ChIP-seq analysis

J. van Helden, M. Defrance, C. Herrmann, D. Puthier, N. Servant, M. Thomas-Chollier, O.Sand

- Tuesday :
  - quick introduction to ChIP-seq and peak-calling (Presentation + Practical session)
  - Functional annotation of peaks
  - Motif analysis of ChIP-seq peaks
  - Peak quality assessment
  - Interpretation of ChIP-seq/ChIP-exo data

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

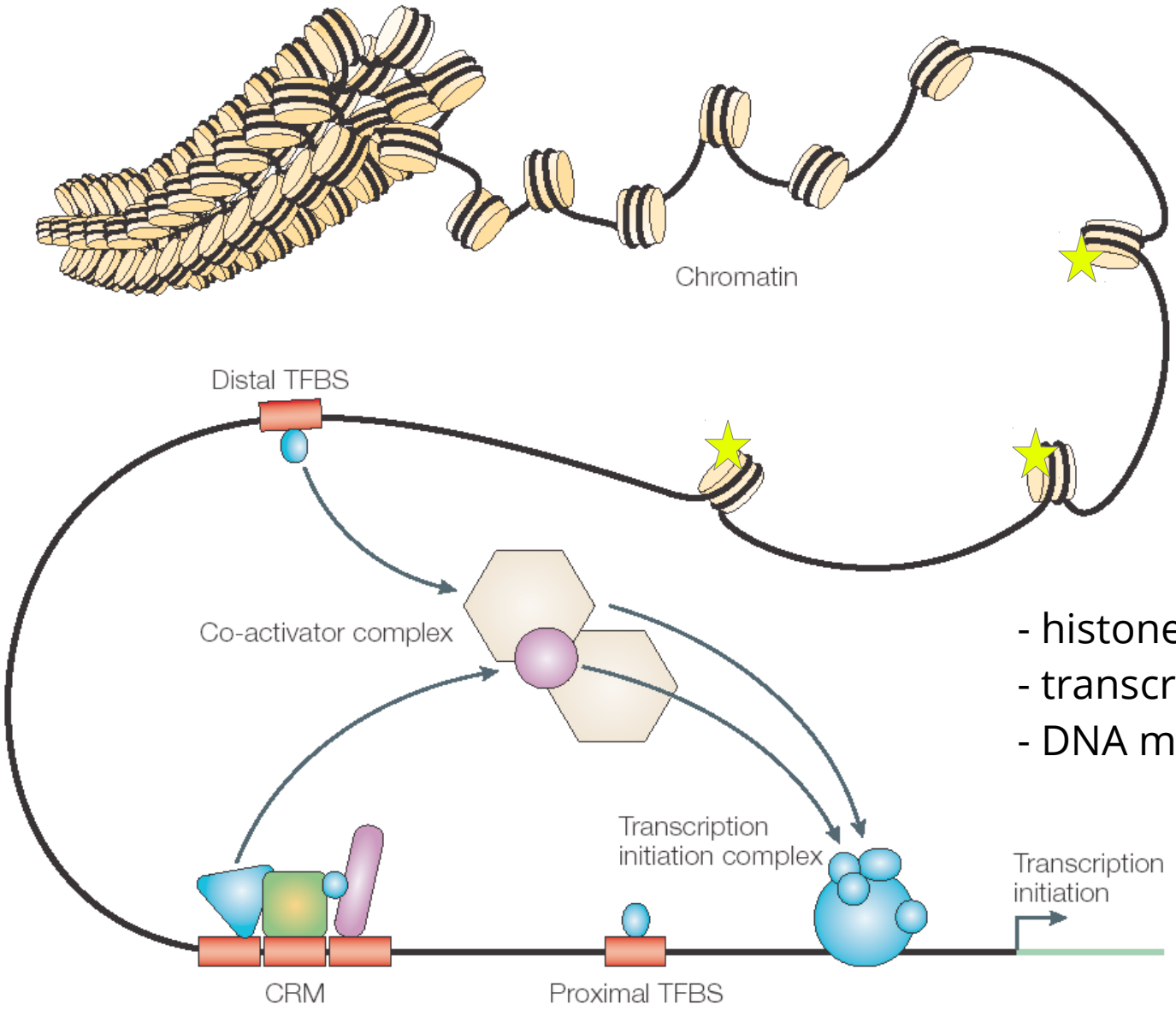
- estrogen-receptor (ESR1) is a key factor in breast cancer developement
- goal of the study: understand the dependency of ESR1 binding on presence of co-factors, in particular GATA3, which is mutated in breast cancers
- approaches: GATA3 silencing (siRNA), ChIP-seq on ESR1 in wt vs. siGATA3 conditions,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

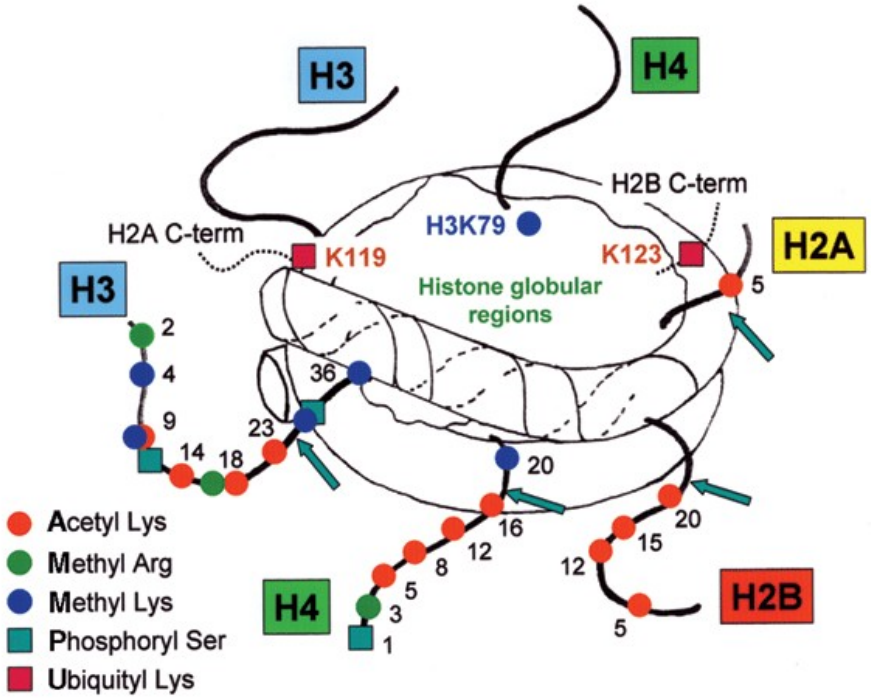
Chromatin – more than just sequence



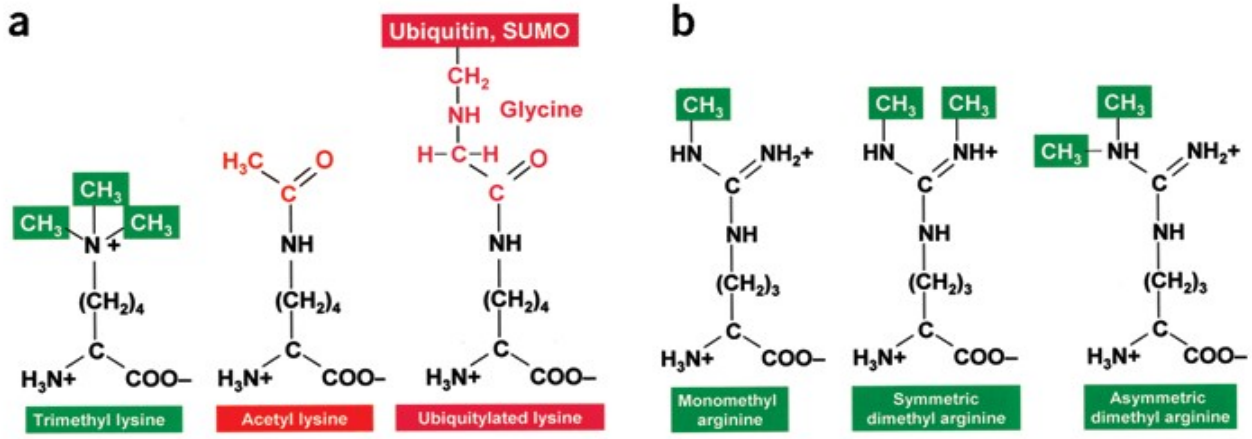
- histone modifications
- transcription factor binding
- DNA methylation

[Wasserman & Sandelin, Nat.Rev.Gen (2004)]

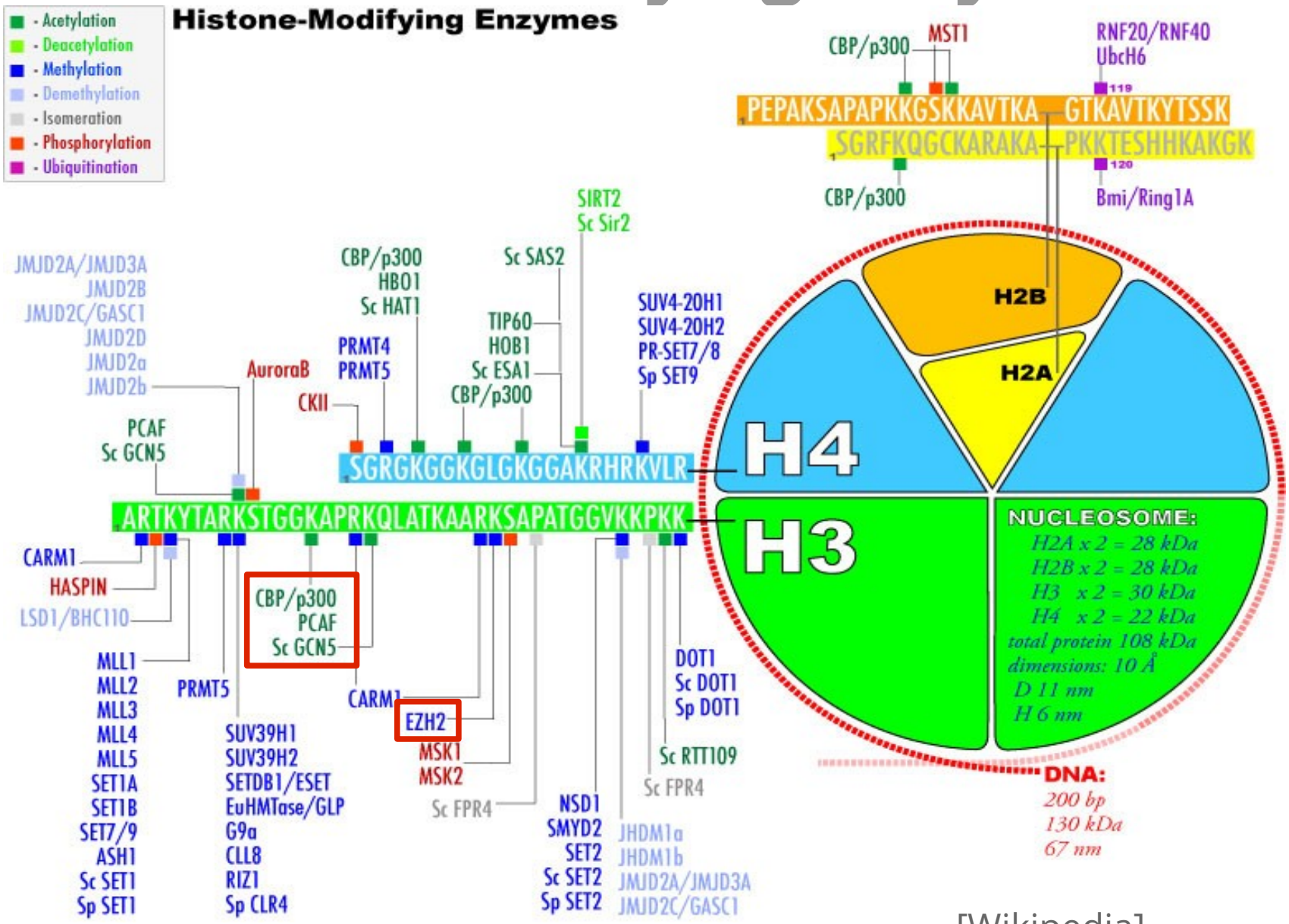
Histone modifications



- histones are subject to post-translational modifications at their N-terminal tail
  - Lysine methylation
  - Lysine/arginine acethylation
  - Serine phosphorylation
  - ubiquitylation

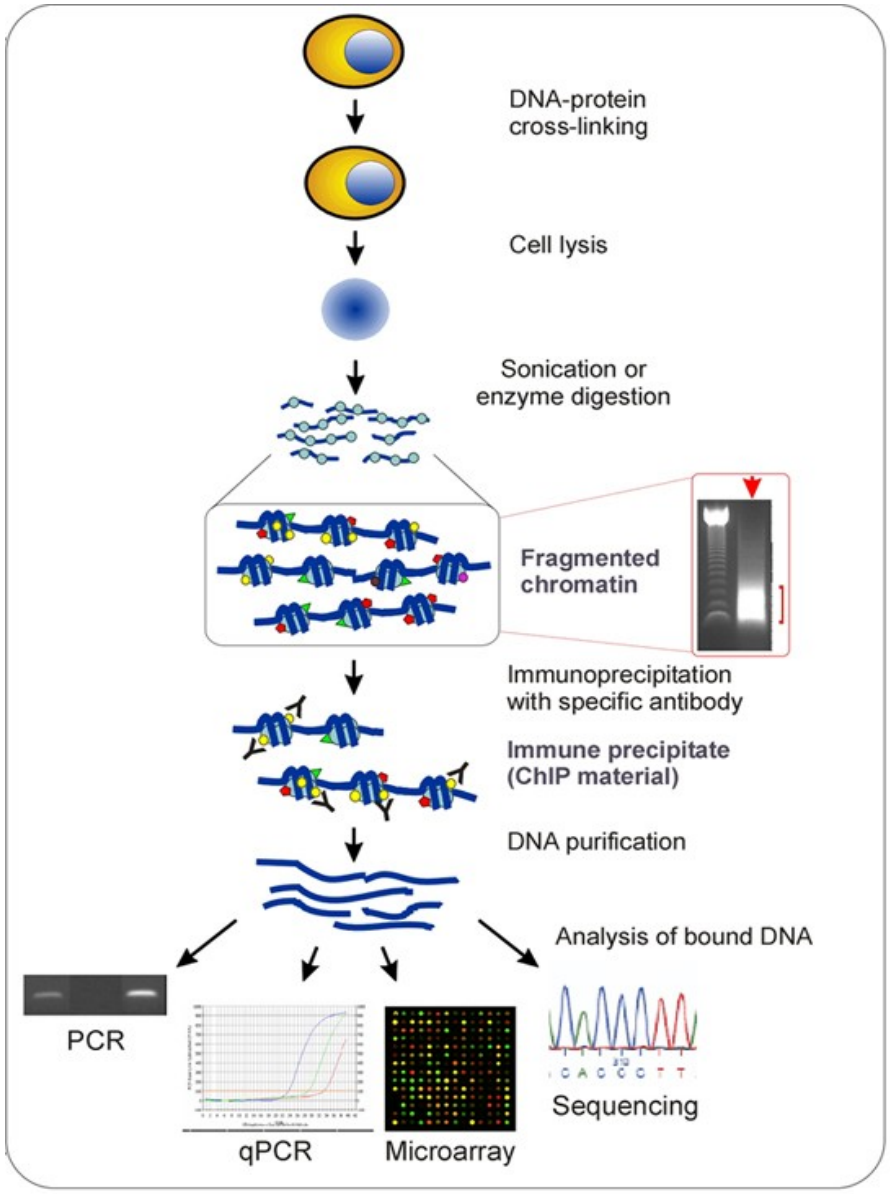


Chromatin binding proteins
Histone modifying enzymes



[Wikipedia]

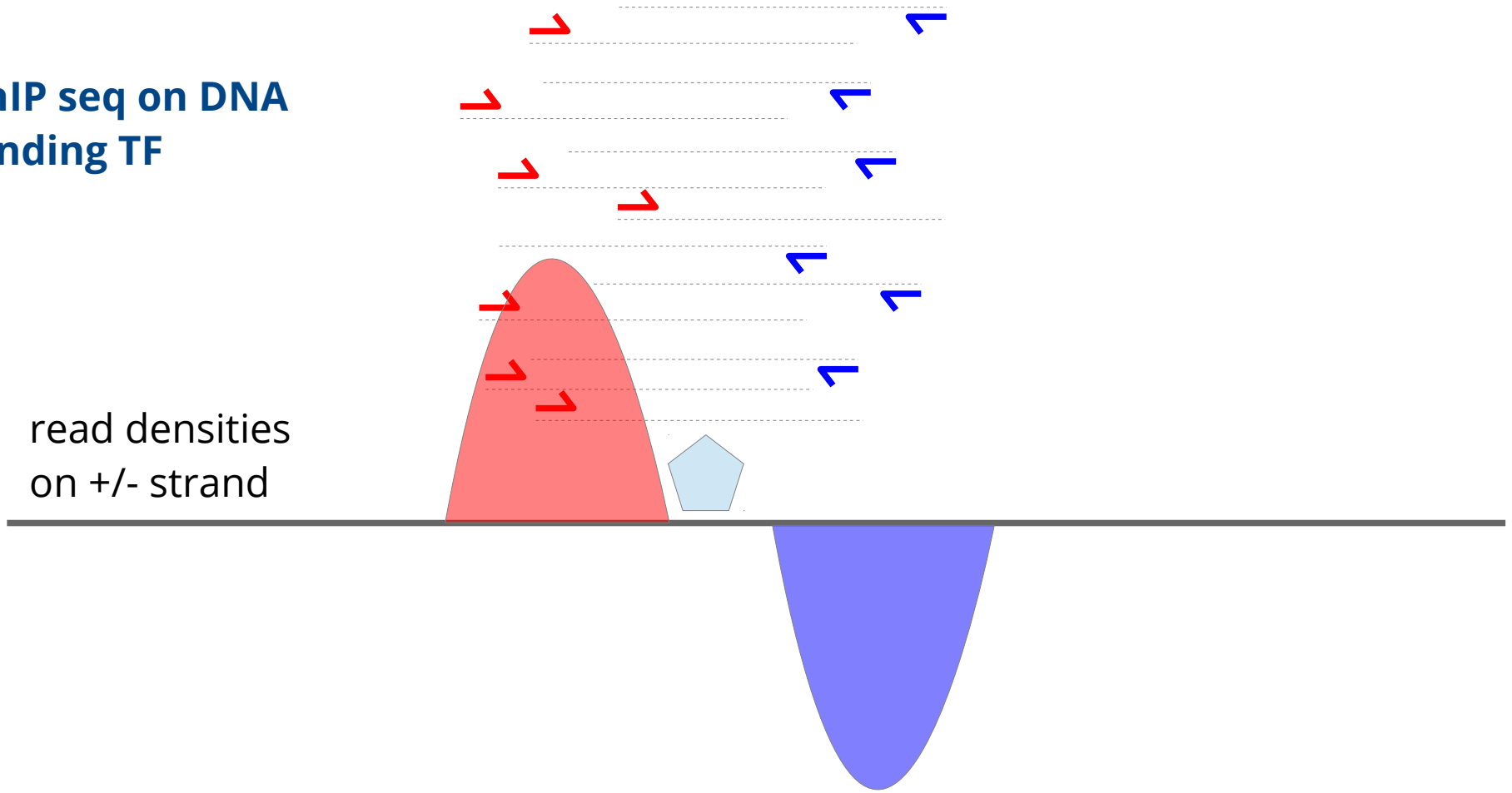
Experimental identification of binding sites



- Chromatin immunoprecipitation (ChIP) followed by
  - sequencing (ChIP-seq)
  - hybridization on array (mostly tilling arrays) ChIP-chip
  - PCR/qPCR
- main challenge : quality/specificity of the antibodies

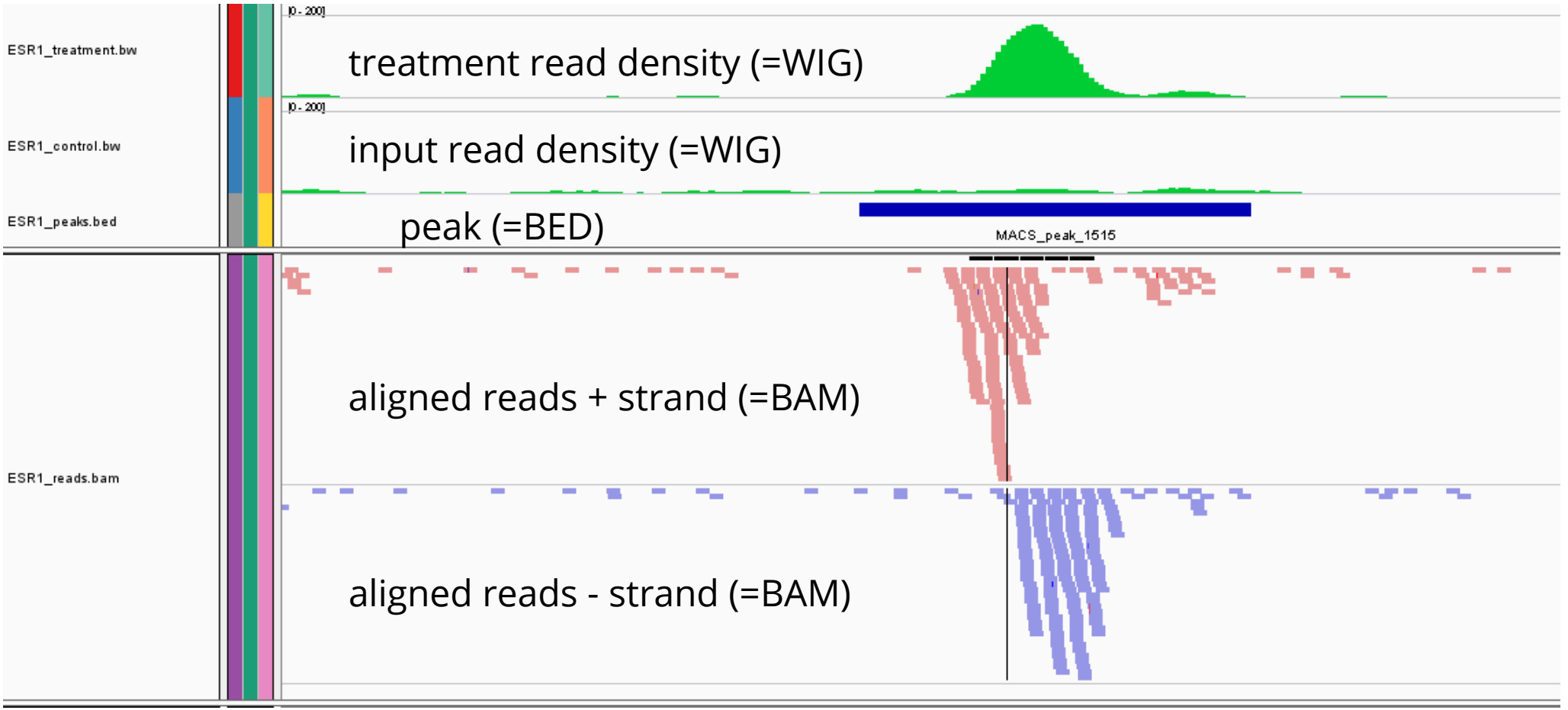
ChIP-seq signal for transcription factors

ChIP seq on DNA binding TF



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

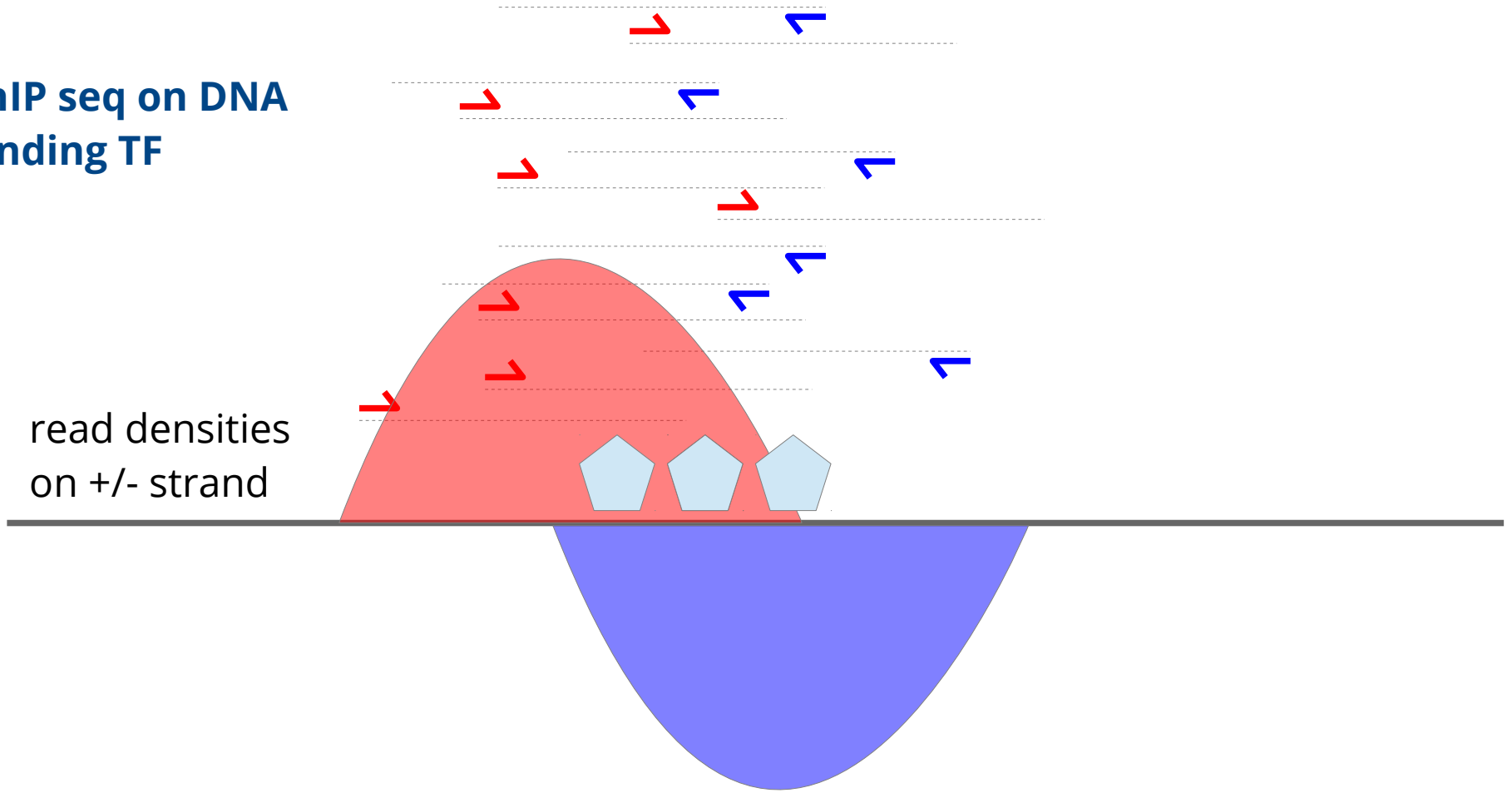
ChIP-seq signal for transcription factors



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

ChIP-seq signal for transcription factors

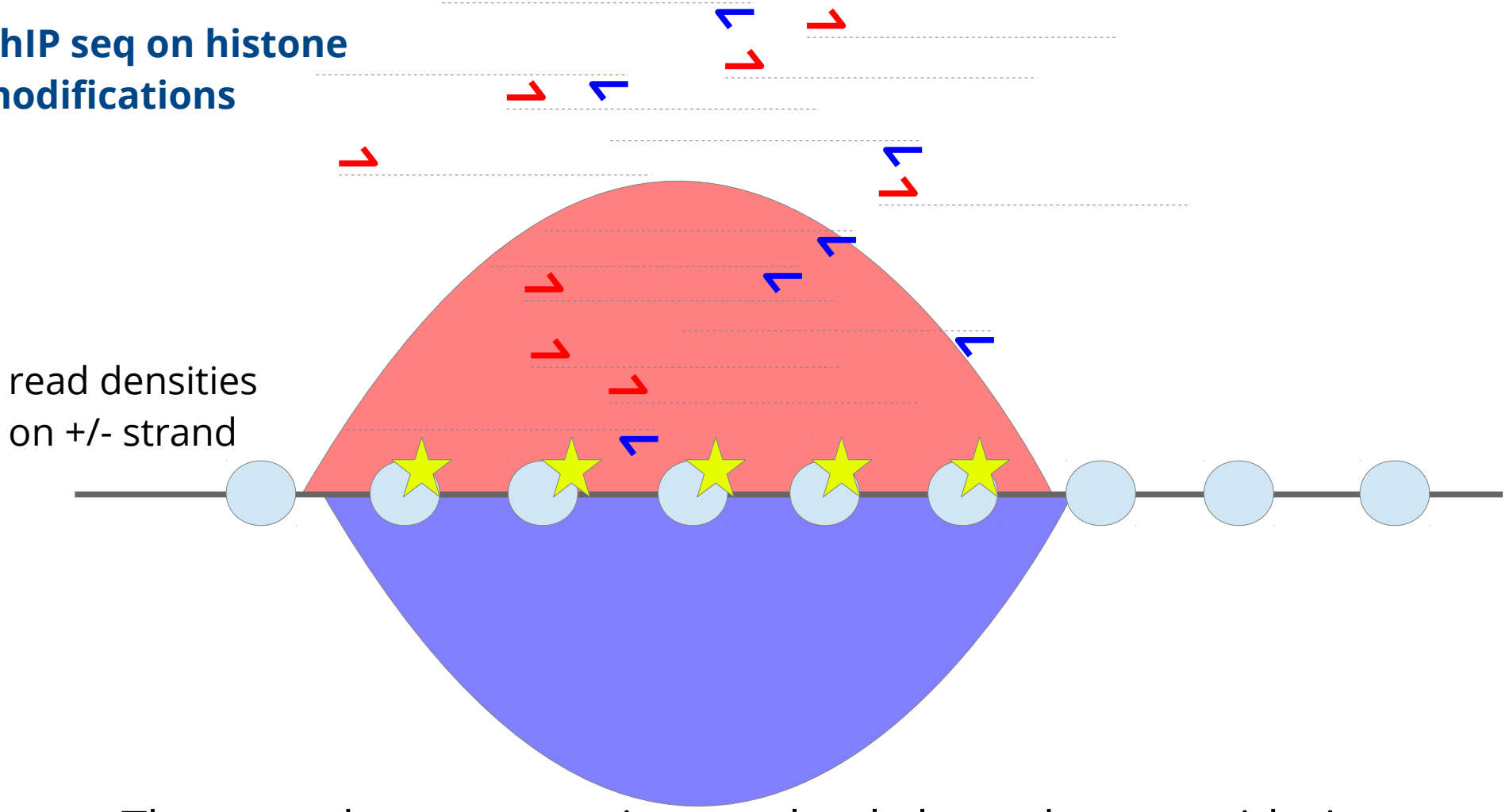
ChIP seq on DNA binding TF



Binding of several TF as complexes tend to blur this asymmetry

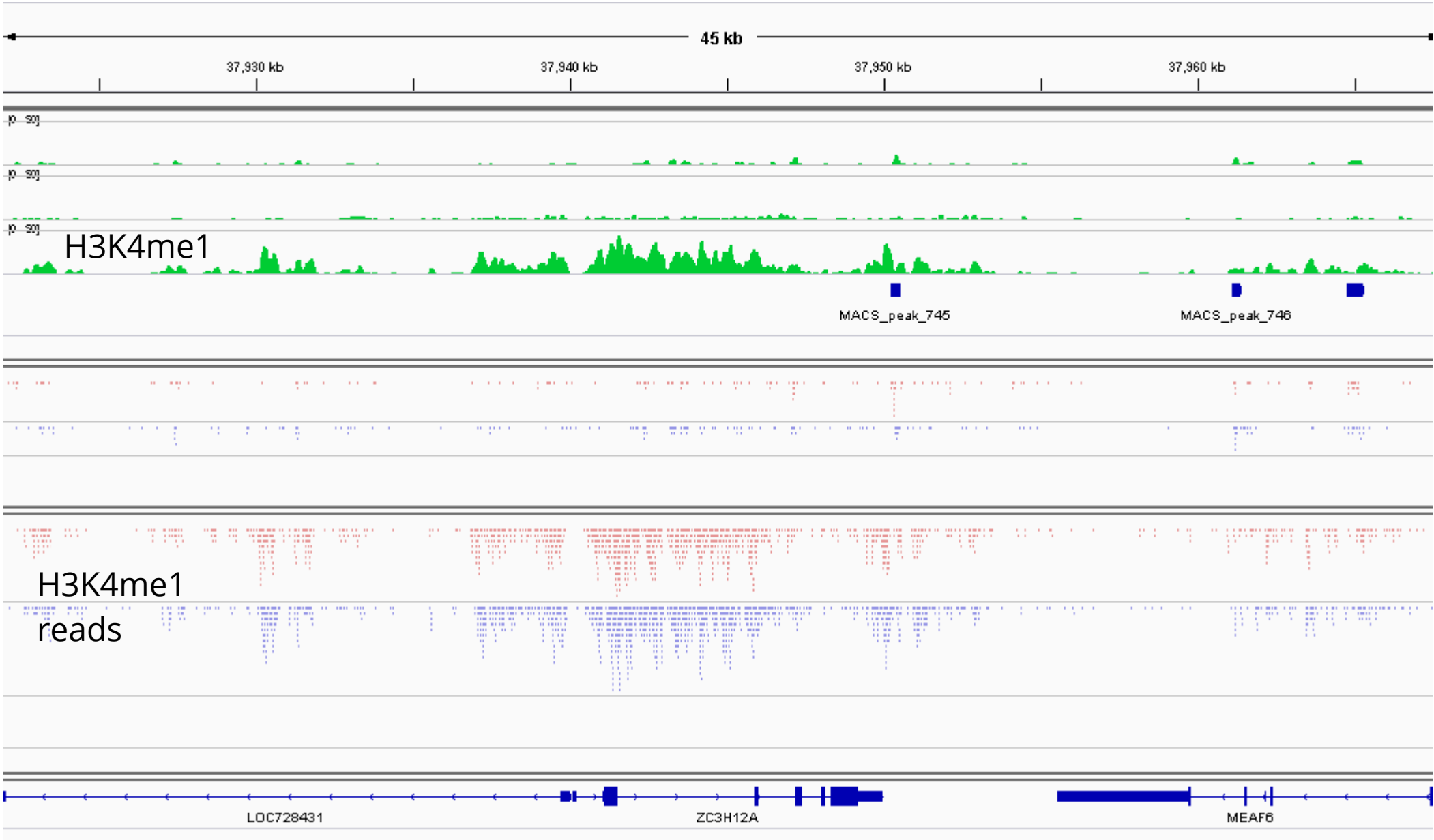
ChIP-seq signal for histone marks

ChIP seq on histone modifications

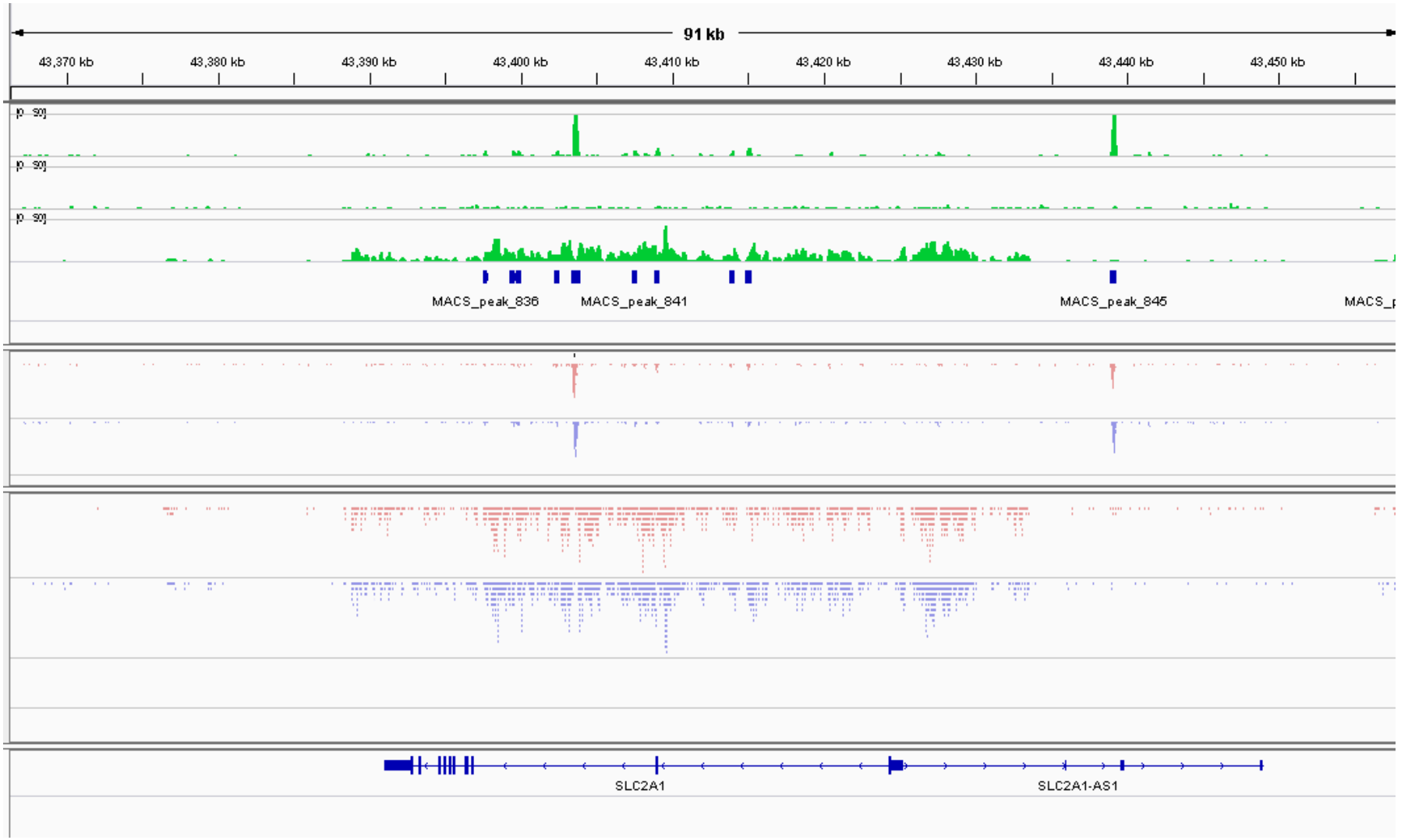


The strand asymmetry is completely lost when considering ChIP datasets for diffuse histone modifications

Real example of ChIP-seq signal



Real example of ChIP-seq signal

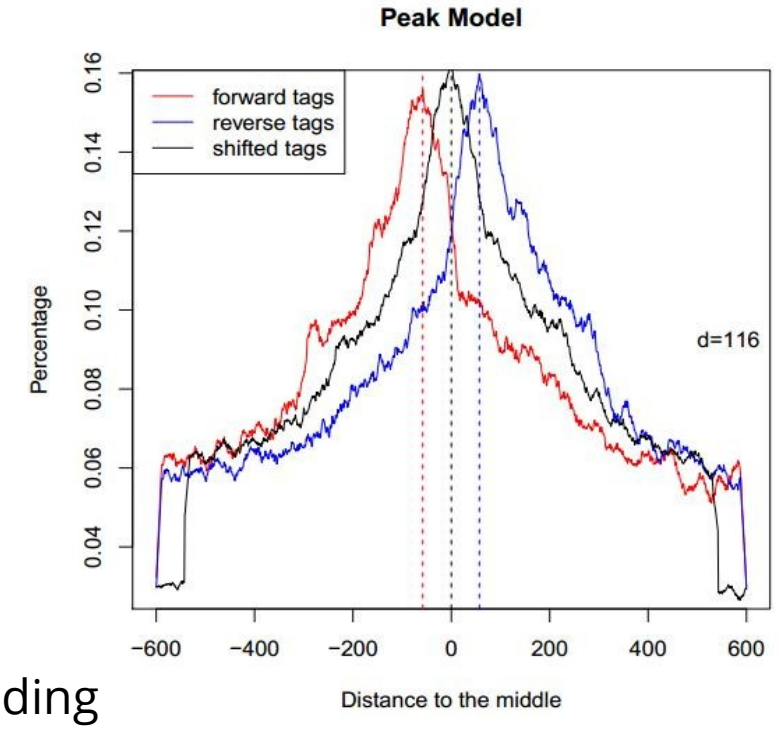


Keys aspects of “peak” finding

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

From aligned reads to binding sites

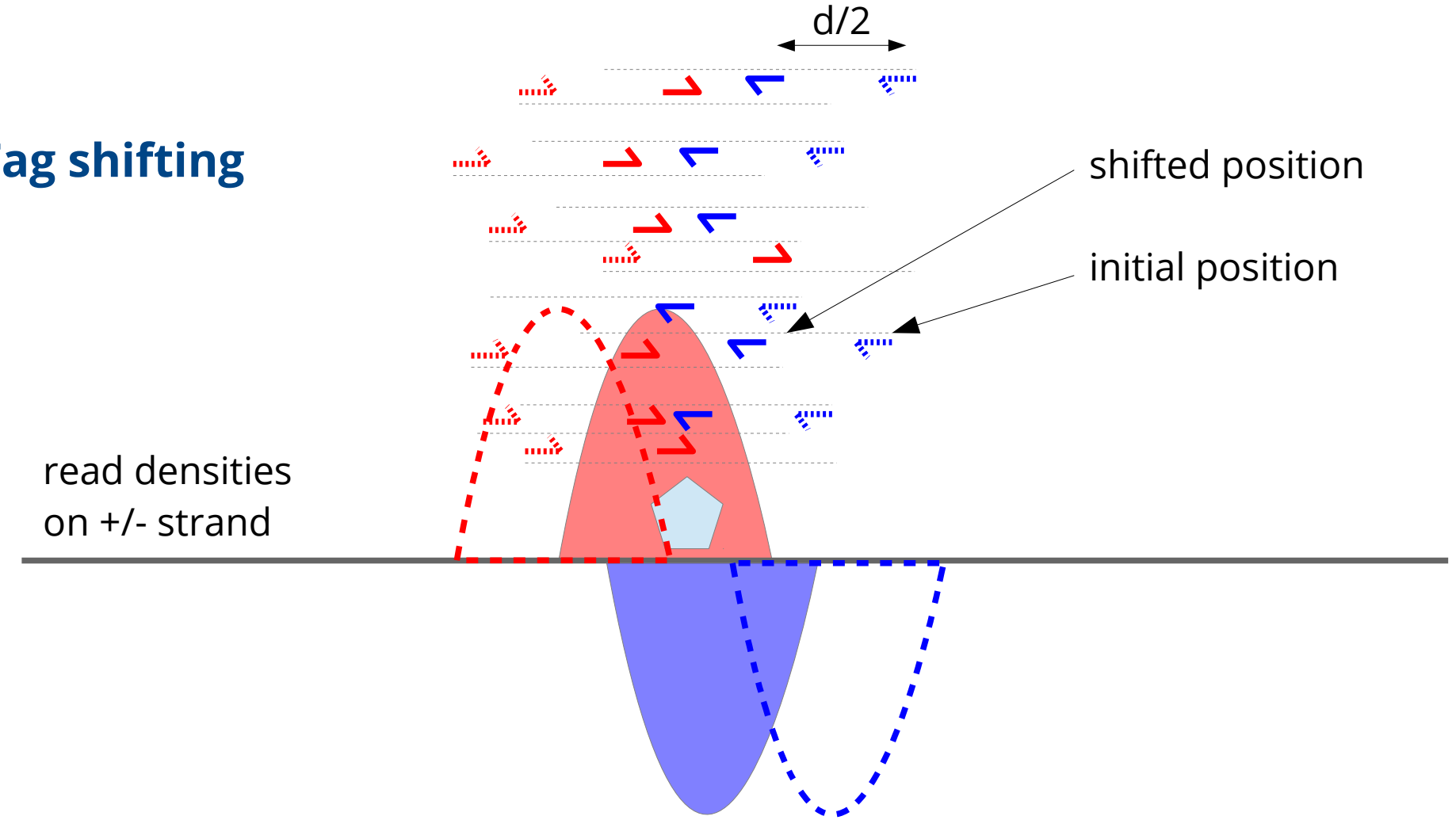
- **Tag shifting vs. extension**
  - positive/negative strand read peaks *do not represent the true location of the binding site*
  - reads can be **shifted** by *d/2* where *d* is the band size (MACS)
    → increased resolution
  - reads can be **elongated** to
    a size of *d* (FindPeaks, PeakSeq,...)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

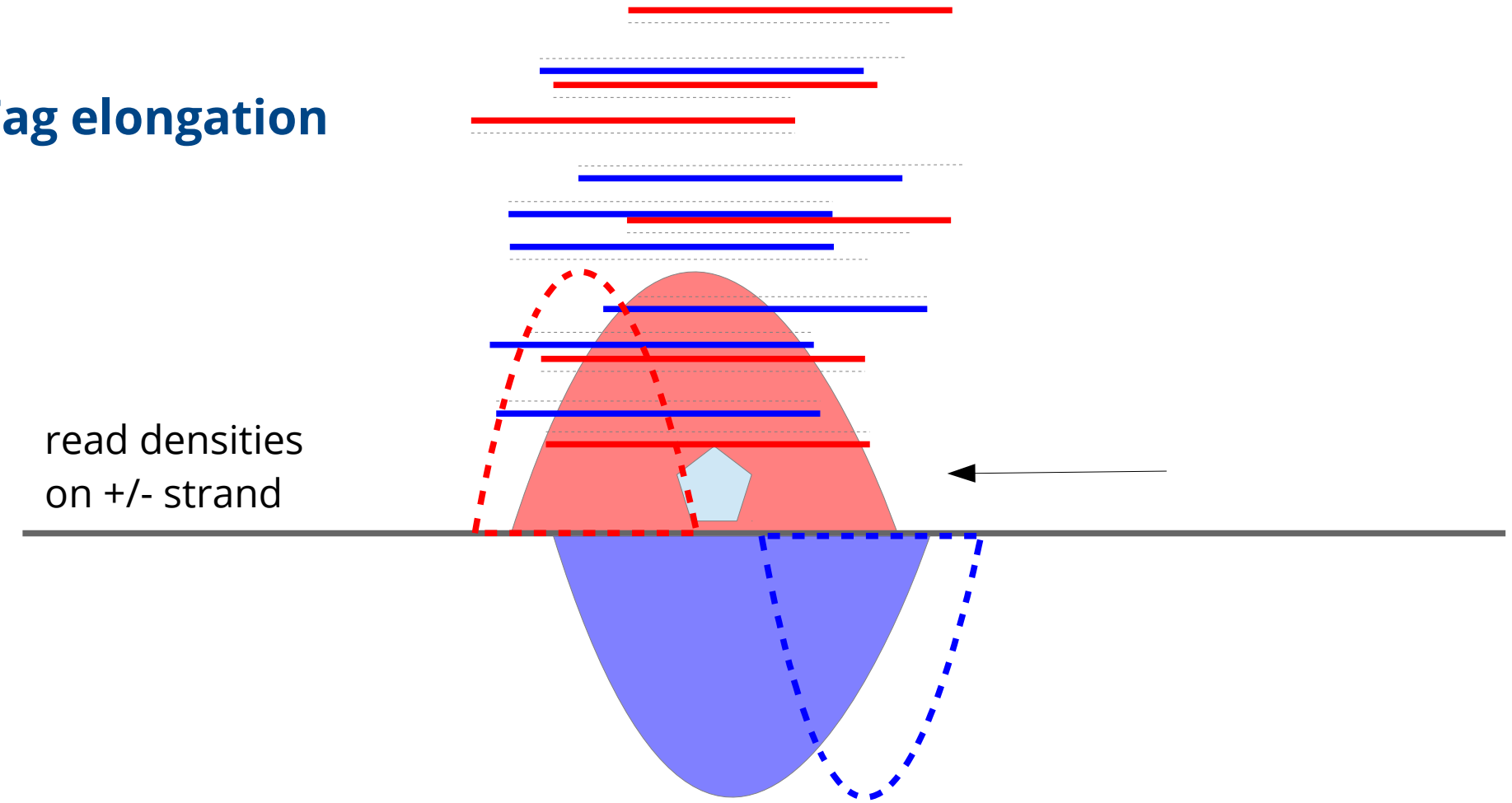
Tag shifting



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

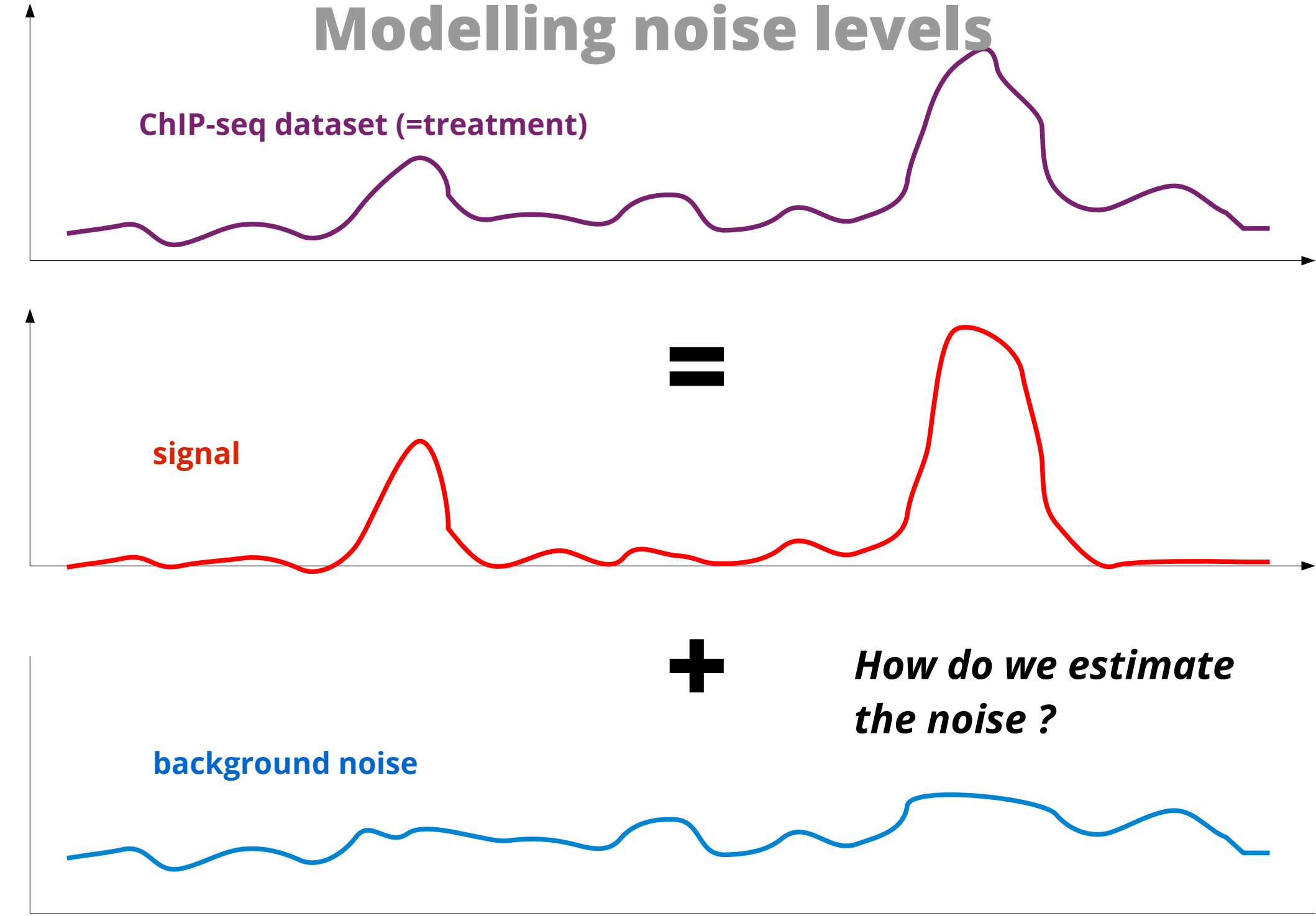
From aligned reads to binding sites

Tag elongation



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

Modelling noise levels



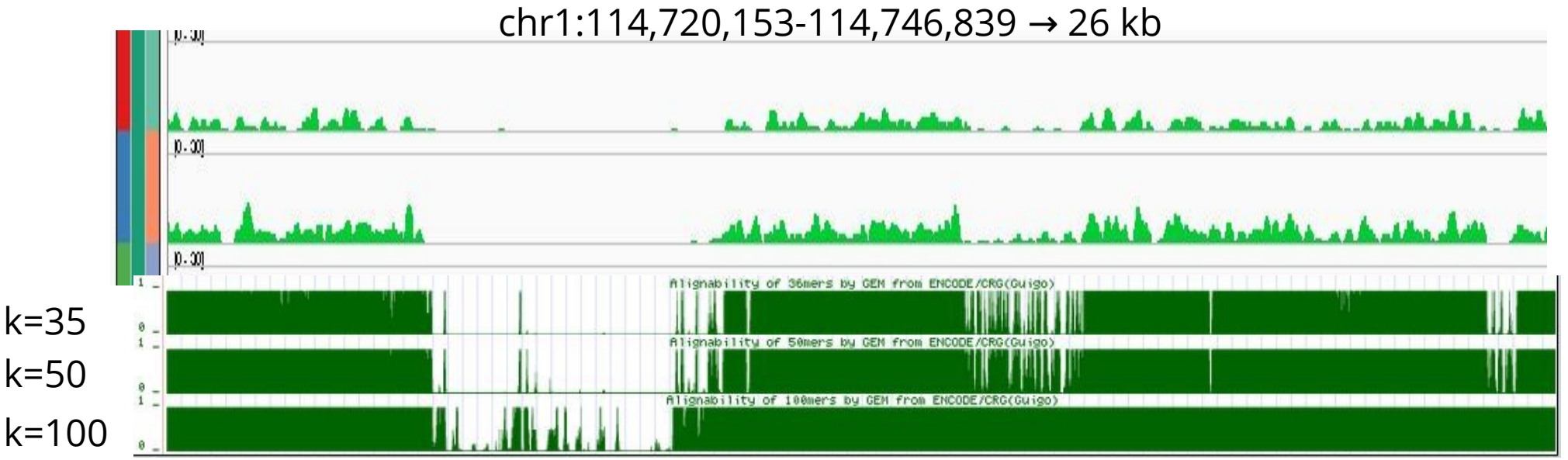
Modelling noise levels

- 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

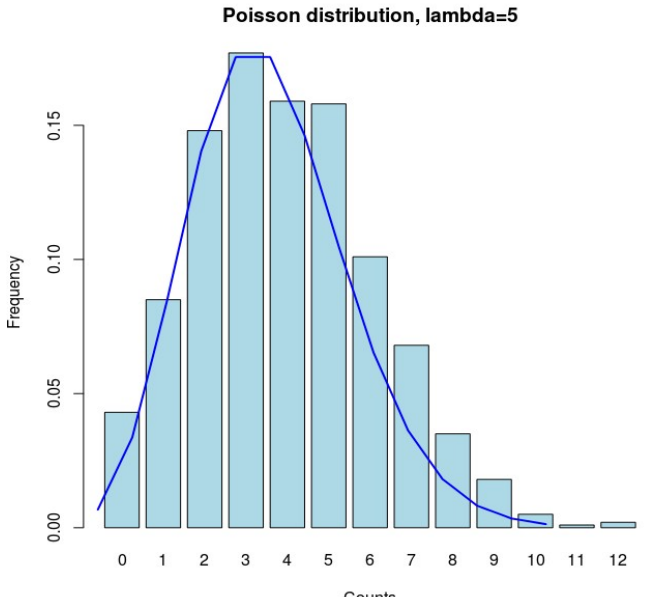
Modelling noise levels

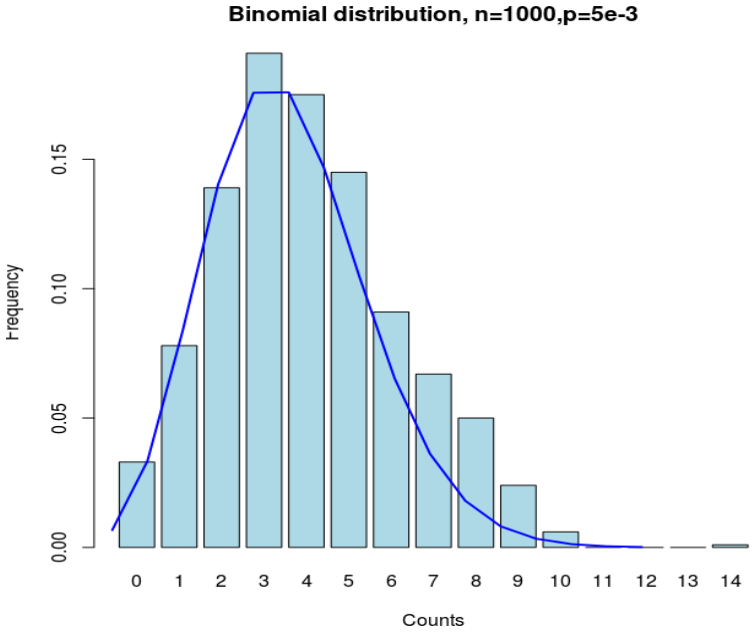
- random distribution of reads in a window of size w modelled using a theoretical distribution
  - **Poisson** distribution
    1 parameter :
    - λ = expected number of reads in window

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

  - **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

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

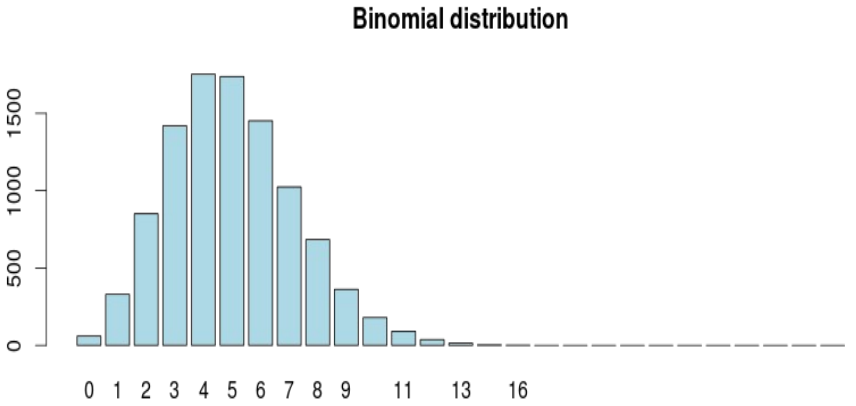


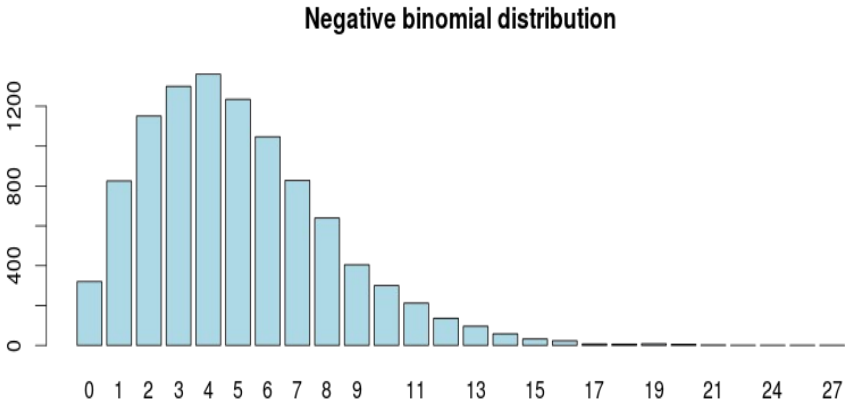


Modelling noise levels

- **Negative Binomial** distribution 2 parameters:
  - *p* = probability to start a read at a particular position
  - r = number of successes
- NB distribution can have arbitrarily large variance

$$Var(X_B) = (1-p)\bar{X} \qquad Var(X_{NB}) = \frac{\bar{X}}{1-p}$$

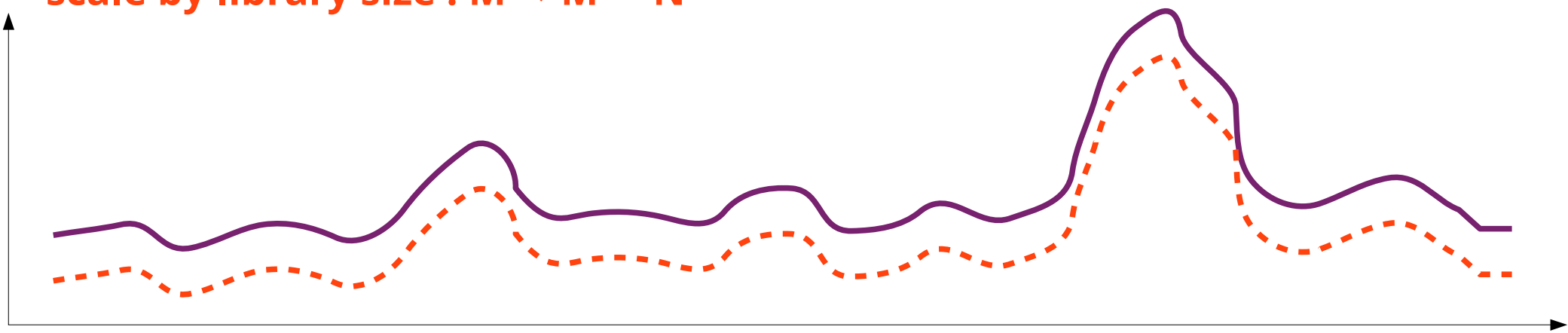




Scaling unequal datasets

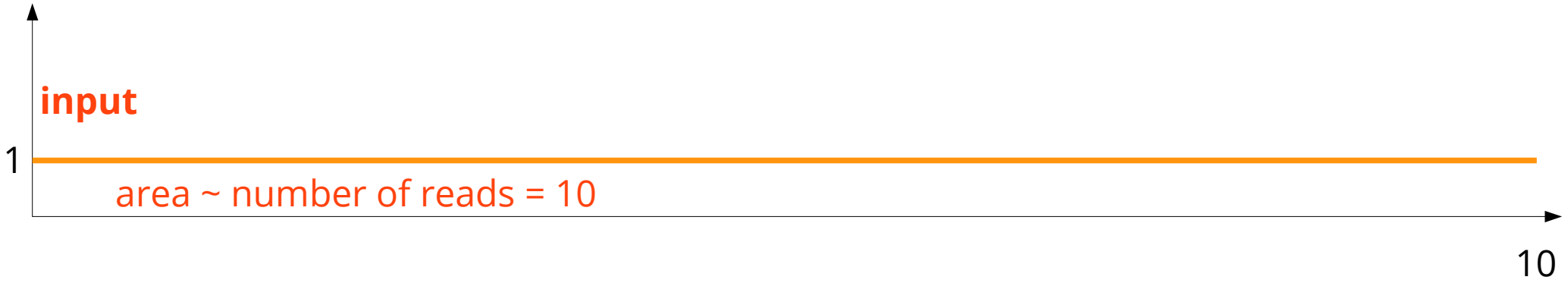
- treatment (=signal + noise) and input (=noise) datasets generally do not have the same sequencing depth → need for normalization
- input dataset should model the noise level in the treatment dataset
- **naïve approach** : upscale/downscale the smaller/larger dataset

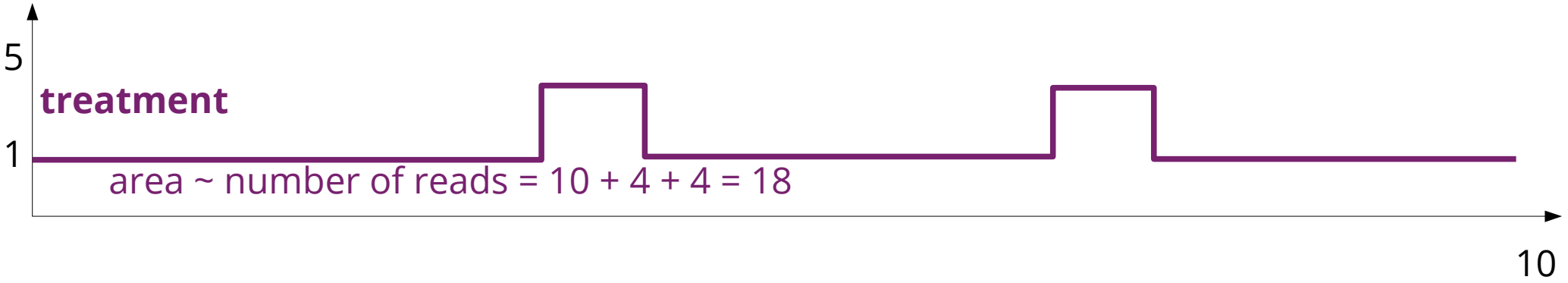
Input : N reads
ChIP-seq dataset → M > N reads
scale by library size : M → M' = N



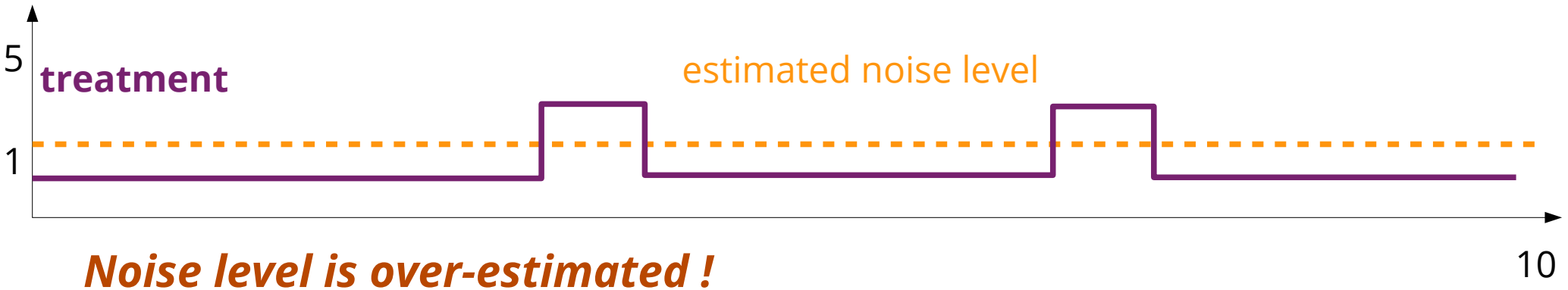
Problem : signal influences scaling factor
More signal (but equal noise) → artificial noise over-estimation

Scaling unequal datasets by library size



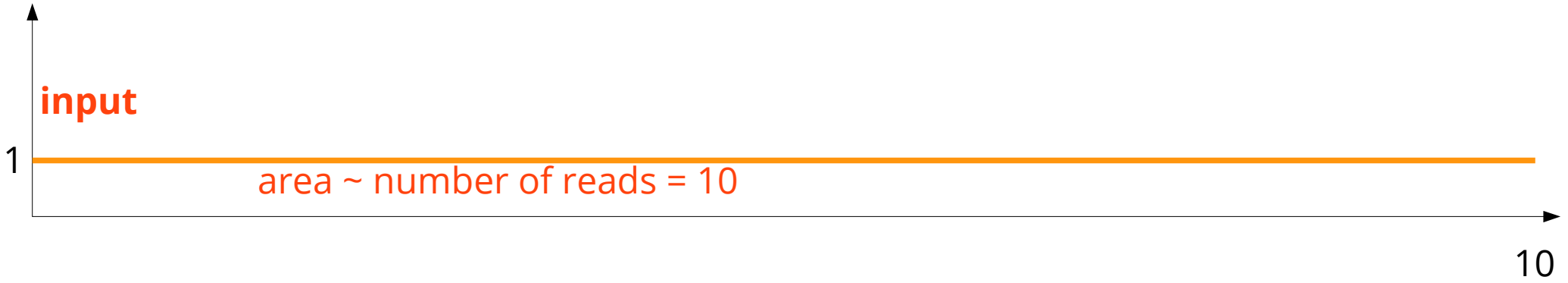


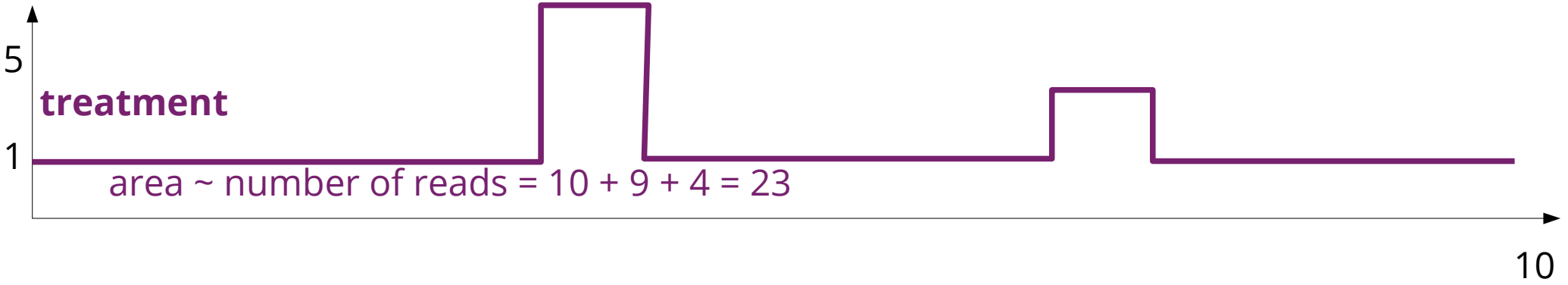
Scaling by library size : upscale input by 18/10 = 1.8



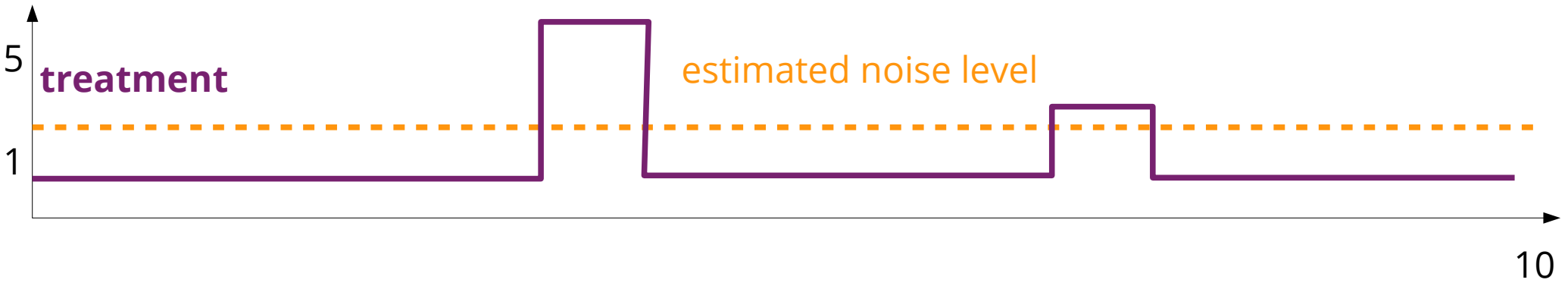
Noise level is over-estimated !

Scaling unequal datasets by library size

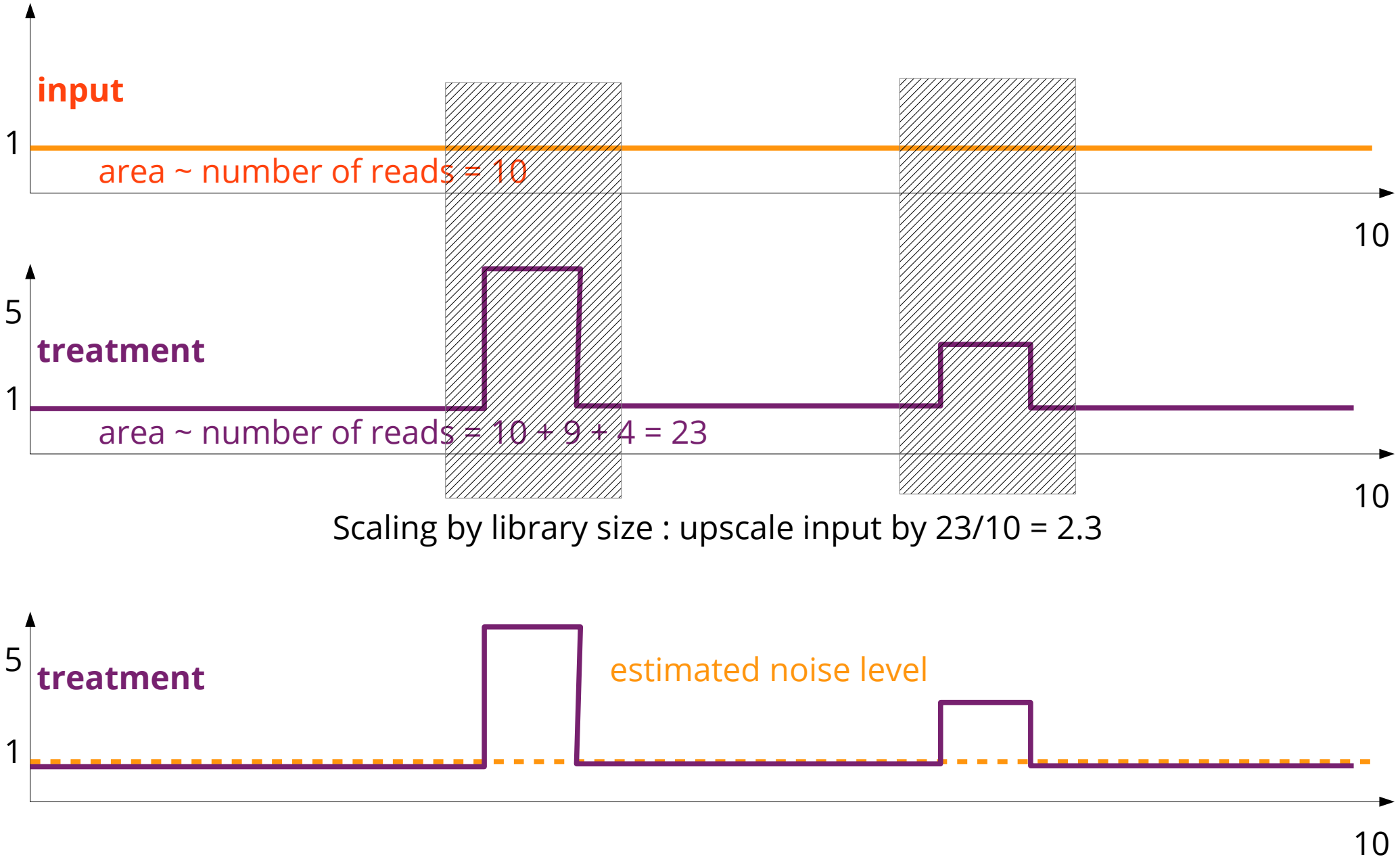




Scaling by library size : upscale input by 23/10 = 2.3



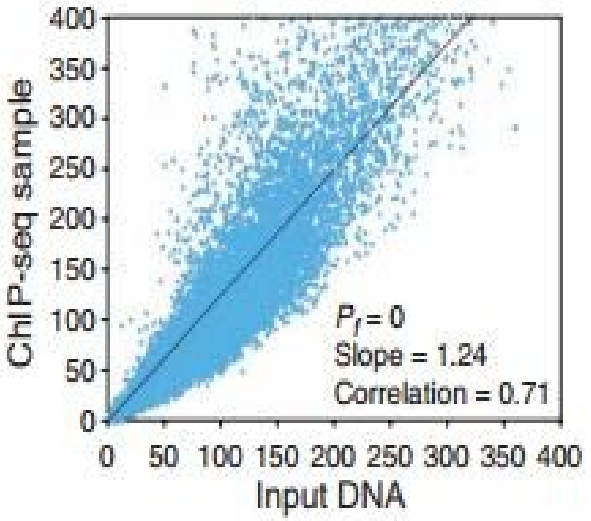
Scaling unequal datasets by library size



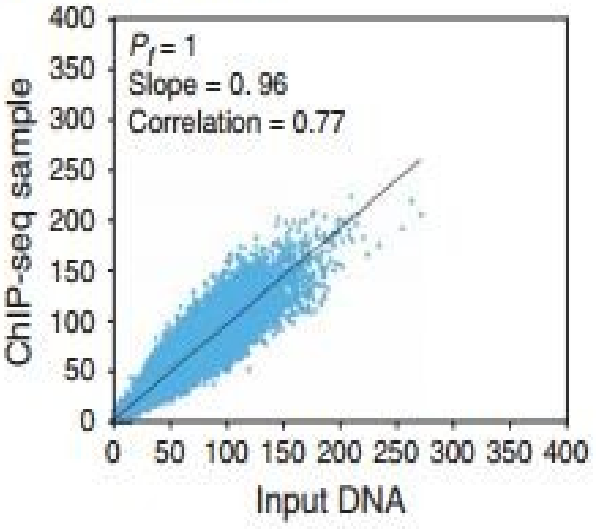
Scaling by library size : upscale input by 23/10 = 2.3

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

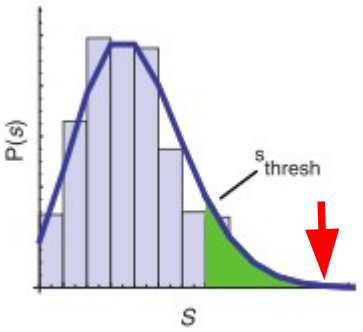
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls

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]

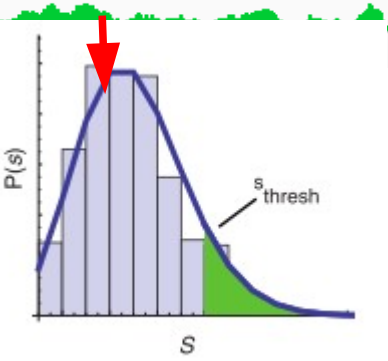
Defining “peaks”

- **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

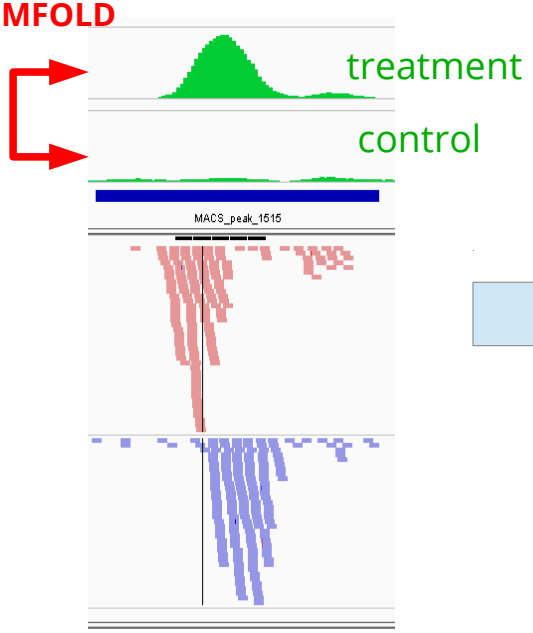
Tags unchanged

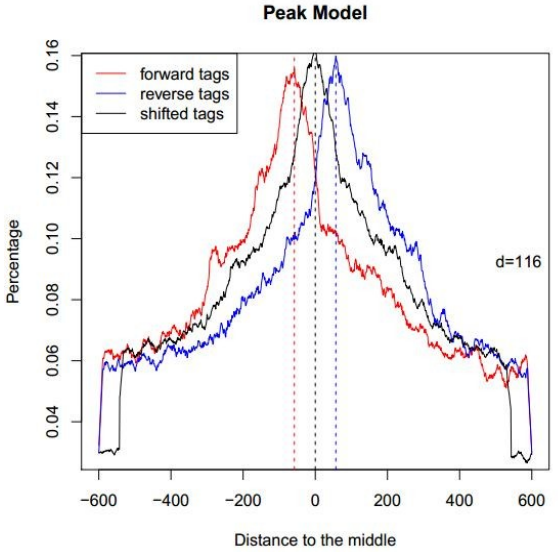
MACS

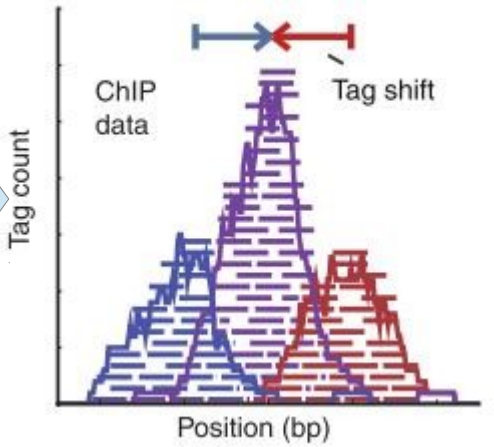
[Zhang et al. Genome Biol. 2008]

- **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

enrichment
> MFOLD



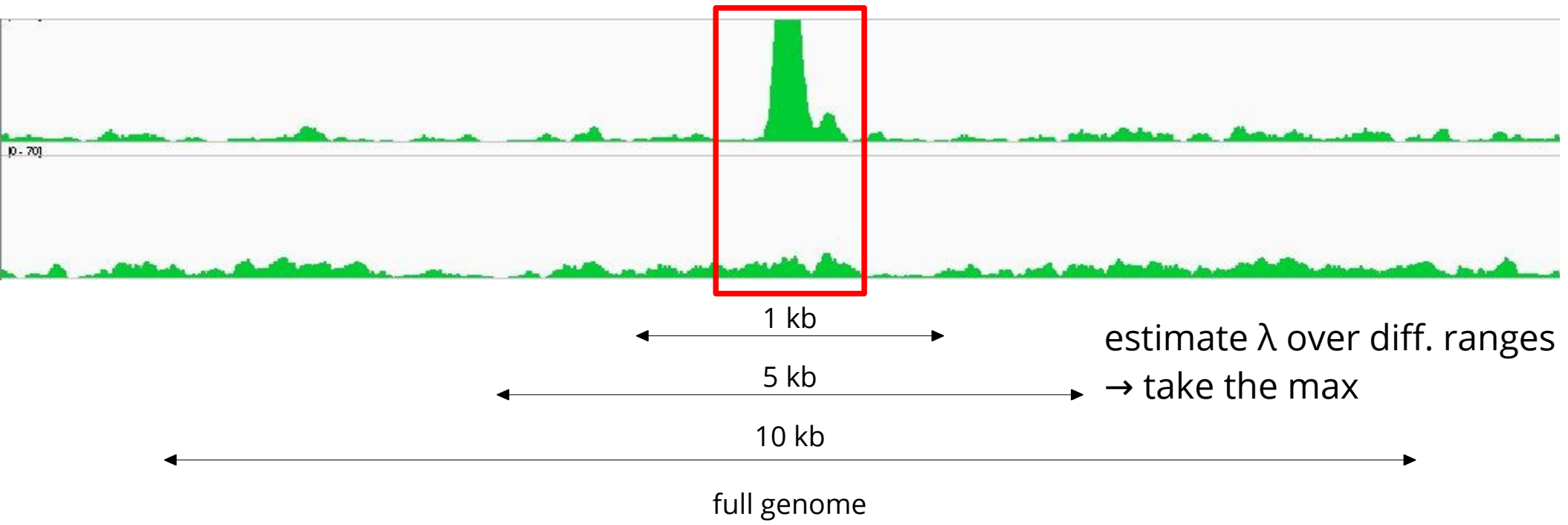




MACS

[Zhang et al. Genome Biol. 2008]

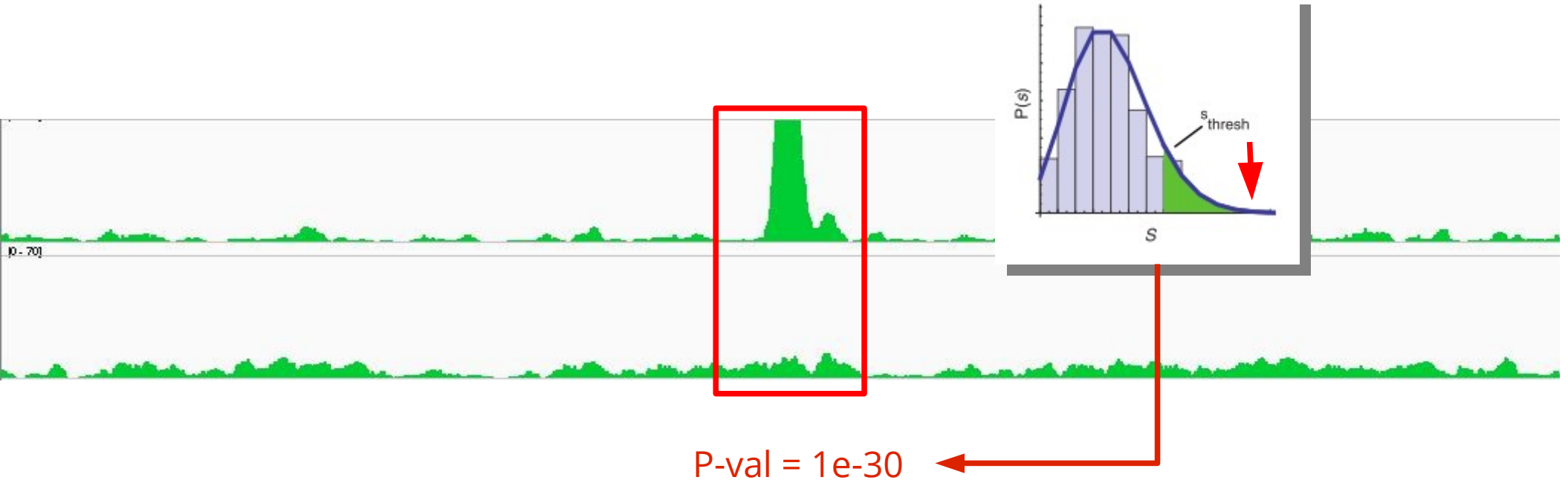
- **Step 2 : identification of local noise parameter**
  - slide a window of size *2*d* across treatment and input
  - estimate parameter λ_{local} of Poisson distribution



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

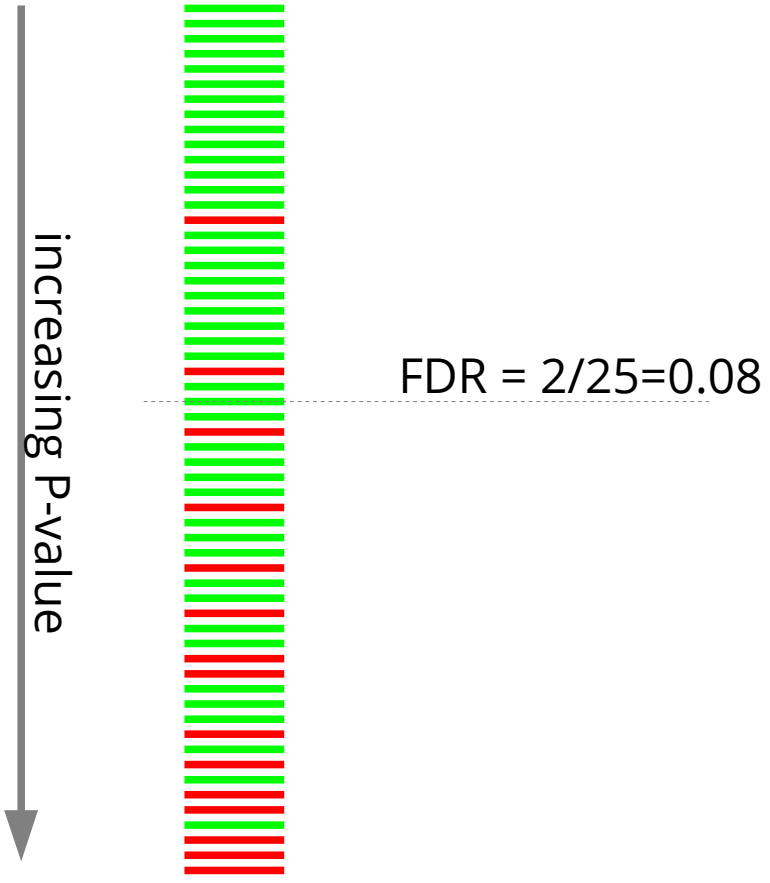


MACS

[Zhang et al. Genome Biol. 2008]

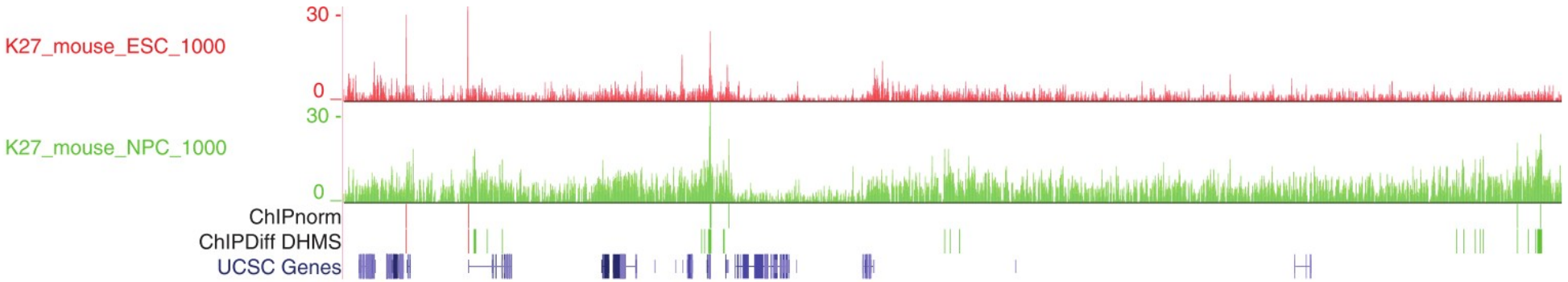
- **Step 4 : estimating FDR**
  - positive peaks (P-values)
  - swap treatment and input; call negative peaks (P-value)

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



Differential enrichment analysis

- ChIP-seq performed under two different conditions
- Question : what are the **differentially** enriched/bound peak regions ?



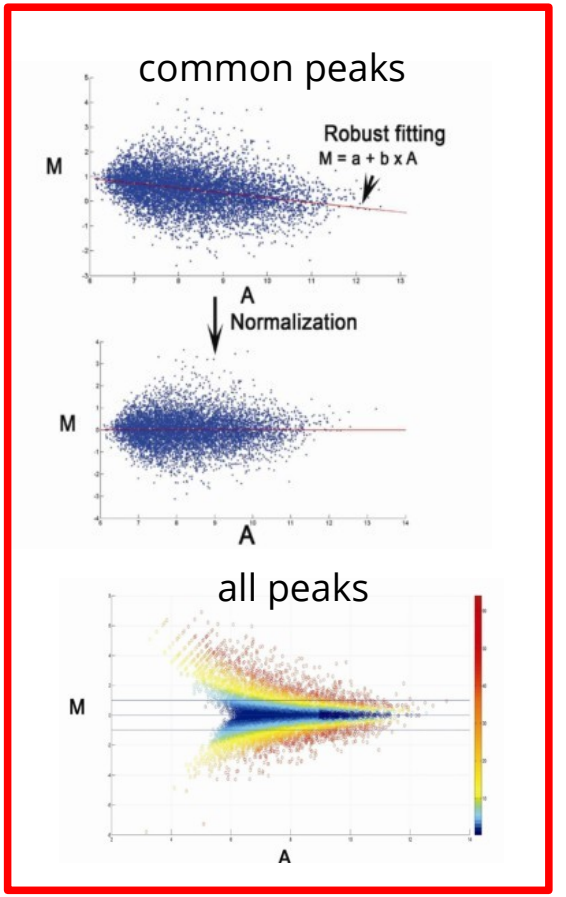
Nair et al., PLOS One 2012

- application to
  - histone modifications
  - DNA methylation

Differential enrichment analysis

- **ChIPDiff** (Xu. et al., Bioinformatics 2008)
  - statistical model : HMM (1 = not enriched; 2 = enriched in sampleA; 3 = enriched in sample B)
  - does not need pre-defined peaks/regions
  - command line executable
- **ChIPnorm** (Nair et al., PLOS One 2012)
  - quantile normalization of enriched-significant bins in both samples
  - requires signal and input datasets for both samples
  - MATLAB program
- **MAnorm** (Shao et al., Genome Biology 2012)
  - MA based normalization of regions containing **common peaks**
    → applied to all regions
  - requires a priori defined peaks for each library
  - MATLAB/R program
- **DiffReps** (Shen et al. PLOS One, 2013)
  - perl executables

MAnorm



Important questions to aks (beforehand...)

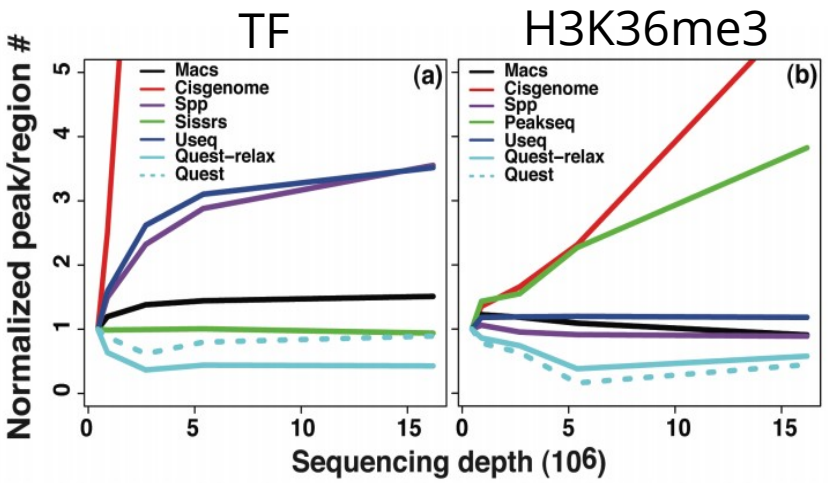
Systematic evaluation of factors influencing ChIP-seq fidelity

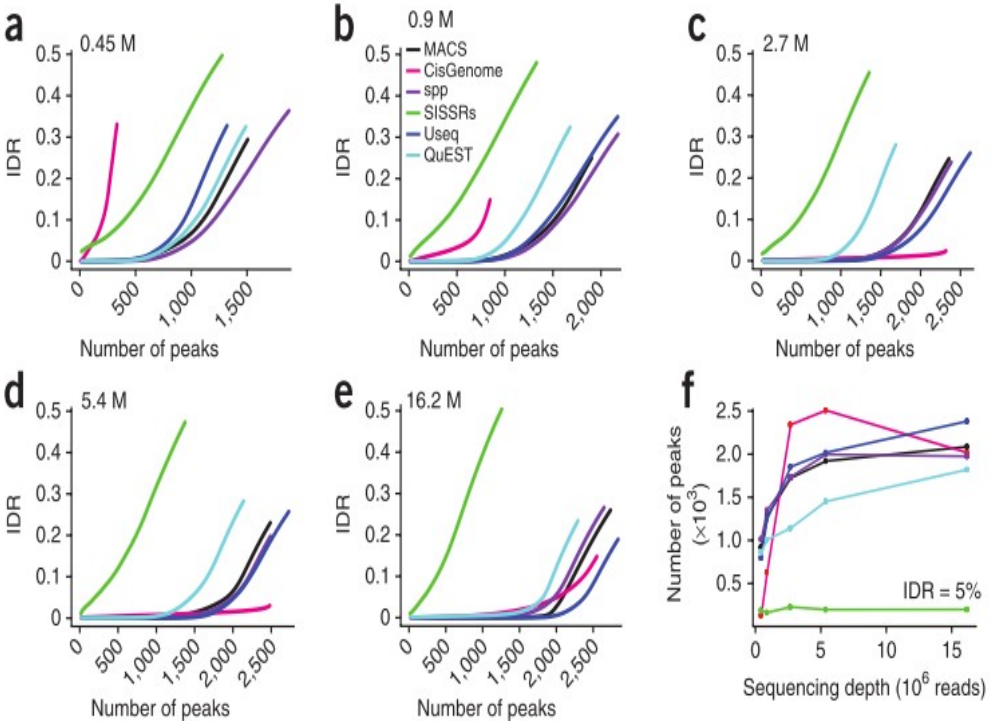
Yiwen Chen[1,12], Nicolas Negre[2,11,12], Qunhua Li[3], Joanna O Mieczkowska[4], Matthew Slattery[2], Tao Liu[1], Yong Zhang[5], Tae-Kyung Kim[6,11], Housheng Hansen He[1], Jennifer Zieba[2], Yijun Ruan[7], Peter J Bickel[8], Richard M Myers[9], Barbara J Wold[10], Kevin P White[2], Jason D Lieb[4] & X Shirley Liu[1]

Nature Methods 2012

Important questions to aks (beforehand...)

- sequencing depth ?
  - library complexity (insufficient depth → insufficient complexity)
  - saturation of ChIP peaks not achieved
- choice of peak caller
  - narrow peaks (TF) or broad peaks (histone modification)
  - sensitivity / specificity
  - **reproducibility** (across replicates) → Irreproducible Discovery Range (IDR)

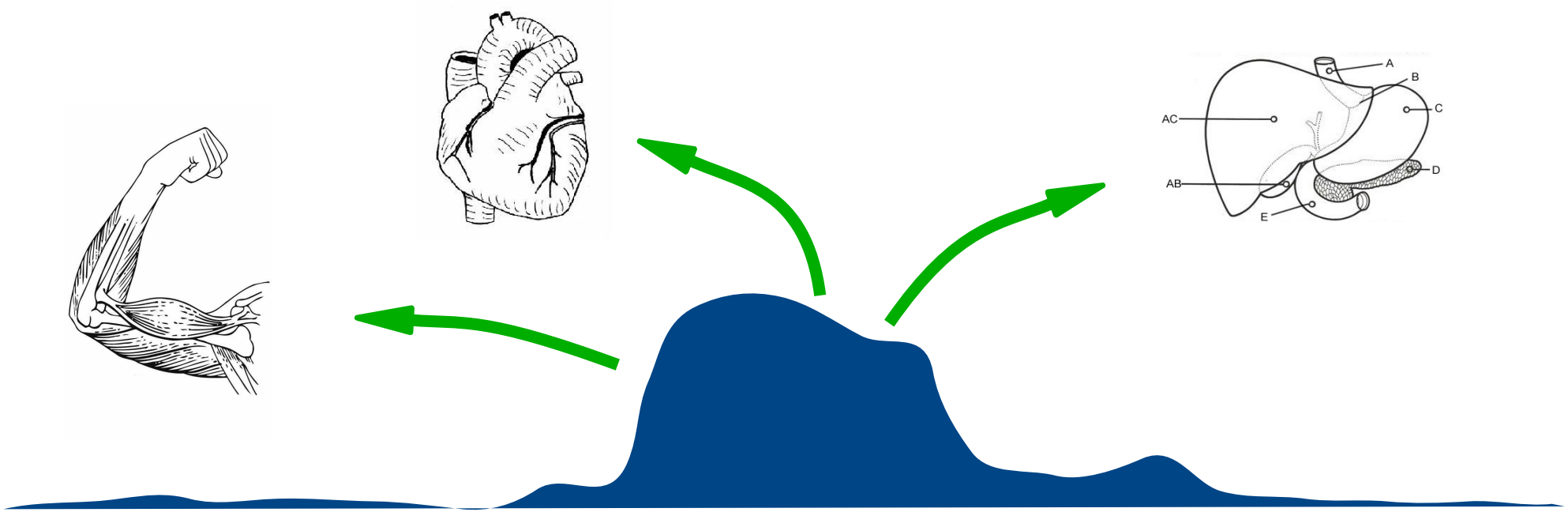




Important questions to aks (beforehand...)

- single vs PE ?
  - PE improves mapping in low-complexity regions
  - improves library complexity
  - but is generally an (financial) overkill ...
- remove redundant reads ?
  - for deeply sequenced libraries, redundant reads are not always artefacts ...

Functional annotation of ChIP-peaks



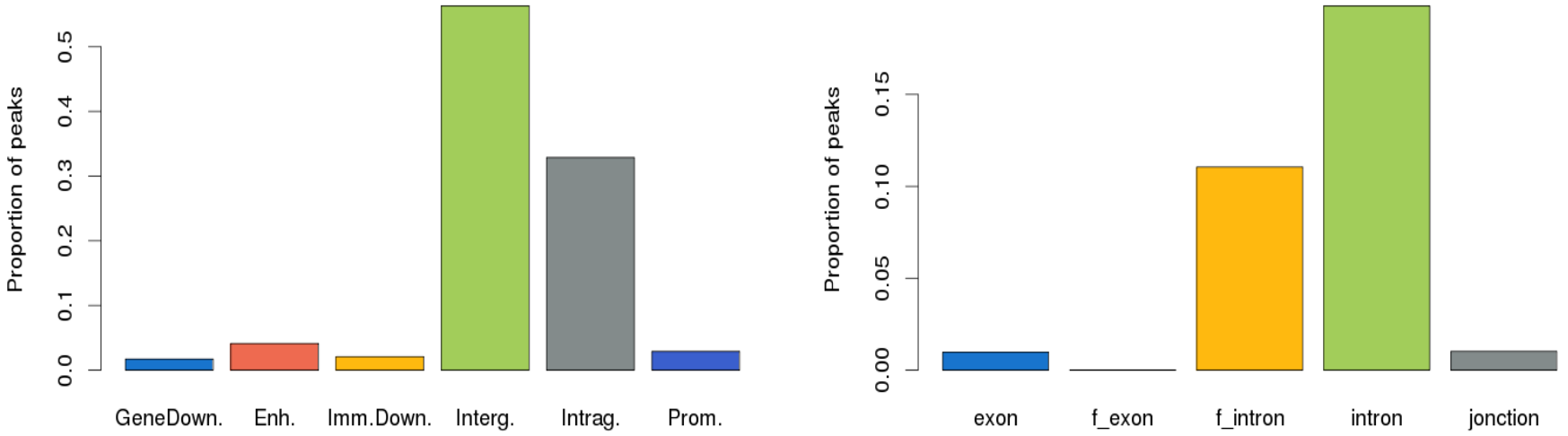
how do we go from peaks to functions ?

Interesting questions to ask

- where do peaks localize ?
  - proximal to TSS ?
  - distal (= enhancer) regions ?
- what are the closest genes (potential targets) ?
- is there a functional enrichment (e.g. GO categories) in genes/regions bound ?

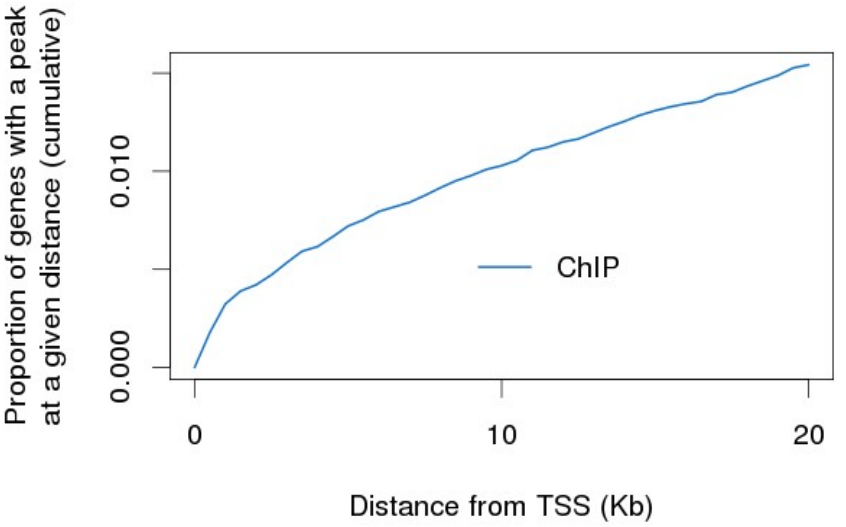
Positional biases

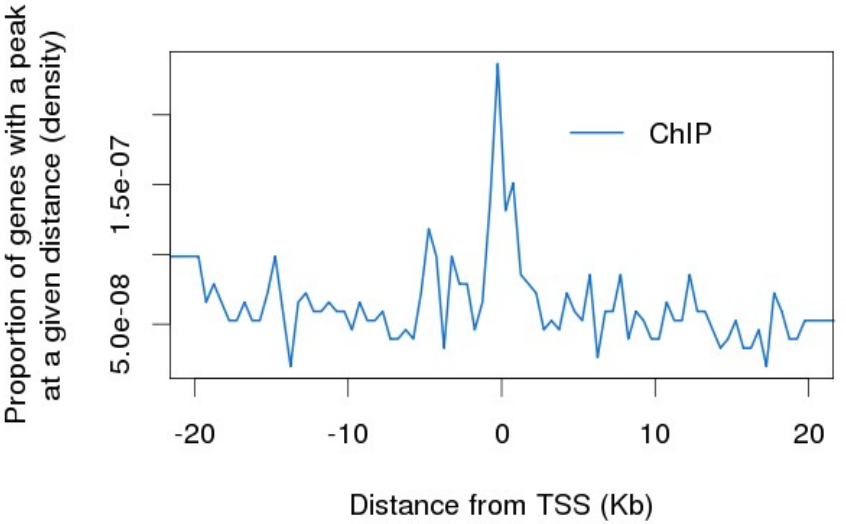
- where do peaks localize ? Proximal promoter ? Intergenic regions ? Intronic regions ?



Positional biases

- Distance to TSS :





Peaks → **Genes → Functions**

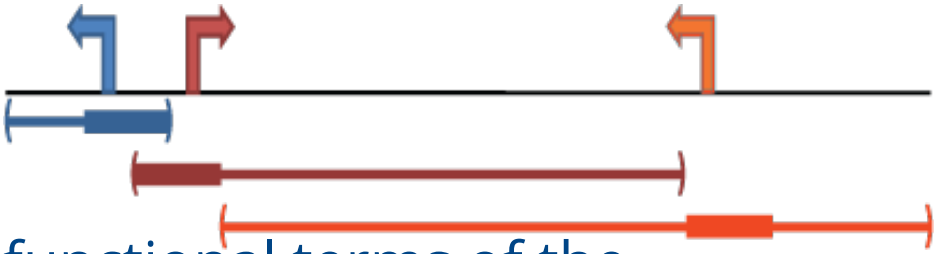
- collect sets of genes
- compute over-represented functional annotations
  - Gene Ontology
  - Phenotypic annotations
  - Biological Pathways
- Typical tools
  - **DAVID** [Huang et al., NAR 2009]
  - **Babelomics** [Medina et al., NAR 2010]

Peaks → Genes → Functions

- **Drawbacks**
  - restricting to proximal regions **discards** a large number of binding events
  - "nearest gene" approach introduces **bias** towards genes with large intergenic regions
    e.g. : "multicellular organism development" : 14% of the genes, but 33% of the genome associated

Genes → Regions ← Peaks

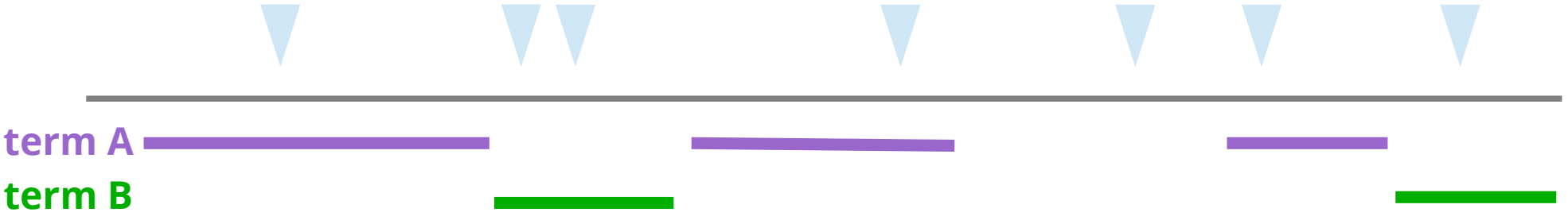
- **Idea** :
  - assign functional annotation to genomic regions
  - use statistics to avoid biases
- assign to each gene a regulatory domain
  - basal (-5kb/+1kb from TSS)
  - extended (up to nearest basal region ; max 1Mb)



- each domain is annotated to the functional terms of the corresponding gene
  → "Functional domains"

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Genes → Regions ← Peaks



*Given that **60%** of the genome is annotated to A, would I randomly expect 3 or more peaks to fall into region A ?*

p > 0.5

*Given that **15%** of the genome is annotated to B, would I randomly expect 3 or more peaks to fall into region B ?*

p = 0.07

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Job description

Job ID: 20111012-public-uXBYVG

Display name: GSM348066_limb_p300_peaks.NEW

Test set: GSM348066_limb_p300_peaks.NEW.bed (3,839 genomic regions)
Show in UCSC genome browser. *What is this?*

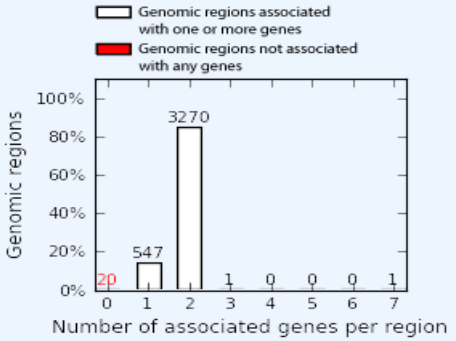
Background: Whole genome background

Assembly: Mouse: NCBI build 37 (UCSC mm9, Jul 2007) *What gene set does GREAT use?*

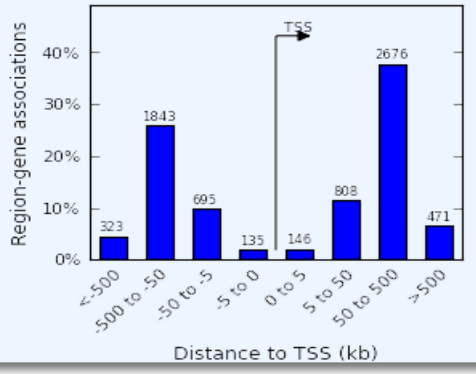
Associated genomic regions: Basal+extension (constitutive 5.0 kb upstream and 1.0 kb downstream, up to 1000.0 kb max extension). Curated regulatory domains are included.
20 of all 3,839 genomic regions (0.5%) are not associated with any genes.
View genomic region-gene associations. *What is this?*
Revise the region-gene association rule. *What is a region-gene association rule?*

Region-gene association graphs:

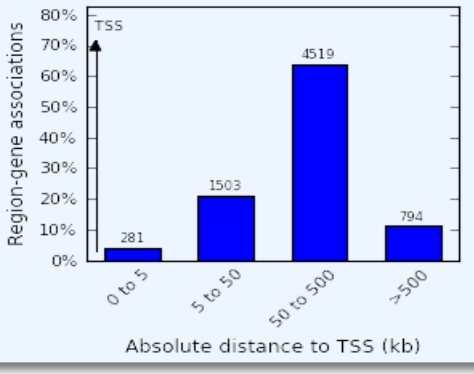
Number of associated genes per region
Download as PDF.



Binned by orientation and distance to TSS
Download as PDF.



Binned by absolute distance to TSS
Download as PDF.



Mouse Phenotype

Global Controls

Table controls: Export ▾ Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set

Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
abnormal limbs/digits/tail morphology	2	**2.0559e-91**	**6.6837e-88**	2.1465	780	20.32%	6	**2.5295e-40**	2.2020	278	681	8.31%
abnormal craniofacial morphology	3	**9.3822e-91**	**2.0334e-87**	2.0082	887	23.10%	10	**8.9231e-36**	2.0382	297	786	8.88%
abnormal limb morphology	5	**2.4990e-80**	**3.2497e-77**	2.3077	604	15.73%	9	**7.4787e-37**	2.4541	202	444	6.04%
abnormal appendicular skeleton morphology	10	**3.0255e-70**	**1.9672e-67**	2.3450	517	13.47%	17	**3.9549e-30**	2.4098	172	385	5.14%
abnormal skeleton extremities morphology	12	**3.2687e-69**	**1.7711e-66**	2.3724	499	13.00%	21	**7.0557e-29**	2.4222	163	363	4.87%
abnormal paw/hand/foot morphology	13	**4.0300e-69**	**2.0156e-66**	2.6813	404	10.52%	23	**5.4918e-28**	2.7186	126	250	3.77%
abnormal head morphology	14	**6.4657e-67**	**3.0029e-64**	2.0134	672	17.50%	25	**2.9042e-27**	2.0562	223	585	6.67%
abnormal digit morphology	18	**1.0543e-61**	**3.8084e-59**	2.6982	358	9.33%	36	**1.2033e-25**	2.7998	109	210	3.26%
abnormal cartilage morphology	23	**7.3728e-58**	**2.0843e-55**	2.3432	430	11.20%	29	**1.1337e-26**	2.5089	140	301	4.19%
abnormal skeleton development	24	**3.5769e-56**	**9.6904e-54**	2.0833	530	13.81%	38	**5.2377e-25**	2.1414	185	466	5.53%
abnormal long bone morphology	25	**4.6593e-56**	**1.2118e-53**	2.3374	419	10.91%	43	**4.9983e-24**	2.3823	140	317	4.19%

GREAT vs. proximal peaks

	GREAT				Proximal 2kb peaks			
	Best GO term	*P-val*	*MGI expression*	*P-val*	*Best GO-term*	*P-val*	*MGI expression*	*P-val*
p300 limb	Embryonic limb morphogenesis	1E-27	TS19 limb	7E-49	Skeletal system development	4E-06	TS19 limb	3E-05
p300 forebrain	CNS development	8E-36	TS17 forebrain	6E-41	Forebrain development	2E-04	TS22 forebrain	3E-07
p300 midbrain	CNS development	1E-12	TS 15 CNS	1E-14	none		none	

- **more specific terms with higher significance**
- **more peaks/genes taken into account**