ChIP-seq analysis

Morgane Thomas-Chollier

Samuel Collombet

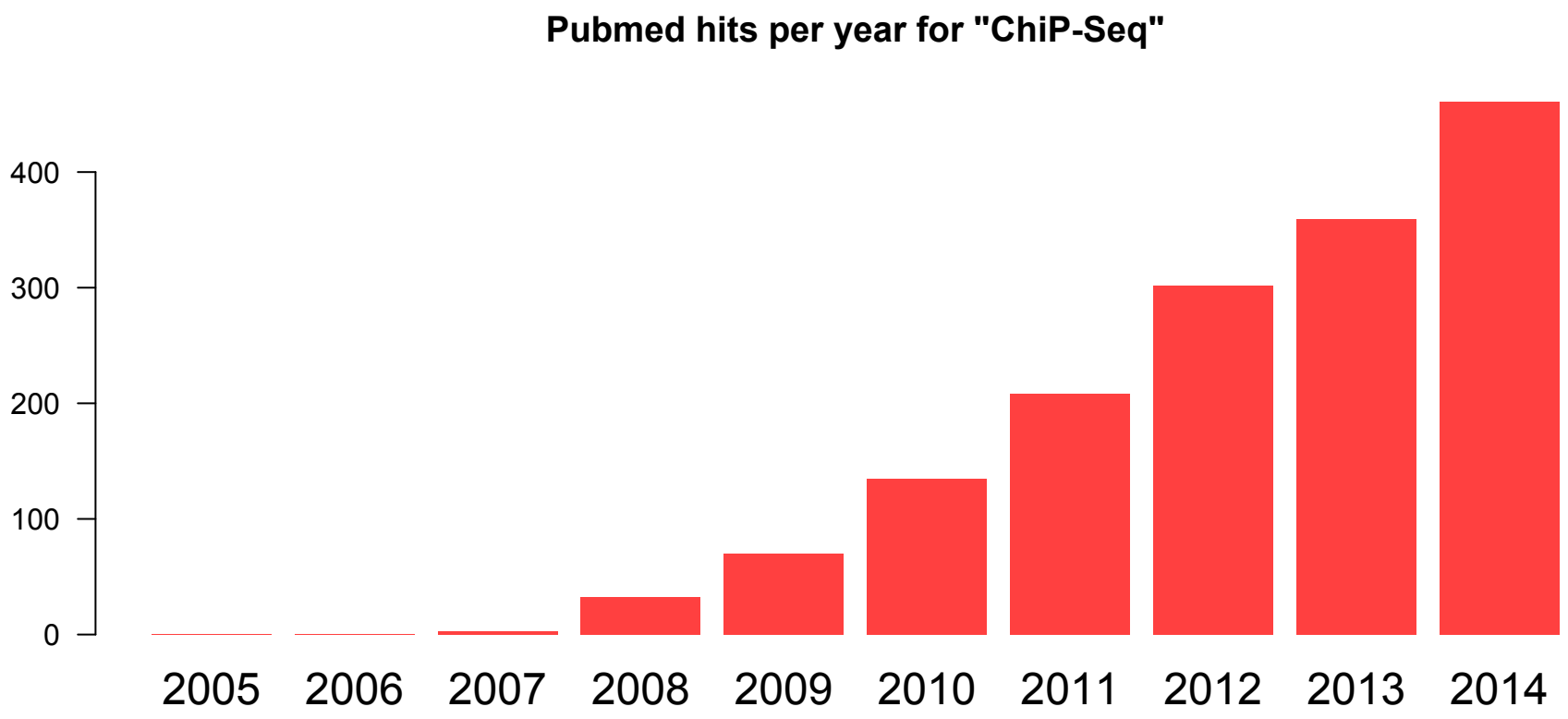
Computational systems biology - IBENS

mthoma IBENS .ens.fr

M2 – Computational analysis of cis-regulatory sequences 2014/2015

Denis Thieffry, Jacques van Helden and Carl Herrmann kindly shared some of their slides.

The ChIP-seq era



- Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein– DNA interactions. Science 316: 1497–1502.
- Barski A, Cuddapah S, Cui K, Roth TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823–837.
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immu- noprecipitation and massively parallel sequenc- ing. Nat Methods 4: 651–657.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieber- man E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.

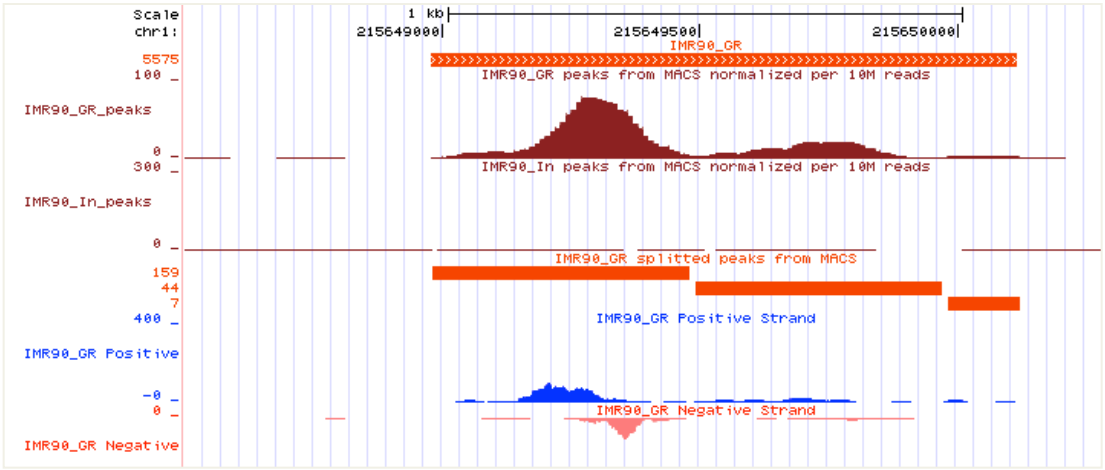
in vivo experimental methods to identify binding sites

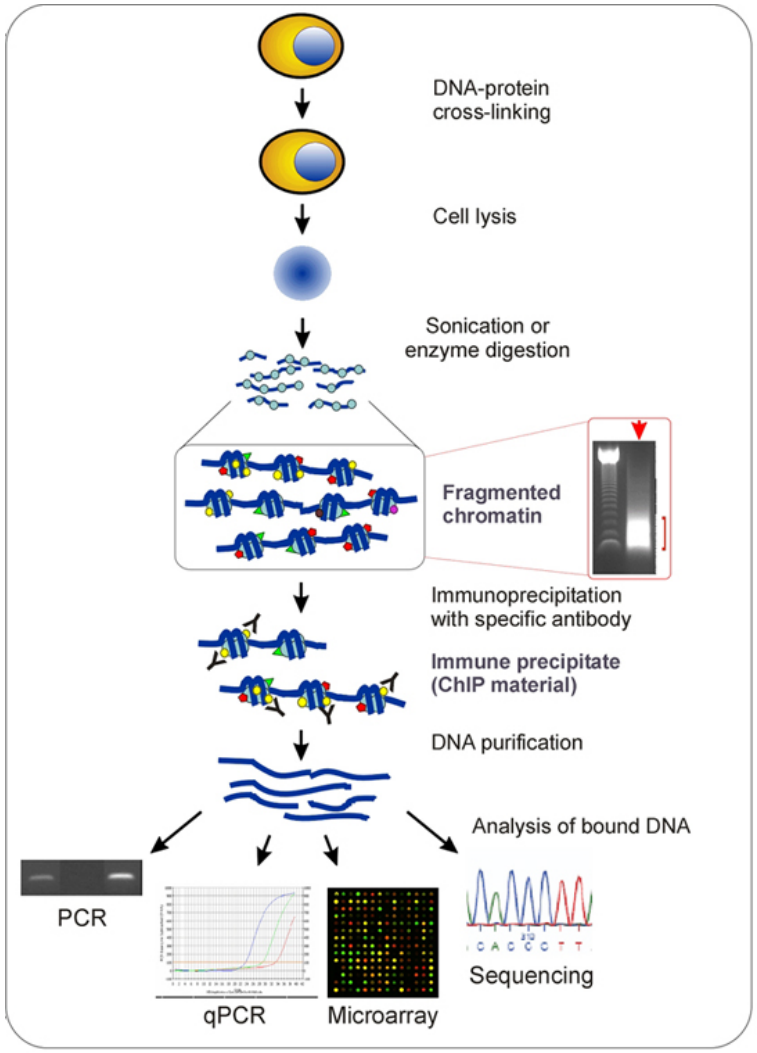
ChIP (=Chromatin Immuno-Precipitation)

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

-small-scale: PCR / qPCR

- large-scale:
  - microarray = **ChIP-on-chip**
  - sequencing = **ChIP-seq**





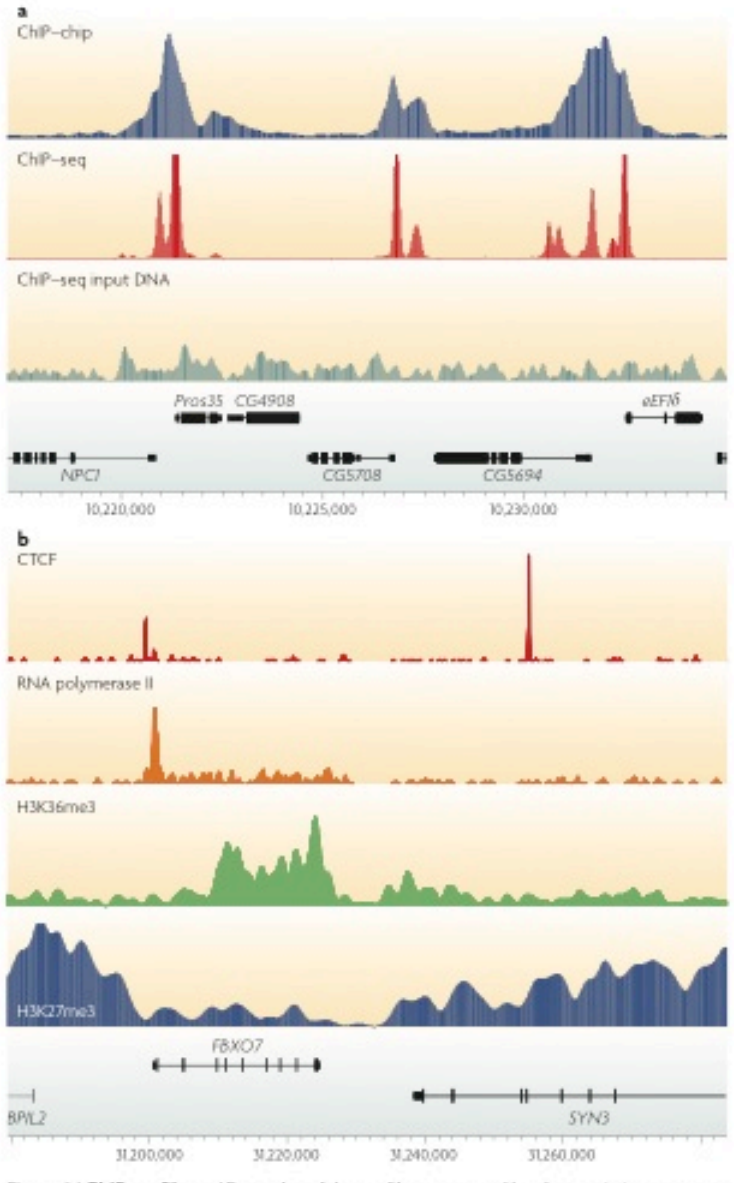
http://www.chip-antibodies.com/

ChIP-seq application

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

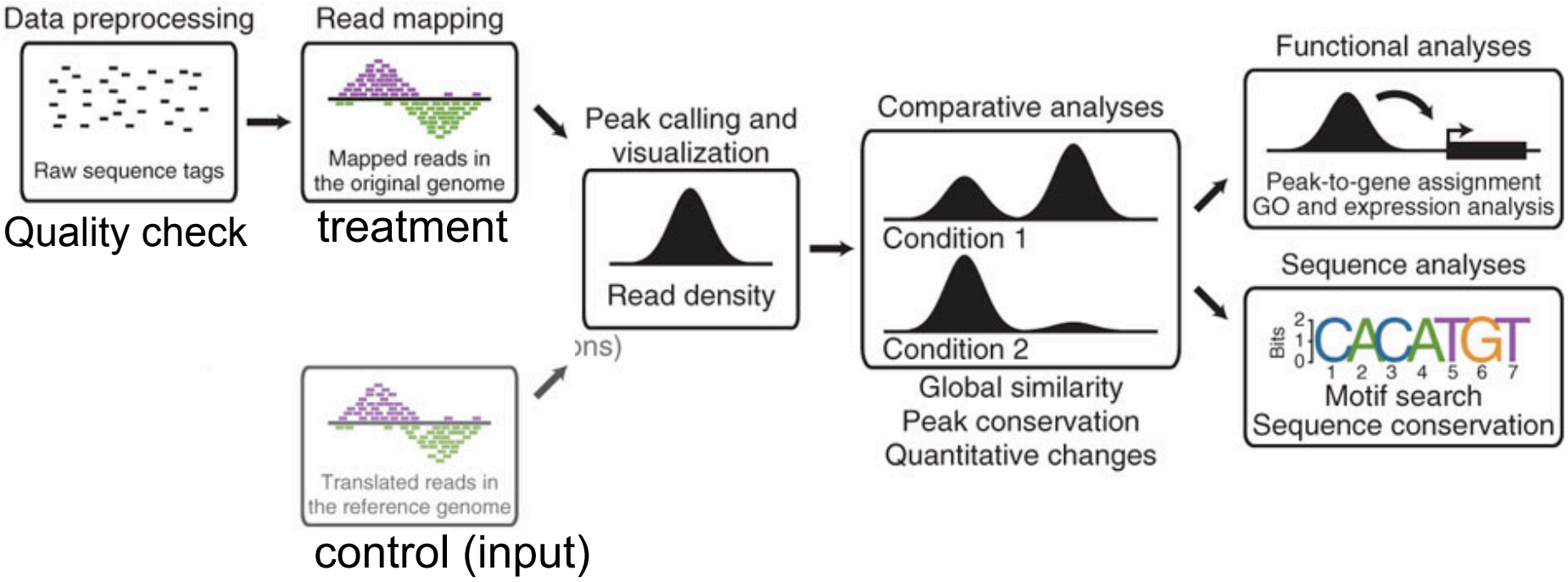
- in a given ***experimental condition*** (cell type, developmental stage,...)

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



Park, Nature reviews 2009

Aim of the course : ChIP-seq analysis workflow

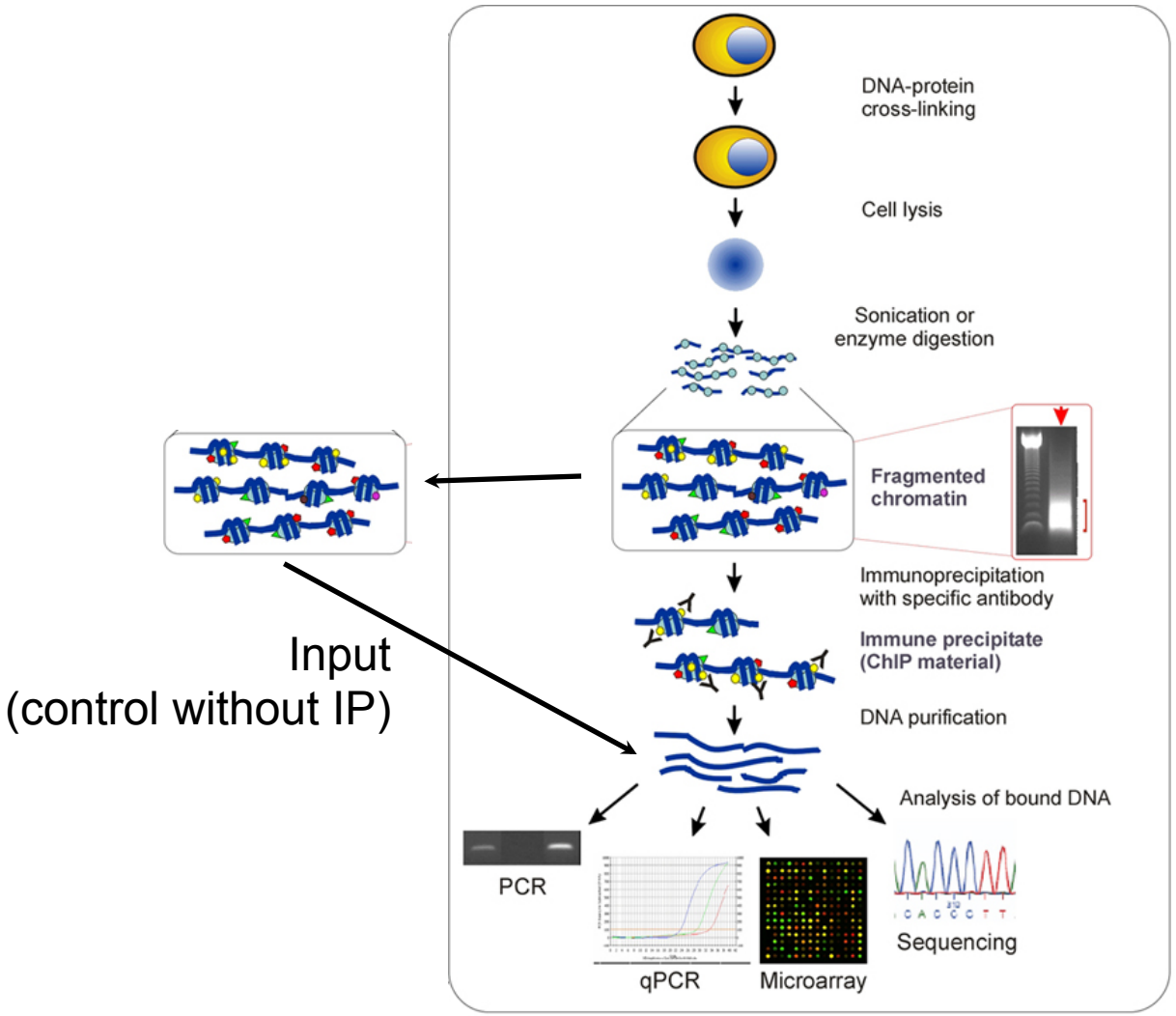


Processing steps

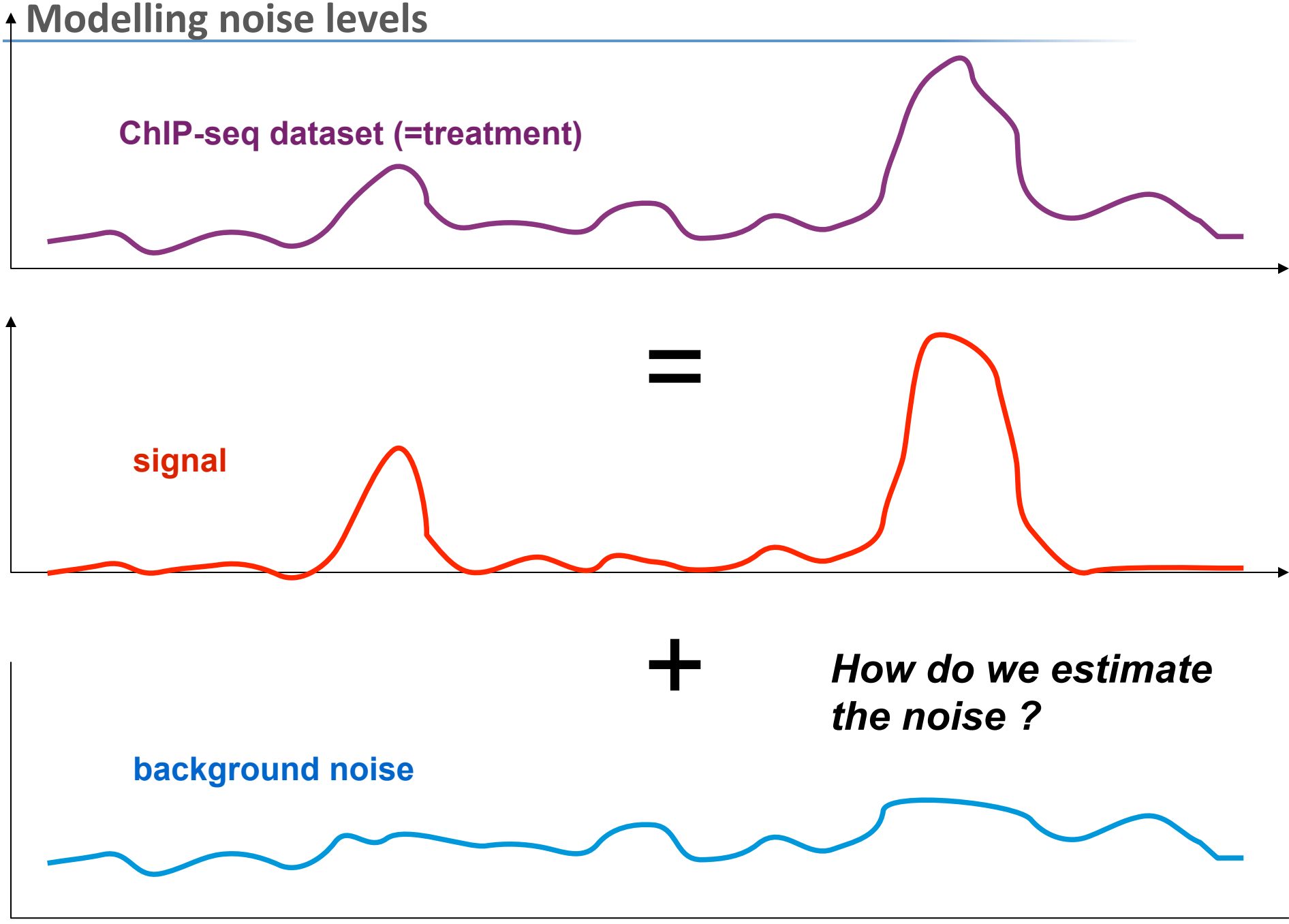
Downstream analyses

Adapted from Bardet et al, Nature Protocols, 2012

Control: input



Modelling noise levels

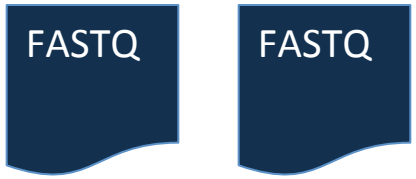


Modelling noise levels

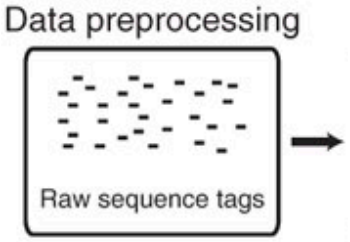
- noise is **not uniform** (chromatin conformation, local biases, mappability)
- input dataset is **mandatory** for reliable local estimation ! (although some algorithms do not require it … :-( )

From sequence reads to peaks

experiment Input



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



Quality check

FASTQ format 1 read = 4 lines

```
@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
```

FASTA format

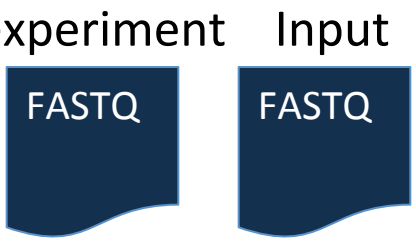
```
>SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
>SRR002012.2 Oct4:5:1:804:348
GTCTGCATTATCTACCAGCACTTCCC
>SRR002012.3 Oct4:5:1:767:334
GCTGTCTTCCCGCTGTTTTATCCCCC
>SRR002012.4 Oct4:5:1:805:329
GTAGTTTACCTGTTCATATGTTTCTG
```

adapted from Wikipedia

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                          |    |        |                              |                     |
 33                         59   64       73                            104                   126
  0                                        40

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
```

From sequence reads to peaks



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

FASTQC

quality check

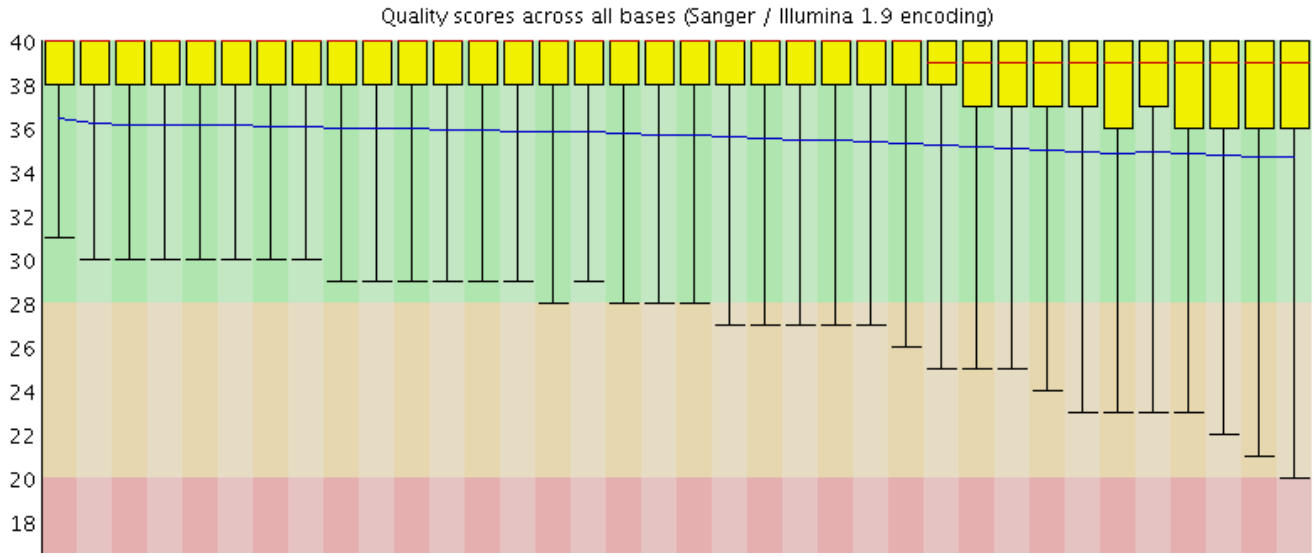
FastQC Report

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/

http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010
http://bioinfo.cipf.es/courses/mda11/lib/exe/fetch.php?media=ngs_qc_tutorial_mda_val_2011.pdf

modEncode Kni Drosophila

FastQC Report

Wed 14 Sep 2011
SRR063881.fq

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

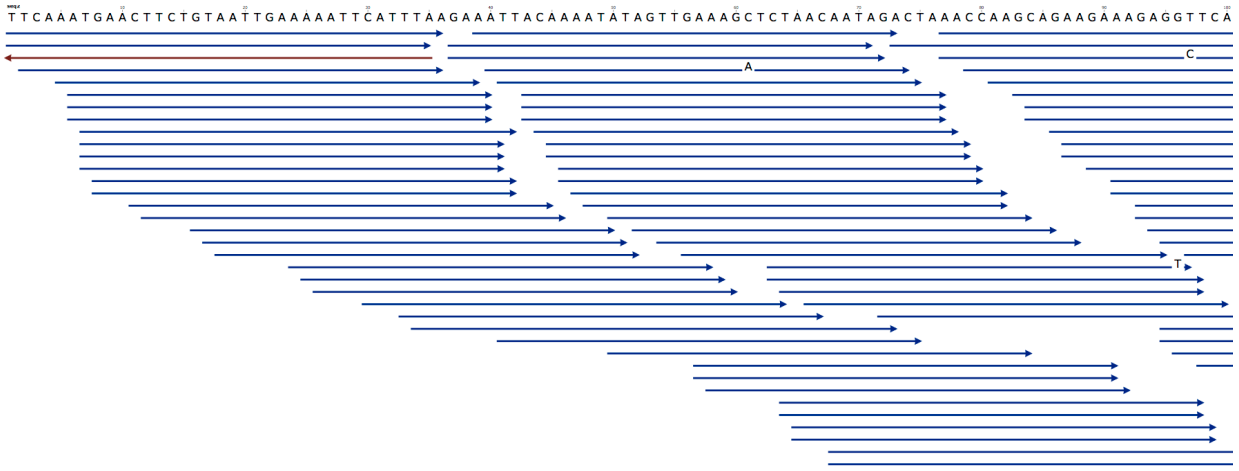
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AATGATACGGCGACCACCGAGATCTACACTCTTTCC	94455	2.3039372679262806	Illumina Single End PCR Primer 1 (100% over 36bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTAGAT	63222	1.5421049383604395	Illumina Single End Apapter 2 (96% over 32bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA	35701	0.8708153554839464	Illumina Single End Apapter 2 (100% over 33bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAA	27284	0.6655087016897003	Illumina Single End Apapter 2 (100% over 33bp)
CCTGCTCCGGCGACCACCGAGATCTACACTCTTTCC	12598	0.30728920333847104	Illumina Single End PCR Primer 1 (100% over 29bp)
CATGATACGGCGACCACCGAGATCTACACTCTTTCC	9924	0.24206525273305185	Illumina Single End PCR Primer 1 (97% over 36bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTAGAA	9073	0.22130774264882905	Illumina Single End Apapter 2 (96% over 32bp)
AATGATACGGCGACCCCCGAGATCTACACTCTTTCC	8711	0.21247787349431826	Illumina Single End PCR Primer 1 (97% over 36bp)
GATCGGACGAGCTCGTCTGCCGTCTTCTGCTTAGAT	8562	0.20884347983679863	Illumina PCR Primer Index 1 (95% over 21bp)
CCTGATACGGCGACCACCGAGATCTACACTCTTTCC	8409	0.20511151856431206	Illumina Single End PCR Primer 1 (100% over 34bp)
GACCACCGAGATCGGAAGAGCTCGTATGCCGTCTTC	6193	0.15105905987261084	Illumina Single End Apapter 2 (100% over 28bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGGAT	4127	0.10066538674217099	Illumina Single End Apapter 2 (100% over 33bp)

From sequence reads to peaks

experiment Input

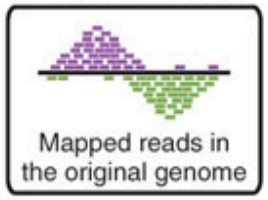
FASTQ FASTQ



FASTQC

Source: http://trac.seqan.de

Read mapping



BED
BAM
SAM

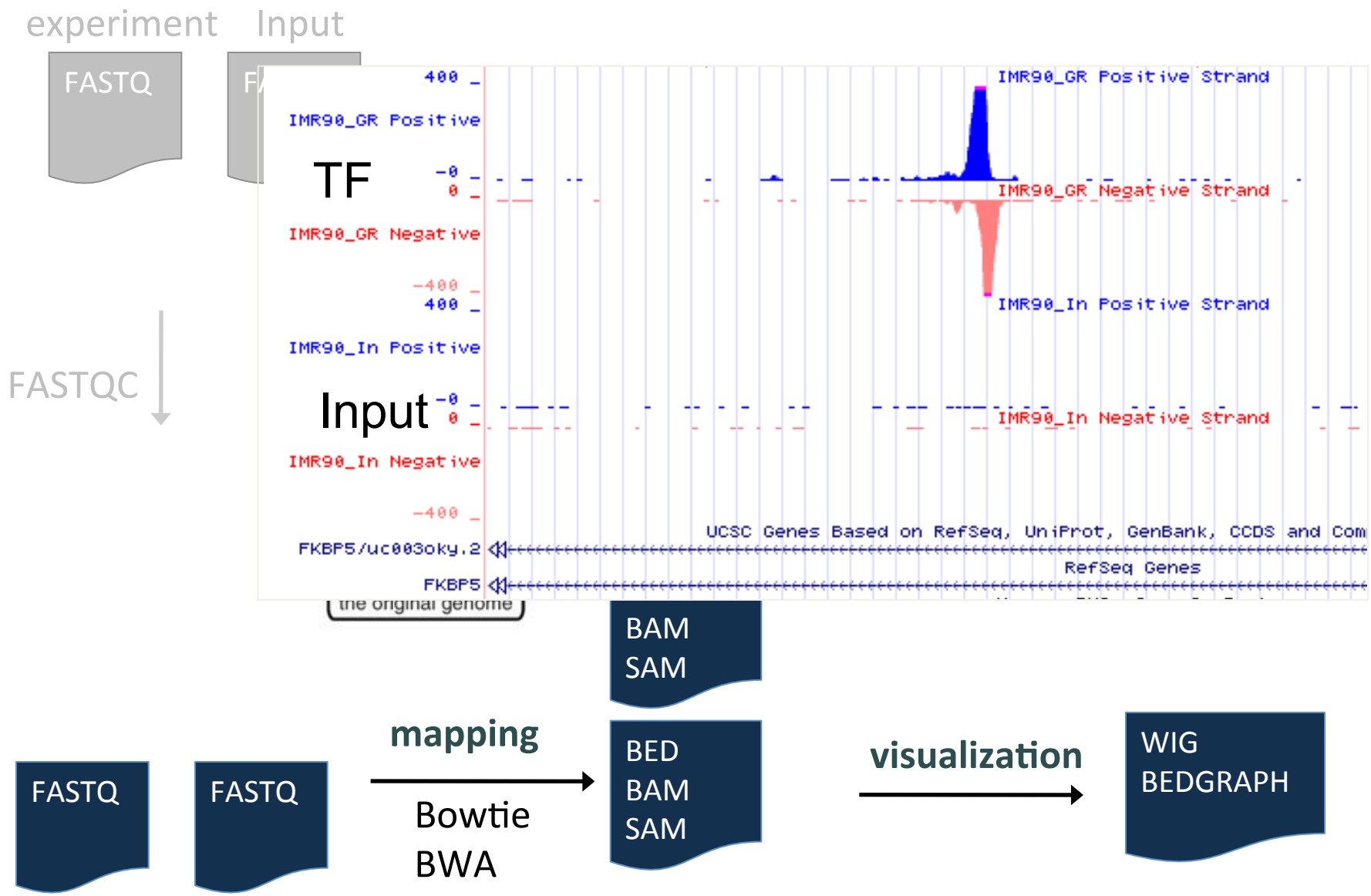
mapping

FASTQ FASTQ

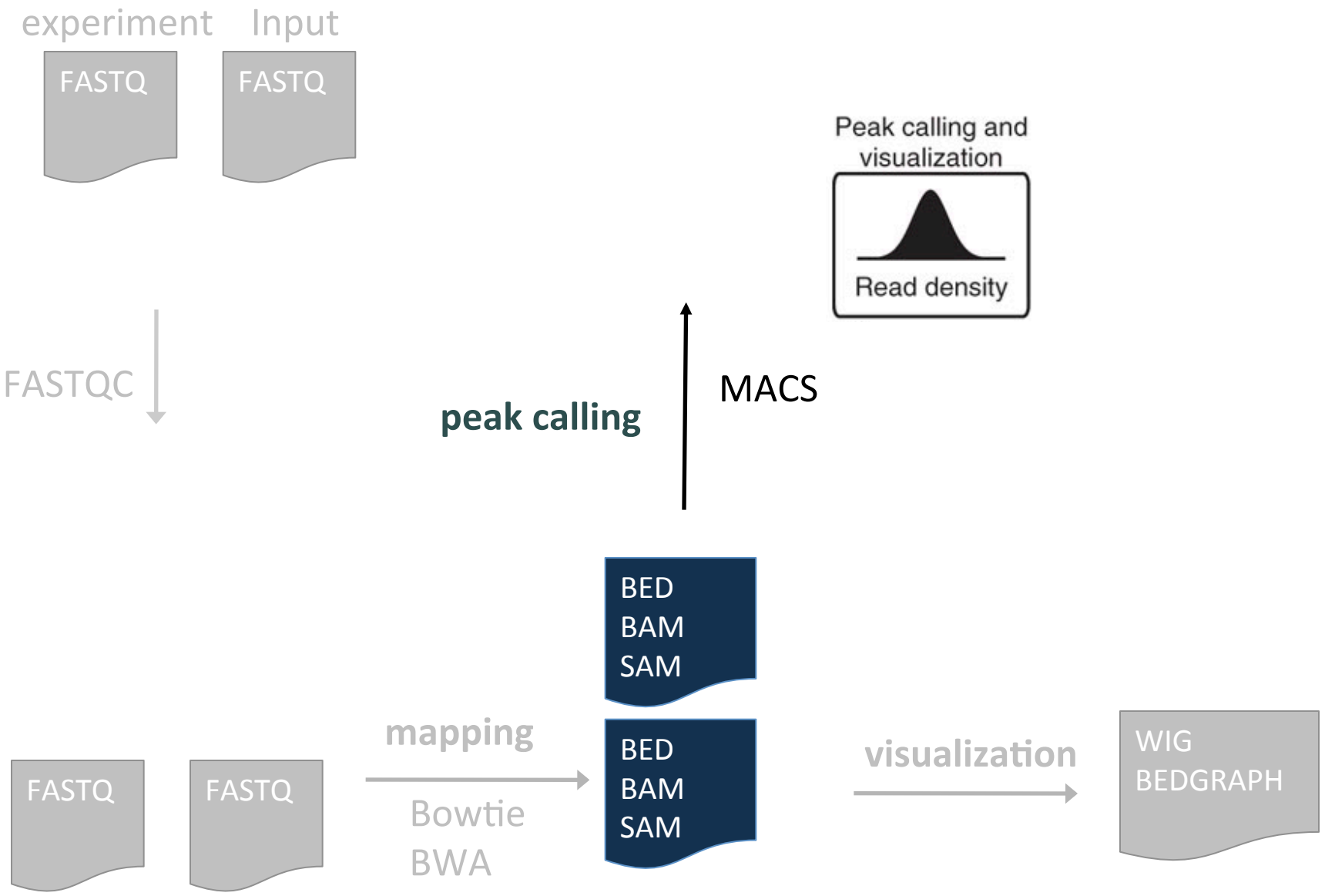
Bowtie
BWA

BED
BAM
SAM

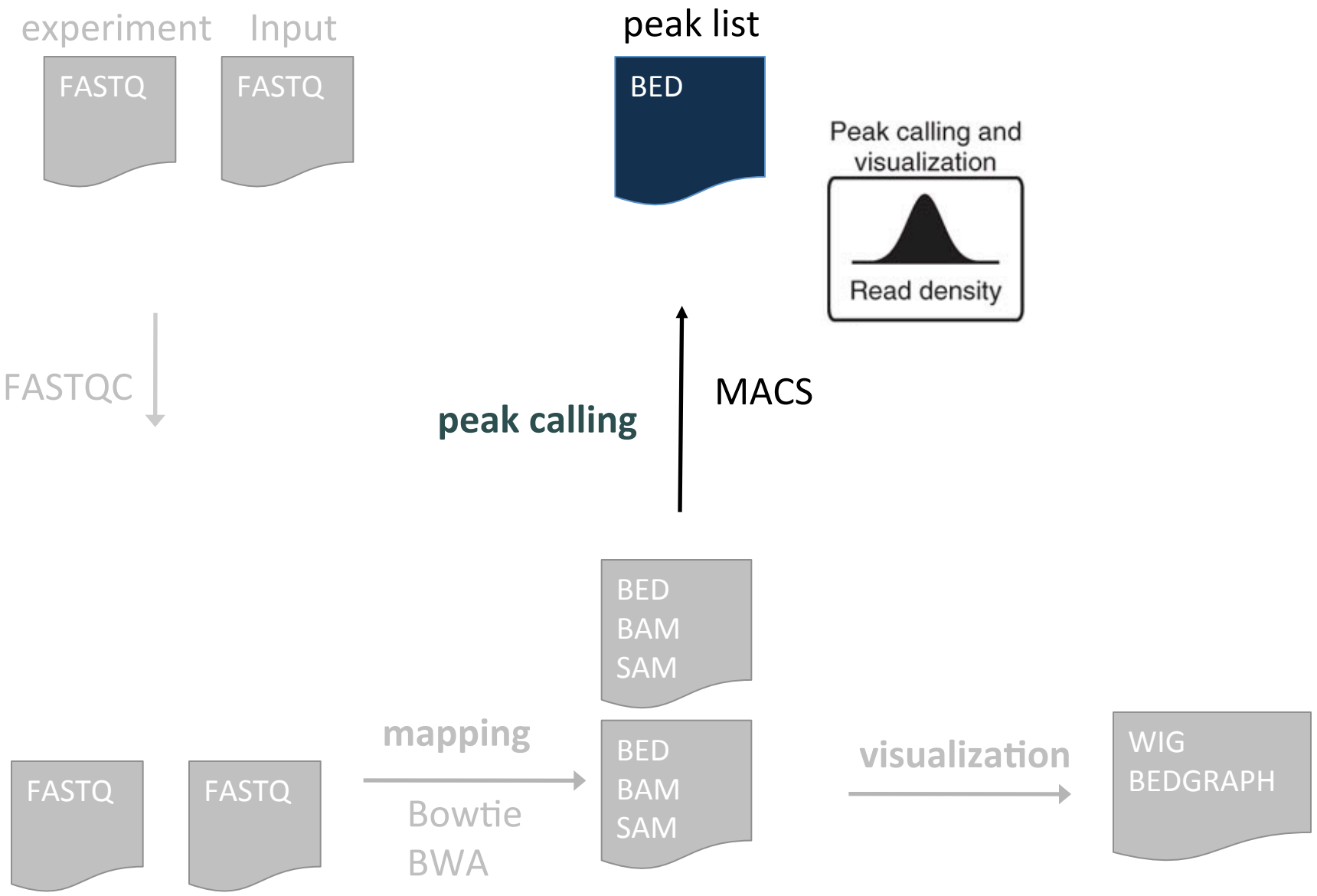
From sequence reads to peaks

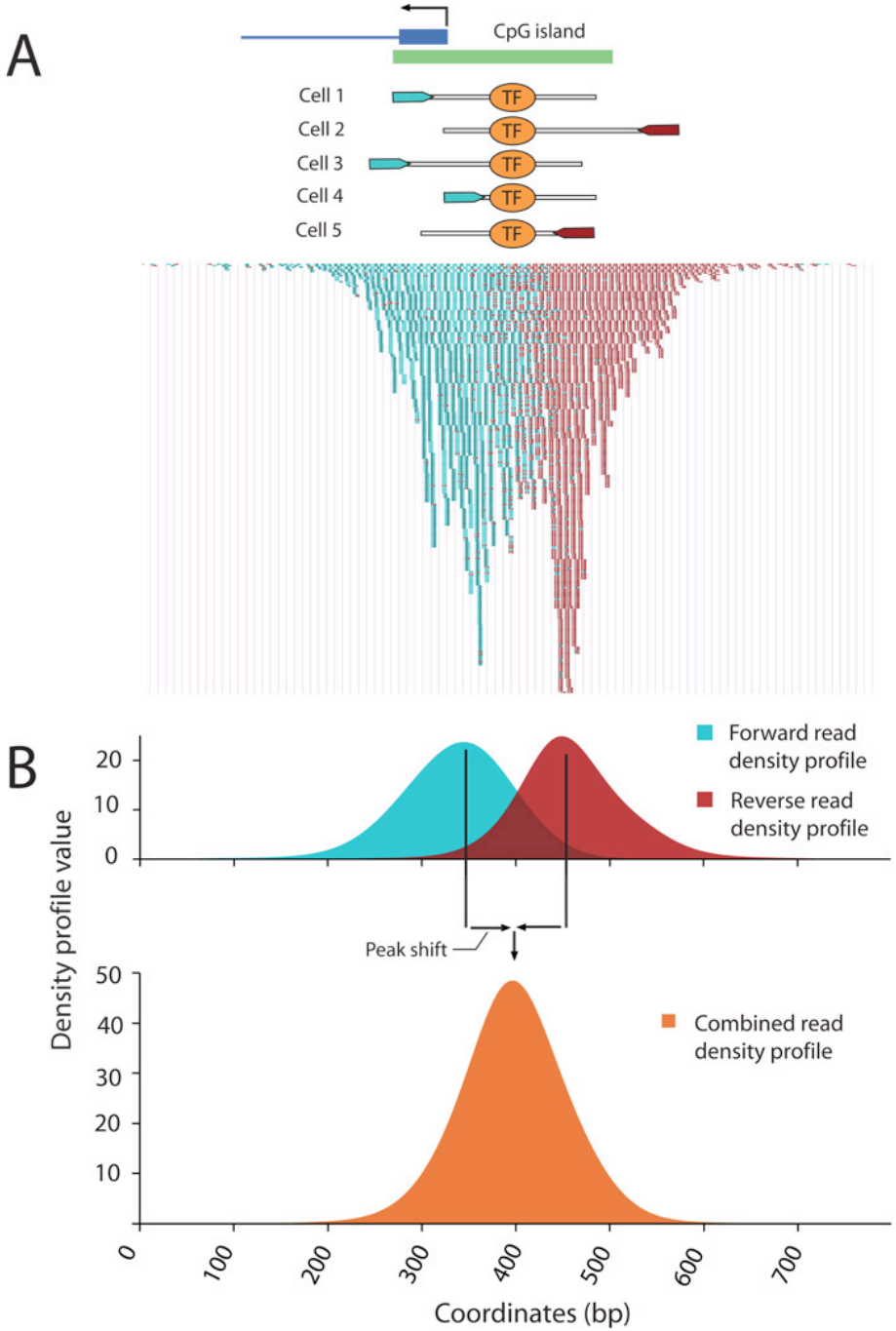


From sequence reads to peaks

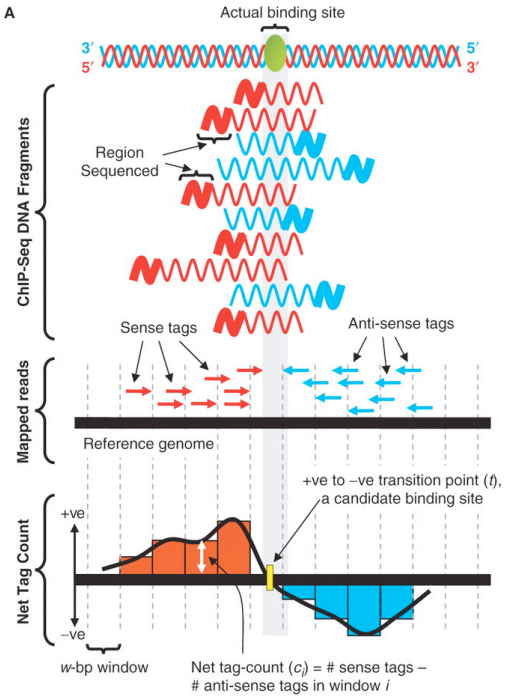


From sequence reads to peaks





mapping

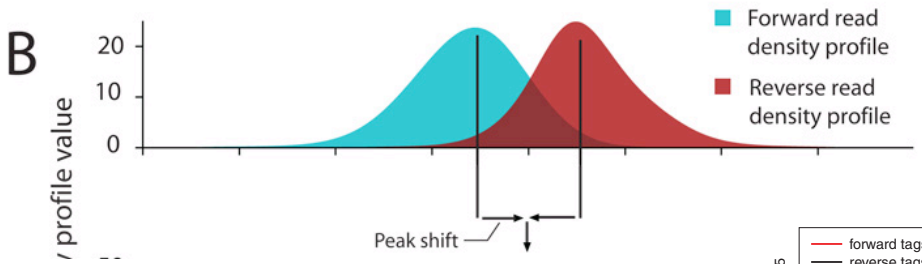


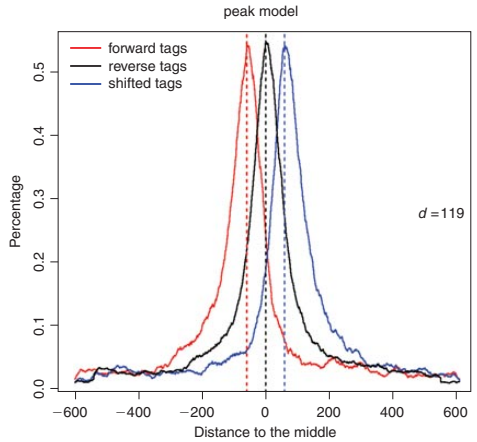
peak-calling

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

From sequence reads to peaks

bimodal enrichment pattern





Two steps strategy :

1 – modelling the read shift size

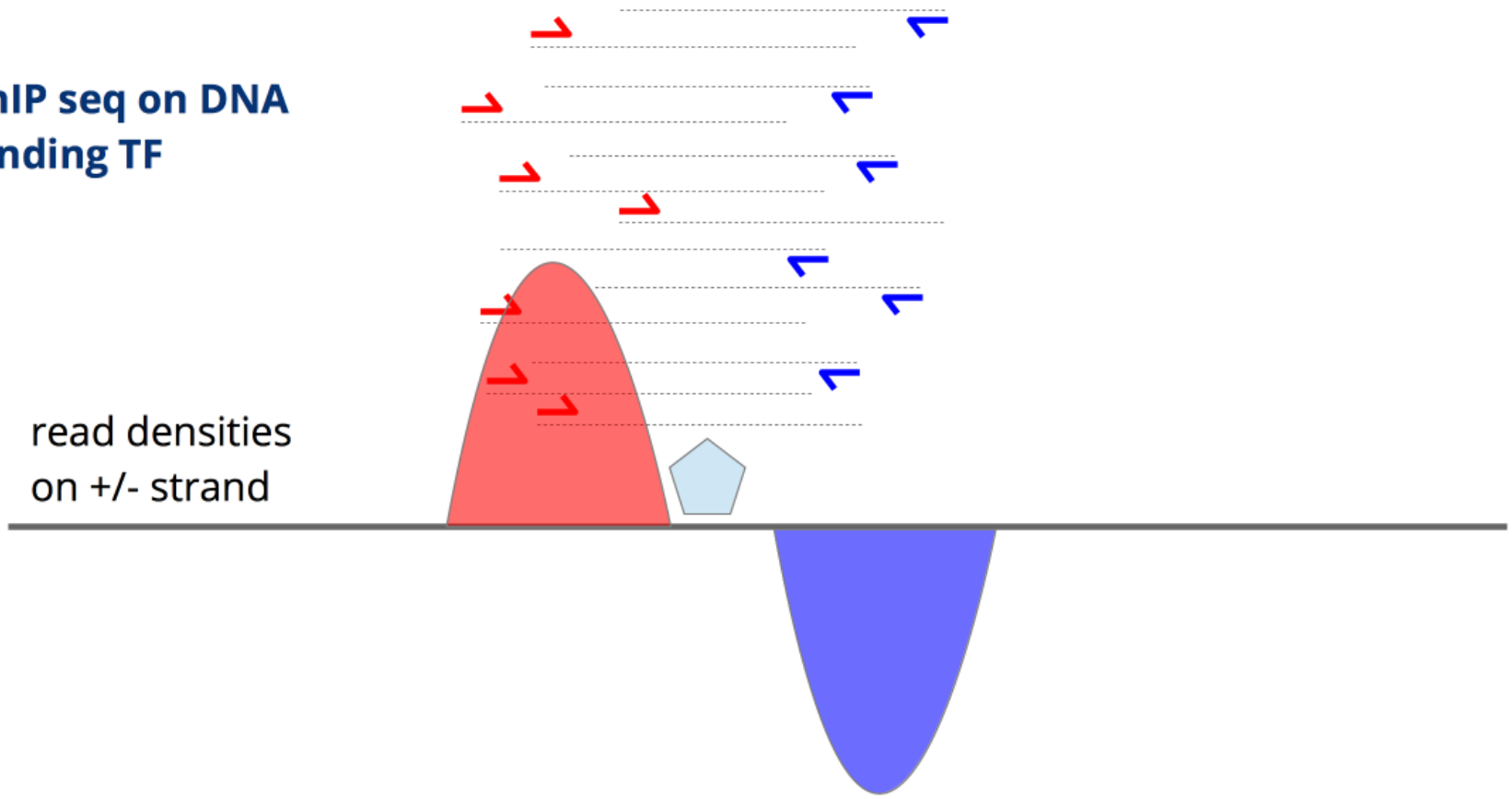
2 – peak calling

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

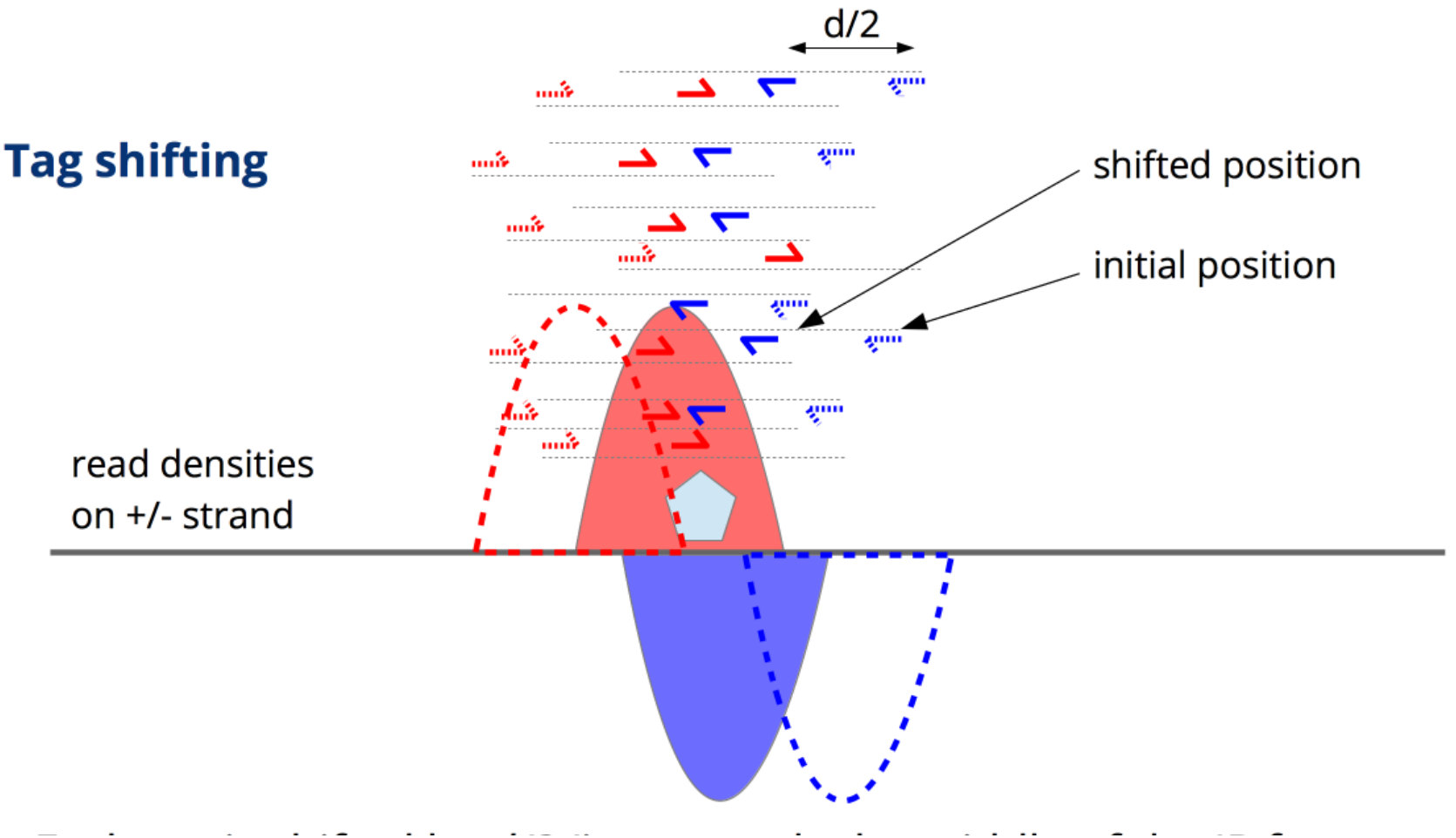
Feng, J., Liu, T., & Zhang, Y. (2011). *Using MACS to Identify Peaks from ChIP-Seq Data, Current Protocols in Bioinformatics*

ChIP-seq signal for transcription factors



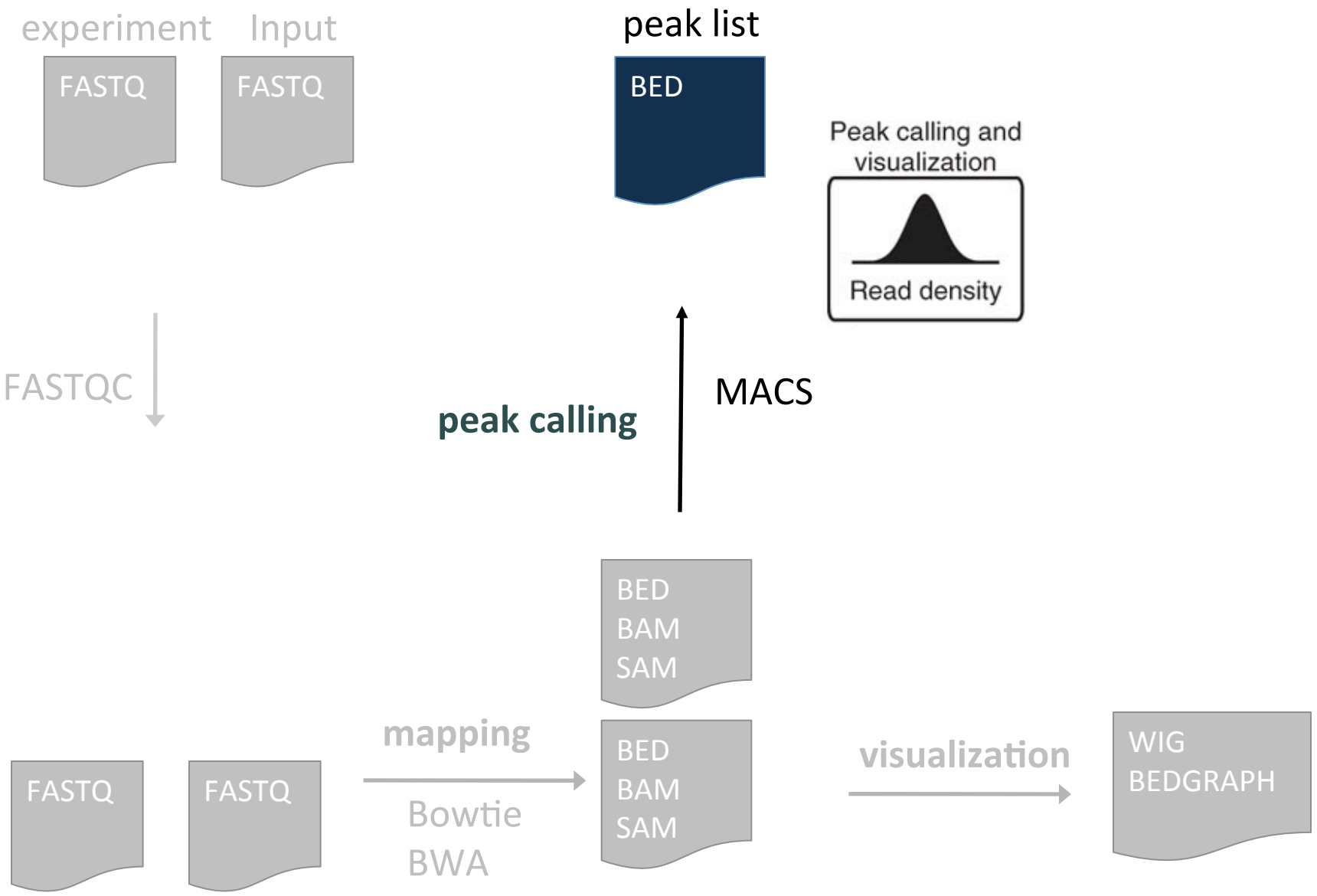
We expect to see a typical strand asymetry in read densities
→ ChIP peak recognition pattern

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 sequence reads to peaks



Peak list (BED file)

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	+

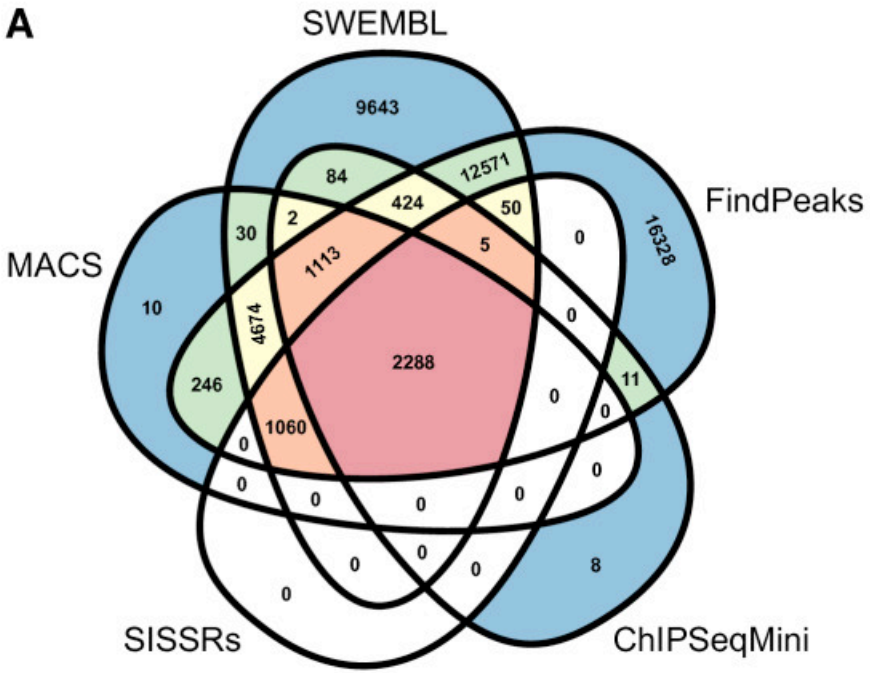
Read mapping programs

- Bowtie (Bowtie2)
- BWA (BWA2)
- STAR

- Generally not having a strong influence on the results
  - » Parameters: retain uniquely mapped reads

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

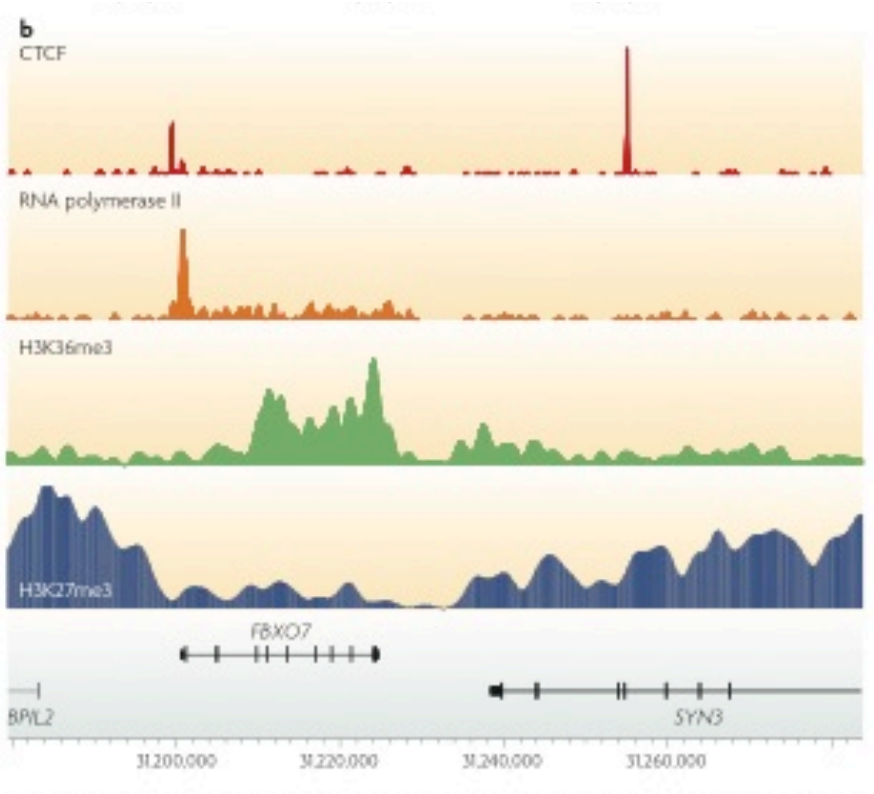


Mali Salmon-Divon *et al*, *BMC Bioinformatics, 2010*

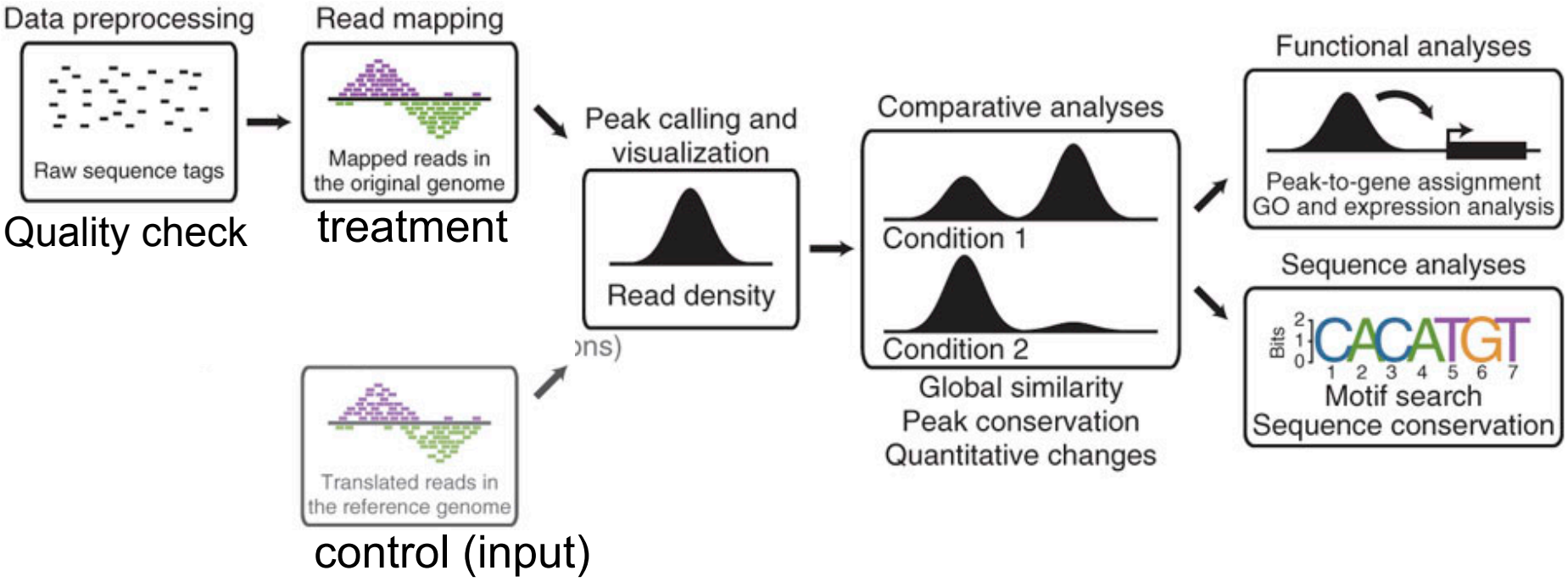
Peak calling programs

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

- Many new programs still developped !



Aim of the course : ChIP-seq analysis workflow

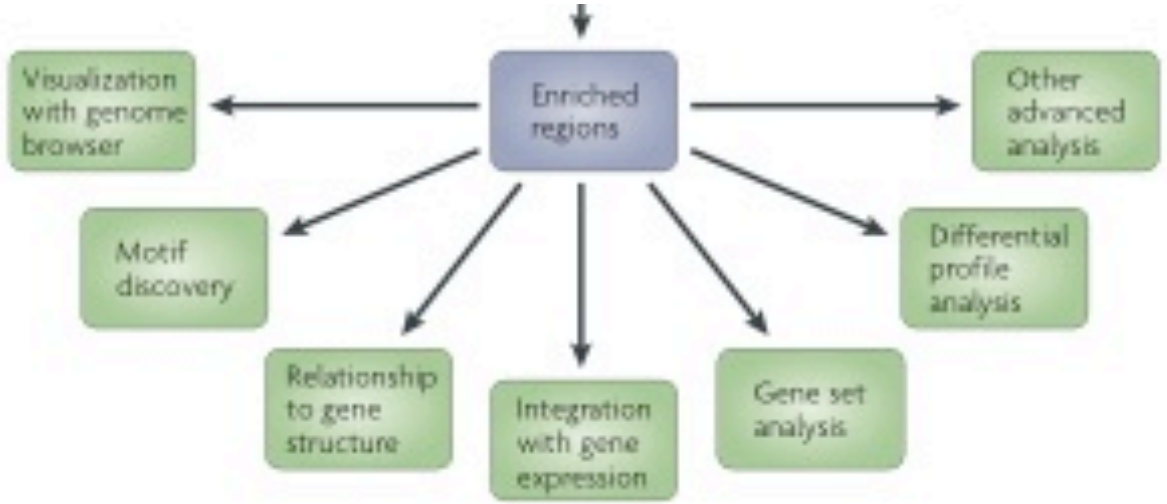


Processing steps

Downstream analyses

Adapted from Bardet et al, Nature Protocols, 2012

Aim of the course



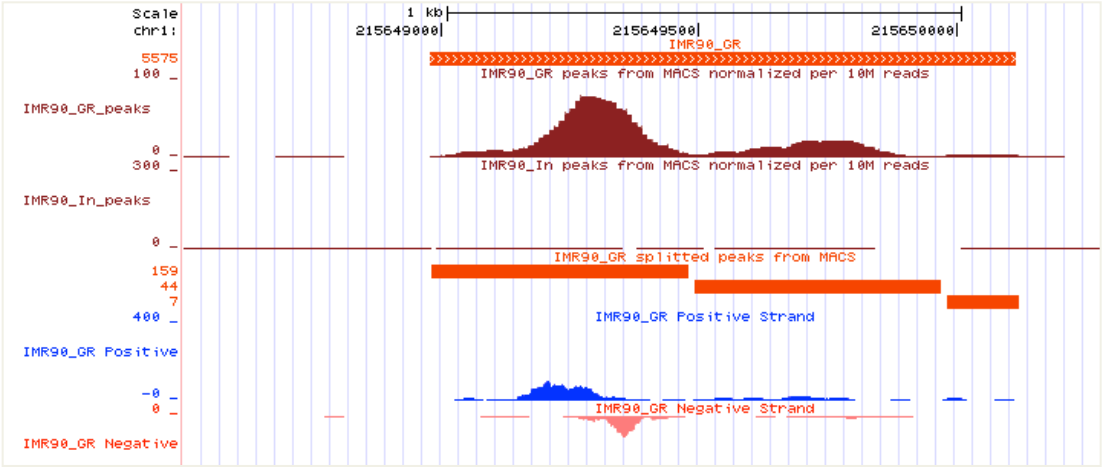
2 – Downstream analyses

- visualisation
- functional annotation of peaks
- motif discovery in peaks

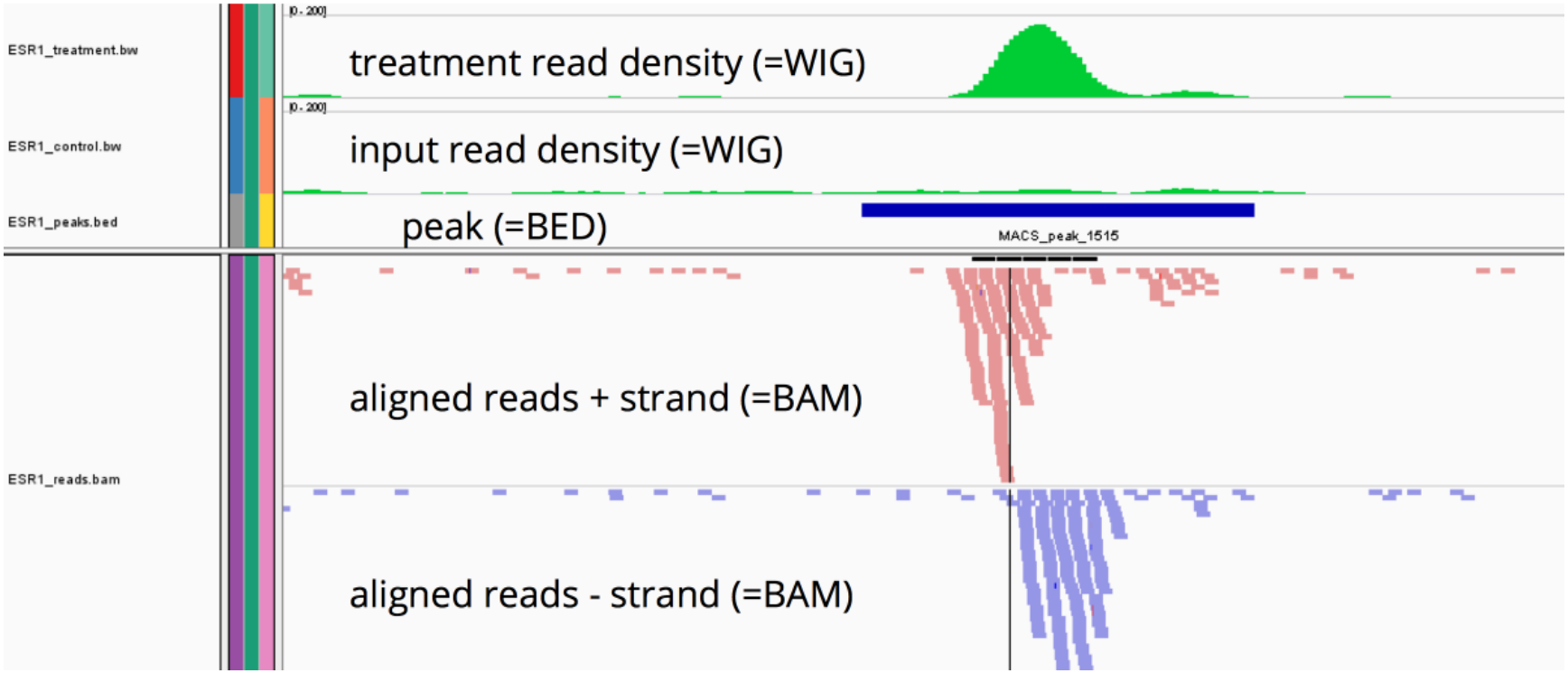
Visualizing in a genome browser

- 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

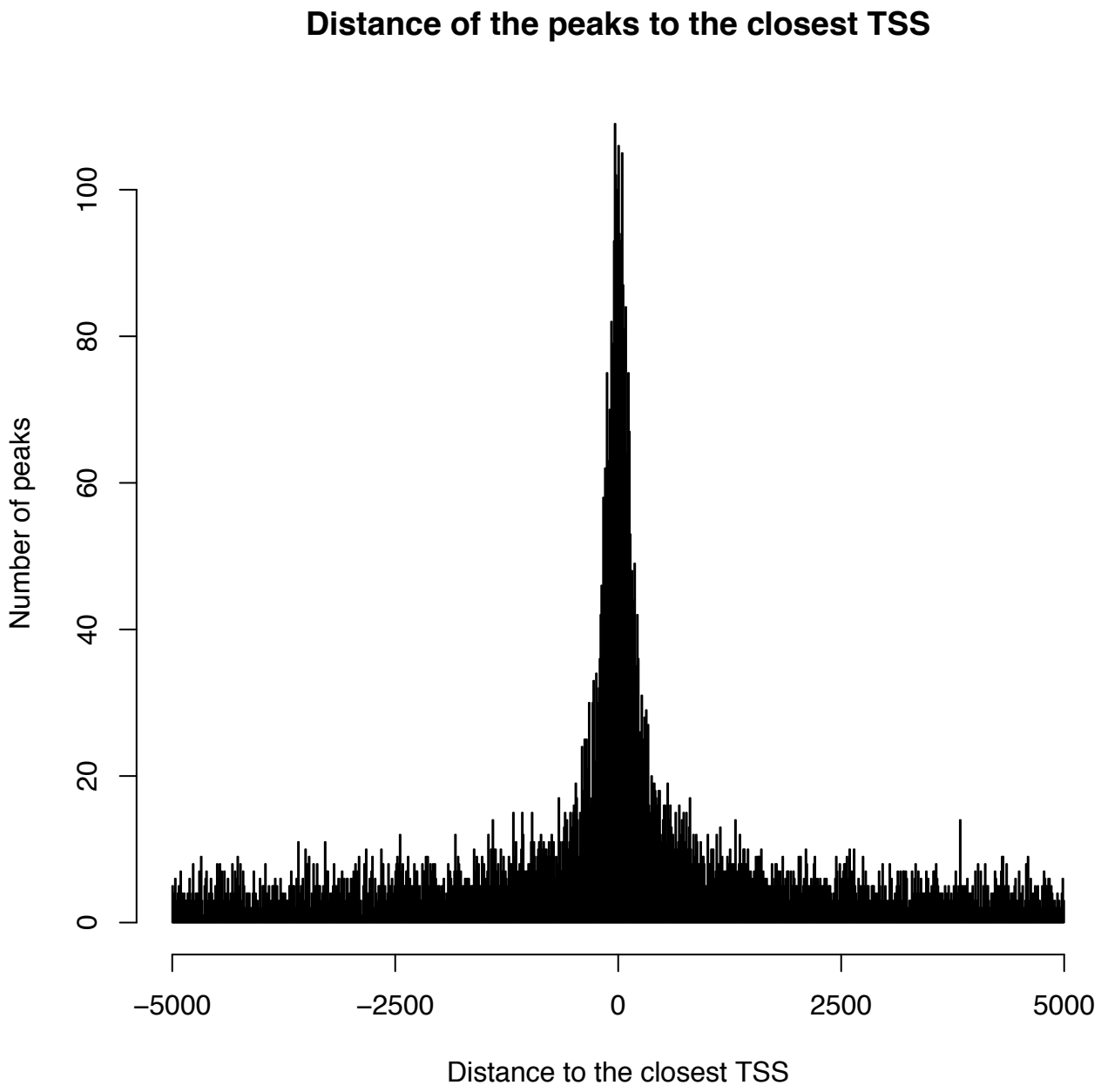
- File formats
  - » BED => simply defines a region (start-end)
  - » WIG, bedgraph => value assigned to each position



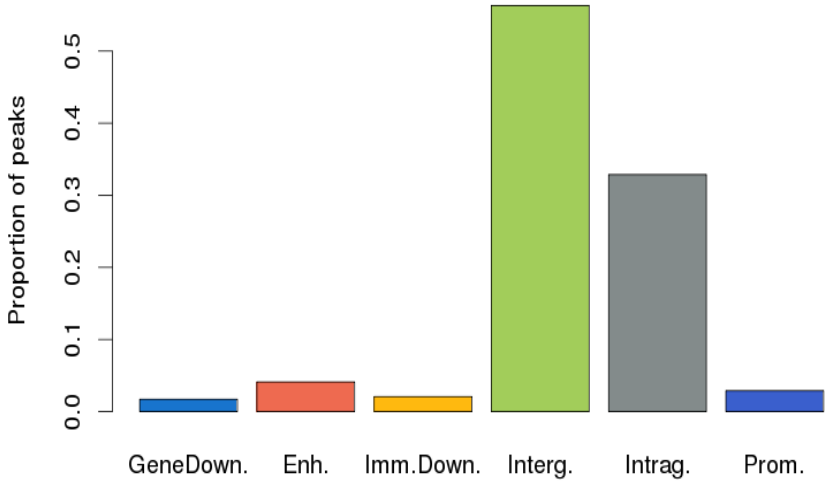
Vizualizing in IGV

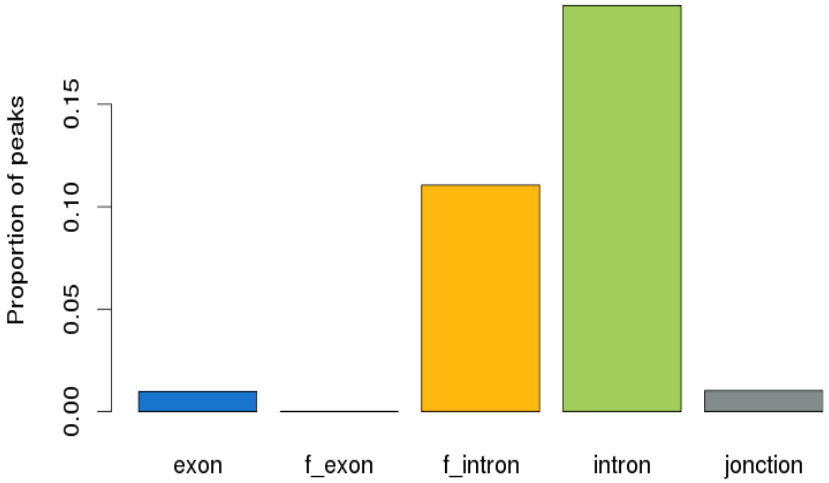


Distance to closest TSS



Localisation of the peaks in the genome





Genes → Regions ← Peaks

- **Idea** :
  - assign functional annotation to genomic regions
  - use statistics to avoid biases
- assign to each gene a regulatory domain
  - basal (-5kb/+1kb from TSS)
  - extended (up to nearest basal region ; max 1Mb)

- each domain is annotated to the functional terms of the corresponding gene
  → **"Functional domains"**

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Genes → Regions ← Peaks

*Given that **60%** of the genome is annotated to A, would I randomly expect 3 or more peaks to fall into region A ?*

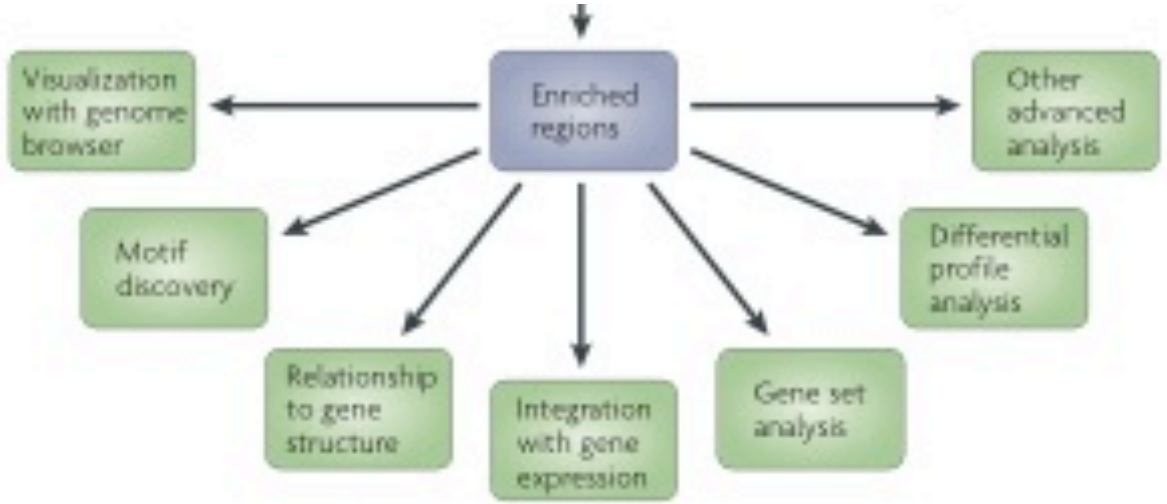
p > 0.5

*Given that **15%** of the genome is annotated to B, would I randomly expect 3 or more peaks to fall into region B ?*

p = 0.07

"GREAT improves functional interpretation of cis-regulatory regions"
McLean et al. Nat. Biotech. (2010)

Aim of the course

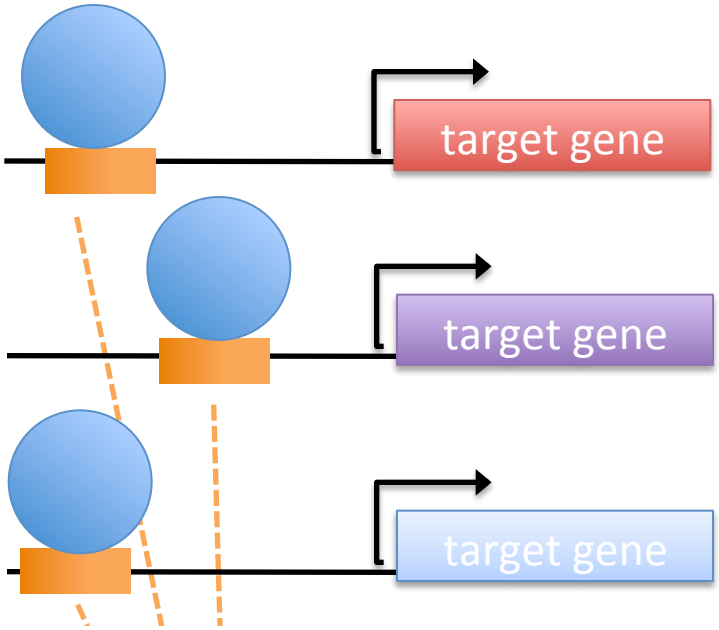


2 – Downstream analyses

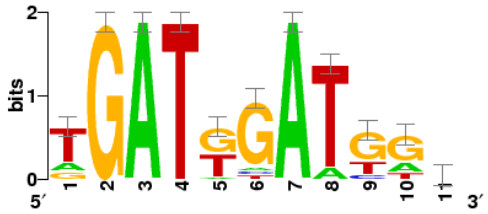
- visualisation
- functional annotation of peaks
- motif discovery in peaks

de novo motif discovery

transcription factor



cis-regulatory elements



binding motif

Problem :

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

de novo motif discovery

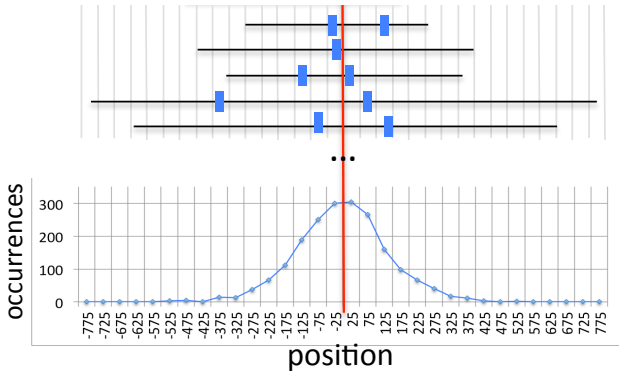
- Find exceptional motifs based on the sequence only

(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**)



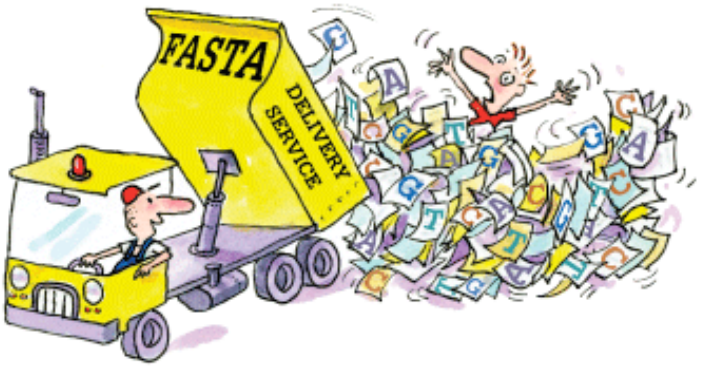
de novo motif discovery

- 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

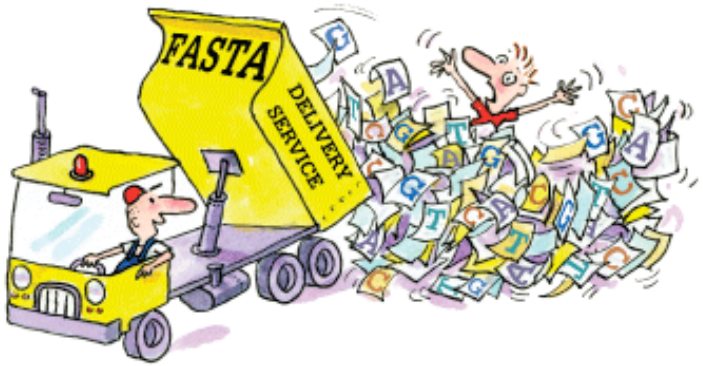


  - dozens of thousands of peaks !!!!!!

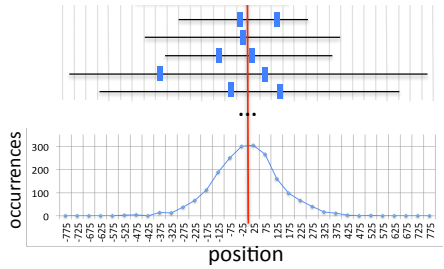
http://www.genomequest.com/landing-pages/ODI-webinar-web.html

New approaches for ChIP-seq datasets

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

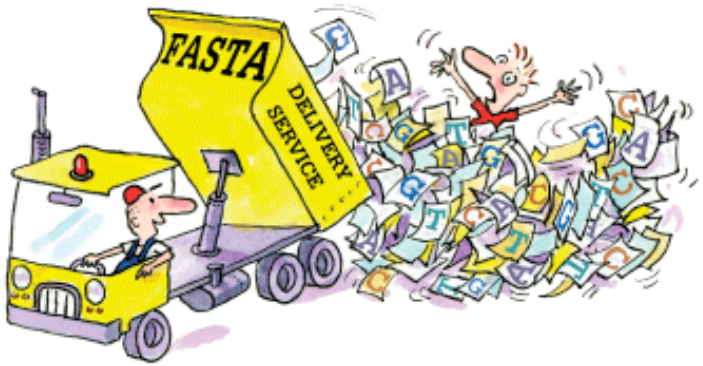


- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias

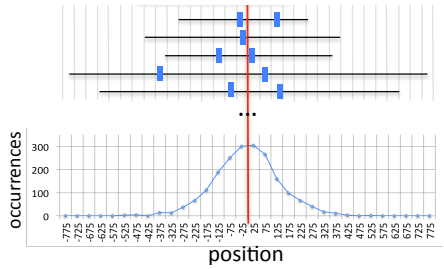


http://www.genomequest.com/landing-pages/ODI-webinar-web.html

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

Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	100kb (web site)	500kb (web site)	unrestricted, but motif discovery restricted to 600 peaks clipped to 100bp	motif discovery restricted to a few hundred base pairs	-
Stand-alone version	yes	yes	no	yes	yes	yes
Tasks						
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	100kb (web site)	500kb (web site)	unrestricted, but motif discovery restricted to 600 peaks clipped to 100bp	motif discovery restricted to a few hundred base pairs	-
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

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

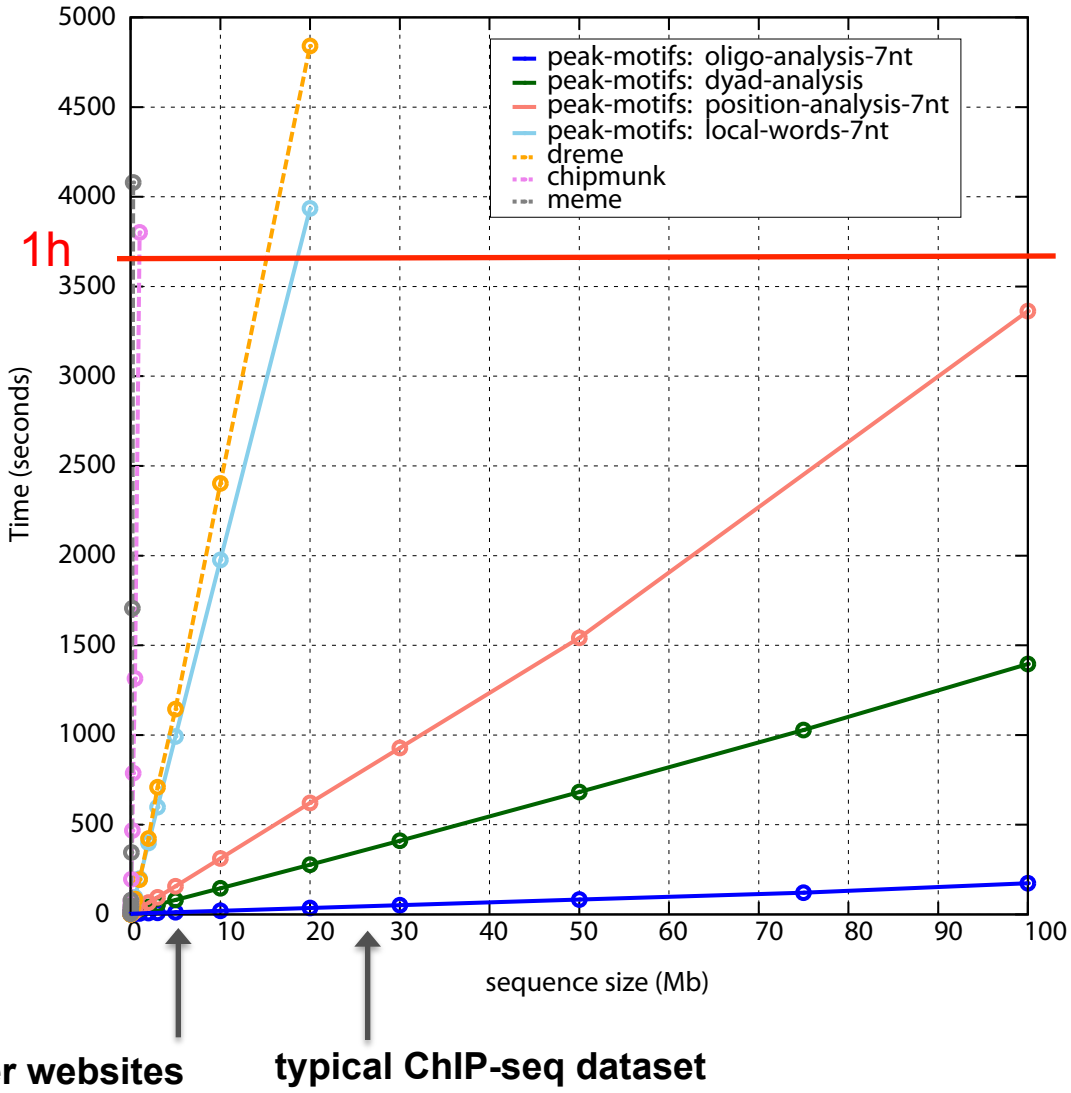
Comparison of tools for ChIP-seq

Program	peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	yes	yes	yes	yes	no	no
Size limitation	unrestricted (Web site tested with 22 Mb)	100kb (web site)	500kb (web site)	unrestricted, but motif discovery restricted to 600 peaks clipped to 100bp	motif discovery restricted to a few hundred base pairs	-
peak finding	no	no	no	no	yes	no
annotation of peak-flanking genes	no	no	yes	no		no
sequence composition (mono- and di-nucleotides)	yes	no	no	no		no
motif discovery	yes	yes	yes	yes	yes	yes
enrichment in motifs from databases	no	no	yes	yes		no
enrichment in discovered motifs	yes	no	no	no		no
peak scoring	no	no	no	yes	yes	no
motif clustering	no	no	no	no		yes
comparison discovered motifs / motif DB	yes	no	no	yes		yes
sequence scanning for site prediction	yes	no	no	yes		no
positional distribution of sites inside peaks	yes	no	yes	no		yes
visualization in genome browsers	yes	no	yes	no		no
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT position-analysis RSAT local-word-analysis + in stand-alone version: MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	no	patser	MAST + AME (enrichment)		no
Motif comparison algorithm	RSAT compare-motifs	no	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	yes	no	yes	no		yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	no	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		no
Motif sizes	variable (multiple word assembly)	user-specified	<=25 for MEME <=12 for Weeder <=13 for ChipMunk			predefined ranges (small, medium, large, extra-large)
Multiple motifs	yes	no	yes	yes		yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

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

RSAT peak-motifs

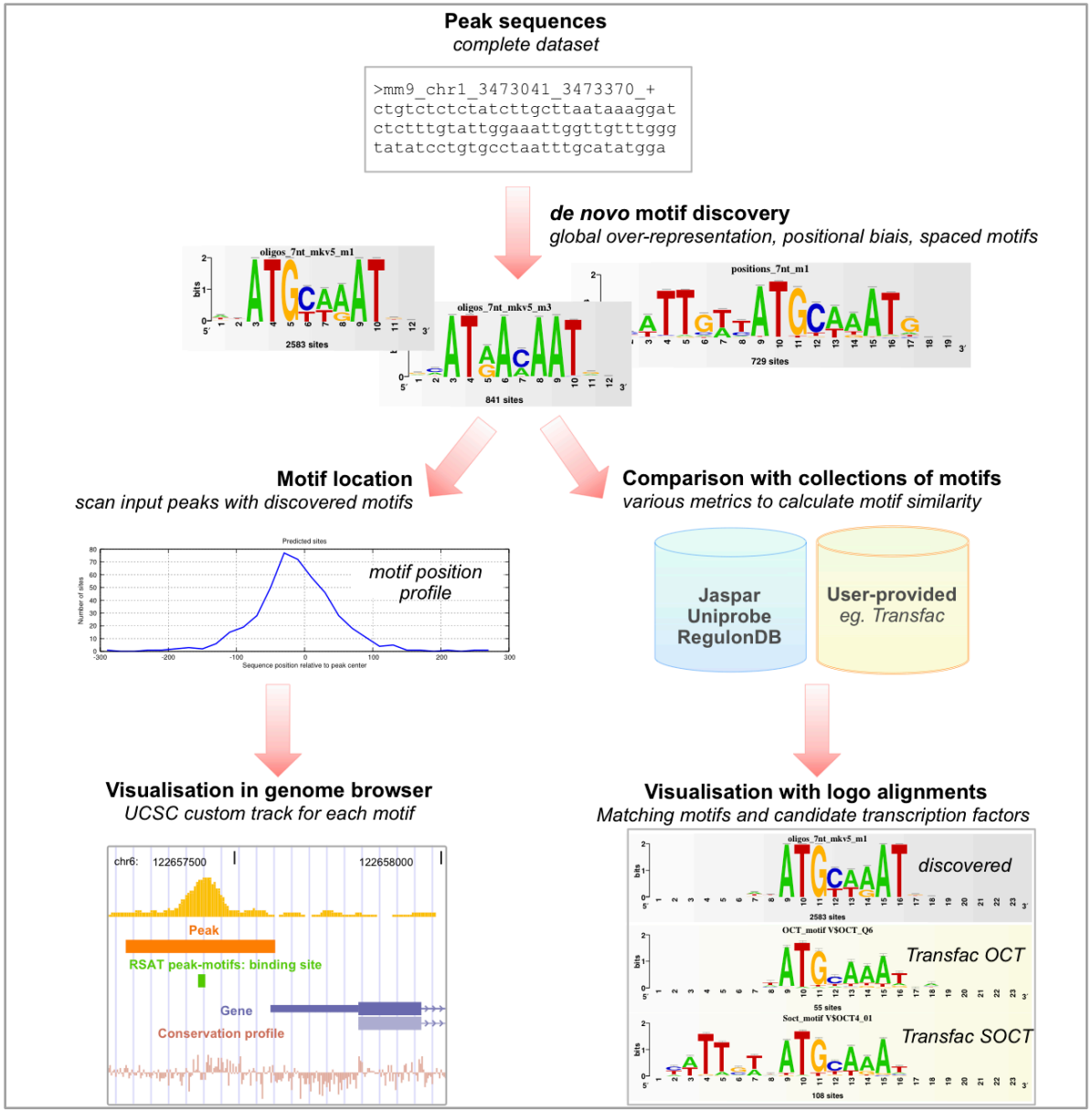
- fast and scalable
- treat full-size datasets



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*

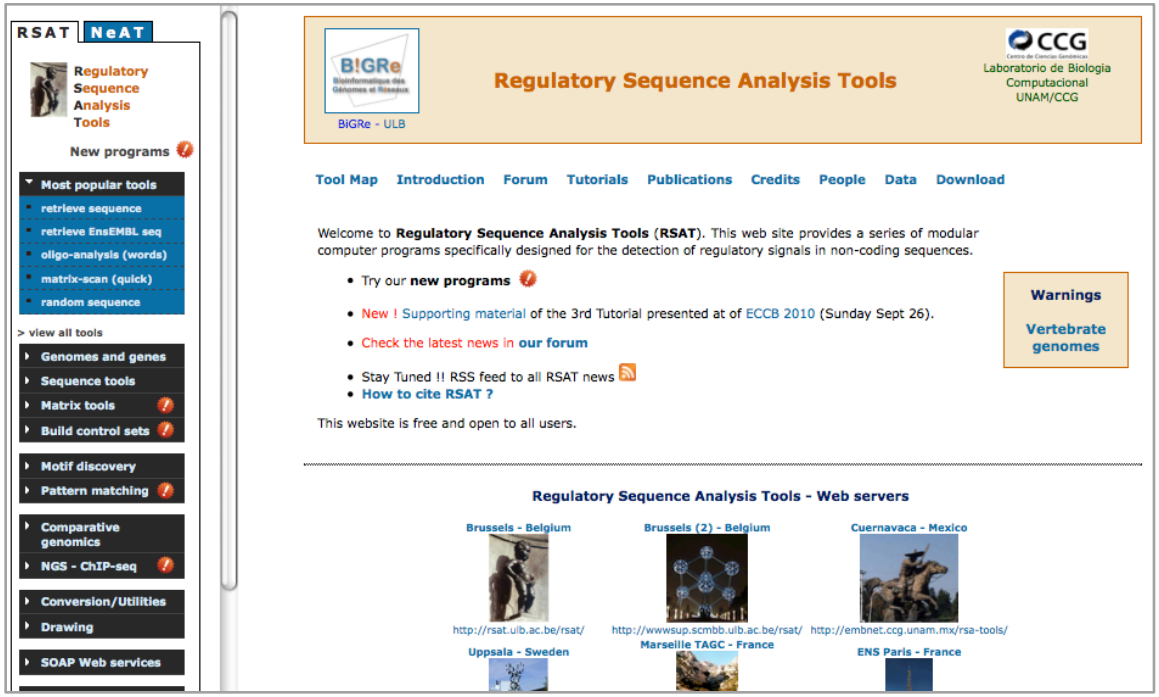
RSAT peak-motifs

- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface

within 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*

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**

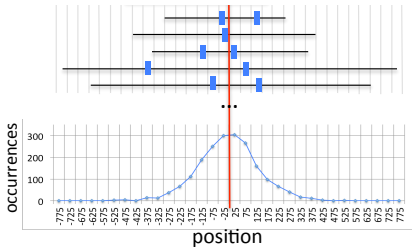
Output ◉ display ○ email

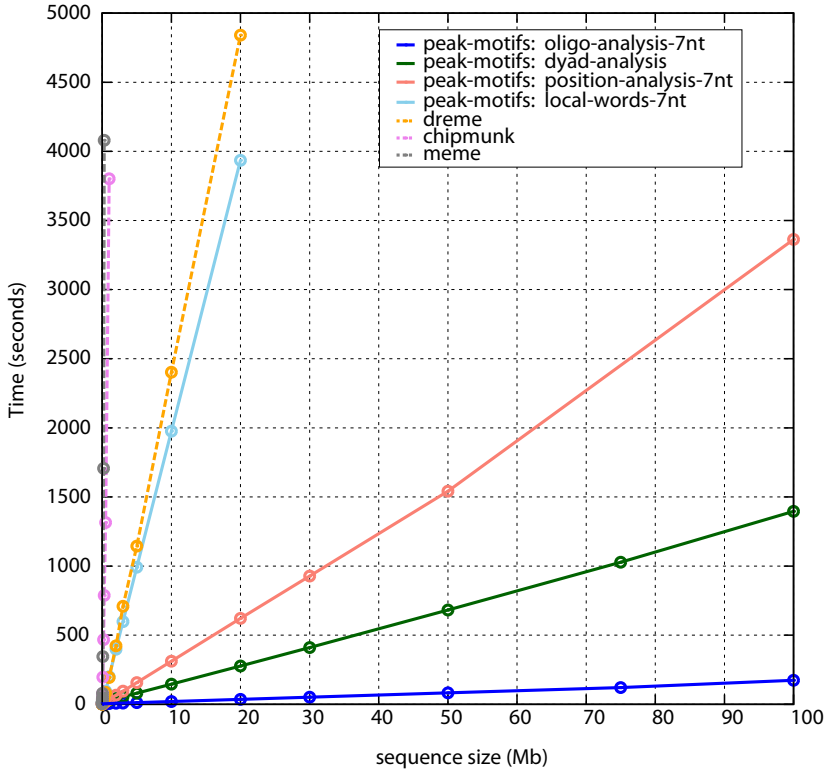
Note: email output is preferred for very large datasets or many comparisons with motifs collections

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

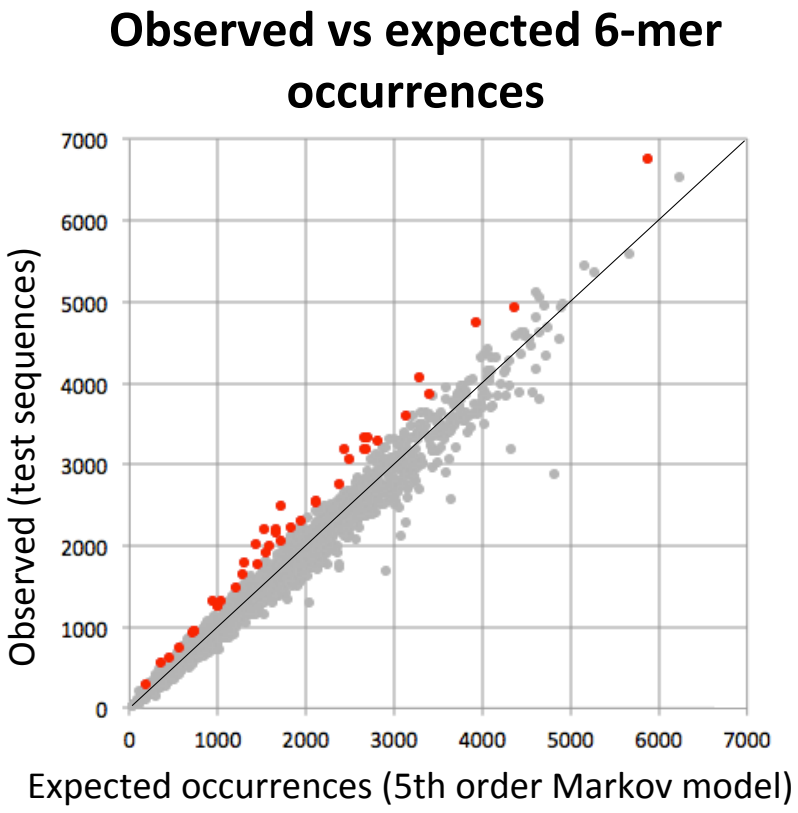




Motif discovery methods: frequency

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

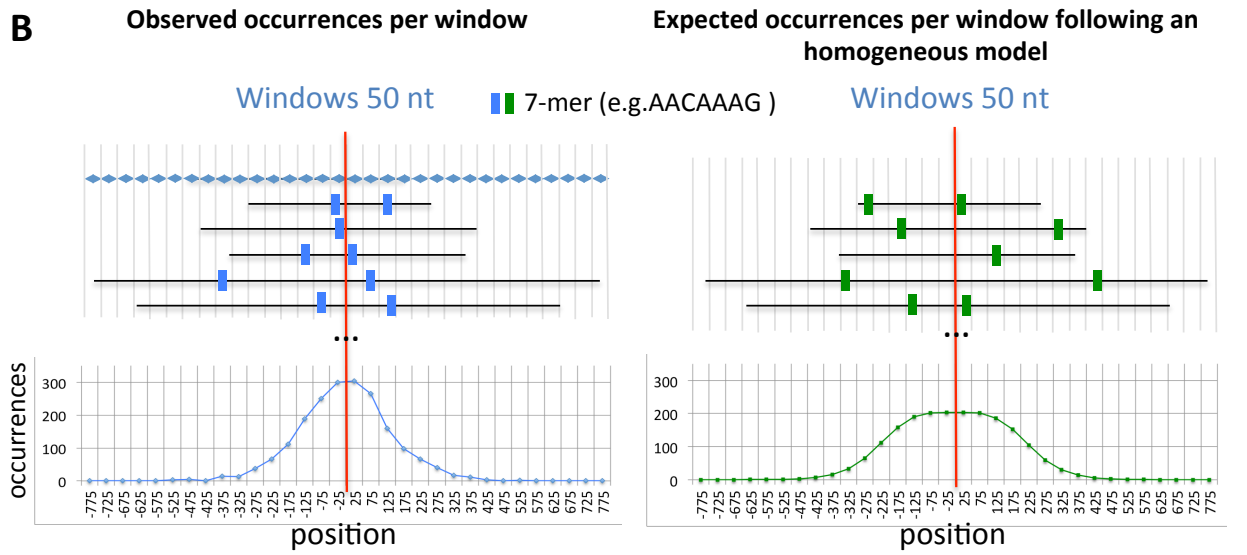
➔ Computation of p-value (binomial) and E-value (multi-testing correction)



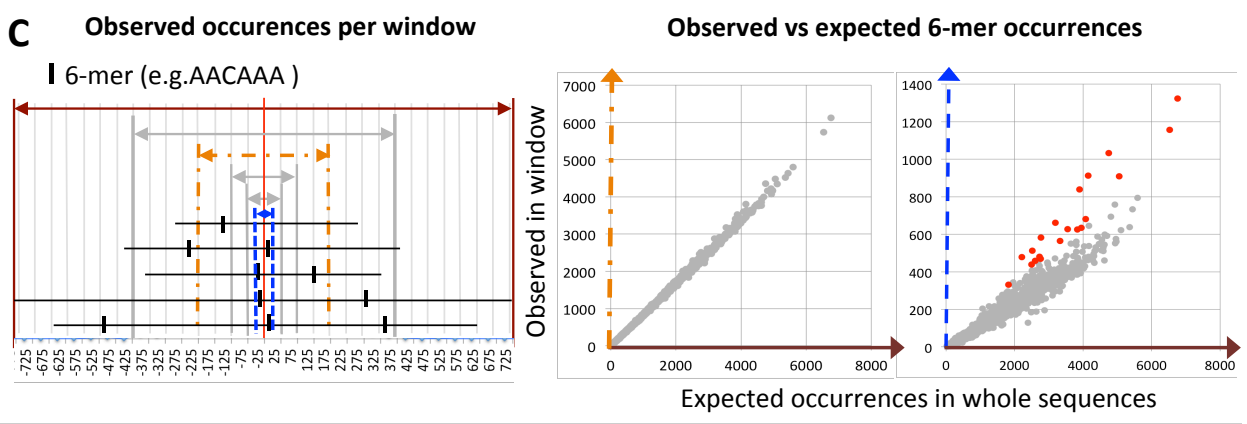
oligo-analysis
dyad-analysis (spaced motifs)

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

Motif discovery methods: positional bias



position-analysis

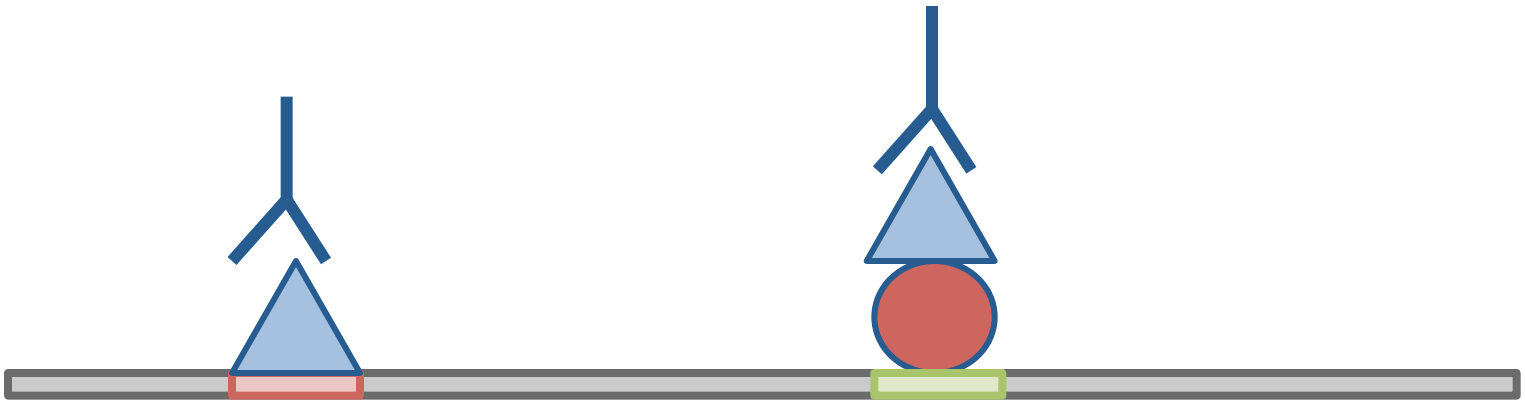


local-words

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

Direct versus indirect binding

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



Direct binding **Indirect binding**

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

To read further …

- **Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data**
  - » Tim Bailey – PLOS Computational Biology 9:11 2013

- **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)