ChIP-seq analysis

J. van Helden, M. Defrance, C. Herrmann, D. Puthier, N. Servant

- Wednesday :
  - quick introduction to ChIP-seq and peak-calling (Presentation + Practical session)
  - Introduction to concepts of motif finding (Presentation)
- Thursday:
  - motif discovery in ChIP-seq peaks (Practical)
  - ChIP-peaks functional annotation (Presentation + Practical)

Datasets used

Research

GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility

Vasiliki Theodorou,[1] Rory Stark,[2] Suraj Menon,[2] and Jason S. Carroll[1,3,4]

[1]Nuclear Receptor Transcription Lab, [2]Bioinformatics Core, Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom; [3]Department of Oncology, University of Cambridge, Cambridge CB2 OXZ, United Kingdom

- estrogen-receptor (ESR1) is a key factor in breast cancer developement
- goal of the study: understand the dependency of ESR1 binding on presence of co-factors, in particular GATA3, which is mutated in breast cancers
- approaches: GATA3 silencing (siRNA), ChIP-seq on ESR1 in wt vs. siGATA3 conditions, chromatin profiling

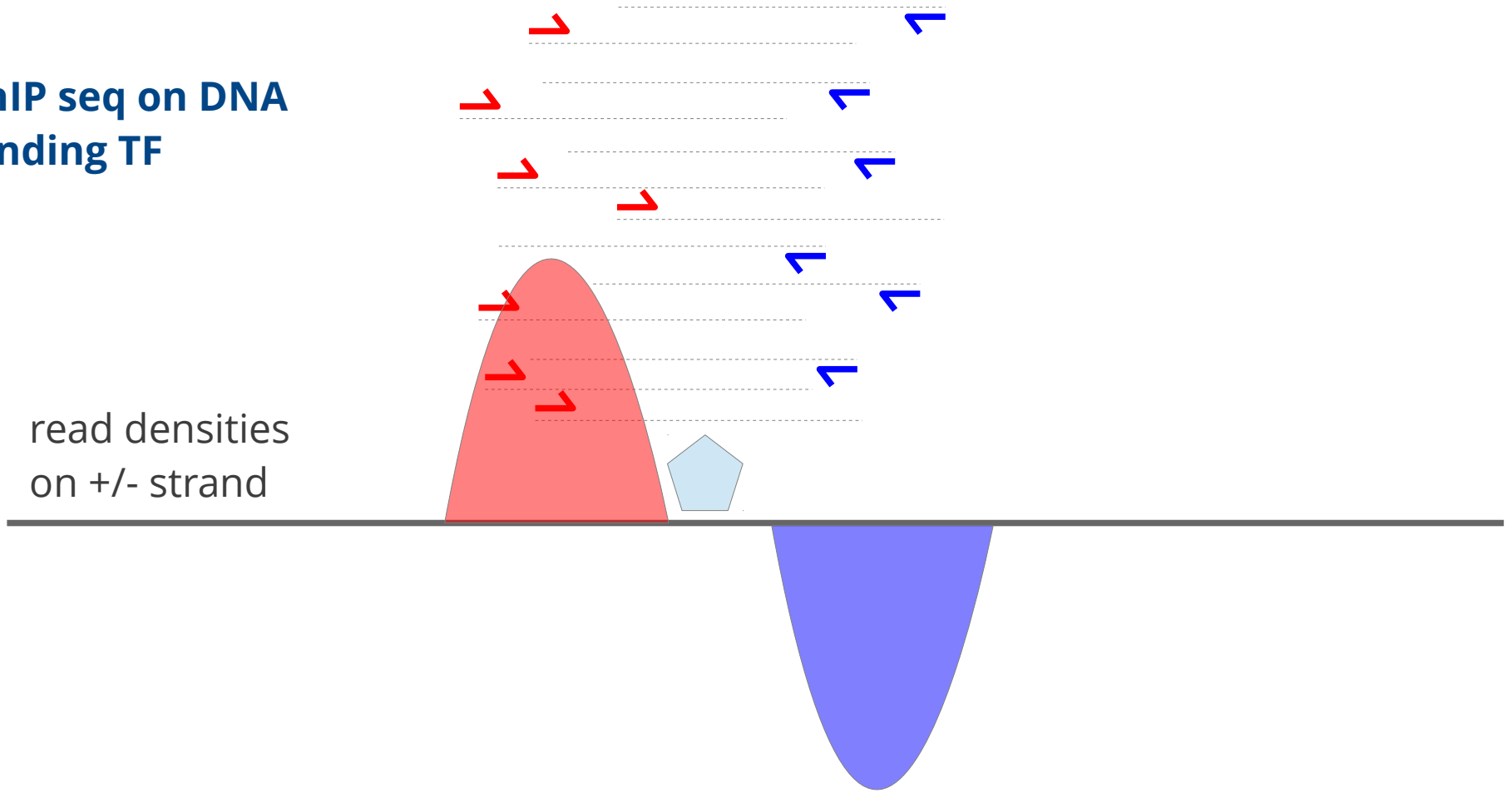
Datasets used

ExpName	CellLine	Replicate	SampleID	SRAExpID	Selected
siNT_ER_E2_r1	MCF-7	r1	GSM986059	SRX176856	X
siGATA_ER_E2_r1	MCF-7	r1	GSM986060	SRX176857	X
siNT_ER_E2_r2	MCF-7	r2	GSM986061	SRX176858	X
siGATA_ER_E2_r2	MCF-7	r2	GSM986062	SRX176859	X
siNT_ER_E2_r3	MCF-7	r3	GSM986063	SRX176860	X
siGATA_ER_E2_r3	MCF-7	r3	GSM986064	SRX176861	X
siNT_FOXA1_Veh_r1	MCF-7	r1	GSM986065	SRX176862	
siGATA_FOXA1_Veh_r1	MCF-7	r1	GSM986066	SRX176863	
GATA3_E2_r1	MCF-7	r1	GSM986067	SRX176864	
GATA3_Veh_r1	MCF-7	r1	GSM986068	SRX176865	
GATA3_E2_r2	MCF-7	r2	GSM986069	SRX176866	
GATA3_Veh_r2	MCF-7	r2	GSM986070	SRX176867	
GATA3_E2_r3	MCF-7	r3	GSM986071	SRX176868	
GATA3_Veh_r3	MCF-7	r3	GSM986072	SRX176869	
GATA3_E2_r4	MCF-7	r4	GSM986073	SRX176870	
GATA3_Veh_r4	MCF-7	r4	GSM986074	SRX176871	
GATA3_E2_r5	MCF-7	r5	GSM986075	SRX176872	
GATA3_Veh_r5	MCF-7	r5	GSM986076	SRX176873	
siNT_H3K27ac_E2_r1	MCF-7	r1	GSM986077	SRX176874	
siGATA_H3K27ac_E2_r1	MCF-7	r1	GSM986078	SRX176875	
siNT_H3K27ac_Veh_r1	MCF-7	r1	GSM986079	SRX176876	
siGATA_H3K27ac_Veh_r1	MCF-7	r1	GSM986080	SRX176877	
siNT_H3K4me1_E2_r1	MCF-7	r1	GSM986081	SRX176878	X
siGATA_H3K4me1_E2_r1	MCF-7	r1	GSM986082	SRX176879	X
siNT_H3K4me1_Veh_r1	MCF-7	r1	GSM986083	SRX176880	
siGATA_H3K4me1_Veh_r1	MCF-7	r1	GSM986084	SRX176881	
siNT_p300_E2_r2	MCF-7	r2	GSM986085	SRX176882	
siGATA_p300_E2_r2	MCF-7	r2	GSM986086	SRX176883	
siNT_p300_Veh_r2	MCF-7	r2	GSM986087	SRX176884	
siGATA_p300_Veh_r2	MCF-7	r2	GSM986088	SRX176885	
ZR751_siNT_ER_E2_r1	ZR751	r1	GSM986089	SRX176886	
ZR751_siGATA_ER_E2_r1	ZR751	r1	GSM986090	SRX176887	
MCF-7_input_r3	MCF-7	r3	GSM986091	SRX176888	X
ZR751_input_r1	ZR751	r1	GSM986092	SRX176889	
ZR751_input_r1	ZR751	r1	GSM986092	SRX176889	

- **ESR1 ChIP-seq in WT & siGATA3 conditions**
  (3 replicates = 6 datasets)
- **H3K4me1 in WT & siGATA3 conditions**
  (1 replicate = 2 datasets)
- **Input dataset in MCF-7**
  (1 replicate = 1 dataset)
- p300 before estrogen stimulation
- GATA3/FOXA1 ChIP-seq
  before/after estrogen stimulation
- microarray expression data, etc ...

ChIP-seq signal for transcription factors

ChIP seq on DNA binding TF



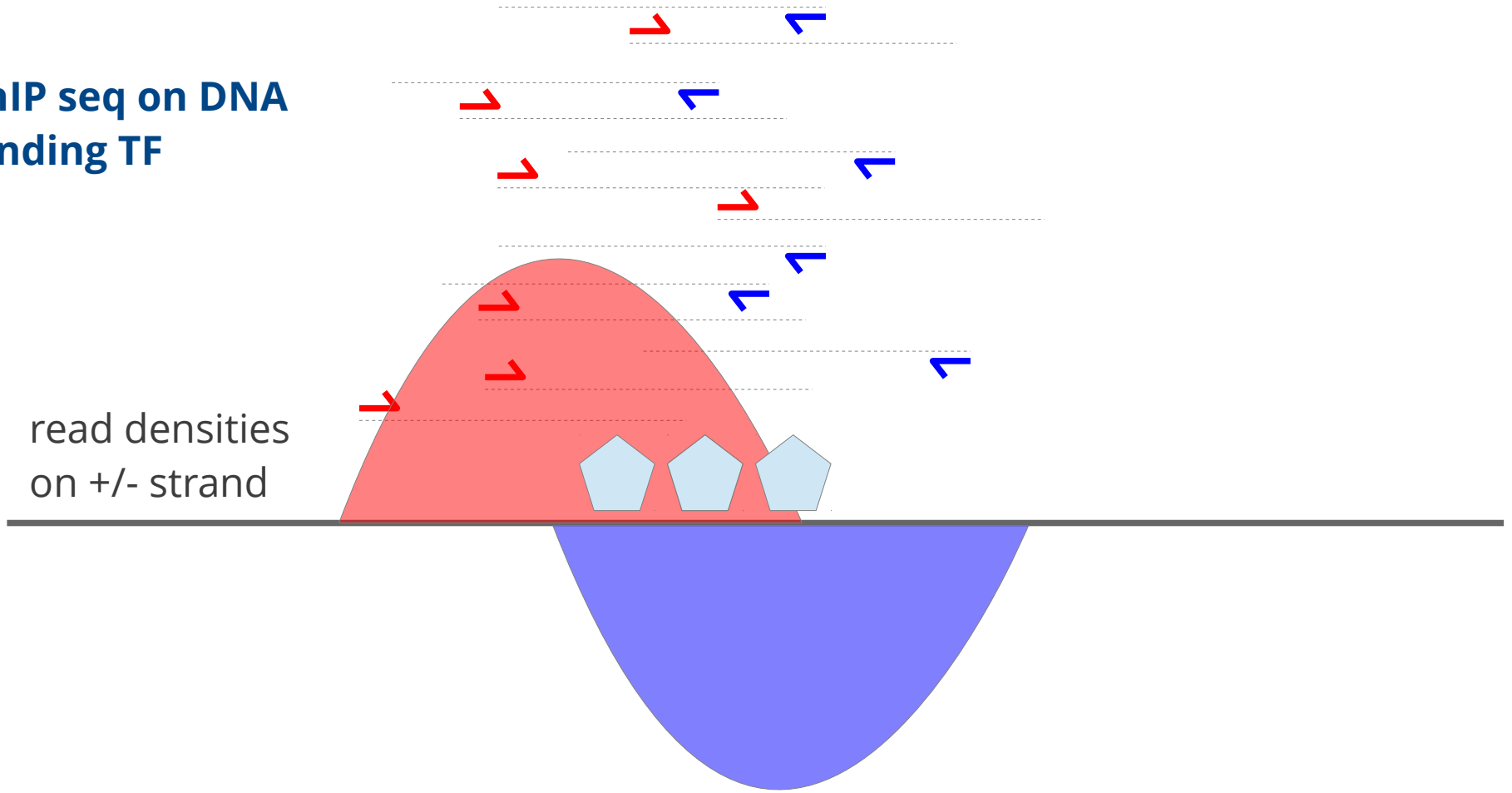
We expect to see a typical strand asymetry in read densities
→ ChIP peak recognition pattern

ChIP-seq signal for transcription factors

(this is the data you are goind to manipulate ...)

ChIP-seq signal for transcription factors

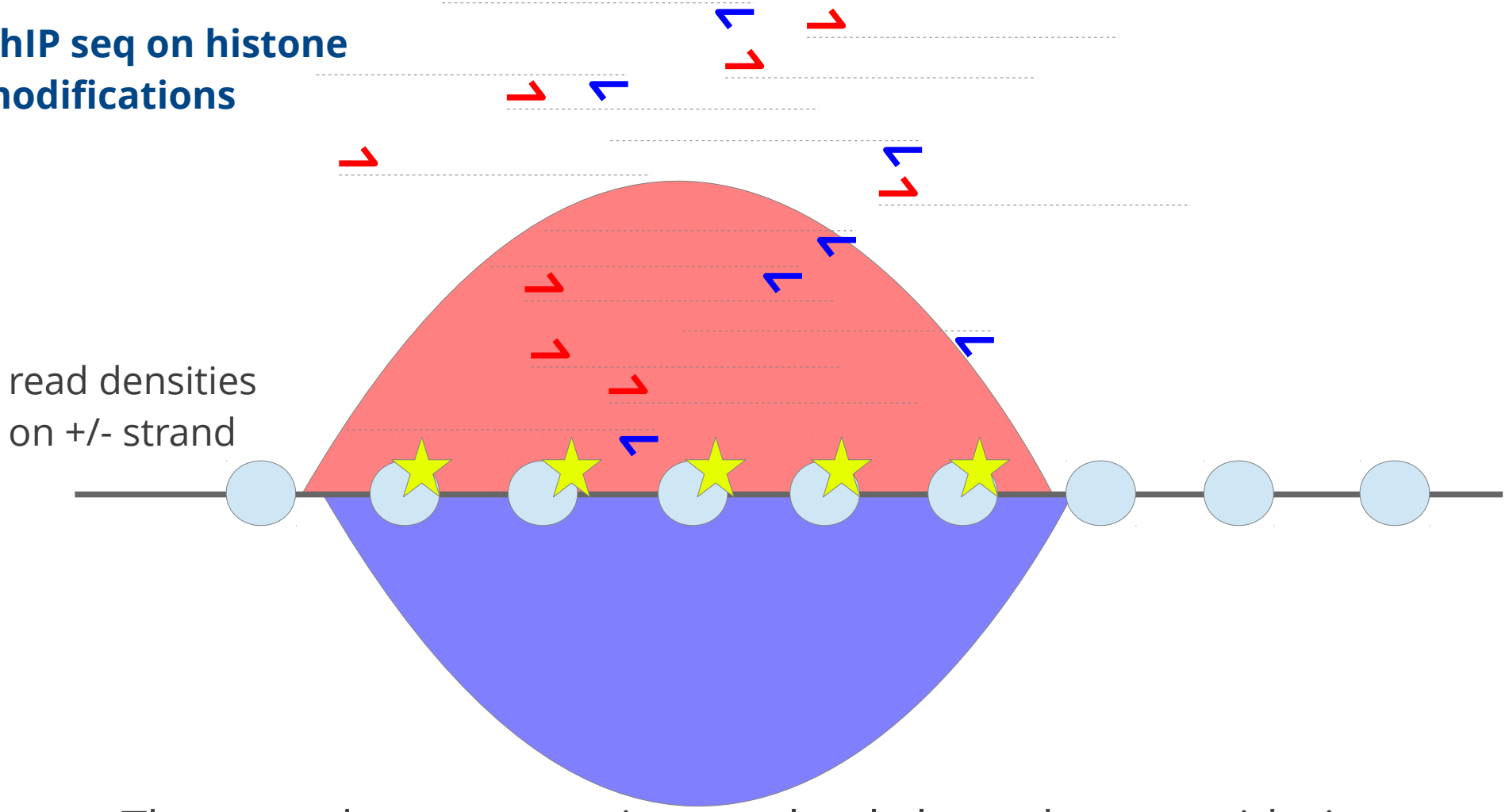
ChIP seq on DNA binding TF



Binding of several TF as complexes tend to blur this asymmetry

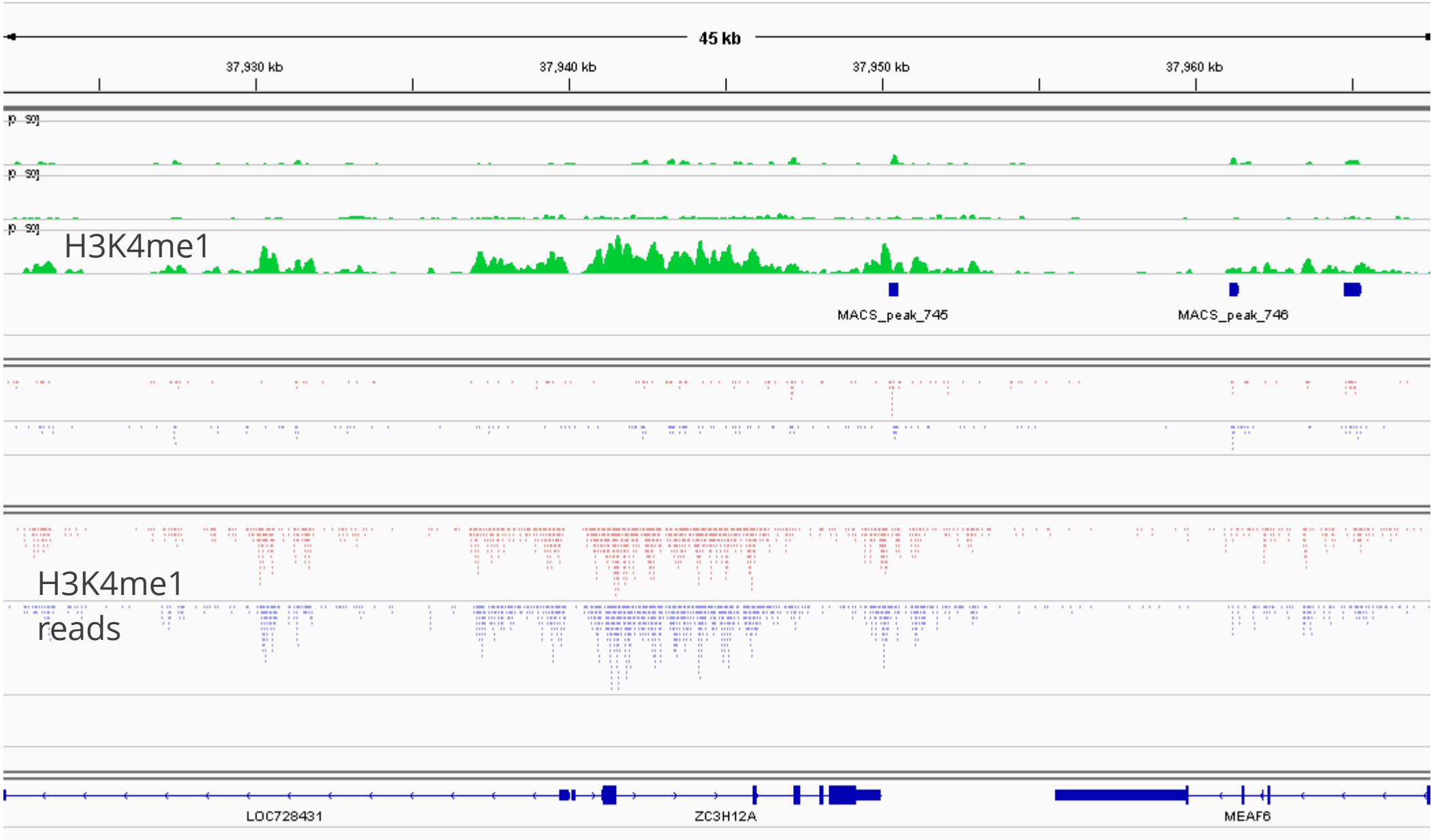
ChIP-seq signal for histone marks

ChIP seq on histone modifications

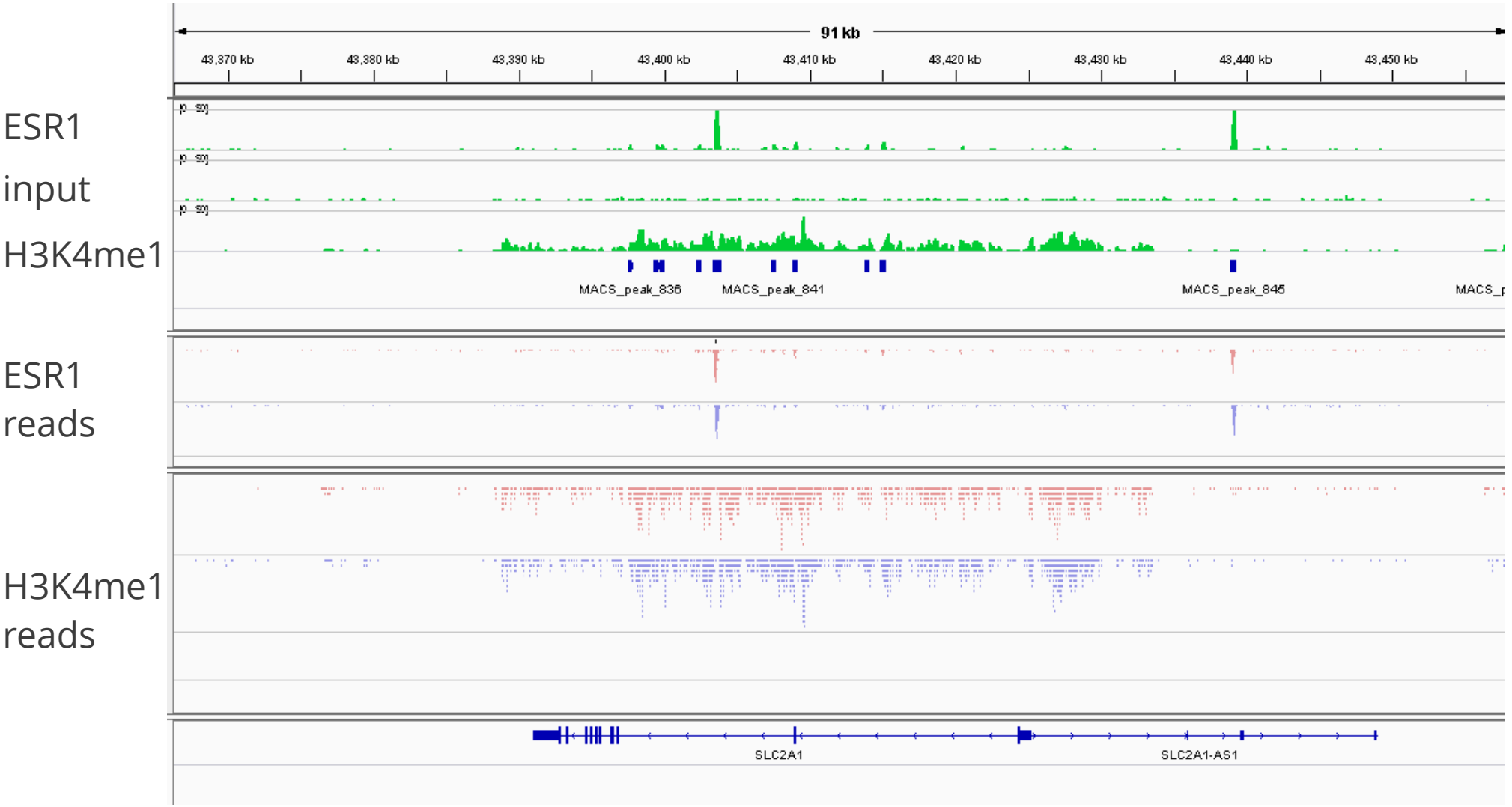


The strand asymmetry is completely lost when considering ChIP datasets for diffuse histone modifications

Real example of ChIP-seq signal



Real example of ChIP-seq signal

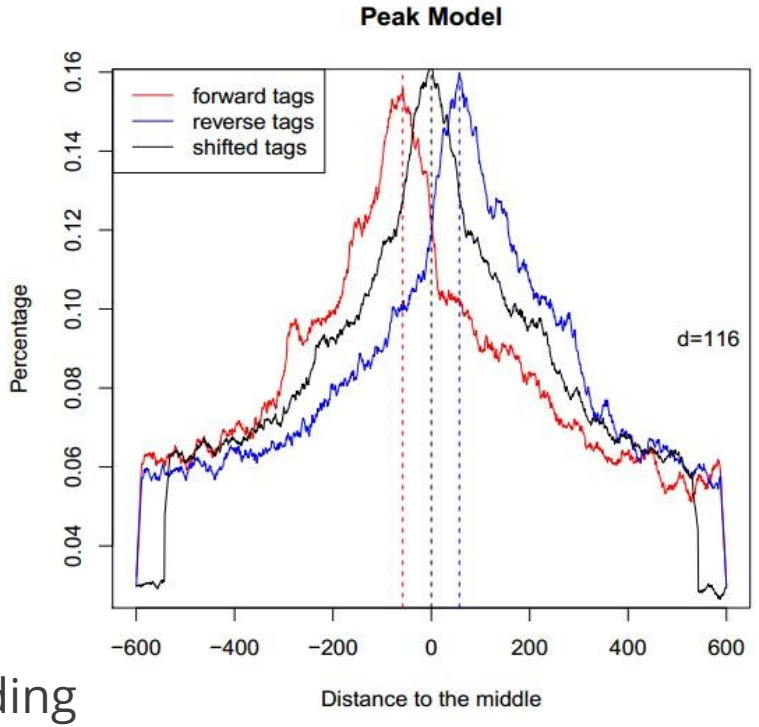


Keys aspects of “peak” finding

- Treating the reads
- Modelling noise levels
- Scaling datasets
- Detecting enriched/peak regions
- Dealing with replicates (→ Exercices)

From aligned reads to binding sites

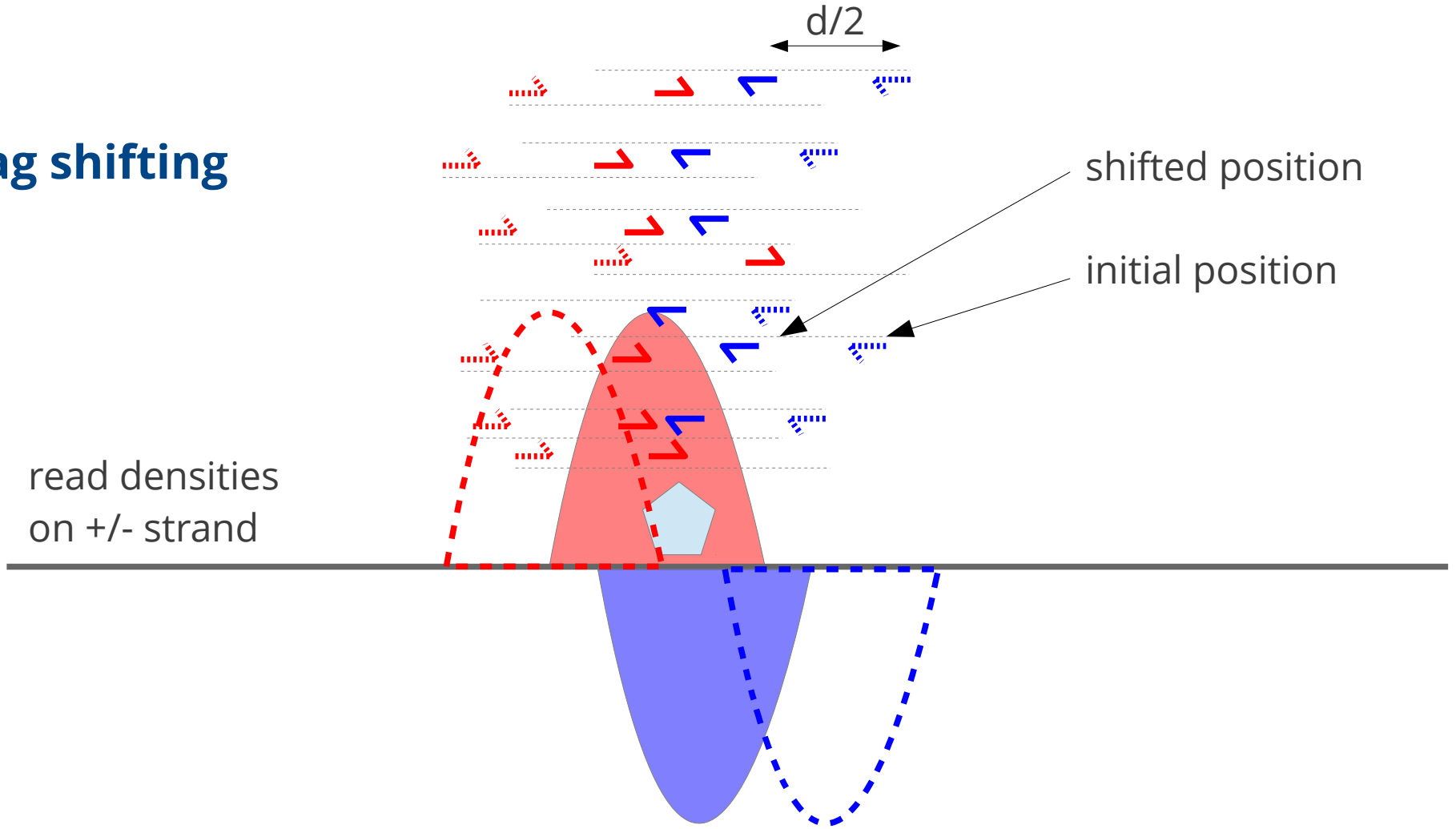
- **Tag shifting vs. extension**
  - positive/negative strand read peaks do not represent the true location of the binding site
  - reads can be **shifted** by *d/2* where *d* is the band size (MACS)
    → increased resolution
  - reads can be **elongated** to
    a size of *d* (FindPeaks, PeakSeq,...)
  - *d* can be estimate from the data (MACS)
    or given as input parameter



example of MACS model building
using top enriched regions

From aligned reads to binding sites

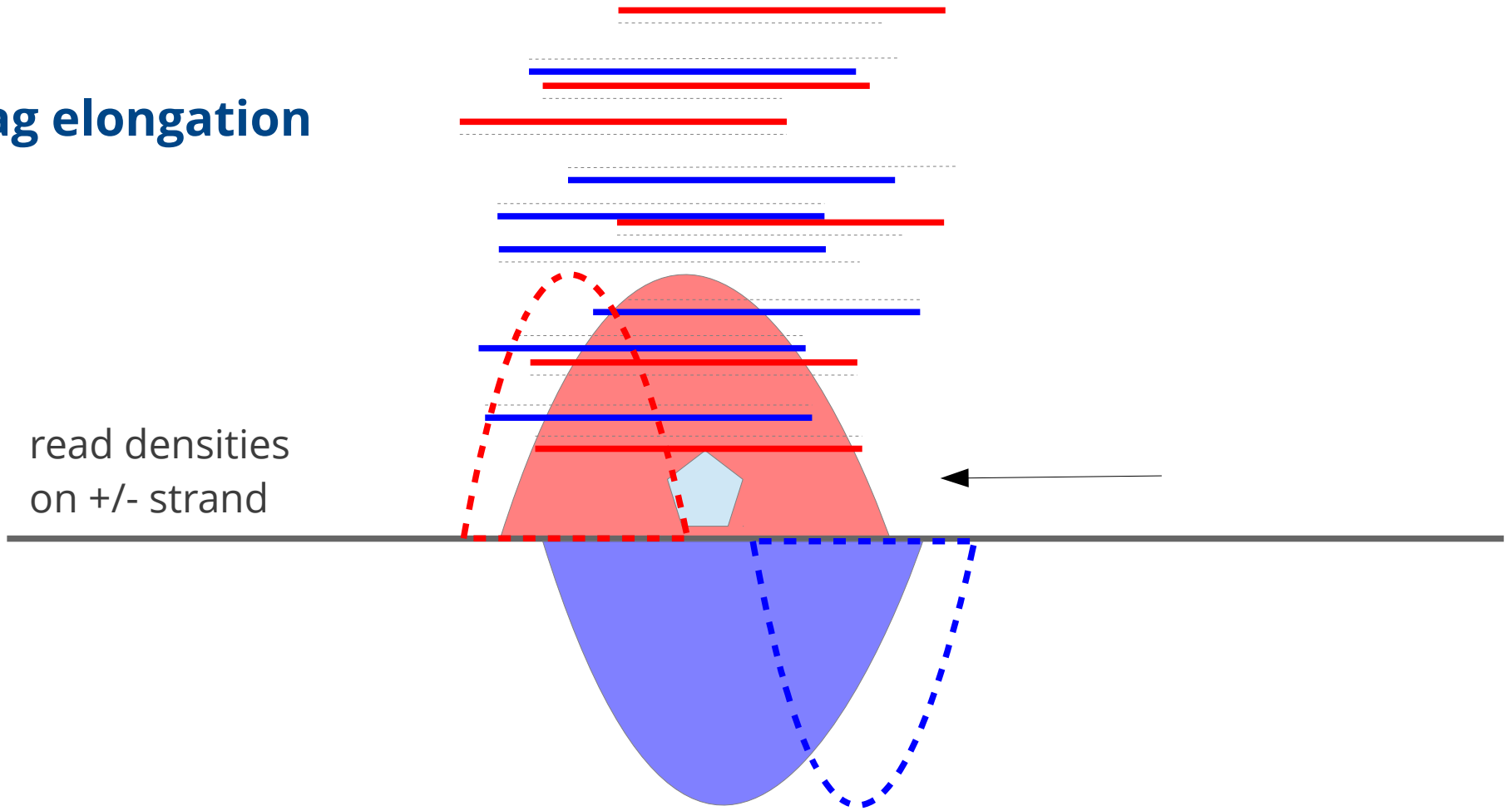
Tag shifting



Each tag is shifted by *d/2* (i.e. towards the middle of the IP fragment) where *d* represent the fragment length

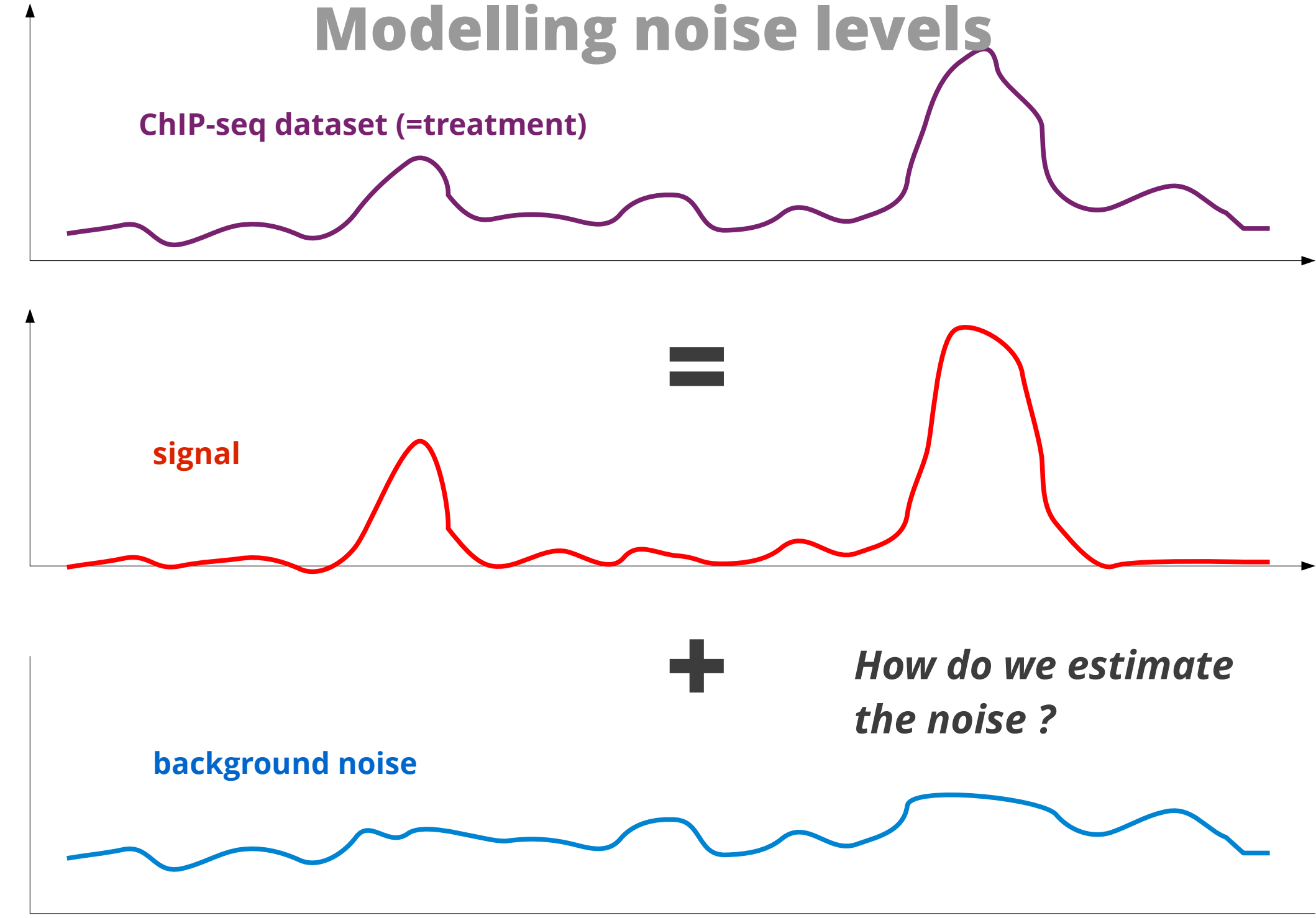
From aligned reads to binding sites

Tag elongation



Each tag is computationaly extended in 3' to a total length of *d*

Modelling noise levels



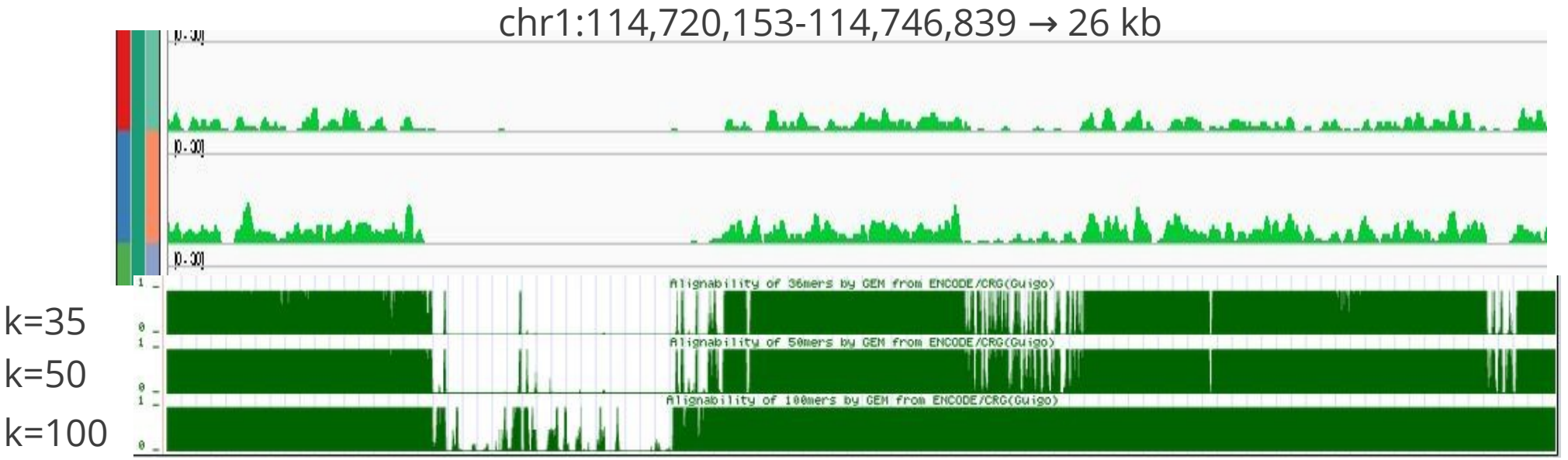
Modelling noise levels

- noise is **not uniform** (chromatin conformation, local biases, mappability)
- input dataset is **mandatory** for reliable local estimation ! (although some algorithms do not require it ... :-( )

chr1:114,720,153-114,746,839 → 26 kb

Modelling noise levels

- the mappability is related to the uniqueness of the k-mers at a particular position of the genome
  - repetitive regions → low uniqueness → low mappability



Longer reads → more uniquely mapped reads

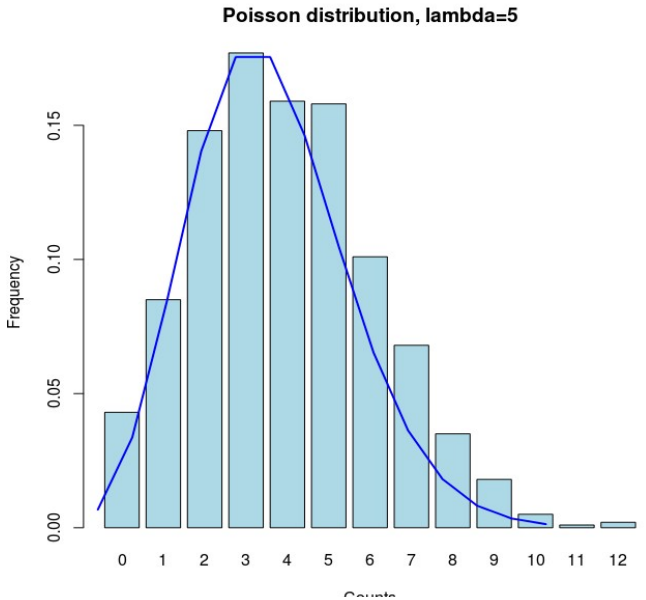
Modelling noise levels

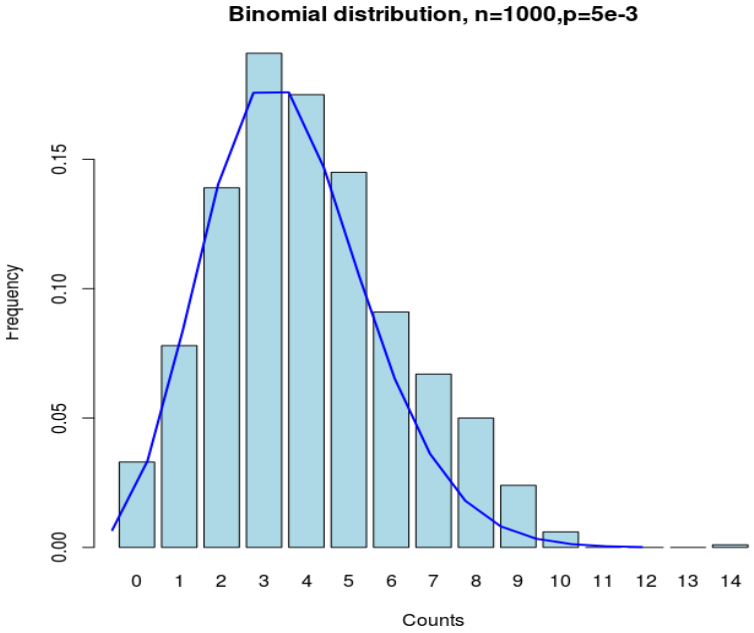
- random distribution of reads in a window of size w modelled using a theoretical distribution
  - **Poisson** distribution
    1 parameter :
    - λ = expected number of reads in window

$$P(X=k)=e^{-k}\frac{\lambda^{k}}{k!}$$

  - **Binomial** distribution
    2 parameters:
    - *p* = probability to start a read at a particular position
    - *n* = number of positions in the window ~ window size
      (assumes no duplicates !)
    - *np* = expected number of reads in window

$$P(X=k)=C_n^k p^k (1-p)^{n-k}$$

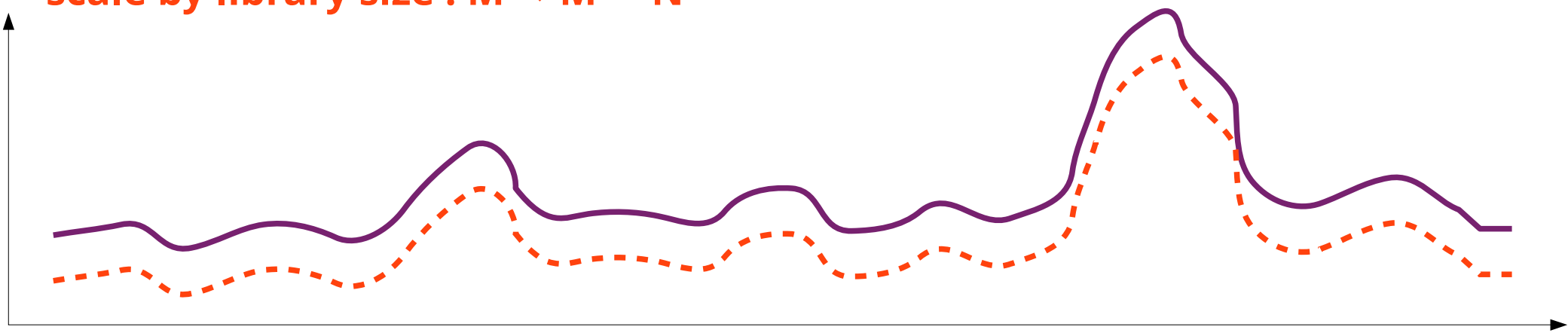




Scaling unequal datasets

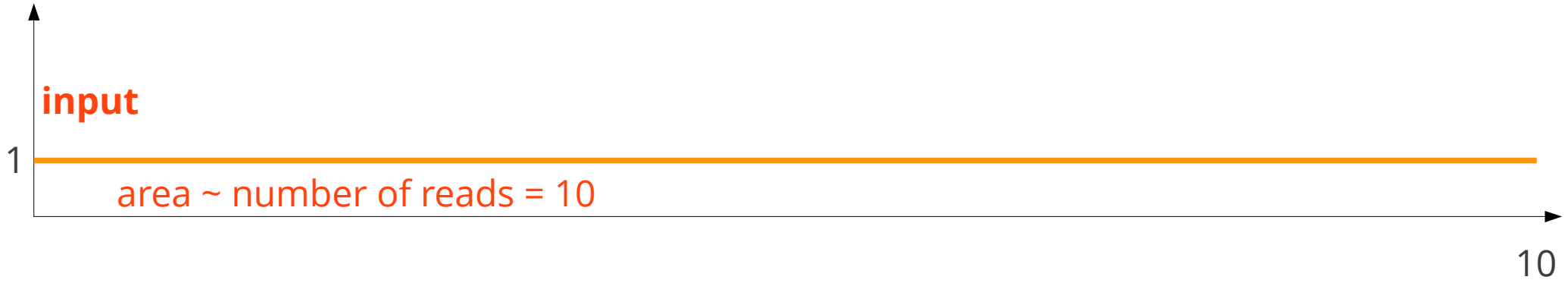
- treatment (=signal + noise) and input (=noise) datasets generally do not have the same sequencing depth → need for normalization
- input dataset should model the noise level in the treatment dataset
- **naïve approach** : upscale/downscale the smaller/larger dataset

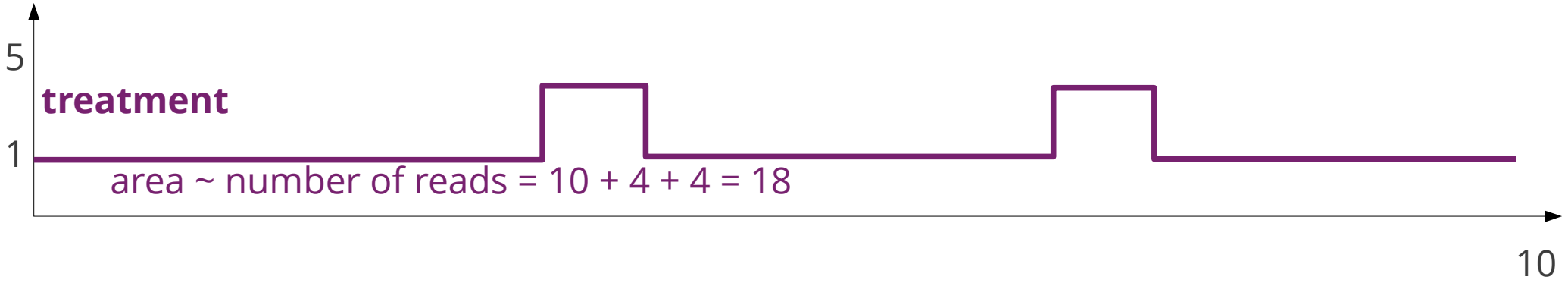
Input : N reads
ChIP-seq dataset → M > N reads
scale by library size : M → M' = N



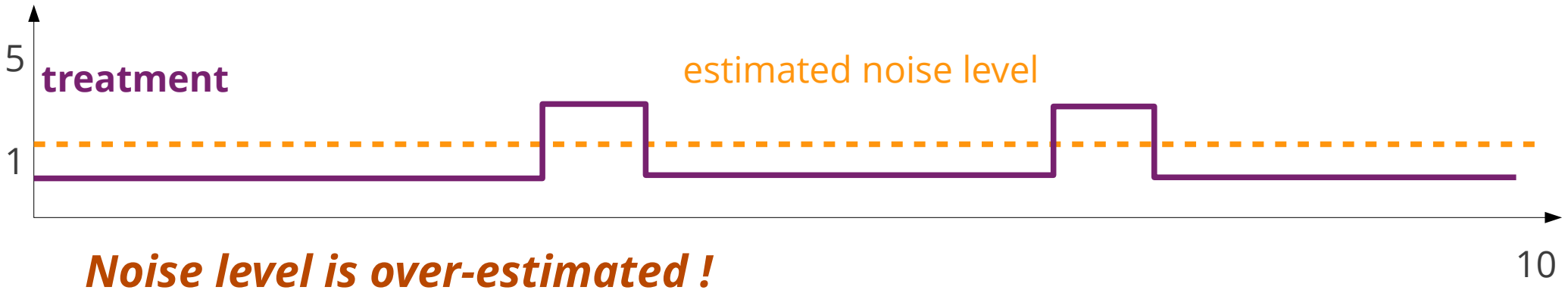
Problem : signal influences scaling factor
More signal (but equal noise) → artificial noise over-estimation

Scaling unequal datasets



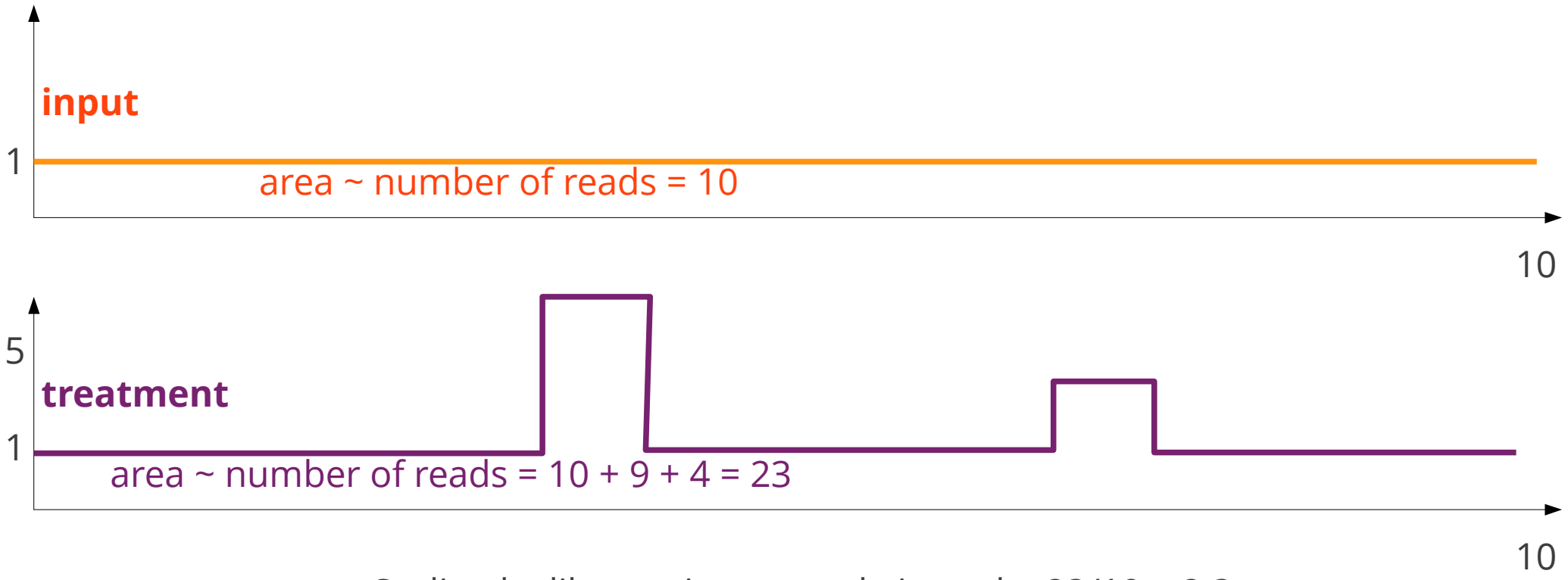


Scaling by library size : upscale input by 18/10 = 1.8



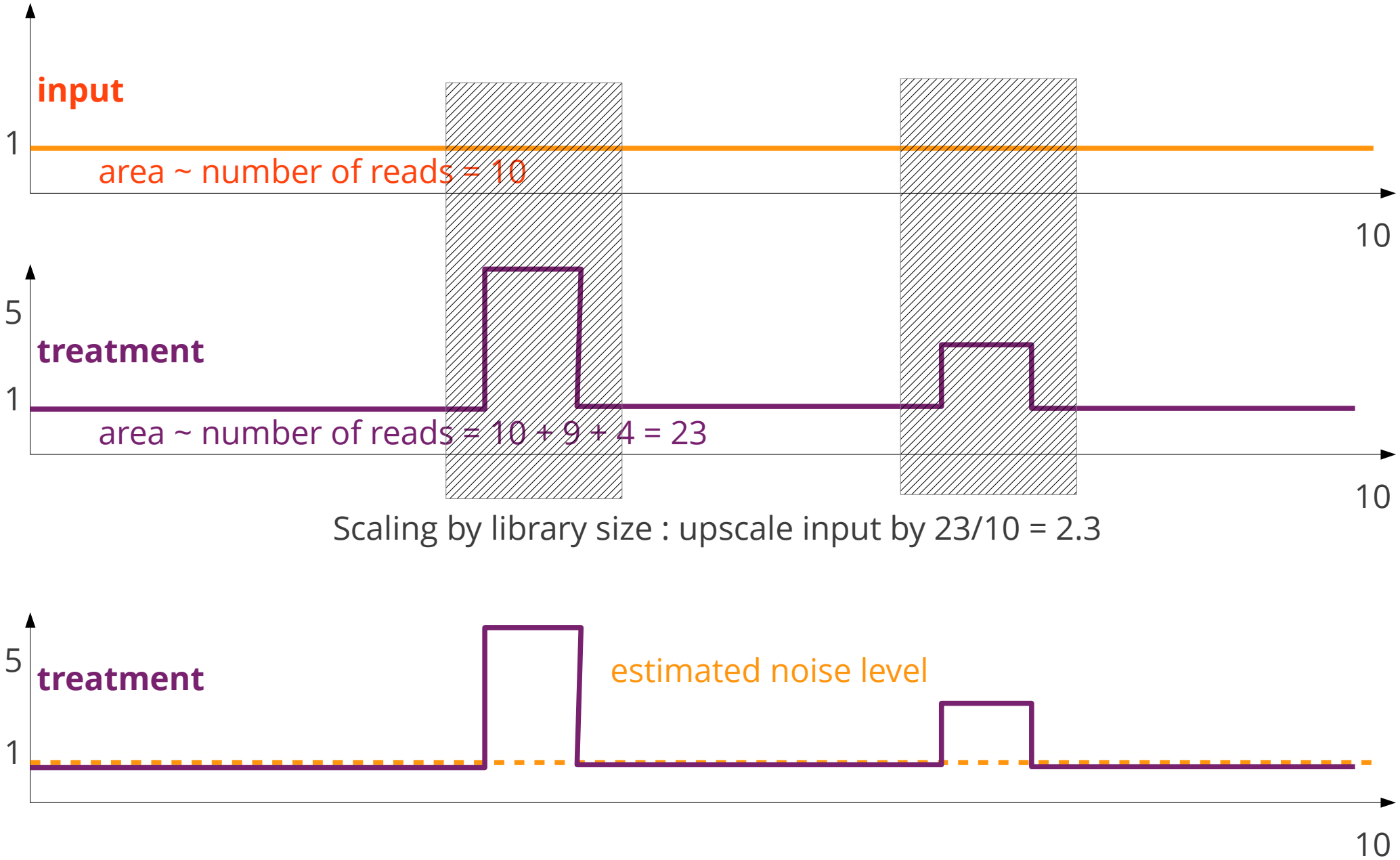
Noise level is over-estimated !

Scaling unequal datasets



Scaling by library size : upscale input by 23/10 = 2.3

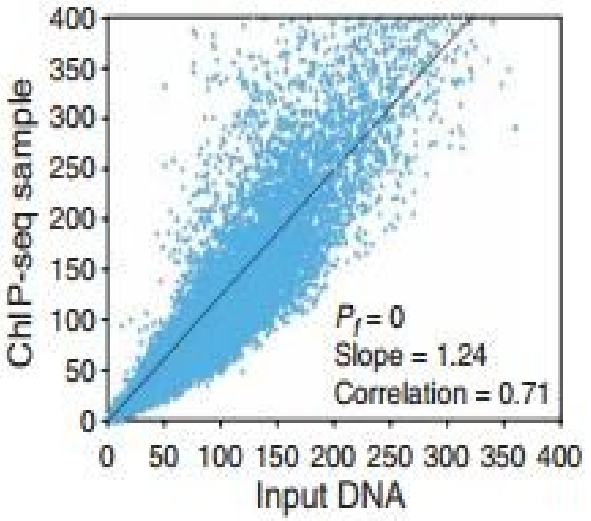
Scaling unequal datasets



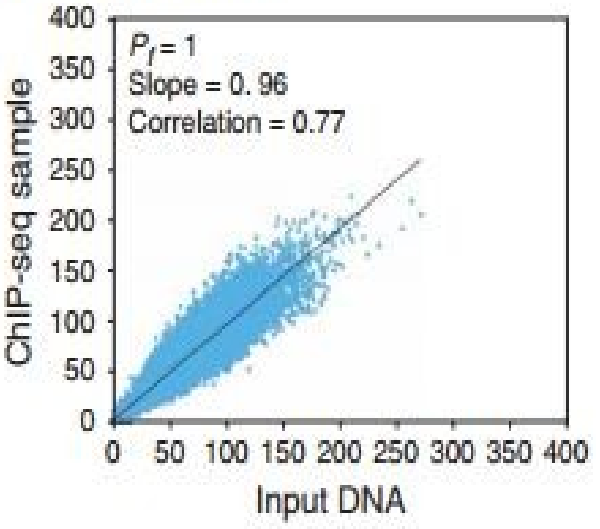
Scaling by library size : upscale input by 23/10 = 2.3

Scaling unequal datasets

- **more advanced** : linear regression by exclusing peak regions (PeakSeq)
- read counts in 1Mb regions in input and treatment



all regions



excluding enriched (=signal) regions

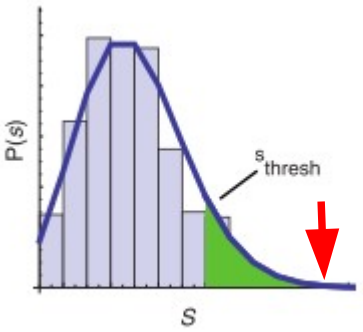
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

Joel Rozowsky[1], Ghia Euskirchen[2], Raymond K Auerbach[3], Zhengdong D Zhang[1], Theodore Gibson[1], Robert Bjornson[4], Nicholas Carriero[4], Michael Snyder[1,2] & Mark B Gerstein[1,3,4]

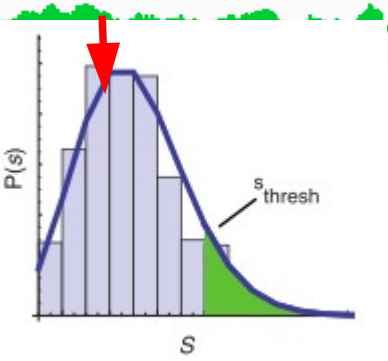
Defining “peaks”

- **Determining “enriched” regions**
  - sliding window across the genome
  - at each location, evaluate the enrichement of the signal wrt. expected background based on the distribution
  - retain regions with P-values below threshold
  - evaluate FDR

Pval < 1e-20



Pval ~ 0.6



	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based duplicate[e]
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
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
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
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
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
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
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
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
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 / Yes
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance					
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation					

Computation for ChIP-seq and RNA-seq studies

Shirley Pepke[1], Barbara Wold[2] & Ali Mortazavi[2]

	Profile
CisGenome v1.1	Strand-specific window scan
ERANGE v3.1	Tag aggregation
FindPeaks v3.1.9.2	Aggregation of overlapped tags
F-Seq v1.82	Kernel density estimation (KDE)
GLITR	Aggregation of overlapped tags
MACS v1.3.5	Tags shifted then window scan
PeakSeq	Extended tag aggregation
QuEST v2.3	Kernel density estimation
SICER v1.02	Window scan with gaps allowed
SiSSRs v1.4	Window scan
spp v1.0	Strand specific window scan

Some methods separate the tag densities into different strands and take advantage of tag asymmetry

Most consider merged densities and look for enrichment

	Profile	Peak criteria[a]	Tag shift
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
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
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation

Tag shift

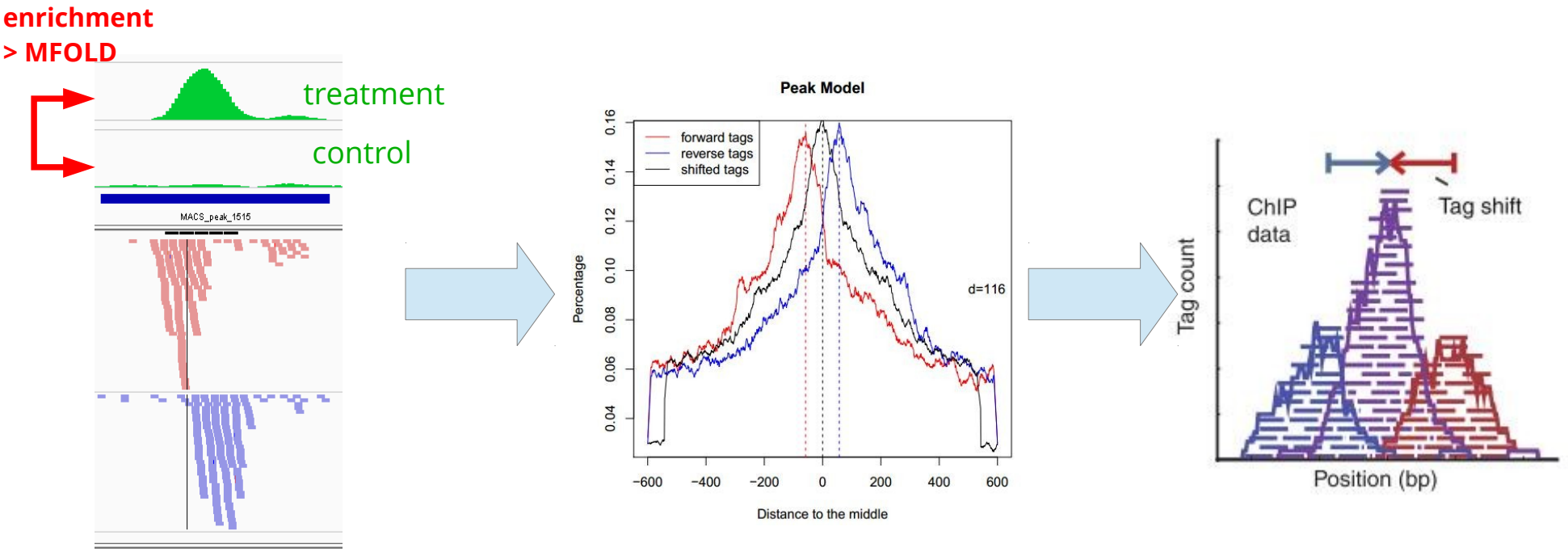
Tag extension

Tags unchanged

MACS

[Zhang et al. Genome Biol. 2008]

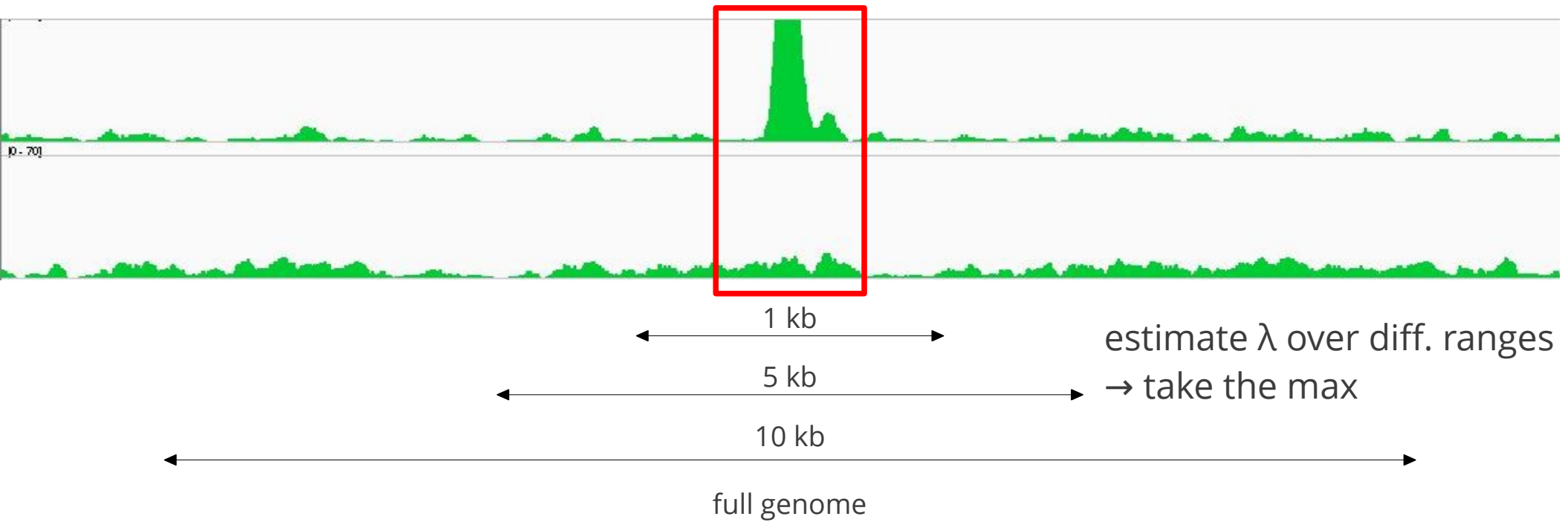
- **Step 1 : estimating fragment length *d***
  - slide a window of size BANDWIDTH
  - retain top regions with MFOLD enrichment of treatment vs. input
  - plot average +/- strand read densities → estimate d



MACS

[Zhang et al. Genome Biol. 2008]

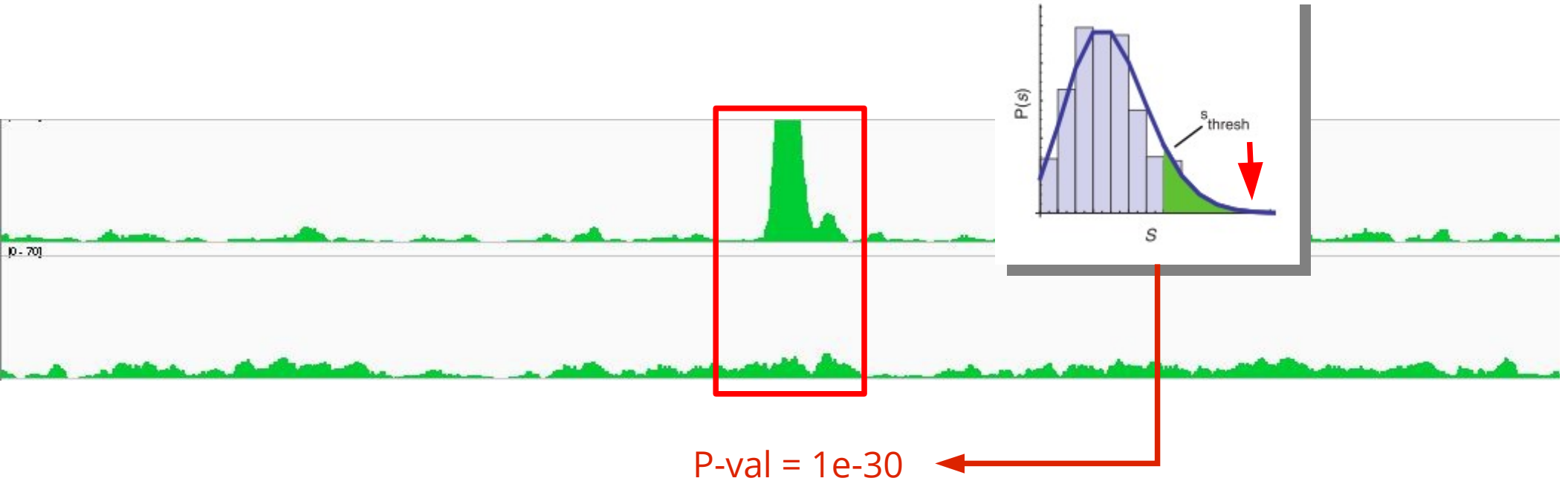
- **Step 2 : identification of local noise parameter**
  - slide a window of size *2*d* across treatment and input
  - estimate parameter λ_{local} of Poisson distribution



MACS

[Zhang et al. Genome Biol. 2008]

- **Step 3 : identification of enriched/peak regions**
  - determine regions with P-values < PVALUE
  - determine summit position inside enriched regions as max density

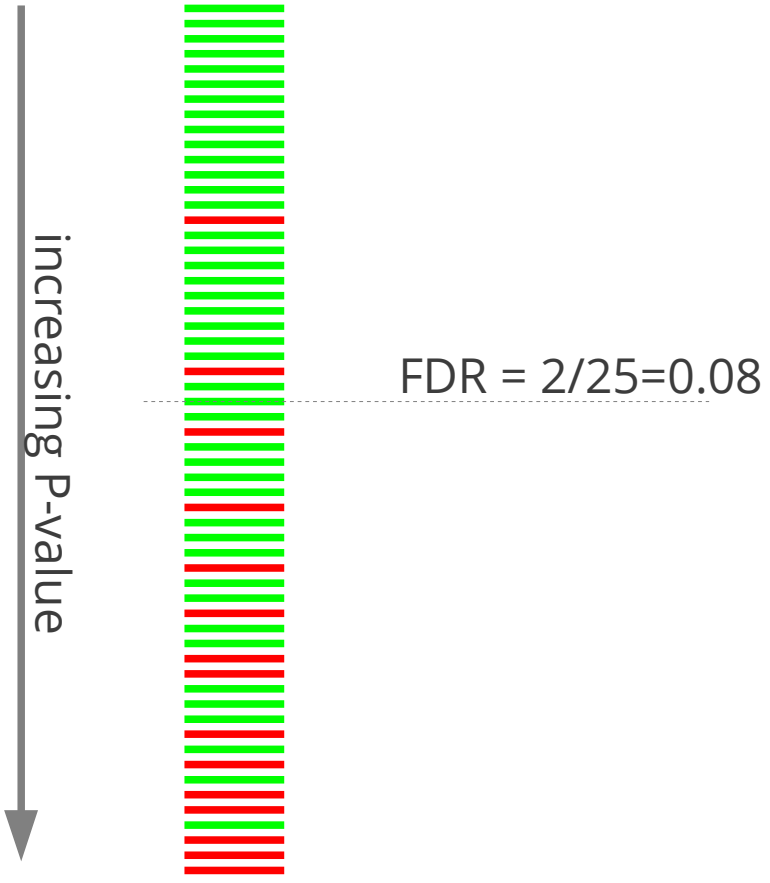


MACS

[Zhang et al. Genome Biol. 2008]

- **Step 4 : estimating FDR**
  - positive peaks (P-values)
  - swap treatment and input; call negative peaks (P-value)

$$FDR(p) = \frac{\#\ \textbf{negative}\ \text{peaks with Pval} < p}{\#\ \textbf{positive}\ \text{peaks with Pval} < p}$$



Program of the Practical Session

Step 0 : Find datasets on Gene Expression Omnibus

Step 1 : retrieving data from Galaxy

Step 2 : data inspection

Step 3 : peak calling using MACS

Step 4 : splitting peaks with PeakSplitter

Step 5 : comparing BED files

Step 6 : visualizing results in IGV

Additional exercices