Bioinformatics Workflow

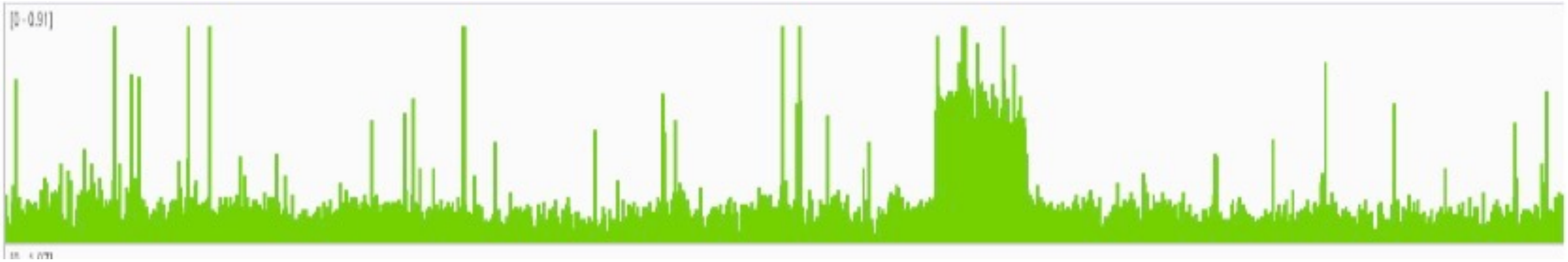
- From aligned reads to peaks -

Carl Herrmann

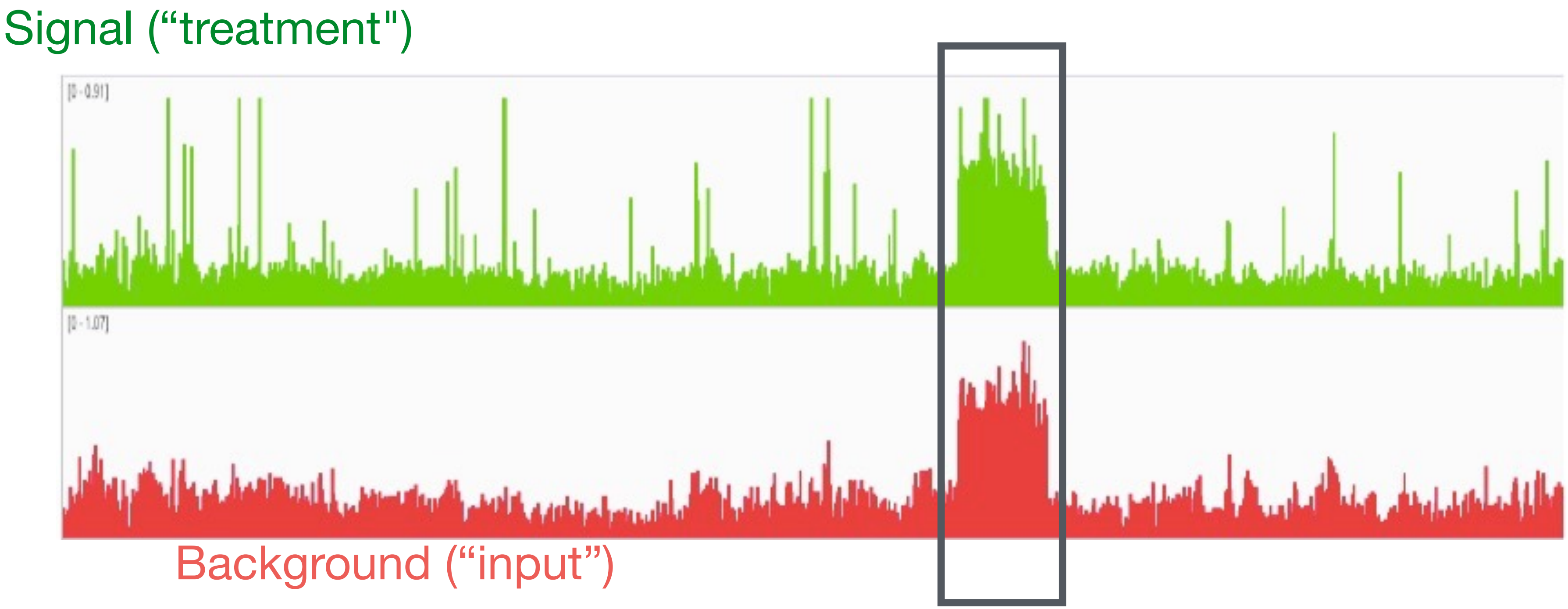
Peak calling: where is the signal?

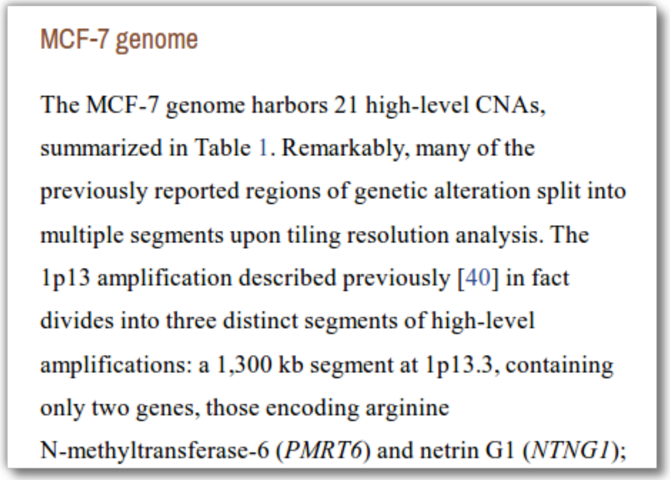
Medizinische Fakultät Heidelberg

Signal (“treatment")



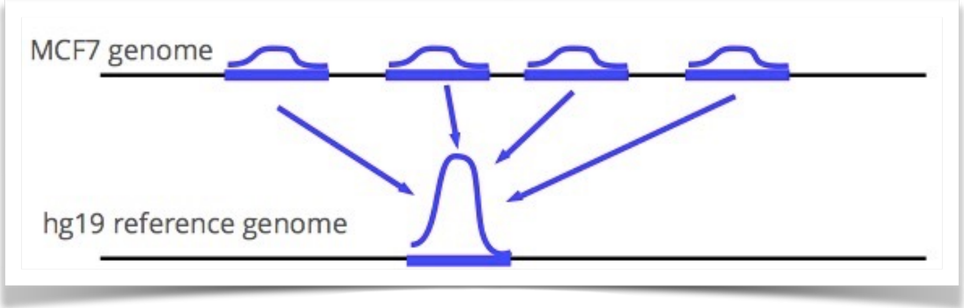
Peak calling: where is the signal?





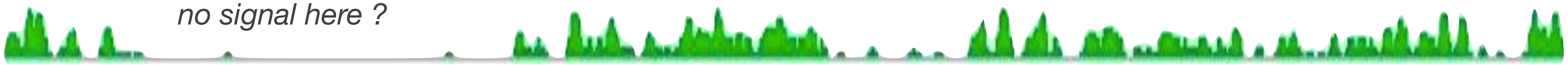
MCF-7 genome

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*);

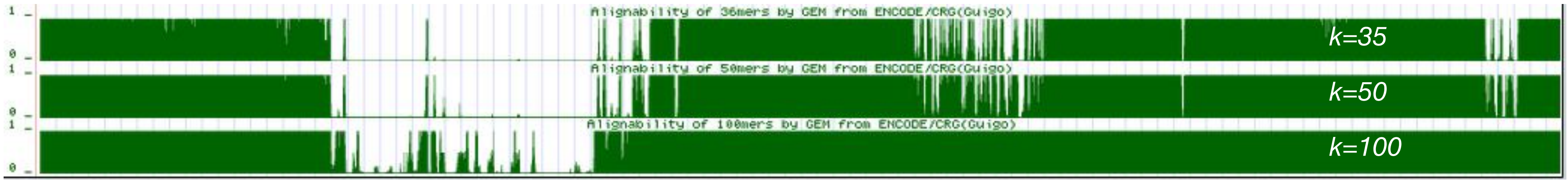


Peak calling: where is the signal?

Signal (“treatment")



Alignability track


1
0
1
0
1
0
Alignability of 36mers by GEM from ENCODE/CRG(Guigo)
Alignability of 50mers by GEM from ENCODE/CRG(Guigo)
Alignability of 100mers by GEM from ENCODE/CRG(Guigo)
k=35
k=50
k=100

- **mappability issue**: alignability track shows, how many times a read from a given position in the genome would align
  - a = 1: read from this genomic locus would ONLY align to this position
  - a = 1/n : read from this position could align to n alternative positions in the genome
- usually only unambiguous reads are kept in the alignment : positions with a < 1 contain no reads at all !

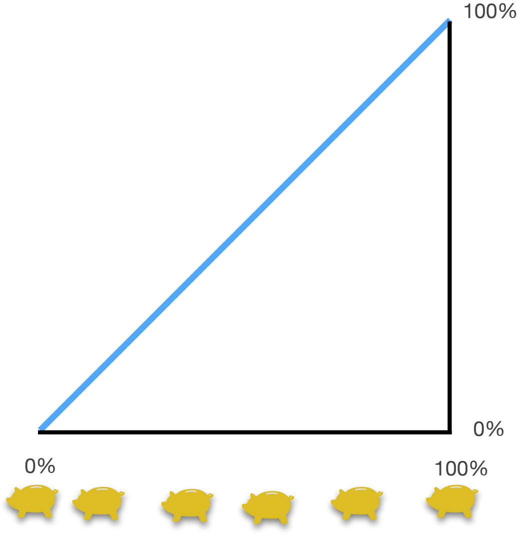
Peak calling: where is the signal?

Availability of a control sample is mandatory !!

→ mock IP with unspecific antibody
→ sequencing of input (=naked) DNA

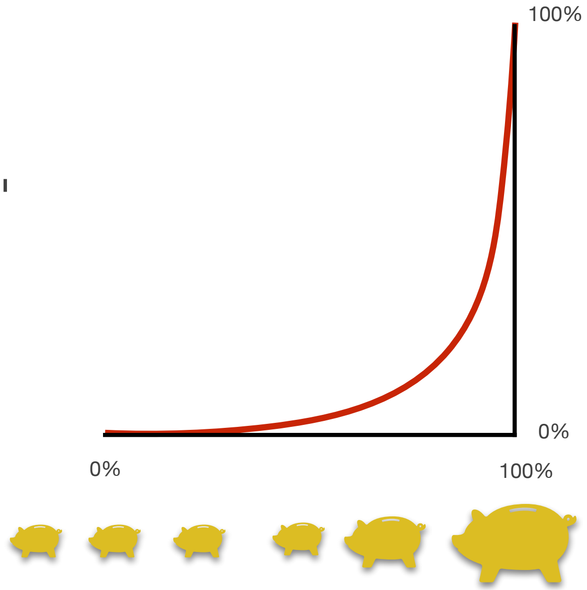
Signal to noise ratio

- Good ChIP-seq experiment:
  - **high** enrichment of signal in **few regions**
  - **low/no** enrichment in **most regions**
- Test this unequal distribution using a Lorenz Curve: cumulative distribution of the signal



even distribution: cumulative curve is a straight line
Good for society / Bad for ChIP-seq

"fingerprint plot"



unequal distribution: cumulative curve has sharp kink
Bad for society / Good for ChIP-seq

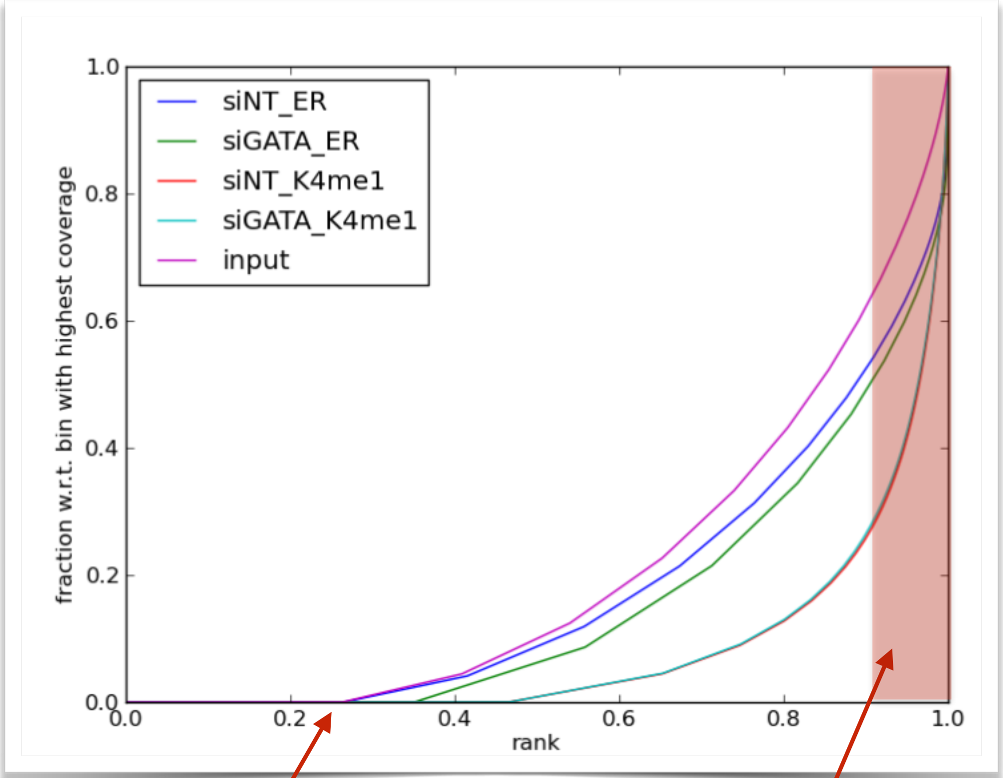
[Diaz et al., deepTools]

Signal to noise ratio

- **Procedure**
  - bin genome into 10 kb regions
  - count reads in each bin from input (Xi) and signal (Yi)
  - total number of reads is Mx and My
  - order Xi and Yi from smallest to largest → X(i) Y(i)
  - plot:

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

  - The more diagonal, the more uniform the signal is (input, bad chip)
  - The more bent, the more focal the signal (good chip)



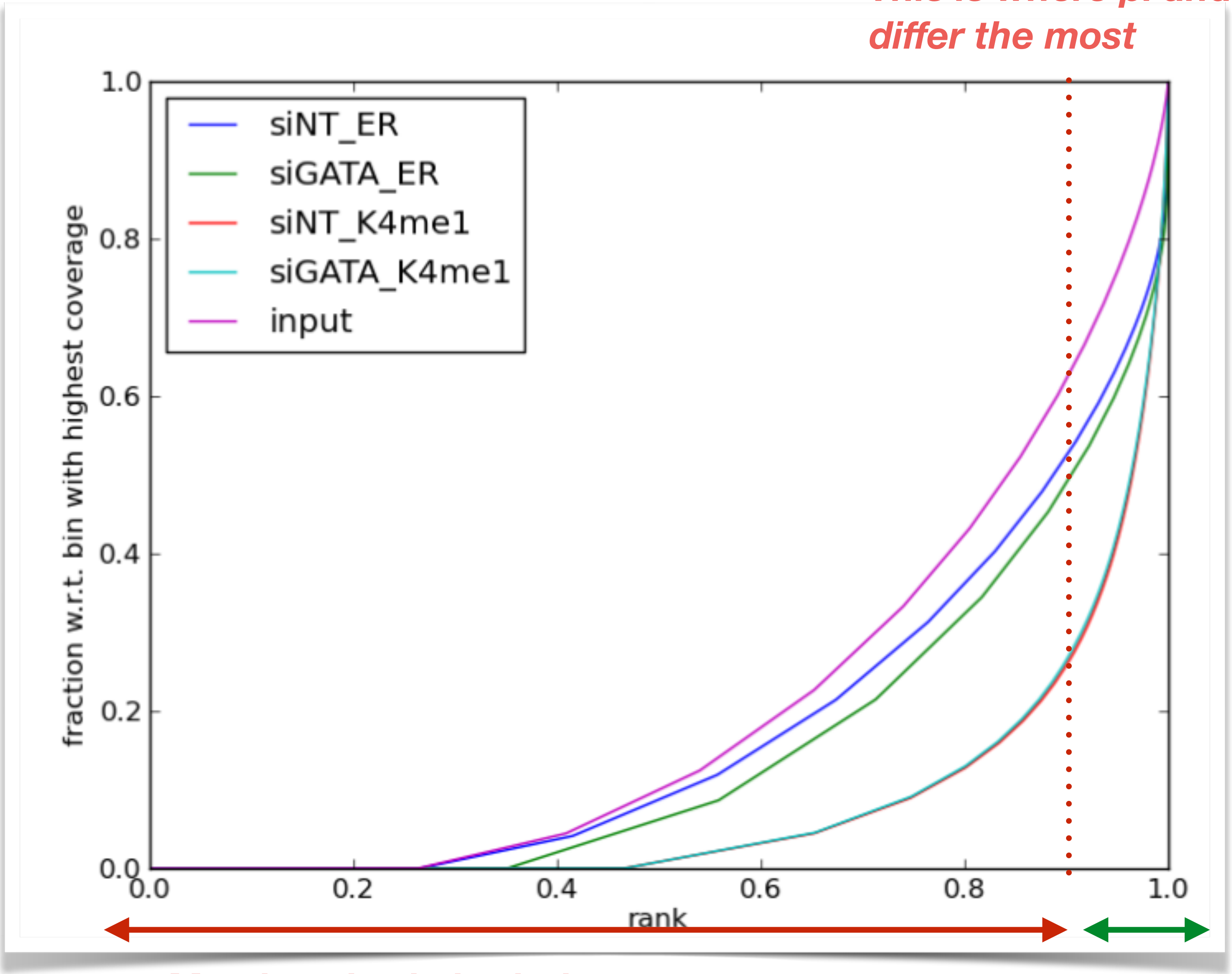
25% of the genome contain no reads

10% of the genome contain 75% of reads

[Diaz et al., deepTools]

Signal to noise ratio

This is where pi and qi differ the most

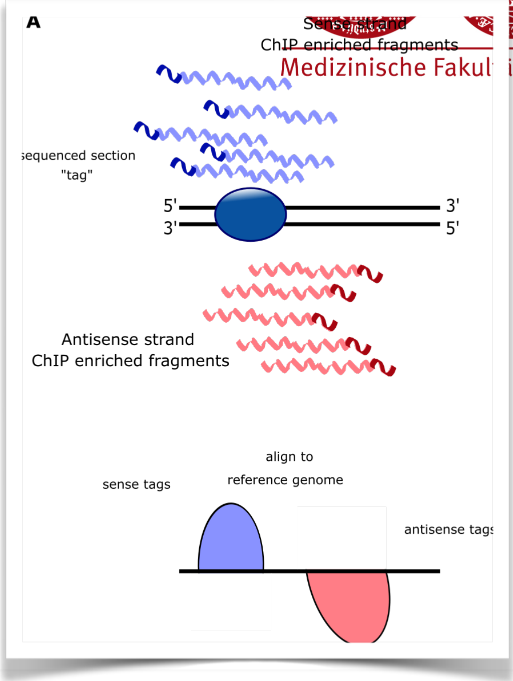


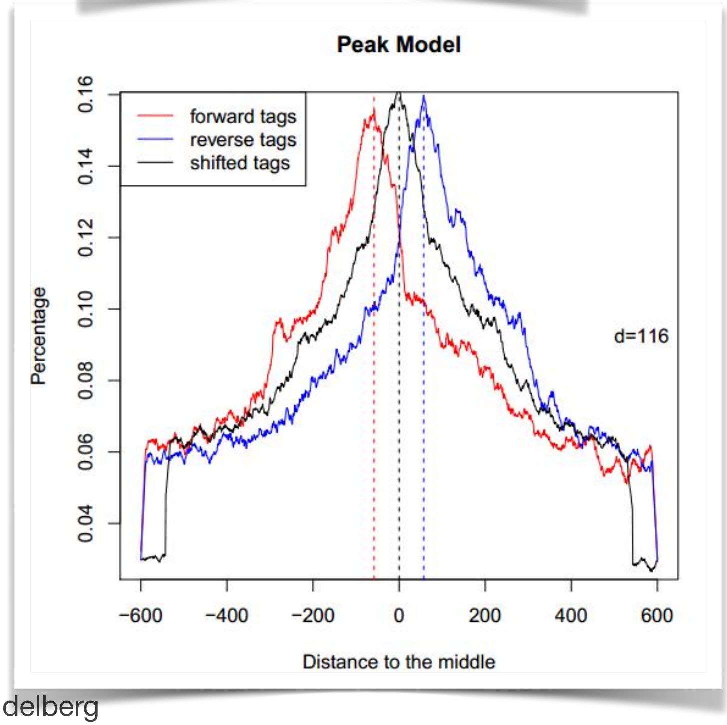
Mostly noise in both datasets → scaling factor using the reads in this range

Treatment contains signal !

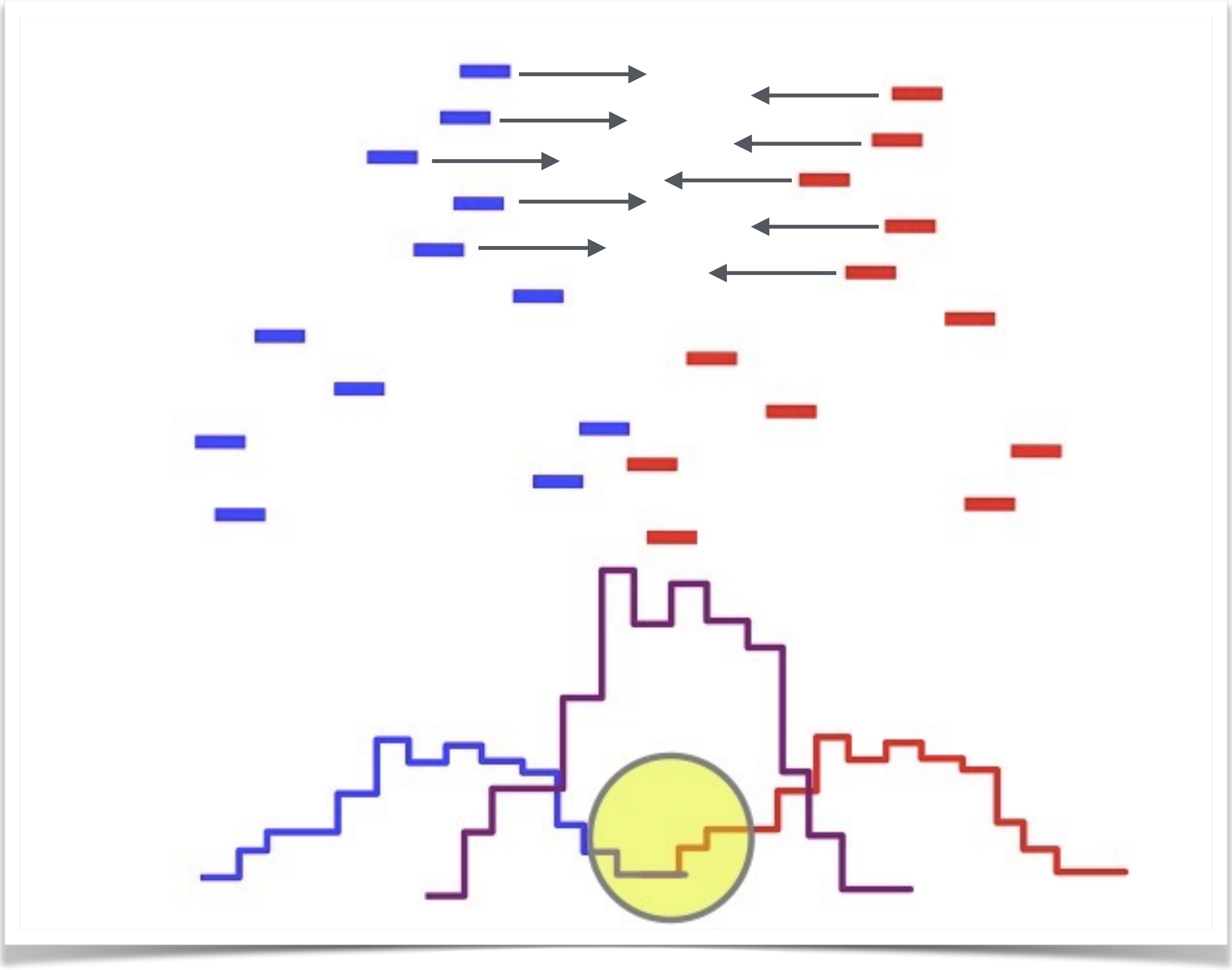
Peak calling (single-end)

- Tag shifting vs tag extension (single-end)
  - read locations do not represent the actual binding site
  - fragment length *d* can be estimated from strand asymmetry
  - reads can be elongated to a size of *d*
  - or: reads can be shifted by length *d/2*



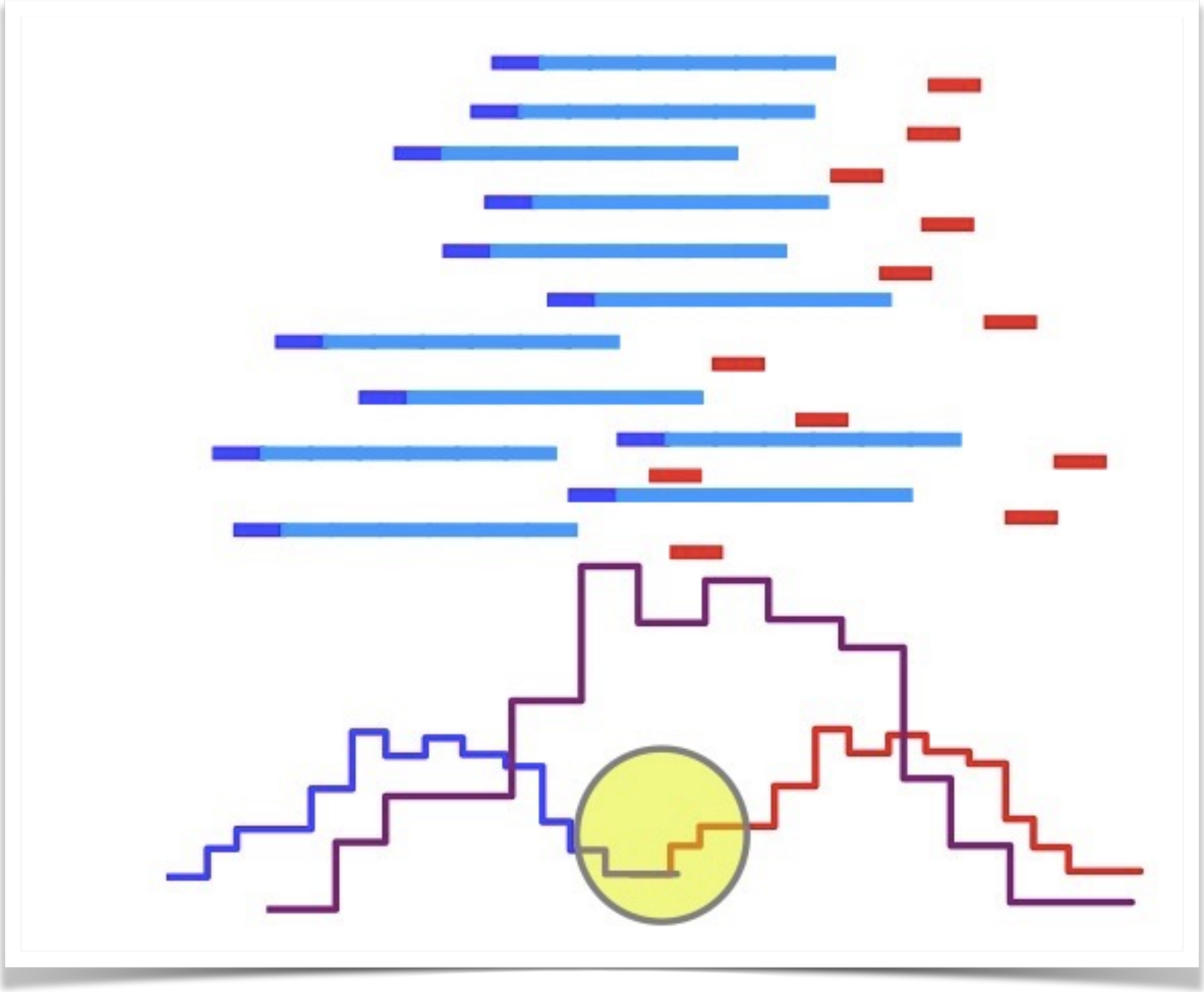


Read shifting



All reads are shifted by *d*/2

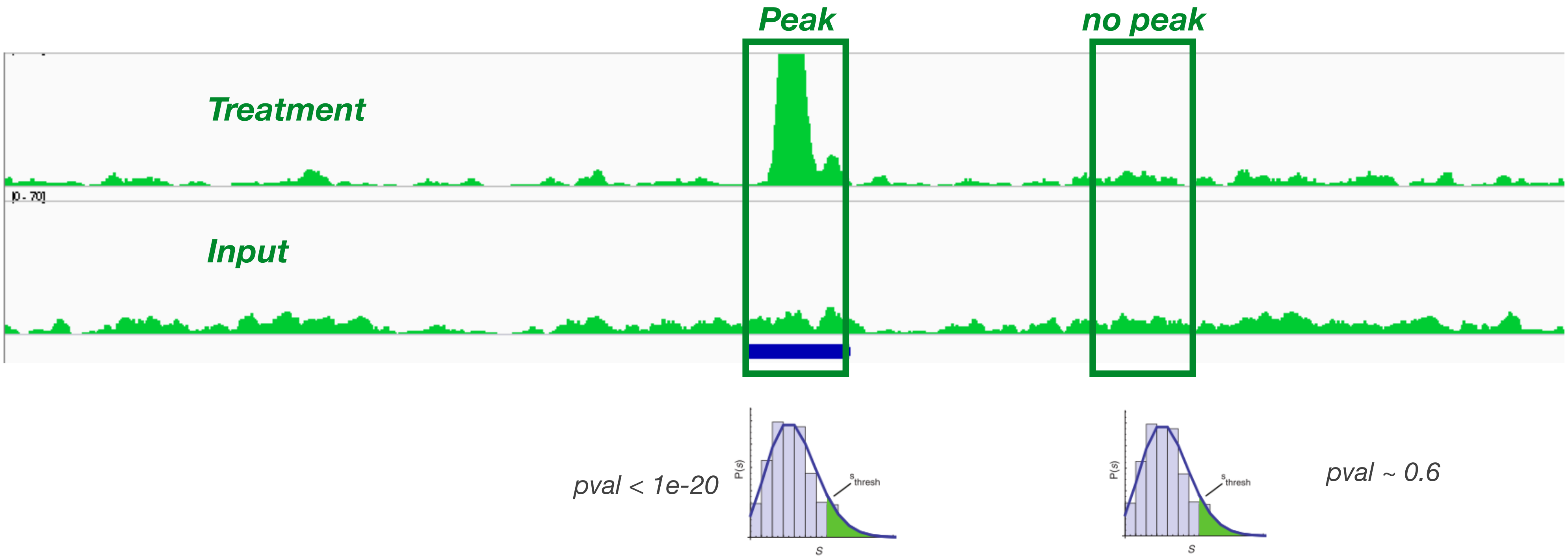
Read extension



All reads are extended to length *d*

Statistical model

- Goal : determine enriched regions
- sliding window across the genome
- at each location, determine the enrichment of the signal vs. background using a Poisson distribution to model expectation
- retain regions below P-value threshold



Reminder: Poisson distribution

- Measures the number of events in a time period for a given **rate** of events
- *Example: number of reads aligning randonly in a region of size w*
- Rate needs to be constant and independent of previous window!

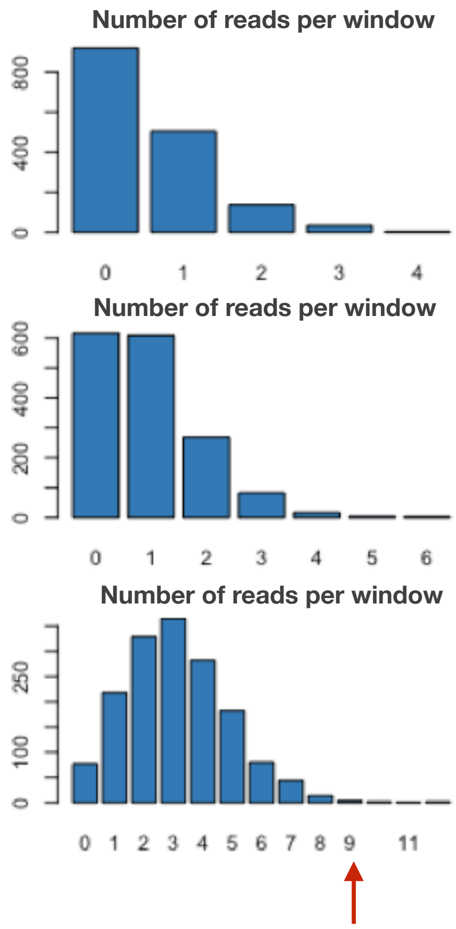
$$\boxed{\begin{aligned} p(k) &= \frac{\lambda^k}{k!}e^{-\lambda} \\ E(X) &= \lambda \\ Var(X) &= \lambda \end{aligned}}$$

λ = average reads per window

$\lambda = 0.56$

$\lambda = 0.94$

$\lambda = 3.125$

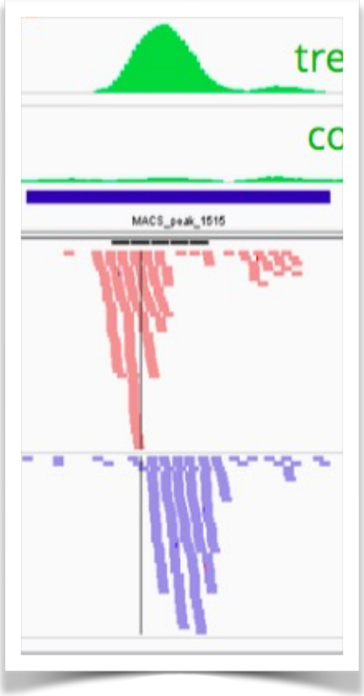


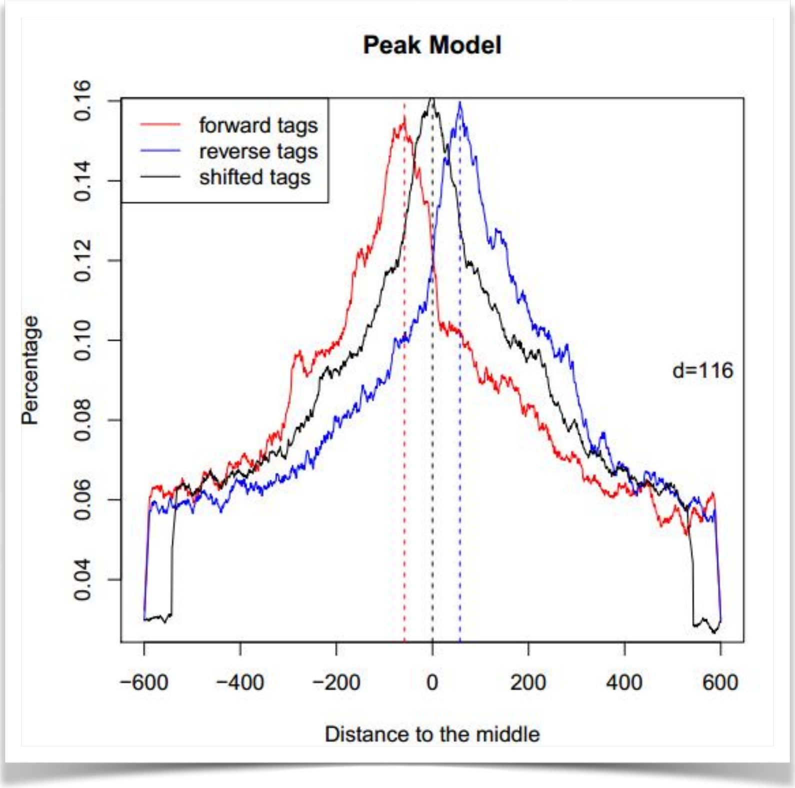
Probability of 9 or more: 0.0015

Example : MACS2

- **Step 1 : estimate fragment length *d***
  - slide a window of length BANDWIDTH
  - retain windows with MFOLD enrichment of treatment / background
  - plot average + / - strand read densities in these windows
  - estimate d

> MFOLD enrichment

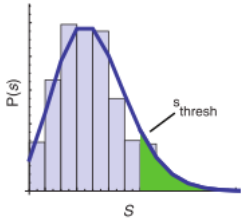


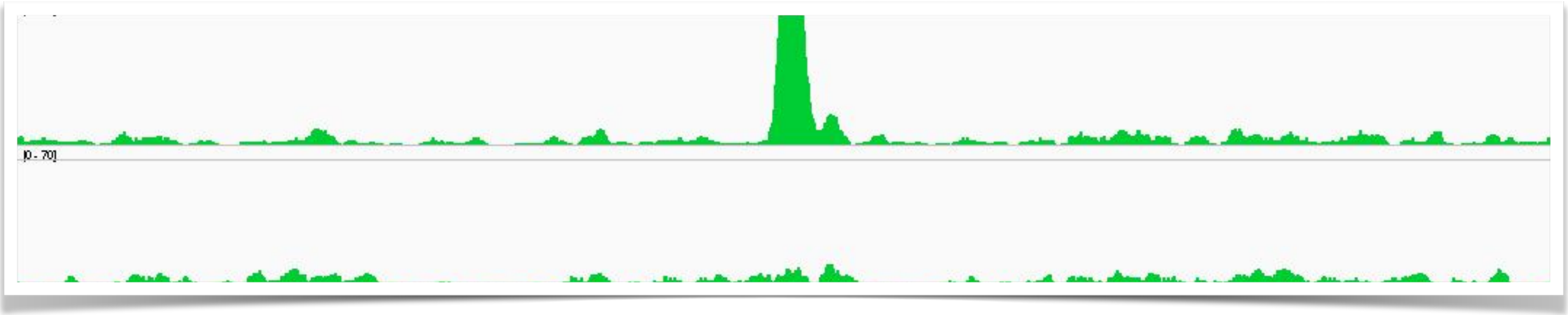


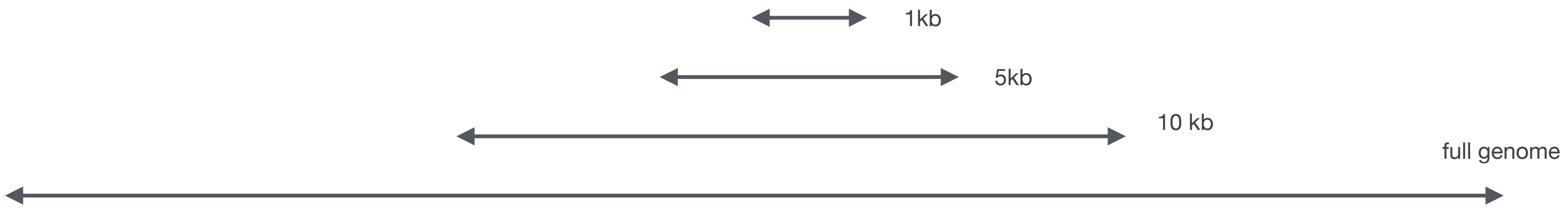
[Zhang et al. , 2008]

Example : MACS2

- **Step 2** : **identification of local noise parameter**
  - slide a window of size 2**d* across treatment and input
  - at each position, estimate parameter λ_{local} (= mean number of read per kb) of **Poisson distribution**







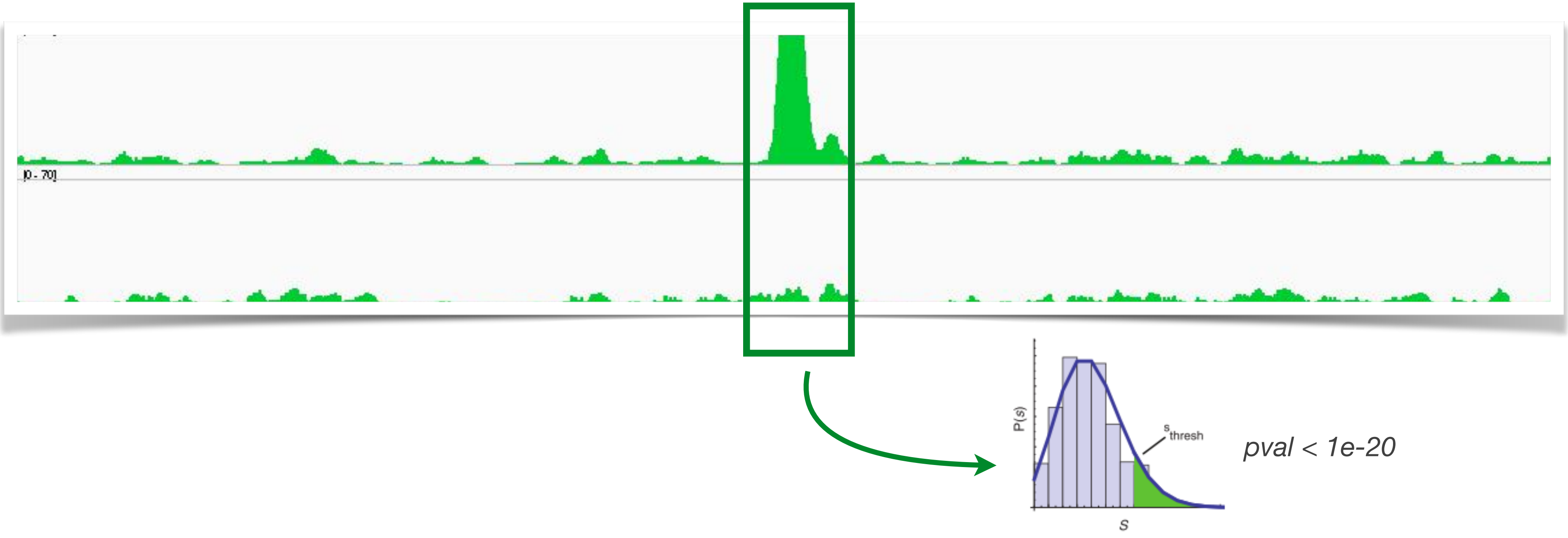
estimate parameter λ_{local} over different ranges, take max.

[Zhang et al. , 2008]

Example : MACS2

- **Step 3 : identification of enriched/peak regions**

  - determine regions with P-values < PVALUE
  - determine **summit position** inside enriched regions as max density



pval < 1e-20

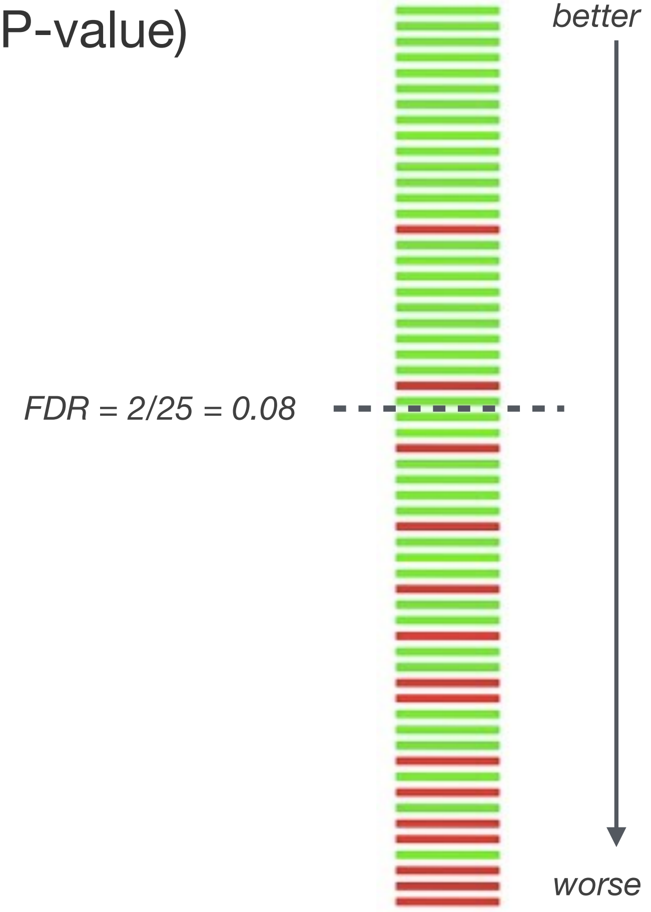
[Zhang et al. , 2008]

Example : MACS2

Medizinische Fakultät Heidelberg

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

$$\text{FDR} = \frac{\text{\# negative peaks with pval} < p}{\text{\# positive peaks with pval} < p}$$



[Zhang et al. , 2008]

Peak calling

- MACS2: typical command

```
macs2 callpeak \
--treatment IP.bam \
--control input.bam \
--name CTCF \
--format BAM \
--keep-dup all \
--gsize 2.7e9 \
--qvalue 0.01 \
--outdir CTCF
```

bam file with IP

bam file with input

name of the experiment (choose freely!)

format of input files (BAM = single-end; BAMPE = paired-end)

should duplicate read be kept? (auto / all)

effective (= mappable) genome size

FDR threshold to call a peak

output directory

Hands on:
ChIP-seq peak calling with MACS2

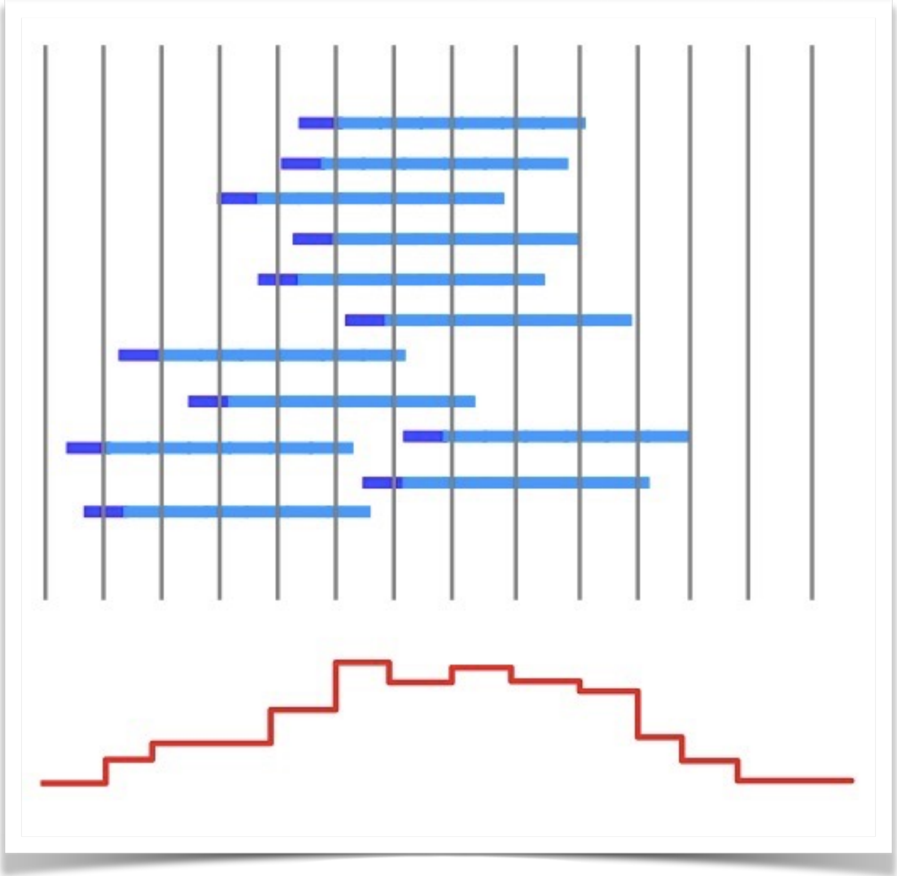
https://hdsu-bioquant.github.io/chipatac2020/05_CHIP_PeakCalling.html

Bioinformatics Workflow
- From aligned reads to signal tracks -

From reads to signal

Medizinische Fakultät Heidelberg

Single-end sequencing



- Reads are extended to 3’ to the estimated/provided fragment length
- Read counts are computed for each bin
- Counts are normalized
  - RPGC: reads per genomic content (normalize to 1x coverage)
  - RPKM: reads per kilobase per million reads per bin
- Tool :

```
bedtools genomecov
or: bamCoverage
```

From reads to signal

```
bamCoverage \
--bam CTCF.bam \
--outFileName CTCF.bw \
--outFileFormat bigwig \
--normalizeUsing RPKM \
--ignoreDuplicates \
--centerReads \
--binSize 200 \
--numberOfProcessors 4
```

output should be in bedgraph format

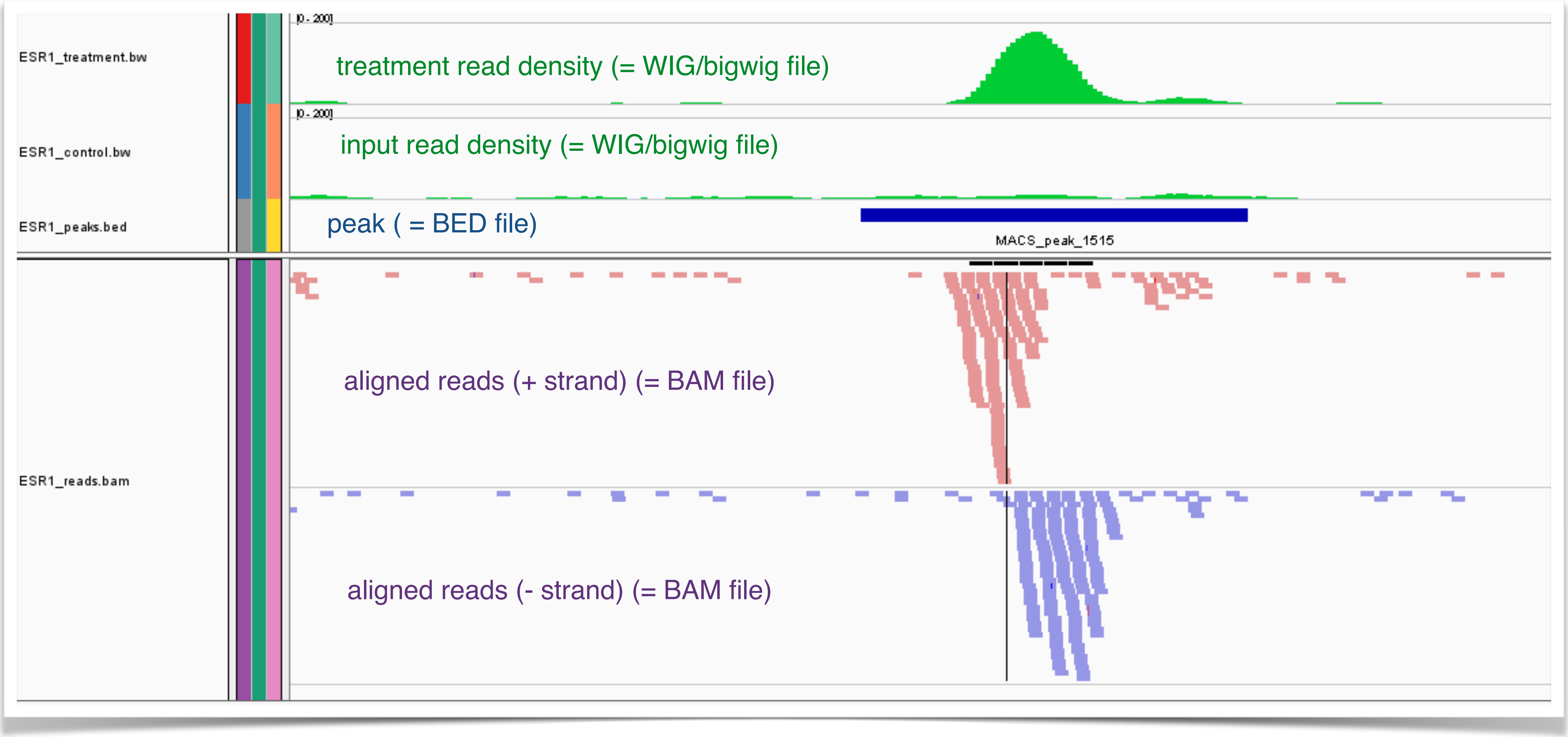
input bam file

fragment extension to 200bp

sort by chromosome and start coordinate; write to output file

Resolution = 200bp

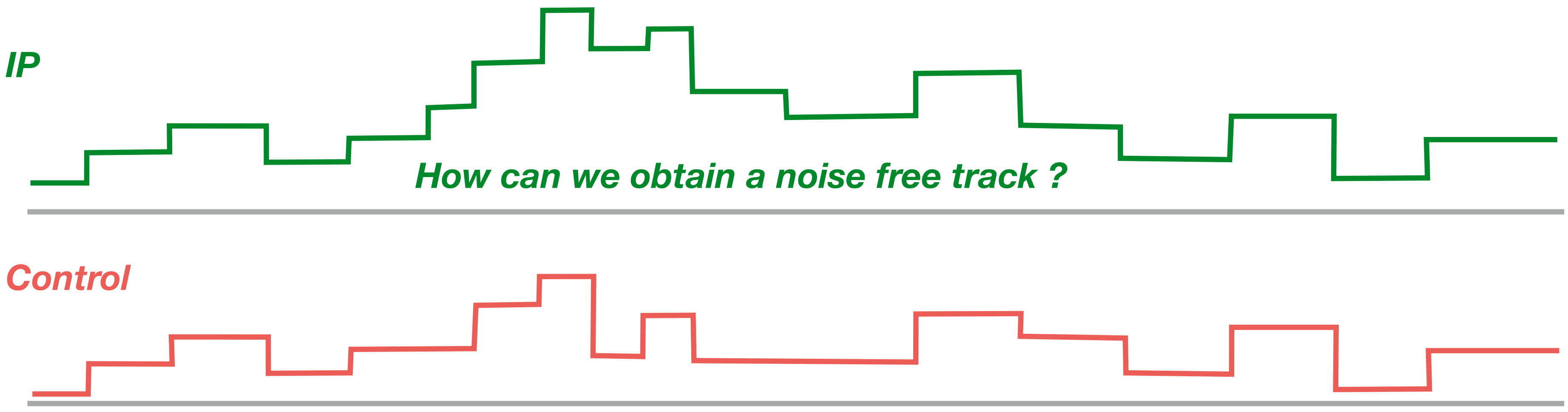
Generating signal



ESR1_treatment.bw
ESR1_control.bw
ESR1_peaks.bed
ESR1_reads.bam
[0 - 200]
[0 - 200]
treatment read density (= WIG/bigwig file)
input read density (= WIG/bigwig file)
peak (= BED file)
MACS_peak_1515
aligned reads (+ strand) (= BAM file)
aligned reads (- strand) (= BAM file)

How can we obtained a single signal track in which the background is subtracted?

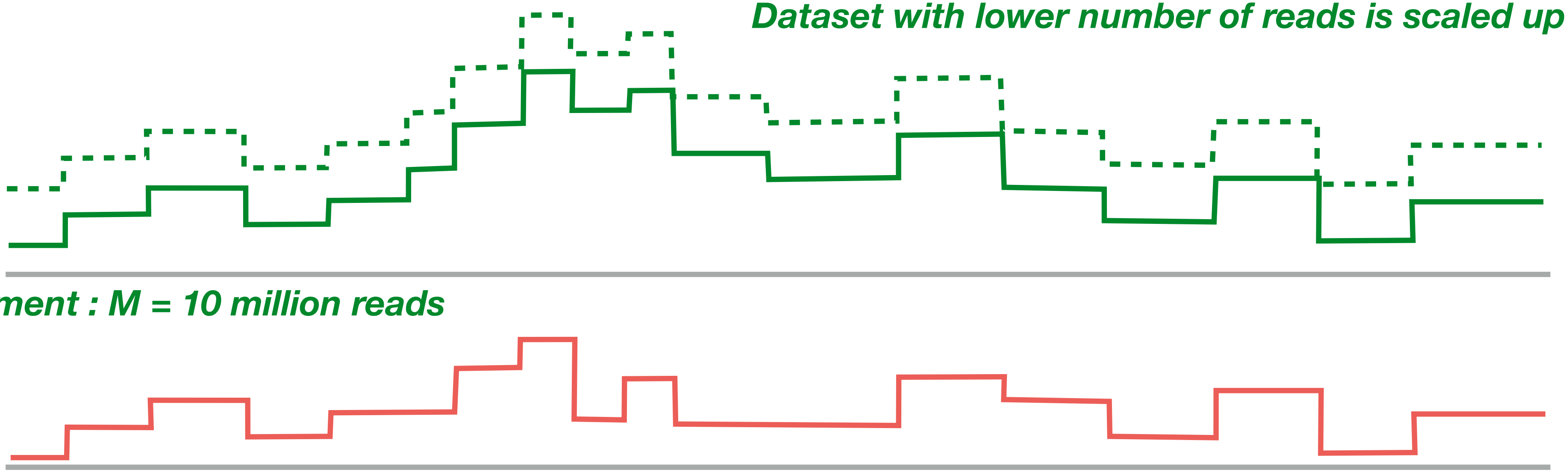
Modelling background noise



- naive subtraction: treatment - input is not possible, because of different sequencing depth
- **Simple solution** : scale library by total number of reads (library size) and perform a relative scaling

$$r = \frac{N_{ctrl}}{N_{IP}} \longrightarrow S_{IP,norm} = r \cdot S_{IP}$$

Modelling background noise



Dataset with lower number of reads is scaled up

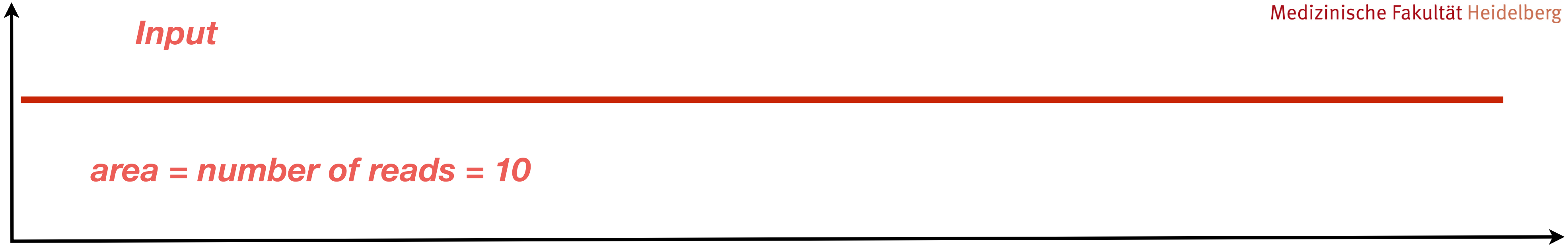
Treatment : M = 10 million reads

Input : M = 12 million reads

- **Problem** : signal influences scaling factor

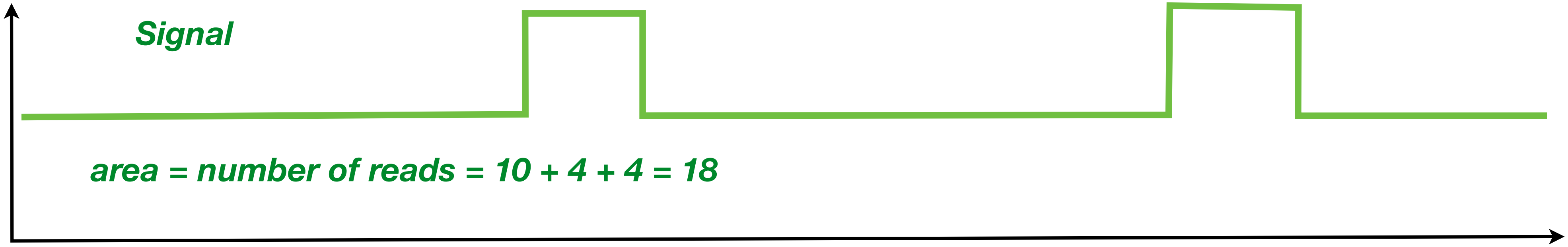
  More signal (but equal noise) → artificial noise over-estimation

Modelling background noise



Input

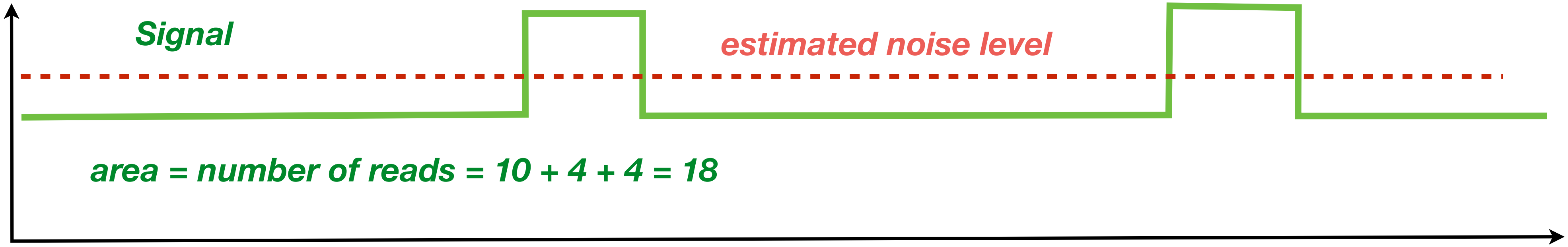
area = number of reads = 10



Signal

area = number of reads = 10 + 4 + 4 = 18

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



Signal

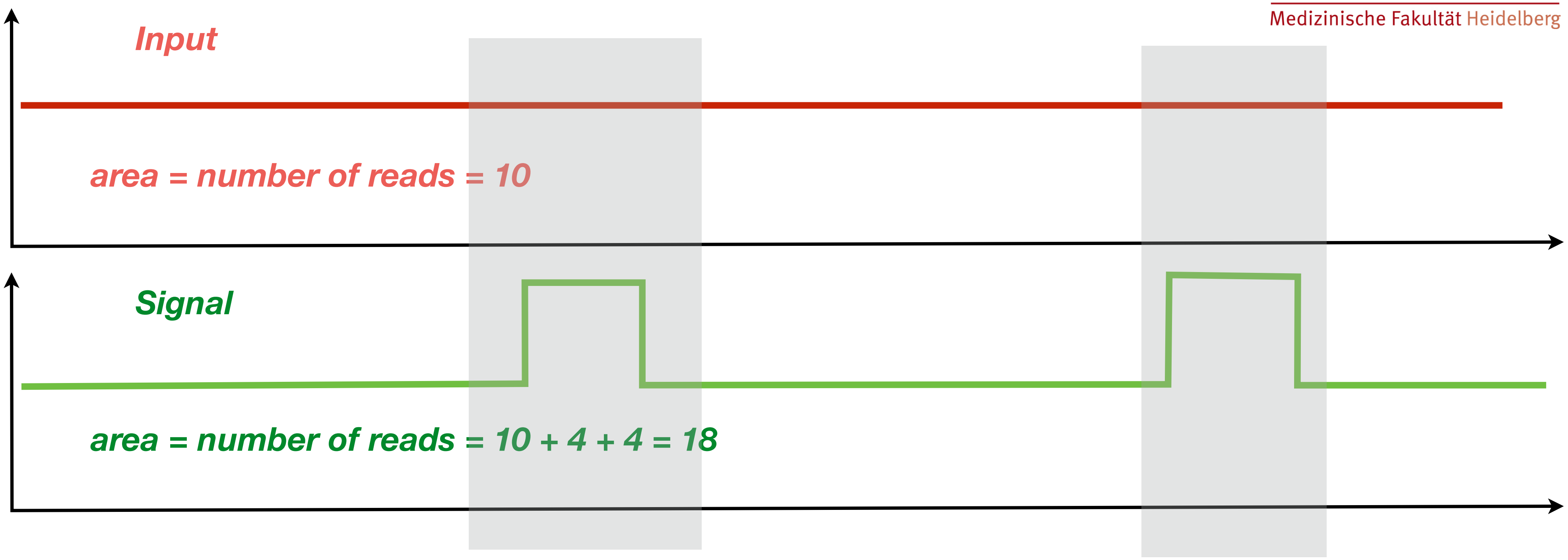
estimated noise level

area = number of reads = 10 + 4 + 4 = 18

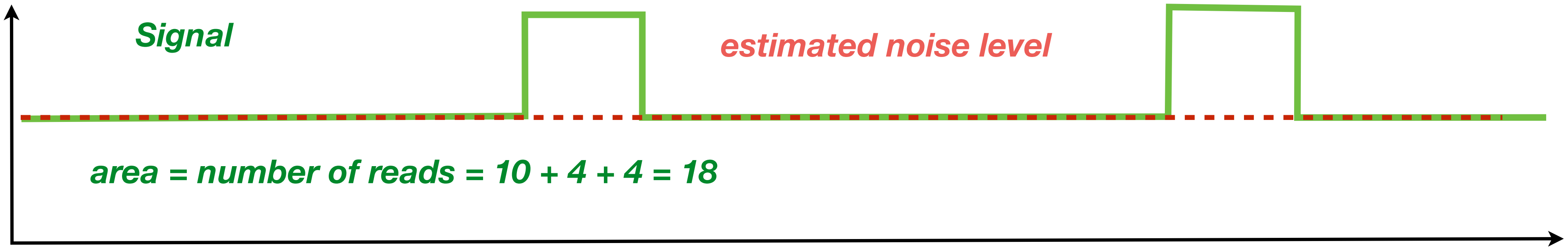
Noise level is over-estimated due to signal

Modelling background noise

Medizinische Fakultät Heidelberg

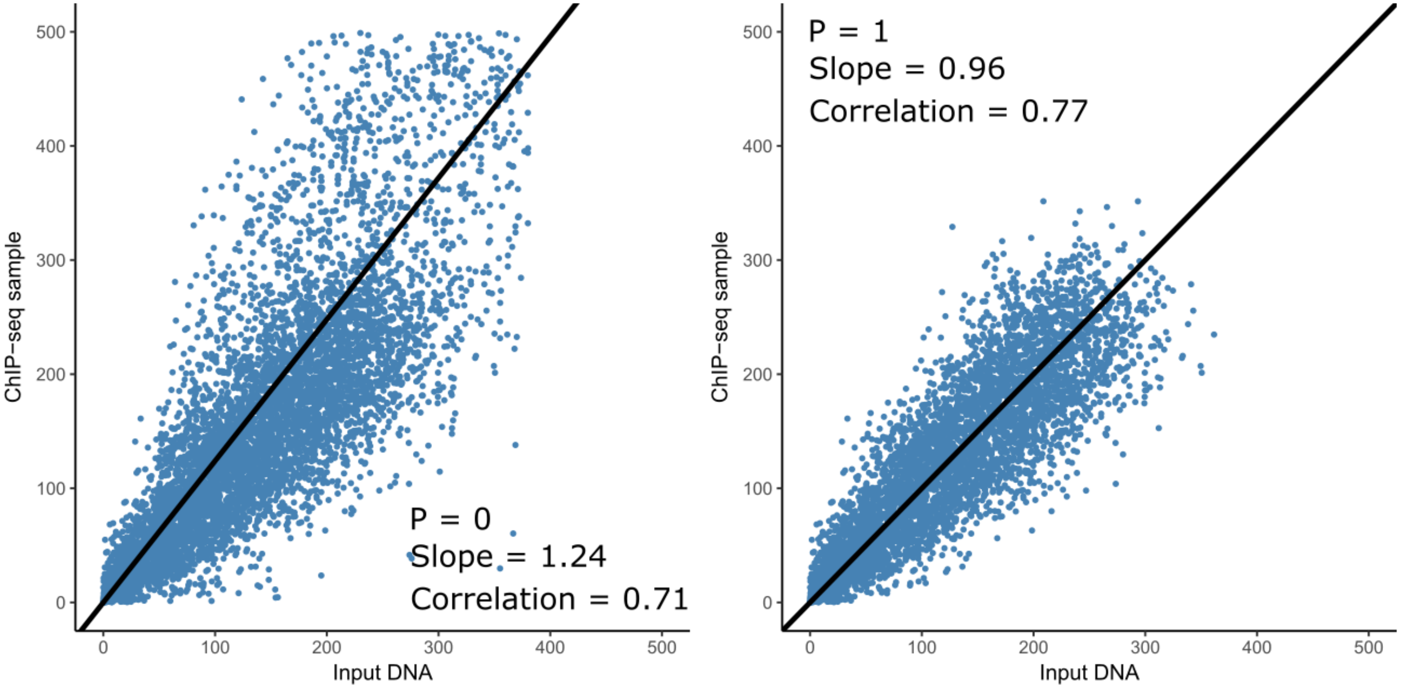

Input
area = number of reads = 10
Signal
area = number of reads = 10 + 4 + 4 = 18

Mask regions containing signal prior to scaling

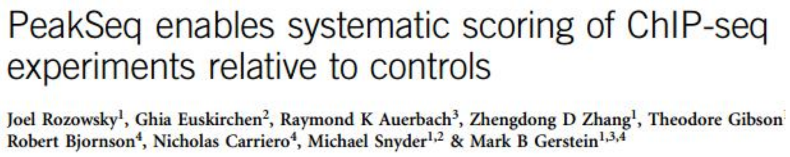


Modelling background noise

- Linear regression by excluding peak regions (PeakSeq)



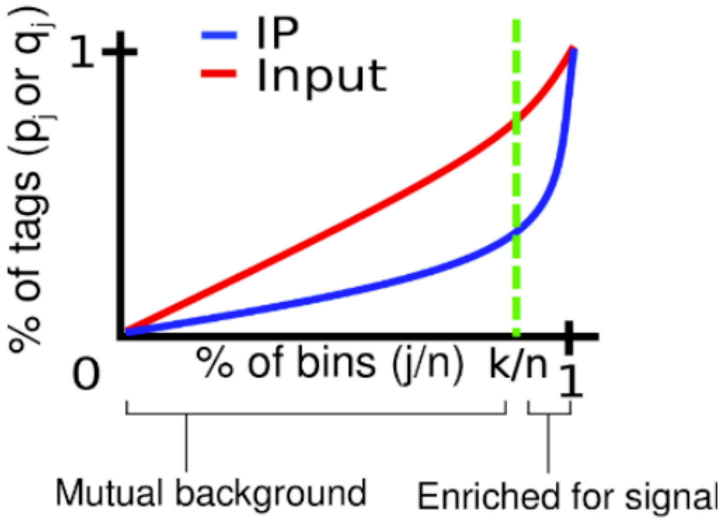
[Figure adapted from Rozowsky et al]



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 Robert Bjornson[4], Nicholas Carriero[4], Michael Snyder[1,2] & Mark B Gerstein[1,3,4]

Modelling background noise



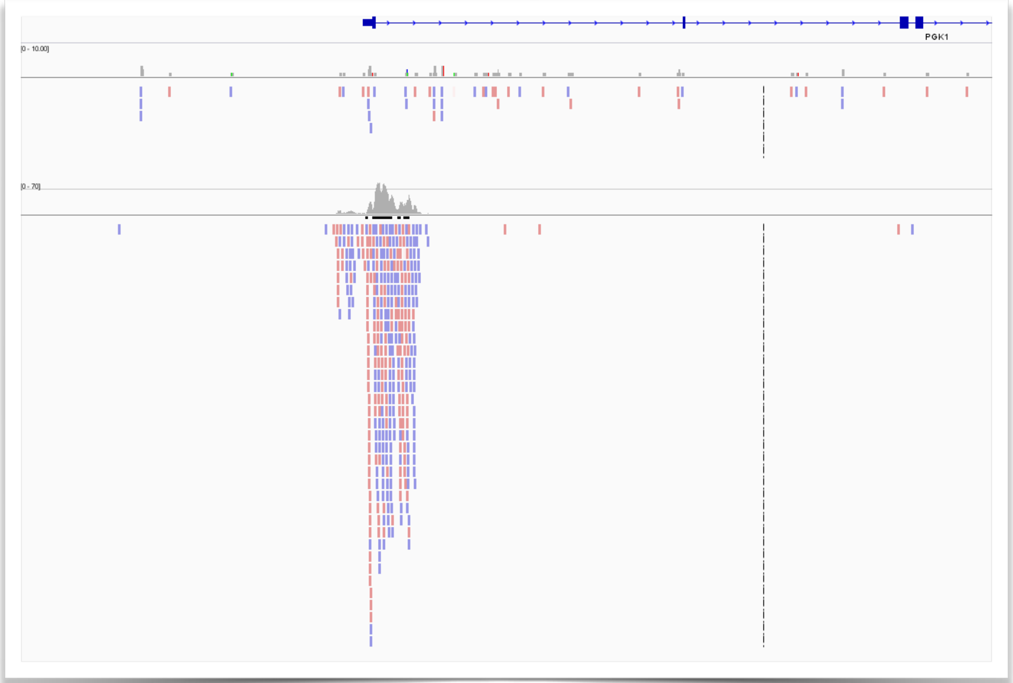
- Signal extraction scaling algorithm (SES, Diaz. et al, 2012)
- Use fingerprint plots to distinguish background noise range / signal range
- Normalize only over the number of reads in the background range

$$r_{back} = \frac{N_{ctrl} \in back}{N_{IP} \in back} \longrightarrow S_{IP,norm} = r_{back} \cdot S_{IP}$$

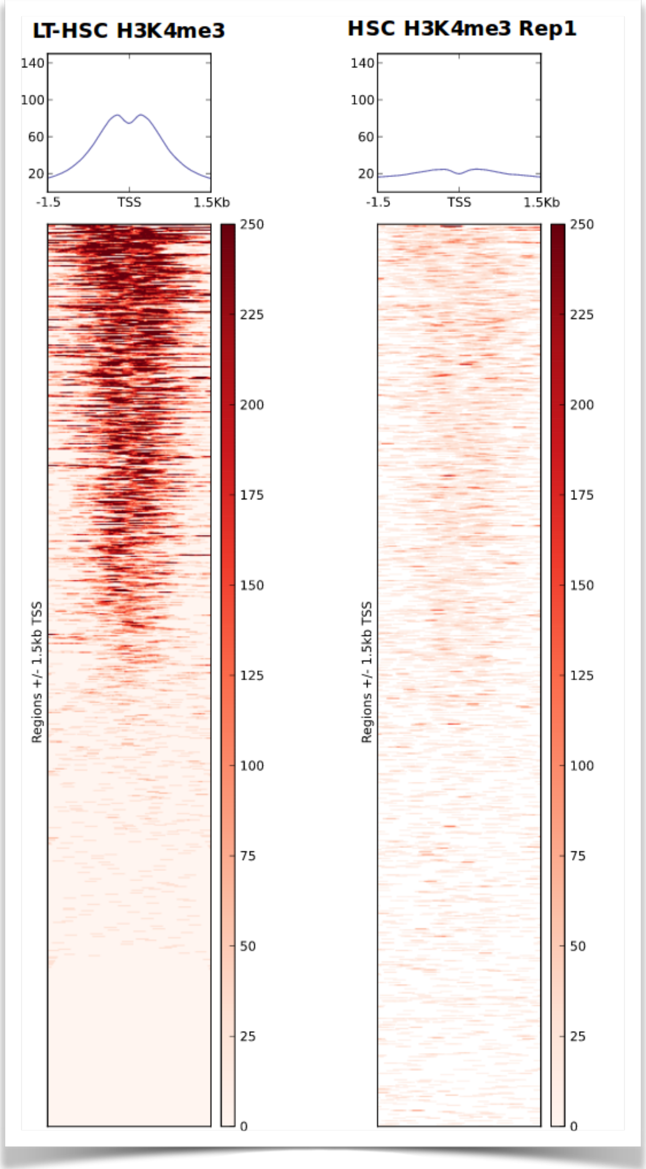
Quality control

- **Qualitative QC :**
  - check your favorite gene / region in IGV
  - heatmap of signal (e.g. at gene promoters)

Specific gene locus



Heatmap of signal at promoters



Quality control

Medizinische Fakultät Heidelberg

- **Quantitative QC :**
  - **fraction of reads in peaks** (FRiP) / SPOT : measures the fraction of reads that fall into the determined peak regions

$$FRiP=\frac{reads \in peaks}{total\ reads}$$

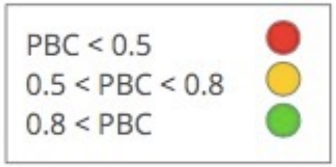
→ dependent on the type of ChIP (TF/histone)

  - **PCR Bottleneck coefficient** (PBC) : measure of library complexity

$$PBC=\frac{N_1}{N_d}$$

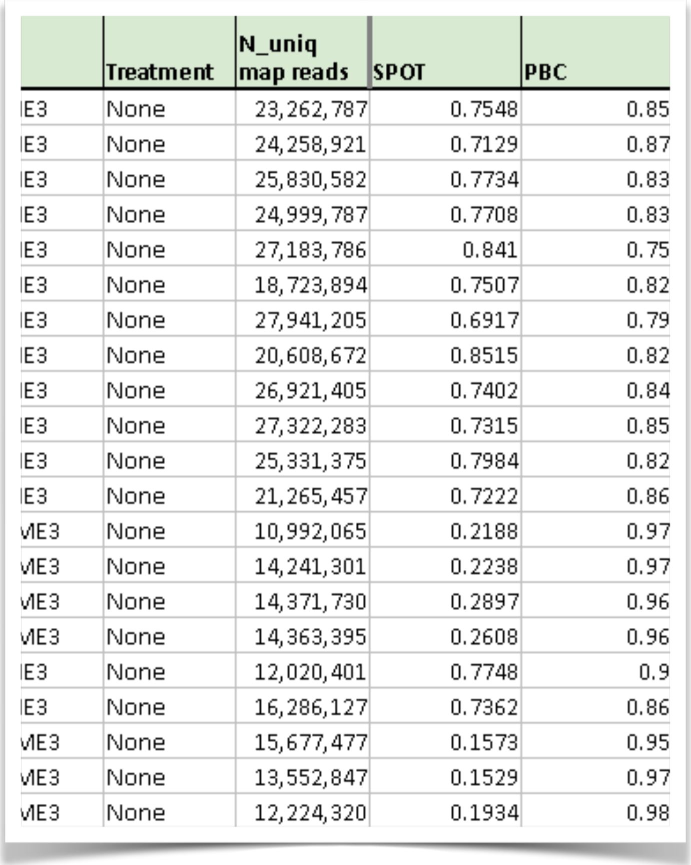
# genomic positions with **one** read aligned

#genomic positions with **one or more** reads



PBC < 0.5 ●
0.5 < PBC < 0.8 ●
0.8 < PBC ●

ENCODE quality measures

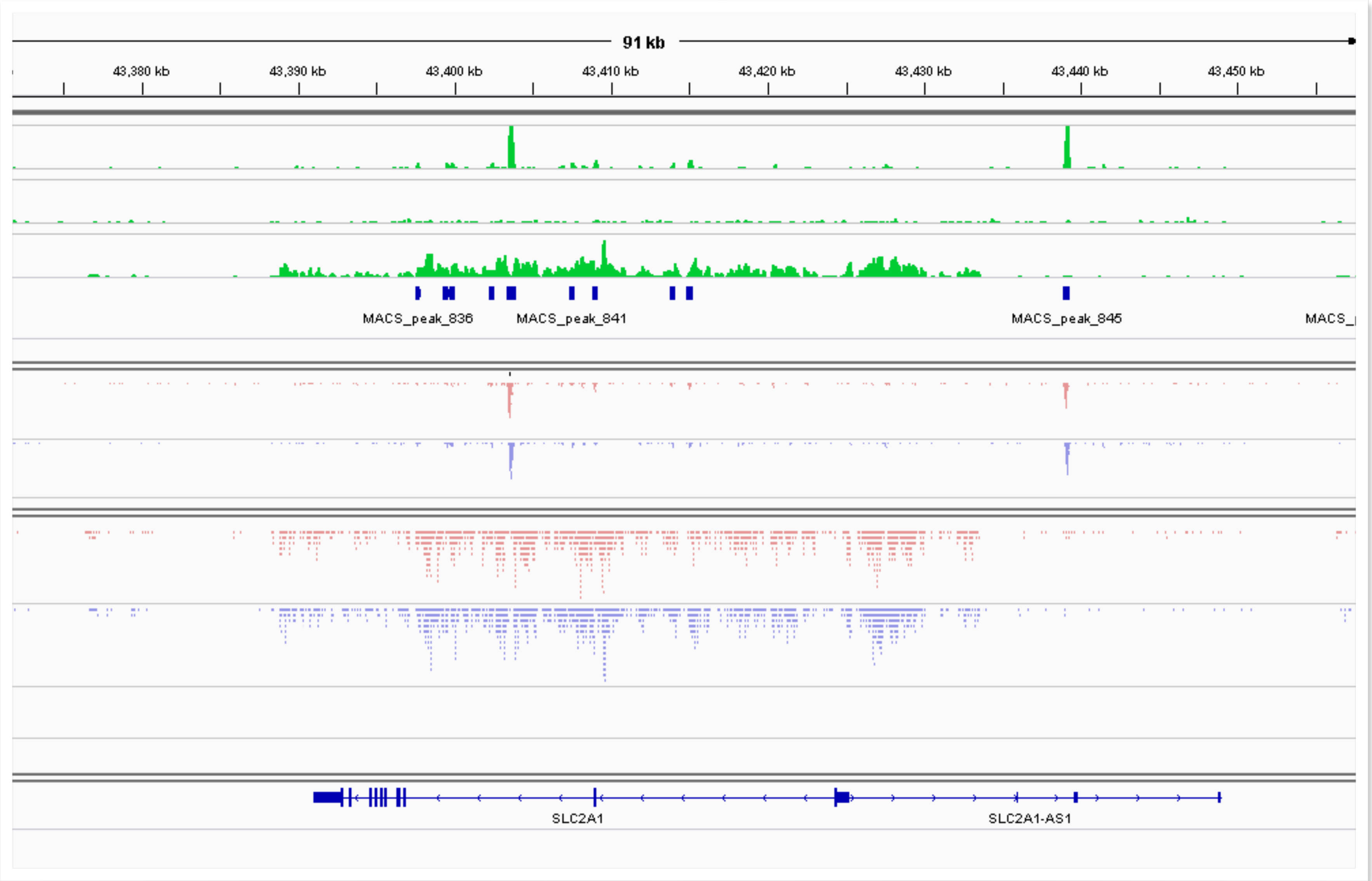


	Treatment	N_uniq map reads	SPOT	PBC
IE3	None	23,262,787	0.7548	0.85
IE3	None	24,258,921	0.7129	0.87
IE3	None	25,830,582	0.7734	0.83
IE3	None	24,999,787	0.7708	0.83
IE3	None	27,183,786	0.841	0.75
IE3	None	18,723,894	0.7507	0.82
IE3	None	27,941,205	0.6917	0.79
IE3	None	20,608,672	0.8515	0.82
IE3	None	26,921,405	0.7402	0.84
IE3	None	27,322,283	0.7315	0.85
IE3	None	25,331,375	0.7984	0.82
IE3	None	21,265,457	0.7222	0.86
ME3	None	10,992,065	0.2188	0.97
ME3	None	14,241,301	0.2238	0.97
ME3	None	14,371,730	0.2897	0.96
ME3	None	14,363,395	0.2608	0.96
IE3	None	12,020,401	0.7748	0.9
IE3	None	16,286,127	0.7362	0.86
ME3	None	15,677,477	0.1573	0.95
ME3	None	13,552,847	0.1529	0.97
ME3	None	12,224,320	0.1934	0.98

From reads to coverage

coverage (bigwig)

reads (bam)



Hands on:
signal tracks and QC

https://hdsu-bioquant.github.io/chipatac2020/07_CHIP_QC.html

https://hdsu-bioquant.github.io/chipatac2020/08_CHIP_bigwig.html