Analyses Chip-Seq

Olivier Sand (CNRS)
Pierre Pericard (Université de Lille)

Cycle Analyse de données de séquençage à haut-débit - 14 et 15/06/2022

Slides mostly from...

Ecole de Bioinformatique Aviesan IFB Inserm (EBAII) 2021

Atelier ChIPseq

Elodie Darbo, INSERM U1218, Bordeaux
Stéphanie Le Gras, IGBMC, Strasbourg
Tao Ye, IGBMC, Strasbourg
Morgane Thomas-Chollier, IBENS, Paris

Contents

Get connected to the galaxy server

URL : https://usegalaxy.fr/join-training/bilille-2022-chipseq/

Authentification et enregistrement => your login and password

TIPS

- **Keep track** of all tools you run. You can for example, create a text file in which you write every tool you run
- **Keep track** of (non-default) parameters you use
- Give **content-explicit names** to the files you're generating
- Give files the **right extension**

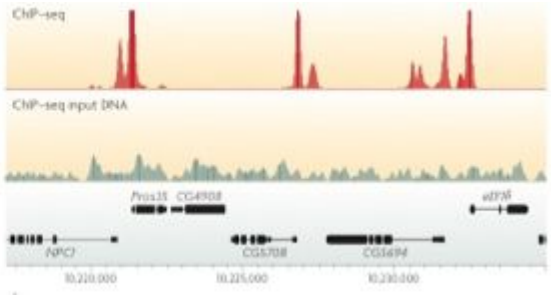
Introduction

ChIP-seq analysis

- Experimental design, Quality Controls, Mapping
- Normalization & peak calling

```
@SRR002012.1 Oct4:5:1:871:340
GGCGCACTTACACCCTACATCCATTG
+
IIIIG1?II;IIIII1IIII1%.I7I
```

?



Reads → **Peaks**

ChIP-seq analysis

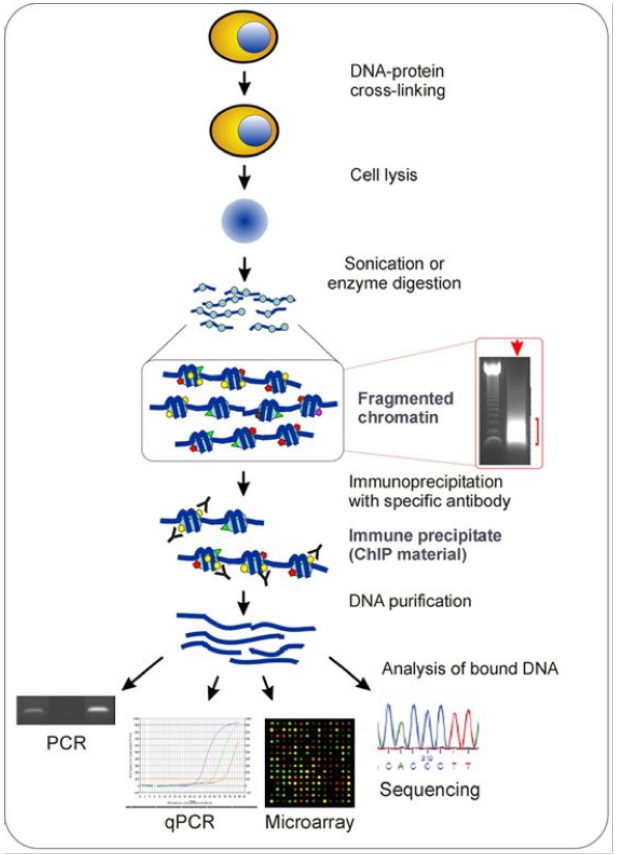
- Experimental design, Quality Controls, Mapping
- Normalization & peak calling
- Motif analysis
- Peak annotation

ChIP-seq

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**

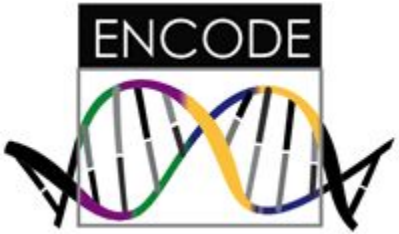


http://www.chip-antibodies.com/

Experimental design

ENCODE

- The Encyclopedia of DNA Elements (ENCODE) Consortium has carried out thousands of ChIP-seq experiments and has used this experience to develop a set of working standards and guidelines



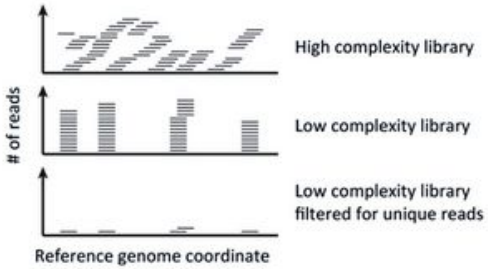
Landt SG, Marinov GK, Kundaje A *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813–1831.

See: https://www.encodeproject.org/about/experiment-guidelines/

Considerations on ChIP

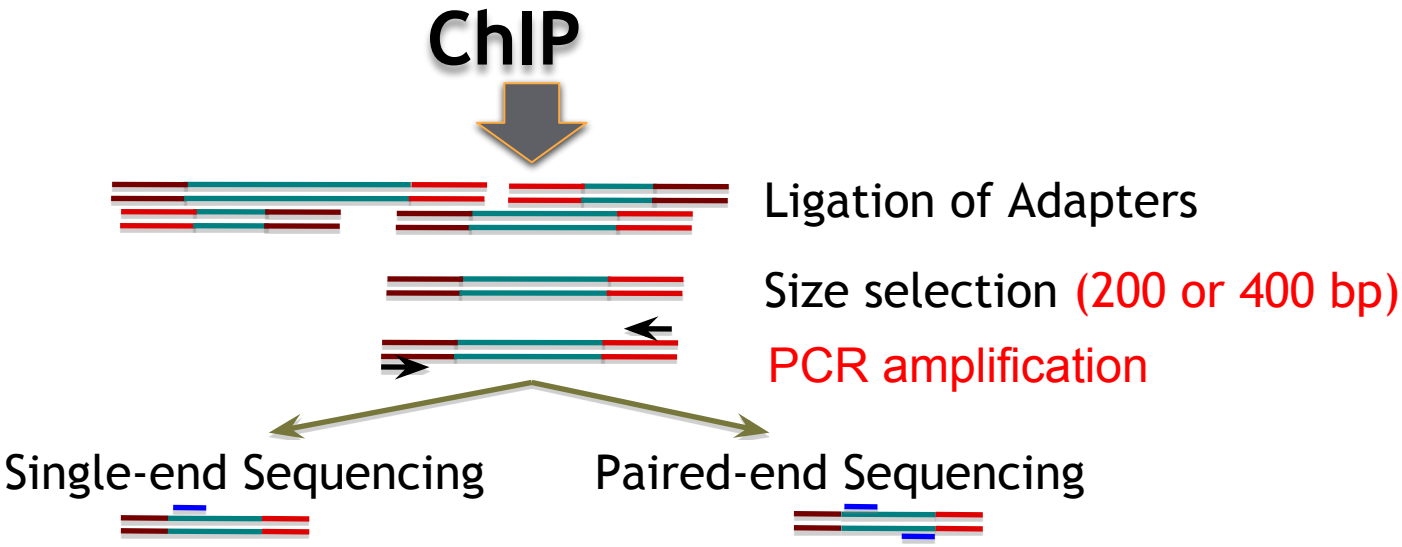
- Antibody
  - Antibody quality varies, even between independently prepared lots of the same antibody (Egelhofer, T. A. *et al*. 2011)
- Number of cells
  - large number of cells are required for a ChIP experiment (limitation for small organisms or precious samples)
- Shearing of DNA (Mnase I, sonication, Covaris): trying to narrow down the size distribution of DNA fragments

→ **Complexity in DNA fragments**



Library prep

- Step between ChIP and sequencing
- Starting material: ChIP sample (1-10ng of sheared DNA)



Sequencing

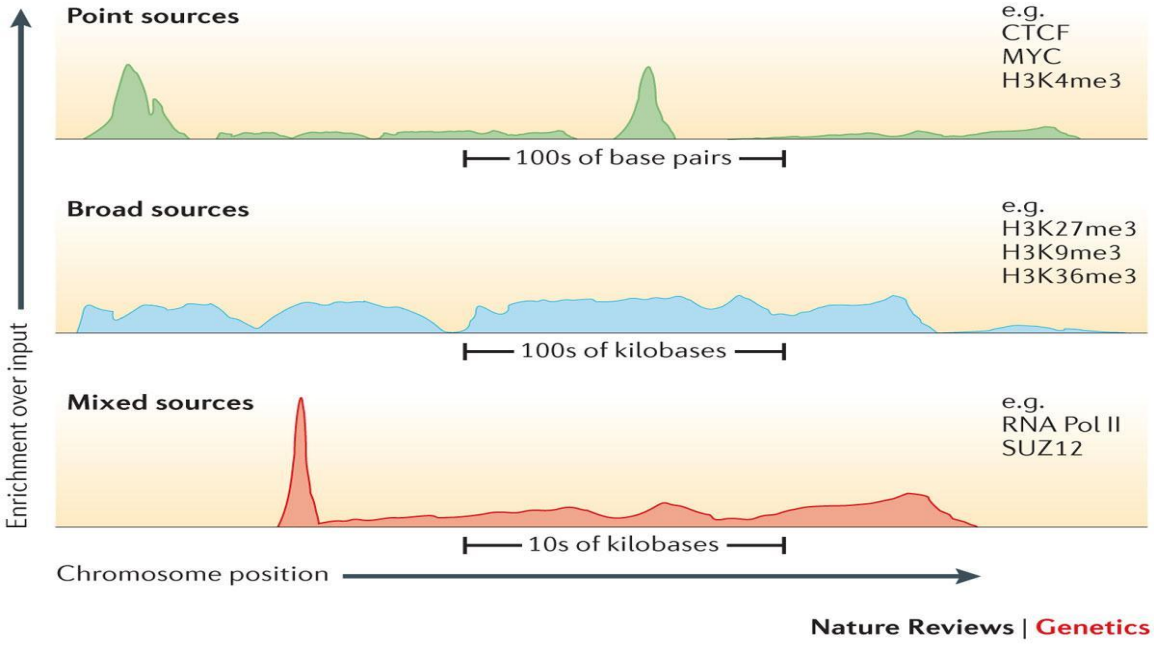
- Sequencer : Illumina HiSeq 4000
- No. of reads per sample:
  - (HiSeq 2500) 4 samples per lane :~41 millions per sample
  - (HiSeq 4000) 8 samples per lane :~43 millions per sample
- Length of DNA fragment : ~200bp
- No. of cycles per run : 50

Single end or paired end ?

- Single end (most of the time until 2016)
- Paired-end (more and more these days)
  - 🙂 Improve identification of duplicated reads
  - 🙂 Better estimation of the fragment size distribution
  - 🙂 Increase the mapping efficiency to **repeated regions**
  - ☹ The price! But 2 x 40bp is affordable

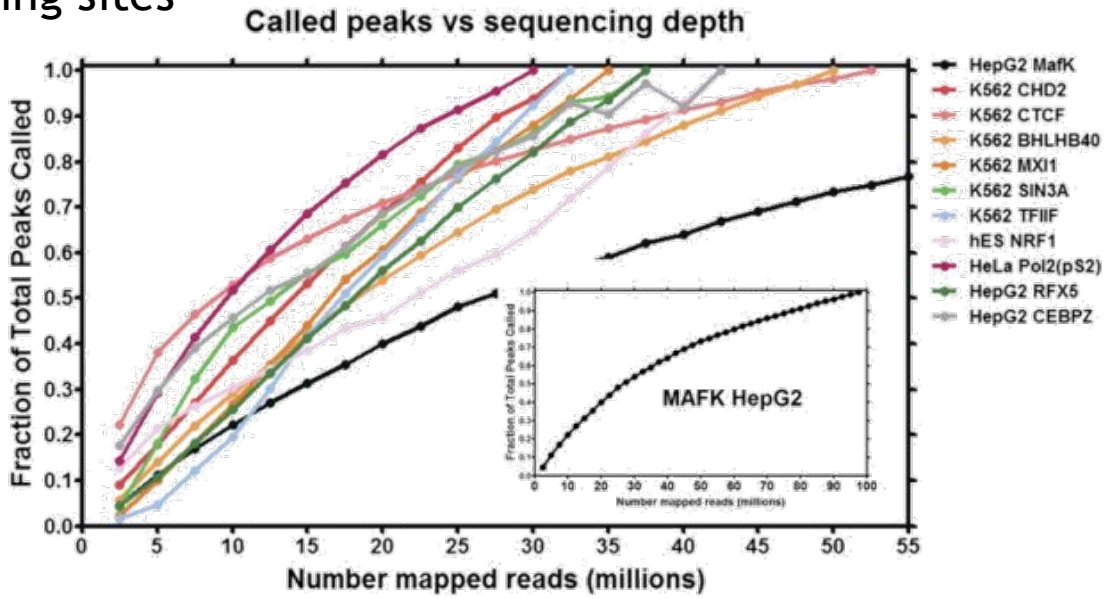
Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein



Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein
  - Number of expected binding sites



Landt et al, 2012

Sequencing depth

- Consider the depth needed depending on:
  - Chipped protein
  - Number of expected binding sites
  - Size of the genome of interest
    Ex:
    - For human genomes
      - 20 million uniquely mapped read sequences for point-source peaks
      - 40 million for broad-source peaks
    - For fly genome: 8 million reads
    - For worm genome: 10 million reads

Controls

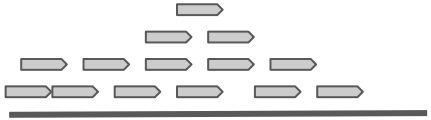
- Used mostly to filter out false positives (high level of noise)
  - Idea: potential false positive will be enriched in both treatment and control.
- A control will fail to filter out false positives if its enrichment profile is very different from the enrichment profile of false positive regions in the treatment sample
- Most commonly used control: Input DNA (a portion of DNA sample removed prior to IP)
- Choice of control is extremely important
- It is recommended to cover the control in a higher extent than the IPs

Why an Input is required ?

- The input is used to model **local noise level**
  - **Accessible regions** are expected to produce **more reads**

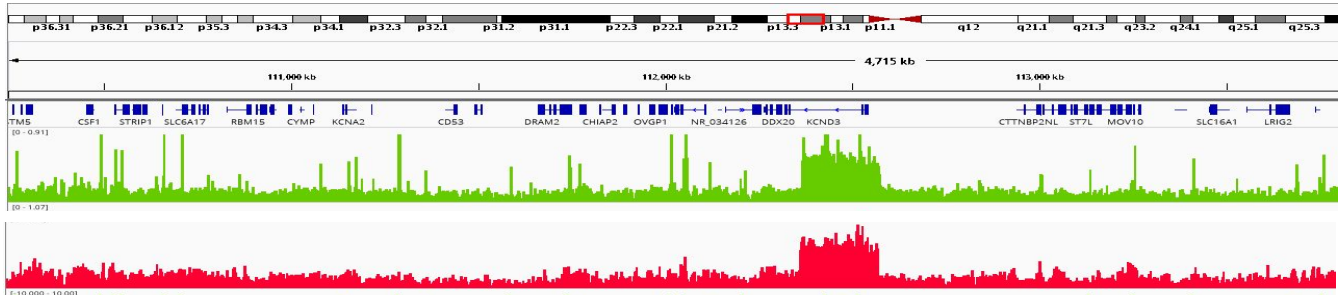
Closed | Open | Closed

Closed | Open | Closed



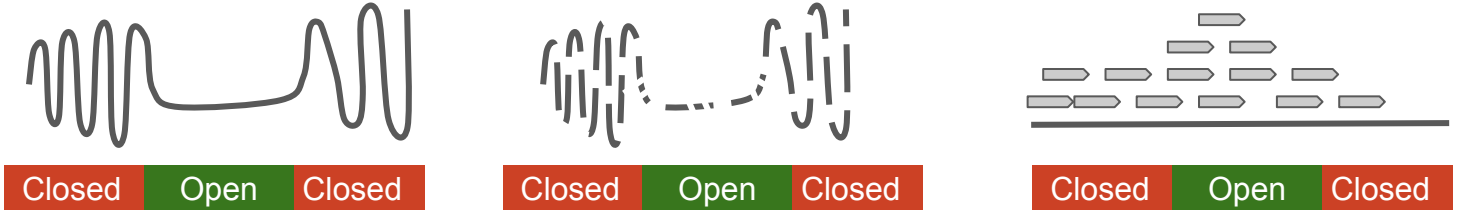
Closed | Open | Closed

  - **Amplified regions (CNV) are expected to produce more reads**



Why an Input is required ?

- The input is used to model **local noise level**
  - **Accessible regions** are expected to produce **more reads**



  - **Amplified regions (CNV) are expected to produce more reads**
  - **Moreover, most peak callers** are configured **with an input as control**

Other controls

- IgG (mock IP): controls for non-specific IP enrichment
  - Problem : low-complexity library (few reads)
- Histone H3 (for H3 variants)
- Uninduced condition (for inducible TFs)
  - Example : Glucocorticoid Réceptor
  - Induced by Dexamethasone (Dex)
  - Control vehicle = Ethanol (EthOH)
- KO of your protein of interest
- …

Replicates

- A minimum of **two** replicates should be carried out per experiment.
- Each replicate should be a **biological** rather than a technical replicate; that is, it results from an independent cell culture, embryo pool or tissue sample.

Data analysed in this course

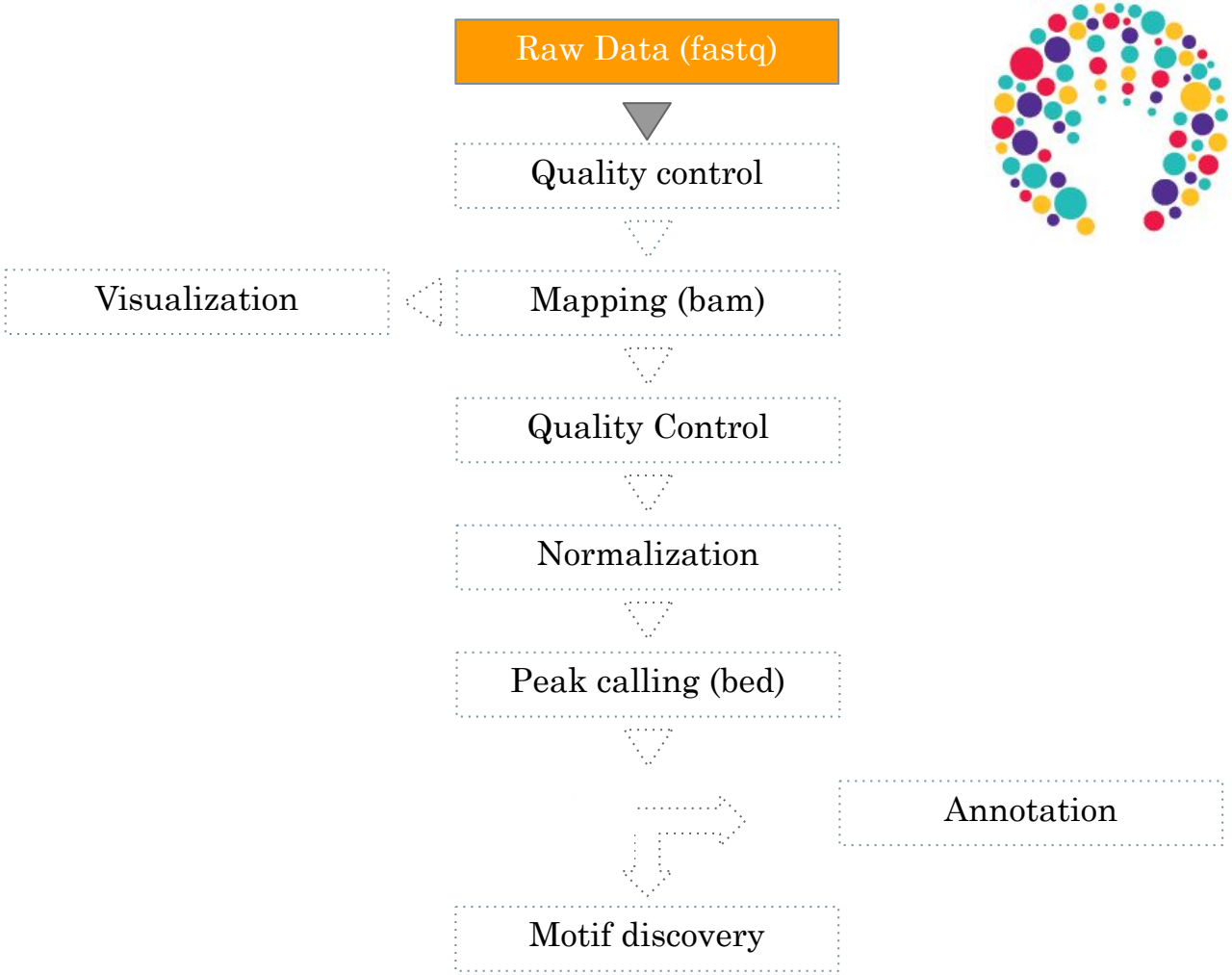
Dataset used

- Wang, C.-Y., T. Jégu, H.-P. Chu, H. J. Oh, and J. T. Lee, 2018 SMCHD1 Merges Chromosome Compartments and Assists Formation of Super-Structures on the Inactive X. Cell 174: 406–421.e25
- https://doi.org/10.1016/j.cell.2018.05.007
- Experiment: H3K27me3, H3K4me3 and CTCF binding
- Control: INPUT DNA

Practical/tutorial to follow

- Galaxy Training Network
- Epigenetics
- Formation of the Super-Structures on the Inactive X
- https://training.galaxyproject.org/training-material/topics/epigenetics/tutorials/formation_of_super-structures_on_xi/tutorial.html

Protocol


Raw Data (fastq)
Quality control
Visualization
Mapping (bam)
Quality Control
Normalization
Peak calling (bed)
Annotation
Motif discovery

Protocol

- Importing ChIP-seq data from data libraries
- [Step 1] Hands-on : Import the data

- Files are in *Données partagées / Shared data*

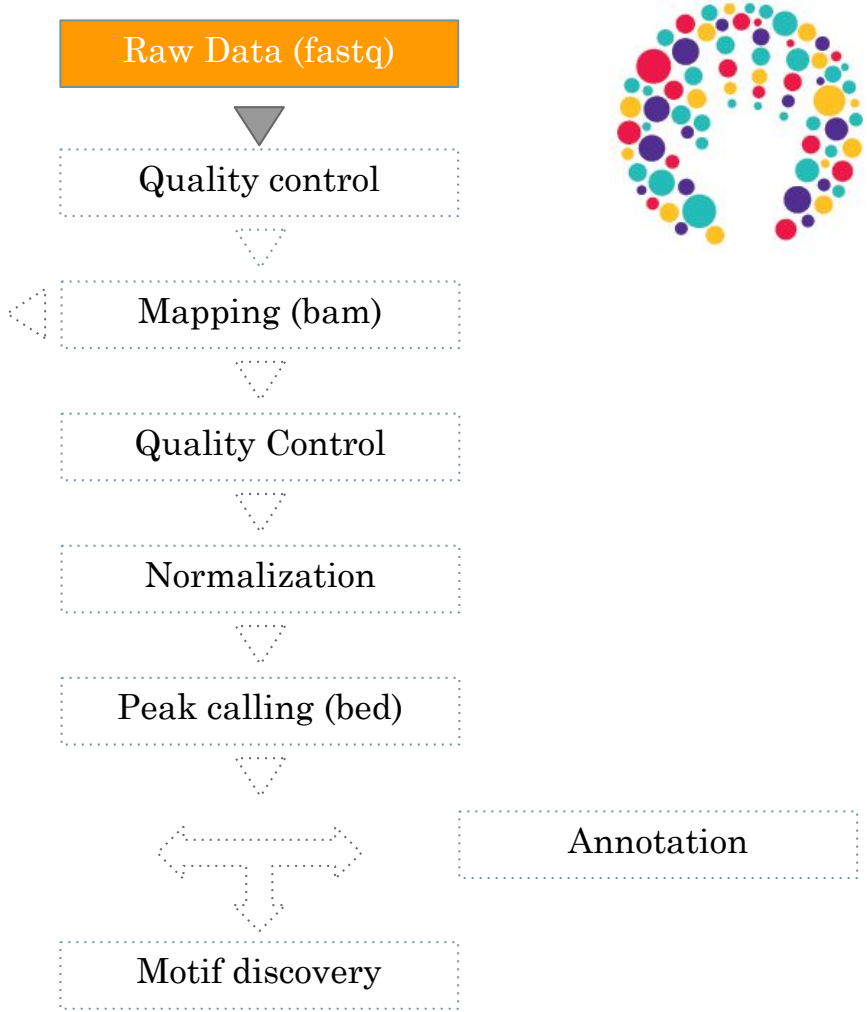
  => *Bibliothèque de données* / Data library

  =>*GTN - Material*

  =>*ChIP-Seq data analysis*

  =>*Formation of the Super-Structures on the Inactive X*

  =>*DOI: 10.5281/zenodo.1324070*



Quality control of the reads

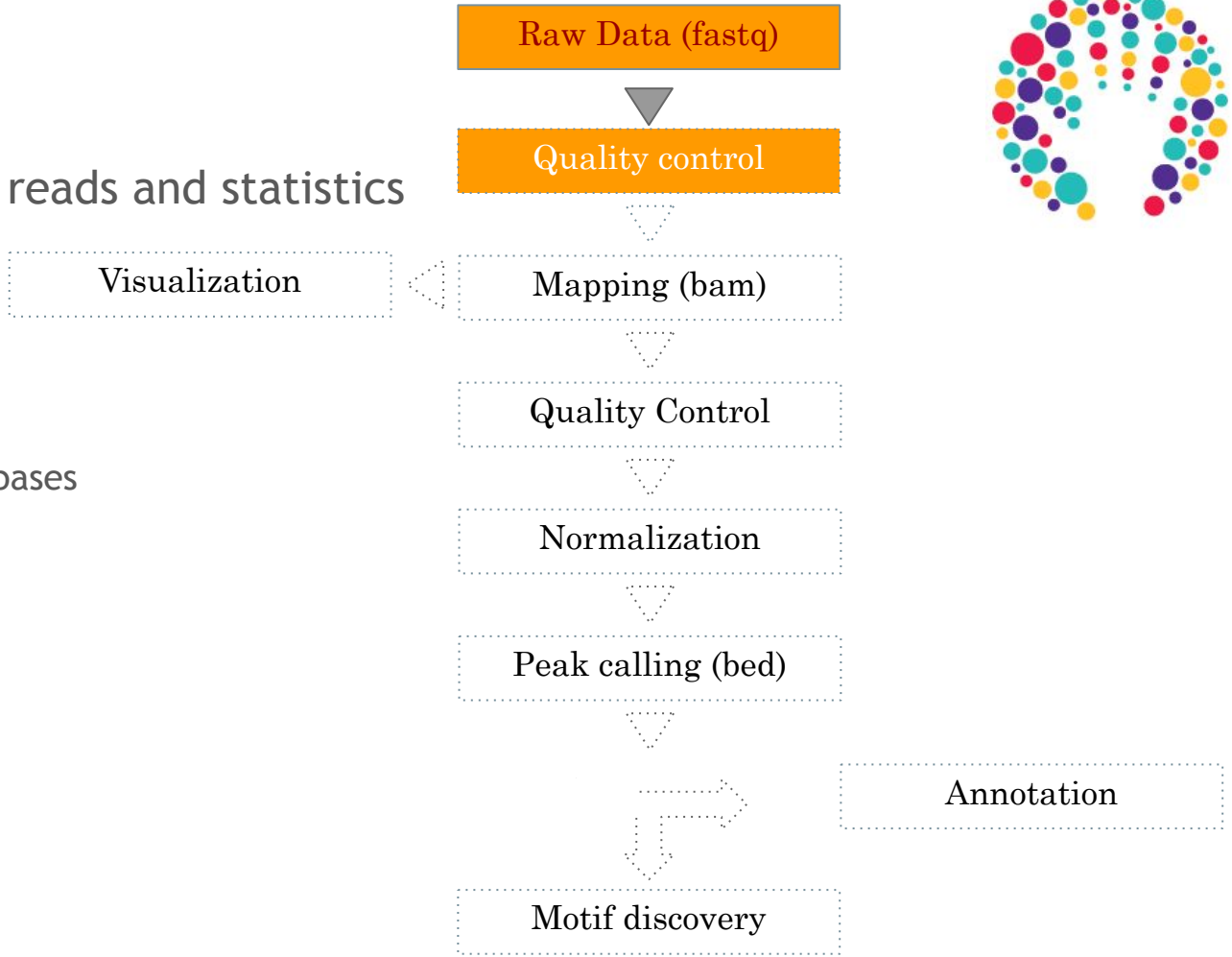
Quality control of the reads

- As for any NGS dataset
- FastQC program (cfr cours Introduction, nettoyage et qualité des données)

Protocol

- Quality control of the reads and statistics

- [Step 1] Hands-on :
  a. Quality control
  b. Trimming low quality bases

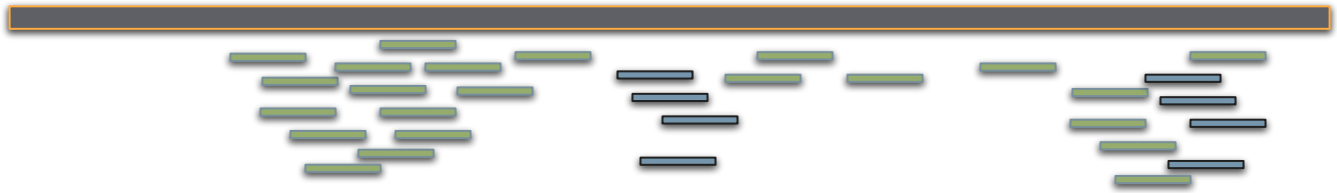


Mapping

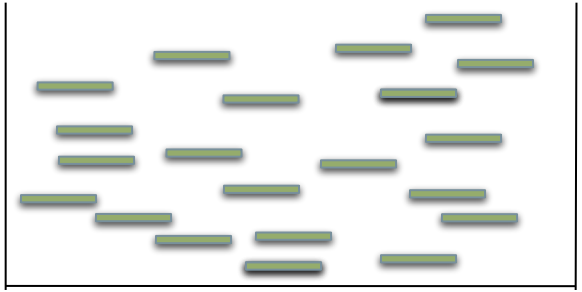
Mapping

- Find out the position of the reads within the reference genome

Ref. Genome

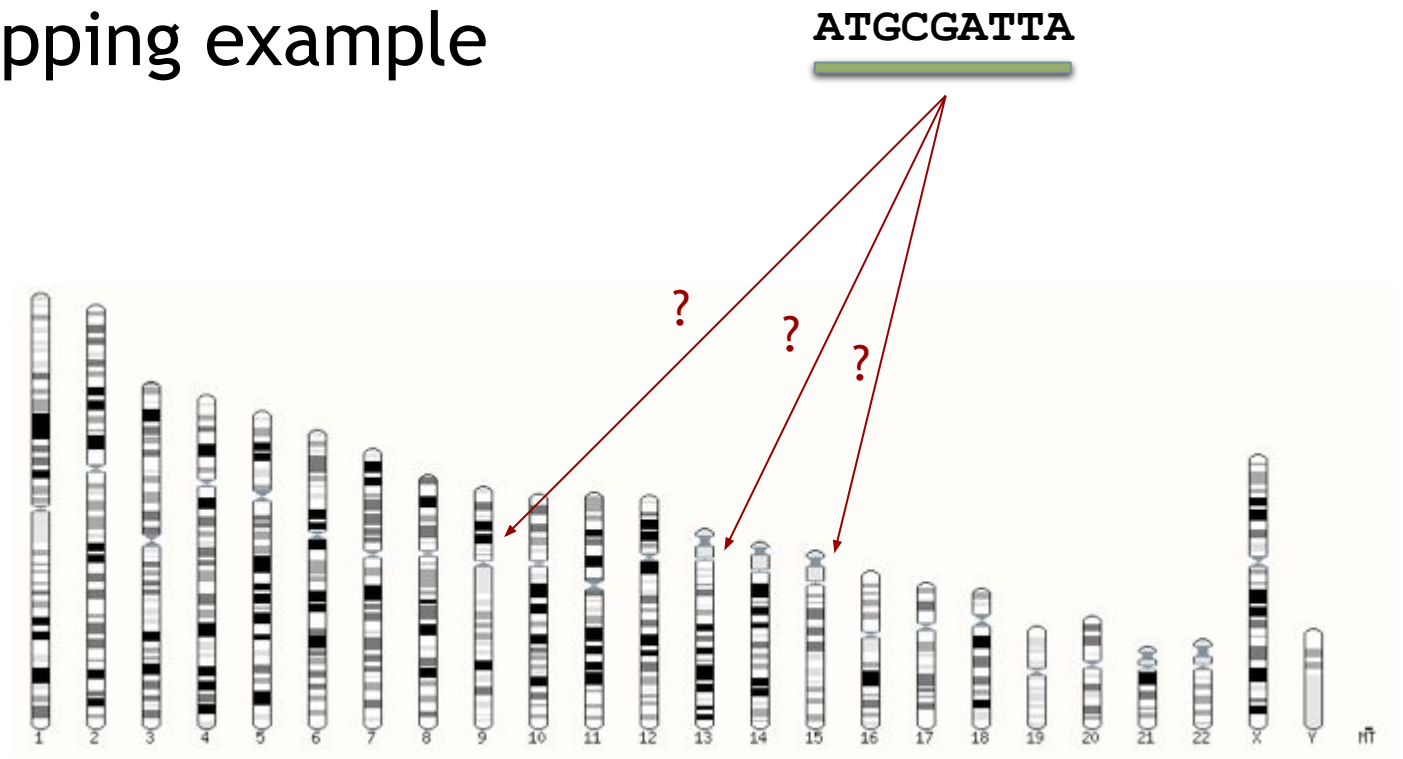


Reads



- One position in the genome
- Many possible positions (Repeat regions, duplicate regions, pseudogenes…)

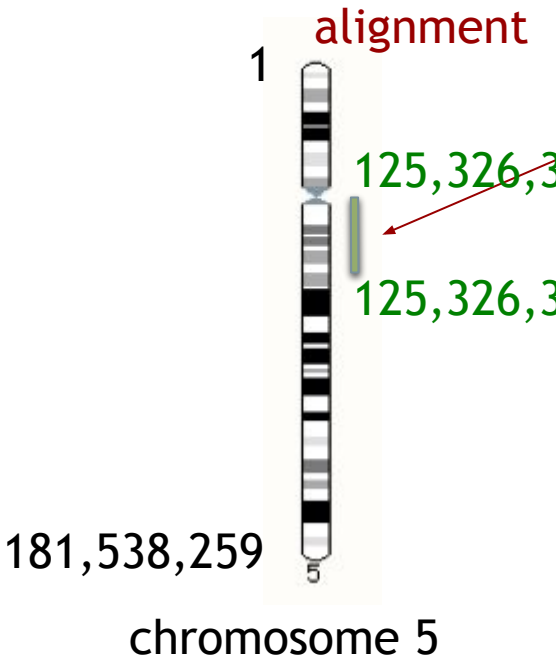
Mapping example



Human chromosomes

Genomic coordinates

ATGCGATTA



Genomic coordinate of the mapped read :

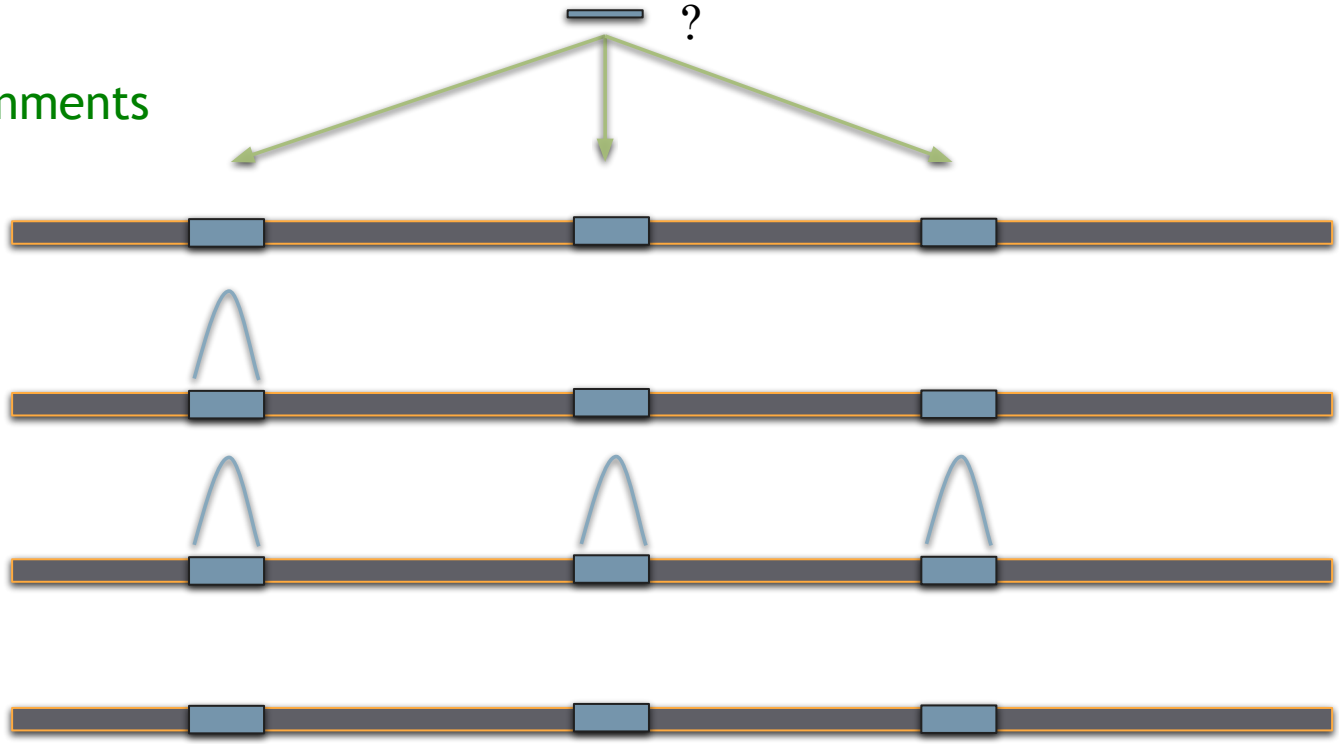
```
chr5   125326318 125326358 +
```

Mapping tool used: Bowtie

- (cfr course “mapping”)
- Designed to align reads if:
  - many of the reads have at least one good, valid alignment,
  - many of the reads are relatively **high-quality**
  - the number of alignments reported per read is small (close to 1)
- Langmead B. et al, Genome Biology 2009
- Langmead B (2010) Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11: Unit 11 17

Duplicated genomic regions

3 possible alignments



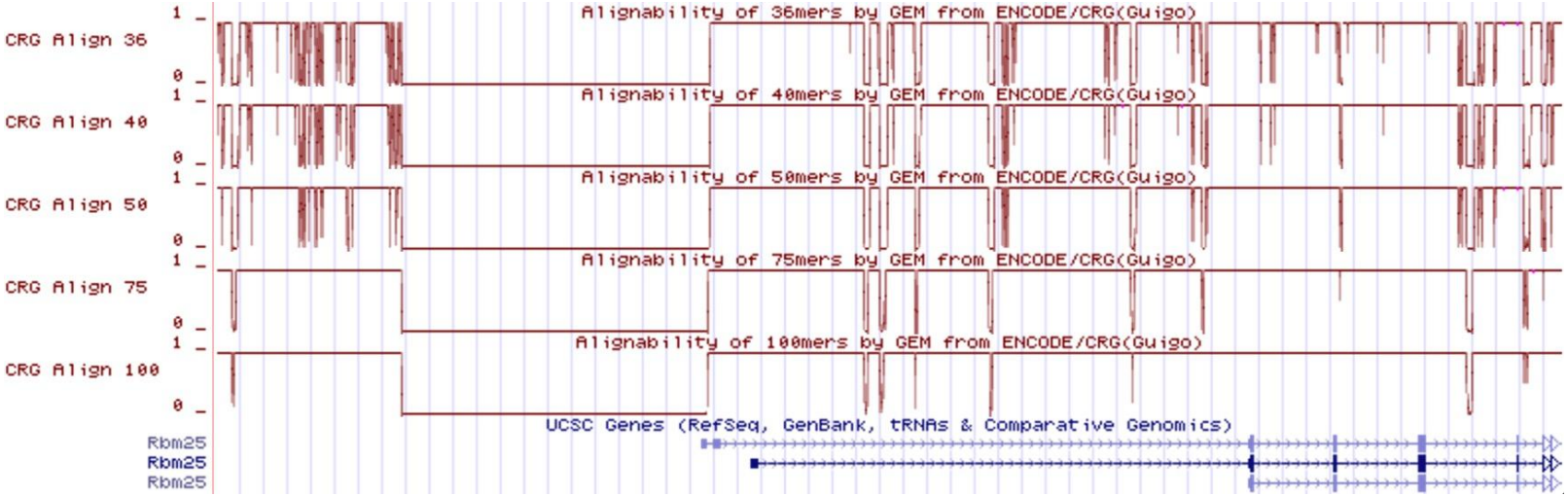
Keep 1 position
best + random

Keep all
possible position

Keep none

Mappability

- Mappability (a): how many times a read of a given length can align at a given position in the genome
  - a=1 (read align once)
  - a=1/n (read align n times)



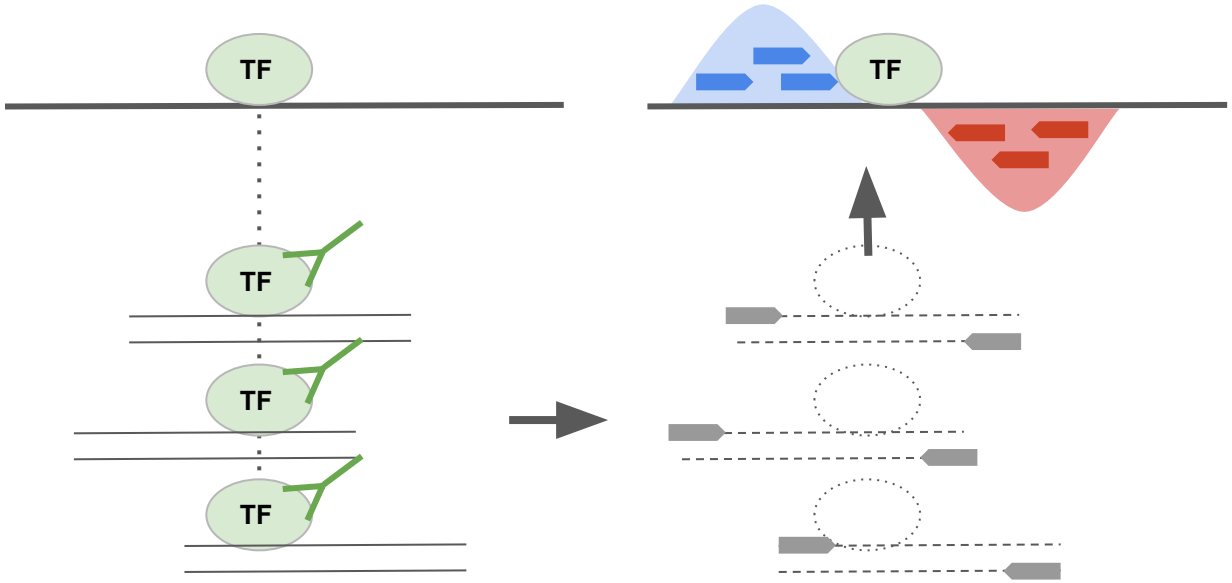
Protocol

- Mapping the reads with Bowtie 2

- [Step 2] Hands-on :
  a. Mapping
  b. Inspect a BAM/SAM file

Mapping: expected signal

- For a transcription factor signal is expected to be **sharp**

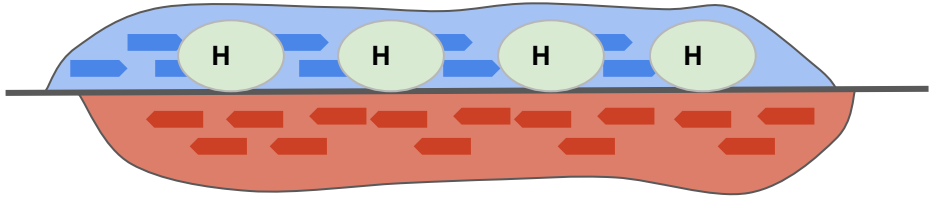


The binding site itself is generally not sequenced !

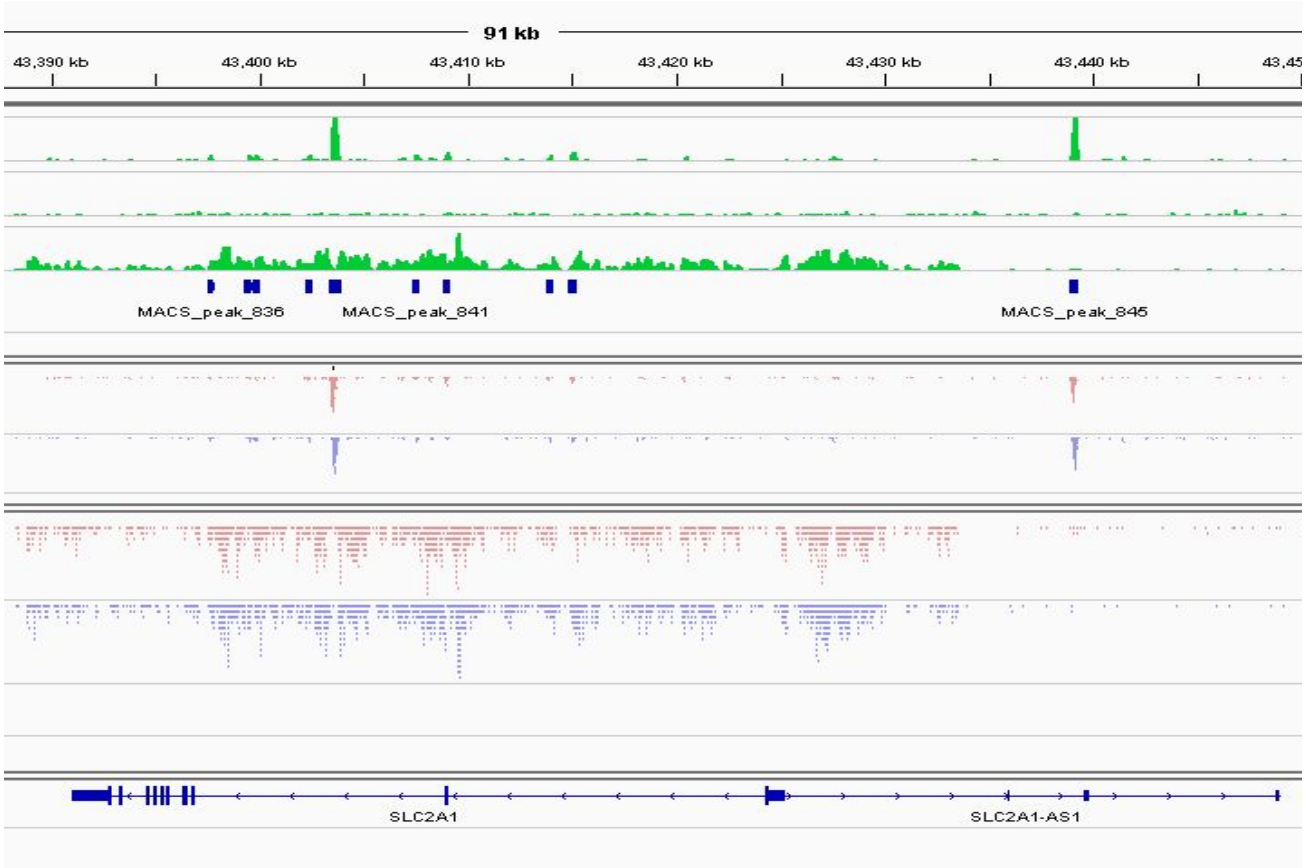
Sense alignments
Rev/comp alignments

Mapping: the expected signal

- For most **histone marks** the signal is expected to be **broad**
- Asymmetry is less/not pronounced
- Peak calling algorithms need to adapt to these various signals



Mapping: observed signal



Trans. Factor (ESR1)

Histone mark (H3K4me1)

Protocol

- Visualization

- [Step 2] Hands-on : Visualization of the reads in IGV

  Alternative: download bam and bai files and use **https://igv.org/app**

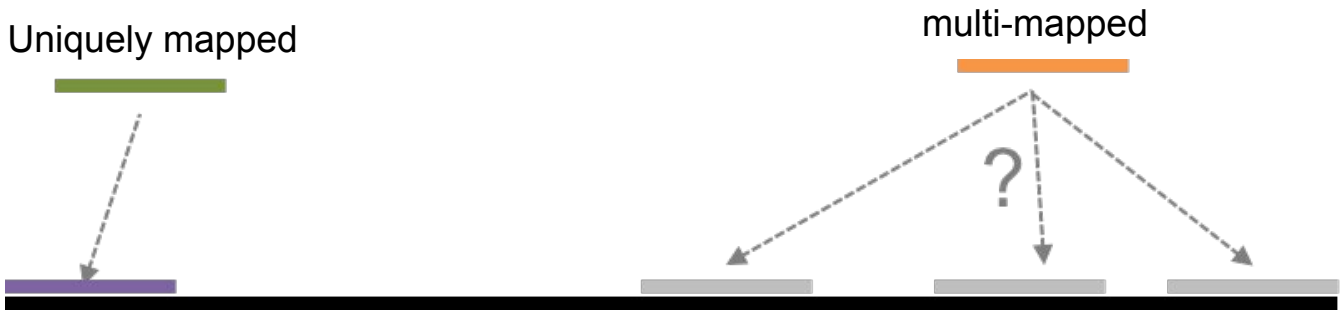
Filtering mapped reads

Which reads to filter ?

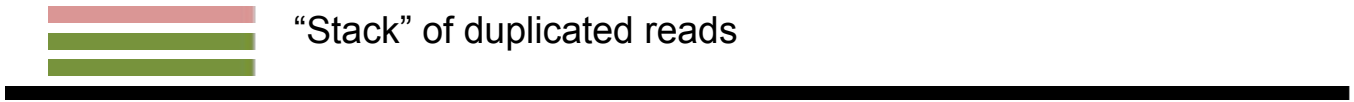
- Low-quality read alignments
  - Tool : samtools
- Multi-mapped reads (unless removed during the mapping step)
  - Tool : samtools
- Duplicated reads (PCR duplicates)
  - Tool : Picard MarkDuplicates

Source of confusion

uniquely mapped reads = reads that “matches” only 1 region in the genome



duplicated reads = reads that “match” at the SAME location (same start, strand)



PCR duplicates

- PCR duplicates
  - Related to poor library complexity
  - The same set of fragments are amplified, may indicate that immuno-precipitation failed
  - Tool to check for
    - FastQC report (duplicate diagram)

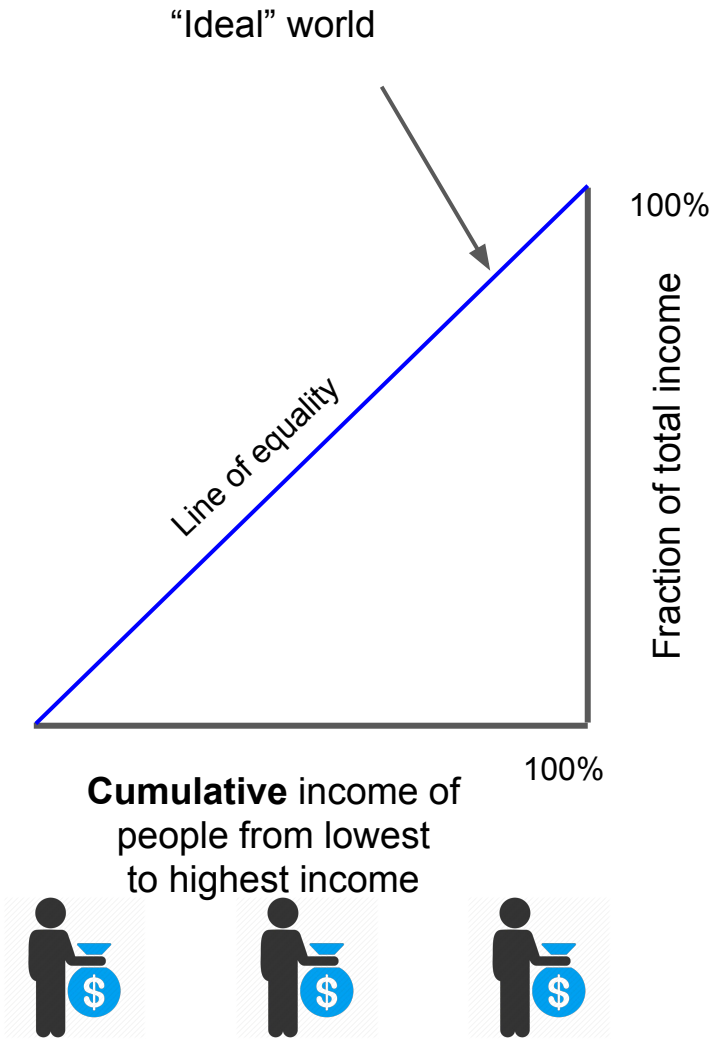
Quality Control on mapped reads

Assessing ChIP quality

- Guidelines from ENCODE
- Various metrics
  - Check **duplicate** rate (see previous Filtering section)
  - Correlation between samples (implemented in **Deeptools multiBamSummary**)
  - Use a **Lorenz Curve** (implemented in **Deeptools plotFingerprint**)

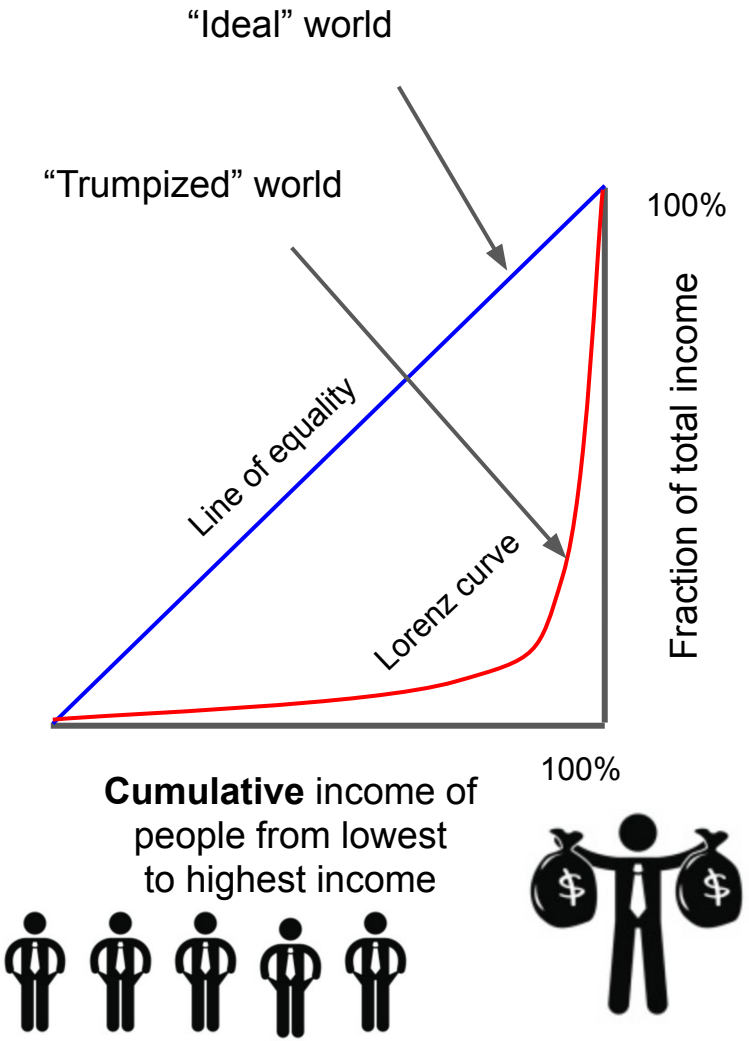
Lorenz curve

- Analyze income among workers by computing cumulative sum.
  - If uniform income distribution :
    - Straight line



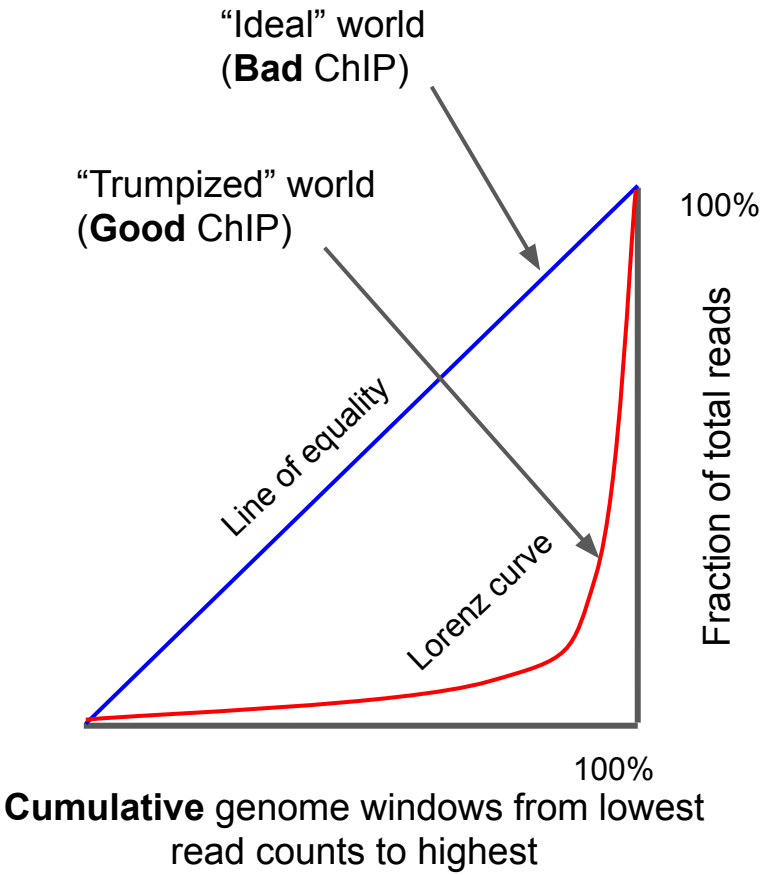
Lorenz curve

- Analyze income among workers by computing cumulative sum.
  - If uniform income distribution :
    - Straight line
  - If they were trumpized
    - Lorenz curve

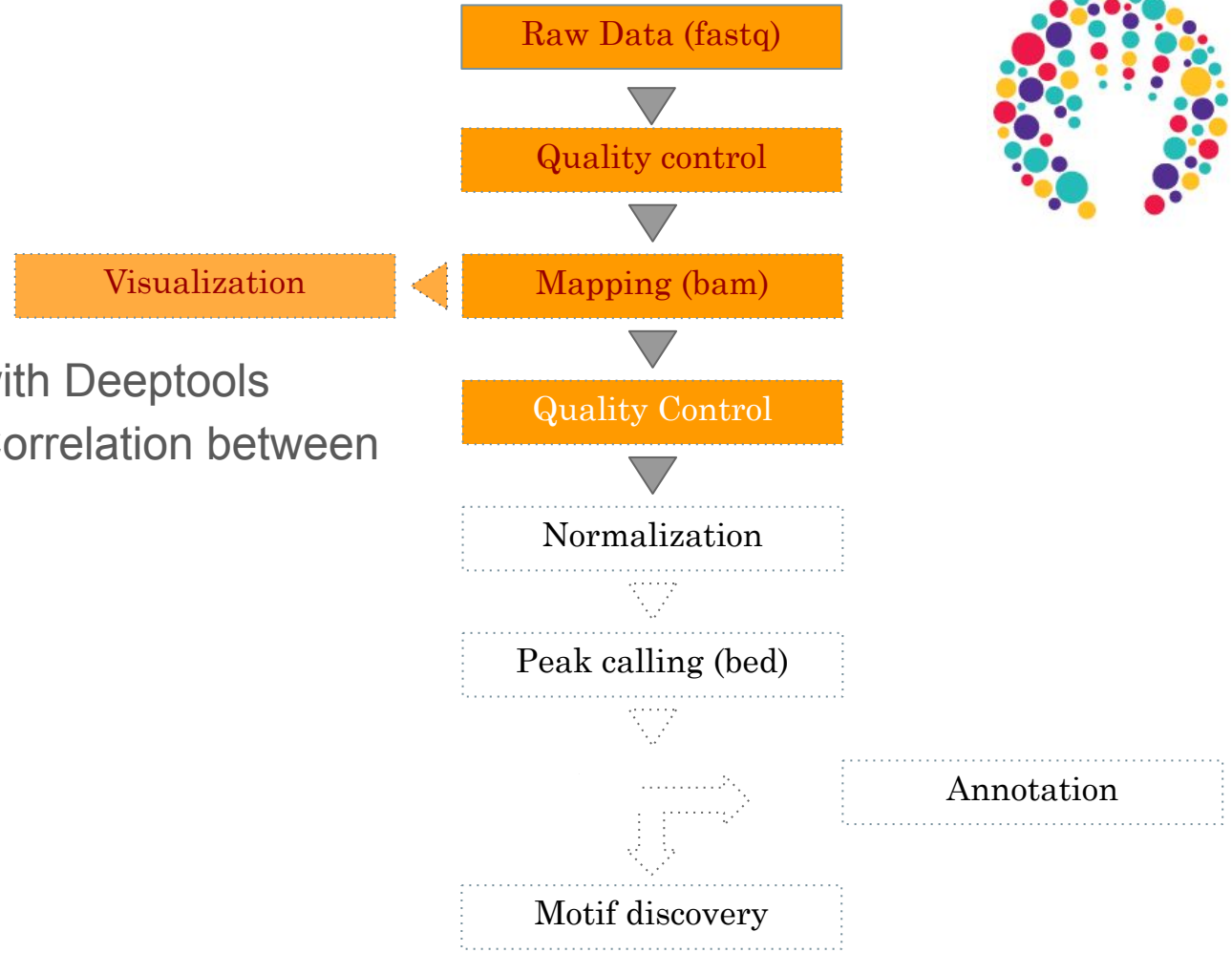


Lorenz curve

- Analyze income among workers by computing cumulative sum.
  - If uniform income distribution :
    - Straight line
  - If they were trumpized
    - Lorenz curve
- Here the workers are the genome windows and incomes are reads

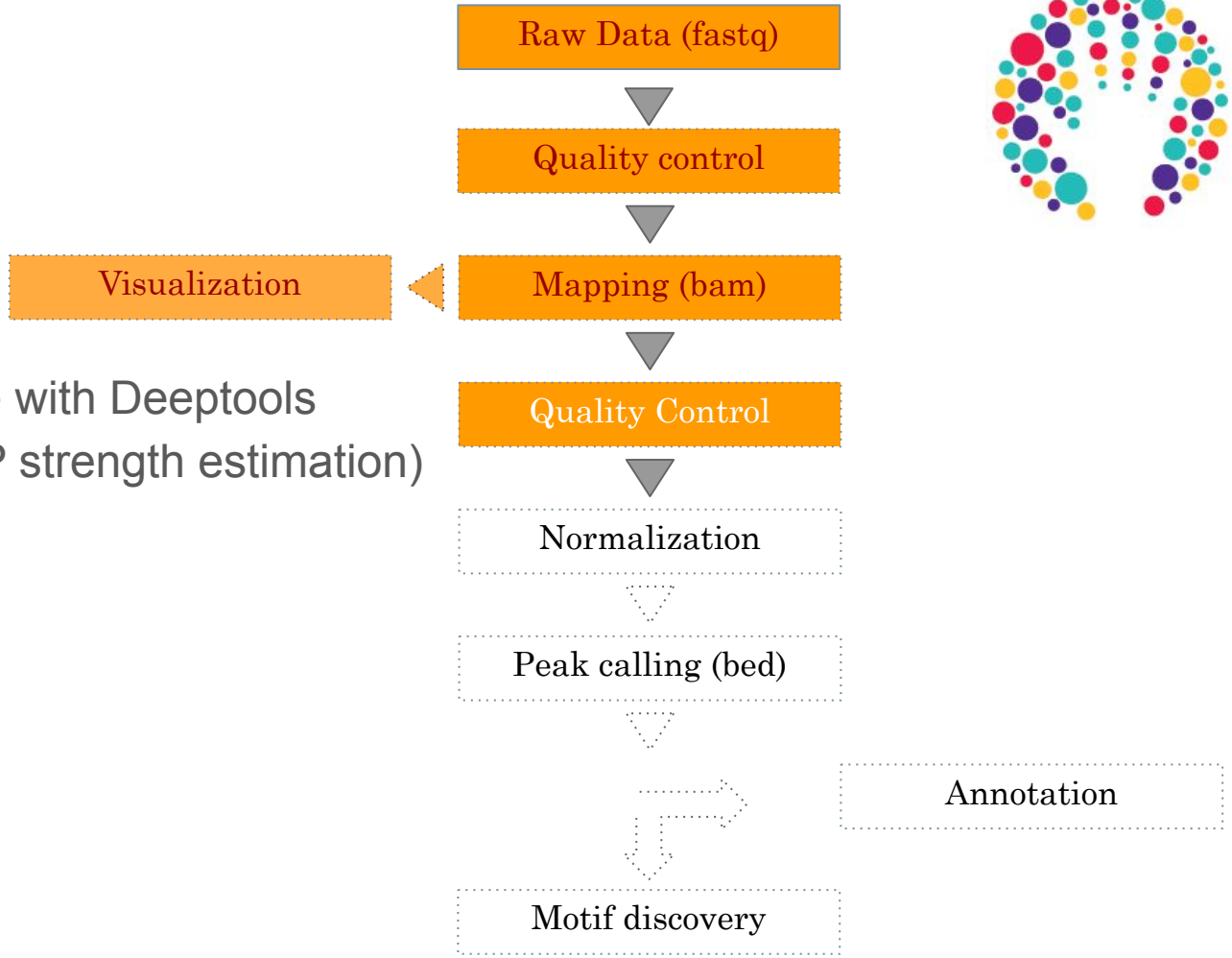


Protocol



- Plot the correlation with Deeptools
- [Step 3] Hands-on: Correlation between samples

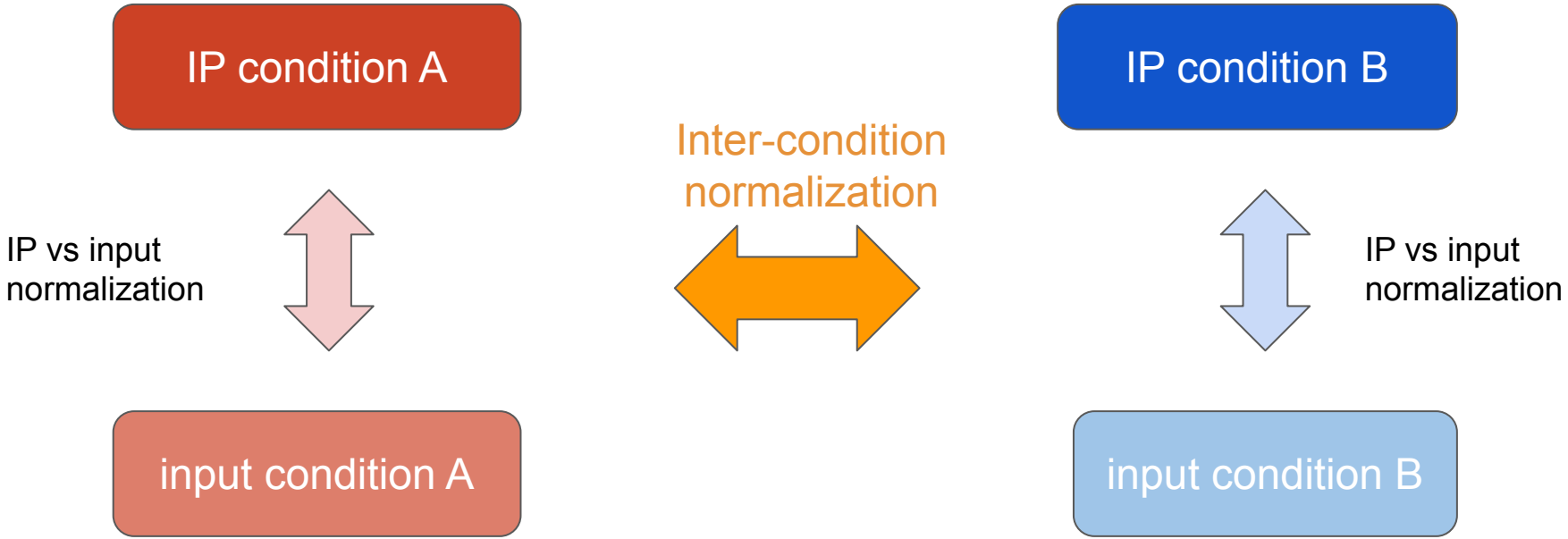
Protocol



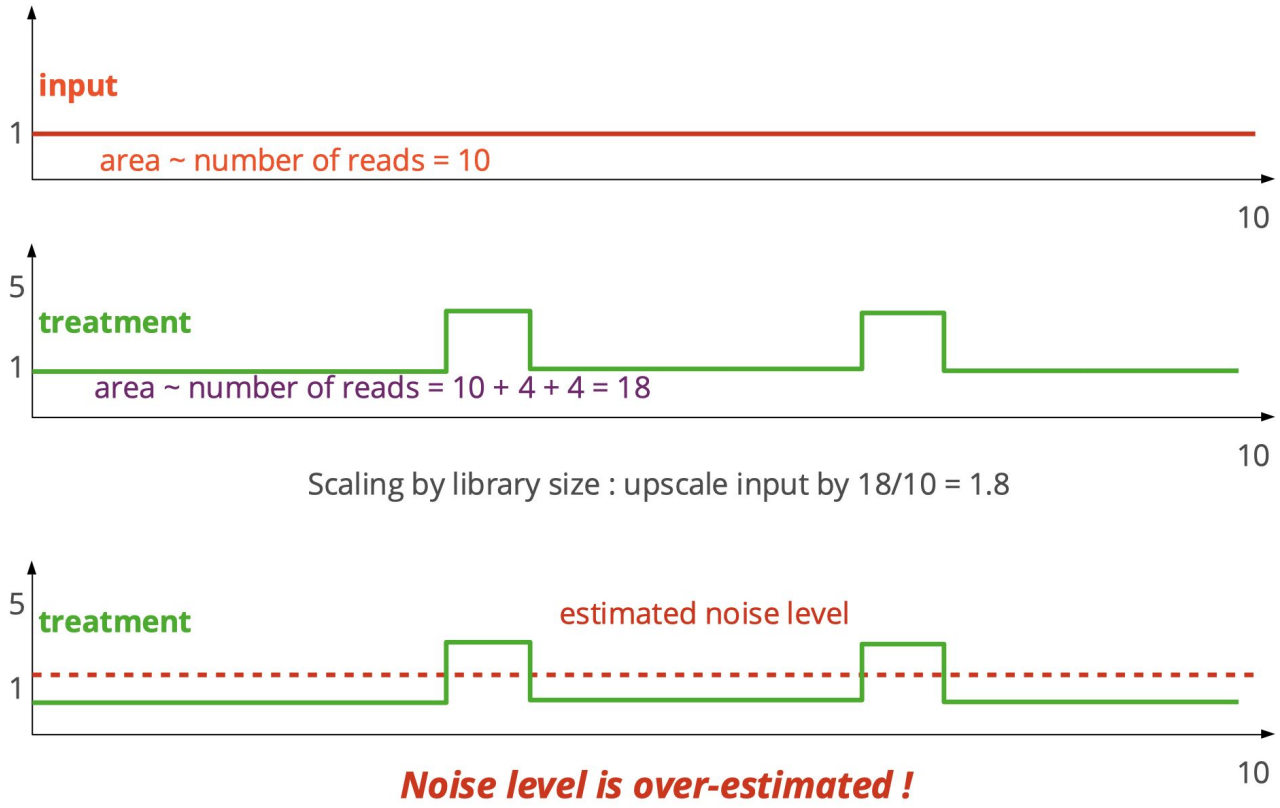
- Plot the Lorenz curve with Deeptools
- [Step 3] Hands-on: IP strength estimation)

Normalization

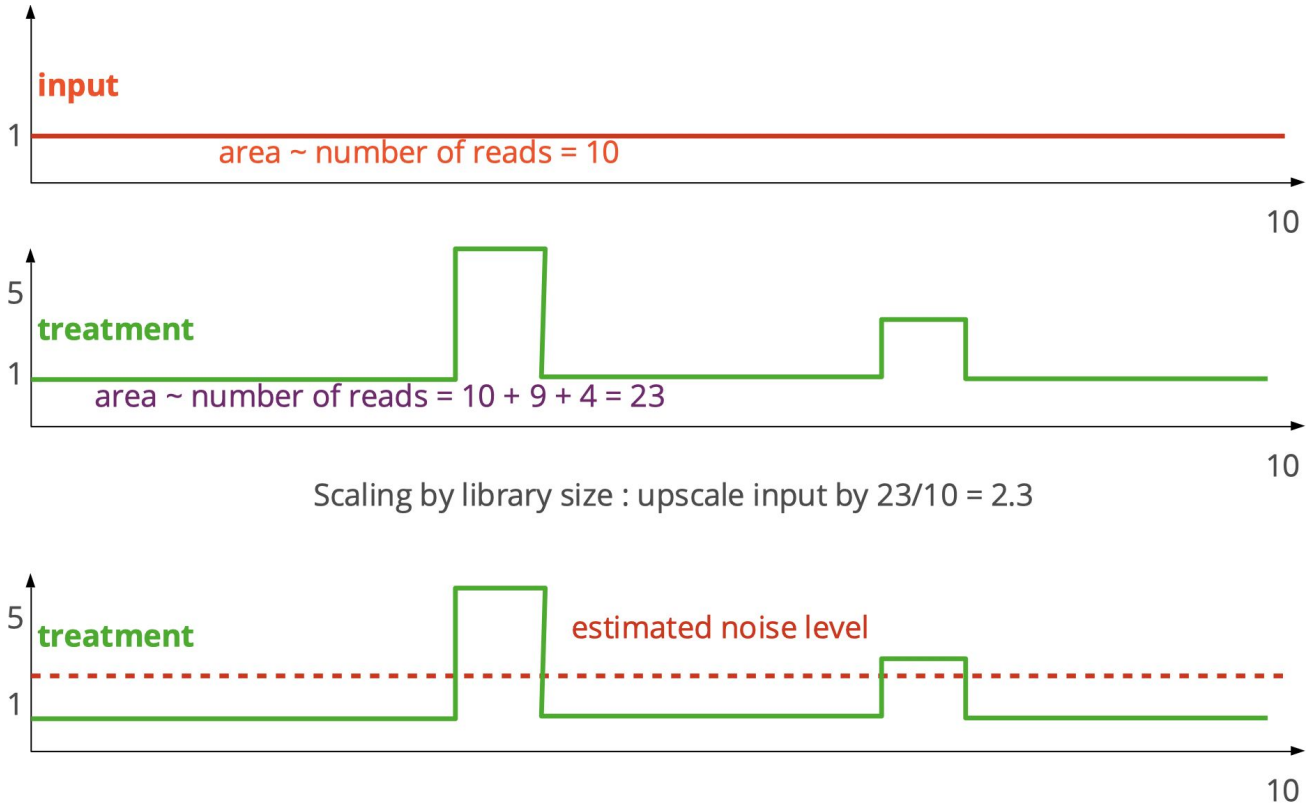
Library size normalization



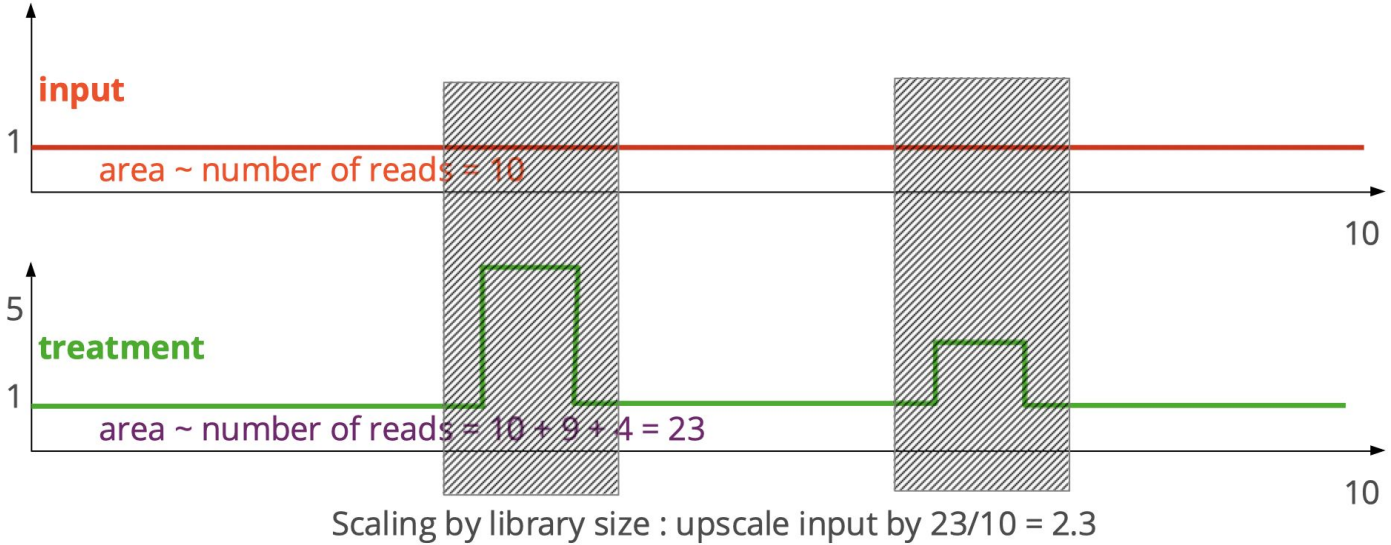
Library size normalization (input vs IP)

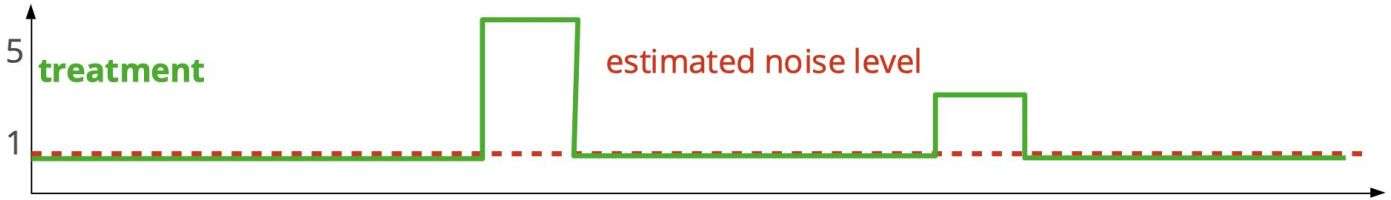


Library size normalization (input vs IP)

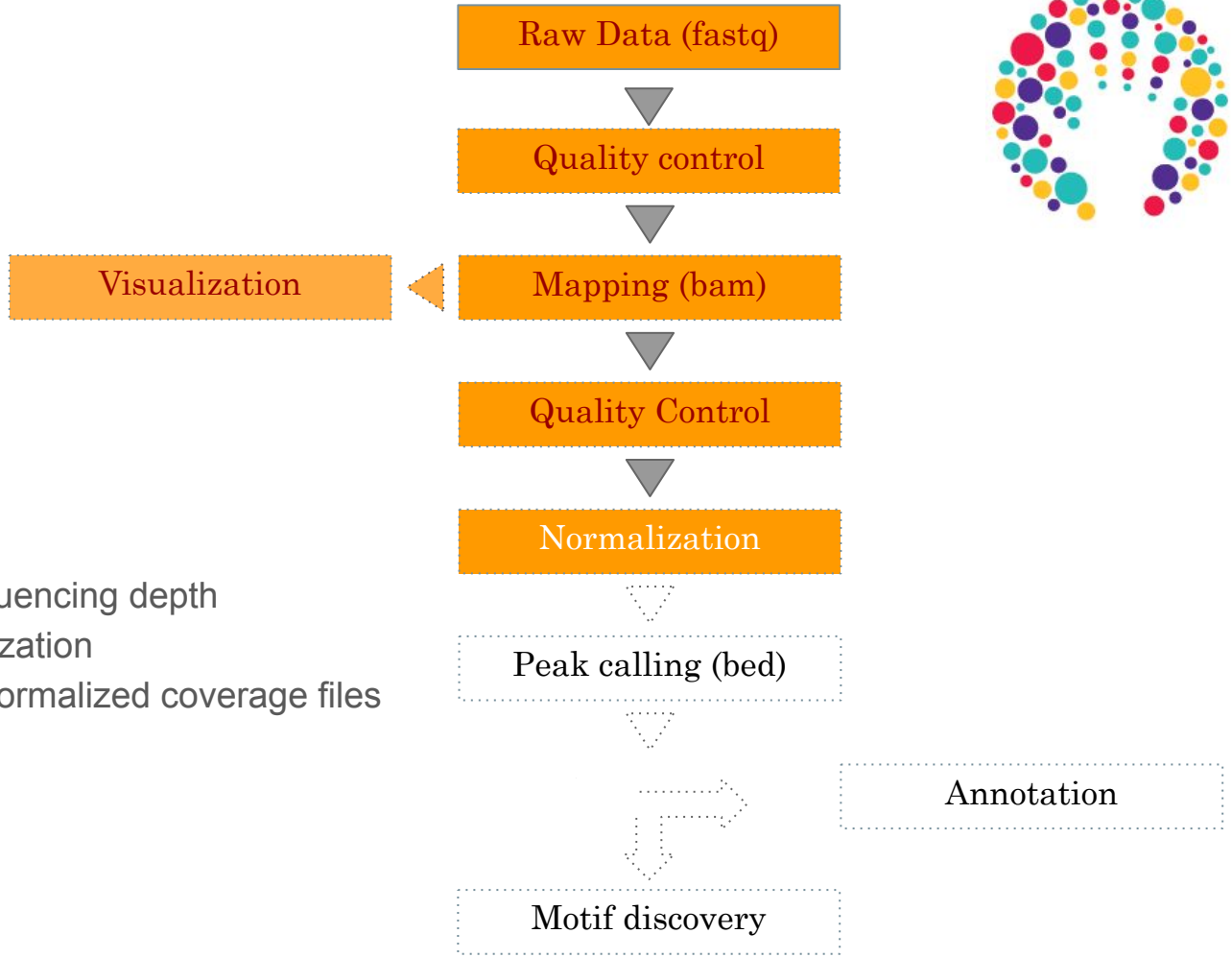


Library size normalization (input vs IP)





Protocol



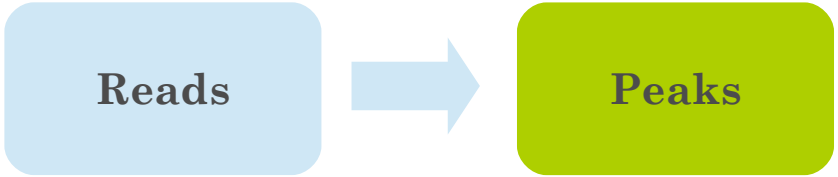
- Normalization
- [Step 4] Hands-on :
  a. Estimation of the sequencing depth
  b. Coverage file normalization
  c. Generation of input-normalized coverage files

Bam files are fat

- **BAM files are fat** as they do contain exhaustive information about read alignments
  - Memory issues (can only visualize fraction of the BAM)

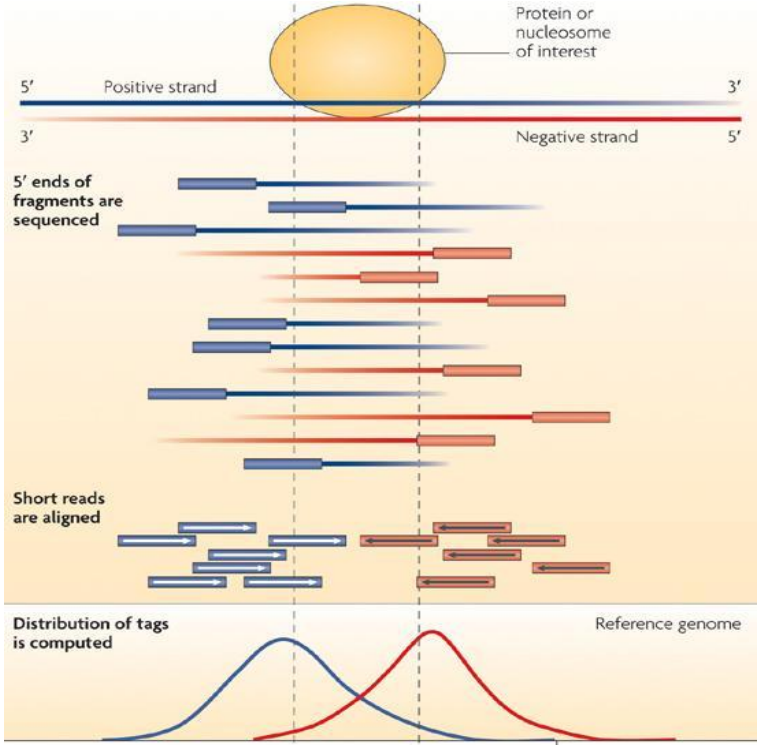
- Need a more **lightweight** file format containing **only genomic coverage** information:
  - ❌ **Wig (not compressed, not indexed)**
  - ✅ **TDF (compressed, indexed)**
  - ✅ **BigWig (compressed, indexed)**

Peak Calling



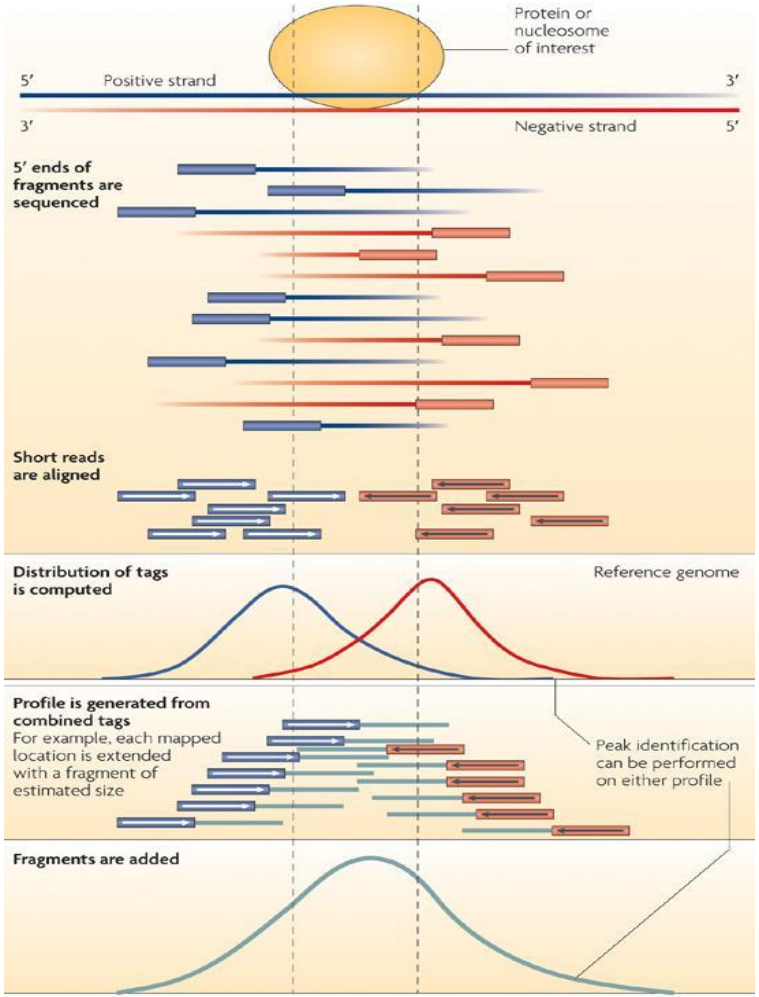
From reads to peaks

- Chip-seq peaks are a mixture of two signals:
  - + strand reads (Watson)
  - - strand reads (Cricks)
- The sequence read density accumulates on forward and reverse strands centered around the binding site



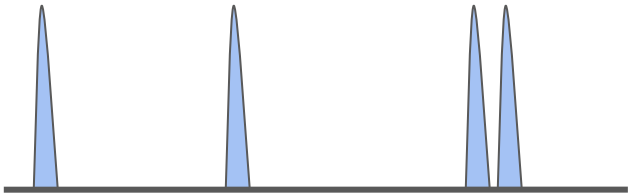
From reads to peaks

- Get the signal at the right position
  - Read shift
  - Extension
- Estimate the fragment size
- Do paired-end

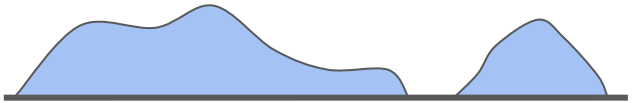


Peak callers

- The **peak caller** should be **chosen** based on
  - Experimental design
    - SE or PE (E.g MACS1.4 vs MACS2)
  - Expected signal
    - Sharp peaks (e.g. Transcription Factors).
      - E.g. MACS
    - Broad peaks (e.g. epigenetic marks).
      - E.g MACS, SICER,...



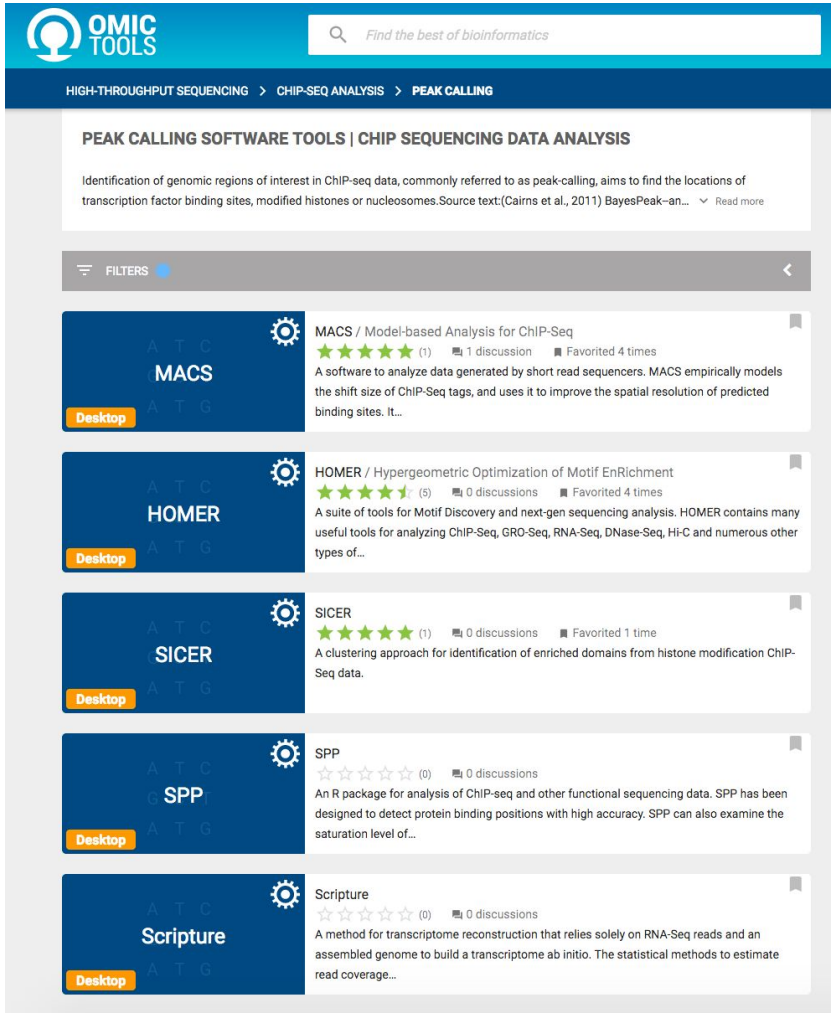
Sharp peaks



Broad peaks

A variety of peak callers

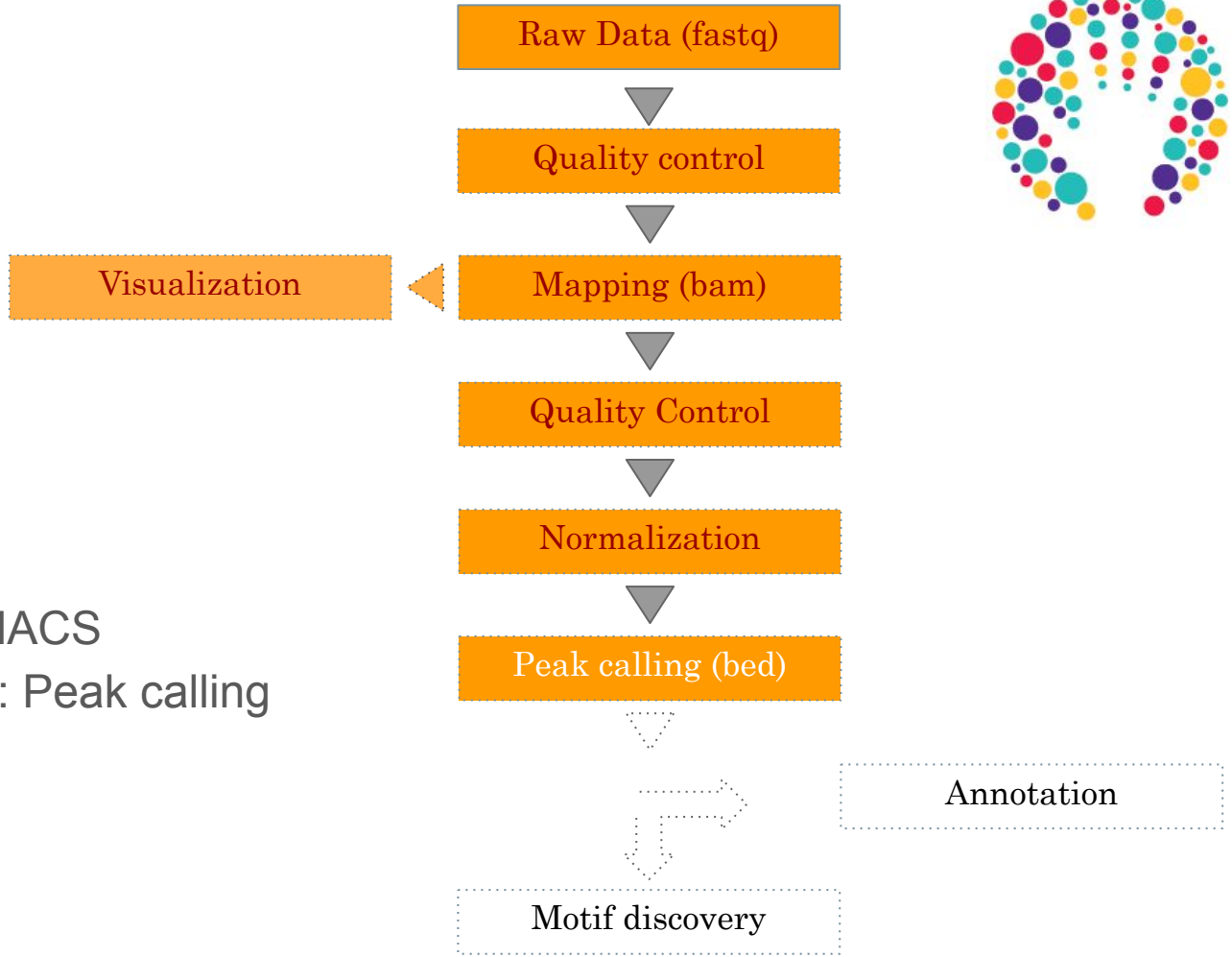
- 60 programs listed on OMICTOOLS
- Most support a control



MACS in summary

- Step 1 : search for candidate regions that look like good peaks, to produce a fine-tuned model of the peaks (d value) to search in Step 2
- Step 2 : actual peak calling
  - sliding window length = 2*d
  - In each window : test if the region is a peak, by comparing the number of reads in the treatment and the expected number of reads
  - Comparison is based on a statistical test with a Poisson distribution, keeping only regions with p-value < threshold
- Step 3 : correction for multiple testing (many windows were tested), calculation of **FDR**

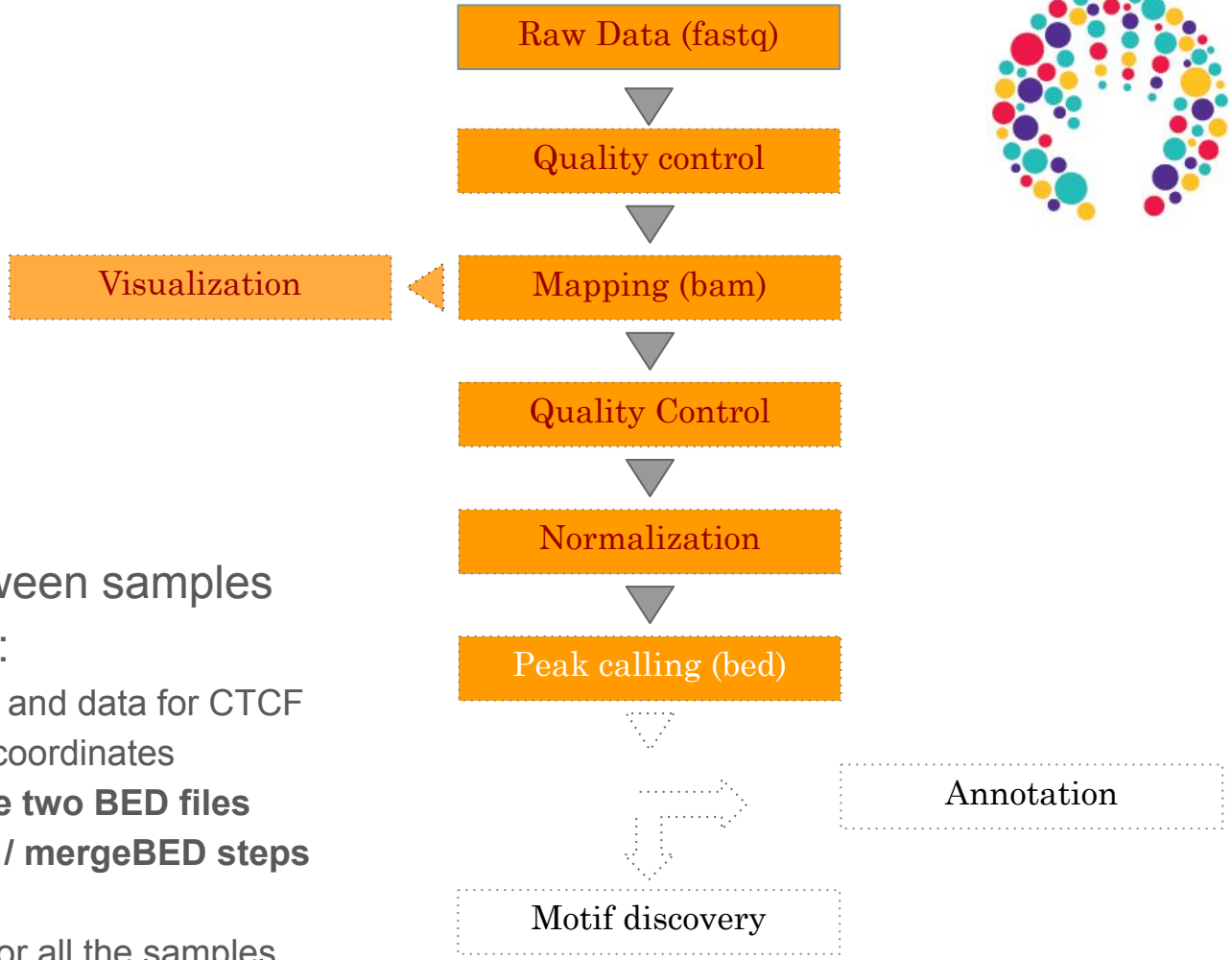
Protocol



- Peak calling with MACS
- [Step 5] Hands-on : Peak calling

Visualize ChIP enrichment

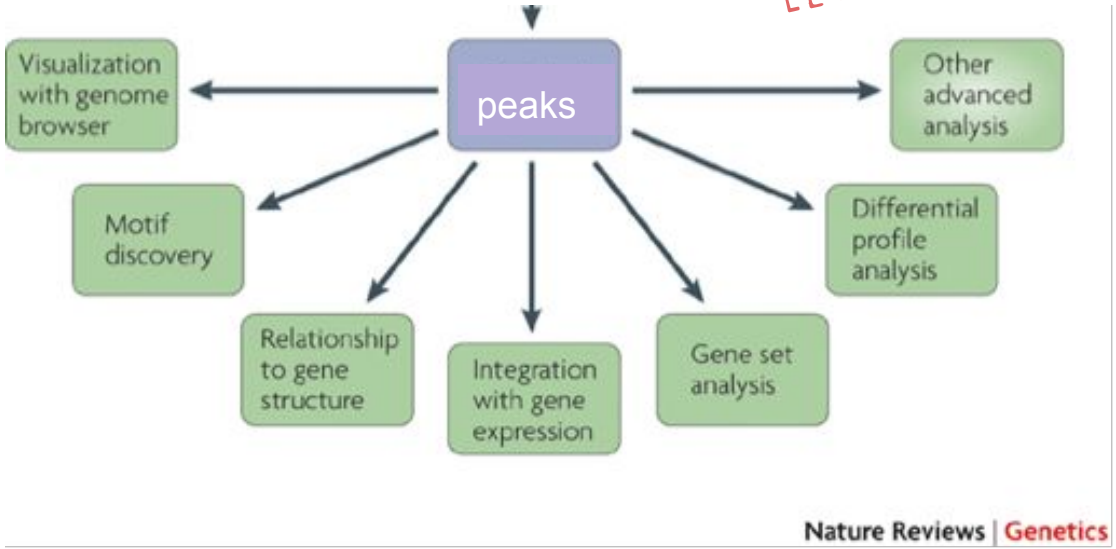
Protocol



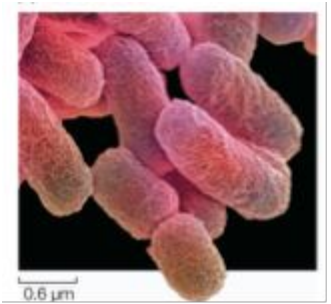
- Plot the signal between samples
- [Step 6] Hands-on :
  a. Prepare the peaks and data for CTCF
  b. Prepare the peak coordinates
     i. **Concatenate two BED files**
     ii. **no sortBED / mergeBED steps**
  c. Plot the heatmap
  d. Plot the heatmap for all the samples

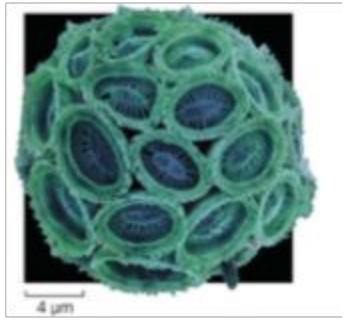
Processing steps are over !

LET'S DO BIOLOGY !!!



Park, Nature reviews 2009





What is the biological question ?

What is the biological question ?

« see if you can find something in the data »

What is the biological question ?

~~« see if you can find something in the data »~~

What is the biological question ?

- **Where** do a transcription factor (TF) bind ?

- **How** do a transcription factor (TF) bind ?
  - Which binding motif(s) (can be several for a given TF !!)
  - Is the binding direct to DNA or via protein-protein interactions ?
  - Are there cofactors (maybe affecting the motif !!), and if so, identify them

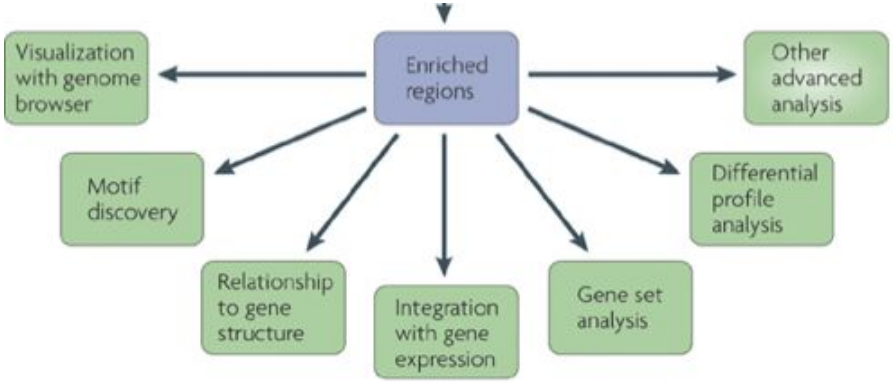
- Which **regulated genes** are directly regulated by a given TF ?

- Where are the **promoters** (PolII) and **chromatin marks** ?

What is the biological question ?

Should drive all « downstream » analyses

Will take time
to « do it all » !!!

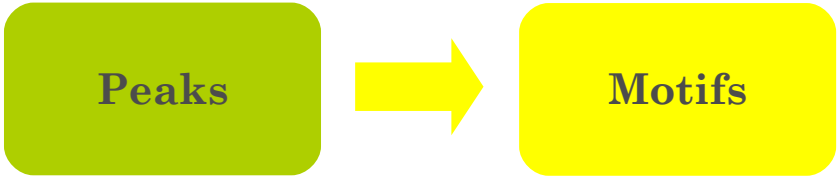


What is the biological question ?

What can be the following experimental work ?

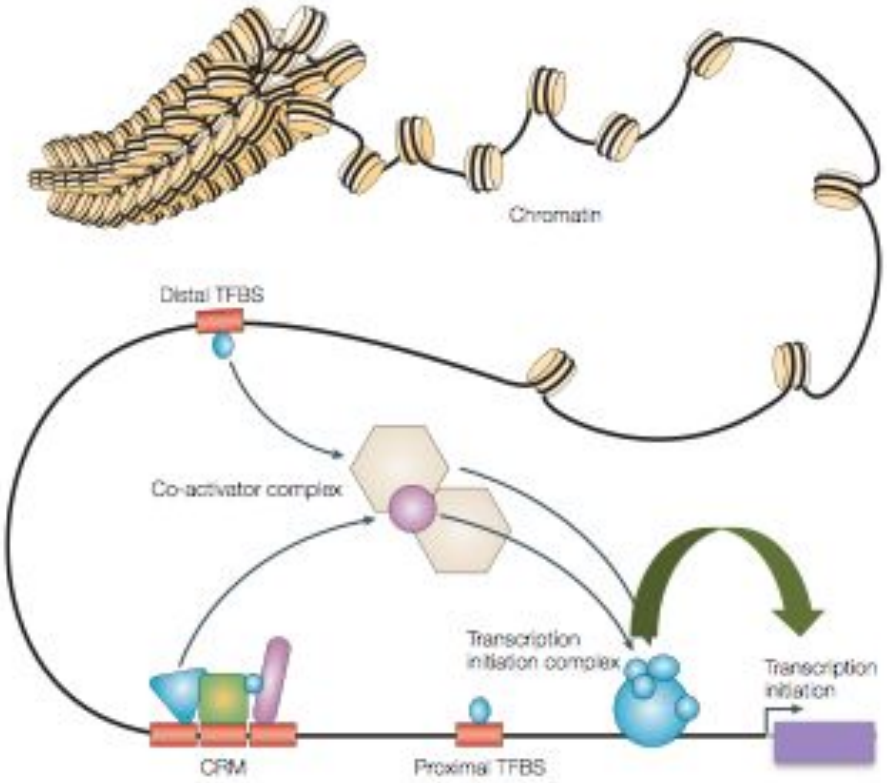
- cell biology (eg: luciferase assay) ?
- in vitro assays (eg: EMSA) ?
- Proteomics (eg: mass spectrometry) ?
- Transgenics ?
- Will depend on
  - the organism
  - available infrastructure

Discovering motifs in peaks



Biological concepts of transcriptional regulation

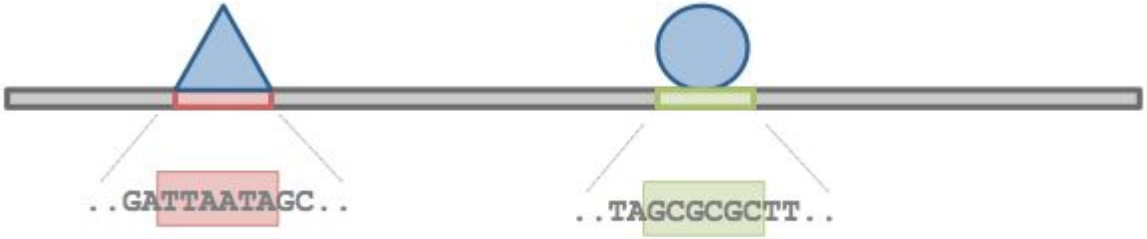
Transcription factors are proteins that modulate (activate/repress) the expression of **target genes** through the binding on **DNA cis-regulatory elements**



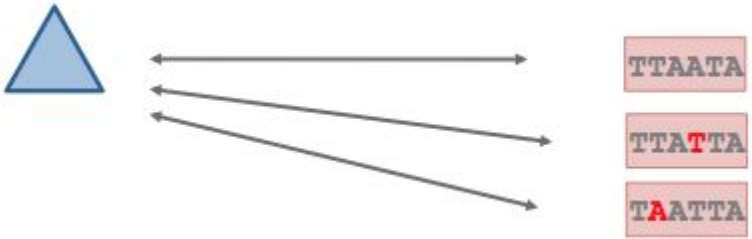
Wasserman et al, Nat Rev Genet, 2004

Transcription factor specificity

How do TF « know » where to bind DNA ?

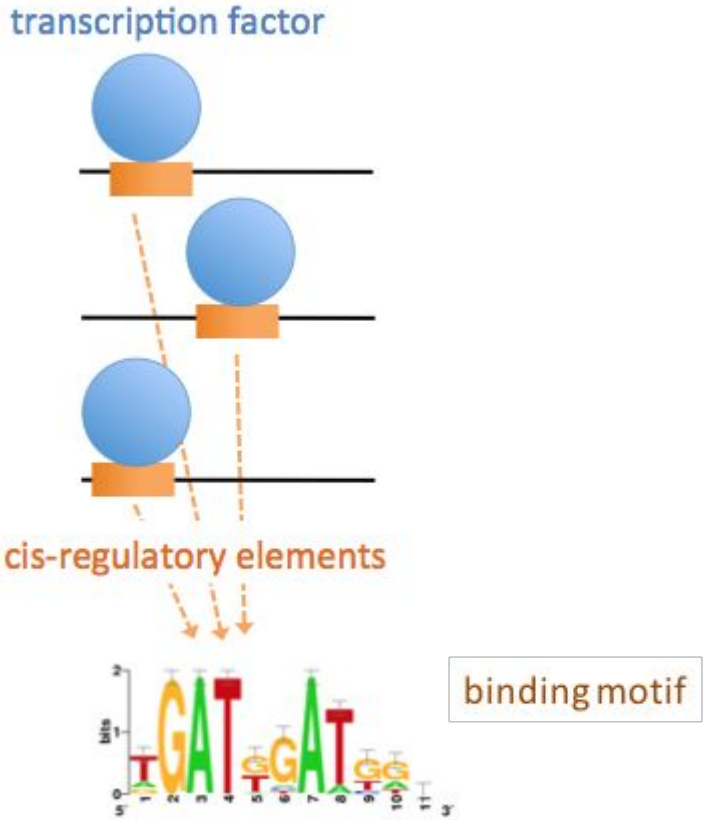


TF recognize TFBS with specific DNA sequences



a given TF is able to bind DNA on TFBSs with different sequences

Binding specificity



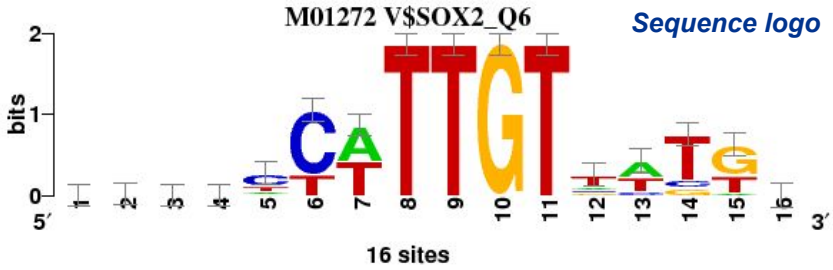
From binding sites to binding motif

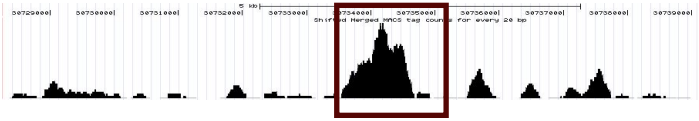
Collection of binding sites
used to build the Sox2 matrix
(TRANSFAC M01272)

```
R15133   GCCCTCATTGTTATGC
R15201   AAACTCTTTGTTTGGA
R15231   TTCACCATTGTTCTAG
R15267   GACTCTATTGTCTCTG
R16367   GATATCTTTGTTTCTT
R17099   TGCACCTTTGTTATGC
R19276   AATTCCATTGTTATGA
R19367   AAACTCTTTGTTTGGA
R19510   ATGGACATTGTAATGC
R22342   AGGCCTTTTGTCCTGG
R22344   TGTGCTTTTGTNNNNN
R22359   CTCAACTTTGTAATTT
R22961   GCAGCCATTGTGATGC
R23679   CACCCCTTTGTTATGC
R25928   TTTTCTATTGTTTTTA
R27428   AAAGGCATTGTGTTTC
```

Position-specific scoring matrix (PSSM)

A	6	7	4	4	2	0	**8**	0	0	0	0	2	**7**	0	1	4
C	2	2	6	5	**9**	**12**	0	0	0	0	0	2	2	2	0	6
G	4	3	2	4	1	0	0	0	0	**16**	0	2	0	2	**9**	3
T	4	4	4	3	**4**	**4**	**8**	**16**	**16**	0	**16**	**9**	**6**	**11**	**5**	2





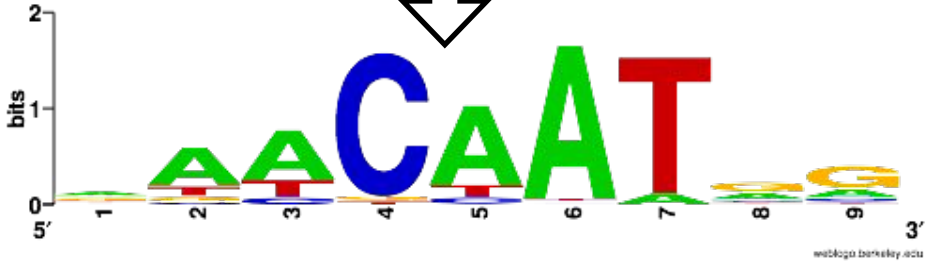
ChIP-seq peaks

```
>mm9_chr1_39249116_39251316_+
gagaggaagggggagaaagagggagggggagGGTGATAGGTAGCCAGGAG
CCAATGGGGGCGTTTTCCTTGTCCAGGCCACTTGCTGGAATGTGAGATGT
AGAATGACCCAAAGAGAGCTGCCAAGACAGAGCTCTGCCCCAGGAATTGA
ACTCAAAGGGTGTCAGAAAGCAGGTGGCCTTTGTGCACCTGGCGCGGGGA
CGTGGCTCCCCTCTTCCGGCTGGTCTAGCCAGGtgcctgcctgcctgcct
gccGTGATCTCTGGACGCCAGTAGAGGGTTGTTGTGGGTTTGGGTGAAAC
ACGCCACCCCTGAGCTCTTCCGCGGGGCTAGCAATCTCCCCATCACCCCA
TTCGCGCTCAGAACCCCCTCAGCGA[illegible]TTAACAGCAGGCCTGGTTCCCCG
```

DNA sequence

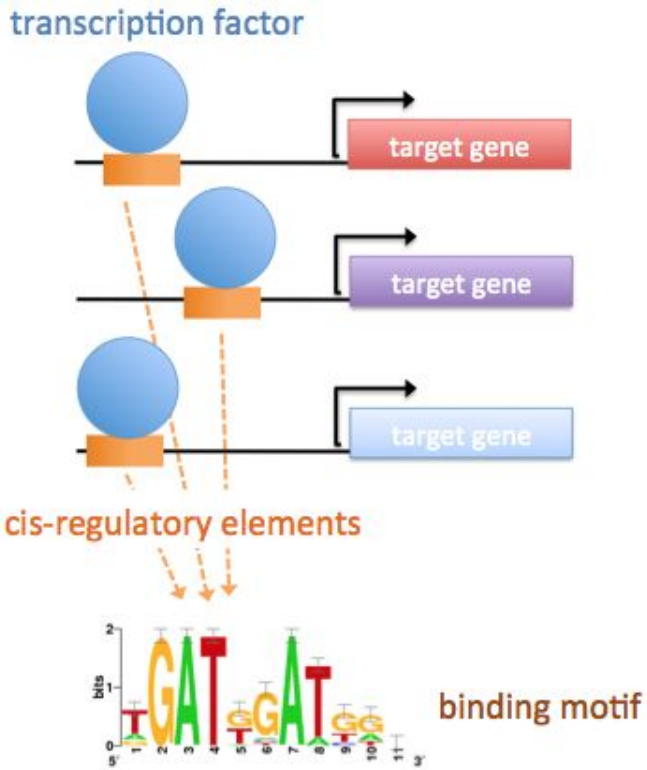
```
A   [24 54 59  0 65 71  4 24  9 ]
C   [ 7  6  4 72  4  2  0  6  9 ]
G   [31  7  0  2  0  1  1 38 55 ]
T   [14  9 13  2  7  2 71  8  3 ]
```

Discovered motif



Motif logo

De novo motif discovery



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

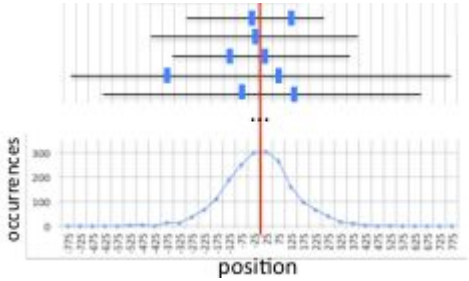
If there is a common regulating factor, can we discover its motif only using these sequences ?

De novo motif discovery

- Find exceptional motifs based on the sequence only

(No prior knowledge of the motif to look for)

- Criteria of exceptionality:
  - ***Over-/under-representation:*** higher/lower frequency than expected by chance
  - ***Position bias:*** concentration at specific positions relative to some reference coordinates (e.g. TSS, peak center, …).



Some motif discovery tools

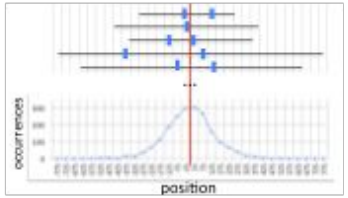
- MEME (Bailey et al., 1994)
- **RSAT oligo-analysis (van Helden et al., 1998)**
- AlignACE (Roth et al. 1998)
- **RSAT position-analysis (van Helden et al., 2000)**
- Weeder (Pavesi et al. 2001)
- MotifSampler (Thijs et al., 2001)
- … many others

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 !!!!!!

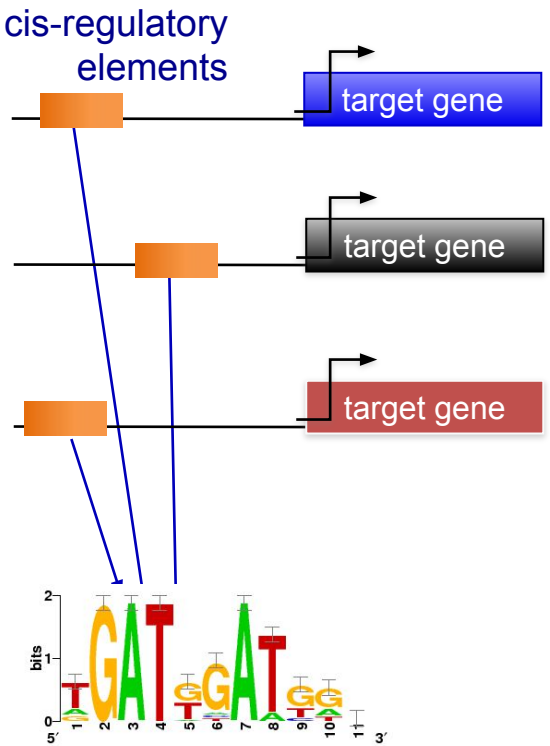
- **the problem is slightly different**
  - promoters: 200-2000bp from co-regulated genes
  - peaks: 300bp, positional bias



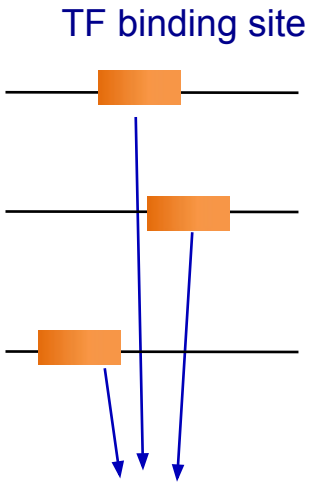
- **motif analysis: not just for specialists anymore !**
  - complete user-friendly workflows

De novo motif discovery

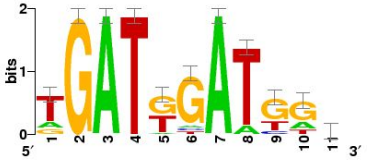
Case 1: promoters of co-expressed genes



Case 2: ChIP-seq peaks



binding motif
(represented as a sequence logo)



Regulatory sequence Analysis Tools (rsat.eu)

Regulatory Sequence Analysis Tools

Welcome to **Regulatory Sequence Analysis Tools** (**RSAT**).

RSAT

This web site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences.

RSAT servers have been up and running since 1997. The project was initiated by **Jacques van Helden**, and is now pursued by the **RSAT team**.

Choose a server

New ! *January 2015*: we are in the process of re-organising our mirror servers into taxon-specific servers, to better suit the drastic increase of available genomes.

RSAT 243 Fungi
maintained by TAGC - Université Aix Marseilles, France

RSAT 9638 Bacteria + Archaea
maintained by RegulonDB - UNAM, Cuernavaca, Mexico

RSAT 92 Metazoa
maintained by platforme ABIMS Roscoff, France

RSAT 186 Protists
maintained by Ecole Normale Supérieure Paris, France

RSAT 39 Plants
maintained by Bruno Contreras Moreira, Spain

RSAT Teaching
maintained by SLU Global Bioinformatics Center, Uppsala, Sweden

Citing RSAT complete suite of tools:

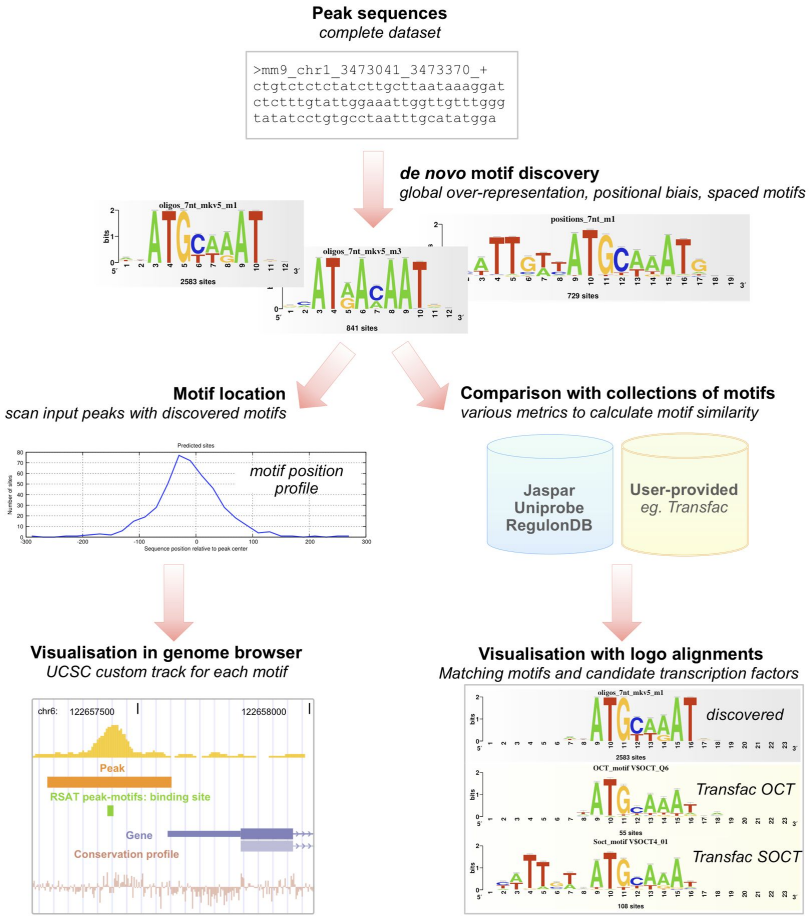
- Thomas-Chollier M, Defrance M, Medina-Rivera A, Sand O, Herrmann C, Thieffry D, van Helden J. (2011) **RSAT 2011: regulatory sequence analysis tools**. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W86-91. [Pubmed 21715389] [Full text]
- Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). **RSAT: regulatory sequence analysis tools**. Nucleic Acids Res. [Pubmed 18495751] [Full text]
- van Helden, J. (2003). **Regulatory sequence analysis tools**. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6. [Pubmed 12824373] [Full text] [pdf]

For citing individual tools: the reference of each tool is indicated on top of their query form.

RSAT logos designed by Mauricio Guzman (http://www.altamirastudio.com/)

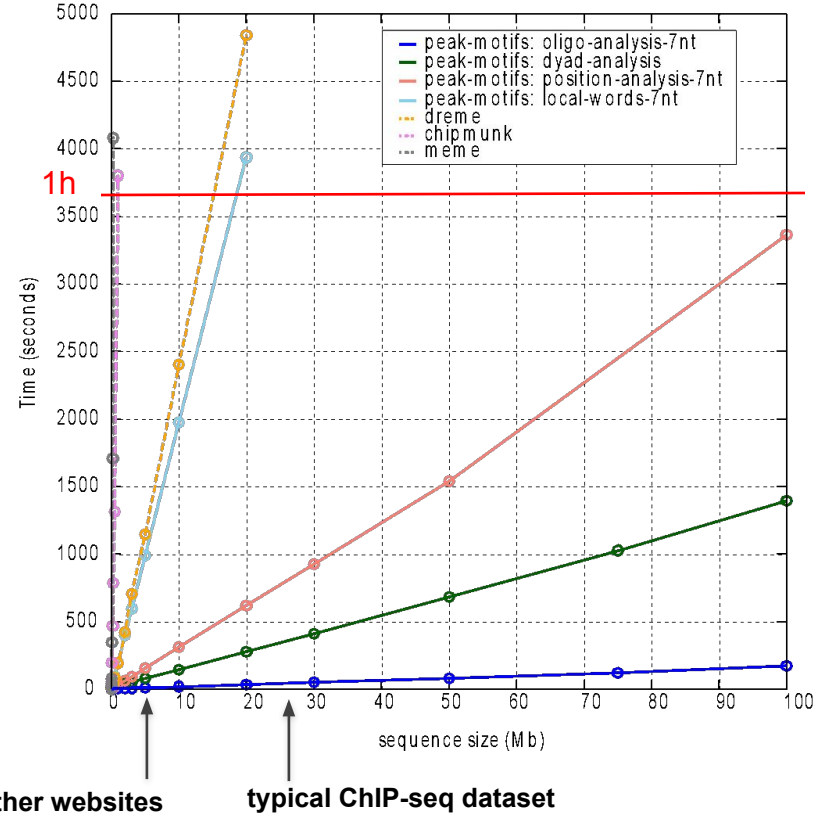
Peak-motifs

- fast and scalable
- treat full-size datasets
- complete pipeline
- web interface
- accessible to non-specialists



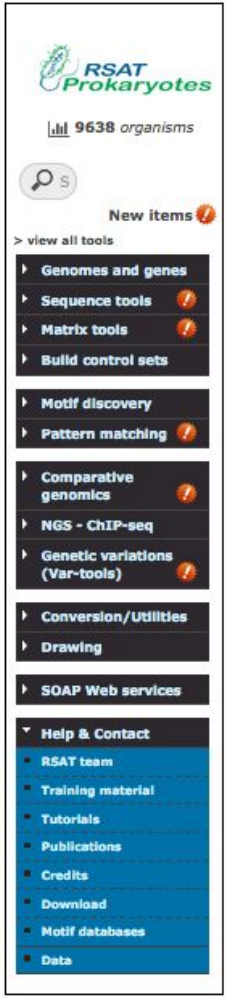
Peak-motifs: why providing yet another tool?

- **fast and scalable**
- **treat full-size datasets**
- **using 4 complementary algorithms**
  - **Global over-representation**
    - **oligo-analysis**
    - **dyad-analysis (spaced motifs)**
  - **Positional bias**
    - **position-analysis**
    - **local-words**



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

RSAT menu



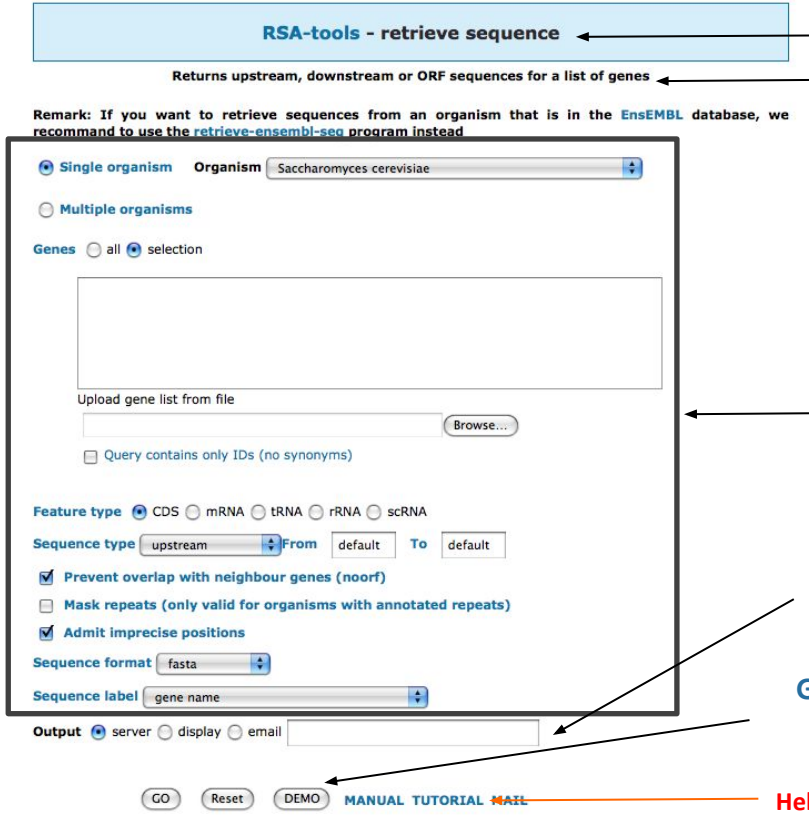
1. Get sequences

2. Run the analysis

3. Visualization

Help: tutorials,

RSAT Web forms



Tool name

Tool description

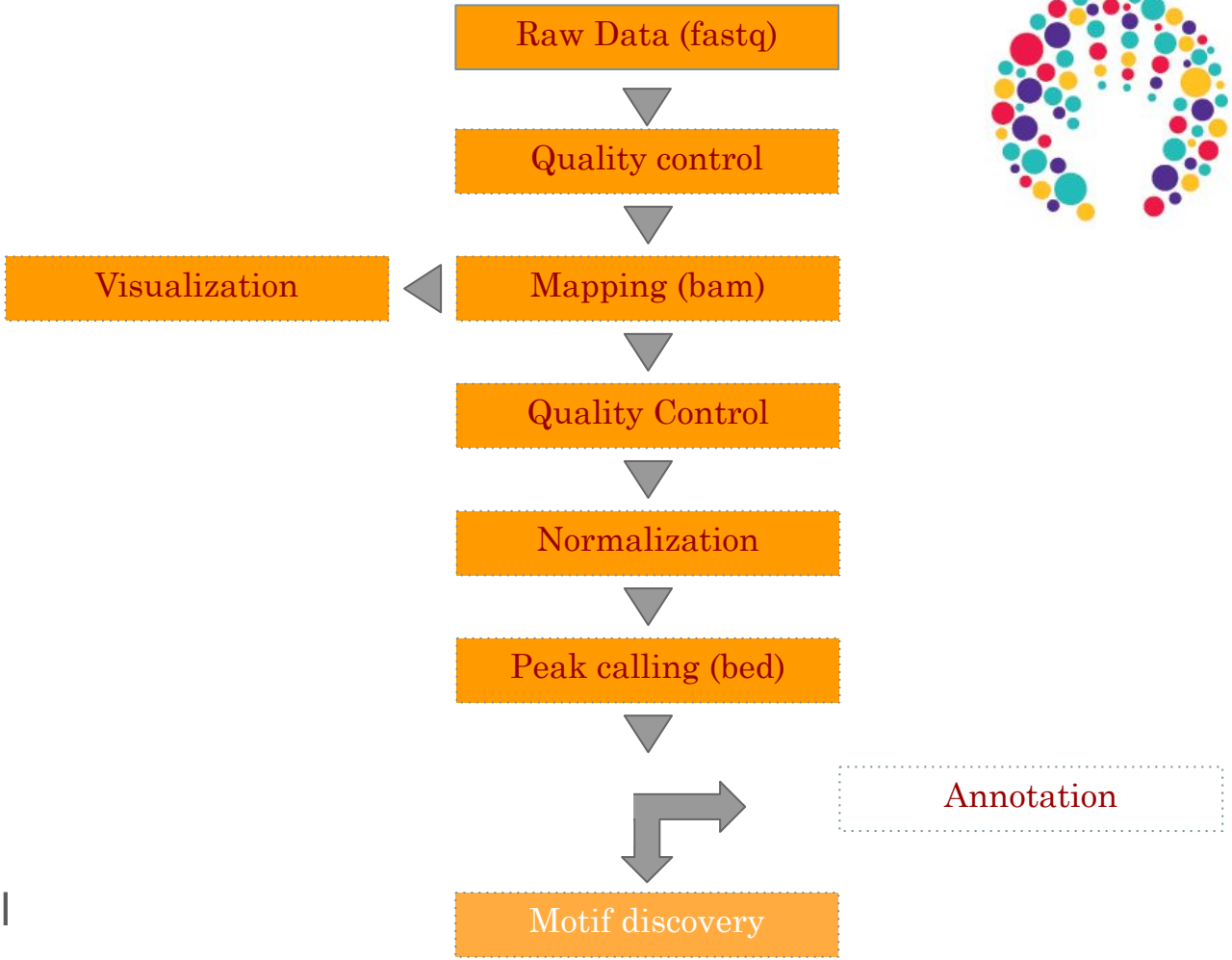
Tool parameters

Output

Go button (launches the analysis)
Demo button (fill in the form for test purposes)

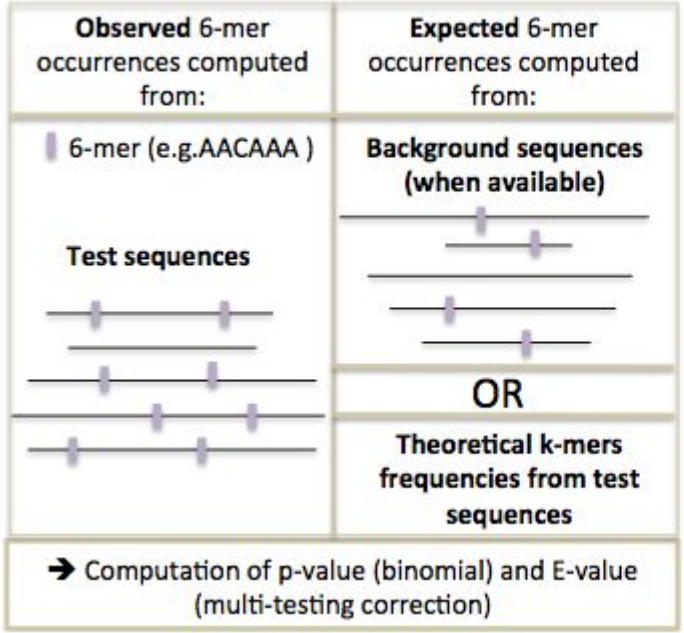
Help

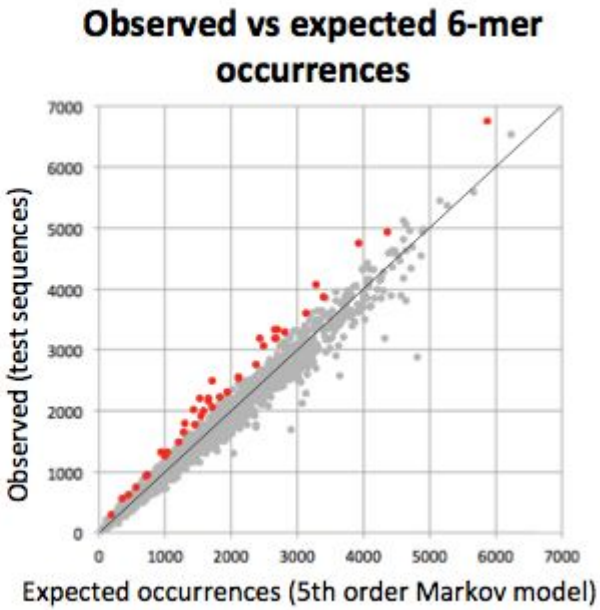
Protocol



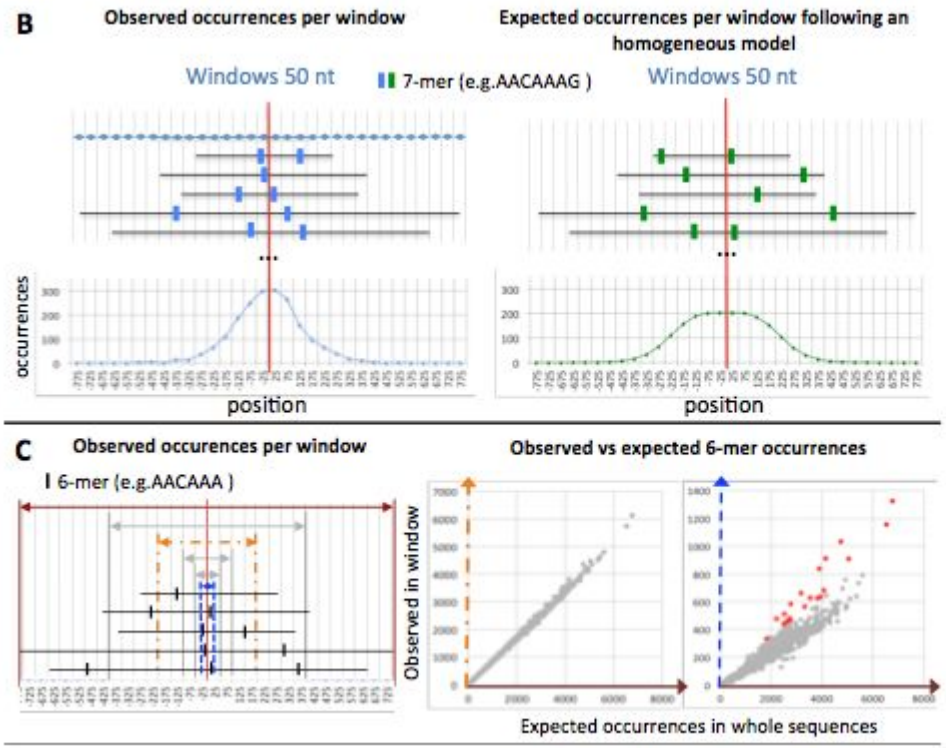
- Motif analysis
- Motif_analysis_practical

Motif discovery: frequency





Motif discovery: positional bias

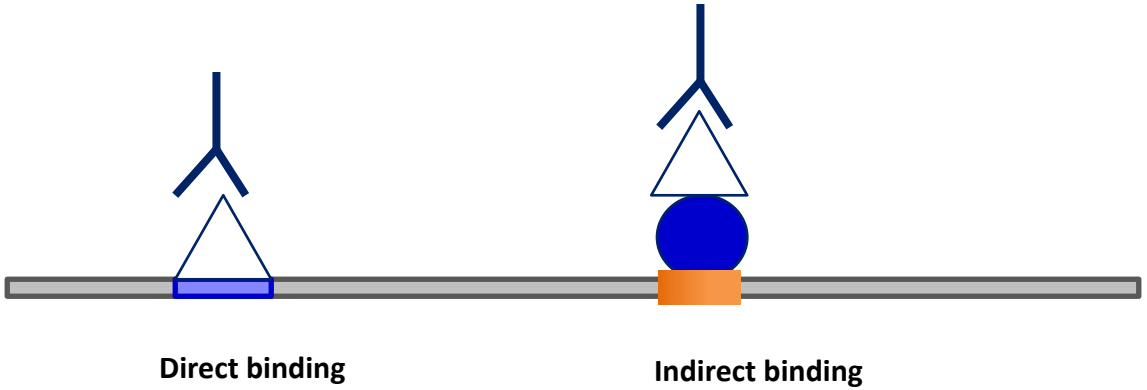


position-analysis

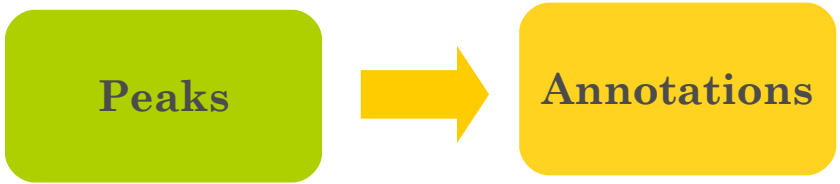
local-words

Direct versus indirect binding

ChIP-seq does not necessarily reveal **direct binding:** The motif of the targeted TF is not always found in peaks!

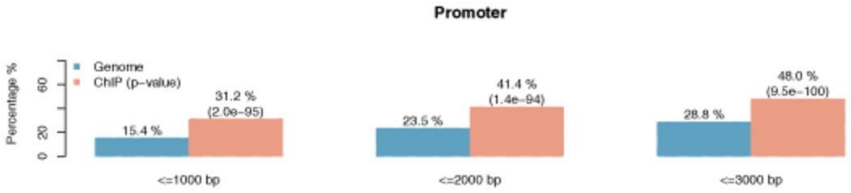


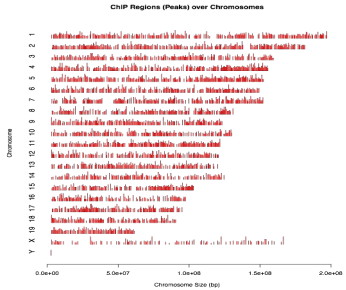
Annotating peaks



Are peaks biased towards any genomic features?

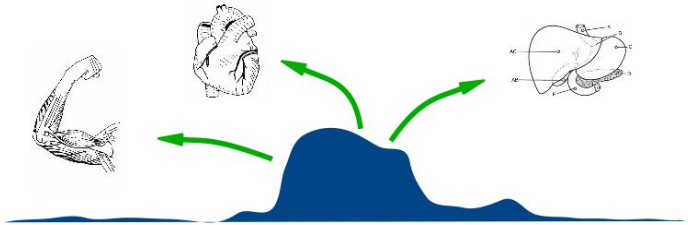
- How are the peaks distributed on the chromosomes?
- Are there genomic features (promoters, intergenic, intronic, exonic regions) enriched in the peaks?
- How are the peaks distributed compared to gene structures (TSS, TTS, introns, exons)?
- How are they distributed compared to the genes?





What are the genes associated to the peaks ?
Are there some functional categories over-represented ?

ChIP-seq peaks → Genes → Ontology terms



GO Molecular Function
GO Biological Process
Disease Ontology
Pathways
…

Various tools available

These tools work with regions (BED files)

- **PAVIS** : https://manticore.niehs.nih.gov/pavis2/
- **GREAT** : http://great.stanford.edu/public/html/

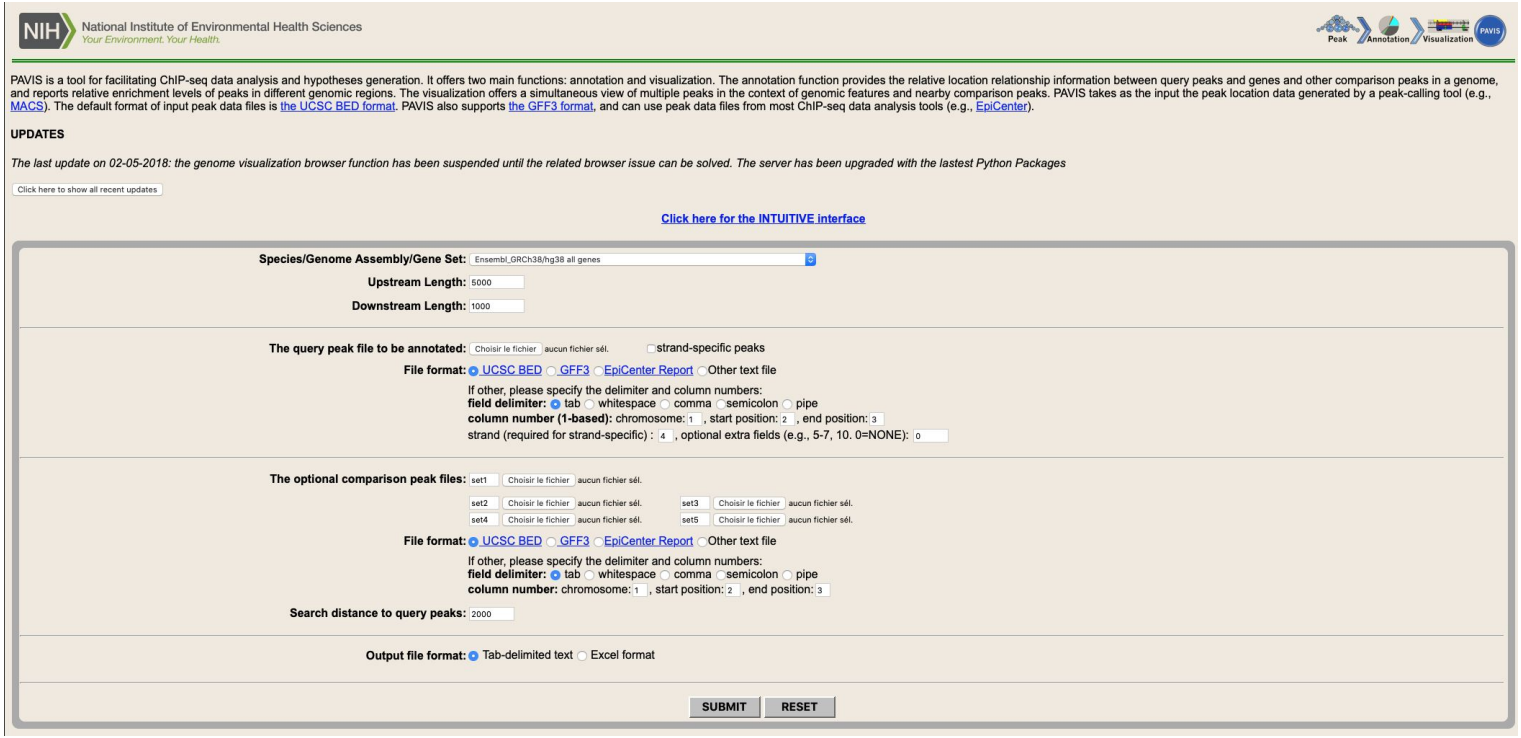
These tools work with gene lists

- **GSEA**: http://www.broadinstitute.org/gsea
- **QuickGO**: https://www.ebi.ac.uk/QuickGO/annotations
- **DAVID:** http://david.abcc.ncifcrf.gov

Warning : rely on the organism annotation and assembly version

=> not all organisms supported by all programs !

PAVIS



Note: mostly model organism genomes supported (human, mouse, rat, fly, zebrafish, worm, yeast, cow, dog, plants)

PAVIS

PAVIS Annotation Report

the Visual Locus Explorer is no longer supported

Your loci have been annotated and stored in the Tab-Separated ASCII file below. With the annotation parameters you specified, 543 of 657 (82.65%) of the loci were successfully associated with genes.

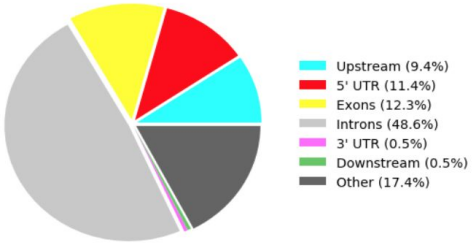
Peak Location Annotation (The Full Annotation File 39.89 kB)

Location	Query Peak	Number	Proportion	EnrichTest1	EnrichTest2	Comparison Peak
Upstream	Q-Upstream	62	9.4%	1.00e+00	9.97e-01	C-Upstream
5' UTR	Q-5UTR	75	11.4%	1.01e-64	6.22e-76	C-5UTR
Exons/CDS	Q-Exon	81	12.3%	1.96e-03	8.71e-09	C-Exon
Introns	Q-Intron	319	48.6%	NA	1.00e+00	C-Intron
3' UTR	Q-3UTR	3	0.5%	1.00e+00	9.96e-01	C-3UTR
Downstream	Q-Downstream	3	0.5%	1.00e+00	1.00e+00	C-Downstream
Unclassified	NA	114	17.4%	NA	NA	NA

The tab delimited form of the table

Note:Upstream length was set to 5000 and Downstream length was set to 1000 (0=no limit).

Distribution of Peaks in Relation to Genes



Click here to download the Pie-Chart in the PDF format

Click here to download the corresponding Bar-Chart in the PDF format

GREAT

Species Assembly

- Human: GRCh37 (UCSC hg19, Feb/2009)
- Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
- Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
- Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set

Can I use a different species or assembly?

Test regions

- BED file: Choisissez un fichier Aucun ...hoisi
- BED data:

What should my test regions file contain?
How can I create a test set from a UCSC Genome Browser annotation track?

Background regions

- Whole genome
- BED file: Choisissez un fichier Aucun ...hoisi
- BED data:

When should I use a background set?
What should my background regions file contain?

Association rule settings

Show settings »

Submit Reset

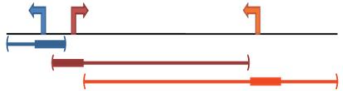
Note: Only human (hg19), mouse (mm9, mm10) and zebrafish (danRer7) genomes are supported

GREAT

Associating genomic regions with genes

GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two step process. First, every gene is assigned a regulatory domain. Then, each genomic region is associated with all genes whose regulatory domain it overlaps.

Basal plus extension



Proximal: 5.0 kb upstream, 1.0 kb downstream, plus Distal: up to 1000.0 kb

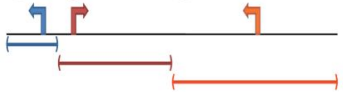
Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

Two nearest genes

within 1000.0 kb

Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction.

Single nearest gene



within 1000.0 kb

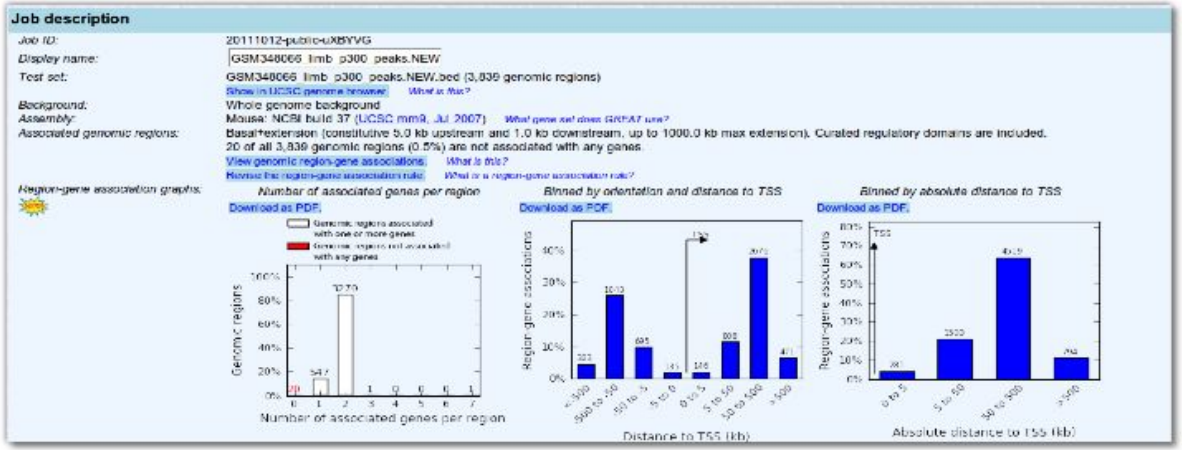
Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

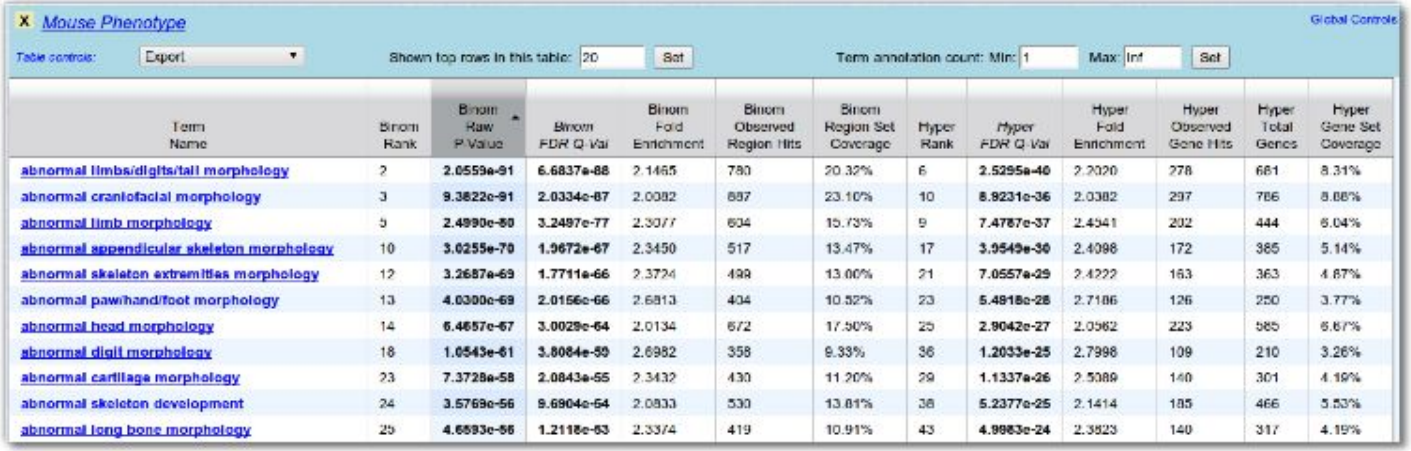
Gene Transcription Start Site (TSS)

Note: Only human (hg19), mouse (mm9, mm10) and zebrafish (danRer7) genomes are supported

GREAT

- Input
  - bed file with peaks
- Output
  - Enriched GO terms and functions





Mouse Phenotype

Global Controls

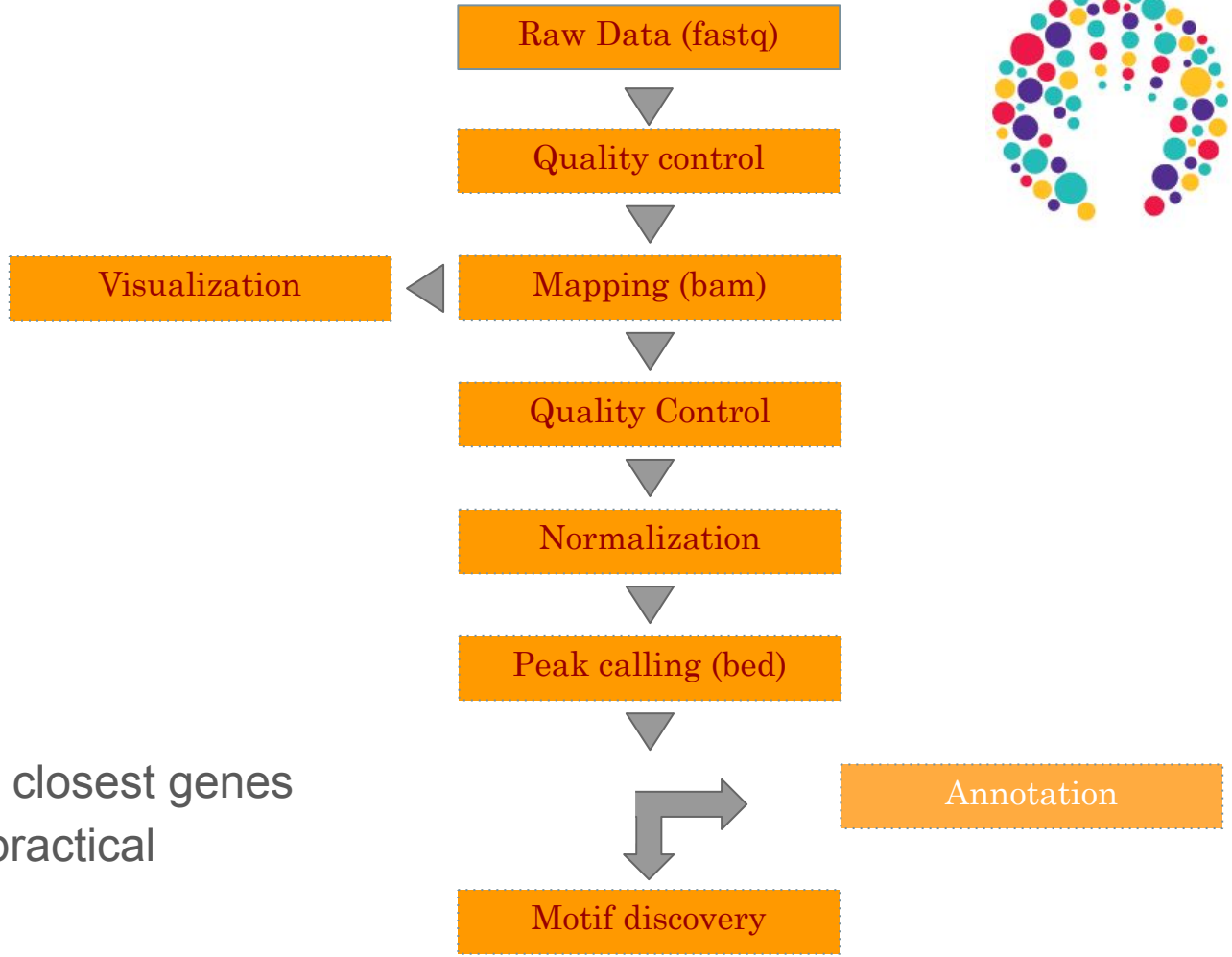
Table controls: Export

Shown top rows in this table: 20 Set

Term annotation count: Min: 1 Max: Inf Set

Term Name	Binom Rank	Binom Raw P Value	Binom FDR Q Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
abnormal limbs/digits/tail morphology	2	2.0559e-91	6.6837e-88	2.1465	780	20.32%	6	2.5295e-40	2.2020	278	681	8.31%
abnormal craniofacial morphology	3	9.3822e-91	2.0334e-87	2.0082	887	23.10%	10	8.9231e-36	2.0382	297	786	8.88%
abnormal limb morphology	5	2.4990e-80	3.2497e-77	2.3077	604	15.73%	9	7.4787e-37	2.4541	202	444	6.04%
abnormal appendicular skeleton morphology	10	3.0255e-70	1.9672e-67	2.3450	517	13.47%	17	3.9549e-30	2.4098	172	385	5.14%
abnormal skeleton extremities morphology	12	3.2687e-69	1.7711e-66	2.3724	499	13.00%	21	7.0557e-29	2.4222	163	363	4.87%
abnormal paw/hand/foot morphology	13	4.0300e-69	2.0156e-66	2.6813	404	10.52%	23	5.4918e-28	2.7186	126	250	3.77%
abnormal head morphology	14	6.4657e-67	3.0029e-64	2.0134	672	17.50%	25	2.9042e-27	2.0562	223	585	6.67%
abnormal digit morphology	18	1.0543e-61	3.8084e-59	2.6982	358	9.33%	36	1.2033e-25	2.7998	109	210	3.26%
abnormal cartilage morphology	23	7.3728e-58	2.0843e-55	2.3432	430	11.20%	29	1.1337e-26	2.5089	140	301	4.19%
abnormal skeleton development	24	3.5769e-56	9.6904e-54	2.0833	530	13.81%	38	5.2377e-25	2.1414	185	466	5.53%
abnormal long bone morphology	25	4.6593e-56	1.2118e-53	2.3374	419	10.91%	43	4.9963e-24	2.3823	140	317	4.19%

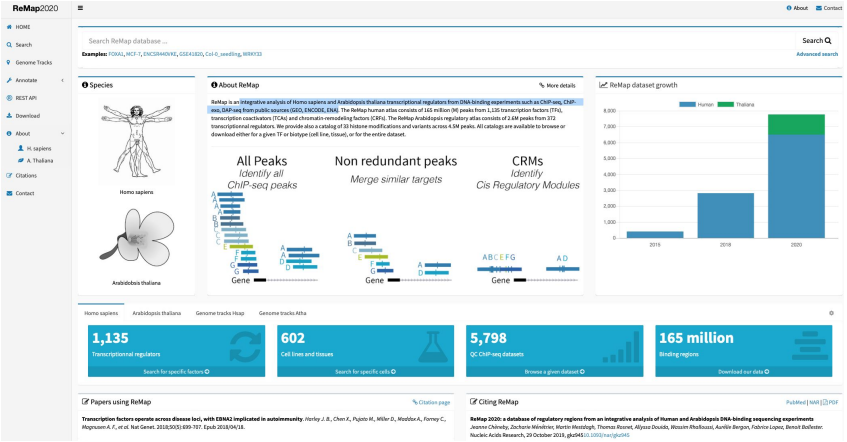
Protocol



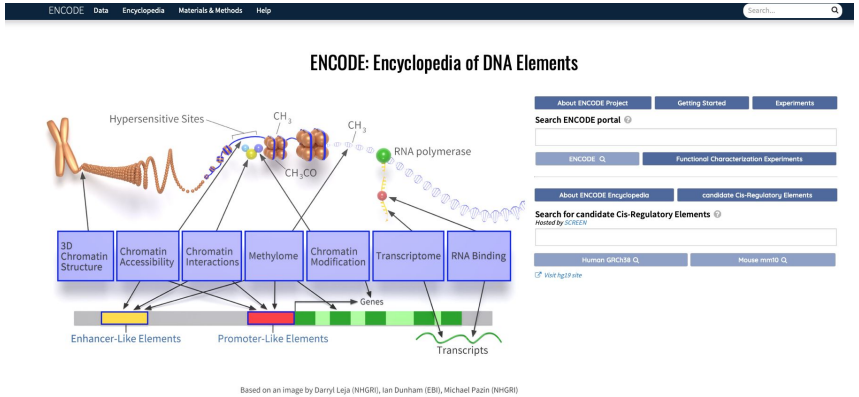
- Associate peaks to closest genes
- Peak_annotation_practical

Other related resources

- ReMAP
  - Is my peak dataset enriched for known TF peaks ?
  - Integrative analysis of H. sapiens and A. thaliana transcriptional regulators from public DNA-binding experiment sources
  - http://remap.univ-amu.fr

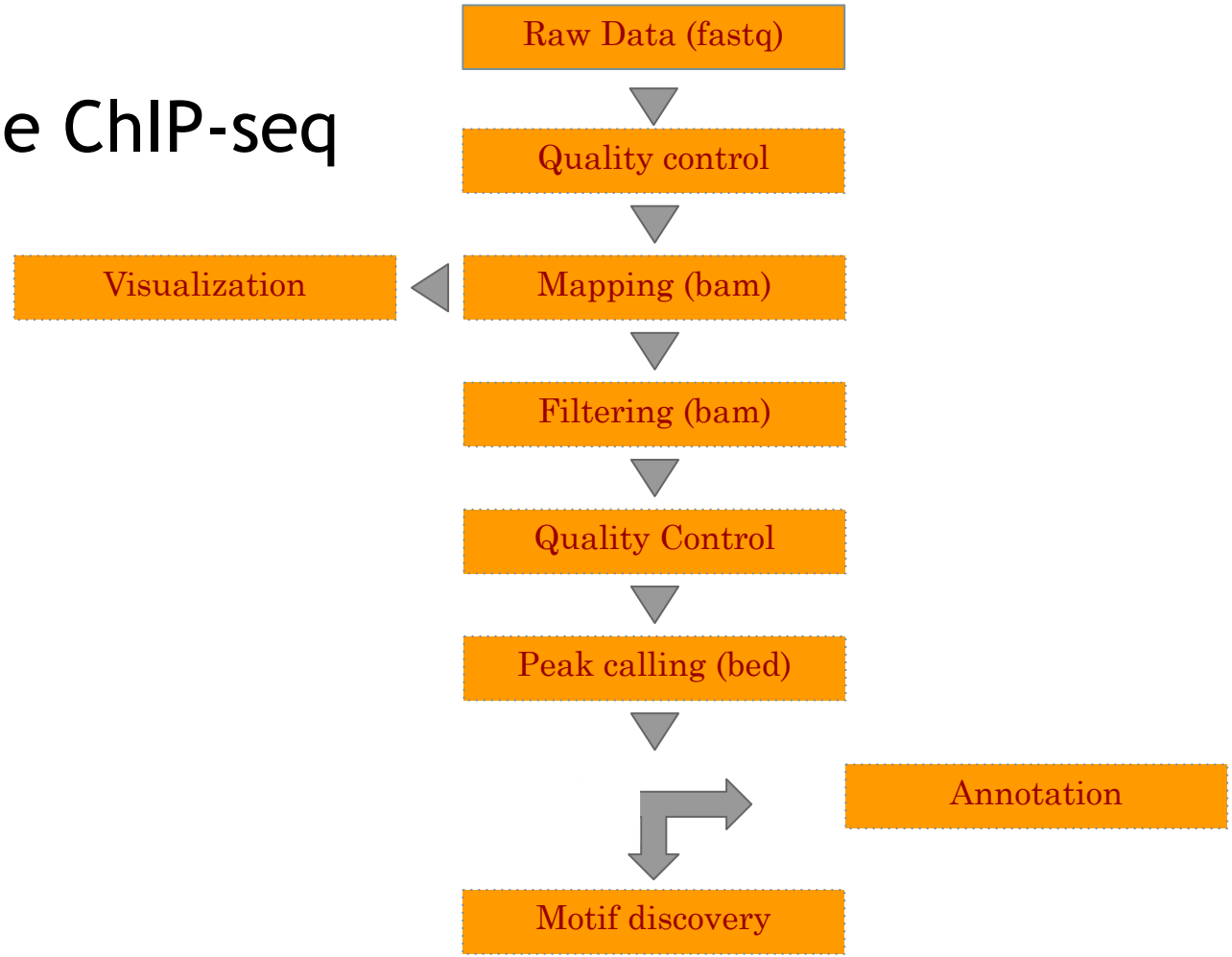


- ENCODE encyclopedia of DNA elements
  - Genomic and transcriptomic annotations
  - https://www.encodeproject.org



Conclusions analyses Chip-Seq

Bilan du pipeline ChIP-seq



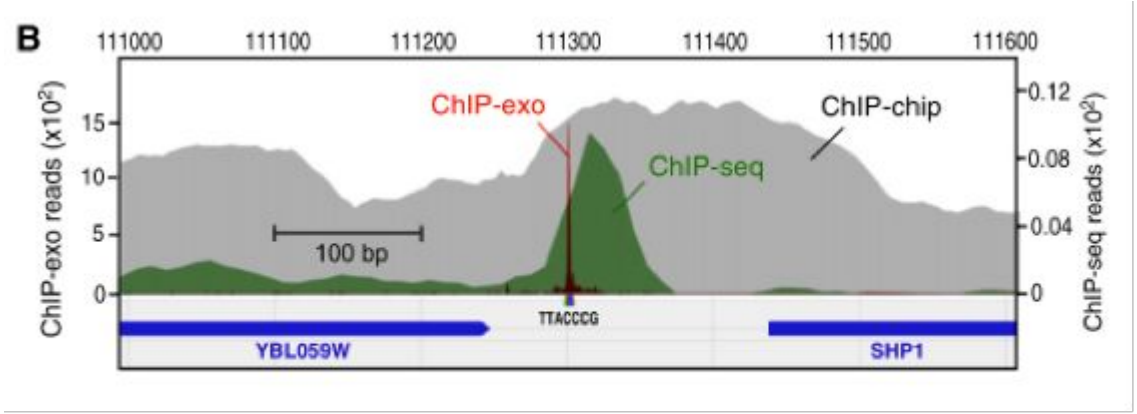
Beyond ChIP-seq : ChIP-exo


crosslink


sonication


antibody


exonuclease



Rhee, H. S., & Pugh, B. F. (2011).

Beyond ChIP-seq

Experimental techniques

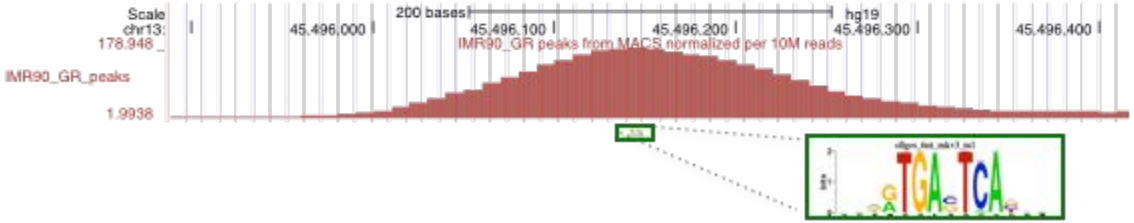
crosslink

sonication

antibody

Improvement aimed

higher resolution => 300bp to 1bp



Beyond ChIP-seq : ChIP-nexus

Experimental techniques


crosslink


sonication

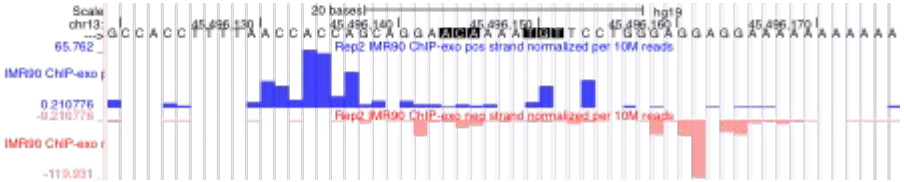

antibody


exonuclease

barcode

Improvement aimed

Get rid of PCR artifacts



Beyond ChIP-seq : native ChIP

Experimental techniques

~~crosslink~~

~~sonication~~

antibody

endonuclease

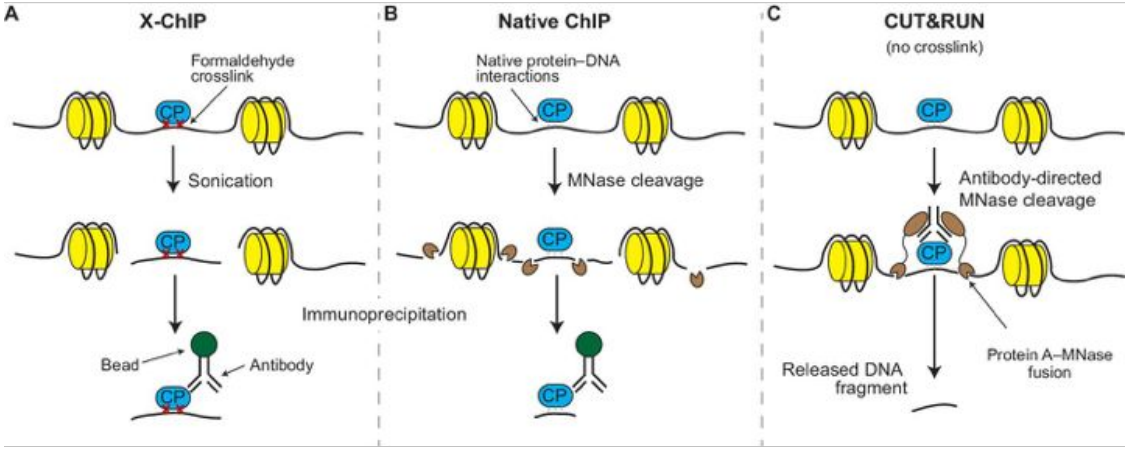
Improvement aimed

Avoid formaldedyde crosslinking

- Formaldehyde crosslinking affects preferentially protein-protein interactions.
- Crosslinking could be the cause of hyper-signaling regions in highly transcribed sites.

Beyond ChIP-seq : native ChIP

= ChIP-seq



No crosslink

No crosslink

CUT&RUN uses the antibodies to guide the cutting activity of the MNase enzyme rather than physically separate wanted from unwanted chromatin fragments

He and Bonasio, Elife, 2017

Beyond ChIP-seq : low-input and single-cell

Experimental techniques

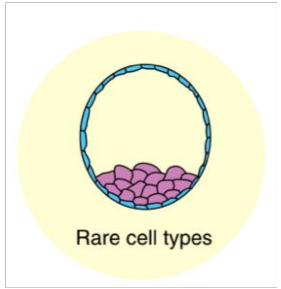

crosslink


sonication


antibody

Improvement aimed

Reduce the amount of starting material (precious samples)



- Low-input: Optimized ChIP-seq protocols => 100-500 cells

Dahl & Gilfillan, Briefings in Functional Genomics, 2017

- Single-cell ChIP-seq : Only one proof-of-concept study, very low coverage

Rotem et al, Nature Biotechnology, 2015

More recent proof-of-concept

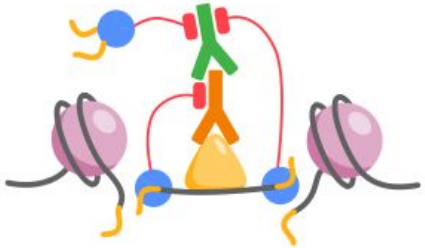
Grosselin et al, Nature Genetics, 2019

Beyond ChIP-seq : Cut&TAG (2019)

CUT&RUN
Cleavage under targets and release using nuclease

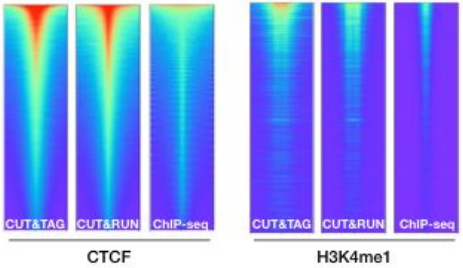
- Cleave adjacent DNA by MNase
- No crosslinking
- Low background

CUT&TAG
Cleavage under targets and tagmentation



- Cleave near antibody site by Tn5
- No crosslinking
- Low background
- Include adaptor ligation
- Adapted for single-cell

Signal profiling at equal read depth from Kaya-Okur *et al.*, 2019



- Antibody to target protein
- Protein A (pA)
- Micrococcal Nuclease (MNase)
- Anti-rabbit antibody (increase pA tethering)
- Hyperactive transposase 5 (Tn5) with adaptors

Beyond ChIP-seq : Cut&TAG (2019)

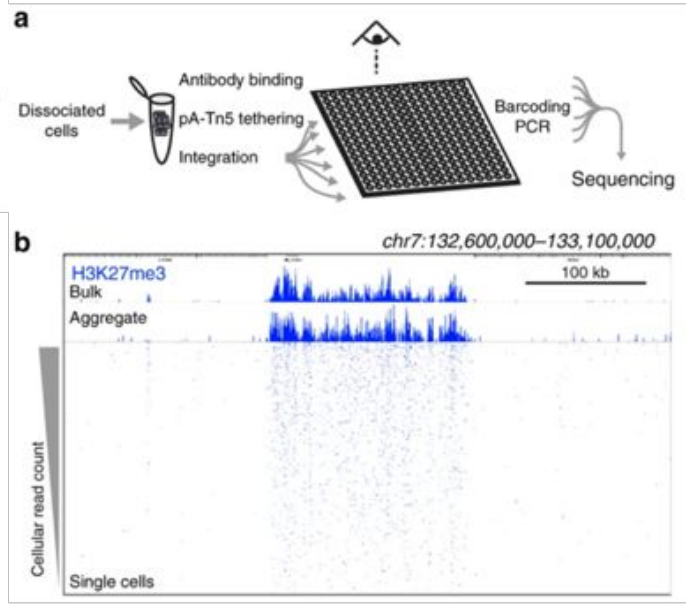

nature COMMUNICATIONS

Article | Open Access | Published: 29 April 2019

CUT&Tag for efficient epigenomic profiling of small samples and single cells

Hatice S. Kaya-Okur, Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad & Steven Henikoff

Nature Communications **10**, Article number: 1930 (2019) | Download Citation



Low background => 3 Million reads sufficient for human….