Statistical and technical details of ChIP-seq analysis

Master in Omics Data Analysis

Beatrice Borsari
University of Vic, Vic
*Computational Biology of RNA Processing, CRG, Barcelona*

Table of contents

- Single-end vs. paired-end sequencing experiments

- The statistics behind peak calling in MACS2

- Metrics to evaluate a ChIP-seq experiment

- How to analyze chromatin states

- Hands-on session and references

How to analyze the enrichment of TFs and histone marks in the genome

- Let's say we want to identify all the genomic locations bound by a specific TF (the green protein)
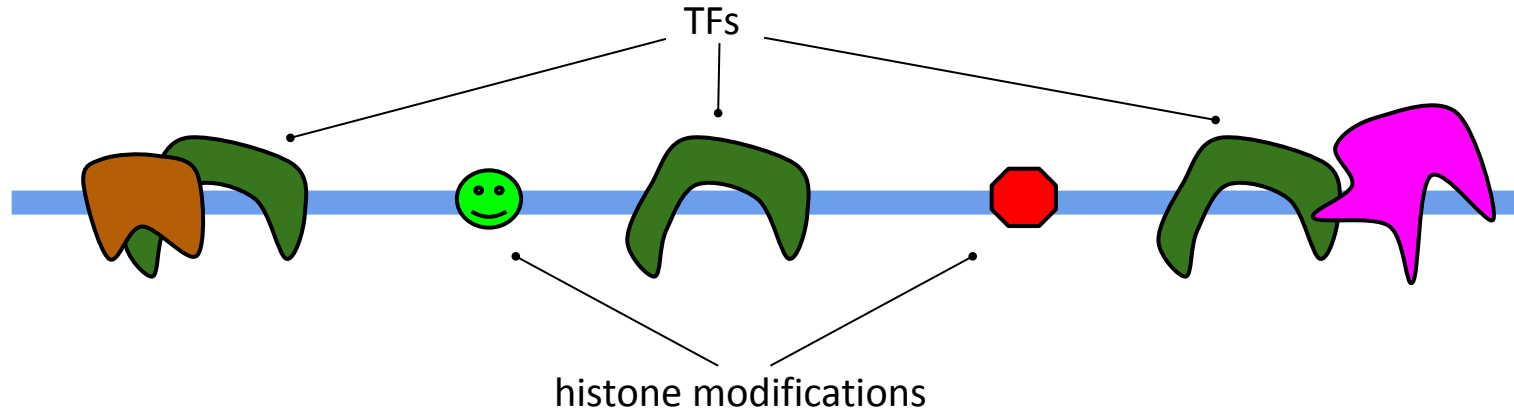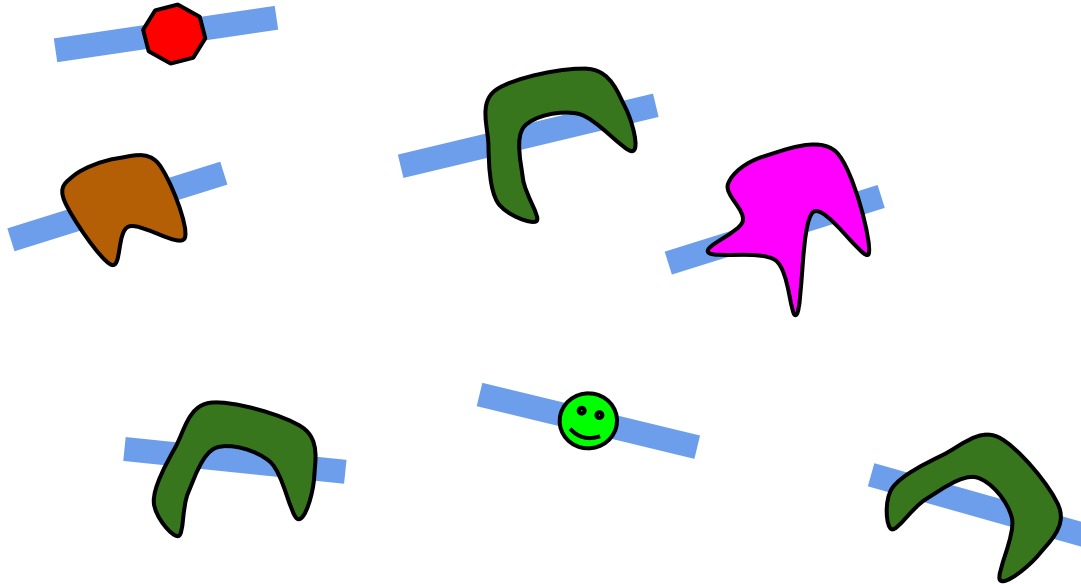
How to analyze the enrichment of TFs and histone marks in the genome

- Let's say we want to identify all the genomic locations bound by a specific TF (the green protein)


TFs

histone modifications
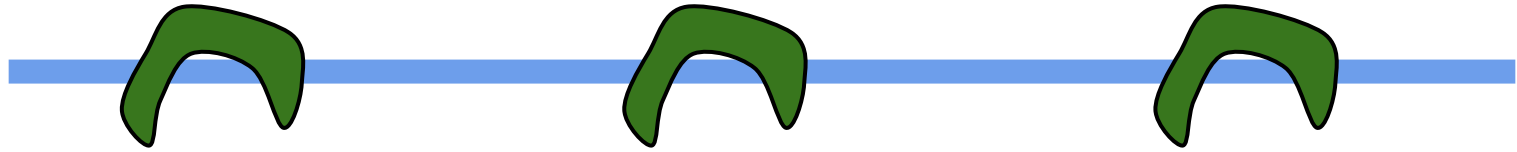
1. We use formaldehyde to glue <u>all</u> the proteins bound to DNA (including the ones we're not interested in) together with DNA

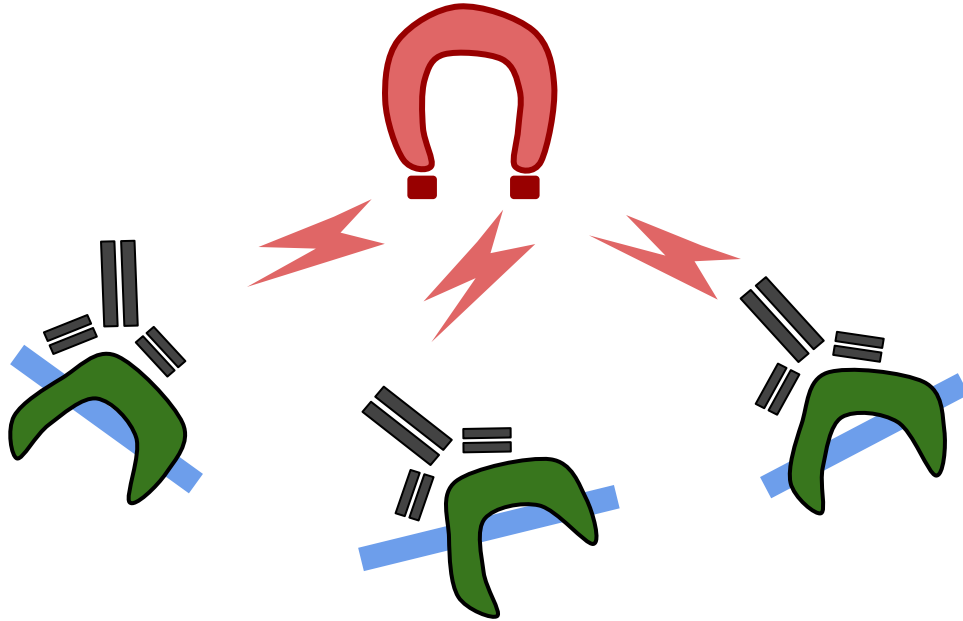2. We cut the DNA up into small (approximately 300 bp) fragments

How to analyze the enrichment of TFs and histone marks in the genome

Antibody attached to a bead

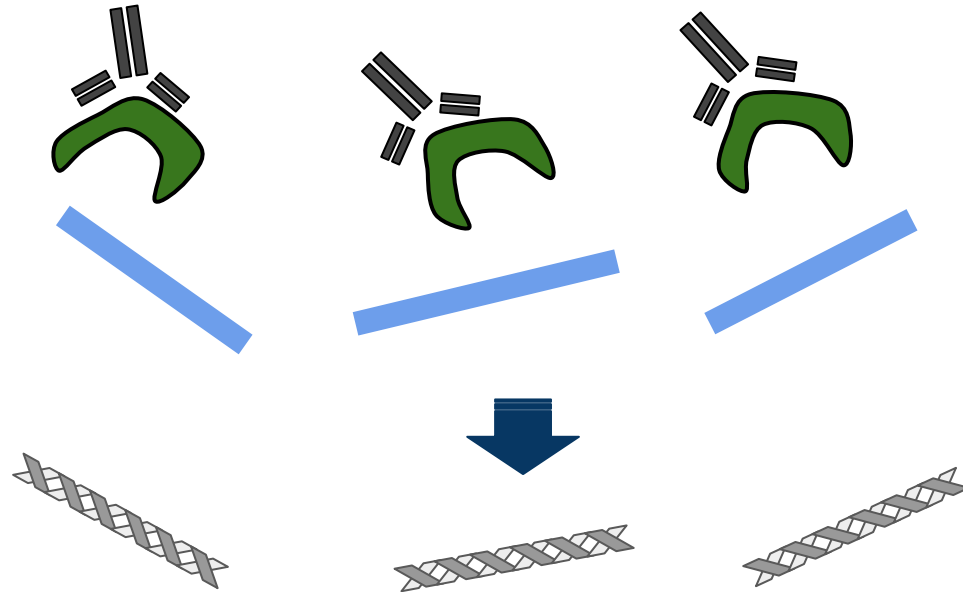3. We isolate the protein we're interested in using an antibody
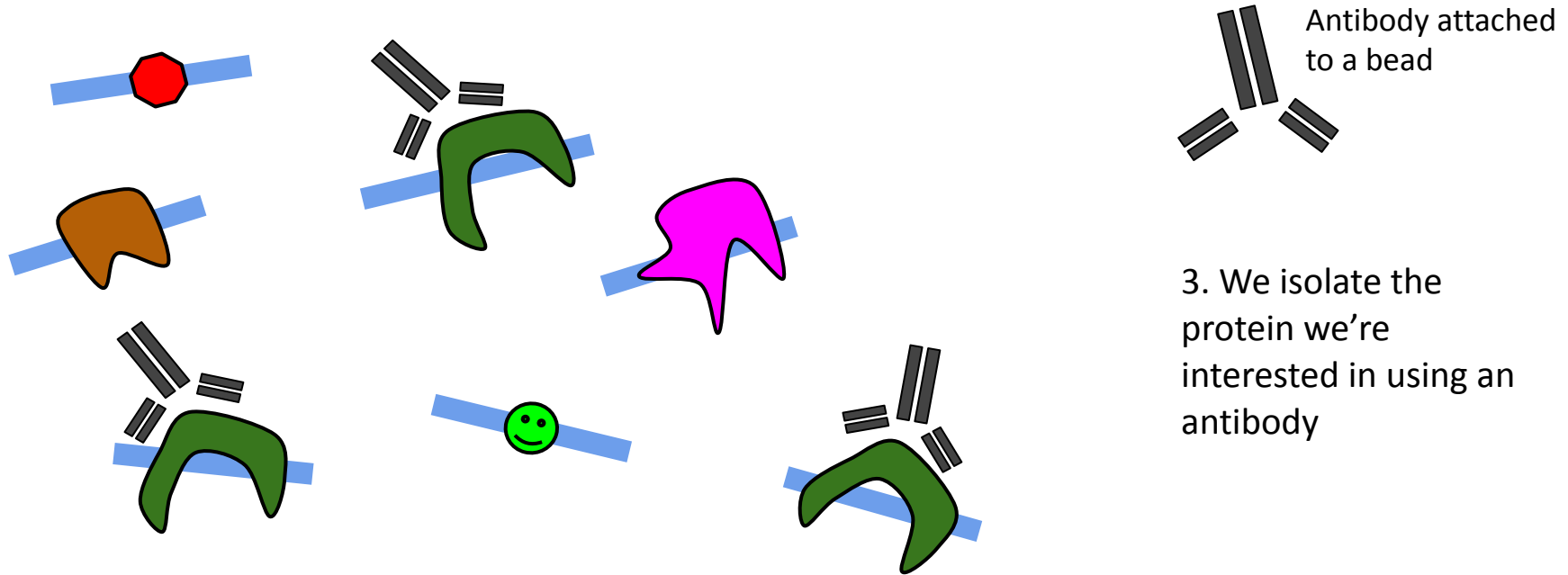
How to analyze the enrichment of TFs and histone marks in the genome

4. We isolate the proteins bound by the first antibody with a second antibody and wash everything else away

How to analyze the enrichment of TFs and histone marks in the genome

5. We reverse the formaldehyde glue by warming up everything and wash away all the proteins, including histones

6. The obtained DNA fragments are sequenced (ChIP-seq) or tested on a microarray platform (ChIP-on-chip)

Hands-on sessions

- [Session 2](#)

- [Session 3.1](#)

Single-end vs. paired-end sequencing experiments
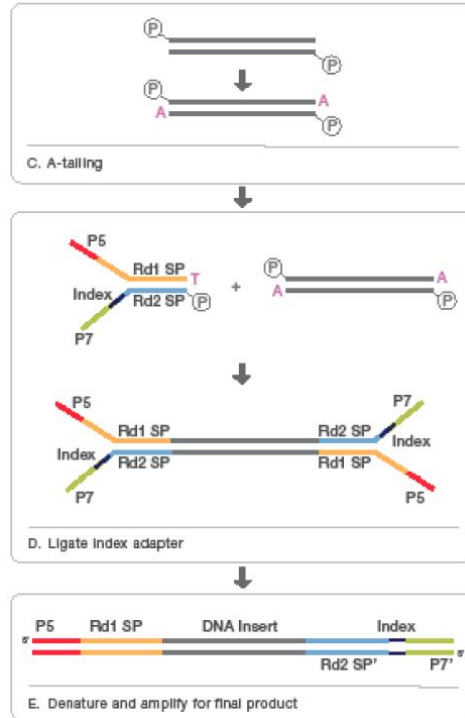
Single-end vs. paired-end sequencing experiments

In single-end ChIP-seq experiments we are sequencing from the 5' end only

Park, 2009

The statistics behind peak calling in MACS2: how much to shift?

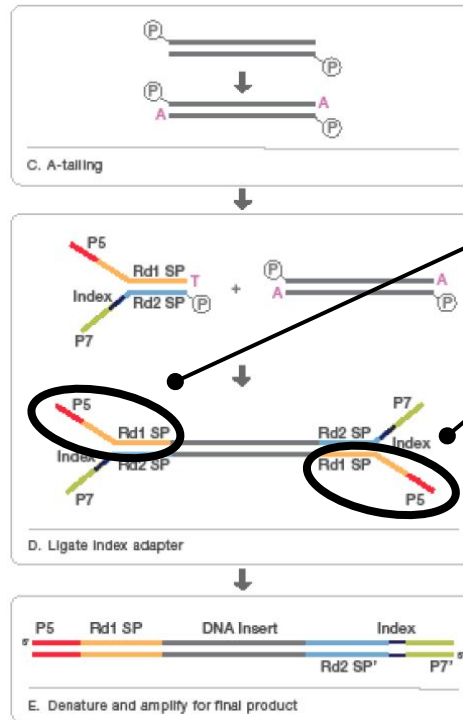Distribution of tags is computed — Reference genome

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
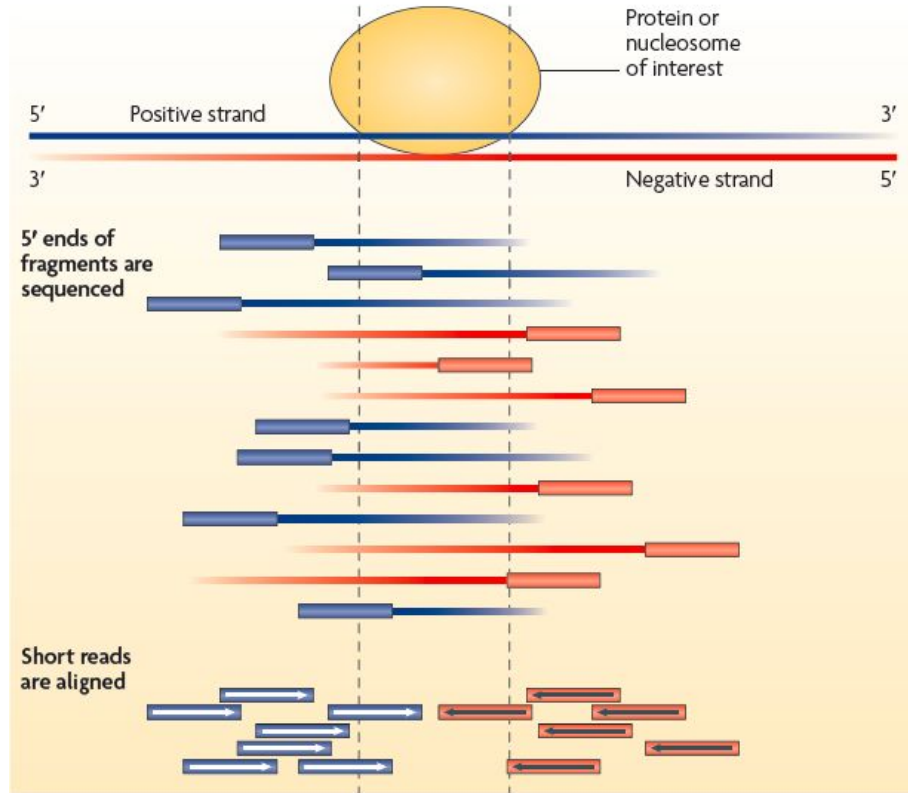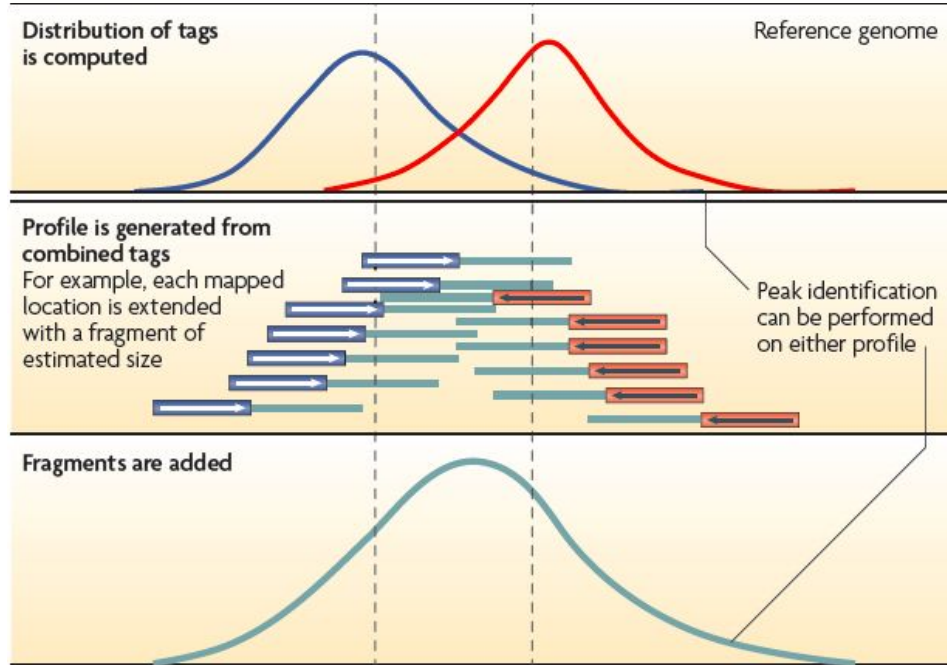
To get the real distribution of reads you can:

1. shift the reads in the direction 5' → 3' (default option)

2. Extend the fragments to reach a fixed fragment length (5' → 3')

- `--no_model` set to TRUE (will not apply the shifting step)
- `--extsize <bp>`

Park, 2009

The statistics behind peak calling in MACS2: how much to shift?

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
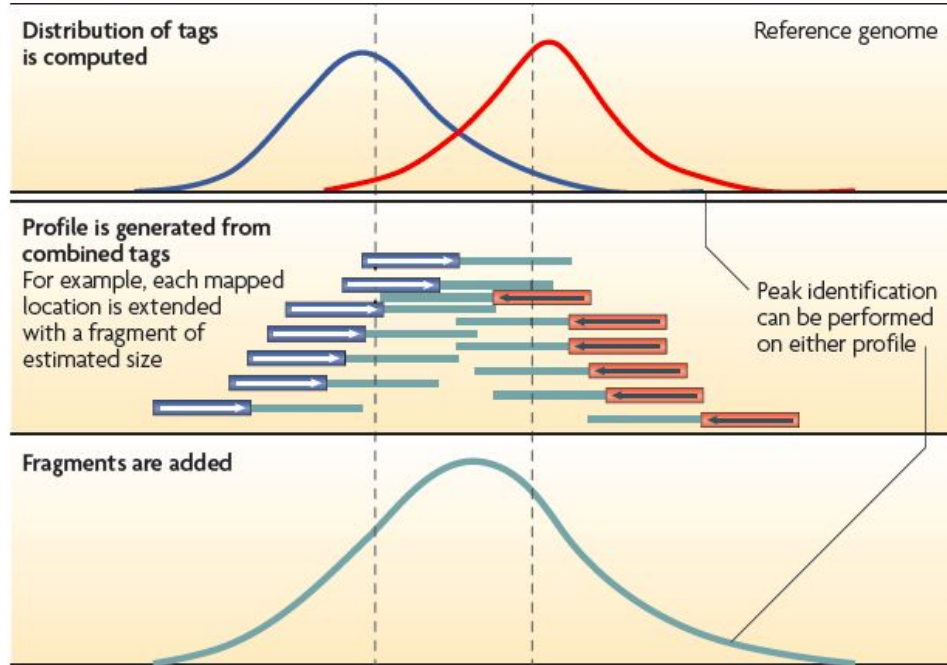
To get the real distribution of reads you can:

1. shift the reads in the direction 5' → 3' (default option)

How much to shift?

2. Extend the fragments to reach a fixed fragment length (5' → 3')

- `--no_model` set to TRUE (will not apply the shifting step)
- `--extsize <bp>`

Park, 2009

The statistics behind peak calling in MACS2: how much to shift?

Running MACS2 function `predictd`, you have to specify:

- the bandwith (`--bw`): half of the sonication size
- a high-confidence fold-enrichment (`--mfold`)

2 * bandwith

ChIP

control

Zhang et al., 2008
https://github.com/taoliu/MACS

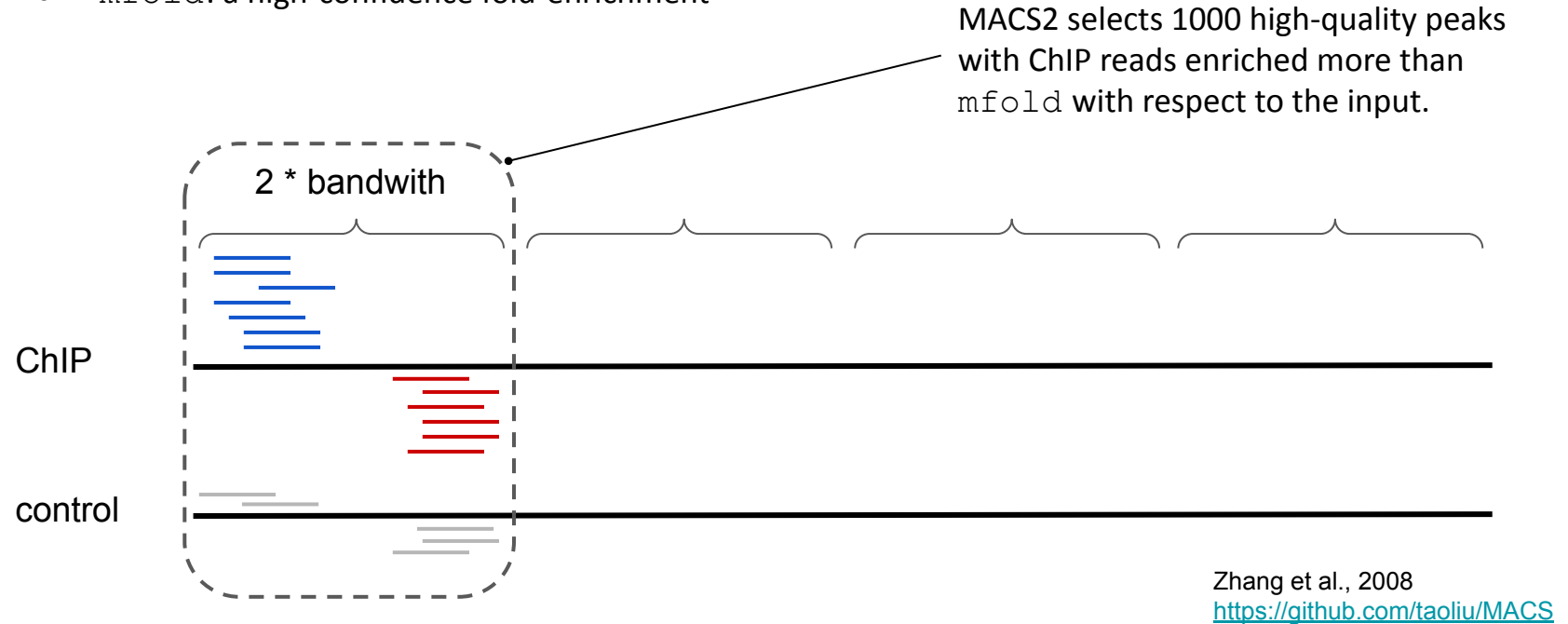# The statistics behind peak calling in MACS2: how much to shift?

Running MACS2 function `predictd`, you have to specify:

- `bandwith`: half of the sonication size
- `mfold`: a high-confidence fold-enrichment

MACS2 selects 1000 high-quality peaks with ChIP reads enriched more than `mfold` with respect to the input.

2 * bandwith

ChIP

control

Zhang et al., 2008
https://github.com/taoliu/MACS

The statistics behind peak calling in MACS2: how much to shift?
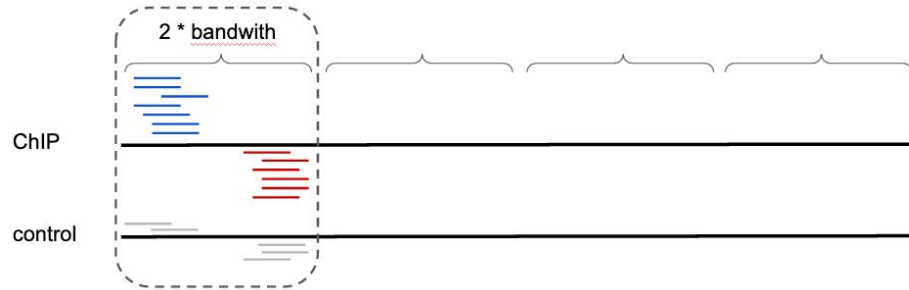
- By aligning the reads of the set of high-quality peaks, MACS2 computes the distance *d* between the summit peaks of the two distributions.

- All the tags will be shifted by d/2 towards the 3'.
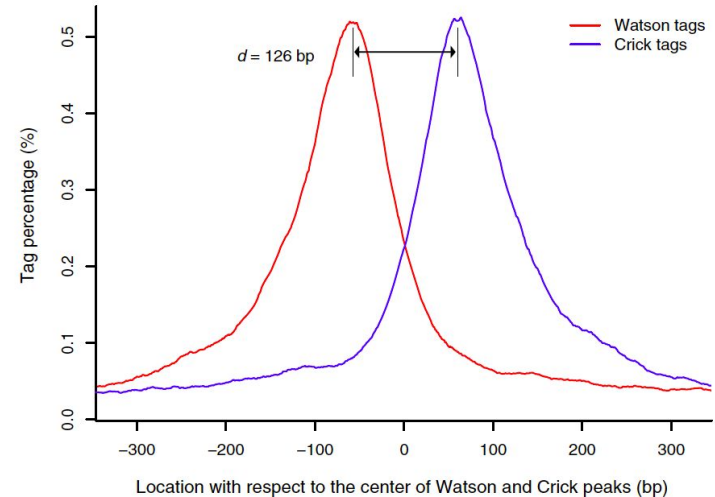
Zhang et al., 2008
https://github.com/taoliu/MACS

Poisson distribution

- gives the probability of a number of events k occurring in a fixed period of time if these events occur with a known average rate (or expected value, λ) and independently of the time since the last event

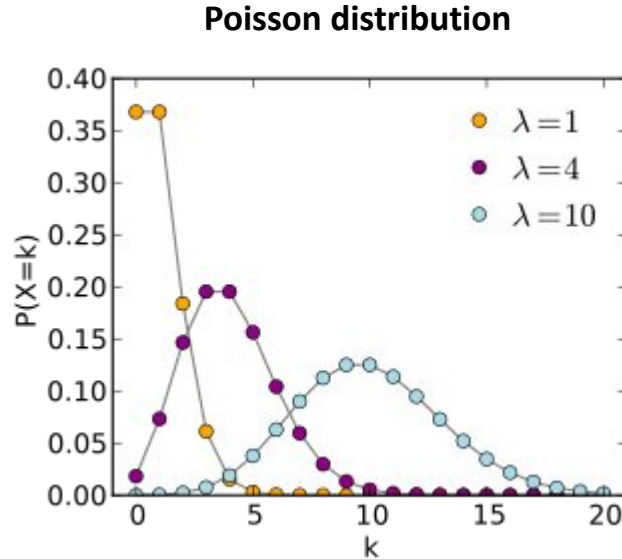- it can also be used for the number of events in other specified intervals such as distance, area or volume

- one parameter (λ) captures both the mean and the variance of the distribution

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Zhang et al., 2008
https://github.com/taoliu/MACS

The statistics behind peak calling in MACS2: how are peaks called?

$$f(k; \lambda) = \Pr(X = k) = \boxed{\frac{\lambda^k e^{-\lambda}}{k!}}$$

2 * **d**

ChIP

k
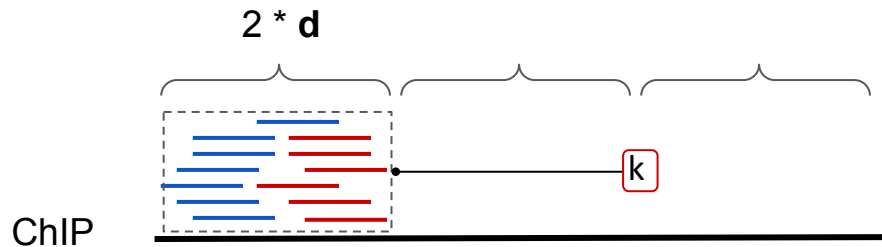
- after shifting the reads by d/2, it slides 2d windows across the genome

- at a given window, the number of events k occurred corresponds to the number of reads found

- what about λ?

Zhang et al., 2008
https://github.com/taoliu/MACS

The statistics behind peak calling in MACS2: how are peaks called?

$$f(k; \lambda) = \Pr(X = k) = \boxed{\frac{\lambda^k e^{-\lambda}}{k!}}$$

2 * d

ChIP

Four types of λ are defined:

- $\lambda_{BG} = \dfrac{\text{total number of \textbf{ChIP reads}}}{\text{genome size}}$

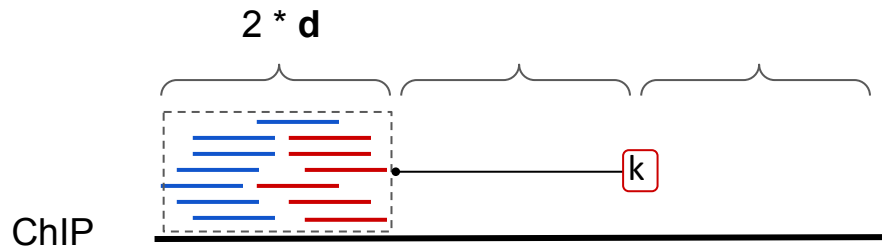- λ_{1k} = estimated with the same formula as λ_{BG} over a window of 1 Kb around the peak summit in the **control**

- λ_{5k} = computed in the control over a window of 5 Kb around the peak summit

- λ_{10k} = computed in the control over a window of 10 Kb around the peak summit

in the control

$$\lambda_{\text{local}} = \boxed{\max(\lambda_{BG}, [\lambda_{1k},] \, \lambda_{5k}, \lambda_{10k})}$$

In the absence of a control, λ_{5k} and λ_{10k} are computed on the ChIP sample, while λ_{1k} is removed from the calculation

Zhang et al., 2008
https://github.com/taoliu/MACS

Metrics to evaluate a ChIP-seq experiment: NRF

Typical ChIP-seq peak

Low-complexity ChIP-seq peak

Landt et al., 2012
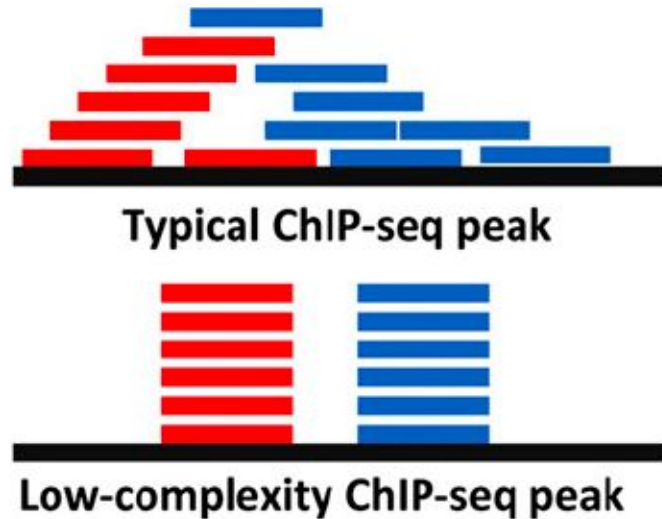
Problems with IP step or library preparation

⬇

PCR-amplification of a limited set of fragments

⬇

High degree of redundancy (or low complexity) in your library

⬇

How to detect this?

Metrics to evaluate a ChIP-seq experiment: NRF

Typical ChIP-seq peak

Low-complexity ChIP-seq peak

Landt et al., 2012

Library complexity:
the fraction of DNA fragments that are non-redundant

Measured by **NRF**

Positions in the genome that uniquely mapped reads map to
_____

total number of uniquely mapped reads

NRF ≥ 0.8 for 10 million uniquely mapped reads
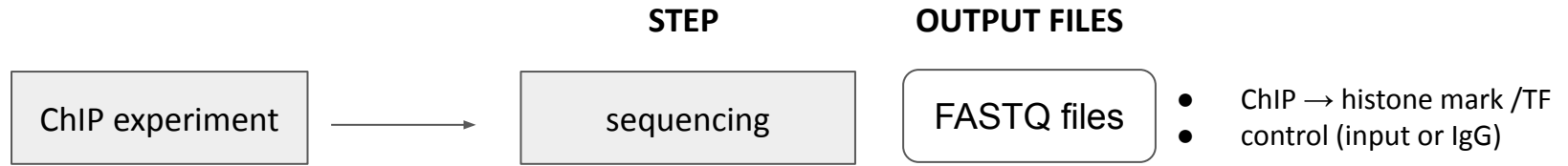
Metrics to evaluate a ChIP-seq experiment: FRiP

The <u>F</u>raction of <u>R</u>eads <u>i</u>n <u>P</u>eaks (FRiP) measures the global enrichment of a ChIP-seq experiment:

$$\frac{\text{Number of mapped reads in peaks}}{\text{Number of mapped reads}}$$

FRiP should be ≥ 0.01 (1%) when calling peaks with MACS2.

Landt et al., 2012

Workflow of ChIP-seq data analysis

STEP

OUTPUT FILES

```
┌─────────────────┐        ┌─────────────────┐        ┌─────────────────┐
│                 │        │                 │        │                 │
│ ChIP experiment │ ─────> │   sequencing    │        │   FASTQ files   │
│                 │        │                 │        │                 │
└─────────────────┘        └─────────────────┘        └─────────────────┘
```

- ChIP → histone mark /TF
- control (input or IgG)

Details can be found here.

Workflow of ChIP-seq data analysis

STEP **OUTPUT FILES**

| ChIP experiment | → | sequencing | FASTQ files | • ChIP → histone mark /TF
• control (input or IgG) |

mapping → BAM files
• ChIP → histone mark /TF
• control (input or IgG)

Details can be found here.

Workflow of ChIP-seq data analysis

STEP **OUTPUT FILES**

ChIP experiment → sequencing

FASTQ files
- ChIP → histone mark /TF
- control (input or IgG)

mapping

BAM files
- ChIP → histone mark /TF
- control (input or IgG)

comparison of ChIP and control → positions where the histone mark / TF is bound in the genome

peak calling

Details can be found here.

Workflow of ChIP-seq data analysis

STEP   **OUTPUT FILES**

ChIP experiment → sequencing → FASTQ files
- ChIP → histone mark /TF
- control (input or IgG)

mapping → BAM files
- ChIP → histone mark /TF
- control (input or IgG)

comparison of ChIP and control → positions where the histone mark / TF is bound in the genome

peak calling

human-readable   binary (compressed)
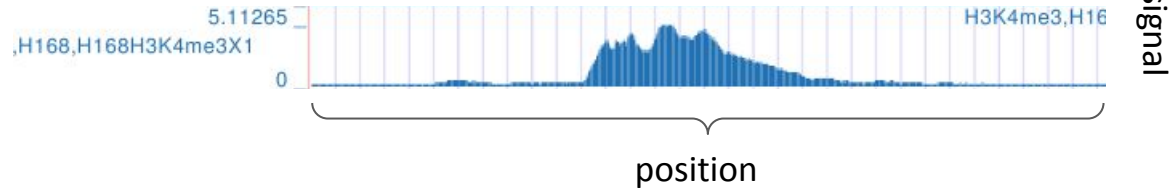
wiggle files → bigWig
BED files → bigBed

Details can be found here.

wiggle (uncompressed) → bigWig (compressed) format

position signal

```
300701  12.5
300702  12.5
300703  12.5
300704  12.5
300705  12.5
```


Details can be found here.

BED (uncompressed) → bigBed (compressed) format

```
chrom     start    end
chr1      778356   779466
chr1      779571   780036
chr1      826622   827025
chr1      827238   827781
chr1      869665   870305
chr1      903908   905506
chr1      909982   910507
chr1      923095   926140
chr1      940046   943376
chr1      958177   961643
chr1      966181   967415
chr1      975903   976702
chr1      997962   1002259
chr1      1012827  1014613
chr1      1019085  1021751
chr1      1024835  1025452
chr1      1032687  1034419
chr1      1040034  1040965
chr1      1041072  1041407
```
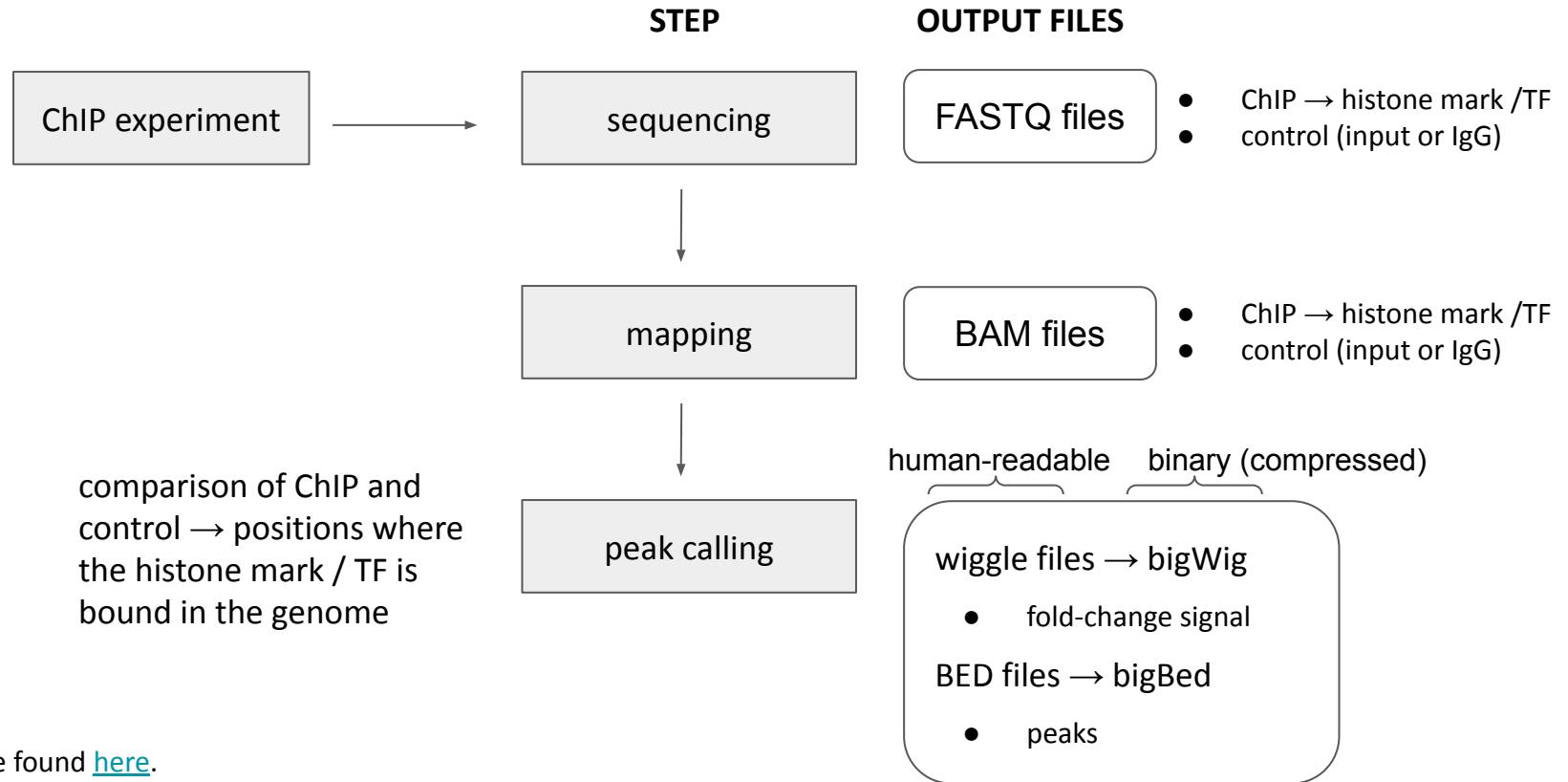
Gives regions (chrom, start, end: compulsory parameters) + additional info (if required).
Can be used to represent genomic segments:

- gene coordinates
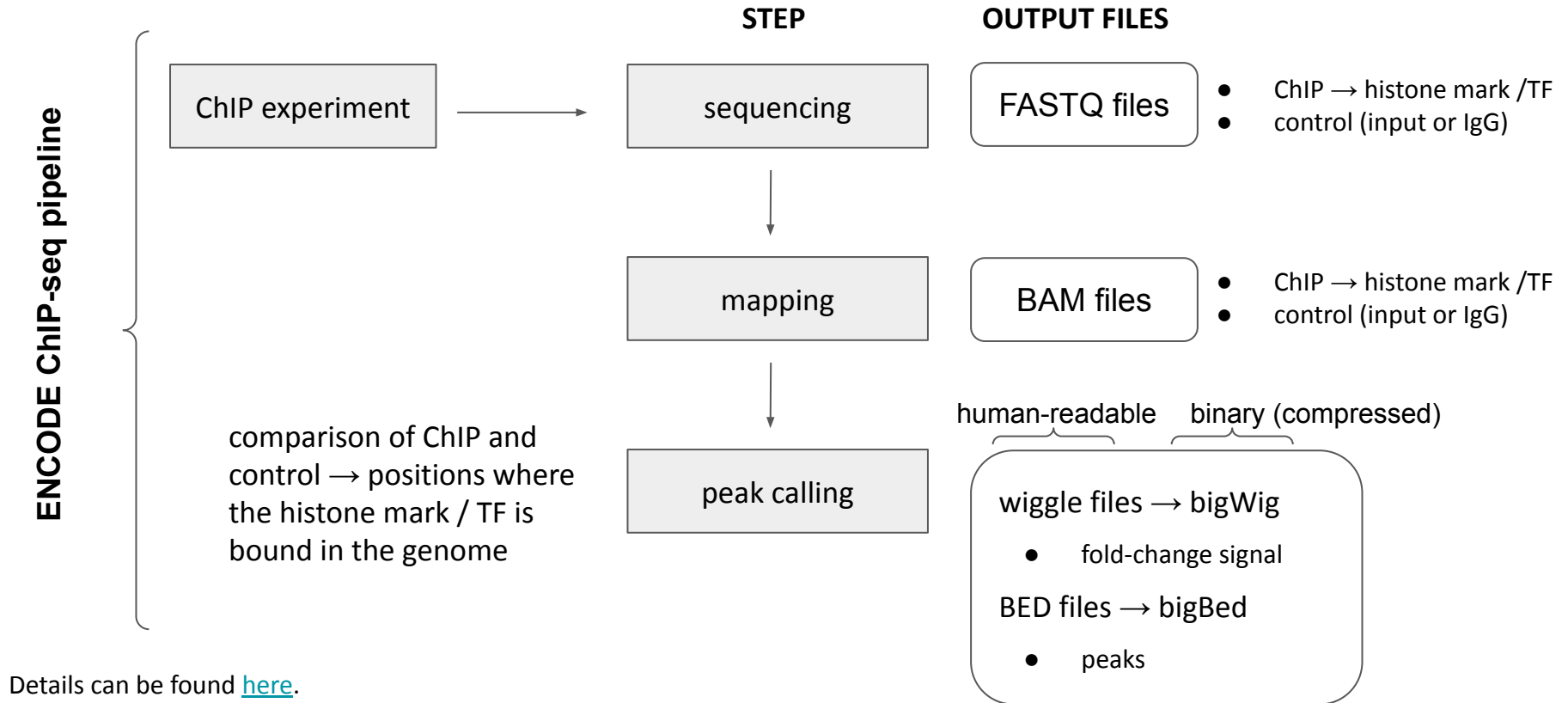- regions where a signal is present (e.g. ChIP-seq peaks)

Details can be found [here](#).

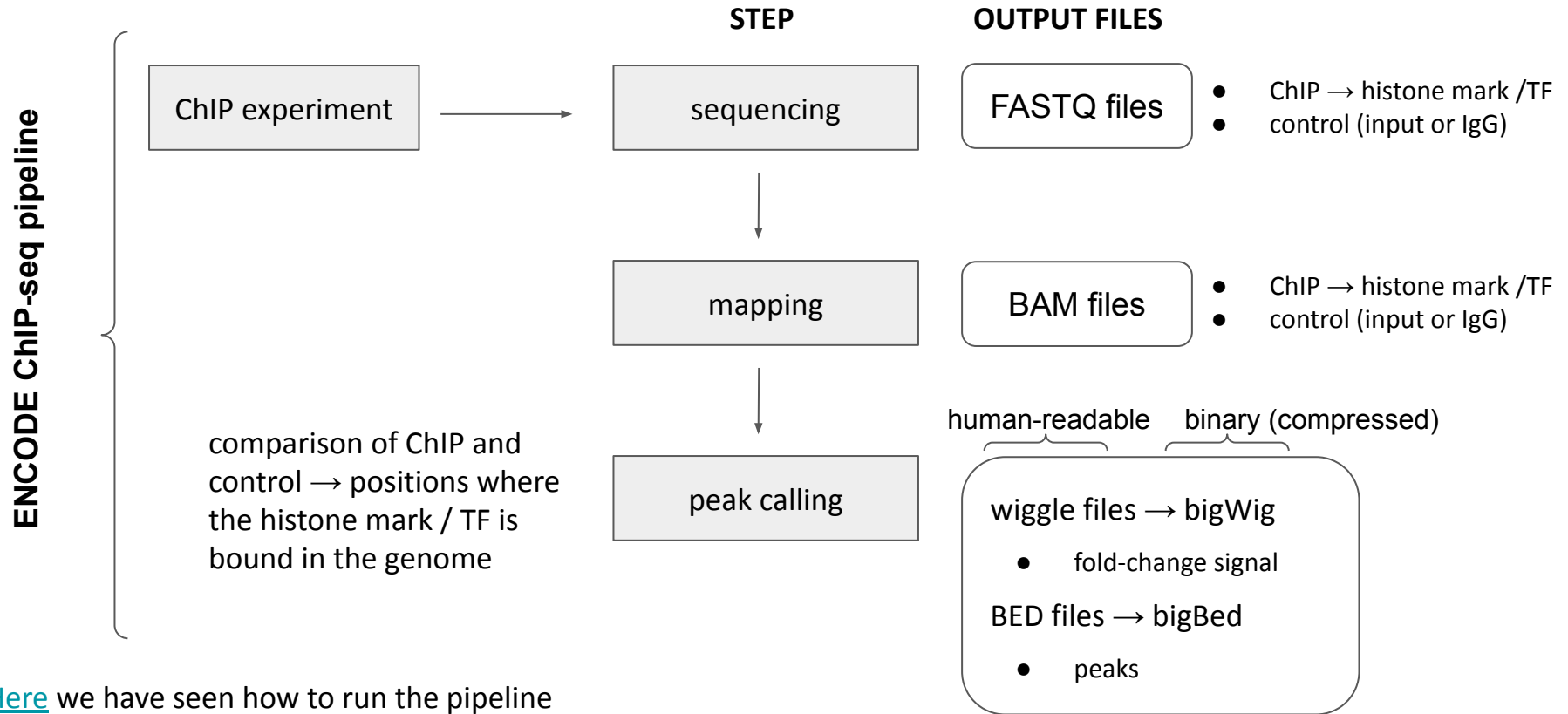# Workflow of ChIP-seq data analysis

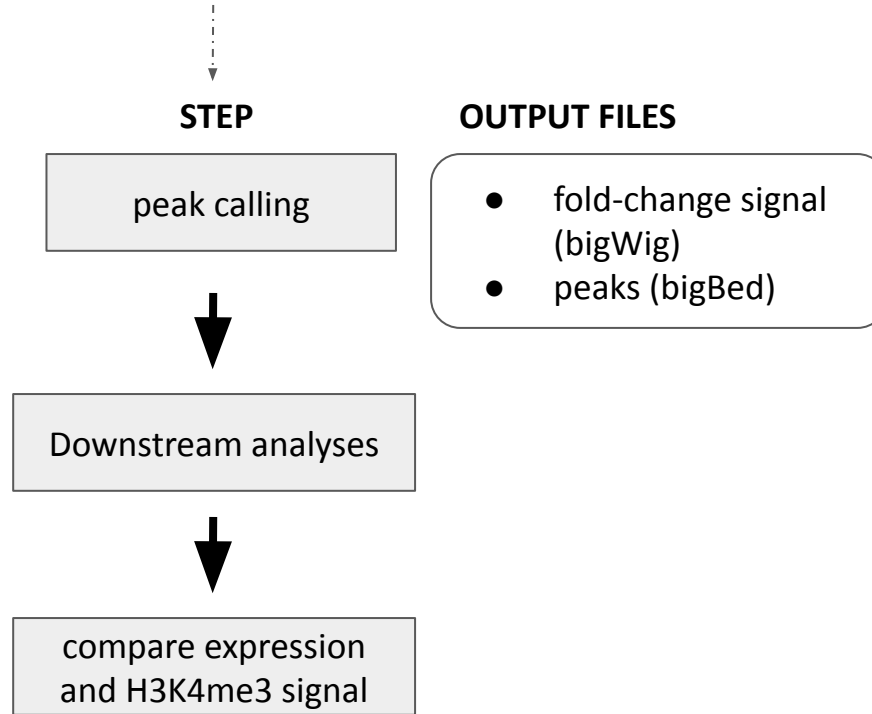Details can be found here.
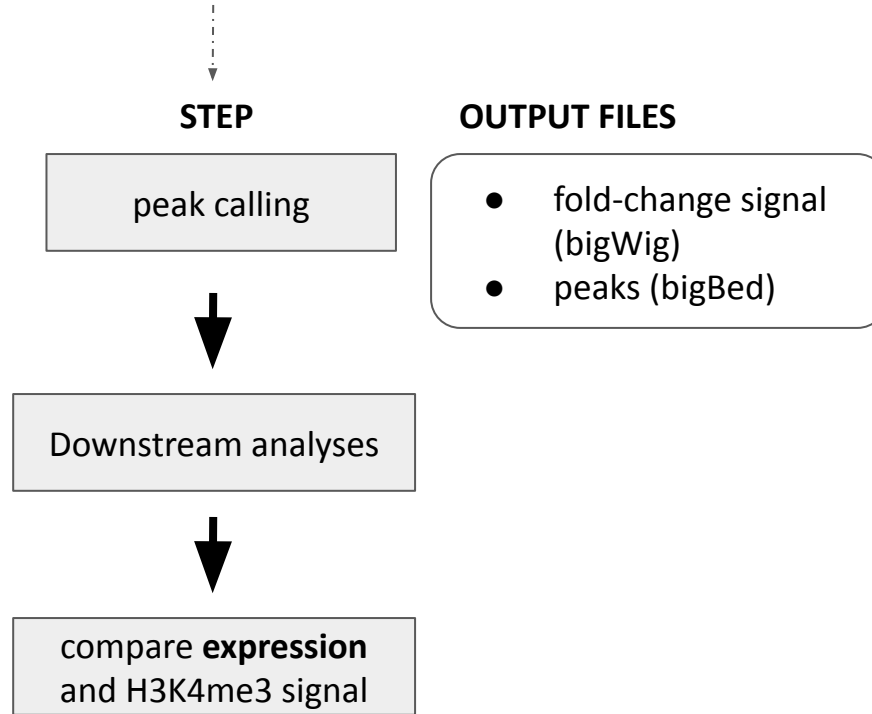
Workflow of ChIP-seq data analysis

ENCODE ChIP-seq pipeline

STEP　　　　**OUTPUT FILES**

ChIP experiment → sequencing

FASTQ files
- ChIP → histone mark /TF
- control (input or IgG)

mapping

BAM files
- ChIP → histone mark /TF
- control (input or IgG)

comparison of ChIP and control → positions where the histone mark / TF is bound in the genome

peak calling

human-readable　　binary (compressed)

wiggle files → bigWig
- fold-change signal

BED files → bigBed
- peaks

Details can be found here.

Workflow of ChIP-seq data analysis

ENCODE ChIP-seq pipeline

STEP

OUTPUT FILES

ChIP experiment → sequencing

FASTQ files
- ChIP → histone mark /TF
- control (input or IgG)

mapping

BAM files
- ChIP → histone mark /TF
- control (input or IgG)

comparison of ChIP and control → positions where the histone mark / TF is bound in the genome

peak calling

human-readable   binary (compressed)

wiggle files → bigWig
- fold-change signal

BED files → bigBed
- peaks

[Here](#) we have seen how to run the pipeline

Downstream analyses

STEP

peak calling

Downstream analyses

compare expression
and H3K4me3 signal

OUTPUT FILES

- fold-change signal
  (bigWig)
- peaks (bigBed)

Downstream analyses

STEP

peak calling

Downstream analyses

compare **expression** and H3K4me3 signal

OUTPUT FILES

- fold-change signal (bigWig)
- peaks (bigBed)

Workflow of RNA-seq data analysis

we will work with TPMs
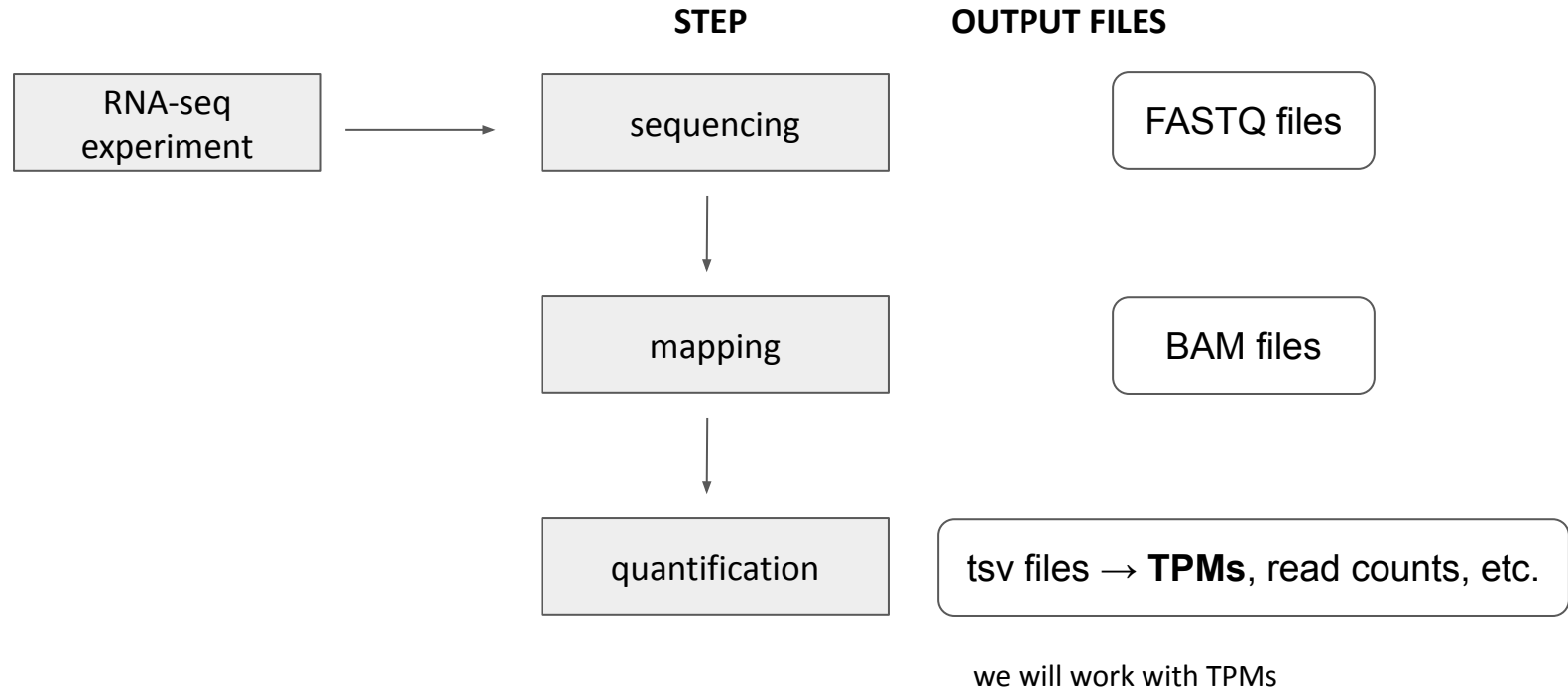
Here we have retrieved the TPM matrices

Downstream analyses
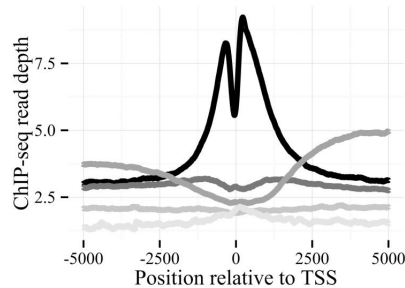
- [Hands-on](#): we'll continue with section 3.2

Downstream analyses

Downstream analyses

↓

compare **expression** and H3K4me3 signal
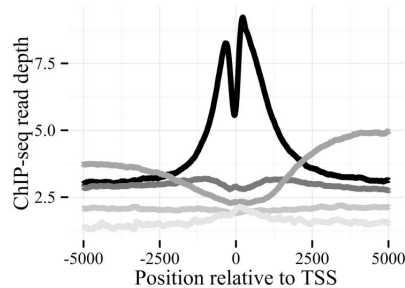
Where is the mark located with respect to a gene?

Downstream analyses

Downstream analyses

↓
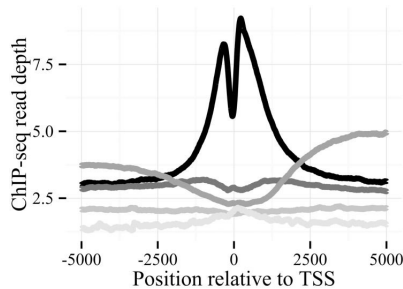
compare **expression** and H3K4me3 signal

Where is the mark located with respect to a gene?
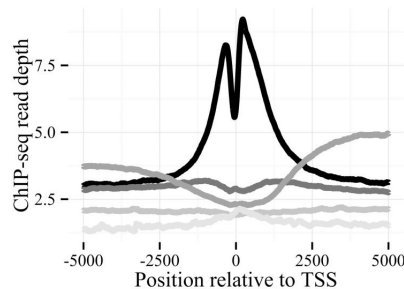
Make your own aggregation plot ([task 2 of section 3.2.](#))

- we have used the genome annotation from Gencode (`tasks 2.1-2.2`)

    - to select protein-coding genes
    - to retrieve the coordinates of the protein-coding genes → BED file

- we have used the TPM matrices to select highly and lowly expressed protein-coding genes in the two tissues (`task 2.3`)

- we have plotted the fold-change signal (bigWig file) over promoter regions (+/- 2 Kb) (`task 2.4`)

Downstream analyses
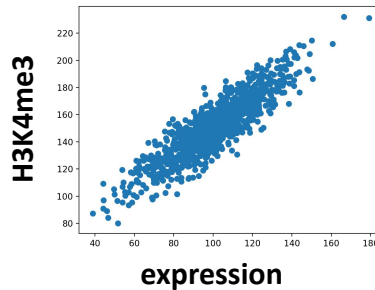
Downstream analyses

Downstream analyses

Chromatin states and the annotation of the genome

The *histone code hypothesis*: specific combinations of chromatin marks encode distinct biological functions

Ernst and Kellis, 2010

Chromatin states and the annotation of the genome

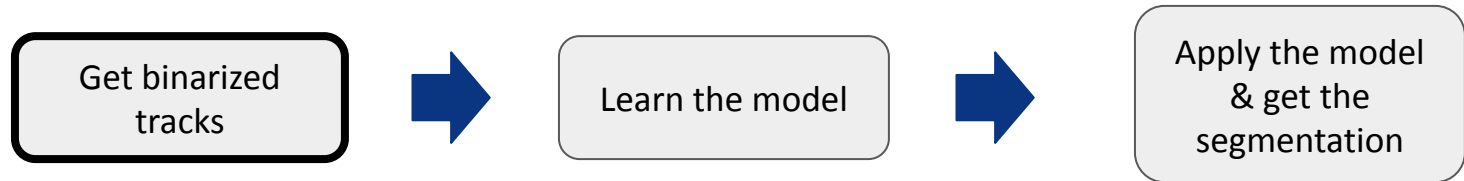| State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	Active promoter
2	12	2	6	9	53	94	95	14	44	1	Weak promoter
3	13	72	0	9	48	78	49	1	10	1	Inactive/poised promoter
4	11	1	15	11	96	99	75	97	86	4	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	Insulator
9	5	0	43	43	37	11	2	9	4	1	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	Polycomb repressed
13	0	0	0	0	0	0	0	0	0	0	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	Repetitive/CNV

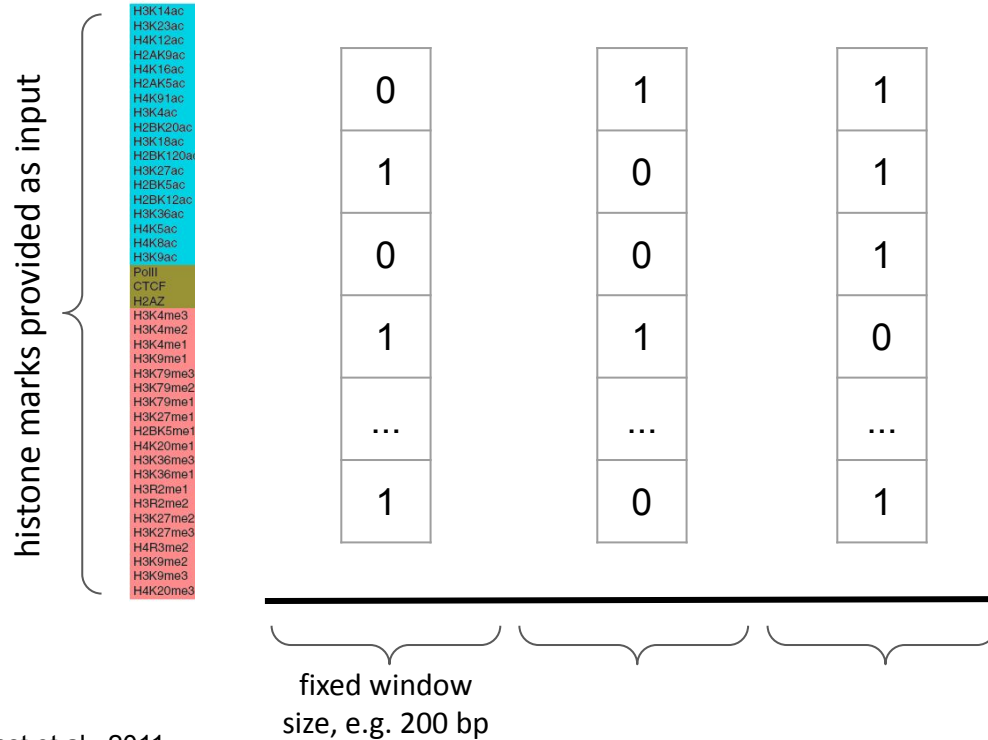Chromatin mark observation frequency (%)

Ernst et al., 2011

- Chromatin state: a combination of histone marks that is biologically meaningful

- ChromHMM is an algorithm based on Hidden Markov Models that segments the genome and assigns chromatin states

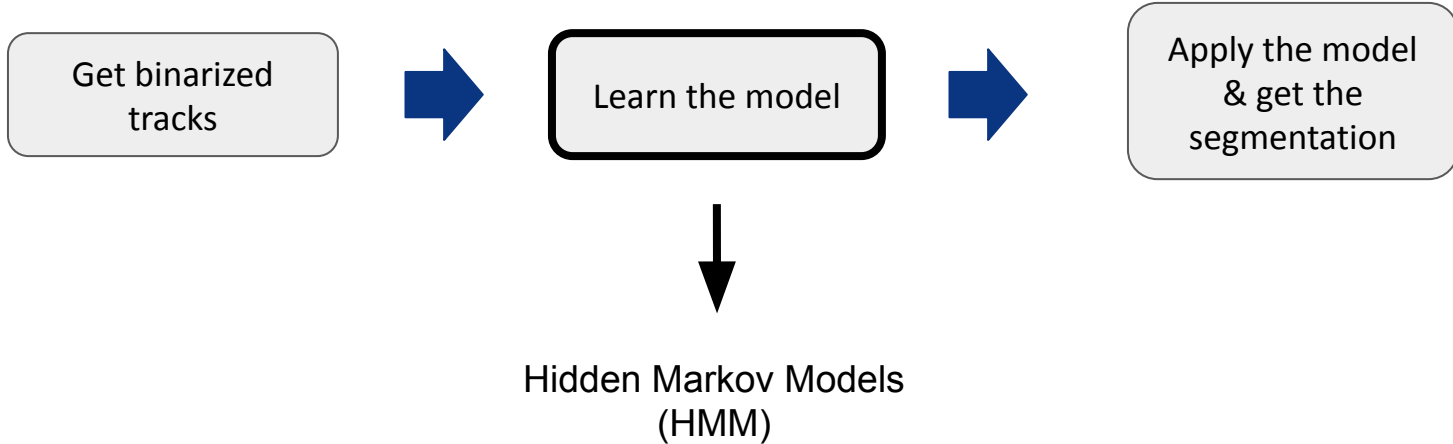# Chromatin states and the annotation of the genome

chromHMM workflow:

Get binarized tracks → Learn the model → Apply the model & get the segmentation

Chromatin states and the annotation of the genome

fixed window size, e.g. 200 bp

histone marks provided as input

Ernst et al., 2011

- The input files are mapped reads, either in BED or BAM format
  - controls are needed as well
  - better to use uniquely mapped reads

- At each region, chromHMM assigns a binary vector of presence / absence of the input marks, similarly to the peak calling procedure

Chromatin states and the annotation of the genome

chromHMM workflow:

Get binarized tracks → Learn the model → Apply the model & get the segmentation

Learn the model ↓

Hidden Markov Models (HMM)

Chromatin states and the annotation of the genome

Markov chain:

- a stochastic model describing a sequence of events in which the probability of each event depends only on the state recorded in the previous event.

- example: register the weather condition day by day

  - if we treat it as a Markov chain, the weather condition in a day depends ONLY on the weather conditions in the day before

- The probability for the 5-days registration "CRRCS" is:

  $P(CRRCS) = P(C) \cdot P(R|C) \cdot P(R|R) \cdot P(C|R) \cdot P(S|C)$

- Some biological examples:

  - presence / absence of CpG islands
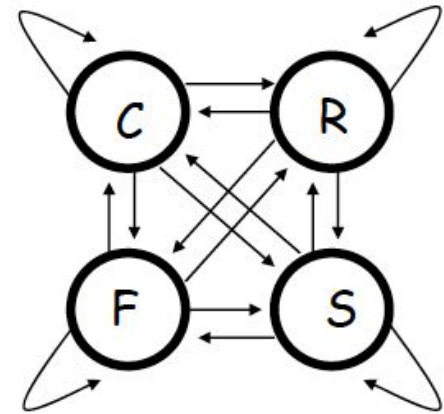 - protein secondary structure (sequence of α chains and β sheets)
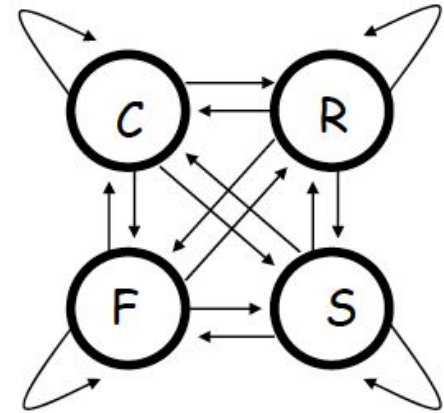
states
C: Clouds
R: Rain
F: Fog
S: Sun

Chromatin states and the annotation of the genome

Hidden Markov Model:

- differently from the Markov chain, in this case the sequence of states is unknown (hidden).
- the goal of a HMM is to infer the sequence of states by interpreting an observable sequence
  - example:
    - observable sequence: primary sequence of a protein (aa residues)
    - hidden path: secondary structure (alternation of α chains and β sheets)
    - question: which is the probability that my observed aa residue (lysine) belongs to a α chain?
  - in our case, we have a <u>multivariate</u> profile (not just one histone mark, but a combination of histone marks)
    - observable sequence: combinations of histone marks (binary presence / absence vector)
   - hidden path: genome annotation in chromatin states

states C: Clouds
R: Rain
F: Fog
S: Sun

Chromatin states and the annotation of the genome

In the case of a Hidden Markov Model:

- before reconstructing the sequence of states (hidden path), you have to *learn* about them:

    - i.e., understand the characteristics of each state
    - in our case: which marks define a specific state?

- During the learning step, it defines:

    - **emission** probabilities: the probability of a histone mark to belong to a specific state

        - e.g. probability of observing a peak of H3K27ac and being in state 8
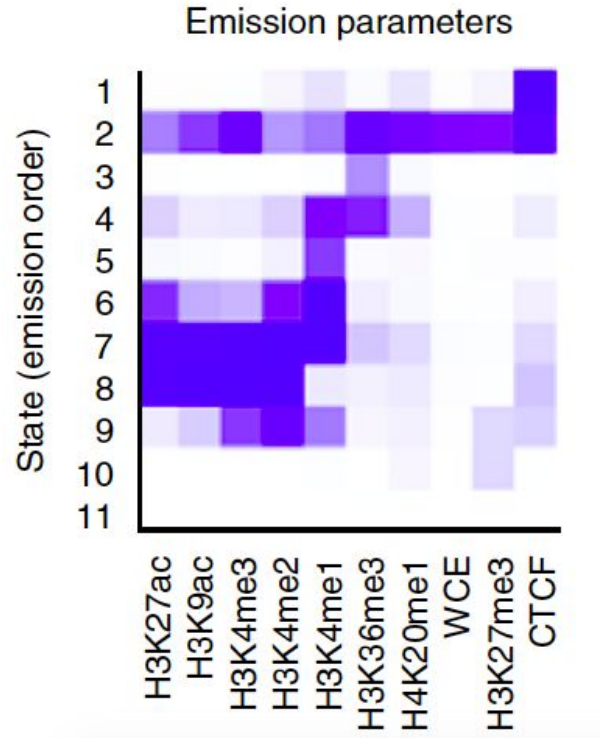
Emission parameters

Ernst and Kellis, 2017

In the case of chromHMM (multivariate HMM):

- You have to specify the number of chromatin states (e.g. 11)

- Besides emission probabilities, during the learning step, it defines:

  - **transition** probabilities: the probability of going from state A in position *i* to state B in position *i+1*

    - e.g. probability that I am in state 8 coming from state 2

Transition parameters

Ernst and Kellis, 2017
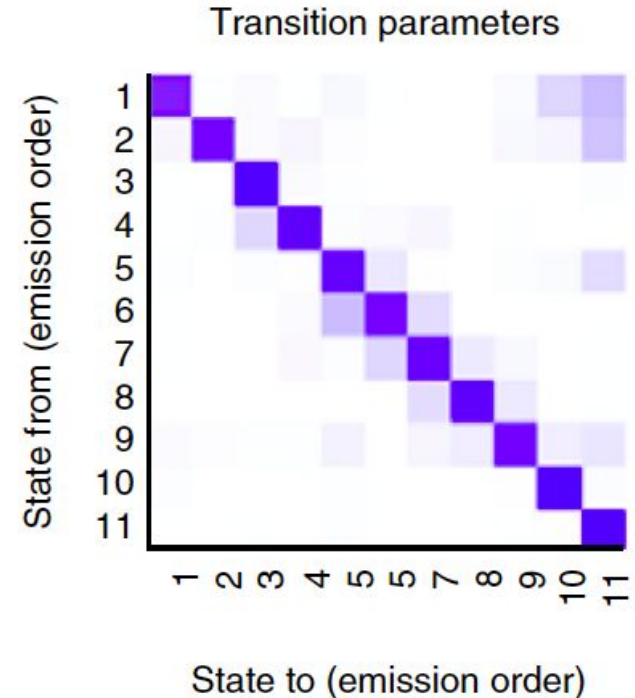
Chromatin states and the annotation of the genome

In the case of chromHMM (multivariate HMM):

- you can learn transition and emission probabilities in one cell type (e.g. K562) and apply the learnt model to another cell type (e.g. HeLa-S3)

- you can learn the model in one cell type and apply it to segment the genome in the same cell type

Chromatin states and the annotation of the genome

chromHMM workflow:

| Get binarized tracks	→	Learn the model	→	Apply the model & get the segmentation
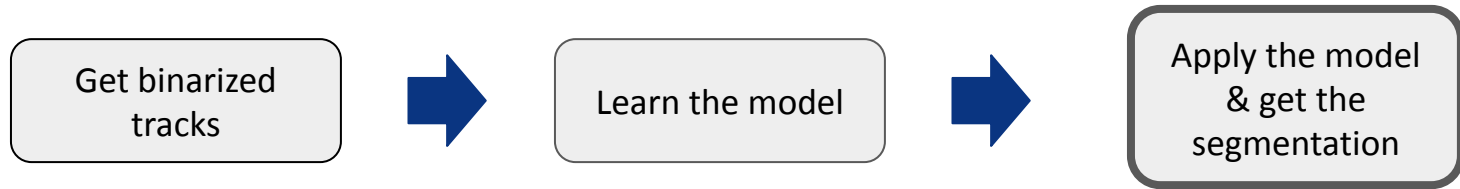
Chromatin states and the annotation of the genome

After the learning step, chromHMM:

- Reconstructs the genome annotation in chromatin states
    - For each genomic segment, computes a posterior probability over different states using a forward-backward algorithm, and assigns the most probable state
- A tutorial on how to run chromHMM can be found in this paper: https://www.ncbi.nlm.nih.gov/pubmed/29120462

Chromatin states and the annotation of the genome

Comparing chromatin states annotation across different cell lines:
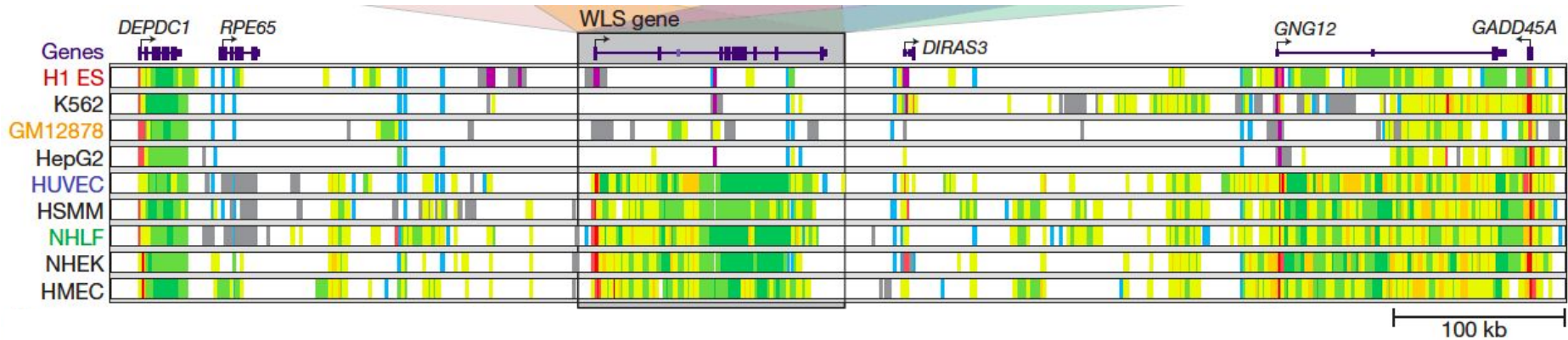
Ernst et al., 2011

Hands-on session

- [Hands-on session 5](#)

- Contact: beatrice.borsari@crg.eu

References

- References:
  - Park 2009, Nat Rev Genet
  - Zhang et al. (2008), Genome Biol
  - Landt et al. (2012), Genome Res
  - Ernst and Kellis (2010), Nat Biotechnol
  - Ernst et al. (2011), Nature
  - Ernst and Kellis (2017), Nat Protoc