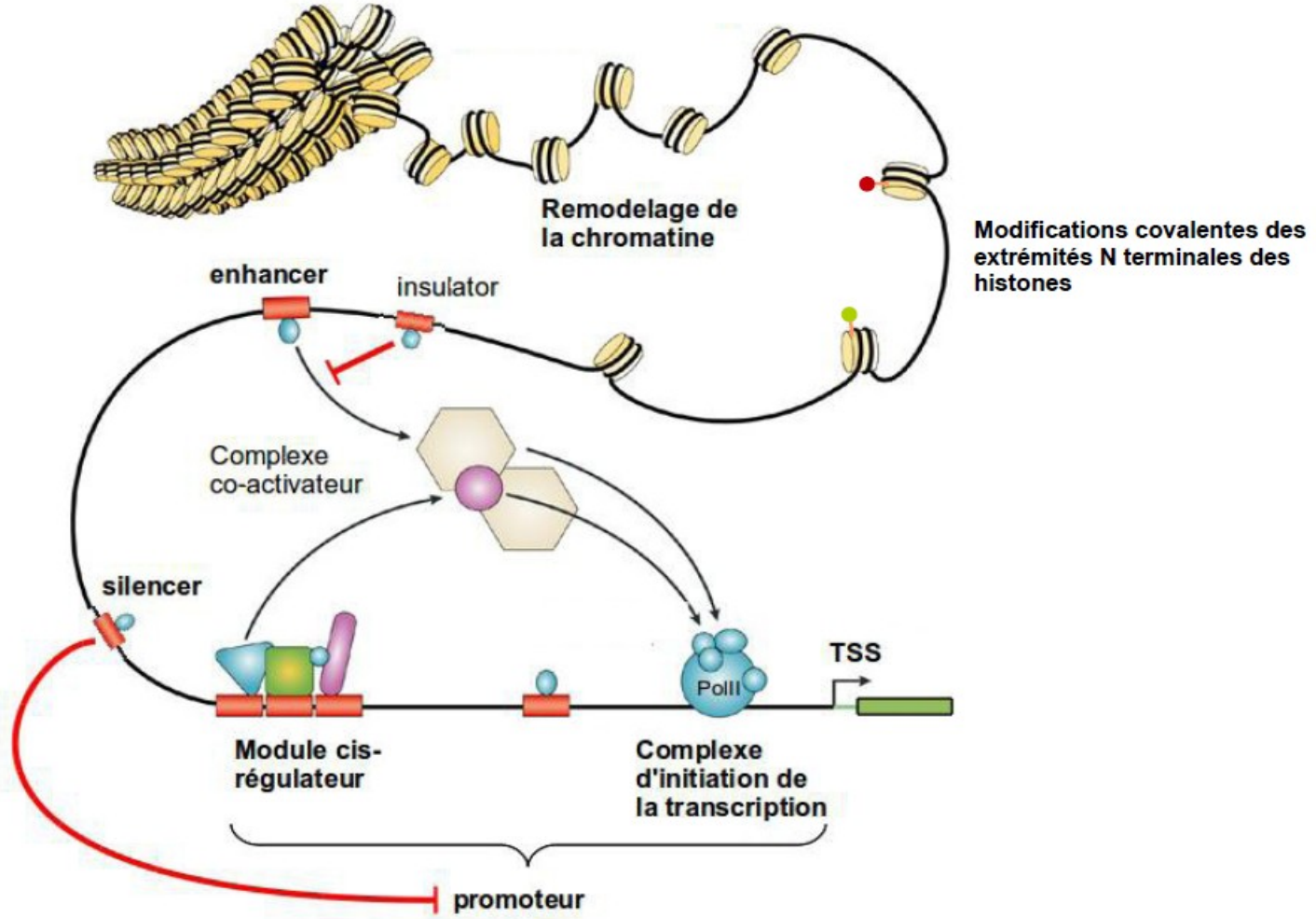
ChIP-seq analysis

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

D. Puthier TAGC – Inserm U1090

http://biow.sb-roscoff.fr/ecole_bioinfo/training_material/chip-seq/documents/presentation_chipseq.pdf

A model of transcriptional regulation

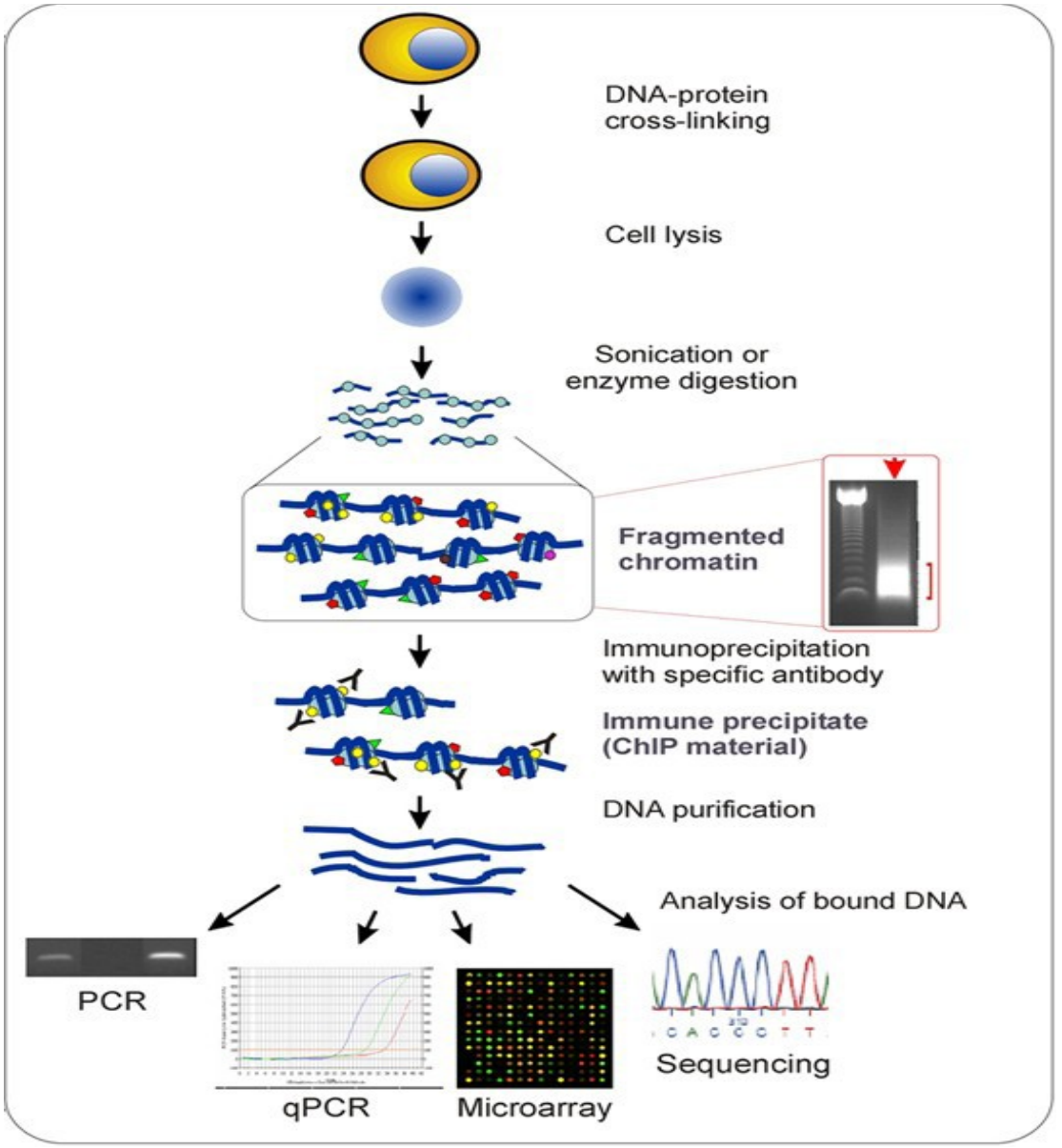


Nat Rev Genet. 2004 Apr;5(4):276-87.

Applied bioinformatics for the identification of regulatory elements.

Wasserman WW, Sandelin A.

Chromatine immuno-precipitation (ChIP)



- Used for:
  - TF localization
  - Histone modifications

ChIP-Seq: technical considerations

- Quality of antibodies: one of the most important factors ('ChIP grade')
  - High sensitivity
    - Fivefold enrichment by ChIP-PCR at several positive-control regions
  - High specificity
    - The specificity of an antibody can be directly addressed by immunoblot analysis (knockdown by RNA-mediated interference or genetic knockout)
  - Polyclonal antibodies may be prefered
    - Offer the flexibility of the recognition of multiple epitopes
- Cell Number
  - Typically
    - 1×10^6 (e.g, RNA polymerase II/histone modifications)
    - 10×10^6 (less-abundant proteins)

Nat Immunol. 2011 Sep 20;12(10):918-22. doi: 10.1038/ni.2117.

ChIP-Seq: technical considerations for obtaining high-quality data.

Kidder BL, Hu G, Zhao K.

ChIP-Seq: technical considerations

- Open chromatin regions are easier to shear
  - Higher background signals

- Two solutions
  - Isotype control antibodies
    - Immunoprecipitate much less DNA than specific antibodies
      - Overamplification of particular genomic regions during the library construction step (PCR)
  - Input
    - Non-ChIP genomic DNA
    - Better control

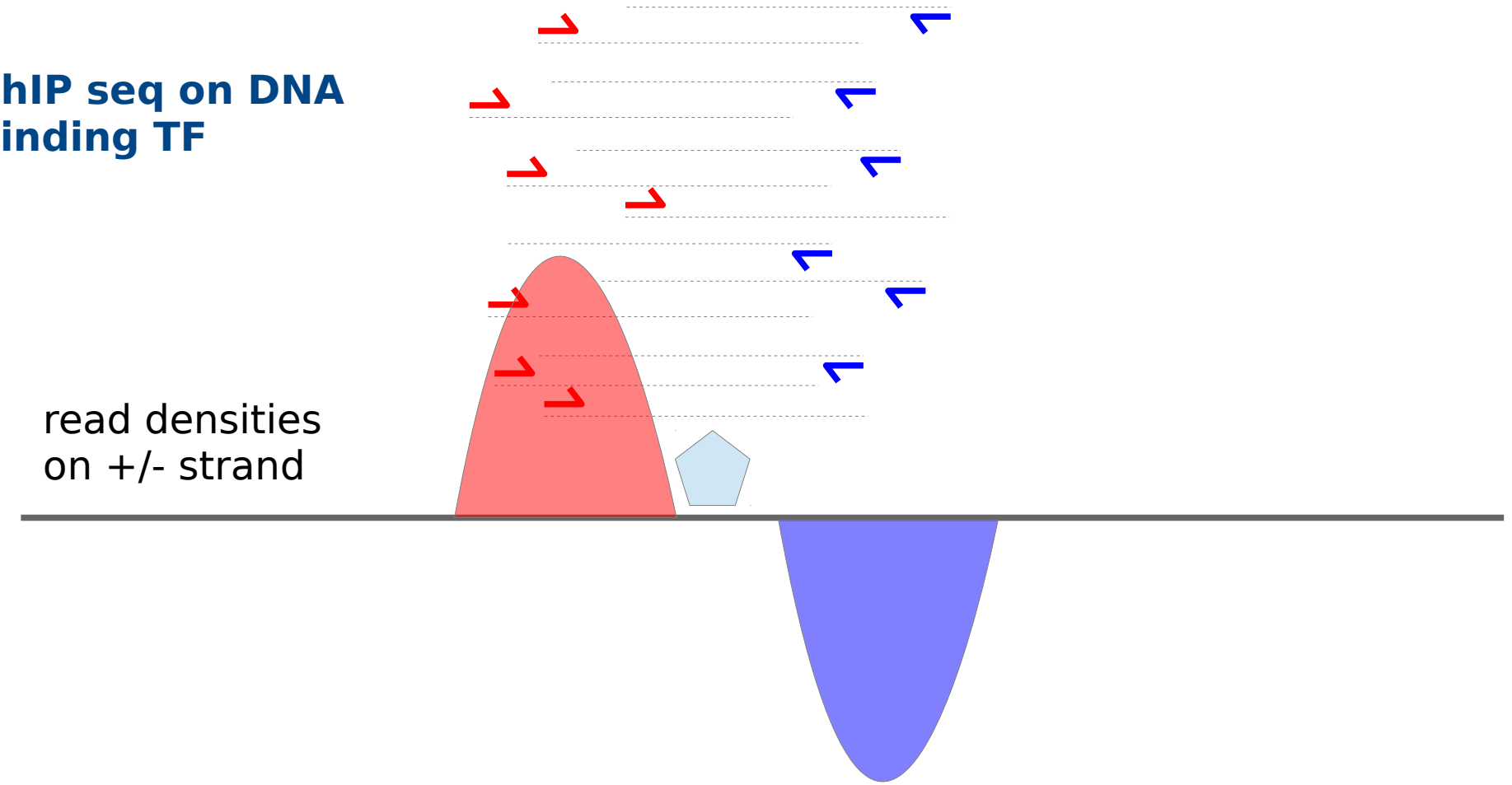
Nat Immunol. 2011 Sep 20;12(10):918-22. doi: 10.1038/ni.2117.

ChIP-Seq: technical considerations for obtaining high-quality data.

Kidder BL, Hu G, Zhao K.

ChIP-seq signal for transcription factors

ChIP seq on DNA binding TF

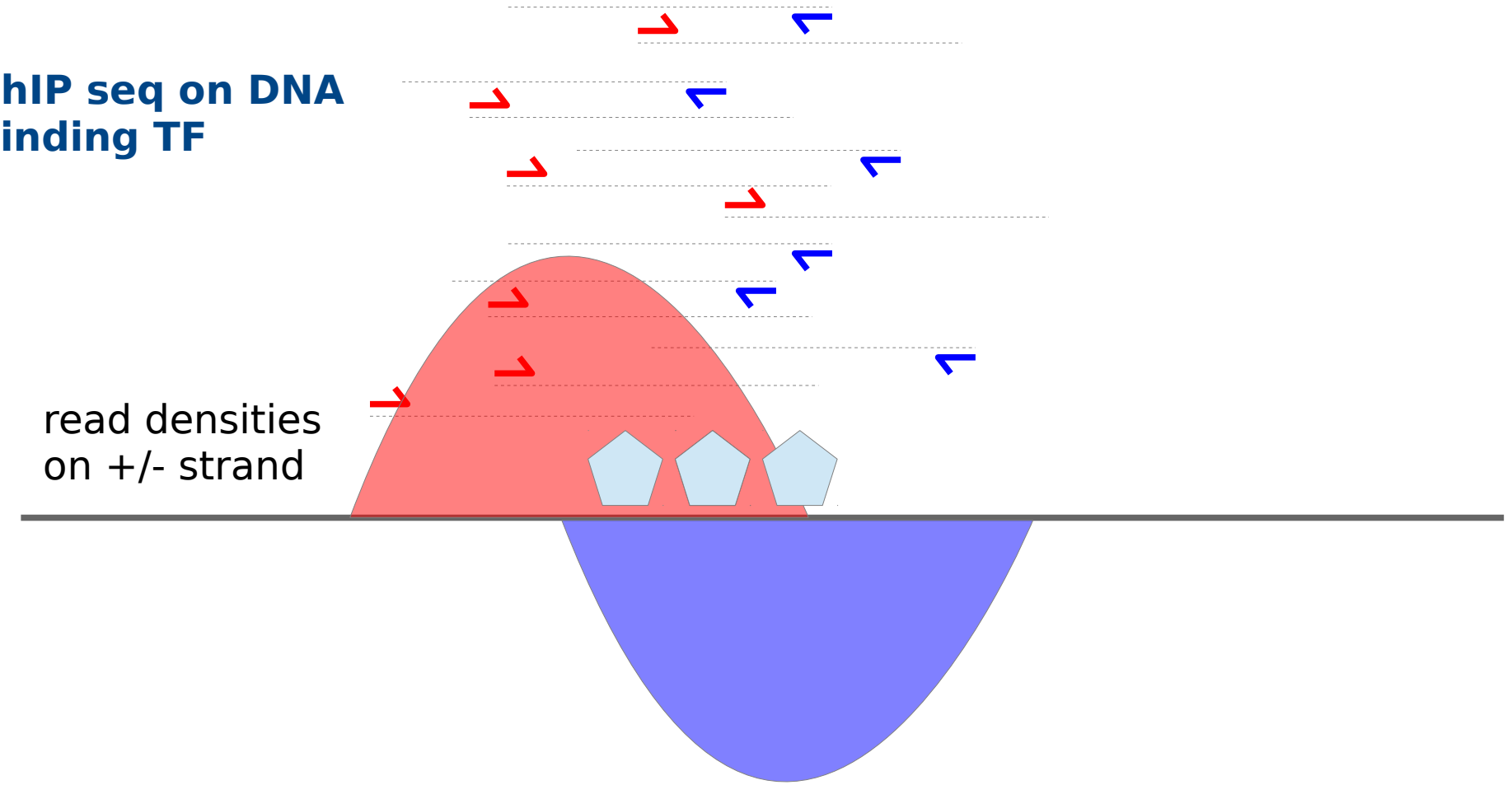


We expect to see a typical strand asymmetry in read densities
→ ChIP peak recognition pattern

ChIP-seq signal for transcription factors

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

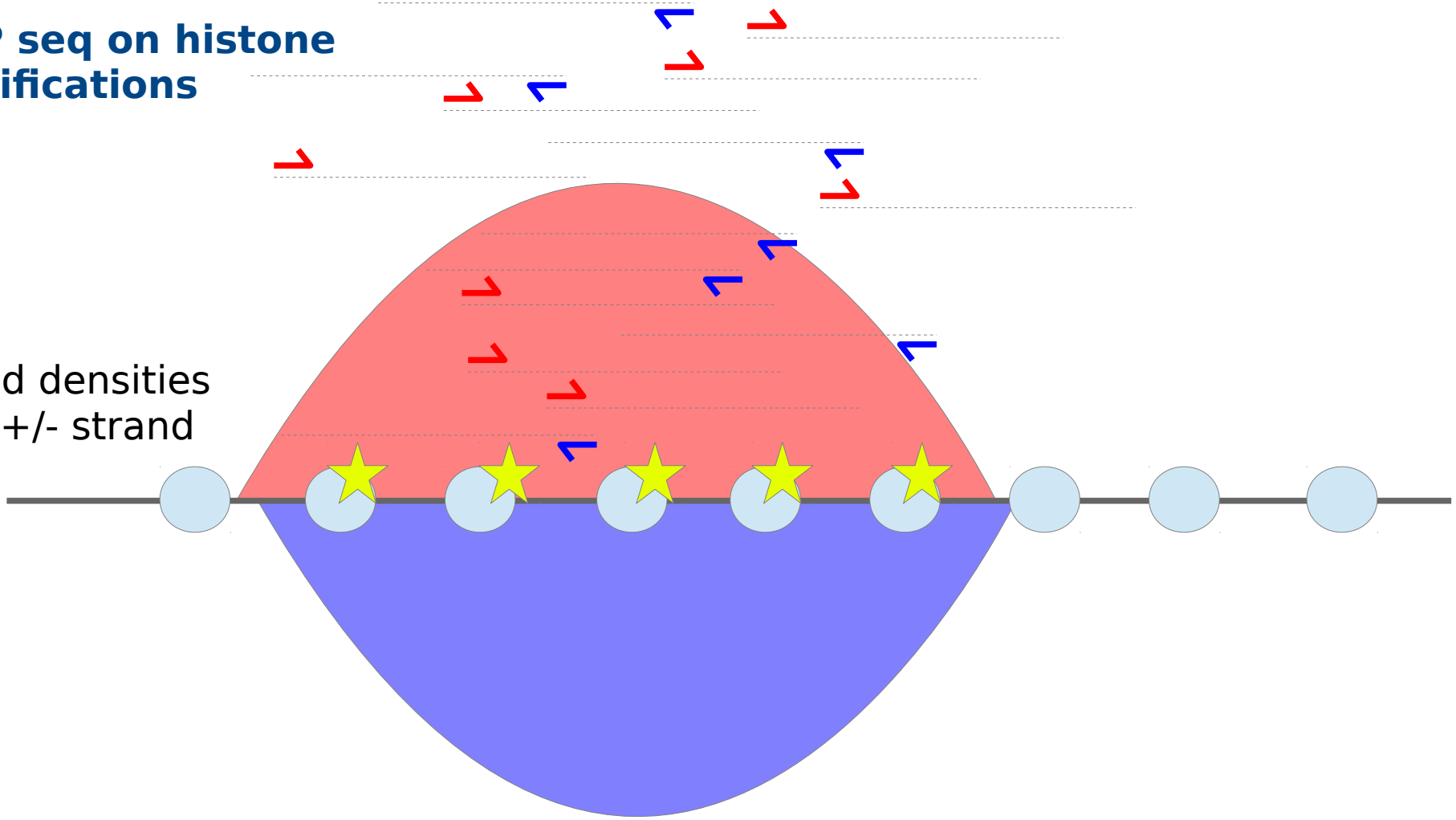
ChIP-seq signal for transcription factors



Binding of several TF as complexes tend to blur this asymmetry

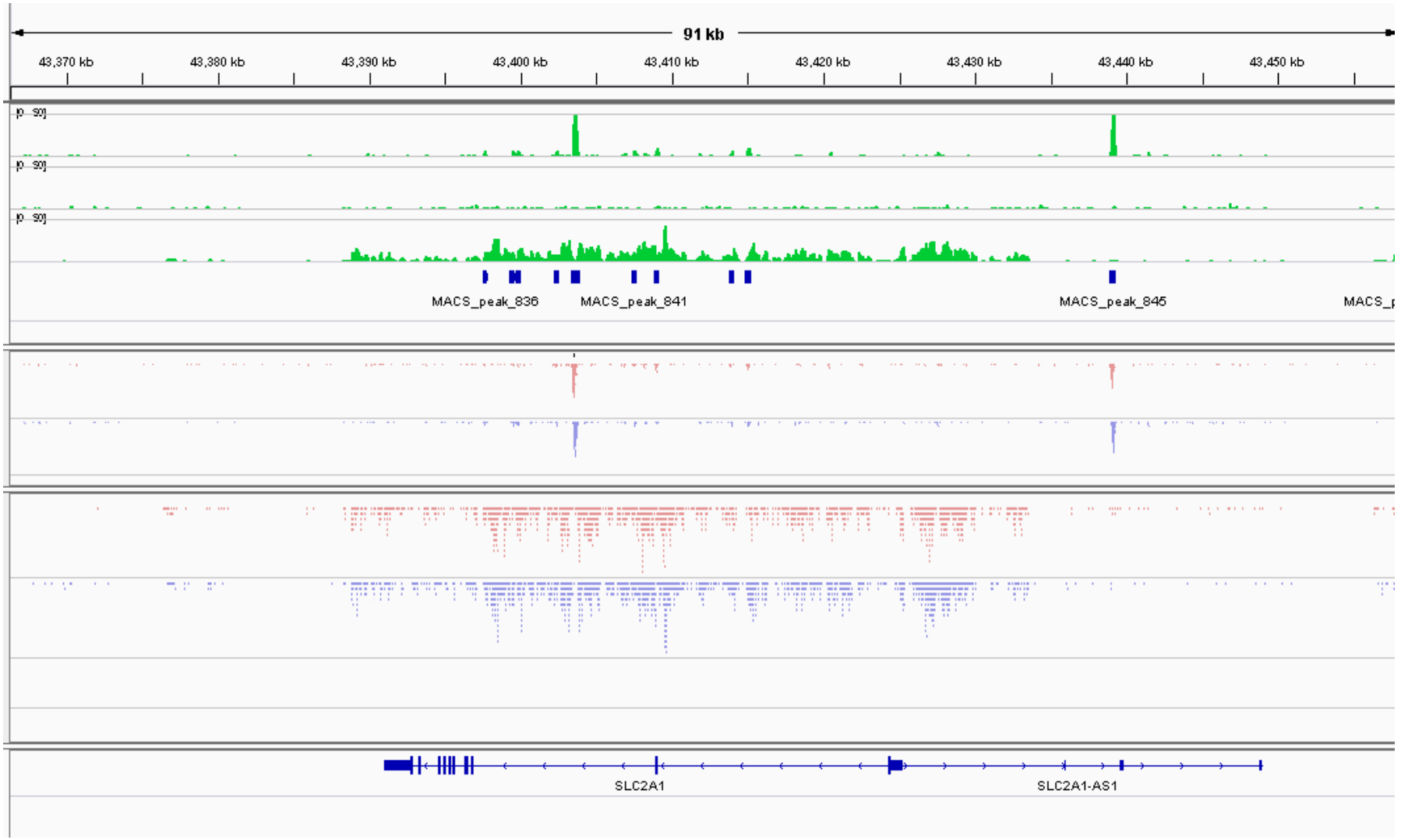
ChIP-seq signal for histone marks

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



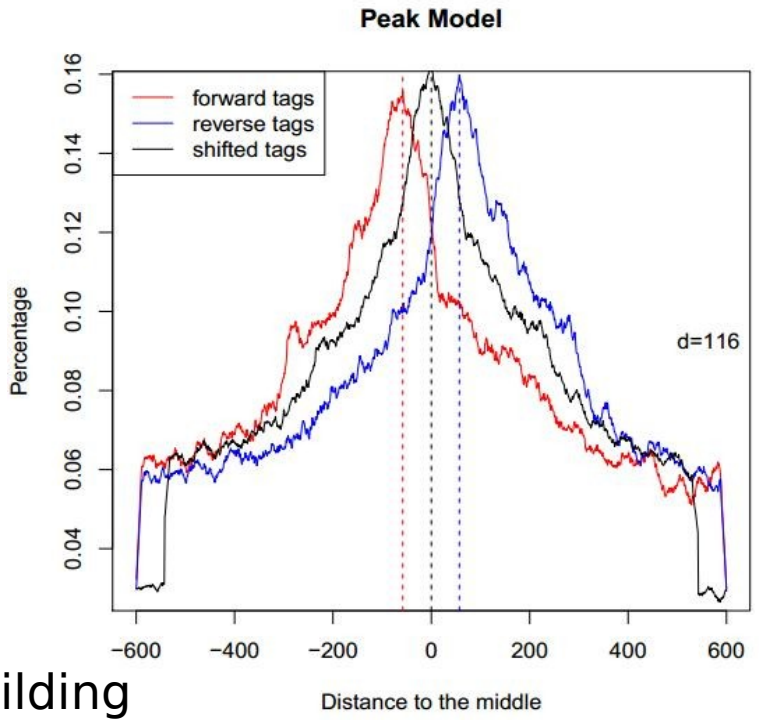
Keys aspects of “peak” finding

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

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

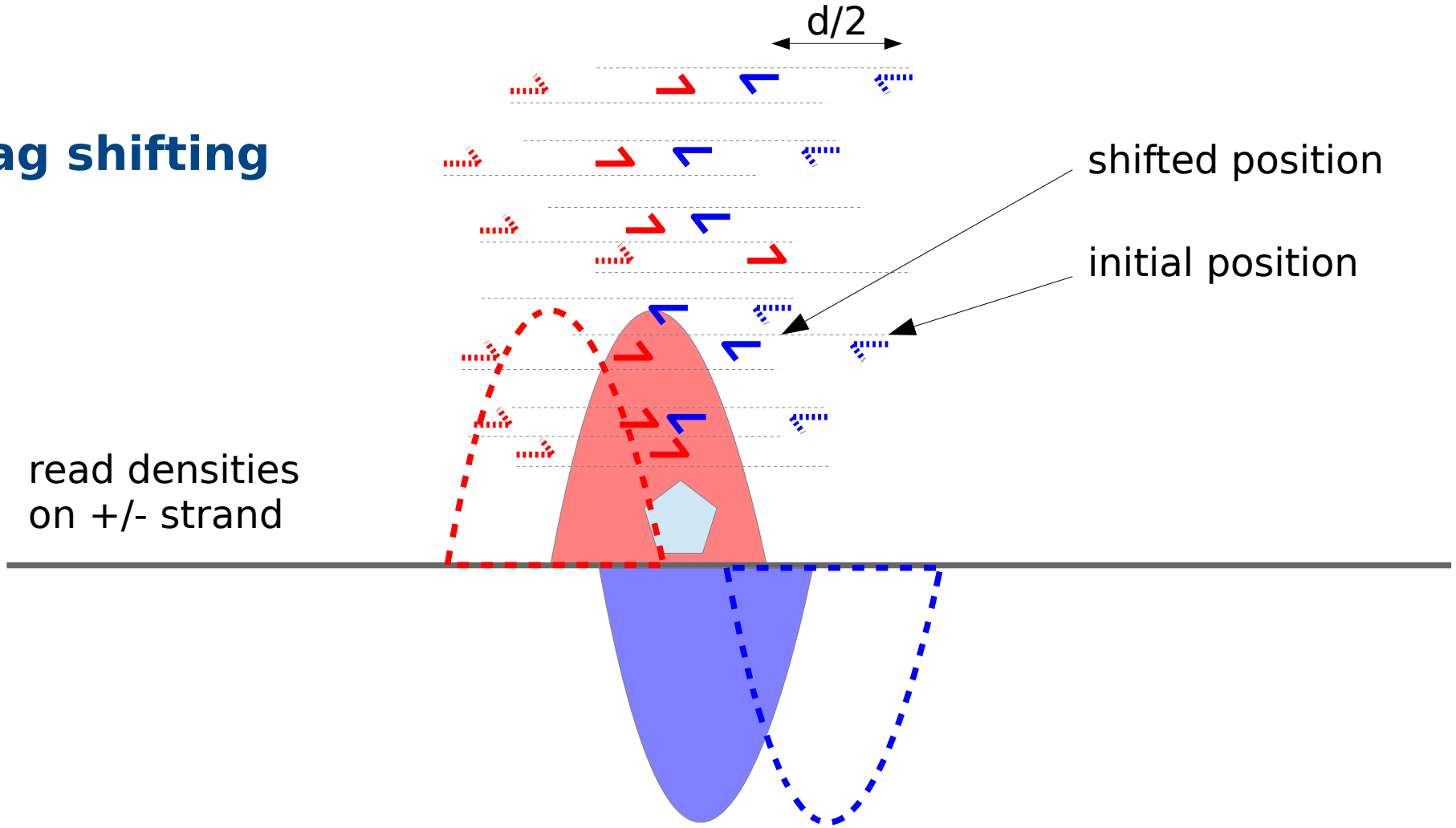
    (MACS) or given as input parameter



example of MACS model building using top enriched regions

From aligned reads to binding sites

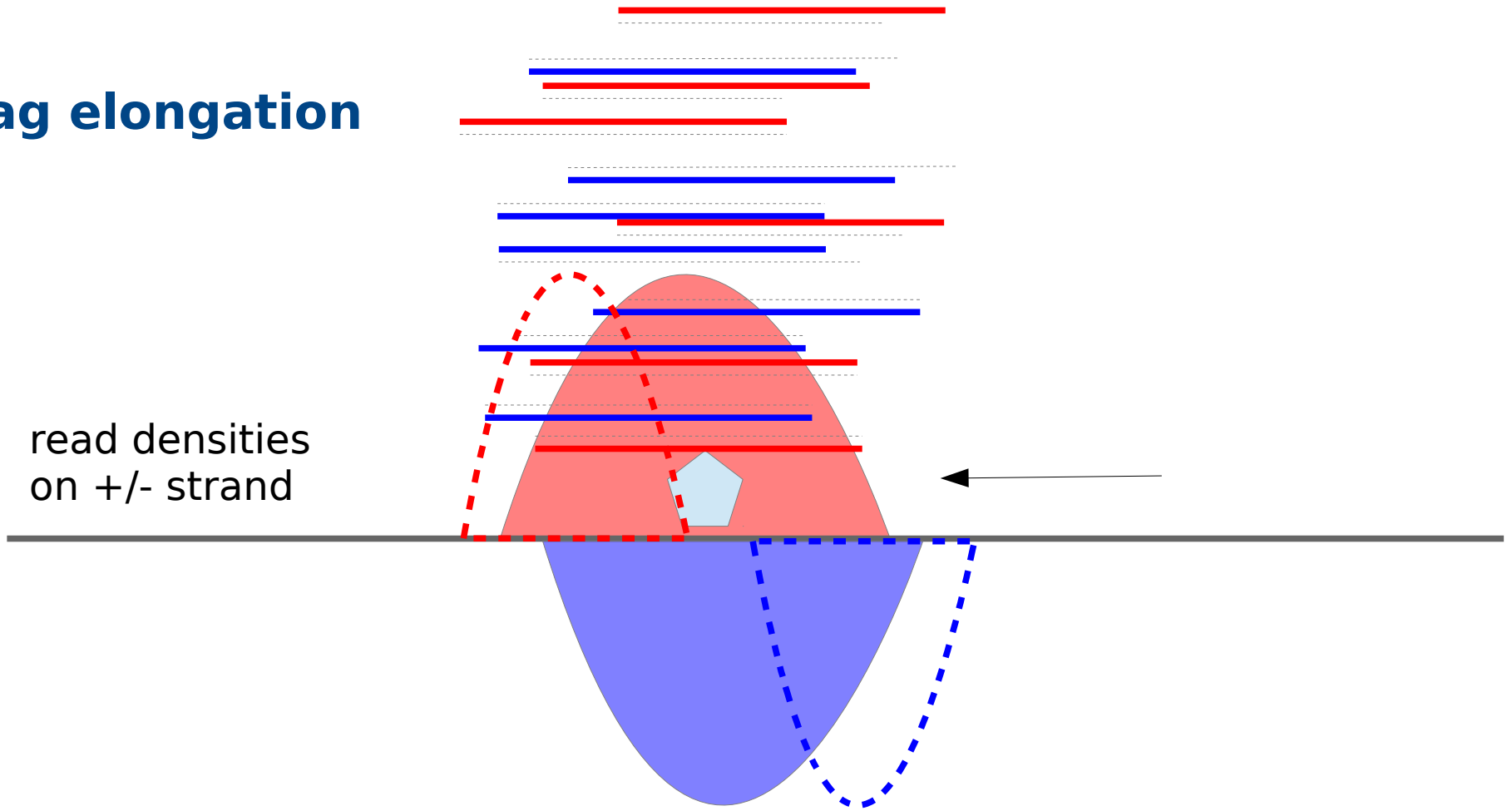
Tag shifting



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

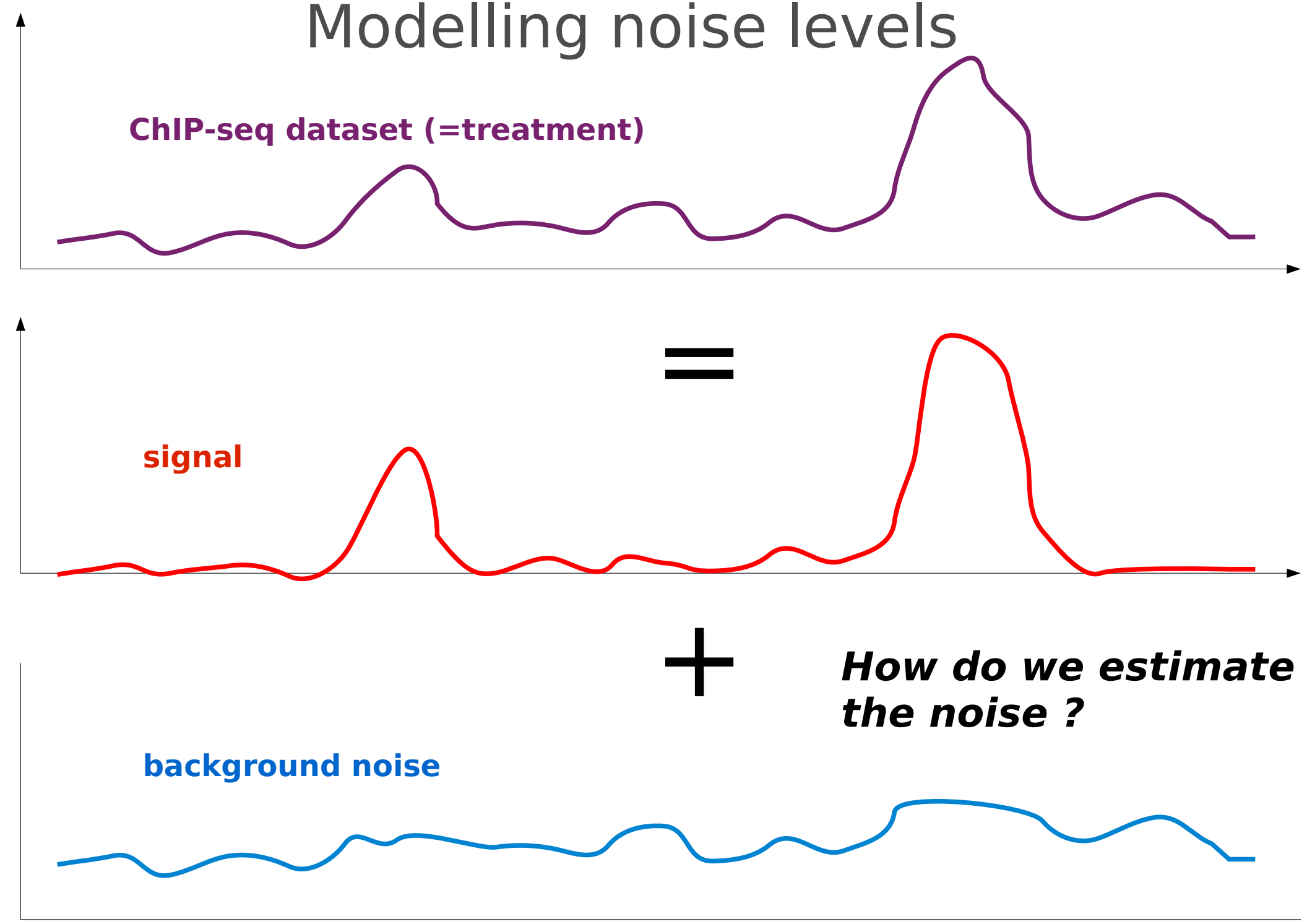
From aligned reads to binding sites

Tag elongation



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

Modelling noise levels



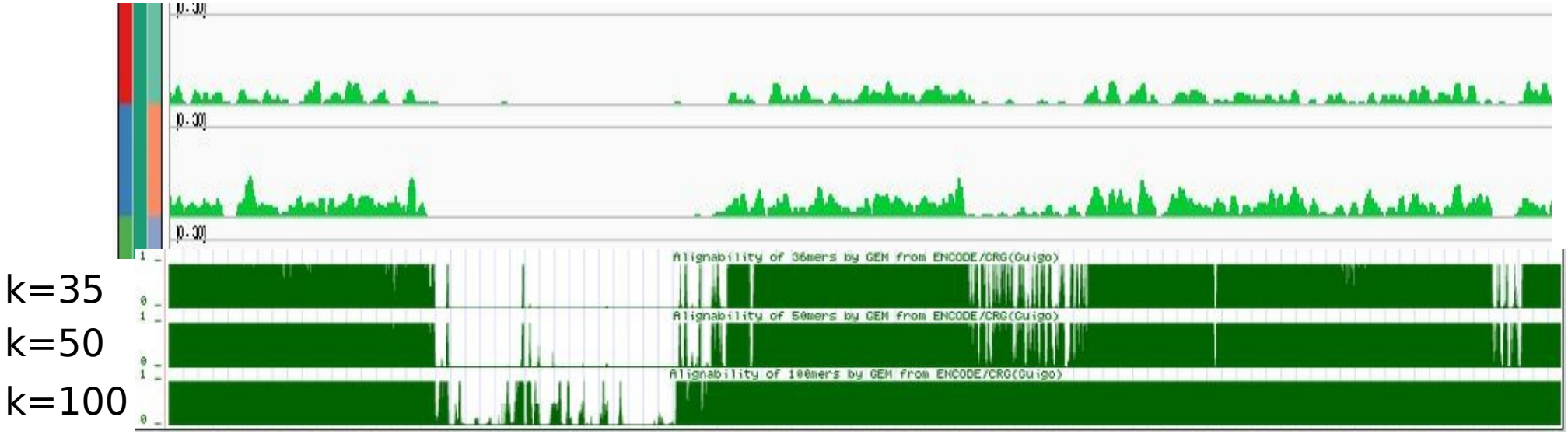
Modelling noise levels

- 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



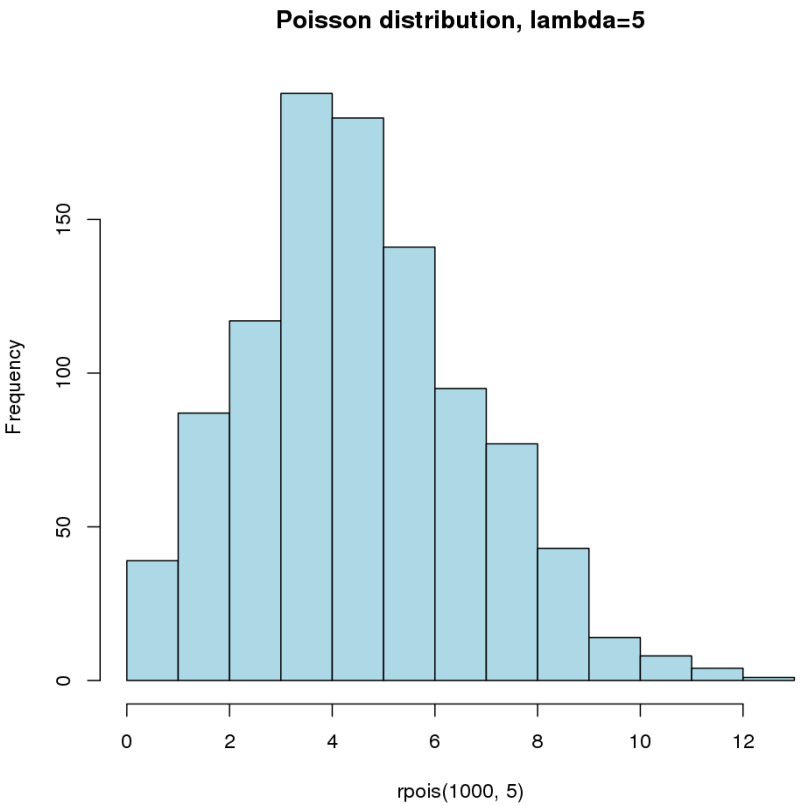
Longer reads → more uniquely mapped reads

- a=1 → read from this position ONLY aligns to this position
- a=1/n → read from this position could align to n locations

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

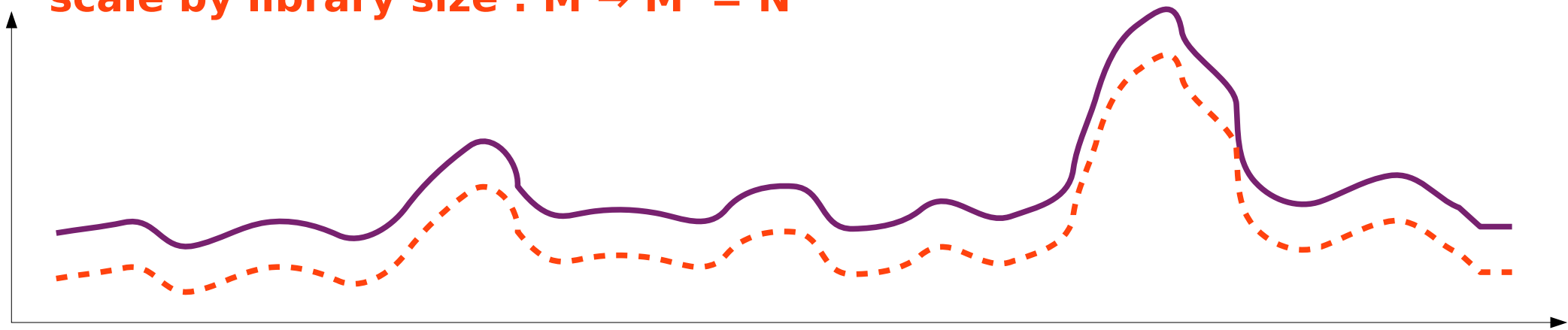
$$P(X=k)=e^{-k}\frac{\lambda^{k}}{k!}$$



Scaling unequal datasets

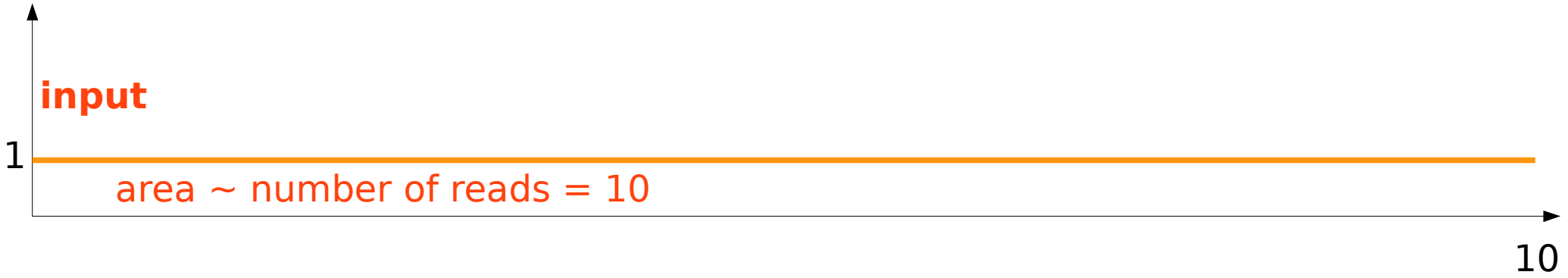
- treatment (=signal + noise) and input (=noise) datasets generally do not have the same sequencing depth → need for normalization
- input dataset should model the noise level in the treatment dataset
- **naïve approach** : upscale/downscale the smaller/larger dataset

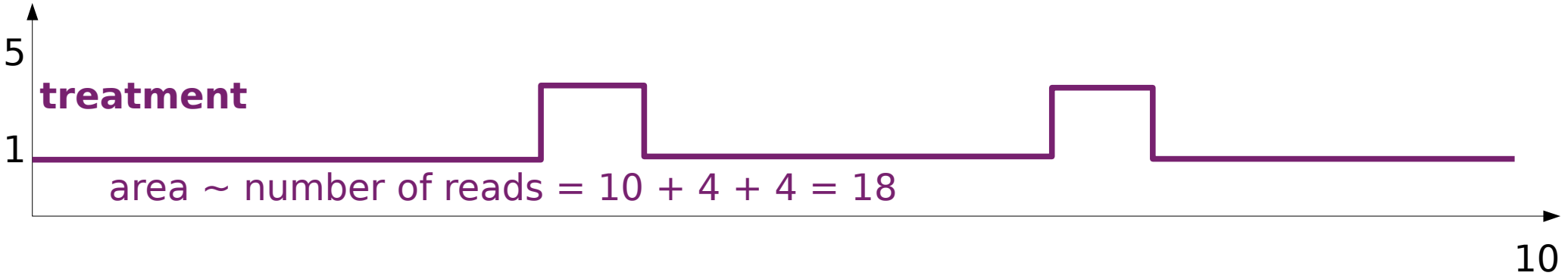
Input : N reads
ChIP-seq dataset → M > N reads
scale by library size : M → M' = N



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

Scaling unequal datasets by library size

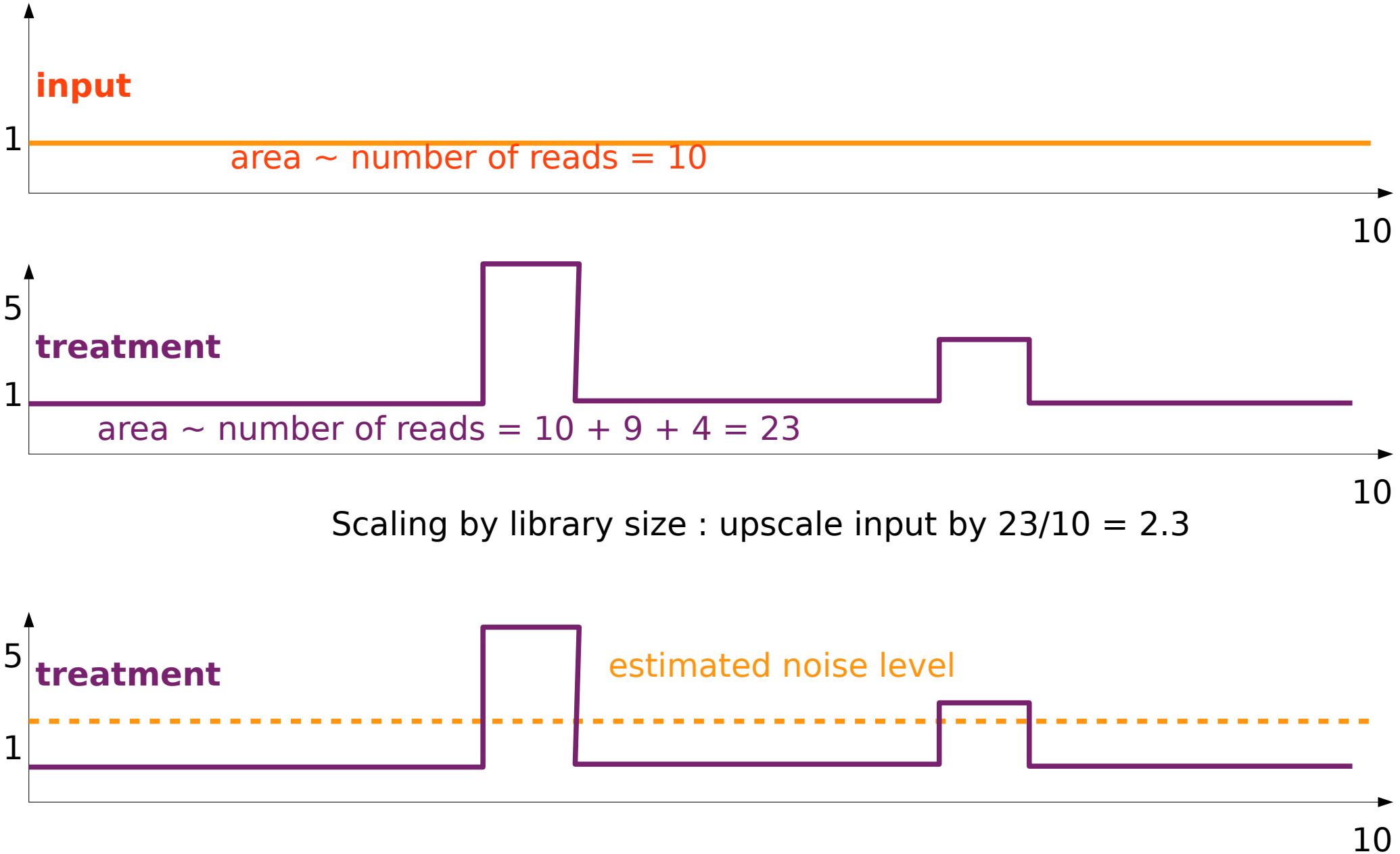




Scaling by library size : upscale input by 18/10 = 1.8

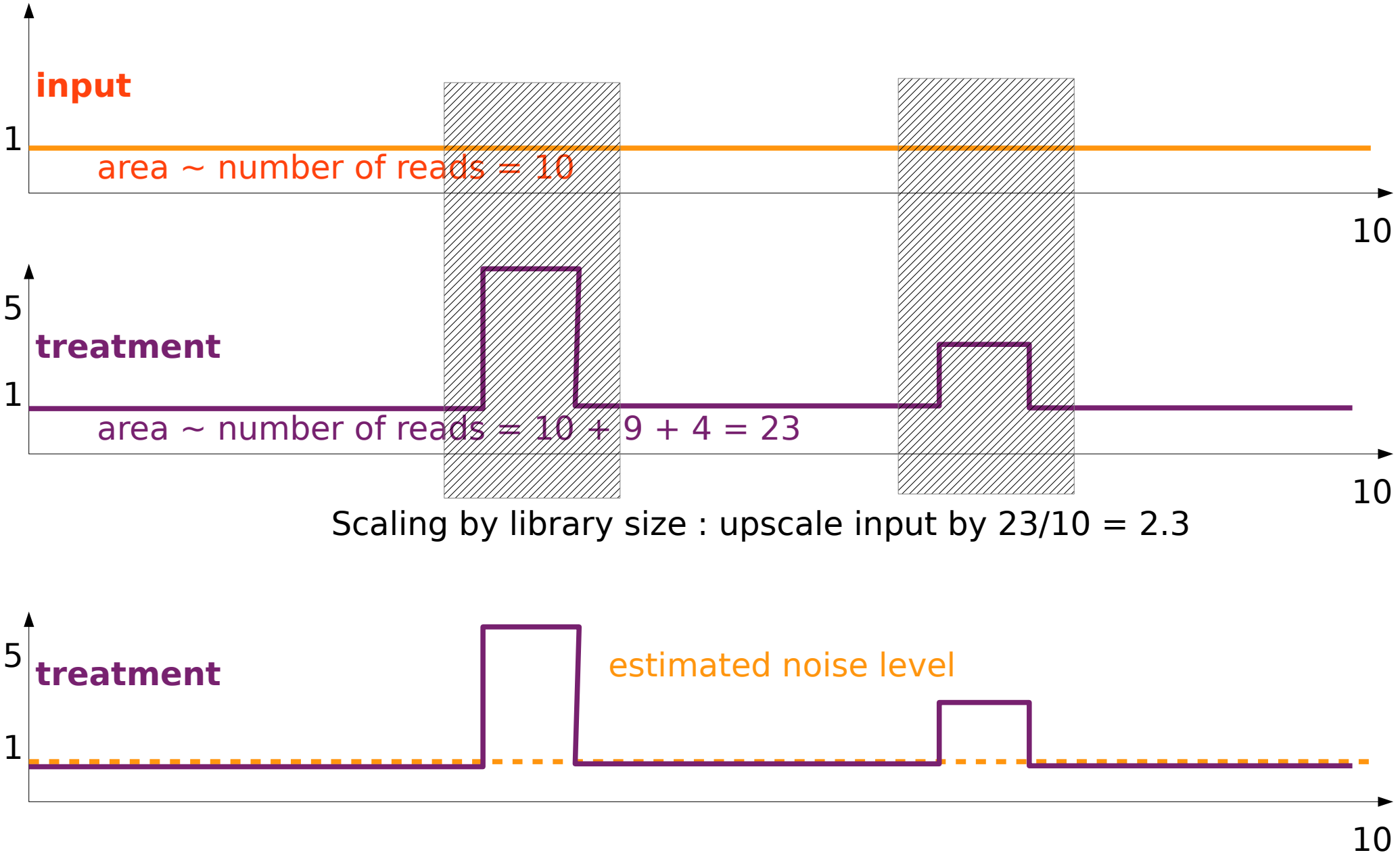
Noise level is over-estimated !

Scaling unequal datasets by library size



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

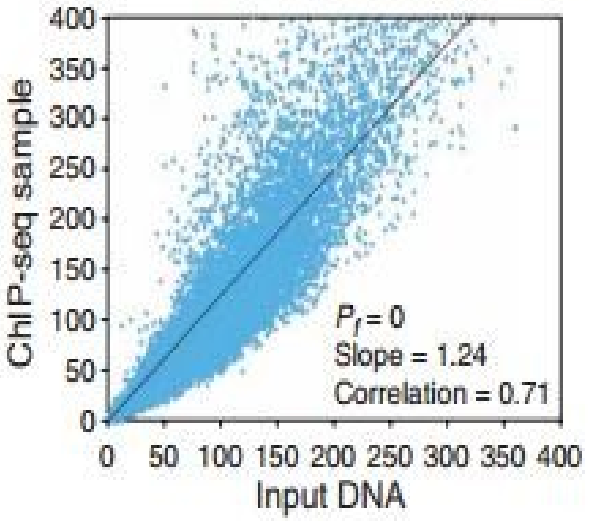
Scaling unequal datasets by library size



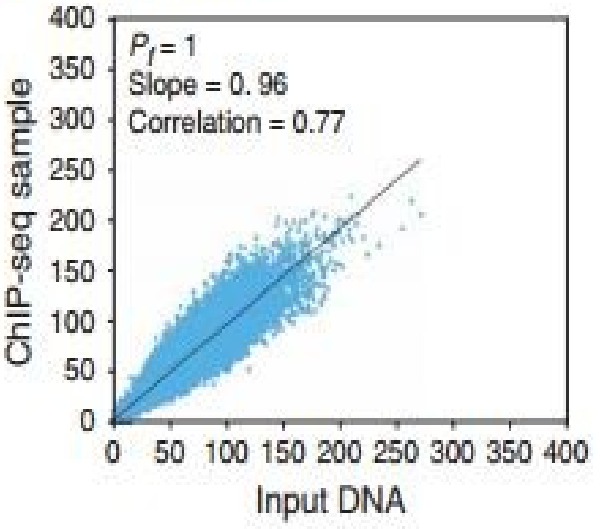
Scaling by library size : upscale input by 23/10 = 2.3

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]

Scaling unequal datasets

- Signal Extraction Scaling (SES)

Input(X)		q		ChiP (Y)		p	abs(p-q)
10	0.05	0.05		4	0.019704433	0.02	0.03
10	0.05	0.10		4	0.019704433	0.04	0.06
10	0.05	0.15		4	0.019704433	0.06	0.09
10	0.05	0.20		4	0.019704433	0.08	0.12
10	0.05	0.25		4	0.019704433	0.10	0.15
10	0.05	0.30		4	0.019704433	0.12	0.18
10	0.05	0.35		4	0.019704433	0.14	0.21
10	0.05	0.40		4	0.019704433	0.16	0.24
10	0.05	0.45		4	0.019704433	0.18	0.27
10	0.05	0.50		4	0.019704433	0.20	0.30
10	0.05	0.55		4	0.019704433	0.22	0.33
10	0.05	0.60		4	0.019704433	0.24	0.36
10	0.05	0.65		4	0.019704433	0.26	0.39
10	0.05	0.70		4	0.019704433	0.28	0.42
10	0.05	0.75		4	0.019704433	0.30	0.45
10	0.05	0.80		4	0.019704433	0.32	0.48
10	0.05	0.85		20	0.098522167	0.41	0.44
10	0.05	0.90		24	0.118226601	0.53	0.37
10	0.05	0.95		30	0.147783251	0.68	0.27
10	0.05	1.00		65	0.320197044	1.00	0.00
200				203			

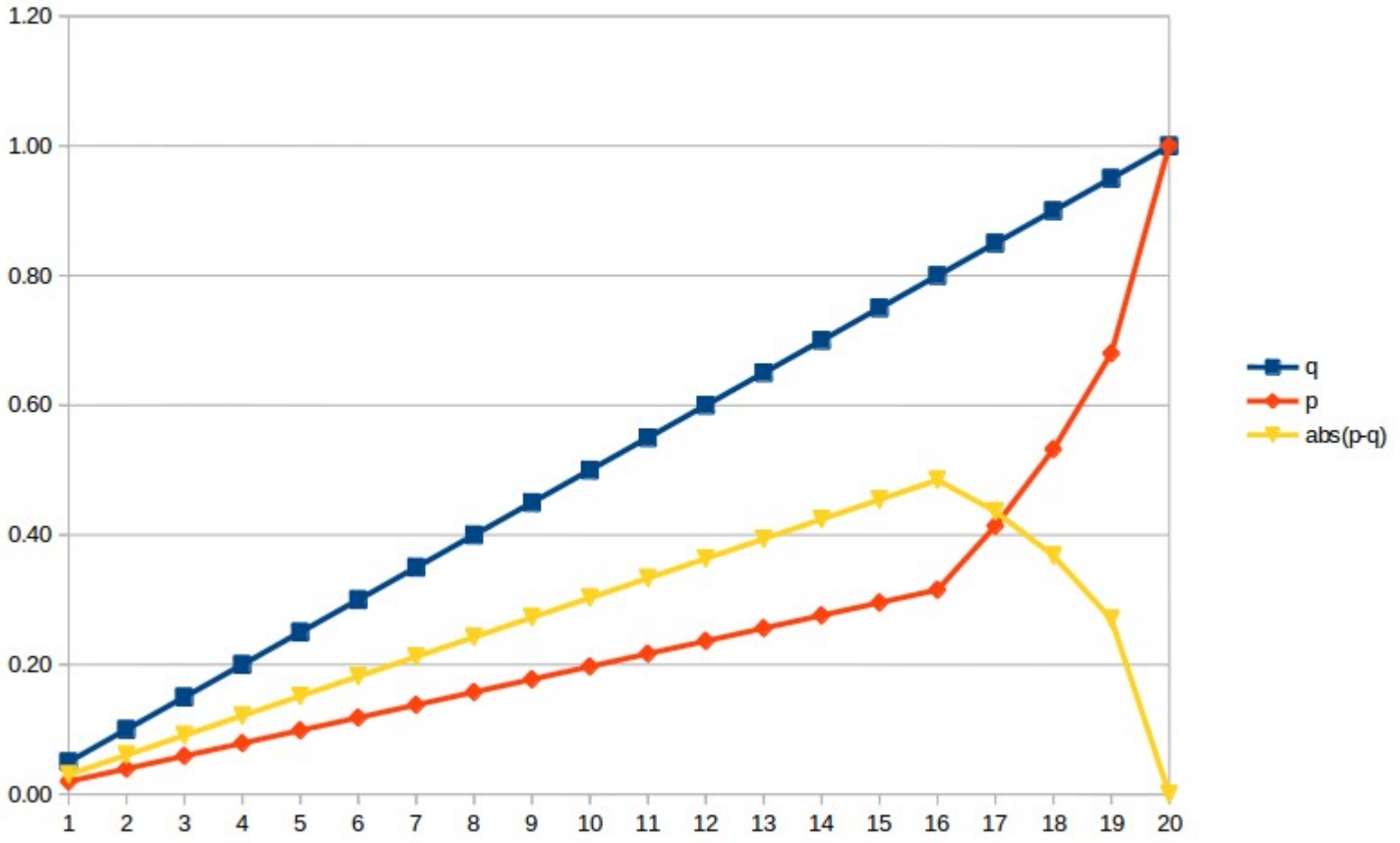
Stat Appl Genet Mol Biol. 2012 Mar 31;11(3):Article 9. doi: 10.1515/1544-6115.1750.

Normalization, bias correction, and peak calling for ChIP-seq.

Diaz A[1], Park K, Lim DA, Song JS.

Scaling unequal datasets

- Signal Extraction Scaling (SES)



Stat Appl Genet Mol Biol. 2012 Mar 31;11(3):Article 9. doi: 10.1515/1544-6115.1750.

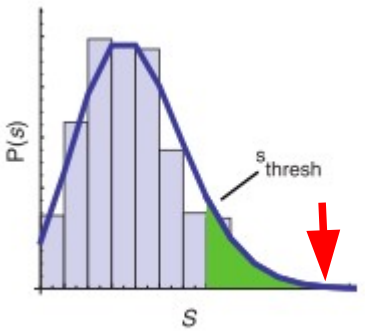
Normalization, bias correction, and peak calling for ChIP-seq.

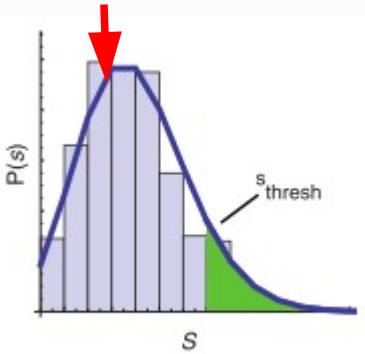
Diaz A[1], Park K, Lim DA, Song JS.

Defining “peaks”

- **Determining “enriched” regions**
  - sliding window across the genome
  - At each location, evaluate the enrichment of the Signal vs background based on Poisson distribution
  - retain regions with P-values below threshold
  - evaluate FDR

Pval < 1e-20



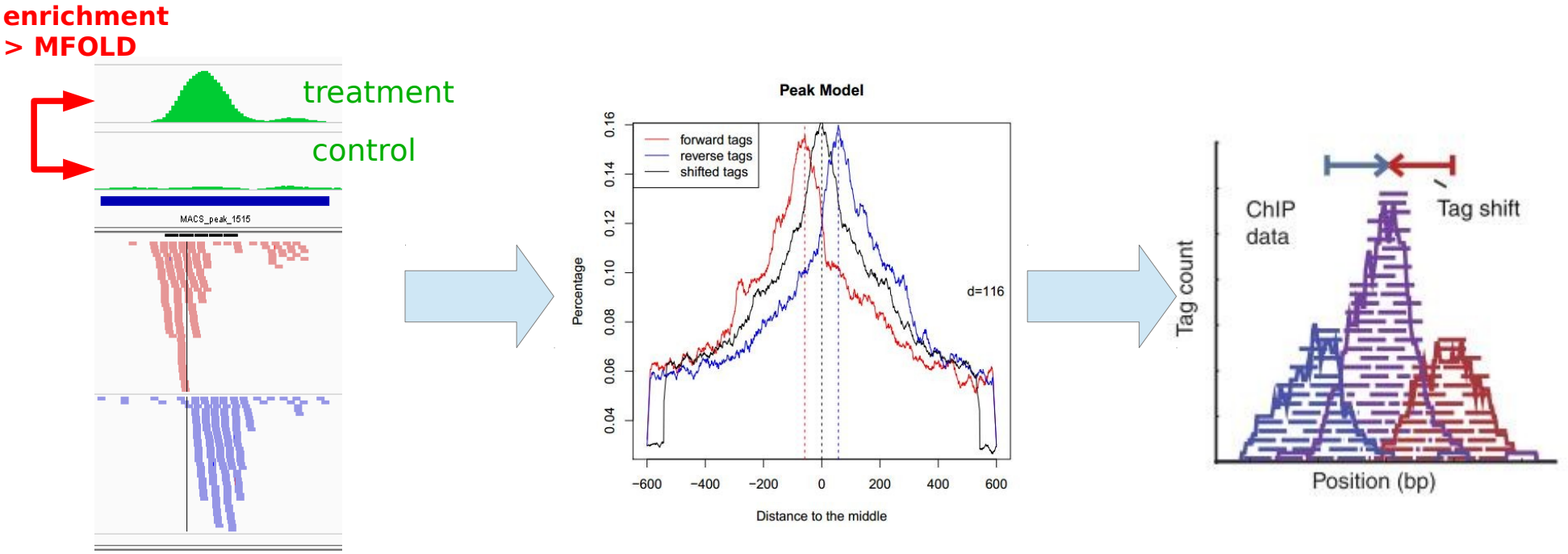


Pval ~ 0.6

MACS

[Zhang et al. Genome Biol. 2008]

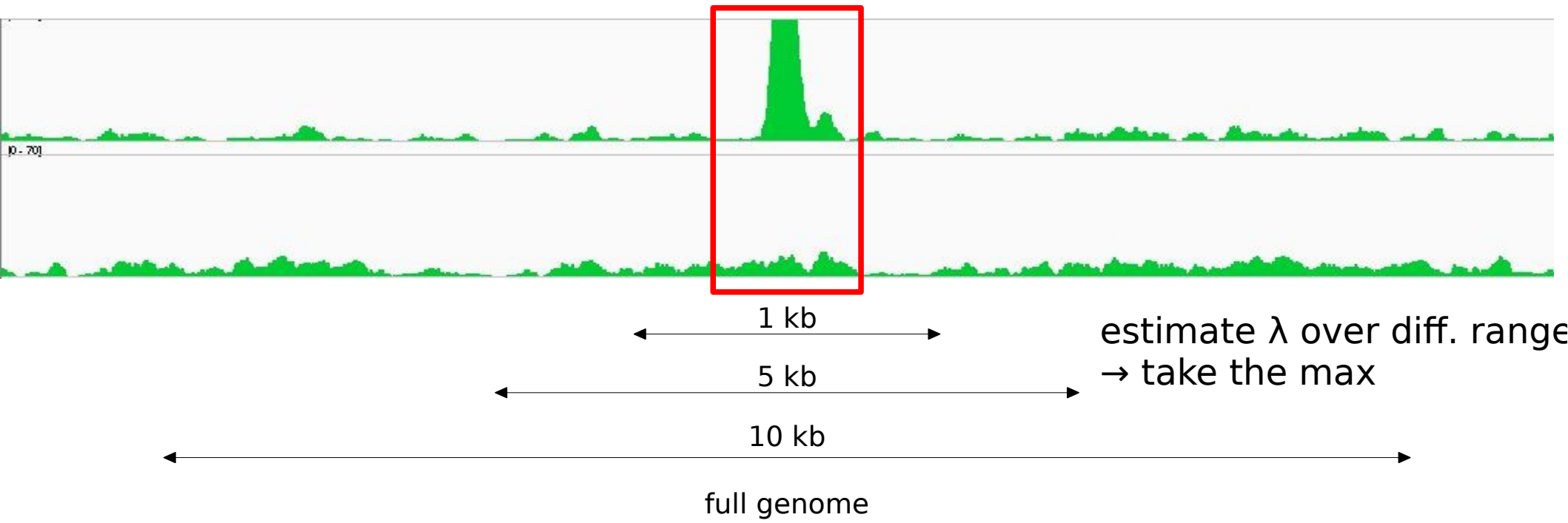
- **Step 1 : estimating fragment length *d***
  - slide a window of size BANDWIDTH
  - retain top regions with MFOLD enrichment of treatment vs. input
  - plot average +/- strand read densities → estimate d



MACS

[Zhang et al. Genome Biol. 2008]

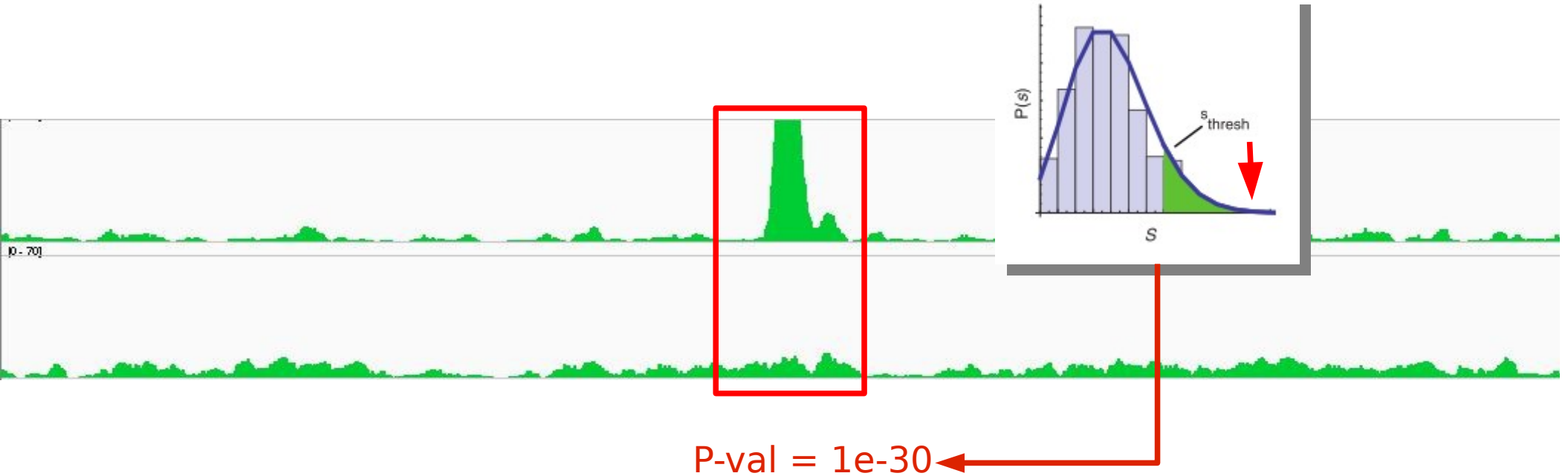
- **Step 2 : identification of local noise parameter**
  - slide a window of size *2*d* across treatment and input
  - estimate parameter λ_{local} of Poisson distribution



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



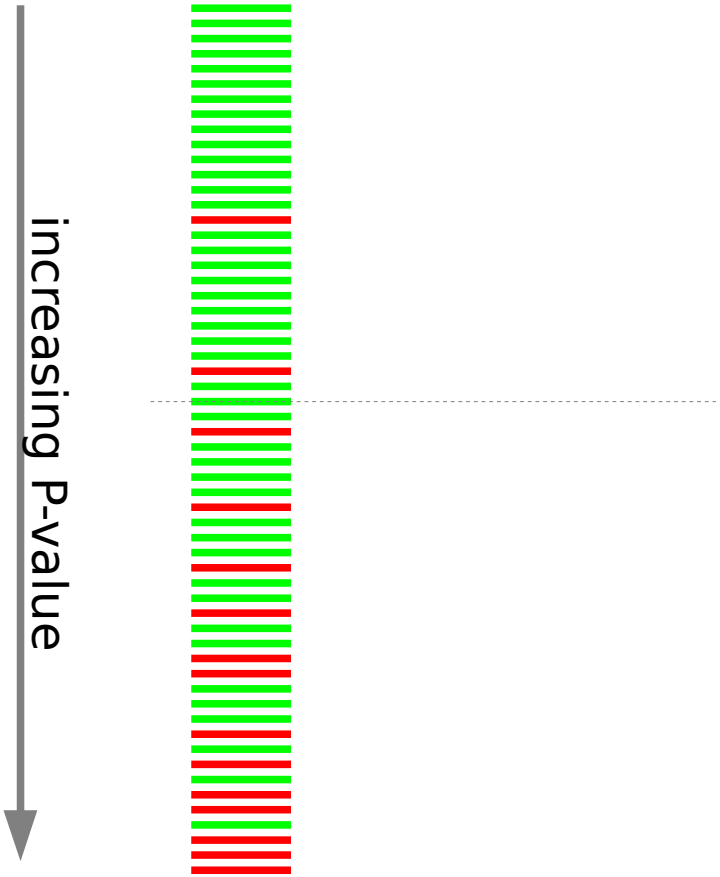
MACS

[Zhang et al. Genome Biol. 2008]

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

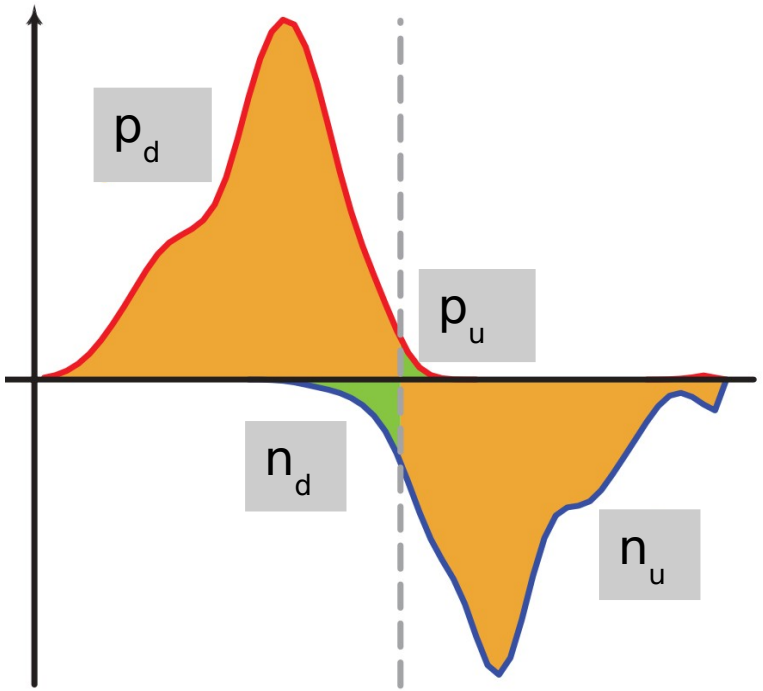
$$\text{FDR}(p) = \frac{\#\ \textbf{observed}\ \text{peaks with Pval} < p}{\#\ \textbf{simulated}\ \text{peaks with Pval} < p \ + \ \#\ \textbf{observed}\ \text{peaks with Pval} < p}$$

FDR = 2/(2+25)=0.074



Peak-Calling: WTD

- Window Tag Density (SPP package)



p_d = positive downstream

p_u = positive upstream

n_d = negative downstream

n_u = negative upstream

$$S_{wtd}(i) = \sqrt{(p_d * n_u)} - (p_u + n_d)$$

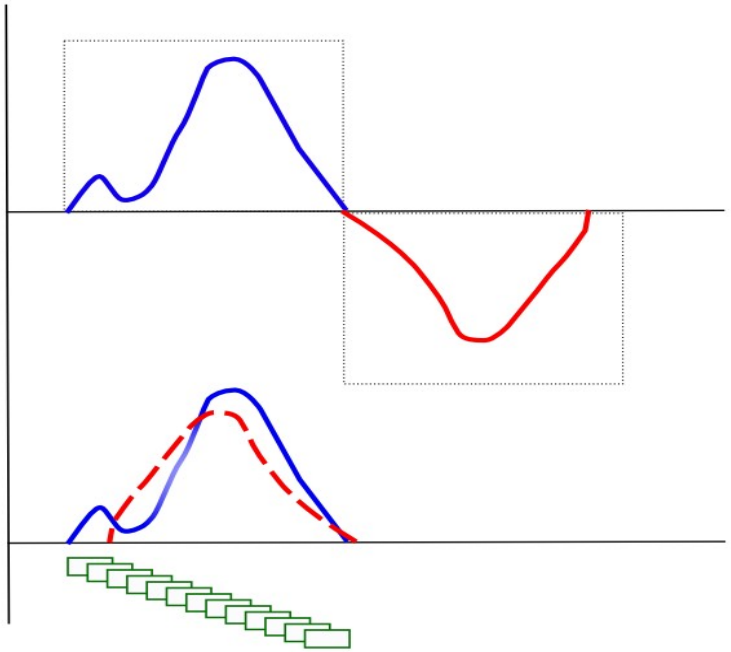
Nat Biotechnol. 2008 Dec;26(12):1351-9. Epub 2008 Nov 16.

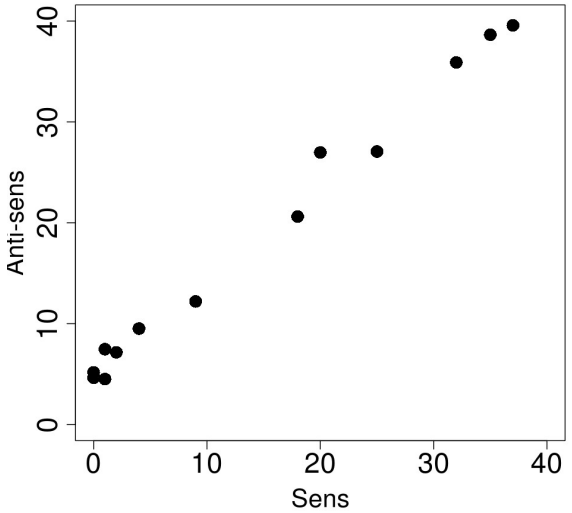
Design and analysis of ChIP-seq experiments for DNA-binding proteins.

Kharchenko PV, Tolstorukov MY, Park PJ.

Peak-Calling: MTC

- Mirror Tag Correlation (SPP package)
  - Strand cross-correlation profile





$$cov(x,y)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$r=\frac{cov(x,y)}{\sqrt{var(x)\,var(y)}}$$

Nat Biotechnol. 2008 Dec;26(12):1351-9. Epub 2008 Nov 16.

Design and analysis of ChIP-seq experiments for DNA-binding proteins.

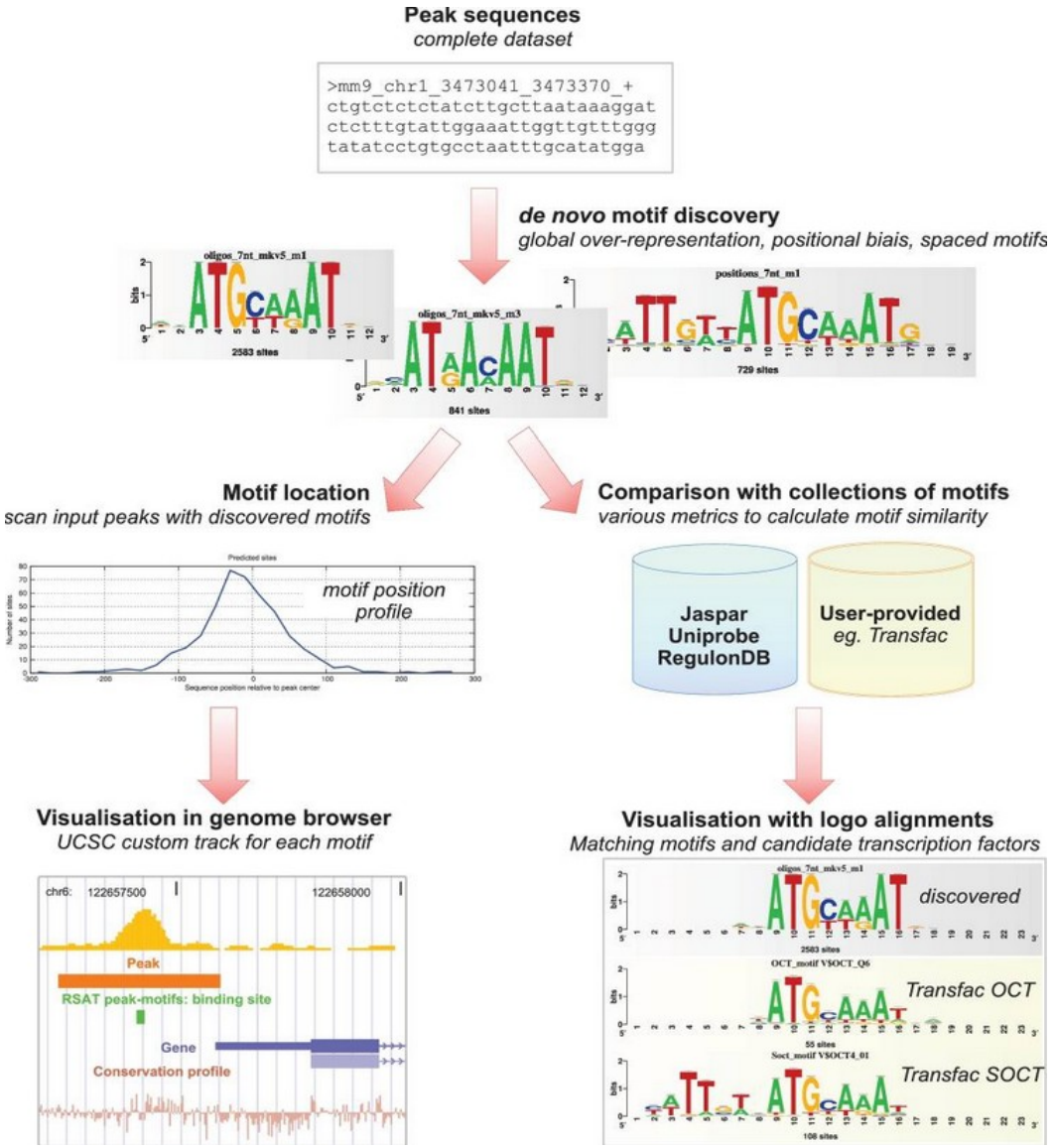
Kharchenko PV, Tolstorukov MY, Park PJ.

	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]

De novo motif discovery (Peak-motifs)



Nucleic Acids Res. 2012 Feb;40(4):e31. doi: 10.1093/nar/gkr1104. Epub 2011 Dec 8.

RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets.

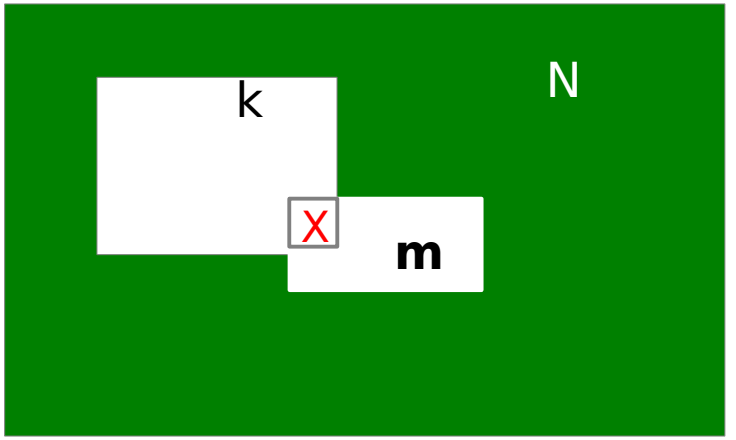
Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J.

Annotating Peaks ?

- Classical approach
  - Associate Peaks to the nearest genes
  - Check if the list of genes is enriched in gene related to :
    - Pathways, GO terms, ...

	Terme	!Terme	
Liste	x	k-x	k
!Liste	m-x	n-(k-x)	N-k
	m (white)	n (black)	N

- N genes in the genome
- m genes associated to a term (e.g. Cell cycle)
  - marked genes
- k genes (associated with peaks)
- If no bias, we expect the same proportion or marked genes in k and in N.
- Hypergeometric test: what is the probability to obtain by chance an intersection containing x or more genes ?



Nearest gene : problem

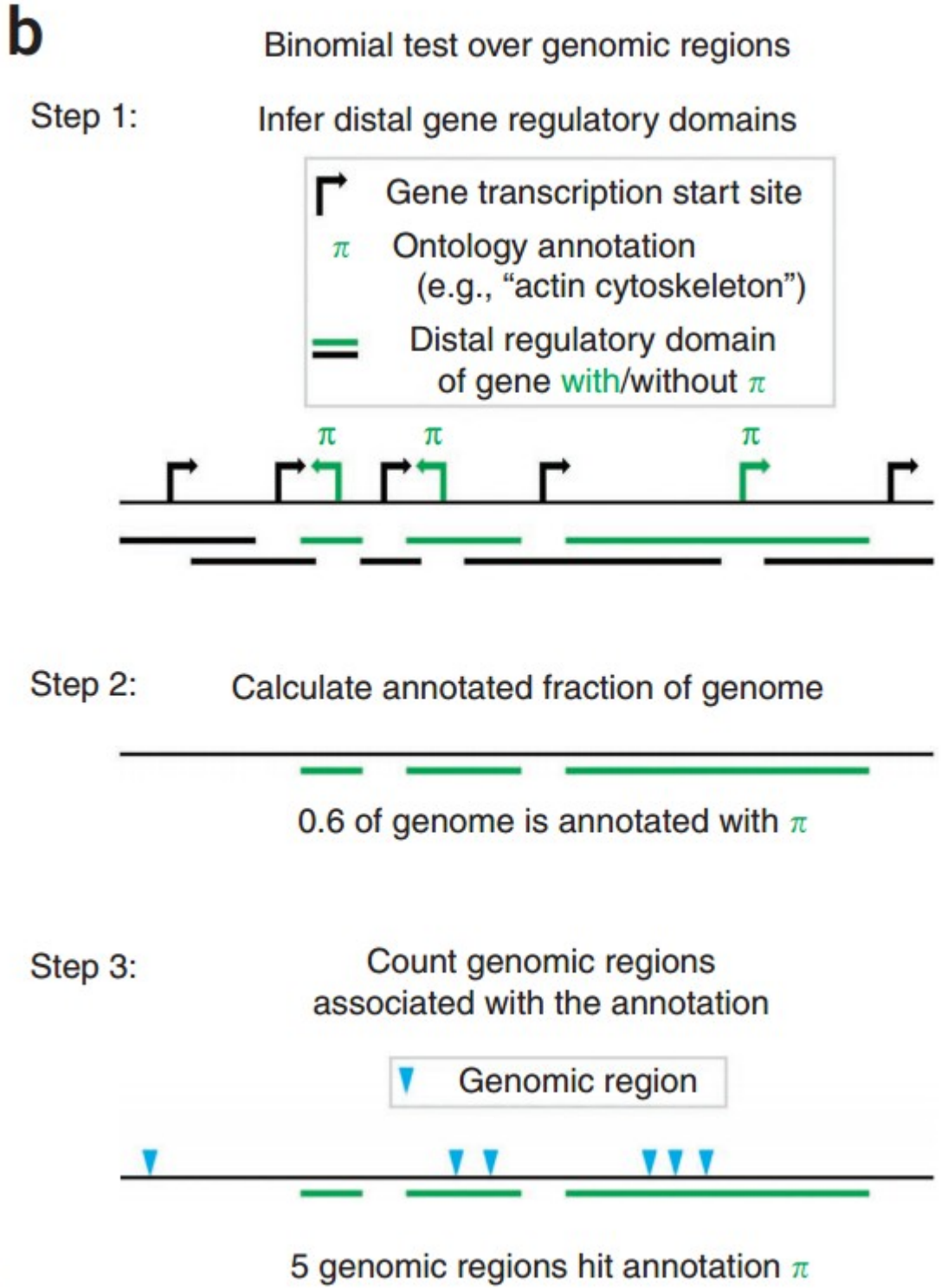
- Problem
  - Associating peaks with gene located at n kb
    - Discards lots of binding events (~ 50%)
  - Associating peaks to the nearest gene
    - Bias for genes within large intergenic regions
      - These genes will tend to be associated frequently with peaks
      - False positive enrichments ('multicellular organismal development')
- Solution
  - GREAT: Annotate genomic regions

GREAT

- GREAT (Genomic Regions Enrichment of Annotations Tool)
  - Define gene regulatory domain around genes
    - User may choose between several solutions
    - E.g single nearest gene

GREAT

- Use a binomial test to check for enrichment



Step 4: Perform binomial test over genomic regions

$n = 6$ total genomic regions

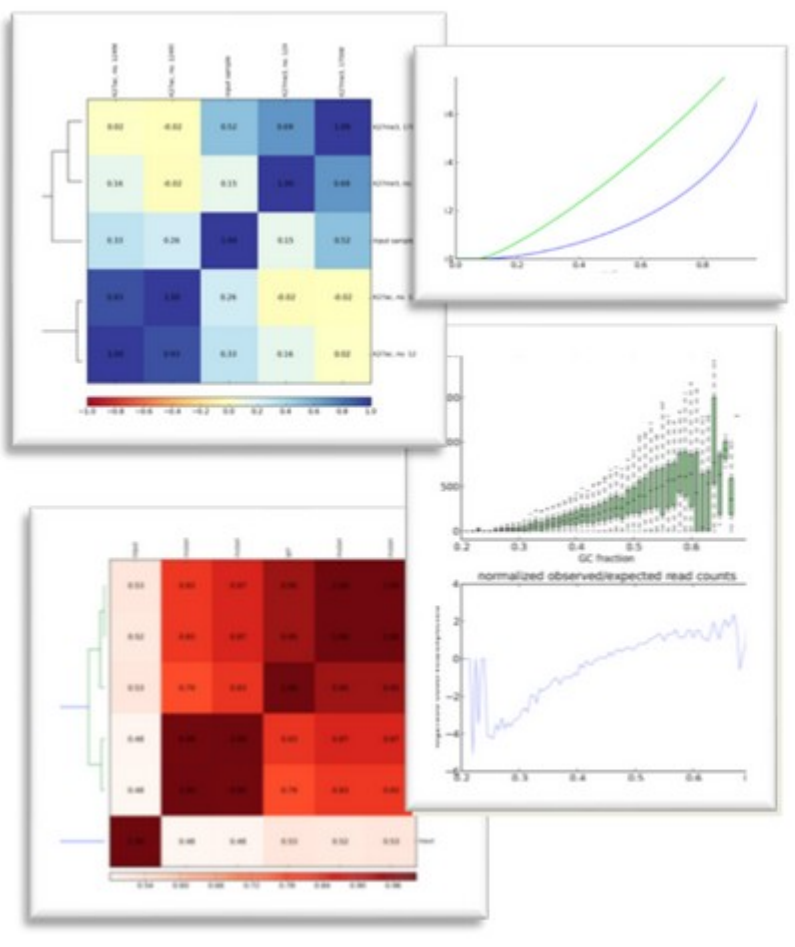
$p_\pi = 0.6$ fraction of genome annotated with π

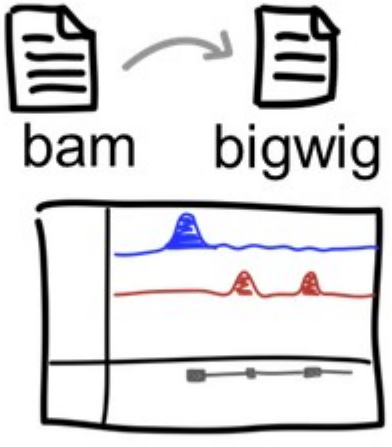
$k_\pi = 5$ genomic regions hit annotation π

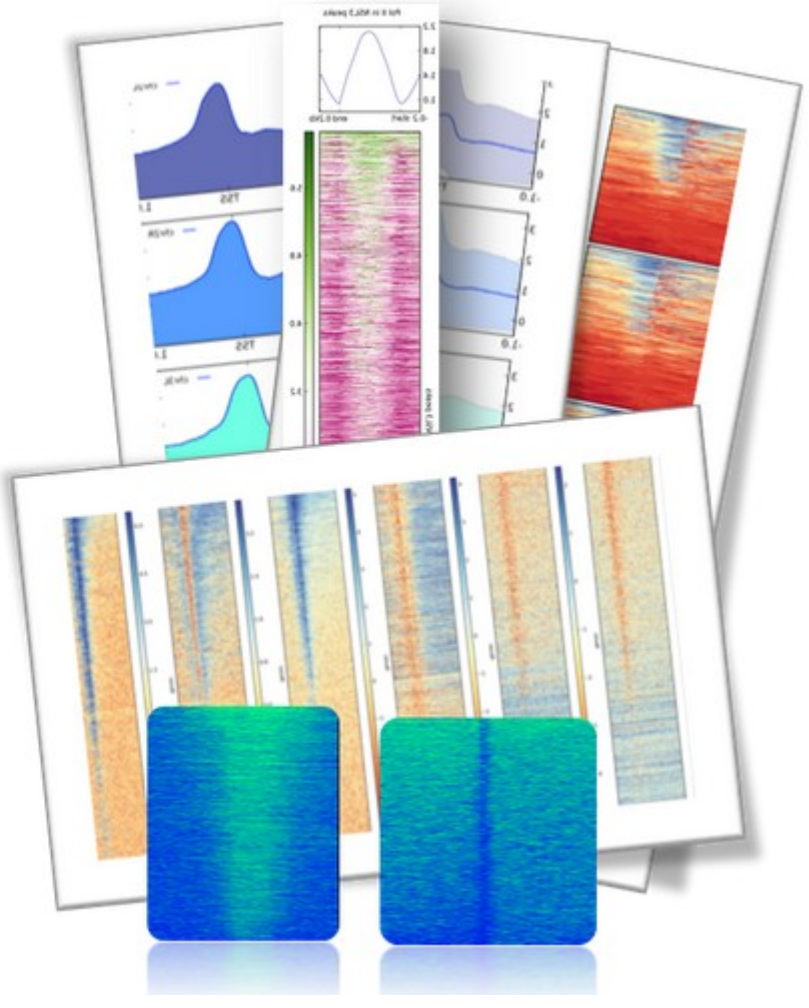
$$P = \Pr_{\text{binom}}(k \geq 5 \mid n = 6, p = 0.6)$$

DeepTools

- DeepTools: user-friendly tools for the normalization and visualization of deep-sequencing data







Data processing
&
file formats

Fastq file format

- Header
- Sequence
- + (optional header)
- Quality (default Sanger-style)

```
@QSEQ32.249996 HWUSI-EAS1691:3:1:17036:13000#0/1 PF=0 length=36
GGGGGTCATCATCATTTGATCTGGGAAAGGCTACTG
+
=.+5:<<<<>AA?0A>;A*A################
@QSEQ32.249997 HWUSI-EAS1691:3:1:17257:12994#0/1 PF=1 length=36
TGTACAACAACAACCTGAATGGCATACTGGTTGCTG
+
DDDD<BDBDB??BB*DD:D################
```

Sanger quality score

- Sanger quality score (Phred quality score): Measure the quality of each base call
  - Based on p, the probality of error (the probability that the corresponding base call is incorrect)
  - Qsanger= -10*log10(p)
  - p = 0.01 <=> Qsanger 20
- Quality score are in ASCII 33
- Note that SRA has adopted Sanger quality score although original fastq files may use different quality score (see: http://en.wikipedia.org/wiki/FASTQ_format)

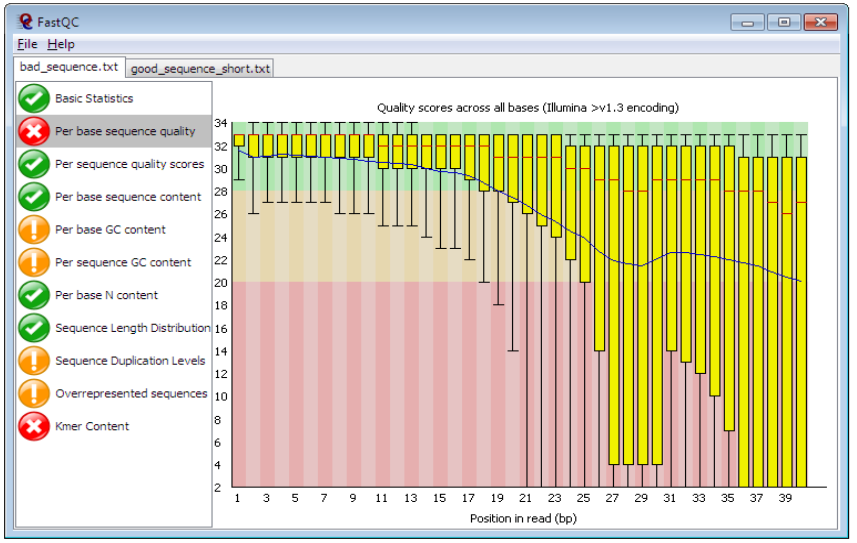
ASCII 33

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

- Storing PHRED scores as single characters gave a simple and space efficient encoding:
- Character ”!” means a quality of 0
- Range 0-40

Quality control for high throughput sequence data

- FastQC
  - GUI / command line
  - http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

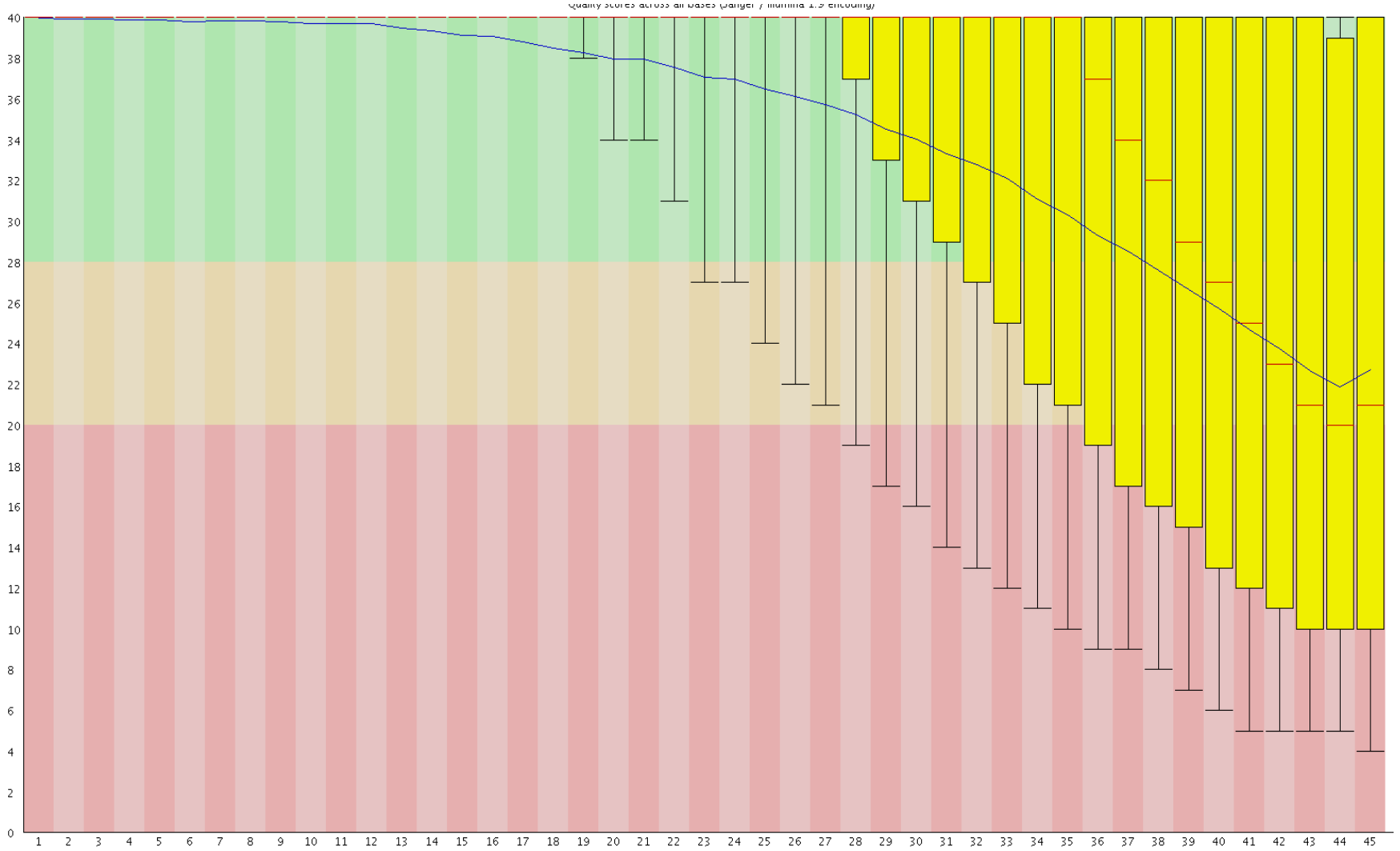


  - ShortRead
    - Bioconductor package

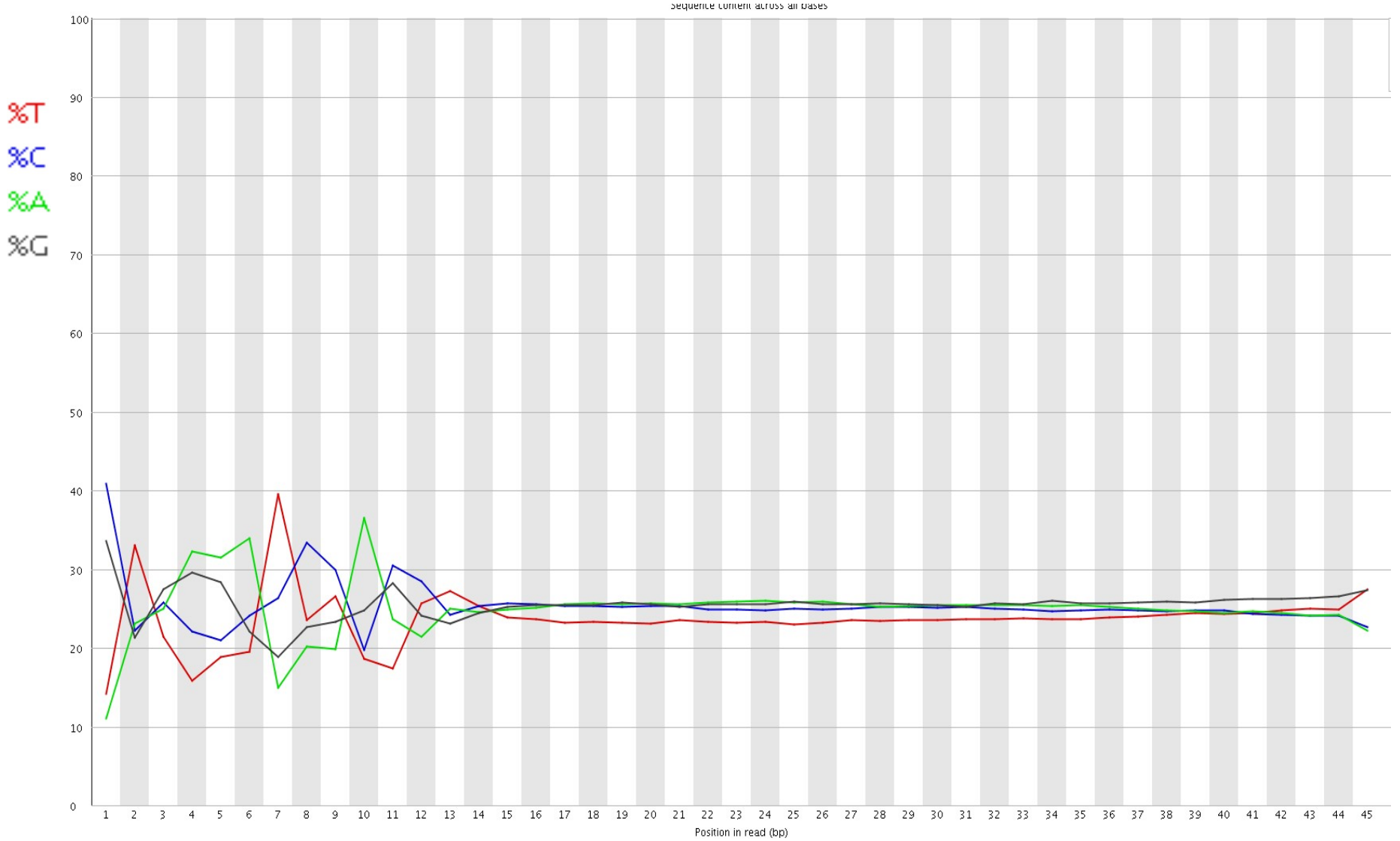
Trimming

- Essential step (at least when using bowtie)
  - Almost mandatory when using tophat
- FASTX-Toolkit
- Sickle
  - Window-based trimming (unpublished)
- ShortRead
  - Bioconductor package
- csfasta_quality_filter.pl
  - SOLiD
    - Mean quality
    - Continuous run of bad colors at the end of the read

Quality control with FastQC

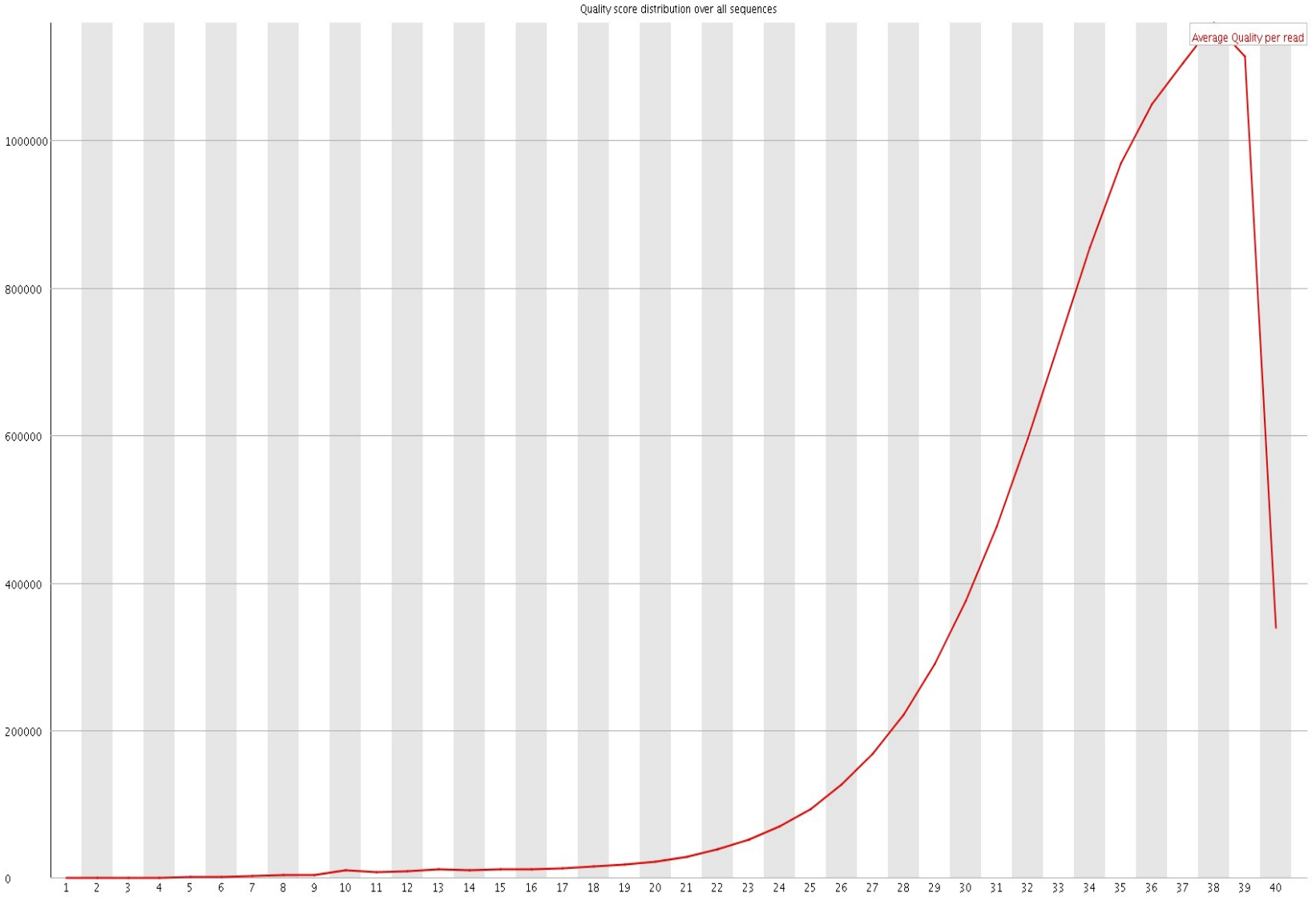


Quality control with FastQC



Position in read

Quality control with FastQC



Nb Reads

Mean Phred Score

Mapping reads to genome: general softwares

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a]Work well for Sanger and 454 reads, allowing gaps and clipping.

[b]Paired end mapping.

[c]Make use of base quality in alignment.dBWA trims the primer base and the first color for a color read.

[e]Long-read alignment implemented in the BWA-SW module. fMAQ only does gapped alignment for Illumina paired-end reads.

[g]Free executable for non-profit projects only

Brief Bioinform. 2010 Sep;11(5):473-83. Epub 2010 May 11.

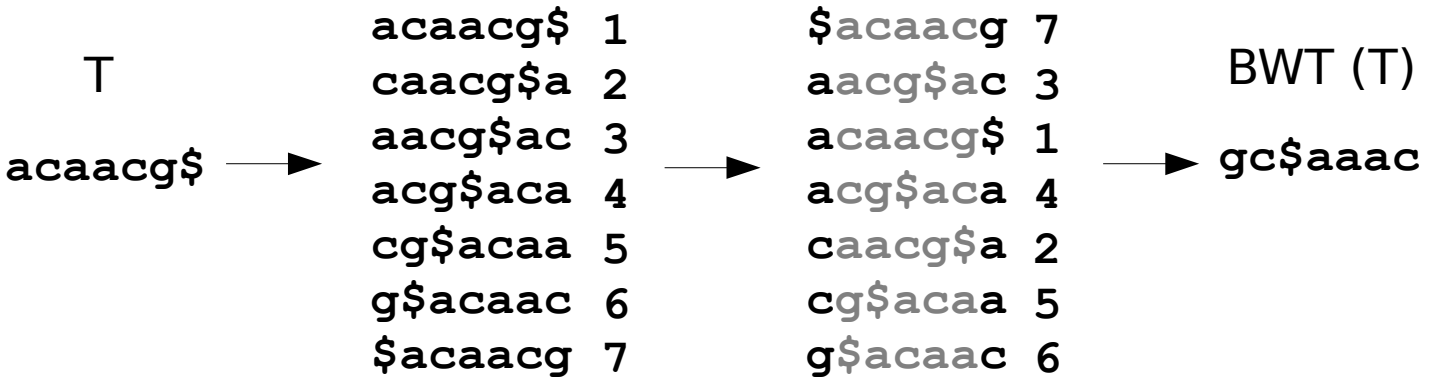
A survey of sequence alignment algorithms for next-generation sequencing.

Li H, Homer N.

Broad Institute, Cambridge, MA 02142, USA. hengli@broadinstitute.org

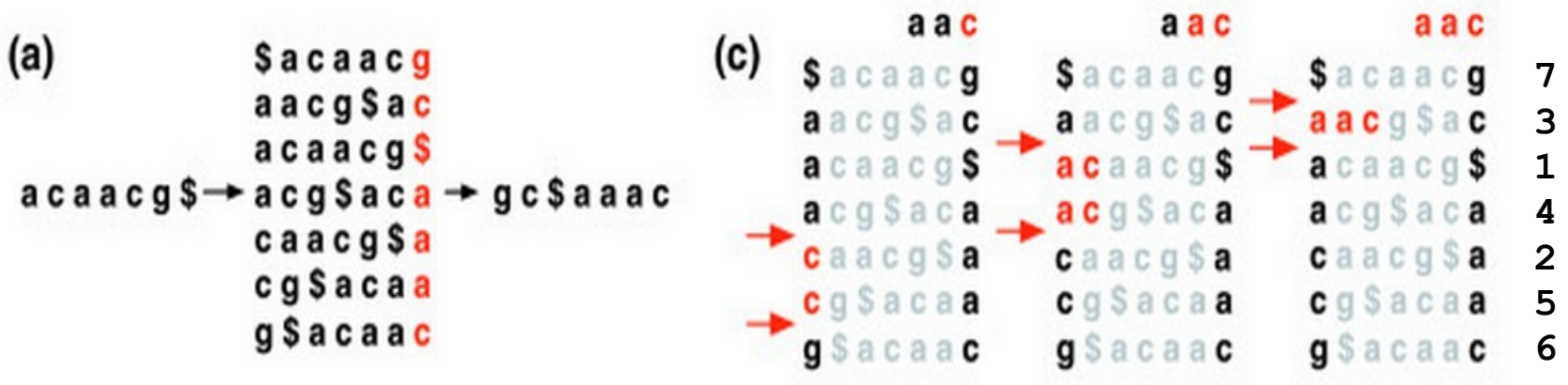
Bowtie principle

- Use highly efficient compressing and mapping algorithms based on Burrows Wheeler Transform (BWT)
- The Burrows-Wheeler Transform of a text T, BWT(T), can be constructed as follows.
  - The character $ is appended to T, where $ is a character not in T that is lexicographically less than all characters in T.
  - The Burrows-Wheeler Matrix of T, BWM(T), is obtained by computing the matrix whose rows comprise all cyclic rotations of T sorted lexicographically.



Bowtie principle

- Burrows-Wheeler Matrices have a property called the Last First (LF) Mapping.
  - The ith occurrence of character c in the last column corresponds to the same text character as the ith occurrence of c in the first column.
  - Example: searching ”AAC” in ACAACG



Genome Biol. 2009;10(3):R25. Epub 2009 Mar 4.

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.

Langmead B, Trapnell C, Pop M, Salzberg SL.

Storing alignment: SAM Format

- Store information related to alignement
  - Read ID
  - CIGAR String
  - Bitwise FLAG
    - read paired
    - read mapped in proper pair
    - read unmapped, ...
  - Alignment position
  - Mapping quality
  - ...

Bitwise flag

- read paired
- read mapped in proper pair
- read unmapped
- mate unmapped
- read reverse strand
- mate reverse strand
- first in pair
- second in pair
- not primary alignment
- read fails platform/vendor quality checks
- read is PCR or optical duplicate

Bitwise flag

- 00000000001 → 2^0 = 1 (read paired)
- 00000000010 → 2^1 = 2 (read mapped in proper pair)
- 00000000100 → 2^2 = 4 (read unmapped)
- 00000001000 → 2^3 = 8 (mate unmapped) …
- 00000010000 → 2^4 = 16 (read reverse strand)

- 00000001001 → 2^0+ 2^3 = 9 → (read paired, mate unmapped)
- 00000001101 → 2^0+2^2+2^3 =13 …
- …

http://picard.sourceforge.net/explain-flags.html

The extended CIGAR string

- Exemple flags:
  - M alignment match (can be a sequence match or mismatch)
  - I insertion to the reference
  - D deletion from the reference
  - http://samtools.sourceforge.net/SAM1.pdf

```
ATTCAGATGCAGTA
ATTCA--TGCAGTA

 5M2D7M
```

Mapping reads

- Main Issues:
  - Number of multihits
    - Issue with short reads → mappability
  - PCR duplicates
    - Warning with ChIP-Seq (library complexity)
  - Number of allowed mismatches
    - Depend on sequence size (sometimes heterogeneous length)
    - Depend of the aligner

Merci