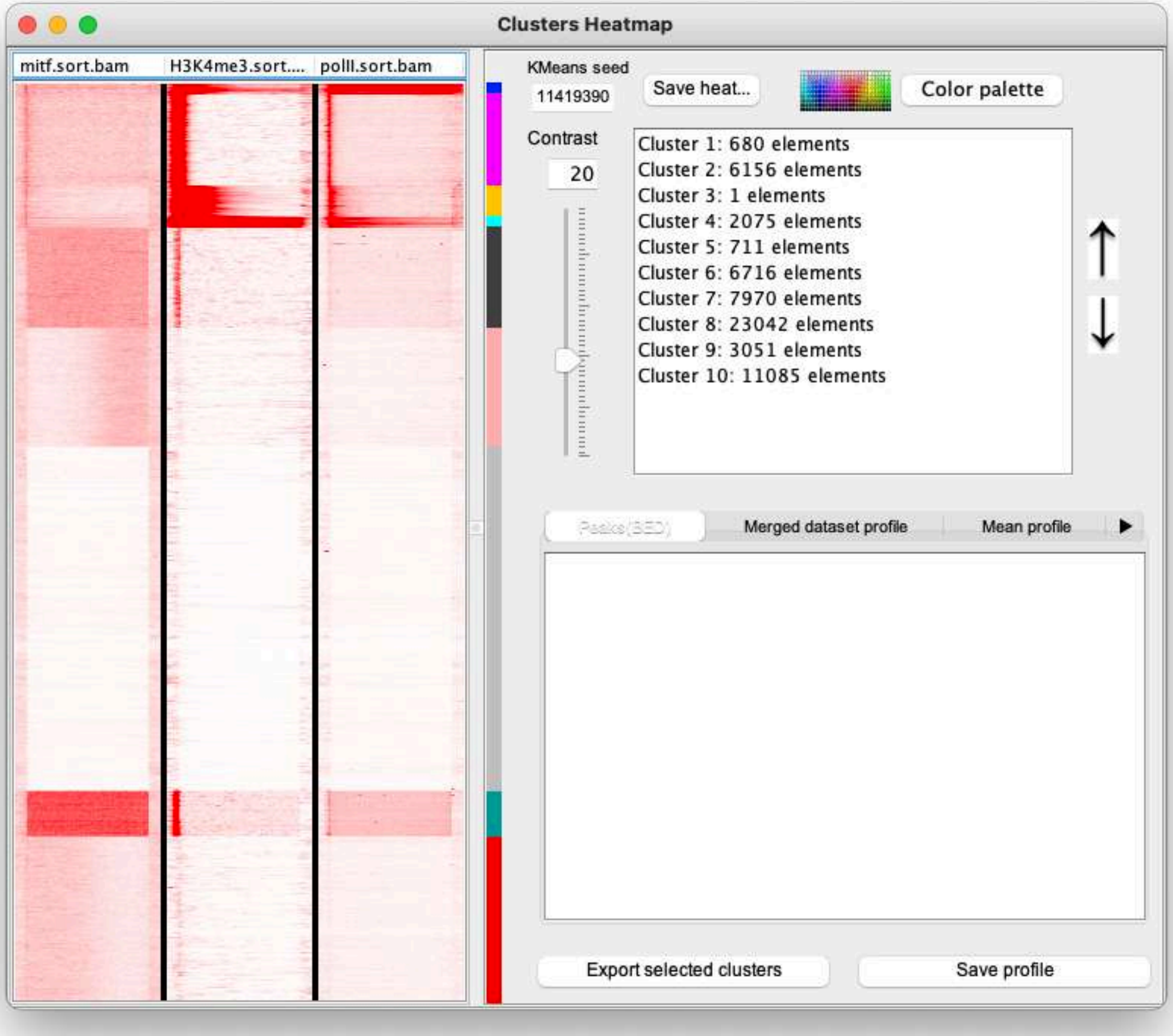
Step 3: clustering

Distribution list:

Clustering Normalization: KMeans linear

Expected Number of Clusters: 10

Exercise 9: Clustering



KMeans seed 11419390

Contrast 20

Cluster 1: 680 elements
Cluster 2: 6156 elements
Cluster 3: 1 elements
Cluster 4: 2075 elements
Cluster 5: 711 elements
Cluster 6: 6716 elements
Cluster 7: 7970 elements
Cluster 8: 23042 elements
Cluster 9: 3051 elements
Cluster 10: 11085 elements

Exercise 9: Clustering

- 1. Export a file with all clusters having MITF, polII and H3K4me3 enrichments (clusters 1, 2, 3, 4 ,5, 9). Save the file as sub-clustering-gene.bed.
  - Do a sub-clustering with the file sub-clustering-gene.bed as reference coordinates (keep same Kmeans seed)
- 2. Additional question:
  - 2.a. Export annotations of cluster 4 generated after last clustering (in question 1.). Save the file as cluster4.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.

ATAC-seq

*<b></b>**A**ssay for **T**ransposase-**A**ccessible **C**hromatin with highthroughput **seq**uencing*

Chromatin accessibility assays

- Chromatin accessibility is the degree to which nuclear macromolecules are able to physically contact chromatinized DNA and is determined by the occupancy and topological organization of nucleosomes as well as other chromatin-binding factors that occlude access to DNA (Klemm et al, 2019)
- Open chromatin regions contains generally transcriptionally active genes
- The accessible genome comprises ~2–3% of total DNA sequence yet captures more than 90% of regions bound by TFs (Thurman et al, 2012)
- Chromatin accessibility is measured by quantifying the susceptibility of chromatin to either enzymatic methylation or cleavage of its constituent DNA
- Chromatin accessibility assays (non exhaustive list)
  - FAIRE-seq
  - DNAse-seq
  - MNAse-seq
  - ATAC-seq

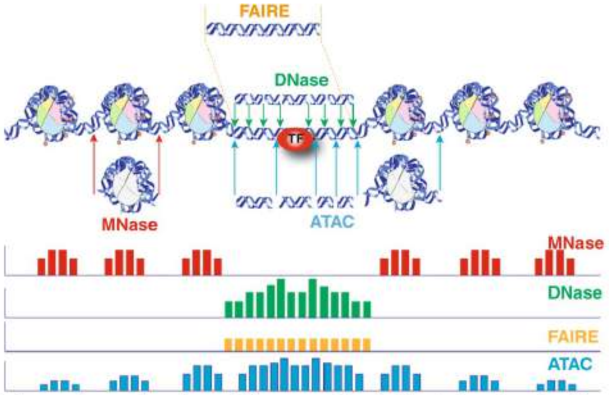
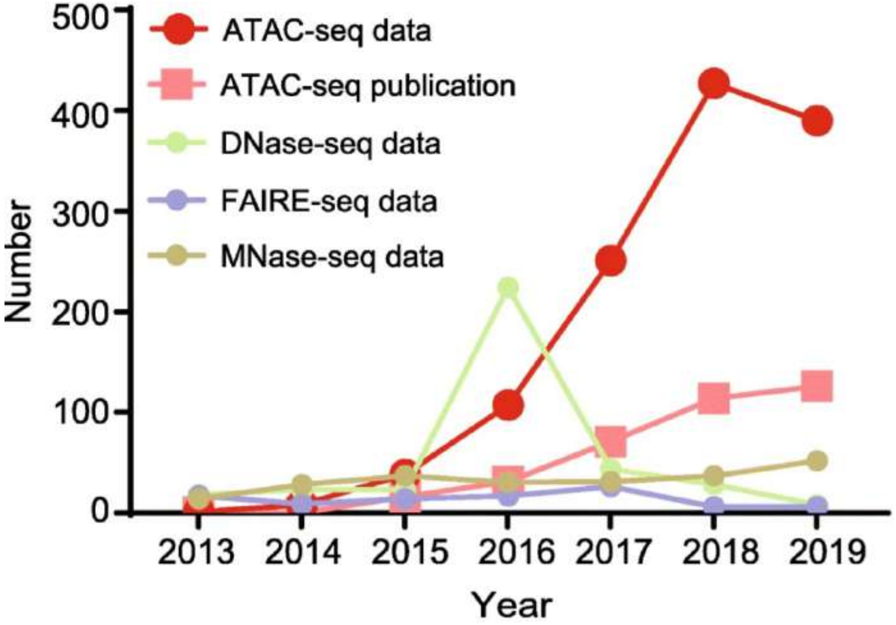


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

(Tsompana and Buck, 2014)

Chromatin accessibility assays

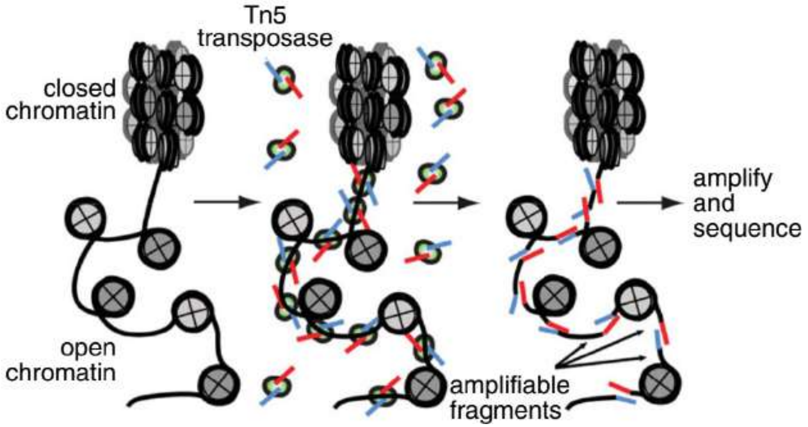
- ATAC-seq has become the most widely used method to detect and analyze open chromatin regions



Yan et al, 2020

ATAC-seq

- Buenrostro et al, 2013
- ATAC-seq is a method for determining chromatin accessibility across the genome
- Transcription factor binding sites and positions of nucleosomes can be identified from the analysis of ATAC-Seq
- Advantages of ATAC-seq over other chromatin accessibility assays:
  - The sensitivity and specificity are comparable to DNase-seq but superior to FAIRE-seq
  - Straightforward and rapidly implemented protocole. ATAC-seq libraries can be generated in less than 3 hours
  - Low number of cells required (500 - 50,000 cells cells)
  - single-cell ATAC-seq (scATAC-seq) (Cusanovich et al, 2015) protocole



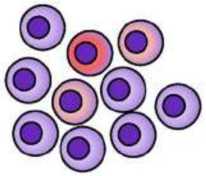
(Buenrostro et al., 2015).

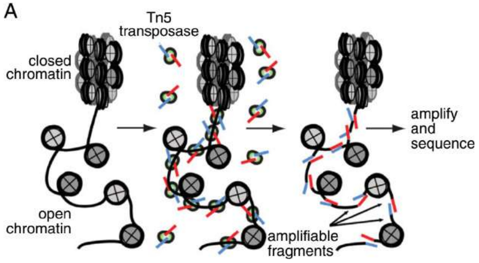
ATAC-seq process

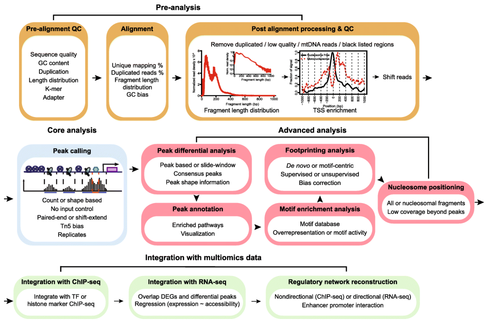
Sample processing

ATAC-seq

Data analysis







Pre-analysis

Pre-alignment QC: Sequence quality, GC content, Duplication, Length distribution, K-mer, Adapter

Alignment: Unique mapping %, Duplicated reads %, Fragment length distribution, GC bias

Post alignment processing & QC: Remove duplicated / low quality / mtDNA reads / black listed regions; Fragment length distribution; TSS enrichment; Shift reads

Core analysis

Peak calling: Count or shape based, No input control, Paired-end or shift-extend, Tn5 bias, Replicates

Advanced analysis

Peak differential analysis: Peak based or slide-window, Consensus peaks, Peak shape information

Peak annotation: Enriched pathways, Visualization

Footprinting analysis: *De novo* or motif-centric, Supervised or unsupervised, Bias correction

Motif enrichment analysis: Motif database, Overrepresentation or motif activity

Nucleosome positioning: All or nucleosomal fragments, Low coverage beyond peaks

Integration with multiomics data

Integration with ChIP-seq: Integrate with TF or histone marker ChIP-seq

Integration with RNA-seq: Overlap DEGs and differential peaks, Regression (expression ~ accessibility)

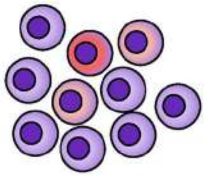
Regulatory network reconstruction: Nondirectional (ChIP-seq) or directional (RNA-seq), Enhancer promoter interaction

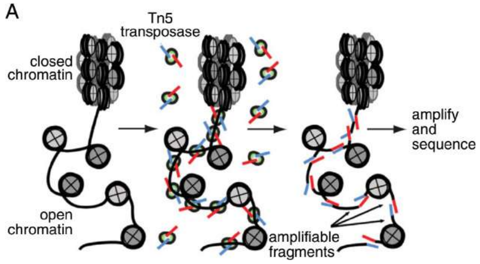
ATAC-seq

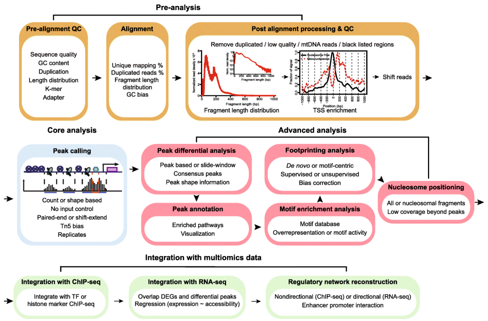
Sample processing

ATAC-seq

Data analysis







Pre-analysis

Pre-alignment QC
Sequence quality
GC content
Duplication
Length distribution
K-mer
Adapter

Alignment
Unique mapping %
Duplicated reads %
Fragment length distribution
GC bias

Post alignment processing & QC
Remove duplicated / low quality / mtDNA reads / black listed regions
Fragment length distribution
TSS enrichment
Shift reads

Core analysis

Peak calling
Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates

Advanced analysis

Peak differential analysis
Peak based or slide-window
Consensus peaks
Peak shape information

Peak annotation
Enriched pathways
Visualization

Footprinting analysis
De novo or motif-centric
Supervised or unsupervised
Bias correction

Motif enrichment analysis
Motif database
Overrepresentation or motif activity

Nucleosome positioning
All or nucleosomal fragments
Low coverage beyond peaks

Integration with multiomics data

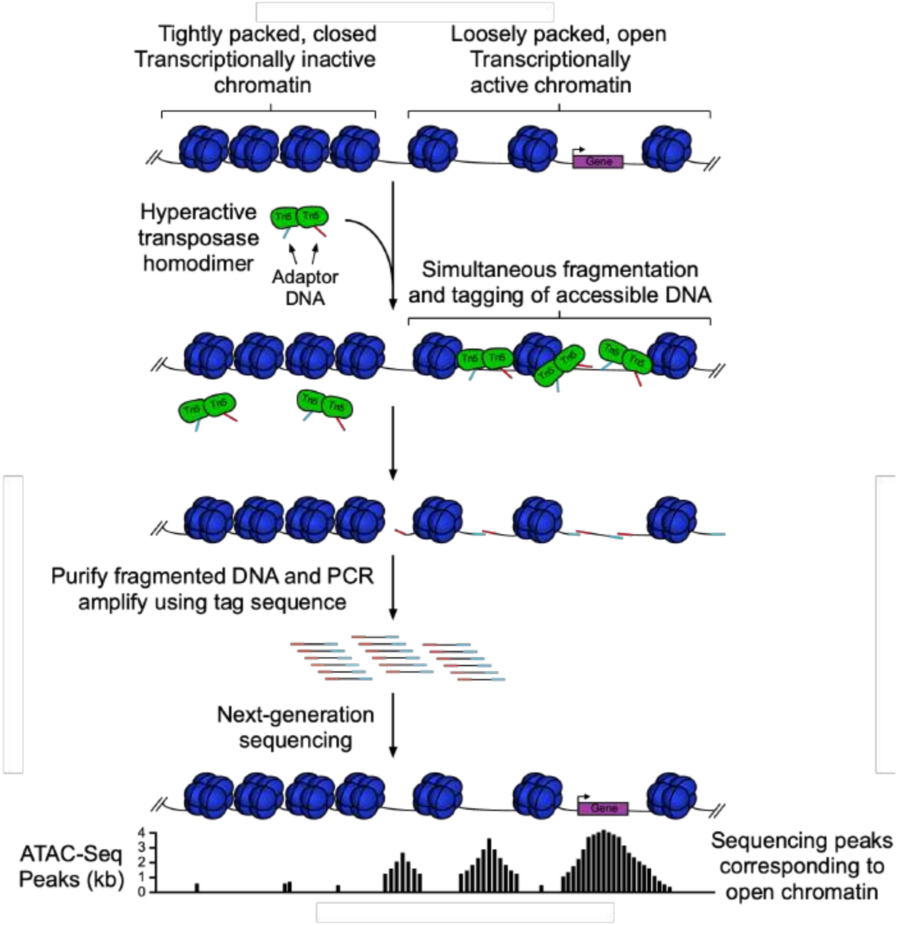
Integration with ChIP-seq
Integrate with TF or histone marker ChIP-seq

Integration with RNA-seq
Overlap DEGs and differential peaks
Regression (expression ~ accessibility)

Regulatory network reconstruction
Nondirectional (ChIP-seq) or directional (RNA-seq)
Enhancer promoter interaction

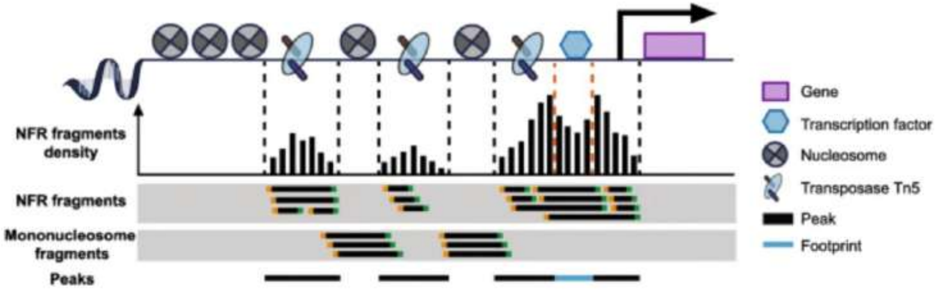
ATAC-seq

- ATAC-seq protocole utilizes a hyperactive Tn5 transposase to insert sequencing adapters into open chromatin regions
- In a process called "tagmentation", Tn5 transposase cleaves and tags double-stranded DNA with sequencing adaptors
- No additional library prep is needed
- Expected results are enrichments of sequenced reads in open chromatin regions as closed chromatin regions, DNA regions bound by TFs or wrapped around nucleosomes should be protected from Tn5 cleavage activity.

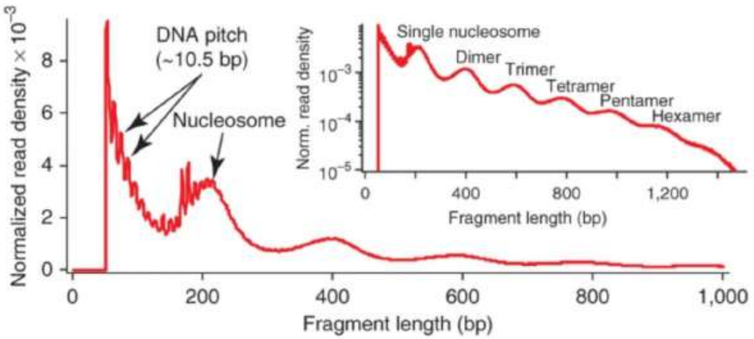


ATAC-seq

- **Paired-end sequencing** so that by looking at the distance between the two reads of a pair, we know in which the chromatin environment (Nucleosome Free Region (NFR), around a mono, di,-nucleosome, around a TF) of the DNA fragment.



Yan et tal, 2020

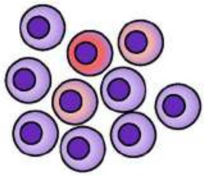


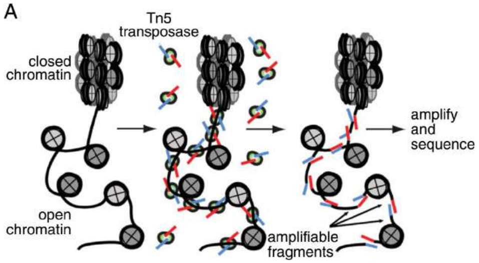
42

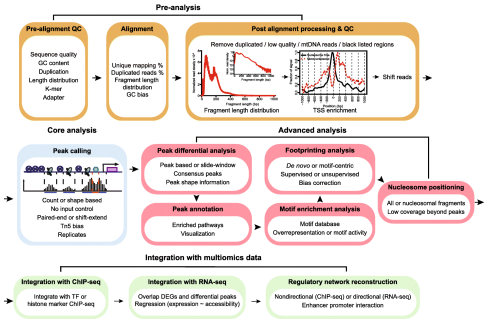
Buenrostro et al, 2023

Analysis of ATAC-seq data

Sample processing → **ATAC-seq** → **Data analysis**







Pre-analysis

Pre-alignment QC: Sequence quality, GC content, Duplication, Length distribution, K-mer, Adapter → Alignment: Unique mapping %, Duplicated reads %, Fragment length distribution, GC bias → Post alignment processing & QC: Remove duplicated / low quality / mtDNA reads / black listed regions; Fragment length distribution; TSS enrichment → Shift reads →

Core analysis

Peak calling: Count or shape based, No input control, Paired-end or shift-extend, Tn5 bias, Replicates

Advanced analysis

Peak differential analysis: Peak based or slide-window, Consensus peaks, Peak shape information

Peak annotation: Enriched pathways, Visualization

Footprinting analysis: *De novo* or motif-centric, Supervised or unsupervised, Bias correction

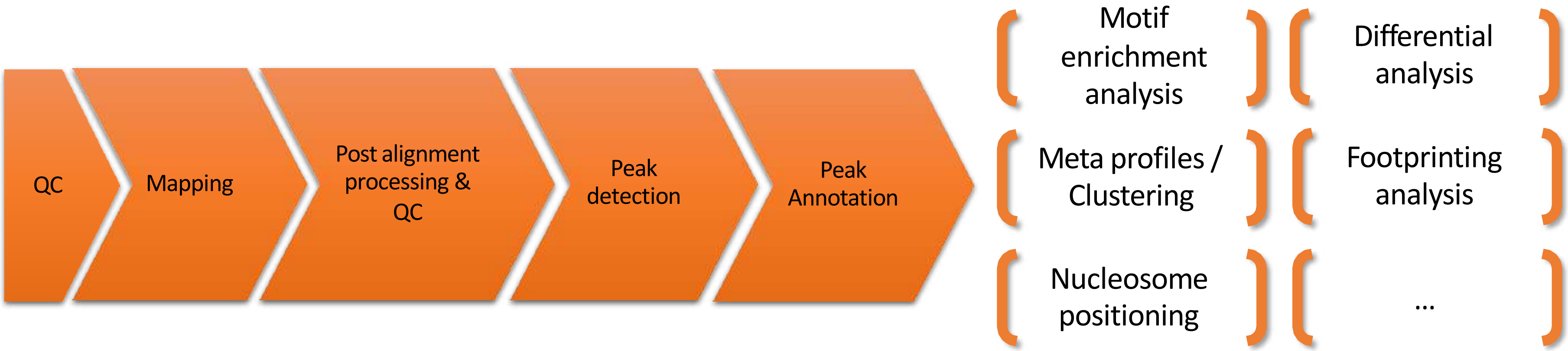
Motif enrichment analysis: Motif database, Overrepresentation or motif activity

Nucleosome positioning: All or nucleosomal fragments, Low coverage beyond peaks

Integration with multiomics data

Integration with ChIP-seq: Integrate with TF or histone marker ChIP-seq → Integration with RNA-seq: Overlap DEGs and differential peaks, Regression (expression ~ accessibility) → Regulatory network reconstruction: Nondirectional (ChIP-seq) or directional (RNA-seq), Enhancer promoter interaction

Analysis of ATAC-seq data



QC

Mapping

Post alignment processing & QC

Peak detection

Peak Annotation

Advanced analysis

Motif enrichment analysis

Differential analysis

Meta profiles / Clustering

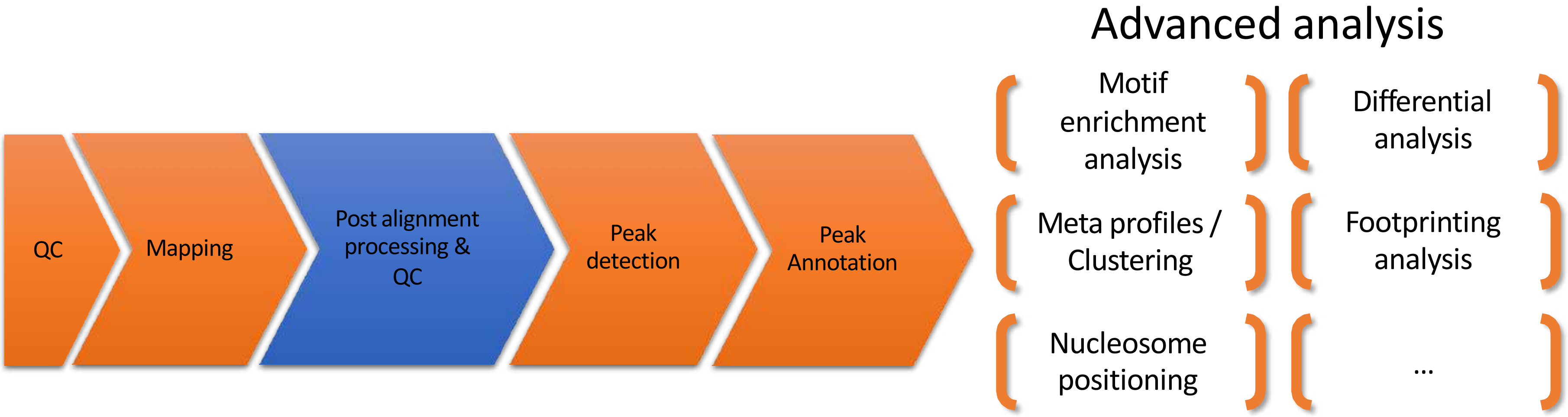
Footprinting analysis

Nucleosome positioning

…

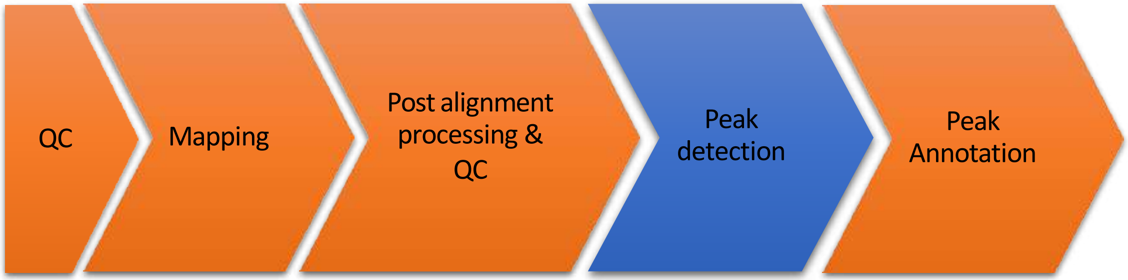
- Overall analysis resemble ChIP-seq data analysis
- Description of particularities of ATAC-seq data analysis

Analysis of ATAC-seq data



- Some cleaning steps are required for ATAC-seq. For example:
  - A large percentage of reads are derived from mitochondrial DNA. These reads are removed as mitochondrial genome is generally not of interest.
    - Omni-ATAC (Corces et al, 2017)

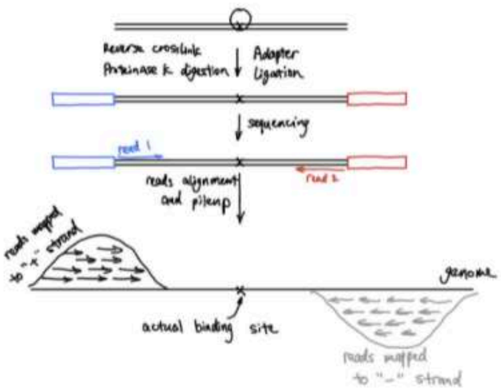
Analysis of ATAC-seq data



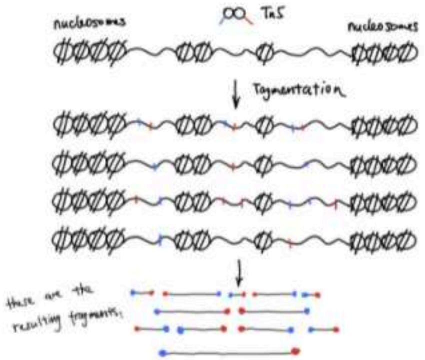
Advanced analysis

- Motif enrichment analysis
- Differential analysis
- Meta profiles / Clustering
- Footprinting analysis
- Nucleosome positioning
- …

ChIP-seq



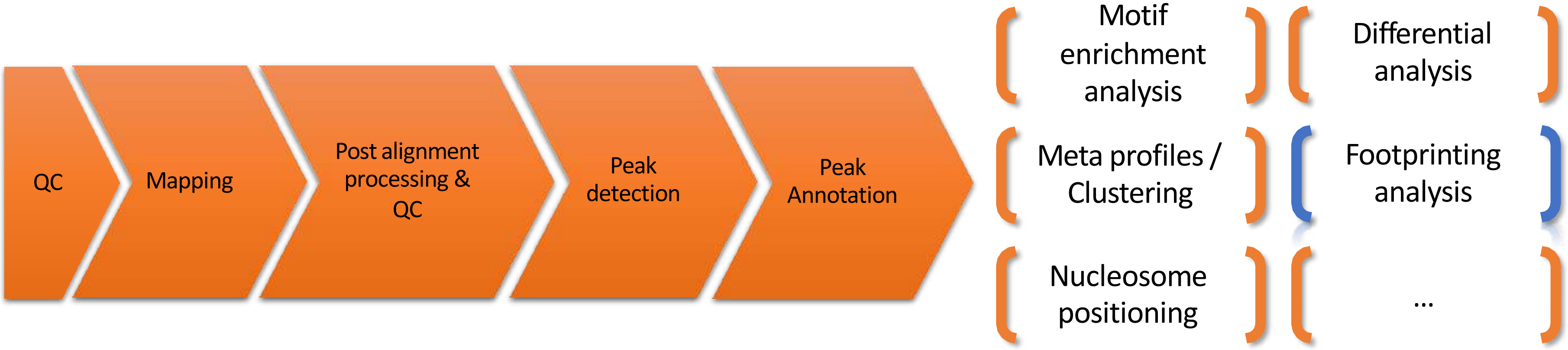
ATAC-seq



Adapted parameters for peak calling (MACS2) : --shift 75 --extsize 150 --nomodel -B --SPMR --keep-dup all --call-summits

Analysis of ATAC-seq data

Advanced analysis



QC

Mapping

Post alignment processing & QC

Peak detection

Peak Annotation

Motif enrichment analysis

Differential analysis

Meta profiles / Clustering

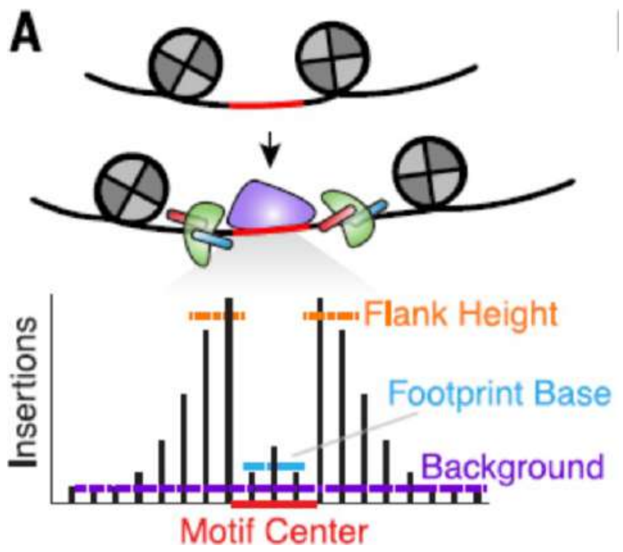
Footprinting analysis

Nucleosome positioning

…

Footprinting analysis

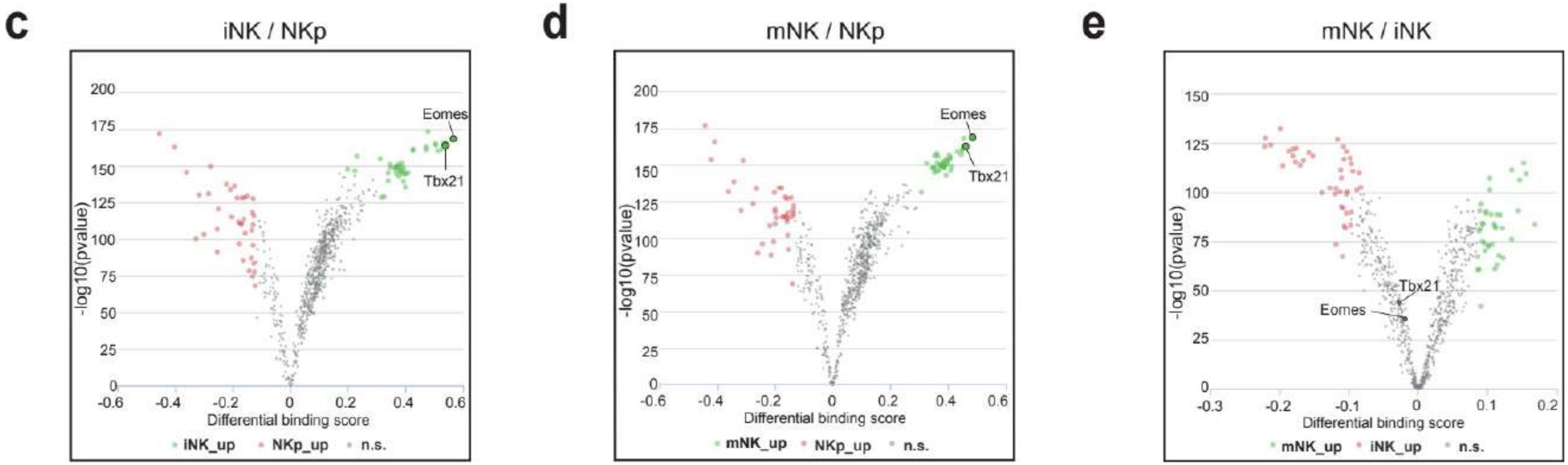
- Tn5 cuts in open chromatin regions
- DNA is protected from cleavage at position of TF binding creating a “notch” in ATAC-seq signal
- Footprinting analysis identifies TF activities
  - Height of the notch reflects TF activity
  - Compare TF activity between different conditions



Corces et al, 2019

Footprinting analysis

- Volcano plots showing differential TF binding activity as predicted by TOBIAS footprinting analysis in ATAC-seq data of NKp, iNK and mNK from Shin et al. (c) iNK vs NKp; (d) mNK vs NKp; (e) mNK vs iNK.
- Each dot represents a TF
- TFs which activity is changing between the two compared developmental stages are colored (see color legend below volcano plots)


c
iNK / NKp
Eomes
Tbx21
-log10(pvalue)
Differential binding score
● iNK_up ● NKp_up ● n.s.
d
mNK / NKp
Eomes
Tbx21
-log10(pvalue)
Differential binding score
● mNK_up ● NKp_up ● n.s.
e
mNK / iNK
Tbx21
Eomes
-log10(pvalue)
Differential binding score
● mNK_up ● iNK_up ● n.s.