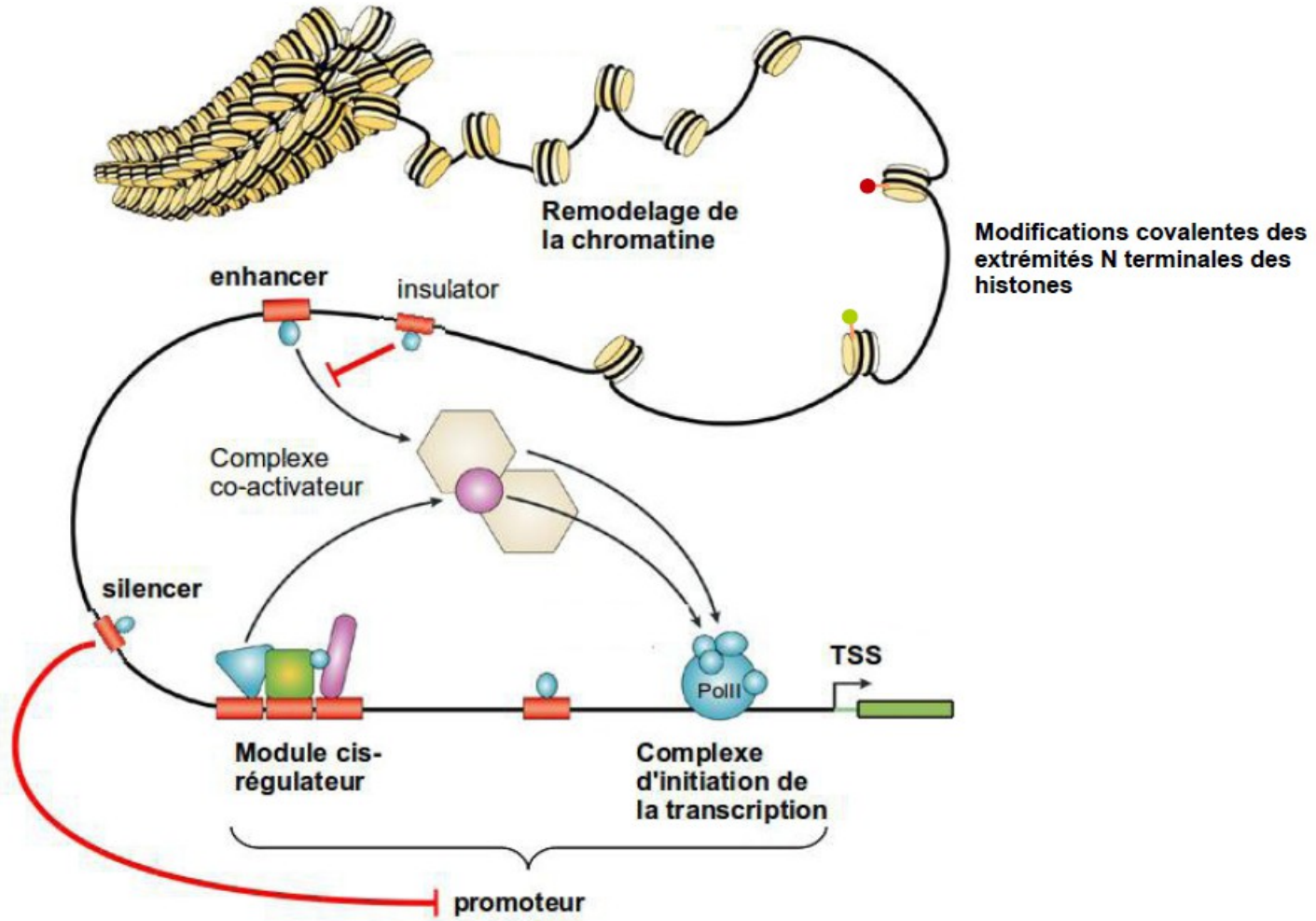
ChIP-seq analysis – D. Puthier

Adapted from "Aviesan Bioinformatic School" (M. Defrance, C. Herrmann, S. Le Gras, J. van Helden, D. Puthier, M. Thomas.Chollier)

About transcriptional regulation and epigenetics

A model of transcriptional regulation



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

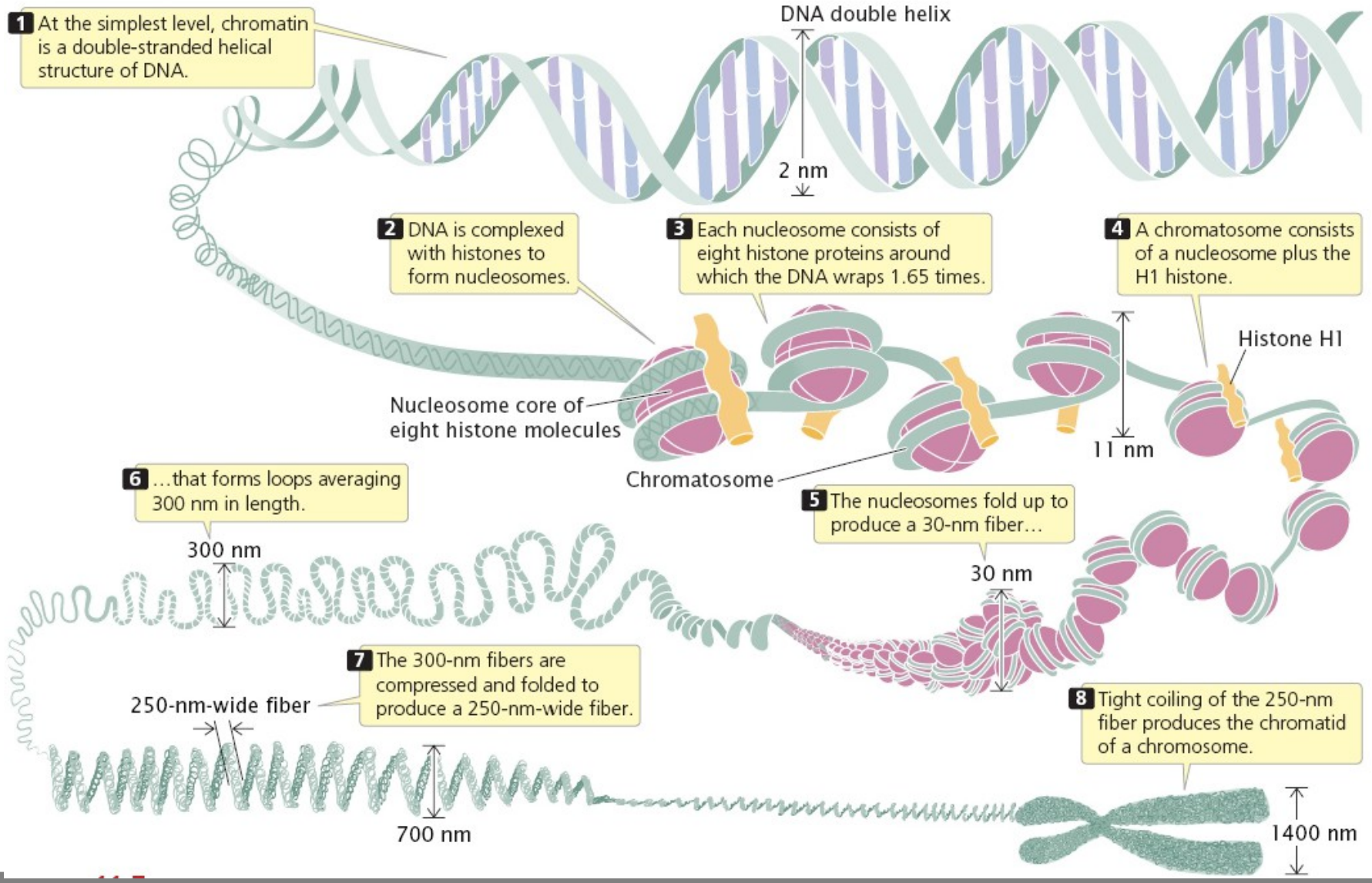
Applied bioinformatics for the identification of regulatory elements.

Wasserman WW, Sandelin A.

Chromatin constraints

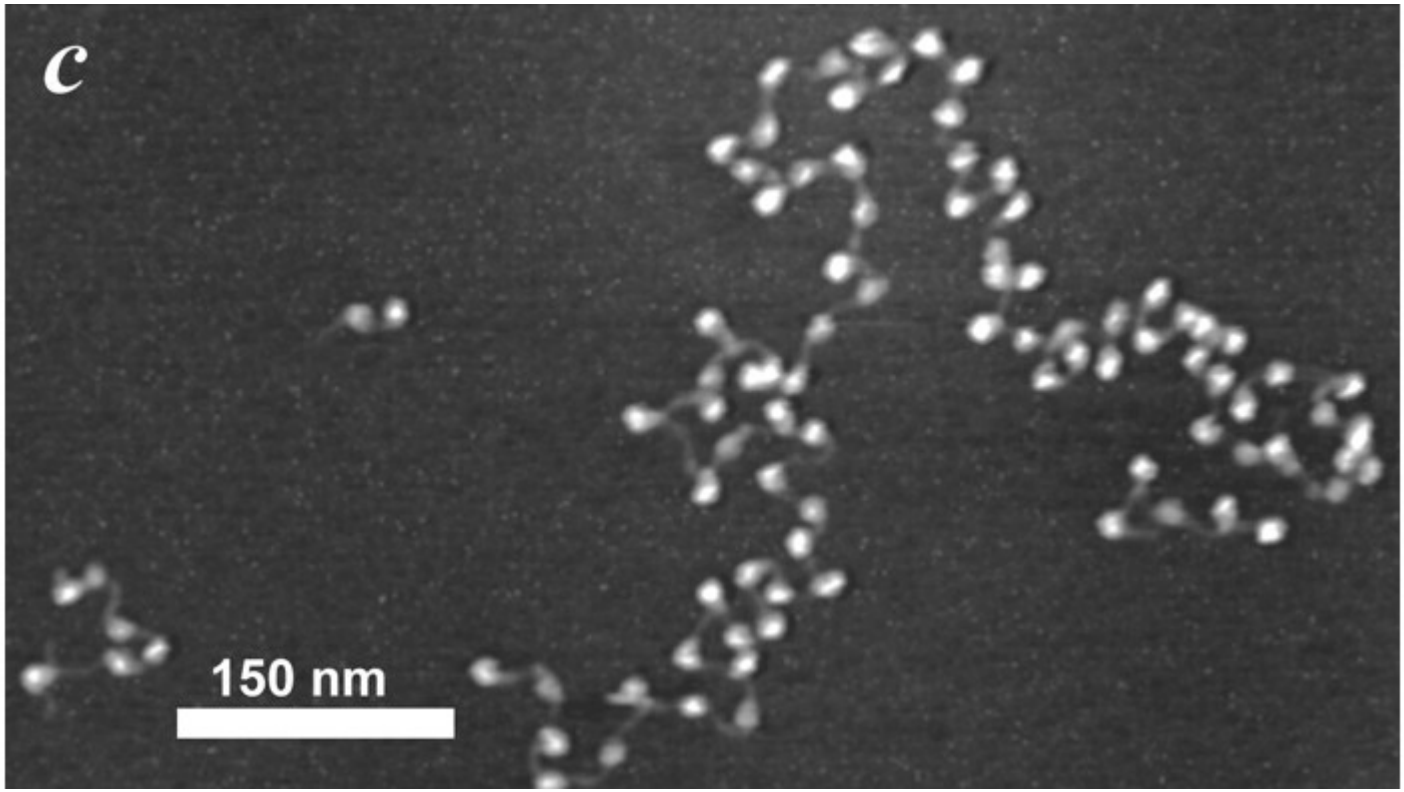
- Each diploid cell contains about 2 meters of DNA
  - High level of compaction required
  - Accessibility required
    - Replication
    - Transcription
    - DNA repair
- Specific machinery required

Chromatin has highly complex structure with several levels of organization



2005. Genetics: A Conceptual Approach, 2nd ed.

Beads on a string



- Figure 4: Chromatin fibers purified from chicken erythrocytes. Each nucleosome (~12-15 nm) is well resolved, along with the linker DNA between the nucleosomes. Given the resolution, other components, if present, such as a transcribing RNA polymerase or transcription factor complexes, should be resolvable

News Physiol Sci. 1999 Aug;14:142-149.

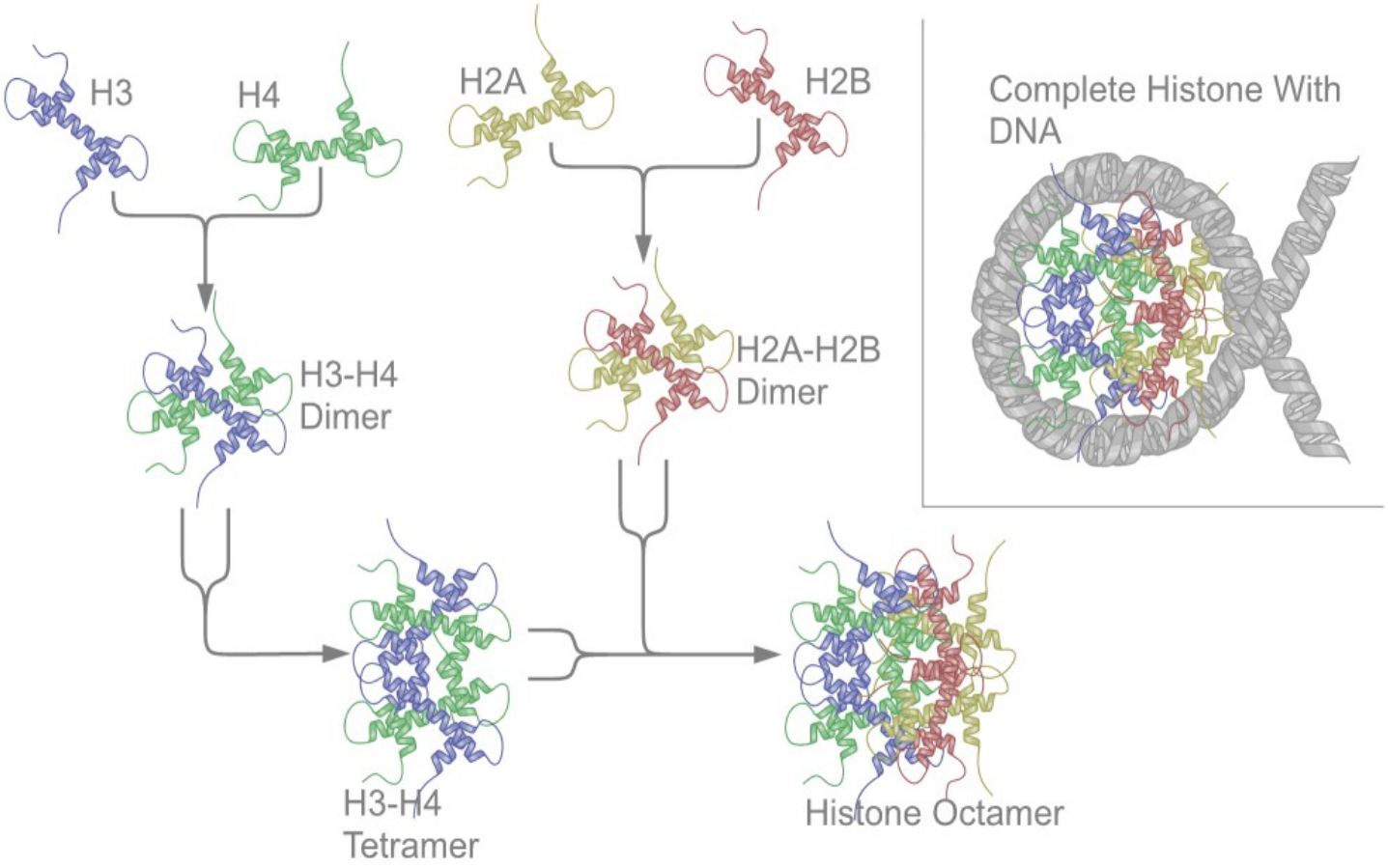
Probing Nanometer Structures with Atomic Force Microscopy.

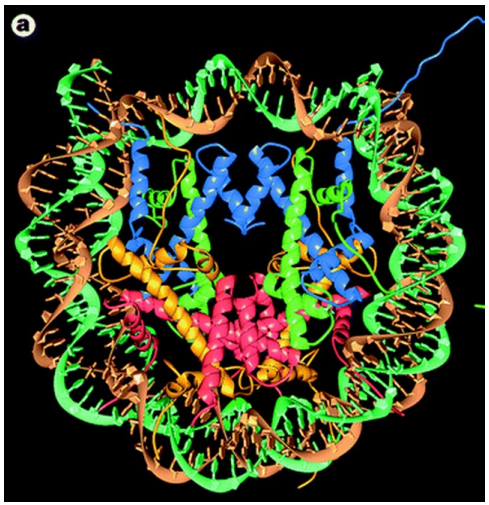
Shao Z.

Histones and nucleosomes

- Histones
  - Small proteins (11-22 kDa)
  - Highly conserved
  - Basic (Arginine et Lysine)
  - N-terminal tails subject to post translational modification
- Nucleosome
  - Octamers of histone
    - (H2A,H2B,H3,H4) x 2
  - 146bp DNA

Nucleosome structure





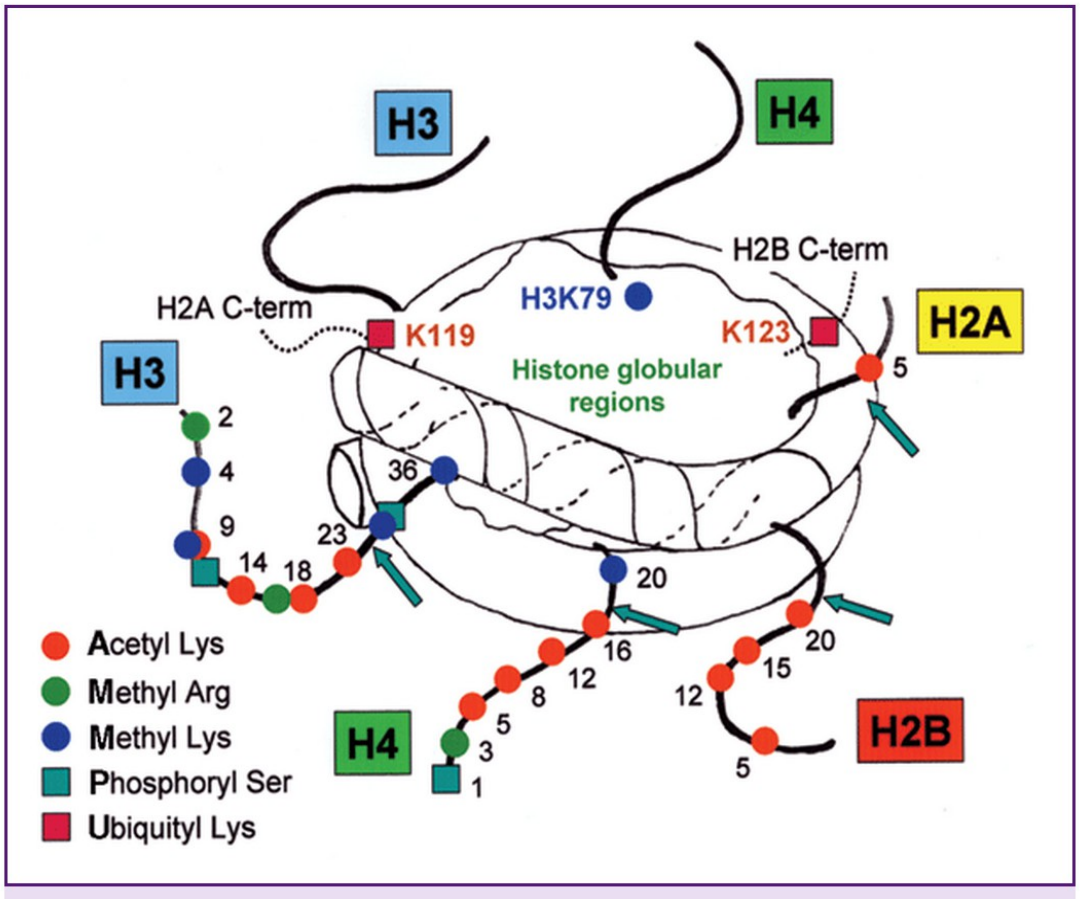
Nature. 1997 Sep 18;389(6648):251-60.

Crystal structure of the nucleosome core particle at 2.8 A resolution.

Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ.

Histone post translational modification

- Lysine acetylation
- Lysine methylation
- Arginine methylation
- Serine phosphorylation
- Threonine phosphorylation
- ADP-ribosylation
- Ubiquitylation
- Sumoylation
- …



Nat Struct Mol Biol. 2005 Feb;12(2):110-2.

Reading signals on the nucleosome with a new nomenclature for modified histones.

Turner BM.

Some alternative modifications

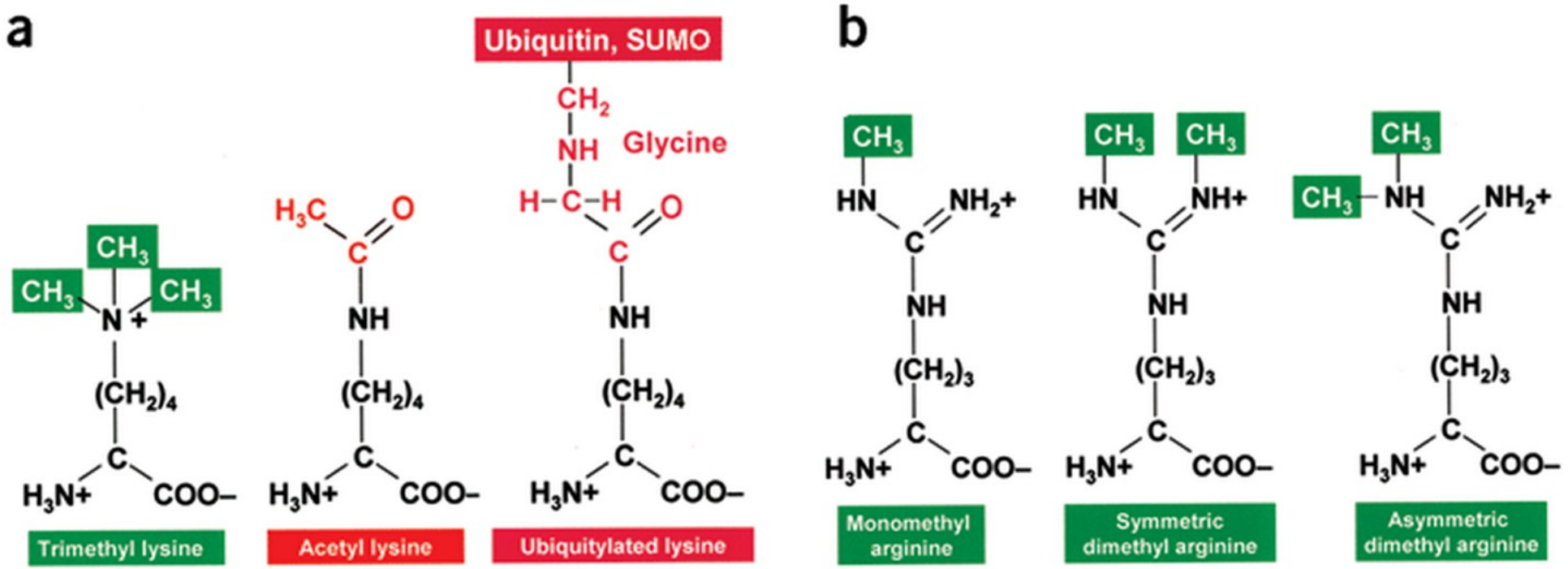


Figure 2. Some alternative modifications to lysine and arginine residues.
(**a**) Lysines can accommodate one, two or three methyl groups (left), always retaining a positive charge; a single acetate group (middle), in which case the positive charge is lost; or can form an isopeptide bond with the C-terminal glycine of ubiquitin or SUMO (right), which also neutralizes the positive charge. (**b**) Arginines can be modified with a single methyl group (left), or with two methyl groups that can be arranged symmetrically (middle) or asymmetrically (right). All forms retain a positive charge.

Nat Struct Mol Biol. 2005 Feb;12(2):110-2.

Reading signals on the nucleosome with a new nomenclature for modified histones.

Turner BM.

The Brno nomenclature

Table 1 The Brno nomenclature for histone modifications

Modifying group	Amino acid(s) modified	Level of modification	Abbreviation for modification[a]	Examples of modified residues[b]
Acetyl-	Lysine	mono-	ac	H3K9ac
Methyl-	Arginine	mono-	me1	H3R17me1
	Arginine	di-, symmetrical	me2s	H3R2me2s
	Arginine	di-, asymmetrical	me2a	H3R17me2a
	Lysine	mono-	me1	H3K4me1
	Lysine	di-	me2	H3K4me2
	Lysine	tri-	me3	H3K4me3
Phosphoryl-	Serine or threonine	mono-	ph	H3S10ph
Ubiquityl-	Lysine	mono-[c]	ub1	H2BK123ub1
SUMOyl-	Lysine	mono-	su	H4K5su[d]
ADP ribosyl-	Glutamate	mono-	ar1	H2BE2ar1
	Glutamate	poly-	arn	H2BE2arn[d]

[a]The use of lowercase letters for the modifications helps distinguish them from either amino acids (identified by their single letter codes) or histones (such as H2A), for which letters are always uppercase.

[b]The nomenclature starts from the left with the histone, then the residue, then the modification. In cases where the modified residue is not known, or not relevant, the modification should follow the histone, for example H4ac and H2Bar1. Multiple modifications can be accommodated by simply extending the listing (for example, H3K4me3K9acS10ph?) for a long as necessary. Because each individual modified residue begins with the uppercase letter specifying the amino acid, and because the modifications themselves are all designated by lowercase letters, the use of commas or dots to separate the individual modified residues in a 'word' specifying multiple modifications is not necessary. On occasion, the presence of an unmodified residue may be an essential component of an information-bearing combination of residues; in this case the residue should be inserted without additions (for example, H3K9S10ph) to indicate H3 unmodified at Lys9 and phosphorylated at Ser10.

[c]Polyubiquitylated histones are designated ubn.

[d]Hypothetical at present.

The nomenclature set out here was devised following the first meeting of the Epigenome Network of Excellence (NoE), at the Mendel Abbey in Brno, Czech Republic. For this reason, it can be referred to as the Brno nomenclature.

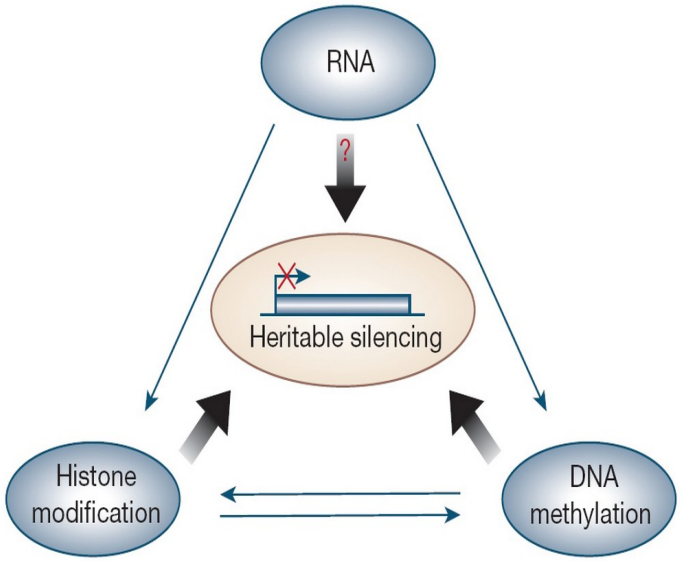
Nat Struct Mol Biol. 2005 Feb;12(2):110-2.

Reading signals on the nucleosome with a new nomenclature for modified histones.

Turner BM.

Epigenetic

- Epigenetics involves genetic control by factors other than an individual's DNA sequence
  - Histone modifications
  - DNA methylation
- Epigenetic modifications may be inherited mitotically or meiotically



Epigenetic Influences and Disease

By: Danielle Simmons, Ph.D. (*Write Science Right*) © 2008 Nature Education

Epigenetic and cancer

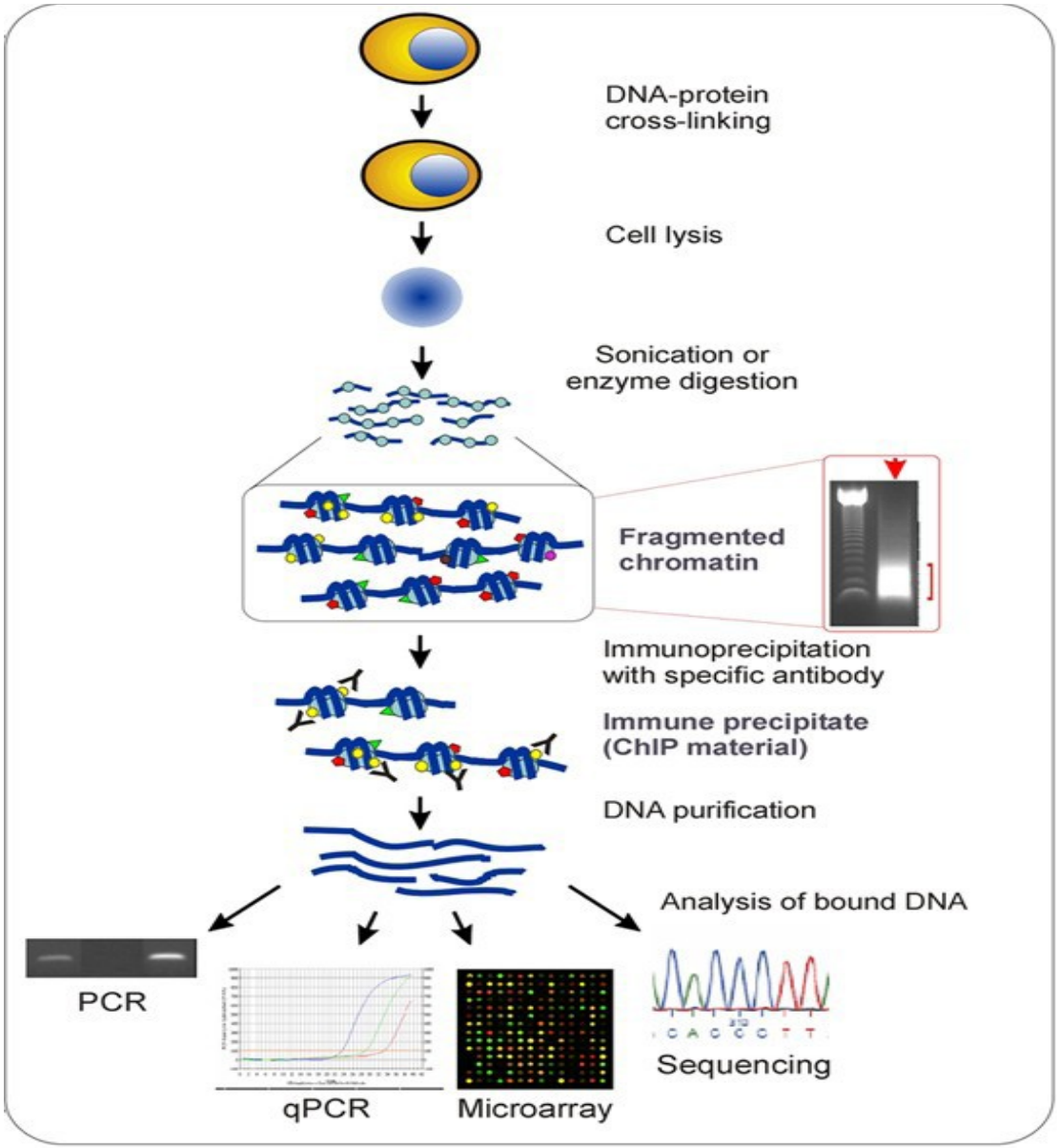
Table 1 Selection of epigenetic genes disrupted in human cancer

Category	Gene	Disruption in cancer	Tumor type
DNA methyltransferases	DNMT3A	Point mutation, translocation, overexpression	AML, MDS, MPD
	DNMT3B	Polymorphism, amplification, overexpression	Colorectal, lung, breast, pancreatic
	DNMT1	Overexpression	Colorectal
Methyl-binding proteins	MBD1	Overexpression	Pancreatic
	MBD2	Overexpression	Glioblastoma
	MBD3	Overexpression	Lung, glioblastoma
	MBD4	Point mutation	Colorectal, stomach, endometrium
DNA hydroxylases	TET1	Translocation	AML
	TET2	Point mutation	AML, MPD, MDS, CMML
Chromatin remodeling complexes	BRG1	Deletion, point mutation	Lung
	BRM	Down regulation	Lung
	ARID1A	Downregulation, deletion, point mutation	Ovarian, endometrioid, gastric, colorectal, breast, cervix, renal, childhood neuroblastoma
	ARID1B	Deletion, point mutation	Childhood neuroblastoma
	ARID2	Point mutation	Lung, hepatocellular carcinoma
	SNF5	Point mutation, deletion, translocation	Rhabdoid

Cancer genomics identifies disrupted epigenetic genes

Laia Simó-Riudalbas · Manel Esteller

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

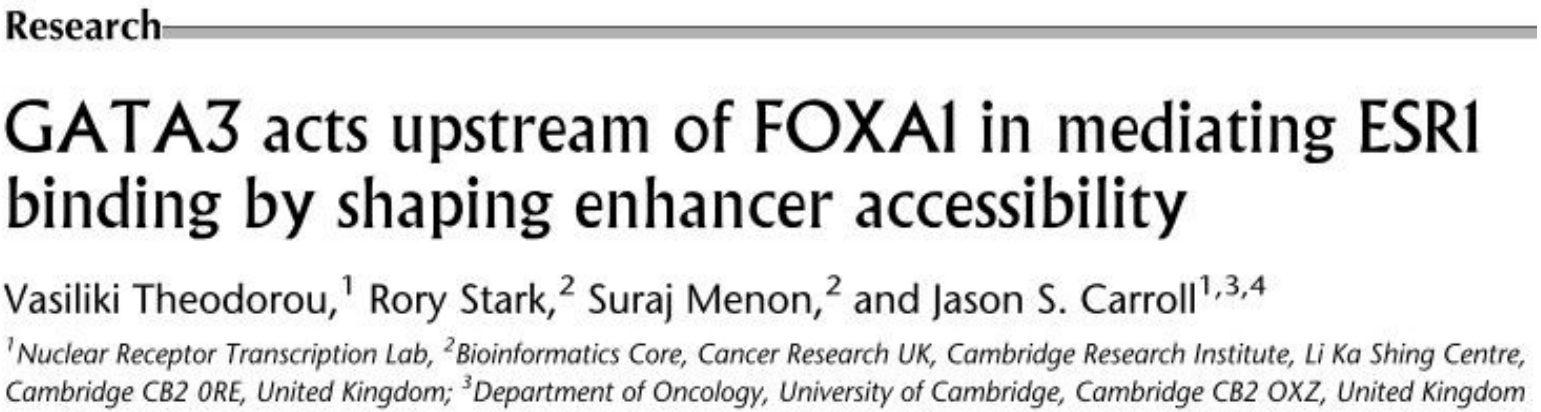
- Open chromatin regions are easier to shear
  - Higher background signals
  - Two solutions
    - Isotype control antibodies
      - Immunoprecipitate much less DNA than specific antibodies
        - Overamplification of particular genomic regions during the library construction step (PCR)
          - Duplicate 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.

Datasets used



Research

GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility

Vasiliki Theodorou,[1] Rory Stark,[2] Suraj Menon,[2] and Jason S. Carroll[1,3,4]

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

- estrogen-receptor (ESR1) is a key factor in **breast cancer 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 ...

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)

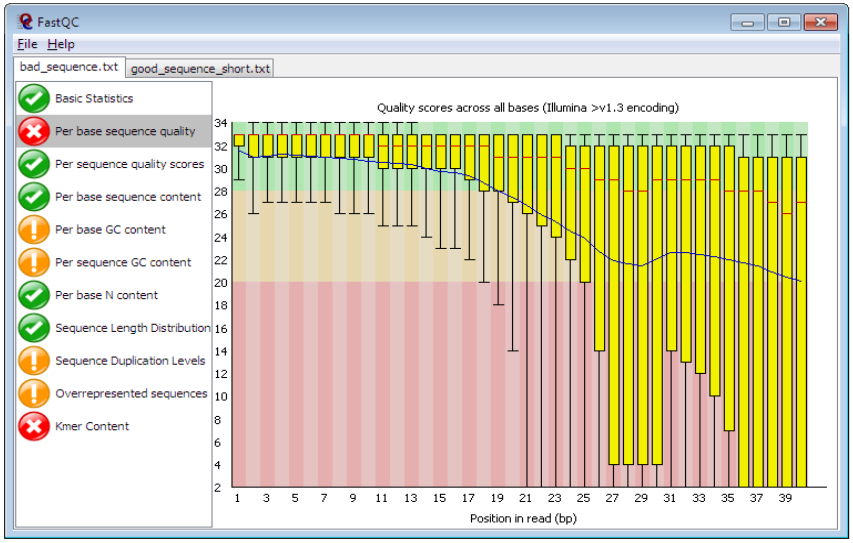
ASCII 33

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

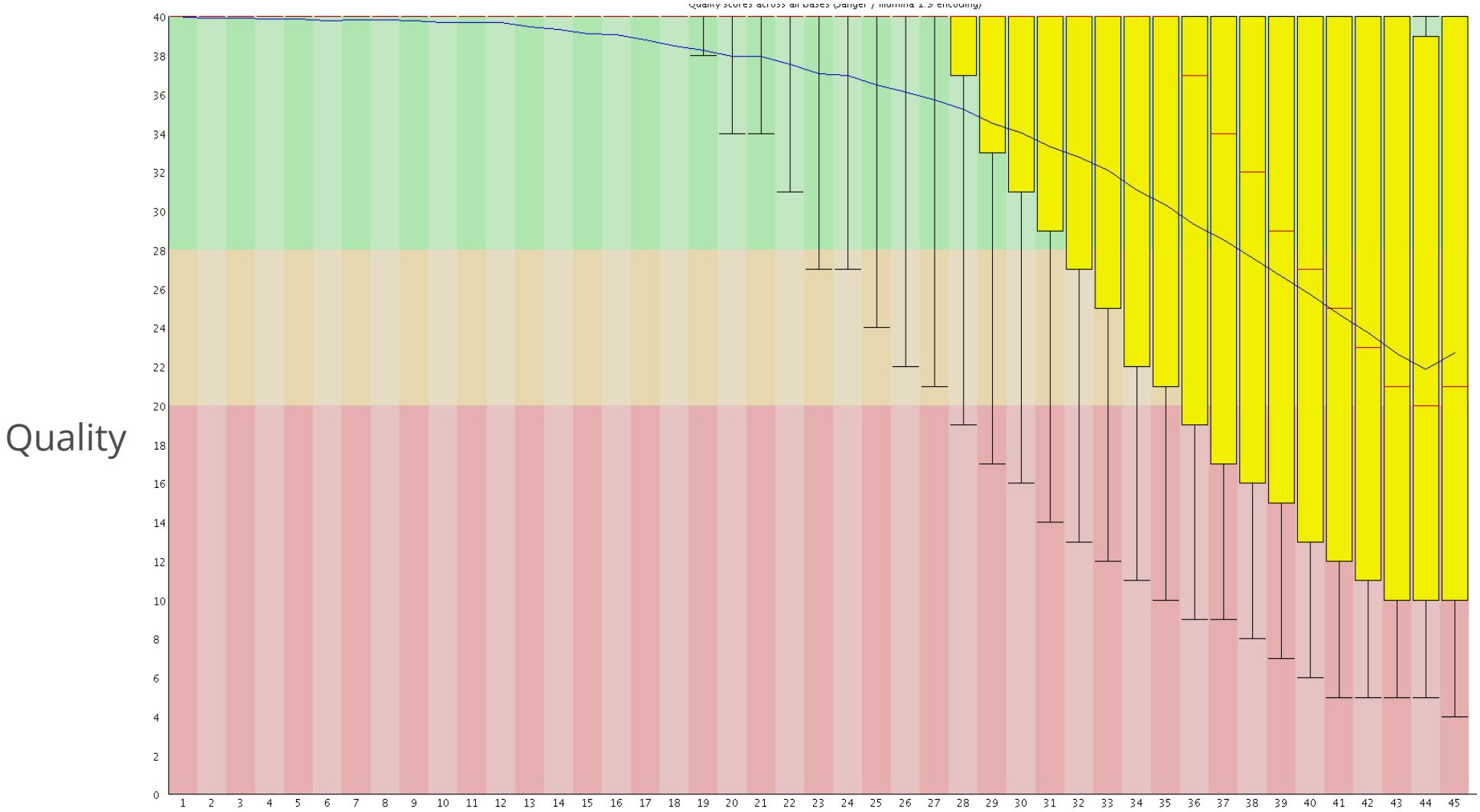


  - ShortRead
    - Bioconductor package

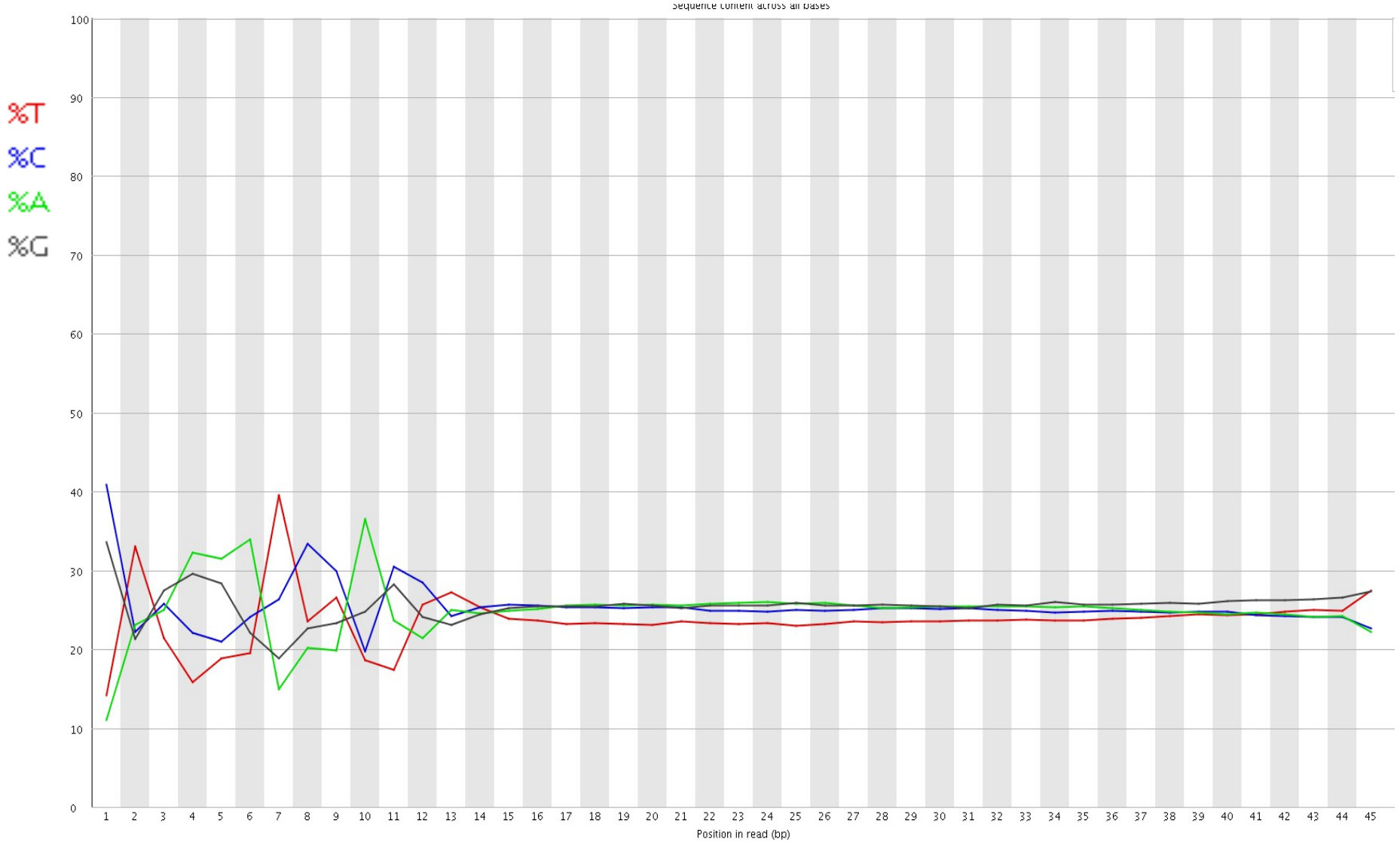
Trimming

- Depending on the aligner this step can be mandatory
- Tools
  - FASTX-Toolkit
  - Sickle
    - Window-based trimming (unpublished)
  - ShortRead
    - Bioconductor package
  - ...

Quality control with FastQC

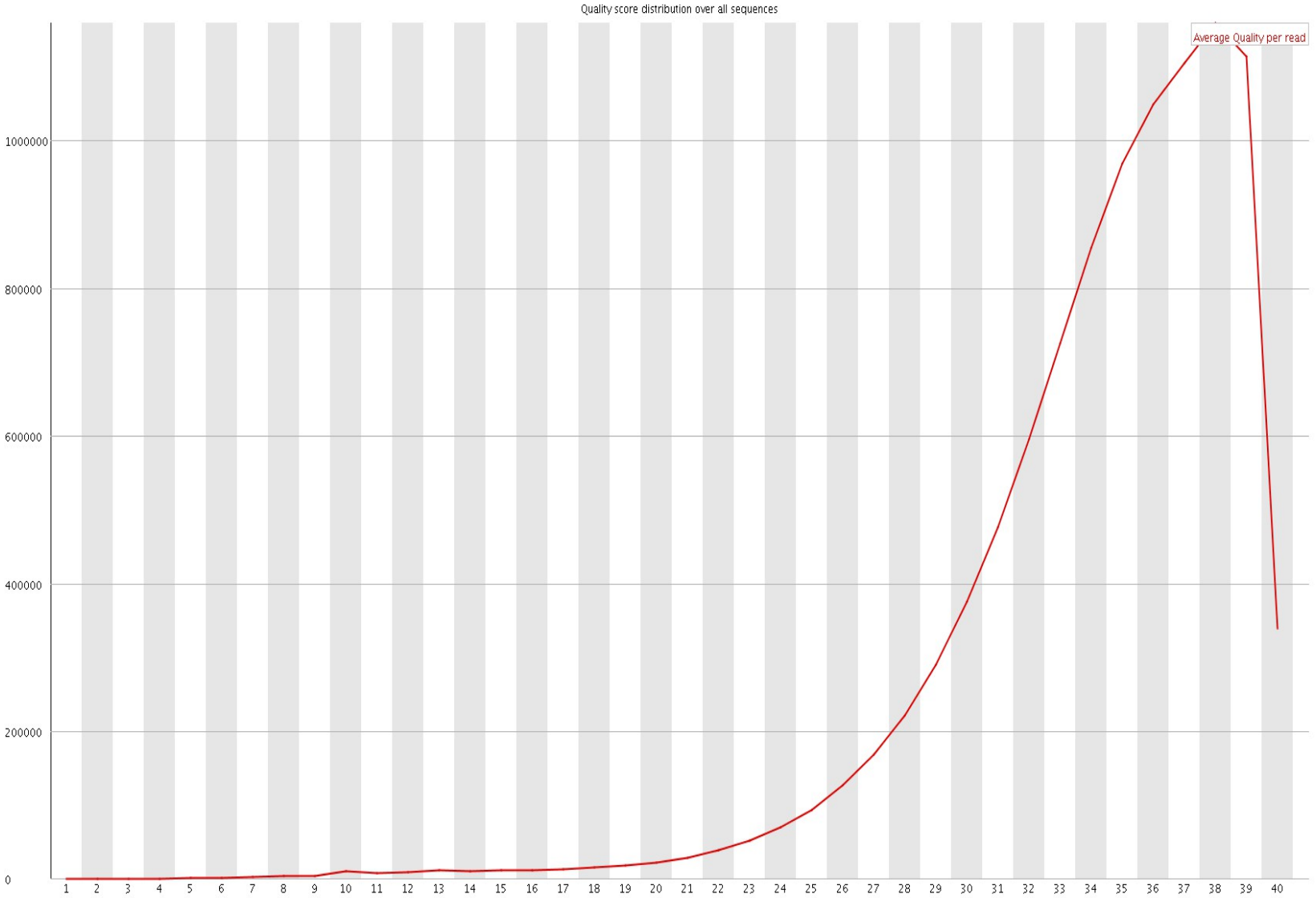


Quality control with FastQC



Position in read

Quality control with FastQC



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.

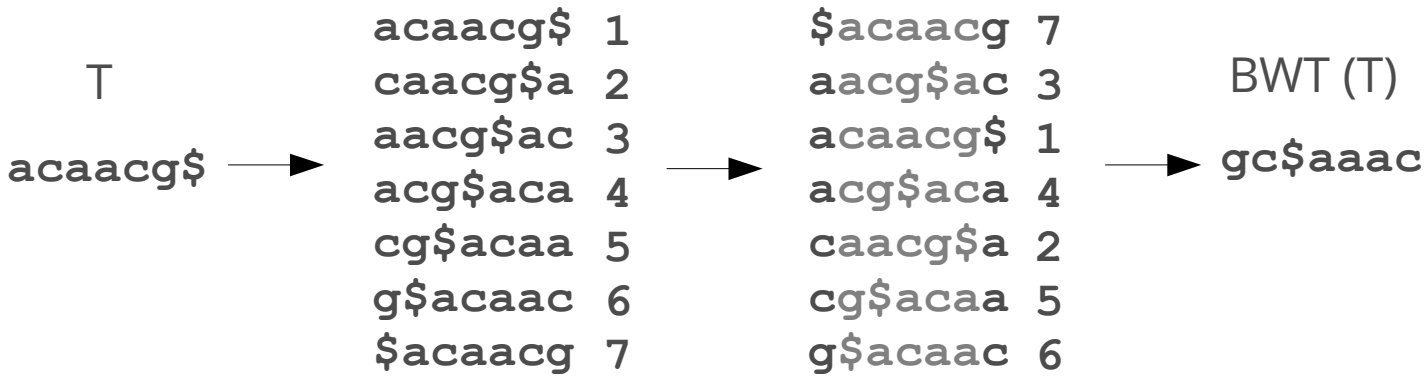
A survey of sequence alignment algorithms for next-generation sequencing.

Li H, Homer N.

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

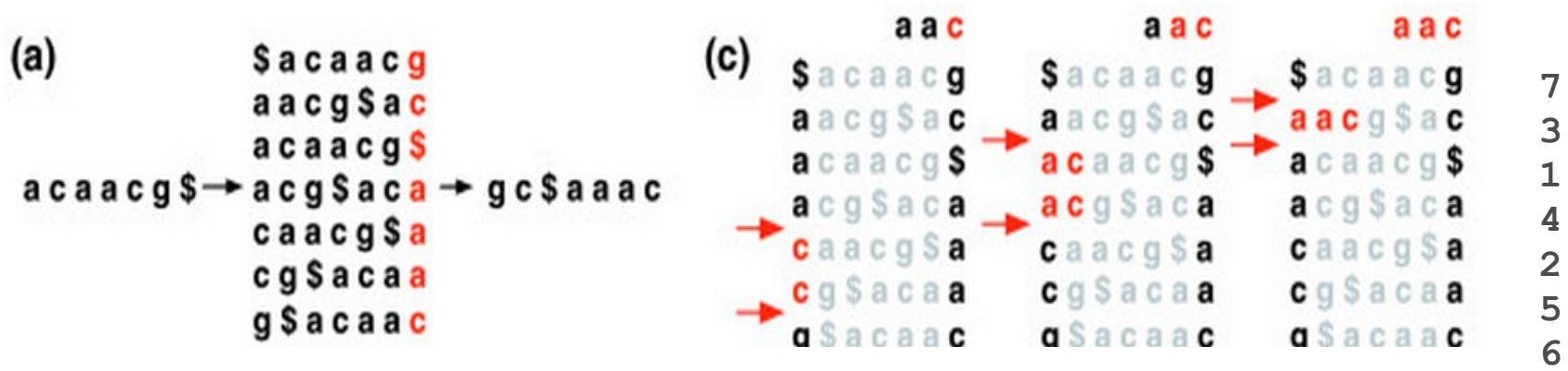
Bowtie principle

- 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

- SAM = 'Sequence Alignment/MAP'
  - BAM: binary/compressed version of SAM
  - Store information related to alignments
    - QNAME : Read ID
    - FLAG: Bitwise Flag
    - RNAME : Reference name (e.g chromosome)
    - POS: start of alignment
    - MAPQ: Mapping Quality
    - CIGAR: CIGAR String
    - RNEXT: Name of the mate
    - ...

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 …
- See: https://broadinstitute.github.io/picard/explain-flags.html

The extended CIGAR string

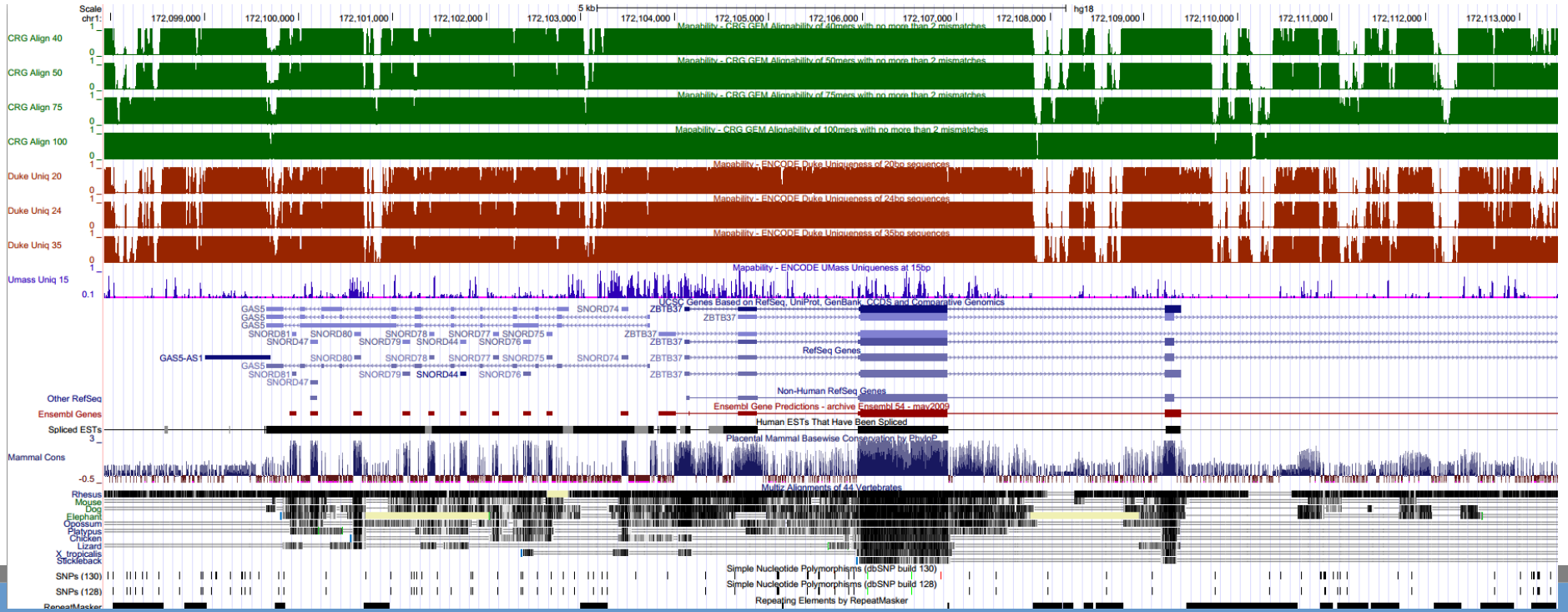
- Example flags:
  - M alignment match (can be a sequence match or mismatch)
  - I insertion to the reference
  - D deletion from the reference
  - http://samtools.sourceforge.net/SAM1.pdf

```
ATTCAGATGCAGTA
ATTCA--TGCAGTA

 5M2D7M
```

Mappability issues

- Mappability: sequence uniqueness of the reference
- Mappability = 1/(#genomic position for a given word)
- Mappability of 1 for a unique k-mer
- Mappability < 1 for a non unique k-mer

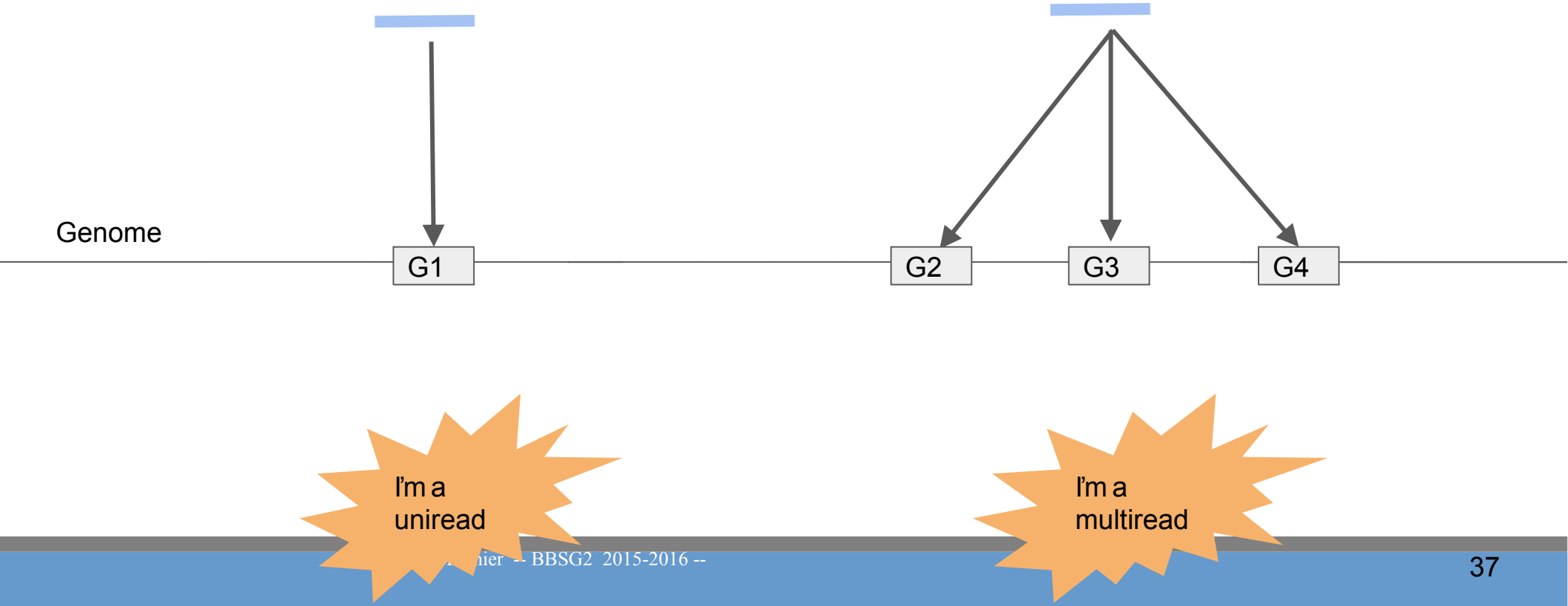


Uniread ? Multireads ?

- Several aligners still use this notion
  - E.g bowtie(1)
- The notion has been superseded by the mapping quality score.
  - Mapping quality score indicates is computed from the probability that alignement is wrong
  - -log10(prob. alignment is wrong)
- It is particularly advised to take into account mapping quality (e.g by selecting high quality alignments from the BAM file)
  - Samtools view -q 30 file.bam

Uniread ? Multireads ?

- First aligners defined the notions of uni-reads and multireads
- An uniread is thought to map to a single position on the genome
- A multiread is thought to map to several position on the genome
  - Which position/gene produced the signal ?



Uniread ? Multireads ?

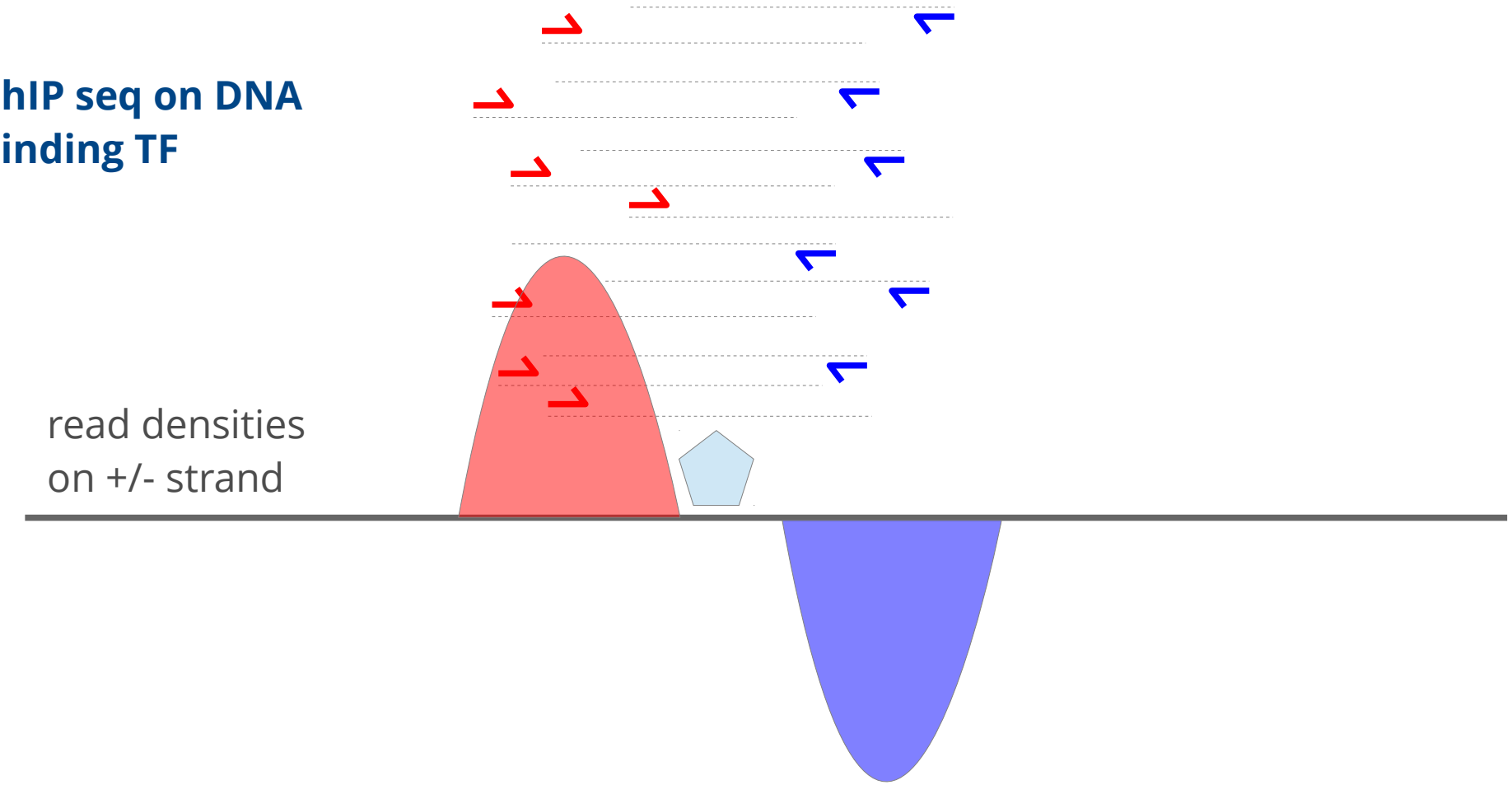
- Several aligners still use this notion
  - E.g bowtie(1)
- The notion has been superseded by the mapping quality score.
  - Mapping quality score indicates is computed from the probability that alignement is wrong
  - -log10(prob. alignment is wrong)
- It is particularly advised to take into account mapping quality (e.g by selecting high quality alignments from the BAM file)
  - Samtools view -q 30 file.bam

PCR Duplicates

- Main Issue in ChIP-Seq:
  - PCR duplicates
    - Related too poor library complexity
    - The same set of fragments are amplified
      - Indicate that Immuno-precipitation failed

ChIP-seq signal for transcription factors on single end dataset

ChIP seq on DNA binding TF



We expect to see a typical strand asymmetry in read densities

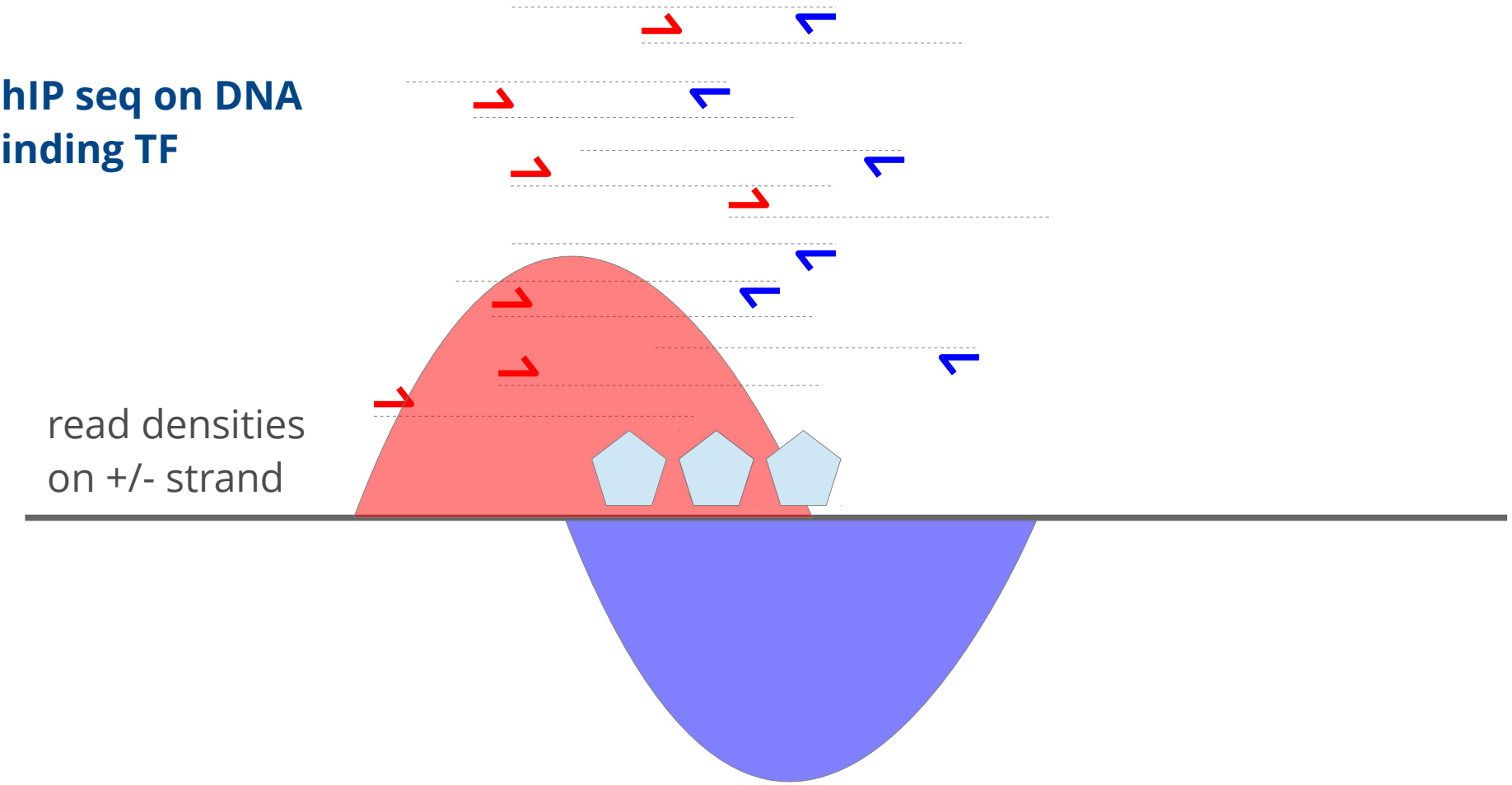
→ ChIP peak recognition pattern (**sharp peaks)**

ChIP-seq signal for transcription factors

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

ChIP-seq signal for transcription factors

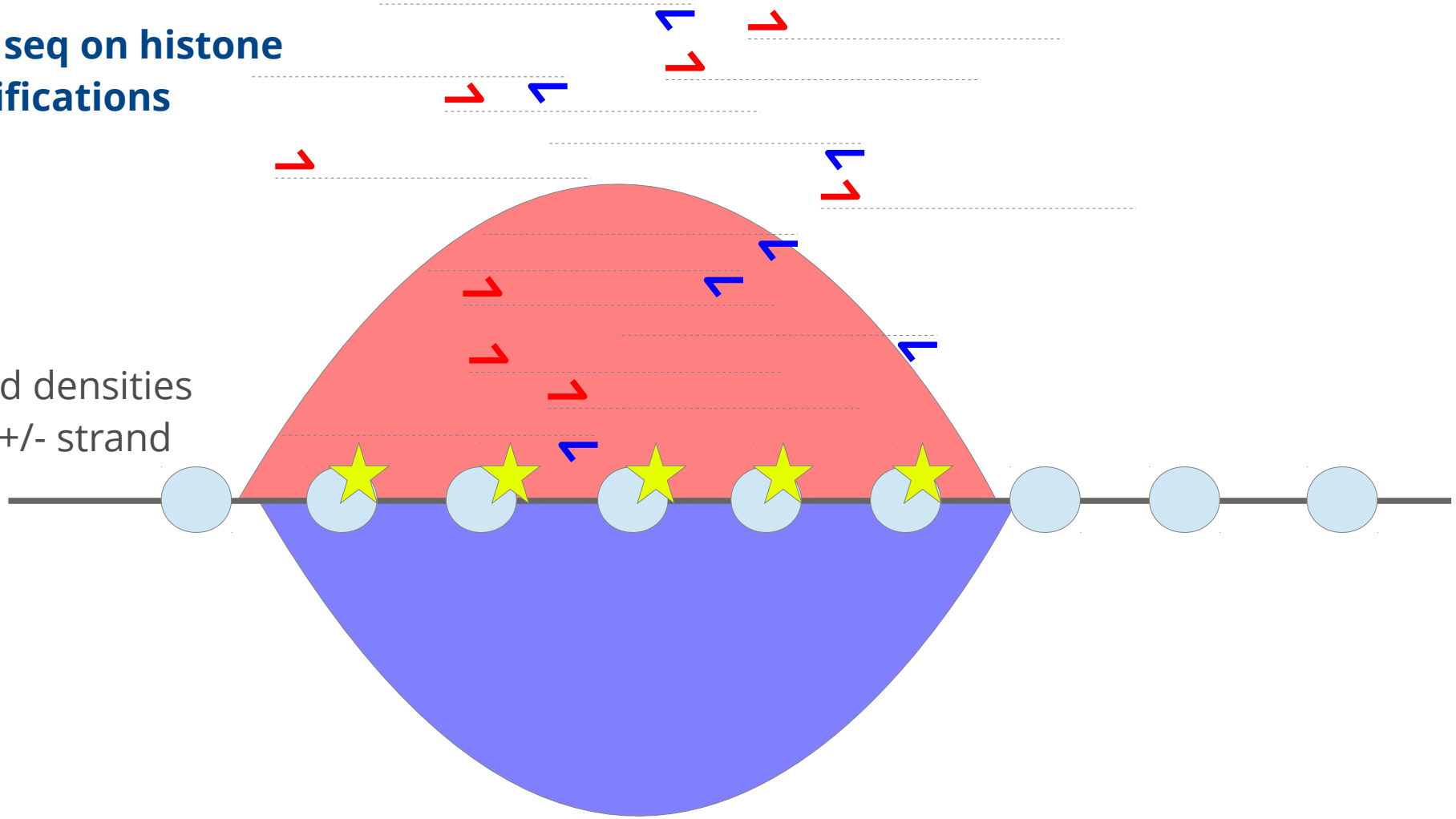
ChIP seq on DNA binding TF



Binding of several TF as complexes tend to blur this asymmetry

ChIP-seq signal for histone marks

ChIP seq on histone modifications



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

Real example of ChIP-seq signal

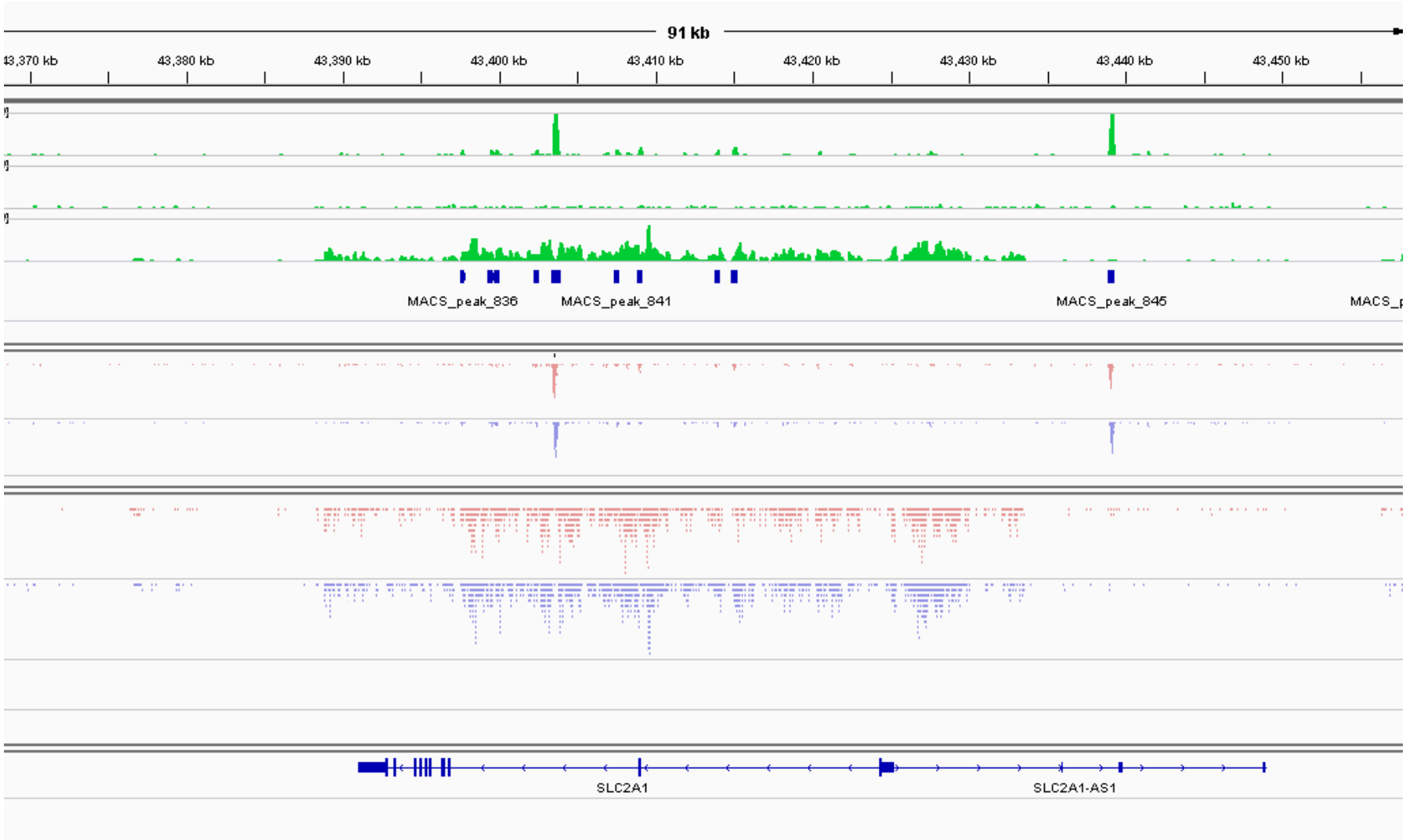
ESR1

input

H3K4me1

ESR1 reads

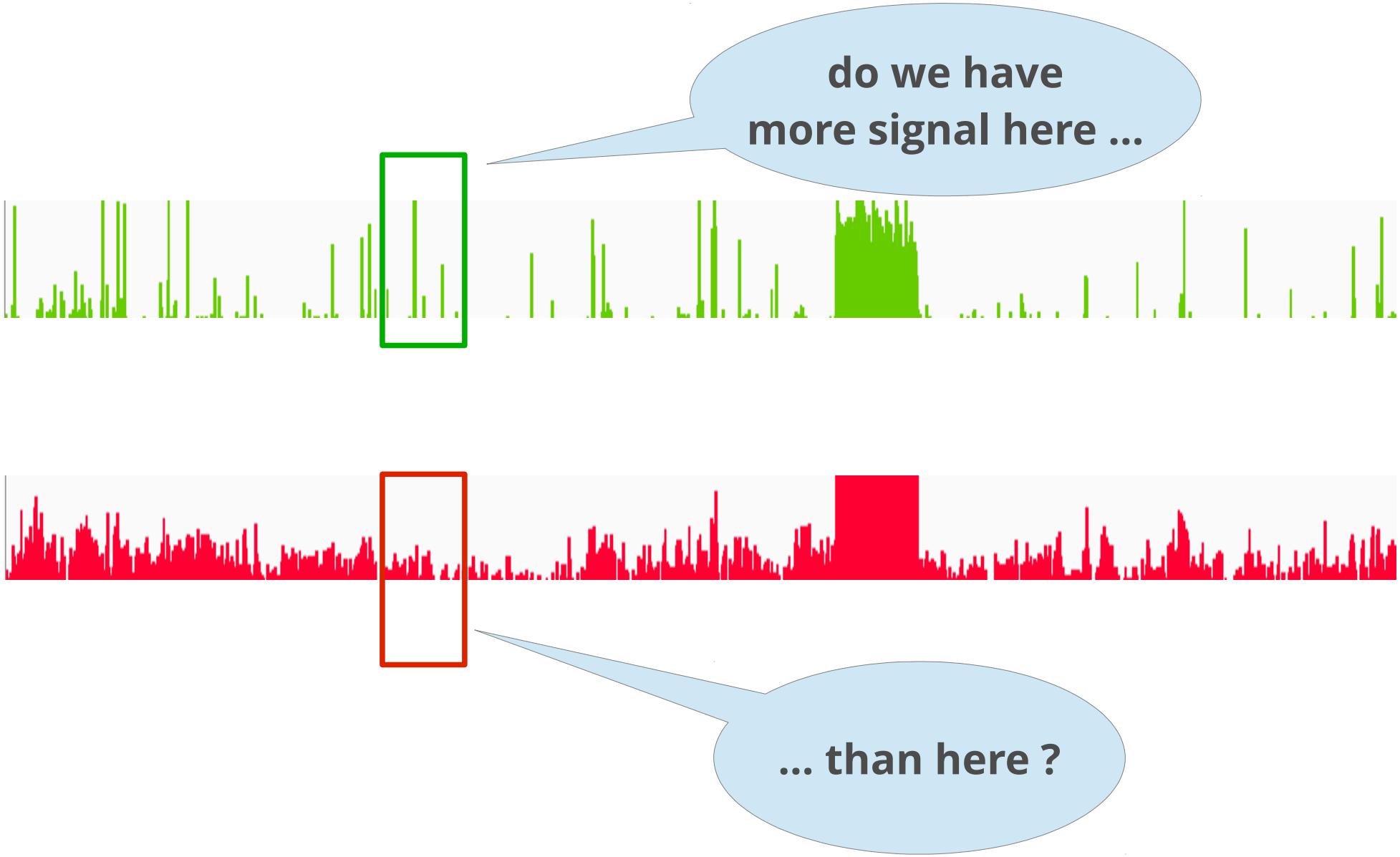
H3K4me1 reads



Keys aspects of “peak” finding

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

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

ChIP-Seq quality control

1. Quality control

- **Quantitative**
  - *Fraction of reads in peaks* (FRiP)

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

→ depends on type of ChIP (TF/histone)

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

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

N_1: Genomic positions with 1 read aligned

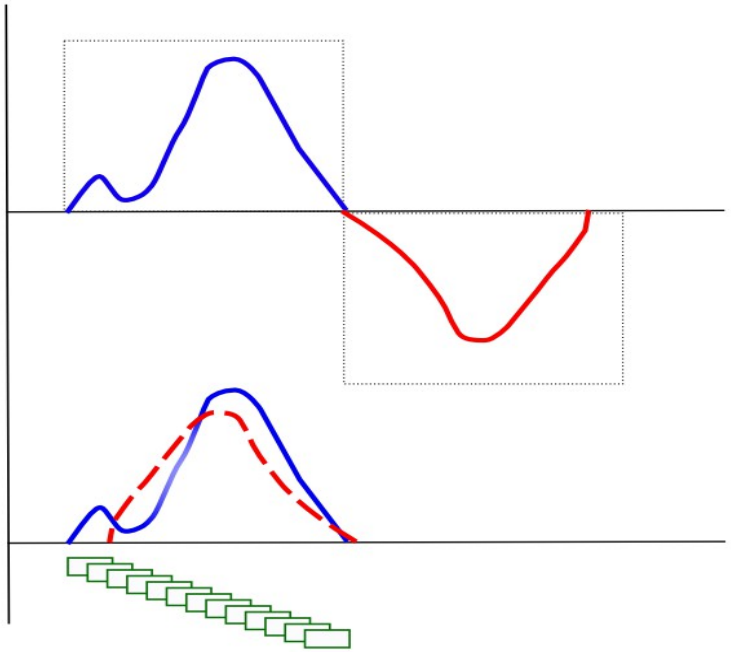
N_d: Genomic positions with ≥ 1 read aligned

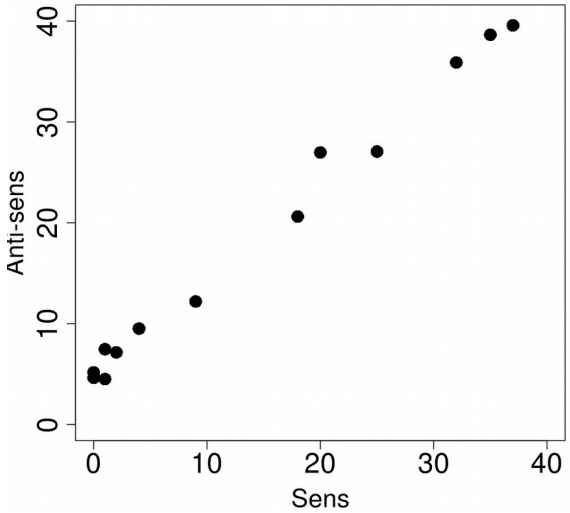
PBC < 0.5 🔴
0.5 < PBC < 0.8 🟡
0.8 < PBC 🟢

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

1. Quality control

- Strand cross-correlation analysis





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

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

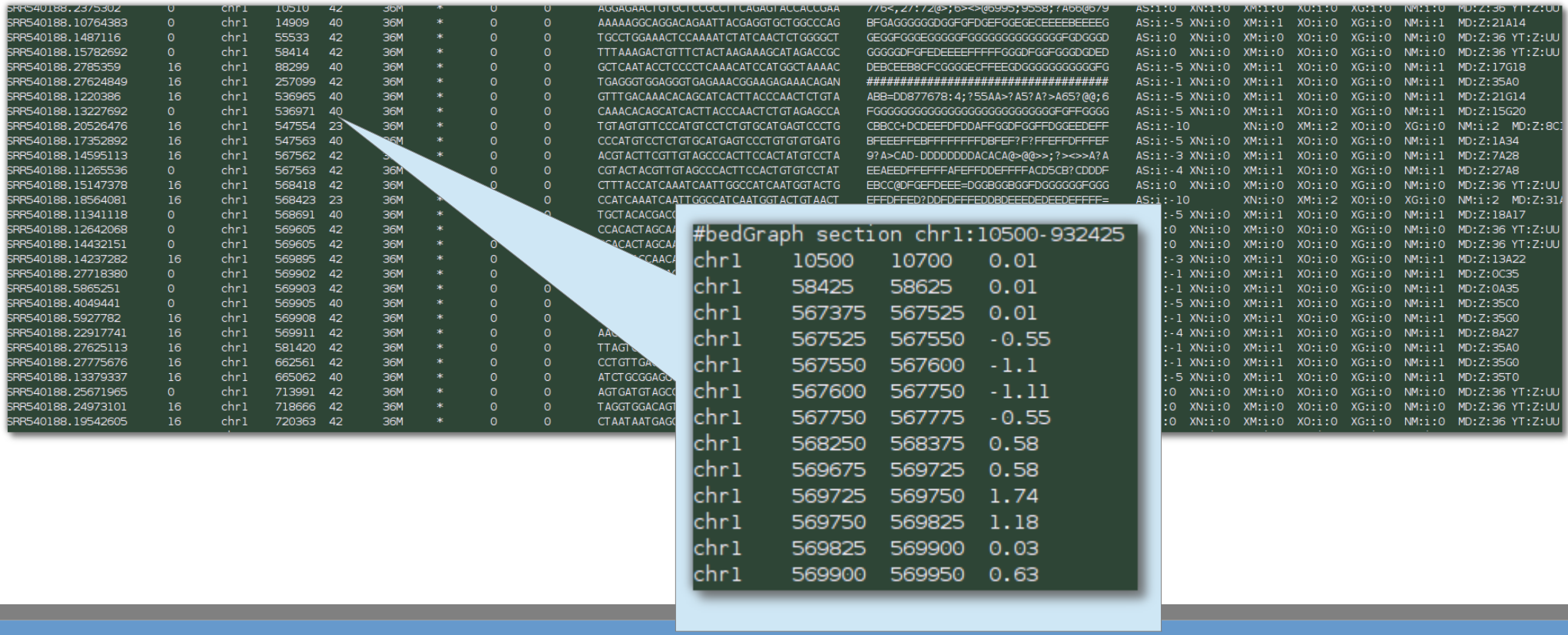
Design and analysis of ChIP-seq experiments for DNA-binding proteins.

Kharchenko PV, Tolstorukov MY, Park PJ.

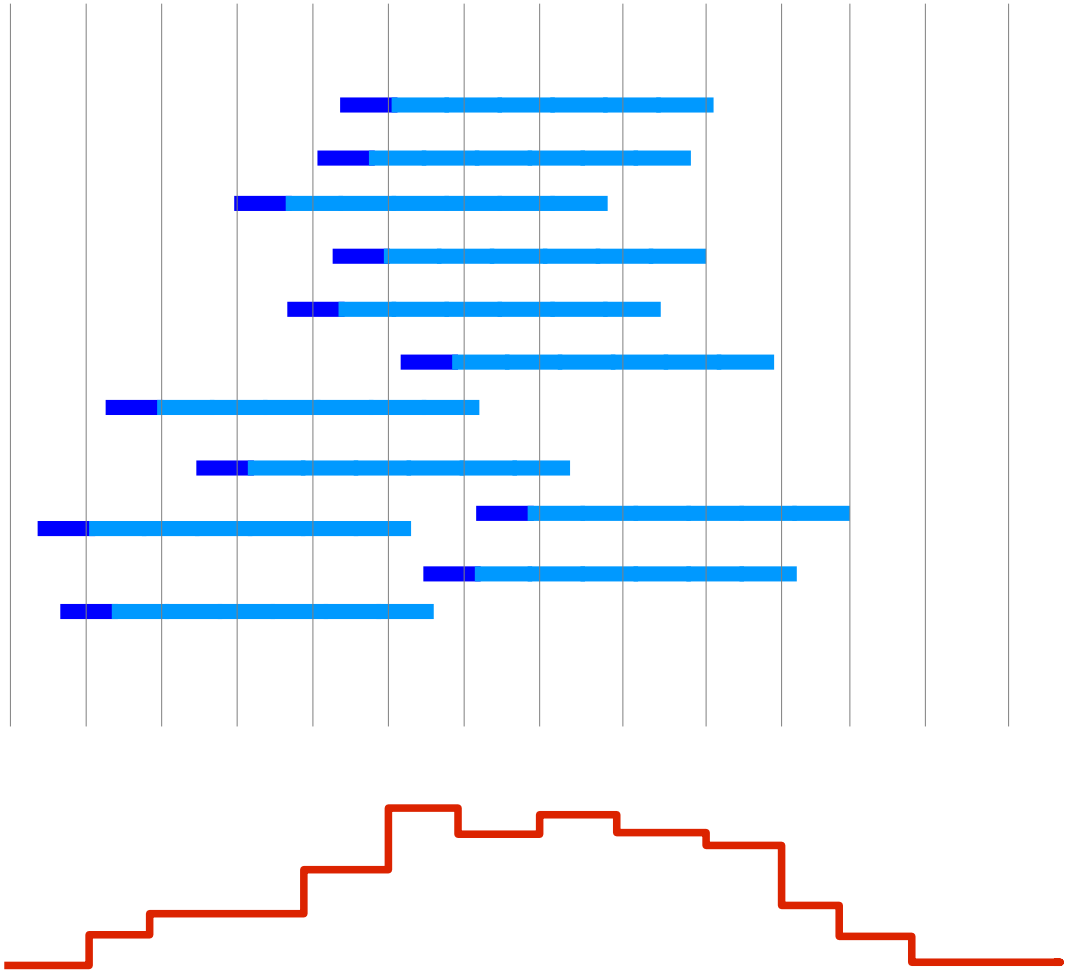
Determine signal coverage

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**



2. from reads to coverage



→ deepTools :
bamCoverage

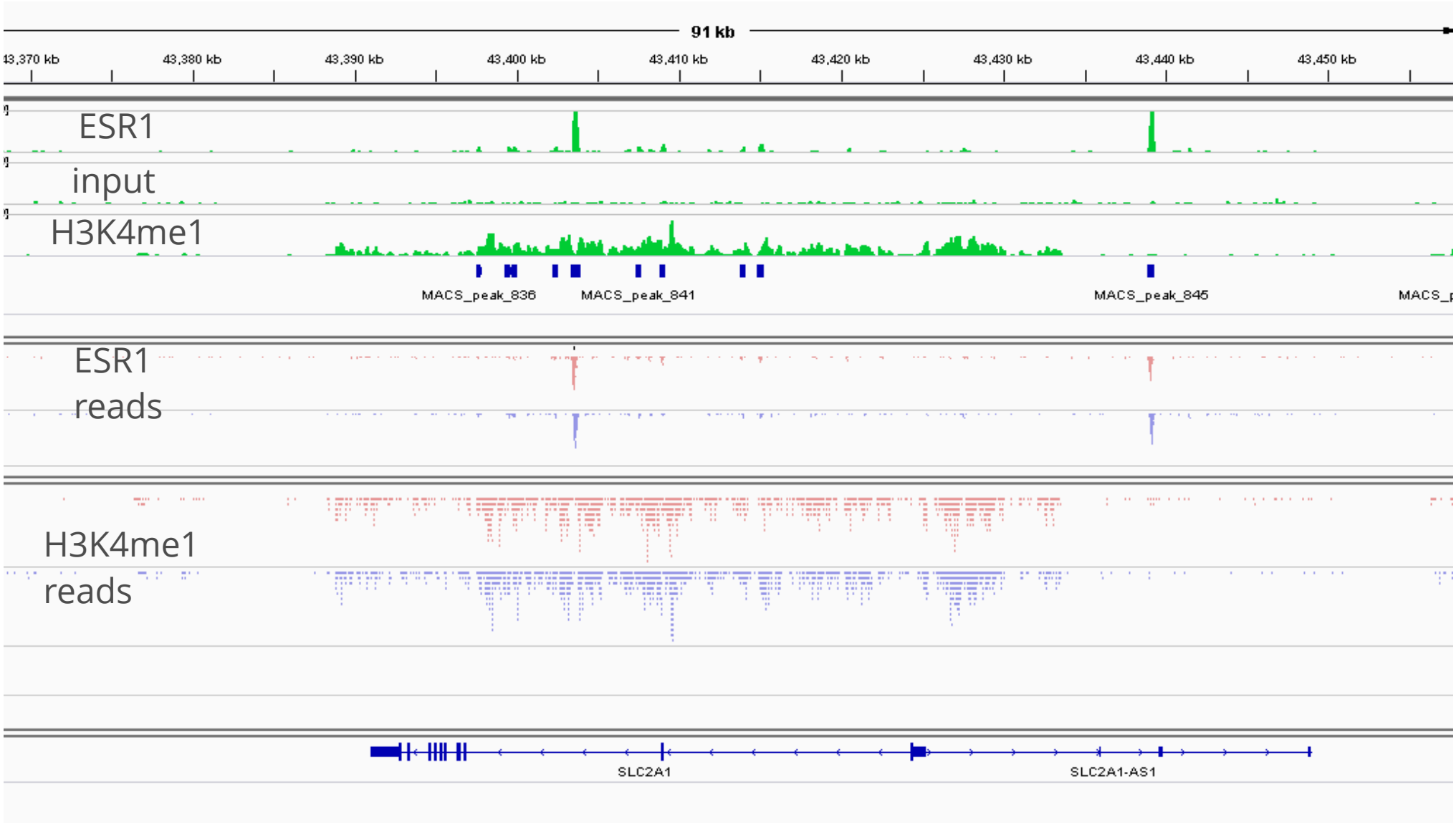
- Reads are extended to 3' to fragment length
- Read counts are computed for each bin
- Counts are normalized
  - reads per genomic content → normalize to 1 x coverage

$$RPGC = \frac{n_{mapped\ reads} \times length_{fragment}}{length_{genome}}$$

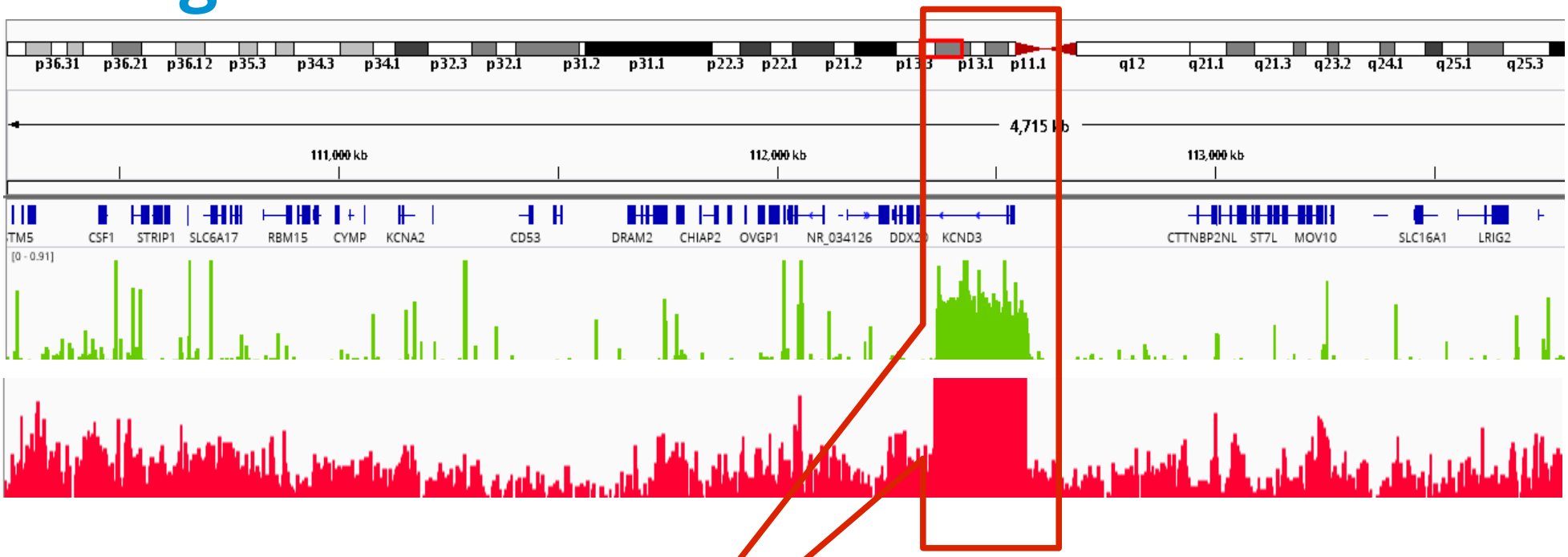
  - reads per kilobase per million reads per bin

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

2. from reads to coverage

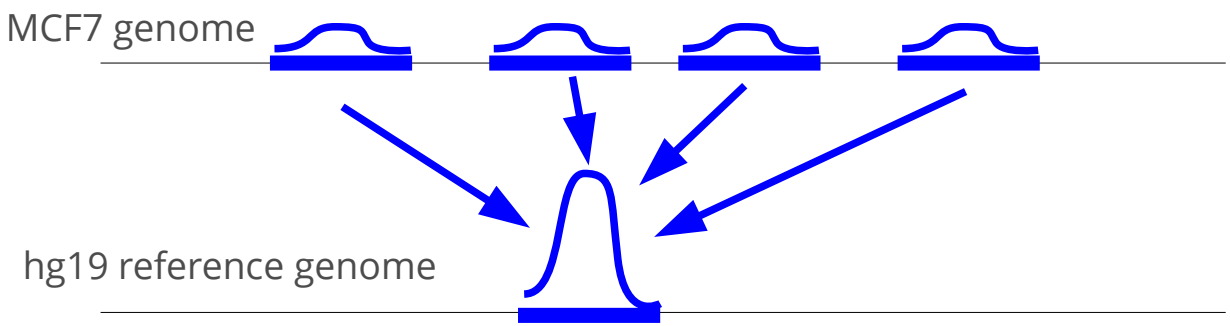


3. signal and noise

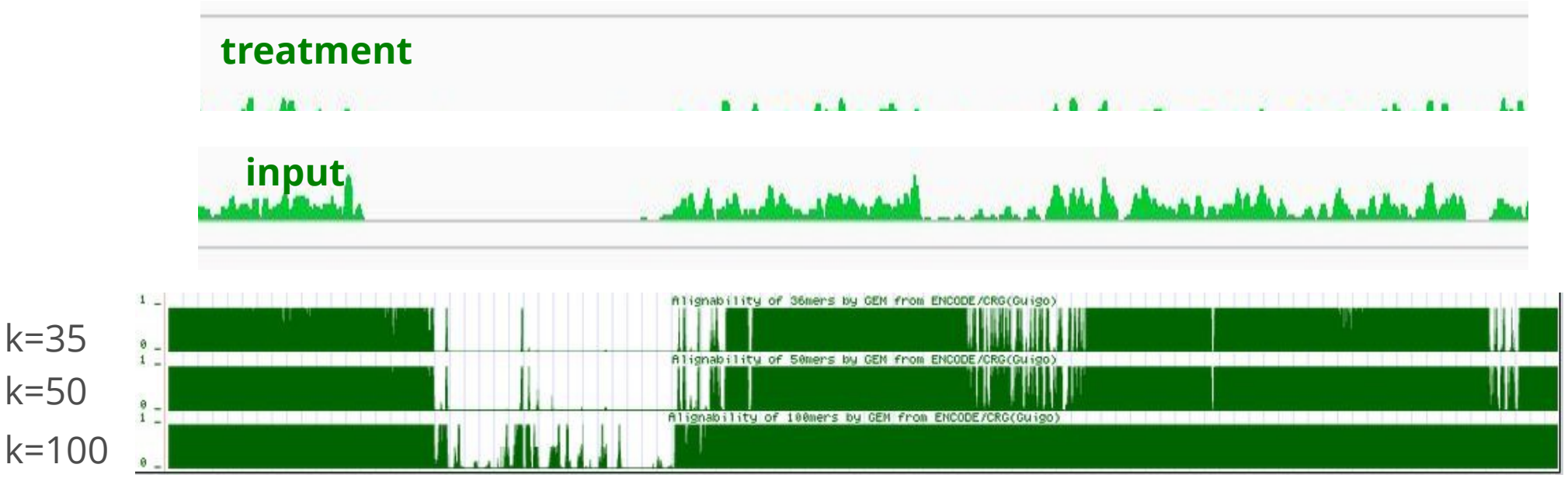


MCF-7 genome

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*);



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

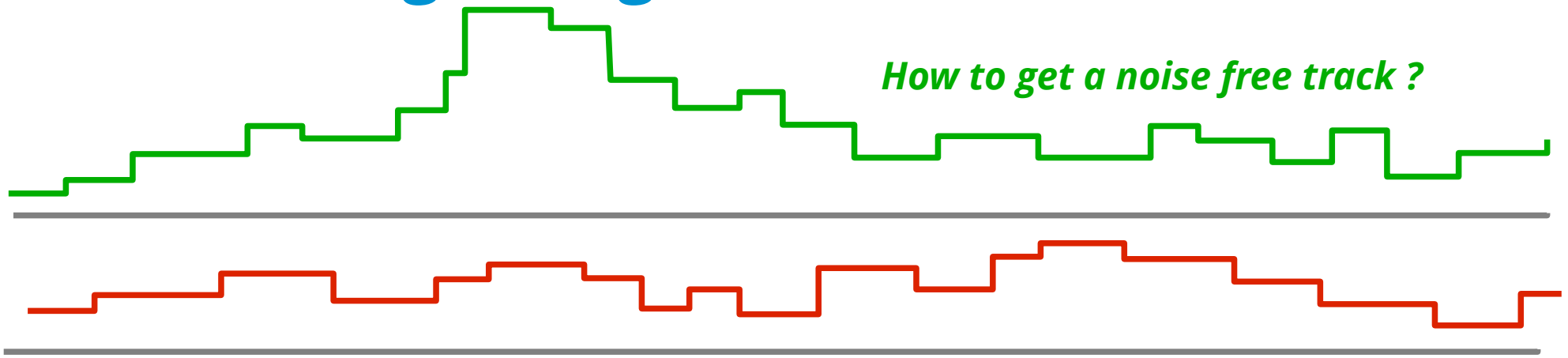
The availability of a control sample in mandatory !

→ mock IP with unspecific antibody

→ sequencing of input (=naked) DNA

→ Preferred

4. modelling background level



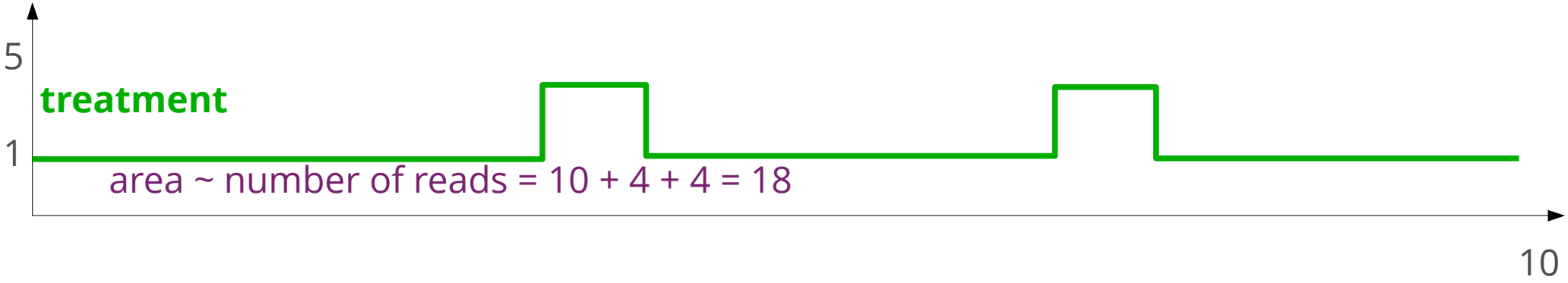
- **naïve subtraction** treatment – input is not possible, because both libraries have different sequencing depth !

- **Solution 1** : before subtraction, scale both libraries by total number of reads (library size)
  - RPGC
  - RPKM

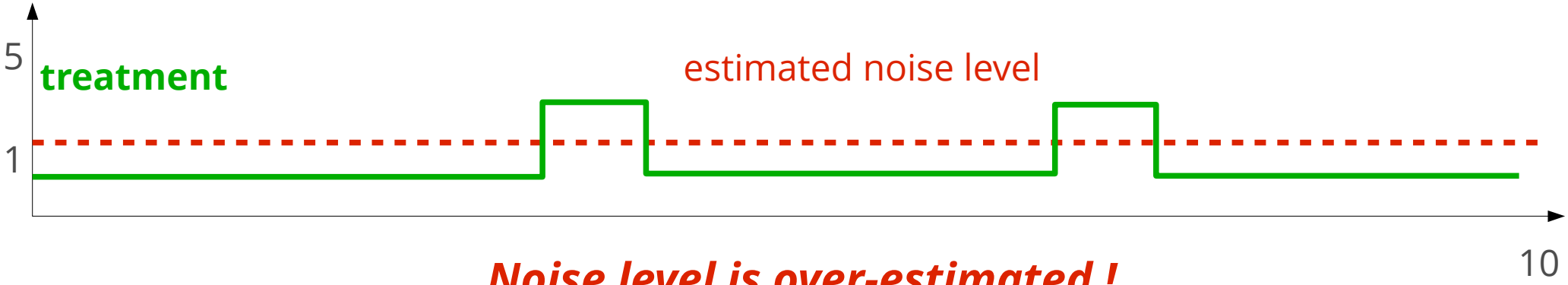
$$RPGC = \frac{n_{mapped\ reads} \times length_{fragment}}{length_{genome}}$$

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

4. modelling background level

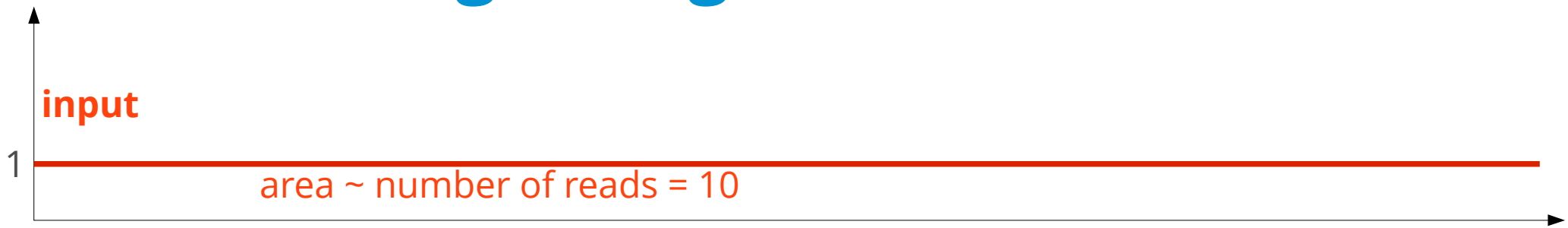


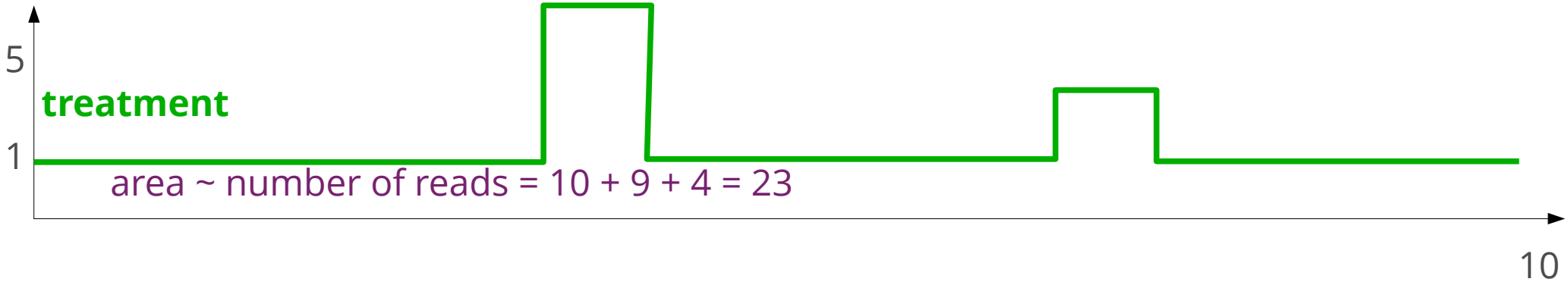
Scaling by library size : upscale input by 18/10 = 1.8



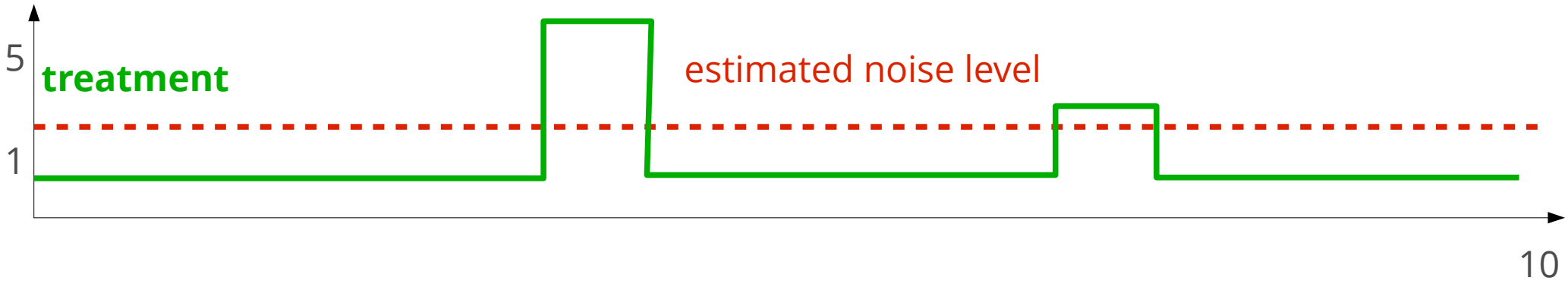
Noise level is over-estimated !

4. modelling background level

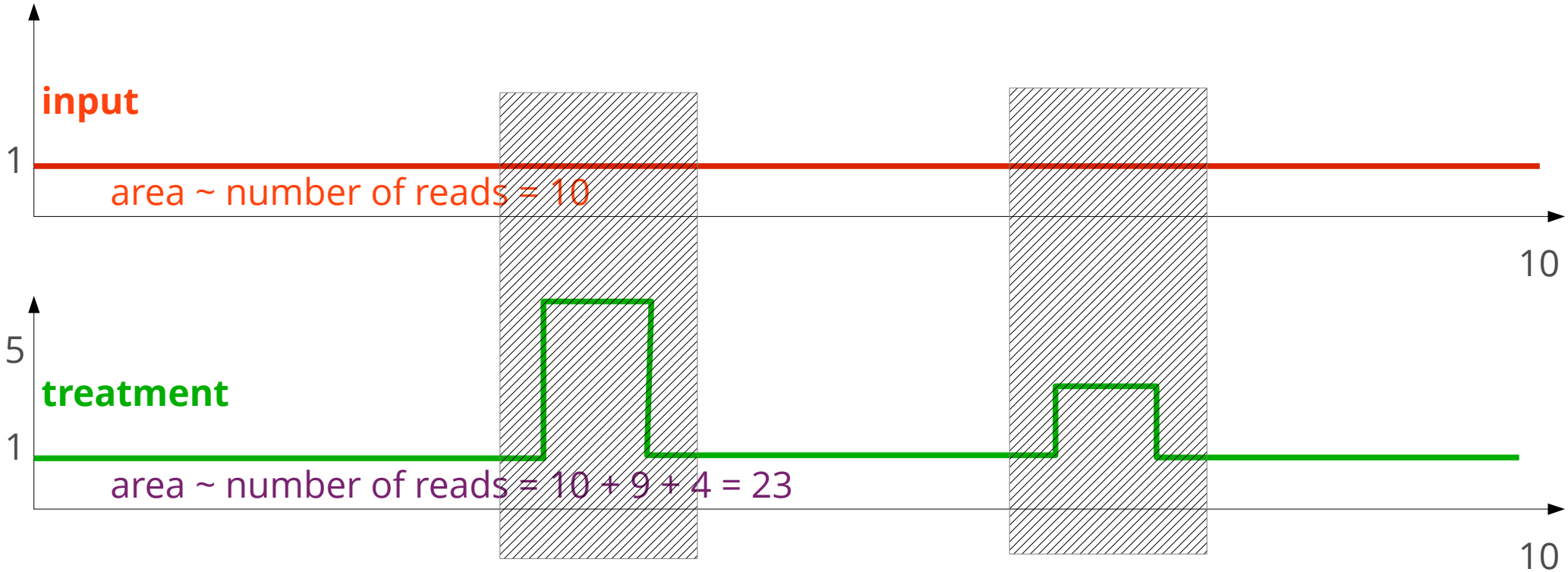




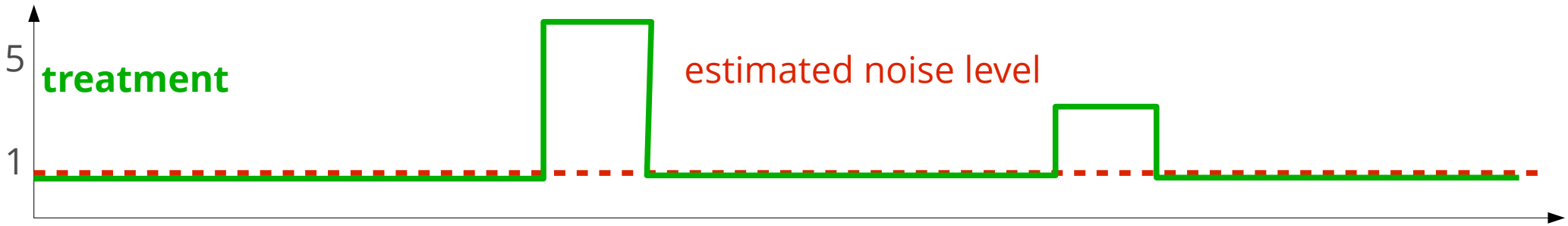
Scaling by library size : upscale input by 23/10 = 2.3



4. modelling background level

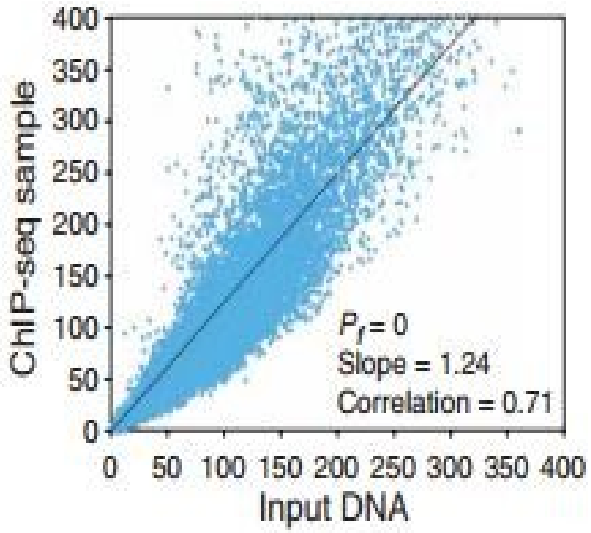


Scaling by library size : upscale input by 23/10 = 2.3

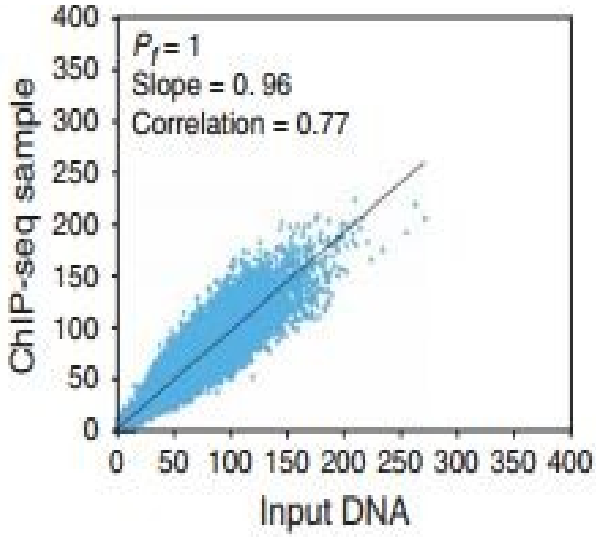


4. modelling background level

- **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 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			

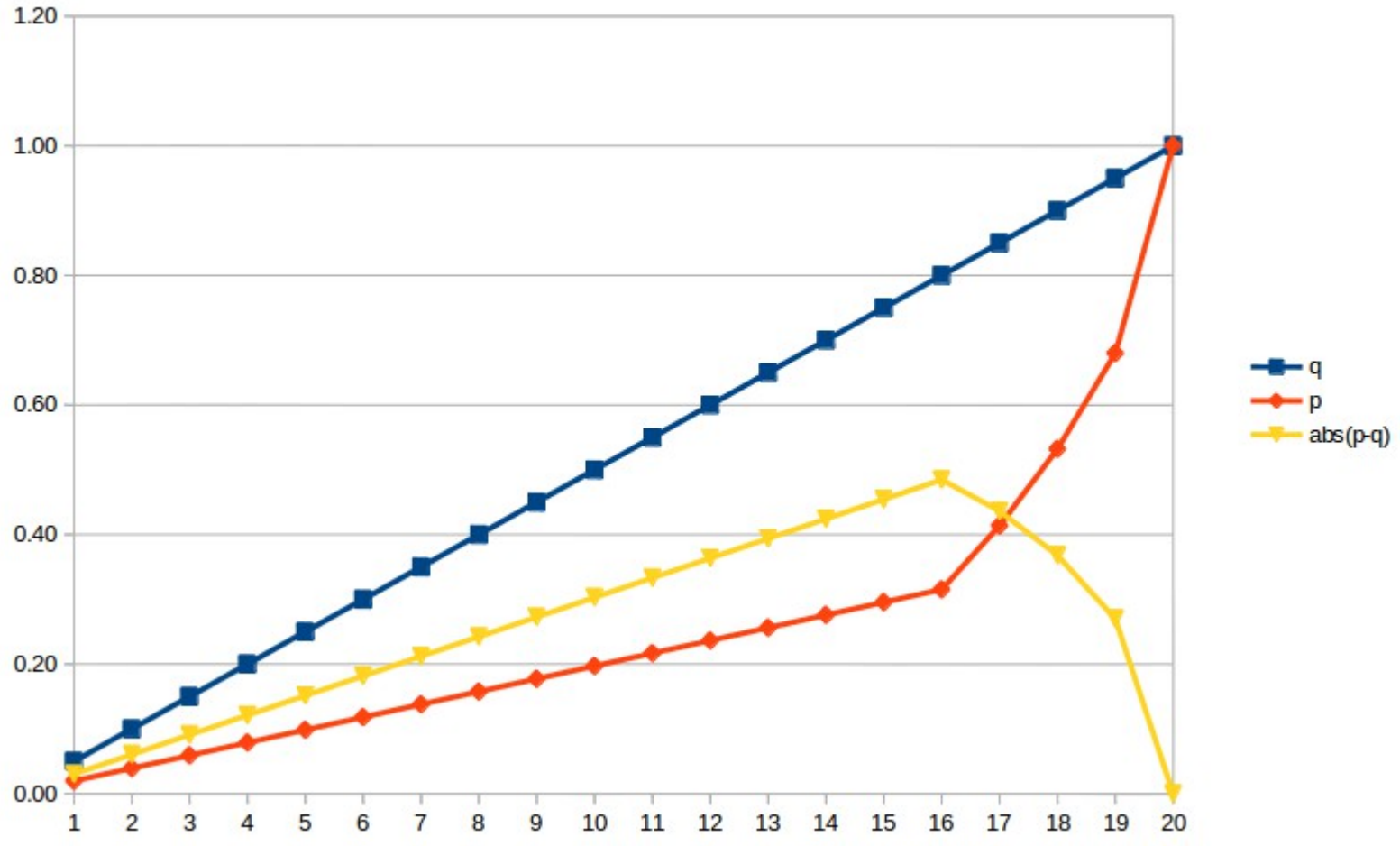
Stat Appl Genet Mol Biol. 2012 Mar 31;11(3):Article 9. doi: 10.1515/1544-6115.1750.

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)



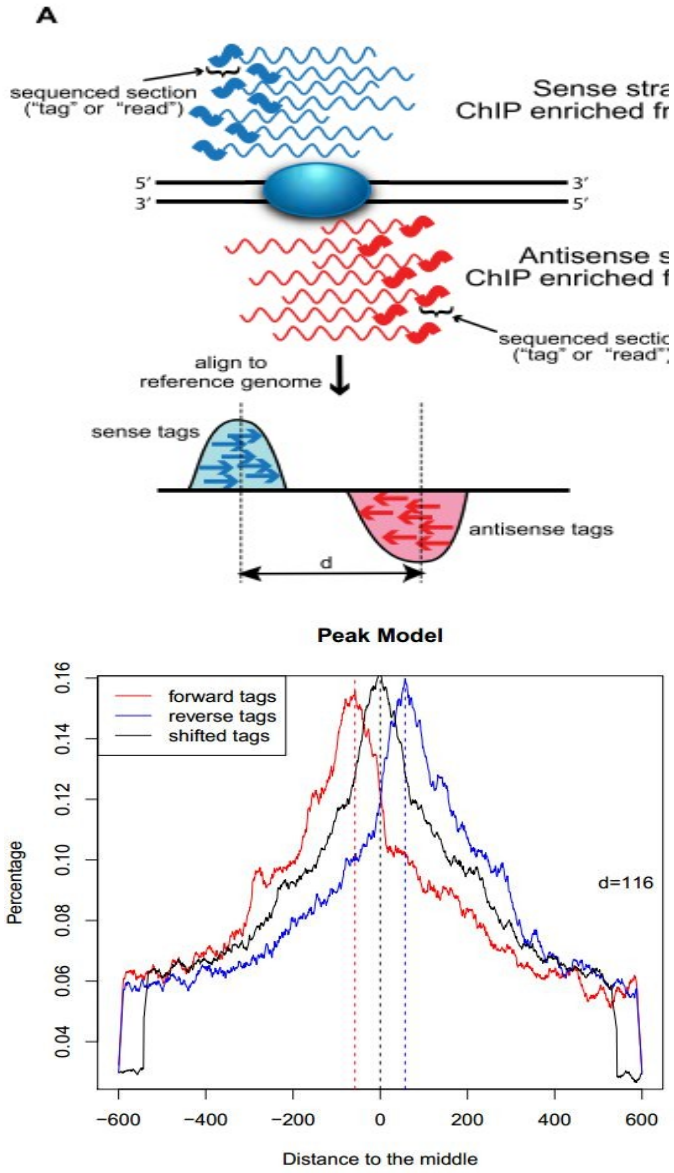
Stat Appl Genet Mol Biol. 2012 Mar 31;11(3):Article 9. doi: 10.1515/1544-6115.1750.

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

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

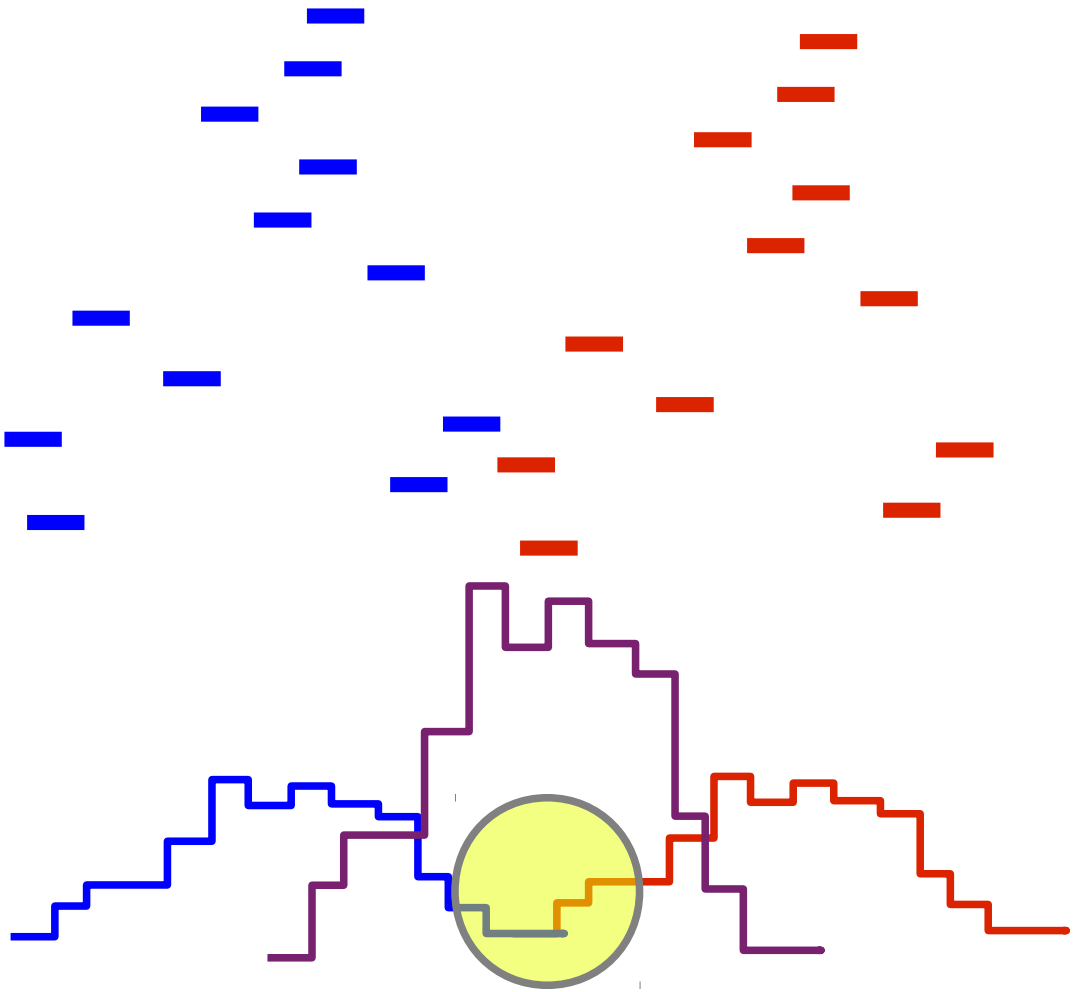
5. from reads to peaks

- **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



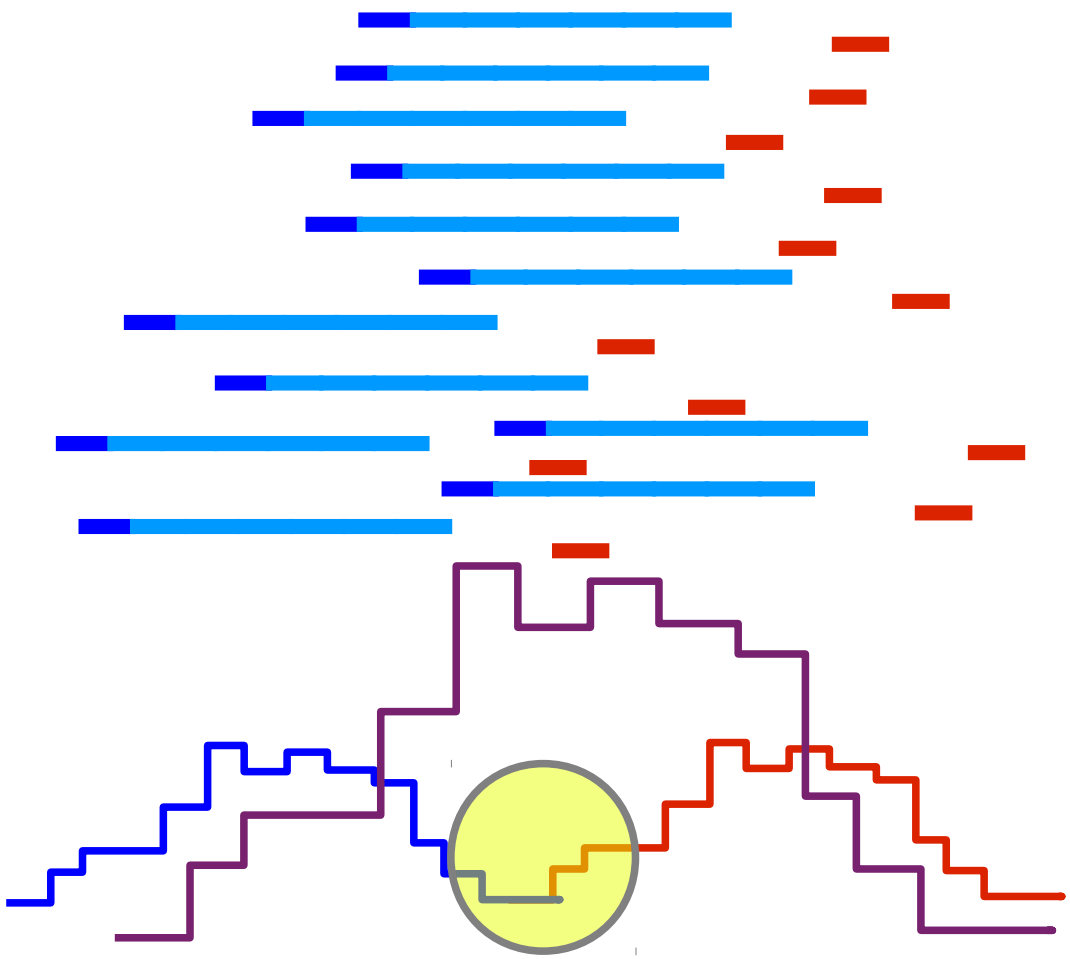
example of MACS model building using top enriched regions

5. from reads to peaks

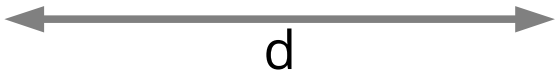


“Read shifting”

5. from reads to peaks



“Read extension”



	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based duplicate[e]
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Optional peak height, ratio to background	Yes / No
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value	No / Yes
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance					
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation					

Computation for ChIP-seq and RNA-seq studies

Shirley Pepke[1], Barbara Wold[2] & Ali Mortazavi[2]

	Profile
CisGenome v1.1	Strand-specific window scan
ERANGE v3.1	Tag aggregation
FindPeaks v3.1.9.2	Aggregation of overlapped tags
F-Seq v1.82	Kernel density estimation (KDE)
GLITR	Aggregation of overlapped tags
MACS v1.3.5	Tags shifted then window scan
PeakSeq	Extended tag aggregation
QuEST v2.3	Kernel density estimation
SICER v1.02	Window scan with gaps allowed
SiSSRs v1.4	Window scan
spp v1.0	Strand specific window scan

Some methods separate the tag densities into different strands and take advantage of tag asymmetry

Most consider merged densities and look for enrichment

	Profile	Peak criteria[a]	Tag shift
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation

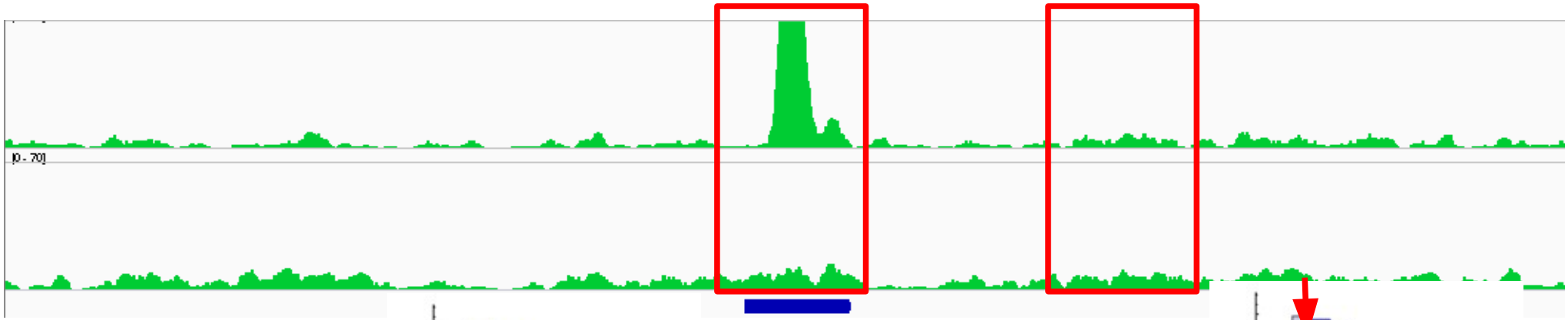
Tag shift

Tag extension

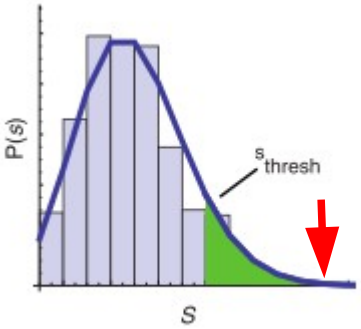
Tags unchanged

5. from reads to peaks

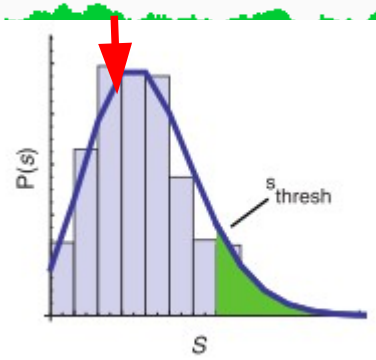
- **Determining “enriched” regions**
  - sliding window across the genome
  - at each location, evaluate the enrichement of the signal wrt. expected background based on the distribution
  - retain regions with P-values below threshold
  - evaluate FDR



Pval < 1e-20



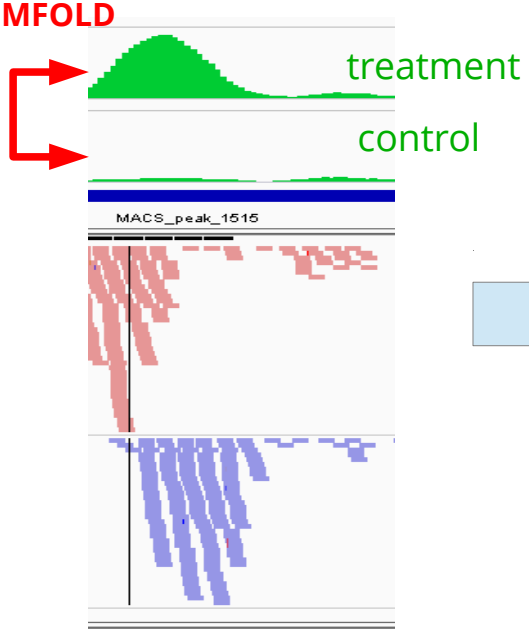
Pval ~ 0.6

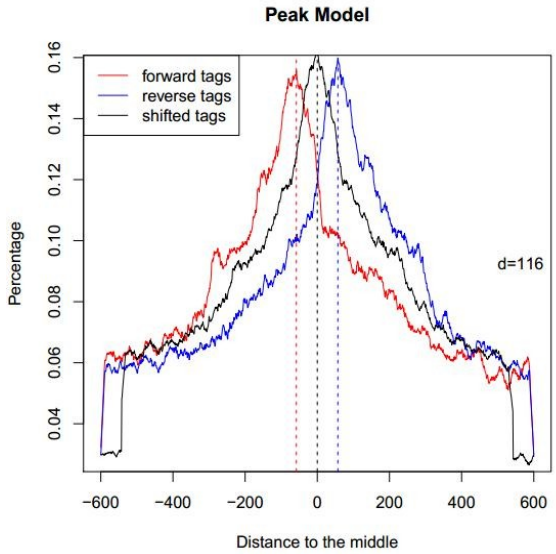


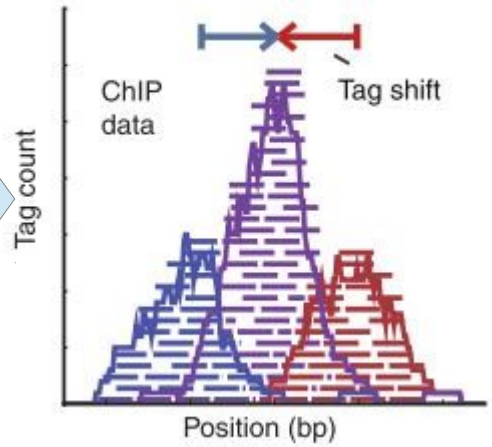
6. MACS [Zhang et al. Genome Biol. 2008]

- **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

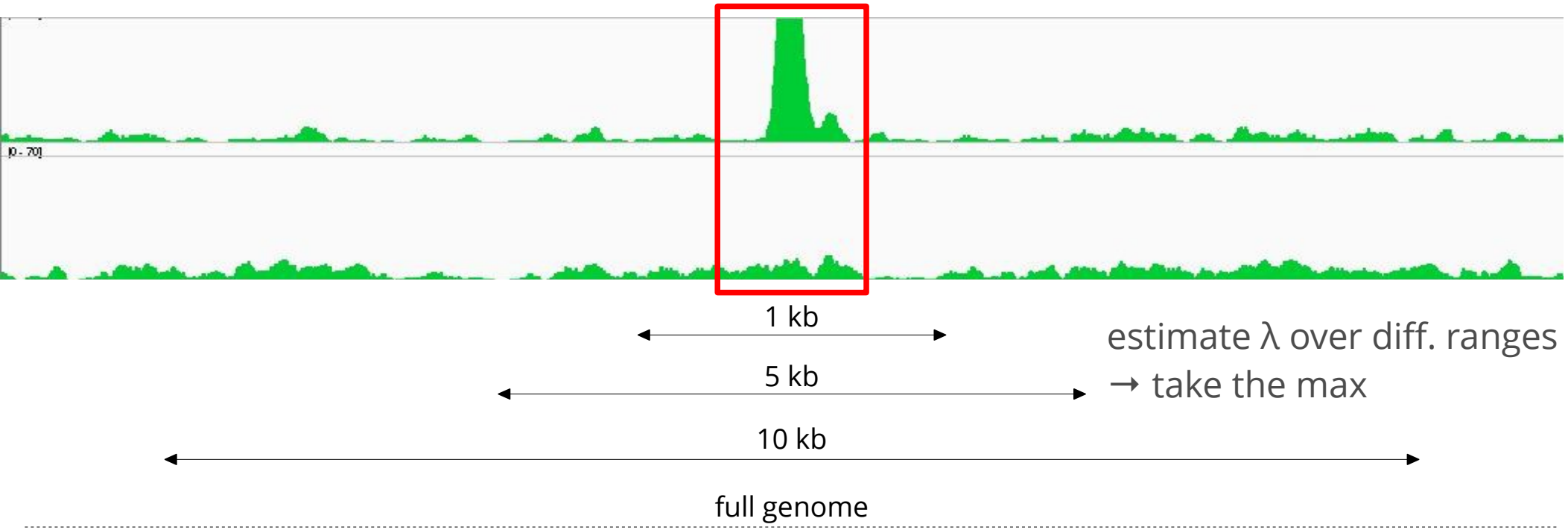






5. MACS [Zhang et al. Genome Biol. 2008]

- **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



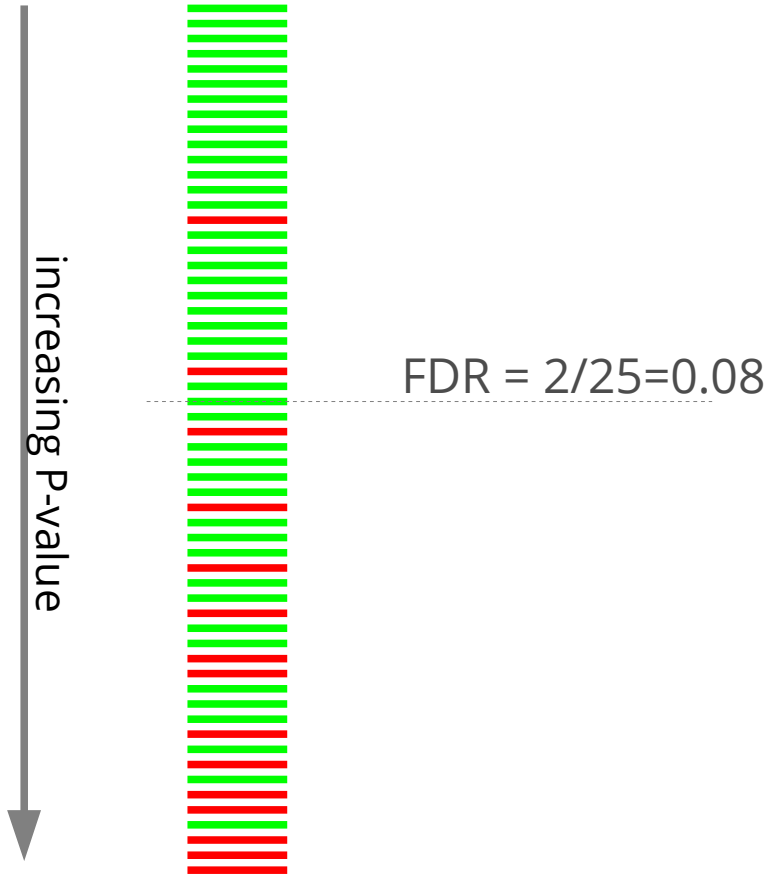
5. MACS [Zhang et al. Genome Biol. 2008]

- **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]

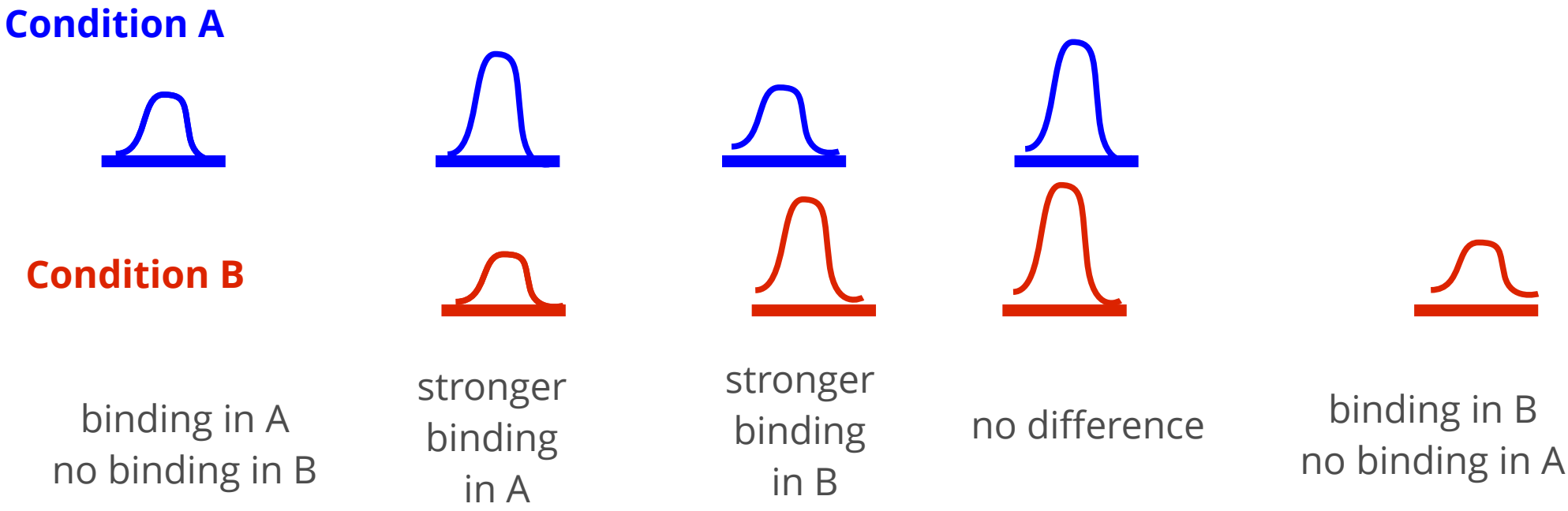
- **Step 4 : estimating FDR**
  - positive peaks (P-values)
  - swap treatment and input; call negative peaks (P-value)

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



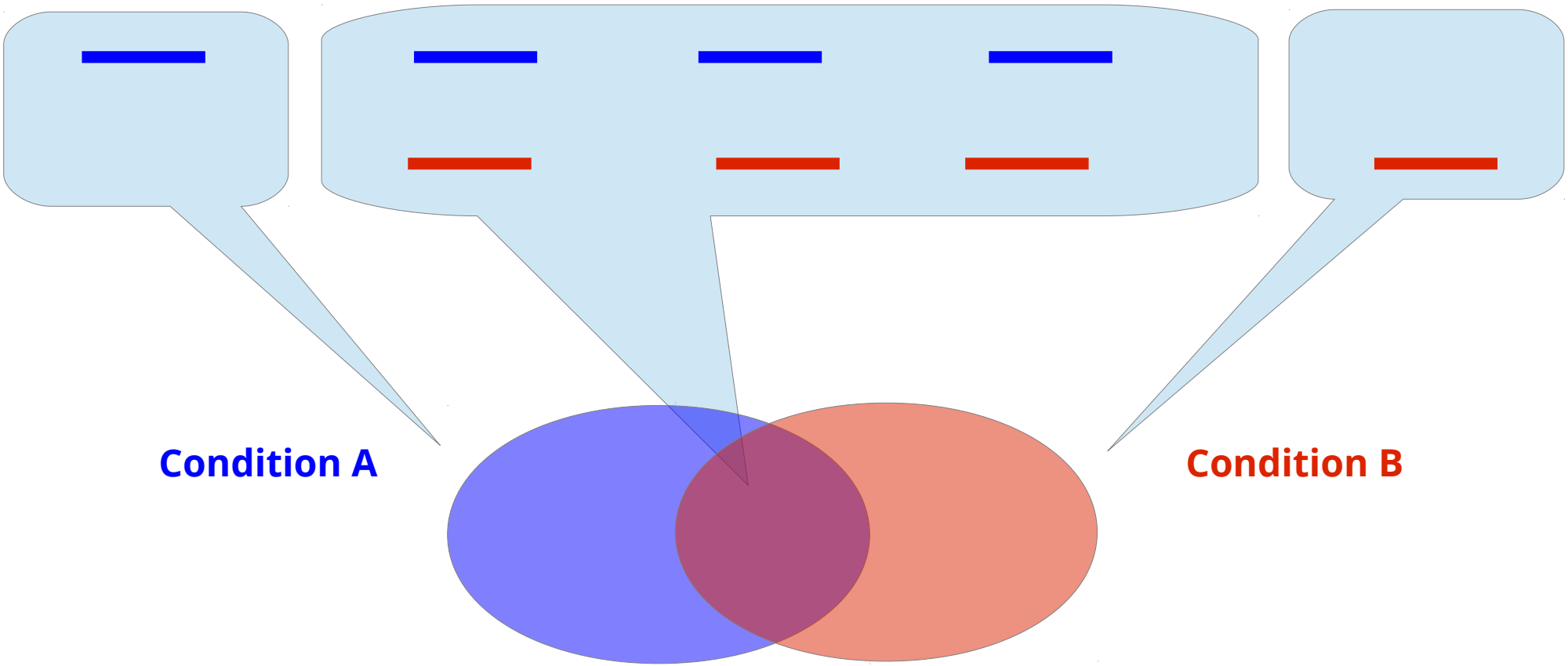
6. differential analysis

- 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



6. differential analysis

- simple approach → compute common and specific peaks

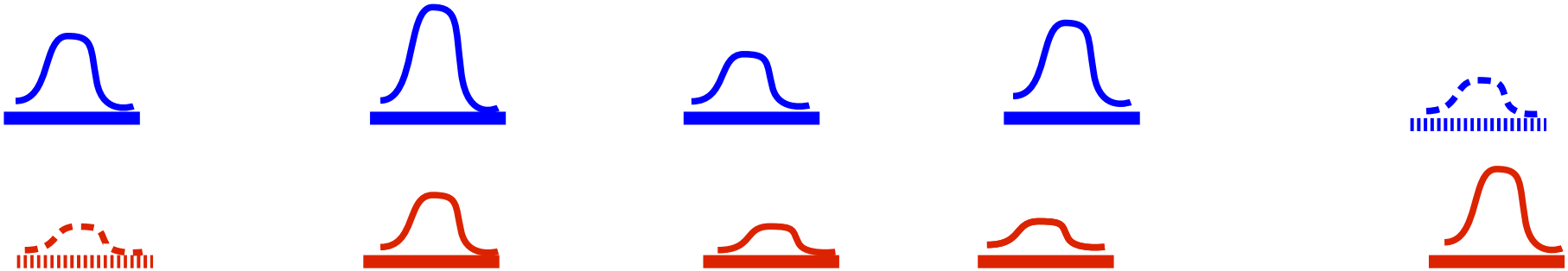


Drawback :

- 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**

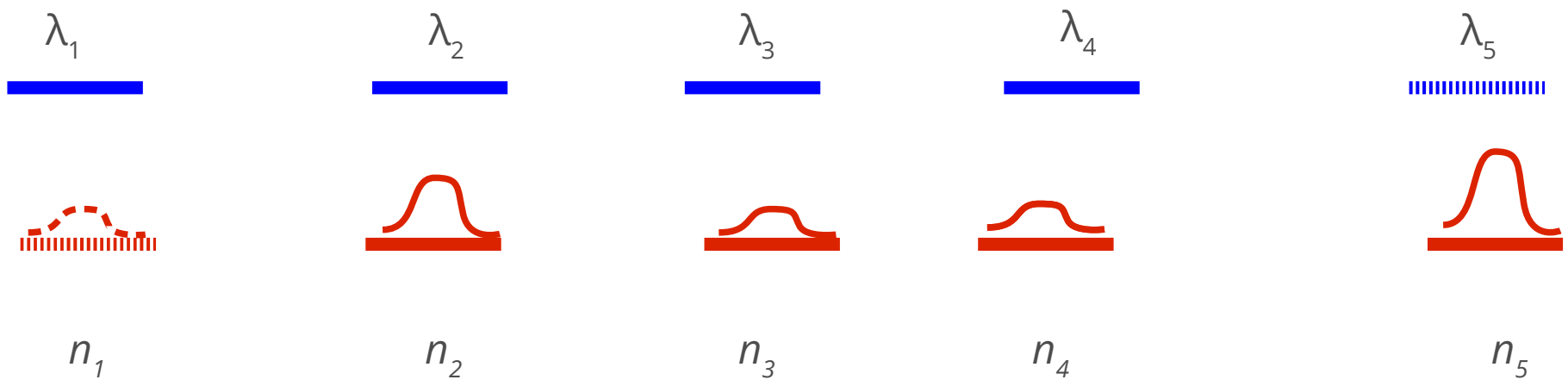


- statistical model
  - **without replicates** : assume simple Poisson model (→ SICER-df)
  - **with replicates** : perform differential test using DE tools from RNA-seq (diffBind using EdgeR, DESeq,...) based on read counts

6. differential analysis

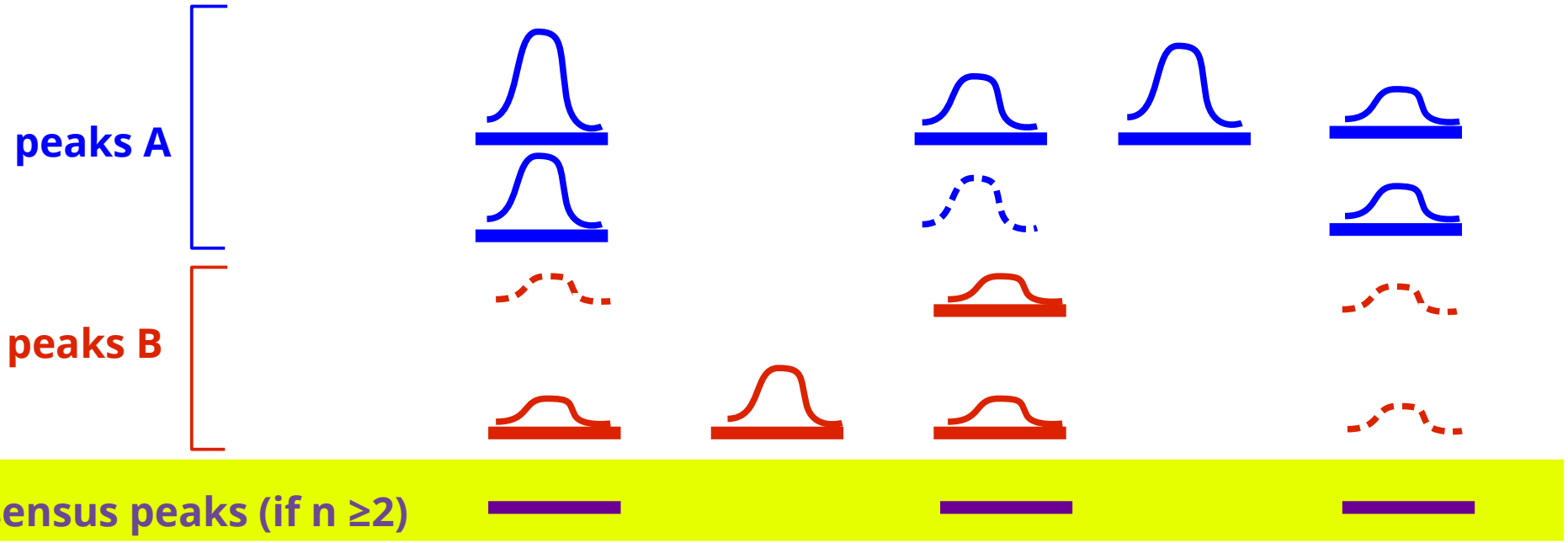
- **without replicates (sicer-df)**
  - consider one condition to be the reference (condition A)
  - call peaks on each condition independently
  - take union of peaks
  - assume Poisson model based on expected number of reads in region
  - compute P-value, log(fold-change)

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

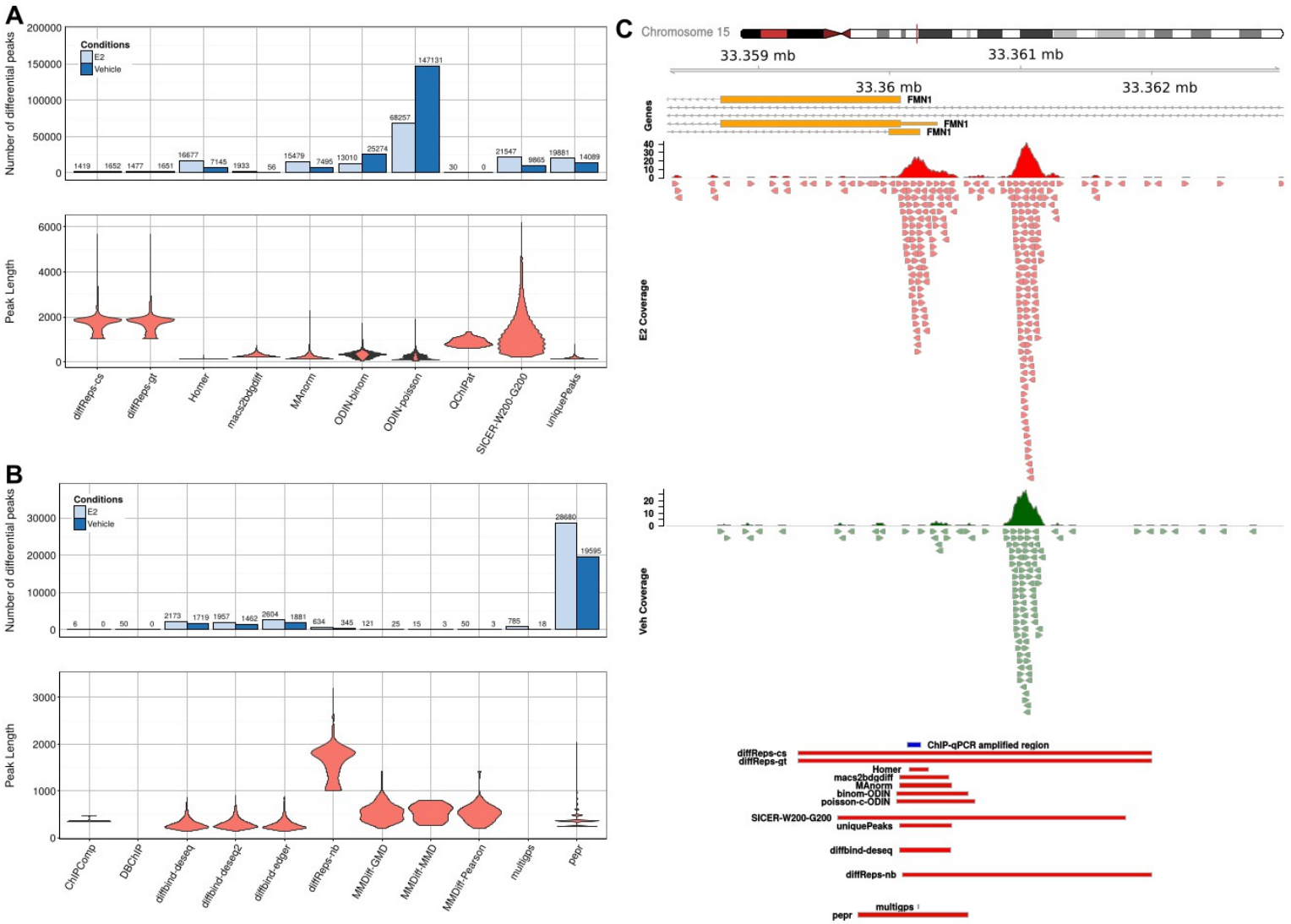


6. differential analysis

- **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)



6. differential analysis



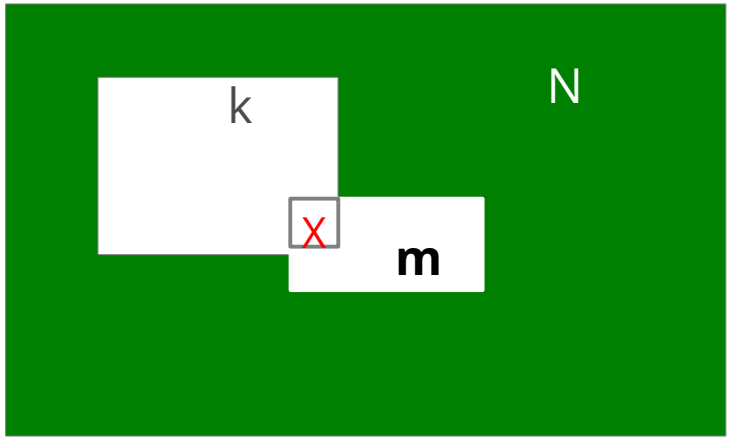
Considerable differences in peak numbers and sizes !

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 ?

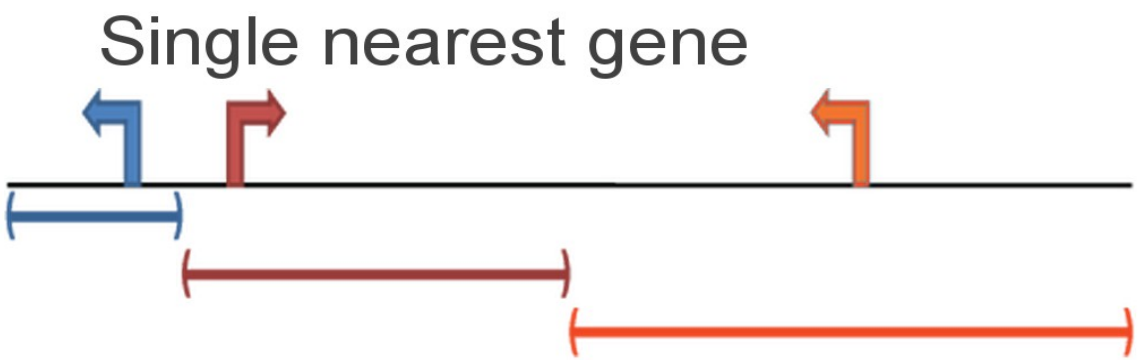


Nearest gene : problem

- 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

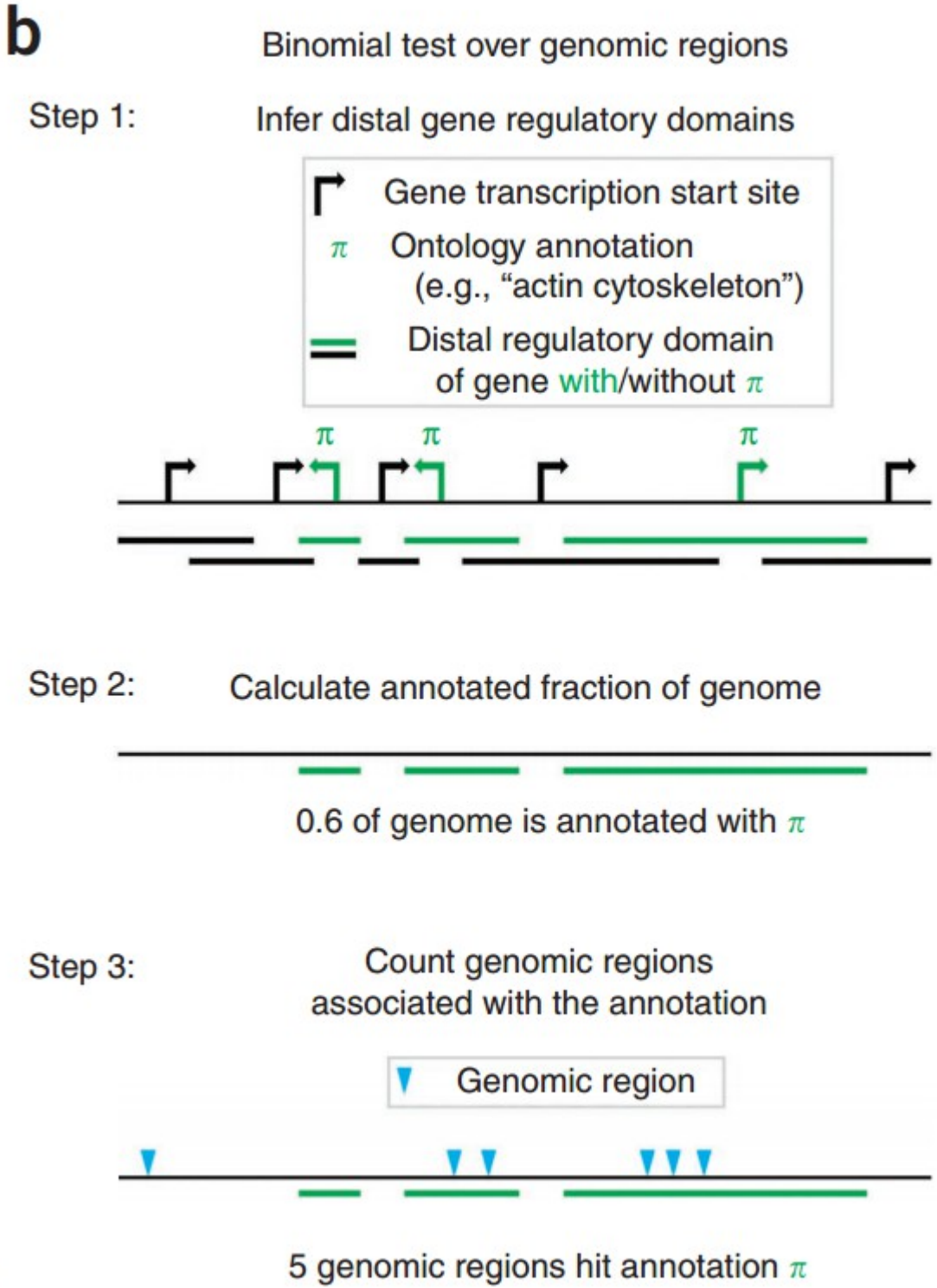
GREAT

- 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

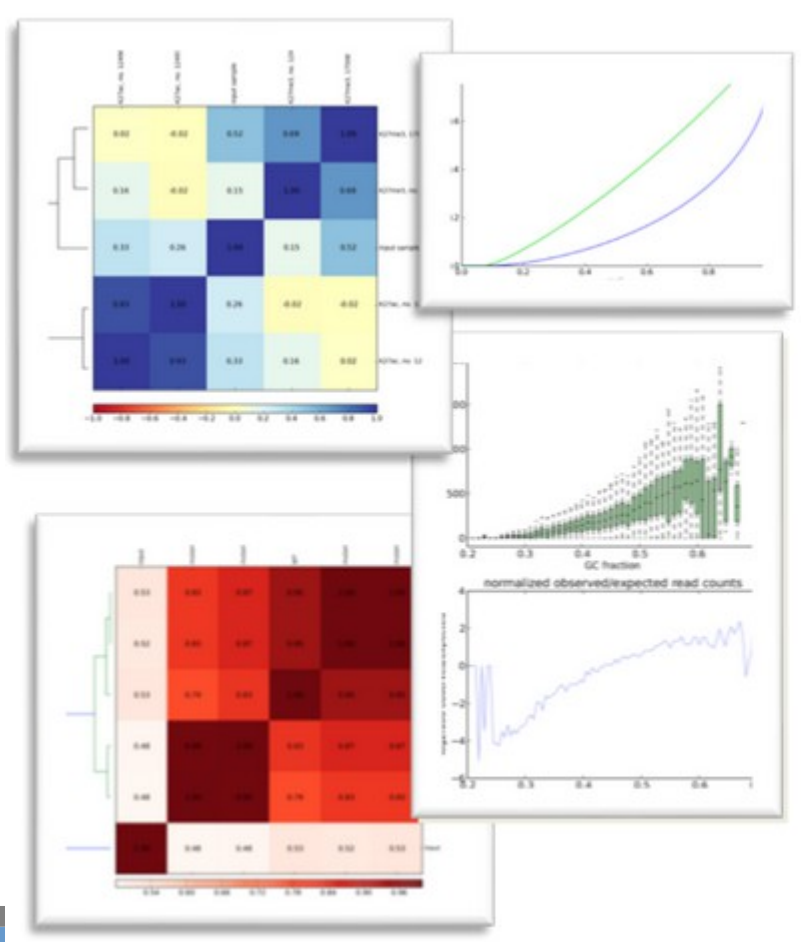
$p_\pi = 0.6$ fraction of genome annotated with π

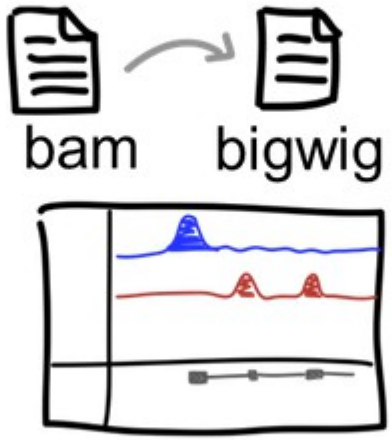
$k_\pi = 5$ genomic regions hit annotation π

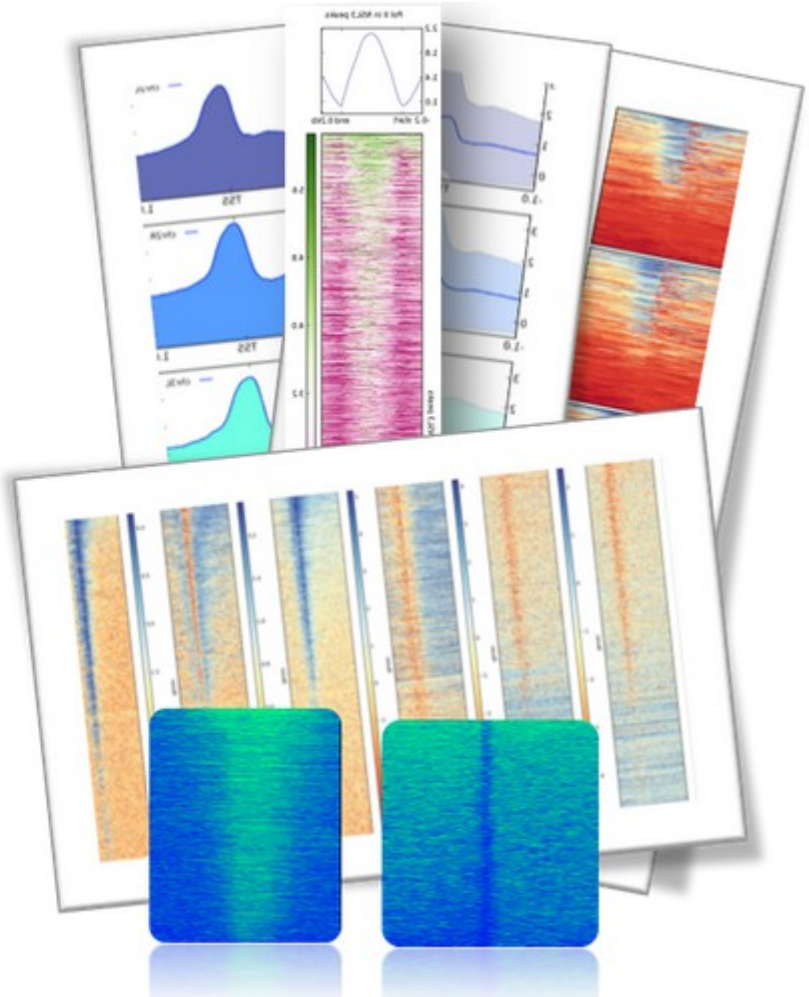
$$P = \Pr_{\text{binom}}(k \geq 5 \mid n = 6, p = 0.6)$$

DeepTools

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







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