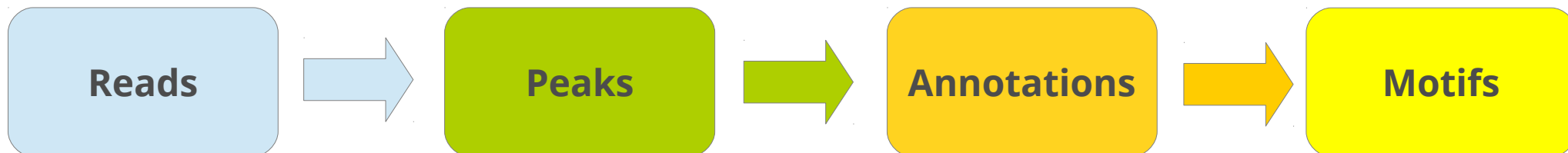


ChIP-seq analysis

M. Defrance, C. Herrmann, S. Le Gras, D. Puthier, M. Thomas.Chollier, J. van Helden

- Data visualization, quality control, normalization & peak calling
 - Presentation (Carl Herrmann)
 - Practical session
- Peak annotation
 - Presentation (Matthieu Defrance)
 - Practical session
- From peaks to motifs
 - Presentation (Jacques van Helden)
 - Practical session



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

- Oestrogen-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

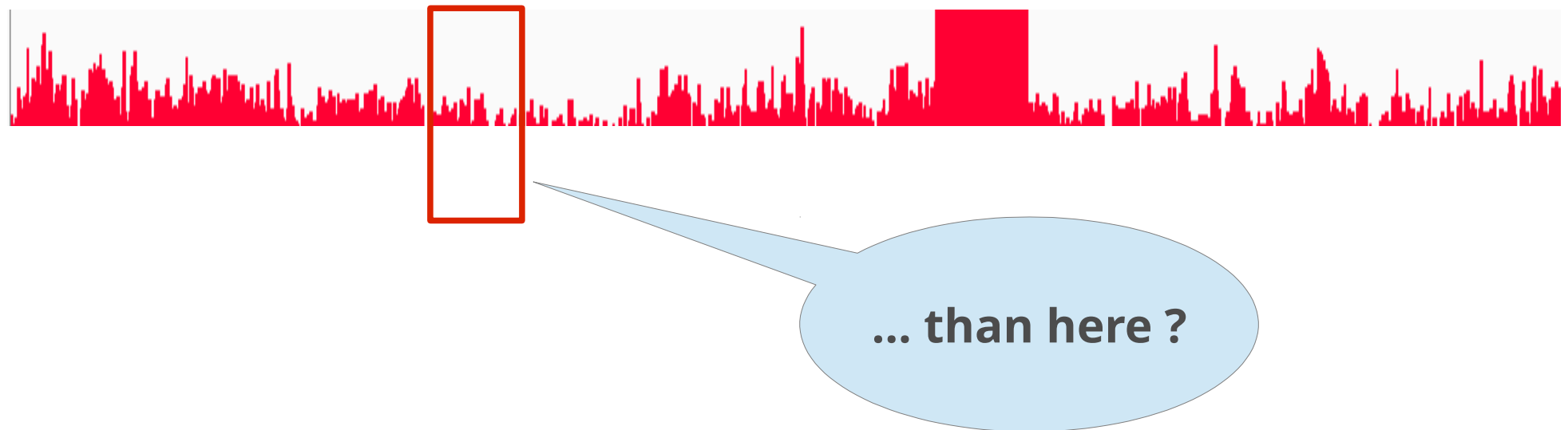
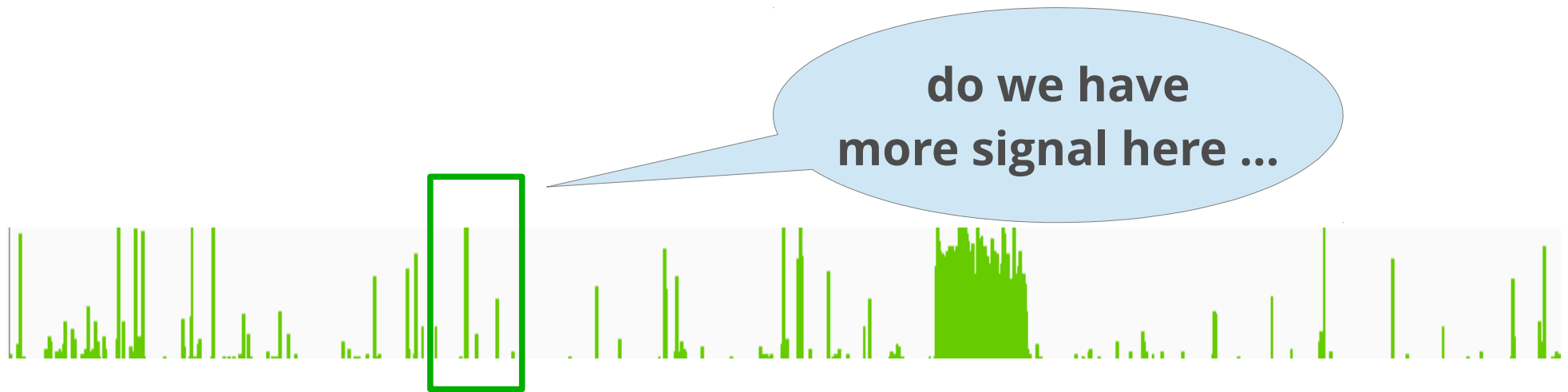
ExpName	CellLine	Replicate	SampleID	SRAExpID	Selected
siNT_ER_E2_r1	MCF-7	r1	GSM986059	SRX176856	X
siGATA_ER_E2_r1	MCF-7	r1	GSM986060	SRX176857	X
siNT_ER_E2_r2	MCF-7	r2	GSM986061	SRX176858	X
siGATA_ER_E2_r2	MCF-7	r2	GSM986062	SRX176859	X
siNT_ER_E2_r3	MCF-7	r3	GSM986063	SRX176860	X
siGATA_ER_E2_r3	MCF-7	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 ...

Hands on !!

Let's have a look at the data

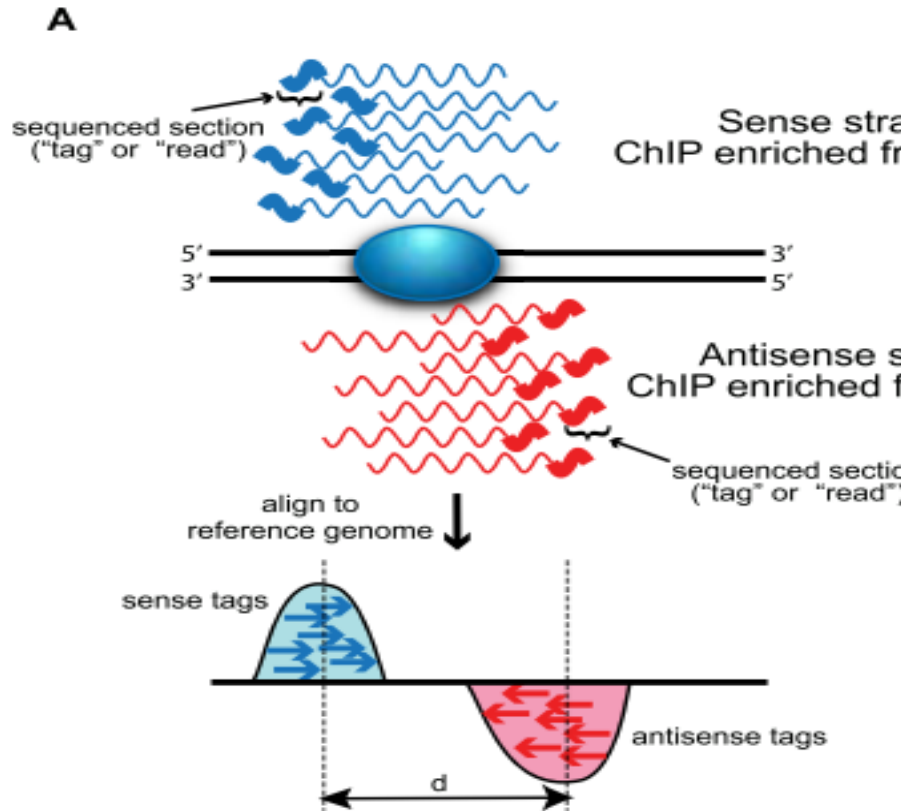
What we want to do



Keys aspects of ChIP-seq analysis

- (1) Quality Control : **do I have signal** ?
- (2) Determine signal **coverage**
- (3) Modelling **noise** levels
- (4) Scaling/**normalizing** datasets
- (5) Detecting enriched **peak** regions
- (6) Performing **differential** analysis

0. Principle of ChIP-seq

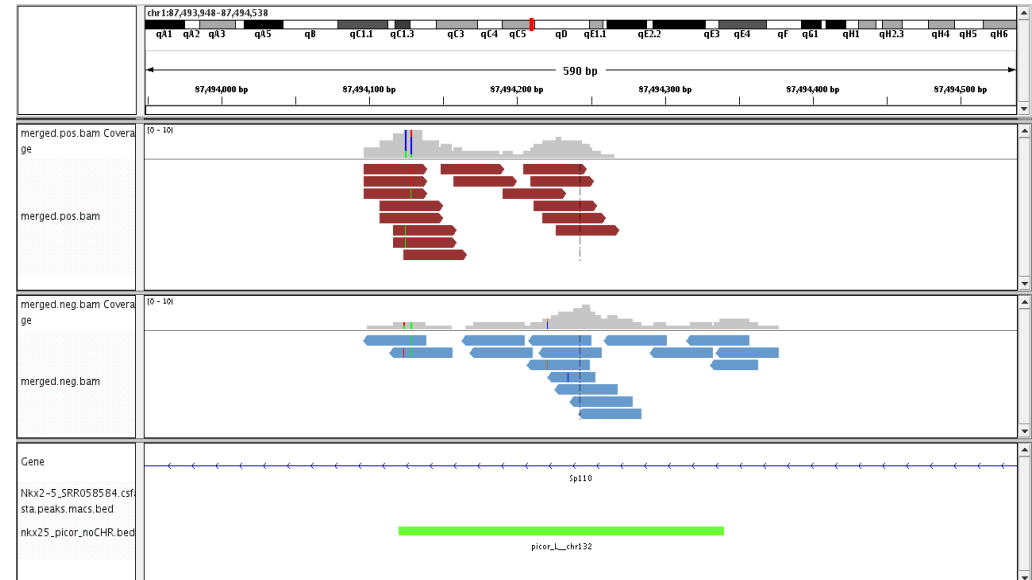
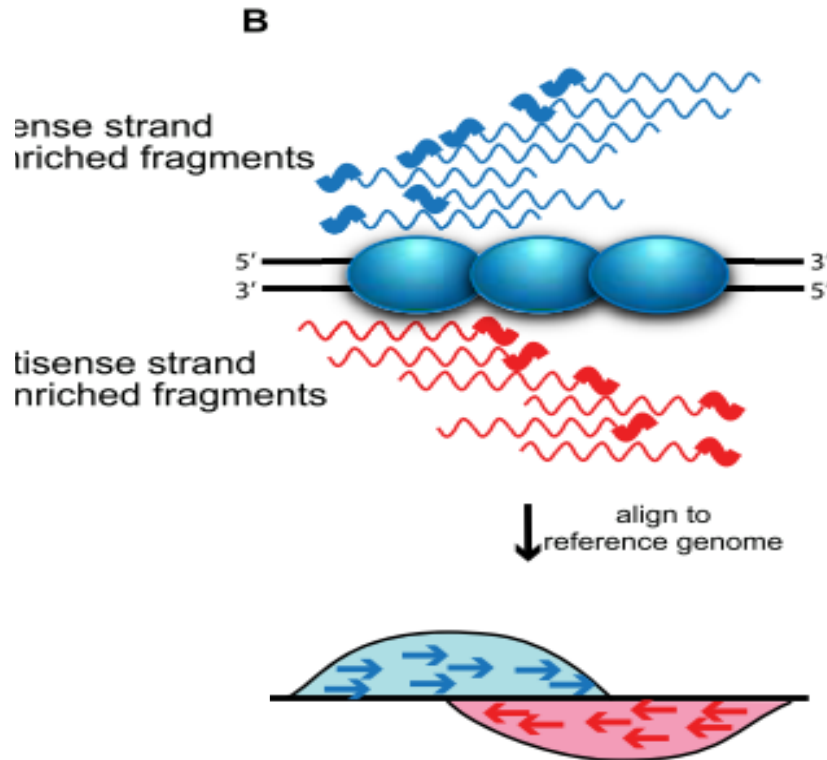


The binding site itself is generally not sequenced !

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

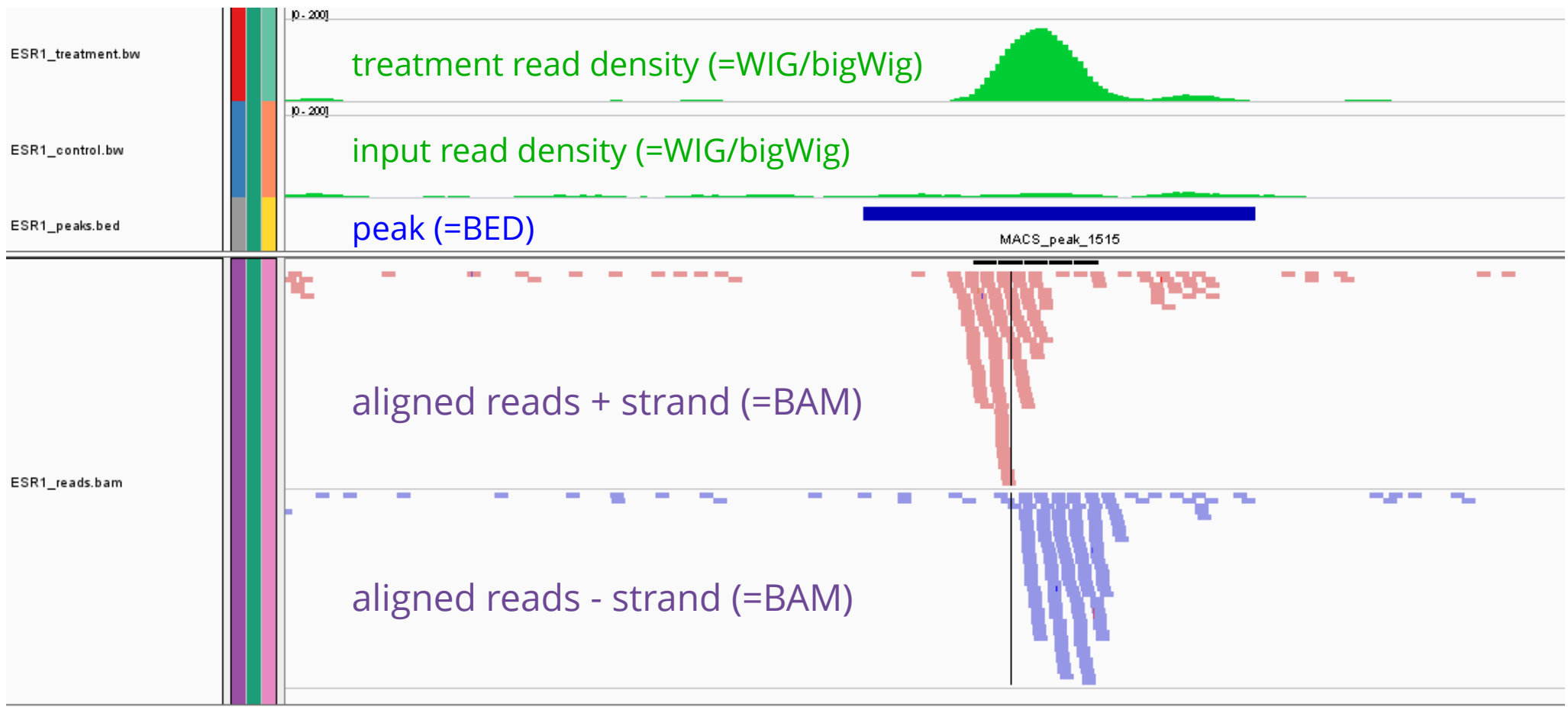
[Wilbanks & Facciotti PLoS One (2010)]

0. Principle of ChIP-seq



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

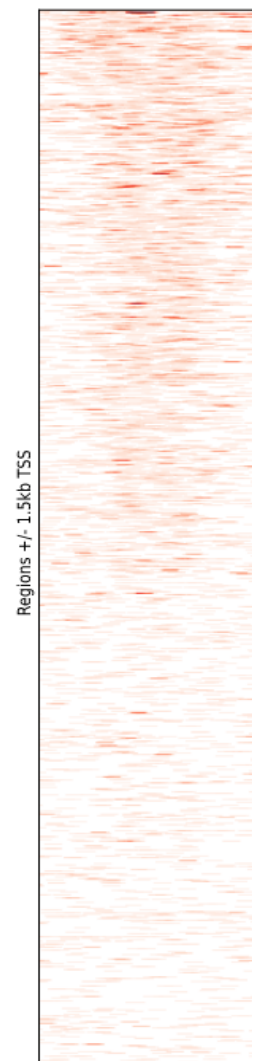
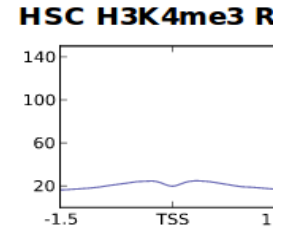
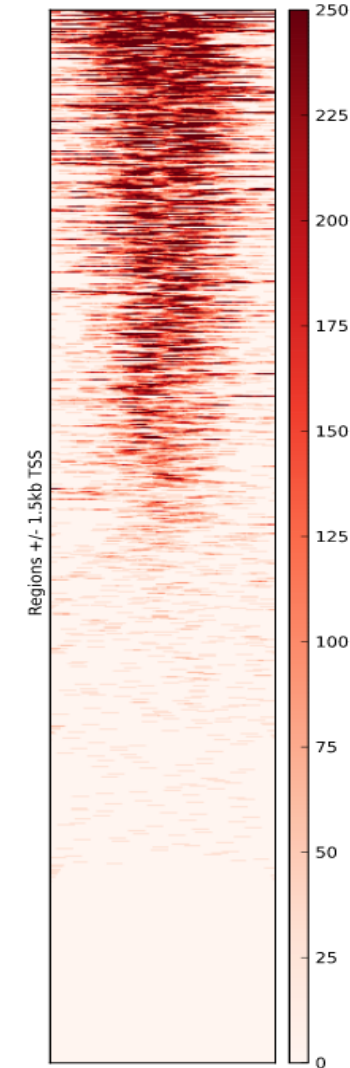
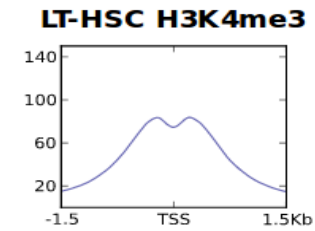
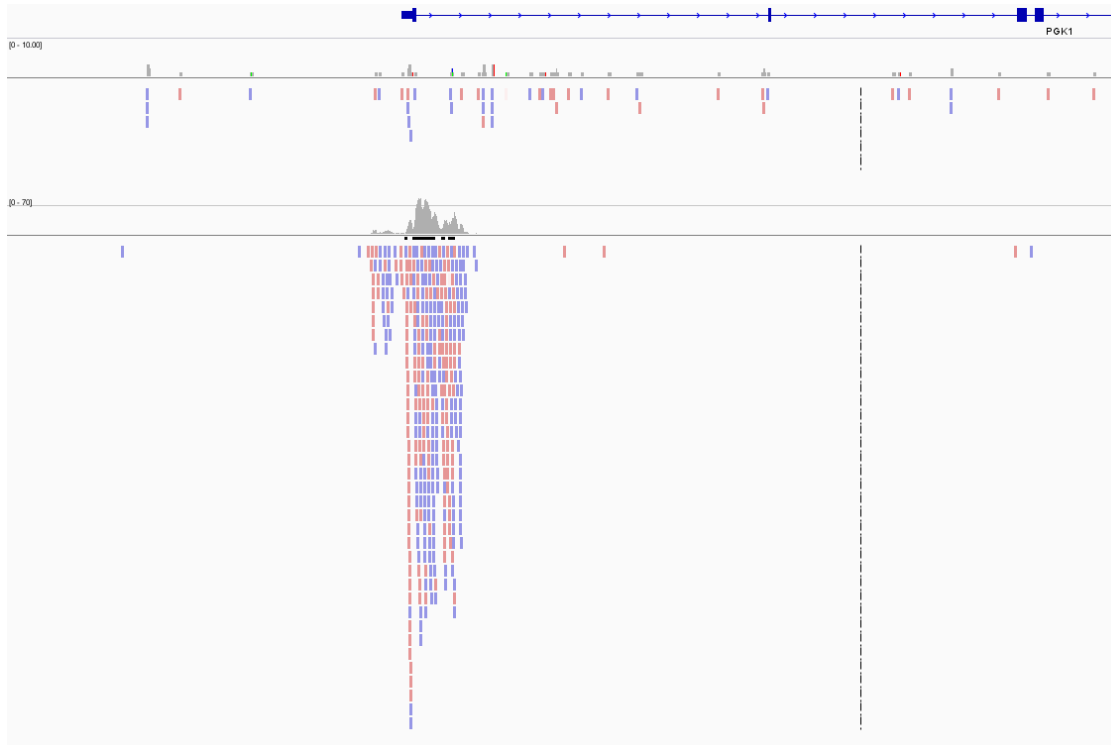
Principle of ChIP-seq



1. Quality control

- Qualitative

- Look at your favourite gene/locus in IGV !
- Heatmap of signal
→ e.g. H3K4me3 at promoters



1. Quality control

- **Quantitative**

- *Fraction of reads in peaks (FRiP)*

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

→ depends on type of ChIP
(TF/histone)

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

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

Genomic positions with 1 read aligned

Genomic positions with ≥ 1 read aligned

PBC < 0.5 ●

0.5 < PBC < 0.8 ●

0.8 < PBC ●

	A	B	C	D	F	J	K
1	Assay	Cell	Target	Treatment	N_uniq map reads	SPOT	PBC
2	TF-ChIP-seq	A549	CTCF	DEX_100nM	24,281,189	0.2361	0.71
3	TF-ChIP-seq	A549	CTCF	DEX_100nM	15,453,361	0.1249	0.41
4	TF-ChIP-seq	A549	GR	DEX_100nM	16,608,102	0.0754	0.91
5	TF-ChIP-seq	A549	GR	DEX_100nM	28,467,922	0.0723	0.44
6	TF-ChIP-seq	A549	POL2	DEX_100nM	19,005,470	0.6166	0.86
7	TF-ChIP-seq	A549	POL2	DEX_100nM	23,115,884	0.5388	0.86
8	TF-ChIP-seq	A549	USF1	DEX_100nM	22,289,881	0.0791	0.87
9	TF-ChIP-seq	A549	USF1	DEX_100nM	12,364,820	0.0517	0.82
10	TF-ChIP-seq	A549	GR	DEX_500pM	19,646,503	0.0105	0.96
11	TF-ChIP-seq	A549	GR	DEX_500pM	15,095,316	0.0109	0.94
12	TF-ChIP-seq	A549	GR	DEX_50nM	19,291,260	0.1289	0.96
13	TF-ChIP-seq	A549	GR	DEX_50nM	16,754,796	0.1426	0.95
14	TF-ChIP-seq	A549	GR	DEX_5nM	20,120,740	0.0343	0.98
15	TF-ChIP-seq	A549	GR	DEX_5nM	20,559,786	0.0641	0.96
16	TF-ChIP-seq	A549	CTCF	EtOH_0.02p	22,672,467	0.1601	0.75
17	TF-ChIP-seq	A549	CTCF	EtOH_0.02p	14,351,615	0.2040	0.42

rget	Treatment	N_uniq map reads	SPOT	PBC
3K4ME3	None	23,262,787	0.7548	0.85
3K4ME3	None	24,258,921	0.7129	0.87
3K4ME3	None	25,830,582	0.7734	0.83
3K4ME3	None	24,999,787	0.7708	0.83
3K4ME3	None	27,183,786	0.841	0.75
3K4ME3	None	18,723,894	0.7507	0.82
3K4ME3	None	27,941,205	0.6917	0.79
3K4ME3	None	20,608,672	0.8515	0.82
3K4ME3	None	26,921,405	0.7402	0.84
3K4ME3	None	27,322,283	0.7315	0.85
3K4ME3	None	25,331,375	0.7984	0.82
3K4ME3	None	21,265,457	0.7222	0.86
3K27ME3	None	10,992,065	0.2188	0.97
3K27ME3	None	14,241,301	0.2238	0.97
3K36ME3	None	14,371,730	0.2897	0.96
3K36ME3	None	14,363,395	0.2608	0.96
3K4ME3	None	12,020,401	0.7748	0.9
3K4ME3	None	16,286,127	0.7362	0.86
3K27ME3	None	15,677,477	0.1573	0.95
3K27ME3	None	13,552,847	0.1529	0.97
3K36ME3	None	12,224,320	0.1934	0.98

<https://www.encodeproject.org/data-standards/2012-quality-metrics/>

2. from reads to coverage

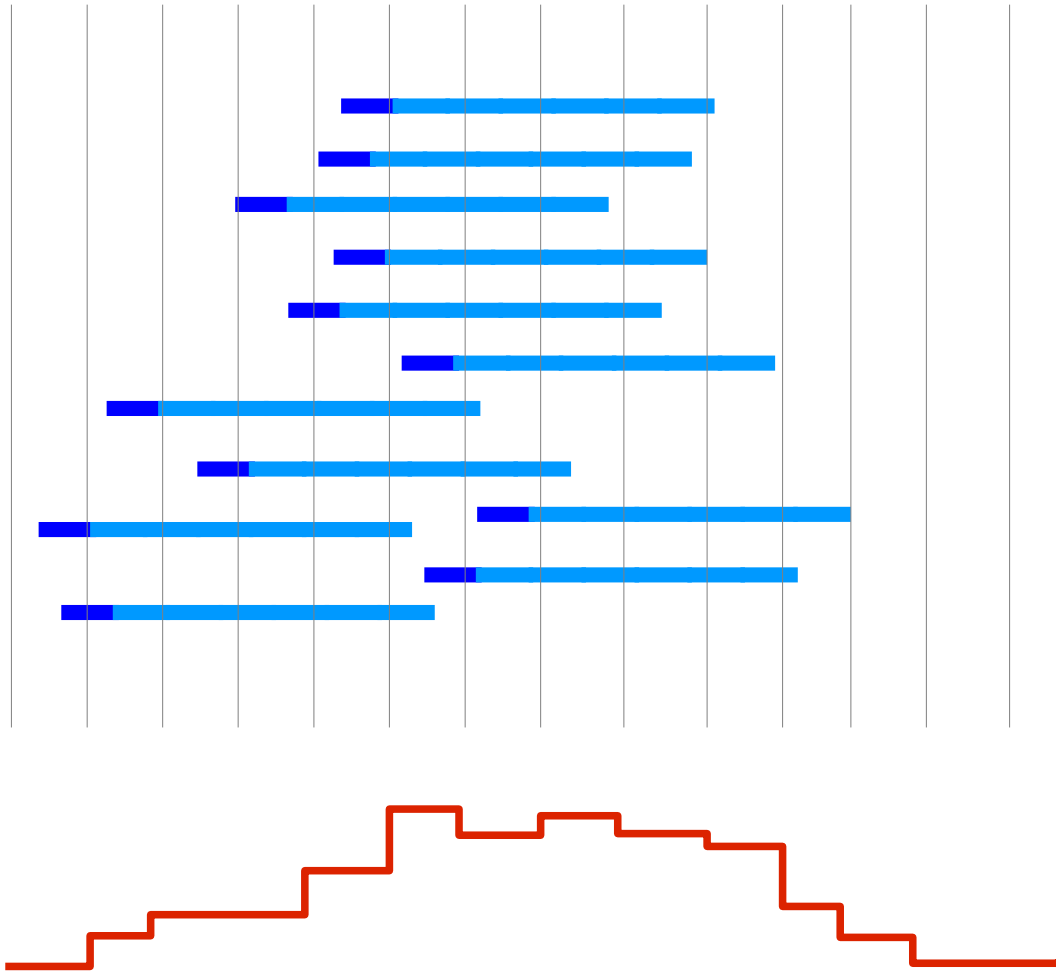
- to visualize the data, we use **coverage plots** (=density of fragments per genomic region)
- need to reduce BAM file to more compact format
→ **bigWig/bedGraph**

SRR540188.2375302	0	chr1	10510	42	36M	*	0	0	AGGAGAACTGTCTCCGCTTCAAGATACCAACGAA	776<,27:7Z@>6>4>@6995;9558;7A6@679	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.10764383	0	chr1	14909	40	36M	*	0	0	AAAAAGGCAGACAGAATTACGAGGTGCTGGCCAG	BFGAGGGGGGGGFGDFGFGGEGCEEEEEEEEEEG	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:21A14	
SRR540188.1487116	0	chr1	55533	42	36M	*	0	0	TGCCTGGAAACTCCAAATCTATCAACTCTGGGGCT	GEGGFGGGGGGGGGGGGGGGGGGGGGGGGGGG	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.15782692	0	chr1	58414	42	36M	*	0	0	TTTAAAGACTGTTTCTACTAAGAAAGCATAGACCGC	GGGGGDFGFEDEEEFFFFFGGDFGGFGGGGGDED	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.2785359	16	chr1	88299	40	36M	*	0	0	GCTCAATACCTCCCTCAACATCCATGGCTAAAC	DEBCEEB8CFGGGGGECFFEEGGGGGGGGGGGGG	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:17G18	
SRR540188.27624849	16	chr1	257099	42	36M	*	0	0	TGAGGTTGGAGGTGAGAAACGGAAGAGAAACAGAN	#####	AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35A0	
SRR540188.1220386	16	chr1	536965	40	36M	*	0	0	GTTTGACAAACACAGCATCACTTACCCAACTCTGT	ABB=DD877678;4;?55AA>?A5?A>A65?@;6	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:21G14	
SRR540188.13227692	0	chr1	536971	40	36M	*	0	0	CAAAACACAGCATCACTTACCCAACTCTGTAGAGCCA	FGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:15G20	
SRR540188.20526476	16	chr1	547554	23	36M	*	0	0	TGTAGTGTTCCTGTCTCTGTGCATGAGTCCCTG	CB8C<DCDEEFDFFDAFFGGDFGGFFDGGEDEFF	AS:i:-10	XN:i:0	XM:i:2	XO:i:0	XG:i:0	NM:i:2	MD:Z:8C1	
SRR540188.17352892	16	chr1	547563	40	36M	*	0	0	CCCATGTCTCTGTGCATGAGTCCCTGTGTGATG	BFEEEEFFBFFFFFFFDBFEF?F?FFFEFFFEFF	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:1A34	
SRR540188.14595113	16	chr1	567562	42	36M	*	0	0	ACGTACTCGTTGTAGCCCACTTCCACTATGTCTCTA	9?A>CAD-DDDDDDDACACA@>@>?><>>A?A	AS:i:-3	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:7A28	
SRR540188.11265536	0	chr1	567563	42	36M	*	0	0	CGTACTACGTTGTAGCCCACTTCCACTGTGTCTCTAT	EEAEEDFFFEFFAFEFDFDEFFFAACD5CB?CDDDF	AS:i:-4	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:27A8	
SRR540188.15147378	16	chr1	568418	42	36M	*	0	0	CTTTACCATCAATCAATTGGCCATCAATGGTACTG	EBCC@DFGEFDEEE=DGGGBGGGGGGGGGGGGGG	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.18564081	16	chr1	568423	23	36M	*	0	0	CCATCAATCAATTGGCCATCAATGGTACTGTAAT	EFFDFED?DDEFFEDDBDEEEDDEEDEFFFEFF	AS:i:-10	XN:i:0	XM:i:2	XO:i:0	XG:i:0	NM:i:2	MD:Z:31A	
SRR540188.11341118	0	chr1	568691	40	36M	*	0	0	TGCTACACGACG		AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:18A17	
SRR540188.12642068	0	chr1	569605	42	36M	*	0	0	CCACACTAGCA		AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.14432151	0	chr1	569605	42	36M	*	0	0	CCACACTAGCA		AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.14237282	16	chr1	569895	42	36M	*	0	0	CCACACTAGCA		AS:i:-3	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:13A22	
SRR540188.27718380	0	chr1	569902	42	36M	*	0	0	CCACACTAGCA		AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:0C35	
SRR540188.5865251	0	chr1	569903	42	36M	*	0	0	CCACACTAGCA		AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:0A35	
SRR540188.4049441	0	chr1	569905	40	36M	*	0	0	CCACACTAGCA		AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35C0	
SRR540188.5927782	16	chr1	569908	42	36M	*	0	0	CCACACTAGCA		AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35G0	
SRR540188.22917741	16	chr1	569911	42	36M	*	0	0	CCACACTAGCA		AS:i:-4	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:8A27	
SRR540188.27625113	16	chr1	581420	42	36M	*	0	0	TTAGTGA		AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35A0	
SRR540188.27775676	16	chr1	662561	42	36M	*	0	0	CCTGTTGA		AS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35G0	
SRR540188.13379337	16	chr1	665062	40	36M	*	0	0	ATCTCGCGAGC		AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:35T0	
SRR540188.25671965	0	chr1	713991	42	36M	*	0	0	AGTGATGTAGC		AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.24973101	16	chr1	718666	42	36M	*	0	0	TAGTGAGACAG		AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU
SRR540188.19542605	16	chr1	720363	42	36M	*	0	0	CTAATAATGAG		AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:36	YT:Z:UU

#bedGraph section chr1:10500-932425

chr1	10500	10700	0.01
chr1	58425	58625	0.01
chr1	567375	567525	0.01
chr1	567525	567550	-0.55
chr1	567550	567600	-1.1
chr1	567600	567750	-1.11
chr1	567750	567775	-0.55
chr1	568250	568375	0.58
chr1	569675	569725	0.58
chr1	569725	569750	1.74
chr1	569750	569825	1.18
chr1	569825	569900	0.03
chr1	569900	569950	0.63

2. from reads to coverage



→ deepTools :
bamCoverage

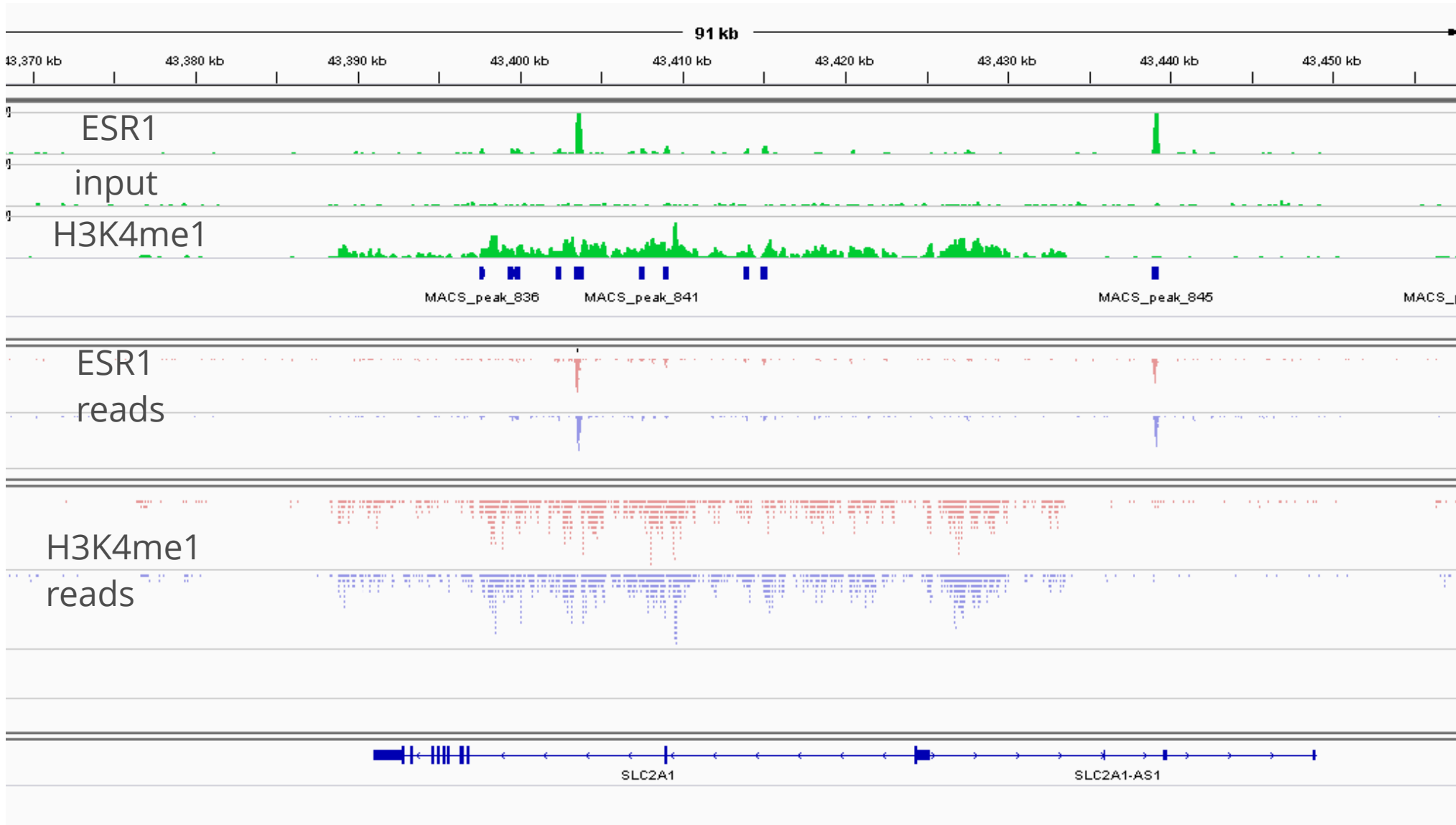
- Reads are extended to 3' to fragment length
- Read counts are computed for each bin
- Counts are normalized
 - reads per genomic content
→ normalize to 1x coverage

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

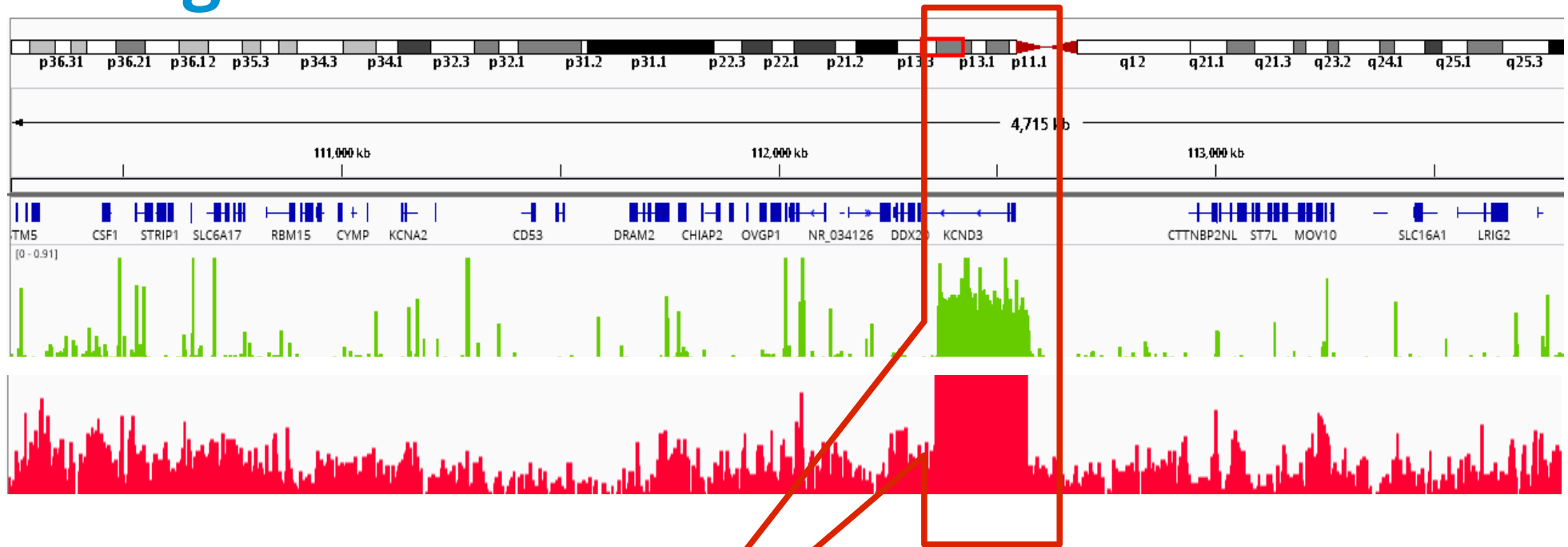
- reads per kilobase per million
reads per bin

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

2. from reads to coverage



3. signal and noise

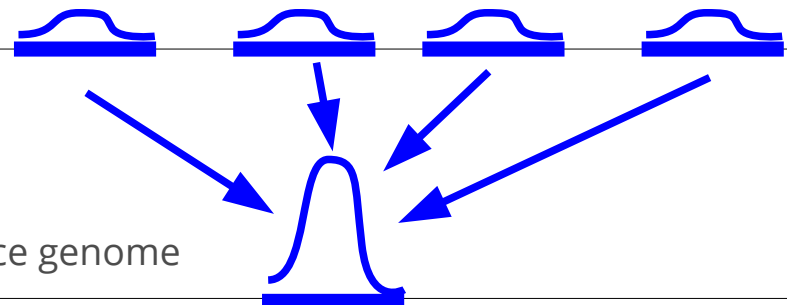


MCF-7 genome

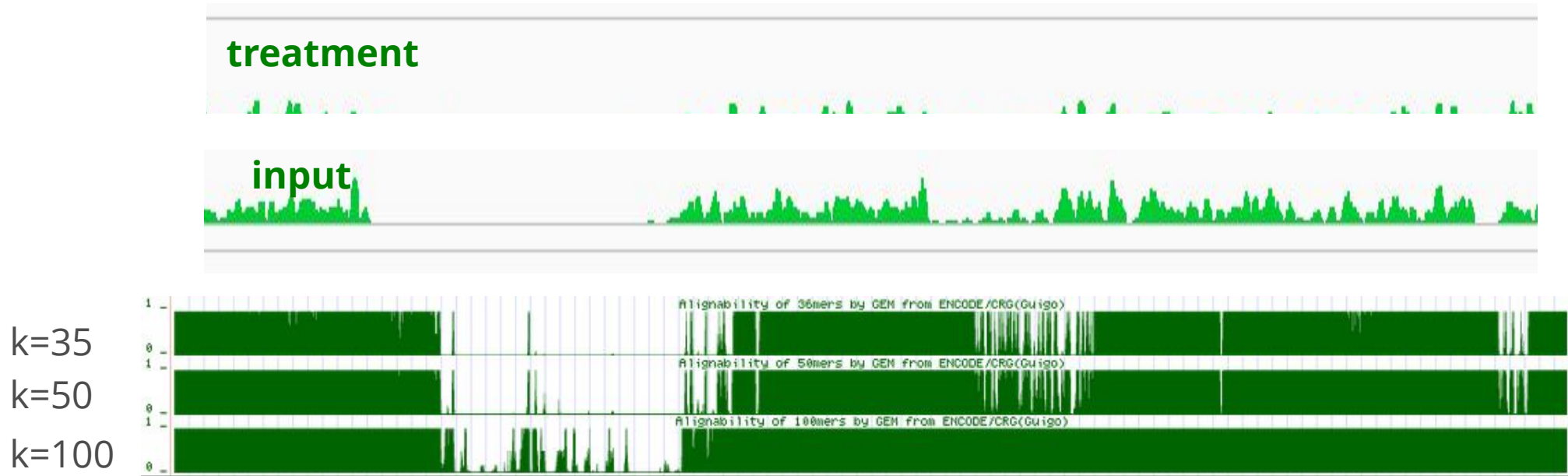
The MCF-7 genome harbors 21 high-level CNAs, summarized in Table 1. Remarkably, many of the previously reported regions of genetic alteration split into multiple segments upon tiling resolution analysis. The 1p13 amplification described previously [40] in fact divides into three distinct segments of 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



3. signal to noise



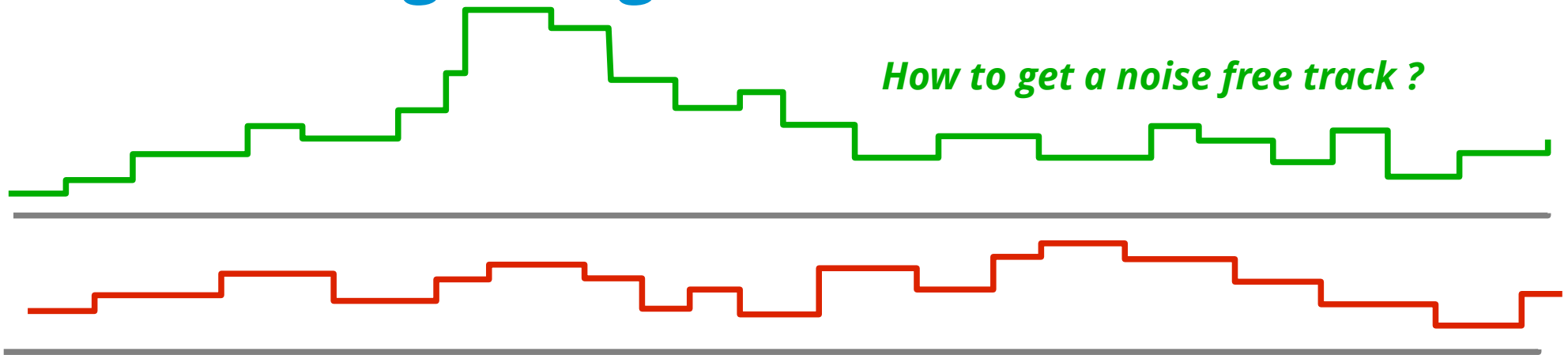
- **Mappability issue** : “alignability” track shows, how many times a read from a given position of the genome would align
 - $a=1$ → read from this position ONLY aligns to this position
 - $a=1/n$ → read from this position could align to n locations→ we usually only keep uniquely aligned reads : **positions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

3. signal to noise

The availability of a control sample in mandatory !

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

4. modelling background level



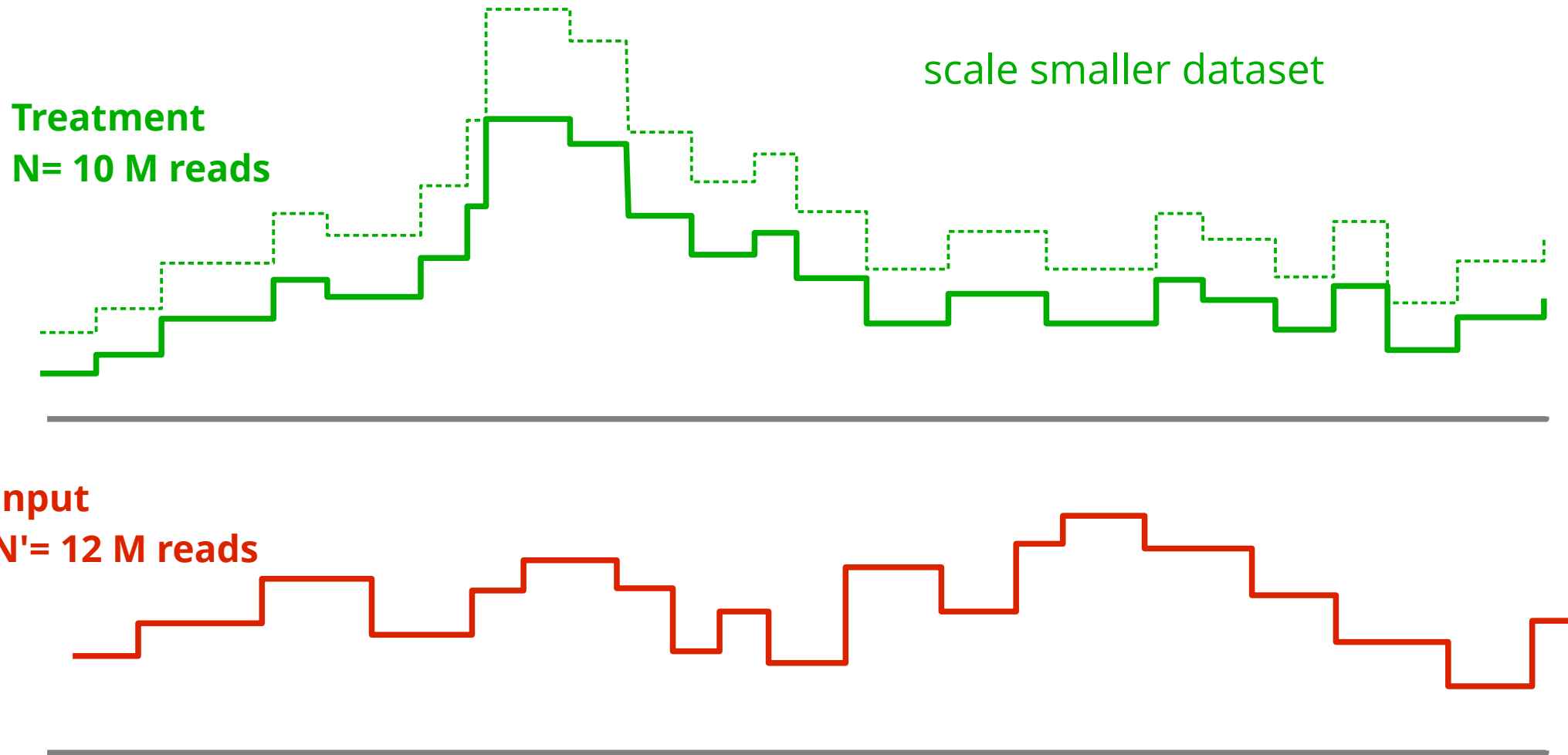
- **naïve subtraction** treatment – input is not possible, because both libraries have different sequencing depth !
- **Solution 1** : before subtraction, scale both libraries by total number of reads (library size)

- RPGC
- RPKM

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

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

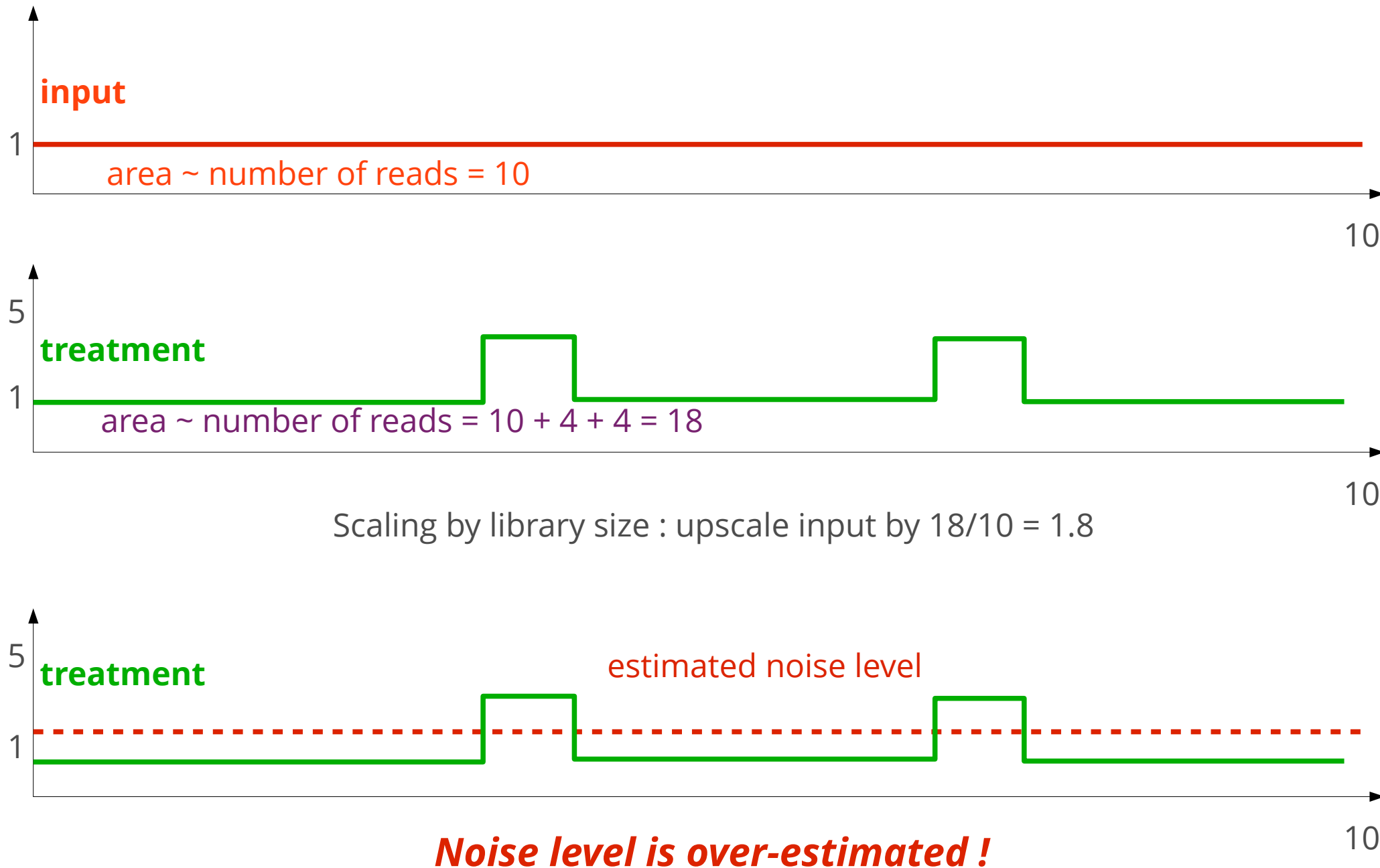
4. modelling background level



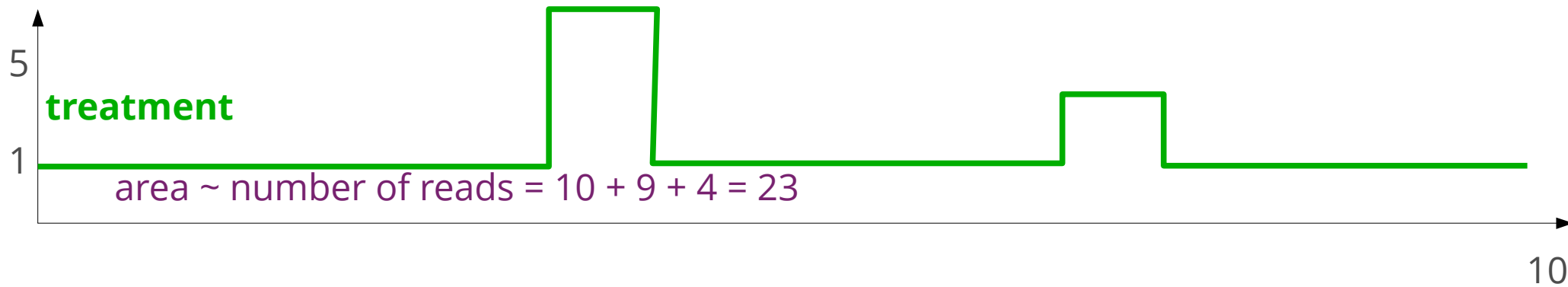
Problem : signal influences scaling factor

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

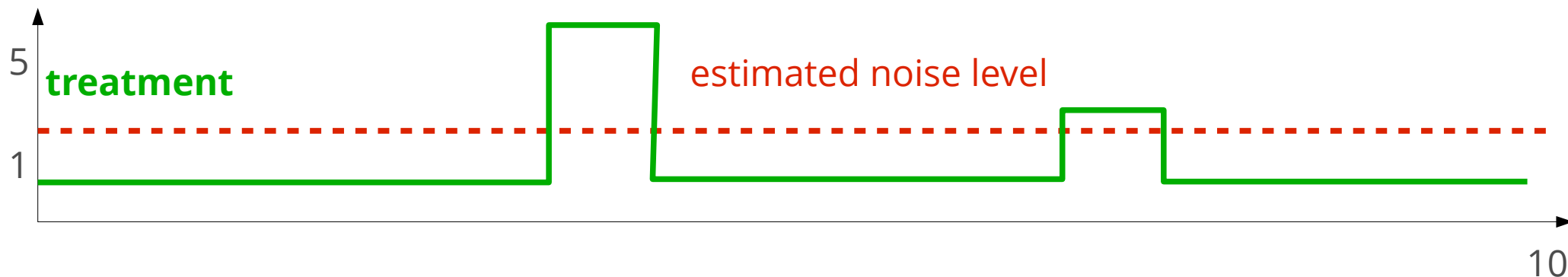
4. modelling background level



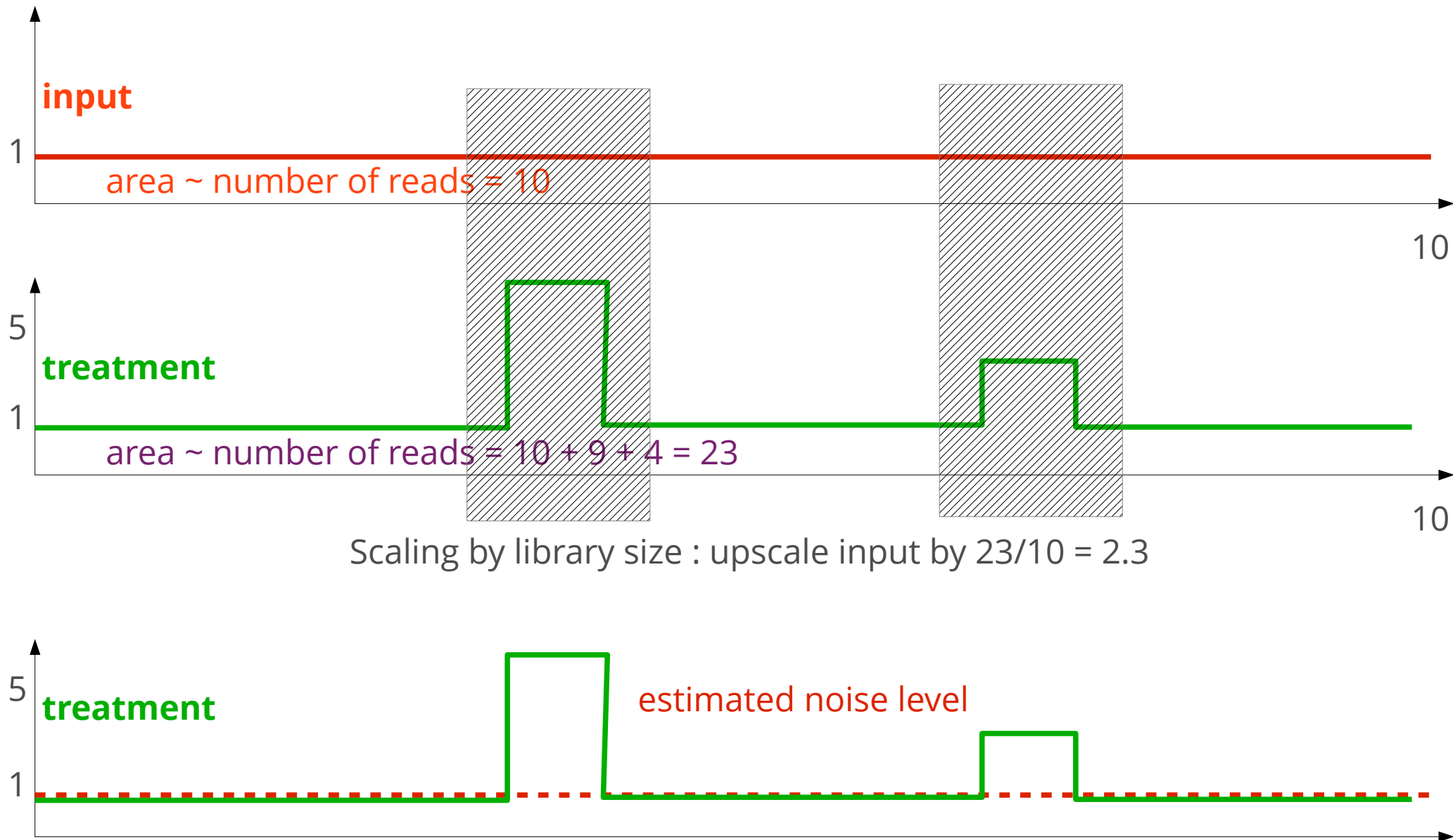
4. modelling background level



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

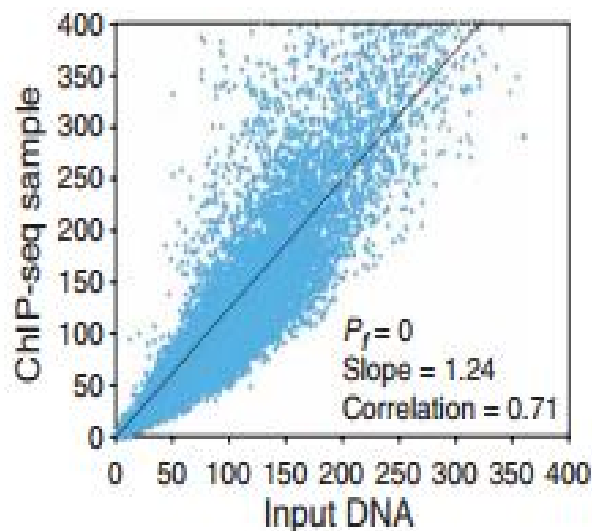


4. modelling background level

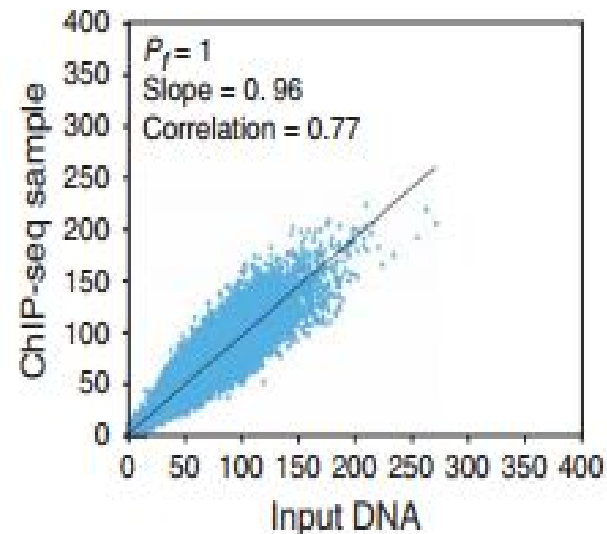


4. modelling background level

- **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 Carriero⁴, Michael Snyder^{1,2} & Mark B Gerstein^{1,3,4}

4. modelling background level

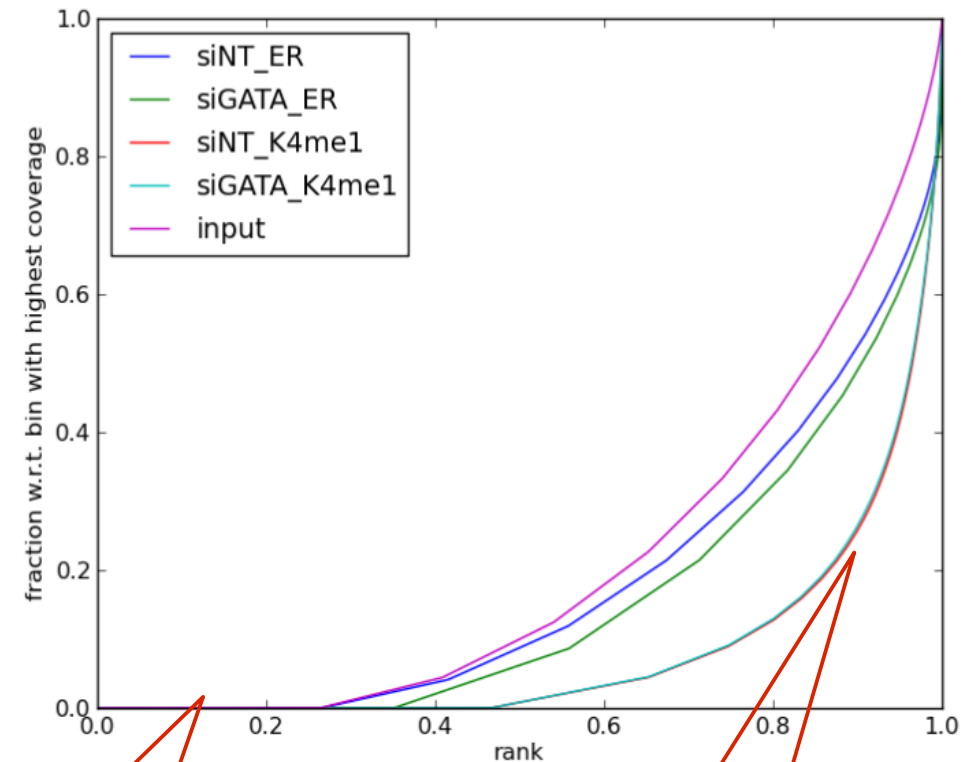
- **Alternative strategy**

(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

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

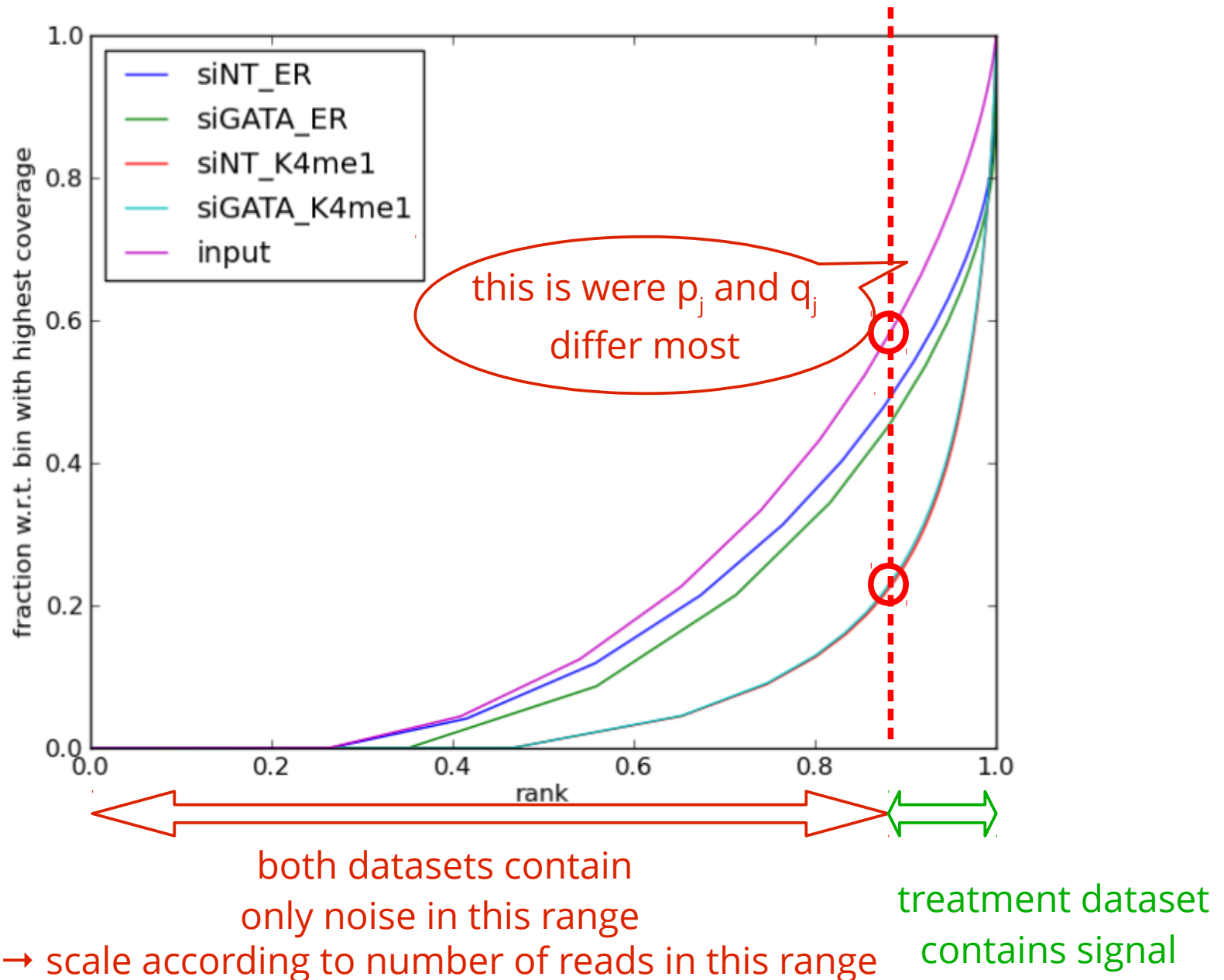
- p_j = proportion of reads in the j less enriched windows



25% of the genome
contains no reads !

90% of the genome
contain ~ 25% of reads

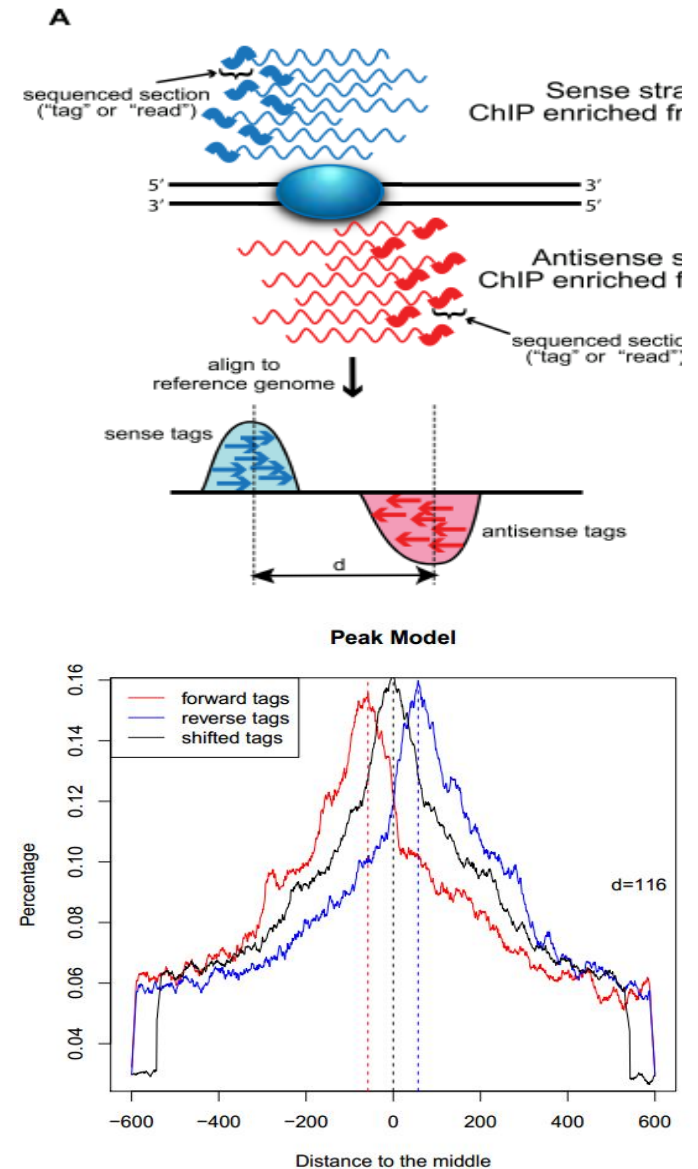
4. modelling background level



5. from reads to peaks

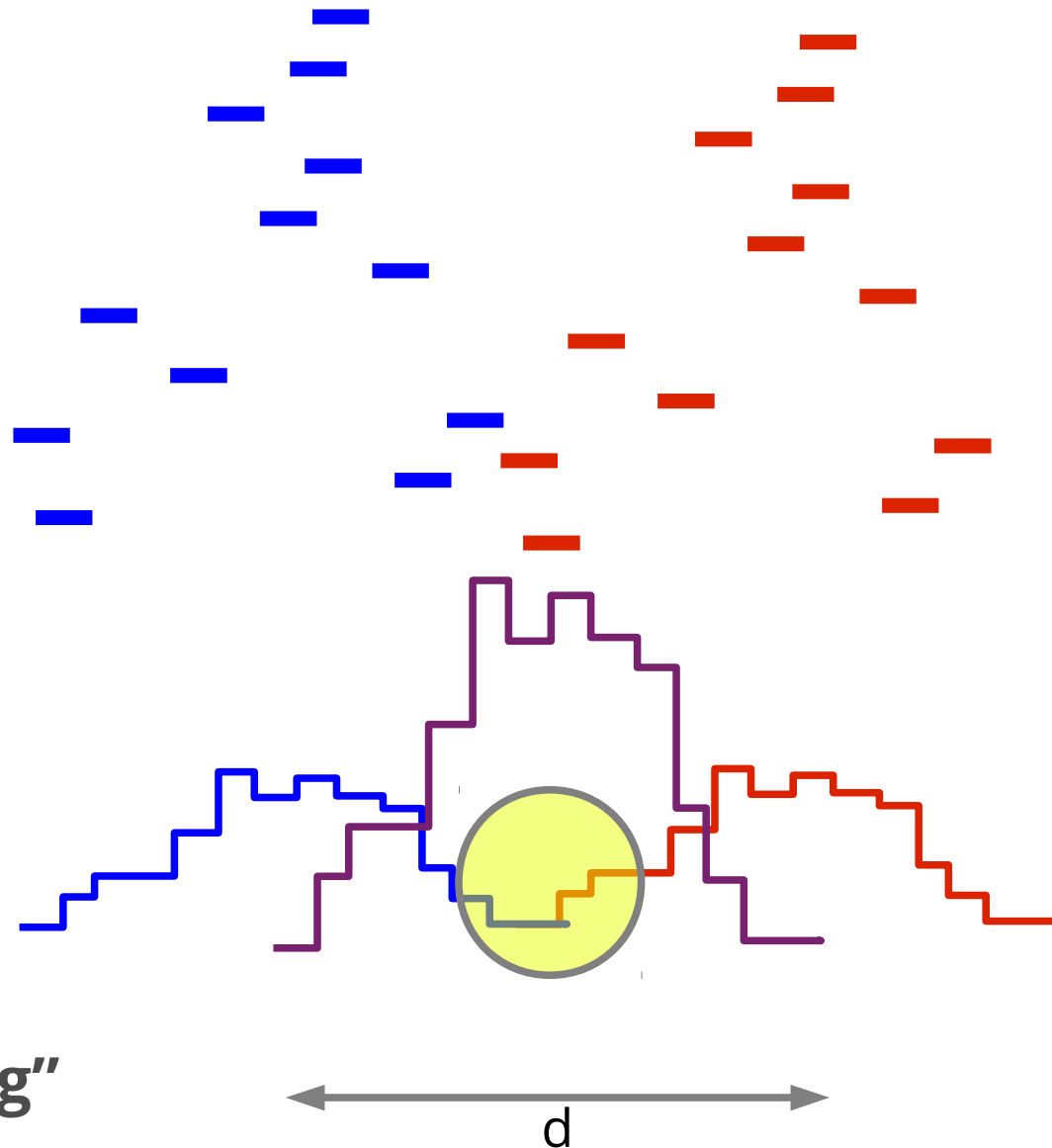
- **Tag shifting vs. extension**

- 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



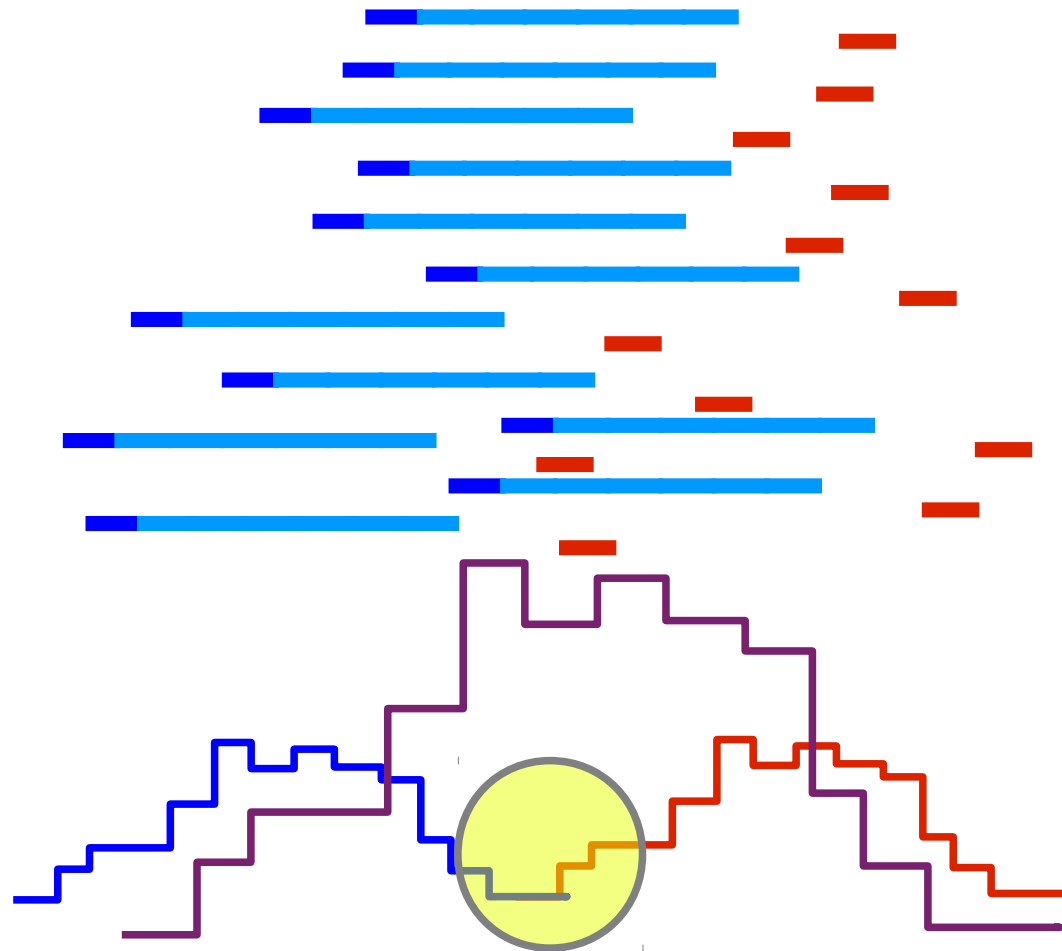
example of MACS model building
using top enriched regions

5. from reads to peaks



“Read shifting”

5. from reads to peaks



"Read extension"



5. From reads to peaks: peak-calling approaches

	Profile	Peak criteria ^a	Tag shift	Control data ^b	Rank by	FDR ^c	User input parameters ^d	Final 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 <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)	<i>s</i> 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 <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, <i>m</i> fold 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)	No / Yes
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region ^f	Average nearest paired tag distance	Used to compute fold-enrichment distribution				
spp v1.0	Strand specific window scan	Poisson <i>P</i> value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling				

Pepke, S., Wold, B. and Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6, S22–32.

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)
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 <i>P</i> value	Estimate from high quality peak pairs
PeakSeq	Extended tag aggregation	Local region binomial <i>P</i> 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	<i>P</i> 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 <i>P</i> value (paired peaks only)	Maximal strand cross-correlation

Tag shift

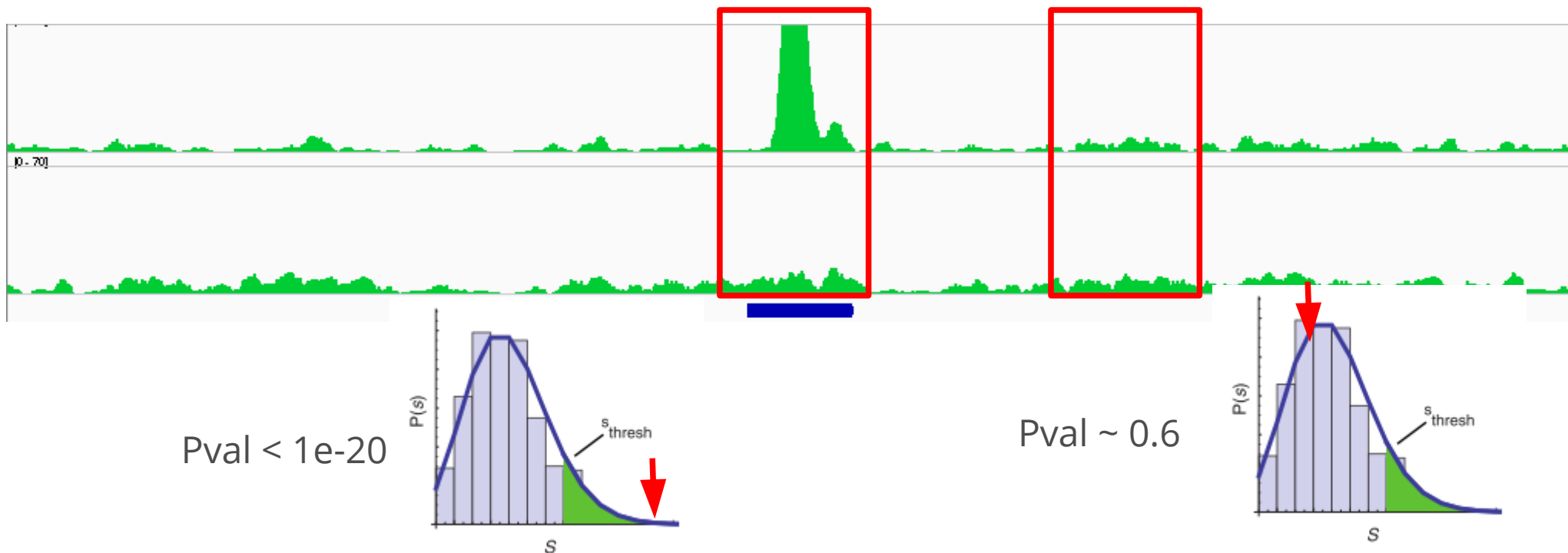
Tag extension

Tags unchanged

5. from reads to peaks

- **Determining “enriched” regions**

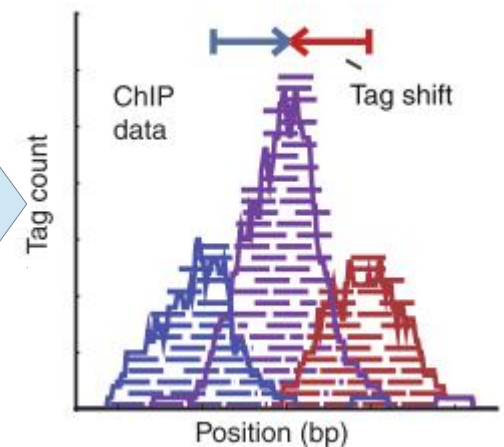
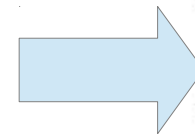
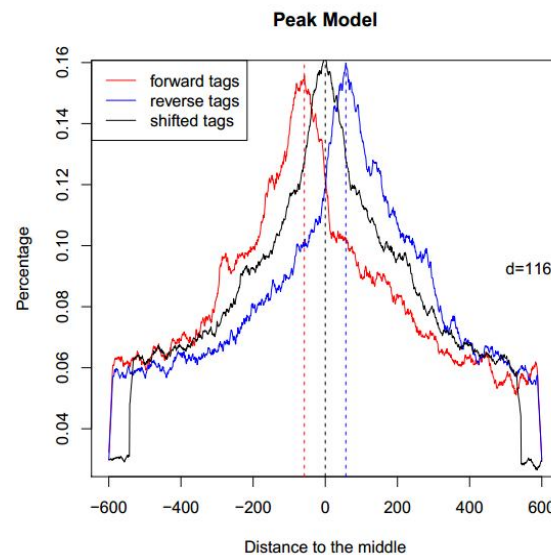
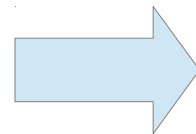
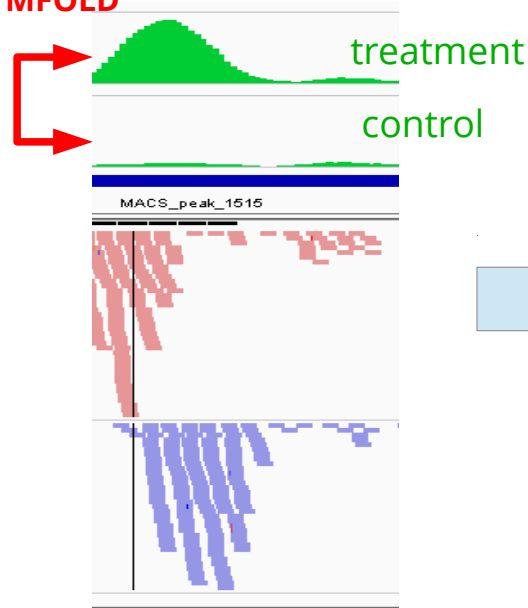
- sliding window across the genome
- at each location, evaluate the enrichment of the signal wrt. expected background based on the distribution
- retain regions with P-values below threshold
- evaluate FDR



6. MACS [Zhang et al. Genome Biol. 2008]

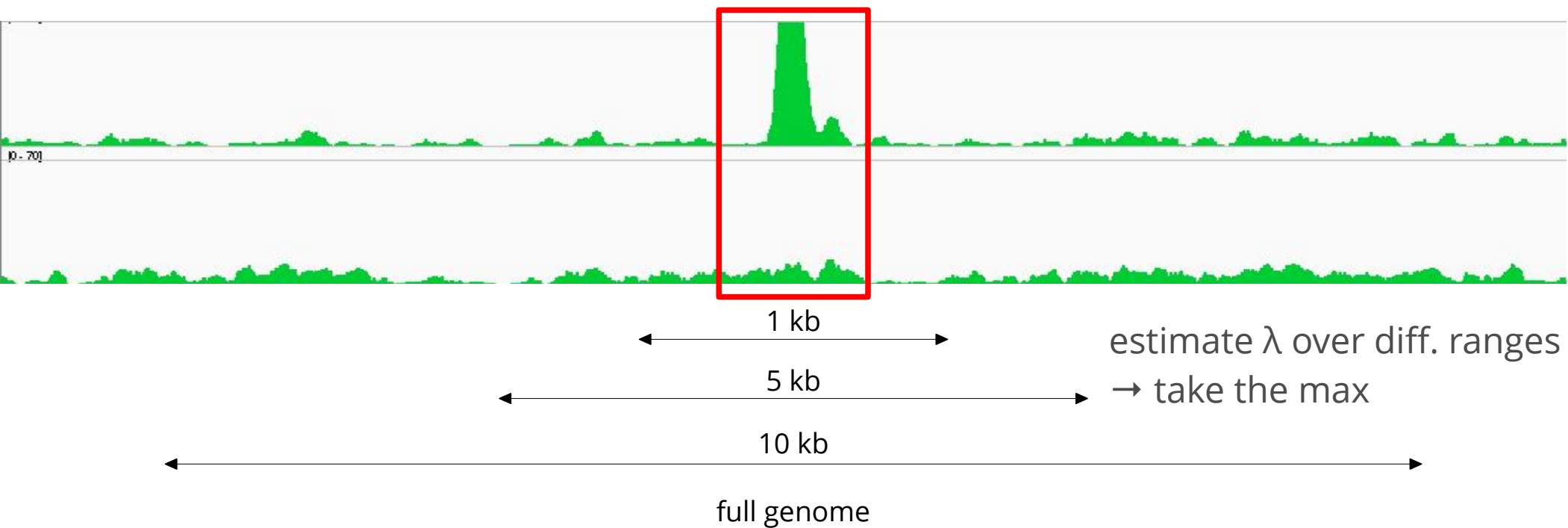
- **Step 1 : estimating fragment length d**
 - slide a window of size **BANDWIDTH**
 - retain top regions with **MFOLD** enrichment of treatment vs. input
 - plot average +/- strand read densities → estimate d

enrichment
> MFOLD



5. MACS [Zhang et al. Genome Biol. 2008]

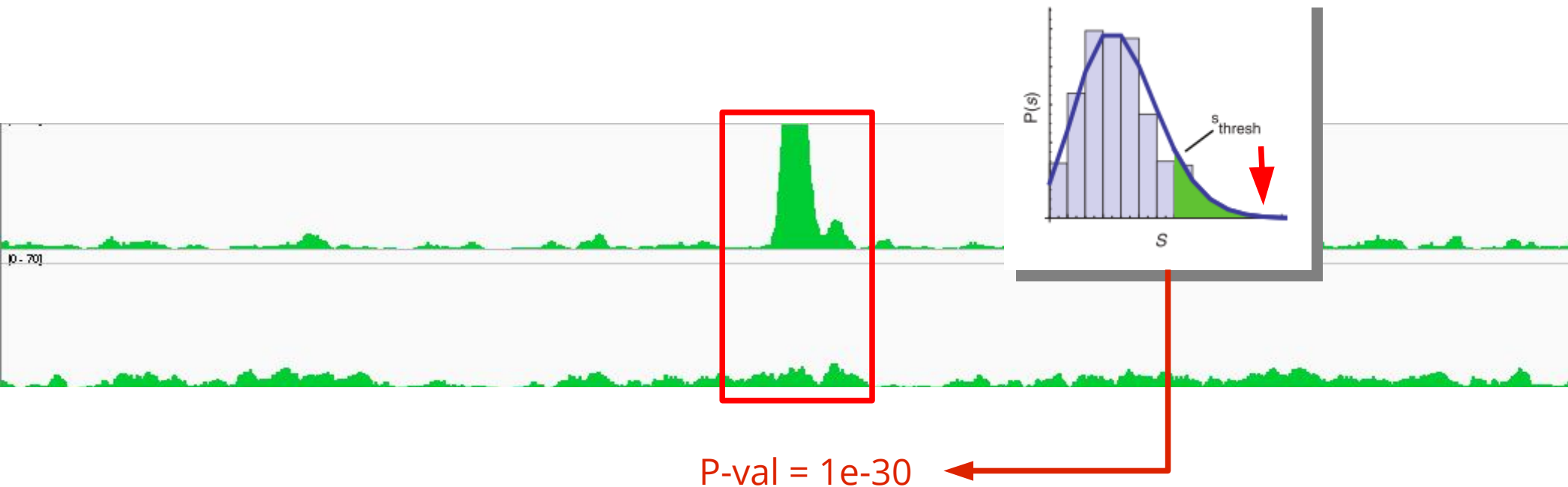
- **Step 2 : identification of local noise parameter**
 - slide a window of size $2*d$ across treatment and input
 - estimate parameter λ_{local} of Poisson distribution



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

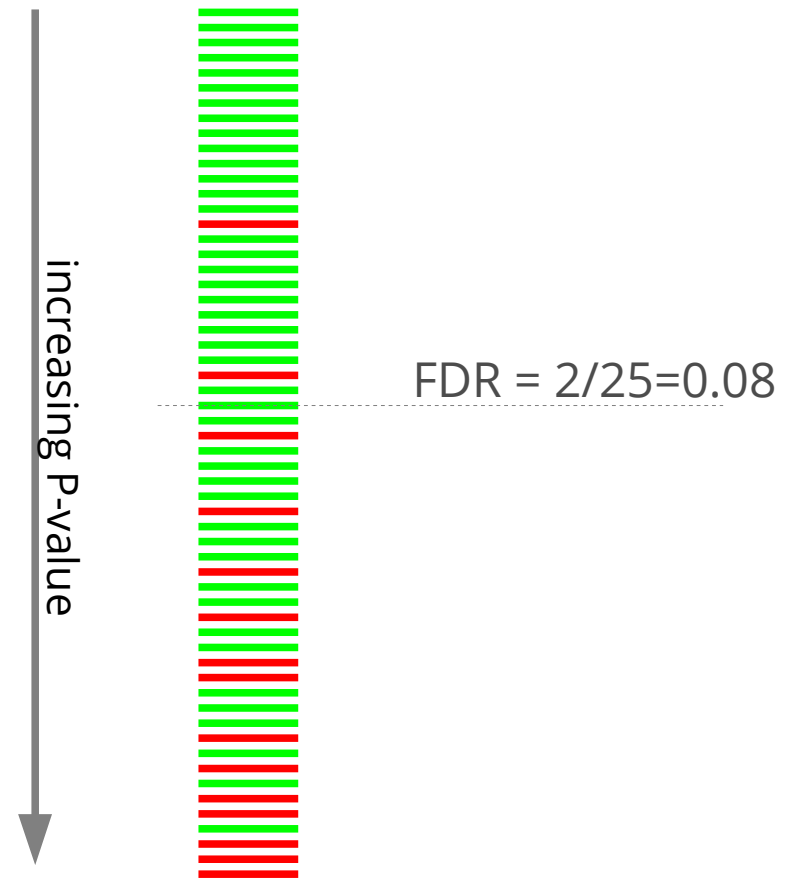


5. MACS [Zhang et al. Genome Biol. 2008]

- **Step 4 : estimating FDR**

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

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



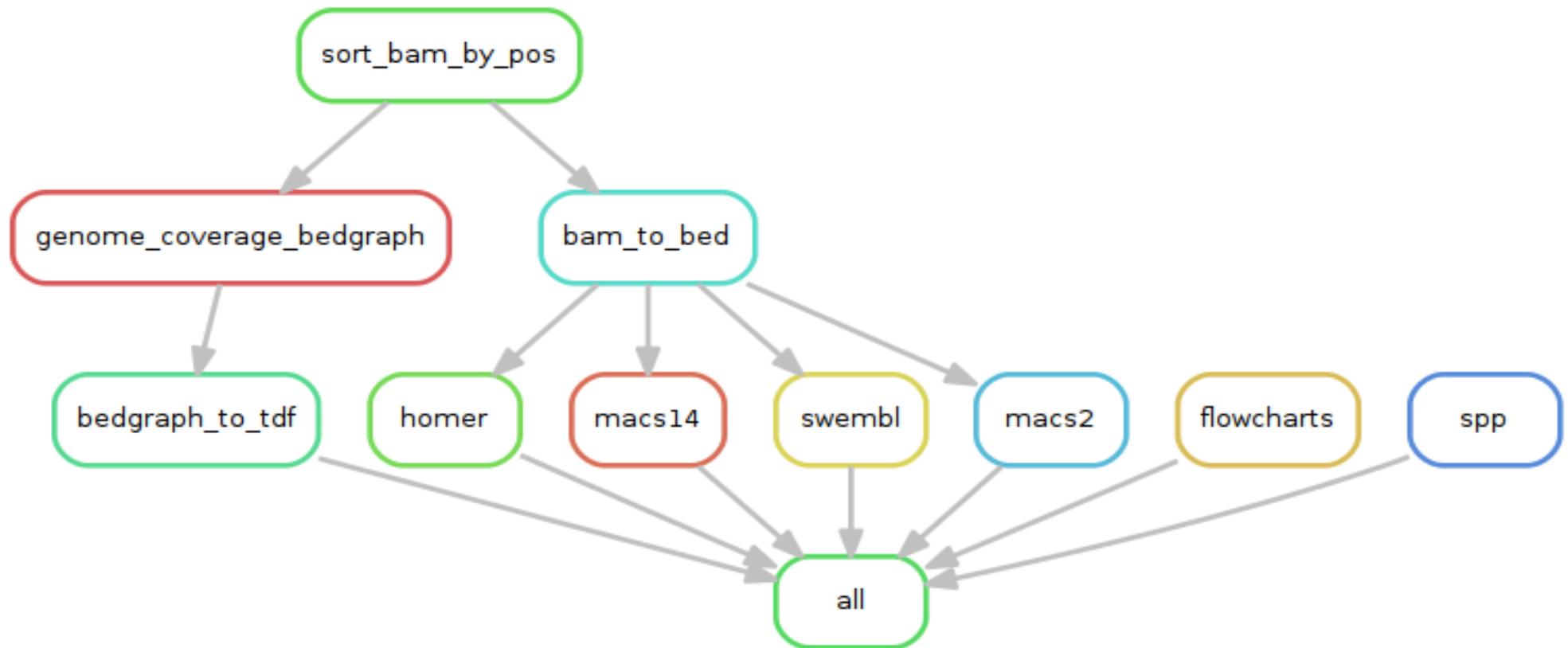
How to choose a peak-caller and tune its parameters ?

- More than 30 peak-callers are currently available (see Pepke et al., 2009 for a review of some of these).
- Some are dedicated to broad peaks (e.g. SICER), others to narrow peaks (e.g. SWEMBL).
- Some peak-caller have different settings for narrow peaks and broad peaks (e.g. MACS, SPP).
- Thresholds affect drastically the number of peaks, and their width (small neighbour peaks can be regrouped in a wider peak).
- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !

Pepke,S., Wold,B. and Mortazavi,A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6, S22–32.

Comparison between peak callers

- Castro-Mondragon, Claire Rioualen, Bruno Contreras and Jacques van Helden (2016). MiMB.
- Workflow to compare peak-callers.



Mutual overlap between peaks (numbers of peaks)

- Castro-Mondragon, Claire Rioualen, Bruno Contreras and Jacques van Helden (2016). MiMB.

Peak-caller	ID	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
Macs2 (default)	f1	2931	3335	2699	2854	3408	532	1298	2224	2704	2848	3263
Macs2 (custom)	f2	2930	9711	5494	8136	9576	544	1340	2767	5020	8840	11973
Macs14 (custom)	f3	2895	7360	6242	7359	9767	535	1325	2659	4510	9518	17225
Homer	f4	2851	8884	6114	18812	16898	534	1328	2743	5091	17125	24503
SPP (fdr=0.01)	f5	2920	9291	6166	15364	24781	544	1333	2751	5104	20399	39018
SPP (fdr=0.001)	f6	534	640	533	534	680	544	532	533	536	557	654
SWEMBL (R=0.1; x=1.2)	f7	1374	1764	1294	1377	1786	534	1352	1343	1340	1368	1502
SWEMBL (R=0.07; x=1.2)	f8	2355	3606	2561	2861	3679	535	1352	2788	2765	2864	3424
SWEMBL (R=0.05; x=1.2)	f9	2852	6302	4224	5316	6697	538	1352	2787	5256	5541	7277
SWEMBL (R=0.02; x=1.2)	f10	2931	9692	6236	16734	20518	544	1352	2788	5256	31867	41904
SWEMBL (R=0.01; x=1.2)	f11	2931	9710	6242	18611	24365	544	1352	2788	5256	31864	92695

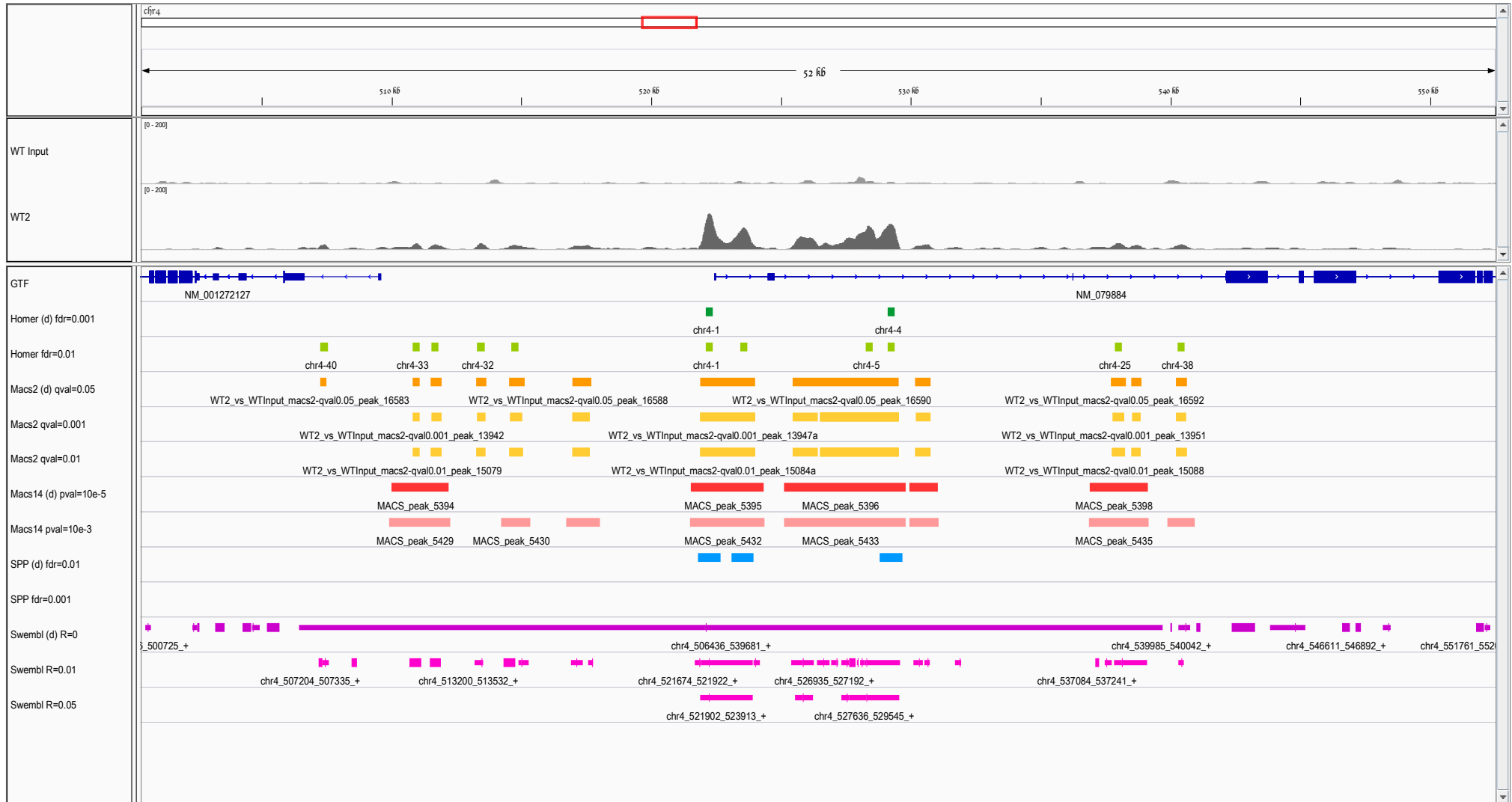
Mutual overlap between peaks (numbers of peaks)

- Castro-Mondragon, Claire Rioualen, Bruno Contreras and Jacques van Helden (2016). MiMB.

Peak-caller	Peaks	ID	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11
Macs2 (default)	2931	f1	1	0.34	0.43	0.15	0.14	0.98	0.96	0.8	0.51	0.09	0.04
Macs2 (custom)	9711	f2	1	1	0.88	0.43	0.39	1	0.99	0.99	0.96	0.28	0.13
Macs14 (custom)	6242	f3	0.99	0.76	1	0.39	0.39	0.98	0.98	0.95	0.86	0.3	0.19
Homer	18812	f4	0.97	0.91	0.98	1	0.68	0.98	0.98	0.98	0.97	0.54	0.26
SPP (fdr=0.01)	24781	f5	1	0.96	0.99	0.82	1	1	0.99	0.99	0.97	0.64	0.42
SPP (fdr=0.001)	544	f6	0.18	0.07	0.09	0.03	0.03	1	0.39	0.19	0.1	0.02	0.01
SWEMBL (R=0.1; x=1.2)	1352	f7	0.47	0.18	0.21	0.07	0.07	0.98	1	0.48	0.25	0.04	0.02
SWEMBL (R=0.07; x=1.2)	2788	f8	0.8	0.37	0.41	0.15	0.15	0.98	1	1	0.53	0.09	0.04
SWEMBL (R=0.05; x=1.2)	5256	f9	0.97	0.65	0.68	0.28	0.27	0.99	1	1	1	0.17	0.08
SWEMBL (R=0.02; x=1.2)	31867	f10	1	1	1	0.89	0.83	1	1	1	1	1	0.45
SWEMBL (R=0.01; x=1.2)	92695	f11	1	1	1	0.99	0.98	1	1	1	1	1	1

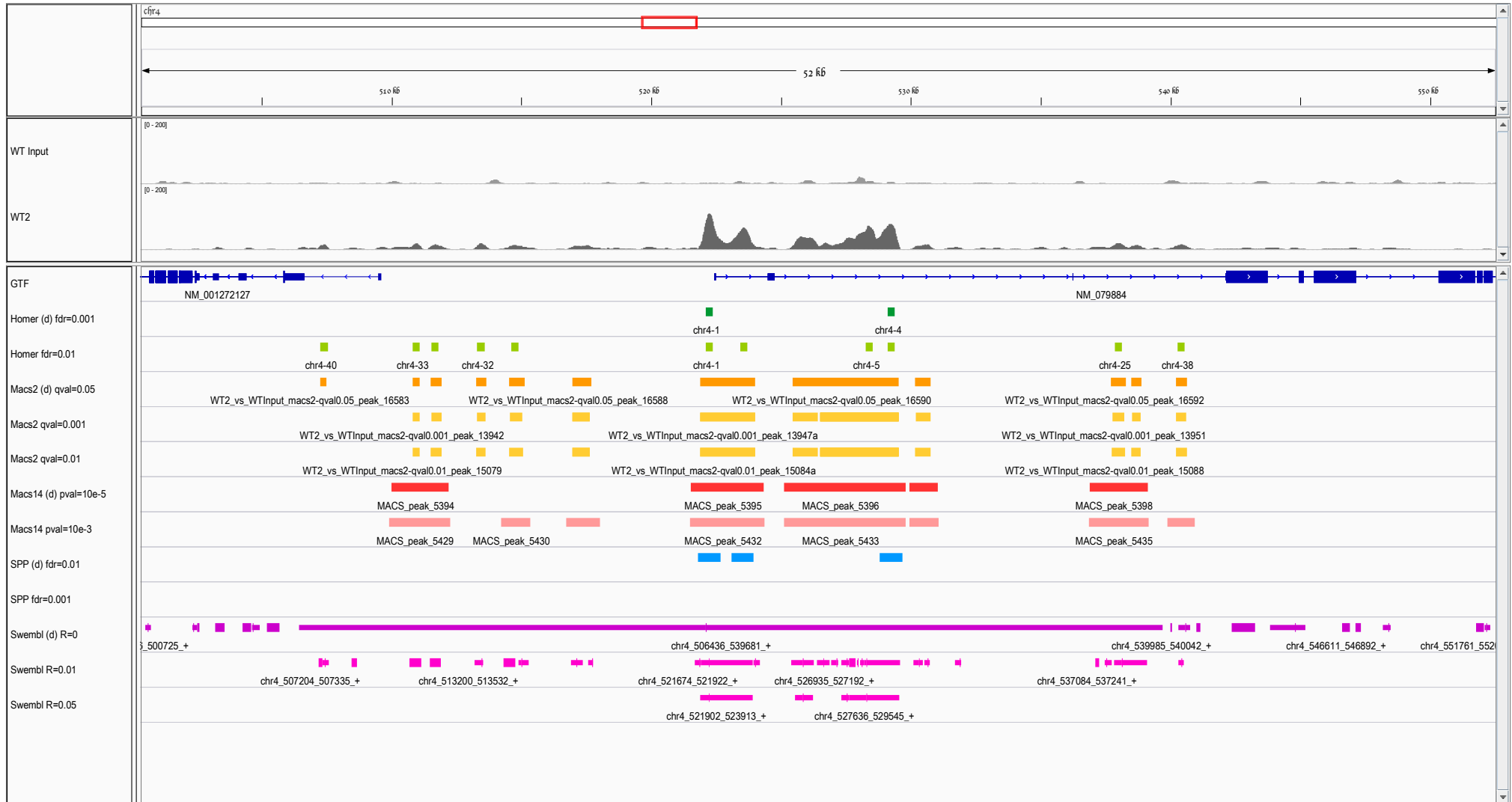
Peak-calling choices do matter !

- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !
- Figure from Claire Rioualen, 2016, data from Andy Saurin (Abd-A in *D.melanogaster*)



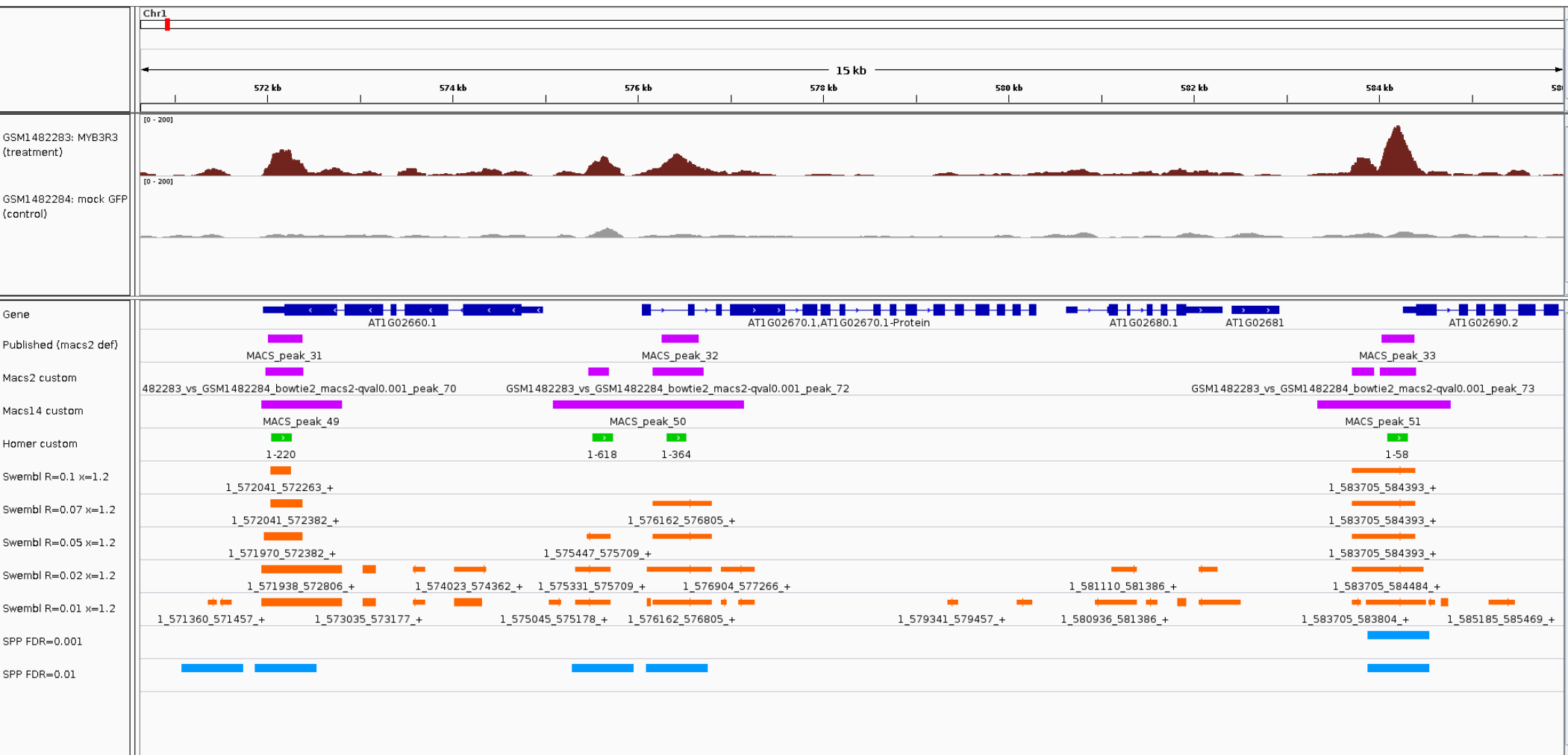
Peak-calling choices do matter !

- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !
- Figure from Claire Rioualen, 2016, data from Andy Saurin (Abd-A in *D.melanogaster*)



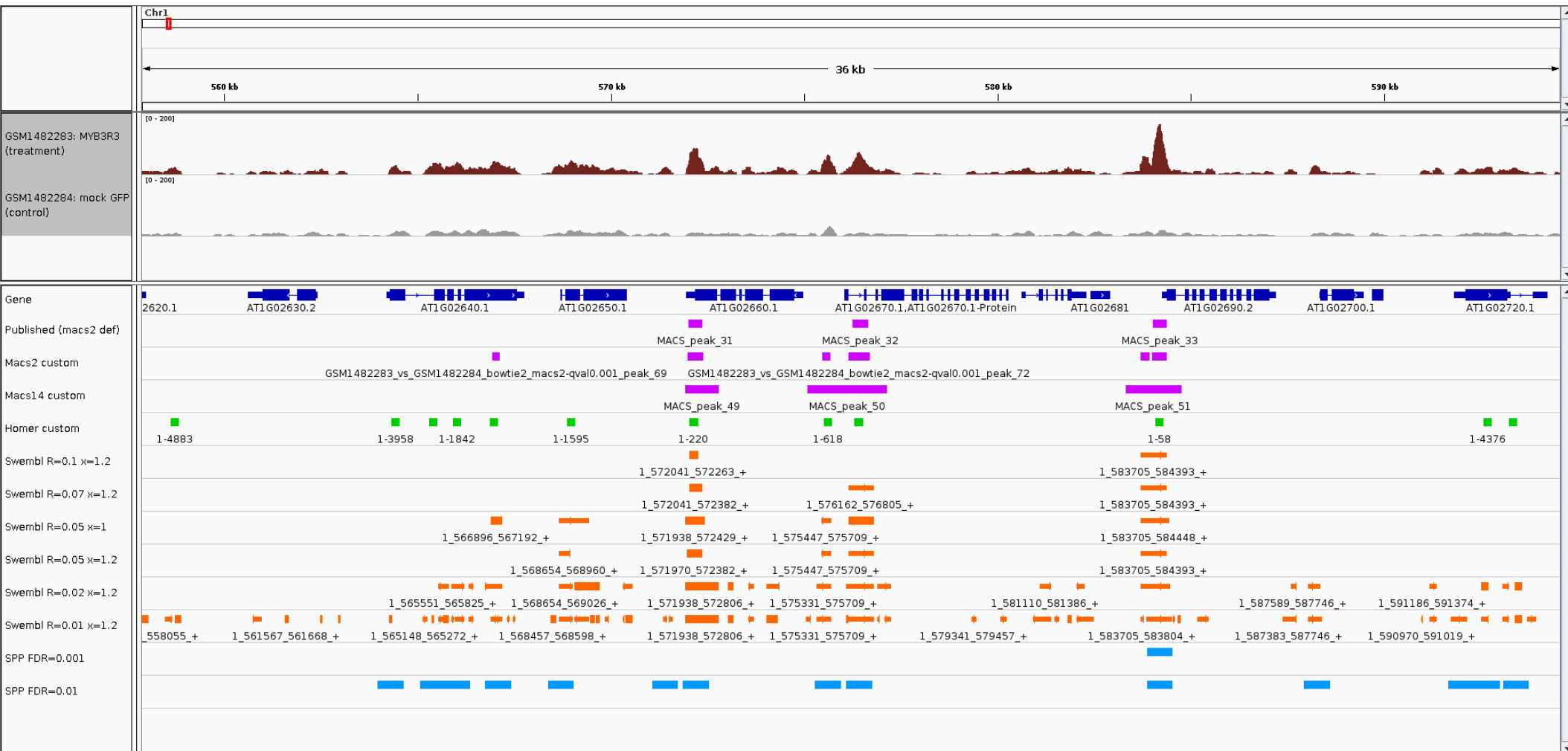
Peak-calling choices do matter !

- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !
- Figure from Claire Rioualen, 2016, MYB3R3 ChIP-seq in *A.thaliana*



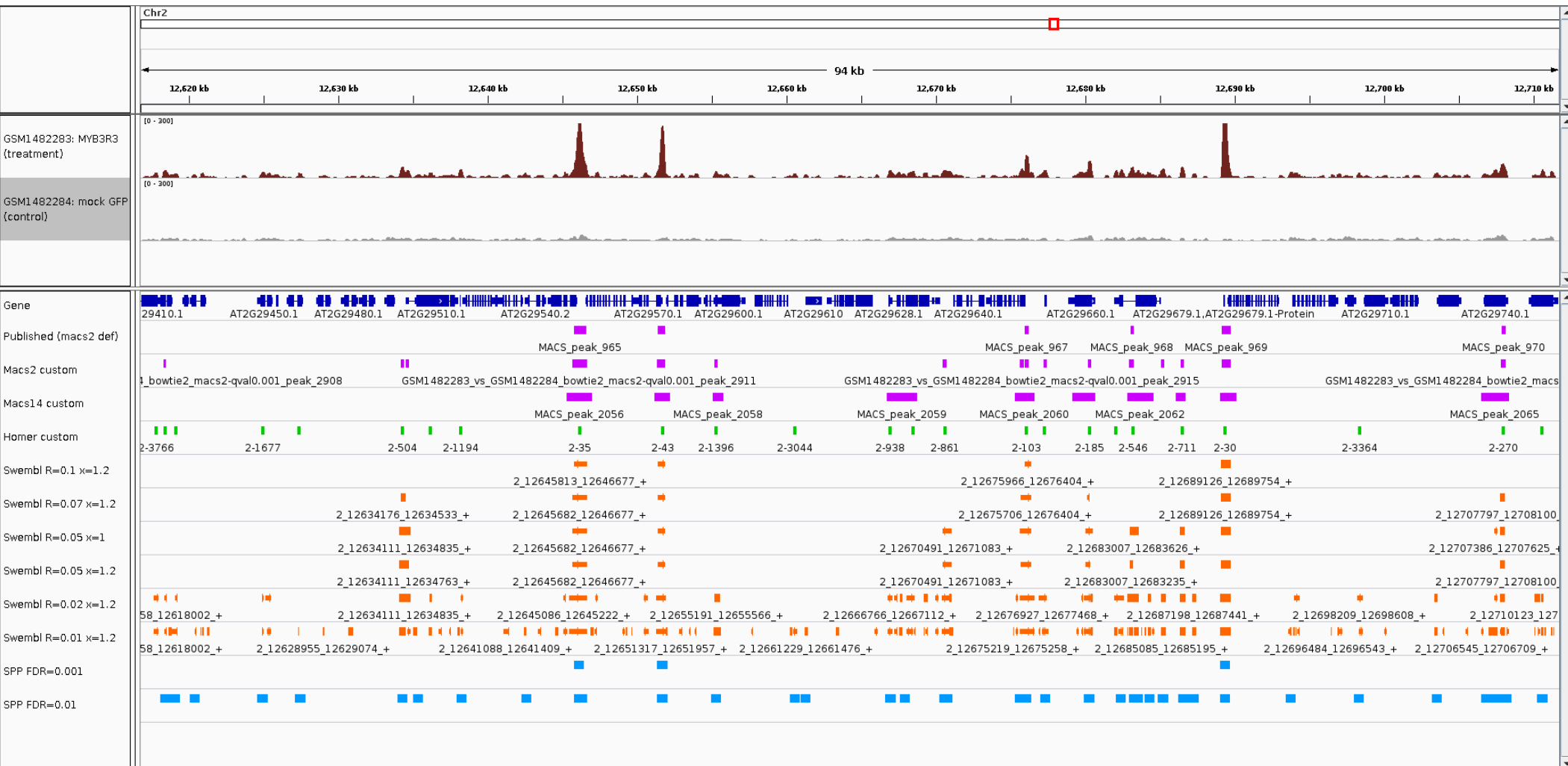
Peak-calling choices do matter !

- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !
- Figure from Claire Rioualen, 2016, MYB3R ChIP-seq in *A.thaliana*



Peak-calling choices do matter !

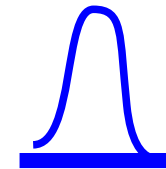
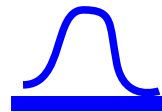
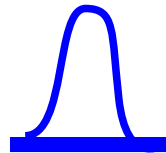
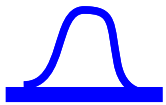
- With the same dataset, the number of peaks can vary from a few hundreds to >100.000, depending on tuning parameters !
- Figure from Claire Rioualen, 2016, MYB3R ChIP-seq in *A.thaliana*



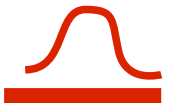
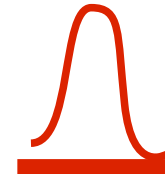
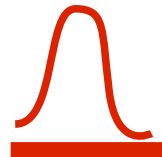
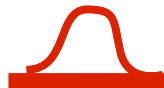
6. differential analysis

- 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

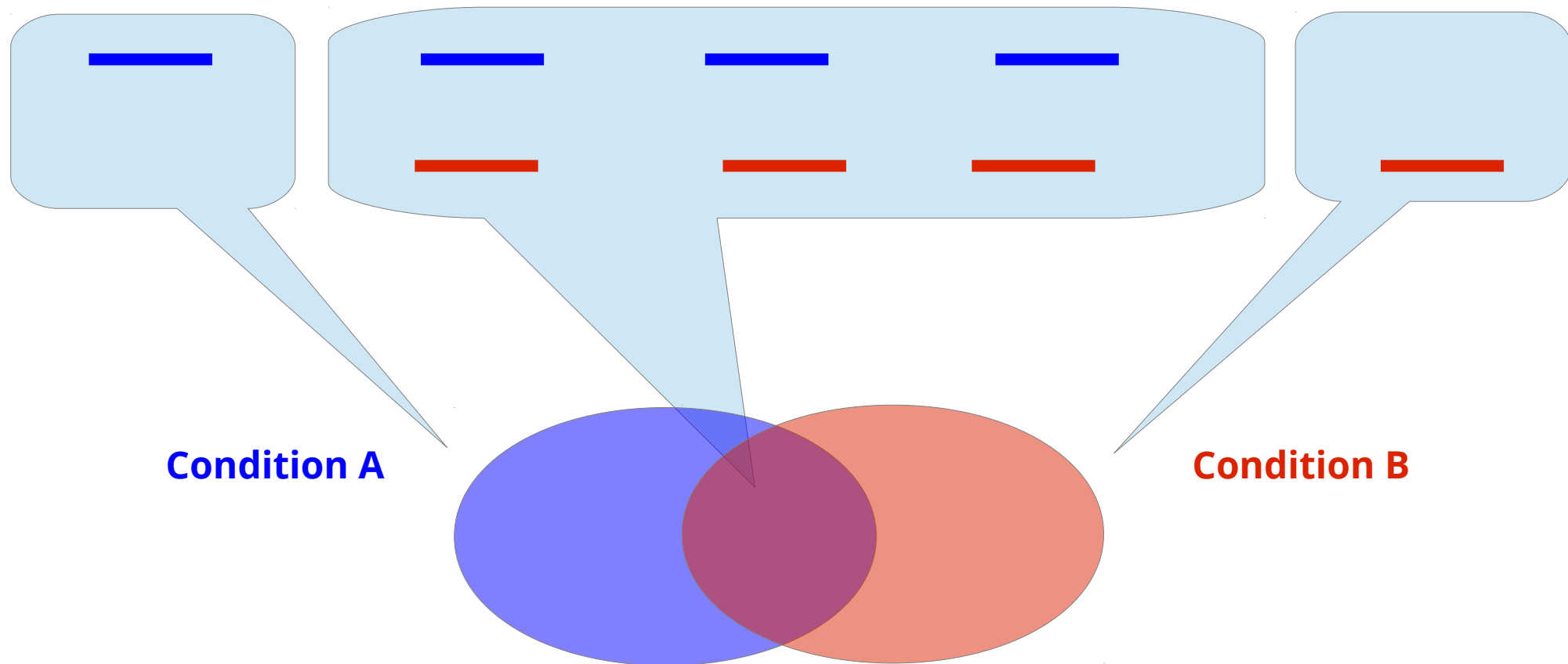
stronger
binding
in B

no difference

binding in B
no binding in A

6. differential analysis

- simple approach → compute common and specific peaks



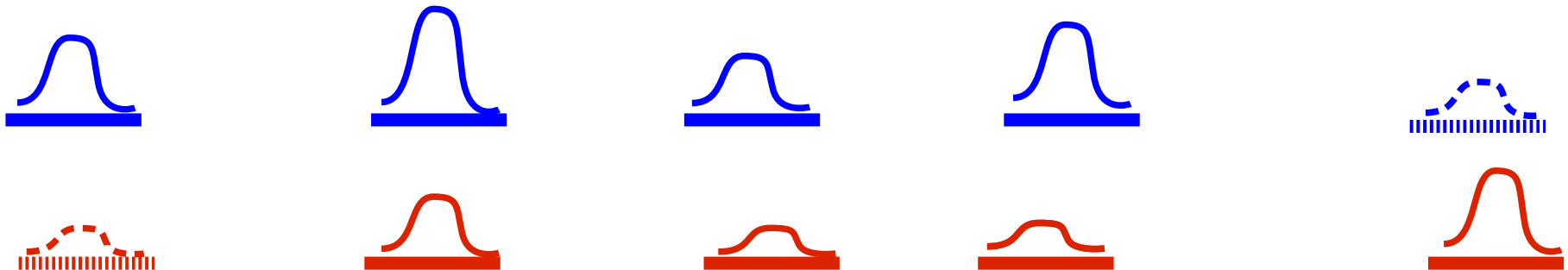
Drawback :

- 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**

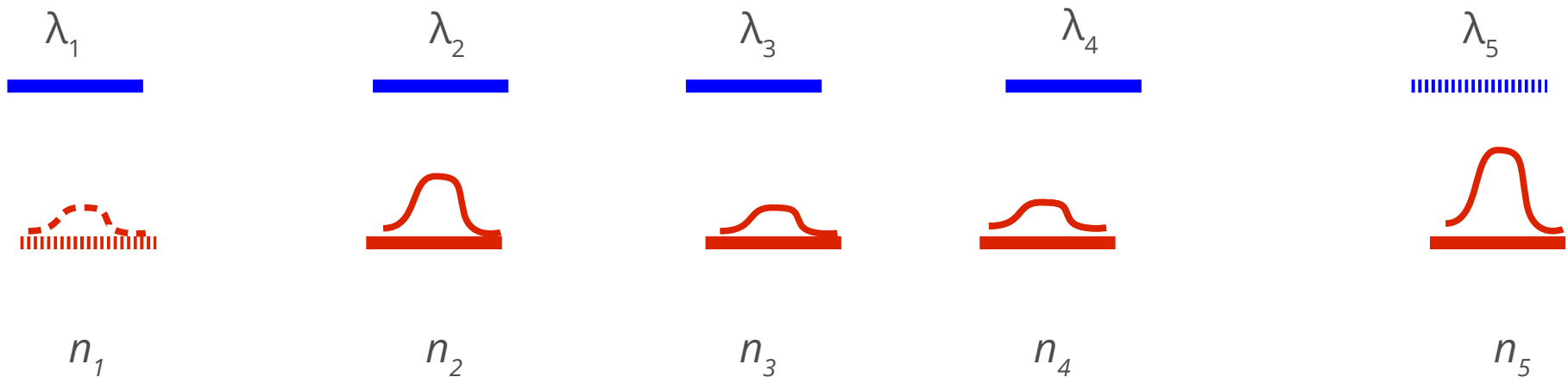
- **without replicates** : assume simple Poisson model (→ SICER-df)
- **with replicates** : perform differential test using DE tools from RNA-seq (diffBind using EdgeR, DESeq,...) based on read counts

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)

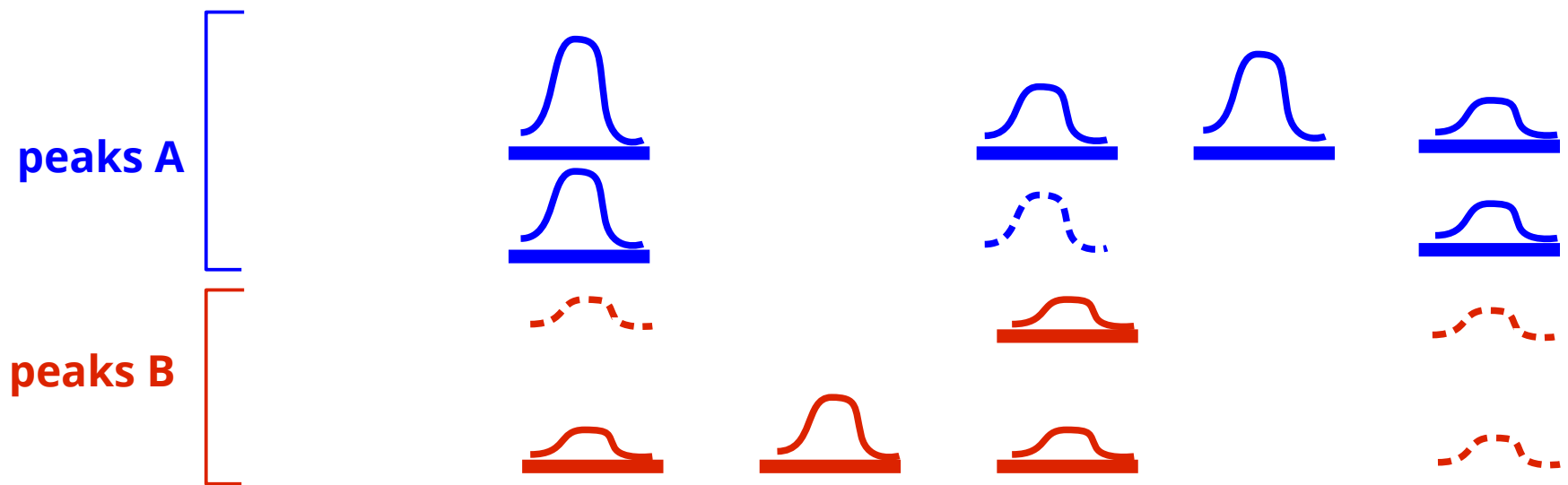
$$\lambda_i = w_i N_A / L_{eff}$$



6. differential analysis

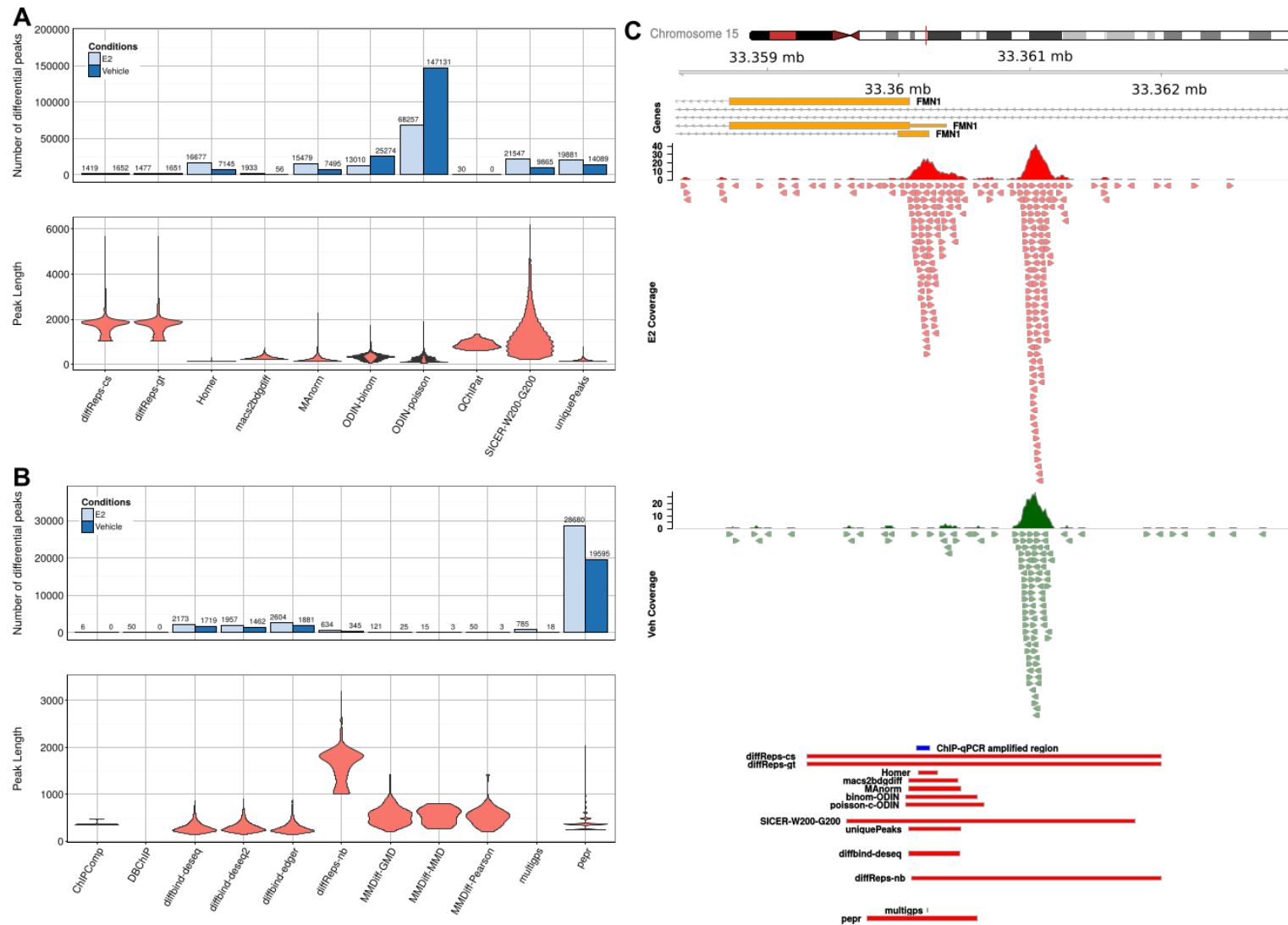
- **with replicates (diffBind)**

- 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)



consensus peaks (if $n \geq 2$)

6. differential analysis



Considerable differences in peak numbers and sizes !

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