

ChIP-seq analysis

M. Defrance, C. Herrmann, D. Puthier, M. Thomas.Chollier

- Tuesday :
 - quick introduction to ChIP-seq and peak-calling (Presentation + tutorial)
- Wednesday:
 - ChIP-peaks functional annotation (Presentation + Practical)
 - motif discovery in ChIP-seq peaks (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}

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

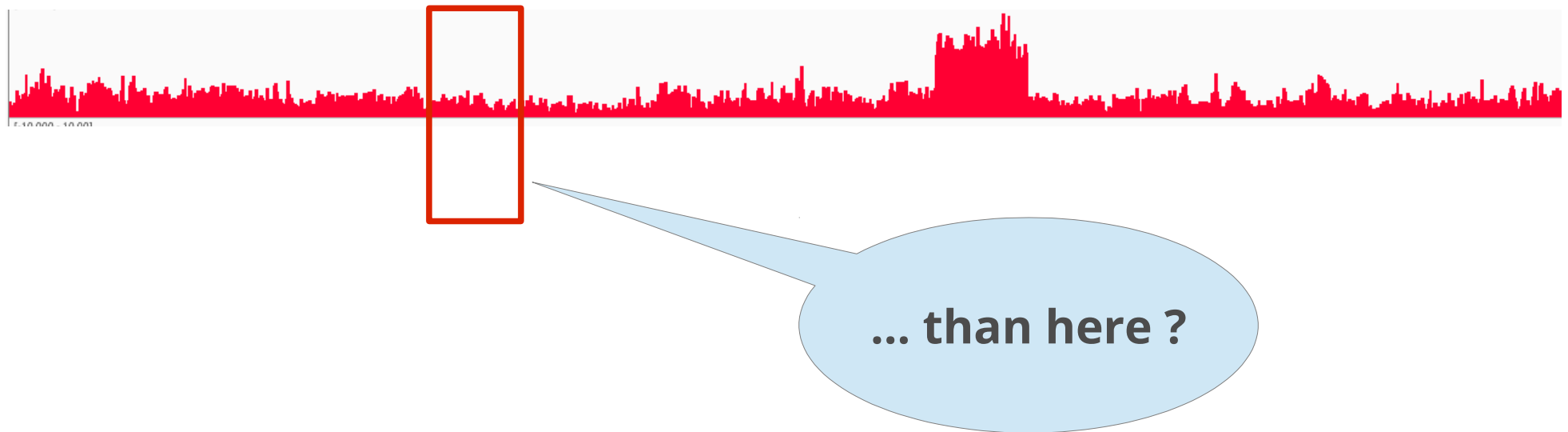
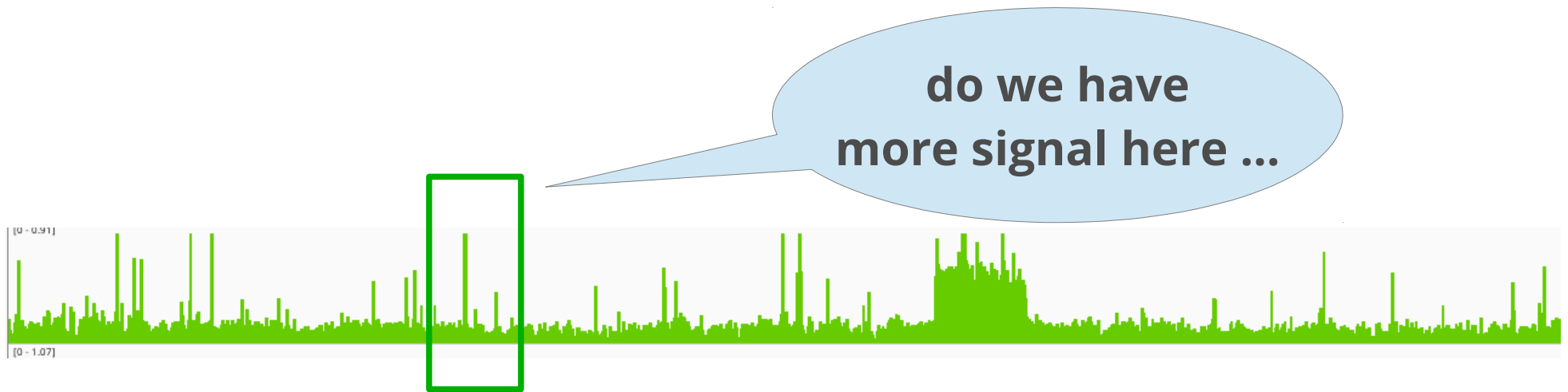
- estrogen-receptor (ESR1) is a key factor in **breast cancer development**
- goal of the study: understand the dependency of ESR1 binding on presence of co-factors, in particular GATA3, which is mutated in breast cancers
- approaches: GATA3 silencing (siRNA), ChIP-seq on ESR1 in wt vs. siGATA3 conditions,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

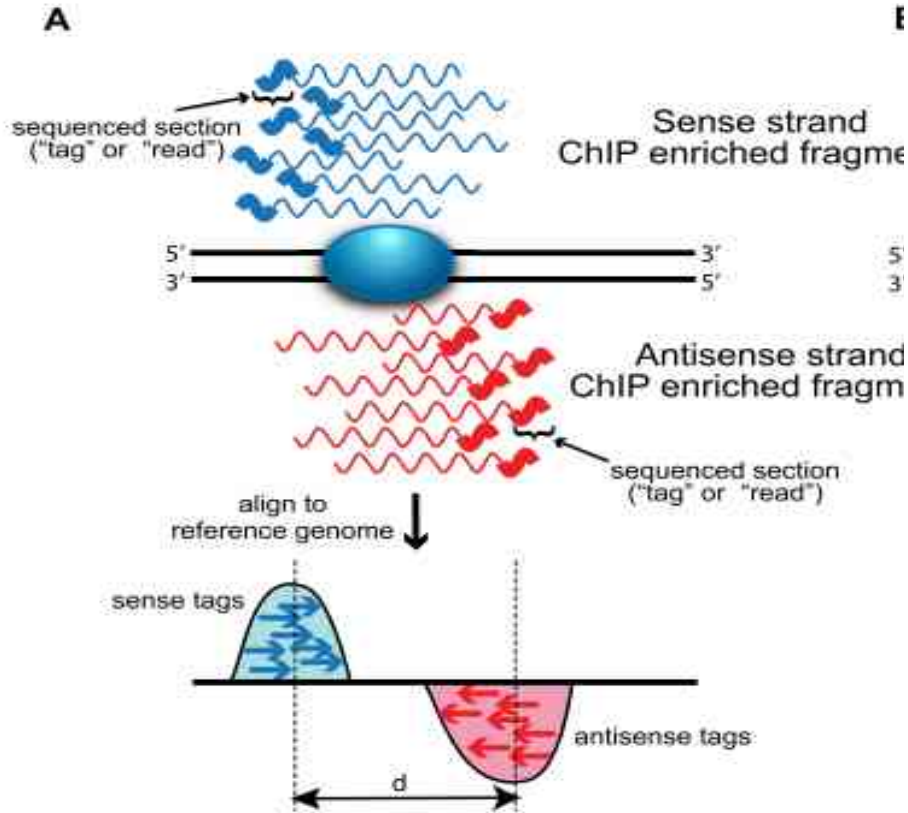
What we want to do



Keys aspects of ChIP-seq analysis

- (1) Determine signal **coverage**
- (2) Modelling **noise** levels
- (3) Scaling/**normalizing** datasets
- (4) Detecting enriched **peak** regions
- (5) Performing **differential** analysis

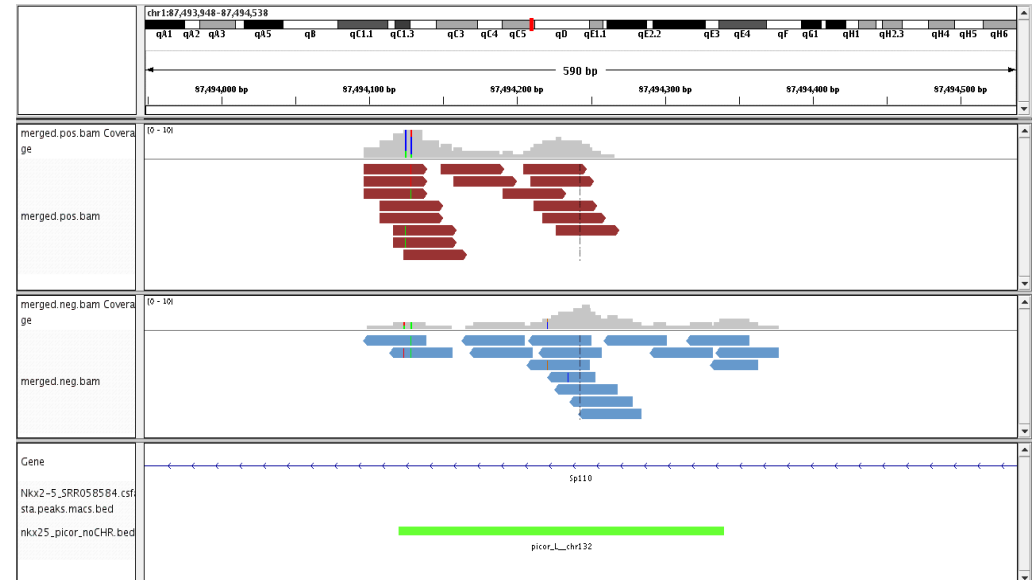
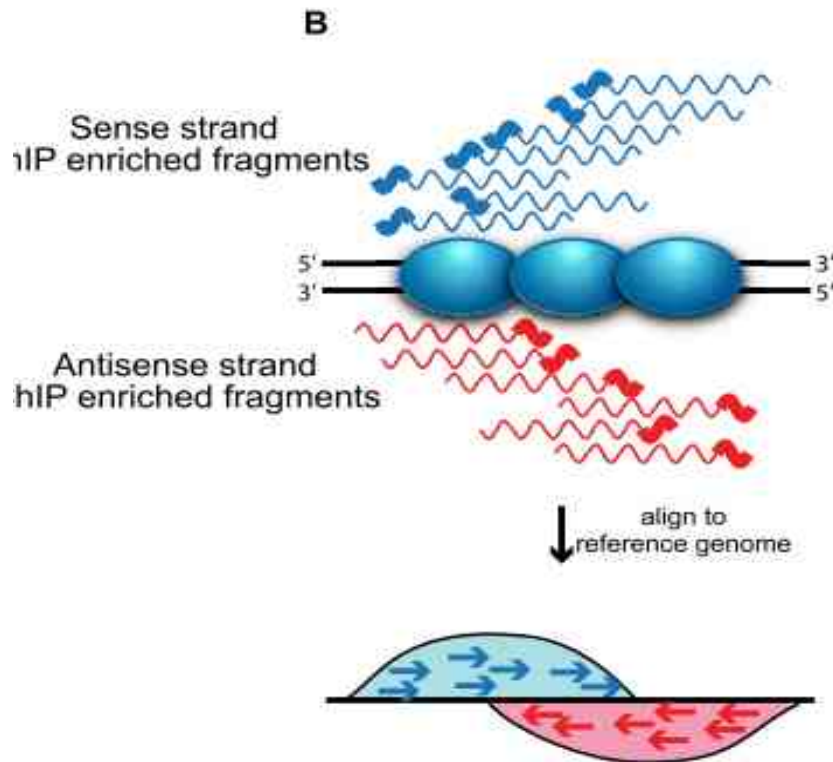
Principle of ChIP-seq



The binding site itself is generally not sequenced !

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

Principle of ChIP-seq



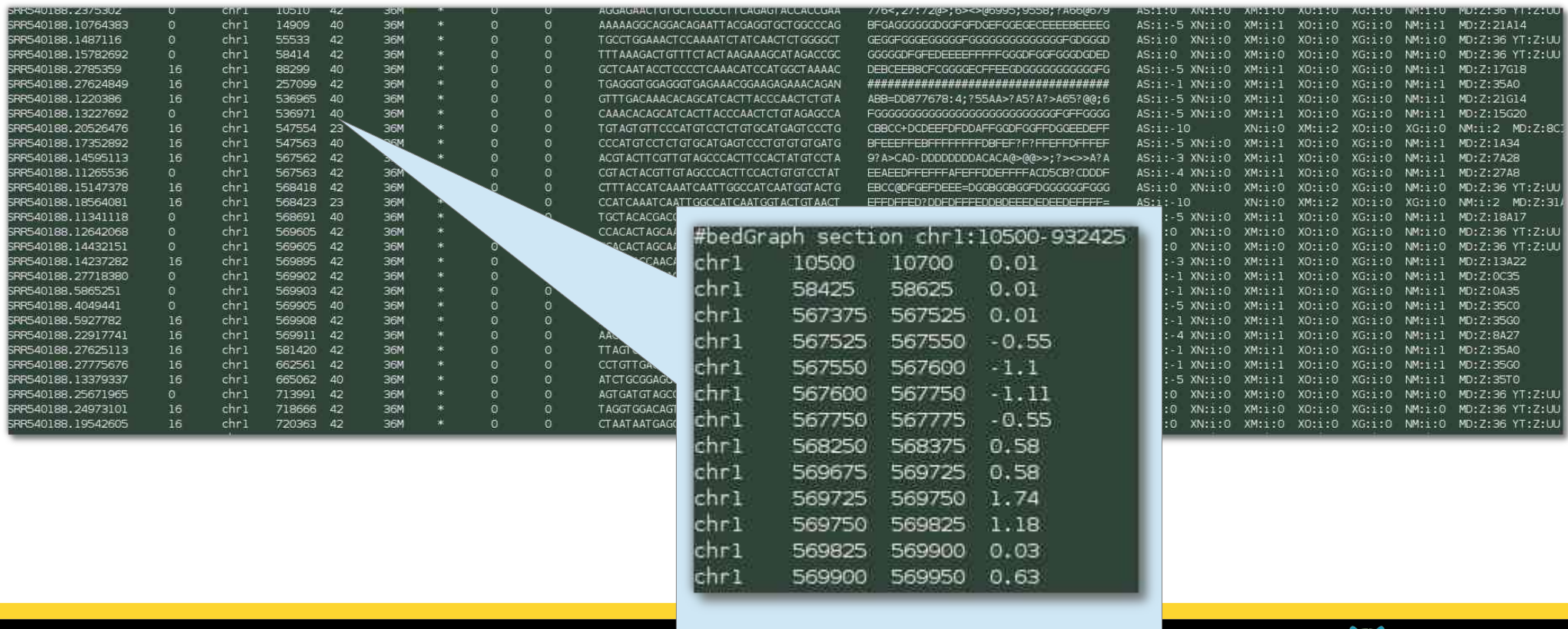
Strand asymmetry is blurred when multiple proteins bind or in case of histone modifications ChIP

Principle of ChIP-seq

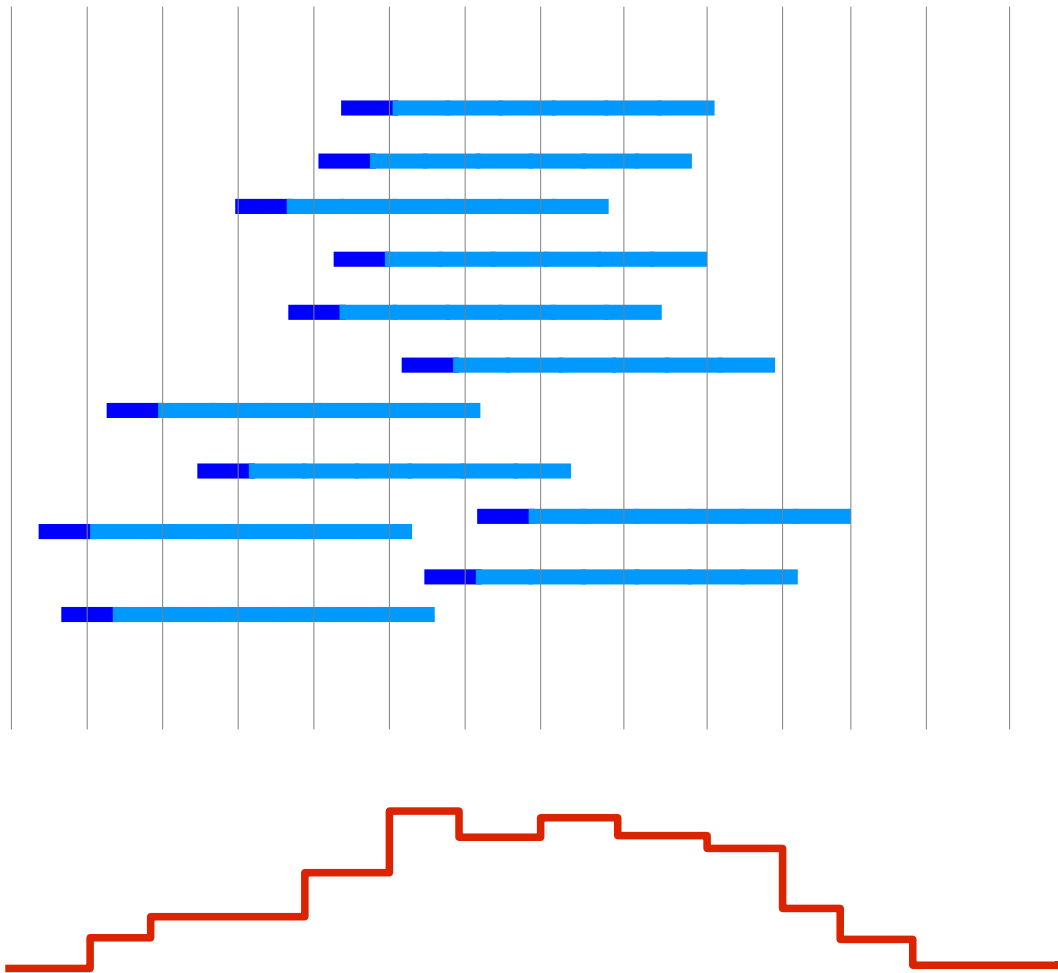


1. from reads to coverage

- to visualize the data, we use **coverage plots** (=density of fragments per genomic region)
- need to reduce BAM file to more compact format
→ **bigWig/bedGraph**



1. from reads to coverage

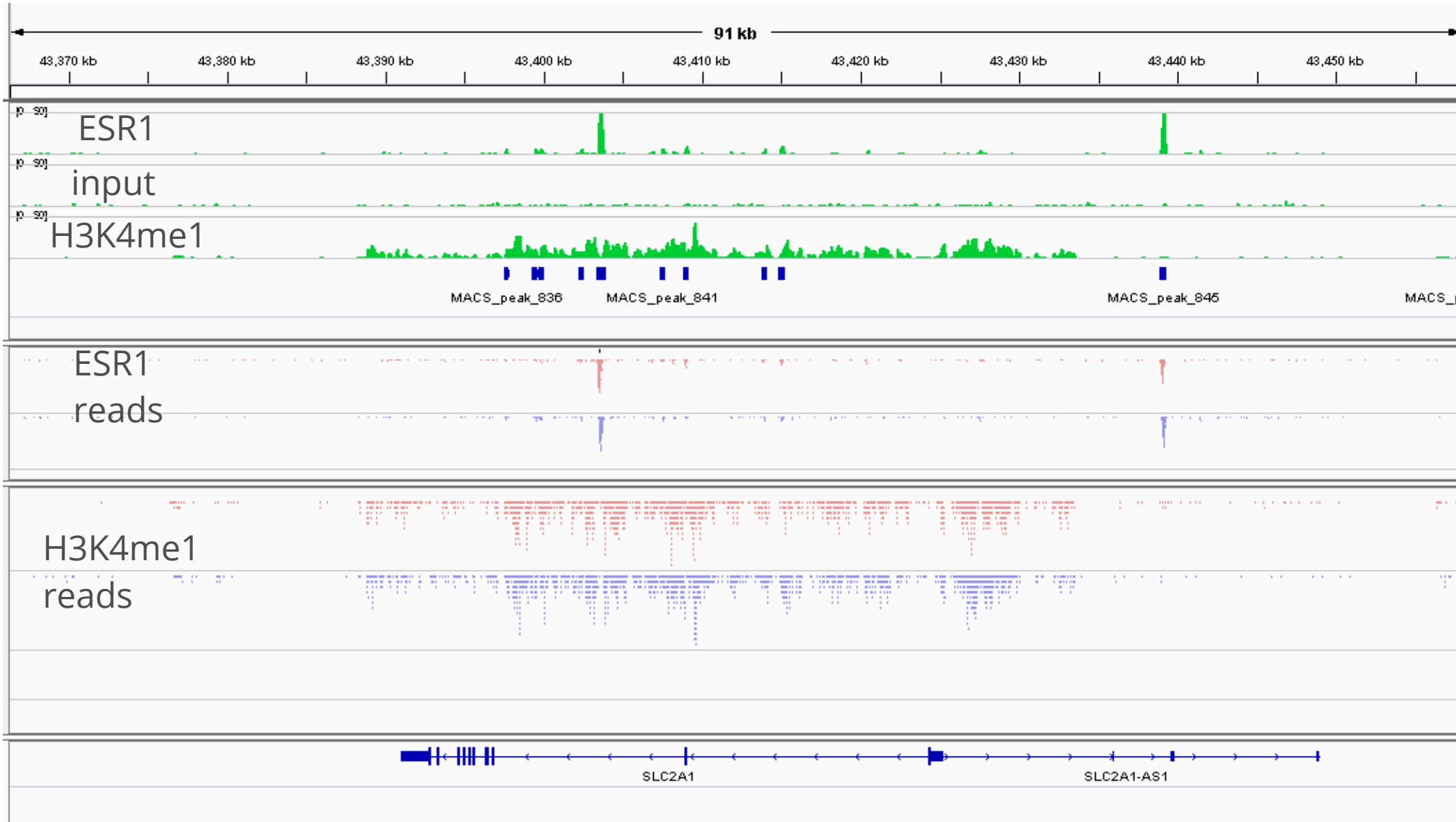


- Reads are extended to 3' to fragment length
- Read counts are computed for each bin
- Counts are normalized
 - reads per genomic content
→ normalize to 1x coverage
- reads per kilobase per million reads per bin

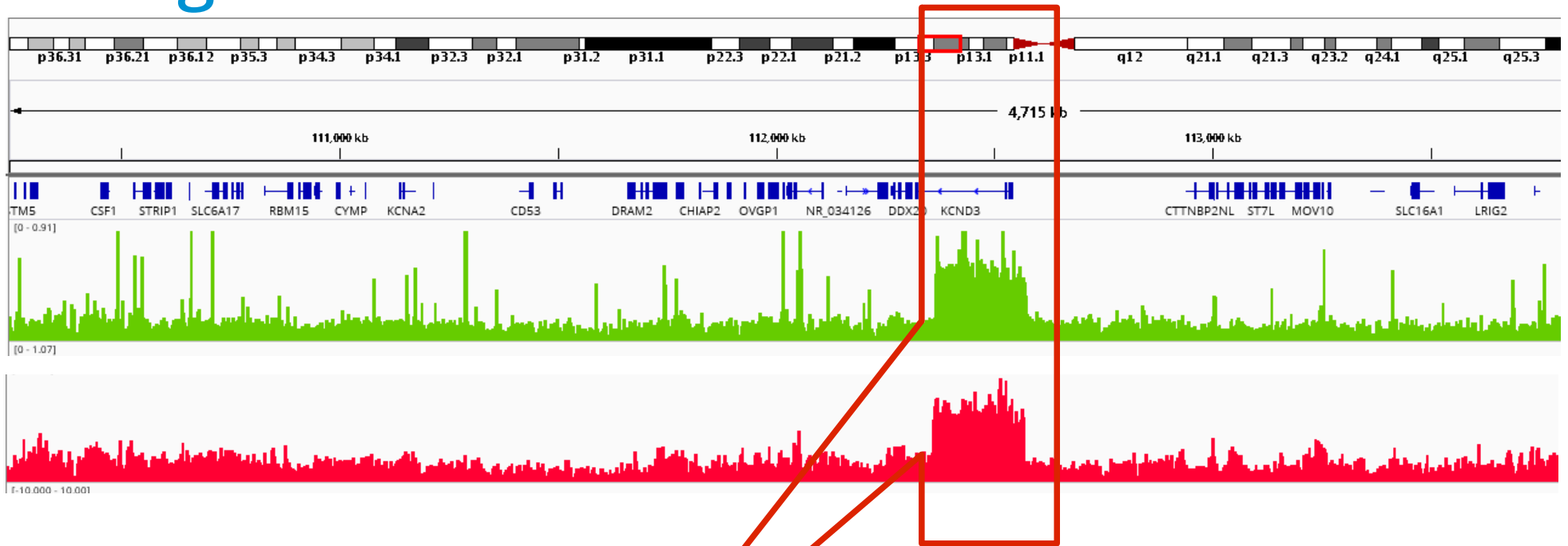
$$SD = \frac{n_{\text{mapped reads}} \times L}{G_{\text{eff}}}$$

$$RPKM = \frac{n_{\text{reads/bin}} \times W_{\text{bin}}}{n_{\text{mapped reads}}}$$

1. from reads to coverage



2. signal and noise

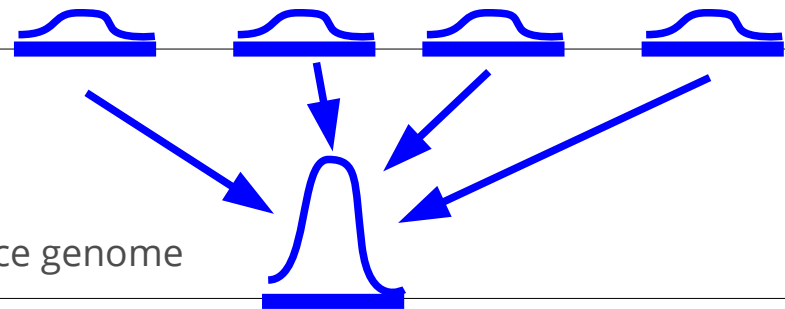


MCF-7 genome

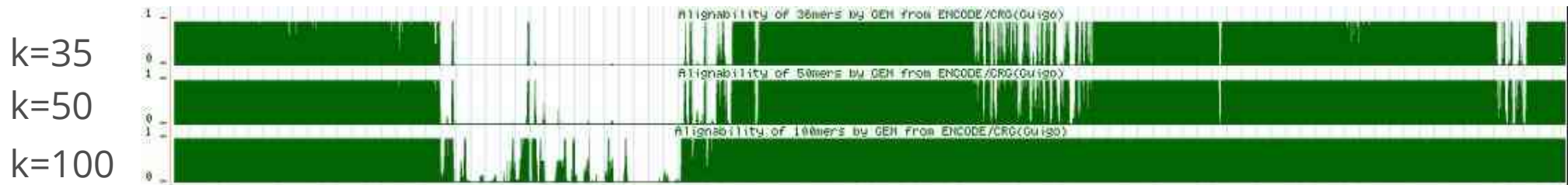
The MCF-7 genome harbors 21 high-level CNAs, summarized in Table 1. Remarkably, many of the previously reported regions of genetic alteration split into multiple segments upon tiling resolution analysis. The 1p13 amplification described previously [40] in fact divides into three distinct segments of high-level amplifications: a 1,300 kb segment at 1p13.3, containing only two genes, those encoding arginine N-methyltransferase-6 (*PMRT6*) and netrin G1 (*NTNG1*);

MCF7 genome

hg19 reference genome



2. signal to noise



- **Mappability issue** : alignability track shows, how many times a read from a given position of the genome would align
 - $a=1$ → read from this position ONLY aligns to this position
 - $a=1/n$ → read from this position could align to n locations→ we usually only keep uniquely aligned reads : **positions with $a < 1$ have no reads left**

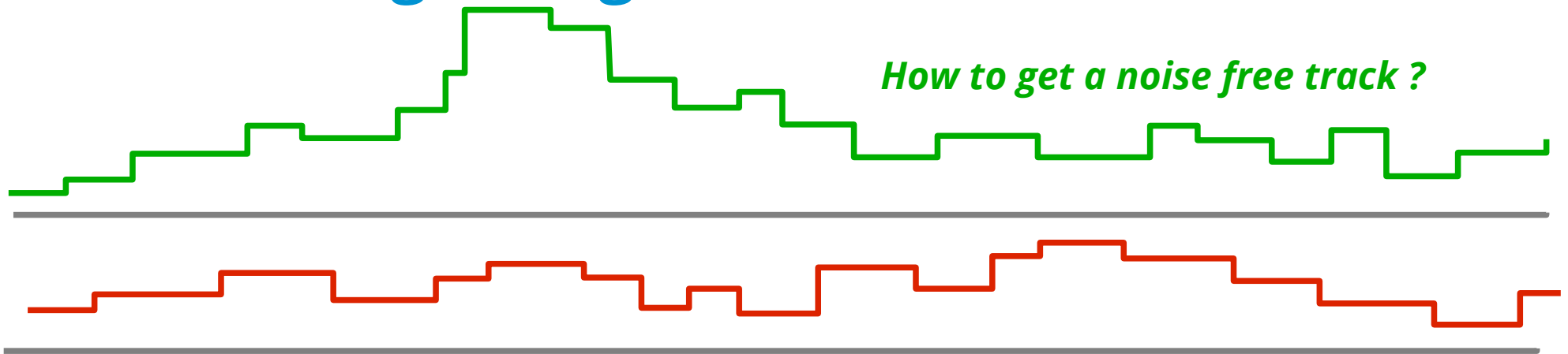
2. signal to noise

The availability of a control sample in mandatory !

→ mock IP with unspecific antibody

→ sequencing of input (=naked) DNA

3. modelling background level

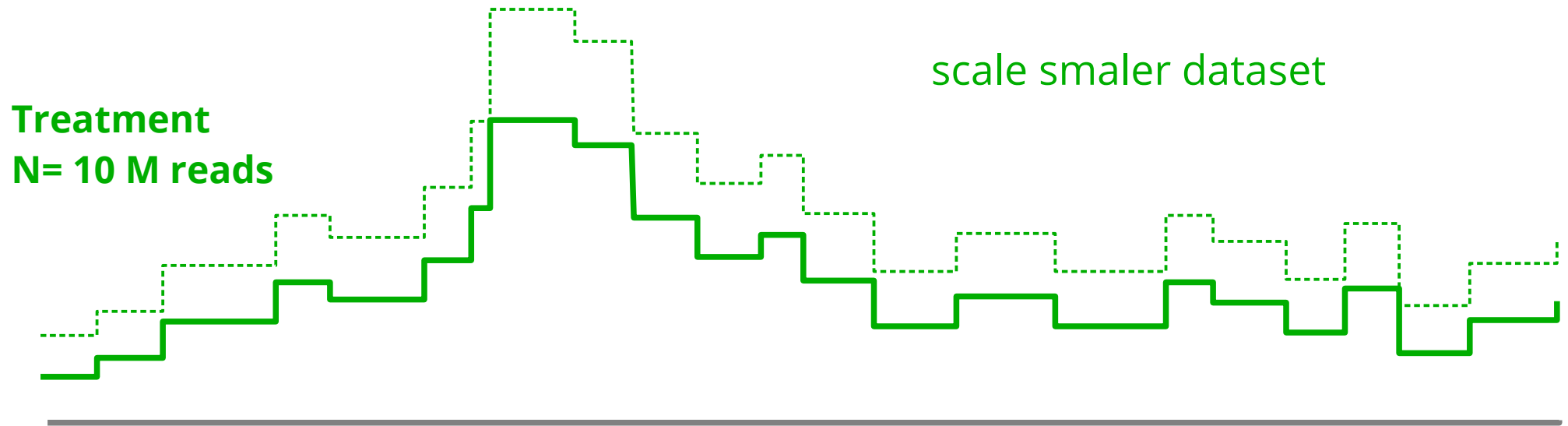


- **naïve subtraction** treatment – input is not possible, because both libraries have different sequencing depth !
- **Solution 1** : before subtraction, scale both libraries by total number of reads (library size)

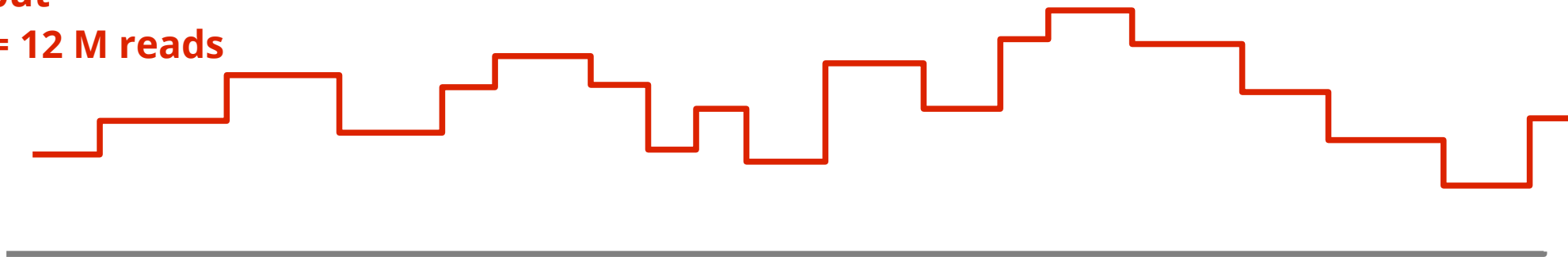
- RPGC
$$SD = \frac{n_{\text{mapped reads}} \times L}{G_{\text{eff}}}$$

- RPKM
$$RPKM = \frac{n_{\text{reads/bin}} \times W_{\text{bin}}}{n_{\text{mapped reads}}}$$

3. modelling background level



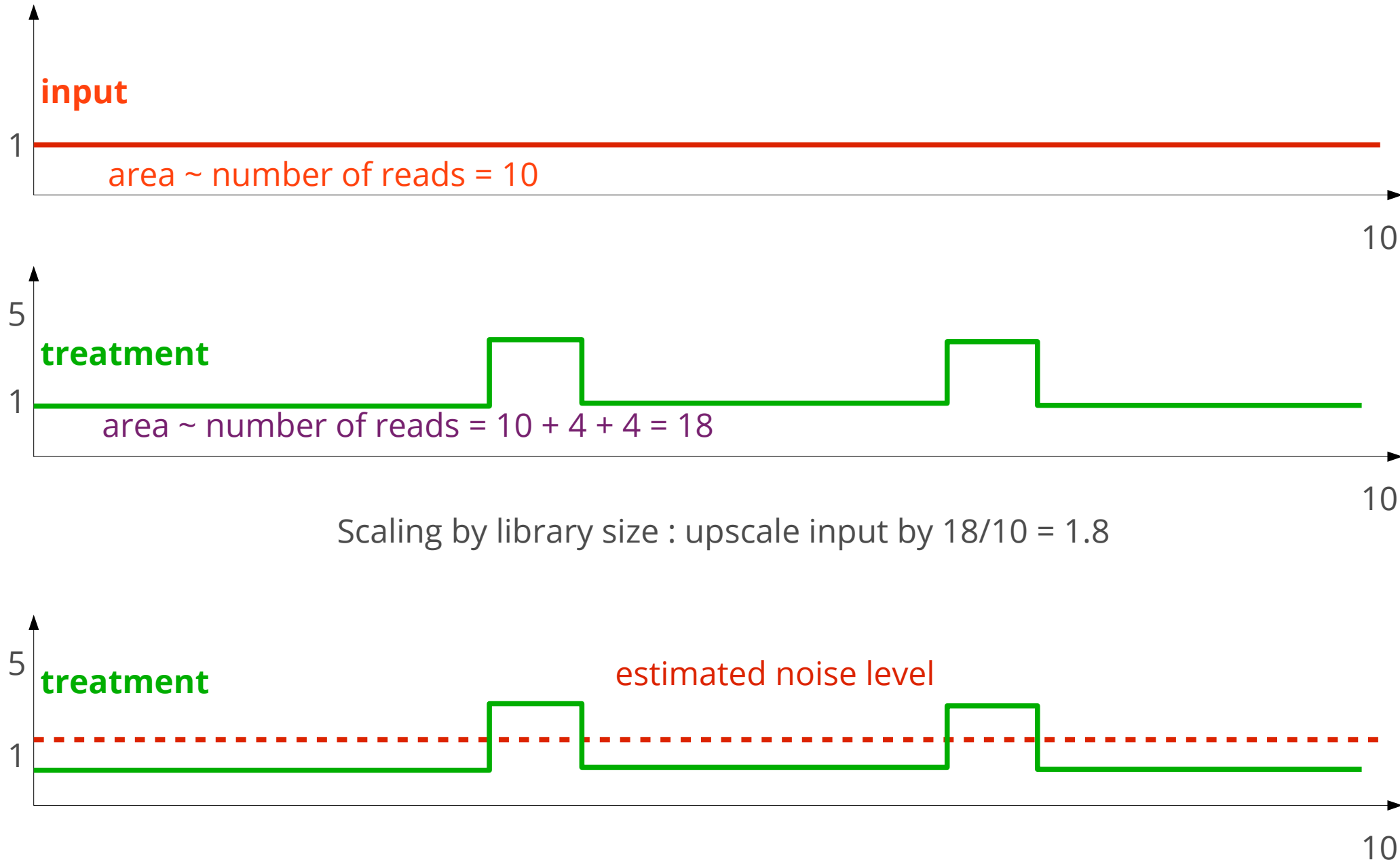
Input
N'= 12 M reads



Problem : signal influences scaling factor

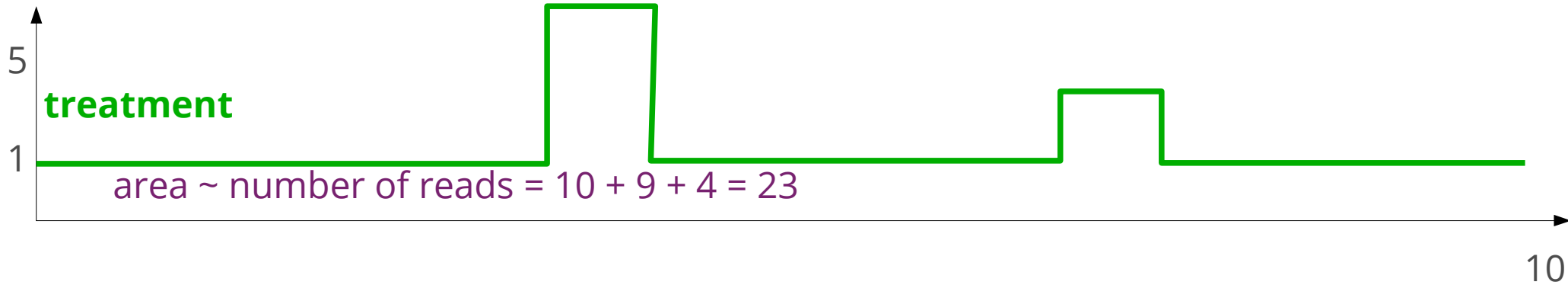
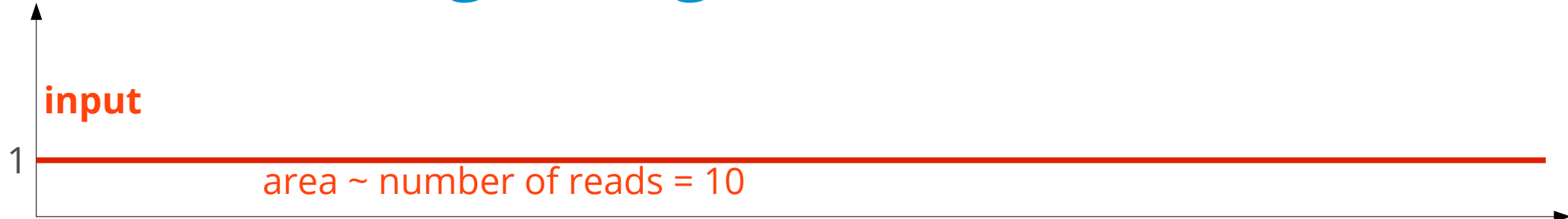
More signal (but equal noise) → artificial noise over-estimation

3. modelling background level

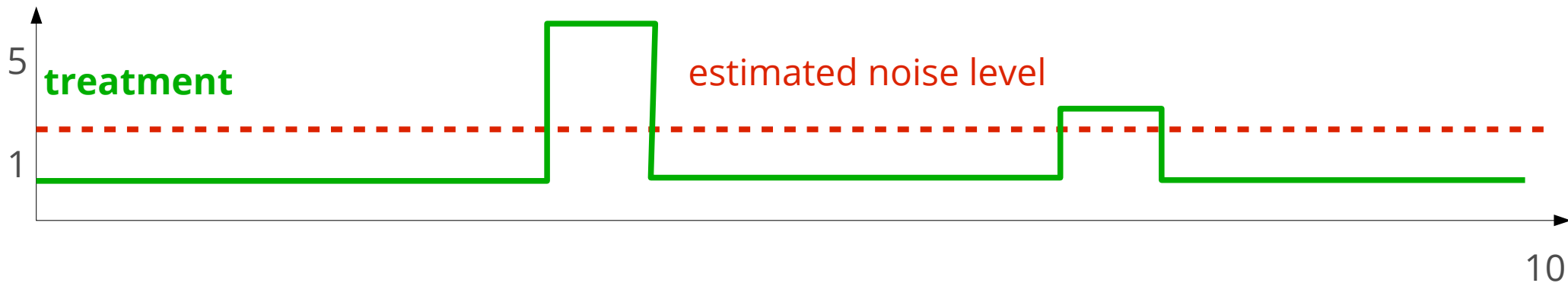


Noise level is over-estimated !

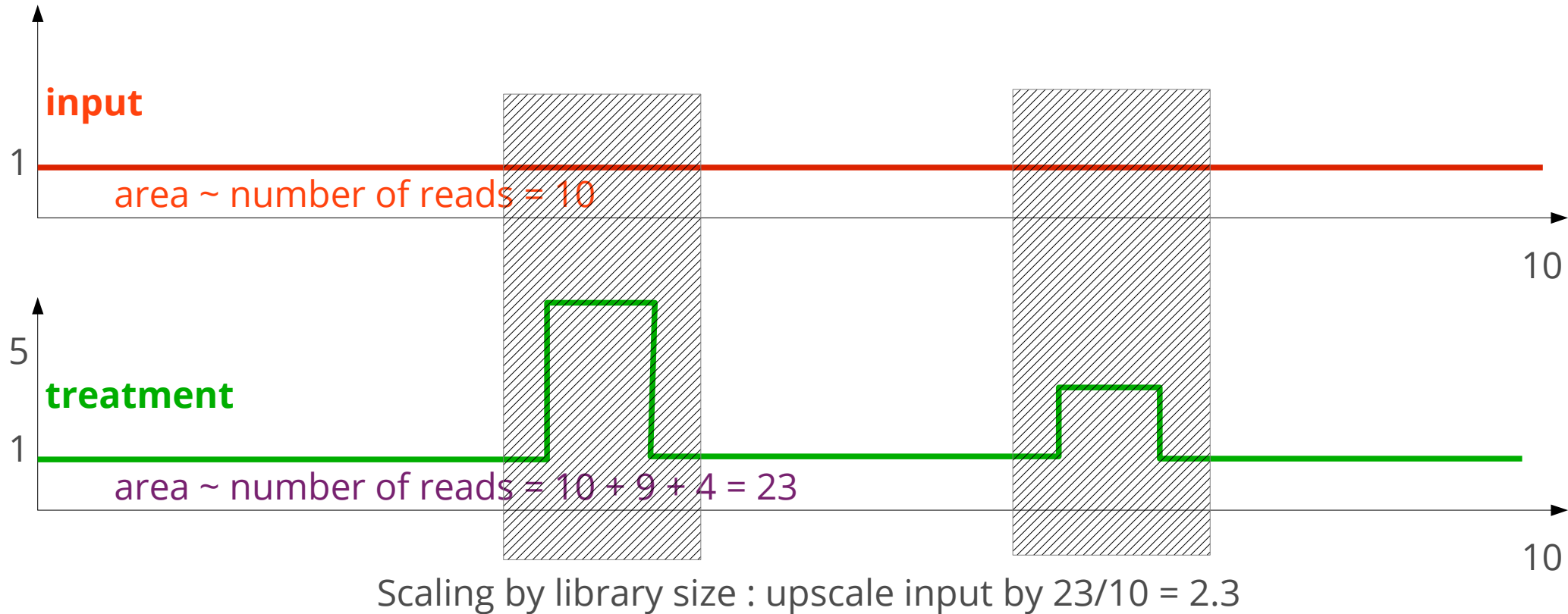
3. modelling background level



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

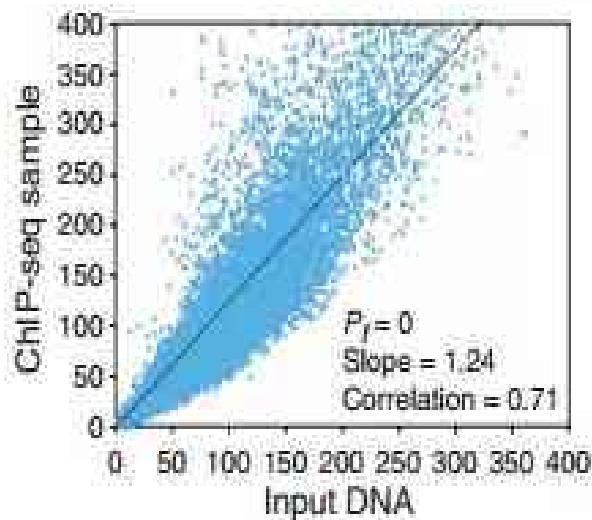


3. modelling background level

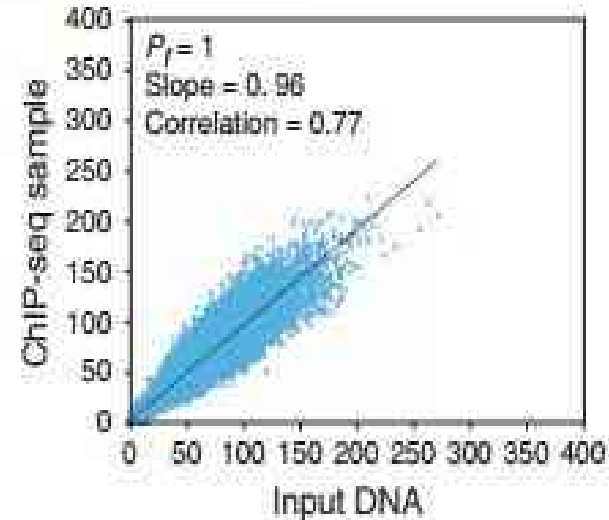


3. modelling background level

- **more advanced** : linear regression by excluding 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 Rozowsky¹, Ghia Euskirchen², Raymond K Auerbach³, Zhengdong D Zhang¹, Theodore Gibson¹, Robert Bjornson⁴, Nicholas Carrero⁴, Michael Snyder^{1,2} & Mark B Gerstein^{1,3,4}

3. modelling background level

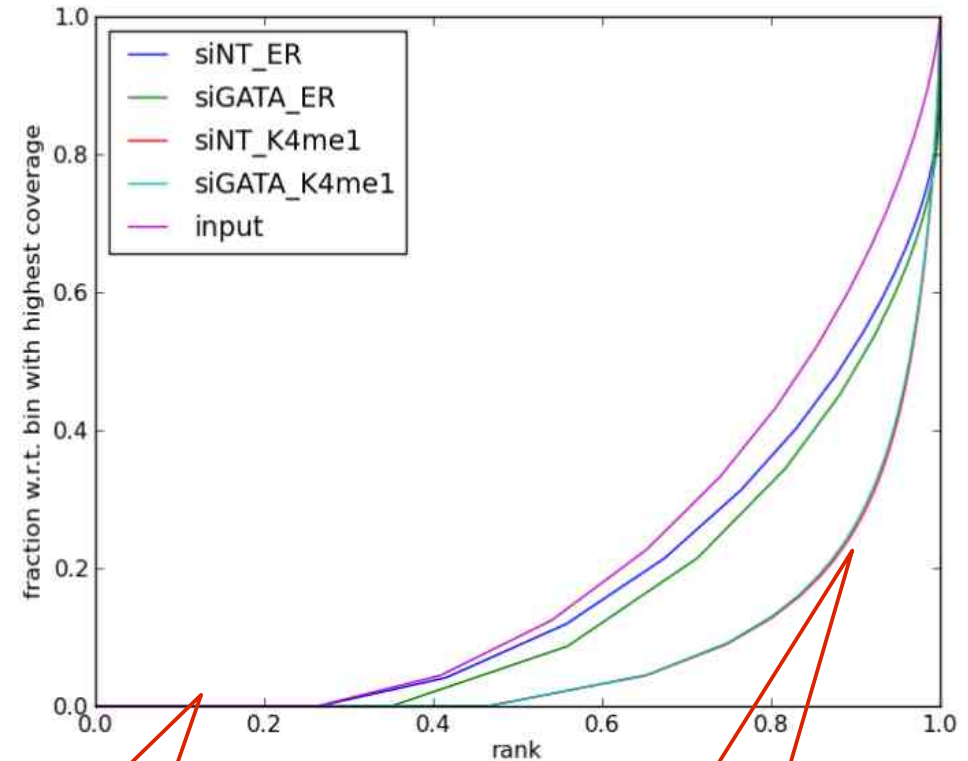
- **Alternative strategy**

(deepTools → Diaz et al.)

1. bin genome into n 10 kb windows
2. count reads in each window for input (X_i) and treatment (Y_i)
3. total number of reads is N_X and N_Y
4. order Y_i from less to most enriched → $Y_{(i)}$
5. define and plot

$$p_j = \sum_{i=1}^j Y_{(i)} / M_Y; q_j = \sum_{i=1}^j X_{(i)} / M_X$$

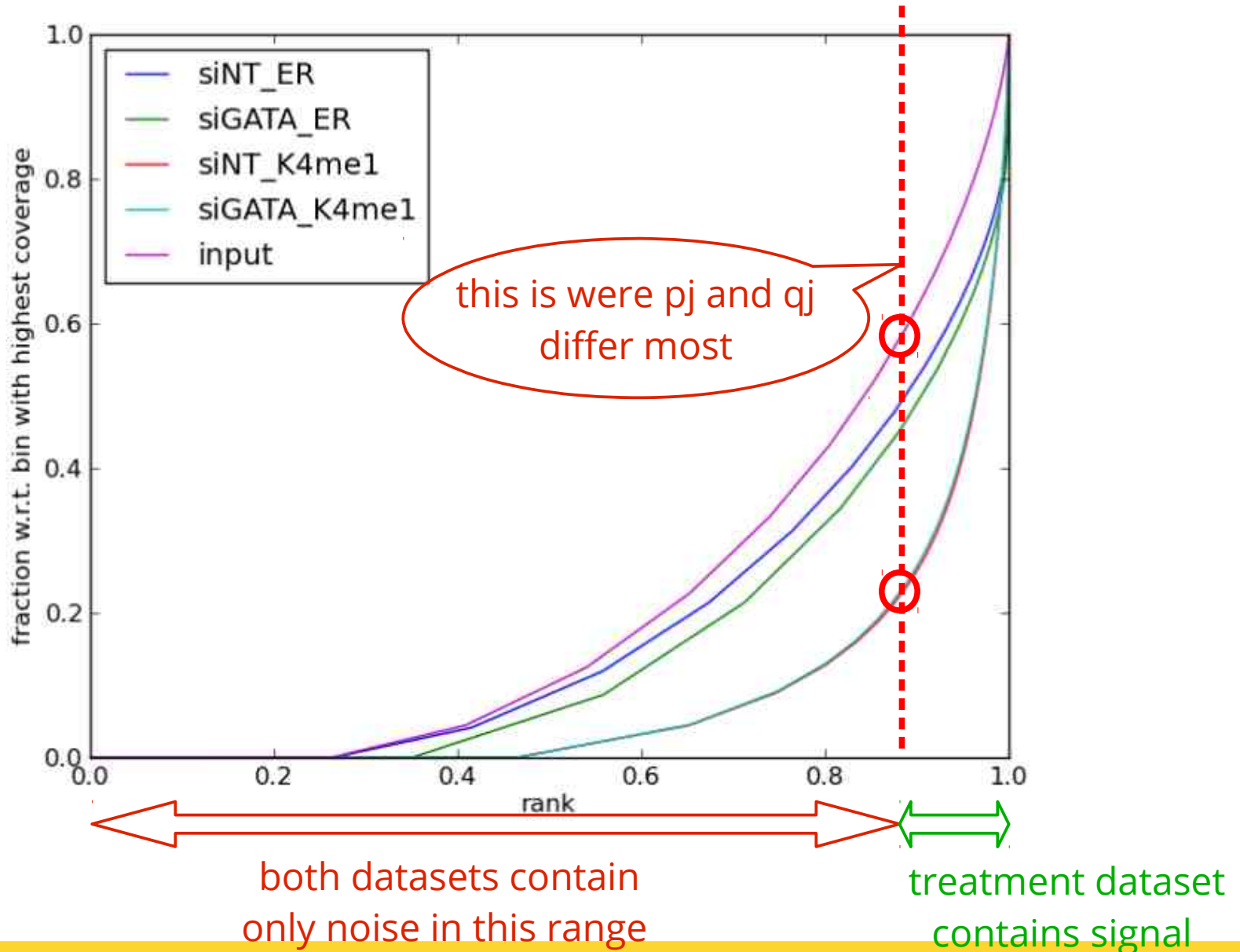
- p_j = proportion of reads in the j less enriched windows



25% of the genome contains no reads !

90% of the genome contain ~ 25% of reads

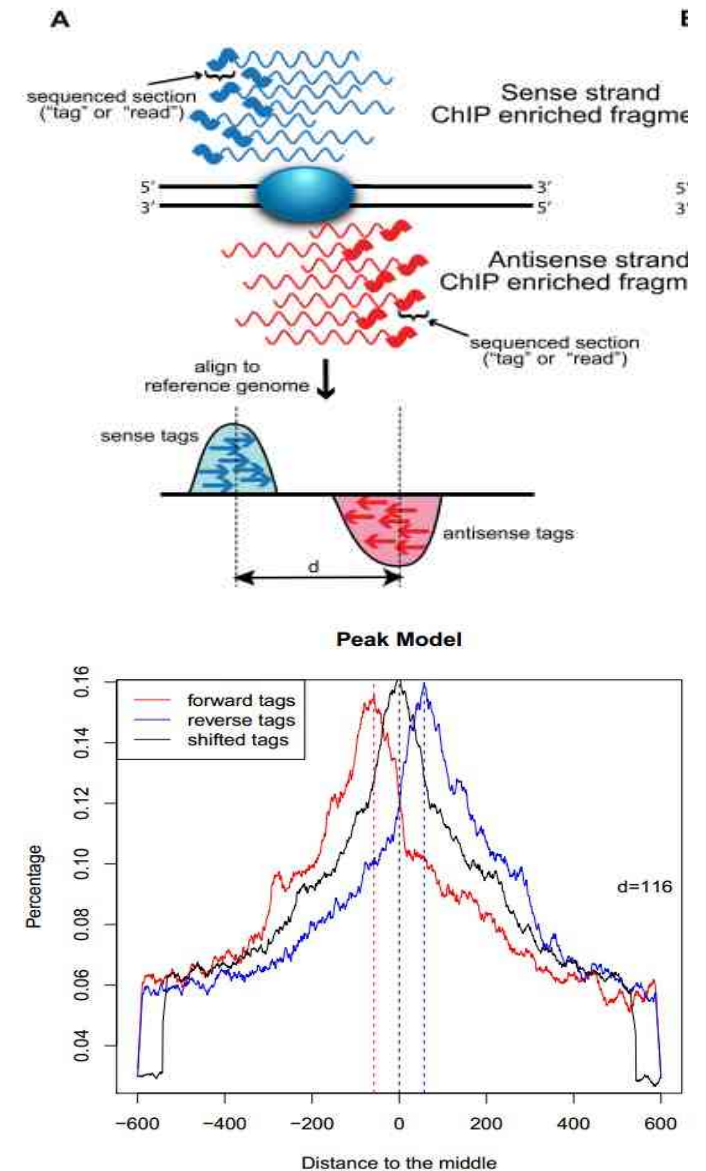
3. modelling background level



4. from reads to peaks

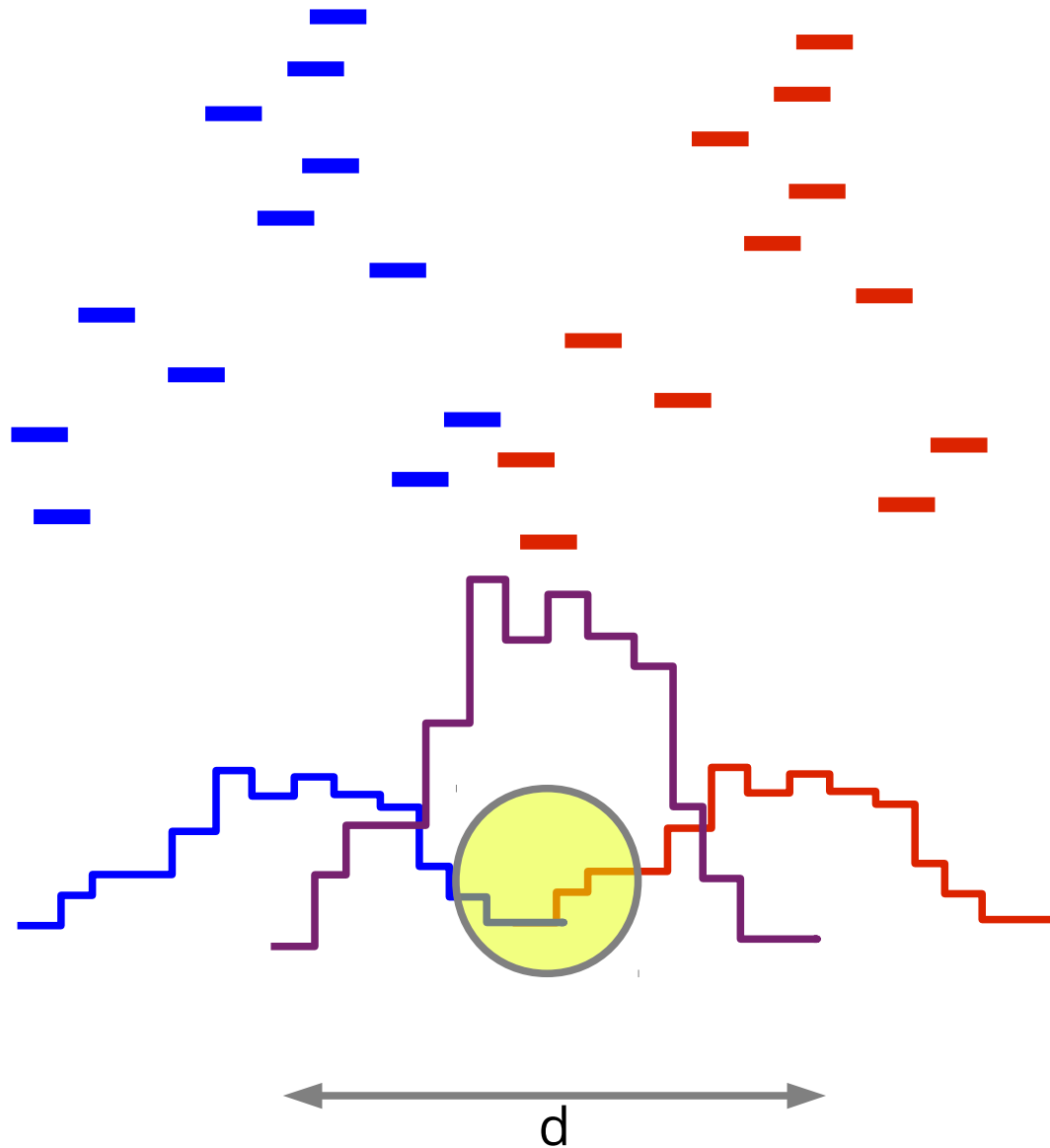
- **Tag shifting vs. extension**

- positive/negative strand read peaks do not represent the true location of the binding site
- fragment length is d and can be estimated from strand asymmetry
- reads can be **elongated** to a size of d
- reads can be **shifted** by $d/2$ → increased resolution



example of MACS model building
using top enriched regions

4. from reads to peaks



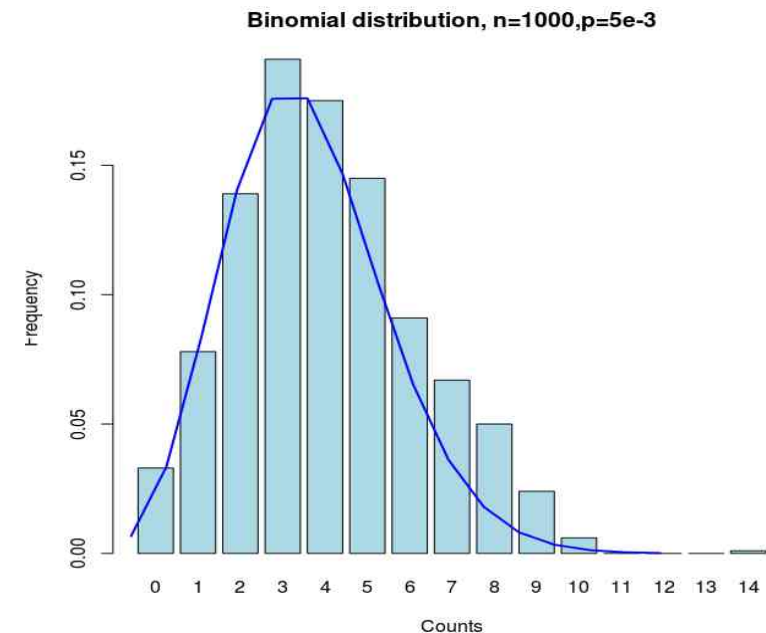
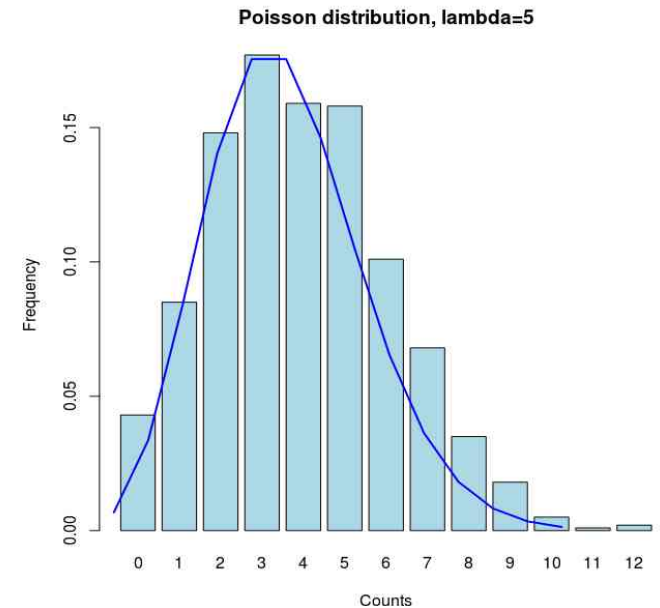
4. from reads to peaks

- 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

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

- **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

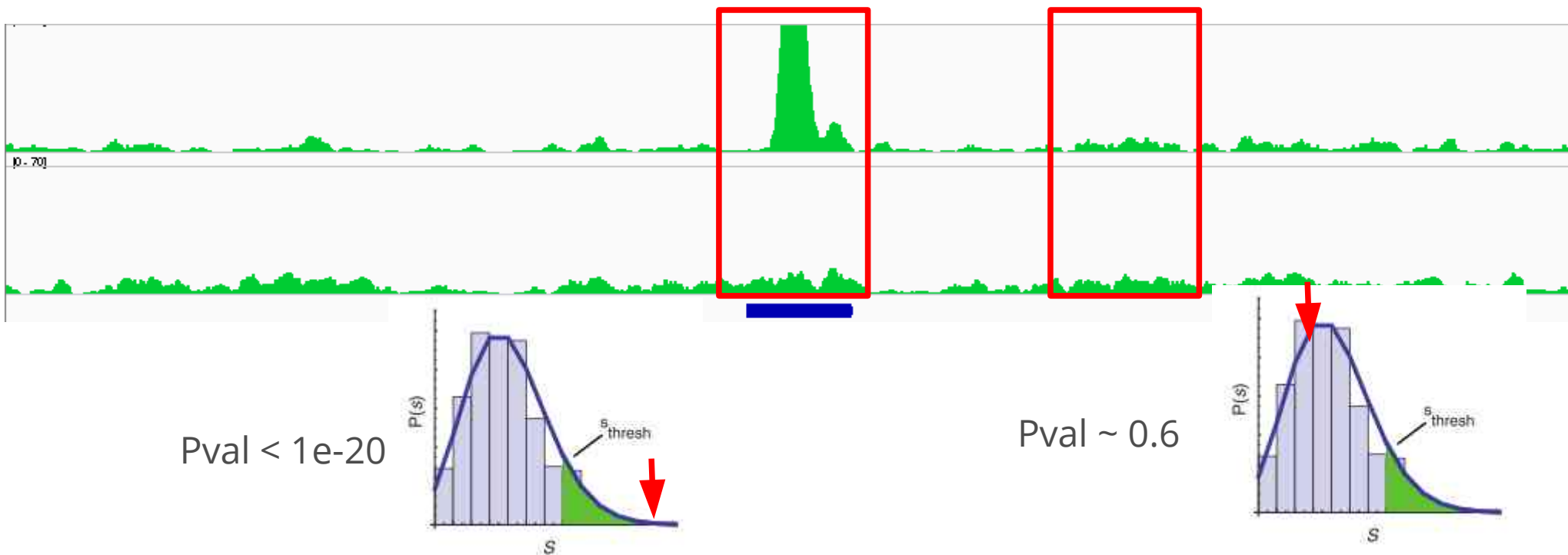
$$P(X=k) = C_n^k p^k (1-p)^{n-k}$$



4. from reads to peaks

- **Determining “enriched” regions**

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



	Profile	Peak criteria ^a	Tag shift	Control data ^b	Rank by	FDR ^c	User input parameters ^d	Strand-based filtering: 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 <i>P</i> values	<i>P</i> 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
GLTR	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 <i>P</i> value	Estimate from high quality peak pairs	Used for Poisson fit when available	<i>P</i> value	1: None 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	<i>P</i> -value threshold, tag length, mfold for shift estimate	No / Yes
PeakSeq	Extended tag aggregation	Local region binomial <i>P</i> value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	<i>q</i> 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	<i>q</i> 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	<i>P</i> value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and <i>P</i> values	<i>q</i> value	1: None 2: From Poisson <i>P</i> values	Window length, gap size, FDR (with control) or <i>F</i> -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 <i>P</i> value (paired peaks only)	Maximal strand cross-correlation					

Computation for ChIP-seq and RNA-seq studies

Shirley Pepke¹, Barbara Wold² & Ali Mortazavi²

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)
----------------	---------------------------------------

GLTR	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)	\pm 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

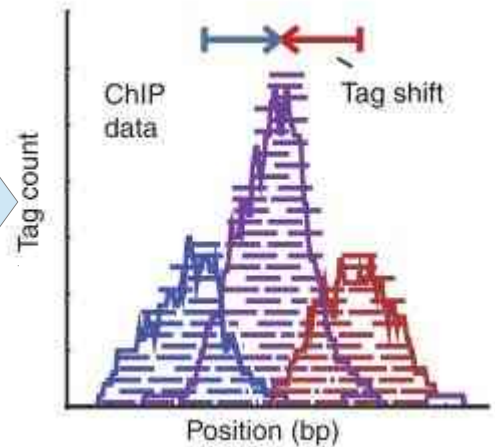
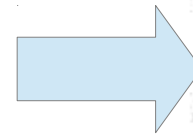
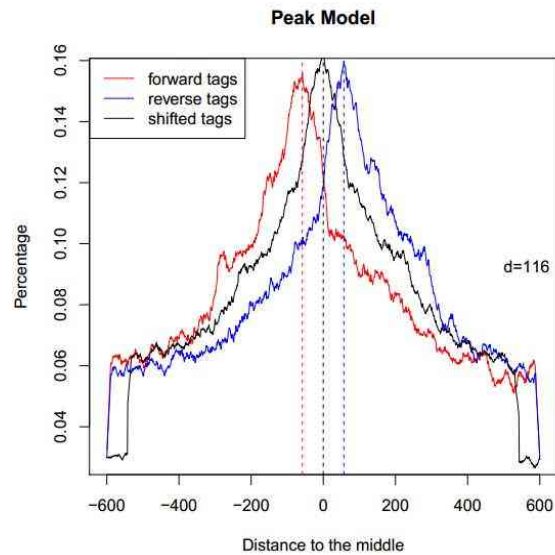
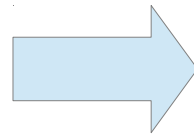
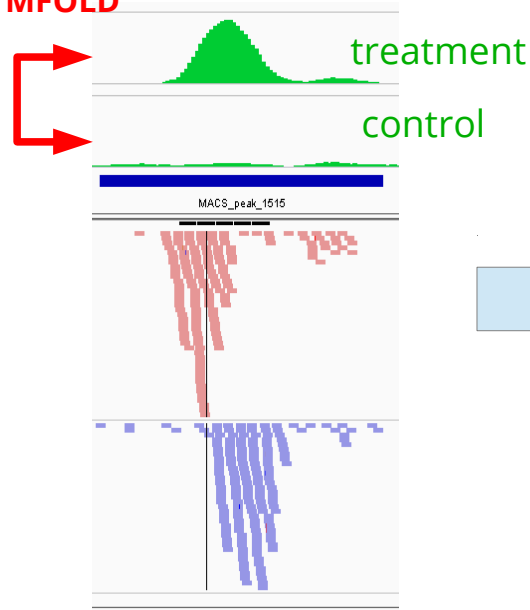
Tags unchanged

5. 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

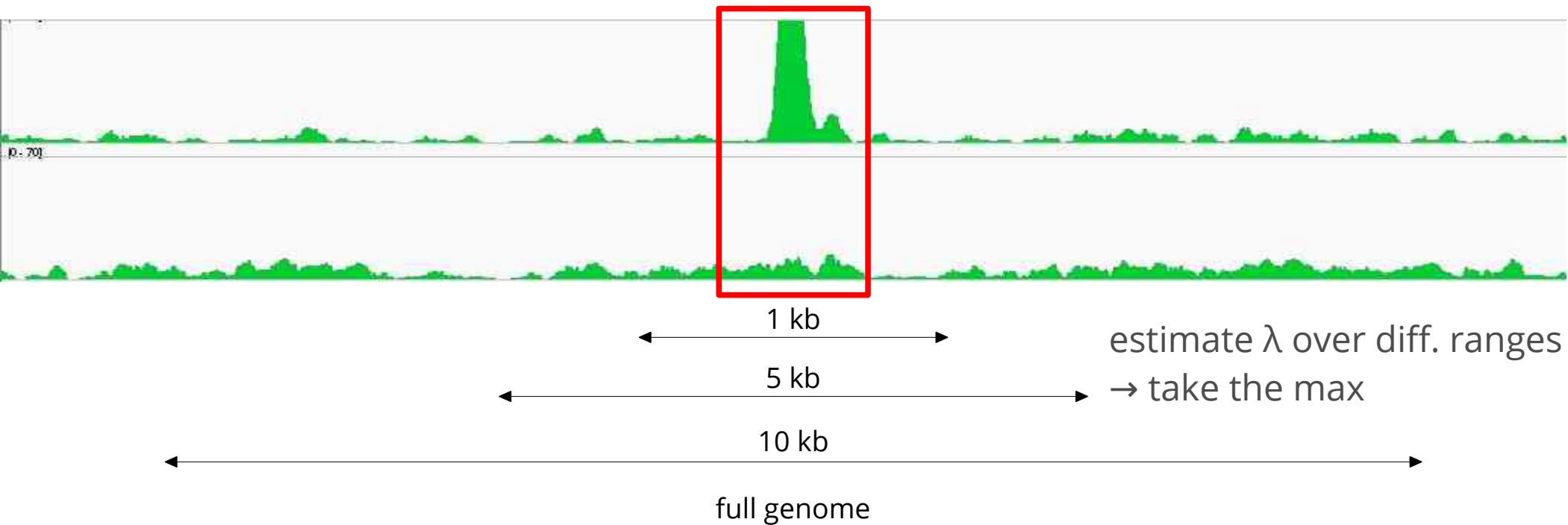
enrichment
> MFOLD



5. 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

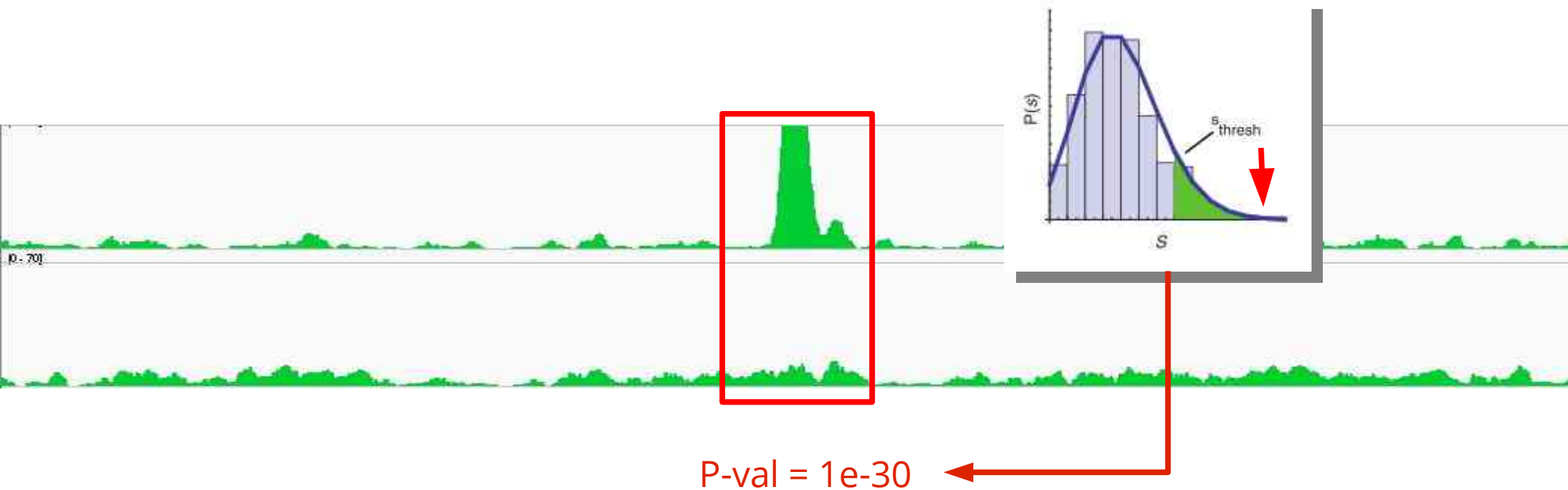


5. MACS

[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



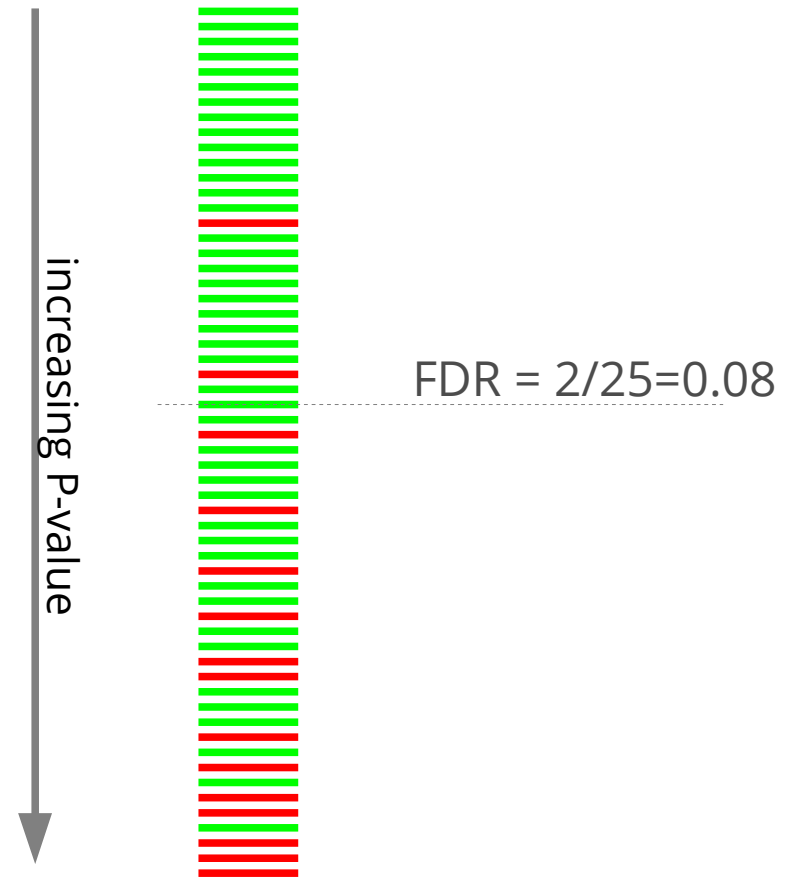
5. MACS

[Zhang et al. Genome Biol. 2008]

- **Step 4 : estimating FDR**

- positive peaks (P-values)
- swap treatment and input; call negative peaks (P-value)

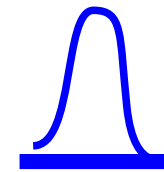
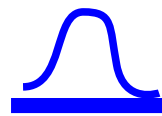
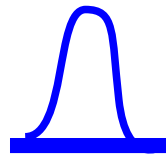
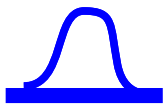
$$\text{FDR}(p) = \frac{\# \text{negative peaks with } P_{\text{val}} < p}{\# \text{positive peaks with } P_{\text{val}} < p}$$



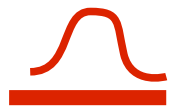
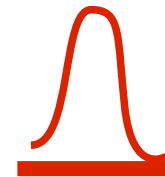
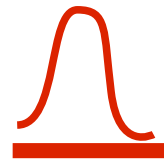
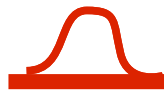
6. differential analysis

- given ChIP-set datasets in different conditions, we want to find **differential binding events** between 2 conditions
 - binding vs. no binding → qualitative analysis
 - weak binding vs. strong binding → quantitative analysis

Condition A



Condition B



binding in A
no binding in B

stronger
binding
in A

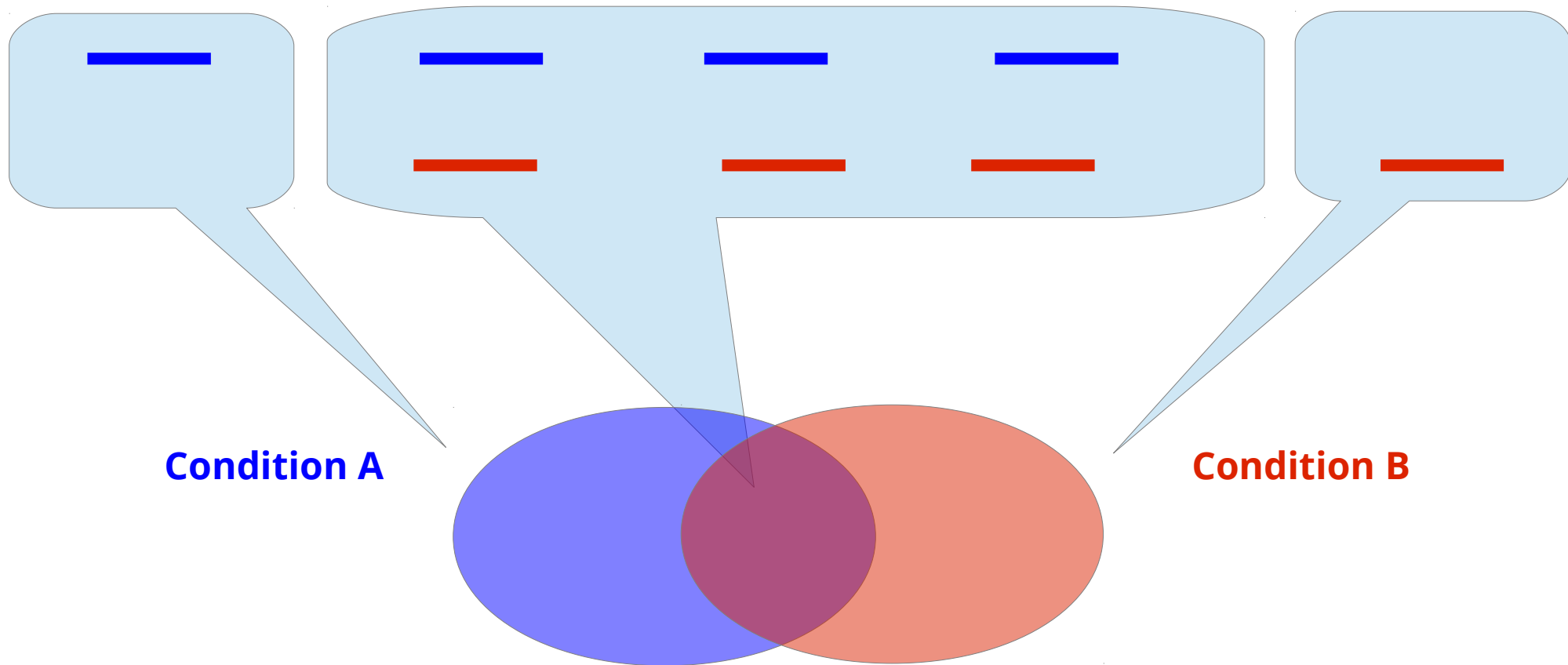
stronger
binding
in B

no difference

binding in B
no binding in A

6. differential analysis

- simple approach → compute common and specific peaks



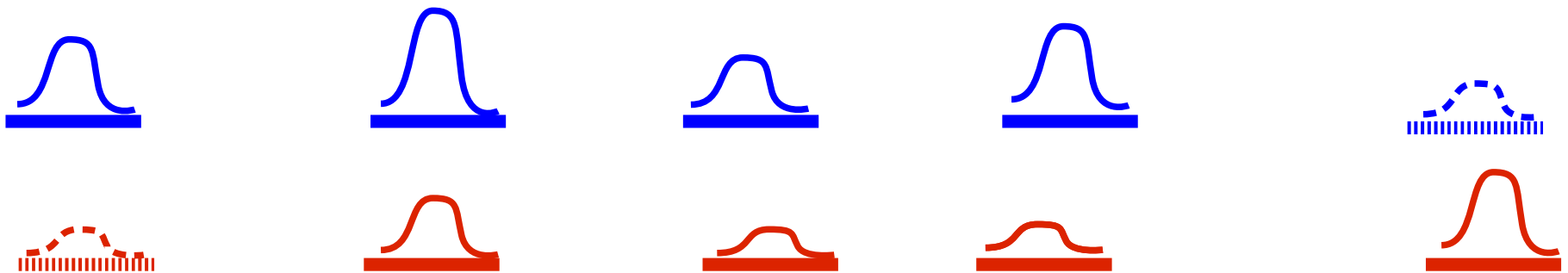
Drawback :

- common peaks can hide **differences in binding intensities**
- **specific peaks** can result from **threshold issues**

6. differential analysis

- **quantitative approach**

- select regions which have signal (union of all peaks)
- in these regions, perform quantitative analysis of differential binding based on **read counts**



- statistical model

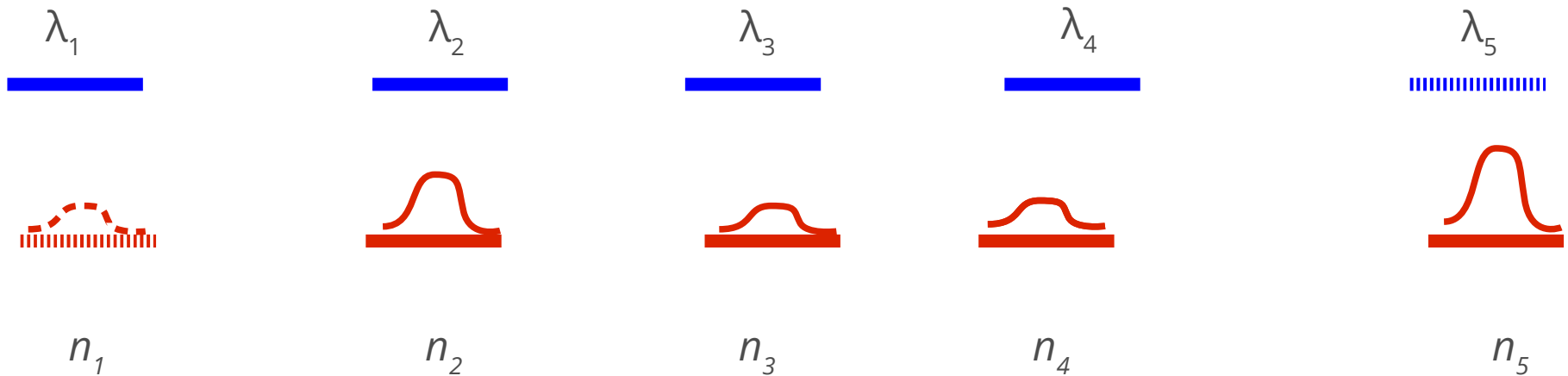
- **without replicates** : assume simple Poisson model (→ SICER-df)
- **with replicates** : perform differential test using DE tools from RNA-seq (diffBind using EdgeR, DESeq,...) based on read counts

6. differential analysis

- **without replicates (sicer-df)**

- consider one condition to be the reference (condition A)
- call peaks on each condition independently
- take union of peaks
- assume Poisson model based on expected number of reads in region
- compute P-value, log(fold-change)

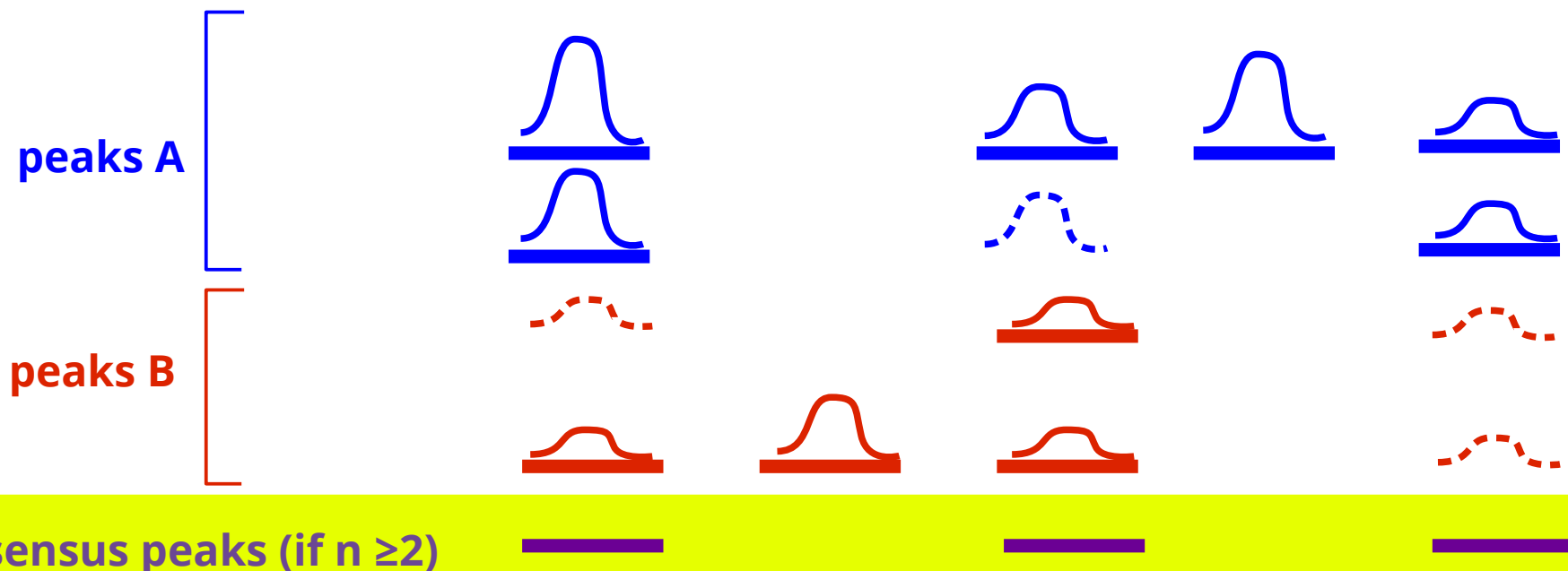
$$\lambda_i = w_i N_A / L_{eff}$$



6. differential analysis

- **with replicates (diffBind)**

- provide list of peaks for replicates A and replicates B
- determine consensus peakset based on presence in at least n datasets
- compute read counts in each consensus peak in each dataset
- run DESeq / EdgeR to determine differential peaks between condition A and B (negative binomial model, variance estimated on replicates)



Program of the Practical Session

Step 0 : Find datasets on Gene Expression Omnibus

Step 1 : Import datasets into your Galaxy history

Step 2 : data inspection : coverage plots, correlation,...

Step 3 : peak calling using MACS

Step 5 : differential analysis

Step 6 : visualizing results in IGV