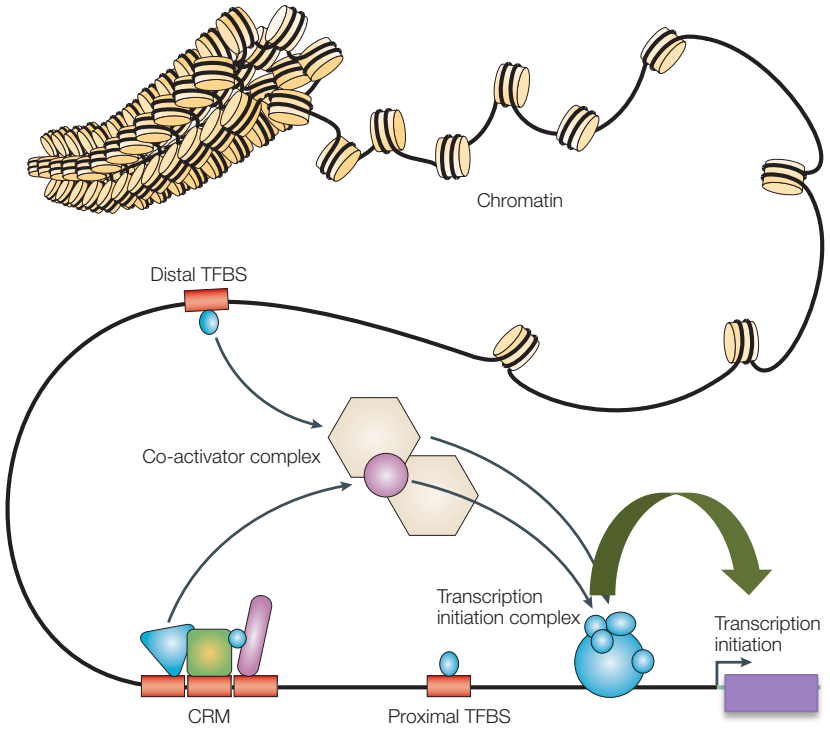
ChIP-seq data analysis

Morgane Thomas-Chollier
mthomas@biologie.ens.fr

Computational Systems Biology
Institut de Biologie de l'Ecole Normale Supérieure, ***Paris, France***

UAEM – Cuernavaca, Mexico – 12th March 2015

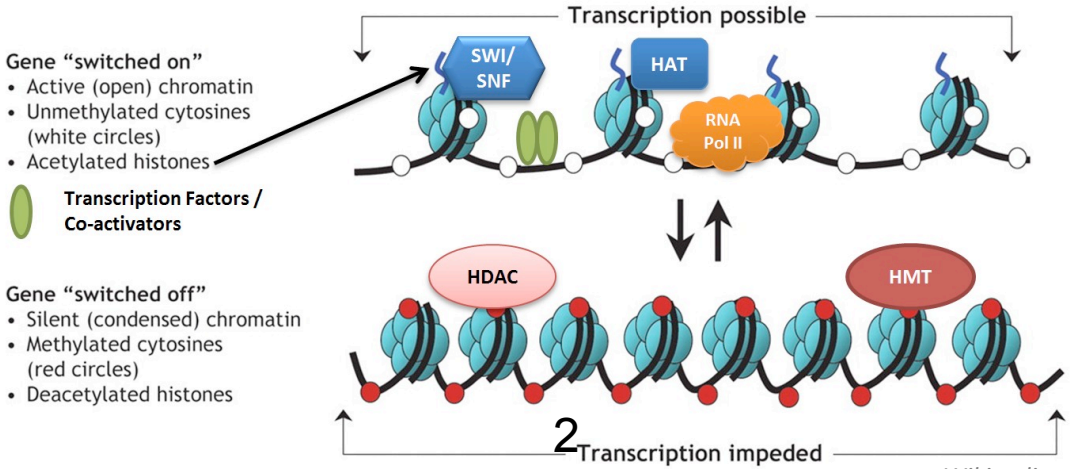
Biological concepts of transcriptional regulation



Wasserman et al, Nat Rev Genet, 2004

Transcription factors are proteins that modulate (activate/repress) the expression of **target genes** through the binding on **DNA cis-regulatory elements**

Chromatin accessibility (open/close) and **histone modifications** (eg: acetylation) also regulate gene expression



Wikipedia

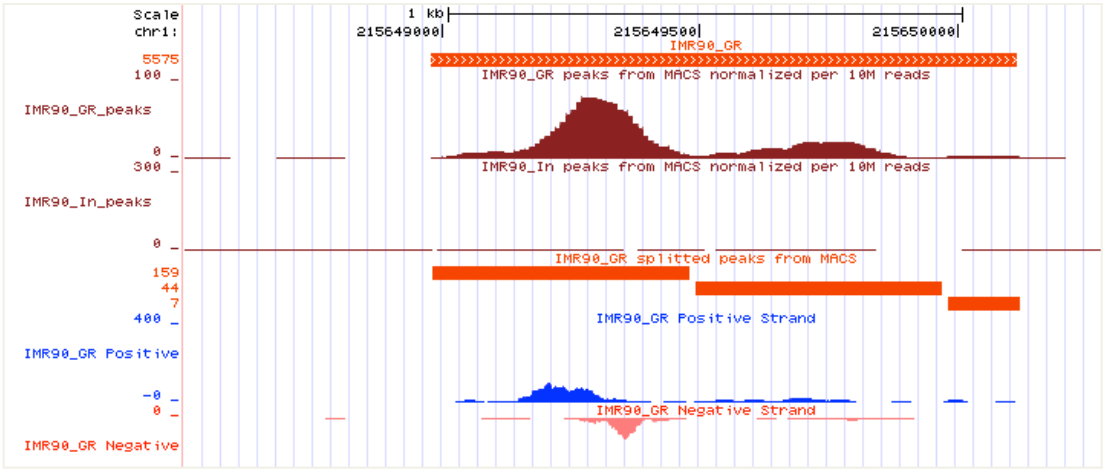
in vivo experimental methods to identify binding sites

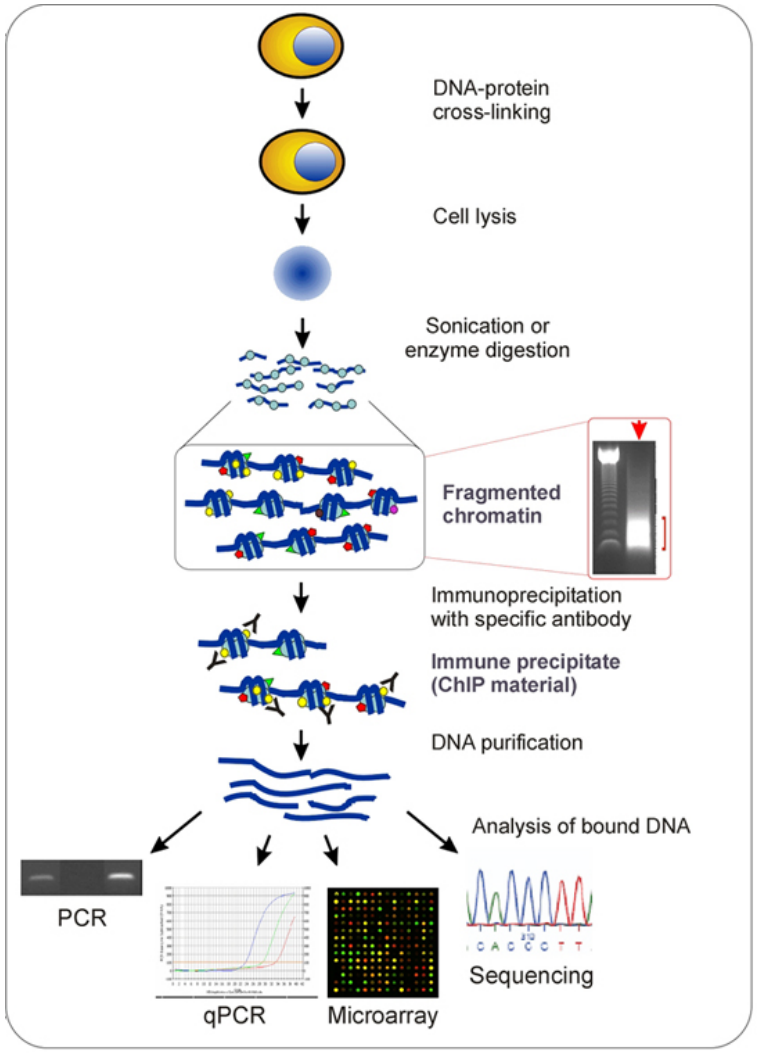
ChIP (=Chromatin Immuno-Precipitation)

differences in **methods to detect** the **bound DNA**

-<u>small-scale</u>: PCR / qPCR

- <u>large-scale</u>:
  - microarray = **ChIP-on-chip**
  - sequencing = **ChIP-seq**





http://www.chip-antibodies.com/

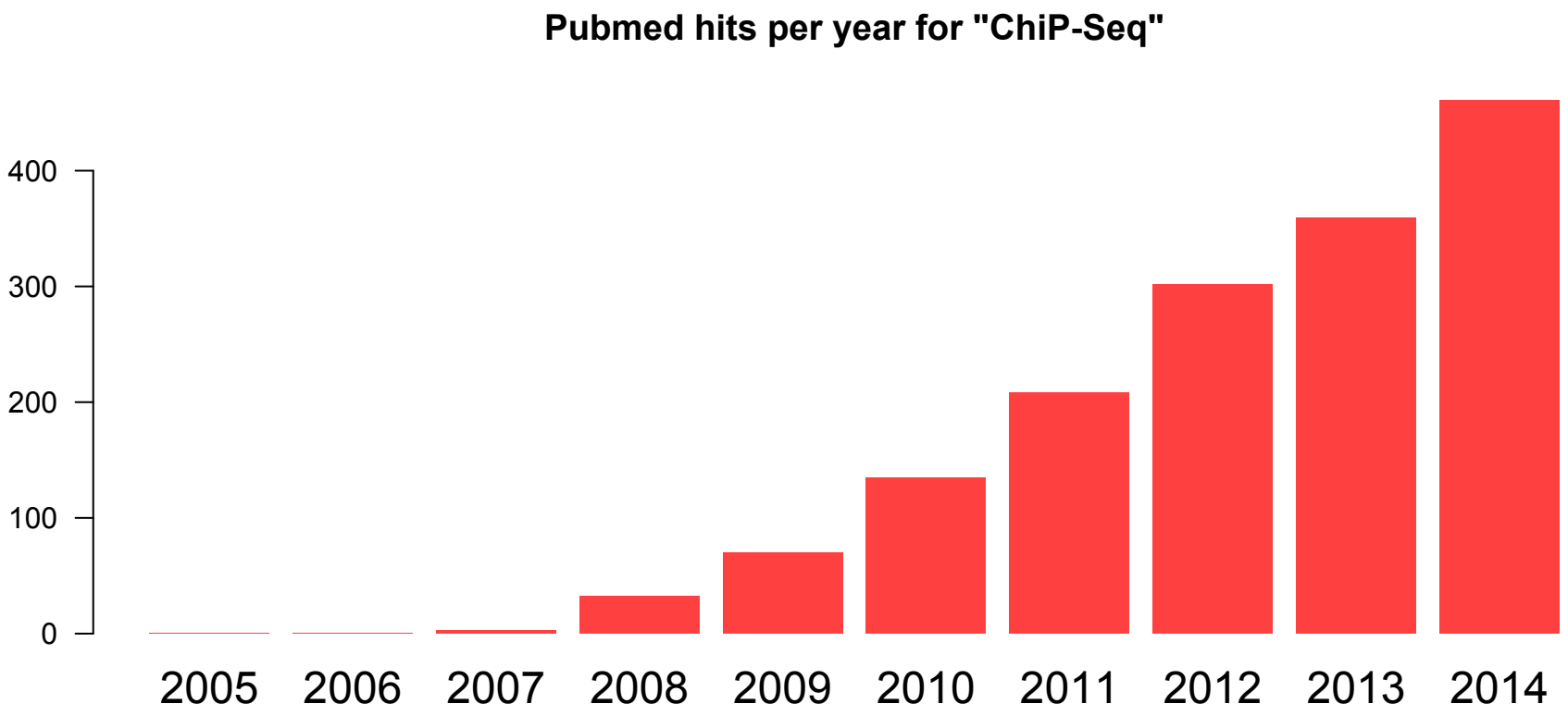
High-troughput sequencing (HTS)
Next Generation Sequencong (NGS)

- Type of data: **millions** of sequences (« reads ») short (50bp-250bp currently, depending on the sequencing machine)
- Reads contain **errors**
- Reduced sequencing time of about **3 days**

Table 1: Performance Parameters of the HiSeq 3000/4000 Systems.[a]

	HiSeq 3000 System	HiSeq 4000 System
Number of Flow Cells per Run	1	1 or 2
Output[b] 2 × 150 bp 2 × 75 bp 1 × 50 bp	 630–750 Gb 315–375 Gb 105–125 Gb	 1300–1500 Gb 650–750 Gb 215–250 Gb
Clusters Passing Filter (Single Reads)	2.1–2.5 billion	4.3–5 billion
Quality Scores	≥ 75% of bases above Q30 at 2 × 150 bp	≥ 75% of bases above Q30 at 2 × 150 bp
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run[c]	up to 6	up to 12
Exomes per Run[d]	up to 90	up to 180
Transcriptomes per Run[e]	up to 50	up to 100

ChIP-seq is a recently but widely adopted technique !

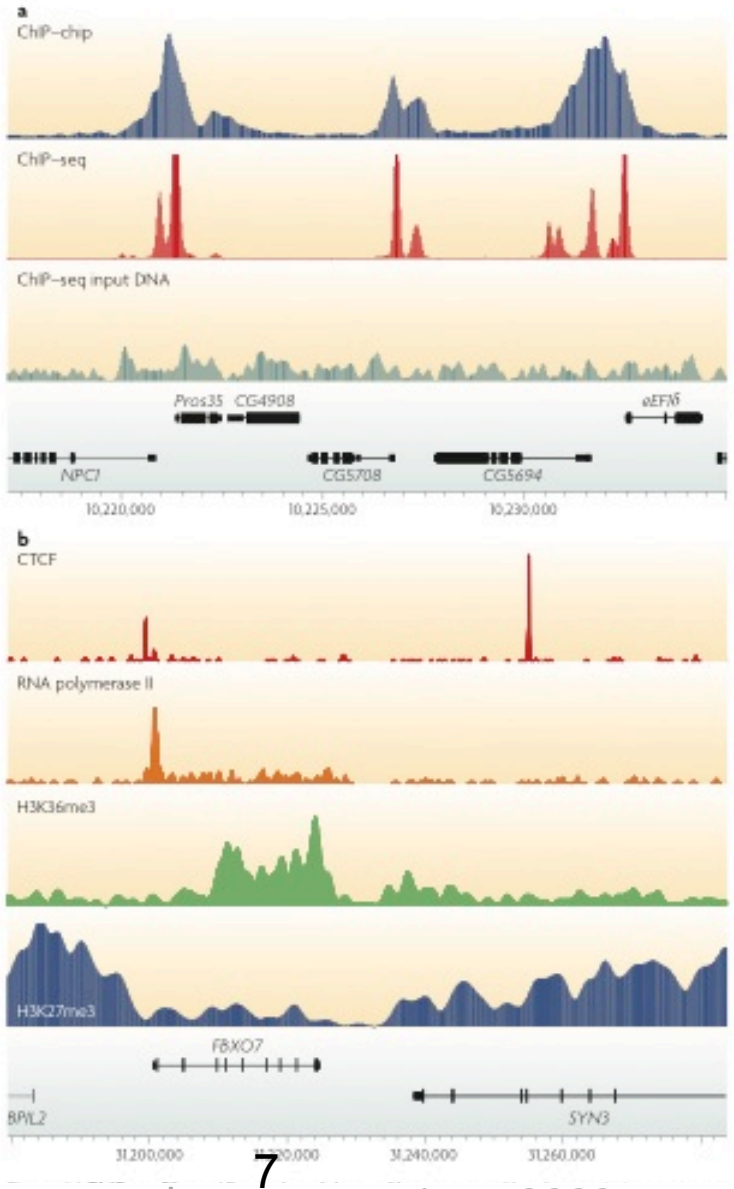


ChIP-seq applications

- find ***all*** regions in the genome bound by
  - a specific **transcription factor**
  - **histones** bearing a specific **modification**

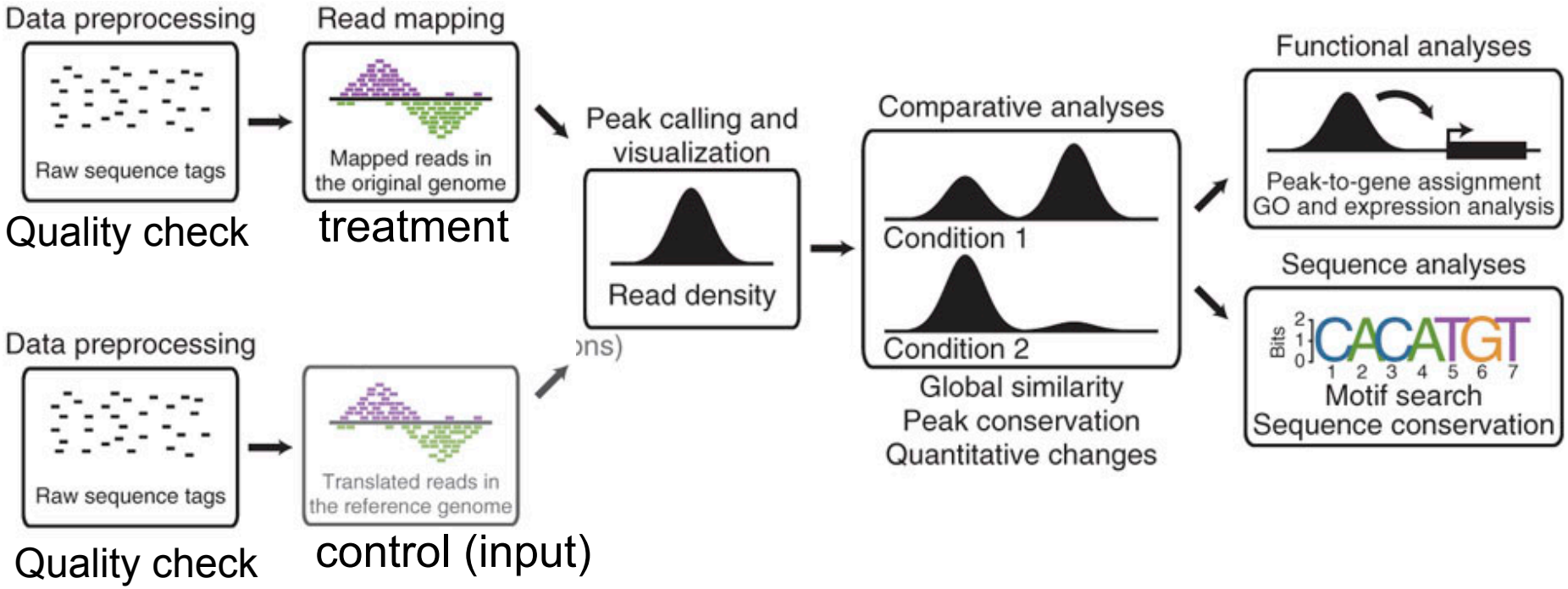
- in a given ***experimental condition*** (cell type, developmental stage,...)

The obtained ChIP-seq **profiles** have **different shapes**, depending on the targeted protein



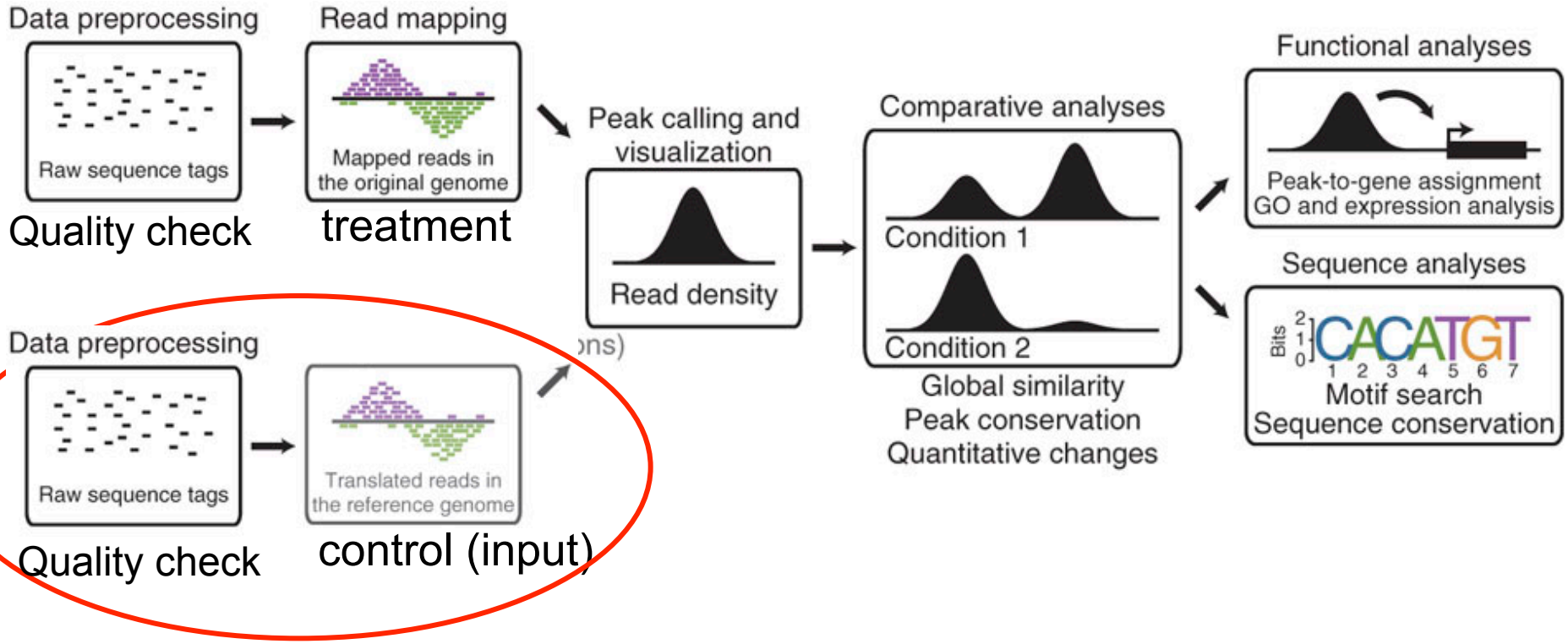
Park, Nature reviews 2009

ChIP-seq analysis workflow



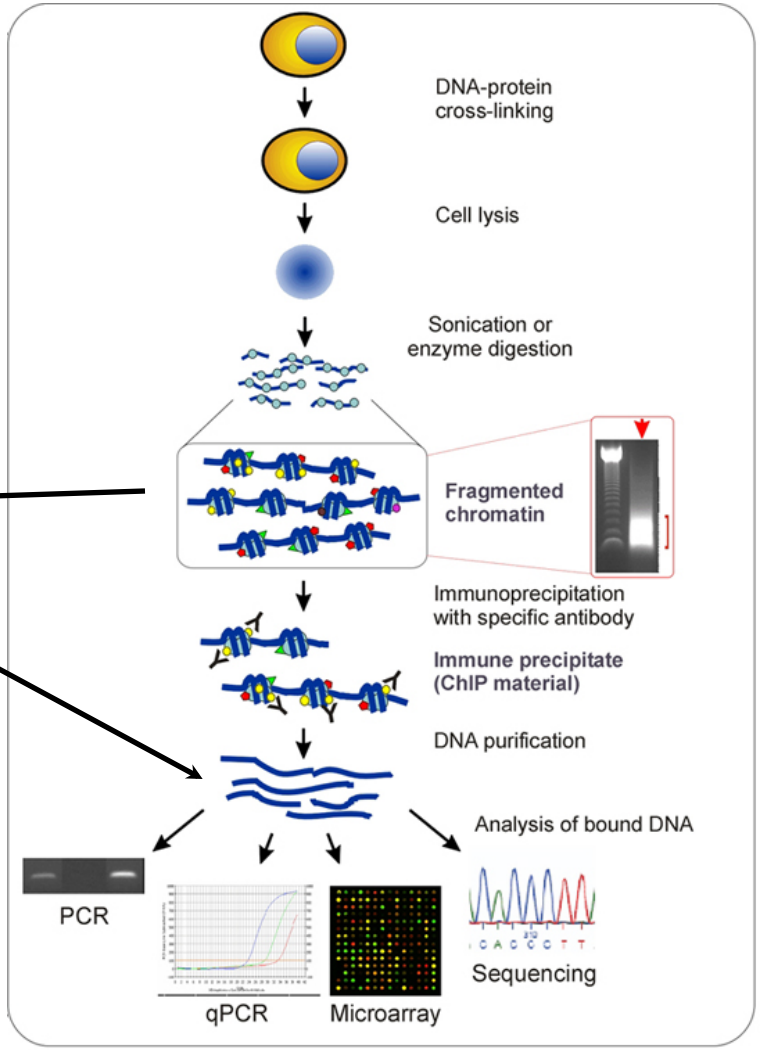
Adapted from Bardet et al, Nature Protocols, 2012

ChIP-seq analysis workflow

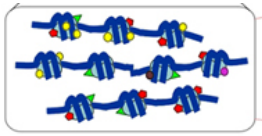


Adapted from Bardet et al, Nature Protocols, 2012

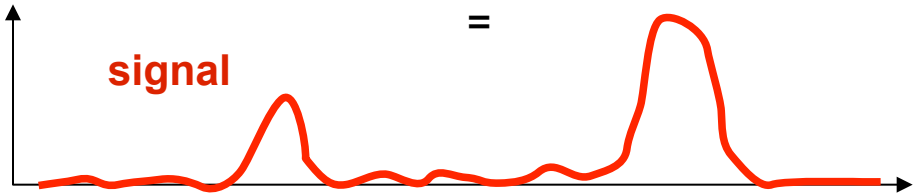
From sequence reads to peaks



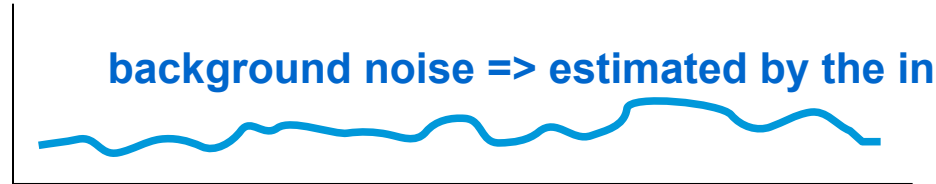
Input
(control without IP)



=

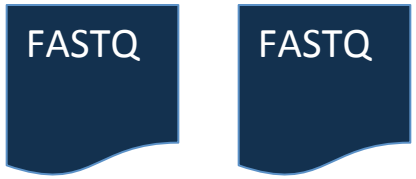


+

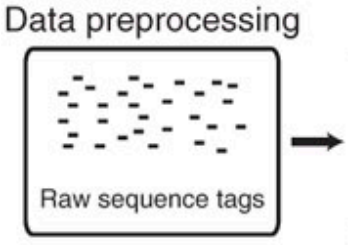


From sequence reads to peaks

experiment Input



sequences (reads length 36 / 50 bp, single-end)
from Illumina

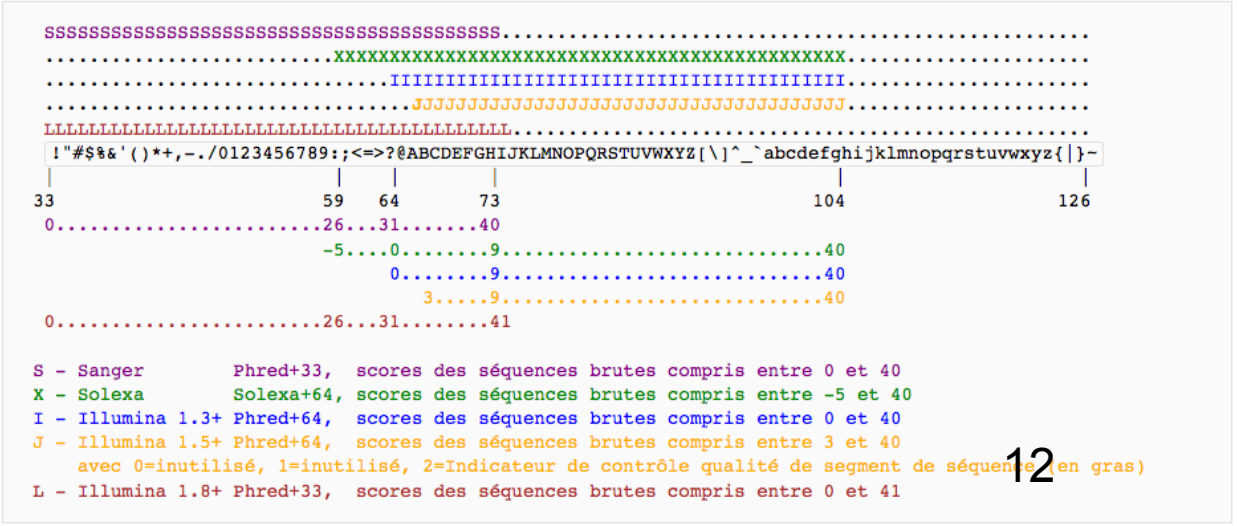


FASTQ format

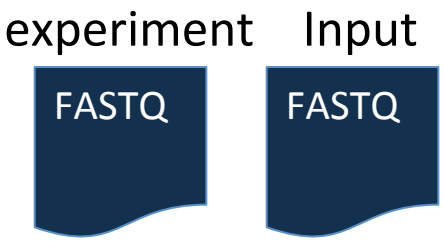
```
@SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
+
IIIIG1?II;IIIII1IIII1%.I7I
@SRR002012.2 Oct4:5:1:804:348
GTCTGCATTATCTACCAGCACTTCCC
+
IIIIIIIII'I2IIIII:)I2II3I0
@SRR002012.3 Oct4:5:1:767:334
GCTGTCTTCCCGCTGTTTTATCCCCC
+
III8IIIIIII3III6II%II*III3
@SRR002012.4 Oct4:5:1:805:329
GTAGTTTACCTGTTCATATGTTTCTG
+
IIIIIII9IIIIII?IIIIIIII7II
```

```
>SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
>SRR002012.2 Oct4:5:1:804:348
GTCTGCATTATCTACCAGCACTTCCC
>SRR002012.3 Oct4:5:1:767:334
GCTGTCTTCCCGCTGTTTTATCCCCC
>SRR002012.4 Oct4:5:1:805:329
GTAGTTTACCTGTTCATATGTTTCTG
```

Source: Wikipedia



From sequence reads to peaks



sequences (reads length 36 / 50 bp, single-end)
from Illumina

FASTQC → **quality check**

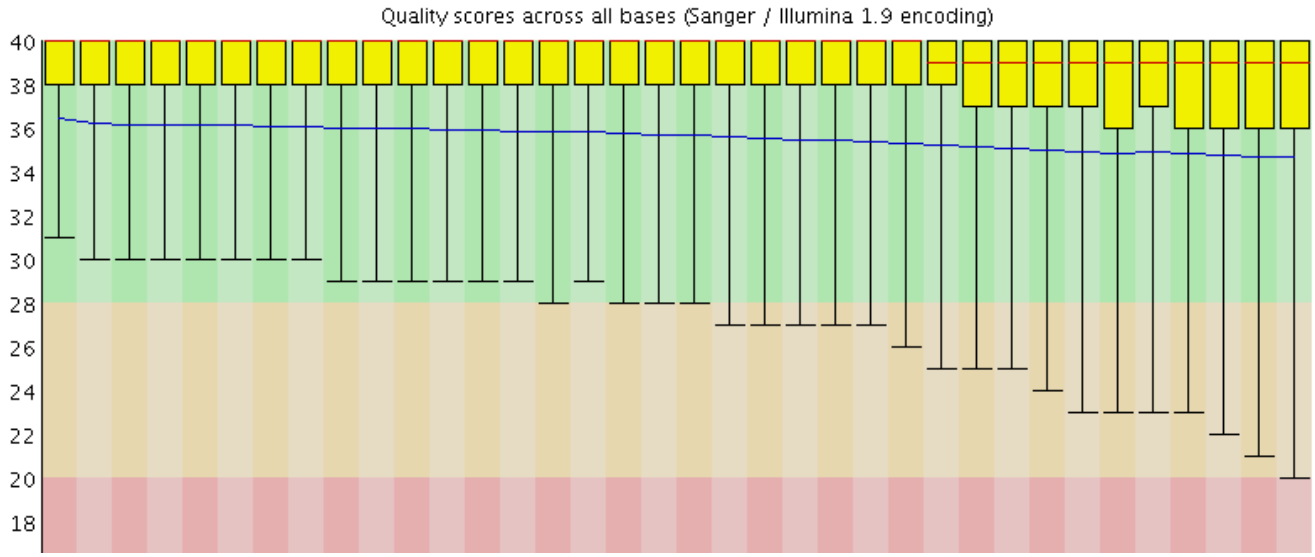
FastQC Report

Fri 1 Jul 201
IMR90_GR.

Summary

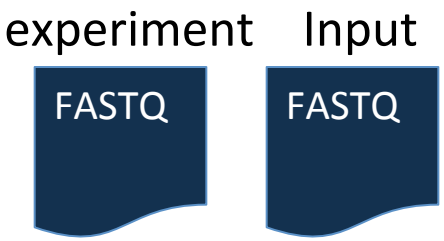
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Per base sequence quality



http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

From sequence reads to peaks



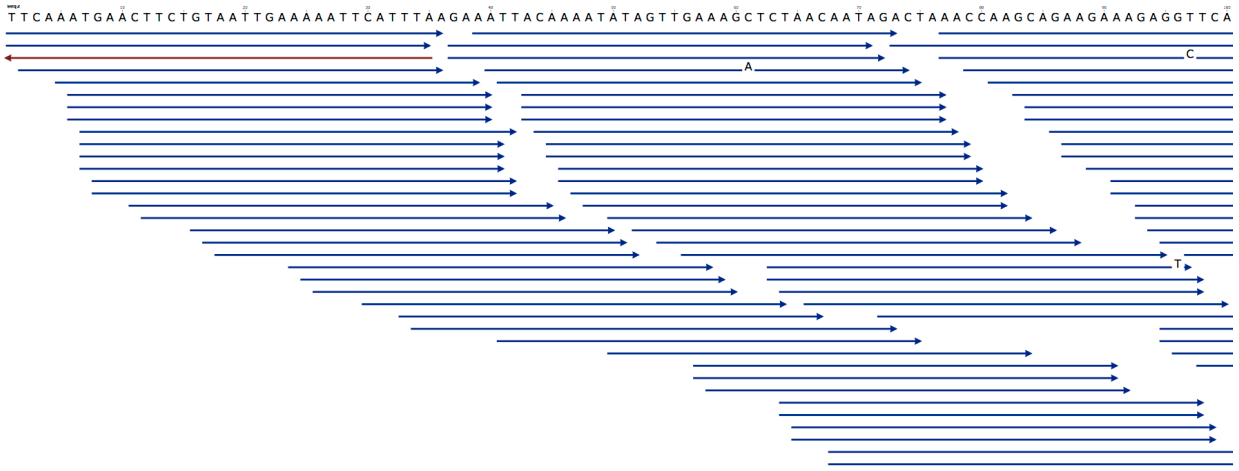
sequences (reads length 36 / 50 bp, single-end)
from Illumina

FASTQC → **quality check**

From sequence reads to peaks

experiment Input

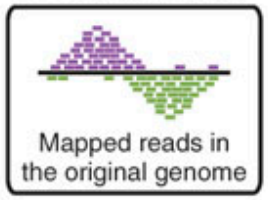
FASTQ FASTQ



FASTQC

Source: http://trac.seqan.de

Read mapping



BAM
SAM

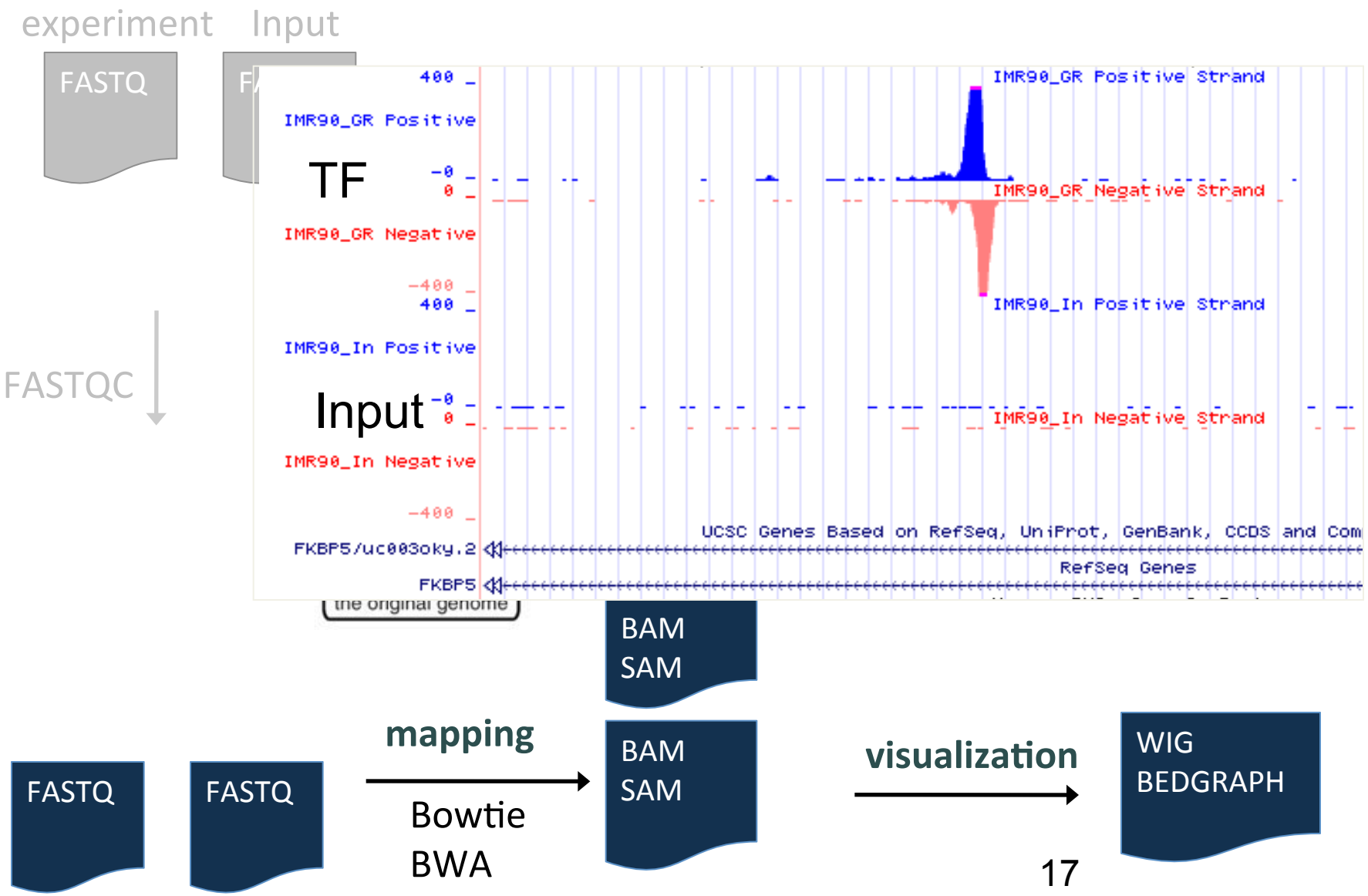
mapping

FASTQ FASTQ

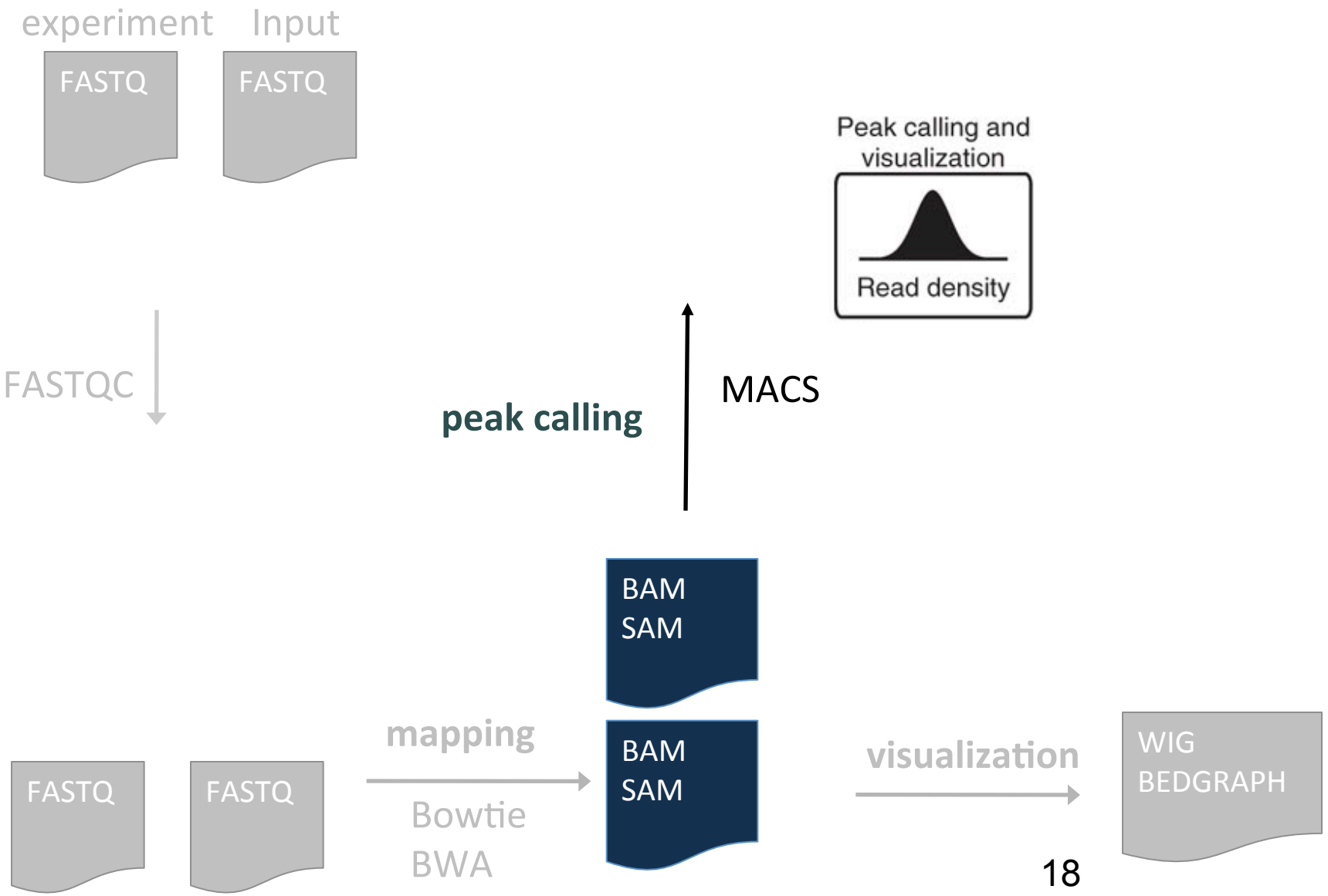
Bowtie
BWA

BAM
SAM

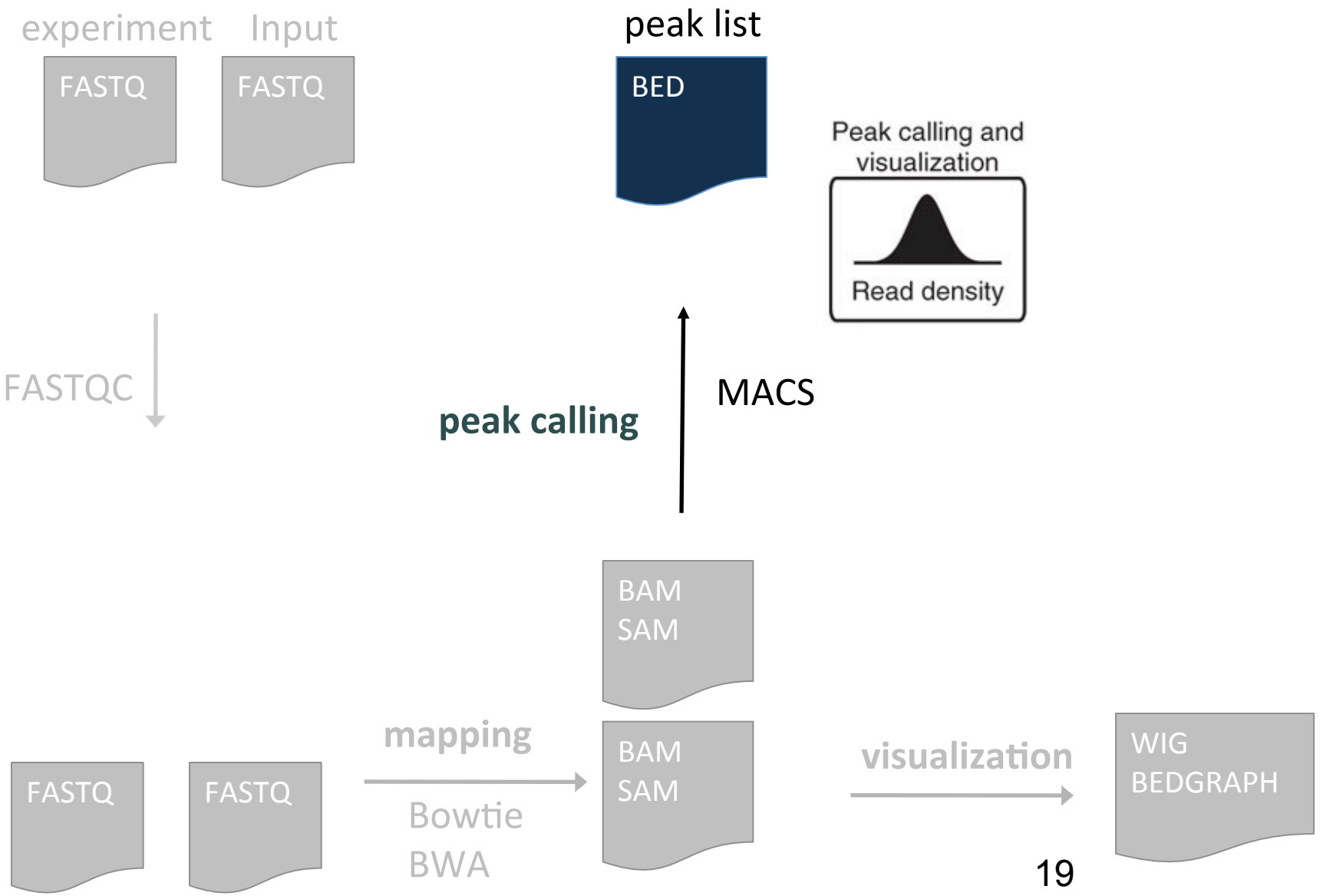
From sequence reads to peaks



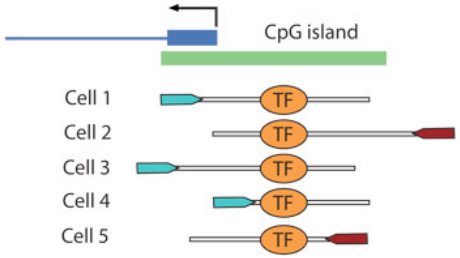
From sequence reads to peaks



From sequence reads to peaks



A

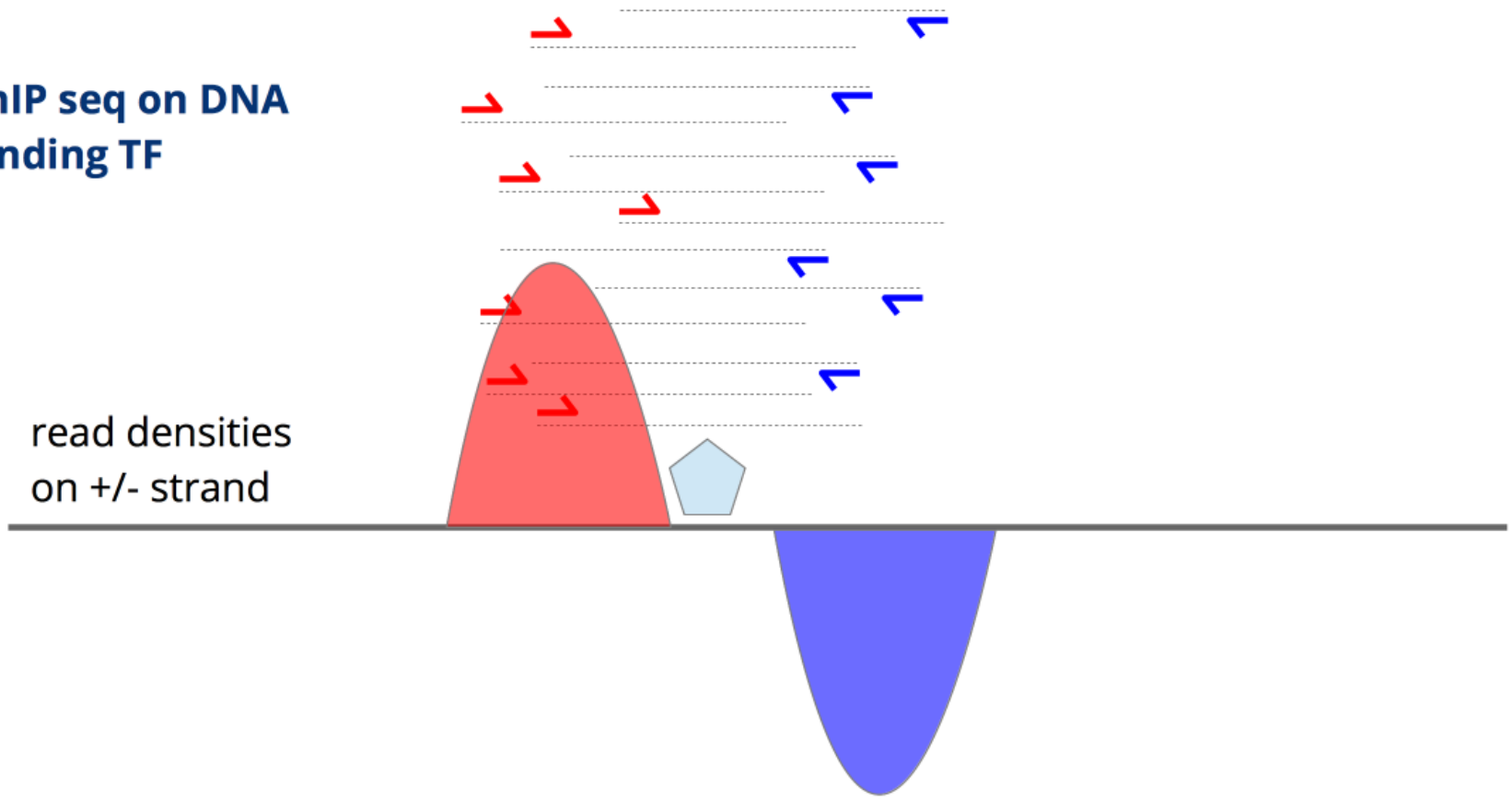


The read « peaks » are not the location of the binding site !

mapping

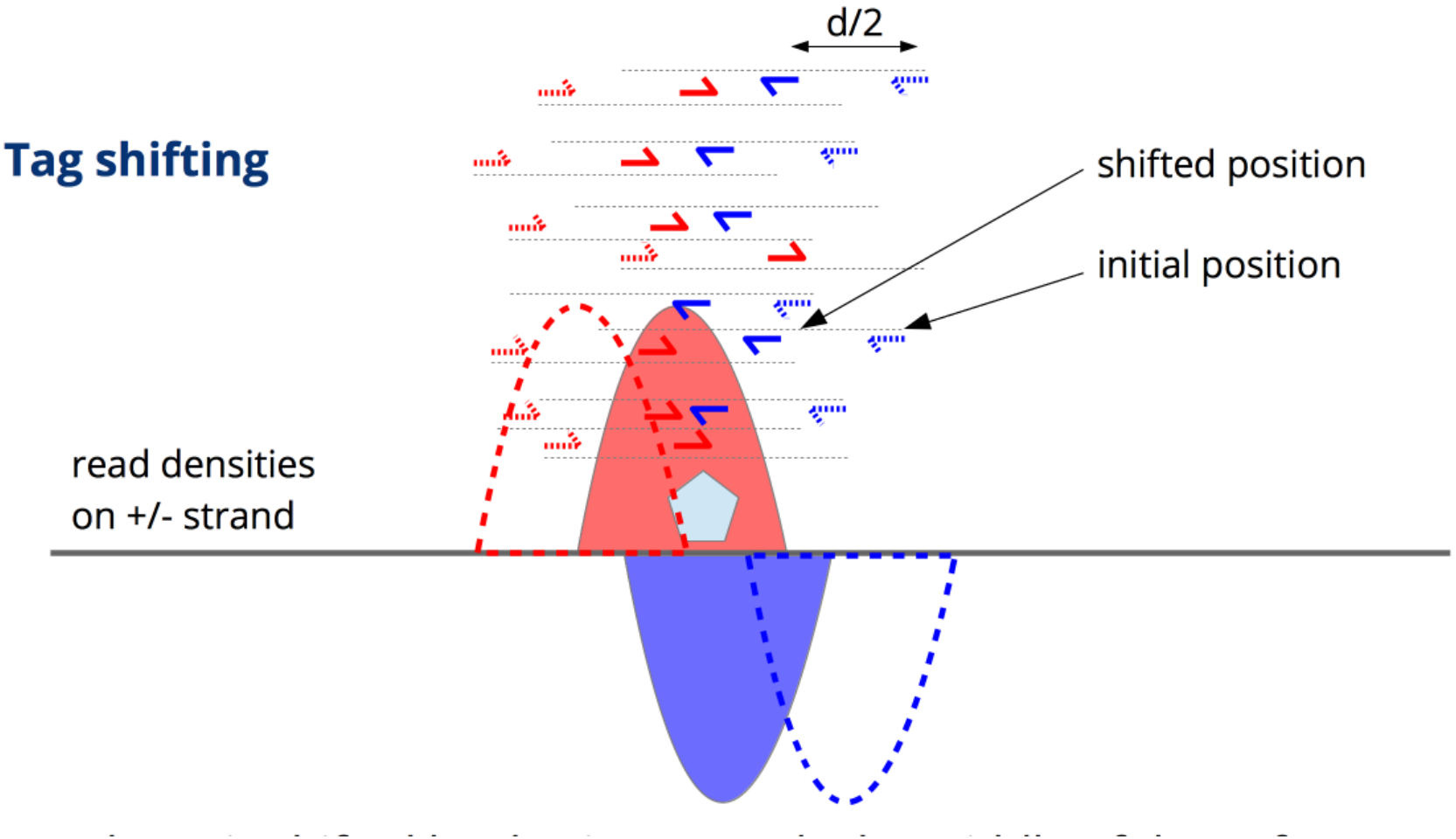
Valouev Nat Methods (2008), Jothi, NAR (2008)

How to determine the position of the TF ?



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

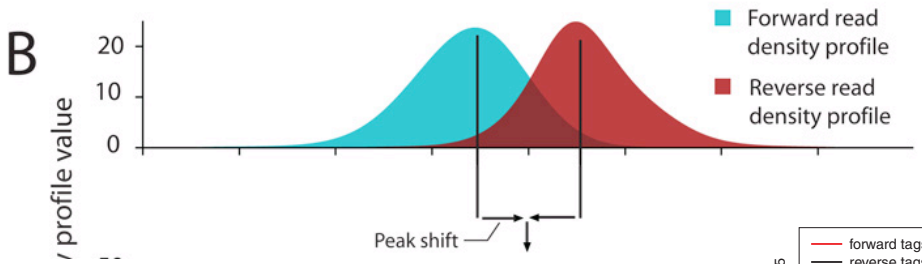
From aligned reads to binding sites

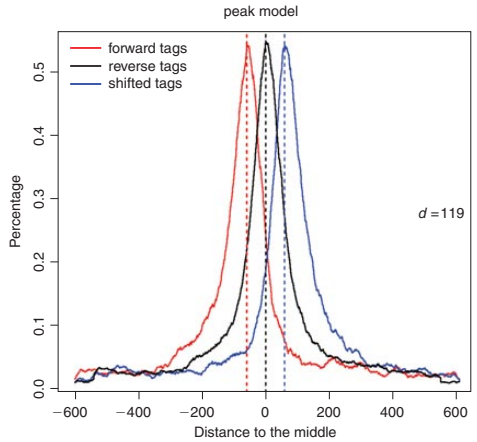


Each tag is shifted by *d/2* (i.e. towards the middle of the IP fragment) where *d* represent the fragment length

From aligned reads to binding sites

bimodal enrichment pattern





MACS: Two steps strategy :

1 – modelling the read shift size

2 – peak calling

1 : search high-quality paired peaks : separates their forward and reverse reads, and aligns them by the midpoint. The distance between the modes of the forward and reverse peaks in the alignment is defined as *d*, and MACS shifts all reads by *d*/2 toward the 3′ ends to better locate the precise binding sites.

2: uses the shift size to search for peaks, Poisson distribution to measure the p-value of each peak, and False Discovery Rate (FDR) calculation using the input data

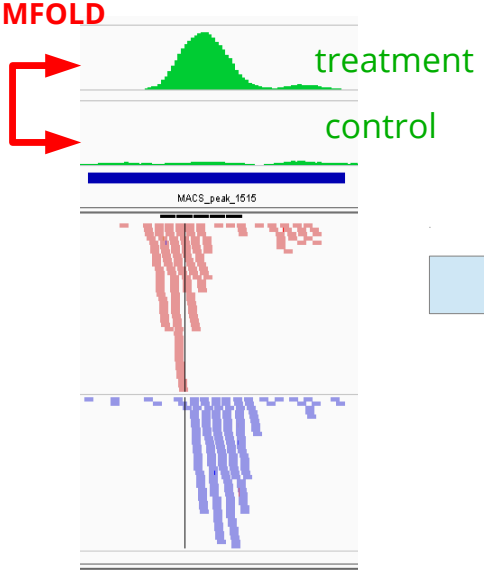
Feng, J., Liu, T., & Zhang, Y. (2011). Using MACS to Identify Peaks from ChIP-Seq Data, Current Protocols in Bioinformatics

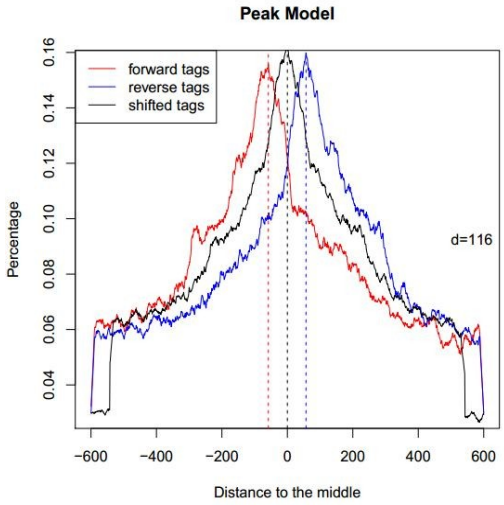
MACS

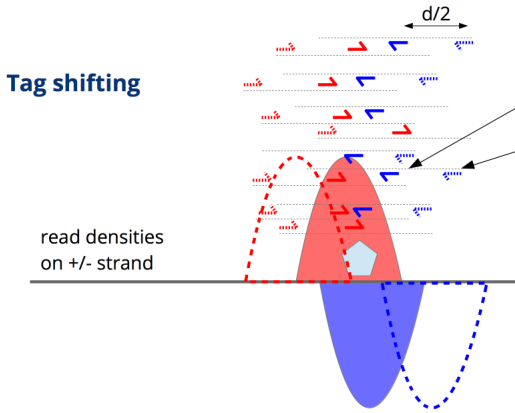
[Zhang et al. Genome Biol. 2008]

- **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



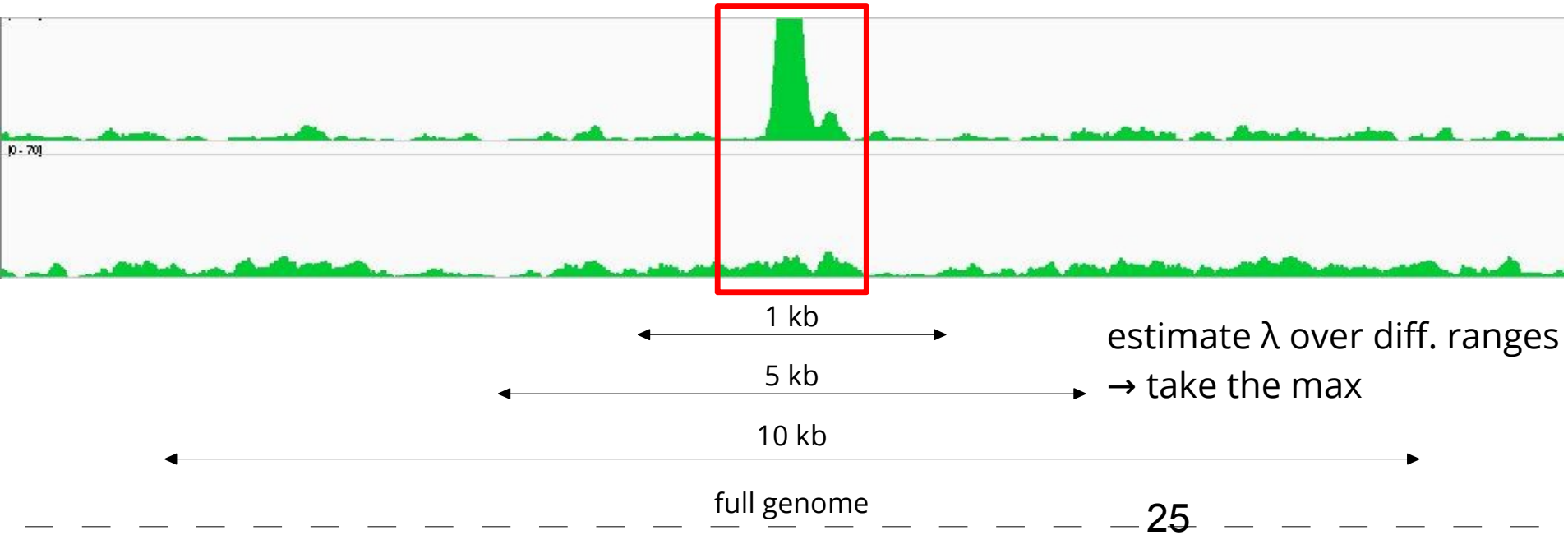




MACS

[Zhang et al. Genome Biol. 2008]

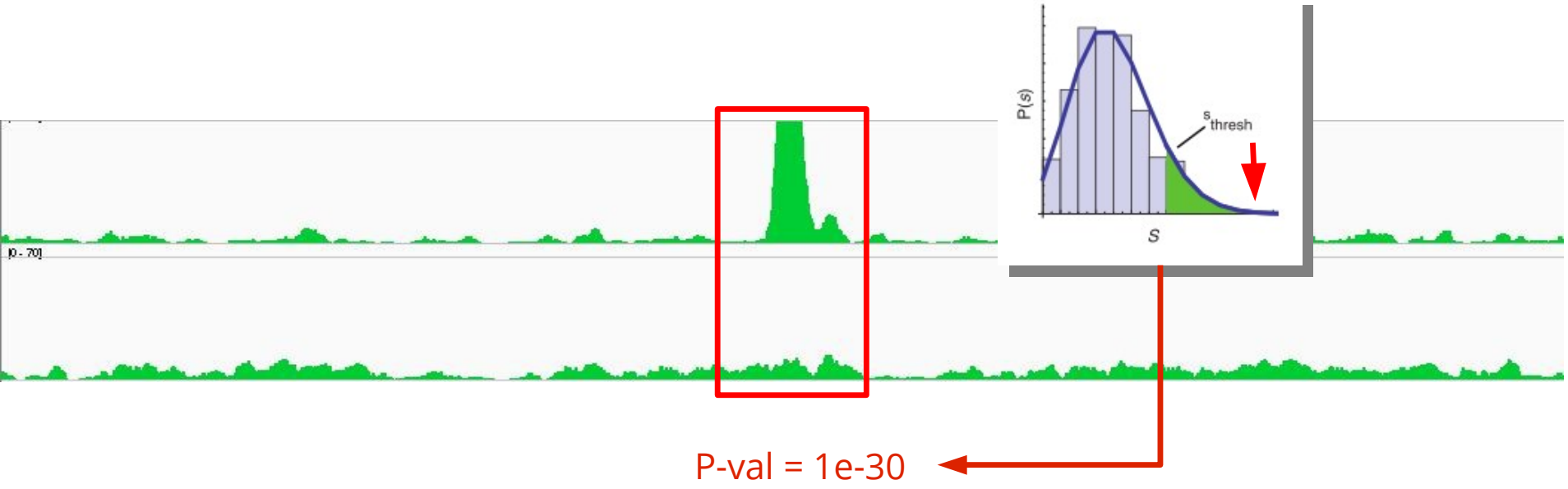
- **Step 2 : identification of local noise parameter**
  - slide a window of size *2*d* across treatment and input
  - estimate parameter λ_{local} of Poisson distribution



MACS

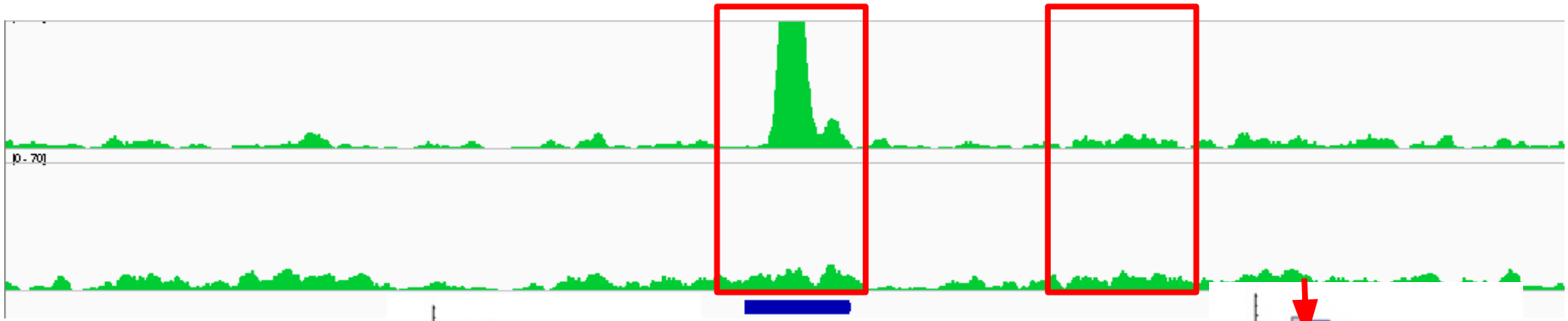
[Zhang et al. Genome Biol. 2008]

- **Step 3 : identification of enriched/peak regions**
  - determine regions with P-values < PVALUE
  - determine summit position inside enriched regions as max density

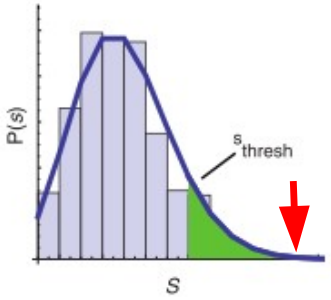


Peak-calling in summary

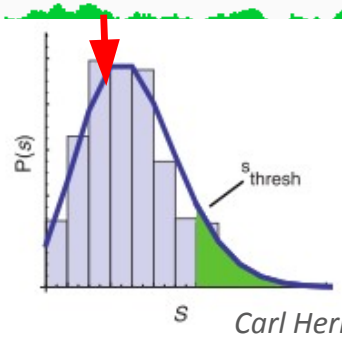
- **Determining "enriched" regions**
  - sliding window across the genome
  - at each location, evaluate the enrichement of the signal wrt. expected background based on the distribution
  - retain regions with P-values below threshold
  - evaluate FDR



Pval < 1e-20

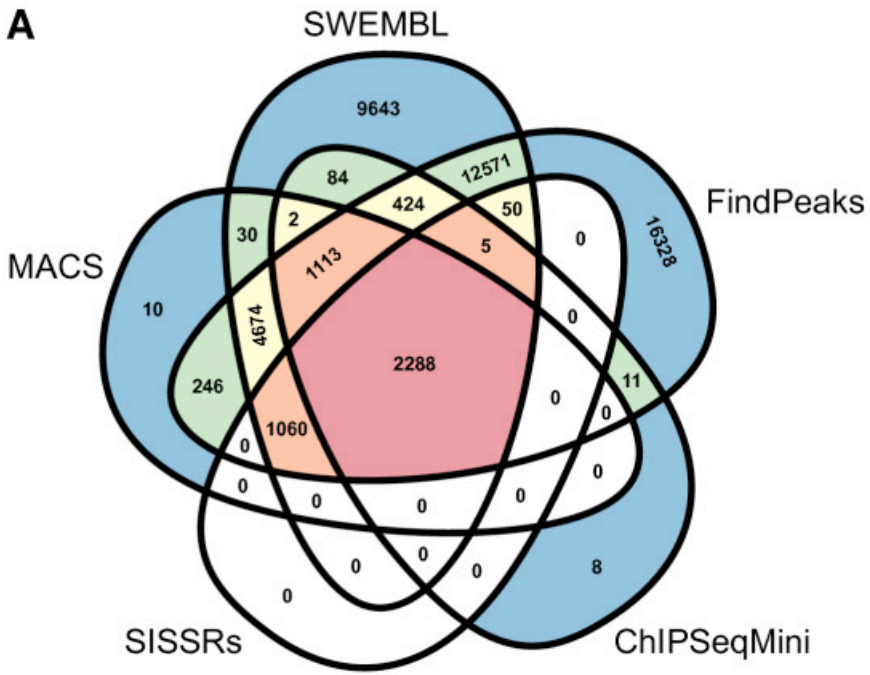


Pval ~ 0.6



Peak-calling programs

- Strong influence on the called peaks
  - Many different programs
  - They do not share the same « default » threshold to retain peaks
  - The top highest peaks are usually common, but the less obvious peaks are often not shared between different peak callers

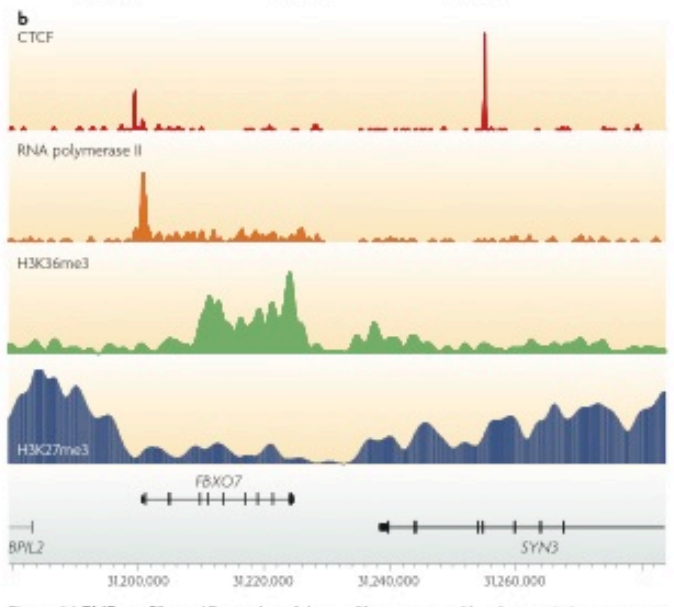


Mali Salmon-Divon *et al*, *BMC Bioinformatics, 201*

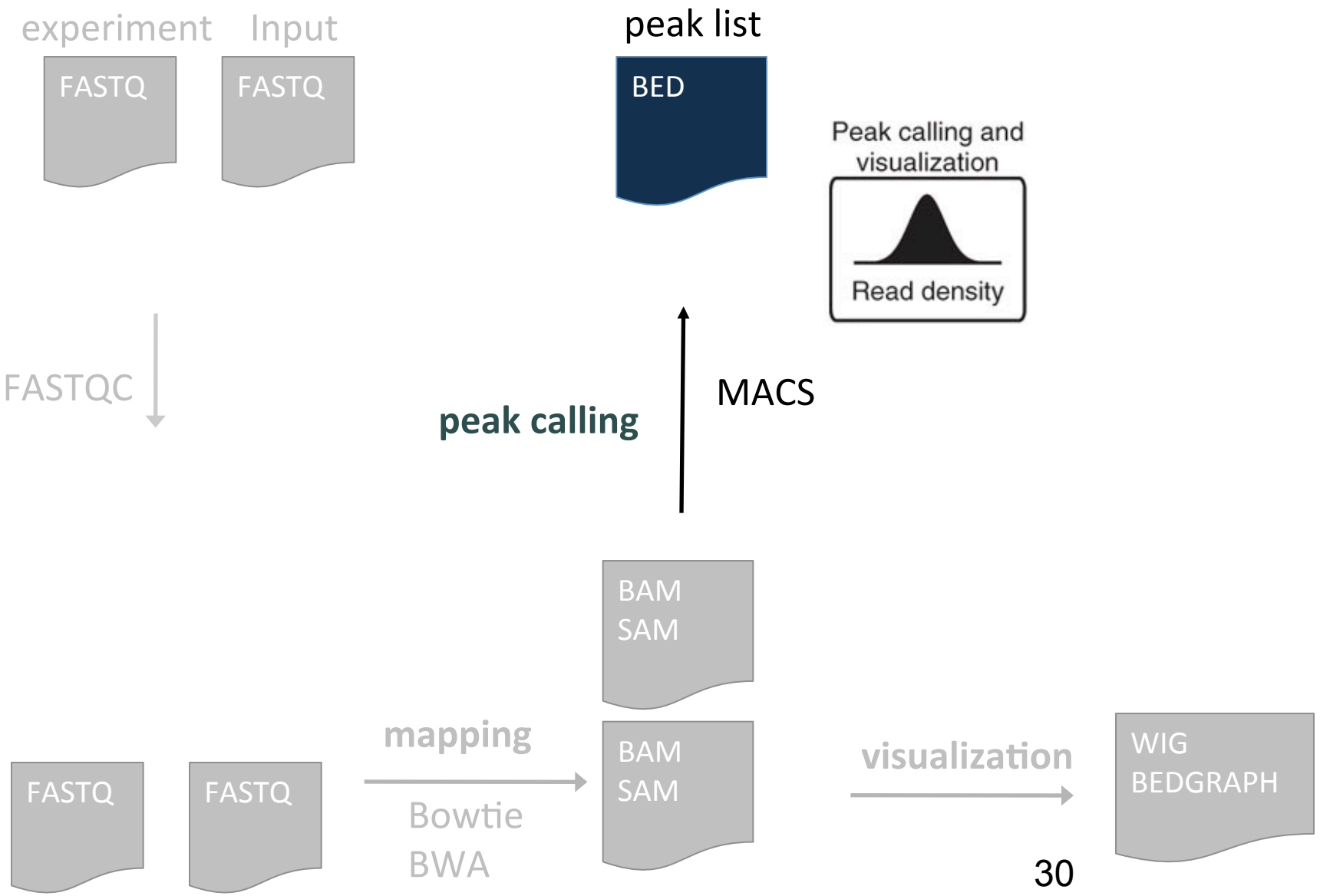
Peak-calling programs

- To be chosen according to type of expected peaks
  - Transcription factors and « sharp » peaks
  - Chromatin marks and « broad peaks »

- Many new programs still being developped !



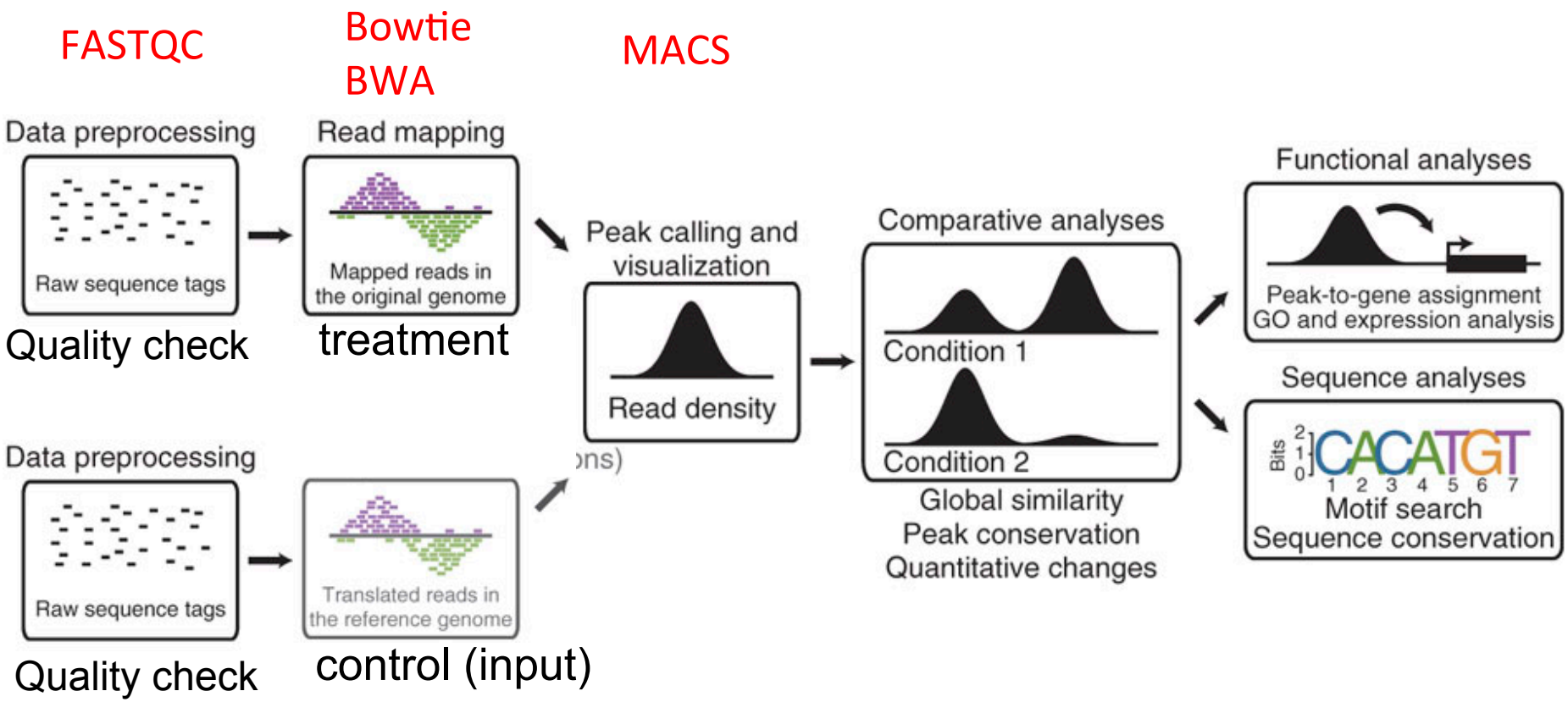
From sequence reads to peaks



Peak list (BED file)

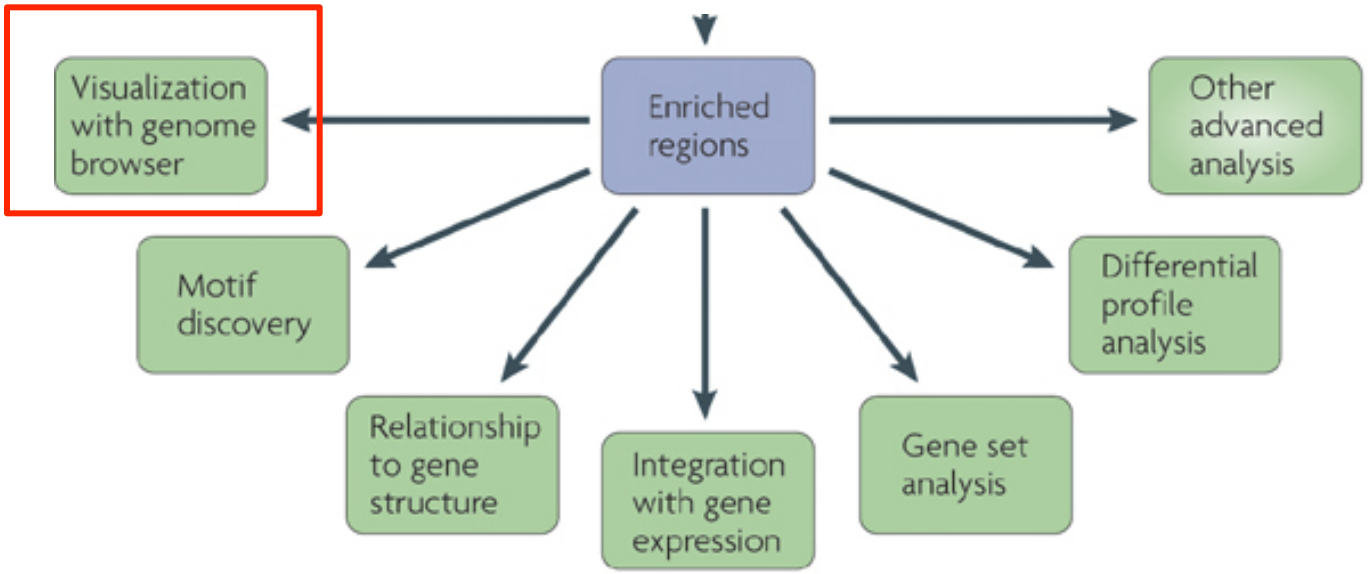
Chromosome	start	end	score1	score2	strand
chr1	145436475	145438649	1478	3206.01	+
chr4	50881	52467	19930	3180.67	+
chr9	31335610	31336400	26372	3170.26	+
chr6	36971531	36973765	22937	3147.85	+
chr4	16234642	16236143	20221	3133.43	+
chr21	40144820	40146203	17188	3131.68	+
chr19	40916830	40918210	13487	3127.46	+
chr4	140477689	140479184	20737	3115.67	+
chr3	12996108	12998488	18417	3108.55	+
chr9	749205	752142	26263	3101.90	+
chr1	11628770	11630411	268	3100.00	+
chr1	153742611	153744775	1556	3100.00	+

ChIP-seq analysis workflow



Adapted from Bardet et al, Nature Protocols, 2012

ChIP-seq analysis workflow: downstream analyses



Park, Nature reviews 2009

Visualizing in a genome browser

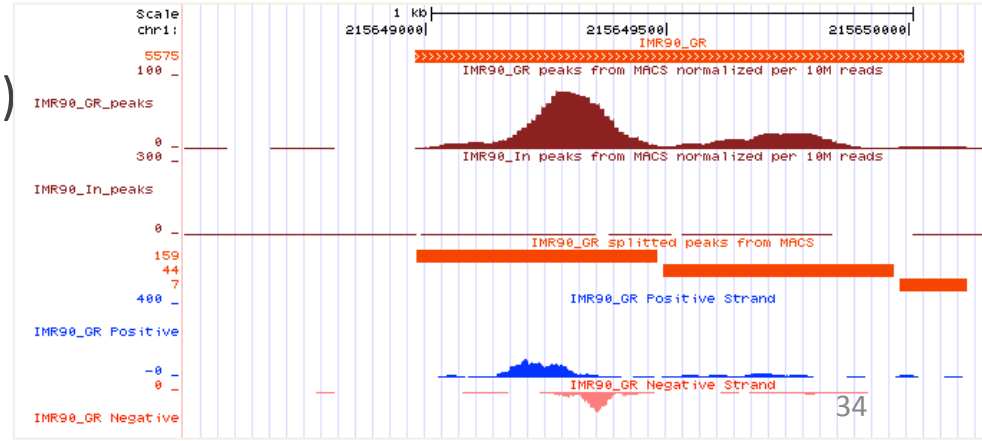
- Local tools (IGV)
  - » Fast
  - » Ideal for sensitive datasets

- web-based tools (UCSC browser) with **custom tracks**
  - » Integrated with many other information (conservation,…)
  - » Easy to share between collaborators

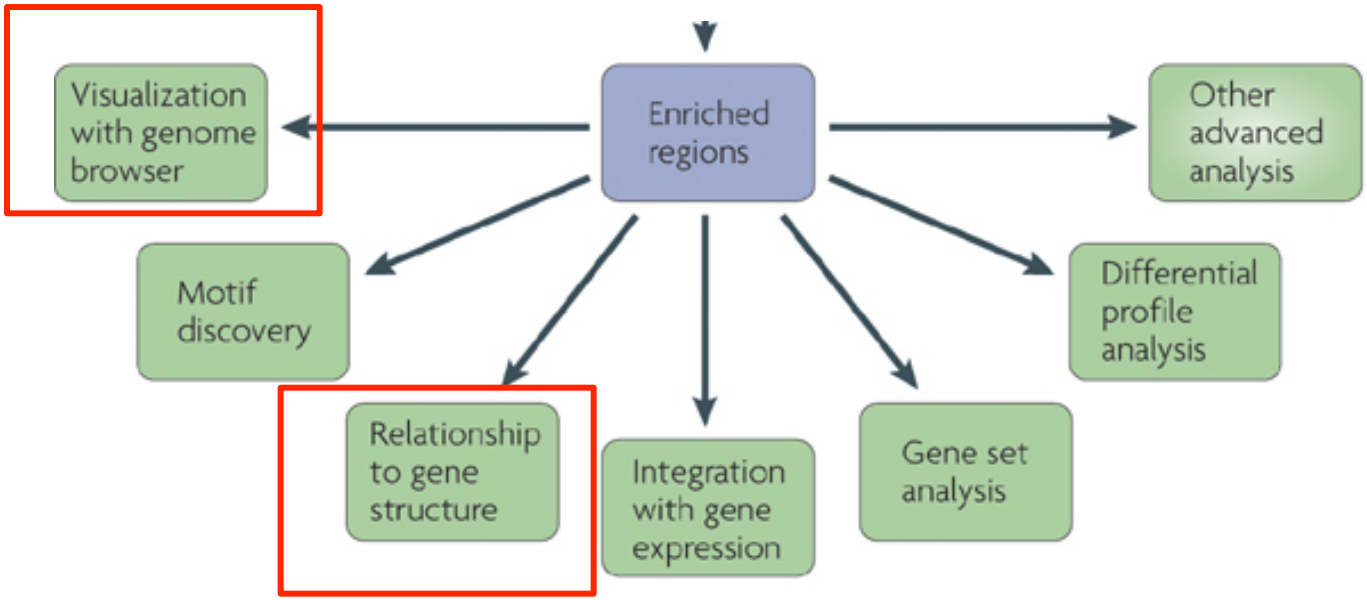
- File formats
  - » BED

  => simply defines a region (start-end)
  - » WIG, bedgraph

  => value assigned to each position

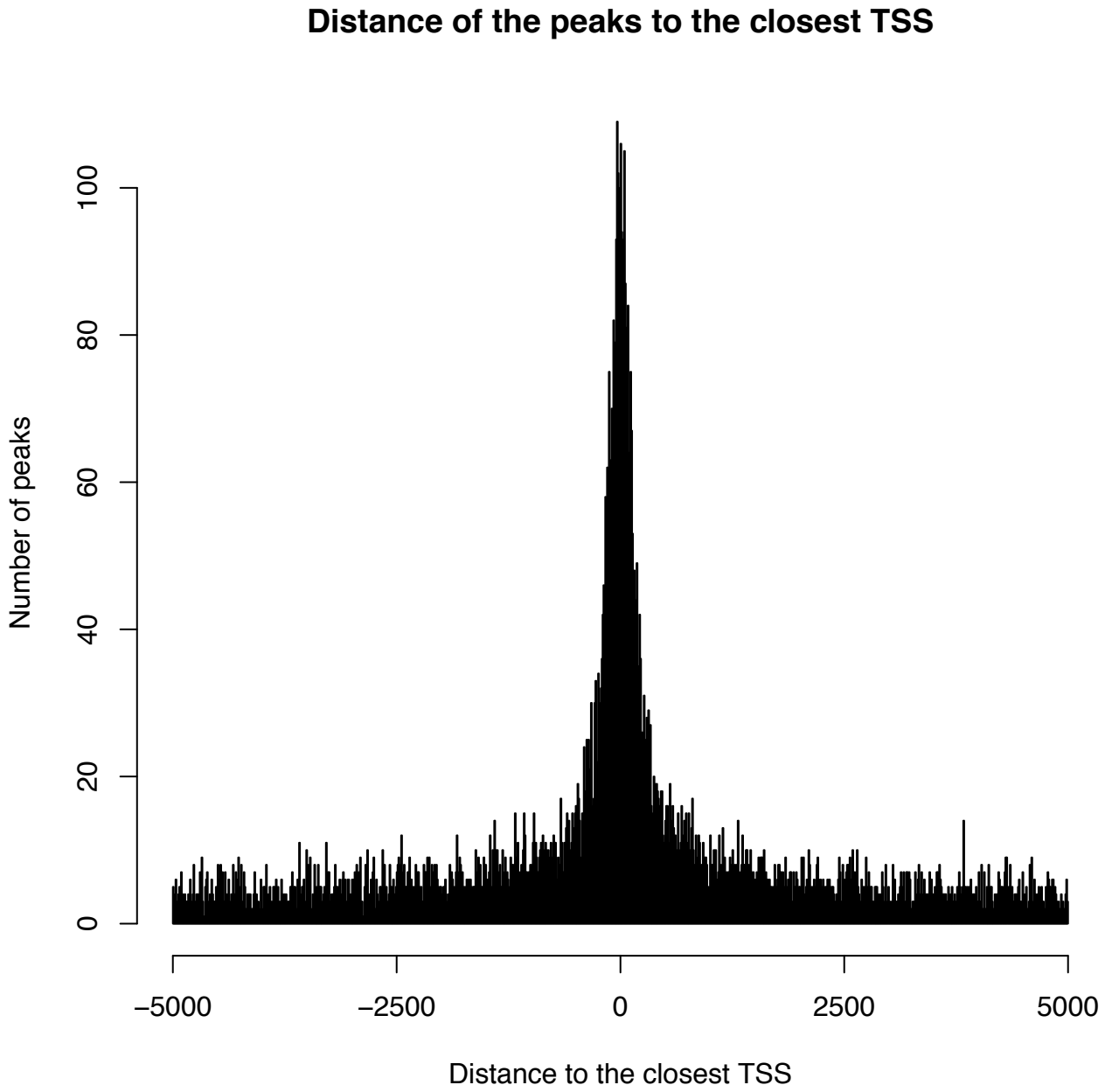


ChIP-seq analysis workflow: downstream analyses

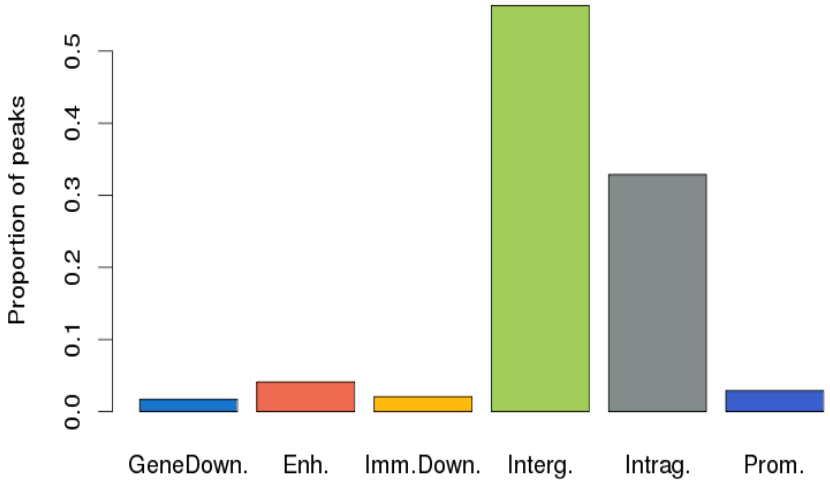


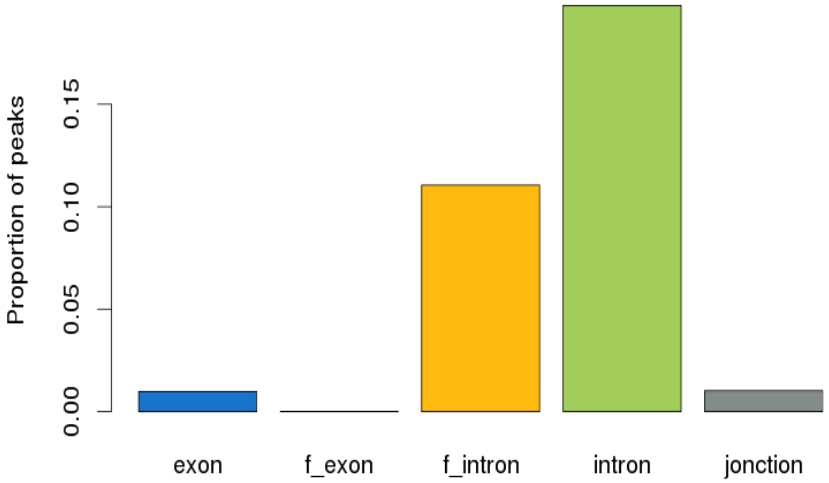
Park, Nature reviews 2009

Distance to closest TSS

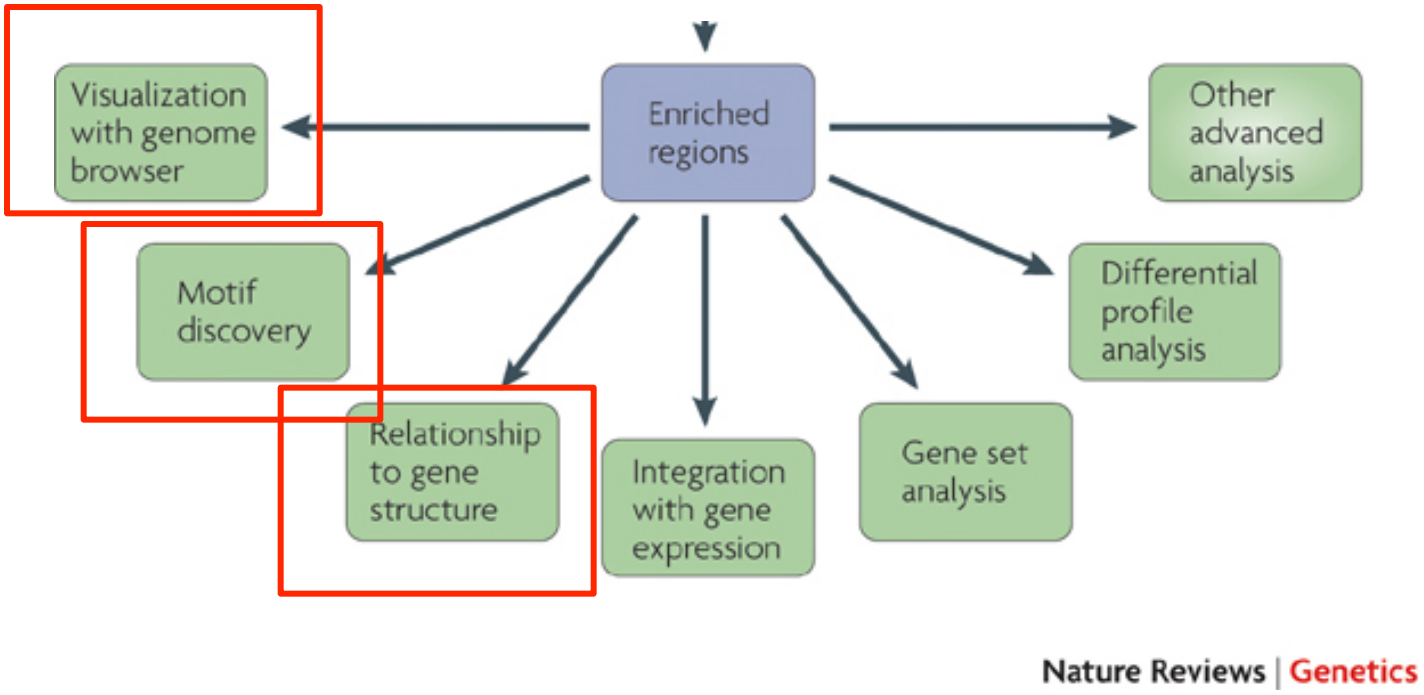


Localisation of the peaks in the genome





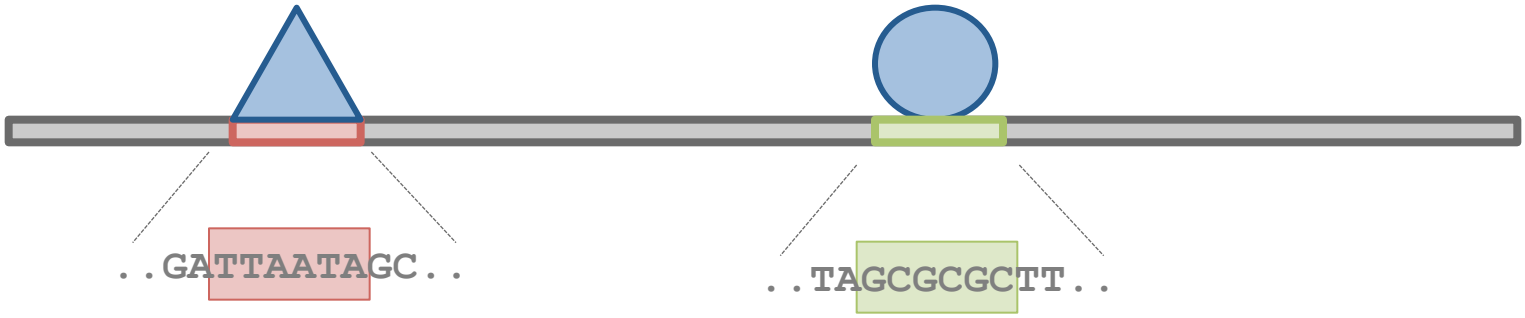
ChIP-seq analysis workflow: downstream analyses



Park, Nature reviews 2009

Transcription factor specificity

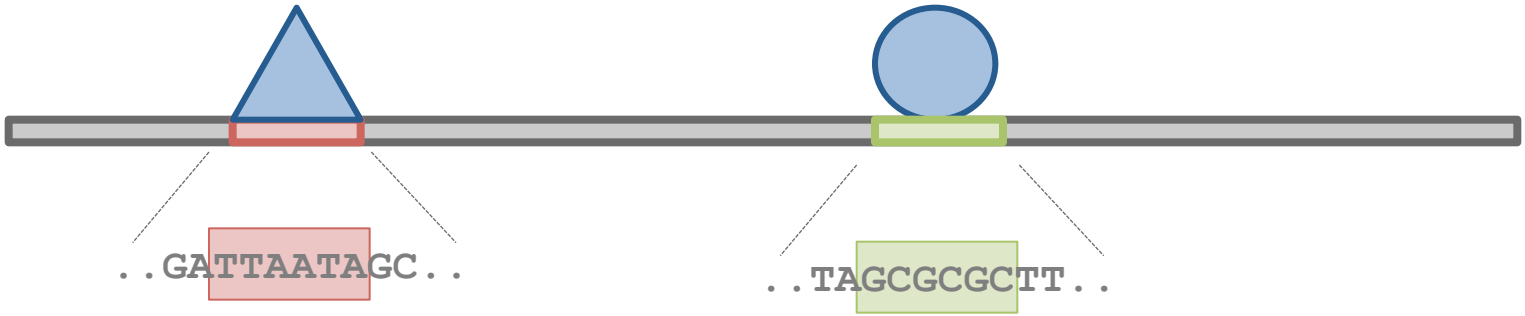
How do TF « know » where to bind DNA ?



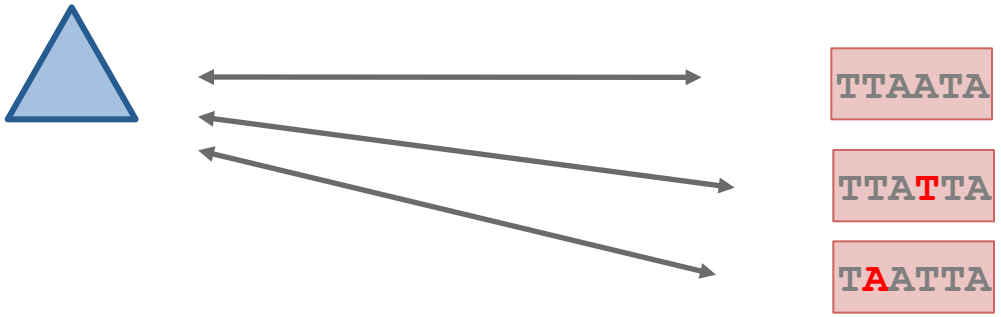
TF recognize TFBS with specific DNA sequences

Transcription factor specificity

How do TF « know » where to bind DNA ?

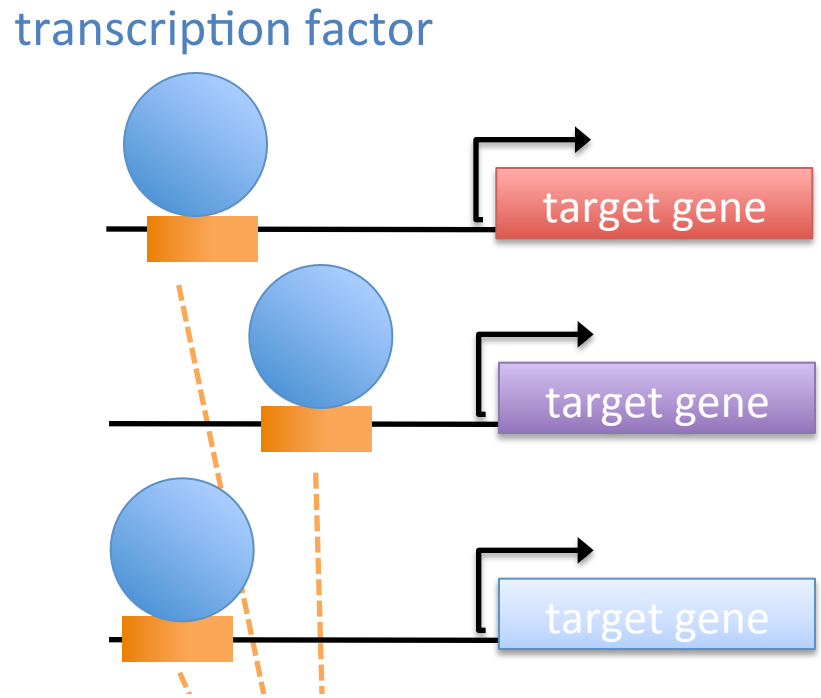


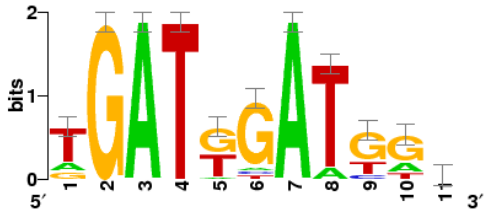
TF recognize TFBS with specific DNA sequences



TFBSs are *degenerate*:
a given TF is able to bind DNA on TFBSs with different sequences

de novo motif discovery





binding motif

Problem :

How can we model/describe the binding specificity of a given TF ?

de novo motif discovery

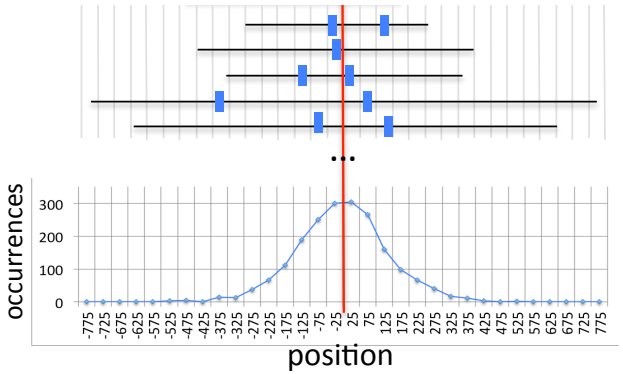
- Find exceptional motifs based on the sequence only

(A priori no knowledge of the motif to look for)

- Criteria of exceptionality:

  – higher/lower frequency than expected by chance (**over-/under-representation**)

  – concentration at specific positions relative to some reference coordinate (**positional bias**)



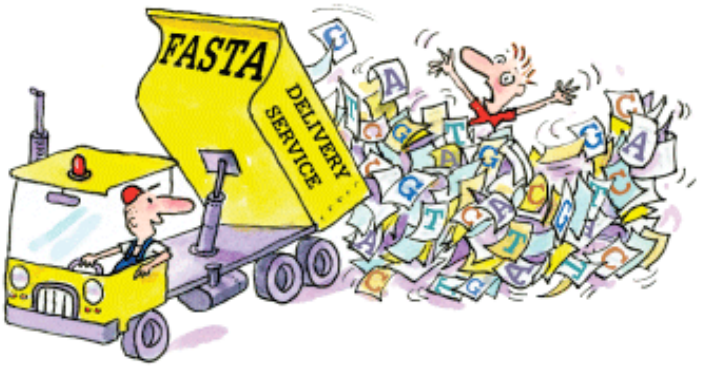
de novo motif discovery

- Tools already exist for a long time !
  - MEME (1994)
  - RSAT oligo-analysis (1998)
  - AlignACE (2000)
  - Weeder (2001)
  - MotifSampler (2001)

Why do we need new approaches for genome-wide datasets ?

New approaches for ChIP-seq datasets

- **Size, size, size**
  - limited numbers of promoters and enhancers

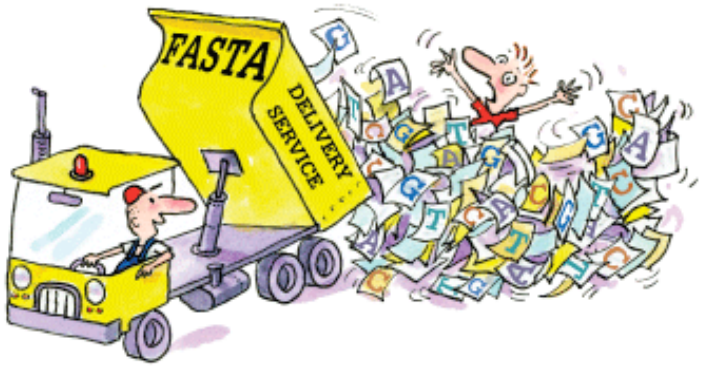


  - dozens of thousands of peaks !!!!!!

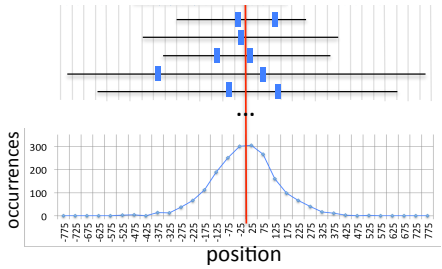
http://www.genomequest.com/landing-pages/ODI-webinar-web.html

New approaches for ChIP-seq datasets

- **Size, size, size**
  - limited numbers of promoters and enhancers
  - dozens of thousands of peaks !!!!!!

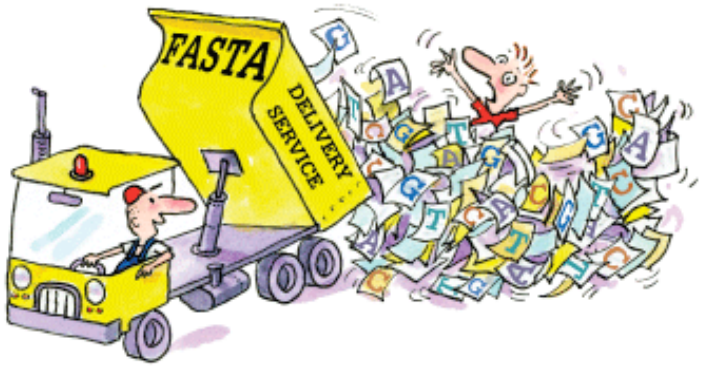


- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias

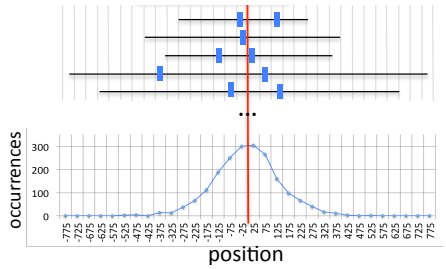


http://www.genomequest.com/landing-pages/ODI-webinar-web.html

New approaches for ChIP-seq datasets



- **Size, size, size**
  - limited numbers of promoters and enhancers
  - dozens of thousands of peaks !!!!!!
- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias



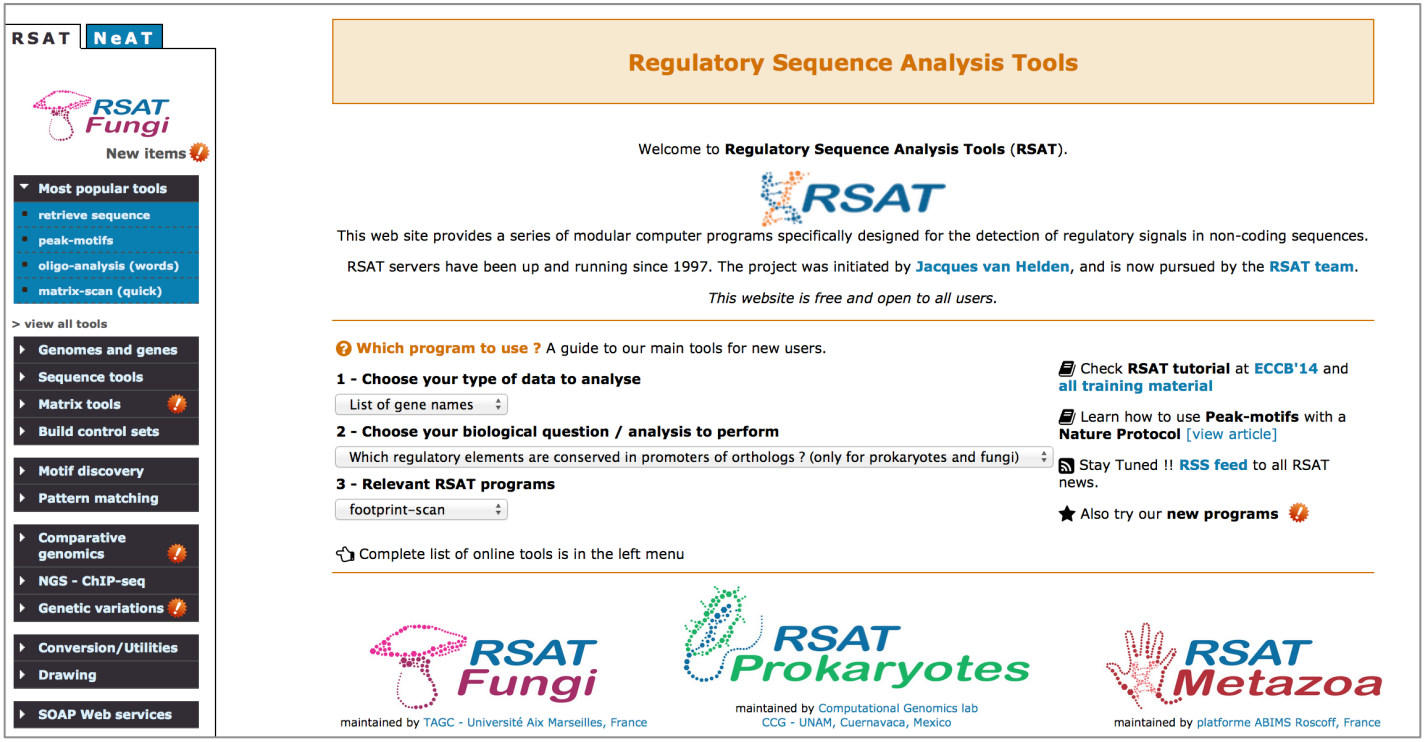
- **motif analysis: not just for specialists anymore !**
  - complete user-friendly workflows

http://rsat.eu

RSAT

Jacques van Helden

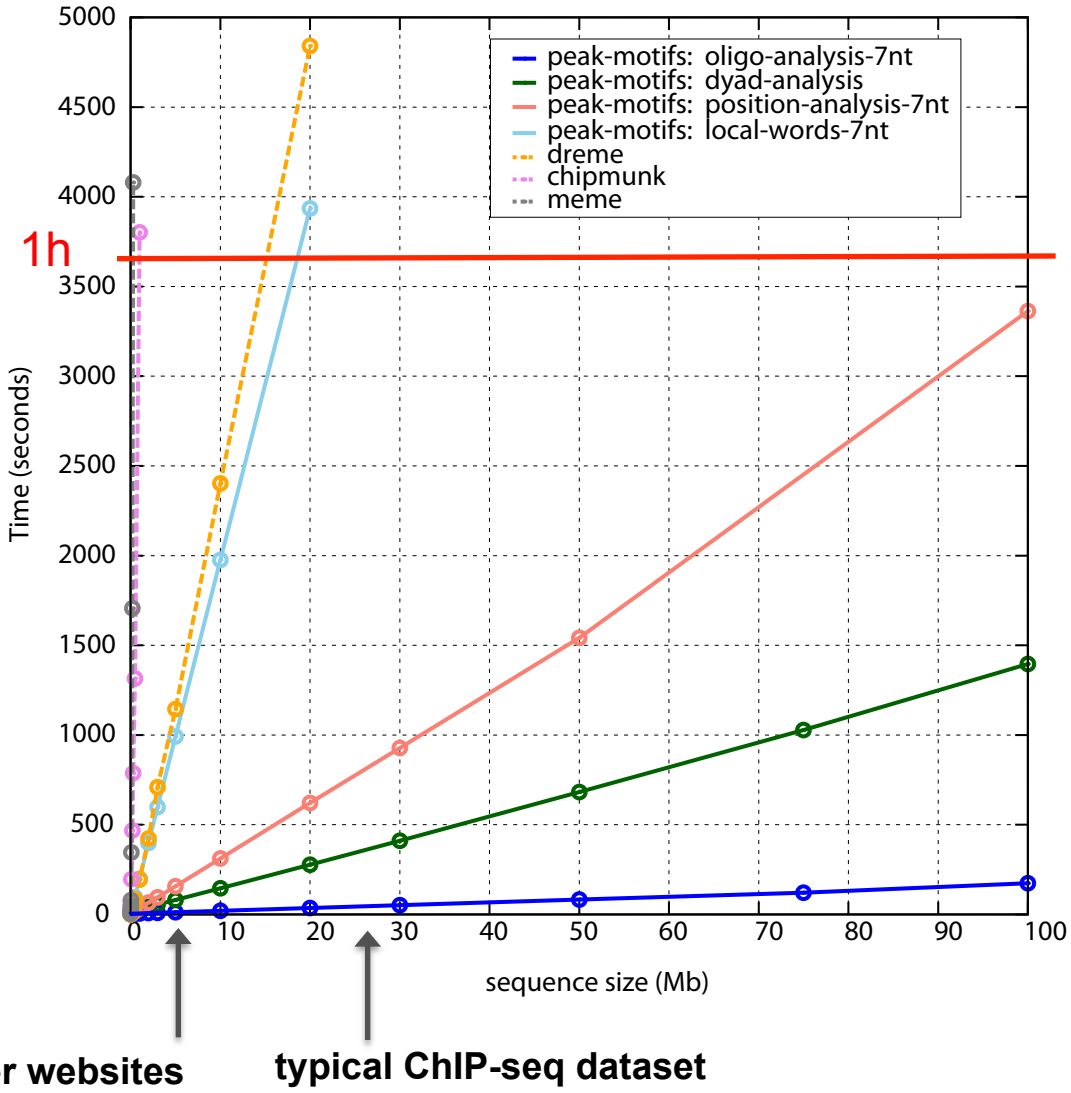
Aix*Marseille université



Thomas-Chollier, Defrance, Medina-Rivera, Sand, Herrmann, Thieffry, van Helden ***Nucleic Acids Research****, 2011*
Medina-Rivera, Abreu-Goodger, Thomas-Chollier, Salgado, Collado-Vides, van Helden ***Nucleic Acids Research****, 2011*
Sand, Thomas-Chollier, van Helden ***Bioinformatics****, 2009*
Thomas-Chollier, Sand*, Turatsinze, Janky, Defrance, Vervisch, van Helden* ***Nucleic Acids Research****, 2008*
Sand, Thomas-Chollier, Vervisch, van Helden ***Nature Protocols****, 2008*
Thomas-Chollier, Turatsinze*, Defrance, van Helden* ***Nature Protocols****, 2008*
van Helden, ***Nucleic Acids Research****, 2003*

RSAT peak-motifs

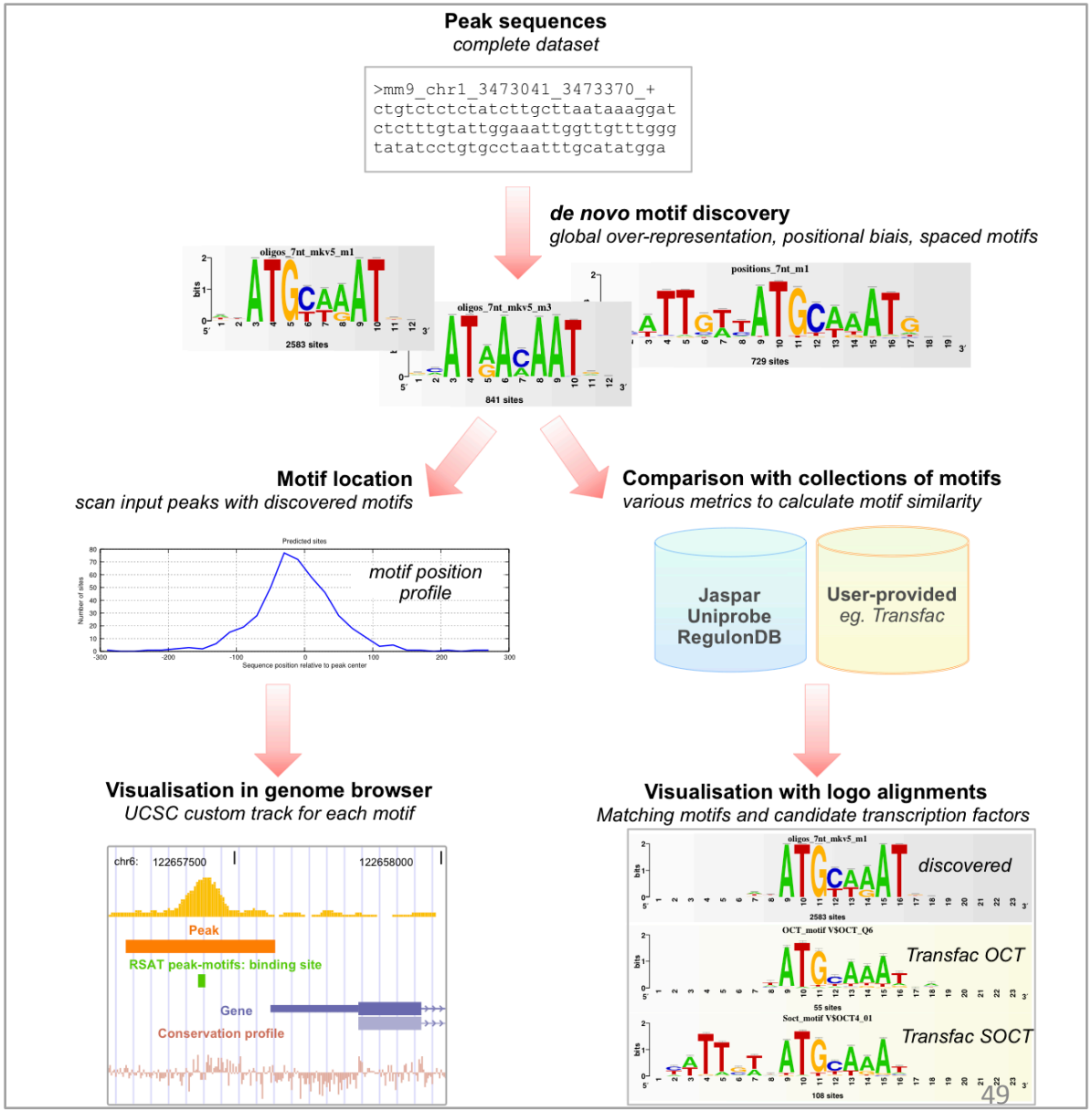
- fast and scalable
- treat full-size datasets



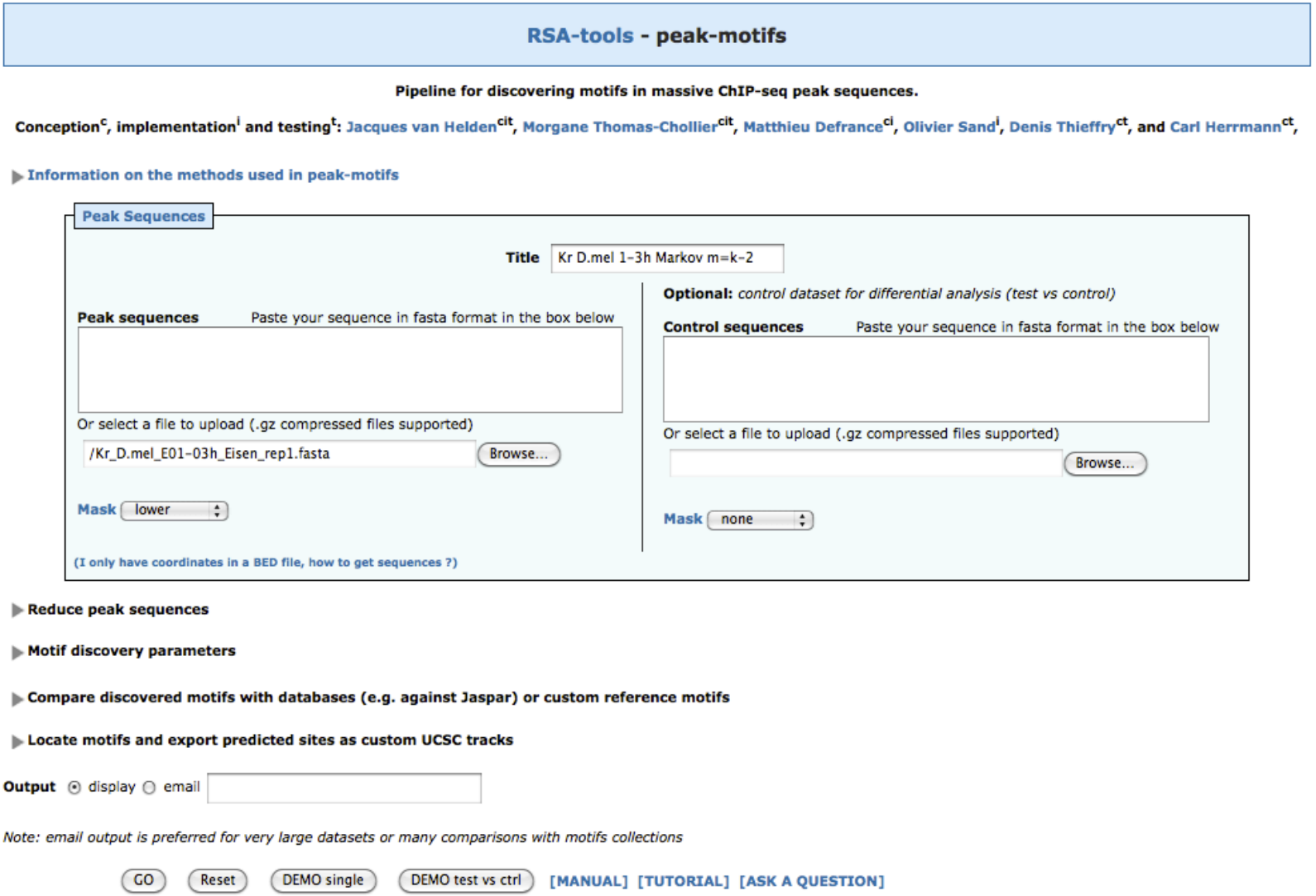
Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*

RSAT peak-motifs

- fast and scalable
- treat full-size datasets
- **complete pipeline**

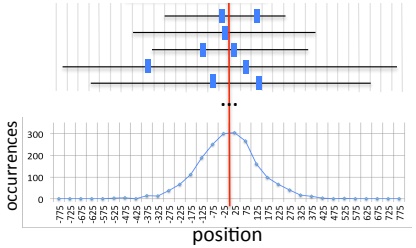


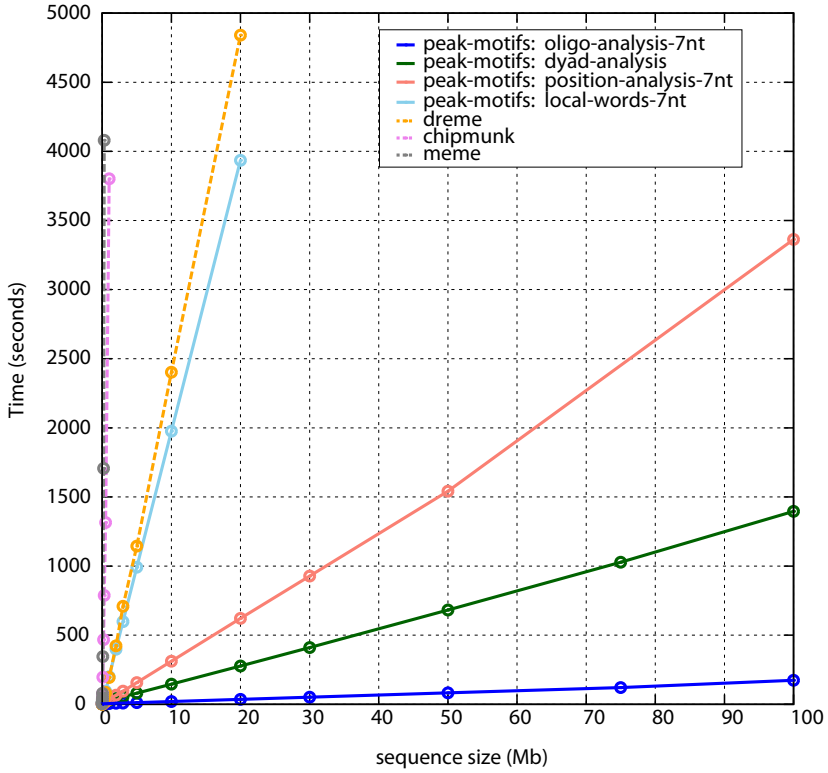
Thomas-Chollier, Herrmann, Defrance, Sand, Thieffry, van Helden ***Nucleic Acids Research****, 2012*


RSA-tools - peak-motifs
Pipeline for discovering motifs in massive ChIP-seq peak sequences.
Conception[c], implementation[i] and testing[t]: Jacques van Helden[cit], Morgane Thomas-Chollier[cit], Matthieu Defrance[ci], Olivier Sand[i], Denis Thieffry[ct], and Carl Herrmann[ct],
Information on the methods used in peak-motifs
Peak Sequences
Title
Kr D.mel 1-3h Markov m=k-2
Peak sequences
Paste your sequence in fasta format in the box below
Or select a file to upload (.gz compressed files supported)
/Kr_D.mel_E01-03h_Eisen_rep1.fasta
Browse...
Mask
lower
(I only have coordinates in a BED file, how to get sequences ?)
Optional: control dataset for differential analysis (test vs control)
Control sequences
Paste your sequence in fasta format in the box below
Or select a file to upload (.gz compressed files supported)
Browse...
Mask
none
Reduce peak sequences
Motif discovery parameters
Compare discovered motifs with databases (e.g. against Jaspar) or custom reference motifs
Locate motifs and export predicted sites as custom UCSC tracks
Output
display
email
Note: email output is preferred for very large datasets or many comparisons with motifs collections
GO
Reset
DEMO single
DEMO test vs ctrl
[MANUAL] [TUTORIAL] [ASK A QUESTION]

RSAT peak-motifs

- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface
- accessible to non-specialists
- **using 4 complementary algorithms**
  - Global over-representation
    - oligo-analysis
    - dyad-analysis (spaced motifs)
  - Positional bias
    - position-analysis
    - local-words

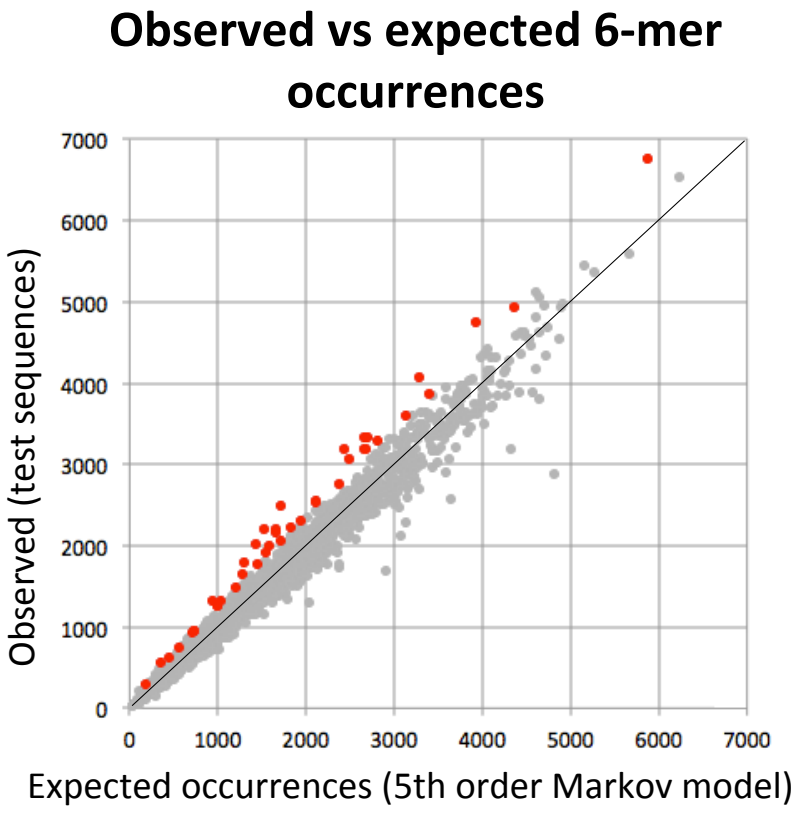




Motif discovery methods: frequency

Observed 6-mer occurrences computed from:	**Expected** 6-mer occurrences computed from:
6-mer (e.g.AACAAA)	**Background sequences (when available)**
Test sequences	OR
	Theoretical k-mers frequencies from test sequences

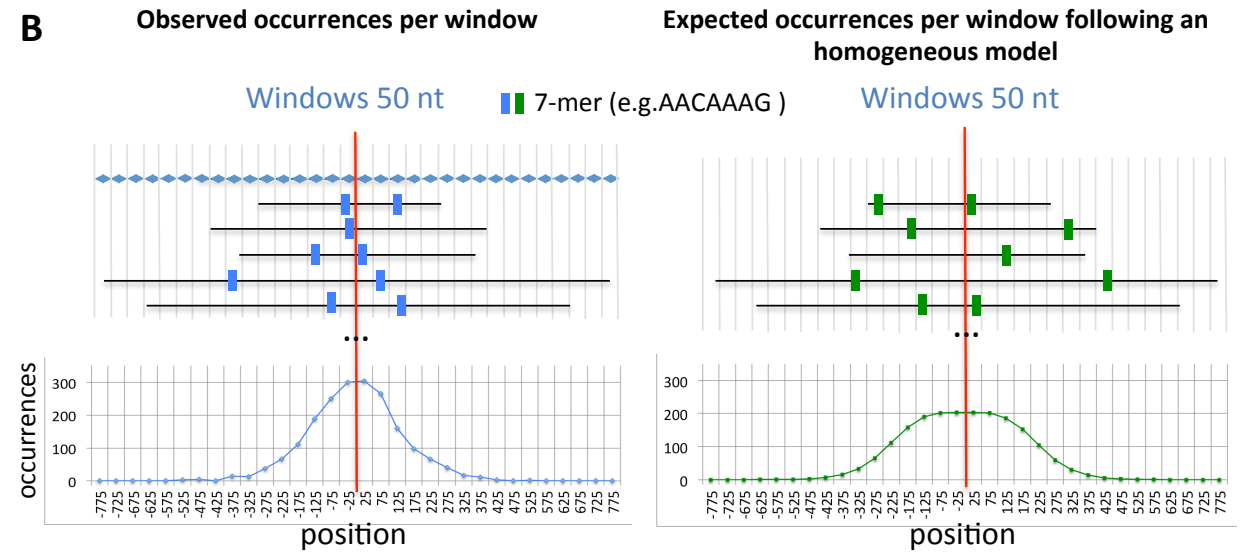
➔ Computation of p-value (binomial) and E-value (multi-testing correction)



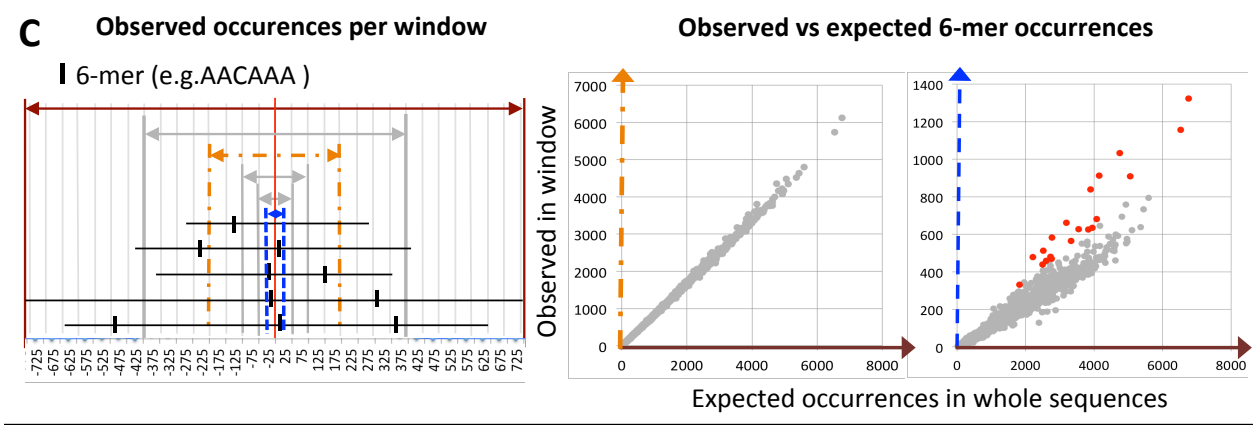
oligo-analysis
dyad-analysis (spaced motifs)

Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden ***Nature Protocols****,2012*

Motif discovery methods: positional bias



position-analysis

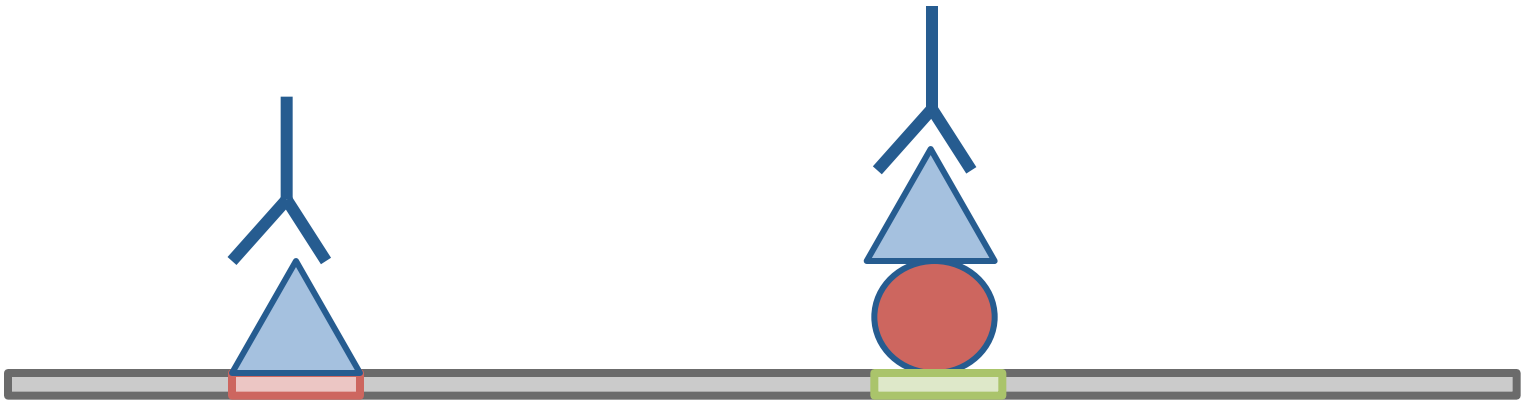


local-words

Thomas-Chollier, Darbo, Herrmann, Defrance, Thieffry, van Helden ***Nature Protocols****,2012*

Direct versus indirect binding

- ChIP-seq does not necessarily reveal **direct binding**



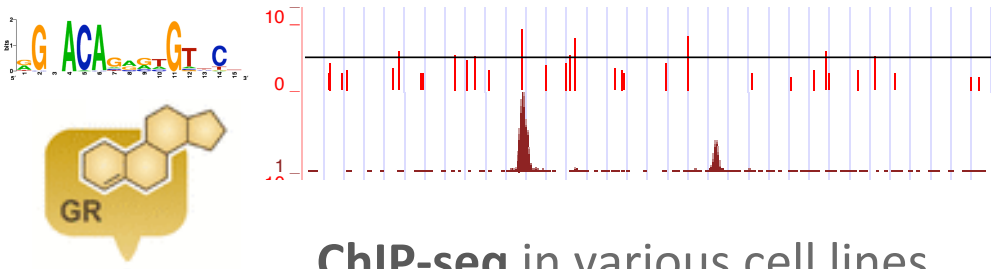
- The motif of the targeted TF is not always found in peaks !

What's next ?

- Practical with the Galaxy platform and RSAT (now)

- Seminar1 (tomorrow at 12 in auditorio de la Facultad)

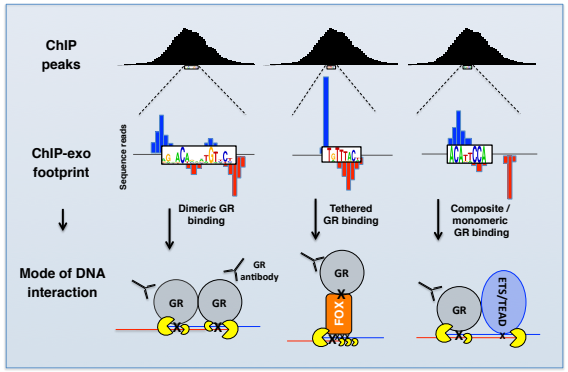
=> biological applications : Glucocorticoid Receptor and histone marks



ChIP-seq in various cell lines
(U2OS, IMR90, Nalm6, K562)

- Seminar2 (Tuesday 17th 11 am in the auditorium of CCG)

=> ChIP-exo



To read further …

- **ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions**
  - » Terrence S. Furey - Nature Reviews Genetics 13, 840-852 (December 2012)

- **ChIP-Seq: advantages and challenges of a maturing technology**
  - » Peter J. Park - Nat Rev Genet. 2009 October; 10(10): 669–680

- **Computation for ChIP-seq and RNA-seq studies**
  - » Shirley Pepke et al - Nature Methods 6, S22 - S32 (2009)